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 fix 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 infer 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 / Explaradatory Variables or input Variable indicated by x.

Eg: 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 Non Linear

→ Linear: Relationship is linear between dependent and independent Variables.

→ Non Linear: 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

} Regression.

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 corelation and Coefficient of determination.

5. useful in Statistical Estimator in demand Curves, Supply Curves, function, Cost function, Consumption function. etc

Comparision between Corelation and Regression.

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 (Letter?) In Corelation (analysis) x and y are random Variables whereas in Regression x is R.V and y is fixed measure.

corelation → 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:

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

β_0, β_1 are parameters

ϵ is Random variable called error term.

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

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

$E(y)$ is Expectation of y

In calculating β_1 we minimized the errors predicting $E(y)$ i.e mean value of y or $E(Y)$

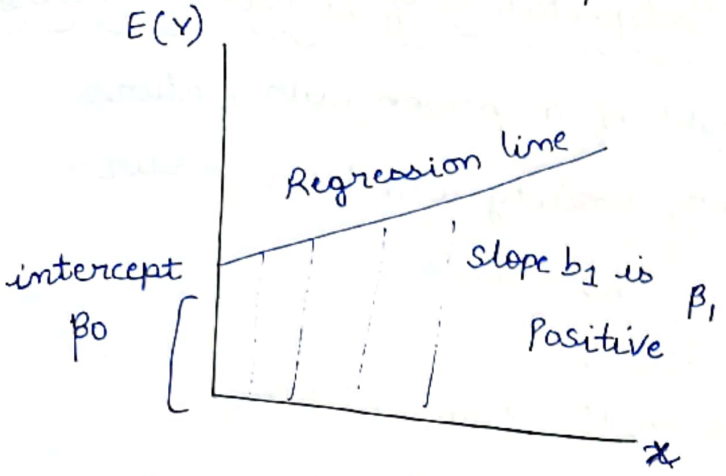
Graph is Straight line

β_0 is y intersect cept of Regression line

β_1 is Slope of the Regression line.

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

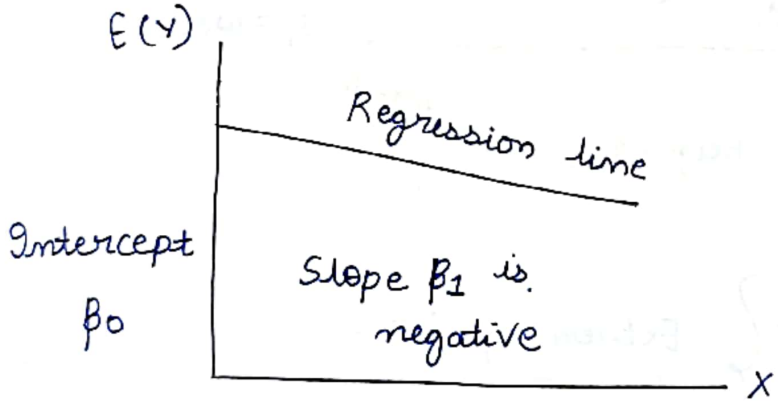
Positive Linear Relationship



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

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

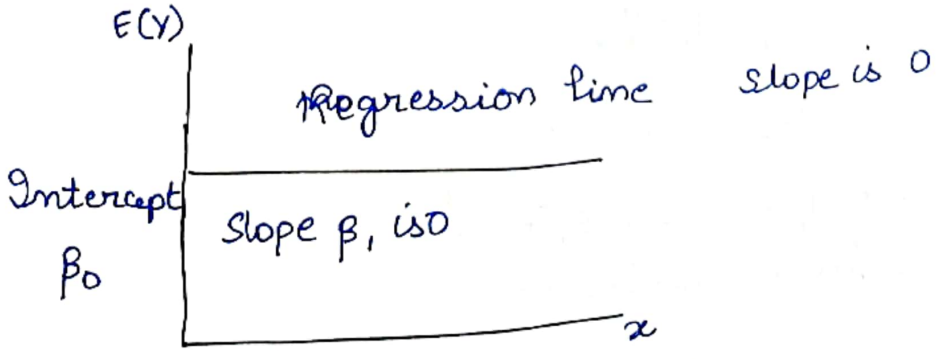
Negative Linear Relationship



$x \uparrow$ y is $\downarrow$

$\because$ Slope is negative

No Relationship.



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

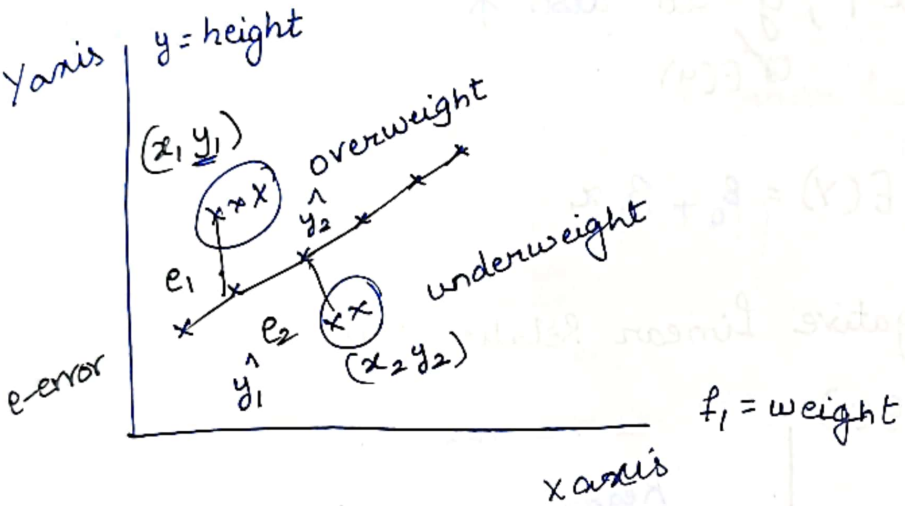
Geometrical Intepretation of Linear Regression.

