CRISP-DM FRAMEWORK

While performing the data analysis we face multiple problems like understanding the problem statement, collecting the correct data, selecting the relevant data, which technique I should follow etc.

CRISO-DM framework helps the data analysts in following a set of procedures to solve any analytical problem irrespective of the industry. Advantage of CRISP-DM is it is not tightly associated with any industry, technology, application and applicable to any line of industry.

History

CRISP-DM (CRoss Industry Standard Process for Data Mining) was conceived in late 1996 by the automotive company Dailmler- Chrystler, the statistical software provider SPSS, and the data warehouse provider NCR (Chapman et al., 2000)

CRISP-DM is a comprehensive data mining methodology and process model that provides anyone—from novices to data mining experts—with a complete blueprint for conducting a data mining project.

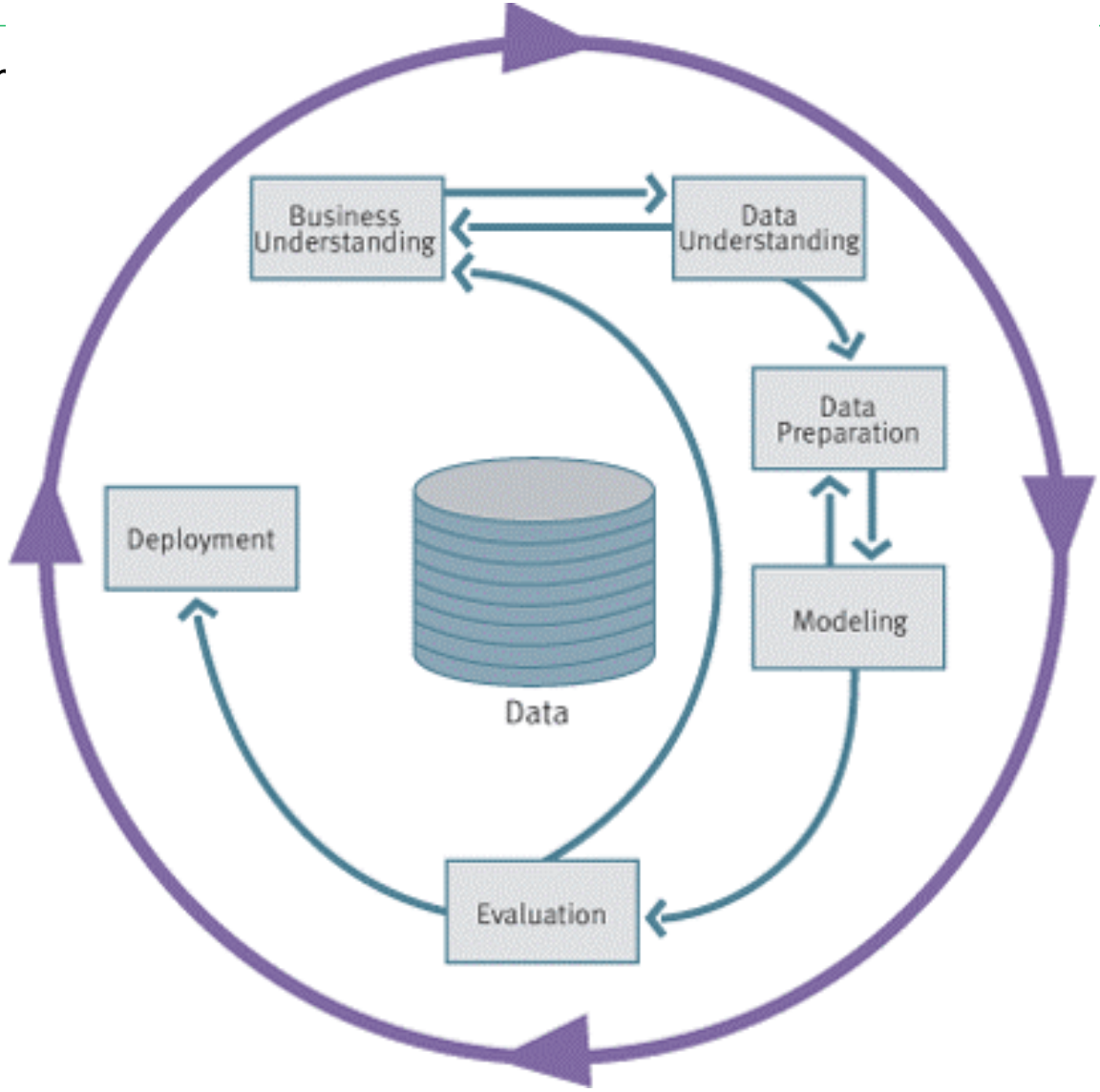
Based on current research CRISP-DM is the most widely used form of data-mining model because of its various advantages which solved the existing problems in the data mining industries.

CRISP-DM FRAMEWORK

- CRISP-DM breaks the process of data mining into six r

    - Business understanding
    - Data understanding
    - Data Preparation
    - Data Modelling
    - Model Evaluation
    - Model Deployment

Above steps in CRISP-DM framework are not necessarily sequential and one can always go back to the previous step if required. It is a continuous process where we can improve the model in the next iteration based on the feedback or learnings from the previous iteration.



CRISP-DM FRAMEWORK

- **Business Understanding**
  - Understanding project objectives and requirements
  - Data mining problem definition
- **Data Understanding**
  - Initial data collection and familiarization
  - Identify data quality issues
  - Initial, obvious results
- **Data Preparation**
  - Record and attribute selection
  - Data cleansing
- **Modeling**
  - Run the data mining tools
- **Evaluation**
  - Determine if results meet business objectives
  - Identify business issues that should have been addressed earlier
- **Deployment**
  - Put the resulting models into practice
  - Set up for repeated/continuous mining of the data

Linear Regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data.

One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest.

This does not necessarily imply that one variable *causes* the other (for example, higher SAT scores do not *cause* higher college grades), but that there is some significant association between the two variables.

Linear Regression

A scatterplot can be a helpful tool in determining the strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatterplot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model.

