Understanding and Predicting Wages: Evidence from Bogotá

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

Link to the GitHub repository: https://github.com/jrconstain/PS1_Group4

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 and 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 (Mejía, 2021, Lora and Benítez, 2021). 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, 2025a). 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 (DANE, 2025b). 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 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.

[1]This figure is obtained by applying the 0.07% of GDP that Mejía (2021) estimate as efficiently recoverable from labor income underreporting to the official 2024 GDP estimate for Colombia (COP 1,706.4 trillion) published by (DANE, 2025a)

Beyond the causal analysis, our predictive exercises demonstrate that wage outcomes can be forecast with reasonable accuracy using relatively parsimonious models. Across multiple specifications, models enriched with nonlinear terms for age and socioeconomic status consistently delivered the lowest prediction errors. In particular, the specification labeled Additional 2 achieved the best performance, with a Root Mean Square Error (RMSE) of 0.45, a result that proved robust under both K-fold and Leave-One-Out Cross-Validation. These findings suggest that predictive models can complement traditional econometric approaches by identifying patterns in the data that flag anomalous observations and, by extension, potential cases of underreporting in income declarations.

Taken together, our findings contribute to both the explanatory and predictive understanding of wage determination in Colombia. The evidence highlights how age and gender shape earnings, how firm and occupational sorting mediate the gender gap, and how predictive models can improve the detection of underreporting in labor income. The remainder of the paper is organized as follows. Section 2 describes the data and variable construction. Section 3 presents the econometric analysis of the age-wage profile, and Section 4 analyzes the gender-earnings gap. Section 5 develops the predictive models and evaluates their performance to close our excercise.

2 Data

We base our exercise on a dataset[2] extracted through a scraping routine to collect all 32,177 observations of individuals surveyed in Bogotá for the 2018 GEIH (*Gran Encuesta Integrada de Hogares*). Conducted by DANE since 2006, the GEIH is Colombia's main source of labor statistics, covering employment, earnings, demographics, and income sources nationwide (BanRep, 2025, DANE, 2025b). From this universe, we restrict the sample to wage earners aged 18 and above, excluding the self-employed and unpaid family workers (N = 9,892). Because the GEIH asks respondents about the number of hours worked during the previous week, values reported may reflect atypical events. We trimmed the top and bottom one percent of the distribution to account for these, excluding 192 cases. This yields our final analytic sample of N = 9,700. We focus on labor income as a more stable outcome largely driven by individual characteristics such as age and education, in contrast to entrepreneurial or informal gains more prone to be affected by external, non-observable factors.

To enrich our analysis, we constructed two additional variables. First, we derived *years of education* from the GEIH questions `p6210` (highest level attained) and `p6210s1` (years completed), by summing the years required to get to the highest level attained and years reported as completed within that level. Second, following (FernÃ¡ndez and Messina, 2018), we computed a measure of potential *experience* as *age* − *years of education* − 6. These transformations allow us to approximate human capital more accurately across individuals.

Table 1 presents the descriptive statistics for the main variables included in our analysis. The average individual in the sample is 36.2 years old with 18.4 years of potential experience, showing little variation between men and women. Notably, women exhibit slightly higher educational attainment than men: they report an average of 12.2 years of education versus 11.5 for men, and their maximum educational level reached is also higher

[2]Assembled by Professor Manuel Fernández and made accessible via the website of Professor Ignacio Sarmiento, Universidad de los Andes. Publicly available at: https://ignaciosarmiento.github.io/GEIH2018_sample/. Importantly, no restrictions (such as a `robots.txt` file) prevented this collection.

(6.20 vs. 6.01 on a seven-point scale). In contrast, labor outcomes show that men earn more on average, both monthly (COP 1,836,380 vs. COP 1,676,747) and hourly (COP 8,945 vs. COP 8,659), and also work more hours per week (49.9 vs. 46.6). Differences in employment structure are also apparent: a greater share of men are formally employed (78.9% vs. 76.2%), men are less likely to work in small firms (18.5% vs. 25.5%), and they tend to be employed in private firms (93.8% vs. 83.8%) and in slightly larger organizations on average (firm size 4.06 vs. 3.82).

