

Complete Data Science and Machine Learning Using Python

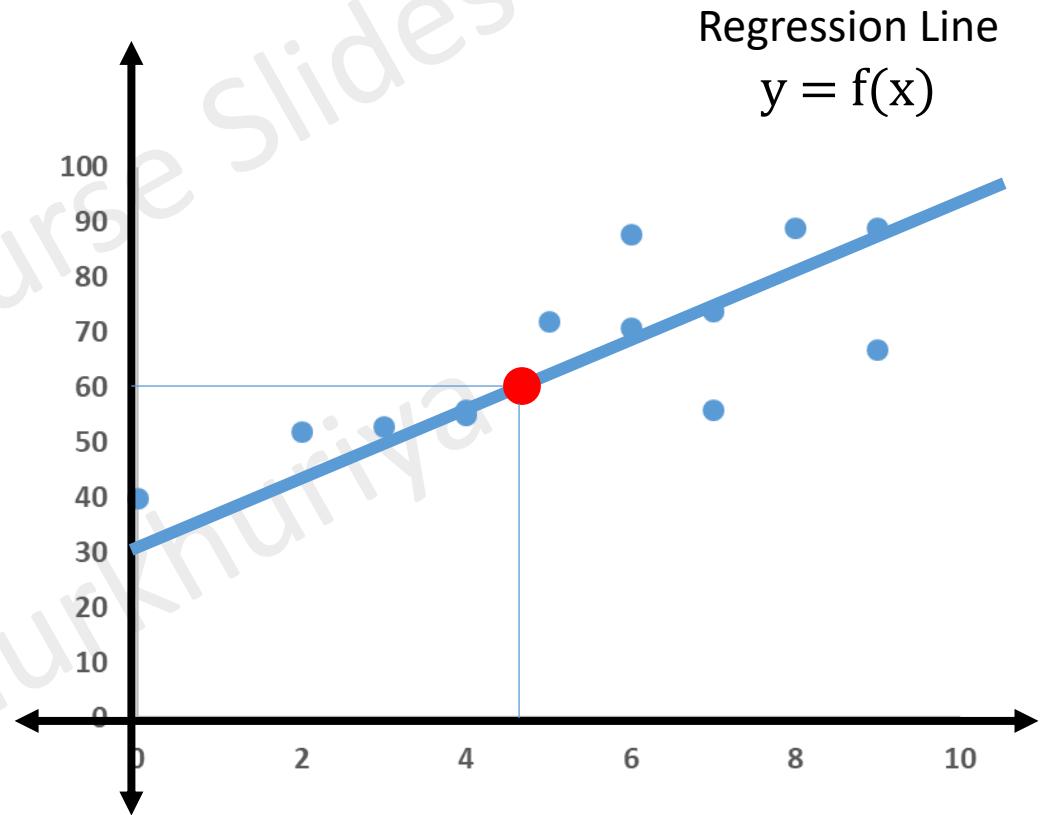
By
Jitesh Khurkhuriya

Linear Regression

Simple Linear Regression

Regression Analysis

- Statistical process for estimating the relationships among variables
- The predictor is a continuous variable
- Relationship between a dependent variable and one or more independent variables (or 'predictors')
- Can also be used to infer causal relationships between dependent and independent variables.



Predicting Continuous Value

REGRESSION

| | |
|----------------------------------|---|
| Ordinal Regression | — Data in Rank Order categories |
| Poisson Regression | — Predicting Event Counts |
| Fast Forest Quantile Regression | — Predicting a Distribution |
| Linear Regression | — Fast Training, Linear Model |
| Bayesian Linear Regression | — Linear Model, Small datasets |
| Neural Network Regression | — Accuracy, Long Training Time |
| Decision Forest Regression | — Accuracy, Fast Training |
| Boosted Decision Tree Regression | — Accuracy, Fast Training, large Memory |

Start

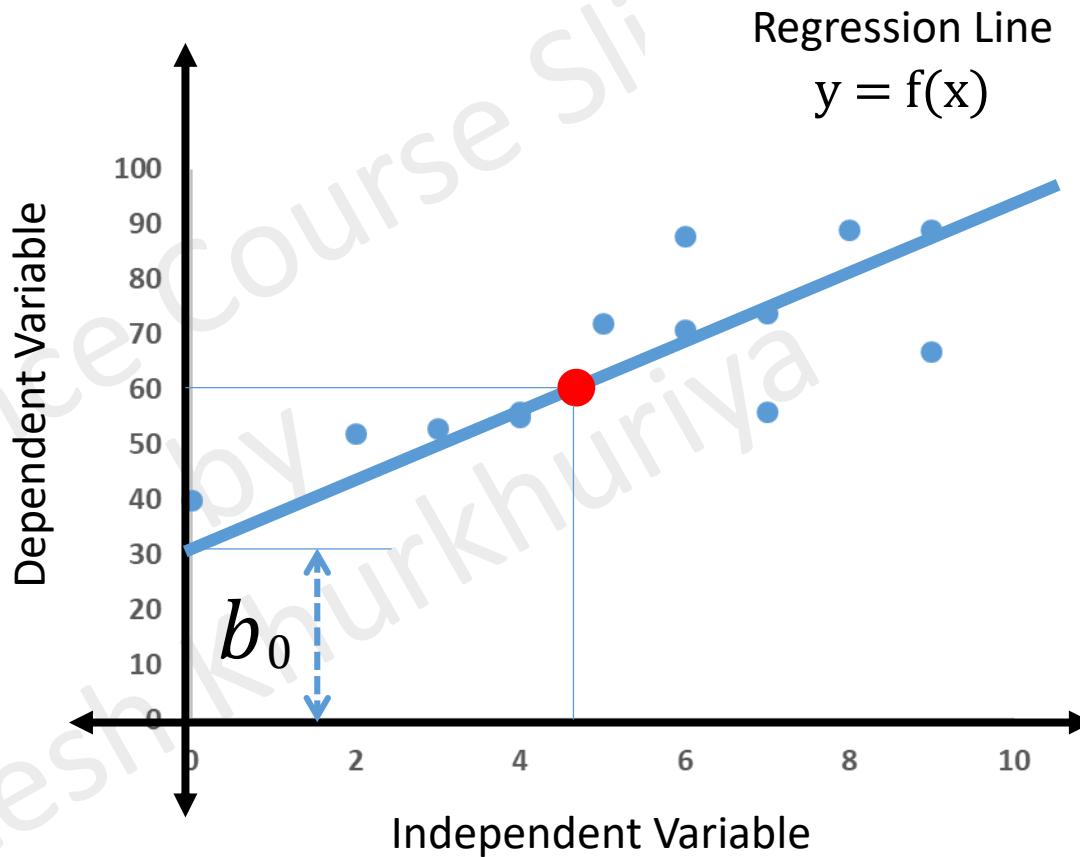
What will be the stock price in future? Should I buy it now?

Simple Linear Regression

Simple Regression :

$$y = b_0 + b_1 x$$

Only one Dependent
Only one Independent



Simple Linear Regression

| Hrs Studied (X) | Marks (Y) |
|--------------------|--------------|
| 0 | 40 |
| 2 | 52 |
| 3 | 53 |
| 4 | 55 |
| 4 | 56 |
| 5 | 72 |
| 6 | 71 |
| 6 | 88 |
| 7 | 56 |
| 7 | 74 |
| 8 | 89 |
| 9 | 67 |
| 9 | 89 |
| 5.38 | 66.31 |
| Mean | |

| X - Mean (A) | Y - Mean (B) | A^2 | A*B |
|-----------------|-----------------|-------|--------|
| -5.38 | -26.31 | 28.99 | 141.66 |
| -3.38 | -14.31 | 11.46 | 48.43 |
| -2.38 | -13.31 | 5.69 | 31.73 |
| -1.38 | -11.31 | 1.92 | 15.66 |
| -1.38 | -10.31 | 1.92 | 14.27 |
| -0.38 | 5.69 | 0.15 | -2.19 |
| 0.62 | 4.69 | 0.38 | 2.89 |
| 0.62 | 21.69 | 0.38 | 13.35 |
| 1.62 | -10.31 | 2.61 | -16.65 |
| 1.62 | 7.69 | 2.61 | 12.43 |
| 2.62 | 22.69 | 6.84 | 59.35 |
| 3.62 | 0.69 | 13.07 | 2.50 |
| 3.62 | 22.69 | 13.07 | 82.04 |
| | | 89.08 | 405.46 |
| Sum | | | |

$$y = b_0 + b_1 x$$

$$b_1 = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2}$$

$$= 405.46 / 89.08$$

$$= 4.55$$

Simple Linear Regression

| Hrs Studied (X) | Marks (Y) |
|-----------------|-----------|
| 0 | 40 |
| 2 | 52 |
| 3 | 53 |
| 4 | 55 |
| 4 | 56 |
| 5 | 72 |
| 6 | 71 |
| 6 | 88 |
| 7 | 56 |
| 7 | 74 |
| 8 | 89 |
| 9 | 67 |
| 9 | 89 |
| 5.38 | 66.31 |
| Mean | |

