Archives of Disease in Childhood 1995; **73**: 270–274

STATISTICS FROM THE INSIDE

16. Multiple regression (2)

M J R Healy

The selection of x-variables

One of the commonest problems of multiple regression, and one of the most difficult, is that of selecting a few useful covariates out of a large selection of potential candidates. If we fit a multiple regression equation and find that one of the regression coefficients is little greater than its standard error, this means that the goodness of prediction as measured by the residual standard deviation has not been reduced to any great extent by including the corresponding x-variable in the regression model. It is then natural to try omitting this x-variable and fitting the data without it.

This kind of approach has been formalised in the various types of *stepwise regression* which are widely available in computer packages. The idea is quite simple. With forward stepwise regression the first step is to fit all the candidates covariates one at a time and to pick the most significant (the one, that is, that produces the greatest reduction in the residual standard deviation). This is then accepted as part of the eventual equation. The remaining covariates are added to the equation one at a time and again the most significant is selected and accepted. The process continues until none of the remaining candidates reaches some predetermined level of significance. Backwards stepwise regression starts by fitting all the candidate covariates simultaneously. The least significant is discarded and the fitting repeated with the remainder. This procedure continues until all the remaining covariates surpass a preset significance level. More generally, the two procedures can be adopted in alternation, adding 'significant' covariates to the equation and discarding those which become 'non-significant'.

Stepwise regression is an appealing method as it appears to offer a means of obtaining an optimal regression equation by a thought-free automatic technique. It has several major drawbacks that are not apparent on the surface. To begin with, different stepwise methods are capable of producing different selections of covariates from the same set of data. None of the methods in fact guarantee the selection of an optimal set, in the sense of one which minimises the residual standard deviation. Most importantly, they inevitably focus on a single selection of covariates, when it is usually the case that several different selections are practically equivalent in terms of goodness of prediction. The stopping rules ('F to remove' and so on) are almost entirely arbitrary, and the ostensible significance levels are so untrustworthy as to be positively misleading. A strong case can be made for preferring a model with fewer covariates, even if the residual mean square is slightly higher than that for a competitor with more covariates. This is rarely if ever taken into account by stepwise regression programs.

Unless the number of potential covariates is extremely large, it is entirely practical to explore all the possible selections. With 20 covariates, there are 1 048 576 possibilities to be compared, but ingenious methodology means that they do not all have to be considered. With care it is possible to examine (say) the best five selections for each total number of covariates, and this gives a much better picture of the overall situation. Software for these calculations is now available is some of the larger packages.

There is, however, one aspect of covariate selection, whatever mechanism is used, which is little appreciated, even by statisticians. Consider confronting a set of 20 potential predictors whose true regression coefficients, unknown to you, are all actually zero. If you calculate all the 1 048 576 possible regressions with their residual standard deviations, the average of these latter will be close to the original standard deviation of the y-variate. But some of them by chance will be larger and some will be smaller, and with this range of possibilities some are likely to be considerably smaller. There will thus be some, perhaps many selections of the covariates which appear to give excellent predictions, and the residual standard deviation reached as a result of the selection procedure will be a gross underestimate of the true value. A regression equation obtained in this way is likely to give extremely disappointing results when applied to a future sample taken from the same population. In much the same way, the estimated regression coefficients of the selected covariates following a stepwise procedure are likely to be larger, maybe considerably larger, than their true values. This is especially important when one of the x's is a measure of exposure and the others are merely possible confounding factors, so that the size of one of the coefficients is especially important.[1] A regression equation with a small number of covariates selected from a larger set must be interpreted with the greatest caution. If at all possible, its implications should be checked using a separate sample of data from the one used in the calculations.

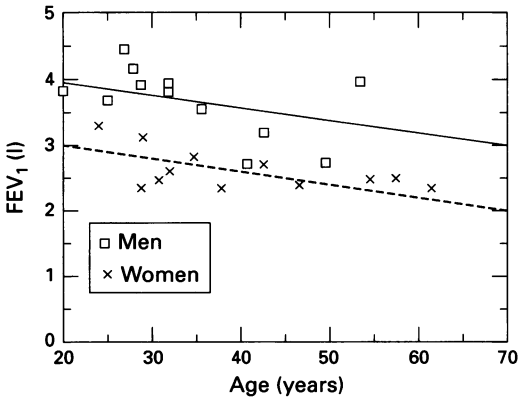
23 Coleridge Court, Milton Road, Harpenden, Herts AL5 5LD

Correspondence to: Professor Healy.

No reprints available.

