Archives of Disease in Childhood 1995; 73: 177–181

STATISTICS FROM THE INSIDE

15. Multiple regression (1)

M J R Healy

Regression equations with two or more x's

In a previous article in this series I described some aspects of the simple regression model, where the mean of a variate y is related to a quantity x in a linear (straight line) fashion. With some formality we can write

$$E(y \mid x)=\alpha+\beta x$$

In this equation α is the value of y where the line crosses the y axis, and is often called the *intercept*; β is the *slope* of the line, the amount of increase in y per unit increase in x. $E(y \mid x)$ is mathematicians' notation for a mean value (misleadingly called an *expectation*) and the vertical bar shows that the mean is that of y for a particular value of x, a *conditional mean* (compare the conditional probabilities that I wrote about in an earlier note). The usual name for y is the *dependent variate*; x goes by various names, notably the *predictor* or *covariate* or (misleadingly, as we shall see) the *independent variate*. Obvious examples are where y might be the response to a drug and x the dose; or y the head circumference of a baby and x the baby's weight. Notice the important assumption of *linearity*; this means that a given change in x corresponds to a fixed change in y, no matter where it starts from.

One use of the regression equation is to *predict* the value of y that might correspond to an observed value of x on a future occasion. The prediction will not of course be perfect, and the observed value of y will differ from that which is predicted by the equation. The difference is usually called a *residual*. The sizes of the residuals can be summarised by quoting their standard deviation (their mean is exactly zero), and this is called the *residual standard deviation* or *residual standard error*. Roughly speaking, around 95% of the residuals can be expected to fall short of twice the residual standard deviation.

It is a natural extension of this idea to use two or more covariates simultaneously to predict the value of y. This leads to a *multiple regression* equation. Starting simply, consider the miniature example in table 1 which shows measurements of height, weight, and chest circumference of 10 army cadets.

The mean chest circumference is 102·6 cm with a standard deviation of 6·78 cm and this suggests that most future measurements of chest circumference from the same population might fall in the range mean ±2 SD, 89 to 116 cm, a width of 27 cm. The standard deviation thus measures our degree of uncertainty concerning the chest circumference of a random individual from this population. We could try to reduce this by predicting chest circumference from height or from weight by doing simple regressions. The standard calculations show that the regression coefficient on height is −0·43 (SE 0·48) giving a residual standard deviation of 6·86 cm; that on weight is +0·92 (SE 0·17) with residual standard deviation of 3·32 cm. Comparing the coefficients with their standard errors, it appears that height is useless as a predictor in this small sample. Weight on the other hand may be quite successful, with a highly significant regression coefficient. These conclusions are confirmed by the reduction (or lack of it) in the residual standard deviation – compared with the previous value of 27 cm, the ±2 SD interval measures 27·4 cm when the subject's height is allowed for, 13·3 cm when weight is allowed for.

What about using them both simultaneously? We need to estimate a relationship of the form

$$E(y \mid x_1,x_2)=\alpha+\beta_1 x_1+\beta_2 x_2$$

where y stands for chest circumference and x_1,x_2 for height and weight respectively. With only two covariates this is not an impossible task for a pocket calculator but multiple regression calculations can become quite heavy and a good computer package is desirable. Most packages provide the same items of information in different guises; I have chosen the Nanostat package (Alphabridge Ltd, 26 Downing Court, London WC1N 1LX) with which I am personally familiar. The computer output is shown in table A.

This is rather a formidable amount of information for a fairly simple problem and it is important not to be intimidated by it. It is most easily read from the bottom up. You will see that the estimated equation can be written as

Chest circumference
=132·66−0·54746×height+0·95709×weight

Table 1 Measurements on army cadets

Height (cm)	*Weight (kg)*	*Chest circumference (cm)*
167·9	71·8	107·3
183·8	75·1	105·2
172·9	58·0	93·4
175·5	58·4	91·9
176·4	67·7	99·8
168·5	75·2	113·4
178·0	71·3	103·7
178·0	67·3	98·1
175·4	75·9	108·4
171·2	65·3	105·2

23 Coleridge Court, Milton Road, Harpenden, Herts AL5 5LD

Correspondence to:
Professor Healy.
No reprints available.

