Correlation and Regression

Sybil L. Crawford

Circulation 2006;114;2083-2088
DOI: 10.1161/CIRCULATIONAHA.105.586495
Circulation is published by the American Heart Association. 7272 Greenville Avenue, Dallas, TX 72514
 Print ISSN: 0009-7322. Online ISSN: 1524-4539

Correlation and Regression

Sybil L. Crawford, PhD

In many health-related studies, investigators wish to assess the strength of an association between 2 measured (continuous) variables. For example, the relation between high-sensitivity C-reactive protein (hs-CRP) and body mass index (BMI) may be of interest. Although BMI is often treated as a categorical variable, eg, underweight, normal, overweight, and obese, a noncategorized version is more detailed and thus may be more informative in terms of detecting associations. Correlation and regression are 2 relevant (and related) widely used approaches for determining the strength of an association between 2 variables. Correlation provides a unitless measure of association (usually linear), whereas regression provides a means of predicting one variable (dependent variable) from the other (predictor variable). This report summarizes correlation coefficients and least-squares regression, including intercept and slope coefficients.

Correlation

Correlation provides a "unitless" measure of association between 2 variables, ranging from −1 (indicating perfect negative association) to 0 (no association) to +1 (perfect positive association). Both variables are treated equally in that neither is considered to be a predictor or an outcome.

Pearson Product-Moment Coefficient of Correlation

The most commonly used version is the Pearson product-moment coefficient of correlation, *r*. Suppose one wants to estimate the correlation between X=BMI, denoted for the i^{th} subject as X_i, and Y=hs-CRP, denoted for the i^{th} subject as Y_i. This is estimated for a sample of size n (i=1, . . . , n) using the following formula[1]:

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

where

$$SS_{xy} = \sum_i (X_i - \bar{X})(Y_i - \bar{Y}),\ SS_{xx} = \sum_i (X_i - \bar{X})^2,$$

and

$$SS_{yy} = \sum_i (Y_i - \bar{Y}).^2$$

Here, $\bar{X}$ indicates the sample mean of X (=BMI), and $\bar{Y}$ the sample mean of Y (=hs-CRP). The numerator of *r* reflects how BMI and hs-CRP co-vary, and the denominator reflects the variability of both BMI and hs-CRP about their respective sample means.

Alternative Correlation Coefficients

The Pearson correlation coefficient assumes that X and Y are jointly distributed as bivariate normal, ie, X and Y each are normally distributed, and that they are linearly related.[2] When these assumptions are not satisfied, nonparametric versions can be used to estimate correlation. These include the Spearman rank correlation coefficient,[2] which is based on a comparison of the ranks of X and Y rather than on the original variables themselves. By using ranks, nonparametric approaches are robust to departures from the assumptions of the Pearson correlation coefficient, as well as to outlying (atypical) observations that may distort the estimated Pearson correlation coefficient. On the other hand, if the assumptions for the Pearson correlation coefficient are met, the nonparametric versions are less efficient. That is, they are less likely to detect an association than the Pearson correlation coefficient. Thus, an alternative to nonparametric correlations is to transform X or Y (or both) to better meet these assumptions. See Erickson and Nosanchuk[3] for a discussion of transformations.

As an example, consider hs-CRP and BMI in Figures 1 and 2. Figure 1A suggests that there is a positive but nonlinear association between hs-CRP and BMI, and Figures 2A and 2B indicate that neither hs-CRP nor BMI is normally distributed; thus, the assumptions for the Pearson correlation coefficient are not met. Consequently, the Spearman rank correlation provides a more appropriate estimate of association. When a natural log transformation is applied to both hs-CRP and BMI to pull in the long right tails, Figure 1B shows a linear association between the log-transformed variables, and Figures 2C and 2D suggest that the log transformation has made each variable's distribution closer to normal. The estimated Pearson correlation of the log-transformed variables is more than one third higher than the corresponding estimate for hs-CRP and BMI, which reflects the greater linearity seen in the scatterplot. Note, however, that the Spearman correlation is identical for the original and trans-

From the University of Massachusetts Medical School, Worcester, Mass.
Correspondence to Sybil L. Crawford, PhD, Preventive and Behavioral Medicine, University of Massachusetts Medical School, 55 Lake Ave N, Shaw Bldg, Room 228, Worcester, MA 01655. E-mail Sybil.Crawford@umassmed.edu
(*Circulation.* 2006;114:2083-2088.)

