

SECTION 5

REGRESSION AND CORRELATION

5.1 INTRODUCTION

In this section we are concerned with relationships between variables. For example:

- How do the sales of a product depend on the price charged?
- How does the strength of a material depend on temperature?
- To what extent is metal pitting related to pollution?
- How strong is the link between inflation and employment rates?
- How can we use the amount of fertiliser used to predict crop yields?

These are essentially two types of problem:

- (i) **CORRELATION** problems which involve measuring the *strength* of a relationship.
- (ii) **REGRESSION** problems which are concerned with the *form* or *nature* of a relationship.

Exercise

Are the five examples above correlation or regression problems?

Answers

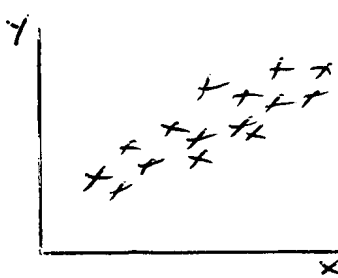
- (i) *Regression*
- (ii) *Regression*
- (iii) *Correlation*
- (iv) *Correlation*
- (v) *Regression*

We will be concerned with relationships between just two variables X and Y which is referred to as **simple** correlation and **simple** regression.

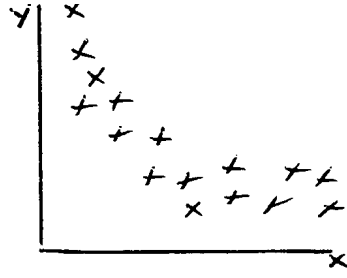
The first step in an analysis should be to plot a **SCATTER DIAGRAM** to visually assess the strength and form of the relationship. (See section 1.4.5 on page 24 for some examples).

5.2 TYPES OF RELATIONSHIP

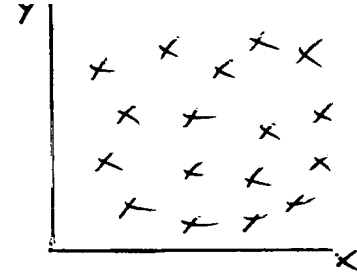
The relationship between two quantitative variables may be described as positive or negative, linear or non-linear.



*Positive linear
(straight line)
relationship*

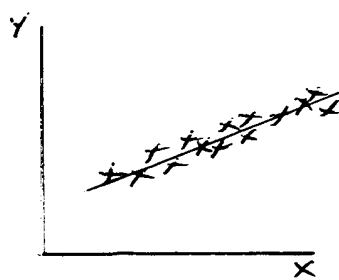


*Negative non-linear
(not straight line)
relationship*

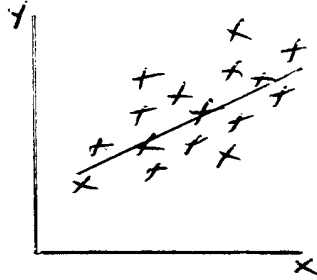


No relationship

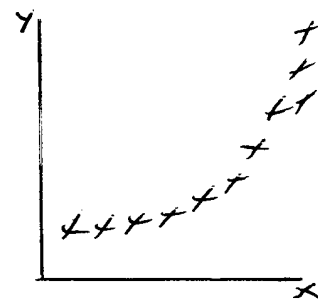
The 'strength' of the relationship refers to the closeness of the points to the underlying curve or straight line.



*Strong positive
linear
relationship*



*Weak positive
linear
relationship*



*Strong non-linear
relationship*

Most of this chapter is concerned with **linear** relationships. Simple linear correlation is covered in 5.3 and simple linear regression in 5.4. An introduction to non-linear relationships is covered in section 5.7.

5.3 CORRELATION

5.3.1 The Correlation Coefficient, r

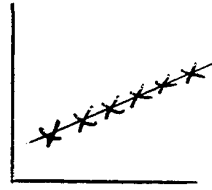
The strength of a relationship can be measured by a correlation coefficient. One of the most widely used is Pearson's Product Moment Correlation Coefficient, denoted r , which provides a measure of the strength of *linear association*.

