Regression self study

BITS Pilani
Pilani Campus

Akanksha Bharadwaj
Asst. Professor, CS/IS

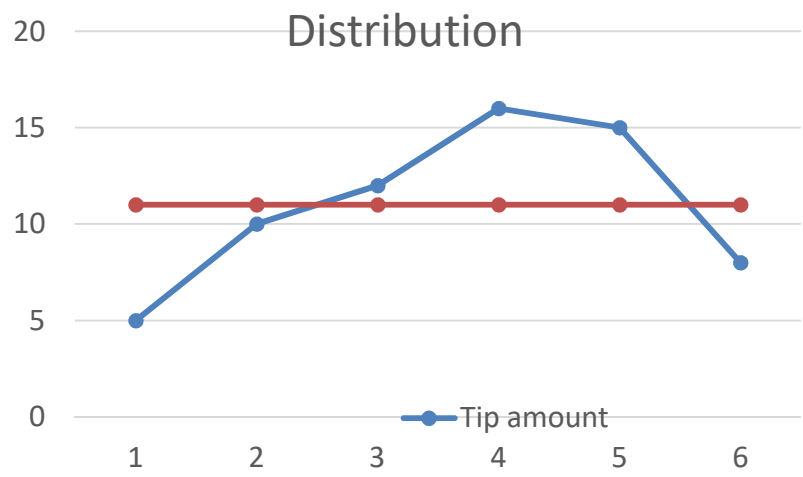
Example Scenario

- Suppose you are a delivery executive and tip is major part of your income. You want to develop a model that will allow you to make prediction about what amount of tip of expect for any given bill amount. So, one day you collect data for 6 deliveries.

S. No	Tip amount
1	5
2	10
3	12
4	16
5	15
6	8

Scenario continued

- Later you realize that you forgot to capture the bill amount and have only tip amount for prediction
- What will you predict the tip will be for 7th order.



- The only thing we can predict is average 11 and this can be the expected outcome for tip of 7th order

Goodness of fit for the tips

- Measure the distance from best fit line
- This distance is referred to as residual or errors
- The **sum** of the **residuals** always equals **zero** (assuming that your line is actually the line of "best fit."

S. No	Tip amount	Distance from mean/ residuals
1	5	5 - 11 = -6
2	10	10 - 11 = -1
3	12	12 - 11 = 1
4	16	16 - 11 = 5
5	15	15 – 11 = 4
6	8	8 – 11 = -3

Squaring the residuals

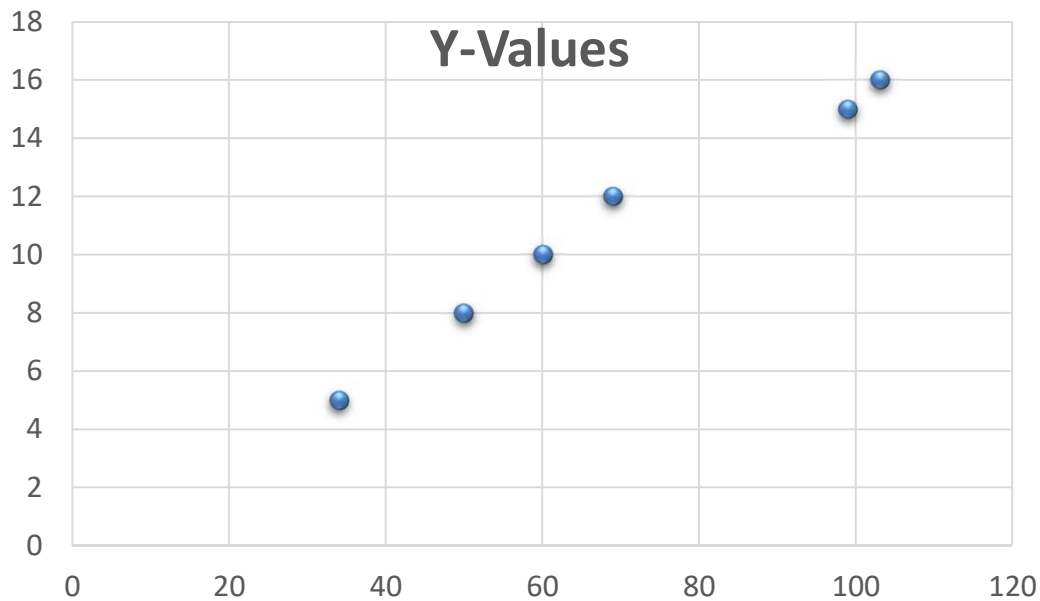
- Sum of Residual^2 = 88
- This is referred to as SSE (Sum of squared errors)
- Here SST= SSE= 88

