CHAPTER EIGHT

CORRELATION AND REGRESSION

Correlation and regression are statistical methods that are commonly used in the medical literature to compare two or more variables. Although frequently confused, they are quite different. **Correlation** measures the association between two variables and quantitates the strength of their relationship. Correlation evaluates only the existing data. **Regression** uses the existing data to define a mathematical equation which can be used to predict the value of one variable based on the value of one or more other variables and can therefore be used to extrapolate between the existing data. The regression equation can therefore be used to predict the outcome of observations not previously seen or tested.

CORRELATION

Correlation provides a numerical measure of the linear or "straight-line" relationship between two continuous variables X and Y. The resulting **correlation coefficient** or "**r value**" is more formally known as the **Pearson product moment correlation coefficient** after the mathematician who first described it. X is known as the **independent** or **explanatory variable** while Y is known as the **dependent** or **response variable**. A significant advantage of the correlation coefficient is that it does not depend on the units of X and Y and can therefore be used to compare any two variables regardless of their units.

An essential first step in calculating a correlation coefficient is to plot the observations in a "scattergram" or "scatter plot" to visually evaluate the data for a potential relationship or the presence of outlying values. It is frequently possible to visualize a smooth curve through the data and thereby identify the type of relationship present. The independent variable is usually plotted on the X-axis while the dependent variable is plotted on the Y-axis. A "perfect" correlation between X and Y (Figure 8-1a) has an r value of 1 (or -1). As X changes, Y increases (or decreases) by the same amount as X, and we would conclude that X is responsible for 100% of the change in Y. If X and Y are not related at all (i.e., no correlation) (Figure 8-1b), their r value is 0, and we would conclude that X is responsible for none of the change in Y.

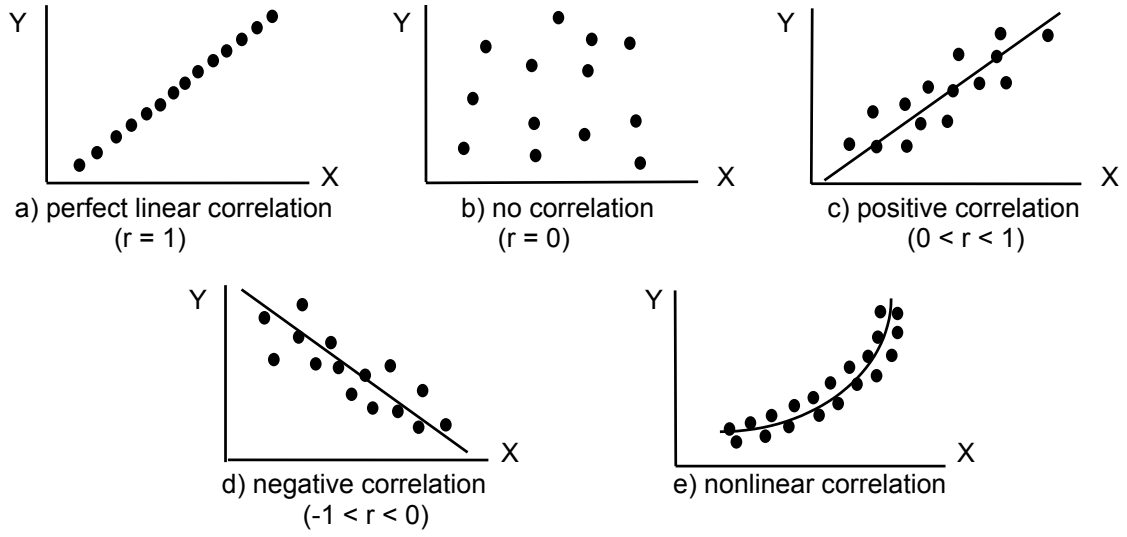


Figure 8-1: Types of Correlations

If the data points assume an oval pattern, the r value is somewhere between 0 and 1, and a moderate relationship is said to exist. A **positive** correlation (Figure 8-1c) occurs when the dependent variable increases as the independent variable increases. A **negative** correlation (Figure 8-1d) occurs when the dependent variable increases as the independent variable decreases or vice versa. If a scattergram of the data is not visualized before the r value is calculated, a significant, but **nonlinear** correlation (Figure 8-1e) may be missed. Because correlation evaluates the linear relationship between two variables, data which

assume a nonlinear or curved association will have a falsely low r value and are better evaluated using a nonlinear correlation method.

