Advanced Statistics: Linear Regression, Part II: Multiple Linear Regression

Keith A. Marill, MD

Abstract

The applications of simple linear regression in medical research are limited, because in most situations, there are multiple relevant predictor variables. Univariate statistical techniques such as simple linear regression use a single predictor variable, and they often may be mathematically correct but clinically misleading. Multiple linear regression is a mathematical technique used to model the relationship between multiple independent predictor variables and a single dependent outcome variable. It is used in medical research to model observational data, as well as in diagnostic and therapeutic studies in which the outcome is dependent on more than one factor. Although the technique generally is limited to data that can be expressed with a linear function, it benefits from a well-developed mathematical framework that yields unique solutions and exact confidence intervals for regression coefficients. Building on Part I of this series, this article acquaints the reader with some of the important concepts in multiple regression analysis. These include multicollinearity, interaction effects, and an expansion of the discussion of inference testing, leverage, and variable transformations to multivariate models. Examples from the first article in this series are expanded on using a primarily graphic, rather than mathematical, approach. The importance of the relationships among the predictor variables and the dependence of the multivariate model coefficients on the choice of these variables are stressed. Finally, concepts in regression model building are discussed. **Key words:** regression analysis; linear models; least-squares analysis; statistics; models, statistical, epidemiologic methods. ACADEMIC EMERGENCY MEDICINE 2004; 11:94–102.

Multiple linear regression is a generalization of simple linear regression in which there is more than one predictor variable. If the investigator suspects that the outcome of interest may be associated with or depend on more than one predictor variable, then the approach using simple linear regression may be inappropriate. A multiple regression model that accounts for multiple predictor variables simultaneously may be used. For example, in the first scenario discussed in Part I of this series, the investigator studied the relationship between the intensity of insulin therapy and the resolution of serum acidosis in patients with diabetic ketoacidosis (DKA). The resolution of acidosis seems to depend on the intensity of insulin therapy, but there may be other important factors too. These could include: the initial severity of the DKA episode, the severity of the patient's underlying disease, the administration of other treatments such as intravenous (IV) fluid, etc. Multiple linear regression allows the investigator to account for all of these potentially important factors in one model. The advantages of this approach are that this may lead to a more accurate and precise understanding of the association of each individual factor with the outcome. It also yields an understanding of the association of all of the factors as a whole with the outcome, and the associations between the various predictor variables themselves.

Expanding the schematic approach introduced in Figure 6 of Part I, the introduction of another predictor variable to the model is represented by the addition of another circle that overlaps the outcome variable circle. This overlap is labeled area C in Part II, Figures 1A and 1B. In general, the addition of the new predictor circle and its overlap, area C, with the outcome circle will increase the total portion of the outcome explained by the regression, areas A + C, and decrease the unknown or residual portion, area B. The new predictor circle and area C may, to some degree, overlap the original predictor circle and area A, depending on the relationship between these two predictor variables. This represents the variable amount of redundancy and collinearity existing between the two predictor variables in the model.

THE MULTIPLE LINEAR REGRESSION MODEL

The multiple linear regression model is built on the same foundation as simple linear regression, and the

From the Division of Emergency Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA.
Received July 24, 2001; revision received July 9, 2002, and April 21, 2003; accepted September 10, 2003.
Series editior: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor–UCLA Medical Center, Torrance, CA.
Based on a didactic lecture, "Concepts in Multiple Linear Regression Analysis, " given at the SAEM annual meeting, St. Louis, MO, May 2002.
Address for correspondence and reprints: Keith A. Marill, MD, Massachusetts General Hospital, 55 Fruit Street, Clinics 115, Boston, MA 02114. Fax: 617-724-0917; e-mail: kmarill@partners.org.
Part I appears on page 87.
doi:10.1197/S1069-6563(03)00601-8

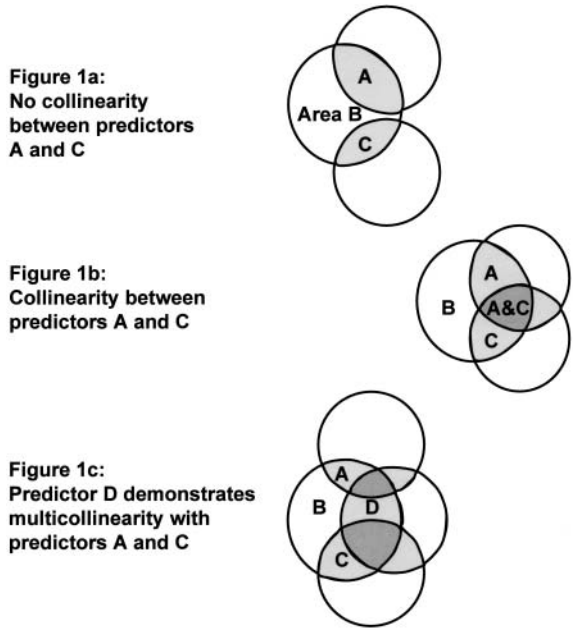

Figure 1. Multivariate schematics.

four fundamental assumptions made with simple linear regression must also be true for multiple linear regression. However, in addition to the concepts discussed thus far for simple linear regression, which remain applicable, a new set of concepts must be introduced. This discussion will concentrate on the situation in which there are two predictor variables and one outcome variable. With a total of three variables, a three-dimensional figure can be used to visualize the data. Models with a larger number of predictor variables follow the same principles, but are more difficult to visualize.

