IMS Health & Quintiles are now IQVIA™

Day - 2
Linear regression analysis

January 9, 2018

What will I learn?

- Concept behind basic regression models
- Difference between correlation and regression
- What is robust model
- How to build a robust regression model and validate the accuracy using R

Cause effect relationship

Regression analysis: To quantify cause – effect relationship

Example:

Let us assume that we are owning a store and in order to increase the sales, we have done lot of advertisement

What to de think is going to happen to the sales after advertisement?

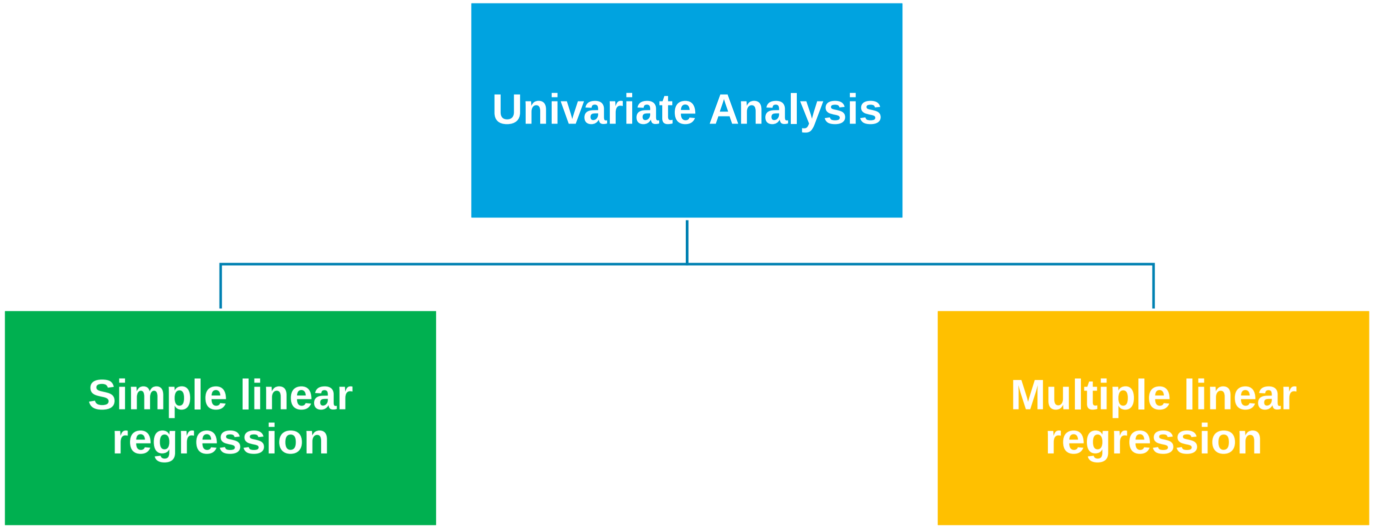
Cause : Advertisement – **Independent variable**

Effect : Increase in sales – **Dependent Variable**

What is quantification?

How much does the sales go up because of advertisement

Types of liner regression analysis



Example :

Type	Dependent Variable	Independent Variable(s)
Simple linear regression	Sales	TV advertisement
Multiple linear regression		TV advertisement , newspaper advertisement and pamphlets

Simple linear regression Equation

The relation ship between the response variable (Y) and predicted variable (X) can be explained on the basis of a **linear model**

