Correlation & Linear Regression

Correlation

- Correlation means a link or connection between two variables.
- The common aim of research is to try to relate a variable of interest to one or more other variables.

Correlation

- We can look for an association between variables e.g. dose of drug and resulting systolic blood pressure, advertising expenditure of a company and end of year sales figures, etc.
- We can then establish a theoretical model to predict the value of one variable from a number of known factors

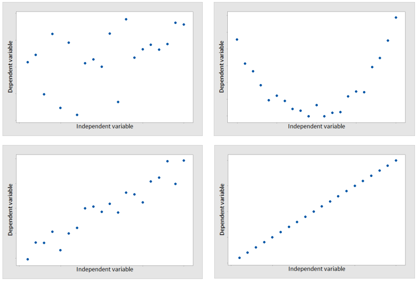
Correlation

- The *independent* variable is the variable which is fixed under the investigator's control, we denote this as x.
- The *dependent* variable is the one which the investigator is trying to estimate or predict, we denote this as y. The y variable is the response variable.

Correlation on a graph

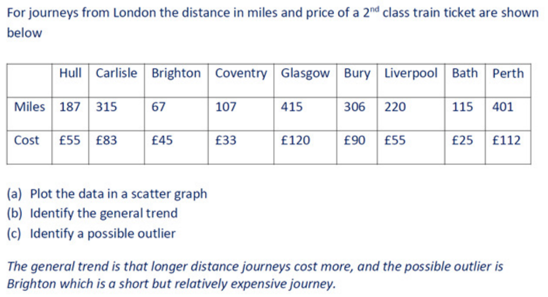
- When we plot data on a scatter plot, we can see if there is a linear association between the two variables.
- Scatter plots do not have to start at zero.

Scatter Plots



Correlation

- Positive correlation means that as one variable gets larger, so does the other.
- Negative correlation means that as one variable gets larger, the other gets smaller.

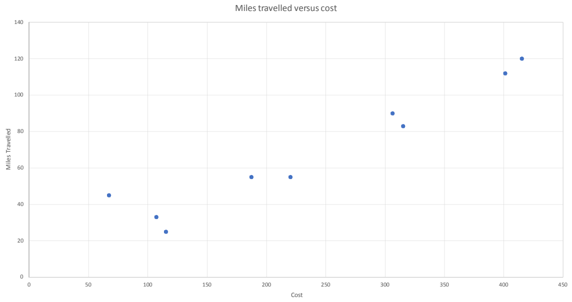


For journeys from London the distance in miles and price of a 2nd class train ticket are shown below

	Hull	Carlisle	Brighton	Coventry	Glasgow	Bury	Liverpool	Bath	Perth
Miles	187	315	67	107	415	306	220	115	401
Cost	£55	£83	£45	£33	£120	£90	£55	£25	£112

(a) Plot the data in a scatter graph
(b) Identify the general trend
(c) Identify a possible outlier

The general trend is that longer distance journeys cost more, and the possible outlier is Brighton which is a short but relatively expensive journey.



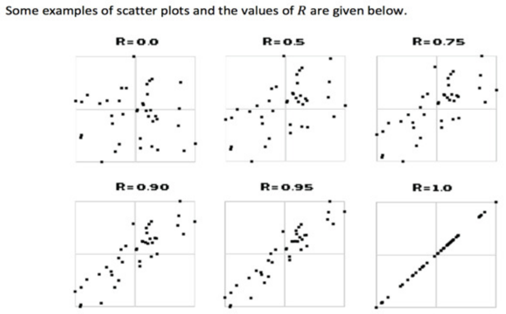
Correlation Coefficient

- The correlation coefficient ‘r’ for a sample and ρ (Greek letter ‘rho’) for a population is known as Pearson’s coefficient and measures the degree of **linear association** between two *numerical* variables.
- The range of possible values of ‘r’ is from -1 to +1
- The correlation is high and positive if the observations lie close to a straight line and therefore give a value of ‘r’ close to 1.
- If the correlation is high and negative, the observations will lie close to the line and the value of ‘r’ will be close to -1.
- If the correlation is close to 0, the observations will be scattered around and there will be little to no correlation between the variables.

