American Wage Statistical Analysis

July 30, 2025

Nguyen Hai Trung Pham - pham6333@mylaurier.ca

Wilfred William - will4727@mylaurier.ca

Table of Contents

Section	Page Number
1. Introduction	**2**
1.1 Purpose	2
1.2 Research Questions	2
2. Data Description	**2 - 3**
2.1 Data Summary	2
2.2 Variables and Their Descriptions	3
3. Exploratory Data Analysis	**4 - 6**
3.1 Summary Statistics	4
3.2 Distribution Plots and Visualizations	4
3.3 Preliminary Conclusions	6
4. Confirmatory Data Analysis	**6 - 12**
4.1 Model and Variable Selection Process	6
4.2 Model Diagnostics and Assumptions	8
5. Key Analyses & Results	**12**
5.1 Do Southerners Earn More or Less Than Others?	12
5.2 What Leads To Higher Wages: Education or Experience?	12
5.3 How Does Marriage Financially Impact A Man vs A Woman?	13
5.4 Financial Returns From Education Across Racial Groups	13
5.5 What Characteristics Do The Highest or Lowest Earners Have?	14
5.6 Simulating The Impact of Increased Unionization and Education	14
6. Conclusion	**15 - 16**
7. R Code	**16**

1. Introduction

1.1 Purpose

The purpose of this analytical report is to use the CPS Wage dataset to investigate and evaluate the factors that influence wages. Additionally, the application of regression techniques discussed in this course (ST362) will be used to build predictive models, test hypotheses, and find patterns. Our goal is to use statistically reliable models to determine outliers and wage relationships in the real world. Both exploratory and confirmatory approaches with multiple linear regression models will be used to discover insights, assess interaction terms, and answer our questions.

1.2 Research Questions

The following is a list of questions that our analysis will answer:

- Does geographic region (South vs. Non-South) impact wages?
- What correlates more with higher wages, more education, or more experience?
- Does marital status affect wages, and does the effect differ by gender?
- Are there differences in financial returns depending on education across races?
- What are the characteristics of the individuals with the highest and lowest wages?
- What is the impact of simulating an increase in unionization and education?

2. Data Description

2.1 Data Summary

The dataset used in this analysis originates from the Current Population Survey (CPS), a monthly household survey administered by the US Census Bureau for the Bureau of Labor Statistics. It is a primary source of information about a population's labour force characteristics, employment, income, and education. The subset used in this project contains 534 individual-level observations of sample data focused specifically on wages and demographic attributes, making it suitable for regression analysis of income factors.

2.2 Variables and Their Descriptions

There's a total of 11 variables in this dataset. Continuous, discrete, indicator, and categorical variables are included in this list, allowing for a comprehensive analysis of the factors affecting earnings. The variables that are part of the dataset are listed below, along with a description of what each of them represents:

- wage: A continuous variable that represents the wage (in dollars) of an individual.
- education: The number of years of formal education an individual has completed.
- experience: Years of work experience associated with an individual.
- age: The age of an individual in years.
- sex: An indicator variable that displays the sex of an individual, where 0 represents male and 1 represents female.
- race: A categorical variable that represents the race of an individual, where 1 represents White, 2 represents Black, and 3 represents Other.
- south: An indicator variable that displays the region an individual lives in, where 1 is living in the South U.S. and 0 is living elsewhere in the U.S.
- union: An indicator variable that displays if an individual's job is unionized, where 1 represents a union job and 0 represents a non-union job.
- marr: An indicator variable that displays an individual's marital status, where 1 represents married and 0 represents not married.
- occupation: A categorical variable that indicates the category that an individual's occupation is associated with. 1 is Professional/Technical, 2 is Managers/Administrators, 3 is Sales Workers, 4 is Clerical/White Collar Workers, 5 is Service Workers, and 6 is Manual Workers.
- sector: A categorical variable which explains the sector of a job for an individual and which indicates 0 is private, 1 is public, and 2 is not-for-profit sector.

