

## Linear Regression:

\* mostly applicable on Data Analysis, Building models

Motivation:

\* purpose is to build a functional relationship (model) between dependent Variable(s) and independent variable(s).

Example:

Business: what is the effect of price on Sales? (can be used to fix the selling price of an item)

↓ Result:

To increase market sale.

Engineering: Can we infer difficult to measure properties of a product from other easily measured variables? (mechanical strength of a polymer from temperature, viscosity or other process variables) - also known as Soft Sensor.

↓

→ Develop a model

→ use the model to predict the mechanical strength, temp., viscosity → also known as Soft Sensor

Software Sensor in literature.

~~This model is used in practise to measure the variables~~

\* This model is used in practise to infer the values of Variables which are difficult to measure using an instrument.

Purpose:

- (i) we are building the model for given purpose
- (ii) purpose is defined depending the area that you are working

\* Regression

→ one one of the widely used Statistical technique.

→ Dependent Variables also known as response Variable, Regress and, Predicted Variable, Output Variable - denoted as variable/s  $y$ .

→ Variable whose output is designed to predict based on model : Symbolic way is output  $y$ .

→ Independent Variable (also known as predictor, Regressor / Exploratory Variables or input Variable indicated by  $x$ ).

e.g. For a plant growth, if we apply fertilizers, fertilizer is independent variable.  
plant growth is Dependent Variable.

### Methods & Types of Regression:

#### Classification of Regression Analysis

- \* univariate: one dependent and one independent variable
- \* multivariate: many dep. and many independent Variables.
- \* Linear vs Nonlinear
  - Linear: Relationship is linear between dependent and independent Variables.
  - Nonlinear: Relationship is nonlinear between dependent and independent Variables.
- \* Simple vs Multiple:
  - Simple: one dependent and one independent Variable (SISO)
  - multiple: one dependent and many independent Variables (miso)

## Linear Regression methods:

→ Simple Linear Regression

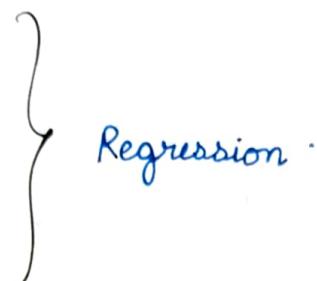
→ Multiple

→ Ridge

→ Principal component

→ Lasso

→ partial Least Squares



## Non Linear Regression methods:

★ Polynomial }  
★ Spline }      Regression  
★ Neural networks

## Simple Linear Regression Model:

(what is Regression?)

→ we can estimate value of one Variable with value of another Variable which is known. The Statistical method which helps us to estimate the unknown value of one Variable from known value of related Variable is called Regression Line of Regression:

→ The line described in average relationship b/w 2 Variables.

→ now-a-days we call this as Estimating line

## Uses of Regression:

1. Estimate relation b/w 2 Variables.  
Eg: like 2 Economic variables like Income and Expectation.
2. Highly valuable tool in Business and economics.
3. Highly used in prediction.
4. we can find relation b/w coefficient of correlation and coefficient of determination.
5. useful in Statistical Estimator in demand curves, Supply curves, function, Cost function, Consumption function etc

## Comparison between Correlation and Regression.

(176)

### 1. Degree of Covariability

→ measures of degree of covariability b/w 2 Variables by regression established by functional relationship is regression (functional) established functional relation b/w dependent and independent Variables so that (former) can be dependent predicted for a given value of the (latter?) In Correlation (analysis) x and y are random Variables whereas in Regression x is R.V and y is fixed measure.

Correlation → Relative Measure.

Absolute Measure → Regression.

## Regression Model:

### Simple Linear Regression Model:

The equation that describes how  $y$  is related to  $x$  and Error term

Eg:

$$\mu = \beta_0 + \beta_1 x + \epsilon$$

$\beta_0, \beta_1$  are parameters

$\epsilon$  is Random variable called error term.

$$E(y) = \beta_0 + \beta_1 x \quad [\text{no error term}]$$

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

$E(y)$  is Expectation of  $y$

In calculating  $\beta_1$ , we minimized the errors predicting  $y$  by  $E(x)$  i.e mean value of  $y$  or  $E(y)$

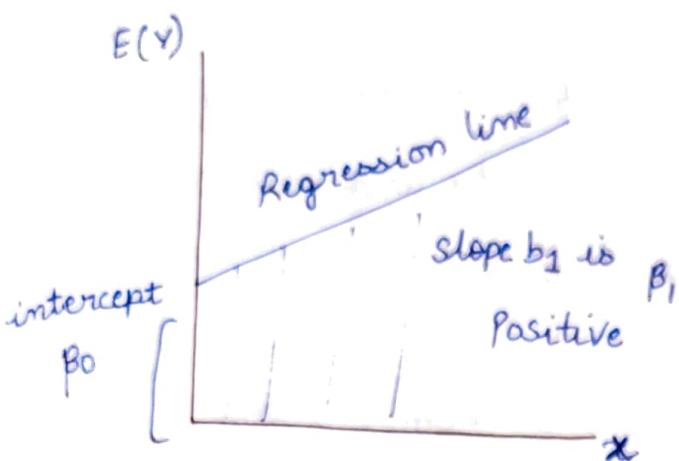
Graph is Straight line

$\beta_0$  is  $y$  intercept of Regression line

$\beta_1$  is Slope of the Regression line.

$E(y)$  is Expected value of  $y$  for an given  $x$  value.

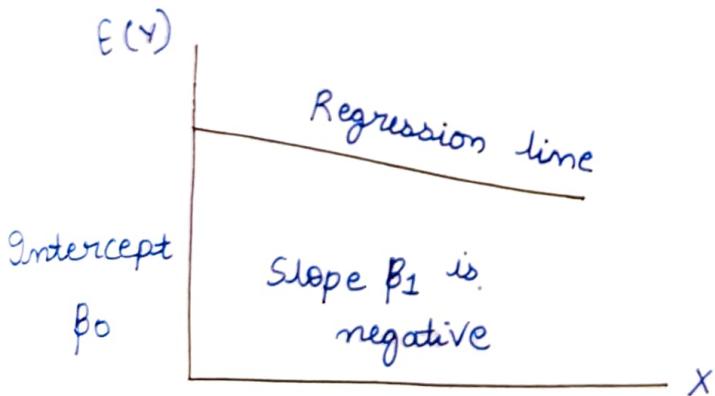
## Positive Linear Relationship



$x \uparrow, y$  is also  $\uparrow$   
or  $E(Y)$

$$E(Y) = \beta_0 + \beta_1 x$$

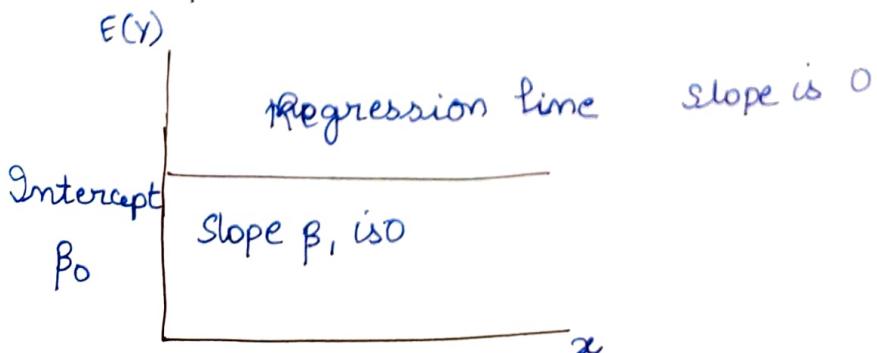
## Negative Linear Relationship



$x \uparrow y$  is  $\downarrow$

$\therefore$  Slope is negative

## No Relationship



for any value of  $x$ ,  $E(Y)$  are same

# Geometrical Interpretation of Linear Regression

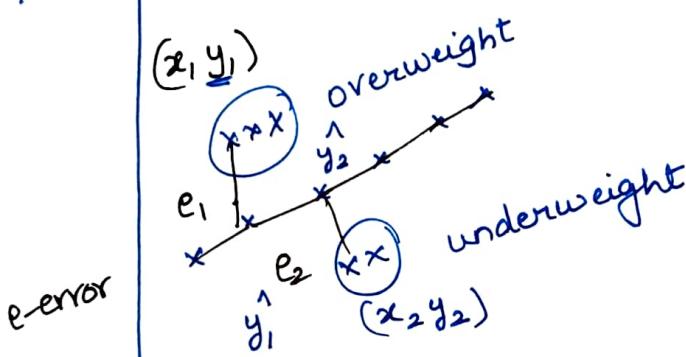
predict height of a person with features  
weight, gender, Ethnicity and hair colour.