$$y = b_0 + b_1 x$$

$$b_1 = 4.55$$

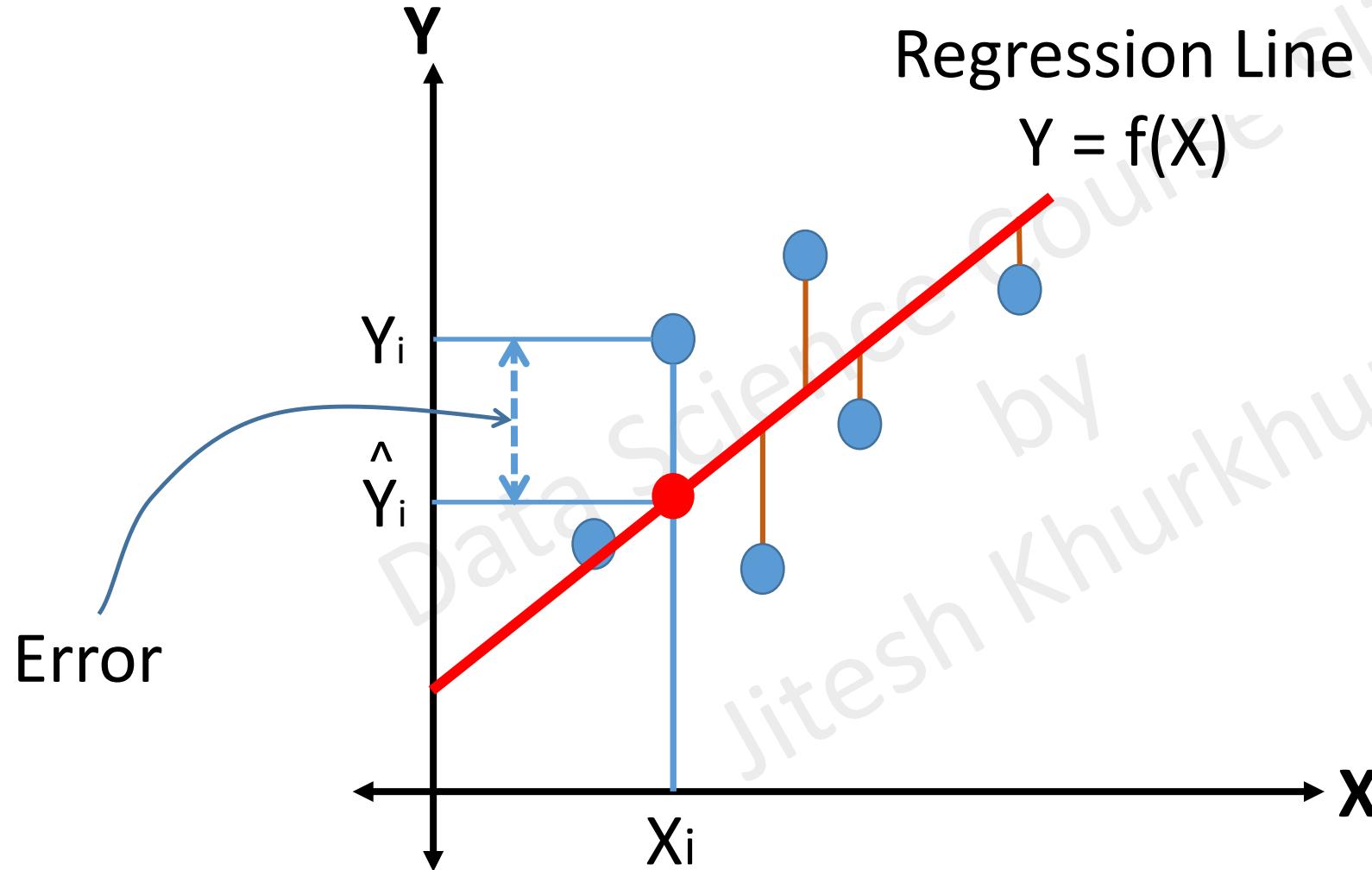
$$b_0 = 41.8$$

$$y = 41.8 + 4.55 x$$



Ordinary Least Square and Error Terms

Ordinary Least Square

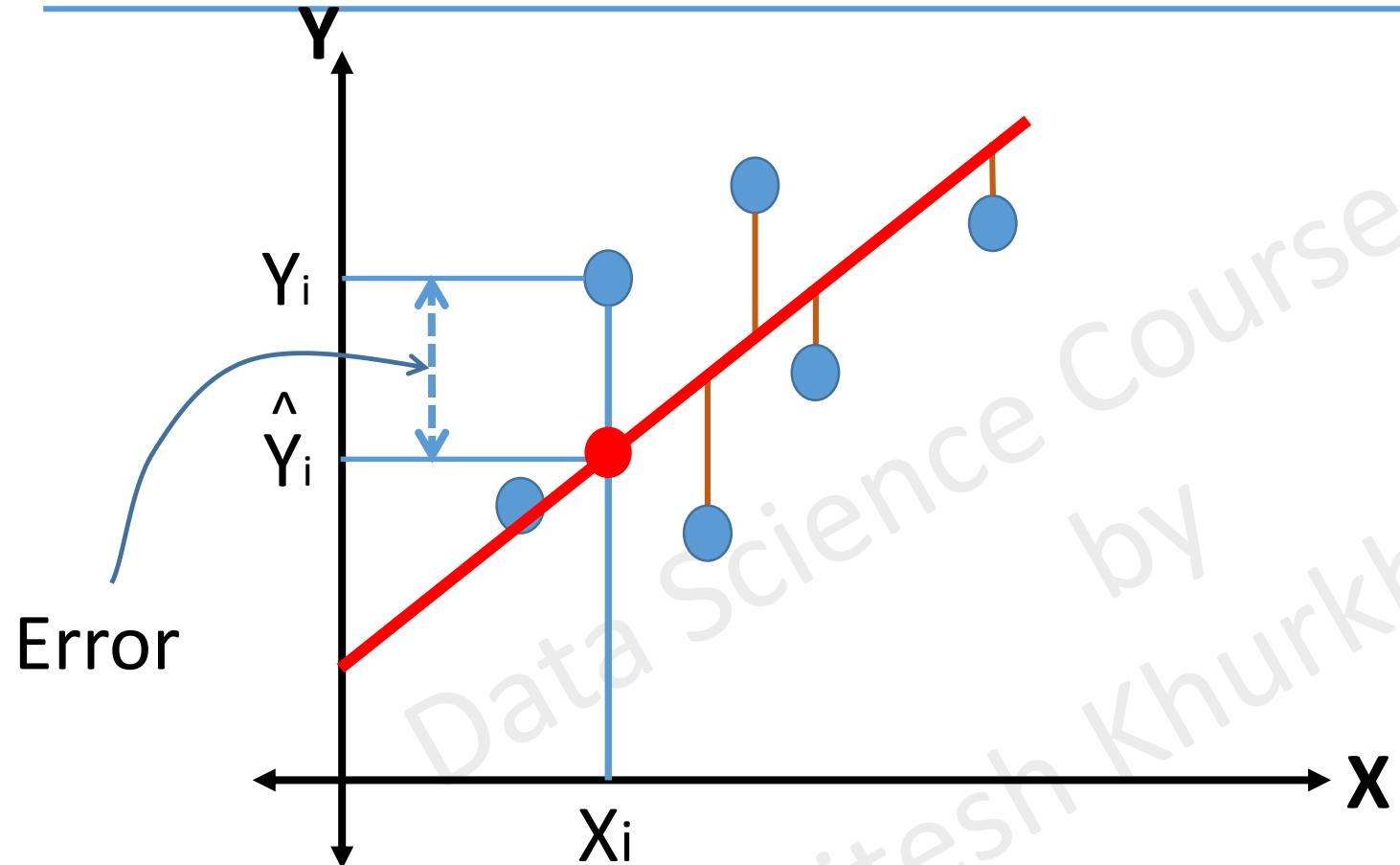


Regression Line
 $Y = f(X)$

Minimum

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Mean Absolute Error



$$\text{MAE} = \frac{1}{n} \sum_{i=0}^n |y_i - \hat{y}_i|$$

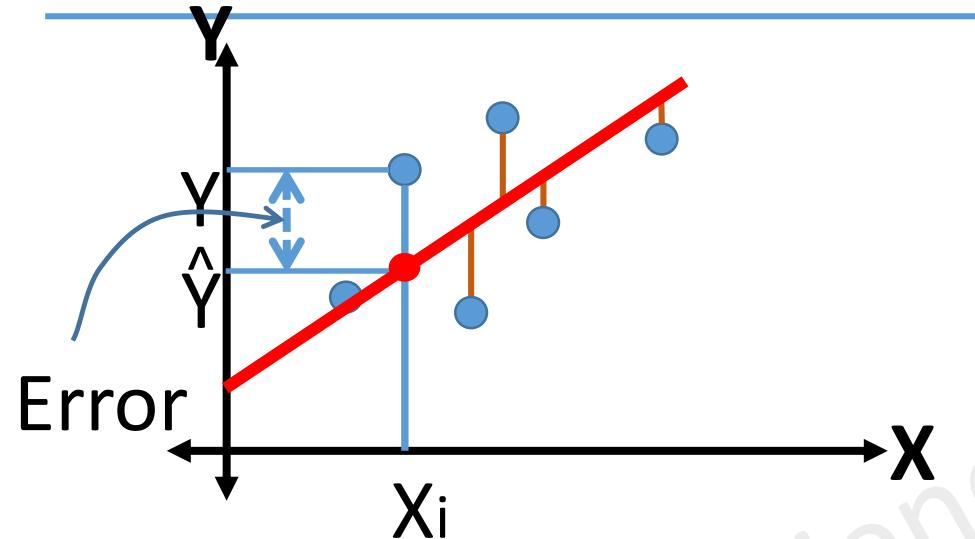
Mean absolute error (MAE) is a quantity used to measure how close forecasts or predictions are to the eventual outcomes.

Root Mean Squared Error

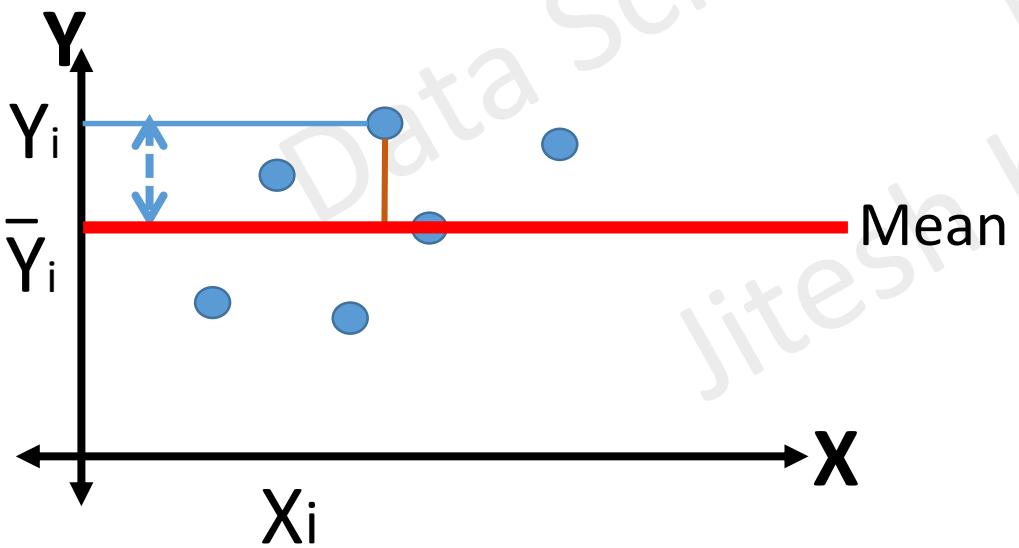
$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=0}^n (y_i - \hat{y}_i)^2}$$

- Very commonly used and makes for an excellent general purpose error metric for numerical predictions.
- Compared to the similar Mean Absolute Error, RMSE amplifies and severely punishes large errors.

Relative Absolute Error



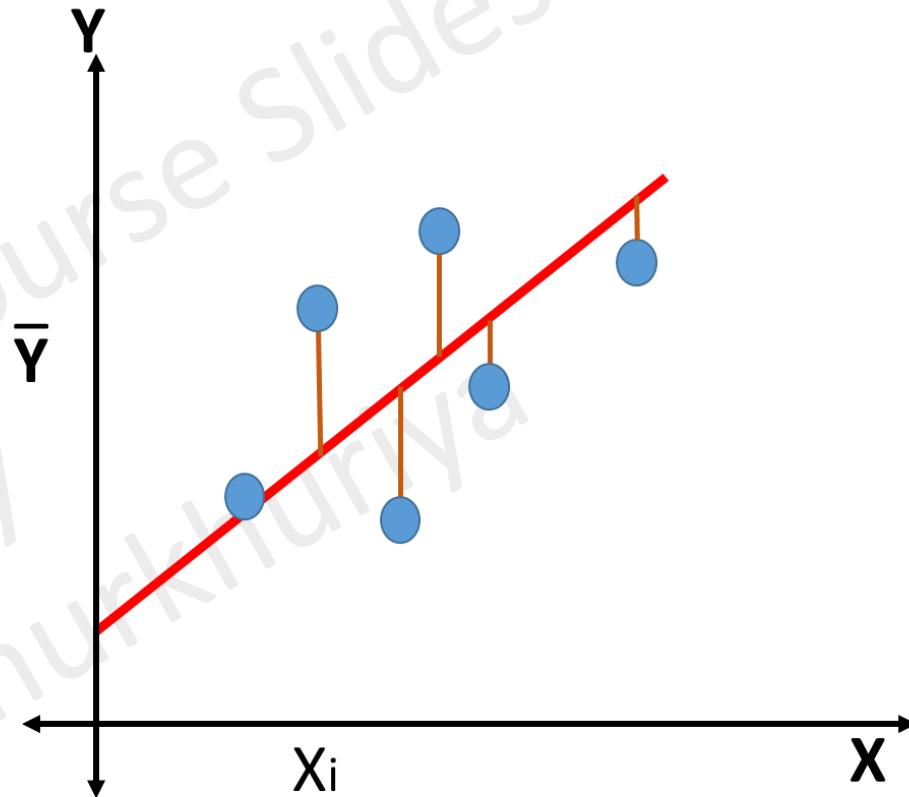
$$\text{RAE} = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{\sum_{i=1}^n |y_i - \bar{y}_i|}$$



R Squared or Coefficient of Determination

Coefficient of Determination

How much (what %) of variation in Y is described by the variation in X?



R-Square With an Example

| Hrs Studied (X) | Marks (Y) |
|-----------------|-----------|
| 0 | 40 |
| 2 | 52 |
| 3 | 53 |
| 4 | 55 |
| 4 | 56 |
| 5 | 72 |
| 6 | 71 |
| 6 | 88 |
| 7 | 56 |
| 7 | 74 |
| 8 | 89 |
| 9 | 67 |
| 9 | 89 |
| 5.38 | 66.31 |
| Mean | |

$$Y = 41.8 + 4.55X$$

| Predicted Marks \hat{Y} |
|------------------------------|
| 41.80 |
| 50.90 |
| 55.45 |
| 60.00 |
| 60.00 |
| 64.55 |
| 69.10 |
| 69.10 |
| 73.65 |
| 73.65 |
| 78.20 |
| 82.75 |
| 82.75 |

| $(Y - \bar{Y})^2$ | $(\hat{Y} - \bar{Y})^2$ |
|-------------------|-------------------------|
| 692.22 | 600.74 |
| 204.78 | 237.47 |
| 177.16 | 117.94 |
| 127.92 | 39.82 |
| 106.30 | 39.82 |
| 32.38 | 3.10 |
| 22.00 | 7.78 |
| 470.46 | 7.78 |
| 106.30 | 53.88 |
| 59.14 | 53.88 |
| 514.84 | 141.37 |
| 0.48 | 270.27 |
| 514.84 | 270.27 |
| 3028.77 | 1844.12 |
| SST | SSR |

