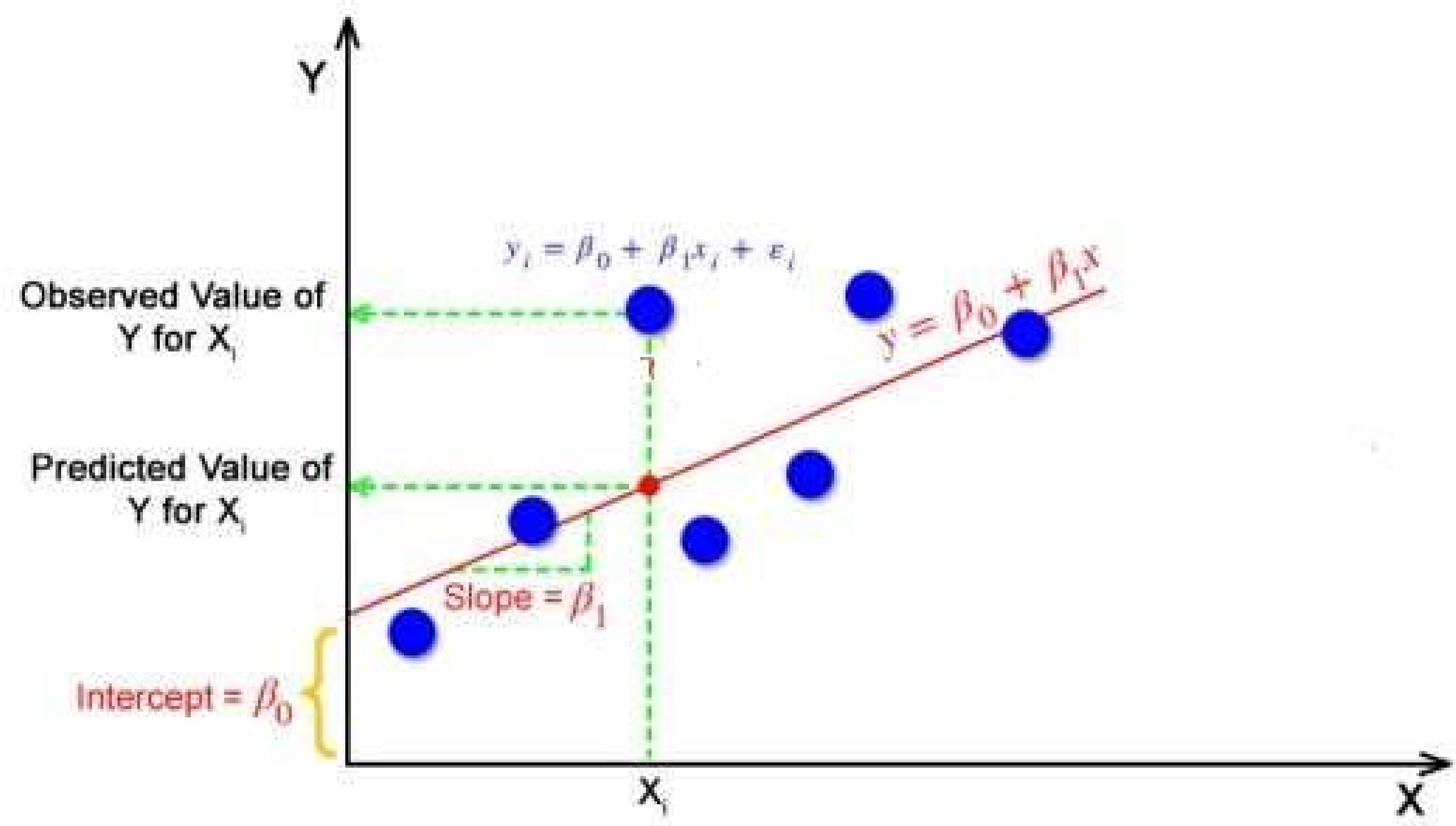
INTRODUCTION TO BUSINESS ANALYTICS

Module 2

Simple Linear Regression

Simple Linear Regression equation

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



Matrix Equations to solve problem using Simple Linear Regression Analysis

By Applying Method of Least squares Criteria, we obtain normal equations in matrix form

$$Y = \beta_0 + \beta_1 x$$

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

All regression problems are solved using the above matrix equations

Matrix approach

Data for simple linear regression

y	x
y_1	x_1
y_2	x_2
$\vdots$	$\vdots$
y_n	x_n

Regression equation

$$Y = \beta_0 + \beta_1 x$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Problem

In an electrochemical machining process, the amount of metal removed (y) is related to the gap (x) between tool and work piece. The following observations are obtained from the experiment. Fit a linear regression model and find its coefficients.