3. Explanatory Data Analysis

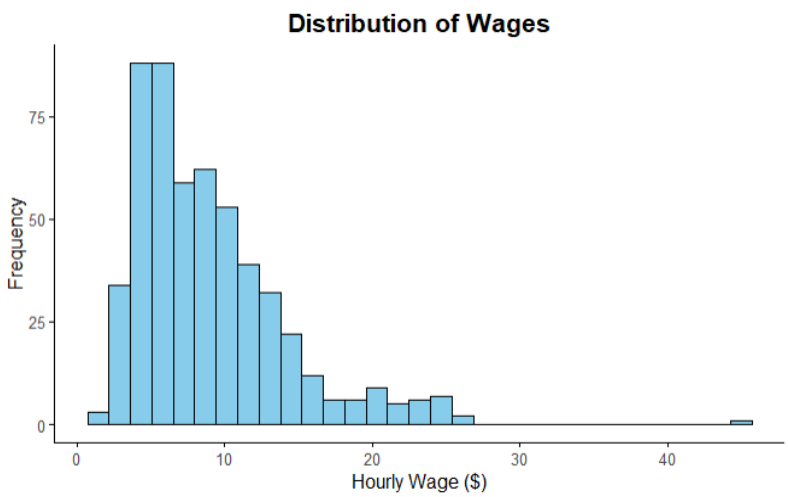
3.1 Summary Statistics

With the use of R, we can compute and extract summary statistics from the dataset. Below is a summarized output of the summary statistics:

Variable	Range	Median	Mean	Insight
Education	2 - 18	12	13.02	Most individuals do some post-secondary education
South	0 or 1	0	0.2921	29.21% of individuals are from the Southern states of America
Sex	0 or 1	0	0.4588	45.88% of the sample is female
Experience	0 - 55	15	17.82	Moderately experienced workforce
Union	0 or 1	0	0.1798	17.98% of individuals are unionized
Wage	1 - 44.5	7.78	9.024	Right-skewed distribution as the mean is greater than the median, possibly driven by a few high earners
Age	18 - 64	35	36.83	Aligns with the working-age population
Race	1, 2, or 3	3	2.699	Heavily skewed to race category 3, which implies that the majority of the sample is not white or black
Occupation	1, 2, 3, 4, 5, or 6	4	4.148	Most individuals are employed at the mid-to-upper occupational levels
Sector	0, 1, or 2	0	0.2753	Most individuals are employed in sector 0, which is the private sector
Marr	0 or 1	0	0.6554	65.54% of individuals are married

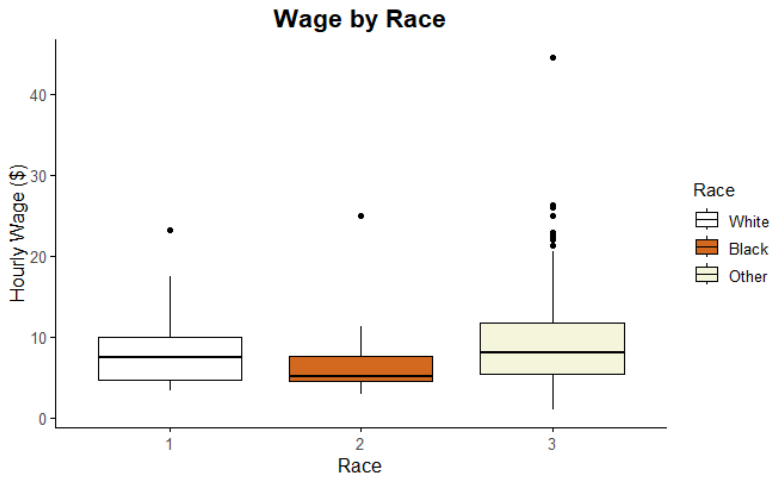
3.2 Distribution Plots

Below are the most relevant and important distribution plots:



This histogram displays the frequency distribution of hourly wages in the dataset. The distribution is right-skewed, indicating that while most individuals earn lower wages (between $5 and $15), a small number earn substantially more, creating a long tail on the right.

This boxplot compares hourly wages between males and females. Males generally have a higher median wage and greater wage dispersion. Females show a lower median, lower wage dispersion, and more outliers on the higher end, supporting the existence of gender-based wage disparities.



This boxplot demonstrates the distribution of wages among three racial categories. In the race category 3 (Other), individuals appear to show higher wages on average than in race categories 1 (White) and 2 (Black). Race category 2 has both a lower median and narrower interquartile range, demonstrating that there is a racial wage gap present in the sample.

