Advanced Statistics: Linear Regression, Part I: Simple Linear Regression

Keith A. Marill, MD

Abstract

Simple linear regression is a mathematical technique used to model the relationship between a single independent predictor variable and a single dependent outcome variable. In this, the first of a two-part series exploring concepts in linear regression analysis, the four fundamental assumptions and the mechanics of simple linear regression are reviewed. The most common technique used to derive the regression line, the method of least squares, is described. The reader will be acquainted with other important concepts in simple linear regression, including: variable transformations, dummy variables, relationship to inference testing, and leverage. Simplified clinical examples with small datasets and graphic models are used to illustrate the points. This will provide a foundation for the second article in this series: a discussion of multiple linear regression, in which there are multiple predictor variables. **Key words:** regression analysis; linear models; least-squares analysis; statistics; models; statistical; epidemiologic methods. ACADEMIC EMERGENCY MEDICINE 2004; 11:87–93.

Linear regression is a mathematical technique that attempts to describe the relationship between two or more variables with a linear or straight-line function. Based on an analysis of the available data or sample, the technique also can be used to draw inferences about a larger population or data set, or to make predictions about future data. Simple linear regression is a subtype of linear regression in which there is a single outcome or dependent variable and a single predictor or independent variable.

Linear regression is a popular technique because many phenomena of interest have a linear relationship, and the technique is able to demonstrate mathematically and visually the relationships between clinically important variables. The linear equation is inherently simple and elegant, and a unique solution usually exists. Furthermore, nonlinear terms can be introduced into the linear framework as needed, primarily to improve the fit to the data and to satisfy the basic assumptions of the model. When a data set cannot be properly described using a linear regression approach, a different and mathematically more complex technique termed ''nonlinear regression'' may be used.

In this article, four clinical questions with associated small theoretical data sets are introduced to illustrate the major points. The clinical questions focus on the treatment of diabetic ketoacidosis (DKA) and aspirin overdose. The four data sets are illustrated graphically in Figures 2 through 5. We will return to these data sets later in the corresponding figures in the second article when a multiple regression approach is taken. The raw data for these examples are listed in Table 1.

FUNDAMENTAL ASSUMPTIONS

Simple linear regression uses the equation for a line to model the relationship between two variables. If z is the outcome variable and x is the predictor variable, then:

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

where k is a coefficient that represents the slope of the linear relationship between the variables x and z, and c is a constant. The constant c is termed the ''z intercept'' because this is the value of z where $x = 0$ and the regression line crosses the z axis. Consider Figure 1, which contains a data set of 12 points and the associated linear regression line, $z = 2x + 1$. This example can be used to review the four necessary assumptions in performing simple linear regression and inference testing:

1. There is some linear relationship between the predictor and outcome variable. As the values of the points increase along the *x*-axis, their values along the *y*-axis also increase. The cloud of points seems to center around a straight line rather than a curve or other shape.

From the Division of Emergency Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA.
Received July 24, 2001; revisions received July 9, 2002, and April 21, 2003; accepted September 10, 2003.
Series editor: 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 II follows on page 94.
doi:10.1197/S1069-6563(03)00600-6

TABLE 1. Data

	X	Y	Z	Log Z
Figure 2	4	4	4	
	8	8	8	
Figure 3	20	4	2	
	30	4	1	
	30	8	6	
Figure 4	4	31	2	0.3
	4	39	8	0.9
	8	28	3	0.48
	8	42	32	1.51
Figure 5	0	0	1	
	0	0	5	
	0	1	13	
	0	1	19	
	1	0	1	
	1	0	3	
	1	1	24	
	1	1	32	

	X	Z	Residual
Figure 7	2	2.5	0.06
	3	3	−0.11
	4	4	0.22
	6	2	−3.13
	7	8	2.2
	8	8	1.53
	9	9	1.86
	10	3	−4.81
	10	10	2.19

2. The variation around the regression line is constant (homoscedasticity). Some points may be farther from the regression line than others. Homoscedasticity means that as the eye moves laterally along the x-axis, the average variation of the points from the regression line stays about the same.

