A Simplified Introduction to Correlation and Regression

K. L. Weldon
Department of Mathematics and Statistics
Simon Fraser University
Burnaby, BC. Canada. V5A 1S6

e-mail: weldon@sfu.ca

Author's Footnote: K. L. Weldon is Associate Professor, Department of Mathematics and Statistics, Simon Fraser University, Burnaby, BC, Canada V5A 1S6 (E-mail: weldon@sfu.ca)

Abstract:

The simplest forms of regression and correlation are still incomprehensible formulas to most beginning students. The application of the technique is also often misunderstood. The simplest and most useful description of the techniques involve the use of standardized variables, the root-mean-square operation, and certain distance measures between points and lines. With simple regression as a correlation multiple, the distinction between fitting a line to points, and choosing a line for prediction, is made transparent. Prediction errors are estimated in a natural way by summarizing actual prediction errors. The connection between correlation and distance is simplified. Few textbooks make use of these simplifications in introducing correlation and regression.

KEY WORDS: Prediction error, Distance, Root-mean-square, Standardized Variables, Standard Deviation

1. INTRODUCTION

The introduction to associations between two quantitative variables usually involves a discussion of correlation and regression. Some of the complexity of the formulas disappears when these techniques are described in terms of standardized versions of the variables. This simplified approach also leads to a more intuitive understanding of correlation and regression. More specifically, the following facts about correlation and regression are simply expressed:

The correlation r can be defined simply in terms of z_x and z_y, $r= \Sigma z_x z_y/n$. This definition also has the advantage of being described in words as the average product of the standardized variables. No mention of variance or covariance is necessary for this. The regression line $z_y=r\ z_x$ is simple to understand. Moreover, the tendency of regression toward the mean is seen to depend directly on r. The appearance of a scatter plot of standardized variables depends only on r. The illusion of higher correlation for unequal standard deviations is avoided. The prediction error of a regression line is the distance of the data points from the regression line. These errors may be summarized by the root mean square of the vertical distances: in standardized variables this is $(1-r^2)^{1/2}$. Correlation is related to the perpendicular distance from the standardized points to the line of slope 1 or -1, depending on the sign of the correlation. In fact the root mean square of these perpendicular distances is $(1-|r|)^{1/2}$.

The key to these simplifications and interpretations is an understanding of the standardization process. For this it is necessary for students to understand that a standard deviation really does measure typical deviations. This is aided by the use of the "n" definition of the standard deviation: $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$

The average squared deviation is apparent, and the square root of this is a natural step to recover the original units. So the standardized data $\frac{(x_i - \overline{x})}{s}$ is the number of these "typical" deviations from the mean. This describes the measurement relative to the (sample) distribution it comes from.

For students who must deal with traditional courses and textbooks using a strictly formula-based approach, it may be necessary to use the suggestion here as a first introduction. Once this simple introduction is accomplished, the more traditional approach could still be used, and shown to yield essentially the same results. The understanding that comes with the simplified approach, along with the calculation ease using statistical software, but with trivially different arithmetic, seems to be a combination that avoids confusion among students. The simplified approach suggested here has been used in many semesters of an introductory course based on the text by Weldon(1986), and no n vs n-1 confusion has been noted by the students or by instructors of subsequent statistics courses.

2. DETAILS

The definition $r= \Sigma z_x z_y/n$ assumes that the "n" definition of the standard deviation is used. A similar definition using the "n-1" definition of the standard deviation would require the n in the denominator to be replaced by n-1. To be able to describe the correlation as the average product of the z's not only simplifies the formula, but allows the student to think of the scatter plot quadrants as determining the correlation. The effect of outliers can be gauged more simply than with the original units formula.

It is important for students to realize that the regression line is not a simple curve-fit to the points, but rather a line designed for prediction. The formula $z_y=r\ z_x$ makes this quite clear since the "curve fit" $z_y=z_x$ is usually a line of greater slope than the regression line. Moreover the lines $z_y=r\ z_x$ and $z_x=r\ z_y$ (or $z_y=(1/r)\ z_x$) are clearly not the same line. Another effect simplified by this approach is that of regression toward the mean, with the predicted z_y less extreme than z_x.

