Modeling Wage Determinants
A Multivariate Analysis of CPS Data

Aleksandr Dudakov, Benjamin Markert, Sebastian Thoning Hofland

2024-12-19

Abstract

This paper examines wage determination using data from the Current Population Survey's Outgoing Rotation Group (CPS ORG) for January 2022. A number of statistical models are employed to ensure a robust analysis. In addition to classical Ordinary Least Squares (OLS) and OLS with robust standard errors, the study incorporates advanced robust regression techniques, including Huber's M-estimator, the MM-estimator with Peña-Yohai initialization, and Least Trimmed Squares (LTS). The analysis also uses the .632 bootstrap method to increase reliability. The results indicate that the simple OLS model with robust standard errors provides sufficient explanatory power. The results reproduce important stylized facts about wage formation, such as the role of age, gender, and education. Finally, the study applies a Linear Mixed Effects model and finds that about a quarter of the wage variance is due to unobserved household-level factors.

University of Milan
Department of Economics, Management and Quantitative Methods (DEMM)
Advanced Multivariate Statistics
Prof. Andrea Cappozzo

UNIVERSITÀ
DEGLI STUDI
DI MILANO

DEMM

Table of Contents

1 Introduction

Understanding the determinants of wages is a central topic in labor economics, as they are a key component of economic inequality, living standards, and economic growth. In simple textbook models, wages are often viewed as a function of education or experience, but in reality, the factors influencing wages are far more complex. Workers with comparable qualifications and experience often experience a significant difference in their incomes, which can be attributed to various factors such as industry or job type.

This disparity makes the analysis of wage determinants an ideal subject for exploration, using multivariate statistical techniques. Modeling the determinants of wages can provide meaningful insights to policymakers, researchers, as well as workers and employers. The disparity in wages is not solely based on individual characteristics, but reflects deeper structural inequalities existing within the labor market, as shown by the persistent gender and racial wage gap.

This study uses data from the Current Populations Survey's Outgoing Rotation Group (CPS ORG) to explore the underlying factors that drive wage disparities. By examining a range of demographic variables alongside employment characteristics, the analysis seeks to uncover patterns that influence wage levels. Moreover, by incorporating regional aspects, the analysis seeks to provide insights into how differences in geographic location and local labor markets influence wage determination.

Wage determination and income inequality has been a widely studied topic over the years, yet several themes remain open for further exploration. One of the most important contributions to this field came in 1958 from Polish economist Jacob Mincer, who developed a model that expanded on simpler income models based on human capital, leading to what is now known as the Mincer Earnings Function (Mincer, 1958). This established a foundational framework for understanding the relationship of wages and human capital. Despite these advancements, there remain several open themes in the field, especially regarding the impact of unobserved factors at the individual or household level, and how these contribute to personal income disparities. Furthermore, the contribution of occupation and industry in explaining wage differences is an area of active investigation.

Throughout this analysis, several multivariate statistical methods are applied with the aim of answering the following questions:

1. What are the key individual-level factors affecting personal income?

2. To what extent do occupation, industry and regional factors contribute to wage variations?

3. How do unobserved household-level factors impact individual wage determination?

2 Data Description

The CPS is a monthly survey conducted by the U.S. Census Bureau and the Bureau of Labor Statistics (BLS) designed to collect comprehensive data on the characteristics of the labor force. The ORG component focuses on collecting detailed information on workers' earnings, making it particularly well suited for income analysis.

The CPS uses a rotating panel design in which households are interviewed for four consecutive months, are out of the sample for eight months, and then reenter the sample for another four months. The fourth and eighth interviews constitute the outgoing rotation groups. During these interviews, additional questions are asked to obtain detailed data on earnings.

For this analysis, we use CPS ORG data for January 2022, focusing on employed individuals aged 16 and older. The data were accessed using the `epiextractr` package in `R`, which facilitates the process of downloading and loading microdata extracts from the Economic Policy Institute (EPI). The EPI CPS ORG extracts are restricted to individuals with a positive earner sample weight and who are in the outgoing rotation months, ensuring a consistent and reliable sample.

Several data transformations and preprocessing steps were performed to prepare the dataset for analysis. The variable `age` was originally coded with a maximum category of *80+*, which contained 71 observations. This category was recoded to 80 to represent age as an integer variable. Given the small number of observations in this category, this change should not significantly affect the results.

The educational attainment variable was recoded from a categorical levels to a numeric variable representing years of education. This mapping is approximate, as detailed in Table 1, and allows for quantitative analysis of education as a predictor variable.

Table 1: Educational Attainment Categories and Corresponding Years of Education

Description	**Years of Education**
Less than 1st grade	0
1st–4th grade	2
5th–6th grade	5
7th–8th grade	7
9th grade	9
10th grade	10
11th grade	11
12th grade, no diploma	12
High school graduate, GED	12
Some college, no degree	13
Associate degree, occupational/vocational	14
Associate degree, academic program	14
Bachelor's degree	16
Master's degree	18
Professional school degree	19
Doctorate degree	21

To ensure that the analysis focused on individuals with meaningful wage data, we filtered the dataset to include only employed individuals and excluded those without wages (a level of job class). An initial assessment revealed missing values for important variables, particularly wage (12.8% missing) and metropolitan status (0.788% missing). To address this, a two-step imputation process was implemented.

First, for individuals with missing wage or metropolitan status in January 2022, we imputed these values using data from other months (2021-2023) for the same individuals. We matched these individuals to their records from other months and selected the closest prior or subsequent non-missing value based on the minimum time difference, taking advantage of the rotating panel design of the CPS. After this step, the missing rate of wage was reduced to 10.3%, while metropolitan status stayed the same with 0.788%.

For the remaining missing values, we used hot deck imputation, a method recommended by the U.S. Census Bureau. This technique replaces missing values with observed responses from similar respondents defined by variables such as age, race, gender, and education. Because the missing rate was not substantial and wages typically do not vary dramatically over short periods of time, we believe that this imputation should not significantly affect the analysis. After hot deck imputation, the missing rates further decreased to 1.07% for wage (131 observations) and 0.123% for metropolitan status (15 observations). These remaining missing values, along with 7 observations where wage was zero, were dropped from the dataset. The deletions are minimal and should not affect the

results of the study.

Additional cleaning steps included excluding unused levels in categorical variables. In the job class variable, the level *Without pay* was already excluded during filtering. In the industry variable, the level *Armed Forces* was removed because it had zero observations. Unused levels in categorical variables were dropped to ensure accuracy in statistical modeling.

The final dataset consists of 12,037 observations and includes the variables described in Table 2, with all levels of categorical variables specified.

Table 2: Description of Variables in the Final Dataset

Variable	**Description**
`age`	Age of the respondent in years (16–80).
`sex`	Sex of the respondent: *Male*, *Female*.
`citizen`	Citizenship status: *Not a US citizen*, *US citizen*.
`race`	Race/ethnicity: *White*, *Black*, *Hispanic*, *Asian*, *Other*.
`married`	Marital status: *Not married*, *Married*.
`metropolitan`	Metropolitan status: *Nonmetropolitan*, *Metropolitan*.
`division`	Census geographic division: *New England*, *Middle Atlantic*, *East North Central*, *West North Central*, *South Atlantic*, *East South Central*, *West South Central*, *Mountain*, *Pacific*.
`education`	Years of education attained (numeric, approximate mapping).
`job_class`	Class of worker: *Government - Federal*, *Government - State*, *Government - Local*, *Private, for profit*, *Private, nonprofit*, *Self-employed, incorporated*, *Self-employed, unincorporated*.
`industry`	Major industry of employment: *Agriculture, forestry, fishing, and hunting*, *Mining*, *Construction*, *Manufacturing*, *Wholesale and retail trade*, *Transportation and utilities*, *Information*, *Financial activities*, *Professional and business services*, *Educational and health services*, *Leisure and hospitality*, *Other services*, *Public administration*.
`wage`	Hourly wage in dollars, including overtime, tips, and commissions.
`education_level`	Original educational attainment categories (see Table 1).
`hhid`	A household identifer.

3 Methodology

3.1 Exploratory Data Analysis

The exploratory data analysis (EDA) aims to understand the characteristics of the dataset and identify patterns that will inform the modeling process. The dataset consists of 12,037 observations from the January 2022 CPS data, focusing on employed individuals.

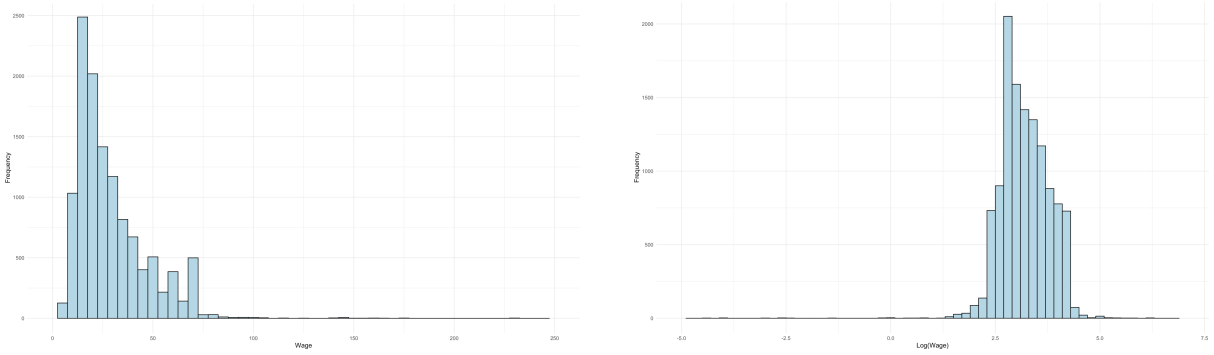
A log transformation was applied to the `wage` variable to account for right skewness and to stabilize the variance.

