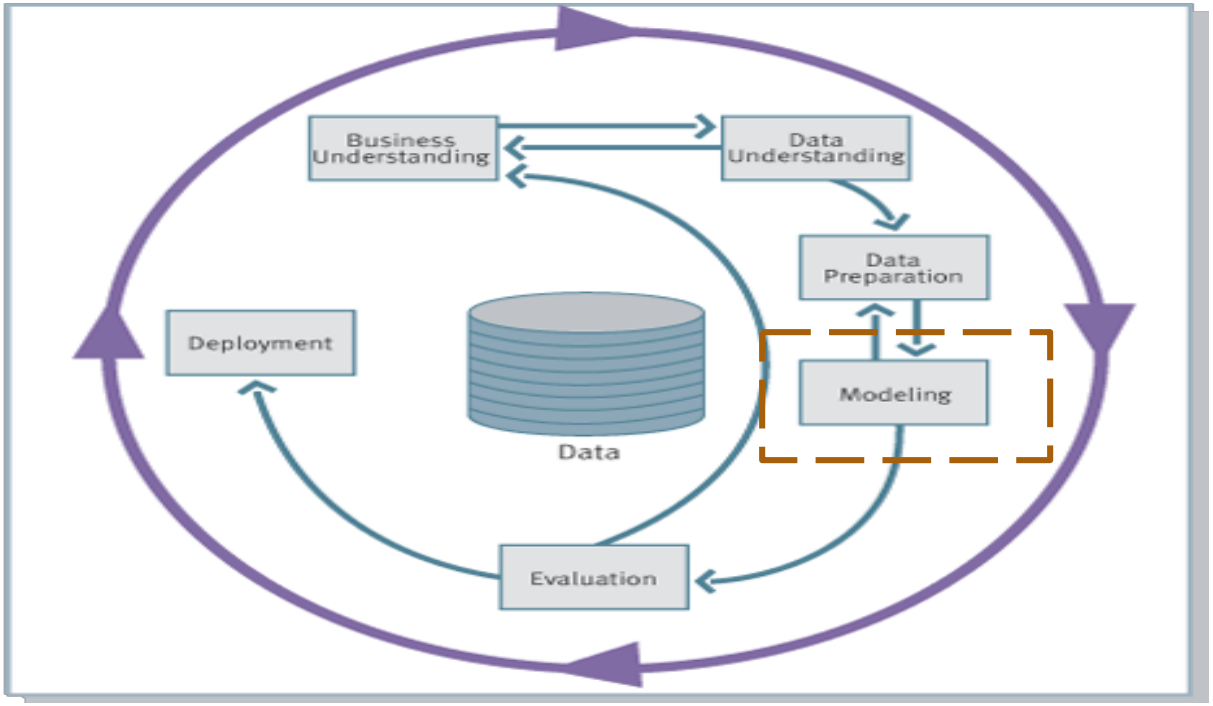
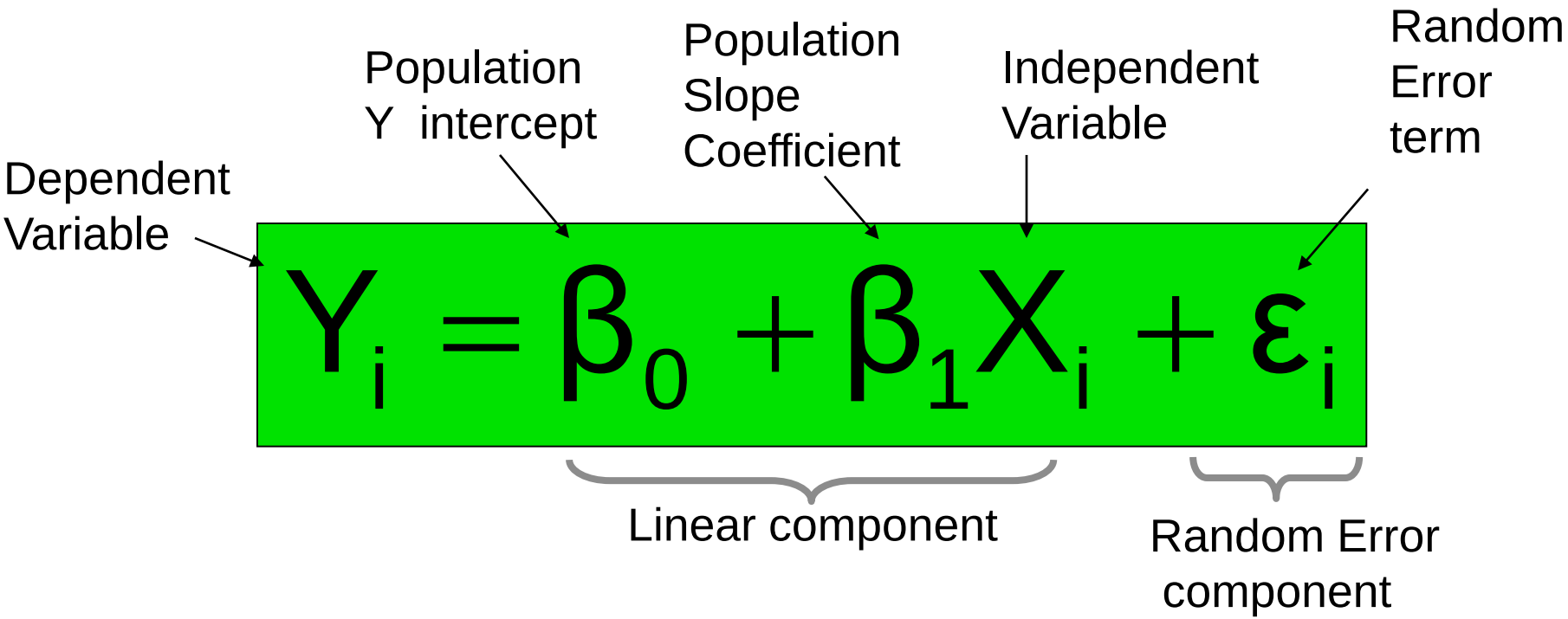
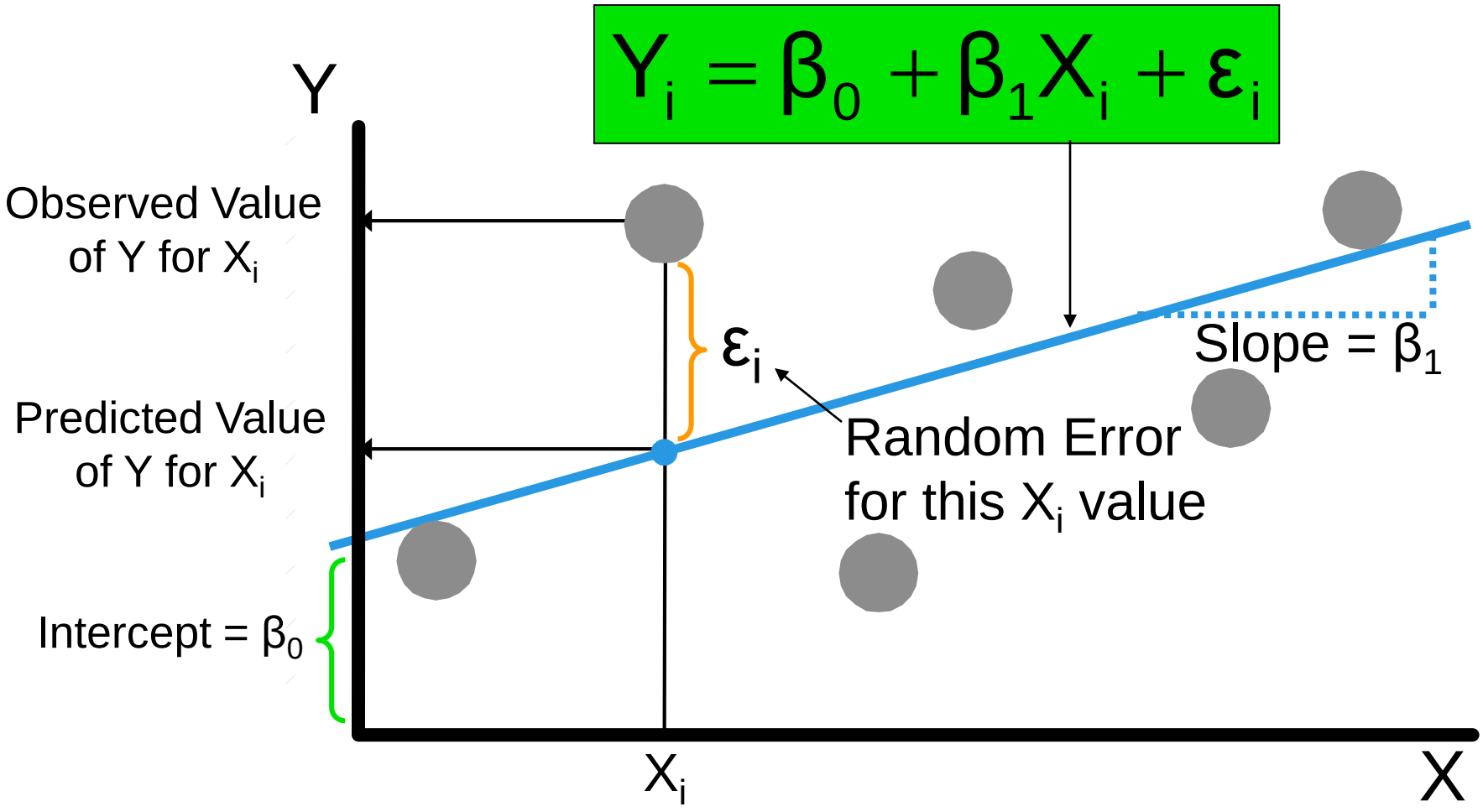
What we have done so far?



