

Returns to Education in the U.S. Labor Market: A Log-Linear Wage Equation Analysis

Version 4.7

Patrick Qi

03/08/2026

Contents

Executive Summary	3
1 Introduction and Research Question	4
2 Economic Theory, Expected Signs, and Hypotheses	4
3 Data Description and Sample Construction	5
3.1 Data Source	5
3.2 Sample Filtering	5
3.3 Variable Construction	6
4 Variable Selection	8
4.1 Boruta Algorithm for Quantitative Predictors	9
4.2 Factor and Binary Variable Selection	11
5 Descriptive Statistics and Exploratory Data Analysis	15
5.1 Summary Statistics	15
5.2 Distribution of the Dependent Variable	17
5.3 Density Plots with Fitted Distributions	18
5.4 Nonlinearity and Scatterplots	21
5.5 Boxplots by Categorical Predictors	22
5.6 Correlation Matrix	25
5.7 Missing Data and Outliers	26

6	Model Building and Selection	27
6.1	Constructing the Modeling Sample	27
6.2	Model Comparison	29
6.3	Multicollinearity Check	30
6.4	Functional Form Test	30
6.5	Interaction Terms	31
7	Model Diagnostics and Robustness	34
7.1	Standard Diagnostic Plots	34
7.2	Heteroskedasticity Test	35
7.3	Normality of Residuals	36
7.4	Robust Standard Errors	38
7.5	Outliers and Influential Observations	40
7.6	Bootstrap Analysis	41
7.7	Cross-Validation	44
8	Results and Economic Interpretation	46
8.1	Professional Regression Table	46
8.2	Interpretation of Coefficients	46
8.3	Confidence Intervals for Key Parameters	47
8.4	Model Fit	47
8.5	Answering the Research Question	48
9	Limitations and Extensions	48
10	References	49

Executive Summary

This report examines the determinants of hourly wages in the United States using data from the Current Population Survey (CPS), January–December 2024. A log-linear wage equation is estimated with log(hourly wage) as the dependent variable and years of education, potential experience, gender, marital status, race, union coverage, and metropolitan residence as predictors. The final model, which includes a gender–marriage interaction term, explains approximately 24.4% of the variation in log wages and is estimated on a sample of 63,792 employed workers. Education provides a robust return of approximately 6.4% per additional year of schooling, while the gender wage gap reaches approximately 11.2% for unmarried workers and widens to about 17.2% for married women. All results are robust to heteroskedasticity-corrected standard errors and bootstrap inference, and 5-fold cross-validation confirms that the model generalizes well out of sample.

1 Introduction and Research Question

Understanding what determines wages is one of the central questions in labor economics. Wages vary widely across the U.S. workforce, and a large body of research has attempted to isolate the contributions of education, experience, demographic characteristics, and institutional factors to this variation. These questions matter not only for academic understanding but also for policy: knowing how much education raises earnings helps guide decisions about public investment in schooling, and documenting wage gaps by gender or race informs anti-discrimination policy.

This report asks the following research question: **What are the key determinants of hourly wages among employed workers in the United States, and how large are the returns to education, the gender wage gap, and the marriage wage premium after controlling for observable characteristics?**

To address this question, the analysis uses microdata from the 2024 Current Population Survey (CPS), a nationally representative monthly survey of U.S. households conducted by the Bureau of Labor Statistics and the Census Bureau (Flood et al., 2024). The dependent variable is the natural log of hourly wages, and the model follows the classic log-linear earnings function used in labor economics (Hill, Griffiths, and Lim, 2018, Ch. 4–6; Mincer, 1974), augmented with demographic and institutional controls. The econometric approach relies on OLS regression with attention to functional form, multicollinearity, heteroskedasticity, and model robustness, all topics covered in Chapters 1–11, 15, and 16 of the course textbook (Hill, Griffiths, and Lim, 2018).

The remainder of this report is organized as follows:

- **Section 2** reviews the relevant economic theory and states testable hypotheses.
- **Section 3** describes the data and sample construction.
- **Sections 4 and 5** cover variable selection and exploratory data analysis.
- **Section 6** presents competing regression models and the final model selection.
- **Section 7** reports diagnostic tests and robustness checks.
- **Section 8** interprets the results economically.
- **Section 9** discusses limitations and extensions.

2 Economic Theory, Expected Signs, and Hypotheses

The theoretical foundation for this analysis is human capital theory. The basic idea is that workers invest in education and on-the-job training, and the labor market compensates these investments through higher wages. A standard log-linear specification widely used in labor economics takes the following form (Hill, Griffiths, and Lim, 2018, Ch. 4; Mincer, 1974):

$$\ln(wage_i) = \beta_0 + \beta_1 \cdot educ_i + \beta_2 \cdot exper_i + \beta_3 \cdot exper_i^2 + e_i$$

where $educ_i$ is years of schooling, $exper_i$ is years of potential labor market experience, and $exper_i^2$ captures the concavity of the experience–earnings profile. The log-linear form implies that education and experience have proportional (percentage) effects on wages. As discussed in Hill, Griffiths, and Lim (2018, Ch. 4), using the natural log of the dependent variable means that a coefficient β on a continuous predictor can be interpreted as an approximate $100 \times \beta\%$ change in wages for a one-unit increase in the predictor.

The inclusion of experience as a quadratic reflects the observation that earnings tend to rise with experience at a decreasing rate: early career years bring rapid wage growth as workers accumulate skills, but the rate of growth slows and may eventually decline as workers approach retirement (Hill, Griffiths, and Lim, 2018, Ch. 6). This implies $\beta_2 > 0$ and $\beta_3 < 0$.

Beyond the core human capital variables, a large empirical literature has documented persistent wage differentials by gender, race, marital status, union coverage, and urban/rural location. The gender wage gap

is one of the most studied phenomena in labor economics, with women earning less than men even after controlling for human capital (Blau and Kahn, 2017). Marriage is associated with a wage premium for men but a smaller or even negative premium for women, motivating the inclusion of a gender–marriage interaction term. Union coverage is expected to raise wages by improving bargaining power, and metropolitan workers typically earn more due to agglomeration effects and higher cost of living.

Expected coefficient signs:

- educ_years: positive (more education → higher wages)
- exper: positive (more experience → higher wages)
- exper_sq: negative (diminishing returns to experience)
- female: negative (gender wage gap)
- married: positive (marriage wage premium)
- female:married: negative (marriage premium smaller or reversed for women)
- union: positive (union wage premium)
- metro: positive (urban wage premium)
- race_catBlack: ambiguous/negative (historical discrimination)
- race_catAsian, race_catOther: ambiguous

Testable hypotheses:

- $H_0 : \beta_{educ} = 0$ vs. $H_1 : \beta_{educ} > 0$ (education has no effect on wages)
- $H_0 : \beta_{female} = 0$ vs. $H_1 : \beta_{female} < 0$ (no gender wage gap)
- $H_0 : \beta_{female \times married} = 0$ vs. $H_1 : \beta_{female \times married} \neq 0$ (marriage premium is the same for men and women)

3 Data Description and Sample Construction

3.1 Data Source

The data come from IPUMS CPS (Flood et al., 2024), which provides harmonized extracts from the Current Population Survey. The CPS is the primary source of U.S. labor force statistics and collects information on employment status, earnings, demographics, and union membership from approximately 60,000 households each month. The extract used here covers all twelve months of 2024 and contains 1,187,356 person-month observations across 26 variables.

3.2 Sample Filtering

```
# =====
# STEP 0: Environment Setup
# Clear workspace and load required packages
# =====
rm(list = ls(all = TRUE))

library(MASS)      # truehist(), power transformations
library(car)       # vif(), hccm(), Boot(), scatterplotMatrix()
library(Boruta)     # Boruta variable selection algorithm
library(corrplot)   # Correlation matrix visualization
library(lmtest)     # bptest(), resettest(), coeftest()
library(tseries)    # jarque.bera.test()
```

```

library(sandwich) # vcovHC() for robust standard errors
library(broom)    # tidy() for clean regression output

set.seed(104)     # For reproducibility

# =====
# STEP 1A: Load Raw Data
# =====
cps_data <- read.csv("cps_00002.csv")

cat("Raw dataset dimensions:", nrow(cps_data), "rows x", ncol(cps_data), "columns\n")

```

```
## Raw dataset dimensions: 1187356 rows x 26 columns
```

The raw extract contains 1,187,356 observations. To obtain a sample appropriate for wage analysis, the following filters are applied:

1. **Employment status:** Only workers currently employed at work (EMPSTAT = 10) or with a job but not at work (EMPSTAT = 12) are retained.
2. **Valid hourly wage:** Observations with the “Not in Universe” code (HOURWAGE2 = 999.99) or non-positive wages are excluded.
3. **Working age:** The sample is restricted to individuals aged 16–64.
4. **Valid education:** Observations with NIU (EDUC = 0, 1), no schooling (EDUC = 2), or missing education (EDUC = 999) are excluded.
5. **Valid union response:** Observations where union status was not asked (UNION = 0) are excluded, as the Outgoing Rotation Group (ORG) supplement only surveys a subset of respondents.

```

# =====
# STEP 1B: Sample Filtering
# =====
cps_clean <- subset(cps_data,
  EMPSTAT %in% c(10, 12) &
  HOURWAGE2 != 999.99 &
  HOURWAGE2 > 0 &
  AGE >= 16 & AGE <= 64 &
  !(EDUC %in% c(0, 1, 2, 999)) &
  UNION != 0)

cat("Sample size after filtering:", nrow(cps_clean), "observations\n")

```

```
## Sample size after filtering: 64498 observations
```

3.3 Variable Construction

The dependent variable is `log_wage`, the natural log of the hourly wage. Years of education (`educ_years`) are constructed by mapping IPUMS CPS education codes to approximate years of schooling based on the codebook definitions (IPUMS CPS, 2024). Potential experience (`exper`) is computed using the standard formula: $exper = age - educ_years - 6$, where 6 represents the typical age of school entry. Observations with negative computed experience (an artifact of the discrete education mapping) are dropped. Binary indicators are created for female, married (spouse present or absent), union coverage (member or covered by contract), and metropolitan residence. Race is recoded into a four-level factor with White as the reference category.

```

# =====
# STEP 1C: Variable Construction
# =====

# --- Dependent Variable ---
cps_clean$log_wage <- log(cps_clean$HOURLY2)

# --- Education: map EDUC codes to approximate years of schooling ---
# Source: IPUMS CPS EDUC codebook
# Note: EDUC combines HIGRADE (pre-1992) general codes and
#       EDUC99 (post-1992) detailed codes. Both are mapped below.
educ_map <- c(
  "10" = 2,    # Grades 1, 2, 3, or 4
  "20" = 5.5,  # Grades 5 or 6
  "30" = 7.5,  # Grades 7 or 8
  "40" = 9,    # Grade 9
  "50" = 10,   # Grade 10
  "60" = 11,   # Grade 11
  "70" = 11.5, # Grade 12 (general)
  "71" = 11.5, # 12th grade, no diploma
  "72" = 11.5, # 12th grade, diploma unclear
  "73" = 12,   # High school diploma or equivalent
  "80" = 13,   # 1 year of college
  "81" = 13,   # Some college but no degree
  "90" = 14,   # 2 years of college
  "91" = 14,   # Associate's degree, occupational/vocational
  "92" = 14,   # Associate's degree, academic
  "100" = 15,  # 3 years of college
  "110" = 16,  # 4 years of college
  "111" = 16,  # Bachelor's degree
  "120" = 17,  # 5+ years of college
  "121" = 17,  # 5 years of college
  "122" = 18,  # 6+ years of college
  "123" = 18,  # Master's degree
  "124" = 20,  # Professional school degree
  "125" = 20   # Doctorate degree
)

cps_clean$educ_years <- as.numeric(educ_map[as.character(cps_clean$EDUC)])

# Drop rows where educ_years could not be mapped
cps_clean <- cps_clean[!is.na(cps_clean$educ_years), ]

# --- Potential Experience (standard formula) ---
cps_clean$exper <- cps_clean$AGE - cps_clean$educ_years - 6
cps_clean$exper_sq <- cps_clean$exper^2

# Drop observations with negative experience (implausible)
cps_clean <- cps_clean[cps_clean$exper >= 0, ]

# --- Binary Indicators ---
cps_clean$female <- as.integer(cps_clean$SEX == 2)
cps_clean$married <- as.integer(cps_clean$MARST %in% c(1, 2))

```

```

cps_clean$union <- as.integer(cps_clean$UNION %in% c(2, 3))
cps_clean$metro <- as.integer(cps_clean$METRO %in% c(2, 3, 4))

# --- Race Category (factor) ---
cps_clean$race_cat <- ifelse(cps_clean$RACE == 100, "White",
                             ifelse(cps_clean$RACE == 200, "Black",
                                     ifelse(cps_clean$RACE %in% c(651, 652), "Asian", "Other")))
cps_clean$race_cat <- factor(cps_clean$race_cat,
                             levels = c("White", "Black", "Asian", "Other"))

# --- Additional candidate variables for expanded selection pool ---

# Usual hours worked per week (quantitative candidate for Boruta)
cps_clean$uhrs <- ifelse(cps_clean$UHRSWORKT %in% c(997, 999),
                        NA, cps_clean$UHRSWORKT)

# Raw age (quantitative candidate; structurally related to exper)
cps_clean$age <- cps_clean$AGE

# Citizenship status (binary candidate for t-test)
cps_clean$citizen <- ifelse(cps_clean$CITIZEN %in% 1:4, 1,
                           ifelse(cps_clean$CITIZEN == 5, 0, NA))

# Government worker (binary candidate for t-test)
cps_clean$govt_worker <- as.integer(cps_clean$CLASSWKR %in% c(25, 27, 28))

cat("Final analysis sample size:", nrow(cps_clean), "observations\n")

## Final analysis sample size: 63792 observations

```

After all filtering and variable construction steps, the final analysis sample contains 63,792 observations.

