Understanding and Predicting Wages: Evidence from Bogotá

Juan José Rojas, Francisco Soler, Jesús Yancy & Juan Otalora

1 Introduction

Colombia has long struggled with insufficient tax revenue: the country consistently collects less and spends more than what its income level would predict (Fergusson & Hofstetter, 2022). A significant portion of this gap stems from tax evasion, estimated at 8.6% of GDP, of which 1.1% corresponds to evasion in personal income taxes (Lora & Mejía, 2019). Among these, labor income underreporting –though relatively modest– still accounts for nearly 1.2 trillion Colombian pesos in lost revenue annually[1], equivalent to about twice the GDP of Guainía and more than double that of Vaupés (DANE, 2025). This paper addresses two main objectives: first, to explore the determinants of labor earnings by examining how individual characteristics–particularly age and gender–shape wage outcomes; and second, to develop predictive models that can identify potential underreporting in tax filings, contributing to the broader goal of reducing evasion.

Our analysis draws on data from the 2018 GEIH survey for Bogotá. From an initial sample of 32,177 individuals, we restrict our focus to salaried workers aged 18 and over, excluding the self-employed and unpaid family workers. This yields a final study sample of 9,892 individuals. We concentrate on wages rather than total income, as salaries tend to be more stable and more strongly correlated with observable individual characteristics. This focus serves both our analytical and predictive goals: it enables precise identification of wage determinants while also providing a robust foundation for models aimed at detecting underreporting among formal employees.

The first set of our empirical results highlights two key patterns. First, the estimated age-wage profile follows a concave, parabolic shape with a peak around age 45, consistent

[1]This figure is obtained by applying the 0.07% of GDP that Lora and Mejía (2019) estimate as efficiently recoverable from labor income underreporting to the official 2024 GDP estimate for Colombia (COP 1,706.4 trillion) published by DANE (2025)

with standard human capital theory. Second, the examination of the gender-earnings gap reveals a more nuanced picture than the 4.7% difference of the unconditional model initially suggests. When controlling for individual characteristics such as education and age, the gap widens to 13%, suggesting that the initial figure understates the disparity–partly because women in the sample are, on average, slightly more educated. Adding controls for firm characteristics (formality, sector, size) reduces the gap to 10%, and further incorporating occupation fixed effects lowers it to 8.4%. Even so, this final gap remains nearly twice the raw estimate, reflecting a mix of selection and discrimination effects that continues to shape gendered labor market outcomes.

[Paragraph 4: Results of the prediction models]

[Paragraph 5: Conclusions and organization of the document]

2 Data

To conduct our academic exercise, we relied on a dataset[2] assembled by Professor Manuel Fernández and made accessible via the website of Professor Ignacio Sarmiento, both faculty members at the Department of Economics at Universidad de los Andes. As data tables were not fully embedded in each page HTML and instead loaded dynamically, we implemented a scraping routine that automatically opened a browser, allowed the table to render, and extracted the content iteratively across its ten pages. In total, we gathered 32,177 observations corresponding to all individuals surveyed in Bogotá as part of the 2018 GEIH (*Gran Encuesta Integrada de Hogares*). The GEIH, conducted by DANE since 2006, is the country's principal source of labor market statistics, covering employment, earnings, demographic characteristics, and income sources across national, regional, departmental, and capital-city levels. Its core is the national labor market survey, and the 2018 edition represents the latest methodological framework available.

Our first filtering criterion was to restrict the sample to individuals aged 18 and above who report earning a salary, i.e., excluding the self-employed and unpaid family workers. We focus specifically on labor income, as opposed to more volatile entrepreneurial or informal gain, because labor earnings are more stable and closely tied to individual characteristics like education and experience. In contrast, earnings from small enterprises are

[2]Publicly available at: https://ignaciosarmiento.github.io/GEIH2018_sample/

influenced by external factors such as market conditions and business fluctuations. After applying these filters, our final analytic sample consisted of approximately 9,800 individuals, whose characteristicsâage, education, earnings, and working hoursâare the foundation for our exploratory analysis.

Describe the data briefly, including its purpose, and any other relevant information.

We will use data for BogotÃ¡ from the 2018 MediciÃ³*ndePobrezaMonetariayDesigualdadReport"th*

The focus of this problem set is on employed individuals older than eighteen (18) years old. Restrict the data to these individuals and perform a descriptive analysis of the variables used in the problem set. Keep in mind that in the data, there are many observations with missing data or 0 wages. I leave it to you to find a way to handle these data.

