Clincial Biostatistics

Regression

Regression analyses

Regression is the rather strange name given to a set of methods for predicting one variable from another. The data shown in Table 1 and come from a student project aimed at estimating body mass index (BMI) using only a tape measure. In the full data, analysed later, we have abdominal circumference, mid upper arm circumference, and sex as possible predictors. We shall start with the female subjects only and will look at abdominal circumference.

BMI, also known as Quetelet's index, is a measure of fatness defined for adults as weight in Kg divided by abdominal circumference in metres squared. Can we predict BMI from abdominal circumference? Figure 1 shows a scatter plot of BMI against abdominal circumference and there is clearly a strong relationship between them. We could try to draw a line on the scatter diagram which would represent the relationship between them and enable us to predict one from the other. We could draw many lines which might do this, as shown in Figure 2, but which line should we choose? The method which we use to do this is **simple linear regression**. This is a method to predict the mean value of one variable from the observed value of another. In our example we shall estimate the mean BMI for women of any given abdominal circumference measurement.

We do not treat the two variables, BMI and abdominal circumference, as being of equal importance, as we did for correlation coefficients. We are predicting BMI from abdominal circumference and BMI is the **outcome**, **dependent**, **y**, or **left hand side** variable. Abdominal circumference is the **predictor**, **explanatory**, **independent**, **x**, or **right hand side** variable. Several different terms are used. We predict the outcome variable from the observed value of the predictor variable.

The relationship we estimate is called **linear**, because it makes a straight line on the graph. A linear relationship takes the following form:

$$\text{BMI} = \text{intercept} + \text{slope} \times \text{abdominal circumference}$$

the intercept and slope are numbers which we estimate from the data. Mathematically, this is the equation of a straight line. The **intercept** is the value of the outcome variable, BMI, when the predictor, abdominal circumference, is zero. The **slope** is the increase in the outcome variable associated with an increase of one unit in the when the predictor.

Table 1. Weight and abdominal circumference in 86 women (data of Malcolm Savage)

Abdominal circum-ference (cm)	BMI (Kg/ht^2)	Abdominal circum-ference (cm)	BMI (Kg/ht^2)	Abdominal circum-ference (cm)	BMI (Kg/ht^2)
51.9	16.30	64.2	19.44	73.1	20.25
53.1	19.70	64.4	19.31	73.2	21.07
54.3	16.96	64.4	18.15	73.2	24.57
57.4	11.99	64.7	20.55	74.0	20.60
57.6	14.04	64.8	15.70	74.1	16.86
57.8	15.16	65.0	18.73	74.4	22.58
58.2	16.31	65.2	18.52	74.7	21.42
58.2	16.17	65.6	21.08	74.8	23.11
59.0	20.08	66.2	17.58	74.8	24.11
59.2	14.81	66.8	18.51	79.3	19.71
59.5	18.02	66.9	18.75	79.7	23.14
59.8	18.43	67.0	19.68	80.0	19.48
59.8	15.50	67.5	18.06	80.3	23.28
60.2	17.64	67.8	21.12	80.4	22.59
60.2	17.54	67.8	20.60	82.2	28.78
60.4	14.18	68.0	19.40	82.2	25.89
60.6	17.41	68.2	22.11	83.2	25.08
60.7	19.44	68.6	19.23	83.9	27.41
61.2	21.63	69.2	19.49	85.2	22.86
61.2	15.55	69.2	20.12	87.8	32.04
61.4	18.37	69.2	24.06	88.3	25.56
62.4	17.69	69.4	19.97	90.6	28.24
62.5	17.64	70.2	19.52	93.2	28.74
63.2	18.70	70.3	23.77	100.0	31.04
63.2	20.36	70.9	18.90	106.7	30.98
63.2	18.04	71.0	20.89	108.7	40.44
63.2	18.04	71.0	17.85		
63.4	17.22	71.2	21.02		
63.8	18.47	72.2	19.87		
64.2	17.09	72.8	23.51		

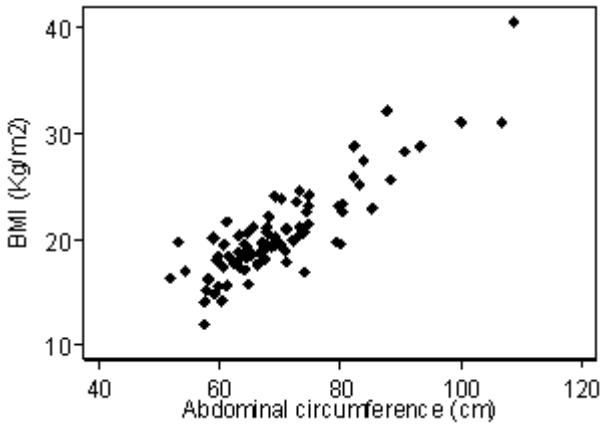

Figure 1. Scatter plot of BMI against abdominal circumference

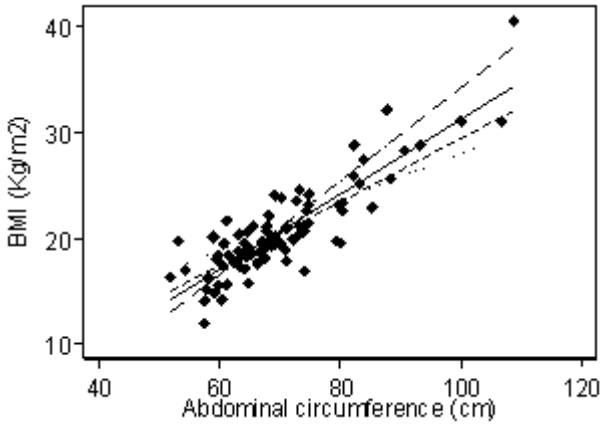

Figure 2. Scatter plot of BMI against abdominal circumference with possible lines to represent the relationship

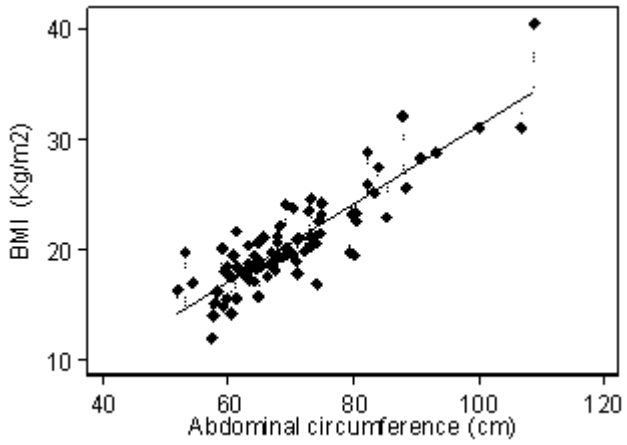

Figure 3. Differences between the observed and predicted values of the outcome variable

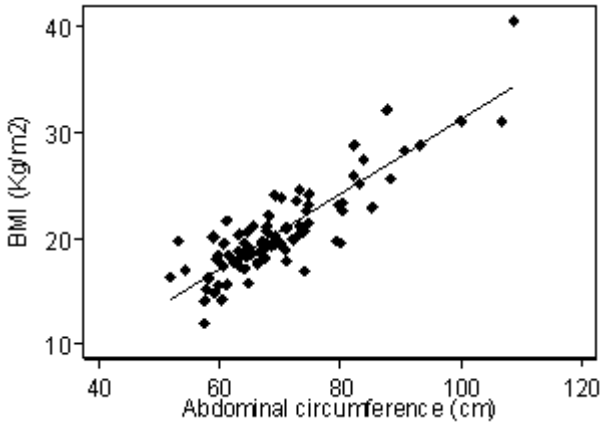

Figure 4. The least squares regression line for BMI and abdominal circumference

To find a line which gives the best prediction, we need some criterion for best. The one we use is to choose the line which makes the distance from the points to the line ***in the y direction*** a minimum. These are the differences between the observed BMI and the BMI predicted by the line. These are shown in Figure 3. If the line goes through the cloud of points, some of these differences will be positive and some negative. There are many lines which will make the sum zero, so we cannot just minimise the sum of the differences. As we did when estimating variation using the variance and standard deviations (Week 1) we square the differences to get rid of the minus signs. We choose the line which will minimise the sum of the squares of these differences. We call this the **principle of least squares** and call the estimates that we obtain the **least squares line** or equation. We also call this estimation by **ordinary least squares** or **OLS**.

