Problem set 1: Predicting income

Big Data and Machine Learning
Febrero 2025

Carlos Manjarres, Juan Felipe Triana, Sebastian Trujillo and Maya Gutiérrez

Github link: https://github.com/MayaGutiBan/BDML_prob_set_1.git

1 Introduction

Income prediction models play a crucial role in economic and public policy analysis, particularly in addressing issues related to tax evasion, social inequality, and fiscal sustainability. In Colombia, tax evasion has reached critical levels. The National Directorate of Taxes and Customs (DIAN) estimate an annual fiscal loss of approximately $65 billion, equivalent to 5.4% of GDP (DIAN, 2022). The most affected tax revenues include income tax and value-added tax (VAT), highlighting the need for improved detection mechanisms.

Moller (2012) argues, that tax collection mechanisms can exacerbate inequality and directly impact social equity. When individuals with fewer resources bear a disproportionately high tax burden, overall government revenues may decline, leading to budgetary constraints that further affect public services and social welfare programs. For example, some studies have established links between tax evasion, informal employment, and unemployment (Calderon Diaz et al., 2020). Informal workers often do not contribute to social security systems, lack health insurance, and fail to pay income taxes, which undermines the financial stability of pension and healthcare systems.

Furthermore, two types of inequalities are present in the colombian tax system: vertical and horizontal inequalities. The former occur when individuals or companies with higher income contribute less than the rest as a proportion of their taxable base; the latter, when having the same amount of income, some pay more than others (Concha et al., 2017). In this regard, the DIAN calculates that, in both forms of evasion, under-declaration in natural persons is higher than 50%. A possible explanation for these inequalities is the variety of existing special exemptions such as non-taxable income, special income, surtaxes, compensations, extraordinary deductions, exempt income and tax credits. Although the exact calculation of what the country loses for this concept is not available, according to approximations made by ECLAC, this percentage is higher by about 20% when compared to Chile, Mexico and Peru (Concha et al., 2017). The collection of taxes on individuals would improve if the effective tax rates increase and if a more uniform treatment is given with very few rates and differential treatments for each source of income.

Given this context, the ability to conduct precise analyses of individuals' incomes is essential for detecting tax fraud and designing effective public policies to mitigate its effects on tax administration. This study develops an income prediction model for Colombian households in 2018, utilizing data from the Gran Encuesta Integrada de Hogares (GEIH).

One key application is identifying cases of income underreporting, thereby improving tax compliance and reducing fiscal losses. Additionally, the model can help pinpoint vulnerable populations requiring social assistance, ensuring better-targeted government interventions. The result of this exercise in predicting incomes shows better fits for conditional models that integrate a variety of controls (specifically demographic variables such as gender, age, socioeconomic class, education level and number of children per household), variables that account for experience and total time

worked as well as a characterization of the work environment (type of work, job title and size of the firm). In the case of the specified model, model 6 was the best fitted model. This model included log transformations of numerical variables for a better fit, showing improvement in predictive performance on the test set and a better generalization to unseen data. However, there are indications that this model is more sensitive to specific data points,which makes it less robust and suggests that there are unaccounted for unseen variables.

The results of this study demonstrate the effectiveness of predictive models in estimating household income, offering insights into the determinants of income levels and the factors influencing tax evasion. These findings underscore the importance of integrating data-driven approaches in economic policymaking to enhance fiscal management and promote social equity.

2 Data description

(a) Describe the data briely, including its purpose, and any other relevant information.

The Gran Encuesta Integrada de Hogares (GEIH) is a survey that has been implemented since 2006 and is collected from probabilistic, stratified, clustered and multistage samples by the *Departamento Administrativo Nacional de Estadística* (DANE). It covers topics such as income, property, savings and investment, employment, access to services, housing, and continuing education, among others. The survey has national coverage that allows obtaining results for capitals, cities and metropolitan areas, large regions and totals by department. For this study, we are evaluating the GEIH data collected in 2018 and the data set used contains all individuals sampled in Bogota.

2.1 Data Compilation Process

(b) Describe the process of acquiring the data and if there are any restrictions to accessing/scraping these data.

The data used in this study was obtained from the website of Ignacio Sarmiento (https://ignaciomsarmiento.github.io/GEIH2018sample/.) and was scraped using RStudio. To extract the information we used the rvest package. However, given that JavaScript was performing additional network requests to retrieve and insert content, a dynamic loading process known as AJAX (Asynchronous JavaScript and XML) or XHR (XML HTTP Request), we had to access the link through the ' network' window and filtering by Fetch/XHR, in order to detect and replicate the network requests that the page makes to obtain the information that feeds the table. We then implemented an automated process using a loop to systematically collect all available records from each data chunk. After gathering the data, we performed a merging procedure to consolidate the extracted information into a unified dataset. In total, the master database consisted of 32,177 observations and 178 variables.

(c) Describe the data cleaning process.

In this problem set, we first created a new variable that counted the number of children (6 years or younger) per household and then restricted the data to individuals who are employed and over 18 years old. The resulting dataset contains 16,542 observations and 179 variables. We selected twenty seven of most relevant variables based on literature reviews and economic intuition, which included a combination of categorical and numerical variables. Next, we identified potential data collection errors by examining the missing values and counted the zero values for numeric variables such as number of hours worked (totalHoursWorked) and Income (Ingtot). For missing values of numeric variables, the mean was computed, and the median was used to replace missing values of categorical

variables. Some variables were renamed for clarity and some new variables were constructed, such as H_Head, an indicator variable that indicates if the person is the head of the household.

We also detected possible outliers in the variable total hours worked. To address this, we considered applying a cutoff for individuals reporting more than 120 hours per week. Also we found, observation with total income equal to zero, employed and over the age of 18. To handle them, we impute the average total income for low wage individuals with the objective to mitigate potential measurement errors.

(d) Describe the variables included in your analysis.

