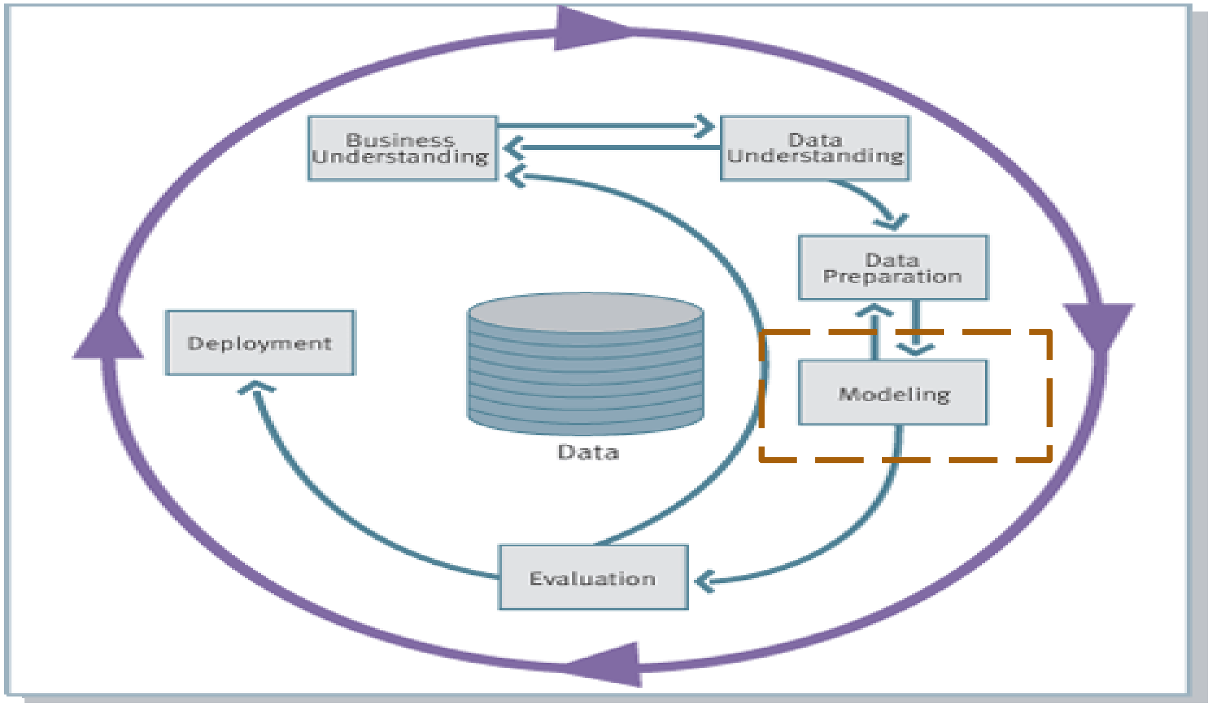
What we have done so far?



