LINEAR REGRESSION ANALYSIS

MODULE – X

Lecture - 32

Heteroskedasticity

Dr. Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

In the multiple regression model

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

it is assumed that

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

$$i.e., Var(\varepsilon_i^2) = \sigma^2,\ Cov(\varepsilon_i \varepsilon_j) = 0,\ i \neq j = 1, 2, ..., n.$$

In this case, the diagonal elements of covariance matrix of ε are same indicating that the variance of each ε_i is same and off-diagonal elements of 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 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

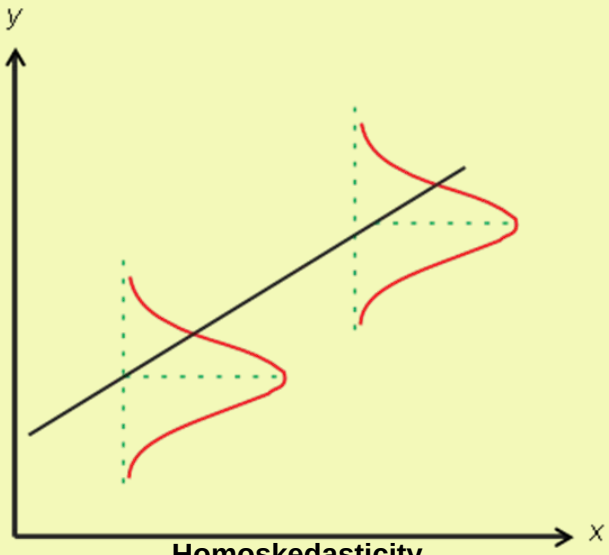
$$Var(\varepsilon_i) = \sigma_i^2,\ i = 1, 2, ..., n$$

and disturbances are pairwise uncorrelated.

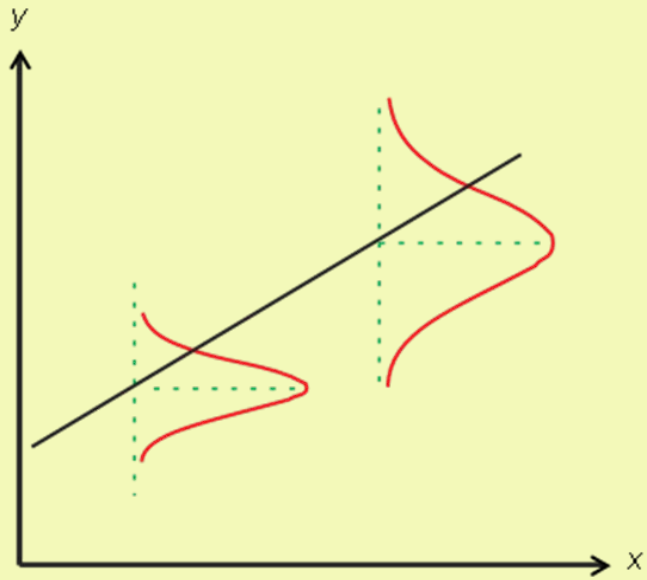
The covariance matrix of disturbances is

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

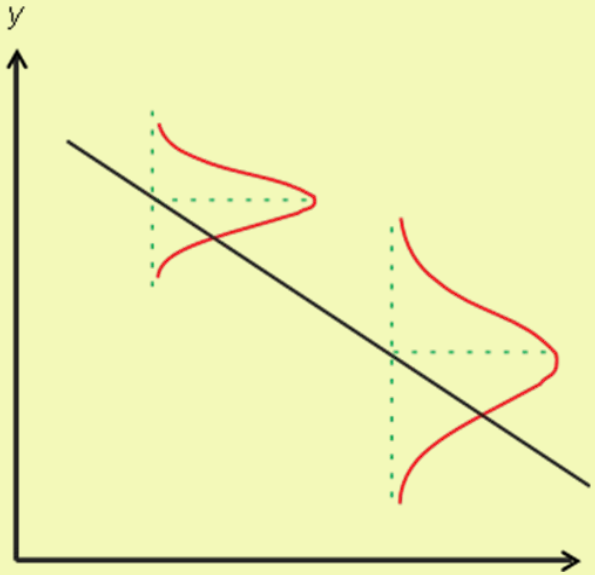
Graphically, the following pictures depict homoskedasticity and heteroskedasticity.



