**B.Sc (Statistics) Sem. IV CC 409 (Linear Models)**

# Topic- Multi-Collinearity and Heteroscedasticity

**Dr. Saurabh**

**Asstt. Professor**

**Dept. of Statistics**

**(email- shaurabh.bhu@gmail.com)**

# MULTI-COLLINEARITY

The term multi-collinearity was given by **Ragnar Frisch**. When explanatory variables are related to each other, it implies existence of multi-collinearity in the model.

**Multi-collinearity can be of two types:**


**1. Perfect Multi-collinearity**

$$\beta_1 X_{1i} + \beta_2 X_{2i} + \ldots\ldots\ldots + \beta_n X_{ni} = 0$$

If none of the coefficients of X, that is, β is equal to zero, then one explanatory variable can be converted into another.


**2. Imperfect Multi-collinearity**

It is the case when one explanatory variable cannot be perfectly converted into another, and the model is:

$$\beta_1 X_{1i} + \beta_2 X_{2i} + \ldots\ldots\ldots + \beta_n X_{ni} + v_i = 0$$

**Causes of Multi-collinearity**

These are given by **Montgomery and Peck**

1. Problems in the specification of the model or nature of specification of the model.
2. Data collection methods may impose a limiting range on the values taken by the regressors in a population.
3. Constraint on the model may also result in the problem of multi-collinearity.
4. Over-determined model also exhibit the problem of multi-collinearity. If number of observations are greater than the number of parameters, we will not be able to get unbiased and unique solutions.
5. Lagged values of the variables included in the model also lead to multi-collinearity. In other words, multi-collinearity is more commonly found in time series data.

**Consequences of Multi-collinearity**

1. Even though the OLS estimates are BLUE, but they have large variances and covariances.
2. Confidence intervals are widened leading to the acceptance of null hypothesis ($H_0$) more frequently.
3. T-ratio becomes statistically insignificant.
4. $R^2$ is very high but t-ratios are insignificant.
5. OLS estimators and their standard errors can be very sensitive to changes in the data.

Estimates' properties at the time of multi-collinearity

1. The estimates might not be able to determine them uniquely.
2. Variance and standard errors will be infinite
3. $\hat{\beta}s$ are unbiased and efficient but have large variances.

**Detection of Multi-collinearity**

Following are the signs which could help locate multi-collinearity in our models:

1. **High $R^2$ and few significant t-ratios**.
2. **High pair wise correlations among the regressors** will also indicate multi-collinearity. This is only a necessary condition and not sufficient, which means that even if the pair wise correlation coefficients are zero, multi-collinearity can exist.

3. **Frisch's Confluence Test**: It regresses the dependent variable on each of the explanatory variable separately and thus obtain all the possible simple regressions and examine the results on the basis of:
    - Apriori knowledge
    - Statistical criteria

    Then we choose that elementary regression which has most plausible result on the basis of apriori and statistical criteria. We gradually then insert additional explanatory variables and examing the effects on individual coefficients, standard error and overall $R^2$.

    If the new explanatory variable improves $R^2$ without rendering the individual coefficient unacceptable, then this new explanatory variable is retained, otherwise it is rejected as a superfluous variable. Also if the new variable changes the signs and the values of individual coefficients considerably, then it is an indication that multi-collinearity is a severe problem in particular data set and needs to be taken care of.

4. **Variance Inflation Factor (VIF)** = $\frac{1}{(1-r_{23}^2)}$

    **Tolerance (TOL)** = $\frac{1}{VIF}$

    If TOL is close to zero, then multicollinearity is present and if it closer to one then less degree of multicollinearity.

    If $r^2 \sim 1$, greater will be the VIF and more is the multicollinearity.