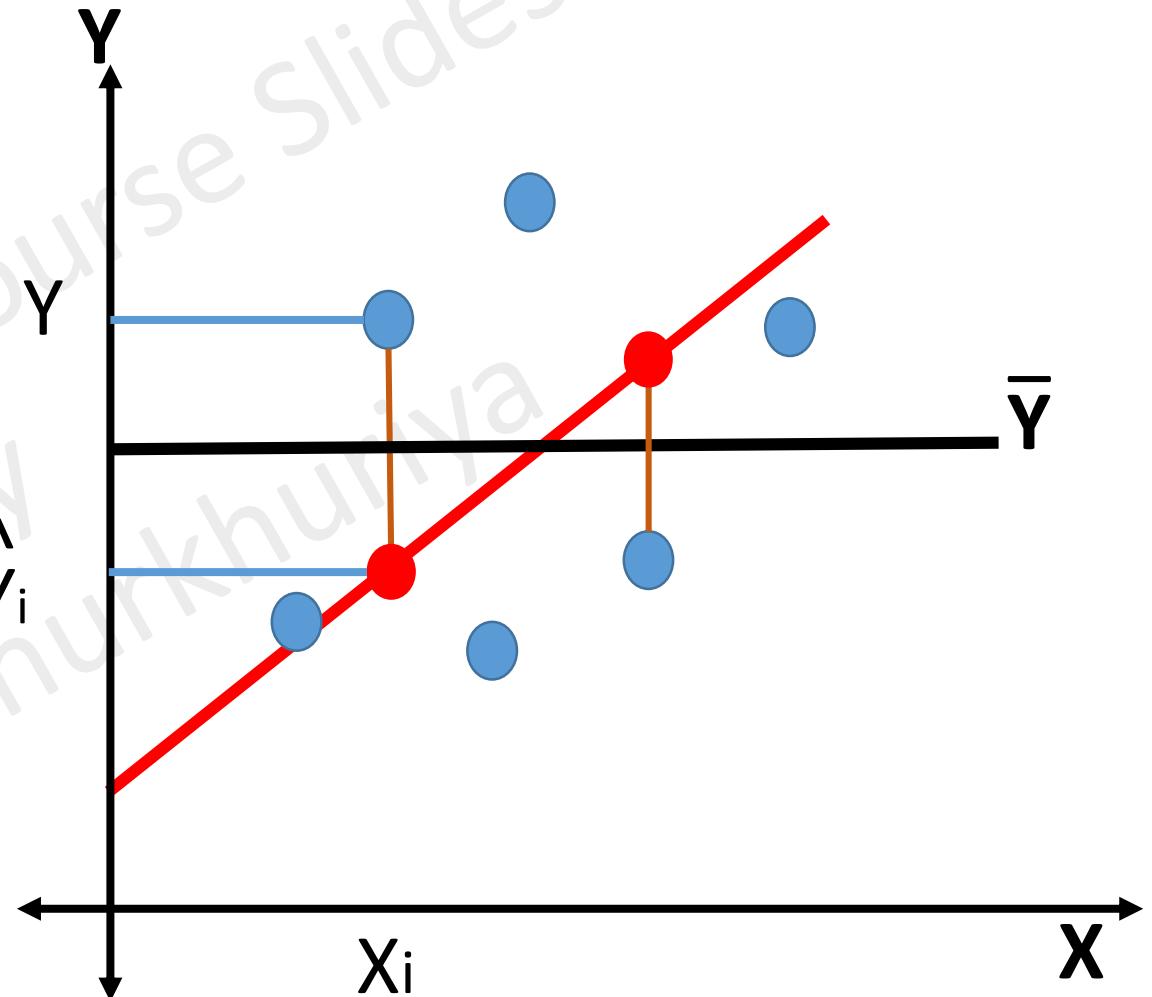
Coefficient of Determination

Sum of Squares Due to Regression

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Total Sum of Squares

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$



R-Square With an Example

| Hrs Studied (X) | Marks (Y) |
|-----------------|-----------|
| 0 | 40 |
| 2 | 52 |
| 3 | 53 |
| 4 | 55 |
| 4 | 56 |
| 5 | 72 |
| 6 | 71 |
| 6 | 88 |
| 7 | 56 |
| 7 | 74 |
| 8 | 89 |
| 9 | 67 |
| 9 | 89 |
| 5.38 | 66.31 |
| Mean | |

$$\begin{aligned} R^2 &= \frac{SSR}{SST} \\ &= \frac{1844.12}{3028.77} \\ &= 0.60886 \end{aligned}$$

Higher the value → Variation in Y is explained by variation in X.

| $(Y - \bar{Y})^2$ | $(\hat{Y} - \bar{Y})^2$ |
|-------------------|-------------------------|
| 692.22 | 600.74 |
| 204.78 | 237.47 |
| 177.16 | 117.94 |
| 127.92 | 39.82 |
| 106.30 | 39.82 |
| 32.38 | 3.10 |
| 22.00 | 7.78 |
| 470.46 | 7.78 |
| 106.30 | 53.88 |
| 59.14 | 53.88 |
| 514.84 | 141.37 |
| 0.48 | 270.27 |
| 514.84 | 270.27 |
| 3028.77 | 1844.12 |
| SST | SSR |

Multiple Linear Regression

Multiple Linear Regression

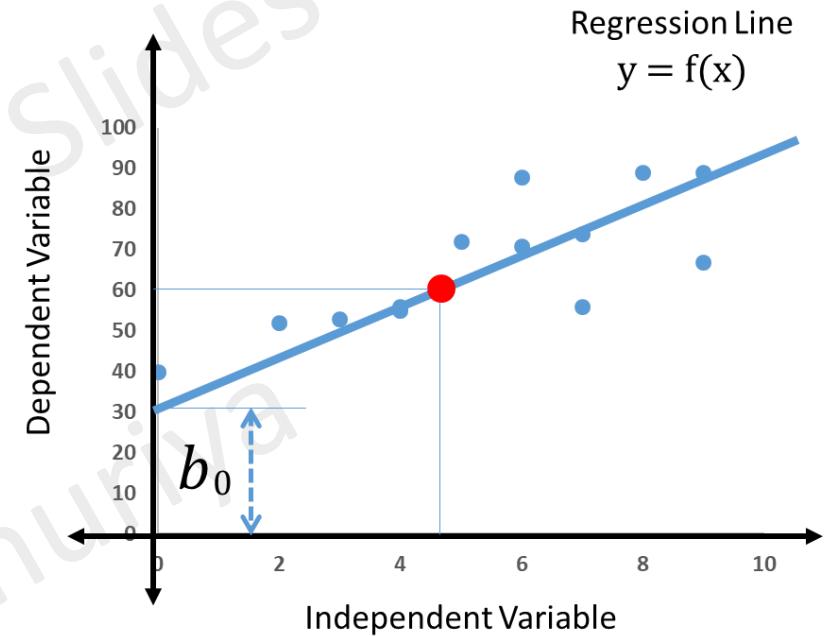
Simple Regression :

$$y = b_0 + b_1 x$$

Only one Dependent
Only one Independent

Multiple Linear Regression :

$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n$$



Multiple Linear Regression

| Hrs Studied (X ₁) | Hrs Slept (X ₂) | Marks (Y) |
|----------------------------------|--------------------------------|--------------|
| 0 | 8 | 40 |
| 2 | 8 | 52 |
| 3 | 7.5 | 53 |
| 4 | 7 | 55 |
| 4 | 9 | 56 |
| 5 | 8.5 | 72 |
| 6 | 9 | 71 |
| 6 | 7 | 88 |
| 7 | 6 | 56 |
| 7 | 7 | 74 |
| 8 | 9 | 89 |
| 9 | 6 | 67 |
| 9 | 9 | 89 |

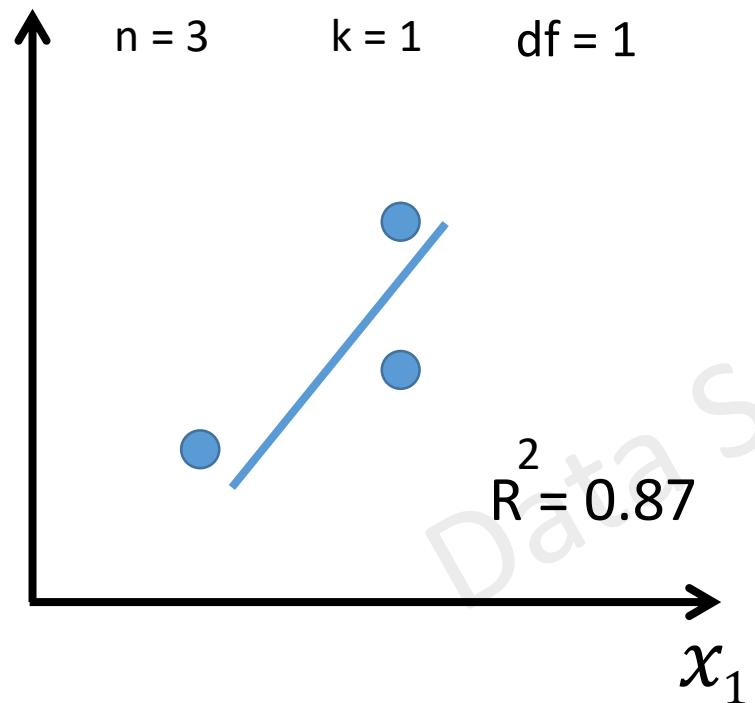
$$y = b_0 + b_1 x_1 + b_2 x_2$$

Dependent Variable
Marks Obtained

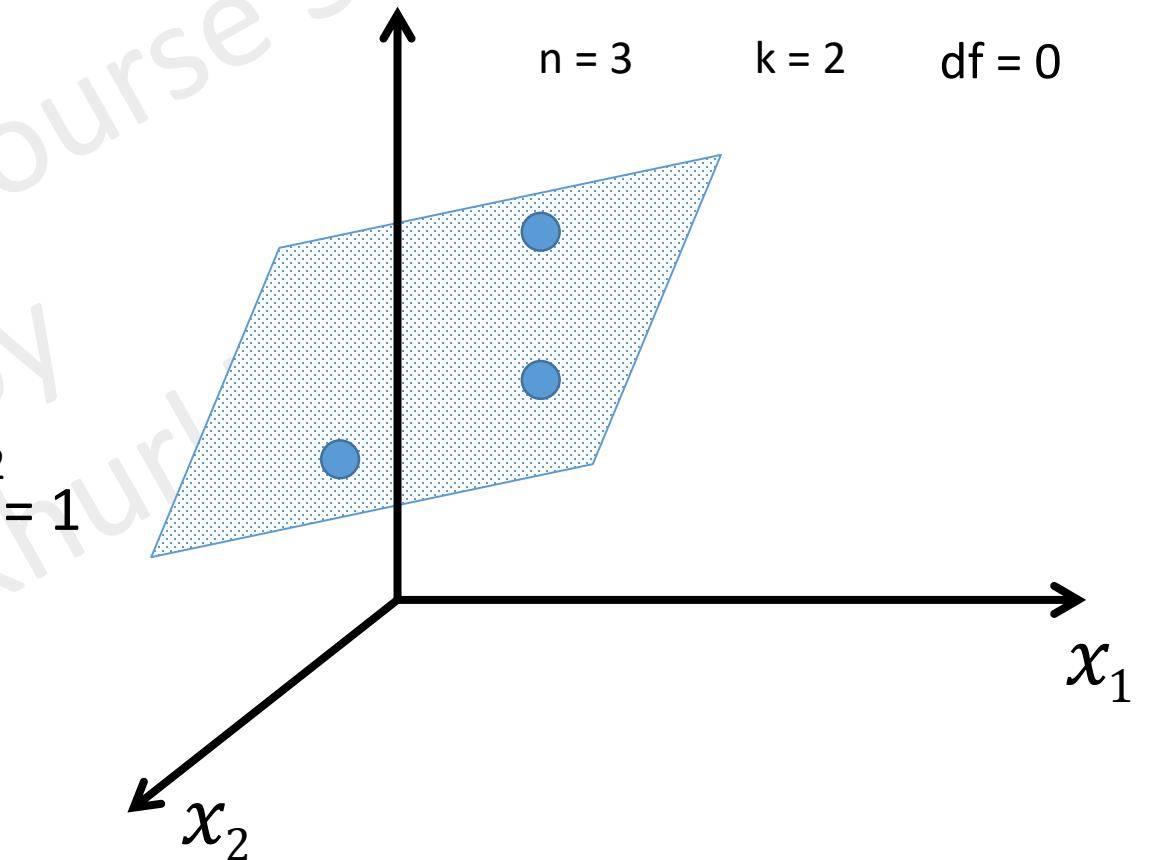
Independent Variable
Hrs Studied
Hrs Slept

Degrees of Freedom ($n - p - 1$)

$$y = b_0 + b_1 x_1$$



$$y = b_0 + b_1 x_1 + b_2 x_2$$



Adjusted R-Squared

$$\bar{R}^2 = 1 - \frac{(1 - R^2) * (n - 1)}{n - p - 1}$$

R = Sample R-Squared

p = Number of independent variables

n = sample size or number of observations

Assumptions of Multiple Linear Regression

Relationship Among Variables

- Linear Relationship
- Multicollinearity
- No Auto-Correlation
- Endogeneity

Behaviour of Data

- Sample Size
- Normality
- Homoscedasticity

Degrees of Freedom

Degrees of Freedom in Statistics

The number of values in the final calculation of a statistic that are free to vary.

