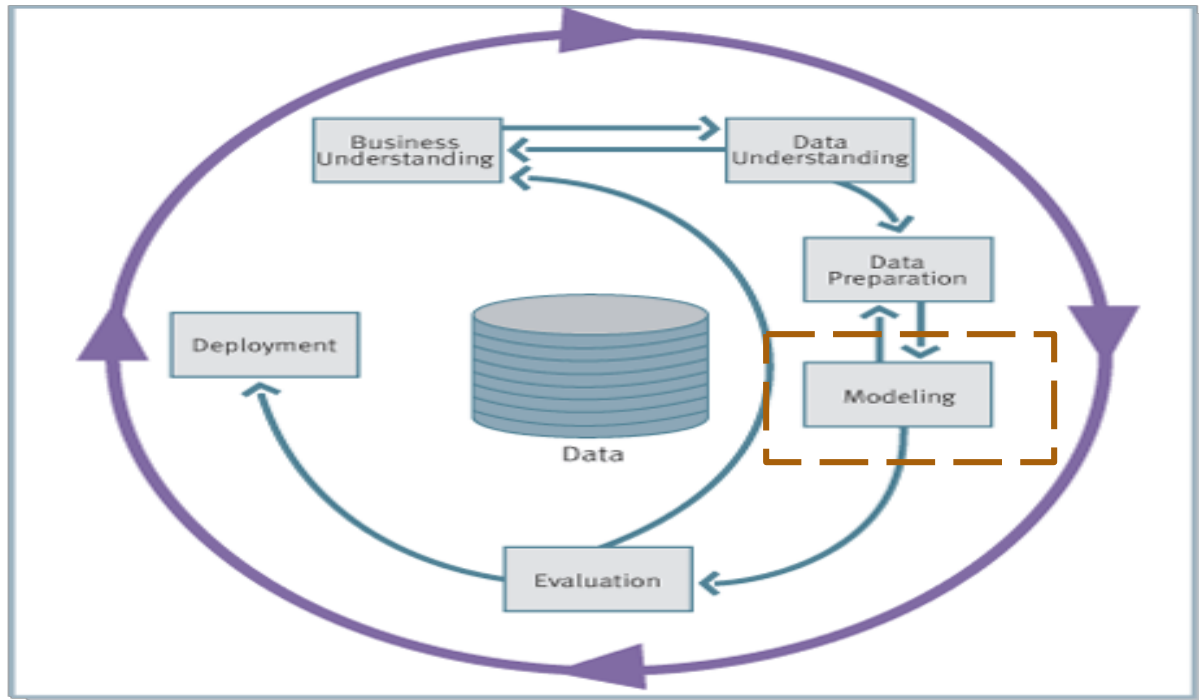


What we have done so far?

---



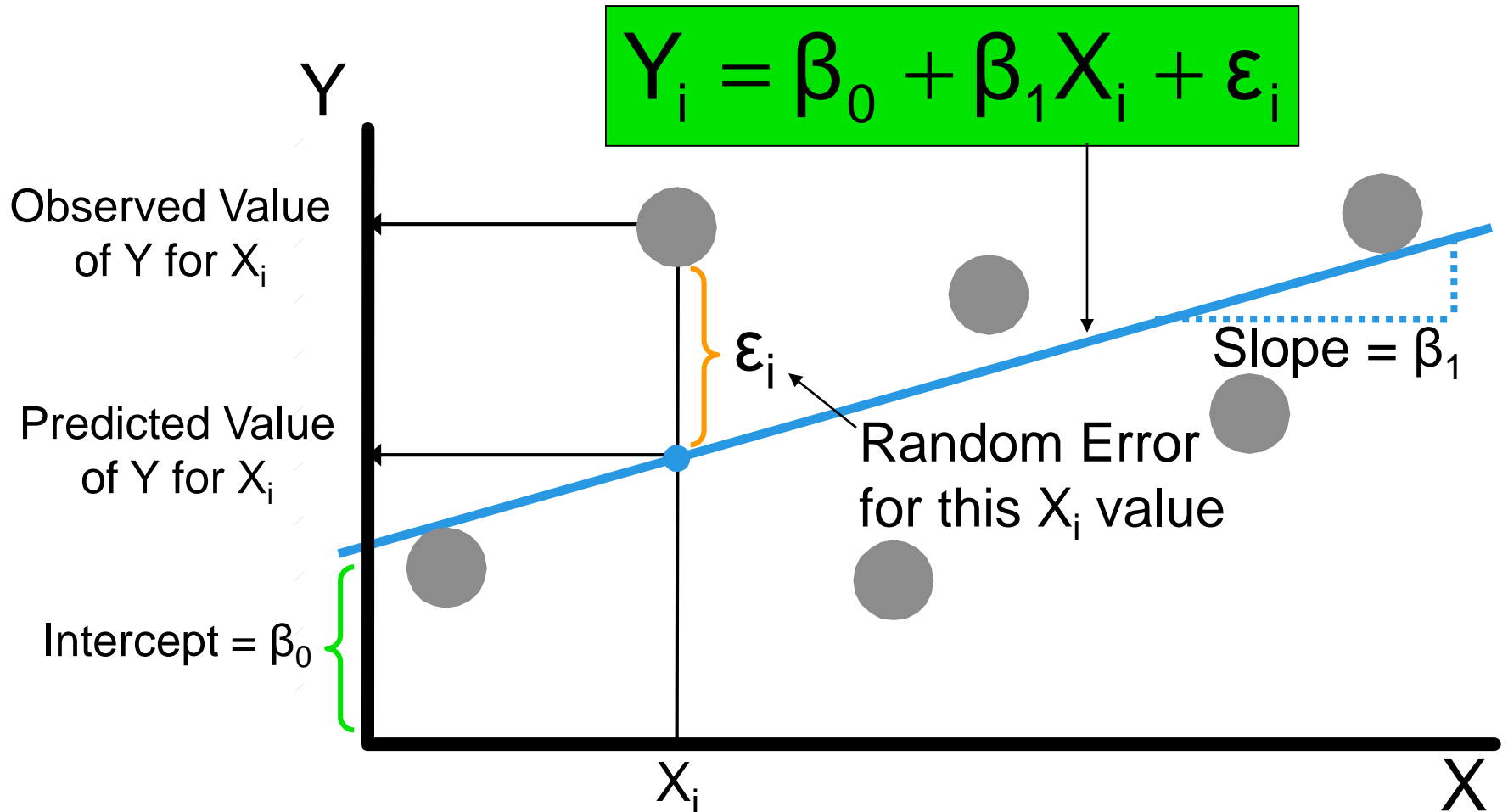
# Simple Linear Regression Model

The diagram illustrates the Simple Linear Regression Model equation,  $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ , with labels and annotations:

- Dependent Variable:**  $Y_i$
- Population Y intercept:**  $\beta_0$
- Population Slope Coefficient:**  $\beta_1$
- Independent Variable:**  $X_i$
- Random Error term:**  $\epsilon_i$

The equation is presented in a green box. Brackets below the equation identify the **Linear component** ( $\beta_0 + \beta_1 X_i$ ) and the **Random Error component** ( $\epsilon_i$ ).

