Problem Set 1: Predicting Income

Big Data y Machine Learning para Economía Aplicada

Gustavo Adolfo Castillo Álvarez (201812166),
Alexander Almeida Ramírez (202225165),
Jorge Luis Congacha Yunda (201920042) y
Jaime Orlando Buitrago González (200612390)

03 de marzo de 2024

1 Introducción

Entre 1991 y 2019, el recaudo del impuesto a la renta aumentó en 7,1 puntos porcentuales (p.p.) dentro de la estructura tributaria de América Latina y el Caribe, consolidándose como la segunda fuente de ingresos (26,6%), después de los impuestos generales sobre bienes y servicios (CEPAL 2023). En contraste, para el último año, Colombia se encontró por encima del promedio de sus pares regionales, pues el impuesto a la renta representó el 32,3% de su estructura tributaria; sin embargo, al compararla carga la tributaria de este impuesto, Colombia está lejos del grupo de países más desarrollados del cual hace parte, pues la renta representa el 6,4% del PIB, un valor lejano del promedio de la OCDE (34%).

Más aún para el 2019, el recaudo del impuesto de renta de las personas correspondió a tan sólo al 1,3% del PIB colombiano según la CEPAL (2023). Este nivel de ingresos tributarios es considerablemente bajo y se debe al complejo sistema tributario que permite mayores exensiones y deducciones para quienes tienen ingresos más altos (Fergusson y Hofstetter 2022), al mismo tiempo que las políticas, factores psicológicos, coyunturas económicas y deficiencias administrativas favorecen la evasión y elusión de este impuesto sobre los ingresos (García, Parra, y Rueda 2021).

Cada reforma tributaria ha intentado mejorar los niveles de recaudo en los últimos años, así mismo, los cambios administrativos en la DIAN también han contribuido a este propósito, buscando un sistema tributario más efciente y progresivo, ya que se cuenta con suficiente información y evidencia para identificar sus principales problemas. No obstante, los comportamientos de las personas, sobre todo aquellos asociados a conductas delictivas (como la elusión y evasión), son más difíciles de identificar ante su carácter ilegal y ético. Instrumentos como el *Machine Learning* (ML) pueden complementar y contribuir a aumentar el recaudo de los impuestos, haciendo visible aquello que las personas buscan ocultar o esconder, es decir, generando predicciones sobre los ingresos de las personas a partir de características individuales, de los hogares, o incluso territoriales.

Así por ejemplo, en Indonesia, a través de modelos predictivos se identificaron potenciales pagos a las deudas por impuestos a partir de los registros administrativos de la autoridad fiscal de ese país (Febriminanto y Wasesa 2022); en Armenia, se pudieron identificar posibles fraudes fiscales a partir de la información reportada por los compradores y vendedores con herramientas de ML (Baghdasaryan et al. 2022); o en Brasil, Sao Paulo, se identificaron potenciales pagadores, ingresos,

el monto de los impuestos y multas a partir de los registros administrativos de autoridad fiscal del municipio (Ippolito y Garcia 2020). Como revisión de literatura indicativa, más no exahustiva, para países con algunas características similares a las de Colombia, es posible identificar las potencialidades de herramientas como el ML en el recaduo de los impuestos asociados a los ingresos.

Es por tanto, que en este Problem Set busca predecir el ingreso por hora de los y las bogotanas mediante modelos que aprovechan las herramientas del ML, haciendo uso de la principal encuesta de hogares colombiana. Más específicamente, se utilizó la *Medición de Pobreza Monetaria y Desigualdad* (Sarmiento-Barbieri 2024) del 2018 para Bogotá, un módulo de la *Gran Encuesta Integrada de Hogares* (GEIH) del DANE, la cual no sólo proporciona información sobre el mercado laboral, ingreso de las personas, características socioeconómicas y territoriales, si no a su vez representa una fuente de información confiable en la cual las personas no tienen incentivos a reportar información falsa sobre su ingreso, pues la encuesta es anónima y no tiene una finalidad tributaria.