OR

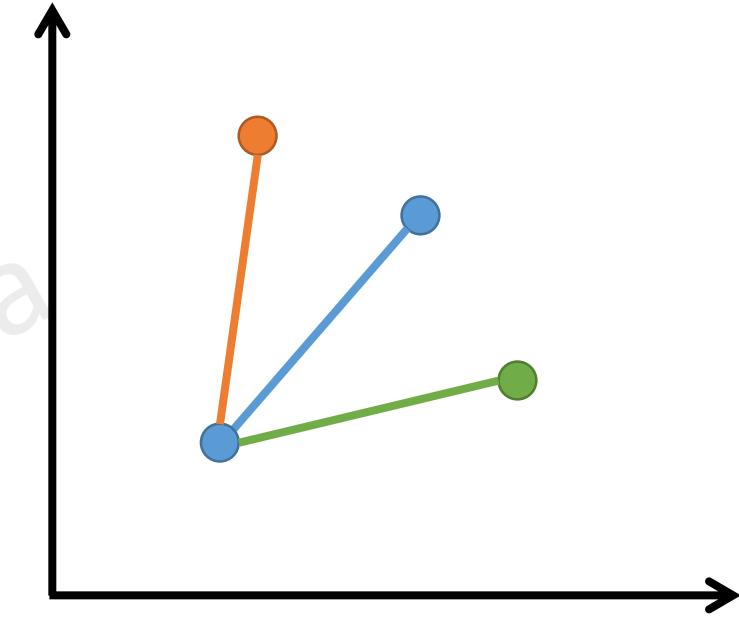
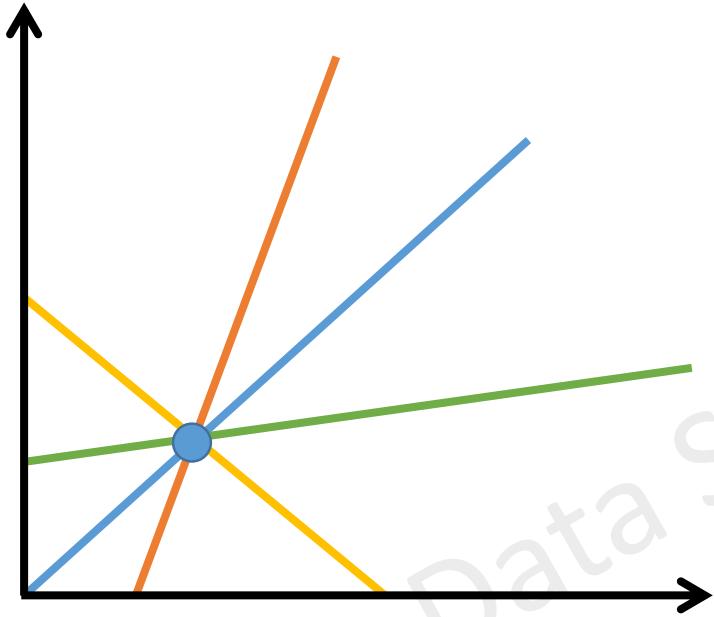
The minimum number of independent coordinates that can specify the position of the system completely.

$$df = n - p - 1$$

Number of Observations

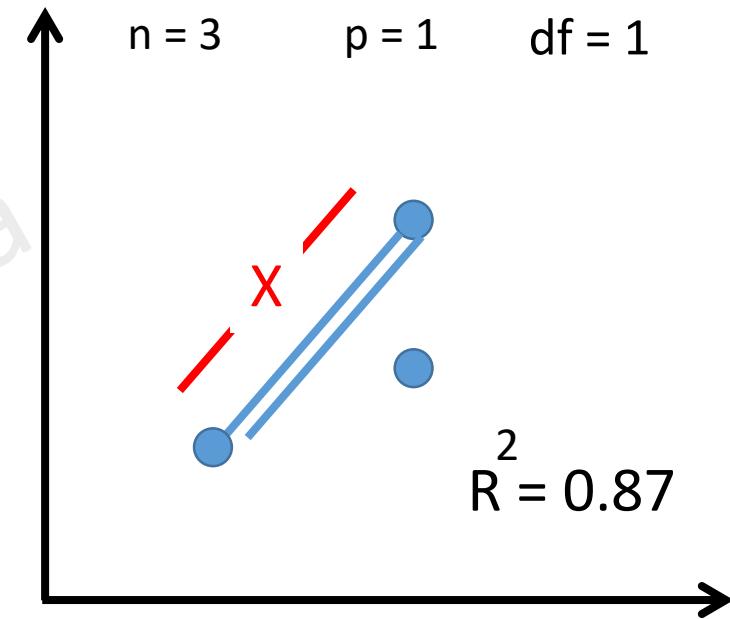
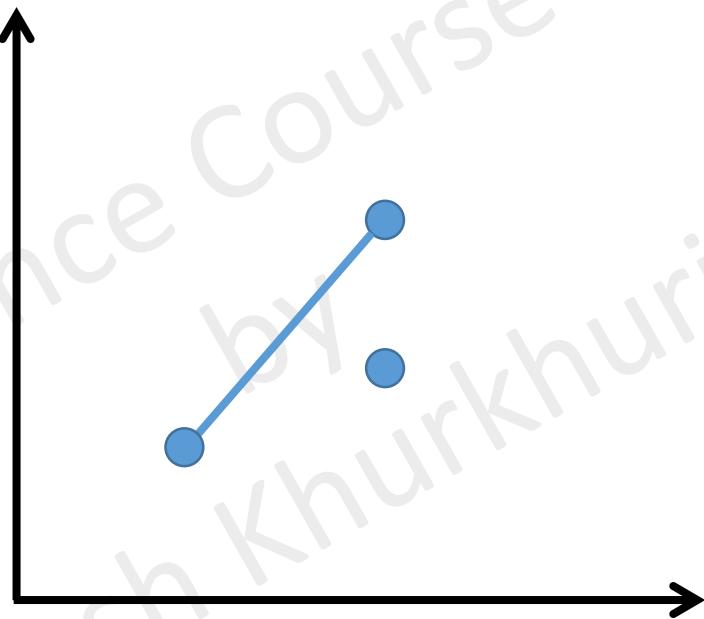
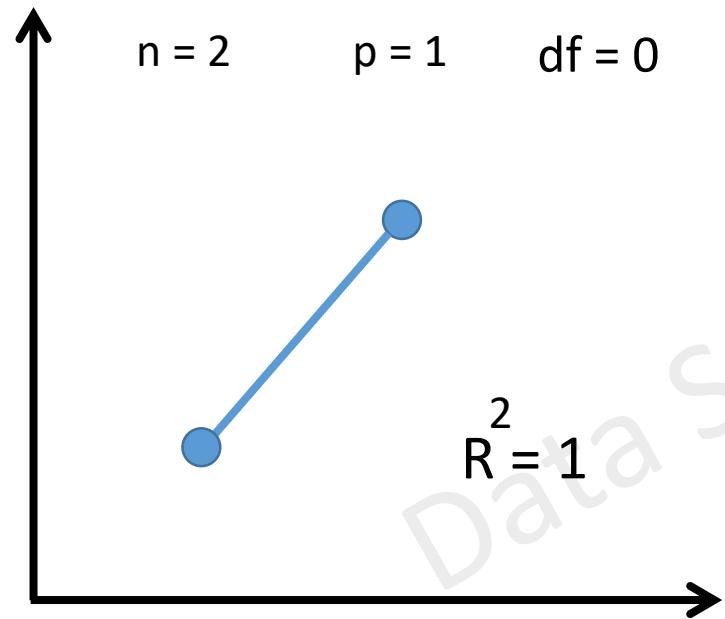
Number of variables

Degrees of Freedom in Statistics



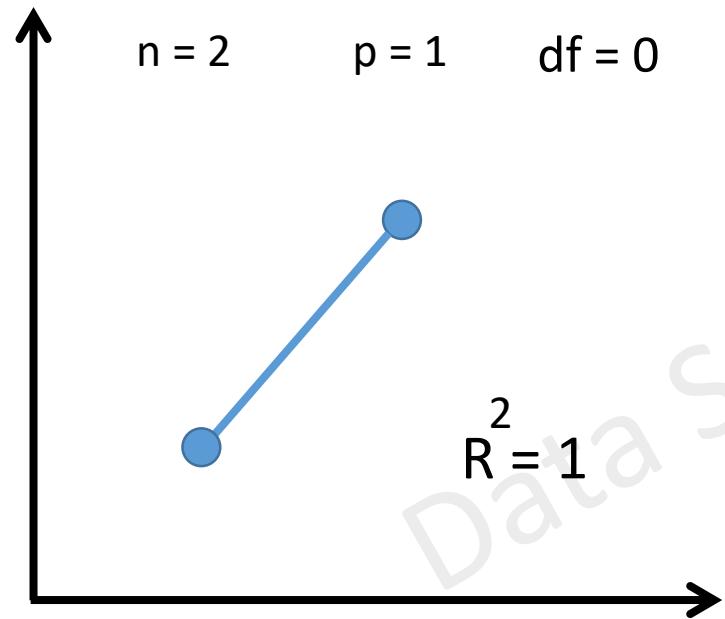
Degrees of Freedom in Statistics ($n - p - 1$)

$$y = b_0 + b_1 x_1$$

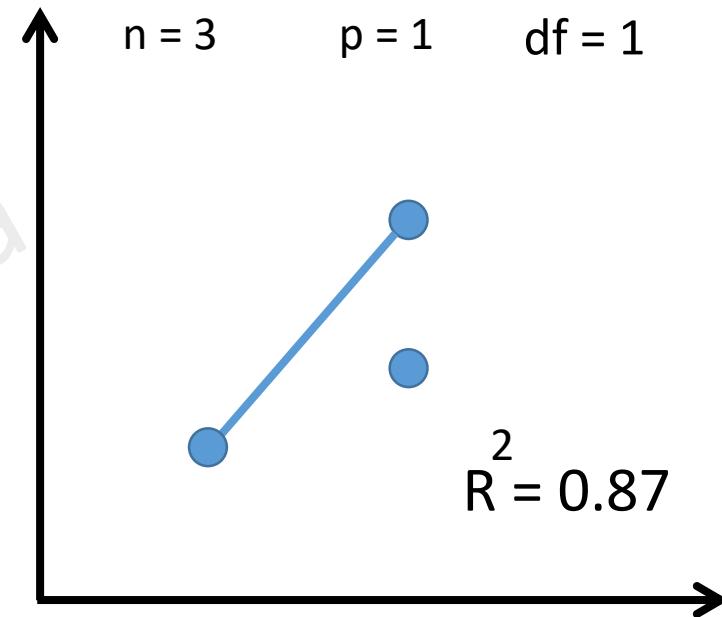


Degrees of Freedom in Statistics ($n - p - 1$)

