Big Data – 202213
Maestría en Economía Aplicada – Universidad de los Andes

Taller 1
junio 27, 2022

Presentación

Integrantes: Ingrid Lorena Molano cód. 200511102
Jorge Eduardo García cód. 201310645
Camilo Villa Moreno cód. 201818624

Repositorio: https://github.com/camilovillam/MECA_BD_PS1

Punto 1 - *Data acquisition*

- *Scrape the data that is available at the following website https://ignaciomsarmiento.github.io/GEIH2018 sample/*

Se realizó el *"scrape"* de la información y el código de este proceso se encuentra en el repositorio citado en la presentación, en la carpeta scripts, en el archivo en formato *".R"*. La base de datos obtenida se encuentra almacenada en la carpeta stores.

- *Are there any restrictions to accessing/scraping these data?*

Si se encontraron restricciones al momento de acceder a la información. La página en la que se enunciaba la información no contenía la tabla. Las tablas provenían de otra URL que llamaba la página del enunciado.

La otra dificultad es que no toda la información estaba en la misma página, por lo que tuvo que identificarse las páginas de las que provenían las tablas para poder crear la base de datos completa. Una vez se identificaron las páginas, se logró establecer una estructura de forma en la que estaba escrita la URL de las páginas de interés y mediante un bucle se obtuvo la base de datos.

- *Using pseudocode describe your process of acquiring the data*

El pseudo código para explicar la adquisición de datos es el siguiente:

```
Página 1 = Leer la URL de la
página("https://ignaciomsarmiento.github.io/GEIH2018_sample/page1.html") y
descargar el html

URL Página de tabla 1 = Consultar la Página 1 para buscar el atributo w3-
include-html
```

```
Página de la tabla 1 = Descargar el html de la URL Página de tabla 1

Tabla página 1 = Consultar la Página de la tabla 1 en busca de una tabla y cargar la tabla

# Ahora que sabemos que las URLs son de la forma https://ignaciomsarmiento.github.io/GEIH2018_sample/pages/geih_page_(N).html donde (N) es el número de la página, hacemos un bucle para cargar todas las páginas

Crear una variable para guardar el data frame con todos los datos

Para cada página entre 1 y 10

      Cargar html de página

      Cargar tabla consultando el html de la página

      Agregar tabla aldata frame

Guardar el data frame en la carpeta stores
```

Punto 2 - *Data Cleaning*

In this problem set, we will focus only on employed individuals older than eighteen (18) working in Bogota´. In this section, you are going to focus on cleaning and describing the data.

- *The data set include multiple variables that can help explain individual income. Guided by your intuition and economic knowledge, choose the most relevant and perform a descriptive analysis of these variables. For example, you can include variables that measure education and experience, given the implications of the human capital accumulation model (Becker, 1962, 1964; and Mincer (1962, 1975).*
- *Note that there are many observations with missing data. I leave it to you to find a way to handle these missing data. In your discussion, describe the steps that you performed cleaning de data, and justify your decisions.*
- *At a minimum, you should include a descriptive statistics table, but I expect tables and figures. Take this section as an opportunity to present a compelling narrative to justify and defend your data choices. Use your professional knowl- edge to add value to this section. Do not present it as a "dry" list of ingredients.*

Pasos para la limpieza base de datos:

1. Primer filtro: Se eliminan las filas de las personas con 18 años o menores dado que el enunciado indica que el estudio debe enfocarse en individuos mayores de 18 años trabajando en Bogotá.
2. Segundo filtro: Se eliminan las personas que no están ocupadas ya que el enunciado indica que el estudio debe enfocarse en individuos que están trabajando.
   Lo anterior arroja una base de datos de personas mayores de 18 años que están ocupadas. Se toman las personas ocupadas ya que se asume que los ocupados son los que se encuentran trabajando, incluyendo los asalariados e independientes. Se toma a los independientes como empleados poque esta clasificación pueden encontrarse

empleados que pagan su seguridad social o personas con contrato de prestación de servicios.

Ilustración 1. Estructura de la población y su desagregado para identificar a los ocupados

Fuente: Tomado de Misión para el Empalme de las Series de Empleo, Pobreza y Desigualdad Mesep (DNP-DANE, 2012)

3. Se identifica la variable ingreso que se va a utilizar en la estimación a lo largo del taller como variable dependiente.

Tabla 1. Variables seleccionadas como ingreso

Variable	Descripción	Razón para la selección
ingtot	Ingresos totales	Se escoge la variable ingreso total, variable que resulta luego de la sumatoria del ingreso observado e imputado, permitiendo una mejor depuración de los datos de los encuestados. Se escoge esta variable en vez de la variable ingreso que indica el salario, pues se filtra por ocupados, que incluyen empleados formales y no formales, empleados e independientes.

Fuente: Elaboración propia

4. Se identifican las variables de interés que se consideran determinantes al momento de estimar el ingreso (variables independientes). Estas variables son:

Tabla 2. Variables identificadas que son determinantes del ingreso

Variable	Descripción	Razón para considerarla como determinante del ingreso
age	Edad	De acuerdo con Muñoz (2004), la edad permite medir la experiencia del trabajador, por lo que, a mayor edad, aumenta el ingreso. Por su parte, Dentré y Herard (2004) interpretan la edad como la acumulación de capital humano, lo que afecta los ingresos de la persona.
sex	Sexo	De acuerdo con Muñoz (2004), el sexo tiene incidencia fuerte sobre los ingresos de las personas por razones culturales. Para la CEPAL (2007), hay una brecha de ingreso entre los hombres y las mujeres por dos razones fundamentales: 1. Las mujeres tienden a estar ocupadas en servicios que, generalmente, son peor remuneradas y pocos protegidas por la legislación. 2. Las mujeres ocupan actividades domésticas luego del casamiento y la maternidad. Y estas diferencias tienden a profundizarse con la vejez. Por otro lado, siguiendo a Hersch (2006), hay una discriminación dentro del mercado laboral, dadas unas características observables de cada individuo (en este caso género), lo que provoca un trato laboral desigual incluso entre personas igualmente productivas, donde se puede evidenciar una discriminación por género, medida en salarios (Cerquera, Arias, Prada, 2019).
p6210	Nivel educativo	De acuerdo con Muñoz (2004), la educación incide en el ingreso de una persona por tres razones: 1. aumenta la productividad de la persona, 2. Da prestigio y 3. Da mayor capacidad, elementos que impactarán de manera positiva el ingreso de la persona. De igual manera, para Figueroa (2010), la educación hace parte del capital humano, entendido este como el conjunto de habilidades y destrezas productivas que se incorporan a los trabajadores, mejorando el nivel de ingreso. Sin embargo, para el mismo autor, esta variable es necesario analizarla con rigurosidad, pues en el caso de los países en vía de desarrollo (exactamente en Perú) hay una paradoja, pues, la educación no es un sistema nivelador de ingresos por dos

Variable	Descripción	Razón para considerarla como determinante del ingreso
		razones: 1. La desigualdad inicial del individuo y 2. La sobrepoblación.
p6210s1	Último grado	Permite calcular los años de educación. Es importante esta variable porque como lo indican Rincón y Jiménez (2011), por cada año adicional de preparación de la persona, se ve un afecto positivo en el ingreso de la misma.
oficio	Tipo de oficio	Esta variable se toma de manera intuitiva, pues si una persona tiene acceso a un mejor empleo, esta obtendrá mejores ingresos, respecto de una que tenga un trabajo de menor grado u otra ocupación. Por ejemplo, para el caso de las personas que estudian, esto se puede constatar con la tasa de retorno de la educación, pues por cada año adicional de educación, la persona obtendrá una mejora en su salario (Lora, 2021), ya sea por las habilidades aprendidas que las llevarán a un mejor cargo o a unas mejoras en sus condiciones laborales.
sizeFirm	Tamaño de la empresa	De manera intuitiva, se escoge esta variable teniendo en cuenta que cuanto mayor es la empresa, el ingreso de la persona tiende a ser mayor, incluso en el mismo cargo. Esto se puede evidenciar, por ejemplo, en las tablas salariales de las disitintas entidades del orden público, donde personas con el mismo perfil y actividades similares, obtienen un mayor ingres respecto de otras. En todo caso, esta variable se considera como determinante para los ingresos en un estudio hecho para la Colombia (Quintero, 2013).
formal	Tipo de trabajo	De acuerdo con el Ministerio del Trabajo, “el trabajo formal representa un ingreso digno y una protección social para el trabajador y su familia, lo que conlleva a una mejor calidad de vida, progreso social y económico, además de una reducción de la pobreza”. En este sentido, si una persona está dentro del mercado laboral formal, en teoría, obtendrá unos mejores ingresos, no exactamente por la actividad desempeñada, sino que este también le permite el acceso al crédito y generar otro tipo de ingresos. Además, teniendo en cuenta que en Bogotá la informalidad ronda alrededor de un 42%

Variable	Descripción	Razón para considerarla como determinante del ingreso
		(DNP, 2018), resulta importante ver el impacto de la variable en ambos escenarios. Nota: se escoge la cifra del 2018, teniendo en cuenta que la base de datos corresponde a ese año.
totalHoursWorked	Total horas de trabajo a la semana	Se infiere que el número de horas trabajadas a la semana afecta el ingreso laboral, pues por cada hora de trabajo adicional realizada, la persona aumentará su ingreso.
p6426	Antigüedad en el trabajo actual (meses)	Esta variable mide exactamente la experiencia laboral de la persona, lo que, al igual que la edad, impactaría de manera positiva el ingreso de las personas (Muñoz, 2004); es decir, una persona adquiere un mayor ingreso, generalmente, en la mediada que tenga mayor experiencia laboral.
p6050	Parentesco con el jefe de familia	En esta se calcula el número de hijos que tiene la persona. Resulta importante esta variable, especialmente en la medición de los ingresos de las mujeres, pues tienen más barreras al momento de conseguir un empleo, dada las limitaciones de tiempo, máxime cuando estas son jefes del hogar (CEPAL, 2007)

Fuente: Elaboración propia

5. Tercer filtro: Se procede a crear una base de datos en la que se deja únicamente las variables identificadas en los pasos 3 y 4 (Función subconjunto).
6. Se evalúan el porcentaje de datos faltantes en la nueva base de datos y se encuentra que el aproximado es un 0% de datos faltantes.
7. Se crean variables proxys que se consideran determinantes en la estimación del ingreso.

