Analysing the factors influencing the wages of the US workforce*

Education as the Strongest Factor While Gender and Race Reflect Inequalities

Jianing Li

November 26, 2024

This study uses a multiple linear regression model to analyze key factors influencing wages among the U.S. workforce, including education, age, gender, race, region, and hours worked. The results show that education is the strongest factor, with higher levels of education leading to significantly higher wages, while gender and race reveal notable disparities, with men and Asian workers earning the most. The study also finds a nonlinear relationship between age and wages, with region and hours worked playing smaller but significant roles. By identifying these drivers, the research provides evidence to inform policies aimed at reducing wage gaps and promoting fairer economic outcomes.

1 Introduction

Income inequality is a persistent social and economic problem with significant implications for individual well-being and societal stability. Understanding how personal characteristics and external factors shape income distribution is essential to addressing its underlying causes. While many studies have explored income inequality, few have systematically analyzed the interaction of these factors. Additionally, existing research often lacks recent data that are both nationally representative and adaptable for diverse analyses. This study addresses these gaps using the 2023 sample from the IPUMS database, which allows users to customize variables and provides high-quality, comprehensive data. Leveraging this resource, the study identifies the main determinants of income inequality through a multiple linear regression model.

The estimand in this study is income. Specifically, it quantifies the effects of education, age, gender, race, region, and hours worked on income. By measuring these effects, the study

---

seeks to uncover the mechanisms driving income inequality and provide a framework for understanding how personal characteristics and external conditions interact to influence income distribution.

The study finds that education is the strongest determinant of income. Higher levels of education lead to significantly higher earnings, with those with a college degree earning much more than other groups. The relationship between age and earnings is non-linear, with earnings peaking between the ages of 40 and 50 before declining, reflecting the impact of the occupational life cycle. In addition, gender and race show disparities, with men and Asians earning the highest average incomes, while women and certain minority groups earn less. Region and hours worked also affect earnings, but their effects are relatively small. Regional income differences reflect differences in economic development and policy environments, while hours worked show diminishing returns beyond a certain point. The findings are important because they reveal the role of education, gender, and race in driving income inequality and provide clear directions for policy interventions. They also provide empirical support for strategies to reduce income inequality, such as improving the allocation of educational resources, promoting gender equality, and increasing economic opportunities for minority groups.

This paper provides a comprehensive analysis of income inequality by examining its key drivers and impacts. Section 2 describes the dataset in detail, highlights its main features, and uses visualizations to present the analysis results. Section 3 focuses on the construction of the multiple linear regression model, validates its effectiveness, and presents the results of the analysis. Section 4 analyzes the research findings, explores the reasons behind the observed patterns, and offers policy recommendations to address income inequality. It also identifies the study's limitations and suggests directions for future research. Section A provides additional details on the model, and methods of survey and sampling, offering more comprehensive support for the study.

2 Data

2.1 Data Overview

We used data from the IPUMS USA database (Ruggles et al. 2024). All analyses were conducted in R (R Core Team 2023). Data simulation, testing, and cleaning were implemented using `tidyverse` (Wickham et al. 2019) and `testthat` (Wickham 2011), with specific tasks performed using `here` (Müller 2020), `arrow` (Richardson et al. 2024), and `readr` (Wickham, Hester, and Bryan 2024). Data manipulation was carried out with `dplyr` (Wickham et al. 2023) and `reshape2` (Wickham 2007), while data visualization was done using `ggplot2` (Wickham 2016), adjusted with `scales` (Wickham, Pedersen, and Seidel 2023). Model training and performance evaluation were supported by `caret` (Kuhn and Max 2008), with model testing and analysis conducted using `broom` (Robinson, Hayes, and Couch 2024). Code formatting and

table presentation were handled by `knitr` (Xie 2024) and `kableExtra` (Zhu 2024), respectively, while regression diagnostics and additional analysis were performed with `car` (Fox and Weisberg 2019).

The IPUMS USA database (Ruggles et al. 2024) is one of the largest collections of microdata from population censuses globally. Supported by organizations such as the National Institutes of Health and the University of Minnesota, it includes data from the American Community Survey (ACS) and other census programs. This resource provides detailed individual-level data, allowing researchers to address specific social, economic, and demographic questions. A key feature of IPUMS USA is its capacity to create customized datasets by selecting variables based on research needs, minimizing the inclusion of irrelevant data.

For this study, data from the 2023 ACS sample in IPUMS USA were used. The selected variables include STATEFIP (state code), SEX (gender), AGE (age), RACE (race), EDUC (educational attainment), EMPSTAT (employment status), UHRSWORK (usual hours worked per week), and INCWAGE (wage and salary income). These variables cover geographic location, demographic characteristics, education, and employment conditions, providing a broad basis for examining factors influencing wage income.

Other databases, such as IPUMS CPS (Flood et al. 2024) and the National Longitudinal Surveys (Bureau of Labor Statistics 2023), were considered but deemed less suitable for this analysis. IPUMS CPS provides detailed labor market data but has a smaller sample size and limited geographic detail. The National Longitudinal Surveys track specific population groups over time but lack broad representation, making them less ideal for cross-sectional studies. In contrast, the 2023 ACS data in IPUMS USA offers larger population coverage and a diverse range of variables, making it a better choice for analyzing wage determinants.

2.2 Measurement

This study utilizes sample data from the 2023 ACS. The dataset is based on a 1% random national sample, ensuring broad representativeness. It includes individuals living in private households as well as those in group quarters, such as student dormitories and care facilities. Weights were applied to the data to account for potential biases introduced by the sampling design and nonresponse.