Circulation is available at http://www.circulationaha.org DOI: 10.1161/CIRCULATIONAHA.105.586495

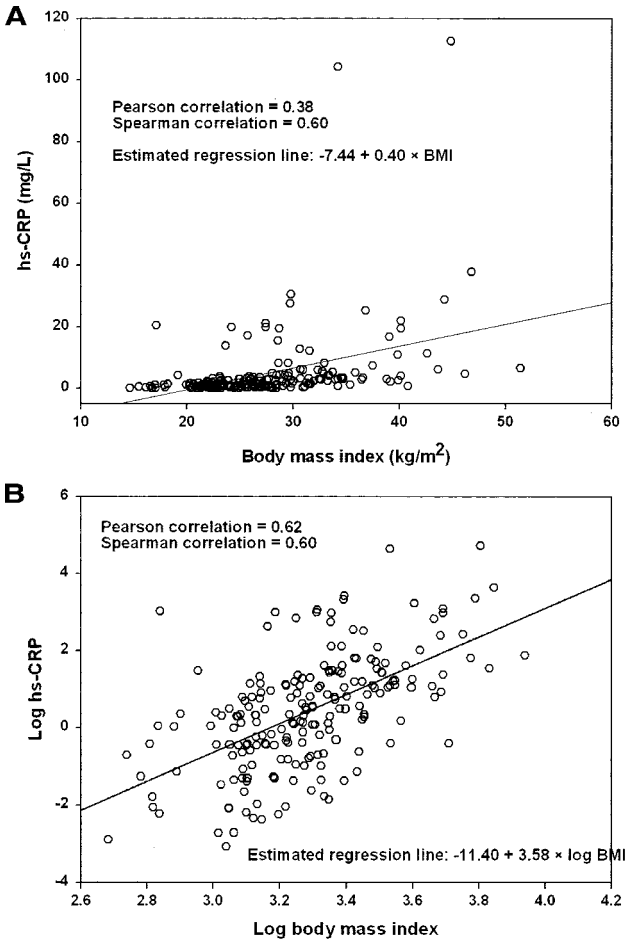

Figure 1. A, Scatterplot of hs-CRP vs BMI, with least-squares linear regression line. B, Scatterplot of natural log-transformed hs-CRP vs natural log-transformed BMI, with least-squares linear regression line.

formed variables, because the log transformation does not change the variables' ranks.

Regression

Regression also indicates whether 2 variables are associated. In contrast to correlation, however, regression considers one variable to be an outcome (dependent variable) and the other to be a predictor variable. As an example, suppose one wants to predict hs-CRP on the basis of BMI. hs-CRP can be modeled as a linear function of BMI, as in Figure 1A:

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

where β_0 is the intercept, β_1 is the slope coefficient for X=BMI, and $e_i = Y_i - (\beta_0 + \beta_1 X_i)$ denotes the residual or error, the part of Y_i that is not explained by the linear function of X_i, $\beta_0 + \beta_1 X_i$. The slope coefficient β_1 indicates the difference in Y that corresponds to a 1-unit difference in X. When X is defined in terms of clinically meaningful units, such as age in years, it facilitates the interpretation of β_1. The above approach assumes a linear association between X and Y. Consequently, it is important to check this assumption, eg, with a scatterplot of Y versus X, before one estimates the intercept and slope; a transformation of X or Y (or both) may be needed, as in the preceding hs-CRP and BMI example.

Least-Squares Estimation

As with correlation, there are different approaches to estimation of a regression line. The most commonly used technique is the method of least squares (sometimes referred to as ordinary least squares to distinguish it from weighted least squares, which is used when observations have different weights from complex sampling designs), which minimizes the sum of the squared residuals or errors (SSE). That is, estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1, respectively, are chosen to minimize

$$SSE = \sum_i [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)]^2 = \sum_i \hat{e}_i^2.$$

The resulting formulas are

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

and

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

The intercept β_0 generally is not of intrinsic interest but is included to estimate β_1 accurately. Note that if X has been centered so that $\bar{X}=0$, then $\hat{\beta}_0 = \bar{Y}$. The numerator for the estimated slope coefficient is identical to the numerator of the estimated Pearson correlation coefficient r; in particular, when r equals 0, $\hat{\beta}_1$ also equals 0. $\hat{\beta}_1$ can be reexpressed as

$$r \times \sqrt{SS_{yy}/SS_{xx}}.$$

Thus, both r and $\hat{\beta}_1$ estimate the linear association between X and Y. Unlike r, however, $\hat{\beta}_1$ is not unitless but reflects the scales of X and Y.

