Chapter 9:

CORRELATION AND REGRESSION

INTRODUCTION

Statistics is often used to investigate the relationship between two (or more) variables of interest. The following are some examples of relations are often studied:

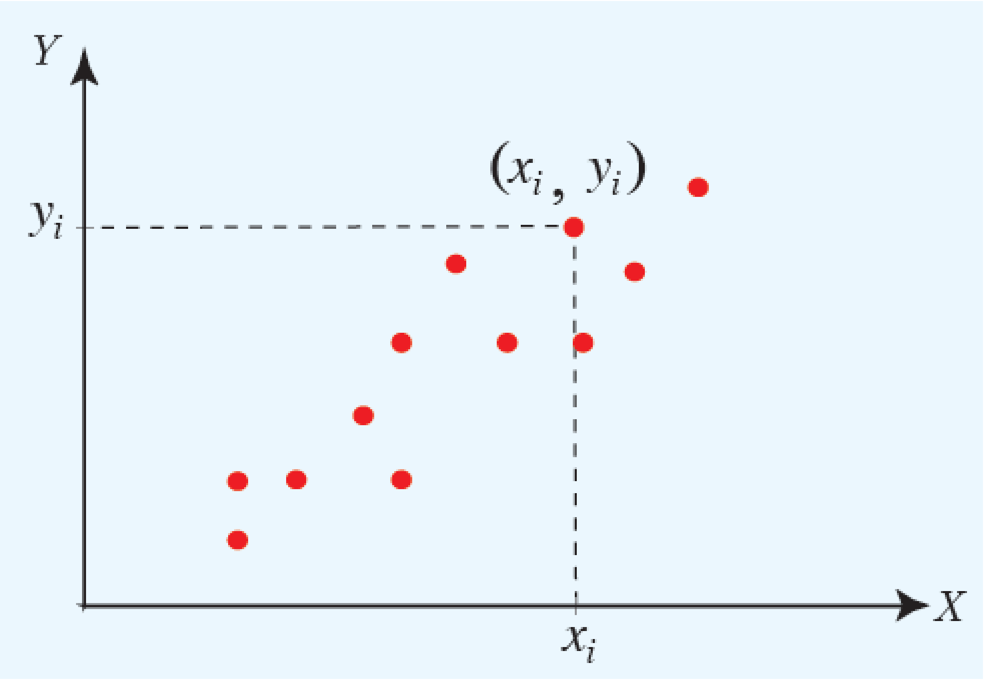
- Is there a relationship between high school grade and the first year college grade point average (GPA)? If so, what is the relationship?
- What is the relationship between the expenditure and income of a Saudi family?
- What is the relationship between the age and blood pressure?
- The relationship between body mass index and systolic blood pressure, or between hours of exercise per week and percent body fat.

In the above examples, we see that there are two basic questions of interest when investigating a pair of variables:

1. **Is there a relationship between the two variables?**
2. **What is the relationship (if any) between the two variables?**

SCATTER PLOT:

Scatter plot is a graph of data, that given in the form of binaries (pairs) (x_i , y_i) , so that each binary is represented by a point in the coordinate plane XoY (i.e., we represent the data by points). It is usually we take the orthogonal coordinates this representation. See the following graph.



CORRELATION COEFFICIENT: (Pearson's Correlation Coefficient)

Let (x_1, y_1), (x_2, y_2),…and (x_n, y_n) be binaries given data. Then the **Pearson's Correlation Coefficient** (or Pearson coefficient of linear correlation) is given by the following relation:

$$\mathbf{r} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\ \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Or using the following relation:

$$\mathbf{r} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\cdot\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\ \sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

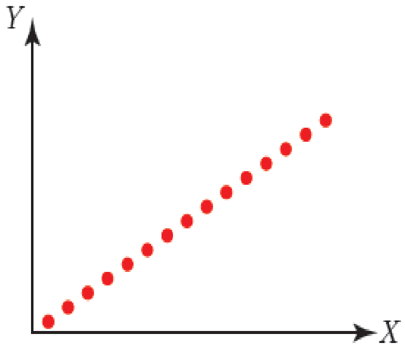
How to interpret the correlation coefficient?

The sign and the absolute value of a correlation coefficient (**r**) describe the direction and the magnitude of the relationship between two variables or two phenomena.

- The value of a correlation coefficient ranges between -1 and 1.
- The greater the absolute value of the correlation coefficient, the greater the correlation between the two variables.
- The strong linear relationship is indicated by a correlation coefficient, that is close to ± 1 or equal to ± 1 , and when the correlation coefficient (**r**) is equal to ± 1 , then one say that the relationship between two variables is complete linear.

