Big Data & Machine Learning
2025-01
Problem Set 1
Juan Pablo Grimaldos – 202122627 & Isabella Garzón – 202122524

I. Introduction

Precise income reporting is crucial for the accurate computation of fiscal policy instruments. Nonetheless, tax fraud persists in the social and economic landscape of most countries in the world. In Colombia, losses of tax revenue attributed to tax fraud accounts for 6% of the GDP each year on average (National Institute of Public Accountants, 2017). Thus, in this country, tax evasion poses a significant challenge to economic stability and social equity, as it undermines public revenue and weakens the ability for fiscal expenditure towards essential goods such as education, healthcare, and infrastructure.

Under-reporting is one of the most common types of tax fraud. It consists of filing a tax return form with a lesser tax base. In Colombia, this phenomenon has occurred massively in the past and persists to this day. Londoño & Avila (2019) used microdata from 1993 to 2016 linked with the leaked "Panama Papers" to prove that marginal percentage increases in the income tax rate produced decreases in reported wealth with a short-term elasticity of two. Thus, the authors find clear evidence of immediate bunching responses to wealth taxes, which implies that individuals lower their reported wealth to reduce their tax burden. This study also finds that a voluntary disclosure scheme between 2015 and 2017 encouraged evaders to disclose 1.7% of GDP in hidden wealth.

A more recent approach to detect tax fraud is through the use of machine learning techniques, which relies on income prediction models (De Roux et al., 2018). By predicting income, it is possible to flag cases of possible fraud and to identify vulnerable groups in the population. To accurately predict income using this approach, it is necessary to apply economic and numerical methods to data that is highly representative of the population parameters we want to estimate. This data should allow us to produce highly precise characterizations of individual profiles that are associated with a given level of income, and the mechanisms that explain the relationship between an individual's characteristics and their wage.

The objective of the present research paper is to produce these highly accurate characterizations to study the relationship between an individual's unobservable features and their hourly wage. To achieve this objective, the present study will construct models that estimate these relationships using data from the GEIH[1] for individuals in Bogotá in 2018. The resources, scripts, and results of this study can be consulted in the following GitHub repository.

Our results indicate that the application of machine learning techniques for income prediction can effectively identify potential groups of individuals whose reported earnings may warrant further investigation by institutions such as the DIAN to flag possible cases of underreporting. The predictive models exhibited high accuracy in estimating nominal hourly wages, leveraging various validation techniques, with the Leave-One-Out-Cross-Validation (LOOCV) method proving to be the most effective in minimizing predictive errors. However, the model with the best predictive performance systematically unpredicts higher income levels, which suggests that institutions such as DIAN should focus investigative efforts on individuals with high income level; as they may be more prone to underreporting earnings, potentially leading to missed tax revenue. The models demonstrated significant precision in predicting nominal hourly wages while accounting for individual and economic characteristics of the Bogotá population in 2018. The estimated "peak-age" at which individuals attain their highest nominal hourly wages is 36.44 years, with a 95% confidence interval ranging from 34.80 to 37.73 years. These results indicate that wage growth occurs predominantly at younger ages, driven by skill accumulation, and declines as work intensity decreases. Furthermore, the analysis of the gender wage gap revealed a persistent disparity in the nominal hourly wage between both sexes. Even after controlling for observable characteristics, women earn, on average, 3.9% less in nominal hourly wages than men. This finding suggests the existence of structural inequalities in the labor market that are not fully explained by differences in job characteristics or qualifications alone.

[1] Gran Encuesta Integrada de Hogares.

II. Data

$$\omega = f(X) + u \tag{1}$$

To correctly identify the data-generating process, it is necessary to construct a Model of individual hourly that can be represented as equation (1), where ω is the hourly wage, and X is a matrix that contains potential explanatory variables. Since the population $f(X)$ is unknown (or understood as a "black box"), our objective is to correctly identify X such that there is a parsimonious yet accurate estimation of the relationship between a characteristic individual and their expected income level given those characteristics.

To estimate this model, we will use data from the National Administrative Department of Statistics' (DANE) "M*edición de Pobreza Monetaria y Desigualdad"* 2018 report, which uses information from the GEIH. It provides information on household income and expenditure, employment status, and demographic characteristics (e.g. age, sex, household-headship status) to estimate monetary poverty and inequality. The data is structured at an individual level and is particularly useful for studying labor market dynamics, wage determinants, and economic disparities. Although the data is representative at national, regional and departmental levels, the study will focus on the data available for individuals in Bogotá.