Regression predictions can be made with the regression equation expressed in original units, but the direct use of $z_y=r\ z_x$ seems a viable alternative. The given value of X is easily converted into a z_x value, the prediction of z_y simply obtained, and then converted back into original units. While this involves a bit more arithmetic, the conceptual simplicity involving only the standardization idea and the r multiplier make this approach preferable for the novice. Note also that the error of prediction can be done using the $\sqrt{1-r^2}$ on the z scale, and transforming to original units.

It is well known that stretching one scale of a scatter plot can increase the apparent correlation (even though the correlation is actually unchanged). Portraying data in their standardized scale removes this illusion. It also makes the point that the correlation does not depend on the scales of the variables. Moreover, "banking to 45°" is recommended for graphical assessment. (Cleveland (1993)).

The distance of a point (x_0,y_0) to a line $y=rx$ in the direction of the y axis is

|y_0-r x_0|. Thus for standardized units we have the root mean squared distance as $\sqrt{\frac{1}{n}\sum\left(z_y - rz_x\right)^2}$. Expanding and using $\frac{1}{n}\sum z^2 = 1$ produces the well-known result that the root mean square distance of the data from the regression line is $\sqrt{1-r^2}$ times the standard deviation that in this case is 1. The condition $\frac{1}{n}\sum z^2 = 1$ only depends on the use of the "n" definition for the standard deviation - with the "n-1" definition of r and the standard deviation, the same result is true.

The minimum (i.e. orthogonal) distance of a point (x_0,y_0) to a line ax+by+c=0 is |ax_0+by_0+c|/$(a^2+b^2)^{1/2}$. The line of slope "(sign of r) 1" in standard units is z_x - (sign of r)z_y=0 and the distance of a point (z_x, z_y) to this line is therefore: $\frac{\left|z_x - (\mathrm{sgn}(r))z_y\right|}{\sqrt{2}}$. The root mean square of these distances is:

$$\frac{1}{\sqrt{2}}\left(\frac{\left(\sum z_x^{\ 2} + \sum z_y^{\ 2} - 2\,\mathrm{sgn}(r)\sum z_x z_y\right)}{n}\right)^{1/2} = \sqrt{1-|r|}.$$

This result appeared in Weldon (1986).

3. "n" DEFINITION OF SAMPLE STANDARD DEVIATION

This "n" definition simplifies a lot of things in teaching statistics. The justification for the more common "n-1" definition is based on the unbiasedness of s^2 for estimating σ^2, which is not really relevant for estimation of σ. One could even question the need for unbiasedness when it costs in terms of mean squared error. The "n" definition is easier to explain and has smaller mean squared error.