If we consider single independent variable to calculate the dependent variable we call it Simple Linear Regression

If we consider more than one independent variable to calculate the dependent variable we call it Multiple Linear Regression

A simple linear regression model attempts to explain the relationship between a dependent and an independent variable using a straight line.

The independent variable is also known as the **predictor** variable. And the dependent variables are also known as the **output** variables

Linear Regression

- In credit risk analytics, let's assume that you need to predict the average amount defaulted by any customer based on different factors such as the credit score of the person, the frequency of using the credit card, the average amount spent during each shopping session, etc.

  Can you tell what would be the dependent variable for this regression problem?

  Can you tell what would be the independent variable for this regression problem?

  What kind of regression problem is this?

Linear Regression

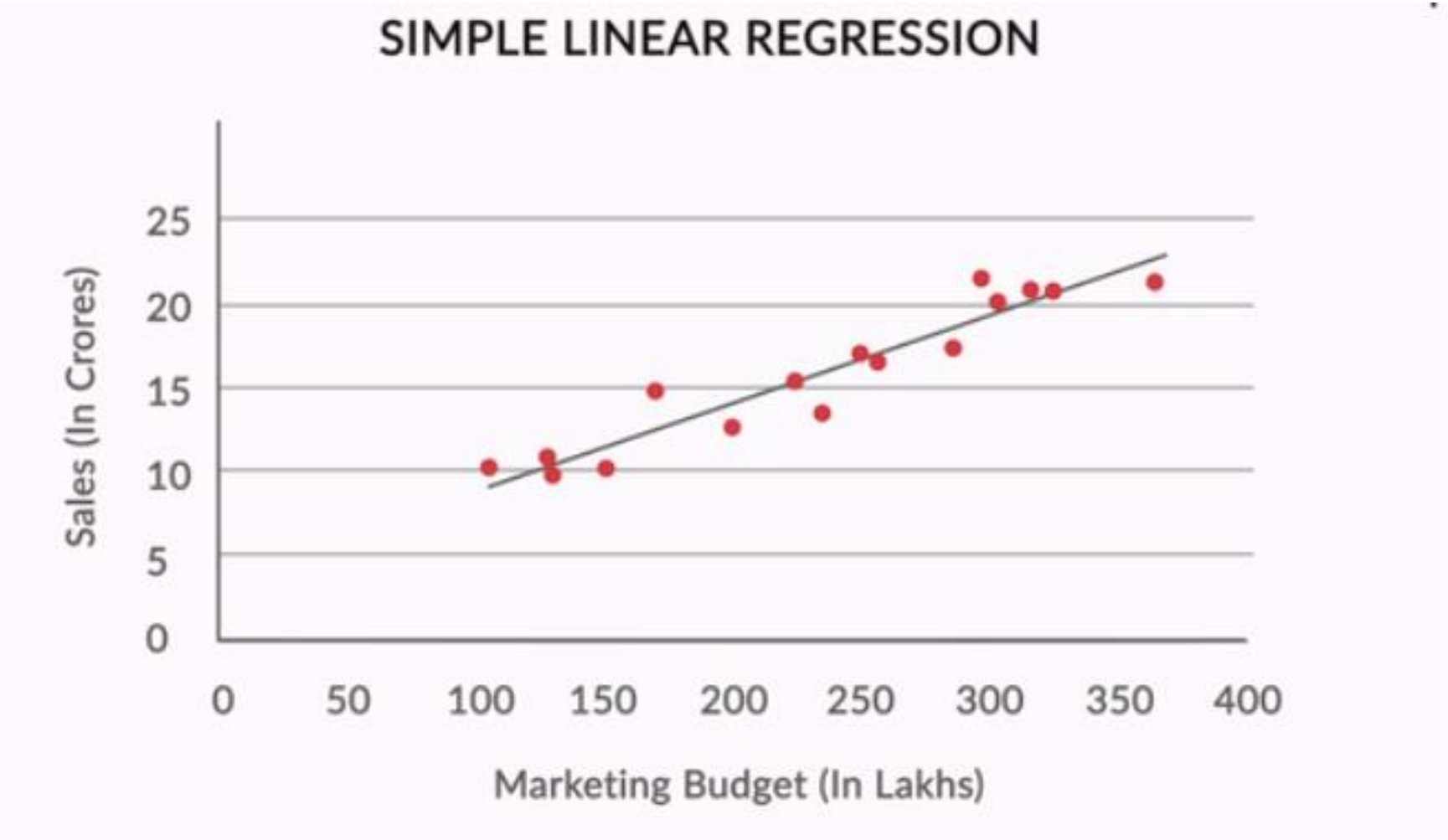
- In credit risk analytics, let's assume that you need to predict the average amount defaulted by any customer based on different factors such as the credit score of the person, the frequency of using the credit card, the average amount spent during each shopping session, etc.

Can you tell what would be the dependent variable for this regression problem? - **Average amount defaulted by customer**