Table 1: Summary statistics for continuous variables

Statistic	Mean	St. Dev.	Min	Max
Age	39.436	13.483	18	94
Income	1,791,931.000	2,668,141.000	15,000.000	85,833,333.000
Work experience	63.758	89.488	0	720
Hours worked	47.403	15.662	1	130
Children per household	0.300	0.566	0	4

For data visualization, we first created box plots for the variables log(total income) , total hours worked, age, self-employment by gender, and maximum education level by gender. Box plots are useful for summarizing the distribution of a variable, identifying central tendencies, and detecting potential outliers. There are extreme values for Total Hours Worked, which supports the decision to consider a cutoff above 120 hours per week. The income distribution, when logged, shows significant dispersion, particularly among higher-income individuals. Finally Self-Employment and Max Education Level by gender suggests possible differences in income distributions between men and women. Next, we created histograms for age, total income, and total hours worked to examine their distributions. Histograms help visualize the frequency of different values, revealing patterns such as skewness or the presence of extreme values. The Age histogram follows a right-skewed distribution, with most individuals concentrated between 20 and 60 years old, this might be because we eliminate the observation of individuals under 18. The Total Hours Worked appears to have extreme values, suggesting the presence of outliers. Finally Total Income (log) shows a relatively normal distribution but with some dispersion at the higher end.

Justification for selected variables

- Demographic variables (gender-female-, socio-economic class-estrato1-, age, number of children per houshold-nmenores-,): Becker also addressed employment discrimination and how factors such as gender, ethnicity and age can influence wages. His research contributed to the understanding of discrimination in the labor market and how it can affect income distribution. Some relevant demographic variables, such as urban or rural labor markets are not accounted for due to the concentration of the analysis to individuals in the city of Bogotá.

> **-Gender:** The recent improvements in women's educational attainment have led to a global rise in their workforce participation. Nonetheless, disparities in earnings persist, disproportionately favouring men regardless of educational achievements. In addition to the underrepresentation of women in the lucrative STEM and ICT fields, the gender pay gap highlights distinct patterns of job mobility between genders. Women are less