Perfect correlations (r value = 1 or -1) are rare, especially in medicine where physiologic changes are due to multiple interdependent variables as well as inherent random biologic variation. Further, the presence of a correlation between two variables does not necessarily mean that a change in one variable necessarily causes the change in the other variable. Correlation does not necessarily imply causation.

The square of the r value, known as the **coefficient of determination** or $\mathbf{r^2}$, describes the proportion of change in the dependent variable Y which is said to be explained by a change in the independent variable X. If two variables have an r value of 0.40, for example, the coefficient of determination is 0.16 and we state that only 16% of the change in Y can be explained by a change in X. The larger the correlation coefficient, the larger the coefficient of determination, and the more influence changes in the independent variable have on the dependent variable.

The calculation of the correlation coefficient is mathematically complex, but readily performed by most computer statistics programs. Correlation utilizes the t distribution to test the null hypothesis that there is no relationship between the two variables (i.e., r = 0). As with any t-test, correlation assumes that the two variables are normally distributed. If one or both of the variables is skewed in one direction or another, the resulting correlation coefficient may not be representative of the data and the result of the t test will be invalid. If the scattergram of the data does not assume some form of elliptical pattern, one or both of the variables is probably skewed (as in Figure 8-1e). The problem of non-normally distributed variables can be overcome by either transforming the data to a normal distribution or using a non-parametric method to calculate the correlation on the ranks of the data (see below).

As with other statistical methods, such as the mean and standard deviation, the presence of a single outlying value can markedly influence the resulting r value, making it appear artificially high. This can lead to erroneous conclusions and emphasizes the importance of viewing a scattergram of the raw data before calculating the correlation coefficient. Figure 8-2 illustrates the correlation between right ventricular end-diastolic volume index (RVEDVI) (the dependent variable), and cardiac index (the independent variable). The correlation coefficient for all data points is 0.72 with the data closely fitting a straight line (solid line). From this, we would conclude that 52% (r^2 = 0.52) of the change in RVEDVI can be explained by a change in cardiac index. There is, however, a single outlying data point on this scattergram, and it has a significant impact on the correlation coefficient. If this point is excluded from the data analysis, the correlation coefficient for the same data is 0.50 (dotted line) and the coefficient of determination (r^2) is only 0.25. Thus, by excluding the one outlying value (which could easily be a data error), we see a 50% decrease in the calculated relationship between RVEDVI and cardiac index. Outlying values can therefore have a significant impact on the correlation coefficient and its interpretation and their presence should always be noted by reviewing the raw data.

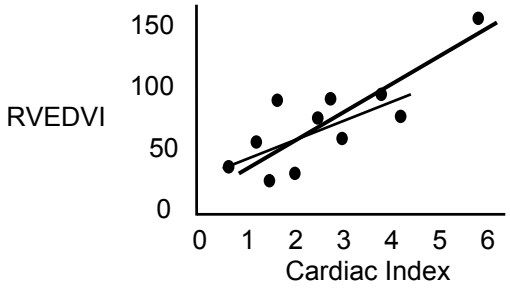


Figure 8-2: Effect of outlying values on correlation

FISHER'S Z TRANSFORMATION

A t test is used to determine whether a significant correlation is present by either accepting or rejecting the null hypothesis (r = 0). When a correlation is found to exist between two variables (i.e., the null hypothesis is rejected), we frequently wish to quantitate the degree of association present. That is, how significant is the relationship? **Fisher's z transformation** provides a method by which to determine whether a correlation coefficient is significantly different from a minimally acceptable value (such as an r value of 0.50). It can also be used to test whether two correlation coefficients are significantly different from each other.

