Universidad de los Andes
Facultad de Economía

Presentado por: Angie Ariza, Daniel Sandoval,
Andrés Díaz, Javier Farfán

Big Data y Machine Learning para Economía Aplicada

Taller 1
2023-02

Link del repositorio en Github: https://github.com/afdiaz27/Taller_1.git

---

Presentado por

Angie Ariza Quitian - 201325848
Andrés Felipe Diaz Barreto - 200610686
Duvan Javier Farfán López – 201317299
Daniel Eduardo Sandoval - 200712968

Introducción

Colombia es considerado un país de ingreso medio. Sin embargo, existe una enorme desigualdad como consecuencia de la disparidad social en los hogares colombianos de y aspectos socioeconómicos como el acceso a servicios básicos (salud, educación, acueducto, etc.) y por supuesto en el ingreso de las familias. Actualmente, Colombia, al igual que en varios países alrededor del mundo, empieza notar la inversión de la pirámide poblacional, lo que significa el envejecimiento de sus habitantes y con esto desafíos frente a los aspectos laborales cuando el 56% de la población se encuentra en la informalidad, según el DANE para julio de 2023, mientras que, en Bogotá, la informalidad se encuentras en el 33%.

Por otra parte, el salario mínimo mensual legal vigente es de $1’116.000, a pesar de esto, muchos individuos y familias, particularmente, las que se encuentran en la informalidad ni siquiera logran generar un salario minino. Estas personas y familias se encuentran en niveles críticos de vulnerabilidad al no cotizar pensión, además de depender del régimen subsidiado de salud.

Este trabajo, se centra particularmente en el ingreso de los hogares en la ciudad de Bogotá, de acuerdo con la información de la Gran Encuesta Integrada de Hogares del año 2018, realizada por el DANE, con el objetivo de analizar el comportamiento de los salarios en relación con la edad, el sexo, entre otras variables sociodemográficas, para estudiar la retribución de la educación sobre el mismo, las brechas salariales de género y otras variables que pueden afectar el ingreso de individuos y familias.

En este ejercicio se realizó el scrapping de la base de datos, limpieza de la misma, un análisis descriptivo de las variables de interés; la estimación del perfil de salario contra la edad; la estimación de la brecha salarial de género y se aplicaron técnicas de verificación del error como lo son el enfoque de conjunto de validación (validation set approach) y LOOCV.

Universidad de los Andes
Facultad de Economía

Presentado por: Angie Ariza, Daniel Sandoval,
Andrés Díaz, Javier Farfán

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.

Universidad de los Andes
Facultad de Economía

Presentado por: Angie Ariza, Daniel Sandoval,
Andrés Díaz, Javier Farfán

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), retomado por Mahnic (2022) 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, Aristizábal & Ángel

Variable	Nombre de la variable	Descripción de la variable	Justificación de uso
			(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. Para Manhic (2022), como lo indica 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.

Variable	Nombre de la variable	Descripción de la variable	Justificación de uso
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.

Variable	Nombre de la variable	Descripción de la variable	Justificación de uso
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)

Variable	Nombre de la variable	Descripción de la variable	Justificación de uso
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

Variable	Nombre de la variable	Descripción de la variable	Justificación de uso
			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

Variable	Nombre de la variable	Descripción de la variable	Justificación de uso
			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 2 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 3 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 4 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).

Adicional a estas estadísticas descriptivas, se generó un summary para todas las variables, sin embargo, se debe notar que el estudio de esta tabla para variables dicótomas no tiene sentido, pues el mínimo es cero y el máximo 1, y la media se ve afectada por la cantidad de unos, sin reflejar información relevante. Por lo anterior, generaron tablas de frecuencia que permitieran ver la distribución de datos de las variables dicótomas. Los resultados de las mismas se pueden ver en la tabla 6.

Tabla 5 Tabla de frecuencia para variables dicótomas de interés

Variable	Descripción	0	1
Female	Hombre = 0 Mujer = 1	7.494	6.792
Clase	Rural = 0 Urbano = 1	-	14.286
Cuentapropia	Asalariadas u otro = 0 Independiente =1	9.994	4.292

La teoría consultada especifica la relación del sector (rural o urbano), influye en la diferencia de salario, sin embargo, la tabla de frecuencias para la variable *clase* muestra que todas las observaciones corresponden al sector urbano de Bogotá, por lo cual está variable deja de ser relevante para el estudio. Así mismo, se puede observar que la base de datos que se usara tiene menor número de mujeres, representando un 47,5% del total de la muestra. También se tiene un 70% asociado a personas asalariadas o con otras formas de relación laboral y 30% a personas independientes, lo que permite pensar en que se podrá hacer una diferenciación de los ingresos por el tipo de vínculo o relación laboral que tengan las personas. Se resalta está ultima variable, pues permitirá tener una aproximación al entendimiento de la retribución de personas asociadas a empresas u otros, así como la diferencia de personas que optan por tener trabajos que dependen exclusivamente de ellos. Según la teoría, esto también puede ser correlacionado con el nivel de informalidad asociado a la actividad laboral, por lo que el entendimiento de estas brechas puede ayudar a la generación de políticas públicas orientadas a garantizar el bienestar de los trabajadores en cada una de las relaciones laborales que tengan.

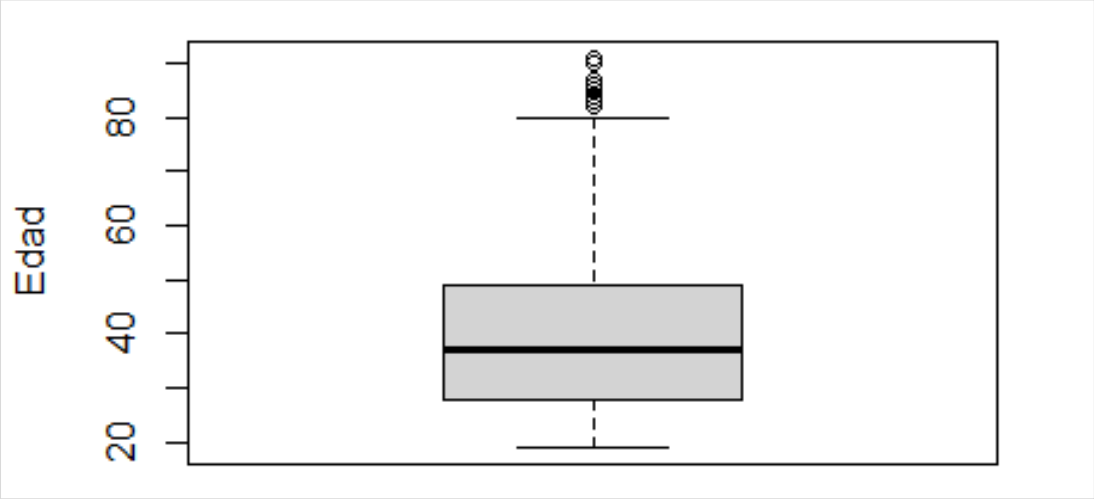
Análisis de la Edad

Una de las variables de mayor interés por la propuesta del ejercicio y su relevancia en los estudios econométricos de la literatura es la edad, por lo que se realizaron un boxplot y un histograma para verificar cual era la distribución de los individuos de la muestra. En la figura 1 se puede ver el boxplot, el cual confirma la información extraída con el comando summary. En esta, se puede observar como la mayoría de las observaciones se concentra entre los 30 y 45 años aproximadamente, representando toda la fuerza laboral y manteniendo la tendencia asociada a las leyes de retiro y jubilación del país. Sin embargo, también se presentan personas entre el tercer y cuarto cuartil, indicando que siguen trabajando y recibiendo una remuneración.