Tabla 3. Variables proxy determinantes del ingreso

Variable	Descripción	Razón para considerarla como determinante del ingreso
años_educ	Años estudiados	Tal vez no considerar el nivel educativo como una categórica, sino calcular el número de años aproximado. Esto es importante porque el nivel terciario puede ser muy diferente, y puede haber importante variación en cantidad de años y salarios. Se pueden aproximar los años estudiados con la información en p6210 y p6210s1 (Nivel educativo y último año aprobado)
exper_pot	Experiencia potencial	Como en la base de dato no se encuentra el número de años de experiencia laboral se utiliza como variable proxy: Experiencia potencial. En la literatura se ha utilizado como proxy de la experiencia la experiencia potencial. Esta nace de restarle a la edad de la

Variable	Descripción	Razón para considerarla como determinante del ingreso
		persona los años que ha estudiado y, además, cinco (5) años – pues en sus años de primera infancia ni estudió ni trabajó.
num_hijos	Número de hijos	Hay evidencia1 que indica que el número de hijos, para las mujeres, es determinante. La variable no está incorporada directamente en la base de datos, pero se puede calcular de manera indirecta con base en la variable P6050 (Parentesco con el jefe actual). Se calcula el número de hijos por hogar y solo a quienes son padres/madres o su pareja se les imputa el número de hijos; a los demás, se les imputa 0.

Fuente: Elaboración propia

Manejo de datos faltantes:

Al filtrar la base de datos con las variables de interés para el objeto de estudio no se identificó el problema de datos faltantes.

Solo en la creación de la variable años de educación (años_educ) se imputa la mediana para datos faltantes.

Estadística descriptiva:

A continuación se presentan las estadísticas descriptivas de las variables seleccionadas.

Tabla 4. Estadísticas descriptivas variables continuas

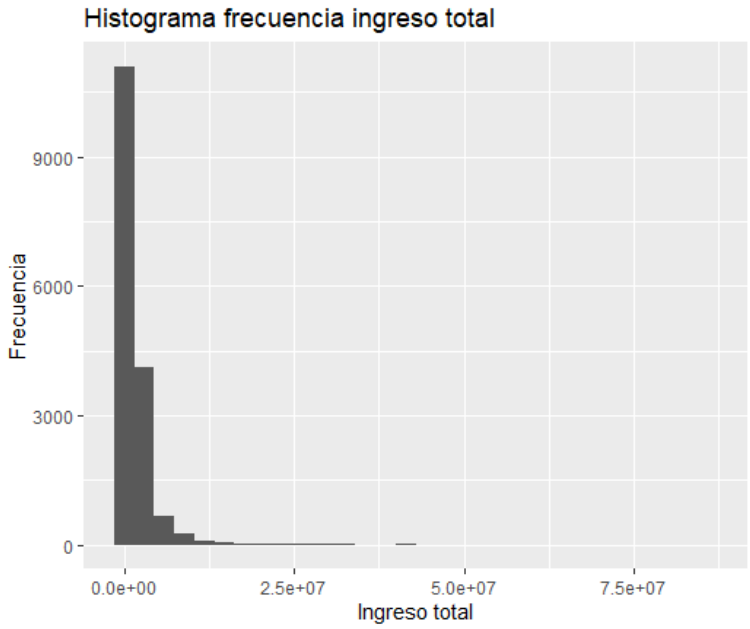
Summary of continuous variables

	n	miss	p.miss	mean	sd	median	p25	p75	min	max	skew
ingtot	16397	0	0	2,00E+06	3,00E+06	1,00E+06	8,00E+05	2,00E+06	0,00E+00	9,00E+07	8.18
age	16397	0	0	4,00E+01	1,00E+01	4,00E+01	3,00E+01	5,00E+01	2,00E+01	9,00E+01	0.47
p6210	16397	0	0	5,00E+00	1,00E+00	5,00E+00	4,00E+00	6,00E+00	1,00E+00	9,00E+00	-0.96
p6210s1	16397	0	0	7,00E+00	4,00E+00	6,00E+00	4,00E+00	1,00E+01	0,00E+00	1,00E+02	1.06
totalHoursWorked	16397	0	0	5,00E+01	2,00E+01	5,00E+01	4,00E+01	5,00E+01	1,00E+00	1,00E+02	0.17
p6426	16397	0	0	5,00E+00	7,00E+00	2,00E+00	6,00E-01	7,00E+00	0,00E+00	6,00E+01	2.33
p6050	16397	0	0	2,00E+00	2,00E+00	2,00E+00	1,00E+00	3,00E+00	1,00E+00	9,00E+00	2.36
años_educ	16397	0	0	1,00E+01	4,00E+00	1,00E+01	9,00E+00	2,00E+01	0,00E+00	3,00E+01	-0.25
exper_pot	16397	0	0	2,00E+01	2,00E+01	2,00E+01	1,00E+01	3,00E+01	-6,00E+00	8,00E+01	0.59
dummy_hijos	16397	0	0	2,00E-01	4,00E-01	0,00E+00	0,00E+00	0,00E+00	0,00E+00	1,00E+00	1.47
num_hijos	16397	0	0	2,00E-01	5,00E-01	0,00E+00	0,00E+00	0,00E+00	0,00E+00	4,00E+00	2.95
age_cuad	16397	0	0	2,00E+03	1,00E+03	1,00E+03	8,00E+02	2,00E+03	4,00E+02	9,00E+03	1.08
años_educ_cuad	16397	0	0	1,00E+02	1,00E+02	1,00E+02	8,00E+01	2,00E+02	0,00E+00	7,00E+02	0.77
exp_pot_cuad	16397	0	0	8,00E+02	9,00E+02	4,00E+02	1,00E+02	1,00E+03	0,00E+00	7,00E+03	1.72

Fuente: Elaboración propia utilizando la base de datos del PS1

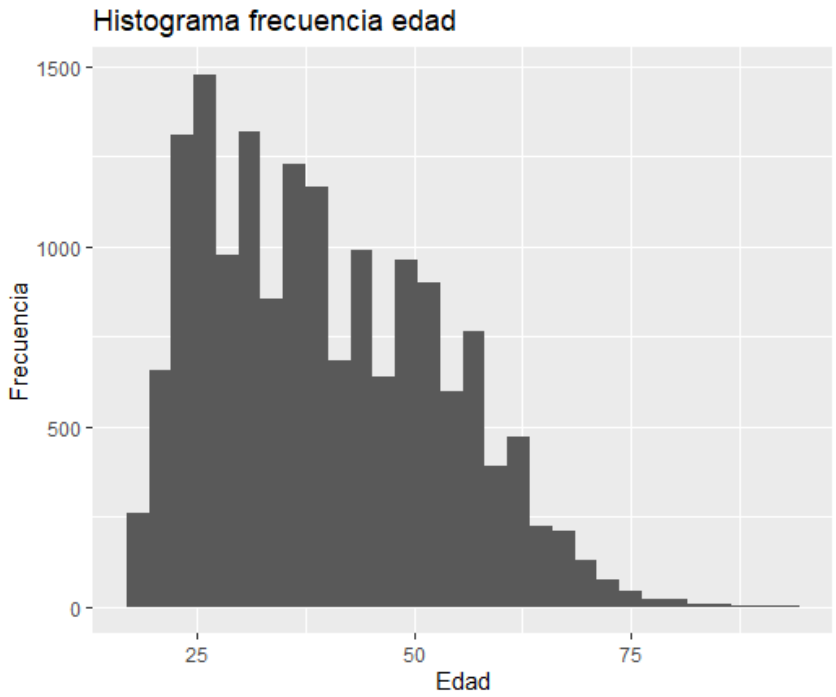
[1] La brecha salarial son los hijos. Artículo del El País, publicado en: https://elpais.com/politica/2018/03/02/actualidad/1520006491_549539.html

Ilustración 2. Histograma frecuencia ingreso total



Fuente: Elaboración propia utilizando la base de datos del PS1

Ilustración 3. Histograma frecuencia ingreso total



Fuente: Elaboración propia utilizando la base de datos del PS1

Tabla 5. Estadísticas descriptivas variables categóricas

```
     ### Summary of categorical variables ###
var       n     miss p.miss level                  freq percent  cum.percent
sex       16397    0 0.0    Mujer                  7715 47.1     47.1
                            Hombre                 8682 52.9     100.0

sizeFirm 16397     0 0.0    Independiente          4025 24.5     24.5
                            2-5 trabajadores       3302 20.1     44.7
                            6-10 trabajadores      1116 6.8      51.5
                            11-50 trabajadores     2039 12.4     63.9
                            Más de 50 trabajadores 5915 36.1     100.0

formal    16397    0 0.0    Informal               6721 41.0     41.0
                            Formal                 9676 59.0     100.0
```

Fuente: Elaboración propia utilizando la base de datos del PS1

Punto 3 - *Age-earnings profile*

A great deal of evidence in Labor economics suggests that the typical worker's age-earnings profile has a predictable path: Wages tend to be low when the worker is young; they rise as the worker ages, peaking at about age 50; and the wage rate tends to remain stable or decline slightly after age 50.

