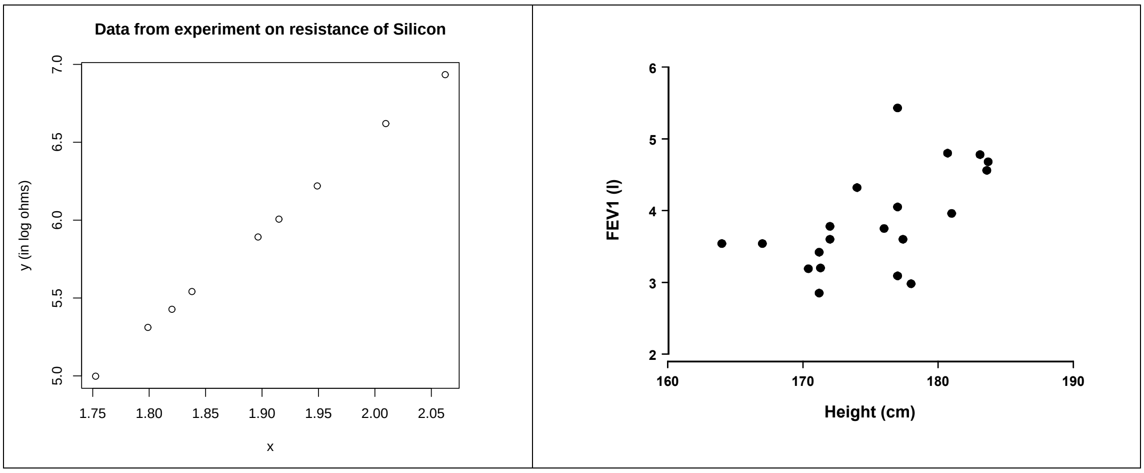
Relating two variables: linear regression and correlation

Introduction

Thus far we have been concerned with the analysis of a single variable – estimating the parameters which define its distribution, assessing the precision of the estimates and trying to decide if values of the parameters are different in separate groups. Quite often we need to consider more than one variable and how they relate to one another. There are ways of assessing the relationship between several variables but we will only consider the case of two variables.

The most obvious way to start to analyse how one variable changes with another is to plot a scatter diagram of the points. Indeed, it is not only obvious, it is an essential first step in any such analysis. In Figure 1 two such scatter plots are shown.

In Figure 1 a) the variable plotted on the vertical axis (generally called the *y* variable) is the log of the electrical resistance of a semi-conductor and the variable plotted on the horizontal axis (the *x*-variable) is a function of temperature. The data clearly indicate that *y* and *x* show very little deviation from a straight line relationship. This kind relationship is quite common in the physical sciences, where linear relationships can often be discerned from the underlying theory



a) resistance and temperature in a semiconductor[†]   b) FEV1 and stature of male medical students[‡]

Figure 1: examples of scatter plots

Figure 1 b) shows the heights (cm) of 20 male medical students on the horizontal axis and the corresponding Forced Expiratory Volume in one second (FEV1) in litres (l) on the vertical axis. This is much more typical of the kind of relationship seen in the biological and clinical sciences. While there seems to be some sort of relationship – taller students tend to have a larger FEV1 – it is by no means as clear cut as that in

[†] Data from chapter 4 of GL Squires, 2001, *Practical Physics, 4th ed.* CUP, Cambridge.

[‡] Data from chapter 11 of M Bland, 2000, *An Introduction to Medical Statistics 3rd ed.*, OUP, Oxford.

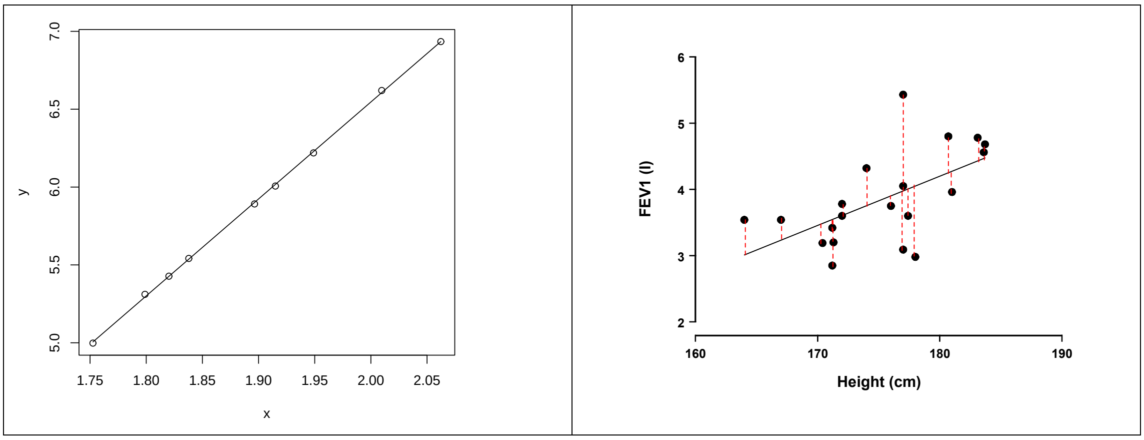
Figure 1 a). There are many instances of students having a larger FEV1 than that of a taller peer. The challenge is to find a quantitative description of this kind of relationship.

