MGTECON 603 - Problem Set 4

(Instructor: Guido Imbens)

Wooyong Park

Collaborators: Cem Kozanoglu, Roberto Gonzalez Tellez, Hanniel Ho, Aileen Wu

October 19, 2025

0 Summary Statistics

The summary statistics of the data are shown in table A.1 in the appendix.

Also, the scatterplot and the correlation matrix of the data are shown in figure A.1 in the appendix.

1 Problem 1

1.1 Construct Age and Wage Variables

I constructed the age variable as age = 1990 - year of birth and the wage variable as wage = exp(log wage).

1.2 Summary Statistics

Based on the results from the R script:

- Mean age: 1955.397 years (standard deviation: 2.905 years)[1]
- Mean wage: \$439.47 (standard deviation: \$364.94)

The age distribution appears to be concentrated around the mean, indicating a relatively homogeneous sample in terms of birth year. The wage distribution shows substantial variation with a standard deviation nearly as large as the mean, suggesting significant income inequality in the sample.

[1]The age variable is constructed as age = 1990 - `yob` in the .txt format dataset. If you consider the `yob` variable to have omitted 1900 in front, then the mean age is 1955.397 - 1900 = 55.397 years.

2 Problem 2

2.1 OLS Regression Results

I estimated the following regression model:

$$wage_i = \beta_0 + \beta_1 \cdot educ_i + \varepsilon_i$$

The regression results show:

	Homoskedastic	White-Robust
(Intercept)	61.1954***	61.1954***
	(2.4623)	(2.5964)
educ	29.6224***	29.6224***
	(0.1868)	(0.2102)
Num. obs.	329509	329509
R^2 (full model)	0.0709	0.0709
Adj. R^2 (full model)	0.0709	0.0709

Notes: The table shows two regression models: one with homoskedastic standard errors and one with Eicker–Huber–White robust standard errors.

Table 1: Regression of wage on years of education

The education coefficient is highly significant ($p < 0.001$). Each additional year of education is associated with approximately \$29.62 higher wages on average.

2.2 Homoskedasticity vs. Heteroskedasticity

The comparison between homoskedastic and heteroskedastic-robust standard errors reveals:

- The homoskedastic standard errors are slightly smaller than the heteroskedastic-robust ones
- The heteroskedastic-robust standard error for the education coefficient increases from 0.1868 to 0.2102

3 Problem 3

3.1 Residual Analysis

To investigate the presence of heteroskedasticity, I computed the residuals and the squared residuals from the OLS regression and analyzed their relationship with years of education. Since it is difficult to list these values for all observations in this document, I attached the scatter plots of the residuals and the squared residuals against the years of education in figure A.2 in the appendix.

3.2 Average Squared Residuals by Education Level

Figure 1 shows the average squared residuals versus years of education, which displays a clear upward trend, indicating heteroskedasticity.

