Factors that Influence the Value of the Coefficient of Determination in Simple Linear and Nonlinear Regression Models

J. A. Cornell and R. D. Berger

Statistics Department and Plant Pathology Department, respectively, University of Florida, Gainesville 32611.

We thank T. Davoli for preparation of the illustrations.
Florida Agricultural Experiment Stations Journal Series Paper 6709.
Accepted for publication 28 February 1986.

ABSTRACT

Cornell, J. A., and Berger, R. D. 1987. Factors that influence the value of the coefficient of determination in simple linear and nonlinear regression models. Phytopathology 77:63-70.

In the fitting of linear regression equations, the coefficient of determination (R^2) is one of the most widely used statistics to assess the goodness-of-fit of the equation. Its value, however, is affected by several factors, some of which are associated more closely with the data collection scheme or the experimental design than with how close the regression equation actually fits the observations. These design factors are: the range of values of the independent variable (X), the arrangement of X values within the range, the number of replicate observations (Y), and the variation among the Y values at each value of X. Another little-known fact is the effect on R^2 of the ratio of the slope of the fitted equation to the estimated standard error of the observations. In nonlinear model fitting, the value of R^2 is best determined by calculating the proportion of the total variation in the observations that cannot be explained by the fitted model and subtracting this proportion from one. Several statistics that are analogous to the standard formula for R^2 in the linear regression case are given and determined to be inappropriate in the nonlinear case. The use of R^2 alone as a model-fitting criterion is often risky and other statistics should be used to assess the goodness of the model when responses from quantitative treatments are analyzed by regression techniques.

Additional key words: coefficient of determination, residuals, standard error.

Linear regression is a commonly used statistical analysis in plant pathology. It has been used, for example, to determine inoculum density/disease intensity relationships (5), survival of pathogens over time (16), growth, sporulation, and infection of pathogens under different environments (9,10), model testing (8), and disease intensity/crop loss relationships (1). Nonlinear regression is used frequently to fit disease proportions over time to various growth models (2,12), disease prediction from environmental parameters (8), crop loss estimation from disease intensity (13), growth, sporulation, and infection of a pathogen with temperature (3), and the relationship of disease intensity to size of experimental plots (7) or to calcium carbonate concentration (4).

For both linear and nonlinear regression, the coefficient of determination is possibly the statistic used most often to assess the goodness-of-fit of empirical models fitted to data. This is because the value of R^2 is provided by every current computer program for regression analysis. Nearly every published article, in which regression analysis was performed, lists the R^2 associated with each equation fitted. The appropriateness of R^2 to assess the goodness of a fitted model is under investigation (11) and, until alternative measures are suggested, it is imperative that the meaning of R^2 and the factors that influence it be understood.

In the fitting of regression models, researchers occasionally raise one or the other of the following two questions when they discover the value of R^2 is extremely low for their model: Why is R^2 so low when the equation seems to fit the data very well? What is the appropriate method to calculate R^2 to determine the goodness-of-fit of a nonlinear model, e.g., exponential models or power functions?

In this article, to address the first question we identify some of the factors in a data set that lower the value of R^2. Our purpose in singling out these factors is twofold: first, to acquaint users of regression techniques of the potential pitfalls that result from relying too heavily on R^2 as a model closeness criterion, and second, to point out that corrective actions to obtain a high R^2 value often are contrary to the principles of good experimental design.

We shall answer the second question by listing several analogous statistics to R^2 that are sometimes provided by current computer programs for regression analysis.

METHODS

Artificial data sets were generated and linear or nonlinear models were fitted by least-squares regression either by hand calculation or by the Statistical Analysis System package (15), using the facilities of the Northeast Regional Data Center of the State University System of Florida in Gainesville.

RESULTS AND DISCUSSION

Factors that affect R^2 in the fitting of simple linear regression equations. In the simple linear regression equation, $Y_i = a + bX_i + e_i$, Y_i is the i^{th} observation of the dependent variable and X_i is the value of the independent variable at which Y_i is observed. The quantities a and b are unknown parameters that represent the intercept and slope of the regression line, respectively. The random error associated with Y_i is termed e_i. The usual assumptions regarding the errors are, that in a population of N values of Y_i, the random errors (e_i) have zero mean, a common variance (σ_e^2), and are independent of one another.

To illustrate the calculations that are required in the analysis of a fitted regression equation, the simple linear regression equation ($Y_i = a + bX_i + e_i$) is fitted to each of two data sets denoted as A and B. The observations (Y_i) are the same in data sets A and B but the ranges of X_i are different (Table 1). The plots of the fitted regression equations are shown in Figure 1.

Included among the entries in Table 1 are the predicted responses ($\hat{Y}_i$) at each X_i obtained with the fitted regression equation. The quantity $Y_i - \hat{Y}_i$, represents the difference between the observed value (Y_i) and the predicted value ($\hat{Y}_i$) at X_i, and this difference is called the residual corresponding to the i^{th} observation. The larger the values of the residuals, the less confident one feels about how well the estimated equation fits the observed values. A numerical

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. § 1734 solely to indicate this fact.

measure, therefore, of how well the model actually fits the data is the variance of the residuals, $s_e^2 = \text{SSE}/(N-2)$. When the residuals are large, s_e^2 is large. Also, the individual residuals can be plotted against the values of X or the values of $\hat{Y}_i$ to ascertain if the linear model is indeed the appropriate choice. In both plots, if the model is correct, the values of the residuals will exhibit random scatter about the line, $Y - \hat{Y} = 0$, and the approximate scatter is uniform for all values of X and/or $\hat{Y}_i$.

The positive square root of s_e^2 (i.e., s_e) is called the estimated standard error of the Y_i values about the regression line (6), that is, $s_e = \sqrt{\Sigma(Y_i - \hat{Y}_i)^2/(N-2)}$ is a function of the residuals, $Y_i - \hat{Y}_i$, about the regression equation. Thus s_e represents a measure of the error with which any observed value of Y is selected from the distribution of Y values at each value of X.

In Figure 1, the two plots of X and Y values for sets A and B differ only in the spread or range of X_i. In set A, the range is $13 - 1 = 12$ units, whereas in set B the range is $8 - 1 = 7$ units. The different ranges of X_i result in different estimates both for a and b in the two fitted regression equations and also different estimates of the error variance (s_e^2). Because $R^2 = 0.9629$ for the fitted equation with data set A is higher than $R^2 = 0.8575$ with data set B, in spite of the fact that the estimated slope, $\hat{b}$, is larger with set B than with set A, we are led to believe the equation $\hat{Y}_i = 10.63 + 3.04X_i$ fits the data in set A better than the equation $\hat{Y}_i = 6.0 + 5.58X_i$ fits the data in set B. Before we can determine if indeed this is the case, we need to define R^2.

