TA3

Arun Kumar Rajasekaran

Linear Regression

Just to extend upon our discussion from last TA

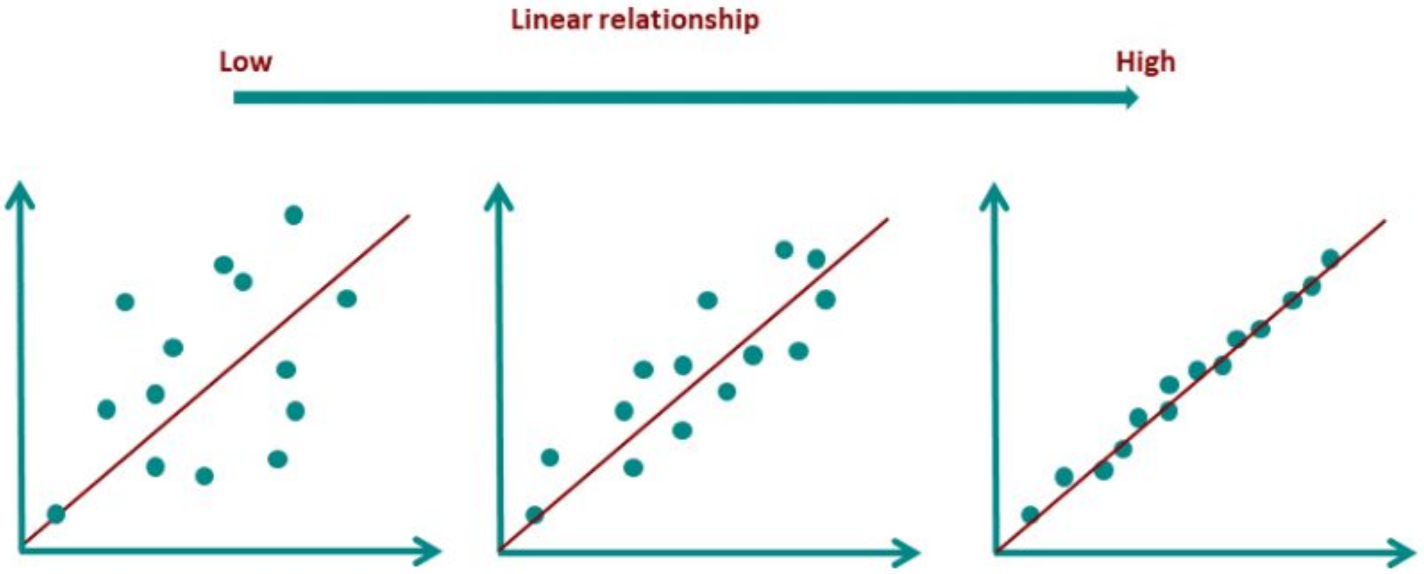
Linear Regression

The regression line can be described by the following equation:

$$\underbrace{\hat{y}}_{\text{Estimated dependent variable}} = \underbrace{b}_{\text{Slope}} \cdot \underbrace{x}_{\text{Independent variable}} + \underbrace{a}_{\text{y intercept}}$$

Definition of "Regression coefficients":

- ***a*** : point of intersection with the *y*-axis
- ***b*** : gradient of the straight line



Assumptions of Linear Regression

Generic 'Least square method' (ref. Next slide)

Step 1: *Calculate the slope 'm' by using the following formula:*

$$m = \frac{n\sum xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

$$b = \frac{\sum(x - \bar{x}) * (y - \bar{y})}{\sum(x - \bar{x})^2}$$

Step 2: *Compute the y-intercept (the value of y at the point where the line crosses the y-axis):*