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

Height is a Real number

$h \in R$

Eg:- 165.6 cm, 153.2 cm, 175.5 cm



weight ↑ height ↑

Exceptions:

overweight, underweight } Extreme points.

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

Error = Actual Value − Predicted Value

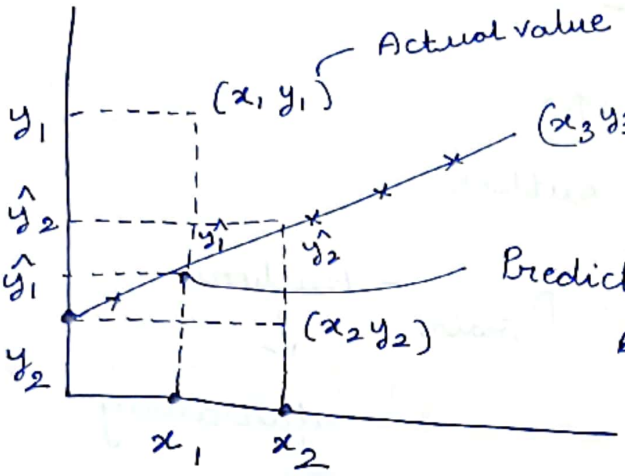
$e_1 = y_1 - \hat{y}_2$

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

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

which ever line is minimized is least Square method [is Principle].

↓
Sum of errors



Error = A.V − P.V

$= y_1 - \hat{y}_1$

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

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

↳ error for x_2

above } regression line → +ve

below } → −ve

Some point on Regression is 0

if we Sum of errors $= e_1^2 + e_2^2 + e_3^2 + \cdots + e_n^2$

→ +ve

$e_1 = 10$

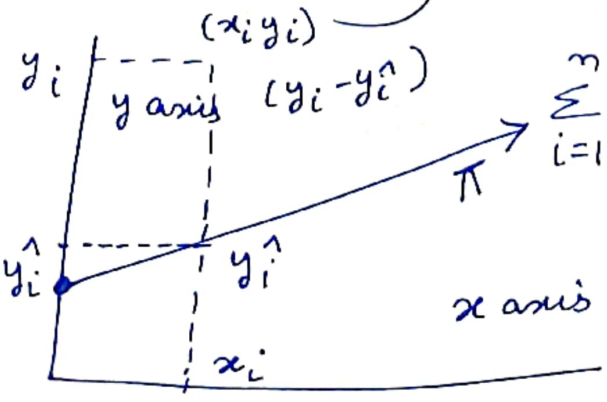
$e_2 = -10$

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

$= 0$

but error exists above line

If not Sum of errors errors may be nullified.



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

Impact of outliers.

For trained data (we find)

Eg: $y = mx + c$

$y_i - \hat{y_i}$ ↑

remove as outliers

$D_{\text{Training Data set}} = D_{\text{Train}} - \text{Outliers}$ ↓ far away

Repeat this procedure.

This is Random Sampling process.

Least Squares Method

- Least Square Criterion

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

$\hat{y}_i$ → predicted value

y_i → Actual value

where:

y_i = Observed value of the dependent variable for the ith observation.

$\hat{y}_i$ = estimated value of dependent variable for ith observation.

In Least Squares method,

(i) we first find errors

(ii) Square of the errors.

(iii) Sum of the Squared errors.

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

$$(y_1-\hat{y}_1)^2 + (y_2-\hat{y}_2)^2 + \cdots (y_n-\hat{y}_n)^2$$

Squared Error (SE) = ~~fy~~

$$y = mx + b$$

$$\Rightarrow (y_1 - (mx_1+b))^2 + (y_2-(mx_2+b))^2 + \cdots\cdots$$

$$(a-b)^2 \qquad\qquad (y_n-(mx_n+b))^2$$

$\hat{y}_1 = mx + b$

$\hat{y}_1 = mx_1 + b$

error = AV − PV

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

$$(y_1-(mx_1+b))^2 \Rightarrow y_1^2 + \underbrace{(mx_1+b)^2}_{(a+b)^2} - 2y_1(mx_1+b)$$

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

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

$$SE = (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 + \ldots$$

$$(y_n - (mx_n + b))^2$$

$$= y_1^2 - 2y_1(mx_1 + b) + (mx_1 + b)^2 +$$

$$y_2^2 - 2y_2(mx_2 + b) + (mx_2 + b)^2 + \ldots\ldots$$

$$+ y_n^2 - 2y_n(mx_n + b) + (mx_n + b)^2$$

$$= y_1^2 - 2x_1y_1m - 2y_1b + m^2x_1^2 + 2mx_1b + b^2 +$$

$$y_2^2 - 2x_2y_2m - 2y_2b + m^2x_2^2 + 2mx_2b + b^2 +$$

$$\ldots\ldots + y_n^2 - 2x_ny_nm - 2y_nb + m^2x_n^2 +$$

$$2mx_nb + b^2$$

Grouping terms

$$= (y_1^2 + y_2^2 + \cdots + y_n^2) - 2m(x_1y_1 + x_2y_2 + \cdots + x_ny_n)$$

$$- 2b(y_1 + y_2 + \cdots + y_n) + m^2(x_1^2 + x_2^2 + \cdots + x_n^2)$$

$$+ 2mb(x_1 + x_2 + \cdots + x_n) + (b^2 + b^2 + \cdots + b^2)$$

$$= n\overline{y^2} - 2mn\,\overline{xy} - 2bn\,\overline{y} + m^2n\,\overline{x^2} +$$

