Linear Regression

The following data* represents the number of units (Units) being replaced in a computer and the corresponding time taken to repair the computer in minutes (Minutes)

independent

dependent

IITP 39Km Railway | 6 AM - 6PM

computers

Units	Minutes
1	23
2	29
3	49
4	64
4	74
5	87
6	96
6	97
7	109
8	119
9	149
9	145
10	154
10	166

60

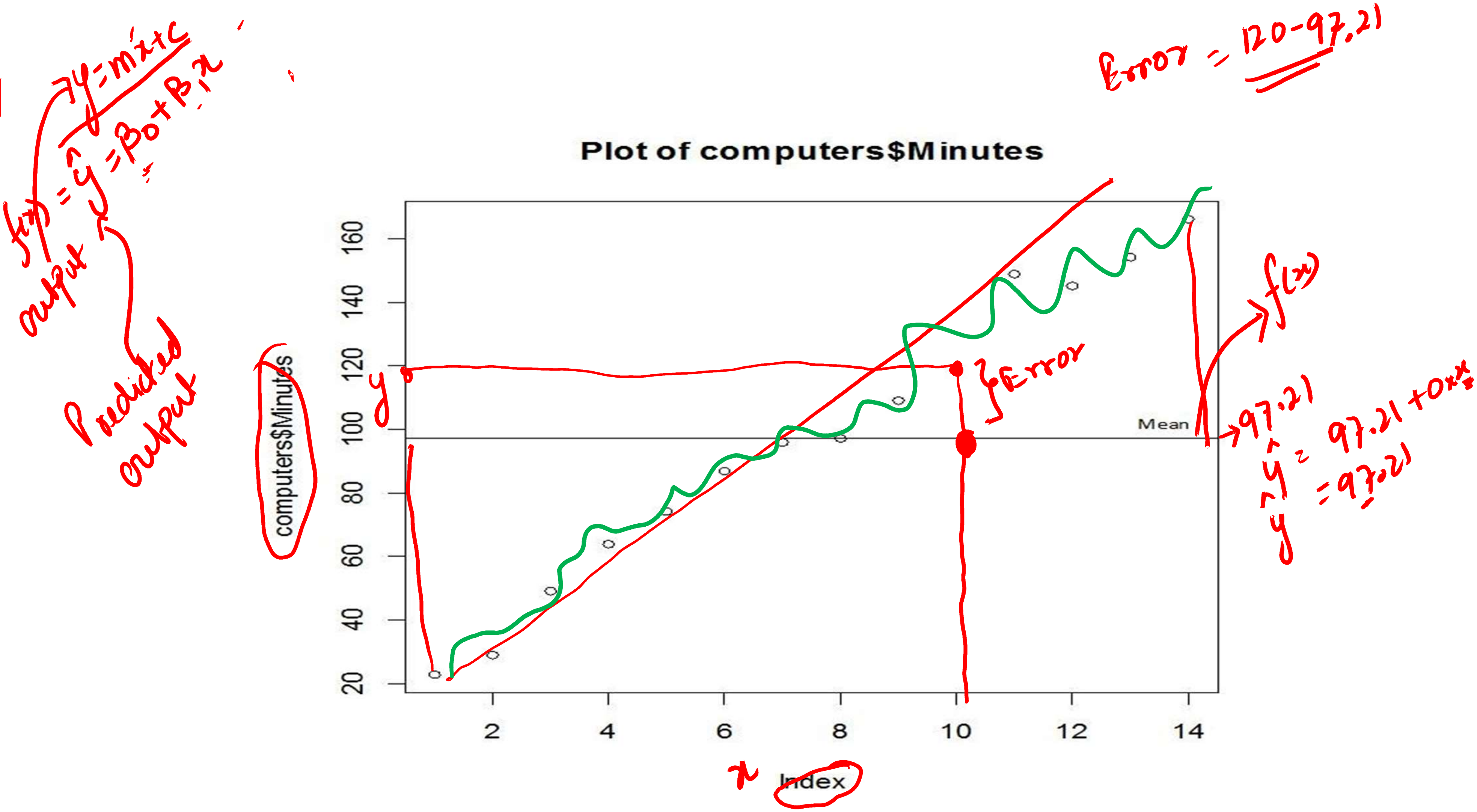
60 wait

If we were to estimate the typical time taken by the computer shop to repair a computer, which of the following would be appropriate?

a) Arithmetic mean = 97.21 minutes

b) Median = 96.50 minutes

Answer: Either Mean or Median can be used when trying to predict the expected value of a single variable.



[Data: 23 29 49 64 74 87 96 97 109 119 149 145 154 166, Mean: 97.21]

Regression Analysis

Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on

minutes units

the relationship between a dependent variable and one or more independent variables (or 'predictors').

