Letter to the Editor

On Comparing Values of Vanderplank's *r*

W. C. Fulton

Assistant professor, Department of Botany and Plant Pathology, Michigan State University, East Lansing, MI 48824. Michigan Agricultural Experiment Station Journal Series Article 9058.
Accepted for publication 29 August 1979.

Vanderplank's *r* (7) has been used by plant pathologists to evaluate data on yield loss (4), sanitation efforts (2), plant spacing effects (6), fungicide effects, and cultivar effects (3). These cited references are only random selections from the recent plant disease literature, and should be considered as examples, not special cases.

Vanderplank (7, pp. 22–27) proposes two methods of estimating *r* from sample data. One method is by a linear regression on time (t) of the natural logarithm (ln) of the proportion of the plant population that is diseased (x) divided by the proportion that is nondiseased ($1 - x$). The slope of the regression line is then taken as an estimate of *r*. The other method is similar, but uses disease proportion data collected at only two times, t_1 and t_2. Those data are again transformed by $y = \ln[x/(1 - x)]$ and *r* is estimated by $r = (y_2 - y_1)/(t_2 - t_1)$. If our intention is merely to calculate an *r* value, then there is no problem. But, if we wish to compare two or more *r* values, or to compare the coefficients of determination of several regression equations of $\ln[x/(1 - x)]$ on t, then a problem arises. For these purposes, we must make assumptions about the functional form of the distribution of the variables. Usually the assumption is a normal distribution, and that is the tacit assumption when testing hypotheses by Vanderplank's method for estimating the standard error of *r* (7, p. 26), or in statements about confidence limits on the slope of the regression line when that approach is used to estimate *r*.

But what is the distribution of $y = \ln [x/(1 - x)]$? Ashton (1) showed that for large samples the expected value (E) of a function, (g) of a random variable (x) is approximated by:

$$E[g(x)] \simeq g[E(x)] \tag{1}$$

and the variance of that function is approximately

$$\text{Var}\,[g(x)] \simeq \left[\frac{dg}{dx}\Big|_{x = E(x)}\right]^2 \quad \text{Var}\,(x) \tag{2}$$

in which the vertical bar indicates that the derivative is to be evaluated at $x = E(x)$. Let $g(x) = y = \ln [x/(1 - x)] = \ln x - \ln (1 - x)$

$$\text{then } E[g(x)] \simeq \ln[E(x)] - \ln [1 - E(x)] \tag{3}$$

$$\text{and Var}\,[g(x)] \simeq \left[\frac{1}{x} + \frac{1}{1 - x}\Big|_{x = E(x)}\right]^2 \text{Var}\,(x) \tag{4}$$

Thus, we have approximations for the mean and the variance of y as functions of those of the parent distribution. Unfortunately, we do not know the distribution of x. We can however, make reasonable assumptions, and consider the implications of these assumptions.

Assumption 1. If the disease is systemic, we might observe a number of plants at random, classifying each as diseased or nondiseased. Here, a reasonable assumption is that we are sampling from a binomial distribution. Under these conditions, the sample proportion, p, of infected plants provides an estimate of E(x), while the variance Var(x) is estimated by $p(1 - p)/n = pq/n$ in which n is the number of plants sampled and $q = (1 - p)$. Substituting these estimates in equations 3 and 4, we get:

$$E\left(\ln \frac{x}{1 - x}\right) \simeq \ln p - \ln q \text{ and}$$

$$\text{Var}\left(\ln \frac{x}{1 - x}\right) \simeq \left[\left(\frac{1}{x} + \frac{1}{1 - x}\right)\Big|_{x = p}\right]^2 \frac{pq}{n}$$

$$= \left[\frac{1}{p} + \frac{1}{q}\right]^2 \frac{pq}{n}$$

$$= \frac{1}{npq} \tag{5}$$

Thus, the variance is a function of both the sample size and the proportion of infected plants in the population.

Assumption 2. If we assume that our estimates of the proportion of plants or tissue diseased follow a normal distribution with mean μ and variance σ^2 estimated by $\bar{x}$ and s^2, we have, on substitution in equation 3.

$$E\left(\ln \frac{x}{1 - x}\right) \simeq \ln \bar{x} - \ln (1 - \bar{x})$$

and on substitution in equation 4.

$$\text{Var}\left(\ln \frac{x}{1 - x}\right) \simeq \left[\frac{1}{x} + \frac{1}{1 - x}\Big|_{x = \bar{x}}\right]^2 s^2$$

$$= \left(\frac{1}{\bar{x}} + \frac{1}{1 - \bar{x}}\right)^2 s^2$$

$$= \frac{(\bar{x} + 1 - \bar{x})^2}{\bar{x}^2 (1 - \bar{x})^2}\, s^2$$

$$= \frac{1}{\bar{x}^2 (1 - \bar{x})^2}\, s^2 \tag{6}$$

Therefore, even with the assumption of normality in the distribution of x, the variance of $\ln [x/(1 - x)]$ is a function of the mean of the proportion of disease.