Height is a Real number

$$h \in \mathbb{R}$$

Eg: 165.6 cm, 153.2 cm, 175.5 cm

Y-axis  $y = \text{height}$



$$f_1 = \text{weight}$$

weight ↑ height ↑

x axis

Exceptions:

overweight }  
under } Extreme points.

Line has to be drawn on all the given points.  
So error is minimized.

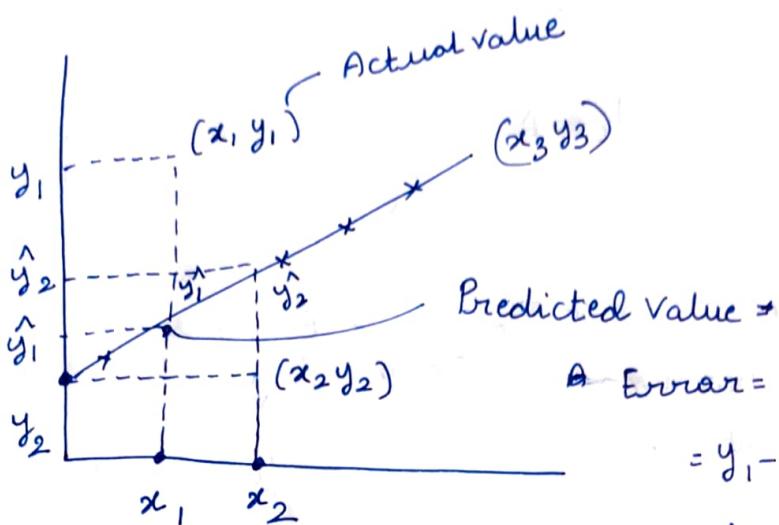
Error = Actual Value - Predicted Value

$$e_1 = y_1 - \hat{y}_1$$

$$e_2 = y_2 - \hat{y}_2$$

Sum of errors =  $e_1^2 + e_2^2 + \dots + e_n^2$   
must be minimized.

which ever line is minimized is least square  
 $\downarrow$   
 Sum of errors method [is principle].



$$\text{Error} = A \cdot V - P \cdot V \\ = y_i - \hat{y}_i$$

$$e_1 = y_1 - \hat{y}_1 = +ve$$

$$e_2 = y_2 - \hat{y}_2 = -ve$$

error for  $x_2$

above regression line  $\rightarrow +ve$   
 below  $\rightarrow -ve$

Some point on Regression is 0

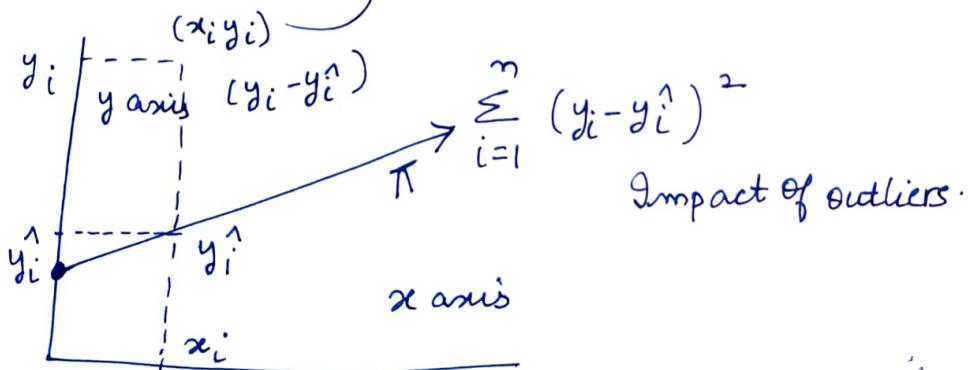
if we Sum of errors =  $e_1^2 + e_2^2 + e_3^2 + \dots + e_n^2$   
 $\rightarrow +ve$

Ex:  
 $e_1 = 10$   
 $e_2 = -10$

$e_1 + e_2 = 10 + (-10) = 0$

but error exists above line

If not Sum of errors errors may be nullified.



Impact of outliers.

For trained data (we find)

$$\text{Eg: } y = mx + c$$

$$y_i - \hat{y}_i \uparrow$$

remove as outliers

$$\begin{matrix} D \\ \text{Training} \\ \text{Data set} \end{matrix} = \begin{matrix} D_{\text{Train}} \\ - \text{Outliers} \\ \downarrow \\ \text{far away} \end{matrix}$$

Repeat this procedure.

This is Random Sampling process.

Least Squares Method

Least Square Criterion

$$\min \sum (y_i - \hat{y}_i)^2$$

where:

$\hat{y}_i$  predicted value  
 $y_i$  Actual

$y_i$  = Observed Value  
for the  $i$ th observation.

$\hat{y}_i$  = Estimated Value of dependent Variable  
for  $i$ th observation.

in Least Squares method,

(i) we first find errors

(ii) Square of the errors

(iii) Sum of the Squared errors.

$$\min \sum_{i=1}^n (y_i - \hat{y}_i)^2$$
$$(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \dots + (y_n - \hat{y}_n)^2$$

Squared Error (SE) =  $\sum (y_i - \hat{y}_i)^2$

$$y = mx + b$$

$$\Rightarrow (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 + \dots$$
$$(a-b)^2 \quad \quad \quad (y_n - (mx_n + b))^2$$

$$\left\{ \begin{array}{l} \hat{y}_1 = mx_1 + b \\ \hat{y}_2 = mx_2 + b \\ \vdots \\ \hat{y}_n = mx_n + b \end{array} \right. \quad \text{error} = AV - PV \quad \left. \right\}$$

$m, b$  are parameters such that Squared error is minimum.

$$(y_1 - (mx_1 + b))^2 \Rightarrow y_1^2 + (mx_1 + b)^2 - 2y_1(mx_1 + b)$$
$$(a+b)^2$$

$$y_1^2 + (mx_1^2 + b^2 + 2mx_1b) - 2y_1x_1m - 2by_1$$

Similarly for  $(y_2 - (mx_2 + b))^2$

$$\begin{aligned}
 SE &= (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 + \dots \\
 &\quad (y_n - (mx_n + b))^2 \\
 &= y_1^2 - 2y_1(mx_1 + b) + (mx_1 + b)^2 + \\
 &\quad y_2^2 - 2y_2(mx_2 + b) + (mx_2 + b)^2 + \dots \\
 &\quad + y_n^2 - 2y_n(mx_n + b) + (mx_n + b)^2 \\
 &= y_1^2 - 2x_1 y_1 m - 2y_1 b + m^2 x_1^2 + 2mx_1 b + b^2 + \\
 &\quad y_2^2 - 2x_2 y_2 m - 2y_2 b + m^2 x_2^2 + 2mx_2 b + b^2 + \\
 &\quad \dots + y_n^2 - 2x_n y_n m - 2y_n b + m^2 x_n^2 + \\
 &\quad 2mx_n b + b^2
 \end{aligned}$$