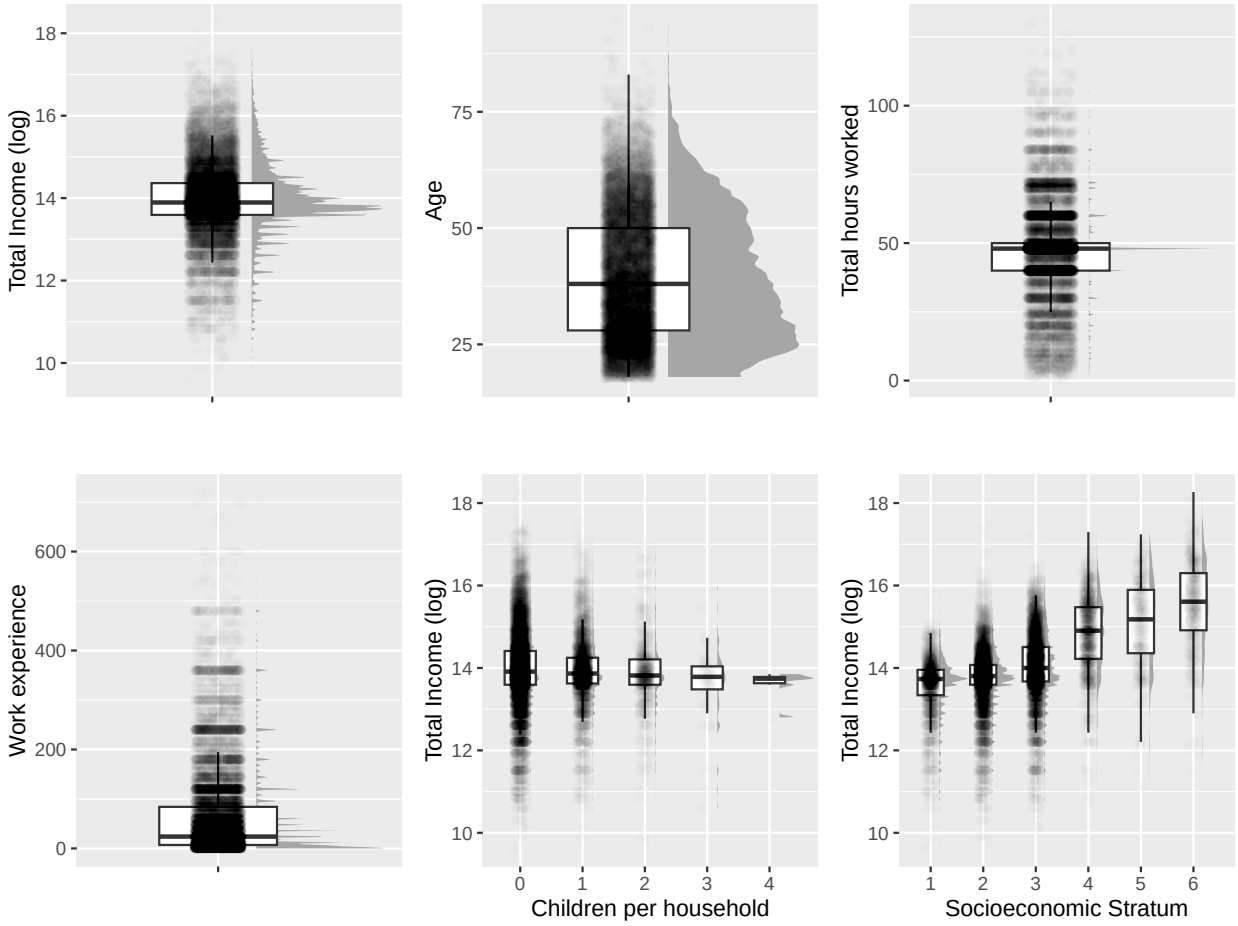

Figure 1: Variable distributions

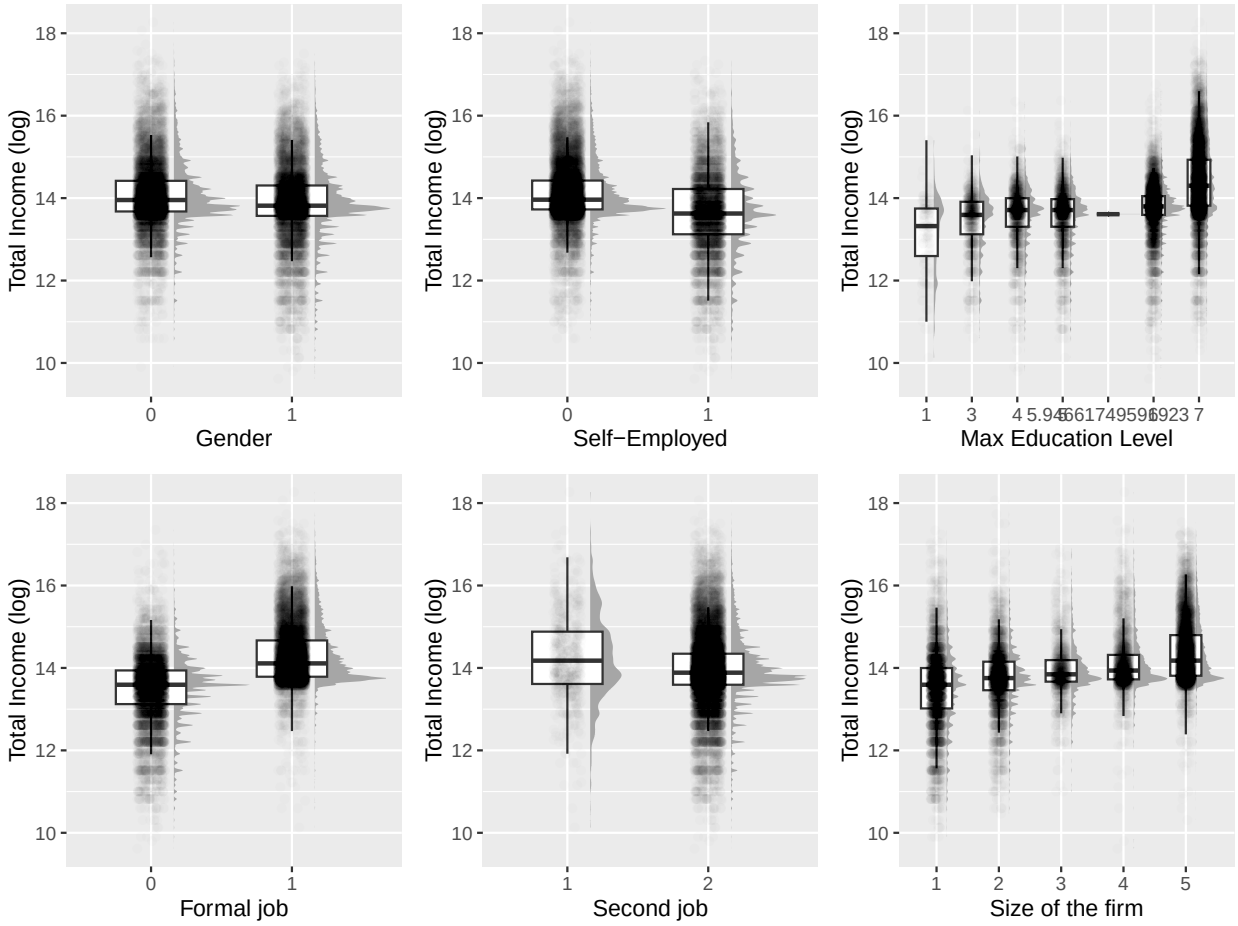

Figure 2: Variable distributions

likely than men to receive promotions or significant wage increases when switching employers.(https://www.oecd.org/en/topics/earnings-by-educational-attainment.html)

-Socioeconomic Class (estrato1): In Colombia, socioeconomic stratification is a classification in strata of residential properties to charge public utilities in a differential manner by strata, allowing subsidies to be assigned and contributions to be collected accordingly. There is a correlation of higher socioeconomic class and intergenerational wealth accumulation effects. Anther potential example are higher socioeconomic class individuals who are more likely to work in professional, managerial, or technical roles, which typically offer higher salaries, while lower socioeconomic class could tend to be individuals that work in manual labor or service jobs, which tend to have lower wages. Furthermore, social networks can play a role, by providing access to better job opportunities, mentorship, and career advancement.

-Number of children per household (nmenores: The number of children per household may affect a persons income through the following channels: on the one hand, the cost of child care and reduced savings may mean that these families have less capacity to save or invest, limiting their ability to generate additional income, Their labor market participation may be limited, especially for mothers, who may reduce their work hours or exit the labor force temporarily to care for children, leading to lower household income. Additionally there are higher probabilities of career interruptions which can result in missed promotions, skill depreciation, and lower lifetime earnings.

- Education level (maxEducLevel): Becker in his book "Education and Earnings" (1975) argued that education is an investment in human capital and that people who invest in their education tend to earn higher wages. Formal education, such as college or graduate degrees, can improve wage prospects.

- Total time worked (totalHoursWorked) and Time worked in company, business, industry (p6426-tiempoTrabajando): In the article "A Theory of the Allocation of Time" (**<empty citation>**)(1965), Becker examines how people make decisions about how to allocate their time between work, leisure and other activities, which has implications for labor income. Similarily, Becker mentions in his article "Investment in Human Capital: A Theoretical Analysis" (1962) that work experience, seen as human capital, affects labor income over time.

- Work Characterization (Type of work -oficio-, Job title -relab-, Size of the firm -sizeFirm-): The type of work could impact income because of the difference between high-paying industries such as finance, technology, and healthcare, typically offer higher wages due to the specialized skills required, high demand for labor, and profitability of the sector. In comparison, low-paying industries like retail, hospitality, and agriculture often offer lower wages due to lower skill requirements, higher labor supply, and lower profit margins. This is related to the skill level required, hierarchy and seniority in the position, and level of specialization required for the work. Similarily, the size of the firm can impact the wages of the workers through resource availability and possibilities of benefits and perks, as well as career advancement opportunities.

