Advanced Multivariate Statistics

Modeling Wage Determinants: A Multivariate Analysis of CPS Data

Aleksandr Dudakov

Benjamin Markert

Sebastian Thoning Hofland

University of Milan

19.12.2024

Problem Statement

Objective:

- Develop and evaluate statistical models to accurately predict wages.
- Analyze the determinants of wages by addressing:
  1. Key individual-level factors affecting personal income.
  2. The contribution of occupation, industry, and regional factors to wage variations.
  3. The role of unobserved household-level factors in wage determination.

Relevance:

- Understanding the determinants of wages is crucial to reducing economic inequality, improving living standards, and promoting economic growth.

Data Description

Data Source:

- Current Population Survey (CPS) Outgoing Rotation Group (ORG), January 2022.
- CPS is a monthly survey designed to collect data on labor force characteristics, with ORG focusing on detailed earnings information. Conducted by the U.S. Census Bureau and Bureau of Labor Statistics.

Key Variables:

- **Demographic:** Age, sex, race, marital status, citizenship.
- **Socioeconomic:** Level of education, job class, industry, metropolitan status.
- **Regional:** Census division.
- **Household:** Household identifier.
- **Target:** Hourly wage including overtime, tips, and commissions.

Data Preprocessing

Handling Missing Data:

- `Wage`: 12.8% missing, `Metropolitan status`: 0.79% missing.
  - **Panel data:** Missing values filled using the closest prior or subsequent observations for the same individual from other months within the CPS.
  - **Hot deck imputation:** Remaining missing values replaced using data from similar respondents grouped by age, gender, race, and education.
- Remaining missing values and zero-wage (131) observations dropped.

Transformations:

- `Age`: Recoded "80+" as 80.
- `Education`: Mapped categories to approximate years of schooling (e.g., Bachelor's = 16).

Final Dataset:

- Employed individuals aged 16+.
- 12,037 observations and 12 variables.

Exploratory Data Analysis

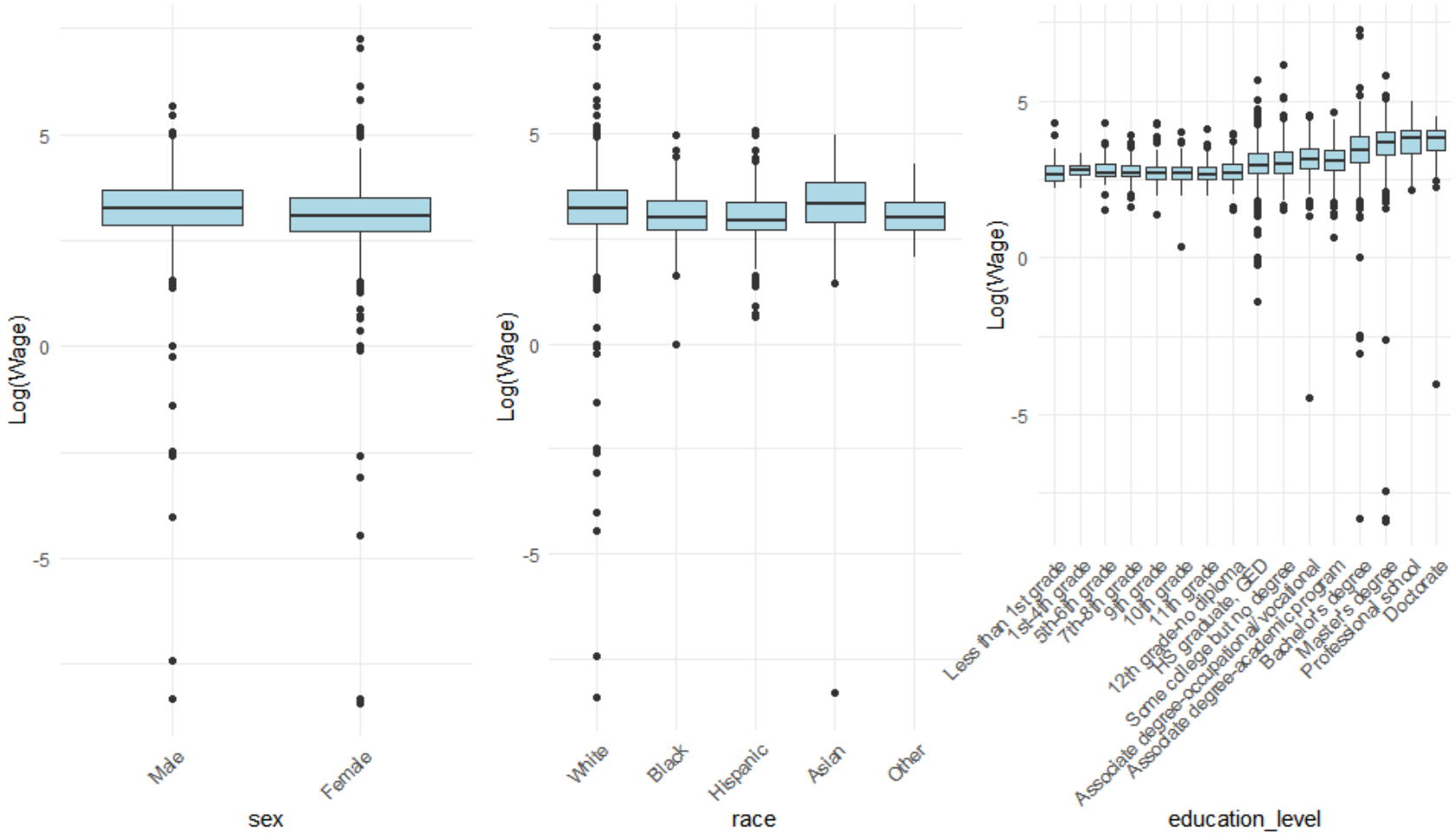
Key Steps:

- Visualized wage distributions: identified significant right skewness. Applied log transformation to `wage` to address skewness and stabilize variance.
- Used boxplots to examine wage variability across categories (e.g., `education`, `race`, `sex`).
- Conducted correlation and multicollinearity analysis: confirmed no multicollinearity ($VIF < 5$ for all variables).

Insights:

- Significant wage differences observed by education level, industry, race, and sex.
- Data characteristics, including outliers and skewness, suggest using robust regression techniques for reliable analysis.

Boxplots of Wage by Categories



Methodology Overview

Modeling Approach:

- **OLS Regression:** Baseline model.
- **Robust Regression:** Addressed outliers, heteroscedasticity, and non-normality using Huber's M-estimator, MM-estimator, and LTS.
- **Linear Mixed Effects (LME):** Incorporated household-level random effects (25% of the total unexplained variance in wages).

Validation:

- Validation set approach: 70%-30% train-test split.
- Test MSE calculated on the test set to evaluate predictive performance.
- .632 bootstrap applied to combine in-bag and out-of-bag errors, providing an unbiased estimate of MSE.

OLS Regression: Overview

Model Specification:

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

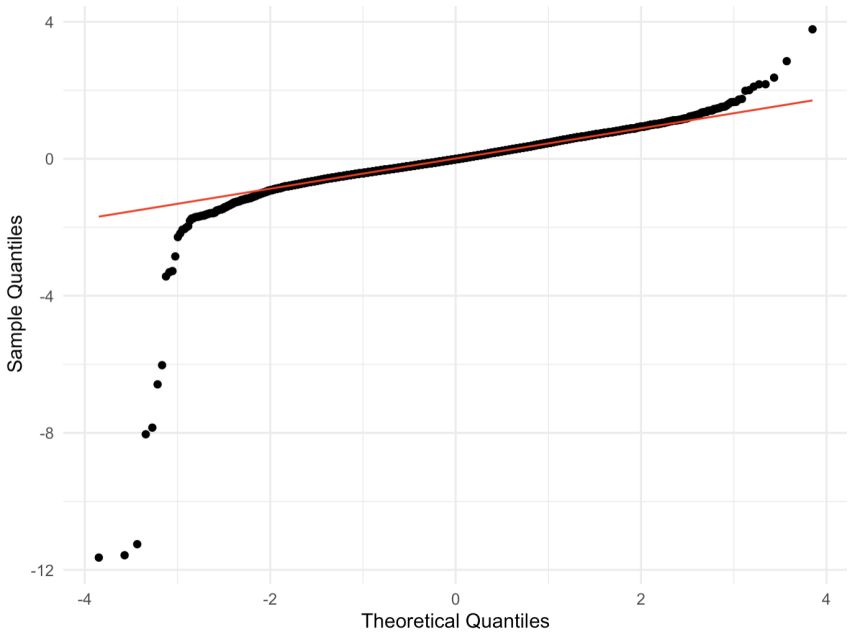
Diagnostics:

- Mild heteroscedasticity detected (Breusch-Pagan test, $p < 0.01$).
- Non-normal residuals confirmed (Shapiro-Wilk test, $p < 0.001$).

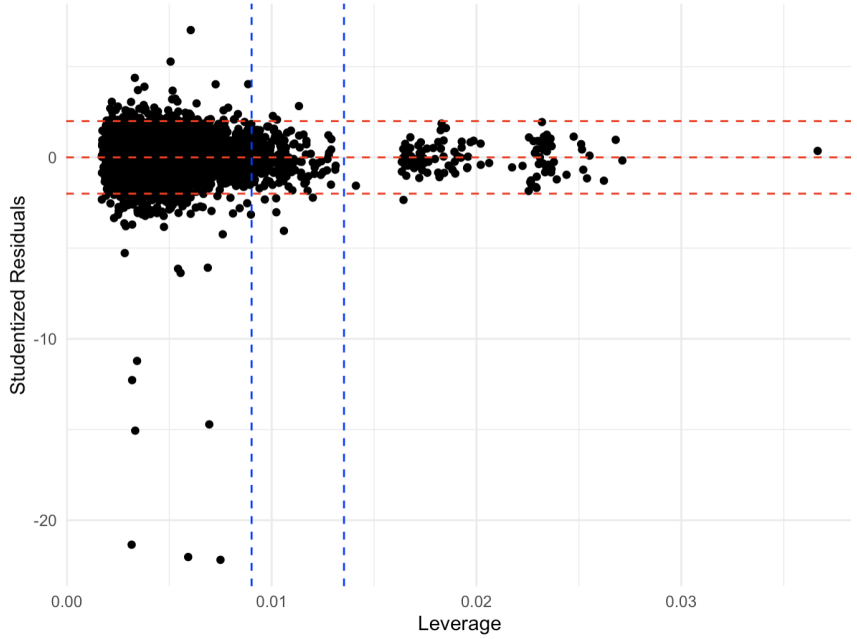
Model Refinements:

- Robust standard errors computed to address heteroscedasticity.
- Best Subset Selection (BSS) suggests the full model is adequate; additional predictors do not significantly improve performance.

OLS Regression: Diagnostics



(a) Normal Q-Q Plot



(b) Residuals vs. Leverage

Observations:

- Residuals show non-normality, especially in the tails.
- Outliers with large residuals detected, potentially affecting estimates.
- Robust regression is recommended to address non-normality and outliers.

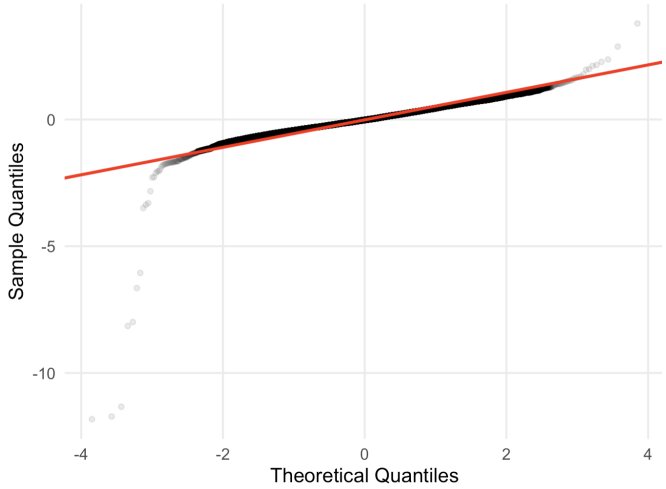
Robust Regression: Methods

- **Huber's M-estimator:**
  - Combines OLS and linear loss to downweight large residuals.
  - Uses iterative reweighted least squares (IRLS) for parameter updates.
- **MM-estimator:**
  - High breakdown point and efficiency.
  - Utilizes Peña-Yohai initialization to handle leverage and outliers.
- **LTS Regression:**
  - Trims a fraction of the largest residuals.
  - Achieves up to 50% tolerance to contamination ($\alpha = 0.5$).

