Violation of Assumptions of Classical Regression Model: Heteroscedasticity

Unit 3

Prepared by

Aditya Korekallu Srinivasa

Scientist (Senior Scale)

ICAR-Indian Agricultural Research Institute
New Delhi

Contents

1 What is Heteroscedasticity?

1.1 Definition and Concept

Heteroscedasticity occurs when the variance of the error term (u_i) in a regression model is not constant across all observations:

$$\text{Var}(u_i|X) \neq \sigma^2$$

This violates the classical regression assumption of **homoscedasticity**, where the error variance is constant for all values of the regressors.

1.2 Why is it a Violation?

Heteroscedasticity leads to inefficient OLS estimators and invalid standard errors, making hypothesis tests unreliable. While OLS coefficients remain unbiased and consistent, they are no longer the Best Linear Unbiased Estimators (BLUE).

1.3 Visual Illustration

A scatterplot of residuals against fitted values or a regressor often shows a "fan" or "cone" shape, indicating increasing or decreasing variance.

2 Reasons for Heteroscedasticity

2.1 Nature of Data

- **Cross-sectional data:** Variance in the dependent variable often increases with the level of an independent variable (e.g., income and expenditure).
- **Time series data:** Periods of instability or structural change can cause changing error variance.

2.2 Model Specification Issues

- **Omitted variables:** Leaving out relevant variables that affect the variance of the dependent variable.
- **Incorrect functional form:** Using a linear model when a log or quadratic form is more appropriate.

2.3 Economic or Physical Causes

- **Income and expenditure:** High-income households have more variable expenditures.
- **Firm size:** Larger firms may have more volatile profits.
- **Agricultural yields:** Variability may increase with landholding or input use.

2.4 Practical Example

In a regression of household consumption on income, the variance of consumption typically increases with income due to more diverse spending patterns.

3 Practical Situations Where Heteroscedasticity Occurs

3.1 Economic Data Examples

- **Income and expenditure surveys:** Expenditure variance increases with income.
- **Firm-level studies:** Output or profit variability increases with firm size.
- **Agricultural production:** Larger farms show greater yield variance.
- **Real estate prices:** Higher price variability in expensive neighborhoods.

3.2 Time Series Examples

- **Inflation rates:** Volatility increases during economic crises.
- **Financial returns:** Volatility clustering in stock returns (ARCH/GARCH effects).

3.3 Graphical Illustration

Residual plots displaying a "funnel" shape (narrow at low values, wide at high values) indicate increasing variance.

4 Detection and Identification of Heteroscedasticity

Detecting heteroscedasticity requires both graphical and formal statistical methods. Below are the main approaches, with guidance on interpretation.

4.1 Graphical Methods

- **Residual Plots:** Plot OLS residuals ($\hat{u}_i$) or squared residuals ($\hat{u}_i^2$) against fitted values or independent variables.
- **Interpretation:**
  - If the spread of residuals is roughly constant, homoscedasticity is plausible.
  - If the spread increases or decreases systematically (fan or cone shape), heteroscedasticity is likely.
- **Example:** In a regression of expenditure on income, if the residual spread widens with income, this suggests heteroscedasticity.

4.2 Formal Statistical Tests

1. Breusch-Pagan Test

- **Procedure:**
  1. Estimate the original regression and obtain OLS residuals ($\hat{u}_i$).
  2. Regress squared residuals ($\hat{u}_i^2$) on the independent variables.
  3. Compute the test statistic: nR^2, where R^2 is from the auxiliary regression.
- **Decision Rule:**
  - Under the null of homoscedasticity, nR^2 follows a chi-square distribution with degrees of freedom equal to the number of regressors.
  - **If the p-value is less than 0.05, reject homoscedasticity.**
- **Interpretation:**
  - A significant result (low p-value) indicates heteroscedasticity.
  - A non-significant result suggests no strong evidence against homoscedasticity.

2. White Test

- **Procedure:**
  1. Regress $\hat{u}_i^2$ on all independent variables, their squares, and cross-products.
  2. Compute nR^2.
- **Decision Rule:**
  - Compare nR^2 to the chi-square distribution.
  - **A p-value below 0.05 indicates heteroscedasticity.**
- **Interpretation:**
  - The White test detects general forms of heteroscedasticity, including nonlinearities and interactions.
  - A significant result signals heteroscedasticity.

3. Goldfeld-Quandt Test

- **Procedure:**
  1. Order data by a suspected variable.
  2. Split into two groups, omitting a middle portion.
  3. Estimate regressions for each group; compute RSS for each.
  4. Calculate $F = \frac{RSS_{large}/df_{large}}{RSS_{small}/df_{small}}$.
- **Decision Rule:**
  - Compare F to the critical value from the F-distribution.

- **If F exceeds the critical value, reject homoscedasticity.**

- **Interpretation:**
  - A significant result (high F) indicates heteroscedasticity.