En este contexto, el presente documento ... (preview of the results an main takeaways)

2 Datos

La GEIH es la principal y tal vez más importante encuesta de hogares con la que cuenta Colombia actualmente. Mensualmente, el DANE recolecta información sobre los ingresos y el mercado laboral de una muestra representativa de la población colombiana, de tal manera, que cada mes se obtienen datos sobre el ingreso y mercado laboral para el ámbito nacional, y anualmente para 23 departamentos y Bogotá, sus capitales y áreas metropolitanas, y otros dominios (rural y urbano) (DANE 2019). La operación estadística tiene como resultado una base de datos anual de aproximadamente 750 mil observaciones o personas, 230 mil hogares y 30 mil viviendas, la cual permite realizar cálculos y estimaciones sobre la población colombiana.

Aunque la GEIH recolecta información sobre los ingresos y el mercado laboral de la población, se realizan poco más de 150 preguntas a los encuestados que capturan información demográfica, económica y social, de tal manera que se expanden las posibilidades de la encuesta. Es así como esta base de datos también permite caracterizar la migración, micronegocios, la transición entre la educación y el trabajo, trabajo infantil, tecnologías de la información y la pobreza monetaria (dentro de los módulos más importantes). Este último módulo es importante para el desarrollo del presente Problem Set, ya que el DANE agrega los ingresos salariales y no salariales per cápita de las unidades de gasto (hogares para simplificar), identificando las personas con ingresos superiores e inferiores a las líneas de pobreza e indigencia definidas en el Comité de Expertos [1]. Es así como la GEIH es también la principal fuente de información para calcular la incidencia, brecha y severidad de esta medición del bienestar de la población.

La GEIH y su módulo sobre la *Medición de Pobreza Monetaria y Desigualdad* están disponibles al público en general a través de la Archivo Nacional de Datos (ANDA) del DANE. Sus módulos anonimizados se pueden descargar y unir para la investigación académica, la toma de decisiones o cualquier propósito individual. En este caso, para el Problem Set no se realizó el descargue de la página del DANE, sino se realizó un *web scraping* de la base de datos filtrada para Bogotá,

[1] Personas o representante de entidades nacionales o de cooperación internacional con la experticia técnica para orientar la operación estadística, cálculos y estimaciones de la pobreza monetaria

de la página web de Ignacio Sarmiento Barbieri (Sarmiento-Barbieri 2024) como caso práctico de extracción de contenidos de una página web, con un formato estructurado.

En la página web se encuentran 10 tablas en formato HTML, las cuales contienen en total las 32.177 observaciones que componen la muestra de personas para Bogotá de la GEIH para 2018, con 179 variables (21 variables adicionales a las presentes en la base de datos del ANDA). Al encontrarse en un formato estructurado (tabla HTML), se realizó el *web scrpaing* haciendo uso del paquete `RSelenium` de `RStudio`, con el cual se automatiza la consulta de cada una de los 10 hipervínculos de la página web, se identifica la tabla en HTML, se almacena y posteriormente se consolida una sola base de datos con las características anteriormente mencionadas.

Como complemento, se descargaron y unieron las bases de datos de la GEIH de 2018 (DANE 2022), de tal manera que se obtuviesen variables complementarias para el desarrollo del Problem Set. Más específicamente, los años de escolaridad, la rama de actividad [2] y la pregunta sobre la posición ocupacional de las personas ocupadas. De esta manera, se obtuvieron variables complementarias como escolaridad como variable continua, experiencia laboral [3], sector económico y posición ocupacional.

Finalmente, se filtró la base de datos con las personas mayores de 18 años, quedando en total 16.542 observaciones o personas para el desarrollo del Problem Set. De esta manera, a continuación se presentan las estadísticas descriptivas de las variables utilizadas en los siguientes puntos.

De las 32177 observaciones, se puede observar la tasa de valores perdidos de las variables de interés:

3 Perfiles de salario por edad

4 Brechas de ingreso por sexo

5 Predicción de salarios

La meta es predecir la brecha salarial del logaritmo del ingreso. Comenzamos estimando el modelo más sencillo de todos, es decir, el modelo univariado:

$$\ln(w) = \beta_1 + \beta_2 Mujer + u \tag{1}$$

El coeficiente $\beta_2 < 0$ indica que las mujeres, en promedio y ceteris praibus, reciben un salario mensual 0.8657713 menos ingresos que los hombres.