5. **Auxillary Regression**: Under this we exclude one variables and regress it on other variables to estimate our $R^2$. Steps followed are:

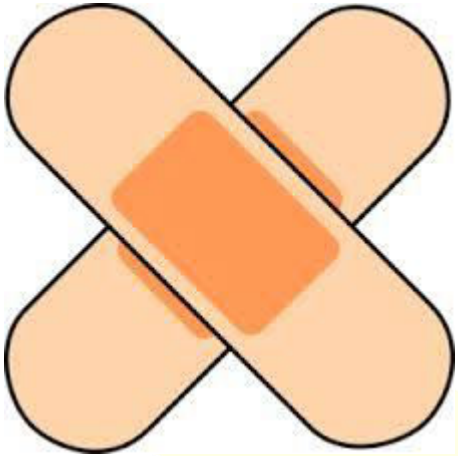


6. **Condition Index**: The method uses Eigen values.

$$\textbf{Condition Index (CI)} = \frac{\textit{Maximum Eigen Value}}{\textit{Minimum Eigen Value}}$$

| | |
|---|---|
| If $100 < CI < 1000$ | Moderate to strong multicollinearity |
| If $CI > 1000$ | Severe multicollinearity |
| | |
| If $10 < \sqrt{CI} < 30$ | Moderate to strong multicollinearity |
| If $\sqrt{CI} > 30$ | Severe multicollinearity |

**Remedies for correcting multi-collinearity**



1. **Using the apriori information**
2. **Dropping a variable.**
3. **Transformation of a variable:** Given a set of data, if there is a problem of multicollinearity, it can also be solved using the first difference form.
4. **Additional or new data**: Since multi-collinearity is a sample phenomenon, it can be reduced by taking another sample or by increasing the sample size.
5. **Reducing polynomial regressions**: having lesser number of polynomial variables in regression equation also takes care of multicollinearity.
6. **Combining cross-section and time series data:** In the face of multi-collinearity, one method of reducing it is pooling of time-series and cross-section data. This is done by estimating regression coefficient from cross-section data and then incorporating them in the original regression equation.

# HETEROSCEDASTICITY

. The term heteroscedasticity means that the variance of error terms is not constant, that is, different error terms have different variances.

$$\sigma_i^2 = f(X_i)$$


So, hetero-scedasticity is the problem of fluctuating variances of error terms. Moreover, the variance of the error terms depends on the value of the explanatory variables. It is a rule in cross-section data and very rate in time series data.

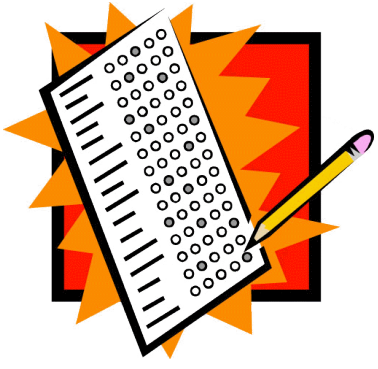
**Causes of Heteroscedasticity**

1. All the models pertaining to learning skills or error learning models exhibit heteroscedasticity.
2. Economic variables such as income and wealth also show heteroscedasticity because as income or wealth increase, so is the discretion to use it.
3. Some economic variables exhibit the skewness in distribution.
4. Data collection technique improves with time. As a result constant variance is not found for all those economic variables where data collecting techniques are changing very fast.
5. Specification errors in the models also lead to heteroscedasticity.
6. Incorrect data transformation methods also show heteroscedasticity.
7. Outliers in the data may result into the problem of heteroscedasticity

**Consequences of heteroscedasticity**

1. The estimators are still linear, unbiased and consistency doesn't change.
2. However, the estimators are no longer BLUE because they are no longer efficient, therefore, they are not the best estimators. The estimators do not have a constant variance.

**Tests of Heteroscedasticity**



1. **Graphical Method**
   Under this method, the error term is plotted against the X-variable and then observe whether there is any systematic pattern or not. If the graph shows some pattern, it would imply that there is heteroscedasticity present in the model.

