

MATH 124 Spring 2005

Lecture: 6

Date: Feb 24, 2005

Skills you should acquire from this lecture:

- Ability to draw and interpret a scatterplot.
- Ability to calculate and interpret correlation.

Related readings in the textbook:

- Section 2.1
- Section 2.2

Scatterplots

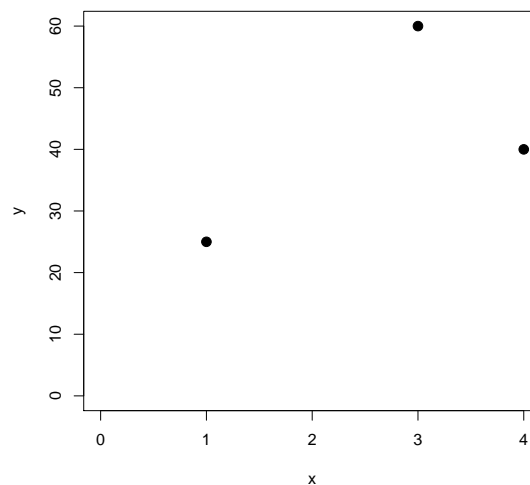
A scatterplot shows the relationship between two quantitative variables each measured on the same set of individuals eg measure both the height and weight for each person using a sample of 20 individuals. One variable falls on the horizontal axis (traditionally this is the explanatory variable) and one variable falls on the vertical axis (traditionally this is the response variable). Each individual is plotted as a point, with the location determined by the values of the two variables.

Example

Suppose we have 3 individuals and we have measured two variables x and y for each of these individuals. Our data table is the following

x	y
4	40
3	60
1	25

A scatterplot of this data looks like the following:



Interpreting a scatterplot

When examining a scatterplot there is several things you might want to consider, including:

- Overall Pattern

Form

No clear pattern

Linear

Curve (non-linear)

Direction

Increasing

Decreasing

Strength

Strong

Weak

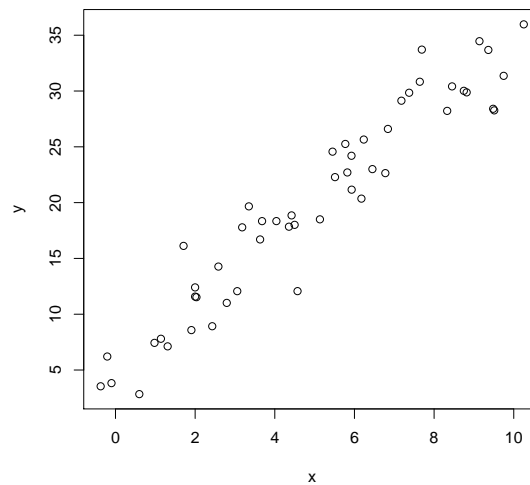
- Deviations from the pattern

Outliers

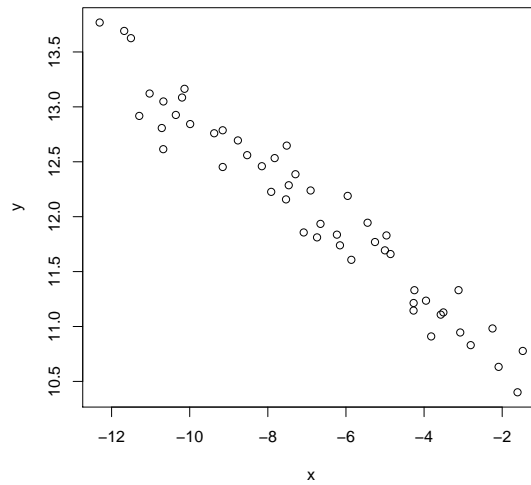
Direction

Direction is typically about deciding whether there is an increasing or decreasing relationship between the two variables being examined.