# 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.

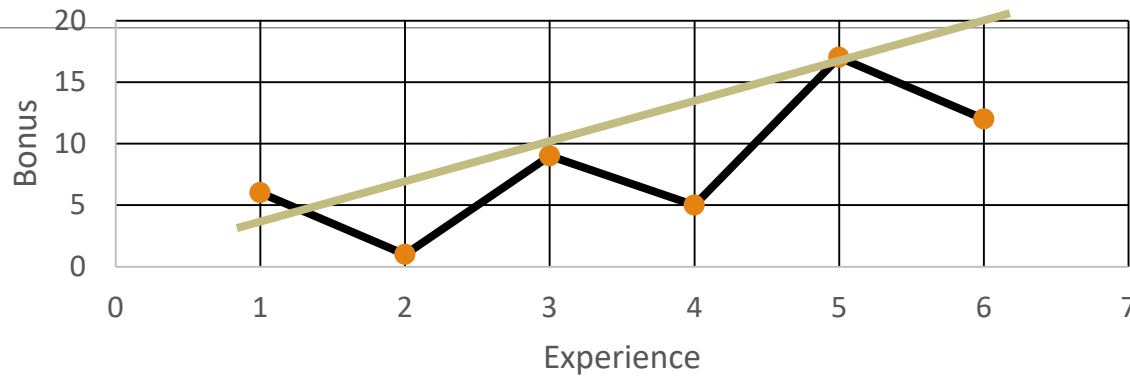
<u>Years of experience <math>x</math></u>	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>
Annual bonus $y$	6	1	9	5	17	12

# 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  $X$ .

Annual\_Bonus



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

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

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

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

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

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

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

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

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

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

	X	Y	(X - $\bar{X}$ )	(Y - $\bar{Y}$ )	$\sum (X - \bar{X})(Y - \bar{Y})$	$\sum (X - \bar{X})^2$
	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

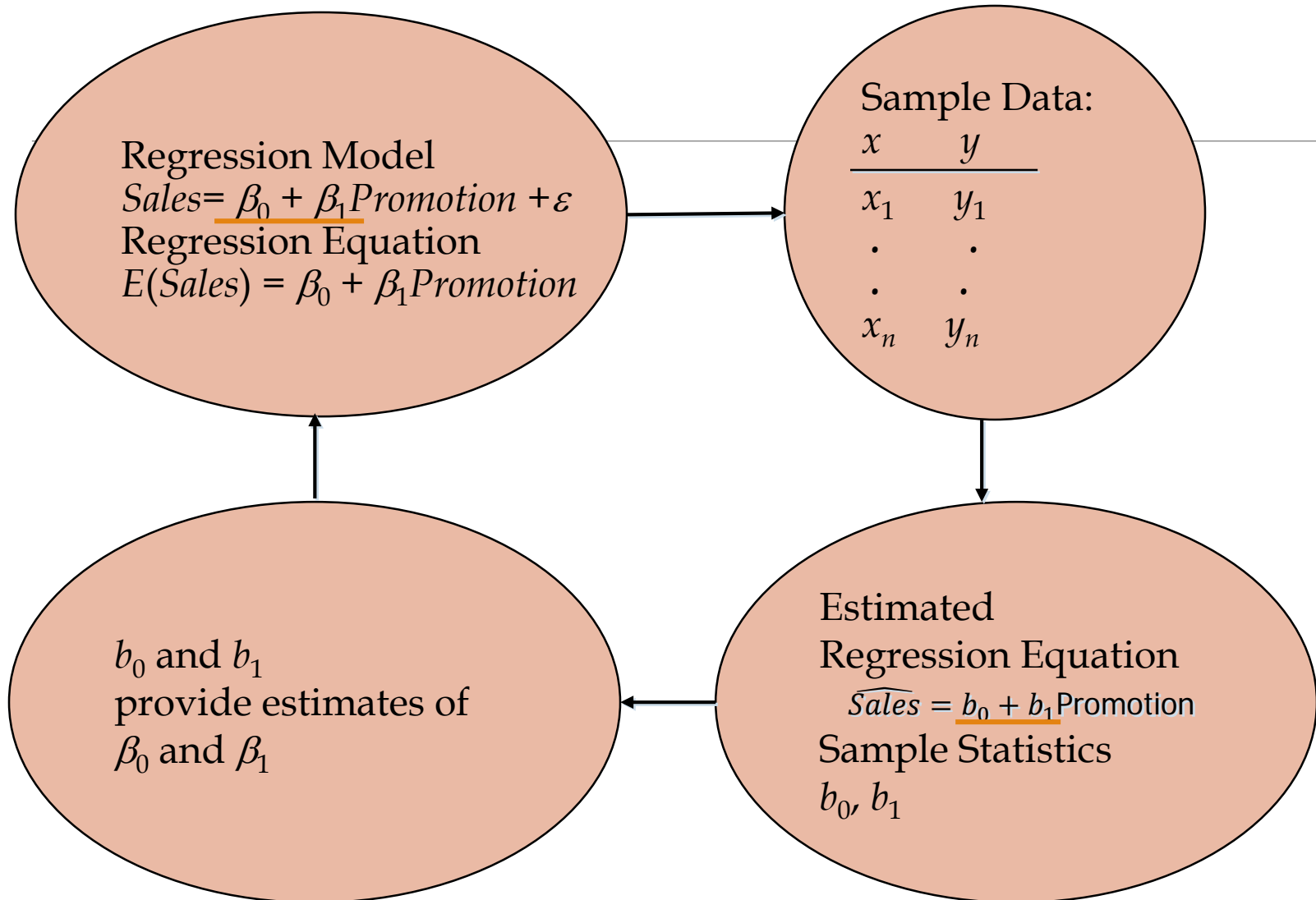
---

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?