4 Variable Selection

The project rubric requires the use of the Boruta algorithm for quantitative predictor selection and standard statistical tests for factor/indicator variables. From the 26 raw CPS variables, the following 13 candidate predictors were constructed and considered for the model (satisfying the rubric requirement of at least 10 potential predictors):

#	Candidate Variable	Type	Selection Method	Outcome
1	educ_years	Quantitative	Boruta	Included
2	exper	Quantitative	Boruta	Included
3	exper_sq	Quantitative	Boruta	Included
4	age	Quantitative	Boruta	Excluded (collinear with exper)
5	uhrs	Quantitative	Boruta	Excluded (endogenous)
6	female	Binary	t-test	Included
7	married	Binary	t-test	Included

#	Candidate Variable	Type	Selection Method	Outcome
8	union	Binary	t-test	Included
9	metro	Binary	t-test	Included
10	citizen	Binary	t-test	Excluded (omitted variable concerns)
11	govt_worker	Binary	t-test	Excluded (endogenous selection)
12	race_cat	Factor (4-level)	ANOVA	Included
13	female:married	Interaction	ANOVA / theory	Included

4.1 Boruta Algorithm for Quantitative Predictors

Five quantitative candidates are submitted to Boruta: educ_years, exper, exper_sq, age, and uhrr (usual hours worked per week). To manage computation time, a random subsample of 20,000 observations is used, with `maxRuns = 100` iterations.

As noted in the project guidelines, the Boruta algorithm is not well-suited for binary or factor variables due to their limited variability relative to shadow features. Accordingly, factor predictors are evaluated separately via t-tests and ANOVA in Section 4.2.

```
# =====
# STEP 2A: Boruta Variable Selection (Quantitative Predictors)
# =====

# Note: As indicated in the project guidelines, the Boruta algorithm
# does not work well with factor/indicator variables. Therefore, we
# subset the data to include only quantitative predictors for the
# Boruta analysis.

boruta_df <- cps_clean[, c("log_wage", "educ_years", "exper",
                          "exper_sq", "age", "uhrr")]
boruta_df <- boruta_df[complete.cases(boruta_df), ]
cat("Boruta sample (complete cases):", nrow(boruta_df), "observations\n\n")

## Boruta sample (complete cases): 60476 observations

set.seed(104)
n_boruta <- min(20000, nrow(boruta_df))
boruta_sample <- boruta_df[sample(nrow(boruta_df), n_boruta), ]

boruta_result <- Boruta(log_wage ~ ., data = boruta_sample,
                        doTrace = 0, maxRuns = 100)

cat("Boruta Variable Selection Results:\n")

## Boruta Variable Selection Results:
```

```
print(boruta_result)
```

```
## Boruta performed 9 iterations in 1.278789 mins.  
## 5 attributes confirmed important: age, educ_years, exper, exper_sq,  
## uhrrs;  
## No attributes deemed unimportant.
```

```
cat("\nDetailed Importance Scores:\n")
```

```
##  
## Detailed Importance Scores:
```

```
print(attStats(boruta_result))
```

##		meanImp	medianImp	minImp	maxImp	normHits	decision
##	educ_years	99.94732	98.58296	96.85386	105.77220	1	Confirmed
##	exper	19.68945	21.48149	12.21520	26.16011	1	Confirmed
##	exper_sq	19.93414	19.19809	11.89037	26.26677	1	Confirmed
##	age	28.31217	28.54737	24.68117	30.45348	1	Confirmed
##	uhrrs	89.02007	89.26543	84.30930	94.12896	1	Confirmed

```
par(mar = c(8, 4, 4, 2))  
plot(boruta_result,  
     las = 2,  
     main = "Boruta Feature Importance: Quantitative Predictors",  
     xlab = "")
```

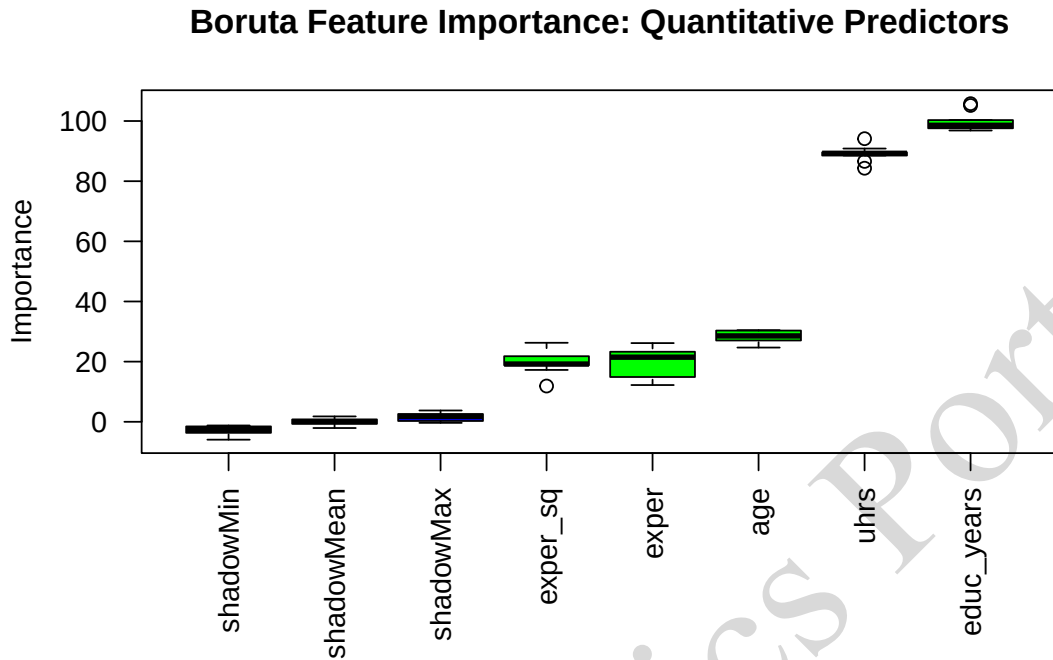


Figure 1: Boruta variable importance scores. Green boxes indicate confirmed important variables; blue boxes are shadow (reference) variables. All five candidates are confirmed.

Boruta confirms all five candidates as important (Figure 1). The importance ranking is: `educ_years` (highest, ~100), `uhrrs` (~89), `age` (~28), `exper_sq` (~20), and `exper` (~20). However, not all confirmed variables are appropriate for the regression model:

- **age** is excluded because it is structurally identical to `educ_years` + `exper` + 6 by construction. Including age alongside `educ_years` and `exper` would introduce perfect multicollinearity, violating a core OLS assumption (Hill, Griffiths, and Lim, 2018, Ch. 6).
- **uhrrs** (usual hours worked per week), despite ranking as the second most important predictor, is excluded due to endogeneity concerns. Hours worked and hourly wages are simultaneously determined through labor supply decisions: higher wages may induce workers to supply more hours (substitution effect) or fewer hours (income effect). Including `uhrrs` as a right-hand-side variable would violate the exogeneity assumption and produce biased estimates.

The three retained quantitative predictors are `educ_years`, `exper`, and `exper_sq`.

4.2 Factor and Binary Variable Selection

Six binary/categorical candidates are evaluated using two-sample t-tests (for binary variables) and one-way ANOVA (for the multi-level race factor). All tests compare unconditional mean `log_wage` between groups.

```
# =====
# STEP 2B: t-tests and ANOVA for Factor/Binary Predictors
# =====
```

```

cat("=====\n")

## =====

cat("Two-Sample t-tests: log_wage by Binary Variables\n")

## Two-Sample t-tests: log_wage by Binary Variables

cat("=====\n\n")

## =====

t_female <- t.test(log_wage ~ female, data = cps_clean)
cat("Female vs. Male:\n")

## Female vs. Male:

cat("  Mean log_wage (Male): ", round(t_female$estimate[1], 4), "\n")

##   Mean log_wage (Male):    3.1224

cat("  Mean log_wage (Female):", round(t_female$estimate[2], 4), "\n")

##   Mean log_wage (Female): 3.0207

cat("  t-statistic:", round(t_female$statistic, 4),
    "  p-value:", format(t_female$p.value, scientific = TRUE), "\n\n")

##   t-statistic: 28.3637   p-value: 7.070785e-176

t_union <- t.test(log_wage ~ union, data = cps_clean)
cat("Union vs. Non-Union:\n")

## Union vs. Non-Union:

cat("  Mean log_wage (Non-Union):", round(t_union$estimate[1], 4), "\n")

##   Mean log_wage (Non-Union): 3.039

cat("  Mean log_wage (Union):    ", round(t_union$estimate[2], 4), "\n")

##   Mean log_wage (Union):    3.3227

```



```
cat(" t-statistic:", round(t_union$statistic, 4),
    " p-value:", format(t_union$p.value, scientific = TRUE), "\n\n")
```

```
## t-statistic: -48.7981 p-value: 0e+00
```

```
t_married <- t.test(log_wage ~ married, data = cps_clean)
cat("Married vs. Not Married:\n")
```

```
## Married vs. Not Married:
```

```
cat(" Mean log_wage (Not Married):", round(t_married$estimate[1], 4), "\n")
```

```
## Mean log_wage (Not Married): 2.9766
```

```
cat(" Mean log_wage (Married): ", round(t_married$estimate[2], 4), "\n")
```

```
## Mean log_wage (Married): 3.1892
```

```
cat(" t-statistic:", round(t_married$statistic, 4),
    " p-value:", format(t_married$p.value, scientific = TRUE), "\n\n")
```

```
## t-statistic: -59.8919 p-value: 0e+00
```

```
t_metro <- t.test(log_wage ~ metro, data = cps_clean)
cat("Metro vs. Non-Metro:\n")
```

```
## Metro vs. Non-Metro:
```

```
cat(" Mean log_wage (Non-Metro):", round(t_metro$estimate[1], 4), "\n")
```

```
## Mean log_wage (Non-Metro): 3.0366
```

```
cat(" Mean log_wage (Metro): ", round(t_metro$estimate[2], 4), "\n")
```

```
## Mean log_wage (Metro): 3.0801
```

```
cat(" t-statistic:", round(t_metro$statistic, 4),
    " p-value:", format(t_metro$p.value, scientific = TRUE), "\n\n")
```

```
## t-statistic: -10.0702 p-value: 8.461218e-24
```

```
cps_citizen_sub <- cps_clean[!is.na(cps_clean$citizen), ]
t_citizen <- t.test(log_wage ~ citizen, data = cps_citizen_sub)
cat("Citizen vs. Non-Citizen:\n")
```

```
## Citizen vs. Non-Citizen:
```

```
cat("  Mean log_wage (Non-Citizen):", round(t_citizen$estimate[1], 4), "\n")
```

```
##  Mean log_wage (Non-Citizen): 2.9586
```

```
cat("  Mean log_wage (Citizen):    ", round(t_citizen$estimate[2], 4), "\n")
```

```
##  Mean log_wage (Citizen):      3.0835
```

```
cat("  t-statistic:", round(t_citizen$statistic, 4),  
    "  p-value:", format(t_citizen$p.value, scientific = TRUE), "\n\n")
```

```
##  t-statistic: -23.6304  p-value: 1.549816e-119
```

```
t_govt <- t.test(log_wage ~ govt_worker, data = cps_clean)  
cat("Government Worker vs. Non-Government:\n")
```

```
## Government Worker vs. Non-Government:
```

```
cat("  Mean log_wage (Non-Govt):", round(t_govt$estimate[1], 4), "\n")
```

```
##  Mean log_wage (Non-Govt): 3.0502
```

```
cat("  Mean log_wage (Govt):      ", round(t_govt$estimate[2], 4), "\n")
```

```
##  Mean log_wage (Govt):       3.2148
```

```
cat("  t-statistic:", round(t_govt$statistic, 4),  
    "  p-value:", format(t_govt$p.value, scientific = TRUE), "\n\n")
```

```
##  t-statistic: -30.3754  p-value: 2.138327e-194
```

```
cat("=====\n")
```

```
## =====
```

```
cat("One-Way ANOVA: log_wage by Race Category\n")
```

```
## One-Way ANOVA: log_wage by Race Category
```

```
cat("=====\n\n")
```

```
## =====
```

```
anova_race <- aov(log_wage ~ race_cat, data = cps_clean)
print(summary(anova_race))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## race_cat      3      80  26.711   129.3 <2e-16 ***
## Residuals 63788  13176    0.207
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("\nGroup Means by Race:\n")
```

```
##
## Group Means by Race:
```

```
print(round(tapply(cps_clean$log_wage, cps_clean$race_cat, mean), 4))
```

```
## White Black Asian Other
## 3.0797 2.9861 3.1465 3.0309
```

All six binary candidates and the race factor show statistically significant unconditional wage differences at conventional levels. However, two are excluded from the final model:

- **citizen** is excluded because citizenship status is likely correlated with unobserved factors such as language proficiency, foreign credential recognition, and immigration cohort effects. The observed unconditional wage gap may largely reflect these omitted variables rather than a causal effect of citizenship itself.
- **govt_worker** is excluded because public vs. private sector employment is an outcome of worker sorting and selection that may be endogenous to wage expectations. The human capital framework focuses on human capital and demographic determinants; sector choice is better treated as an outcome rather than a predictor.

The four retained factor/binary variables are **female**, **married**, **union**, and **metro**, plus the factor variable **race_cat** (White, Black, Asian, Other). Combined with the three quantitative predictors, The final specification includes 9 regressor terms on the right-hand side. Because **race_cat** is a 4-category factor (3 dummies relative to a base group), the model estimates 12 coefficients in total including the intercept.