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

Describe the data cleaning process and

Descriptive the variables included in your analysis. At a minimum, you should include a descriptive statistics table with its interpretation. However, I expect a deep analysis that helps the reader understand the data, its variation, and the justification for your data choices. Use your professional knowledge to add value to this section. Do not present it as a "dry" list of ingredients.

3 Age-wage profile

To better understand the determinants of wages, we examine the age-wage profile, which captures how average earnings vary with an individual's age (Startz, 2014). There is a long tradition in labor economics of describing this profile as an inverted U-shape across most labor markets and occupations, reflecting the tendency for wages to rise with age until peaking–typically around the early 50s–before declining as workers approach retirement. This age-wage pattern is a core input in many labor market models.

In this subsection we are going to estimate the *Age-wage profile* for the individuals in this sample:

$$\log(w) = \beta_1 + \beta_2 \,\text{Age} + \beta_3 \,\text{Age}^2 + u \tag{1}$$

Table 1. Age–Wage regression (dependent variable: log(hourly wage))

	Dependent variable:
	Log_Total_hour_salary
age	0.069^{***}
	(0.004)
age2	-0.001^{***}
	(0.00004)
Constant	7.339^{***}
	(0.068)
Observations	9,700
R^2	0.047
Adjusted R^2	0.047
Residual Std. Error	0.705 (df = 9697)
F Statistic	239.598^{***} (df = 2; 9697)
Note:	*p<0.1; **p<0.05; ***p<0.01

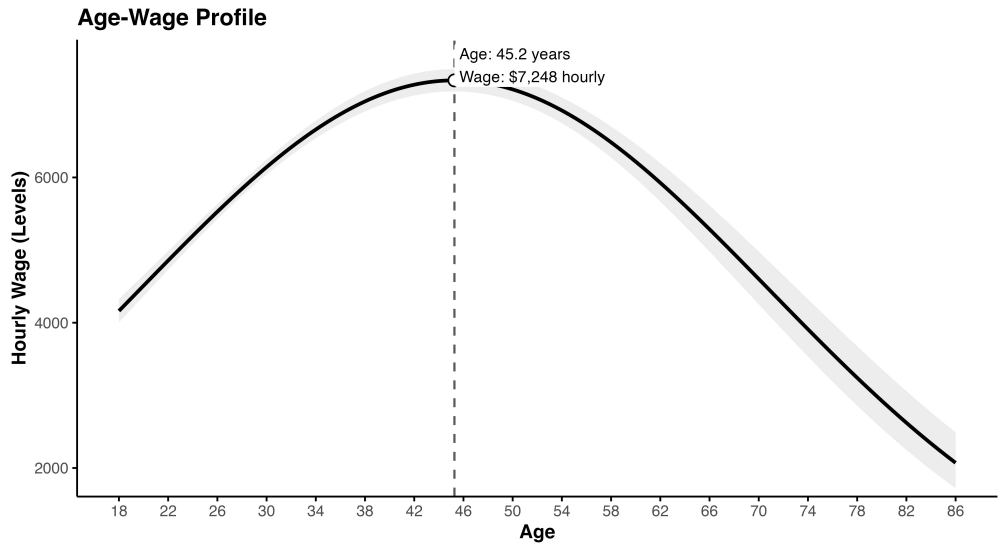
3.0.0.1 An interpretation of the coefficients and its significance. Both age terms are precisely estimated with the expected signs: $\widehat{\beta}_2 > 0$ and $\widehat{\beta}_3 < 0$, implying a concave life-cycle profile.

3.0.0.2 A discussion of the model's in-sample fit. As expected for a univariate cross-section, the model explains a modest share of dispersion ($R^2 = 0.047$; residual s.e. = 0.705 log points). Most variation reflects education, occupation/industry, region and firm/job heterogeneity not modeled here. The goal is to transparently characterize the *shape* and the turning point of the profile within the observed age support.

3.0.0.3 A plot of the estimated age-earnings profile and the "peak age" (with bootstrap CIs). Under equation, the profile attains its maximum at

$$a^* = -\frac{\widehat{\beta}_2}{2\,\widehat{\beta}_3} = \mathbf{45.2} \text{ years.}$$

Figure 1. Estimated age–earnings profile with 95% pointwise bootstrap band. The dashed line marks the peak age; the dot annotates the peak wage.