Solution:

y (in grams)	28	20	25	11	17
x (in mm)	5	3	4	1	2

$$Y = \beta_0 + \beta_1 x$$

$$X = \begin{bmatrix} 1 & 5 \\ 1 & 3 \\ 1 & 4 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \quad Y = \begin{bmatrix} 28 \\ 20 \\ 25 \\ 11 \\ 17 \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Using the matrix approach we get $\beta = \begin{bmatrix} 7.6 \\ 4.2 \end{bmatrix}$

The estimated fitted line is **y = 7.6 + 4.2 x**

2) Do Simple linear Regression analysis using the data

x	1.1	1.4	1.3	1.5
y	89	94	91	95

$y = \beta_0 + \beta_1 x$

$$x = \begin{bmatrix} 1 & 1.1 \\ 1 & 1.4 \\ 1 & 1.3 \\ 1 & 1.5 \end{bmatrix} \quad y = \begin{bmatrix} 89 \\ 94 \\ 91 \\ 95 \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$[x^T x] = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1.1 & 1.4 & 1.3 & 1.5 \end{bmatrix} \begin{bmatrix} 1 & 1.1 \\ 1 & 1.4 \\ 1 & 1.3 \\ 1 & 1.5 \end{bmatrix} = \begin{bmatrix} 4 & 5.3 \\ 5.3 & 7.11 \end{bmatrix}$$

$$[x^T y] = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1.1 & 1.4 & 1.3 & 1.5 \end{bmatrix} \begin{bmatrix} 89 \\ 94 \\ 91 \\ 95 \end{bmatrix} = \begin{bmatrix} 369 \\ 490.3 \end{bmatrix}$$

$$[x^T x]^{-1} [x^T y] = \begin{bmatrix} 20.314 & -15.14 \\ -15.14 & 11.428 \end{bmatrix} \begin{bmatrix} 369 \\ 490.3 \end{bmatrix} = \begin{bmatrix} 72.724 \\ 16.4884 \end{bmatrix}$$