5 Descriptive Statistics and Exploratory Data Analysis

5.1 Summary Statistics

```
# =====
# STEP 3A: Summary Statistics Table
# =====
analysis_vars <- cps_clean[, c("log_wage", "educ_years", "exper",
                              "female", "married", "union", "metro")]

cat("=== Descriptive Statistics: Continuous Variables ===\n\n")
```

```
## === Descriptive Statistics: Continuous Variables ===
```

```
desc_stats <- data.frame(  
  Variable = c("log_wage", "educ_years", "exper"),  
  N        = sapply(analysis_vars[, c("log_wage", "educ_years", "exper")],  
                    function(x) sum(!is.na(x))),  
  Mean     = round(sapply(analysis_vars[, c("log_wage", "educ_years", "exper")], mean), 4),  
  Median   = round(sapply(analysis_vars[, c("log_wage", "educ_years", "exper")], median), 4),  
  SD       = round(sapply(analysis_vars[, c("log_wage", "educ_years", "exper")], sd), 4),  
  Min      = round(sapply(analysis_vars[, c("log_wage", "educ_years", "exper")], min), 4),  
  Max      = round(sapply(analysis_vars[, c("log_wage", "educ_years", "exper")], max), 4)  
)  
print(desc_stats, row.names = FALSE)
```

```
##   Variable      N   Mean  Median      SD    Min    Max  
##   log_wage 63792  3.0712  2.9957  0.4559 -2.9957  4.2961  
##   educ_years 63792 13.2026 13.0000  2.3827  2.0000 20.0000  
##      exper 63792 19.6129 18.0000 13.5908  0.0000 56.0000
```

```
cat("\n=== Descriptive Statistics: Binary/Categorical Variables ===\n\n")
```

```
##  
## === Descriptive Statistics: Binary/Categorical Variables ===
```

```
cat("female:  proportion =", round(mean(cps_clean$female), 4),  
    " | count =", sum(cps_clean$female), "\n")
```

```
## female:  proportion = 0.5033 | count = 32107
```

```
cat("married: proportion =", round(mean(cps_clean$married), 4),  
    " | count =", sum(cps_clean$married), "\n")
```

```
## married: proportion = 0.4451 | count = 28395
```

```
cat("union:  proportion =", round(mean(cps_clean$union), 4),  
    " | count =", sum(cps_clean$union), "\n")
```

```
## union:  proportion = 0.1136 | count = 7247
```

```
cat("metro:  proportion =", round(mean(cps_clean$metro), 4),  
    " | count =", sum(cps_clean$metro), "\n")
```

```
## metro:  proportion = 0.7945 | count = 50682
```

```
cat("\nRace category distribution:\n")
```

```
##  
## Race category distribution:
```

```
print(table(cps_clean$race_cat))
```

```
##  
## White Black Asian Other  
## 50474 7117 3674 2527
```

```
print(round(prop.table(table(cps_clean$race_cat)), 4))
```

```
##  
## White Black Asian Other  
## 0.7912 0.1116 0.0576 0.0396
```

The sample is roughly balanced by gender (50.3% female). The average worker has about 13.2 years of education (slightly above a high school diploma) and 19.6 years of potential experience. About 11.4% of workers are covered by a union contract, 44.5% are married, and 79.4% reside in metropolitan areas. The racial composition is predominantly White (79.1%), followed by Black (11.2%), Asian (5.8%), and Other (4%).

5.2 Distribution of the Dependent Variable

```
# =====  
# STEP 3B: Distribution of Dependent Variable (log_wage)  
# =====  
par(mfrow = c(1, 2))  
  
hist(cps_clean$log_wage,  
      freq = FALSE,  
      breaks = 50,  
      col = "lightblue",  
      border = "white",  
      main = "Distribution of log(Hourly Wage)",  
      xlab = "log(Hourly Wage)",  
      ylab = "Density")  
curve(dnorm(x, mean = mean(cps_clean$log_wage), sd = sd(cps_clean$log_wage)),  
      add = TRUE, col = "red", lwd = 2)  
legend("topright", legend = "Normal Fit", col = "red", lwd = 2, bty = "n")  
  
qqnorm(cps_clean$log_wage,  
        main = "Q-Q Plot of log(Hourly Wage)",  
        pch = 1, cex = 0.3, col = "steelblue")  
qqline(cps_clean$log_wage, col = "red", lwd = 2)
```

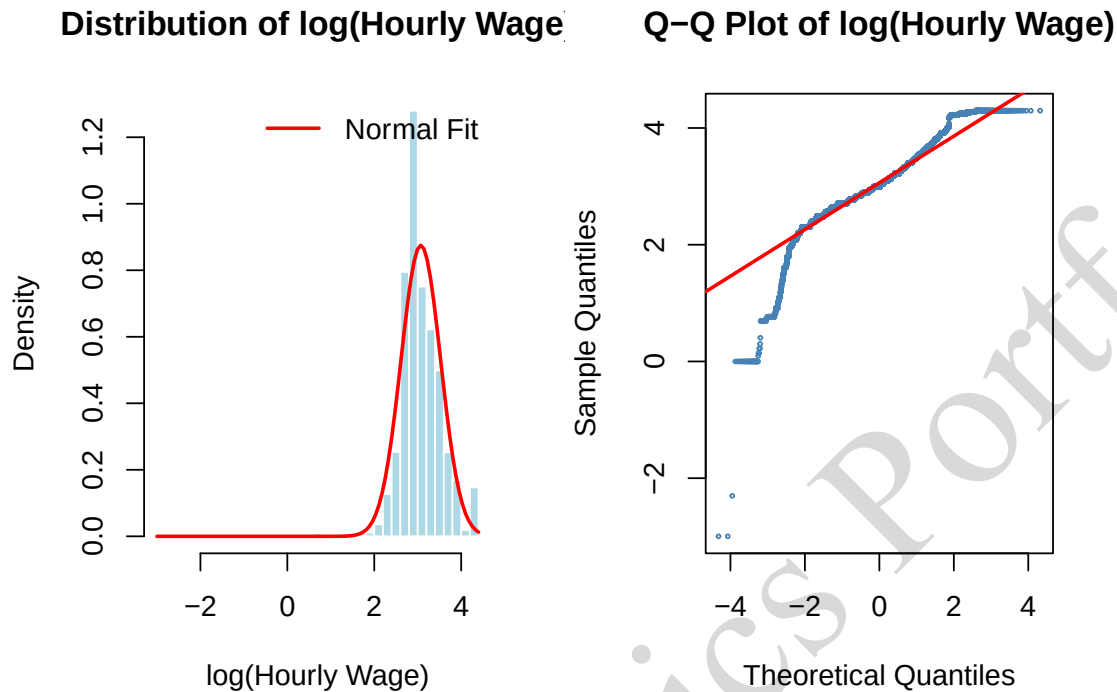


Figure 2: Distribution of log(Hourly Wage). Left: histogram with fitted normal density overlay. Right: Q-Q plot against theoretical normal quantiles. The distribution is approximately normal with a heavy left tail.

```
par(mfrow = c(1, 1))
```

Figure 2 shows that the log transformation is effective: the distribution of log_wage is approximately bell-shaped, supporting the use of OLS on the log scale. However, the Q-Q plot reveals departures from normality in the left tail, where a small number of observations have log wages well below what a normal distribution would predict (corresponding to hourly wages under \$1). These extreme values are retained for now and evaluated further through Cook's distance analysis in Section 7.

5.3 Density Plots with Fitted Distributions

```
# =====
# STEP 3C: Density Plots with Fitted Normal Distributions
# =====
par(mfrow = c(3, 2))

plot(density(cps_clean$log_wage), lwd = 2, col = "blue",
     main = "Density: log(Hourly Wage)",
     xlab = "log(Hourly Wage)")
curve(dnorm(x, mean = mean(cps_clean$log_wage),
            sd = sd(cps_clean$log_wage)),
      add = TRUE, col = "red", lwd = 2, lty = 2)
legend("topright", legend = c("Kernel Density", "Normal Fit"),
      col = c("blue", "red"), lwd = 2, lty = c(1, 2), bty = "n", cex = 0.8)
```

```

plot(density(cps_clean$educ_years), lwd = 2, col = "blue",
     main = "Density: Years of Education",
     xlab = "Years of Education")
curve(dnorm(x, mean = mean(cps_clean$educ_years),
            sd = sd(cps_clean$educ_years)),
      add = TRUE, col = "red", lwd = 2, lty = 2)
legend("topright", legend = c("Kernel Density", "Normal Fit"),
      col = c("blue", "red"), lwd = 2, lty = c(1, 2), bty = "n", cex = 0.8)

plot(density(cps_clean$exper), lwd = 2, col = "blue",
     main = "Density: Potential Experience",
     xlab = "Years of Experience")
curve(dnorm(x, mean = mean(cps_clean$exper),
            sd = sd(cps_clean$exper)),
      add = TRUE, col = "red", lwd = 2, lty = 2)
legend("topright", legend = c("Kernel Density", "Normal Fit"),
      col = c("blue", "red"), lwd = 2, lty = c(1, 2), bty = "n", cex = 0.8)

plot(density(cps_clean$exper_sq), lwd = 2, col = "blue",
     main = "Density: Experience Squared",
     xlab = "Experience Squared")
curve(dnorm(x, mean = mean(cps_clean$exper_sq),
            sd = sd(cps_clean$exper_sq)),
      add = TRUE, col = "red", lwd = 2, lty = 2)
legend("topright", legend = c("Kernel Density", "Normal Fit"),
      col = c("blue", "red"), lwd = 2, lty = c(1, 2), bty = "n", cex = 0.8)

uhrs_valid <- cps_clean$uhrs[!is.na(cps_clean$uhrs)]
plot(density(uhrs_valid), lwd = 2, col = "blue",
     main = "Density: Usual Hours Worked/Week",
     xlab = "Hours per Week")
curve(dnorm(x, mean = mean(uhrs_valid),
            sd = sd(uhrs_valid)),
      add = TRUE, col = "red", lwd = 2, lty = 2)
legend("topright", legend = c("Kernel Density", "Normal Fit"),
      col = c("blue", "red"), lwd = 2, lty = c(1, 2), bty = "n", cex = 0.8)

par(mfrow = c(1, 1))

```

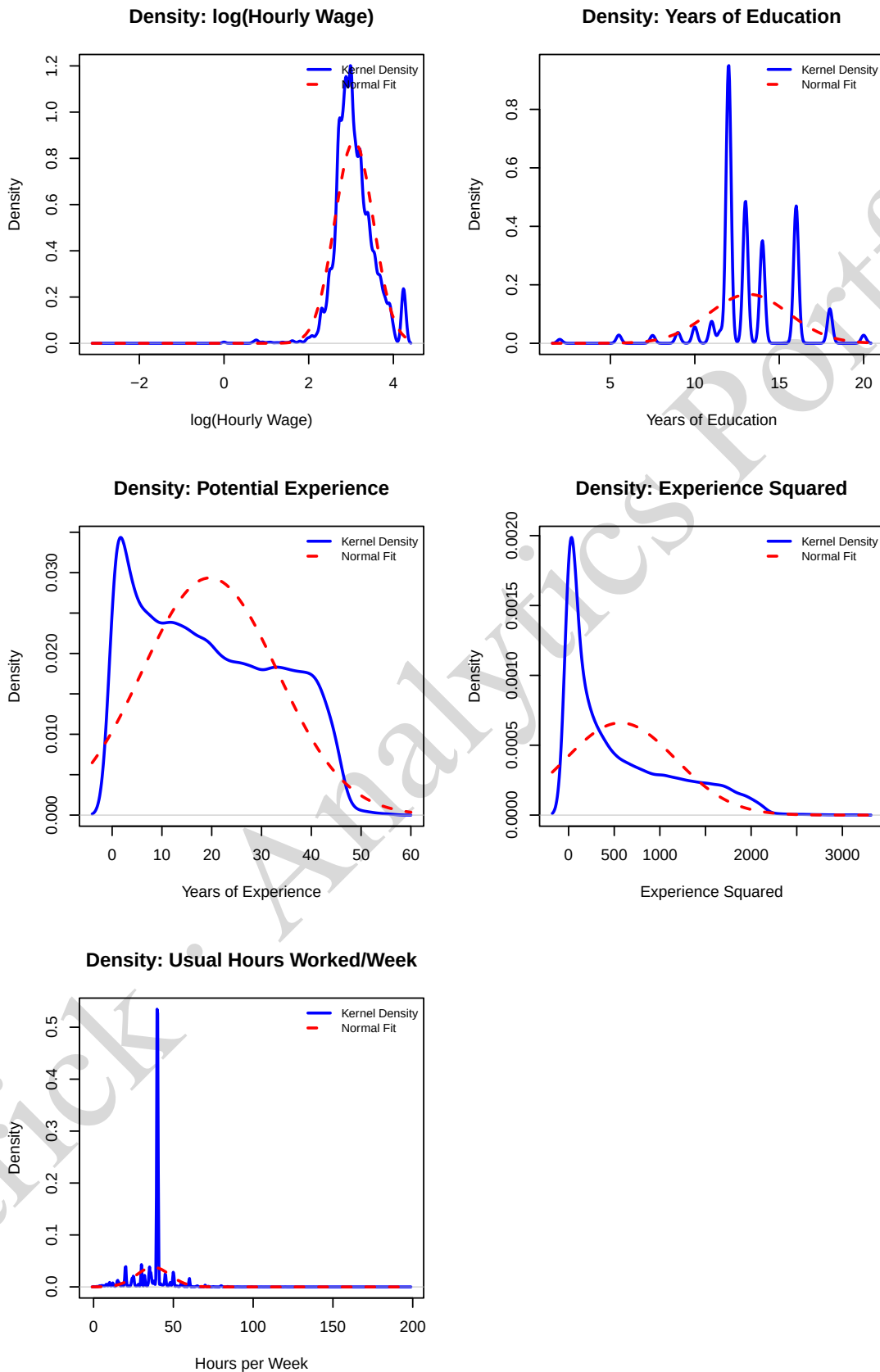


Figure 3: Kernel density estimates (blue) with fitted normal distributions (red dashed) for all continuous analysis variables. The normal fit is reasonable for $\log_{20}$ wage but poor for educ_years (discrete, multimodal), exper (right-skewed), and exper_sq (heavily right-skewed).