3. The variation of the data around the regression line follows a normal distribution at all values of the predictor variable. If one examines the data points at any particular value of x, they will form a bell-shaped or normal curve around the value of the regression line at that point. The majority of points will be close to the regression line, and fewer points will be farther away.

4. The deviation of each data point from the regression line is independent of the deviation of the other data points. The value of one point and its relationship to the regression line has no relationship or bearing on the value of another point in the dataset.

If these four assumptions are met, then the model is considered valid and optimal. Now the methodology of simple linear regression is examined using some specific simplified clinical research scenarios.

THE METHOD OF LEAST SQUARES

Consider a small, retrospective study in which the investigator reviewed the resolution rate of acidosis in DKA as a function of the intensity of intravenous (IV) insulin therapy. There were two patients: one received 4 units per hour of IV insulin, and the other received 8 units per hour. Examining Figure 2, we can see that, after four hours of therapy, the serum bicarbonate increased by 4 mEq/L in the first patient and 8 mEq/L in the second. The regression equation $z = 1x$ perfectly describes the relationship in this small data set. In this case, $z = 0$ at the z intercept where $x = 0$, and thus the constant c in Equation 1 equals zero. Each unit per hour of insulin therapy seems to be associated with a 1-mEq/L increase in the serum bicarbonate level after four hours of therapy.

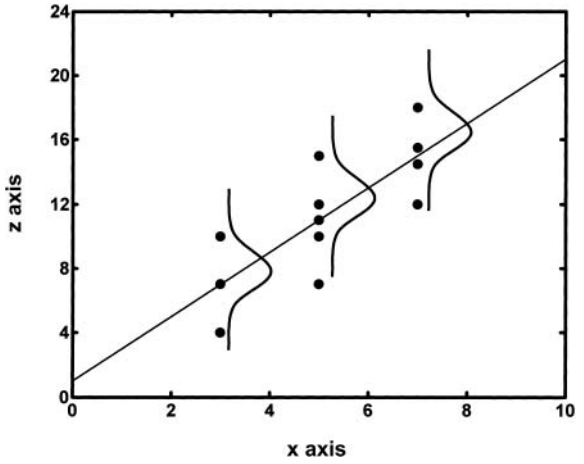

Figure 1. The four assumptions of linear regression.

The relationship in the data is usually not perfectly linear. In a separate observational study, the investigator studied the improvement in acidosis after four hours in three patients with DKA as a function of their initial respiratory rate. The investigator hypothesized that patients who were able to sustain an elevated respiratory rate might be able to "hyperventilate their way out" of DKA. Figure 3 depicts the data and the associated regression line. The slope is positive, which

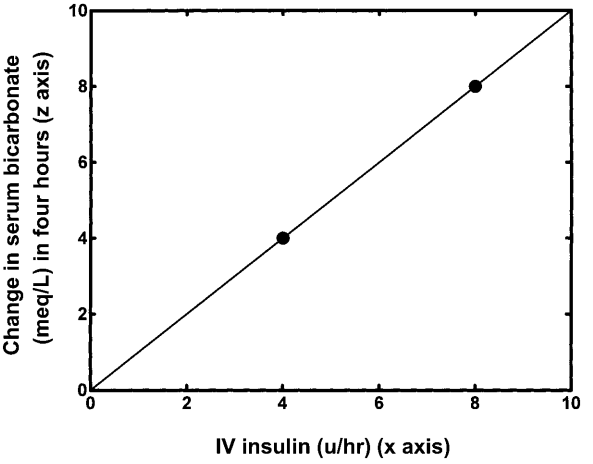

Figure 2. $z = 1x$, $R^2 = 1.0$. IV = intravenous.