There are many computer programs which will estimate the least squares equation and for the data of Table 1 this is

$$\text{BMI} = -4.15 + 0.35 \times \text{abdominal circumference}$$

This line is shown in Figure 4. The estimate of the slope, 0.35, is also known as the **regression coefficient**. Unlike the correlation coefficient, this is not a dimensionless number, but has dimensions and units depending on those of the variables. The regression coefficient is the increase in BMI per unit increase in abdominal circumference, so is in kilogrammes per square metre per centimetre, BMI being in Kg/m^2 and abdominal circumference in cm. If we change the units in which we measure, we will change the regression coefficient. For example, it we measured abdominal circumference in metres, the regression coefficient would be 35 $Kg/m^2/m$. The intercept is in the same units as the outcome variable, here Kg/m^2.

In this example, the intercept is negative, which means that when abdominal circumference is zero the BMI is negative. This is impossible, of course, but so is zero abdominal circumference. We should be wary of attributing any meaning to an intercept which is outside the range of the data. It is just a convenience for drawing the best line within the range of data that we have.

Confidence intervals and P values in regression

We can find confidence intervals and P values for the coefficients subject to assumptions. These are that deviations from line should have a Normal distribution with uniform variance. (In addition, as usual, the observations should be independent.)

For the BMI data, the estimated slope = 0.35 $Kg/m^2/cm$, with 95% CI = 0.31 to 0.40 $Kg/m^2/cm$, $P<0.001$. The P value tests the null hypothesis that in the population from which these women come, the slope is zero. The estimated intercept = –4.15 Kg/m^2, 95% CI = –7.11 to –1.18 Kg/m^2. Computer programs usually print a test of the null hypothesis that the intercept is zero, but this is not much use. The P value for the slope is exactly the same as that for the correlation coefficient.

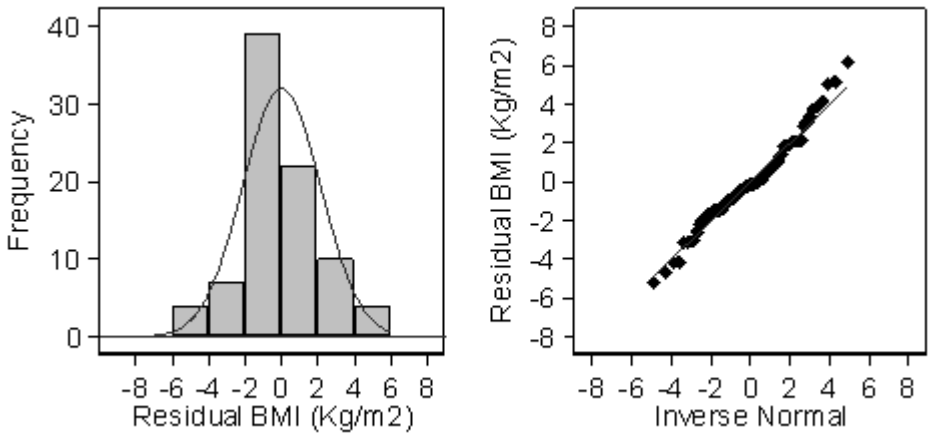

Figure 5. Histogram and Normal plot for residuals for the BMI and abdominal circumference data

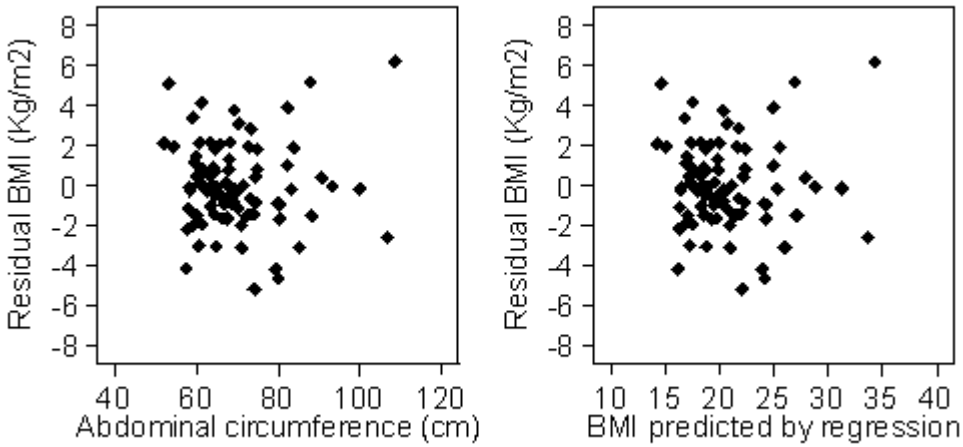

Figure 6. Scatter plot of residual BMI against abdominal circumference and against the regression estimate

Table 2. 24 hour energy expenditure (MJ) in groups of lean and obese women (Prentice *et al.*, 1986, cited by Altman, 1991)

Lean	Obese
6.13	8.79
7.05	9.19
7.48	9.21
7.48	9.68
7.53	9.69
7.58	9.97
7.90	11.51
8.08	11.85
8.09	12.79
8.11	
8.40	
10.15	
10.88	

Testing the assumptions of regression

For our confidence intervals and P values to be valid, the data must conform to the assumptions that deviations from line should have a Normal distribution with uniform variance. The observations must be independent, as usual. Finally, our model of the data is that the line is straight, not curved, and we can check how well the data match this.

We can check the assumptions about the deviations quite easily using techniques similar to those used for t tests. First we calculate the differences between the observed value of the outcome variable and the value predicted by the regression, the regression estimate. We call these the **deviations from the regression line**, the **residuals about the line**, or just **residuals**. These should have a Normal distribution and uniform variance, that is, their variability should be unrelated to the value of the predictor.

We can check both of these assumptions graphically. Figure 5 shows a histogram and a Normal plot for the residuals for the BMI data. The distribution is a fairly good fit to the Normal. We can assess the uniformity of the variance by simple inspection of the scatter diagram in Figure 4. There is nothing to suggest that variability increases as abdominal circumference increases, for example. It appears quite uniform. A better plot is of residual against the predictor variable, as shown in Figure 6. Again, there is no relationship between variability and the predictor variable. Figure 6 also shows a plot of the residual against the regression estimate, the value predicted by the regression. Some books prefer this version of the plot. As you can see, the actual plot is identical, only the horizontal scale is changed. The plot of residual against predictor should show no relationship between mean residual and predictor if the relationship is actually a straight line. If there is such a relationship, usually that the residuals are higher or lower at the extremes of the plot than they are in the middle, this suggests that a straight line is not a good way to look at the data. A curve might be better.

Dichotomous predictor variables

Table 2 shows 24 hour energy expenditure (MJ) in groups of lean and obese women. In week 4, we analysed these data using the two sample t method. We can also do this by regression. We define a variable = 1 if the woman is obese, = 0 if she is lean.