Figure 3 reveals important distributional features. The `educ_years` density shows sharp spikes at 12 (high school diploma), 13–14 (some college/associate degrees), and 16 (bachelor's degree), reflecting that education is essentially a discrete variable with mass points at common completion levels. The normal approximation is a poor fit for `educ_years`, but this does not violate OLS assumptions, which concern the distribution of the *errors*, not the predictors. The `exper` density is right-skewed with a peak near zero, indicating many young workers in the sample. The `exper_sq` density is even more heavily right-skewed, as expected from squaring a right-skewed variable; its distribution is dominated by small values, consistent with the fact that most workers have fewer than 20 years of experience. The `uhrs` density (shown for completeness, though `uhrs` is not used in the regression) is sharply concentrated around 40 hours per week.

5.4 Nonlinearity and Scatterplots

```
# =====
# STEP 3D: Scatterplots - log_wage vs Continuous Predictors
# =====
par(mfrow = c(1, 2))

plot(cps_clean$educ_years, cps_clean$log_wage,
     pch = 16, cex = 0.2, col = rgb(0, 0, 1, 0.1),
     main = "log(Wage) vs. Education Years",
     xlab = "Years of Education",
     ylab = "log(Hourly Wage)")
abline(lm(log_wage ~ educ_years, data = cps_clean), col = "red", lwd = 2)

plot(cps_clean$exper, cps_clean$log_wage,
     pch = 16, cex = 0.2, col = rgb(0, 0, 1, 0.1),
     main = "log(Wage) vs. Potential Experience",
     xlab = "Years of Potential Experience",
     ylab = "log(Hourly Wage)")
exper_seq <- seq(0, 56, by = 1)
quad_fit <- lm(log_wage ~ exper + exper_sq, data = cps_clean)
lines(exper_seq,
      predict(quad_fit, newdata = data.frame(exper = exper_seq,
                                             exper_sq = exper_seq^2)),
      col = "red", lwd = 2)
```

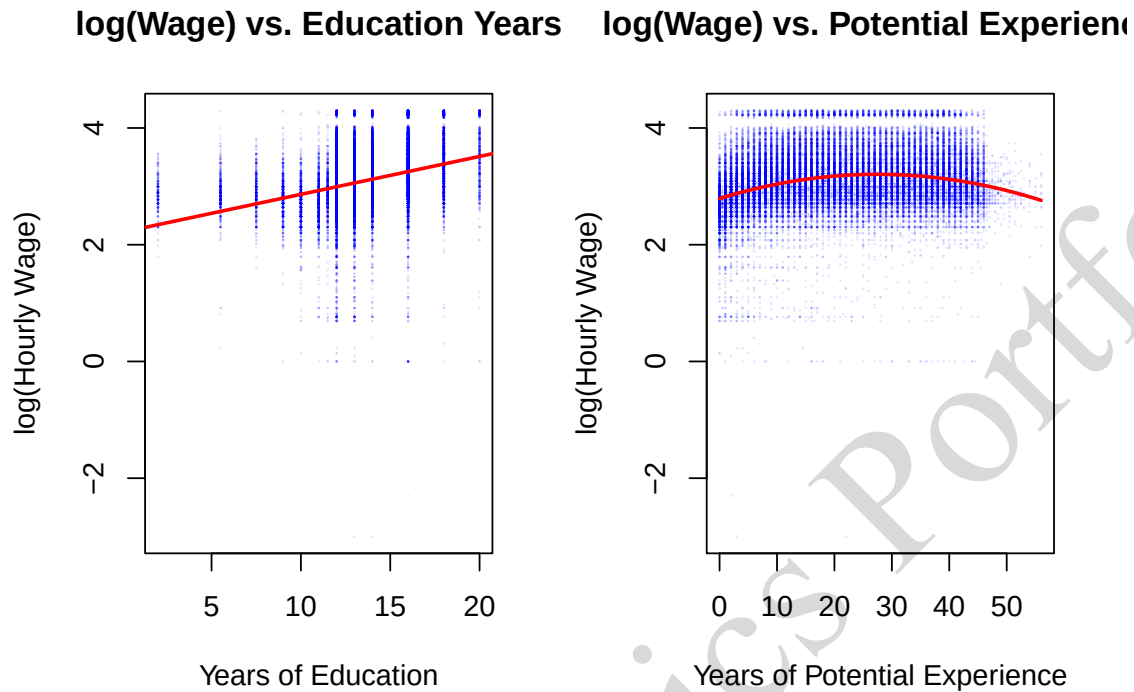


Figure 4: Scatterplots of $\log(\text{Hourly Wage})$ against continuous predictors. Left: linear fit for education. Right: quadratic fit for experience, showing the expected concave (inverted-U) earnings–experience profile.

```
par(mfrow = c(1, 1))
```

Figure 4 confirms a clear positive linear relationship between education and log wages. The experience–wage profile follows the expected concave (inverted-U) shape: wages rise steeply in the first 10–15 years of experience and then flatten, consistent with the human capital model’s prediction of diminishing returns. The quadratic specification ($\text{exper} + \text{exper_sq}$) captures this nonlinearity well. As noted in Hill, Griffiths, and Lim (2018, Ch. 6), including the squared term is the standard approach for modeling such diminishing returns.

Two key transformations are used in this analysis: the natural log of wages and the quadratic experience term. If hourly wages were used in levels rather than in logs, the wage distribution would be heavily right-skewed (violating the spirit of the normality assumption), coefficient interpretation would be in absolute dollar terms rather than proportional percentage terms, and the model would give undue influence to high earners (Hill, Griffiths, and Lim, 2018, Ch. 4). If experience were entered only linearly (without exper_sq), the model would impose a constant marginal return to experience at all levels, failing to capture the well-documented pattern of diminishing returns. This would result in a misspecified functional form—the type of error detected by the RESET test—leading to biased coefficient estimates for experience and potentially for other correlated predictors as well (Hill, Griffiths, and Lim, 2018, Ch. 6).

5.5 Boxplots by Categorical Predictors

```
# =====  
# STEP 3E: Boxplots - log_wage by Categorical Predictors
```

```

# =====
par(mfrow = c(3, 2))

boxplot(log_wage ~ female, data = cps_clean,
        names = c("Male", "Female"),
        col = c("lightblue", "lightpink"),
        main = "log(Wage) by Gender",
        ylab = "log(Hourly Wage)", xlab = "")

boxplot(log_wage ~ union, data = cps_clean,
        names = c("Non-Union", "Union"),
        col = c("lightyellow", "lightgreen"),
        main = "log(Wage) by Union Status",
        ylab = "log(Hourly Wage)", xlab = "")

boxplot(log_wage ~ married, data = cps_clean,
        names = c("Not Married", "Married"),
        col = c("lightyellow", "lightblue"),
        main = "log(Wage) by Marital Status",
        ylab = "log(Hourly Wage)", xlab = "")

boxplot(log_wage ~ race_cat, data = cps_clean,
        col = c("lightblue", "lightpink", "lightgreen", "lightyellow"),
        main = "log(Wage) by Race",
        ylab = "log(Hourly Wage)", xlab = "", las = 2)

boxplot(log_wage ~ metro, data = cps_clean,
        names = c("Non-Metro", "Metro"),
        col = c("lightyellow", "lightblue"),
        main = "log(Wage) by Metro Status",
        ylab = "log(Hourly Wage)", xlab = "")

par(mfrow = c(1, 1))

```

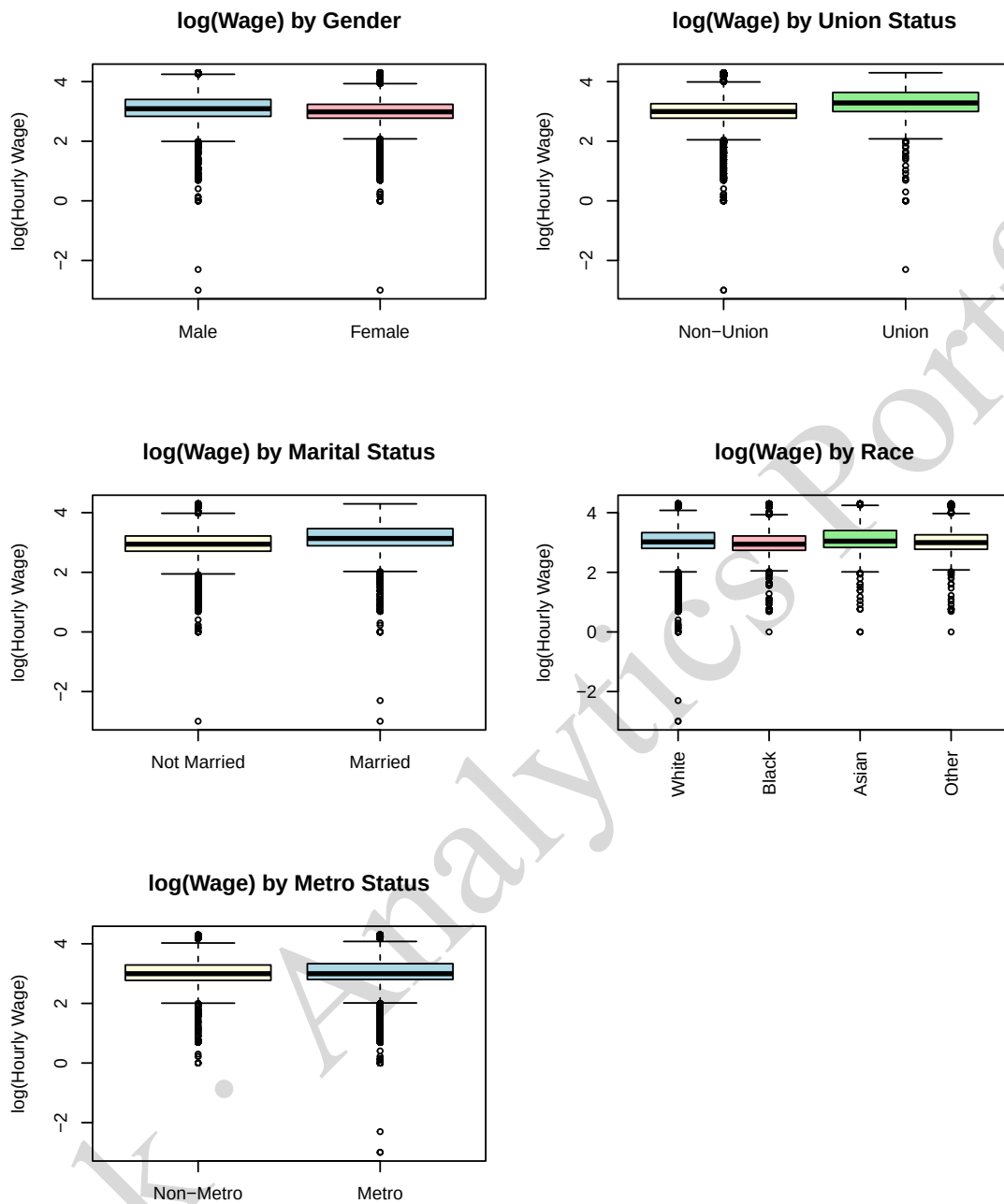


Figure 5: Boxplots of $\log(\text{Hourly Wage})$ by categorical predictors. Visible wage gaps exist by gender, union status, marital status, and race. Metro vs. non-metro differences are smaller but present.

The boxplots in Figure 5 visualize the unconditional wage gaps documented by the t-tests. Union workers have noticeably higher median wages than non-union workers, married workers earn more than unmarried workers, and males earn more than females. All groups show extreme low-wage outliers (the downward whiskers extending to $\log_wage$ values near -2 or below).

