HETEROSCEDASTICITY

BY

SANJEEV KUMAR

Department of Economics, Chaudhary Charan Singh University, Meerut

Meaning

In the multiple regression model

$$y = X\beta + \varepsilon,$$

it is assumed that

$$V(\varepsilon) = \sigma^2 I,$$

i.e.,

$$Var(\varepsilon_i^2) = \sigma^2,$$

$$Cov(\varepsilon_i \varepsilon_j) = 0, \; i \neq j = 1, 2, \ldots, n.$$

In this case, the diagonal elements of the covariance matrix of ε are the same indicating that the variance of each ε_i is same and off-diagonal elements of the covariance matrix of ε are zero indicating that all disturbances are pairwise uncorrelated. This property of constancy of variance is termed as **homoskedasticity** and disturbances are called as **homoskedastic disturbances.**

In many situations, this assumption may not be plausible, and the variances may not remain the same. The disturbances whose variances are not constant across the observations are called **heteroskedastic disturbance,** and this property is termed as **heteroskedasticity.** In this case

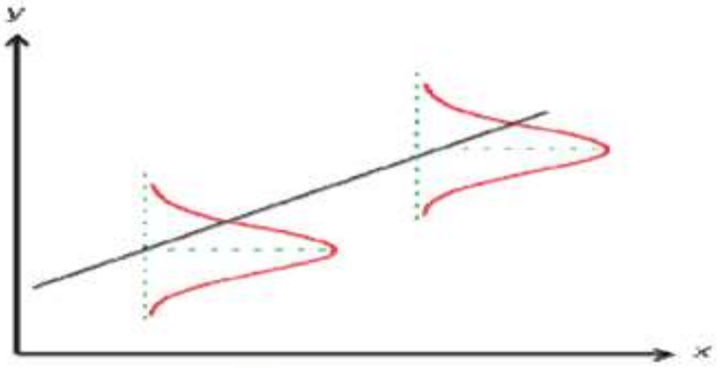
$$Var(\varepsilon_i) = \sigma_i^2,\ i = 1, 2, ..., n$$

and disturbances are pairwise uncorrelated.

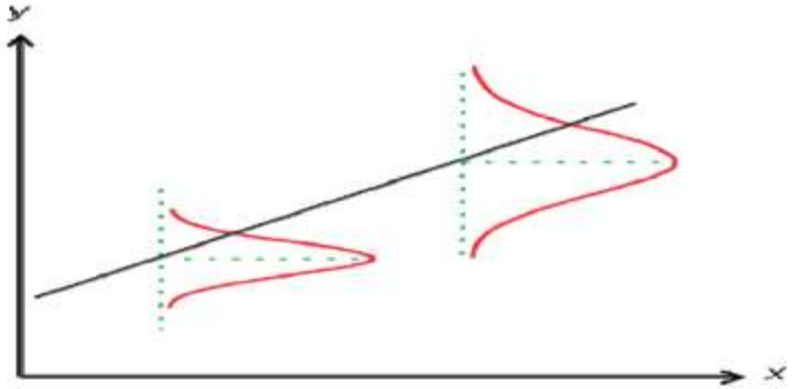
The covariance matrix of disturbances is

$$V(\varepsilon) = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_n^2) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}.$$

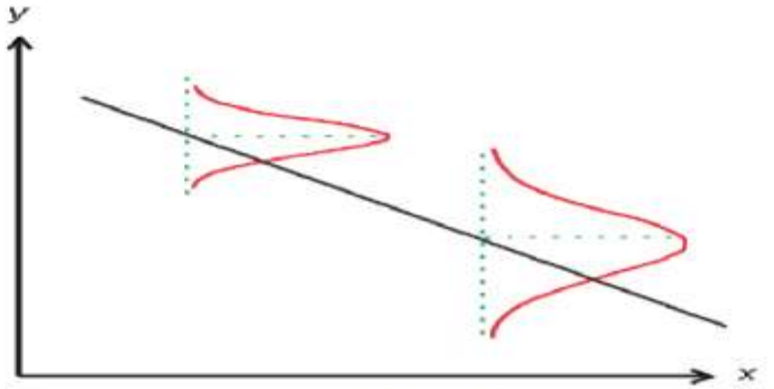
Graphically, the following pictures depict homoskedasticity and heteroskedasticity.