Definition of R^2. In the calculation of the summary statistics (Table 1), the quantity SS_Y is a measure of the variation in the Y_i values about their mean, $\overline{Y}$. In other words, SS_Y is a measure of the uncertainty in predicting Y without taking X into consideration. Similarly, SSE is a measure of the variation in the values of Y_i, or the uncertainty in predicting Y, when a regression model containing the variable X is employed. A reasonable measure of the effect of X in explaining the variation in Y is R^2, calculated either as

$$R^2 = 1 - \text{SSE}/\text{SS}_Y \text{ or } R^2 = (\text{SS}_Y - \text{SSE})/\text{SS}_Y \qquad (1)$$

The quantity $(\text{SS}_Y - \text{SSE})$ is equal to $\hat{b}\text{SS}_{XY}$ and is the regression sums of squares (S.S. Regression), that is, the variation in the Y_i values explained, or accounted for, by the fitted regression equation $\hat{Y}_i = \hat{a} + \hat{b}X_i$.

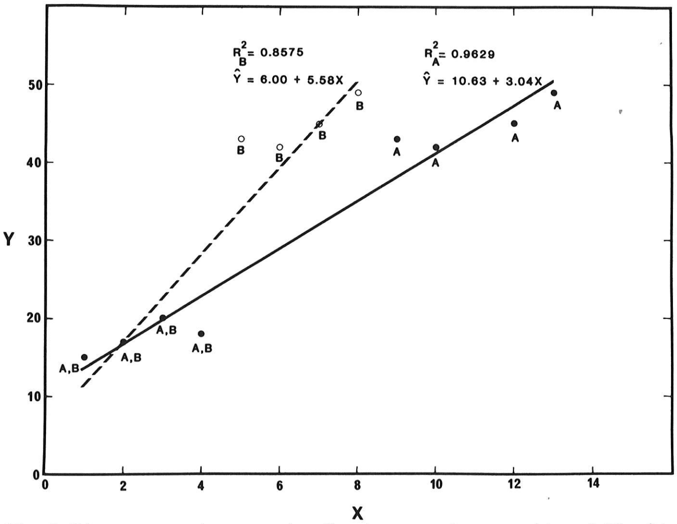

Fig. 1. Linear regression equation fitted to two data sets (A and B) with identical Y values. The different ranges of X cause different estimates of slopes, intercepts, and R^2 values.

TABLE 1. Calculations needed to obtain the fitted regression equations and other summary statistics for two data sets

Observations	Data set A					Data set B			
Y_i	X_i	$Y_i - \overline{Y}$	$X_i - \overline{X}$	$\hat{Y}_i$	$Y_i - \hat{Y}_i$	X_i	$X_i - \overline{X}$	$\hat{Y}_i$	$Y_i - \hat{Y}_i$
15	1	−16.125	−5.75	13.67	1.33	1	−3.5	11.58	3.42
17	2	−14.125	−4.75	16.70	0.30	2	−2.5	17.17	− 0.17
20	3	−11.125	−3.75	19.74	0.26	3	−1.5	22.75	− 2.75
18	4	−13.125	−2.75	22.78	−4.78	4	−0.5	28.33	−10.33
43	9	11.875	2.25	37.96	5.04	5	0.5	33.92	9.08
42	10	10.875	3.25	40.99	1.01	6	1.5	39.50	2.50
45	12	13.875	5.25	47.06	−2.06	7	2.5	45.08	− 0.08
49	13	17.875	6.25	50.10	−1.10	8	3.5	50.67	− 1.67
$\Sigma Y_i =$				249.0				249.0	
$\overline{Y}$ =				31.125				31.125	
$\Sigma X_i =$				54.0				36.0	
$\overline{X}$ =				6.75				4.5	
$\Sigma(Y_i - \overline{Y})^2 = \text{SS}_Y =$				1,526.875				1,526.875	
$\Sigma(X_i - \overline{X})^2 = \text{SS}_X =$				159.50				42.0	
$\Sigma(X_i - \overline{X})(Y_i - \overline{Y}) = \text{SS}_{XY} =$				484.25				234.5	
$\Sigma(Y_i - \hat{Y}_i)^2 = \text{SSE} =$				56.67				217.58	
Estimate of the intercept $\hat{a} = \overline{Y} - \hat{b}\overline{X} =$				10.63				6.00	
Estimate of the slope $\hat{b} = \text{SS}_{XY}/\text{SS}_X =$				3.04				5.58	
S. S. Regression $\text{SS}_Y - \text{SSE} = \hat{b}\text{SS}_{XY} =$				1,470.21				1,309.29	
Coefficient of determination $R^2 = 1 - \text{SSE}/\text{SS}_Y =$				0.9629				0.8575	
Estimate of σ_e^2: $s_e^2 = \text{SSE}/(N-2) =$				9.44				36.26	
Slope/standard error $\hat{b}/s_e =$				0.9894				0.9267	
Fitted regression equation				$\hat{Y}_i = 10.63 + 3.04X_i$				$\hat{Y}_i = 6.00 + 5.58X_i$	

The quantity R^2 is interpreted therefore as the *proportion of the total variation associated with the use of the independent variable* X. Thus, the closer R^2 is to one, the greater is the proportion of the total variation in the Y values that is explained (or accounted for) by introducing the independent variable X into the regression equation. But does this acutally mean that the higher the value of R^2 the better the model fits the data?

In a descriptive sense, formula (1) shows that R^2 is an estimator of the strength of the relationship between Y and X because R^2 is directly related to the sum of squares for regression, which is itself a function of the estimated slope, that is, S. S. Regression $= \hat{b}\mathrm{SS}_{XY}$. But we shall see that the magnitude of the strength of this relationship, R^2, can also be influenced by the choice of X values and the sample size. In particular, we will discuss the following cases as they pertain to factors that affect R^2: 1) the number of replicated observations of Y and the variation in these Y values at each setting of X for a given set of X values, 2) the range of the X (largest value minus the smallest value), 3) the arrangement of X values within a given range, and 4) the value of the slope estimate ($\hat{b}$) in relation to the standard error (s_e). To address each of these four cases, we shall use pairs of data sets to illustrate the behavior of R^2. In three of the four cases, the fitted model is the same for each data set and equivalent fits are obtained, as measured by $s_e^2 = \mathrm{SSE}/(N-2)$, yet the R^2 values for the two models differ. A rule that states how the value of R^2 is affected for each case is given.

