Circulation
JOURNAL OF THE AMERICAN HEART ASSOCIATION

American Heart Association®
Learn and Live℠

Multiple Linear Regression: Accounting for Multiple Simultaneous Determinants of a Continuous Dependent Variable

Bryan K. Slinker and Stanton A. Glantz

Circulation 2008;117;1732-1737
DOI: 10.1161/CIRCULATIONAHA.106.654376
Circulation is published by the American Heart Association. 7272 Greenville Avenue, Dallas, TX 72514

Statistical Primer for Cardiovascular Research

Multiple Linear Regression

Accounting for Multiple Simultaneous Determinants of a Continuous Dependent Variable

Bryan K. Slinker, DVM, PhD; Stanton A. Glantz, PhD

In many cardiovascular experiments and observational studies, multiple variables are measured and then analyzed and interpreted to provide biomedical insights. When these data lend themselves to analyzing the association of a continuous dependent (or response) variable to 2 or more independent (or predictor) variables, multiple regression methods are appropriate. Multiple regression differs from ANOVA, in which the predictors are represented as "factors" with multiple discrete "levels." In this report, we focus on multiple regression to analyze data sets in which the response variable is continuous; other methods, such as logistic regression and proportional hazards regression, are useful in cases in which the response variable is discrete.[1]

Although many studies are designed to explore the simultaneous contributions of multiple predictors to an observed response, the data are often analyzed by relating each of the predictor variables, 1 at a time, to a single response variable with the use of a series of simple linear regressions. However, although 2-dimensional data plots and separate simple regressions are easy to visualize and interpret, multiple regression analysis is the preferred statistical method.[1–5] We want to reach correct conclusions not only about which predictors are important and the size of their effects but also about the structure by which multiple predictors simultaneously relate to the response. Often, we also want to know whether the multiple predictors that influence a response or outcome do so independently or whether they interact.[6] Finally, although only 1 or 2 predictors may interest us, our analysis often must adjust for other influences (ie, confounding effects).

A series of simple regressions cannot accomplish these tasks; if we want to examine the simultaneous effects of multiple predictors on a response, we must use a method that treats them accordingly. Conducting a series of simple regression analyses when multiple regression analysis is called for may lead to erroneous conclusions about the contribution of each of multiple predictor variables because this approach does not account for their simultaneous contributions. As a result, a predictor may be deemed important when it is not, or, conversely, a predictor may appear unrelated to the response when examined alone but relate strongly when considered simultaneously with other predictors.

From Simple Linear Regression to Multiple Regression

Simple linear regression involves estimating the straight line

$$\hat{Y}=b_0+b_1X \tag{1}$$

where $\hat{Y}$ is the predicted value of the response variable, Y, at a given value of the predictor variable, X. The intercept, b_0, estimates the value of the response when the predictor is 0, and the slope, b_1, estimates the average change in the response for a unit change in the predictor. The "best" estimate for this line is the one that minimizes the sum (denoted by the Greek letter Σ) of squared residuals, SS_{res}, between the observed values of Y and the values of $\hat{Y}$ predicted from Equation 1 across the corresponding values of X (thus, this is called ordinary least squares regression):

$$SS_{res}=\sum[Y-\hat{Y}]^2 \tag{2}$$

The question immediately arises whether the relationship between X and Y is "statistically significant," in other words, whether knowing the value of X allows predicting Y better than just knowing the mean value of Y or, equivalently, whether the slope of the regression line for the underlying population is different from 0. To test the "null hypothesis" that this slope is 0, we compare the magnitude of the statistic b_1 to the precision with which it is estimated, its standard error, s_{b_1}. A smaller standard error indicates higher certainty in the value of the estimate. We use this information to compute a *t* statistic

$$t_{b_1}=\frac{b_1}{s_{b_1}} \tag{3}$$

t_{b_1} will be large if b_1 is large compared with s_{b_1}. If t_{b_1} (with $n-2$ degrees of freedom; n is the number of observations) exceeds the maximum expected under the null hypothesis of no relationship between Y and X, we conclude that the slope

From the Veterinary and Comparative Anatomy, Pharmacology, and Physiology Department, Washington State University, Pullman, Wash (B.K.S.), and Center for Tobacco Control Research and Education, Cardiovascular Research Institute, Department of Medicine, University of California at San Francisco, San Francisco (S.A.G.).