Fitting a line to the data

A quantitative description for the data in Figure 1 a) is readily obtained by fitting a straight line to the data. While in this case this might reasonably be done by eye, it is better, especially in less clear-cut cases, to have an objective algorithm for doing this. Recall that a straight line has the equation

$$y = \alpha + \beta x$$

where α is the *intercept* and β is the *slope*. The term 'fitting a line' simply means some sort of algorithm for finding suitable values for α and β.



a) Line of best fit for semiconductor data

b) line of best fit for FEV1 data

Figure 2: lines fitted to the data in figure 1

The most widely used algorithm for finding α and β is the *Method of Least Squares*. This has been applied to obtain the lines in Figure 2. It is easier to see how it works by considering its application in Figure 2 b). A line is drawn on the figure and the vertical distance of each point from the line is found – the dashed lines in Figure 2 b). It is usual to count the distances of the points above the line as positive but as negative for points below the line. Simply adding these up would lead to cancellation between the positive and negative distances. To overcome this, the distances are squared (and therefore all positive) before be being added up – a device reminiscent of the way a standard deviation is computed. All possible lines are considered and the one which gives the smallest value for this *sum of squares* is the fitted line. Of course, you cannot actually try drawing all possible lines – the method works by applying methods from the calculus.

If all data looked like Figure 1 a) then this would be the end of the matter – least squares has provided an objective way to calculate the best fitting line and its meaning would be clear. However, what has been achieved by fitting a line such as that shown in Figure 2 b)? Least squares by itself is essentially a geometrical tool and to make sense of its use in cases such as Figure 2 b) requires a statistical context.

A statistical perspective – the method of linear regression

In order to make progress when considering the statistical relationship between two variables, let's go back to the case of one variable. The fundamental idea was that the variable was measured on a population and we observed a sample from that population. The distribution of the variable in that population is defined in terms of parameters and for a Normal distribution these are the mean, μ and standard deviation, σ.

For the data in Figure 1 b), the distribution of FEV1 in the population could also be thought of as having a mean and standard deviation. However, in order to acknowledge the fact that FEV1 is likely to be larger in larger people, it is too restrictive to consider the mean of this population to be a single number, μ. Rather, we should expect the mean to depend on the height, x, of the individual: i.e the mean is a function of height $\mu(x)$. In order to keep things as simple as possible we also assume that the way the mean depends on height is as simple as possible, namely a straight line – so:

$$\mu(x) = \alpha + \beta x \, .$$

This dependence of the mean of the y variable on the x variable is known as the *regression of y on x*. Other forms of dependence, such as something more complicated than a straight line or dependence on more than one x variable, are possible but will not concern us presently.

It would be possible to allow the other parameter, σ, to depend on height, but in most applications this is not found to be necessary and we will assume that the distribution of the FEV1 has the same standard deviation regardless of the height of the student.

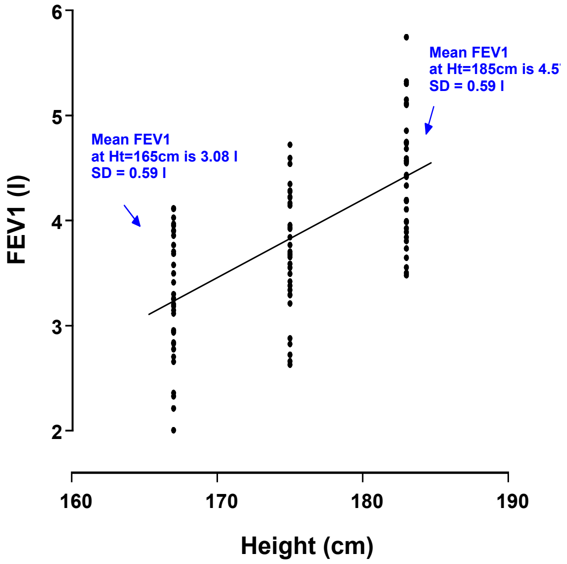
To summarise, we have approached the analysis of two variables by considering the following.

- The mean of the population of one variable depends linearly on the value of the other variable. So, e.g. we consider the FEV1s for the population of students who have a given height and assume that this *mean* varies linearly with height. We do *not* assume that the FEV1 of a given student depends on his height via a straight line relationship – a glance at Figure 1 b) shows that this would be untenable.

- We also assume that the spread of FEV1 about this mean is measured by a standard deviation, σ, about the line and that this does not change with height.

- Consequently, in a regression analysis there are three parameters to estimate – the intercept and slope of the line defining the mean, α and β (note now why Greek letters were chosen for these quantities) and the standard deviation about the line, σ.

In Figure 3 some artificial data have been generated using parameter values similar to those which obtain in Figure 1 b). This shows how data from the approach just outlined would look for three groups of students. Within each group all the students have the same height but the heights differ between the groups. Note that the *mean* FEV1 increases linearly with height but note that FEV1 values for individuals do not inevitable increase with the height of the individual. Note also that with the present approach the spread about the line is the same for all the heights.