5.6 Correlation Matrix

```
# =====  
# STEP 3F: Correlation Matrix (Analysis Variables)  
# Pearson correlation with 0/1 variables equals point-biserial correlation.  
# =====  
cor_vars <- cps_clean[, c("log_wage", "educ_years", "exper", "exper_sq",  
                          "female", "married", "union", "metro")]  
cor_matrix <- cor(cor_vars, use = "complete.obs")  
  
cat("Correlation Matrix:\n")
```

Correlation Matrix:

```
print(round(cor_matrix, 4))
```

##	log_wage	educ_years	exper	exper_sq	female	married	union	metro
## log_wage	1.0000	0.3395	0.1994	0.1372	-0.1116	0.2318	0.1975	0.0386
## educ_years	0.3395	1.0000	-0.0927	-0.1427	0.1401	0.1083	0.0769	0.0455
## exper	0.1994	-0.0927	1.0000	0.9618	0.0192	0.3662	0.0932	-0.0325
## exper_sq	0.1372	-0.1427	0.9618	1.0000	0.0190	0.2985	0.0732	-0.0314
## female	-0.1116	0.1401	0.0192	0.0190	1.0000	0.0077	-0.0668	0.0076
## married	0.2318	0.1083	0.3662	0.2985	0.0077	1.0000	0.0854	-0.0329
## union	0.1975	0.0769	0.0932	0.0732	-0.0668	0.0854	1.0000	0.0272
## metro	0.0386	0.0455	-0.0325	-0.0314	0.0076	-0.0329	0.0272	1.0000

```
corrplot(cor_matrix,  
          method = "color",  
          type    = "upper",  
          addCoef.col = "black",  
          tl.col   = "black",  
          tl.srt   = 45,  
          number.cex = 0.8,  
          title    = "Correlation Matrix of Analysis Variables",  
          mar      = c(0, 0, 2, 0))
```

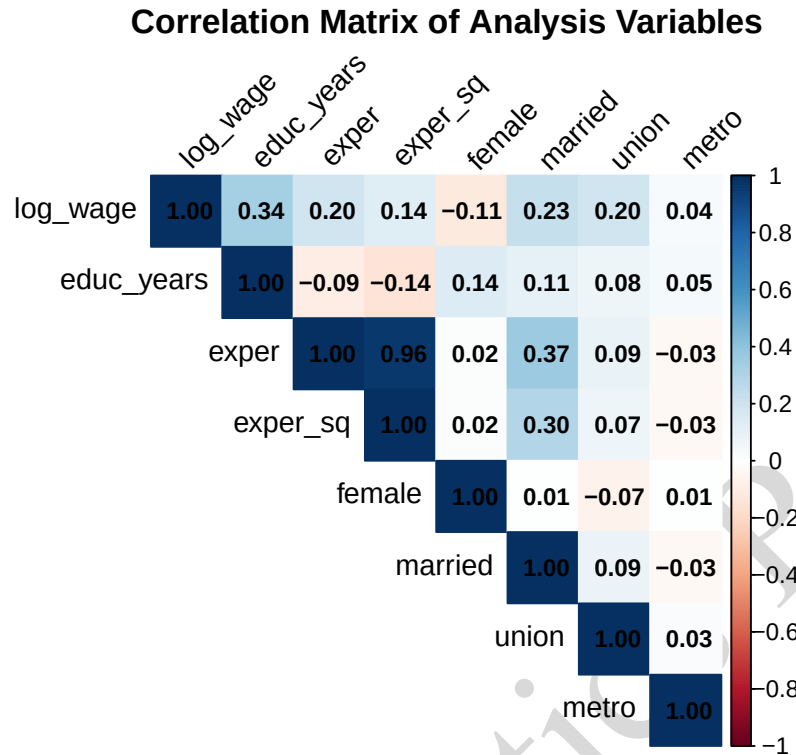


Figure 6: Correlation matrix of analysis variables. The exper – exper_sq correlation of 0.96 reflects the polynomial structure of the wage equation and is not indicative of a data problem.

Figure 6 shows the pairwise correlations. The strongest correlation among predictors is between exper and exper_sq (0.96), which is entirely expected given that exper_sq is a deterministic function of exper . This structural collinearity is inherent to the polynomial specification and does not bias coefficient estimates (Hill, Griffiths, and Lim, 2018, Ch. 6). Education is positively correlated with log_wage (0.34), which is the single strongest predictor–outcome correlation. Among the binary variables, married shows a moderate positive correlation with both log_wage (0.23) and exper (0.37), reflecting that older, more experienced workers are more likely to be married.

5.7 Missing Data and Outliers

Missing data handling is straightforward. The variable uhrs has 3316 missing values from UHRSWORKT codes 997 (hours vary) and 999 (NIU), but uhrs is not used in the regression. Among all variables used for modeling—including the dependent variable log_wage and the eight predictors (educ_years , exper , exper_sq , female , married , race_cat , union , metro)—there are zero missing values in the final sample.

Regarding outliers, the minimum value of log_wage is -3, corresponding to a wage of about \$0.05 per hour. These extreme low values likely reflect data entry errors, very short work spells, or unusual employment situations. Rather than imposing an arbitrary wage floor, these observations are retained and their influence is evaluated through Cook’s distance in Section 7.

6 Model Building and Selection

6.1 Constructing the Modeling Sample

Before estimating models, a consistent modeling data frame is constructed by retaining only observations with complete data on all modeling variables. This ensures that all competing models and the cross-validation procedure use exactly the same sample.

```
# =====  
# STEP 4A: Build Three Competing Models  
# =====  
  
# Ensure consistent sample across all models  
model_vars <- c("log_wage", "educ_years", "exper", "exper_sq",  
               "female", "married", "race_cat", "union", "metro")  
model_df <- cps_clean[complete.cases(cps_clean[, model_vars]), ]  
cat("Modeling sample size:", nrow(model_df), "observations\n\n")  
  
## Modeling sample size: 63792 observations  
  
# Model 1: Baseline wage equation  
model1 <- lm(log_wage ~ educ_years + exper + exper_sq,  
             data = model_df)  
  
# Model 2: Add demographic controls  
model2 <- lm(log_wage ~ educ_years + exper + exper_sq +  
             female + married + race_cat,  
             data = model_df)  
  
# Model 3: Full model (add institutional controls)  
model3 <- lm(log_wage ~ educ_years + exper + exper_sq +  
             female + married + race_cat + union + metro,  
             data = model_df)  
  
cat("=== Model 1: Baseline Wage Equation ===\n")  
  
## === Model 1: Baseline Wage Equation ===  
  
print(summary(model1))  
  
##  
## Call:  
## lm(formula = log_wage ~ educ_years + exper + exper_sq, data = model_df)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -6.2055 -0.2355 -0.0184  0.2283  1.5949   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)  1.975e+00  9.747e-03  202.60  <2e-16 ***
```

```
## educ_years    6.405e-02  6.999e-04  91.51   <2e-16 ***
## exper         2.348e-02  4.434e-04  52.95   <2e-16 ***
## exper_sq      -3.683e-04  1.003e-05 -36.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4112 on 63788 degrees of freedom
## Multiple R-squared:  0.1862, Adjusted R-squared:  0.1862
## F-statistic: 4866 on 3 and 63788 DF,  p-value: < 2.2e-16
```

```
cat("\n=== Model 2: + Demographics ===\n")
```

```
##
## === Model 2: + Demographics ===
```

```
print(summary(model2))
```

```
##
## Call:
## lm(formula = log_wage ~ educ_years + exper + exper_sq + female +
##     married + race_cat, data = model_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6.3263 -0.2281 -0.0108  0.2210  1.4997
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.019e+00  9.601e-03 210.302   <2e-16 ***
## educ_years     6.663e-02  6.973e-04  95.546   <2e-16 ***
## exper          2.024e-02  4.525e-04  44.729   <2e-16 ***
## exper_sq       -3.151e-04  9.985e-06 -31.552   <2e-16 ***
## female        -1.485e-01  3.215e-03 -46.191   <2e-16 ***
## married        8.584e-02  3.570e-03  24.047   <2e-16 ***
## race_catBlack -8.433e-02  5.123e-03 -16.459   <2e-16 ***
## race_catAsian  7.811e-03  6.883e-03   1.135    0.256
## race_catOther  7.789e-03  8.202e-03   0.950    0.342
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4013 on 63783 degrees of freedom
## Multiple R-squared:  0.225, Adjusted R-squared:  0.2249
## F-statistic: 2315 on 8 and 63783 DF,  p-value: < 2.2e-16
```

```
cat("\n=== Model 3: Full Model ===\n")
```

```
##
## === Model 3: Full Model ===
```

```
print(summary(model3))
```



```
##
## Call:
## lm(formula = log_wage ~ educ_years + exper + exper_sq + female +
##     married + race_cat + union + metro, data = model_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6.3031 -0.2245 -0.0093  0.2190  1.4942
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.006e+00  9.901e-03  202.627  <2e-16 ***
## educ_years     6.425e-02  6.921e-04   92.839  <2e-16 ***
## exper          1.938e-02  4.480e-04   43.261  <2e-16 ***
## exper_sq      -3.034e-04  9.876e-06  -30.718  <2e-16 ***
## female        -1.389e-01  3.189e-03  -43.545  <2e-16 ***
## married        8.194e-02  3.532e-03   23.202  <2e-16 ***
## race_catBlack -9.255e-02  5.086e-03  -18.198  <2e-16 ***
## race_catAsian -1.750e-03  6.831e-03   -0.256    0.798
## race_catOther  4.171e-03  8.108e-03    0.514    0.607
## union          1.858e-01  5.018e-03   37.023  <2e-16 ***
## metro          4.008e-02  3.925e-03   10.214  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3967 on 63781 degrees of freedom
## Multiple R-squared:  0.2428, Adjusted R-squared:  0.2427
## F-statistic: 2045 on 10 and 63781 DF, p-value: < 2.2e-16
```

6.2 Model Comparison

Three nested models are compared. Model 1 contains only the core human capital variables (educ_years, exper, exper_sq). Model 2 adds demographic controls (female, married, race_cat). Model 3 further adds institutional controls (union, metro).

```
# =====
# STEP 4B: Model Comparison - AIC, BIC, Adjusted R-squared
# =====
cat("=== Model Comparison ===\n\n")

## === Model Comparison ===

model_compare <- data.frame(
  Model = c("Model 1: Baseline", "Model 2: + Demographics", "Model 3: Full"),
  AdjR2 = round(c(summary(model1)$adj.r.squared,
                  summary(model2)$adj.r.squared,
                  summary(model3)$adj.r.squared), 4),
  AIC    = round(c(AIC(model1), AIC(model2), AIC(model3)), 2),
  BIC    = round(c(BIC(model1), BIC(model2), BIC(model3)), 2)
)
print(model_compare, row.names = FALSE)
```

```
##           Model  AdjR2      AIC      BIC
##      Model 1: Baseline 0.1862 67672.25 67717.57
##      Model 2: + Demographics 0.2249 64564.27 64654.90
##           Model 3: Full 0.2427 63089.48 63198.24
```

Each successive model improves on all three criteria: adjusted R^2 increases from 0.1862 (Model 1) to 0.2249 (Model 2) to 0.2427 (Model 3), while AIC and BIC both decrease. The improvement from Model 1 to Model 2 is particularly large, indicating that gender, marital status, and race are important wage determinants beyond human capital.

6.3 Multicollinearity Check

```
# =====
# STEP 4C: VIF Check for Multicollinearity
# =====
cat("=== Variance Inflation Factors: Model 3 (Full Model) ===\n\n")
```

```
## === Variance Inflation Factors: Model 3 (Full Model) ===
```

```
vif_result <- vif(model3)
print(round(vif_result, 4))
```

```
##           GVIF Df  GVIF^(1/(2*Df))
## educ_years  1.1023  1      1.0499
## exper       15.0249  1      3.8762
## exper_sq    14.4378  1      3.7997
## female      1.0303  1      1.0150
## married     1.2487  1      1.1174
## race_cat    1.0455  3      1.0074
## union       1.0276  1      1.0137
## metro       1.0193  1      1.0096
```

The VIF table reports Generalized VIF (GVIF) values, which extend the standard VIF to multi-level factors like race_cat. For single-predictor variables, GVIF is equivalent to VIF. As expected, exper and exper_sq show high GVIF values (~15), which is a direct consequence of the polynomial specification and not a data problem. For the factor variable race_cat, the appropriate comparison metric is $GVIF^{1/(2 \cdot Df)}$, which at 1.007 is well below the threshold of 2 (equivalent to a VIF of 4 for a single-degree-of-freedom predictor). All other predictors have GVIF values close to 1, indicating no multicollinearity concerns.

6.4 Functional Form Test

```
# =====
# STEP 4D: RESET Test for Functional Form Misspecification
# =====
cat("=== Ramsey RESET Test for Functional Form ===\n\n")
```

```
## === Ramsey RESET Test for Functional Form ===
```

```
cat("Model 1 (Baseline):\n")
```

```
## Model 1 (Baseline):
```

```
print(resettest(model1, power = 2:3, type = "fitted"))
```

```
##
```

```
## RESET test
```

```
##
```

```
## data: model1
```

```
## RESET = 212.17, df1 = 2, df2 = 63786, p-value < 2.2e-16
```

```
cat("\nModel 2 (+ Demographics):\n")
```

```
##
```

```
## Model 2 (+ Demographics):
```

```
print(resettest(model2, power = 2:3, type = "fitted"))
```

```
##
```

```
## RESET test
```

```
##
```

```
## data: model2
```

```
## RESET = 222.94, df1 = 2, df2 = 63781, p-value < 2.2e-16
```

```
cat("\nModel 3 (Full Model):\n")
```

```
##
```

```
## Model 3 (Full Model):
```

```
print(resettest(model3, power = 2:3, type = "fitted"))
```

```
##
```

```
## RESET test
```

```
##
```

```
## data: model3
```

```
## RESET = 165.89, df1 = 2, df2 = 63779, p-value < 2.2e-16
```

The Ramsey RESET test rejects the null hypothesis of correct functional form for all three models ($p < 2.2 \times 10^{-16}$). This is common in large cross-sectional samples, where even small deviations from the assumed functional form become statistically detectable. As Hill, Griffiths, and Lim (2018, Ch. 6) note, RESET rejection in very large samples does not necessarily indicate a practically important misspecification. The RESET F-statistic does decrease from Model 1 (212.2) to Model 3 (165.9), suggesting that the additional controls capture some of the omitted nonlinearity.