Simple Linear Regression Model

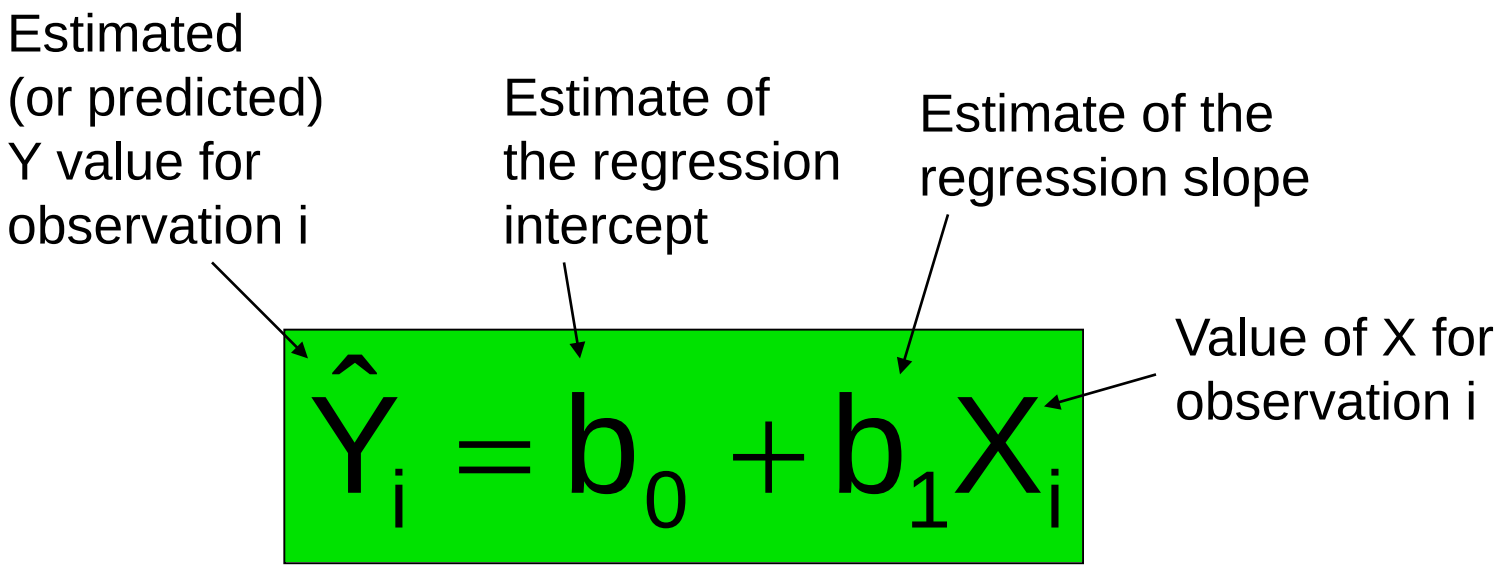


Simple Linear Regression Model



Simple Linear Regression Equation (Prediction Line)

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



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

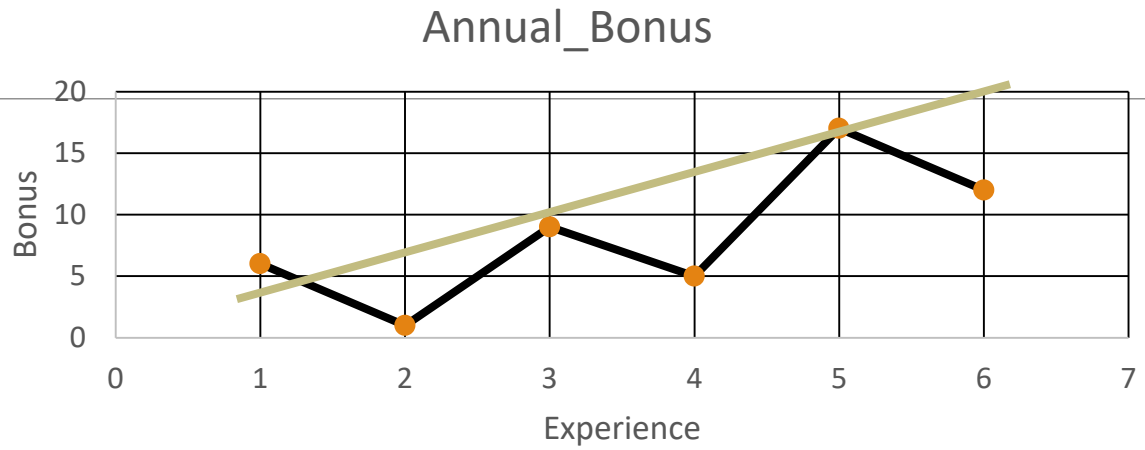
The annual bonuses ($1,000s) of six employees with different years of experience were recorded as follows. We wish to determine the straight line relationship between annual bonus and years of experience.

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

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

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

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

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

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

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

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

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

	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

$$b_1 = \frac{37}{17.5} = 2.114 \qquad b_0 = \frac{50}{6} - 2.114 * \frac{21}{6} = 0.9343$$

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

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)

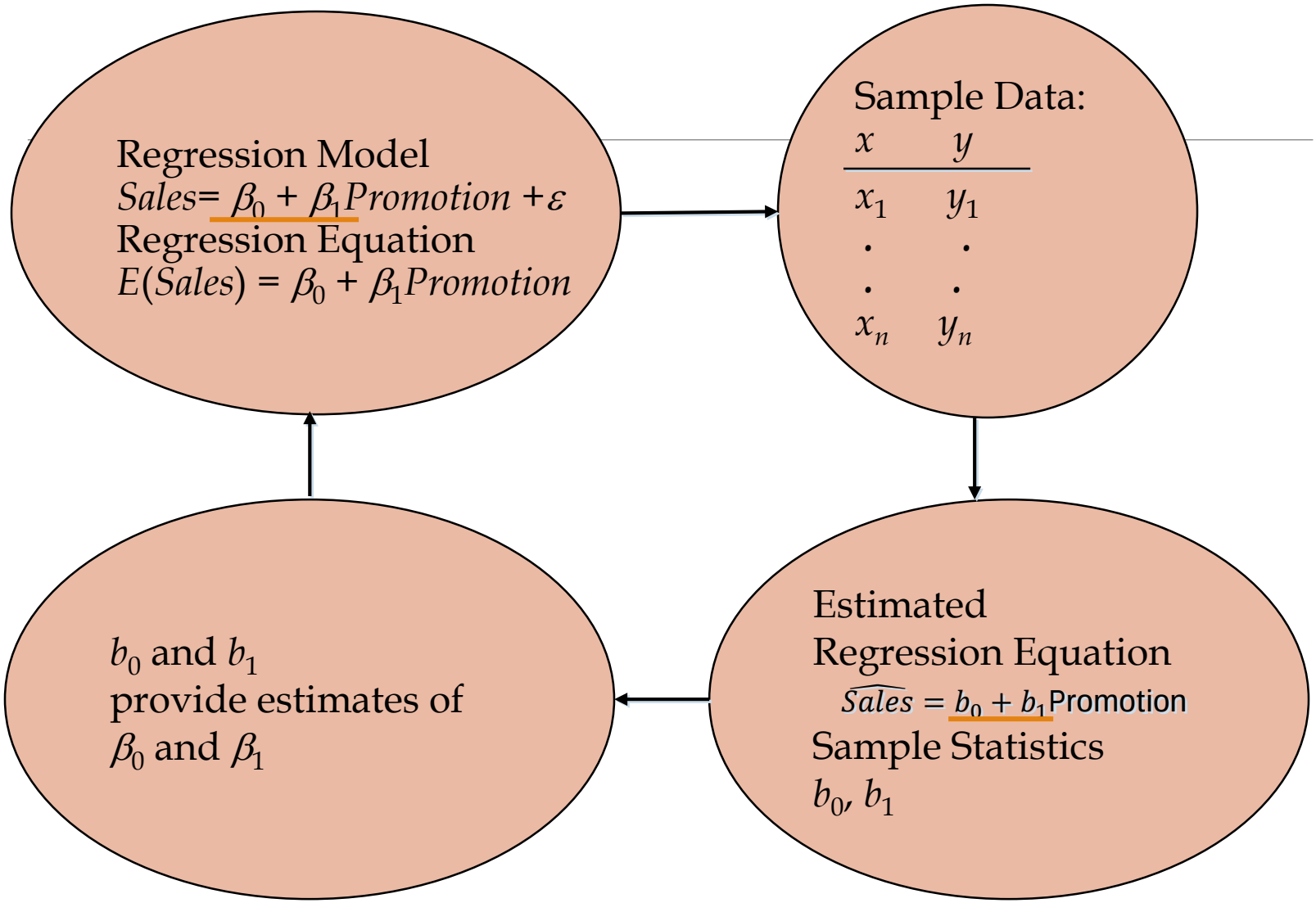
S.No.	Rev	Prom	S.No.	Rev	Prom
1	5	1	13	16	7
2	6	1.8	14	17	8.1
3	6.5	1.6	15	18	8
4	7	1.7	16	18	10
5	7.5	2	17	18.5	8
6	8	2	18	21	12.7
7	10	2.3	19	20	12
8	10.8	2.8	20	22	15
9	12	3.5	21	23	14.4
10	13	3.3	22	7.1	1
11	15.5	4.8	23	10.5	2.1
12	15	5	24	15.8	4.75

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.

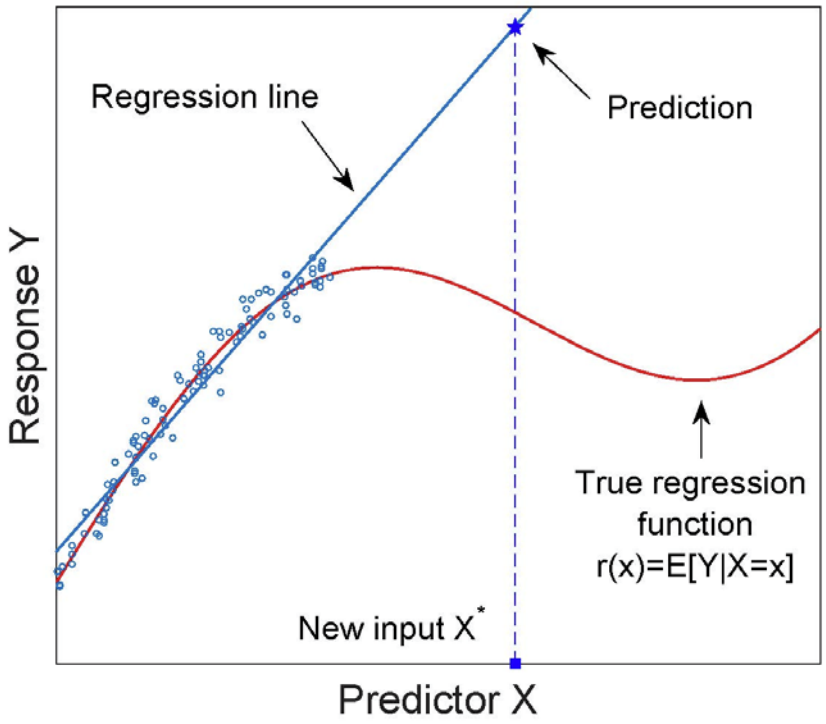
b1 = Expected change in the revenue for every one unit increase in promotion.