Correspondence to Stanton A. Glantz, PhD, University of California, 530 Parnassus, Room 366, San Francisco, CA 94143-1390. E-mail glantz@medicine.ucsf.edu

(*Circulation*. 2008;117:1732-1737.)

Circulation is available at http://circ.ahajournals.org

DOI: 10.1161/CIRCULATIONAHA.106.654376

is significantly different from 0, meaning that knowing X contributes to our ability to predict Y.

This approach generalizes directly to multiple predictor variables. For example, the simplest multiple regression equation relates a single continuous response variable, Y, to 2 continuous predictor variables, X_1 and X_2:

$$\hat{Y}=b_0+b_1X_1+b_2X_2 \tag{4}$$

where $\hat{Y}$ is the value of the response predicted to lie on the best-fit regression plane (the multidimensional generalization of a line). The intercept, b_0, is the plane's reference position; it defines the value of Y when both X_1 and $X_2=0$. The regression coefficient b_1 quantifies the sensitivity of Y to change in X_1, adjusting for the effect of X_2 on Y. Similarly, b_2 quantifies the sensitivity of Y to change in X_2, adjusting for the effect of X_1 on Y.

As in simple linear regression, we evaluate whether individual predictors affect the response using *t* tests; for each regression coefficient b_j we compute

$$t_{b_j}=\frac{b_j}{s_{b_j}} \tag{5}$$

t_{b_j} will be large if the magnitude of b_j is large compared with the precision with which it is estimated, s_{b_j}. If t_{b_j} (with $n-k-1$ degrees of freedom, where k is the number of predictor variables) exceeds the maximum value expected under the null hypothesis of no relationship between Y and X_j, we conclude that X_j contributes significantly to the observed response in Y, adjusting for the effects of the other predictor variables.

An Example: Ice Cream Consumption

Diet contributes to cardiovascular risk, and therefore we may want to identify significant determinants of the consumption of certain foods. For example, Figure 1 shows data from a study of the determinants of ice cream consumption[7] in which we want to relate consumption (C, pints per person) to both mean outdoor temperature (T, °F) and weekly family income (I, $).

First, we separately examine the linear relationships between consumption and temperature and between consumption and income using simple regressions. For the former (Figure 1A), we estimate

$$\hat{C}=0.21\ \text{pints/person}+0.0031\ \text{pints/person/°F}\cdot T \tag{6}$$

$s_{b_T}=0.0005$ and $t_{b_T}=0.0031/0.0005=6.502$, which (with $30-2=28$ degrees of freedom) yields $P<0.001$. Thus, we conclude that for each 1°F rise in outdoor temperature, ice cream consumption increases, on the average, by 0.0031 pints per person, consistent with our visual impression of the data.

In contrast, simple regression of C on I suggests that ice cream consumption is not significantly associated with family income, as shown below,

$$\hat{C}=0.32\ \text{pints/person}+0.0005\ \text{pints/person/\$/week}\cdot I \tag{7}$$

because $s_{b_I}=0.002$ and $t_{b_I}=0.0005/0.002=0.25$, yielding $P=0.8$. Thus, we conclude that Equation 7 is no better explanation of the observed ice cream consumption than no fit at all, consistent with our visual impression of the data (Figure 1B).

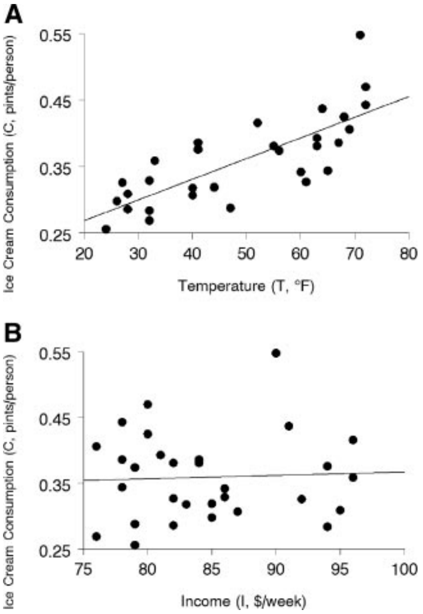

Figure 1. A, Scatterplot of ice cream consumption (C) vs temperature (T) showing the best-fit simple regression line as described by Equation 6 in the text. The overall fit is significant (*P*<0.001), and thus we conclude that ice cream consumption is associated with temperature. B, Scatterplot of ice cream consumption (C) vs income (I) showing the best-fit simple regression line as described by Equation 7 in the text. The overall fit is not significant (*P*=0.80), and thus we conclude that ice cream consumption is not associated with income. n=30 households.