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

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

**b0 = Expected sales when Promotion expense is zero.**

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

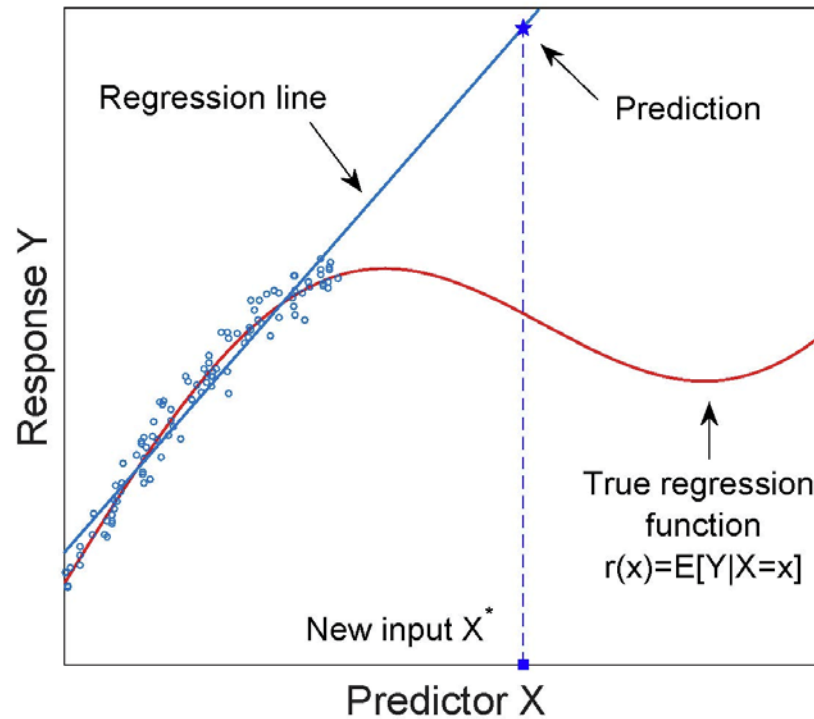
---

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

Ans: 12.735

$$\text{Expected Sales} = 6.83 + 1.81 * \text{Promotion}$$

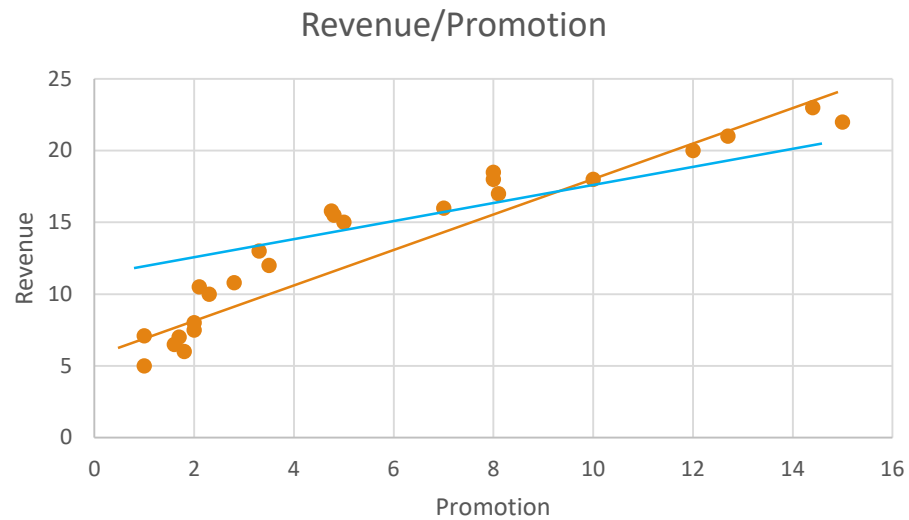
# Danger of Extrapolation



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

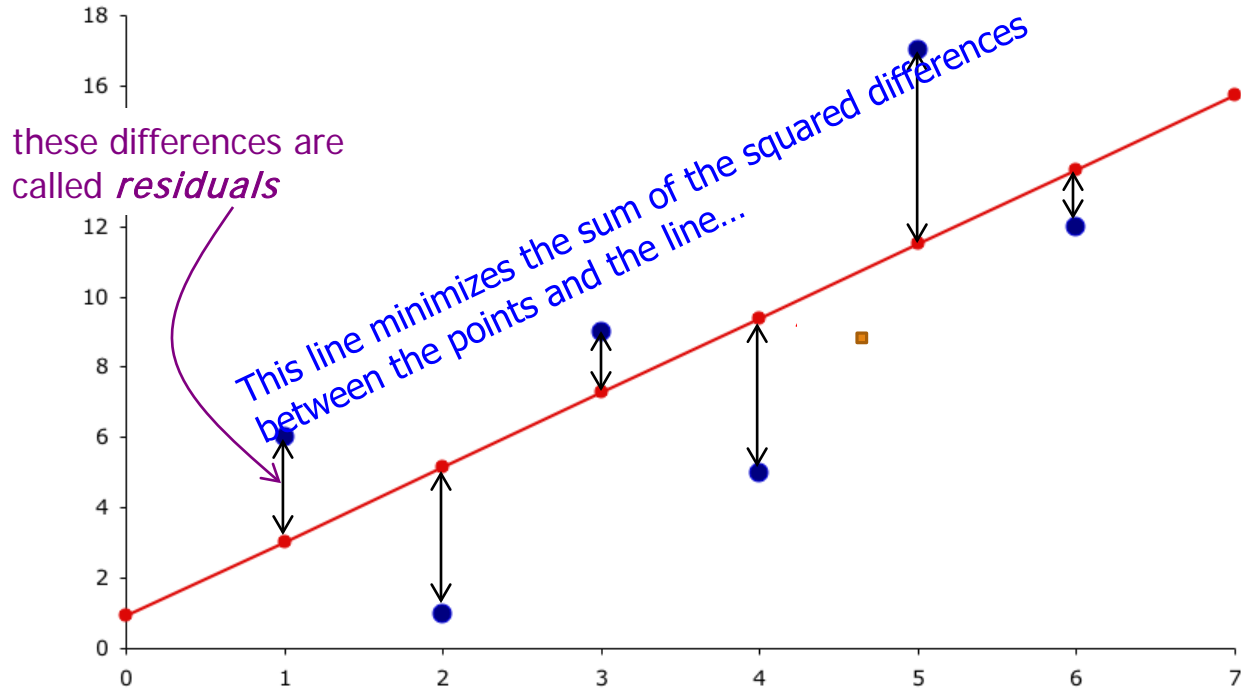
# Statistical v/s Mathematical Relationship

---



$$\text{Revenue} = 6.83 + 1.181 * \text{Promotion} / \text{Revenue} = 6.83 + 1.181 * \text{Promotion} + \epsilon$$

# Best Fit Line



---

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

# How well model fits the data?

---

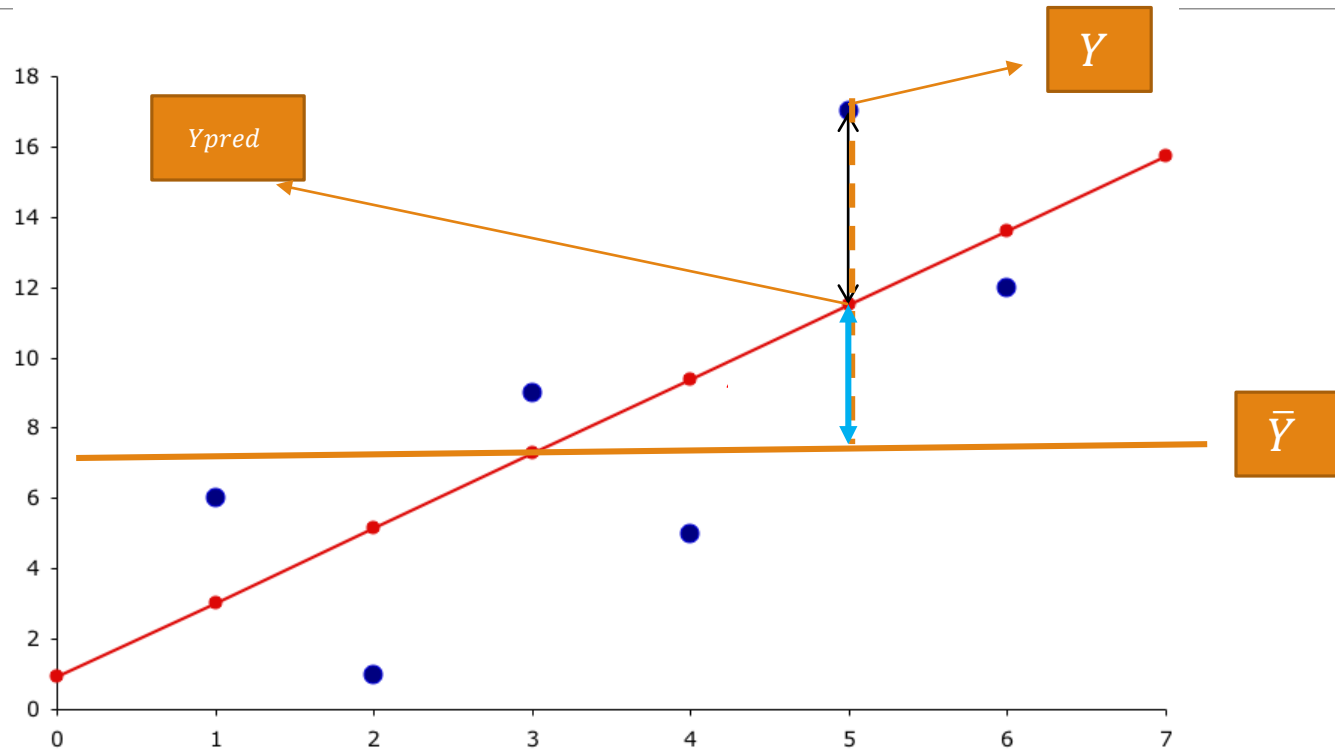
- R-squared ( $R^2$ )