$$y = b_0 + b_1 x_1$$

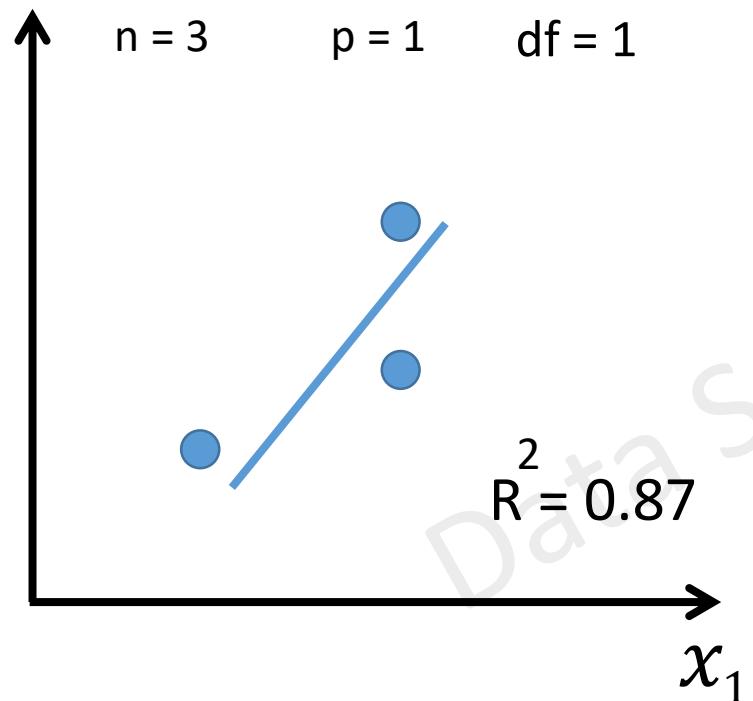


$$df = 0 \rightarrow 1$$
$$R^2 = 1 \rightarrow 0.87$$

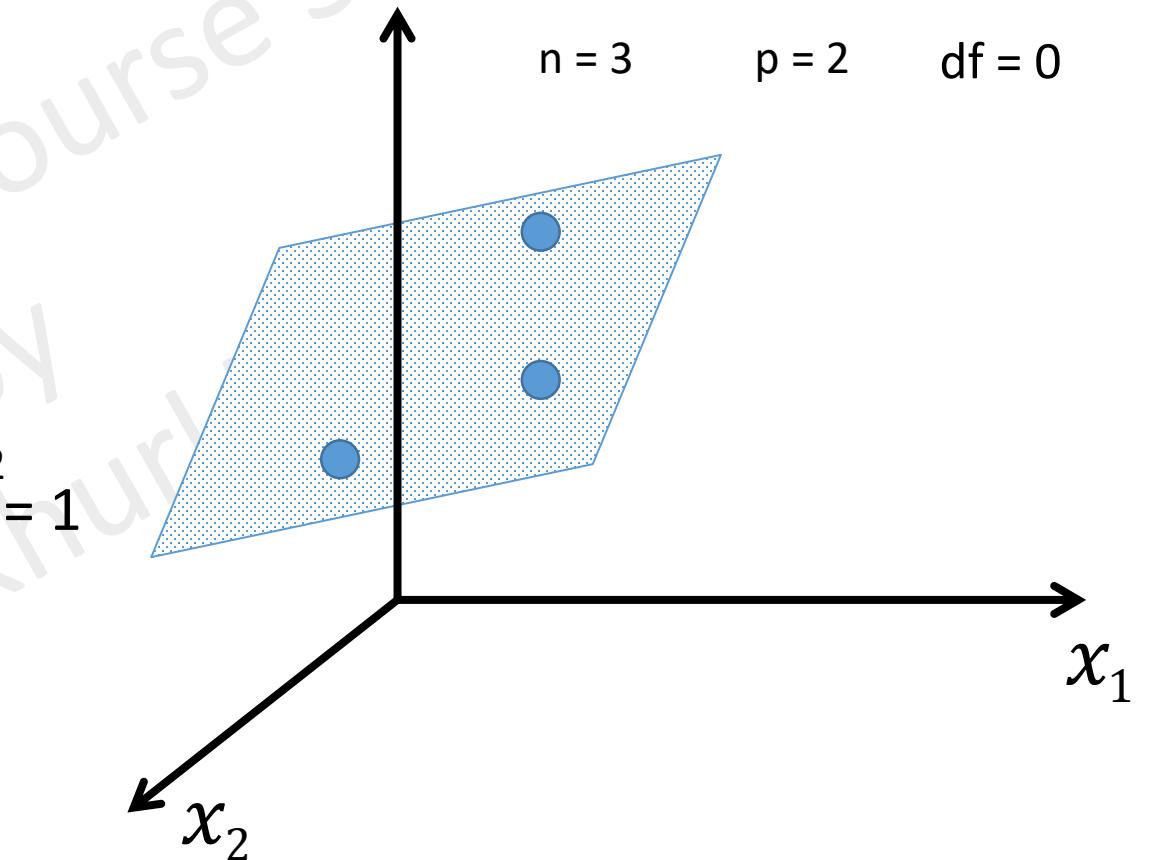


Degrees of Freedom in Statistics ($n - p - 1$)

$$y = b_0 + b_1 x_1$$



$$y = b_0 + b_1 x_1 + b_2 x_2$$



Adjusted R-Squared

$$\bar{R}^2 = 1 - \left[\frac{(1 - R^2) * (n - 1)}{n - p - 1} \right]$$

R = Sample R-Squared

p = Number of independent variables

n = sample size or number of observations

Adjusted R-Squared

Lower value of
Adjusted R-Squared



$$\bar{R}^2 = 1 - \left[\frac{(1 - R^2) * (n - 1)}{n - p - 1} \right]$$

Increase in this term

Lower Denominator due
to higher value of p.

If the R-Squared does not
increase significantly.

R = Sample R-Squared

p = Number of independent variables

n = sample size or number of observations

Adjusted R-Squared

| N | p | R-Squared | Adjusted R-Squared |
|----|----|-----------|--------------------|
| 50 | 10 | 0.80 | 0.75 |
| 50 | 12 | 0.82 | 0.76 |
| 50 | 15 | 0.83 | 0.75 |
| 50 | 20 | 0.84 | 0.73 |

$$\bar{R}^2 = 1 - \frac{(1 - R^2) * (n - 1)}{n - p - 1}$$

Multiple Linear Regression Assumptions

Assumptions of Multiple Linear Regression

Relationship Among Variables

- Linear Relationship
- No Multicollinearity
- No Auto-Correlation
- Endogeneity

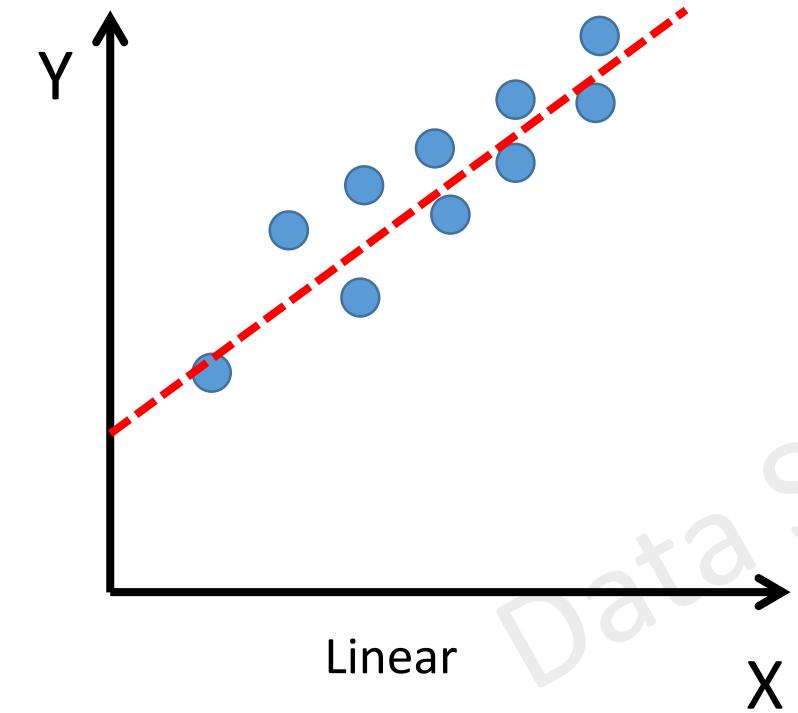
Behaviour of Data

- Sample Size
- Normality
- Homoscedasticity

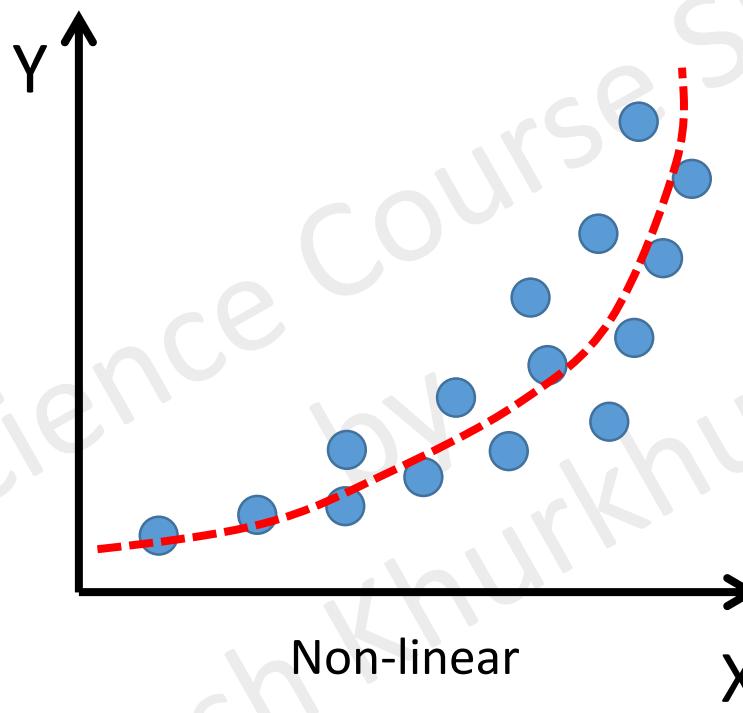
Linear Relationship

- Dependent and Independent Features have linear relationship
- Can be Positive or Negative correlation
- Can be checked using Pearson Correlation Coefficient as well as visualisation

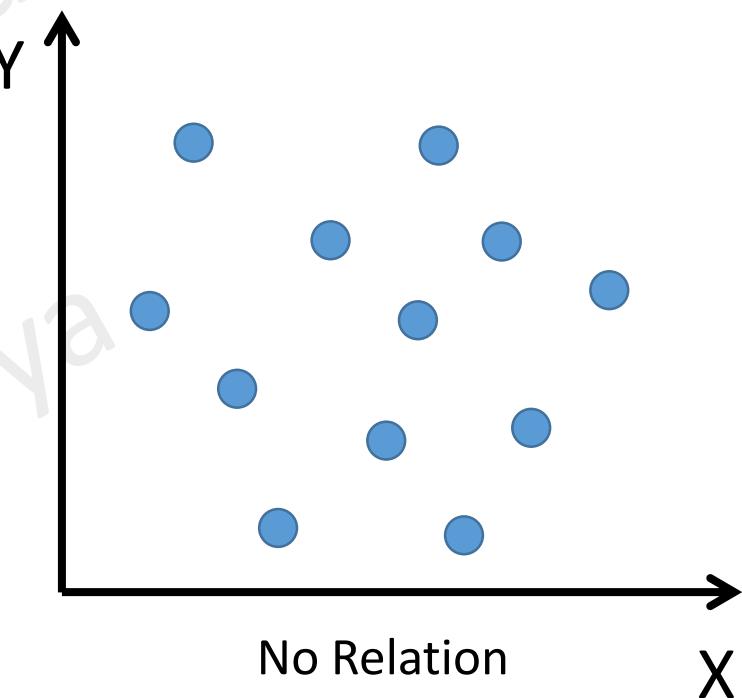
Linear Relationship



Linear

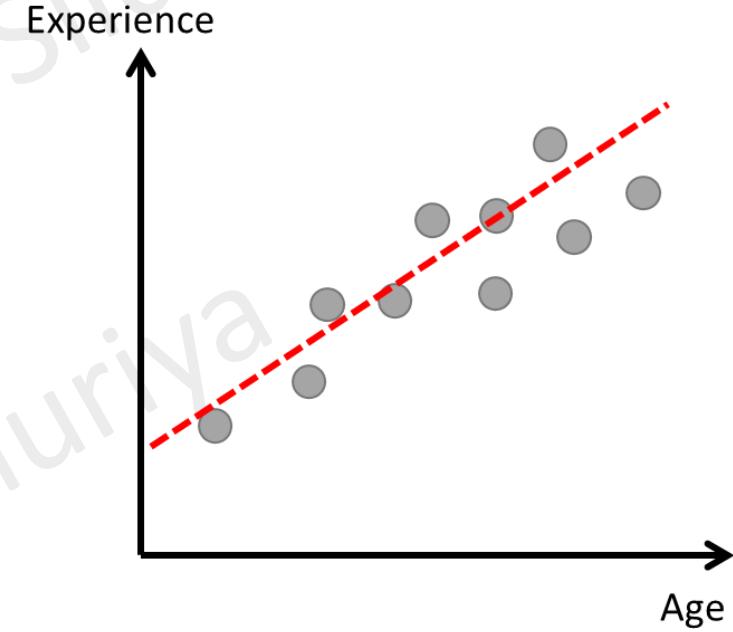
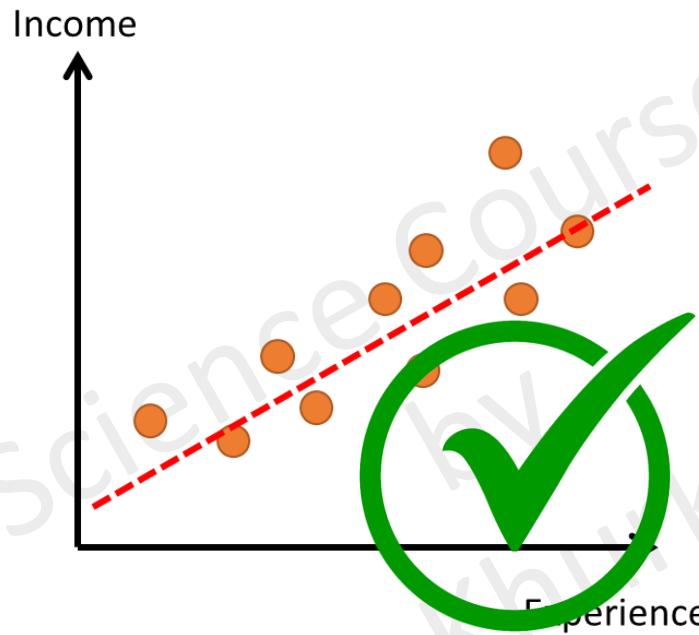
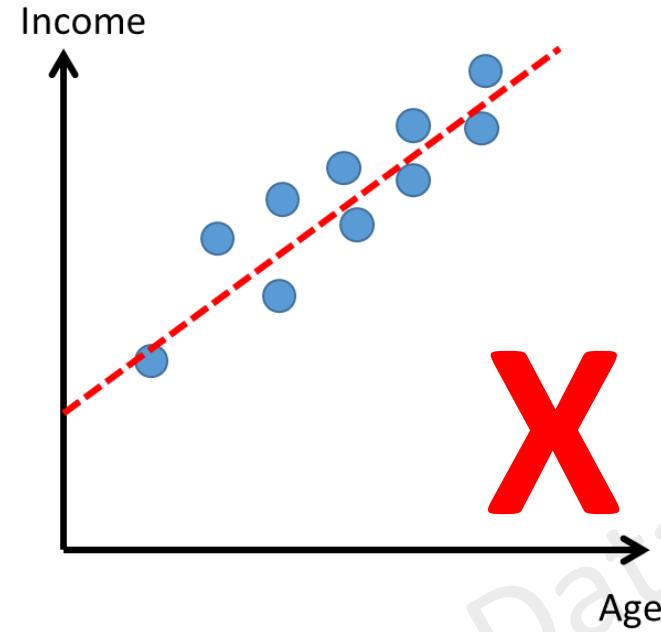