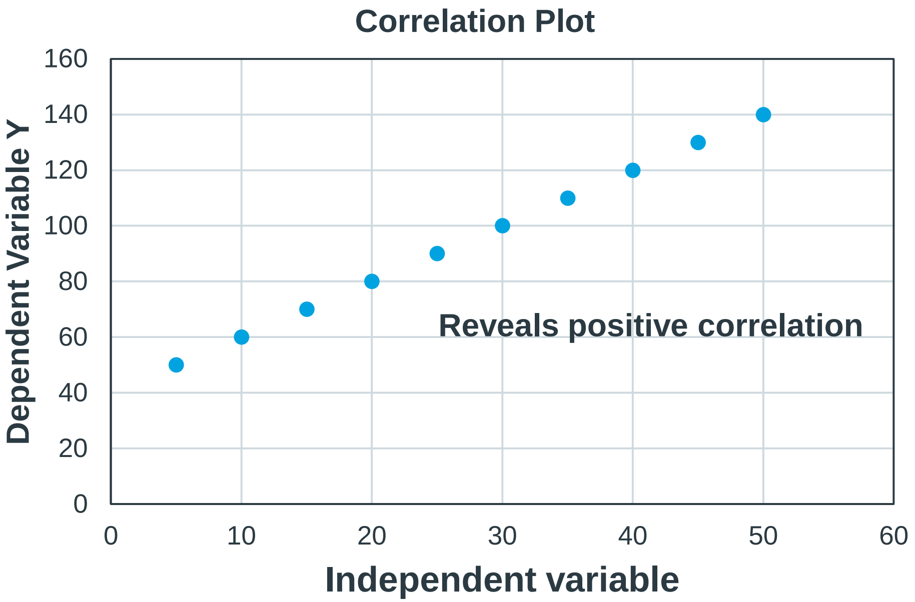
$$Y = a + b(x) + e$$

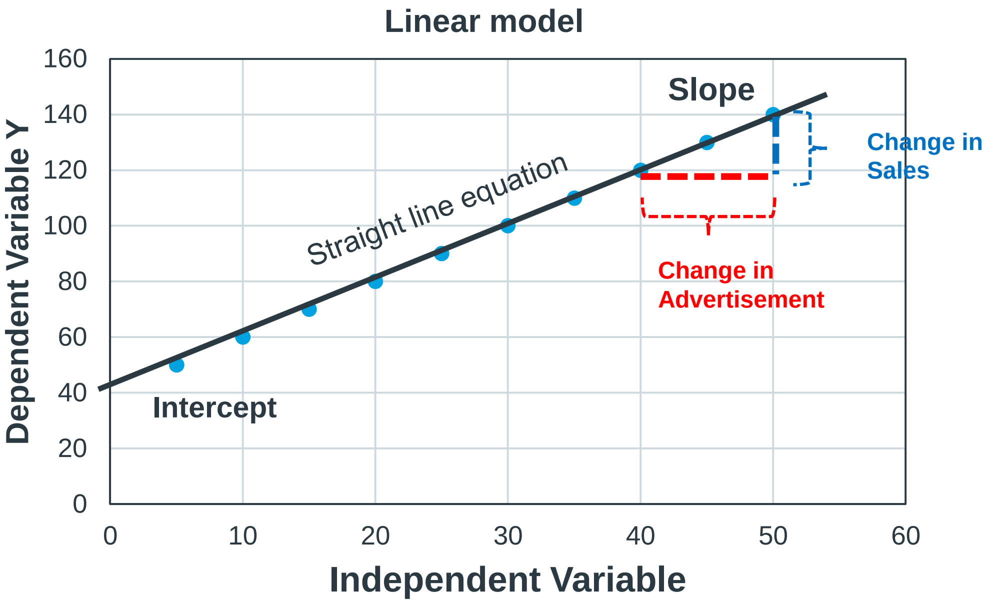
Intercept

Error

Slope

Difference between each data points and the regression line





How linear regression works?