Simple Linear Regression Model

Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

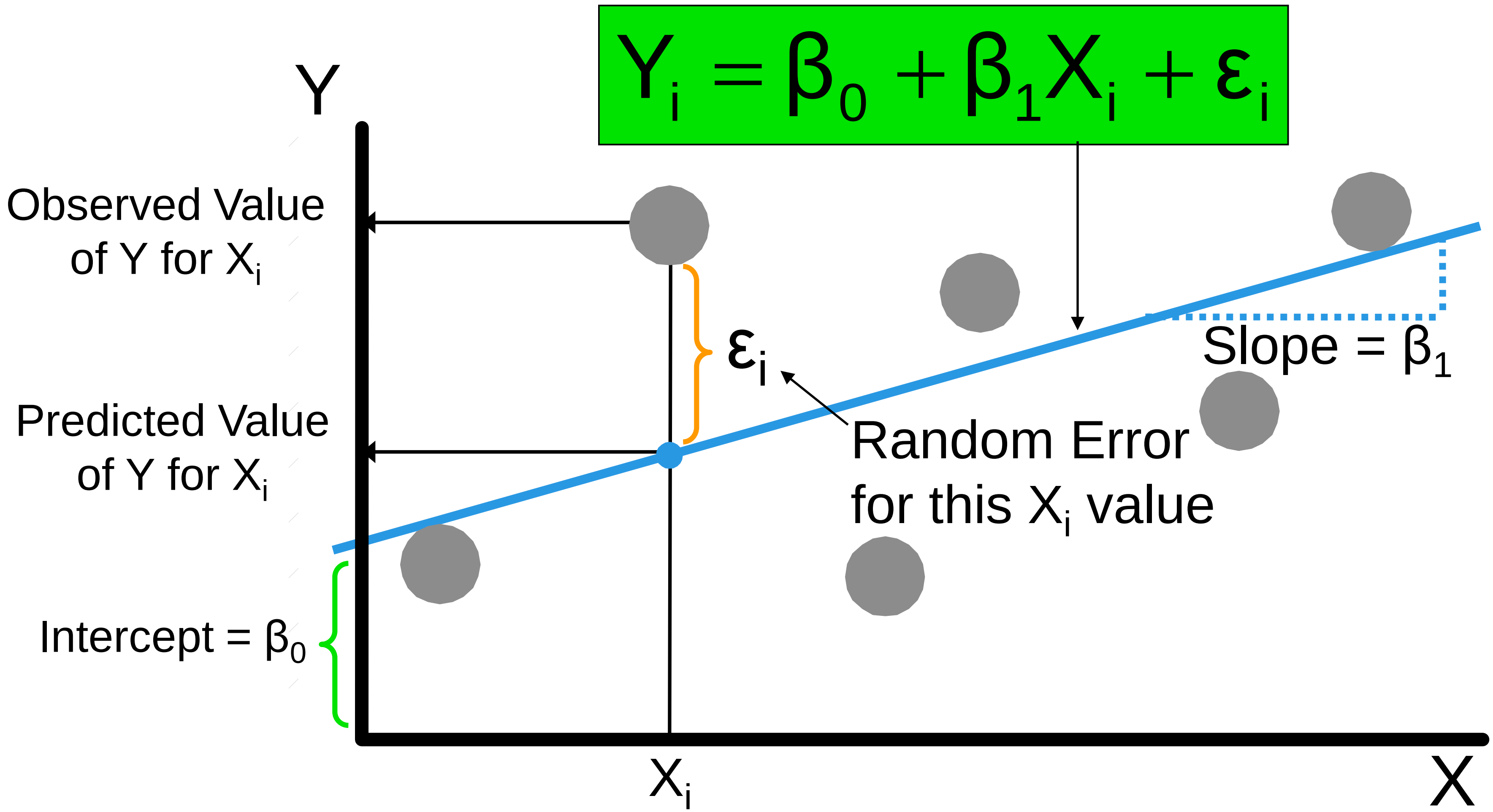
Random Error term

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Linear component

Random Error component

Simple Linear Regression Model



Simple Linear Regression Equation (Prediction Line)

The simple linear regression equation provides an estimate of the population regression line.

Estimated (or predicted) Y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

Value of X for observation i

$$\hat{Y}_i = b_0 + b_1 X_i$$

The Least Squares Method

b_0 and b_1 are obtained by finding the values that minimize the sum of the squared differences between Y and $\hat{Y}$:

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$

Example

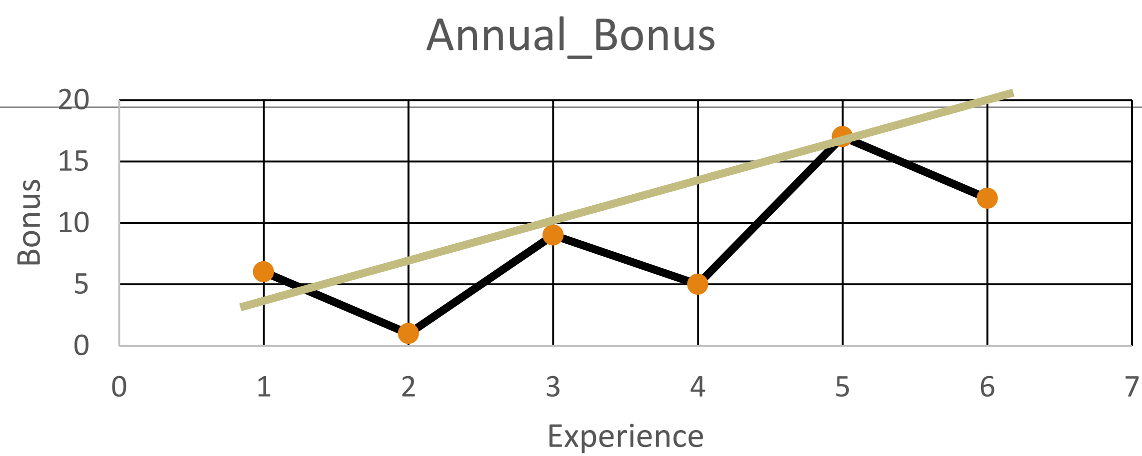
The annual bonuses ($1,000s) of six employees with different years of experience were recorded as follows. We wish to determine the straight line relationship between annual bonus and years of experience.

Years of experience x	1	2	3	4	5	6
Annual bonus y	6	1	9	5	17	12

Interpretation of the Slope and the Intercept

b_0 is the estimated mean value of Y when the value of X is zero.

b_1 is the estimated change in the mean value of Y as a result of a one-unit increase in X.



$$\hat{Y} = 2x + 3$$

X	Y	$\hat{Y}$	$(Y - \hat{Y})$	$(Y - \hat{Y})^2$
1	6	5	1	1
2	1	7	-6	36
3	9	9	0	0
4	5	11	-6	36
5	17	13	4	16
6	12	15	-3	9
				98

$$\hat{Y} = 2.114x + 0.934$$

	Y	$\hat{Y}$	$(Y - \hat{Y})$	$(Y - \hat{Y})^2$
1	6	3.048	2.952	8.714304
2	1	5.162	-4.162	17.32224
3	9	7.276	1.724	2.972176
4	5	9.39	-4.39	19.2721
5	17	11.504	5.496	30.20602
6	12	13.618	-1.618	2.617924
				81.10476

$$SS_{XY} = \sum \left(X - \bar{X}\right)\left(Y - \bar{Y}\right)$$

$$SS_{XX} = \sum \left(X - \bar{X}\right)^2$$

$$b_1 = \frac{SS_{XY}}{SS_{XX}}$$

$$b_0 = \bar{Y} - b_1 \bar{X} = \frac{\sum Y}{n} - b_1 \frac{\sum X}{n}$$

$$SS_{XY} = \sum (X - \overline{X})(Y - \overline{Y})$$

$$SS_{XX} = \sum (X - \overline{X})^2$$

$$b_1 = \frac{SS_{XY}}{SS_{XX}}$$

$$b_0 = \overline{Y} - b_1 \overline{X} = \frac{\sum Y}{n} - b_1 \frac{\sum X}{n}$$

	X	Y	$(X - \overline{X})$	$(Y - \overline{Y})$	$\sum (X - \overline{X})(Y - \overline{Y})$	$\sum (X - \overline{X})^2$
	1.00	6.00	-2.50	-2.33	5.83	6.25
	2.00	1.00	-1.50	-7.33	11.00	2.25
	3.00	9.00	-0.50	0.67	-0.33	0.25
	4.00	5.00	0.50	-3.33	-1.67	0.25
	5.00	17.00	1.50	8.67	13.00	2.25
	6.00	12.00	2.50	3.67	9.17	6.25
Total	21.00	50.00			37.00	17.50

$$b_1 = \frac{37}{17.5} = 2.114$$

$$b_0 = \frac{50}{6} - 2.114 * \frac{21}{6} = 0.9343$$

$$\hat{Y} = 2.114x + 0.9343$$

R output

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	6.83086	0.64958	10.52	4.80e-10	***
Prom	1.18101	0.09148	12.91	9.64e-12	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.946 on 22 degrees of freedom

Multiple R-squared: 0.8834, Adjusted R-squared: 0.8781

F-statistic: 166.7 on 1 and 22 DF, p-value: 9.636e-12

Q. 1 Develop a simple linear regression model between (Y) Revenue and (X) Promotion. What is the expected change in the revenue for every one unit increase in promotion?

