

Regression Modelling

Interesting Hypothesis

- Vegetarians miss fewer flights.
- Women use camera phone more than men.
- Left handed men earn more money!
- Smokers are better sales people.
- Those who whistle at work place are more efficient.

Regression enables us to test few of these Hypotheses

What is Regression?

Regression is a tool for finding **existence of an association** relationship between a dependent variable (Y) and one or more independent variables(X_1, X_2, \dots, X_n) in a study.

The relationship can be **linear** or **non-linear**.

- Mathematical relationship is an exact relationship.

$$Y = \beta_0 + \beta_1 X$$

- Statistical relationship is not an exact relationship.

$$Y = \beta_0 + \beta_1 X + e$$

(This is the Population Regression Function)

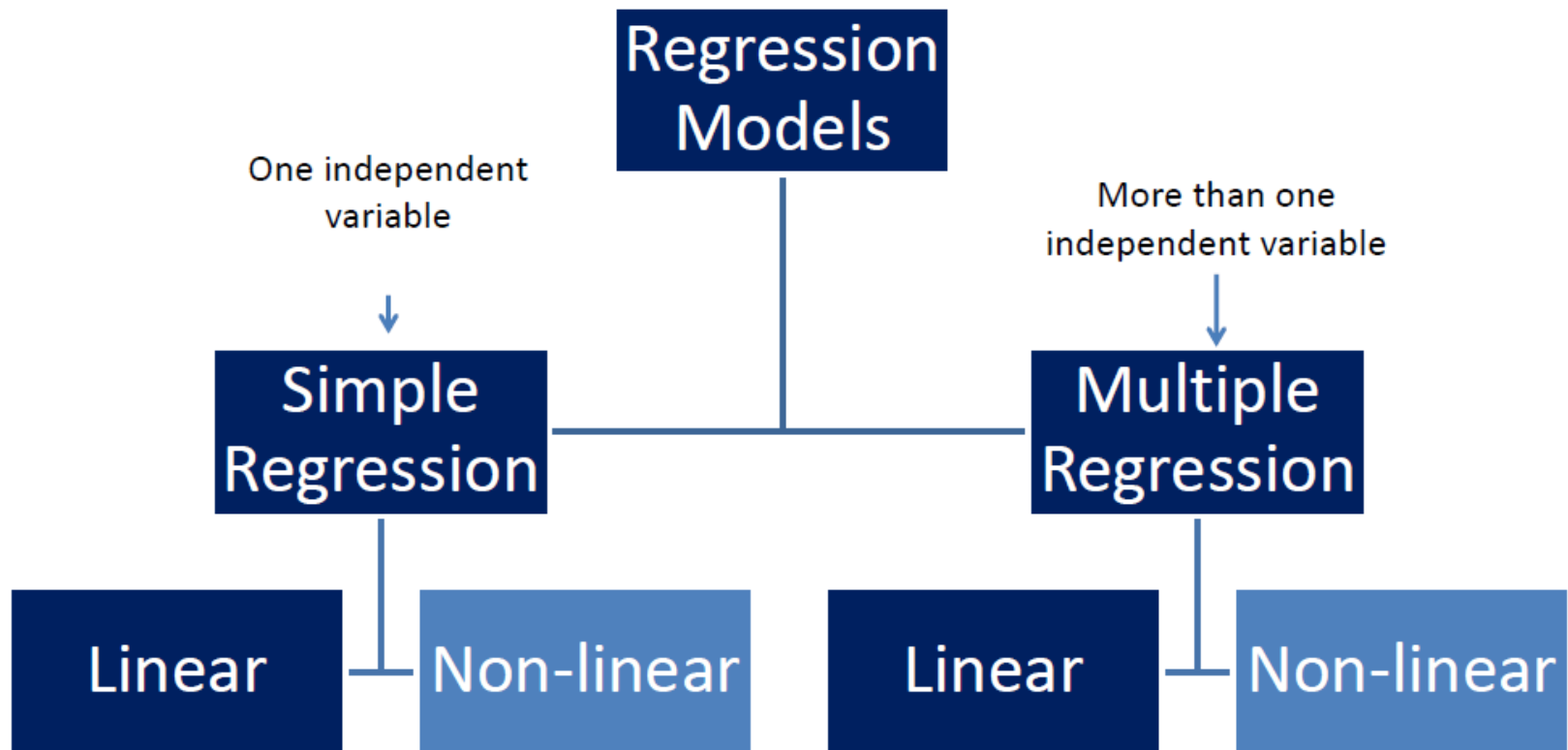
