Heteroscedasticity

Major Points to Consider:

- Nature of the problem of heteroscedasticity and its consequences
- Reasons for heteroscedasticity
- Testing for presence of heteroscedasticity
- Remedial measures

➢ **Nature of the Problem:**

The problem of heteroscedasticity arises when the variance of the random disturbance term is not constant across the observations.

Consider the model: $Y_i = \alpha + \beta X_i + u_i$ It is assumed that,

$E(u_i|X_i) = 0$ for all i (assumption of conditional zero mean); $E(u_i^2|X_i) = \sigma^2$ for all i (assumption of conditional constant variance - homoscedasticity; $E(u_i u_j|X_i) = 0$ for all $i \neq j$ (assumption of conditional independence - no autocorrelation)

If $E(u_i^2) = \sigma_i^2$ - if the variance varies across observations, there is problem of heteroscedasticity.

The problem is likely to be more common in case of cross-sectional than time-series data.

Example: Let us take the example of the relationships between consumption expenditure and income at household level. Since the households with low income do not have much flexibility in spending, their consumption expenditure is unlikely to vary considerably. On the other hand, there may be large flexibility in spending for the high-income households. This means that the deviation of actual consumption expenditure from the mean may be different depending on the level of income. It is very likely that the higher income households will have larger deviation around the mean than the lower income households leading to the problem of **heteroscedasticity.**

However, while heteroscedasticity is generally observed for the cross-sectional data, the problem can occur in time-series data as well, particularly when the dependent variable changes significantly from the beginning to the end of the series. For example, if we model the sales of a newly introduced electronic product from the first sales to the present, the number of units sold will be vastly different. Further, while modeling time-series data, if the measurement errors change over time, heteroscedasticity can be present because the regression model includes measurement error in the random disturbance term. For example, if the measurement errors decrease over time as better methods are introduced, one may expect the error variance to decrease.

Pure versus Impure heteroscedasticity

One can categorize heteroscedasticity into ***pure*** and ***impure*** types. Pure heteroscedasticity refers to the cases where the correct model (in respect of functional form or the choice of variables) is specified and yet the variance of the random disturbance term is not constant. On the other hand, impure heteroscedasticity refers to the cases where the model is incorrectly specified and the variance of the random disturbance term is not constant. For example, when an important variable is left out in a model, the omitted effect is absorbed in the random disturbance term. If the effect of the omitted variable varies throughout the observed range of data, it can show the signs of heteroscedasticity. In such cases, the problem is not of pure heteroscedasticity but of ***specification error***. Hence, it is important to determine whether there is pure or impure heteroscedasticity because the solutions are different. If the problem is of impure form (i.e., of specification error), one needs to identify the important variable(s) that have been left out of the model or the correct functional form and refit the model.

- **Consequences of Heteroscedasticity:**

In the presence of heteroscedasticity, the OLS estimators of the coefficients (i.e., alpha and beta) will be unbiased, consistent and asymptotically normally distributed, but not efficient, i.e., they are not ***BLUE***. If the presence of heteroscedasticity is ignored and the OLS estimators are used for the regression coefficients, the properties of ***unbiasedness*** and ***consistency*** still are not violated, but these OLS estimators are ***no longer efficient***. It is possible to find an alternative linear and unbiased estimator of these coefficients that have lower variances than the OLS estimators. Further, the estimated variances and covariance of the OLS estimators would be ***biased*** and ***inconsistent***. Hence, the usual statistical tests of hypotheses will no longer be valid.

While heteroscedasticity does not cause bias in the OLS estimators of the coefficients, it does make them less precise. Lower precision increases the likelihood that the coefficient estimates are further away from the correct population value. Heteroscedasticity tends to result in the p-values that are smaller than what they should be. This is so because heteroscedasticity increases the variance of the estimators of the coefficients, but the OLS procedure does not detect this increase. Consequently, the OLS method calculates the t-statistics and the F-statistics using the underestimated variances. This problem can lead to the conclusion that a model is statistically significant when it is actually not so.