3 Age wage earning

The age-wage profile describes the relationship between a worker's age and their earnings over their career. It is often upward-sloping, reflecting the accumulation of human capital through experience and training. According to Mincer (1974), earnings typically increase with age as workers gain skills, knowledge, and productivity. The human capital theory (Becker, 1964) explains that individuals invest in education and on-the-job training early in their careers, leading to lower initial wages but higher earnings growth over time. Additionally, firms may offer lower starting wages as part of implicit long-term labor contracts that reward workers in later years (Lazear, 1979).

Studies have empirically shown that wages tend to be low for younger workers, increase steadily, and peak around the age of 50 before stabilizing or slightly declining (Murphy & Welch, 1990). The decline after 50 is often attributed to factors such as diminishing returns to experience, lower adaptability to new technologies, and potential health-related productivity declines (Johnson Neumark, 1996). These patterns are observed consistently in labor market data across different countries and industries, reinforcing the robustness of the age-earnings profile in economic research.

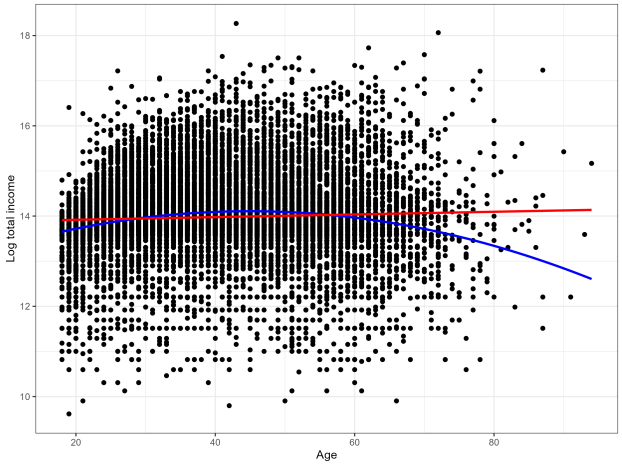

Figure 3: Income-Age scatter plot

To support these findings empirically, we are going to use data from the 2018 Medición de Pobreza Monetaria y Desigualdad from the GEIH for Bogotá. The sample is restricted for employed people over 18 years old and accounts for their total income when it is different to cero. We exclude observations with zero income to focus only on employed individuals, as those with no earnings are often out of the labor market (e.g., students, retirees, unemployed without benefits). Including them would distort the estimated relationship between age and wages, leading to biased results. Additionally, zero-income observations can create econometric issues such as heteroskedasticity and model misspecification, making standard regression techniques less reliable, specially when in order to reduce the variance of income we use the logarithm of income.

Figure 3 illustrates the relationship between these two variables in the sample. At first glance, no clear pattern emerges due to the high variance in income, as reflected in the nearly flat linear fit. However, upon closer examination (particularly for higher income levels) an inverted U-shape becomes apparent in the scatter plot. This pattern is captured by the quadratic fit, which reveals a slight but noticeable curvature in the relationship.

A semilogarithmic Log-Lin labor market model was used to estimate the effect of age on wages. A

quadratic term was added in order to estimate the age-wage profile in a way that captures both the overall trend and potential nonlinearities in the relationship. The estimation of the beta parameters was done by Ordinary Least Squares (OLS).

Table 2: Age-wage regression

	log_ingtot	
	Age	Age^2
	(1)	(2)
age	0.003***	0.056***
	(0.001)	(0.003)
I(ageˆ2)		−0.001***
		(0.000)
Constant	13.859***	12.849***
	(0.021)	(0.058)
Observations	16,542	16,542
R^2	0.002	0.023
Adjusted R^2	0.002	0.023
Residual Std. Error	0.871 (df = 16540)	0.862 (df = 16539)
F Statistic	34.645*** (df = 1; 16540)	191.842*** (df = 2; 16539)
Notes:	***Significant at the 1 percent level.	
	**Significant at the 5 percent level.	
	*Significant at the 10 percent level.	

In the linear model (1), the coefficient on age (0.003) is positive and statistically significant at the 1% level, suggesting that wages tend to increase with age. However, the effect is quite small, indicating that a simple linear specification may not adequately capture the full dynamics of the age-wage relationship. In model (2), which includes a quadratic term for age, the coefficient on age (0.058) becomes larger, while the coefficient on age^2 (-0.001) is negative and highly significant. This suggests an inverted U-shaped relationship, where wages initially rise with age but eventually reach a peak and then decline. This finding aligns with standard and mentioned above labor economics theories, which suggest that productivity and earnings increase with experience but may taper off due to factors such as skill depreciation or reduced labor market mobility in later years.

The inclusion of the quadratic term significantly improves the model's explanatory power. The R^2 increases from 0.002 to 0.024, indicating that while the linear model explains only 0.2% of the variance in wages, the quadratic model accounts for 2.4%, a twelvefold improvement. The residual standard error also decreases from 0.877 to 0.867, and the F-statistic rises substantially, reinforcing the idea that the quadratic model provides a much better fit. Overall, these results highlight the importance of considering nonlinearities in wage determination, as the simple linear model fails to capture the full earnings trajectory over the life cycle.

This analysis suggests that there is a peak-age where on average, total income is at its maximum level. Given the quadratic estimation above, we can compute this age for the sample. To do this,

we use the fact that the maximum value of income occurs when the first order derivate of (2) equals cero

Since the model is

$$\log(\text{income}) = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{age}^2 \tag{1}$$

The maximum value (peak) occurs where the derivative equals zero:

$$\frac{d}{d\text{age}} \left(\beta_1 \cdot \text{age} + \beta_2 \cdot \text{age}^2 \right) = 0 \tag{2}$$

Solving for age:

$$max(age) = \frac{-\beta_1}{2\beta_2} \tag{3}$$

To obtain a robust estimate of this peak, we use bootstrapping, which allows us to account for sampling variability and improve inference. The bootstrap results indicate that the mean estimated peak age is 44.8 years, with a 95% confidence interval of (43.80, 45.85). This suggests that, on average, wages tend to rise until around mid-40s, after which they gradually decline, confirming the inverted U-shaped age-wage profile observed in the data.