This measure is independent of scale and always lies in the range -1 to $+1$;

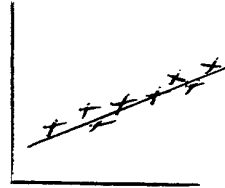
- -1 if there is a perfect negative linear relationship
- $+1$ if there is a perfect positive linear relationship.

Some illustrations showing scatter diagrams together with values of r are:

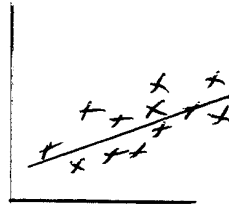
$r = 1$ Perfect positive linear association



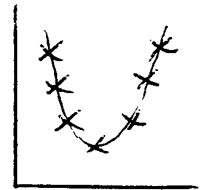
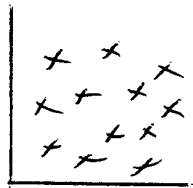
$r = 0.8$ Strong positive linear association



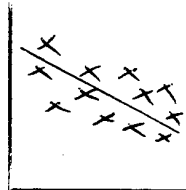
$r = 0.4$ Weak positive linear association



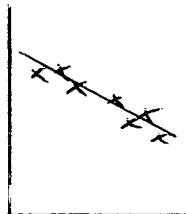
$r = 0$ No linear association



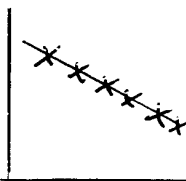
$r = -0.4$ Weak negative linear association



$r = -0.8$ Strong negative linear association



$r = -1$ Perfect negative linear association



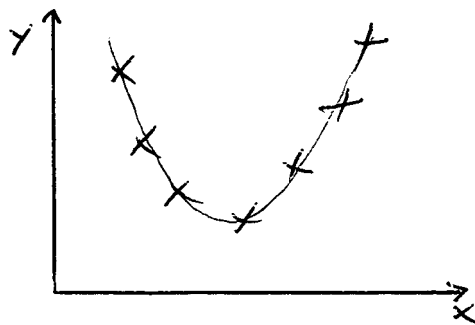
5.3.2 Calculation of r

To calculate r either:

- (i) Use the computational formula for hand calculations - see the example in section 5.6.1.
- (ii) Use a package such as Minitab or a spreadsheet such as Excel.
- (iii) Use the correlation button (usually labelled r) available on some calculators.

5.3.3 Interpretation of r

- (i) r measures the **LINEAR** relationship between 2 variables. If a relationship is non-linear, r can be very misleading.



Strong relationship
(non-linear) but data
would give a value of
r close to zero.

Plot a scatter diagram first. If the relationship is non-linear, it may be possible to make it linear by transforming one or both of the variables (eg. take logs) as described in section 5.7.

- (ii) A high value of r (close to +1 or -1) **DOES NOT IMPLY ANY CAUSAL LINK** between the variables. It just means that if one variable is relatively high in value then the other tends to be relatively high as well (or low if r is negative). There may be a third variable which is causing the changes in both X and Y. Examples of these so-called **spurious correlations** are given below.

X : milk consumption per capita }
Y : incidence of cancer } Common factor = degree of development

X : birth rates in Holland }
Y : size of stork population } Common factor = time

Correlation coefficients can suggest where causes may lie but to show that changing one variable **causes** changes in the other variable requires a *controlled experiment*.

5.3.4 Testing the Significance of r

r measures the strength of relationship found in a sample. An important question is whether the size of r suggests the variables really are related (in the population). For example, if data on the heights and IQ's of 40 individuals yields a correlation coefficient of 0.08 then, although r is not exactly zero, the only sensible conclusion is that height and IQ are really not related. On the other hand, a correlation of 0.99 calculated from 40 pairs of values would clearly suggest that the variables really are related. What about less extreme values - say a correlation of 0.6 or -0.4 ? Are these just chance sampling effects?

The table below gives critical values of r (ignoring the sign). A calculated coefficient must be larger than the critical value before we can conclude that the variables really are related.

Table 5.1 Critical values for the Pearson Product Moment Correlation Coefficient, r at the 5% level.