- *In the data set, multiple variables describe income. Choose one that you believe is the most representative of the workers' total earnings, justifying your selection.*

Se escoge la variable ingreso total, variable que resulta luego de la sumatoria del ingreso observado e imputado, permitiendo una mejor depuración de los datos de los encuestados. Esta variable mide la totalidad de los ingresos de las personas provenientes de distintas fuentes; es decir, no solamente mide los ingresos de las personas que se encuentran dentro del mercado laboral sino también incluye a las personas que se encuentran fuera de la formalidad y reciben otros tipos de ingresos.

Para realizar este ejercicio, se podría utilizar la variable "ingreso laboral" que mide los ingresos monetarios por concepto de labores de cada persona; es decir, el ingreso percibido por realizar una actividad de manera consistente y bajo un conjunto de formalidades y prerrogativas. Esto permitiría focalizarse en el interés de la teoría económica expuesta en el ejercicio que es el salario y la edad, en donde a mayor edad, se adquiere un mayor salario, teniendo en cuenta elementos como el nivel de educación y la experiencia con el pasar de los años. Sin embargo, el mercado laboral bogotano para el año 2018 se componía en un 41.8% por trabajadores informales (DNP, 2018), lo que dejaría por fuera del análisis a un segmento importante de la población en el análisis. De igual manera, la variable "ingreso total" incluye tanto el mercado formal como informal, lo que permitiría hacer ambas mediciones y tener una mejor explicación del salario.

Tabla 6. Estadísticas descriptivas variable ingtot

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	800000	1059878	1778855	1736544	85833333

Fuente: Elaboración propia utilizando la base de datos del PS1

De acuerdo con la variable "Ingreso Total", un colombiano tiene un mínimo de ingresos de 0$ y un máximo de $85'833.333, por lo que se puede decir que hay una brecha muy alta entre la persona que menos y más gana en Bogotá (datos atípicos). De igual manera, el ingreso que se encuentra en la mitad (mediana) es de $1059878. Por último, se puede decir, de acuerdo con los datos recolectados, que una persona en Colombia tiene un ingreso de $1'778855.

- *Based on this estimate using OLS the age-earnings profile equation:*

$$Income = \beta 1 + \beta 2Age + \beta 3Age2 + u \qquad (2)$$

En esta parte del ejercicio se crea la variable age2, la cual es la edad al cuadrado. Este proceso es importante para explicar, siguiendo la teoría económica laboral (Muñoz, 2004), el comportamiento del ingreso respecto de la edad; es decir, una persona con el pasar de sus años, adquiere distintas habilidades mediante la educación y la experiencia laboral que le permiten obtener un mejor ingreso, lo que se le denomina capital humano (Torres, 2019). Sin embargo, dada la forma cóncava de la función y siguiendo la teoría económica, se puede observar que el ingreso de la persona aumenta, pero este es menor respecto del año anterior, y, en un punto de su vida, medida en años, obtiene su máximo de ingreso, y luego estos empiezan a decrecer aún más. Esto se puede explicar por el deterioro de la salud de la persona, previsión social y/o modificaciones de la fuerza del mercado laboral (Miralles, 2010).

- *How good is this model in sample fit?*

Regresión del modelo

Tabla 7 Regresión Ingreso versus edad y edad al cuadrado

	Dependent variable:
	ingtot
age	88,845.290***
	(9,142.668)
age2	-775.479***
	(105.316)
Constant	-385,016.200**
	(184,585.700)
Observations	16,397
R^2	0.016
Adjusted R^2	0.016
Residual Std. Error	2,664,061.000 (df = 16394)
F Statistic	131.992*** (df = 2; 16394)
Note:	*p<0.1; **p<0.05; ***p<0.01

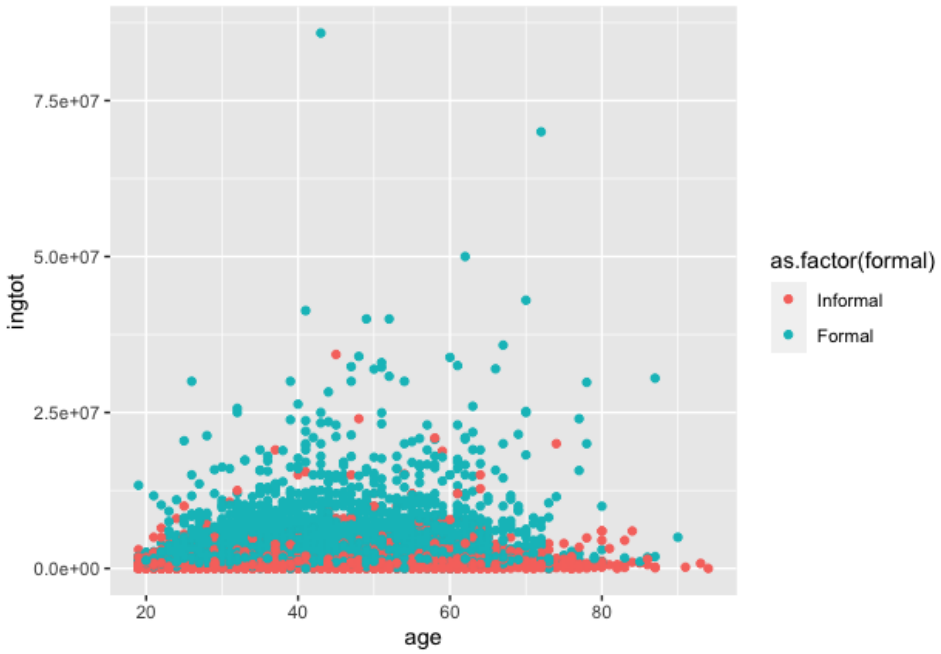
Fuente: Elaboración propia utilizando la base de datos del PS1

Esta modelo se hizo con 16.393 observaciones. Este modelo lo que busca es predecir el ingreso teniendo en cuenta la edad del individuo. Así, el ingreso de una persona se obtiene de multiplicar age (edad de la persona) por B2 (88.845,290), menos el valor de multiplicar su edad al cuadrado (age2) por B3 (-775.479) y luego se resta el valor del intercepto o constante (-385.016,200). De igual manera, hay dependencia global en el modelo, dado que el *P Valor de la Prueba F* es cercano a 0. Por último, el R^2 muestra que los cambios en el ingreso son explicados por la edad en un 16%, lo que muestra un modelo poco explicativo, dado que, de acuerdo con la literatura expuesta en la selección de variables, el ingreso depende también de otras variables.

Por otro lado, de acuerdo con la teoría económica y el enunciado expuesto, se dice que las personas que adquieren un mayor nivel de capital humano (Torres, 2019), tienden a tener mayores ingresos. Se podría decir que las personas que poseen estas características se encuentran dentro del mercado laboral formal. Así, la teoría se estaría cumpliendo, ya que en Bogotá, de acuerdo con los datos suministrados, las personas que tienen un mayor ingreso son las que se encuentran dentro del mercado formal, como lo muestra la Ilustración 4.

En conclusión, es un modelo poco explicativo y predictivo, dado que el ingreso es explicado también por otras variables. Se cumple la teoría de que a mayor edad, suponiendo que esta implica mayor capital humano, el ingreso de la persona mejora, lo que se demuestra en la Ilustración 5. Por último, teniendo en cuenta los trabajos realizados, es preferible que la variable dependiente ingreso sea transformada en forma logarítmica, dado su comportamiento gráfico.