$R^2$  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  $SST = \sum_{i=1}^n (y - \bar{y})^2$

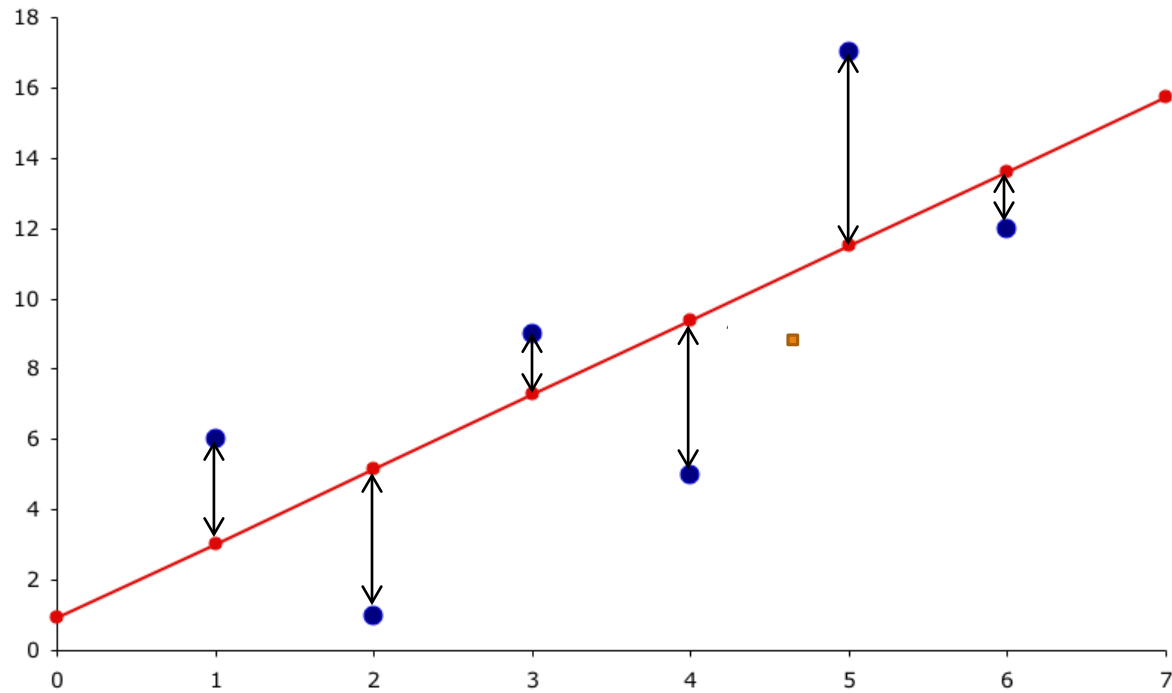
Sum of Squares Error  $SSE = \sum (y - \hat{y})^2$

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

$$SST = SSR + SSE$$

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

# 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

---

Is there a statistically significant relationship between Rev and Prom at  $\alpha = 0.1$ ?

# T-test for Relationship Between x and y (*cont'd*)

---

Ho: asserts  $\beta_1 = 0$  (no linear relationship exists)

Ha: asserts  $\beta_1 \neq 0$  (linear relationship exists)

T-test based on t-distribution with n-2 degrees of freedom:

$$t = \frac{(b_1 - \beta_1)}{s_{b_1}}$$

$$s_{b_1} = \frac{s}{\sqrt{\sum x^2 - (\sum x)^2 / n}}$$

When null hypothesis true  $t = b_1 / S_{b_1}$  follows t-distribution with n-2 df.

Large values  $S_{b_1}$  indicate estimate of slope  $b_1$  *unstable*

Small values  $S_{b_1}$  indicate estimate of slope  $b_1$  *precise*

100(1-alpha)% confidence interval for slope  $\beta_1$ ,  
where  $t_{n-2}$  based on 2 degrees of freedom:

$$b_1 \pm (t_{n-2})(s_{b_1})$$

---

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

P-value

t stat

# Interval Estimate

---

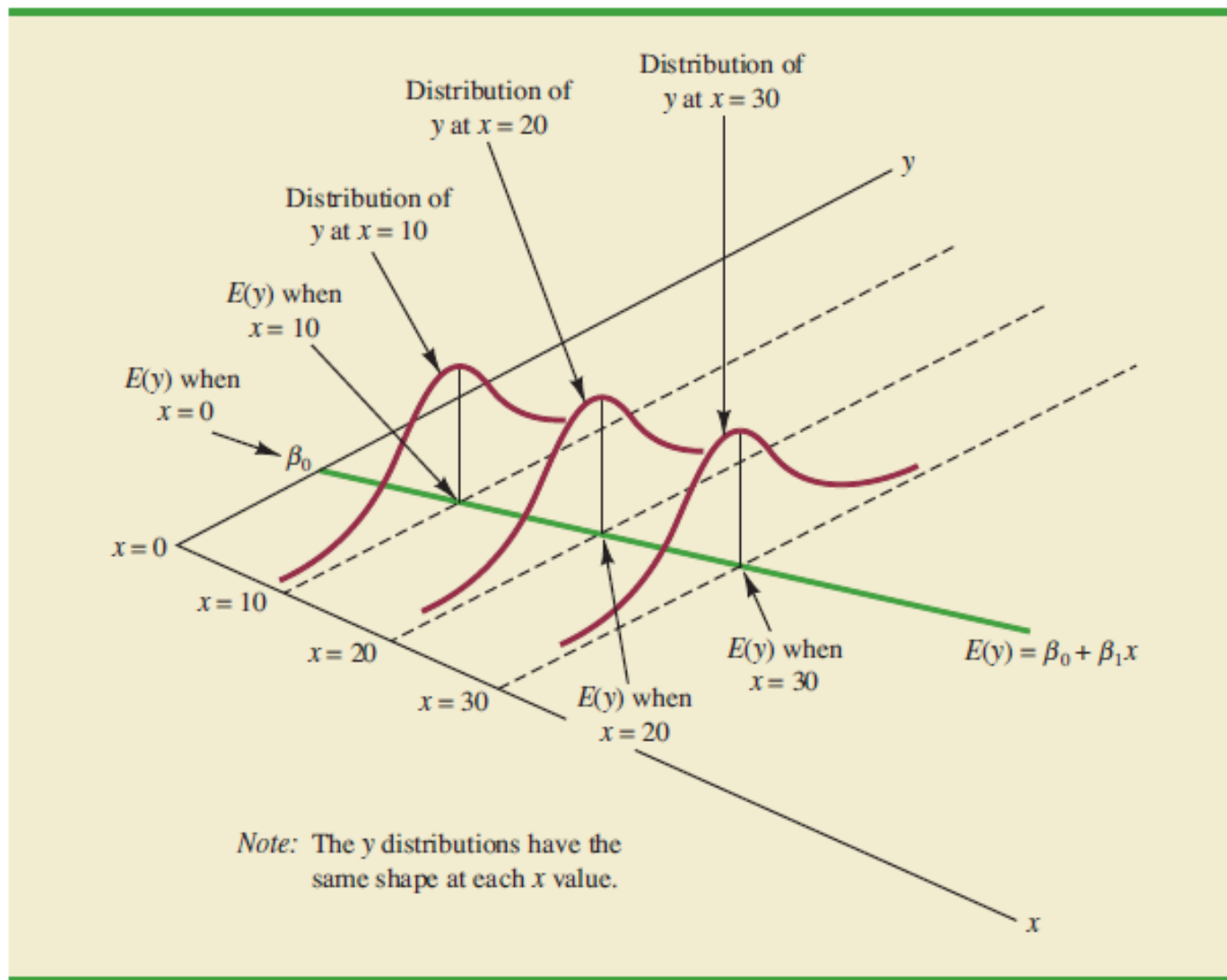
## Prediction Interval of y Given x

$$y_p \pm t_{n-2}(s) \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

What is the value of the rating likely to be of a cereal at the sugar level of 14.

$x_p$	given value of x, for which prediction being made
$y_p$	point estimate of y, for given value of x
$t_{n-2}$	multiplier associated with sample size and confidence level
s	standard error of estimate

FIGURE 14.6 ASSUMPTIONS FOR THE REGRESSION MODEL



# Assumptions

---

L

**Linear Function:** The mean of the response,  $E(Y_i)$ , at each value of the predictor,  $x_i$ , is a linear function of the  $x_i$ .

I

**Independent:** The errors,  $\epsilon_i$ , are Independent. (Not a problem in cross-sectional data)

N

**Normally Distributed:** The errors,  $\epsilon_i$ , at each value of the predictor,  $x_i$ , are Normally distributed.

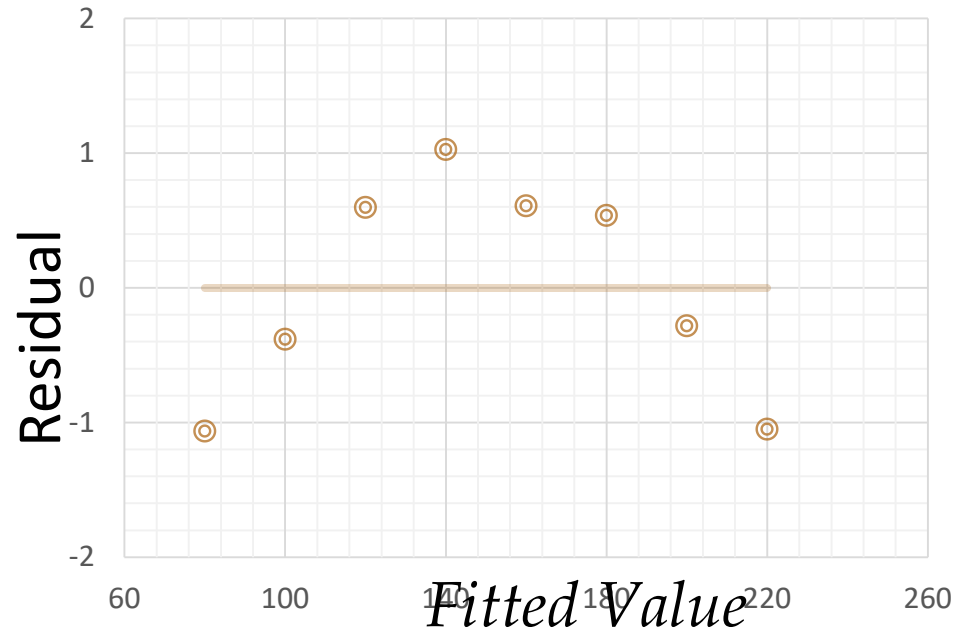
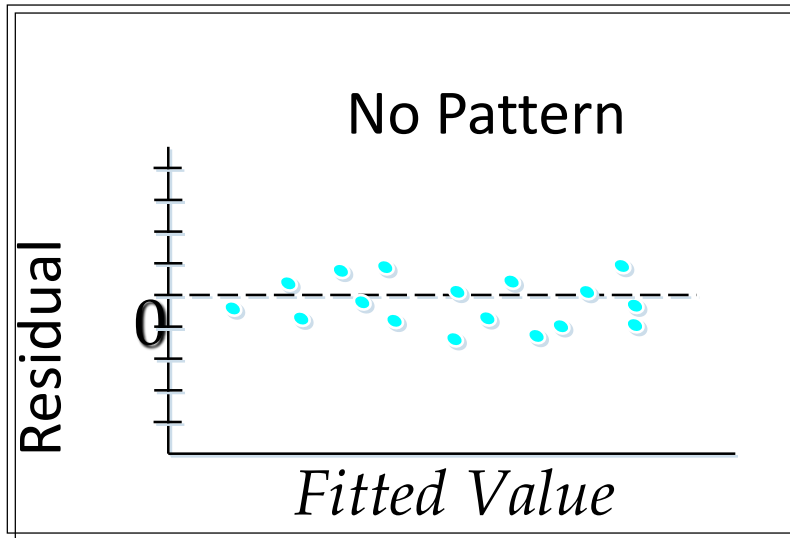
E

**Equal variances (denoted  $\sigma^2$ ):** The errors,  $\epsilon_i$ , at each value of the predictor,  $x_i$ , have Equal variances.

