

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 con 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 exenciones 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 eficiente 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 exhaustiva, para países con algunas características similares a las de Colombia, es posible identificar las potencialidades de herramientas como el ML en el recaudo 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 *Gran Encuesta Integrada de Hogares* (GEIH) del DANE para el 2018 (Sarmiento-Barbieri 2024), 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

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:

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

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 $\text{woman} \sim x_1 + x_2 + \dots$, donde las x_i son todas aquellas demás variables de control para corregir el potencial sesgo de variable omitida. Luego nos quedamos con los residuales woman_res , y ejecutamos una segunda regresión en la que estimemos $\log_wage \sim x_1 + x_2 + \dots$, y guardamos estos residuales, \log_wage_res .

Finalmente ejecutamos la segunda regresión univariada $\log_wage_res \sim \text{woman_res}$ y en principio deberíamos obtener el mismo coeficiente de haber ejecutado el modelo completo.

<i>Dependent variable:</i>		
	Ln Salario	
	(1)	(2)
age		0.012*** (0.001)
womanMujer	-0.144*** (0.016)	-0.124*** (0.014)
relab		0.126*** (0.020)
Constant	14.076*** (0.011)	13.898*** (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

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.
- 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

Fuente: Cálculos propios a partir de @geih

Figura 1: Título de la gráfica