If we carry out regression:

energy = 8.07 + 2.23 × obese

slope: 95% CI = 1.05 to 3.42 MJ, P=0.0008.

Compare this with the two sample t method:

Difference (obese – lean) = 10.298 – 8.066 = 2.232.

95% CI = 1.05 to 3.42 MJ, P=0.0008.

The two methods give identical results. They are shown graphically in Figure 7.

The assumptions of two sample t method are that

1. energy expenditure follows a Normal distribution in each population,
2. variances are the same in each population.

The assumptions of regression are that

1. differences between observed and predicted energy expenditure follow a Normal distribution,
2. variances of differences are the same in whatever the value of the predictor.

These are the same. The energy expenditure predicted for a group by the regression is equal to the mean of the group and if the differences from the group mean follow a Normal distribution, then so do the residuals about the regression line, which goes through the two means. If variances are the same for the two values of the predictor, then they are the same in the two groups.

Multiple regression

In this section I expand the idea of regression to describe using more than one predictor variable.

I illustrated simple linear regression using the prediction of body mass index (BMI) from abdominal circumference in a population of adult women. Figure 8 shows scatter diagrams of BMI against abdominal circumference and of BMI against mid upper arm circumference. This time both men and women are included in the sample. The regression equations predicting BMI from abdominal circumference and from mid upper arm circumference are:

	BMI =	–1.35 +	0.31×abdomen
95% CI		–3.49 to 0.78	0.28 to 0.33
			P<0.001

	BMI =	–4.59 +	1.09×arm
95% CI		–7.12 to –2.07	0.98 to 1.20
			P<0.001

Both abdominal and arm circumference are highly significant predictors of BMI. Could we get an even better prediction if we used both of them? Multiple regression enables us to do this. We can fit a regression equation with more than one predictor:

	BMI =	–5.94 +	0.18×abdomen	+	0.59×arm
95% CI		–8.10 to –3.77	0.14 to 0.22		0.45 to 0.74
			P<0.001		P<0.001

This multiple regression equation predicts BMI better than either of the simple linear regressions. We can tell this because the standard deviation of the residuals, what is left after the regression, is 2.01 Kg/m^2 for the regression on abdomen and arm together, whereas it is 2.31 and 2.36 Kg/m^2 for the separate regressions on abdomen and on arm respectively.

The regression equation was found by an extension of the least squares method described for simple linear regression. We find the coefficients which make the sum of the squared differences between the observed BMI and that predicted by the regression a minimum. This is called **ordinary least squares** regression or **OLS** regression.

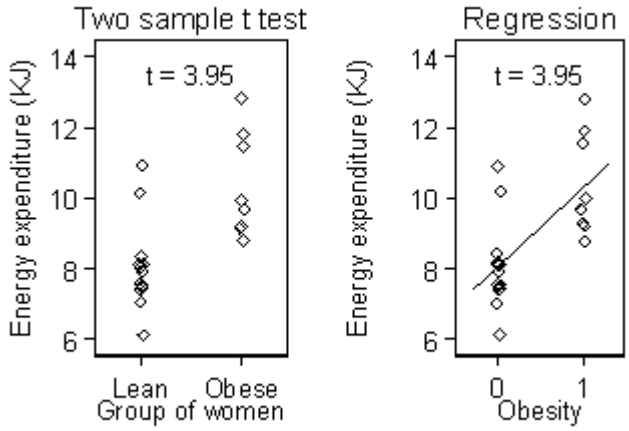

Figure 7. Equivalence of regression and two sample t method for comparing the mean energy expenditure in two groups of women

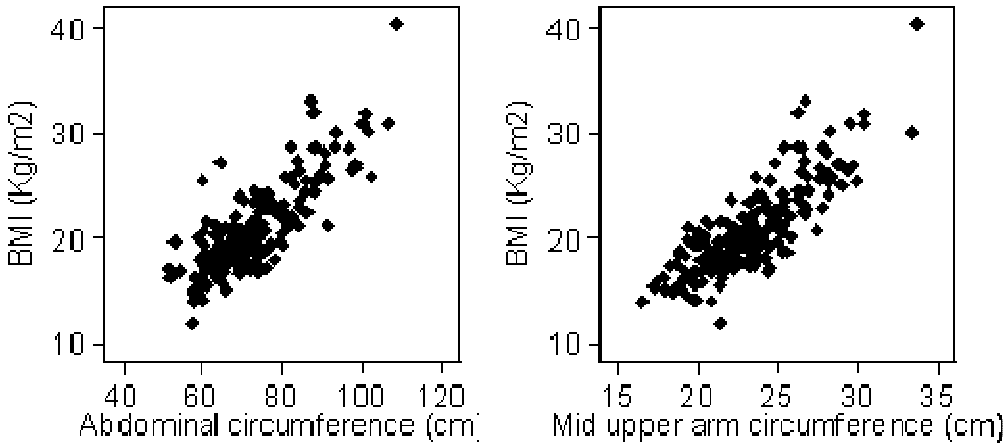

Figure 8. BMI against abdominal circumference and arm circumference in 202 adults

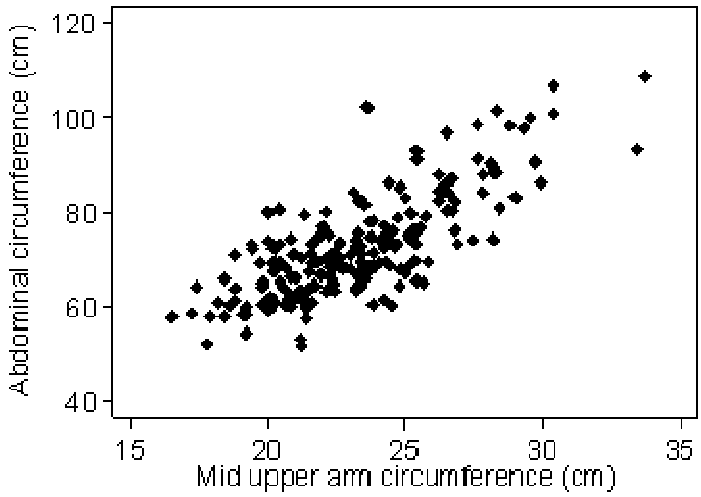

Figure 9. Abdominal circumference against mid upper arm circumference in 202 adults

Although both variables are highly significant, the coefficient of each has changed. Both coefficients have got closer to zero, going from 0.305 to 0.178 for abdomen and from 1.089 to 0.582 for arm circumference. The reason for this is that abdominal and arm circumferences are themselves related, as Figure 9 shows. The correlation is r = 0.77, P<0.001. Abdominal and arm circumferences each explains some of the relationship between BMI and the other. When we have only one of them in the regression, it will include some of the relationship of BMI with the other. When both are in the regression, each appears to have a relationship which is less strong than it really is.

Each predictor also reduces the significance of the other because they are related to one another as well as to BMI. We cannot see this from the P values, because they are so small, but the t statistics on which they are based are 20.64 and 19.97 for the two separate regressions and 8.80 and 8.09 for the multiple regression. Larger t statistics produce smaller P values. It is quite possible for one of the variables to become not significant as a result of this, or even for both of them to do so. We usually drop variables which are not significant out of the regression equation, one at the time, the variable with the highest P value first, and then repeat the regression.

