Problem Set 1: Big Data and Machine Learning for Applied Economics

September 16, 2024

1 Introduction

The following work focuses on the application of basic machine learning and data analysis techniques to predict individual income, taking into account observable variables that may correlate with wages. In this case, the model will be trained and tested using the information available at https://ignaciomsarmiento.github.io/GEIH2018_sample/, which contains data for Bogotá from the 2018 "Medición de Pobreza Monetaria y Desigualdad Report" that utilizes information from the GEIH.

The main objective it´s to construct a model of the form:

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

Where w is the hourly wage, and X it's the matrix of features that will be use to explain the interest variable. In this case, the proposed model will be linear, that means that $f(X)$ follows the form $f(X) = \beta X$, where β it's the vector that will be estimated by training.

Numerous authors have explored wage behavior in Colombia. For instance, Gómez [**?**] identified the city as a pivotal factor in wage inequality, highlighting the significant role that industrial structure plays in shaping urban disparities. Additionally, recognizing the impact of wage estimation on tax policy, researchers like Kugler [**?**] demonstrated that a 10% increase in payroll taxes during the 1980s and 1990s in Colombia resulted in notable reductions in both formal wages and employment, suggesting that reducing taxes could boost demand for lower-skilled workers.

Given that salary is the most critical component of labor compensation and a fundamental element of an organization's strategy [**?**], numerous studies have explored advanced modeling techniques beyond linear approaches, such as Bayesian Gaussian process regression and neural networks [**?**]. For example, Sanabria [**?**] developed a predictive model that leverages natural language processing, specifically utilizing Colombian Twitter data, to enhance salary estimation accuracy.

2 Data

2.1 Describing Data

The data come from the Gran Encuesta Integrada de Hogares (GEIH) conducted by DANE in Colombia. This survey compiles detailed information on the socioeconomic conditions of households and individuals in the country. Specifically, it contains demographic characteristics such as age, gender, education level, marital status, among others, as well as information on labor conditions, income and housing conditions. The dataset is divided into 10 sub-parts, consisting of approximately 32,000 records with 179 different variables. Within them we can find household variables such as its identifier, location, persons belonging to the household, etc.... In addition to individual variables such as those mentioned above.

These data are of great importance when defining relevant variables related to salary, as we find all types of data related to income and labor sector, such as gender, age, working hours and education level. In addition, GEIH data are consistent, complete and accessible when performing Machine Learning or data analysis tasks.

2.2 Describing the Process

Within the data acquisition we found no challenging constraints to get the data. We used 'selenium' a library that automates processes in Python. This automator was combined with a for loop in python, to enter each link related to each subset of data, then by the HTML configuration we could identify that each row corresponded to a line of text, therefore we used a function that took each line of text and saved it as a row in a dataframe, this information is accessed through the XPATH, that is, the characteristics of the web page, this is used in combination with a timer that temporarily stops the process, this with the intention of not identifying that a bot is being used in the web page. At the end each dataframe generated by each link within the loop is concatenated into a general dataframe, this process is repeated for the data dictionary and labels.

2.3 Data Cleaning

In this data cleaning process several processes were taken into account due to the number of variables within the dataset.

1. The context of the problem and the data was taken into account, therefore the dataset was filtered with employed persons over 18 years of age, in addition to eliminating identifiers since they are not necessary when analyzing the data and geographic data since all the data are from Bogota and therefore are not significant variables.

2. The quality of the data is taken in a univariate way, a filtering of variables with an amount of nulls above 30% is performed, that is to say that the columns with an amount greater than 30% in null values are eliminated from the dataset, since they are columns that generate noise and are not complete and consistent for a correct data modeling. In addition, reiterative columns or columns related to the household were eliminated when the problem to be solved is of an individual nature.

3. The last process was multicollinearity issues between predictor variables within the model, it was decided to leave variables that did not have such a strong relationship between them to avoid multicollinearity problems, the first step was to review variables with VIF above 10 to separate them from the dataset and review them individually in a correlation matrix, then those highly correlated variables were removed from the dataset. Since they are redundant variables among themselves, a common problem with the GEIH data.