In linear regression, homogeneous variances are necessary to make valid probabliity statements about the results. When the variances are nonhomogenous, as they are under either assumption 1 or 2 above, a weighted regression analysis is required (1,5) in which the weights applied to the observations typically are the reciprocals of the variance at the observed value of the variable.

It is a common (and necessary) practice to violate many of the underlying assumptions when applying statistical techniques to biological realities. Some of these assumptions tend to be of minor importance. To test the importance of ignoring the heterogeneity of variance in $\ln [x/(1 - x)]$ when estimating Vanderplank's *r* by nonweighted linear regression, a computer simulation of sampling an epidemic was constructed. The advantages of using a simulation

00031-949X/79/000206$03.00/0

are that experimental error can be eliminated from the observations and that the exact nature of the underlying distribution is known. In weighted regression analysis the relative, not the absolute, weight applied to an observation is important. By comparing equation 5 to equation 6 with n and s^2 held constant and setting the maximum weight equal to one, it can be seen that the relative weight applied to an observation under assumption 2 would always be less than or equal to the relative weight which would be applied under assumption 1. Therefore, if it were important to weight observations under assumption 1, it would be even more important to weight them under assumption 2. In this simulation, an epidemic of a systemic disease, was assumed to progress at a rate of $r = 0.46$ per unit per day (cf 7, Fig. 3.1).

On the first day of the experiment, the proportion of diseased plants was set at $x = 0.005$. Therefore, $y = \ln [x/(1 - x)] = 5.293$. Each day thereafter, y was determined by adding $r = 0.46$ to the previous value of y. Values for x after day one were computed from $x = e^y/(1 \times e^y)$ (Table 1).

For each of the 11 successive days, 1,000 random numbers were drawn from a population uniformly distributed between 0 and 1. Each of these random numbers was considered to be a plant picked at random from a population of plants. For each day's sample, those random numbers which were less than the value of x for that day were considered to represent diseased plants. Those with values higher than x were considered to be disease free. Thus, on each day, i, we have a binomial distribution with parameter $\theta = x_i$. The proportion of the random numbers chosen on that day which were less than x_i provide an estimate, $\hat{p}_i$, of the parameter θ. Weights were computed as $w = npq$, the reciprocal of the variance, equation 5, and multiplied by the logit of $\hat{p}_i$.

A typical "experiment" is shown in Fig. 1A. The dotted line is the theoretical line for $r = 0.46$. The points are the sample proportions transformed to logits. The solid line is the unweighted least squares regression line, and the weighted least squares line is represented by the dashed line. In this case, all three lines are similar; the weighted regression line is nearly coincident with the theoretical line. Values of r were 0.472 for the weighted regression and 0.503 for the nonweighted regression.

Two hundred such experiments were conducted. Of these, the extreme values for r computed by unweighted least squares were $r = 0.400$ and $r = 0.564$. The corresponding values computed by weighted regression were $r = 0.430$ and $r = 0.510$. These "worst case" examples are shown in Fig. 1B. Considerable departure from the theoretical slope is evident in the unweighted least squares lines, with the weighted regression lines much closer to the true value.

With the 200 sets of data, 100 paired comparisons of r were made using a standard Student's t test for homogeneity of regression. At $\alpha = 0.05$, nine of the comparisons indicated significant differences when the regressions were nonweighted. By contrast, only four of the weighted regressions showed significant differences. That is, confidence statements about the weighted regression r's are valid, but those for unweighted regression are invalid, since twice as many comparisons of random variation were determined to be significantly different from one another as the chosen α level would suggest.

TABLE 1. The progress of a simulated epidemic of 11 days with an $r = 0.46$, x = the proportion of diseased plants

Day	x	ln [x/(1−x)]
1	.005	−5.293
2	.008	−4.833
3	.012	−4.373
4	.020	−3.913
5	.031	−3.453
6	.047	−2.993
7	.073	−2.533
8	.112	−2.073
9	.166	−1.613
10	.240	−1.153
11	.334	−0.693

Similar experiments with the two-point approach gave indications of significant differences which were in agreement with the chosen level for the t test. The two-point approach requires several estimates of p at each of the two sampling times. Each estimate was based on the examination of 1,000 "plants" as described above. An equal number of estimates was made at each sampling time. Valid results were obtained when seven or only three estimates were made with the true infection levels set at .009 and .443. Results also were correct when seven estimates were made at .009 and .250 relative incidence.

