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

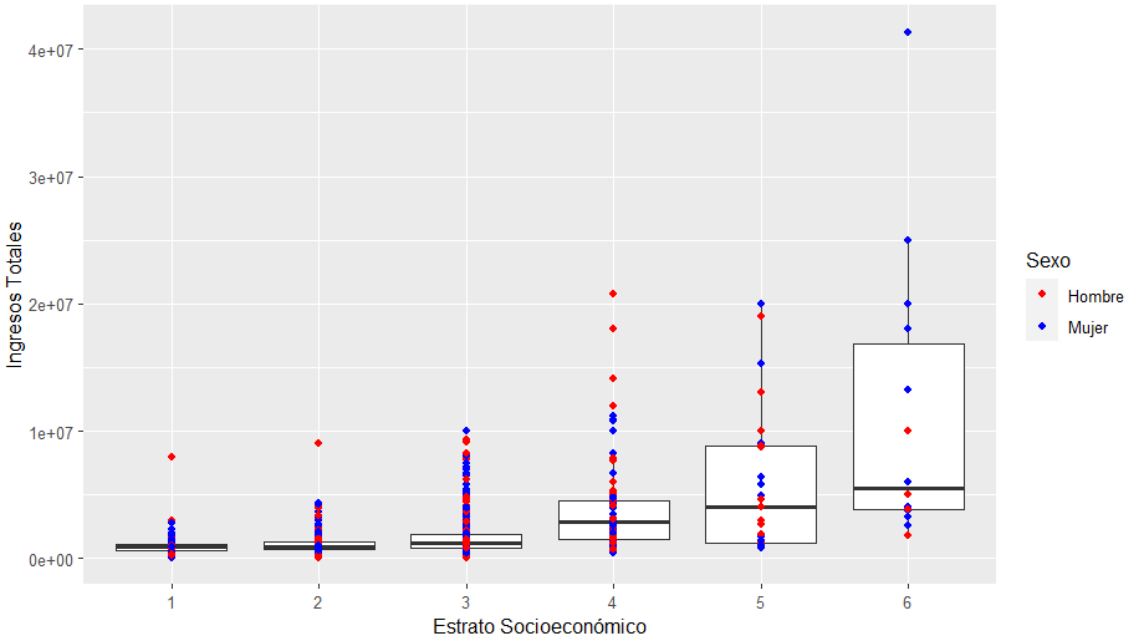
Statistic	Ingresos Totales	Edad	Meses Experiencia Laboral
N	1,551	1,551	1,551
Mean	1,587,665	40	69
St. Dev.	2,233,349	14	95
Min	20,000	18	0
Median	1,003,357	38	30
Max	41,333,333	84	600

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:

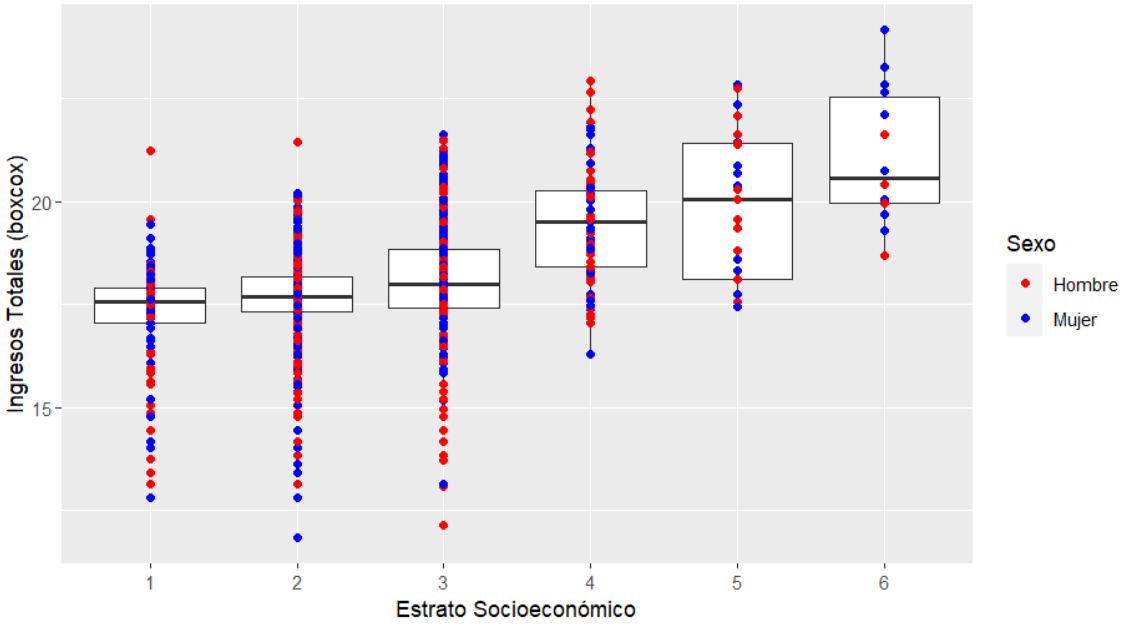
Figure 1: Cajas y Bigotes - Ingreso vs. Estrato por Sexo