$$2mbn\,\overline{x} + nb^2$$

Proof:

$$\overline{y}_2 = \frac{y_1^2 + y_2^2 + \cdots + y_n^2}{n}$$

$$ny_2 = y_1^2 + y_2^2 + \cdots + y_n^2 \quad \text{Sub}$$

$$\overline{xy} = \frac{x_1 y_1 + x_2 y_2 + \cdots + x_n y_n}{n}$$

$$n\overline{xy} = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$$

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

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

$$\overline{x^2} = \frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}$$

$$x_1^2 + x_2^2 + \cdots + x_n^2 = nx^2$$

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

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

Standard error $= n\overline{y^2} - 2mn\bar{x}\bar{y} - 2bn\bar{y} +$

[SE] $m^2 n\overline{x^2} + 2mbn\bar{x} + nb^2$

$y = mx + b$

m, b are parameters

find m, b

differentiate partially

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

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

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

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

[2n common]

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

$$m\overline{x^2} + b\bar{x} = \overline{xy} \quad \{\div \text{ with } \bar{x}\}$$

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

$mx + b = y$

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

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

differentiating partially w.r.t b [constant 0]

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

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

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

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

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

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

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

we get max. value

∴ line must pass through $\left(\frac{\bar{x}^2}{\bar{x}}, \frac{\bar{x}\bar{y}}{\bar{x}}\right)$ and $(\bar{x}, \bar{y})$

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

x_1 y_1 x_2 y_2

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

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

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

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

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

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

$$\text{Var}(x) = E(x-\bar{x})^2$$

$$= E(x^2 - 2x\bar{x} + \bar{x}^2)$$

$$= E(x^2) - 2(\bar{x})^2 + (\bar{x})^2$$

$$= \overline{x^2} - (\bar{x})^2$$

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

$\therefore$ 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

$$\bar{y} = b_0 + b_1 \bar{x}$$ → Parameters.

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

$$b_0 = \bar{y} - b_1 \bar{x}$$

Multiple Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_x x_p + \epsilon$$

If more than 1 independent variable

Multiple Regression Equation

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_x x_p$$

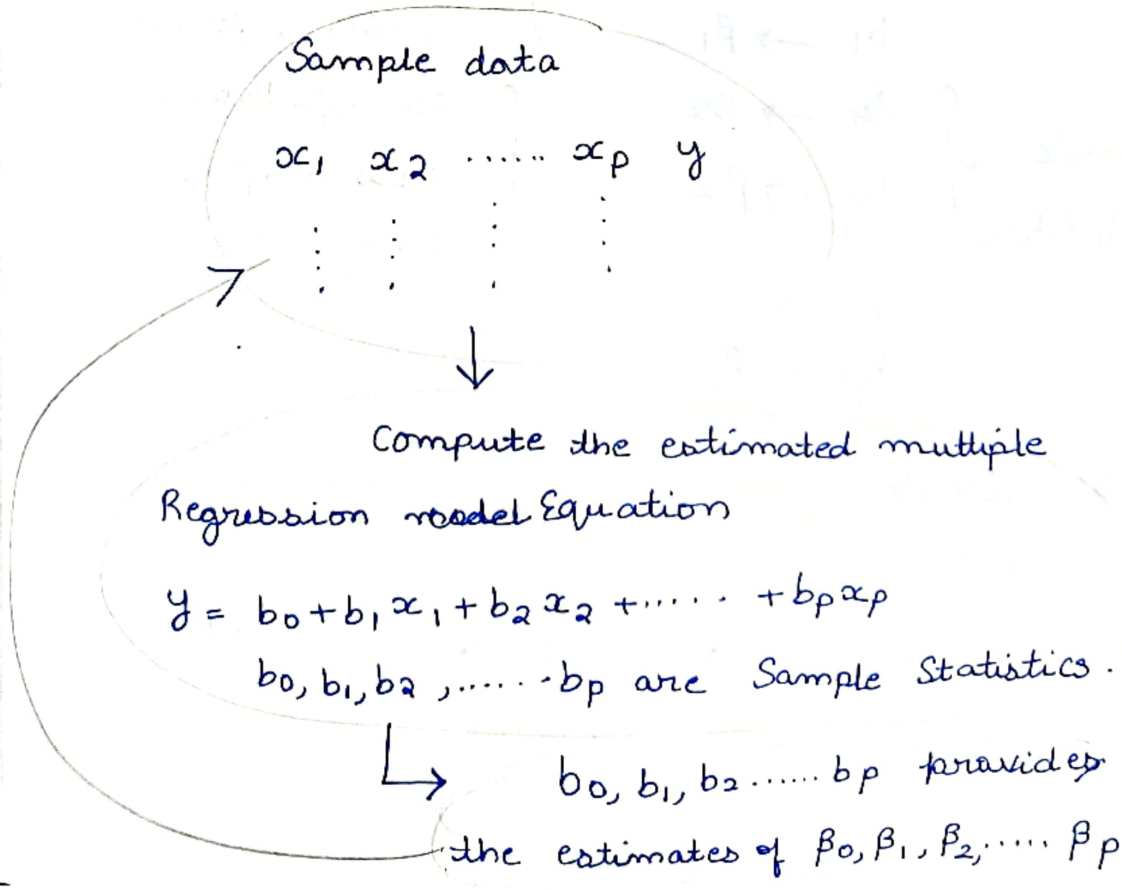
$$[E(\epsilon) = 0]$$

$\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ are unknown parameters

The Estimation process for Multiple Regression

To find unknown parameters from the population we collect Sample data

y is dependent variable



find How can we decide $\beta_0, \beta_1, \beta_2, \ldots \beta_n$ 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 β_0 and β_1

