A Practical Look at Correlation

Rudy A. Gideon and Carol A. Ulsafer

Dept. of Mathematical Sciences, University of Montana, Missoula, MT, 59812

Abstract

The word correlation in general indicates that two quantities are related and somehow linked together. Unfortunately, the Pearson correlation coefficient measures only linear relationships and yet has become synonymous with the general concept of correlation. This is entirely too narrow and so there is a pressing need for a correlation function that is independent of linearity and measures the general relationship between variables that are related in any manner whatsoever. The desired new correlation measure should also mimic Pearson's correlation coefficient when the data are bivariate normal in distribution. Although it has not been widely recognized, the square root of the coefficient of determination can fill this role. For actual experiments in which the data are adequate, that is, allow the response variable to follow the shape of the model and allow a measure of local variation, the estimate of the coefficient of determination considered below appears to work quite well. This estimate does not rely on least squares theory. In addition, a second rank based robust correlation measure is presented. The focus of this paper is to introduce not only a wider concept of correlation but also a general, easy to understand estimation procedure which can be used in teaching the beginning student and to give practitioners of statistics a helpful tool in studying the relationship between variables. A third degree polynomial model, a real-data nonlinear regression model, and the bivariate normal have been chosen to illustrate the significance of the new correlation coefficients.

1. Introduction

The coefficient of determination is defined as $\frac{s_y^2 - s^2}{s_y^2}$ or $1 - \frac{s^2}{s_y^2}$, where s_y^2 is the overall variance of the dependent variable and s^2 is the local variance for a fixed x, the independent variable. This parameter is well known but not commonly used. In this work the square root of this quantity is considered as a general correlation coefficient that works for both linear and nonlinear models. For real data problems, the variance s^2, assumed homogeneous throughout this paper, is unknown so an estimator is needed. In the case of nonhomogeneous error variance, a weighted average of the error variances would have to be used. Ideally the estimator should be the same for all models.

The regressor or independent variable is denoted by x and the response or dependent variable by y. The data points are indicated with subscripts and the collection of n

points is (x, y). For two data points (x_1, y_1) and (x_2, y_2), with $\bar{y}$ the average of the two y-values, $(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 = \frac{1}{2}(y_1 - y_2)^2$. The right hand side is unbiased and has one degree of freedom for the estimate of the variance of the y variable. Assume the model is $y = f(x) + error$ and let the error be denoted by e. Then $y_1 = f(x_1) + e_1$ and $y_2 = f(x_2) + e_2$. The random or residual error of the model is denoted by $\text{var}(e) = s^2$. Hence, if $x_1 = x_2$, the best possible case, then $\frac{1}{2}(y_1 - y_2)^2 = \frac{1}{2}(e_1 - e_2)^2$ and this last term estimates s^2 with one degree of freedom. If the x-data is distributed throughout the area of interest for the model, with paired values close together, then a good measure of correlation can be obtained from these ideas. These simple observations form the foundation for the definitions.

2. A new generalized correlation coefficient derived from an estimate of the coefficient of determination

Let the x data be numbered from 1 to n, from least to greatest, and plotted with the corresponding y values. Thus, x_1 is the smallest x and y_1 is the response value for that x, x_2 is the next smallest and is paired with y_2 and so forth. The y-values are arranged by the x-values they are paired with. Let n2 be the greatest integer in n/2 (the notation n2 is used for simplification).

With the data labeled as indicated above, and letting $\bar{y}$ be the mean of all the y-data, the definition of the estimate of the coefficient of determination is

$$r_{gcc}^2 = 1 - \frac{2(n-1)\sum_{i=1}^{n2}(y_{2i-1} - y_{2i})^2 / 2}{n\sum_{i=1}^{n2}((y_{2i-1} - \bar{y})^2 + (y_{2i} - \bar{y})^2)}. \qquad (2.1)$$

The square root of this is a correlation estimate, called the generalized correlation coefficient. A nonparametric version, which is robust, and uses the individual terms in the above summations is appropriate when some outliers or a few highly variable points are in the data. It is presented below. Because the summations terminate at n2, if the number of observations is odd, the last actual data point is not used. Thus, n in formula (2.1) refers to the number of data points in the summations not the number of observations.

Several observations are now made to motivate, justify, and explain this definition.