S. No	Tip amount	Distance from mean/ residuals	Residual^2
1	5	5 - 11 = -6	36
2	10	10 - 11 = -1	1
3	12	12 - 11 = 1	1
4	16	16 - 11 = 5	25
5	15	15 – 11 = 4	16
6	8	8 – 11 = -3	9

- The goal of simple linear regression is to find the best fit line that minimizes the SSE

Example: Add Bill amount to sample data

- Suppose we are able to get Bill amount for our previous tip data that was collected.
- **Step1**- Draw scatter plot

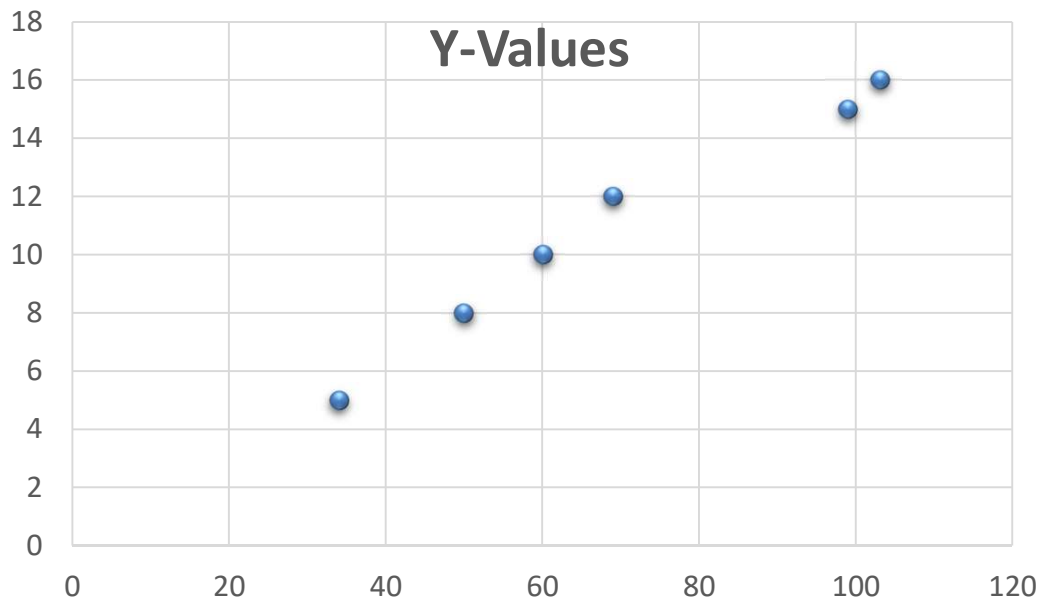


S. No	Tip amount (y)	Bill amount (x)
1	5	34
2	10	60
3	12	69
4	16	103
5	15	99
6	8	50

Example

Step2- Look for a visual line

Does the data seem following a line?



Example

Step3- Correlation coefficient (optional step)

S. No	Tip amount (y)	Bill amount (x)	xy
1	5	34	170
2	10	60	600
3	12	69	828
4	16	103	1648
5	15	99	1485
6	8	50	400
sum	66	415	5131
sum of squares	814	32427	