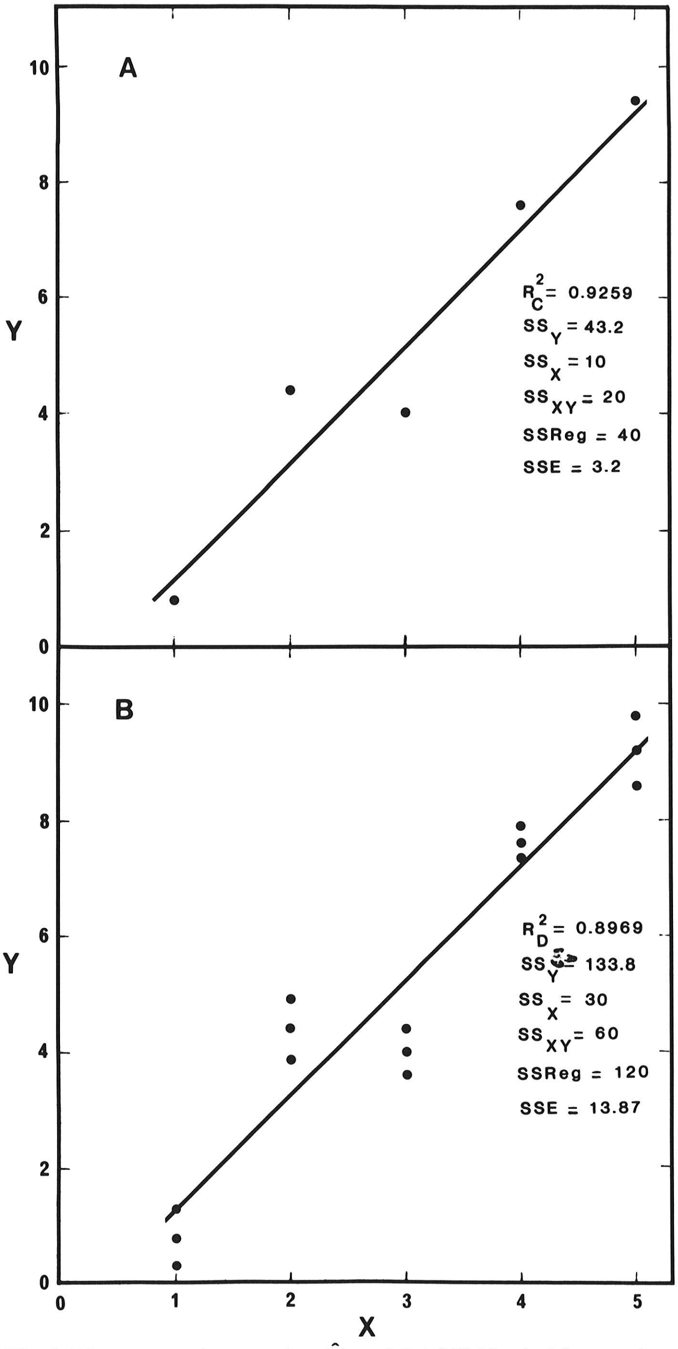

Fig. 2. Linear regression equation ($\hat{Y} = -0.8 + 2X$) identical for two data sets. **A,** Set C with one observation of Y at each X value. **B,** Set D with triplicate observations of Y. The estimated variance ($s_e^2 = 1.066$) is the same for both data sets. The value of R^2 decreases as the number of replicate observations increases.

Case 1—sample size effect on R^2. Data sets C and D produced the same fitted model: $\hat{Y} = -0.8 + 2X$, which is shown in Figure 2 along with the summary statistics. In data set D, three observations ($n = 3$) were taken at each setting of X and the average of each set of triplicate Y values is equal to the Y value in set C at the corresponding value of X. The estimate of the unexplained variation in the Y values at each value of X is the same ($s_e^2 = 1.066$) with both data sets, which means the regression equation fits the data values in both sets equally well. Note that only with set D, which has replicated Y values, can the value of s_e^2 be checked. The lower R^2 with data set D is of interest even though the fitted model and the measure of closeness of the model to the data, s_e^2, are identical with both sets. The lower R^2 with set D leads one to believe the fitted model ($\hat{Y} = -0.8 + 2X$) is not as good with set D as it is with set C. Although it is true that the fitted model explains less of the total variation in the Y values of set D than in set C, it is important to realize that the variation among the replicated Y values at each value of X, although contained in the total sum of squares of the Y values, is not accounted for in the regression sum of squares. Hence, by collecting replicate Y values at one or more values of X in an attempt to obtain a valid estimate of the error variance (a principle of good experimental design), this action will result in lowering the value of R^2. However, in lowering the value of R^2 by collecting replicate Y values, one gains something in return. Using the replicated observations only, one can obtain an unbiased estimate of σ_e^2. With the unbiased estimate of σ_e^2, one can test for

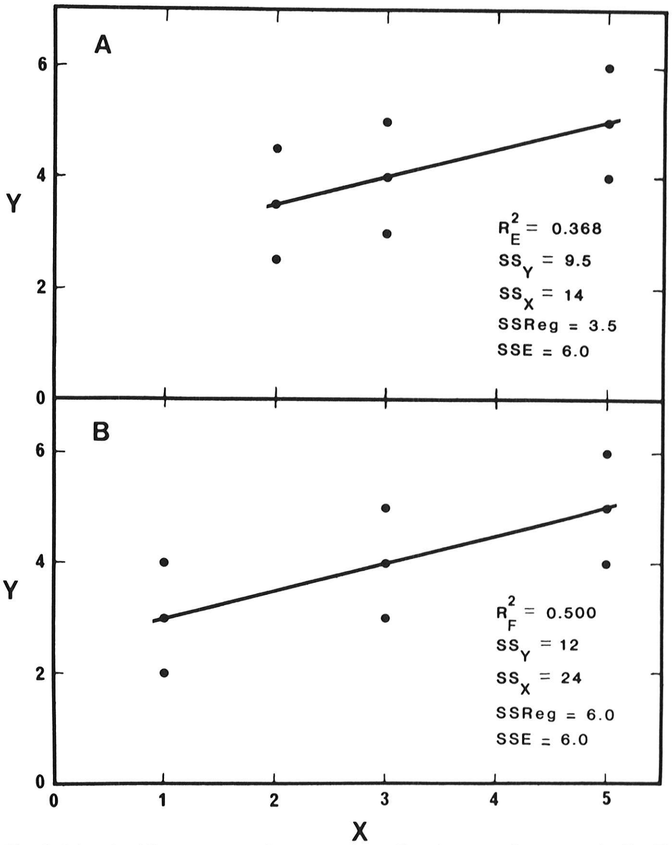

Fig. 3. Identical linear regression equations fitted to two data sets. **A,** Set E with three units in the range of X. **B,** Set F with four units. The R^2 value increases as the range of X increases.

adequacy of the fitted regression equation. This test is described after Rule 3. Without replicate Y values, the test is not possible.

In retrospect then, in the case of the two models fitted to sets C and D, the values of s_e^2 were the same, therefore, the models possessed equivalent fits to their respective data sets. Yet R^2 with set D was 0.0294 lower in value than R^2 with set C because of the variation among the replicate Y values in set D.

Rule 1. The value of R^2 for a fitted model (at a fixed value of $\hat{b}/s_e$) is reduced when more than one ($n > 1$) Y value is collected at each value of X. The ratio $\hat{b}/s_e$ is a measure of the strength of the linear association between Y and X relative to the unexplained variation in Y. The reducing effect of collecting $n > 1$ replicate values of Y at each X on R^2 becomes less for large values of $\hat{b}/s_e (\geqslant 5)$.

Case 2—the effect of the range of X on R^2. In the initial example, the Y values were identical with data sets A and B, however, in set A the data was taken from a larger range of X (12 units) than in set B (range = 7 units). This larger range was responsible for the higher R^2 associated with the closer fitting model of set A. The difference in the ranges of X values resulted in not only different models for each set, but also different estimates of the variance (s_e^2) as well as different R^2 values.