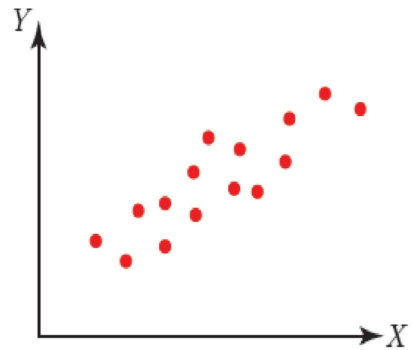
- The weak linear relationship is indicated by a correlation coefficient, that is close to zero or equal to zero, and when the correlation coefficient (**r**) equal to zero, then one says not, that doesn’t represent a relationship between the two variables, because it is possible that the relationship between the two variables is not linear (see upcoming drawings for models of the correlation).
- A positive correlation means that if one variable gets bigger value, the other variable tends to get bigger value also, i.e. the relationship between the two variables is positive monotone.
- A negative correlation means that if one variable gets bigger, the other variable tends to get smaller, i.e. the relationship between the two variables is negative monotone.

The scatterplots below show how different patterns of data produce different degrees of line correlation.



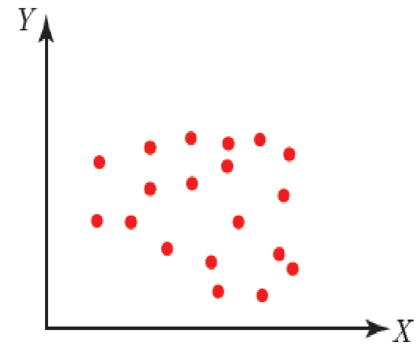
Maximum positive correlation

($\mathbf{r}$ = 1.0)



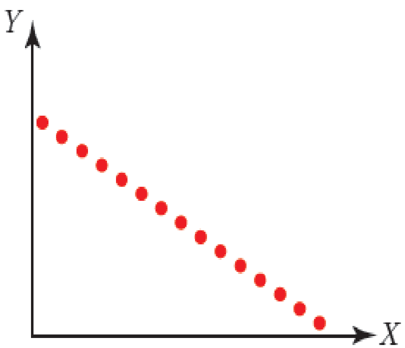
Strong positive correlation

($\mathbf{r}$ = 0.80)



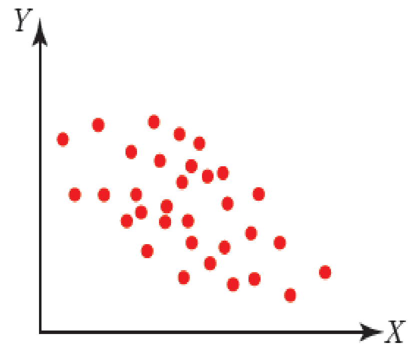
Very Weak correlation

($\mathbf{r}$ = 0.25)



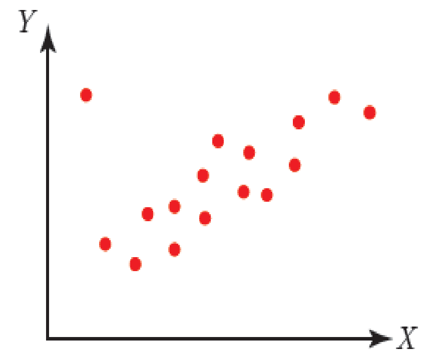
Maximum negative correlation

($\mathbf{r}$ = -1.0)



Weak negative correlation

($\mathbf{r}$ = -0.45)



Strong correlation & outlier

($\mathbf{r}$ = 0.7)

Several points are evident from the scatterplots.

- When the slope of the line in the plot is negative, the correlation is negative; and vice versa.
- The strongest correlations ($\mathbf{r} = 1.0$ and $\mathbf{r} = -1.0$) occur when data points fall *exactly* on a straight line.
- The correlation becomes weaker as the data points become more scattered.
- If the data points fall in a random pattern with unclear direction, the correlation is equal to zero or very close to zero.
- Correlation is affected by outliers. Compare the second scatterplot with the last scatterplot. The single outlier in the last plot greatly reduces the correlation (from 0.80 to 0.70).

SIMPLE LINEAR CORRELATION

- There are many statistical tests to determine the strength and the significance of the linear relationship between X and Y. In general, we might use the following rule to determine the strength of the linear relationship.

- The square value of correlation coefficient ($\mathbf{r}$) is called the coefficient of determination and one denoted it by $\mathbf{r}^2$.

ASSESSMENT OF CORRELATION STRENGTH

The Relationship between the two variables X and Y (or phenomena)	The Range of $\mathbf{r}$
No linear or $S_X = 0$ or $S_Y = 0$	$\mathbf{r} = 0$
Very weak	$0 < \lvert \mathbf{r} \rvert \leq 0.30$
Weak (an acceptable degree of linearity)	$0.30 < \lvert \mathbf{r} \rvert \leq 0.50$
Moderately strong linear	$0.50 < \lvert \mathbf{r} \rvert \leq 0.70$
Strong (the linearity very clear)	$0.70 < \lvert \mathbf{r} \rvert \leq 0.86$
Very Strong (high degree of linearity)	$0.86 < \lvert \mathbf{r} \rvert < 1$
Complete (all points are located on one straight)	$\lvert \mathbf{r} \rvert = 1$