The aforementioned data was obtained by web scraping from *https://ignaciomsarmiento.github.io/GEIH2018_sample/*. This website hosts a sample of data from the M*edición de Pobreza Monetaria y Desigualdad"* 2018 report, partitioned in 10 dataset chunks with equal dimensions. Each data chunk sources its information from another website. For this reason, when inspecting each data chunk's HTML source code, it is possible to identify the tag that performs the data's embedding[2]. This tag identifies the page(s) where the data is being sourced from. With this information, we utilized the *'rvest'* package in R, which allows for automatic extraction of HTML elements from web pages. Based on the similar characteristics of the tags in each chunk, a base URL was defined and then iterated over to construct each data-containing URL dynamically. This allowed us to retrieve each table and compile them in a single dataset. Since the dataset is publicly available on the website, no formal permissions are required to perform scrapping. Nonetheless, the present method is vulnerable to website structure changes, since major changes in the website's HTML structure can cause the script to fail.

Once the raw data was collected, various preprocessing steps were applied to ensure its usability for analysis. In particular, the analysis of this study focuses on employed individuals aged 18 or older. According to the DANE's (2018) methodological document for the GEIH's report, the variable '*ocu'* (employment status) indicates if an individual is either occupied (employed) or unoccupied (unemployed, economically inactive, or below the working age). Thus, the dataset was initially filtered by '*ocu'* to exclude unoccupied individuals and subsequently filtered to exclude individuals younger than 18.

Moreover, several key variable transformations were performed to improve the precision of the estimates using in the model and ensure their statistical interpretability. The variable that indicated the self-reported sex was recoded to a new binary variable ('*female'*) that indicates if the individual reported itself as a female or not. The dataset also contains nominal and real hourly income variables, both of which exhibit highly right-skewed distributions. The nominal hourly income variable is the predetermined result variable in all major models of this study. The objective of the inclusion of real wages was to remove the effect of inflation over the model estimations. These variables were log-transformed, which partially reduced skewness and made models that used them interpretable to the percentage change.

A significant portion of the dataset contains missing values, particularly in wage-related variables. These missing values were identified and summarized using the *'dplyr'* package in R. This identification process suggests that most missing values are concentrated in wage-related variables, particularly for informal workers. This pattern is expected, as informal workers often lack well-defined

[2] For instance, in the first data chunk, the tag is

salaries or structured payment periods. Given the right-skew in the distribution of log hourly wage variables, separate median imputations were performed for formal and informal workers.

Moreover, extreme values in the continuous variables can also distort the analysis. For this reason, in order to keep data from the data generating process we wish to estimate (and thus reducing the impact of these observations on our prediction error), winsorization caps outliers at the 97.5th percentile. This ensures that inconveniently high values do not distort the analysis. This was performed for all continuous variables except age.

Table 1. Descriptive statistics – continuous variables

Statistic	N	Mean	St. Dev.	Min	Max
Log. Nominal hourly wage	16,397	8.57	0.55	5.79	10.32
Log. Real hourly wage	16,397	8.48	0.53	5.02	10.20
Nominal hourly wage	16,397	6,437.12	5,528.65	326.67	30,303.02
Real hourly wage	16,397	5,819.77	4,917.24	151.91	26,956.21
Age	16,397	39.63	13.39	19	94
Weekly work hours	16,397	47.21	14.91	1.00	84.00

Note: This table presents the descriptivte statistics for the continuous variables used in this project.

Table 1 displays the descriptive statistics of the continuous variables selected for the present investigation. Two result variables will be utilized: log nominal hourly wages and log real hourly wages. Both variables have similar means but are distributed differently. The exponentiated wage variables reflect the actual reported income values of the sample. It is interesting to note how the mean hourly income for the nominal variable (COP $6,437.12) indicates that the individuals of the sample (if assumed to work the maximum legal 48 weekly work hours in 2018) would earn a monthly salary of approximately COP $1,235,904, almost twice the minimum monthly salary for the year in which the data was collected (Decree 2269 of 2017). Nonetheless, there is substantial heterogeneity in the sample; both exponentiated variables reflect standard deviations of COP $5,528.65 and COP $4,917.24, respectively. The effects of this heterogeneity over the distribution of our result variable are reduced when performing the logarithmic transformation, as seen in Table 1. The sample contains individuals from 19 to 94 years of age. The sample mean age is 39.63 with a standard deviation of 13.39, which indicates that, on average, working individuals in the sample are below the pension age. The weekly work hours variable has a mean of 47.21, suggesting that most individuals in the sample worked the legal weekly work hour limit for 2018. Nonetheless, the range and standard error of this variable suggests possible diverse employment conditions.