Regression analysis focuses on building a model that can be used to predict the value of the dependent variable based on the predictor variables.

The regression model is represented using a mathematical model of form $y = f(X)$, where y is the dependent variable and X is the set of predictor variables ($x_1, x_2 \ldots x_n$).

Regression Analysis

$f(x) = \beta_0 + \beta_1 x_1 + \epsilon$

Error

dependent

features → Parameter

Independent Variable

Linear form: $f(X) = \beta_0 + \beta_1 x_1 \cancel{+ \beta_2 x_2 + \ldots + \beta_n x_n} + \epsilon$

Non-linear form: $f(X) = \beta_0 + \beta_1 {x_1}^{p_1} + \beta_2 {x_2}^{p_2} + \ldots + \beta_n {x_n}^{p_n} + \epsilon$

Regression Analysis

Some commonly used types of linear forms are:

Simple linear form – Here there is one predictor and one dependent variable : $f(X) = \beta_0 + \beta_1 x_1 + \epsilon$

Multiple linear form – Here there are multiple predictor variables and one dependent variable:

$f(X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$

Some commonly used types of non-linear forms are:

Polynomial form: $f(X) = \beta_0 + \beta_1 {x_1}^{p_1} + \beta_2 {x_2}^{p_2} + \ldots + \beta_n {x_n}^{p_n} + \epsilon$

Quadratic form: $f(X) = \beta_0 + \beta_1 {x_1}^2 + \beta_2 x_1 + \epsilon$

Logistic form: $f(x) = \frac{1}{1+e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}} + \epsilon$

classification

Where $\beta_0, \beta_1, \beta_2 \ldots \beta_n$ are said to be the regression coefficients and ϵ accounts for the error in prediction. The regression coefficients and the error in prediction are real numbers.

Regression Analysis

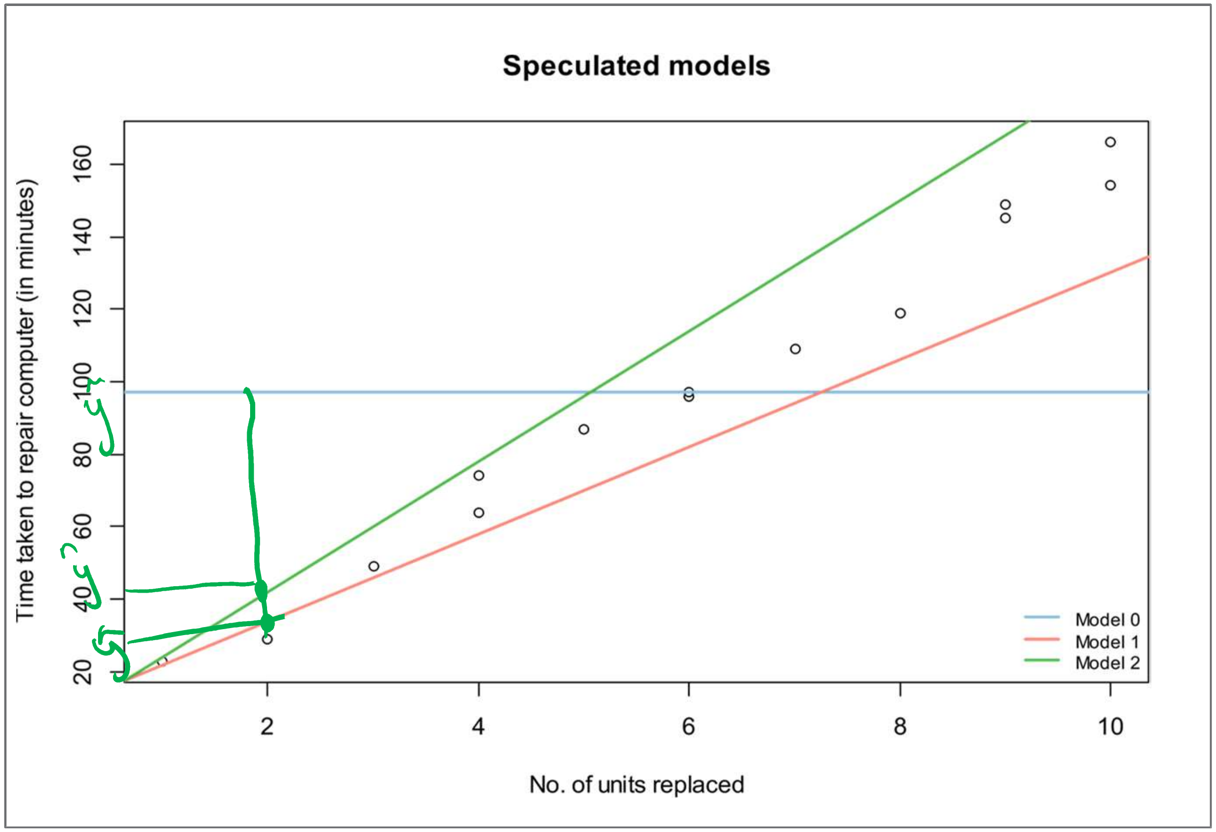
Time taken to repair the computer = $\beta_0 + \beta_1 *$ No. of units replaced