Table 1. Individual and job characteristics: overall and by sex.

	All (N=9,700)	Men (N=4,875)	Women (N=4,825)
Panel A. Individual characteristics			
Age (years)	36.232 (11.986)	35.955 (12.147)	36.513 (11.815)
Experience (years)	18.373 (13.352)	18.429 (13.360)	18.316 (13.345)
Socioeconomic level (1–6)	2.508 (0.976)	2.438 (0.949)	2.579 (0.998)
Maximum educational attainment[a](1–7)	6.104 (1.100)	6.013 (1.115)	6.196 (1.077)
Years of education	11.866 (3.999)	11.530 (4.010)	12.206 (3.959)
Panel B. Earnings & work hours			
Labor income (monthly, COP)	1,756,975 (2,412,230)	1,836,380 (2,493,475)	1,676,747 (2,324,769)
Hourly labor income (COP/hour)	8,803 (12,563)	8,945 (12,728)	8,659 (12,393)
Total hours worked (week)	48.274 (10.522)	49.949 (10.391)	46.582 (10.382)
Panel C. Employment structure (%)			
Formal employment	77.6	78.9	76.2
Small firm ($\leq$5 employees)	22.0	18.5	25.5
Private employees	88.8	93.8	83.8
Firm size[b]	3.94 (1.32)	4.06 (1.18)	3.82 (1.44)

Notes: Each numeric cell reports the *Mean* on the first line and the *Standard Deviation* (in parentheses) on the second line. Monetary values in Colombian pesos (COP). *(a)* Firm size is an ordinal scale; higher values indicate larger firms. *(b)* Maximum educational attainment ranges from 1 to 7 (with 6 = complete secondary, 11 years; 7 = tertiary). One observation is missing in this variable for the overall sample (All: $N = 9{,}699$).

3 Age-wage profile

To better understand the determinants of wages, we begin by examining the age-wage profile–a core building block of human-capital models since Mincer (1974)–which captures how average earnings vary with an individual's age (Startz, 2014). This profile typically follows an inverted U-shape, with wages increasing until peaking in the early 50s, and then gradually declining as retirement nears. We estimate the following quadratic Mincerian specification, a standard reduced-form approach that captures the life-cycle wage pattern and allows identification of the turning point:

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

where w denotes hourly wages, and $\beta_2 > 0$ and $\beta_3 < 0$ under the canonical concave profile.

Table 2 show the results of our estimation. Both age coefficients exhibit the expected signs, implying a concave life-cycle profile consistent with human-capital accumulation and job-ladder mechanisms (Mincer, 1974, Card, 1999). The point estimates (0.069 on age; -0.001 on age^2) imply that hourly wages increase with age but at a diminishing marginal rate, a pattern prevalent across labor markets and cohorts (Murphy and Welch, 1990, Lagakos et al., 2018). As expected in a cross-sectional setting, the explanatory power of the model is modest ($R^2 = 0.047$; residual s.e. $= 0.705$ log points). The remaining dispersion likely reflects variation in factors such as education, occupation, firm characteristics, and match quality-determinants emphasized in both empirical and structural literatures (Heckman et al., 2006).

Table 2. Unconditional Age–Wage Profile regression.