Table 3 presents summary statistics for the continuous variables. The `wage` variable has a mean higher than the median, indicating right skewness, and a large standard deviation relative to the mean. The `log_wage` variable is less skewed, suggesting that the log transformation normalizes the distribution.

Table 3: Summary Statistics for Continuous Variables

Variable	**Mean**	**SD**	**Median**	**Min**	**Max**	**Skewness**
Age	43.35	14.67	42.00	16.00	80.00	0.20
Education (years)	14.20	2.75	14.00	0.00	21.00	−0.20
Wage ($)	29.55	28.20	23.81	0.0002	1,442.31	27.40
log wage	3.21	0.63	3.17	−8.41	7.27	−2.52

Figure 1a shows the histogram of `wage` and Figure 1b shows `log_wage`. The wage distribution is highly right-skewed, with most observations clustered at lower wage values and a long tail extending to higher wages. The bulk of workers receive a relatively low wage (less than 20$ in an hour), while a small share of people receive a disproportionate high wage. The log transformation effectively reduces the skewness, resulting in a more symmetric distribution that approximates normality.



(a) Histogram of Wage (b) Histogram of log(Wage)

Figure 1: Histogram of wage and log(Wage)

To examine the relationship between `log_wage` and categorical variables, boxplots were constructed and are presented in Figure 2. The boxplots illustrate how median wages and wage variability differ across categories of key variables such as `industry`, `job_class`, `education_level`, `race`, and `sex`.

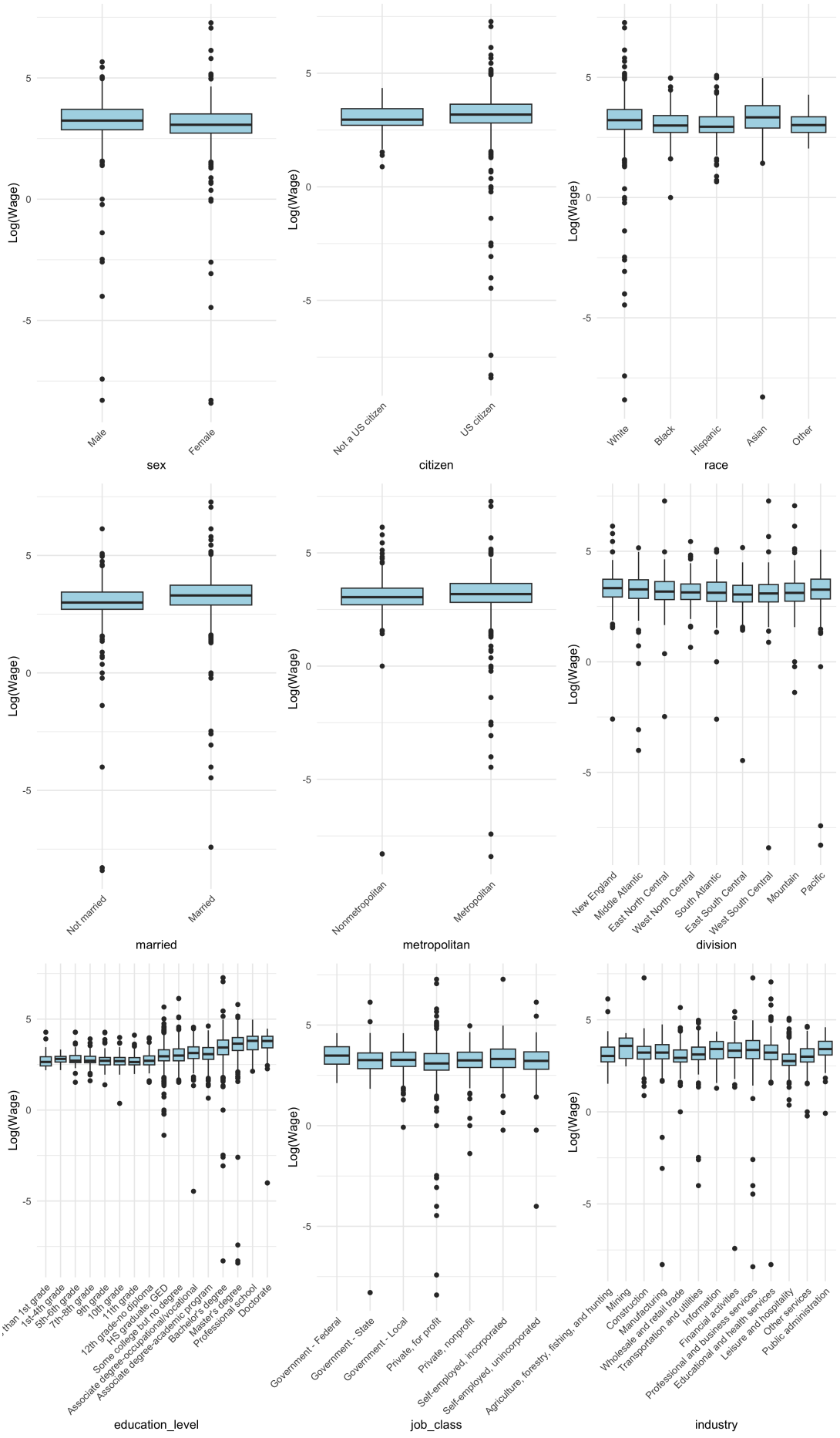
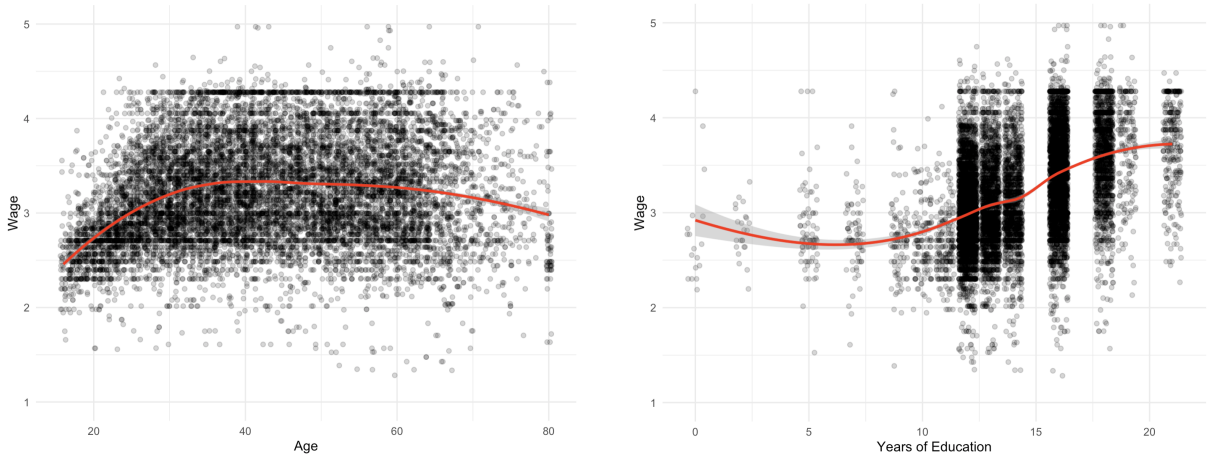


Figure 2: Boxplots of log(Wage) by Categorical Variables

The boxplots reveal that industries such as *Professional and Business Services* and *Information* have higher median log wages, while *Leisure and Hospitality* exhibits the lowest median wage. The category Job Class seems distorted by outliers. This is especially evident when looking at the job class *Private, for profit*, indicating the vague definition of the different job classes. Higher education levels correspond to higher median wages, demonstrating the positive association between education and earnings. Here the possible presence of outliers in some of the high-education categories is also evident. Males have higher median wages than females, indicating a potential gender wage gap. Racial disparities are also evident, with Asian and White individuals earning higher median wages compared to Black and Hispanic individuals.

Scatter plots of `log_wage` versus `age` and `education` are shown in figures 3a and 3b, respectively (the 37 most extreme outliers have been excluded for visual clarity). The plots include LOESS smoothing curves. Wages tend to increase with age up to about age 40 and then plateau or decline slightly, reflecting typical career progression. There is a positive correlation between education and wages, with wages increasing significantly after high school (12 years of education) and accelerating further at higher levels of education.

Another notable observations is that the wage distribution exhibits clustering at specific values, likely due to rounding in survey responses, as individuals often report approximate figures rather than precise earnings. Additionally, employers frequently set wages at round numbers, which further contributes to the observed pattern. Also there could be an effect from institutional factors such as certain tax-brackets and unionized negotiations which means that many people earn the same.