Dependiendo del tipo de análisis y con un análisis mucho más profundo, esto podría servir para tomar decisiones sobre el incremento de la edad de pensión, o para evidenciar las falencias en el sistema, pues dichas observaciones pueden pertenecer a personas que nunca pudieron acceder a un sistema de seguridad social, y por eso se ven obligadas a trabajar en edades avanzadas. El boxplot de la figura 1 no permite despreciar todas las observaciones de individuos entre los 45 y 80 años. Estas observaciones, además, permitirán corroborar la ecuación y teoría de

Mincer, en la que se argumenta que, alcanzado el máximo de retribución de salario, este tiende a disminuir por cuestiones de capacidad y productividad, así como de condiciones de retiro y obsolescencia de conocimiento, especialmente en los últimos años, donde los avances tecnológicos superan la velocidad de aprendizaje de las personas.

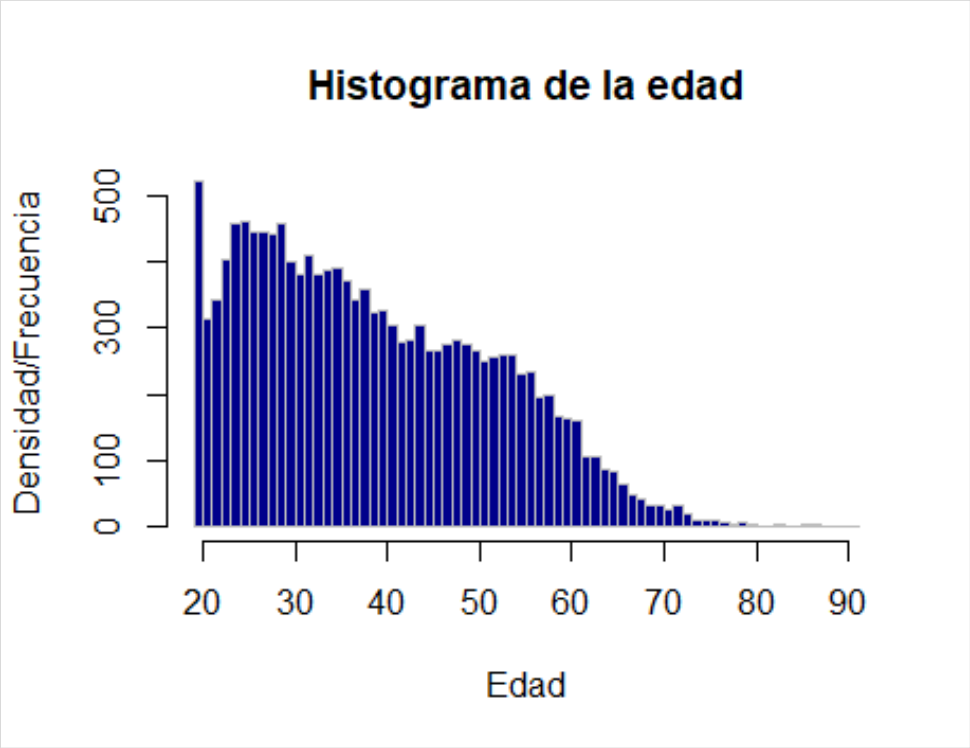
Figura 1 Boxplot de la edad



Para revisar la distribución de la edad en la muestra se realizó un histograma, el cual se puede ver en la Figura 2. Dado que la base de datos se sesgo por la omisión de personas menores a 18 años, se puede ver como la distribución está recargada hacia la izquierda. Sin embargo, esto no sería muy diferente si no se truncara la información, pues los menores de edad no deberían ejercer actividades laborales y la naturaleza de la GEIH 2018 es tener información de personas habilitadas para trabajar. En futuras ocasiones, y con la disponibilidad de la información, poseer todo el espectro de información laboral, incluso en menores de edad, podría representar el entendimiento del trabajo infantil en Colombia.

Al igual que la salida de estadísticas descriptivas y el boxplot de la figura 1, en el histograma de la edad se puede apreciar como hay mayor densidad de observaciones en personas con edades entre los 25 y 50 años. Es de esperarse que dad esto, el pico de salario máximo para Colombia este en este rango, tal como lo sugiere la literatura. Dado que las condiciones laborales y de educación en Colombia han cambiado, se espera que dicho pico, con información del 2018, este disminuyendo en comparación con años anteriores, es decir, las facilidades de acceso a la educción pueden hacer que el pico de máximo salario sea alcanzado antes en comparación con generaciones anteriores. Esta premisa será verificada en el inciso 3 de este trabajo.

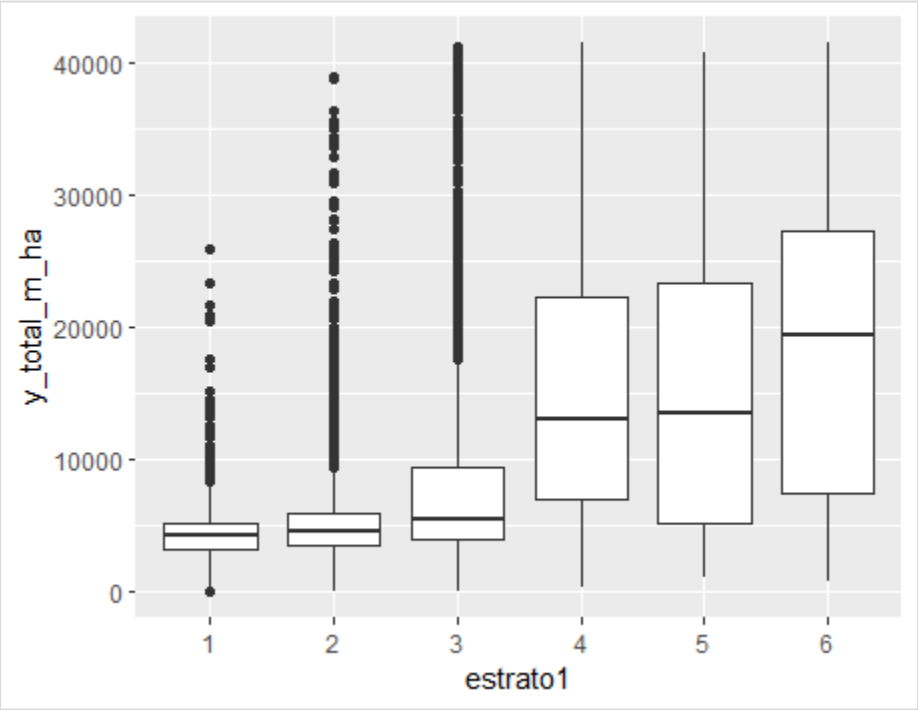
Figura 2 Distribución de la variable edad representada por un histograma