# Detecting Nonlinearity of Regression Function

## Plot Residuals versus Fitted Value

- Random Cloud around 0  $\Rightarrow$  Linear Relation
- U-Shape or Inverted U-Shape  $\Rightarrow$  Nonlinear Relation



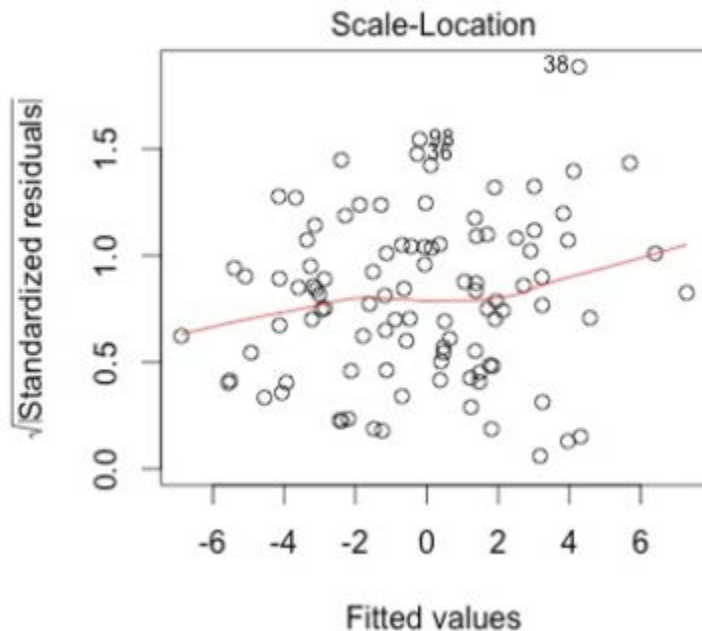
# Constant Variance

Plot Residuals versus Predicted Values

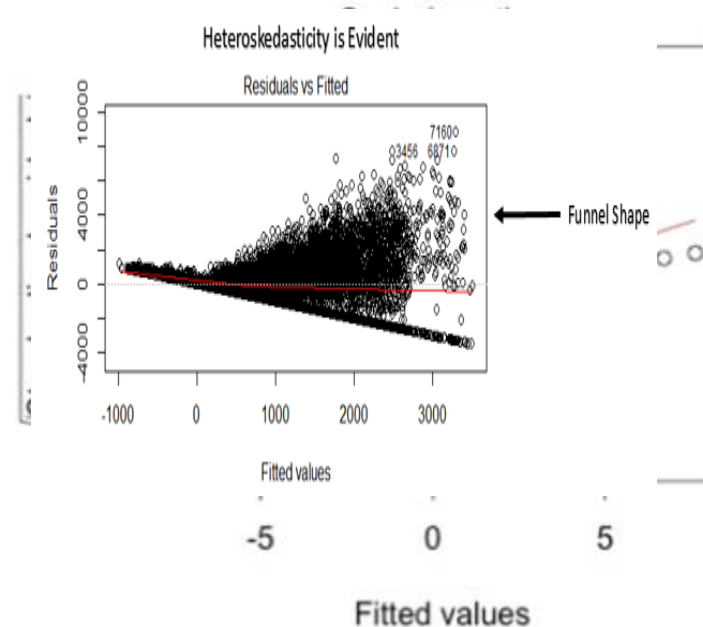
- Random Cloud around 0  $\Rightarrow$  Linear Relation
- Funnel Shape  $\Rightarrow$  Non-constant Variance

Plot absolute Residuals, squared residuals, or square root of absolute residuals

**Homoskedasticity is Evident**



**Heteroskedasticity is Evident**

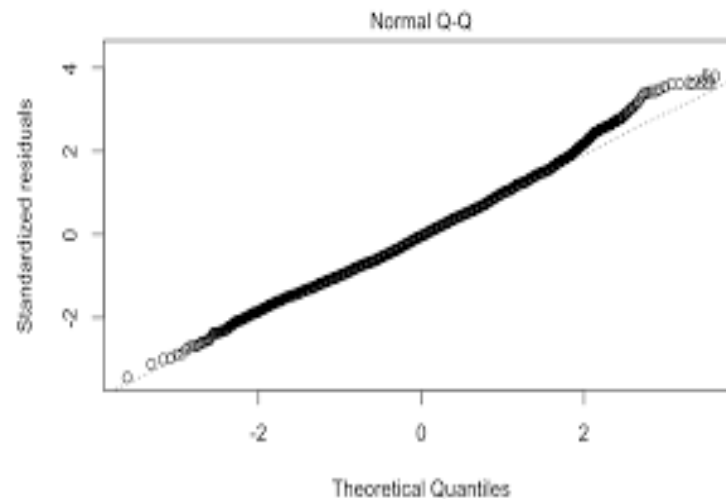


# Normality Assumption

---

## Normal Probability Plot

Quantile-Quantile plot: quantiles of particular distribution against quantiles of standard normal distribution



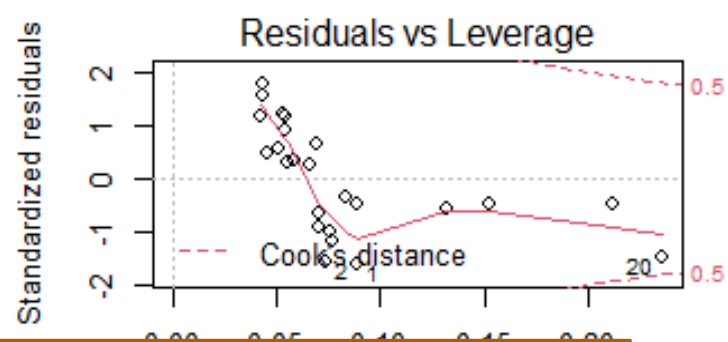
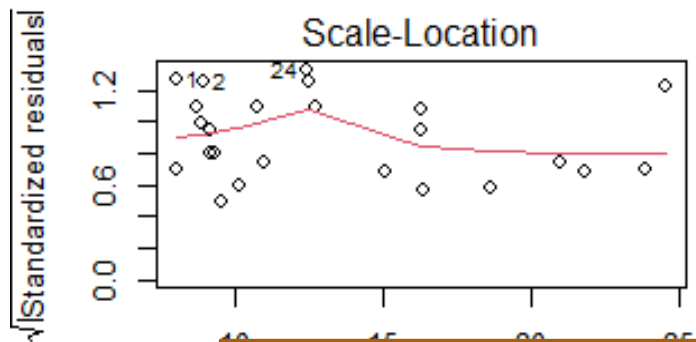
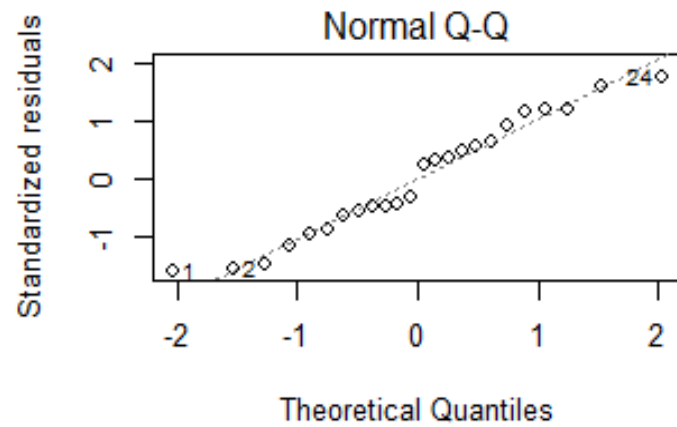
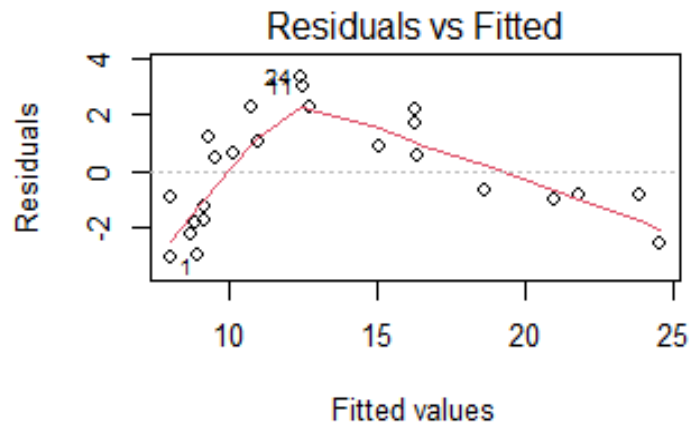
If the points lie majorly on the line, errors are assumed to be normally distributed.