$$y = 72.724 + 16.488x$$

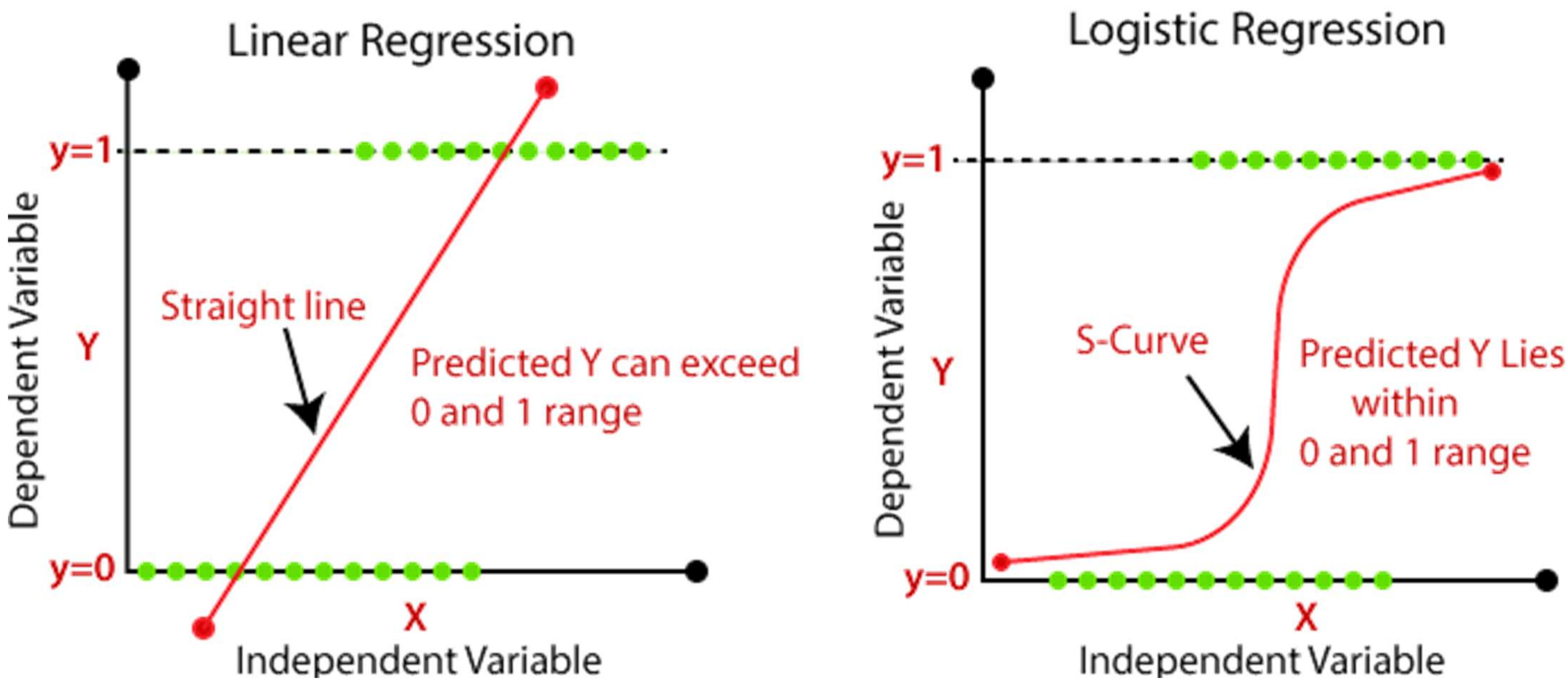
INTRODUCTION TO BUSINESS ANALYTICS

Module 2

Logistic Regression

Logistic Regression

- Linear Regression is used to handle regression problems whereas Logistic regression is used to handle the classification problems.
- Linear regression provides a continuous output but Logistic regression provides logical output (Binary output)

