

CHOICE OF VARIABLES FOR CORRELATION ANALYSIS
WHEN MEASUREMENT ERROR IS PRESENT

F. Thomas Lloyd^{1/}

Abstract.--Although correlation analysis is widely used, it is not well known that using variables measured with error causes the correlation coefficient to be underestimated (or biased toward zero). Knowledge of the form of this bias will help in the selection of variables to be used in studying soil and site productivity relationships. An example using site index shows that bias can be appreciable.

Additional keywords: Bias, site index, variance ratios, *Pinus taeda*.

Planners of studies and interpreters of published results should be aware that the familiar and frequently used correlation coefficient is underestimated (or biased toward zero) when the variables used to compute the coefficient are measured with error. This paper presents a method of approximating the magnitude of this bias. The methodology should not, however, be used to adjust computed correlation coefficients that are to be published. Rather, it is an experimental design tool which when used in the establishment of new studies can aid selection of variables and measurement techniques that produce the least amount of bias. It can also be used to aid interpretation of published data by assessing whether actual correlations are substantially greater than those presented.

RESULTS

Measurement error as used in this work is the random fluctuation of repeated measurements on the same experimental unit about the actual (but unobservable) value of the variable. The average of these plus and minus fluctuations approaches zero as the number of measurements increases. This error is different from measurement bias, where for measurement bias the mean value of repeated measurements does not approach the actual value as the number of remeasurements is increased. So the definitions of a pair of measurement errors,

$$\epsilon = q - Q \quad (1)$$

and

$$\delta = m - M, \quad (2)$$

are simply differences between the observed values (q, m) and the actual values (Q, M).

^{1/} Biometrician, Southeastern Forest Experiment Station, USDA Forest Service, Charleston, S. C.

The question addressed below is whether the correlation coefficient (ρ) for the population of measured values is equal to the correlation coefficient (P) for the unobservable variables (Q, M). One might expect them to be equal because of the intuitive feeling that the plus and minus errors will cancel in the long run. Although this is generally true when estimating population means, the variance component due to measurement error does not drop out of the variance expression for the measured responses (q, m); and, the definition of the correlation coefficient is a function of these variances, plus the covariance of the two variables. For the measured variables q and m , the correlation coefficient is

$$\rho = \frac{\text{Covariance of } q \text{ and } m}{\sqrt{(\text{Variance of } q)(\text{Variance of } m)}} \quad (3)$$

If the expressions

$$q = Q + \epsilon$$

and

$$m = M + \delta$$

from Equations (1) and (2) are substituted into the definition in Equation (3), then

$$\rho = P / \sqrt{(1 + \lambda_q)(1 + \lambda_m)} \quad (4)$$

where, as previously stated, P is the desired correlation between Q and M and the new quantities λ_q and λ_m are ratios of the variances of the measurement errors (ϵ, δ) divided by the corresponding population variances for Q and M , respectively. The remainder of this paper concentrates on interpretation and computation of these λ -ratios.

INTERPRETATION AND CALCULATION OF λ

As stated, the λ 's are ratios of variances. When measurement error is absent from both variables, the corresponding measurement error variances are equal to zero, making the λ 's equal zero and, from Equation (4), ρ equal to P --in other words, no bias. The percent bias, that is, the difference between ρ and P divided by P times 100, is calculated from the expression $100 [1 - 1/\sqrt{(1 + \lambda_q)(1 + \lambda_m)}]$. Conversely, if measurement error variances are extremely large relative to the corresponding population variances, the percent bias as calculated above approaches 100. Since the correlation coefficient is bounded by ± 1 , a 100 percent bias means a zero correlation for the measured responses. Of course, the practical limit on the percent bias will be something less than this 100 percent extreme.

In addition to quantifying the effect of measurement error, the λ -ratios show the importance of knowing the variances of Q and M for the populations being sampled. This is an experimental design consideration that is often overlooked. It is clear from the definition of λ that bias could be unacceptably large if experimental units are sampled from a restricted population (that is, one with a small variance) even when measurement error appears to be within acceptable limits. Conversely, the bias produced by measurement error can be reduced by sampling from a population with a greater range in the Q and M values.

The use of the above results requires the estimation of measurement error variances and population variances. It is possible to get estimates (or approximations) from a variety of sources. For example, knowledge about data ranges, either from general experience or from previously collected data can be used. A simple calculation like $(R/4)^2$, where R is a range estimate based on knowledgeable experience, would be a useful approximation. A better variance estimate could be obtained by using a table like that in Snedecor and Cochran (1967, p. 40) and an observed range from a sample of a given size. Another source could be variance component estimates from an analysis of variance on a candidate variable measured and analyzed for another purpose. Experiments with subsampling could be especially useful for estimating measurement error variance. Sometimes class frequency data is available from past inventories. An estimate of population variance can be obtained from this data by assuming a uniform distribution within each class. The above is not meant to be instructive, nor is the list of possibilities exhaustive. The point being made is not to be overly concerned about the properties of the variance estimators, but to ingeniously use available data to estimate these variances. Often we know more about the populations being sampled than we will admit. Also, small, specially designed studies could be conducted to supply data for estimating the λ 's.

EXAMPLE

Suppose you want to measure the degree of relationship between site productivity and some soil variable in a population of plots on which 15-year-old stands of loblolly pine (*Pinus taeda* L.) are growing. What percent bias would result from measurement error in site index estimation alone?

The standard error of a site index prediction, base age 50, obtained from measuring 4 sample tree heights in a 15-year-old stand of Coastal Plain loblolly pine is estimated as 3.05 feet (Lloyd, in press). An estimate of population variance can be obtained by specifying a distribution of plots by site index classes for the population being sampled and assuming uniform distributions within classes. For example, a population of plots with 5 percent in the 70-foot site class, 10 percent in the 80-foot class, and 85 percent in the 90-foot class produces an estimated population standard

deviation of 5.86 feet. The resulting estimate of the λ -ratio is

$$\lambda_q = \frac{(3.05)^2}{(5.86)^2} = 0.27,$$

which produces a percent bias of 11 percent. Intuitively, the variance of site index predictions is greater for 10-year-old stands. Using the same site index variance predictor as that used above, except for 10-year-old stands, the

$$\lambda_q = \frac{(4.2)^2}{(5.86)^2} = 0.51,$$

and this value of λ_q (with $\lambda_m = 0$) produces a 19 percent bias.

It is not uncommon for soil variables to have large measurement error problems. For the purpose of illustration, assume some candidate soil variable has measurement errors comparable to that for site index from 10-year-old stands. Then the combined percent bias is 34 percent. This says that if the true correlation coefficient was, say, 0.95, the measured site and soil variables would on the average predict a correlation of 0.63. Expressing the strength of the above relationship as the percentage of variation in one variable explained by the other (R-square), the difference between the true fit and that from the measured variables is 51 (90-39) percent.

CONCLUSIONS

Caution should be used when interpreting correlation statistics from soil-site studies or when designing studies in which correlation (or regression) will be used because measurement error causes bias. An example using site index illustrates that the percent bias can be appreciable. Bias is reduced by careful selection of variables and measurement tools, and by an awareness of the population variances of the experimental units being studied. Estimating the percent bias requires computing measurement error variances and population variances from supplementary data sources. Ingenuity should be used in the estimation of these variances.

LITERATURE CITED

- Lloyd, F. T. (in press). Precision and the probability of misclassification in site index estimation. Forest Science.
- Snedecor, G. W. and W. G. Cochran. 1967. Statistical Methods. Iowa State Univ. Press, Ames, Iowa. p. 593.