There is another possible predictor variable in the data, sex. Figure 10 shows BMI for men and women. This difference is not significant using regression of BMI on sex, or an equivalent two sample t test, P = 0.5. If we include sex in the regression, as described for the energy expenditure data, using the variable 'male' = 1 if male and = 0 if female, we get

BMI	=	–6.44	+	0.18×abdomen	+	0.64×arm	–	1.39×male
95% CI		–8.49 to –4.39		0.14 to 0.22		0.50 to 0.78		–1.94 to –0.84
				P<0.001		P<0.001		P<0.001

This time the coefficients, confidence intervals and, although you can't tell, the P values, for abdomen and arm are hardly changed. This is because neither is closely related to sex, the new variable in the regression. Male has become significant. This is because including abdominal and arm circumference as predictors removes so much of the variation in BMI that the relationship with sex becomes significant. Mean BMI is lower for men than women ***of the same abdominal and arm circumference*** by 1.39 units. When we have continuous and categorical predictor variables together, regression is also called **analysis of covariance** or **ancova**, for historical reasons.

Testing the assumptions of multiple regression

We have to make the same assumptions for multiple linear regression as for simple linear regression. For our confidence intervals and P values to be valid, the data must conform to the assumptions that deviations from line should have a Normal distribution with uniform variance. The observations must be independent. Finally, our model of the data is that the relationship with each of our predictors is adequately represented by a straight line rather than a curve.

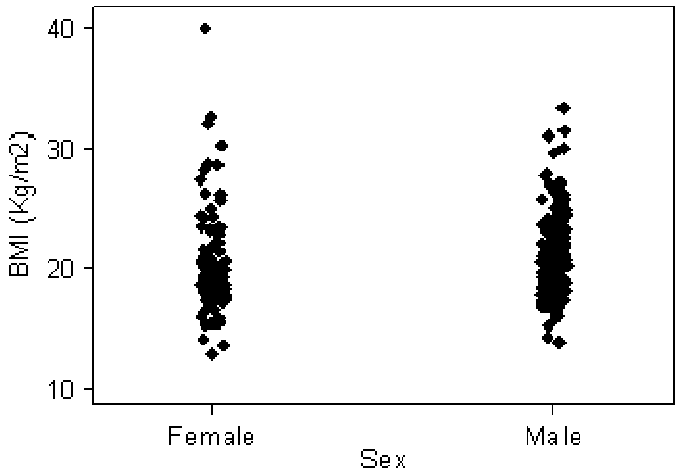

Figure 10. BMI by sex in 202 adults

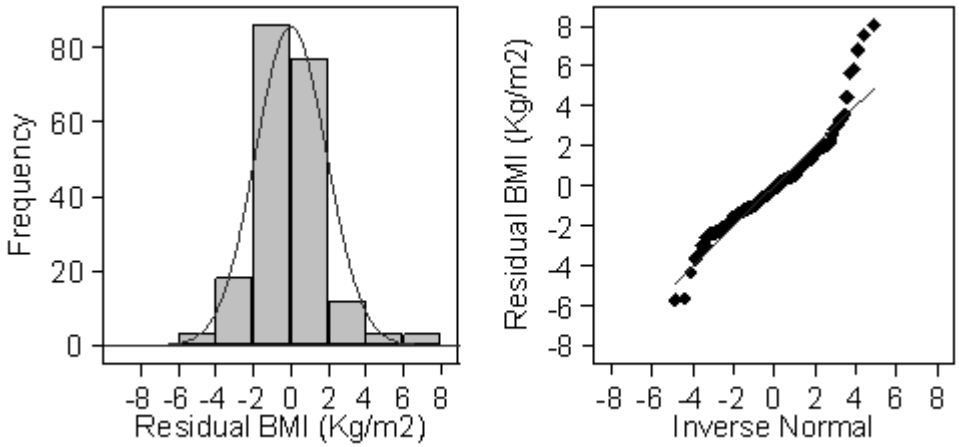

Figure 11. Residual BMI after regression on abdominal and arm circumference and sex, for 202 adults

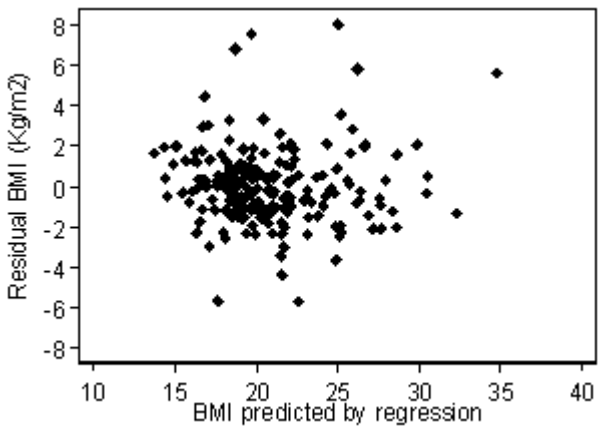

Figure 12. Residual BMI after regression on abdominal and arm circumference and sex against the regression estimate, for 202 adults

We can check these assumptions in the same way as we did for simple linear regression. First we calculate the residuals, the differences between the observed value of the outcome variable and the value predicted by the regression. These should have a Normal distribution and uniform variance, that is their variability should be unrelated to the value of the predictors. We can use a histogram and a Normal plot to check the assumptions of a Normal distribution (Figure 11). For these data, there is a small departure from a Normal distribution, because the tails are longer than they should be. This is seen both in the histogram and by the way the Normal plot departs from the straight line at either end. There is little skewness, however, and regression is fairly robust to departures from a Normal distribution. It is difficult to transform to remove long tails on either side of the distribution. If we plot the residual against the predicted value, the regression estimate, we can see whether there is an increase in variability with increasing magnitude. Figure 12 shows that there is no such relationship in this example.

When there are departures from the Normal distribution or uniform variance, we can try to improve matters by a suitable transformation of the outcome variable (Week 5). These problems usually go together and a transformation which removes one usually removes the other as well. I give an example for the asthma trial below.

Regression lines which are not straight

We can fit a curve rather than a straight line quite easily. All we need to do is to add another term to the regression. For example, we can see whether the relationship between BMI and abdominal circumference is better described by a curve. We do this by adding a variable equal to the square of abdominal circumference:

BMI =	16.03	–	0.16 × abdomen	+	0.0030 × abdomen^2
95% CI	4.59 to 27.47		–0.45 to 0.14		0.0011 to 0.0049
			P=0.3		P=0.003

The abdomen variable is no longer significant, because the abdomen and the abdomen squared are very highly correlated, which makes the coefficients difficult to interpret. We can improve things by subtracting a number close to the mean abdominal circumference. This makes the slope for abdomen easier to interpret. In this case, the mean abdominal circumference is 72.35 cm, so I have subtracted 72 from before squaring:

BMI =	0.59	+	0.27 × abdomen	+	0.0030 × (abdomen – 72)^2
95% CI	–1.85 to 3.03		0.24 to 0.31		0.0011 to 0.0049
			P<0.001		P=0.003

The coefficient for the squared term is unchanged, but the linear term is changed. We have evidence that the squared term is a predictor of BMI and we could better represent the data by a curve. This is shown in Figure 13.

Using multiple regression for adjustment

You will often see the words ‘adjusted for’ in reports of studies. This almost always means that some sort of regression analysis has been done, and if we are talking about the difference between two means this will be multiple linear regression.