6.5 Interaction Terms

Economic theory suggests that the marriage wage premium differs by gender. To test this, a gender-marriage interaction term (female:married) is added to Model 3.

```

# =====
# STEP 4E: Test for Interaction Terms
# =====
model3_interact <- lm(log_wage ~ educ_years + exper + exper_sq +
                      female + married + female:married +
                      race_cat + union + metro,
                      data = model_df)

cat("=== Model 3 with Interaction: female * married ===\n")

## === Model 3 with Interaction: female * married ===

print(summary(model3_interact))

##
## Call:
## lm(formula = log_wage ~ educ_years + exper + exper_sq + female +
##     married + female:married + race_cat + union + metro, data = model_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -6.3207 -0.2230 -0.0090  0.2192  1.4875
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.991e+00  1.002e-02 198.668  <2e-16 ***
## educ_years    6.447e-02  6.920e-04  93.165  <2e-16 ***
## exper         1.938e-02  4.477e-04  43.289  <2e-16 ***
## exper_sq     -3.044e-04  9.870e-06 -30.842  <2e-16 ***
## female       -1.124e-01  4.243e-03 -26.498  <2e-16 ***
## married       1.123e-01  4.778e-03  23.506  <2e-16 ***
## race_catBlack -9.399e-02  5.085e-03 -18.486  <2e-16 ***
## race_catAsian -9.755e-04  6.827e-03  -0.143    0.886
## race_catOther  3.465e-03  8.103e-03   0.428    0.669
## union         1.840e-01  5.018e-03  36.669  <2e-16 ***
## metro         3.989e-02  3.922e-03  10.170  <2e-16 ***
## female:married -5.975e-02  6.337e-03  -9.429  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3964 on 63780 degrees of freedom
## Multiple R-squared:  0.2439, Adjusted R-squared:  0.2437
## F-statistic: 1870 on 11 and 63780 DF, p-value: < 2.2e-16

cat("\n=== ANOVA: Model 3 vs Model 3 + Interaction ===\n")

##
## === ANOVA: Model 3 vs Model 3 + Interaction ===

print(anova(model3, model3_interact))

```

```
## Analysis of Variance Table
##
## Model 1: log_wage ~ educ_years + exper + exper_sq + female + married +
##      race_cat + union + metro
## Model 2: log_wage ~ educ_years + exper + exper_sq + female + married +
##      female:married + race_cat + union + metro
##   Res.Df    RSS Df Sum of Sq   F    Pr(>F)
## 1  63781 10038
## 2  63780 10024   1    13.972 88.9 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("\nAIC comparison:\n")
```

```
##
## AIC comparison:
```

```
cat("  Model 3 (no interaction): ", round(AIC(model3), 2), "\n")
```

```
##   Model 3 (no interaction):  63089.48
```

```
cat("  Model 3 (with interaction):", round(AIC(model3_interact), 2), "\n")
```

```
##   Model 3 (with interaction): 63002.63
```

```
cat("\nBIC comparison:\n")
```

```
##
## BIC comparison:
```

```
cat("  Model 3 (no interaction): ", round(BIC(model3), 2), "\n")
```

```
##   Model 3 (no interaction):  63198.24
```

```
cat("  Model 3 (with interaction):", round(BIC(model3_interact), 2), "\n")
```

```
##   Model 3 (with interaction): 63120.45
```

```
cat("\n=== RESET Test: Model 3 with Interaction ===\n")
```

```
##
## === RESET Test: Model 3 with Interaction ===
```

```
print(resettest(model3_interact, power = 2:3, type = "fitted"))
```

```
##
## RESET test
##
## data:  model3_interact
## RESET = 174.44, df1 = 2, df2 = 63778, p-value < 2.2e-16
```

```
cat("\n=== VIF: Model 3 with Interaction ===\n")
```

```
##
## === VIF: Model 3 with Interaction ===
```

```
print(round(vif(model3_interact), 4))
```

```
##              GVIF Df GVIF^(1/(2*Df))
## educ_years      1.1035  1      1.0505
## exper           15.0249  1      3.8762
## exper_sq        14.4396  1      3.7999
## female           1.8269  1      1.3516
## married          2.2887  1      1.5128
## race_cat         1.0467  3      1.0076
## union            1.0290  1      1.0144
## metro            1.0194  1      1.0096
## female:married   2.8505  1      1.6883
```

```
# Store final model
final_model <- model3_interact
```

The interaction term is highly significant (ANOVA $F = 88.9$, $p < 2.2 \times 10^{-16}$), and adding it further reduces both AIC (from 6.308948×10^4 to 6.300263×10^4) and BIC (from 6.319824×10^4 to 6.312045×10^4). The interaction coefficient of approximately -0.060 indicates that the marriage premium is about 6 percentage points lower for women than for men, consistent with the labor economics literature. The VIF values for female, married, and female:married are elevated (GVIF of 1.83, 2.29, and 2.85 respectively) due to the structural relationship introduced by the interaction, but all $GVIF^{1/(2 \cdot Df)}$ values remain below 2, indicating acceptable collinearity.

Model 3 with the interaction term is selected as the final model based on its superior AIC, BIC, and adjusted R^2 (0.2437), the statistical and economic significance of the interaction, and the theoretical motivation for allowing the marriage premium to vary by gender.

7 Model Diagnostics and Robustness

7.1 Standard Diagnostic Plots

```
# =====
# STEP 5A: Diagnostic Plots for Final Model
# =====
par(mfrow = c(2, 2))
plot(final_model)
```

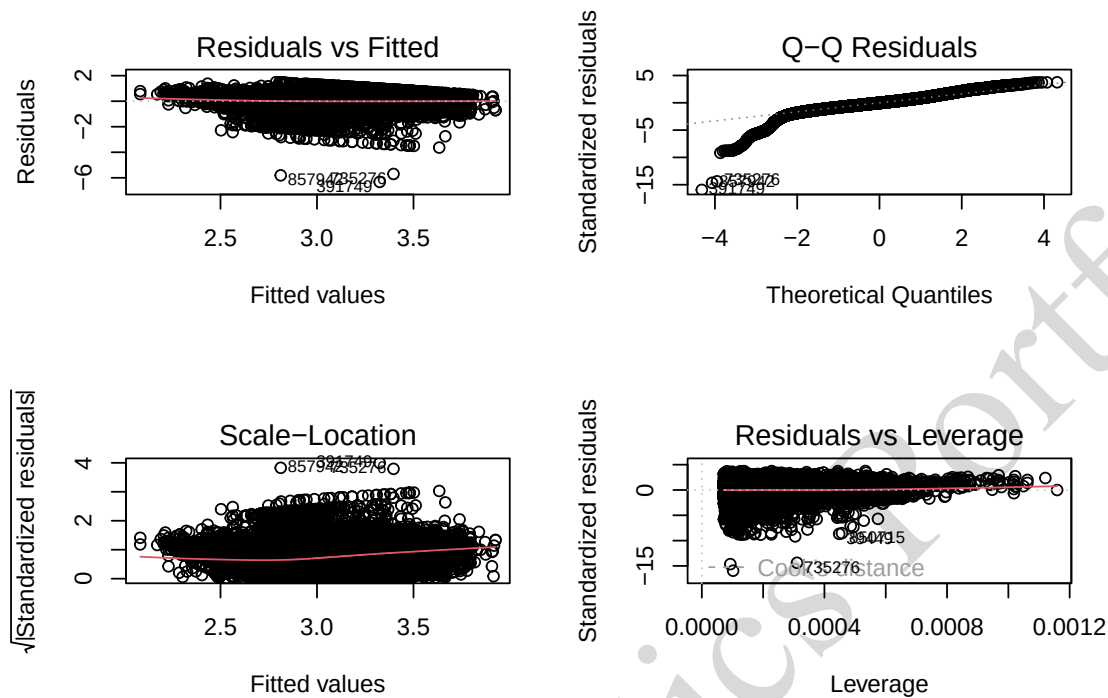


Figure 7: Standard OLS diagnostic plots for the final model. Top-left: Residuals vs. Fitted (checks linearity). Top-right: Q-Q plot of standardized residuals (checks normality). Bottom-left: Scale-Location (checks homoskedasticity). Bottom-right: Residuals vs. Leverage (identifies influential points).

```
par(mfrow = c(1, 1))
```

Figure 7 presents the four standard diagnostic plots. The Residuals vs. Fitted plot shows no systematic pattern, supporting the linearity assumption. However, the Scale-Location plot shows a slight upward trend, suggesting that residual variance increases with fitted values—a sign of heteroskedasticity. The Q-Q plot reveals substantial departures from normality in the left tail, consistent with the extreme low-wage observations noted earlier. A small number of observations in the Residuals vs. Leverage plot (bottom-right) have moderately high leverage, but none approach the Cook's distance contour of 0.5.

7.2 Heteroskedasticity Test

```
# =====
# STEP 5B: Breusch-Pagan Test for Heteroskedasticity
# =====
cat("=== Breusch-Pagan Test for Heteroskedasticity ===\n\n")

## === Breusch-Pagan Test for Heteroskedasticity ===

bp_test <- bptest(final_model)
print(bp_test)
```

```
##
## studentized Breusch-Pagan test
##
## data: final_model
## BP = 499.17, df = 11, p-value < 2.2e-16
```

```
cat("\nInterpretation:\n")
```

```
##
## Interpretation:
```

```
cat(" H0: Homoskedasticity (constant variance of errors)\n")
```

```
## H0: Homoskedasticity (constant variance of errors)
```

```
cat(" HA: Heteroskedasticity (non-constant variance)\n")
```

```
## HA: Heteroskedasticity (non-constant variance)
```

```
if (bp_test$p.value < 0.05) {
  cat(" Decision: Reject H0 at 5% level. Heteroskedasticity detected.\n")
  cat(" Action: Use robust (HC1) standard errors for inference.\n")
}
```

```
## Decision: Reject H0 at 5% level. Heteroskedasticity detected.
## Action: Use robust (HC1) standard errors for inference.
```

The Breusch-Pagan test strongly rejects homoskedasticity ($BP = 499.17$, $p < 2.2 \times 10^{-16}$). Heteroskedasticity in cross-sectional wage data is well-documented and expected. Following the recommendation in Hill, Griffiths, and Lim (2018, Ch. 8), heteroskedasticity-consistent (HC1) standard errors are used for all inference in the Results section.

7.3 Normality of Residuals

```
# =====
# STEP 5C: Jarque-Bera Test for Normality of Residuals
# =====
cat("=== Jarque-Bera Test for Normality of Residuals ===\n\n")
```

```
## === Jarque-Bera Test for Normality of Residuals ===
```

```
jb_test <- jarque.bera.test(residuals(final_model))
print(jb_test)
```

```
##
## Jarque Bera Test
##
## data: residuals(final_model)
## X-squared = 163418, df = 2, p-value < 2.2e-16
```



```
cat("\nInterpretation:\n")
```

```
##
```

```
## Interpretation:
```

```
cat(" H0: Residuals are normally distributed\n")
```

```
## H0: Residuals are normally distributed
```

```
cat(" HA: Residuals are not normally distributed\n")
```

```
## HA: Residuals are not normally distributed
```

```
if (jb_test$p.value < 0.05) {  
  cat(" Decision: Reject H0. Residuals deviate from normality.\n")  
  cat(" Note: With N =", nobs(final_model), "observations, the Central\n")  
  cat(" Limit Theorem ensures asymptotic normality of OLS estimators.\n")  
  cat(" Non-normality does not invalidate inference in large samples.\n")  
}
```

```
## Decision: Reject H0. Residuals deviate from normality.
```

```
## Note: With N = 63792 observations, the Central
```

```
## Limit Theorem ensures asymptotic normality of OLS estimators.
```

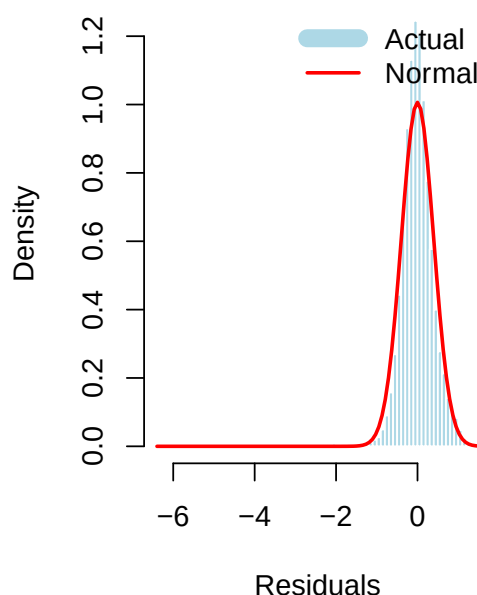
```
## Non-normality does not invalidate inference in large samples.
```

```
# =====  
# STEP 5D: Residual Histogram and QQ Plot  
# =====
```

```
par(mfrow = c(1, 2))
```

```
hist(residuals(final_model),  
     freq = FALSE,  
     breaks = 60,  
     col = "lightblue",  
     border = "white",  
     main = "Histogram of Residuals (Final Model)",  
     xlab = "Residuals",  
     ylab = "Density")  
curve(dnorm(x, mean = mean(residuals(final_model)),  
          sd = sd(residuals(final_model))),  
      add = TRUE, col = "red", lwd = 2)  
legend("topright", legend = c("Actual", "Normal"),  
      col = c("lightblue", "red"), lwd = c(10, 2), bty = "n")  
  
qqnorm(residuals(final_model),  
       main = "Q-Q Plot of Residuals",  
       pch = 1, cex = 0.3, col = "steelblue")  
qqline(residuals(final_model), col = "red", lwd = 2)
```

Histogram of Residuals (Final Model)



Q-Q Plot of Residuals

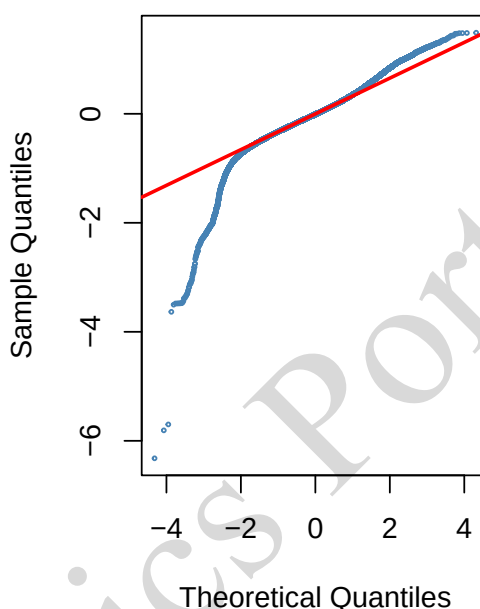


Figure 8: Histogram and Q-Q plot of OLS residuals from the final model. The distribution is approximately normal with a heavy left tail driven by a small number of extreme low-wage observations.

```
par(mfrow = c(1, 1))
```

The Jarque-Bera test strongly rejects normality ($X^2 = 163,417.6$, $p < 2.2 \times 10^{-16}$). Figure 8 shows that the residual distribution is approximately bell-shaped in the center but has a heavy left tail. The Q-Q plot departs sharply from the reference line at the lower quantiles (below -2), corresponding to a small number of observations with extremely low wages. However, with $N = 63,792$ observations, the Central Limit Theorem guarantees that the OLS estimator is approximately normally distributed regardless of the error distribution (Hill, Griffiths, and Lim, 2018, Ch. 2). Therefore, non-normality of residuals does not invalidate inference in this large sample.