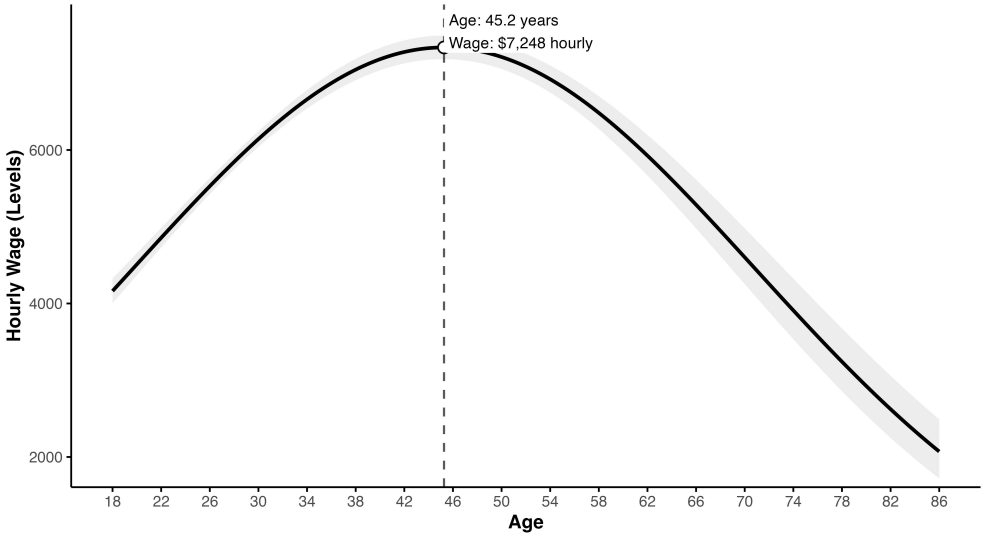
	Dependent variable:
	Log Hourly Wage
Age	0.069^{***}
	(0.004)
Age^2	-0.001^{***}
	(0.00004)
Constant	7.339^{***}
	(0.068)
Observations	9,700
R^2	0.047
Adjusted R^2	0.047
Residual Std. Error	0.705 (df = 9697)
F Statistic	239.598^{***} (df = 2; 9697)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 1 displays the estimated age–wage trajectory for the full sample of 9,700 workers (men and women combined), revealing a smooth concave pattern. Under the quadratic model, the turning point is given by:

$$a^* \;=\; -\frac{\widehat{\beta}_2}{2\,\widehat{\beta}_3},$$

which yields a peak at **45.2 years**. At this peak, the fitted hourly wage level is approximately \$7,248 COP/hour. To assess the precision of this estimate, we implement a nonparametric pairs bootstrap with $B = 1{,}000$ resamples. The resulting distribution of a^* yields a standard error of 0.672 years, negligible bias (-0.027), and a 95% percentile interval of [44.07, 46.65], reflecting the relative flatness of the profile near its peak.

Figure 1. Estimated Age–Wage profile.



Notes: The dashed line marks the peak age; the dot annotates the peak wage.

4 The gender earnings GAP

The gender wage gap has long been a central topic in labor economics, with evidence pointing to persistent disparities that cannot be fully explained by productivity differences alone (Goldin, 2014). To estimate the effect of being female on hourly wages, we rely on the Frisch-Waugh-Lovell (FWL) theorem, which allows us to partial out sets of controls and isolate the residual contribution of gender. Table 3 reports our estimates across four specifications, beginning with an unconditional regression and progressively introducing controls for individual characteristics (years of education, age, and age^2), firm attributes (formality, sector, and firm size), and occupation.

Table 3. Female-Wage unconditional and conditional regressions.

	No Controls	Individual controls	Firm controls	Occupation controls
	(1)	(2)	(3)	(4)
Female (=1) / Residual FWL	-0.047^{***}	-0.138^{***}	-0.105^{***}	-0.084^{*}
	(0.015)	(0.011)	(0.011)	(0.011)
Constant	8.752^{***}	0.000	0.000	0.000
	(0.010)	(0.006)	(0.005)	(0.005)
Controls education/age	No	Yes	Yes	Yes
Formality, sector and firm size	No	No	Yes	Yes
Type of occupation	No	No	No	Yes
Observations	9,700	9,700	9,700	9,700
R^2	0.001	0.016	0.010	0.006
Adjusted R^2	0.001	0.016	0.009	0.006
Residual Std. Error (df = 9698)	0.721	0.545	0.512	0.463
F Statistic (df = 1; 9698)	10.270^{***}	155.092^{***}	93.887^{***}	55.306^{*}

*p<0.1; **p<0.05; ***p<0.01

The unconditional regression (1) indicates that women earn on average 4.7 percent less per hour than men, what at first glance may appear as a modest difference. However, when comparing women and men with similar observable characteristics (years of education, age, and age squared), as in specification (2), the gap jumps to 13.8%. This occurs

because, as shown in the descriptive statistics, women in our sample tend to be more educated than men. Thus, failing to account for these differences leads the unconditional estimate to understate the true wage disparity.

