Chapter Nine: Linear Regression & Correlation

When you completed this chapter, you will be able to:
- ✓ identify the relationship between variables;
- ✓ understand the concepts of Least Squares Method;
- ✓ recognise the limitations of applying the Linear Regression Equation;
- ✓ calculate the Correlation Coefficient, Determination Coefficient and Rank Coefficient;
- ✓ interpret the meaning of the coefficients.

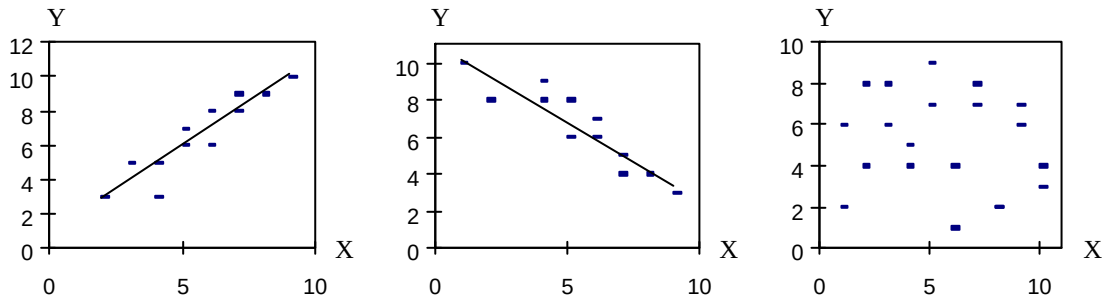
Reference(s): Mason Chapters 12 and 15 (p.563-565), Berenson Chapter 17, Owen Chapter 23
Exercise(s): Seminars 21 and 22, Mason Chapter12 Exercises 3, 5, 13, 23, 30,
Mason Chapter15 Exercise 25

Linear regression and correlation analysis is a study of determining both the ***nature*** and the ***strength*** of a relationship between two variables.

e.g.1	sales of soft drinks	outdoor temperature
	safety stock	lead time on order
	project completion time	no. of worker per project
	Dependent Variable (Y)	**Independent Variable (X)**

Regress Y (dependent variable) on X (independent variable)

Step 1: plot the data set to get an impression of the relationship between the variables, if exist. The data plot is called ***scatter*** *diagram*.



Direct Relationship
X↑, Y↑

Inverse Relationship
X↑, Y↓

No Relationship

Step 2: Determine the equation for the regression line.

$$\hat{y} = a + b\,x$$

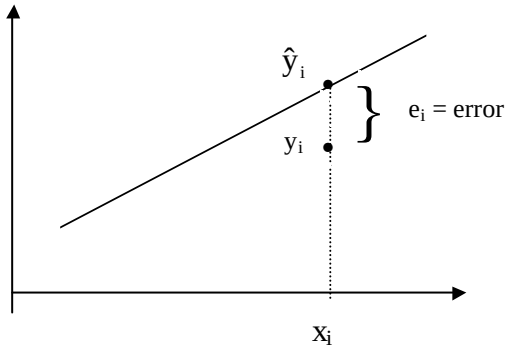
Estimated value of Y

y-intercept

slope of the regression line

Goodness of Fit for the Regression line

To minimize the error between the estimated points and the actual observed points.



Methods of Least Squares

Min $\sum e_i^2$

$S = \sum e_i^2$
$= \sum (y_i - \hat{y}_i)^2$
$= \sum (y_i - (a + bx_i))^2$

So, for minimize S, we want: $\partial S/\partial a = 0$, and $\partial S/\partial b = 0$
Hence,

$$b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{Sxy}{Sxx}$$

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

$$a = \bar{y} - b\,\bar{x}$$

then the equation of the regression line is:

$$\hat{y} = a + b\,x$$

e.g.2 An experiment was set up to investigate the variation of the specific heat of a certain chemical with temperature. Two measurements of the specific heat were taken at each of a series of temperature.