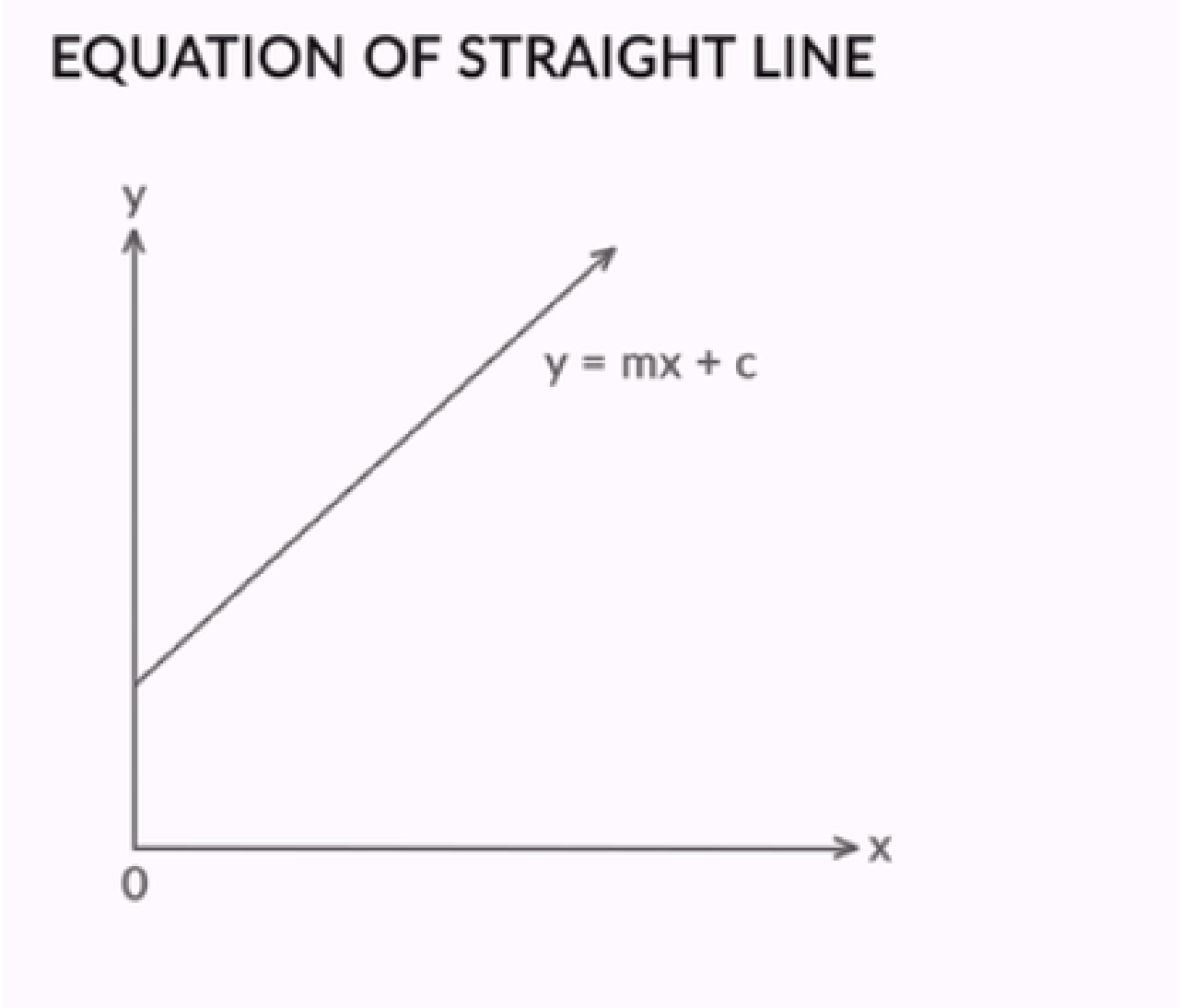
Can you tell what would be the independent variable for this regression problem? - **credit score of the person, the frequency of using the credit card, the average amount spent during each shopping session**