Regression Nomenclatures

<u>Dependent Variable</u>	<u>Independent Variable</u>
Explained Variable	Explanatory variable
Regressand	Regressor
Predictand	Predictor
Endogenous Variable	Exogenous Variable
Controlled Variable	Control Variable
Target Variable	Stimulus Variable
Response Variable	

Dependent and Independent Variables

- Terms dependent and independent does not necessarily imply a causal relationship between two variables.
- Regression is **not** designed to capture **causality**.
- Purpose of regression is to **predict the value of dependent variable given the value(s) of independent variable(s)**.

Types of Regression

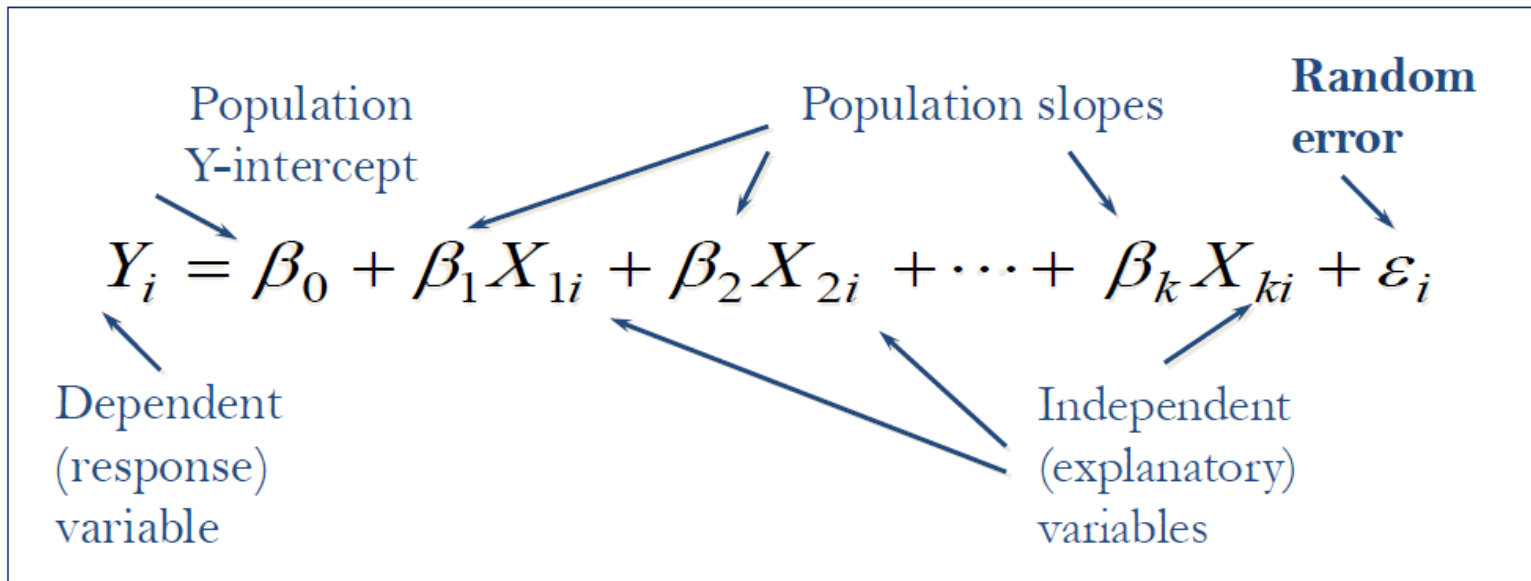


Types of Linear Regression

- Simple Linear Regression: $Y = B_0 + B_1 X + e$
- Multiple Linear Regression: $Y = B_0 + B_1 X + B_2 X_1 + B_3 X_2 + \dots + B_k X_k + e$