Number of pairs of observations n	Critical Value r^*
4	0.950
5	0.878
6	0.811
7	0.754
8	0.707
9	0.666
10	0.632
12	0.576
14	0.532
16	0.497
18	0.468
20	0.444
25	0.397
30	0.361
35	0.335
40	0.312
50	0.279

As an example of using the above table, suppose that we had 16 observations on each of our variables (ie. $n = 16$). We would then require the correlation coefficient to be either greater than 0.497 or less than -0.497 in order to conclude that there was a linear relationship between the variables.

Note that the fewer the observations we have on each variable, the closer we require r to be to ± 1 in order to conclude that our variables are linearly related. Conversely, the more data we have, the less extreme r has to be.

Exercise

Can you conclude that in the example in 5.6 really are related?

*[Answer: The calculated correlation is $r = +0.855$ and, since $n = 10$, the tabulated critical value is 0.632. Thus the correlation is significantly high and we can conclude that production and spoilage rates **really are related.**]*

5.4 REGRESSION

5.4.1 The Dependent Variable and the Explanatory Variable

Here we are concerned with the **form** of the relationship between the variables. This can be summarised by an equation that enables us to predict or estimate the values of one variable (Y - the **dependent** variable) given values of the other variable (X - the independent or **explanatory** variable). The first thing to do then is to decide which variable is which. To summarise:

Regression analysis is concerned with how the values of Y depend on the corresponding values of X. The variable that is eventually to be predicted or estimated should be labelled Y.

Exercise

Which are the dependent variables in examples 1, 2 and 5 in section 5.1.

[Answers: Ex1 Sales; Ex2 Extent of borrowing; Ex3 Crop yield.]

5.4.2 Finding the Best-Fitting Line

In practice, most relationships are not exact - there will be a certain amount of scatter about a line. We require the equation of the line that fits the data best. This could be done by fitting a line by eye but a better approach is to use the **Method of Least Squares**. This is a mathematical technique that gives the slope and the intercept of the best-fitting line in the sense that it minimises the errors involved in using the line to predict values of Y.

In order to find the slope, b, and the intercept, a, of this **least-squares regression line** which in general is denoted:

$$Y = a + bX$$

then, as with calculating r in 5.3.2, either:

- (i) Use the computational formulae for hand calculations - see the example in section 5.6.2.
- (ii) Use a package such as Minitab or spreadsheet such as Excel.
- (iii) Use the regression buttons (usually labelled a and b) available on some calculators.

To illustrate how to interpret and use the line, consider the following.

Example

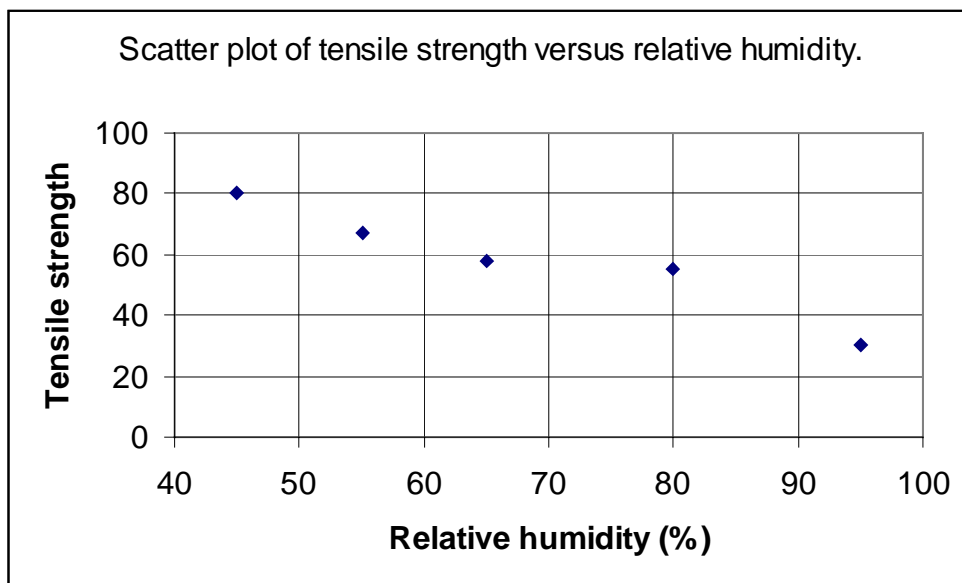
It is suspected that there is some relationship between the relative humidity and tensile strength of a certain material. The following measurements are obtained.