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

Ans: 12.735

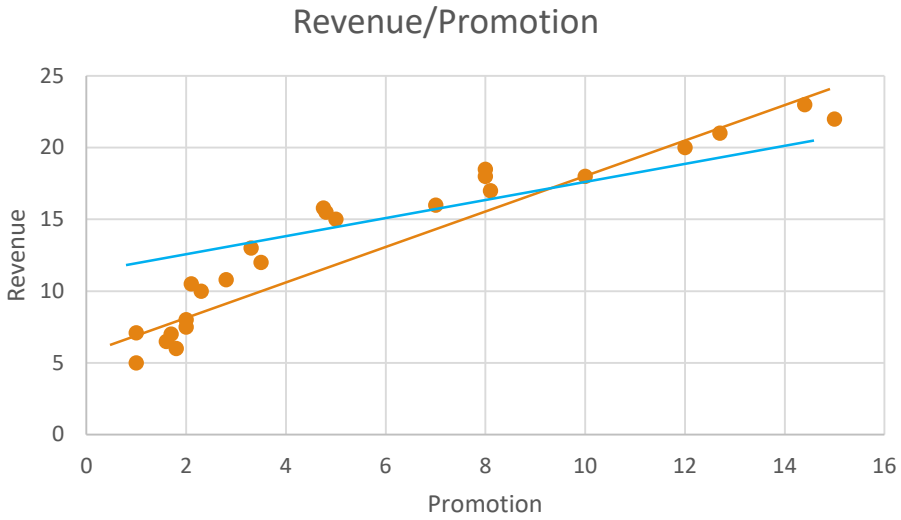
Expected Sales =
6.83+1.81*Promotion

Danger of Extrapolation



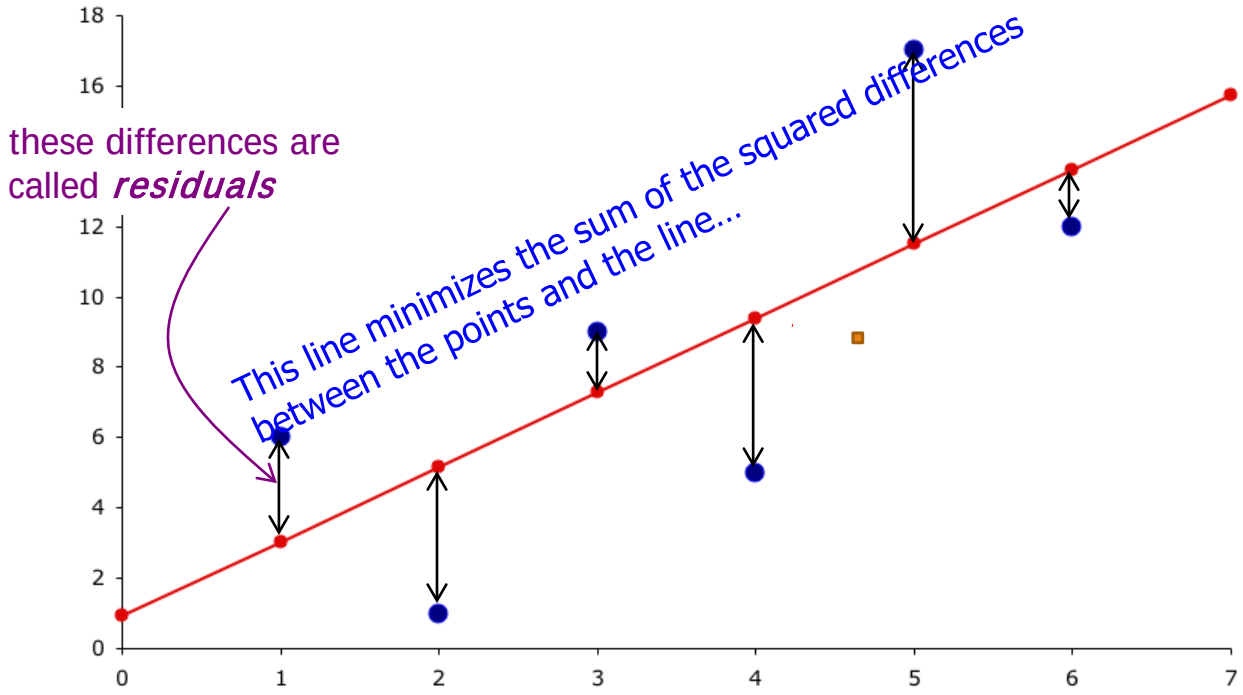
Range of Promotion Variable is [1,15].
Perdiction for values outside this range could be misleading.

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)

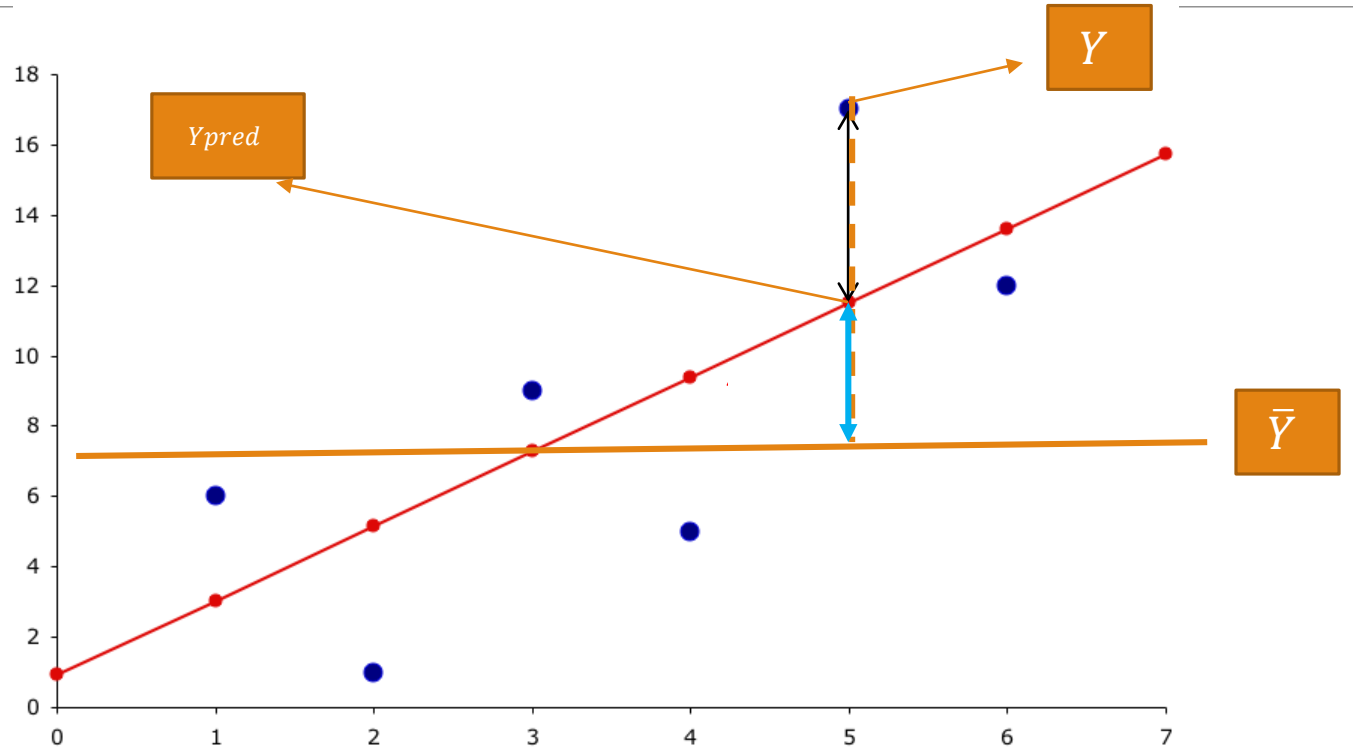
  R^2 is proportion of variability in response variable explained

