Problem Set 1: Predicting Income

Paula Ramos, Karen Uribe-Chaves
Juan D. Urquijo

June 26 2022 -
Repository: *Github*

1 *Data acquisition*

(a) Scrape the data that is available at the following website: *Link*

Desarrollo: Ver código R - Data Acquisition.

(b) Are there any restrictions to accessing/scraping these data?

Desarrollo: No hay restricciones dado que al correr el protocolo de exclusión para robots o "robots.txt" no existe, por lo tanto se es libre para scrappear la página.

(c) Using pseudocode describe your process of acquiring the data

Desarrollo: Se realizan un número de pasos para poder adquirir la información:

1. Se identifica si la página web es estática o dinámica. En este caso la página es dinámica, por lo tanto se requiere buscar dentro del desarrollador el link personalizable para cada "chunk" de datos. El link es: *Pagina para web scrapping*

2. Al identificar el link, se toma como referencia el primer "chunk" y se extrae la tabla a partir de la función *html table*; y se valida que la tabla contenga la información necesaria.

3. Dado que los pasos anteriores, únicamente se realizaron para el primer "chunk", se construye un *for* en R para realizar el mismo proceso en cada uno de los chunks (del 1 al 10) y así, para extraer toda las tablas.

4. Una vez se obtienen todas las tablas de cada "chunk" se unen con un *rbind*

5. Finalmente, se valida con la fuente principal de datos que la base extraida por medio de *web scrapping* esté completa y sea consistente. En caso afirmativo, se procede con el ejercicio.

2 *Data Cleaning*

(a) 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).

Desarrollo: De acuerdo a la teoría económica, y en el análisis de los principales determinantes del ingreso de los individuos se ha encontrado que la educación y la ocupación son dos de los determinantes más importantes del nivel de ingresos de los hogares. En este punto, Su y Heshmati (2013), encuentran que estos dos factores ejercen efectos heterogéneos en los diferentes percentiles de la distribución del ingreso en China. Además, ante la diferencia entre áreas rurales y urbanas, el tipo de oficio ofrece evidencia de las demandas de trabajo en las diferentes áreas.

De esta manera se seleccionan las primeras variables de la Gran Encuesta Integrada de Hogares de 2008 (GEIH) obtenida por medio de web-scrapping:

- Escolaridad: la variable "maxEducLevel" obtenida de la base, es una variable categórica que toma valores desde 1 hasta 9, indicando el grado más alto de educación alcanzado por el individuo encuestado; siendo 1 el nivel más bajo.
- Ocupación: la variable "oficio" obtenida de la base, es una variable categórica que toma valores desde 1 hasta 99, permitiendo identificar la ocupación general a la que pertenece el individuo. Las variables identificadas pueden incluir: (1) Químicos, (2) Arquitectos o Ingenieros, (3) Dibujantes, (4) Pilotos, etc.
- Tipo de vinculación al mercado laboral: Producto de la evidencia relacionada con la ocupación, se considera que el tipo de inserción al mercado laboral se relaciona con la ocupación del individuo. De esta manera se toma la variable "formal", que toma el valor de 1 si el individuo pertenece al mercado formal (cotiza a seguridad social) y 0 de lo contrario. Esta variable permitiría reconocer las diferencias entre estos dos grupos y estimar funciones de ingreso para ambos tipos de trabajadores.

Por otro lado, Wodon (2000) a partir de la Encuesta Nacional de Hogares de Bangladesh encuentra que la ubicación del individuo (rural / urbano) afecta los retornos de la educación, y por tanto el ingreso de las personas. Siguiendo esta evidencia se validan las siguientes variables:

- Área: Se valida si en la muestra con la variable "clase" los individuos todos pertenencen al área urbana. Dado que la muestra seleccionada es únicamente en Bogotá, no se encontró individuos en el área rural.
- Estrato: A pesar que no se encontraron individuos en el área rural, consideramos que es necesario validar si la ubicación del individuo dentro de la ciudad genera algún efecto en la distribución del ingreso y por tanto en la predicción del mismo. Por lo anterior, se tomó la variable "estrato1" que nos permite identificar el estrato socioeconómico del individuo, acorde a la teoría económica.

Finalmente, siguiendo la función de ingresos de Mincer que analiza las características de la inversión de capital humano (educación, experiencia), tipo de empleo, género, raza, edad, etc.; se analizan las siguientes variables:

- Experiencia laboral: La variable "p6426" permite identificar los meses de experiencia laboral del individuo. Dado que la medición de la experiencia puede ser complicada y no exacta, se suele utilizar la aproximación de la variable con el tiempo o antigüedad en el empleo (o empleos similares).
- Edad: Según la teoría económica, se espera que a mayor edad, mayor ingreso; pero ante los rendimientos marginales de la experiencia ya mencionados se toma la variable al cuadrado.
- Sexo: La desigualdad de ingresos y heterogeneidad en el mercado laboral, hace necesario identificar los efectos de ser mujer en el salario total percibido.

De esta manera se tienen las principales variables que se cree afectan el ingreso del individuo, además de identificadores como el hogar, la relación con el jefe del hogar, si tiene un segundo trabajo y el tamaño de la empresa.

(b) 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.

Desarrollo: Los pasos realizados para la eliminación de los *missing values (NAs)* se describen a continuación:

- En primer lugar, se toma la base extraida por *web scrapping* y se analizan las variables contenidas en ella por medio del comando `skim`. Utilizando el diccionario, identificamos las variables categóricas para volverlas tipo factor. Esto permite que no se tomen como un valor numérico y puedan interpretarse en futuras estimaciones.
- Posterior a esto filtramos la base para tener únicamente los individuos mayores de 18 años y que son ocupados; lo que nos deja con una cantidad de NAs menor. Una vez se filtra y se seleccionan las variables del punto (a) procedemos a lidiar con los *missing values*, sacando la cantidad de NAs restantes por variable.
- Al tener la cantidad de NAs por variable, es posible calcular el porcentaje de observaciones faltantes para cada una de ellas, lo que nos brinda una medida estandarizada para conocer el porcentaje de *missing values* para cada variable.
- Calculados los NAs se procede a la imputación de variables. En este caso, lo primero que se realiza es identificar cuantas observaciones de "ingtot" eran menores o iguales a cero. Para esta variable, el método de imputación elegido, al ser una variable numérica, es a partir de la media de las ingresos del hogar, identificado con la variable "directorio".
- Una vez realizada la imputación, validamos que no quede ningún ingreso menor o igual a cero. En este caso, se encontró una observación faltante y se imputó con la media del total de la muestra. A través de una nueva validación, se tiene que no quedan *missing values* restantes en la base a estudiar.

(c) 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.

Desarrollo:

Las estadísticas descriptivas son piezas de información que permiten comprender y representar un conjunto de datos. Se utilizan para describir las principales características de la información numérica y categórica de la base de datos en estudio. Iniciando con las variables numéricas, encontramos:

Table 1: Estadísticas Descriptivas Variables Numéricas

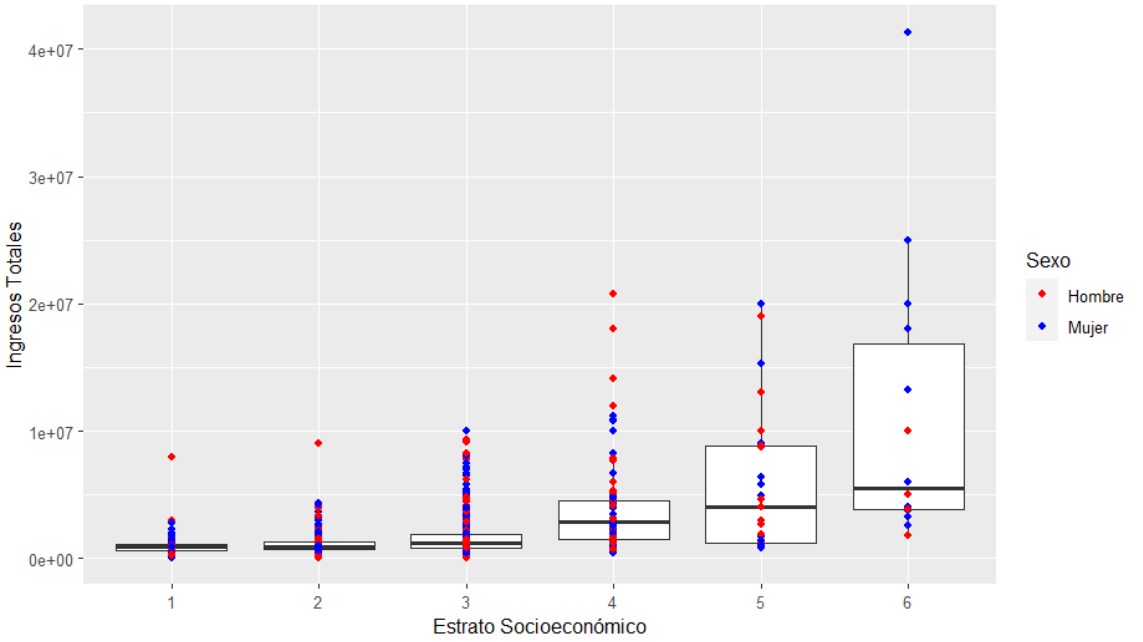
Statistic	Ingresos Totales	Edad	Meses Experiencia Laboral
N	1,551	1,551	1,551
Mean	1,587,665	40	69
St. Dev.	2,233,349	14	95
Min	20,000	18	0
Median	1,003,357	38	30
Max	41,333,333	84	600

En la Table 1 se encuentra que para un total de 1.551 observaciones, los ingresos totales de la población de la muestra tienen un promedio de $1.587.665 COP y una desviación estándar de $2.233.394 COP. El ingreso mínimo encontrado en la muestra estudiada es de $20.000 COP y el máximo de 41.333.333 COP, con una mediana de $1.003.357 COP; mostrando gran dispersión en los datos.

Por su parte, en la misma muestra se encuentra que la edad promedio es de 40 años con una desviación estándar de 14 años. La persona más joven de la muestra tiene 18 años, dado el filtro realizado para tomar únicamente los mayores de 18 años, y la persona de mayor edad tiene 84 años, con una mediana de 38. Finalmente, en lo referente a las variables numéricas se analiza los meses de experiencia laboral. Se encuentra que en la muestra se tiene un promedio de experiencia laboral de 69 meses con una desviación estándar de 95 meses. La mínima experiencia laboral es de 0 meses, y la máxima de 600 meses, con una mediana de 30 meses de experiencia que de nuevo denota alta dispersión, y tiene sentido con la distribución del ingreso observada.

Las variables categóricas por su parte, nos permiten identificar las características de ciertos grupos asociados a las variables numéricas. Para interés de la estimación, en primer lugar se analiza el ingreso comparado con el estrato socioeconómico y el sexo, estas dos últimas como variables categóricas:

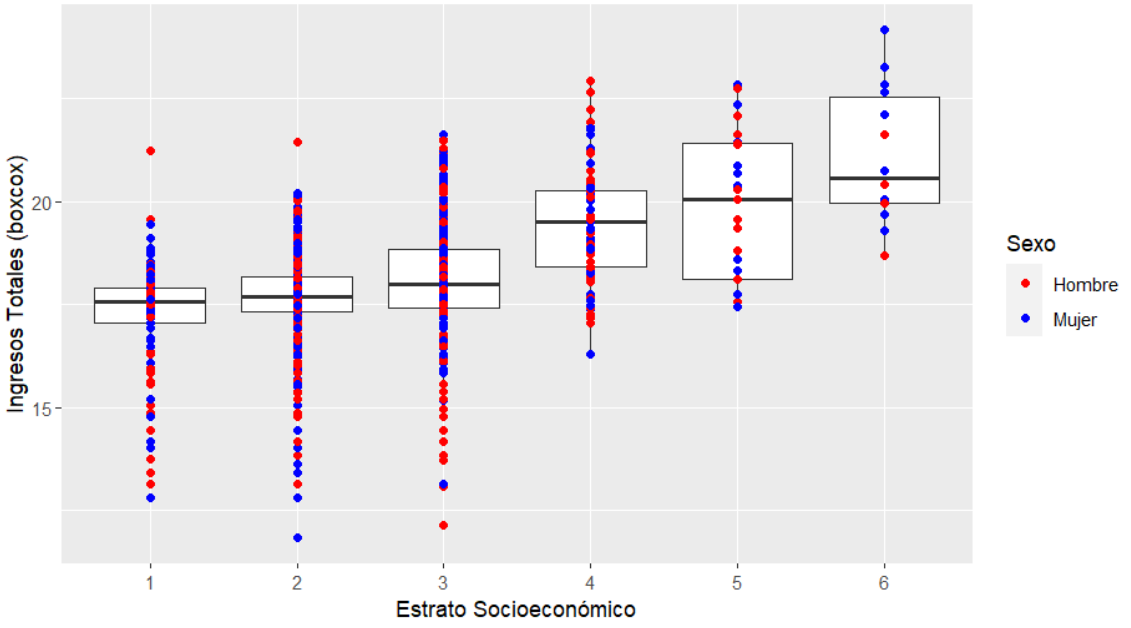
Figure 1: Cajas y Bigotes - Ingreso vs. Estrato por Sexo



Las dimensiones de las cajas en la Figure 1 están determinada por la distancia del rango intercuartílico (IQR), que es la diferencia entre el primer y tercer cuartil; lo que nos brinda una idea de que tan dispersos se encuentran los datos. Por ejemplo, para el estrato 1, el IQR es muy pequeño y por tanto indica que la dispersión del salario en este estrato es baja; lo que contrasta con el IQR del estrato 6 en donde la dispersión es mucho mayor. Los bigotes, por su parte, nos muestran el límite para detectar valores atipicos. Como punto de comparación, se observa que en el estrato 6 existe un valor atítpico que corresponde al máximo de $41.333.333 COP del salario, y además nos indica que este salario lo recibe un hombre (punto azul).

Al observar la Figure 2, se identifican valores atípicos y por tanto procedemos a realizar una transformación boc-cox del ingreso total, para que nos permita observar la distribución de manera más clara. Se obtiene:

Figure 2: Cajas y Bigotes - Ingreso (Box-cox) vs. Estrato por Sexo



El ajuste de la distribución en la Figure 2 nos permite ver el crecimiento de los ingresos y la dispersión de los mismos de manera más clara por estrato y sexo, además de reducir los valores atípicos que distorsionaban la distribución. En el estrato 1 y 2, se observa que son dos mujeres las que tienen mayor ingreso, lo que contrasta con el estrato 6 que siguen siendo los hombres los que lideran la distribución en este grupo. Por su parte, la mediana representada por el centro de la caja, nos permite identificar la asimetría de las distribuciones. En este caso, para el estrato 1, 4 y 5 se observa asimetría negativa, lo que indica que los datos se concentran en la parte superior de la distribución, y por tanto se espera que la media sea menor. En contraste, para los estratos 2, 3 y 6, la asimetría es positiva indicando que los datos se concentran en la parte interior de la distribución, y la media para estos grupos es mayor a su mediana.

Consideramos que es relevante entender la formalidad de la economía, y por tanto observar la distribución del ingreso y la edad. Para esto se construyó un gráfico de dispersión. La Figure 3 no nos muestra con claridad la dispersión de los datos, sin embargo se puede extraer que los trabajadores formales (que cotizan) tienen mayor salario que aquellos informales, sobre todo en la edad entre los 40 y 60 años. Para observar mejor la distribución, se realiza de nuevo el análisis con el ingreso transformado por medio de box-cox.

