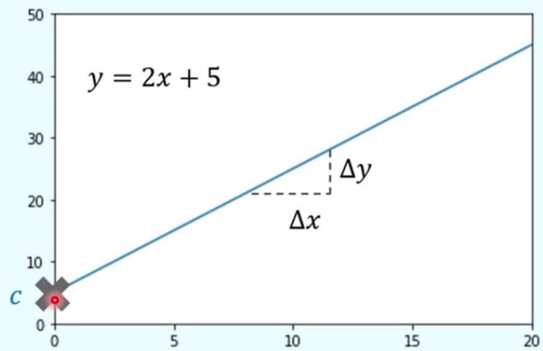
Module 6 > Topic 1 > Video 1

Introduction to Linear Regression

Introduction to Predictive Statistics

Descriptive Statistics	Vs	Inferential Statistics	Vs	Predictive Statistics
Objective is to organize, summarize and describe the given data		Objective is to make inference from the sample and make generalization about the population		Objective is to predict based on the existing data
Common tools used are 1. Visualization such as bar charts, line charts, box plots etc. 2. Statistical summary measures such as mean, median, mode, standard deviation, variance, etc.		Common tools used are 1. Probability distribution 2. Hypothesis testing, ANOVA etc.		Common tools used are 1. Linear Regression 2. Logistic Regression, Linear Discriminant Analysis

Linear Equations



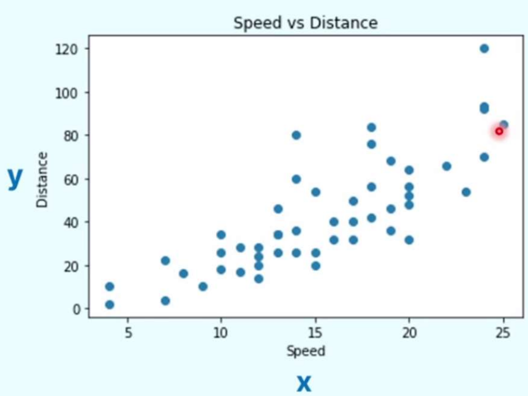
Linear equations can be written as

$$y = m.x + c$$

$$where, slope, m = \frac{y_2 - y_1}{x_2 - x_1} or \frac{\Delta y}{\Delta x}$$

$$c = constant\ or\ intercept$$

Understanding Simple Linear Regression



Speed in mph is the speed at which the car is moving

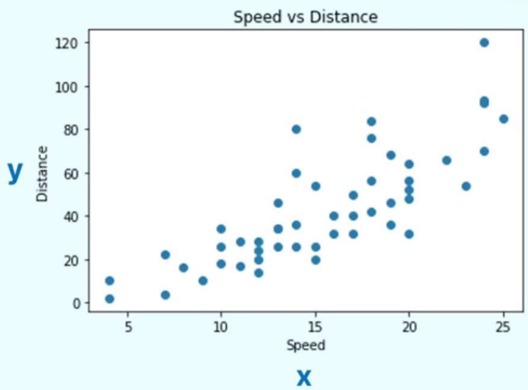
Distance in ft is the braking distance.

Braking distance is the distance between the point where the brake was applied and the point where the car stopped

Can we predict braking distance if we know the speed?

Internshala Trainings

Understanding Simple Linear Regression



Intercept (c) **Slope (m)**

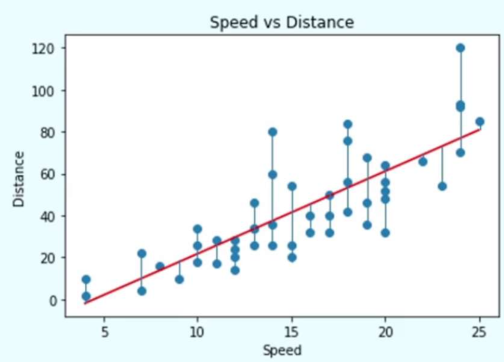
$$Y = \beta_0 + \beta_1 * X$$

$$Braking\ Distance = \beta_0 + \beta_1 * Speed$$

If we know the coefficients, β_0 and β_1, then we should be able to predict the braking distance from speed.

So how to estimate the coefficients?

