IAPT: 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.

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**.

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		

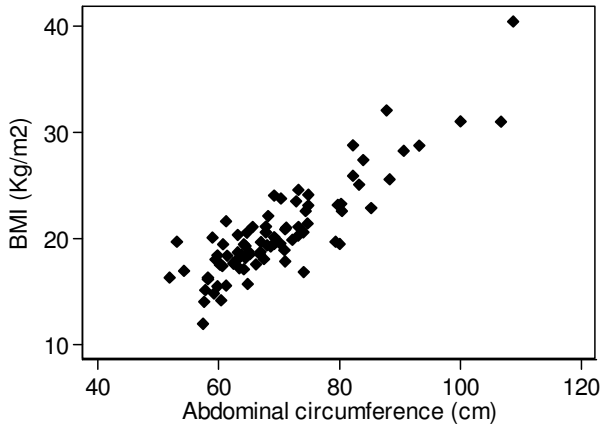

Figure 1. Scatter plot of BMI against abdominal circumference

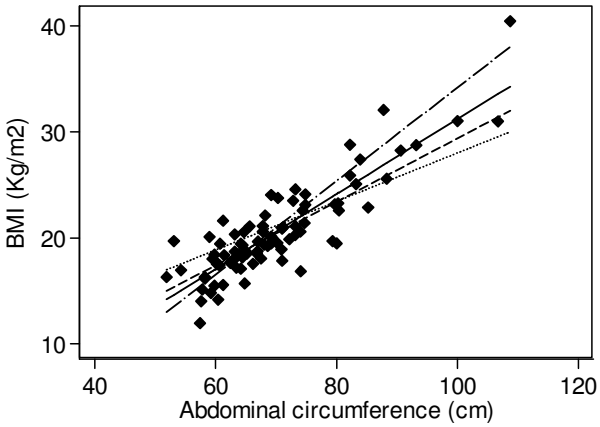

Figure 2. Scatter plot of BMI against abdominal circumference with possible lines to represent the relationship

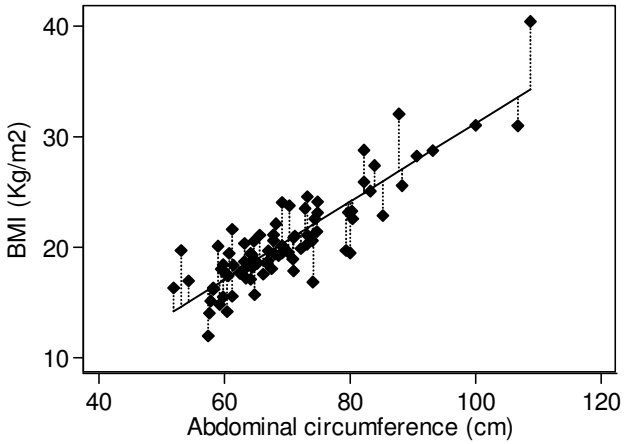

Figure 3. Differences between the observed and predicted values of the outcome variable

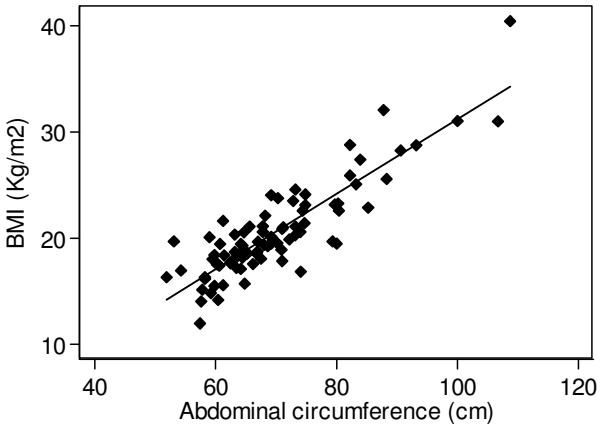

Figure 4. The least squares regression line for BMI and abdominal circumference

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.

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.