Relative humidity (%)	Tensile strength
45	80
55	67
65	58
80	55
95	30

Investigate the relationship between relative humidity and tensile strength if it is required to be able to predict tensile strength for various levels of relative humidity.

First the data are plotted in a scatter diagram.

The **dependent** variable (in this case *Tensile Strength*) is plotted on the vertical axis and the **explanatory** (or independent or regressor) variable (in this case *Relative Humidity*) is plotted on the horizontal axis.



From the plot it is seen that there is a clear negative relationship between the variables, since as relative humidity increases, tensile strength decreases. The relationship can also be seen to be linear since the plotted points scatter around a straight line.

The equation of the regression line for predicting tensile strength can be found to be

$$Y = 118.9 - 0.90X$$

or Estimated tensile strength = $118.9 - 0.90 \times$ relative humidity(%).

5.4.3 Interpretation of a and b

(i) ***Slope***

In general this tells us how we expect Y to change, on average, if X is increased by 1 unit

In this example, $b = -0.90$. Thus, for every additional % in relative humidity, tensile strength will *decrease* by an average of 0.90 units.

(If the slope were *positive*, we would expect Y to *increase* as X increases.)

(ii) ***Intercept***

This is the value of Y predicted for $X = 0$.

In this example, $a = 118.9$ which means that at zero relative humidity, tensile strength is estimated to be 118.9. In most applications, the intercept has no useful practical interpretation. It just serves to fix the line.

5.4.4 Predictions

Predicted values of Y for given values of X are obtained by simple substituting the value of X into the regression equation.

For example, the tensile strength for 70% relative humidity, is predicted to be

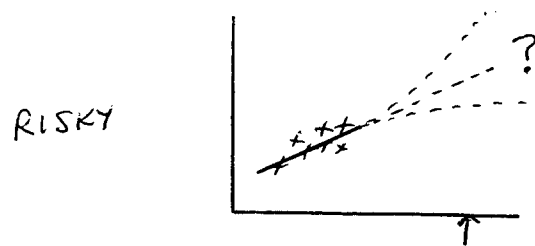
$$\begin{aligned} Y &= 118.9 - (0.90)(70) \\ &= 118.9 - 63 = \mathbf{55.9 \text{ units.}} \end{aligned}$$

and the tensile strength for 25% relative humidity, is predicted to be

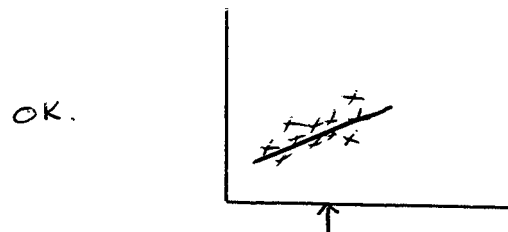
$$\begin{aligned} Y &= 118.9 - (0.90)(25) \\ &= 118.9 - 22.5 = \mathbf{96.4 \text{ units.}} \end{aligned}$$

Extrapolation and Interpolation

The last prediction for a relative humidity of 25% is EXTRAPOLATION since we are predicting for a value of X outside the range of the data (which is for relative humidity values between 45% and 95%). Effectively we are assuming that the linear relationship continues outside the observed range. Unless we have firm evidence to support this, we should treat this prediction very cautiously



On the other hand, the first prediction for 70% is well within the observed range and is an INTERPOLATION. The prediction here should be fairly accurate.



5.5 A FULL REGRESSION ANALYSIS

5.5.1 Checklist of Stages

Stage 1 Define the Variables

Which is the dependent variable (Y) and which is the explanatory variable (X)?

Stage 2 Plot a Scatter Diagram

Is the relationship linear? If not, see section 5.7. There is little point in finding the equation of the best-fitting *line* if you should really be fitting a *curve*.