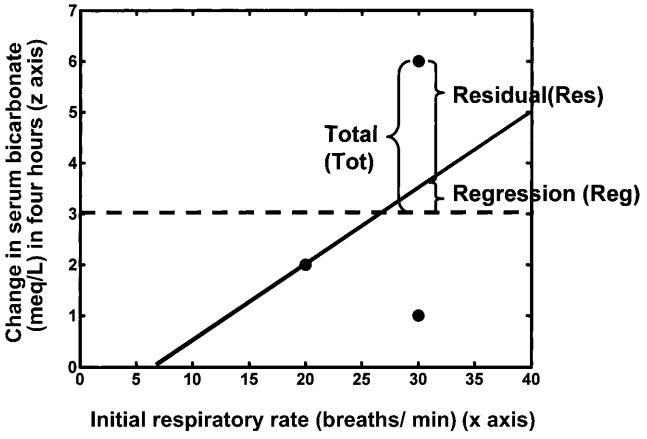

Figure 3. $z = 0.15x - 1$, $R^2 = 0.11$.

suggests that patients who present with higher respiratory rates do seem to improve more rapidly. In this case, the regression line does not touch all of the data points, but rather is the "best fit" for the data. This suggests that whereas the initial respiratory rate may "explain" part of the improvement in the observed acidosis, there may be other clinical and experimental factors involved. Perhaps the intensity of insulin therapy also is associated with the improvement in acidosis as we saw in Figure 2, or perhaps our measurements of the respiratory rate or serum bicarbonate are imprecise and there is a large amount of error. How is the best-fit regression line determined, and how can we quantitate the relative strength of the association between the predictor variable and the outcome?

The method of least squares is most commonly used to find the best fit in linear regression. In Figure 3, there are three data points with outcome values 1, 2, and 6. The mean of the outcome values is 3, and a broken straight line representing the mean outcome value for the three points has been drawn horizontally for the sample. Each data point has a certain variation from the mean outcome value. A regression line also has been drawn across the figure. The variation of each data point from the mean outcome can be represented by the sum of the vertical distance from the mean outcome line to the regression line (reg), and the distance from the regression line to the data point. This latter distance is referred to as the "residual" (res). The total distance (tot) of each point from the mean outcome value thus can be apportioned between the regression and the residual. The goal in formulating the regression line is to maximize the portion attributed to the regression and to minimize the residual for all of the data points. To be mathematically precise, the sum of all of the residuals squared is minimized—thus the method of least squares.

In linear regression, it often is necessary to calculate the sum of a value for all of the points in the data set. To do this in our example, it may be desirable to label each patient enrolled with a unique number starting with 1 for the first patient and ending with the highest number for the last patient. If we use the letter p to denote the patient number, then in this example, there are three patients with values p = 1, 2, 3. The Greek symbol sigma, $\sum$, is used to denote the summation of a mathematical value. Thus, if we wish to sum the residual distance for all of the patients, we would write $\sum_{p=1}^{3}$res, where the term p = 1 below the sigma denotes the first patient in the series, and the number 3 above the sigma denotes the last patient whose value is included in the summation. In this example, $\sum_{p=1}^{3}\text{res} = 0 + 2.5 + 2.5 = 5$, where each of the three numbers 0, 2.5, and 2.5 represents the residual distance from the regression line to the data point for each of the three patients, respectively.

In the method of least squares, our interest is minimizing the sum of the residuals squared, $\sum_{p=1}^{3}(\text{res})^2$. In this example, $\sum_{p=1}^{3}(\text{res})^2 = 0^2 + (2.5)^2 + (2.5)^2 = 12.5$. This summation of squared terms often is abbreviated as $\sum_{p=1}^{3}(\text{res})^2 = SS_{res}$. Similarly, the summation of the distances from the mean line to the regression line, and from the mean line to the data points can be squared. For this example, these values would be depicted and abbreviated as follows: $\sum_{p=1}^{3}(\text{reg})^2 = SS_{reg} = (1)^2 + (0.5)^2 + (0.5)^2 = 1.5$ and $\sum_{p=1}^{3}(\text{tot})^2 = SS_{tot} = 1^2 + (2)^2 + (3)^2 = 14$.