4 Gender wage gap

4.1 Conditional and Unconditional Models

The two models used to estimate the gender wage gap are the following:

Unconditional:

$$\log(\text{ingtot}_i) = \beta_0 + \beta_1 \text{female}_i + \varepsilon_i \tag{4}$$

Conditional :

$$\begin{aligned} \log(\text{ingtot}_i) = \beta_0 &+ \beta_1 \text{female}_i + \beta_2 \text{age}_i + \beta_3 \text{age}_i^2 + \beta_4 \text{cuentaPropia}_i + \beta_5 \text{estrato1}_i \\ &+ \beta_6 \text{formal}_i + \beta_7 \text{maxEducLevel}_i + \beta_8 \text{parentesco\_jhogar}_i + \beta_9 \text{tiempo\_trabajando}_i \\ &+ \beta_{10} \text{otro\_trabajo}_i + \beta_{11} \text{relab}_i + \beta_{12} \text{sizeFirm}_i + \beta_{13} \text{totalHoursWorked}_i \\ &+ \beta_{14} \text{oficio}_i + \beta_{15} \text{otro\_ingreso}_i + \beta_{16} \text{nmenores}_i + \varepsilon_i \end{aligned} \tag{5}$$

With these two models, our aim is to decompose the gender gap and assess how much of it is attributable to observed differences versus potentially discriminatory factors. The purpose of the unconditional model is to estimate the raw association between gender and income without controlling for other factors, providing a baseline measure of the overall disparity. In contrast, the conditional model introduces control variables relevant to the dependent variable. By doing so, we can account for systematic differences in income that are not directly related to gender, allowing us to better isolate the true effect of gender on earnings. Comparing both models helps determine whether the observed gap is primarily driven by differences in characteristics (e.g., education, occupation, hours worked) or if an unexplained residual gap persists, which could indicate potential discrimination or other unobserved factors.

Table 3: Gender wage regression

	log_ingtot Unconditional Model	log_ingtot Conditional Model
	(1)	(2)
female1	-0.189^{***}	
	(0.013)	
female		-0.143^{***}
		(0.011)
Constant	14.064^{***}	13.501^{***}
	(0.009)	(0.151)
Controls	NO	YES
Observations	16,542	16,542
R^2	0.012	0.588
Adjusted R^2	0.012	0.585
Residual Std. Error	0.867 (df = 16540)	0.562 (df = 16422)
F Statistic	195.073^{***} (df = 1; 16540)	196.746^{***} (df = 119; 16422)

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.
Controls for age, education, experience, sector, job title, hours worked, firm size and children.

In the first place, from the results of the linear models we know that both coefficients associated with the *female* variable are statistically significant at the 1% level, indicating that gender has a meaningful impact on the logarithm of total income. Secondly, in both models, being female is associated with a decrease in total income. In the unconditional model, being female reduces total income by approximately 18.9%, *ceteris paribus*, while in the conditional model, the reduction is 14.3%, *ceteris paribus*.

The persistence of this negative effect, even after controlling for relevant variables, suggests that the income gap between genders may be driven by multiple factors. One possible explanation is selection bias, where women systematically sort into lower-paying occupations or sectors due to structural constraints or personal preferences. Another possibility is labor market discrimination, where equally qualified women receive lower wages than their male counterparts due to biases in hiring, promotion, or wage-setting practices. The reduction in the magnitude of the gender gap coefficient when controls are included implies that some of the observed disparity can be attributed to differences in observable characteristics. However, the persistence of a significant negative coefficient even in the conditional model suggests that discrimination or unobserved factors may still be at play.

Table 4: Gender wage regression using FWL theorem

	log_ingtot Conditional Model	res_log_ingtot Conditional model with FWL
	(1)	(2)
female	-0.143^{***}	
	(0.011)	
res_female		-0.143^{***}
		(0.011)
Constant	13.501^{***}	-0.000
	(0.151)	(0.004)
Controls	NO	YES
Observations	16,542	16,542
R^2	0.588	0.010
Adjusted R^2	0.585	0.010
Residual Std. Error	0.562 (df = 16422)	0.560 (df = 16540)
F Statistic	196.746^{***} (df = 119; 16422)	166.643^{***} (df = 1; 16540)
Notes:	***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level. Controls for age, education, experience, sector, job title, hours worked, firm size and children.	

From table 4 we know that the estimators from FWL and FWL with bootstrap are the same. Both are statistically significant, and have equal bias. How ever the estimator from FWL with bootstrap have an slightly lower residual std error.