3.3 Preliminary Conclusions

From these initial observations, we can see that:

- Wage disparities exist by sex and race. Males generally earn more than females, even after accounting for education and experience. Among races, individuals labeled "Other" tend to have higher wages than White and Black groups.

- The wage distribution is right-skewed due to a small number of high earners. While the dataset suits linear modeling, transformations and interaction terms may improve fit given the skewness and categorical variables.

- Outliers are present but do not significantly affect overall wage trends.

These findings highlight the impact of outliers, the importance of controlling for multiple variables simultaneously, and the need for multivariable regression transformation.

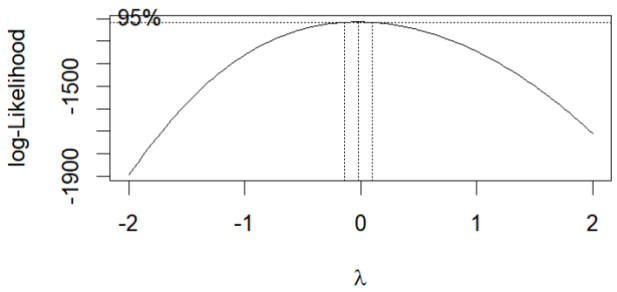
4. Confirmatory Data Analysis

4.1 Model and Variable Selection Process

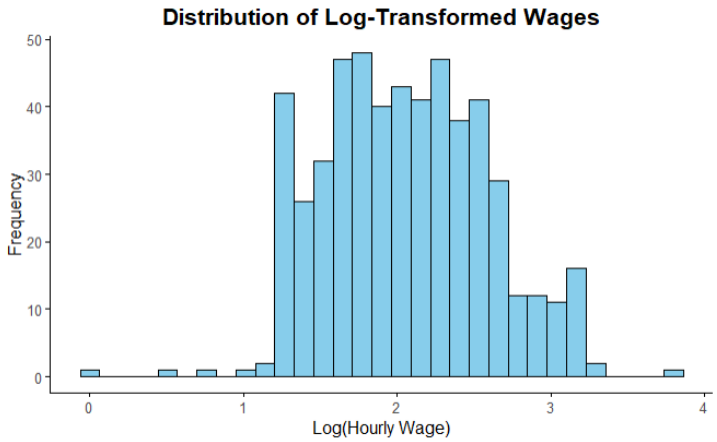
Model Transformations:

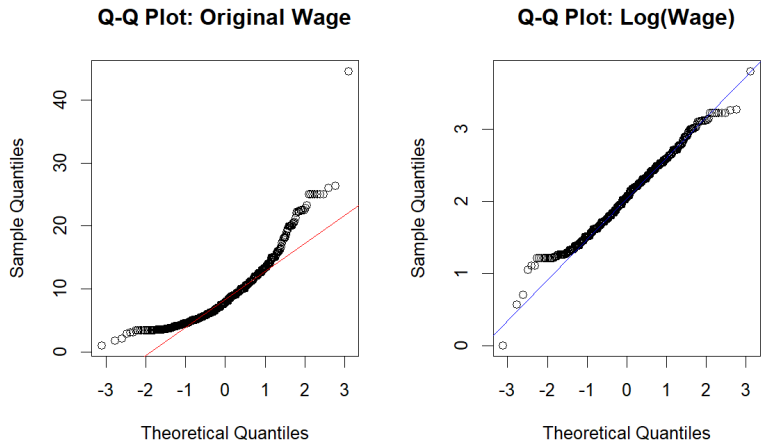
Before executing the final modeling, all categorical variables were changed to factors in R to be

encoded appropriately for regression. Due to skewness and heteroscedasticity present within the wage data, a Box-Cox transformation analysis was performed to find the most appropriate transformation. Given the Box-Cox analysis (Figure 4), the lambda was approximately 0, therefore suggesting that a log transformation should be employed on the response variable.



The log transformation yielded a slightly improved adjusted R^2 value, going from 0.30 to 0.34, and a significantly lower residual standard error, decreasing from 4.3 to 0.4 in the model. Therefore, this transformed model explains more variability, displays homoscedasticity, and is a more normally distributed model as seen in the graphs below, aligning better with the linear regression assumptions and being a more reliable model.