Data collection for the ACS involved structured questionnaires covering a wide range of topics, such as demographics, education, employment, and income. For example, UHRSWORK (usual hours worked per week) captures labor input based on respondents' reports of their average weekly work hours. The smallest geographic unit in the dataset is the Public Use Microdata Area (PUMA), which contains at least 100,000 residents and falls within state boundaries. This structure protects respondent privacy while supporting regional analysis.

Table 1: Cleaned Data Overview

INCWAGE	UHRSWORK	education_level	region	age	gender	race_group
105000	40	Above_Bachelor	Northeast	49	Female	Asian
73000	40	Bachelor	South	29	Male	White
30000	45	Below_High_School	West	45	Female	Other
37000	40	High_School	West	27	Male	Other
1200	10	Some_College	South	22	Male	Black

While the ACS methodology ensures data reliability and extensive coverage, variables like income and work hours rely on self-reported responses, which may introduce minor inaccuracies. Nonetheless, the use of sampling weights and robust design make the dataset a reliable foundation for studying wage income and its associated factors. A detailed explanation of the survey methodology and sampling design is provided in the appendix.

2.3 Dataset Description

The sample of the cleaned dataset is shown in Table 1, including the following variables. UHRSWORK and INCWAGE are retained from the raw data, while the other variables are newly constructed. During the data cleaning process, only individuals aged 18-65 and those with an employment status of "Employed" were included. This ensures the dataset focuses on studying the wage determinants of the employed labor force. Below is a description of each variable and how it was constructed:

- **UHRSWORK**: Reports the number of hours per week the respondent usually worked if they were employed during the reference period.

- **INCWAGE**: Represents each respondent's total pre-tax wage and salary income earned as an employee during the previous year.

- **education_level**: Derived from the EDUC variable, which indicates respondents' educational attainment as measured by the highest year of school or degree completed. It is grouped into five categories: Below High School (includes all levels below high school), High School (completed grade 12), Some College (completed 1-2 years of college), Bachelor's Degree (completed 4 years of college), and Above Bachelor (more than 5 years of higher education).

- **age**: Filtered from the AGE variable to include only respondents aged 18 to 65. This variable records the respondent's age in years as of their last birthday.

- **region**: Constructed from the STATEFIP variable to represent geographic regions. Valid state codes (1–56) were grouped into four traditional U.S. regions: Northeast, Midwest, South, and West. Each state was assigned to its corresponding region for further regional analysis.

4

- **gender**: Based on the SEX variable, with original codes recoded as "Male" and "Female."

- **race_group**: Constructed from the RACE variable and grouped into four categories: White, Black, Asian, and Other. White and Black remain consistent with the original classifications. Asian includes Chinese, Japanese, and Other Asian categories, while Other includes all other races and mixed-race groups.

2.4 Data Results

Figure 1 illustrates the distribution of demographic characteristics and their relationship with average income across gender, race, region, and education level. Pie charts display the proportional distribution of each group, while bar charts show corresponding average income levels, providing a straightforward depiction of these relationships.

Chart (a) shows the gender distribution: females account for 48.9% of the population, and males 51.1%, indicating a nearly equal split. Chart (b) shows that males earn an average income of $82,840, compared to $57,146 for females.

Chart (c) presents the racial composition, with White individuals forming the largest group at 67%, followed by Black individuals at 18.2%, Asians at 6.8%, and other racial groups at 8.1%. Chart (d) shows that Asians have the highest average income at $96,574, followed by White individuals at $73,778. Black individuals and other racial groups earn $51,221 and $56,604, respectively.

Chart (e) shows the regional distribution. The West represents the largest share at 29%, followed by the Northeast (26.6%), the South (24.1%), and the Midwest (20.3%). Chart (f) indicates that the Midwest has the lowest average income at $62,046, while the Northeast has the highest at $78,138. The South and West have average incomes of $66,954 and $75,293, respectively.

Chart (g) shows education levels: 16.5% of the population has less than a high school education, 30.4% has completed high school, 23.5% has some college education, 26.2% holds a bachelor's degree, and 9.5% has education beyond a bachelor's degree. Chart (h) highlights the connection between education and income. Those with less than a high school education earn $36,149 on average, while individuals with a bachelor's degree or higher earn $87,465 and $124,144, respectively. Income increases steadily with higher levels of education.

Figure 2 examines the relationship between income, work hours, and age, incorporating both individual-level data and aggregated trends.