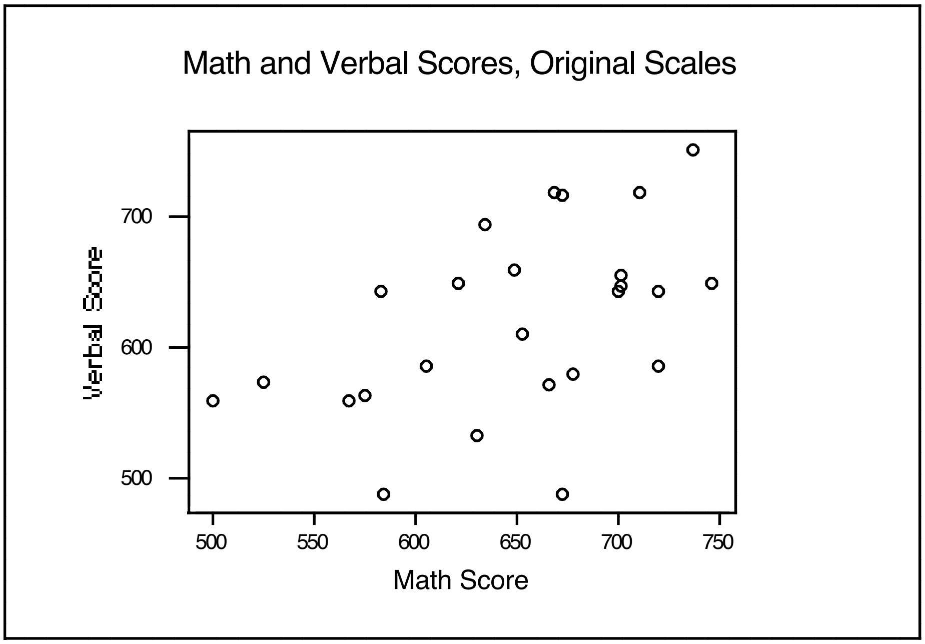
Some instructors are reluctant to use the "n" definition of the sample standard deviation, since it complicates the discussion of the t-statistic. But actually, if the t-statistic is defined in terms of the "n" definition of the sample standard deviation, the divisor "n-1" appears in its proper place as a degrees-of-freedom factor, preparing the student in a natural way for the chi-square and F-statistics: $t = \frac{(\bar{x}-\mu)}{s/\sqrt{n-1}}$

s in this formula is $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ as before. Nevertheless, for instructors who wish to stick with the "n-1" definition, the approach given in the rest of this paper to correlation and regression will still hold together.

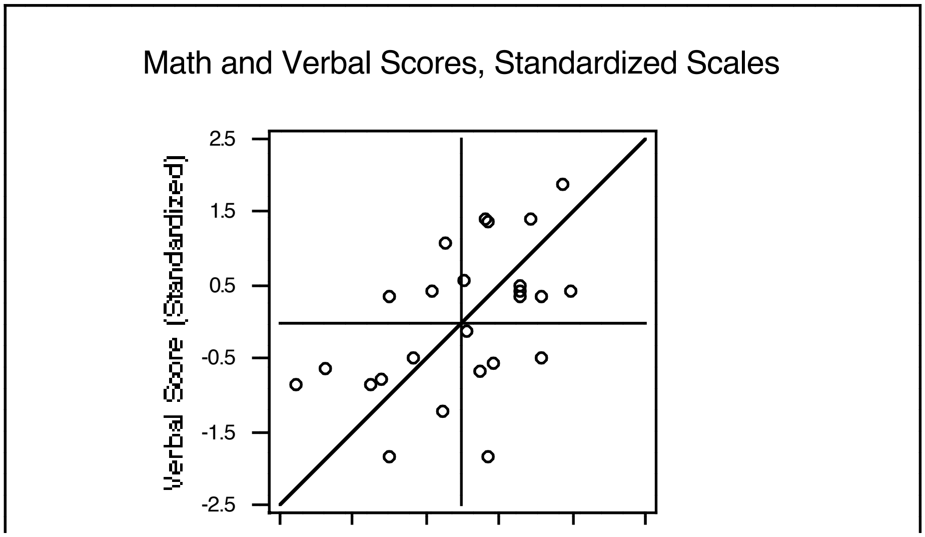
Another reason for avoiding the n-definition is the confusion that might be caused by the majority preference for the n-1 definition in other textbooks. However, once the idea of standard deviation is understood through the simplest approach, the existence of variations may not be so disturbing. The n-1 definition can be viewed in regression contexts as an "improvement" on the n-definition, and its extensions to the multi-parameter case would likely be accepted without too much consternation.