The equation for the regression model now represents a flat plane. Letting z be the outcome variable and x and y be the predictor variables, we have:

$$z = k_1x + k_2y + c \qquad \text{(equation 1)}$$

where k_1 and k_2 are the constant coefficients for x and y, respectively, and c is the z intercept at $x = y = 0$. k_1 and k_2 determine the tilt of the plane along the x- and y-axes, respectively. Note that the outcome variable, z, is a linear function of each of the predictor variables, x and y, and this forces the regression model to be a flat plane with no curves or bending. Figure 2A demonstrates a regression plane where $k_1 = k_2$.

The plane that fits the data best can again be found using the least-squares technique described in the first article in this series. This approach finds the plane that minimizes the sum of the residuals squared. The residual value for each data point equals the actual value of z at that point minus the corresponding predicted value of z on the regression plane (Figure 5A). The optimal coefficients c, k_1, and k_2 are found such that the regression plane has the proper

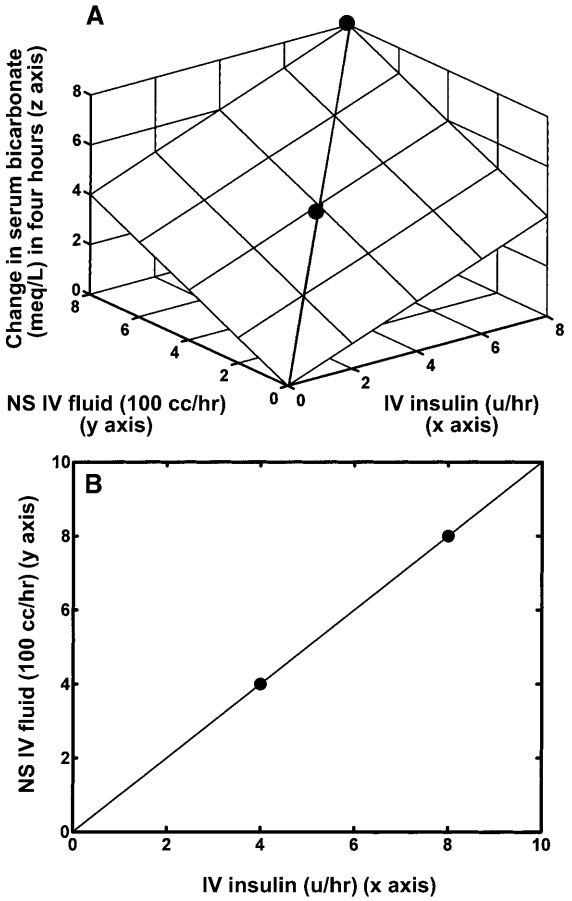

Figure 2. (**A**) $z = 0.5x + 0.5y$, $R^2 = 1.0$. (**B**) $y = 1x$, $R_{pred}^2 = 1.0$. NS = normal saline; IV = intravenous.

elevation and tilt that minimizes SS_{res}. This approach leads to three equations and three unknowns, and there usually is a unique solution. It is still true that $SS_{tot} = SS_{reg} + SS_{res}$ and $R^2 = SS_{reg}/SS_{tot}$, but the meaning of these equations has changed somewhat. SS_{reg} includes the contribution of both of the predictor variables to the regression, not each one individually. R is now called the multiple correlation coefficient. R^2, which is called the coefficient of determination, suggests what proportion of the variation in the outcome variable can be attributed to both of the predictor variables in the linear model as a whole. Referring to the schematics in Figures 1A and 1B, $R^2 = (A + C)/(A + B + C)$, where the overlap, if any, of areas A and C would only be counted once.

In simple linear regression, a test for whether the relationship in the regression model is statistically significant and unlikely to be due to chance is equivalent to a t-test in which the ratio of the slope to its standard error (SE) is computed and checked for significance. In multiple regression, this test has a different meaning because there are multiple predictor variables and multiple slopes. Instead, an analysis of variance (ANOVA) is used to test for the significance of the model as a whole. Schematically, this is equivalent to comparing the size of area A + C

versus area B in Figures 1A and 1B after adjusting for the number of predictor variables and data points. If the regression area A + C is relatively large compared with the residual area B, then it is concluded that the predictors taken together are associated with the outcome beyond mere chance. Mathematically, a comparison is made between the mean $SS_{reg} = MS_{reg} = SS_{reg}/k_{tot}$, which includes contributions from all of the predictor variables, and the mean $SS_{res} = MS_{res} = SS_{res}/(n - k_{tot} - 1)$, where n is the number of data points and k_{tot} is the number of predictor variables. $F = MS_{reg}/MS_{res}$, and if, after accounting for the appropriate degrees of freedom, F is sufficiently large, then the null hypothesis is rejected and it is concluded that the multiple linear model explains some of the variation in the outcome variable.

RELATIONSHIPS AMONG THE PREDICTORS

Perhaps the most important difference in multiple versus simple linear regression is that the multiple regression model includes the linear relationships among the predictor variables themselves. These relationships, termed "multicollinearity," can have a tremendous effect on the model coefficients and the precision with which they are known. To illustrate this, we return to the simple univariate predictor models in Figures 2 through 5 of Part I of this series, and now include multivariate data with two predictor variables. Figures 2A through 5A of this article, Part II, correspond to Figures 2 through 5 of Part I and include the same data points; however, a second predictor variable, y, has been added. How does inclusion of an additional predictor variable affect the regression model? The answer is "It depends"—it depends on whether there is a linear relationship between the new predictor variable and the predictor variable or variables that already are present in the model.

