Correlation and prediction: Interpreting the significance of r

Lysle E. Johnston, Jr., DDS, PhD

An interesting by-product of the growing popularity of the personal computer in contemporary orthodontic practice is the ease with which it can be used to generate cephalometric treatment forecasts (growth, surgery, etc.). As a result, a whole new generation of orthodontists must now come to grips with a venerable question: easy or hard, expensive or cheap, are predictions worth doing? Given the reasonable assumption that orthodontics is a useful service, some might regard a little electronic showmanship as an acceptable and entertaining means of "educating" the patient. The more intellectually fastidious, however, might not agree. Moreover, should the treatment results differ markedly from the forecast that "sold the case," the patient might well decide to examine the discrepancy in a court of law. For those who are new to the Delphic art of cephalometric prediction, a few brief comments about accuracy are perhaps in order.

The purpose of a prediction scheme is to forecast change. Accordingly, it is reasonable to expect the predicted increments to bear some sort of relationship—preferably a close relationship—to the changes that actually occur. The clinician, therefore, should have some way of judging, on the one hand, how close, and, on the other, how close is close enough. Correlation is the statistical tool most commonly used to answer these questions.

The Pearson product-moment coefficient of linear correlation, ρ, is a dimensionless index of the extent to which two characteristics—for example, a predictor variable and an increment of change—vary together. Like ρ, its sample estimate, r, varies between +1 (a perfect positive relationship) and -1 (a perfect negative relationship). Statistical significance—the probable existence of some type of relationship—can be inferred from the test statistic

$$r\sqrt{(N-2)/(1-r^2)}$$

which has a *t*-distribution with N-2 degrees of

Abstract

Given the growing popularity of cephalometric programs for the personal computer, it is once again necessary for the specialty to confront the problem of prediction accuracy. The strength of the relationships upon which a prediction scheme is based is often assessed by means of the coefficient of linear correlation, *r*. Although it is common to judge the practical significance of a relationship by squaring the correlation coefficient, the present paper argues that the *index of forecasting efficiency*, the percentage reduction in error, is not only the more appropriate index, but also one that is easy to infer directly from *r*.

Key Words

Prediction • Correlation • Index of forecasting efficiency

Submitted: January 1993 **Accepted for publication:** June 1993 The Angle Orthod 63:273-276

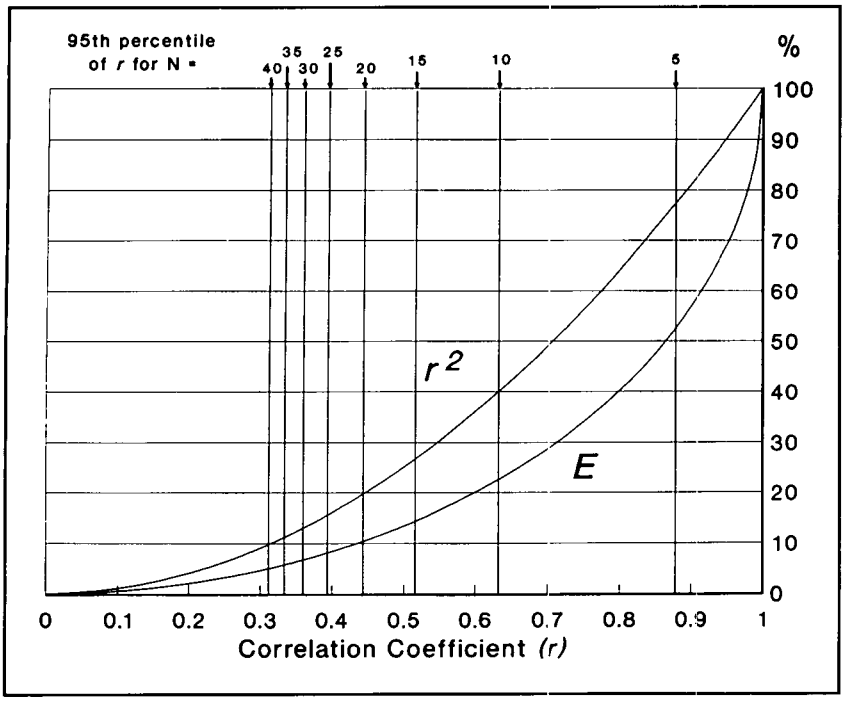

Figure 1

Figure 1
***E* and 100(r^2) expressed as functions of *r*. Note that for any level of *r*, *E* is markedly lower than 100(r^2). The vertical lines denote the 95th percentile of *r* for various common sample sizes (N).**

freedom. As demonstrated in Figure 1, even with a relatively modest sample size (say, 30 or so), the correlation between predictor and criterion variables can be quite low and still be statistically significant. As a result, even if the data argue that the relationship is real, one is left with the problem of deciding whether or not it has any "practical" significance.