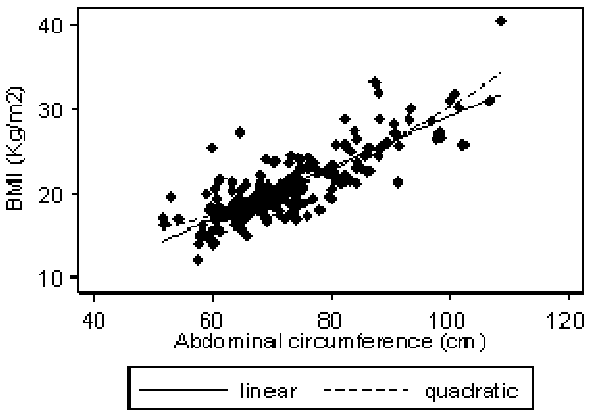

Figure 13. BMI and abdominal circumference, showing the simple linear regression line and the quadratic curved line

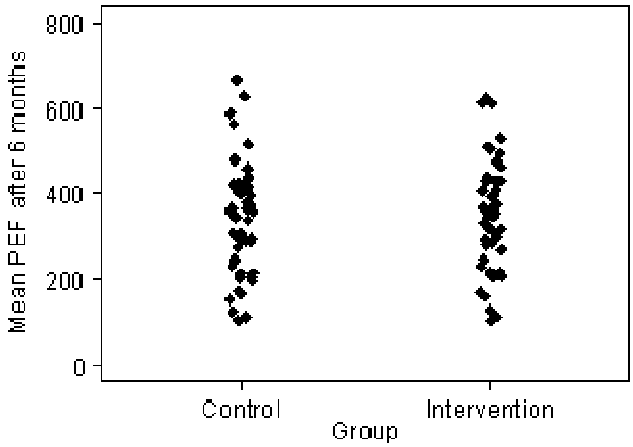

Figure 14. Mean of one-week diary peak expiratory flow six months after training by an asthma specialist nurse or usual care (data of Levy *et al.*, 2000)

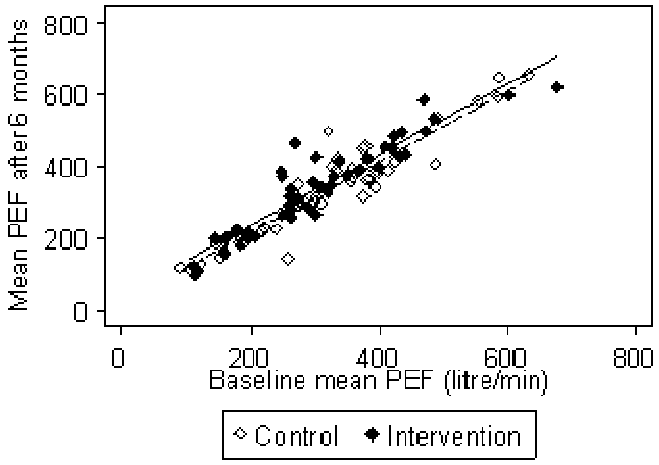

Figure 15. Mean PEF after 6 months against baseline PEF for intervention and control asthmatic patients, with fitted analysis of covariance lines (data of Levy *et al.*, 2000)

In clinical trials, regression is often used to adjust for prognostic variables and baseline measurements. For example, Levy *et al.* (2000) carried out a trial of education by a specialist asthma nurse for patients who had been taken to an accident and emergency department due to acute asthma. Patients were randomised to have two one-hour training sessions with the nurse or to usual care. The measurements were one week peak expiratory flow and symptom diaries made before treatment and after three and six months. We summarised the 21 PEF measurement (three daily) to give the outcome variables mean and standard deviation of PEF over the week. We also analysed mean symptom score. The primary outcome variable was mean PEF, shown in Figure 14. There is no obvious difference between the two groups and the mean PEF was 342 litre/min in the nurse intervention group and 338 litre/min in the control group. The 95% CI for the difference, intervention – control, was –48 to 63 litre/min, P=0.8, by the two-sample t method.

However, although this was the primary outcome variable, it was not the primary analysis. We have the mean diary PEF measured at baseline, before the intervention, and the two mean PEFS are strongly related. We can use this to reduce the variability by carrying out multiple regression with PEF at six months as the outcome variable and treatment group and baseline PEF as predictors. If we control for the baseline PEF in this way, we might get a better estimate of the treatment effect because we will remove a lot of variation between people.

We get:

PEF@6m =	18.3	+	0.99 × PEF@base	+	20.1 × intervention
95% CI	–10.5 to 47.2		0.91 to 1.06		0.4 to 39.7
			P<0.001		P=0.046

Figure 15 shows the regression equation (or analysis of covariance, as the term is often used in this context) as two parallel lines, one for each treatment group. The vertical distance between the lines is the coefficient for the intervention, 20.1 litre/min. By including the baseline PEF we have reduced the variability and enabled the treatment difference to become apparent.

There are clear advantages to using adjustment. In clinical trials, multiple regression including baseline measurements reduces the variability between subjects and so increase the power of the study. It makes it much easier to detect real effects and produces narrower confidence intervals. It also removes any effects of chance imbalances in the predicting variables.

Is adjustment cheating? If we cannot demonstrate an effect without adjustment (as in the asthma nurse trial) is it valid to show one after adjustment? Adjustment can be cheating if we keep adjusting by more and more variables until we have a significant difference. This is not the right way to proceed. We should be able to say in advance which variables we might want to adjust for because they are strong predictors of our outcome variable. Baseline measurements almost always come into this category, as should any stratification or minimisation variables used in the design. If they were not related to the outcome variable, there would be no need to stratify for them. Another variable which we might expect to adjust for is centre in multi-centre trials, because there may be quite a lot of variation between centres in their patient populations and in their clinical practices. We might also want to adjust for known important predictors. If we had no baseline measurements of PEF, we would want to

adjust for height and age, two known good predictors of PEF. We should state before we collect the data what we wish to adjust for and stick to it.

In the PEF analysis, we could have used the differences between the baseline and six month measurements rather than analysis of covariance. This is not as good because there is often measurement error in both our baseline and our outcome measurements. When we calculate the difference between them, we get two lots of error. If we do regression, we only have the error in the outcome variable. If the baseline variable has a lot of measurement error or there is only a small correlation between the baseline and outcome variables, using the difference can actually make things worse than just using the outcome variable. Using analysis of covariance, if the correlation is small the baseline variable has little effect rather than being detrimental.

Transformations in multiple regression

In the asthma nurse study, a secondary outcome measure was the standard deviation of the diary PEFs. This is because large fluctuations in PEF are a bad thing and we would like to produce less variation, both over the day and from day to day. Figure 16 shows SD at six months against SD at baseline by treatment group. Figure 17 shows the distribution of the residuals after regression of SD at six months on baseline SD and treatment. Clearly the residuals have a skew distribution and the standard deviation of the outcome variable increases as the baseline SD increases. We could try a log transformation. This gives us a much more uniform variability on the scatter diagram (Figure 18) and the distribution of the residuals looks a bit closer to the Normal. The multiple regression equation is

logSD@6m	= 2.78	+ 0.017 × SD@base		– 0.42 × intervene
95% CI	2.48 to 3.08	0.010 to 0.024		–0.65 to –0.20
		P<0.001		P<0.001

We estimate that the mean log SD is reduced by –0.42 by the intervention, whatever the baseline SD. Because we have used a log transformation, we can back transform just as we did for the difference between two means (Week 5). The antilog is exp(–0.42) = 0.66. We interpret this as that the mean standard deviation of diary PEF is reduced by a factor of 0.66 by the intervention by the specialist asthma nurse. We can antilog the confidence interval, too, giving 0.52 to 0.82 as the confidence interval for the ratio of nurse SD to control SD.