(a) Scatter Plot of Wage vs. Age

(b) Scatter Plot of Wage vs. Education

Figure 3: Scatter plots

A correlation analysis was conducted to assess relationships among numeric variables

and to detect potential multicollinearity issues. The correlation matrix is presented in Table 4. The correlation between `education` and `log_wage` is moderate (0.39), indicating that higher education is associated with higher wages. Other correlations are relatively low, suggesting that multicollinearity is not a significant concern.

Table 4: Correlation Matrix of Numeric Variables

	Age	**Education**	**log(Wage)**	**Female**	**Citizen**	**Married**	**Metro**
Age	1.00	0.07	0.13	−0.01	0.05	0.32	−0.05
Education	0.07	1.00	0.39	0.08	0.16	0.16	0.09
log(Wage)	0.13	0.39	1.00	−0.12	0.06	0.19	0.08
Female	−0.01	0.08	−0.12	1.00	0.06	−0.05	−0.00
Citizen	0.05	0.16	0.06	0.06	1.00	−0.03	−0.08
Married	0.32	0.16	0.19	−0.05	−0.03	1.00	−0.02
Metropolitan	−0.05	0.09	0.08	−0.00	−0.08	−0.02	1.00

Variance Inflation Factors (VIFs) were calculated to further assess multicollinearity, with all VIF values being below 5. This confirms that multicollinearity is not a significant issue in the dataset.

In summary, the EDA suggests key predictors such as `education`, `industry`, `job_class`, `race`, and `sex` show significant associations with `log_wage`, justifying their inclusion in the modeling process. The absence of strong multicollinearity among the predictors ensures reliable coefficient estimates in the regression models. The presence of outliers and possible heteroscedasticity suggest that robust regression methods and advanced modeling techniques may be beneficial in the analysis.

3.2 Model Building and Analysis

In this section, we model the determinants of personal income using a variety of statistical techniques. We begin with Ordinary Least Squares (OLS) regression as a baseline model to examine the relationships between the log-transformed wage (`log_wage`) and the predictor variables. To address potential violations of OLS assumptions, such as heteroscedasticity, non-normality, and the presence of outliers, we employ robust regression methods. In addition, we use Linear Mixed-Effects (LME) models to account for hierarchical structures and unobserved heterogeneity at the household level.

Data Splitting and Evaluation Approach For model evaluation, we adopt a validation set approach by randomly splitting the dataset into training and test sets in a

70%/30% ratio. This allows us to evaluate the predictive performance of the models on unseen data. The training set consists of 8,425 observations and the test set contains 3,612 observations.

To evaluate the performance of the models, we calculate the Mean Squared Error (MSE) on the test set. The test MSE is calculated as

$$\text{Test MSE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left(\log \text{wage}_i - \log \widehat{\text{wage}}_i \right)^2,$$

where n_{test} is the number of observations in the test set, $\texttt{log\_wage}_i$ is the actual log wage, and $\texttt{log\_}\hat{\texttt{wage}}_i$ is the predicted log wage.

In addition, we use the .632 bootstrap method to compute the test MSE for both OLS and robust regressions in order to assess the reliability of their generalization error.

3.2.1 Ordinary Least Squares Regression

Model Specification Several model specifications were explored to determine the most appropriate model for the data. The final OLS model was selected based on the statistical significance of the predictors, goodness-of-fit measures, and theoretical considerations. The dependent variable is `log_wage`.

The model is specified as

$$\begin{aligned} \texttt{log\_wage}_i = \beta_0 &+ \beta_1 \texttt{age}_i + \beta_2 \texttt{age}_i^2 + \beta_3 \texttt{sex}_i + \beta_4 \texttt{citizen}_i \\ &+ \beta_5 \texttt{race}_i + \beta_6 \texttt{married}_i + \beta_7 \texttt{metropolitan}_i + \beta_8 \texttt{division}_i \\ &+ \beta_9 \texttt{education}_i + \beta_{10} \texttt{job\_class}_i + \beta_{11} \texttt{industry}_i + \varepsilon_i, \end{aligned}$$

where ε_i are the error terms.

Estimation Results The OLS model was estimated on the training data. Table 8 presents the estimated coefficients, with standard errors in parentheses below the estimates, along with statistical significance levels. The results are reported in column (1), which shows the standard OLS estimates.

The adjusted R^2 of the model is 0.2366, indicating that approximately 23.66% of the variability in `log_wage` is explained by the model. The test Mean Squared Error (MSE) of the model, estimated on the validation set, is 0.3150.

Model diagnostics We assessed the OLS model assumptions through diagnostic plots and statistical tests.

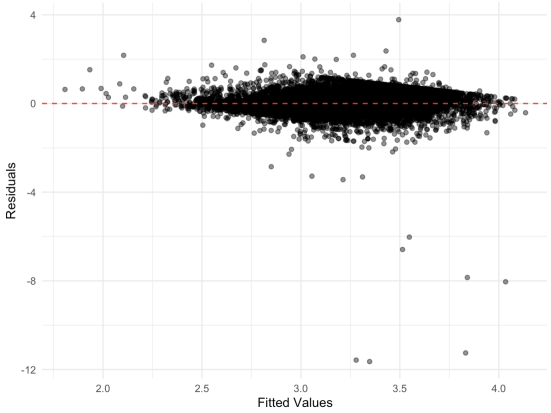
- **Residuals vs. Fitted Values** (Figure 4a): Residuals are centered around zero but have a slight funnel shape, indicating mild heteroscedasticity.
- **Scale-Location Plot** (Figure 4b): Shows a slight curvature, indicating some heteroscedasticity.
- **Normal Q-Q Plot** (Figure 4c): Deviations from the reference line, especially in the tails, indicate deviations from normality.
- **Residuals vs. Leverage** (Figure 4d): A few observations have high leverage and large residuals, indicating potential influence on model estimates.
- **Histogram of Residuals** (Figure 4e): The residuals are approximately symmetric, but have heavier tails than a normal distribution.

Tests for Normality and Heteroscedasticity

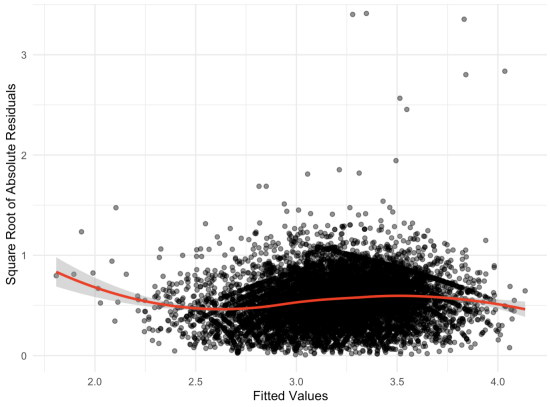
- **Shapiro-Wilk Test**: The test yields $W = 0.7807$ with a p value $< 2.2 \times 10^{-16}$, which leads us to reject the null hypothesis of normality in the residuals.
- **Breusch-Pagan Test**: The test yields $BP = 85.282$ with $df = 37$ and a p value of 1.113×10^{-5}. We find significant evidence of heteroscedasticity at the 1% significance level.

Robust Standard Errors Given the evidence of non-normality and heteroscedasticity, we computed robust standard errors using the HC1 estimator. The robust standard errors are reported in column (2) of Table 8. The significance levels remain largely unchanged, suggesting that heteroscedasticity does not significantly affect the estimated standard errors.

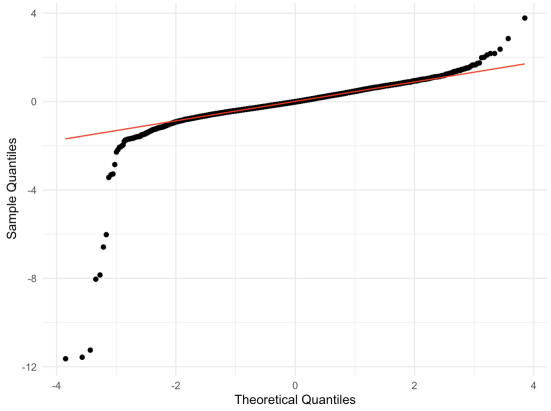
Best Subset Selection To explore whether a simpler model could achieve similar predictive performance, we performed Best Subset Selection (BSS) using an exhaustive search over all 37 predictors. The optimal model size corresponds to 25 predictors, yielding a test MSE of 0.3147, which is slightly lower than the test MSE of the full model of 0.3150