(a) The ratio of the terms in the two summations, $\frac{(y_{2i-1} - y_{2i})/2}{(y_{2i-1} - \bar{y})^2 + (y_{2i} - \bar{y})^2}$, would be one if $\bar{y}$ were the average of y_{2i-1} and y_{2i}. Because $\bar{y}$ is the grand mean, the ratio is always less than one, and the less it is the more indication that the data are dependent.

(b) The sum in the denominator divided by $n-1$ is the sample variance of the y-data for n even. For n odd the last data point is deleted because it has no paired point.

(c) If the data is bivariate normal with correlation coefficient r and population variances s_x^2 and s_y^2, then for x_{2i-1} and x_{2i} very close together, the expectation of $\frac{1}{2}(y_{2i-1}-y_{2i})^2$ is close to the conditional variance of y given x, $s^2=(1-r^2)s_y^2$. It follows that the expectation of the numerator sum is close to $\frac{n}{2}(1-r^2)s_y^2$. Because the expectation of the denominator divided by $n-1$ estimates s_y^2, it follows that the ratio in r_{gcc} is close in expectation to $1-r^2$ and, one minus this ratio, estimates r^2. If $r=0$, then the expectation of $\frac{1}{2}(y_{2i-1}-y_{2i})^2$ is s_y^2 because the two y - values are now independent and the ratio estimates one.

(d) From comment (c) it is clear that the formula for r_{gcc} mimics the regression version of the correlation after a model has been fit. That is, $1-\frac{SS(residual)}{SS(total)}$ is the square of the correlation coefficient. (SS denotes sum of squares.)

(e) When a curvilinear or nonlinear model, $y=f(x)+error$, is analyzed with pairs of repeat points for the regressor variable x, then as stated above each term in the numerator sum is estimating the residual variance, $\text{var}(e)$. It follows that r_{gcc}^2 is estimating the fraction of variation explained by the "correct " model.

(f) A formula for r_{gcc} is given if there is a designed experiment with repeat points; it is based on the within or error sum of squares in a one-way analysis of variance. Let the repeat points be $\{y_{ij}\}_{j=1}^{n_i}, i=1,2,\ldots,k$ where there are k groups of repeat points for k different x values with group i having n_i points. Let

$$\hat{s}_e^2=\frac{\sum_{i=1}^{k}\sum_{u=1}^{n_i}(y_{iu}-\bar{y}_i)^2}{\sum_{i=1}^{k}n_i-k}$$

and let $\hat{s}_y^2$ be the sample variance for y.

Then $r_{gcc}^2=1-\frac{\hat{s}_e^2}{\hat{s}_y^2}$.

3. Illustration of the generalized correlation coefficient by simulations

Three examples are now given.

3.1 Bivariate normal

First the bivariate normal is used to illustrate properties of r_{gcc}. Below is a table of the difference between Pearson's r_p and the generalized correlation coefficient r_{gcc}. As expected, r_{gcc} holds up very well for sample sizes over 50 and correlations over 0.5. It starts to be less accurate for smaller sample sizes, less than 20, and smaller correlations. There were 500 simulations per comparison.

Difference between $\lvert r_p \rvert$ and r_{gcc} over the bivariate normal distribution						
r	sample size 100		sample size 50		sample size 20	
	mean	SD	mean	SD	mean	SD
0.9	0.00455	0.01601	0.01086	0.02646	0.03330	0.06130
0.7	0.01080	0.06046	0.02099	0.09336	0.05667	0.16561
0.5	0.02140	0.12228	0.04817	0.18489	0.07384	0.23728
0.3	0.03096	0.17972	0.03420	0.22055	0.02080	0.27146
The standardized bivariate normal was used above; below the SDs were 1.5						
0.9	0.00399	0.01698	0.01089	0.02634	0.03468	0.05741
0.7	0.00540	0.05495	0.02142	0.09220	0.06090	0.17269
0.5	0.03875	0.13187	0.03405	0.16977	0.06069	0.22753
0.3	0.02833	0.18861	0.02476	0.21679	0.04866	0.26067