COLLINEARITY

In Figure 2 of Part I of this series, the investigator studied the association of the intensity of intravenous (IV) insulin therapy with the rate of resolution of DKA in two diabetic patients. It was found that the predictor and outcome variables were proportional, and for every one unit per hour of insulin therapy, there was an associated 1-mEq/L increase in the serum bicarbonate level after four hours of therapy. The researcher now returns to the data and investigates whether the intensity of IV normal saline (NS) fluid therapy also is associated with the rate of DKA resolution.

The outcome variable, the increase in the serum bicarbonate after four hours of therapy, is graphed as a function of the intensity of insulin and IV NS therapy for the two study patients in Figure 2A, Part II. Notice that the resolution of DKA seems to be associated with the intensity of both insulin and fluid therapy. Recall that the original equation for the relationship between insulin therapy and the improvement in serum bicarbonate was $z = x$. Would a correct model with two predictor variables now be $z = x + y$? No. If the intensity of insulin therapy is 4-units per hour and IV NS hydration is 4 in 100 mL/hr units, then the improvement in serum bicarbonate would be $4 + 4 = 8$ mEq/L after four hours of therapy, instead of 4 mEq/L as in the figure. This would overestimate the improvement in outcome. A more correct model would be $z = 1/2x + 1/2y$. By adding the y variable to the model, the value of the coefficient in the x-axis has decreased by 50% from 1 to 0.5. Why is this so? It is because the data displays collinearity in the x,y plane.

Figures 2B through 5B are two-dimensional graphs of the same data in Figures 2A through 5A, but only the x- and y-axes are displayed. This allows a clear display of the relationship between the predictor variables x and y in each data set. Notice that the value of y varies with the value of x in Figure 2B. The patient who received more insulin therapy also received more IV NS hydration. They increase together linearly, and thus display positive collinearity. In the simple linear model of Part I Figure 2, the entire improvement in the serum bicarbonate was associated only with the insulin therapy, whereas in the multivariable model in Part II, Figure 2A, we chose to apportion the improvement in the serum bicarbonate equally between the insulin and IV NS treatments. When IV NS treatment is included in the model, the improvement in serum bicarbonate associated with treatment must be shared between two therapies, insulin and NS. When collinearity is present, the magnitude of the predictor coefficients change depending on which predictor variables are included in the model. In particular, when an increase in one predictor variable is associated with an increase in another predictor, there is positive collinearity. When there is positive collinearity, the value of positive predictor coefficients will tend to decrease as more predictors are included in the model. The association with the outcome must be shared among multiple predictors.

The change in the regression model associated with the addition of a new predictor variable can be even more dramatic. Consider the second example in Part I of this series: the investigator examined the hypothesis that a higher initial respiratory rate may be associated with a greater improvement in DKA. It was found in Part I, Figure 3, that patients with higher initial respiratory rates demonstrated greater improvement. The investigator now realizes, however, that the intensity of insulin therapy should also be included in the analysis.

Part II, Figure 3A, is a graph of the improvement in serum bicarbonate as a function of the initial

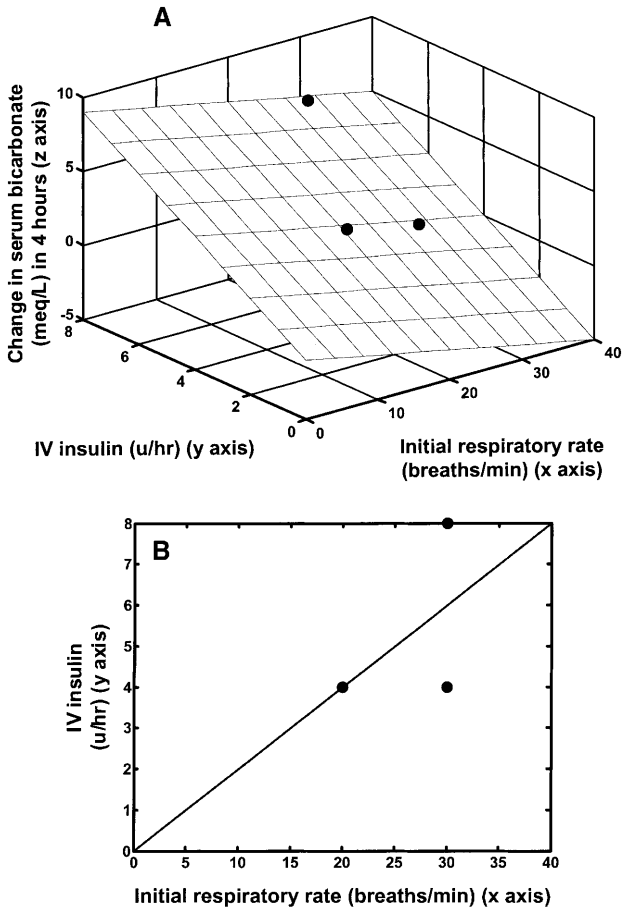

Figure 3. (**A**) $z = -0.1x + 1.25y - 1$, $R^2 = 1.0$. (**B**) $y = 0.2x$, $R_{pred}{}^2 = 0.25$. IV = intravenous.

respiratory rate and the intensity of insulin treatment. As expected, the increase in serum bicarbonate is greater in patients who receive more insulin. Interestingly, the improvement in serum bicarbonate is now less in patients with higher initial respiratory rates. The sign of the x coefficient relating the initial respiratory rate to the improvement in serum bicarbonate has changed from positive to negative. Once again, this change has occurred because there is collinearity between the two predictor variables, the initial respiratory rate and the intensity of insulin therapy. Patients with higher initial respiratory rates presumably have more severe disease and are treated more aggressively with insulin. This association and positive collinearity are demonstrated in the plot relating the two predictor variables, x and y (Figure 3B). The greater improvement in serum bicarbonate originally attributed to a higher initial respiratory rate in the univariate analysis actually seems to be due to more aggressive insulin therapy. After analyzing the same data as in Part I, but with two predictor variables instead of one, the investigator finds that there is no longer evidence suggesting that patients can hyperventilate their way out of DKA.