On the basis of these separate analyses, we conclude that ice cream consumption increases as outdoor temperature increases but is not influenced by family income. Each of these separate analyses, however, assumes that no other important predictors of ice cream consumption are present to confound each analysis. In most studies, this situation will not be the case.

Now, we use multiple regression to estimate the simultaneous effects of T and I on C,

$$\hat{C}=b_0+b_TT+b_II \tag{8}$$

The Table shows the results of a computer program used to fit this equation to the data, yielding

$$\hat{C}=-0.113\ \text{pints/person}+0.0035\ \text{pints/person/°F}\cdot T+0.0035\ \text{pints/person/\$/week}\cdot I \tag{9}$$

Figure 2 represents this regression equation as a plane fit through the data (the data points are not plotted). Comparing the 2 separate simple regression results (Equations 6 and 7)

Table. Results of Regression Analysis

Independent Variable	Regression Coefficient b_j	Standard Error s_{b_j}	t_{b_j}	*P*
Intercept	−0.1132	0.1083	−1.045	0.3051
Temperature	0.0035	0.0004	7.963	<0.0005
Income	0.0035	0.0012	3.017	0.0055

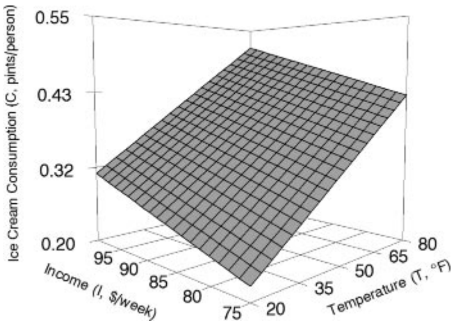

Figure 2. Three-dimensional plot of the best-fit multiple regression plane relating ice cream consumption (C) to both temperature (T) and income (I), as described by Equation 9 in the text. One edge of the plane (in the C vs T direction) has a positive slope, and we conclude, as we did in relation to the simple regression analysis shown in Figure 1A, that increasing ice cream consumption is associated with increasing temperature ($P<0.001$). The other edge of the plane (in the C vs I direction) also has a positive slope, and we conclude, in contrast to our conclusion in relation to the simple regression analysis shown in Figure 1B, that increasing ice cream consumption is also associated with increasing income ($P<0.01$).

with that of the multiple regression (Equation 9), we see that the estimates of b_T, quantifying the temperature effect, differ only slightly (0.0031 in Equation 6 versus 0.0035 in Equation 9), whereas the estimates of b_I, quantifying the income effect, differ by an order of magnitude (0.0005 in Equation 7 versus 0.0035 in Equation 9). From the fit of Equation 9, we estimate $s_{b_T}=0.0004$ pints/person/°F and $s_{b_I}=0.0012$ pints/person/$/week, and therefore $t_{b_T}=7.963$ ($P<0.001$) and $t_{b_I}=3.017$ ($P<0.01$). In contrast to the conclusions we drew from separate simple regressions, we now conclude that ice cream consumption, C, is significantly associated with both outdoor temperature, T, and family income, I. Ice cream consumption increases as outdoor temperature increases (b_T is positive), and, independently, after adjustment for temperature, ice cream consumption also increases as family income increases (b_I is positive).

With simultaneous consideration of both predictors, the multiple regression analysis is more revealing because, after adjustment for the association between T and C, it identifies an association between I and C that was masked in the separate simple regressions.

In summary, multiple linear regression and the associated statistics, b_j, s_{b_j}, and t_{b_j}, allow us to judge the magnitude and quality of the relationship between a response variable, Y, and 2 or more predictors, X_1, X_2, . . . , X_k. Using the individual t_{b_j}, we also make inferences about the statistical significance of each predictor, adjusting for the effects of the other predictors. These inferences and judgments are made under the assumption that the regression equation correctly specifies the true relationship among these variables; that is, Y is related linearly to the X_j, and Y is related only to the predictors, X_1, X_2, . . . , X_k, included in the equation. We further assume that the residuals are normally distributed and have equal variance across the predictor data space. Interpreting a multiple regression analysis thus requires careful examination of other aspects of the character of the fit and the relationship among variables to evaluate these assumptions.