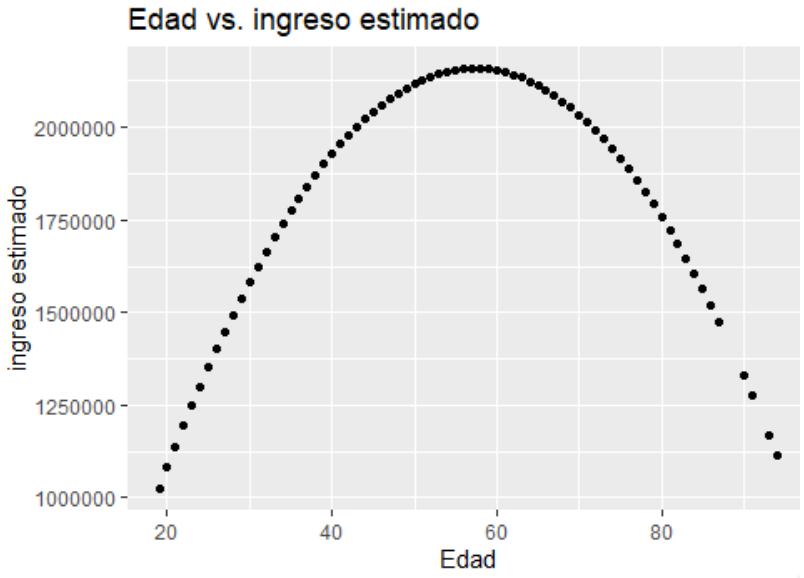
Ilustración 4 Distribución de ingreso para el mercado formal e informal



Fuente: Elaboración propia utilizando la base de datos del PS1

- *Plot the predicted age-earnings profile implied by the above equation.*

Ilustración 5. Ingreso estimado versus edad



Fuente: Elaboración propia utilizando la base de datos del PS1

Con los datos suministrados, la ilustración anterior muestra un comportamiento cóncavo de la edad respecto del ingreso, confirmándose lo expuesto por la teoría económica; es decir, el comportamiento decreciente del ingreso respecto de la edad, teniendo un ingreso mínimo a la edad de los 18 años (desde esa edad se hace la medición), alcanzando su nivel máximo a los 57.28 años, para luego aplanarse durante unos años e iniciar un descenso por las condiciones como la salud y modificación de la fuerza laboral, entre otros, que los lleva "a competir con los jóvenes que inician su vida laboral, y como respuesta a esta situación recurren a la generación de empleo informal" (Ocampo, 2010), por lo que sus ingresos por actividades realizadas directamente, tienden a disminuir.

- *What is the "peak age" suggested by the above equation? Use bootstrap to calculate the standard errors and construct the confidence intervals.*

Luego de hacer lo correspondiente, el "peak age" es de 57.28 años de edad; es decir, la persona alcanza su máximo de ingreso a los 57.28 años de edad.

Ilustración 6. Bootstrap para calcular el error estándar del peak age

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = Datos_P3, statistic = eta_mod.fn, R = 1000)


Bootstrap Statistics :
    original   bias    std. error
t1*  57.2841 1.153608    5.790346
```

Fuente: Elaboración propia utilizando la base de datos del PS1

Luego de realizar el bootstrap, se confirma el peak-age de 57.28 años de edad y se observa que la desviación estándar es de 5.79 años de edad; es decir, los datos difieren con respecto al valor promedio de edad en 5.79 años de edad.

Intervalos de confianza

Asumiendo todo lo demás constante, se espera que el ingreso de una persona en Bogotá tenga su punto máximo, en promedio, cuando este tenga entre 45.93 años y 68.633 años d edad, con un nivel de confianza del 95%

Punto 4 - *The earnings GAP*

Most empirical economic studies are interested in a single low dimensional parameter, but determining that parameter may require estimating addi- tional "nuisance" parameters to estimate this coefficient consistently and avoid omit- ted variables bias. Policymakers have long been concerned with the gender earnings gap.

- *Estimate the unconditional earnings gap*

$$log(Income) = \beta 1 + \beta 2 Female + u \qquad (3)$$

Tabla 8. Regresión logaritmo del ingreso versus la variable dicótoma mujer

	(1)
mujer	-0.196***
	(0.014)
Constant	14.071***
	(0.009)
Observations	16,138
R2	0.012
Adjusted R2	0.012
Residual Std. Error	0.871 (df = 16136)
F Statistic	202.761*** (df = 1; 16136)

Fuente: Elaboración propia utilizando la base de datos del PS1

Según lo anterior se tiene el siguiente modelo:

$$\log(ingtot) = 14.071 - 0.196\,mujer + u$$

Este modelo es acorde a lo esperado, ya que el beta de la variable mujer presenta el signo negativo, lo cual es congruente con los estudios de brecha género, en los que se tiene que ser mujer afecta negativamente el nivel de ingresos.

- *How should we interpret the β2 coefficient? How good is this model in sample fit?*

Dado que se trata de un modelo donde log-lin (variable dependiente en logaritmo y

variable independiente en formato lineal - discreta), el coeficiente $\beta 2$ se interpreta de la forma que si la variable mujer toma el valor de 1 (es mujer), se genera un cambio en el ingreso equivalente al 100* $\beta 2$ porciento. De este modo, se tiene que una mujer tiene un ingreso 19.6% menor que el de los hombres.

- *Estimate and plot the predicted age-earnings profile by gender. Do men and women in Bogota´ have the same intercept and slopes?*

Se prueban varias regresiones para identificar el mejor modelo que estime la relación entre el logaritmo del ingreso y la edad por género. Estas regresiones se presentan en la siguiente tabla.

Tabla 9. Regresiones estimadas para escoger el modelo que permita estimar el "peak age" por genero

	Dependent variable:			
	ln_ing			
	(1)	(2)	(3)	(4)
mujer	-0.196***	-0.195***	-0.207***	0.268***
	(0.014)	(0.014)	(0.014)	(0.043)
age		0.002***	0.057***	0.065***
		(0.001)	(0.003)	(0.003)
age_cuad			-0.001***	-0.001***
			(0.00003)	(0.00003)
mujer_age				-0.012***
				(0.001)
Constant	14.071***	13.983***	12.932***	12.680***
	(0.009)	(0.023)	(0.061)	(0.064)
Observations	16,138	16,138	16,138	16,138
R2	0.012	0.014	0.034	0.042
Adjusted R2	0.012	0.013	0.034	0.042
Residual Std. Error	0.871 (df = 16136)	0.871 (df = 16135)	0.862 (df = 16134)	0.858 (df = 16133)
F Statistic	202.761*** (df = 1; 16136)	110.646*** (df = 2; 16135)	190.505*** (df = 3; 16134)	178.609*** (df = 4; 16133)
Note:				*p<0.1; **p<0.05; ***p<0.01

Fuente: Elaboración propia utilizando la base de datos del PS1

$$\log(ingtot) = 12.138 + 0.268\ mujer + 0.065\ edad - 0.001\ edad^2 - 0.012\ mujer\_edad + u$$

Se escoge este modelo porque es el que mayor coeficiente de determinación presenta (R2=0.042).

Se resalta que en este modelo cambia el signo de la variable *mujer*, el cual deja de ser negativo y pasa a ser positivo. Este efecto debe ser contrarrestado con el signo de la variable interacción *mujer_age*, el cual es negativo.

El modelo seleccionado tiene una variación en el intercepto y en la pendiente si es mujer. De este modo, el modelo cambia de intercepto y pendiente si es mujer o es hombre.

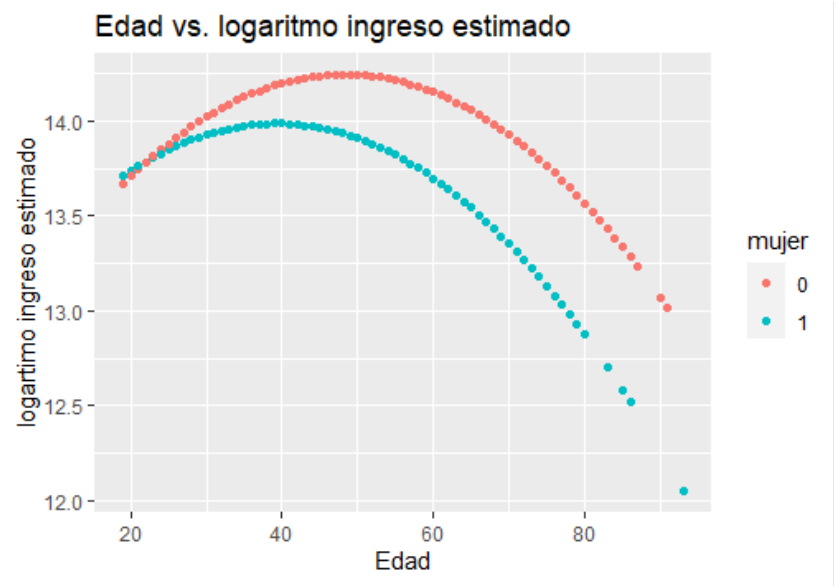
Tabla 10. Pendiente e intercepto si es hombre o mujer

Género	**Intercepto**	**Pendiente**
Mujer	12.680+0.268=12,948	0.065
Hombre	12.680	0.053

Fuente: Elaboración propia utilizando la base de datos del PS1

La diferencia entre el intercepto y la pendiente se evidencia en la siguiente gráfica:

Ilustración 7. Logaritmo del ingreso estimado según la edad diferenciado por el genero



Fuente: Elaboración propia utilizando la base de datos del PS1

En la gráfica se identifica que si es mujer antes de los 20 años (aproximadamente), las mujeres ganan más que en los hombres, hallazgo que no es esperado según los estudios donde la mujer gana menos, pero que es acorde al signo positivo de la variable *mujer* en la regresión.

La tendencia anterior se invierte luego de los 20´s años de edad y el hombre gana más que la mujer, resultado acorde al signo de la variable interacción *mujer_age*.

De igual forma, la gráfica evidencia que la edad en la mujer alcanza su ingreso máximo es menor que la edad en la que el hombre alcanza su ingreso máximo. Según la gráfica esta diferencia es de aproximadamente 10 años.

- *What is the implied "peak age" by gender?. Use bootstrap to calculate the standard errors and construct the confidence intervals. Do these confidence intervals overlap?*

Para calcular el "*peak age"* se deriva y se iguala a cero. Luego se calcula mediante la técnica Bootstrap el error estándar.

Ilustración 8. Bootstrap para calcular el error estándar del peak age si es mujer

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = datosGEIH_P4, statistic = peakage_m.fn, R = 1000)


Bootstrap Statistics :
    original     bias    std. error
t1* 39.31479 0.02985336   0.6262007
```