Correlation Coefficient – How is it calculated?

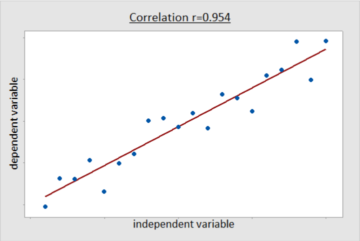
- We can calculate ‘r’ using a computer. There is a mathematical formula but we don’t need to worry about its origin in this course.

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{(\sum(x_i - \overline{x})^2)(\sum(y_i - \overline{y})^2)}}$$

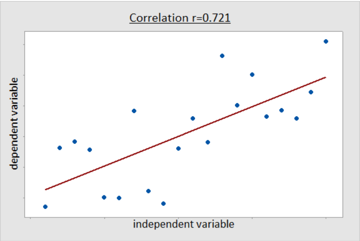


Some examples of scatter plots and the values of *R* are given below.

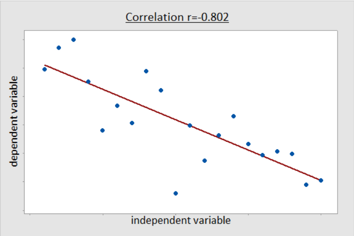
Correlation



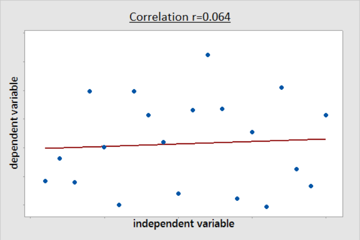
Correlation



Correlation



No Correlation



Being careful

- Remember, two variables could be highly correlated by coincidence (coincidental correlation) but not causal meaning that one does not cause the other. E.g. house prices and size of the hole in the ozone layer.
- There may be no direct connection between highly correlated variables (known as spurious correlation)
- Indirect correlation – where variables don't depend on each other but on a third variable. e.g. high correlation between infant mortality and the extent of overcrowding in a certain town between 1st and 2nd WW
- N.B. Low correlation does not necessarily mean a low degree of association (relationship may be non-linear)

Question 1

A research study has reported a correlation of –0.59 between the eye colour (brown, green, blue) of experimental animals and the amount of nicotine that is fatal to the animal when consumed. This indicates …

(a) nicotine is less harmful to one eye colour than others

(b) lethal dose decreases as eye colour changes

(c) eye colour of animals must always be considered in assessing the effect of nicotine consumption

(d) further study required to explain this correlation

(e) correlation is not an appropriate measure of association

Question 2

If the correlation between body weight and annual income were high and positive, we would conclude that …

(a) high incomes cause people to eat more food

(b) low incomes cause people to eat less food

(c) high income people tend to spend a greater proportion of their income on food than low income people, on average

(d) high income people tend to be heavier than low income people, on average

(e) high incomes cause people to gain weight

Linear Regression – Line of best fit

- The line of best fit is a line drawn to best fit the data. When drawing the line of best fit by hand, note that:
  - It should have roughly the same number of points above as below
  - It does not have to go through any of the data points exactly.
  - It does not have to go through the origin of the graph

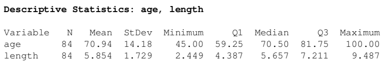
Line of Best Fit

- We can develop an equation to predict the dependent variable from knowledge of predictor variable(s)
- general equation of a straight line y = mx + c, where *c* is the intercept and *m* is the slope or gradient, of the line
- We fit this line by eye so it is subjective
- On a computer, linear regression fits a straight line to the data using method of least squares

Practical Example

- data from a study of foetal development
- date of conception (and hence age) of the foetus is known accurately
- height of the foetus (excluding the legs) is known from ultrasound scan
- age and length of the foetus are clearly related
- aim is to model the length and age data and use this to assess whether a foetus of known age is growing at an appropriate rate