Factors with more than two levels

We can use any categorical variable as a predictor. We do not have to restrict ourselves to those with only one level, such as intervention or control, but can also use categorical variables with more than two categories. For example, Table 3 shows some data from a study of six patients with prostate cancer being treated to reduce the size of their tumours. Can we estimate the relationship between tumour size and portal vein transit time? Figure 19 shows a scatter diagram. We might be tempted to calculate a regression line of transit time on tumour size, or correlation coefficient between them, but this could be highly misleading. The observations are not independent, because the measurements on the same person will be more like one another than they are like those on another person. Also, we are interested in whether reduced tumour size is associated with reduced blood flow, not whether people with larger tumour have greater blood flow. We would like to look at the relationship between tumour size and blood flow within the same subject.

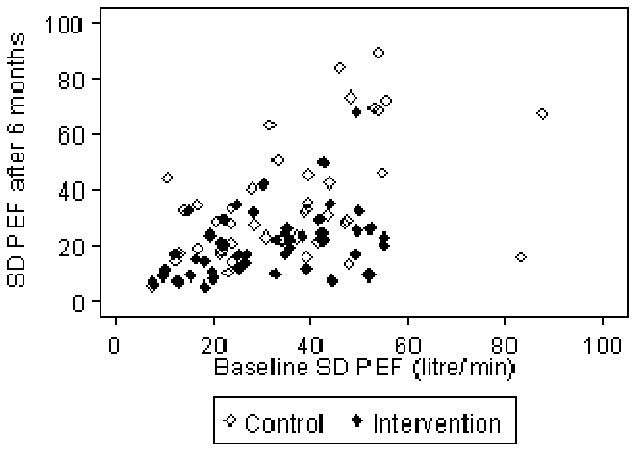

Figure 16. Standard deviation of diary PEF after six months, by baseline standard deviation and treatment group

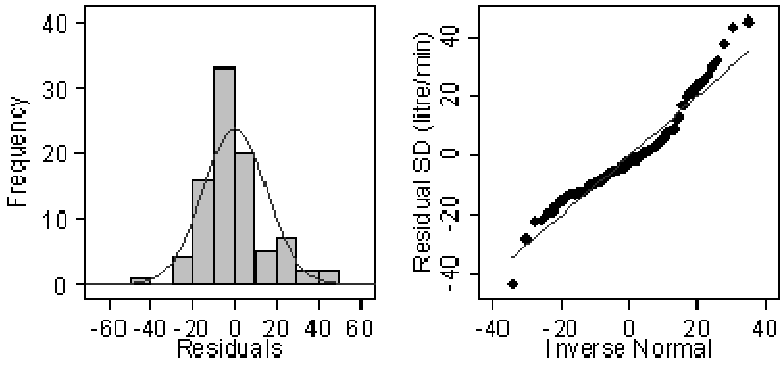

Figure 17. Residual standard deviation of diary PEF after six months after regression on baseline SD and treatment group

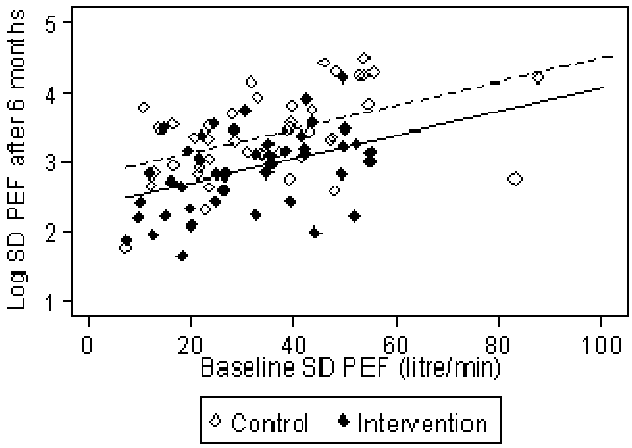

Figure 18. Log transformed standard deviation of diary PEF after six months, by baseline standard deviation and treatment group

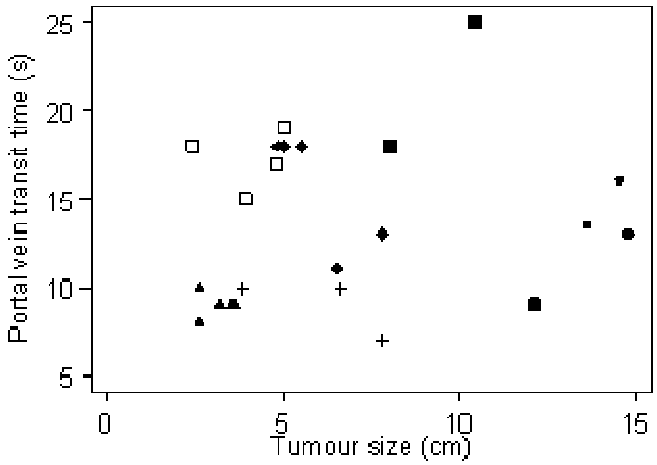

Figure 19. Serial measurements of tumour size by CT scan and portal vein blood flow (transit time in sec) for six patients, each symbol representing a different patient

Table 3. Serial measurements of tumour size by CT scan and portal vein blood flow (transit time in sec) for six patients (data of Oliver Byass)

Subject 1			Subject 2			Subject 3		
Time	CT	PV	Time	CT	PV	Time	CT	PV
0	10	*	1	12.1	9	1	14.7	13
1	8	*	2	10.4	25	2	14.8	13
2	7.8	13	3	9.4	*	3	14.5	*
3	6.5	11	4	7.2	*	4	14.5	16
4	5.5	18	5	8	18	5	14.1	*
5	4.8	18	6	8.2	*	6	13.6	13.5
6	5	18						

Subject 4			Subject 5			Subject 6		
Time	CT	PV	Time	CT	PV	Time	CT	PV
1	5	19	1	3.6	9	1	7.8	7
2	3.9	15	2	2.6	10	2	6.6	10
3	4.8	17	3	2.6	8	3	5.5	*
4	2.4	18	4	3.2	9	4	4.5	*
			5	3.5	9	5	3.8	10

CT = tumour size (cm) by CT scan, PV = portal vein transit time
* = missing data

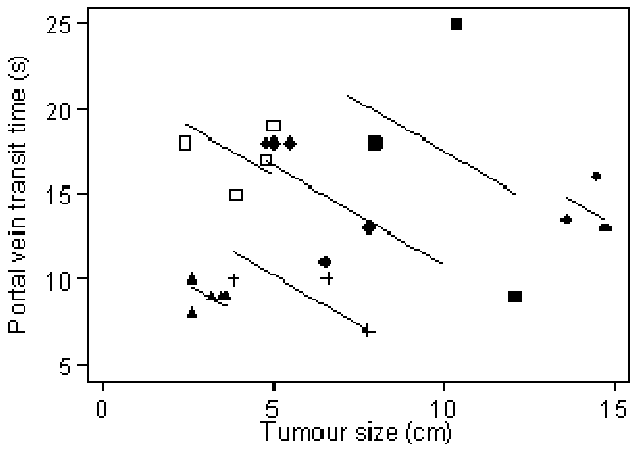

Figure 20. Serial measurements of tumour size by CT scan and portal vein blood flow (transit time in sec) for six patients, showing the analysis of covariance model

Table 4. Results of a factorial clinical trial of antidepressant drug counselling and information leaflets to improve adherence to drug treatment: patients reporting continuing treatment at 12 weeks (Peveler *et al.*, 1999)

Leaflet	Drug counselling		Total
	Yes	No	
Yes	32/53 (60%)	22/53 (42%)	54/105 (51%)
No	34/52 (65%)	20/55 (36%)	54/108 (50%)
Total	66/105 (63%)	42/108 (39%)	