What kind of regression problem is this? – **Multiple regression**

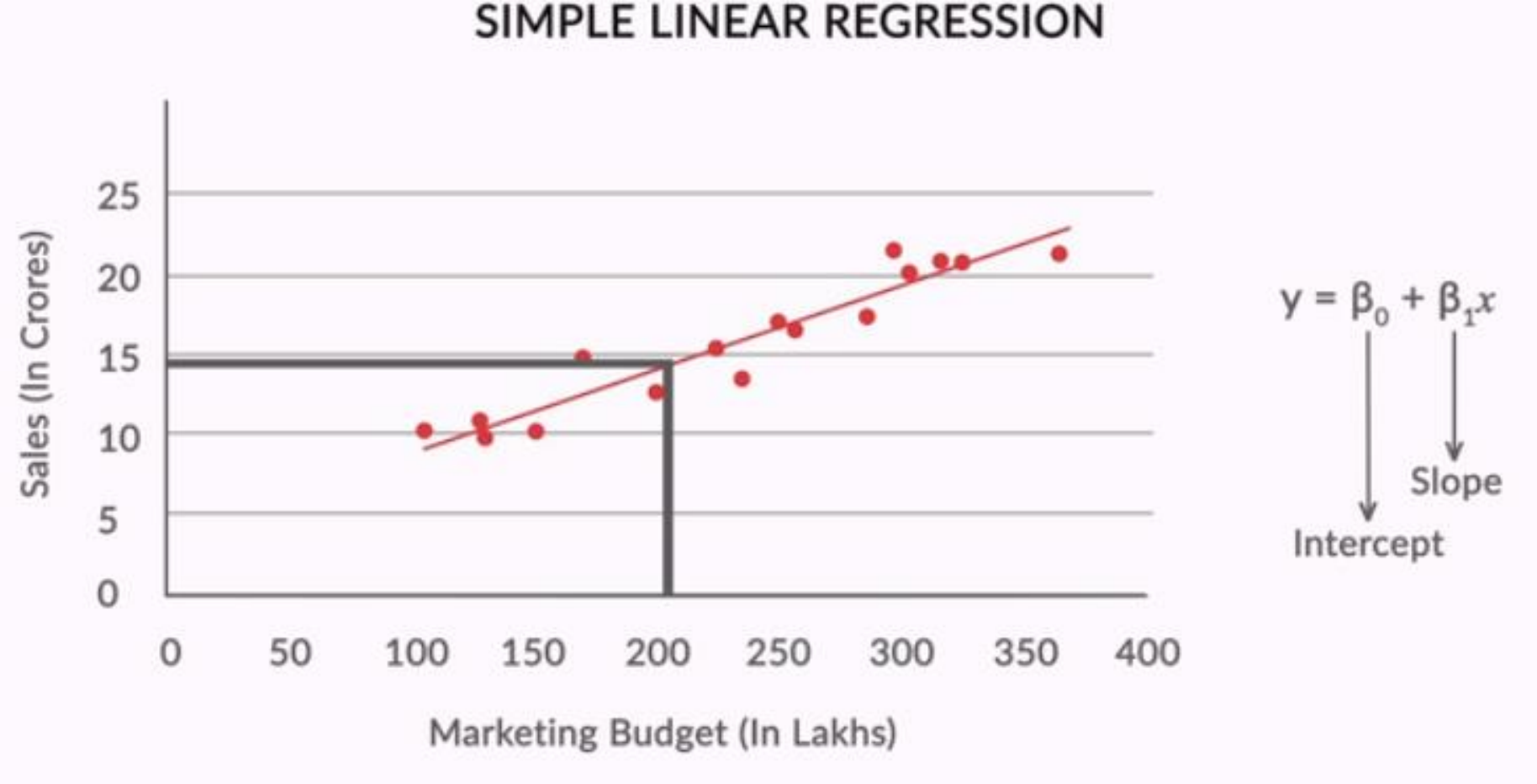
Linear Regression
- Regression Line



Linear Regression
- Regression Line

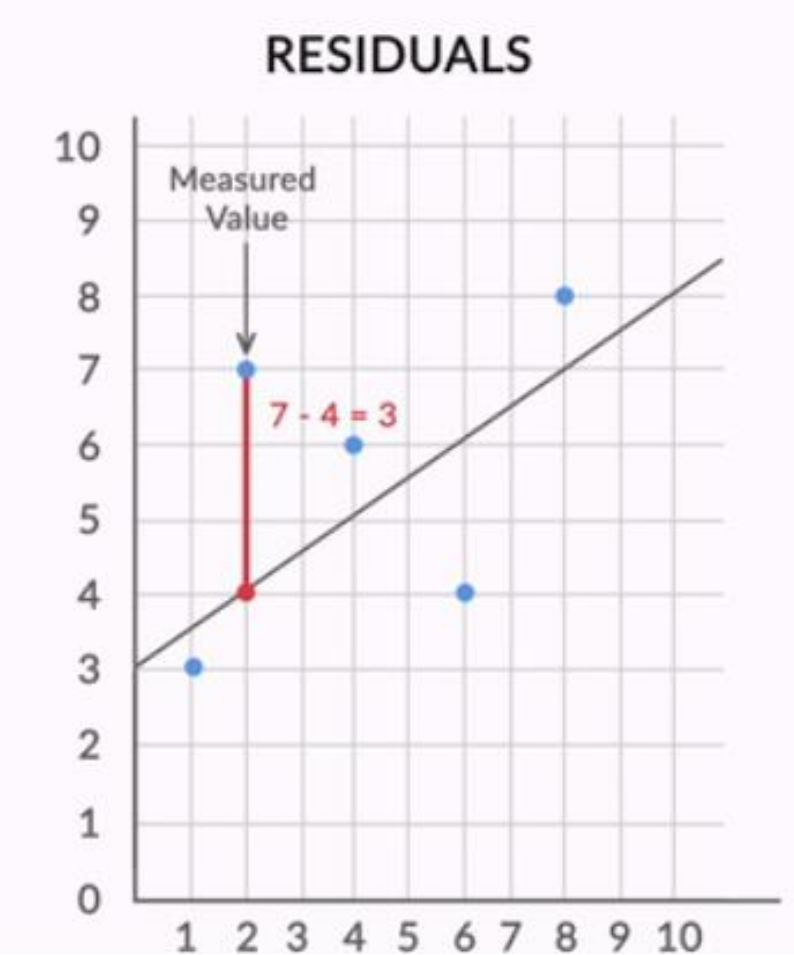


Linear Regression
- Regression Line

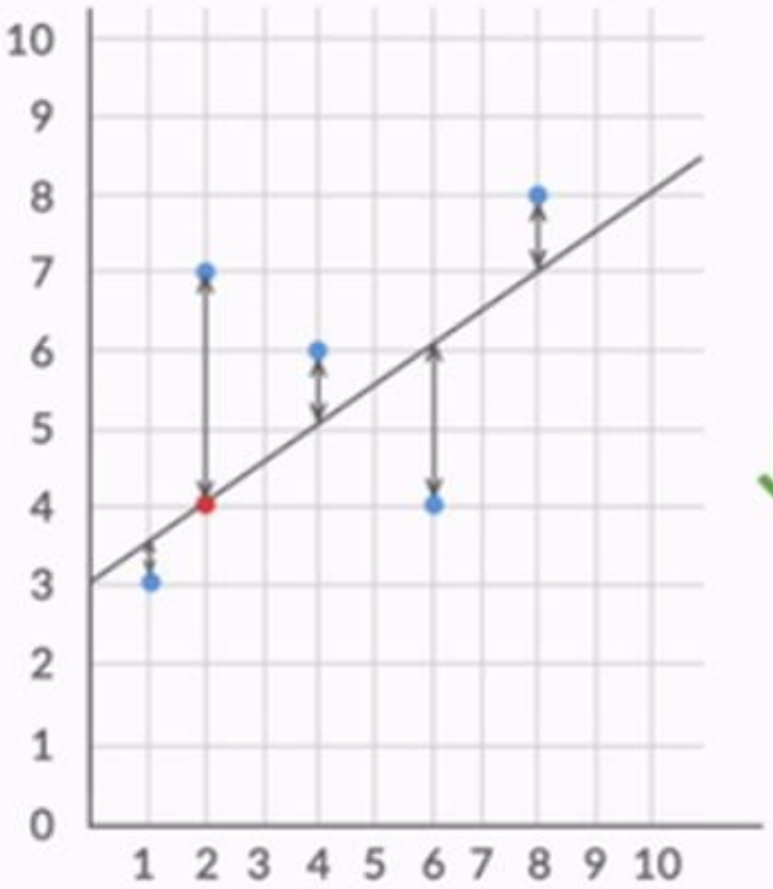


Linear Regression
- Best Fit Line

In regression, there is a notion of a best-fit line — the line which fits the given scatter-plot in the best way. Let's look at how you can define the notion of a best-fit line.