Model 0: Time taken to repair the computer = *97.21 (mean)*

Model 1: Time taken to repair the computer = $10 + 12 *$ No. of units replaced

Model 2: Time taken to repair the computer = $6 + 18 *$ No. of units replaced

Regression Analysis



Regression Analysis

model0

```
model0
```

Units replaced	Observed time taken	Expected value	Expected - Observed value
1	23	97.21429	74.2142857
2	29	97.21429	68.2142857
3	49	97.21429	48.2142857
4	64	97.21429	33.2142857
4	74	97.21429	23.2142857
5	87	97.21429	10.2142857
6	96	97.21429	1.2142857
6	97	97.21429	0.2142857
7	109	97.21429	-11.7857143
8	119	97.21429	-21.7857143
9	149	97.21429	-51.7857143
9	145	97.21429	-47.7857143
10	154	97.21429	-56.7857143
10	166	97.21429	-68.7857143

Error

$(\hat{y} - y)$

Error = Sum(minutes mean - computer$minutes)

= -8.25e-14

Error = Sum((minutes mean - computer$minutes)^2)

= 27768.36

Sum of Squared Errors

Regression Analysis

model1

Units replaced	Observed time	Expected time	Expected - Observed value
1	23	22	-1
2	29	34	5
3	49	46	-3
4	64	58	-6
4	74	58	-16
5	87	70	-17
6	96	82	-14
6	97	82	-15
7	109	94	-15
8	119	106	-13
9	149	118	-31
9	145	118	-27
10	154	130	-24
10	166	130	-36

Error = Sum((minutesmean - computer$minutes)^2)
= 4993

Regression Analysis

model2

Units replaced	Observed time	Expected time	Expected - Observed value
1	23	24	1
2	29	42	13
3	49	60	11
4	64	78	14
4	74	78	4
5	87	96	9
6	96	114	18
6	97	114	17
7	109	132	23
8	119	150	31
9	149	168	19
9	145	168	23
10	154	186	32
10	166	186	20

ε

Error = Sum((minutesmean - computer$minutes)^2)
= 5001

Regression Analysis

The best fit model is obtained by solving the linear regression model $\hat{y} = (\beta_0 + \beta_1 x) + \varepsilon$

To determine the coefficients β_0 and β_1 such that the error (as shown below) is minimum:

$$\text{Error} = \Sigma(\hat{y} - y)^2$$

$$0 = \text{Error} = \Sigma(\beta_0 + \beta_1 x - y)^2$$

$$\frac{\partial \text{Error}}{\beta_0} =$$

$$\frac{\partial \text{Error}}{\beta_1} =$$

Differentiating the equation $\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$ w.r.t b_0 and equating to 0 we get

no. of Rows

$$b_1 = \frac{[\sum_{i=1}^{n}(x_i y_i)] - n\bar{x}\bar{y}}{[\sum_{i=1}^{n} x_i^2] - n\bar{x}^2}$$

– mean
bar

Similarly, differentiating the equation $\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$ w.r.t b_1 and equating to 0 we get