Suppose we examine the effect of varying the range of X on the value of R^2 by considering first the case where the fitted models for the two data sets are again identical in form, $\hat{Y}_i = 2.5 + 0.5X_i$. In Figure 3, two data sets (E and F) each contain triplicate Y values at each setting of X. The range of X in data set F is larger ($5 - 1 = 4$ units) than for set E ($5 - 2 = 3$ units). Since, in both data sets, the same number of Y values were collected at the respective low, middle, and high levels of X, then SS_X with set F exceeds SS_X with set E because the range of X is greater with set F.

The range of X affects SS_X and affects the sum of squares for regression since

$$\begin{aligned}\text{S.S. Regression} &= \text{SS}_Y R^2 \\ &= \hat{b}^2 \text{SS}_X\end{aligned}$$

where $\hat{b}$ is the estimated slope of the regression line. Since the slopes of the two fitted models with data sets E and F are identical, then

$$\left[\text{S. S. Regression (F)} = \hat{b}^2 \text{SS}_{X(\text{F})}\right] > \left[\text{S. S. Regression (E)} = \hat{b}^2 \text{SS}_{X(\text{E})}\right] \qquad (2)$$

If the value of the ratio SS_X/SS_Y with set F exceeds the value of this ratio with set E, R^2 with set F exceeds R^2 with set E:

$$\left[R_{\text{F}}^2 = \hat{b}^2 \text{SS}_{X(\text{F})}/\text{SS}_{Y(\text{F})}\right] > \left[R_{\text{E}}^2 = \hat{b}^2 \text{SS}_{X(\text{E})}/\text{SS}_{Y(\text{E})}\right]$$

even though the fitted models are the same and their fits are equivalent (i.e., $s_e^2 = 0.857$) in the two data sets. Thus, when fitting a simple linear regression model, the greater one spreads out the X values (thus directly affecting SS_X) relative to the spread of the Y values, the larger the ratio SS_X/SS_Y becomes and for a fixed value of $\hat{b}$, the higher will be the value of R^2.

A cautionary note: The occurrence of an observation (Y_i) at an extreme X_i can influence the R^2 disproportionately. For example, a model fitted to a cluster of 5–20 observations may have an $R^2 < 0.2$; but with the addition of a single observation at an extreme X_i outside the cluster, the R^2 could increase to be > 0.95. The researcher must be assured that response values at the extremes of the range of X are good observations rather than being outliers.

If an outlier Y value exists within the range of X, the outlier observation will have a larger residual than will the other observations. The residuals of outliers make both SSE and SS_Y large. Because $R^2 = 1 - \text{SSE}/\text{SS}_Y$, with outliers where both SSE and SS_Y increase simultaneously, the ratio SSE/SS_Y is forced toward unity so that R^2 approaches zero. Hence, outlier observations within the range of X have a reducing effect on R^2.

Now let us consider the case in which two data sets (G and H) again have different ranges of X, but here the models possess the same value for the ratio, $\hat{b}/s_e$. In Rule 1, it was stated that because the ratio $\hat{b}/s_e$ is a measure of the strength of the linear relationship between Y and X relative to the standard error of the Y values, the larger the value of $\hat{b}/s_e$, the greater is the effect of X on Y. Since the two fitted models have the same $\hat{b}/s_e$ value, X affects Y equally in the two sets. However, the range of X in set G is $5 - 1 = 4$ units and in set H, the range is $3.5 - 2.5 = 1$ unit (Fig. 4).

The estimated slope of the regression equation with set H is almost three times as large ($5.72/2.06 = 2.777$) as the estimated slope for set G. However, the magnitude of the estimate of the unexplained variation in Y, s_e^2, is approximately eight ($3.8837/0.5053 = 7.7$) times greater with set H than with set G. Although both models possess a value of 2.9 for the ratio $\hat{b}/s_e$, R^2 is 52% larger (0.9655 vs. 0.637) with set G than with set H because of the greater range of X values in set G. In other words, X has four times more leverage in set G than in set H and this produces a higher R^2 value with set G.

Rule 2. For N equally spaced values of X with range W, suppose a single observation of Y is taken at each X. If the relative measure of linear association of Y on X is fixed to be b/σ_e, then the value of R^2 increases with increasing range W. If N is allowed to increase indefinitely, R^2 converges in value to $t/(1 + t)$ where $t = (b/\sigma_e)^2 W^2/12$. This formula for the value of R^2 is given in (14).

Case 3—the effect of the arrangement of X values in a given range on R^2. Data sets I and J each have the same range of X but in set I the values are equally spaced, whereas in set J the values of X are at the two extremes of the range (Fig. 5). With both data sets, the fitted model is $\hat{Y}_i = 2 + 2.5X_i$. The higher R^2 with data set J is caused by two factors; the lower value of s_e^2 and the higher value of SS_X with set J. If the value of s_e^2 had been the same in both data sets, the R^2 with set J would still exceed R^2 with set I since SS_X is maximized by taking half of the observations ($N/2$) at each of the extreme values of X.

Rule 3. The value of R^2 for a fixed range W of X is maximum if when N is even, $N/2$ observations are collected at each end point of W. If N is odd, collecting $(N + 1)/2$ observations at one end point and $(N - 1)/2$ observations at the other end point will maximize R^2. If N is allowed to increase indefinitely, R^2 converges in value to $v/(1 + v)$ where $v = (b/\sigma_e)^2 W^2/4$, as shown in (14).

The attempt to maximize SS_X (and R^2) by selecting two extreme values of X and collecting data only at these values can only be

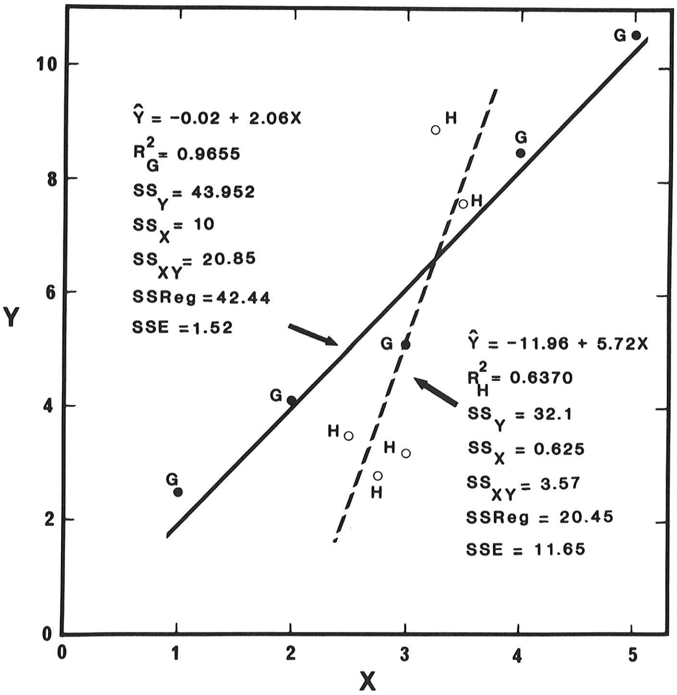

Fig. 4. Linear regression equations fitted to two data sets with equivalent ratios of $\hat{b}/s_e$. In Set G (dots and solid line), the range of X is four units. In Set H (circles and dashed line), range of X is one unit. The R^2 value increases as the range of X increases.