Análisis de las horas trabajadas a la semana y su relación con el estrato

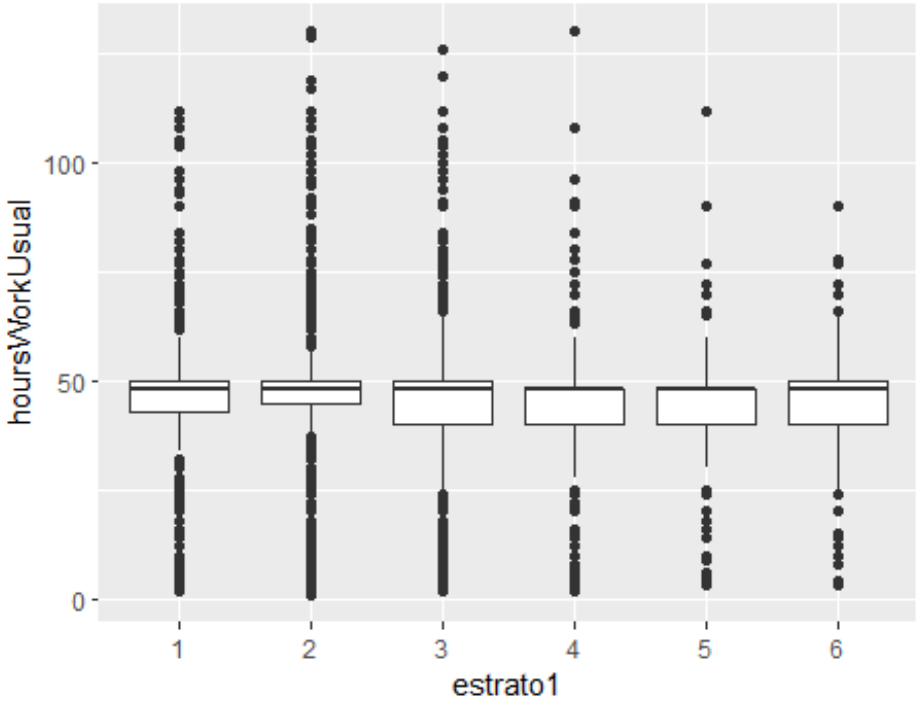
Como se ha mencionado a lo largo de este documento, el estrato despierta diversas opiniones sobre si es una medida justa se segregación de personas por características asociadas a sus perfiles sociodemográficos y los lugares donde se encuentra el lugar de residencia de los mismos. Sin embargo, muchas veces el estrato puede estar asociado a la capacidad adquisitiva de las personas. Por tal motivo, se decidió comparar el estrato con el salario percibido por los individuos incluidos en la encuesta y que son objeto del análisis de este trabajo. En la figura 3, el boxplot de estrato contra salario por hora permite ver como a medida que aumenta el estrato (mejores condiciones económicas, sociales y financieras en teoría), la mediana del salario recibido por hora también aumenta, siendo en los estratos 1, 2 y 3 similar e inferior a 10.000 COP/hora (salario mensual esperado de 1.600.000 COP/mes aproximadamente). También se destaca como los estratos 4 y 5 tienen una mediana similar, pero no muy diferente a la de los primeros estratos. Esto es una justificación a los lineamientos de la DIAN y el DANE en el establecimiento de la categoría de clase alta en Colombia, pues los datos para la ciudad de Bogotá muestran que en Colombia los salarios altos y por encima de los 10 millones no son usuales. El boxplot de la figura 3 permite tener una idea de la relación entre el estrato y el salario, sin embargo, debe ser estudiada con cuidado, pues en una regresión podría presentar endogeneidad, ya que las condiciones de vivienda se ven afectadas por el nivel de ingresos de una persona o de la familia.

Figura 3 Boxplot de comparación entre el salario por hora y el estrato



Partiendo de lo anterior, resulta interesante estudiar la relación entre los estratos, el ingreso percibido y el total de horas trabajadas por semana. Para ello se realizó un boxplot de estrato contra horas trabajas, el cual se puede ver en la figura 4. Dado e establecimiento de jornada laboral para Colombia de 48 horas por semana, se puede apreciar como para todos los estratos la mediana está cercana a ese valor. Sin embargo, al revisar la información de la figura 3, se puede ver que por el mismo tiemo trabajado, los estratos 1, 2 y 3 reciben una menor remuneración, lo que puede ser el resultado de una menor preparación por falta de acceso de educación terciaria e incluso por costumbres sociales. Con estos dos boxplots se podría inferir como la preparación y las condiciones financieras si influyen en el salario, por lo que se justifica la inclusión de las variables en la base de datos para realizar análisis de incisos posteriores. Además de lo anterior, y de acuerdo a la ecuación de Mincer, se podría asociar a los estratos 1, 2 y 3 con una menor productividad, teniendo en cuenta la teoría de capital humano. Está ultima conclusión debe ser revisada a profundidad para evitar caer en una relación espuria que despierte opiniones controversiales en foros públicos.

Figura 4 Boxplot de horas usualmente trabajadas contra estrato



Análisis del salario

El salario es la variable objetivo del problem set, por lo que se realiza el análisis mediante la comparación de bloxplots e histogramas. En el primer bloxblot de la figura 5 muestra cómo se distribuyen los salarios por hora, sin embargo, se nota como hay algunos valores por encima del tercer quartil que estiran el grafico, permitiendo pensar que hay valores atípicos. Cuando se grafica el segundo boxlot que se muestra en la figura 6, se puede ver como esos valores se normalizan y ya no parecen outliers. En este caso, son los valores cercanos a cero los que alteran la distribución de dichos datos. En ese gráfico se corroborar porque en la literatura se prefiere usar el logaritmo del salario para estimar los modelos de regresión que involucran la explicación de dicha variable.