Influential Observations:

Outlier analysis is conducted using five influence measures: Leverage, Cook's Distance, DFFITS, DFBETAS, and COVRATIO. All observations that exceeded thresholds for three or more influence measures were removed, resulting in 28 influential points to be identified and discarded. The resulting model provided an adjusted R^2 value of 0.46 and a residual standard error of 0.36, both values once again showing increased reliability and variability in the model.

Variable Selection:

To follow up, we followed through with a stepwise AIC selection in R to identify a robust model to predict log(wage) with the fewest predictor variables. Here, the model with the lowest AIC value is considered the best-fitted model. At the end of this process, the model with the lowest AIC had a value of -1012.94, and retained the variables occupation, sex, age, education, union, south, marriage, race, and sector. Essentially, only removing experience and deeming it as not significant or contributing to the model's fit. This model has an adjusted R^2 value of 0.46 and a residual standard error of 0.36, maintaining the same values as before but with one less predictor. A stepwise BIC selection and Adjusted R^2 tests were also conducted and provided similar results. We selected the model because it has the best overall predictive accuracy and model fit, while balancing complexity and interpretability.

Final Model:

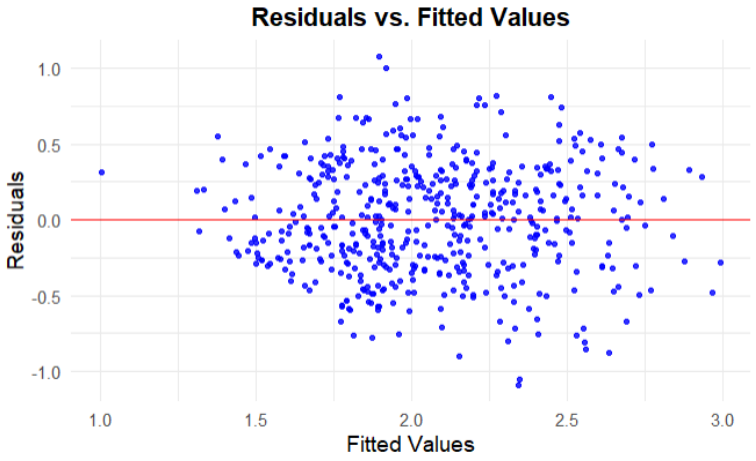
So the final model that we will be using is:

log(wage) ~ occupation + sex + age + education + union + south + marriage + race + sector

<u>4.2 Model Diagnostics and Assumptions</u>

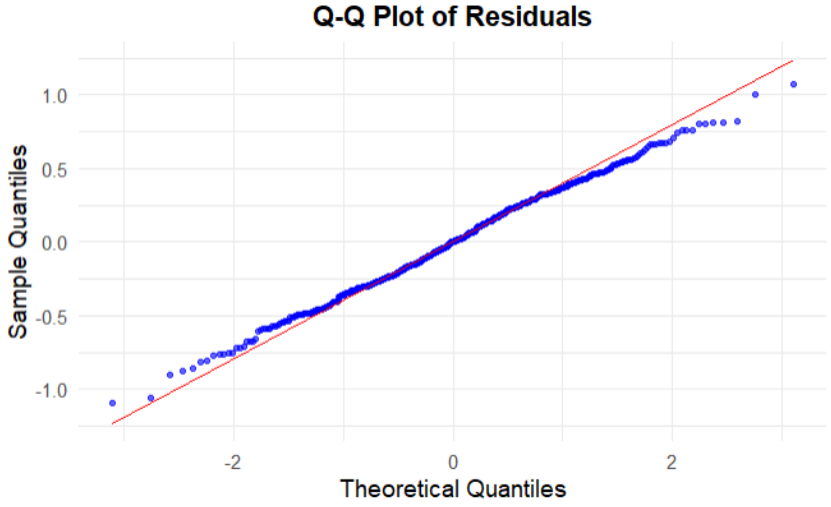
Linearity:

In the residuals versus fitted values plot below, residuals are roughly scattered around 0 without a clear pattern or curve. Although there is slightly more spread around the middle fitted values, there is no obvious funnel shape. This satisfies the assumptions of linearity between predictors and log-transformed wage.