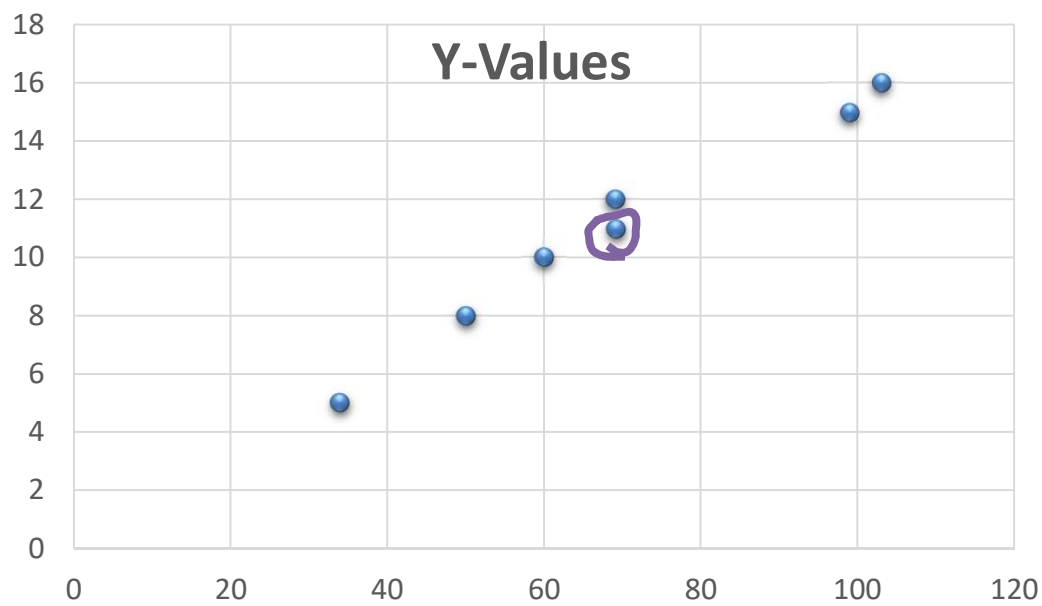
r is coming as **0.988**

In this case relationship between x and y is strong and positive

Example

Step4: Centroid

- For this take the mean of tip amount (11) and mean of bill amount (69.17)
- **The best fit regression line/least squares line has to pass through this centroid**



Example

To get the line of equation calculate b_1 and b_0

b_1 is 0.152

b_0 is 0.486

S. No	Tip amount (y)	Bill amount (x)	Bill deviation	Tip deviation	Deviation products
1	5	34	-35.17	-6	211.02
2	10	60	-9.17	-1	9.17
3	12	69	-0.17	1	-0.17
4	16	103	33.83	5	169.15
5	15	99	29.83	4	119.32
6	8	50	-19.17	-3	57.51
Mean	11	69.16666667			566
sum of squares			3722.8334		

Interpretation

- b_1 tells that for every dollar increase in bill amount tip amount increases by 0.15 dollar

- b_0 tells that when bill amount is zero dollar tip expected is 0.486. This may or may not make any sense in real world

Error

- Earlier with one variable and line through mean as best fit SSE was 88
- For our current model SSE is reduced considerably to **1.948344**

S. No	Observed Tip amount (y)	Observed Bill amount (x)	Predicted tip amount	residuals/error
1	5	34	5.654	-0.654
2	10	60	9.606	0.394
3	12	69	10.974	1.026
4	16	103	16.142	-0.142
5	15	99	15.534	-0.534
6	8	50	8.086	-0.086
sum of squares				**1.948344**

- SST= SSE + SSR
- SSR is sum of squares due to regression and value is 86.051656 (i.e. 88 – SSE)

COEFFICIENT OF DETERMINATION (12.5)

$$r^2 = 1 - \frac{SSE}{SS_{yy}} = 1 - \frac{SSE}{\Sigma y^2 - \frac{(\Sigma y)^2}{n}}$$

Note: $0 \le r^2 \le 1$

- For tip example SSR and value is 86.051656
- SST is 88
- So, COD is 0.9778
- We can say that 97.78% of the total sum of squares can be explained by using the estimated regression equation to predict tip amount. The remainder is error.
- We want the SSE to be as small as possible so that model can be a **good fit.**

Question

Ten-Year Sales Data for Huntsville Chemicals

Year	Sales ($ millions)
2000	7.84
2001	12.26
2002	13.11
2003	15.78
2004	21.29
2005	25.68
2006	23.80
2007	26.43
2008	29.16
2009	33.06

Find the best fit line

Year x	Sales y	x^2	xy
2000	7.84	4,000,000	15,680.00
2001	12.26	4,004,001	24,532.26
2002	13.11	4,008,004	26,246.22
2003	15.78	4,012,009	31,607.34
2004	21.29	4,016,016	42,665.16
2005	25.68	4,020,025	51,488.40
2006	23.80	4,024,036	47,742.80
2007	26.43	4,028,049	53,045.01
2008	29.16	4,032,064	58,553.28
2009	33.06	4,036,081	66,417.54
$\Sigma x = 20{,}045$	$\Sigma y = 208.41$	$\Sigma x^2 = 40{,}180{,}285$	$\Sigma xy = 417{,}978.01$

$$b_1 = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{(417{,}978.01) - \frac{(20{,}045)(208.41)}{10}}{40{,}180{,}285 - \frac{(20{,}045)^2}{10}} = \frac{220.17}{82.5} = 2.6687$$

$$b_0 = \frac{\sum y}{n} - b_1\frac{\sum x}{n} = \frac{208.41}{10} - (2.6687)\frac{20{,}045}{10} = -5{,}328.57$$