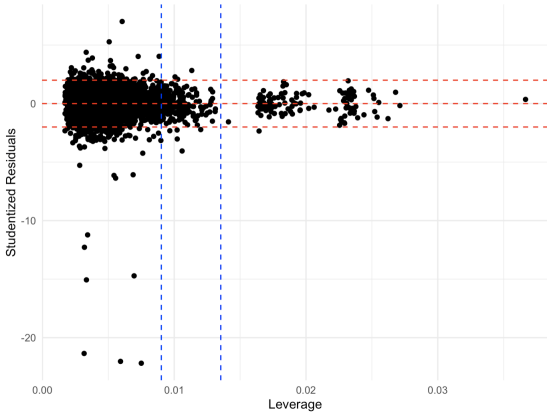
(a) Residuals vs. Fitted Values



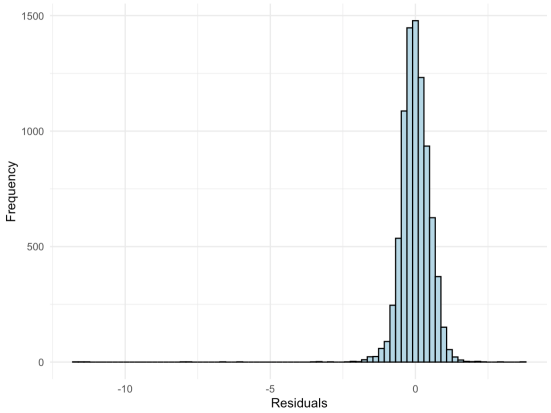
(b) Scale-Location Plot



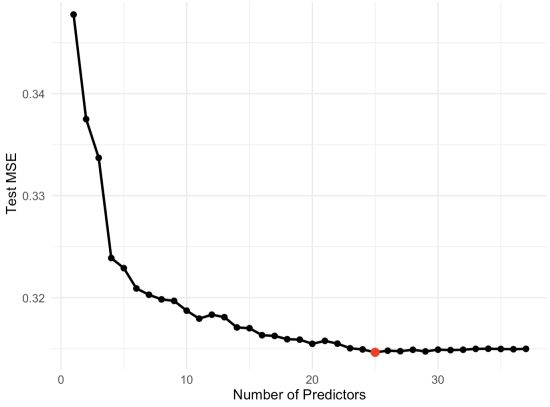
(c) Normal Q-Q Plot



(d) Residuals vs. Leverage



(e) Histogram of Residuals



(f) BSS: Test MSE by Model Size

Figure 4: OLS Regression Diagnostics

(see Figure 4f). The negligible difference indicates that the inclusion of additional predictors does not lead to overfitting. The excluded variables are mainly levels of the job class categorical variables with insignificant coefficients. Therefore, we continue with the original OLS model as it adequately captures the relationships without overfitting.

3.2.2 Robust Regression

The presence of outliers and violations of normality in the residuals, as identified in the EDA and the diagnostic analysis of the OLS model, suggests that standard linear regression may not provide reliable estimates. To address these issues, we employ robust regression methods that are designed to be less sensitive to outliers and violations of model assumptions such as non-normality of residuals. Specifically, we consider three robust regression techniques: Huber's M-estimator, MM-estimator with Peña-Yohai initialization, and Least Trimmed Squares (LTS) regression. These methods offer different approaches to mitigate the influence of outliers and achieve robust parameter estimates.

Huber's M-Estimator Huber's M-estimator is a robust regression method that reduces the influence of outliers by using a loss function that is less sensitive to extreme values than the squared loss used in OLS. The method minimizes the objective function

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\left(\frac{r_i}{\sigma}\right), \tag{1}$$

where $r_i = y_i - \mathbf{x}_i^\top \boldsymbol{\beta}$ are the residuals, σ is a scale parameter, and $\rho(\cdot)$ is

$$\rho(u) = \begin{cases} \frac{1}{2}u^2, & \text{if } |u| \le k, \\ k|u| - \frac{1}{2}k^2, & \text{if } |u| > k. \end{cases} \tag{2}$$

where $k > 0$ is a tuning parameter that controls the threshold between quadratic and linear behavior. The Huber loss is quadratic for small residuals (like OLS) and linear for large residuals, reducing the influence of outliers.

The estimator is computed iteratively using weighted least squares, where the weights depend on the residuals from the previous iteration:

$$w_i = \begin{cases} 1, & \text{if } |u_i| \le k, \\ \frac{k}{|u_i|}, & \text{if } |u_i| > k, \end{cases} \tag{3}$$

with $u_i = \frac{r_i}{\sigma}$. The iterative reweighted least squares (IRLS) algorithm updates the parameter estimates until convergence.

MM-estimator with Peña-Yohai Initialization The MM-estimator is a robust regression method that combines high breakdown point estimators with high efficiency. It consists of three steps:

1. **Initial Estimation**: Obtain a robust initial estimate of the parameters using an S-estimator with a high breakpoint (e.g., 50%). The S-estimator minimizes the scale of the residuals, given by

$$\hat{\boldsymbol{\beta}}_0 = \arg\min_{\boldsymbol{\beta}} \sigma(\boldsymbol{\beta}),$$

where $S(\boldsymbol{\beta})$ is a scale estimate that satisfies

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i}{S(\boldsymbol{\beta})}\right) = b, \tag{4}$$

where b controls the breakpoint. The S-estimator that we use is Turkey's biweight function.

2. **Refinement**: Improve the efficiency of the estimator by applying an M-estimator with a redescending ψ function starting from the initial estimate.

3. **Final Estimation**: Adjust the scale and obtain the final parameter estimates that achieve high efficiency in the normal model.

The Peña-Yohai initialization uses an initial estimator that is robust to high leverage points and outliers in the response variable.

The MM-estimator solves

$$\sum_{i=1}^{n} \psi\left(\frac{r_i}{\hat{\sigma}}\right)\mathbf{x}_i = \mathbf{0}, \tag{5}$$

where $\psi(\cdot)$ is a bounded, redescending function and $\hat{\sigma}$ is a robust estimate of the scale.

Least Trimmed Squares (LTS) Regression The LTS regression estimator is a robust method that minimizes the sum of the least squares residuals, effectively trimming a fraction of the largest residuals. Formally, it solves

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{h} \left(r_{(i)}\right)^2, \tag{6}$$

where $r_{(i)}$ are the residuals and $h = \lfloor n\alpha \rfloor$ where $0.5 \leq \alpha \leq 1$ controls the proportion of observations used in the fitting process. The LTS estimator has a high breakpoint, up to 50% when $\alpha = 0.5$, meaning it can tolerate up to 50% contamination in the data without breaking down. We are using $\alpha = 0.5$.

LTS regression focuses on fitting the model to the majority of the data by excluding the most extreme residuals, which are likely to be outliers.

Estimation Results and Comparison The robust regression models were estimated on the training data using the same model specification as the OLS model. The estimated coefficients for the robust methods are presented in Table 8 in columns (3) through (5).

Overall, the coefficients obtained from the robust regression methods are similar to those obtained from the OLS model, indicating that the main conclusions regarding the relationships between the predictors and the log-transformed wage remain consistent. However, the robust methods provide more reliable estimates in the presence of outliers and heteroscedasticity.

Diagnostic plots for the robust regression models were examined to assess fit, detect outliers, and evaluate the effectiveness of robustness in dealing with extreme observations.

- **Residuals vs. Fitted Values and Scale-Location Plots**: For all robust models, these plots show mild heteroscedasticity, indicating adequate model fit. The results are consistent with those observed in the OLS model.

- **Q-Q Plots of Residuals**: The Q-Q plots for the robust models closely match the OLS model for the central part of the distribution, with small deviations at the tails. These deviations correspond to vertical outliers that are successfully downweighted by the robust methods. This adjustment ensures that the models maintain the assumption of approximate normality in the residuals while minimizing the influence of extreme values.

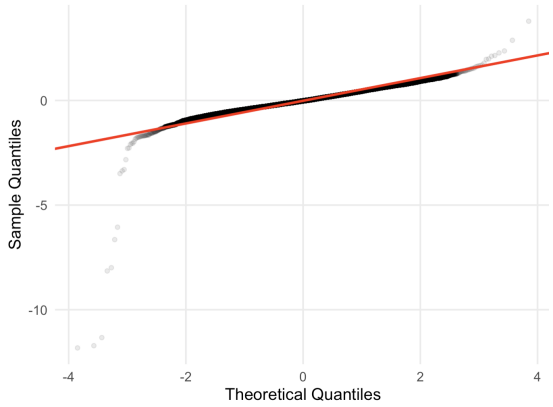
- **Outlier Maps**: Outlier maps for the robust models (Huber's M-estimator, LTS regression, and MM-estimator) highlight the relationship between the standardized residuals and the leverage (or Mahalanobis distance). Figure 5 shows the Q-Q plots and outlier maps for the LTS regression and the MM-estimator. Residuals vs. fitted values and scale-location plots are similar to OLS and are omitted for brevity.

  - The dashed lines in the outlier maps serve as visual thresholds for identifying

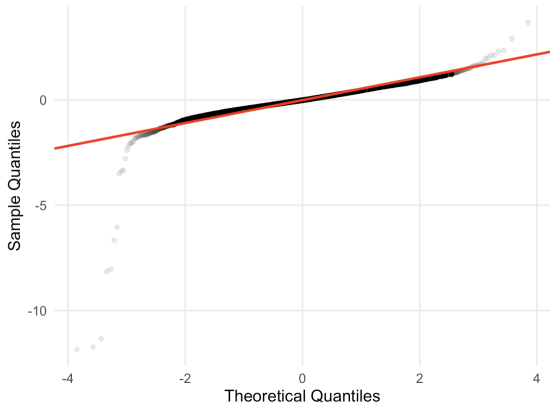
potential outliers. The horizontal red dashed lines at standardized residuals of -2 and 2 indicate observations with unusually large residuals. The vertical blue dashed lines, plotted at multiples (2 and 3 times) of the mean leverage, highlight observations with high leverage that could disproportionately influence the model estimates.

