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: Justificación Ecuación de Mincer

(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 filtró la base...
- Se validaron
- Se identificó

(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

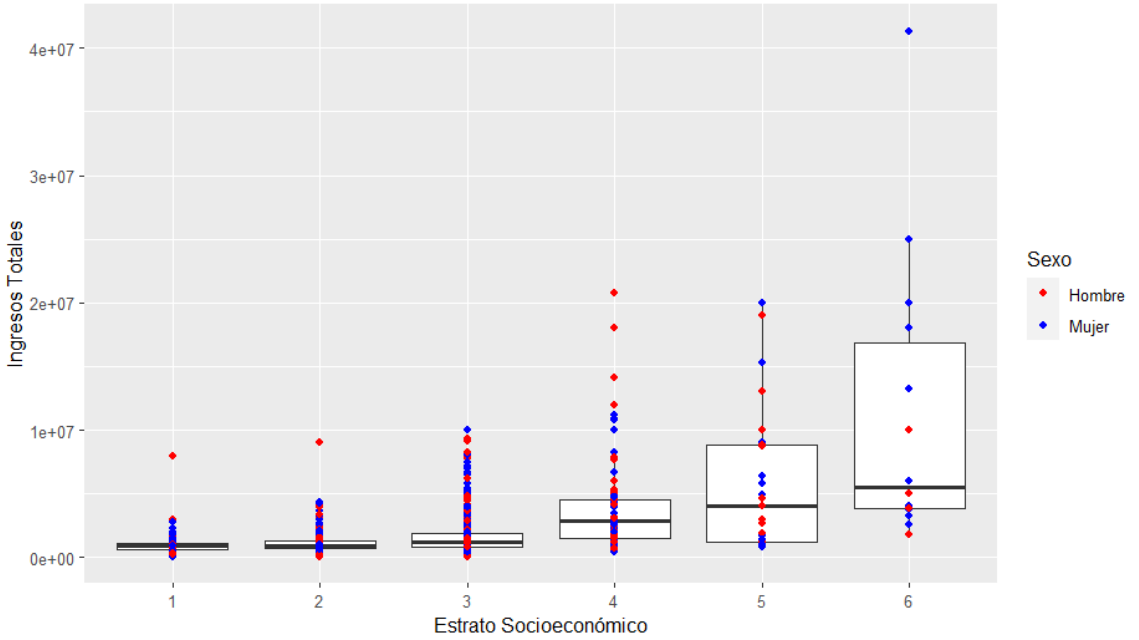
Statistic	Ingresos Totales	Edad	Meses Experiencia Laboral
N	1,551	1,551	1,551
Mean	1,587,665	40	69
St. Dev.	2,233,349	14	95
Min	20,000	18	0
Pctl(25)	800,000	28	8
Median	1,003,357	38	30
Pctl(75)	1,569,226	50	84
Max	41,333,333	84	600

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.

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. 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.

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:

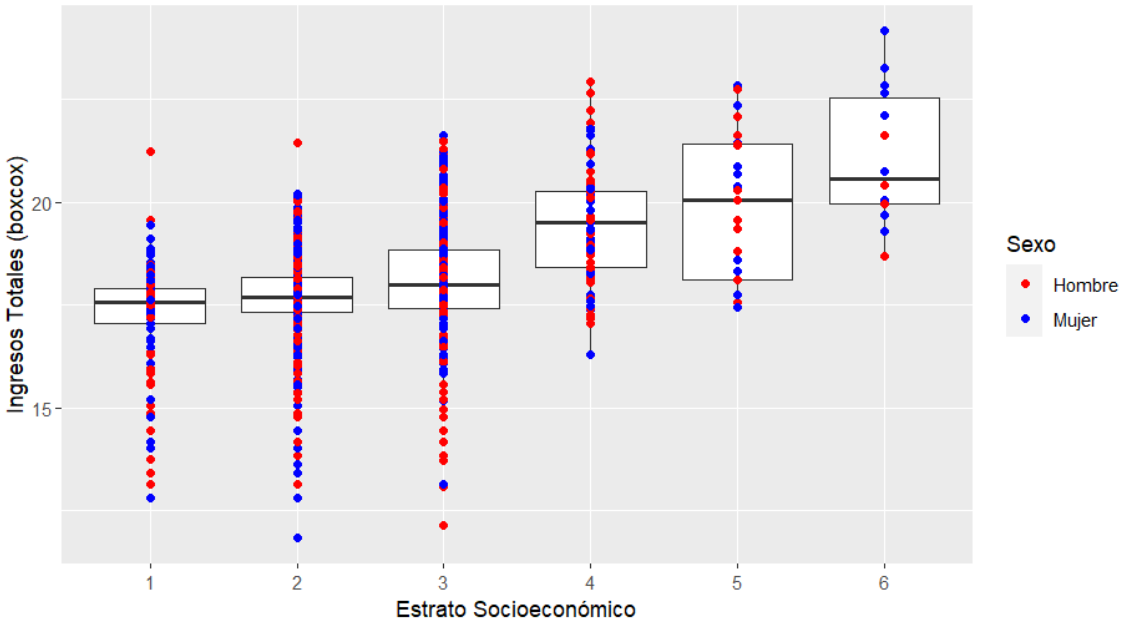
Figure 1: Cajas y Bigotes - Ingreso vs. Estrato por Sexo



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. En este punto el estrato 6 tiene 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 1, se identifican valores atípicos y por tanto procedemos a realizar una transformación boc-cox del ingreso total, para que nos permita observar la gráfica de manera más clara. Se obtiene:

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



El ajuste de la distribución 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.