Mean FEV1
at Ht=165cm is 3.08 l
SD = 0.59 l

Mean FEV1
at Ht=185cm is 4.5
SD = 0.59 l

Figure 3: artificial data from the standard regression set-up

Notice how this approach to analysing the data has introduced an asymmetry into the analysis – we are considering the distribution of FEV1 *given* the height of the student. While this sometimes means that regression is not the right tool to use in the analysis of two variables, it is far more common to find that the problem to hand is best posed in this asymmetric way.

What do we estimate?

For a single variable we can estimate the population mean, μ and population standard deviation, σ by the mean and standard deviation from a sample. In a regression analysis things are no different in principle. However the population mean now depends on height as $\alpha + \beta x$ and we now need to estimate two parameters, α and β, the population intercept and slope. If a line is fitted to the data as in Figure 2, then the algorithm will produce values for the slope and intercept of the least squares line fitted to the sample. So if the line fitted to the sample has equation $y = a + bx$ then a is the *sample* intercept and b is the *sample* slope and these estimate α and β respectively. Finally we need an estimate of the spread about the line, namely σ.

This is estimated by a quantity *s* related to the minimised sum of squares about the fitted line, although the details need not concern us.

It is worth noting that when you calculate these estimates, they should be given units. The intercept is on the same scale as the *y* variable and so has the same units. This is also true of the standard deviation about the line, The slope has units of *y* per *x*, so in the example in Figure 1 b) it will be in litres per cm.

How good are the estimates and testing a hypothesis

The intercept parameter α is often of little direct interest. In practice, because a regression analysis is aimed at elucidating the relationship between *y* and *x* it is the slope parameter, β, which is of the greatest interest. This is because it measures the rate at which the mean of the *y* variable changes as the *x* variable changes.

We saw in an earlier part of the course that it is important to know how good our sample mean, *m* is as an estimate of the population mean μ and that we could measure this by the standard error. In just the same way we can ask how good the sample slope, *b* is as an estimate of β. The answer is again a standard error, albeit one calculated using a slightly more complicated formula. The complexity need not concern us, as Minitab does the calculation for us and the interpretation of the standard error is the same as for the standard error of the mean.

So far we have discussed regression analysis in terms of the meaning of the approach, the parameters used and their estimation and little has been said about testing any hypotheses. This is probably as it should be, as hypothesis testing has a limited role here. The one hypothesis which is often of interest is whether β=0. The reason for this is that if β=0 then the mean of the *y* variable does not change with the *x* variable. In other words, at least in terms of this form of linear dependence, if β=0 then there is no association between the *y* and *x* variables. As with the *t*-test the appropriate standard error plays an important role in testing this hypothesis. The details are beyond our present scope and we leave Minitab to do the work.

Estimation in Minitab

If a linear regression is fitted in Minitab then the Session Window will look like that shown Figure 4.

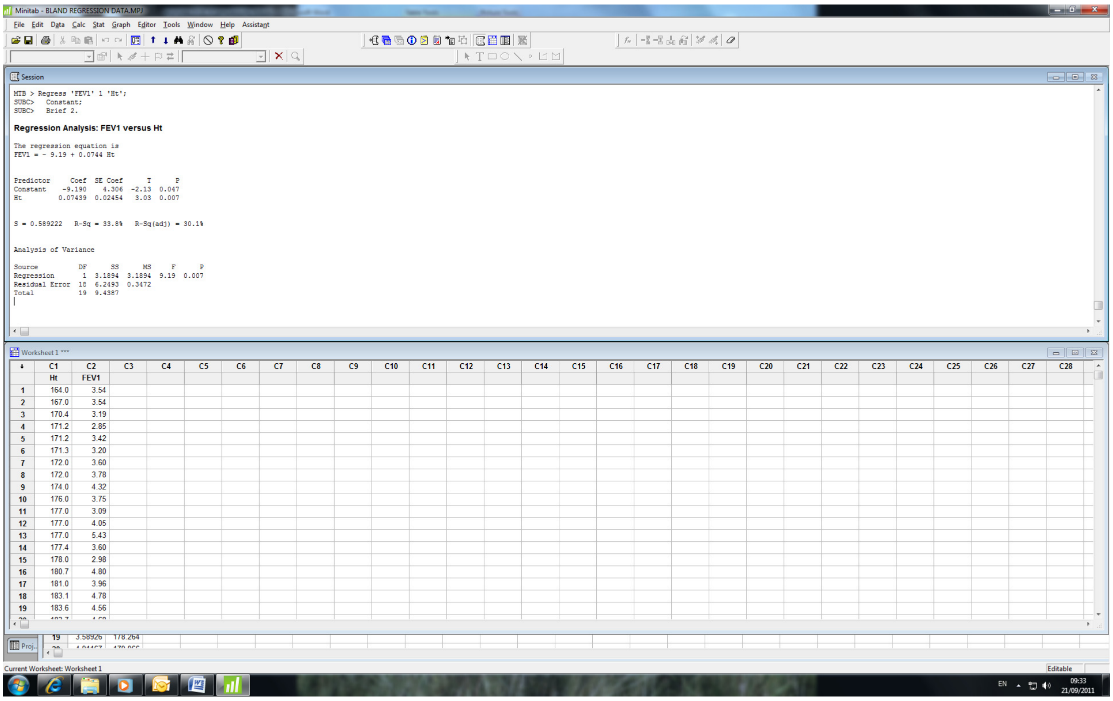
The regression command in Minitab will, in fact, fit a multiple regression with several *x* variables. While this does not concern us it explains some apparent redundancy in the output shown in Figure 4. All of the material from the heading `Analysis of Variance` can be ignored. When there is just one *x* variable (but not when there are more) this section merely repeats information already given in the first part of the output. The section `Unusual Observations` tries to draw attention to possibly erroneous or outlying values. While this is laudable, this particular automation of the important process of inspecting data for outlying values highlights rather too many points and we will not use it.