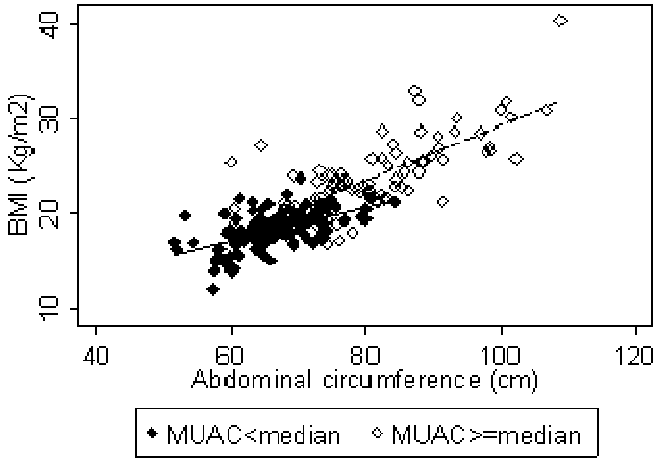

Figure 21. Interaction between abdominal and arm circumference in their effects on BMI

We can do this by multiple regression or analysis of covariance (which are the same thing). Essentially, we fit a model which has parallel lines relating transit time and tumour size for each subject separately. To do this, we need to fit subject as a predictor. We cannot just put the variable subject number into a regression equation as if it were an interval variable, because there is no sense in which Subject 2 is greater than Subject 1. We do not want to assume that our categories are related in this way. Instead, we define **dummy variables** or **indicator variables** which enable us to estimate a different mean for each category. We do this by defining a set of variables, each of which is 0 or 1. For the example, we need five dummy variables:

sub1 = 1 if Subject 1, 0 otherwise,
sub2 = 1 if Subject 2, 0 otherwise,
sub3 = 1 if Subject 3, 0 otherwise,
sub4 = 1 if Subject 4, 0 otherwise,
sub5 = 1 if Subject 5, 0 otherwise.

If all of these variables are zero, then we have Subject 6. We need five dummy variables to represent a categorical variable with six categories. Subject 6 is called the **reference category**.

We then do regression on our continuous predictor variable and all the dummy variables:

$$PV = 22.5 - 1.17\times CT + 6.7\times sub1 + 8.2\times sub2 - 0.6\times sub3 - 9.9\times sub4 - 6.4\times sub5$$
$$\qquad\qquad P=0.05 \quad P=0.06 \quad P=0.1 \quad P=0.8 \quad P=0.001 \quad P=0.01$$

We should ignore the individual tests for the coefficients, because they do not mean much. What we want to know is whether there is any evidence that subject as a whole has an effect. (It would be very surprising if it didn't, because they are six different people.) To do this, we get a combined F test for the factor, which is beyond the scope of this course. Here the test statistic is F = 6.83 with 1 and 5 degrees of freedom, P = 0.001. The fitted lines are shown in Figure 20.

Most statistical computer programs will calculate the dummy variables for you. You need to specify in some way that the variable is categorical, using terms such as 'factor' or 'class variable' for a categorical variable and 'covariate' or continuous' for quantitative predictors.

Dichotomous outcome variables and logistic regression

There are other forms of regression which enable us to do similar things for other kinds of variables. Logistic regression allows us to predict the proportion of subjects who will have some characteristic, such as a successful outcome on a treatment, when the outcome variable is a yes or no, dichotomous variable.

For our first example, Table 4 shows the results of a clinical trial of two interventions which it was hoped would improve adherence to antidepressant drug treatment in patients with depression (Peveler *et al.*, 1999). Two different interventions, antidepressant drug counselling and an information leaflet, were tested in the same trial. The trial used a factorial design where subjects were allocated to one of four treatment combinations:

1. counselling and leaflet
2. counselling only
3. leaflet only
4. neither intervention

The outcome variable was whether patients continued treatment up to 12 weeks. The authors reported that

> '66 (63%) patients continued with drugs to 12 weeks in the counselled group compared with 42 (39%) of those who did not receiving counselling (odds ratio 2.7, 95% confidence interval 1.6 to 4.8; number needed to treat=4). Treatment leaflets had no significant effect on adherence.' (Peveler *et al.*, 1999)

How did they come to these conclusions? We might think that we would take the total row of the table and use the method of estimating the odds ratio for a two by two table described in Week 6. The problem with this is that if both variables have an effect then each will affect the estimate for the other. We use logistic regression instead.

Our outcome variable is dichotomous, continue treatment yes or no. We want to predict the proportion who continue treatment from whether they were allocated to the two interventions, counselling and leaflet. We would like a regression equation of the form:

$$\text{proportion} = \text{intercept} + \text{slope} \times \text{counselling} + \text{slope} \times \text{leaflet}$$

The problem is that proportions cannot be less than zero or greater than one. How can we stop our equation predicting impossible proportions? To do this, we find a scale for the outcome which is not constrained. Odds has no upper limit, so it can be greater than one, but it must be greater than or equal to zero. Log odds can take any value. We therefore use the log odds of continuing treatment, rather than the proportion continuing treatment. We call the log odds the **logit** or **logistic transformation** and the method used to fit the equation

$$\text{log odds} = \text{intercept} + \text{slope}_1 \times \text{counselling} + \text{slope}_2 \times \text{leaflet}$$

is called **logistic regression**. The slope for counselling will be the increase in the log odds of continuing treatment when counselling is used compared to when counselling is not used. It will be the log of the odds ratio for counselling, with both the estimate and its standard error adjusted for the presence or absence of the leaflet. If we antilog, we get the adjusted odds ratio.

The fitted logistic regression equation for the data of Table 4, predicting the log odds of continuing treatment, is:

$$\text{log odds} = -0.559 + 0.980 \times \text{counselling} + 0.216 \times \text{leaflet}$$

This is calculated by a computer-intensive technique called maximum likelihood. This finds the values for the coefficients which would make the data observed the most likely outcome. We can find 95% confidence intervals for the coefficients and P values testing the null hypothesis that the coefficient is zero. These are seldom of interest for the intercept. For counselling, the 95% CI is 0.426 to 1.53, P=0.001. For the leaflet, we have 95% CI = –0.339 to 0.770, P=0.4.

If we antilog this equation, we get an equation predicting the odds:

$$\text{odds} = 0.57 \times 2.66^{\text{counselling}} \times 1.24^{\text{leaflet}}$$

because when we antilog, things which are added become multiplied and two numbers which are multiplied become one number raised to the power of the other (see separate Note on Logarithms). This is actually quite easy to interpret, although it doesn't look it. The variable for counselling is zero if the subject did not receive counselling, or one if the subject received counselling. Any number raised to the power zero is equal to one and so $2.66^0 = 1$. Any number raised to the power one is just the number itself and so $2.66^1 = 2.66$. Hence if the subject has counselling, the odds of continuing treatment is multiplied by 2.66, so 2.66 is the odds ratio for counselling. Similarly, the odds ratio for continuing treatment if given the leaflet is 1.24. The 95% confidence intervals for these odds ratios are 1.53 to 4.64 and 0.71 to 2.16 respectively.

The odds ratio for counselling is described as being adjusted for the presence or absence of the leaflet, and the odds ratio for the leaflet is described as being adjusted for counselling.

Interactions

The estimates produced in the previous section were made using all the observations. They were made assuming that the odds ratio for counselling was unaffected by the presence or absence of the leaflet and that the odds ratio for the leaflet was unaffected by the presence or absence of counselling. We can ask whether the presence of the leaflet changes the effect of counselling by testing for an interaction between them.

To do this we define an interaction variable. We can define this to be equal to one if we have both counselling and leaflet, zero otherwise. The counselling and leaflet variables are both 0 or 1. If we multiply the counselling and leaflet variables together, we get the interaction variable:

Interaction = counselling × leaflet.