Figure 3: Dispersión - Ingreso vs. Edad y Formalidad

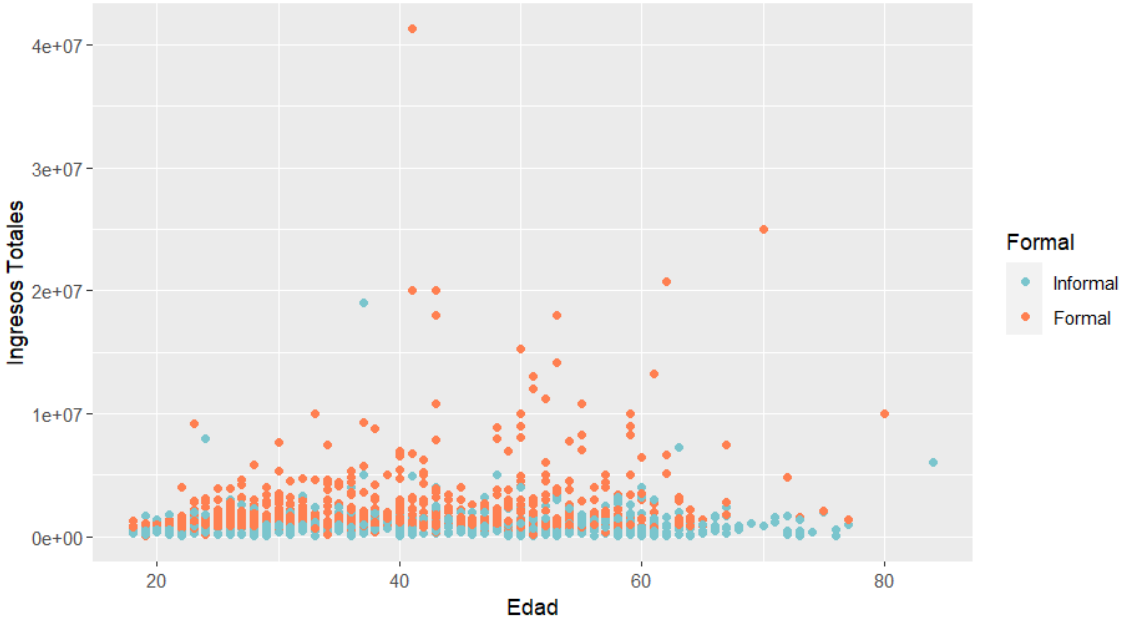
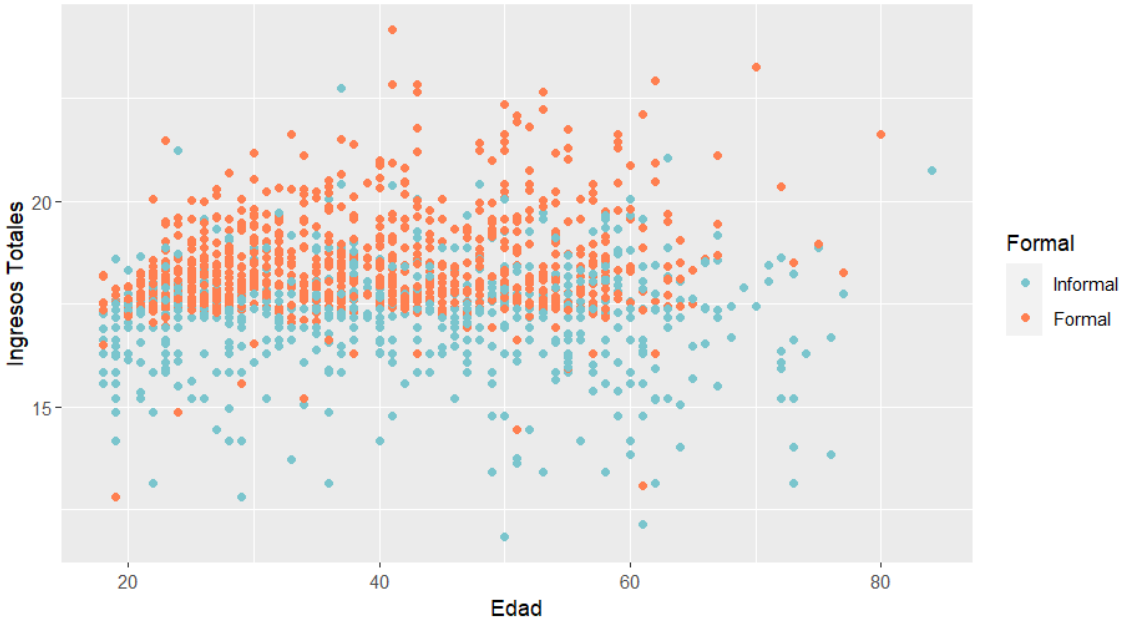


Figure 4: Dispersión - Ingreso (boxcox) vs. Edad y Formalidad



La Figure 4 con los valores del ingreso escalado, dan cuenta de mayores ingresos para los trabajadores formales sistematicamente para todas las edades. En otras palabras, el ingreso de los trabajadores informales de Bogotá es menor y más inequitativo que el de los ocupados formales.

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.

(a) 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.

Desarrollo: Se tomó como punto de referencia para la medición del ingreso de los trabajadores la variable "ingtot" que corresponde a los Ingresos Totales del individuo perteneciente al hogar. Este corresponde a:

$$Ingtot = Ingtotob + Ingtotes \tag{1}$$

Donde:

Ingtot: Ingreso total por persona que resulta de sumar cada una de las fuentes de ingresos tanto observadas como imputadas

Ingtotob: Ingreso total observado por persona que resulta de sumar los ingresos percibidos en las siguientes fuentes: ingreso monetario primera actividad (impa), ingreso segunda actividad (isa), ingreso en especie (ie), ingreso monetario desocupados e inactivos (imdi) e ingresos provenientes de otras fuentes no laborales (iof) (intereses, pensiones, ayudas, cesantias, arriendos y otros)

Ingtotes: Ingreso total por persona que resulta de sumar cada una de las fuentes de ingresos imputadas a los registros faltantes

Esto nos permite reducir las imputaciones y considerar todas las fuentes de ingreso para cada individuo.

(b) Based on this estimate using OLS the age-earnings profile equation:

$$Income = \beta_1 + \beta_2 Age + \beta_3 Age^2 + u \tag{2}$$

Table 2: Age-earnings profile

	Dependent variable:	
	Ingreso Total	Ingreso Total_(box cox)
	(1)	(2)
Edad	79,117.390***	0.095***
	(24,766.020)	(0.015)
Edad^2	−708.042**	−0.001***
	(289.505)	(0.0002)
Constant	−307,848.700	16.006***
	(491,941.800)	(0.295)
Observations	1,551	1,551
R^2	0.018	0.026
Adjusted R^2	0.016	0.025
Residual Std. Error (df = 1548)	2,215,012.000	1.328
F Statistic (df = 2; 1548)	13.885***	20.878***
Note:		*p<0.1; **p<0.05; ***p<0.01