Grouping terms

$$\begin{aligned}
 &= (y_1^2 + y_2^2 + \dots + y_n^2) - 2m(x_1 y_1 + x_2 y_2 + \dots + x_n y_n) \\
 &\quad - 2b(y_1 + y_2 + \dots + y_n) + m^2(x_1^2 + x_2^2 + \dots + x_n^2) \\
 &\quad + 2mb(x_1 + x_2 + \dots + x_n) + (b^2 + b^2 + \dots + b^2) \\
 &= n\bar{y}^2 - 2mn\bar{x}\bar{y} - 2bn\bar{y} + m^2\bar{x}^2 + \\
 &\quad 2mbn\bar{x} + nb^2
 \end{aligned}$$

$$\left\{ \begin{array}{l} \bar{y}_2 = \frac{y_1^2 + y_2^2 + \dots + y_n^2}{n} \\ \text{Proof:} \end{array} \right.$$

$$\boxed{ny_2} = \underbrace{y_1^2 + y_2^2 + \dots + y_n^2}_{\text{Sub}} \quad \underline{\underline{=}}$$

$$\bar{xy} = \frac{x_1y_1 + x_2y_2 + \dots + x_ny_n}{n}$$

$$n\bar{y} = x_1y_1 + x_2y_2 + \dots + x_ny_n$$

$$\bar{y} = \frac{y_1 + y_2 + \dots + y_n}{n}$$

$$n\bar{y} = y_1 + y_2 + \dots + y_n$$

$$\bar{x^2} = \frac{x_1^2 + x_2^2 + \dots + x_n^2}{n}$$

$$x_1^2 + x_2^2 + \dots + x_n^2 = n\bar{x}^2$$

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

$$n\bar{x} = x_1 + x_2 + \dots + x_n$$

$$\text{Standard error} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2} = \sqrt{\frac{1}{n} \sum_{i=1}^n (m\bar{x} + b - \bar{y})^2} = \sqrt{\frac{1}{n} \sum_{i=1}^n (m^2 \bar{x}^2 + 2mb\bar{x} + b^2 - 2m\bar{x}\bar{y} - 2b\bar{y} + \bar{y}^2)} = \sqrt{m^2 \bar{x}^2 + 2mb\bar{x} + b^2 - 2m\bar{x}\bar{y} - 2b\bar{y} + \bar{y}^2}$$

$$y = mx + b$$

$m, b$  are parameters

find  $m, b$

differentiate partially

$$\frac{d}{dx} (x^n) = nx^{n-1}$$

$$\frac{d}{dx} (1) = 1$$

$$\frac{\partial (SE)}{\partial m} = -2n\bar{x}\bar{y} + 2m\bar{x}^2 + 2bn\bar{x} = 0$$

$$\frac{\partial (SE)}{\partial b} = -2n\bar{x}\bar{y} + 2m\bar{x}^2 + 2bn\bar{x} = 0$$

[2m common]

$$- \bar{xy} + \bar{mx^2} + b\bar{x} = 0$$

$$\bar{mx^2} + b\bar{x} = \bar{xy}$$

$$m \frac{\bar{x^2}}{\bar{x}} + b = \frac{\bar{xy}}{\bar{x}}$$

{ ÷ with  $\bar{x}$  }

$$mx + b = y$$

one Point  $\left( \frac{\bar{x^2}}{\bar{x}}, \frac{\bar{xy}}{\bar{x}} \right)$

$$SE = \frac{ny^2}{x} - 2mn\bar{x}\bar{y} - 2bn\bar{y} + m^2n\bar{x}^2 + 2mbn\bar{x} + nb^2$$

differentiating partially wrt b [constant 0]

$$\frac{\partial (SE)}{\partial b} = -2n\bar{y} + 2mn\bar{x} + 2nb = 0$$

$$= -\bar{y} + m\bar{x} + b = 0$$

$y = mx + b$  (passing through  $\bar{x}, \bar{y}$ )

$$\bar{y} = m\bar{x} + b$$

another point  $(\bar{x}, \bar{y})$

$$\frac{dy}{dx} = 0$$

$$\frac{d^2y}{dx^2} < 0$$

$\therefore$  line must pass

through

$$\left( \frac{\bar{x}^2}{\bar{x}}, \frac{\bar{xy}}{\bar{x}} \right) \text{ and } (\bar{x}, \bar{y})$$

we get max.  
value

SLOPE:  $\left( \frac{\bar{x}^2}{\bar{x}}, \frac{\bar{xy}}{\bar{x}} \right), (\bar{x}, \bar{y})$

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\bar{y} - \frac{\bar{xy}}{\bar{x}}}{\bar{x} - \frac{\bar{x}^2}{\bar{x}}} = \frac{\bar{x}\bar{y} - \bar{xy}}{(\bar{x})^2 - \bar{x}^2}$$

$$m = \frac{\bar{xy} - \bar{x}\bar{y}}{\bar{x}^2 - (\bar{xy})^2} = \frac{\text{Cov}(x, y)}{\text{Var}(x)}$$

$$\text{Cov}(x, y) = E(x - \bar{x})(y - \bar{y})$$

$$= E(xy - x\bar{y} - \bar{x}y + \bar{x}\bar{y})$$

$$E(xy) = \bar{x}\bar{y} - \frac{E}{\bar{x}\bar{y}} (xy) = \bar{y}\bar{x} - \cancel{\bar{x}\bar{y}} + \cancel{\bar{x}\bar{y}}$$

$$\begin{aligned}\text{Var}(x) &= E(x - \bar{x})^2 \\ &= E(x^2 - 2x\bar{x} + \bar{x}^2) \\ &= E(x^2) - 2(\bar{x})^2 + (\bar{x})^2 \\ &= \bar{x}^2 - (\bar{x})^2\end{aligned}$$

$$m = \frac{\bar{xy} - \bar{x}\bar{y}}{\bar{x}^2 - (\bar{x})^2}$$

$\therefore x$  is Covariance of  $(x, y)$  and Variance of  $x$

$$y = mx + b$$

$$y = b_0 + b_1 x$$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$y$  intercept for estimated Linear Regression

$$\begin{aligned}\bar{y} &= b_0 + b_1 \bar{x} && \rightarrow \text{Parameters} \\ b_0 &= \bar{y} - b_1 \bar{x}\end{aligned}$$

## Multiple Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon$$

If more than 1 independent variable

### Multiple Regression Equation

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

$$[E(\epsilon) = 0]$$

$\beta_0, \beta_1, \beta_2, \dots, \beta_p$  ~~are~~  
are unknown  
parameters

The Estimation process for Multiple Regression

To find unknown parameters from the population we collect Sample data

$y$  is dependent variable

Sample data

$$x_1 \ x_2 \ \dots \ x_p \ y$$

⋮ ⋮ ⋮ ⋮



Compute the estimated multiple  
Regression model Equation

$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p$$

$b_0, b_1, b_2, \dots, b_p$  are Sample Statistics.

↳  $b_0, b_1, b_2, \dots, b_p$  provides  
the estimates of  $\beta_0, \beta_1, \beta_2, \dots, \beta_p$

How can we decide  $\beta_0, \beta_1, \beta_2, \dots, \beta_n$   
are equal to 0 or not

→ By using Significant test

Similar to Linear Regression model -

Simple vs multiple Regression.

In Simple linear regression,  $b_0$  and  $b_1$ , were the  
Sample Statistics used to estimate the  
parameters  $\beta_0$  and  $\beta_1$ .

Multiple regression parallels this Statistical  
inference process, with  $b_0, b_1, b_2, \dots, b_p$  denoting  
the Sample Statistics used to estimate the  
parameters  $\beta_0, \beta_1, \beta_2, \dots, \beta_p$

$$b_0 \rightarrow \beta_0$$

$$b_1 \rightarrow \beta_1$$

$$b_2 \rightarrow \beta_2$$

$$b_3 \rightarrow \beta_3$$

:

$$b_p \rightarrow \beta_p$$

Sample Statistics }      } Going to predict  
                                  } the population parameters  
                                  } with help of  
                                  } Sample Statistics

In Simple regression we have only 1 independent Variable

multiple → more than 1 independent Variable

## Example Example : Trucking Company:

[source: Statistics for Business and Economics, 2012, Anderson]

As an illustration of multiple regression analysis, we will consider a problem faced by the Trucking Company.

A major portion of business involves deliveries throughout its local area.