Fuente: Elaboración propia utilizando la base de datos del PS1

Ilustración 9. Bootstrap para calcular el error estándar del peak age si es hombre

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = datosGEIH_P4, statistic = peakage_h.fn, R = 1000)


Bootstrap Statistics :
    original    bias    std. error
t1* 48.24924 0.0695564   0.7453415
```

Fuente: Elaboración propia utilizando la base de datos del PS1

Con el error estándar se calcula el intervalo de confianza al 95%, por lo que se obtienen los siguientes resultados:

Tabla 11. "Peak age" por genero e intervalos de confianza del error estándar

Género	***Peak age***	**Error Estándar**	**Intervalo de confianza**	
			Límite inferior	**Límite superior**
Mujer	39.31479 ≅ 39,3 años	0.6262007	38.08743	40.54214
Hombre	48.24924 ≅ 48,2 años	0.7453415	46.76914	49.72935

Fuente: Elaboración propia utilizando la base de datos del PS1

En este caso los intervalos no se superponen.

- *Equal Pay for Equal Work? A common slogan is "equal pay for equal work". One way to interpret this is that for employees with similar worker and job characteristics, no gender earnings gap should exist. Estimate a conditional earnings gap that incorporates control variables such as similar worker and job characteristics (X).*

- *Estimate the conditional earnings gap log(Income) = $\beta 1 + \beta 2 Female + \theta X + u$*

Se prueban diferentes modelos con diferentes variables de control. A continuación, se presentan los modelos estimados:

Tabla 12. Modelos estimados para identificar la brecha por genero condicional con variables de control

	Dependent variable:			
	ln_ing			
	(1)	(2)	(3)	(4)
mujer	-0.018	-0.009	-0.005	-0.002
	(0.032)	(0.031)	(0.031)	(0.031)
age	-0.006*	-0.006*	-0.004	-0.016***
	(0.003)	(0.003)	(0.003)	(0.003)
age_cuad	0.001***	0.001***	0.001***	0.001***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
mujer_age	-0.003***	-0.004***	-0.004***	-0.004***
	(0.001)	(0.001)	(0.001)	(0.001)
num_hijos	0.014			
	(0.015)			
años_educ	-0.052***	-0.052***	-0.098***	
	(0.012)	(0.012)	(0.007)	
años_educ_cuad	0.004***	0.004***	0.005***	
	(0.0004)	(0.0004)	(0.0002)	
factor(p6210)3	0.112	0.113		-0.166***
	(0.070)	(0.070)		(0.061)
factor(p6210)4	-0.019	-0.019		-0.403***
	(0.086)	(0.086)		(0.065)
factor(p6210)5	-0.139	-0.138		-0.507***
	(0.098)	(0.098)		(0.068)
factor(p6210)6	-0.171	-0.171		-0.383***
	(0.105)	(0.105)		(0.072)
factor(p6210)9	-0.338	-0.339		-0.708
	(0.616)	(0.616)		(0.621)
factor(oficio)2	-0.005	-0.004	-0.003	-0.115
	(0.144)	(0.144)	(0.144)	(0.146)
factor(oficio)3	-0.347**	-0.346**	-0.372***	-0.625***
	(0.144)	(0.144)	(0.144)	(0.145)
factor(oficio)4	1.159***	1.158***	1.155***	0.983***
	(0.286)	(0.286)	(0.286)	(0.289)
factor(oficio)5	-0.014	-0.014	-0.014	-0.084
	(0.177)	(0.177)	(0.177)	(0.179)
factor(oficio)6	0.030	0.031	0.033	-0.007
	(0.149)	(0.149)	(0.149)	(0.151)
factor(oficio)7	-0.163	-0.162	-0.164	-0.303*
	(0.155)	(0.155)	(0.155)	(0.156)
factor(oficio)8	0.084	0.085	0.086	-0.019
	(0.147)	(0.147)	(0.147)	(0.149)
factor(oficio)9	0.199	0.200	0.203	0.108
	(0.157)	(0.157)	(0.158)	(0.159)
factor(oficio)11	-0.083	-0.082	-0.082	-0.208
	(0.145)	(0.145)	(0.145)	(0.147)
factor(oficio)12	0.164	0.165	0.165	0.101
	(0.145)	(0.145)	(0.146)	(0.147)
factor(oficio)13	-0.285**	-0.285**	-0.288**	-0.388***
	(0.143)	(0.143)	(0.143)	(0.144)
factor(oficio)14	-0.845***	-0.839***	-0.863***	-1.065***
	(0.286)	(0.286)	(0.287)	(0.289)
factor(oficio)15	-0.053	-0.052	-0.051	-0.166
	(0.160)	(0.160)	(0.160)	(0.161)
factor(oficio)16	-0.190	-0.189	-0.201	-0.377**
	(0.149)	(0.149)	(0.150)	(0.151)
factor(oficio)17	-0.241	-0.241	-0.261	-0.420***
	(0.159)	(0.159)	(0.159)	(0.161)
factor(oficio)18	-0.395**	-0.394**	-0.409**	-0.593***
	(0.163)	(0.163)	(0.163)	(0.164)

factor(oficio)19	-0.258*	-0.257*	-0.263*	-0.409***
	(0.146)	(0.146)	(0.147)	(0.148)
factor(oficio)20	-0.353	-0.353	-0.372	-0.498
	(0.454)	(0.454)	(0.454)	(0.459)
factor(oficio)21	0.208	0.209	0.202	0.065
	(0.142)	(0.142)	(0.142)	(0.144)
factor(oficio)30	-0.025	-0.024	-0.033	-0.188
	(0.151)	(0.151)	(0.151)	(0.152)
factor(oficio)31	0.189	0.189	0.186	0.206
	(0.211)	(0.211)	(0.211)	(0.213)
factor(oficio)32	-0.320**	-0.319**	-0.346**	-0.583***
	(0.151)	(0.151)	(0.151)	(0.152)
factor(oficio)33	-0.354**	-0.353**	-0.378***	-0.615***
	(0.144)	(0.144)	(0.144)	(0.145)
factor(oficio)34	-0.387**	-0.387**	-0.414***	-0.659***
	(0.160)	(0.160)	(0.160)	(0.162)
factor(oficio)35	-0.357	-0.357	-0.379	-0.613**
	(0.239)	(0.239)	(0.239)	(0.241)
factor(oficio)36	-0.664***	-0.664***	-0.687***	-0.952***
	(0.166)	(0.166)	(0.167)	(0.168)
factor(oficio)37	-0.550***	-0.549***	-0.576***	-0.815***
	(0.149)	(0.149)	(0.149)	(0.150)
factor(oficio)38	-0.520***	-0.520***	-0.546***	-0.794***
	(0.146)	(0.146)	(0.146)	(0.147)
factor(oficio)39	-0.417***	-0.416***	-0.440***	-0.666***
	(0.142)	(0.142)	(0.142)	(0.143)
factor(oficio)40	-0.202	-0.200	-0.217	-0.430**
	(0.169)	(0.169)	(0.170)	(0.171)
factor(oficio)41	-0.324**	-0.323**	-0.345**	-0.571***
	(0.143)	(0.143)	(0.143)	(0.144)
factor(oficio)42	-0.180	-0.179	-0.193	-0.373**
	(0.161)	(0.161)	(0.161)	(0.163)
factor(oficio)43	0.199	0.200	0.178	-0.039
	(0.196)	(0.196)	(0.196)	(0.198)
factor(oficio)44	-0.098	-0.097	-0.116	-0.315**
	(0.146)	(0.146)	(0.146)	(0.147)
factor(oficio)45	-0.509***	-0.508***	-0.532***	-0.758***
	(0.142)	(0.142)	(0.142)	(0.143)
factor(oficio)49	-0.057	-0.054	-0.086	-0.292
	(0.286)	(0.286)	(0.287)	(0.289)
factor(oficio)50	-0.117	-0.117	-0.141	-0.357**
	(0.173)	(0.173)	(0.173)	(0.175)
factor(oficio)51	-0.219	-0.217	-0.234	-0.472***
	(0.154)	(0.154)	(0.154)	(0.155)
factor(oficio)52	-0.502*	-0.503*	-0.506*	-0.758***
	(0.286)	(0.286)	(0.287)	(0.289)
factor(oficio)53	-0.389***	-0.389***	-0.412***	-0.641***
	(0.143)	(0.143)	(0.143)	(0.144)
factor(oficio)54	-0.394***	-0.393***	-0.412***	-0.653***
	(0.143)	(0.143)	(0.144)	(0.145)
factor(oficio)55	-0.544***	-0.544***	-0.558***	-0.812***
	(0.144)	(0.144)	(0.144)	(0.145)
factor(oficio)56	-0.477***	-0.478***	-0.499***	-0.729***
	(0.167)	(0.167)	(0.167)	(0.169)
factor(oficio)57	-0.465***	-0.464***	-0.489***	-0.731***
	(0.147)	(0.147)	(0.147)	(0.148)
factor(oficio)58	-0.466***	-0.465***	-0.493***	-0.725***
	(0.143)	(0.143)	(0.143)	(0.144)
factor(oficio)59	-0.466***	-0.465***	-0.491***	-0.739***
	(0.144)	(0.144)	(0.144)	(0.145)
factor(oficio)60	0.349	0.348	0.301	0.047
	(0.454)	(0.454)	(0.455)	(0.459)
factor(oficio)61	-0.294	-0.294	-0.311*	-0.511***