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 variation in expenditure on food increases because the choice and varieties in food increases, in general upto 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 same irrespective of the income of the person. This may not generally be a correct assumption in real situations. Rather the consumption pattern changes and hence the variance of *y* and so the variances of disturbances will not remain constant. In general, it will be increasing as income increases.

In another example, suppose in a simple linear regression model, *x* denotes the number of hours of practice for typing and *y* denotes the number of typing errors per page. It is expected that the number of typing mistakes per page decreases as the person practices more. The homoskedastic disturbances assumption implies that the number of errors per page will remain same irrespective of the number of hours of typing practice which may not be true is practice.

Possible reasons for heteroskedasticity

There are various reasons due to which the heteroskedasticity is introduced in the data. Some of them are as follows:

1. The nature of phenomenon under study may have an increasing or decreasing trend. For example, the variation in consumption pattern on food increases as income increases, similarly the variation in number of typing mistakes decreases as the number of hours of typing practice increases.

2. Sometimes the observations are in the form of averages and this introduces the heteroskedasticity in the model. For example, it is easier to collect data on the expenditure on clothes for the whole family rather than on a particular family member. Suppose in a simple linear regression model

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}, i = 1,2,...,n,\ j = 1,2,...,m_i\ ,$$

y_{ij} denotes the expenditure on cloth for the j^{th} family having m_j members and x_{ij} denotes the age of the i^{th} person in j^{th} family. It is difficult to record data for individual family member but it is easier to get data for the whole family. So y_{ij}'s are known collectively.

Then instead of per member expenditure, we find the data on average expenditure for each family member as

$$\overline{y}_i = \frac{1}{m_j}\sum_{j=1}^{m_j} y_{ij}$$

and the model becomes

$$\overline{y}_i = \beta_0 + \beta_1 \overline{x}_i + \overline{\varepsilon}_i.$$

It we assume $E(\varepsilon_{ij}) = 0, Var(\varepsilon_{ij}) = \sigma^2$, then $E(\overline{\varepsilon}_i) = 0,\ Var(\overline{\varepsilon}_i) = \frac{\sigma^2}{m_j}$ which indicates that the resultant variance of disturbances does not remain constant but depends on the number of members in a family m_j. So heteroskedasticity enters in the data. The variance will remain constant only when all m_j 's are same.

3. Sometimes the theoretical considerations introduces the heteroskedasticity in the data. For example, suppose in the simple linear model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, 2, ..., n$$

y_i denotes the yield of rice and x_i denotes the quantity of fertilizer in an agricultural experiment. It is observed that when the quantity of fertilizer increases, then yield increases. In fact, initially the yield increases when quantity of fertilizer increases. Gradually, the rate of increase slows down and if fertilizer is increased further, the crop burns. So notice that β_1 changes with different levels of fertilizer. In such cases, when β_1 changes, a possible way is to express it as a random variable with constant mean $\overline{\beta}_1$ and constant variance θ^2 like

$$\beta_{1i} = \overline{\beta}_1 + v_i,\ i = 1, 2, ..., n$$

with

$$E(v_i) = 0, Var(v_i) = \theta^2, E(\varepsilon_i v_i) = 0.$$

So the complete model becomes

$$\begin{aligned} y_i &= \beta_0 + \beta_{1i} x_i + \varepsilon_i \\ \beta_{1i} &= \overline{\beta}_1 + v_i \\ \Rightarrow y_i &= \beta_0 + \overline{\beta}_1 x_i + (\varepsilon_i + x_i v_i) \\ &= \beta_0 + \overline{\beta}_1 x_i + w_i \end{aligned}$$

where

$$w_i = \varepsilon_i + x_i v_i$$

is like a new random error component.