$$c = y - mx$$

Step 3: *Substitute the values in the final equation:*

$$y = mx + c$$

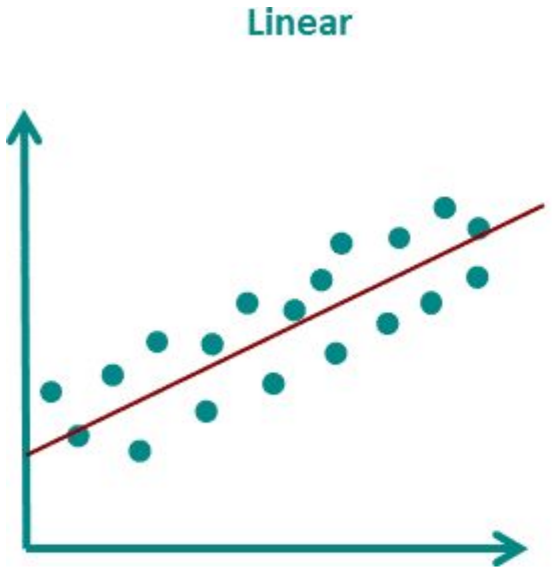
Assumptions of Linear Regression

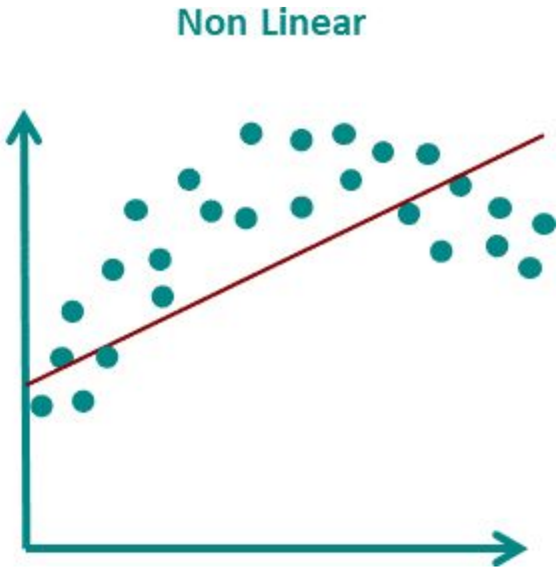
https://www.technologynetworks.com/informatics/articles/calculating-a-least-squares-regression-line-equation-example-explanation-310265

Assumptions of Linear Regression

- Linearity: There must be a linear relationship between the dependent and independent variables.
- Homoscedasticity: The residuals must have a constant variance.
- Normality: Normally distributed error
- No multicollinearity: No high correlation between the independent variables
- No auto-correlation: The error component should have no auto-correlation