Caveats and Considerations

Linearity

The multiple regression equation (Equation 4) estimates the additive effects of X_1 and X_2 on the response. It further specifies that each predictor is related linearly to the response through its regression coefficient, b_1 and b_2 (ie, the "slopes"). In simple linear regression, one can assess linearity by looking at a plot of the data points. In multiple regression, one can examine scatterplots of Y and of residuals versus the individual predictor variables. If a nonlinearity appears, one may be able to incorporate into the model an appropriate linearizing transform[1] or use nonlinear regression.[1,8]

Model Misspecification Bias

If the form of the regression equation is not correct (such as when substantial nonlinearities are ignored) or important predictor variables are left out of the equation, the estimates of those regression coefficients that are included in the equation will be biased. This situation occurred in our example: Excluding temperature as a predictor of ice cream consumption changed the estimate of the regression coefficient for the effect of family income by an order of magnitude ($b_I=0.0005$ in Equation 7 versus 0.0035 in Equation 9). Similarly, if we incorrectly specify the nature of the relationship between predictor variables (eg, additive versus interactive effects) or between the response and a predictor (eg, linear versus nonlinear relationships), we will bias the estimated regression coefficients. The magnitude of these biases will vary from problem to problem, and it is impossible to know with certainty that no model misspecification biases exist. One must guard against bias by systematically examining all data that were collected and applying sound judgment based on one's knowledge of the basic and clinical science underpinning the study, more often than not using multiple regression instead of simple regression.

Normality and Equal Variance

Multiple regression assumes that the residuals are normally distributed and have equal variance across the predictor data space. These assumptions are typically evaluated with the use of graphical methods and related statistics to assess the residuals.[1–5,9] If these assumptions do not hold, the response variable can sometimes be transformed so that the assumptions will hold for the transformed data,[1,9] although this approach is subject to the caveat that interpretations are now, strictly speaking, made with respect to transformed data. Alternatively, robust regression methods,[9] bootstrap methods,[10] or the mixed-effects regression method discussed below can be used to estimate regression parameters and their standard errors.

Multicollinearity

Each regression coefficient attempts to quantify the independent effect of the corresponding predictor on the response. However, when the multiple predictor variables are correlated with each other, which is often the case when dealing with biological or clinical data, this will not be the case. Correlation between the predictor variables reduces the precision of the estimates of the individual regression coefficients and therefore "inflates" the associated standard errors, s_{b_j}. Another way to think of this is that high correlation among predictor variables means that we use redundant information to predict the response; as a consequence, we are less certain of the independent effect of any 1 predictor. This problem is called *multicollinearity*[1] and should be of concern if the correlation between a pair of predictor variables is above about 0.9; depending on the specific data set and regression equation, multicollinearity might be an important consideration even with weaker correlations. Several diagnostics, including the so-called variance inflation factor, can be used to more fully evaluate multicollinearities, especially those arising in more complex models.[1,4,11] Severe multicollinearity can paradoxically yield a significant overall regression model fit in which none of the individual regression coefficients, b_j, are significant (because of the "inflation" of the s_{b_j}) and, in the extreme, can yield nonsensical estimates of 1 or more of the b_j.

Numerous ad hoc statistical approaches to dealing with multicollinearity are available.[1,4,12,13] However, sometimes you simply have to drop 1 or more of the "redundant" variables from the regression equation. Alternatively, to the extent that one can experimentally manipulate variables, multicollinearity can sometimes be mitigated with careful experimental design to reduce the correlation among predictor variables.[1,13]