4.2 Peak ages by gender

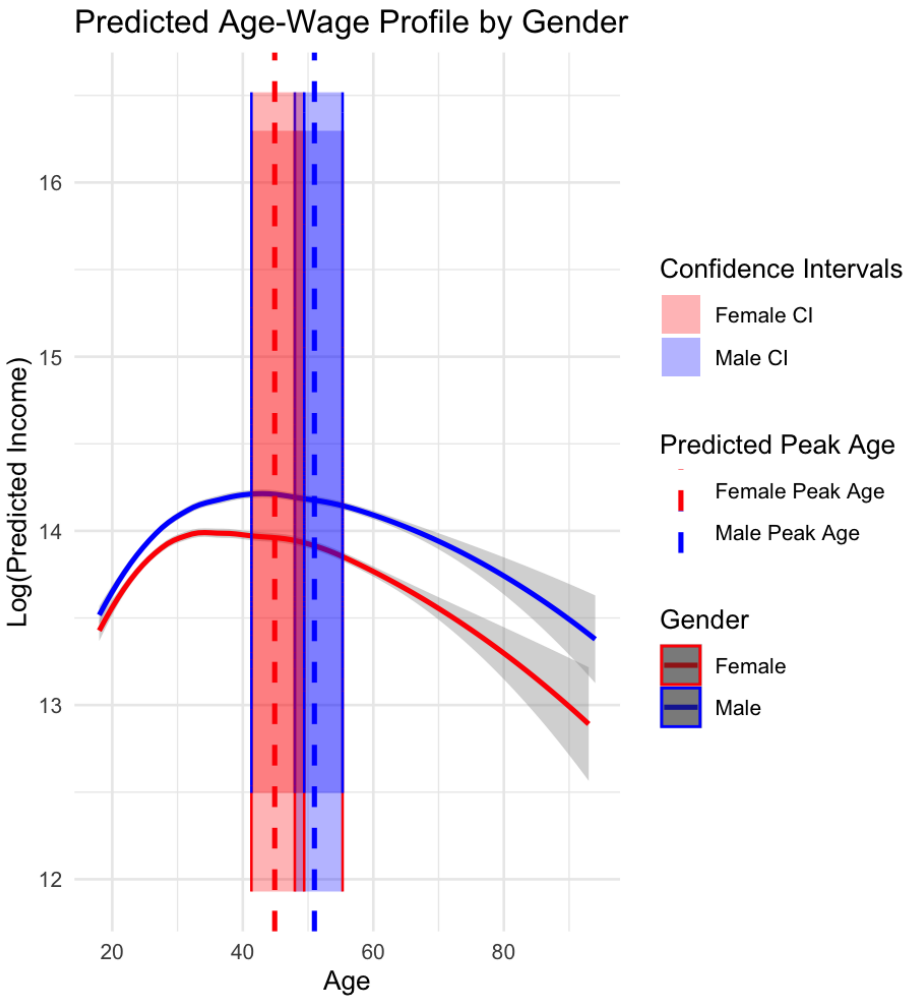

Figure 4: Peak ages by gende

Figure 3 highlights a persistent gender income disparity. First of all, men have consistently higher predicted income, than women at any given age. Also, men tend to peak later in their careers, the estimated peak age for women is approximately 48 years, with a standard error of 2.55. Although men estimated peak age is approximately 53 year with std error of 1.91.

Figure 3, combined with previous statistical analysis done in section 4.1, strongly supports the claim that women suffer from gender-based wage discrimination. It highlights how the wage gap is not merely due to personal choices or differences in experience but rather a systemic disadvantage faced by women in the labor market.

5 Predicting income

In this section we report and compare the predictive performance in terms of the RMSE of seven different model specifications that explore non-linearities and complexity of relationships between

predictive variables.

The following models were calculated:

Model 1: Simple linear model with categorical variables

$$\log(\text{ingtot}_i) = \beta_0 + \beta_1 \text{female}_i + \varepsilon_i \tag{6}$$

Model 2: Non conditional non-linearity

$$\log(\text{ingtot}_i) = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{age}_i^2 + \varepsilon_i \tag{7}$$

Model 3: Conditional non-linearity

$$\begin{aligned}\log(\text{ingtot}_i) = \beta_0 &+ \beta_1 \text{female}_i + \beta_2 \text{age}_i + \beta_3 \text{age}_i^2 + \beta_4 \text{cuentaPropia}_i + \beta_5 \text{estrato1}_i \\ &+ \beta_6 \text{formal}_i + \beta_7 \text{maxEducLevel}_i + \beta_8 \text{parentesco\_jhogar}_i + \beta_9 \text{tiempo\_trabajando}_i \\ &+ \beta_{10} \text{otro\_trabajo}_i + \beta_{11} \text{relab}_i + \beta_{12} \text{sizeFirm}_i + \beta_{13} \text{totalHoursWorked}_i \\ &+ \beta_{14} \text{oficio}_i + \beta_{15} \text{otro\_ingreso}_i + \beta_{16} \text{nmenores}_i + \varepsilon_i\end{aligned} \tag{8}$$

Model 4: Interaction effects

$$\begin{aligned}\log(\text{ingtot}_i) = \beta_0 &+ \beta_1 \text{female}_i + \beta_2 (\text{female}_i \times \text{age}_i) + \beta_3 \text{age}_i + \beta_4 \text{age}_i^2 \\ &+ \beta_5 \text{cuentaPropia}_i + \beta_6 (\text{cuentaPropia}_i \times \text{female}_i) + \beta_7 \text{estrato1}_i + \beta_8 (\text{estrato1}_i \times \text{female}_i) \\ &+ \beta_9 \text{formal}_i + \beta_{10} \text{maxEducLevel}_i + \beta_{11} \text{parentesco\_jhogar}_i + \beta_{12} \text{tiempo\_trabajando}_i \\ &+ \beta_{13} \text{otro\_trabajo}_i + \beta_{14} \text{relab}_i + \beta_{15} \text{sizeFirm}_i + \beta_{16} \text{totalHoursWorked}_i \\ &+ \beta_{17} \text{oficio}_i + \beta_{18} \text{otro\_ingreso}_i + \beta_{19} \text{nmenores}_i + \varepsilon_i\end{aligned} \tag{9}$$

Model 5: Higher-order age terms

$$\begin{aligned}\log(\text{ingtot}_i) = \beta_0 &+ \beta_1 \text{female}_i + \beta_2 \text{age}_i + \beta_3 \text{age}_i^2 + \beta_4 \text{age}_i^3 + \beta_5 \text{age}_i^4 \\ &+ \beta_6 \text{cuentaPropia}_i + \beta_7 \text{estrato1}_i + \beta_8 \text{formal}_i + \beta_9 \text{maxEducLevel}_i \\ &+ \beta_{10} \text{parentesco\_jhogar}_i + \beta_{11} \text{tiempo\_trabajando}_i + \beta_{12} \text{otro\_trabajo}_i + \beta_{13} \text{relab}_i \\ &+ \beta_{14} \text{sizeFirm}_i + \beta_{15} \text{totalHoursWorked}_i + \beta_{16} \text{oficio}_i + \beta_{17} \text{otro\_ingreso}_i + \beta_{18} \text{nmenores}_i + \varepsilon_i\end{aligned} \tag{10}$$

Model 6: Log transformations

$$\begin{aligned}\log(\text{ingtot}_i) = \beta_0 &+ \beta_1 \text{female}_i + \beta_2 \log(\text{age}_i) + \beta_3 \text{cuentaPropia}_i + \beta_4 \text{estrato1}_i \\ &+ \beta_5 \text{formal}_i + \beta_6 \text{maxEducLevel}_i + \beta_7 \text{parentesco\_jhogar}_i + \beta_8 \text{tiempo\_trabajando}_i \\ &+ \beta_9 \text{otro\_trabajo}_i + \beta_{10} \text{relab}_i + \beta_{11} \text{sizeFirm}_i + \beta_{12} \log(\text{totalHoursWorked}_i) \\ &+ \beta_{13} \text{oficio}_i + \beta_{14} \text{otro\_ingreso}_i + \beta_{15} \text{nmenores}_i + \varepsilon_i\end{aligned} \tag{11}$$

