variables.

HETEROSCEDASTICITY :

One of the important assumption of the OLS method is that the disturbance term follows normal distribution with zero mean and constant variance. i.e., $U_i \sim N(0, \sigma_u^2)$. This means that the variance of the disturbance term remains same for all the values of the explanatory variables. Symbolically —

$$\text{var}(U) = E(U^2) = \sigma_u^2$$

This assumption is known as homoscedasticity. But var(U) may not remain the same for all the observations. Thus, we may get $\text{var}(U) \neq \sigma_u^2$. This condition of non-homogeneity of variance is called as heteroscedasticity.

It should be noted that heteroscedasticity, generally arises more in cross-sectional data than in time series data. In cross-sectional data, we are generally concern with the members of population at a given point of time. These members may be of different sizes which leads to the rise of the problem of heteroscedasticity. On the other hand, in time series data, the variable tends to be similar orders of magnitude because one generally collects data for the same entity over a period of time.

CAUSES OF HETEROSCEDASTICITY :

The Different causes of heteroscedasticity are as follows —

(i) As the value of the independent variables increases, there may be an increase in the disturbance term.

(ii) Measurement errors can also be a cause of heteroscedasticity as because some respondents might provide more accurate response than others.

10-6-22

(iii) The disturbance term is introduced in the regression model because of the omitted variables. In case of omitted variables also the disturbance term may vary with explanatory variables.

(iv) Heteroscedasticity may arise due to incorrect specification of the regression model. The non-inclusion of some important variables may be the reason of the non-constant variance of the disturbance term.

(v) Heteroscedasticity may also occur if there are sub-population differences.

CONSEQUENCES :

Consequences of heteroscedasticity are —

(1) The OLS estimators will have no bias :

Give that the model is —

$$Y_i = \beta X_i + U_i \quad \text{with}$$

$$E(U_i) = 0$$

$$E(U_i^2) = \sigma_u^2$$

~~$E(U_i U_j) =$~~

$$E(U_i U_j) = 0, \quad i \neq j$$

Now, we know that —

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$= \frac{\sum x_i (\beta X_i + U_i)}{\sum x_i^2} \quad \text{putting the value of } Y_i = \beta X_i + U_i$$

$$= \frac{\beta \sum x_i X_i + \sum x_i U_i}{\sum x_i^2}$$

$$= \frac{\beta \sum x_i^2 + \sum x_i U_i}{\sum x_i^2}$$

$$= \beta + \frac{\sum x_i U_i}{\sum x_i^2}$$

taking expectation on both sides, we get —

$$E(\hat{\beta}) = \beta + \frac{\sum x_i E(U_i)}{\sum x_i^2}$$

$$= \beta + 0 \quad | \because E(U_i) = 0$$

$$= \beta$$

Thus, $\hat{\beta}$ is an unbiased estimator of β.

① Variance of OLS coefficients will be incorrect :

We know,

$$\text{var}(\hat{\beta}) = E\left\{\hat{\beta} - E(\hat{\beta})\right\}^2$$

$$= E\left(\beta + \frac{\sum x_i u_i}{\sum x_i^2} - \beta\right)^2$$

$$= E\left(\frac{\sum x_i u_i}{\sum x_i^2}\right)^2$$

$$= \frac{1}{(\sum x_i^2)^2} \cdot E(\sum x_i u_i)^2$$

$$= \frac{1}{(\sum x_i^2)^2} \cdot E\left(x_1^2 u_1^2 + x_2^2 u_2^2 + \ldots\ldots + 2x_1 x_2 u_1 u_2 + \ldots\right)$$

$$= \frac{1}{(\sum x_i^2)^2}\left\{x_1^2 E(u_1^2) + x_2^2 E(u_2^2) + \ldots\ldots + 2x_1 x_2 E(u_1 u_2) + 2x_1 x_3 E(u_1 u_3) + \ldots\right\}$$

$$= \frac{1}{(\sum x_i^2)^2}\left(x_1^2 \sigma_u^2 + x_2^2 \sigma_u^2 + x_3^2 \sigma_u^2 + \ldots 0 + 0 \ldots 0\right)$$

$$= \frac{1}{(\sum x_i^2)^2} \cdot \sigma_u^2 \left(x_1^2 + x_2^2 + \ldots + x_n^2\right)$$

$$= \frac{1}{(\sum x_i^2)^2} \cdot \sigma_u^2 \cdot \sum_{i=1}^{n} x_i^2$$

$$= \frac{\sigma_u^2 \cdot \sum x_i^2}{(\sum x_i^2)^2}$$

$$= \sigma_u^2 \cdot \frac{1}{\sum x_i^2}$$

substituting the value of $\hat{\beta} = \beta + \frac{\sum x_i u_i}{\sum x_i^2}$ and $E(\hat{\beta}) = \beta$