To this end, it is often suggested that the correlation coefficient be squared and multiplied by 100 to yield an estimate of the percentage of the "variance" (actually, the sum of squares) of one variable that is shared ("accounted for") by the other. Although r^2 (the "coefficient of determination") thus provides something of a theoretical basis for evaluating the strength of the relationship between two variables, a "sum of squares" has no obvious intuitive meaning. Fortunately, if the relationship between two variables is being explored for the purpose of *predicting* one from the other, a concrete meaning easily can be imputed to *r*.

A prediction method is successful to the extent that it reduces error *vis-à-vis* some alternative. In orthodontics, the most common alternative to prediction is to "bet on the mean" – to expect the average growth increment, the average late mesial shift, the average amount of relapse, the average soft-tissue response, etc. The error of this method would be the standard deviation (S_Y) of the increment of change. If, instead, one were to use a linear regression equation to predict the individual increment, a measure of the error of this method would be the standard error of regression ($S_{Y|X}$, the root mean squared deviation from the regression line). The percentage reduction in prediction error achieved by the regression equation would thus be given by $100[(S_Y-S_{Y|X})/S_Y]$, a statistic that is sometimes referred as the *index of forecasting efficiency*, *E*.[1] For example, if $S_{Y|X}$ is half the size of S_Y, then the efficiency of a prediction based on this relationship would be 50%. Fortunately, it is not necessary to know either $S_{Y|X}$ or S_Y to calculate this useful percentage.

It is easy to demonstrate that a close approximation of *E* can be obtained directly from *r*: $E = [(N-2)/(N-1)](1-\sqrt{1-r^2})100$. With increasing N, (N-2)/(N-1) goes to unity, and the index becomes approximately $100(1-\sqrt{1-r^2})$. *E* thus constitutes an easily calculated, intuitive means of interpreting the practical significance of a given level of correlation. To provide a graphic comparison of *E* and 100(r^2) as measures of the potential utility of a relationship, the two statistics have been plotted as a function of *r*. Figure 1 shows that, for any given *r*, 100(r^2) is consistently larger than *E*. Thus, if it be granted that reduction in error, rather than sum of squares accounted for, is the "gold standard," and then it is clear that r^2 is a "rule of thumb" that greatly overestimates a correlation's practical value. More to the point, the index of forecasting efficiency demonstrates that it takes a surprisingly high linear correlation between dependent and independent variables to effect a marked reduction in error.

For example, a number of mixed dentition analyses use the width of the lower incisors – and a correlation of about 0.65 – to predict the size of the unerupted buccal segments.[2-5] Judged from the standpoint of *E*, these popular prediction schemes would thus have standard errors only about 25% smaller than the standard deviation of the mean buccal-segment width. For example, in one of these studies[5], $S_{Y|X}$ = 0.85 mm and S_Y = 1.12 mm. Forecasting efficiency, therefore, would equal 100[(1.12-0.85)/1.12] = 24.1%, an index that can also be calculated directly from the correlation coefficient. It may be inferred from the Figure that a 50% reduction in prediction error (i.e., *E* = 50%) would not be seen until *r* exceeds 0.85. In the context of cephalometric prediction, the correlation between double determinations in an error study is only slightly larger.[6] As a result, *E* = 50-60% may constitute very nearly the technical limit of cephalometric prediction, regardless of method.[6] But how can this pessimistic appraisal be reconciled with the fact that many contemporary prediction schemes are said to produce an accurate, clinically useful forecast?

The post-treatment face is merely the pre-treatment face modified by a relatively small increment of change. As a result, there is commonly a high, but spurious, part/whole correlation between measurements of size obtained before and after treatment. As a result, a prediction based on the pre-treatment face may *resemble* the actual outcome without accounting for any of the variance attributable to the increment of change. Given the error of the cephalometric method and the high level of correlation required to achieve an efficient prediction, informed skepticism and caution would seem to be in order. *Caveat emptor.*

Author Address
Lysle E. Johnston, Jr.
Department of Orthodontics and
Pediatric Dentistry
School of Dentistry
The University of Michigan
Ann Arbor, MI 48109-1078

L.E. Johnston, Jr. is the Robert W. Browne Professor of Dentistry and Chairman, Department of Orthodontics and Pediatric Dentistry, The University of Michigan, Ann Arbor, MI.