The categorical variables used in this study include sex, female household head status (which indicates if the individual is a female household head), education level (incomplete primary, complete primary, incomplete secondary, complete secondary, tertiary, or none), formality status of employment, firm size in which the individual works in (2-5 workers, 6-10 workers, 11-50 workers, >50 workers) and employment sector (private sector, public sector, domestic, self-employed, employer, family worker, worker in other's businesses, day laborer, other). The sample is relatively balanced in terms of sex, with 52.95% male and 47.05% female participation. Notably, only 15.81% of the individuals in the sample are female household heads, which reflects potential gender disparities in economic leadership within the households. Individuals in the sample are also relatively educated, with 74% having completed at least secondary education and 42.18% attaining tertiary education. However, despite this high level of educational attainment, informal employment remains significant, comprising 40.99% of the sample. Heterogeneity in the labor market is represented in the sample, with significant self-employment (24.55%) and microenterprise work (20.14%), while 36.07% are in large firms. Employment takes place mostly in the private sector (56.31%), with 30.90% self-employed and only 3.84% in the public sector. These categorical variables reflect crucial demographic and employment characteristics that can have a high explicative power over the underlying mechanisms that associate an individual with a certain level of income.

III. Age-Wage Profile

The relationship between an individual's age and their corresponding wage has been thoroughly discussed in economic literature. Authors like Li (2023) suggest a positive relationship between age and nominal hourly wage, while the vast remaining majority insists that the relationship between both variables follows an "inverse-U" path. To correctly construct individual characterization profiles that will yield accurate income predictions, we estimated and compared Model (2), (3), (4), (5).

$$Log(\omega^N) = \beta_1 + \beta_2 Age + u \quad (2)$$

$$Log(\omega^N) = \gamma_1 + \gamma_2 Age + \gamma_3 Age^2 + \varepsilon \quad (3)$$

$$Log(\omega^R) = \beta_1 + \beta_2 Age + u \quad (4)$$

$$Log(\omega^R) = \gamma_1 + \gamma_2 Age + \gamma_3 Age^2 + \varepsilon \quad (5)$$

Where ω^N denotes the nominal hourly wage, and ω^R denotes the real hourly wage. The results of estimating Models (2) through (5) are displayed on Table 2.

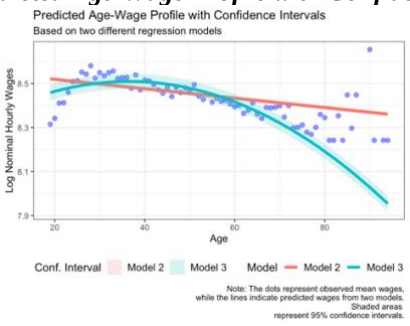
Table 2. Regression Results: Nominal and Real Income

	Dependent variable:			
	Log Nominal Hourly Wage		Log Real Hourly Wage	
	Model 2	Model 3	Model 4	Model 5
	(1)	(2)	(3)	(4)
Age	-0.002***	0.012***	-0.002***	0.009***
	(0.0002)	(0.001)	(0.0002)	(0.001)
Age Squared		-0.0002***		-0.0001***
		(0.00001)		(0.00001)
Constant	8.562***	8.287***	8.451***	8.247***
	(0.008)	(0.024)	(0.008)	(0.022)
Observations	14,993	14,993	14,993	14,993
R^2	0.007	0.017	0.005	0.011
Adjusted R^2	0.007	0.017	0.005	0.011

Note: *p<0.1; **p<0.05; ***p<0.01

This table presents the estimation results for nominal and real income models, considering age and its quadratic term. Standard errors are in parentheses.

Figure 1. Predicted Age-Wage Profile with Confidence Intervals



Preliminarily, Models (2), (3), (4), and (5) were ran to examine non-studentized model fits. After thoroughly inspecting the residuals for these models, it was found that significant observations had an error and leverage profile that, according to Greene's (2003) interpretation of the student criteria, would classify them as an outlier in these models. These observations (a total of 1,404) were removed from

the sample. This was only performed once in this research, as further removal of observations would significantly reduce sample size and increase model testing variance.