[x_i include because of heteroscedasticity

Suppose, $E(u_i^2) = k_i \sigma_u^2$

~~$\text{var}(\hat{\beta}_2) = \frac{1}{(\sum x_i^2)^2} \sum x_i^2 u_i$~~

$$\therefore \text{var}(\hat{\beta}) = \frac{1}{(\sum x_i^2)^2}\left(x_1^2 \cdot k_1 \sigma_u^2 + x_2^2 k_2 \sigma_u^2 + \ldots\ldots + x_n^2 k_n \sigma_u^2\right)$$

$$= \frac{1}{(\sum x_i^2)^2} \cdot \sigma_u^2 \left(x_1^2 k_1 + x_2^2 k_2 + \ldots\ldots + x_n^2 k_n\right)$$

$$= \frac{\sigma_u^2}{(\sum x_i^2)^2} \cdot \sum x_i^2 k_i$$

$$= \frac{\sigma_u^2}{\sum x_i^2} \cdot \frac{\sum k_i x_i^2}{\sum x_i^2}$$

if x_i^2 and k_i are positively correlated and if $\frac{\sum k_i x_i^2}{\sum x_i^2}$ is greater than unity, than the var$(\hat{\beta})$ under heteroscedasticity will be greater than the variance under homoscedasticity.

14-6-22

DETECTION OF HETEROSCEDASTICITY :

The various method that can be adopt to detect heteroscedasticity are —

1. INFORMAL METHODS :

(a) Nature of the problem :

Very often the nature of the problem under consideration suggests whether heteroscedasticity is likely to be encountered. As a matter of fact, in cross-sectional data involving heterogeneous units, heteroscedasticity may be the rule rather than the exception. Thus, in a cross-sectional analysis involving the investment expenditure in relation to sales, rate of interest etc., heteroscedasticity is generally expected if small-, medium-, and large size firms are sampled together.

(b) Graphical Method :

If there is no a priori or empirical information about the nature of heteroscedasticity, in practice one can do the regression analysis on the assumption that there is no heteroscedasticity and then do a postmortem examination of the residual squared $\hat{u}_i^2$ to see if they exhibit any systematic pattern. An example examination of the $\hat{u}_i^2$ may reveal patterns such as those shown in fig.

In fig (a), we see that there is no systematic pattern between the two variables, suggesting that perhaps no heteroscedasticity is present in the data. fig (b) to (e), however, exhibit definite patterns. For instance, fig (c) suggest a linear relationship, whereas fig (d) and (e) indicate a quadratic relationship between $\hat{u}_i^2$ and $\hat{Y}_i$.