Dependent variable CHEST

	df	SS	MS	F	p
Regression	2	388.09	194.05	51.82	0.0001
Residual	7	26.214	3.7448		
Total	9	414.30	46.034		

Variance accounted for = 91.9%

Residual s.d. taken to be 1.9352

Regression coefficients

	b	se	t
HEIGHT	-0.54746	0.13455	-4.069
WEIGHT	0.95709	0.098896	9.678
Constant term	132.66	23.851	5.562

Table A Regression of chest circumference on height and weight

where the coefficients are taken from the column marked b*. Looking at the standard errors of the coefficients and the corresponding t values, it is apparent that both of them are highly significant, the coefficient of height being negative.

This last illustrates a very important fact about the coefficients in a multiple regression equation (they are called *partial regression coefficients*). The negative height coefficient might seem to say that taller men tend to have smaller chest circumferences, contrary to intuition. This is not at all the correct inference. Each coefficient in a multiple regression measures the effect of its x variable when the other x's in the equation have been allowed for ('partialled out' is the phrase that is sometimes used). The height coefficient measures the effect of height on chest circumference among men all of a given weight; it is very plausible that there will be some short fat ones and some tall thin ones, giving rise to the negative relationship in the equation.

We need a measure of the closeness of the regression relationship, and this will be provided by the amount of scatter of the observed points away from the regression. This in turn is measured by the *residuals*, the departures of the observed y values from those predicted by the regression equation. As we have seen, the scatter of the residuals can be summarised by their standard deviation, the *residual standard deviation* which you will see given in the computer output in table A. The actual residuals in our example are given in table 2 and the residual variance can be obtained by summing their squares and dividing by the degrees of freedom. We have estimated a constant term or *intercept* and two slope coefficients and these use up 3 degrees of freedom, leaving 7 for the residual SD. We can see that the residual SD of 1·9352 cm is much smaller than the original SD of 6·78 cm and indeed less than that of 3·32 cm achieved by predicting from weight alone.

The residual SD can also be obtained, along with other information, from the *analysis of variance* table which is at the top of the computer output (I introduced the analysis of variance in an earlier article in this series but did not have the nerve actually to show one). A more precise name for this would be an analysis of sums of squares. You can check that the Total sum of squares in the column headed SS is just the usual sum of squares of the 10 chest circumferences about their mean with its $(10-1)=9$ degrees of freedom. This is usually calculated by hand by first obtaining the crude sum of squares (simply the sum of the squared observations) and then subtracting the 'correction for the mean', a term depending only upon the mean value and the sample size. In a very similar way the sum of squares of the residuals can be got by taking the sum of squares about the mean and subtracting a term depending upon the regression coefficients. This latter term is the one labelled 'Regression' in the analysis of variance table. Since two unknown coefficients have been estimated, it has 2 degrees of freedom. Both the residual degrees of freedom and the residual sum of squares can be calculated by making the corresponding columns of the table add up. Then dividing the residual sum of squares by its degrees of freedom we get the residual mean square, which is another name for the variance of the residuals. The square root of this is the residual standard deviation. You may like to check that the sum of the squares of the residuals in table 2 is equal (rounding errors apart) to that given by the indirect method of calculation represented by the analysis of variance table.

We can also get a mean square from the Regression line in the table. Comparing this with the residual mean square provides an F test of the hypothesis that *both* of the true coefficients are zero. In this example, the F value is very highly significant as would be expected.

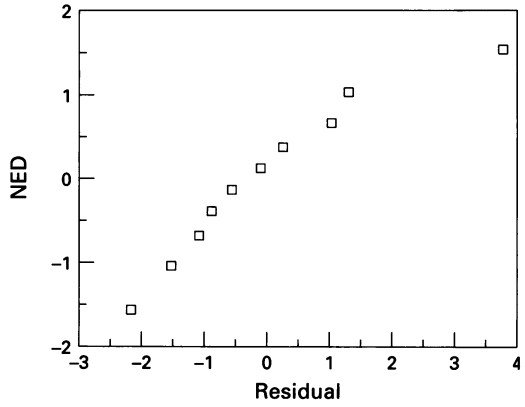
Table 2 Chest circumferences

Observed	*Predicted*	*Residuals*
107·3	109·46	−2·16
105·2	103·91	+1·29
93·4	93·51	−0·11
91·9	92·47	−0·57
99·8	100·88	−1·08
113·4	112·38	+1·02
103·7	103·45	+0·25
98·1	99·62	−1·52
108·4	109·28	−0·88
105·2	101·43	+3·77

Goodness of fit

How well have we succeeded in fitting the variation in our data? There are various ways of answering this question. We have already seen that the original standard deviation of 6·78 cm has been reduced to 1·94 cm. Squaring these figures, the original variance was 46·0 and this has been reduced to 3·7. The reduction in variance is 42·3 which is 92% of the starting value – we can say that 92% of the original variance has been 'accounted for' by the regression relationship, a figure which appears in the computer output in table A. I note that most computer packages at this point provide a quantity called the *multiple correlation coefficient*, denoted by R^2. This is actually the ratio

*For clarity I have repeated the coefficients from the computer output with their five significant figures. In real life, these should be rounded off to three significant figures or so.