Coefficient of Determination

A unitless estimate of the strength of the linear association between Y and X is given by the coefficient of determination, also known as R^2. R^2 is the proportion of variance in the outcome Y accounted for by the linear function of the predictor X, ie, the fitted value $= \hat{\beta}_0 + \hat{\beta}_1 X$, and is estimated as $(SS_{yy} - SSE)/SS_{yy} = 1 - (SSE/SS_{yy})$. SSE is the amount of variability in the outcome Y that is "left over," ie, not explained by the linear function of the predictor X. Note that the estimated Pearson correlation coefficient equals the square root of R^2; R^2 ranges from 0 (no linear association) to 1 (perfect linear association, whether positive or negative). A related quantity is the residual mean square $\hat{\sigma}^2$, the variance of the residuals, or equivalently, the variability of Y about the estimated regression line. For a regression with a single predictor variable, this is computed as $SSE/(n-2)$.[1] For a given data set, the smaller $\hat{\sigma}^2$ is, the larger R^2 is; $\hat{\sigma}^2$ is not unitless, however, but varies with the scale of the observed data.

Checking Assumptions: Regression Diagnostics

The above formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$ can be used to estimate a regression line regardless of the distributions of X and Y.

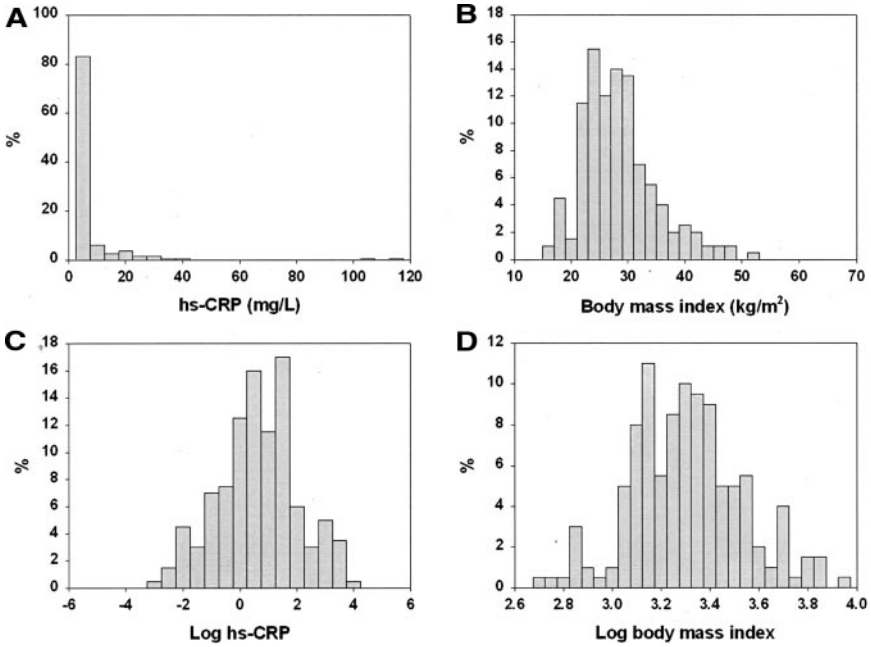

Figure 2. A, Histogram of hs-CRP. B, Histogram of BMI. C, Histogram of natural log-transformed hs-CRP. D, Histogram of natural log-transformed BMI.

Assumptions required for inferences with regard to the coefficients and estimation or prediction from the regression line, however, include the following: (1) normally distributed residuals with a mean of zero; (2) constant variance of the residuals; and (3) independence of residuals from different observations.