Figura 5 Boxplot del salario por hora

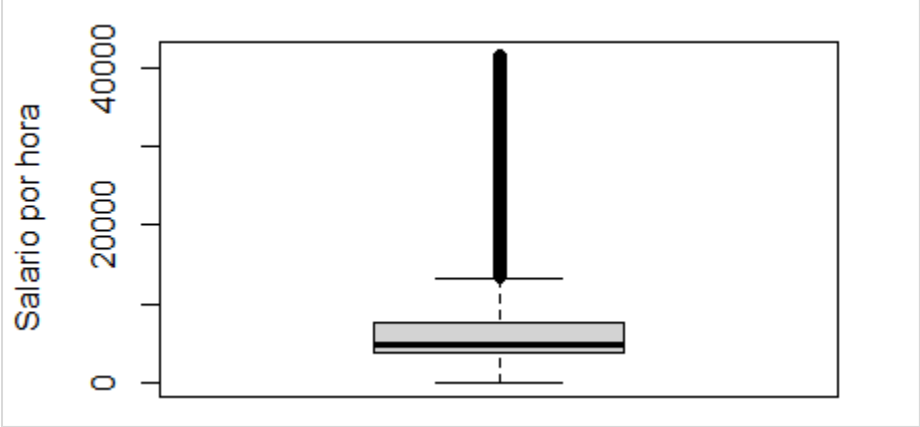
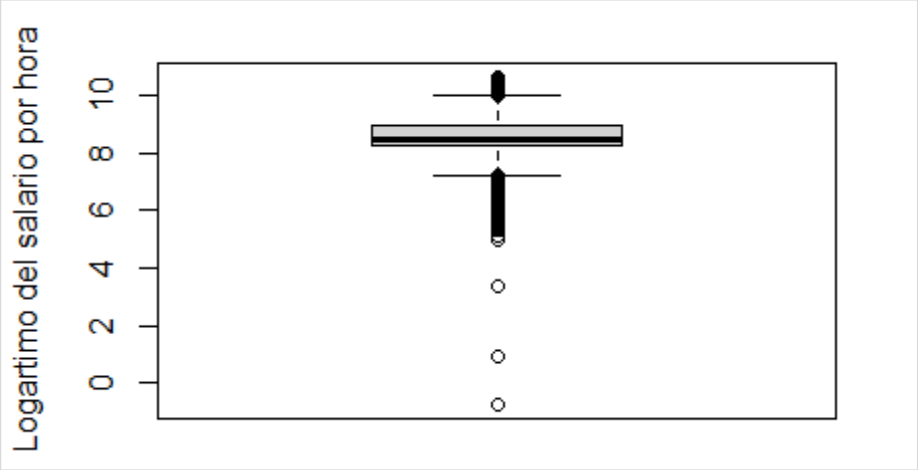
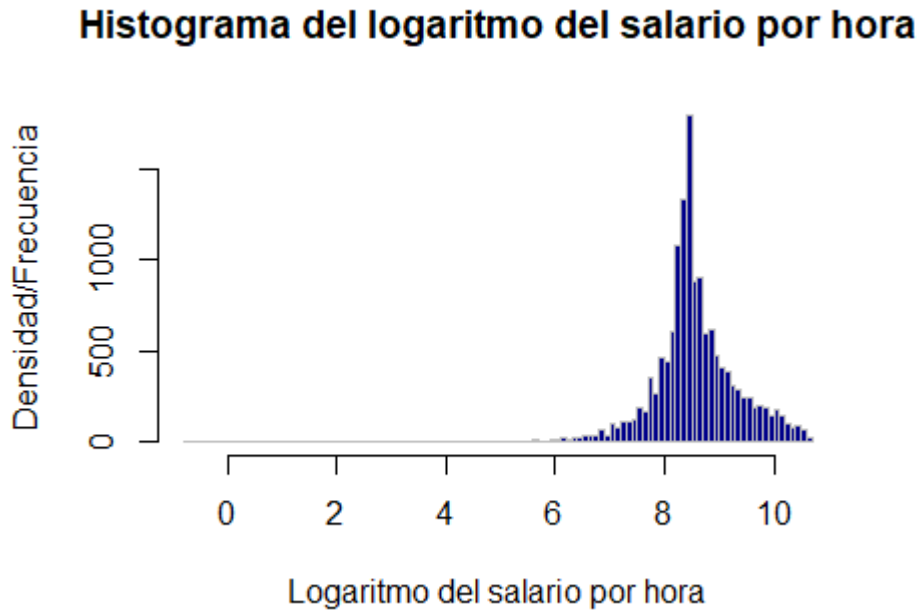


Figura 6 Boxplot del logaritmo del salario por hora



Finalmente, en la figura 7 se puede ver cómo cambia la distribución de observaciones asemejándose más a la normal, sin embargo, tiene una mayor curtosis, lo que la aleja de dicha distribución. Adicional a lo anterior, un mejor tratamiento de los datos cercanos a cero podría mejorar la distribución.

Figura 7 Histograma del logartimo del salario



E

Universidad de los Andes
Facultad de Economía

Presentado por: Angie Ariza, Daniel Sandoval, Andrés Díaz, Javier Farfán

Estimación de perfil edad – salarios

3.1 Regresión

Teniendo en cuenta la regresión realizada entre el logaritmo del salario, la edad y la edad al cuadrado representado en el siguiente modelo:

$$Logaritmo\ del\ salario = \beta_0 + \beta_1 Edad + \beta_2 Edad^2 + u$$

Este modelo permite identificar un aspecto importante sobre el salario, y es el punto de inflexión que puede tener. Al regresar el modelo se obtuvo los siguientes resultados:

```
================================================
                        Dependent variable:
                    ---------------------------
                                y
------------------------------------------------
age                          0.050***
                             (0.003)

age2                         -0.001***
                             (0.00003)

Constant                     7.691***
                             (0.056)

------------------------------------------------
Observations                  14,286
R2                             0.028
Adjusted R2                    0.028
Residual Std. Error      0.734 (df = 14283)
F Statistic          203.246*** (df = 2; 14283)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

Los coeficientes, incluyendo la constante del intercepto son significativos al 99% en el modelo. Así mismo, la incertidumbre del modelo equivale al 1% por el mismo valor de la significancia. En ese sentido, para esta regresión es posible afirmar que el salario depende de la edad. Según Mincer, esta variable puede ser asociada a la experiencia, por lo que se tendría que hacer una evaluación más profunda para eliminar las diferencias que puede tener la educación al momento de la estimación del salario.

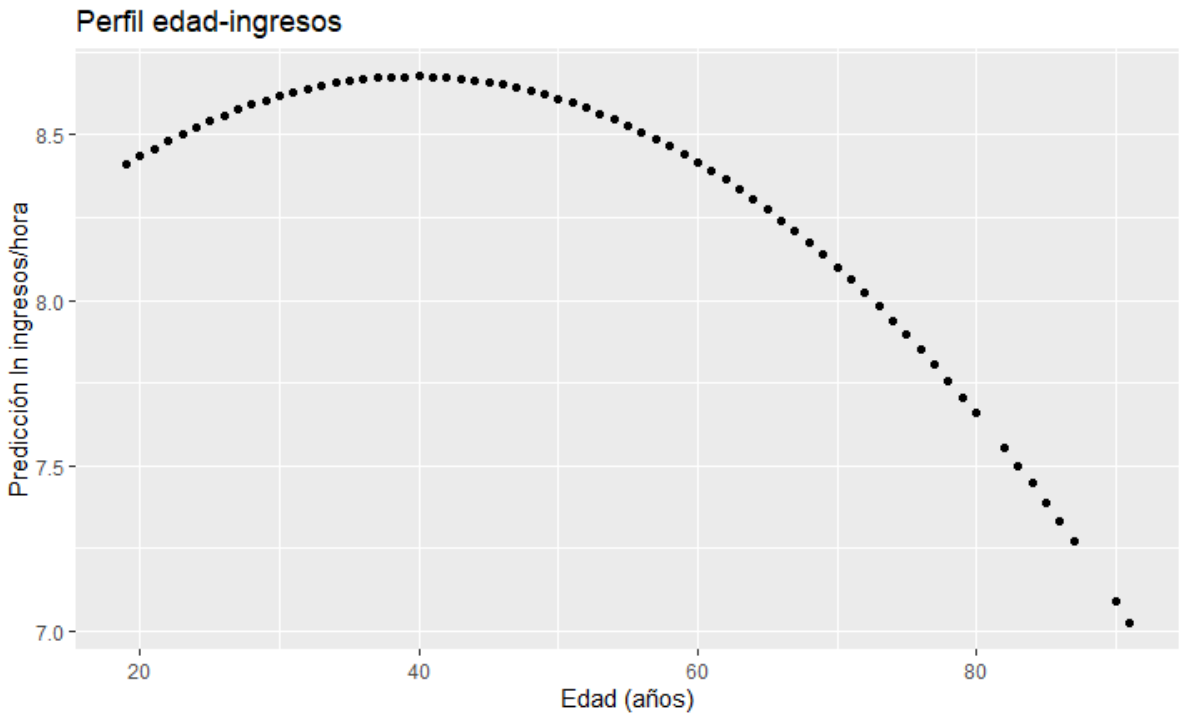
El F estadístico es igualmente significativo al 99%, lo que indicaría que al menos una de las variables explica este modelo. El R^2 y el R^2 ajustado, son pequeños teniendo en cuenta que no se trata un modelo rico en variables independientes. Con un mayor número de variables independientes y controles, el R^2 aumentaría.

Teniendo en cuenta que el modelo está expresado en una función cuadrática, la lectura de los resultados no puede tener en cuenta efectos marginales, ya que el efecto marginal se calcula a partir de la derivada de este modelo propuesto.

Hasta este punto, solo contados con dos coeficientes de β_1 y β_2 de signo positivo para β_1 y negativo para el β_2. Este signo negativo es el signo esperado (característica de una función parabólica cóncava, es decir, que abre hacia abajo), ya que logra confirmar la teoría económica en la cual los individuos logran aumentar su salario, hasta un punto máximo y desde ahí decrece por razones como la aparición de nuevas tecnologías, falta de capacitación, obsolescencia, perdida de la productividad, entre otras ya mencionadas anteriormente.

Presentación perfiles edad-salarios y "edad-pico"

Figura 8 Perfil Edad Salario



La gráfica 8, evidencia como el comportamiento del $\beta_2 Edad^2$ es igual a una función cóncava y corresponde a la predicción del logaritmo del salario y es consistente con la teoría económica. En este modelo, el resultado predicho determina que, en la ciudad de Bogotá, el punto máximo para que un individuo aumente porcentualmente su salario es en promedio a los 39,6 años, desde ese punto, empieza a decrecer.

Para realizar la estimación de los intervalos de confianza se usó Bootstrap, usando un R de 1.000 y una semilla de 1000, obteniendo el siguiente resultado:

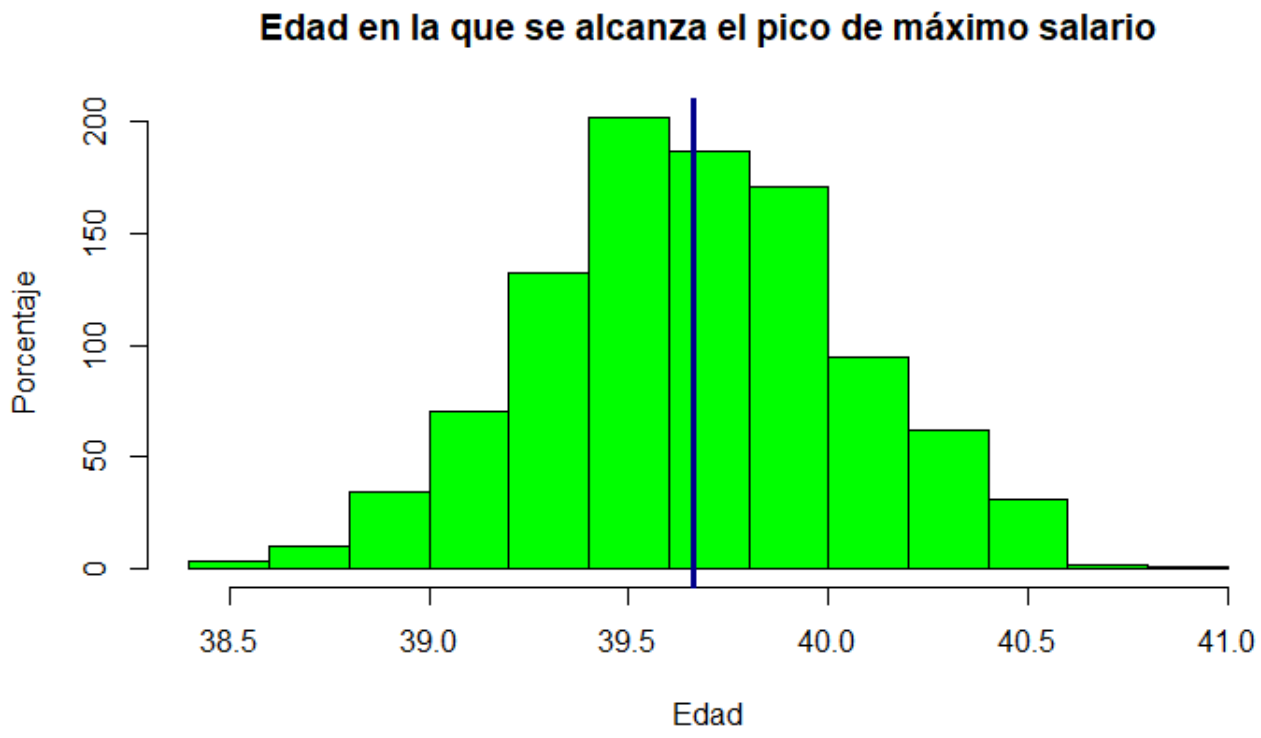
```
Bootstrap Statistics :