The key part of the output is the fitted line, reported as

```
The regression equation is
```

```
FEV1 = - 9.19 + 0.0744 Ht
```

Of course, this really should read 'Mean FEV1 =…' but the output overlooks this point. From this we see that the sample intercept, *a*, is -9.19 l and the sample slope, *b*, is 0.0744 l/cm.



```
MTB > Regress 'FEV1' 1 'Ht';
SUBC>   Constant;
SUBC>   Brief 2.
```

Regression Analysis: FEV1 versus Ht

```
The regression equation is
FEV1 = - 9.19 + 0.0744 Ht


Predictor     Coef  SE Coef      T      P
Constant    -9.190    4.306  -2.13  0.047
Ht         0.07439  0.02454   3.03  0.007


S = 0.589222   R-Sq = 33.8%   R-Sq(adj) = 30.1%


Analysis of Variance

Source          DF      SS      MS     F      P
Regression       1  3.1894  3.1894  9.19  0.007
Residual Error  18  6.2493  0.3472
Total           19  9.4387
```

↓	C1	C2
	Ht	FEV1
1	164.0	3.54
2	167.0	3.54
3	170.4	3.19
4	171.2	2.85
5	171.2	3.42
6	171.3	3.20
7	172.0	3.60
8	172.0	3.78
9	174.0	4.32
10	176.0	3.75
11	177.0	3.09
12	177.0	4.05
13	177.0	5.43
14	177.4	3.60
15	178.0	2.98
16	180.7	4.80
17	181.0	3.96
18	183.1	4.78
19	183.6	4.56

19	3.58926	178.264

Figure 4: Minitab output after fitting a regression line to the data from Figure 1 b)

The next part of the output, namely

```
Predictor      Coef  SE Coef      T      P
Constant     -9.190    4.306  -2.13  0.047
Ht          0.07439  0.02454   3.03  0.007
```

is also important.

The column headed `Coef` simply repeats the values of *a* and *b*, albeit with more significant figures. Note that the term `Constant` is used to identify what we have called the intercept. The standard error of *b* is given under the heading `SE Coef` as 0.02454. The test of the hypothesis β=0 is based on the *t*-statistic 3.03[†] given under `T`, with the corresponding P-value given as 0.007 under `P`.

The last part of the output which is of relevance is

```
S = 0.589222   R-Sq = 33.8%   R-Sq(adj) = 30.1%
```

The value of `S` (which is what we have called *s*), namely 0.589 l is the estimate of the standard deviation which measures the spread of the FEV1 about the fitted line.

[†] This is, in fact, `Coef/SE Coef` = 0.07439/0.02454.

In summary the key items we usually need to extract from the output are:

- the estimated slope and intercept, given under `Coef`;
- the standard error of the slope, given under `SE Coef`;
- the P-value for the test of the hypothesis $\beta=0$;
- the standard deviation about the line, given as `S`.

It is perhaps worth considering some points regarding these quantities.

1. The estimate of *a* is -9.19 l: how can something in litres be negative? Remember that the intercept is the mean FEV1 when the individual has height equal to 0. Consequently we should not expect the value to make sense directly. It is for this reason that little attention is focused on the intercept. If we omit the intercept then the fit of the line is distorted and we could get quite the wrong value for the slope, so the intercept needs to be there, it is just that its value needs to be interpreted appropriately.

2. The value P=0.007 indicates that the data provide very strong evidence that the mean FEV1 *does* depend on the height of the student, i.e. that the data discredit the hypothesis $\beta=0$. This does not mean that FEV1 is determined once height is known – Figure 1 b) shows this is far from the case. However it shows that the mean FEV1 changes with height, and the value of *b* shows that the mean FEV1 increases by about 0.074 l for each 1 cm increase in height.

3. The variation left in the FEV1 values once the effect of the height of the student has been taken into account is measured by *s*. In many analyses we make no explicit use of *s* but it is used implicitly when we test hypotheses, derive confidence intervals or assess predictions.