Normal plot of chest circumference residuals (NED=Normal Equivalent Deviate).

of the Regression sum of squares to the Total sum of squares. It has the drawback that adding a further predictor to the equation inevitably increases R^2, making the fit look better when this may be unjustified. As an example, the simple regression of chest circumference on height gives an R^2 of 0·09 (R=0·30 – sounds quite good), whereas the residual variance is actually larger than that with no covariate – the variance accounted for is – 2·4%! Elaborate correction formulas have been derived to cope with this, but their use seems to me to be unnecessary. The sums of squares are really no more than convenient steps in the calculations – the mean squares are the interpretable quantities, and the index of goodness of fit is best based upon them. The multiple correlation coefficient is a leftover from the early days of statistics, when correlation and coefficients for measuring it were all the rage, and it is nowadays best avoided.

The use either of R^2 or the percentage of variance accounted for is subject to two important qualifications. First, it must not be assumed that a very high value, as in our example, implies a fit that is close enough for a given practical application. Knowing height and weight, we can predict a new subject's chest circumference with an uncertainty of around 4 cm either way. This is not too bad for tailoring purposes but may be insufficient in more demanding circumstances. The actual size of the residual standard deviation is far more important than its relation to some other quantity. Moreover, both measures tacitly assume that the denominator, the original variance or sum of squares, is a meaningful quantity, and this in turn assumes that both the y's and the x's constitute a random sample from a suitable population. We shall see later that this is by no means always the case.

A more searching examination of the goodness of fit of the regression involves inspection of the individual residuals, which we have seen in table 2 (any statistical package worthy of the name will calculate these for you). This is best done graphically. One question is whether one or two of the points lie unduly far away from their predicted values – the residual in the last line of table 2 looks a little large, for example. A Normal plot of the residuals is shown in the figure, and this confirms that the 10th observation may be a little extreme – did this sturdy fellow really weigh only 65 kg? It will often be worth plotting the residuals against each of the x's to see whether the assumption of a linear relationship is a plausible one.

The analysis of variance

The analysis of variance table has not played a major part in the example above, but it is worth a little further study if only for future reference. To begin with, we can use it to track the results of fitting one covariate and then another. If we use the Nanostat program to fit weight as a single covariate, the resulting analysis of variance table is shown in table B.

The Total sum of squares is of course the same as before and the regression has accounted for an amount 326·10 with 1 degree of freedom. If you go back to the original analysis of variance for the regression with two x variables (table A), you will see that the regression on both height and weight accounted for an amount 388·09 with 2 degrees of freedom. This means that adding height to the equation using weight alone has accounted for an *extra sum of squares* of 388·09−326·10=61·99 with 1 degree of freedom. We can thus write the original analysis of variance table in an extended form as in table C.

The F ratio of 16·55 for height after weight has 1 degree of freedom on top and so must be the square of a t value. Indeed, if you will look at the original regression output in table A, you will see that the t value for the height coefficient is 4·069, which is just the square root of 16·55. To this extent, the two analyses are telling exactly the same story. However, the corresponding t value of 9·678 for weight is by no means the square root of 29·58, the F ratio in the first row of table C; for this we need the alternative split of the 2 degrees of freedom for regression obtained by doing the simple regression on height first and then adding weight (table D).