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

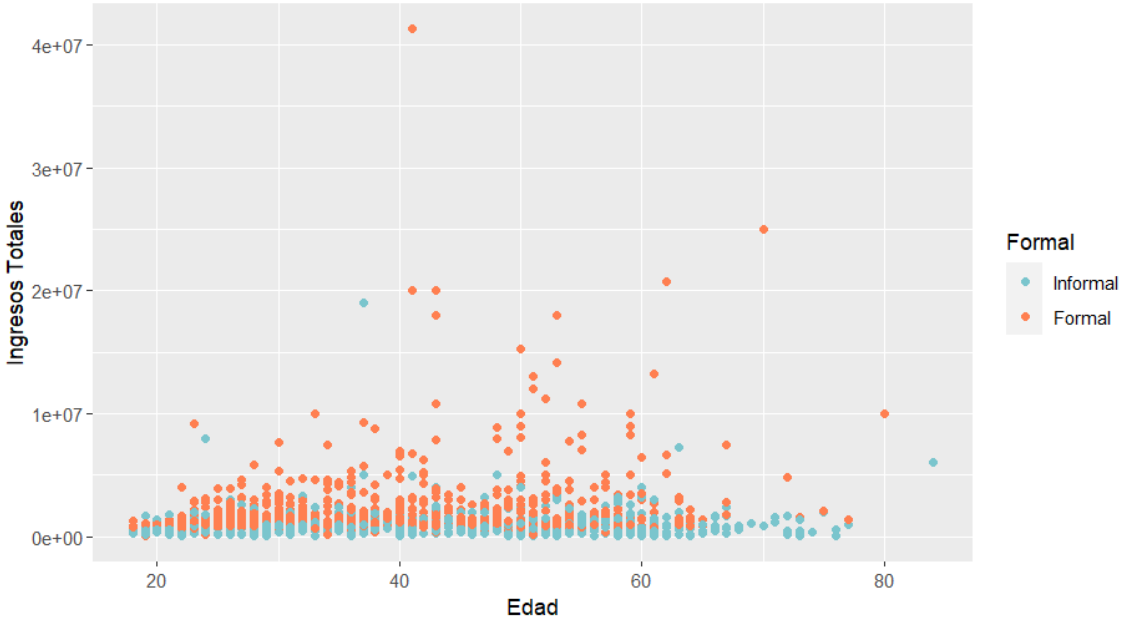
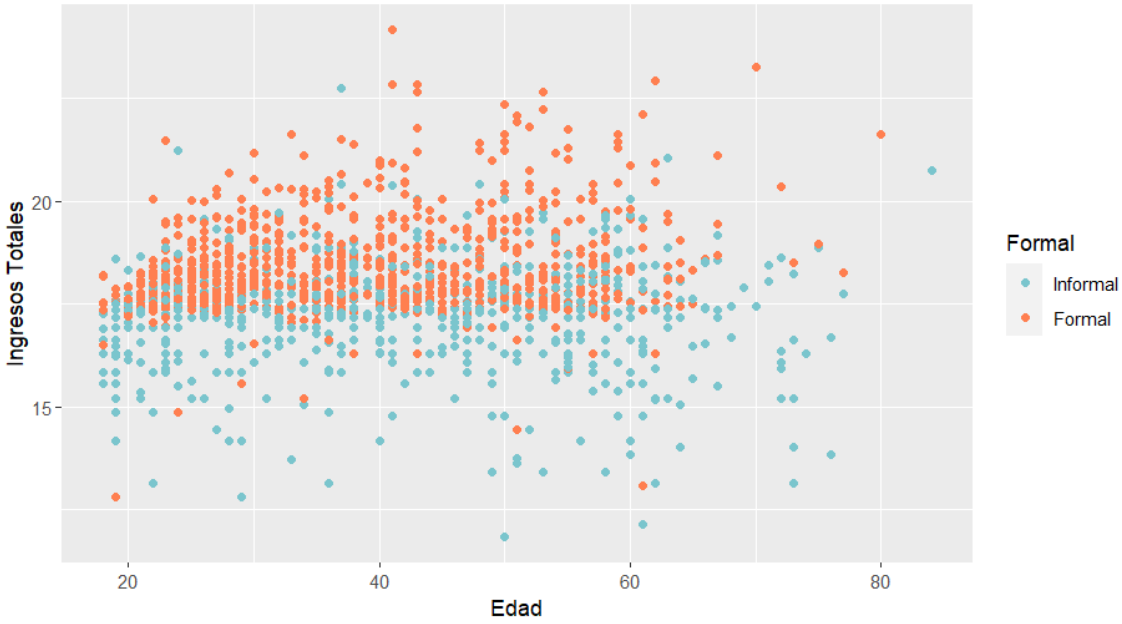