These assumptions should be checked before any inferences are made from the estimated regression line. For example, to assess whether residuals are normally distributed, a statistical test (eg, the Kolmogorov-Smirnov χ^2 test[2]) can be done to compare the estimated distribution to a normal distribution. Related graphical checks include a histogram of the estimated residuals and a normal probability plot, also known as a quantile-quantile plot, of the observed residual quantiles versus quantiles that would be expected under a normal distribution[4]; the latter plot will approximate a straight line if the assumption of normality is met. Also, a scatterplot of the estimated residuals versus the fitted values should have a "cloud" pattern, which indicates no increase or decrease in the variability of the residuals as X increases (ie, constant variance), and no curvilinear pattern that suggests a nonlinear association of X and Y.[5] In addition, influential observations can be detected with diagnostic tools available in most statistical software packages, such as Cook's distance,[4,6] which indicates for each observation how much the estimated regression coefficients would change if that observation were omitted and the regression coefficients reestimated; a value of at least 1 indicates a highly influential observation. Although an influential observation often will have a large, outlying residual, this is not guaranteed to occur, because an extremely influential observation may "pull" the regression line toward itself and hence have a relatively small residual. A more detailed discussion of leverage and influence is beyond the scope of this report.

Continuing the previous hs-CRP and BMI example, the estimated regression line for hs-CRP as a linear function of BMI is hs-CRP$=-7.44+0.40\times$BMI, with an R^2 value of 0.20. Residuals from the regression of hs-CRP on BMI, seen in Figure 3A, are not normally distributed and exhibit a large, positive outlier. The scatterplot of residuals versus fitted values (Figure 4A) demonstrates increasing variability in the residuals with larger fitted values. The Kolmogorov-Smirnov χ^2 statistic is statistically significant ($P<0.01$), which indicates a departure of the estimated residual distribution from normality. Moreover, one observation has a Cook's distance >1, which indicates high influence on the estimated regression line.

The corresponding estimated line from regressing log hs-CRP on log BMI is log hs-CRP$=-11.40+3.58\times$log BMI, with an R^2 value of 0.37. The proportion of variance explained almost doubles when the variables are transformed, which reflects the improvement in linearity. The histogram of the residuals from the regression of log hs-CRP on log BMI, seen in Figure 3B, is closer to bell-shaped and has no outliers, and there is no significant departure from normality (the probability value for the corresponding Kolmogorov-Smirnov χ^2 statistic$=0.12$). The scatterplot of residuals versus fitted values (Figure 4B) indicates constant variance of the residuals across the range of fitted values. In addition, none of the observations have a Cook's distance of at least 1. Note that the scales on the *y* axis, which indicate the scales of the 2 sets of residuals, are not comparable because the original data are on different scales.

As seen in this example, transforming either the outcome or the predictor (or both) often solves one or more problems, including nonlinear associations, outlying values, and nonconstant variance of residuals. Nonlinear associations also may be modeled with polynomial regression, expanding the right-hand side of the equation to include terms for X^2, X^3, and so on.[7]

Estimation and Prediction

In addition to determining magnitudes of association, the estimated regression line can be used to estimate the average

