Problem Set 1: Predicting Income

Paula Ramos, Karen Uribe-Chaves
Juan D. Urquijo

June 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...

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

Desarrollo: A partir del uso del código...

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: Estadísticas Descriptivas

Table 1: Estadisticas Descriptivas

Statistic	N	Mean	St. Dev.	Min	Max
Ingresos Totales	1,551	1,575,344.000	2,236,424.000	1,564,583.000	41,333,333.000
Experiencia	1,551	68.937	95.086	84	600
Edad	1,551	39.703	13.532	18	84

3 *Age-earnings profile*

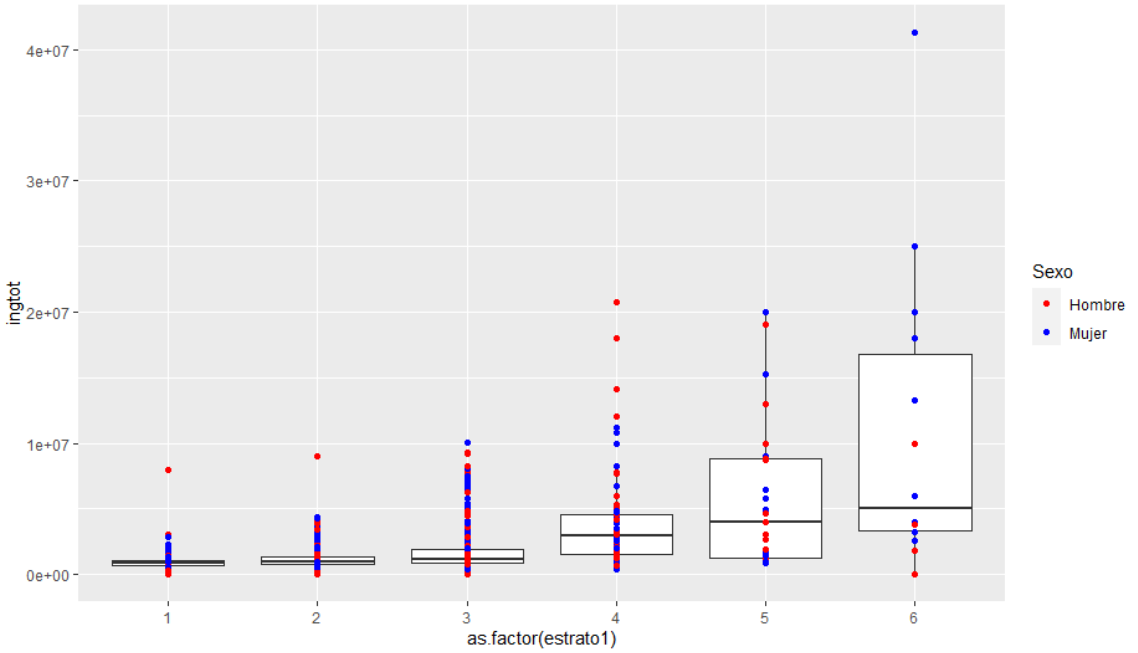
A great deal of evidence in Labor economics suggests that the typical worker's age-earnings profile has a predictable path: Wages tend to be low when the worker is young; they rise as the worker ages, peaking at about age 50; and the wage rate tends to remain stable or decline slightly after age 50.

(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: Regresión y correlaciones.

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

Figure 1: Ingresos vs. Estrato y Sexo



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

How good is this model in sample fit?

Desarrollo: Pruebas y gráficas