Figure 4: Dispersión - Ingreso (boxcox) vs. Edad y Formalidad



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?

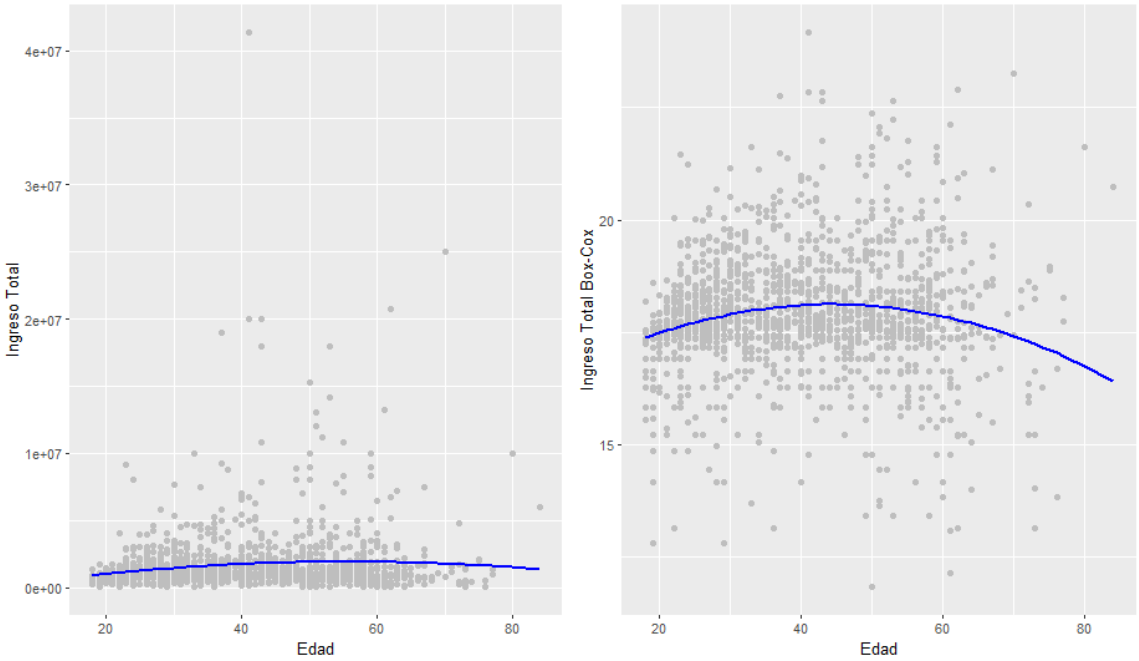
Desarrollo: La importancia de conocer si un modelo tiene buen ajuste a los datos, radica en proporcionar buenas respuestas a la pregunta de investigación. En este caso en particular, el modelo de regresión lineal estimado por *OLS*, teóricamente minimiza la distancia entre la línea ajustada y todos los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