EXAMPLE:

The results of a class of 10 students on midterm exam marks (X) and on the final examination marks (Y) are as follows:

The values of X	77	54	71	72	81	94	96	99	83	67
The values of Y	82	38	78	34	47	85	99	99	79	68

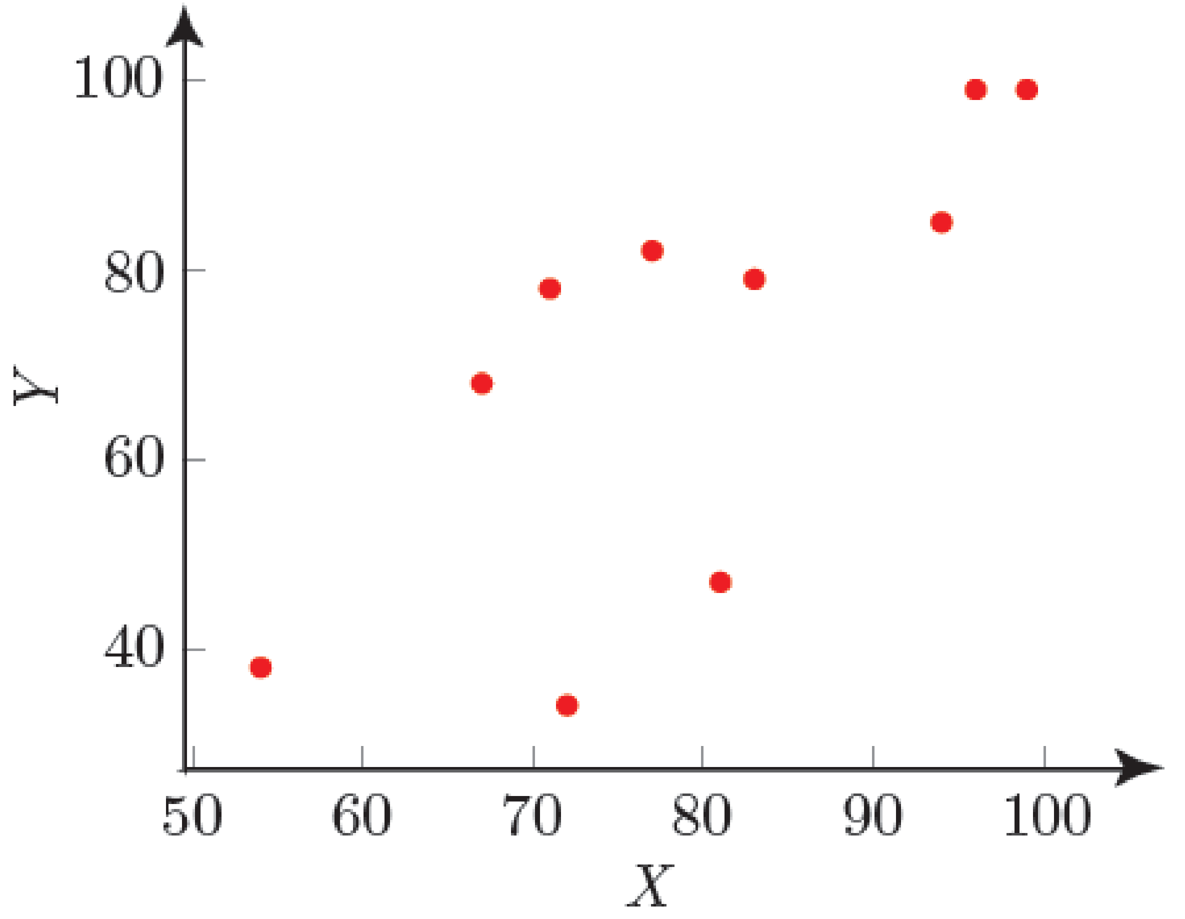
a. Represent the given data on the scatter plot.

b. Is there a linear relationship (linear association) between X and Y? Is it positive or negative?

c. Calculate the correlation coefficient (**r**).

Solution: We have:

For a) The scatter plot for the given data is:



For b) The scatter plot suggests that there is a positive linear association between X and Y. since there is a linear trend for which the value of Y linearly increases when the value of X increases.