Using regression to make predictions

One of the applications of regression is its use to use one variable to predict the value of another. This sounds quite useful: some clinically important variable might be difficult or very invasive to measure and predicting it from other, related variables would be attractive. Equally variables which will only become apparent in the future, such as survival time, might be predicted from variables known at presentation. In practice these advantages are more apparent than real. Any proper attempt at prediction needs to take into account the natural variability in the system and this often places wide limits on the prediction made for an individual. Despite this a more detailed consideration of prediction methods is instructive.

On the basis of the sample in Figure 1 b) how would we predict FEV1 for a student if we knew nothing about the student? The best value we could use would be the mean of the FEV1 values, namely 3.86 l. How would this change if we were told that the student had a height of h cm – e.g. we might be told h = 180 cm?

A student with height of 180 cm is amongst the taller students, so we would expect them to have an FEV1 above the average. In one sense this does not alter our approach - we would still use the mean but we could now use the regression analysis to quote a mean that is specifically for those students with this height. Our approach assumes that the mean FEV1 for a student of height h is $\alpha + \beta h$, which in this example would be $\alpha + 180\times\beta$.

Of course, as $\alpha + 180\times\beta$ depends on the *parameters* α and β this cannot be calculated because we never know the value of a parameter. Instead we have to use the estimates of α and β, a and b which we have obtained from the regression analysis. So the prediction we can calculate is:

$$a + 180\times b = -9.19 + 180 \times 0.07439 = 4.20 \text{ l.}$$

As we might expect for a taller than average student, his FEV1 is also above the average. This simple use of the estimated regression equation is all that is needed for a prediction comprising a single value.

Intervals for prediction

The matter becomes more complicated if we wish to place limits on this prediction. The problem is that two possible pairs of limits can be computed and their interpretation is very different.

The first type of interval is a confidence interval – exactly the same in principle as we have already calculated for a mean of a single variable. In this case we acknowledged the uncertainty in the sample mean, m say, as an estimate of μ, by computing an interval (m_L, m_U) within which we 'expect' μ to lie. An important observation here is that as the size of the sample from which the interval is calculated increases so the width of the interval decreases. This reflects the fact that from a bigger sample we will have a more precise estimate of μ. The width of the interval reflects only one source of variation, namely the uncertainty of m as an estimate of μ. The situation is the same in regression – a and b are estimates of α and β and this uncertainty can be quantified. If we estimate the mean of the population of students with height h by $a + bh$ then the uncertainty in a and b will naturally give rise to uncertainty in this estimate and this can be quantified through a confidence interval. Actually computing the confidence interval is a little more complicated and has some features which need a little thought to understand. For example the width of the interval varies with the value of h. However, we do not dwell on this feature. As in the case for a single variable, the larger the sample the narrower will be the confidence interval.

In most cases where prediction is contemplated it is likely to be applied to an individual. While a prediction consisting of a single value is useful, it is much more helpful if an interval can be given in which the value for the individual is likely to lie. It is very important when computing such an interval that it is not confused with a confidence interval. That a confidence interval is inappropriate can be appreciated by realising that it makes no sense computing an interval for an individual which can be made as precise as possible by making the sample on which *a* and *b* are estimated as large as possible. To appreciate this consider figure 2 b). Even once the height of the student is prescribed there remains considerable variation in the value of FEV1, so any sensible interval for the prediction of the value for an individual cannot shrink to zero width under any circumstances.

The issue is that while the prediction for an individual is $a + bh$, the variability of this quantity, *as a prediction for an individual*, comes not only from the uncertainty in *a* and *b* as estimates of α and β, but also from the variability an individual will have about its mean – something measured by the estimate of the standard deviation about the line, *s*. This gives rise to a different way to compute the interval, and a different name for the interval: the terms *prediction interval* or *tolerance interval* are encountered. Minitab will compute such interval and indeed can graph the intervals across a range of heights *h*.

Example

For a student with height 180 cm we have seen that the estimated FEV1 is 4.20 l. The intervals Minitab computes are:

95% confidence interval	95% prediction interval
3.83, 4.57 l	2.91, 5.49 l

The 95% confidence interval indicates that the mean FEV1 for students with height 180 cm is within 3.83l to 4.57l with 95% confidence. The prediction interval indicates that 95% of students with height 180 cm have FEV1 values between 2.91 l and 5.49 l.

Figure 5 shows these intervals, together with the raw data and fitted line, for all heights *h* from 165 to 185 cm. Note that the confidence intervals are curved, indicating that our estimate of the mean FEV1 is better near the 'centre' of the data. The prediction interval is almost a straight line because the interval is dominated by the intrinsic variability in the data.

Using a computer to generate data that are similar to that in Figure 1 b), but now with a sample of size 1000 rather than 20, the prediction for the student with height 180 cm becomes 4.18 l and the intervals become

95% confidence interval	95% prediction interval
4.14, 4.23 l	3.08, 5.29 l

The prediction interval has changed little but the confidence interval is much narrower. This illustrates the fact that we now have much more information about the mean FEV1 at this height but that the natural variability of students remains largely unchanged.