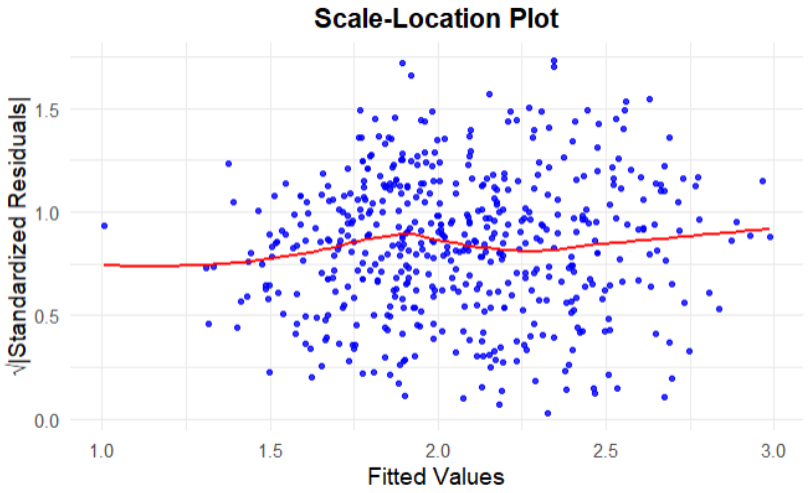
Normality:

All the points in the Q-Q plot below lie roughly on the 45-degree reference line. There are only slight deviations at the tails, but overall, this satisfies the assumption that residuals are normally distributed.



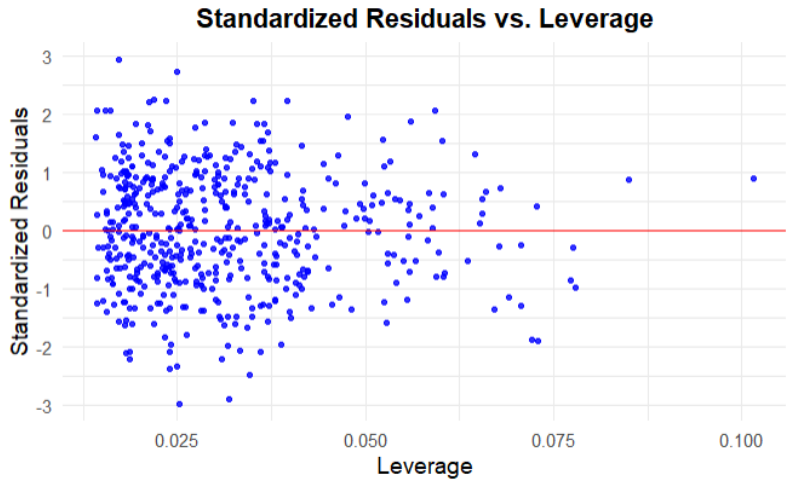
Homoscedasticity:

The scale-location plot below of standardized residuals versus fitted values is used to confirm homoscedasticity. We see in the graph below that the red line is relatively flat with a slight upward trend at higher fitted values. This indicates that there's very mild heteroscedasticity as fitted values increase. However, since the spread of residuals is primarily consistent, we can say that the assumption of constant variance has been reasonably met.



Influential Observations:

Looking at the graph below, the leverage values are mostly low, with no extreme leverage points. Standardized residuals are mostly in the magnitudinal range of 2, with no points beyond the magnitude of 3. Residuals seem randomly scattered with no obvious influential points, combining high leverage and large residuals. Overall, no high-leverage influential points are present, and the model is not being heavily influenced by individual observations.



Multicollinearity:

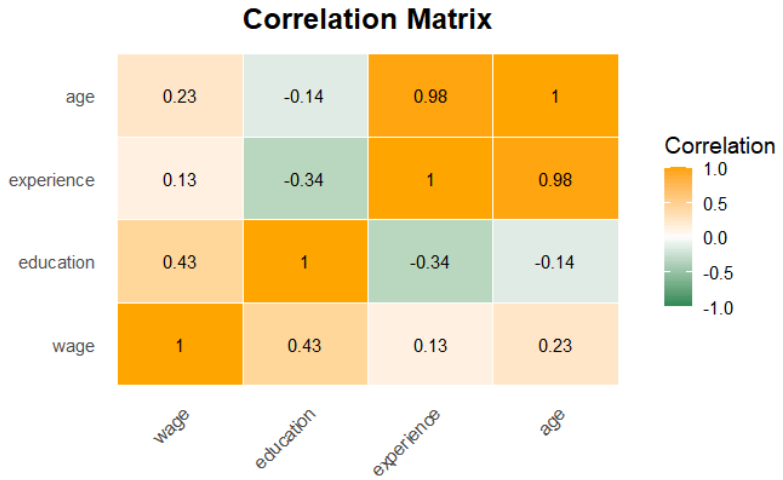
To assess multicollinearity, we must look at the VIF values. If the VIF value of a predictor is greater than 5 or 10, then it displays multicollinearity and should be removed from the model.