For c) To calculating the coefficient of correlation $(\boldsymbol{r})$ we will create the following table:

i	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	77	82	-2.4	11.1	5.76	123.21	-26.64
2	54	38	-25.4	-32.9	645.16	1082.41	835.66
3	71	78	-8.4	7.1	70.56	50.41	-59.64
4	72	34	-7.4	-36.9	54.76	1361.61	273.06
5	81	47	1.6	-23.9	2.56	571.21	-38.24
6	94	85	14.6	14.1	213.16	198.81	205.86
7	96	99	16.6	28.1	275.56	789.61	466.46
8	99	99	19.6	28.1	384.16	789.61	550.76
9	83	79	3.6	8.1	12.96	65.61	29.16
10	67	68	-12.4	-2.9	153.76	8.41	35.96
Total	794	709	0	0	1818.4	5040.9	2272.4

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{n} = \frac{794}{10} = 79.4 \ , \ \bar{y} = \frac{\sum_{i=1}^{10} y_i}{n} = \frac{709}{10} = 70.9$$

$$\sum_{i=1}^{10} (x_i - \bar{x})^2 = 1818.4 \ , \ \sum_{i=1}^{10} (y_i - \bar{y})^2 = 5040.9 \ and \ \sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y}) = 2272.4$$

Then the correlation coefficient is:

$$\mathbf{r} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} ((x_i - \bar{x}^2} \cdot \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} = \frac{2272.4}{\sqrt{1818.4}\sqrt{5040.9}} = 0.75056 \approx 0.75$$

Alternatively, we can use the relation:

$$\mathbf{r} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\cdot\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\ \sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

i	x_i	x_i^2	y_i	y_i^2	$x_i \cdot y_i$
1	77	5929	82	6724	6314
2	54	2916	38	1444	2052
3	71	5041	78	6084	5538
4	72	5184	34	1156	2448
5	81	6561	47	2209	3807
6	94	8836	85	7225	7990
7	96	9216	99	9801	9504
8	99	9801	99	9801	9801
9	83	6889	79	6241	6557
10	67	4489	68	4624	4556
Total	794	64862	709	55309	58567

$$\mathbf{r} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\cdot\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\ \sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

$$= \frac{585670 - 794 \times 709}{\sqrt{648620 - (794)^2}\ \sqrt{553090 - (709)^2}} = 0.75056$$

Based on our rule, there is a strong positive linear relationship between *X* and *Y*. (The values of *Y* increase when the values of *X* increase).

SIMPLE LINEAR REGRESSION

Much of mathematics is devoted to studying variables that are deterministically related. Saying that X and Y are related in this manner means that once we are told the value of X, the value of Y is completely specified. For example, suppose the cost for a small pizza at a restaurant is SR10 plus SR 2 per topping. If we let $X =$ toppings and $Y=$ price of pizza, then $Y = 10 + 2X$. If we order a 3-topping pizza, then $Y = 10 + 2\left(3\right) = 16$ SR.

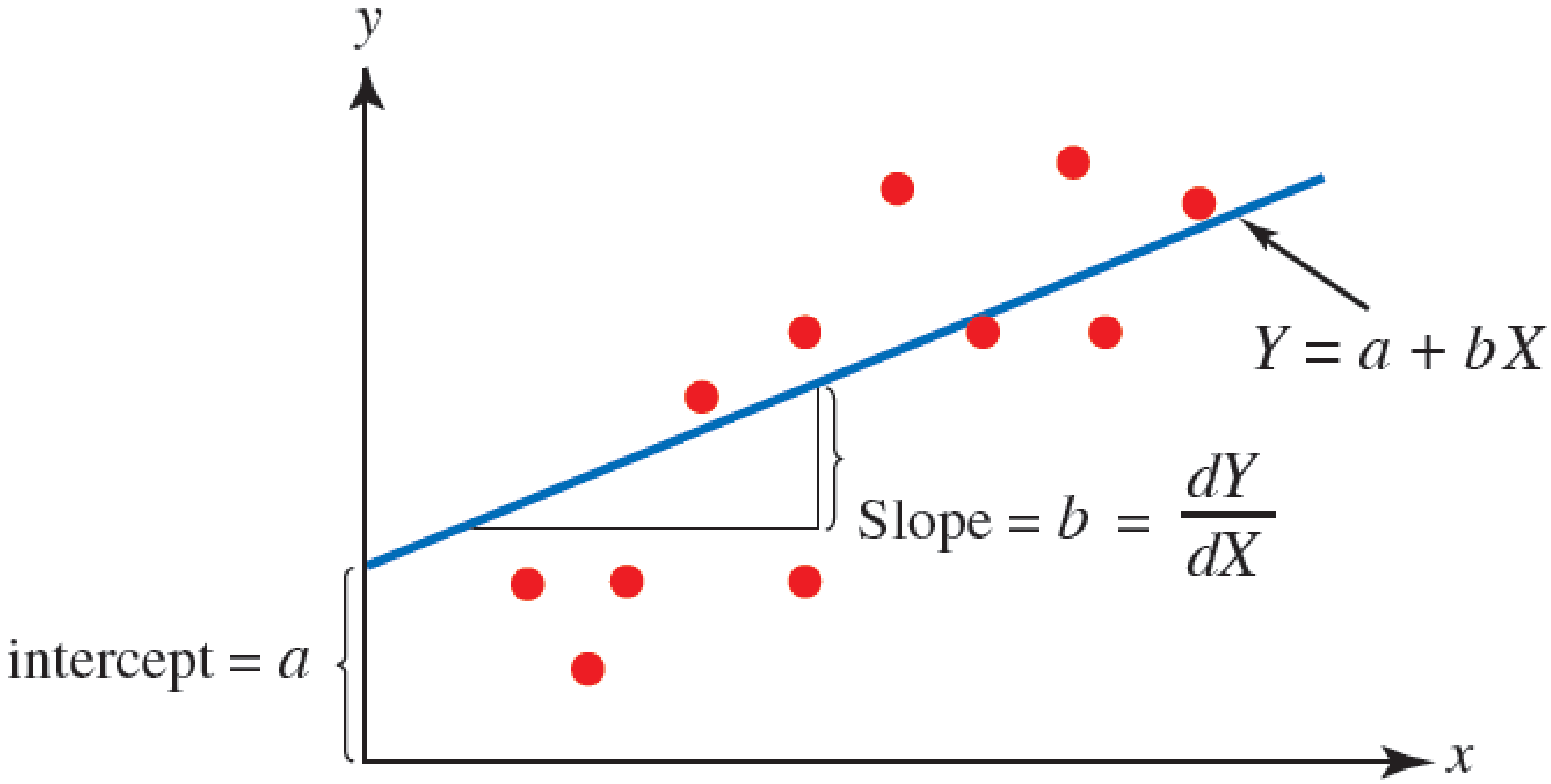
The simple linear regression line of a population describing the linear relationship between explanatory variable X and the response variable Y is given by the following relation:

$$\boxed{Y = a + bX + \mathcal{E}}$$

Where:

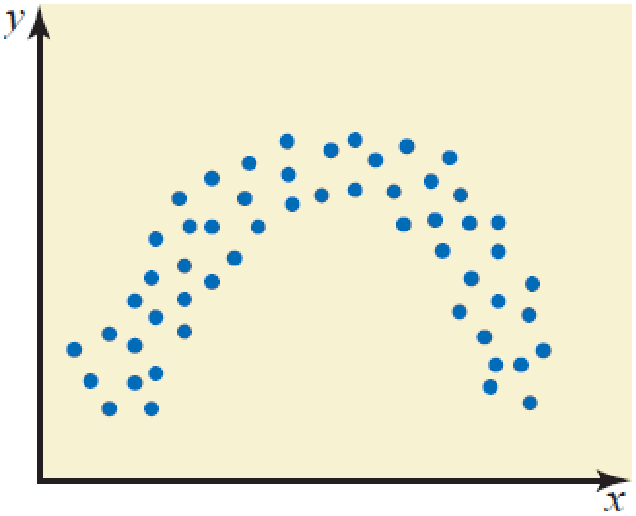
- $\mathcal{E}$ is a normal random variable with zero expectation $\mathrm{E}(\mathcal{E}) = 0$. This term $(\mathcal{E})$ in the form of simple regression line makes the regression analysis as a probabilistic approach.

- a and b are the parameters of the simple regression line, where a is a constant term (intercept) and b is the coefficient of the variable X (slope).

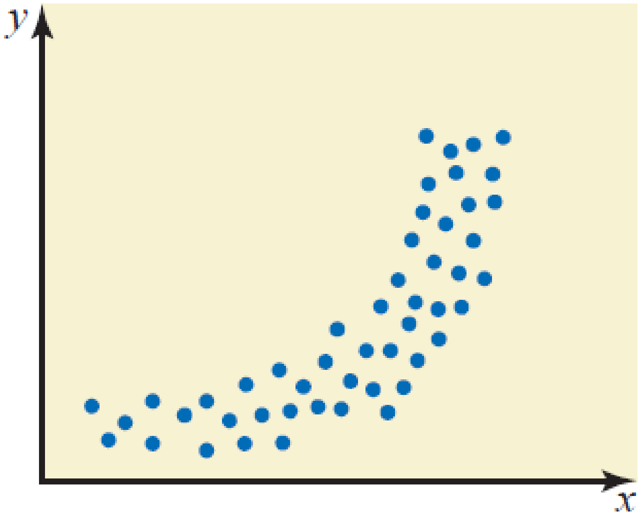

y
$Y = a + bX$
Slope = b = $\frac{dY}{dX}$
intercept = a
x

A Note on the Use of Simple Linear Regression

We should apply linear regression with caution. When we use simple linear regression, we assume that the relationship between two variables is described by a straight line. In the real world, the relationship between variables may not be linear. Hence, before we use a simple linear regression, it is better to construct a scatter diagram and look at the plot of the data points. We should estimate a linear regression model only if the scatter diagram indicates such a relationship. See this graph



(a)



(b)

THE METHOD OF LEAST SQUARES FOR ESTIMATING *a* and *b*

$$\hat{Y} = \hat{a} + \hat{b}\,X$$

where the coefficients $\hat{a}\ and\ \hat{b}$ cab be estimated as:

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \quad \text{Or by the relation} \quad \hat{b} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

and

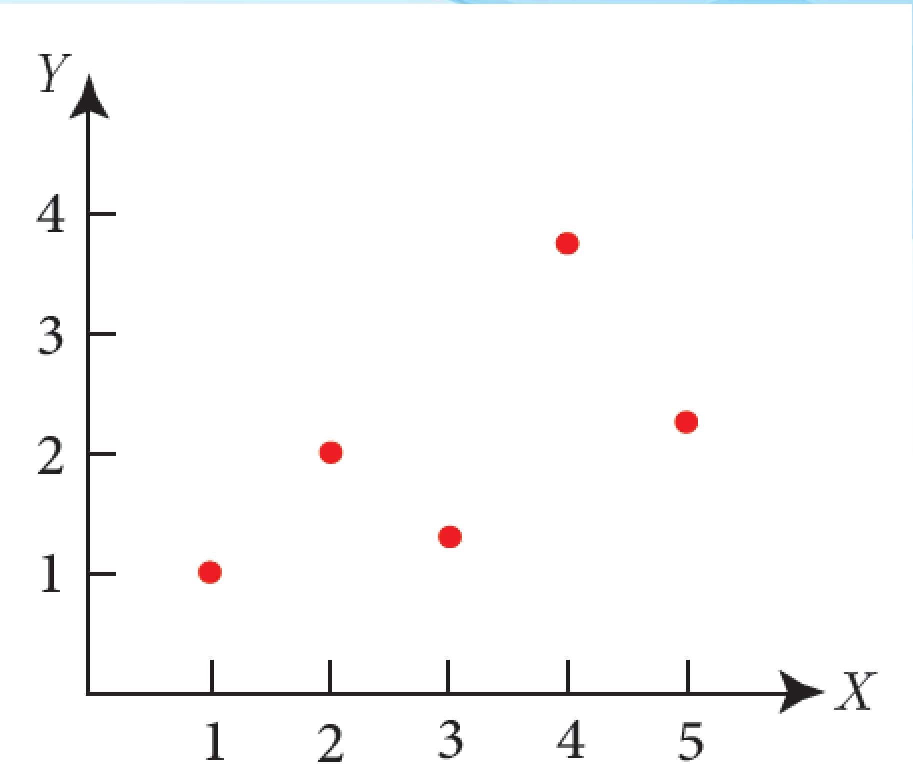
$$\hat{a} = \overline{y} - \hat{b}\,\overline{x} \quad \text{Or by the relation} \quad \hat{a} = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

EXAMPLE:

The example data given below:

X	Y
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25

Scatter plot:



You can see that there is a positive relationship between *X* and *Y*. If you were going to predict *Y* from *X*, the higher the value of *X*, the higher your prediction of *Y*.

a. Calculate the correlation coefficient between X and Y

b. Estimate the simple linear regression line $\hat{Y} = \hat{a} + \hat{b}X$

c. Find the value of Y when X=6?

a. From the given data, we have:

i	x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
1	1	1	1	1	1
2	2	2	4	4	4
3	3	1.3	9	1.69	3.9
4	4	3.75	16	14.0625	15
5	5	2.25	25	5.0625	11.25
------	$\sum x_i = 15$	$\sum y_i = 10.3$	$\sum x_i^2 = 55$	$\sum y_i^2 = 25.815$	$\sum x_i \cdot y_i = 35.15$

Then the linear correlation coefficient is given by:

$$\mathbf{r} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\cdot\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\ \sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

$$= \frac{5(35.15) - (15)(10.3)}{\sqrt{[5(55) - (15)^2][5(25.815 - (10.3)^2)]}} = 0.63$$

b. Linear regression interested in finding the best-fitting straight line through the points.

The best-fitting line is the simple regression line given by:

$$\hat{Y} = \hat{a} + \hat{b}\, X$$

The coefficients $\hat{b}$ and $\hat{a}$ can be estimated by using the forms:

$$\hat{b} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

And

$$\hat{a} = \overline{y} - \hat{b}\, \overline{x}$$

i	x_i	y_i	$(x_i - \overline{x})$	$(y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})(y_i - \overline{y})$
1	1	1	-2	-1.06	4	2.12
2	2	2	-1	-0.06	1	0.06
3	3	1.3	0	-0.76	0	0
4	4	3.75	1	1.69	1	1.69
5	5	2.25	2	0.19	4	0.38
Total	15	10.3	0	0	10	4.25