Chart (a) is a scatter plot of work hours versus income, with each point representing an individual. The red quadratic regression line indicates a positive correlation between income and work hours, where longer hours generally lead to higher income. However, the regression line primarily reflects a linear trend, suggesting a stable relationship between income and work hours without significant signs of deceleration. Additionally, income variability increases with
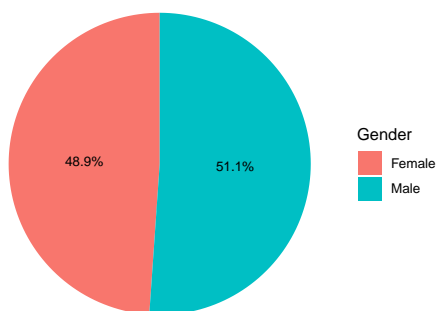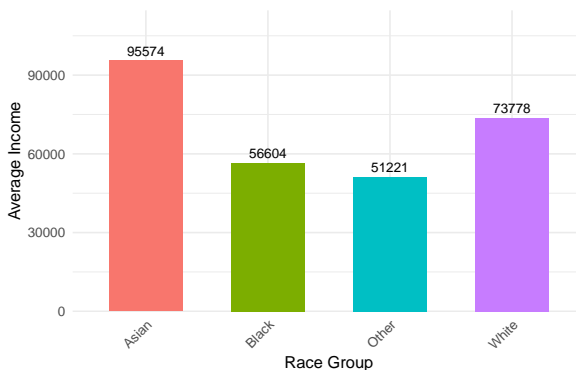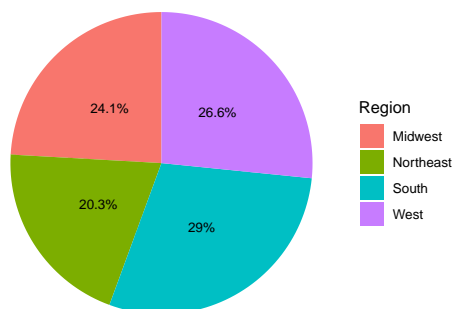
(a) Gender Proportion
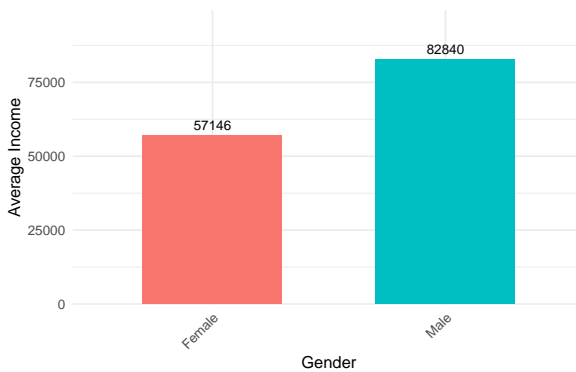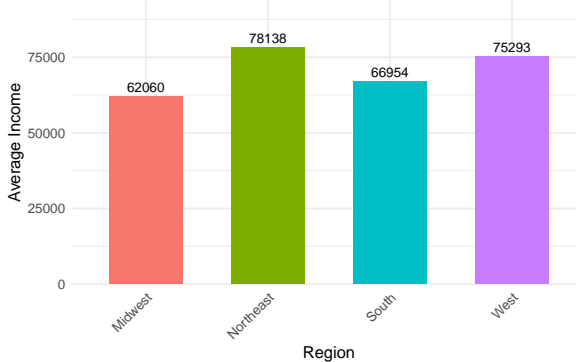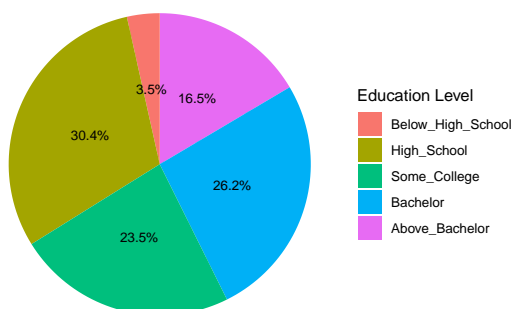
(b) Average Income by Gender
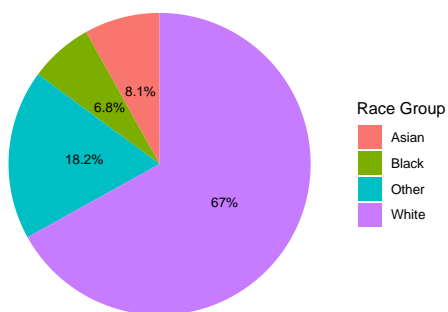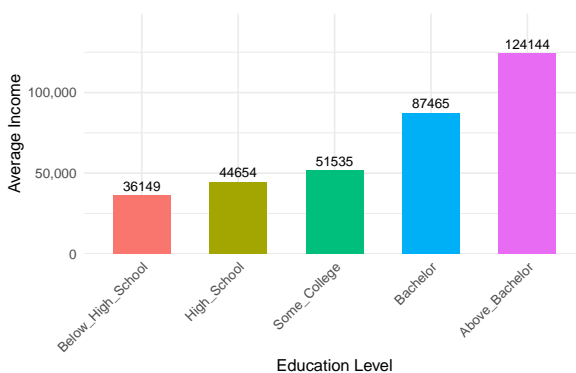
(c) Race Proportion

(d) Average Income by Race
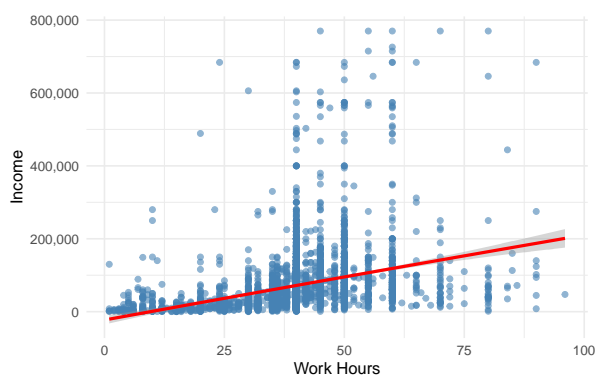
(e) Region Proportion