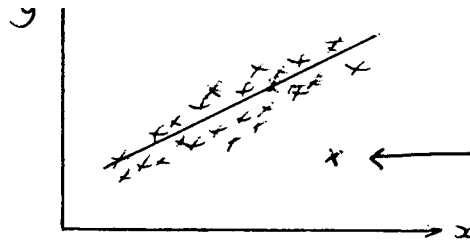
Also look for any 'strange' points lying a long way from the rest.

Unusual observations (or 'outliers'.)

These can be of two types:

– Large residual

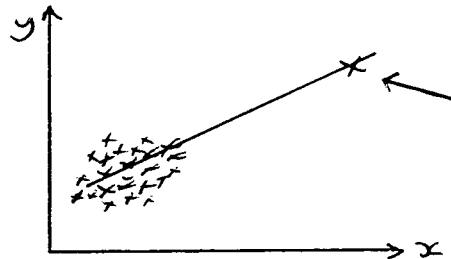
i.e. An anomalous point which must be investigated. It is quite likely to be an error.



– Influential observation

i.e. A point having a very large effect on the final result.

Re-run the analysis without it.



Stage 3 Obtain the Equation of the Regression Line

Use a package or calculator to find the slope and intercept of the line of best fit.

Stage 4 Interpretation

From the output, identify and interpret:

- the significance of the regression
- the slope (and possibly the intercept)
- the R^2 statistic.

These will be explained in the example in 5.5.2.

Stage 5 Examine the Residuals

A residual is the difference between the observed value for Y at a given value of X and the corresponding value predicted by the regression equation for that value of X. They can be used to decide whether the fitted regression equation is reasonable as in the example below.

Stage 6 Carry out any required Predictions

If everything seems OK, predict Y for given values of X as required, being mindful of *extrapolating* too far beyond the range of data.

5.5.2 Worked Example using Excel

A major airline wants to estimate the relationship between the number of reservations and the actual number of passengers who show up for flight XYZ. They can use this information to estimate how many reservations the airline should take for the flight in the future.

Information gathered over 12 randomly selected days for flight XYZ is given in the table below:

<i>Day</i>	<i>No. of Reservations</i>	<i>No. of Passengers</i>
1	250	210
2	548	405
3	156	120
4	121	89
5	416	304
6	450	320
7	462	319
8	508	410
9	307	275
10	311	289
11	265	236
12	189	170

The airline is particularly interested in how many passengers to expect for a day on which 350 reservations have been made, and for a day on which 80 reservations have been made.

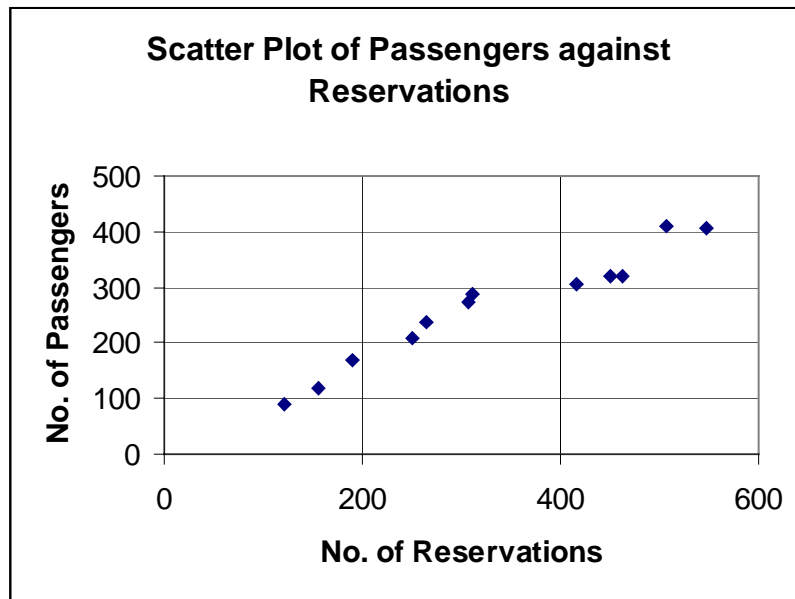
Stage 1

The *dependent variable* in this example is *the number of passengers*.

and the *explanatory variable* is *the number of reservations*.