Example

Data:

Revenue generated (in million of rupees) from a product; Promotion Expenses (in million of rupees)

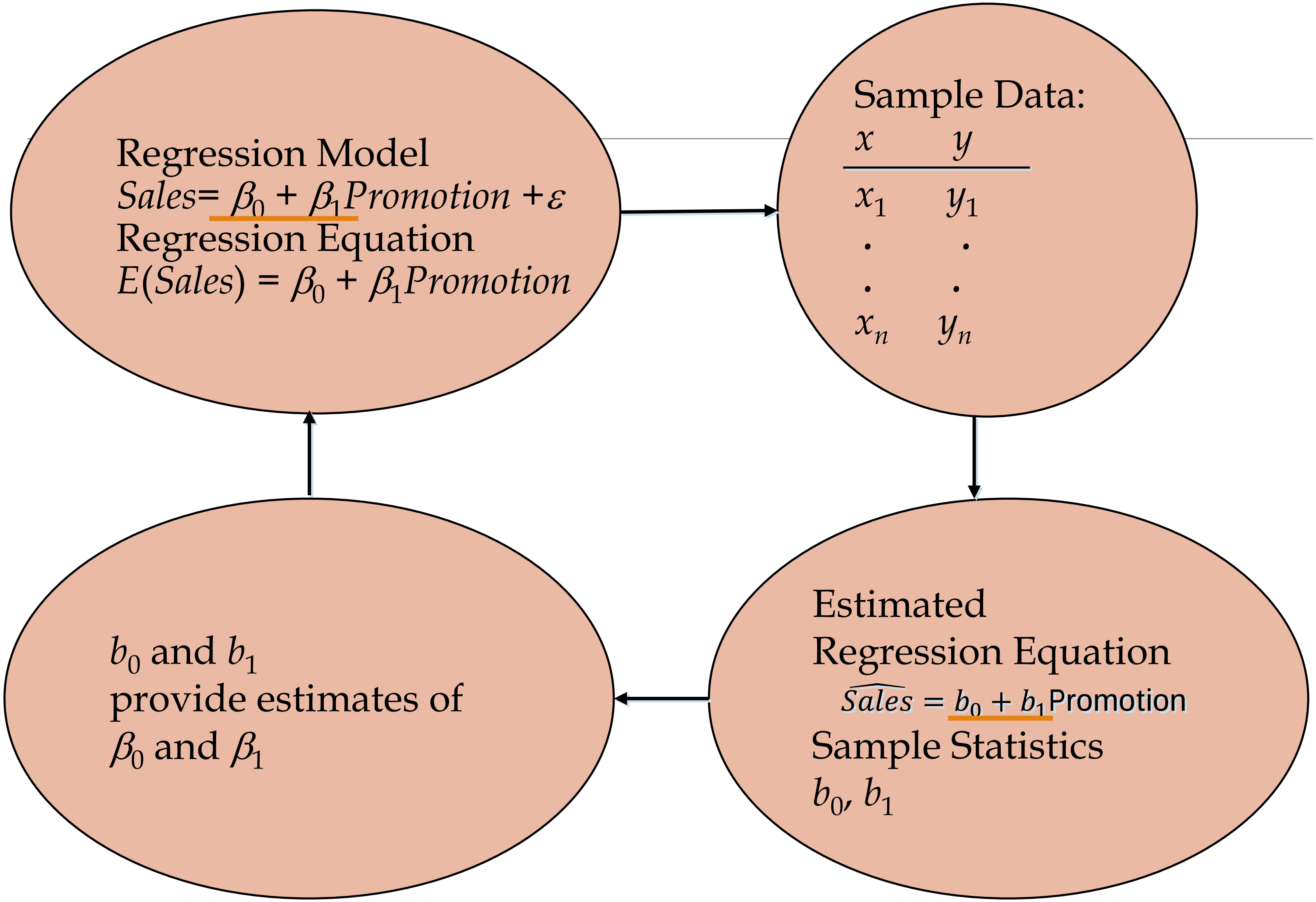
S.No.	Rev	Prom	S.No.	Rev	Prom
1	5	1	13	16	7
2	6	1.8	14	17	8.1
3	6.5	1.6	15	18	8
4	7	1.7	16	18	10
5	7.5	2	17	18.5	8
6	8	2	18	21	12.7
7	10	2.3	19	20	12
8	10.8	2.8	20	22	15
9	12	3.5	21	23	14.4
10	13	3.3	22	7.1	1
11	15.5	4.8	23	10.5	2.1
12	15	5	24	15.8	4.75