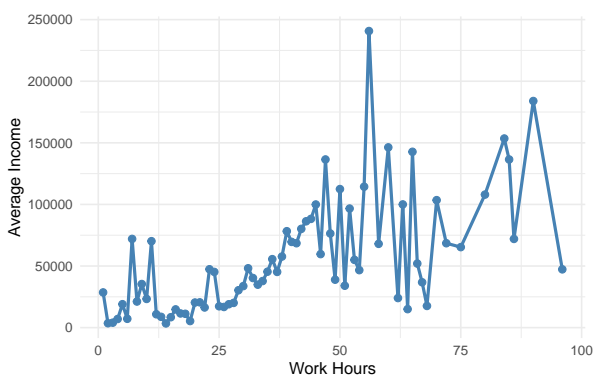(f) Average Income by Region

(g) Education Proportion

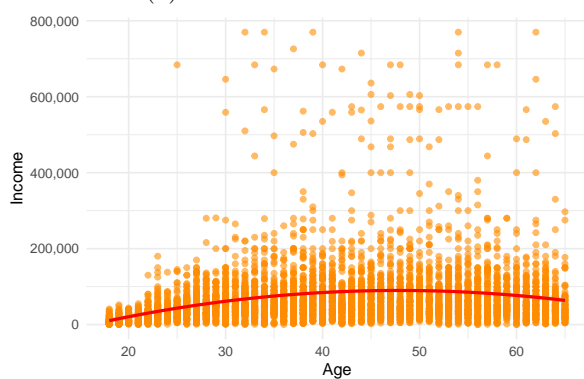(h) Average Income by Education Level

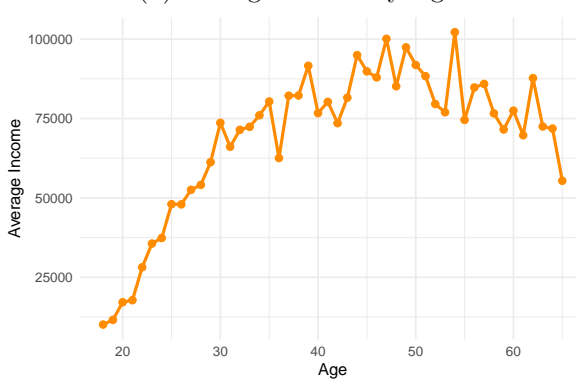Figure 1: Demographics Distribution and Average Income by Group

(a) Work Hours vs Income

(b) Average Income by Age

(c) Age vs Income

(d) Average Income by Work Hours

Figure 2: Income Trends by Age and Work Hours

longer work hours, particularly beyond 50 hours per week, where fluctuations become more pronounced. Chart (b) illustrates the relationship between work hours and average income. While income generally increases with work hours, especially between 20 and 50 hours per week, it becomes unstable beyond 50 hours.

Chart (c) is a scatter plot of age versus income, with the red quadratic regression line showing that income increases with age before eventually declining. Chart (d) depicts the trend of average income by age. The line chart shows that average income steadily rises with age, peaking in the late 40s and early 50s, and gradually declines as individuals approach retirement age.
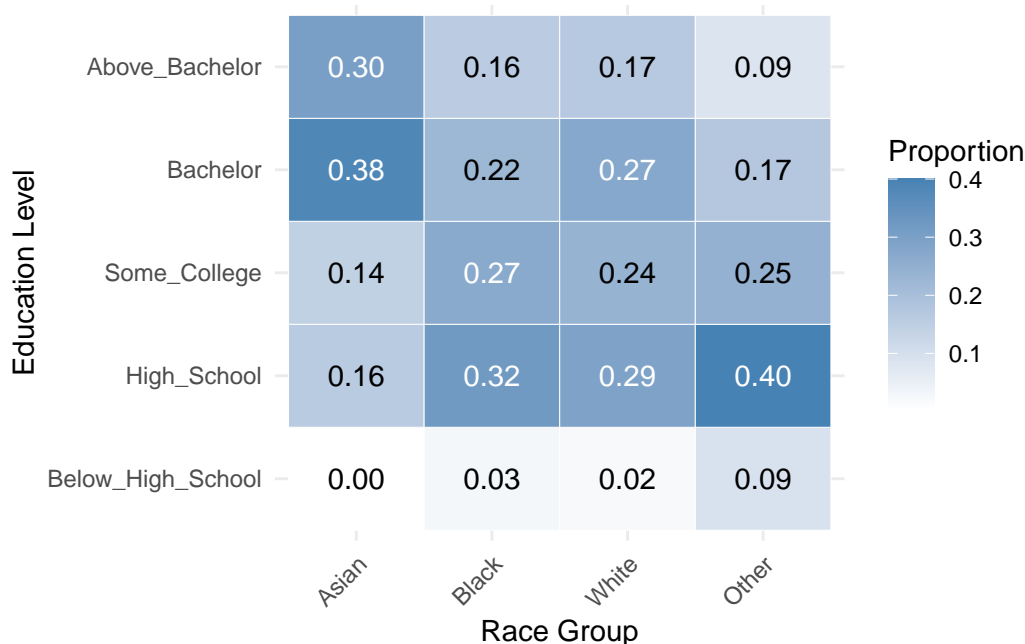

Figure 3: Proportion of Education Levels by Race Group

Figure 3 illustrates the proportion of different education levels across racial groups. The horizontal axis represents racial groups, including "Asian," "Black," "White," and "Other," while the vertical axis represents education levels, ranging from "Below High School" to "Above Bachelor." The intensity of blue indicates the proportion, with darker shades representing higher values.