Stage 2

A scatter plot of the number of passengers against the number of reservations is given below. It is clear from the diagram that there is a positive relationship between the variables (ie. as the number of reservations increases so does the number of passengers). It is also reasonably clear that there is a linear trend in the data.



Stage 3

The Excel output for this problem is as follows:

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.9692
R Square	0.9393
Adjusted R Square	0.9332
Standard Error	26.2649
Observations	12

(Note that the 'Multiple R' value is just the correlation coefficient.)

C

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	106665.788	106665.788	154.623	0.000
Residual	10	6898.462	689.846		
Total	11	113564.250			

B

	Coefficients	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	33.0721	19.929	1.659	0.128
No. of Reservations	0.690468	0.056	12.435	0.000

A

Much of the above output is not relevant to this course. The three things to look out for are indicated.

The intercept of the line of best fit is 33.0721 and the slope is 0.690468. Hence the regression line is given by

A → $Y = 33.0721 + 0.690468X$

or (after rounding)

$$\text{Number of passengers} = 33.1 + 0.690 \times \text{Number of reservations.}$$

Stage 4

Significance

B → This is a probability that indicates whether the relationship is real.

If '**significance**' ≤ 0.05 , the regression is statistically significant at the 5% level. The line does represent a real relationship and we would PROCEED with the rest of the analysis.

If '**significance**' > 0.05 , the regression is not significant. There is no evidence of a real relationship between the variables and the analysis would STOP here.

(Actually this is exactly the same as using Table 5.1 to test the correlation coefficient - denoted Multiple R in the Excel output.)

In our example, the significance of 0.000 is less than 0.05 so we proceed to interpret and use the line.

Interpretation of the Slope

$$b = 0.690$$

This indicates that every extra reservation leads to an additional 0.690 of a passenger. Put another way, for every additional 100 reservations we would expect (approximately) an additional 69 passengers.

Interpretation of the Intercept

$$a = 33.1$$

This indicates that on a day on which no reservations are made we would expect 33.1 passengers. Again, this is not particularly useful.

C

The R² Statistic

This statistic is simply the correlation coefficient squared and multiplied by 100 to convert to a percentage.

It gives the percentage of variation in the dependent variable (Y) that can be explained by variation in the explanatory variable (X).

Here R² = 93.9%.

This indicates that 93.9% of the variation in the dependent variable, number of passengers, is explained by the explanatory variable, number of reservations. In other words, 93.9% of the variation in the number of passengers on these 12 flights can be attributed to the fact that different numbers of reservations were made. This leaves only 6.1% to be explained by other factors (including pure chance effects). This suggests that the model gives a good fit to the data.

In general, the closer R² is to 100%, the better is the model. An R² of 100% indicates a perfect model.

Stage 5

Residuals are defined as:

$$\text{Residual} = (\text{Actual value of Y}) - (\text{Predicted value of Y})$$

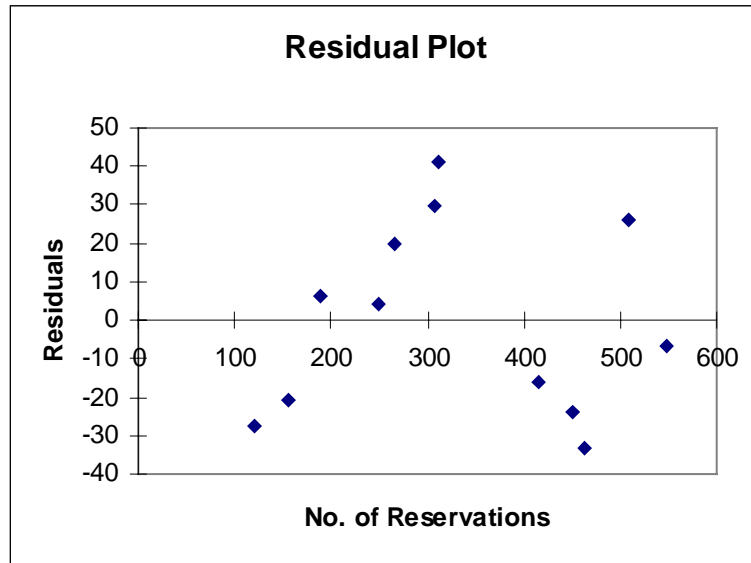
To assess the reasonableness of the fitted regression equation, these residuals should be plotted against the corresponding values of X. If the residual plot shows *random scatter about zero* then the regression equation can be used for prediction etc. If there is some obvious pattern in the residual plot, it may be that the relationship is not linear and a transformation may help (see section 5.7). The *residuals* and *residual plot* for this passenger data are:

Observation	X	Actual No. of Passengers	Predicted No. of Passengers	Residuals
1	250	210	205.7	4.3
2	548	405	411.4	-6.4
3	156	120	140.8	-20.8
4	121	89	116.6	-27.6
5	416	304	320.3	-16.3
6	450	320	343.8	-23.8
7	462	319	352.1	-33.1
8	508	410	383.8	26.2
9	307	275	245.0	30.0
10	311	289	247.8	41.2
11	265	236	216.0	20.0
12	189	170	163.6	6.4

=210 – 205.7

=405 – 411.4

etc.



There is no obvious pattern here so a linear relationship looks reasonable and we can carry on to stage 6 if required.

Stage 6

Predictions

From the line of best fit, for a day on which 350 reservations were made, the predicted number of passengers

$$= 33.1 + (0.690)(350) = 274.6$$

So we would expect **275** passengers on flight XYZ.

For a day on which 80 reservations were made, the predicted number of passengers

$$= 33.1 + (0.690)(80) = 88.3$$

So we would now expect **88** passengers on flight XYZ.

The first prediction is an example of interpolation and so should be acceptable since R^2 is high. The second prediction is an example of extrapolation and so it is based on the assumption that the relationship between the variables remains linear down to 80 reservations. In this case we need to be more cautious. The problem here is that whilst the relationship between our variables is linear (or close to linear) for number of reservations in the range 121-548, the relationship may not be linear outside this range.

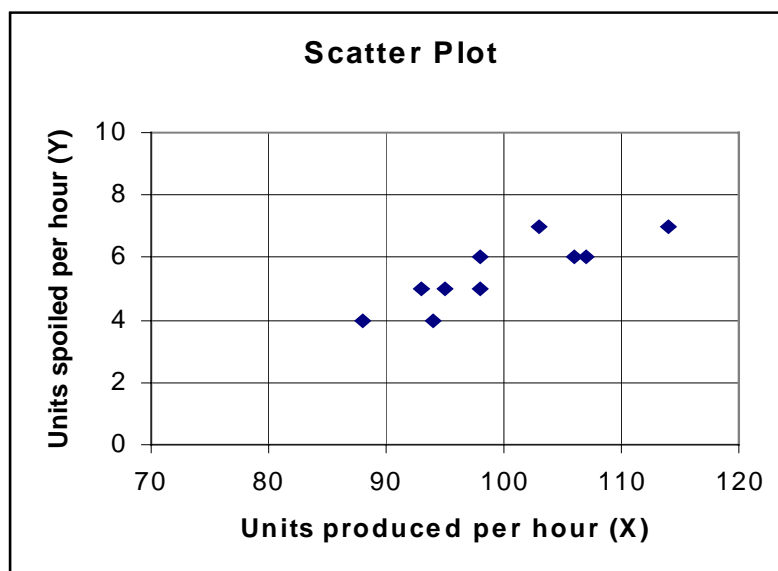
5.6 USING THE COMPUTATIONAL FORMULAE

This section is included for reference only. As far as this course is concerned, it is assumed you will use a package to do the calculations.

5.6.1 Correlation

To illustrate the calculation of r we consider the following data which gives the hourly numbers of units spoiled and the hourly number of items produced for 10 press operators.