recommended if one knows beforehand that the relationship between Y and X is a straight line. When this is not known and further, if a test of the nonlinearity of the relationship is desired, then at least three distinct values of X must be used. When there are exactly three X values, the middle value is preferably at or near the middle of the range of X values. The test for nonlinearity involves the following steps:

1. Replicate observations must be collected at one or more levels of X.

For illustrative purposes, let us denote the replicated Y values at X_i by Y_{ij}, where $j = 1, 2, \dots, n_i$ and let there be k levels of X_i, i.e., $X_1, X_2, \dots, X_k$ so that $n_1 + n_2 + \dots + n_k = N$, where $n_i \geqslant 1$. That is, if at X_1 three values of Y are collected, they are denoted as Y_{11}, Y_{12}, and Y_{13}.

2. An estimate of the variability of the Y_{ij} values about their mean, $\overline{Y}_i$, is calculated using the replicated observations at X_i, as

$$\mathrm{SSPE}_i = \Sigma_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2.$$

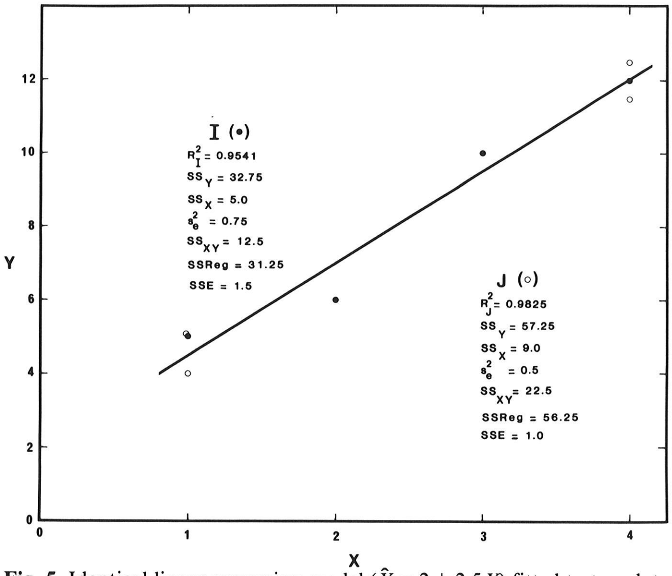

Fig. 5. Identical linear regression model ($\hat{Y} = 2 + 2.5X$) fitted to two data sets (I and J). The R^2 is higher with set J because the observations are grouped at the extremes of the range of X.

SSPE_i is called the sum of squares pure error at X_i and SSPE_i has associated with it $n_i - 1$ df. Pooling or summing the SSPE_i across the k levels of X_i, we obtain the sum of squares pure error

$$\mathrm{SSPE} = \Sigma_{i=1}^{k}\ \Sigma_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2.$$

SSPE has $\Sigma_{i=1}^{k}(n_i - 1) = (N - k)$ df.

3. Subtract SSPE from SSE, where SSE is obtained from the analysis of the N observations as in Table 1, to get SS Lack of Fit, that is

$$\text{SS Lack of Fit} = \text{SSE} - \text{SSPE}.$$

SS Lack of Fit has $(N - 2) - (N - k) = k - 2$ df.

4. The test for nonlinearity (or the test of the hypothesis, lack of fit equals zero) of the fitted model is performed by calculating

$$\mathrm{F} = \frac{\text{SS Lack of Fit}/(k-2)}{\text{SSPE}/(\mathrm{N} - k)},$$

and then comparing the calculated F value to a tabled value of F with $k - 2$ and $(N - k)$ df in the numerator and in the denominator, respectively, for some level of significance, say $\alpha = 0.05$. Nonlinearity is suspected when the calculated F value exceeds the tabular value.

Case 4—the effect of the ratio b/σ_e on R^2 for a fixed value of N and range W. Data sets K, L, and M each consist of only five values of Y, one at $X = 1, 2, 3, 4$, and 5 (Fig. 6). In each of the three data sets, the fitted models possess equivalent fits in terms of producing the same value for the estimate of unexplained variation in Y, i.e., $s_e^2 = 1.067$. The strength of the linear relationship between Y and X, as measured by the ratio $\hat{b}/s_e$ however, increases with increasing slope estimate as one proceeds from K to L to M. This increase in slope for a fixed value of s_e produces an increase in R^2 for sets K, L, and M of 0.6667, 0.9259, and 0.9697, respectively.

Rule 4. When two simple linear regression equations are fitted to their respective data sets where the ranges of X are the same and the fitted models are equivalent in terms of s_e^2, then the fitted model having the greater slope ($\hat{b}$) will also possess the higher value of R^2.

To summarize, through the use of several data sets to which simple regression models were fitted, the value of R^2 was shown to be affected by the different data collection schemes. Specifically, the factors of the sampling scheme that were seen to have a