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

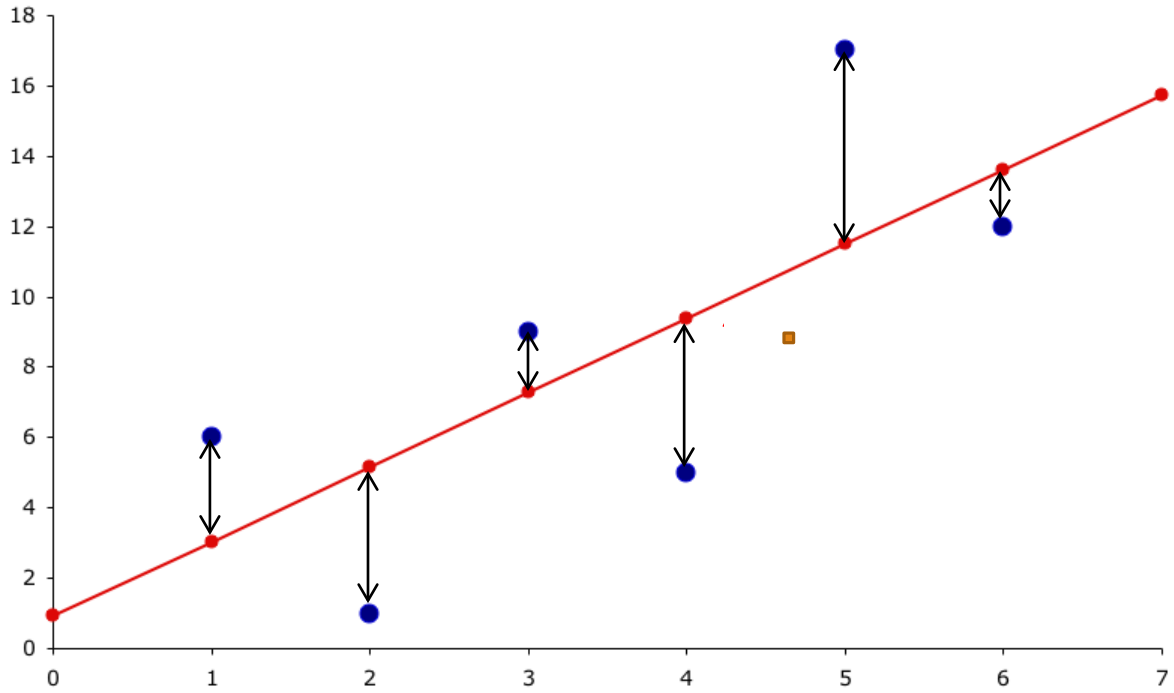
Sum of Squares Error SSE = $\sum(y - \hat{y})^2$

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

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:

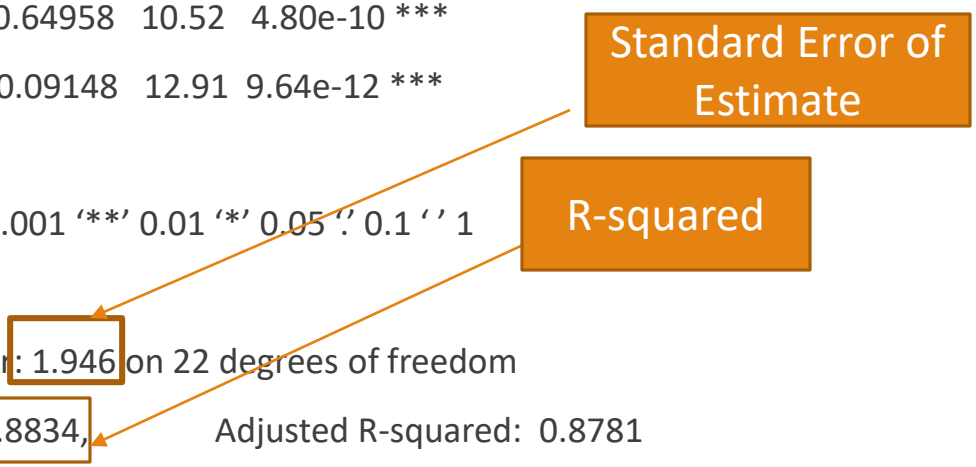
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	6.83086	0.64958	10.52	4.80e-10	***
Prom	1.18101	0.09148	12.91	9.64e-12	***

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



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:

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

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

When null hypothesis true $t = b_1 / S_{b1}$ follows t-distribution with n-2 df.

Large values S_{b1} indicate estimate of slope b_1 *unstable*

Small values S_{b1} indicate estimate of slope b_1 *precise*

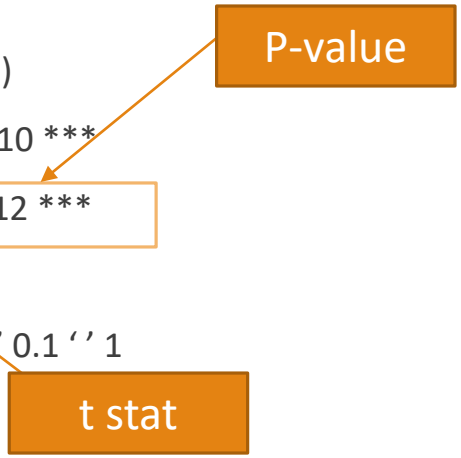
100(1-alpha)% confidence interval for slope β_1, where t_{n-2} based on 2 degrees of freedom:

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

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

Interval Estimate

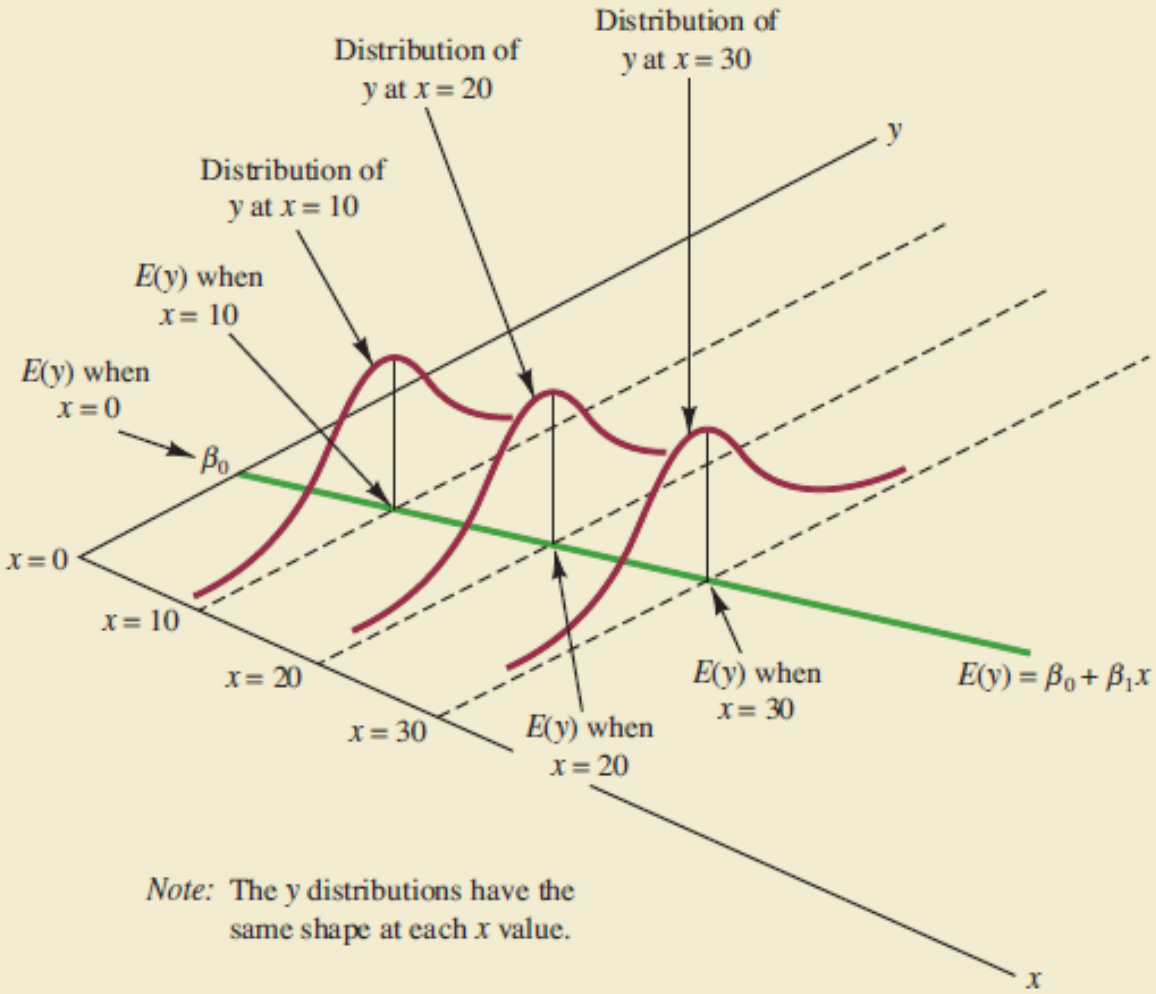
Prediction Interval of y Given x

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

WWhat 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



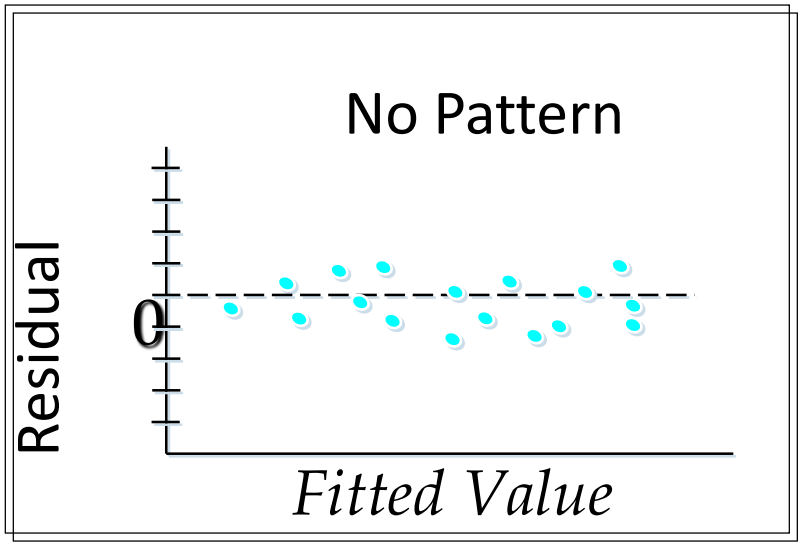
Assumptions

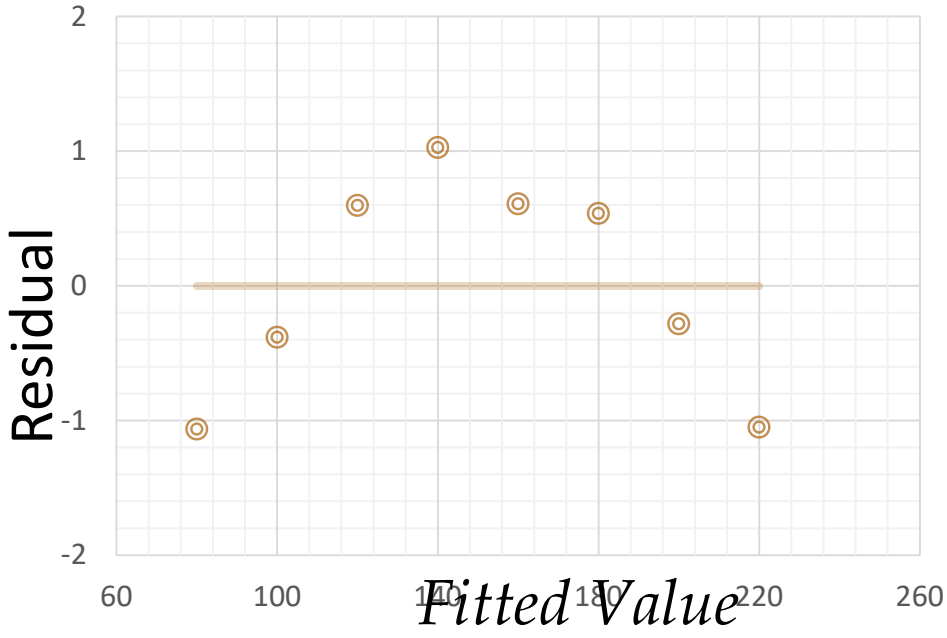
L	**Linear Function:** The mean of the response, E(Yi), at each value of the predictor, xi, is a linear function of the xi.
I	**Independent:** The errors, ϵi, are Independent. (Not a problem in cross-sectional data)
N	**Normally Distributed:** The errors, ϵi, at each value of the predictor, xi, are Normally distributed.
E	**Equal variances (denoted σ2):** The errors, ϵi, at each value of the predictor, xi, have Equal variances.

Detecting Nonlinearity of Regression Function

Plot Residuals versus Fitted Value

- Random Cloud around 0 ⇒ Linear Relation
- U-Shape or Inverted U-Shape ⇒ Nonlinear Relation





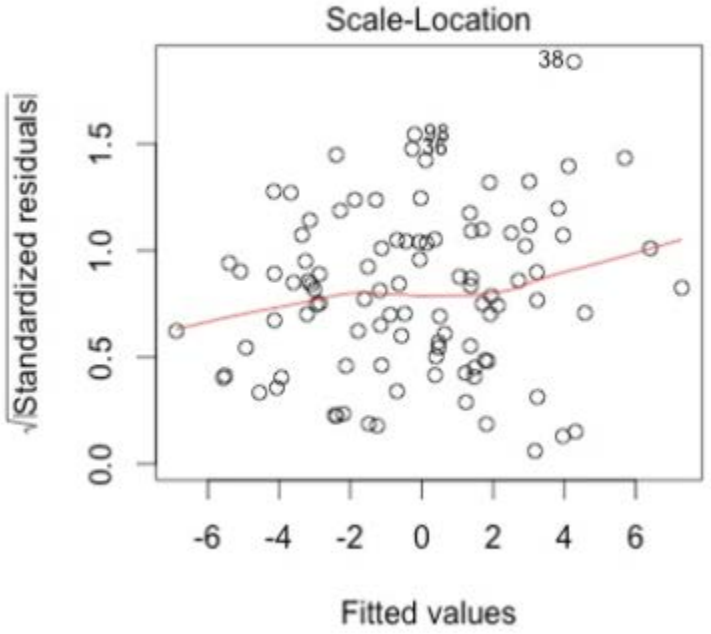
Constant Variance

Plot Residuals versus Predicted Values

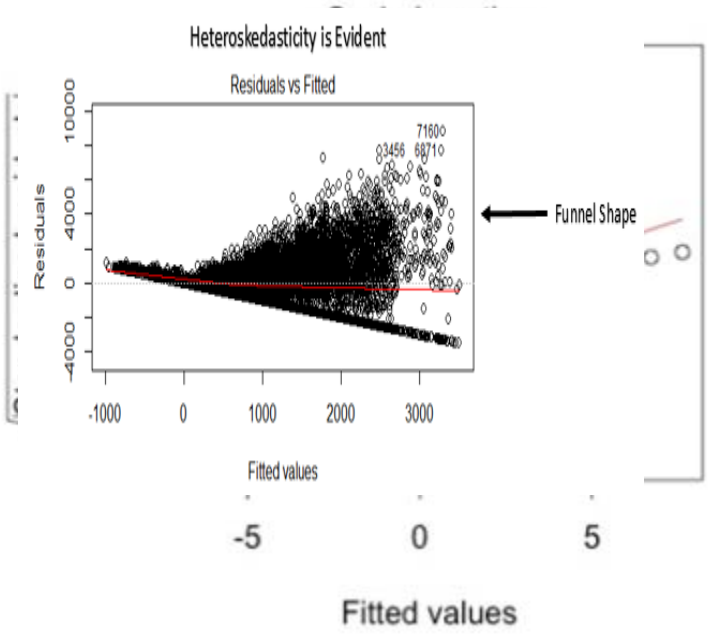
- Random Cloud around 0 $\Rightarrow$ Linear Relation
- Funnel Shape $\Rightarrow$ Non-constant Variance

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

Homoskedasticity is Evident



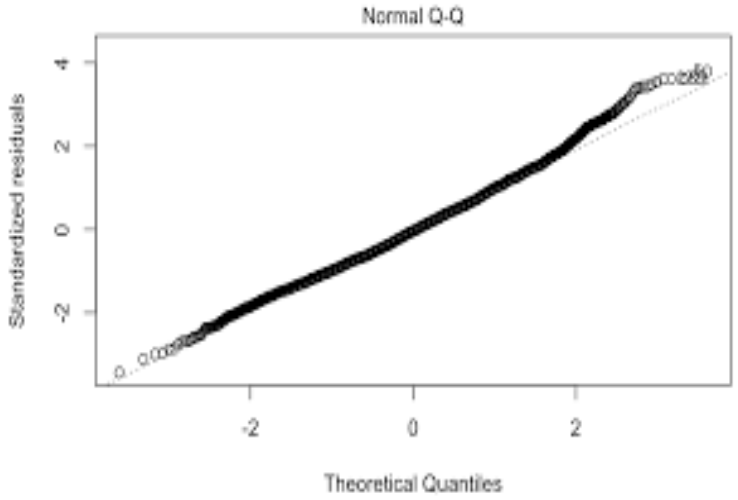
Heteroskedasticity is Evident



Normality Assumption

Normal Probability Plot

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

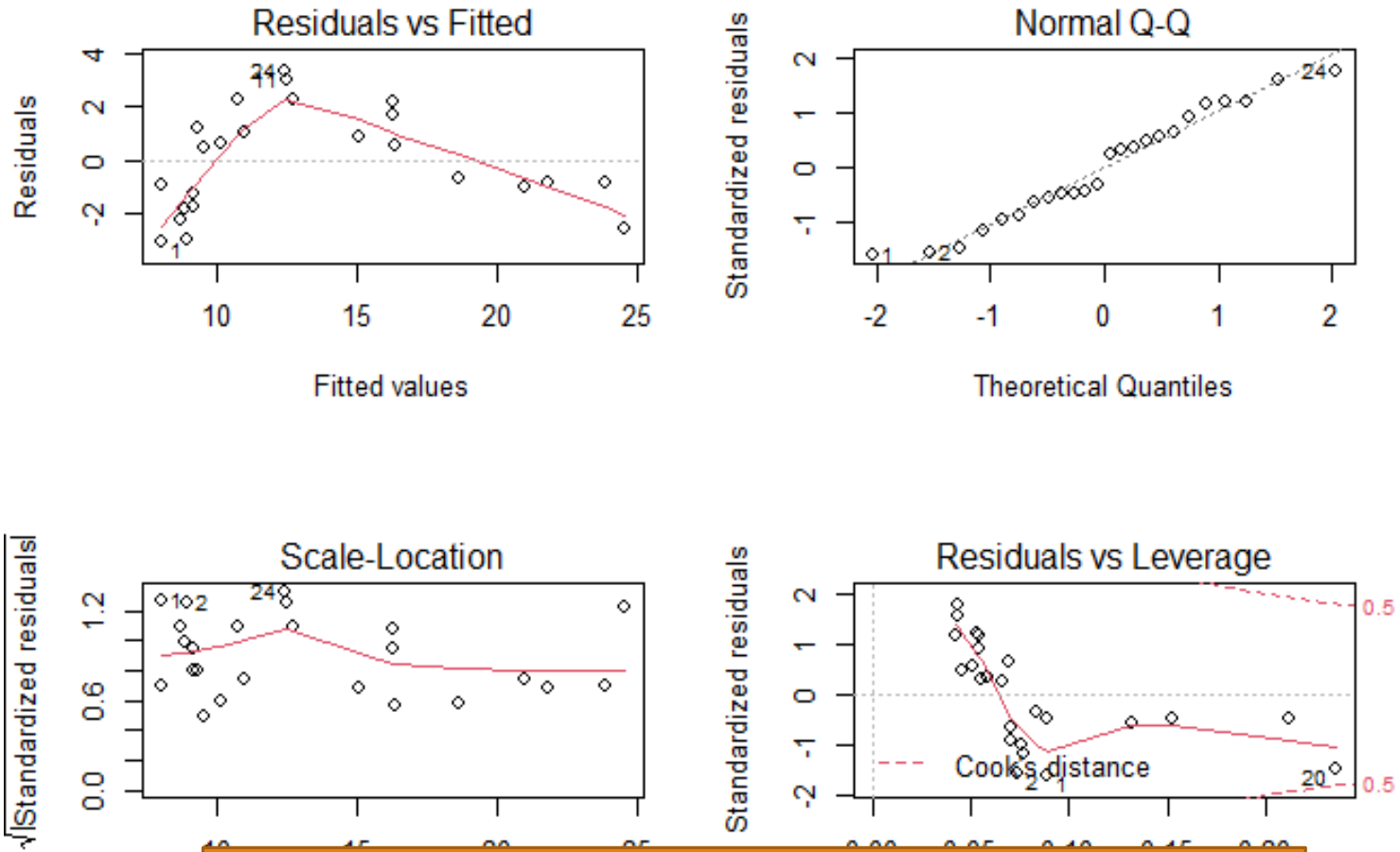


If the points lie majorly on the line, errors are assumed to be normally distributed.