Non-linear



No Relation

No Multicollinearity

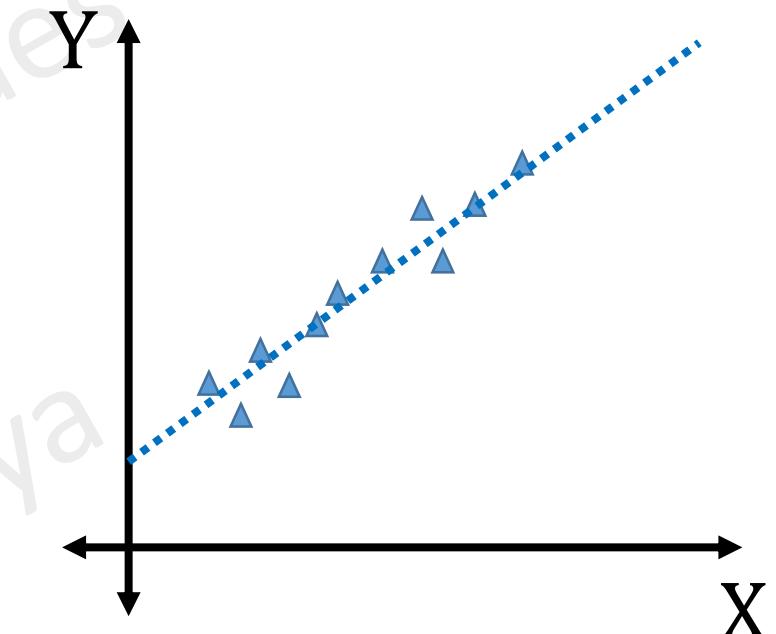


Statistically Correlated

- Strength of the correlation – Coefficient of Correlation
- Direction of correlation – Sign of the Coefficient

Pearson Correlation
Coefficient

$$r = \frac{\sum (x - \bar{x}) * (y - \bar{y})}{(N - 1) * \sigma_x * \sigma_y}$$



Correlation Coefficient Matrix

| Age | Experience | Education Received | Salary |
|-----|------------|--------------------|-----------|
| 32 | 8 | 6 | \$ 8,000 |
| 40 | 15 | 8 | \$ 12,000 |
| 35 | 6 | 8 | \$ 10,000 |

| | Age | Experience | Education Received | Salary |
|--------------------|-----|------------|--------------------|--------|
| Age | 1 | 0.9 | 0.2 | 0.7 |
| Experience | 0.9 | 1 | 0.15 | 0.72 |
| Education Received | 0.2 | 0.15 | 1 | 0.85 |
| Salary | 0.7 | 0.72 | 0.85 | 1 |

Auto-Correlation

Data Science Course Slides
by
Jitesh Khurkhuriya

Auto-Correlation

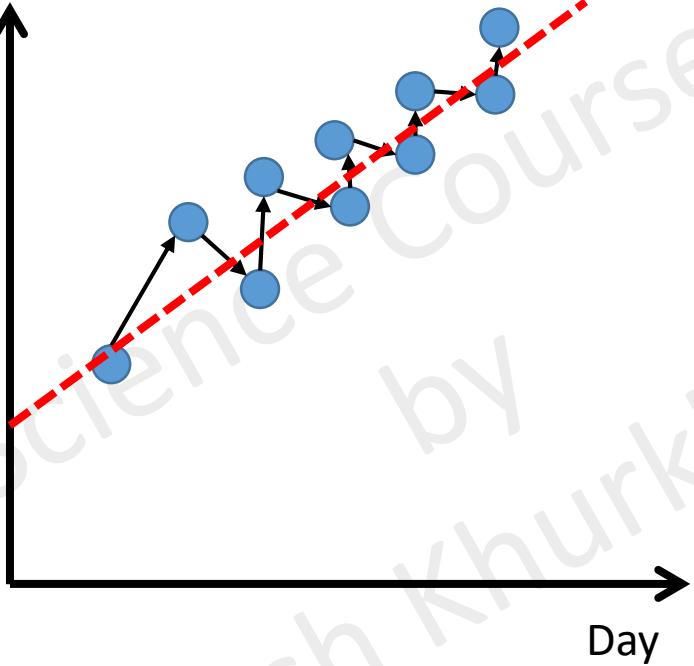
The value of one record for the same variable or feature is dependent on the value from the same column but of different record.



| X1 | X2 | Y |
|----------|----------|----|
| Value-11 | Value-21 | Y1 |
| Value-12 | Value-22 | Y2 |
| Value-13 | Value-23 | Y3 |

Auto-Correlation

Stock Price



Measure of Autocorrelation

$$\text{ACF, } \rho_k = \frac{\sum_{t=k+1}^T (x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^T (x_t - \bar{x})^2}$$

Number of time Units or Lag k

Lagged Data

Autocorrelation for lag of 1.

k = 1

$$\text{ACF, } \rho_1 = \frac{(x_2 - \bar{x})(x_1 - \bar{x})}{(x_2 - \bar{x})^2}$$

$$(33 - 34)(32 - 34)$$

$$\text{ACF, } \rho_1 = \frac{(33 - 34)(32 - 34)}{(33 - 34)^2}$$

| Day | Temperature |
|-----|-------------|
| 1 | 32 |
| 2 | 33 |
| 3 | 33 |
| 4 | 36 |
| 5 | 33 |
| 6 | 37 |

Autocorrelation

k = 4

$$\frac{\sum_{t=k+1}^T (x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^T (x_t - \bar{x})^2}$$

| Day | Temperature |
|-----|-------------|
| 1 | 32 |
| 2 | 33 |
| 3 | 33 |
| 4 | 36 |
| 5 | 33 |
| 6 | 37 |
| 7 | |
| 8 | |



Autocorrelation

| t0 | t-1 | t-2 | t-3 | t-4 |
|-----|-----|-----|-----|-----|
| 8 | NaN | NaN | NaN | NaN |
| 14 | 8 | NaN | NaN | NaN |
| 36 | 14 | 8 | NaN | NaN |
| 56 | 36 | 14 | 8 | NaN |
| 84 | 56 | 36 | 14 | 8 |
| 94 | 84 | 56 | 36 | 14 |
| 106 | 94 | 84 | 56 | 36 |
| 110 | 106 | 94 | 84 | 56 |
| 93 | 110 | 106 | 94 | 84 |
| 67 | 93 | 110 | 106 | 94 |
| 35 | 67 | 93 | 110 | 106 |
| 37 | 35 | 67 | 93 | 110 |
| 36 | 37 | 35 | 67 | 93 |
| 34 | 36 | 37 | 35 | 67 |
| 28 | 34 | 36 | 37 | 35 |
| 39 | 28 | 34 | 36 | 37 |
| 17 | 39 | 28 | 34 | 36 |

Sliding Window Approach

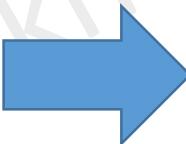
| t0 | t-1 | t-2 | t-3 | t-4 |
|-----|-----|-----|-----|-----|
| 8 | NaN | NaN | NaN | NaN |
| 14 | 8 | NaN | NaN | NaN |
| 36 | 14 | 8 | NaN | NaN |
| 56 | 36 | 14 | 8 | NaN |
| 84 | 56 | 36 | 14 | 8 |
| 94 | 84 | 56 | 36 | 14 |
| 106 | 94 | 84 | 56 | 36 |
| 110 | 106 | 94 | 84 | 56 |
| 93 | 110 | 106 | 94 | 84 |
| 67 | 93 | 110 | 106 | 94 |
| 35 | 67 | 93 | 110 | 106 |
| 37 | 35 | 67 | 93 | 110 |
| 36 | 37 | 35 | 67 | 93 |
| 34 | 36 | 37 | 35 | 67 |
| 28 | 34 | 36 | 37 | 35 |
| 39 | 28 | 34 | 36 | 37 |
| 17 | 39 | 28 | 34 | 36 |



pandas.shift()

Autocorrelation Function

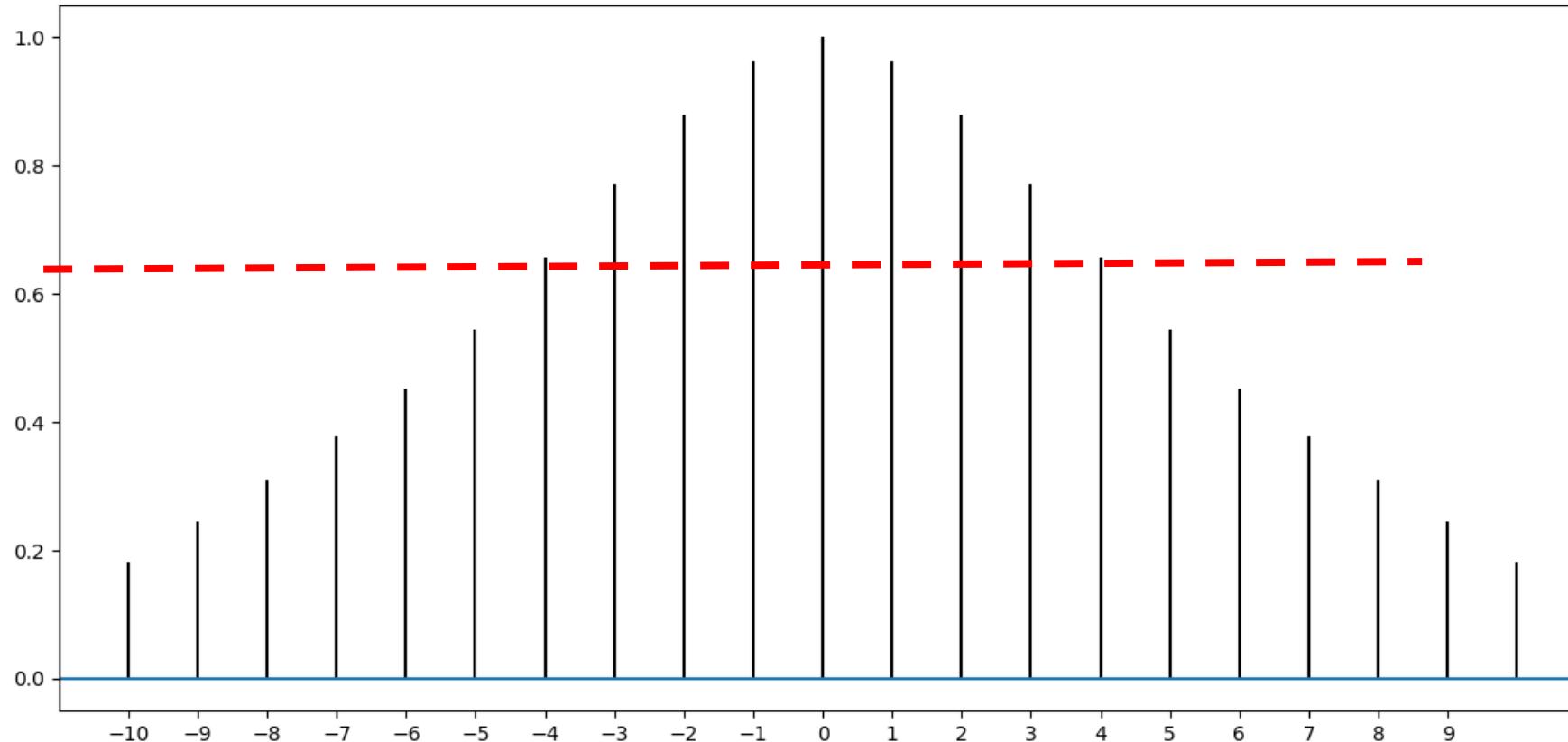
| t0 | t-1 | t-2 | t-3 | t-4 |
|-----|-----|-----|-----|-----|
| 8 | NaN | NaN | NaN | NaN |
| 14 | 8 | NaN | NaN | NaN |
| 36 | 14 | 8 | NaN | NaN |
| 56 | 36 | 14 | 8 | NaN |
| 84 | 56 | 36 | 14 | 8 |
| 94 | 84 | 56 | 36 | 14 |
| 106 | 94 | 84 | 56 | 36 |
| 110 | 106 | 94 | 84 | 56 |
| 93 | 110 | 106 | 94 | 84 |
| 67 | 93 | 110 | 106 | 94 |
| 35 | 67 | 93 | 110 | 106 |
| 37 | 35 | 67 | 93 | 110 |
| 36 | 37 | 35 | 67 | 93 |
| 34 | 36 | 37 | 35 | 67 |
| 28 | 34 | 36 | 37 | 35 |
| 39 | 28 | 34 | 36 | 37 |
| 17 | 39 | 28 | 34 | 36 |



$$\frac{\sum_{t=k+1}^T (x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^T (x_t - \bar{x})^2}$$