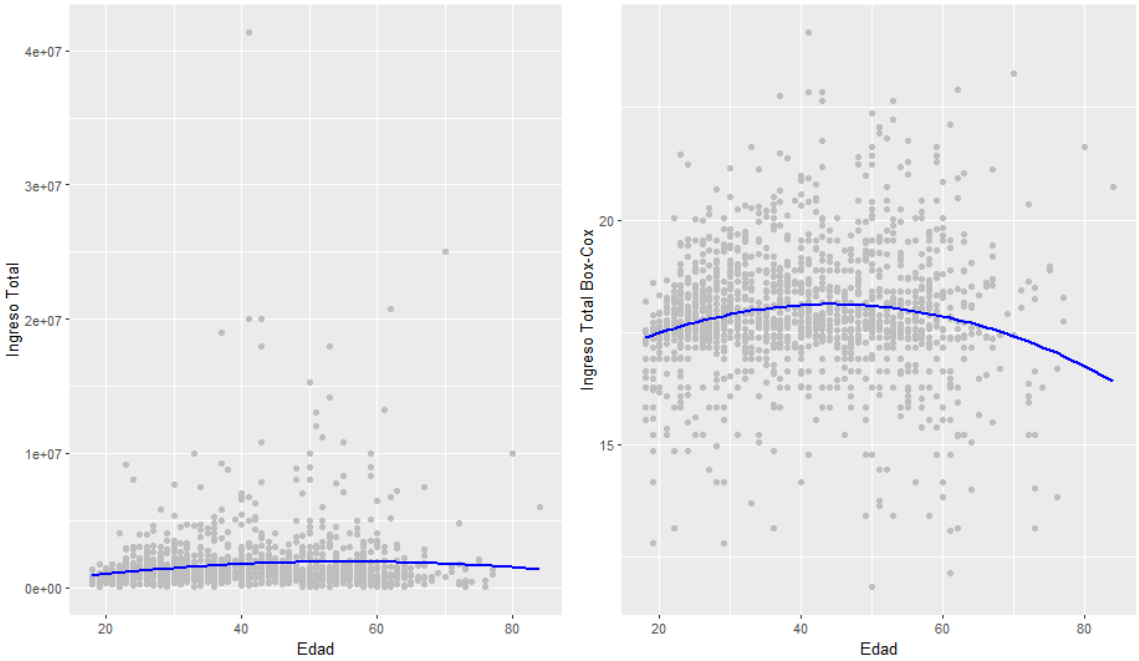
How good is this model in sample fit?

Desarrollo: La importancia de conocer si un modelo tiene buen ajuste a los datos, radica en proporcionar buenas respuestas a la pregunta de investigación. En este caso en particular, el modelo de regresión lineal estimado por *OLS*, teóricamente minimiza la distancia entre la línea ajustada y todos los puntos de datos, o en otras palabras, minimiza la suma de los errores al cuadrado.

Para conocer que tan buen ajuste tiene el modelo, se puede analizar su R^2 que refleja la bondad del ajuste de un modelo al ingreso total, que es la variable que se quiere explicar. En la Table 2, se observa que el R^2 para el modelo 1 sin transformaciones, es de 0.018, lo que nos indica que el modelo tiene un poder de explicación bajo en relación con la variablidad de los datos. Del mismo modo, para el modelo con la variable de ingreso transformada por box-cox se observa que el R^2 mejora hasta 0.026, pero sigue siendo bajo. Para entender intuitivamente como se construye el R^2 , su ecuación asociada es:

$$R^2 = \frac{\text{Variacion Explicada del Modelo}}{\text{Total Variación del Modelo}}$$

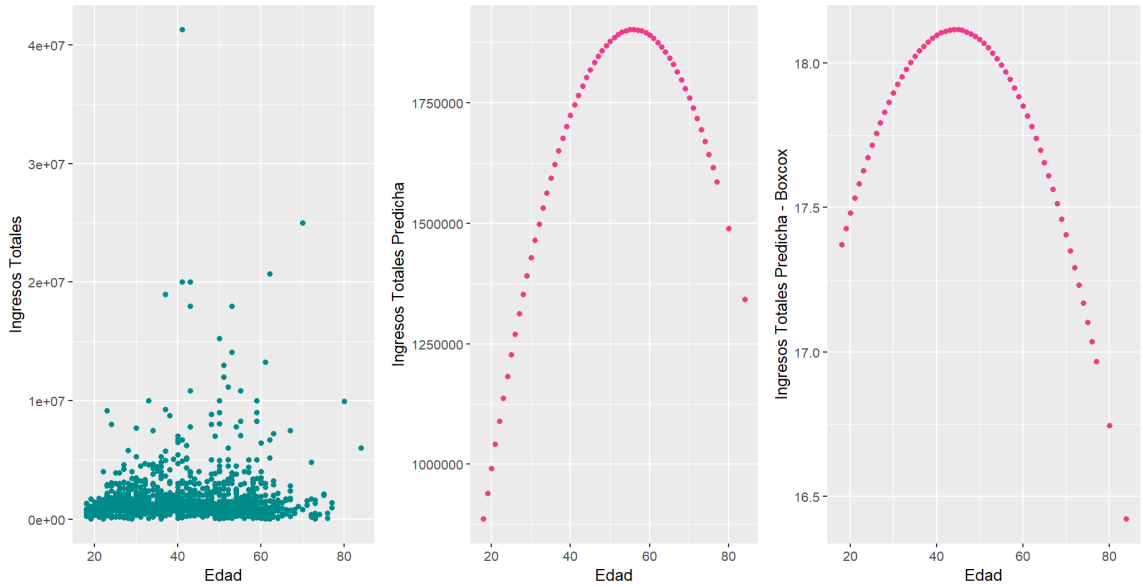
Figure 5: Fit de los Modelos de Age Earnings Profile (modelo 1)



De la misma manera encotramos que el ajuste del modelo no es bueno por la gráfica de las predicciones comparada con la variable de ingreso total para ambas estimaciones. La Figure 5 nos muestra que la línea de predicción (línea azul), tiene bajo poder de predicción sobre los datos reales de la muestra; afirmando lo ya encontrado por el R^2.

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

Figure 6: Distribución de Ingreso y Edad Predicha



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

Desarrollo: La edad óptima según la estimación (1) correspondiente al ingreso total, es de 55.87 años, lo que quiere decir que el individuo alcanza su ingreso máximo a esta edad óptima, a partir de la cual cada año adicional representa un ingreso menor. Esto refleja los rendimientos marginales de la edad con respecto al ingreso y está alineado a la teoría económica.

Al realizar el mismo cálculo con el ingreso transformado, se obtiene una edad óptima de 44.30 años, que tiene la misma interpretación, sin embargo es menor dada la menor dispersión en los datos.

Intervalos de Confianza:

Table 3: Intervalos de Confianza - Modelo (1) Ingreso Total

	2.5 %	97.5 %
(Intercept)	-1,272,791.000	657,094.000
Edad	30,538.910	127,695.900
Edad^2	-1,275.906	-140.178

Table 4: Intervalos de Confianza - Modelo (2) Ingreso Total Box-cox

	2.5 %	97.5 %
(Intercept)	15.427	16.585
Edad	0.066	0.124
$Edad^2$	-0.001	-0.001

Los intervalos de confianza indican dónde es probable que resida el parámetro de población. En otras palabras, es un rango de valores que describe la incertidumbre que rodea a una estimación. En este caso la Table 3 y 4 nos muestran el valor del ingreso con respecto a la variable independiente de edad y su transformación. Por ejemplo, con un 95% de confianza el ingreso total debe ubicarse entre $30,538 COP y $127,695 COP con respecto a la edad del individuo. Si bien el valor del salario en la transformación del ingreso box-cox no es directamente interpretable, se espera que la estimación se ubique entre 15.42 y 16.58 a un 95% de confianza.

4 *The Earnings GAP*

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

a) Estimate the unconditional earnings gap

$$log(Income) = \beta_1 + \beta_2 Female + u \tag{3}$$

Desarrollo: Estimación

Table 5: Unconditional Earnings Gap

	Dependent variable:
	Log_(Ingreso Total)
Female = 1	-0.141^{***}
	(0.042)
Constante	13.966^{***}
	(0.029)
Observations	1,551
R^2	0.007
Adjusted R^2	0.007
Residual Std. Error	0.828 (df = 1549)
F Statistic	11.164^{***} (df = 1; 1549)
Note:	*p<0.1; **p<0.05; ***p<0.01

b) How should we interpret the β_2 coefficient? How good is this model in sample fit?

Desarrollo: De acuerdo con los resultados de la regresión, el coeficiente de β_2 es negativo, lo que signfica que el hecho de ser mujer (female = 1) disminuye 14,1% el ingreso percibido, este resultado es signficativo con un pvalue menor al 1%. Sin embargo, el modelo especificado solo tiene en cuenta una de las variables que pueden afectar el ingreso, por lo que la bondad de ajuste del modelo es baja, con tan solo 0,7% de explicación de la variable de interés (ingeso total), teniendo en cuenta el R^2 y el R^2 ajustado.