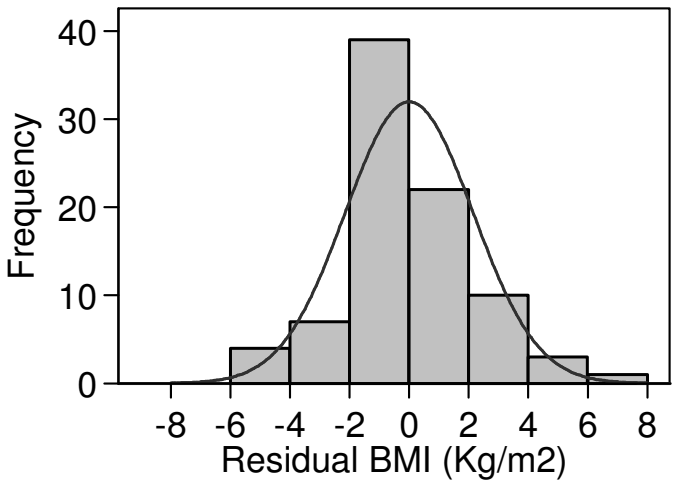

Figure 5. Histogram and Normal plot for residuals for the BMI and abdominal circumference data

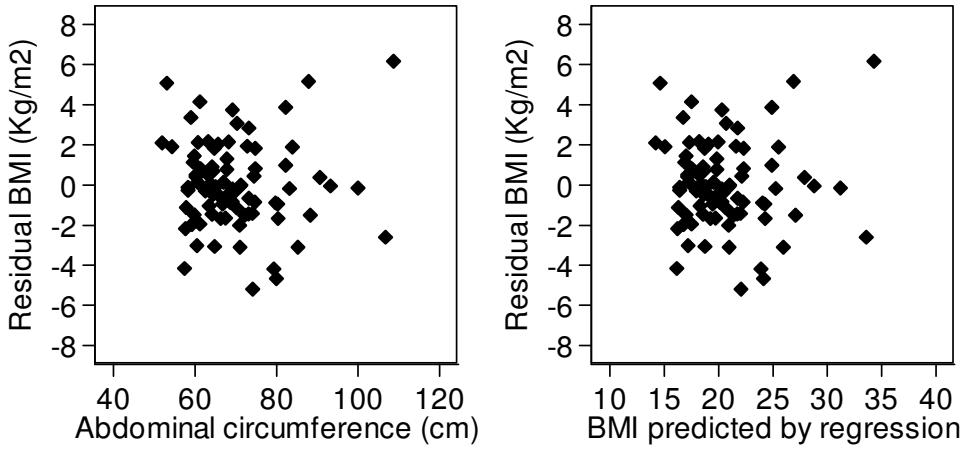

Figure 6. Scatter plot of residual BMI against abdominal circumference and against the regression estimate

We can check both of these assumptions graphically. Figure 5 shows a histogram 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.

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 7 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.

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 8 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.

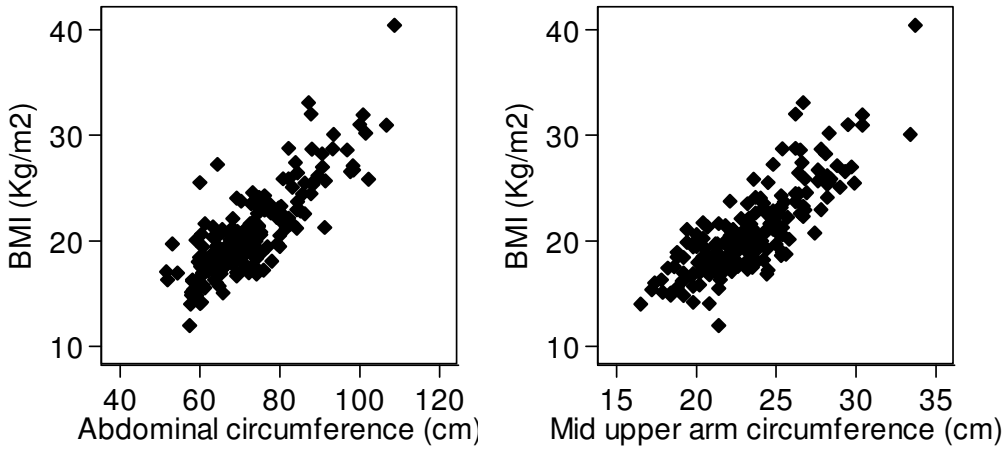

Figure 7. BMI against abdominal circumference and arm circumference in 202 adults