Homoskedasticity



Heteroskedasticity ($Var(y)$ increases with x)



Heteroskedasticity ($Var(y)$ decreases with x)

Examples: Suppose in a simple linear regression model, x denote the income and y denotes the expenditure on food. It is observed that as the income increases, the expenditure on food increases because of the choice and varieties in food increase, in general, up to a certain extent. So the variance of observations on y will not remain constant as income changes. The assumption of homoscedasticity implies that the consumption pattern of food will remain the same irrespective of the income of the person. This may not generally be a correct assumption in real situations. Instead, the consumption pattern changes and hence the variance of y and so the variances of disturbances will not remain constant. In general, it and will be increasing as income increases.

Reasons for the Problem of Heteroscedasticity

follows:

(*i*) One of the sources of heteroscedasticity is grouping. Data from large scale surveys are often published in grouped form with an different number of entities in different groups. Working with group averages in such cases give rise to heteroscedastic disturbances.

(*ii*) There may be certain outlying observation in the data which would increase or decrease error variance.

(*iii*) The problem of heteroscedasticity often arises because the scale of a variable varies enormously within the sample.

(*iv*) Following the error-learning model, as people learn, their errors of behaviour become smaller over time example, typing errors.

(*v*) As incomes grow, people have more discretionary income and hence more scope for choice about the disposition of their income. Hence, σ_i^2 is likely to increase with income. Thus, in the regression of saving on income we may find that σ_i^2 is increasing with income because people have more choices about their savings behaviour. Similarly, growth oriented companies are likely to show more variability in their dividend payout ratio than established companies.

(*vi*) Another source of heteroscedasticity is skewness in the distribution of one or more regressors included in the model. Examples are economic variables such as income, wealth. It is known that the distribution of income and wealth in capitalistic economy is uneven, with the bulk of the income and wealth being owned by a few at the top.

(*vii*) Other sources of heteroscedasticity can also arise because of (*a*) incorrect data transformation and (*b*) incorrect functional form, example linear versus log-linear models

Consequencesof Heteroscedasticity

If $E(u_i^2) \neq \sigma_u^2$ *i.e.* problem of heteroscedasticity is present then we have the following consequences.

1. The coefficient of the estimates will be statistically unbiased. Let us take a two variable model $Y_i = \beta_0 + \beta_1 X_i + u_i$

***Proof*:** Apply OLS $\hat{\beta}_1 = \dfrac{\Sigma x_i y_i}{\Sigma x_i^2}$

$$= \frac{\Sigma x_i (\beta_1 x_i + u_i)}{\Sigma x_i^2}$$

$$= \frac{\beta_1 \Sigma x_i^2}{\Sigma x_i^2} + \frac{\Sigma x_i u_i}{\Sigma x_i^2}$$

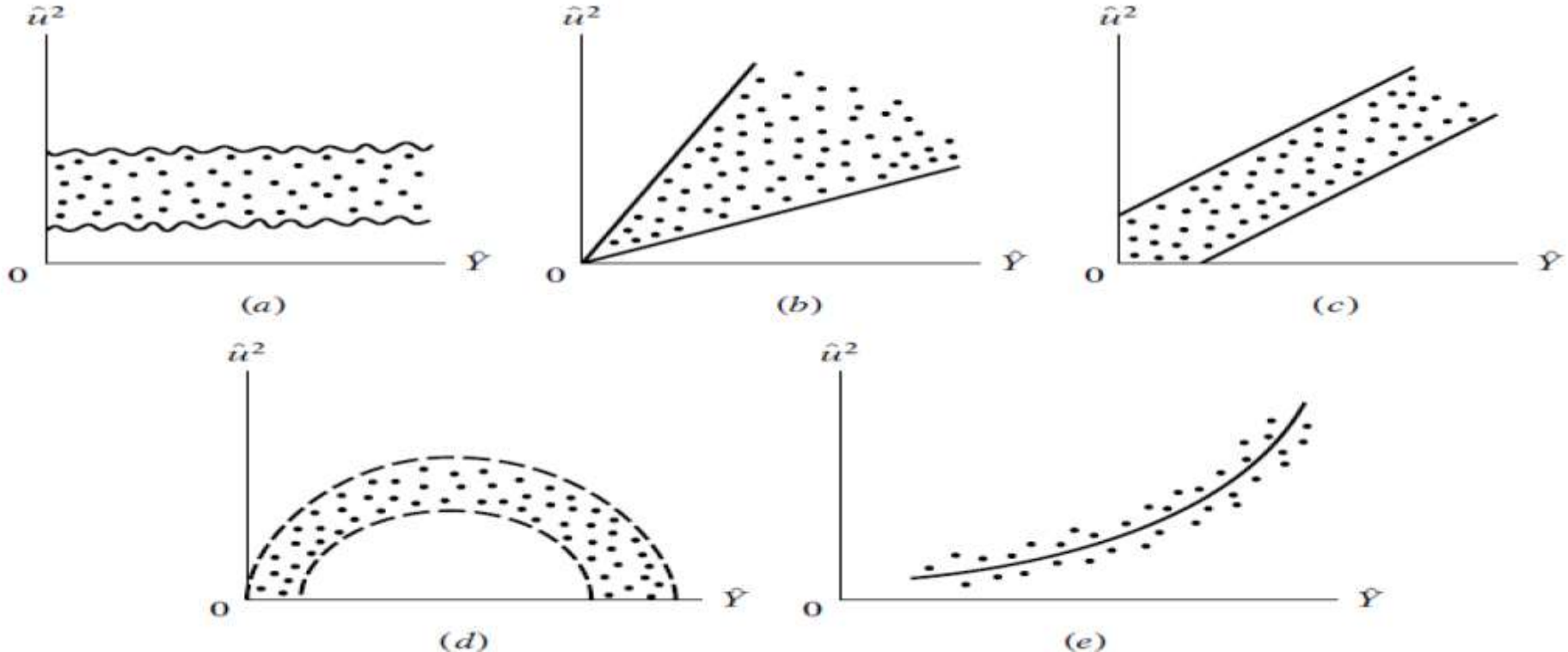
$$= \beta_1 + \frac{\Sigma x_i u_i}{\Sigma x_i^2}$$