The results shown in Table 2 suggest that the relationship between the logarithmic hourly wage (both nominal and real) and age is non-linear. When incorporating a polynomial age term to Models (2) and (4) on Models (3) and (5), the model's sample fit increases substantially; the additional squared variable led reduced the model's residual sum of squares (RSS), which in turn, fosters a notably higher R-squared and thus lower sample bias (reducible error) for Models (3) and (5) relative to Models (2) and (4). In the linear models (Model 2 and Model 4), age has a negative statistically significant coefficient at the 1% level, suggesting that income decreases with age when a quadratic relationship is not considered, which directly contradicts Li (2023). Nonetheless, in the models including the quadratic term (Model 3 and Model 5), the relationship changes: the age coefficient is positive and significant (0.012 in Model 3 and 0.009 in Model 5), while the quadratic term is negative and significant (-0.0002 and -0.0001, respectively). This suggests that the relationship between age and income follows an inverted U-shape where income increases up to a point and then declines with further increases in age. Although the model fit increases when the quadratic term is added, it is still considerably low across Model (3)'s and Model (5)'s specifications (0.017 and 0.011, respectively). This suggests that there are other factors (variables) that also play a crucial role in determining income. Furthermore, it is interesting to note that models that used Log(ω^N) as their dependent variable had a higher goodness of sample fit to those that used Log(ω^R).

Figure 1 displays the predicted age-income profile considering the nominal hourly variable models (Model 2 and Model 3), and their 95% confidence intervals. In this graph, it is clear that the quadratic model better captures the curvature of the age-income relationship in the sample, compared to the linear model, which underestimates income at middle ages and overestimates it at older ages. In contrast, Model (3) seems to underestimate income only at older ages. These results are consistent with economic theory; income tends to rise with experience until it peaks, after which it may decline due to factors such as skill obsolence or reduced work intensity.

The information of these models could be useful to predict, at which age, on average, does an individual maximize their hourly income. Since the relationship between age and hourly income is quadratic and best explained when using nominal hourly income, the predicted "peak-age" can be plausibly estimated by optimizing Model 3 with respect to age. This would yield equation (3.1).

$$Peak - age = \frac{-\gamma_1}{2\gamma_2} \tag{3.1}$$

Using the coefficients obtained from running Model (3), the predicted "peak-age" for the individuals in our sample is approximately 36.44 years. According to this model, this would be the age at which, on average, an individual is expected to maximize their nominal hourly wage. This is consistent with the information in Figure 1, which shows that the predicted age values peak around an individual's late 30s. However, to understand the variability of this statistic, it is necessary to estimate its distribution using the empirical distribution of the data in the sample. To do this, a bootstrap estimation was performed with $B = 1000$ replications. The bootstrap consisted of calculating Model (3)'s coefficients and then estimating equation (3.1) using data from a sample of size 14,993 with replacement. The procedure yielded a bootstrap mean of 36.39, 95% confidence interval (CI) [34.80, 37.73] with standard error 0.77. This confidence interval suggests that we can be 95% confident based on this sample that the population "peak-age" is between 34.80 and 37.73 years of age. On average, we expect the population "peak-age" to be within 1.96 standard deviations of the bootstrap mean, 36.39. The bootstrap estimates are consistent with the OLS estimate, with a difference of approximately 0.05 years. The bootstrap error has been affected by a mild asymmetry and bias. This is expected to be minimized as sample size increases.

The estimated "peak-age" of approximately 36.44 years suggests that, on average, workers reach their highest nominal hourly salary in their late 30s. The bootstrap confidence interval reinforces the robustness of this estimate. These results highlight the previously noted relationship between nominal hourly wages and age, possibly explaining a declining labor mobility with age after key moments for career growth, salary negotiations, and workforce planning.