- The maps show that outliers, both leverage (although there are almost no bad leverage points) and vertical, are appropriately assigned reduced weights, as indicated by the transparency of the points. The darker (more opaque) points reflect observations with high weights (closer to 1), suggesting that these observations fit the model assumptions well.
- Vertical outliers are roughly equally distributed above and below zero, and their influence is effectively minimized by robust weighting.
- Good leverage points (observations with high leverage that support the regression line) retain weights close to 1, highlighting their positive contribution to the model.

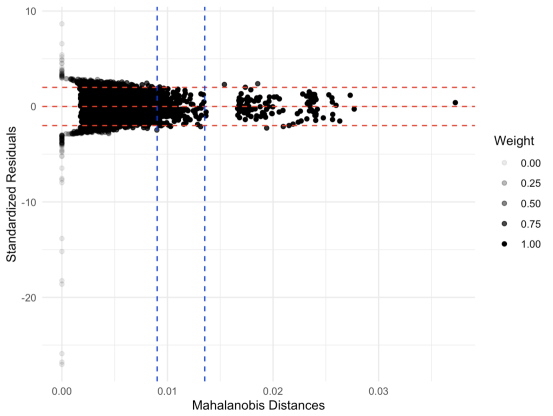
The robust regression methods provide parameter estimates that are less sensitive to outliers and deviations from model assumptions than OLS. While coefficients and predictive accuracy are similar across models, robust methods increase reliability by reducing the influence of extreme observations. The **MM-estimator with Peña-Yohai initialization** shows slightly tighter clustering of residuals around zero and better control of leverage points compared to other methods. This suggests slightly better handling of outliers and improved robustness to leverage effects. However, the differences are minimal and the overall results, both in terms of model coefficients and predictive accuracy, are very similar to those obtained from the OLS model.



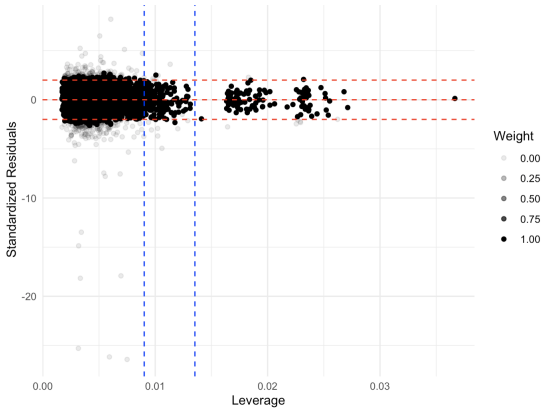
(a) Normal Q-Q Plot for MM-estimator



(b) Normal Q-Q Plot for LTS Regression



(c) Outlier Map for MM-estimator



(d) Outlier Map for LTS Regression

Figure 5: Robust Regression Diagnostics. Only MM-estimator and LTS Regression are shown.

Test MSE Comparison The test Mean Squared Errors (MSE) for the robust regression models are presented in Table 5. The results indicate that the robust methods achieve predictive performance very similar to the OLS model, with test MSEs nearly identical.

Table 5: Test Mean Squared Error for Regression Models

Model	**Test MSE**
Ordinary Least Squares (OLS) Regression	0.3150
Best Subset Selection (BSS) Regression	0.3147
Huber's M-estimator Robust Regression	0.3147
MM-estimator Robust Regression with Peña–Yohai Initialization	0.3149
Least Trimmed Squares (LTS) Robust Regression	0.3150
Linear Mixed Effects (LME) Regression	0.3173

Given the similarity in predictive performance and model coefficients, and given the potential benefits of robustness, the use of robust regression methods such as the MM-estimator may still be advantageous in empirical analyses, especially when there is concern about the presence of outliers or violations of classical linear regression assumptions. However, in this particular analysis, the robust methods did not yield significantly different results from the OLS model, suggesting that the influence of outliers and assumption violations may be limited in this dataset.

3.2.3 Linear Mixed-Effects Models

To account for potential hierarchical structures and unobserved heterogeneity at the household level, we use Linear Mixed-Effects (LME) models. These models allow for both fixed effects, which represent the average relationships across individuals, and random effects, which capture the variability attributable to group-level factors - in this case, households.

Because individuals within the same household may share unobserved characteristics that affect their wages, such as shared economic resources or social networks, including random effects for households can improve model accuracy and inference.

The LME model is specified as:

$$\begin{aligned}\texttt{log\_wage}_{ij} = \beta_0 &+ \beta_1 \texttt{age}_{ij} + \beta_2 \texttt{age}_{ij}^2 + \beta_3 \texttt{sex}_{ij} + \beta_4 \texttt{race}_{ij} + \beta_5 \texttt{citizen}_{ij} \\ &+ \beta_6 \texttt{married}_{ij} + \beta_7 \texttt{metropolitan}_{ij} + \beta_8 \texttt{division}_{ij} + \beta_9 \texttt{education}_{ij} \\ &+ \beta_{10} \texttt{job\_class}_{ij} + \beta_{11} \texttt{industry}_{ij} + u_j + \varepsilon_{ij},\end{aligned}$$

where:

- i indexes individuals, j indexes households.
- $u_j \sim N(0, \sigma_u^2)$ is the random intercept for household j, capturing unobserved household-level effects.
- $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ is the individual-level residual error.

We considered several candidate LME models with different random effect structures, including nested random effects for divisions and households. However, these models did not significantly improve the model fit compared to the simpler model with only random

intercepts at the household level. Therefore, we chose the simpler model for ease of interpretation.

Estimation Results The estimation results for the fixed effects of the LME model are presented in column (6) of Table 8. The estimated fixed effects are largely consistent with those of the OLS model, indicating the robustness of the estimated effects across different modeling approaches.

The test mean square error (MSE) for the LME model is 0.3173, which is slightly higher than the test MSE for the OLS model, which is 0.3150. This suggests that the inclusion of household-level random effects does not significantly improve the predictive performance of the model.

Variance Components The estimated variance components of the LME model are as follows:

- **Household level variance** (random intercept): $\sigma_u^2 = 0.07506$;
- **Residual Variance at Individual Level**: $\sigma_\varepsilon^2 = 0.21912$.

To assess the proportion of total variance that is due to household-level effects, we compute the Percentage of Variation due to Random Effects (PVRE):

$$\text{PVRE} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2} \times 100\% = \frac{0.07506}{0.07506 + 0.21912} \times 100\% = 25.51\%.$$

This indicates that about 25.51% of the unexplained variance in `log_wage` is due to unobserved differences between households, while the remaining 74.49% is due to individual-level factors.

Model Diagnostics We performed diagnostic checks to assess the assumptions of the LME model (see Figure 6). Diagnostic plots for the LME model indicate that the residuals behave similarly to those in the OLS model.

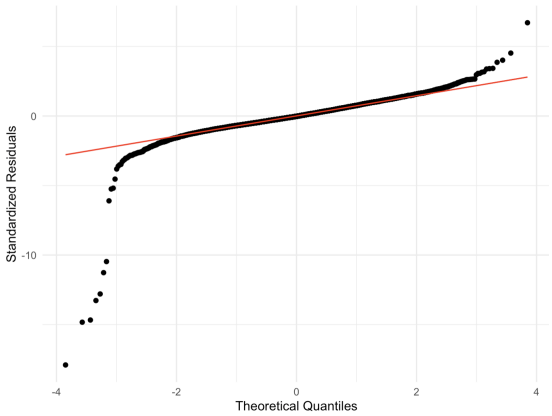
- **Residuals vs. Fitted Values**: The standardized residuals are centered around zero with no obvious patterns, indicating that the model adequately captures the central trend. A slight funnel shape indicates mild heteroscedasticity.

- **Q-Q Plot of Standardized Residuals**: The residuals generally follow the theoretical quantiles, with deviations in the tails indicating possible non-normality due to outliers.

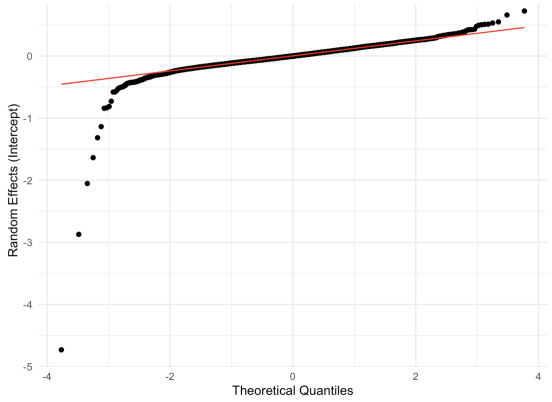
- **Q-Q Plot of Random Effects**: The random intercepts for households agree well with the assumption of normality, with slight deviations in the tails.

To deal with potential outliers and violations of model assumptions, we explored robust Linear Mixed Effects models based on M-estimators. These models produced results very similar to the standard LME model, with no significant differences in parameter estimates or predictive performance. Therefore, we do not report these results.

The inclusion of household-level random effects in the LME model provides no improvement over the OLS model in terms of predictive performance, as indicated by the slightly higher test MSE. However, the LME model accounts for unobserved heterogeneity at the household level by capturing common factors that influence wages. About one-quarter of the variance in wages can be attributed to unobserved household-level effects. Nevertheless, individual-level factors remain the primary determinants of wage variation in the data.