- **Reasons for Presence of Heteroscedasticity:**

- Presence of outliers in the sample (particularly when it is of small size)
- Specification bias
  - Exclusion of variable(s)
  - Inappropriate Functional form

This causes impure heteroscedasticity or specification error

- Skewness in the distribution of explanatory variable(s)
- Incorrect data transformation
- Error-learning behaviour of the researcher

- Changes in data collection techniques
- Pooling of samples with structural differences

➢ **Testing for Presence of Heteroscedasticity:**

(a) Graphical Tests-Scatter Diagram of Squared Residuals

Before carrying out any formal statistical tests of heteroscedasticity, it is useful to examine the patterns of the residuals visually. The squares of the residuals obtained by applying the OLS method of estimation to the model $Y_i = \alpha + \beta X_i + u_i$ are to be plotted against the independent variable that is suspected to cause heteroscedasticity. If more than one independent variables are suspected to cause the problem of heteroscedasticity (in case of a multiple regression model), the squares of the residuals are to be plotted against each of these independent variables. One may also plot the squares of the residuals against the estimated values of the dependent variables.

If it appears that the residuals are roughly of the same size for all the values of the independent variable(s), it is generally safe to assume that heteroscedasticity is not severe enough. On the other hand, if the scatter plot shows a particular pattern, it signals the presence of the problem of heteroscedasticity. If the plot shows some uneven envelope of residuals, so that the width of the envelope is considerably larger for some values of the independent variable(s) than for others, it signals the presence of heteroscedasticity. However, this is not a formal test of heteroscedasticity. It only suggests whether heteroscedasticity may exist. Hence, one should also carry out formal tests to confirm presence of the same statistically.

(b) Formal Statistical Tests for Heteroscedasticity

1. Park Test

Model: $Y_i = \alpha_1 + \alpha_2 X_i + u_i$

Assumption: $\sigma_i^2 = \sigma^2 X_i^{\beta} e^{v_i}$ (this means that the variance of the random disturbance term varies with the independent variable X)

<u>Steps to be followed:</u>

- Apply the method of OLS to estimate the model: $Y_i = \alpha_1 + \alpha_2 X_i + u_i$ and get $\hat{u}_i^2$
- Use $\hat{u}_i^2$ as a proxy for σ_i^2
- Estimate the model: $\ln(\hat{u}_i^2) = \ln(\sigma^2) + \beta \ln X_i + v_i$
- Null Hypothesis: $\beta = 0$
- If β is statistically significant, the problem of heteroscedasticity is present.

<u>Limitations of the Park Test:</u>

- The random disturbance term in the auxiliary regression model (**v_i**) may not satisfy the OLS assumptions and may itself have heteroscedasticity problem.
- The test is based on the assumption of a particular form of heteroscedasticity - $\sigma_i^2 = \sigma^2 X_i^{\beta} e^{v_i}$

2. Glejser Test

This test is similar to the Park Test. However, the Glejser test suggests regressing $|\hat{u}_i|$ on the explanatory variable that is closely associated with σ_i^2.

Model: $Y_i = \alpha_1 + \alpha_2 X_i + u_i$

Assumption: The test is based on the following functional forms of heteroscedasticity:

(i) $|\hat{u}_i| = \theta_1 + \theta_2 X_i + v_i$; (ii) $|\hat{u}_i| = \theta_1 + \theta_2 \sqrt{X_i} + v_i$; (iii) $|\hat{u}_i| = \theta_1 + \theta_2 \frac{1}{X_i} + v_i$;

(iv) $|\hat{u}_i| = \theta_1 + \theta_2 \frac{1}{\sqrt{X_i}} + v_i$; (v) $|\hat{u}_i| = \sqrt{\theta_1 + \theta_2 X_i} + v_i$; (vi) $|\hat{u}_i| = \sqrt{\theta_1 + \theta_2 X_i^2} + v_i$

<u>Steps to be followed:</u>

- Apply the method of OLS to estimate the model: $Y_i = \alpha_1 + \alpha_2 X_i + u_i$ and get $|\hat{u}_i|$
- Use $|\hat{u}_i|$ as a proxy for σ_i^2 and estimate the auxiliary regression model.
- Null Hypothesis: $\theta_2 = 0$
- If θ_2 is statistically significant, the problem of heteroscedasticity is present.