Table1: Comparison of the new nonlinear correlation coefficient with Pearson's

When r becomes zero or near zero, the expectation of the ratio in the definition (2.1) of r_{gcc} becomes just the ratio of two quantities both estimating s_y^2, and this ratio can become greater than one. In this case just the ratio could be examined because zero is the only possible estimate of r. A short comparison was done of the mean square error of r_{gcc} and r_p with standard deviations both 1.5. For r at 0.5, 0.75, and 0.9 and sample sizes of 50 and 100 with 50 simulations each, the mean square error of r_{gcc} was roughly twice that of r_p. For example for n =100, $r = 0.9$ and 50 simulations the average mean square error of r_p was 0.000197 and of r_{gcc} was 0.00033. With the same n and 50 simulations but $r = 0.5$ the average mean square error of r_p was 0.0023 while r_{gcc} was 0.0104. The above shows that this new general correlation coefficient does reasonably well for bivariate normal for moderate to large sample sizes and the correlation not too small. It is only marginally less valuable as a

correlation estimate than r_p in this case, but its true worth is demonstrated in the next two examples.

3.2 A cubic model

The second simulation example has the model $y = f(x) + error$ where $f(x) = 2x^3 + 3x^2 - 5x - 6$. This cubic has local minimum and maximum at $-\frac{1}{2} \pm \frac{\sqrt{39}}{6}$, or approximately 0.54083 and -1.54083, respectively. The three roots are -2, -1, and 1.5. It diverges to minus infinity as x decreases and to plus infinity as x increases. Two types of data each with 26 observations will be considered. In the first, a designed experiment, there are two repeat y observations for fixed x-values spread uniformly over the area of interest. In the second there is one y observation for each randomly chosen x. The repeat point design allows for the best estimate of correlation. This number of observations seemed sufficient to illustrate the desired features. If there are more repeat points than two at some x values then of course the numerator in the definition of r_{gcc} needs to be adjusted using formula (f) above.

These ideas are related to pure error in Draper and Smith in the discussion of fits of regression models and so the reader can resort to that area for further insight. In any case, given an x, $y = f(x) + error$. To study the properties of this new correlation coefficient, both well-behaved and erratic errors are introduced. The former are generated via the normal distribution and the latter via the Cauchy distribution. The value of the new robust rank based correlation coefficient, defined below, is demonstrated when the errors are Cauchy.

In this model x is restricted to the interval -3.5 to 2.5 so the data could show an upward trend and Pearson's correlation coefficient would be significant. The formula for r_{gcc} can be changed to compute a correlation coefficient adjusted for linear regression. All that is necessary is to replace the denominator by the residual sum of squares from the linear regression and the $n-1$ replaced by $n-2$. The examples contain this adjustment; in other words, is there still correlation when a linear component has been taken into consideration? This example was constructed for a positive answer.

The x values for the designed case were from -3.5 to 2.5 by increments of 0.5 and for each x there were two random observations. For a finer grid the results for the new generalized correlations would be even better. The errors had means of zero and scale factors of 5,10, and 15. Both the Normal and Cauchy distributions were used. The statistical package S-Plus was used for the simulations. The sample size was 26 with two observations at each point and 500 simulations were run for each distribution and scale setting. The means and standard deviations of the 500 simulations are listed in Table 2 below for five correlations:

- Pearson's
- r_{gcc}
- the generalized correlation when adjusted for linear regression, $r_{gcc} - adj$ (in which the denominator of r_{gcc} is replaced by the residual sum of squares and degrees of freedom are $n - 2 = 24$)
- the robust correlation estimate $r_{gd} - rob$ (defined after the examples)
- The fifth correlation coefficient appearing in the last row is a symmetrized version of the generalized correlation coefficient. It is also explained below.

This 3rd degree curvilinear model has population CODs depending upon how the x-values are generated and the form of the error variable. These are computed for several cases so that the simulations can be better judged. In the second case, the 26 x-values are from a uniform random variable, U, over the interval (−3.5, 2.5). In all cases homogeneous variation is assumed on y for a fixed x and it is denoted by s^2.

In this second case the variation formula can be used:
$Var(Y) = Var(E(Y|X)) + E(Var(Y|X))$. Now $Var(Y|X) = s^2$ and so the second term is just s^2. The first term is $Var(f(X)) = E(f(X)^2) - E^2(f(X))$. The first term is a function of the moments of U up to the 6th power and the second up to the third. The formula for the moments of U is $E(X^k) = \frac{(2.5)^{k+1} - (-3.5)^{k+1}}{(k+1)(2.5+3.5)}$, $k = 1,2,\ldots,6$. This was used to obtain $Var(Y) = 122.121 + s^2$. Then the COD is $1 - (s^2/(122.121 + s^2))$. The square root of this COD is the correlation; it is illustrated for $s = 5, 10, 15$. The values are 0.9111, 0.7415, and 0.5931, respectively.