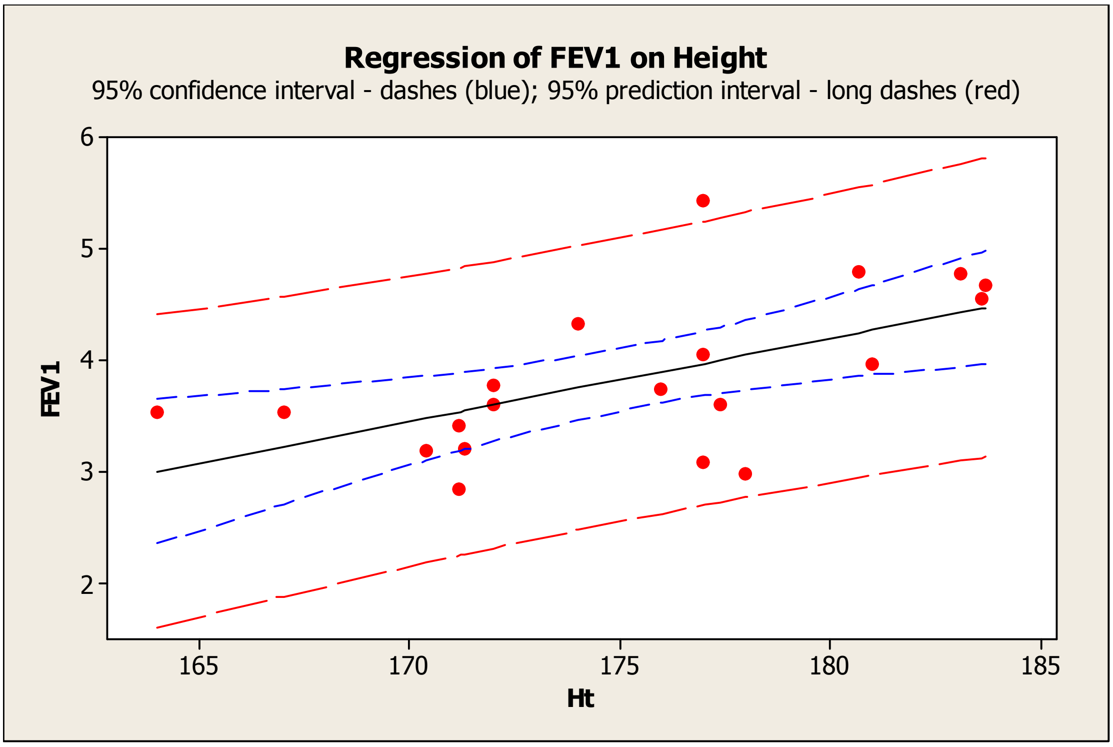


Figure 5: 95% confidence and prediction intervals for the data in figure 1 b)

Pitfalls and Assumptions

Pitfalls

As with any statistical technique, the method can be applied inappropriately. Some comments of relevance to this are given below.

1. The technique, as usually applied in the medical sciences, is rather a weak, empirical technique which deduces a relationship between two variables from that which can be apprehended in the sample itself. There are some exceptions, for example in pharmacology where a regression analysis might be guided by some sort of compartmental model.

2. It follows that it would be unwise to use the regression line, such as that fitted in Figure 2 b), for other kinds of data, even if they appear similar. So, for example, the line fitted in Figure 2 b) should not be extrapolated for use in, e.g., children, older males or female students.

3. Beware of outlying or unusual values, as these can have a noticeable influence on the estimates you obtain. Problems of this nature can often be identified when assessing the assumptions underlying the technique, which will be discussed below.

4. The regression of FEV1 on Height is shown in Figure 4 and gives rise to the equation which might loosely be written:
   FEV1 = $a + b\ h$ or in this instance FEV1 = -9.19 + 0.07439 h

It is tempting to rearrange this equation to obtain

$h = -a/b + 1/b$ FEV1 or in this instance $h = 123.54 + 13.44$ FEV1

However, if you perform the regression of height on FEV1 in Minitab you obtain the output in Table 6, which gives a different equation. Why is this?

The answer lies in the looseness of the above equation. It should have read Mean FEV1 = $a + b\ h$ and so the rearrangement would have led to the equation $h = -a/b + 1/b$ Mean FEV1, which is not in the form Mean h = $c + d$ FEV1.

The regressions of FEV1 on height and of height on FEV1 are quite different[†] and essentially address different questions. It is up to the analyst to ensure that the correct approach is chosen.

Regression Analysis: Ht versus FEV1

```
The regression equation is
Ht = 158 + 4.54 FEV1


Predictor      Coef  SE Coef      T      P
Constant    157.864    5.870  26.89  0.000
FEV1          4.542    1.499   3.03  0.007
```

Figure 6: Minitab output for regression of Height on FEV1

Assumptions

Problems can also arise in the application of regression because the method makes some assumptions and in particular instance s these may not be met. Assessing the assumptions gives rise to a large and important field of statistics known as *regression diagnostics* but we will consider only some simple aspects here.

Before indicating how the assumptions are checked it is useful to reiterate what the assumptions are.

1. That the *mean* of the y variable at a given value of the x variable changes linearly with x.

[†] Not wholly different – there are some points of overlap. For example the P-value for the slope is the same in Figure 4 and Table 6. We will not pursue this point here.