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

$b_0 \rightarrow \beta_0$

$b_1 \rightarrow \beta_1$

~~$b_1 \rightarrow \beta_2$~~

$b_2 \rightarrow \beta_2$

$\vdots$

$b_p \rightarrow \beta_p$

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

Explo 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	no. of deliveries y = Travel Time (hours)
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 → Independent Variable.

dependent

↳ delivers → Independent.

x1 miles travelled

Dependent Variable: travel-time

Independent Variables:

x1 = miles travelled

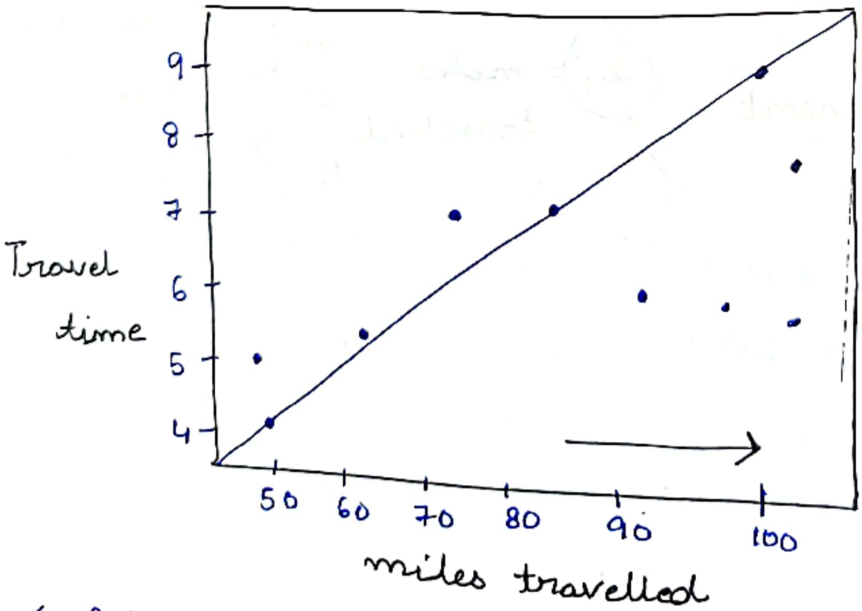
n_of_deliveries.

Let's understand the relation b/w x1 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

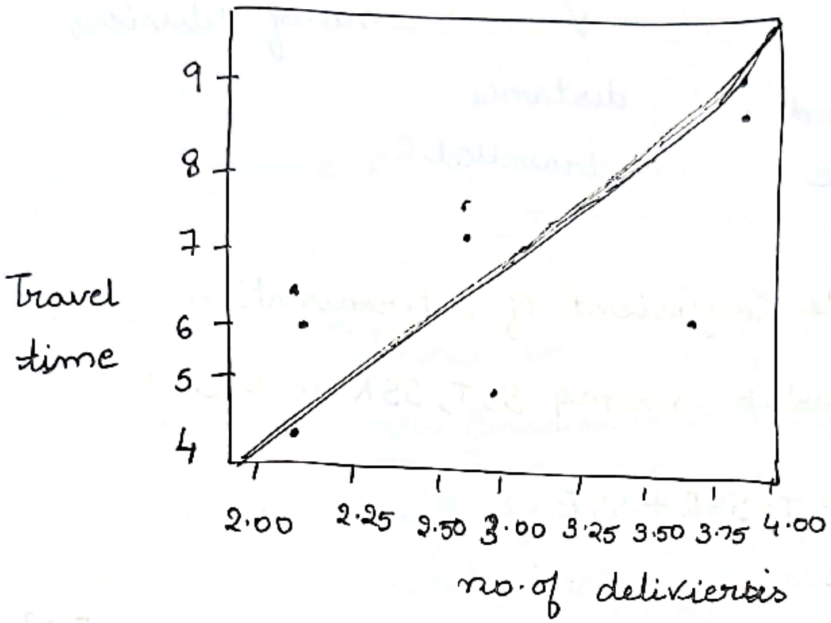


✓ Relation between one independent and one dependent Variable.

Travel time depends on miles travel

now Consider no. of deliviers

Simple linear regression with number of deliveries.

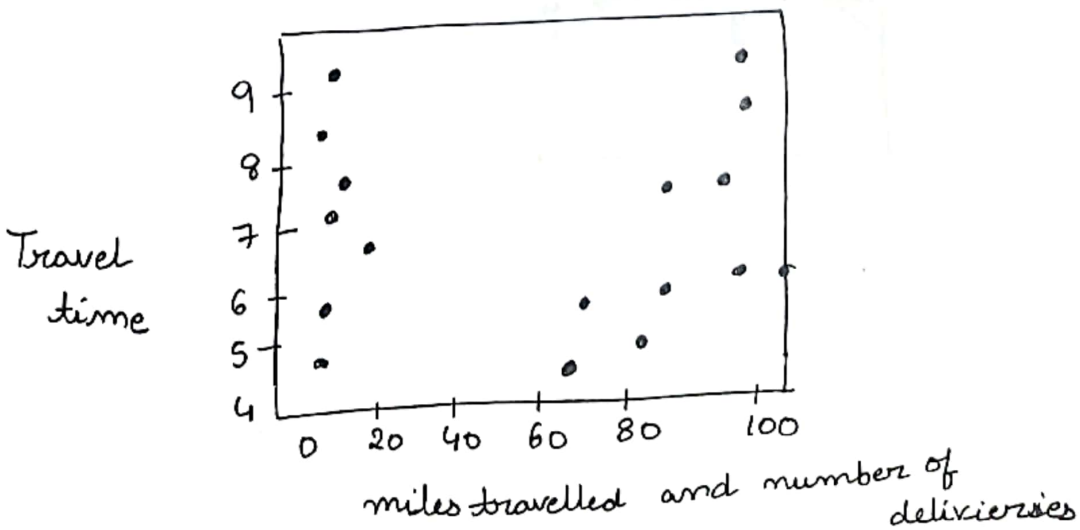


linear regression

There is a +ve correlation

? why correlation → If there is no Co relation there is no relation between dependent and independent variable So regression model is not used in such cases.