<u>Limitations of the Glejser Test:</u>

- The random disturbance term in the auxiliary regression model may have non-zero mean and also be heteroscedastic.
- The functional forms specified in (v) and (vi) are non-linear in parameters. Hence, these functional forms cannot be estimated by applying the OLS method.

3. Goldfeld–Quandt Test

Model: $Y_i = \alpha + \beta X_i + u_i$

Assumption: $\sigma_i^2 = \sigma^2 X_i^2$ i.e., the heteroscedastic variance is positively related to the explanatory variable.

<u>Steps to be followed:</u>

- Arrange the observations in ascending order of X

- Omit **c** number of central observations and divide the remaining (n-c) observations into two groups with each being of (n-c)/2 observations.

- Fit separate regression models for each of the groups and obtain RSS_1 and RSS_2

- Each RSS has $\frac{n-c}{2}-k$ or $\frac{n-c-2k}{2}$ degrees of freedom.

- Compute the test statistic: $\lambda = \frac{RSS_2/df}{RSS_1/df} \sim F_{\left[\frac{n-c-2k}{2},\frac{n-c-2k}{2}\right]}$

- Null hypothesis: $\sigma_1^2 = \sigma_2^2$ (homoscedasticity)

- If the computed value of the F-Statistic is greater than its critical value at the chosen level of significance, the null hypothesis of homoscedasticity is rejected.

Limitations of Goldfeld-Quandt Test:

- Power of the test depends on and is sensitive to **c** (the number of central observations omitted)**.** Introducing ***c*** is intended to increase the power of the test. However, increase in ***c*** leads to lower degrees of freedom in the estimation with each sub-sample and this tends to lower the power of the test. Hence, there is a trade-off in choosing ***c***.

  Judge et al. suggest c = 4 if n = 30 (13%), whereas Goldfeld-Quandt suggest for **c** = 8 if n = 30 (25%). Overall no more than a third of the central observations should be dropped.

  One possible choice is: $c \approx \left[\frac{n}{3} - 2k\right]$

4. White's General Test for Heteroscedasticity

Model: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$

Steps to be followed:

- Estimate the regression model by applying the method of OLS and obtain $\hat{u}_i^2$

- Estimate the auxiliary regression model:

  $$\hat{u}_i^2 = \gamma + \delta_1 X_{1i} + \delta_2 X_{2i} + \delta_3 X_{1i}^{\ 2} + \delta_4 X_{2i}^{\ 2} + \delta_5 X_{1i} X_{2i} + v_i$$

- Calculate the test statistic $\lambda = nR^2 \sim \chi^2$ with degrees of freedom equal to the number of regressors in the auxiliary model (excluding the intercept), i.e. 5 here.

- Null hypothesis: $H_0 : \delta_1 = \delta_2 = .. = \delta_5 = 0$ i.e., there is no heteroscedasticity.

- If the null hypothesis is rejected, there is heteroscedasticity problem.

Important Points:

- White test is very general as it does not specify any specific form of heteroscedasticity.
- Due to its generality, the test may help in identifying other specification errors as well. If no cross-product terms are present in the auxiliary regression, it is a test of pure heteroscedasticity. On the other hand, if the cross-product terms are present, it is a test of both pure heteroscedasticity and specification error.

Limitations of the White's General Test:

- If one or more of the explanatory variables are dummy variables, one must be careful while specifying the auxiliary regression model. This is so because, in case of dummy variable(s), the higher order(s) of the variable(s) will have the same value. Hence, if both of them are included in the auxiliary regression, there will be perfect multicollinearity. Therefore, one should exclude the higher order(s) of the dummy variable(s) from the auxiliary regression.
- If there are a large number of explanatory variables in the original model, the number of explanatory variables in the auxiliary regression could even exceed the number of observations. In this case, one must exclude some of the explanatory variables from the auxiliary regression, but choice of the appropriate set of variables to be included is a critical task.
- The power of the test may be low in some cases (depending on specification of the auxiliary regression model).
- The test is non-constructive in that if the null hypothesis is rejected, the results of the test do not provide any guidance for the remedial measures.