Table 1 FEV_1 *and age in men and women*

	Men		Women	
	Age	FEV_1	*Age*	FEV_1
	20	3·87	24	3·30
	25	3·68	32	2·61
	32	3·81	31	2·47
	43	3·20	58	2·52
	54	3·97	62	2·35
	27	4·45	47	2·42
	29	3·89	35	2·83
	36	3·57	29	2·35
	41	2·73	43	2·73
	50	2·74	38	2·36
	32	3·88	55	2*50
	28	4·14	29	3·13
Means	34·8	3·661	40·3	2·631



FEV_1 *against age in men and women.*

One question that is frequently asked is which of a number of covariates (or which subset) is the most *important*. The exact meaning of this question is by no means clear, and there is only a tiny statistical literature on the subject, much of it inaccessible.[2] It can certainly not be asserted that one covariate is more important than another just because one of them was the first of the two to be selected in a forward stepwise procedure, nor are the 'standardised beta's' produced by some regression programs particularly relevant. Indeed, asking for the relative importance of two covariates tacitly supposes that only one of them will eventually be used, and this suggests that models which include both covariates should not be considered.

Regression with counted proportions

In a previous article dealing with data transformations, I mentioned the log odds or *logit* transformation of a proportion p:

$$\text{logit}(p) = \log_e\left(\frac{p}{1-p}\right)$$

If the quantity to be predicted or explained is a counted proportion p (so many 'successes' out of so many 'trials'), linear regression of p on a covariate x may be a rather implausible model as it is capable of predicting values of p outside the range 0 to 1. Transforming the p's to logits gets rid of this drawback, as the logit variable can go from minus infinity to infinity. Linearity of the regression of logit(p) on a covariate cannot be assumed, any more than it can with a continuous y-variate, but the transform is found in practice to provide a well fitting model for data in a large number of instances.

Just as ordinary simple regression can be extended to multiple regression with two or more predictors, so multiple logistic regression can also be used. The fitting procedure is arithmetically more complicated, but this is taken care of by suitable computer packages (the GLIM package (NAG Ltd, Oxford OX2 8DR) is one of the most satisfactory).

Apart from this, the analogy between ordinary and logistic regression is complete and the interpretation of the regression equation, with all its problems, is essentially the same for both. The regression coefficients can be interpreted by remembering that the logit of a proportion p is the logarithm of the corresponding odds. Suppose that a particular covariate x has a coefficient b; then a unit increase in x increases the log odds by an amount b. This means that the odds themselves are increased by a factor of e^b. Thus if b=1·6, say, the odds are increased by a factor of $e^{1·6}$=4·95.

One new feature does arise in the logistic case. The distribution of the errors (the variation of the observed p's about their true values) is binomial rather than Normal, and this means that a theoretical rather than an estimated error variance is available. As a result, the analogue of the residual sum of squares (which is known as the *deviance*) can be referred to tables of χ^2 and used as a test of the goodness of fit of the model. Note, however, that the test is one of the omnibus type that tries to guard against all possible kinds of misfit and so is liable to lack power in reacting to any particular kind such as systematic curvature of the relationship. It must also be borne in mind that the test assumes 'large' denominators, though as usual little guidance is available as to how large is 'large'. At any rate, the test cannot be used with so-called binary data where the denominators in the 'proportions' are all 1's.

Analysis of covariance

In an earlier note, I mentioned a type of analysis which is usually known as an analysis of covariance. The appropriate situation may most easily be appreciated from an example such as that in table 1 which shows values for forced expiratory volume in one second (FEV_1) in samples of men and women. It will be seen from the table that the difference between the means amounts to just over 1 litre. However, FEV_1 declines with age, and the women are on average some 5·5 years older than the men. How much of the difference can be ascribed to the age discrepancy between the two groups?*

The data are shown graphically in the figure. Regression lines of FEV_1 on age have been fitted to the two samples with the slopes constrained to be equal, so that the lines are parallel. The fair comparison, the mean

*I am aware that FEV_1 is usually standardised against height rather than age. Please accept my non-clinician's examples with a degree of indulgence.

Table 2 Heights of Swedish boys

Age (years)	Height (cm)	Social class
4·2026	99·00	1·0000
4·0684	109·50	2·0000
4·0301	106·00	3·0000
4·1862	107·00	1·0000
4·1369	108·50	2·0000
4·1314	108·00	1·0000
4·0411	111·50	1·0000
4·1834	102·00	3·0000
3·9808	104·00	3·0000
4·0602	104·50	1·0000
...	...	...

