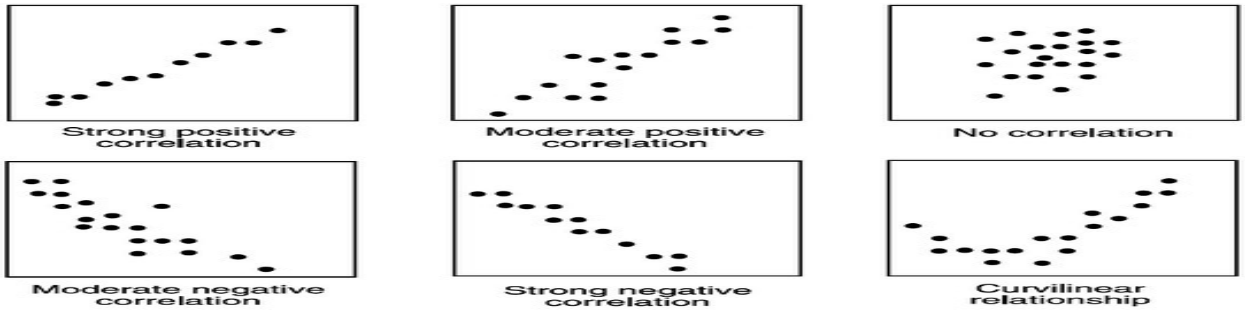
Scatter Diagram

Scatter diagrams or plots provides a graphical representation of the relationship of two continuous variables

Be Careful - Correlation does not guarantee causation. Correlation by itself does not imply a cause and effect relationship!



Judge strength of relationship by width or tightness of scatter

Determine direction of the relationship, e.g. If X increases, and Y decreases; it is negative correlation, similarly if X increases, and Y increases, it is positive correlation

Scatter Plot can show Strong positive correlation, Moderate positive correlation, No correlation, Moderate negative Correlation, Strong

Negative correlation, curvilinear relation.

Correlation Analysis

Correlation Analysis measures the degree of linear relationship between two variables

Range of correlation coefficient	-1 to +1
Perfect positive relationship	+1
Perfect negative relationship	-1

No Linear relationship	0

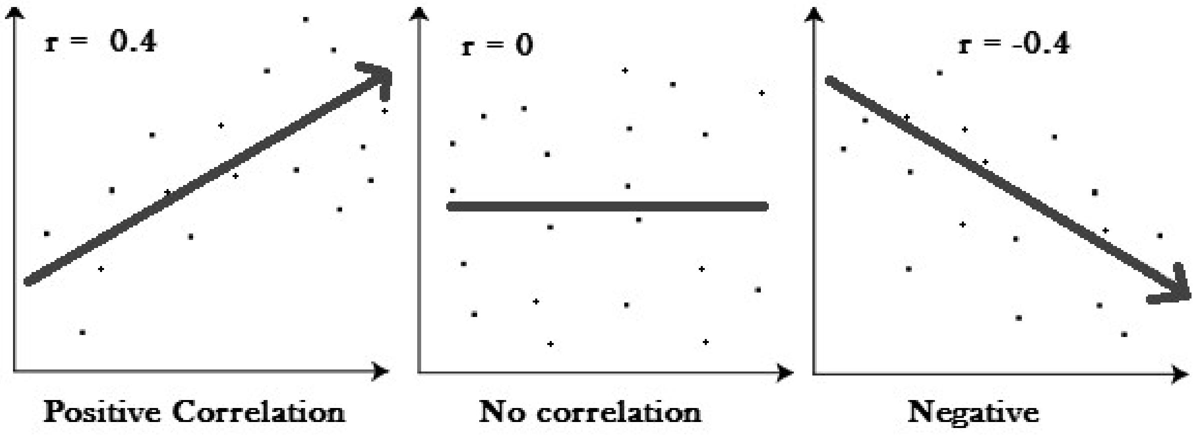
If the absolute value of the correlation coefficient is greater than 0.85, then we say there is a good relationship

- Example: r = 0.87, r = -0.9, r = 0.9, r = -0.87 describe good relationship
- Example: r = 0.5, r = -0.5, r = 0.28 describe poor relationship

Correlation values of -1 or 1 imply an exact linear relationship. However, the real value of correlation is in quantifying less than perfect relationships

We can perform regression analysis, which attempts to further describe this type of relationship, if the correlation is good between the 2 variables

Correlation Analysis:



Positive correlation: r>0

Negative correlation: r<0

No correlation: r=0

$r = (n (\sum xy)-(\sum x) (\sum y))/ (sqrt ([n\sum x^2-(\sum x)^2] [n\sum y^2-(\sum y)^2]))$

Linear Regression Model

The equation that represents how an independent variable is related to a dependent variable and an error term is a regression model

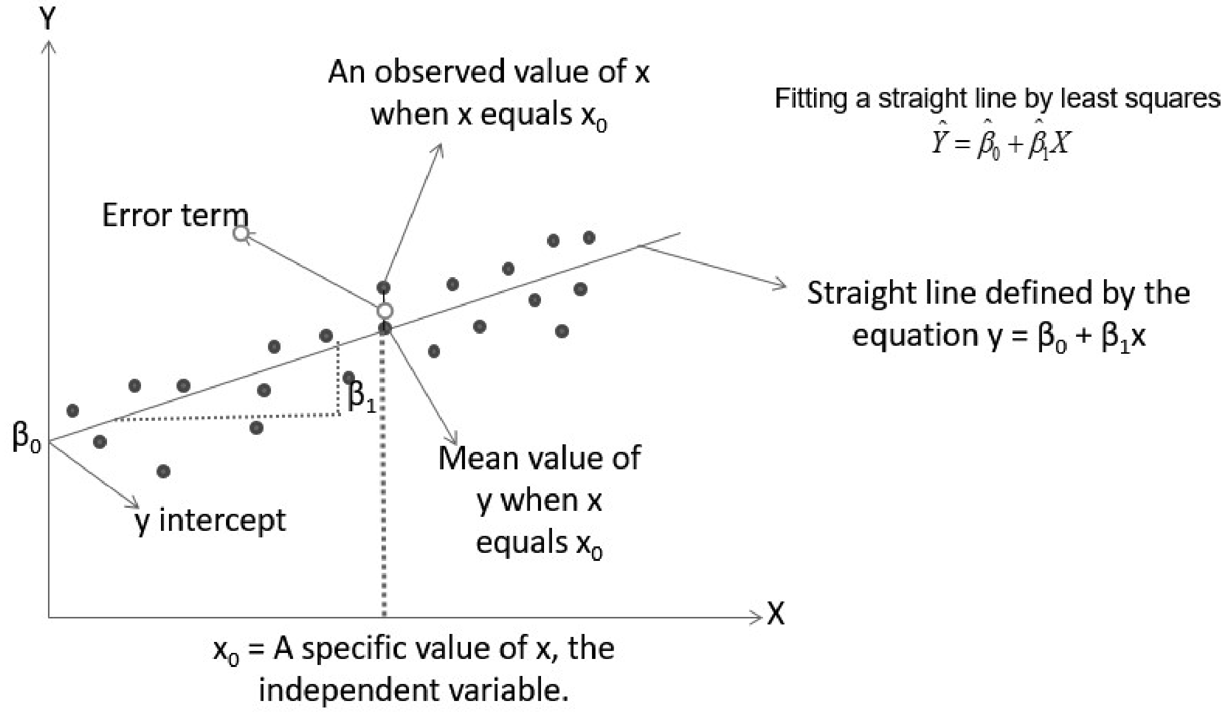
$$y = \beta_0 + \beta_1 x + \varepsilon$$

Where, β0 and β_1 are called parameters of the model,

ε is a random variable called error term.

β o= ($(\sum y)(\sum x^2)$-(∑x) $(\sum xy)$)/ ([n∑x^2-(∑x)^2])

β 1= ($(\sum xy)$-(∑x) $(\sum y)$)/ ([n∑x^2-(∑x)^2])



Y
An observed value of x when x equals x_0
Fitting a straight line by least squares
$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
Error term
Straight line defined by the equation $y = \beta_0 + \beta_1 x$
β_1
β_0
Mean value of y when x equals x_0
y intercept
X
x_0 = A specific value of x, the independent variable.