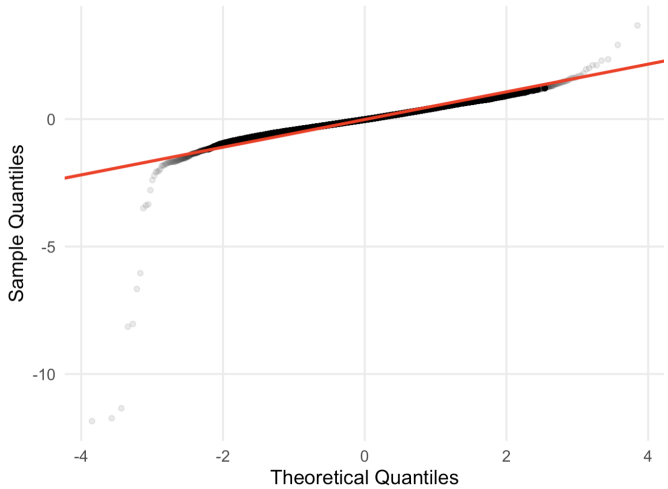
Key Points:

- Robust methods effectively downweight outliers while retaining valid leverage points.
- MM-estimator provides better leverage control and tighter residual clustering, enhancing reliability.

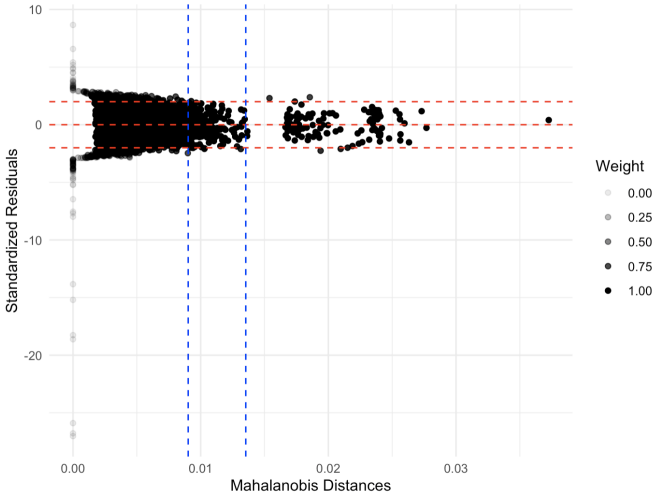
Robust Regression: Diagnostics



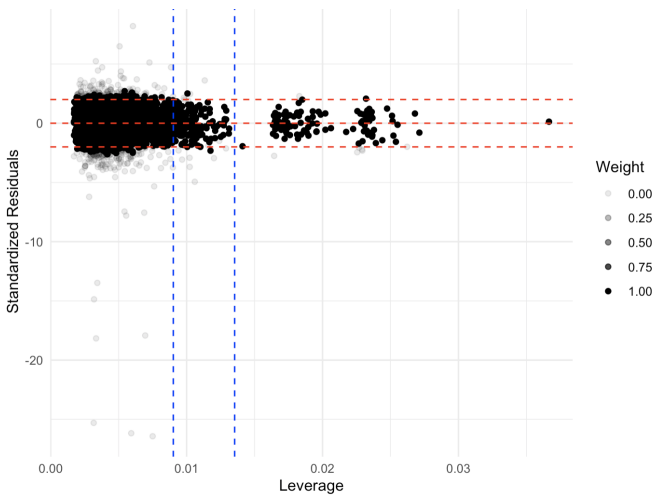
(a) Normal Q-Q Plot for MM-estimator



(b) Normal Q-Q Plot for LTS Regression



(c) Leverage vs. standardized residuals plot for MM-estimator



(d) Leverage vs. standardized residuals plot for LTS Regression

Robust Regression: Results

Model	Test MSE
OLS Regression	0.3150
Huber's M-estimator	0.3147
MM-estimator	0.3149
LTS Regression	0.3150

Key Points:

- Robust methods achieve nearly identical MSEs to OLS
- Coefficients remains stable, indicating minimal impact of outliers on prediction

Linear Mixed-Effects Model

Model Specification:

$$\texttt{log\_wage}_{ij} = \beta_0 + \beta_1 \texttt{age}_{ij} + \cdots + \beta_{11} \texttt{industry}_{ij} + u_j + \varepsilon_{ij}$$

where:

- $u_j \sim N(0, \sigma_u^2)$: Household-level random intercept.
- $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$: Individual-level residual error.