5. Breusch-Pagan Lagrange Multiplier Test

The test statistic is easy to calculate and a variety of alternatives can be covered

Model: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i$

Alternative specification for auxiliary regressions:

(1) $\sigma_i^2 = \alpha_1 + \alpha_2 Z_{2i} + \alpha_3 Z_{3i} + ... + \alpha_p Z_{pi} + v_i$

(2) $\sigma_i = \alpha_1 + \alpha_2 Z_{2i} + \alpha_3 Z_{3i} + ... + \alpha_p Z_{pi} + v_i$

(3) $\ln(\sigma_i^2) = \alpha_1 + \alpha_2 Z_{2i} + \alpha_3 Z_{3i} + ... + \alpha_p Z_{pi} + v_i$

Here, all or some of X can serve as Z

Steps to be followed:

- Estimate the original model and obtain the estimated residuals
- Use $\hat{u}_i^2$, $|\hat{u}_i|$ and $\ln(\hat{u}_i^2)$ for estimating auxiliary regression models with respect to (1), (2) and (3) respectively.

- Compute the test statistic: $\lambda = nR^2 \sim \chi^2_{(p-1)})$
- Null Hypothesis: $H_0 : \alpha_2 = \alpha_3 = ... = \alpha_p = 0$
- When the null hypothesis is rejected, the estimated model suffers from the problem of heteroscedasticity.
- The more accurately one knows the exact causes of heteroscedasticity, better is the power of the test.

Alternative Approach (Commonly Followed):

- Assumption: $\sigma_i^2 = \alpha_1 + \alpha_2 Z_{2i} + \alpha_3 Z_{3i} + ... + \alpha_p Z_{pi} + v_i$
- Null Hypothesis: $H_0 : \alpha_2 = \alpha_3 = ... = \alpha_p = 0$
- Estimate the original model and obtain the estimated residuals.
- Compute $\tilde{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n}$ (the MLE estimator of σ^2)
- Compute $q_i = \frac{\hat{u}_i^2}{\tilde{\sigma}^2}$
- Estimate the auxiliary regression model: $q_i = \alpha_1 + \alpha_2 Z_{2i} + \alpha_3 Z_{3i} + ... + \alpha_p Z_{pi} + v_i$
- Obtain explained sum of squares (ESS) from the auxiliary regression.
- Compute the test statistic: $\lambda = \frac{ESS}{2} \sim \chi^2_{(p-1)})$
- If the null hypothesis is rejected, the estimated model suffers from the problem of heteroscedasticity.

Limitations of the Lagrange Multiplier Test:

- The test assumes that the error variance is a linear function of one or more of the explanatory variables. Thus, if heteroscedasticity exists, but the error variance is a non-linear function of the explanatory variable(s), this test will not be valid.

6. Spearman's Rank Correlation Test

Model: $Y_i = \alpha + \beta X_i + u_i$

Steps to be followed:

- Estimate the model and obtain the residuals $\hat{u}_i$
- Ignore the sign of $\hat{u}_i$ and rank $|\hat{u}_i|$ and X_i (or $\hat{Y}_i$) in ascending or descending order
- Compute Spearman's Rank Correlation Coefficient by using the formula:

$$r_s = 1 - 6\left[\frac{\sum_{i=1}^{n} d_i^{\,2}}{n(n^2-1)}\right]$$

Here, d_i= Difference in ranks of $|\hat{u}_i|$ and X_i

Derivation of the Expression

Let us consider **n** pairs of observations of the two variables **X** and **Y**. If the values of these observations are ranked in ascending or descending order, both **X** and **Y** take values of the first **n** natural numbers. Hence, $X_i = 1,2,......,n$ and $Y_i = 1,2,......,n$

⇨ $\sum X_i = \sum Y_i = \dfrac{n(n+1)}{2}$

⇨ $\bar{X} = \bar{Y} = \dfrac{(n+1)}{2}$

Similarly, $\sum X_i^2 = \sum Y_i^2 = \dfrac{n(n+1)(2n+1)}{6}$

Hence,

$$\sigma_X^2 = \frac{\sum X_i^2}{n} - \left(\frac{\sum X_i}{n}\right)^2 = \left(\frac{(n+1)(2n+1)}{6}\right) - \left(\frac{n+1}{2}\right)^2$$

$$= \frac{n^2-1}{12} = \sigma_Y^2 \qquad (\because \sum X_i^2 = \sum Y_i^2 ; \bar{X} = \bar{Y})$$