The next step is to find the correct coefficient k and constant c in the regression equation, which produce a regression line that has the smallest possible SS_{res}. The problem is solved using two differential equations,[1,2] and two interesting findings occur as a direct result of this process. First, except for the unusual situation in which all of the data have the same value of x, there is always a unique solution or single line that is best. Second, the line that is found to be best always satisfies the following equation[1,2]:

$$SS_{tot} = SS_{reg} + SS_{res} \qquad \text{(equation 2)}$$

Thus, the sum of all the squared distances from the mean line, SS_{tot}, can be apportioned between regression and residual components. If Equation 2 is divided by SS_{tot}, then we have:

$$1 = \frac{SS_{reg}}{SS_{tot}} + \frac{SS_{res}}{SS_{tot}} \qquad \text{(equation 3)}$$

The ratio SS_{reg}/SS_{tot} can be viewed as the portion of the variation in the outcome variable that can be attributed to, or explained by, the regression model, and the other portion, SS_{res}/SS_{tot}, can be considered the error or unexplained portion. Together, the two portions add up to one.

It is also true that:

$$\frac{SS_{reg}}{SS_{tot}} = R^2 \qquad \text{(equation 4)}$$

where R equals the Pearson correlation coefficient. So R^2 can be used as one measure of the strength of the linear relationship between the two variables and the validity of the linear model. For example, in Figure 2, $R^2 = 1$ because there is a perfect linear relationship, and if there is no linear relationship, then $R^2 = 0$. In Figure 3, $R^2 = 0.11$ which means the residual portion, SS_{res}/SS_{tot}, is: $1 - 0.11 = 0.89$. The variation in the initial respiratory rate only seems to "explain" 11/100, or 11% of the improvement in the serum bicarbonate. This suggests that there are other important predictors of acidosis resolution besides the initial respiratory rate, or that there is a large amount of error in the measurements.

VARIABLE TRANSFORMATIONS

In addition to studying the resolution of acidosis in patients with DKA, the investigator also is interested in the length of the patient stay in the intensive care unit (ICU). Perhaps the required intensity of insulin therapy also can predict the duration of subsequent ICU admission. Four patients who required ICU admission were retrospectively enrolled, and the duration of their ICU stay is plotted as a function of the initial intensity of insulin therapy in Figure 4A. There seems to be an association of the intensity of insulin therapy with the duration of the ICU admission, but there also is a problem.

Notice the pattern of the data points around the regression lines in Figures 1 and 4A. As previously noted, in Figure 1, there is homoscedasticity because the spacing of the points around the regression line stays about the same as the eye moves along the line. In Figure 4A, the distance of the points from the regression line increases to the right, suggesting the shape of a funnel on its side. This violates the assumption of homoscedasticity. When the native data do not satisfy this condition, variable transforms may be used to produce a regression equation that is satisfactory.

Although linear regression is, by definition, a process of linear modeling, it is possible to introduce nonlinear terms to the linear mathematical framework by transforming variables. The primary motivation to perform variable transforms is to improve the regression fit and to satisfy the necessary regression assumptions such as homoscedasticity. Logarithmic transforms are particularly useful in this regard because they differentially compress the spread in the data at high and low values of the transformed variable. In our example, we define a new variable q, where q is the log of the length of ICU admission, or