Multiple regression.



Multiple regression:

Eg: $\hat{y} = -.869 + .0611 x_1 + .923 x_2$

- $\hat{y}$: Travelled time
- x_1: distance travelled
- x_2: no. of deliveries

multiple coefficient of determination

Relationship among SST, SSR and SSE

$$SST = SSR + SSE$$

where

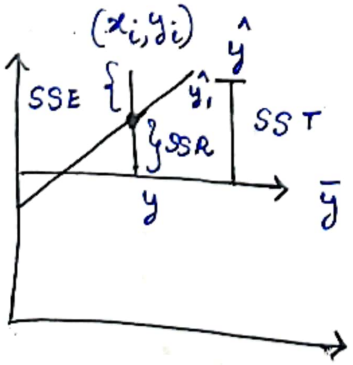
SST = total Sum of Squares $= \Sigma (y_i - \bar{y})^2$

SSR = Sum of Squares due to regression

$= \Sigma (\hat{y}_i - \bar{y})^2$

SSE = Sum of Squares due to error

$= \Sigma (y_i - \hat{y}_i)^2$



Adjusted Multiple Coefficient of Determination:

$$SST = SSR + SSE$$

$$SSR = SST - SSE$$

↳ decreased if

↓

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 → Independent Variables

$y \to x_1$, R^2 ↑ ↙	Adj R^2 ↑
$y \to x_1 x_2$ ↑	↑
$y \to x_1 x_2 x_3$ (x_3 is noise data)	

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

R^2 ↑ $y \rightarrow x_1 x_2 x_3$ ↓ Adj R^2

If R^2 and Adj R^2 are same we need not add variables

If R^2 is 0.9, Adj R^2 = 0.3 → we can add more values to fill 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 ~~above~~ 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
	incr.	decreasing

Adj

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 Signi-ficant 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 Seperate 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:

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

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

$H_0 = \beta_1 = \beta_2 = \cdots = \beta_p = 0$

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

Test Statistic:

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

Relation Rejection rule:

p Value approach — Reject H_0 if p-Value $\leq \alpha$

Critical Value approach — Reject H_0 if $F \geq F_\alpha$

where

F_α 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 β_1 ↓ dependent Variable

$\beta_2 = 0$

no relation between x_2 and y

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

Eg:

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

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

t test for individual Significance.

for any parameter β_i $H_0: \beta_i = 0$ — β_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

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

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

$t = 0.661135 / 0.009888 = 6.18$

$t = 0.9234 / 0.2211 = 4.18$

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

$t_{\alpha/2} = 2.262$

$\alpha = 5\% = 0.05$

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

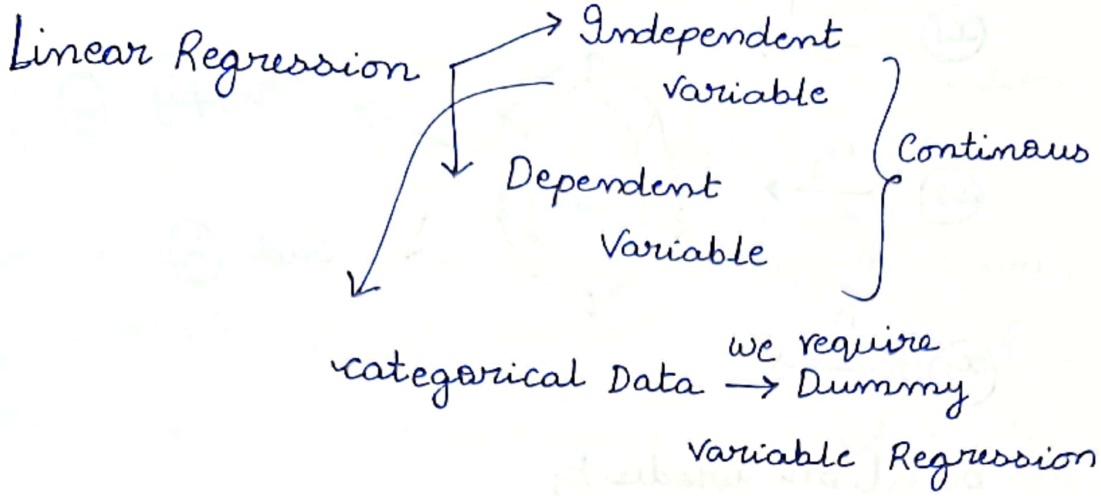
$6.18 > 2.262$

reject H_0 in Case of 1st variable

$4.176 > 2.262$

In Case of 2nd 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 Continous and nature of the regression line is (nature) linear

