

Lecture 26: Heteroskedasticity: concept, sources and consequences for OLS

Econometrics ■ Unit II: Multicollinearity and heteroskedasticity

Kapil Kumar

Department of Statistics
Chaudhary Charan Singh University, Meerut

B.Sc. (Hons) Statistics, Semester V
Academic Session 2026–2027

Previous lecture

Lecture 25: Limitations of ridge regression; summary of key formulas.

Today

- Concept
- Consequences of heteroskedasticity for OLS
- Tests for heteroskedasticity

- The classical model assumes that every disturbance has the same variance: $\text{Var}(\varepsilon_i) = \sigma^2$ for all i . This is called **homoskedasticity** (equal spread).
- If the variances differ,

$$\text{Var}(\varepsilon_i) = \sigma_i^2, \quad i = 1, \dots, n,$$

the disturbances are **heteroskedastic**.

- Keeping the disturbances uncorrelated, the covariance matrix is diagonal:

$$\text{Var}(\varepsilon) = \Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2).$$

- This is the GLS model of Chapter 2 with $\sigma^2\Omega = \Sigma$.

1. **Scale differences in cross-section data.** Expenditure varies much more among rich households than among poor households, because the rich have more choice. Large firms show larger fluctuations in output than small firms.
2. **Learning.** As people learn, their errors of behaviour (for example, typing errors) become smaller, so the variance falls over time.
3. **Grouped data.** Suppose each observation is the average of m_g individual values, each with independent error of variance σ^2 . The error of the group average is $\bar{\varepsilon}_g = \frac{1}{m_g} \sum_{i=1}^{m_g} \varepsilon_{gi}$, so

$$\text{Var}(\bar{\varepsilon}_g) = \frac{1}{m_g^2} \sum_{i=1}^{m_g} \text{Var}(\varepsilon_{gi}) = \frac{1}{m_g^2} \cdot m_g \sigma^2 = \frac{\sigma^2}{m_g}.$$

Groups of different sizes therefore have different error variances.

4. **Outliers and wrong functional form** can also create unequal spread of the residuals.

(1) OLS is still unbiased. $b = \beta + (X'X)^{-1}X'\epsilon$ and $E(\epsilon) = 0$ give $E(b) = \beta$. Heteroskedasticity does not affect this.

(2) The variance formula changes. By Rule P4,

$$\text{Var}(b) = (X'X)^{-1}X'\Sigma X(X'X)^{-1},$$

not $\sigma^2(X'X)^{-1}$.

A simple case

For $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, let $S_{xx} = \sum_i (x_i - \bar{x})^2$. The OLS slope is

$$b_1 = \frac{\sum_i (x_i - \bar{x}) y_i}{S_{xx}}.$$

Substituting $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ and using $\sum_i (x_i - \bar{x}) = 0$ and $\sum_i (x_i - \bar{x}) x_i = S_{xx}$,

$$b_1 = \beta_1 + \frac{\sum_i (x_i - \bar{x}) \varepsilon_i}{S_{xx}}.$$

Since the ε_i are uncorrelated with variances σ_i^2 ,

$$\text{Var}(b_1) = \frac{\sum_i (x_i - \bar{x})^2 \sigma_i^2}{S_{xx}^2}.$$

Only if every $\sigma_i^2 = \sigma^2$ does this reduce to the usual σ^2/S_{xx} . If the large variances occur where $(x_i - \bar{x})^2$ is large, the true variance exceeds what the usual formula suggests.

(3) s^2 is biased. By Rule P6, $E(e'e) = \text{tr}(M\Sigma)$, so

$$E(s^2) = \frac{\text{tr}(M\Sigma)}{n - k},$$

which is generally not equal to any common σ^2 .

(4) Tests and intervals are invalid. Because the standard errors are wrong, the usual t and F tests and confidence intervals are unreliable. The error may go in either direction.

(5) OLS is not efficient. By Aitken's theorem, GLS with $\Omega = \Sigma$ has smaller variance. OLS is no longer BLUE.

The null hypothesis and a preliminary check

In each test the null hypothesis is homoskedasticity:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2.$$

- A preliminary check is a plot of squared residuals e_i^2 against $\hat{y}_i$ or against a suspected variable.
- A funnel shape suggests heteroskedasticity.
- The formal tests follow.

Practice

- Exercises 4.1 and 4.2 (Exercises and Solutions booklet).

Next lecture

Lecture 27: Bartlett's test.