Multiple Linear Regression Model (MLRM)

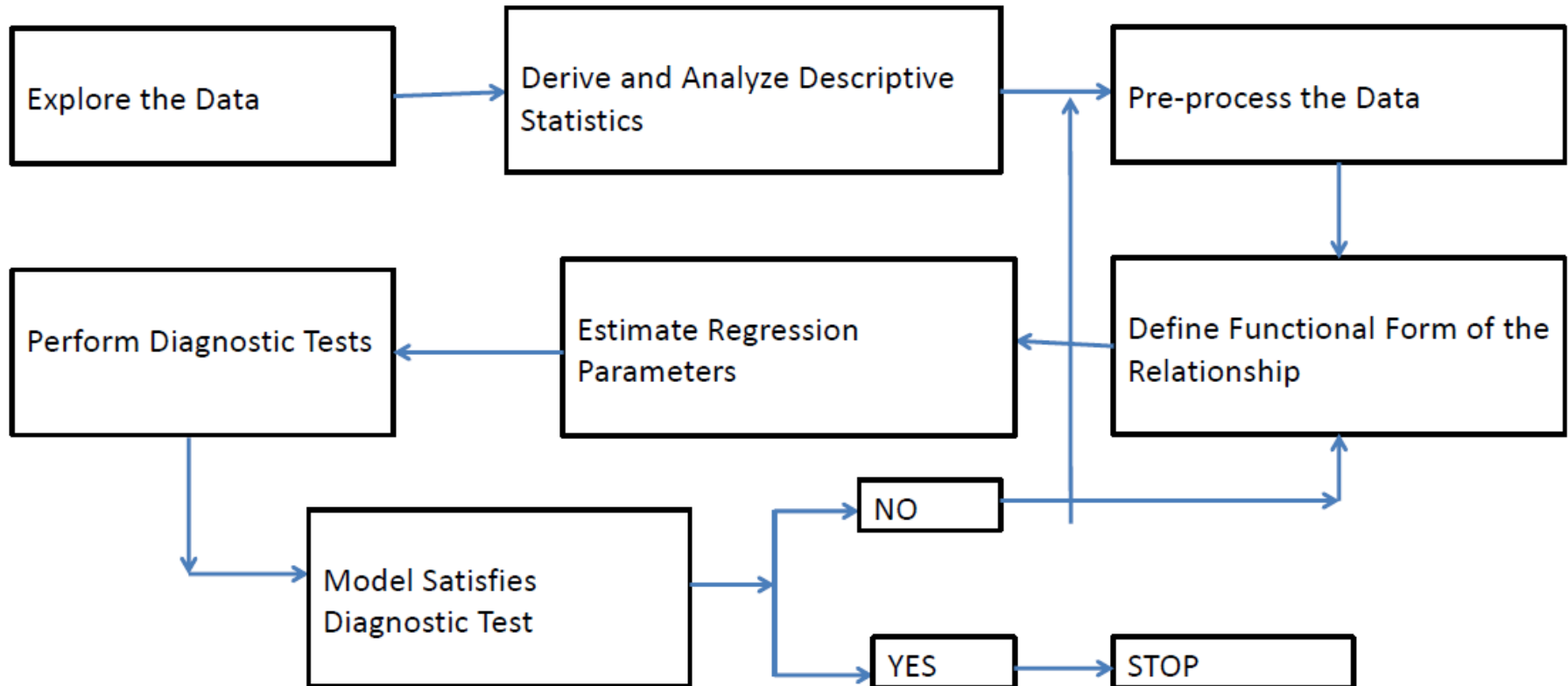
Relationship between 1 dependent & 2 or more independent variables is a **linear function**.