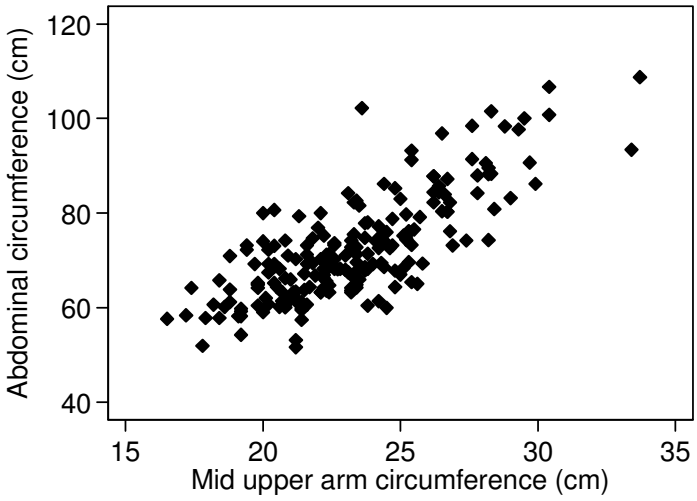

Figure 8. Abdominal circumference against mid upper arm circumference in 202 adults

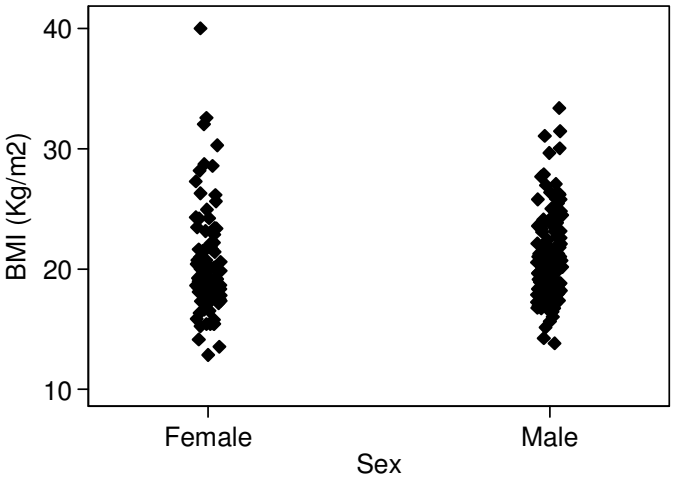

Figure 9. BMI for women and men

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 9 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.

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.

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 10. 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

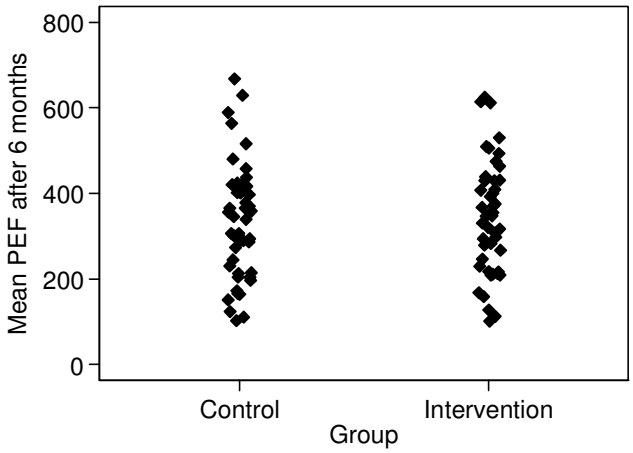

Figure 10. Mean of one-week diary peak expiratory flow six months after training by an asthma specialist nurse or usual care (data of Levy *et al.*, 2000)

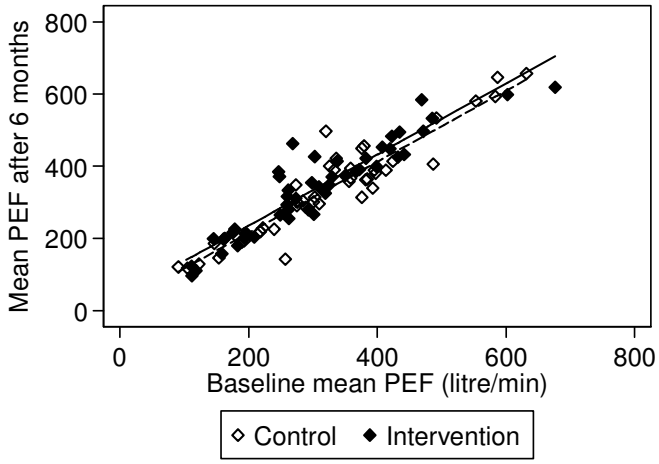

Figure 11. Mean PEF after 6 months against baseline PEF for intervention and control asthmatic patients, with fitted analysis of covariance lines (data of Levy *et al.*, 2000)

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 11 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.

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:

- logistic regression for dichotomous outcome variables,
- Cox regression for survival analysis,
- 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,

Reference

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.