Predictor	DF	GVIF^(1/(2*Df))
Occupation	5	1.12
Sex	1	1.14
Age	1	1.09
Education	1	1.33
Union	1	1.07
South	1	1.03
Marriage	1	1.05
Race	2	1.02
Sector	2	1.09

Based on the analysis of the table above, the VIF values are significantly below the threshold of 5 or 10. Therefore, we do not encounter any problem with multicollinearity and don't need to remove any variables from our current model.

Correlation:

Looking at the Correlation Matrix below, we can see that all the pairs of continuous variables primarily show low correlation values. The only pair that differs from this trend is between age and experience, which has a correlation value of 0.98. This logically makes sense, as the older you are, the more time you'd have to gain experience. However, this is not a concern since "experience" is not included in the model as a predictor.



Model Fit and ANOVA:

The adjusted R^2 value of 0.4647 indicates that the model explains roughly 46% of the variation in hourly wages, which is strong given the complexity of wage determination. Since the model has a large F-statistic value of 32.21, this indicates that the model as a whole is significantly better than a model with no predictors. Also, the p-value less than 0.05 allows us to reject the null hypothesis and state that at least one of the variables has a significant effect on wages.

To assess how much each predictor contributes to the regression model, an ANOVA table was obtained. According to the ANOVA table, most predictor variables had large F-values and low p-values, with sex, age, and education proving to be the most powerful predictors. However, the Marriage, Race, and Sector variables had lower F-statistics, but still had low p-values less than 0.05, showcasing their statistical significance.

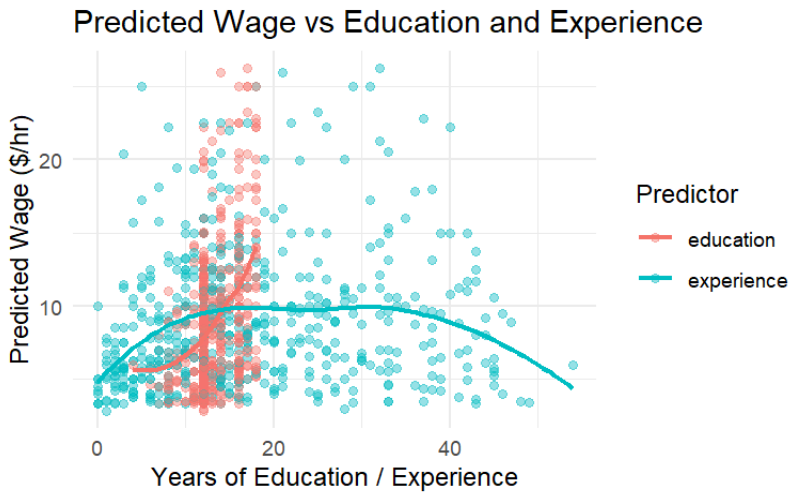
5. Key Analyses & Results

5.1 Do Southerners Earn More or Less Than Others?

By looking at the average wages grouped by region in R, it can be observed that those living in the South are earning $7.63/hour, which is considerably less than those living outside the South, who earn $9.46/hour on average. The regional difference in wages indicates that there are potentially economic differences by geography, even when controlling for the other predictors.