2.4 Descriptive Analysis

Along with the cleaning steps we have issues of reviewing the relationship between variables with high VIF, in this case we can see how variables are redundant with each other, such as 'formal' and 'informal' as they are opposite columns and it is not necessary that both remain within the dataset. Another example is 'p6210' and 'maxEducLevel' which have the same answer but from different questions.

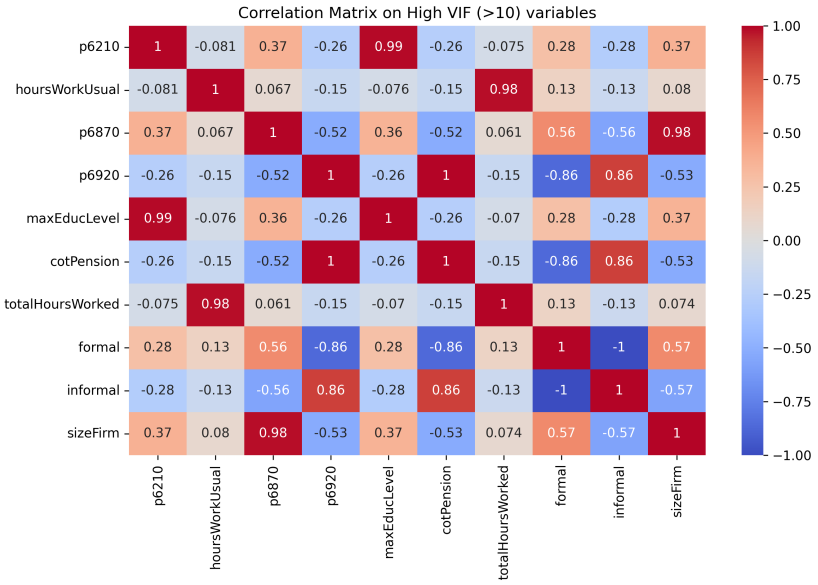

Figure 1: Correlation Matrix with High VIF

We also saw high cardinality in categorical variables, such as 'oficio' where it is seen that so many categories would generate noise within the data modeling and high bias to the results.

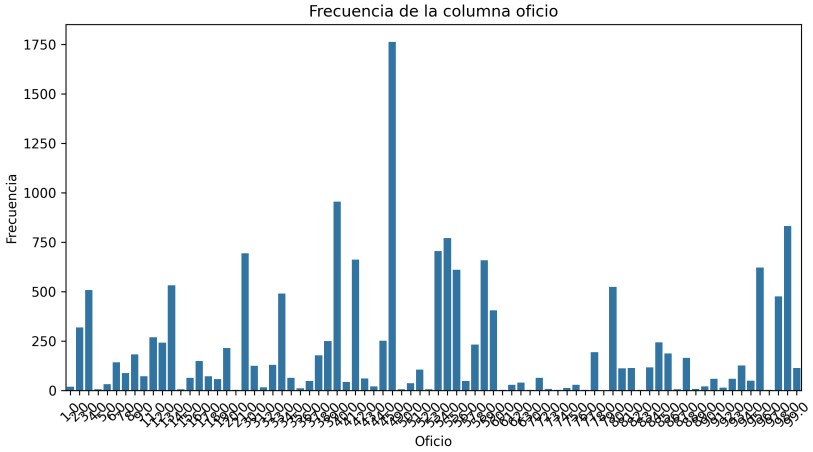

Figure 2: High Cardinality in 'oficio' variable

Finally, our target variable has certain descriptive statistics that indicate that it may have certain outliers, it has an average too far from the median, also the IQR is small compared to the standard deviation, therefore our target variable is treated in a way that takes the data up to the 95th quartile, as can be seen in Figure 3.