In terms of the overall performance of the tested models, models 3–6 have significantly lower RMSE values compared to Models 1 and 2, indicating that they perform much better in predicting log(ingtot). Additionally, Models 1 and 2 have low R^2 values (0.012 and 0.023, respectively), indicating that they explain very little of the variance in log(ingtot). Models 3–6 have much higher R^2 values (ranging from 0.588 to 0.597), indicating that they explain a substantial portion of the variance in log(ingtot). The Adjusted R^2 values are slightly lower than the R^2 values, which is

Table 5: Model comparison

	log_ingtot					
	(1)	(2)	(3)	(4)	(5)	(6)
female1	−0.189***		−0.143***	−0.060	−0.144***	−0.133***
	(0.013)		(0.011)	(0.037)	(0.011)	(0.011)
Constant	14.064***	12.849***	13.501***	13.484***	12.932***	12.631***
	(0.009)	(0.058)	(0.151)	(0.151)	(0.284)	(0.163)
Controls	NO	NO	YES	YES	YES	YES
Test set RMSE	0.867	0.862	0.595	0.594	0.593	0.593
Observations	16,542	16,542	16,542	16,542	16,542	16,542
R^2	0.012	0.023	0.588	0.591	0.589	0.597
Adjusted R^2	0.012	0.023	0.585	0.588	0.586	0.594
Residual Std. Error	0.867 (df = 16540)	0.862 (df = 16539)	0.562 (df = 16422)	0.560 (df = 16415)	0.561 (df = 16420)	0.555 (df = 16423)
F Statistic	195.073*** (df = 1; 16540)	191.842*** (df = 2; 16539)	196.746*** (df = 119; 16422)	188.038*** (df = 126; 16415)	194.233*** (df = 121; 16420)	206.344*** (df = 118; 16423)

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.
Controls for age, education, experience, sector, job title, hours worked, firm size and children.

expected as they penalize the addition of unnecessary variables. However, they remain high for Models 3–6, confirming that these models are well-specified.

All models have significant F-statistics (p ¡ 0.01), indicating that the models are statistically significant overall. However we can conclude that models 1 and 2 are simple models with poor predictive performance and low explanatory power, while models 3–6 are more complex and perform significantly better in terms of both prediction (lower RMSE) and explanation (higher R^2 and Adjusted R^2).

The lowest prediction error is achieved by Model 6, with an RMSE of 0.5927824. This model includes logarithmic transformations, specifically log(age) and log(totalHoursWorked) and all the relevant controls. This suggests that non-linear transformations and flexible functional forms improve the model's predictive accuracy.

5.1 Potential tax evasion

If the outliers are due to data issues (e.g., extreme or incorrect values), the DIAN should investigate these individuals. On the other hand, if the outliers are due to model limitations, they may not represent actual anomalies but rather areas where the model can be improved. If the outliers are concentrated in specific subgroups (e.g., certain job titles, sectors, or education levels), the model may need refinement to better capture these relationships. In order to examine each of these cases we computed the prediction errors in the test sample, and examine its distribution, shown in the following graph:

By defining outliers as observations with errors greater than ±2 standard deviations, the number of outliers identified was 259 observations (5,2% of the test set).

Characteristics of Outliers:

From the previous graph, we can see that the model's errors for the outliers are consistent with its errors for the full test set for the three variables shown (Age, Experience and Total Hours Worked), meaning the outliers are not extreme or unusual in terms of prediction errors. This implies that the model is consistent and behaves similarly for both the full test set and the outliers. The fact that the errors follow the same trend suggests that the model does not have a systematic bias (e.g., consistently overpredicting or underpredicting) for the outliers.

It also tells is of the nature of the outliers, specifically that the outliers may not be true anomalies but rather observations that are at the extremes of the data distribution, as is the case of outliers