To develop better work Schedules, the managers want to estimate the total daily travel time for their drivers.

preliminary data for butler trucking:

Driving assignment       $x_1$  = miles travelled       $y$  = Travel Time (hours)

	Independent Variable	$x_1$ = miles travelled	no. of deliveries	dependent	
0	1	100	4	9.3	
1	2	50	3	4.8	
2	3	100	4	8.9	
3	4	100	2	6.5	
4	5	50	2	4.2	
5	6	80	3	6.2	
6	7	75	3	7.4	
7	8	65	4	6.0	
8	9	90	3	7.6	
9	10	90	2	6.1	

$x_1$  miles travelled

↳ deliveries → independent .

$x_1$  miles travelled

Dependent Variable: travel-time

Independent Variables:

$x_1$  = miles travelled

n-of-deliveries -

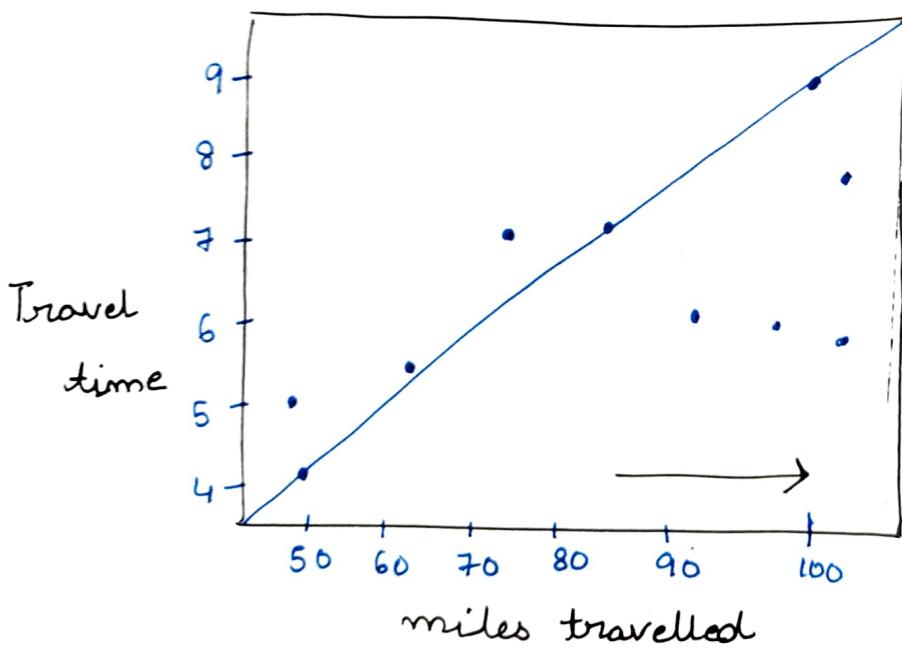
lets

understand the relation b/w  $x_1$  miles travelled  
and travel time.

we need to draw Scatter plot:

Scatter diagram of preliminary Data for  
Trucking  $x_1$

Simple linear regression with miles travelled

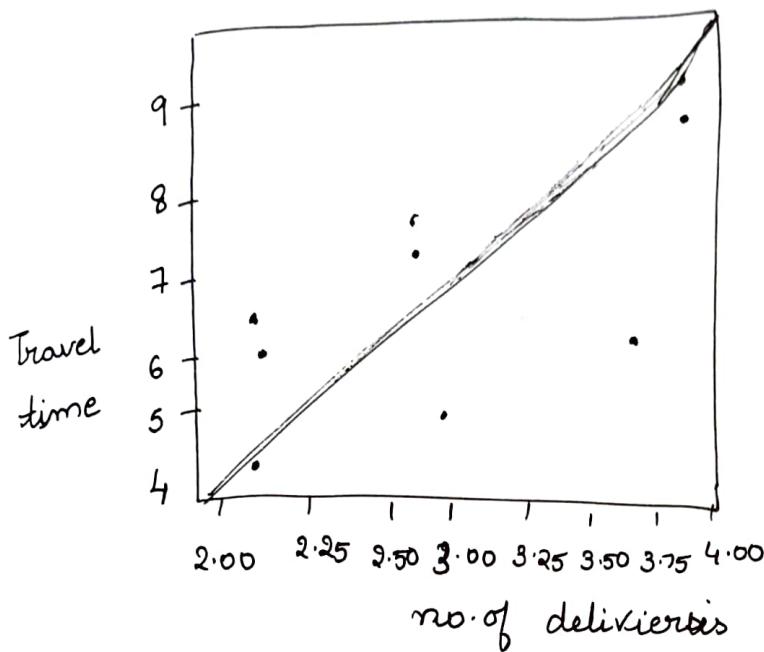


✓ Relation between one independent and one dependent Variable.

Travel time depends on miles travel

now consider no. of deliveries

Simple linear regression with number of deliveries.

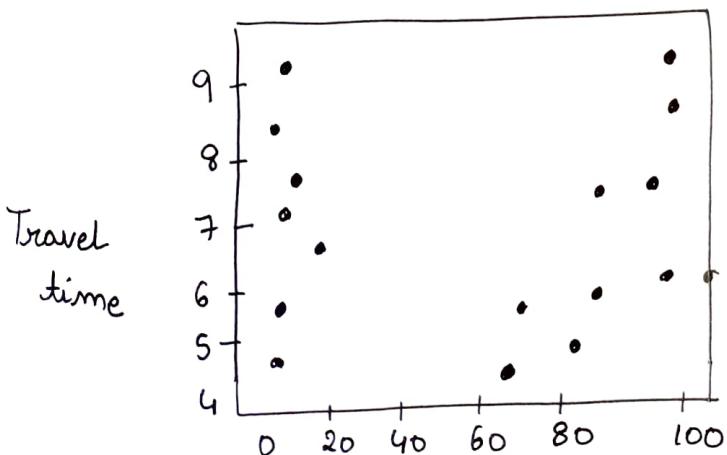


linear regression

There is a +ve correlation

? why no correlation  $\rightarrow$  If there is no correlation there is no relation between dependent and independent variable so regression model is not used in such cases.

Multiple regression.



miles travelled and number of deliveries

Multiple regression:

$$\text{Eq: } \hat{y} = -869 + 0.611x_1 + 0.923x_2$$

↓      ↓      ↓  
travelled      distance      no. of deliveries  
time              travelled

multiple Coefficient of determination

Relationship among SST, SSR and SSE

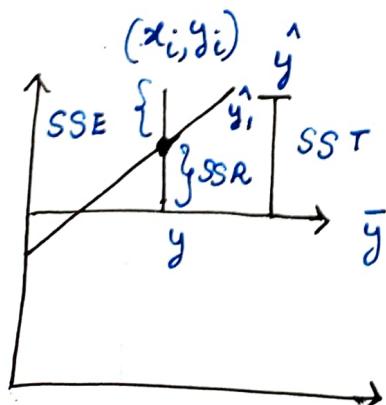
$$SST = SSR + SSE$$

where

$$SST = \text{total Sum of Squares} = \sum (y_i - \bar{y})^2$$

$$\begin{aligned} SSR &= \text{Sum of Squares due to regression} \\ &= \sum (\hat{y}_i - \bar{y})^2 \end{aligned}$$

$$\begin{aligned} SSE &= \text{Sum of Squares due to error} \\ &= \sum (y_i - \hat{y}_i)^2 \end{aligned}$$



## Adjusted Multiple Coefficient of Determination

$$SST = SSR + SSE$$

$$SSR = SST - SSE$$

$\downarrow$   
↳ decreased

$$\text{Increase } R^2 = \frac{SSR}{SST}$$

purpose of Adjusted  $R^2$

If we add new Independent Variable

Adjusted  $R^2$  will check if new independent variable is explaining variable or noise variable

$n$  = number of Observations

$p$  = denoting the no. of independent Variables

Adjusted  $R^2$

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

$$R^2_a = 1 - (1 - 0.904) \frac{10-1}{10-2-1} = 0.88$$

Supposely

$y$  is dependent

$x_1, x_2, x_3, x_4 \rightarrow$  Independent variables

$y \rightarrow x_1, R^2$

$\uparrow \swarrow$

Adj  $R^2$

$\uparrow$

$y \rightarrow x_1, x_2 \uparrow$

$\uparrow$

$y \rightarrow x_1, x_2, x_3$  ( $x_3$  is noise data)

noisy value implies it will not explain  $y$ ,  
and disturbs existing relation

$$y \rightarrow x_1 x_2 x_3$$
$$\begin{array}{c} R^2 \\ \uparrow \end{array} \qquad \begin{array}{c} \text{Adj } R^2 \\ \downarrow \end{array}$$

If  $R^2$  and  $\text{Adj } R^2$  are same we need not add Variables

If  $R^2$  is 0.9  $\rightarrow$  we can add more values to fill  
 $\text{Adj } R^2 = 0.3$  the Gaps

Adj multiple Coefficient Vs multiple Coefficient

Every time you add a independent Variable to a model, the  $R$ -Squared increases, even if the independent Variable is insignificant. It never declines. whereas Adjusted  $R$ -Squared increases only when independent Variable is Significant and affects dependent Variable.