In the first or the designed case, two, or in general, an equal number observations are taken at each of the 13 equally distant points. The X variable is treated as a discrete random variable with each of the 13 points having an equally likely probability. Let normal Z represent the error with zero mean and variance s^2. Then for x_i values −3.5(0.5) 2.5,

$$E(Y) = E(f(X) + Z) = E(f(X) + 0 = \frac{1}{13}\sum_{i=1}^{13} f(x_i) = -3.$$

$$E(Y^2) = E((f(X) + Z)^2) = E(f^2(X)) + E(2Z)E(f(X)) + E(Z^2) =$$

$$\frac{1}{13}\sum_{i=1}^{13} f^2(x_i) + 0 + s^2 = 234 + s^2.$$

$$Var(Y) = E(Y^2) - E^2(Y) = 234 + s^2 - (-3)^2 = 225 + s^2.$$

Now the COD is $1-(s^2/(225+s^2))$. The square roots, or the correlations, of these CODs are illustrated for $s = 5, 10, 15$. The values are 0.9487, 0.8321, and 0.7071 respectively. Compare these to the mean of the simulations for r_{gcc} below.

Two additional correlation coefficients are included in the following tables for comparison sake: $r_{gd}-rob$ and r_{sgcc} are the notations for the robust rank based and symmetric correlation coefficient. In order to maintain the flow of the information and to allow the reader to become more familiar with the main ideas, the definitions and explanations of these come after the examples.

Correlation coefficients for a cubic model, paired repeat points							
corr		Normal scale factors			Cauchy scale factors		
coef		5	10	15	5	10	15
Pearson	mean	0.717	0.631	0.535	0.391	0.270	0.242
	SD	0.041	0.078	0.114	0.256	0.227	0.213
r_{gcc}	mean	0.951	0.825	0.699	0.497	0.387	0.345
	SD	0.019	0.076	0.160	0.293	0.253	0.230
$r_{gcc}-adj$	mean	0.899	0.690	0.574	0.427	0.352	0.333
	SD	0.040	0.154	0.233	0.262	0.224	0.212
$r_{gd}-rob$	mean	0.945	0.850	0.765	0.822	0.728	0.668
	SD	0.050	0.188	0.233	0.170	0.218	0.240
r_{sgcc}	mean	0.894	0.755	0.630	0.516	0.416	0.402
	SD	0.034	0.084	0.161	0.231	0.235	0.228

Table 2: Correlation comparisons for designed experiment: four correlations and two error distributions

Using the Cauchy distribution may seem extreme, but in reality it may actually mimic some experimental situations better than the probably too good normal distribution. It is well known that 10-20 percent of most real life data is questionable; many methods require considerable expertise and expenditure of time deciding which of these data to retain and this makes the results somewhat arbitrary. The robust correlation with Cauchy error shows that reasonable results can be had without manipulation of the data. A study of this table makes it clear that all four new correlations add a considerable amount of information about the relationship between the x and y data before any model is proposed. For the true cubic model a typical random sample is shown in Figure 1 with normal error and scale factor 10.