Let us assume that **d**

⇨ $d_i = X_i - Y_i$

Hence, $\sum d_i^2 = \sum \{(X_i - \bar{X}) - (Y_i - \bar{Y})\}^2 \quad (\because \bar{X} = \bar{Y})$

Or, $\sum d_i^2 = \sum (X_i - \bar{X})^2 + \sum (Y_i - \bar{Y})^2 - 2\sum \{(X_i - \bar{X})(Y_i - \bar{Y})\}$

Or,

$$\frac{\sum d_i^2}{n} = \frac{\sum (X_i - \bar{X})^2}{n} + \frac{\sum (Y_i - \bar{Y})^2}{n} - \frac{2\sum \{(X_i - \bar{X})(Y_i - \bar{Y})\}}{n}$$

$$= \sigma_X^2 + \sigma_Y^2 - 2\,cov(X,Y) = 2\sigma_X^2 - 2\,cov(X,Y) \qquad (\because \sigma_X^2 = \sigma_Y^2)$$

Hence, $cov(X,Y) = \sigma_X^2 - \dfrac{\sum d_i^2}{2n}$

Or,

$$r_s == \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{cov(X,Y)}{\sigma_X^2} = 1 - \frac{\sum d_i^2}{2n\sigma_X^2} == 1 - \frac{12\sum d_i^2}{2n(n^2-1)} = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

- Compute the test statistic (for n>8 and assuming that the population rank correlation coefficient is zero): $\lambda = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \sim t_{(n-2)}$
- The null hypothesis: $H_0 : r_s = 0$
- If the null hypothesis is not rejected, the residuals satisfy the OLS assumption of homoscedasticity.
- If there are more than one independent variables, the rank correlation coefficient can be computed between $|\hat{u}_i|$ and each of the independent variables separately and their statistical significance can be tested.

Limitations of Rank Correlation Coefficient:

- Assumes equal distance across the observations while ranking them and this may affect the correlation coefficient.

7. Koenker-Bassett Test

Model: $Y_i = \alpha + \sum_{j=1}^{k} \beta_j X_{ji} + u_i$

Steps to be followed:

- Estimate the model and obtain the residuals $\hat{u}_i$ and estimated value of Y_i ($\hat{Y}_i$)
- Estimate the auxiliary regression model: $\hat{u}_i^2 = \theta_1 + \theta_2 \hat{Y}_i^2 + v_i$
- Null Hypothesis: $H_0 : \theta_2 = 0$
- If the null hypothesis is rejected, there is problem of heteroscedasticity.

Limitations of Koenker-Bassett Test:

- In case of multiple regression models, the test fails to identify the variables that cause the problem of heteroscedasticity. It only suggests that the estimated model suffers from the problem of heteroscedasticity

These statistical tests in general specify if the random disturbance term is heteroscedastic, but not the manner in which the problem exits. Hence, one should identify the source of the non-constant variance to resolve the problem accordingly. One may start with a variable that has a large range.

- **Remedial Measures**

Redefining the variables

If there are large differences across the observations, one can redefine the variables to reduce the impact of the size differential. Often this approach is preferred as it involves the least amount of tinkering with the original data. It adjusts only the specific variables that need to be redefined in a manner that often makes sense. Indeed, this may also help in improving the model beyond merely removing the heteroscedasticity. In many cases, log transformation of the variables can also resolve the problem of heteroscedasticity as it reduces the scale.

Weighted regression

Weighted regression is a method that assigns each data point a weight based on the variance of its fitted value. The idea is to give small weights to observations associated with higher variances to shrink their squared residuals. Such weighted regression minimizes the sum of the weighted squared residuals. When the correct weights are assigned, the problem of heteroscedasticity can be resolved. However, weighted regression may involve more data manipulation because it applies the weights to all the variables. Finding the theoretically correct weights can be difficult and less intuitive.

