Problem Set 1: Predicting Income

Gianluca Cicco Bilbao 202020881
Adrián Suárez García 202123771
Sergio Delgado Quevedo 202212287
Andrez Felipe Guerrero 202424503

September 7, 2025

GitHub Repository:
https://github.com/BigData-ML-G5/Problem-Set-1-Predicting-Income

Contents

1 Introduction

Accurately predicting individual hourly wages is key for public policy (taxation, targeting, and social protection) and also a natural testbed to compare econometric and machine-learning approaches. This study builds and compares wage-prediction models using GEIH–Bogotá 2018 microdata, with a linear, Mincer-style specification as the benchmark. Labor-economics research shows that human capital (schooling and potential experience) is central for wage determination, with concave age–earnings profiles and robust returns to education [5, 3]. In Colombia, institutional features such as a binding minimum wage and high informality segment the wage distribution and create mass around the minimum, affecting both levels and dispersion [1].

Against this background, our contribution is purely predictive. We contrast an interpretable econometric baseline (Mincer-type OLS with controls for occupation, industry, firm size, and formality) with flexible algorithms that capture nonlinearities and interactions without prespecification. This is relevant because, when the goal is to minimize out-of-sample error, machine-learning methods often add value by relaxing functional assumptions and exploiting high-dimensional features [6]. In related labor applications, tree-based tools—random forests and boosting—have shown advantages for classifying/predicting wage-related outcomes and policy exposure, including identifying workers affected by minimum-wage changes [4].

Methodologically, we estimate OLS models and compare them with ML algorithms. All comparisons rely on out-of-sample validation: a 70/30 train–test split (pre-set seed), RMSE/MAE metrics, and LOOCV for the two best models. We also report variable importance and inspect tail errors to discuss whether outliers reflect mis-measurement, under-reporting, or unmodelled heterogeneity.

The GEIH–Bogotá 2018 survey fits this task due to its urban sample size and rich sociodemographic and job-related covariates. We model log hourly wages (computed from labor income and effective hours) and document cleaning choices, exclusions, and treatment of extreme values. This setting lets us test whether flexible models leverage structural features of Bogotá's labor market—such as mass around the minimum wage and formal–informal segmentation—that may limit strictly linear specifications [1].

Our findings confirm several key hypotheses. The analysis reveals a statistically significant concave age-wage profile consistent with life-cycle theory, with wage level peaking around age 54, implying age is a key aspect to explain wage level, yet high variance on our model estimation strongly suggest other variables affecting wage must be considered. Furthermore, we quantify a persistent gender wage gap that, while partially explained by observable characteristics, remains significant, especially present when performing the same job. Our estimations for age-wage profile discriminated by gender suggest a similar average wage level at young age, but highly differentiated entering mid age, with women peaking around age 47, while male peaking at 60, with men having significantly higher wages. Finally, the predictive evaluation demonstrates that flexible models incorporating complex, non-linear interactions substantially outperform simpler linear specifications in terms of out-of-sample error, highlighting the importance of feature engineering for this task.

2 Data

2.1 Data Source and Acquisition

The data for this study is a sample from the 2018 Gran Encuesta Integrada de Hogares (GEIH) for the city of Bogotá, made publicly available on the following website `https://ignaciomsarmiento.github.io/GEIH2018_sample/]`. The acquisition was performed using web scraping techniques with the `rvest` package in R, iterating through ten distinct HTML pages available on the main page, each containing a chunk of the complete data sample, each presented table and obtained itself from a GET request of an outer web page; we replicated this request through all chunks to consolidate the final dataset.

2.2 Data Cleaning and Sample Preparation

The raw dataset underwent a comprehensive, detailed cleaning process. First, considering the GEIH columns are named through codes non-representative of the variable's meaning, and that some others were crafted from the original data, we employed the data set's dictionary, available at https://ignaciomsarmiento.github.io/GEIH2018_sample/ddi-documentation-spanish-608.pdf and renamed each column. Second, we created 4 crucial variables for our further analysis due to their social relevance to wage levels, as we discuss furtherly: number of minors, household head, household head female, and age squared; each represent the amount of minors in a household, if the person is the head of the household, if that household head is female, and the square of each person's age, respectivelly. After filtering for individuals `18 years or older` with positive hours worked, we selected the main variables for our investigation and filtered the data set to just keep those; section 2.3 explains to detail these selections. The final dependent variable for this study is `ingreso_laboral_horas_actuales` (hourly labor income). This variable was deliberately selected to isolate earnings derived exclusively from labor. Other available income measures in the survey included non-labor-related sources such as rents or pensions, which would have confounded the analysis by introducing relationships not relevant to our objective of modeling wages. By focusing specifically on labor income, we ensure that our dependent variable aligns with the Mincerian framework and other foundational theories in labor economics.