(a) Normal Q-Q Plot for Standardized Residuals



(b) Normal Q-Q Plot of Random Effects (Household)

Figure 6: LME Model Diagnostics. Residuals vs. fitted values is similar to OLS, so it is omitted.

3.2.4 Bootstrap Methods

To obtain more reliable estimates of parameter variability and predictive performance, especially in the presence of non-normality we use bootstrap methods. Bootstrapping is a resampling technique that approximates the sampling distribution of an estimator

by repeatedly resampling with replacement from the observed data and recalculating the estimator for each resampled dataset. This approach is particularly advantageous when the theoretical distributions are complex or unknown, and it is transformation preserving, i.e., it maintains the properties of the transformed variables.

Bootstrap Confidence Intervals for Mean and Median We first apply bootstrapping methods to estimate confidence intervals for the mean and median of the wage variables, both in their original scale (`wage`) and in their log-transformed scale (`log_wage`). Using the bias-corrected and accelerated (BCa) bootstrap intervals, which adjust for both bias and skewness in the bootstrap distribution, we perform $B = 1000$ bootstrap resampling. Our goal is to demonstrate the flexibility of bootstrap methods, which provide valid confidence intervals even for non-normal data and statistics such as the median, by using their transformation-preserving nature.

Table 6 summarizes the sample estimates and 95% confidence intervals obtained by normal theory and bootstrap BCa methods.

Table 6: Comparison of Confidence Intervals for Mean and Median of `wage` and `log_wage`

	Mean		Median
	Normal Theory 95% CI	Bootstrap BCa 95% CI	Bootstrap BCa 95% CI
`wage`	[29.05, 30.06]	[29.13, 30.24]	[23.09, 24.02]
`log_wage`	[3.195, 3.217]	[3.196 3.218]	[3.139, 3.179]

For `log_wage`, the bootstrap BCa confidence interval closely matches the normal theory interval, reflecting the approximate normality of the log-transformed wage. In contrast, for `wage`, the bootstrap interval is slightly wider and asymmetric, capturing the skewness present in the original wage distribution. The bootstrap's transformation-respecting property ensures that confidence intervals remain valid under transformations, making it particularly useful for non-normally distributed variables such as `wage`, where normal theory intervals may not fully capture the underlying variability.

Bootstrap Estimation of the Test Mean Squared Error To evaluate the predictive performance of our regression models without relying solely on a validation set, we use the *.632 bootstrap estimator* for the test Mean Squared Error (MSE). The .632 bootstrap method combines in-sample and out-of-sample error estimates from bootstrap samples to approximate the expected test error, effectively addressing potential biases associated with training on resampled data.

For each bootstrap sample $b = 1, \ldots, B$, we perform the following steps:

1. **Model Fitting**: Fit the model on the bootstrap sample (in-bag data) to obtain estimated coefficients $\hat{\beta}^{(b)}$.

2. **In-Bag MSE**: Compute the in-bag MSE:

$$\text{MSE}_{\text{in-bag}}^{(b)} = \frac{1}{n_{\text{in-bag}}} \sum_{i \in \text{in-bag}} \left(y_i - \hat{y}_i^{(b)} \right)^2,$$

   where $\hat{y}_i^{(b)} = x_i^\top \hat{\beta}^{(b)}$.

3. **Out-of-Bag MSE**: Compute the out-of-bag MSE:

$$\text{MSE}_{\text{out-of-bag}}^{(b)} = \frac{1}{n_{\text{out-of-bag}}} \sum_{i \in \text{out-of-bag}} \left(y_i - \hat{y}_i^{(b)} \right)^2.$$

4. **.632 Bootstrap MSE**: Combine the in-bag and out-of-bag MSEs:

$$\text{MSE}_{.632}^{(b)} = 0.368 \times \text{MSE}_{\text{in-bag}}^{(b)} + 0.632 \times \text{MSE}_{\text{out-of-bag}}^{(b)}.$$

The overall .632 bootstrap estimate of the test MSE is then obtained by averaging over all bootstrap samples:

$$\text{MSE}_{.632} = \frac{1}{B} \sum_{b=1}^{B} \text{MSE}_{.632}^{(b)}.$$

We apply this procedure to both the OLS regression model and the robust MM-estimator regression model, with $B = 300$ bootstrap replications. Bootstrapping pairs (resampling observations with both predictors and response) is used to preserve the relationship between variables.

The .632 bootstrap estimates of the test MSE for the models are presented in Table 7, along with the previously obtained test MSEs from the validation set approach.

Table 7: Comparison of Test MSE Across Models

Model	Validation Set Test MSE	.632 Bootstrap Test MSE
OLS Regression	0.3150	0.3113
Robust MM-Estimator Regression	0.3149	0.3014

We observe that the .632 bootstrap estimates are very similar to the validation set MSEs, demonstrating the reliability of the bootstrap method for estimating test error. The robust MM-estimator regression model shows a slightly lower test MSE compared to the OLS model, indicating a marginal improvement in predictive performance.

Hypothesis Testing for Difference in Test MSE Using the bootstrap distribution of the test MSE differences between the two models, we construct a confidence interval to test whether the difference is statistically significant. Specifically, we compute the differences $\Delta\text{MSE}^{(b)} = \text{MSE}_{\text{MM}}^{(b)} - \text{MSE}_{\text{OLS}}^{(b)}$ for each bootstrap sample and obtain the $(1-\alpha)$ confidence interval from the empirical quantiles of the bootstrap differences (see Figure 7 for the histogram of bootstrap MSE differences).

For $\alpha = 0.05$, the 95% bootstrap confidence interval for the difference in the test MSE is $[-0.060, 0.030]$. Since this interval includes zero, we conclude that there is no statistically significant difference between the test MSEs of the OLS and robust MM-estimator models at the 5% significance level. This indicates that both models perform comparably in terms of predictive accuracy.

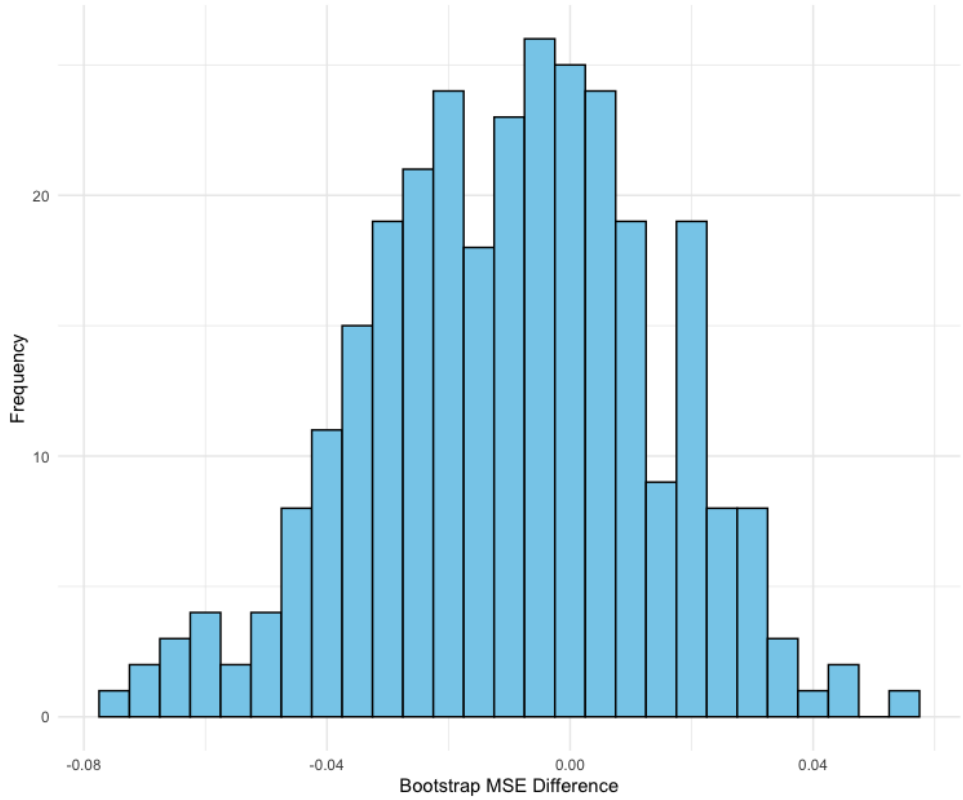


Figure 7: Histogram of Bootstrap MSE Differences

The use of bootstrap methods has several advantages in our analysis:

- **Transformation-Respecting Confidence Intervals**: Bootstrapping provides confidence intervals that are valid under transformations of the data. This is evident in the comparison between `wage` and `log_wage`, where the bootstrap BCa intervals for `log_wage` are close to normal theory intervals due to the approximate normality after transformation.

- **Test MSE Estimation Without Validation Set**: The .632 bootstrap estimator allows us to estimate the test MSE without relying on a separate validation set or cross-validation. This is particularly useful when data splitting may not be desirable due to sample size constraints.

- **Statistical Inference on Predictive Performance**: By constructing confidence intervals for the difference in test MSEs, we can formally test hypotheses about the comparative predictive performance of different models, which is not straightforward with traditional validation set approaches or cross-validation.