Remedial Measures

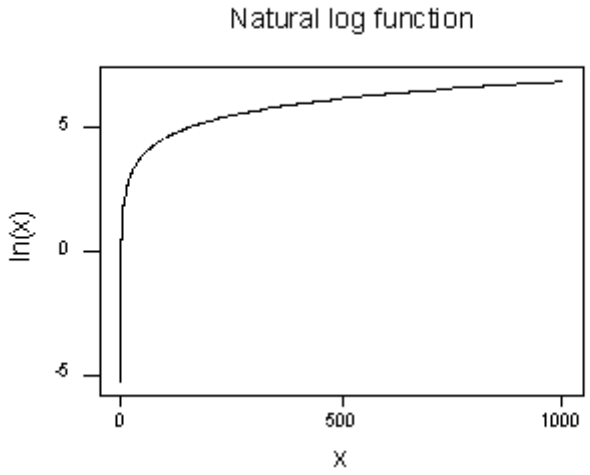
- Nonlinear Relation – Add polynomials, fit exponential regression function, or transform *X*
- Non-Constant Variance – Weighted Least Squares, transform *Y* and/or *X*, or fit Generalized Linear Model
- Non-Normality of Errors – Transform Y or fit Generalized Linear Model
- Non-Independence of Errors – 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	$Log_{10}(X)$
1	0
10	1
100	2

$$\hat{Y} = b_0 + b_1 x \qquad \widehat{Y_n} = b_0 + b_1(x+1)$$

$$\widehat{Y_n} - \hat{Y} = b_1$$

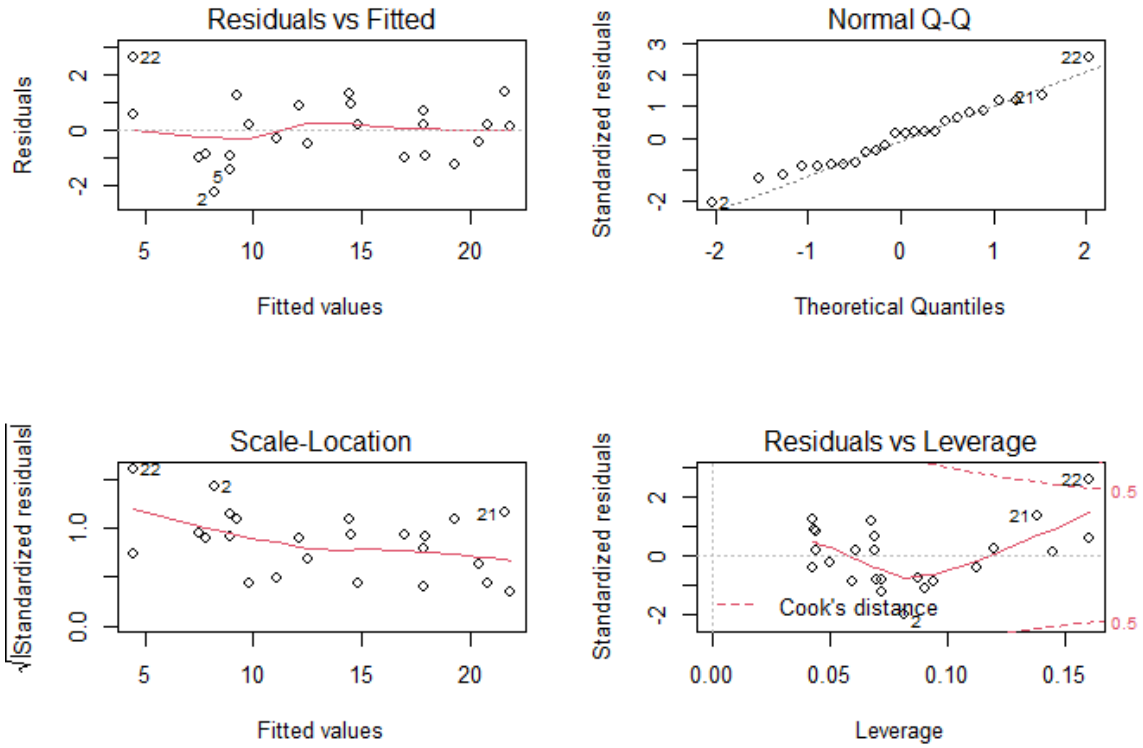
$$\hat{Y} = b_0 + b_1 ln x \qquad \widehat{Y_n} = b_0 + b_1 \ln(kx)$$

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

Interpretation:

In general, a *k*-fold increase in the predictor *x* is associated with a b1×ln(k) change in the mean of the response *y*.

Revised Model

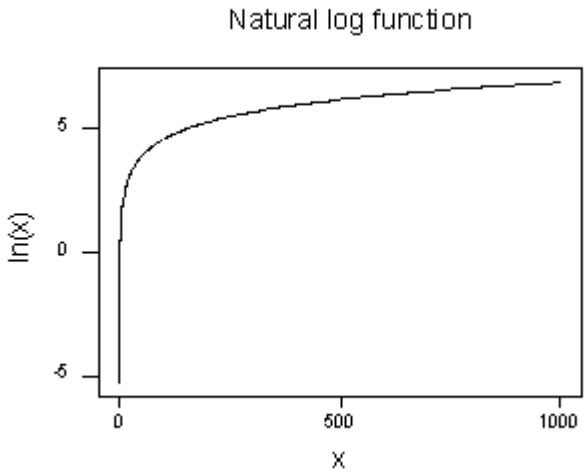


Sales= bo+b1(Promotion) ⟶ Sales= bo+b1*LN(Promotion)

Model Comparison

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

Log-Transformation in Y



$$\hat{Y} = b_0 + b_1 x \qquad \widehat{Y_n} = b_0 + b_1(x+1)$$

$$\widehat{Y_n} - \hat{Y} = b_1$$

$$\widehat{\ln Y} = b_0 + b_1 x \qquad \widehat{\ln Y_n} = b_0 + b_1(x+1)$$

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

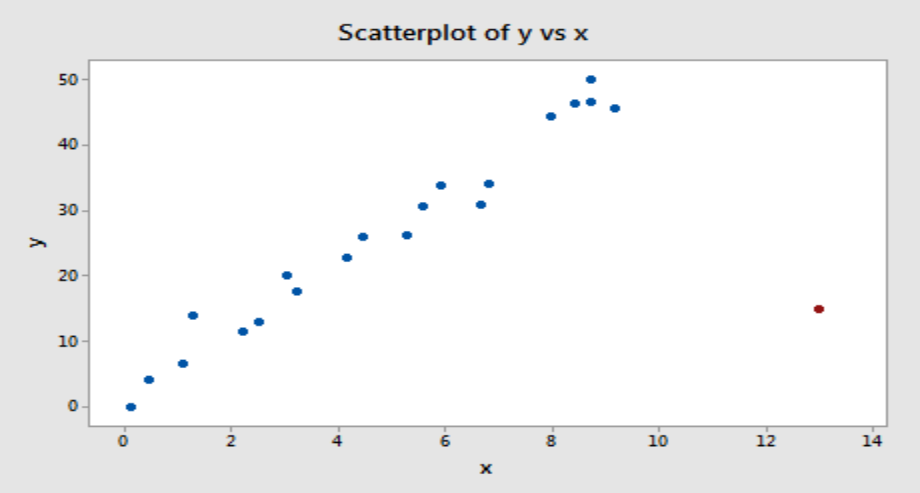
$$\widehat{Y_n} / \hat{Y} = e^{b_1}$$

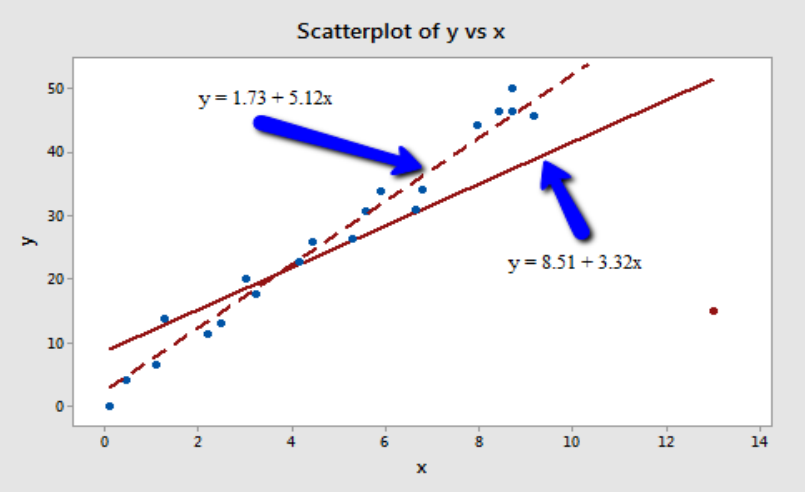
Each 1-unit increase in *X* multiplies the expected value of *Y* by e^{b_1} .