Equation of the Trend Line: $\hat{y} = -5{,}328.57 + 2.6687x$

- Company want to predict sales for the year 2012 using the equation of the trend line developed from their historical time series data.

$$\hat{y}(2012) = -5{,}328.57 + 2.6687(2012) = 40.85$$

STANDARD ERROR OF THE ESTIMATE

- Residuals represent errors of estimation for individual points. With large samples of data, residual computations become laborious.

SUM OF SQUARES OF ERROR	$SSE = \Sigma(y - \hat{y})^2$

COMPUTATIONAL FORMULA FOR SSE	$SSE = \Sigma y^2 - b_0 \Sigma y - b_1 \Sigma xy$

- The **standard error of the estimate**, denoted s_e, is *a standard deviation of the error of the regression model* and has a more practical use than SSE

STANDARD ERROR OF THE ESTIMATE	$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$

- The reason why we divide by n - 2 and not n - 1 has to do with the degrees of freedom issue.

How is the standard error of the estimate used?

- The standard error of the estimate is a standard deviation of error.
- One of the assumptions for regression states that for a given x the error terms are normally distributed.
- Because the error terms are normally distributed, s_e is the standard deviation of error, and the average error is zero, approximately 68% of the error values (residuals) should be within 0 +- $1s_e$ and 95% of the error values (residuals) should be within 0 +- $2s_e$.
- By having knowledge of the variables being studied and by examining the value of s_e , the researcher can often make a **judgment** about the **fit** of the **regression model** to the data by using s_e .

- In addition, some researchers use the standard error of the estimate to identify outliers. They do so by looking for data that are outside +-2s_e or +-3s_e .

- The standard error of the estimate provides a single measure of error, which, if the researcher has enough background in the area being analyzed, can be used to understand the magnitude of errors in the model.

Testing the Overall Model

- It is common in regression analysis to compute an ***F* test** to determine the overall significance of the model.

In the case of simple regression analysis, $F = t^2$. Thus, for the airline cost example, the F value is

$$F = t^2 = (9.43)^2 = 88.92$$

The F value is computed directly by

$$F = \frac{SS_{reg}/df_{reg}}{SS_{err}/df_{err}} = \frac{MS_{reg}}{MS_{err}}$$

where

$df_{reg} = k$
$df_{err} = n - k - 1$
k = the number of independent variables

- The values of the sum of squares (SS), degrees of freedom (df), and mean squares (MS) are obtained from the analysis of variance table

Airline example

Analysis of Variance

Source	DF	SS	MS	F	p
Regression	1	2.7980	2.7980	89.09	0.000
Residual Error	10	0.3141	0.0314		
Total	11	3.1121			

The F value for the airline cost example is calculated from the analysis of variance table information as

$$F = \frac{2.7980/1}{.3141/10} = \frac{2.7980}{.03141} = 89.09$$

- The difference between this value (89.09) and the value obtained by squaring the *t* statistic (88.92) is due to rounding error.
- This output value means it is highly unlikely that the population slope is zero

Multiple Linear Regression

When k = 2, the model becomes $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

The normal equations obtained from least squares principles are

$$\sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_{1i} + \beta_2 \sum_{i=1}^{n} X_{2i}$$

$$\sum_{i=1}^{n} X_{1i} Y_i = \beta_0 \sum_{i=1}^{n} X_{1i} + \beta_1 \sum_{i=1}^{n} X_{1i}^{\ 2} + \beta_2 \sum_{i=1}^{n} X_{1i} X_{2i}$$

$$\sum_{i=1}^{n} X_{2i} Y_i = \beta_0 \sum_{i=1}^{n} X_{2i} + \beta_1 \sum_{i=1}^{n} X_{1i} X_{2i} + \beta_2 \sum_{i=1}^{n} X_{2i}^{\ 2}$$