The interaction variable is zero if either counselling or leaflet is zero, so is one only when both are one. We can add interaction to the logistic regression equation:

$\log \text{odds} = \text{intercept} + \text{slope}_1 \times \text{counselling} + \text{slope}_2 \times \text{leaflet} + \text{slope}_3 \times \text{interaction}$

If we fit coefficients to the data in Table 4, we get:

log odds = –0.560 + 0.981×counselling + 0.217×leaflet – 0.002×interaction			
95% CI	0.203 to 1.78	–0.558 to 0.991	–1.111 to 1.107
	P=0.01	P=0.6	P=1.0

Compare the model without the interaction:

log odds = –0.559 + 0.980 × counselling + 0.216 × leaflet		
95% CI	0.426 to 1.53	–0.339 to 0.770
	P=0.001	P=0.4

The estimates of the treatment effects are unchanged by adding this non-significant interaction but the confidence intervals are wider and P values bigger. We do not need the interaction in this trial and should omit it.

If we did decide to keep the interaction, the estimate of the effect of counselling would be modified by the presence or absence of the leaflet. The interaction variable is equal to counselling multiplied by leaflet. We could write the equation as

log odds = –0.560 + 0.981×counselling + 0.217×leaflet – 0.002×counselling×leaflet

The total effect of counselling is then 0.981 – 0.002×leaflet, i.e. it 0.981 if there is no leaflet and 0.981 – 0.002 = 0.979 if there is a leaflet.

We can do the same thing for continuous outcome variables. Above, the relationship between BMI and abdominal circumference was assumed to be the same for males and for females. This may not be the case. Men and women are different shapes and the slope of the line describing the relationship of BMI to abdominal circumference may differ between the sexes. If this is the case, we say that there is an interaction between abdominal circumference and sex. We can investigate this using multiple regression.

We want our equation to be able to estimate different slopes for males and females. We create a new variable by multiplying the abdominal circumference by the variable 'male', which = 1 for a male and = 0 for a female. We can add this to the multiple regression on abdominal circumference, arm circumference, and sex:

BMI = –6.44 + 0.18×abdomen + 0.64×arm – 1.39×male
P<0.001 P<0.001 P<0.001

Adding the interaction term:

BMI = –7.95 + 0.21×abdomen + 0.63×arm + 1.63×male – 0.04×male×abdomen
P<0.001 P<0.001 P=0.4 P=0.1

The coefficients for both abdomen and male are changed by this and male becomes not significant. The interaction term is not significant, either. However, we will consider what the coefficients mean before going on to complete our analysis of these data. For a female subject, the variable male = 0 and so male×abdomen = 0. The coefficient for abdominal circumference is therefore 0.21 Kg/m^2 per cm. For a male subject, the variable male = 1 and so male×abdomen = abdomen. The coefficient for abdominal circumference is therefore 0.21 – 0.04 = 0.17 Kg/m^2 per cm. (When we did not include the interaction term, the coefficient was between these two values, 0.18 Kg/m^2 per cm.) Looking at this another way, the coefficient for abdominal circumference can be rewritten as 0.21 – 0.04×male. If the interaction is not significant, we usually drop it from the model, as it makes things more complicated without adding to our predictive power.

We can add other interactions to our model, between sex and arm circumference and between abdominal and arm circumference. In each case, we do this my multiplying the two variables together. The only one which is statistically significant is that between abdominal and arm circumference:

BMI = 8.45 – 0.02×abdomen + 0.03×arm – 1.22×male + 0.0081× abdomen×arm
P<0.8 P<0.9 P<0.001 P=0.01

If the interaction is significant, both the main variables, abdominal and arm circumference, must have a significant effect on BMI, so we ignore the other P values. The coefficient of abdominal circumference now depends on arm circumference, so it becomes –0.02 + 0.0081×arm circumference. The interaction is illustrated in Figure 21, where we show the regression of BMI on abdominal circumference separately for subjects with mid upper arm circumference below and above the median. The slope is steeper for subjects with larger arms.

Sample size

Multiple regression methods may not work well with small samples. We should always have more observations than variables, otherwise they won't work at all. However, they may be very unstable if we try to fit several predictors using small samples. The following rules of thumb are based on simulation studies. For multiple regression, we should have at least 10 observations per variable. For logistic regression, we should have at least 10 observations with a 'yes' outcome and 10 observations with a 'no' outcome per variable. Otherwise, things may get very unstable.

Types of regression

Multiple regression and logistic regression are the types of regression most often seen in the medical literature. There are many other types for different kinds of outcome variable. Those which you may come across include:

- Cox regression for survival analysis, Week 8,
- ordered logistic regression for outcome variables which are qualitative with ordered categories,
- multinomial regression for outcome variables which are qualitative with unordered categories,
- Poisson regression for outcome variables which are counts,
- negative binomial regression for outcome variables which are counts with extra sources of variability,

Pitfalls in multiple regression

The outcome variable in multiple linear regression should be a continuous, interval scale measurement. Discrete quantitative variables may be used when there are a lot of possible values, as such variables may be treated as continuous. We should use multiple regression when the outcome variable is dichotomous, categorical, or discrete with only a few possible values. There are special methods of regression for such data, such as logistic regression for dichotomous variables, ordered logistic regression for ordered categories, multinomial regression for unordered categories, and Poisson regression and negative binomial regression for counts.

We should not use regression for survival times where not all the times are known because the event has yet to happen. We can use Cox regression for data of this type, described in Week 8.

For continuous data, we should always check the assumptions of uniform variance and Normal distribution for the residuals. Uniform variance is the more important assumption. If these are not met, the fitted relationship may not be the best and significance tests and confidence intervals may not be valid.

We should beware of using regression where there is an inadequate number of observations. We should always have more observations than variables. If we have the same number of variables as observations, we can predict the values of the outcome variable in the sample exactly, whatever the model. The regression tells us nothing useful about the population as a whole. We also need sufficient observations to estimate the variability about the regression. A useful rules of thumb for multiple regression is that we should have at least ten observations per predictor variable. It is so easy to do logistic regression with lots of predictor variables that we are often tempted to use too many. Logistic regression is a large sample method and we should make sure that we have enough data for fitting the coefficients. Programs do not usually warn you when you do not. They will produce estimates, confidence intervals and P values without any warning that they are unreliable.

We should not use categorical predictor variables as if they were quantitative, we should use dummy variables instead. For example, if we coded our six patients in Table 3 as patient = 1, 2, 3, 4, 5, and 6, we should not fit a regression model with patient as a variable, unless we use a program which calculates dummy variables for us.

We must always beware of lack of independence among the observations, such as multiple observations on same subject. We can allow for this using methods which take this data structure into account, such as robust standard errors.

References

Altman DG. (1991) *Practical Statistics for Medical Research*. Chapman and Hall, London.

Levy ML, Robb M, Allen J, Doherty C, Bland JM, Winter RJD. (2000) A randomized controlled evaluation of specialist nurse education following accident and emergency department attendance for acute asthma. *Respiratory Medicine* 94, 900-908.

Peveler R, George C, Kinmonth A-L, Campbell M, Thompson C. Effect of antidepressant drug counselling and information leaflets on adherence to drug treatment in primary care: randomised controlled trial. *BMJ* 1999; **319**: 612-615.

Prentice AM, Black AE, Coward WA, Davies HL, Goldberg GR, Murgatroyd PR, Ashford J, Sawyer M, Whitehead RG. (1986) High-levels of energy-expenditure in obese women. *British Medical Journal* **292**, 983-987.