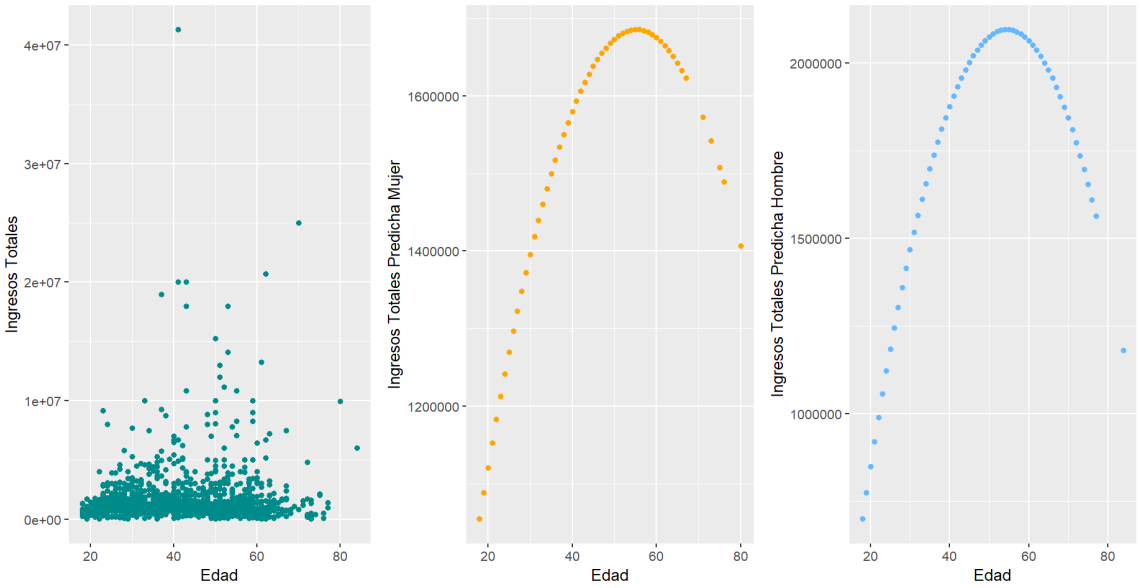
c) Estimate and plot the predicted age-earnings profile by gender. Do men and women in Bogotá have the same intercept and slopes?

Desarrollo: Si se comparan los interceptos y los coeficientes del modelo age earnings profile para cada género, se puede concluir que son diferentes (Tabla 6).Así mismo, de acuerdo con la Figura 7, se evidencia que el ingreso se incrementa a medida que la edad aumenta; no obstante, el salario empieza a crecer a un ritmo cada vez menor, llegando a un punto óptimo, después del cual empieza a decrecer (razón por la cual incluir la $edad^2$ es pertinente).

Table 6: Estimaciones del modelo edad-ingreso por sexo

	Dependent variable:	
	Ingreso Total	
	Female (1)	Male (2)
Edad	50,289.230	114,280.800***
	(33,936.650)	(36,097.800)
Edad^2	−455.185	−1,048.945**
	(408.879)	(413.287)
Constant	296,845.900	−1,016,314.000
	(658,129.100)	(726,859.800)
Observations	725	826
R^2	0.009	0.026
Adjusted R^2	0.006	0.024
Residual Std. Error	1,938,270.000 (df = 722)	2,427,762.000 (df = 823)
F Statistic	3.346** (df = 2; 722)	11.145*** (df = 2; 823)
Note:		*p<0.1; **p<0.05; ***p<0.01

Figure 7: Distribución de Ingreso y Edad Predicha por Sexo



En relación a los resultados, se evidencia que la pendiente del modelo condicionado a female = 1, es de $296,845 COP, mientras que para el coeficiente de la edad se puede observar que, un año adicional en una mujer, incrementa $50,289 COP el ingreso. Ahora bien, para edad^2 el coeficiente es de -$455. Sin embargo, ningún coeficiente es significativo estadísticamente y la bondad de ajuste del modelo es baja, con un R^2 y el R^2 ajustado menor al 1%.

Para el modelo condicionado a female = 0, se evidencia que el intercepto es de -$1,016,314 COP (No significativo estadísticamente). Por su parte, el coeficiente de la edad muestra que un año adicional en un hombre afecta positivamente el ingreso en $114,280 COP, con un p value menor al 1%. Por ultimo, la variable edad^2 afecta negativamente el ingreso en -$1,048 COP, significativo estdísticamente con un p value menor al 5%. La bondad de ajuste del modelo es baja, pero mayor que la estimación del modelo condicionado a female = 1, con un R^2 y el R^2 ajustado de 2,6% y 2,4% respectivamente.

d) 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?

Desarrollo: La edad óptima para las mujeres es de 55 años y para los hombres de 54 años, lo que indica que solo existe un año de diferencia entre la edad optima, en la cual se máximizan los ingresos de mujeres y hombres.

Intervalos de Confianza:

Table 7: Intervalos de Confianza - Female

	2.5 %	97.5 %
(Intercept)	-995,229.300	1,588,921.000
Edad	-16,337.070	116,915.500
Edad^2	-1,257.918	347.549

Table 8: Intervalos de Confianza - Male

	2.5 %	97.5 %
(Intercept)	-2,443,031.000	410,403.200
Edad	43,426.240	185,135.400
Edad^2	-1,860.165	-237.724

Al usar la metodología de Bootstrap para hallar los intervalos de confianza, se puede concluir para el modelo condicionado a female = 1, que existe una alta probabilidad (nivel de confianza de 95%) de que el parámetro poblacional de la edad respecto al ingreso, se encuentre entre un rango de -$16,337 COP y $116,916 COP, mientras que, el intervalo para la variable edad^2 se ubica entre -$1258 COP y $348 COP.

Con respecto al modelo condicionado a female = 0, se puede evidenciar que entre $43,426 COP y $185,135 COP, se encuentra el valor real del parámetro de la edad en relación al ingreso, con 95% de certeza, respecto a la variable edad^2 el rango se situa entre -$1,860 COP y -$238 COP.

5 *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).

a) Estimate the conditional earnings gap $log(Income) = \beta_1 + \beta_2 Female + \theta_X + u$

Table 9: Conditional Earnings Gap

	Dependent variable:
	log_income
Female = 1	-0.178^{***}
	(0.033)
Formal = 1	0.497^{***}
	(0.036)
Edad	0.053^{***}
	(0.007)
Edad^2	-0.001^{***}
	(0.0001)
Estrato 2	0.081
	(0.054)
Estrato 3	0.194^{***}
	(0.057)
Estrato 4	0.771^{***}
	(0.090)
Estrato 5	1.140^{***}
	(0.139)
Estrato 6	1.754^{***}
	(0.183)
Max. Educ. Level = Primary Incomplete	0.614^{***}
	(0.199)
Max. Educ. Level = Primary Complete	0.615^{***}
	(0.193)
Max. Educ. Level = Secundary Incomplete	0.605^{***}
	(0.191)
Max. Educ. Level = Secundary Complete	0.747^{***}
	(0.188)
Max. Educ. Level = Terciary	1.075^{***}
	(0.190)
	0.001^{***}
	(0.0002)
Constant	11.555^{***}
	(0.235)
Observations	1,551
R^2	0.412
Adjusted R^2	0.406
Residual Std. Error	0.640 (df = 1535)
F Statistic	71.728^{***} (df = 15; 1535)
Note:	*p<0.1; **p<0.05; ***p<0.01

El modelo tiene una bondad de ajuste relativamente alta, teniendo en cuenta los modelos analizados hasta el momento, con un R^2 de 41.2% y un R^2 ajustado de 40.6%, lo cual quiere decir que las variables incluidas explican el ingreso en un porcentaje alredor del 40%. En la estimación (log-lin) todos los coeficientes son significativos estadísticamente, y se puede evidenciar lo siguiente: (1) β_1 el intercepto del modelo es de \$11,555 COP, (2) β_2 para las mujeres (female = 1), el ingreso disminuye 17,8%, (3) β_3 para las personas empleadas formalmente, el ingreso aumenta 49,7%, (4) β_4 en relación a la edad, un año más aumenta el ingreso 5,3%, (5) β_5 con respecto a la edad^2, se puede decir que la relación con el ingreso es negativa en 0,1%, (6) β_6 - β_{11} con respecto al estrato, se evidencia que a medida que aumenta el estrato, se incrementa el ingreso, (7) β_{12} - β_{16} en cuanto a el nivel de educación, se evidencia que a medida que se incrementan los años de educación, aumenta el ingreso, y (8) β_{17} a medida que aumenta el tiempo de permanencia en un trabajo/negocio/industria, el ingreso se incrementa en 1%.

b) Use FWL to repeat the above estimation, where the interest lies on β_2. Do you obtain the same estimates?