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

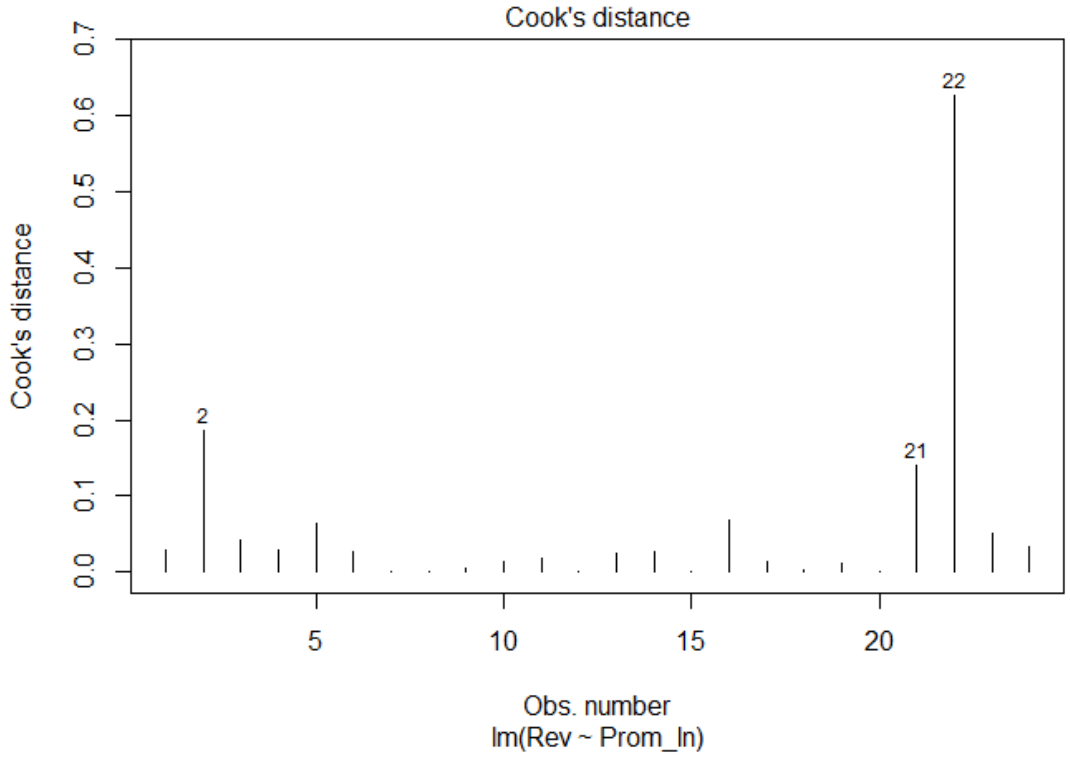
Cook's distance (Cook, 1977) measures the change in the regression parameters and thus how much the predicted value of the dependent variable changes for all the observations in the sample when a particular observation is excluded from sample for the estimation of regression parameters.

Rule of Thumb:

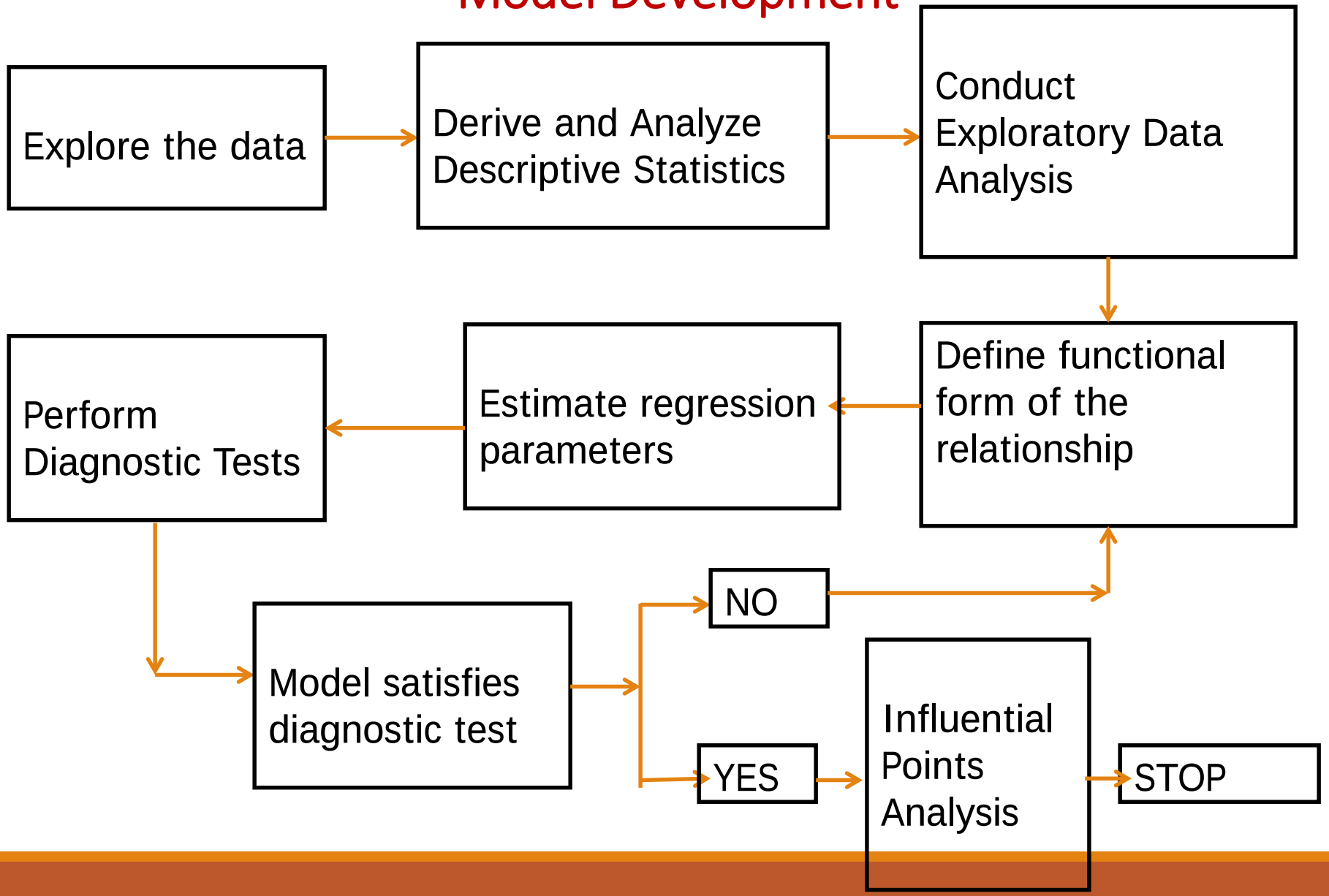
influential observations have Cook's Distance > 1.0

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

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


Cook's distance
Cook's distance
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
2
21
22
5
10
15
20
Obs. number
lm(Rev ~ Prom_ln)

Model Development



Things to do in Multiple Linear Regression:

1) Interpretation of regression coefficients

2) Multi- Collinearity

3) Variable Selection

4) Inclusion of Qualitative Predictor Variables

The Multiple Regression Model

First, recall Simple Regression Model:

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

Multiple Regression Model extends idea:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + \varepsilon$$

where, β_1, β_2, …, β_m are model parameters whose true value remains unknown and ε represents error term

Example

Data on amount of money spent (Y) by customers at an e-commerce portal, monthly income (X1) and family size (X2) is collected for 200 customers. Build a regression model.

S.No	Family Size	Income	Amount Spent
1	2	77040	1725
2	2	48000	644
3	5	77281	2010
4	3	95881	1094
5	2	92760	1947
6	1	118201	2136
7	3	112200	1498
8	4	59401	1255
9	2	152400	1913
10	4	114120	849
11	3	22080	696
12	4	16200	168
13	2	52560	636
14	2	79561	1143
15	4	22560	180
16	2	64681	1114
17	2	54960	748
18	1	22801	449

Model Summary

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.956e+02  1.371e+02   2.886 0.004336 **
Income       1.740e-02  1.064e-03  16.358  < 2e-16 ***
Family.Size -1.294e+02  3.827e+01  -3.381 0.000869 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 600.1 on 197 degrees of freedom
Multiple R-squared:  0.5841,    Adjusted R-squared:  0.5799
F-statistic: 138.3 on 2 and 197 DF,  p-value: < 2.2e-16
```

$$\text{Amou}nt = 395.6 + 0.0174 * Income - 129.4 * \text{Family.Size}$$

For every one unit increase in Income amount spent increases by 0.0174 when the variable Family.size is kept constant, and for one unit increase in Family.Size the amount spent decreases by 129.4 when income is kept constant.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.956e+02  1.371e+02   2.886 0.004336 **
Income       1.740e-02  1.064e-03  16.358  < 2e-16 ***
Family.Size -1.294e+02  3.827e+01  -3.381 0.000869 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 600.1 on 197 degrees of freedom
Multiple R-squared:  0.5841,    Adjusted R-squared:  0.5799
F-statistic: 138.3 on 2 and 197 DF,  p-value: < 2.2e-16
```

Model Performance Measures

Adjusted R-squared

SEE

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-m-1} = 1 - (1 - 0.5841)\frac{(200-1)}{(200-1-2)} = 0.579$$

		R-squared	Adjusted R-squared
Model 1	Y1~X2	0.9557	0.9493
Model 2	Y1~X1+X2	0.9573	0.9431

Temperature (Celsius) X1	Price of Dough X2	Price of Pizza Y1
21	1	5
15	3	12
16	6	15
21	8	19
27	12	24
24	15	27
21	17	29
23	21	31
21	26	36

F-test for Significance of Overall Regression Model

Ho: $\beta_1 = \beta_2 \ldots\ldots \beta_m = 0$, Model: $y = \beta_0 + \varepsilon$

Ha: At least one $\beta_i \neq 0$,

- Observed F-statistic F_{obs} = MSR/MSE follows $F_{m,\ n-m-1}$ distribution
- F-test is right-tailed test, since values non-negative
- Ho rejected when p-value small
- Where p-value = $p(F_{m,\ n-m-1} > F_{obs})$, represents area in tail to right of observed value

t-test for Significance of Individual Variable

The null and alternative hypotheses in the case of individual independent variable and the dependent variable Y is given, respectively, by

H_0: There is no relationship between independent variable X_i and dependent variable Y

H_A: There is a relationship between independent variable X_i and dependent variable Y

Alternatively,

H_0: $\beta_i = 0$

H_A: $\beta_i \neq 0$

$$t = \frac{\hat{\beta}_i - 0}{S_e(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{S_e(\hat{\beta}_i)}$$

The corresponding test statistic is given by

Checking Significance

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.956e+02  1.371e+02   2.886 0.004336 **
Income       1.740e-02  1.064e-03  16.358  < 2e-16 ***
Family.Size -1.294e+02  3.827e+01  -3.381 0.000869 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 600.1 on 197 degrees of freedom
Multiple R-squared:  0.5841,    Adjusted R-squared:  0.5799
F-statistic: 138.3 on 2 and 197 DF,  p-value: < 2.2e-16
```

Individual Variable Significance

Overall Model Significance

Example

You are an HR analyst for a company. You want to build a model for predicting Salary of the employees. To build your model, you use the data available with the company i.e. work experience of an employee. You are also interested to explore whether gender of an employee has an impact on the salary.

The data in below table provides salary, gender, and work experience (WE) of 27 workers in a firm. In Table gender = 1 denotes female and 2 denotes male and WE is the work experience in number of years

S. No.	Gender	WE	Salary	S. No.	Gender	WE	Salary
1	1	2	6800	15	2	1	17700
2	1	3	8700	16	2	6	34700
3	1	1	9700	17	2	7	38600
4	1	3	9500	18	2	7	39900
5	1	4	10100	19	2	7	38300
6	1	6	9800	20	2	3	26900
7	2	3	19100	21	2	4	31800
8	2	4	28000	22	1	5	8000
9	2	3	25700	23	1	5	8700
10	2	1	20350	24	1	3	6200
11	2	4	30400	25	1	3	4100
12	2	1	19400	26	1	2	5000
13	2	2	22100	27	1	1	4800
14	2	1	20200				

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     8157       3968   2.056  0.05039 .
WE              3090       1013   3.050  0.00535 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10160 on 25 degrees of freedom
Multiple R-squared:  0.2712,    Adjusted R-squared:  0.242
F-statistic: 9.302 on 1 and 25 DF,  p-value: 0.005355
```

Does Salary depend on the gender of the employee?