       original        bias       std. error
t1*   7.690522730   -2.013065e-04   6.009782e-02
t2*   0.049727512    2.079892e-05   3.220967e-03
t3*  -0.000626812   -4.036597e-07   3.962613e-05
```

Presentado por: Angie Ariza, Daniel Sandoval, Andrés Díaz, Javier Farfán

A partir de los errores estándar se obtuvieron los siguientes intervalos de confianza;

Variable	IC
b0	0.1177917
b1	0.006313095
b2	7.766721e-05

El pico de salario para la ciudad de Bogotá según la información de la muestra se puede ver a continuación:}



La estimación de del pico se hace obteniendo la derivada de la función regresada, teniendo como formula de cálculo que el máximo de salario es -B1/2B2, que como se había expresado antes, con los coeficientes hallados se obtiene un pico a los 39,6 años. Según la literatura, el pico de salario ha venido disminuyendo porque las personas tienen mayor facilidad de acceder a la educación de calidad. Esto puede representar un problema, puesto que el ahorro con el tiempo decrece y se traduce en personas de la tercera edad empobrecidas.

Estimación de la brecha salarial de género e interpretación de las estimaciones

Para estimar la brecha de salario incondicional de género, se utilizó el siguiente modelo:

$$\log(w) = \beta_0 + \beta_1 Female + \mu,$$

donde log(w) = logaritmo natural del salario por hora, *Female* = la variable dicótoma que indica si el individuo es mujer (1 = mujer o 0 + hombre) y μ = los errores del modelo de regresión. No se han utilizado variables de control para realizar el análisis de brecha salarial incondicional de género.

Presentado por: Angie Ariza, Daniel Sandoval, Andrés Díaz, Javier Farfán

	Variable dependiente:
	Ingreso por hora
Mujer	-0.077***
	(0.012)
Constante	8.606***
	(0.009)
Observaciones	14,286
R2	0.003
R2 Ajustado	0.003
Error residual estándar	0.744 (df = 14284)
Estadístico F	38.167*** (df = 1; 14284)
Nota:	*p<0.1; **p<0.05; ***p<0.01

Los resultados de la regresión indican que si el individuo en cuestión es mujer tendría un ingreso por hora equivalente al 7.7% menos que un individuo que es hombre en la ciudad de Bogotá para el año 2018. La regresión presenta una desviación estándar de 0.012, lo cual presenta una variabilidad razonable para el análisis. Los resultados de la regresión nos indican que a un nivel de significancia del 1%, se rechaza la hipótesis nula del modelo, es decir, se rechaza la hipótesis que no existe una relación entre el género femenino y el ingreso por hora.

De acuerdo con la revisión de la literatura presentada en la sección titulada "Descripción del proceso de limpieza y selección preliminar de variables" y con el slogan común que indica que, a iguales condiciones, los trabajadores deberían obtener un ingreso similar, se procedió a realizar la estimación de la brecha de salario condicional de género donde se emplearon variables de control. Las variables de control utilizadas son: *age*, *age2, informal*, *maxEducLevel*, *sizeFirm,* y *relab.* De este modo, se utilizó el siguiente modelo:

$$\log(w) = \beta_1 + \beta_2 Female + \beta_3 Age + \beta_4 Age^2 + X\lambda + \mu,$$

donde log(w) = logaritmo natural del salario por hora, *Female* = la variable dicótoma que indica si el individuo es mujer (1 = mujer o 0 + hombre), *Age* = edad del individuo, Age^2 = edad al cuadrado, $X\lambda$ = un conjunto de controles que contiene el remanente de variables explicada anteriormente (*informal*, *maxEducLevel*, *sizeFirm,* y *relab)* y μ = los errores del modelo de regresión.

Para realizar el análisis se utilizó el método de Mínimos Cuadrados Ordinarios (MCO) para estimar la regresión con controles. También empleó el modelo Frisch-Waugh-Lovell (FWL) para realizar una comparación entre modelos. Los resultados son:

	Variable dependiente:	
	Ordinary Least Squares	FWL

Presentado por: Angie Ariza, Daniel Sandoval, Andrés Díaz, Javier Farfán

	(Sin controles)	(Con controles)	(3)
Mujer (OLS)	-0.076960***	-0.127306***	
	(0.012457)	(0.010096)	
Mujer (FWL)			-0.127306***
			(0.010090)
Constante	8.605672***	7.076159***	-0.000000
	(0.008589)	(0.077882)	(0.004888)
Observaciones	14,286	14,286	14,286
R2	0.002665	0.384299	0.011022
R2 Ajustado	0.002595	0.383479	0.010952
Error residual estándar	0.743570 (df = 14284)	0.584602 (df = 14266)	0.584233 (df = 14284)
Note:		*p<0.1; **p<0.05; ***p<0.01	

Como se puede evidenciar en la tabla de resultados expuesta anteriormente, el coeficiente del modelo de MCO con controles presenta un valor de -12.73%. Esto indica que la brecha de salario condicional indica que, si un individuo es mujer, su ingreso es 12.73% menor que un hombre cuando se controla por variables tales como la edad, la experiencia, la informalidad, el tamaño de la firma, y el tipo de ocupación. El coeficiente por controles indica que la brecha salarial condicional para las mujeres en Bogota en 2018 es mayor que la brecha salarial incondicional.