For example, suppose we wish to compare cardiac index (CI) with RVEDVI and pulmonary artery occlusion pressure (PAOP) in 100 patients to determine whether changes in RVEDVI or PAOP correlate better with changes in CI. Assume the calculated correlation coefficient between CI and RVEDVI is 0.60, and that between CI and PAOP is 0.40. An r value of 0.60 is clearly better than an r value of 0.40, but is this difference significant? We can use Fisher's z transformation to answer this question.

The CI, RVEDVI, and PAOP data that were used to calculate the correlation coefficients all have different means and standard deviations and are measured on different scales. Thus, before we can compare these correlation coefficients, we must first transform them to the **standard normal z distribution** (such that they both have a mean of 0 and standard deviation of 1). This can be accomplished using the following formula or by using a z transformation table (available in most statistics textbooks):

$$z(r) = 0.5 \cdot \ln \frac{1+r}{1-r}$$

where r = the correlation coefficient and z(r) = the correlation coefficient transformed to the normal distribution

After transforming the correlation coefficients to the normal (z) distribution, the following equation is used to calculate a critical z value, which quantitates the significance of the difference between the two correlation coefficients (the significance of the critical value can be determined in a normal distribution table):

$$z = \frac{z(r_1) - z(r_2)}{\sqrt{1/(n-3)}}$$

If the number of observations (n) is different for each r value, the equation takes the form:

$$z = \frac{z(r_1) - z(r_2)}{\sqrt{1/(n_1-3) + 1/(n_2-3)}}$$

Using these equations for the above example, (where $r_{(CI\ vs\ RVEDVI)}$ =0.60 and $r_{(CI\ vs\ PAOP)}$ =0.40), $z_{(CI\ vs\ RVEDVI)}$ =0.693 and $z_{(CI\ vs\ PAOP)}$ =0.424. The critical value of z which determines whether 0.60 is different from 0.40 is therefore:

$$z = \frac{0.693 - 0.424}{\sqrt{1/(100-3)}} = \frac{0.269}{0.1} = 2.64$$

From a normal distribution table (found in any statistics textbook), a critical value of 2.64 is associated with a significance level or p value of 0.008. Using a p value of < 0.05 as being significant, we can state that the correlation between CI and RVEDVI is statistically greater than that between CI and PAOP.

Confidence intervals can be calculated for correlation coefficients using Fisher's z transformation. The transformed correlation coefficient, z(r), as calculated above, is used to derive the confidence interval. In order to obtain the confidence interval in terms of the original correlation coefficient, however, the interval must then be transformed back. For example, to calculate the 95% confidence interval for the correlation between CI and RVEDVI (r=0.60, z(r)=0.693), we use a modification of the standard confidence interval equation:

$$z(r) \pm 1.96 \times \sqrt{1/(n-3)}$$

where z(r) = the transformed correlation coefficient, and 1.96 = the critical value of z for a significance level of 0.05

Substituting for z(r) and n:

$$0.693 \pm (1.96)(0.1)$$

$$0.693 \pm 0.196$$

0.497 to 0.889

Converting the transformed correlation coefficients back results in a 95% confidence interval of 0.46 to 0.71. As the $r_{(CI\ vs\ PAOP)}$ of 0.40 resides outside these confidence limits, we confirm our conclusion that a correlation coefficient of 0.60 is statistically different from one of 0.40 in this patient population.

CORRELATION FOR NON-NORMALLY DISTRIBUTED DATA

As discussed above, situations arise in which we wish to perform a correlation, but one or both of the variables is non-normally distributed or there are outlying observations. We can transform the data to a normal distribution using a logarithmic transformation, but the correlation we then calculate will actually be the correlation between the logarithms of the observations and not that of the observations themselves. Any conclusions we then make will be based on the transformed data and not the original data.