To handle cases of missing data in this variable, we implemented a conditional mean imputation strategy based on occupation (`oficio`), age, and gender. As discussed in the descriptive analysis, these predictors are strong indicators of an individual's wage. Finally, we applied a natural logarithm to the completed income variable to create our final target, `log_ingreso_laboral_horas_actuales`. This transformation is standard practice as it helps to stabilize the variance of the variable, reduce the influence of extreme outliers, and allows for the coefficients in our models to be interpreted as semi-elasticities (approximate percentage changes).

2.3 Descriptive Statistics

The final cleaned dataset consists of **16,356 observations**. Table 1 presents a detailed summary of the key variables used throughout the analysis, including measures of central tendency, dispersion, and inline histograms that visualize the distribution of each variable. The average log hourly wage is approximately 8.8, and the average worker is 39.2 years old. The sample is balanced by gender and is primarily composed of formal workers (60%).

Table 1: Descriptive Statistics of Key Variables

Variable	**N**	**Mean**	**Std. Dev.**	**Min**	**Max**
`log_ingreso_laboral_horas_actuales`	16,356	8.452	0.724	4.129	12.288
`age`	16,356	41.472	12.001	18	88
`hombre`	16,356	0.528	0.499	0	1
`formal`	16,356	0.748	0.434	0	1
`estrato1`	16,356	3.220	1.391	1	6
`maximo_nivel_educativo`	16,356	4.671	1.802	1	7

Further exploratory analysis, summarized in Figure 1, reveals several key relationships between hourly income and worker characteristics. The plots show clear positive trends between income and socioeconomic level, education, company size, and tenure. These visual insights confirm the relevance of these variables as predictors and controls in our models.

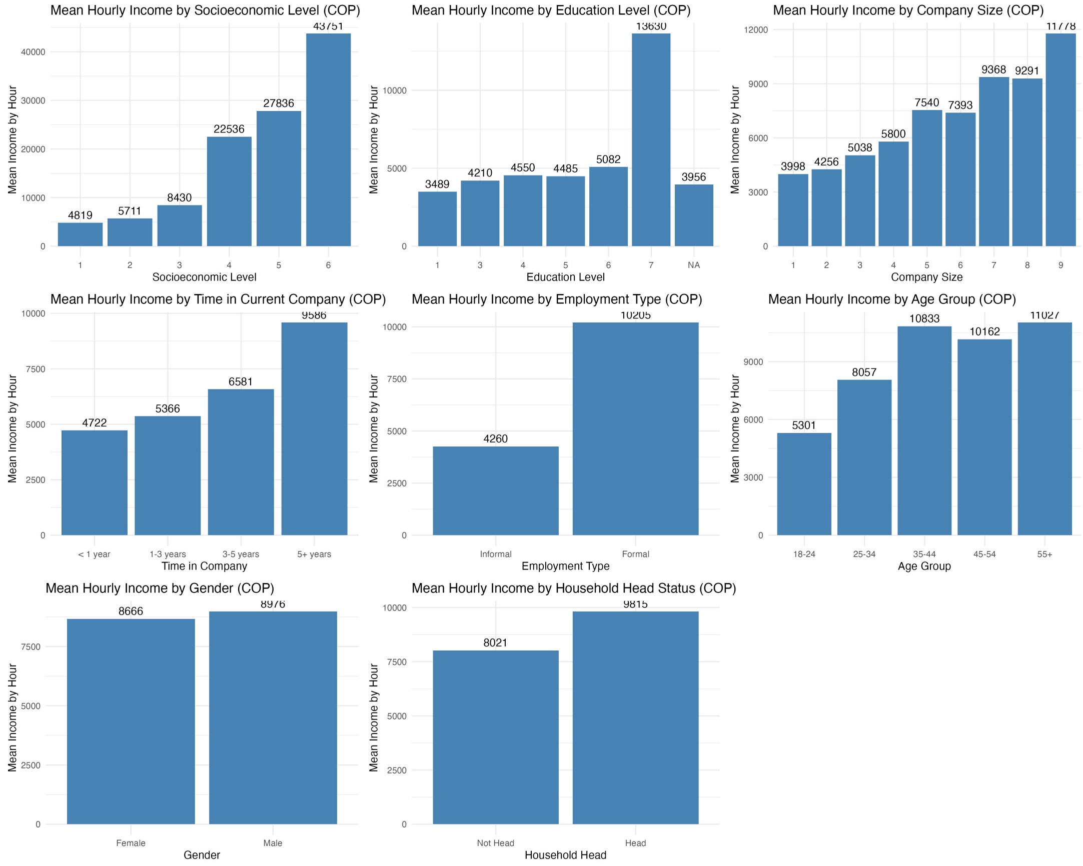

Figure 1: Exploratory analysis of mean hourly income by various socioeconomic and job characteristics.