Cuando se utiliza el modelo FWL, el coeficiente es igual al modelo MCO con controles. Existe una pequeña variación en el error residual estándar en línea con una diferencia en grados de libertad. Todos los resultados de los modelos son estadísticamente significativos a un nivel de 1%.

Por otro lado, también se empleó el modelo FWL con *Bootstrap*, un modelo que permite cuantificar la incertidumbre asociada con los estimadores o coeficientes. El *Bootstrap* en cuestión está partiendo de la semilla 198 y realiza 1,000 repeticiones. Los resultados del *Bootstrap* presentan el mismo coeficiente que FWL y MCO con controles, pero el modelo presenta un error residual estándar menor que los otros modelos, presentando un error residual estándar de 0.0101 (vs. el error residual estándar de MCO de 0.585 y el error residual estándar de FWL de 0.584).

Bootstrap Statistics:

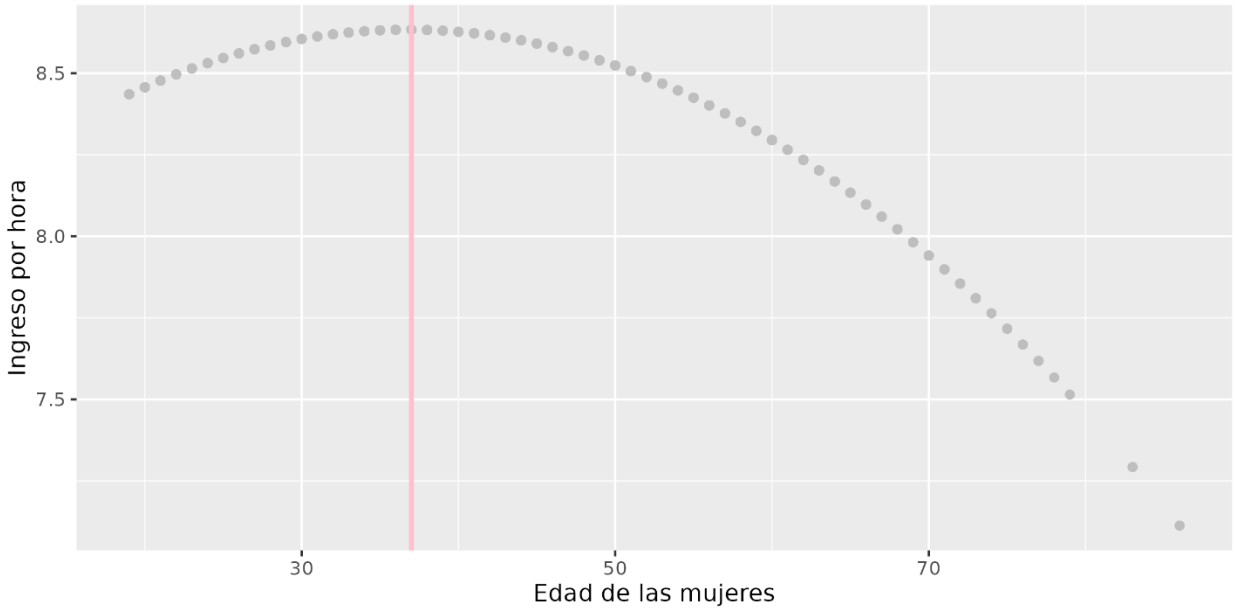
	Original	Varianza	Error Residual Estándar
t1*	-0.127306	-0.0002559765	0.01010661

Perfiles edad-salarios y "edad-pico" por género

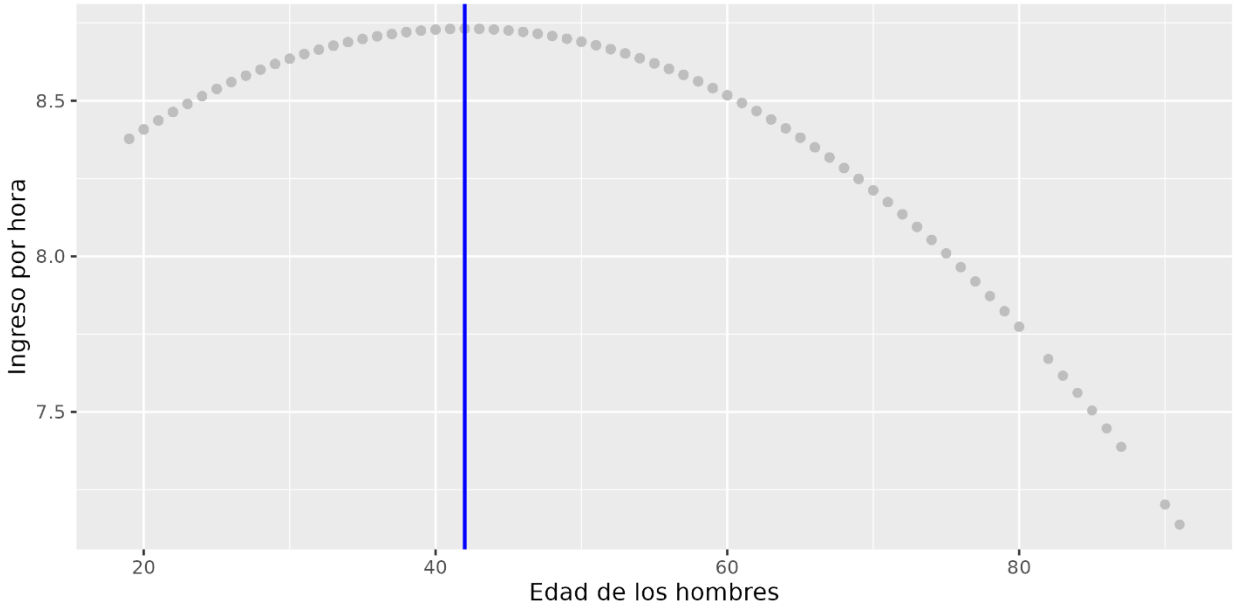
A continuación, se encuentra la gráfica que representa el perfil de la edad y el ingreso de las mujeres. Esta se creó utilizando un modelo de regresión lineal que tenía como variable dependiente el ingreso por hora, y que tenía como variables independientes la edad del individuo y la edad del individuo al cuadrado. En este análisis se encontró que la edad pico donde una mujer obtiene mayores ingresos es a la edad de 37 años con un ingreso de 8.63.

Universidad de los Andes
Facultad de Economía

Presentado por: Angie Ariza, Daniel Sandoval,
Andrés Díaz, Javier Farfán



La gráfica presentada a continuación muestra el perfil de la edad y el ingreso de los hombres. Esta se creó utilizando un modelo de regresión lineal que tenía como variable dependiente el ingreso por hora, y que tenía como variables independientes la edad del individuo y la edad del individuo al cuadrado. En este análisis se encontró que la edad pico donde un hombre obtiene mayores ingresos es a la edad de 42 años con un ingreso de 8.73.



Con base en este análisis de los perfiles de salario-edad para diferentes géneros, se puede concluir que las mujeres llegan a una edad pico de máximo salario antes que los hombres pero que su salario pico es menor al de los hombres.

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 6. Especificaciones evaluadas en el método de validación cruzada: conjunto de validación

Modelo	Especificación[1][2]
Modelo 1	$\log(w) = \beta_0$
Modelo 2	$\log(w) = \beta_0 + \beta_1 Age + \beta_2 Age^2 + u$
Modelo 3	$\log(w) = \beta_0 + \beta_1 Female + u$
Modelo 4	$\log(w) = \beta_0 + \beta_1 Female + X\lambda + u$
Modelo 5	$\log(w) = \beta_0 + \beta_1 Female + \beta_2 Age + X\lambda + u$
Modelo 6	$\log(w) = \beta_0 + \beta_1 Female + \beta_2 Age + \beta_3 Age^2 + \beta_4 Female \cdot Age + \beta_5 Female \cdot Age^2 + X\lambda + u$
Modelo 7	$\log(w) = \beta_0 + \beta_1 Female + \beta_2 Age + \beta_3 Age^2 + X\lambda + u$
Modelo 8	$\log(w) = \beta_0 + \beta_1 Female + \beta_2 Age + \beta_3 Age^2 + \beta_4 Age^3 + u$
Modelo 9	$\log(w) = \beta_0 + \beta_1 Female + \beta_2 Age + \beta_3 Age^2 + \beta_4 Age^3 + \beta_5 Age^4 + u$
Modelo 10	$\log(w) = \beta_0 + \beta_1 Female + \beta_2 Age + \beta_3 Age^2 + \beta_4 Age^3 + X\lambda + u$

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:

[1] En algunos modelos se utilizan un conjunto de controles identificado con la variable X, que contiene las siguientes variables: *informal, maxEducLevel, sizeFirm, relab*

[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.

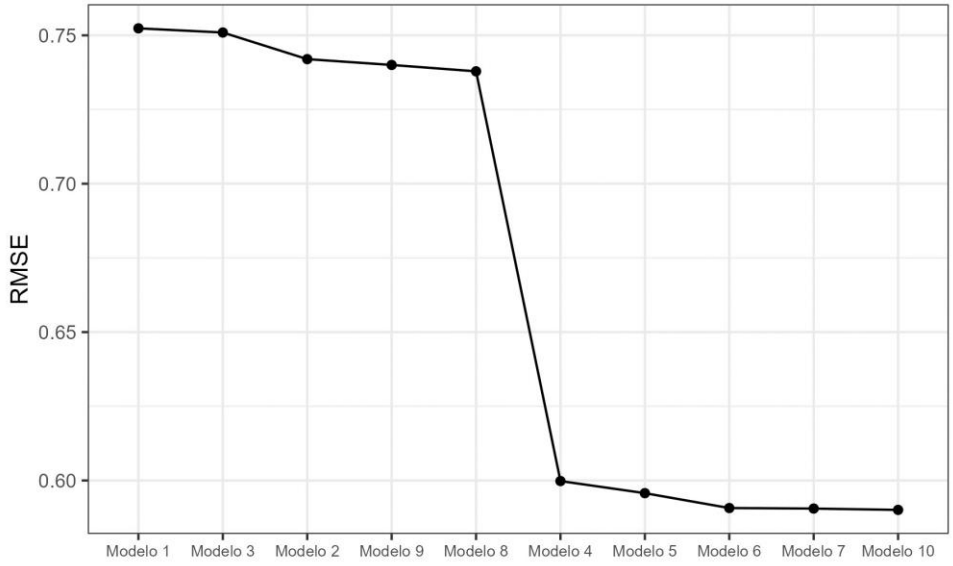


Figura 9. RMSE de prueba de las especificaciones sujetas a evaluación

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:

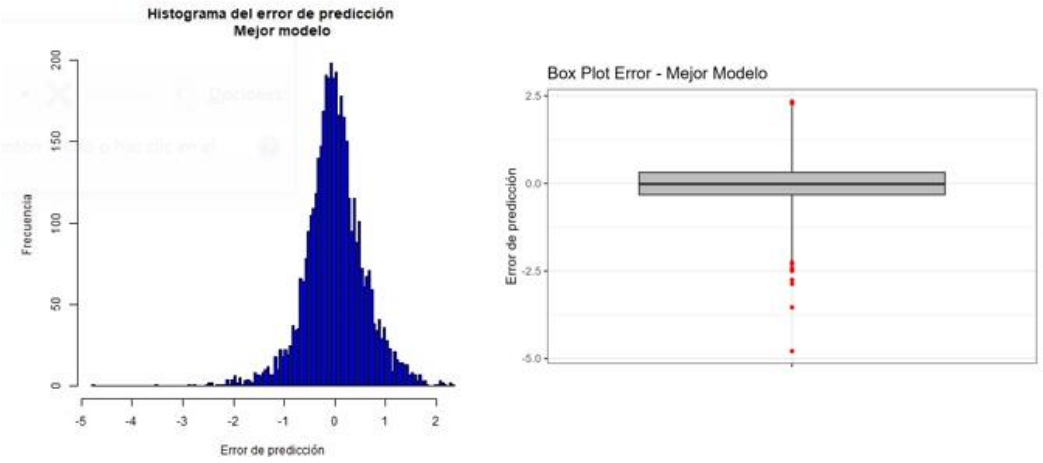


Figura 10. Distribución del error de predicción en el training set – Validation Set Approach.

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:

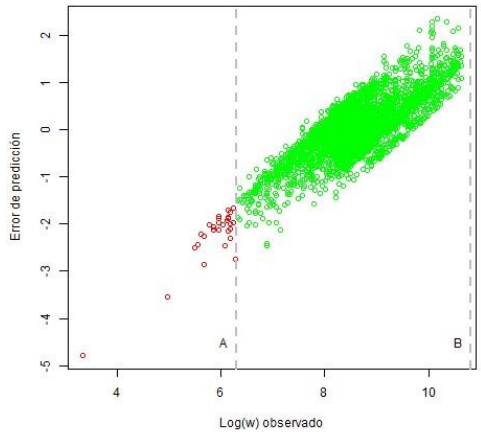


Figura 11. Diagrama de dispersión entre el error de predicción y el logaritmo del salario por hora observado en el training set – Validation Set Approach.

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.

Estimación del error del modelo usando Leave-One-out cross validation - LOOCV

Del análisis realizado por KFOOLD, se concluyó que los mejores modelos son el #7 y el # 10. Por lo cual se realiza el procedimiento de Leave-one-out cross validation con esos modelos. Una vez creadas las recetas y los workflow para cada uno, se ejecutó el bucle en el código que permitiera la realización del LOOCV.

Luego de varios días de espera y de dificultades con las maquinas que ejecutaban el procedimiento, se obtuvieron los siguientes errores asociados a los dos modelos:

Error obtenido en la salida del modelo 7 por LOOCV