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

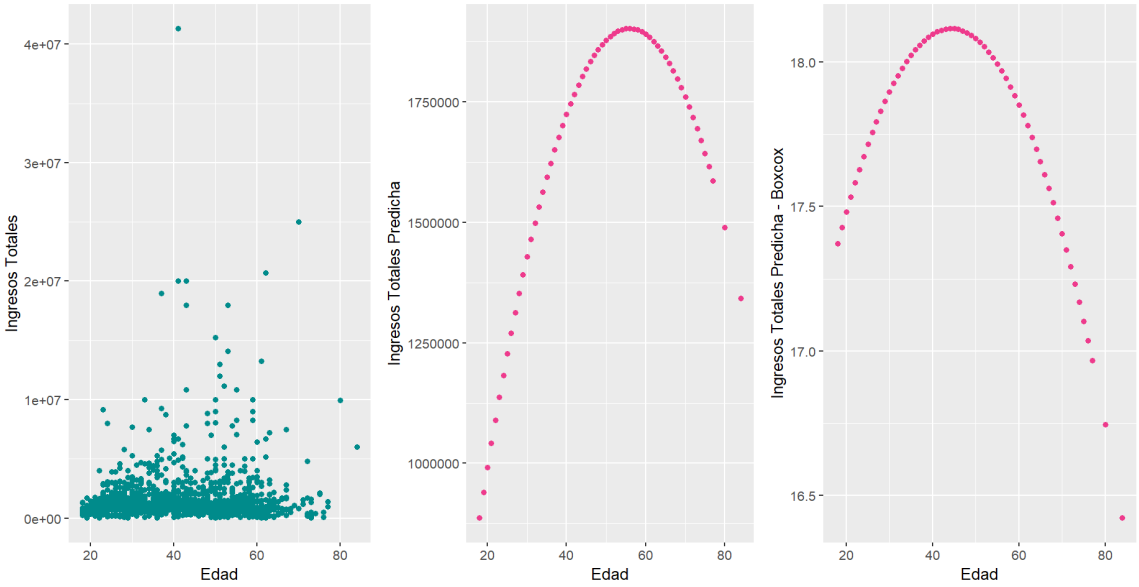
De la misma manera encotramos que el ajuste del modelo no es bueno por la gráfica de las predicciones comparada con la variable de ingreso total para ambas estimaciones. La Figure 5 nos muestra que la línea de predicción (línea azul), tiene bajo poder de predicción sobre los datos reales de la muestra;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

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



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% el ingreso total debe ubicarse entre $30,538 COP y $127,695 COP con respecto a la edad del individuo.

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

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

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

Desarrollo: La edad óptima para Mujer es de XX y para Hombre de XX lo que indica que...

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?

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).

Table 5: Unconditional Earnings Gap

	Dependent variable:
	Log_(Ingreso)
Female	-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

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

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

(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"?

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.)

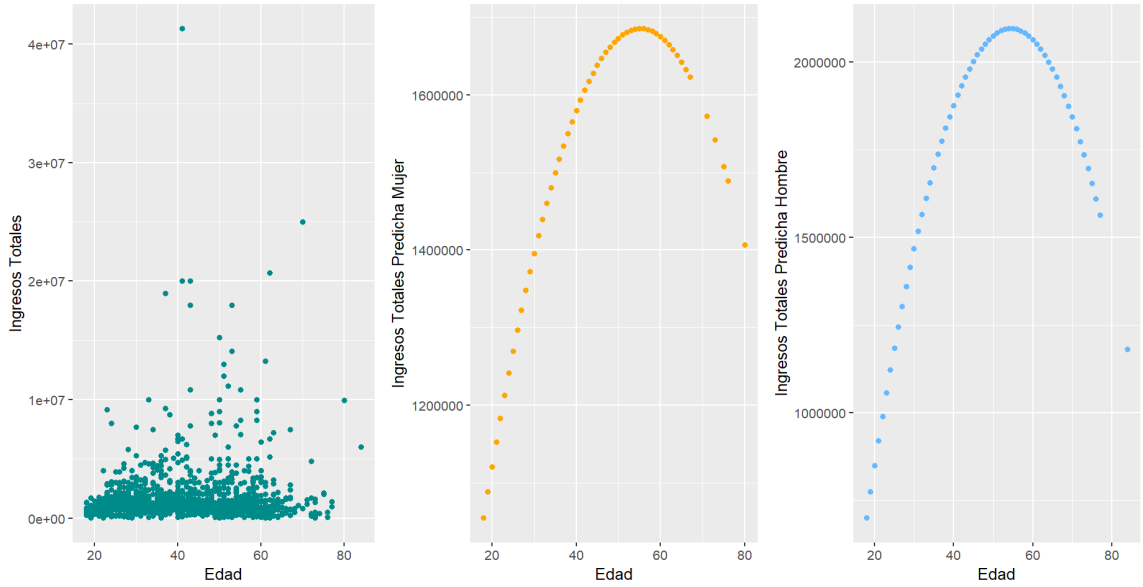
i. Estimate a model that only includes a constant. This will be the benchmark.

ii. Estimate again your previous models

iii. 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.

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

Figure 7: Distribución de Ingreso y Edad Predicha por Sexo



v. 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?

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

(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.

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