<u>Question:</u> For the Model $Y_i = \beta X_i + u_i$ with $var(u_i) = \sigma^2 Z_i^2$, prove that the WLS estimator of β has lower variance than its OLS estimator, where the weight is $w_i = \frac{1}{Z_i}$

<u>Answer:</u> Model: $Y_i = \beta X_i + u_i$ with $var(u_i) = \sigma^2 Z_i^2$

The OLS estimator of β: $\hat{\beta} = \frac{\sum X_i Y_i}{\sum X_i^2}$ and $var(\hat{\beta}) = \frac{\sum X_i^2 \, var(u_i)}{\left(\sum X_i^2\right)^2} = \frac{\sigma^2 \sum X_i^2 Z_i^2}{\left(\sum X_i^2\right)^2}$

If we divide the entire equation by Z$_i$, $\frac{Y_i}{Z_i} = \beta \frac{X_i}{Z_i} + \frac{u_i}{Z_i}$ or $y_i = \beta x_i + v_i$

Here $var(v_i) = \frac{var(u_i)}{Z_i^2} = \frac{\sigma^2 Z_i^2}{Z_i^2} = \sigma^2$ (constant – homoscedasticity)

The WLS estimator of β: $\beta^* = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i(\beta x_i + v_i)}{\sum x_i^2} = \beta + \frac{\sum x_i v_i}{\sum x_i^2}$

$$var(\beta^*) = E(\beta^* - \beta)^2 = E\left(\frac{\sum x_i v_i}{\sum x_i^2}\right)^2 = \frac{\sum x_i^2 E(v_i^2)}{\left(\sum x_i^2\right)^2} = \frac{\sigma^2 \sum x_i^2}{\left(\sum x_i^2\right)^2} = \frac{\sigma^2}{\sum x_i^2} \; [\because E(u_i u_j) = 0]$$

Hence, $\frac{var(\beta^*)}{var(\hat{\beta})} = \frac{\frac{\sigma^2}{\sum x_i^2}}{\frac{\sigma^2 \sum X_i^2 Z_i^2}{\left(\sum X_i^2\right)^2}} = \frac{\left(\sum X_i^2\right)^2}{\left(\sum X_i^2 Z_i^2\right)\left(\sum x_i^2\right)} = \frac{\left(\sum X_i^2\right)^2}{\left(\sum X_i^2 Z_i^2\right)\left(\sum\left(\frac{X_i}{Z_i}\right)^2\right)}$

Let, $X_iZ_i = a_i$ and $\frac{X_i}{Z_i} = b_i$

⇨ $a_ib_i = X_i^2$

Hence, $\frac{var(\beta^*)}{var(\hat{\beta})} = \frac{\left(\sum a_ib_i\right)^2}{\left(\sum a_i^2\right)\left(\sum b_i^2\right)}$

According to Cauchy-Schwartz inequality, $\left(\sum a_ib_i\right)^2 < \left(\sum a_i^2\right)\left(\sum b_i^2\right)$

⇨ $var(\beta^*) < var(\hat{\beta})$

When $a_i = \theta b_i$

$\frac{var(\beta^*)}{var(\hat{\beta})} = \frac{\left(\sum a_ib_i\right)^2}{\left(\sum a_i^2\right)\left(\sum b_i^2\right)} = \frac{\left(\sum a_ib_i\right)^2}{\left(\sum a_i^2\right)\left(\sum b_i^2\right)} = 1$, i.e., $var(\beta^*) = var(\hat{\beta})$

When $a_i = \theta b_i$, $\frac{a_i}{b_i} = \frac{X_iZ_i}{\frac{X_i}{Z_i}} = Z_i^2 = \theta$

i.e., $var(u_i) = \sigma^2 Z_i^2 = \theta\sigma^2$ (constant – homoscedasticity)

Heteroscedasticity consistent covariance matrix (HCCM) estimation

White developed a method for obtaining consistent estimates of the variances and covariance of the OLS estimators. This is known as the heteroscedasticity consistent covariance matrix (HCCM) estimator. Most statistical packages have an option to calculate the HCCM.