In the table below, adjusted  $r$ -Squared is maximum when we included two Variables. It declines when third Variable is added whereas  $r$ -Squared increases when we included third Variable. It means third Variable is insignificant to the model.

Variables	R-Squared	Adjusted R-Squared
1	67.5	67.1
2	85.9	84.2
3	88.9	81.7
	—	decreasing
Adj	mscr.	

Adjusted r-Squared can be negative when r-Squared is close to zero.

Adjusted r-Squared Value always be less than or equal to r-Squared value.

### Testing for Significance:

F test

t-test

The F test is used to determine whether a Significant relationship exists between the dependent Variable and the Set of all the independent Variables; we will refer to the F test as the test for overall Significance.

If the f test shows an overall Significance, the t test is used to determine whether each of the individual independent Variable is Significant.

A Separate t test is conducted for each of the independent Variables in the model; we refer to each of these t tests as a test for individual Significance.

F test:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon$$

The hypotheses for the F test involve the parameters of the multiple regression model.

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$$

$H_a$ : one or more of the parameters is not equal to zero.

Test Statistic

$$F = \frac{MSR}{MSE}$$

Rejection rule:

P Value approach

Reject  $H_0$  if  $P\text{-value} \leq \alpha$

Critical Value approach

Reject  $H_0$  if  $F \geq F_\alpha$

where

$F_\alpha$  is based on an F distribution with p degrees of freedom in the numerator and  $n-p-1$  degrees of freedom in the denominator

$$\beta_1 = 0$$

There is no relation between  $x_1$  and  $y$



Coefficient of  $\beta_1$ , dependent variable

$$\beta_2 = 0$$

no relation between  $x_2$  and  $y$

$$MSR = \frac{SSR}{\text{degree of freedom}} = \frac{21.608}{2} = 10.804$$

e.g.: freedom

$$MSE = \frac{SSE}{n-P-1} = \frac{2.299448}{10-2-1} = 0.328491$$

$$F = \frac{MSR}{MSE} = \frac{10.804}{0.328491} = 32.88$$

t test for individual Significance

for any parameter  $\beta_i$ ,  $H_0: \beta_i = 0$  —  $\beta_i$

$$H_a: \beta_i \neq 0$$

Test Statistic  $t = \frac{b_i}{S_{b_i}}$

$$t = \frac{b_i - \beta_i}{S_{b_i}} \quad [\text{if } \beta_i = 0]$$

Rejection Rule:

P-Value approach Reject  $H_0$  if P-Value  $\leq \alpha$

Critical Value approach Reject  $H_0$  if  
 $t \leq t_{\alpha/2}$  or  $t \geq t_{\alpha/2}$

where  $t_{\alpha/2}$  is based on a t distribution  
with  $n-p-1$  degrees of freedom

t Test for individual Significance

$$b_1 = 0.061135 \quad S_{b_1} = 0.009888$$

$$b_2 = 0.9234 \quad S_{b_2} = 0.2211$$

$$t = 0.661135 / 0.009888 = 6.18$$

$$t = 0.9234 / 0.2211 = 4.18$$

Reject  $H_0$  if  $t \leq -t_{\alpha/2}$  or  $t \geq t_{\alpha/2}$

$$t_{\alpha/2} = 2.262$$

$$\alpha = 5\% = 0.05$$

$$6.18 > 2.262$$

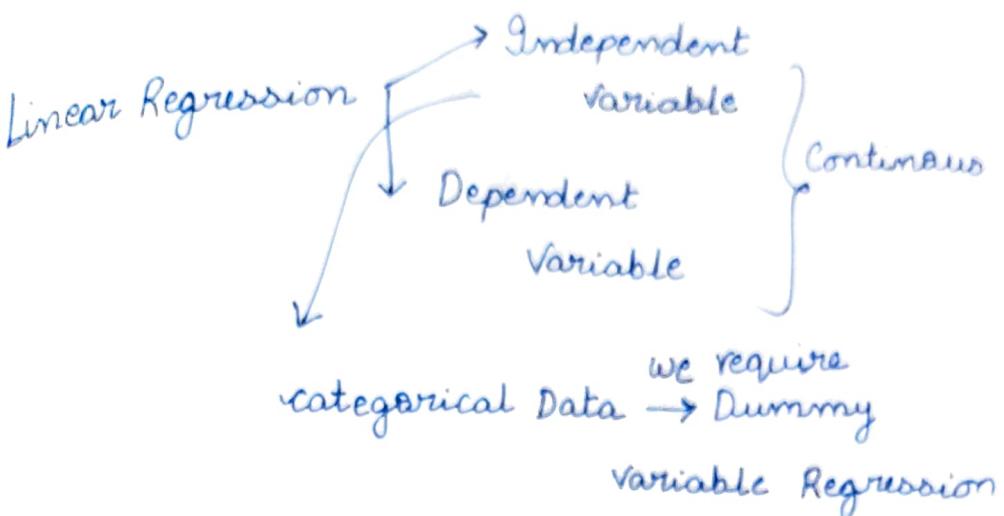
$$\frac{\alpha}{2} = 0.025$$

reject  $H_0$  in case of  
1<sup>st</sup> variable

$$4.176 > 2.262$$

In Case of 2<sup>nd</sup> variable Reject  $H_0$

## Logistic Regression:



There may be a possibility that dependent variable is categorical data. In such cases we need Logistic Regression.

If Dependent Variable → Categorical

The essential difference between these two is that Logistic Regression is used when the dependent Variable is binary in nature. In contrast, Linear regression is used when the dependent Variable is Continuous and nature of the regression line is nature.

Dependent Variable in Regression applications; need of Logistic Regression

Dependent Variable  
Gender: m or F  
Success or Failure

buy or not buy  
categorical  
In linear regression

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

Gender: male or  
female dependent  
variable

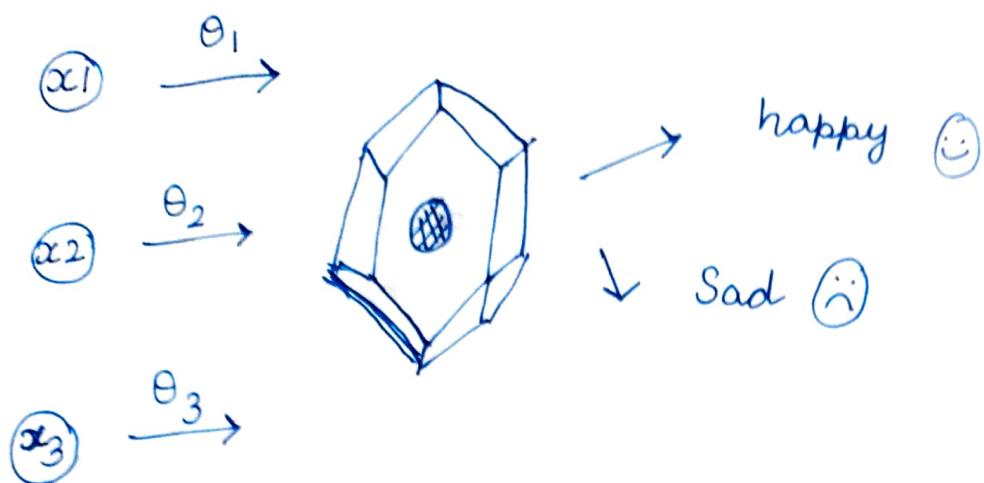
categorical  
data

independent

Quality of the product  
good or bad -

Dummy Variable  
Regression

## Need of Logistic Regression.



based on product;

dependent variable is categorical data

## Approval of Credit Card:

approves credit card or not

$y=1$  bank approved

$y=0$  bank rejects

we can estimate the probability of bank to approval of credit card

## Pay off Personal Loan:

will pay loan or not.