Linear regression works based on Ordinary Least Square method

<u>Observation</u> : For one unit increase in Independent variable, dependent variable increased by two units

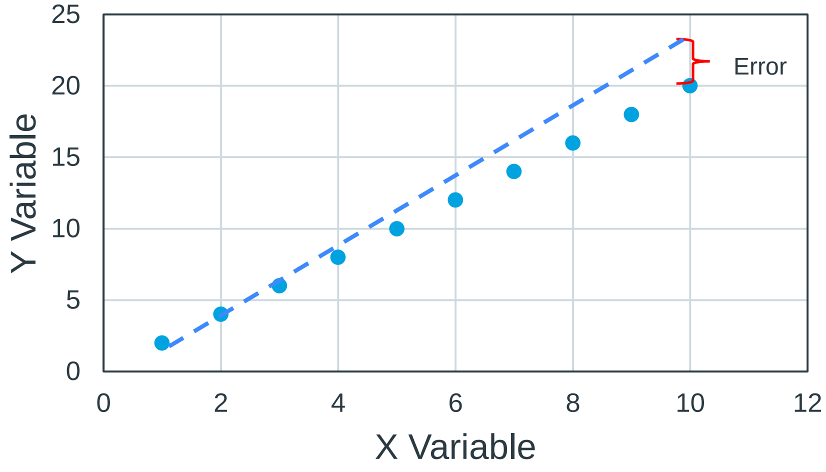
Ordinary Least Square (*OLS*) Estimates

Ordinary Least Regression (OLS) try's to identify ***best possible line*** by minimizing the error

What is best possible line?