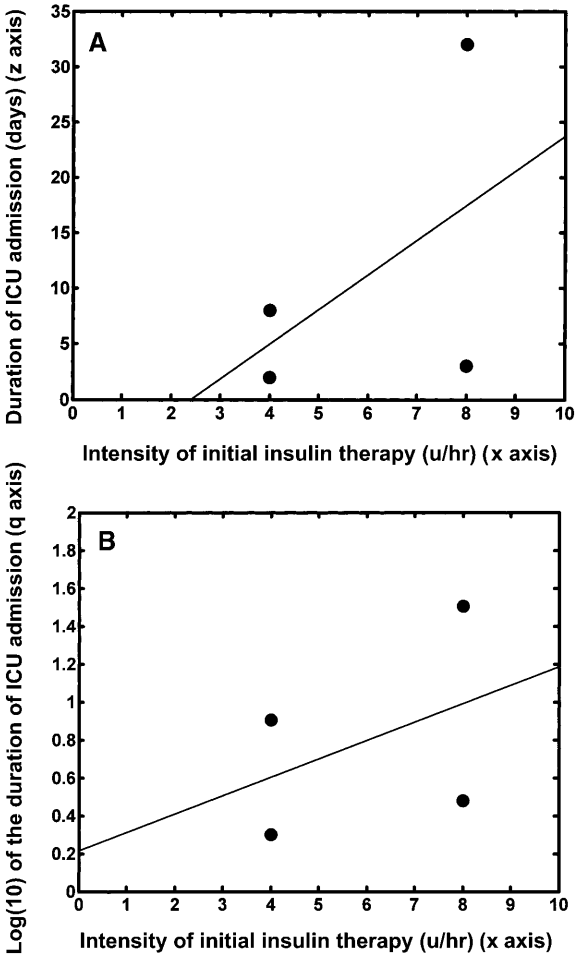

Figure 4. (A) $z = 3.125x - 7.5$. **(B)** $q = 0.097x + 0.213$, where $q = \log_{10} z$ $R^2 = 0.18$. ICU = intensive care unit.

$q = \log z$. Thus, if the length of ICU stay is 32 days, then $z = 32$ and $q = \log(32) = 1.5$. Next, we try a new transformed regression equation:

$$\log z = q = kx + c$$

which is graphed in Figure 4B. We can see this relationship satisfies the necessary linear regression assumptions because the variation of the data points around the regression line is fairly constant moving along the line. The large spread in the data due to the patient who stayed in the ICU for 32 days has been relatively compressed down. Thus, the regression equation that includes a logarithmic transform is a better fit than the original linear equation. Q can be modeled as a linear function of x, and the results eventually can be transformed back to the variable z.

In a similar fashion, a linear regression equation also can be developed after transforming the independent predictor variable. For example, let $z = c + kx$ as in simple linear regression, and now let $x = s^2$, so that x is a transform of the variable s. Although the equation remains linear with respect to x, it is

now a quadratic equation with respect to the underlying predictor variable s: $z = c + ks^2$. In summary, transforms allow one to extend the use of the well-developed mathematical framework in linear regression to model some data sets, which otherwise would not satisfy the necessary assumptions. Extreme care must be taken when performing and interpreting such transformations because the results often are not intuitive. In our example, we have now modeled the log of the length of the ICU stay as a function of the initial intensity of insulin therapy, but the log of the length of the ICU stay is a number with no units of measure and no clear clinical meaning.

DUMMY VARIABLES AND INFERENCE TESTING

Prescott et al. demonstrated that salicylate clearance can be enhanced in the overdose setting with both bicarbonate infusion and forced diuresis, and alkalinization of the urine is an important factor in enhancing excretion.[3] It also is known that alkalinization of the urine is difficult to achieve in the setting of hypokalemia, because the renal tubules will retain potassium in lieu of hydrogen ions. An investigator postulated that a potassium infusion may be helpful to clear salicylate in acute aspirin overdose regardless of the initial serum potassium level. A small study of salicylate overdose was designed with eight laboratory animals divided into four groups of two. Each group of animals was assigned to receive an identical acute salicylate overdose, subsequent infusion of a standardized IV fluid regimen, and one of four treatments: placebo, potassium, bicarbonate, or both agents.