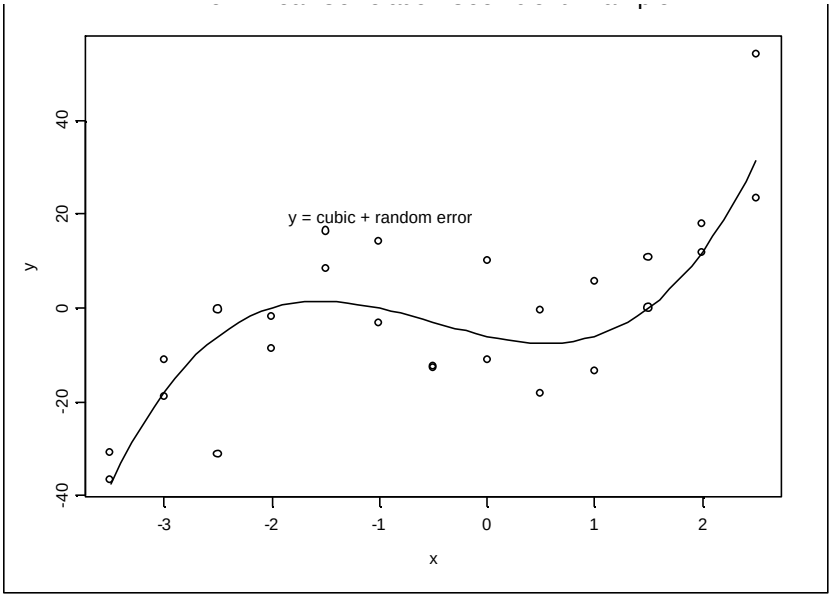

Figure 1: The cubic model with repeat observations

Table 3 below mimics the Table 2 above in which scale factors 5, 10, and 15 are used for both the normal and Cauchy distributions. The x data were selected randomly by using 26 observations from the uniform distribution over -3.5 to 2.5 and again 500 simulations were run. As one would expect the correlation information about the x-y relationship is less because of confounding between the error and the model, but the general pattern about the comparison of the effectiveness of the correlations remains much the same. A bit unexpectedly the robust correlation coefficient did better than expected but had lots of variation for the Cauchy case. Additional runs yielded similar reults.

Correlation coefficients for a cubic model, random uniform x data							
corr		Normal scale factor			Cauchy scale factor		
coef		5	10	15	5	10	15
Pearson	mean	0.584	0.568	0.495	0.219	0.264	0.202
	SD	0.055	0.098	0.129	0.208	0.218	0.199
r_{gcc}	mean	0.925	0.707	0.623	0.377	0.373	0.336
	SD	0.033	0.137	0.200	0.241	0.240	0.220
$r_{gcc}-adj$	mean	0.887	0.566	0.522	0.365	0.341	0.337
	SD	0.052	0.231	0.265	0.229	0.220	0.214
$r_{gd}-rob$	mean	0.847	0.781	0.707	0.716	0.699	0.657
	SD	0.194	0.209	0.288	0.222	0.224	0.261
r_{sgcc}	mean	0.862	0.649	0.569	0.376	0.414	0.366
	SD	0.041	0.132	0.192	0.225	0.235	0.220

Table 3: Correlation comparisons for random x - data: four correlations and two error distributions

3.3 A nonlinear regression model

This example is in all three editions of Draper and Smith. They used it to show how to estimate and test the parameters of a nonlinear regression. Here it is used to illustrate the generalized correlation coefficient, r_{gcc}. The model is $Y = a + (0.49 - a)e^{-b(X-8)} + e$. The parameters estimated are a and b and the error is e. The estimates were $\hat{a} = 0.39$ and $\hat{b} = 0.10$, so the fitted equation is $Y = 0.39 + 0.10e^{-0.10(X-8)}$. All of the data appears in the book so the new correlations can be computed. They are $r_p = -0.8651$, r_{gcc} =0.9466, $r_{gcc} - adj$=0.7725, and $r_{gd} - rob$= 0.9323. The new correlations deal with magnitude only so none of them are negative. If preferred, a negative sign could be appended; however, for many cases a negative sign would not make sense. It is clear from the three new correlation coefficients that the relationship between x, length of time, and y, amount of Chlorine, is not simply linear. First, r_{gcc} is larger in magnitude than r_p. Second, $r_{gcc} - adj$ is still large even after the simple linear regression effect is deleted. Third, the fact that $r_{gd} - rob$ is large may indicate that there may be a few data points of large variation .

