# Calculating correlation coefficients with repeated observations: Part 1—correlation within subjects

J Martin Bland, Douglas G Altman

*This is the twelfth in a series of occasional notes on medical statistics*

In an earlier *Statistics Note*[1] we commented on the analysis of paired data where there is more than one observation per subject, as shown in table I. We pointed out that it could be highly misleading to analyse such data by combining repeated observations from several subjects and then calculating the correlation coefficient as if the data were a simple sample. This note is a response to several letters about the appropriate analysis for such data.

TABLE I—*Repeated measurements of intramural pH and $Paco_2$ for eight subjects*[2]

| Subject | pH | $Paco_2$ | Subject | pH | $Paco_2$ |
|---|---|---|---|---|---|
| 1 | 6·68 | 3·97 | 5 | 7·30 | 4·32 |
| 1 | 6·53 | 4·12 | 5 | 7·37 | 3·23 |
| 1 | 6·43 | 4·09 | 5 | 7·27 | 4·46 |
| 1 | 6·33 | 3·97 | 5 | 7·28 | 4·72 |
| 2 | 6·85 | 5·27 | 5 | 7·32 | 4·75 |
| 2 | 7·06 | 5·37 | 5 | 7·32 | 4·99 |
| 2 | 7·13 | 5·41 | 6 | 7·38 | 4·78 |
| 2 | 7·17 | 5·44 | 6 | 7·30 | 4·73 |
| 3 | 7·40 | 5·67 | 6 | 7·29 | 5·12 |
| 3 | 7·42 | 3·64 | 6 | 7·33 | 4·93 |
| 3 | 7·41 | 4·32 | 6 | 7·31 | 5·03 |
| 3 | 7·37 | 4·73 | 6 | 7·33 | 4·93 |
| 3 | 7·34 | 4·96 | 7 | 6·86 | 6·85 |
| 3 | 7·35 | 5·04 | 7 | 6·94 | 6·44 |
| 3 | 7·28 | 5·22 | 7 | 6·92 | 6·52 |
| 3 | 7·30 | 4·82 | 8 | 7·19 | 5·28 |
| 3 | 7·34 | 5·07 | 8 | 7·29 | 4·56 |
| 4 | 7·36 | 5·67 | 8 | 7·21 | 4·34 |
| 4 | 7·33 | 5·10 | 8 | 7·25 | 4·32 |
| 4 | 7·29 | 5·53 | 8 | 7·20 | 4·41 |
| 4 | 7·30 | 4·75 | 8 | 7·19 | 3·69 |
| 4 | 7·35 | 5·51 | 8 | 6·77 | 6·09 |
| 5 | 7·35 | 4·28 | 8 | 6·82 | 5·58 |
| 5 | 7·30 | 4·44 | | | |

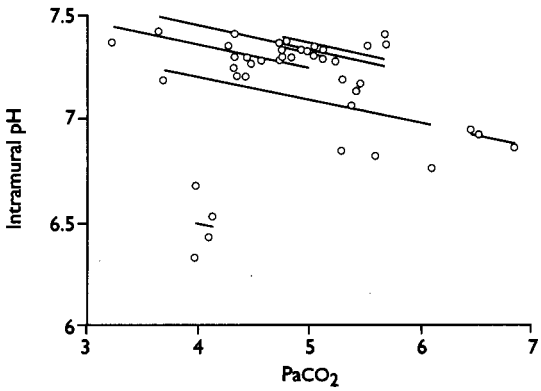
The choice of analysis for the data in table I depends on the question we want to answer. If we want to know whether subjects with high values of intramural pH also tend to have high values of $Paco_2$ we are interested in whether the average pH for a subject is related to the subject's average $Paco_2$. We can use the correlation between the subject means, which we shall describe in a subsequent note. If we want to know whether an increase in pH within the individual was associated with an increase in $Paco_2$ we want to remove the differences between subjects and look only at changes within.

To look at variation within the subject we can use multiple regression. We make one of our variables, pH or $Paco_2$, the outcome variable and the other variable and the subject the predictor variables. Subject is treated as a categorical factor using dummy variables [3 4] and so has seven degrees of freedom. We use the analysis of variance table [3 4] for the regression (table II),

TABLE II—*Analysis of variance for the data in table I*

| Source of variation | Degrees of freedom | Sum of squares | Mean square | Variance ratio (F) | Probability |
|---|---|---|---|---|---|
| Subjects | 7 | 2·9661 | 0·4237 | 48·3 | <0·0001 |
| $Paco_2$ | 1 | 0·1153 | 0·1153 | 13·1 | 0·0008 |
| Residual | 38 | 0·3337 | 0·0088 | | |
| Total | 46 | 3·3139 | 0·0720 | | |

which shows how the variability in pH can be partitioned into components due to different sources. This method is also known as analysis of covariance and is equivalent to fitting parallel lines through each subject's data (see figure). The residual sum of squares




*pH against $Paco_2$ for eight subjects, with parallel lines fitted for each subject*

in table II represents the variation about these lines. We remove the variation due to subjects (and any other nuisance variables which might be present) and express the variation in pH due to $Paco_2$ as a proportion of what's left:

$$\frac{\text{Sum of squares for } Paco_2}{\text{Sum of squares for } Paco_2 + \text{residual sum of squares}}$$

The magnitude of the correlation coefficient within subjects is the square root of this proportion. For table II this is:

$$\sqrt{\frac{0{\cdot}1153}{0{\cdot}1153+0{\cdot}3337}} = 0{\cdot}51$$

The sign of the correlation coefficient is given by the sign of the regression coefficient for $Paco_2$. Here the regression slope is −0·108, so the correlation coefficient within subjects is −0·51. The P value is found either from the F test in the associated analysis of variance table, or from the *t* test for the regression slope. It doesn't matter which variable we regress on which; we get the same correlation coefficient and P value either way.

If we incorrectly calculate the correlation coefficient ignoring the fact that we have 47 observations on only 8 subjects, we get −0·07, P=0·7. Hence the correct analysis within subjects reveals a relation which the incorrect analysis misses.

**Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE**
J Martin Bland, *reader in medical statistics*

**Medical Statistics Laboratory, Imperial Cancer Research Fund, PO Box 123, London WC2A 3PX**
Douglas G Altman, *head*

Correspondence to:
Dr Bland.

*BMJ* 1995;310:446

1 Bland JM, Altman DG. Correlation, regression, and repeated data. *BMJ* 1994;308:896.
2 Boyd O, Mackay CJ, Lamb G, Bland JM, Grounds RM, Bennett ED. Comparison of clinical information gained from routine blood-gas analysis and from gastric tonometry for intramural pH. *Lancet* 1993;341:142-6.
3 Altman DG. *Practical statistics for medical research.* London: Chapman and Hall, 1991.
4 Armitage P, Berry G. *Statistical methods in medical research.* 3rd ed. Oxford: Blackwell, 1994.