IV. The Gender Earnings GAP

$$Log(\omega^N) = \beta_1 + \beta_2 Female + u \quad (6)$$

$$Log(\omega^N) = \gamma_1 + \gamma_2 Female + \gamma_3 X + \varepsilon \quad (7)$$

$X = [Age, Age^2, Employment\ Sector, Female\ Household\ Head, Weekly\ Hours\ Worked, Formality\ Status, Max.\ Education\ Level]$

Another potentially related that could explain the relationship between an individual's characteristics and thus contribute to identify the data generating process is an individual's sex. Policymakers have been concerned with a possible wage gap between individuals based on their sex. To initially explore this relationship, Model (6) was estimated. Model (6) contemplates a sex variable that indicates if the individual in the sample is a female or not. Attempting to explore this relationship more accurately, Model (7) was also calculated. This model incorporates a series of controls included in matrix X to account for individual and job characteristics that may influence income. Specifically, these controls employment sector, whether the individual is the female head of household, total weekly hours worked, t employment formality status, the firm size of the company in which the observation works, and the maximum education level attained by the individual. The employment sector matters because industries have different wage structures, based on their specific characteristics. Female household head status, could show the impact in earnings, considering that women may have additional constraints on time and work flexibility. Weekly hours worked help distinguish wage differences due to labor effort. Employment formality is relevant since formal jobs offer higher wages due to job stability and labor benefits, additionally, women tend to be overrepresented in the informal sector and this could influence the gender gap. In relation with the employment sector, firm size is relevant because larger firms tend to pay higher wages. Finally, maximum education level attained, is a key determinant of wages, as it reflects differences in skills and qualifications.

Model (6) will represent the unconditional model of our analysis, and Model (7) will represent our conditional model. Model (7) will also be estimated using Frisch-Waugh-Lovell (FWL) decomposition. The conditional FWL model's sex will also be estimated using bootstrap to gain insights on the variability of the statistic, which, in this case, represents the mean percentage difference in the nominal hourly salaries between women and men.

Table 3. Regression Results: Wage Gap Models

	Dependent variable:		
	Log Nominal Hourly Wage		Residualized Log Nominal Hourly Wage
	Model 6	Model 7	Model 7 FWL
	(1)	(2)	(3)
Female	**-0.009***	**-0.039*****	
	(0.006)	(0.005)	
Female FWL			**-0.039*****
			(0.005)
Constant	8.482***	7.814***	-0.000
	(0.004)	(0.024)	(0.002)
Observations	14,993	14,993	14,993
R^2	0.0002	0.421	0.004
Adjusted R^2	0.0001	0.420	0.004
Residual Std. Error	0.338 (df = 14991)	0.257 (df = 14983)	0.257 (df = 14991)
F Statistic	2.942* (df = 1; 14991)	1,209.138*** (df = 9; 14983)	63.691*** (df = 1; 14991)
Note:			*p<0.1; **p<0.05; ***p<0.01

Model 7 includes additional controls: Age, Employment Sector, Female Household Head, Weekly Hours Worked, Formality, Firm Size, and Education. FWL refers to the Frisch-Waugh-Lovell decomposition. Bootstrap models account for resampling variability.

Table 3 displays the regression estimation results output for Model (6) and (7), also including Model (7)'s FWL estimation, which essential yields the same coefficient. The FWL approach to Model (7) removes the variability in the result variable explained by the control variables in matrix X. The equivalence in coefficients is thus proof of the relationship between FWL and OLS estimators when "controlling" for the effect of covariates. When estimating the wage gap using FWL and bootstrapping using $B = 1000$ replications, the mean coefficient obtained was -0.0391, 95% CI [-0.0483, -0.0300], and standard error 0.0048. Both the OLS and the FWL variations of Model (7) are included within the bootstrap CI, assuring the robustness of the estimations. The CI implies that we can be 95% confident, based on the data of this sample, that the true population age gap is between 4.83% and 3%. The sample error suggests that the predictions for the wage gap using this model will, on average, either underestimate or overestimate the conditional wage gap by 0.48 percentage points. This standard error is marginally lower than the standard error reported by the Model (7) FWL variation without bootstrap (0.005).

The estimated coefficient for the sex variable in Model (6) is -0.009, statistically significant at the 10% level. Moreover, the coefficient for female becomes -0.039, and its statistical significance increase considerable to the 1% level after controlling for covariates in Model (7). This increase in magnitude and statistical significance suggests that observable individual characteristics (such as education, job formality, and industry sector) have an influence on the estimated gender wage gap. In the unconditional model suggests that, on average, women earn 0.9% less than men in nominal hourly wages, ceteris paribus. Furthermore, in the conditional model, after controlling for relevant covariates, the estimated wage gap in hourly nominal salaries increases up to 3.9% on average, ceteris paribus. This indicates that there is a larger unexplained component of the wage gap.