These examples have demonstrated effects due to collinearity and confounding between two predictor variables, and they are represented schematically in Figure 1B, in which the two predictor areas A and C overlap. The term ''multicollinearity'' is used to describe the same types of collinear effects that can occur among three or more predictor variables in a data set. Both of these examples demonstrated positive collinearity. Sometimes, the value of one predictor variable may decrease as the other predictor variable increases. This would be negative colinearity. Data sets with many predictor variables may contain complex multicollinearities with both positive and negative collinear relationships.

NO COLLINEARITY

In Part I, Figure 4B, the investigator determined that the log of the duration of the intensive care unit (ICU) stay for patients with DKA varied with the initial intensity of insulin therapy: $\log z = kx + c$. Perhaps there are other factors that might also help explain the duration of ICU admission, such as the patient's age, comorbid illnesses, or secondary infections. In Part II, Figure 4A, the investigator graphed the log of the ICU stay as a function of two predictor variables, the initial intensity of insulin therapy, x, and the patient's age in years, y. When comparing the new figure with two

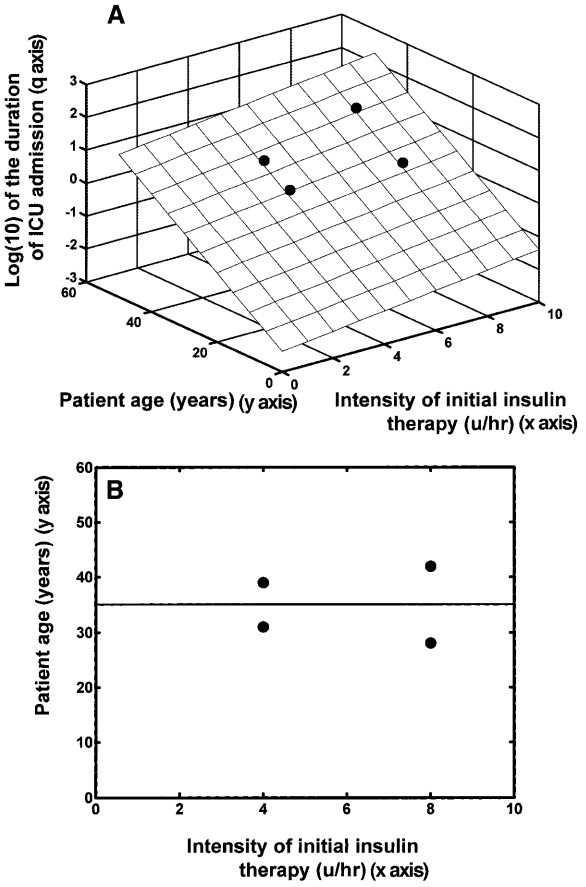

Figure 4. (**A**) $q = 0.097x + 0.074y - 2.37$, $R^2 = 1.0$. (**B**) $y = 0x + 35$, $R_{pred}{}^2 = 0$. ICU = intensive care unit.

predictor variables, Part II, Figure 4A, to the previous graph with one predictor variable, Part I, Figure 4b, note that although a new coefficient in the y-plane representing patient age has been added, there is no change in the x-axis coefficient relating insulin therapy to ICU stay. Inspection of the relationship between the two predictor variables in Part II, Figure 4B reveals that there is no linear relationship in the x,y plane between the two predictor variables, because the slope of the regression line is zero. The intensity of initial insulin therapy is unrelated to patient age, and thus there is no collinearity between the two variables. As a result, each predictor is independently associated with the outcome, and the inclusion of the age predictor has no bearing on the association or coefficient of insulin therapy with ICU stay. This situation is represented schematically in Figure 1A. Also notice in Part II, Figure 4A that the regression plane now fits the data perfectly, and all of the residuals are zero. Addition of a predictor variable that demonstrates no collinearity with the other predictors usually will improve the model by reducing the residuals without altering the coefficients that already are present.

ASSESSING COLLINEARITY

How is the degree of collinearity among two predictor variables or multicollinearity among multiple predictor variables assessed? The degree of collinearity between two predictor variables is quantified by their correlation coefficient, R_{pred}^2. The correlation coefficient of one predictor variable with another can be labeled R_{pred}^2 to distinguish it from the coefficient of determination of all of the predictors with the outcome, which remains R^2. Returning to Part II, Figure 2B, it can be observed that there is complete collinearity in the x,y plane representing the two predictors, because the data forms a straight line, ($R_{pred}^2 = 1$). In Figure 3B, there is partial collinearity, ($R_{pred}^2 = 0.25$), and in Figure 4B, there is no collinearity, ($R_{pred}^2 = 0$). When there are more than two predictor variables, multicollinearity can be assessed by determining the R_{pred}^2 or coefficient of determination of the predictor variable of interest with the other predictor variables. This essentially is a regression of one predictor variable with all of the others, and it represents a regression among the predictors within the larger regression model. Schematically, it represents the total proportion of the predictor-of-interest circle that is overlapped by other predictor circles in the model (Figure 1).