Consequences of Heteroscedasticity

- OLS estimators are
  - Still linear
  - Still unbiased
  - NOT *minimum variance*
- Variances of OLS estimators are biased
  - May be positive bias (overestimate) or negative (underestimate)
- Hypothesis tests using t and F distributions are unreliable

GRAPHICAL METHOD

- In this method Residual square (r_i^2 or $r_i^{2'}$)is plotted against the predicted value of the dependent variable. If the plot doesn't show any pattern then heteroscedasticity is said to be absent otherwise it is said to be absent.
- In the figures on next slide first figure (a) shows case of homoscedastic data whereas other figures (b, c, d & e) are the examples of heteroscedastic data [1].



Tests for heteroskedasticity

The presence of heteroskedasticity affects the estimation and test of hypothesis. The heteroskedasticity can enter into the data due to various reasons. The tests for heteroskedasticity assume a specific nature of heteroskedasticity. Various tests are available in the literature, e.g.,

1. Bartlett test
2. Breusch Pagan test
3. Goldfeld Quandt test
4. Glesjer test
5. Test based on Spearman's rank correlation coefficient
6. White test
7. Ramsey test
8. Harvey Phillips test
9. Szroeter test
10. Peak test (nonparametric) test

Spearman's rank correlation test

It d_i denotes the difference in the ranks assigned to two different characteristics of the i^{th} object or phenomenon and n is the number of objects or phenomenon ranked, then the Spearman's rank correlation coefficient is defined as

$$r = 1 - 6\left(\frac{\sum_{i=1}^{n} d_i^2}{n(n^2-1)}\right); \quad -1 \le r \le 1.$$

This can be used for testing the hypothesis about the heteroskedasticity.

Consider the model

$$y_i = \beta_0 + \beta_1 X_i + \varepsilon_i .$$

1. Run the regression of y on X and obtain the residuals e.
2. Consider $|e_i|$.
3. Rank both $|e_i|$ and X_i (or $\hat{y}_i$) in an ascending (or descending) order.
4. Compute rank correlation coefficient r based on $|e_i|$ and X_i (or $\hat{y}_i$).
5. Assuming that the population rank correlation coefficient is zero and $n > 8$, use the test statistic

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

which follows a t-distribution with $(n-2)$ degrees of freedom.

6. The decision rule is to reject the null hypothesis of heteroskedasticity whenever $t_0 \ge t_{1-\alpha}(n-2)$.

If there are more than one explanatory variables, then rank correlation coefficient can be computed between $|e_i|$ and each of the explanatory variables separately and can be tested using t_0.

GLEJSER TEST

- The Glejser test is similar to Park's test. Instead of one Glejser had used different functional forms to model error variance (or its estimate) over explanatory variables. If any of the model comes out to be significant then heteroscedasticity is said to be present.

The test procedure is as follows:

1. Use OLS and obtain the residual vector e on the basis of available study and explanatory variables.
2. Choose Z and apply OLS to

$$|e_i| = \delta_0 + \delta_1 Z_i^h + v_i$$

where v_i is the associated disturbance term.

3. Test $H_0 : \delta_1 = 0$ using t-ratio test statistic.
4. Conduct the test for $h = \pm 1, \pm\frac{1}{2}$. So the test procedure is repeated four times.