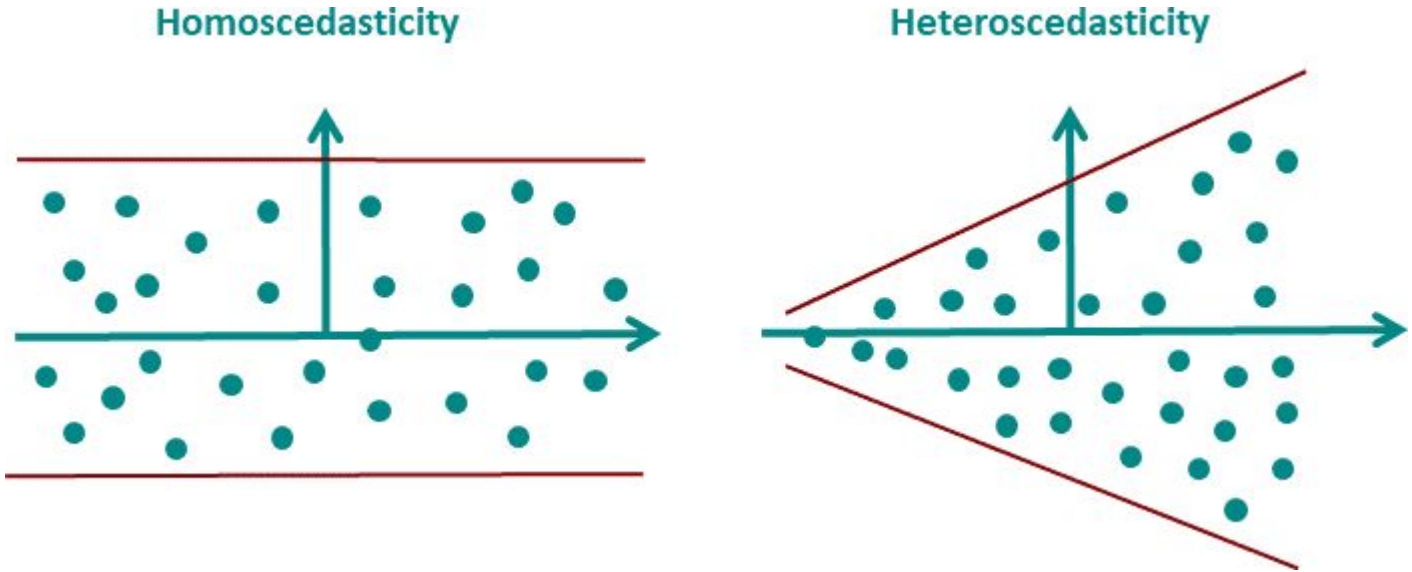
1. Linearity





In linear regression, a straight line is drawn through the data. This straight line should represent all points as good as possible. If the points are distributed in a non-linear way, the straight line cannot fulfill this task.

2. Homoscedasticity



Since in practice the regression model never exactly predicts the dependent variable, there is always an error. This very error must have a constant variance over the predicted range.