A company wants to estimate how Sales is impacted by Promotion expense.

Collect some data for promotional expense and corresponding sales.

Estimate the relationship between Sales and Promotional Expense.

Estimation Process



Develop a simple linear regression model between (Y) Revenue and (X) Promotion. What is the expected change in the revenue for every one unit increase in promotion?

bo

b1

$$\hat{y} = b_0 + b_1 x$$

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	6.83086	0.64958	10.52	4.80e-10	***
Prom	1.18101	0.09148	12.91	9.64e-12	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.946 on 22 degrees of freedom

Multiple R-squared: 0.8834, Adjusted R-squared: 0.8781

F-statistic: 166.7 on 1 and 22 DF, p-value: 9.636e-12

b0 = Expected sales when Promotion expense is zero.

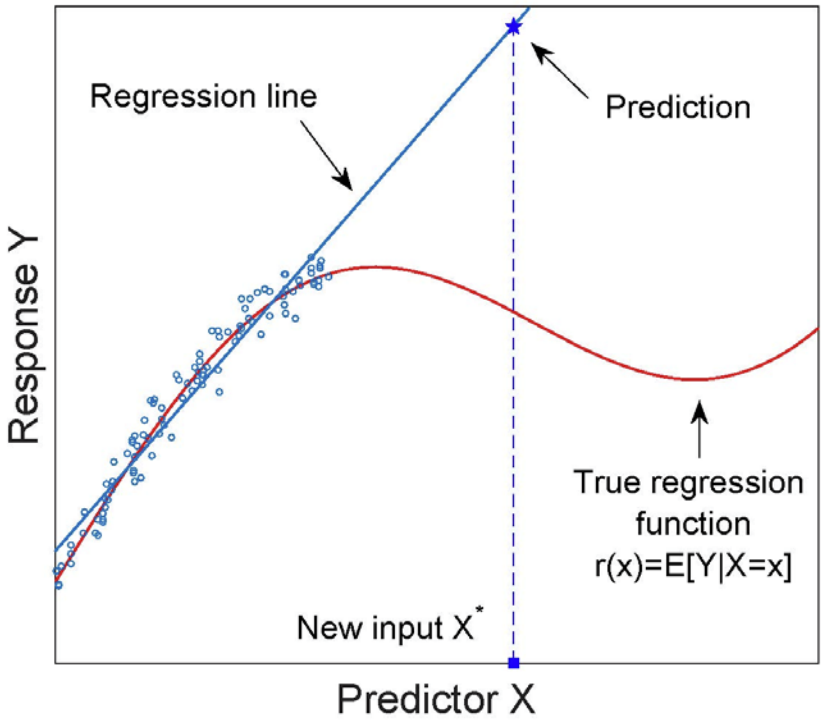
b1 = Expected change in the revenue for every one unit increase in promotion.

What is the expected value of Rev at Prom =5?