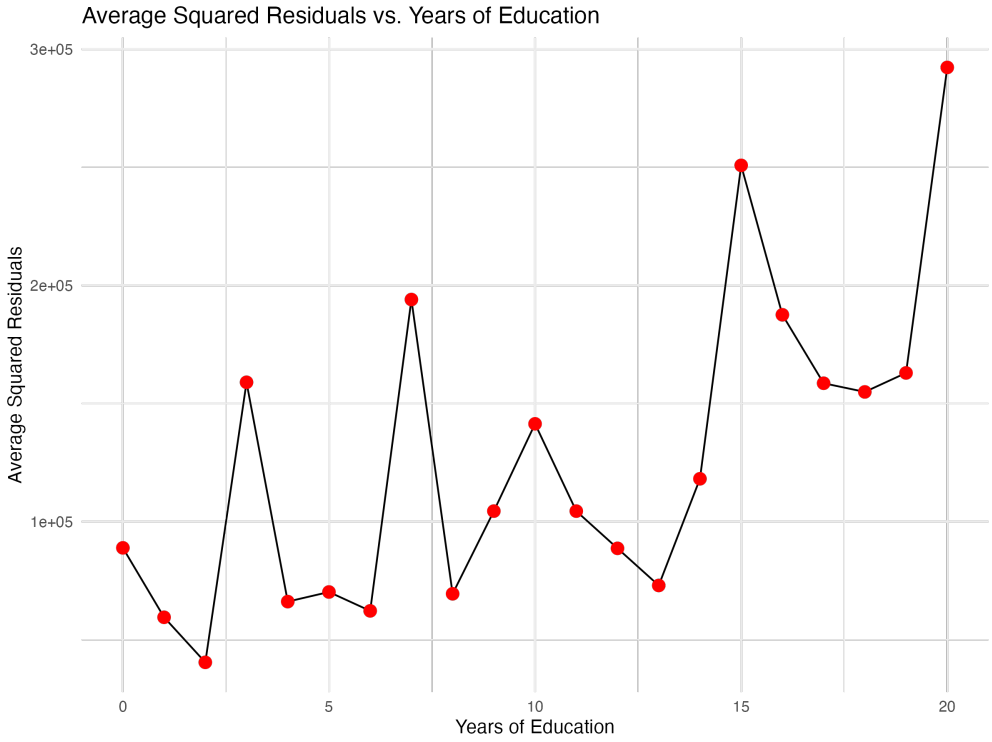


Figure 1: Average Squared Residuals vs. Years of Education

This upward trend in the variance of residuals across education levels provides strong evidence of heteroskedasticity. The assumption of constant variance (homoskedasticity) is violated, as the error variance appears to increase with the level of education.

3.3 Economic Interpretation

This heteroskedasticity pattern makes economic sense. At lower education levels, wage outcomes are likely more constrained and homogeneous, leading to lower variance. However, at higher education levels, there are likely more diverse career paths and opportunities, resulting in greater wage dispersion and higher variance in the residuals.

4 Problem 4

As in table 1 in the appendix, the homoskedastic standard errors are slightly smaller than the heteroskedastic-robust ones. This is expected from figure 1 and the fact that mean years of education is 12.77 years. Whereas the homoskedastic standard errors compute $\hat{\mathbb{V}}(\hat{\beta}_1) = \frac{\sum_{i=1}^{N} \hat{\varepsilon}_i^2}{(N-2)\sum_{i=1}^{N}(x_i-\bar{x})^2}$, the heteroskedastic-robust standard errors compute $\hat{\mathbb{V}}(\hat{\beta}_1) =$

$\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2\hat{\epsilon}_i^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2\sum_{i=1}^{N}(x_i-\bar{x})^2}$. This implies that the EHW standard errors give more weights to the observations that are further away from the mean of the years of education. Since the squared residuals are much higher for higher years of education, the EHW standard errors are larger for higher years of education.

5 Problem 5

5.1 Bootstrap Analysis with Small Samples (n=20)

I conducted a bootstrap analysis using 10,000 bootstrap samples of size 20 to investigate the coverage properties of confidence intervals under both homoskedastic and heteroskedastic assumptions.

Results:

- Homoskedastic model coverage: 90.83%
- Heteroskedastic-robust model coverage: 90.62%

Both models show coverage rates close to the nominal 95% level, though slightly below. It is closer to the nominal 95% level for the homoskedastic model, although the difference is not significantly large. I believe this is becaues the sample size is small, so the bootstrap standard errors are not very reliable.

6 Problem 6

6.1 Bootstrap Analysis with Medium Samples (n=200)

For samples of size 200:

- Homoskedastic model coverage: 91.52%
- Heteroskedastic-robust model coverage: 94.70%

6.2 Bootstrap Analysis with Large Samples (n=2000)

For samples of size 2000:

- Homoskedastic model coverage: 91.86%
- Heteroskedastic-robust model coverage: 95.32%

6.3 Analysis of Coverage Results

The bootstrap results reveal important patterns:

1. **Sample Size Effects**: As sample size increases from 20 to 200 to 2000, coverage rates generally improve and move closer to the nominal 95% level.

2. **Heteroskedasticity Robustness**: The heteroskedastic-robust standard errors consistently provide better coverage rates than homoskedastic standard errors, particularly as sample size increases. This confirms the presence of heteroskedasticity in the data.

3. **Coverage Convergence**: The heteroskedastic-robust model achieves coverage rates very close to 95% for larger samples (94.70% for n=200, 95.32% for n=2000), while the homoskedastic model remains below 95% even for large samples.

The analysis reveals clear evidence of heteroskedasticity in the wage-education relationship, with error variance increasing with education level. The bootstrap simulations demonstrate that heteroskedastic-robust standard errors provide better coverage properties than homoskedastic standard errors, particularly for larger samples.