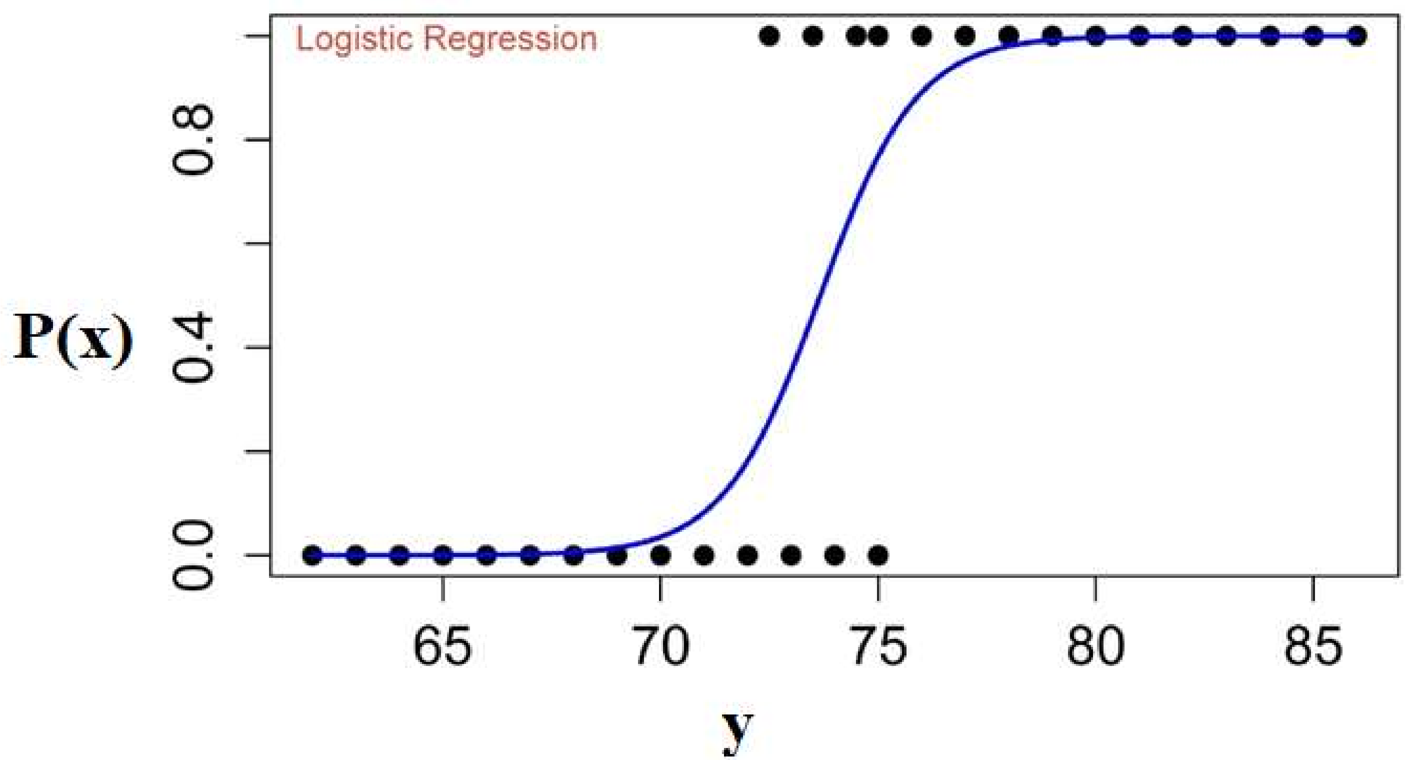


Sigmoid function in Logistics Regression

Sigmoid function is used in Logistic Regression

If P is the probability of success, then $\mathrm{P(x)} = \dfrac{1}{1+e^{-y}}$

Where $\mathbf{y = \beta_0 + \beta_1 x}$



Other Formulas used in Logistic Regression

P = Probability of Success, then

$$\ln\left(\frac{P}{1-P}\right) = y$$

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

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x$$

ln = natural logarithm of a number

Sigmoid function $$P(x) = \frac{1}{1+e^{-y}}$$

Problem

The data set of probability of pass of 5 students in a exam and the number of hours of study are given in the table below.

No of hours of study	Probability of Pass (%)
29	0.248
27	0.00454
33	88.08
28	0.0336
36	99.966

a) Apply logistic regression and calculate probability of pass for the student who studied 34 hours.

b) At least, how many hours a student should study so that he pass the course with the probability of more than 75 percent?

Solution

The data set is

No of hours of study	Probability of Pass (%)
29	0.248
27	0.00454
33	88.08
28	0.0336
36	99.966

No of hours of study	Probability of Pass (%)	Actual Probability of Pass	Binary value (Pass/Fail)
29	0.248	0.00248	0
27	0.00454	0.0000454	0
33	88.08	0.8808	1
28	0.0336	0.000336	0
36	99.966	0.99966	1

Solution – Contd….

No of hours of study (x)	Actual Probability of Pass (P)	$\ln\left(\frac{P}{1-P}\right) = y$
29	0.00248	-6
27	0.0000454	-10
33	0.8808	2
28	0.000336	-8
36	0.99966	8

Now apply simple linear regression formulas for $y = \beta_0 + \beta_1 x$

$$\beta = (X'X)^{-1} X'y$$

Using the matrix approach we get $\beta = \begin{bmatrix} -64.29 \\ 2.0124 \end{bmatrix}$

The estimated fitted line is **y = -64.29 + 2.0124 x**

Logistic Regression

$$p = \frac{1}{1+e^{-y}}$$

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

$$\ln\left(\frac{p}{1-p}\right) = y = \beta_0 + \beta_1 x$$

No: of hours of study	Probability of pass %	$y = \ln\left(\frac{p}{1-p}\right)$
29	0.248	−6
27	0.00454	−10
33	88.08	2
28	0.0336	−8
36	99.966	+8