Influential Data Points

An observation may be "unusual" in that it has an extreme location in the data space, as judged in relation to the location of the bulk of the other observations. These unusual observations may heavily influence the magnitude of the estimate of 1 or more of the regression coefficients, b_j. These points are called *outliers* if their location is unusual in the direction of the response variable (ie, unusually high or low values of Y) or *leverage points* if their location is unusual in the direction of a predictor variable (ie, unusually high or low values of X_j). Examining scatterplots of data will help to identify such points. In addition, formal "regression diagnostics," such as Cook's distance, the diagonal values of the "hat matrix," and Studentized residuals have been developed to help identify unusual observations and quantify their potential influence.[1,4,5,9,11] Although these diagnostic statistics can help to identify influential observations, particularly in multiple regression analyses in which neither the observations nor their effect can be easily visualized in the multivariate data space, they cannot tell us what to do. In many cases, these influential observations result from simple data entry errors, such as transpositions, and are easily corrected. In other cases, the influential observations reflect a problem of model misspecification, such as ignoring nonlinearity, and correcting the misspecification will reduce their influence. Accordingly, one should not use regression diagnostics to justify excluding otherwise valid observations from analysis simply to avoid their influence. On the other hand, influential observations can lead to erroneous results, and therefore their presence and effect should be evaluated and understood.

Extensions of Multiple Linear Regression

Beyond simple extension of multiple regression to include additional continuous predictor variables, numerous other useful extensions to the basic procedure are available.

Variable Selection

Several procedures, known collectively as variable selection methods, have been developed to select a "best" multiple regression model that includes a subset of predictor variables drawn from a larger pool of candidate predictors.[1,3,4,14] Most of these techniques are based on incremental changes in SS_{res} (Equation 2) as predictor variables are (1) added sequentially to a model, starting from nothing (forward selection); (2) subtracted sequentially from a model, starting with all candidates included (backward elimination); or (3) more commonly, selected by stepwise regression, a strategy that proceeds as with forward selection, but each time a variable is added a backward elimination step occurs to test whether any variables entered previously can be removed. These techniques can be useful adjuncts in a multipronged strategy to identify an appropriate multiple linear regression model; they allow examination of many possible regression models to look for consistency of model identification with the use of multiple methods. It is important, however, to avoid rote application of these methods, particularly for large data sets containing many possible predictor variables in which multicollinearity may be a problem. Severe multicollinearity will play havoc with the order of selection or elimination of variables with the use of these methods, and one must be cautious in inferring relative importance of predictors on the basis of their order of selection.

Interactions Among Predictor Variables

The 2-predictor multiple regression equation (Equation 4) specifies that predictors X_1 and X_2 are additive in their respective effects. Often, however, 2 or more predictor variables interact (ie, synergize or compete) to determine a response. Indeed, determining whether effects are additive or not is often the reason for conducting a study, and one may therefore wish to explore whether 2 predictors interact (ie, are not simply additive) in their relation to the response. The simplest interaction between 2 predictors is introduced into a multiple regression equation through the product of the 2 variables.[1–3] For example, if we wish to express an interaction between predictors X_1 and X_2 in Equation 4, we use

$$(10) \qquad \hat{Y}=b_0+b_1X_1+b_2X_2+b_3X_1X_2$$

If the regression coefficient, b_3, is statistically significant, then we have evidence that the relationship between Y and X_1 depends on the value of X_2 (or vice versa, depending on how we view the underlying biology).

This basic concept can be generalized to more complicated interaction models. However, this must be done thoughtfully. Introducing several product terms into a complex multiple regression problem, particularly if the expanded set of predictors is then subjected to variable selection methods, can yield misleading results. In addition, product terms will highly correlate with each of the variables used to create the product, artificially introducing multicollinearity.

Categorical Predictor Variables

We have shown how to relate a continuous response variable, Y, to multiple continuous predictor variables, such as X_1 and X_2. Often, however, we may want to also include predictor variables that are categorical, such as gender, ethnicity, or treatment group. It is possible to do so by including in the regression equation a set of "dummy" (or "indicator") variables that take on values of 0, 1, or −1 to represent the levels of categorical information. The simplest example is a variable that has only 2 categories, such as chocolate and vanilla. If our focus, for example, is on the difference in consumption of chocolate ice cream with reference to vanilla, we create a dummy variable F (for flavor) defined as 0 if vanilla and 1 if chocolate (this is called *reference coding*; the reference group is coded with 0). To introduce this into our analysis and determine whether ice cream consumption differs for vanilla and chocolate flavors, we add F to the regression equation,