So

$$E(w_i) = 0$$

$$Var(w_i) = E(w_i^2)$$

$$= E(\varepsilon_i^2) + x_i^2 E(v_i^2) + 2x_i E(\varepsilon_i v_i)$$

$$= \sigma^2 + x_i^2\theta^2 + 0$$

$$= \sigma^2 + x_i^2\theta^2.$$

So variance depends on *i* and thus heteroskedasticity is introduced in the model. Note that we assume homoskedastic disturbances for the model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ \beta_{1i} = \overline{\beta}_1 + v_i$$

but finally ends up with heteroskedastic disturbances. This is due to theoretical considerations.

4. The skewness in the distribution of one or more explanatory variables in the model also causes heteroskedasticity in the model.
5. The incorrect data transformations and incorrect functional form of the model can also give rise to the heteroskedasticity problem.
6. Sometimes the study variable may have distribution such as Binomial or Poisson where variances are the functions of means. In such cases, when the mean varies, then the variance also varies.

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 literature for testing the presence of heteroskedasticity, 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.

We discuss the Bartlett's test.

Bartlett's test

It is a test for testing the null hypothesis

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

This hypothesis is termed as the **hypothesis of homoskedasticity**. This test can be used only when replicated data is available.

Since in the model

$$y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \varepsilon_i,\ E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma_i^2,\ i = 1, 2, ..., n,$$

only one observation y_i is available to find σ_i^2, so the usual tests can not be applied. This problem can be overcome if replicated data is available. So consider the model of the form

$$y_i^* = X_i \beta + \varepsilon_i^*$$

where y_i^* is a m_i x 1 vector, X_i is m_i x k matrix, β is k x 1 vector and ε_i^* is m_i x 1 vector. So replicated data is now available for every y_i^* in the following way:

$y_1^* = X_1 \beta + \varepsilon_1^*$ consists of m_1 observations

$y_2^* = X_2 \beta + \varepsilon_2^*$ consists of m_2 observations

$\vdots$

$y_n^* = X_n \beta + \varepsilon_n^*$ consists of m_n observations.

All the individual model can be written as

$$\begin{pmatrix} y_1^* \\ y_2^* \\ \vdots \\ y_n^* \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_1^* \\ \varepsilon_2^* \\ \vdots \\ \varepsilon_n^* \end{pmatrix}$$

or

$$y^* = X\beta + \varepsilon^*$$

where y^* is a vector of order $\left(\sum_{i=1}^{n} m_i\right) \times 1$, X is $\left(\sum_{i=1}^{n} m_i\right) \times k$ matrix, β is $k \times 1$ vector and ε^* is $\left(\sum_{i=1}^{n} m_i\right) \times 1$ vector.

Application of OLS to this model yields

$$\hat{\beta} = (X'X)^{-1} X' y^*$$

and obtain the residual vector

$$e_i^* = y_i^* - X_i \hat{\beta}.$$

Based on this, obtain

$$s_i^2 = \frac{1}{m_i - k} e_i^{*\prime} e_i^*$$

$$s^2 = \frac{\sum_{i=1}^{n} (m_i - k) s_i^2}{\sum_{i=1}^{n} (m_i - k)}.$$

Now apply the Bartlett's test as

$$\chi^2 = \frac{1}{C}\sum_{i=1}^{n}(m_i - k)\log\frac{s^2}{s_i^2}$$

which has asymptotic χ^2 – distribution with $(n-1)$ degrees of freedom where

$$C = 1 + \frac{1}{3(n-1)}\left[\sum_{i=1}^{n}\left(\frac{1}{m_i - k}\right) - \frac{1}{\sum_{i=1}^{n}(m_i - k)}\right].$$

Another variant of Bartlett's test

Another variant of Bartlett's test is based on likelihood ratio test statistic. If there are *m* independent normal random samples where there are n_i observations in the i^{th} sample. Then the likelihood ratio test statistic for testing

$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_i^2 = ... = \sigma_m^2$ is

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

where $\bar{y}_i$ and s_i^2 are the sample mean and sample variance respectively of the i^{th} sample.

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

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

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

To obtain an unbiased test and a 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

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

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

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