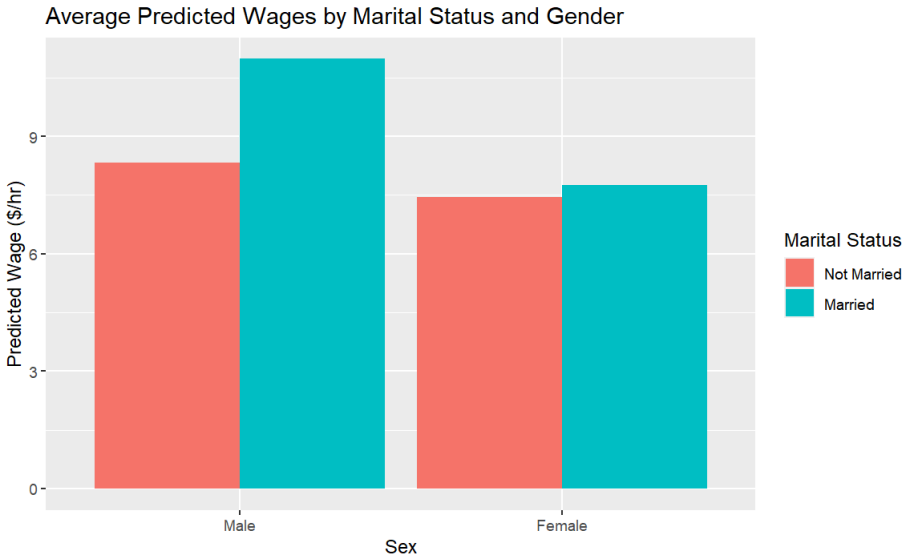
5.2 What Leads To Higher Wages: Education or Experience?

The visualization below contrasts predicted wages against education and experience, which provides an interesting comparison. Firstly, education has a positive, exponential trend, with wages climbing quickly in each later phase of education, somewhat continuously increasing. Second, experience has an inverted-U relationship: wages increase with early career experience, then eventually plateau and can even decrease after about 30 years of experience. Therefore, while both contribute to wages, education appears to offer more sustained wage growth. Experience helps early in one's career, but has diminishing returns for more experienced workers.



5.3 How Does Marriage Financially Impact A Man vs A Woman?

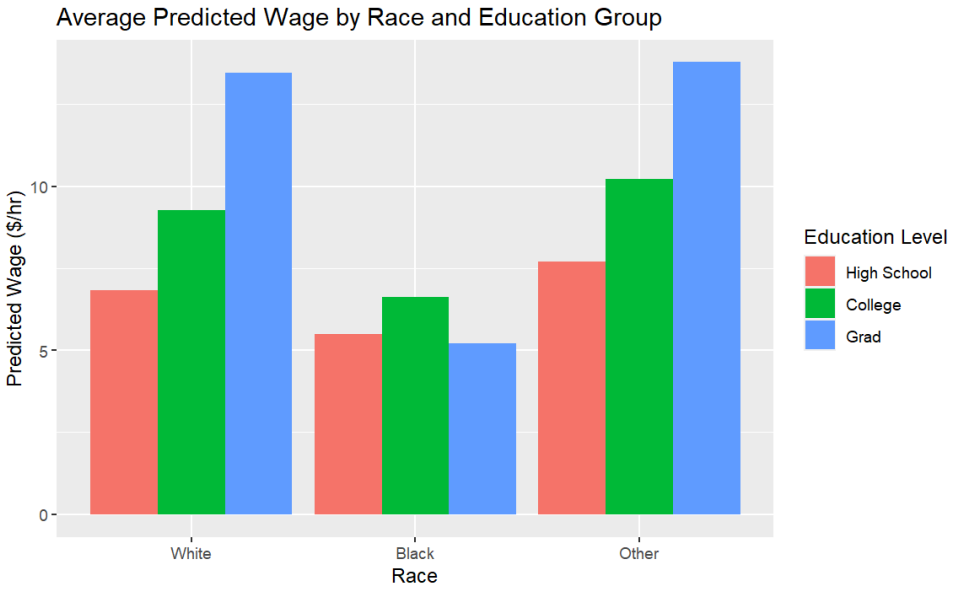
The graph below shows that men who are married earn significantly greater wages compared to their unmarried counterparts, while women only receive a small increase in wages after marriage. This difference between wages by marital status can arise from matters such as traditional gender roles, employer treatment, or uneven distribution of household responsibilities.



5.4 Financial Returns From Education Across Racial Groups

Looking at the bar chart below, it is evident that the amount of wages an individual will earn based on education varies substantially across racial groups. Individuals who are White and of other races benefit most from graduate education, with predicted wages increasing significantly.

By comparison, Black individuals have lower returns at all educational levels and even a decrease in wages at the graduate level. These patterns show that market disparities and systemic inequalities don't simply benefit White workers, but are specifically a detriment to Black people.