	(0.184)	(0.184)	(0.184)	(0.186)
factor(oficio)62	-0.394**	-0.393**	-0.417**	-0.652***
	(0.173)	(0.172)	(0.173)	(0.174)
factor(oficio)63	-0.116	-0.120	-0.147	-0.336
	(0.454)	(0.454)	(0.455)	(0.459)
factor(oficio)70	-0.274*	-0.274*	-0.292*	-0.505***
	(0.161)	(0.161)	(0.161)	(0.163)
factor(oficio)72	-0.536**	-0.530**	-0.541**	-0.786***
	(0.258)	(0.258)	(0.258)	(0.261)
factor(oficio)73	0.182	0.185	0.186	-0.111
	(0.454)	(0.454)	(0.455)	(0.459)
factor(oficio)74	-0.495**	-0.492**	-0.522**	-0.742***
	(0.226)	(0.226)	(0.226)	(0.228)
factor(oficio)75	-0.786***	-0.786***	-0.818***	-1.021***
	(0.182)	(0.182)	(0.183)	(0.184)

factor(oficio)76	-0.306	-0.308	-0.295	-0.572
	(0.380)	(0.380)	(0.380)	(0.384)
factor(oficio)77	-0.437***	-0.436***	-0.452***	-0.692***
	(0.148)	(0.148)	(0.148)	(0.149)
factor(oficio)78	0.156	0.162	0.111	-0.039
	(0.627)	(0.627)	(0.627)	(0.634)
factor(oficio)79	-0.507***	-0.506***	-0.527***	-0.765***
	(0.144)	(0.144)	(0.144)	(0.145)
factor(oficio)80	-0.636***	-0.635***	-0.649***	-0.895***
	(0.153)	(0.153)	(0.153)	(0.154)
factor(oficio)81	-0.481***	-0.480***	-0.498***	-0.730***
	(0.153)	(0.153)	(0.153)	(0.154)
factor(oficio)82	-0.643	-0.642	-0.689	-0.893*
	(0.454)	(0.454)	(0.455)	(0.459)
factor(oficio)83	-0.414***	-0.412***	-0.433***	-0.672***
	(0.152)	(0.152)	(0.152)	(0.153)
factor(oficio)84	-0.381***	-0.380***	-0.397***	-0.641***
	(0.147)	(0.147)	(0.147)	(0.148)
factor(oficio)85	-0.335**	-0.333**	-0.360**	-0.588***
	(0.148)	(0.148)	(0.148)	(0.149)
factor(oficio)86	0.142	0.141	0.127	-0.141
	(0.307)	(0.307)	(0.308)	(0.311)
factor(oficio)87	-0.435***	-0.435***	-0.450***	-0.700***
	(0.149)	(0.149)	(0.150)	(0.151)
factor(oficio)88	-0.566**	-0.565**	-0.591**	-0.799***
	(0.258)	(0.258)	(0.258)	(0.260)
factor(oficio)89	-0.426**	-0.426**	-0.439**	-0.710***
	(0.196)	(0.196)	(0.197)	(0.198)
factor(oficio)90	-0.537***	-0.535***	-0.560***	-0.793***
	(0.162)	(0.162)	(0.163)	(0.164)
factor(oficio)91	-0.615***	-0.614***	-0.634***	-0.872***
	(0.216)	(0.216)	(0.216)	(0.218)
factor(oficio)92	-0.423***	-0.421***	-0.447***	-0.682***
	(0.163)	(0.163)	(0.163)	(0.164)
factor(oficio)93	-0.483***	-0.482***	-0.499***	-0.735***
	(0.152)	(0.152)	(0.152)	(0.153)
factor(oficio)94	-0.677***	-0.676***	-0.702***	-0.912***
	(0.169)	(0.169)	(0.169)	(0.170)
factor(oficio)95	-0.361**	-0.360**	-0.374***	-0.622***
	(0.144)	(0.143)	(0.144)	(0.145)
factor(oficio)96	0.190	0.190	0.215	0.011
	(0.454)	(0.454)	(0.454)	(0.459)
factor(oficio)97	-0.572***	-0.571***	-0.596***	-0.837***
	(0.144)	(0.144)	(0.144)	(0.145)
factor(oficio)98	-0.438***	-0.437***	-0.459***	-0.700***
	(0.143)	(0.143)	(0.143)	(0.144)
factor(oficio)99	-0.696***	-0.697***	-0.722***	-0.950***
	(0.153)	(0.153)	(0.153)	(0.154)
exp_pot_cuad	-0.001***	-0.001***	-0.001***	-0.001***
	(0.00005)	(0.00005)	(0.00005)	(0.00004)
sizeFirm2-5 trabajadores	0.160***	0.161***	0.160***	0.160***
	(0.016)	(0.016)	(0.016)	(0.016)
sizeFirm6-10 trabajadores	0.217***	0.217***	0.215***	0.211***
	(0.023)	(0.023)	(0.023)	(0.023)
sizeFirm11-50 trabajadores	0.240***	0.240***	0.238***	0.238***
	(0.020)	(0.020)	(0.020)	(0.020)
sizeFirmMás de 50 trabajadores	0.339***	0.340***	0.338***	0.352***
	(0.018)	(0.018)	(0.018)	(0.018)
totalHoursWorked	0.011***	0.011***	0.011***	0.011***
	(0.0003)	(0.0003)	(0.0003)	(0.0003)
formalFormal	0.368***	0.368***	0.368***	0.375***
	(0.014)	(0.014)	(0.014)	(0.014)
mujer:num_hijos	-0.024			
	(0.022)			
Constant	13.192***	13.187***	13.416***	13.806***
	(0.173)	(0.173)	(0.165)	(0.169)
Observations	16,138	16,138	16,138	16,138
R2	0.519	0.519	0.517	0.508
Adjusted R2	0.516	0.516	0.515	0.505
Residual Std. Error	0.610 (df = 16038)	0.610 (df = 16040)	0.611 (df = 16045)	0.617 (df = 16042)
F Statistic	174.980*** (df = 99; 16038)	178.583*** (df = 97; 16040)	187.035*** (df = 92; 16045)	174.069*** (df = 95; 16042)
Note:		*p<0.1; **p<0.05; ***p<0.01		

Fuente: Elaboración propia utilizando la base de datos del PS1

Se escoge el modelo No. 3, pues se descarta la variable num_hijos y la interacción con mujer, que no resultan representativos en los modelos. De igual forma se opta por utilizar el modelo con años de educación (variable continua) y no la variable categórica de nivel de educación, ya que se considera que existe una gran diferencia en años de educación en el nivel terciario.

- *Use FWL to repeat the above estimation, where the interest lies on β2. Do you obtain the same estimates?*

A continuación se presenta la aplicación del teorema FWL. Se encuentra que se cumple el teorema y se obtiene el mismo estimador.

Tabla 13. Aplicación Teorema FWL

	Dependent variable:	
	ln_ing	res_ing
	(1)	(2)
mujer	-0.005	
	(0.031)	
mujer_age	-0.004***	
	(0.001)	
res_mujer		-0.005
		(0.031)
res_age_mujer		-0.004***
		(0.001)
Observations	16,138	16,138
R2	0.517	0.012
Adjusted R2	0.515	0.012
Residual Std. Error	0.611 (df = 16045)	0.609 (df = 16135)
F Statistic	187.035*** (df = 92; 16045)	97.001*** (df = 2; 16135)
Note:		*p<0.1; **p<0.05; ***p<0.01

Fuente: Elaboración propia utilizando la base de datos del PS1

- *How should we interpret the β2 coefficient? How good is this model in sample fit? Is the gap reduced? Is this evidence that the gap is a selection problem and not a "discrimination problem"?*

El coeficiente de la variable mujer se reduce cuando se incorporan las variables de control, esto se observa al comparar el beta con el modelo inicial en el que solo se incluía la variable mujer como variable independiente.

$$\log(\text{ingtot}) = 14.071 - 0.196\,\text{mujer} + u$$

En el modelo anterior, el beta de la variable mujer indicaba que las mujeres ganaban 19.6% menos que los hombres.

En el modelo escogido con variables de control para identificar la brecha salarial condicionada, se encuentra que se reduce el beta de la variable mujer, por lo que se tiene que las mujeres ganan un 0,5% menos ingresos que los hombres.

En este modelo de brecha salarial condicionada, también se incluía la variable interacción mujer con edad (*mujer_age*), la cual indica que a medida que la mujer aumenta de edad (cumple un año más) sus ingresos se reducen en un 0,4% en comparación a lo de un hombre.

A pesar de que la brecha salarial se disminuye cuando se incorporan variables de control, la brecha salarial se mantiene. Por lo que se considera que la brecha salarial no se trata únicamente de un problema de selección.

Punto 5 - Predicting earnings.