Dependent Variable in Regression applications; need of Logistic Regression

Dependent Variable
Gender : m or F
Success or Failure

buy or not buy

In linear regression

categorical

$y = a + b_1 x_1 + b_2 x_2$

0, 1

independent

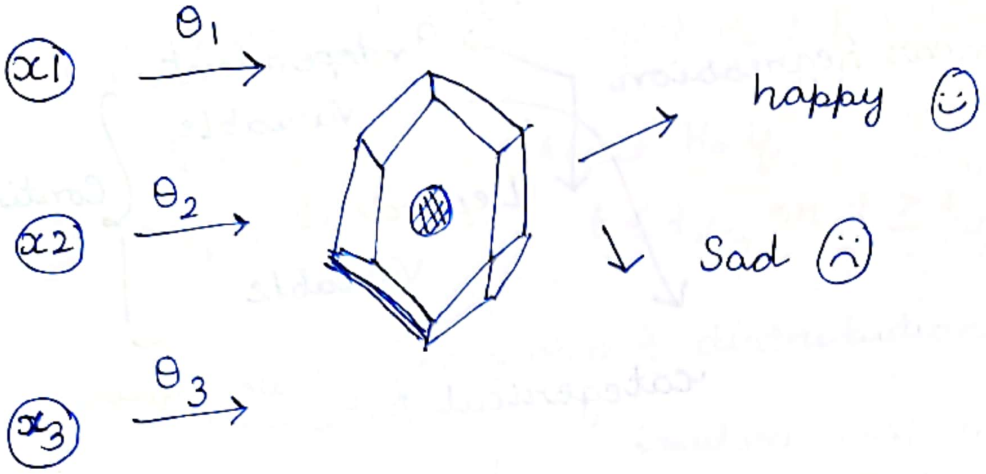
Gender: male or female dependent variable

categorical data

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 estime the probability of bank to approval of credit card

- Payoff 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:

Management thinks that annual Spending (independent) at Simmons Stores and whether a Customer has a Simmons Credit Card are two Variables that might be helpful in predicting wheather 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 (x_1)	Card (x_2)	Coupon (y)
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

x_1, x_2 → independent variables

y → categorical variable / dependent variable

Coupon: 0 → 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 coupon 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 $x_1, x_2, \cdots, x_p$.

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

y is Coupon

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

↳ independent Variables.

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

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

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

E(y) maximum value of E(y) is 1

minimum value of E(y) is 0

Because of the interpolation

Linear
Error

↳ normal distribution

dependent variable.

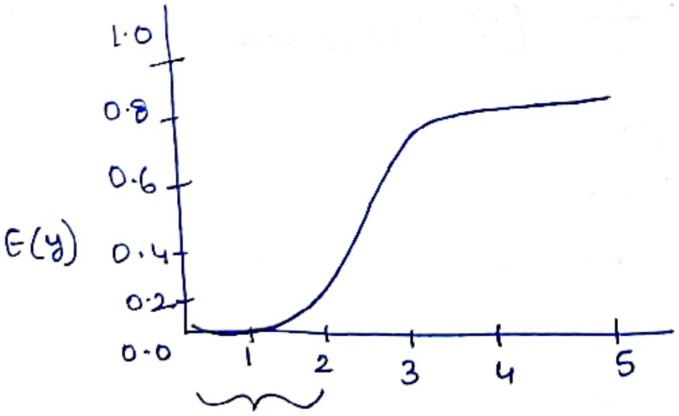
y coupon { 1, 0 } Binomial distribution

If dependent variable of is binomial distribution we must apply logistic regression and the error term is follows Binomial ; Linear cant be used

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

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

Logistic regression equation for β_0 and β_1



3

x is ↑ , $E(y)$ is almost 1

x is below 1 $\xrightarrow{E(y)}$ 0

Estimating the Logistic Regression Equation.

In Simple linear and multiple regression the least Squares method is used to Compute $b_0, b_1, b_2, \cdots, b_p$ as estimates of the model parameters $(0, 1, \cdots, p)$.

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.

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

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.

$\beta_0, \beta_1, \cdots \beta_p \rightarrow$ Population } Parameters
$b_0, b_1, \cdots, b_p \rightarrow$ 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

$P(Y=1 \mid x_1=2, x_2=0) = 0.1880$

coupon — independent

prob. that customer using coupen

x_1 = annual Spending

x_2 = Simmon's Credit Card

$x_1 = 2$ → 2000$

$x_2 = 0$ → not using Simmon's Credit card

$$\hat{y} = \frac{e^{-2.14637 + 0.34163(2) + 1.09873(0)}}{1 + e^{-2.14637 + 0.34164(2) + 1.09873(0)}}$$

(x_1 = 2nd term, x_2 = 3rd term)

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

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

$$= 0.1880$$

$$P(y=1 \mid x_1=2 \mid x_2=1) = 0.4099$$

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

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	0.7931
	no	0.1413	0.1880	0.2457	0.3144	0.3922	0.4759	0.5610

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 freedom (df): 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