Table 11: Estimación modelo original y modelo teorema FWL

	Dependent variable:	
	log_income (Logaritmo del Ingreso Total)	res_y (Residuos Ingreso Total)
	(1)	(2)
Female= 1	-0.178^{***}	
	(0.033)	
Formal = 1	0.497^{***}	
	(0.036)	
Edad	0.053^{***}	
	(0.007)	
Edad^2	-0.001^{***}	
	(0.0001)	
Estrato 2	0.081	
	(0.054)	
Estrato 3	0.194^{***}	
	(0.057)	
Estrato 4	0.771^{***}	
	(0.090)	
Estrato 5	1.140^{***}	
	(0.139)	
Estrato 6	1.754^{***}	
	(0.183)	
Max. Educ. Level = Primary Incomplete	0.614^{***}	
	(0.199)	
Max. Educ. Level = Primary Complete	0.615^{***}	
	(0.193)	
Max. Educ. Level = Secundary Incomplete	0.605^{***}	
	(0.191)	
Max. Educ. Level = Secundary Complete	0.747^{***}	
	(0.188)	
Max. Educ. Level = Terciary	1.075^{***}	
	(0.190)	
Tiempo trabajado en empresa actual (meses)	0.001^{***}	
	(0.0002)	
res_f (Residuos Female)		-0.178^{***}
		(0.033)
Constant	11.555^{***}	0.000
	(0.235)	(0.016)
Observations	1,551	1,551
R^2	0.412	0.019
Adjusted R^2	0.406	0.018
Residual Std. Error	0.640 (df = 1535)	0.637 (df = 1549)
F Statistic	71.728^{***} (df = 15; 1535)	29.671^{***} (df = 1; 1549)
Note:		*p<0.1; **p<0.05; ***p<0.01

Se obtiene el mismo coeficiente de β_2, a través del teorema de FWL.

(c) How should we interpret the β_2 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"?

β_2 cuando female es igual a 1, se reduce el ingreso en 17,8%, es decir, el hecho de ser mujer afecta de manera negativa el ingreso (Modelo log - lin).

El modelo estimado con la muestra originalmente se encuentra alrededor de 40%, mientras que utilizando el teorema FWL tiene una bondad de ajuste baja de aproximadamente 2%.

Dado que los datos son obtenidos a través de una encuesta, es posible que exista sesgo de selección, en el cual tanto hombres como mujeres pueden dar valores menores o mayores de sus salarios reales. Ahora bien, concluir que es un problema de discriminación es complejo, ya que cada persona empleada tiene un salario distinto en función del capital humano acumulado que posee y los tipos de oficio que esté dispuesta a realizar. Sin embargo, manteniendo las otras variables constantes (formal, edad, estrato y nivel educativo, tiempo de experiencia en trabajo actual), bajo este modelo es posible afirmar que existe una relación inversa entre el ingreso y ser mujer.

6 *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.

(a) 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.)

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

Desarrollo: Cabe mencionar que su coeficiente es igual a la media de la variable dependiente y que se estimó con la muestra train.

Table 12: Modelos Benchmark

	Dependent variable:		
	Ingreso Total	Ingreso Total Box-cox	Logaritmo Ingreso Total
	(1)	(2)	(3)
Constant	1,520,399.000***	17.878***	13.889***
	(56,586.740)	(0.040)	(0.025)
Observations	1,085	1,085	1,085
R^2	0.000	0.000	0.000
Adjusted R^2	0.000	0.000	0.000
Residual Std. Error (df = 1084)	1,863,930.000	1.306	0.807
Note:			*p<0.1; **p<0.05; ***p<0.01

Estimate again your previous models

Table 13: Modelos Estimados Anteriormente - Muestra Entrenamiento

	Dependent variable:		
	Ingreso Total	Logaritmo Ingreso Total	
	(1)	(2)	(3)
Edad	113,554.400***		0.062***
	(25,318.240)		(0.009)
Edad^2	−1,180.240***		−0.001***
	(295.968)		(0.0001)
Estrato 2			0.098
			(0.064)
Estrato 3			0.186***
			(0.068)
Estrato 4			0.802***
			(0.106)
Estrato 5			0.980***
			(0.166)
Estrato 6			1.459***
			(0.236)
Max. Educ. Level = Primary Incomplete			0.336
			(0.230)
Max. Educ. Level = Primary Complete			0.397*
			(0.223)
Max. Educ. Level = Secundary Incomplete			0.337
			(0.219)
Max. Educ. Level = Secundary Complete			0.492**
			(0.216)
Max. Educ. Level = Terciary			0.798***
			(0.217)
Tiempo trabajado en empresa actual (meses)			0.001***
			(0.0002)
Female = 1		−0.145***	−0.190***
		(0.049)	(0.039)
Formal = 1			0.473***
			(0.042)
Constant	−915,575.700*	13.956***	11.692***
	(503,479.800)	(0.033)	(0.279)
Observations	1,085	1,085	1,085
R^2	0.024	0.008	0.391
Adjusted R^2	0.023	0.007	0.382
Residual Std. Error	1,842,786.000 (df = 1082)	0.805 (df = 1083)	0.635 (df = 1069)

Note: *p<0.1; **p<0.05; ***p<0.01

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.

Desarrollo:

Table 14: Nuevos Modelos Propuestos - Muestra de Entrenamiento

	Dependent variable:				
	Ingreso Total - Boxcox				
	(1)	(2)	(3)	(4)	(5)
Máx Educ. Level (Años)	0.153*	0.169*	0.136*	0.128*	−0.128*
	(0.012)	(0.012)	(0.012)	(0.012)	(0.037)
Experiencia Potencial (años)	0.051*	0.046*	0.048*	0.051*	−0.128*
	(0.007)	(0.007)	(0.007)	(0.007)	(0.027)
Experiencia Potencial^2(años)	−0.001*	−0.001*	−0.001*	−0.001*	−0.001*
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Horas semanales trabajadas - Ocu. Principal		0.023*	0.018*	0.018*	0.019*
		(0.002)	(0.002)	(0.002)	(0.002)
Segundo Trabajo = 1		0.698*	0.740*	0.733*	0.633*
		(0.209)	(0.196)	(0.195)	(0.189)
Female = 1			−0.232*	−0.213*	−0.284*
			(0.066)	(0.066)	(0.078)
Trabajador Cuentra Propia = 1			−0.251*	−0.065	−0.138
			(0.084)	(0.107)	(0.103)
Informal = 1			−0.714*	−0.519*	−0.416*
			(0.079)	(0.094)	(0.092)
Microempresa = 1				−0.552*	−0.451*
				(0.133)	(0.129)
Tamaño de la firma (2 a 5 wks)				0.164	0.101
				(0.119)	(0.116)
Tamaño de la firma (6 a 10 wks)				−0.154	−0.093
				(0.139)	(0.135)
Tamaño de la firma (11 a 50 wks)				−0.202	−0.196
				(0.108)	(0.104)
Edad					0.185*
					(0.027)
Hombre*Educación terciaria					−0.498*
					(0.092)
Mujer*Educación terciaria					−0.362*
					(0.100)
Constant	15.602*	14.291*	15.359*	15.515*	15.134*
	(0.189)	(0.224)	(0.228)	(0.232)	(0.242)
Observations	1,085	1,085	1,085	1,085	1,085
R^2	0.168	0.243	0.341	0.353	0.398
Adjusted R^2	0.166	0.239	0.336	0.346	0.390
Residual Std. Error	1.193 (df = 1081)	1.139 (df = 1079)	1.064 (df = 1076)	1.057 (df = 1072)	1.020 (df = 1069)
F Statistic	72.877* (df = 3; 1081)	69.115* (df = 5; 1079)	69.630* (df = 8; 1076)	48.736* (df = 12; 1072)	47.186* (df = 15; 1069)
Note:					p<0.1; p<0.05; **p<0.01