Table 1: Descriptive Statistics for y_salary_m_hu

Statistics	Value
Number of Observations	9892
Mean	7945.17
Standard Deviation	11607.18
Minimum	151.00
25% Percentile	3797.00
Median (50%)	4475.50
75% Percentile	7291.00
Maximum	291666.00

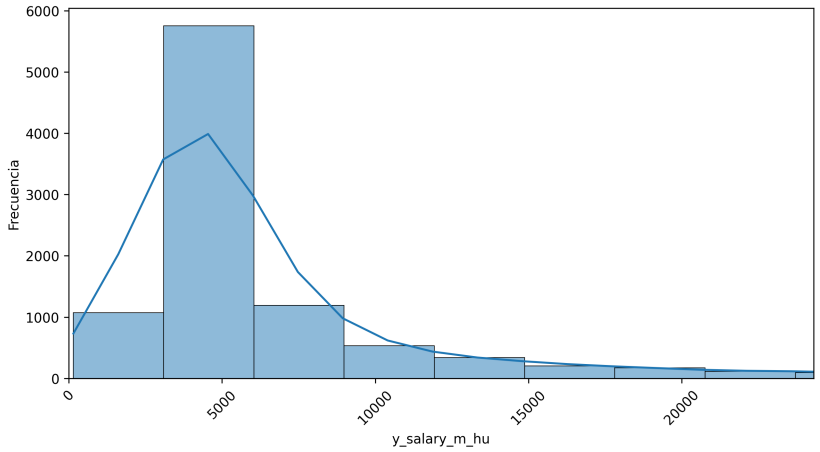

Figure 3: Distribution of the target variable

3 Age-wage profile

In this section, an estimation of a linear regression model will be carried out, following the form:

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

Where w corresponds to the wage obtained and Age indicates the person's age. For this purpose, the Python programming language was used with the previously mentioned dataset, with the exception that the entirety of the observations was utilized, considering that we are working with a single variable. Using the sklearn library, it was possible to perform the regression through the ordinary least squares method, with the following results:

Dep. Variable:	log_wage	**R-squared:**	0.047
Model:	OLS	**Adj. R-squared:**	0.047
Method:	Least Squares	**F-statistic:**	244.4
Date:	Sun, 15 Sep 2024	**Prob (F-statistic):**	2.51e-104
Time:	09:15:57	**Log-Likelihood:**	-10718.
No. Observations:	9964	**AIC:**	2.144e+04
Df Residuals:	9961	**BIC:**	2.146e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P> |t|	[0.025	0.975]
const	7.2921	0.066	111.062	0.000	7.163	7.421
age	0.0650	0.004	18.513	0.000	0.058	0.072
age_2	-0.0007	4.36e-05	-16.231	0.000	-0.001	-0.001

Note that the table shows that the model was estimated using the OLS method, with a total of 9964 observations. In particular, for this type of model, it is important to focus on the results that indicate the model's performance, such as R^2, AIC, and BIC. For these values, we see that the fit does not seem to perform well, indicating that it is not possible to make a good prediction of the wage by only considering age and the quadratic term of age. This is likely due to the model having a significant bias.

The coefficients obtained are $\beta_0 = 7.2921$, $\beta_{Age} = 0.0650$, and $\beta_{Age^2} = -0.0007$. This indicates that most of the weight in the regression is carried by the first two terms (the intercept and age). Specifically, an increase of one year in age is accompanied by only a small increase from the terms associated with age. This behavior suggests that the model is prioritizing the logarithmic wage average, implying that the null model (or a model based on the average) may provide a better fit for the data. This could occur because age alone, even when considering its quadratic term, is not necessarily a strong predictor for explaining wages in this model.

Although the metrics have shown that the fit is not the best, studies have been proposed that suggest that: "Wages tend to be low when the worker is young; they rise as the worker ages, peaking at about age 50; and the wage rate tends to remain stable or decline slightly after age 50". Therefore, the predicted wage by the model was plotted against the real data for an age interval. The results are shown below:

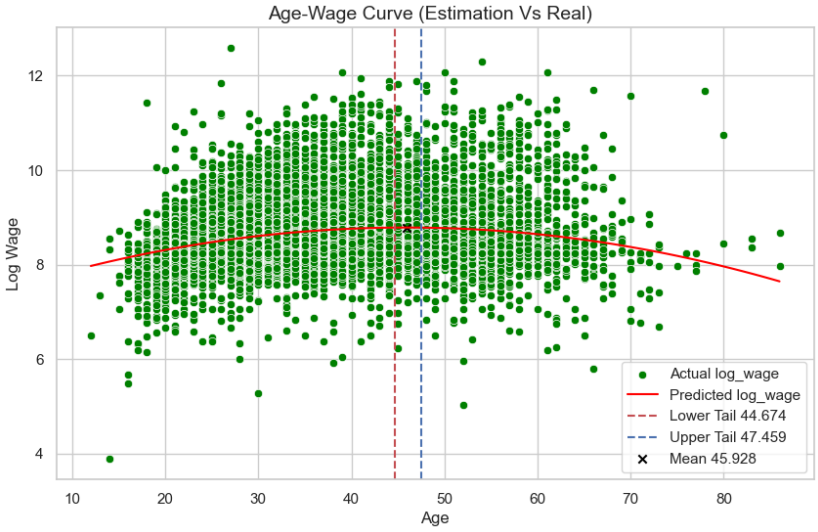

Figure 4: Age-Wage Curve

An estimation of a confidence interval was conducted to determine the peak age at which the highest salary is obtained, as shown in the previous figure. This estimation was performed using the Bootstrap method with 10,000 iterations. The peak age was derived by differentiating the regression equation with respect to age in order to find the maximum value.

In this way, we find that the age at which the maximum expected salary is reached corresponds to 45.928 years, with a lower bound of 44.674 years and an upper bound of 47.459 years. Therefore, our model exhibits the expected behavior commonly seen in the literature. However, this does not imply that accurate predictions can be made using age as the sole variable.

4 The gender earnings

In this section, we analyze the gender wage gap using both an unconditional and conditional regression model. We first estimate the unconditional wage gap, and then include worker and job characteristics to observe how the gap changes.

We begin by estimating the following equation:

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

Where Female is an indicator variable equal to 1 if the individual is identified as female.

Next, we estimate a conditional earnings gap model by incorporating controls such as occupation, education level, firm size, and age. The model is:

$$\log(w) = \beta_1 + \beta_2 \text{Female} + \beta_3 \text{Occupation} + \beta_4 \text{Education} + \beta_5 \text{MicroFirm} + \beta_6 \text{Age} + \beta_7 \text{Age}^2 + u \quad (4)$$

Where:

- Occupation: a categorical variable representing the individual's occupation, with reference to "Chemists, Physicists (Professionals and Technicians)".
- Education: the highest level of education achieved, with reference to "none".
- MicroFirm: a dummy variable that takes the value 1 if the firm has 5 or fewer employees, 0 otherwise.
- Age and Age^2: capture the nonlinear relationship between age and salary.

Table 2: Models comparation

	Dependent variable:	
	log_salario	
	Simple model	Conditional model
	(1)	(2)
sex	-0.0449^{***}	-0.0853^{***}
	(0.0145)	(0.0122)
Constant	8.6412^{***}	8.3974^{***}
	(0.0102)	(0.1716)
Observations	9,891	9,891
R^2	0.0010	0.5225
Adjusted R^2	0.0009	0.5182
Residual Std. Error	0.7214 (df = 9889)	0.5009 (df = 9803)
Note:		*p<0.1; **p<0.05; ***p<0.01

The results of the analysis, including the estimated coefficients from both the unconditional and conditional models, are presented in Table 2. The primary focus is on the coefficient β_2, which corresponds to the variable "Female." In the unconditional model, the coefficient on "Female" is -0.0449, indicating that women earn approximately 4.49% less than men without controlling for any other factors. This difference is statistically significant, with a p-value less than 0.01. However, when additional controls such as job and worker characteristics are included in the conditional model, the coefficient for "Female" becomes -0.0853. This suggests that women earn about 8.53% less than men, implying that the gender wage gap widens once these characteristics are accounted for. Again, the result is statistically significant with a p-value below 0.01.

