**Topic 2: Supplementary notes. About Scatter plots.**

**URL**

**http://www.uow.edu.au/student/qualities/statlit/module3/5.4interpret/index.html**

**5.4 Scatterplots: How can you interpret a scatter plot?**

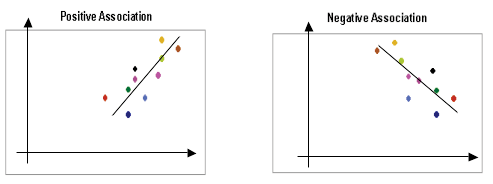
To read a scatter plot you need to look for the overall pattern. This tells you something about the **direction, form** and **strength**of the relationship.

**(a) Direction**

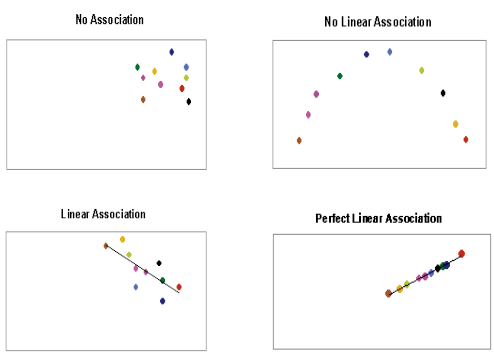
*Positive gradient:* When the larger values of the horizontal (explanatory) variable are associated with larger values of the vertical (response) variable. As the explanatory variable increases, so does the response variable. Can you see how the data, as we move from left to right, are gradually rising?

*Negative gradient:* When the larger values of the explanatory variable are associated with smaller values of the response variable. As the explanatory variable increases, the response variable decreases. Can you see how the data, as we move from left to right, are gradually decreasing?

(In both cases we always use a consistent method - "explanatory variable increases" means that we move from left to right - what mathematicians call 'moving in the positive direction'.)

Notice that in each of the following two diagrams a trend line has been superimposed on the scatterplot - it helps gives an overall view of the direction in which the data points are sitting.   
  


**(b) Form**

We need to know whether there is association or not, and whether it is linear or not. The relationship might be linear or curved or there might be no underlying form. In this course we will mainly concentrate on linear relationships, but we must be aware of the existence of non-linear ones.   
  


**(c) Strength**

The strength of the pattern is related to how tightly clustered the points are around the underlying form. We often use phrases like those following to describe the strength of the relationship, whether negative or positive. These phrases are of course, subjective.

|  |  |
| --- | --- |
| **(near) zero correlation** statlit - zero correlation | **"weak" correlation** statlit - weak correlation |
| **"moderate" positive correlation** statlit - moderate correlation | **"strong" positive correlation** statlit - strong correlation neg |
| **"moderate" negative correlation** statlit - strong correlation neg | **"strong" negative correlation** statlit - mdoerate correlation neg |

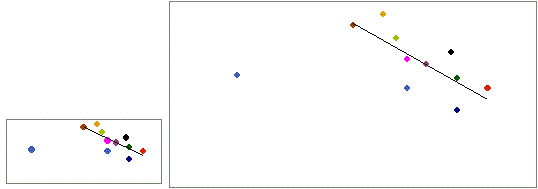
**(d) Outliers and influential points**

You can also look for individual points that fall outside the overall pattern of the scatter plot. Outliers can have a big influence on correlation. These should be examined (as far as possible) to determine whether they are real data values, or some kind of data error. It is quite common for a researcher to perform two analyses - the first analysis with outliers remaining in the data set, the second with them removed.

The implications of removing/retaining the outlier must be clearly stated (it is unethical to simply erase a data point because it is not in the mainstream pattern!). Reasons and justification for any action must be clearly enunciated.

|  |  |
| --- | --- |
| statlit - outlier | If the blue outlier were to be removed, we would have a data set with a high level of association. As it is, the outlier has a significant effect on the level of association. |

|  |  |
| --- | --- |
| statlit - influential | In this graph, influential points lie in the same direction as the major part of the data set, but are a long way removed.  For the graph at left questions would have to be asked as to why there is a gap, and whether there are special characteristics causing the two clusters to arise. |

Note that changes in the scale of a graph **do not** change the strength of a pattern. Below are two scatter plots of the same data each drawn using different scales. Changes in the relative scale might appear to change the strength of the pattern but note that the line showing the best fit for the trend of the relationship is similar in both cases.  
  


