Review

Statistics review 7: Correlation and regression

Viv Bewick[1], Liz Cheek[1] and Jonathan Ball[2]

[1]Senior Lecturer, School of Computing, Mathematical and Information Sciences, University of Brighton, Brighton, UK
[2]Lecturer in Intensive Care Medicine, St George's Hospital Medical School, London, UK

Correspondence: Viv Bewick, v.bewick@brighton.ac.uk

Published online: 5 November 2003 *Critical Care* 2003, **7**:451-459 (DOI 10.1186/cc2401)
This article is online at http://ccforum.com/content/7/6/451

Abstract

The present review introduces methods of analyzing the relationship between two quantitative variables. The calculation and interpretation of the sample product moment correlation coefficient and the linear regression equation are discussed and illustrated. Common misuses of the techniques are considered. Tests and confidence intervals for the population parameters are described, and failures of the underlying assumptions are highlighted.

Keywords coefficient of determination, correlation coefficient, least squares regression line

Introduction

The most commonly used techniques for investigating the relationship between two quantitative variables are correlation and linear regression. Correlation quantifies the strength of the linear relationship between a pair of variables, whereas regression expresses the relationship in the form of an equation. For example, in patients attending an accident and emergency unit (A&E), we could use correlation and regression to determine whether there is a relationship between age and urea level, and whether the level of urea can be predicted for a given age.

Scatter diagram

When investigating a relationship between two variables, the first step is to show the data values graphically on a scatter diagram. Consider the data given in Table 1. These are the ages (years) and the logarithmically transformed admission serum urea (natural logarithm [ln] urea) for 20 patients attending an A&E. The reason for transforming the urea levels was to obtain a more Normal distribution [1]. The scatter diagram for ln urea and age (Fig. 1) suggests there is a positive linear relationship between these variables.

Correlation

On a scatter diagram, the closer the points lie to a straight line, the stronger the linear relationship between two variables. To quantify the strength of the relationship, we can calculate the correlation coefficient. In algebraic notation, if we have two variables x and y, and the data take the form of n pairs (i.e. $[x_1, y_1]$, $[x_2, y_2]$, $[x_3, y_3]$ … $[x_n, y_n]$), then the correlation coefficient is given by the following equation:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where $\bar{x}$ is the mean of the x values, and $\bar{y}$ is the mean of the y values.

This is the product moment correlation coefficient (or Pearson correlation coefficient). The value of r always lies between −1 and +1. A value of the correlation coefficient close to +1 indicates a strong positive linear relationship (i.e. one variable increases with the other; Fig. 2). A value close to −1 indicates a strong negative linear relationship (i.e. one variable decreases as the other increases; Fig. 3). A value close to 0 indicates no linear relationship (Fig. 4); however, there could be a nonlinear relationship between the variables (Fig. 5).

For the A&E data, the correlation coefficient is 0.62, indicating a moderate positive linear relationship between the two variables.

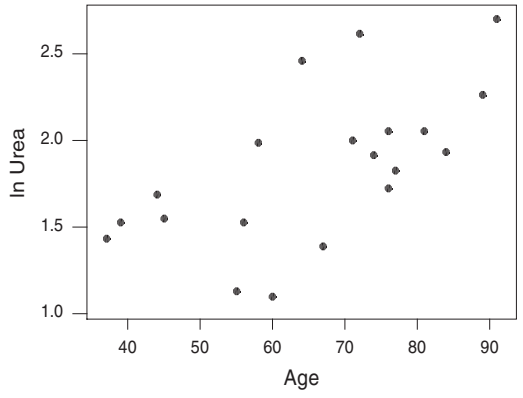
A&E = accident and emergency unit; ln = natural logarithm (logarithm base e).