Our results suggest that while the robust MM-estimator regression model has a slightly lower test MSE than the OLS model, the difference is not statistically significant. This implies that both models have similar predictive capabilities on our dataset. The consistency of parameter estimates across bootstrap samples and models reinforces the robustness of our conclusions regarding the determinants of personal income.

Discussion of Results We used a variety of statistical models to ensure robustness and to account for potential violations of standard regression assumptions. Starting with Ordinary Least Squares (OLS) regression, we observed slight deviations from normality and the presence of outliers in the residuals. Diagnostic tests indicated heteroscedasticity and non-normality, which could potentially bias standard errors and affect inference.

To address these concerns, we applied robust regression techniques, including Huber's M-estimator, MM-estimator with Peña-Yohai initialization, and Least Trimmed Squares (LTS) regression. The robust models produced coefficient estimates and test Mean Squared Errors (MSE) that were very similar to those of the OLS model. This consistency suggests that the impact of outliers and assumption violations on our OLS estimates is minimal.

Given the potential for unobserved heterogeneity at the household level, we extended our analysis using Linear Mixed Effects (LME) models. The LME model included household-level random effects to capture common factors among individuals within the same house-

hold. The estimated Percentage of Variation due to Random Effects (PVRE) was approximately 25.51%, indicating that household-level factors account for a modest portion of the unexplained variance in `log_wage`. However, the majority of the variance (about 74.49%) is explained by unobserved individual-level characteristics. The fixed effects from the LME model remained consistent with the OLS estimates, reinforcing the robustness of our findings across different modeling frameworks.

To further validate our results and provide more reliable estimates of predictive performance, we used bootstrap methods. The .632 bootstrap estimator allowed us to estimate the test MSE without relying on a separate validation set, providing an advantage in terms of statistical efficiency. The bootstrap results confirmed that the robust MM-estimator did not significantly outperform the OLS model in predictive accuracy. The 95% bootstrap confidence interval for the difference in test MSE between the MM-estimator and the OLS model included zero, indicating no statistically significant difference. This formal test reinforces the conclusion that the OLS model is sufficient for our analysis.

Given the consistent results across models and the lack of significant improvement from more complex or robust methods, we select the OLS regression model with robust standard errors for further economic interpretation. The OLS model offers simplicity and interpretability while adequately capturing the relationships between the predictors and the log-transformed wage.

4 Analysis

In this section we present the economic analysis of the regression results from the OLS model and the variance components from the linear mixed effects model.

- **Education**: Each additional year of education is associated with an increase in expected wages of about 7.3%, holding other factors constant. This effect is highly significant ($p < 0.001$). This is completely in line with stylized empirical facts; higher education leads, on average, to higher wages.

- **Age**: Age has a positive effect on wages, but as can be seen on the negative sign on the coefficient in from of Age^2 there are diminishing returns to age and in fact the effect of age becomes negative. The age with the highest income can be calculated as the top point of a concave parabola as approximately 49 years. The positive affect of age stems from the experience that a person gets from being on the job market. The

negative effect is probably a result of people working less and prioritizing different things in life as they get older.

- **Sex**: Women earn significantly less than men, with a coefficient indicating a reduction in expected wage of about 13.7% ($p < 0.001$), suggesting a gender wage gap. Understanding this effect, it is important to keep in mind, that this is after taken all the other variables into consideration. In this way, this is a very powerful result.

- **Citizenship**: Being a U.S. citizen is associated with an approximate 4.5% higher expected wage, although it is only significant at the 10% level ($p < 0.1$).

- **Race**: Black individuals earn significantly less than white individuals, with a coefficient indicating an approximate 9.4% lower expected wage ($p < 0.001$). Hispanic individuals have a negative coefficient of -0.055, corresponding to an approximate 5.5% decrease in expected wage ($p < 0.05$). Asian individuals have a small negative coefficient ($\beta = -0.02$), suggesting an approximate 2% lower expected wage compared to White individuals, but this effect is not statistically significant ($p = 0.401$). Thus, there is a significant wage gap between Black, Hispanic and White individuals in this sample, while wage differences between Asian and White individuals are not statistically significant.

- **Marital Status**: Married individuals have a higher expected wage than unmarried individuals, with an increase of about 8.2% ($p < 0.001$). An interesting way to further explore the effect of marital status, would be to include an interaction term in between marital status and sex. Traditional gender roles would probably imply a negative effect of marriage on wage for woman and the opposite for men.

- **Metropolitan Status**: Living in a metropolitan area is associated with an approximate 9.4% increase in expected wages ($p < 0.001$). This is as expected since many of the high-paying jobs are in the big cities.

- **Division**: Individuals in certain divisions (e.g., East South Central, West South Central, Mountain) earn less than those in the reference division (New England), with varying degrees of statistical significance. These results are very hard to interpret because the divisions has so many different states in them that there is not any coherent story to tell about them from an economic perspective.

- **Job Class**: Working in state or local government is associated with lower wages compared to federal government jobs, with coefficients indicating approximate decreases of 21.7% ($p < 0.001$) for state government and 13.3% ($p < 0.001$) for local government. Private for-profit jobs have a negative coefficient of -0.074, indicating an approximate 7.4% decrease in expected wage relative to federal government jobs, although this effect is marginally significant ($p = 0.060$). Private nonprofit jobs and self-employed positions (both incorporated and unincorporated) also have negative coefficients, indicating lower wages relative to federal government jobs, but the statistical significance varies.

- **Industry**: Industries such as mining, construction, manufacturing, public administration, information and financial activities are associated with higher wages, while wholesale and retail trade, leisure and hospitality, and other services are associated with lower wages.

The findings of the analysis closely align with Mincer (1958), stating that the development of human capital through education and experience is a key determinant of income. More recent literature, such as Borjas and Van Ours (2010) confirm the results obtained. Diminishing returns to experience found in the model are confirmed in the literature. The significant gender and racial wage gaps, where women and minority groups earn significantly less is confirmed by Borjas and Van Ours (2010), while native-born workers tend to earn more. While being married is associated with a higher wage, which according to Mincer (1958) can be viewed as an investment in human capital, as married individuals may have greater access to stable household resources and social networks, Polachek (2007) highlights that the gender wage gap between married men and women is significantly larger than between single men and women.

The PVRE of 25.51% suggests that household-level factors account for a non-negligible fraction of the variance in wages. This may capture unobserved household characteristics such as shared economic resources, social networks, or other family-level effects. However, the majority of the variance is still explained by individual-level characteristics, highlighting the importance of these factors in wage determination.

It is noteworthy that about 33% of the individuals in the dataset live in one-person households (49% - in two-person households, 12% - in three-person households, and so on, counting only employed individuals), which limits the strength of household-level random effects since there is no intra-household variability for these individuals. Nevertheless, the

inclusion of household random effects remains valuable for capturing the shared variance among multi-person households.

5 Conclusion

This study investigated the determinants of wages using a range of statistical methods to ensure robustness and address potential violations of standard regression assumptions. While initial diagnostic tests revealed minor issues with normality and heteroscedasticity in the OLS regression, robust methods such as Huber's M-estimator, MM-estimator, and Least Trimmed Squares produced results that were consistent with OLS. These findings indicate that the OLS estimates were not significantly affected by outliers or assumption violations. We determined individual-level factors, such as education and age as key determinants, while other demographic factors such as gender, being married, and race also significantly affect wage. These results are consistent across OLS and robust regression models. It was shown that occupation, industry, and regional variables contribute to the variation of wages, reflecting structural disparities in the labor market. However, due to their broad definition it is difficult to draw concrete conclusions. The inclusion of household-level random effects in Linear Mixed Effects (LME) models revealed that approximately 25.51% of wage variance is attributed to unobserved household-level factors, while the majority remains at the individual level. This highlights the importance of environmental and socio-economic factors influencing individual wages, going beyond observable characteristics. Bootstrap validation further confirmed the robustness of the results, showing no significant improvement in predictive accuracy when using more complex estimators.

While the findings are robust, future research could try to investigate the unobserved household-level factors more in detail, given that these account for a significant variance in wages. Incorporating additional household-level variables in a dynamic context through an analysis of longitudinal data, and investigating the effects of interaction among demographic, occupational, and regional factors could give more insights in our understanding of wage determination. More thorough investigation of disparities among gender and race could help to further eliminate persistent wage gaps and supporting policymaking for a more equitable labor market.