Among Asians, the proportion of highly educated individuals is relatively high, with 38% holding a bachelor's degree and 30% having education beyond a bachelor's degree, while the proportion of those with "Below High School" education is 0%. In the Black group, "High School" accounts for the largest proportion at 32%, followed by "Some College" at 27%, while those with "Above Bachelor" degrees make up only 16%. In the White group, "High School" accounts for 29%, with 27% holding a bachelor's degree and 24% having "Some College"

education, while 17% have "Above Bachelor" degrees. Among the "Other" group, 40% have a "High School" education, 25% have "Some College," 17% hold a bachelor's degree, and only 9% have "Above Bachelor" degrees.

Overall, the distribution of education levels varies across racial groups. Asians have a higher proportion of highly educated individuals compared to other groups, while "High School" education is more prevalent among Black and "Other" groups. This reflects the complex relationship between race and educational background.

3 Model

The goal of our modelling strategy is to quantify the contributions of key factors—such as weekly working hours, age (and its quadratic term), education level, region, gender, and race—to variations in income (log-transformed). By estimating the relative impact of these predictors, the model provides insights into potential social and demographic disparities in income.

Here we briefly describe the multiple linear regression model used to investigate these relationships. The model captures both linear and nonlinear effects (e.g., through the quadratic term for age), and all categorical variables are treated as factors to account for group-level differences.

3.1 Model set-up

This study use Multiple Linear Regression (MLR) to model the relationship between income and various predictors, implemented using R's `lm` function. The dataset is divided into training and testing sets using the `createDataPartition` function from the `caret` package, with 60% allocated for model training and parameter estimation and 40% for testing to evaluate predictive performance.

Multiple regression models rely on several assumptions: linearity, meaning a linear relationship exists between independent and dependent variables; homoskedasticity, requiring constant error variance across predictors; independence of errors, meaning residuals are uncorrelated; normality, where residuals follow a normal distribution; and independence of independent variables, ensuring no significant multicollinearity. These assumptions are discussed in Section A.1 and evaluated using the diagnostic plots shown in Figure 5.

The final model is displayed below:

$$
\begin{aligned}
\log(\text{INCWAGE}_i) = {} & \beta_0 + \beta_1 \text{UHRSWORK}_i + \beta_2 \text{region}_i \\
& + \beta_3 \text{education\_level}_i + \beta_4 \text{age}_i + \beta_5 \text{age}_i^2 \\
& + \beta_6 \text{gender}_i + \beta_7 \text{race\_group}_i + \epsilon_i
\end{aligned} \tag{1}
$$

Table 2: Training and Testing Data Evaluation Results

Metric	Training	Testing
RMSE	0.775	0.782
MAE	0.552	0.555
R^2	0.469	0.495

- β_0 is the coefficient for the intercept.

- β_1 is the coefficient for the continuous variable UHRSWORK_i, which measures weekly hours worked.

- β_2 is the coefficient corresponding to the categorical variable region_i, which includes the levels Midwest, Northeast, South, and West. Midwest is the reference level.

- β_3 is the coefficient corresponding to the categorical variable education_level_i, which includes the levels Above Bachelor, Bachelor, Below High School, High School, and Some College. Above Bachelor is the reference level.

- β_4 and β_5 are the coefficients for the linear and quadratic terms of age_i, capturing the nonlinear relationship between age and income.

- β_6 is the coefficient for the binary variable gender_i, which includes the levels Male and Female. Female is the reference level.

- β_7 is the coefficient corresponding to the categorical variable race_group_i, which includes the levels Asian, Black, Other, and White. Asian is the reference level.

- ϵ_i is the error term, capturing the deviation of the observed value from the predicted value due to unobserved factors.

3.1.1 Model justification

The selection of variables and model structure was based on theory and data characteristics. Given the right-skewed nature of the income distribution in Figure 4, the dependent variable is log-transformed (`log(INCWAGE)`) to improve model fit. Furthermore, to capture the nonlinear relationship between age and income evident in Figure 2, a quadratic term for age (`age^2`) is included in the model. Categorical variables, such as region, education level, race group, and gender, are represented using dummy variables. Each category is compared to a reference category, allowing the regression model to capture and interpret group-level differences.

Table 3: Estimated coefficients of the final model

	Coefficients	Lower.CI	Upper.CI
(Intercept)	7.1943	6.9446	7.4440
UHRSWORK	0.0400	0.0380	0.0421
regionNortheast	0.1287	0.0632	0.1941
regionSouth	-0.0192	-0.0794	0.0410
regionWest	0.1560	0.0927	0.2192
education_levelBachelor	-0.2131	-0.2814	-0.1448
education_levelBelow_High_School	-1.0844	-1.2151	-0.9537
education_levelHigh_School	-0.7492	-0.8182	-0.6802
education_levelSome_College	-0.5750	-0.6462	-0.5037
age	0.0985	0.0869	0.1101
I(age^2)	-0.0010	-0.0011	-0.0008
genderMale	0.2632	0.2186	0.3078
race_groupBlack	-0.1014	-0.2166	0.0137
race_groupOther	-0.1290	-0.2230	-0.0350
race_groupWhite	-0.0007	-0.0834	0.0821

During the model selection process, an initial model included an interaction term between education level and race group (education_level:race_group), based on the association observed in Figure 3, while other variables remained unchanged. This interaction term aimed to evaluate income differences across racial groups at the same education level. However, most interaction terms were not statistically significant and provided minimal improvement to model fit metrics, such as adjusted R^2 and AIC. To simplify the model and enhance interpretability, this interaction term was excluded from the final specification.