This widening suggests the presence of a selection issue, as women might be underrepresented in higher-paying roles or industries. Furthermore, the persistent negative coefficient on the "Female" variable even after controlling for observable characteristics hints at the possibility of a discrimination problem, where women are paid less than men for comparable roles and qualifications. Therefore, it is likely that both selection and discrimination contribute to the gender wage gap observed in the data.

Comparing the two models reveals that the unconditional model has a very low explanatory power, with an R-squared of only 0.0010. In contrast, the conditional model shows much better explanatory power, with an R-squared of 0.5225. This improvement indicates that variables such as occupation, education, firm size, and age play a significant role in explaining wage differences. The inclusion of these additional factors in the conditional model also leads to a lower residual standard error, which further demonstrates a better in-sample fit.

Figure 5: Peak ages by gender

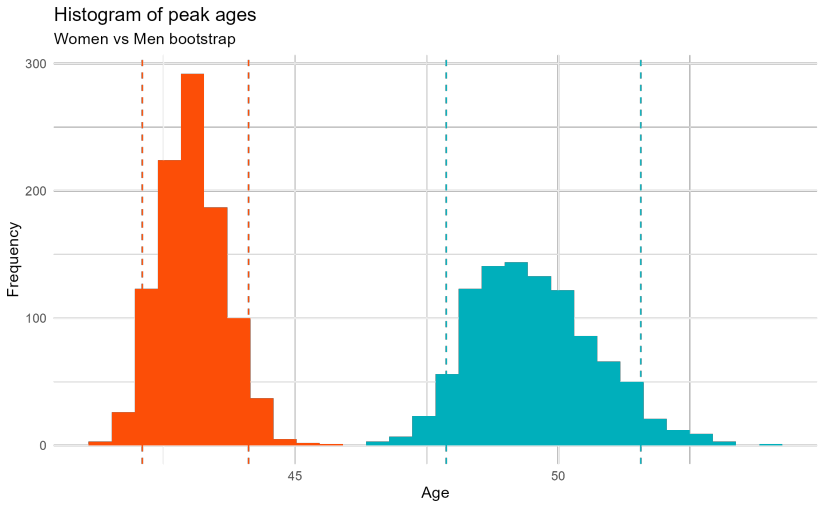

Figure 6: Peak ages by gender bootstrap histograms

On the other hand, the results of the bootstrap analysis for peak ages in men and women provide the following statistics: for men, the original peak age estimate is 49.36 years, with a bias of 0.178137 and a standard error of 1.164219. For women, the original peak age estimate is 43.02 years, with a bias of 0.0277057 and a standard error of 0.618308.

These findings suggest that men and women reach their highest earning potential at different stages of their careers, with men reaching their peak around 6 years later than women. The relatively low standard errors indicate that the estimates are precise. However, this difference in peak ages might reflect several potential issues. One possible explanation is that men and women follow different career trajectories due to factors such as career breaks for caregiving, which may disproportionately affect women. Another possible explanation is the existence of discrimination in terms of career advancement opportunities, which might cause women to experience stagnation in their earnings at earlier ages.

5 Predicting earnings

To evaluate the performance of the different estimates, the data was divided into two parts. In the first part, the data was used for training and the second for testing. This was done by creating a seed with the *set.seed(123)* command in R

The following models were estimated. The first 3 models correspond to estimates made in previous sections.

$$\log(w) = \beta_1 + \beta_2 \cdot \text{Age} + \beta_3 \cdot \text{Age}^2 + u \tag{5}$$

$$\log(w) = \beta_1 + \beta_2 \cdot \text{Female} + u \tag{6}$$

$$\hat{u}_{\log(w)} = \beta_2 \hat{u}_{\text{Female}} + \hat{v} \tag{7}$$

$$\begin{aligned}\log(w) = \beta_1 + \beta_2 \cdot \text{Female} + \beta_3 \cdot \text{MaxEducLevel} + \beta_4 \cdot \text{Occupation} + \\ \beta_4 \cdot \text{MicroEntreprise} + \beta_5 \cdot \text{Age} + \beta_6 \cdot \text{Age}^2 + u\end{aligned} \tag{8}$$