1. We say that the relationship between two variables is a *positive association* if we observe that as the value of one variable increases so does the value of the other variable. A typical scatterplot in this case looks something like the following:



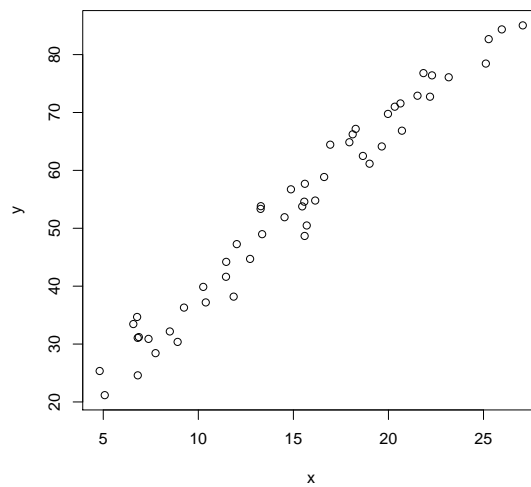
2. We say that the relationship between two variables is a *negative association* if we observe that as the value of one variable increases the value of the other variable decreases. A typical scatterplot in this case looks something like the following:



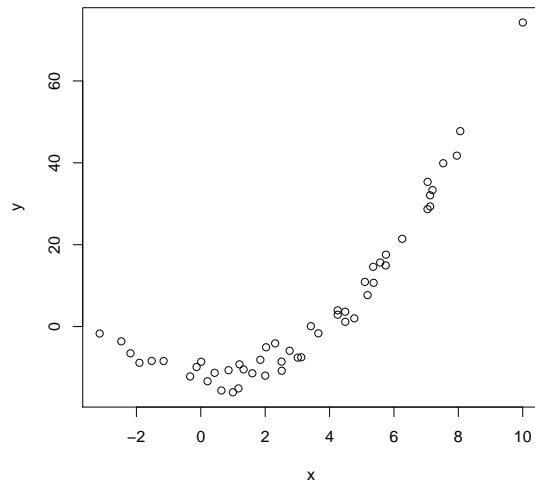
Form

Form is about determining the shape of the relationship between the two variables being examined.

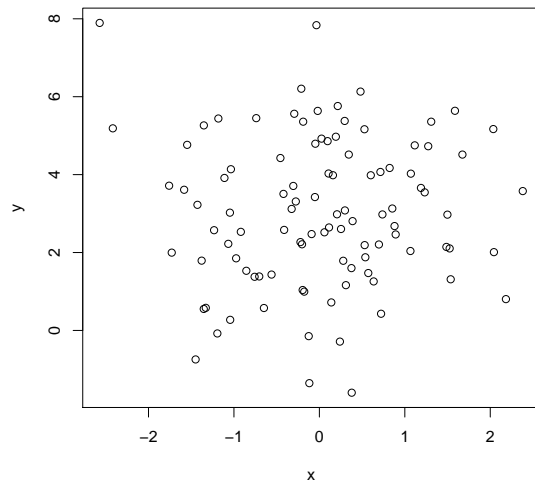
1. We say that the form is *linear* if the points seem to lie along a line



2. We say that the form is *non-linear* if the points seem to lie along a curve of some kind



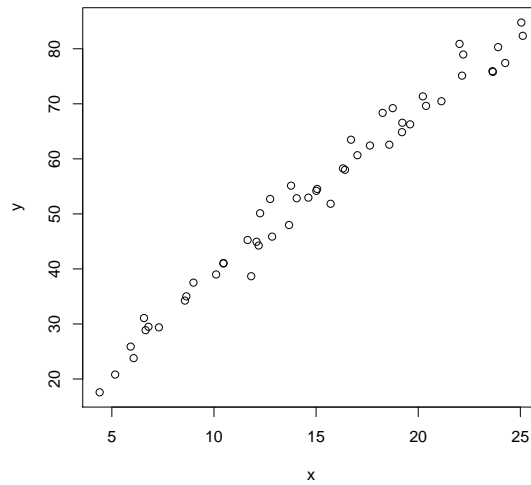
3. We say that the form has *no pattern* if there does not appear to be a clear relationship between the two variables.