Table 1

Age and ln urea for 20 patients attending an accident and emergency unit

Subject	Age (years)	ln urea
1	60	1.099
2	76	1.723
3	81	2.054
4	89	2.262
5	44	1.686
6	58	1.988
7	55	1.131
8	74	1.917
9	45	1.548
10	67	1.386
11	72	2.617
12	91	2.701
13	76	2.054
14	39	1.526
15	71	2.002
16	56	1.526
17	77	1.825
18	37	1.435
19	64	2.460
20	84	1.932

Figure 1

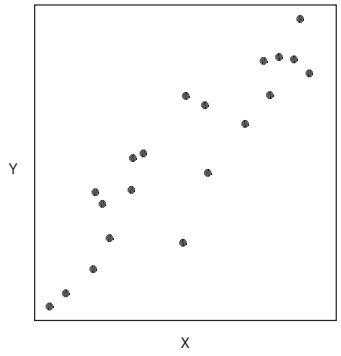


Scatter diagram for ln urea and age

Hypothesis test of correlation

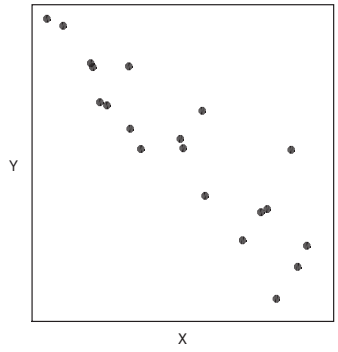
We can use the correlation coefficient to test whether there is a linear relationship between the variables in the population as a whole. The null hypothesis is that the population correlation coefficient equals 0. The value of r can be compared with those given in Table 2, or alternatively exact *P* values can be obtained from most statistical packages. For the A&E data, r=0.62 with a sample size of 20 is greater than the value highlighted bold in Table 2 for *P*=0.01, indicating a *P* value of less than 0.01. Therefore, there is sufficient evidence to suggest that the true population correlation coefficient is not 0 and that there is a linear relationship between ln urea and age.

Figure 2



Correlation coefficient (r) = +0.9. Positive linear relationship.

Figure 3



Correlation coefficient (r) = −0.9. Negative linear relationship.

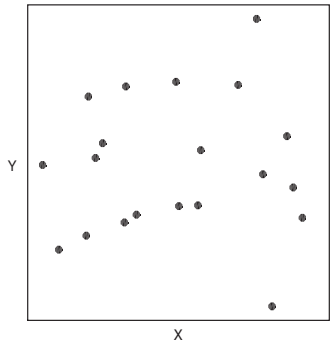
Confidence interval for the population correlation coefficient

Although the hypothesis test indicates whether there is a linear relationship, it gives no indication of the strength of that relationship. This additional information can be obtained from a confidence interval for the population correlation coefficient.

To calculate a confidence interval, r must be transformed to give a Normal distribution making use of Fisher's z transformation [2]:

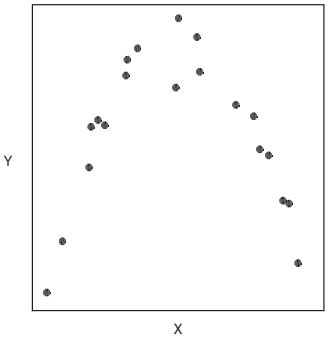
$$z_r = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right)$$

Figure 4



Correlation coefficient (r) = 0.04. No relationship.

Figure 5



Correlation coefficient (r) = –0.03. Nonlinear relationship.

The standard error [3] of z_r is approximately:

$$\frac{1}{\sqrt{n-3}}$$

and hence a 95% confidence interval for the true population value for the transformed correlation coefficient z_r is given by $z_r - (1.96 \times \text{standard error})$ to $z_r + (1.96 \times \text{standard error})$. Because z_r is Normally distributed, 1.96 deviations from the statistic will give a 95% confidence interval.

For the A&E data the transformed correlation coefficient z_r between ln urea and age is:

$$\frac{1}{2}\log_e\left(\frac{1+0.62}{1-0.62}\right)=0.725$$

The standard error of z_r is:

$$\frac{1}{\sqrt{20-3}}=0.242$$

The 95% confidence interval for z_r is therefore $0.725 - (1.96 \times 0.242)$ to $0.725 + (1.96 \times 0.242)$, giving 0.251 to 1.199.