for Actual probability divide by 100

$$X = \begin{bmatrix} 1 & \cancel{26}\ 29 \\ 1 & 27 \\ 1 & 33 \\ 1 & 28 \\ 1 & 36 \\ \Sigma & \end{bmatrix} \qquad y = \begin{bmatrix} -6 \\ -10 \\ 2 \\ -8 \\ 8 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 5 & 153 \\ 153 & 4739 \end{bmatrix} \qquad X'y = \begin{bmatrix} -14 \\ -314 \end{bmatrix}$$

$$\begin{bmatrix} 4739/286 & -153/286 \\ -153/286 & 5/286 \end{bmatrix}$$

$$X'X^{-1} = \begin{bmatrix} 16.569 & -0.534 \\ -0.534 & 0.017[illegible]4 \end{bmatrix}$$

$$(X'X^{-1})(X'y) = \begin{bmatrix} -64.29 \\ \cancel{1.84284} \\ 2.0124 \end{bmatrix}$$

$\cancel{\beta = (X^t X)^{-1}}$

a) Probability of pass for the student who studied 34 hours -

$$y = \beta_0 + \beta_1 x$$
$$= -64.29 + 2.012(34)$$
$$= 4.118$$

$$P = \frac{1}{1+e^{-y}}$$
$$= \frac{1}{1+e^{-4.118}}$$
$$= \underline{\underline{0.9839}}$$

(98.39% of pass) for student who studied 34 hrs

(b) Atleast how many hours a student should study so that he pass the cours with probability of more than 75%

$x \rightarrow$ no: of hours

$P = 0.75$

$$y = \ln\left(\frac{P}{1-P}\right)$$
$$= \underline{\underline{1.098}}$$

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

$$x = \frac{y - \beta_0}{\beta_1}$$
$$= \frac{1.098 + 64.29}{2.0129}$$
$$= \underline{\underline{32.492}}$$

minimum 32.5 hrs has to study

INTRODUCTION TO BUSINESS ANALYTICS

Module 2

Time Series Analysis

Time Series Analysis

Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time.

In time series analysis, analysts record data points at consistent intervals over a set period of time rather than just recording the data points intermittently or randomly.

It is also called Forecasting

This technique is used in

1) Forecasting stock prices
2) Forecasting Sales
3) Forecasting whether
4) Discovering Seasonal patterns

Time Series – Problem

The stock prices of five days are given in the table. Apply autoregressive method of first order to predict sales for 6^{th} day.

Day	1	2	3	4	5
Sales	5	6	8	9	10

Time Series Analysis

Ex)

day	Stock price
1	5
2	6
3	8
4	9
5	10

y x

$y = f(x)$

Autoregressive method (Order 1)

y → current stock price

x → previous stock price

y	x
5	–
6	5
8	6
9	8
10	9

→

y	x
6	5
8	6
9	8
10	9

Prediction of Stock price for 6th data

$y = \beta_0 + \beta_1 x$

$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ $y = \begin{bmatrix} 6 \\ 8 \\ 9 \\ 10 \end{bmatrix}$ $x = \begin{bmatrix} 1 & 5 \\ 1 & 6 \\ 1 & 8 \\ 1 & 9 \end{bmatrix}$

$x^T x = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 5 & 6 & 8 & 9 \end{bmatrix} \begin{bmatrix} 1 & 5 \\ 1 & 6 \\ 1 & 8 \\ 1 & 9 \end{bmatrix} = \begin{bmatrix} 4 & 28 \\ 28 & 206 \end{bmatrix}$

$x^T y = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 5 & 6 & 8 & 9 \end{bmatrix} \begin{bmatrix} 6 \\ 8 \\ 9 \\ 10 \end{bmatrix} = \begin{bmatrix} 33 \\ 240 \end{bmatrix}$

$\beta = (x^T x)^{-1} (x^T y) = \begin{bmatrix} 5.15 & -0.7 \\ -0.7 & 0.1 \end{bmatrix} \begin{bmatrix} 33 \\ 240 \end{bmatrix} = \begin{bmatrix} 1.95 \\ 0.9 \end{bmatrix}$

$\beta = \begin{bmatrix} 1.95 \\ 0.9 \end{bmatrix}$

$x_6 = 10$

$y = 1.95 + 0.9x$

$y_6 = 1.95 + 0.9 \times 10 = \cancel{2.85}\ \underline{10.95}$