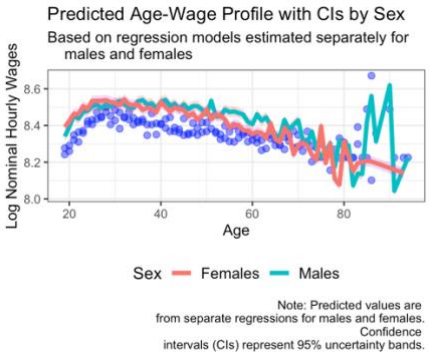
This increase in the estimated wage gap in model (7) suggests that women with similar observed characteristics to men still experience lower wages, which implies the existence of structural inequalities in the labor market. The conditioning on the covariates in matrix X influenced the model's sample fit significantly; the R-squared increase from 0.0002 in Model (6) to 0.421 in Model (7). This indicates that, with the incorporation of the additional characterization variables to the model, it was possible to reduce the sample bias of the model, which implies more accurate predictions. Furthermore, the increase in the R-squared suggests that a substantial portion of wage variation is explained by the control variables, but the persistent and significant negative coefficient for the sex variable (Female) suggests that gender-related factors are still at play.

Therefore, is it possible to argue that the gender wage gap could result from factors such as selection bias, discrimination, or a combination of both? If women in the sample were to be selecting themselves into lower-paying jobs or industries, the estimated wage gap could partially reflect a self-selection effect rather than outright wage discrimination. Nonetheless, since Model (7) controls for sector of employment, formality, and weekly hours worked, the fact that the estimated coefficient becomes more negative suggests that selection bias alone is not able to fully explain the wage gap. The increase in the magnitude of the coefficient reflects a structural effect in wages, where women are systematically paid less than men (even when controlling for observable characteristics). Therefore, it is most likely that the gender wage gap results from a combination of both selection and discrimination mechanisms.

$$Log(\omega^N) = \gamma_1 + \gamma_2 Female + \gamma_3 X + \varepsilon \mid Female_i = 0 \quad (7.1)$$

$$Log(\omega^N) = \gamma_1 + \gamma_2 Female + \gamma_3 X + \varepsilon \mid Female_i = 1 \quad (7.2)$$

Figure 2. Predicted Age-Wage Profile with Confidence Intervals by Sex



Considering the existence of a potential sex-wage gap motivated by discriminatory and selection bias, it is pertinent to now construct a predicted age-wage profile with the information in the sample in order to examine the relationship between this gap and the individual's age with their nominal hourly wage. To estimate these peak-ages by sex, the models (7.1) and (7.2) were estimated using FWL decomposition and their age coefficients were used to compute equation (3.2) for $B = 1000$ replications using bootstrapping. Model (7.1) is Model (7) but calculated only for the subset of individuals in the sample who identify as male. Model (7.2) is Model (7) but calculated only for the subset of individuals in the sample who identify as females.

The predicted age-wage profile segregated by sex (in this case, by Model 7.1 and 7.2) can be observed in Figure 2. Both men and women exhibit an inverted U-shaped wage trajectory, consistent with the findings on section III. For both groups, wages seem to rise early in the career due to experience accumulation, peak at some point, and then decline slightly. The wage peak appears to be slightly later for women. This suggests that there may exist delayed earnings growth effects, attributed to factors like sectoral differences, selection, or discrimination (as discussed in section III). According to Figure 2, it is between ages 35-60 where the wage gap becomes more pronounced between men and women: reflecting possible glass ceiling effects. Post-retirement age data (65+) is very noisy, with male wages showing high volatility even after controlling for outliers twice during this whole investigation. This may suggest that the model fit may have a higher non-sample variance. Although graphically having low bias, the model's non-reducible error can hinder its predictive performance in other scenarios. The wage gap seems to persist until late adulthood (70-80 years of age), where differences are less pronounced, but data starts to become extremely variable. Although there appears to be a wage gap between both sexes, the CIs on the graph for men and women overlap substantially, which may suggest that gendered peak-age differences might not be statistically significant. This hypothesis is reinforced when analyzing the bootstrap distribution values for the peak-age estimation by sex. Men's mean peak age is estimated at 50.76 years, 95% CI [49.11, 52.48] and standard error 0.88. Women's mean peak age is estimated at 52.29 years, with a considerably larger standard error of 1.95, and a wider 95% CI of [49.03, 56.66]. The CIs for men and women have a substantial overlap, which also suggests that the estimated peak ages are not significantly different at the confidence levels. Furthermore, the difference between the mean peak ages of men and women (52.29 – 50.76 = 1.53 years) is small relative to the standard error of the female estimate (1.95); implying that the difference between peak ages may not be statistically robust.