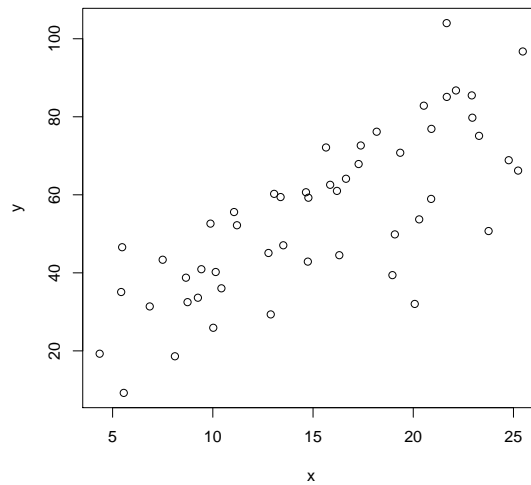
Strength

Strength is about how strongly the points on the scatterplot seem to follow the particular form

1. We say that the relationship is *strong* if the points seem to follow a specific form very closely.

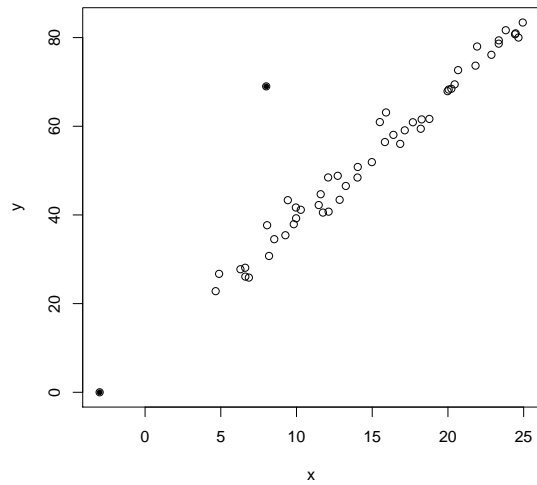


2. We say that the relationship is *weak* if the points seem to follow a specific form but not closely



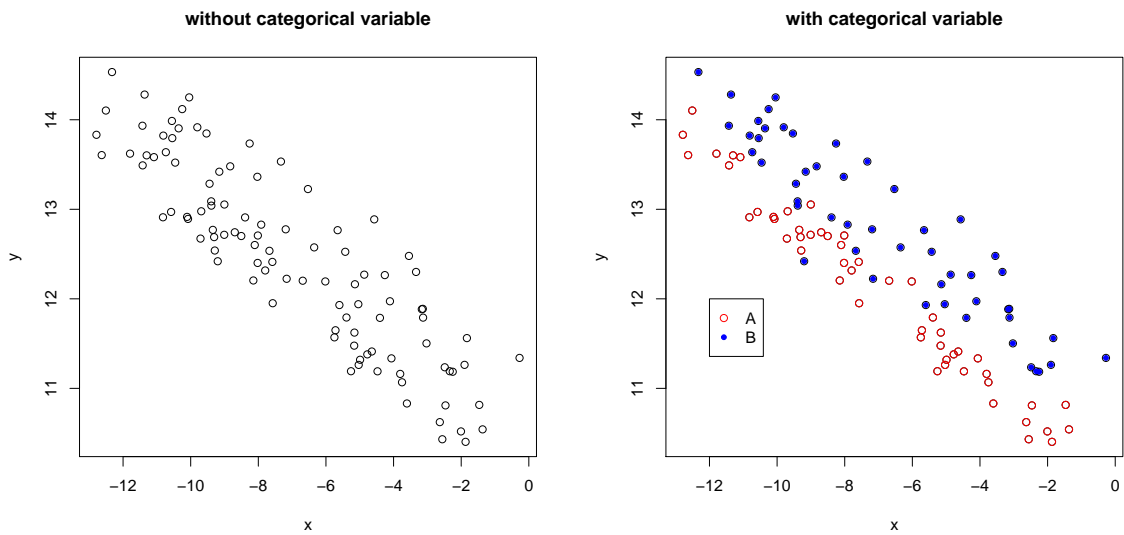
Outliers

On a scatterplot outliers are points that seem to deviate from the rest of the points, either by being far away from the general mass of the point cloud or clearly off the underlying form.