Temp ^oC	50	60	70	80	90	100
Specific heat	1.60 1.64	1.63 1.65	1.67 1.67	1.70 1.72	1.71 1.72	1.71 1.74

Find the regression line and estimate the specific heat when the temperature is 75^oC.

Sol] *Identify the dependent and independent variable,*

Temp - Independent variable (X).
Specific heat - Dependent variable (Y).

From the data set, we have

$n = 12,\ \sum x = 900,\ \sum y = 20.16,\ \sum xy = 1519.9,\ \sum x^2 = 71000.$

Hence, $\bar{x} = \sum x/n = 75,\ \bar{y} = \sum y/n = 1.68$

The slope of the regression line is:

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

$$= \frac{1519.9 - (12)(75)(1.68)}{71000 - (12)(75^2)} = 0.002257$$

Use your calculator to find those figures to verify the answers

The y-intercept of the regression line is:
$a = \bar{y} - b\,\bar{x}$
$= 1.68 - (0.002257)(75) = 1.5107$

$\therefore$ The Regression line: $\hat{y} = 1.5107 + 0.002257\,x$

Prediction using Regression line:

When $x = 75,\ \hat{y} = 1.5107 + 0.002257(75) = 1.68$ (correct to 2 decimal places)

The ***prediction*** of specific heat is 1.68 (units)

e.g.3 A random sample of eight introductory accounting texts yielded the figures shown in the table for annual sales (in thousands) and price (in dollars).

Sales	12.2	18.6	29.2	15.7	25.4	35.2	14.7	11.1
Price	29.2	30.5	29.7	31.3	30.8	29.9	27.8	27.0

Find the regression line and estimate the sales when the price is $30.

Sol] *Identify the dependent and independent variable,*

Sales - Dependent variable (Y).
Price - Independent variable (X).

$n = 8$, $\sum x = 236.2$, $\sum y = 162.1$, $\sum xy = 4825.35$, $\sum x^2 = 6989.16$.

Hence, $\bar{x} = \sum x/n = 29.5250$, $\bar{y} = \sum y/n = 29.2625$.

The slope of the regression line is:

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

$$= \frac{4825.35 - (8)(29.5250)(20.2625)}{6989.16 - (8)(29.5250)^2} = 2.5625$$

The y-interpret of the regression line is:

$a = \bar{y} - b\,\bar{x}$

$= 20.2625 - (2.5625)(29.5250) = -55.3959$

$\therefore$ The Regression line: $\hat{y} = -55.3959 + 2.5625\,x$

Prediction using Regression line:

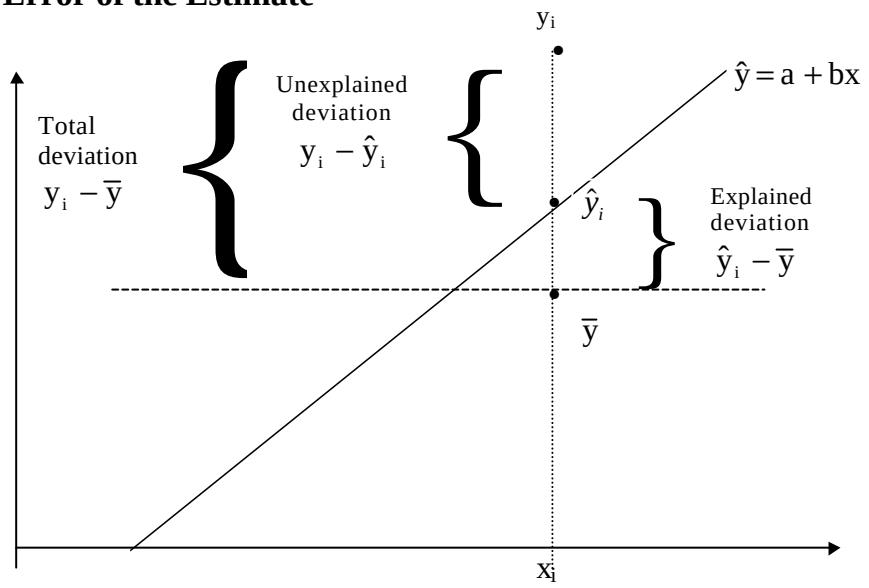
When $x = 30$, $\hat{y} = -55.3959 + 2.5625\,(30)$
$= 21.48$ (correct to 2 decimal places)

The ***prediction*** of sales is 21.48 (in thousands)

Remark: 1. The estimated regression equation is *valid only* ***over the same range*** *as the one from which the sample was taken* initially.