Ordinary Least Squares



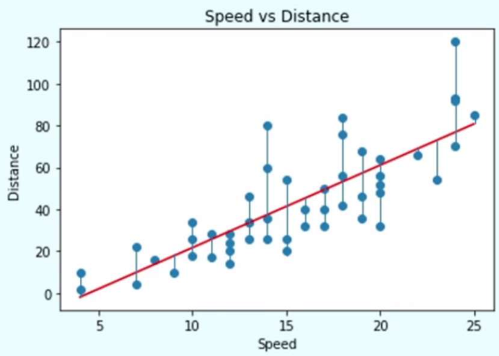
We expect the regression line to pass through most data points

The sum of squares against this line, residual sum of squares will be the least

$$Residual\ Sum\ of\ Squares = \sum_i (y_i - \hat{y}_i)^2$$

$$\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} * x_i$$

Ordinary Least Squares



By minimizing the residual sum of function, the coefficients can be computed as the following:

$$\widehat{\beta_1} = \frac{\sum_i (y_i - \bar{y}) * (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1} * \bar{x}$$

$$\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} * x_i$$

Multiple Linear Regression

We can extend the simple linear regression idea to estimate Y based on set of X variables

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \cdots + \beta_n * X_n + \varepsilon$$

Linear Regression

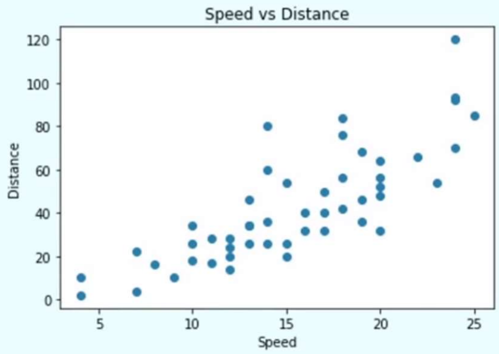
- Statistical method to estimate a linear relationship between a Y variable and a set of X variables
- The estimated linear relationship summarizes the change in Y variable (aka. Response or Dependent variable) given a unit change in the X variables. (aka. Predictors or Independent variables)
- The estimate is a useful function to predict the Y variables given new set of X values, given all underlying assumptions hold good.
- Simple Linear Regression – one x variable
- Multiple Linear Regression – multiple x variables

Module 6 | Topic 1 | Video 2

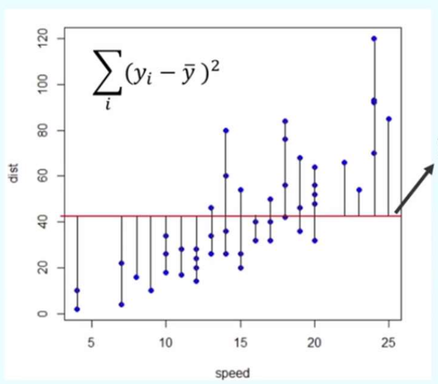
Goodness of Fit

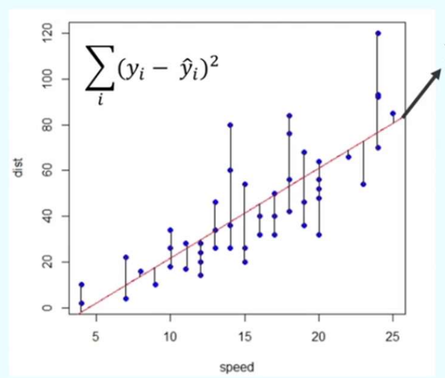
Goodness of fit (or) R-squared

Goodness of fit is a measure of how good a model is



Goodness of fit (or) R-squared





$$Total\ Sum\ of\ Squares = \sum_i (y_i - \bar{y})^2$$

$$Residual\ Sum\ of\ Squares = \sum_i (y_i - \hat{y}_i)^2$$

Goodness of fit (or) R-squared

$$Total\ Sum\ of\ Squares = \sum_i (y_i - \bar{y})^2 \qquad Residual\ Sum\ of\ Squares = \sum_i (y_i - \hat{y}_i)^2$$

Since the best fit line or the regression line would pass through most of the actual data points, we expect

$$Residual\ Sum\ of\ Square < Total\ Sum\ of\ Square$$

$$\frac{Residual\ Sum\ of\ Squares}{Total\ Sum\ of\ Square} < 1 \qquad 0 < 1 - \frac{Residual\ Sum\ of\ Squares}{Total\ Sum\ of\ Square} < 1$$

$$R^2 = 1 - \frac{Residual\ Sum\ of\ Squares}{Total\ Sum\ of\ Square}$$

Interpreting the R-squared value

$$R^2 = 1 - \frac{Residual\ Sum\ of\ Squares}{Total\ Sum\ of\ Square}$$

$if\ R^2 \cong 1$ The error is minimal therefore the model is good

$if\ R^2 \cong 0$ The error is high therefore the model is not good
The model is only as good as the mean value

Goodness of fit (or) R-squared

- R-square is a measure of goodness of fit for a regression model.
- It explains the relationship between the dependent Y variable and the independent x variables.

$$R^2 = 1 - \frac{Residual\ sum\ of\ squares}{Total\ sum\ of\ squares}$$

- $1\text{-}R^2$ is the residual or error which is what the model cannot explain due reasons such as lack of relationship between the x and y variables, or due to insufficient x variables etc.
- R^2 is also the square of the Pearson correlation coefficient R in simple linear regression, which ranges from -1 to +1, and takes a value between 0 and 1. 0 represents no relation between the x and y variables, while 1 represent a perfect relation.