QUANTIFYING UNCERTAINTY OF THE COEFFICIENTS

In the experiment described in Part I, Figure 5, the investigator administered either placebo, potassium, bicarbonate, or both agents to each of four groups of animals with experimental salicylate overdose. Evaluating the results with respect to potassium infusion alone in Part I, Figure 5, it was found that salicylate clearance was higher in the animals that received potassium. The investigator now takes a multivariate approach and analyzes salicylate clearance as a function of both potassium and bicarbonate treatment in Part II, Figure 5A. Similar to the potassium predictor variable, the bicarbonate variable, y, is given dummy variable values of 0 or 1, corresponding to the absence or presence of bicarbonate infusion, respectively. Inspection of Part II, Figure 5B reveals that, by design,

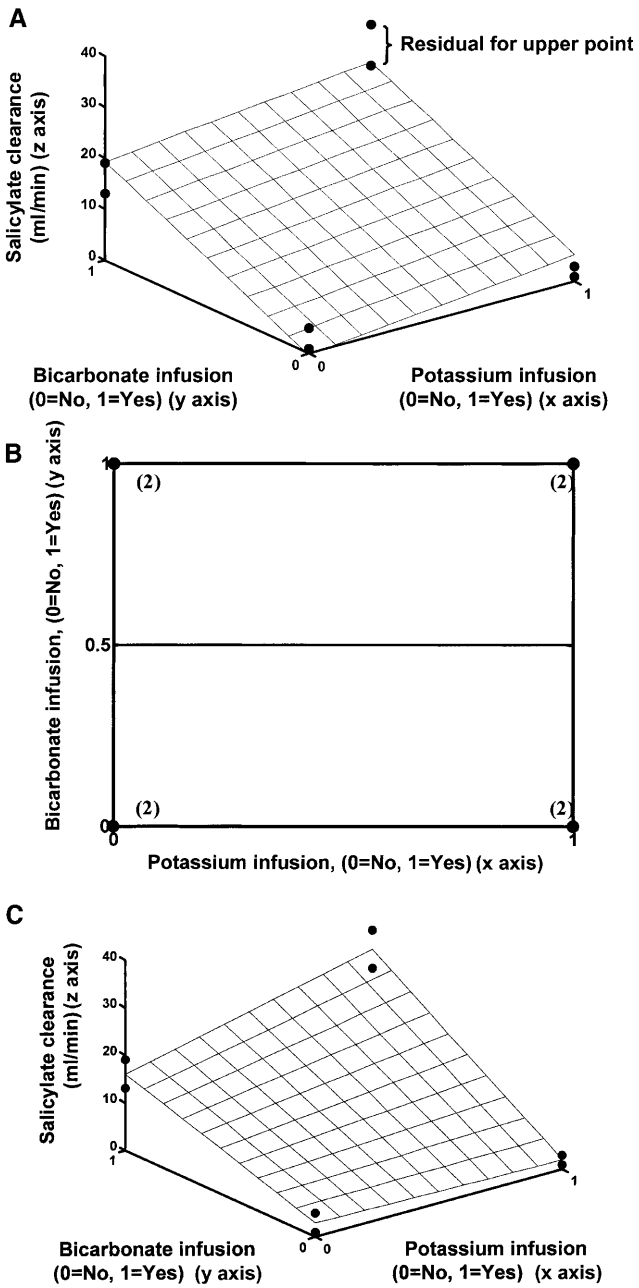

Figure 5. (**A**) $z = 5.5x + 19.5y - 0.25$, $R^2 = 0.85$. (**B**) $y = 0x + 0.5$, $R_{pred}^2 = 0$. (**C**) $z = -x + 13y + 13xy + 3$, $R^2 = 0.94$.

there is no linear relationship or collinearity between the two predictor variables, potassium and bicarbonate infusion, and each data point represents two animals that received identical treatment. Thus, addition of the bicarbonate variable to the analysis in Part II, Figure 5A causes no change in the value of the potassium coefficient, 5.5, from the original model in Part I, Figure 5.

In Part I, the investigator also determined the SE of the potassium coefficient, 8.7, and used this to perform inference testing and to calculate the 95% confidence interval (CI) of the coefficient, −15.8 to 26.8. How is the SE of a predictor coefficient determined in the multiple regression model, and how is it affected by the inclusion of other predictor variables?

It was stated in Part I that the SE of a coefficient in simple linear regression is:

$$\mathrm{SE(coefficient)}_{\mathrm{simplereg}} = \left[\frac{\mathrm{SS_{res}}}{(n-2)\sum_{1}^{n}(X-X_{\mathrm{mean}})^2}\right]^{1/2} \qquad \text{(equation 2)}$$

In multiple linear regression, a similar formula is used, but a modification must be made to account for possible multicollinearity. When collinearity is present among two or more predictor variables, there is additional uncertainty in the value of their coefficients.

Returning to Figure 1, whenever there is a linear relationship between two variables in the experimental sample, their circles overlap. The degree or strength of overlap will have some uncertainty when inferences are made on a different or larger population. Specifically, when two predictor variables exhibit collinearity, their circles overlap, as in Figure 1B. Due to this overlap, the size of the individual areas A and C are less certain. The extent of overlap of one predictor variable with all of the others is quantified by its multiple correlation coefficient with the other predictors, R_{pred}^2.