The diagram shows the Multiple Linear Regression Model equation:
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$
 with the following labels and arrows:

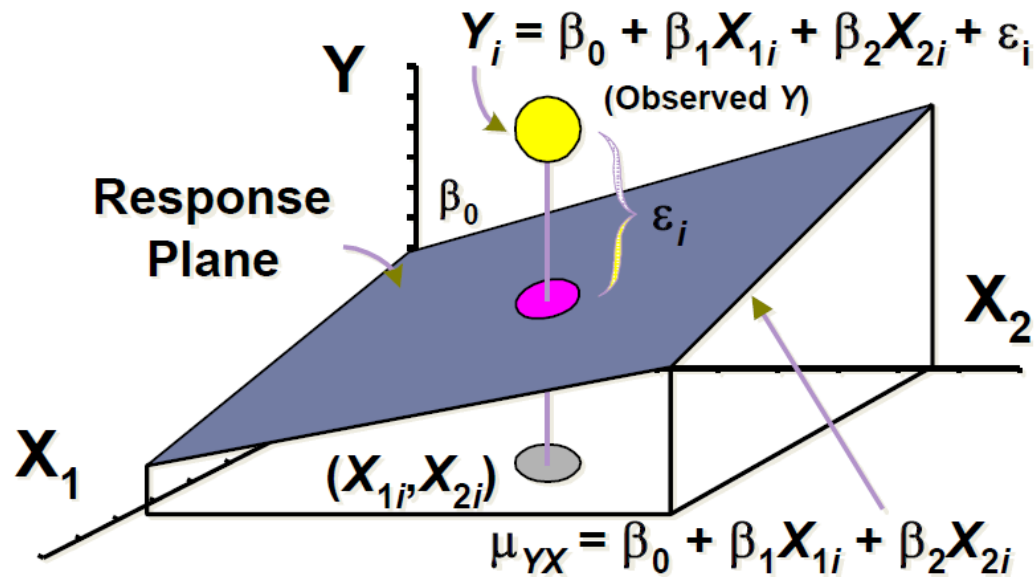
- Population Y-intercept**: Points to β_0
- Population slopes**: Points to $\beta_1, \beta_2, \dots, \beta_k$
- Random error**: Points to ε_i
- Dependent (response) variable**: Points to Y_i
- Independent (explanatory) variables**: Points to $X_{1i}, X_{2i}, \dots, X_{ki}$

MLRM Development Process



MLRM Prediction Equation

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}$$



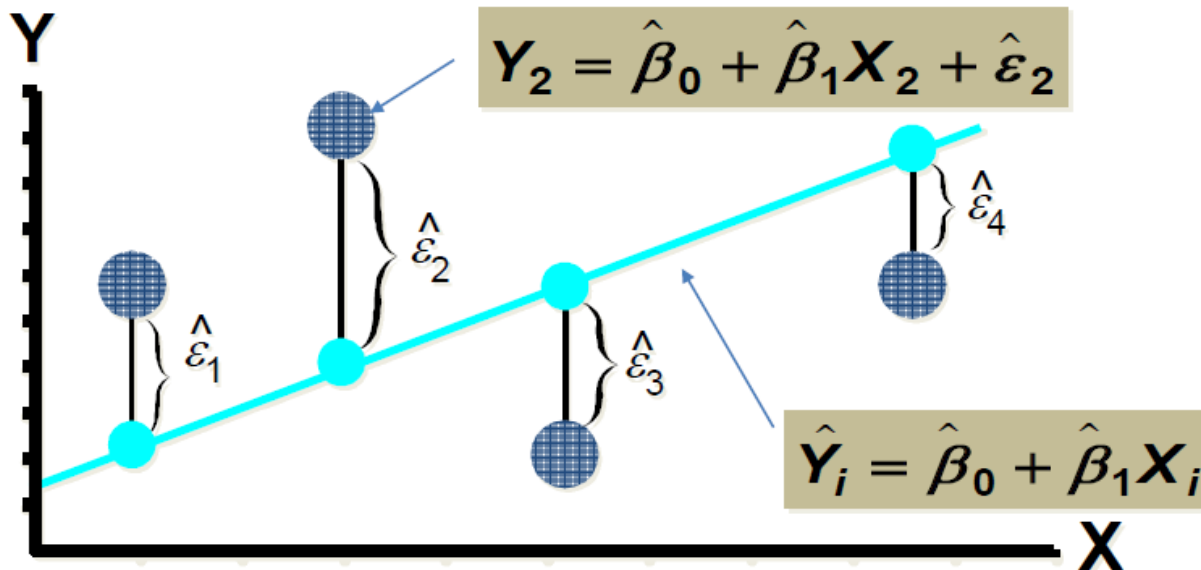
MLRM Model Assumptions

- The error term, e_i , follows a **normal distribution**.
- For different values of X , the variance of e_i is constant(**Homoscedasticity**).
- There is **no Multi-collinearity** (no perfect linear relationship among explanatory variables).
- There is **no autocorrelation** between two e_i values.

