

# Problem Set 4

Cem Kozanoglu

October 19, 2025

Collaborators: Wooyong Park, Roberto Gonzalez-Tellez, Hanniel Ho and Aileen Wu.

## 1 Summary Statistics

We observe that the ages of the sample are uniformly distributed between 51 and 60, with the wages showing a right-skewed distribution, indicating that a majority of individuals earn lower wages while a smaller number earn significantly higher wages.

The mean and standard deviation of the variables are presented in the table below:

Variable	Mean	Standard Deviation
Age	55.40	2.90
Wage	439.47	364.94

Table 1: Summary Statistics

The following graphs depict the distributions of ages and wages respectively:

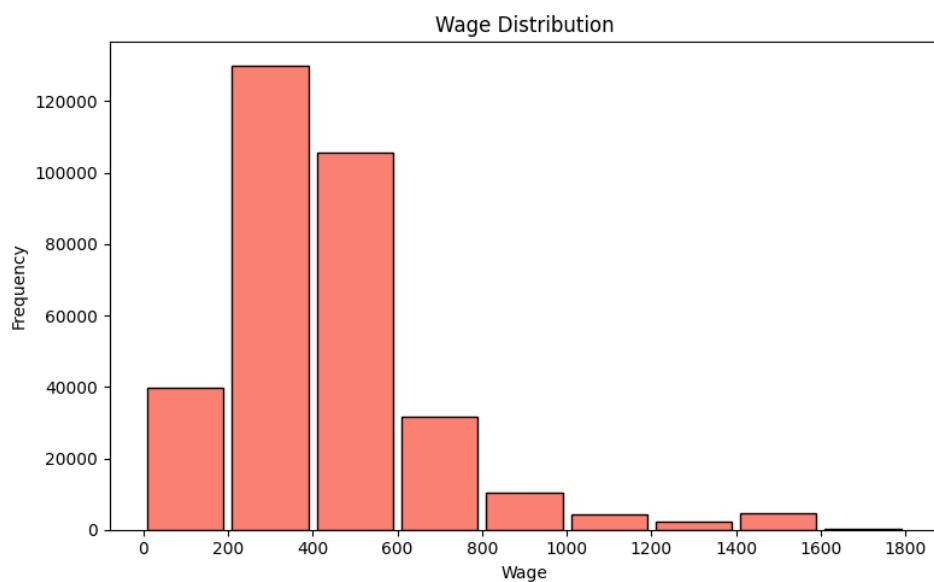


Figure 1: Wage Distribution

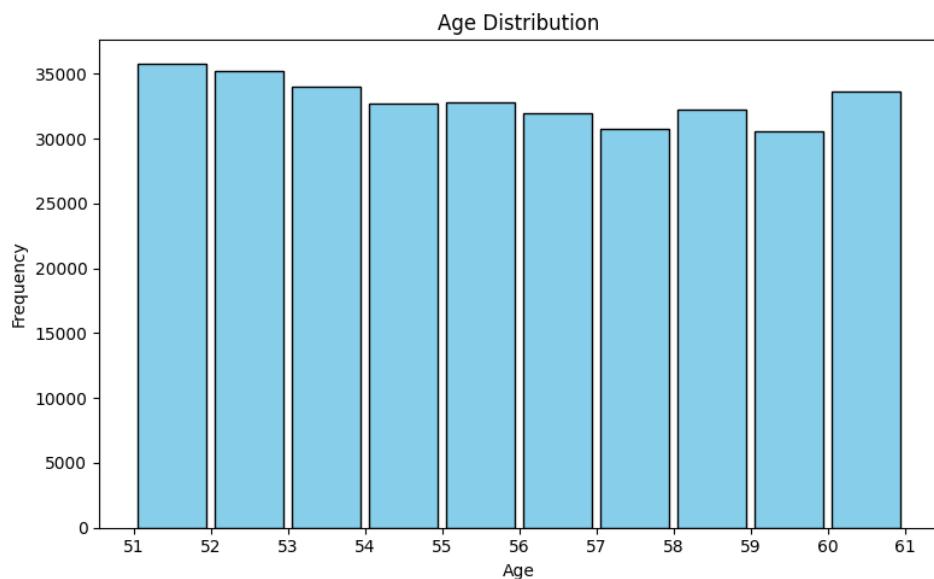


Figure 2: Age Distribution

## 2 Full regression

I ran a regression of wage on education with the following model specification:

$$\text{wage}_i = \beta_0 + \beta_1 \cdot \text{educ}_i + \epsilon_i$$

The following table presents the results of the full regression analysis, and the standard errors are calculated assuming homoskedasticity:

	Homoskedastic	White-Robust
(Intercept)	61.1954*** (2.4623)	61.1954*** (2.5964)
educ	29.6224*** (0.1868)	29.6224*** (0.2102)
Num. obs.	329509	329509
R <sup>2</sup>	0.0709	0.0709
Adj. R <sup>2</sup>	0.0709	0.0709

We can clearly conclude that education has a positive and statistically significant effect on wages at the 1% significance level. Specifically, each additional year of education is associated with an average increase in wages of approximately 30 units (we are not told the unit explicitly in the problem set but I assume that it corresponds to weekly earnings in dollars).

We also observe a higher heteroskedasticity-robust standard error for the education coefficient compared to the homoskedastic standard error (0.2102 vs. 0.1868). The interpretation of this will be discussed in detail in Section 4 below.