2. Even there is a relationship between X, and Y, it *does not imply X* ***causes*** *Y.*

The Standard Error of the Estimate



The difference between y_i and the mean of Y-value ($\bar{y}$) is often called the Total Deviation of y, $(y_i - \bar{y})$.

The deviation, $\hat{y}_i - \bar{y}$, is the part of the total deviation that is "explained" by the regression line, so it is called Explained Deviation.

The deviation, $y_i - \hat{y}_i$ is the error for the i^{th} sample observation, and since we have no basis for explaining why it occurred, so it is called Unexplained Deviation.

Total Deviation = Unexplained Deviation + Explained Deviation

$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$

$\Sigma(y_i - \bar{y})^2 = \Sigma(y_i - \hat{y}_i)^2 + \Sigma(\hat{y}_i - \bar{y})^2$

Total Variation = Unexplained Variation + Explained Variation

SST = SSE + SSR

The measure of the variability about the regression line is the value of SSE divided by its **degree of freedom**.

Since SSE is the sum of the squared errors, this measure is thus the sample variation of the values of e_i. The number of degrees of freedom in this measure is **n-2**. Hence,

Variance of estimates:

$$S_e^2 = \sum \frac{(y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}$$

Standard error of the estimates:

$$S_e = \sqrt{\frac{1}{n-2}\sum (y_i - \hat{y}_i)^2} = \sqrt{\frac{SSE}{n-2}}$$

where :

$$\begin{aligned} SSE &= \sum (y_i - \hat{y}_i)^2 \\ &= \sum [(y_i - (a + bx_i)]^2 \\ &= \sum [(y_i - \bar{y}) - b(x_i - \bar{x})]^2 \\ &= \sum (y_i - \bar{y})^2 - 2b\sum (x_i - \bar{x})(y_i - \bar{y}) + b^2 \sum (x_i - \bar{x})^2 \\ &= S_{yy} - 2bS_{xy} + b^2 S_{xx} \qquad \text{since } b = S_{xy}/S_{xx} \\ &= S_{yy} - bS_{xy} \end{aligned}$$

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x^2 - n\bar{x}^2$$
$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y^2 - n\bar{y}^2$$
$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum xy - n\bar{x}\bar{y}$$

Class Exercise 1

A marketing manager of a soft drink company is studying the effect of its latest advertising campaign. People chosen at random were called and asked how many cans of the soft drinks they had bought in the past week and how many times they had seen the advertisements in the past week. Information is shown below:

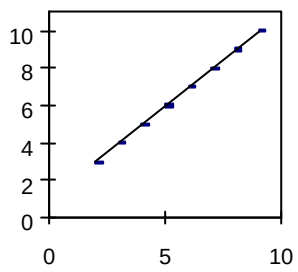
Number of Ads seen	2	6	9	6	4	9
Cans purchased	1	4	7	8	5	8

a) State the dependent and independent variables.
b) The manager would like to regress the number of cans purchased on the number of advertisements seen. Find the regression equation.
c) The manager used the regression equation to predict the number of cans purchased for a man who has seen the advertisements 15 times last week. What is the prediction? Is the prediction reliable? Why?

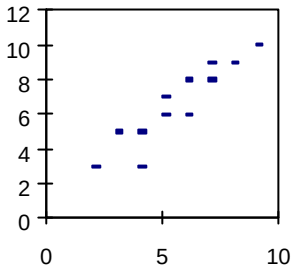
Linear Correlation Analysis

Linear Regression is concentrated on describing *the **nature** of the relationship* between two variables.