Ans: 12.735

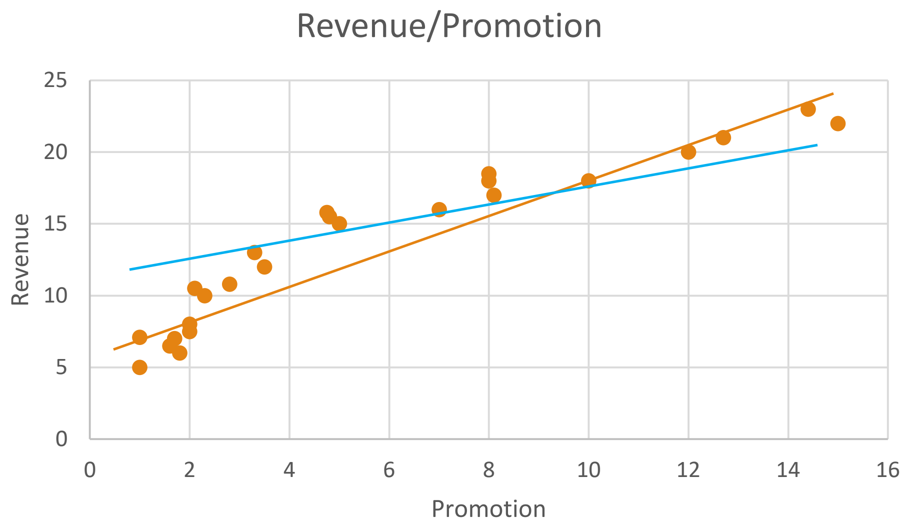
Expected Sales =
6.83+1.81*Promotion

Danger of Extrapolation



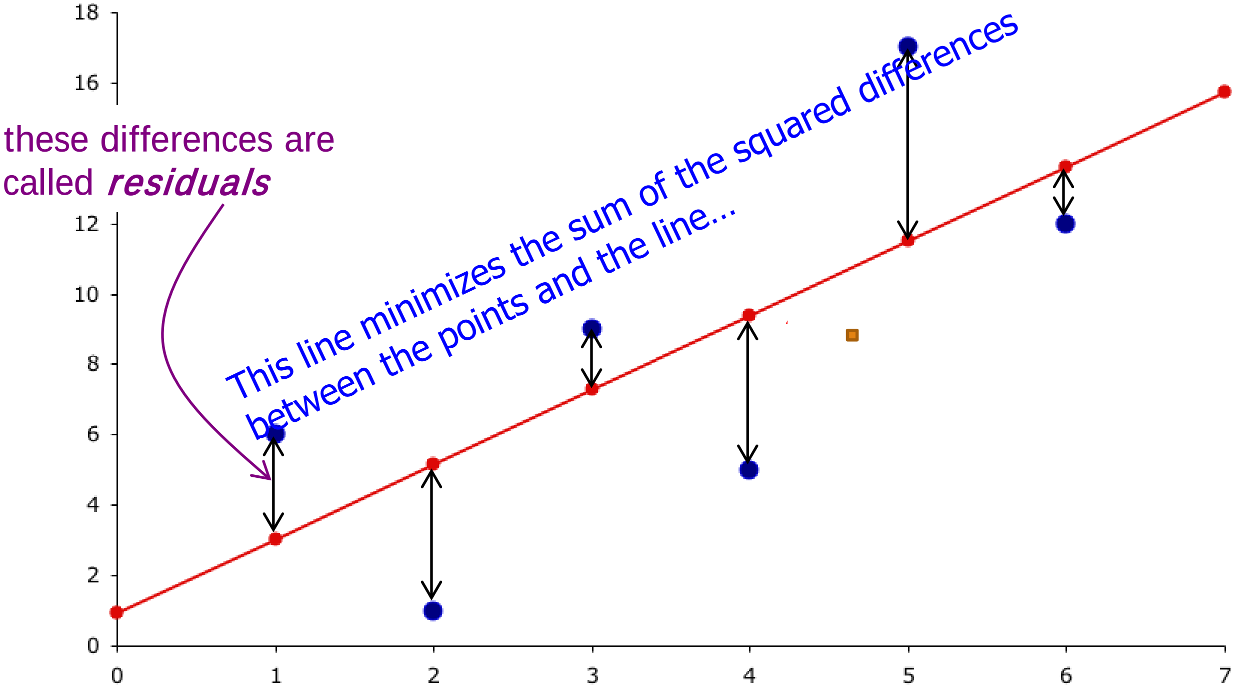
Range of Promotion Variable is [1,15].
Perdiction for values outside this range could be misleading.