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

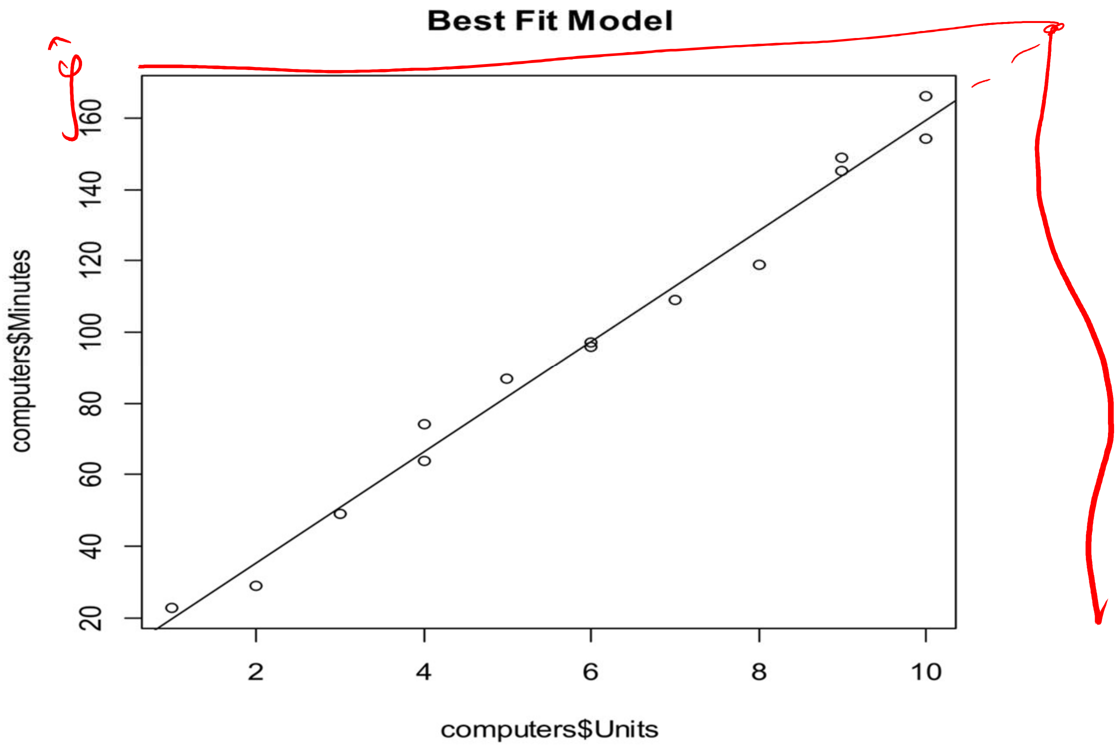
Here, b_0 and b_1 are the regression coefficients for the sample, $\bar{x}$ and $\bar{y}$ are the sample means of the variables X and Y respectively.

Best fit Model

Using the previously obtained formula, b_0 and b_1 for the given sample dataset is determined as 4.162 (approximately) and 15.509 (approximately) respectively as shown.

```
#Computing values for b0 and b1
x <- computers$Units
y <- computers$Minutes
xiyi<- x*y
n <- nrow(computers)
xmean <- mean(computers$Units)
ymean <- mean(computers$Minutes)
numerator <- sum(xiyi)-n*xmean*ymean
denominator <- sum(x^2)-n*(xmean^2)
b1 <- numerator/denominator
b0 <- ymean-b1*xmean
#values of b0 and b1
b0
[1] 4.161654
b1
[1] 15.50877
```

The plot of this model along with the given data is as shown below




```
#Plot of Best fit Model: minutes = 4.161654 + 15.50877*units to be replaced
plot(computers$Units,computers$Minutes, main="Model 1")
abline(b0,b1)
```

Best fit Model

The following code snippet shows the number of units replaced, observed time taken, expected time taken (based on the model) and the difference between predicted and observed values for the best fit model. The total sum of squared errors for the best fit model is 348.8484.

```
#Best fit model Units, Observed time, Expected time, Difference between observed and expected time
best_fit <- data.frame(matrix(data=c(computers$Units,computers$Minutes,(b0+b1*computers$Units),
                              ((b0+b1*computers$Units)-computers$Minutes)),ncol=4))
colnames(best_fit) <- c("Units replaced", "Observed time", "Expected time",
                        "Expected - Observed value")
best_fit
Units replaced Observed time Expected time Expected - Observed value
             1            23      19.67043                -3.3295739
             2            29      35.17920                 6.1791980
             3            49      50.68797                 1.6879699
             4            64      66.19674                 2.1967419
             4            74      66.19674                -7.8032581
             5            87      81.70551                -5.2944862
             6            96      97.21429                 1.2142857
             6            97      97.21429                 0.2142857
             7           109     112.72306                 3.7230576
             8           119     128.23183                 9.2318296
             9           149     143.74060                -5.2593985
             9           145     143.74060                -1.2593985
            10           154     159.24937                 5.2493734
            10           166     159.24937                -6.7506266
```

```
#Sum of squared errors for Best fit model
sum((((b0+b1*computers$Units)-computers$Minutes))^2)
[1] 348.8484
```

m0 - 27
m1 = 4997
m2 = 5001
$\boxed{m3 = 348}$

Best Fit Model

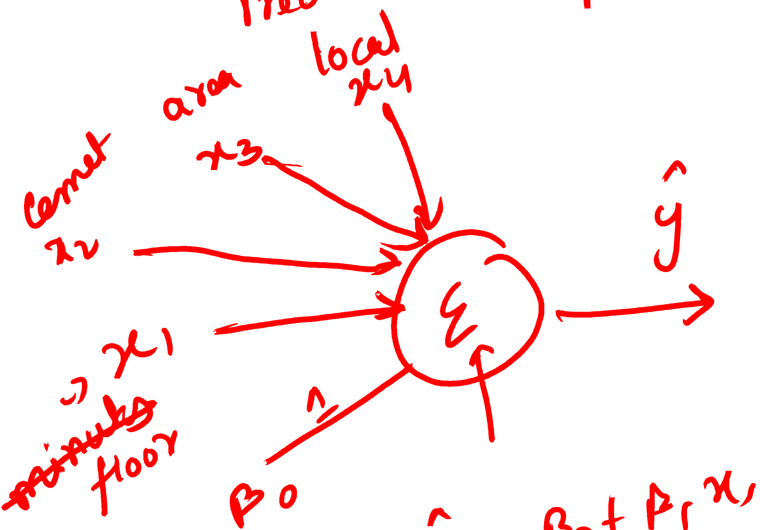
The best fit linear regression model can also be obtained in R using the lm() command as shown below:

Syntax for the lm() function : lm(dependent variable ~ predictor variable)

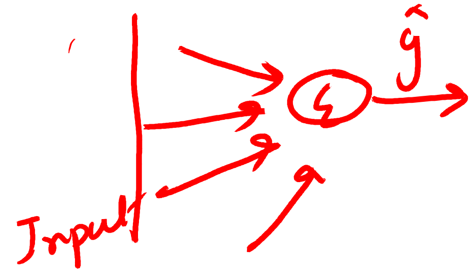
```
lm.model <- lm(computers$Minutes~computers$Units)
lm.model

Call:
  lm(formula = computers$Minutes ~ computers$Units)

Coefficients:
  (Intercept)  computers$Units
4.162            15.509
```

The value of coefficients b_0 and b_1 indicate that it takes approximately 15.509 minutes to replace a unit and a fixed time of approximately 4.162 minutes to understand a given repair.

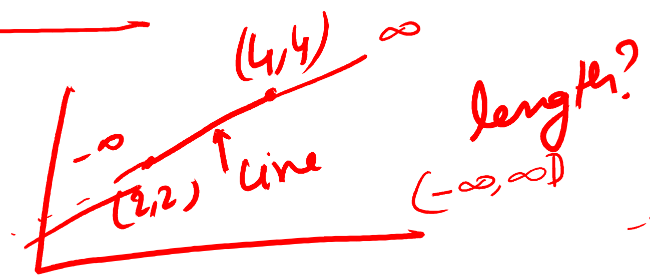
Predict Price of a flat

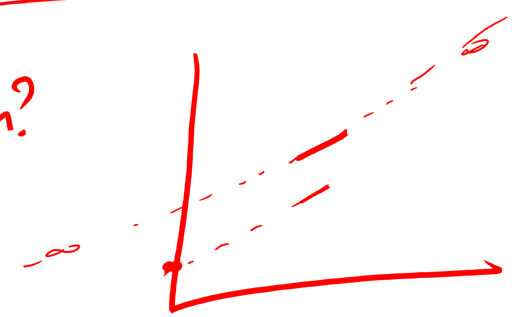


$\hat{y} = \beta_0 + \beta_1 x_1$



SVM





Logistic Regression

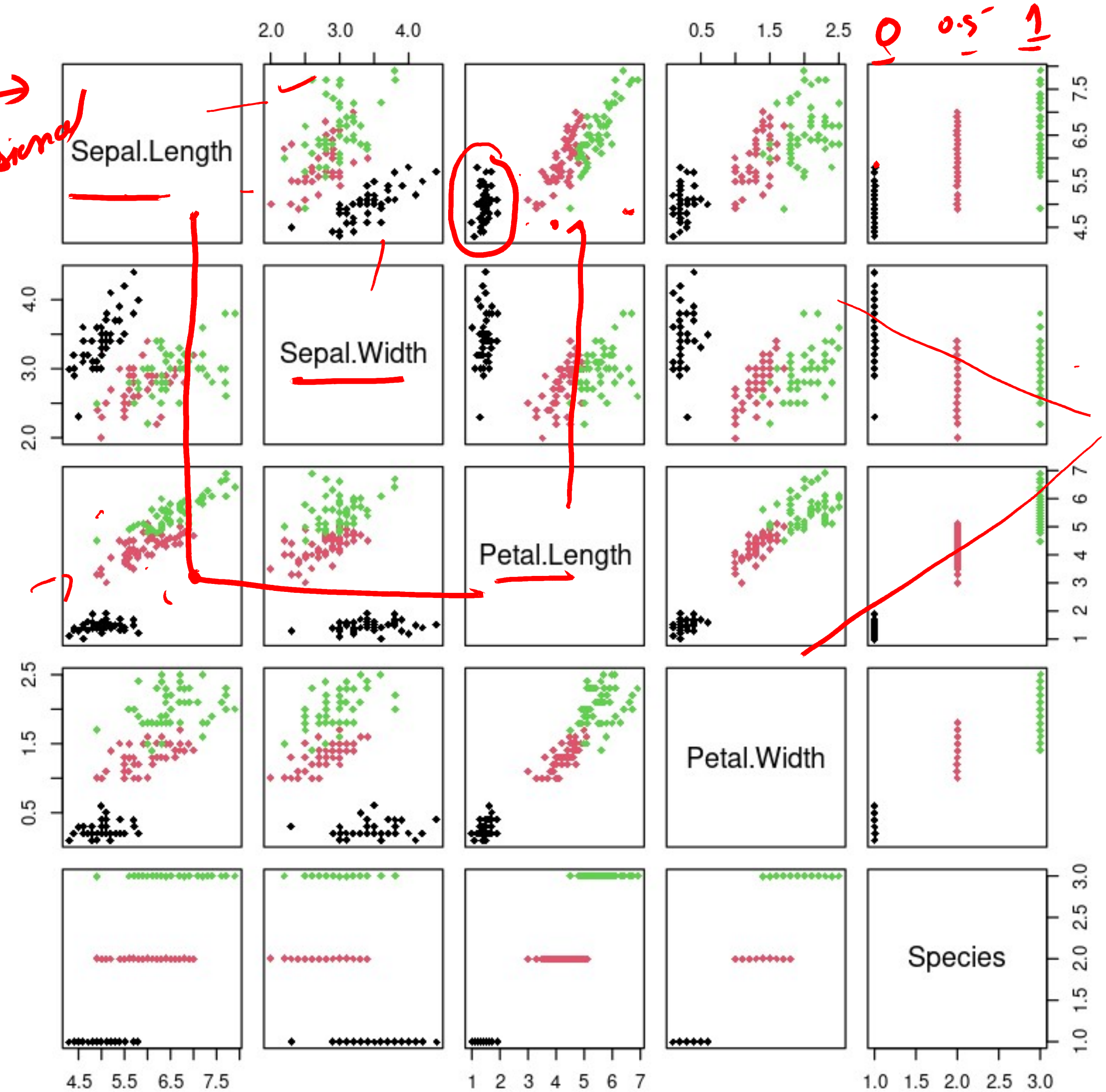
Iris Dataset

S.L, S.W, P.L, P.W

multidimensional Plot

Sepal Lengh	Sepal Width	Petal Length	Petal Width	Class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
5.6	2.5	3.9	1.1	Iris-versicolor
5.9	3.2	4.8	1.8	Iris-versicolor
6.3	2.9	5.6	1.8	Iris-virginica
6.5	3	5.8	2.2	Iris-virginica