The increased uncertainty due to collinearity also can be visualized with inspection of the linear regression plane. Compare Figures 2 and 4, which demonstrate the extremes of complete and no collinearity between the x and y predictor variables. Imagine that the regression plane is a piece of cardboard that is balanced on the data points in space. In Figure 2A, the cardboard easily can be tilted or rotated around the line connecting the two data points, whereas in Figure 4A, the cardboard sits on a stable platform of four points spaced apart. As the data points become more linearly oriented in the x,y predictor plane, the cardboard becomes less stable and more easily tilted over the line that represents the regression of the x and y predictors. This is analogous to the simple linear regression model in which the slope of the regression line becomes less certain as the spacing of the points along the x-axis is decreased.

The increase in the SE of a predictor coefficient due to multicollinearity is quantified by the square root of the variance inflation factor (VIF), where:

$$\mathrm{VIF} = \frac{1}{1-R_{\mathrm{pred}}^2} \qquad \text{(equation 3)}$$

R_{pred}^2 can have any value from 0 to 1. The greater the overlap and collinearity of the predictor of interest with the other predictors, then the greater is R_{pred}^2 and the VIF. Recall that the SE is the square root of the variance. To determine the SE of a predictor coefficient in multiple linear regression, we multiply the formula from simple linear regression times the square root of the VIF to obtain:

$$\mathrm{SE(coefficient)}_{\mathrm{multreg}} = \left[\frac{\mathrm{SS_{res}}}{(n-k_{\mathrm{tot}}-1)\sum_{1}^{n}(X-X_{\mathrm{mean}})^2} * \frac{1}{1-R_{\mathrm{pred}}^2}\right]^{1/2} \qquad \text{(equation 4)}$$

where the denominator of the original formula has been slightly modified to account for the number of predictor variables, k_{tot}. What happens to the SE of the potassium coefficient in the salicylate clearance experiment when the bicarbonate predictor variable is included in the regression model? We already have determined from Figure 5B that there is no collinearity between the potassium and bicarbonate predictor coefficients. Therefore, R_{pred}^2 for the potassium coefficient is zero, and the VIF $= 1/(1-0) = 1$. So there is no inflation of the variance or the SE. Does this mean that the SE remains unchanged? No. A review of Figure 1A reveals that when a second predictor variable represented by area C is included in the model, the area representing the residuals or uncertainty, area B, becomes smaller. The new information serves to increase the certainty of the model, and this is manifested by a decrease in the total error or residuals, $\mathrm{SS_{res}}$. According to equation 4, if the $\mathrm{SS_{res}}$ decreases, then the SE of the coefficient also decreases. By including the bicarbonate coefficient, the SE of the potassium coefficient decreases from 8.7 to 3.8, and the span of the 95% CI of the coefficient decreases from −15.8 to 26.8 to the narrower range of −4.3 to 15.3. It was previously noted that inclusion of a noncollinear predictor variable in the regression model usually adds new information and decreases the total uncertainty or $\mathrm{SS_{res}}$. It is now apparent that this generally leads to a decrease in the SE of the other predictor coefficients.

INCREASING POWER

The concept in the paragraph above has important implications for inference testing and univariate

versus multivariate analysis. In general, the power to determine a statistically significant effect is based on the magnitude of the effect and the degree of uncertainty in the results. If the magnitude of the effect is relatively large with respect to the uncertainty of the results, then the null hypothesis is rejected and statistical significance is concluded. The power of the test can only be increased by either increasing the magnitude of the effect or decreasing the uncertainty in the results. Perhaps the most common approach to decreasing uncertainty and increasing power is by collecting more data and increasing the number of data points, n. Including another noncollinear independent predictor variable in an analysis is another technique that can be used to decrease uncertainty and increase power without increasing n. In essence, instead of collecting more data points, the investigator is using more information from each data point that already is included to improve the precision of the model. In the example above, a t-test can be used in the original univariate model in Part I, Figure 5 to determine the effect of potassium infusion on salicylate clearance. In Part II, Figure 5A, the investigator has moved to a multivariate approach that includes both the potassium and bicarbonate therapies. The comparable multivariate statistical test would be a two-way ANOVA. Although the magnitude of the potassium therapy effect is the same in both tests, the power to determine its statistical significance may be increased using the ANOVA approach.

LEVERAGE REVISITED

The multivariate model depicted in Part II, Figure 5A is an improvement over the univariate model in Part I, Figure 5, as evidenced by a decrease in the SS_{res} from 905 to 145 and a corresponding increase in R^2 from 0.06 to 0.85. Careful inspection of the graphs, however, reveals that the two animals that received both potassium and bicarbonate had a salicylate clearance that was remarkably higher than the other three groups. These two data points are exerting upward leverage in both the univariate and multivariate regression models. The evaluation of leverage in multivariate regression is comparable with that in simple linear regression. In multivariate regression, Cook's distance corresponds to the combined change in all of the predictor coefficients and the z intercept when the data point in question is removed. If Cook's distance is relatively large for one or more data points as compared with the others, then those points may have a disproportionate influence on the regression model.

INTERACTION EFFECT

Why did the two animals that received both potassium and bicarbonate have such a high salicylate clearance? It could be because those particular animals happened to have highly efficient kidneys, or perhaps there was a dosing or measurement error. An alternative explanation would be that there is an interaction between the two treatments. Bicarbonate infusion alone may alkalanize the urine and increase salicylate excretion somewhat, and potassium alone may have little effect. Potassium infusion and the presence of excess renal potassium combined with bicarbonate infusion may allow bicarbonate to alkalanize the urine to a much greater extent. The effect of the combined treatment would be greater than the sum of the individual treatments alone.