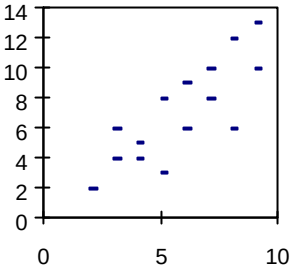
Linear Correlation is *to determine the **strength** of the linear relationship* between these variables. Strength refers to the degree of association between variables.



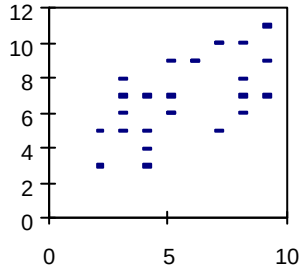
Very Strong Correlation



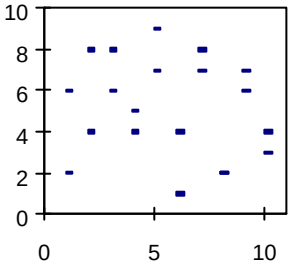
Strong Correlation



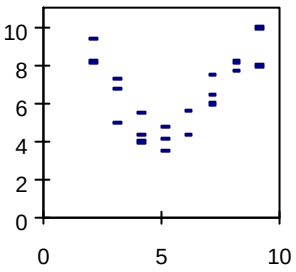
Medium Correlation



Week Correlation



No Correlation



No Linear Correlation

To measure the strength of relationship of two variables, the measurement *Population Correlation Coefficient, **r**,* is used.

$$\rho = \frac{\text{Covariance of x and y}}{(\text{s.d. of x})(\text{s.d. of y})}$$

$$= \frac{c[x, y]}{\delta_x \delta_y}$$

where : $c[x,y] = E[(x- \bar{x})(y- \bar{y})]$

Remark : if two variables are independent, then c[x,y] = 0

In all estimation problems, we use a sample statistic to estimate a population parameter. The sample statistics is called the *Sample Correlation Coefficient, r*.

$$r = \frac{\text{sample covariance of x and y}}{(\text{sample s.d. of x})\,(\text{sample s.d. of y})}$$

$$= \frac{c_s[x, y]}{s_x s_y}$$

$$= \frac{\sum (x-\bar{x})(y-\bar{y}) \Big/ (n-1)}{\sqrt{\sum (x-\bar{x})^2 \Big/ (n-1)}\sqrt{\sum (y-\bar{y})^2 \Big/ (n-1)}}$$

$$= \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

r must range from ***-1*** to ***+1***. *Negative values corresponding to lines with* ***negative*** *slopes*, i.e. x ↑, y ↓. *Positive values corresponding to lines with* ***positive*** *slopes*, i.e. x↑, y↑.

A perfect linear relationship appears in the sample when $r = \pm 1$, where all sample points lie on a straight line.

If $|r| \geq 0.7$, then a ***strong*** linear relationship can be concluded, otherwise ***weak*** relationship is concluded.

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

$$r^2 = \frac{S^2_{xy}}{S_{xx} S_{yy}}$$

$$= \frac{bS_{xy}}{S_{yy}} \qquad \text{since } b = S_{xy}/S_{xx}$$

$$= \frac{S_{yy} - (S_{yy} - bS_{xy})}{S_{yy}}$$

$$= \frac{S_{yy}}{S_{yy}} - \frac{S_{yy} - bS_{xy}}{S_{yy}}$$

$$= \frac{SST}{SST} - \frac{SSE}{SST}$$

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

$$= \textbf{\textit{Coefficient of Determination}}$$

Coefficient of Determination, r^2, is *the **percentage of data variation explained** by the regression equation.*

If $r^2 \geq 0.5$, it means ***at least 50%*** of data variation is explained by the estimated regression line, the regression equation is concluded to be ***good fit*** *for the sample data.*

e.g.4 Refer to e.g. 2,

Temp °C	50	60	70	80	90	100
Specific heat	1.60	1.63	1.67	1.70	1.71	1.71
	1.64	1.65	1.67	1.72	1.72	1.74

Calculate the sample coefficient of determination and the sample coefficient of correlation. And then interpret the answers.