2. **Park Test**: Under this test, we assume that variance is a function of the explanatory variable (X), that is,

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

Taking log on both sides, we get;

Log $\sigma_i^2$ = log$\sigma^2$ + $\beta$log$X_i$ + $v_i$

Log $\widehat{u_i^2}$ = $\alpha$ + $\beta$log$X_i$ + $v_i$

We cannot directly observe $\sigma_i^2$ . so for all heteroscedasticity test, we will be using $\widehat{u_i^2}$.

Steps followed under this method are:

Run OLS and calculate $\hat{u}_i$

Run regression of Log $\widehat{u_i^2}$ on log $X_i$ as given in the above equation.

Apply t-test and if $\beta_1$ turns out to be significantly different from zero, then it implies heteroscedasticity is present.

3. **Gleizer Test**: Error term is related to the independent variable through different functional forms. According to him, following functional forms can be tried and heteroscedasticity can be tested:
   - $\widehat{u_i^2} = \beta_1 + \beta_2 X_i + v_i$
   - $\widehat{u_i^2} = \beta_1 + \beta_2 \sqrt{X_i} + v_i$
   - $\widehat{u_i^2} = \beta_1 + \beta_2 \frac{1}{X_i} + v_i$
   - $\widehat{u_i^2} = \beta_1 + \beta_2 \frac{1}{\sqrt{X_i}} + v_i$
   - $\widehat{u_i^2} = \sqrt{\beta_1 + \beta_2 X_i} + v_i$
   - $\widehat{u_i^2} = \sqrt{\beta_1 + \beta_2 {X_i}^2} + v_i$

We then judge the statistical significance of $\beta_1$ & $\beta_2$ by any standard test. If they are estimated to be statistically different from zero, then heteroscedasticity is present.

There can be two possibilities:

| | |
|---|---|
| **If $\beta_1 = 0$ & $\beta_2 \neq 0$** | **Pure heteroscedasticity** |
| **If $\beta_1 \neq 0$ & $\beta_2 \neq 0$** | Mixed heteroscedasticity |


4. **Spearman's Rank Correlation Test**: Following steps are followed to conduct this test:
**Step 1**: Fit the regression to the data and obtain the residual $\hat{u}_i$.

**Step 2:** Ignoring the sign of $\hat{u}_i$ by taking the absolute values, we rank both $|\hat{u}_i|$ and $X_i$s either in ascending order or descending order.