$$\begin{aligned}\log(w) = \beta_1 + \beta_2 \cdot \text{Female} + \beta_3 \cdot \text{MaxEducLevel} + \\ \beta_4 \cdot \text{TotalHoursWorked} + \beta_5 \cdot \text{Age} + \beta_6 \cdot \text{Age}^2 + u\end{aligned} \tag{9}$$

$$\begin{aligned}\log(w) = \beta_1 + \beta_2 \cdot \text{Female} + \beta_3 \cdot \text{MaxEducLevel} + \\ \beta_4 \cdot \text{TotalHoursWorked} + \beta_5 \cdot \text{Age} + \beta_6 \cdot \text{Age}^2 + \beta_6 \cdot \text{Occupation}^2 + u\end{aligned} \tag{10}$$

$$\begin{aligned}\log(w) = \beta_1 + \beta_2 \cdot \text{Female} + \beta_3 \cdot \text{MaxEducLevel} + \beta_4 \cdot \text{TotalHoursWorked} \\ +\beta_5 \cdot \text{Age} + \beta_6 \cdot \text{Age}^2 + \beta_6 \cdot \text{Occupation}^2 + \beta_7 \cdot \text{Age} \cdot \text{Female} + \beta_6 \cdot \text{Age}^2 \cdot \text{Female} + u\end{aligned} \tag{11}$$

$$\begin{aligned}\log(w) = \beta_1 + \beta_2 \cdot \text{Female} + \beta_3 \cdot \sum_{m=1}^{3} \beta_4^m Age_i^m + \sum_{m=0}^{3} \beta_5^m (MaxEducLevel_i \times Age_i^m) + \\ \beta_5 \cdot \text{TotalHoursWorked} + \beta_6 \cdot \text{TotalHoursWorked} \cdot \text{Female} + \beta_7 \cdot \text{Occupation} + \\ \beta_8 \cdot \text{Occupation} \cdot \text{Female} + + \beta_9 \cdot \sum_{m=1}^{3} \beta_{10}^m Age_i^m + \sum_{m=0}^{3} \beta_{11}^m (Female_i \times Age_i^m) + u\end{aligned} \tag{12}$$

$$\begin{aligned}\log(w) = \beta_1 + \beta_2 \cdot \text{Female} + \sum_{m=1}^{3} \beta_3^m Age_i^m + \sum_{m=0}^{3} \beta_4^m (MaxEducLevel_i \times Age_i^m) + \\ \sum_{m=0}^{2} \beta_5^m (Female_i \times TotalHoursWorked_i^m) + \beta_6 \cdot \text{Occupation} + \\ \beta_7 \cdot \text{Occupation} \cdot \text{Female} + \beta_8 \cdot \sum_{m=1}^{3} \beta_9^m Age_i^m + \sum_{m=0}^{3} \beta_{10}^m (Female_i \times Age_i^m) + u\end{aligned} \tag{13}$$

$$\begin{aligned}\log(w) = \beta_1 + \beta_2 \cdot \text{Female} + \sum_{m=1}^{3} \beta_3^m Age_i^m + \sum_{m=0}^{4} \beta_4^m (MaxEducLevel_i \times Age_i^m) + \\ \sum_{m=0}^{3} \beta_5^m (Female_i \times TotalHoursWorked_i^m) + \sum_{m=0}^{3} \beta_6^m (Occupation_i \times TotalHoursWorked_i^m) \\ + \sum_{m=0}^{3} \beta_7^m (Occupation_i \times MaxEducLevel_i^m) + \beta_8 \cdot \text{Occupation} \cdot \text{Female} \\ +\beta_9 \cdot \sum_{m=1}^{3} \beta_{10}^m Age_i^m + \sum_{m=0}^{3} \beta_{11}^m (Female_i \times Age_i^m) + u\end{aligned} \tag{14}$$

Five additional models are estimated to provide flexibility to the model, since polynomial functions and interactions with the variables considered relevant are estimated. However, it is recognized that very complex models can directly affect the variance of the model.