In this data set there are 26 degrees of freedom for pure error and the sum of squares for this is (using the formula in (f) above) SSPE = 0.002367 (the book gave .0024), and the sum of squares for the response variable, amount of chlorine, is 0.03950 with 43 degrees of freedom (n is 44). This more refined generalized correlation coefficient is $\sqrt{1 - \frac{43(0.002367)}{26(0.03950)}}$ = 0.9491. Compare to 0.9466 above. So the suggested new generalized correlation seems to do well again even without being calculated with all repeat points. It is also possible to test that the suggested model is reasonably good by using the sum of squares about the regression, usually called residual SS. This SS is 0.00501, called SS(model) to distinguish it from the new correlation, (book value was 0.0050) with 42 degrees of freedom as suggested in the book. There should be very little correlation left if the model is good. A correlation is obtained by comparing SSPE to SS(Model) as follows: substitute in the denominator of the COD the error variation computed for the chosen model, SS(model). This is just like the linear regression check used earlier to see if there is correlation after a linear fit. The COD is $1 - \frac{42(0.002367)}{26(0.00501)}$ = 0.2368. This compares well to the approximate F test that was given in Draper and Smith. (SS(model)-SSPE)/16 = 0.000165 (the book was 0.00016) and SSPE/26 =0.00009104. So an approximate F(16,26) test is the ratio 0.000165/0.00009104 = 1.81. The upper 0.95 F point is 2.05. So the model gives a reasonably good fit. The book rounds everything to two digits as that is the accuracy of the data, and so their F was 1.8. In any case it is clear from their work or from this new correlation method, whose after regression COD was 0.2368, that the model adequately fits the data.

This reasonably good, but not great fit, may be due to the 3 to 5 points that were unusually far from the fitted model which was confirmed from the fact that the robust correlation coefficient was so large.

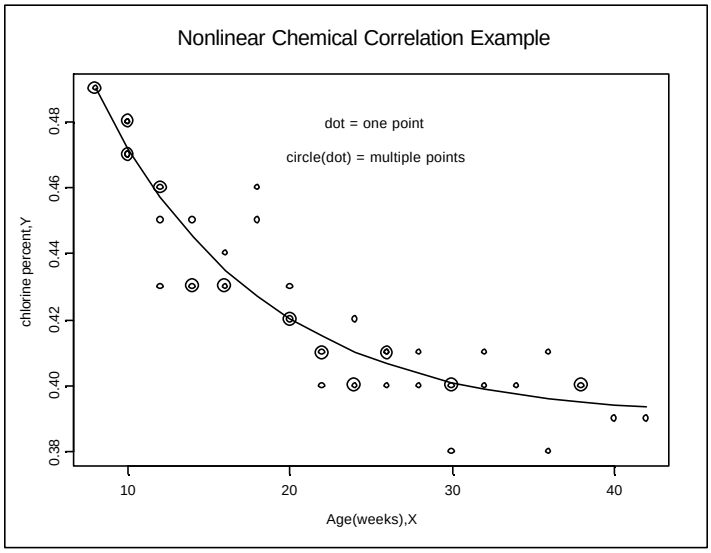

Figure 2: Chemical example with fitted curve and data points

The ordered $x-$data is:
8,8,10,10,10,10,12,12,12,12,14,14,14,16,16,16,18,18,20,20,20,22,22,22,24,24,24,26,26, 26,28,28,30,30,30,32,32,34,36,36,38,38,40,42
and the corresponding values for y are:
0.49,0.49,0.48,0.47,0.48,0.47,0.46,0.46,0.45,0.43,0.45,0.43,0.43,0.44,0.43,0.43,0.46,0.45,
0.42,0.42,0.43,0.41,0.41,0.40,0.42,0.40,0.40,0.41,0.40,0.41,0.41,0.40,0.40,0.40,0.38,0.41,0.40,0.40,0.41,0.38,0.40,0.40,0.39,0.39.

One further comparison is to compute the correlation of the model with the data, which is known as R^2. This is done by replacing the numerator in the GCC definition by the Residual SS from the model which was 0.00501and the SS of the y-data is 0.03950. So the model R^2 is $1-\dfrac{43(0.00501)}{42(0.03950)} = 0.8701$. The square root of this is 0.9328. For a good model fit this correlation should be close to r_{gcc} which from above is 0.9434. The difference between these two correlations is very small so this suggests that a realistic fit has been obtained.

3.3.1 A permutation test

In today's computer world it is not necessary to obtain the exact distribution of the test statistic. If it is assumed no relationship exists between variables, and here that

includes nonlinear relationships, then a permutation test can be done. In the third example, under a null hypothesis of no relationship, any permutation of the y-data is as equally likely as any other. It is impossible to produce all 44 factorial possibilities and then to see where the observed is. Instead a set of 1000 random permutations of the data was drawn and the new correlation computed on each permutation of the y, chlorine data. The correlation values were then ordered and the observed correlation was located. Inference can be drawn to reject an uncorrelated hypothesis if the observed correlation is in, say, the upper 5% set of correlation values. For r_{gcc} the observed value 0.9466 far exceeded the upper 99% point. For the robust version of the correlation, the observed value 0.9323 again far exceeded the upper 99% point.