Feasible generalized least squares (FGLS) estimator

The method of GLS can potentially resolve the problem of heteroscedasticity. However, the GLS estimator requires σ_i to be known for each of the observations. In order to make the GLS estimator feasible, one can use the sample data to obtain an estimate of σ_i for each observation in the sample. Subsequently, the GLS estimator can be applied using the estimates of σ_i . This is called the Feasible Generalized Least Squares Estimator or the FGLS estimator.

Thus, instead of assuming the structure of heteroscedasticity, one may estimate the structure of heteroscedasticity from the OLS estimators.

Example: Assume that we have the following linear regression model:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

It is assumed that the error variance is a linear function of X_{1i} and X_{2i} and the rest of the assumptions of the classical linear regression model hold. Hence, we can assume the following heteroscedasticity structure:

$$var(u_i) = \sigma_i^2 = \theta_0 + \theta_1 X_{1i} + \theta_2 X_{2i} + v_i$$

In order to obtain the FGLS estimates of the parameters one needs to proceed as follows.

- Regress Y_i against a constant, X_{1i}, and X_{2i} applying the OLS method
- Calculate the residuals from this regression $\hat{u}_i$ and square these residuals $\hat{u}_i^2$
- Regress the squared residuals $\hat{u}_i^2$ and estimate the auxiliary regression model.
- Use the estimates of the coefficients to calculate the predicted values $\hat{\sigma}_i^{\ 2}$. This is an estimate of the error variance for each observation.
- Check the predicted values. For any predicted value that is non-positive, replace it with the squared residual for that observation. This ensures that the estimate of the variance is a positive number.
- Find the square root of the estimate of the error variance $\hat{\sigma}_i$ for each observation.
- Calculate the weight $w_i = 1/\hat{\sigma}_i$ for each observation.
- Multiply Y_i, , X_{1i}, and X_{2i} for each observation by its weight.
- Regress w_iY_i on w_i, w_iX_{1i}, and w_iX_{2i} using the OLS method.

Properties of the FGLS Estimator

If the assumed model of heteroscedasticity is a reasonable approximation of the true heteroscedasticity, then the FGLS estimator has the following properties:

- It is non-linear.
- It is biased in small samples.
- It is asymptotically more efficient than the OLS estimator.
- Monte Carlo studies suggest that it tends to yield more precise estimates than the OLS estimator. However, if the assumed model of heteroscedasticity is not a reasonable approximation of true heteroscedasticity, the FGLS estimator will yield worse estimates than the OLS estimator.