2. Homoscedasticity

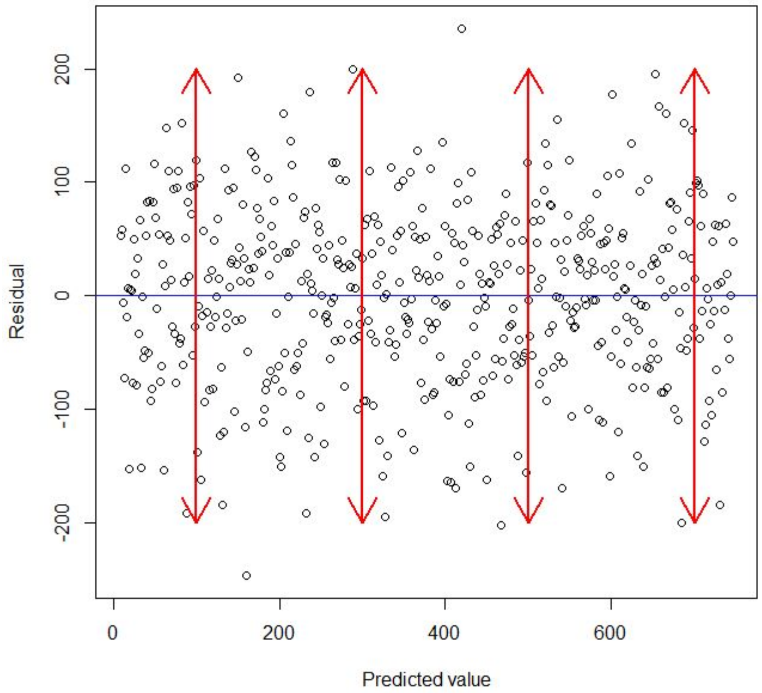
Advanced analysis include,

***F*-Test**

Modified Levene Test

Breusch-Pagan Test

Bartlett's Test



2. Homoscedasticity

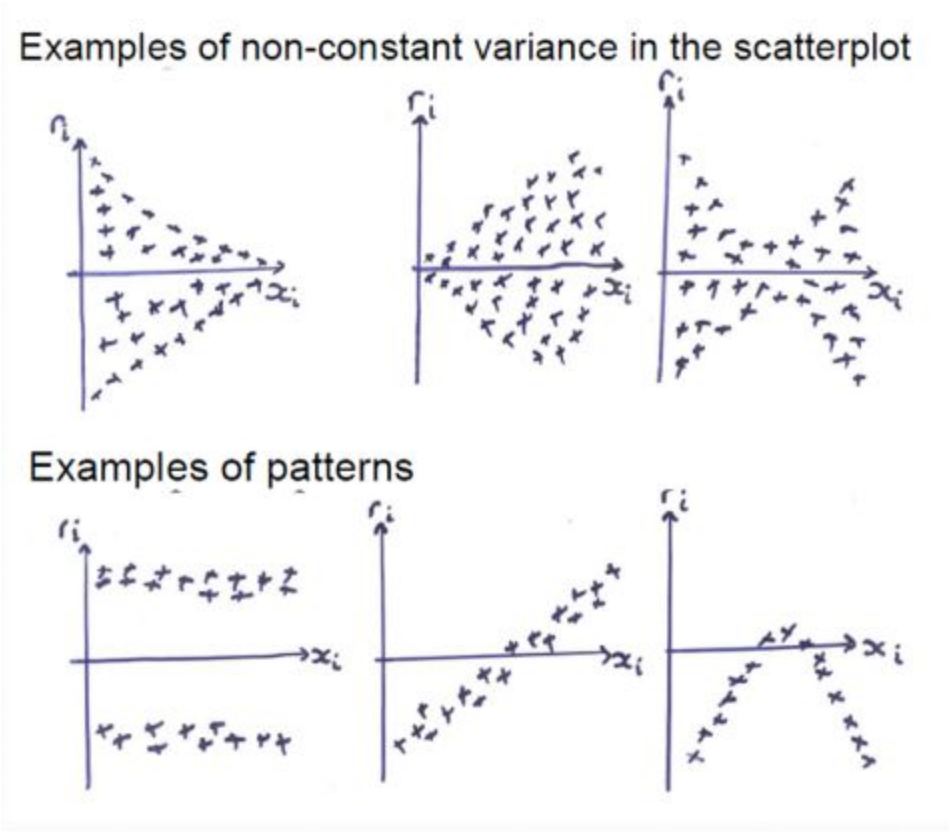
Advanced analysis include,

***F*-Test**

Modified Levene Test

Breusch-Pagan Test

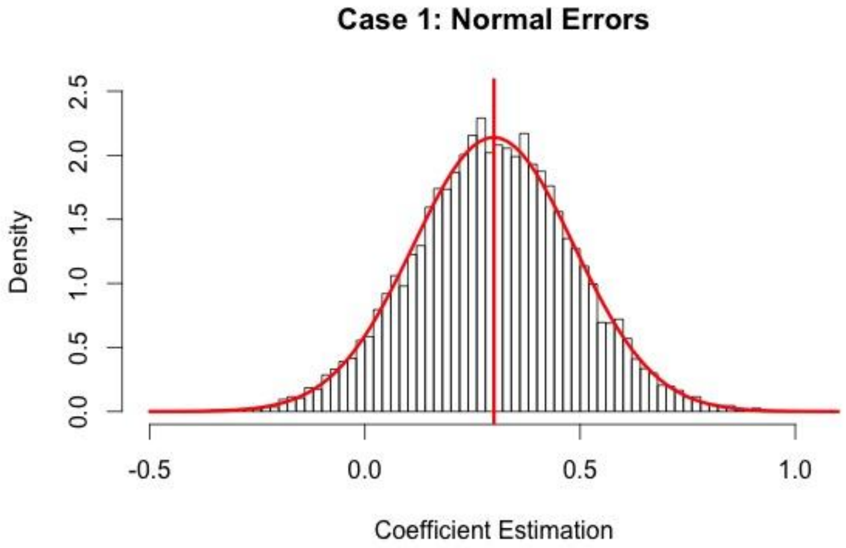
Bartlett's Test

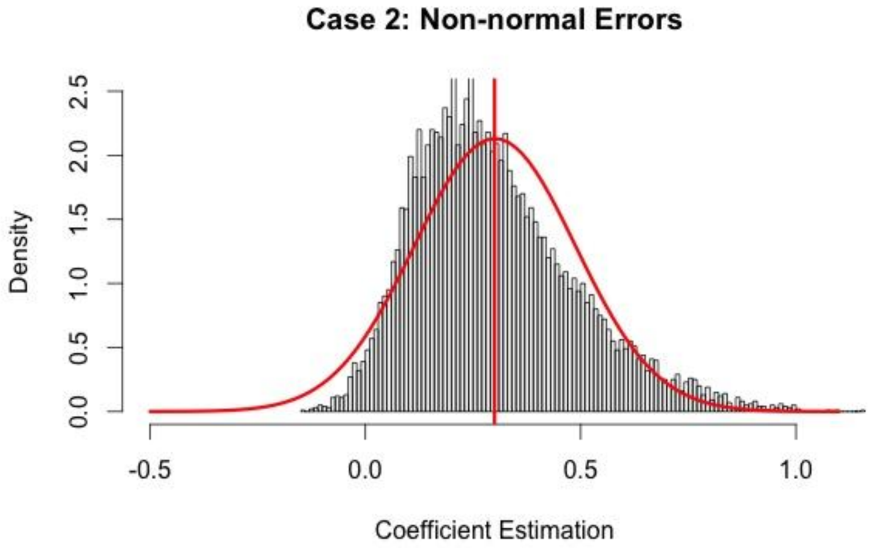


2. Homoscedasticity (how to fix?)

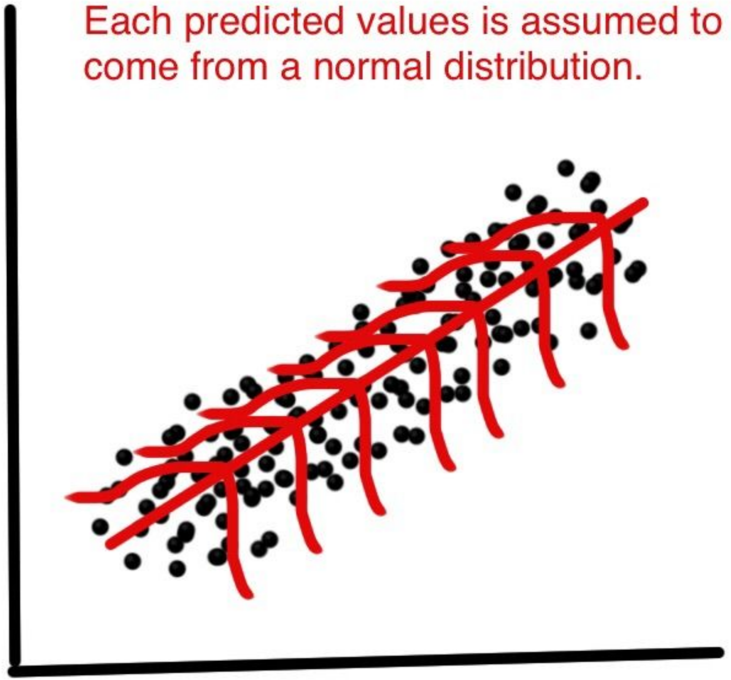
- **Variance stabilizing transformation:** A transformation of the outcome used to correct non-constant variance is called a “variance stabilizing transformation.” common transformations are the natural logarithm, square root, inverse, and Box-Cox
- Advanced methods such as weighted or generalized least squares can be used to handle non-constant variance.
- Non-constant variance may co-occur with non-linearity and/or non-normality.

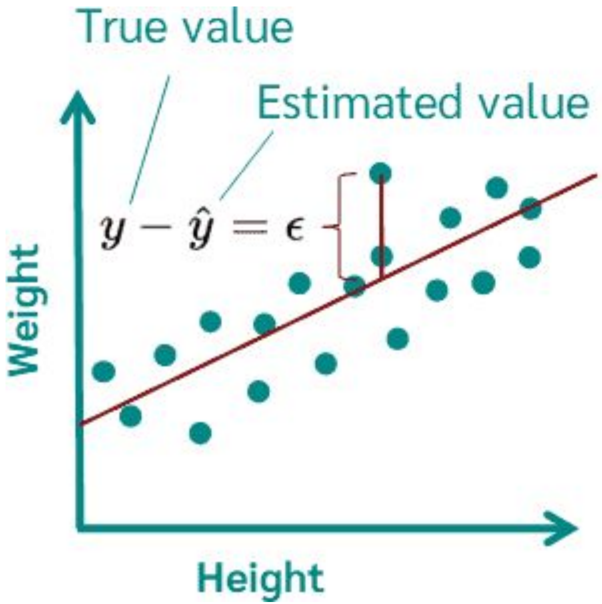
3. Normal distribution of error





3. Normal distribution of error





Error epsilon

$$y = b \cdot x + a + \boxed{\epsilon}$$

Multiple LR vs Multivariate

Simply…

“Regression analysis results in a formula of the form Y=a+bX. A multiple regression has more than one X in one formula. A multivariate regression has more than one Y, but in different formulae. And a multivariate multiple regression has multiple X’s to predict multiple Y’s with each Y in a different formula, usually based on the same data.”

A bit more, equations…

Simple regression pertains to one dependent variable (y) and one independent variable (x): $y=f(x)$

Multiple regression (aka multivariable regression) pertains to one dependent variable and multiple independent variables: $y=f(x1,x2,...,xn)$

Multivariate regression pertains to multiple dependent variables and multiple independent variables: $y1,y2,...,ym=f(x1,x2,...,xn)$

.

TA Open Discussion

Large language models