Regression Analysis

R-squared-also known as Coefficient of determination, represents the % variation in output (dependent variable) explained by input variables/s or Percentage of response variable variation that is explained by its relationship with one or more predictor variables

Higher the R^2, the better the model fits your data

R^2 is always between 0 and 100%

R squared is between 0.65 and 0.8 => Moderate correlation

R squared in greater than 0.8 => Strong correlation

R^2=SSR/SST= (SSR/(SSR+SSE))

0<=R^2<=1

Mathematically

SSR =$\sum(\hat{y}$-$\overline{y})^2$ → measure of an explained variation

SSE =$\sum(y - \hat{y})^2$→ measure of an unexplained variation

SST = SSR+SSE =$\sum$(y-$\overline{y})^2$ → measure of total variation in y

Regression Analysis

Prediction and Confidence Interval are types of confidence intervals used for predictions in regression and other linear models

Prediction Interval: Represents a range that a single new observation is likely to fall given specified settings of the predictors

Confidence interval of the prediction: Represents a range that the mean response is likely to fall given specified settings of the predictors

The prediction interval is always wider than the corresponding confidence interval because of the added uncertainty involved in predicting a single response versus the mean response

Regression Techniques – Multiple Linear Regression

Y-continuous, x – Multiple & continuous

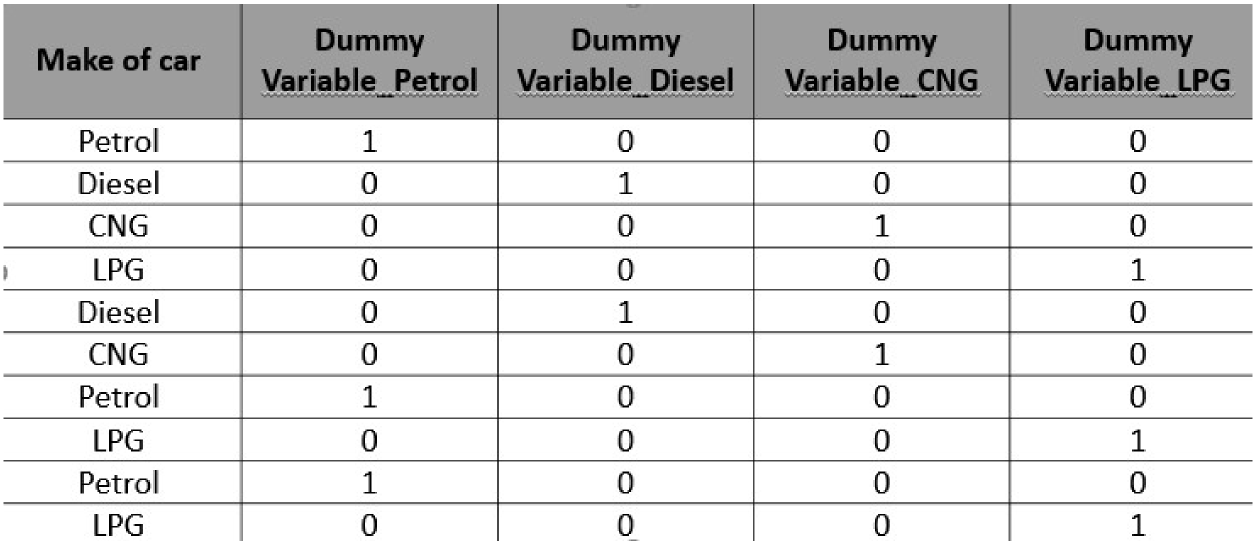
We apply Multiple linear Regression

Y-continuous, x – Multiple & discrete

We create dummy variable for discrete component and

We then apply Multiple linear Regression

Multiple Linear Regression – Dummy Variable



Make of car	Dummy Variable_Petrol	Dummy Variable_Diesel	Dummy Variable_CNG	Dummy Variable_LPG
Petrol	1	0	0	0
Diesel	0	1	0	0
CNG	0	0	1	0
LPG	0	0	0	1
Diesel	0	1	0	0
CNG	0	0	1	0
Petrol	1	0	0	0
LPG	0	0	0	1
Petrol	1	0	0	0
LPG	0	0	0	1

Multiple Regression Model

DATA : CARS, 81 observations, *"cars.csv"*

- VOL = cubic feet of cab space

- HP = engine horsepower
- MPG = average miles per gallon
- SP = top speed, miles per hour
- WT = vehicle weight, hundreds of pounds

Our interest is to model the MPG of a car based on the other variables.

Model and Assumptions

Our Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_k X_k + \varepsilon$

Linear

Independent

Normal

Equal Variance

Linearity (Assumptions about the form of the model):

- Linear in parameters
- Assumptions about the errors:
- IID Normal (Independently & identically distributed)
- Zero mean

- Constant variance (Homoscedasticity)
- If no constant variance (HETEROSCEDASTICITY)
- Independent of each other. If not independent, it is called as AUTO CORRELATION problem
- Assumptions about the predictors:
- Non-random
- Measured without error
- Linearly independent of each other. If not it is called as COLLINEARITY problem
- Assumptions about the observations:
- Equally reliable