Another method is to perform the correlation on the ranks of the data using **Kendall's (tau)** or **Spearman's (rho) rank correlation**. Both are non-parametric methods in which the data is first ordered from smallest to largest and then ranked. The correlation coefficient is then calculated using the ranks. Rank correlations can also be used in the situation where one wishes to compare ordinal or discrete variables (such as live versus die, disease versus no disease) with continuous variables (such as cardiac output). These rank correlation methods are readily available on most computer statistics packages. The traditional Pearson correlation coefficient is a stronger statistical test, however, and should be used if the data is normally distributed.

REGRESSION

Regression analysis mathematically describes the dependence of the Y variable on the X variable and constructs an equation which can be used to predict any value of Y for any value of X. It is more specific and provides more information than does correlation. Unlike correlation, however, regression is not scale independent and the derived regression equation depends on the units of each variable involved. As with correlation, regression assumes that each of the variables is normally distributed with equal variance. In addition to deriving the regression equation, regression analysis also draws a **line of best fit** through the data points of the scattergram. These "regression lines" may be linear, in which case the relationship between the variables fits a straight line, or nonlinear, in which case a polynomial equation is used to describe the relationship.

Regression (also known as **simple regression**, **linear regression, or least squares regression**) fits a straight line equation of the following form to the data:

$$Y = a + bX$$

where Y is the dependent variable, X is the single independent variable, a is the Y-intercept of the regression line, and b is the slope of the line (also known as the **regression coefficient**).

Once the equation has been derived, it can be used to predict the change in Y for any change in X. It can therefore be used to extrapolate between the existing data points as well as predict results which have not been previously observed or tested.

A t test is utilized to ascertain whether there is a significant relationship between X and Y, as in correlation, by testing whether the regression coefficient, b, is different from the null hypothesis of zero (no relationship). If the correlation coefficient, r, is known, the regression coefficient can be derived as follows:

$$b = r \frac{\text{standard deviation}_Y}{\text{standard deviation}_X}$$

The regression line is fitted using a method known as "**least squares**" which minimizes the sum of the squared vertical distances between the actual and predicted values of Y. Along with the regression equation, slope, and intercept, regression analysis provides another useful statistic: the **standard error of the slope**. Just as the standard error of the mean is an estimate of how closely the sample mean approximates the population mean, the standard error of the slope is an estimate of how closely the measured slope approximates the true slope. It is a measure of the "goodness of fit" of the regression line to the data and is calculated using the standard deviation of the **residuals**. A residual is the vertical distance of each data point from the least squares fitted line (Figure 8-3).

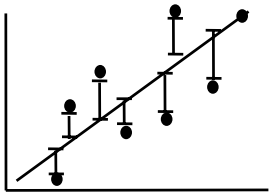


Figure 8-3: Residuals with least squares fit regression line

Residuals represent the difference between the observed value of Y and that which is predicted by X using the regression equation. If the regression line fits the data well, the residuals will be small. Large residuals may point to the presence of outlying data which, as in correlation, can significantly affect the validity of the regression equation.

The steps in calculating a regression equation are similar to those for calculating a correlation coefficient. First, a scattergram is plotted to determine whether the data assumes a linear or nonlinear pattern. If outliers are present, the need for nonlinear regression, transformation of the data, or non-parametric methods should be considered. Assuming the data are normally distributed, the regression equation is calculated. The residuals are then checked to confirm that the regression line fits the data well. If the residuals are high, the possibility of non-normally distributed data should be reconsidered.

When reporting the results of a regression analysis, one should report not only the regression equation, regression coefficients, and their significance levels, but also the standard deviation or variance of each regression coefficient and the variance of the residuals. A common practice is to "standardize" the regression coefficients by converting them to the standard normal (z) distribution. This allows regression coefficients calculated on different scales to be compared with one another such that conclusions can be made independent of the units involved. **Confidence bands** (similar to confidence intervals) can also be calculated and plotted along either side of the regression line to demonstrate the potential variability in the line based on the standard error of the slope.