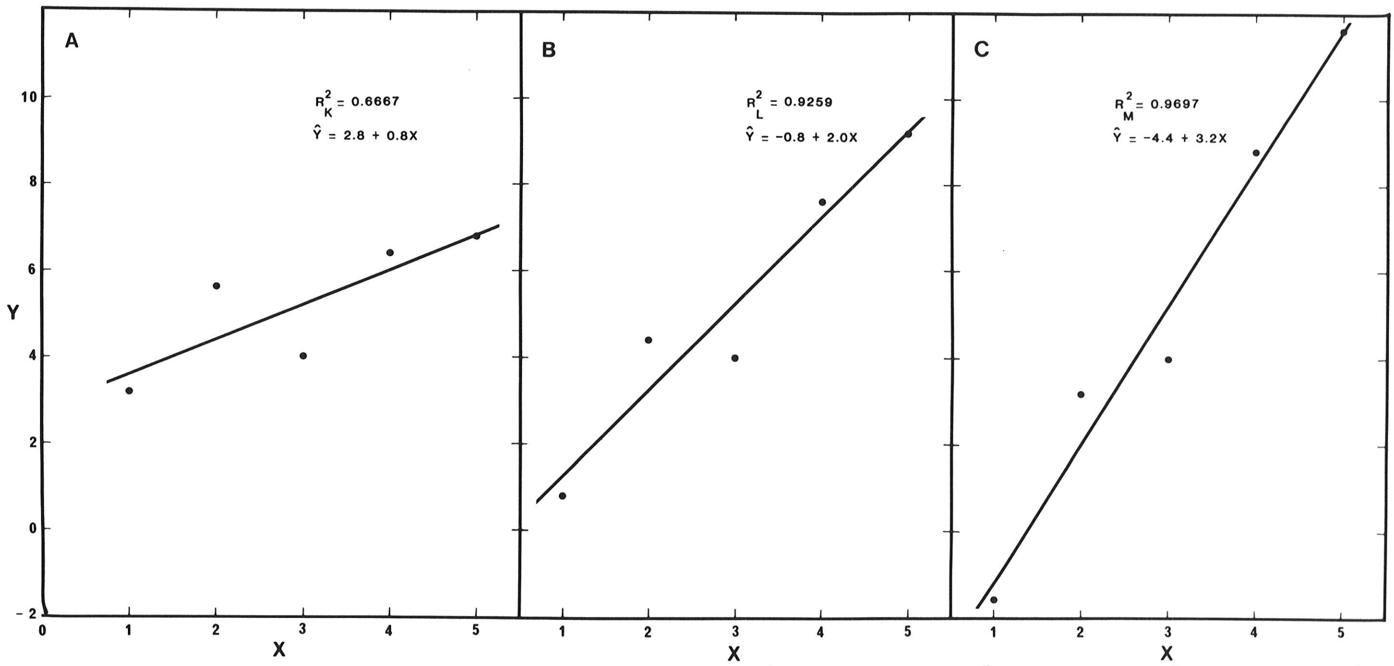

Fig. 6. Linear regression equations fitted to three data sets with equivalent variance ($s_e^2 = 1.067$). **A,** Set K with $\hat{b} = 0.8$. **B,** Set L with $\hat{b} = 2$. **C,** Set M with $\hat{b} = 3.2$. The R^2 value increases as slope ($\hat{b}$) increases.

reducing effect on R^2 were, replicated Y values at one or more values of X, defining a small range of X values at which to collect the observations, and, by observing the Y's at intermediate values of X across the range. In emphasizing the latter two factors, minimum range and intermediate X values, our intent is primarily to point out why an experimenter may have obtained an R^2 that is considered lower than expected and is not to suggest ways to avoid getting a low R^2.

Suppose one can choose the specific X values that might be termed reasonable to investigate the possible presence of a linear relationship between Y and X, how then should they proceed? Keeping in mind rules 1, 2, and 3, we recommend the following for fitting a simple regression equation:

1. Collect replicate observations of Y at each setting of X so as to be able to calculate an estimate of σ_e^2. Keep in mind that multiple observations of Y at each X have a reducing effect on the R^2, therefore, R^2 should not be relied on as the sole model-fitting criterion.

2. Select the range of X to be as large as possible, with the assurance that the relationship between Y and X is linear over the range. If one is not sure of the existence of a strict linear relationship between Y and X, then increasing the range of X increases the likelihood of detecting nonlinearity in the relationship of Y on X if such exists. To detect this nonlinearity, collect one or more observations near the middle of the interval of X values and test for model lack of fit. If lack of fit is present, the simple linear regression model is not an adequate representation of the true relationship between Y and X and the equation should be upgraded by the addition of terms like cX^2 and dX^3.

3. For a fixed range of X, observations collected near the end points of the range will make larger the quantity $SS_X = \Sigma(X_i - \overline{X})^2$ and this increases the precision of the slope estimate of the fitted model.

The statistical significance of R^2. In simple regression, the statistical significance of R^2 is determined by testing the hypothesis that the slope (b) of the regression equation is zero. The test on the slope is performed by first calculating the F ratio [F = S. S. Regression/(SSE/($N-2$))]. Then the value of the F ratio for the regression model is compared with the tabular value of $F_{(1,N-2,\alpha)}$, where $1 = \text{df}$ in the numerator, $N-2 = \text{df}$ in the denominator, and α = the significance level. When the F ratio of the model is equal or greater than the tabular F value, the hypothesis of zero slope is rejected, i.e., a statistically significant regression has been obtained. An approximately equivalent test on R^2 would involve calculating $R_\alpha^2 = F_{(1,N-2,\alpha)}/[F_{(1,N-2,\alpha)} + (N-2)]$. If the R^2 of the model is equal to or greater than R_α^2, then the regression equation is statistically significant. However, as N becomes large (>50), the value of $F_{(1,N-2,\alpha)}$ becomes small and along with N in the denominator, R_α^2 can be very small. For example, with $\alpha = 0.05$ level of significance and $N = 50$, $F_{(1,48,0.05)} = 4.03$ and $R_\alpha^2 = (4.03/52.03) = 0.0775$. With these parameters, an R^2 of only $\geqslant 0.0775$ is required to be statistically significant. Such a test on R^2 is sometimes meaningless since low R_α^2 values, however statistically significant, are intuitively unappealing unless some explanation is given (e.g., replicated Y values, etc.) concerning how the low value of R^2 resulted.

An R^2 statistic to measure the goodness-of-fit of nonlinear models. In this section, we extend the discussion on the use of R^2 as a measure of the goodness-of-fit but now the models are nonlinear in the unknown parameters. Although our discussion will center on the fitting of two simple nonlinear equations, the arguments are applicable for more complicated nonlinear equations.

Consider the following nonlinear models,

$$Y_i = aX_i^b e_i \qquad \text{(multiplicative errors)} \tag{3}$$

and

$$Y_i = a\exp(bX_i + e_i) \qquad \text{(exponential errors).} \tag{4}$$

Typically, to avoid the difficulties of having to use a nonlinear least-squares procedure to estimate the parameters a and b in equations 3 and 4, one takes the logarithms of both sides of the equalities in equations 3 and 4 to linearize the models, so that for equation 3:

$$y_i = \alpha + bx_i + e_i \tag{5}$$

where $y_i = \log_{10}(Y_i)$, $\alpha = \log_{10}(a)$, and $x_i = \log_{10}(X_i)$; and for equation 4:

$$y_i = \alpha + bX_i + e_i \tag{6}$$

where $y_i = ln(Y_i)$, $\alpha = ln(a)$, and X_i is not transformed. Then using linear or ordinary least squares, the estimates of the parameters α and b in equations 5 and 6 are calculated and the antilog of $\hat{\alpha}$ along with the estimate $\hat{b}$ are substituted into equations 3 and 4 to produce the prediction equations, e.g., for equation 3:

$$\hat{Y}_i = 10^{\hat{y}_i} = 10^{\hat{\alpha}} X_i^{\hat{b}} \tag{7}$$

where $10^{\hat{\alpha}}$ is the antilog of $\hat{\alpha}$, and for equation 4:

$$\hat{Y}_i = \exp(\hat{y}_i) = \exp(\hat{\alpha})\exp(\hat{b}X_i). \tag{8}$$

The explanatory power or goodness-of-fit of the fitted models (Eq. 7 and 8) in terms of the proportion of total variation in Y_i by using X_i in the models can once again be measured by the coefficient of determination. Previously, the formula for R^2, given in equation 1, was $R^2 = 1 - \text{SSE}/\text{SS}_Y$; this equation can also be written in terms of the Y_i, the predicted values $\hat{Y}_i$, and the mean $\overline{Y}$, as:

$$R^2 = 1 - \Sigma(Y_i - \hat{Y}_i)^2/\Sigma(Y_i - \overline{Y})^2. \tag{9}$$

To compare the quantity in equation 9 to other statistics also used to measure the goodness-of-fit of empirical models, we denote the quantity in equation 9 as R_1^2.