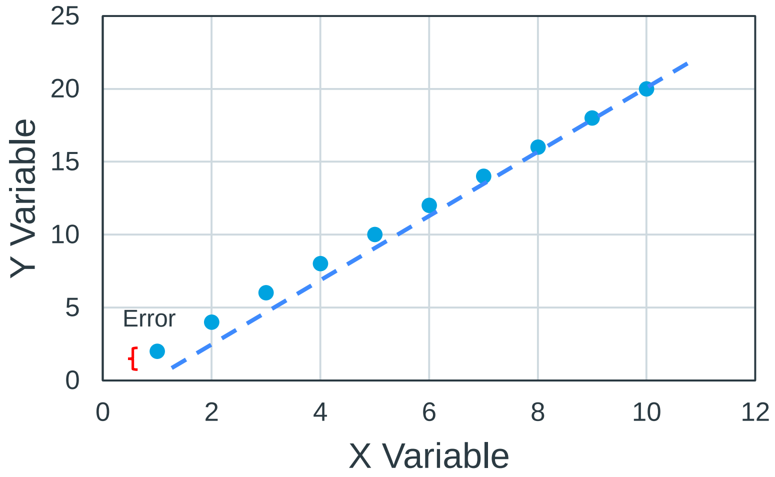
👎

Scenario 1



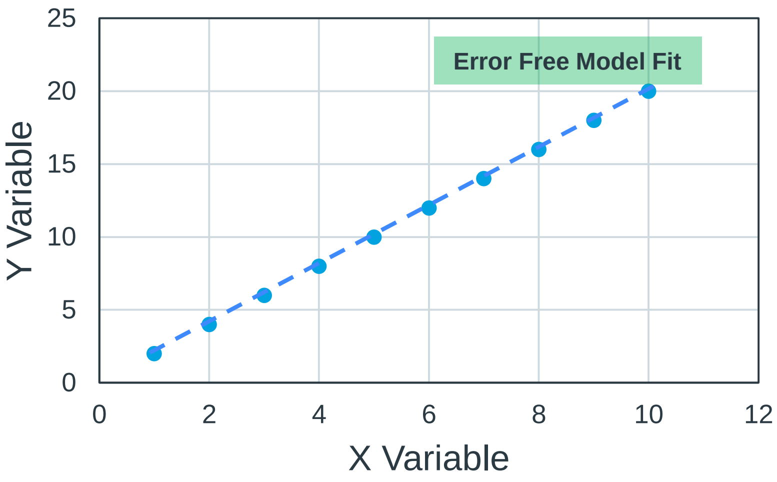
👎

Scenario 2



👍

Scenario 3



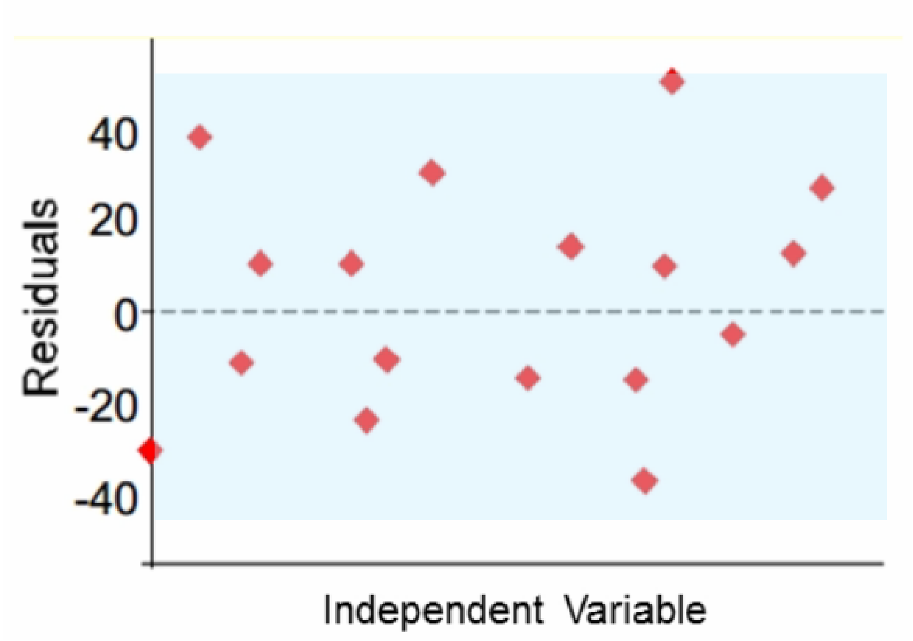
Criteria's for a robust model

- The data should be a random sample of the population
- **Dependent and independent variables should have a linear relationship (Linearity)**
- **Avoid correlation between the independent variables (Avoid Multicollinearity)**
- **Residuals should have constant variance (Homoscedasticity)**
- There should not be any association between the errors

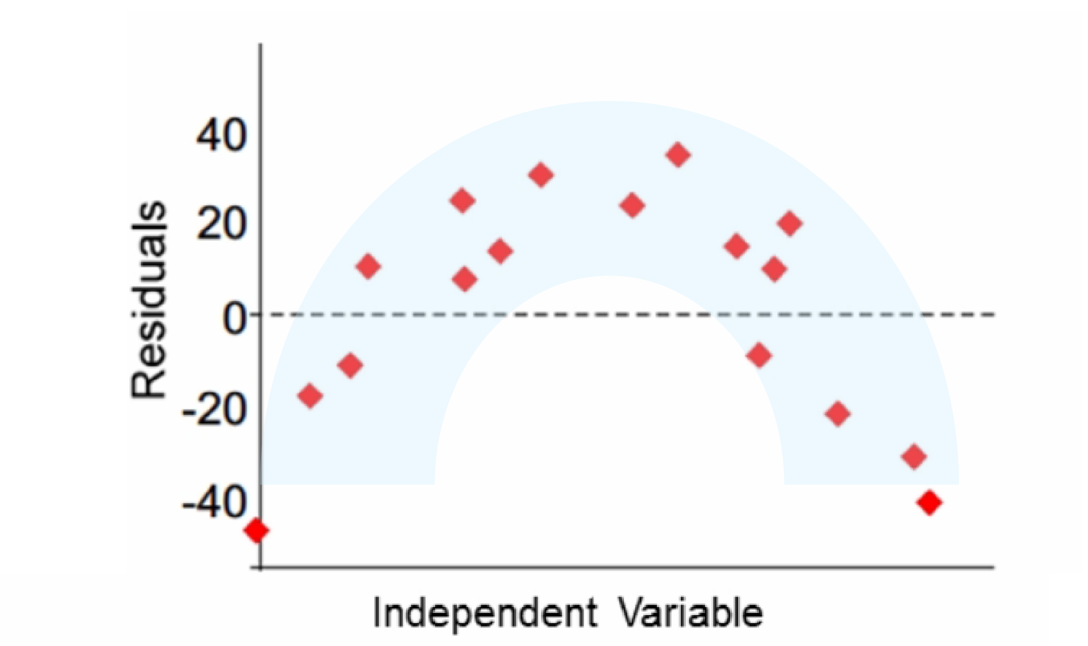
Linearity

Dependent variable should have linear relationship with the independent variable

No Pattern



Curvilinear Pattern



In case of violation : Transformation has to be done / Non-linear regression has to be used