# Remedial Measures

---

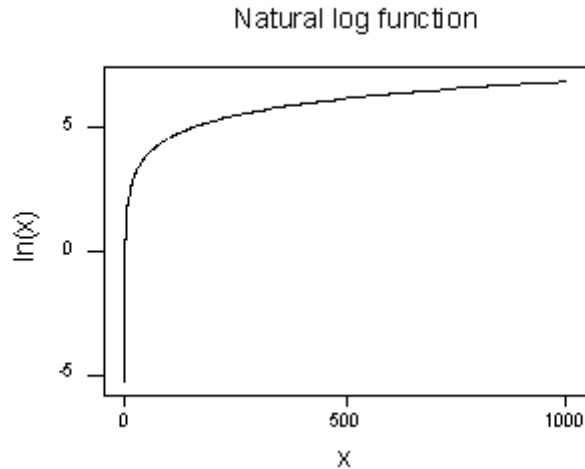
- Nonlinear Relation – Add polynomials, fit exponential regression function, or transform  $X$
- Non-Constant Variance – Weighted Least Squares, transform  $Y$  and/or  $X$ , or fit Generalized Linear Model
- Non-Normality of Errors – Transform  $Y$  or fit Generalized Linear Model
- Non-Independence of Errors – Transform  $Y$  or use Generalized Least Squares

# Sales Vs Promotion Expense



From Graph 1, Only Linearity assumption seems violated  
Remedial: Transform X to LN(X)

# Log-Transformation in X



X	$\text{Log}_{10}(X)$
1	0
10	1
100	2

Interpretation:

In general, a  $k$ -fold increase in the predictor  $x$  is associated with a  $b_1 \times \ln(k)$  change in the mean of the response  $y$ .

---

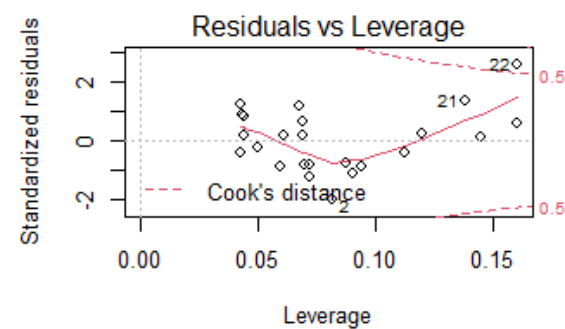
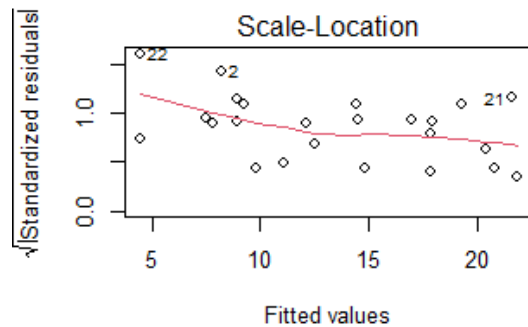
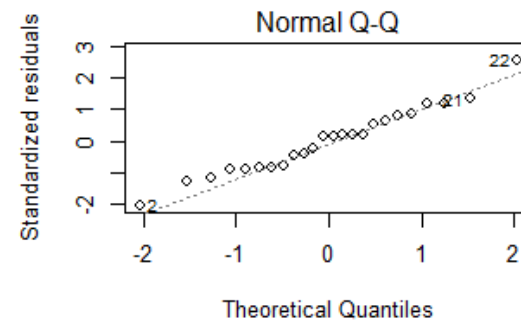
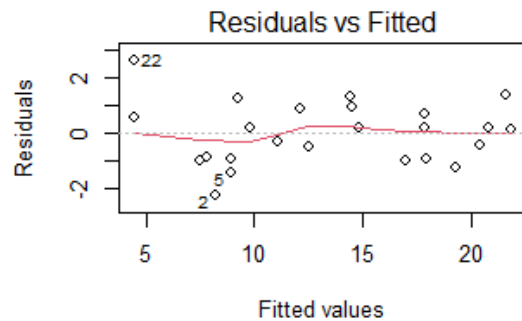
$$\hat{Y} = b_0 + b_1 x \quad \hat{Y}_n = b_0 + b_1(x+1)$$

$$\hat{Y}_n - \hat{Y} = b_1$$

$$\hat{Y} = b_0 + b_1 \ln x \quad \hat{Y}_n = b_0 + b_1 \ln(kx)$$

$$\hat{Y}_n - \hat{Y} = b_1 \ln(k)$$

# Revised Model



Sales =  $b_0 + b_1(\text{Promotion})$   $\longrightarrow$  Sales =  $b_0 + b_1 * \text{LN}(\text{Promotion})$

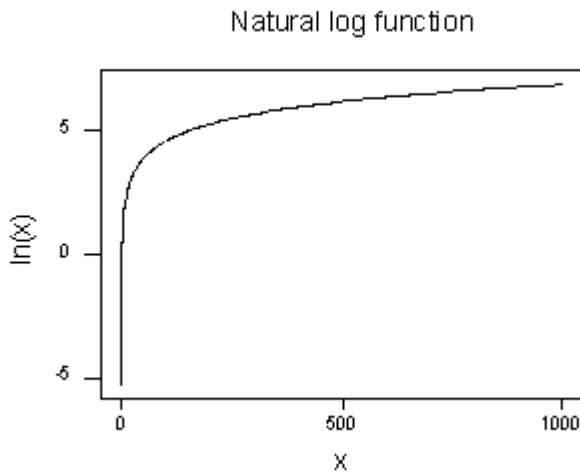
# Model Comparison

---

Model without transformation	Model with Log transformation
R-squared: 0.8834	R-squared: 0.9604
SEE: 1.946	SEE: 1.134

# Log-Transformation in Y

---



Each 1-unit increase in  $X$  multiplies the expected value of  $Y$  by  $e^{b_1}$ .