2. The spread of the data about this line is constant, that is it does not change as *x* changes.

3. The deviations from the line follow a Normal distribution (strictly this assumption is only needed if you intend to compute confidence or prediction intervals for the estimates or predictions or to perform hypothesis tests such as testing β=0).

Assessing the linearity assumption: an essential first step is to draw a scatter plot of the *y* variable against the *x* variable. This can be assessed by eye to see if the assumption is plausible and that no other form of relationship is suggested by the data. Techniques which go beyond this are surprisingly difficult and quickly become rather technical.

Assessing the spread about the line: this can also be assessed from a scatterplot but defining quantities known as *residuals* helps here. For each point in the dataset there is a residual. They are shown diagramatically in Figure 2 b): the residual is the vertical distance of a point from the fitted line, i.e. the lengths of the dashed red lines in the Figure. Residuals are positive for points above the line and negative for points below the line. The reason residuals are important in regression diagnostics is that if the fitted line truly reflects the structure of the data then the residuals are a sample from a distribution with population mean equal to zero and they all have the same standard deviation. The most useful way residuals are used is graphically. For example if the assumption of constant standard deviation is true then a plot of the residuals against the height of the individual should show points with a spread that changes little with height. The method will be exemplified in the practical session.

Assessing the Normality of the deviations from the line: as the residuals are essentially the deviations from the line then this assessment amounts to checking that the residuals come from a common Normal distribution. As described at the beginning of the course the best way of doing this is using a Normal Probability plot for the residuals.

Notice how no assumptions are made concerning the *x* variable. It can be discrete or continuous, observed from a population or determined by the experimenter.

Correlation

Correlation is a statistical concept which goes together with regression and the two are closely related. The idea behind correlation is that there should be some way to quantitatively express the difference between the data shown in Figure 7. The data in the left-hand panel seems to conform to a straight line more closely than the data in the right-hand panel. A regression line could be fitted to either panel but there is a difference between the two sets of data and there may be circumstances when quantifying this is useful. The correlation coefficient attempts to do this and it will now be described.

There are several correlation coefficients defined in the literature. The one we consider is the most commonly encountered and is known as the *product-moment correlation* or the *Pearson correlation* and is universally given the symbol *r*. Strictly speaking, and in keeping with our convention on Greek and Roman letters, this is the *sample* correlation and the underlying population correlation is given the Greek letter ρ (rho).

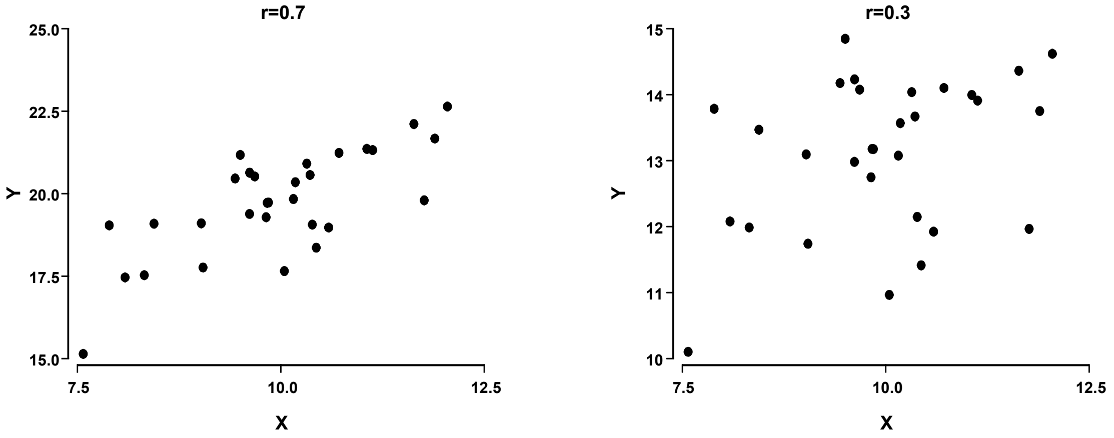


Figure 7: two sets of fictitious data showing (left) closer conformity to a line and (right) greater dispersion.

Properties of the correlation

The correlation coefficient *r* has the following properties:

1. it always takes values between -1 and 1;
2. if the points were to lie exactly on a straight line then *r* would be either -1 or 1;
3. a value of 0 corresponds to no *linear* relation between the variables;
4. it can be computed for data which comprise pairs of continuous variables.

Negative values of *r* correspond to data in which the *y* variable tends to decrease as the *x* variable increases, whereas for positive values of *r* the *y* and *x* variables tend to increase or decrease together. The ‘strength’ of relationship between the variables is unaffected by the sign of *r*, which simply reflects the direction of the relationship. This is illustrated in Figure 8, where the data on the left have correlation -0.7 and those on the right 0.7. Similarly data with correlation 1 lie exactly on a straight line with positive slope whereas data with *r*=-1 lie on a line with negative slope.