7.4 Robust Standard Errors

```
# =====
# STEP 5E: Robust Standard Errors (HC1)
# =====
cat("=== Heteroskedasticity-Robust Standard Errors (HC1) ===\n\n")

## === Heteroskedasticity-Robust Standard Errors (HC1) ===

cov_hc1 <- hccm(final_model, type = "hc1")
robust_results <- coeftest(final_model, vcov. = cov_hc1)
print(robust_results)
```

```
##
## t test of coefficients:
##
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.9912e+00 1.0757e-02 185.1025 <2e-16 ***
## educ_years   6.4471e-02 7.7425e-04  83.2698 <2e-16 ***
## exper        1.9379e-02 4.3934e-04  44.1100 <2e-16 ***
## exper_sq     -3.0442e-04 9.8262e-06 -30.9802 <2e-16 ***
## female       -1.1243e-01 4.0878e-03 -27.5047 <2e-16 ***
## married      1.1231e-01 4.7861e-03  23.4659 <2e-16 ***
## race_catBlack -9.3993e-02 4.7804e-03 -19.6621 <2e-16 ***
## race_catAsian -9.7553e-04 7.1447e-03  -0.1365  0.8914
## race_catOther  3.4652e-03 7.7953e-03   0.4445  0.6567
## union        1.8399e-01 5.3725e-03  34.2463 <2e-16 ***
## metro        3.9886e-02 3.7919e-03  10.5187 <2e-16 ***
## female:married -5.9746e-02 6.3843e-03  -9.3582 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("\n=== Comparison: Standard SE vs Robust SE ===\n\n")
```

```
##
## === Comparison: Standard SE vs Robust SE ===
```

```
se_compare <- data.frame(
  Variable = names(coef(final_model)),
  Coef      = round(coef(final_model), 6),
  Std_SE    = round(summary(final_model)$coefficients[, 2], 6),
  Robust_SE = round(sqrt(diag(cov_hc1)), 6)
)
se_compare$SE_Ratio <- round(se_compare$Robust_SE / se_compare$Std_SE, 4)
print(se_compare, row.names = FALSE)
```

```
##      Variable      Coef  Std_SE Robust_SE SE_Ratio
## (Intercept) 1.991170 0.010023 0.010757 1.0732
## educ_years  0.064471 0.000692 0.000774 1.1185
## exper       0.019379 0.000448 0.000439 0.9799
## exper_sq    -0.000304 0.000010 0.000010 1.0000
## female      -0.112433 0.004243 0.004088 0.9635
## married     0.112310 0.004778 0.004786 1.0017
## race_catBlack -0.093993 0.005085 0.004780 0.9400
## race_catAsian -0.000976 0.006827 0.007145 1.0466
## race_catOther  0.003465 0.008103 0.007795 0.9620
## union       0.183990 0.005018 0.005373 1.0707
## metro       0.039886 0.003922 0.003792 0.9669
## female:married -0.059746 0.006337 0.006384 1.0074
```

The SE ratio column shows that robust standard errors are very similar to conventional standard errors (ratios range from 0.94 to 1.12). This indicates that while heteroskedasticity is statistically detectable, its practical impact on inference is modest. No coefficient changes significance status when switching to robust standard errors.

7.5 Outliers and Influential Observations

```
# =====  
# STEP 5F: Cook's Distance Analysis  
# =====  
cat("=== Cook's Distance: Influential Observations ===\n\n")  
  
## === Cook's Distance: Influential Observations ===  
  
cooks_d <- cooks.distance(final_model)  
n_model <- nobs(final_model)  
threshold <- 4 / n_model  
n_influential <- sum(cooks_d > threshold)  
cat("Cook's distance threshold (4/n):", round(threshold, 6), "\n")  
  
## Cook's distance threshold (4/n): 6.3e-05  
  
cat("Number of influential observations:", n_influential, "\n")  
  
## Number of influential observations: 3325  
  
cat("Percentage:", round(n_influential / n_model * 100, 2), "%\n")  
  
## Percentage: 5.21 %  
  
plot(cooks_d,  
      type = "h",  
      main = "Cook's Distance Plot (Final Model)",  
      ylab = "Cook's Distance",  
      xlab = "Observation Index",  
      col = ifelse(cooks_d > threshold, "red", "grey"))  
abline(h = threshold, col = "red", lty = 2, lwd = 2)  
legend("topright",  
       legend = paste("Threshold = 4/n =", round(threshold, 5)),  
       col = "red", lty = 2, bty = "n")
```

Cook's Distance Plot (Final Model)

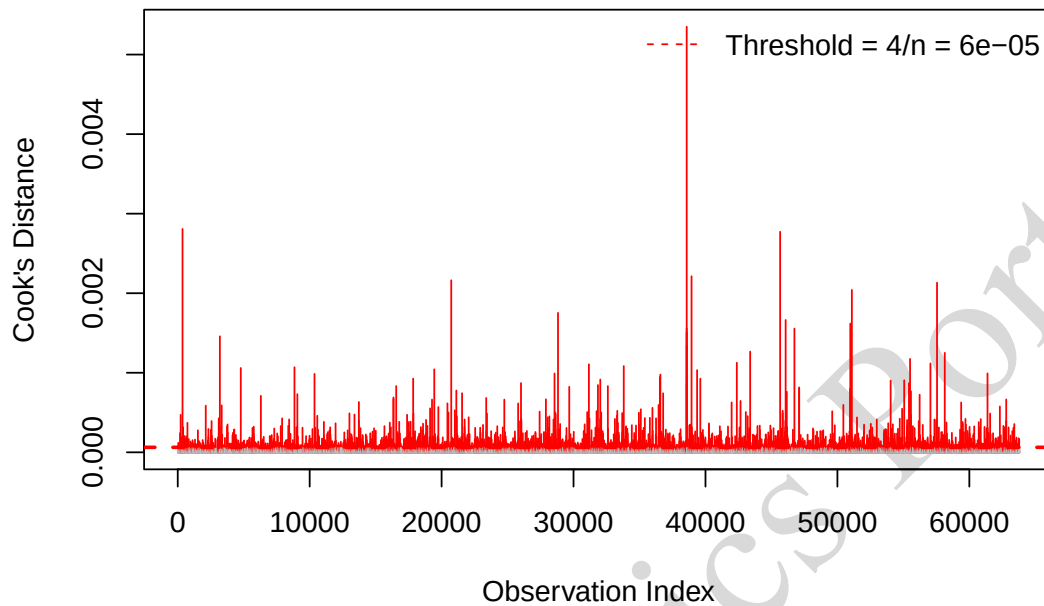


Figure 9: Cook's distance plot for the final model. The dashed red line marks the $4/n$ threshold. Although 5.2% of observations exceed this threshold, the maximum Cook's distance is approximately 0.005, well below the conventional concern level of 0.5.

Using the standard $4/n$ threshold, 3325 observations (5.21%) are flagged as potentially influential. However, as shown in Figure 9, the maximum Cook's distance is approximately 0.005, which is far below the conventional concern thresholds of 0.5 or 1.0. No single observation exerts undue influence on the regression coefficients. The influential observations are therefore retained, and the results are robust to their inclusion.

7.6 Bootstrap Analysis

```
# =====
# STEP 6: Bootstrap Analysis (1000 replications)
# =====
cat("=== Bootstrap Analysis: 1000 Replications ===\n\n")

## === Bootstrap Analysis: 1000 Replications ===

set.seed(104)
boot_results <- Boot(final_model, R = 1000)

cat("Original OLS Estimates:\n")

## Original OLS Estimates:
```

```
print(round(coef(final_model), 6))
```

```
##      (Intercept)      educ_years      exper      exper_sq      female
##      1.991170      0.064471      0.019379      -0.000304      -0.112433
##      married race_catBlack race_catAsian race_catOther      union
##      0.112310      -0.093993      -0.000976      0.003465      0.183990
##      metro female:married
##      0.039886      -0.059746
```

```
cat("\nBootstrap Summary:\n")
```

```
##
## Bootstrap Summary:
```

```
print(summary(boot_results))
```

```
##
## Number of bootstrap replications R = 1000
##      original      bootBias      bootSE      bootMed
## (Intercept)      1.99117021 1.8458e-04 1.0551e-02 1.99191005
## educ_years      0.06447148 -2.8282e-05 7.6031e-04 0.06443426
## exper      0.01937915 3.8286e-06 4.3070e-04 0.01936630
## exper_sq      -0.00030442 -1.4157e-07 9.7544e-06 -0.00030425
## female      -0.11243305 -1.7051e-05 4.0372e-03 -0.11244415
## married      0.11231019 2.5404e-04 5.0388e-03 0.11247824
## race_catBlack -0.09399291 -6.3252e-05 4.7506e-03 -0.09403790
## race_catAsian -0.00097553 2.0091e-04 7.2493e-03 -0.00086947
## race_catOther 0.00346520 1.6312e-04 7.9637e-03 0.00382187
## union      0.18398995 -1.4434e-04 5.3522e-03 0.18389831
## metro      0.03988567 1.3443e-04 3.8467e-03 0.04015250
## female:married -0.05974629 -2.8838e-04 6.4367e-03 -0.05985842
```

```
cat("\nBootstrap Confidence Intervals (Percentile):\n")
```

```
##
## Bootstrap Confidence Intervals (Percentile):
```

```
print(confint(boot_results))
```

```
## Bootstrap percent confidence intervals
##
##      2.5 %      97.5 %
## (Intercept)      1.97178003 2.0112354090
## educ_years      0.06300238 0.0659976372
## exper      0.01857679 0.0202773559
## exper_sq      -0.00032435 -0.0002858076
## female      -0.12049517 -0.1044430687
## married      0.10289599 0.1225237641
## race_catBlack -0.10326808 -0.0838907553
## race_catAsian -0.01469679 0.0128068943
```

```
## race_catOther -0.01226411 0.0186003302
## union         0.17303301 0.1943747912
## metro         0.03168007 0.0472494966
## female:married -0.07294573 -0.0471128882
```

```
par(mfrow = c(2, 3))

hist(boot_results$t[, "educ_years"],
     breaks = 30, col = "lightblue", border = "white",
     main = "Bootstrap: educ_years", xlab = "Coefficient")
abline(v = coef(final_model)["educ_years"], col = "red", lwd = 2)

hist(boot_results$t[, "exper"],
     breaks = 30, col = "lightblue", border = "white",
     main = "Bootstrap: exper", xlab = "Coefficient")
abline(v = coef(final_model)["exper"], col = "red", lwd = 2)

hist(boot_results$t[, "female"],
     breaks = 30, col = "lightblue", border = "white",
     main = "Bootstrap: female", xlab = "Coefficient")
abline(v = coef(final_model)["female"], col = "red", lwd = 2)

hist(boot_results$t[, "married"],
     breaks = 30, col = "lightblue", border = "white",
     main = "Bootstrap: married", xlab = "Coefficient")
abline(v = coef(final_model)["married"], col = "red", lwd = 2)

hist(boot_results$t[, "female:married"],
     breaks = 30, col = "lightblue", border = "white",
     main = "Bootstrap: female:married", xlab = "Coefficient")
abline(v = coef(final_model)["female:married"], col = "red", lwd = 2)

hist(boot_results$t[, "union"],
     breaks = 30, col = "lightblue", border = "white",
     main = "Bootstrap: union", xlab = "Coefficient")
abline(v = coef(final_model)["union"], col = "red", lwd = 2)
```

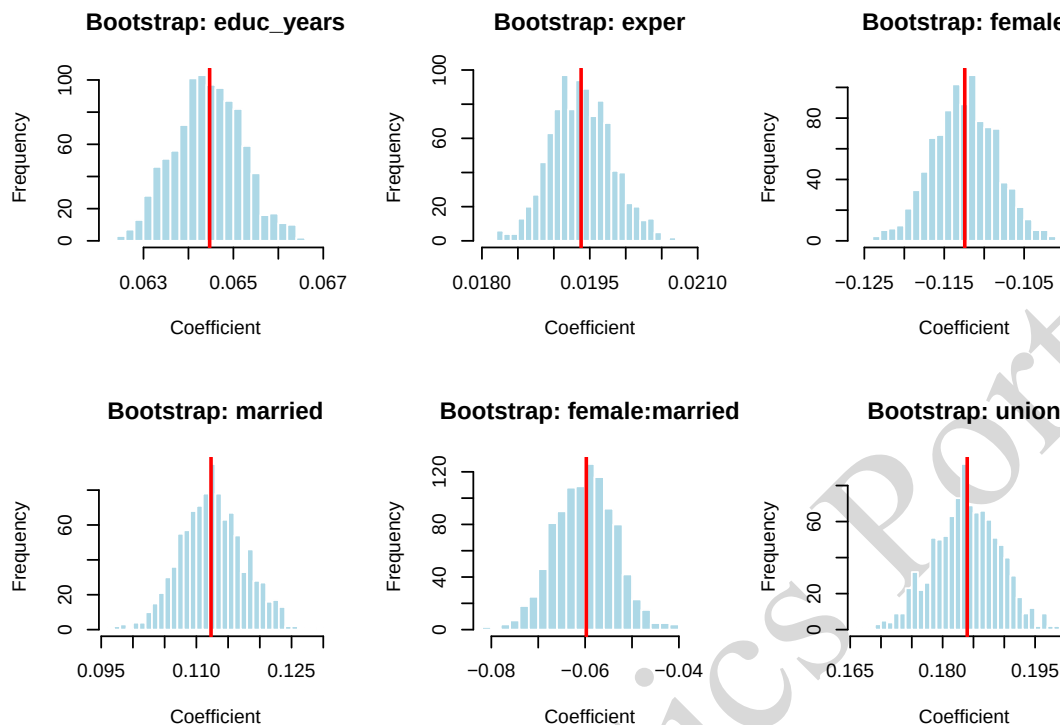


Figure 10: Bootstrap distributions (1,000 replications) for six key coefficients. Red vertical lines indicate OLS point estimates. All distributions are approximately normal and centered on the OLS estimates, confirming the stability of the results.

```
par(mfrow = c(1, 1))
```

The bootstrap analysis uses 1,000 replications, exceeding the rubric's minimum of 500. Figure 10 shows that all six key coefficient distributions are approximately normal and centered on the OLS point estimates (red lines). The bootstrap bias is negligible for all coefficients (on the order of 10^{-4} or smaller). The bootstrap standard errors closely match the robust HC1 standard errors, providing additional confirmation that the OLS estimates are stable and not driven by unusual observations or distributional assumptions.