Now we turn to prediction. You built a couple of models in the previous section using your knowledge as an applied economist, the task here is to assess the predictive power of these models.

- *Split the sample into two samples: a training (70%) and a test (30%) sample. Don't forget to set a seed (in R, set.seed(10101), where 10101 is the seed.)*

Se realiza la partición de la base de datos en las muestras de entrenamiento (training) y de prueba (test).

Tabla 14. Partición de muestras de entrenamiento

	Test = TRUE	**Test = FALSE**	**Total**
n	4919	11478	16397
%	30%	70%	100

Fuente: Elaboración propia utilizando la base de datos del PS1

Se comparan ambas tablas, para saber qué tan parecidos son ambos grupos:

Tabla 15. Comparación de las particiones

Comparación de las particiones de entrenamiento (70%) y prueba (30%)					
	FALSE		TRUE		p-test
n	11478		4919		
ingtot (mean (SD))	1786588.96	(2652067.35)	1760809.51	(2761352.70)	0.573
age (mean (SD))	39.66	(13.39)	39.55	(13.39)	0.622
mujer = 1 (%)	5346	(46.6)	2369	(48.2)	0.065

años_educ (mean (SD))		11.42	(4.36)	11.47	(4.34)	0.457
formal = Formal (%)		6766	(58.9)	2910	(59.2)	0.815
totalHoursWorked (mean	(SD))	47.47	(15.55)	47.38	(15.77)	0.739
exper_pot (mean (SD))		23.24	(15.11)	23.07	(15.07)	0.514
num_hijos (mean (SD))		0.18	(0.48)	0.18	(0.48)	0.484
p6426 (mean (SD))		5.37	(7.47)	5.31	(7.49)	0.657

Fuente: Elaboración propia utilizando la base de datos del PS1

Estas variables ya se han explicado previamente; solo vale la pena destacar que ambos grupos son estadísticamente similares en todas las variables, de acuerdo a los p-valores observados previamente (test de diferencia de medias).

- *Estimate a model that only includes a constant. This will be the benchmark.*

  La medición empleada para la comparación de los diferentes modelos de predicción es el MSE (error promedio cuadrático). Para el caso del modelo simple, que solo incluye una constante, el MSE es 7.624183e+12, que será el valor de referencia.

- *Estimate again your previous models*

  La siguiente tabla presenta los diferentes modelos que se estimarán, tanto en este inciso como en el siguiente. Es decir, son modelos previamente estimados en las secciones previas de este taller y nuevos modelos con variables, interacciones y formas funcionales:

Tabla 16. Modelos y variables regresoras

Modelo	**Variables regresoras**
1	1
2	edad
3	edad + edad_cuad
4	edad + edad_cuad + edad:formal
5	mujer
6	mujer + edad
7	mujer + edad + edad_cuad
8	mujer + edad + edad_cuad + mujer:edad
9	mujer + edad + edad_cuad + mujer:edad + años_educ + años_educ_cuad + oficio + exp_pot_cuad + sizeFirm + totalHoursWorked + formal
10	edad + edad_cuad + mujer + años_educ + años_educ_cuad + exper_pot + exp_pot_cuad + oficio + sizeFirm + formal + totalHoursWorked + p6426 + num_hijos
11	edad:mujer + edad_cuad + años_educ + años_educ_cuad + exper_pot:mujer + exp_pot_cuad + oficio + sizeFirm + formal + totalHoursWorked + p6426 + num_hijos:mujer
12	edad:mujer + edad_cuad:mujer + años_educ:mujer + años_educ_cuad:mujer + exper_pot:mujer + exp_pot_cuad:mujer + oficio:mujer + sizeFirm:mujer + formal:mujer + totalHoursWorked:mujer + p6426:mujer + num_hijos:mujer
13	edad:años_educ + edad_cuad + mujer + años_educ_cuad + exper_pot + exp_pot_cuad + edad:oficio + sizeFirm + formal:totalHoursWorked + p6426 + edad:num_hijos

14	edad:mujer + poly(edad,3) + poly(edad,4) + años_educ:mujer + poly(años_educ,3) + poly(años_educ,4) + exper_pot:mujer + poly(exper_pot,3) + poly(exper_pot,4) + oficio:mujer + sizeFirm:mujer + formal:mujer + poly(totalHoursWorked,3) + poly(totalHoursWorked,4) + p6426:mujer + poly(p6426,3) + poly(p6426,4) + num_hijos:mujer + poly(num_hijos,3)
15	edad:años_educ:exper_pot:totalHoursWorked:num_hijos:mujer +edad_cuad +poly(edad,5):poly(años_educ,5):poly(exper_pot,5):poly(totalHoursWorked,5) + años_educ_cuad +exp_pot_cuad +oficio +sizeFirm +formal + p6426

Fuente: Elaboración propia utilizando la base de datos del PS1

- *In the previous sections, the estimated models had different transformations of the dependent variable. At this point, explore other transformations of your independent variables also. For example, you can include polynomial terms of certain controls or interactions of these. Try at least five (5) models that are increasing in complexity.*

  Ver tabla presentada en el inciso anterior.

- *Report and compare the average prediction error of all the models that you estimated before. Discuss the model with the lowest average prediction error.*

  La siguiente table presenta el error de predicción promedio cuadrático (MSE) para los 15 modelos estimados.

Tabla 17. Predicción MSE

Modelo	MSE
1	7,62418E+12
2	7,53388E+12
3	7,51486E+12
4	6,89602E+12
5	7,59541E+12
6	7,50604E+12
7	7,48571E+12
8	7,4527E+12
9	5,05272E+12
10	5,03085E+12
11	5,02562E+12
12	5,06399E+12
13	5,11321E+12
14	5,02371E+12
15	2,55077E+15

Fuente: Elaboración propia utilizando la base de datos del PS1

- *For the model with the lowest average prediction error, compute the leverage statistic for each observation in the test sample. Are there any outliers, i.e., observations with high leverage driving the results? Are these outliers potential people that the DIAN should look into, or are they just the product of a flawed model?*

Como resultado de este procedimiento, se genera una matriz como la que se ilustra a continuación. Se presentan los primeros 10 elementos a modo de ejemplo:

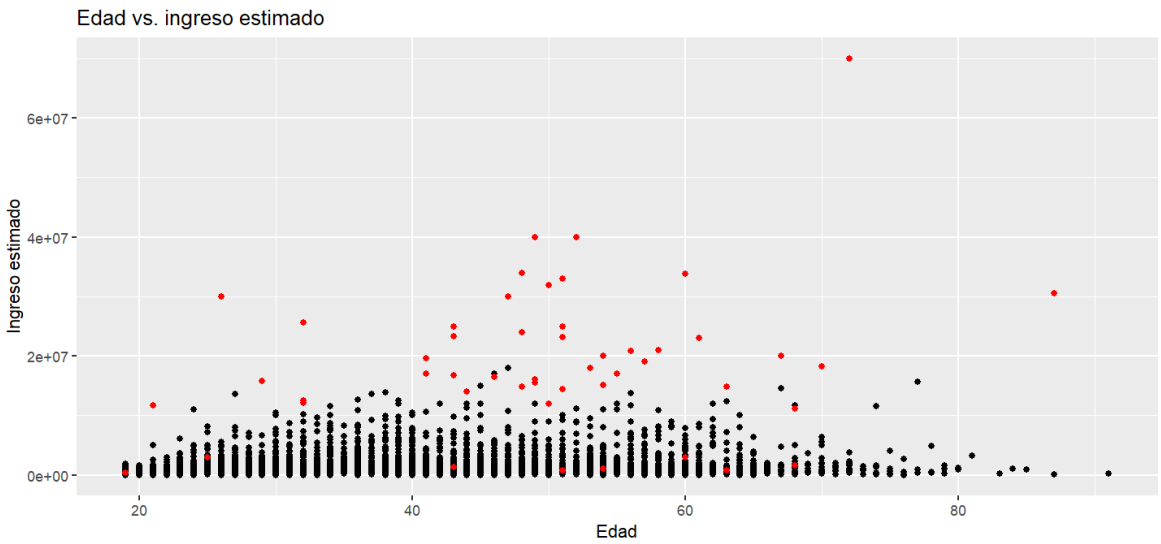
Tabla 18. Cálculo de las mediciones u, h y alpha

Cálculo de las mediciones u, h y alpha para una muestra de 10 observaciones				
Elemento_j	**u**	**h**	**alpha**	**abs_alpha**
1	-5,721.002	0.005	-5,748.844	5,748.844
2	-436,548.700	0.021	-446,052.500	446,052.500
3	-428,051.000	0.006	-430,454.400	430,454.400
4	-2,023,970.000	0.010	-2,043,857.000	2,043,857.000
5	291,748.400	0.017	296,721.100	296,721.100
6	690,224.400	0.003	692,600.200	692,600.200
7	-122,653.000	0.008	-123,591.600	123,591.600
8	-150,049.300	0.045	-157,166.700	157,166.700
9	-888,819.200	0.007	-894,778.600	894,778.600
10	337,152.200	0.033	348,785.100	348,785.100

Fuente: Elaboración propia utilizando la base de datos del PS1