150 Records



Learning Function

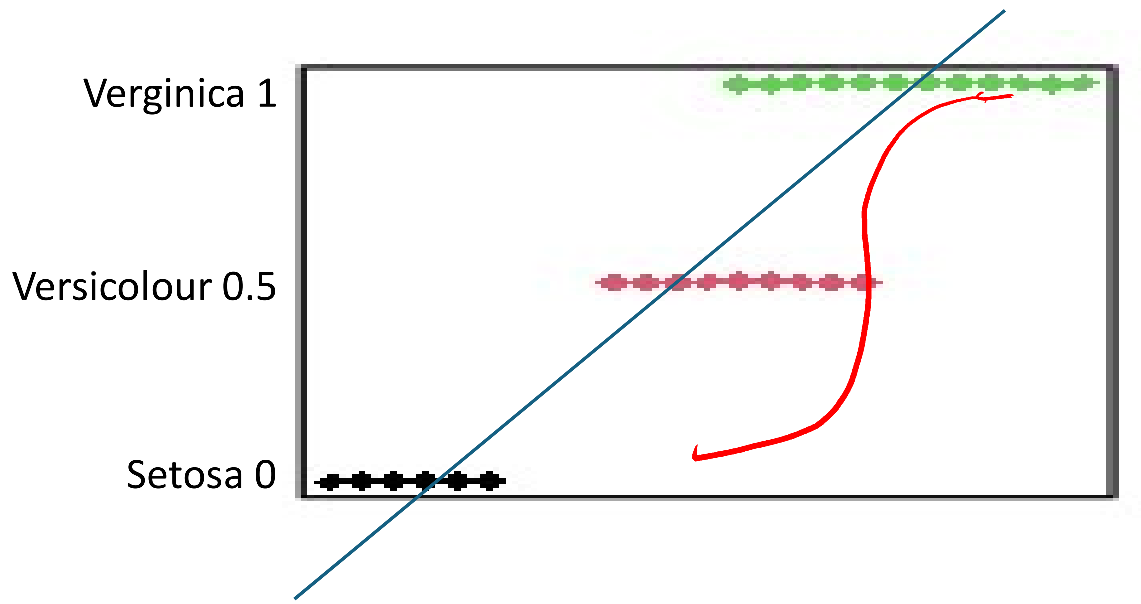
Multiple Linear Regression

$$F(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

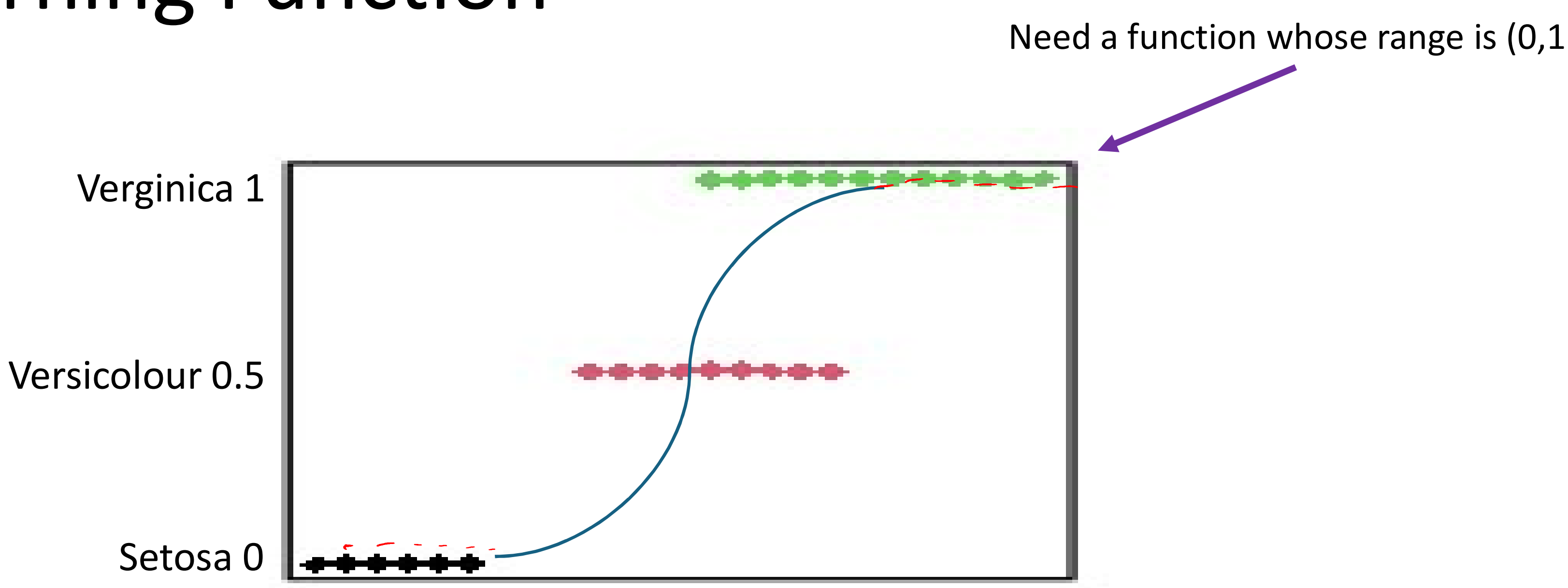
Simple Linear Regression

$$F(x) = \beta_0 + \beta_1 x_1$$

Range: (-∞ , ∞)



Learning Function



Odds

Odds of my winning to loosing is 2 is to 8.

- **Odds –** Something happening to something not happening

$$Odds = \frac{Something\ happening}{Something\ not\ happening} = \frac{Win}{Loose}$$

- **Probability** – Something happening to all the set of events

$$P_{win} = \frac{Something\ happening}{(Something\ happening) + (Something\ not\ happening)} = \frac{Win}{Win + Loose}$$