## Applying for job

dependent variable

Getting a job

not getting a job

Example:

Let us consider an application of Logistic regression involving a direct mail promotion being used by Simmons Stores.

Simmons owns and operates a national chain of women's apparel stores.

5000 copies of an expensive four-color sales catalog have been printed, and each catalog includes a coupon that provides a \$50 discount on purchases of \$200 or more.

The catalogs are expensive and Simmons would like to send them to only those customers who have the highest probability of using the coupon.

Variance:

independent  
management thinks that annual spending at Simmons Stores and whether a customer has a Simmons Credit Card are two variables that might be helpful in predicting whether a customer who receives the catalog will use the coupon.

Simmons conducted a pilot study using a random sample of 50 Simmons Credit Card customers and 50 other customers who do not have a Simmons Credit Card.

Simmons sent the catalog to each of the 100 customers selected.

At the end of a test period, Simmons noted whether the customer used the coupon or not?

Data (10 customer out of 100)

customer	Spending	Card	Coupon
1	2.291	1	0
2	3.215	1	0
3	2.135	1	0
4	3.924	0	0
5	2.528	1	0
6	2.473	0	1
7	2.384	0	0
8	7.076	0	0
9	1.182	1	1
10	3.345	0	0

categorical  
variable  
 $y$

Coupon

not using coupon

Explanation of Variables:

Amount of each customer spent last year at Simmons is shown in thousands of dollars and credit card info coded as 1 ✓ 0 no.

In the <sup>coupon</sup> column, a 1 is recorded if the sampled customer used the coupon and 0 if not.

Logistic Regression Equation:

- \* If the two values of the dependent variable  $y$  are coded as 0 or 1, the value of  $E(y)$  in equation given below provides the probability that  $y=1$  given a particular set of values for the independent variables  $x_1, x_2, \dots, x_p$ .

$$E(y) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}}$$

$y$  is Coupon

$P(y=1/x_1, x_2, \dots, x_p)$

↳ independent Variables.

$$\frac{e^\infty}{1+e^\infty} = \frac{\infty}{\infty} \quad (\because \text{indeterminant value})$$

$$\downarrow \frac{e^\infty}{e^\infty (1+e^{-\infty})} = \frac{1}{1+0} = 1 \quad \downarrow \text{max value}$$

$$\frac{e^{-\infty}}{1+e^{-\infty}} = \frac{e^{-\infty}}{e^{-\infty}(1+e^{-\infty})} = \frac{1}{1+e^0} = 0$$

$E(y)$  maximum value of  $E(y)$  is 1  
minimum value of  $E(y)$  is 0

Because of the interpretation  
linear  
Error

↳ normal distribution

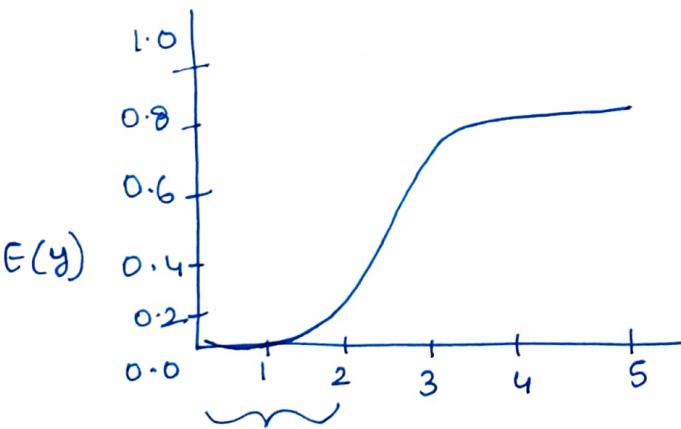
$y$  coupon      } dependent variable  
                Binomial distribution

If dependent Variable of binomial distribution,  
we must apply logistic regression and  
the error term follows Binomial ; Linear can't  
be used

Because of the interpretation of  $E(y)$  as a  
probability, the logistic regression is

$$E(y) = P(y=1 | x_1, x_2, x_3, \dots, x_p)$$

Logistic regression equation for  $\beta_0$  and  $\beta_i$ ,



$x$  is  $\uparrow$ ,  $E(y)$  is almost 1

$x$  is below 1  $\downarrow$   $E(y) \approx 0$

## Estimating the Logistic Regression Equation.

In simple linear and multiple regression the least squares method is used to compute  $b_0, b_1, b_2, \dots, b_p$  as estimates of the model parameters  $(0, 1, \dots, p)$ .

The nonlinear form of the logistic regression equation makes the method of computing estimates more complex.

We use maximum likelihood estimation in case of logistic regression.

$$\hat{y} = \text{estimate of } P(y=1|x_1, x_2, \dots, x_p) = \frac{e^{b_0 + b_1 x_1 + \dots + b_p x_p}}{1 + e^{b_0 + b_1 x_1 + \dots + b_p x_p}}$$

Here  $y$  is dependent variable  
(coupon)

$y$  has provides an estimate of the probability that  $y=1$ , given a particular set of values for the independent variables.

$$\beta_0, \beta_1, \dots, \beta_p \rightarrow \text{Population Parameters}$$
$$b_0, b_1, \dots, b_p \rightarrow \text{Sample }$$

estimate population parameters with help of sample parameters.

## Logistic Regression Analysis:

Eg) Vote for a Democrat or Republican

Social Sciences

Marketing

Health Care

Managerial use

dependent  
variable

$$P(Y=1 | X_1=2, X_2=0) = 0.1880$$

↓      ✓ independent  
coupon

prob. that customer

$x_1 = 2$   
↓  
2000\$

using coupon

$x_2 = 0$   
↓  
not using

$x_1$  = annual Spending

Simmons' Credit Card

$x_2$  = Simmons' Credit Card

$$\hat{y} = \frac{e^{-2.14637 + 0.341643(2) + 1.09873(0)}}{1 + e^{-2.14637 + 0.34164(2) + 1.09873(0)}}$$

$$= \frac{e^{-1.4631}}{1 + e^{-1.4631}}$$

$$= \frac{0.2315}{1.2315}$$

$$= 0.1880 \checkmark$$

$$P(y=1 | x_1=2 | x_2=1) = 0.4099$$

$$= \frac{e^{-0.3644}}{1+e^{-0.3644}} = \frac{0.6946}{1.6946} = 0.4099$$

person have more credit card is having more probability (using coupon).

It appears that the probability of using the coupon is much higher for customers with a Simmons Credit card.

#### Annual Spending

	\$ 1000	\$ 2000	\$ 3000	\$ 4000	\$ 5000	\$ 6000	\$ 7000
credit card	yes	0.3305	0.4099	0.4943	0.5791	0.6594	0.7315
	no	0.1413	0.1880	0.2457	0.3144	0.3922	0.4759
							0.5616

#### Testing for Significance:

$$H_0 : \beta_1 = \beta_2 = 0$$

$H_a$ : one or both of the parameters is not equal to 0.

G Statistics: The test for overall Significance.

Degrees of df freedom : No. of Independent Variables

If the null hypothesis is true, the Sampling distribution of G follows a chiSquare dist. with df to equal no. of independent Variables in the model.