$$\widehat{Y} = b_0 + b_1x \quad \widehat{Y}_n = b_0 + b_1(x+1)$$

$$\widehat{Y}_n - \widehat{Y} = b_1$$

$$\widehat{\ln Y} = b_0 + b_1x \quad \widehat{\ln Y}_n = b_0 + b_1(x+1)$$

$$\widehat{\ln Y}_n - \widehat{\ln Y} = b_1$$

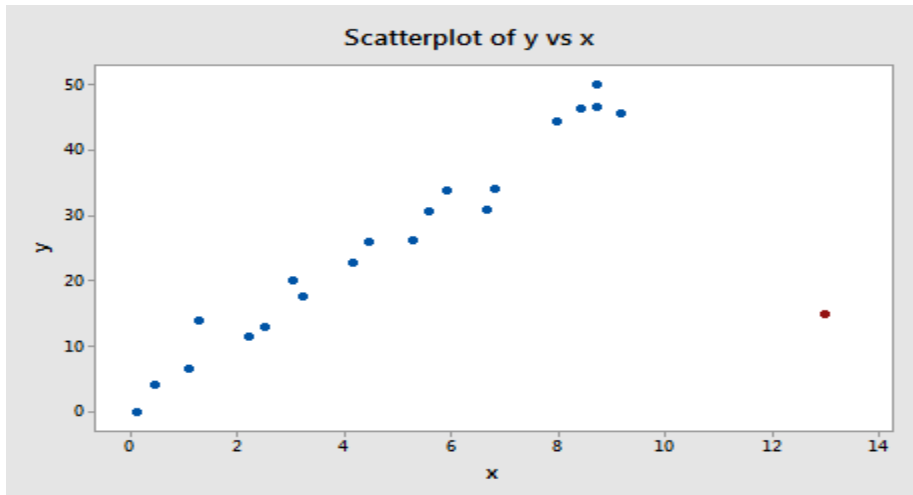
$$\widehat{Y}_n / \widehat{Y} = e^{b_1}$$

## Influential Observations

---

- .
- A data point is **influential** if it unduly influences any part of a regression analysis, such as the predicted responses, the estimated slope coefficients, or the hypothesis test results.
- We would like for a regression model to be representative of all of the sample observations, not an artifact of a few.
- If the influence points are bad values, then they should be eliminated from the sample.
- If they are not bad values, there may be nothing wrong with these points, but if they control key model properties we would like to know it, as it could affect the end use of the regression model.

# Influential points



## Model Summary

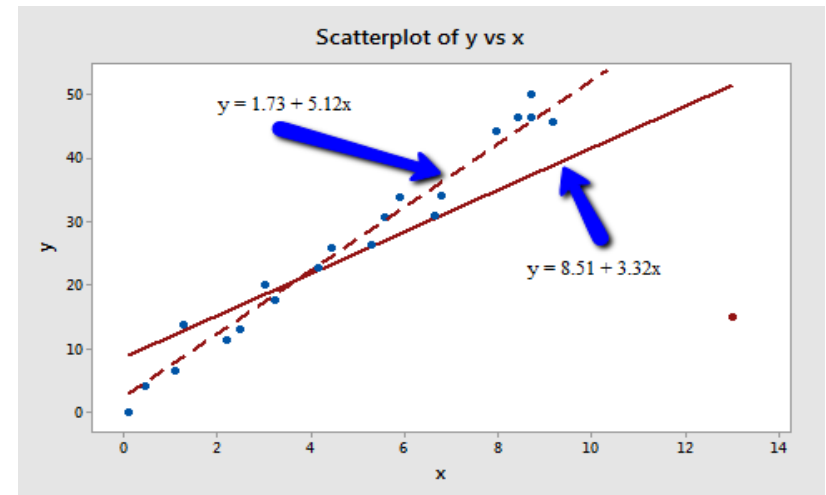
S	R-sq	R-sq(adj)	R-sq(pred)
10.4459	55.19%	52.84%	19.11%

## Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	8.50	4.22	2.01	0.058	
x	3.320	0.686	4.84	0.000	1.00

## Regression Equation

$$y = 8.50 + 3.320 x$$



## Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.59199	97.32%	97.17%	96.63%

## Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.73	1.12	1.55	0.140	
x	5.117	0.200	25.55	0.000	1.00

## Regression Equation

$$y = 1.73 + 5.117 x$$

# Influential Observations

---

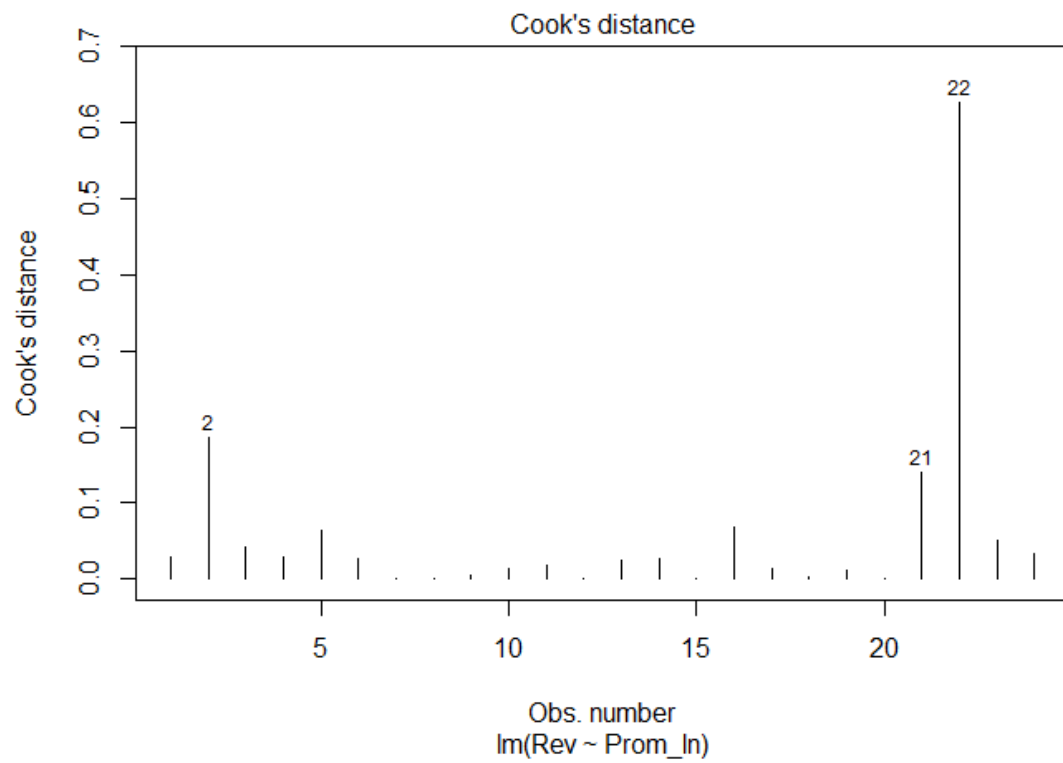
Cook's distance (Cook, 1977) measures the change in the regression parameters and thus how much the predicted value of the dependent variable changes for all the observations in the sample when a particular observation is excluded from sample for the estimation of regression parameters.

Rule of Thumb:

influential observations have Cook's Distance  $> 1.0$

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(m+1)s^2} \left[ \frac{h_i}{(1-h_i)^2} \right]$$

- $(y_i - \hat{y}_i)$   $i_{th}$  residual
- $s$  standard error of the estimate
- $h_i$  leverage of  $i_{th}$  observation
- $m$  number of predictors



# Model Development