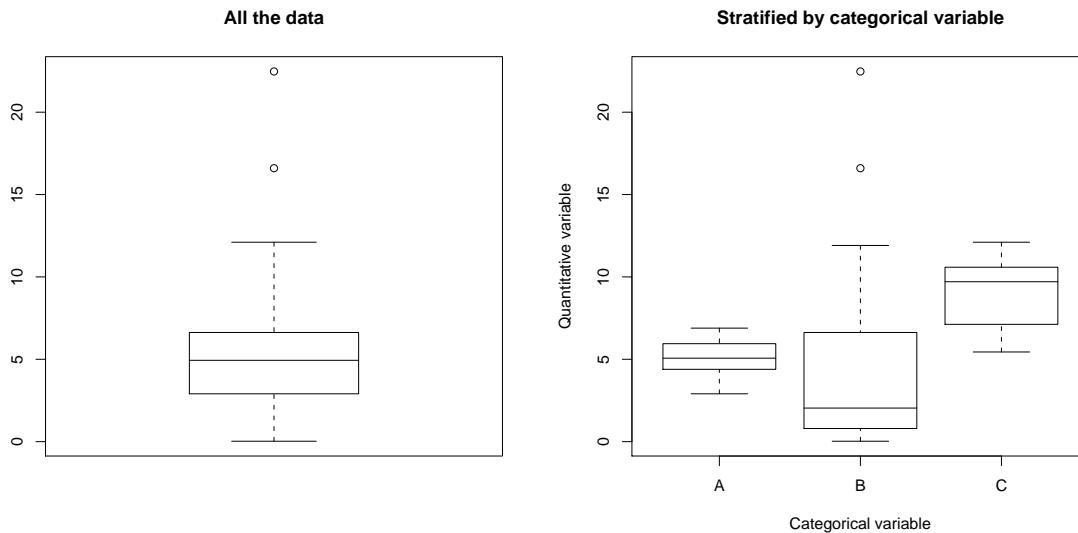
Adding a categorical variable to a scatterplot

Sometimes along with the two quantitative variables, we are also interested in the values of a categorical variable and its relationship to the two quantitative variables. We add the categorical variable to the scatterplot by using different plotting symbols or colors. Adding this to the graph may well add explanatory information to your plot.



Comparing a quantitative and a categorical variable using boxplots

Boxplots allow us to compare quantitative and categorical variables. Specifically, we draw boxplots of the quantitative variable where each boxplot is created using all individuals with a specific value for the categorical variable. Differences in the shape, size and position of the boxplots show whether there are differences in the quantitative variable between values of the categorical variable. In other words, it allows you to see whether there is an association between the categorical variable and the quantitative variable.



Correlation

Correlation is a summary statistic which measures the strength and direction of the linear relationship between two quantitative variables each measured on the same individuals (or objects). Suppose our two variables are x and y . Then the formula for correlation is

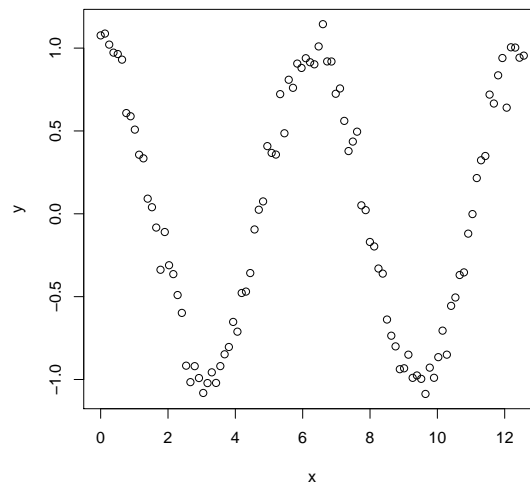
$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

where s_x and s_y are the sample standard deviations of x and y respectively. \bar{x} and \bar{y} are the sample means.