$$\hat{C}=b_0+b_TT+b_II+b_FF \tag{11}$$

b_F quantifies the average difference in chocolate ice cream consumption with reference to vanilla, after adjustment for the effects of temperature and income; computing t_{b_F} allows us to judge whether this effect is statistically significant. Alternatively, we could conduct the same analysis by defining F=1 if vanilla and −1 if chocolate, which is often called *effects coding*. With this coding, b_F quantifies the deviation of the response for chocolate and vanilla from the average of both chocolate and vanilla. For simple problems, reference (1, 0) coding is often more straightforward to interpret. For more complicated problems, the choice of coding will depend on many factors that are beyond the scope of this review. This dummy variable method generalizes to cases in which the factor of interest has more than 2 levels by creating a set of dummy variables totaling 1 fewer than the number of levels of the factor.[1–3]

Repeated Measurements Within the Same Subjects

Sometimes repeated measurements are made within individual subjects to establish a relationship between 2 or more variables in each of multiple subjects, and then these multiple relationships are pooled into 1 data set.[15,16] If the regression analysis does not account for the different subjects, both the estimates of the regression coefficients, b_j, and the estimates of error, including the s_{b_j}, will be biased and could be in error.

Consider, for example, the data shown in Figure 3, in which response, Y, and predictor, X, were measured in 6 subjects. If we use simple regression to fit the pooled data with the use of Equation 1, ignoring the subjects, we estimate $b_X=+2.16$ with $s_{b_X}=0.478$. Computing t_{b_X} leads to the conclusion that a significant positive relationship exists between Y and X ($P<0.001$; Figure 3A).

The data, however, suggest the opposite, a relationship with a negative slope when considered within each subject. In effect, the model is misspecified by excluding the subjects, and bias is introduced because the relative location of each subject's response is ignored.

A simple (but, as we will see, usually unsatisfactory) approach to account for variation from subject to subject is to represent the subjects by a adding a set of dummy variables to the regression equation.[1,15,16] For example, for these 6 subjects, we define a set of 5 dummy variables using effects coding, $S_i=1$ if subject i (i=1 to 5), −1 if subject 6, 0 otherwise, and write

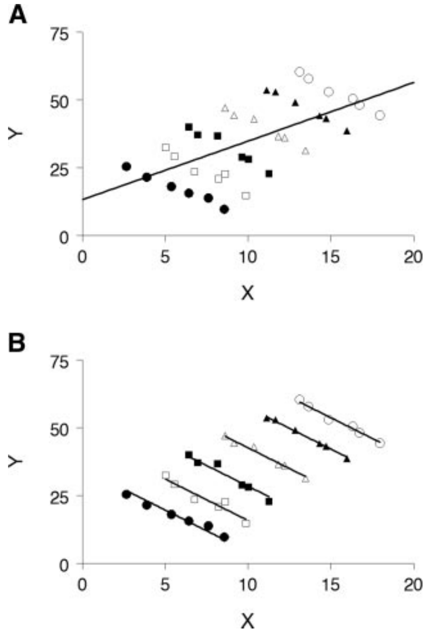

Figure 3. A, Scatterplot of values of Y and X measured in 6 subjects (each subject has a different symbol). When we use simple regression to fit these data using Equation 1, we conclude that Y significantly increases as X increases ($P<0.001$). B, The same data shown in A but now fit with the use of Equation 12, which is a multiple regression equation that includes a set of dummy variables to account for the fact that data were collected within 6 individual subjects. When we account for the effect of different subjects, we now correctly estimate that Y decreases significantly ($P<0.001$) as X increases, within each subject, just as our visual impression of the data suggests.

$$\hat{Y}=b_0+b_XX+\sum b_{S_i}S_i \tag{12}$$

Where the sum (Σ) is from 1 to 5.[1] We estimate $b_X=-3.05$ with $s_{b_X}=0.118$ ($P<0.001$; Figure 3B). Thus, after accounting for the different lines within each of the subjects, we now find a negative slope (with a smaller standard error), which much better describes the relationship within individual subjects. Note that Equation 12 estimates b_X assuming that a common slope is present across all subjects, which may or may not be appropriate depending on the specific problem (interactions can be introduced to allow each subject's line to have its own slope).