5.1 Estimación con FWL

En la primera etapa, *partialling-out*, ejecutamos dos regresiones. Suponiendo que la variable de interés es la dicotómica de $Mujer$, entonces en la primera regresión buscamos estimar `woman~x1+x2+...`, donde las `xi` son todas aquellas demás variables de control para corregir el potencial sesgo de variable omitida. Luego nos quedamos con los residuales `woman_res`, y

[2] Código a dos dígitos de la Clasificación Industrial Internacional Uniforme de todas las actividades Económicas (CIIU).

[3] Variable *proxy* construida como Edad-Años de escolaridad-Años de ingreso al sistema educativo (6 años)

	Dependent variable:	
	Ln Salario	
	(1)	(2)
age		0.012***
		(0.001)
womanMujer	−0.144***	−0.124***
	(0.016)	(0.014)
relab		0.126***
		(0.020)
Constant	14.076***	13.898***
	(0.011)	(0.192)
Observations	9,964	9,963
R^2	0.009	0.453
Adjusted R^2	0.008	0.448
Residual Std. Error	0.775 (df = 9962)	0.579 (df = 9876)
F Statistic	86.083*** (df = 1; 9962)	94.912*** (df = 86; 9876)
Note:		*p<0.1; **p<0.05; ***p<0.01
		Controles: oficio, maxEduclevel

Cuadro 1: Estimando brecha de género

ejecutamos una segunda regresión en la que estimemos `log_wage~x1+x2...`, y guardamos estos residuales, `log_wage_res`.

Finalmente ejecutamos la segunda regresión univariada `log_wage_res~woman_res` y en principio deberíamos obtener el mismo coeficiente de haber ejecutado el modelo completo.

6 Referencias bibliográficas

Baghdasaryan, V., H. Davtyan, A. Sarikyan, y Z. Navasardyan. 2022. «Improving Tax Audit Efficiency Using Machine Learning: The Role of Taxpayer's Network Data in Fraud Detection». *Applied Artificial Intelligence* 36 (1): 2012002. https://doi.org/10.1080/08839514.2021.2012002.

CEPAL. 2023. *Panorama fiscal de América Latina y el Caribe 2023*. CEPAL.

DANE. 2019. «Medición de Pobreza Monetaria y Desigualdad 2018». [Base de datos].`https://microdatos.dane.gov.co/index.php/catalog/608`

DANE. 2022. «Gran Encuesta Integrada de Hogares 2018. Empalmada». [Base de datos].`https://microdatos.dane.gov.co/index.php/catalog/758`

Febriminanto, R., y M Wasesa. 2022. «Machine Learning Analytics for Predicting Tax Revenue Potential». *Indonesian Treasury Review* 7 (3): 193-205.

Fergusson, L., y M. Hofstetter. 2022. «The Colombian tax system: A diagnostic review and proposals for reform». UNDP.

García, M., O. Parra, y F. Rueda. 2021. «Features of tax structure and tax evasion in Colombia». *Apuntes Contables*, n.º 28: 17-40.

Ippolito, A., y A. Garcia. 2020. «Tax Crime Prediction with Machine Learning: A Case Study in the Municipality of Sao Paulo». *ICEIS 2020* 1: 452-59.

Sarmiento-Barbieri, I. 2024. «Problem Set 1. BDML». [Base de datos].`https://ignaciomsarmiento.github.io/GEIH2018_sample/`

7 Ejemplos

Para incrustar código y resultados de la consola

```
summary(cars)
```

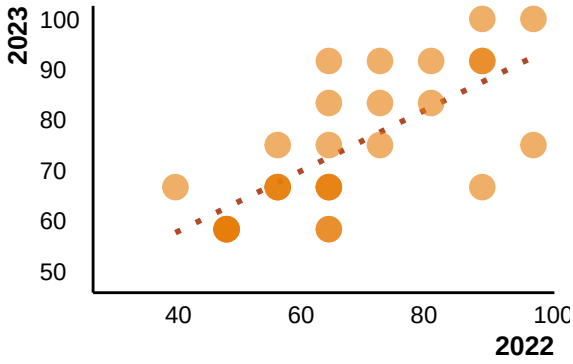
```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Para incluir ecuaciones

$$w = f(X) + u$$

Para incluir gráficas

Figura 1: Título de la gráfica



Fuente: Cálculos propios a partir de @geih