Table 2

5% and 1% points for the distribution of the correlation coefficient under the null hypothesis that the population correlation is 0 in a two-tailed test

	r values for two-tailed probabilities (*P*)			Two-tailed probabilities (*P*)	
Sample size	0.05	0.01	Sample size	0.05	0.01
3	1.00	1.00	23	0.41	0.53
4	0.95	0.99	24	0.40	0.52
5	0.88	0.96	25	0.40	0.51
6	0.81	0.92	26	0.39	0.50
7	0.75	0.87	27	0.38	0.49
8	0.71	0.83	28	0.37	0.48
9	0.67	0.80	29	0.37	0.47
10	0.63	0.76	30	0.36	0.46
11	0.60	0.73	40	0.31	0.40
12	0.58	0.71	50	0.28	0.36
13	0.55	0.68	60	0.25	0.33
14	0.53	0.66	70	0.24	0.31
15	0.51	0.64	80	0.22	0.29
16	0.50	0.62	90	0.21	0.27
17	0.48	0.61	100	0.20	0.26
18	0.47	0.59	110	0.19	0.24
19	0.46	0.58	120	0.18	0.23
20	0.44	**0.56**	130	0.17	0.23
21	0.43	0.55	140	0.17	0.22
22	0.42	0.54	150	0.16	0.21

Generated using the standard formula [2].

We must use the inverse of Fisher's transformation on the lower and upper limits of this confidence interval to obtain the 95% confidence interval for the correlation coefficient. The lower limit is:

$$\frac{e^{2\times0.251}-1}{e^{2\times0.251}+1}$$

giving 0.25 and the upper limit is:

$$\frac{e^{2\times1.199}-1}{e^{2\times1.199}+1}$$

giving 0.83. Therefore, we are 95% confident that the population correlation coefficient is between 0.25 and 0.83.

The width of the confidence interval clearly depends on the sample size, and therefore it is possible to calculate the sample size required for a given level of accuracy. For an example, see Bland [4].

Misuse of correlation

There are a number of common situations in which the correlation coefficient can be misinterpreted.

One of the most common errors in interpreting the correlation coefficient is failure to consider that there may be a third variable related to both of the variables being investigated, which is responsible for the apparent correlation. Correlation does not imply causation. To strengthen the case for causality, consideration must be given to other possible underlying variables and to whether the relationship holds in other populations.

A nonlinear relationship may exist between two variables that would be inadequately described, or possibly even undetected, by the correlation coefficient.

A data set may sometimes comprise distinct subgroups, for example males and females. This could result in clusters of points leading to an inflated correlation coefficient (Fig. 6). A single outlier may produce the same sort of effect.

It is important that the values of one variable are not determined in advance or restricted to a certain range. This may lead to an invalid estimate of the true correlation coefficient because the subjects are not a random sample.

Another situation in which a correlation coefficient is sometimes misinterpreted is when comparing two methods of measurement. A high correlation can be incorrectly taken to mean that there is agreement between the two methods. An analysis that investigates the differences between pairs of observations, such as that formulated by Bland and Altman [5], is more appropriate.

Regression

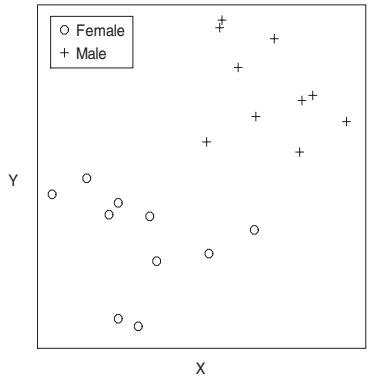
In the A&E example we are interested in the effect of age (the predictor or x variable) on ln urea (the response or y variable). We want to estimate the underlying linear relationship so that we can predict ln urea (and hence urea) for a given age. Regression can be used to find the equation of this line. This line is usually referred to as the regression line.

Note that in a scatter diagram the response variable is always plotted on the vertical (y) axis.