The evaluation of the final model's performance was conducted on both training and testing datasets. As shown in tbl-test, the RMSE and MAE values are slightly higher for the testing data compared to the training data (0.786 vs. 0.773 for RMSE, and 0.568 vs. 0.543 for MAE), reflecting a modest decline in predictive accuracy when applied to unseen data. Similarly, the R^2 values are 0.488 for the training set and 0.468 for the testing set, indicating that the model captures a moderate proportion of the variability in the log-transformed income. Additionally, all variables included in the final model were statistically significant, with p-values below 0.05, further supporting the robustness of the selected predictors. Overall, the final model achieves a practical balance between complexity, interpretability, and predictive performance.

3.1.2 Model Results

Table 3 presents the effects of weekly working hours, age, region, education level, gender, and racial group on log income, along with their 95% confidence intervals. The estimated coefficients indicate varying degrees of influence among these factors.

Education level has the strongest impact on income. Compared to those with a high school education (the reference group), individuals with education beyond a bachelor's degree have a coefficient of 1.084, showing a significant increase in income. Similarly, those with a bachelor's degree and some college education have coefficients of 0.8713 and 0.5094, respectively, suggesting higher education correlates with higher income.

Age and its quadratic term reveal a nonlinear relationship. The positive coefficient for age (0.0985) indicates income increases with age. However, the negative coefficient for the quadratic term (-0.0010) shows this growth slows over time, eventually declining. This pattern reflects typical career income trajectories, with rapid growth early on, stability in middle age, and potential decreases later in life.

Gender also plays a significant role. Male income is significantly higher than female income, with a coefficient of 0.2632, even after controlling for other factors.

The effect of region varies. Compared to the Midwest (reference group), incomes in the West and Northeast are higher, with coefficients of 0.1560 and 0.1287. The coefficient for the South is -0.0192, close to zero, indicating little difference. This may relate to regional economic conditions and policies.

Weekly working hours show a modest effect on income, with a coefficient of 0.040. While longer hours increase income, the impact is relatively small, potentially due to wage structures and overtime policies.

The effects of racial group are more complex. Compared to Asian individuals (reference group), income for Black and "Other" racial groups is notably lower, with coefficients of -0.1014 and -0.1290. The coefficient for White individuals is nearly zero (-0.0007), indicating no significant difference. These patterns reflect complexities in racial income disparities.

In summary, education level has the largest influence on income, followed by age and gender. Region and racial group also play roles, while the effect of weekly working hours is smaller. These results provide a clear picture of how socioeconomic factors affect income and offer important context for further research and policymaking.

4 Discussion

This paper uses a multiple linear regression model to analyze key factors that influence income, including education, age, gender, race, region, and hours worked. The study is based on data from the 2023 American Community Survey (ACS), cleaned and processed for visualization and analysis. The model quantifies the impact of these variables and explores their significance, showing how personal characteristics and external factors shape the income distribution. Education, age, gender, and race are treated as endogenous variables, reflecting individual attributes, while region and work hours are considered exogenous, capturing the influence of external conditions on income. Through the analysis of these variables, the study

provides insights into the roots of income inequality and offers evidence to inform policymaking.

4.1 Education and Age: Core Drivers of Income

Education is a key factor influencing income, with a clear positive correlation between education level and earnings. Graduate degree holders earn significantly more, while a bachelor's degree also boosts income. However, those with only 1-2 years of college earn similar incomes to high school graduates, underscoring the importance of completing college. Higher education plays a crucial role in enhancing economic potential and securing high-paying jobs. Most income differences stem from variations in bachelor's, graduate, and doctoral education levels, while pre-college education has a smaller impact.

Age also has a nonlinear impact on income. The model indicates that income peaks between ages 40 and 50 before gradually declining. The quadratic term captures this trend, reflecting how different stages of a career affect earnings. Early in their careers, individuals experience rapid income growth as they gain experience and promotions. In later stages, productivity declines, and income typically decreases as retirement approaches.

Education and age shape an individual's earning potential. These findings are valuable for career planning and offer guidance for policymakers. Individuals can enhance their competitiveness in the job market by pursuing a bachelor's degree or higher. Governments can increase education funding and provide financial aid to expand access to higher education. Additionally, support programs for those nearing retirement can help mitigate the impact of declining wages on their quality of life.

4.2 Gender and Race: Social Inequalities in Income Distribution

Gender inequality in income distribution is evident. Studies show that men earn significantly more than women, even when controlling for education, age, and other factors. This disparity may stem from gender bias and occupational segregation. Women are often expected to take on more family responsibilities and face potential career interruptions due to childbirth (Francine D. Blau and Kahn 2017). Additionally, men are overrepresented in many high-paying professions, contributing to lower average wages for women (Francine D. Blau, Brummund, and Liu 2013).

Racial disparities, on the other hand, highlight another form of structural inequality. Research indicates that Asians earn the highest incomes, followed by Whites, while Black individuals and those from "other" racial groups earn the least. These differences can be attributed to historical and systemic factors, such as racial discrimination in the labor market and the underrepresentation of minorities in high-paying industries (Queneau 2009). Educational disparities between racial groups, as shown in Figure 3, also contribute to these income gaps.