V. Predicting Earnings

$$\underset{M = [Age, Age^2, Age^3, Female]}{Log(\omega^N) = \gamma_0 + \gamma_1 M + \varepsilon} \tag{8}$$

$$\underset{Z = [Age, Age^2, Age^3, maxEducLevel, maxEducLevel^2, WeeklyHoursWorked, WeeklyHoursWorked^2, Female, EmploymentSector, Formal, SizeFirm]}{Log(\omega^N) = \gamma_0 + \gamma_1 Z + \varepsilon} \tag{9}$$

$$Log(\omega^N) = \gamma_0 + \gamma_1 K + \varepsilon \quad (10)$$
$$K = [Age, Age^2, (Age \cdot Female), EmploymentSector, WeeklyHoursWorked, \gamma_7 maxEducLevel]$$

$$Log(\omega^N) = \gamma_0 + \gamma_1 S + \varepsilon \quad (11)$$
$$S = [Age, Age^2, Female, (Formal \cdot Female), EmploymentSector, WeeklyHoursWorked, \gamma_7 maxEducLevel, Formal]$$

$$Log(\omega^N) = \gamma_0 + \gamma_1 W + \varepsilon \quad (12)$$
$$W = [Age, Age^2, Female, (Formal \cdot Female), maxEducLevel, Formal, HeadFemale, (maxEducLevel \cdot Female), (Formal \cdot WeeklyHoursWorked), (Age \cdot HeadFemale), (EmploymentSector \cdot WeeklyHoursWorked)]$$

Following the inference-based estimations previously performed, this section will evaluate the predictive power of different earnings models. To do so, the sample is split into a training set (70%) and a testing set (30%). A seed was implemented to ensure reproducibility of the results. To explore the predictive performance, Models 1 to 7 were complemented with additional models (See Models 8 to 12), which build upon the variables contained in the previously defined matrix X. The complementary Models (8) to (12) proposed capture non-linear relationships between the independent variables and the result variable and also introduce interaction terms between independent variables. By incorporating polynomial terms and interactions effects, the models aim to potentially reduce prediction errors and enhance model performance.

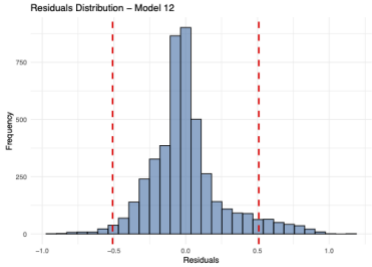
Table 4. Predictive Performance Results: RMSE

Model	RMSE
Model 12	0.254
Model 9	0.255
Model 7	0.258
Model 11	0.261
Model 10	0.304
Model 4	0.314
Model 3	0.325
Model 6	0.327
Model 8	0.339
Model 2	0.342
Model 5	0.353

Best Model: Model 12 | Lowest RMSE: 0.2536

Both sets of models (the linear and non-linear estimations) were considered when using *'caret'* package in R, to report and compare the predictive performances in terms of the RMSE. The results in Table 4 show that Model 12 achieves the lowest RMSE (0.254), making it the most accurate for predicting earnings, this model integrates multiple interactions, including between education and gender, formality and hours worked, age and female household headship. This result suggests that earnings are shaped by complex interdependencies rather than simple linear effects. Also, models with a low RMSE, such as Model 9 and 7 perform well. In contrast, simpler models like Model 2 and 5, have a higher RMSE values, 0.353 and 0.342 respectively, highlighting the limitation of basic specifications.