$$\bar{x} = \frac{15}{5} = 3 \text{ and } \bar{y} = \frac{10.3}{5} = 2.06$$

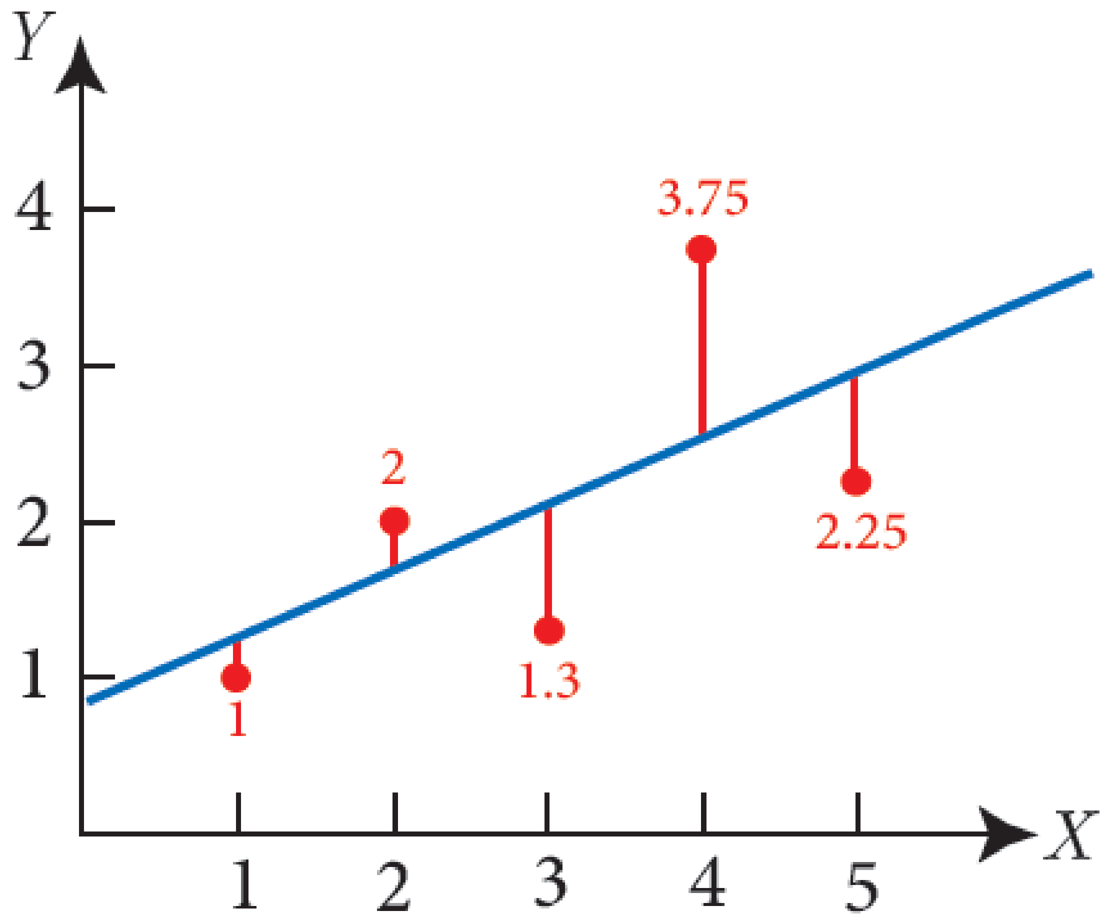
$$\hat{b} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{4.25}{10} = 0.425$$

And

$$\hat{a} = \bar{y} - \hat{b}\,\bar{x} = 2.06 - (0.425)(3) = 0.785$$

Hence, the estimated simple linear regression model is:

$$\hat{Y} = 0.785 + 0.425\,X$$


Y
4
3
2
1
3.75
2
2.25
1.3
1
1
2
3
4
5
X

c. To find the value of Y when X=6, we use the regression equation as follows:

$$\hat{Y} = 0.758 + 0.425X$$

When X=6 we have:

$$\hat{Y} = 0.758 + 0.425(6)$$

$$= 3.3$$

Chapter 10:

MULTIPLE REGRESSION

MULTIPLE LINEAR REGRESSION

$y = \beta_0 + \beta_1\, x_1$ Simple linear regression

$y = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \ldots + \beta_n\, x_n$ Multiple linear regression

y is the dependent variable and x_i are the independent variables

We will use the independent variables to predict the dependent variable

Example

A,B, and C are the independent products and cost is the dependent variable. The data are presented in the following table

Months	cost	A	B	C
1	44439	515	541	928
2	43936	929	692	711
3	44464	800	710	824
4	41533	979	675	758
5	46343	1165	1147	635
6	44922	651	939	901
7	43203	847	755	580
8	43000	942	908	589
9	40967	630	738	682
10	48582	1113	1175	1050
11	45003	1086	1075	984
12	44303	843	640	828
13	42070	500	752	708
14	44353	813	989	804
15	45968	1190	823	904
16	47781	1200	1108	1120
17	43202	731	590	1065
18	44074	1089	607	1132
19	44610	786	513	839

Solution and explanation

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.803398744
R Square	0.645449542
Adjusted R Square	0.57453945
Standard Error	1252.763898
Observations	19

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	42856229.89	14285409.96	9.102365067	0.001126532
Residual	15	23541260.74	1569417.383		
Total	18	66397490.63			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	35102.90045	1837.226911	19.10645889	6.11198E-12	31186.94398	39018.85691	31186.94398	39018.85691
A	2.065953296	1.664981779	1.240826369	0.23372682	-1.482871361	5.614777953	-1.482871361	5.614777953
B	4.176355531	1.681252566	2.484073849	0.025287785	0.592850514	7.759860548	0.592850514	7.759860548
C	4.790641037	1.789316107	2.677358695	0.017222643	0.976804034	8.604478041	0.976804034	8.604478041