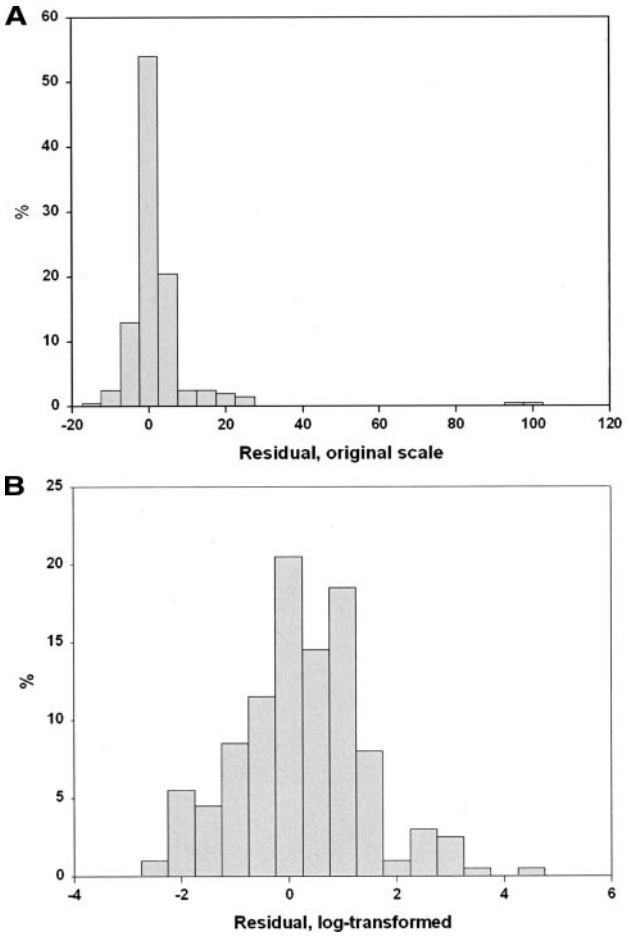

Figure 3. A, Histogram of residuals from least-squares linear regression of hs-CRP on BMI. B, Histogram of residuals from least-squares linear regression of natural log-transformed hs-CRP on natural log-transformed BMI.

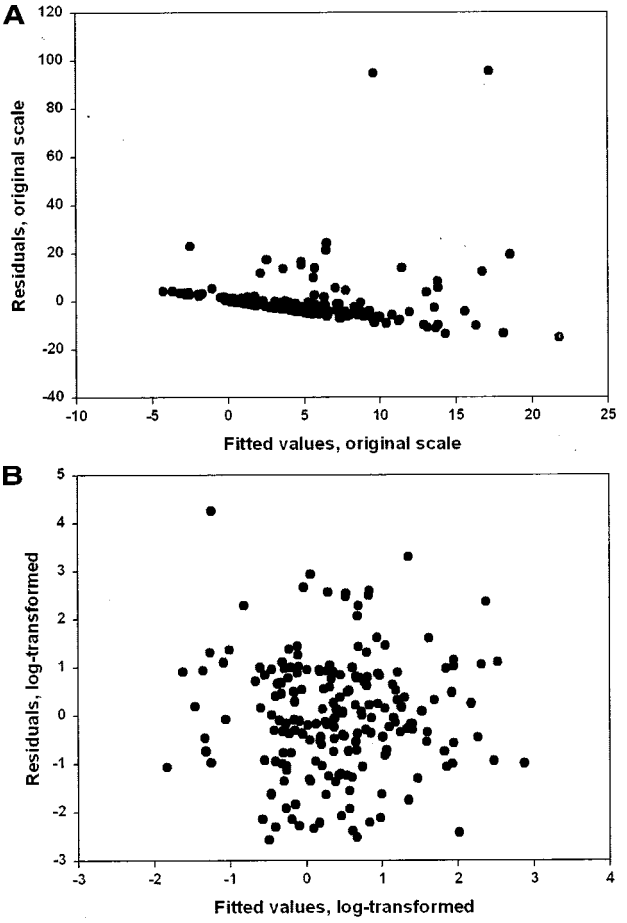

Figure 4. A, Scatterplot of residuals from least-squares linear regression of hs-CRP on BMI vs corresponding fitted values. B, Scatterplot of residuals from least-squares linear regression of natural log-transformed hs-CRP on natural log-transformed BMI vs corresponding fitted values.

Y at a specified value of X. In the preceding example, we can estimate the average (mean) hs-CRP concentration at, say, BMI=25 kg/m^2. Using the estimated regression line on the untransformed variables, this would be estimated as $-7.44+0.40\times25=2.56$ mg/L. In addition, we can predict the hs-CRP concentration for an individual patient with a BMI of 25 kg/m^2, also given by 2.56 mg/L. The corresponding estimate on the log hs-CRP scale is $-11.40+3.58\times\log(25)=0.12$.