14-6-22

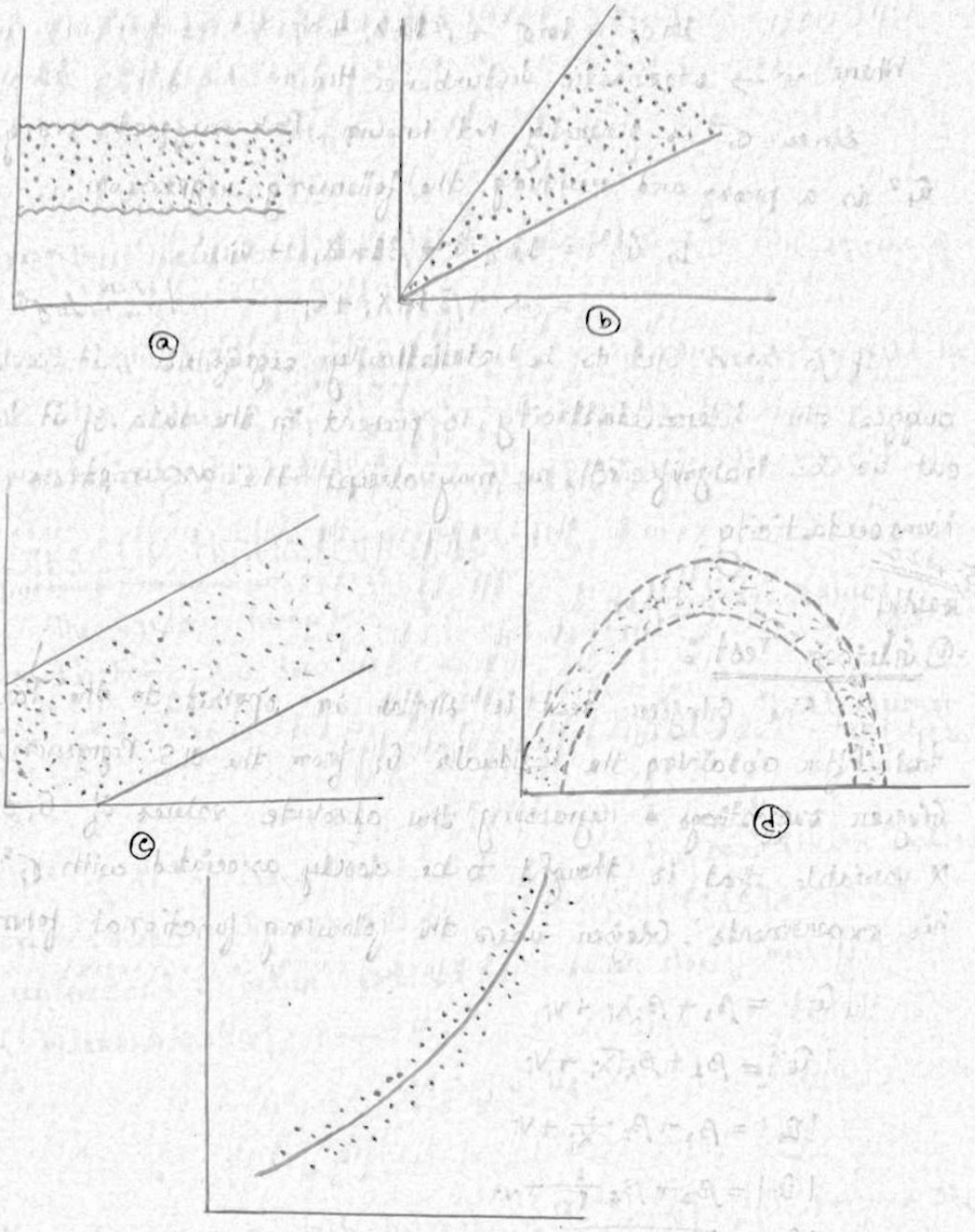

Fig: Hypothetical Patterns of estimated squared residuals.

§2. FORMAL METHODS:

(a) Park Test:

Park formalize the graphical method by suggesting that σ_i^2 is some function of the explanatory variable X_i. The functional form he suggests is —

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

$$\ln \sigma_i^2 = \ln \sigma^2 + \beta \ln X_i + v_i$$

Where $v_i \rightarrow$ stochastic disturbance term

Since σ_i^2 is generally not known. Park suggests using $\hat{u}_i^2$ as a proxy and running the following regression:

$$\ln \hat{u}_i^2 = \ln \sigma^2 + \beta \ln X_i + v_i$$
$$= \alpha + \beta \ln X_i + v_i \quad \Big| \text{ where, } \alpha \rightarrow \ln \sigma^2$$

If β turns out to be statistically significant, it would suggest the heteroscedasticity is present in the data. If it turns out to be insignificant, we may accept the assumption of homoscedasticity.

15-6-22

~~Goldfel~~

(b) Glejser Test:

The Glejser test is similar in spirit to the Park test. After obtaining the residuals $\hat{u}_i$ from the OLS Regression, Glejser suggests regressing the absolute values of $\hat{u}_i$ on the X variable that is thought to be closely associated with σ_i^2. In his experiments, Glejser uses the following functional forms —

$$|\hat{u}_i| = \beta_1 + \beta_2 X_i + v_i$$
$$|\hat{u}_i| = \beta_1 + \beta_2 \sqrt{X_i} + v_i$$
$$|\hat{u}_i| = \beta_1 + \beta_2 \cdot \frac{1}{X_i} + v_i$$
$$|\hat{u}_i| = \beta_1 + \beta_2 \frac{1}{\sqrt{X_i}} + v_i$$
$$|\hat{u}_i| = \sqrt{\beta_1 + \beta_2 X_i} + v_i$$
$$|\hat{u}_i| = \sqrt{\beta_1 + \beta_2 X_i^2} + v_i$$

Where, v_i is the error term.