Properties of correlation

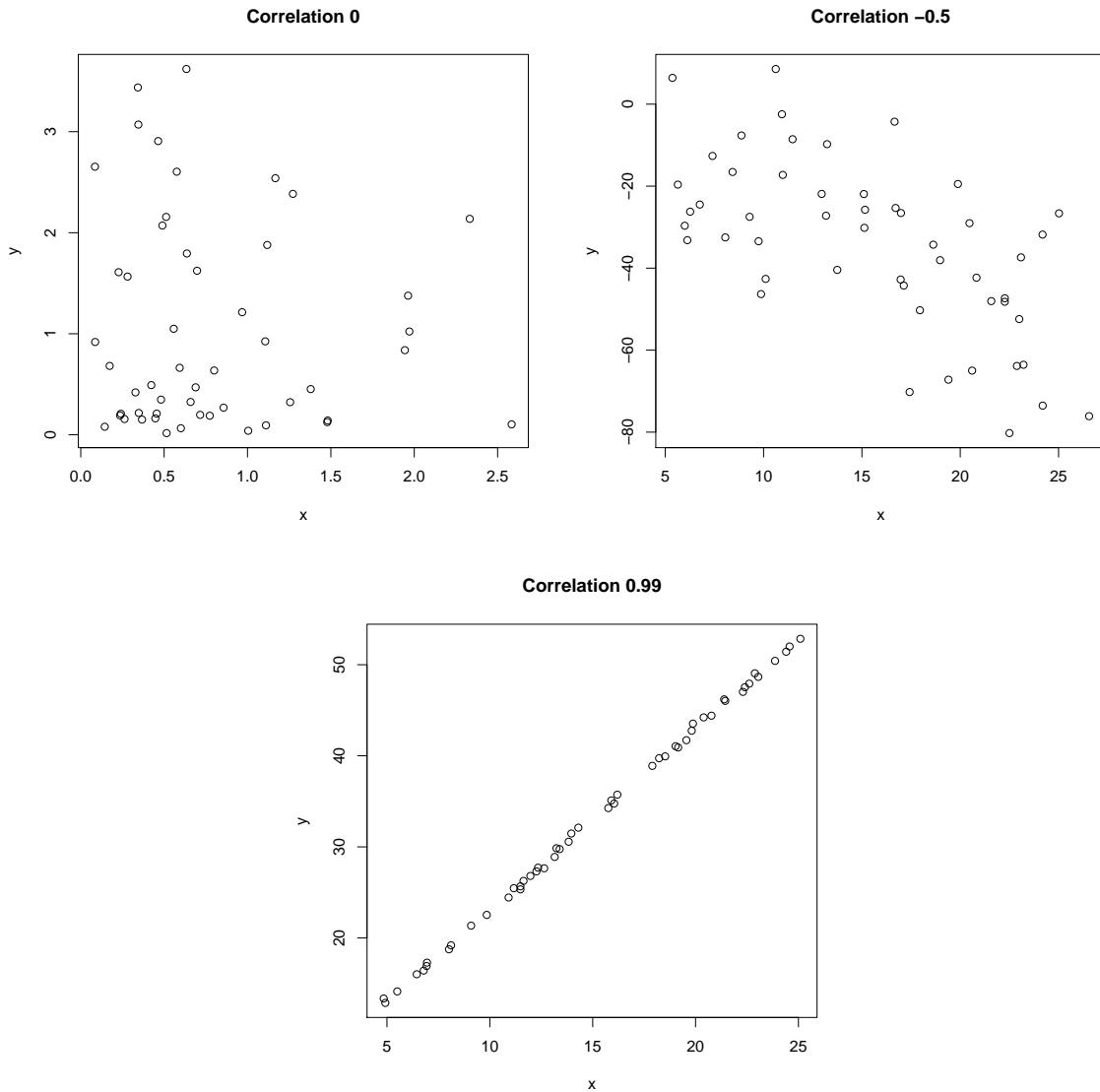
1. If $r > 0$ we say that there is a positive relationship between x and y
2. If $r < 0$ we say that there is a negative relationship between x and y

3. The smallest r can be is -1 and the largest is 1 . ie $-1 \leq r \leq 1$
4. If the correlation is close to 1 we say that there is a strong positive linear relationship between the two variables
5. If the correlation is close to -1 we say that there is a strong negative linear relationship between the two variables
6. If the correlation is close to 0 we say that there is a weak relationship between the variables
7. Correlation has no units of measurement
8. Correlation describes only the strength of the linear relationship. Non-linear relationships are not described no matter how strong. eg this situation would have correlation 0 , but there is a clear and strong non linear relationship



9. Correlation can be affected by outliers
10. Correlation does not take into consideration which variable is the explanatory variable and which the response variable.

Some examples



An alternative formula for hand calculation of correlation

When computing correlation by hand you may find it easier to use this version of the formula. In fact it is strongly recommended that you only ever use this formula when computing correlation for homeworks and in exams, forget the other version entirely !!!!

$$r = \frac{1}{n-1} \left(\frac{1}{s_x} \right) \left(\frac{1}{s_y} \right) \left(\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} \right)$$