The difference between estimation of an average and prediction for an individual subject lies in the associated variability. The estimated variance of an estimate of a mean at X=x* is given by

$$\hat{\sigma}^2\left[\frac{1}{n}+\frac{(x^*-\bar{X})^2}{SS_{xx}}\right]$$

which increases with $\hat{\sigma}$ and with the distance between x* and the observed sample mean for X.[1] That is, the estimate of the mean is less precise for larger values of the residual mean square (variability of Y about the regression line) and as the value of x* is farther from the center of the observed data. The variance for a prediction at X=x* is equal to

$$\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(x^*-\bar{X})^2}{SS_{xx}}\right],$$

which equals the variance for an estimated mean plus $\hat{\sigma}^2$.[1] Thus, predicting Y for an individual at a given X value is less precise than estimating the mean at the same X value. This can be seen graphically in Figure 5. Estimates of both the mean log hs-CRP and predicted log hs-CRP across the range of log BMI values are given by the estimated regression line (solid line). The 95% CIs for mean log hs-CRP and for predicted log hs-CRP also are presented; the CIs for predictions for an individual are much wider than those for the mean. Both CIs are wider for extreme values of log BMI than for log BMI values nearer the sample mean.

Figure 1A indicates that for values of BMI <18.6 kg/m^2, linear regression on the untransformed data produces negative estimates of hs-CRP (for the mean or for an individual patient), which are invalid for this outcome. In contrast, negative estimates of hs-CRP can be backtransformed with exponentiation, ie, the antilog, to produce estimates on the original scale, which are guaranteed to be above zero because of the nature of the antilog transformation. This suggests another possible advantage of working with transformed variables.

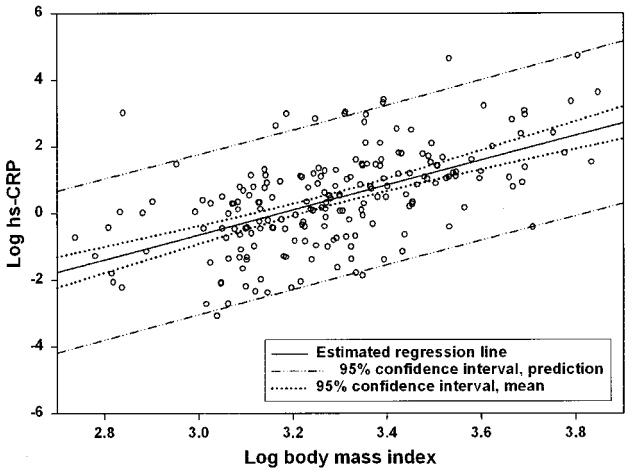

Figure 5. Scatterplot of natural log-transformed hs-CRP vs natural log-transformed BMI, with least-squares linear regression line and 95% CIs for prediction and for mean estimation.

Alternatives to Least-Squares Estimation

Ordinary least-squares regression is widely used, in part because of its ease of computation and also because it has desirable properties when the assumptions are met.[7] Because the regression line is estimated by minimizing the squared residuals, however, outlying values can exert a relatively large impact on the estimated line. With the advent of computers, alternative methods have been developed that are computationally more demanding but are more robust to outliers. Some techniques reduce the influence of outliers by replacing squared residuals with other functions of the residuals or minimizing the median of the squared residuals rather than the sum (see Rousseeuw and Leroy[8]). Other approaches are nonparametric, such as Tukey's resistant lines[3] or Theil's method.[2] It is difficult to generalize some of these approaches to the setting with multiple predictor variables, however.

Additional Considerations and Cautions

Extrapolation

Even when an estimated regression line provides a good fit to the observed data, it is important not to extrapolate beyond the range of the sample, because the estimated line may not be appropriate. For example, as seen in Figure 1A, estimates of Y from the regression line may be invalid for extreme X values. Alternatively, the relation between X and Y may become nonlinear outside the range of the sample.

Study Design and Interpretation of Estimates

Estimates of correlation and R^2 depend not only on the magnitude of the underlying true association but also on the variability of the data included in the sample (see Weisberg[4]). In the preceding hs-CRP and BMI example, the estimated Pearson correlation of log hs-CRP and log BMI in the full sample is 0.62. If we restrict the sample to the middle 2 quartiles of log BMI, thereby artificially decreasing the SD of log BMI from 0.23 to 0.08, the corresponding estimated correlation is 0.31, an underestimate. Conversely, if we include only women in the top and bottom log BMI quartiles (which yields an SD of log BMI of 0.31), the estimated correlation is 0.70, an overestimate. In the first instance, because the variation in X is constrained to be too small, the variation in Y ignoring X (ie, the horizontal spread in Figure 2 for the middle half of the data) is close to the variation in Y accounting for X, ie, the variation about the regression line. Consequently, the estimated proportion of explained variance in Y is deflated. The reverse occurs in the second instance. Thus, estimates that are not computed from a random sample from the entire range of the variables may not reflect the true correlation.