Avoid Multicollinearity

There should not be any association between the independent variables

Variables	**VIF
TV Ad	< 7
Newspaper Ad	< 7
Pamphlet	< 7

👍

👎

Variables	VIF
TV Ad	> 7
Newspaper Ad	> 7
Pamphlet	> 7

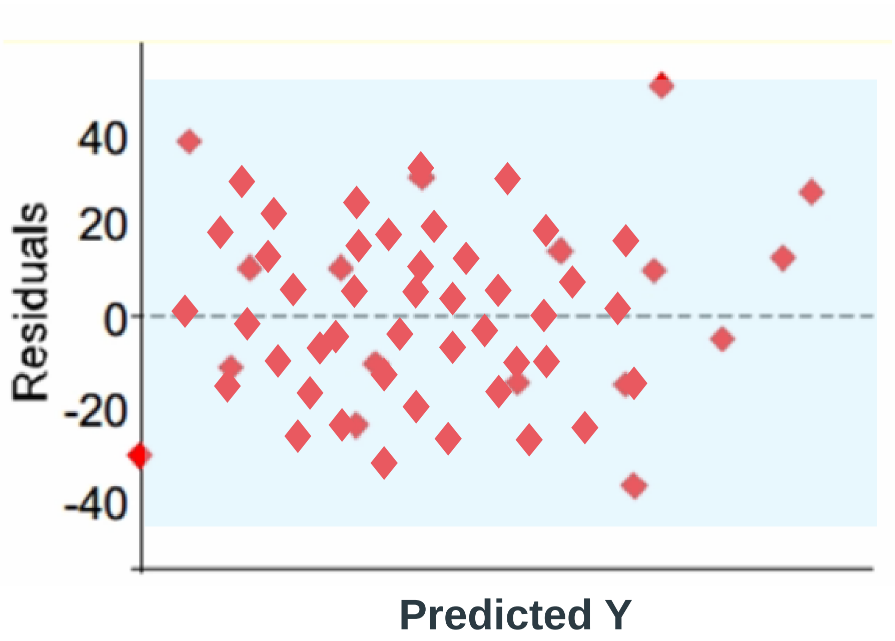
**** Variance Inflation Factor**

In case of violation : Drop the problematic variable / PCA or FA or Ridge regression has to be used