Including firm-level characteristics in (3)-such as whether the job is formal or informal, whether it belongs to the public or private sector, and the size of the firm–reduces the gap to 10.5%. This reflects the fact that women are more likely to self-select into smaller or less formal firms, which generally pay lower wages. Finally, specification (4) adds controls for occupation, further reducing the gap to 8.4 percent. This highlights that occupational sorting-with women disproportionately concentrated in lower-paying professions-explains part of the observed disparity.

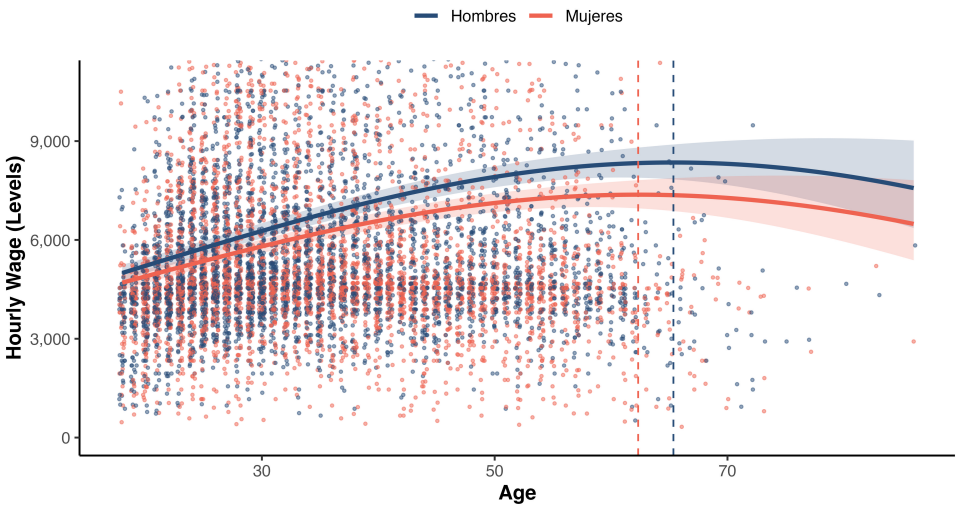
Using FWL tends to provide wrong standard errors due to the incapacity to recognize the complete set of estimators that were omitted in the final regression, that is why bootstrap is used to recalculate properly this values. Doing so for model 4 (all controls), we can observe a bias of -0.0002, which is really small and shows a good performance of the initial estimation. Also, the confidence intervale is between -10.8% and -6% at a 95% level; validating the statistical significance of the estimator.

Overall, the sequential introduction of controls reveals a combination of selection and discrimination effects. Selection is evident in women's greater representation in smaller, less formal firms and lower-paying occupations. Yet, even after accounting for these compositional differences, a residual gap of 8.4% percent remains. This indicates that women with comparable education, age, firm characteristics, and professional roles still earn significantly less than men, providing strong evidence against the notion of "equal pay for equal work."

4.1 Gender-Age Gap

Additionally, analyzing the age-wage peak separately for each gender could potentially reveal the interaction between two factors: age and gender, which may disproportionately affect older women. This was solved by adding into model 3 interactions between Age and Age^2 with $Female$. In Figure 2, we observe that the peak age at which men and women reach their maximum wages is at 62 for women and 65 for men. Prior research suggests that this U-shaped pattern, typically showing a peak in the fifties, largely reflects the cross-sectional nature of wage data. Wages do not necessarily decline at older ages; rather, they tend to remain relatively stable, while the apparent decline is driven by partial retirement and gradual reductions in labor supply among older workers (Luong and Hébert, 2009, Casanova, 2012, Scarfe et al., 2023). A more detailed discussion of why these estimated peaks change substantially once controls are introduced is provided in the Appendix.

Figure 2. Age-Wage Profile by sex: Conditional Estimation.