Linear Regression
- Best Fit Line



RESIDUALS

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

Intercept Slope

$$e_i = y_i - y_{pred}$$

Ordinary Least Squares Method:

$$e_1^2 + e_2^2 + \_\_\_ + e_n^2 = \text{RSS (Residual Sum Of Squares)}$$

$$RSS = (Y_1 - \beta_0 - \beta_1 X_1)^2 + (Y_2 - \beta_0 - \beta_1 X_2)^2 + \_\_\_\_ + (Y_n - \beta_0 - \beta_1 X_n)^2$$

$$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

- Since now you know that the best-fit line is obtained by minimising a quantity called Residual Sum of Squares (RSS), this is the best time to be introduced to what is known as the **cost function**.

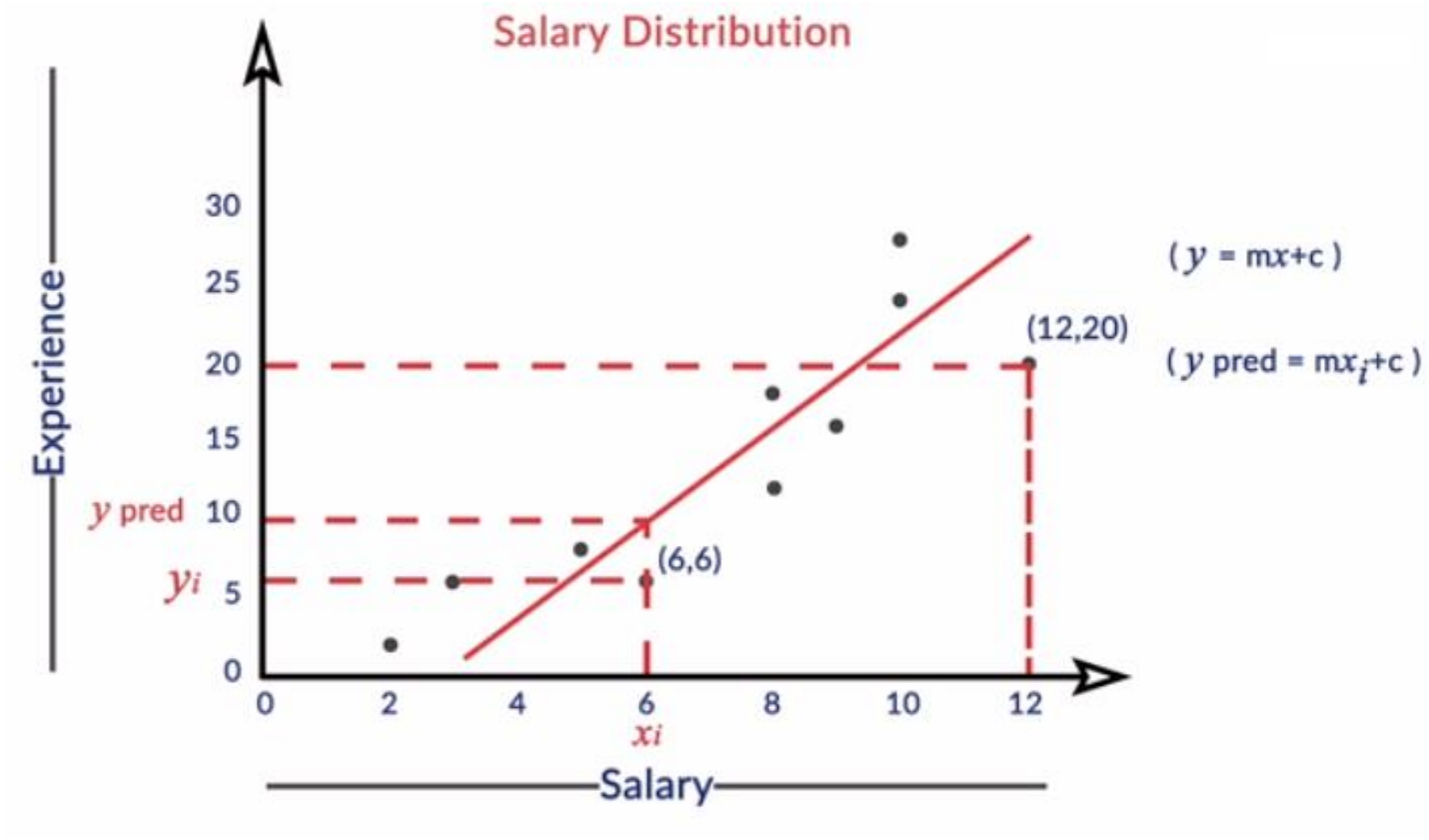
Linear Regression
- Best Fit Line

Sample Data

Years of experience	Salary (lakhs per annum)
6	6
12	20
3	6
8	12
9	16
10	24
2	2
5	8
8	18
10	28

Linear Regression
- Best Fit Line

Linear Regression
- Best Fit Line



Linear Regression
- Best Fit Line

$$J(m,c) = [\, y_1-(mx_1 + c)\,]^2 + \ldots\ldots$$

$$\frac{\partial J}{\partial m} = 0$$

$$\frac{\partial [\, y_1-(mx_1 + c)\,]^2 + \ldots\ldots}{\partial m}$$

$$2(\, y_1 - mx_1 - c\,)\,\boxed{(-x_1)}\ldots$$

Linear Regression
- Best Fit Line

$$J(m,c) = [\, y_1-(mx_1 + c)\,]^2+.....$$

$$\frac{\partial J}{\partial m} = 0$$

$$\frac{\partial [\, y_1-(mx_1 + c)\,]^2_{+.....}}{\partial m}$$

$$2(\, y_1 - mx_1 - c\,)\ (-x_1).....$$

$$= 0$$

$$J(m,c) = [\, y_1-(mx_1 + c)\,]^2+.....$$

$$\frac{\partial J}{\partial c} = 0$$

$$\frac{\partial [\, y_1-(mx_1 + c)\,]^2_{+.....}}{\partial c}$$

$$2(\, y_1 - mx_1 - c\,)\ \boxed{(-1)} +.....$$

Linear Regression
- Best Fit Line

$$J(m,c) = [\, y_1-(mx_1 + c)\,]^2 + \ldots\ldots$$

$$\frac{\partial J}{\partial m} = 0$$

$$\frac{\partial [\, y_1-(mx_1 + c)\,]^2 + \ldots\ldots}{\partial m}$$

$$2(\, y_1 - mx_1 - c\,)\ (-x_1) \ldots\ldots$$

$$= 0$$

$$J(m,c) = [\, y_1-(mx_1 + c)\,]^2 + \ldots\ldots$$

$$\frac{\partial J}{\partial c} = 0$$

$$\frac{\partial [\, y_1-(mx_1 + c)\,]^2 + \ldots\ldots}{\partial c}$$

$$2(\, y_1 - mx_1 - c\,)\ (-1) + \ldots\ldots$$

$$= 0$$

Linear Regression
- Best Fit Line

- By replacing the x values and y values from the table into the above equations and equating the both we can find out the m, c values.