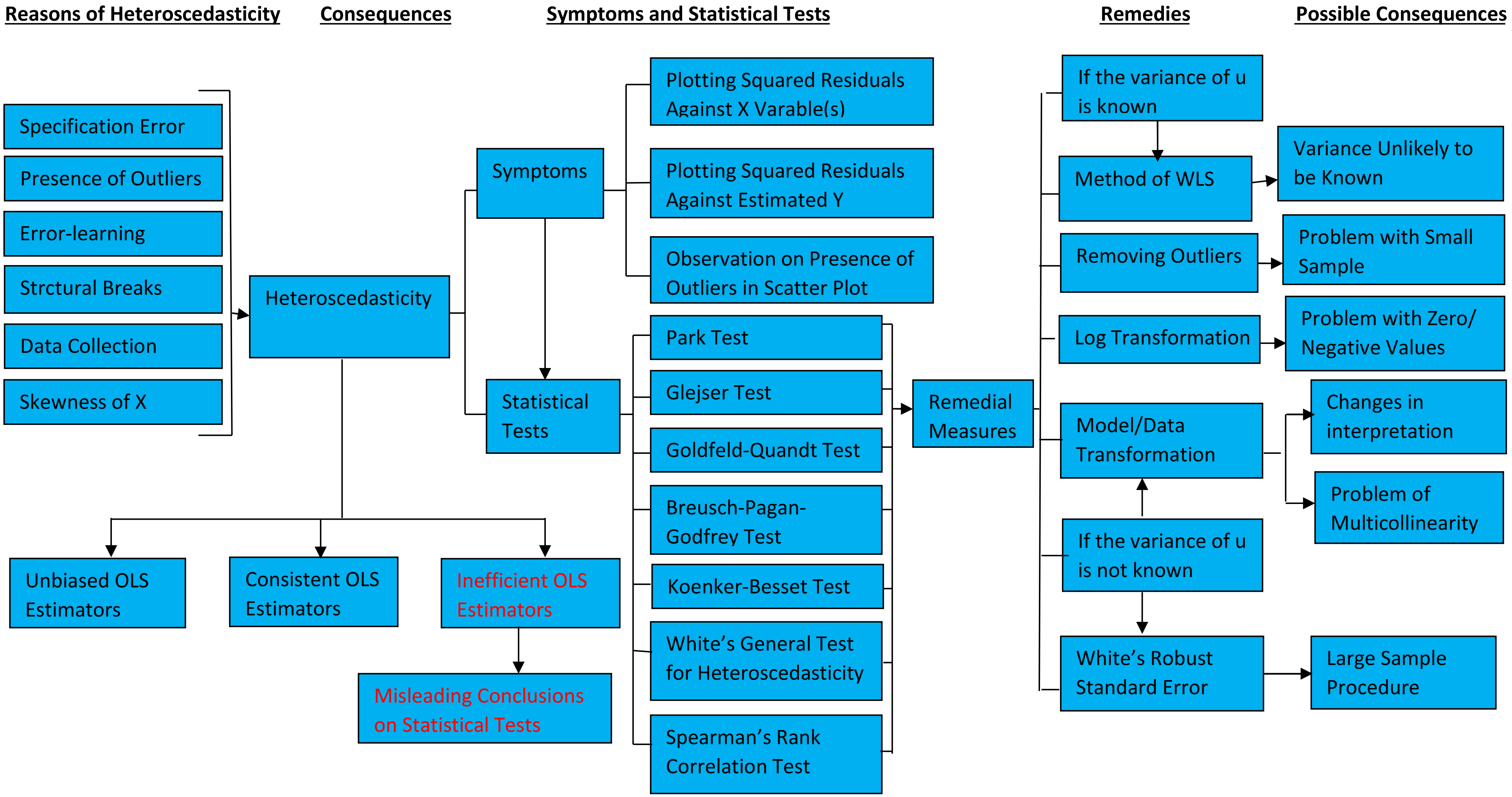
Transforming the dependent variable

Here, original data are transformed into different values that produce good looking residuals. In many situations, a Box-Cox transformation on the dependent variable is used. Again, this process can also involve manipulation of the data. Sometimes, it also makes interpretation of the results very difficult.

- **Concluding Remarks**

There are many different reasons for heteroscedasticity. Identifying the cause(s) and resolving the problem to resolve heteroscedasticity can require extensive subject-area knowledge. In most cases, remedial measures for severe heteroscedasticity are necessary. However, if the primary goal is to predict the dependent variable rather than estimating the specific effects of the independent variable(s), one may not need to correct the non-constant variance. Further, getting more precise estimates with an alternative estimator requires specifying approximate heteroscedasticity structure. Otherwise, alternative estimators can yield estimates that are worse than the original OLS estimator.

Heteroscedasticity



Reasons of Heteroscedasticity

Consequences

Symptoms and Statistical Tests

Remedies

Possible Consequences

Specification Error

Presence of Outliers

Error-learning

Strctural Breaks

Data Collection

Skewness of X

Heteroscedasticity

Unbiased OLS Estimators

Consistent OLS Estimators

Inefficient OLS Estimators

Misleading Conclusions on Statistical Tests

Symptoms

Plotting Squared Residuals Against X Varable(s)

Plotting Squared Residuals Against Estimated Y

Observation on Presence of Outliers in Scatter Plot

Statistical Tests

Park Test

Glejser Test

Goldfeld-Quandt Test

Breusch-Pagan-Godfrey Test

Koenker-Besset Test

White's General Test for Heteroscedasticity

Spearman's Rank Correlation Test

Remedial Measures

If the variance of u is known

Method of WLS

Removing Outliers

Log Transformation

Model/Data Transformation

If the variance of u is not known

White's Robust Standard Error

Variance Unlikely to be Known

Problem with Small Sample

Problem with Zero/ Negative Values

Changes in interpretation

Problem of Multicollinearity

Large Sample Procedure