Eq. 5	Eq. 6	Eq. 7	Eq. 8	Eq. 9	Eq. 10	Eq. 11	Eq. 12	Eq. 13	Eq. 14
0.5031	0.5093	0.7887	0.2586	0.3438	0.2595	0.2594	0.2562	0.2555	0.3408

Table 3: MSE of Equations 5 to 14

The table 3 presents the results of estimating the Mean Square Error for each of the previous specifications in the test data. It can be observed that the model of equation 11 is the one with the lowest MSE. For this reason, the distribution of the errors in this model will be explored.

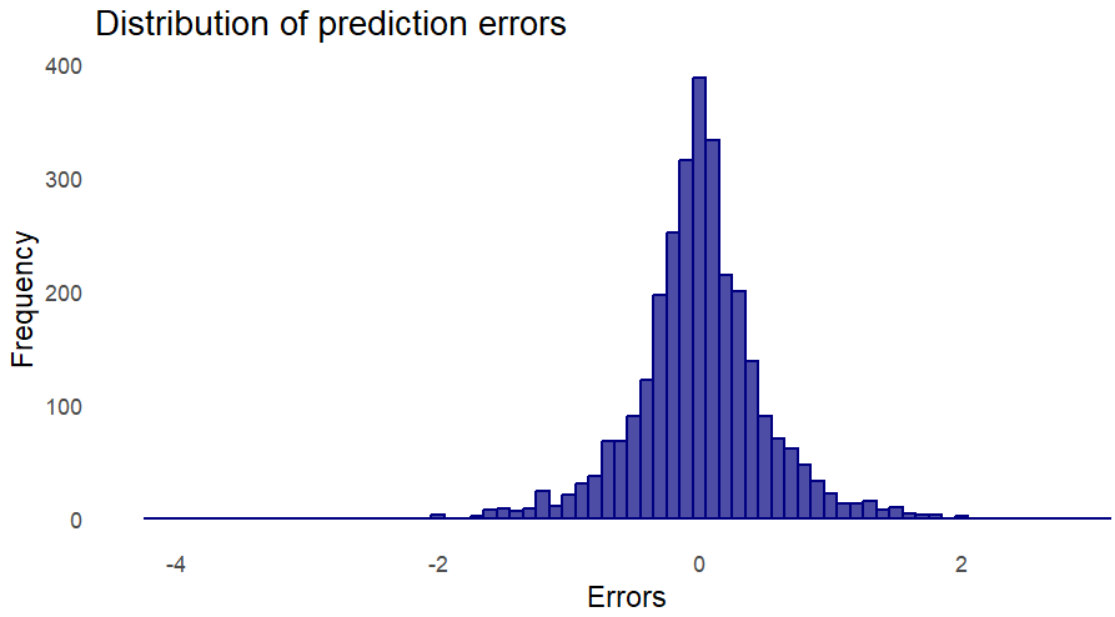

Figure 7: Distribution of prediction errors

The errors behave normally, however an outlier can be seen in the left tail of the graph. This indicates that the model did not predict this observation. This can happen for different reasons, such as unobserved variables that directly affect that individual, an overfitting or underfitting of the model, an individual behaving differently than his peers. Given that only one observation in particular is observed that deviates from normal behavior, it would be relevant to expand this information to understand in depth the reasons why the explanatory variables do not allow explaining his wage.

Using more data to train the model can be useful to reduce bias. For this reason, the Leave One Out technique is used, which consists of training the model with the n-1 data and testing it with the missing observation. In the following table, the MSEs of the best two specifications consisting of equation 10 and 11 are presented.

Eq. 10	Eq. 11
0.5001	0.4756

Table 4: MSE of the two best specifications

In the table 4 it is observed that the MSE changes for the two models, this increase for both models. This change occurs as LOOCV allows the model to train better on the data and allows for a less biased model. However, it is important to highlight that LOOCV has less bias, but greater variance due to the multiples estimations that are performed. For this reason, other approaches such as k fold validation are recommended