Supported by N.I.D.R. grant DE08716; this help is greatly appreciated. In addition, the author wishes to thank Drs. Carroll-Ann Trotman, Fedon Livieratos, and James McNamara for their constructive comments during the preparation of this manuscript.

References

1. Guilford JP. Fundamental statistics in psychology and education. New York: McGraw-Hill Book Co., pp. 222-3, 1942.
2. Ballard ML, Wylie WL. Mixed dentition case analysis—estimating size of unerupted permanent teeth. Am J Orthod 1947;33:754-9.
3. Bolton WA. Disharmony in tooth size and its relation to the analysis and treatment of Malocclusion. 1958;28:113-130.
4. Hixon EH, Oldfather RE. Estimation of the sizes of unerupted cuspid and bicuspid teeth. Angle Orthod 1958;28:236-40.
5. Tanaka MM, Johnston LE. The prediction of the size of unerupted canines and premolars in a contemporary orthodontic population. J Am Dent Assoc 1974;88:798-801.
6. Johnston LE. A statistical evaluation of cephalometric prediction. Angle Orthod 1968;38:284-304.

Commentary: Correlation and prediction

Brian G. Leroux, PhD; Douglas S. Ramsay, DMD, PhD, MSD

The article by Johnston provides a useful service by recommending the evaluation of the percentage of explained deviation (not variance) between two variables by using the index *E*. The percentage of explained variance (100 times r^2) may be difficult to interpret because variance is measured in squared units instead of the original units of measurement. This occurs because the variance for a sample is the sum of the squared deviations of the values from their mean which is then divided by the number of values minus one. Thus, for example, the mean amount of time it takes a group of people to read this commentary would be measured in minutes but the variance would be in squared minutes. However, the square root of the variance returns the estimate of variability to the actual measurement units; this transformed estimate of variability is called the standard deviation.

The relationship between r^2 and *E* is analogous to the relationship between variance and standard deviation. The mathematical analogy is made clear if we relate the complementary proportion of unexplained variance ($1 - r^2$) to the complementary proportion of unexplained standard deviation ($1 - E/100$). Just as standard deviation is the square root of the variance, $1 - E/100$ is the square root of $1 - r^2$. Likewise, for the same reasons that standard deviation is easier to interpret than variance, it may be easier to understand the predictive utility of an observed correlation using *E* rather than r^2. It should be remembered, however, that the predictive value of a correlation is not altered by the scale on which one chooses to measure it, although it might be easier to interpret on one scale than another.

The current article goes on to highlight the small percentage of standard deviation explained with

moderate degrees of correlation. It correctly stresses the limited practical (i.e., predictive) significance of a moderate correlation between two cephalometric measures. Clearly, there are many factors involved in the cephalometric prediction of growth and treatment effects and it is probably unrealistic to expect a single variable to be of much predictive value. The future of orthodontic predictions should focus on multiple factors that, when taken together, provide a greater predictive value than any one factor alone.

Authors' Response

Although the bulk of Drs. Leroux and Ramsay's comments need no reply, other than my thanks, I feel compelled to respond to their closing statement that, "The future of orthodontic predictions should focus on multiple factors that, when taken together, provide a greater predictive value than any one factor alone." On the face of it, this would seem a logical, perceptive admonition. Indeed, my first paper on the subject of prediction (which appeared in the pages of this journal some 25 years ago) featured just such a multivariate approach. Unfortunately, multiple regression proved inadequate to the task then, and this failure along with many others before and since provide empirical evidence that true prediction (i.e., the ability to account for individual variation in future increments of growth) may exceed the capabilities of the cephalometric technique. There is also considerable theoretical justification for this conclusion.

For example, why should we assume that measurements derived from radiographic shadows must contain, either individually or collectively, useful information about the probable pattern of future growth? In my opinion, it would be remarkable if the biology of facial growth were that simple. In turn, any lack of information would be compounded by the unavoidable consequences of cephalometric error (whose magnitude may approach that of the growth increments we would predict) and sensitivity to initial conditions (a basic element of chaos theory). Regardless of technology, therefore, I would argue that these limiting factors will conspire to defeat even our most imaginative attempts at generating efficient, individualized cephalometric growth predictions. Once again, *caveat emptor*.

Lysle E. Johnston