The estimated wage-age trajectories for both genders display a similar shape, with the primary difference being the scale of earnings at each age-women reaching a peak salary of $7,373 per hour, and men $8,350, almost 14% less. Statistically speaking, these peaks are not significantly different, as a hypothesis test using the t-Student method yields a p-value of 0.6. This further suggests that, when controlling for the factors discussed earlier, men and women follow similar age-wage patterns; however, gender still influences the overall level of earnings, with men generally earning higher salaries throughout their careers.

5 Predicting wages

Now, moving away from the standard econometrical view that only focuses on causality and estimators analysis, we will try to predict the salary per hour using various specifications, including the ones presented before.

Validation Set Approach (VS): Using a sample of 70% of the total sample (6794 observations) for training and 30%(2906) for testing, we selected 8 specification to try to predict salary per hour (That will be listed in the appendix).

Table 4. RMSE of Predicted Models using Validation Set Approach.

Model	RMSE VS
Additional 2	0.4502
Additional 1	0.4543
Base 2	0.4706
Additional 4	0.4713
Additional 5	0.5372
Additional 3	0.5405
Base 0	0.7069
Base 1	0.7230

In Table 4, we observe the Root Mean Square Error (RMSE) of the specifications listed in the appendix (, organized from the lowest to the highest. Overall we can see that, besides Model Base 0 and 1, the models have a good perfomance calculating the

LogHourlyWage. Calculating the relative error in percentage, by evaluating each RMSE with e and subtracting -1, we can observe that the first 6 models predictions are around 56% and 70% difference of the original value, and the last two, around 100%. Model Additional 2 showed the lowest RMSE with 0.4502. This specification is an improvement compared to the conditional gender earnings gap presented in previous sections intended to increase its predicting potential. Specifically, this model adds variables such as Age to the 3 and 4 for more complexity and non-linearity and also the socioeconomical status, a variables that while predicting the gaps in gender might be as relevant as other persona characteristics, for predicting the wages is really crucial since it classifies people into categories for economical status depending on where you live.

To identify the observations that “missed the mark", we defined these as those points that were more than 2 standard deviations away from the mean; a total of 167 observations “missed the mark" in the prediction.

Table 5. Observations that "missed the mark" main characteristics

Statistic	N	Mean	St. Dev.	Min	Max
Monthly Wage	167	5,227,224.000	7,907,487.000	60,000.000	60,100,000.000
Total Hours Worked	167	45.826	14.212	10	84
Age	167	38.108	12.973	18	71
Male	167	0.515	0.501	0	1
Socialeconomical Status	167	3.084	1.333	1	6
Years of Education	167	13.862	4.362	0	22
Experience	167	18.246	13.996	0	60

In Table 5 we can observe the main characteristics of these 167 observations in the most relevant variables in our dataset. Overall, we can not identify clear patterns in variables such as Age, Male, Socialeconomical Status, Years of education or Experience, since they are well between the same distributions of all the observations as showed during previos sections. Nonetheless it is important to identify the behaviour of Monthly Wage to see any patterns regarding the Salary.

Figure 3. Laboral Wage of Prediction Error Outliers

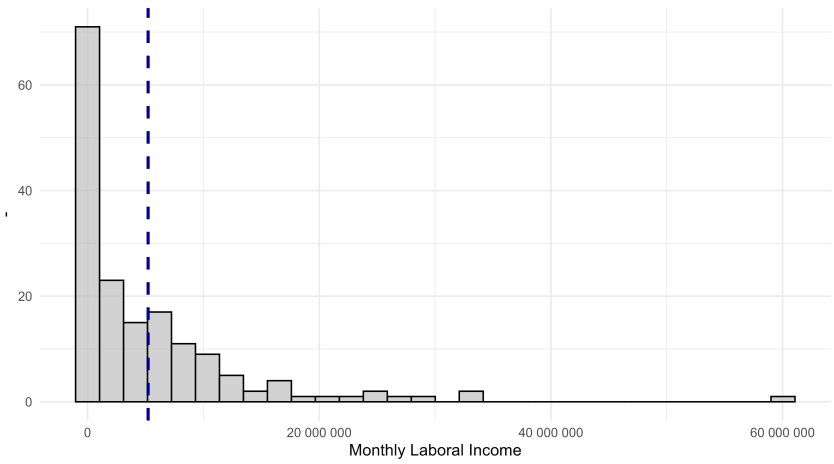


Figure 3 shows that most of the observation presenting prediction errors tend to be allocated in the left wing of the distribution, reporting less than $5000000 in Monthly Laboral Wage. It is important to note that, when most of the distributions is around the same values, the fact that this predictions miss the values in such a significant way, raise

an alarm to look into the values reported, to avoid any kind of under reporting for tax evation.