In practice, one can choose any value of h. For simplicity, we choose $h = 1$.

- The test has only asymptotic justification and the four choices of h give generally satisfactory results.
- This test sheds light on the nature of heteroskedasticity.

PARK'S TEST

Park had modeled the error variance as a function of explanatory variables defined as:

$$\sigma_i^2 = \sigma^2 X_i^{\beta} e^{v_i}$$

or, $$\log_e \sigma_i^2 = \log_e \sigma^2 + \beta \log_e X_i + v_i$$

Where v_i is homoscedastic error term.

However, since σ_i^2 is unknown Park had been suggested the use of r_i^2 in its place. If β comes out to be significant then heteroscedasticity is said to be present in the data.

Another variant of Bartlett's test

Another variant of Bartlett's test is based on the likelihood ratio test statistic

$$u = \sum_{i-1}^{m} \left(\frac{s_i^2}{s^2} \right)^{n_i/2}$$

where

$$s_i^2 = \frac{1}{n_i} \sum_{j-1}^{n_i} \left(y_{ij} - \overline{y}_i \right)^2 , \; i = 1, 2, \ldots, m; \; j = 1, 2, \ldots, n_i$$

$$s^2 = \frac{1}{n} \sum_{i-1}^{m} n_i s_i^2$$

$$n = \sum_{i-1}^{m} n_i \; .$$

To obtain an unbiased test and modification of $-2 \ln u$ which is a closer approximation to χ^2_{m-1} under H_0, Bartlett test replaces n_i by $(n_i - 1)$ and divide by a scalar constant. This leads to the statistic

$$M = \frac{(n-m) \log \hat{\sigma}^2 - \sum_{i-1}^{m} (n_i - 1) \log \hat{\sigma}_i^2}{1 + \frac{1}{3(m-1)} \left[\sum_{i-1}^{m} \left(\frac{1}{n_i - 1} \right) - \frac{1}{n-m} \right]}$$

which has a χ^2 distribution with $(m-1)$ degrees of freedom under H_0 and

$$\hat{\sigma}_i^2 = \frac{1}{n-1} \sum_{j-1}^{n_i} (y_{ij} - \overline{y}_i)^2$$

$$\hat{\sigma}^2 = \frac{1}{n-m} \sum_{i-1}^{m} (n_i - 1) \hat{\sigma}_i^2 .$$

In experimental sciences, it is easier to get replicated data, and this test can be easily applied. In real-life applications, it is challenging to get replicated data, and this test may not be applied. This difficulty is overcome in Breusch Pagan test.

Goldfeld Quandt test

This test is based on the assumption that σ_i^2 is positively related to X_{ij}, i.e., one of the explanatory variables explains the heteroskedasticity in the model. Let j^{th} explanatory variable explains the heteroskedasticity, so

$$\sigma_i^2 \propto X_{ij}$$

or $\sigma_i^2 = \sigma^2 X_{ij}$.

The test procedure is as follows:

1. Rank the observations according to the decreasing order of X_j.
2. Split the observations into two equal parts leaving c observations in the middle.

   So each part contains $\frac{n-c}{2}$ observations provided $\frac{n-c}{2} > k$.
3. Run two separate regression in the two parts using OLS and obtain the residual sum of squares SS_{res1} and SS_{res2}.
4. The test statistic is

   $$F_0 = \frac{SS_{res2}}{SS_{res1}}$$

   which follows $F-$distribution, i.e., $F\left(\frac{n-c}{2}-k, \frac{n-c}{2}-k\right)$ when H_0 true.
5. The decision rule is to reject H_0 whenever $F_0 > F_{1-\alpha}\left(\frac{n-c}{2}-k, \frac{n-c}{2}-k\right)$.

- This test is a simple test, but it is based on the assumption that one of the explanatory variables helps in determining the heteroskedasticity.
- Then the test is an exact finite sample test.
- The only difficulty in this test is that the choice of c is not obvious. If a large value of c is chosen, then it reduces the degrees of freedom $\frac{n-c}{2}-k$, and the condition $\frac{n-c}{2} > k$ may be violated.

BREUSCH–PAGAN–GODFREY TEST

- The goldfled–Quandt test depend upon the correct selection of the value of c and the correct explanatory variable according to which observations are to be arranged.
- To overcome this difficulty Breusch–Pagan–Godfrey defined another test.