y = 0.0528x + 3.3525

SLOPE	0.0528
INTERCEPT	3.3525
RSS	28.77190461

Linear Regression
- Best Fit Line

$$TSS = (Y_1 - \bar{Y})^2 + \ldots\ldots + (Y_n - \bar{Y})^2$$

$$\text{Or} \quad \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

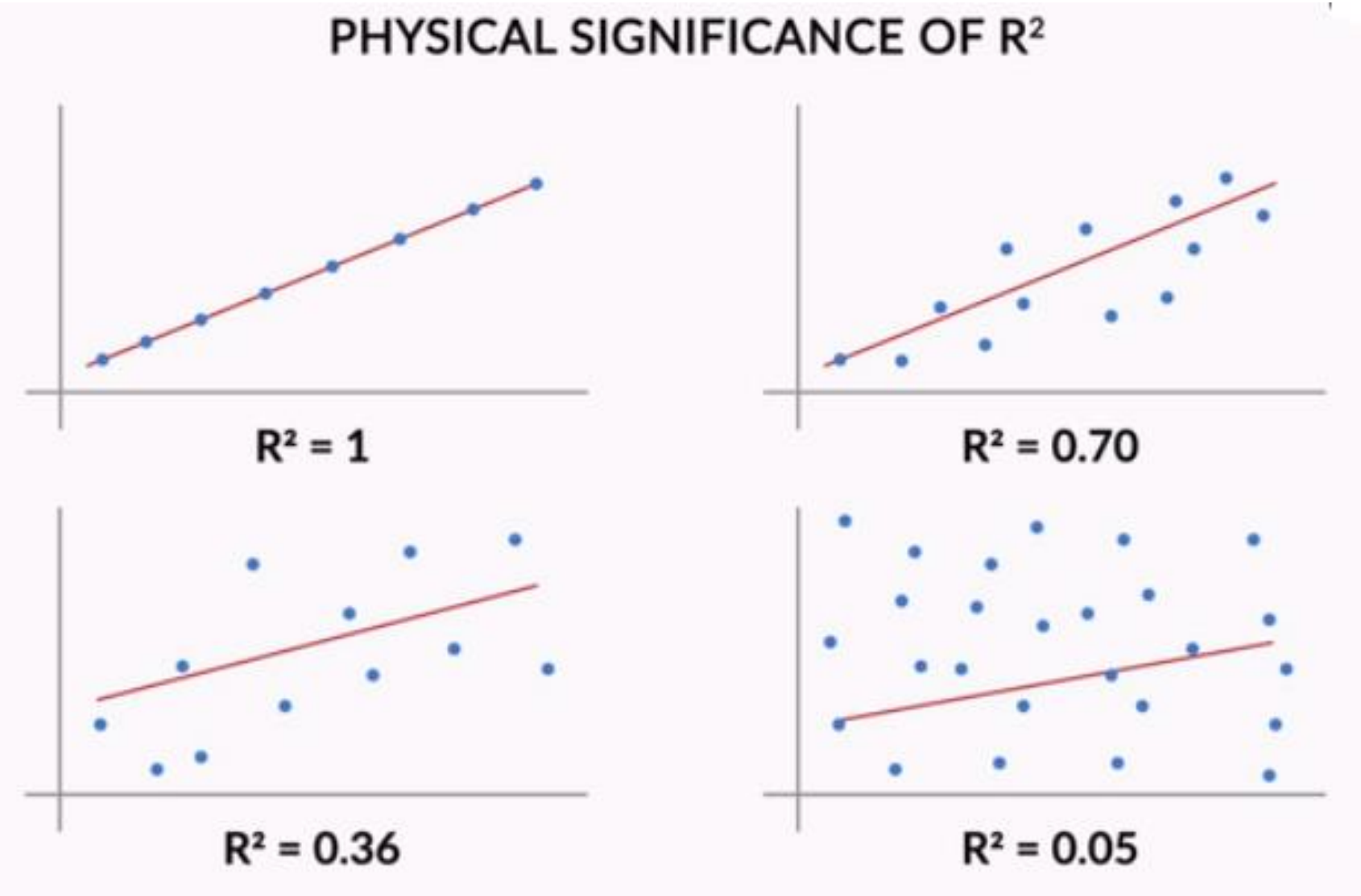
$$R^2 = 1 - \frac{RSS}{TSS}$$

Where

RSS - Residual sum of squares

TSS - Total sum of squares

Linear Regression
- Best Fit Line



Multiple Variable Linear Regression

While performing the data analysis we face multiple problems like understanding the problem statement, collecting the correct data, selecting the relevant data, which technique I should follow etc.