# Interpreting Odds ratios in Logistic Regression

Odds Ratio:

probability to odds to log of odds

$p^*$  of Success of Some event

$$p = 0.8$$

$$\text{prob. of failure} = 1 - 0.8 = 0.2$$

$$\text{The odds of Success} = \frac{P}{1-P} = \frac{0.8}{0.2} = \frac{4}{1} = 4$$

$$P = 0.5 \rightarrow 50-50$$

$$q = 0.5 \quad \frac{P}{1-P} = \frac{0.5}{0.5} = \frac{1}{1}$$

Transformation is monotonic



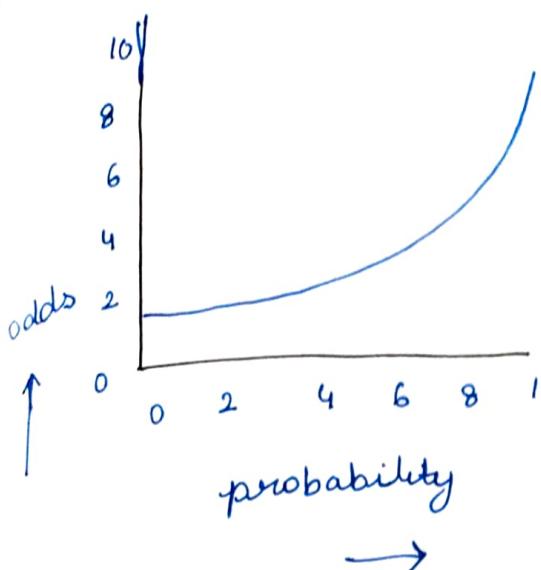
from prob. to odds

→ Prob	odds
↑	↑
e.g. .001	.001001

{prob. ranges from 0 to 1

$$p = 0 \quad \text{odds} = \frac{p}{1-p} = \frac{0}{1-0} = 0 \quad \text{odds ranges} \quad 0 \text{ to } \infty \}$$

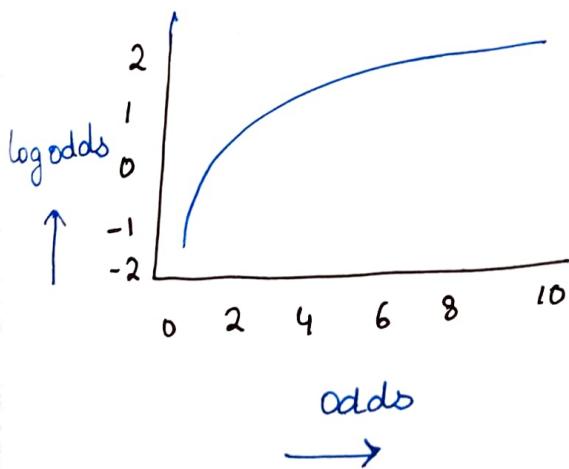
$$p = 1 \quad \text{odds} = \frac{1}{1-1} = \frac{1}{0} = \infty$$



Log Transformation is monotonic

from Odd to log odd

$$\begin{array}{cccc}
 p & \text{odds} & \longleftrightarrow & \log \text{odd} \\
 .001 & .001001 & -6.906755 \\
 \uparrow & & & \uparrow
 \end{array}$$



Logistic regression model allows us to establish the relation b/w a binary outcome Variable (dependent) and Group of predicted Variables (Independent)

(Logistic)

$$\text{logit}(P) = \log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

↓  
from [MLE]

$$P = p(y=1)$$

$x_1, x_2, \dots, x_k$

predictor Variables

Independent Variable

Exponentiate and take the multiplicative inverse  
of both sides

$$\frac{1-P}{P} = \frac{1}{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)} \quad \checkmark$$

$$\left[ \because \log_e\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \right]$$

$\overbrace{\phantom{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}}^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}$

$$\frac{1}{P} - 1 = \frac{1}{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}$$

$$\frac{1}{P} = 1 + \frac{1}{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}$$

↓

use above eq. we have

partial out the fraction on LHS of eq and  
add 1 to both Sides.

change 1 to a common denominator

$$\frac{1}{P} = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k) + 1}{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}$$

finally, take the multiplicative inverse again  
to obtain the formula for the prob.  $P(Y=1)$

$$P = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}.$$

$$p = \underbrace{0}_{\text{to}} \underbrace{1}_{\cdot}$$

Interpreting the Logistic Regression Equation.

P is 0 to 1 Prob(Success)

$$P = 0.8 \quad q = 1 - P = 1 - 0.8 = 0.2$$

odd  $\rightarrow 0$  to  $\infty$

$$\text{odd (Success)} = \frac{P}{1-P} = \frac{P}{q} = \frac{0.8}{0.2} = 4 = \frac{4}{1}$$

The odds of Success are 4 to 1

$$\text{Odds (failure)} = \frac{q}{P} = \frac{0.2}{0.8} = 0.25 = \frac{1}{4}$$

$$\text{i.e. } \frac{1}{4} = 0.25 \text{ and } \frac{1}{0.25} = 4$$

→ dependent Variable (Coupon)

$$\text{Odds} = \frac{P(y=1 | x_1, x_2, \dots, x_p)}{P(y=0 | x_1, x_2, \dots, x_p)}$$

$$= \frac{P(y=1 | x_1, x_2, \dots, x_p)}{1 - P(y=1 | x_1, x_2, \dots, x_p)} = \frac{P}{1-P} = \frac{\text{Success}}{\text{not Success}}$$

Odd ratio :

$$\text{Odds Ratio} = \frac{\text{Odds}_1}{\text{Odds}_0}$$

1<sup>st</sup> Level  
0<sup>th</sup> Level  
↓  
not having Credit card  
having the credit card.

measures the impact on the odds of a one-unit increase in only one of the independent variables

Interpretation:

\* for eg: Suppose we want to compare the Odds of using the coupon for customers who Spend \$2000 annually and have a Simmons Credit card ( $x_1=2$  and  $x_2=1$ ) to the Odds of \$2000 annually and do not have a Simmons Credit card ( $x_1=2$  and  $x_2=0$ )

compare  $\begin{cases} x_1 = 2 & x_2 = 1 \\ x_1 = 2 & x_2 = 0 \end{cases}$  } we are interested in  
 odds of using the coupon interpreting the effect  
 of a one-unit increase  
 in the independent  
 variable  $x_2$

$$\text{odd } S_1 = \frac{p(y=1 | x_1=2, x_2=1)}{1 - p(y=1 | x_1=2, x_2=1)}$$

$$\text{estimate of odd } S_1 = \frac{0.4099}{1 - 0.4099} = 0.6946$$

$$\text{odd } S_0 = \frac{p(y=1 | x_1=2, x_2=0)}{1 - p(y=1 | x_1=2, x_2=0)}$$

$$\text{estimate of odd } S_0 = \frac{0.1880}{1 - 0.1880} = 0.2315$$

$$\text{Estimated odd ratios} = \frac{0.6946}{0.2315} = 3.00$$

The estimated odds in favour of using the coupon for customers who spent \$2000 last year and have a Simmons Credit Card are 3 times greater than the estimated odds in favour of using the coupon for customers who spent \$2000 last year and do not have a Simmons Credit card.

The Odds ratio for each independent variable is computed while holding all the other independent variables constant.

$x_1$  = Spending

$x_2$  = Simmons' credit card.

But it does not matter what constant values are used for other independent variables.