4. AN EXAMPLE

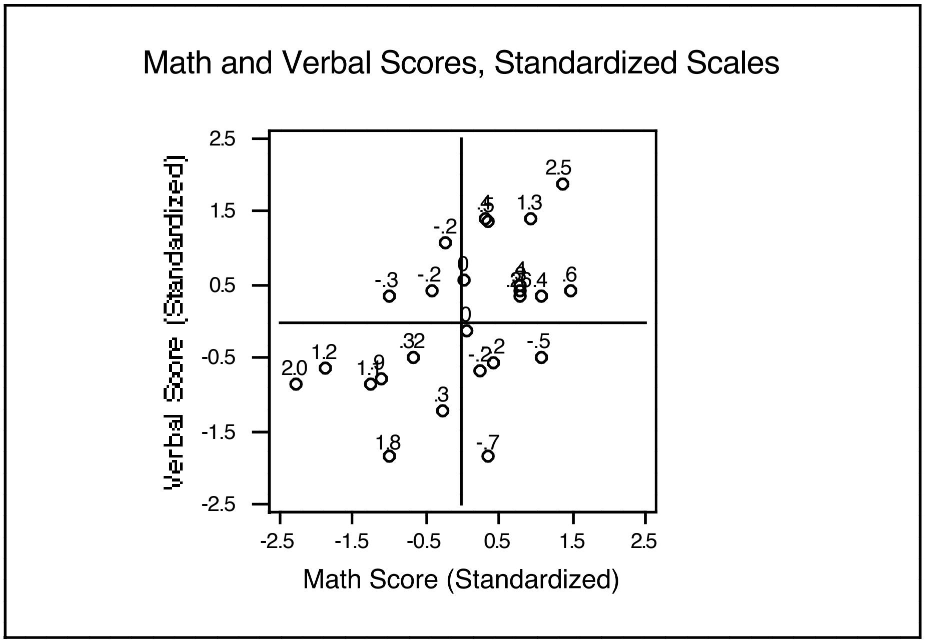
To illustrate the above formulas, consider the following data set relating performance in mathematics with performance on a verbal test. The data was sampled from a larger data set in MINITAB (See Reference). The first step for the student is to plot the data, and MINITAB produces a plot as follows using the default scaling.



If the variables are centered, and the scales equalized, this plot becomes:



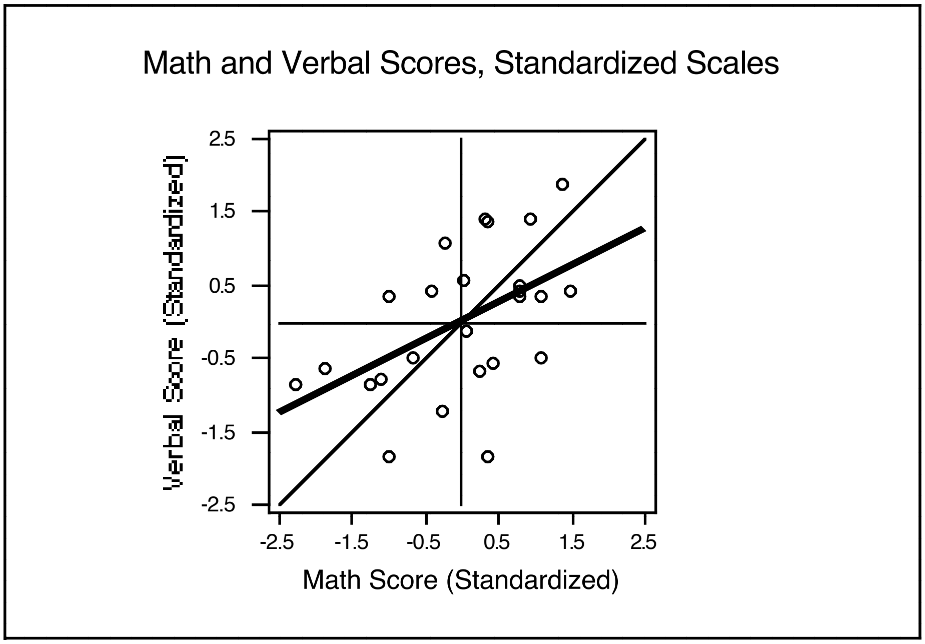
In this scale, it can be seen that the perpendicular distances of the points from the line of slope one (which may be called the point-fit line or the SD line) are usually less than one but greater than a half. The formula says that the root mean square of these distances is $= \sqrt{1-|r|}$, and in this case the sample correlation r is approximately 0.5 so that the rms distance is 0.7, which is about what we expect from visualizing the graph. Another observation from the graph could be made concerning the average product. The average product is the correlation, and the idea of this can be gleaned from a graph such as the following, in which the points are annotated with the product of the standard scores:



The fact that the average product is 0.5 is not obvious, but one can at least see which quadrants must have the largest contributions to the average in order that the correlation be positive.