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.

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.

Figure 3: Dispersión - Ingreso vs. Edad y Formalidad

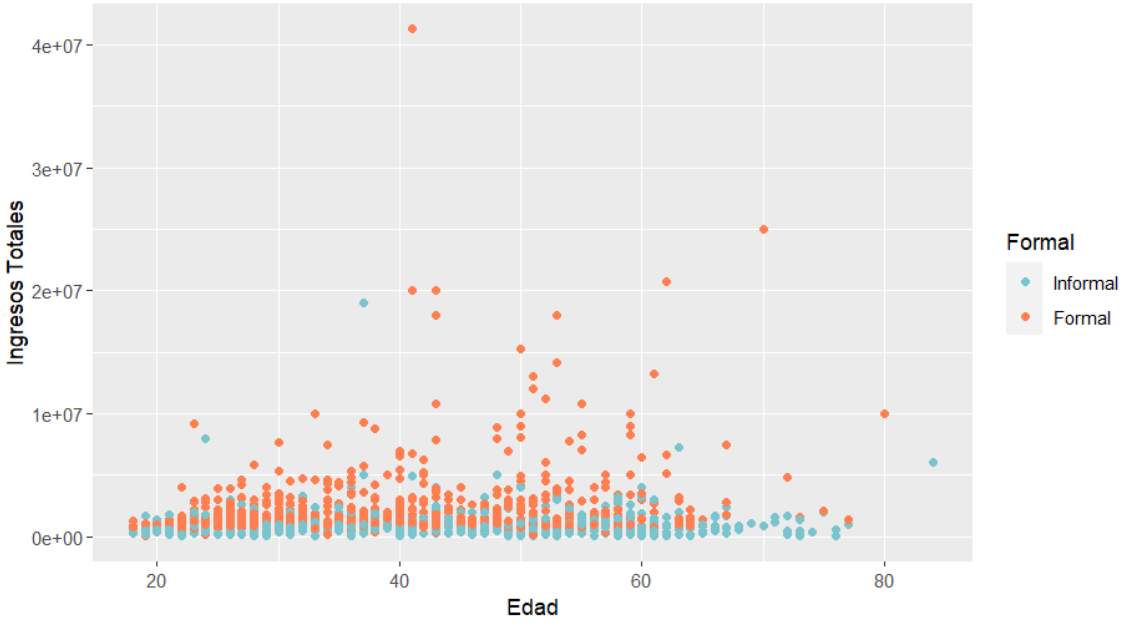
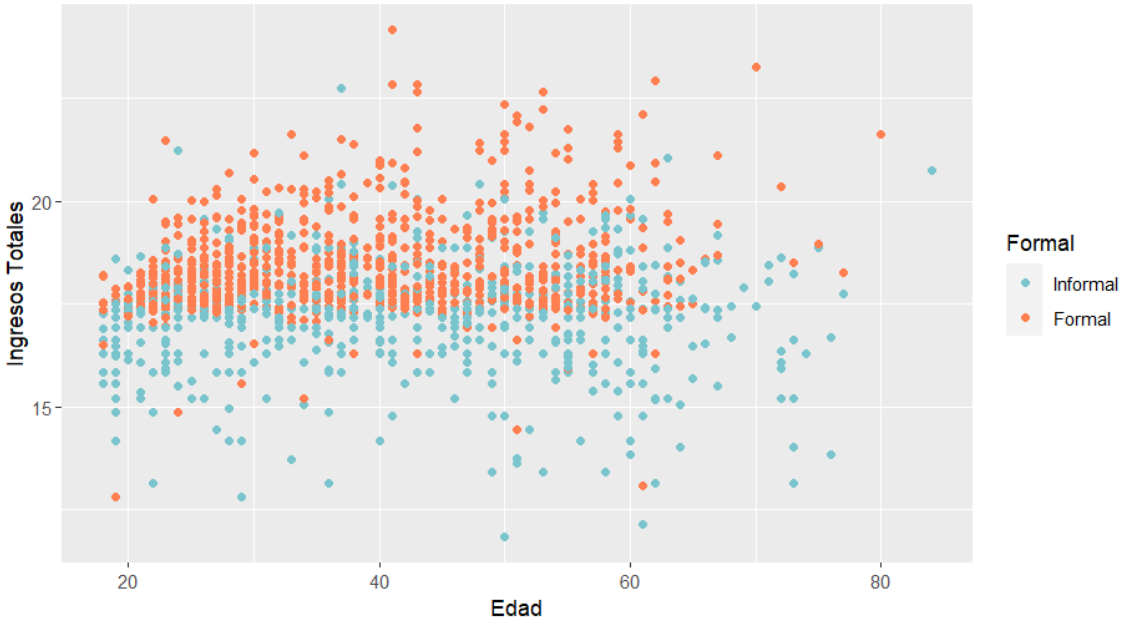


Figure 4: Dispersión - Ingreso (boxcox) vs. Edad y Formalidad



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: Ingresos totales = Suma de Ingreso + Ingreso imputado

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

$$Income = \beta_1 + \beta_2 Age + \beta_3 Age^2 + u \tag{1}$$

Table 2:

	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: Según el R2 del modelo el fit no es tan bueno.