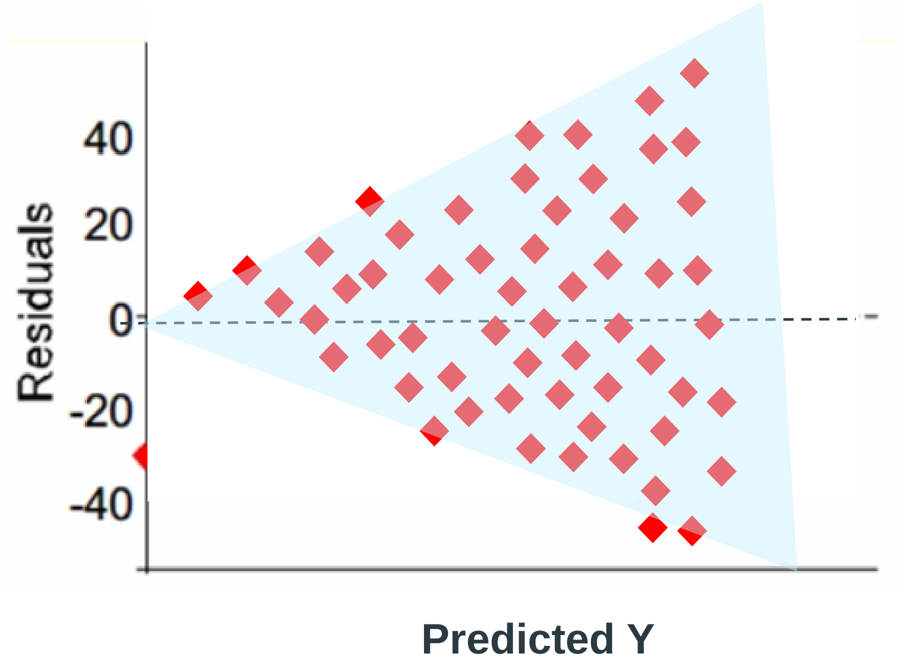
Homoscedasticity

If the residuals have constant variance, we should not see any relationship between residuals and predicted Y

Homoscedasticity



Heteroscedasticity



In case of violation : Transformation has to be done / WLS method has to be used instead of OLS

Performance evaluation vs model validation

Model performance evaluation : It is an assessment of how accurate the model is, and how well it answers the business question framed

Statistical evaluation

- How well is the model “predicting”/”explaining” ?
- **Metric :** Classification table / Confusion matrix

Business evaluation

- Are the relationship captured by the model intuitive and explainable?
- **Metric :** Look for business explanation

Model Validation : It is assessment of how valid and applicable the model is, beyond the sample on which it was generated

Training dataset

- Typically models should be **build on the training data set**

Test dataset

- Developed model should be used on the test data set **to ensure the general applicability of the model**

Performance evaluation technique

Commonly used validation metrics are

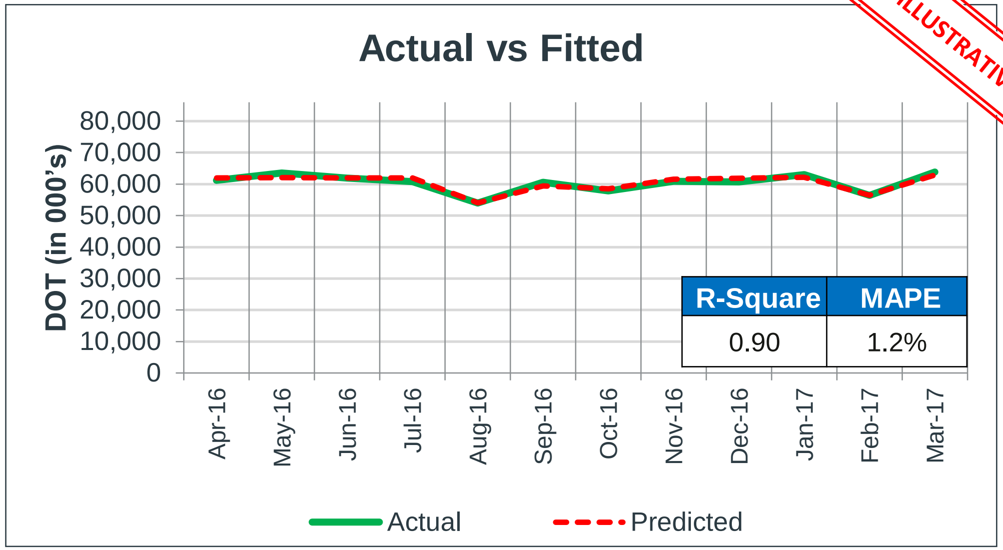
- **R^2** Explains the amount of variation in Y (Dependent Variable) because of X (Independent Variable)

We also look at :