The investigator first analyzed the results with respect to the infusion of potassium, and a scattergram of the eight data points with an associated linear regression line is depicted in Figure 5. Notice that on the *x*-axis, the two values are 0 or 1, and these are associated with the absence or presence of a potassium infusion, respectively. In this situation, the variable x was used as a ''dummy variable.'' To use a dummy variable in simple linear regression, the sample is divided according to the presence or absence of a particular characteristic. A value of $x = 1$ is associated with the presence of the characteristic, and a value of $x = 0$ is associated with its absence. There are no data with values of x outside of 0 and 1. The use of dummy variables often is simpler and preferred for analysis and inference testing. In this example, the investigator could have used the total amount of potassium infused in milliequivalent units for the x-axis, but the dummy variable was preferred. For other data that have no associated units of measure, such as the answer to a ''yes'' or ''no'' question, a dummy variable must be used.

The linear regression line drawn in Figure 5 clearly has a slope different than zero. Taking a statistical

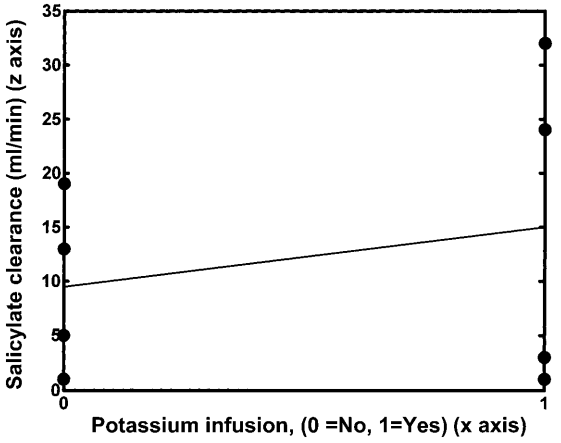

Figure 5. $z = 5.5x + 9.5$, $R^2 = 0.06$.

approach, however, we may ask how certain are we of the derived value of this slope? Based on this small sample, can we conclude that potassium infusion increases the clearance of salicylate in this animal species? The slope or coefficient in the regression has a value, and an associated uncertainty or standard error (SE). The SE of the coefficient is the standard deviation of the possible values of the coefficient in the theoretical population from which this particular experimental sample of animals was drawn. The formula for the standard error of the coefficient is:

$$\mathrm{SE\ (coefficient)} = \left[\frac{SS_{res}}{(n-2)\sum_1^n (x - x_{mean})^2}\right]^{1/2} \quad \text{(equation 5)}$$

where n is the total number of subjects in the sample, x is the value of x for each of the subjects, and x_{mean} is the mean value of x for all of the subjects. In this example, $n = 8$, the values of x are 0, 0, 0, 0, 1, 1, 1, 1, and $x_{mean} = 0.5$. The SE of the coefficient is high if the residual values and SS_{res} are high and there is a large amount of variation in the data that is not accounted for in the regression. Conversely, the SE is relatively low if n is high and there are a lot of subjects, and if the values of x are widely spread out along the x-axis. A wider base of data along the x-axis increases the certainty of the regression slope.

If the value of the slope is large and its SE is small, then we can be confident that it is not only different from zero in this particular sample, but it is likely to be different from zero if we repeat the experiment with another sample of animals. Thus, regression analysis can be used for inference testing. This particular case represents the situation of a Student's t-test. The p-value represents the probability of obtaining a slope whose absolute value is equal to or greater than the one actually obtained under the null hypothesis that the true slope is zero. If the p

value is small, then the result obtained would be unlikely to occur under the null hypothesis, and the null hypothesis would be rejected. In this example, p = 0.56, so the null hypothesis is not rejected. Alternatively, the 95% confidence interval (95% CI) of the slope or predictor coefficient can be calculated using the formula:

$$95\%\,\text{CI for the coefficient}\,k = [k - (t)(SE), k + (t)(SE)] \quad \text{(equation 6)}$$

where k and SE are the value and SE of the coefficient of interest, and t is the t value with p = 0.05 and the appropriate degrees of freedom. The 95% CI for the potassium infusion coefficient is –15.8 to 26.8. Because the 95% CI spans zero, it is not concluded that the potassium coefficient is significantly different from zero.