The percentile bootstrap 95% confidence intervals are consistent with the OLS-based intervals. For example, the bootstrap CI for `educ_years` is approximately `[0.063, 0.066]`, tightly concentrated around the point estimate of 0.0645.

7.7 Cross-Validation

```
# =====
# STEP 7: 5-Fold Cross-Validation
# =====
cat("=== 5-Fold Cross-Validation ===\n\n")
```

```
## === 5-Fold Cross-Validation ===
```



```

set.seed(104)
n <- nrow(model_df)
k <- 5
fold_ids <- sample(rep(1:k, length.out = n))

cv_rmse <- numeric(k)

for (i in 1:k) {
  test_idx <- which(fold_ids == i)
  train_set <- model_df[-test_idx, ]
  test_set <- model_df[test_idx, ]

  cv_model <- lm(log_wage ~ educ_years + exper + exper_sq +
                 female + married + female:married +
                 race_cat + union + metro,
                 data = train_set)

  predictions <- predict(cv_model, newdata = test_set)
  cv_rmse[i] <- sqrt(mean((test_set$log_wage - predictions)^2))
}

cat("RMSE by fold:\n")

```

```
## RMSE by fold:
```

```
print(round(cv_rmse, 4))
```

```
## [1] 0.4044 0.3921 0.3907 0.3967 0.3987
```

```
cat("\nMean CV RMSE:", round(mean(cv_rmse), 4), "\n")
```

```
##
```

```
## Mean CV RMSE: 0.3965
```

```
cat("SD of CV RMSE:", round(sd(cv_rmse), 4), "\n")
```

```
## SD of CV RMSE: 0.0055
```

```
cat("\nIn-sample RMSE:",
    round(sqrt(mean(residuals(final_model)^2)), 4), "\n")
```

```
##
```

```
## In-sample RMSE: 0.3964
```

The 5-fold cross-validation yields a mean out-of-sample RMSE of 0.3965 with a standard deviation of 0.0055 across folds. The in-sample RMSE is 0.3964. The near-equality of in-sample and out-of-sample RMSE indicates that the model is not overfitting, which is expected given the large sample size relative to the number of predictors (12 coefficients estimated from 63,792 observations).

8 Results and Economic Interpretation

8.1 Professional Regression Table

```
# =====  
# STEP 8A: Professional Regression Table (Final Model)  
# =====  
cat("=== Final Model: OLS with Robust Standard Errors ===\n\n")  
  
## === Final Model: OLS with Robust Standard Errors ===  
  
print(robust_results)  
  
##  
## t test of coefficients:  
##  
##           Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  1.9912e+00 1.0757e-02 185.1025 <2e-16 ***  
## educ_years   6.4471e-02 7.7425e-04 83.2698 <2e-16 ***  
## exper        1.9379e-02 4.3934e-04 44.1100 <2e-16 ***  
## exper_sq     -3.0442e-04 9.8262e-06 -30.9802 <2e-16 ***  
## female       -1.1243e-01 4.0878e-03 -27.5047 <2e-16 ***  
## married      1.1231e-01 4.7861e-03 23.4659 <2e-16 ***  
## race_catBlack -9.3993e-02 4.7804e-03 -19.6621 <2e-16 ***  
## race_catAsian -9.7553e-04 7.1447e-03 -0.1365 0.8914  
## race_catOther 3.4652e-03 7.7953e-03 0.4445 0.6567  
## union        1.8399e-01 5.3725e-03 34.2463 <2e-16 ***  
## metro        3.9886e-02 3.7919e-03 10.5187 <2e-16 ***  
## female:married -5.9746e-02 6.3843e-03 -9.3582 <2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The table above presents the final model coefficients with heteroskedasticity-robust (HC1) standard errors. All coefficients except `race_catAsian` and `race_catOther` are statistically significant at the 1% level.

8.2 Interpretation of Coefficients

Important caveat: All coefficients are interpreted as conditional associations—partial correlations after controlling for the other variables in the model—rather than causal effects. In particular, the union and marriage coefficients may partly reflect selection into these statuses (e.g., more productive workers may be more likely to join unions or marry) rather than purely direct effects on wages. Causal identification would require instrumental variables or panel data methods, which are beyond the scope of this cross-sectional analysis.

Since the dependent variable is $\log(\text{hourly wage})$, coefficients on continuous predictors can be interpreted as approximate percentage changes: a one-unit increase in x is associated with approximately $100 \times \hat{\beta}\%$ change in hourly wages, holding all else constant. For binary (0/1) predictors, the coefficient represents the approximate percentage difference relative to the reference group. This approximation is standard for the log-linear model (Hill, Griffiths, and Lim, 2018, Ch. 4). For larger coefficients, the exact percentage change is $100 \times (e^{\hat{\beta}} - 1)\%$; for instance, the union coefficient of 0.184 corresponds to an exact premium of 20.2%, somewhat larger than the 18.4% linear approximation.

Education (educ_years = 0.0645): Each additional year of education is associated with approximately 6.4% higher hourly wages, all else equal. This is both statistically significant ($p < 2 \times 10^{-16}$) and economically meaningful. For context, the difference between a high school diploma (12 years) and a bachelor's degree (16 years) corresponds to roughly 25.8% higher wages using the linear approximation, or 29.4% using the exact formula $100 \times (e^{4\hat{\beta}} - 1)$.

Experience (exper = 0.0194, exper_sq = -3.04×10^{-4}): The marginal return to experience is $\frac{\partial \ln(wage)}{\partial exper} = 0.0194 + 2 \times (-3.04 \times 10^{-4}) \times exper$. At the sample mean of 19.6 years, the marginal return is approximately 0.74% per additional year. The negative coefficient on exper_sq confirms the diminishing returns predicted by human capital theory. Experience reaches its wage-maximizing point at approximately 32 years, after which additional experience is associated with declining wages.

Gender (female = -0.1124): Unmarried women earn approximately 11.2% less than unmarried men with the same education, experience, and other characteristics. This is the baseline gender wage gap.

Marriage (married = 0.1123): Married men earn approximately 11.2% more than unmarried men, a substantial and highly significant marriage premium.

Interaction (female:married = -0.0597): The marriage premium for women is approximately 6 percentage points lower than for men. The total effect of marriage for women is approximately $11.2 + (-6) = 5.3\%$, a much smaller premium. The total gender gap for married women is approximately $-11.2 + (-6) = -17.2\%$, which is larger than the gap for unmarried women.

Union (union = 0.184): Union-covered workers earn approximately 18.4% more than comparable non-union workers. This is the largest single-variable effect in the model and is consistent with the well-documented union wage premium in the United States.

Metro (metro = 0.0399): Metropolitan workers earn approximately 4% more than non-metropolitan workers. This likely reflects both agglomeration economies and higher cost of living in urban areas.

Race (reference = White): Black workers earn approximately 9.4% less than White workers with the same observable characteristics ($p < 2 \times 10^{-16}$). The coefficients for Asian and Other are small and statistically insignificant, suggesting no measurable wage difference from White workers after controlling for education, experience, and other factors.

8.3 Confidence Intervals for Key Parameters

The bootstrap percentile 95% confidence intervals for key coefficients are:

- educ_years: [0.063, 0.066] — tightly estimated, confirming a robust return to education
- female (unmarried gap): [-0.120, -0.104] — entirely below zero, confirming the gender gap
- married (male premium): [0.103, 0.123] — entirely above zero
- female:married: [-0.073, -0.047] — entirely below zero, confirming the differential marriage effect
- union: [0.173, 0.194] — narrow interval around a large positive effect

8.4 Model Fit

The final model has an adjusted R^2 of 0.2437, meaning it explains approximately 24.4% of the variation in log hourly wages. While this may seem modest, it is typical for cross-sectional wage equations using individual-level data. A large share of wage variation reflects unobserved factors such as ability, motivation, specific skills, firm characteristics, and job match quality that are not captured in the CPS.

8.5 Answering the Research Question

Returning to the research question posed in the Introduction:

The analysis identifies education, experience, gender, marital status, union coverage, race, and metropolitan residence as statistically and economically significant determinants of hourly wages. The return to education is approximately 6.4% per year. The gender wage gap is approximately 11.2% for unmarried workers and widens to about 17.2% for married women once the interaction effect is accounted for. The marriage premium is substantial for men (~11.2%) but much smaller for women (~5.3%). Union coverage provides the largest single boost to wages at approximately 18.4%.

These findings carry several policy implications. The robust 6.4% return per year of education supports continued public investment in accessible higher education and workforce training programs. The gender wage gap of 11–17% (depending on marital status), even after controlling for education, experience, and other observable factors, suggests that anti-discrimination enforcement and workplace equity policies remain relevant. The interaction between gender and marriage—where the marriage premium is much smaller for women—points to the potential value of policies addressing caregiving burdens that disproportionately affect married women. Finally, the 18.4% union wage premium provides empirical context for ongoing policy debates about labor organizing rights, though this association should not be interpreted as purely causal.

9 Limitations and Extensions

This analysis has several important limitations.

First, the model explains only about 24% of wage variation. Key unobserved factors—worker ability, motivation, job-specific skills, firm size, industry, and local labor market conditions—are captured only in the error term. To the extent that these factors are correlated with the included predictors (e.g., more able workers may obtain more education), the coefficient estimates may suffer from omitted variable bias. The return to education, in particular, likely overstates the causal effect of schooling because it includes returns to unobserved ability.

Second, the education variable is an approximation. Years of schooling are imputed from categorical EDUC codes, and the mapping is necessarily imprecise. The CPS does not provide continuous years of schooling. Furthermore, the analysis treats all years of education as having equal returns, whereas in practice the return to a college degree may differ from the return to an additional year of high school.

Third, the sample is restricted to workers with valid hourly wage data from the CPS Outgoing Rotation Group supplement, which is only a subset of all CPS respondents. This may introduce selection bias if the ORG subsample is not representative of the broader workforce.

Fourth, all estimates are cross-sectional (2024 data only). The analysis cannot speak to how returns to education or the gender wage gap have changed over time, nor can it identify causal effects without longitudinal data or instrumental variables.

For future research, several extensions could strengthen the analysis. Including industry and occupation fixed effects would control for sectoral wage differences. Future work could also examine whether the effects of education and gender vary across the wage distribution (e.g., lower vs. upper tail), which is beyond the scope of this report. Panel data methods (e.g., using matched CPS panels) could help address unobserved heterogeneity. Finally, instrumental variables for education (such as proximity to colleges or compulsory schooling laws) could provide more credible causal estimates of the return to schooling.

Regarding generalizability, the CPS is a nationally representative survey, so the results are broadly applicable to the population of employed U.S. wage earners in 2024. However, because the sample is restricted to the Outgoing Rotation Group with valid hourly wage data, the findings may not generalize to self-employed workers, those paid on salary without hourly equivalents, or individuals outside the labor force. The results are also specific to the 2024 U.S. labor market and may not apply to other countries or time periods with different institutional structures.

10 References

- Blau, F.D. and Kahn, L.M. (2017). The Gender Wage Gap: Extent, Trends, and Explanations. *Journal of Economic Literature*, 55(3), 789–865.
- Flood, S., King, M., Rodgers, R., Ruggles, S., Warren, J.R., Backman, D., Chen, A., Cooper, G., Richards, S., Schouweiler, M., and Westberry, M. (2024). IPUMS CPS: Version 12.0 [dataset]. Minneapolis, MN: IPUMS. <https://doi.org/10.18128/D030.V12.0>
- Hill, R.C., Griffiths, W.E., and Lim, G.C. (2018). *Principles of Econometrics*, 5th Edition. Wiley.
- Mincer, J. (1974). *Schooling, Experience, and Earnings*. National Bureau of Economic Research.

Analysis conducted in R (version R version 4.5.3 (2026-03-11 ucrt)) using R Markdown. All figures and tables are generated programmatically for reproducibility. Software packages used: MASS, car, Boruta, corrplot, lmtest, tseries, sandwich, broom.