Now the F value of 93·66 for weight after height is just the square of the t value of 9·678 from the original analysis in table A. Both the

	df	SS	MS	F	p
Regression	1	326.10	326.10	29.58	0.0006
Residual	8	88.206	11.026		
Total	9	414.30			

Table B Regression of chest circumference on weight

Regression on	df	SS	MS	F	p
Weight	1	326.10	326.10	29.58	0.0006
then Height	1	61.99	61.99	16.55	0.0048
Residual	7	26.214	3.7448		
Total	9	414.30			

Table C Regression of chest circumference on weight, plus height

Regression on	df	SS	MS	F	p
Height	1	37.355	37.355	0.79	0.3992
then Weight	1	350.74	350.74	93.66	0.0000
Residual	7	26.214	3.7448		
Total	9	414.30	46.034		

Table D Regression of chest circumference on height plus weight

Table 3 Rectal temperature, axillary temperature, and heart rate

Rectal	*Axillary*	*Heart rate*
36·8	36·0	73
36·4	35·8	74
37·6	36·8	83
37·3	36·5	84
37·0	36·5	82
37·2	36·7	81
38·0	37·3	92
36·8	36·3	74
37·5	36·7	85
37·7	36·9	83

t values in the original analysis assess the significance of the corresponding x variable *after the other one has been allowed for*. Note that we need two analysis of variance tables for analysing the same body of data.

A rather similar looking set of data is shown in table 3. This shows the rectal and axillary temperatures and heart rates of 10 babies. If we do the regression of heart rate on the two temperature variables the computer output is shown in table E. (Notice the absurd value of the constant term – the estimated mean heart rate when both temperatures are zero. A more sensible display of the results would have used (temperature −37) for each of the x variables.) Looking at the coefficients and their standard errors, this looks pretty disappointing – the two t values are around 1 and far from any kind of meaningful significance level. It appears that heart rate is not predictable from the temperature variables. But now look at the analysis of variance. The F ratio is significant beyond the 0·002 level! This says that the individual null hypotheses $\beta_1=0$ and $\beta_2=0$ are both quite plausible, while the joint hypothesis ($\beta_1=0$, $\beta_2=0$) is not. Have we arrived at a contradiction within a single method of statistical analysis?

By no means. If we construct one of the progressive analyses of variance we get the analysis in table F. Rectal temperature by itself is quite an effective predictor, but adding in axillary temperature does very little extra good.

But the other progressive analysis of variance is quite similar (table G). Now axillary temperature by itself is a good predictor but if it is in the equation, there is no point in adding in rectal temperature. The two temperature variables are essentially telling the same story. Either one alone is an excellent predictor; given either one, there is no need for the other.

This example along with the previous one illustrates the most important lesson to be borne in mind when confronted with the results of a multiple regression analysis. The magnitude, the significance, and the interpretation of a partial regression coefficient all depend upon what other covariates are included in the equation. There is in general no such thing as *the* effect of an x on a y; we need to know what other x's are involved and whether they are being controlled for. We should really write the multiple regression equation in the form

$$E(y \mid x_1,x_2)=\alpha+\beta_{1\cdot2}x_1+\beta_{2\cdot1}x_2$$

so that each of the coefficients refers to the other x variable as well as to its own.

This underlines the essential problem of interpreting observational (as opposed to experimental) data, such as are the rule in epidemiological investigations. With such data it is always difficult to ensure that all the relevant x variables have been considered and properly allowed for. In an experimental setting, the effect of one x variable on another can be eliminated by a careful choice of treatments, and the possibility of overlooked x variables can be coped with by the device of *randomisation*, a topic I hope to return to in a later article. This is the reason why experimental findings are, potentially at least, more firmly based than those of purely observational studies.

It is worth inquiring a little more closely into the cause of the trouble in the last example. The problem arises from the fact that the two x variables are closely correlated – this is the statistical version of my remark that they are both telling the same story. Two such covariates (for reasons based upon the underlying mathematics) are said to be almost *collinear*. One effect of collinearity is to produce imprecise estimates of the coefficients, with large standard errors. Although the computer output does not disclose the fact, the estimates themselves are also closely correlated – if one is too big by chance, the other will almost surely be too small. The opposite of collinear is *orthogonal*; orthogonal variables are essentially uncorrelated. The reason for introducing these new terms is that the x variables do not (as we shall see) have to be statistical variates for which correlation language is appropriate.

When two covariates are nearly collinear, it is often helpful to do the regression on the difference between them as one predictor and their sum or mean as the other. These new covariates for our example are shown in table 4. The regression output using them is shown in table H.

