DATA SCIENCE – BIVARIATE ANALYSIS

HOMOSCEDASTICITY & HETEROSCEDASTICITY

HOMOSCEDASTICITY

Purpose:

- It ensures that the variance of the errors in a regression model is constant across all levels of the independent variables.
- This is a key assumption for ordinary least squares (OLS) regression, allowing for accurate and reliable statistical inferences.

Rule of Thumb:

- Ideally, the spread of residuals should be evenly distributed around the regression line.
- When examining a residual plot (residuals vs. predicted values), look for a random, scattered pattern with no discernible trend.

When to Use:

- When you want to perform OLS regression and ensure that the standard errors of the coefficients are reliable.
- When you expect the variability of the dependent variable to be consistent across the range of independent variables.

Points:

I. It allows for reliable t-tests and F-tests.
II. It allows for valid confidence intervals.
III. It means that the OLS estimators are the best linear unbiased estimators (BLUE).
IV. It means that the variance of the error terms are equal.
V. It simplifies statistical analysis.
VI. It is a key assumption for many statistical models.

HETEROSCEDASTICITY

Purpose:

- It describes a situation where the variance of the errors in a regression model is not constant.

- It indicates that the variability of the dependent variable changes across the range of independent variables.

Rule of Thumb:

- In a residual plot, heteroscedasticity often manifests as a cone-shaped or fan-shaped pattern, where the spread of residuals increases or decreases with the predicted values.

When to Use:

- You don't "use" heteroscedasticity; it's a condition that exists in your data.
- You need to identify and address it to ensure the validity of your statistical inferences.
- when the variance of the error terms are not equal.

Points:

i. It leads to biased standard errors, making statistical tests unreliable.
ii. It can result in incorrect conclusions about the significance of regression coefficients.
iii. It indicates that the OLS estimators are not efficient.
iv. It often occurs when dealing with data that has a wide range of values.
v. Techniques like weighted least squares or robust standard errors are needed to correct for it.
vi. It is a common problem in econometric studies.

Which is Better?

- Homoscedasticity is generally preferred because it allows for more reliable statistical inferences.
- Heteroscedasticity is a problem that needs to be addressed. It doesn't mean the data is useless, it just means that standard OLS regression may provide misleading results.
- Therefore, if possible, data that is homoscedastic is better for basic OLS regression. When heteroscedasticity is found, there are methods to correct for it, and those methods should be employed.