Variance Components:

- Household variance: $\sigma_u^2 \approx 0.08$
- Residual variance: $\sigma_\varepsilon^2 \approx 0.22$
- **PVRE:** 25.51% of unexplained wage variance is attributed to household-level effects.

Key Results:

- Fixed effects align closely with OLS, demonstrating robustness.
- Household random effects reveal unobserved heterogeneity but offer minimal predictive improvement.

LME Model: Results and Diagnostics

Key Results:

- **Residuals:** Follow theoretical quantiles with slight tail deviations.
- **Random Effects:** Household intercepts align with normality, with slight tail deviations.
- Robust LME models showed no improvement over standard LME in estimates or performance.

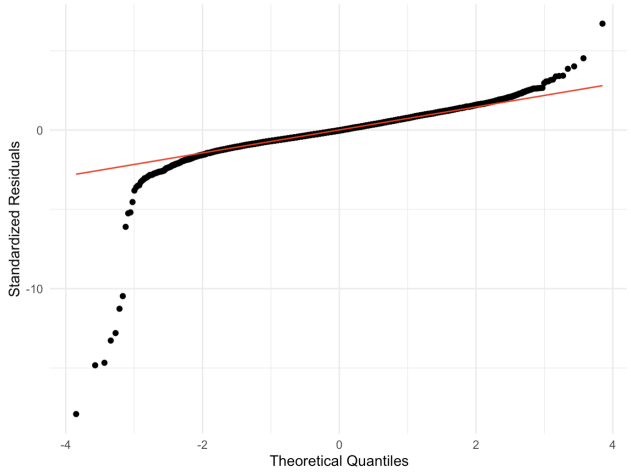


Figure: Normal Q-Q Plot for Standardized Residuals

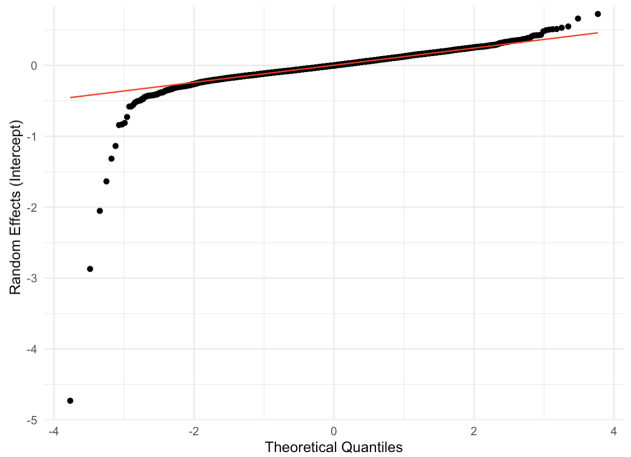


Figure: Q-Q Plot of Household Random Effects

.632 Bootstrap MSE:

- Combines in-bag and out-of-bag errors for reliable test MSE estimation.
- Steps:
  1. Fit the model to bootstrap samples (in-bag data).
  2. Compute in-bag and out-of-bag MSEs:

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

  3. Average across all bootstrap samples for final estimate.
- Applied to OLS and MM-estimator models with $B = 300$ replications.

Key Results:

- .632 bootstrap MSE is consistent with validation set results.
- MM-estimator slightly outperforms OLS.

Hypothesis Testing for Test MSE Differences

Bootstrap Hypothesis Testing:

- Constructed bootstrap distribution of MSE differences:

$$\Delta\text{MSE}^{(b)} = \text{MSE}_{\text{MM}}^{(b)} - \text{MSE}_{\text{OLS}}^{(b)}.$$

- 95% confidence interval from empirical quantiles:

$$[-0.060, 0.030].$$

- Interval includes zero: no statistically significant difference at 5% level.