El alpha (leverage) no dice nada por sí solo, y debe interpretarse de manera relativa, en comparación a los demás valores. Por lo tanto, organizamos de mayor a menor las observaciones de acuerdo con el valor absoluto de su alpha. Con fines ilustrativos, incluimos en la gráfica de dispersión las 50 observaciones con mayor alpha. De nuevo, no hay un umbral determinado para definir cuando una observación se considera un outlier, y dependerá de cada caso qué tratamiento debe dársele. Con el presente conjunto de datos, sería pertinente eliminar dos o tres outliers que generan un peso importante

Ilustración 10. Edad versus ingresoestimado



Fuente: Elaboración propia utilizando la base de datos del PS1

(a) Repeat the previous point but use K-fold cross-validation. Comment on simi-

larities/differences of using this approach.

La siguiente tabla compara los MSE usando ambos enfoques (1. Partición 70/30 de la base de datos. 2. Validación cruzada *k-fold* con K=10.)

Tabla 19. Comparación MSE

Modelo	MSE	MSE_CV_10-fold
1	7,62418E+12	7,10893E+12
2	7,53388E+12	6,94012E+12
3	7,51486E+12	6,92651E+12
4	6,89602E+12	6,39179E+12
5	7,59541E+12	7,0839E+12
6	7,50604E+12	6,98221E+12
7	7,48571E+12	6,9343E+12
8	7,4527E+12	6,89353E+12
9	5,05272E+12	4,54516E+12
10	5,03085E+12	4,55818E+12
11	5,02562E+12	4,5418E+12
12	5,06399E+12	4,56417E+12
13	5,11321E+12	4,64285E+12
14	5,06535E+12	4,55666E+12
15	2,55077E+15	5,38717E+14

Fuente: Elaboración propia utilizando la base de datos del PS1

Como se puede observar en la tabla, con el método de validación cruzada "K-fold", con K=10, se obtienen en general MSE menores a los obtenidos con la partición 70/30 de la muestra. La principal desventaja de la partición 70/30 es que depende mucho de la aleatoriedad con que se haya generado la muestra, y los resultados pueden variar considerablemente si la muestra fuera otra. El modelo con menor error cuadrático promedio fue el 11, el cual termina siendo nuestro modelo recomendado para la predicción:

$$modelo_{11} = \text{lm}(\text{ingtot} \sim \text{edad:mujer} + \text{edad}^2 + \text{años}_{\text{educ}} + \text{años}_{\text{educ}}^2 + \text{exper\_pot:mujer} + \text{exp}_{\text{pot}}^2 + \text{oficio} + \text{sizeFirm} + \text{formal} + \text{totalHoursWorked} + \text{antigüedad\_empleo} + \text{num\_hijos:mujer})$$

El modelo incorpora las variables de interés que presentamos previamente, además de interacciones que consideramos relevantes con la variable mujer.

(b) *LOOCV*. With your preferred predicted model (the one with the lowest average prediction error) perform the following exercise:

i. Write a loop that does the following:

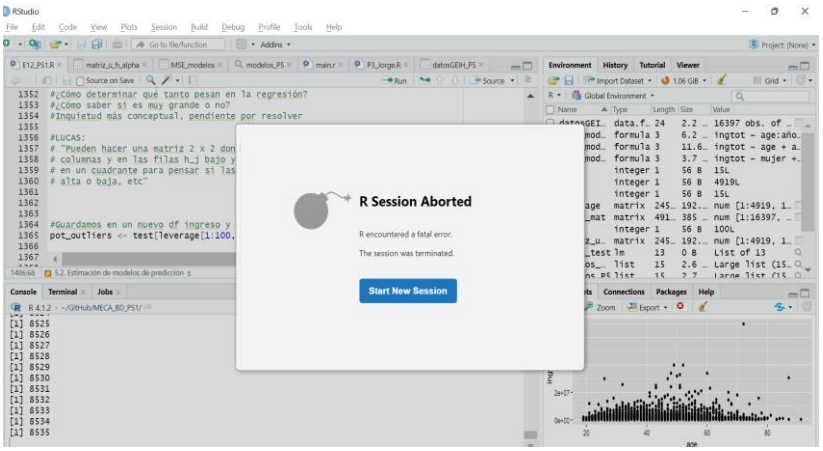
- Estimate the regression model using all but the $i-th$ observation.

- Calculate the prediction error for the $i-th$ observation, i.e. $(y_i - \hat{y}_i)$
- Calculate the average of the numbers obtained in the previous step to get the average mean square error. This is known as the Leave-One-Out Cross-Validation (LOOCV) statistic.

ii. Compare the results to those obtained in the computation of the leverage statistic

Teniendo en cuenta que el modelo 11 es relativamente complejo (incluye múltiples términos e interacciones), el procedimiento de LOOCV es bastante intensivo computacionalmente. Con el computador disponible no se logró completar el cálculo, generándose un error después de 8535 ciclos:

Ilustración 11. Error modelo 11



En general, para un modelo más sencillo, LOOCV se tomó alrededor de 30 minutos, mientras que el de K-fold se calcula en cuestión de segundos. Lo más importante en este caso es que la diferencia entre uno y otro es muy pequeña (de nuevo, para el caso de un modelo sencillo, el #2, fue de 0,33%). En esta línea, no se justifica hacer el LOOCV, por lo intensivo en cómputo, y por las desventajas teóricas en cuanto al trade-off sesgo – varianza.

En resumen, la validación cruzada K-fold nos entrega los mejores resultados, pues es mucho menos intensiva en cómputo si se compara con LOOCV, no depende tanto de la conformación aleatoria de la muestra de entrenamiento y de prueba, como sucede con la partición 70/30, y representa un buen balance entre sesgo y varianza, sin que ninguno de estos dos valores sea extremadamente alto.

BIBLIOGRAFÍA

Bolaños, A. (2018). La brecha salarial son los hijos. *El País*, pág. https://elpais.com/politica/2018/03/02/actualidad/1520006491_549539.html.

Cerquera, O., Arias, C., Prado, J. (2020). La Brecha Salarial por género en Colombia y en el Departamento de Caldas. Recuperado el 24 de junio del 2022, de https://www.redalyc.org/journal/3578/357863806006/html/

DNP-DANE. (2012). *Misión para el Empalme de las Series de Empleo, Pobreza y Desigualdad* . Bogotá D.C: DNP-DANE.

Figueroa, A. (2010). ¿Mejora la distribución del ingreso con la educación? El caso de Perú. Recuperado el 25 de junio del 2022, de https://repositorio.cepal.org/bitstream/handle/11362/11422/102115136_es.pdf?sequence=1&isAllowed=y

Giraldo, C., Cardona, D. Ser viejo en colombia tiene su costo laboral. Recuperado el 25 de junio del 2022, de http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0124-81462010000200005

Guataquí, J., García, A., Rodríguez, M. (2009). Estimaciones de los determinantes de los ingresos laborales en Colombia con consideraciones diferenciales para asalariados y cuenta propia. Recuperado el 24 de junio del 2022, de https://repository.urosario.edu.co/bitstream/handle/10336/10851/5756.pdf

Laez, F., Jiménez, M. (2011). La importancia de la educación para reducir la inequidad. Recuperado el 24 de junio del 2022, de https://www.uv.mx/cienciahombre/revistae/vol24num1/articulos/educacion/

Lora, E. (2021). La rentabilidad de estudiar y las ocupaciones en riesgo. Recuperado el 25 de junio del 2022, de https://www.eltiempo.com/economia/sectores/rentabilidad-de-estudiar-y-ocupaciones-en-riesgo-626880

Madrigal, B. (2009). Capital humano e intelectual: su evaluación. Recuperado el 24 de junio del 2022, de https://www.redalyc.org/pdf/2190/219016838004.pdf

Muñóz, C. (2004). Determinantes del ingreso y del gasto corriente de los hogares. Recuperado del 23 de junio del 2022, de http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0124-59962004000100008#:~:text=Lo%20que%20se%20espera%20es,ingreso%20debido%20a%20razones%20culturales

SA. (2018). Mercado laboral urbano – resultados 2018: Bogotá. Recuperado el 24 de junio del 2022, de https://colaboracion.dnp.gov.co/CDT/Estudios%20Econmicos/3%20Informe%20Bogota%202018.pdf

Quintero, E. S. (2013). *Factores determinantes de los salarios en Colombia - Trabajo Fin de Máster*. Universidad de Zaragoza.