MULTIPLE LINEAR REGRESSION

PREDICT: SALES

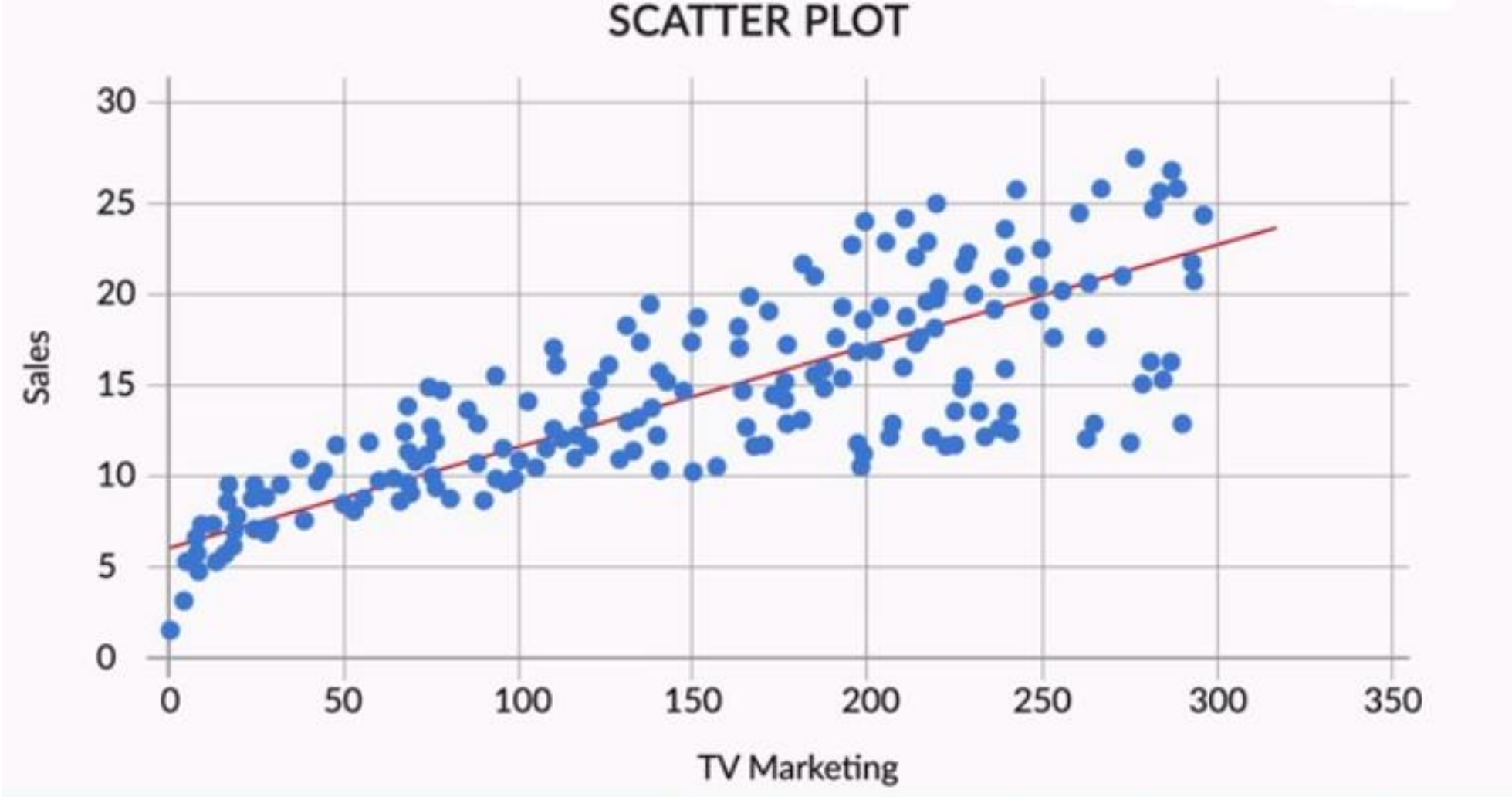
1. TV marketing
2. Newspaper marketing
3. Radio marketing

MULTIPLE LINEAR REGRESSION

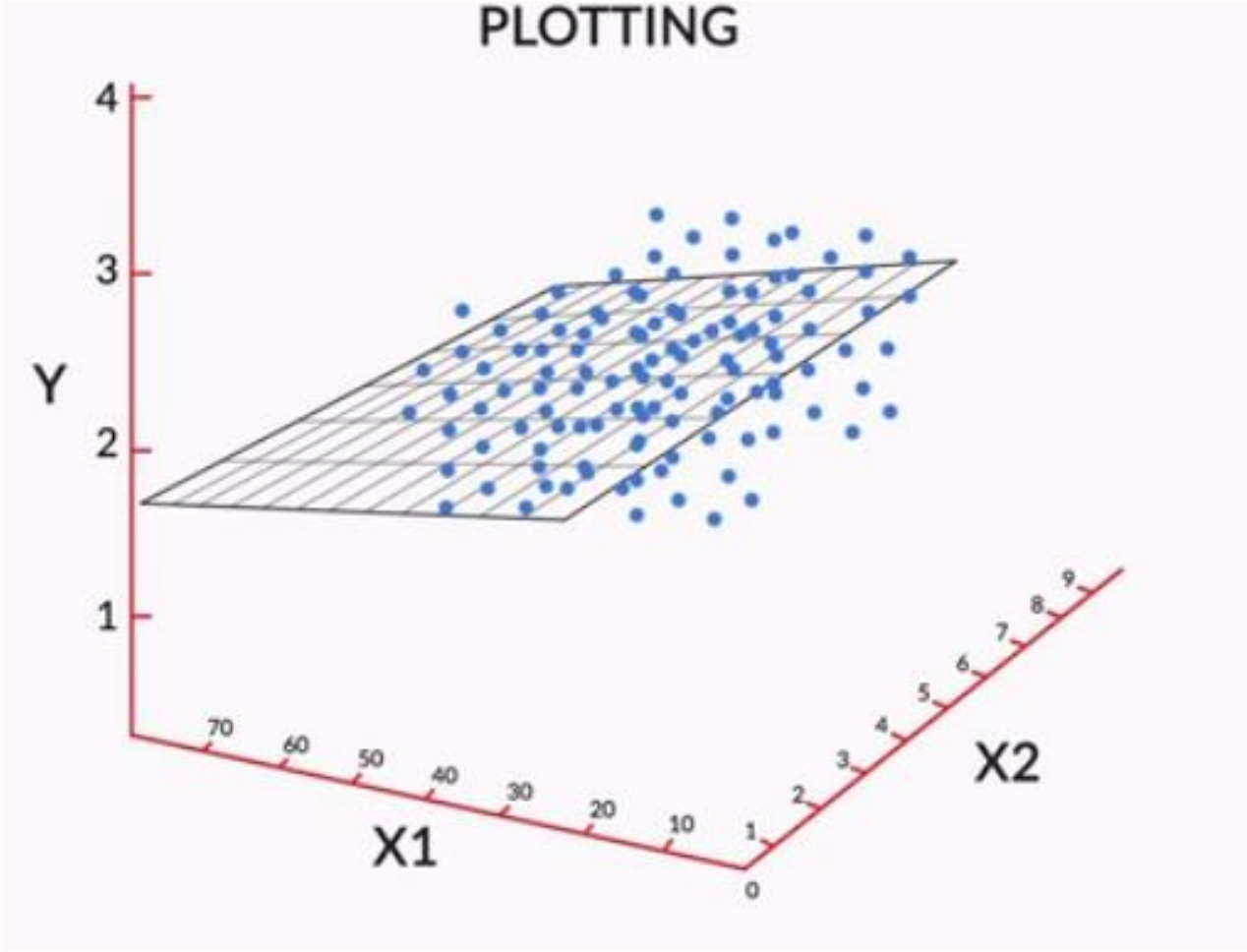
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