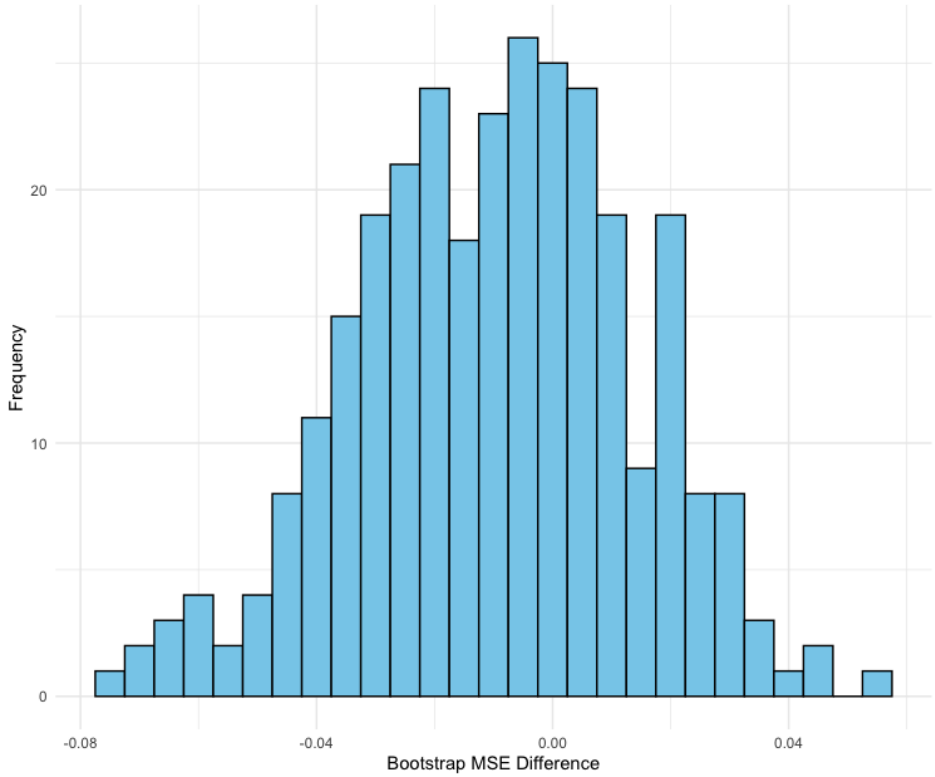


Figure: Histogram of Bootstrap MSE Differences

Analysis: Key Findings

Key Determinants of Wages:

- **Education:** Each additional year of education increases expected wages by 7.3% ($p < 0.001$), consistent with empirical evidence.
- **Age:** Wages peak at around 49 years; diminishing returns and eventual decline suggest a balance of experience and reduced labor supply.
- **Sex:** Women earn 13.7% less than men, indicating a significant gender wage gap ($p < 0.001$).
- **Race:** Black and Hispanic individuals earn 9.4% and 5.5% less than White individuals, respectively ($p < 0.05$); wage differences for Asians are not significant.
- **Metropolitan Status:** Living in metropolitan areas increases wages by 9.4% ($p < 0.001$).
- **Job Class and Industry:** Federal jobs and industries like mining and finance offer higher wages, while retail and hospitality show significant wage deficits.

Analysis: Household Effects and Variance

Household-Level Random Effects:

- **PVRE:** 25.51% of unexplained wage variance is due to household-level factors, capturing shared characteristics like economic resources or social networks.
- Variability is limited by the high proportion of single-person households (33%), reducing intra-household comparisons.
- Individual-level factors dominate wage determination, but household effects highlight socio-economic influences beyond observable variables.

Conclusion and Future Research

Conclusions:

1. Individual-level factors, such as education and age were identified as key determinants, while other demographic factors such as gender, being married, and race also significantly affect wage.
2. 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.
3. 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.

Future Directions:

- Explore unobserved household-level factors using longitudinal data.
- Study interaction effects among demographic, occupational, and regional variables for deeper insights.

Thank you!

Questions?