Gender and race are exogenous factors, beyond individual control, and the wage disparities they cause reflect systemic inequality. Addressing the gender pay gap requires targeted policies, such as promoting pay transparency and increasing women's participation in high-paying industries. To reduce racial income disparities, policymakers can implement measures like providing education and employment support for minority groups, encouraging diversity hiring, and strictly enforcing anti-discrimination laws. Cultural change is also crucial; reducing biases against women and racial groups can help narrow wage gaps and promote fairness.

4.3 Region and Working Hours: External Environmental Impact on Income

Region affects income. Studies show that incomes are higher in the Northeast and West and lowest in the Midwest. This may be due to differences in economic development and industry structures. Developed regions offer more high-paying jobs, while less developed areas have limited options. To reduce regional income gaps, the government can support industries in the Midwest and attract high-value sectors. Improving infrastructure and better allocating educational resources can also create more quality jobs in underdeveloped areas.

Work hours are positively linked to income, but the returns diminish. Beyond 50 hours per week, income becomes more unstable. This suggests that working longer hours does not always lead to higher pay. Policies like overtime pay rules or limits on work hours may play a role. The government should introduce better labor laws, such as capping excessive work hours. Encouraging businesses to innovate and improve efficiency can reduce reliance on long hours. These measures can protect workers' rights while promoting a fairer and healthier labor market.

4.4 Weaknesses and next steps

The data used in this study has certain limitations. It relies on survey responses, which can include self-reporting bias. For example, respondents may overestimate or underestimate their income or working hours. Additionally, the original data collection only classified gender as male or female. Non-binary and transgender groups were not included, leaving some populations unrepresented. Future research could incorporate more diverse datasets, such as administrative records or third-party data, to improve accuracy. Using more detailed classifications could also better reflect overlooked groups and enhance inclusivity.

Some important variables were not included in the analysis, such as industry classification. While industry significantly affects income, its complex structure and lack of standardized categories made it difficult to analyze. This exclusion might have introduced bias into the results. To simplify the model, smaller racial groups, such as Indigenous, and multiracial individuals, were categorized as "Other." Although this made the analysis manageable, it overlooked finer details in racial differences. Future studies could refine these categories and include more detailed variables when sample sizes allow.

The model also has limitations. With fewer variables, it cannot fully explain the complexities of income distribution. This study is based on U.S. data from 2023, so its findings may not apply to other countries or time periods. Future research could include more variables, such as industry, family background, and regional policies, to improve explanatory power. Combining time-series and cross-country data would also help study income trends and disparities on a broader s1cale.