To describe the interaction effect in addition to the individual effects of each of the two treatments, an interaction term is added to the regression equation to yield:

$$z = k_1x + k_2y + k_3xy + c \qquad \text{(equation 5)}$$

where k_3 is the coefficient of the interaction term, xy. Part II, Figure 5C demonstrates the new model that includes the interaction term for the salicylate clearance experiment. The interaction term adds shape to the previously flat plane. Imagine that the regression plane is a flat piece of paper instead of cardboard. If we pick up the corner of the paper farthest from the origin and allow the paper to curve down to the origin, then this is the shape associated with inclusion of a positive interaction term in the regression equation.

The x variable coefficient representing potassium infusion has changed from +5.5 to −1, and the y coefficient has changed from 19.5 to 13 in the new model incorporating the interaction effect. The x and y coefficients have changed as a result of some expected colinearity of each with the new xy term. This is depicted schematically in Figure 1C, in which the new interaction effect is represented by area D, and it partly overlaps areas A and C. In summary, the improvement in salicylate clearance attributed to potassium infusion in Part I, Figure 5 may actually be the result of both bicarbonate infusion alone and the combined effects of bicarbonate and potassium infusion together.

SE OF THE COEFFICIENT: TWO COMPETING EFFECTS

The multivariate model depicted in Figure 5C appears to fit the data best, and a small, negative effect of potassium therapy alone is suggested by the potassium coefficient of −1. What is the effect of inclusion of the interaction term on the SE of the potassium coefficient? In one sense, the new interaction term has improved the fit of the model, as evidenced by a decrease in the SS_{res} to 60 and a corresponding increase in R^2 to 0.94. This is represented schematically in Figure 1C by the portion

of area D that does not overlap the predictor areas A and C and results in a decrease in the residual area, B. This decrease in the uncertainty of the model leads to a decrease in the SE of the other predictors such as the potassium coefficient.

The interaction term also has some collinearity with the potassium coefficient, and this is represented by the overlap of area D with area A in Figure 1C. New explanatory information is not added in this area, instead, the collinearity represents redundancy in the two predictors. This redundancy leads to uncertainty in the distribution of the association of these predictors with the outcome variable. The size of area A becomes less certain. The extent of collinearity between the predictor variable x and the other predictors y and xy is quantified by its coefficient of determination, R_{pred}^2, which is now 0.50. Then, according to equation 3, VIF $= (1/1 - 0.50) = 2.0$. This means that the SE of the coefficient relating potassium infusion to salicylate clearance is increased or inflated by a factor of $[VIF]^{1/2} = [2.0]^{1/2} = 1.41$, or 41%, when the interaction variable is included in the model.

Thus, inclusion of the collinear interaction term has two competing effects on the SE of the potassium coefficient. New information is included in the model, and this is manifested by a decrease in the total error or residuals, SS_{res}, and a decrease in the SE of the potassium predictor coefficient. Conversely, the interaction term is partly collinear with the potassium predictor, and its inclusion increases the uncertainty and SE of the potassium effect. In this example, the overall effect is that the SE of the potassium coefficient increases from 3.8 to 3.9, and the 95% CI of the potassium coefficient is −11.8 to 9.8 after inclusion of the interaction term. In general, the SE of existing predictor coefficients may increase or decrease with the inclusion of a new colinear predictor in the regression model. The total effect depends on the relative amount of new explanatory information added versus the extent of collinearity and redundancy the new predictor displays with each of the previously existing predictor variables.

CHOOSING PROPER VARIABLES

The art and science of building the multiple regression model require active collaboration between the clinical or laboratory scientist and the statistician. The general goal should be the inclusion of all predictor variables that add substantial independent information while avoiding excessive collinearity or overlap. When multicollinearity exists, which usually is the case in medical research, the predictor variable coefficients can be biased or their SEs can be increased by inclusion of either too few or too many variables in the regression model.

Consider the following examples. Suppose one is trying to predict the likelihood of myocardial infarction (MI) in patients who present to the emergency department. The researcher may tabulate a list of the predictor variables, each with its own individual univariate statistic, such as an odds ratio (OR) and an associated 95% CI. Let us also assume that some of the predictors are positively correlated, such as the degree of elevation of the electrocardiogram (ECG) ST segment and the serum troponin level. Each individual univariate statistic will attribute all of the overlap in predictive value with the other predictors to the particular predictor of interest. As we move down the list of predictors, the entire overlap with the other variables is attributed to each individual predictor variable in turn. Consequently, although each univariate statistic is numerically correct, it is biased toward a higher value, and the association of the predictor variables as a whole with the outcome variable will seem larger than it truly is. When there is positive collinearity among positive predictor variables, each predictor coefficient will be highest when viewed in a univariate model.

Conversely, consider the same situation, but in this case, an excessive number of collinear variables are included in a single multivariate logistic model in which the outcome is the probability of an MI, which varies between 0 and 1. In addition to a history of smoking, the investigators also measured the history of coughing, frequency of visitation to a drinking establishment, and dental coloration. It is expected that all of these variables would display positive collinearity with the smoking history. Based on our experience and previous science, we know that, fundamentally, smoking might lead to an increase in the likelihood of an acute MI, but the other variables likely would not.