At the peak, the fitted hourly wage level is approximately \$7,248 (COP per hour), matching the annotation in Figure 1. To quantify uncertainty, we implement a nonparametric pairs bootstrap resampling individuals ($B = 1{,}000$). The bootstrap distribution of a^* yields a standard error of 0.672 years, negligible bias (-0.027), and a **95% percentile confidence interval** of [**44.07, 46.65**]. Near the top the curve is very flat, so these bounds mainly affect the *location* of the peak rather than its *height*.

4 The gender earnings GAP

Policymakers have long been concerned with the gender wage gap, and is going to be our focus in this subsection.

(a) Begin by estimating and discussing the unconditional wage gap:

$$\log(w) = \beta_1 + \beta_2 \,\text{Female} + u \tag{2}$$

where *Female* is an indicator that takes one if the individual in the sample is identified as female.

(b) *Equal Pay for Equal Work?* A common slogan is "equal pay for equal work". One way to interpret this is that for employees with similar worker and job characteristics,

no gender wage gap should exist. Estimate a conditional earnings gap incorporating control variables such as similar worker and job characteristics. That is estimate an equation of the form

$$\log(w) = \beta_1 + \beta_2 \,\text{Female} + \theta X + u \tag{3}$$

where X is the vector of worker and job characteristics. Think deeply on the role of controls. When estimating the equation do it:

i. First, using FWL

ii. Second, using FWL with bootstrap. Compare the estimates and the standard errors.

(c) Next, plot the predicted age-wage profile and estimate the implied "peak ages" with the respective confidence intervals by gender.

When presenting and discussing your results, include:

- An estimating equation, explaining the included control variables (beware of "bad controls").
- A regression table, with the estimates side by side of the conditional and unconditional wage gaps, highlighting the coefficient of interest. Controls, should not be included in the table but dutifully noted.[3]
- An interpretation of the "Female" coefficients, a comparison between the models, and the in-sample fit.
- A discussion about the implied peak ages and their statistical similarity/difference.
- A thoughtful discussion about the unconditional and conditional wage gap, seeking to answer if the changes in the coefficient are evidence of a selection problem, a "discrimination problem," a mix, or none of these issues.

5 Predicting wages

In the previous sections, you estimated some specifications with inference in mind. In this subsection, we will evaluate the predictive power of these specifications.

[3]Tip: Look how applied papers construct their results tables. These papers usually present comparable results in the same table with coefficients side by side, which helps the reader follow the discussion.

(a) Split the sample into two: a training (70%) and a testing (30%) sample. (Use the seed 10101 to achieve reproducibility.)

(b) Report and compare the predictive performance in terms of the RMSE of all the previous specifications with at least five (5) additional specifications that explore non-linearities and complexity.

(c) In your discussion of the results, comment:

   i. About the overall performance of the models.

   ii. About the specification with the lowest prediction error.

   iii. For the specification with the lowest prediction error, explore those observations that seem to "miss the mark." To do so, compute the prediction errors in the test sample, and examine its distribution. Are there any observations in the tails of the prediction error distribution? Are these outliers potential people that the DIAN should look into, or are they just the product of a flawed model?

(d) **LOOCV.** For the two models with the lowest predictive error in the previous section, calculate the predictive error using Leave-one-out-cross-validation (LOOCV). Compare the results of the test error with those obtained with the hold-out set approach and explore the potential wins like the influence statistics. (*Note: Some implementations and computations within this subsection take time, depending on your coding skills, plan accordingly!*)

6 Additional Guidelines

- **Document.** Submit a .pdf document in Brightspace under the activity 'Problem Set 1: Predicting Income'.

- **Slides for in-class presentation:** In addition to the .pdf document, each team must prepare four sets of slides (one for each section of the problem set, omitting the introduction) to present in class. These must be uploaded to the activity 'Slides: PS1' in Brightspace.

  – File name format: `nombre_equipo_##` (use leading zero for teams numbered below 10). Example for team 1:

    - `data_equipo_01` (Data)
    - `age_equipo_01` (Age-wage profile)
    - `gap_equipo_01` (Equal Pay for Equal Work?)
    - `pred_equipo_01` (Predicting wages)
  - Respect these file names exactly.
  - Maximum 10 minutes per presentation (approximately 3 slides).

Note. A general reference on tax administration gaps (not required here, but cited in the original document) is available at: https://www.irs.gov/newsroom/the-tax-gap.