Multiple Linear Regression

- For k = 2, based on the sample data the model can be written as $Y = b_0 + b_1X_1 + b_2X_2 + \varepsilon$
- The normal equations obtained from least squares principles are

$$\sum_{i=1}^{n} Y_i = nb_0 + b_1\sum_{i=1}^{n} X_{1i} + b_2\sum_{i=1}^{n} X_{2i}$$

$$\sum_{i=1}^{n} X_{1i}Y_i = b_0\sum_{i=1}^{n} X_{1i} + b_1\sum_{i=1}^{n} X_{1i}^{2} + b_2\sum_{i=1}^{n} X_{1i}X_{2i}$$

$$\sum_{i=1}^{n} X_{2i}Y_i = b_0\sum_{i=1}^{n} X_{2i} + b_1\sum_{i=1}^{n} X_{1i}X_{2i} + b_2\sum_{i=1}^{n} X_{2i}^{2}$$

Question

EXAMPLE 12-1 Wire Bond Strength

In Chapter 1, we used data on pull strength of a wire bond in a semiconductor manufacturing process, wire length, and die height to illustrate building an empirical model. We will use the same data, repeated for convenience in Table 12-2, and show the details of estimating the model parameters. A three-dimensional scatter plot of the data is presented in Fig. 1-15. Figure 12-4 shows a matrix of two-dimensional scatter plots of the data. These displays can be helpful in visualizing the relationships among variables in a multivariable data set. For example, the plot indicates that there is a strong linear relationship between strength and wire length.

Table 12-2 Wire Bond Data for Example 12-1

Observation Number	Pull Strength y	Wire Length x_1	Die Height x_2	Observation Number	Pull Strength y	Wire Length x_1	Die Height x_2
1	9.95	2	50	14	11.66	2	360
2	24.45	8	110	15	21.65	4	205
3	31.75	11	120	16	17.89	4	400
4	35.00	10	550	17	69.00	20	600
5	25.02	8	295	18	10.30	1	585
6	16.86	4	200	19	34.93	10	540
7	14.38	2	375	20	46.59	15	250
8	9.60	2	52	21	44.88	15	290
9	24.35	9	100	22	54.12	16	510
10	27.50	8	300	23	56.63	17	590
11	17.08	4	412	24	22.13	6	100
12	37.00	11	400	25	21.15	5	400
13	41.95	12	500				