Regression Results

Table 8: Regression Results

	OLS	OLS (HC1)	M (Huber's)	MM	LTS	LME	Bootstrap
Intercept	1.434^{***} (0.089)	1.434^{***} (0.087)	1.275^{***} (0.074)	1.253^{***} (0.075)	1.263^{***} (0.067)	1.428^{***} (0.09)	1.434^{***} (0.087)
Age	0.037^{***} (0.003)	0.037^{***} (0.002)	0.04^{***} (0.002)	0.04^{***} (0.002)	0.039^{***} (0.002)	0.037^{***} (0.002)	0.037^{***} (0.003)
Age^2	-0.000^{***} (0.000)	-0.000^{***} (0.000)	-0.0004^{***} (0.000)	-0.0004^{***} (0.000)	-0.0004^{***} (0.000)	-0.0004^{***} (0.000)	-0.0004^{***} (0.000)
Sex (Female)	-0.137^{***} (0.013)	-0.137^{***} (0.013)	-0.151^{***} (0.011)	-0.154^{***} (0.011)	-0.156^{***} (0.01)	-0.14^{***} (0.012)	-0.138^{***} (0.013)
Citizen (US citizen)	0.045^{*} (0.025)	0.045^{*} (0.024)	0.049^{**} (0.021)	0.044^{**} (0.021)	0.044^{**} (0.019)	0.045^{*} (0.026)	0.045^{*} (0.024)
Married	0.082^{***} (0.013)	0.082^{***} (0.015)	0.071^{***} (0.011)	0.065^{***} (0.011)	0.065^{***} (0.01)	0.084^{***} (0.014)	0.084^{***} (0.014)
Metropolitan	0.094^{***} (0.017)	0.094^{***} (0.02)	0.082^{***} (0.014)	0.08^{***} (0.014)	0.077^{***} (0.013)	0.093^{***} (0.018)	0.093^{***} (0.019)
Education	0.073^{***} (0.002)	0.073^{***} (0.003)	0.08^{***} (0.002)	0.081^{***} (0.002)	0.083^{***} (0.002)	0.072^{***} (0.002)	0.074^{***} (0.003)
Division (East North Central)	-0.066^{**} (0.028)	-0.066^{***} (0.025)	-0.075^{***} (0.024)	-0.068^{***} (0.024)	-0.083^{***} (0.021)	-0.062^{**} (0.03)	-0.067^{***} (0.026)
Division (East South Central)	-0.146^{***} (0.031)	-0.146^{***} (0.026)	-0.153^{***} (0.026)	-0.146^{***} (0.026)	-0.153^{***} (0.024)	-0.142^{***} (0.033)	-0.145^{***} (0.025)
Division (Middle Atlantic)	-0.045 (0.03)	-0.045 (0.032)	-0.012 (0.025)	-0.003 (0.025)	-0.006 (0.023)	-0.039 (0.032)	-0.045 (0.032)
Division (Mountain)	-0.078^{***} (0.027)	-0.078^{***} (0.023)	-0.079^{***} (0.023)	-0.07^{***} (0.023)	-0.079^{***} (0.02)	-0.078^{***} (0.029)	-0.078^{***} (0.024)
Division (Pacific)	-0.009 (0.026)	-0.009 (0.028)	0.024 (0.022)	0.035 (0.022)	0.03 (0.02)	-0.003 (0.028)	-0.009 (0.028)

Continued on next page

	OLS	OLS (HC1)	M (Huber's)	MM	LTS	LME	Bootstrap
Division (South Atlantic)	−0.085***	−0.085***	−0.089***	−0.078***	−0.084***	−0.079***	−0.084***
	(0.026)	(0.022)	(0.021)	(0.021)	(0.019)	(0.027)	(0.021)
Division (West North Central)	−0.095***	−0.095***	−0.104***	−0.097***	−0.107***	−0.094***	−0.096***
	(0.028)	(0.024)	(0.024)	(0.024)	(0.021)	(0.03)	(0.023)
Division (West South Central)	−0.114***	−0.114***	−0.117***	−0.106***	−0.123***	−0.11***	−0.114***
	(0.028)	(0.024)	(0.023)	(0.024)	(0.021)	(0.03)	(0.024)
Race (Asian)	−0.021	−0.021	0.003	0.01	0.021	−0.01	−0.022
	(0.025)	(0.035)	(0.021)	(0.021)	(0.019)	(0.026)	(0.034)
Race (Black)	−0.094***	−0.094***	−0.107***	−0.105***	−0.116***	−0.095***	−0.095***
	(0.02)	(0.017)	(0.017)	(0.017)	(0.015)	(0.021)	(0.017)
Race (Hispanic)	−0.055***	−0.055***	−0.057***	−0.057***	−0.052***	−0.055***	−0.056***
	(0.02)	(0.018)	(0.017)	(0.017)	(0.015)	(0.021)	(0.016)
Race (Other)	−0.025	−0.025	−0.035	−0.028	−0.029	−0.017	−0.031
	(0.066)	(0.048)	(0.055)	(0.055)	(0.049)	(0.067)	(0.05)
Industry (Construction)	0.07	0.07	0.076*	0.07*	0.08**	0.078	0.068
	(0.049)	(0.047)	(0.041)	(0.042)	(0.037)	(0.049)	(0.047)
Industry (Educational and health services)	−0.034	−0.034	−0.033	−0.034	−0.034	−0.016	−0.035
	(0.047)	(0.045)	(0.04)	(0.04)	(0.036)	(0.048)	(0.046)
Industry (Financial activities)	0.049	0.049	0.075*	0.072*	0.071*	0.063	0.049
	(0.05)	(0.051)	(0.042)	(0.042)	(0.038)	(0.05)	(0.049)
Industry (Information)	0.062	0.062	0.094*	0.1*	0.048	0.082	0.062
	(0.064)	(0.063)	(0.053)	(0.054)	(0.049)	(0.064)	(0.066)
Industry (Leisure and hospitality)	−0.153***	−0.153***	−0.139***	−0.14***	−0.148***	−0.14***	−0.152***
	(0.049)	(0.046)	(0.041)	(0.042)	(0.037)	(0.05)	(0.048)
Industry (Manufacturing)	0.046	0.046	0.08**	0.081**	0.094**	0.063	0.046
	(0.048)	(0.049)	(0.04)	(0.041)	(0.037)	(0.049)	(0.051)
Industry (Mining)	0.16*	0.16*	0.198***	0.163**	0.287***	0.173*	0.15*
	(0.092)	(0.087)	(0.077)	(0.077)	(0.07)	(0.091)	(0.083)
Industry (Other services)	−0.124**	−0.124**	−0.098**	−0.1**	−0.12***	−0.112**	−0.126**
	(0.052)	(0.052)	(0.044)	(0.044)	(0.04)	(0.052)	(0.054)

Continued on next page

	OLS	OLS (HC1)	M (Huber's)	MM	LTS	LME	Bootstrap
Industry (Professional and business services)	0.072 (0.047)	0.072 (0.046)	0.092** (0.04)	0.094** (0.04)	0.074** (0.036)	0.091* (0.048)	0.073 (0.047)
Industry (Public administration)	0.144** (0.056)	0.144*** (0.054)	0.133*** (0.047)	0.131*** (0.047)	0.134*** (0.042)	0.155*** (0.056)	0.143*** (0.055)
Industry (Transportation and utilities)	−0.026 (0.05)	−0.026 (0.052)	0.003 (0.042)	0.001 (0.043)	−0.003 (0.038)	−0.002 (0.05)	−0.024 (0.053)
Industry (Wholesale and retail trade)	−0.086* (0.047)	−0.086* (0.045)	−0.082** (0.04)	−0.079** (0.04)	−0.1*** (0.036)	−0.073 (0.048)	−0.087* (0.046)
Jobclass (Government - Local)	−0.133*** (0.042)	−0.133*** (0.033)	−0.116*** (0.035)	−0.11*** (0.035)	−0.117*** (0.031)	−0.132*** (0.041)	−0.132*** (0.034)
Jobclass (Government - State)	−0.217*** (0.044)	−0.217*** (0.049)	−0.171*** (0.037)	−0.174*** (0.037)	−0.185*** (0.033)	−0.202*** (0.043)	−0.217*** (0.048)
Jobclass (Private, for profit)	−0.074* (0.04)	−0.074** (0.033)	−0.072** (0.033)	−0.07** (0.033)	−0.075** (0.03)	−0.072* (0.039)	−0.075** (0.035)
Jobclass (Private, nonprofit)	−0.055 (0.045)	−0.055 (0.037)	−0.052 (0.038)	−0.046 (0.038)	−0.055 (0.034)	−0.063 (0.044)	−0.055 (0.039)
Jobclass (Self-employed, incorporated)	−0.094* (0.049)	−0.094** (0.046)	−0.095** (0.041)	−0.088** (0.041)	0.021 (0.037)	−0.106** (0.049)	−0.096** (0.049)
Jobclass (Self-employed, unincorporated)	−0.088* (0.046)	−0.088* (0.045)	−0.071* (0.039)	−0.067* (0.039)	−0.043 (0.035)	−0.091** (0.046)	−0.088* (0.045)

Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

R Code

The R code used for data processing, analysis, and modeling throughout this study is available on GitHub. You can find the complete code at the following link:

GitHub Repository: AMS Wage Analysis

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Borjas, G. J., & Van Ours, J. C. (2010). *Labor economics*. McGraw-Hill/Irwin Boston.

Dikta, G., & Scheer, M. (2021). *Bootstrap methods: With applications in r*. Springer.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in r*. Springer.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th). Pearson Prentice Hall.

Koller, M. (2016). Robustlmm: An r package for robust estimation of linear mixed-effects models. *Journal of Statistical Software*, *75*(6), 1–24.

Maronna, R. A., Martin, D. R., Yohai, V. J., & Salibián-Barrera, M. (2019). *Robust statistics: Theory and methods (with r)*. John Wiley & Sons.

Mincer, J. (1958). Investment in human capital and personal income distribution. *Journal of Political Economy*, *66*(4), 281–302.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in s and s-plus*. Springer.

Polachek, S. W. (2007, November). Earnings over the life cycle: The mincer earnings function and its applications [IZA Discussion Paper No. 3181].