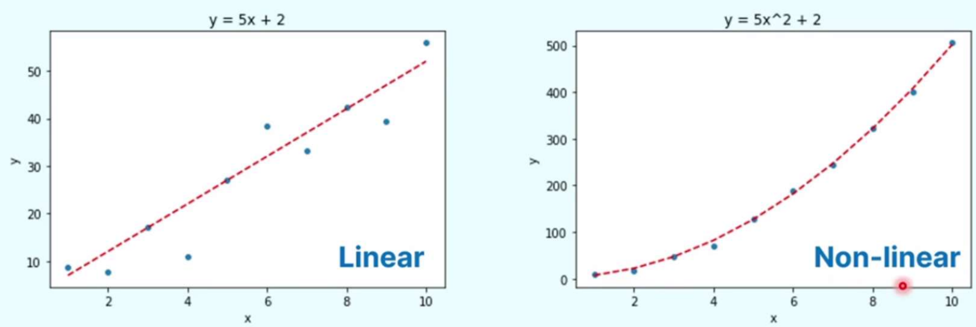
Module 6 | Topic 1 | Video 3

Linear Regression Assumptions and Conditions

Internshala Trainings

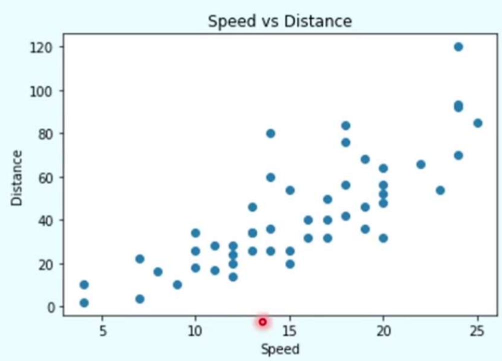
Assumptions

1. **Linearity:** relationship between the X and Y variables must be linear



Assumptions

1. **Linearity:** relationship between the X and Y variables must be linear
2. **Independence:** observations are independent of each other
3. **Normality:** For an fixed value of X, Y is normally distributed



Internshala Trainings

Assumptions

03:28 / 03:29

1. **Linearity:** relationship between the X and Y variables must be linear
2. **Independence:** observations are independent of each other
3. **Normality:** For an fixed value of X, Y is normally distributed
4. **No Multicollinearity:** X variables are not correlated to each other
5. **Normality of residuals:** residuals must be normally distributed

Module 6 › Topic 2 › Video 5

Advanced Topics in Multiple Linear Regression

Multi-Collinearity

- Multi-collinearity is a phenomenon in which, relation exist among two or more supposedly independent variables.

$$Carat = f(x, y, z, density)$$

	carat	price	x	y	z
carat	1.000000	0.921591	0.975094	0.951722	0.953387
price	0.921591	1.000000	0.884435	0.865421	0.861249
x	0.975094	0.884435	1.000000	0.974701	0.970772
y	0.951722	0.865421	0.974701	1.000000	0.952006
z	0.953387	0.861249	0.970772	0.952006	1.000000

Internshala Trainings

$$Price = f(Carat, \ldots, x, y, z)$$

- Explaining which variable is contributing to the change in the y variable becomes difficult with increasing multicollinearity among the x-variables
- Neither goodness of fit nor the model significance may suffer but interpretability becomes difficult
- Coefficients of some of the multicollinear variables may not be significant. One way to deal with that is to remove the collinear variables from the final model.

Multi-Collinearity

- Multi-collinearity is a phenomenon in which, relation exist among two or more supposedly independent variables.
- Fundamental assumption in linear regression is that the independent variables can explain the dependent variable independently of each other.
- As multi-collinearity increases, the ability to interpret the model using fewer variables diminishes.

Adjusted R-square

$$Diamond\ Price = f(Carat, Cut, Clarity, Color, \dots)$$

$$+$$

$$f(Pizzas\ eaten\ for\ dinner)$$

$$+$$

$$f(Places)$$

$$+$$

$$f(anything\ else\ that\ you\ think\ \dots)$$

Adjusted R-square

Given the regression uses least squares method to predict y variable, the model may improve with every additional x variable even due to chance. With increasing number of x variables, the model may start to overfit which can artificially inflate the R-square value. This would, therefore, undermine the model.

However, adjusted R-square improves only when the x variable explains the dependent variable sufficiently better than chance. Adjusted R-square is calculated as the following, where n is the number of observations and k is the number of independent variables.

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2) * (n - 1)}{(n - k - 1)}$$