Sol] use X represents the temp
and Y represents specific heat.

Coefficient of Determination =

$$r^2 = \frac{(\sum xy - n\bar{x}\bar{y})^2}{(\sum x^2 - n\bar{x}^2)(\sum y^2 - n\bar{y}^2)}$$

we have, n = 12,

$\sum x = 900$, $\sum y = 20.16$, $\sum xy = 1519.9$,
$\sum x^2 = 71000$, $\sum y^2 = 33.8894$

$$r^2 = \frac{[(1519.9) - (12)(900/12)(20.16/12)]^2}{[71000 - (12)(900/12)^2][33.8894 - (12)(20.16/12)^2]}$$
$$= \frac{62.41}{(3500)(0.0206)}$$
$$= 0.8656$$

$\Rightarrow$ <u>86.56% of the sample variability in Specific Heat is explained</u> by its linear dependence Temperature.
$\Rightarrow$ since $r^2 = 0.8656 > 0.5$, <u>the regression line is fit for the sample data</u>.

Coefficient of Correlation =

$$r = \sqrt{r^2}$$
$$= \sqrt{0.8656}$$
$$= 0.93$$

$\Rightarrow$ <u>strong linear positive relationship</u> between Specific Heat and Temperature.

e.g.5 The following data represent the price of a six-pack of Coca-Cola and a pound of chicken in a supermarket in 10 selected districts.

District	Price	
	Coca Cola	Chicken
Kwun Tong	2.89	2.57
Kowloon Bay	2.08	2.06
Kowloon Tong	2.07	2.49
Mongkok	2.55	1.49
T.S.T	3.61	2.75
Central	2.14	1.68
Tsuen Wan	2.85	1.29
Chai Wan	2.29	2.17
Tai Po	5.01	3.99
Shatin	2.47	1.90

Calculate the sample coefficient of determination and the sample coefficient of correlation. And then interpret the answers.

Sol] Use X represents the price of Coca-Cola,

and Y represents the price of chicken.

$$\text{Coefficient of Determination} = r^2 = \frac{(\sum xy - n\bar{x}\bar{y})^2}{(\sum x^2 - n\bar{x}^2)(\sum y^2 - n\bar{y}^2)}$$

we have, $n = 10$,

$\sum x = 27.96$, $\sum y = 22.39$, $\sum xy = 67.5173$,

$\sum x^2 = 85.6452$, $\sum y^2 = 55.5567$.

$$r^2 = \frac{[(67.5173) - (10)(2.796)(2.239)]^2}{[85.6452 - (10)(2.796)^2][55.5567 - (10)(2.239)^2]}$$

$$= \frac{(4.91486)^2}{(7.46904)(5.42549)}$$

$= 0.5961$

$\Rightarrow$ 59.61% of the sample variability in Y is explained by its linear dependence X.
$\Rightarrow$ since $r^2 = 0.5961 > 0.5$, the regression line is fit for the sample data.

Coefficient of Correlation =

$r = \sqrt{r^2}$

$= \sqrt{05961}$

$= 0.7721$

$\Rightarrow$ strong linear positive relationship between X and Y.

Spearman' s Coefficient of Rank Correlation

Rank Correlation Coefficient, r_s, measures the degree of correlation that exists between ***two sets of ranks*** rather than their ***actual numerical values***.

To calculate r_s, first rank the x' s among themselves, giving rank 1 to the largest (or smallest), rank 2 to the second largest (or smallest), and so on. Then rank the y' s similarly. Find the sum of the squares of the difference, d, between the ranks of x' s and y' s .