3 Age-Wage Profile

Following labor economics theory, we estimate the age–wage profile using a linear regression model including standard quadratic specification, as follows:

$$\log(\text{LaborIncome}_i) = \alpha + \beta_1 \, \text{age}_i + \beta_3 \, \text{Age}_i^2 + \varepsilon_i. \tag{1}$$

The results of the OLS regression are presented in Table 2, while Figure 2 provides a graphical representation of the estimated relationship between age and hourly income. The curve highlights the increasing returns to experience at younger ages, followed by a concave pattern that captures the decline in wages at later stages of the working life cycle.

Table 2: Estimation of the Age-Wage Profile

	log(Hourly Wage)
Variable	(1)
Age	0.044***
	(0.002)
Age^2	-0.0004***
	(0.00003)
Constant	7.780***
	(0.049)
Observations	16,356
R^2	0.043

Note: Dependent variable is `log(ingreso_laboral_horas_actuales)`. Standard errors from OLS. *** p¡0.01.

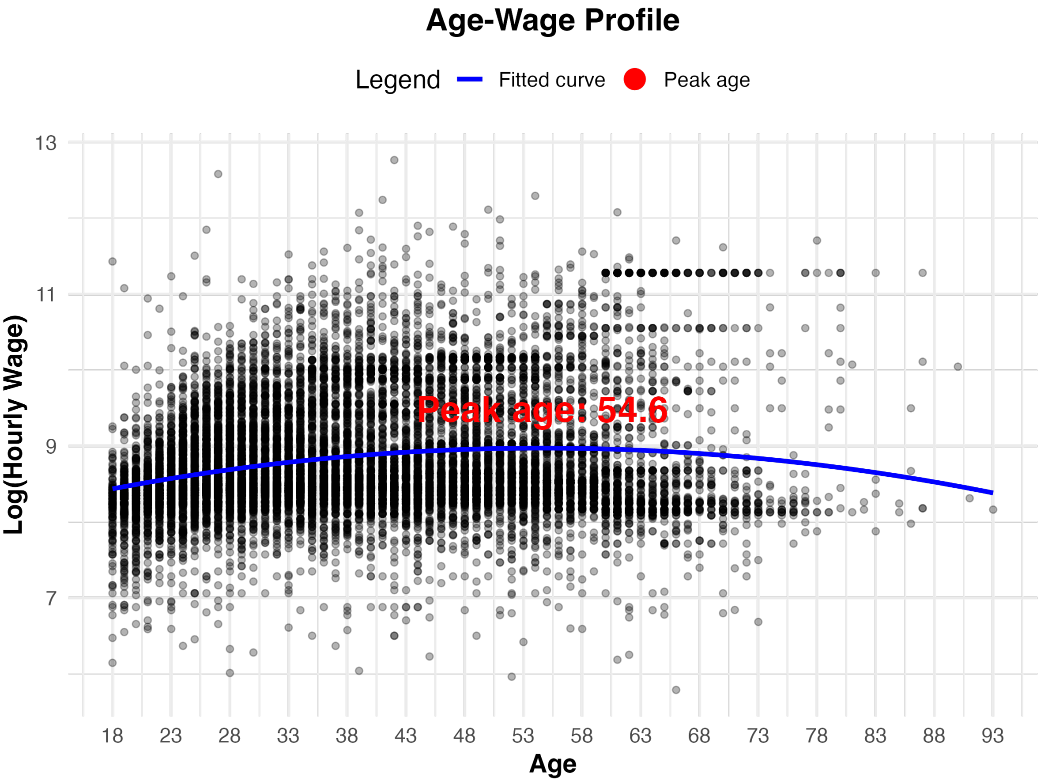

Figure 2: Estimated Age-Wage Profile. The graph illustrates the predicted relationship between worker age and hourly income, highlighting the peak age where earnings tend to be maximized.