Vanderplank's (7) admonition to interpret estimates of r with caution has not always been heeded. The previous paragraphs have shown that comparison of nonweighted linear regression estimates of r's will show significant differences where none exist. The use of weighted linear regression or the two-point estimation method avoids this problem. It should be emphasized, however, the two-points method must be based on the means of several independent samples at each of the two points, since under these circumstances the central limit theorem applies. Weights were calculated herein as the reciprocal of the variance of the distribution, specifically, $w_i = n_i p_i q_i$. Here $n_i p_i$, and q_i were known. In experimental work, the weights can be computed by first estimating p_i using nonweighted linear regression, then recomputing the regression with weights based on this estimate p_i, of p_i as $\hat{W}_i = n_i \hat{p}_i \hat{q}_i$.

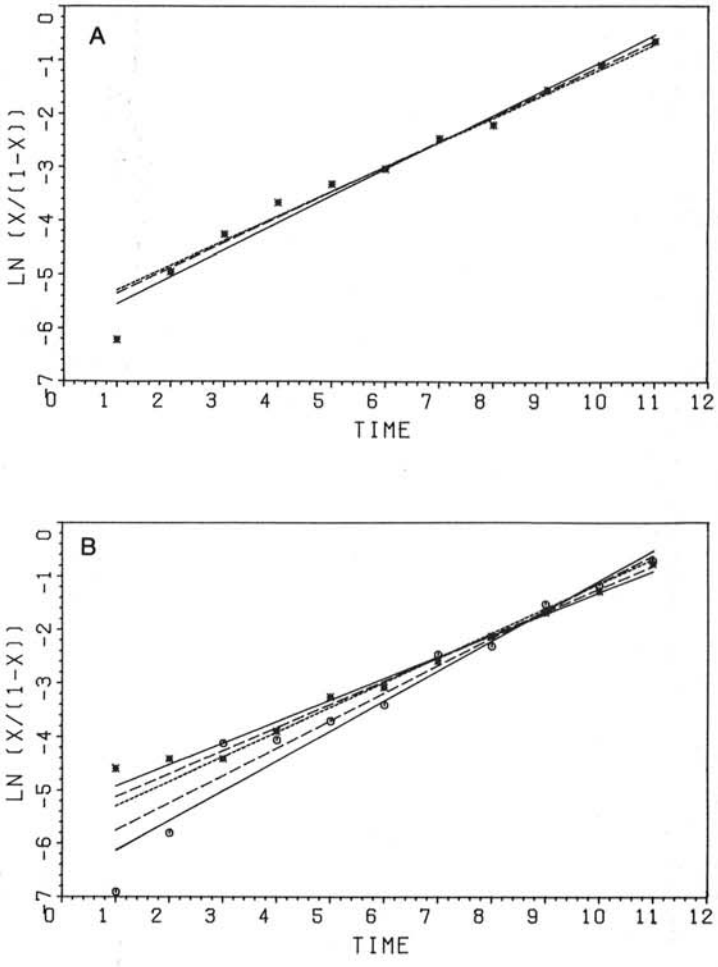

Fig. 1. Linear regressions of the logit transformation of plant disease incidence from a simulated epidemic over time. **A,** A typical example of the 200 epidemics generated. **B,** The most divergent pair of the 200 epidemics generated; ie, that simulated epidemic which gave the maximum estimate for r and that which gave the minimum estimate for r. Legend: solid line = unweighted least squares line; long dashes = weighted least squares line; and short dashes = true epidemic.

LITERATURE CITED

1. ASHTON, W. D. 1972. The Logit Transformation. Hafner Publ. Co., New York. 88 pp.
2. BERGER, R. D. 1977. Application of epidemiological principles to achieve plant disease control. Annu. Rev. Phytopathol. 15:165-183.
3. FRY, W. E. 1978. Quantification of general resistance of potato cultivars and fungicide effects for integrated control of potato late blight. Phytopathology 68:1650-1655.
4. JAMES, W. C., L. C. CALLBECK, W. A. HODGSON, and C. S. SHIH. 1971. Evaluation of a method used to estimate loss in yield of potatoes caused by late blight. Phytopathology 61:1471-1476.
5. STEEL, R. G. D., and J. H. TORRIE. 1960. Principles and Procedures of Statistics. McGraw-Hill, New York. 481 pp.
6. STRANDBERG, J. O., and J. M. WHITE. 1978. *Cercospora apii* damage of celery—effects of plant spacing and growth on raised beds. Phytopathology 68:223-226.
7. VANDERPLANK, J. E. 1963. Plant Diseases: Epidemics and Control. Academic Press, New York. 349 pp.