Equation of a straight line

The equation of a straight line is given by $y = a + bx$, where the coefficients a and b are the intercept of the line on the y axis and the gradient, respectively. The equation of the regression line for the A&E data (Fig. 7) is as follows: ln urea $= 0.72 + (0.017 \times \text{age})$ (calculated using the method of least squares, which is described below). The gradient of this line is 0.017, which indicates that for an increase of 1 year in age the expected increase in ln urea is 0.017 units (and hence the expected increase in urea is 1.02 mmol/l). The predicted ln urea of a patient aged 60 years, for example, is $0.72 + (0.017 \times 60) = 1.74$ units. This transforms to a urea level of $e^{1.74} = 5.70$ mmol/l. The y intercept is 0.72, meaning that if the line were projected back to age = 0, then the ln urea value would be 0.72. However, this is not a meaningful value because age = 0 is a long way outside the range of the data and therefore there is no reason to believe that the straight line would still be appropriate.

Figure 6



Subgroups in the data resulting in a misleading correlation. All data: r = 0.57; males: r = –0.41; females: r = –0.26.

Method of least squares

The regression line is obtained using the method of least squares. Any line $y = a + bx$ that we draw through the points gives a predicted or fitted value of y for each value of x in the data set. For a particular value of x the vertical difference between the observed and fitted value of y is known as the deviation, or residual (Fig. 8). The method of least squares finds the values of a and b that minimise the sum of the squares of all the deviations. This gives the following formulae for calculating a and b:

$$a = \bar{y} - b\bar{x} \qquad b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

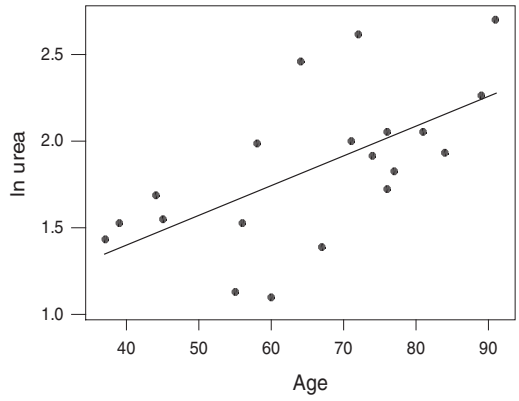
Usually, these values would be calculated using a statistical package or the statistical functions on a calculator.

Hypothesis tests and confidence intervals

We can test the null hypotheses that the population intercept and gradient are each equal to 0 using test statistics given by the estimate of the coefficient divided by its standard error.

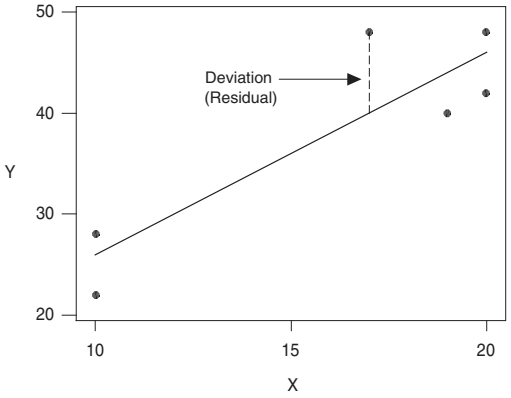
$$\text{The standard error of the intercept} = s\sqrt{\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$$

Figure 7



Regression line for ln urea and age: ln urea = 0.72 + (0.017 × age).

Figure 8



Regression line obtained by minimizing the sums of squares of all of the deviations.

and for the gradient $= \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

where $s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - b\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-2)}}$

The test statistics are compared with the t distribution on n – 2 (sample size – number of regression coefficients) degrees of freedom [4].

The 95% confidence interval for each of the population coefficients are calculated as follows: coefficient ± (t_{n-2} × the standard error), where t_{n-2} is the 5% point for a t distribution with n – 2 degrees of freedom.

For the A&E data, the output (Table 3) was obtained from a statistical package. The *P* value for the coefficient of ln urea (0.004) gives strong evidence against the null hypothesis, indicating that the population coefficient is not 0 and that there is a linear relationship between ln urea and age. The coefficient of ln urea is the gradient of the regression line and its hypothesis test is equivalent to the test of the population correlation coefficient discussed above. The *P* value for the constant of 0.054 provides insufficient evidence to indicate that the population coefficient is different from 0. Although the intercept is not significant, it is still appropriate to keep it in the equation. There are some situations in which a straight line passing through the origin is known to be appropriate for the data, and in this case a special regression analysis can be carried out that omits the constant [6].