Autocorrelation Function (ACF)

pyplot.acorr()



Assumptions of Multiple Linear Regression

Relationship Among Variables

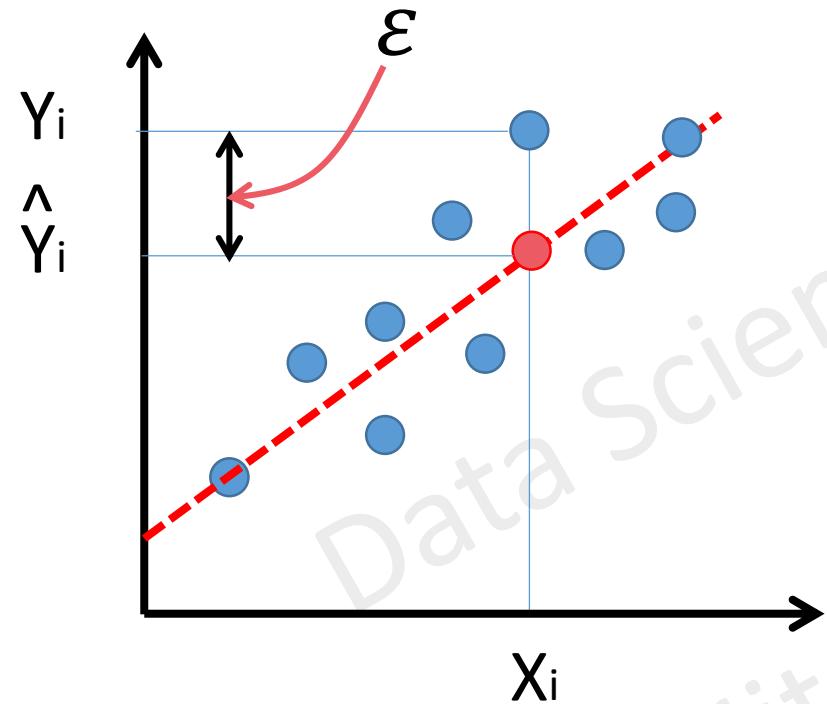
- Linear Relationship
- No Multicollinearity
- No Auto-Correlation
- Endogeneity

Behaviour of Data

- Sample Size
- Normality
- Homoscedasticity

Endogeneity

- Situations in which an explanatory/independent variable is correlated with the error term.



$$y = b_0 + b_1 x_1$$

$$y_i = \hat{y}_i + \varepsilon$$

$$y_i = b_0 + b_1 x_1 + \varepsilon$$

$$\varepsilon = f(x) \quad ??$$

Endogeneity

| Wheelbase | Length | Engine Size | Horsepower | Price |
|-----------|--------|-------------|------------|-------|
| 88.6 | 168.8 | 130 | 111 | |
| 94.5 | 171.2 | 152 | 154 | |
| 99.8 | 176.6 | 109 | 102 | |
| 99.4 | 176.6 | 136 | 115 | |

$$\text{Price} = 8215 + 11.5 * \text{Wheelbase} + 7.8 * \text{Length} + 2.8 * \text{EngineSize} + 3.2 * \cancel{\text{HP}}$$

Omitted Variable Bias

| Wheelbase | Length | Engine Size | Horsepower | Price |
|-----------|--------|-------------|------------|-------|
| 88.6 | 168.8 | 130 | 111 | |
| 94.5 | 171.2 | 152 | 154 | |
| 99.8 | 176.6 | 109 | 102 | |
| 99.4 | 176.6 | 136 | 115 | |

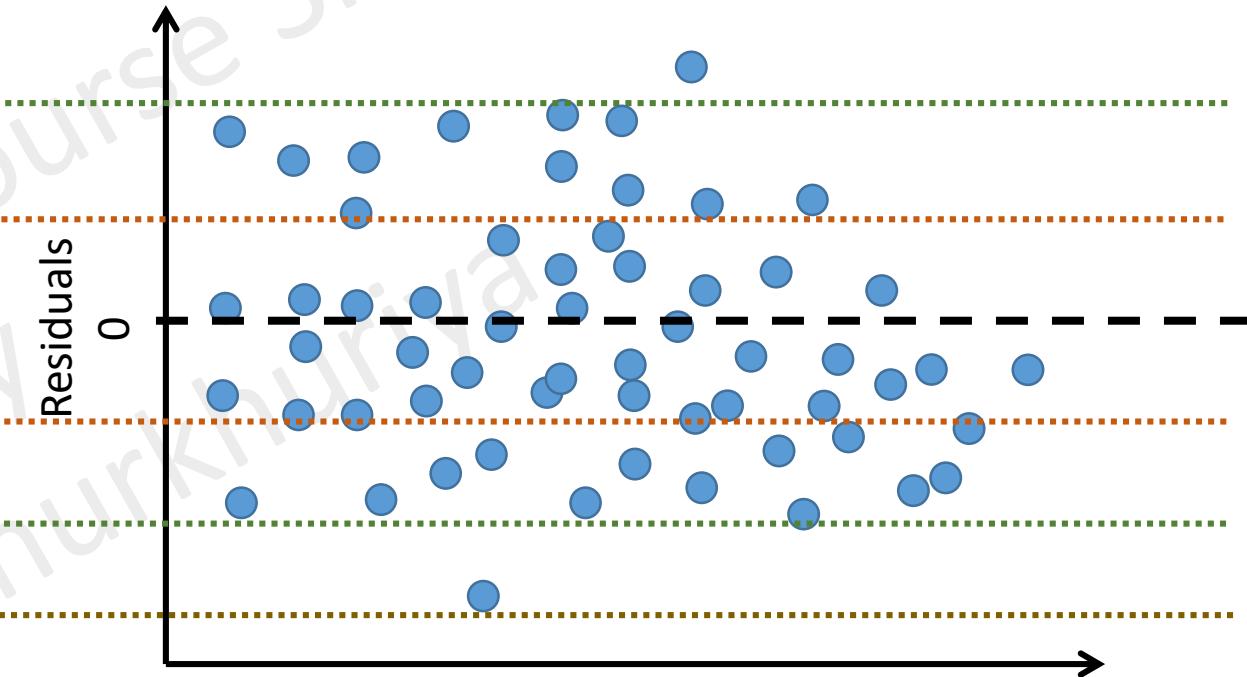
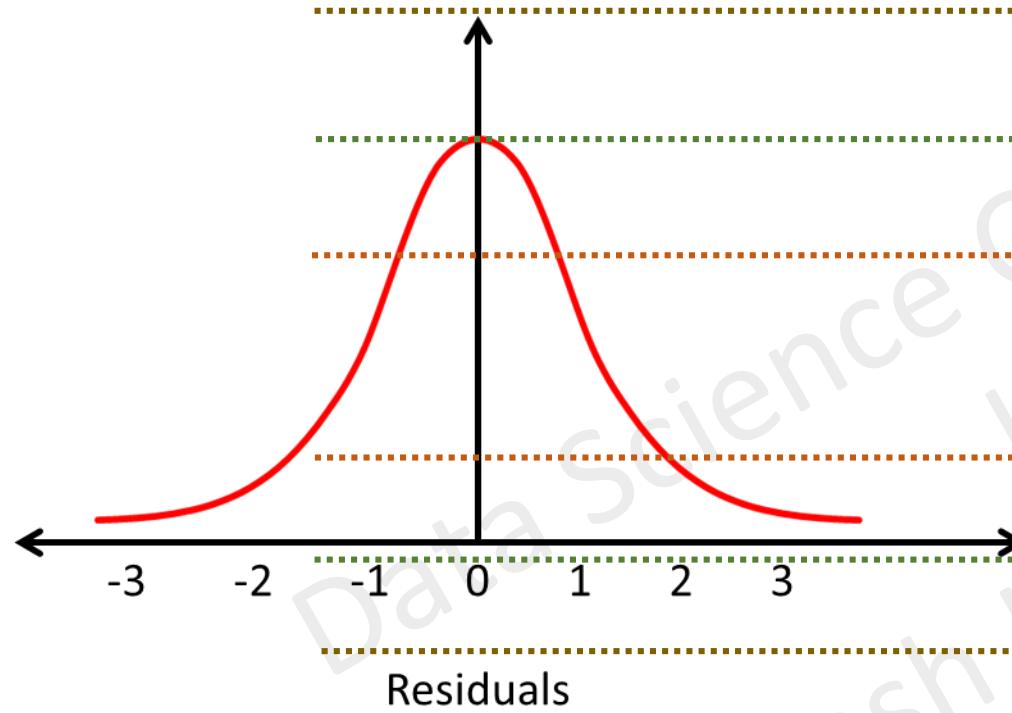
$$\text{Price} = 8215 + 11.5 \cdot \text{Wheelbase} + 7.8 \cdot \text{Length} + 2.8 \cdot \text{EngineSize} + \epsilon$$

Price ~ EngineSize ~ Horsepower ~ error

QuickFix

- Start with All the variable
- Remove unwanted ones using Adjusted R-Squared or feature selection methods

Normality of Residuals



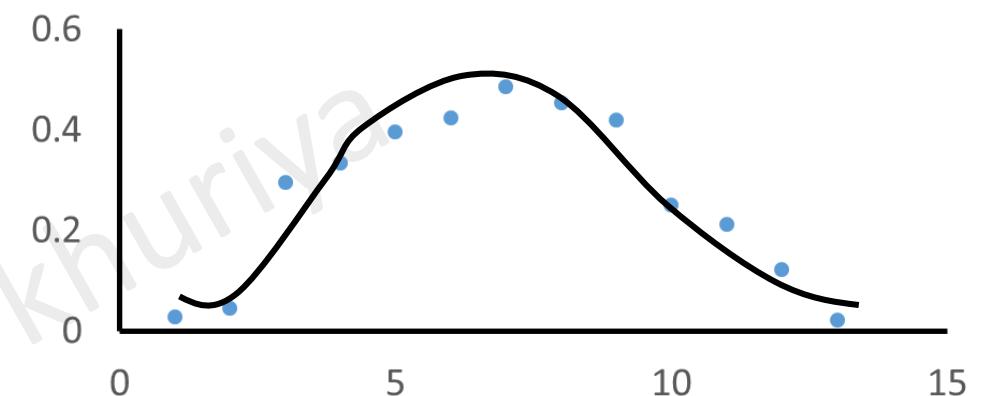
Normality of Residuals

| Hrs Studied (X) | Marks (Y) |
|-----------------|-----------|
| 0 | 40 |
| 2 | 52 |
| 3 | 53 |
| 4 | 55 |
| 4 | 56 |
| 5 | 72 |
| 6 | 71 |
| 6 | 88 |
| 7 | 56 |
| 7 | 74 |
| 8 | 89 |
| 9 | 67 |
| 9 | 89 |

| Marks predicted |
|-----------------|
| 41.80 |
| 50.90 |
| 55.45 |
| 60.00 |
| 60.00 |
| 64.55 |
| 69.10 |
| 69.10 |
| 73.65 |
| 73.65 |
| 78.20 |
| 82.75 |
| 82.75 |

| Residuals |
|-----------|
| -1.80 |
| 1.10 |
| -2.45 |
| -5.00 |
| -4.00 |
| 7.45 |
| 1.90 |
| 18.9 |
| -17.65 |
| 0.35 |
| 10.8 |
| -15.75 |
| 6.25 |

$$y = 41.8 + 4.55 x$$



Residuals or Errors should be normally distributed.