The regression output confirms the expected concave relationship, with a positive coefficient for `age` and a negative one for `age`^{2}. The term associated with age indicates that, on average, an increase in 1 year of age implies a 4.4% increase in hourly wage. Nonetheless, the quadratic term implied by age squared is negative, meaning that the marginal effect of age on wages diminishes over time. Based on these estimates, the earnings profile peaks at approximately 54.6 years of age, after which the effect of age on wages turns negative. In other words, while workers wage grows early in their careers due to experience gain, this growth slows down as they approach middle age, eventually declining slightly after the mid-50s. Furthermore, the `R`^{2} value indicates that age alone explains 4.4% of wage variation, a predictable outcome as economic and social theory suggest multiple other variables are fundamental to explain someone's salary, such as education and occupation. This is evident in Figure 2, as we observe a great variance around the fitted curve.

Furthermore, to assess the uncertainty of this estimate, a non-parametric bootstrap with 1,000 replications was performed. This procedure yielded a 95% confidence interval for the peak age of [**51.21, 57.69**], which does not span an excessively wide range, giving us confidence in the stability of our estimate. The confidence intervals for the `age` and `age`^{2} coefficients themselves were also confirmed to not contain zero, reinforcing the statistical significance of the concave relationship. We confirm with our confidence intervals that, even with age being a key aspect explaining salary, there are many other variables to be considered, thus further exploratory analysis is done to better understand what explains people's wage level in Bogota.

4 The Gender Earnings Gap

The selection of control variables for the conditional model is grounded in established labor-economics research. Human Capital Theory [5] motivates the inclusion of education (experience (proxied by `age` and its square, and `maximo_nivel_educativo`) as primary determinants of productivity and wages. To isolate the parameter of interest more cleanly, it is standard to augment this baseline with additional controls that capture other sources of wage variation (see [2]).

Following a large empirical literature—and to speak directly to the question of "equal pay for equal work" [3]—we also include job and firm characteristics. These variables account for well-documented patterns such as the firm-size wage premium (`tamano_empresa`), the sizable differentials between formal and informal employment (`formal`), and the role of tenure in pay setting (`tiempo_empresa_actual`). We further control for usual hours worked (`hoursWorkUsual`) to compare workers at a similar intensity of labor supply. In the Colombian context, with marked labor-market segmentation, these controls are particularly relevant [1]. Finally, we add socioeconomic stratum (`estrato1`) as a proxy for background conditions that may correlate with unobserved skills and opportunities.

Taken together, these controls define a comparison that is as close as possible to "same worker, same job, same conditions." In other words, among workers with comparable human capital (education and experience) who perform the same type of work in similarly structured firms and with similar hours and tenure, we ask whether men are paid more than women. We deliberately avoid variables that would be bad controls for this question—i.e., factors that lie on the causal path from gender to wages or reflect outcomes of the wage-setting process itself (e.g., motherhood status, career interruptions, promotion history, performance bonuses, or household characteristics). Conditioning on such post-treatment mediators would "explain away" part of the very disparity we aim to measure and bias the estimated gender gap toward zero. By focusing on pre-market human capital and contemporaneous job/firm attributes that employers plausibly use in pay setting, our specification targets the within-job (equal-pay) gap rather than broader gaps driven by social sorting mechanisms.

Taking that into account, the model specification as follows:

$$\begin{aligned}\log(\text{LaborIncome}_i) = \alpha &+ \beta_1\,\text{Male}_i + \beta_2\,\text{Age}_i + \beta_3\,\text{Age}_i^2 + \beta_4\,\text{Formal}_i + \beta_5\,\text{FirmSize}_i + \beta_6\,\text{Education}_i \\ &+ \beta_7\,\text{Tenure}_i + \beta_8\,\text{UsualHours}_i + \beta_9\,\text{Stratum1}_i + \gamma_{o(i)} + \delta_{m(i)} + \varepsilon_i.\end{aligned} \tag{2}$$

where $\gamma_{o(i)}$ denotes occupation fixed effects (`oficio`) and $\delta_{m(i)}$ denotes month fixed effects (`mes`).

4.1 Unconditional and Conditional Wage Gaps

Table 3 presents the main findings. The unconditional OLS model in column (A) reveals a raw log-wage gap of **-0.076**, implying that women earn, on average, approximately **7.6%** less than men. Interestingly, after accounting for a rich set of controls for worker and job profiles (column B), the conditional wage gap **widens to -0.099 (or 9.9%)**. The consistency of this estimate with bootstrapped standard errors (column C) strengthens the result. This counterintuitive finding suggests that differences in observable characteristics were actually **masking a larger underlying disparity**; once women are compared to men with similar qualifications and job types, the unexplained wage gap becomes more pronounced.