Analysis of variance

As stated above, the method of least squares minimizes the sum of squares of the deviations of the points about the regression line. Consider the small data set illustrated in Fig. 9. This figure shows that, for a particular value of x, the distance of y from the mean of y (the total deviation) is the sum of the distance of the fitted y value from the mean (the deviation explained by the regression) and the distance from y to the line (the deviation not explained by the regression).

The regression line for these data is given by y = 6 + 2x. The observed, fitted values and deviations are given in Table 4. The sum of squared deviations can be compared with the total variation in y, which is measured by the sum of squares of the deviations of y from the mean of y. Table 4 illustrates the relationship between the sums of squares. Total sum of squares = sum of squares explained by the regression line + sum of squares not explained by the regression line. The explained sum of squares is referred to as the 'regression sum of squares' and the unexplained sum of squares is referred to as the 'residual sum of squares'.

This partitioning of the total sum of squares can be presented in an analysis of variance table (Table 5). The total degrees of freedom = n – 1, the regression degrees of freedom = 1, and the residual degrees of freedom = n – 2 (total – regression degrees of freedom). The mean squares are the sums of squares divided by their degrees of freedom.

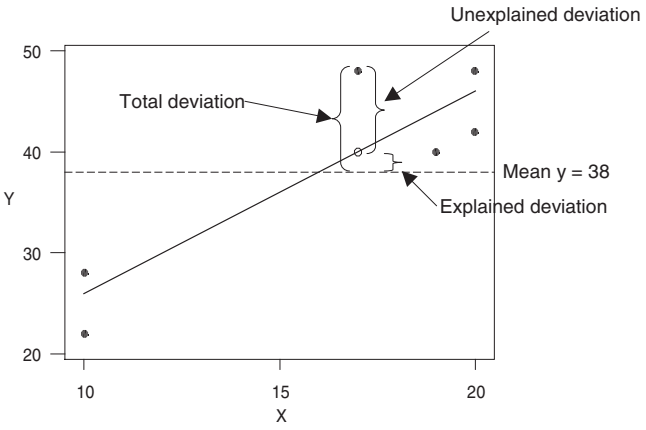
If there were no linear relationship between the variables then the regression mean squares would be approximately the same as the residual mean squares. We can test the null hypothesis that there is no linear relationship using an F test. The test statistic is calculated as the regression mean square divided by the residual mean square, and a *P* value may be obtained by comparison of the test statistic with the F distribution with 1 and n – 2 degrees of freedom [2]. Usually, this analysis is carried out using a statistical package that will produce an exact *P* value. In fact, the F test from the analysis of variance is equivalent to the t test of the gradient for

Table 3

Regression parameter estimates, *P* values and confidence intervals for the accident and emergency unit data

	Coefficient	Standard error of coefficient	t	*P*	Confidence interval
Constant, or intercept	0.72	0.346	2.07	0.054	−0.01 to +1.45
ln urea	0.017	0.005	3.35	0.004	0.006 to 0.028

Figure 9



Total, explained and unexplained deviations for a point.

regression with only one predictor. This is not the case with more than one predictor, but this will be the subject of a future review. As discussed above, the test for gradient is also equivalent to that for the correlation, giving three tests with identical *P* values. Therefore, when there is only one predictor variable it does not matter which of these tests is used.

The analysis of variance for the A&E data (Table 6) gives a *P* value of 0.006 (the same *P* value as obtained previously), again indicating a linear relationship between ln urea and age.

Table 5

Analysis of variance for a small data set

Source of variation	Degrees of freedom	Sum of squares	Mean square	F	*P*
Regression	1	456	456	15.2	0.018
Residual	4	120	30		
Total	5	576			

Coefficent of determination

Another useful quantity that can be obtained from the analysis of variance is the coefficient of determination (R^2).

$$R^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}}$$

It is the proportion of the total variation in y accounted for by the regression model. Values of R^2 close to 1 imply that most of the variability in y is explained by the regression model. R^2 is the same as r^2 in regression when there is only one predictor variable.