A Summary Statistics

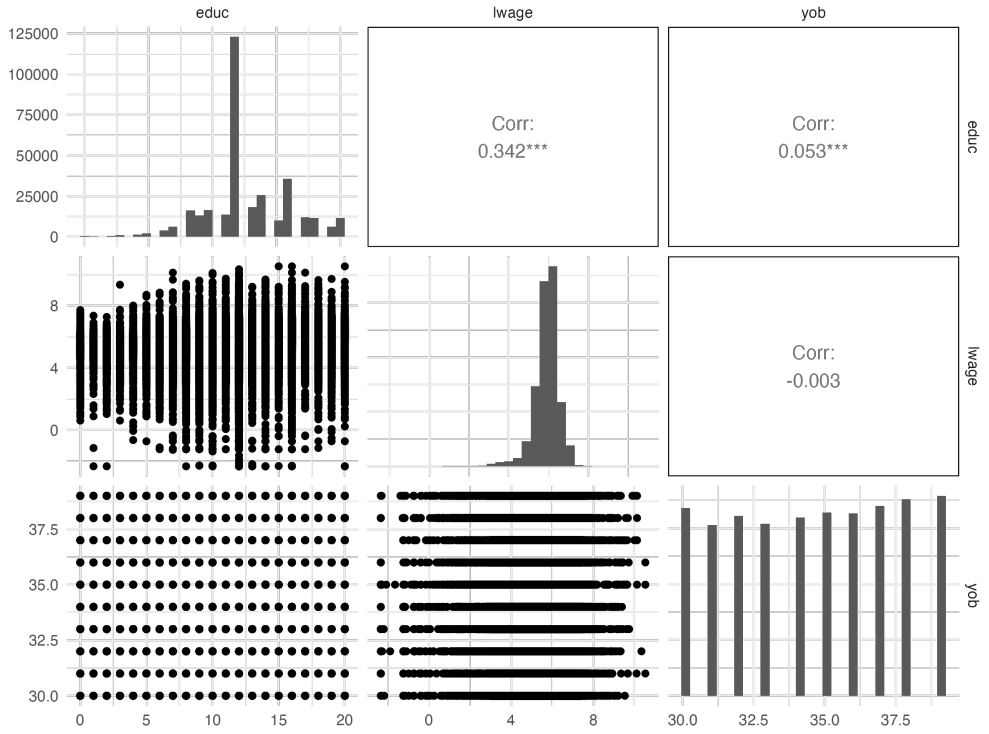
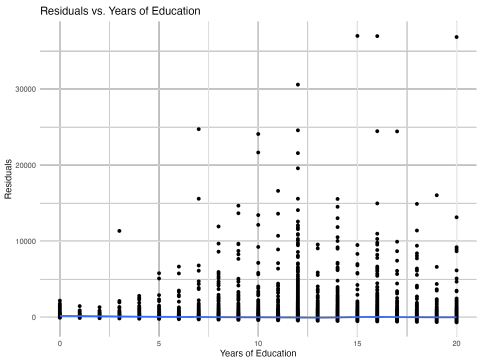
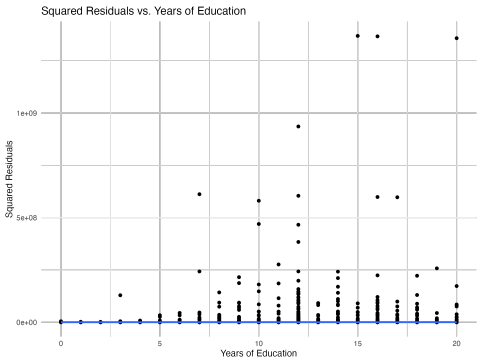


Figure A.1: Scatterplot Matrix

B Residual Analysis



(a) Residuals vs. Years of Education



(b) Squared Residuals vs. Years of Education

Figure A.2: Residual Analysis

variable	numNA	numZeros	fracZeros	mean	sd	min	max	10%	20%	30%	40%	50%	90%	95%	99%	99.9%
educ	0	598	0.0018	12.7699	3.2812	0.0000	20.0000	9.0000	11.0000	12.0000	12.0000	12.0000	17.0000	19.0000	20.000	20.0000
lwage	0	3	0.0000	5.8999	0.6788	-2.3418	10.5321	5.2343	5.5218	5.7243	5.8472	5.9525	6.5567	6.8004	7.274	7.9495
yob	0	0	0.0000	34.6028	2.9050	30.0000	39.0000	30.0000	32.0000	33.0000	34.0000	35.0000	39.0000	39.0000	39.000	39.0000

Table A.1: Summary Statistics