The range of the predictor variable also affects the standard error of the estimated regression slope, computed as $\hat{\sigma}/\sqrt{SS_{xx}}$, which decreases as the variability in X increases; consequently, the slope is estimated with the greatest precision if one samples X entirely at the minimum and maximum possible values.[7] Clearly, such a design is not optimal, however, for detecting departures from assumptions, eg, nonlinearity.

Categorical Versus Continuous Variables

When a variable is continuous, treating it as a continuous variable typically retains more information than collapsing it to an ordinal categorical variable.[9] In some cases, however, the latter version may be preferable. Consider the example of alcohol consumption. In some populations, there may be a large percentage with no consumption, which leads to a large "spike" at the value 0; hence, there may be no straightforward transformation that satisfies the assumptions of correlation or linear regression. Here, it may be more useful to categorize alcohol consumption as an ordinal variable, eg, zero consumption and quartiles of nonzero consumption, and to use ANOVA rather than linear regression. As another example, consider years of education. A difference of 1 year often has a different impact depending on whether the reference point is, say, 11 years compared with 13 years. In this case, a categorized ordinal variable may provide a better fit to the data. Moreover, categorized variables may be more interpretable in clinical settings.[10]

Confounding

The above discussion assumes there is only a single predictor variable of interest. The association between X and Y, however, may be due in part to the contribution of additional variables that are related to both X and Y, ie, confounding variables. For example, the estimated association between BMI and hs-CRP may be due in part to age, because both BMI and hs-CRP are themselves positively related to age. The methods summarized above can be expanded to include multiple predictors, and associations between X and Y that adjust for these confounding factors can be estimated. Returning to the hs-CRP and BMI example, a partial (age-adjusted) correlation between hs-CRP and BMI can be computed; for the Pearson correlation, this is done by regressing hs-CRP on age, regressing BMI on age, and computing the Pearson correlation of the 2 sets of residuals, ie, the component of hs-CRP that is unrelated to age and the component of BMI that is unrelated to age. Similarly, an age-adjusted slope for BMI can be estimated by adding age as a predictor to the linear regression model. A regression model

with multiple predictors is referred to as multiple regression. A later article in this series will address both partial correlation and multiple regression.

Discussion

Correlation and regression are 2 widely used approaches for determining the strength of association between 2 variables. Regression also is used for predicting an outcome from a predictor variable. Estimates are easily obtained in a variety of statistical software packages. For both methods, it is important to assess whether the assumptions are valid before one draws conclusions from the estimates. If assumptions are not satisfied, options include applying transformations to better meet the assumptions or using nonparametric versions. Both correlation and regression are easily generalized to the situation with multiple predictor variables.

Disclosures

None.

References

1. McClave JT, Dietrich II FH. *Statistics*. San Francisco, Calif: Dellen; 1985.
2. Daniel WW. *Applied Nonparametric Statistics*. 2nd ed. Boston, Mass: PWS-KENT; 1990.
3. Erickson BH, Nosanchuk TA. *Understanding Data*. 2nd ed. Toronto, Canada: University of Toronto Press; 1992.
4. Weisberg S. *Applied Linear Regression*. New York, NY: Wiley; 1980.
5. Tabachnick BG, Fidell LS. *Using Multivariate Statistics*. 3rd ed. New York, NY: Harper Collins; 1996.
6. Belsley DA, Kuh E, Welsch RE. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York, NY: Wiley; 1980.
7. Draper N, Smith H. *Applied Regression Analysis*. 2nd ed. New York, NY: Wiley; 1981.
8. Rousseeuw PJ, Leroy AM. *Robust Regression and Outlier Detection*. New York, NY: Wiley; 1987.
9. Ragland DR. Dichotomizing continuous outcome variables: dependence of the magnitude of association and statistical power on the cutpoint. *Epidemiology*. 1992;3:434–440.
10. Mazumdar M, Glassman JR. Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. *Stat Med*. 2000;19:113–132.

KEY WORDS: statistics ■ epidemiology ■ computers