BREUSCH–PAGAN–GODFREY TEST

1. Fit the regression model $Y = X\beta + \epsilon$ using OLS method and obtain the residuals $r_1, r_2, \ldots, r_n$.
2. Obtain the estimate of $\sigma 2$ using: $\tilde{\sigma}^2 = \sum_{i=1}^{n} r_i^2/n$
3. Construct the variable p_i using: $p_i = r_i^2/\tilde{\sigma}^2$
4. Regress the p_i over the Z_j's. Some or all X_j's may serve as Z_j's.

$$p_i = \alpha_0 + \alpha_1 Z_{1i} + \alpha_2 Z_{2i} + \cdots + \alpha_m Z_{mi} + v_i$$

Where vi is the homoscedastic error term.

5. Obtain the Explained (Regression) Sum of Square (ESS) and define: $\Theta = ESS/2$
6. If Θ exceeds the $\chi^2_{(m-1)}$ at given level of significance then heteroscedasticity is said to be present.

REMEDIAL MEASURES

As explained earlier, heteroscedasticity does not destroy the unbiasedness and consistency properties of the OLS estimators, but they are no longer efficient. This lack of efficiency makes the usual hypothesis testing procedure of dubious value. Therefore, remedial measures may be called for. There are two approaches to correct the problem of heteroscedasticity they are

A) when $\sigma_{u_i}^2$ is known and

B) when $\sigma_{u_i}^2$ is not known.

(B) When σ_u^2 is not known :

Method I : Data Transformation Method

When σ_u^2 is not known there is a method to obtain consistent estimators of the variances and covariances of OLS estimators even if there is a problem of heteroscedasticity.

To explain, let us take the model

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Let us consider several assumption about the pattern of heteroscedasting proportional.

Method II : Log Transformation : A log transformation such as

$$\ln Y_i = \beta_0 + \beta_1 \ln X_i + u_i$$

Very often it reduces heteroscedasticity when compared with the regression $Y_i = \beta_0 + \beta_1 X_i + u_i$, because log transformation compresses the scale in which variable are measured.

An advantage of the log transformation model is that the slope coefficient β_1 measures the elasticity of Y with respect to X.

To conclude, the above discussion of the remedial measures, which of the transformation discussed will work depends on the nature of the problem and the severity of heteroscedasticity.

Method II : Weighted Least Square Method

When $\sigma_{u_i}^2$ is known, the most useful method of dealing with heteroscedasticity is by means of weighted least square

Consider a two variable model

$$Y_i = \beta_0 + \beta_1 X_i + U_i$$

Usual or unweighted least square method consists in minimising residual sum of square $\Sigma e_i^2 = \Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$ with respect to unknowns. This method gives same weight to each e_i^2. For example points 'A', 'B' and C will have same weight in computing Σe_i^2 (see Figure 7.5). In this case the e_i^2 associated with point 'C' will dominate the residual sum of square.

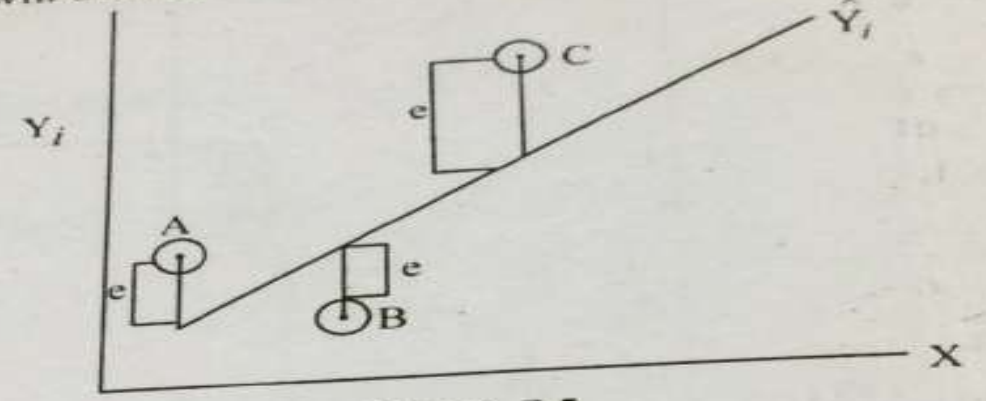


Figure 7.5

Method of weighted least square take into account the extreme points such as 'C' by minimising not usual or unweighted sum of squares but the following residual sum of square.

$$\text{Min } \Sigma e_i^2 = \Sigma W_i (Y_i - \hat{\beta}_0^* - \hat{\beta}_1^* X_i)^2$$

where W_i = weights

$\hat{\beta}_0^*$ and $\hat{\beta}_1^*$ are weighted least square estimators. W_i are chosen in such