For the A&E data, $R^2 = 1.462/3.804 = 0.38$ (i.e. the same as 0.62^2), and therefore age accounts for 38% of the total variation in ln urea. This means that 62% of the variation in ln urea is not accounted for by age differences. This may be due to

Table 4

Small data set with the fitted values from the regression, the deviations and their sums of squares

x (mean x = 16)	y (mean y = 38)	Fitted y = 6 + 2x	Unexplained deviation = y − fitted y	Explained deviation = fitted y − mean y	Total deviation = y − mean y
10	22	26	−4	−12	−16
10	28	26	2	−12	−10
20	42	46	−4	8	4
20	48	46	2	8	10
19	40	44	−4	6	2
17	48	40	8	2	10
		Sum of squares	120	456	576

Table 6

Analysis of variance for the accident and emergency unit data

Source of variation	Degrees of freedom	Sum of squares	Mean square	F	*P*
Regression	1	1.462	1.462	11.24	0.004
Residual	18	2.342	0.130		
Total	19	3.804			

inherent variability in ln urea or to other unknown factors that affect the level of ln urea.

Prediction

The fitted value of y for a given value of x is an estimate of the population mean of y for that particular value of x. As such it can be used to provide a confidence interval for the population mean [3]. The fitted values change as x changes, and therefore the confidence intervals will also change.

The 95% confidence interval for the fitted value of y for a particular value of x, say x_p, is again calculated as fitted $y \pm (t_{n-2} \times$ the standard error). The standard error is given by:

$$s\sqrt{\left[\frac{1}{n}+\frac{(x_p-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]}$$

Fig. 10 shows the range of confidence intervals for the A&E data. For example, the 95% confidence interval for the population mean ln urea for a patient aged 60 years is 1.56 to 1.92 units. This transforms to urea values of 4.76 to 6.82 mmol/l.

The fitted value for y also provides a predicted value for an individual, and a prediction interval or reference range [3] can be obtained (Fig. 10). The prediction interval is calculated in the same way as the confidence interval but the standard error is given by:

$$s\sqrt{\left[1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]}$$

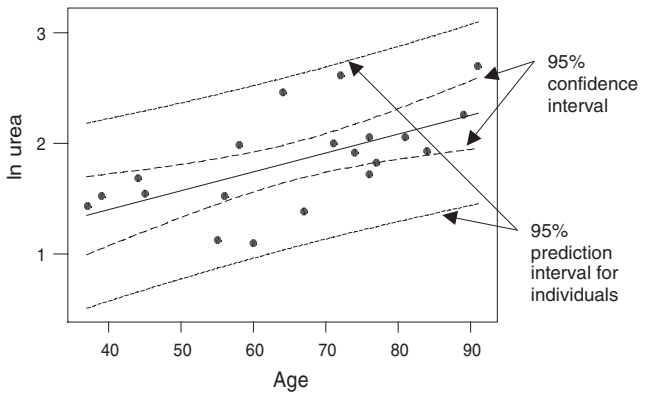
For example, the 95% prediction interval for the ln urea for a patient aged 60 years is 0.97 to 2.52 units. This transforms to urea values of 2.64 to 12.43 mmol/l.

Both confidence intervals and prediction intervals become wider for values of the predictor variable further from the mean.

Assumptions and limitations

The use of correlation and regression depends on some underlying assumptions. The observations are assumed to be independent. For correlation both variables should be random variables, but for regression only the response variable y must be random. In carrying out hypothesis tests or calculating confidence intervals for the regression parameters, the response variable should have a Normal distribution and the variability of y should be the same for each value of the predictor variable. The same assumptions are needed in testing the null hypothesis that the correlation is 0, but in order to interpret confidence intervals for the correlation coefficient both variables must be Normally distributed. Both correlation and regression assume that the relationship between the two variables is linear.