Another feature of the "average product" definition of the correlation is the ability to detect outliers. From the graph it can be observed that a point at (-1.5, 1.5) seems not to belong to the oval shaped scatter, even though the individual values of -1.5 and 1.5 on either variable are not unusual. Such a point would be in the extreme lower tail of a dotplot of the products of the standardized variables, confirming from this definition of correlation that the point is unusual. Note that the addition of this one point would reduce the sample correlation from .50 to .38.

The regression line for predicting the Verbal Score from the Math Score is, in the standardized scale, z_V=r z_M, and graphically it can be shown as:



This shows clearly the difference between the "fit-to-the-points" line, and the regression line for predicting V from M. Moreover, the line for predicting M from V is clearly a different line.

As a final step in using the z_V=r z_M regression equation, it is necessary to convert the regression line back to the original units. First we need to record the mean and SD of each variable: mean(V)=619 SD(V)=71; mean(M)=649, SD(M)=65. Then substituting directly into z_V=r z_M, one gets (V-619)/71=0.5 (M-649)/65. For example, if M=700, the right side is 0.5*51/65 = .39 so that the predicted V is 619+.39*71 = 647. For many predictions, one may need the explicit equation: V=619+71*0.5*(M-649)/65 which simplifies to V=265+.54M. Compared to the arithmetic of formulas for the mean and slope, which tend to obscure the operation, this is relatively straightforward.

5. RELATED WORK

Most textbooks introduce correlation and regression via formulas. For example, Moore and McCabe (1993) use the n-1 definition of the correlation, and define the regression slope in terms of the unstandardized variables. Freedman, Pisani and Purves (1998) do better but neither text use the fact of two regression lines to make the point that a regression line is not a fit to points but rather aimed at prediction.

The explicit interpretation of correlation in terms of distance does not appear in the "Thirteen ways" summary article by Rodgers and Nicewander (1988), nor in the follow-up papers by Rovine and von Eye (1997) and Nelsen (1998), even though this interpretation appears one of the most intuitive.

6. SUMMARY

When data is expressed in standardized form, correlation and regression methods can be described very simply. The difference between fitting a line to points, and regression, is clarified by this simpler presentation. The use of n-1 in formulas for the standard deviation and the correlation coefficient is an unnecessary complication.

REFERENCES

1. Cleveland, W.S. (1993) *Visualizing Data*. Chapter 3. Hobart Press. Summit, NJ.
2. Weldon, K.L. (1986) Statistics: A Conceptual Approach. Chapter 5, p 144. Prentice-Hall, Englewood Cliffs, NJ.
3. Minitab, Inc. 3081 Enterprise Drive, State College, Pa. USA. 16801-3008.
4. Moore, D.S. and McCabe, G.P. (1993), *Introduction to the Practice of Statistics*.Second Edition. Chapter 2. Freeman. New York, NY.
5. Freedman, D, Pisani,R. and Purves, R. (1998) Third Edition. *Statistics*. Norton. New York, NY.
6. Rodgers, J.L. and Nicewander, W.A. (1988), Thirteen Ways to Look at the Correlation Coefficient, *The American Statistician*, 42, 59-66.
7. Rovine, M.J. and von Eye, A (1997), A 14th Way to Look at a Correlation Coefficient: Correlation as the Proportion of Matches, *The American Statistician*, 51, 42-46.
8. Nelsen, R.B. (1998), Correlation, Regression Lines, and Moments of Inertia, *The American Statistician*, 52, 343-345.