$p = 0.8$

prob. of failure $= 1 - 0.8 = 0.2$

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

$p = 0.5 \rightarrow 50\text{-}50$

$q = 0.5$

$$\frac{p}{1-p} = \frac{0.5}{0.5} = \frac{1}{1}$$

Transformation is monotonic

↓

from prob. to odds

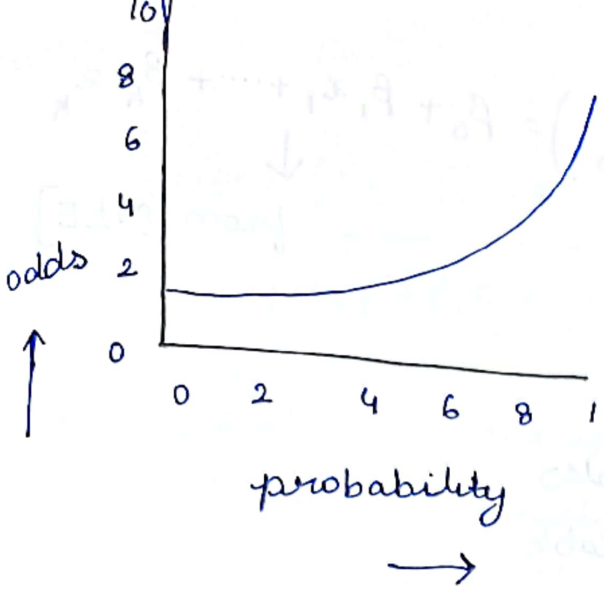
→ Prob ↑ odds ↑

Eg: .001 .001001

{prob. ranges from 0 to 1

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

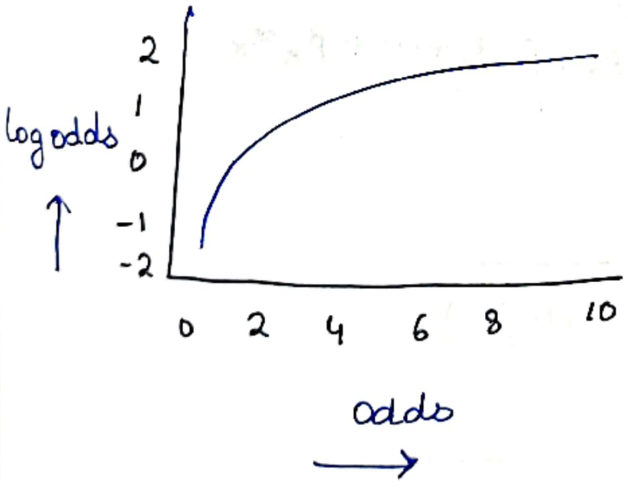
$p = 1$ $\quad \text{odds} = \frac{1}{1-1} = \frac{1}{0} = \infty$



Log Transformation is monotonic

from Odd to log odd

P	odds ⟷	log odd
.001	.001001	-6.906755



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

(logistic)

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K$$

p: $p = p(y=1)$

from [MLE]

$x_1, x_2, \cdots x_K$

predictor Variables

Independent Variable

Exponentiate and take the multiplicative inverse of both Sides

$$\frac{1-p}{p} = \frac{1}{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K)}$$

$$\left[\because \log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K\right]$$

$$e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K}$$

$$\frac{1}{p} - 1 = \frac{1}{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K)}$$

$$\frac{1}{p} = 1 + \frac{1}{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K)}$$

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

$$\frac{1}{p} = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K) + 1}{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K)}$$

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

$$p = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K)}$$

$p = 0$ to 1

Interpreting the logistic Regression Equation.

Prob is 0 to 1 Prob(Success)

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

odd $\rightarrow$ 0 to ∞

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

The odds of Success are 4 to 1

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

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

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

(dependent Variable (Coupon))

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

Odd ratio :

$$\text{Odds Ratio} = \frac{\text{Odds}_1}{\text{Odd}\,S_0}$$

1^{St} Level

0^{th} 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 using the coupon for customers who Spend $2000 annually and do not have a Simmons Credit card ($x_1 = 2$ and $x_2 = 0$).

Compare odds of using the coupon

$x_1 = 2$ $x_2 = 1$

$x_1 = 2$ $x_2 = 0$

} we are interested in interpreting the effect of a one-unit increase in the independent Variable x_2

$$\text{Odds}_1 = \frac{P(y=1 | x_1=2, x_2=1)}{1 - P(y=1 | x_1=2, x_2=1)}$$

$$\text{estimate of odds}_1 = \frac{0.4099}{1-0.4099} = 0.6946$$

$$\text{Odds}_0 = \frac{P(y=1 | x_1=2, x_2=0)}{1 - P(y=1 | x_1=2, x_2=0)}$$

$$\text{estimate of Odds}_0 = \frac{0.1880}{1-0.1880} = 0.2315$$

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

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 = Simmon's 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 (x2) using $3000, instead of $2000, as the value for annual Spending variable (x1), we would Still Obtain the Same value for the estimated odds ratio (3.00)

$$\text{odd } s_1 = \frac{0.4943}{1-0.4943} = \frac{0.4943}{0.5057} = 0.97745699$$

$$\text{odd } s_0 = \frac{0.2457}{1-0.2457} = \frac{0.2457}{0.7543} = 0.325732$$

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

Relationship b/w odd ratios and Coefficient of independent Variables

	Coef.	Std Err	Z	P>|z|
Const	-2.1464	0.5772	-3.7183	0.0002
x_2 card	1.0987	0.4447	2.4707	0.0135
x_1 spending	0.3416	0.1287	2.6551	0.0079

Estimated odds ratio $= e^{b_1} = e^{0.341643} = 1.41$

$= e^{b_2} = e^{1.09873} = 3.00$ ✓

Odd ratio $= e^{\beta_i}$

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 eo. 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.

G, Z — Logistic Regression.

F, T — Linear Regression.

Linear Regression Model VS Logistic Regression Model.

Linear Regression model:

$Y_1 = X_1 + X_2 + \cdots + X_n$ (Dependent Variable; independent Variable)

where,

Y_1 is Continous data

Independent Variables = non metric and metric

non metric → discrete

metric → continous

Logistic Regression model

$Y_1 = X_1 + X_2 + \cdots + X_n$ (Independent; dependent)

where,

Y_1 = Binary nonmetric → Discrete

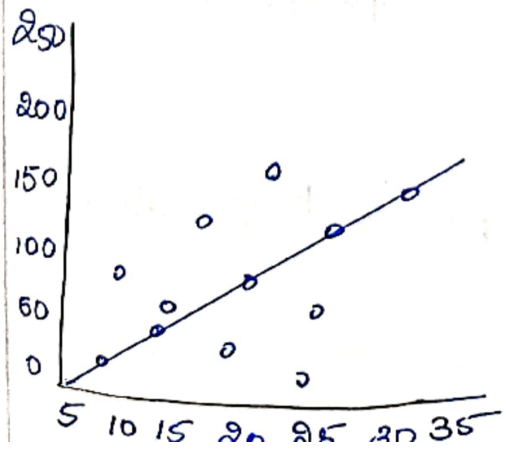
Independent Variables = non metric and metric