$$P_{Loose} = \frac{Something\ not\ happening}{(Something\ happening) + (Something\ not\ happening)} = \frac{Loose}{Win + Loose}$$

$$\boxed{Odds = \frac{Something\ happening}{Something\ not\ happening} = \frac{Win}{Loose} = \frac{Win\ /(Win+Loos\ \ )}{Loose\ /(Win+Loose)} = \frac{P_{win}}{P_{Loose}}} = \frac{P_{win}}{1-P_{win}}$$

Odds

loosing
~~Winning~~

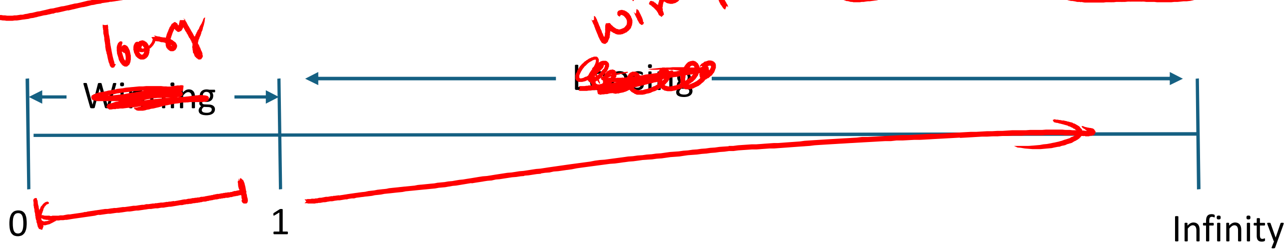
$\frac{1}{1}$ $\frac{2}{1}$

- Odd(1:2) = $\frac{1}{2}$ =0.5
- Odd(1:4) = $\frac{1}{4}$ =0.25
- Odd(1:8) = $\frac{1}{8}$ =0.125
- Odd(1:16) = $\frac{1}{16}$ =0.0625
- Odd(1:32) = $\frac{1}{32}$ =0.03125