Inclusion of multiple extraneous collinear variables that are not associated with the outcome variable in the multiple regression model would not be expected to bias the predictor variable of interest, smoking history. In practice, the situation is often complex and these variables may actually be associated with the outcome for a variety of unanticipated reasons. For example, history of coughing may be a better measure of cigarette use than the reported smoking history, and bar patrons may suffer from second-hand smoke. In this situation, inclusion of these variables in the model may alter the history of smoking coefficient. Regardless of any bias that may occur, inclusion of nonpredictive collinear variables will generally inflate the variance and uncertainty of the predictor of interest. Finally, it is interesting to consider the possible consequences of removing the smoking history variable from the model. The model might still demonstrate an excellent R^2 with the remaining extraneous variables, which could now be viewed as positively biased.

There are numerous manual and automated methods for building multiple linear regression models.[1,2] As one can appreciate from the examples above, methods that rely on univariate screening[3,4] and

automated stepwise techniques[5] are prone to bias and scientific error. The use of all automated predictor variable selection methods is discouraged. Even when using manual methods to choose predictor variables, there is active debate regarding whether investigators should generally use the smallest number of predictor variables that yield a good fit to the data, or whether all scientifically credible predictor variables should be included to maximize the predictive value of the model.[6,7] Perhaps the best and simplest approach is to design an experiment to collect data on the most important known and suspected fundamental predictor variables based on current scientific knowledge. The investigator then can examine the results to confirm that the data satisfy the assumptions of the linear model, and that all of the included predictor variables contribute unique information and do not demonstrate excessive multicollinearity. Regardless of the approach used to construct the regression model, the results should eventually be validated using either a similar, but separate, set of data, or using another method such as cross-validation.[8,9]

MULTICOLLINEARITY: DEALING WITH IT

Sometimes multicollinearity among important predictors is inevitable. How can the investigator deal with this? One approach would be to collect more data to decrease collinearity. Sometimes, however, this may be difficult or impossible. For example, we know that patients with elevated troponin are more likely to have an elevated ST segment on ECG, and finding enough patients with an elevated troponin and normal ST segment may be difficult. Collinear predictors may be combined to form a summary variable or score. The Goldman criteria are an example of a combined score composed of multiple collinear risk factors for heart disease used to evaluate cardiac risk in patients undergoing noncardiac surgical procedures.[10] Another approach would be to study different predictors that may be more fundamental and display less collinearity. Instead of measuring patient age and history of hypertension and elevated cholesterol as predictors of an MI or angina, the investigator might instead measure the degree of luminal narrowing on cardiac catheterization. This is a variation of the concept used in the technique of principal component analysis.[11] Finally, there are modifications of the least-squares approach such as ridge regression. This analytic technique introduces bias in the estimates of highly collinear variable coefficients in exchange for a decrease in their uncertainty.[12–14]

CONCLUSIONS

Most problems in clinical medicine are multivariate. Consequently, a univariate approach in research analysis often is flawed and may produce quantitatively or qualitatively incorrect predictor coefficients, and incorrect conclusions with inference testing. A multivariate approach often is required. Multiple linear regression is a useful technique for modeling many phenomena in medical research. For data sets that meet the necessary assumptions, it offers a well-developed model that usually can be solved exactly, yielding estimates of the predictor variable coefficients and their SE or uncertainty. This can lead to a better understanding of the relative effects and importance of the predictors of interest, and allows the investigator to prognosticate the outcome of future data. Applications in clinical medicine include models to determine diagnosis, prognosis, and therapeutic outcomes.

The author thanks Doctors Elaine Rabin, Lillian Oshva, Lisa Campanella, Ellen Weber, Lewis Goldfrank, and the statistical editors of *Academic Emergency Medicine* for their thoughtful encouragement, suggestions, challenges, and support.

References

1. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996; 15:361–87.
2. Nick TG, Hardin JM. Regression modeling strategies: an illustrative case study from medical rehabilitation outcomes research. Am J Occup Ther. 1999; 53:469–70.
3. Sun GW, Shook TL, Kay GL. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. J Clin Epidemiol. 1996; 49:907–16.
4. Cardo DM, Culver DH, Ciesielski CA, et al. A case–control study of HIV seroconversion in health care workers after percutaneous exposure. N Engl J Med. 1997; 337:1485–90.
5. Derksen S, Keselman HJ. Backward, forward and stepwise automated subset-selection algorithms—frequency of obtaining authentic and noise variables. Br J Mathemat Stat Psychol. 1992; 45:265–82.
6. Wears RL, Lewis RJ. Statistical models and Occam's razor. Acad Emerg Med. 1999; 6:93–4.
7. Kepermann N, Willits N. In response to ''Statistical models and Occam's razor.'' Acad Emerg Med. 2000; 7:100–3.
8. Efron B. Computer-intensive methods in statistics. Sci Am. 1983; 248:116–30.
9. Efron B, Gong G. A leisurely look at the bootstrap, the jackknife, and cross-validation. Am Stat. 1983; 37:36–48.
10. Goldman L, Caldera DL, Nussbaum SR, Multifactorial index of cardiac risk in noncardiac surgical patients. N Engl J Med. 1977; 297:845–50.
11. Using principal components to diagnose and treat multicollinearity. In: Glantz SA, Slinker BK. Primer of Applied Regression and Analysis of Variance, 2nd ed. New York, NY: McGraw-Hill, 2001, pp. 219-37.
12. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970; 12:55–67.
13. Hoerl AE, Kennard RW. Ridge regression: applications to nonorthogonal problems. Technometrics 1970; 12:69–82.
14. Ridge regression. In: Myers RH. Classical and Modern Regression with Applications, 2nd ed. Pacific Grove, CA: Duxbury Press, 1990, pp. 392-411.