Accounting for subject variability in this way and estimating regression coefficients and standard errors by ordinary least squares regression treats the subjects as a fixed effect, with "fixed" meaning that these individuals are the only "levels" of the factor (subjects) of interest, which constrains statistical inference to only the specific subjects studied. Most of the time, however, we want to make statistical inferences that extrapolate to an entire population on the basis of data collected in our sample of subjects; to do so correctly, we must treat the subjects as a random effect (ie, as a random sample of the population of interest). A common way to do this is with mixed-effects regression,[17,18] where "mixed" refers to the inclusion of both random and fixed effects. Although it is beyond the scope of this review to delve into details (another article in this series will address this topic in the context of longitudinal analysis), we reestimated this example using a mixed-effect regression; $b_X = -3.04$ with $s_{b_x} = 0.118$, which are nearly identical to those obtained above. If we turn our attention to b_0, however, we see that although the estimates of b_0 are similar (66.78 versus 66.68), s_{b_0} is much larger in the mixed-model result (10.14 versus 1.23 with the use of ordinary least squares). This mixed-model regression approach is usually necessary to correctly estimate uncertainty when repeated observations exist within subjects.

Summary

More often than not, when one's impulse is to conduct a series of separate simple regressions involving the same response variable, multiple regression should be used instead. The flexibility of multiple regression allows elegant, insightful, and often the only correct analysis. Simple nonlinearities and interaction effects can be introduced to extend the utility of this method well beyond that of simple regression. As with any multivariate statistical technique, however, it is possible to make substantial errors if the method is applied blindly without appropriate consideration of the underlying assumptions, correlations among predictors, influential observations, and thoughtful exploration of model structure.

Disclosures

None.

References

1. Glantz SA, Slinker BK. *Primer of Applied Regression and Analysis of Variance*. 2nd ed. New York, NY: McGraw-Hill; 2000.
2. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. *Applied Regression Analysis and Multivariable Methods*. 3rd ed. Boston, Mass: Duxbury Press; 1997.
3. Kutner MH, Nachtsheim CJ, Neter J, Li W. *Applied Linear Statistical Models*. 5th ed. New York, NY: McGraw-Hill; 2004.
4. Myers RH. *Classical and Modern Regression with Applications*. 2nd ed. Boston, Mass: Duxbury Press; 1990.
5. Weisberg S. *Applied Linear Regression*. 3rd ed. New York, NY: John Wiley & Sons Inc; 2005.
6. Slinker BK. The statistics of synergism. *J Mol Cell Cardiol*. 1998;30: 723–731.
7. Kadiyala KR. Testing for the independence of regression disturbances. *Econometrica*. 1970;38:97–117.
8. Ratkowsky DA. *Handbook of Nonlinear Regression Models*. New York, NY: Marcel Dekker; 1990.
9. Montgomery DC, Peck EA, Vining GG. *Introduction to Linear Regression Analysis*. 3rd ed. New York, NY: John Wiley & Sons Inc; 2001.
10. Chernick MR. *Bootstrap Methods: A Practitioner's Guide*. New York, NY: John Wiley & Sons Inc; 1999.
11. Belsley DA, Kuh E, Welsch RE. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York, NY: John Wiley & Sons Inc; 1980.
12. Smith G, Campbell F. A critique of some ridge regression methods. *J Am Stat Assoc*. 1980;75:74–81.
13. Slinker BK, Glantz SA. Multiple regression for physiological data analysis: the problem of multicollinearity. *Am J Physiol*. 1985;249:R1–R12.
14. Hocking RR. The analysis and selection of variables in linear regression. *Biometrics*. 1976;32:1–49.
15. Slinker BK, Wu Y, Brennan AJ, Campbell KB, Harding JW. Angiotensin IV has mixed effects on cardiac function and speeds relaxation. *Cardiovasc Res*. 1999;42:660–669.
16. Peng J, Raddatz K, Molkentin JD, Wu Y, Labeit S, Granzier H, Gotthardt M. Cardiac hypertrophy and reduced contractility in hearts deficient in the titin kinase region. *Circulation*. 2007;115:743–751.
17. Feldman HA. Families of lines: random effects in linear regression analysis. *J Appl Physiol*. 1988;64:1721–1732.
18. Hedeker D, Gibbons RD. *Longitudinal Data Analysis*. New York, NY: John Wiley & Sons Inc; 2006.

Key Words: biostatistics ■ data analysis ■ statistics