non metric → Discrete

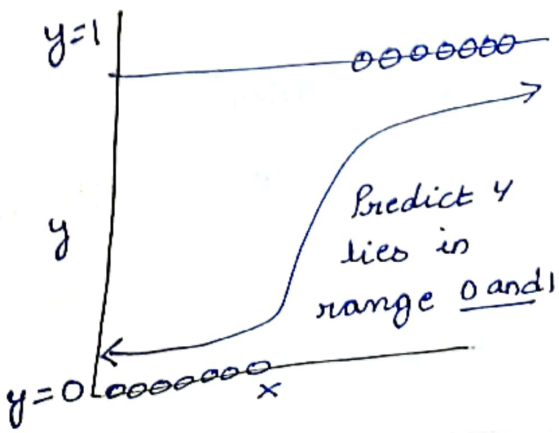
metric → continous

Graphical Representation:

Linear Regression



Logistic Regression



correspondence of primary Elements of model fit

Linear Regression	Logistic
Total Sum of Squares SST	-2LL of base model → log likelihood
Error Sum of Squares SSE	-2LL of proposed model
F test of model fit $F = \frac{MSR}{MSE} = \frac{SSR/K}{SSE/n-K-1}$	G Test chi-Square test -2LL difference
Coefficient of determination (R^2)	Pseudo R^2 measures
Regression Sum of Squares.	Difference of -2LL for base and proposed models

Base model:

There is no independent Variable

Determination of Coefficients

Linear Regression

R^2

$r^2 = SSR / SST$

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 = loglikelyhood

-2LL = -2LL of base model

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

Testing for Overall Significance:

Linear:	Logistic
F-test of model fit	G-test
$F = MSR / MSE$	$G = -2 \ln \left[\text{likelihood without a variable} / \text{with variable} \right]$

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	$2LL_{null} - (2LL_{model})$
normally distributed	Binomially

Linear Reg. assumes that residuals are approxily 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	maximum likelihood 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.	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

Date:

15/2/24

Motivation for Regression

Galton's theory:

1850's - scientists. → Darwin's theory (inspiration)

↓

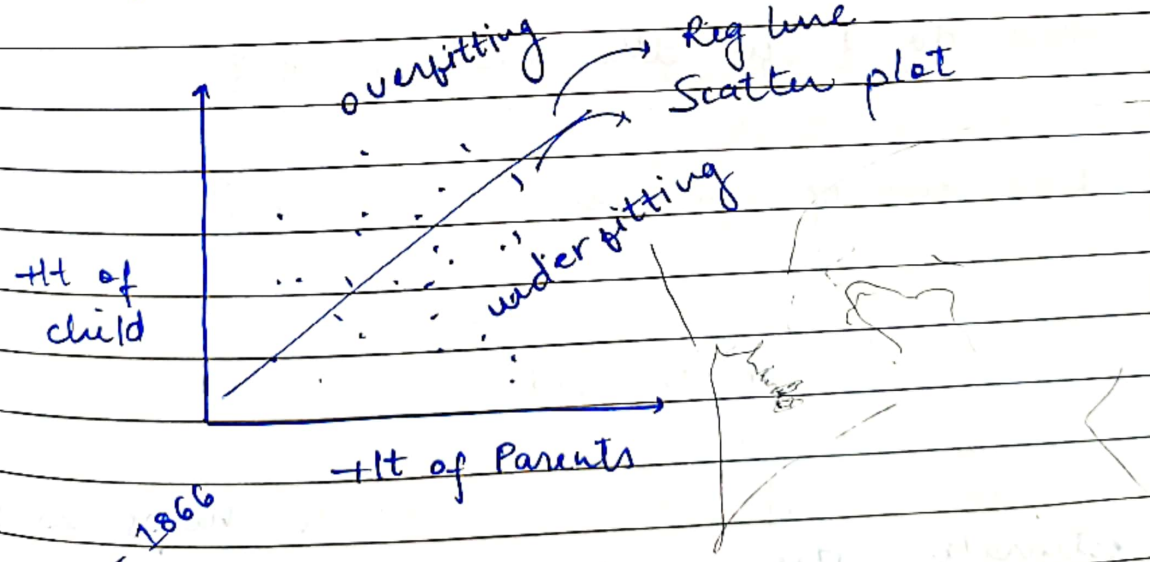
plants also have life

Mean → Theory of evolution.

202 families → 905 adults.

Male + 1.8 x Female. why ?

height of child ⟶ Parent's height



1st paper - 1866 on Regression

why Gaussian distribution

Date: / /

Examples for Simple linear Regression:

1. A Hospital may be interested in finding how, treatment cost of a patient varies with body weight of the person

2. Restaurants would like to know the relⁿships b/n the customer waiting time after placing the order & net promotr score (nps)

3. Bank would like to understand the impact of unemployement rate on % of non performing assets.

* (Non-performing Assets)

* Examples for Multi

1. Salary of MBA Stadents at the time of Graduation may depend on factors such as their academic perfomance, prior work experience, communication skills &

2. Market share of brand may depend on factor such as price, promotion expenses, competitors price, ...

Date: / /

Logistic Regression

Classification problems

1) Human research dept. of an industry may try to predict whether an applicant would accept the job offer or not

2) Sentiment abt a product / a service in social media can be classified as +ve, -ve or neutral which enables an organization to understand the sentiments abt their product / service. Orgnizations may like to understand resons for -ve sentiment if exists & take necessary actions.