$$Sales = \beta_0 + \beta_1 \cdot TV + \beta_2 \cdot Newspaper + \beta_3 \cdot Radio$$

Multiple Variable Linear Regression



Multiple Variable Linear Regression



Multiple Variable Linear Regression

Till now you have understood that the variable 'Newspaper' is probably not required in the model. Is it good to drop "Newspaper" from our model? How do you take that decision?

Multiple Variable Linear Regression

- You saw that the variable 'Newspaper' is slightly correlated with 'Sales', though it is not a very significant variable in the overall model (its coefficient, compared to TV and Radio, is also quite small).

- Now, how do you decide whether the variable 'Newspaper' should stay in the model? More importantly, can you say that 'Newspaper' significantly affects 'Sales'?

- Find out the coefficient of 'NewsPaper' individually against 'Sales' and along with all other coefficients combined. See the difference.

Multiple Variable Linear Regression

- What is ‘P-Value’?

- **P-Value** signifies how much null hypothesis is true? If it is high null hypothesis is also high and vice versa.
- In this example null hypothesis is that there will not be any change in the output if we keep a variable or feature.
- If that is the case why should we keep that in the model? We can remove that feature.
- We can identify such features looking at P-Value. If it is high remove those

Multiple Variable Linear Regression – Multi Collinearity

- **Problem Statement:**
- Consider a real estate company has a data set of the prices in the region of Delhi. It wishes to use the data to optimise the sale prices of the properties, based on important factors such as area, bedrooms, parking, etc.
- Essentially the company wants:
- To identify the variables affecting house prices, e.g. area, number of rooms, bathrooms, etc.
- To create a linear model that quantitatively relates house prices with variables such as the number of rooms, area, number of bathrooms, etc.
- To know the accuracy of the model, i.e. how well these variables predict house prices.
- You can calculate the multi collinearity in the model using VIF and understand the important variables

Multiple Variable Linear Regression

- **VIF - Variance Inflation Factor**
- VIF can be a good metrics to look at to tackle multicollinearity between variables. The VIF value can range from 1 to any higher value which indicates that if the particular variable is taken into the model how much it is contributing to the multicollinearity.

Multiple Variable Linear Regression

- **Variance Inflation Factor - A Useful Measure of Multicollinearity**

- **Multicollinearity** refers to a situation where multiple predictor variables are correlated with each other. Since multiple variables are involved, you cannot use the rather simplified 'correlation coefficient' to measure collinearity (it only measures the correlation between two variables).

- Thus, you need a metric such as VIF(Variance Inflation Factor) to measure the correlation of one variable with multiple variables.

- For example, consider a set of predictor variables x1,x2,x3....xnx1,x2,x3....xn. The VIF value of x1x1 is calculated by building a multiple linear regression model with x1x1 as the target variable and all the others (x2,x3....xnx2,x3....xn) as the predictors. The VIF of x1x1 is then computed using the r-squared value of this model R1R1, i.e.:

Multiple Variable Linear Regression

$$VIF(x_1) = \frac{1}{1-R_1^2}$$

- If the value of R1R1is high, such as 0.90 (which implies that x1x1 can be predicted using the other predictors), the VIF value will be high as well, and vice-versa. Since x1x1 can be predicted using all other predictor variables, it is advisable to remove it from the model to avoid making the model unnecessarily complex.

- **Note:** In a later module, Model Selection, you will study the notion of model complexity and related concepts in detail.

- Thus, VIF is a simple and useful metric used to measure collinearity or correlation between multiple variables.

Multiple Variable Linear Regression

- To summarise, you learned the following important concepts:
- Model building using p-Value and VIF
- Why scaling of variables is important
- Evaluation of model using r-squared and adjusted r-squared

- In this case study, you only had a few variables, and one could eliminate variables by manually looking at the p-value, VIF values etc.

Multiple Variable Linear Regression

- Though in most real-life problems, you may have a much larger set of variables, say more than 100. In that case, selecting variables manually might take forever and will be a quite mundane process.

- Thus, we need an automatic way to eliminate the unnecessary variables from the model. Let's see in the next lecture how we can automate variable selection using a technique called RFE (Recursive Feature Elimination).

Multiple Variable Linear Regression

- Let's now understand about the model evaluation metric **adjusted r-squared** and why that is a more trustworthy metric than r-squared

- **adjusted r-squared** penalizes the model if there are unnecessary features thus giving little low value compared

Multiple Variable Linear Regression

The Adjusted R-Square is the modified form of R-Square that has been adjusted for the number of predictors in the model. It incorporates model's degree of freedom. The adjusted R-Square only increases if the new term improves the model accuracy.

$$R^2 \text{ adjusted} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where

R^2 = Sample R square

p = Number of predictors

N = total sample size