A Appendix {-}

A.1 Model details

(a) Income Distribution

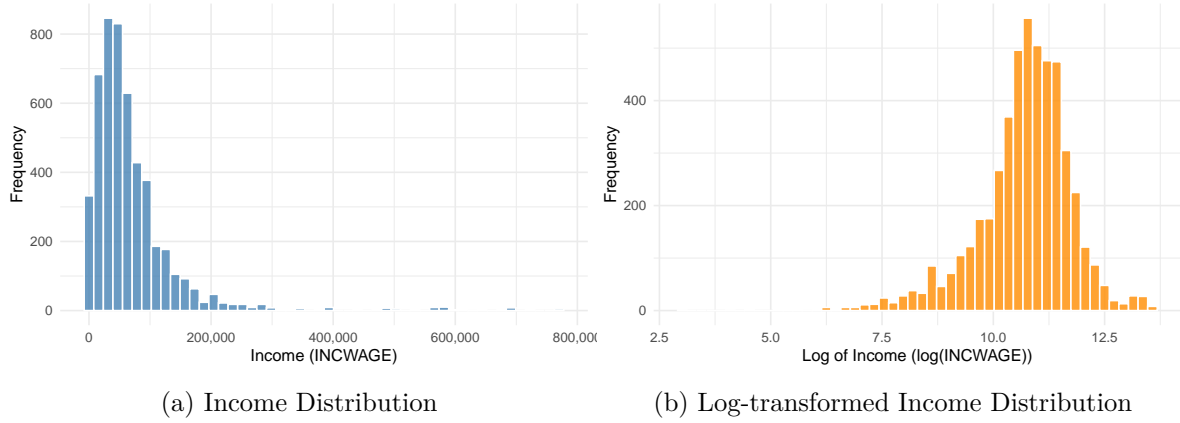(b) Log-transformed Income Distribution

Figure 4: Distribution of Income: Original vs Log-transformed

The model assumptions were checked using Figure 5. First, the Residuals vs Fitted plot tests the linearity assumption, showing that residuals are randomly scattered around the fitted values without any clear pattern, indicating that the linear relationship is valid. Second, the Normal Q-Q plot checks for the normality of residuals. Most points align closely with the diagonal line, suggesting the residuals are approximately normally distributed, with only slight deviations at the tails. Third, the Scale-Location plot evaluates homoskedasticity, and the red line remains mostly flat, suggesting that the variance of the residuals is fairly consistent. Finally, the Residuals vs Leverage plot identifies potential outliers or high-leverage points, indirectly testing the independence of errors and the independence of variables. Most points have low leverage, with only a few high-leverage points that may need attention. Overall, the assumptions are largely met, though there are some limitations that could be addressed with further refinements in the analysis.
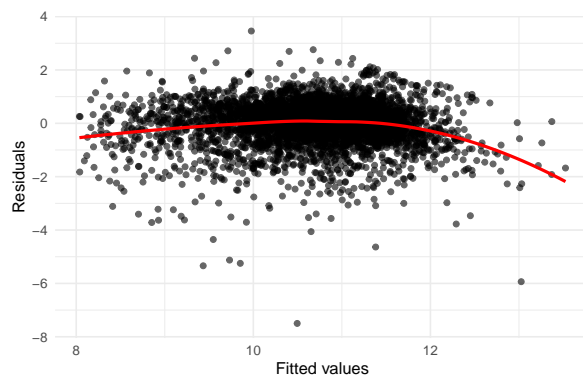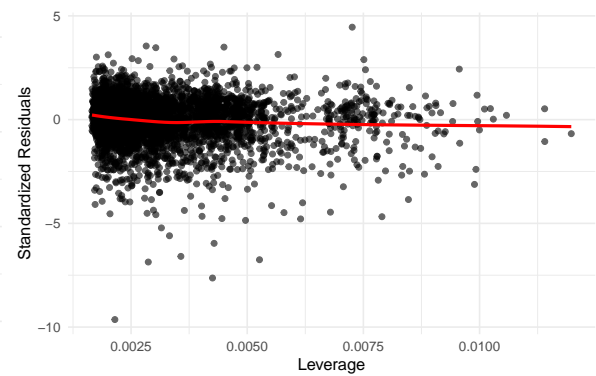
(a) Residuals vs Fitted
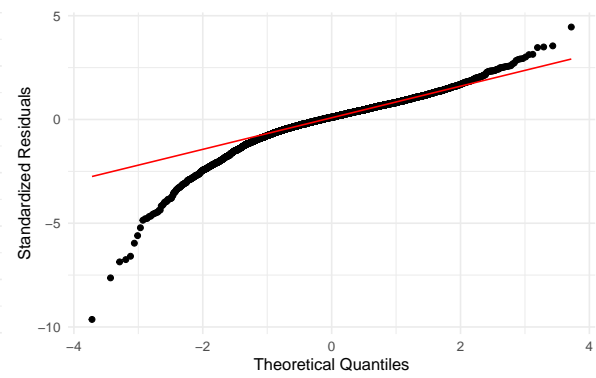
(b) Normal Q-Q Plot
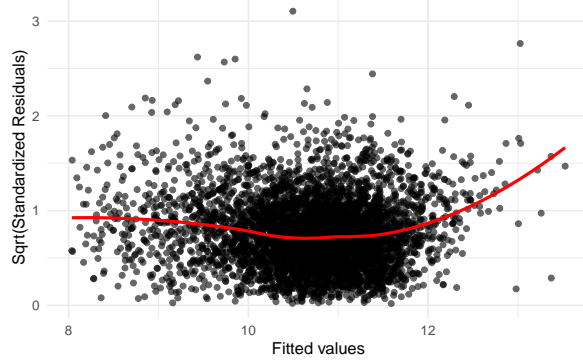
(c) Scale-Location

(d) Residuals vs Leverage

Figure 5: Model Assumption Checks

References

Blau, Francine D, Peter Brummund, and Albert Yung-Hsu Liu. 2013. "Trends in Occupational Segregation by Gender 1970–2009: Adjusting for the Impact of Changes in the Occupational Coding System." *Demography* 50: 471–92.

Blau, Francine D., and Lawrence M. Kahn. 2017. "The Gender Wage Gap: Extent, Trends, and Explanations." *Journal of Economic Literature* 55 (3): 789–865. https://doi.org/10.1257/jel.20160995.

Bureau of Labor Statistics, U. S. Department of Labor. 2023. "National Longitudinal Survey of Youth 1979 Cohort, 1979-2020 (Rounds 1-29)." Columbus, OH: Produced; distributed by the Center for Human Resource Research (CHRR), The Ohio State University.

Flood, Sarah, Miriam King, Renae Rodgers, Steven Ruggles, J. Robert Warren, Daniel Backman, Annie Chen, et al. 2024. "IPUMS CPS: Version 12.0 [Dataset]." Minneapolis, MN: IPUMS. https://doi.org/10.18128/D030.V12.0.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression.* Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Kuhn, and Max. 2008. "Building Predictive Models in r Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. https://doi.org/10.18637/jss.v028.i05.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files.* https://CRAN.R-project.org/package=here.

Queneau, Hervé. 2009. "Trends in Occupational Segregation by Race and Ethnicity in the USA: Evidence from Detailed Data." *Applied Economics Letters* 16 (13): 1347–50. https://doi.org/10.1080/13504850701367346.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'.* https://CRAN.R-project.org/package=arrow.

Robinson, David, Alex Hayes, and Simon Couch. 2024. *Broom: Convert Statistical Objects into Tidy Tibbles.* https://CRAN.R-project.org/package=broom.

Ruggles, Steven, Sarah Flood, Matthew Sobek, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Renae Rodgers, and Megan Schouweiler. 2024. "IPUMS USA: Version 15.0 [dataset]." Minneapolis, MN: IPUMS, University of Minnesota. https://doi.org/10.18128/D010.V15.0.

Wickham, Hadley. 2007. "Reshaping Data with the reshape Package." *Journal of Statistical Software* 21 (12): 1–20. http://www.jstatsoft.org/v21/i12/.

———. 2011. "Testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

———. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan,

Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data.* https://CRAN.R-project.org/package=readr.

Wickham, Hadley, Thomas Lin Pedersen, and Dana Seidel. 2023. *Scales: Scale Functions for Visualization.* https://CRAN.R-project.org/package=scales.

Xie, Yihui. 2024. *Knitr: A General-Purpose Package for Dynamic Report Generation in r.* https://yihui.org/knitr/.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax.* https://CRAN.R-project.org/package=kableExtra.