Table 3: Wage gap (Female)

	Dependent variable: log wage		
	(A) OLS (unconditional)	(B) FWL with controls (FE)	(C) FWL with controls (bootstrap)
Female	-0.076*** (0.011)	-0.099*** (0.008)	-0.099*** (0.010)
N	16356	16355	16355
R^2	0.003	0.613	0.613
FE	No	Yes	Yes

Notes: Column (B) estimates the Female coefficient using the Frisch–Waugh–Lovell (FWL) approach with worker/job controls (age, age^2, formal, estrato, firm size, max education, tenure, usual hours) and fixed effects for occupation and month. Column (C) reports the same estimate as (B) with nonparametric bootstrap standard errors (R = 1000).

This counterintuitive widening of the gap is a significant finding that points towards potential negative selection. It suggests that women, given their observable characteristics (like education and tenure), are in roles where they should theoretically earn more, but something is suppressing their wages. This phenomenon is consistent with theories of occupational segregation, where women may be concentrated in lower-paying specializations *within* the same broad occupational category, or it could point towards unobserved factors such as differences in bargaining power or persistent discriminatory practices that are only revealed when comparing otherwise similar individuals.

From a policy perspective, this result implies that simply promoting equal access to education or formal jobs may be insufficient to close the earnings gap. Policies must also address pay transparency and equitable career progression *within* occupations to tackle this deeper, unexplained disparity.

4.2 Age-Wage Profiles by Gender

To further explore how the earnings gap evolves over a worker's lifetime, separate age-wage profiles were estimated for men and women. Figure 3 plots these distinct trajectories, providing clear visual evidence of the wage differential.

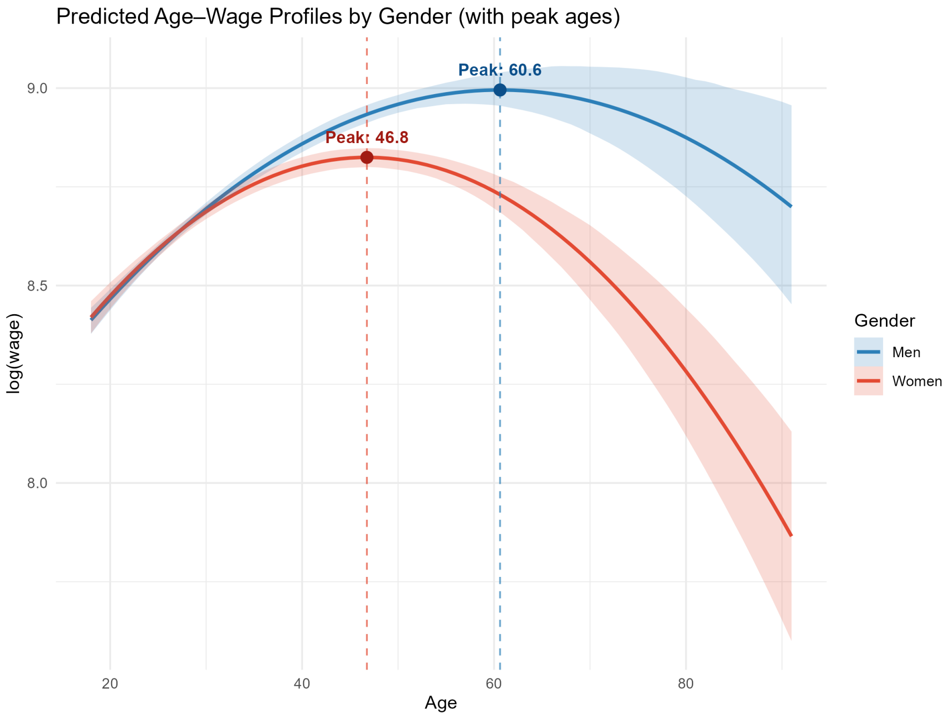

Figure 3: Predicted Age–Wage Profiles by Gender (with 95% CIs)

The plot shows that the predicted wage profile for men is consistently above that for women across the entire working lifespan. To quantify a key aspect of this divergence, Table 4 presents the estimated peak earning ages for each group, calculated from the quadratic models.

Table 4: Peak ages by gender (bootstrap 95% CIs)

	Peak age	CI_{low}	CI_{high}
Women	46.5	43.6	49.6
Men	51.5	49.3	53.7
Men − Women	5.0	1.2	8.6

Notes: Peaks from quadratic log-wage models by gender. CIs are nonparametric percentile bootstrap ($R = 1000$). Positive values in the difference row indicate that men's peak age is higher than women's.