### 3 Squared Residuals by year

To investigate the presence of heteroskedasticity in the regression model, I plotted the squared residuals against the years of education. The plot is shown below:

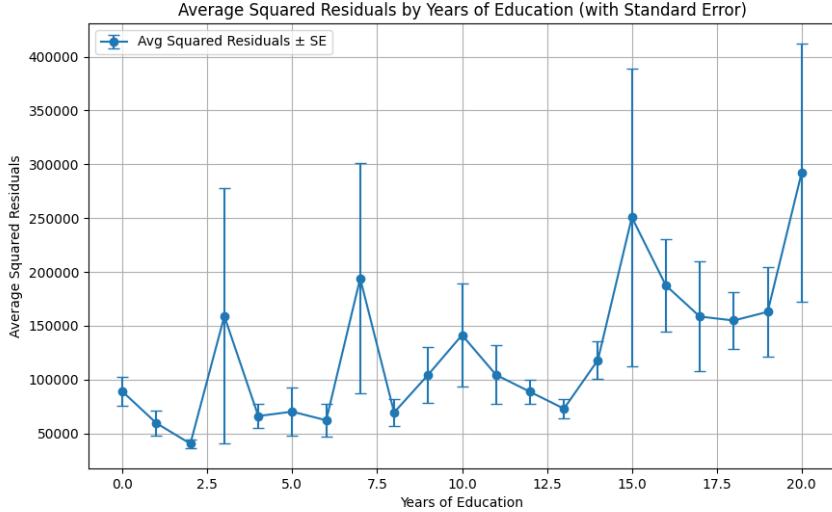


Figure 3: Squared Residuals vs Education

The plot shows that both the average squared residuals and the standard error of the averages increase with the years of education, suggesting the presence of heteroskedasticity in the model. This indicates that the variance of the error terms is not constant across all levels of education, which violates one of the key assumptions of the classical linear regression model.

Two interesting observations can be made from this plot.

First, there is a general upward trend in the average squared residuals as the years of education increase, indicating that individuals with higher education levels tend to have more variability in their wages. This could be due to a variety of factors, such as differences in job types, industries, or other unobserved characteristics that are correlated with education, and is in line with our priors about what we might expect.

Second, an interesting observation is that we can see a spike in the average squared residuals in the years 3, 7 and 15 which indicate that our models is relatively 'worse' at modelling the outcomes for people at this education level. This might be due to the fact that these education levels correspond to the year *before* educational milestones (e.g., completion of certain degrees or certifications) that could lead to greater variability in wages. Consider that if someone has gone through 15 years of education, they are likely to have benefitted significantly from the human capital aspects of education without the

'signalling' benefits that come with actually completing the education and obtaining a degree. This may be leading some individuals to successfully take advantage of their education in the labor market, while others may not be able to take advantage without the signalling advantages that the degree brings, resulting in higher variability in wages for this group.

## 4 Homoskedastic vs Heteroskedastic Standard Errors

As mentioned in Section 2, the heteroskedasticity-robust standard error for the education coefficient (0.2102) is higher than the homoskedastic standard error (0.1868). This difference can be attributed to the presence of heteroskedasticity in the model, as indicated by the plot of squared residuals against years of education in Section 3. This result is also consistent with the data we are working with, as wages tend to exhibit heteroskedasticity due to various factors such as differences in job types, industries, and individual characteristics that are correlated with education.

First let's consider how the two standard errors are calculated. The homoskedastic standard error assumes that the variance of the error terms is constant across all levels of the independent variable (education in this case). Under this assumption, the formula for the homoskedastic standard error is:

$$\hat{V}_{homoskedastic} = \frac{1}{N} \sum_{i=1}^n \hat{\epsilon}_i^2 \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

The heteroskedasticity-robust standard error relaxes this assumption and allows for the variance of the error terms to vary across different levels of the independent variable. The formula is:

$$\hat{V}_{heteroskedastic} = \frac{\frac{1}{N} \sum_{i=1}^n \hat{\epsilon}_i^2 (X_i - \bar{X})^2}{N \cdot (\frac{1}{N} \sum_{i=1}^n (X_i - \bar{X})^2)^2}$$

In this case, the presence of heteroskeasticity leads to a larger standard error because values further from the mean of education (i.e., individuals with very low or very high education levels) tend to have larger squared residuals, as seen in the plot in Section 3.

This results in a higher overall variance estimate when using the heteroskedasticity-robust formula, leading to a larger standard error for the education coefficient. Mathematically, this is because while the homoskedastic estimator calculates the average of the squared residuals, the heteroskedasticity-robust estimator weights these squared residuals by the squared deviation of each observation from the mean of education. As a result, observations with larger deviations (which tend to have larger squared residuals due to heteroskedasticity) contribute more to the overall variance estimate, leading to a higher standard error.

## 5 Subsample Confidence Intervals and Coverage Probabilities

I sampled 10,000 subsamples of sizes 20, 200, and 2000 from the original dataset without replacement. For each subsample, I estimated the regression of wage on education and constructed 95% confidence intervals for the coefficient of education using both homoskedastic and heteroskedastic standard errors. I then calculated the coverage probabilities by determining the proportion of these confidence intervals that contained the true population coefficient for education.

The coverage probabilities for both homoskedastic and heteroskedastic standard errors across different sample sizes are summarized in the table below:

Sample Size	Homoskedastic Coverage Probability	Heteroskedastic Coverage Probability
20	88.3%	87.2%
200	90.7%	94.2%
2000	91.8%	95.5%

Table 2: Coverage Probabilities by Sample Size

We observe that as the sample size increases, the coverage probabilities for the heteroskedastic standard errors approach the nominal level of 95%, while the homoskedastic

standard errors consistently underperform, especially in smaller samples. This indicates that heteroskedasticity-robust standard errors provide more reliable inference in the presence of heteroskedasticity, particularly as the sample size grows. It is also a further corroboration of our earlier finding that the model exhibits heteroskedasticity, as the robust standard errors yield better coverage probabilities.