1- Look at P-Values: if the p-value of a product is greater than 0.05 then that product does not make a significant effect on predicting the dependent variable.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.803398744
R Square	0.645449542
Adjusted R Square	0.57453945
Standard Error	1252.763898
Observations	19

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	42856229.89	14285409.96	9.102365067	0.001126532
Residual	15	23541260.74	1569417.383		
Total	18	66397490.63			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	35102.90045	1837.226911	19.10645889	6.11198E-12	31186.94398	39018.85691	31186.94398	39018.85691
A	2.065953296	1.664981779	1.240826369	0.23372682	-1.482871361	5.614777953	-1.482871361	5.614777953
B	4.176355531	1.681252566	2.484073849	0.025287785	0.592850514	7.759860548	0.592850514	7.759860548
C	4.790641037	1.789316107	2.677358695	0.017222643	0.976804034	8.604478041	0.976804034	8.604478041

Then, we will exclude using the values for the independent variable for product A. Product B and C both have p-values less than 0.05

So we need to rerun the multiple regression excluding the values of product A.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.780421232
R Square	0.609057299
Adjusted R Square	0.560189461
Standard Error	1273.715391
Observations	19

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	40439876.29	20219938.14	12.46335684	0.000545638
Residual	16	25957614.34	1622350.896		
Total	18	66397490.63			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	35475.30255	1842.860853	19.25012543	1.72346E-12	31568.61207	39381.99304	31568.61207	39381.99304
B	5.320968077	1.429095476	3.72331182	0.001849065	2.291421005	8.350515149	2.291421005	8.350515149
C	5.417137848	1.745311646	3.103822668	0.006825007	1.717242442	9.117033255	1.717242442	9.117033255

Predict the monthly cost for

1200 A models

800 B models

1000 C models

$$y = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \beta_3\, x_3$$

$$= 35475.30255+800*5.320968077+1000*5.417137848$$

=45149.21

2- Look at the 95% confidence interval: if the 95% C.I. of a product includes the zero, then this product does not make a significant effect on predicting the dependent variable.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.803398744
R Square	0.645449542
Adjusted R Square	0.57453945
Standard Error	1252.763898
Observations	19

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	42856229.89	14285409.96	9.102365067	0.001126532
Residual	15	23541260.74	1569417.383		
Total	18	66397490.63			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	35102.90045	1837.226911	19.10645889	6.11198E-12	31186.94398	39018.85691	31186.94398	39018.85691
A	2.065953296	1.664981779	1.240826369	0.23372682	-1.482871361	5.614777953	-1.482871361	5.614777953
B	4.176355531	1.681252566	2.484073849	0.025287785	0.592850514	7.759860548	0.592850514	7.759860548
C	4.790641037	1.789316107	2.677358695	0.017222643	0.976804034	8.604478041	0.976804034	8.604478041

Then, we will exclude using the values for the independent variable for product A when predicting the dependent variable (cost).

Applications using Excel 2019

Data 6

Multiple linear regression

Chapter 11:

ANALYSIS OF VARIANCE (ANOVA)

Analysis of Variance (ANOVA)

- One-way analysis of variance (abbreviated one-way ANOVA) is a technique that can be used to compare means of more than two samples (using the F distribution).
- The one-way ANOVA is used to test for differences among at least three groups, since the two-group case can be covered by a t-test.

$$H_0:\ \mu_1 = \mu_2 = \cdots = \mu_n$$

$$H_1:\ \mu_1 \neq \mu_2 \neq \cdots \neq \mu_n = \text{ at least two of them are not equal}$$

Example

A	B	C	D	E
25	25	24	20	14
21	28	24	17	15
21	24	16	16	13
18	25	21	19	11

$$H_0\colon\ \mu_A = \mu_B = \mu_C = \mu_D = \mu_E$$

H_1: at least two of them are not equal

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
A	4	85	21.25	8.25		
B	4	102	25.5	3		
C	4	85	21.25	14.25		
D	4	72	18	3.333333		
E	4	53	13.25	2.916667		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	331.3	4	82.825	13.04331	8.93E-05	3.055568
Within Groups	95.25	15	6.35			
Total	426.55	19				

We can make a decision in two ways:

1- Look at the P-value

if the p-value is greater than 0.05 then we can conclude that all means are equal. If not, we will accept the alternative hypothesis.

In this example:

$$P - value = 8.93E - 0.5 < 0.05$$

So we will accept the alternative hypothesis that is at least 2 means are different.

2- Look at the F value

if the F-crit is greater than F-value then we can conclude that all means are equal.

If not, we will accept the alternative hypothesis.

In this example:

$$F - crit = 3.06 < F - value = 13.04$$

So we will accept the alternative hypothesis that is at least 2 means are different.

Applications using excel 2019

Data 7

ANOVA