**5.5 Scatterplots: The correlation coefficient, r**

The strength of linear association can be measured using a number - the correlation co-efficient. Correlation measures the strength and direction of a linear relationship. Rather than use subjective word descriptors such as "strong positive correlation", r gives a numerical measure.

The correlation coefficient, r, has a specific range of values:

statlit - r range

Note that:

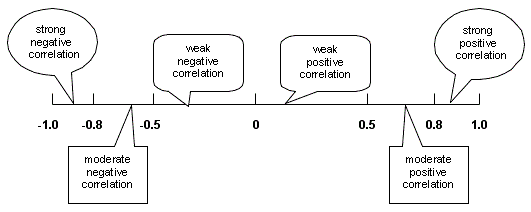
* *r*  never lies outside this range, therefore *r* = 2 is a nonsense answer whose only explanation can be "I made an arithmetic error".
* *r* =1 is perfect positive correlation and all the data points lie exactly on a straight line with positive gradient.
* *r* = -1 likewise is perfect negative correlation.
* *r*  is often measured or referred to as a percentage. In this case, the range is from -100% to 100% (remembering that 100% is the same as 1)

How do I find the *r* value for a data set?

Steps you need to follow:

1. draw the scatterplot;
2. draw the trend line which describes the direction of the data;
3. evaluate how closely the cloud of data points clusters around the line;
4. determine what *r*  value and what word descriptor best suits the data cloud.

The following diagram has a number line of *r* values to help you assigning the numbers and the word descriptors.



Consider the following examples of scatterplots.

These have the cloud of data points and a trend line fitted to show the direction of the data.

It would be helpful for you to memorise these to assist you describe your own data sets:

|  |  |
| --- | --- |
| zero correlation | moderate positive correlation  r=0.5 |
| statlit - corr 1 | statlit - corr 2 |
| strong positive correlation  r=0.8 | strong positive correlation  r=0.95 |
| statlit - corr 3 | statlit - corr 4 |
| And now consider some negative gradients: | |
| weak negative correlation  r=-0.40 | moderate negative correlation  r=-0.65 |
| statlit - corr 5 | statlit - corr 6 |
| moderate negative correlation  r=-0.75 | strong negative correlation  r=-0.85 |
| statlit - corr 7 | statlit - corr 8 |

|  |
| --- |
| **SCENARIO [Moore, 1995]**  *Archaeopteryx* is an extinct animal that possessed both scales and feathers and at one stage was thought to be the 'missing link' between lizards and birds. Only six fossil specimens exist and they vary greatly in size. As a result, there has been a lot of discussion about whether the fossils all belong to one species or to different species. In order to help answer this question, data from the length (cm) of the femur (a leg bone) was plotted against the length of the humerus (a bone in the arm) on a scatter plot. Data were available for five of the specimens.  **Comment:** If the specimens belong to the same species and the differences are due to differences in size because of age, then the points should show a positive (but not necessarily linear) relationship. If any of the plotted points was an outlier from the bivariate pattern shown by the other points, this might suggest (but not prove) that the point represented a specimen from a different species.  statlit - humerus |

**Question**

What does the scatterplot indicate?  
  
(a) No association indicates that five separate species are present.  
(b) An outlier can be observed indicating the presence of two separate species  
(c) There is a strong positive association indicating that the specimens belong to one species.  
(d) It is not possible to make any statement about the possible number of species present from the data provided.

**Answer**

Did you select C as your answer? Well done! There are no obvious outliers. The researchers (Houck, M. A., Gauthier, et.al 1990) also produced scatter plots of the length of other bones such as the ulna, tibia and teeth against femur length and obtained similar relationships. These results led them to conclude that the fossils were consistent with a single species of different sizes none of which was fully grown.