MLRM Methodology

Ordinary Least Square (OLS) Methodology, minimizes the sum of square of errors.

$$\text{LS minimizes } \sum_{i=1}^n \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \dots + \hat{\varepsilon}_n^2$$



Interpretation of MLRM Coefficients

- The intercept, β_0 , is the mean value of the dependent variable Y , when the independent variable $X=0$
- The slope, β_i , is the change in the value of the dependent variable, Y , for unit change in the independent variable X_i , keeping all other X s constant (controlled).

MLRM Model Diagnostics

- Test for overall model fitness (R-Square and Adjusted R-Square)
- Test for overall model statistical significance (F test test)
- Test for statistical significance of individual explanatory variables (t test)
- Test for Normality and Homoscedasticity of residuals
- Test for Multi-collinearity and Auto Correlation

Test for overall Model Fitness:

Coefficient of Multiple Determination R^2

- R Square ranges from 0 to 1

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

R^2 = Multiple coefficient of multiple determination

R^2 is the percentage of the variation of y explained by

x_1, x_2, \dots, x_k

Test for overall Model Fitness: Adjusted R²

- Inclusion of additional explanatory variables will increase the R² value
- By introducing an additional explanatory variable, we increase the numerator of the expression for R² while the denominator remains the same.
- To correct this defect, we adjust the R² by taking into account the degrees of freedom

$$R_A^2 = 1 - \frac{SSE/(n - k - 1)}{SST/(n - 1)}$$

R_A^2 = Adjusted R - Square

n = number of observations

k = number of explanatory variables

Test for overall Model Significance: F-Test

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

H_A : Not all β values are zero

Test for overall significance of multiple regression model.

Checks if there is a statistically significant relationship between Y and any of the explanatory variables (X_1, X_2, \dots, X_k).

Analysis of Variance					
Source	D F	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	7E+14	1.48E+14	55.2	<.0001
Error	52	1.40E+14	2.69E+12		
Corrected Total	57	8.80E+14			

Test for individual explanatory variables (t-test)

$$H_0 : \beta_i = 0$$

$$H_A : \beta_i \neq 0$$

$$t = \frac{\hat{\beta}_i}{S_e(\hat{\beta}_i)}$$

T- test:

By rejecting the null hypothesis, we can claim that there is a statistically significant relationship between the response variable Y and explanatory variable X_i .

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	493778	1520007	0.32	0.7466
Ratings	1	651375	90775	7.18	<.0001
Price	1	-1833.67081	127.24677	-14.41	<.0001
Num_new_features	1	547167	85234	6.42	<.0001
Stock_market_ind	1	-105.38867	106.01603	-0.99	0.3248
Market_promo_budget	1	100.92891	8.38744	12.03	<.0001

Test for Normality and Homoscedasticity of Residuals

- **Normality:** The errors should be normally distributed, we test for normality of residuals and the p-value should be more than 0.05 (we accept H_0)
 - **Test:** Anderson-Darling (AD) Test, Shapiro-Wilk normality test
- **Homoscedasticity:** At each level of the predictor variables, the variance of the residual term should be constant and the p-value should be more than 0.05 (we accept H_0)
 - **Test:** Breusch-Pagan Test, Cook-Weisberg Test

Test for Multicollinearity

- **Multicollinearity:** High correlation between explanatory variables is called Multi-collinearity
 - **Test:** Variance Inflation Factor (VIF), is a relative increase in variance in S.E. of beta because of collinearity , VIF should be less than 1.7. To reduce the Multicollinearity, one of the ideal measures is to drop the variable causing multi-collinearity

Test for Auto-Correlation

- **Auto Correlation:** For any two observations of Y_i , error terms should be not correlated
 - **Test:** Durbin Watson Test, test statistic ranges from 0-4. When $DW=2$, no auto-correlation exists. Values away from 2, show that auto-correlation exists.