LOOCV. For the top 2 specifications with the lowest RMSE, resampling methods such as K-fold Cross-Validation (K-fold CV) and Leave One Out Cross-Validation (LOOCV) were used to validate the initial prediction value and to identify if these RMSE are consistent and the models do capture a good prediction or if we are over-fitting the models, making them good for that specific partition but not as good for any other kind of sample. For the case of K-fold Cross-Validation, k=10 was used, as is standard within the discipline. This was done to verify whether the use of a more computationally lightweight model yields results that do not differ significantly or, alternatively, to analyze potential influential observations that might be biasing the results in certain partitions.

Table 6. RMSE of Predicted Models using Validation Set Approach, K-fold CV, and LOOCV

Model	RMSE_VS	RMSE_K-fold CV	RMSE_LOOCV
Additional 2	0.4502	0.4486	0.4493
Additional 1	0.4543	0.4543	0.4533

In Table 6, we present the RMSE results from K-fold CV and LOOCV, compared to the original RMSE obtained from the validation set (VS) approach. We observe that the predictions from the VS approach are consistent with those from the more sophisticated training-testing methods. If the results had shown significant differences between VS and the other approaches, it might suggest that some influential observations were impacting the prediction model within the specific partition used in the first approach. This would warrant an investigation into why these observations have such an effect on the modelâs predictions and whether their inclusion or exclusion is beneficial for the overall training process.

References

BanRep (2025). Notas metodológicas de series estadísticas históricas: Mercado laboral. Recuperado el 7 de septiembre de 2025.

Card, D. (1999). The causal effect of education on earnings. In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics*, volume 3A, pages 1801–1863. Elsevier (North-Holland), Amsterdam.

Casanova, M. (2012). Wage and earnings profiles at older ages. Working Paper 2012-001, Human Capital and Economic Opportunity Working Group, University of Chicago, Chicago, IL.

DANE (2025a). Gran encuesta integrada de hogares (geih) históricos. Recuperado el 7 de septiembre de 2025.

DANE (2025b). Pib por departamento: Información 2024 preliminar. Recuperado el 7 de septiembre de 2025.

Fergusson, L. and Hofstetter, M. (2022). The colombian tax system: A diagnostic review and proposals for reform. Technical Report UNDP LAC PDS No. 28, United Nations Development Programme (UNDP).

FernÃ¡ndez, M. and Messina, J. (2018). Skill premium, labor supply, and changes in the structure of wages in latin america. *Journal of Development Economics*, 135:555–573.

Heckman, J. J., Lochner, L. J., and Todd, P. E. (2006). Earnings functions, rates of return and treatment effects: The mincer equation and beyond. In Hanushek, E. A. and Welch, F., editors, *Handbook of the Economics of Education*, volume 1, chapter 7, pages 307–458. Elsevier, Amsterdam.

Lagakos, D., Moll, B., Porzio, T., Qian, N., and Schoellman, T. (2018). Life-cycle wage growth across countries. *Journal of Political Economy*, 126(2):797–849.

Lora, E. and Benítez, M. (2021). Impuesto de renta a las personas naturales. In Lora, E. and Mejía, L. F., editors, *Reformas para una Colombia post-COVID-19: Hacia un nuevo contrato social*, chapter 4. Fedesarrollo, Bogotá.

Luong, M. and Hébert, B.-P. (2009). Age and earnings. *Perspectives*, (75-001-X):5.

Mejía, L. F. (2021). Elementos para reducir la evasión y la elusión tributaria. In Lora, E. and Mejía, L. F., editors, *Reformas para una Colombia post-COVID-19: Hacia un nuevo contrato social*, chapter 8. Fedesarrollo, Bogotá.