Para los nuevos modelos propuestos se utilizó una variable para experiencia potencial, ya que es complicado encontrar una variable que muestre cuántos años ha trabajado una persona en realidad. Por ello, en la literatura se ha utilizado como proxy de la experiencia la experiencia potencial. Esta nace de restarle a la edad de la persona los años que ha estudiado y, además, restarle cinco (5) años pues en sus años de primera infancia ni estudió ni trabajó.

Para esto se usó la variable p6210 (¿Cuál es el nivel educativo más alto alcanzado por....y el último año o grado) y se transformó dandole a cada categoria unos años de estudio para

poder calcular la proxy de experiencia potencial. Todos los modelos fueron estimados con la muestra Train y teniendo como variable dependiente la transformación box cox del ingreso total.

Además, para crear la variable Máx. Educ Level (Años) se supuso que Primary Incomplete son 4 años de estudio, Primary Complete son 5, Secundary Incomplete son 10, Secundary Complete son 11 y Terciary son 15.

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

Desarrollo: A continuación una tabla que comparara el MSE de los cinco modelos estimados. Este cálculo se obtuvo de hacer la predicción de los modelos sobre la muestra Train y la muestra Test.

Table 15: MSE - Nuevos Modelos

	Train	Test
Modelo 1	1.418	1.613
Modelo 2	1.291	1.616
Modelo 3	1.123	1.407
Modelo 4	1.103	1.395
Modelo 5	1.026	1.303

El modelo con el error de predicción menor es el modelo 5, el modelo más complejo propuesto. Cuando se comparan los resultados en cada muestra podemos ver que siguen la misma tendencia. Al complejizar más el modelo se vería un cambio en la tendencia del MSE obtenido de la muestra Test.

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?

Se estima el modelo seleccionado (No. 5) con el data set de prueba, y se evidencia que las siguientes variables: experiencia potencional, experiencia potencial^2, dicótoma que indica si es mujer, indicadora de trabajo por cuenta propia, dicótoma de informalidad, indicadores de trabajo en microecmpresa y la interacción hombre e ir a la universidad, tienen coeficientes negativos y significativos estadísticamente. Por otra parte, las variables con coeficientes positivos y significativos estadísticamente son horas de trabajo a la semana, el tamaño de firma 2 y la edad (Tabla 16). Es importante señalar que estos resultados son similares a los obtenidos en el modelo incial con el data set de entrenamiento, eso es una buena señal sobre la consistencia del modelo. Además, el modelo tiene una bondad de ajuste de 42% (R^2) y ajustada por grados de libertad 40% (R^2ajustado).

Table 16: Modelo 5 con Muestra de Prueba

	Dependent variable:
	Ingreso Total Box-cox
Máx Educ. Level (Años)	−0.061
	(0.059)
Experiencia Potencial (años)	−0.143***
	(0.043)
Experiencia Potencial^2(años)	−0.0003
	(0.0002)
Horas semanales trabajadas - Ocu. Principal	0.012***
	(0.004)
Segundo Trabajo = 1	−0.236
	(0.318)
Female = 1	−0.285**
	(0.129)
Trabajador Cuentra Propia = 1	−0.381**
	(0.177)
Informal = 1	−0.508***
	(0.154)
Microempresa = 1	−0.490**
	(0.214)
Tamaño de la firma (2 a 5 wks)	0.353*
	(0.208)
Tamaño de la firma (6 a 10 wks)	−0.046
	(0.204)
Tamaño de la firma (11 a 50 wks)	−0.276
	(0.169)
Edad	0.188***
	(0.041)
Hombre*Educación terciaria	−0.504***
	(0.156)
Mujer*Educación terciaria	−0.283
	(0.172)
Constant	14.576***
	(0.420)
Observations	466
R^2	0.423
Adjusted R^2	0.404
Residual Std. Error	1.106 (df = 450)
F Statistic	21.975*** (df = 15; 450)
Note:	*p<0.1; **p<0.05; ***p<0.01

Se calcularon los residuales estandarizados (uj) y el leverage statistic (hj) de la estimación del modelo seleccionado (No. 5). Posterior a ello, se hallan los alphas (α) = uj/(1-hj) para cada observación en el data set de test.

Luego, se define una variable nueva, denominada cuadrante, en la cual se clasifican los alphas teniendo en cuenta dos condiciones:

1). Las observaciones con valores de uj por debajo del percentil p25 (uj bajos) y con valores de hj por debajo de p25 (hj bajos) ó

2). Las observaciones con valores de uj por encima del percentil p75 (uj altos) y con valores de hj por debajo de p75 (hj altos).

Dentro de la nueva variable cuadrante se encuentran 59 observaciones de las 466 que componen la muestra, es decir, el 12.6%.

Table 17: Head table of leverage statistic for each observation

	uj_1	hj_1	alphas	cuadrante
1	0.56	0.05	0.64	1
2	-0.65	0.01	-0.72	1
3	-1.46	0.06	-1.66	0
4	-0.43	0.03	-0.48	0
5	-1.06	0.03	-1.19	0
6	0.91	0.05	1.03	1

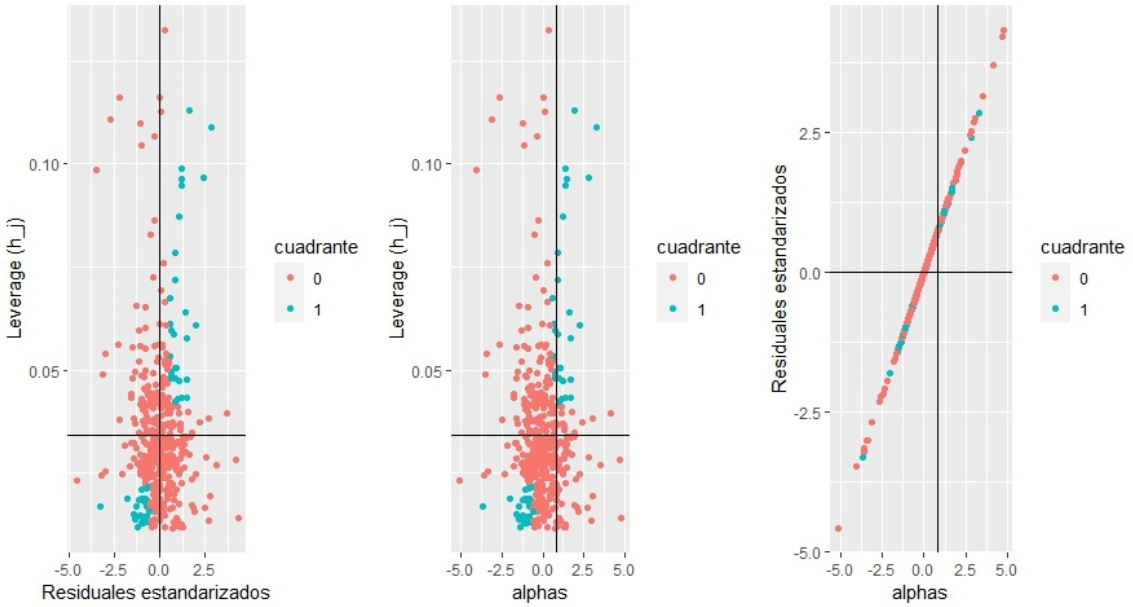
Table 18: Tail table of leverage statistic for each observation

	uj_1	hj_1	alphas	cuadrante
461	-0.25	0.04	-0.28	0
462	-0.53	0.01	-0.59	0
463	-0.99	0.02	-1.11	1
464	-0.24	0.05	-0.27	0
465	1.09	0.05	1.23	1
466	0.71	0.02	0.80	0