Figure 3. Distribution of Errors in the Model With Best Performance



To further assess the performance of Model 12, prediction errors were computed, and its distribution was examined. In figure 3, the graph displays the residual distribution for this model, serving as a diagnostic tool to evaluate the predictive accuracy and identify potential outliers. The distribution appears approximately normal, with most residuals concentrated around zero, suggesting that the model performs well in predicting income. The dashed red lines mark a threshold of ±0.51, corresponding to two standard deviations from the mean. Residuals that fall beyond this range are classified as potential outliers, as they deviate significantly from the expected values. Note that most outliers are concentrated in the right tail of the distribution. This is almost a "mirror" of the distribution of the log nominal hourly income values. This is extremely insightful, since it indicates that the prediction errors "mirror" the asymmetry in the income data. In economic terms, this suggests that there are more unexpected discrepancies in observations with higher levels of income. This implies that our best fitting model (Model 12) consistently mispredicts higher levels of income. This would constitute an "irreducible" error term in the estimation of these models, since Model 12 is one of the most flexible models of all our research. This suggests that, even by making the model more complex and flexible, there is a specific shortcoming of these models to predict high levels of income.

From a policy perspective, this suggests that the DIAN should carefully investigate people with high income levels. This is because, if our models underpredict these incomes, the DIAN could be missing out on potential tax revenue. This is highly consistent with the literature and antecedents mentioned in the introduction. Although some discrepancies may come from model limitations (such as omitted variables or measurement errors), this analysis yields systematic patterns in the residuals which may highlight the correct group where underreporting of earnings may become more prevalent.

Table 5. Comparison LOOCV and Validation Set Approach

Model	RMSE
Model 12: Validation set approach	0.254
Model 12: LOOCV	0.253
Model 12: Leverage	0.253
Model 9: Validation set approach	0.255
Model 9: LOOCV	0.254
Model 9: Leverage	0.254

Note: This table contains the LOOCV estimation for the two models with the lowest prediction errors, comparing with the validation set approach.

$$\widehat{\gamma_1}^{-j} = \widehat{\gamma_1} - \frac{1}{1-h_j}(K'K)^{-1}K_j\widehat{\varepsilon_j} \tag{13}$$

Table 5 compares the Root Mean Squared Error (RMSE) for two model, specifically Model 12 and 9, the ones with the lowest predictive errors. This is done using different validation approaches: Validation Set Approach (0.254), LOOCV (0.253) and Leverage (0.253). Model 12 consistently outperforms Model 9 across all methods, indicating better predictive accuracy. Among the validation techniques, LOOCV and Leverage yields the lowest RMSE, suggesting they provide a more reliable estimate of prediction error by reducing bias. In contrast, the Validation Set Approach shows the highest RMSE. These results are consistent with the previous residual analysis, where Model 12 shows a lower testing bias. The differences in RMSE across validation methods also highlight the potential influence of outliers, as LOOCV mitigates their impacts by evaluating each observation individually. Note that the LOOCV and Leverage approaches yield the same RMSE. This is because there is an extremely strong relationship between both estimation methods. Equation (13) shows how the Leverage computation method allows us to calculate the prediction error for the outcome by using a model estimated without the *j-th* observation. This is exactly what the LOOCV does for a particular observation *j*. Both the Leverage and LOOCV measures are useful to assess the predictive performance of a linear regression model. Moreover, the Leverage measure allowed us to compute this predictive performance without the cross-validation approach that implied running the constructed model N times. Overall, Model 12 is the most robust and accurate for income prediction.

Bibliography

Departamento Administrativo Nacional de Estadística. (2018). Gran Encuesta Integrada de Hogares - GEIH - 2018. Retrieved from https://microdatos.dane.gov.co/index.php/catalog/547

De Roux, D., Perez, B., Moreno, A., Villamil, M. D. P., & Figueroa, C. (2018, July). Tax fraud detection for under-reporting declarations using an unsupervised machine learning approach. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 215-222).

Greene, W. H. (2003). ECONOMETRIC ANALYSIS.

Li, M. (2023). Age and hourly wage: How aging affects earning. Advances in Economics, Management and Political Sciences, 3, 413-422. https://www.researchgate.net/publication/370036076_Age_and_Hourly_Wage_How_Aging_affect_Earning

Londoño-Vélez, J., & Avila-Mahecha, J. (2019). Can Wealth Taxation Work in Developing Countries? Quasi-Experimental Evidence from Colombia.

National Institute of Public Accountants of Colombia. (2017, July 13). Colombia loses the equivalent of more than 6% of GDP due to tax evasion. Retrieved from https://incp.org.co/publicaciones/infoincp-publicaciones/impuestos/nacionales/2017/07/colombia-pierde-equivalente-mas-del-6-del-pib-evasion/