Figure 10



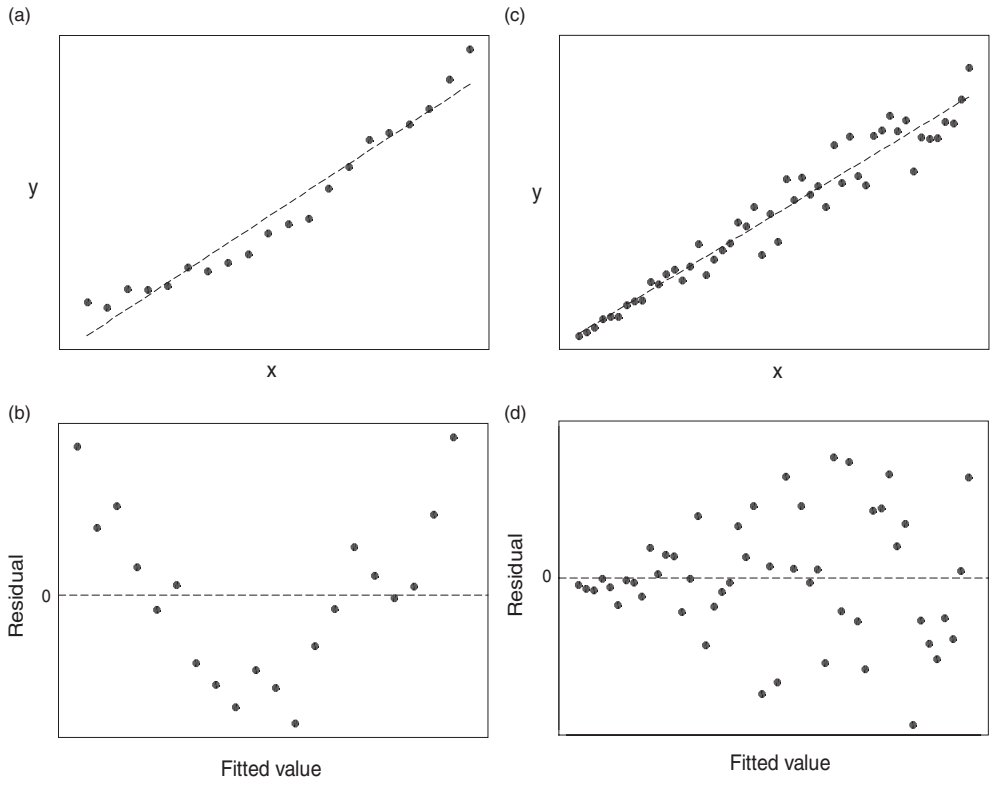
Regression line, its 95% confidence interval and the 95% prediction interval for individual patients.

A scatter diagram of the data provides an initial check of the assumptions for regression. The assumptions can be assessed in more detail by looking at plots of the residuals [4,7]. Commonly, the residuals are plotted against the fitted values. If the relationship is linear and the variability constant, then the residuals should be evenly scattered around 0 along the range of fitted values (Fig. 11).

In addition, a Normal plot of residuals can be produced. This is a plot of the residuals against the values they would be expected to take if they came from a standard Normal distribution (Normal scores). If the residuals are Normally distributed, then this plot will show a straight line. (A standard Normal distribution is a Normal distribution with mean = 0 and standard deviation = 1.) Normal plots are usually available in statistical packages.

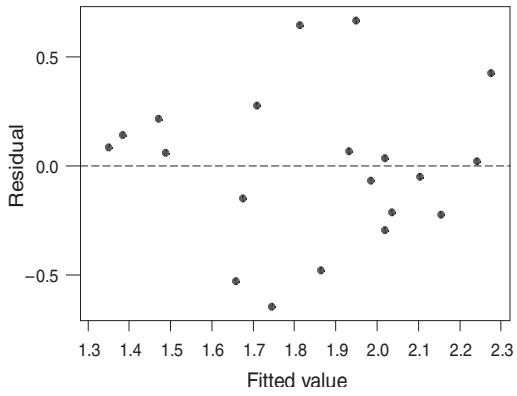
Figs 12 and 13 show the residual plots for the A&E data. The plot of fitted values against residuals suggests that the assumptions of linearity and constant variance are satisfied. The Normal plot suggests that the distribution of the residuals is Normal.