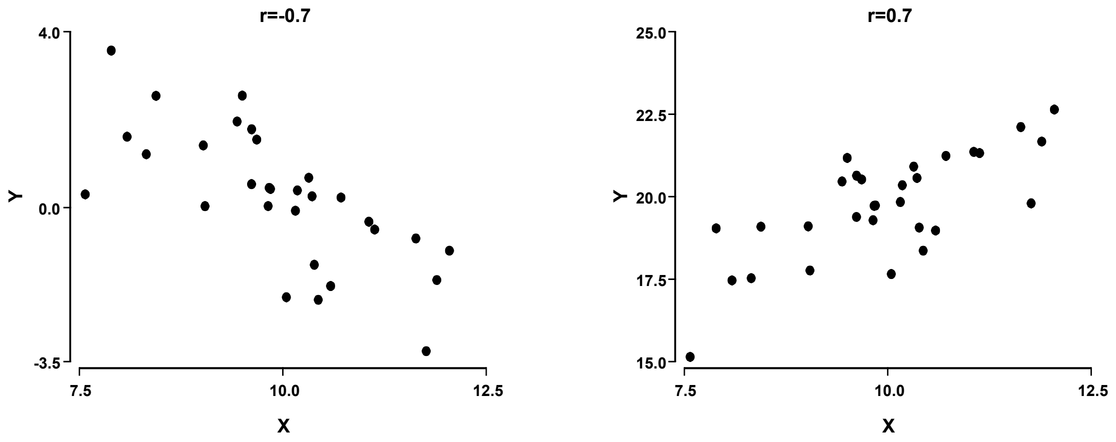


Figure 8: two sets of fictitious data with (left) correlation -0.7 and (right) correlation 0.7.

One of the problems with the correlation coefficient is that while the extreme values are readily interpreted, matters are less easy with intermediate values. In Figure 9 data with correlation coefficients 0 and 0.3 are shown and it is clear that there is little to distinguish the two.

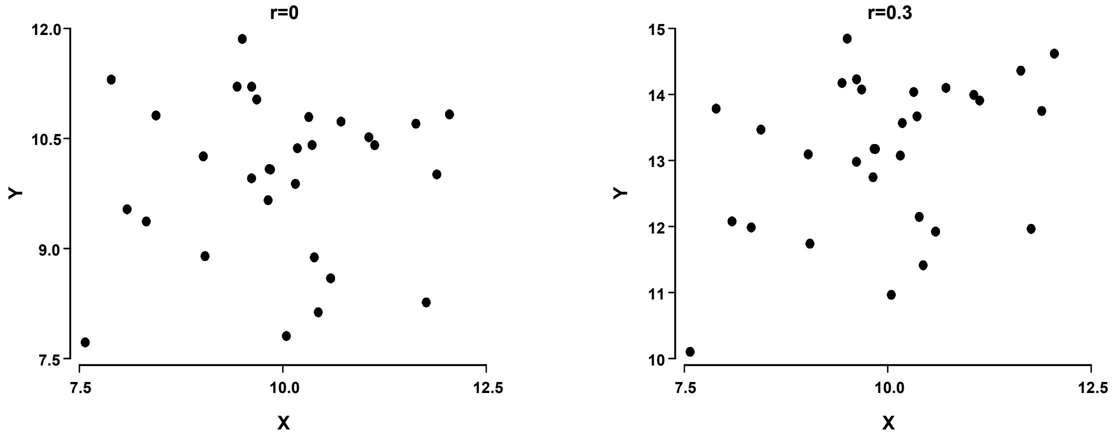


Figure 9: two sets of fictitious data with (left) correlation 0 and (right) correlation 0.3.

Hypothesis tests for the correlation coefficient

Statements such as '… the variables exhibited a correlation r=0.35 (P=0.07)' are often encountered in the literature. The P-value indicates that a hypothesis has been tested, but which one? In fact it is a test that the population correlation ρ=0, i.e. that there is no (linear) relation between the variables. In fact, this turns out to be exactly the same test as testing β=0 in a regression of y on x, so the test and the P value is the same as would be obtained in an output such as that in Figure 4. In order to perform this test it is necessary that one of the variables has a Normal distribution. If both variables are Normal then a confidence interval for ρ can be computed but as the value of the correlation coefficient is awkward to interpret it is not much used.

One problem with the practical use of the correlation coefficient is that the difficulty of interpreting the value of the correlation gives rise to a tendency among users to use the test of ρ=0 in dichotomous way – as establishing that there is or there is not a relationship between variables. This is unfortunate because even apparently quite weak levels of correlation, such as those seen the right hand panel of Figure 9 can be deemed significant if the sample size if large enough. For example, a value of r=0.3 will give P<0.05 if the sample is larger than 41. It is important to realise that such a ‘significant’ result essentially discredits the hypothesis that the two variables are unrelated, not that they are necessarily closely related. When interpreting a correlation coefficient it is useful to bear the right hand panel of Figure 9 in mind as a reminder of the difficulty in interpreting intermediate values of r.

In general terms care needs to be exercised in the use of correlation. The statistical relationship between two variables is often too complicated to permit its summary by a single value. While this level of succinctness may be useful when there are many variables to consider, regression usually provides a more comprehensible method for assessing how two variables are related.