Note that this is exactly the same regression as before – the intercept is the same, so is the residual standard error and so (if you work them out) are the predicted values. We have simply expressed it in terms of more convenient variables. It will be seen that there is (much as might be expected) a very

Dependent variable HR

	df	SS	MS	F	p
Regression	2	265.06	132.53	17.89	0.0018
Residual	7	51.843	7.4061		
Total	9	316.90	35.211		

Variance accounted for = 79.0%

Residual s.d. taken to be 2.7214

Regression coefficients

	b	se	t
RECTAL	4.4243	6.6465	0.666
AXILLARY	7.5867	7.4047	1.025
Constant term	-360.91	76.744	-4.703

Table E Regression of heart rate on rectal and axillary temperature

Regression on	df	SS	MS	F	p
Rectal	1	257.28	257.28	34.52	0.0004
then Axillary	1	7.78	7.78	1.05	0.3396
Residual	7	51.843	7.4061		
Total	9	316.90	35.211		

Table F Regression of heart rate on rectal plus axillary temperature

Regression on	df	SS	MS	F	p
Axillary	1	261.78	261.78	37.99	0.0003
then Rectal	1	3.28	3.28	0.44	0.5284
Residual	7	51.843	7.4061		
Total	9	316.90	35.211		

Table G Regression of heart rate on axillary plus rectal temperature

Table 4 Temperature difference, mean temperature, and heart rate

Difference	*Mean*	*Heart rate*
0·8	36·40	73
0·6	36·10	74
0·8	37·20	83
0·8	36·90	84
0·5	36·75	82
0·5	36·95	81
0·7	37·65	92
0·5	36·55	74
0·8	37·10	85
0·8	37·30	83

```
Dependent variable HR

                df        SS          MS          F          p
Regression       2       265.06      132.53      17.89      0.0018
Residual         7       51.843      7.4062
-------------------------------------------------------------------
Total            9       316.90      35.211

Variance accounted for =  79.0%

Residual s.d. taken to be          2.7214

Regression coefficients

                          b           se           t

DIFF                   -1.5812      6.9553      -0.227
MEAN                    12.011      2.1230       5.658
Constant term          -360.91      76.744      -4.703
```

Table H Regression of heart rate on difference and mean of rectal and axillary temperatures

significant regression on the mean of the two temperatures, while the difference between them has no appreciable effect. This kind of approach is often useful with other pairs of closely correlated covariates, notably systolic and diastolic blood pressure.

Curve fitting

Multiple regression in its standard form, as described above, is a handy technique to have available but does not play a major part in everyday medical statistics. It can, however, be generalised in all sorts of directions. As a first example of this, note that the covariates in a regression equation are completely unrestricted. Suppose that we relate a body measurement to age, but that the relationship when plotted is obviously not a straight line. We shall probably need age itself as a covariate but there is nothing to stop us including age^2, $log(age)$, $\sqrt{age}$, or combinations of these and other variables related to age itself. The only restriction is that the equation must only contain the coefficients as simple multipliers – terms such as $e^{\beta x}$ or $\log(\alpha+\beta x)$ require more complicated methodology.

The form of curve that is usually tried in an attempt to fit a curved line is the polynomial

$$E(y)=\beta_0+\beta_1 t+\beta_2 t^2+\beta_3 t^3+\ldots$$

with as many terms as necessary. In principle this can be fitted as a multiple regression equation, with $x_1=t$, $x_2=t^2$ and so on. In practice there are difficulties. When higher powers are introduced, the successive terms can become closely collinear, leading to large standard errors. Some improvement in this respect can be had by choosing the origin of the t variable somewhere near the middle of the data points. As well as this, polynomials often behave irregularly near the ends of the data, and they certainly should never be extrapolated outside the range of the actual data points. Royston and Altman have recently pointed out that there is no need to be restricted to integer powers of t, and that permitting the use of terms in $1/t$, $\sqrt{t}$ etc gives improved flexibility.[1] Berkey *et al* have used a curve which can be written as

$$E(y)=\beta_0+\beta_1 x+\beta_2 \log x+\beta_3/x+\beta_4/x^2$$

to fit height measurements on children aged from 8 to 18 years.[2] Other curves which have not been much exploited are the trigonometric series

$$E(y)=\beta_0+\beta_1 \sin t+\beta_2 \cos t+\beta_3 \sin 2t+\beta_4 \cos 2t+\ldots$$

The usefulness of these is not confined to periodic phenomena.

There are in fact an enormous number of possible curves which can be fitted using multiple regression. Choice of a curve, including the number of terms to include in a polynomial or other series, should be guided by careful inspection of the residuals.

1 Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied Statistics* 1994; **43:** 429–67.
2 Berkey CS, Laird NM, Valadian I, Gardner J. The analysis of longitudinal growth data with covariates. In: Tanner JM, ed. *Auxology 88: Perspectives in the science of growth and development.* London: Smith-Gordon, 1989: 31–9.