Figure 11



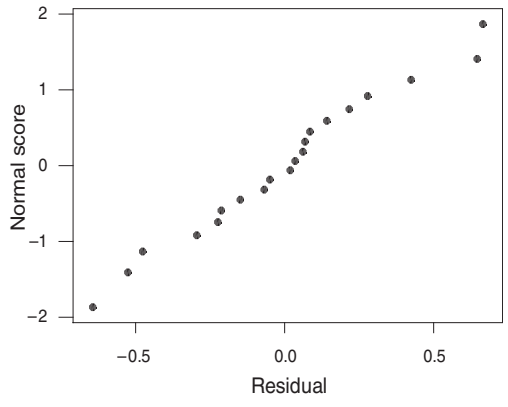
(a) Scatter diagram of y against x suggests that the relationship is nonlinear. **(b)** Plot of residuals against fitted values in panel a; the curvature of the relationship is shown more clearly. **(c)** Scatter diagram of y against x suggests that the variability in y increases with x. **(d)** Plot of residuals against fitted values for panel c; the increasing variability in y with x is shown more clearly.

Figure 12



Plot of residuals against fitted values for the accident and emergency unit data.

Figure 13



Normal plot of residuals for the accident and emergency unit data.

When using a regression equation for prediction, errors in prediction may not be just random but also be due to inadequacies in the model. In particular, extrapolating beyond the range of the data is very risky.

A phenomenon to be aware of that may arise with repeated measurements on individuals is regression to the mean. For example, if repeat measures of blood pressure are taken, then patients with higher than average values on their first reading

will tend to have lower readings on their second measurement. Therefore, the difference between their second and first measurements will tend to be negative. The converse is true for patients with lower than average readings on their first measurement, resulting in an apparent rise in blood pressure. This could lead to misleading interpretations, for example that there may be an apparent negative correlation between change in blood pressure and initial blood pressure.

Conclusion

Both correlation and simple linear regression can be used to examine the presence of a linear relationship between two variables providing certain assumptions about the data are satisfied. The results of the analysis, however, need to be interpreted with care, particularly when looking for a causal relationship or when using the regression equation for prediction. Multiple and logistic regression will be the subject of future reviews.

Competing interests

None declared.

References

1. Whitley E, Ball J: **Statistics review 1: Presenting and summarising data.** *Crit Care* 2002, **6:**66-71.
2. Kirkwood BR, Sterne JAC: *Essential Medical Statistics*, 2nd ed. Oxford: Blackwell Science; 2003.
3. Whitley E, Ball J: **Statistics review 2: Samples and populations.** *Crit Care* 2002, **6:**143-148.
4. Bland M: *An Introduction to Medical Statistics*, 3rd ed. Oxford: Oxford University Press; 2001.
5. Bland M, Altman DG: **Statistical methods for assessing agreement between two methods of clinical measurement.** *Lancet* 1986, **i:**307-310.
6. Zar JH: *Biostatistical Analysis*, 4th ed. New Jersey, USA: Prentice Hall; 1999.
7. Altman DG: *Practical Statistics for Medical Research*. London: Chapman & Hall; 1991.

This article is the seventh in an ongoing, educational review series on medical statistics in critical care.

Previous articles have covered 'presenting and summarizing data', 'samples and populations', 'hypotheses testing and *P* values', 'sample size calculations', 'comparison of means' and 'nonparametric means'.

Future topics to be covered include:

Introduction to correlation and regression
Chi-squared and Fishers exact tests
Analysis of variance
Further non-parametric tests: Kruskal–Wallis and Friedman
Measures of disease: PR/OR
Survival data: Kaplan–Meier curves and log rank tests
ROC curves
Multiple logistic regression.

If there is a medical statistics topic you would like explained, contact us at editorial@ccforum.com.