difference in FEV_1 between a man and a woman of the same age, is seen to be given by the vertical distance between the two lines, which is measured by the difference between the constant terms in the two regression equations. We can estimate this very conveniently by a multiple regression equation. One of the x-variables will obviously be age. The other should be a made-up *dummy* variable which takes the value 0 for all the men and 1 for all the women – let us call it z. Now consider the regression equation for the whole of the data

$$E(FEV_1)=\alpha+\beta_1\times age+\beta_2\times z$$

Because the value of z is either 0 or 1, this is equivalent to two equations:

for men

$$E(FEV_1)=\alpha+\beta_1\times age$$

and for women

$$E(FEV_1)=(\alpha+\beta_2)+\beta_1\times age$$

These are the equations of two parallel lines each with slope β_1 and the vertical distance between them is given by the difference between the constant terms, which is β_2. My Nanostat computer program produces the following estimates for the coefficients:

	b	SE	t
Age	−0·019058	0·0070390	−2·707
Z	−0·92518	0·16057	−5·762
Constant term	4·3231	0·26828	16·114

```
Dependent variable HT

Caution - x's linearly dependent

                df          SS          MS           F          p
Regression       3      470.88      156.96        9.91     0.0000
Residual       734       11631.      15.846
-------------------------------------------------------------------
Total          737       12102.      16.421

Variance accounted for =   3.5%

Residual s.d. taken to be          3.9808

Regression coefficients

                              b           se            t

AGE                      9.4221      1.8088        5.209
SC1                     0.31039     0.37834        0.820
SC2                     0.61457     0.35472        1.733
SC3                     0.00000     0.00000     9999.000
Constant term            66.356      7.3447        9.035

Table A Regression of height on age and social class
```

The vertical distance between the lines is thus −0·9252, with 95% confidence interval −1·2591 to −0·5911. The age effect is significant at the 1% level, but it is not large enough to reduce the estimated mean difference by more than a small amount.

The analysis above relies heavily on the parallelism of the two regression lines – it is this which has enabled us to make a general statement about the gender difference which is valid across the whole age range of the data. But is the assumption of parallelism justified? We can investigate this by constructing and fitting a second dummy variable which is equal to 0 for all the men and to age for all the women. Calling this zz, the multiple regression equation

$$E(FEV_1)=\alpha+\beta_1\times age+\beta_2\times z+\beta_2\times zz$$

is equivalent to the two equations:

for men

$$E(FEV_1)=\alpha+\beta_1\times age$$

and for women

$$E(FEV_1)=(\alpha+\beta_2)+(\beta_1+\beta_2)\times age$$

so that the two lines have different slopes and different constant terms. The computer produces the following estimates:

	b	SE	t
Age	−0·0127819	0·0011110	−2·504
Z	−1·94654	0·55407	−2·645
ZZ	−0·014618	0·014351	1·019
Constant term	4·6275	0·401246	11·527

The coefficient of zz is β_3 which measures the difference between the separate slopes. It is scarcely larger than its standard error, so that there is no serious evidence from the data for lack of parallelism.

In this analysis, age is known as a *confounding factor*, one which influences the outcome variable but which is not of direct interest to the investigator. Confounding factors can usually be avoided in an experimental setup, either by equalising them or by randomisation, but they are ubiquitous in observational studies and multiple regression provides a powerful methodology for investigating and at least to some extent eliminating their effects.

Predictors with categories

In the analysis of covariance above, one of the predictors in the regression equation was sex, a classification with two categories. To include this within the multiple regression model, we constructed a dummy variable which took the value 0 for males and 1 for females. The same kind of technique can be used to cope with a classification with more than two categories (the statistical jargon refers to it as a *factor* with more than two *levels*), but some care is needed. Table 2 shows the start of some data on the heights of 4 year old Swedish boys (I am indebted to Dr G Tanner-Lindgren for access to these data). The ages actually range from 3·75 to 4·25 years, and the children are classified into three socioeconomic groups – social classes for short – labelled 1, 2, and 3. The question at issue is whether social class affects

```
Dependent variable HT

Caution - x's linearly dependent

                df          SS          MS           F          p
Regression       2      40.924      20.462        1.25     0.2880
Residual       735       12061.      16.410
-------------------------------------------------------------------
Total          737       12102.      16.421

Variance accounted for =   0.1%

Residual s.d. taken to be          4.0509

Regression coefficients

                              b           se            t

SC1                     0.27415     0.38494        0.712
SC2                     0.56707     0.36086        1.571
SC3                     0.00000     0.00000     9999.000
Constant term            104.59     0.27311       382.950

Table B  Regression of height on social class
```

HT

	df	SS	MS	F	P
Between groups	2	40.922	20.461	1.25	0.288(
Within groups	735	12061.	16.410		
Total	737	12102.	16.421		

Pooled within-group S.D. = 4.0509 (3.9% of mean)
Between-group variance component = 0.016634

SC

	N	Mean	S.E.M.
1	223	104.86	0.27127
2	295	105.16	0.23585
3	220	104.59	0.27311

Table C One way analysis of variance, height by social class

height, and it is natural to try to relate height to social class in some kind of regression analysis. We shall want to include age in the regression to allow for the fact that the mean age in the three social classes will not be exactly the same – as in the previous example, age is acting as a confounding factor.