Statistical v/s Mathematical Relationship



Revenue = 6.83+1.181*Promotion / Revenue = 6.83+1.181*Promotion+ε

Best Fit Line



What proportion of the variation in Rev is explained by Prom?

How well model fits the data?

- R-squared (R^2)

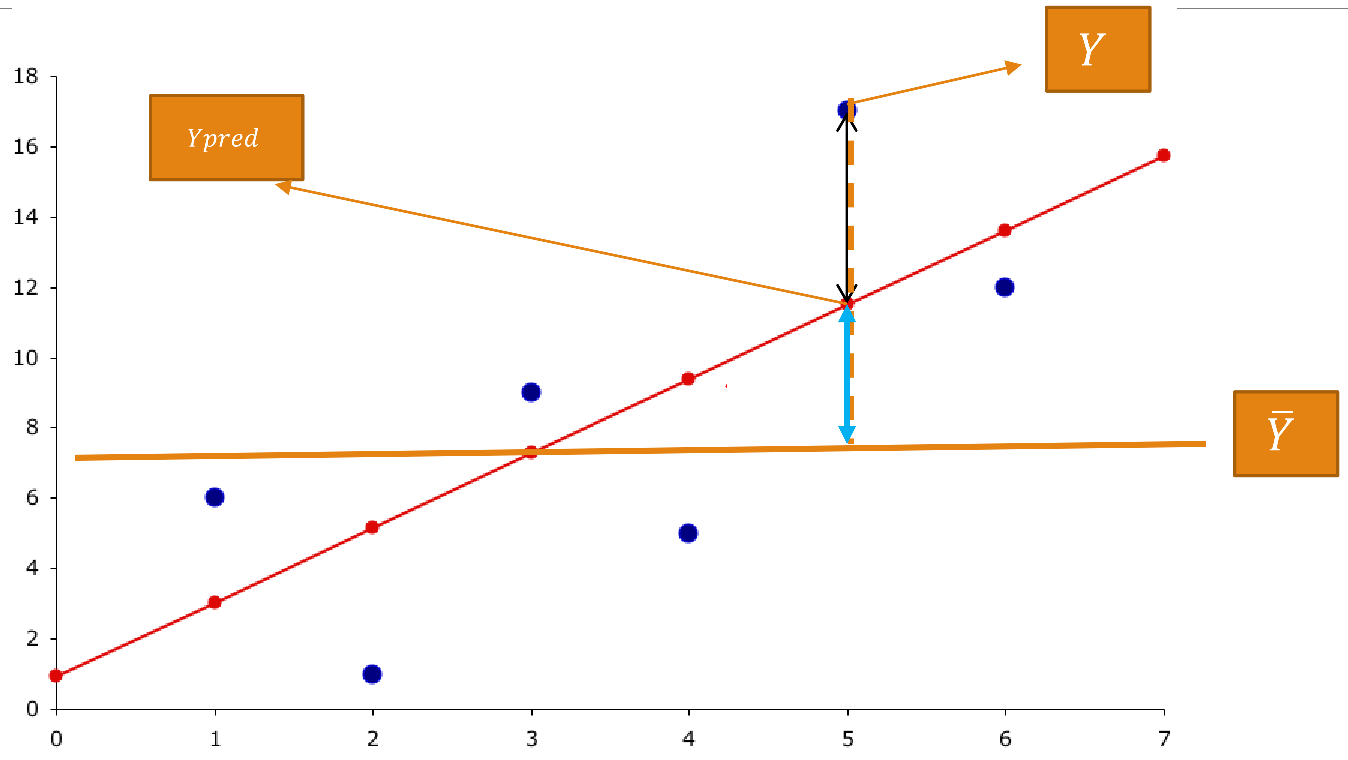
  R^2 is proportion of variability in response variable explained

  by regression model

- SEE (Standard Error of the Estimate)

  Average difference between the predicted response and the actual value.