The Rank Correlation Coefficient:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

r_s must also range from ***-1*** to ***+1***, it is interpreted as same as r.

Remark: For instance, if the third and fourth largest values of the variables are the same, there are lies in rank, assign each the rank (3+4)/2 = 3.5.

e.g.6 Calculate Spearman' s Coefficient of Rank Correlation for the following data, and then interpret the answer.

Number of hours studied	8	5	11	13	10	5	18	15	2	8
Grade in Examination	56	44	79	72	70	54	94	85	33	65

Sol]

Number of hours studied	Grade in Examination	Rank of x	Rank of y	d	d^2
8	56	4.5	4	0.5	0.25
5	44	2.5	2	0.5	0.25
11	79	7	8	-1.0	1.00
13	72	8	7	1.0	1.00
10	70	6	6	0.0	0.00
5	54	2.5	3	-0.5	0.25
18	94	10	10	0.0	0.00
15	85	9	9	0.0	0.00
2	33	1	1	00	0.00
8	65	4.5	5	-0.5	0.25
				$\sum d^2$	3.00

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

$$= 1 - \frac{6(3.00)}{10(10^2-1)}$$

$$= 0.98$$

⇒ strong linear positive relationship between hours of study and examination result.

e.g. 7 Refer to e.g.5, calculate Spearman' s Coefficient of Rank Correlation, and then interpret the result.

Sol]

Price of coca-cola (x)	Price of chicken (y)	Rank of x	Rank of y	d	d^2
2.89	2.57	8	8	0	0
2.08	2.06	2	5	-3	9
2.07	2.49	1	7	-6	36
2.55	1.49	6	2	4	16
3.61	2.75	9	9	0	0
2.14	1.68	3	3	0	0
2.85	1.29	7	1	6	36
2.29	2.17	4	6	-2	4
5.01	3.99	10	10	0	0
2.47	1.90	5	4	1	1
				$\sum d^2$	102

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6(102)}{(10)(10^2-1)}$$

$$= 0.3818$$

⇒ weak linear positive relationship between price of coca-cola and price of chicken.

Class Exercise 2

Refer to class exercise 1,

a) compute the Coefficient of Correlation, and comment on the nature of the relationship between these variables,

b) what percentage of the data variation of the number of cans purchased is explained by the regression equation?

Class : _____ **Name** : _____ **No.** : _____

Class Exercise 1 (Solution)

a) dependent variable = cans purchases (Y)

independent variable = no. of advertisements seen (X)

b) $\sum x = 36$, $\sum y = 33$,
$\sum x^2 = 254$, $\sum y^2 = 219$, $\sum xy = 229$, $n = 6$

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

$= [229 - (6)(6)(5.5)] / [254 - (6)(6)^2]$

$= 0.8158$

$a = \bar{Y} - b\bar{X}$

$= 5.5 - (0.8158)(6)$

$= 0.6053$

$\hat{y} = 0.6053 + 0.8158\ x$

c) when X = 15

$\hat{y} = 0.6053 + 0.8158\ (15)$

$\hat{y} = 12.84$ (cans)

No, the prediction is not reliable,

since the X-value is out of the data range (X = 2 to 9) from which the regression equation is established.

Class : _____ **Name** : _____ **No.** : _____

Class Exercise 2 (Solution)

a) $r = \dfrac{(\sum xy - n\bar{x}\bar{y})}{\sqrt{(\sum x^2 - n\bar{x}^2)(\sum y^2 - n\bar{y}^2)}}$

$= [229 - (6)(6)(5.5)] / \sqrt{[(254 - (6)(6^2))(219 - (6)(5.5^2))]}$

$= 0.8212$

$\Rightarrow$ strong positive linear relationship

b) $r = 0.8212$,

$r^2 = 0.8212^2 = 0.6743$

$\Rightarrow$ 67.43% of data variation can be explained by the regression equation.
$\Rightarrow$ since $r^2 = 0.6743 > 0.5$, the regression line is fit for the sample data.