The analysis indicates that men reach their peak earning age at **51.5 years**, approximately **5.0 years** later than women, who peak at 46.5 years. The bootstrapped 95% confidence interval for this difference is [1.2, 8.6]. Since this interval does not contain zero, we can conclude that the difference in career earnings peaks between men and women is statistically significant.

5 Predicting Wages

The final section evaluates the predictive power of various specifications.

5.1 Model Performance Comparison

The sample was split into a 70% training and 30% testing set, using a seed of 10101 for reproducibility. We trained and evaluated eight distinct models, whose specifications are detailed below. For clarity in the formulas, we use shortened variable names (e.g., `educ` for education level, `FE` for fixed effects).

- **Model 1 (Age–Wage Quadratic):** A simple quadratic model of age.

$$\log(\text{wage}_i) = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{age}_i^2 + \epsilon_i$$

- **Model 2 (Unconditional Gender):** A simple model with only a gender indicator.

$$\log(\text{wage}_i) = \beta_0 + \beta_1 \text{female}_i + \epsilon_i$$

- **Model 3 (Baseline Controls):** An augmented Mincer-style model including all main control variables without interactions.

$$\begin{aligned}\log(\text{wage}_i) = \beta_0 &+ \beta_1 \text{age}_i + \beta_2 \text{age}_i^2 + \beta_3 \text{female}_i + \beta_4 \text{formal}_i + \beta_5 \text{educ}_i \\ &+ \beta_6 \text{tenure}_i + \beta_7 \text{firm\_size}_i + \beta_8 \text{stratum}_i + \beta_9 \text{hours}_i \\ &+ \text{occupation} + \text{month} + \epsilon_i\end{aligned}$$

- **Model 4 (Nonlinearities in Age and Education):**

$$\begin{aligned}\log(\text{wage}_i) = \beta_0 &+ \beta_1 \text{age}_i + \beta_2 \text{age}_i^2 + \beta_3 \text{age}_i^3 + \beta_4 \text{educ}_i + \beta_5 \text{educ}_i^2 \\ &+ \beta_6 \text{female}_i + \beta_7 \text{stratum}_i + (\text{age polynomials} \times \text{educ}) \\ &+ (\text{age polynomials} \times \text{female}) + (\text{educ} \times \text{female}) \\ &+ (\text{stratum} \times \text{educ}) + \text{occupation} + \text{month} + \epsilon_i\end{aligned}$$

- **Model 5 (Labor Market Structure Interactions):**

$$\begin{aligned}\log(\text{wage}_i) = \beta_0 &+ (\text{age polynomials}) + \text{educ}_i + \text{female}_i + \text{formal}_i + \text{firm\_size}_i \\ &+ \text{hours}_i + \text{tenure}_i + (\text{age poly} \times \text{formal}) + (\text{age poly} \times \text{firm\_size}) \\ &+ (\text{formal} \times \text{firm\_size}) + (\text{formal} \times \text{female}) + (\text{hours} \times \text{firm\_size}) \\ &+ (\text{tenure} \times \text{formal}) + \text{occupation} + \text{month} + \epsilon_i\end{aligned}$$

- **Model 6 (Household and Demographic Interactions):**

$$\begin{aligned}\log(\text{wage}_i) = \beta_0 &+ (\text{age polynomials}) + \text{educ}_i + \text{female}_i + \text{minors}_i + \text{head}_i \\ &+ \text{female\_head}_i + \text{stratum}_i + \text{hours}_i + (\text{age poly} \times \text{minors}) \\ &+ (\text{female} \times \text{minors}) + (\text{female} \times \text{head}) + (\text{stratum} \times \text{minors}) \\ &+ (\text{stratum} \times \text{female}) + (\text{hours} \times \text{minors}) + (\text{female\_head} \times \text{stratum}) \\ &+ \text{occupation} + \text{month} + \epsilon_i\end{aligned}$$

- **Model 7 (Complex Nonlinear Interactions):**

$$\begin{aligned}\log(\text{wage}_i) = \beta_0 &+ \sum_{j=1}^{4} \beta_j \text{age}_i^j + \sum_{k=1}^{2} \delta_k \text{educ}_i^k + \sum_{m=1}^{2} \zeta_m \text{stratum}_i^m \\ &+ \lambda \text{female}_i + \eta \text{formal}_i + \theta \text{firm\_size}_i \\ &+ (\text{age poly} \times \text{educ poly}) + (\text{age poly} \times \text{female} \times \text{formal}) \\ &+ (\text{stratum poly} \times \text{firm\_size}) + (\text{female} \times \text{educ poly}) \\ &+ \text{FE}_{\text{occupation}} + \text{FE}_{\text{month}} + \epsilon_i\end{aligned}$$