wing
~~Loosing~~

- Odd(2:1) = $\frac{2}{1}$ = 2
- Odd(4:1) = $\frac{4}{1}$ = 4
- Odd(8:1) = $\frac{8}{1}$ = 8
- Odd(16:1) = $\frac{16}{1}$ = 16
- Odd(32:1) = $\frac{32}{1}$ = 32

6000:1



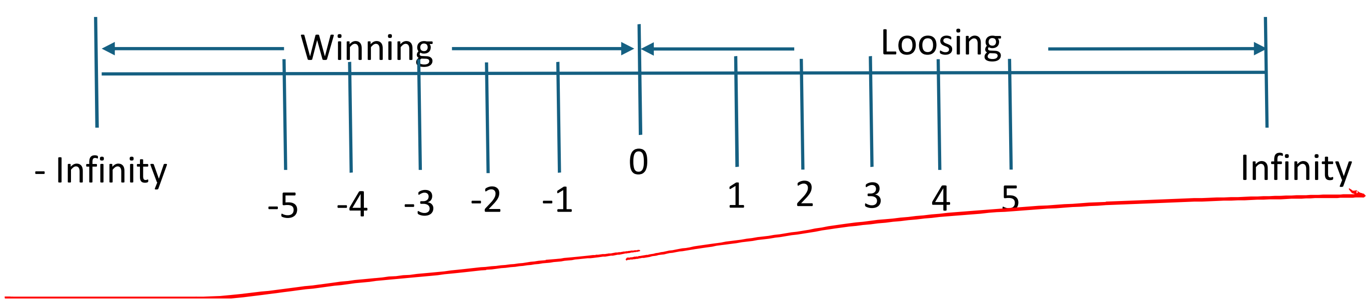
Log of Odds

Winning

- Odd(1:2) = $\frac{1}{2}$ =0.5 = log (0.5) = -1
- Odd(1:4) = $\frac{1}{4}$ =0.25 = log (0.25) = -2
- Odd(1:8) = $\frac{1}{8}$ =0.125 = log (0.125) = -3
- Odd(1:16) = $\frac{1}{16}$ =0.0625 = log (0.0625) = -4
- Odd(1:32) = $\frac{1}{32}$ =0.03125 = log (0.03125) = -5

Loosing

- Odd(2:1) = $\frac{2}{1}$ = 2 = log (2) = 1
- Odd(4:1) = $\frac{4}{1}$ = 4 = log (4) = 2
- Odd(8:1) = $\frac{8}{1}$ = 8 = log (8) = 3
- Odd(16:1) = $\frac{16}{1}$ = 16 = log (16) = 4
- Odd(32:1) = $\frac{32}{1}$ = 32 = log (32) = 5



Log of Odds

$$Odds = \frac{P_{win}}{P_{Loose}}$$

$$\log(Odds) = \log\left(\frac{P_{win}}{P_{Loose}}\right)$$

$$\log(Odds) = \log\left(\frac{P_{win}}{1 - P_{win}}\right)$$

f

$$\log\left(\frac{P_{win}}{1 - P_{win}}\right) = \beta_0 + \beta_1 x$$

$$\frac{P_{win}}{1 - P_{win}} = e^{\beta_0 + \beta_1 x}$$

$$P_{win} = (1 - P_{win})e^{\beta_0 + \beta_1 x}$$

$$P_{win} = 1e^{\beta_0 + \beta_1 x} - P_{win}e^{\beta_0 + \beta_1 x}$$

$$P_{win} + P_{win}e^{\beta_0 + \beta_1 x} = 1e^{\beta_0 + \beta_1 x}$$

$$P_{win}(1 + e^{\beta_0 + \beta_1 x}) = 1e^{\beta_0 + \beta_1 x}$$

(0-1)

$$P_{win} = \frac{e^{\beta_0 + \beta_1 x}}{(1 + e^{\beta_0 + \beta_1 x})}$$

Divide numerator and Denominator by $e^{\beta_0 + \beta_1 x}$

$$\boxed{P_{win} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}}$$

logistic Regression

Logistic Function/Sigmoid Function