(c) Goldfeld - Quandt Test:

This popular method is applicable if one assumes that the heteroscedasticity variance, σ_i^2, is positively related to one of the explanatory variables in the regression model. For simplicity, consider the usual two-variable model —

15-6-22

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

Suppose σ_i^2 is positively related to X_i as

$$\sigma_i^2 = \sigma^2 X_i^2$$

where $\sigma^2 \rightarrow$ constant

Assumption postulates that σ_i^2 is proportional to the square of the X variable.

If equation is appropriate, it would mean σ_i^2 would be larger than values of X_i. If that turns out to be case, heteroscedasticity is most likely to be present in the model.

REMEDIES OF HETEROSCEDASTICITY :

The various measures that can be adopted to correct heteroscedasticity are discussed below —

a. When σ_i^2 is known : The Method of Generalized Least Squares (GLS) Estimator

If the variance is non constant, then the best linear unbiased estimator (BLUE) is the generalised least squares (GLS) estimator. To understand it better, ~~eens~~ let us consider the following model with heteroscedasticity —

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad ; \quad i = 1, 2, 3, \ldots n$$

$$\text{var}(u_i) = \sigma_i^2$$

Now, to derive the GLS estimator of the model, first find the error variance σ_i^2 and take square root of the error variance σ_i. This gives the standard deviation of the error. Then divide every term on both sides of the regression equation by this standard deviation of the error σ_i. This gives the following regression equation with a new transformed model

$$\frac{Y_i}{\sigma_i} = \beta_1 \cdot \frac{1}{\sigma_i} + \beta_2 \cdot \frac{X_{2i}}{\sigma_i} + \beta_3 \frac{X_{3i}}{\sigma_i} + u_i/\sigma_i$$

$$\Rightarrow Y_i^* = \beta_1 X_{1i}^* + \beta_2 X_{2i}^* + \beta_3 X_{3i}^* + u_i^* \quad ; i = 1, 2, 3, \ldots n$$

Here, $Y_i^* = \frac{Y_i}{\sigma_i}$; $X_{1i}^* = \frac{1}{\sigma_i}$; $X_{2i}^* = \frac{X_{2i}}{\sigma_i}$; $X_{3i}^* = \frac{X_{3i}}{\sigma_i}$; $u_i^* = \frac{u_i}{\sigma_i}$

The error variance for the transformed model is $\text{Var}(u_i) = \text{var}\left(\frac{u_i}{\sigma_i}\right)$

$= \frac{\text{var}(u_i)}{\sigma_i^2}$

$= 1$,

which is constant

That why heteroscedasticity can be remove from the model.

18-6-22

b. When σ_i^2 is unknown : The method of feasible generalised least square estimator (FGLS)

In this method, we use the sample data to obtain an estimate of σ and then applying the GLS method using the estimates of σ. When we do that we have a different estimator called the feasible generalised least square estimator.

For eg, let us consider the following general linear regression model with heteroscedasticity

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + U_i$$

$$\text{var}(U_i) = \sigma_i^2 \qquad i = 1, 2, 3, \ldots n.$$

Further assume that the error variance is ~~regres~~ a linear function of X_2 and X_3

$$\text{var}(u_i) = \sigma^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} \qquad i = 1, 2, 3, \ldots n$$

To obtain FGLS estimates of the parameters β_1, β_2 and β_3, following steps should be taken.

1. Run the original OLS regression and obtain the residual $\hat{u}_i$.
2. Square these residuals $\hat{u}_i^2$, and regress it on a constant, X_{i2} and X_{i3}, by using OLS.
3. Then use the estimates of α_1, α_2 and α_3 in order to calculate the predicted values of $\hat{\sigma}_i^2$.
4. Since we can't have a negative variance, any predicted value that is non-positive must be replaced with the squared residual for that observation.

18-6-22

5. Take the square root of the estimate of the error variance, $\hat{\sigma}_i$ for each observation.

6. Find the weight $w_i = 1/\hat{\sigma}_i$ for each observation.

7. Multiply Y_i, X_{i2} and X_{i3} for each observation by this weight

8. Finally regress $w_i Y_i$ on $w_i X_{i2}$ and $w_i X_{i3}$ using OLS.