R-squared



R-squared

Sum of Squares Total $SST = \sum_{i=1}^{n} \left(y - \overline{y}\right)^2$

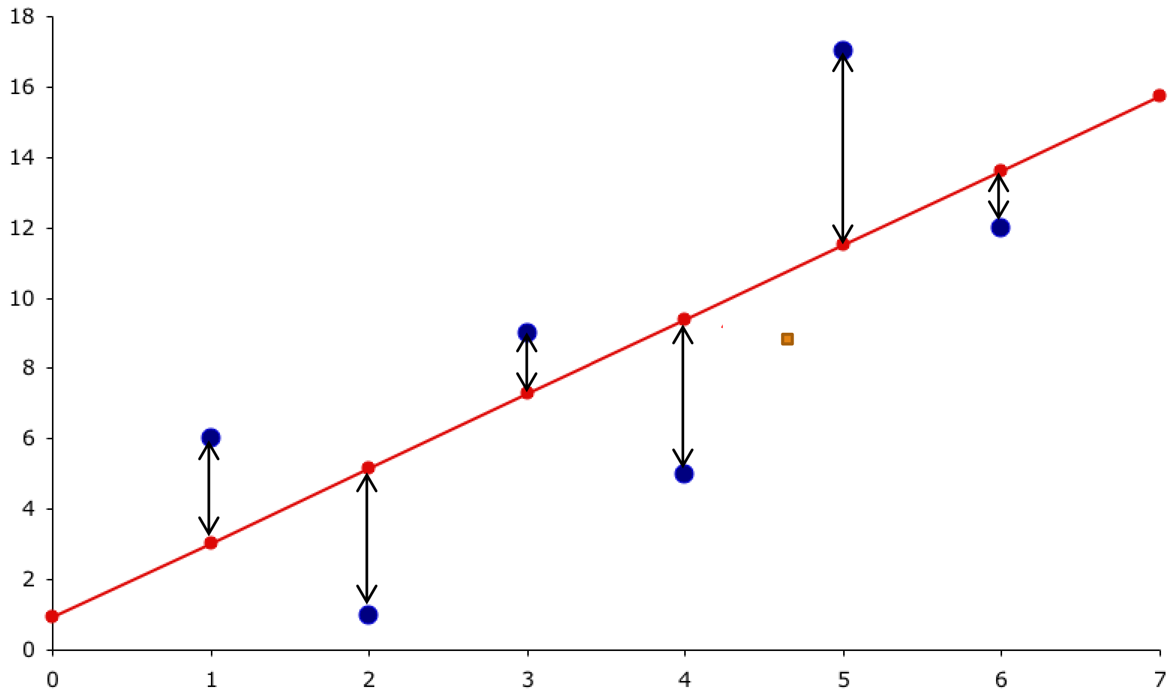
Sum of Squares Error SSE = $\sum(y - \hat{y})^2$

Sum of Squares Regression, $SSR = \sum_{i=1}^{n} \left(\hat{y} - \overline{y}\right)^2$

SST = SSR + SSE

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

SEE



Measures Accuracy of predictions

$$s = \sqrt{SSE/(n - m - 1)}$$

Average difference between the predicted response and the actual value.

What proportion of the variation in Rev is explained by Prom?

```
Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)  6.83086    0.64958   10.52   4.80e-10 ***
Prom         1.18101    0.09148   12.91   9.64e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.946 on 22 degrees of freedom
Multiple R-squared:  0.8834,        Adjusted R-squared:  0.8781
F-statistic: 166.7 on 1 and 22 DF,  p-value: 9.636e-12
```

Standard Error of Estimate

R-squared