It may be instructive to visualize a schematic representation of simple linear regression as in Figure 6. Let the central circle represent the total variation in the outcome variable, z, and the upper lateral circle represent the total variation in the predictor variable, x. The area of overlap of the two circles, area A, represents the portion of the outcome that is "explained" by the regression of the predictor variable. Then, area B, the remaining area in the outcome circle outside of the overlap, represents the variability in the outcome due to all other possible factors that are not explained by the predictor variable. Area B represents the residuals in the regression. In this schematic, the correlation coefficient, $R^2 = A/A + B$, represents the portion of the total outcome explained by the predictor variable. It is interesting to note that the SE of the coefficient is a function of the residuals, or area B, but not the regression, area A. No matter how large area A is, and how much of the outcome is associated with the predictor variable, the uncertainty in the regression coefficient is primarily a function of the portion of the outcome that the regression doesn't explain, the residual area B. When a t-test is performed, a comparison is made between the relative sizes of area A and B after adjusting for the total number of data points, *n*. If the regression portion is relatively large compared with the residual portion, then the null hypothesis that there is no relationship between the predictor variable and the outcome is rejected. It is concluded that there is a linear relationship between the predictor and the outcome.

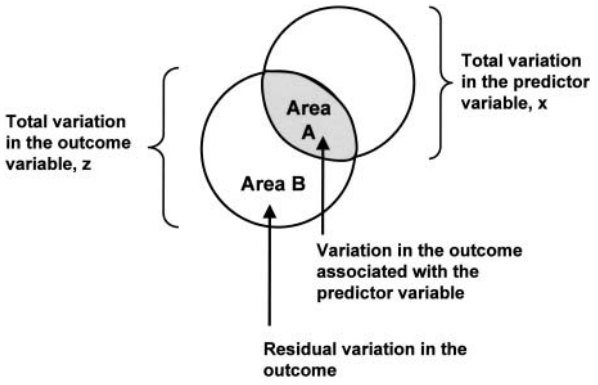

Figure 6. Simple linear regression schematic.

LEVERAGE

When a regression model is constructed, it is critical to examine the data and fit of the model visually. In our first example, the investigator collected data on two patients with DKA to assess the relationship between the intensity of insulin therapy and the resolution of metabolic acidosis. Suppose the investigator decided to confirm the initial result with a larger sample of patients. Figure 7A shows a graph of a larger, identical study of patients with DKA. The increase in serum bicarbonate after four hours of IV insulin therapy is plotted as a function of the intensity of insulin therapy, and the associated linear regression line has been