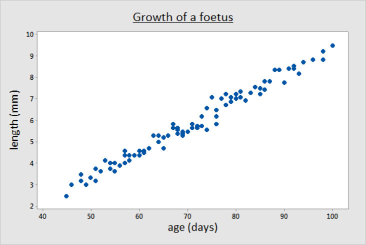
Descriptive Statistics



```
Descriptive Statistics: age, length

Variable   N   Mean  StDev  Minimum     Q1  Median     Q3  Maximum
age       84  70.94  14.18    45.00  59.25   70.50  81.75   100.00
length    84  5.854  1.729    2.449  4.387   5.657  7.211    9.487
```

Graphical Assessment of Data



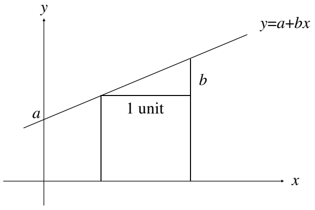
Linear Regression Model

- from the plot it would appear that age and length are strongly related, possibly in a linear way
- a straight line can be expressed mathematically in the form

$$y = a + bx$$

- where *b* is the slope, or gradient of the line, and *a* is the intercept of the line with the *y*-axis

Modelling a Straight Line



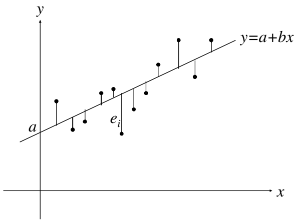
Fitting a Regression Line

- if the data lay on a straight line and there was no random variation about that line, it would be simple to draw an approximate straight line on the scatter-plot
- this is not the case with real data
- for a given value of the explanatory variable there will be a range of observed values for the response variable
- different assessors would estimate different regression lines
- in order to have objective results, it is necessary to define some criteria in order to produce the 'best fitting straight line' for a given set of data

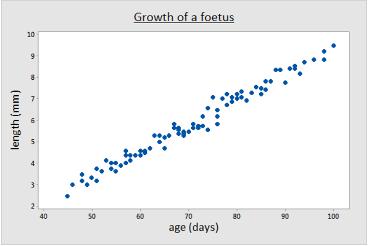
Method of Least Squares

- This is used to define the position of the line which, on average, has all the points as close to it as possible.
- The best fitting line is called the ***least squares linear regression line***
- The vertical distances between each point and the fitted line are called the *residuals* and are used in estimating the variability about the fitted line

Least Squares Line



Graphical Assessment of Data



Interpretation of Results

- The regression equation is …

$$length = -2.66 + 0.12 \times age$$

- The line cuts the y – axis at -2.66 and the gradient/slope is 0.12. This implies that as the age of the foetus increases by one day, the length increases by 0.12mm
- Calculate what the estimated length for a foetus of age 85 days would be

$$length = -2.66 + (0.12 \times 85) = 7.54$$

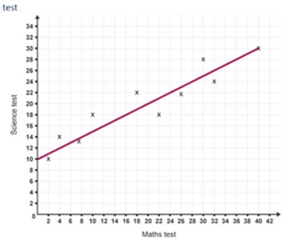
Model Predictions

- A prediction interval gives the range of values between which the value for an individual is likely to lie: (7.01 to 8.08mm)
- We can use the regression model to assess whether a foetus of known age is growing at an appropriate rate.
- For example, consider a foetus of age 85 days
- does the measured length lie within the normal range i.e. between 7.01 and 8.08mm?
- if measured length is <7.01mm, there is evidence that the foetus is not growing as it should
- if measured length if >8.08mm, is the foetus larger than expected? Is the actual age (and due date) wrong?

Uses of Regression Lines in prediction

- The value of the independent variable (*x*) should be within in the range of the given data. If we go beyond the range, it’s not obvious that there is still a linear relationship.
- Remember that the predicted value of the dependent variable (*y*) is only an estimate.