The first trap to avoid is that of including a single x-variable for social class which takes the values 1, 2, and 3. Doing this involves the tacit assumption that the difference in mean height between classes 1 and 2 is the same as that between classes 2 and 3, and this may well not be so. The numbers 1, 2, and 3 in the third column of the table are in fact not what they seem; they are ordinal numbers and should really be written as 1st, 2nd, and 3rd. In presenting data of this type, it is a good idea to avoid temptation by labelling the categories A, B, and C rather than numerically.

As a better method, let us construct three dummy variables to correspond to the three social classes. The first dummy variable will take the value 1 for all children in social class 1 and 0 for all others; the second will take the value 1 for all the children in social class 2 and 0 for all the others; the third, the value 1 for all the children in social class 3 and 0 for all the others. Doing this and invoking the Nanostat regression command produces the results in table A.

Various things are at once apparent. There is an obscure warning message at the top of the computer output; the SC3 predictor has an odd looking regression coefficient; and although there are four x-variables in the equation, the analysis of variance shows only 3 degrees of freedom for regression. Clearly something peculiar has happened.

The explanation can be found by pointing out that our construction of dummy x-variables had been too enthusiastic. Each individual child must belong to one or other of the three social classes. It follows that each child will have one of its dummy variables equal to 1 and the other two equal to 0. The three dummy variables for any particular child must thus add up to 1, and it follows that if you tell me the values of any two of them, I can deduce the value of the third without looking at the data. Although there are three constructed predictors, there are only two pieces of information. This argument should not be unfamiliar; it exactly parallels that for assigning (n−1) degrees of freedom to the sum of squares of n deviations for the mean when estimating a variance or standard deviation. The three categories of the social class classification entitle us to only 2 degrees of freedom, and this is reflected in the regression line of the analysis of variance.

The situation can be looked at in another way. For a particular child, the predicted value from the regression equation can be written as

$$a+b_1\times age+b_2\times SC1+b_3\times SC2+b_4\times SC3$$

where one of the SC variables has the value 1 and the other two the value 0. Now suppose we were to add some quantity, 10 say, to each of the social class b's and at the same time subtract 10 from the intercept a. It will be seen that the predictions from the regression are entirely unchanged. When a set of x-variables are linked by an exact relationship – when the x's are linearly dependent, as the jargon has it – the corresponding coefficients are not uniquely determined by the data, and an equally good fit can be had using quite different values for them. We are at liberty to give one of the coefficients an arbitrary value, and the other coefficients along with the constant term can be adjusted to compensate. The computer program, of course, has to make a specific choice, and it has set the SC3 coefficient exactly equal to 0. The zero standard error for this coefficient indicates that it is not an empirical estimate but rather an imposed arbitrary value.

So what is the use of an analysis with this kind of arbitrary outcome? I showed just now that we can add any constant value to the three social class coefficients provided that we subtract the same value from the constant term. But doing this does not affect the quantities that we are really interested in, the *differences* between the social class b's. These can be estimated uniquely from the data. As presented, the coefficients for SC1 and SC2 in the computer output represent the differences in age adjusted mean height between classes 1 and 2 and the 'reference class' 3. It can be seen that they are taller by 0·3 and 0·6 cm on average, but that these amounts are not large when compared with their standard errors. It may be noted that the class 2 boys are actually slightly taller on average than those in class 1; though the effect is far from significant, it would be totally concealed by an analysis which treated the social class codes as if they were values of a continuous measurement.

Some further light may be case by considering for illustrative purposes an analysis of the same data without allowing for age, using just the three social class dummy x-variables. The Nanostat computer output is shown in table B and the outcome is very little changed.

But the data structure is now rather simpler than before. We have in fact three independent samples and these constitute a one way classification, a generalisation of that for which the unpaired t test is used. Readers of one of my previous notes may recall that an appropriate method of analysis of such a structure is the one way analysis of variance, and the Nanostat

computer output from such an analysis is shown in table C.

A comparison of the two analyses is quite instructive. It can be seen that the analyses of variance are effectively identical. Moreover, the group means in the second analysis can be obtained from the first by simply adding the appropriate SC coefficients to the constant term. In fact, the analysis of variance can be regarded simply as a special case of the multiple regression analysis. The same can be shown to be true of two way and more complex analyses, and also *a fortiori* of the simpler *t* tests which are special cases of the analysis of variance. Multiple regression underlies almost all the most useful statistical techniques, so that its theoretical as well as its practical importance can scarcely be overrated.

1 Raab GM. Selecting confounders from covariates. *Journal of the Royal Statistical Society A* 1994; **157:** 271–83.
2 Healy MJR. Measuring importance. *Stat Med* 1990; **9:** 633–7.