- **Model 8 (High-Order "Kitchen Sink"):**

$$\begin{aligned}\log(\text{wage}_i) = \beta_0 &+ (\text{age polynomials}) + \text{educ}_i + \text{female}_i + \text{formal}_i + \text{firm\_size}_i \\ &+ \text{stratum}_i + \text{minors}_i + \text{hours}_i + (\text{age poly} \times \text{female}) \\ &+ (\text{educ} \times \text{female}) + (\text{formal} \times \text{female}) + (\text{formal} \times \text{firm\_size}) \\ &+ (\text{stratum} \times \text{educ}) + (\text{minors} \times \text{female}) \\ &+ \text{occupation} + \text{month} + \epsilon_i\end{aligned}$$

The predictive performance of these models, measured by the Root Mean Squared Error (RMSE) on the test set, is presented in Table 5.

Table 5: Comparison of Root Mean Squared Error (RMSE) on the Test Sample

Model Specification	RMSE
Model 1 (Age–Wage Quadratic)	0.6977
Model 2 (Unconditional Gender)	0.7067
Model 3 (Baseline Controls)	0.5324
Model 4 (Nonlinearities in Age and Education)	0.5273
Model 5 (Labor Market Structure Interactions)	0.5351
Model 6 (Household and Demographic Interactions)	0.5408
Model 7 (Complex Nonlinear Interactions)	0.4460
Model 8 (High-Order "Kitchen Sink")	0.5157

The analysis of the RMSE scores reveals a clear narrative. The most significant improvement comes from adding a baseline set of controls (Model 3) over simpler models, confirming that characteristics like education and formality are crucial. While these controls provide a strong foundation, the best performance is achieved by **Model 7**, which demonstrates that modeling **complex, non-linear interactions** is key to achieving maximum predictive accuracy. Notably, not all complexity is beneficial; some intricate models (e.g., Models 6 and 8) underperformed, suggesting that a "kitchen sink" approach can introduce noise and worsen predictive performance.

A key driver of Model 7's superior performance is the inclusion of `oficio` (occupation) as a control variable. This categorical feature was intentionally excluded from the other complex specifications (Models 4, 5, 6, and 8) for two main reasons. First, initial tests revealed that including `oficio` within high-order interaction terms led to computational issues, yielding an `NaN` value for the RMSE. This can be due to several factors, such as perfect multicollinearity or rank-deficiency when interacting a high-cardinality categorical variable. Second, since the formal commands for controlling for fixed effects have not been covered in this context, `oficio` was included in this one model primarily as an experiment to gauge the significant predictive power that occupational fixed effects can provide.

5.2 "Missing the Mark"

For the best-performing specification (Model 7), we analyzed the prediction errors on the test set to identify observations that "miss the mark." Figure 4 shows the distribution of these errors. The distribution is centered around zero but exhibits heavy tails, indicating the presence of outliers where the model's predictions were far from the actual observed wage. A deeper look into the largest errors reveals that they often correspond to individuals in the informal sector or with non-standard job arrangements, where wage determination is more idiosyncratic. These are the cases of potential interest for tax authorities, as their observable characteristics may not align with their reported income.