For instance, if we (considered) computed the Odds ratio for Simmons' credit card

Variable ( $x_2$ ) using \$3000, instead of \$2000, as the value for annual Spending variable ( $x_1$ ), we would still obtain the same value for the estimated Odds ratio (3.00)

$$\text{odd } S_1 = \frac{0.4943}{1 - 0.4943} = \frac{0.4943}{0.5057} = 0.97745699$$

$$\text{odd } S_0 = \frac{0.2457}{1 - 0.2457} = \frac{0.2457}{0.7543} = 0.325732$$

$$\text{Estimated odd ratio} = \frac{0.97745699}{0.325732} = \underline{\underline{3}}$$

Relationship b/w odd ratios and Coefficient of independent variables

	Coef.	Std Err	Z	P >  Z
Const	-2.1464	0.5772	-3.7183	0.00022
$x_2$ (card)	1.0987	0.4447	2.4707	0.0135
$x_1$ (spending)	0.3416	0.1287	2.6551	0.0079

$$\begin{aligned}\text{Estimated Odds ratio} &= e^{b_1} = e^{0.341643} \\ &= 1.41 \\ &= e^{b_2} = e^{1.09873} = 3.00 \checkmark \\ \text{Odd ratio} &= e^{\beta_i}\end{aligned}$$

Effect of a change of more than one unit in odd ratio.

The odd ratios for an independent Variable represents the change in the Odds for a one unit change in the independent Variable holding all the other independent Variables constant.

Suppose that we want to consider the effect of a change of more than one unit, say  $c$  units.

For instance, suppose in the Simmons ex. that we want to compare the odds of using the coupon for customers who spend \$ 5000 annually ( $x_1=5$ ) to the odds of using the coupon for customers who spend \$ 2000 annually ( $x_1=2$ )

In this case  $c=5-2=3$  and corresponding estimated odds ratio is very useful.

$$e^{cb_1} = e^{3(0.341643)} = e^{1.0249} = 2.79$$

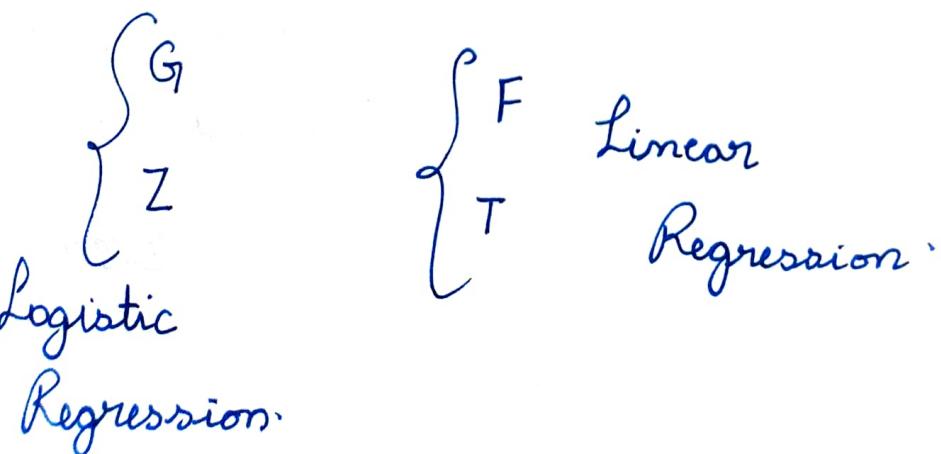
This result indicates that the estimated odds of using the coupon for customers who spend \$5000 annually is 2.79 times greater than the estimated odds of using the coupon for customers who spend \$2000 annually.

In other words, the estimated odds ratio for an increase of \$3000 in annual spending is 2.79

### G vs Z

Because of the unique relationship b/w the estimated coefficients in the model and the corresponding odd ratios, the overall test for Significance based upon the G Statistic is also a test of overall Significance for the Odds ratio.

In addition, the Z test for the individual Significance of a model parameter also provides a Statistical test of Significance for the corresponding odds ratio.



# Linear Regression Model vs Logistic Regression Model

Linear Regression model:

$$y_1 = x_1 + x_2 + \dots + x_n$$

Dependent Variable

independent Variable

where,

$y_1$  is Continuous data

Independent Variables = non metric and  
metric

↓  
discrete

↳ continuous

Logistic Regression model

$$y_1 = x_1 + x_2 + \dots + x_n$$

Independent

dependent

where,

$y_1$  = Binary nonmetric

Independent Variables = non metric and  
metric

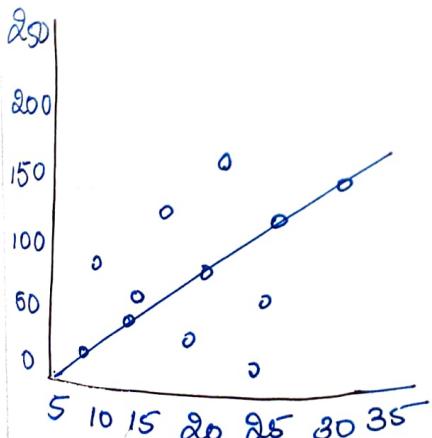
↗ Discrete

↓  
Discrete

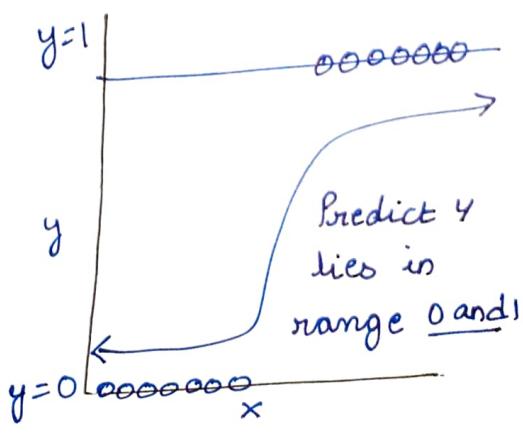
↳ continuous

## Graphical Representation:

Linear  
Regression



Logistic  
Regression



correspondence of primary Elements of model fit

linear

logistic

Regression.

Total Sum of Squares  
SST

Error Sum of Squares  
SSE

F test of model fit

$$F = \frac{mSR}{mSE} = \frac{SSR/K}{SSE/(n-K-1)}$$

Coefficient of determination ( $R^2$ )

Regression Sum of Squares.

-2LL of base model

loglikelihood

-2LL of proposed model

G Test

chi-Square test

-2LL difference

Pseudo  $R^2$  measure

Difference of -2LL for base and proposed models

Base model:

There is no independent variable

## Determination of Coefficients

### Linear Regression

$R^2$

$$R^2 = \frac{SSR}{SST}$$

where

$SSR$  = Sum of Squares due to Regression

$SST$  = total Sum of Squares

### Log likelihood:

\* Measures used in Logistic Regression.

\* Lack of predictive fact

\* SSE in linear Regression

\* If you (have) less loglikelihood then it is a good model

\* To Compare models

### Logistic Regression:

$$R^2 \text{ Logit} = \frac{-2LL_{\text{null}} - (-2LL_{\text{model}})}{-2LL_{\text{null}}}$$

where  $LL$  = loglikelihood

$-2LL = -2LL$  of base model

$-2LL_{\text{model}} = -2LL$  of proposed model

Testing for Overall Significance:

Linear:

F-test of model fit

$$F = \frac{MSR}{MSE}$$

Logistic

G-test

$$G = -2 \ln \left[ \frac{\text{likelihood without a Variable}}{\text{likelihood with Variable}} \right]$$

Significance of Each independent Variable

t-test

$$\frac{\hat{\beta}_1}{\text{Std Error}(\hat{\beta}_1)}$$

Wald test

Linear Regression

Logistic

Error Sum of Squares

-2LL of proposed model

$$SSR = \sum (y_i - \bar{y}_i)^2$$

Diff. b/w log likelihood

$$SST - SSE$$

$$\Delta LL_{\text{null}} - (\Delta LL_{\text{model}})$$

normally

distributed

Binomially

linear Reg. assumes  
that residuals are  
approx. equal for all  
predicted variable  
values.

logistic Reg. does not  
need residuals to  
be equal for each  
level of the  
predicted dep. Variable  
values

based on Least Square  
estimation

Reg. coefficients should be  
chosen in such a  
way that it minimizes  
the sum of the squared  
distances of each  
observed response to  
its fitted value.

maximum  
likelihood estimation

maximizes the  
probability of  $y$   
given  $x$  (likelihood)  
with MLE, the  
computer uses  
diff. "iterations" in  
which it tries  
different solutions  
until it gets the  
max. likelihood  
estimates