Inclusion of a categorical variable in Regression Model

If a categorical variable has k categories; create k-1 dummy variables.

The category for which there is no dummy variable is called base category.

Example

Gender Variable Contains two categories: 1(Female); 2 (Male)

Create 'Gender2' a dummy variable representing Females

Gender 2 = 1 if Gender =1; Gender 2 = 0 if Gender != 1

$Y = \beta_0 + \beta_1 \times$ Gender2

Gender2=1 ; $Y1 = \beta_0 + \beta_1$

Gender 2=0 ; $Y2 = \beta_0$

β_1 = Y1-Y2

The coefficient to a specific category represents the change in the value of Y from base category.

Does Salary depend on the gender of the employee?

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    27543       1543  17.848 9.73e-16 ***
Gender2       -19927       2315  -8.608 6.02e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5977 on 25 degrees of freedom
Multiple R-squared:  0.7477,    Adjusted R-squared:  0.7376
F-statistic:  74.1 on 1 and 25 DF,  p-value: 6.022e-09
```

Salary of female employees is on average (-19927) of male employees

Solution

Let the regression model be:

$Y = \beta_0 + \beta_1 \times \text{Gender2} + \beta_2 \times \text{WE}$

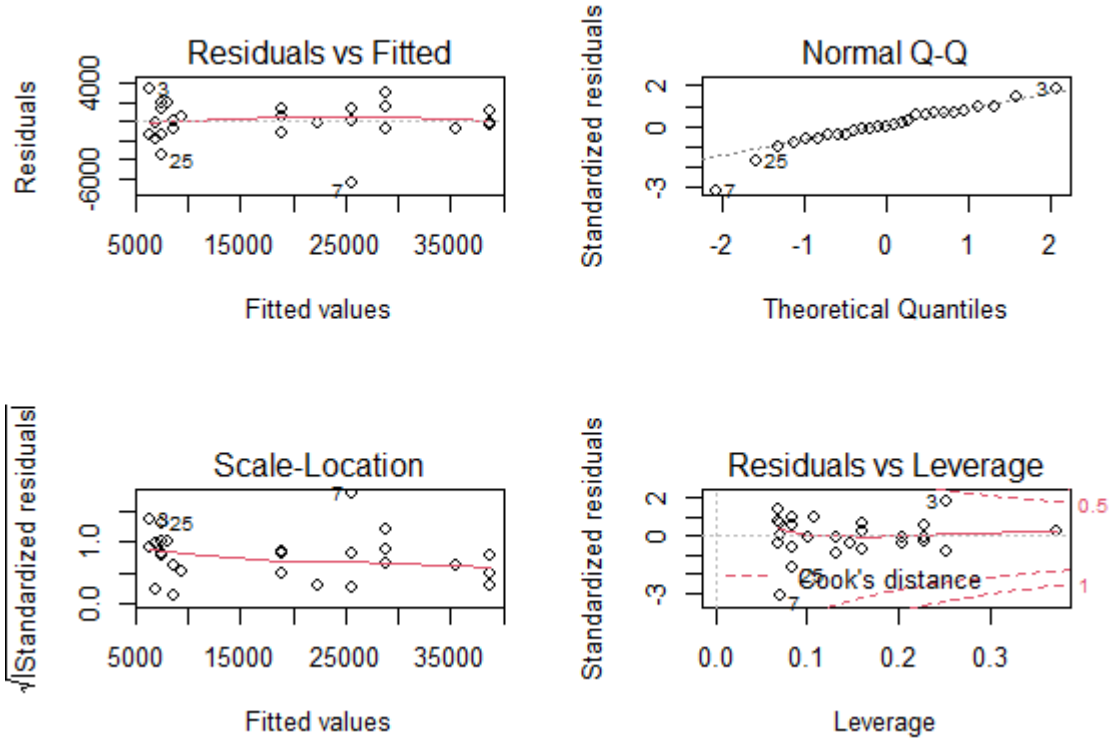
R output:

Gender 2:
1: Female
0: Male

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  18365.3     1427.2  12.868 2.91e-12 ***
WE            2549.5      322.6   7.904 3.91e-08 ***
Gender2     -18821.9     1252.5 -15.027 1.04e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3214 on 24 degrees of freedom
Multiple R-squared:   0.93,     Adjusted R-squared:  0.9241
F-statistic: 159.4 on 2 and 24 DF,  p-value: 1.388e-14
```

Diagnostics: Check for Assumptions



Assumptions

Absence of Multi-Collinearity: Predictor Variables should not be correlated with each other.

L	**Linear Function:** The mean of the response, E(Yi), at each value of the predictor, xi, is a linear function of the xi.
I	**Independent:** The errors, ϵi, are Independent. (Not a problem in cross-sectional data)
N	**Normally Distributed:** The errors, ϵi, at each value of the predictor, xi, are Normally distributed.
E	**Equal variances (denoted σ2):** The errors, ϵi, at each value of the predictor, xi, have Equal variances.

Multicollinearity

Multicollinearity is condition where two or more predictors are correlated.

Effects of multi collinearity

- Data set with severe multicollinearity may have significant F-test, while having no significant t-tests for predictors.
- The interpretation of coefficients as measuring marginal effects is unwarranted in the presence of correlated variables.

(b) Regression of Y on X_1

$\hat{Y} = 23.500 + 5.375X_1$

Source of Variation	SS	df	MS
Regression	231.125	1	231.125
Error	188.750	6	31.458
Total	419.875	7	

(c) Regression of Y on X_2

$\hat{Y} = 27.250 + 9.250X_2$

Source of Variation	SS	df	MS
Regression	171.125	1	171.125
Error	248.750	6	41.458
Total	419.875	7	

When the predictor variables are uncorrelated, the coefficients remain the same when other predictor variables are included.

Example -1

Case i	Crew Size X_{i1}	Bonus Pay (dollars) X_{i2}	Crew Productivity Y_i
1	4	2	42
2	4	2	39
3	4	3	48
4	4	3	51
5	6	2	49
6	6	2	53
7	6	3	61
8	6	3	60

Correlation between X1 and X2 is 0.

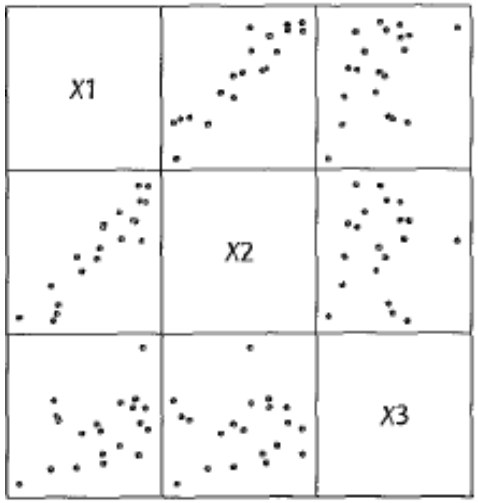
Example -1

Case i	Crew Size X_{i1}	Bonus Pay (dollars) X_{i2}	Crew Productivity Y_i
1	4	2	42
2	4	2	39
3	4	3	48
4	4	3	51
5	6	2	49
6	6	2	53
7	6	3	61
8	6	3	60

Subject i	Triceps Skinfold Thickness X_{i1}	Thigh Circumference X_{i2}	Midarm Circumference X_{i3}	Body Fat Y_i
1	19.5	43.1	29.1	11.9
2	24.7	49.8	28.2	22.8
3	30.7	51.9	37.0	18.7
...	...	...	...	...
18	30.2	58.6	24.6	25.4
19	22.7	48.2	27.1	14.8
20	25.2	51.0	27.5	21.1

Variables in Model	b_1	b_2
X_1	.8572	—
X_2	—	.8565
X_1, X_2	.2224	.6594
X_1, X_2, X_3	4.334	−2.857

(a) Scatter Plot Matrix of X Variables



(b) Correlation Matrix of X Variables

$$r_{XX} = \begin{bmatrix} 1.0 & .924 & .458 \\ .924 & 1.0 & .085 \\ .458 & .085 & 1.0 \end{bmatrix}$$

- Effect on the coefficient of X1 in the presence of X2 and X3.
- The coefficient of x2 changed direction in the presence of X1 and X3

Variance Inflation Factor

Variance Inflation Factors (VIFs) report presence of multicollinearity

$$VIF_i = \frac{1}{1 - R_i^2}$$

R_i^2 is R^2 obtained by regressing x_i against other predictors

Note: R_i^2 large when x_i highly correlated with other predictors

$VIF_i > 10$ indicates severe multicollinearity

Variance Inflation Factor

- $R_i^2 = 0$ leads to minimum value for $VIF_i = 1$
- Alternately, VIF_i increases without bound as R_i^2 approaches 1
- Where x_i uncorrelated with predictors, $VIF_i = 1$ .In general, $VIF_i > 5$ and