**Step 3**: Now compute the Spearman's Rank Correlation

$r_s = 1 - 6[\frac{\sum d_i^2}{n(n^2-1)}]$

**Step 4**: Now assuming that the population correlation coefficient is zero, we apply the t-test.

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$$

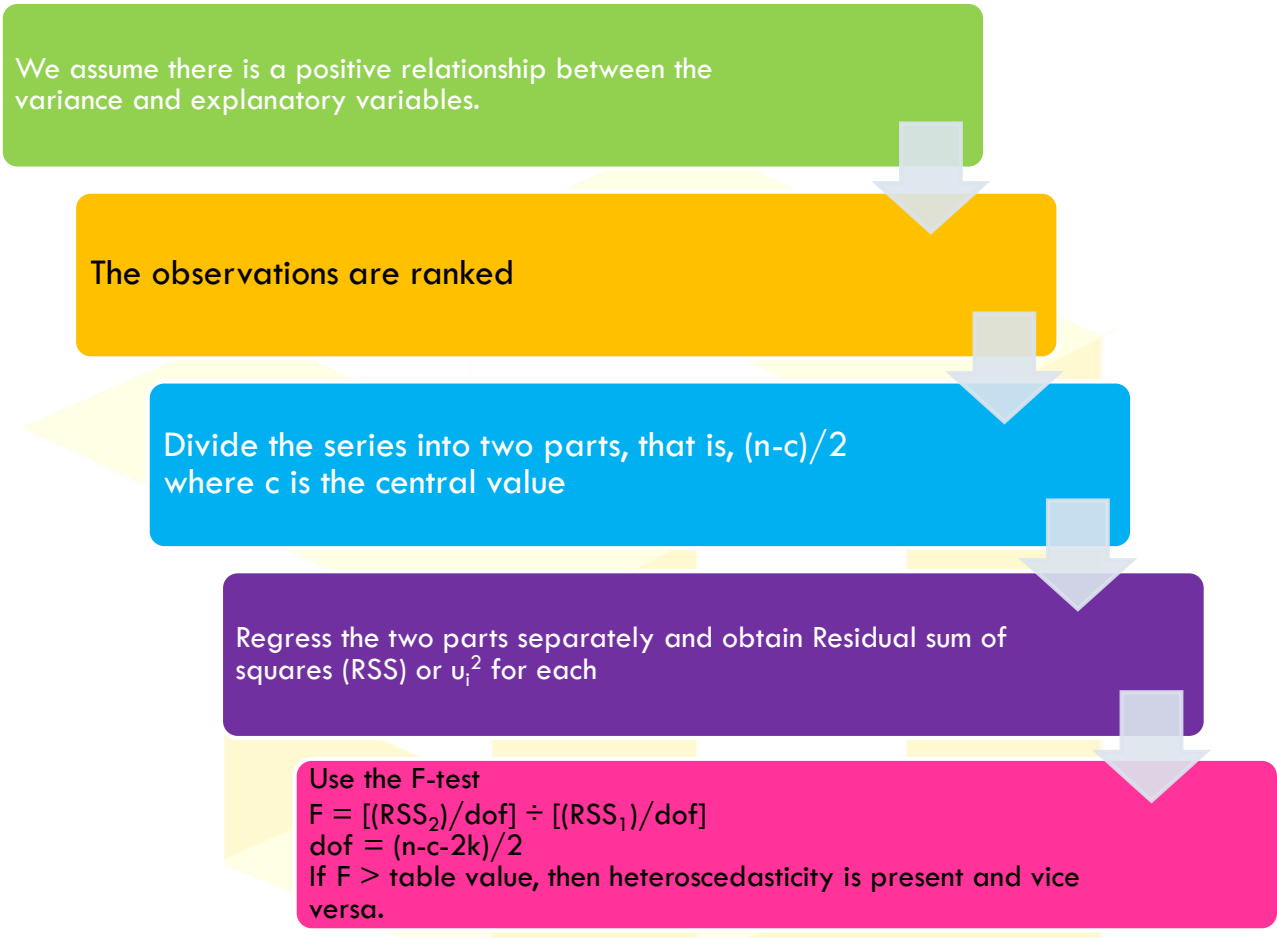
**For n-2 degrees of freedom.**
**If the computed t-value is greater than the table value, heteroscedasticity is present.**
**If computed t-value is less than the table value, then there is no correlation between $X_i$ and $u_i$.**

5. **Goldfeld-Quandt test:** This test is only applicable to sample size greater than or equal to 30.

The following assumptions are taken into consideration while applying this test:

a. $\sigma_{ui}^2 = \sigma^2 X_i^2$
b. assumes that $u_i$ are not auto-correlated
c. Error term follows a normal distribution.
d. Number of observations is atleast twice the number of parameters.

We follow the underlined steps while applying this test:




6. **White's General Heteroscedasticity Test**: The main merit of this test is it has no assumptions like normality.

Steps to perform this test are:

a. Calculate $\widehat{u_i}s$
b. Regress $\widehat{u_i}$ on our explanatory variable X.
c. $\widehat{u^2}_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{2i}^2 + \alpha_5 X_{3i}^2 + \alpha_6 X_{2i} X_{3i} + v_i$
d. Calculate $R^2$, that is, goodness of fit.
e. Use $\chi^2$ test to estimate the model.
f. Test for heteroscedasticity and specification errors. If no cross products are present, that is, ($X_{2i}X_{3i}$), then pure heteroscedasticity. If cross products are present, then it reveals the presence of heteroscedasticity and specification bias.