4. Park and Glejser Tests

- **Park Test:**
  - Regress $\log(\hat{u}_i^2)$ on $\log(X_i)$.
  - A significant slope coefficient suggests heteroscedasticity.

- **Glejser Test:**
  - Regress $|\hat{u}_i|$ on X_i or a function of X_i.
  - A significant coefficient indicates heteroscedasticity.

4.3 Summary Table: Interpreting Test Results

Test	Procedure
Breusch-Pagan	Regress squared residuals on regressors; compute nR^2
White	Regress squared residuals on regressors, their squares, and cross-products; compute
Goldfeld-Quandt	Split sample, compare group variances using F-test
Park/Glejser	Regress $\log(\hat{u}_i^2)$ or $|\hat{u}_i|$ on suspected variable(s)

4.4 Practical Guidance

- Always begin with graphical analysis; it may reveal patterns missed by formal tests.
- Use multiple tests for confirmation, as each has different sensitivity.
- A significant test result (low p-value or large F statistic) means the null of homoscedasticity is rejected, indicating heteroscedasticity.
- Non-significant results suggest no strong evidence of heteroscedasticity, but do not guarantee its absence.

5 Consequences of Heteroscedasticity

5.1 Theoretical Consequences

- **Unbiasedness:** OLS estimators remain unbiased and consistent.
- **Inefficiency:** OLS is no longer efficient (not BLUE); other estimators may have smaller variance.
- **Incorrect standard errors:** OLS standard errors are biased, leading to unreliable t- and F-tests.

- **Invalid inference:** Confidence intervals and hypothesis tests may be misleading.

5.2 Practical Consequences

- **Over- or underestimation of significance:** Variables may appear insignificant (or significant) due to incorrect standard errors.
- **Misleading policy conclusions:** Policy recommendations based on incorrect inference can be problematic.
- **Reduced predictive accuracy:** Model predictions may be less reliable, especially at data extremes.

5.3 Example

If error variance increases with income, OLS standard errors for the income coefficient are underestimated, overstating statistical significance.

6 Remedies for Heteroscedasticity

6.1 Data Transformation

- **Logarithmic transformation:** Taking logs of the dependent variable (and/or regressors) often stabilizes variance.
- **Other transformations:** Square root or reciprocal transformations may help, depending on the form of heteroscedasticity.

6.2 Model Specification

- **Add omitted variables:** Including relevant variables that explain variance may reduce heteroscedasticity.
- **Use appropriate functional form:** Specify the model according to the true relationship.

6.3 Robust Estimation

- **Heteroscedasticity-robust standard errors:** Use White's or other robust standard errors for valid inference without changing OLS coefficients.
- **Weighted Least Squares (WLS):** If the form of heteroscedasticity is known, weight observations inversely proportional to the error variance.
- **Generalized Least Squares (GLS):** More general approach when the structure of heteroscedasticity is known or can be modeled.

6.4 Interpretation and Cautions

If the primary goal is prediction and the model fits well, mild heteroscedasticity may not be a major concern. For inference, addressing heteroscedasticity is essential.

Glossary

- **Heteroscedasticity**: The condition where the variance of the error term is not constant across observations.
- **Homoscedasticity**: The classical assumption that the error variance is constant.
- **Breusch-Pagan Test**: A formal test for heteroscedasticity based on regressing squared residuals on regressors.
- **White Test**: A general test for heteroscedasticity using squared residuals and all possible cross-products of regressors.
- **Weighted Least Squares (WLS)**: An estimation method that assigns weights to observations to correct for heteroscedasticity.
- **Robust Standard Errors**: Adjusted standard errors that remain valid in the presence of heteroscedasticity.
- **Goldfeld-Quandt Test**: A test for heteroscedasticity based on comparing variances in subgroups of the data.

Practice Questions

1. Define heteroscedasticity. How does it violate the classical regression assumptions?
2. List and explain the main causes of heteroscedasticity in applied econometric work.
3. Give practical examples where heteroscedasticity is likely to occur in economic data.
4. Describe and interpret at least three methods for detecting heteroscedasticity in a regression model.
5. What are the consequences of heteroscedasticity for OLS estimation and inference?
6. Discuss at least three remedies for heteroscedasticity. When should each be used?
7. Is heteroscedasticity always a problem? Explain with reference to prediction versus inference.
8. How can robust standard errors be used to address heteroscedasticity?
9. Explain the use and interpretation of the Breusch-Pagan and White tests.
10. What is weighted least squares and when is it appropriate to use?

References

- Gujarati, D. N., & Porter, D. C. (2010). *Basic Econometrics* (5th ed.). McGraw-Hill. [See especially Chapter 11: Heteroscedasticity]
- Wooldridge, J. M. (2013). *Introductory Econometrics: A Modern Approach* (5th ed.). Cengage Learning. [See discussion in the chapter on multiple regression and heteroscedasticity]