```
# A tibble: 1 × 3
  .metric   .estimator  .estimate
  <chr>     <chr>           <dbl>
1 rmse      standard        0.585
```

Universidad de los Andes
Facultad de Economía

Presentado por: Angie Ariza, Daniel Sandoval, Andrés Díaz, Javier Farfán

Error obtenido en la salida del modelo 10 por LOOCV

```
# A tibble: 1 × 3
.metric  .estimator  .estimate
 <chr>   <chr>          <dbl>
1 rmse    standard       0.585
```

Los errores de los dos modelos por LOOCV son 0,585 si solo se compara a nivel de tres decimales. Lo anterior se justifica dado que los dos modelos, tanto el #7 como el #10, son muy parecidos (el modelo #10 tiene un término cubico asociado a la edad). Esta tendencia también es similar cuando se usó el KFOOLD y los errores de los dos modelos eran muy similares.

A continuación, se puede ver la comparación entre los errores hallados por KFOOLD y con LOOCV para los dos modelos:

Modelo # 7		**Modelo # 10**	
KFOOLD	**LOOCV**	**KFOOLD**	**LOOCV**
0,59	0,585	0,59	0,585

Para los dos modelos, la estimación del error con LOOCV es menor, sin embargo, la diferencia es mínima. En la ejecución del procedimiento se pudo evidenciar la ventaja de usar KFOOLD pues es mucho más rápido, evita la carga computacional y los resultados son muy similares a LOOCV. Usando el enfoque de validación, los dos modelos son capaces de predecir con un error estándar similar dado que los dos métodos usados muestran el mismo error. La predicción no se verá afectada por valores dentro de la muestra ya que al ser cada una probada en LOOCV y obtener un valor similar al usar KFOOLD de puede afirmar que las predicciones no tienen una desviación significativa para ninguno de los dos casos.

Al comparar con los dos métodos de estimación de errores, se puede evidenciar el efecto de la estadística de influencia. Esta se centra en ver el efecto de cada observación en un modelo estadístico y sus afectaciones en las estimaciones generales. En LOOCV sucede los mismo, pues el modelo iterativo usa una observación como test mientras usa el resto de observaciones para entrenar el modelo de regresión a usar. En este caso específico, las observaciones en promedio no presentan desviaciones significativas en el modelo, arrojando el mismo error para los dos métodos. Es importante resaltar esto, pues KFOOLD usa grupos de datos para entrenar y otros para predecir, arrojando estimaciones similares de errores. Las desviaciones significativas de LOOCV contra KFOOLD puede evidenciar el efecto de la estadística de influencia de algunas observaciones, con posible aplicación en la identificación de valores atípicos que alteren las predicciones de los modelos.

Para este caso no se presentan observaciones influyentes debido a que los errores por los dos métodos son muy similares.

Conclusiones

Entre los principales resultados, destacamos que las personas de estratos 1,2 y 3 trabajan la misma cantidad de horas de que las personas de estratos 4,5 y 6, a pesar de usar la misma cantidad de tiempo, los estratos bajos ganan menos en relación a los estratos altos, ejemplificando la desigualdad con la se abrió esta introducción. La distribución de

salario, ante estas diferencias entre el nivel de ingresos, asume los ingresos altos como valores atípicos, y por ello, es la importancia de utilizar y el logaritmo del salario, en el que los datos ya presentan una distribución más similar a una distribución normal.

La inclusión de controles, como son la formalidad del empleo, el nivel de educación, el tamaño de la firma en la que trabaja y el tipo de oficio, mejora en un 20%, en términos de la magnitud del RMSE, el desempeño predictivo del modelo.

El punto máximo del salario que alcanza un individuo en Bogotá es muy rápido, 22 años antes de alcanzar la edad de pensión. Esto representa un problema para la generación de ingresos, el ahorro y por supuesto el sistema de seguridad social del país.

La brecha salarial en Bogotá, evidencia un problema estructural de acceso a aspectos socioeconómicos que hace que una mujer, aunque trabaje más tiempo, gana un 12% menos que un hombre.

El error de predicción se explica parcialmente por la presencia de datos atípicos en el salario. De acuerdo con el enfoque de conjunto de validación, la contribución de estos datos atípicos en el error del modelo con el mejor desempeño era del 10.3%. Sin embargo, gran parte del error no se puede explicar por esta razón, por lo que el modelo ajustado con este método no es el adecuado para identificar posibles casos de fraude.

Bibliografía

Aristizábal Lopera, T., & Ángel López, E. (2017). *Efecto de la escolaridad sobre el salario en Colombia*. https://repository.eafit.edu.co/handle/10784/12111?locale-attribute=es

CEPAL. (2007). *Las brechas entre el campo y la ciudad en Colombia, 1990-2003, y propuestas para reducirlas.* CEPAL.

DANE. (2020). *BRECHA SALARIAL DE GÉNERO EN COLOMBIA*.

Guataquí, J. C., García, A. F., & Rodríguez, M. (2009). *ESTIMACIONES DE LOS DETERMINANTES DE LOS INGRESOS LABORALES EN COLOMBIA CON CONSIDERACIONES DIFERENCIALES PARA ASALARIADOS Y CUENTA PROPIA*.

Iregui, A. M., Melo, L., & Ramírez, M. T. (2010). Formación e incrementos de salarios en Colombia_ un estudio microeconómico apartir de una encuesta a nivel de firma. *REPORTES DEL EMISOR*, 1–8.

Mahnic, P. (2022). Eduación y crecimiento económico: considerando no linealidades en la ecuación de Mincer. *Económica*, *68*, 027. https://doi.org/10.24215/18521649e027

Otero-Cortés, A., & Acosta-Ariza, E. (2022). Desigualdades en el mercado laboral urbano-rural en Colombia, 2010-2019. *Revista CS*, *Especial*, 173–219. https://doi.org/10.18046/recs.iespecial.4939

PAGE, M. (2022). *ESTUDIO DE REMUNERACIÓN 2022*.

Universidad de los Andes
Facultad de Economía

Presentado por: Angie Ariza, Daniel Sandoval,
Andrés Díaz, Javier Farfán