To assess the goodness-of-fit of linear regression equations of the form in equations 5 or 6 or of a similar form with several X's, three statistics that are analogous to R_1^2 and that are used by some, are

$$R_2^2 = \Sigma(\hat{Y}_i - \overline{Y})^2/\text{SS}_Y \tag{10}$$

$$R_3^2 = \Sigma(\hat{Y}_i - \overline{\hat{Y}})^2/\text{SS}_Y \tag{11}$$

$$R_4^2 = 1 - \Sigma(r_i - \bar{r})^2/\text{SS}_Y \tag{12}$$

In equation 11, $\overline{\hat{Y}} = \Sigma\hat{Y}_i/N$ and in equation 12, $r_i = Y_i - \hat{Y}_i$ (the i^{th} residual). What we would like to know is to measure the goodness-of-fit of the nonlinear fitted models (Eq. 7 and 8), are the statistics R_1^2, R_2^2, R_3^2, and R_4^2 equivalent? If they are not, which of the statistics is the appropriate one for nonlinear models such as those in equations 7 and 8, or any other form of nonlinear model for that matter? Before we attempt to answer these questions, let us illustrate the calculations necessary to compute the values of the four R_i^2, $i = 1, 2, 3, 4$, when a model of the form of equation 3 is fitted to the data set in Table 2. The first stage in the model-fitting exercise is to transform the Y_i and X_i values to their $\log_{10}$ values (Table 2). The fitted model for $\log_{10}(Y_i)$ as a linear function of $\log_{10}(X_i)$ is:

$$\hat{y}_i = 0.2485 + 1.757x_i \tag{13}$$

with $R^2 = 0.9968$, which was calculated as in equation 1 (Fig. 7). Note that the value of $R^2 = 0.9968$ refers to the proportion of the total variation in $y_i(\log_{10}(Y_i))$ that is explained by the fitted model (Eq. 13), but this is not the proportion of the total variation in the original Y_i values explained by the fitted model.

The parameter estimates $\hat{\alpha} = 0.2485$ and $\hat{b} = 1.757$ in equation 13 are used to obtain estimates of the original parameters $\hat{a} = 10^{\hat{\alpha}}$, and $\hat{b} = \hat{b}$, which when substituted into equation 7 produce the prediction equation $\hat{Y}_i = 1.772X_i^{1.757}$. Values of $\hat{Y}_i$ corresponding to values of X_i are calculated using the prediction equation or by computing $\hat{Y}_i = 10^{\hat{y}_i}$. These values of $\hat{Y}_i$ are listed in Table 2 and are used to calculate the values of R_1^2, R_2^2, R_3^2, and R_4^2 using the formulas (Eqs. 9–12, respectively). For our data set, these values ar: $R_1^2 = 0.9969$, $R_2^2 = 0.9293$, $R_3^2 = 0.9289$, and $R_4^2 = 0.9973$. The four R_i^2 statistics, which are equivalent in the linear regression case, have a range of values equal to $R_4^2 - R_3^2 = 0.0684$, for this nonlinear regression problem. And although we have not observed with our example the particular extreme values that are possible with some of the R_i^2 values, it is possible for both R_2^2 and R_3^2 to exceed unity. Furthermore, $R_4^2 = R_1^2$ only if $\bar{Y} = Y$; $R_4^2 > R_1^2$ for all other cases. Because R_1^2 in equation 9 is always less than or equal to unity, it is the obvious choice and is thus the appropriate statistic to measure goodness-of-fit for both linear and nonlinear regression models.

Since R_1^2 is our choice to assess the fit of a nonlinear model, the question may be asked, should the researcher by concerned with the settings of the variable X to raise R_1^2 as in the linear regression case? Again, a single observation (Y) at each setting of X will produce a higher R_1^2 value than if replicate observations are taken. Also, a wide range of X values will make R_1^2 higher than a narrow range of X values. However, the optimal arrangement of X values within the range depends upon the particular form of the power or exponential model that is being fitted. Unlike the linear regression case where collecting observations at the extremes of the range of X_i variables increases the value of R_1^2, the optimal settings of the X variable to estimate the unknown parameters in a nonlinear model depend upon the actual values of the parameters; as such, the optimal settings of X cannot be specified without first stating the form of the nonlinear model. In most cases, the X values should be set at equal-sized intervals throughout the range.

In this paper, we have focused on the use of the coefficient of determination (R^2) as a measure of goodness-of-fit of regression equations. In the literature, the coefficient of determination appears, perhaps more often than any other statistic, as a regression model criterion. Unfortunately, it is also a misunderstood statistic. Because estimation of the observation variance is often just as

TABLE 2. Calculation of four analogous R^2 statistics for a fitted nonlinear model[z] ($\hat{Y}_i = 1.772X_i^{1.757}$)

		Log10							
Y_i	X_i	y_i	x_i	$\hat{Y}_i$	$Y_i - \hat{Y}_i$	$\hat{Y}_i - \bar{Y}$	$\hat{Y}_i - \bar{\hat{Y}}$	$r_i - \bar{r}$	$Y_i - \bar{Y}$
2.0	1	0.301	0.000	1.772	0.228	−47.014	−46.270	−0.516	−46.786
5.1	2	0.708	0.301	5.992	−0.892	−42.794	−42.050	−1.636	−43.686
19.9	4	1.299	0.602	20.258	−0.358	−28.528	−27.784	−1.102	−28.886
52.0	7	1.716	0.845	54.164	−2.164	5.378	6.122	−2.908	3.214
71.5	8	1.854	0.903	68.490	3.010	19.704	20.448	2.266	22.714
86.0	9	1.934	0.954	84.241	1.759	35.455	36.199	1.015	37.214
105.0	10	2.021	1.000	101.377	3.623	52.591	53.335	2.879	56.214
$\Sigma Y_i = 314.5$				$\Sigma \hat{Y}_i = 336.294$	$\Sigma(Y_i - \hat{Y}_i)^2$	$\Sigma(\hat{Y}_i - \bar{Y})^2$	$\Sigma(\hat{Y}_i - \bar{\hat{Y}})^2$	$\Sigma(r_i - \bar{r})^2$	SS_Y
$\bar{Y} = 48.786$				$\bar{\hat{Y}} = 48.042$	= 30.939 (1)	= 9,295.53 (2)	= 9,291.66 (3)	= 27.067 (4)	= 10,002.95 (5)

$$\bar{r} = \Sigma(Y_i - \hat{Y}_i)/7$$

$$= 0.744$$

Formulas: $R_1^2 = 1 - [(1)/(5)] = 0.9969$; $R_2^2 = (2)/(5) = 0.9293$; $R_3^2 = (3)/(5) = 0.9289$; $R_4^2 = 1 - [(4)/(5)] = 0.9973$

[z] If a model is linear in the parameters (e.g., $Y_i = a + bX_i + e_i$) and the model contains a constant term (e.g., a), then the following equalities are true: 1) $\Sigma Y_i = \Sigma \hat{Y}_i$; 2) $\bar{r} = 0$; 3) (1) = (4), hence $R_1^2 = R_4^2$; 4) (2) = (3), hence $R_2^2 = R_3^2$; and 5) (2) = (5) − (1) = (5) − (4) = (3), hence $R_1^2 = R_2^2 = R_3^2 = R_4^2$.