5.5 What Characteristics Do The Highest or Lowest Earners Have?

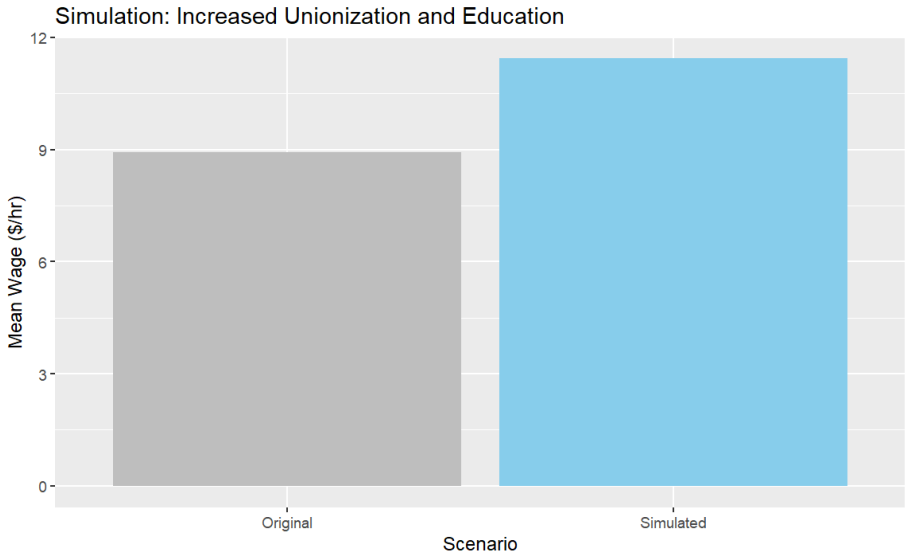
Based on the table below that showcases the characteristics of the highest and lowest outliers after cleaning the data, we see that Men in minority groups (outside of Black) are on either side of the spectrum of earnings. Both individuals also work in different occupations, however, the most notable differences are the age, education, and experience gaps. The difference between the two individuals is 5 years of education, 36 years of age, and 31 years of experience.

Characteristics of Highest vs. Lowest Predicted Wage Profiles

	education	experience	age	sex	race	union	occupation	sector	wage
169	17	32	55	Male	Other	Non-Union	Professional/Technical	Private	26.29
93	12	1	19	Male	Other	Non-Union	Manual	Private	2.85

5.6 Simulating The Impact of Increased Unionization and Education

To assess potential policy interventions, we modeled a simulation to increase unionization by 20% and average education by two years. Projected wages increased from $8.92/hour to $11.43/hour, a $2.51/hour total increase. This increase outlines the power of access to unions and investment in education on wage outcomes.



6. Conclusion

Findings

The analysis revealed that education, occupation, union membership, and sex are strong predictors of wage. The final log-linear model explained approximately 46% of the variation in wages. Males, non-Southerners, and union members generally earned more, while wage returns varied by race and education level. Simulations showed that increasing education and unionization could significantly raise average wages.

Advantages

The model improved the residual behavior due to the log transformation of the dependent variable. The final model exhibited less skewness, and the important linear regression assumptions were met with regard to multicollinearity and constant variance within groups. The exclusion of influential points in the final model helped to make the model more robust, while the step-wise procedure ensured a balance between accuracy and simplicity of model prediction.

Limitations

There were potential factors affecting the outcome that were unmeasured in the model. For example, industry or workplace impacts, job performance, and cost of living, there were assignments of linear relationships in all cases, which could oversimplify the real-world behavior of these relationships. Experience was omitted as a factor because of its strong correlation with age, however, it is a major factor we acknowledged was important to consider.

Discussion

Overall, the model stated about the wages is straightforward, interpretable, and still identifies and conceptualizes the contractors we found. It efficiently contrasts education, union availability, and demographic factors. Future work can support an alternative to the simple model that can investigate the use of non-linear models, which allow the subtlety of available info to be modeled with fewer assumptions, and/or include interaction terms likely to improve the final model's explanatory power while maintaining the advantages of a more parsimonious model and evaluative framework.

7. R Code

https://drive.google.com/file/d/1VOAKBA7DSK1kwf6-S7Dy6OC9CGo-En6s/view?usp=sharing