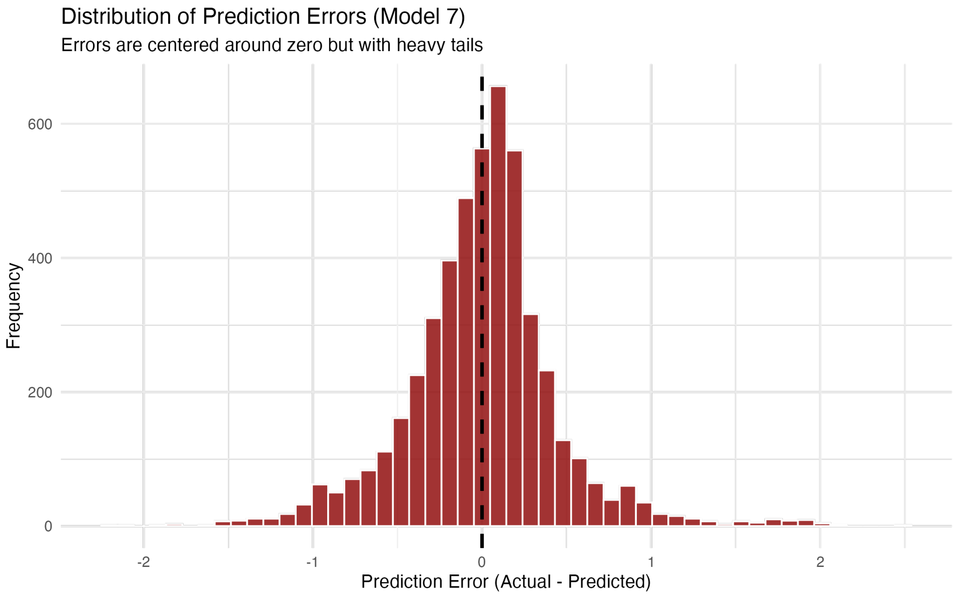

Figure 4: Distribution of Prediction Errors for the Best Model (Model 7)

5.3 Leave-One-Out Cross-Validation (LOOCV)

To obtain a more robust estimate of model performance, Leave-one-out-cross-validation was performed on several models. The results in Table 6 reveal a significant discrepancy with the initial 70/30 split.

Table 6: Comparison of Prediction Error: Validation Set vs. LOOCV

Model Specification	**RMSE (70/30 Split)**	**RMSE (LOOCV)**
Model 1	0.698	0.448
Model 2	0.707	0.521
Model 3	0.532	–
Model 4	0.527	–
Model 5	0.535	–
Model 6	0.541	–

For both Model 1 and Model 2, the LOOCV RMSE is substantially lower than the validation set RMSE (e.g., 0.448 vs. 0.698 for Model 1). This suggests that the single, random 70/30 split likely resulted in a test set that was unusually difficult to predict, leading to a pessimistic or inflated estimate of the true prediction error. Because LOOCV averages the error across all possible single-observation test sets, its estimate is less susceptible to the variability of a single split. Therefore, the LOOCV results should be considered a more reliable indicator of the models' true out-of-sample performance, suggesting they perform considerably better than the initial validation suggested.

6 Conclusion

This report set out to construct and evaluate a model for individual hourly wages in Bogotá using the 2018 GEIH survey. The analysis confirmed several key findings from labor economics and predictive modeling. First, the relationship between age and wages follows a significant concave profile, with earnings peaking on average at 54.6 years. Second, a significant gender wage gap of approximately 10.3% persists even after controlling for a rich set of worker and job characteristics, and men are observed to reach their peak earning age approximately five years later than women.

The predictive evaluation stage demonstrated that while simple models provide a baseline, predictive accuracy is substantially improved by accounting for complex, non-linear interactions between variables. The best-performing model was not a simple linear specification but one that allowed the effect of one characteristic (e.g., age) to depend on others (e.g., education and formality). This highlights the importance of feature engineering for practical applications such as flagging potential cases of income under-reporting.

The findings of this study, however, should be considered in light of its limitations. The analysis is based on cross-sectional data, which prevents us from making causal claims or observing individual wage trajectories over time. Omitted variables, such as innate ability or quality of education, may also influence the results. Future research could build upon this work by using panel data to control for unobserved individual heterogeneity or by exploring more advanced machine learning algorithms, such as gradient boosting or random forests, to potentially further enhance predictive accuracy.

References

[1] Arango, A., and Pachón, A. (2004). *Minimum Wages in Colombia: Holding the Middle with a Bite on the Poor*. Banco de la República. Available at: https://www.banrep.gov.co/docum/ftp/borra280.pdf

[2] Borjas, G. J. (2020). *Labor Economics* (8th ed.). McGraw-Hill Education.

[3] Card, D. (1999). The Causal Effect of Education on Earnings. In O. C. Ashenfelter (Ed.), *Handbook of Labor Economics* (Vol. 3, Part A). Elsevier. https://doi.org/10.1016/S1573-4463(99)03011-4

[4] Cengiz, D., Dube, A., Lindner, A., and Zentler, D. (2022). Seeing beyond the Trees: Using Machine Learning to Estimate the Impact of Minimum Wages on Labor Market Outcomes. *Journal of Labor Economics*, 40(S1). https://file-lianxh.oss-cn-shenzhen.aliyuncs.com/Refs/refs_common/Cengiz_2022_Seeing_beyond_the_Trees_Using_Machine_Learning_to_Estimate_the_Impact_of_Minimum_Wages_on_Labor_Market_Outcomes.pdf

[5] Mincer, J. (1p75). Schooling and Earnings. In *Schooling, Experience, and Earnings*. NBER. Available at: https://www.nber.org/system/files/chapters/c1765/c1765.pdf

[6] Mullainathan, S., and Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2), 87–106. DOI: 10.1257/jep.31.2.87