HomoScedasticity

Data Science Course Slides
by
Jitesh Khurkhuriya

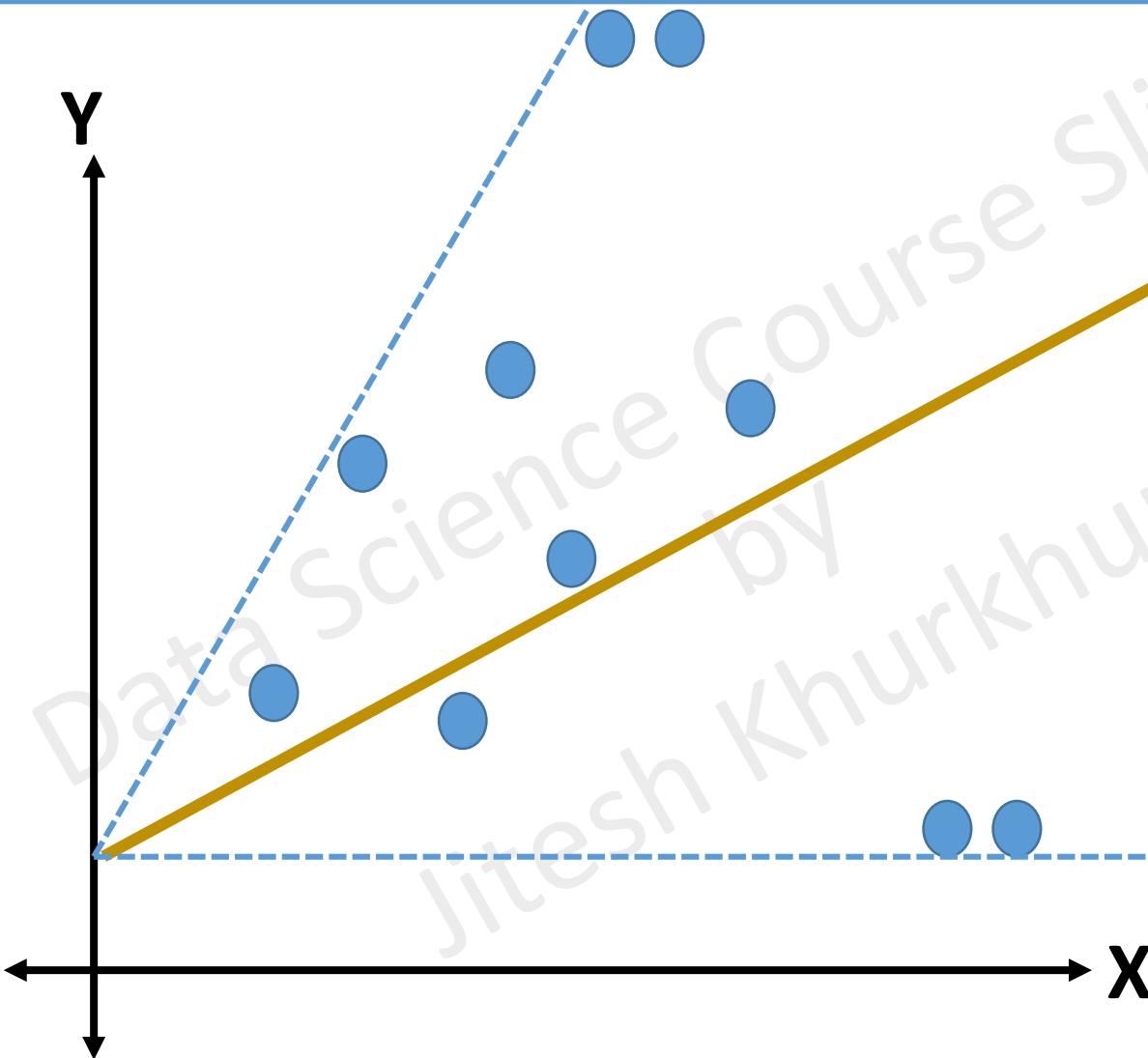
Homoscedasticity

HomoScedasticity

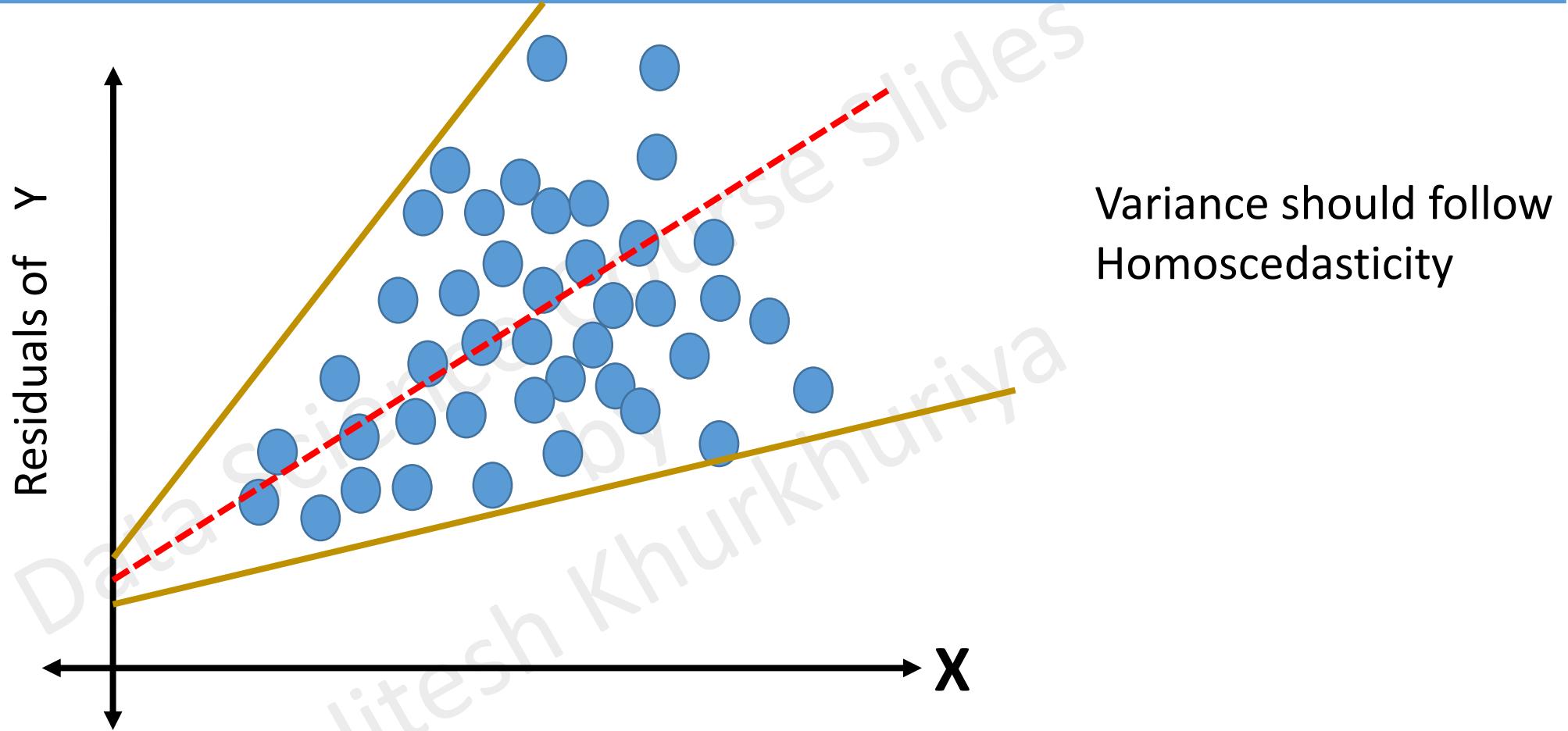
Same

Variance/Spread

Ordinary Least Square

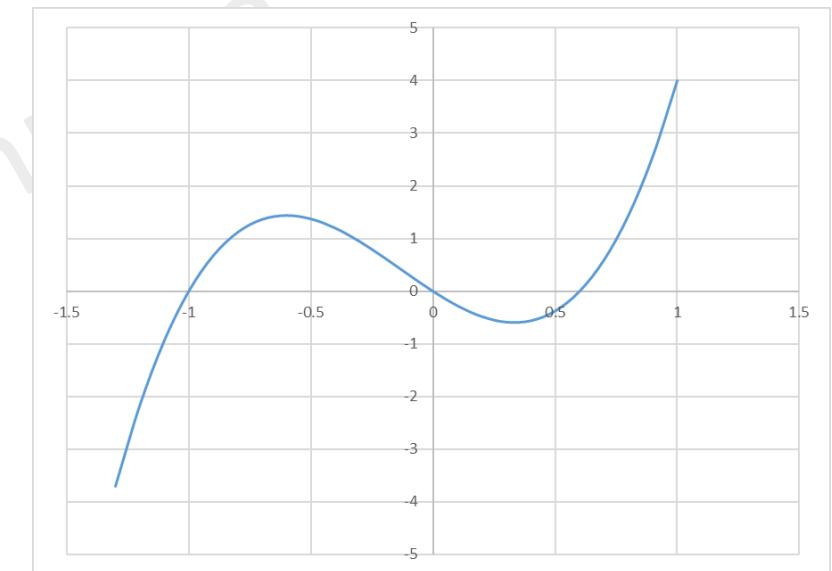


Heteroscedasticity



Remedies

- Rebuild the model with new predictors
- Look for outliers
- Variable transformation using Log or Power transformation
- Consider Polynomial or other regression algorithm



Dummy Variable Trap

Multiple Linear Regression

| Hrs Studied (X ₁) | Hrs Slept (X ₂) | Marks (Y) |
|----------------------------------|--------------------------------|--------------|
| 0 | 8 | 40 |
| 2 | 8 | 52 |
| 3 | 7.5 | 53 |
| 4 | 7 | 55 |
| 4 | 9 | 56 |
| 5 | 8.5 | 72 |
| 6 | 9 | 71 |
| 6 | 7 | 88 |
| 7 | 6 | 56 |
| 7 | 7 | 74 |
| 8 | 9 | 89 |
| 9 | 6 | 67 |
| 9 | 9 | 89 |

$$y = b_0 + b_1 x_1 + b_2 x_2$$

Dependent Variable
Marks Obtained

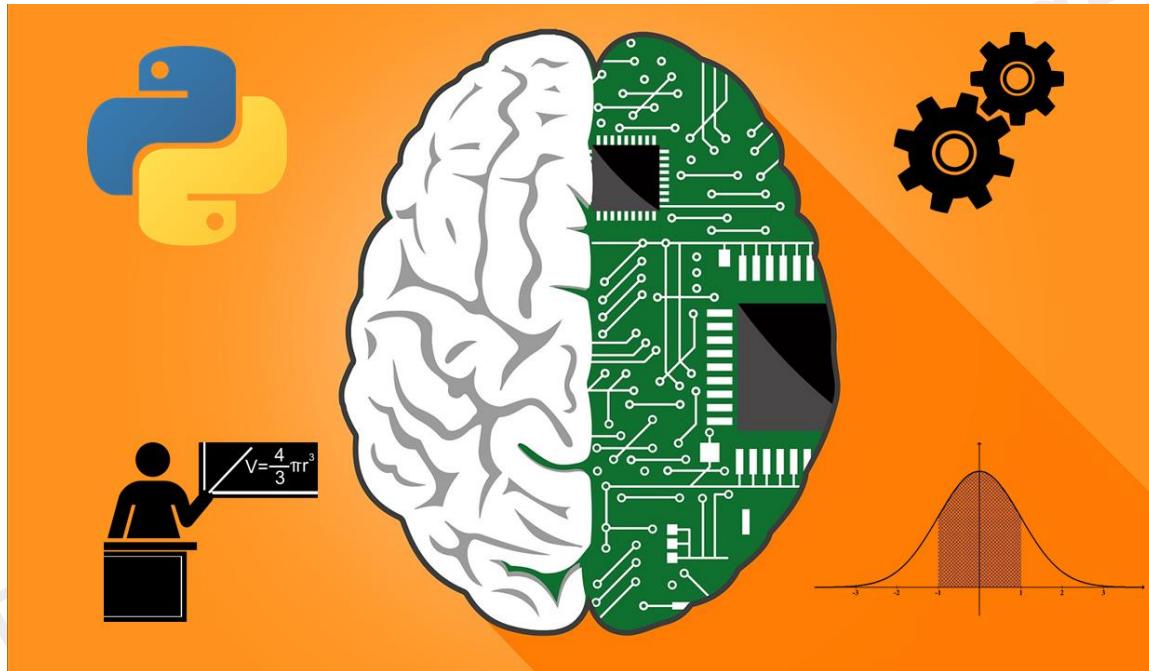
Independent Variable
Hrs Studied
Hrs Slept

Dummy Variable Trap

| Hrs Studied (X ₁) | Hrs Slept (X ₂) | Math (X ₃) | Science (X ₄) | Art (X ₅) | Marks (Y) |
|----------------------------------|--------------------------------|---------------------------|------------------------------|--------------------------|--------------|
| 0 | 8 | 1 | 0 | 0 | 40 |
| 2 | 8 | 0 | 1 | 0 | 52 |
| 3 | 7.5 | 0 | 0 | 1 | 53 |
| 4 | 7 | 1 | 0 | 0 | 55 |
| 4 | 9 | 1 | 0 | 0 | 56 |
| 5 | 8.5 | 1 | 0 | 0 | 72 |
| 6 | 9 | 0 | 1 | 0 | 71 |
| 6 | 7 | 0 | 0 | 1 | 88 |
| 7 | 6 | 0 | 0 | 1 | 56 |
| 7 | 7 | 0 | 1 | 0 | 74 |
| 8 | 9 | 0 | 1 | 0 | 89 |
| 9 | 6 | 1 | 0 | 0 | 67 |
| 9 | 9 | 0 | 0 | 1 | 89 |

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4$$

Complete Data Science and Machine Learning Using Python



Thank You!