Mincer, J. (1974). *Schooling, Experience, and Earnings*. Columbia University Press for the National Bureau of Economic Research, New York.

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

Scarfe, R., Singleton, C., Sunmoni, A., and Telemo, P. (2023). The age-wage-productivity puzzle: Evidence from the careers of top earners. *Economic Inquiry*.

Startz, R. (2014). Age-earnings profile. In *Encyclopedia of Education Economics & Finance*. SAGE Publications.

6 Appendix

6.1 Specifications for prediction in "Predicting Wages"

Base 0:

$$\text{Log Hourly Wage}_i = \beta_0 + \beta_1 \,\text{Age}_i + \beta_2 \,(\text{Age}_i)^2 + \varepsilon_i$$

Base 1:

$$\text{Log Hourly Wage}_i = \beta_0 + \beta_1 \,\text{Female}_i + \varepsilon_i$$

Base 2:

$$\begin{aligned}\text{Log Hourly Wage}_i = \beta_0 &+ \beta_1 \,\text{Female}_i + \beta_2 \,\text{Years of Education}_i \\ &+ \beta_3 \,\text{Age}_i + \beta_4 \,(\text{Age}_i)^2 \\ &+ \beta_5 \,\text{Formal Business}_i + \beta_6 \,\text{Size Firm}_i \\ &+ \beta_7 \,\text{Laboral Relation}_i + \beta_8 \,\text{Profession}_i + \varepsilon_i\end{aligned}$$

Additional 1:

$$\begin{aligned}\text{Log Hourly Wage}_i = \beta_0 &+ \beta_1 \,\text{Female}_i + \beta_2 \,\text{Max Educational Year}_i \\ &+ \beta_3 \,\text{Experience}_i + \beta_4 \,(\text{Experience}_i)^2 \\ &+ \beta_5 \,\text{Formal Business}_i + \beta_6 \,\text{Size Firm}_i \\ &+ \beta_7 \,\text{Laboral Relation}_i + \beta_8 \,\text{Profession}_i + \varepsilon_i\end{aligned}$$

Additional 2:

$$\begin{aligned}\text{Log Hourly Wage}_i = \beta_0 &+ \beta_1 \,\text{Female}_i + \beta_2 \,\text{Years of Education}_i \\ &+ \beta_3 \,\text{Age}_i + \beta_4 \,(\text{Age}_i)^2 \\ &+ \beta_5 \,(\text{Age}_i)^3 + \beta_6 \,(\text{Age}_i)^4 \\ &+ \beta_7 \,\text{Formal Business}_i + \beta_8 \,\text{Size Firm}_i \\ &+ \beta_9 \,\text{Laboral Relation}_i + \beta_{10} \,\text{Profession}_i \\ &+ \beta_{11} \,\text{Socioeconomical Status}_i + \varepsilon_i\end{aligned}$$

Additional 3:

$$\begin{aligned}\text{Log Hourly Wage}_i = \beta_0 &+ \beta_1 \,\text{Female}_i + \beta_2 \,\text{Years of Education}_i \\ &+ \beta_3 \,\text{Age}_i + \beta_4 \,(\text{Age}_i)^2 \\ &+ \beta_5 \,(\text{Age}_i \times \text{Female}_i) \\ &+ \beta_6 \,(\text{Age}_i^2 \times \text{Female}_i) \\ &+ \beta_7 \,(\text{Age}_i \times \text{Years of Education}_i) + \varepsilon_i\end{aligned}$$