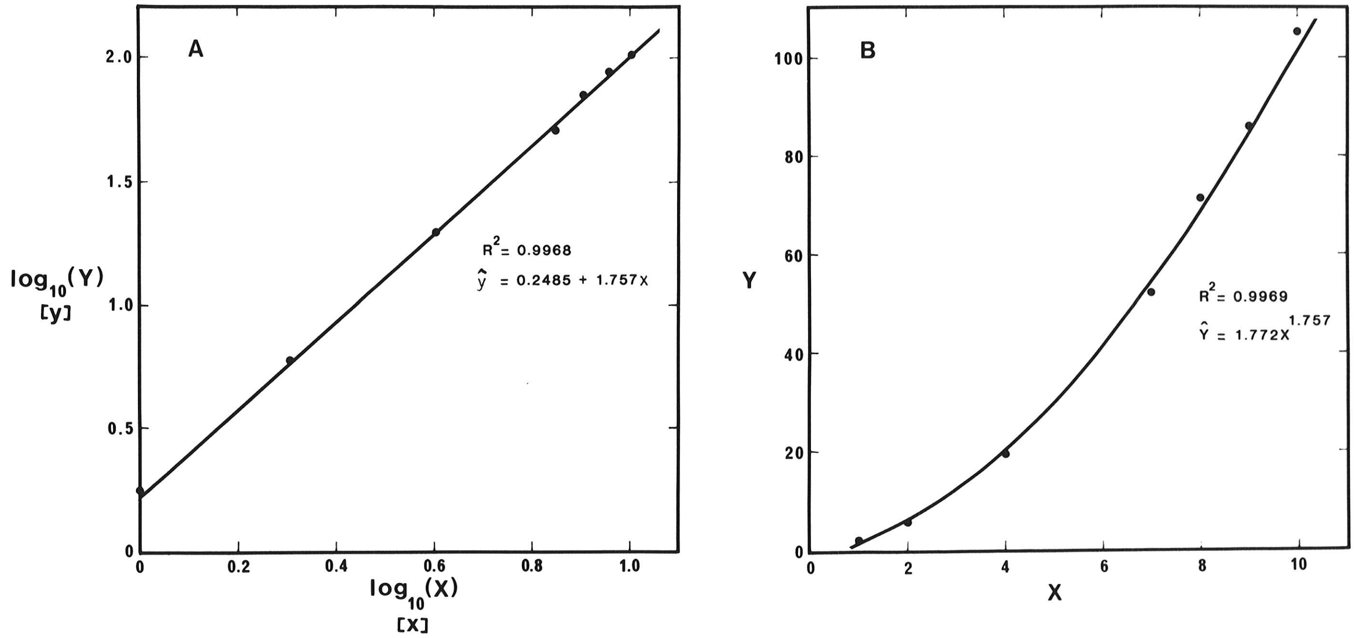

Fig. 7. Fitted regression equations for a nonlinear model. **A,** Linear regression of log10 (Y) vs. log10 (X). **B,** Plot of nonlinear multiplicative model to the original X and Y values. Note the slight difference in R^2 values for the two cases.

important as model fitting, an accompanying statistic that measures the proportional reduction in the variance estimate is the adjusted R^2:

$$R_A^2 = 1 - (1 - R^2)(N - 1)/(N - p - 1)$$
$$= 1 - \hat{\sigma}_e^2 / (SS_Y / (N - 1)) \qquad (14)$$

where p + 1 is the number of parameters (including the intercept a) in the linear or nonlinear model. The quantity $\hat{\sigma}_e^2$ is the observation variance estimate obtained from the fitting of the model. The $SS_Y/(N - 1)$ is the variance estimate assuming $Y = \overline{Y} + e_i$. Thus, when the model explains a significant amount of the behavior of the Y variable, $\hat{\sigma}_e^2$ will be small relative to $SS_Y/(N - 1)$. Like R^2, R_A^2 has an upper bound of unity. We recommend R_A^2 be used as a model criterion along with R^2, because the two criteria provide supplemental information concerning the fit of a regression equation.

LITERATURE CITED

1. Backman, P. A., and Crawford, M. A. 1984. Relationship between yield loss and severity of early and late leafspot diseases of peanut. Phytopathology 74:1101-1103.
2. Berger, R. D. 1981. Comparison of the Gompertz and logistic equations to describe disease progress. Phytopathology 71:716-719.
3. Bonde, M. R., Peterson, G. C., and Duck, N. B. 1985. Effects of temperature on sporulation, conidial germination, and infection of maize by *Peronosclerospora sorghi* from different geographical areas. Phytopathology 75:122-126.
4. Campbell, R. N., Greathead, A. S., Myers, D. F., and de Boer, G. J. 1985. Factors related to control of clubroot of crucifers in the Salinas Valley of California. Phytopathology 75:665-670.
5. Dillard, H. R., and Grogan, R. G. 1985. Relationship between sclerotial spatial pattern and density of *Sclerotinia minor* and the incidence of lettuce drop. Phytopathology 75:90-94.
6. Draper, N. R., and Smith, H. 1981. Applied Regression Analysis. 2nd ed. John Wiley & Sons, New York. 709 pp.
7. Gerik, T. J., Rush, C. M., and Jeger, M. J. 1985. Optimizing plot size for field studies of Phymatotrichum root rot of cotton. Phytopathology 75:240-243.
8. Grove, G. G., Madden, L. V., Ellis, M. A., and Schmitthenner, A. F. 1985. Influence of temperature and wetness duration on infection of immature strawberry fruit by *Phytophthora cactorum*. Phytopathology 75:165-169.
9. Johnson, D. A., and Skotland, C. B. 1985. Effects of temperature and relative humidity on sporangium production *Pseudoperonospora humili* on hop. Phytopathology 75:127-129.
10. Kellam, M. K., and Coffey, M. D. 1985. Quantitative comparison of the resistance to Phytophthora root rot in three avocado rootstocks. Phytopathology 75:230-234.
11. Kvalseth, T. O. 1985. Cautionary note about R^2. The Am. Stat. 39:279-285.
12. Madden, L. V. 1980. Quantification of disease progress. Prot. Ecol. 2:159-176.
13. Madden, L. V., Pennypacker, S. P., Antle, C. E., and Kingsolver, C. H. 1981. A loss model for crops. Phytopathology 71:685-689.
14. Ranney, G. B., and Thigpen, C. C. 1981. The sample coefficient of determination in simple linear regression. The Am. Stat. 35:152-153.
15. SAS Institute. 1982. SAS User's Guide: Statistics. Cary, NC. 584 pp.
16. Spotts, R. A. 1985. Environmental factors affecting conidial survival of five pear decay fungi. Plant Dis. 69:391-392.