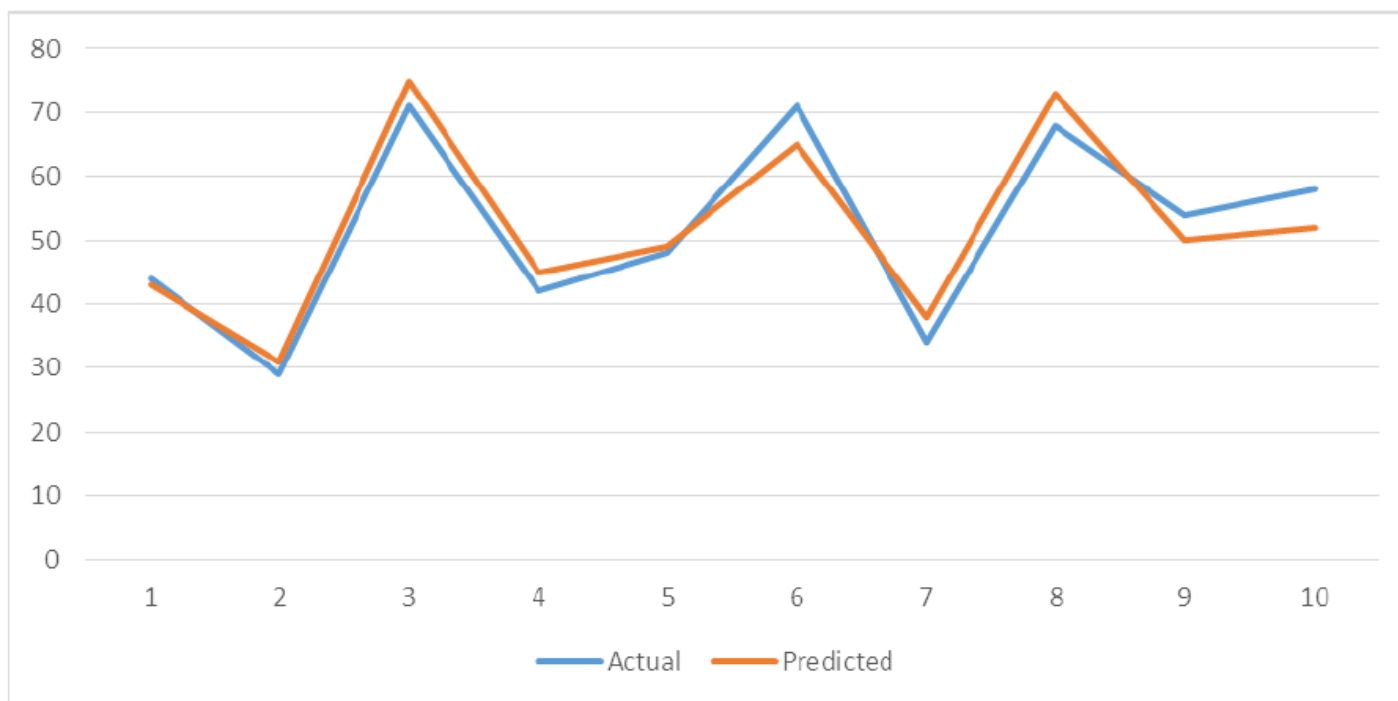
Mean Absolute Percentage Error (MAPE)

The mean absolute percentage error (**MAPE**), also known as mean absolute percentage deviation (MAPD), is a measure of prediction accuracy of a forecasting method in statistics. For example, if the MAPE is 5, on average, the forecast is off by 5%. The equation is:

$$\frac{\sum |(y_t - \hat{y}_t) / y_t|}{n} \times 100, (y_t \neq 0)$$

where y_t equals the actual value, \hat{y}_t equals the fitted value, and n equals the number of observations.

Actual Vs Predicted



Thank You