Additional 4:

$$\begin{aligned}\text{Log Hourly Wage}_i = \beta_0 &+ \beta_1 \,\text{Female}_i + \beta_2 \,\text{Years of Education}_i \\ &+ \beta_3 \,\text{Age}_i + \beta_4 \,(\text{Age}_i)^2 \\ &+ \beta_5 \,(\text{Age}_i)^3 + \beta_6 \,(\text{Age}_i)^4 \\ &+ \beta_7 \,\text{Experience}_i + \beta_8 \,(\text{Experience}_i)^2 \\ &+ \beta_9 \,(\text{Experience}_i)^3 + \beta_{10} \,(\text{Experience}_i)^4 \\ &+ \beta_{11} \,(\text{Experience}_i)^5 \\ &+ \beta_{12} \,\text{Formal Business}_i + \beta_{13} \,\text{Size Firm}_i \\ &+ \beta_{14} \,\text{Laboral Relation}_i + \beta_{15} \,\text{Profession}_i \\ &+ \beta_{16} \,\text{Socioeconomical Status}_i + \beta_{17} \,\text{College}_i \\ &+ \beta_{18} \,(\text{Age}_i \times \text{Female}_i) \\ &+ \beta_{19} \,(\text{Age}_i^2 \times \text{Female}_i) + \varepsilon_i\end{aligned}$$

Additional 5:

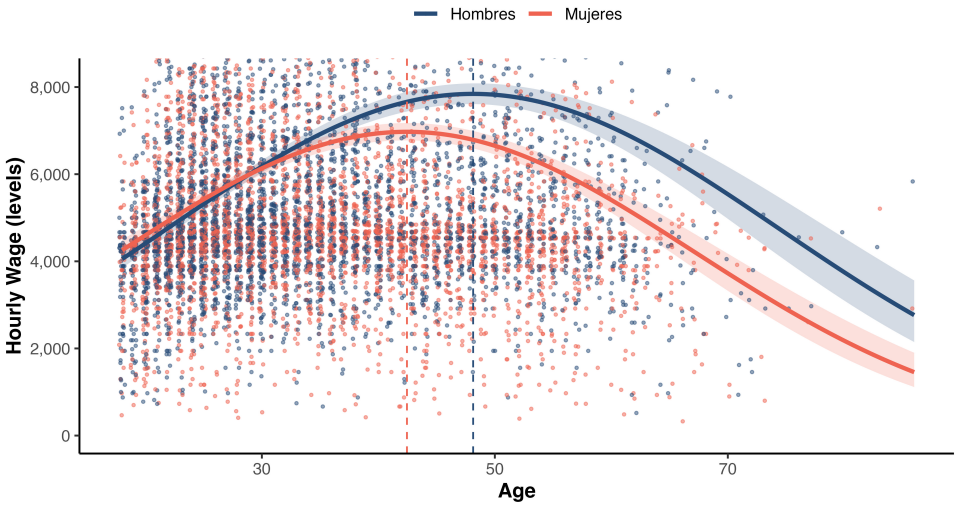
$$\begin{aligned}\text{Log Hourly Wage}_i = \beta_0 &+ \beta_1 \,\text{Experience}_i + \beta_2 \,(\text{Experience}_i)^2 \\ &+ \beta_3 \,(\text{Experience}_i)^3 + \beta_4 \,(\text{Experience}_i)^4 \\ &+ \beta_5 \,(\text{Experience}_i)^5 + \beta_6 \,(\text{Experience}_i)^6 \\ &+ \beta_7 \,(\text{Experience}_i)^7 + \beta_8 \,\text{Profession}_i \\ &+ \beta_9 \,(\text{Profession}_i \times \text{Experience}_i) + \varepsilon_i\end{aligned}$$

6.2 Age-Wage Profile - an Unconditional Estimation

Additionally, we wanted to present how the peak ages would change in the initial unconditional model for Gender Earnings Gap.

This results reflect the condition of the gap before controlling the observations and identifying the possible mix effect described in the main body of the paper.

Figure 4. Age-Wage Profile by sex: Unconditional Estimation



In figure 4 we can observe how men and women tend to have equal salaries, but then women start to fall behind, which is consistent with the estimation of model 2 in the gender age gap that included individual characteristics; women tend to be underpaid, in reference to men, when taking into account age characteristics, with this indication that around 40 years old, the divergence between the two gender starts.

Lastly, this unconditional version, speaks directly to the Figure 1, since having an almost perfectly distributed sample in gender (50% women and 50% men) could allow us to average the two peaks (48.1 for men and 42.4 for women) and find the same peak age for all the sample, 45.2)