25. The graph below shows the correlation between marks in the maths test and marks in the science test

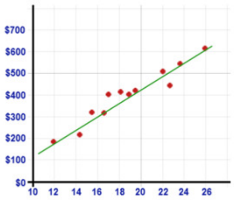


(a) Using the line of best fit, estimate a pupil’s mark on the Science Test if they scored 0 on the Maths test

(b) Using the line of best fit, estimate a pupil’s mark on the Science Test if they scored 40 on the Maths test

(c) Using the information from part (a) and (b), work out the gradient and intercept of the line of best fit

The graph below shows how the age of some fun runners affects the amount of sponsorship money raised



(a) Estimate the correlation coefficient

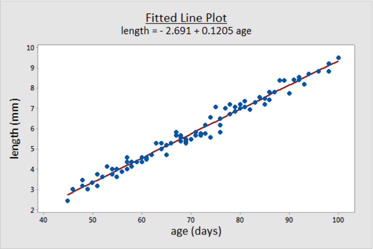
(b) Using the line of best fit, estimate the amount raised by a 16-year-old and a 26-year-old

(c) Calculate the gradient and find the line of best fit (you can use that the intercept is - $180)

(d) Use the line of best fit to estimate the amount raised by a 13-year-old

(e) Why can the line of best fit not be used to estimate the amount raised by a 30-year-old?

Fitted Line



Exercise

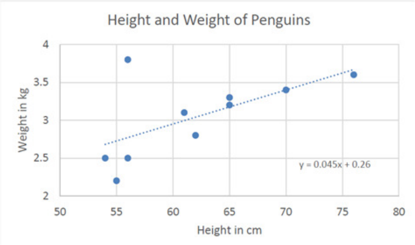
- Use the calculated least squares linear regression line to estimate the size of a foetus at the following gestation times:
  - (a) 2 days
  - (b) 60 days
  - (c) 100 days
  - (d) 300 days
- For each of your estimated lengths, state whether or not you believe the estimate to be accurate.

Coefficient of Determination

- This is a measure of the amount of variability in the data which is explained by the regression line and is usually expressed as a percentage.
- If we have a high r^2 it means that the regression line fits the data well and therefore the predicted values for the dependant variable will be more accurate.
- The correlation (*r*) between foetal size and age is 0.988
- The coefficient of determination is r^2 (i.e. 0.988×0.988) which is 97.6% meaning that the model is a good fit and we should get reliable estimates of length.
- We can think of this as 97.6% of the variability in length is explained by the variability in the ages. So what about the other 2.4%?
- Factors such as genetics, diet habits of the mother etc.

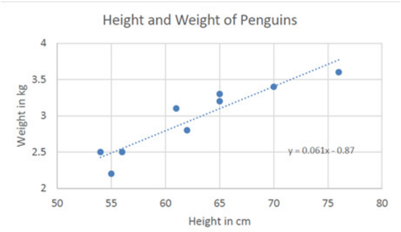
Outliers

- Outliers can have a large effect on the line of best fit and can dramatically weaken the correlation coefficient.
- In some cases, the outlier should be removed before fitting the regression line.



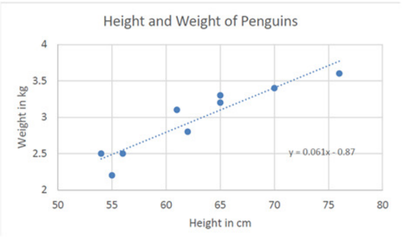
The correlation coefficient = 0.615
Weight in kg of penguins = 0.26 + 0.0145 x Height

Estimate the weight of a penguin with height 55cm



We have now made the plot again but removed the outlier.
Correlation coefficient = 0.931
Weight in kg of penguins = -0.87 + 0.061 x Height

Now estimate again the weight of a penguin with height 55cm



For which weight might we expect to predict more accurately, a height of 65cm or a height of 80cm?

Hypothesis Testing

- A hypothesis is a prediction about the data.
- The **null hypothesis** ($\boldsymbol{H_0}$) is a statement that indicates nothing unusual is happening, for example ‘this coin produces exactly 50% heads’
- The **alternative hypothesis** ($\boldsymbol{H_1}$) is a statement about the data that might be true, for example ‘this coin produces more than 50% heads’.
- Rather than try and prove the alternative hypothesis, which is generally not possible, the process is instead to show the null hypothesis is unlikely.

Hypothesis Testing

- the linear relationship between two variables is significant if there is evidence to suggest that ρ is significantly different from zero
- for the hypothesis test
- observed value of *r* is 0.864 and associated p-value is <0.001 (calculated by MINITAB using >Stat >Basic Statistics >Correlation)
- Conclusion: reject H_0 and conclude that there is evidence to suggest that there is a linear relationship of size on age