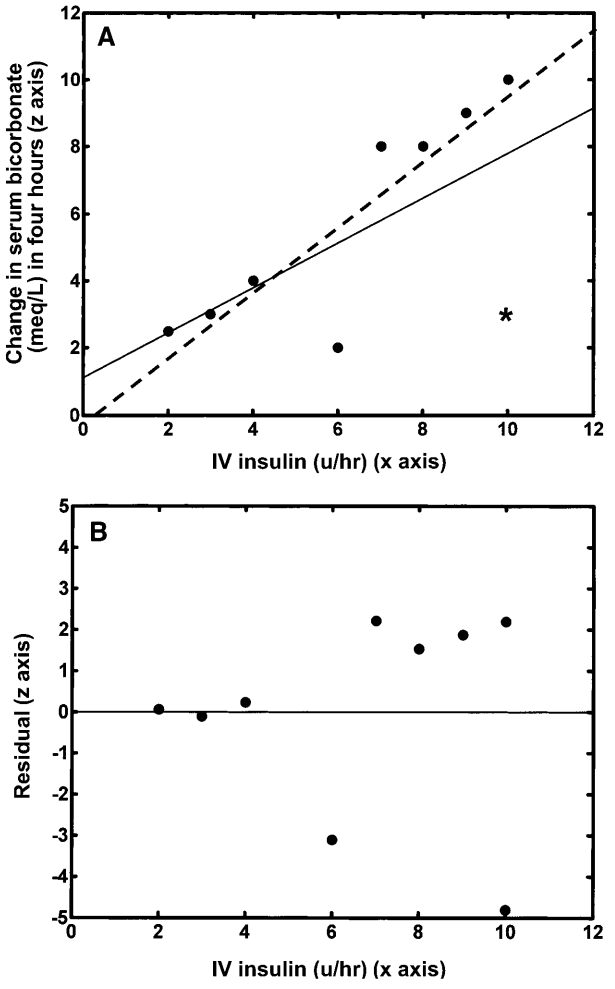

Figure 7. (A) Regression lines with (solid) and without (broken) the leverage point (10,3). **(B)** Residuals versus independent predictor variable, intravenous (IV) insulin therapy.

added. To focus on and to further assess the relationship between the data points and the regression line, it often is instructive to plot the residual differences between the data points and the regression line as a function of the predictor variable, x. Figure 7B demonstrates the ''plot of the residuals'' for the data originally graphed in Figure 7A. In Figure 7B, the line $y = 0$ now represents the regression line, and each data point is represented by its residual distance from this line.

Note that in Figures 7A and 7B there are two outlier points, one in the middle of the data set, and one toward the edge. Both of these outliers can have a disproportionate effect on the regression model when compared with the other data, but outliers on the edge may be more troublesome. The middle outlier will tend to raise or lower the entire regression line by some amount that will change the value of the constant c. It will not, however, greatly affect the slope of the regression line. The lateral outlier, however, has tilted the line down toward its side. It is exerting leverage on the regression and has a large effect in decreasing the value of the slope, k. Note the increase in the slope of the regression line when this single point is removed from the analysis in Figure 7A. The investigator should recheck this outlying data point to be sure that it is not an error, and perhaps collect more data in its area to see if it represents a trend.

Regression tools used to assess the amount of leverage exerted by an individual data point include the studentized residual and Cook's distance.[4,5] The studentized residual of a particular point is the original residual value that has been standardized and adjusted for the leverage exhibited by that point on the regression. A residual is standardized when it is divided by the standard deviation of all the residuals in the dataset. Cook's distance quantifies the combined change in the slope and z intercept as a whole when the point in question is removed from the analysis. If the studentized residual and Cook's distance are relatively large for a given point in a data set, then that point may be exerting undue leverage on the regression model.

CONCLUSIONS

Simple linear regression is a mathematical technique used to create a linear model of the association between a single predictor variable and an outcome measurement. The method of least squares is commonly used to determine the slope and constant of the regression equation, and four assumptions must be satisfied to produce a valid model. The data should always be graphed and visually inspected to ensure that the assumptions are satisfied, and transformations of either the predictor or outcome variables can be performed as necessary. The researcher should also check for and investigate any outlying data points, particularly those that exert undue leverage. A valid model can be used to perform univariate inference testing such as Student's t-test, and predictions about future data can be made.

References

1. Glantz SA, Slinker BK. The first step: understanding simple linear regression. In: Primer of Applied Regression and Analysis of Variance, 2nd ed. New York, NY: McGraw-Hill, 2001, pp 10–49.
2. Draper NR, Smith H. Fitting a straight line by least squares. In: Applied Regression Analysis, 3rd ed. New York, NY: John Wiley & Sons, 1998, pp 15–46.
3. Prescott LF, Balali-Mood M, Critchley JA, Johnstone AF, Proudfoot AT. Diuresis or urinary alkalinisation for salicylate poisoning? Br Med J. 1982; 285:1383–6.
4. Glantz SA, Slinker BK. Primer of Applied Regression and Analysis of Variance, 2nd edition. New York, NY: McGraw-Hill, 2001, pp 133–44.
5. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. Applied Regression Analysis and Other Multivariable Methods, 3rd ed. Pacific Grove, CA: Duxbury Press, 1998, pp 212–37.