Operator	A	B	C	D	E	F	G	H	I	J
Units produced per hour (X)	94	98	106	114	107	93	98	88	103	95
Units spoiled per hour (Y)	4	5	6	7	6	5	6	4	7	5



The relationship seems to be linear and positive.

The correlation coefficient is given by

$$r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$$

where $S_{XX} = \sum x^2 - (\sum x)^2/n,$

$$S_{YY} = \sum y^2 - (\sum y)^2/n,$$

and $S_{XY} = \sum xy - (\sum x)(\sum y)/n.$

Operator	x	y	x ²	y ²	xy
A	94	4	8836	16	376
B	98	5	9604	25	490
C	106	6	11236	36	636
D	114	7	12996	49	798
E	107	6	11449	36	642
F	93	5	8649	25	465
G	98	6	9604	36	588
H	88	4	7744	16	352
I	103	7	10609	49	721
J	95	5	9025	25	475
Totals	996	55	99752	313	5543

$$\begin{aligned}
 n &= 10 & \sum y &= 55 \\
 \sum x &= 996 & \sum y^2 &= 313 \\
 \sum x^2 &= 99752 \\
 \sum xy &= 5543
 \end{aligned}$$

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} = 99752 - \frac{(996)^2}{10} = 550.4$$

$$S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n} = 313 - \frac{(55)^2}{10} = 10.5$$

and
$$S_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} = 5543 - \frac{(996)(55)}{10} = 65.0$$

$$\therefore r = \frac{65.0}{\sqrt{(550.4)(10.5)}} = \boxed{+0.855}$$

5.6.2 Regression

The Method of Least Squares gives the slope (b) and the intercept (a) of the best-fitting lines as:

$$b = \frac{S_{XY}}{S_{XX}}$$

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

where S_{XY} and S_{XX} are as defined in 5.6.1.

The following data are the travelling times (Y in minutes) and distance travelled (X in km) for 12 trips made by a photocopier repair man.

x	y	x^2	xy
2	5	4	10
7	13	49	91
8	10	64	80
3	7	9	21
5	9	25	45
4	8	16	32
9	15	81	135
10	16	100	160
10	17	100	170
7	13	49	91
7	12	49	84
5	9	25	45
$\sum x = 77$	$\sum y = 134$	$\sum x^2 = 571$	$\sum xy = 964$

$$\bar{x} = \frac{77}{12} = 6.417 \quad \bar{y} = \frac{134}{12} = 11.167 \quad n = 12$$

Then

$$S_{XY} = 964 - \frac{(77)(134)}{12}$$
$$= 104.167$$

$$S_{XX} = 571 - \frac{77^2}{12}$$
$$= 76.917$$

Thus

$$b = \frac{104.167}{76.917}$$

$$= \boxed{1.354}$$

and

$$a = 11.167 - (1.354)(6.417)$$

$$= \boxed{2.477}$$

So the line of best fit is given by the equation

$$\underline{Y = 2.5 + 1.35X}$$

or

$$\underline{\text{Estimated Time} = 2.5 + (1.35) \times \text{Distance}}$$

Exercise

Use the computational formulae to confirm that the regression line for predicting units spoiled per hour (Y) from units produced per hour (X) is given by:

$$Y = -6.26 + 0.118X .$$

(The data and most of the preliminary calculations are in 6.6.1.)

5.7 NON-LINEAR REGRESSION

So far we have considered only linear relationships between variables. Before calculating the correlation coefficient or a regression line, this must be checked with a scatter plot. If there is a curved relationship it can often be made linear simply by TRANSFORMING one (or both) of the variables.

Some common transformations are:

Use $\log(Y)$ and X
 Y and $\log(X)$
 $\log(Y)$ and $\log(X)$
 Y and $1/X$
 Y and \sqrt{X}
 Y and X^2

In practice, try a few transformations and see which has the best linearising effect using scatter diagrams. Check the correlation coefficient as well – the closer it is to ± 1 , the greater is the linear relationship between the variables. (Actually, it is more usual to check the equivalent R^2 value – the closer it is to 100%, the better.) We will look at some examples in the tutorial session.

Having linearised the data, proceed with the regression analysis as before.