MULTIPLE LINEAR REGRESSION

Multiple linear regression is used when there is more than one independent variable to explain changes in the dependent variable. The equation for multiple linear regression takes the form:

$$Y = a + b_1X_1 + b_2X_2 + ... + b_nX_n$$

where a = the Y intercept, and b_1 through b_n are the regression coefficients for each of the independent variables X_1 through X_n.

The dependent variable Y is thus a "weighted average" based on the strength of the regression coefficient, b, of each independent variable X. Once the multiple regression analysis is completed, the regression coefficients can be ordered from smallest to largest to determine which independent variable contributes the most to changes in Y. Multiple linear regression is thus most appropriate for continuous variables where we wish to identify which variable or variables is most responsible for changes in the dependent variable.

A computerized multiple regression analysis results in the regression coefficients (or "regression weights"), the standard error of each regression coefficient, the intercept of the regression line, the variance of the residuals, and a statistic known as the **coefficient of multiple determination** or "**multiple R**" which is analogous to the Pearson product moment correlation coefficient "r". It is utilized in the same manner as the coefficient of determination (r^2) to measure the proportion of change in Y which can be attributed to changes in the dependent variables ($X_1...X_n$). The statistical test of significance for R is the F distribution instead of the t distribution as in correlation.

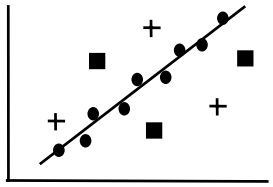
Multiple regression analysis is readily performed by computer and can be performed in several ways. **Forward selection** begins with one independent (X) variable in the equation and sequentially adds additional variables one at a time until all statistically significant variables are included. **Backward selection** begins with all independent variables in the equation and sequentially removes variables until only the statistically

significant variables remain. **Stepwise regression** uses both forward and backward selection procedures together to determine the significant variables. **Logistic regression** is used when the independent variables include numerical and/or nominal values and the dependent or outcome variable is dichotomous, having only two values. An example of a logistic regression analysis would be assessing the effect of cardiac output, heart rate, and urinary output (all continuous variables) on patient survival (i.e., live versus die; a dichotomous variable).

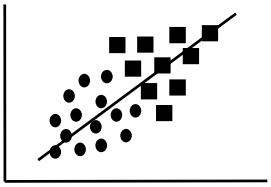
POTENTIAL ERRORS IN CORRELATION AND REGRESSION

Two special situations may result in erroneous correlation or regression results. First, multiple observations on the same patient should not be treated as independent observations. The sample size of the data set will appear to be increased when this is really not the case. Further, the multiple observations will already be correlated to some degree as they arise from the same patient. This results in an artificially increased correlation coefficient (Figure 8-4a). Ideally, a single observation should be obtained from each patient. In studies where repetitive measurements are essential to the study design, equal numbers of observations should be obtained from each patient to minimize this form of statistical error.

Second, the mixing of two different populations of patients should be avoided. Such grouped observations may appear to be significantly correlated when the separate groups by themselves are not (Figure 8-4b). This can also result in artificially increased correlation coefficients and erroneous conclusions.



a) Multiple observations on a single patient (solid dots) leading to an artificially high correlation.



b) Increased correlation with groups of patients which, by themselves, are not correlated.

Figure 8-4: Potential errors in the use of regression

SUGGESTED READING

1. O'Brien PC, Shampo MA. Statistics for clinicians: 7. Regression. Mayo Clin Proc 1981;56:452-454.
2. Godfrey K. Simple linear regression. NEJM 1985;313:1629-1636.
3. Dawson-Saunders B, Trapp RG. Basic and clinical biostatistics (2nd Ed). Norwalk, CT: Appleton and Lange, 1994: 52-54, 162-187.
4. Wassertheil-Smoller S. Biostatistics and epidemiology: a primer for health professionals. New York: Springer-Verlag, 1990:53-59.
5. Altman DG. Statistics and ethics in medical research. V -analyzing data. Brit Med J 1980; 281:1473-1475.
6. Conover WJ, Iman RL. Rank transformations as a bridge between parametric and nonparametric statistics. Am Stat 1981; 35:124-129.