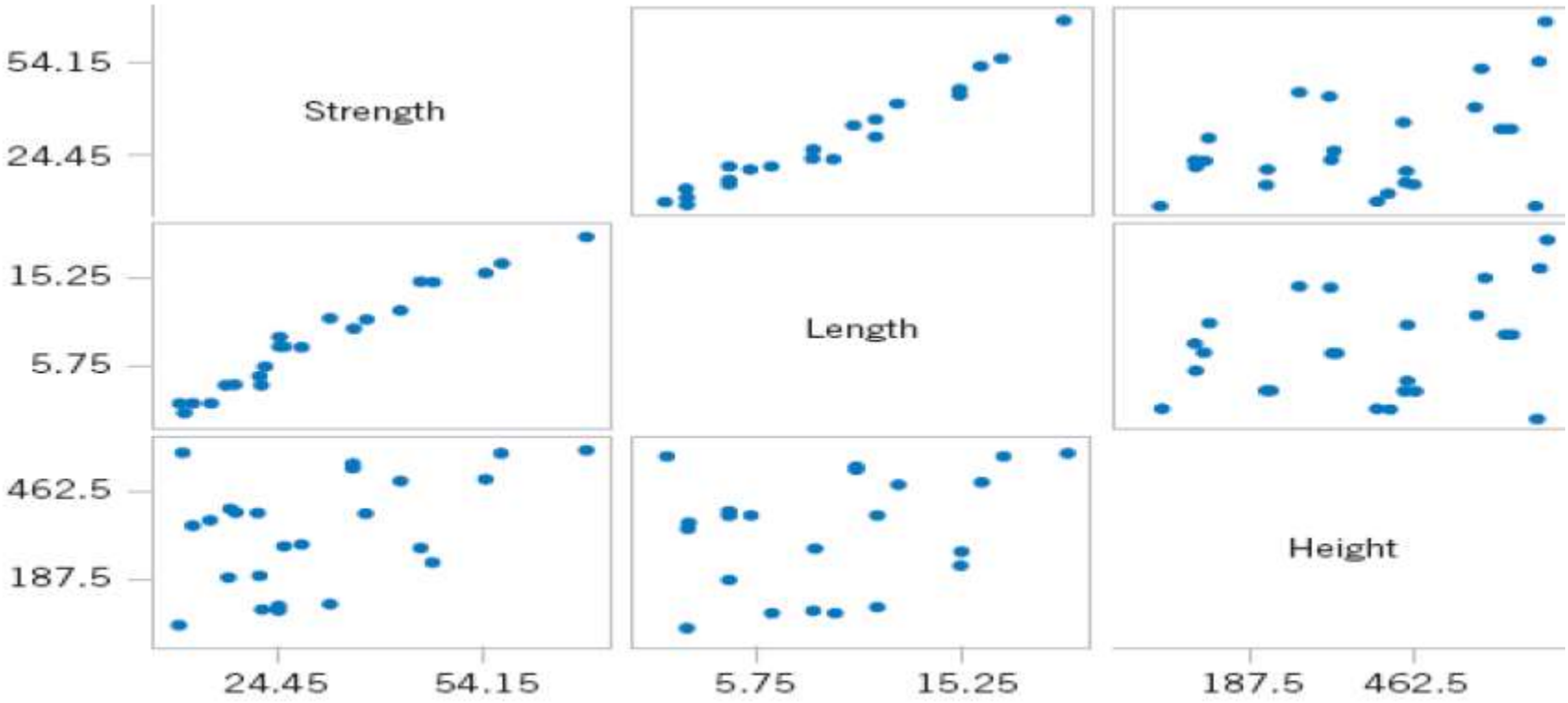


Figure 12-4 Matrix of scatter plots (from Minitab) for the wire bond pull strength data in Table 12-2.

Solution

$$\text{Pull strength} = b_0 + b_1 \text{Wire length} + b_2 \text{Die height} + \varepsilon$$

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \varepsilon$$

The normal equations are

$$25b_0 + 206b_1 + 8294b_2 = 725.82$$

$$206b_0 + 2396b_1 + 77177b_1 = 8008.47$$

$$8294b_0 + 77177b_1 + 3531848b_2 = 274816.71$$

Solving these normal equations we get

$b_0 = 2.26379$, $b_1 = 2.74427$, $b_2 = 0.01253$

The fitted regression line is

$Y = 2.26379 + 2.74427\, X_1 + 0.01253\, X_2$

- The predicted probabilities can be greater than 1 or less than 0
  - Probabilities, by definition, have max =1; min = 0;
  - This is not a big issue if they are very close to 0 and 1
- The error terms vary based on size of X-variable ("heteroskedastic") –
  - There may be models that have lower variance – more "efficient"
- The errors are not normally distributed because Y takes on only two values
  - Creates problems for
  - More of an issue for statistical theorists

Logistic Regression

- **The model that describes the S-type curve is as follows**

- **Let ‘p’ be the probability that an event ‘Y’ occurs, ie., P(Y=1)**

- **Let ‘1 -p’ be the probability that an event ‘Y’ do not occurs, ie., P(Y=0)**

- The solution of the Logistic Regression Model can be obtained from **Maximum Likelihood Method**.

- However, direct method of estimation may be difficult because of complexity in function and should be solved iteratively using computers.

$$\text{Odds (Event)} = \frac{\text{Probability (Event)}}{1 - \text{Probability (Event)}}$$

$$\text{Probability (Event)} = \frac{\text{Odds (Event)}}{1 + \text{Odds (Event)}}$$

Example

- Probability of a success = 0.8
- Probability of a failure = 0.2

Odds of success

$$\text{Odds Ratio} = \frac{\text{Probability of success}}{\text{Probability of failure}} = \frac{0.8}{0.2}$$

$$= 4$$

The probability ranges from 0 to 1

Odds ranges from o to +∞

Why to take this trouble transforming from probability to log odds?

- Usually difficult to model a variable which has restricted range such as probability ($0 \leq p \leq 1$)

Reason

- This transformation is an attempt to get around this restricted range problem

- It maps probability ranging between 0 and 1 to log odds ranging from $-\infty$ to $+\infty$

Another reason

- Among all of the infinitely many choices of transformation, the log of odds is one of the easiest to understand and interpret. This transformation is called **Logit transformation**.