- **Fit Chart – Actual Vs Fitted Values**

- **MAPE – Mean Absolute Percentage Error**

ILLUSTRATIVE

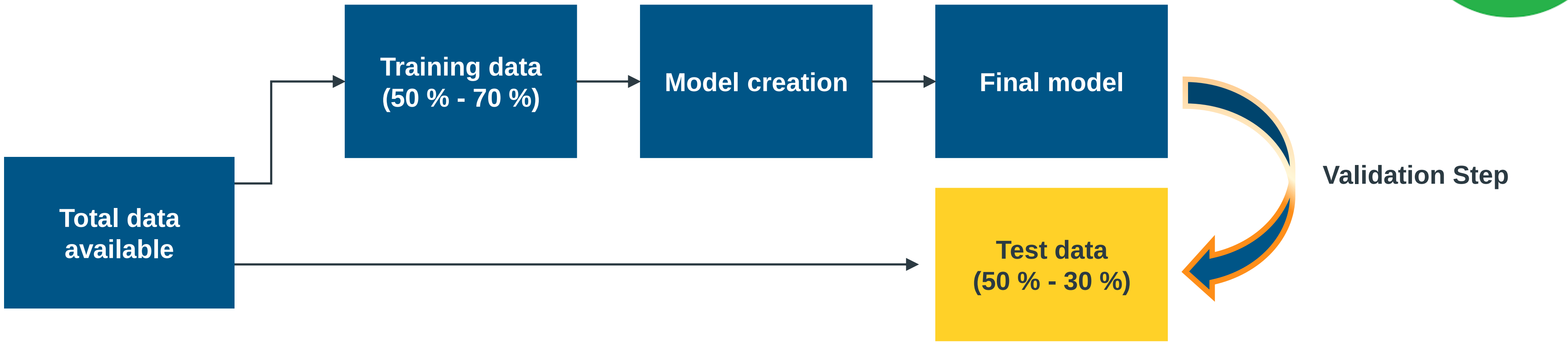


R-Square	MAPE
0.90	1.2%

Model validation technique

It is assessment of how valid and applicable the model is, beyond the sample on which it was generated

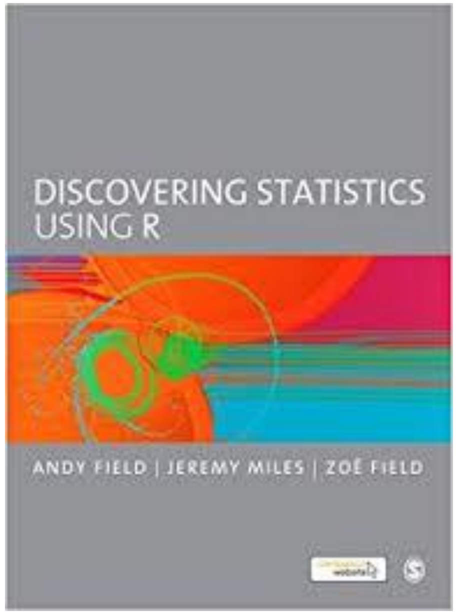
Steps involve in validation process :



Reference :

Discovering Statistics Using R

- Andy Field



DISCOVERING STATISTICS USING R

ANDY FIELD | JEREMY MILES | ZOË FIELD

Practical Session – Implementation in R

Appendix

OLS Estimates

Mathematical calculation

The ordinary least square regression find the best possible line by looking ate the error (or the difference between the points on each line and actual value Y) and minimizing the sum of their squares.

Why sum of square?

To over positive and negative differences

Mathematically, minimize $$Q = \sum_{i=1}^{N} (Y_i - b_0 - b_1 X_i)^2$$

Using differential calculus, we will get

Intercept $$b_o = \frac{\Sigma X_i^2 \Sigma Y_i - \Sigma X_i \Sigma X_i Y_i}{n\Sigma X_i^2 - (\Sigma X_i)^2}$$

Beta coefficients $$b_1 = \frac{n\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i}{n\Sigma X_i^2 - (\Sigma X_i)^2}$$

Please Note : These estimates are called as OLS estimates
(all these calculation will be taken care by R in the back end)

<u>Take home point</u> : If we calculate the Intercept (b_0) and beta coefficient (b_1) by using this formula then the line we generate will automatically be the best possible line.