Graficamente se pueden evidenciar si existen outliers con un alto leverage (color azul) en la Figura 8. Por lo anterior, el gobierno debería fijarse en los individuos que cuentan con un alto leverage y un alto valor en el residual del modelo (p75), así como en aquellos con un bajo leverage y un bajo valor en el residual del modelo (p25); en especial en aquellos individuos que declaran ingresos por debajo de lo que predice el modelo (error de predicción), relacionado con las coordenadas de y, y que a su vez impactan los resultados por debajo de lo que deberían (leverage statistic), relacionado con las coordenadas x.

(b) Repeat the previous point but use K-fold cross-validation. Comment on similarities/differences of using this approach.

Figure 8: Outliers



Desarrollo: Para la validación cruzada utilizando la técnica de k-fold, se asumió que su parametro k sería igual a cinco. Además se estió un modelo 6 que tuviera una especificación muy compleja y que no tuviera sentido ecnómico para validar que el procedimiento se hizo adecuadamente.

La especificación del modelo 6 fue:

$$\begin{aligned} Ingtotboxcox = maxEducLevel + exppotp6210 + exppotp6210^2 + \\ poly(hoursWorkUsual, 9) + dostrabajo + female + cuentaPropia + \\ informal * age + microEmpresa + sizeFirm + relab + poly(age, 8) + \\ female * cuentaPropia + college * female \end{aligned} \tag{4}$$

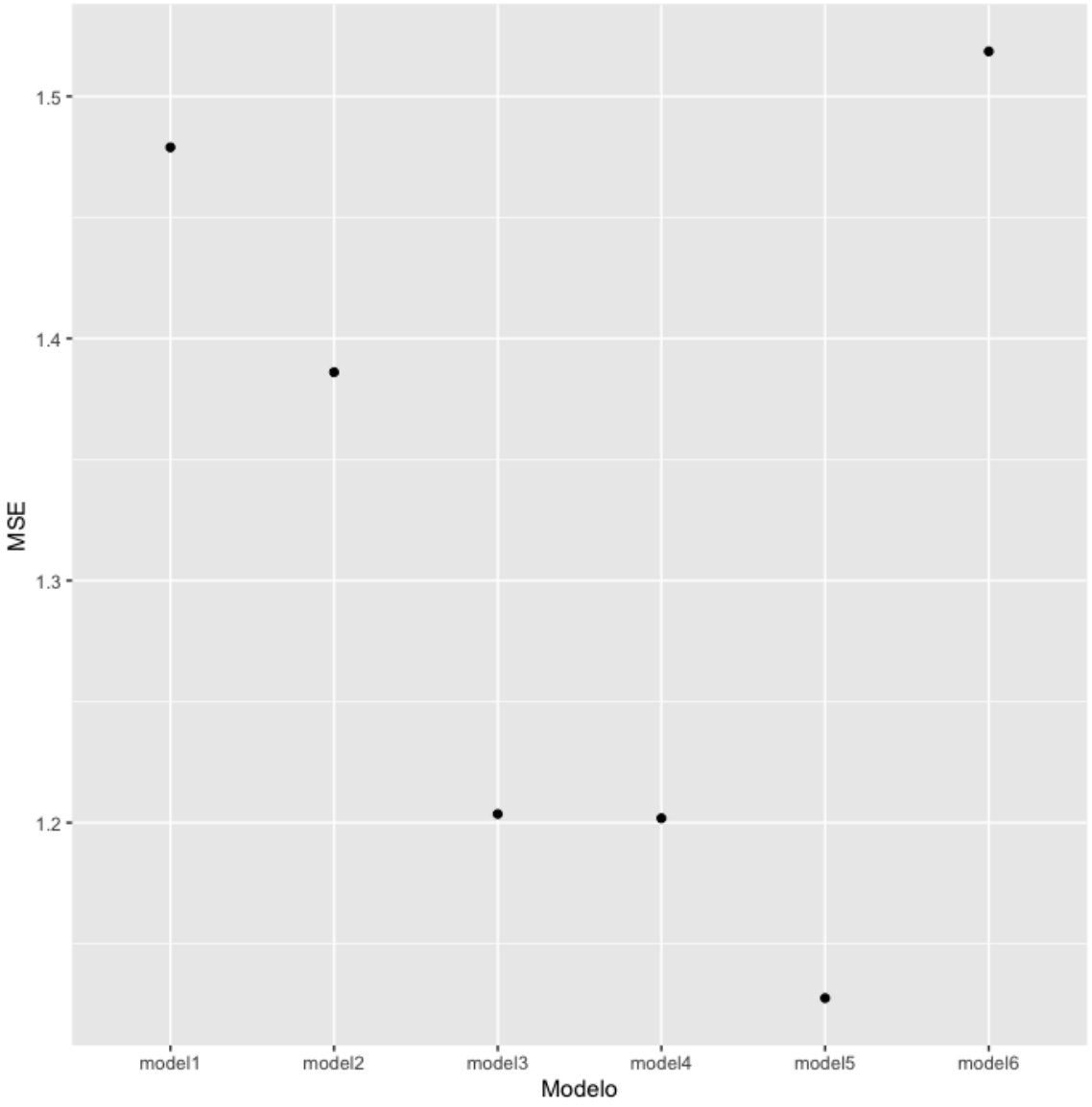
A continuación una tabla que resume los principales estadísticos del proceso para cada modelo.

Table 19: Principales estadísticos del K-fold cross-validation

	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD	MSE
Modelo 1	1.22	0.18	0.86	0.04	0.02	0.03	1.48
Modelo 2	1.18	0.23	0.85	0.07	0.04	0.04	1.39
Modelo 3	1.10	0.33	0.81	0.09	0.04	0.06	1.20
Modelo 4	1.10	0.34	0.81	0.03	0.05	0.03	1.20
Modelo 5	1.06	0.38	0.77	0.08	0.07	0.05	1.13
Modelo 6	1.23	0.32	0.77	0.26	0.11	0.04	1.52

El siguiente gráfico compara los seis modelos estimados y sus errores de predicción resultantes de haber utilizado la técnica de k-fold cross validation.

Figure 9: K-Fold Cross Validation



Como se puede observar, haber utilizado esta técnica redujo en gran medida los errores de predición de los modelos. Además, es evidente que para modelos mal especificados, aunque sean muy complejos, no pueden ser tomados en cuenta por su gran error de predicción.

(c) 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.

Desarrollo:
Para el LOOCV se utilizó el modelo 5 pues era el que tenía un menor error de predicción. Como resultado de este proceso para la muestra test y se obtuvo para este modelo un MSE promedio igual a 1.279341

ii. Compare the results to those obtained in the computation of the leverage statistic
De acuerdo al leverage statistic para el modelo 5 (hj) y la validación hecha LOOCV se encuentra que las estadisticas calculadas arrojan los mismos resultados.

	LOOCV	Leverage
cv	1.28	1.28

7 *Referencias*

Su, B., Heshmati, A. (2013). Analysis of the Determinants of Income and Income Gap between Urban and Rural China. *SSRN Electronic Journal*. doi:10.2139/ssrn.2210822

Wodon, Q. T. (1999). Microdeterminants of Consumption, Poverty, Growth, and Inequality in Bangladesh. *Policy Research Working Papers*. doi:10.1596/1813-9450-2076