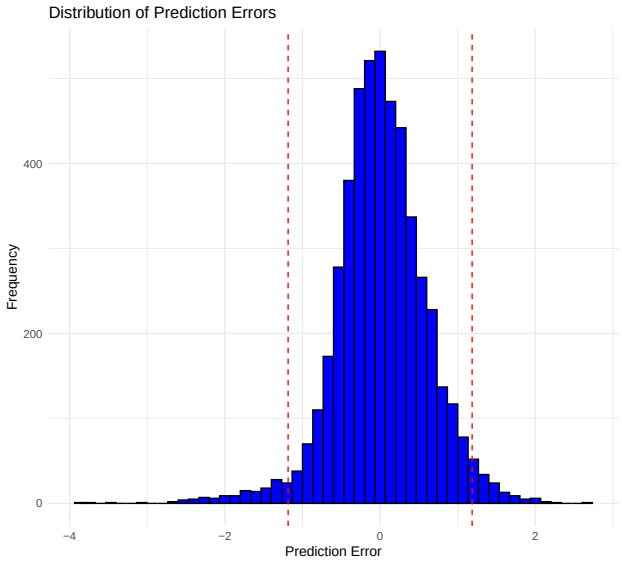

Figure 5: Distribution of prediction errors

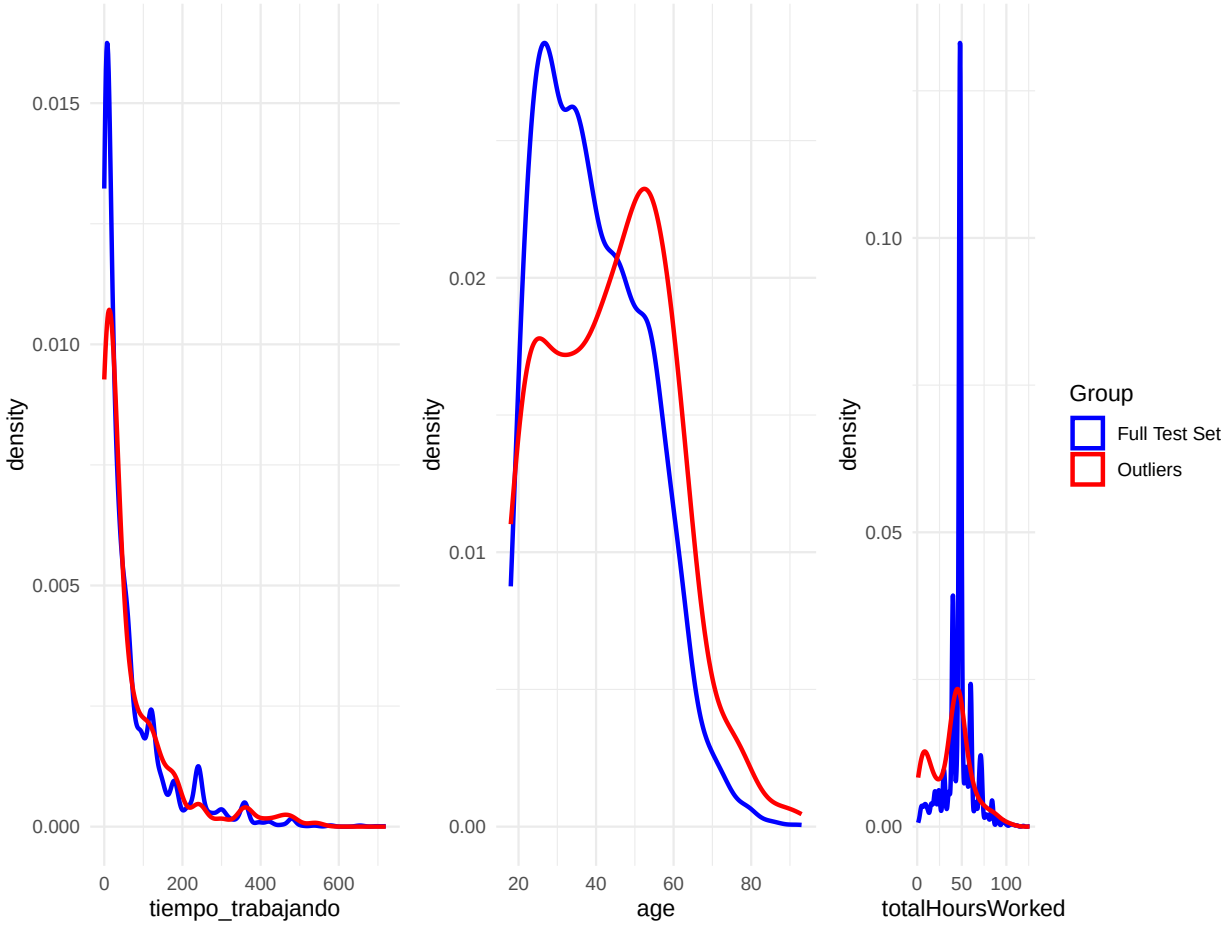

Figure 6: Comparison of income distribution: Full Test Set vs. Outliers

Table 6: Comparison of Model Performance

Metric	Model 1	Model 2
Test RMSE	0.5934259	0.5927824
Validation Set RMSE (LOOCV)	0.5634864	0.5575462
Number of Influential Observations	737	801

that may have extreme values in Total Hours Worked and Experience.

The model's inability to predict these observations accurately may be due to:

- Inherent variability: The outliers may represent natural variability in the data that the model cannot fully capture.

-Missing features: The model may lack important variables that could help explain the behavior of these outliers.

- Non-linear relationships: The outliers may follow non-linear patterns that the model (e.g., a linear model) cannot adequately capture.

Therefore, the DIAN should not necessarily look into investigating outliers with extreme values for potential data issues like the observations that miss the mark, and instead refine the model by adding other variables, interaction terms, or more flexible functional forms.

In terms of the comparison of the two models with the lowest RMSE, model 5 and 6 where selected and LOOCV was performed to further examine the model prediction. Model 6 has a slightly lower RMSE (0.5928 vs. 0.5934), indicating a marginal improvement in predictive performance on the test set. In terms of the validation set RMSE (using LOOCV), model 6 again performs better (0.5575 vs. 0.5635), suggesting that it generalizes slightly better to unseen data. The number of influential observations was calculated using Cooks distance. Model 6 has more influential observations (801 vs. 737), which might indicate that it is more sensitive to specific data points. However, the increase in influential observations suggests it might be more sensitive to outliers or specific data points, which could make it less stable. If robustness is a concern, model 6 being overly influenced by certain data points. Otherwise, if pure predictive accuracy is the goal, model 6 seems is the better choice.

References

Becker, G. S. (1964). *Human capital: A theoretical and empirical analysis, with special reference to education*. University of Chicago Press.

Calderon Diaz, M. A., Manrique Chaparro, O. L., & Chloe Jade Day, S. (2020). Informality and tax evasion: An experimental test among female entrepreneurs in bogota [Epub Oct 08, 2021]. *Semest. Econ.*, *23*(55), 239–257. https://doi.org/10.22395/seec.v23n55a11

Concha, T., Ramírez, J. C., & Acosta, O. L. (2017). *Tributación en colombia: Reformas, evasión y equidad* (Notas de estudio No. LC/TS.2017/137 and LC/BOG/TS.2017/1) (Distribución: Limitada). United Nations Economic Commission for Latin America and the Caribbean (CEPAL). Santiago. Retrieved October 15, 2023, from URL

DIAN. (2022). *Informe 2022 rendición de cuentas prospectiva 2023-2026*. https://www.dian.gov.co/entidad/Documents

Lazear, E. P. (1979). Why is there mandatory retirement? *Journal of Political Economy*, *87*(6), 1261–1284.

Mincer, J. (1974). *Schooling, experience, and earnings*. NBER.

Moller, L. C. (2012). Fiscal policy in colombia: Tapping its potential for a more equitable society. *World Bank Policy Research Working Paper*, (6092). Retrieved October 15, 2023, from https://ssrn.com/abstract=2085145

Murphy, K. M., & Welch, F. (1990). Empirical age-earnings profiles. *Journal of Labor Economics*, *8*(2), 202–229.