4. The robust generalized nonlinear correlation coefficient

The robust version of the new generalized correlation coefficient is now explained. This work relies on the correlation estimation system (CES) that is currently being written up and submitted for publication. Detailed information is given in the Web site www.math.umt/gideon. Below is given the computation procedure and its result for the chosen examples.

This regression method via correlation (correlation regression) is performed on two sets of order statistics so as to estimate the ratio s_e/s_y as an entity rather than estimating numerator and denominator individually. The slope in the regression will be the estimate. For a robust estimate a robust correlation coefficient must be used; this work uses the Greatest Deviation Correlation Coefficient (Gideon and Hollister, 1987). Both the Greatest Deviation and Kendall's correlation coefficients are more robust than the Spearman correlation coefficient. So Kendall's would be a reasonable substitute. In order to eliminate any concern about ties, the tied value method appearing in Gideon and Hollister (1987) or in the papers on the Web site must be used. This method is completely general, working in all cases so that ties are no longer of any concern; the general background is given in the Web papers. An appendix gives S-Plus or R code that can be used.

In brief, the correlation regression idea is as follows. Let the index i represent the i^{th} order statistic. Assume there are two sets of data $t_i = m_1 + s_1 z_{1i}$ and $w_i = m_2 + s_2 z_{2i}$, $i = 1,2\ldots, n$ where z_{1i} and z_{2i} are the i^{th} order statistics from samples of standard normal variables. If e_i is the difference between z_{1i} and z_{2i} then

$z_{1i} + e_i = \dfrac{t_i - m_1}{s_1} + e_i = z_{2i}$ so substituting above for z_{2i},

$w_i = (m_2 - \dfrac{s_2}{s_1} m_1) + \dfrac{s_2}{s_1} t_i + s_2 e_i$, $i = 1,2,\cdots, n$. Thus the slope, s_2/s_1, is a ratio of scale parameters. Using any correlation coefficient r, the regression is done for data $\{t_i, w_i\}_{i=1}^n$ by solving for s in a "scale" correlation coefficient equation $r(t^o, w^o - st^o) = 0$, where t^o and w^o represent the ordered data in vector form. The

ratio s_2/s_1 is estimated by s. The "scale" correlation coefficient equation makes the t variable uncorrelated with the residuals of the regression in the sense of whatever correlation coefficient is being used.

In the actual situation the scale parameter for w is denoted by s_e and for t by s_y. Let the data be subscripted as in the definition of r_{gcc}, with $w_i = \left| \frac{y_{2i-1} - y_{2i}}{\sqrt{2}} \right|$, and $t_i = \sqrt{(y_{2i-1} - \bar{y})^2 + (y_{2i} - \bar{y})^2}$, $i = 1,2,\ldots,n2$; recall that n2 is essentially n/2. All the data are now ordered within their own group and a simple correlation regression is performed ordered w on ordered t. The slope of this regression is an estimate of s_e/s_y. As stated above, the robust Greatest Deviation Correlation Coefficient, r_{gd}, will be used. Least squares regression or Pearson's correlation coefficient regression is not appropriate as it is not robust. The scale regression equation, $r_{gd}(t^o, w^o - st^o) = 0$, is solved for slope s. Again the superscript means the data are ordered. Then the estimate of the correlation coefficient squared, or the COD, is $1 - s^2$. The fact that this works well for messy data is demonstrated above in the examples with Cauchy distribution error. The one anomaly in this method is that for small sample sizes or small correlation coefficients, s can be greater than one, indicating that no relationship has been found. In this case zero is used for the estimate of correlation in the simulations and these may or may not be entered into the averages and standard deviations.

A population definition of GCC, r_{gcc}, is motivated from the fact that all the information about the relationship between x and y is contained in s_e and s_y. The quantity $1 - s_e/s_y$ could be defined as the correlation coefficient. For the cases s_e =0 and 1, this correlation coefficient is 1 and 0, respectively. It also stays between 0 and 1. However, to be consistent with the historical development and normal theory the correlation coefficient is defined to be $\sqrt{1-(s_e^2/s_y^2)}$. Then if $y = f(x) + e$ and s_e^2 is the variance of the error, the definition of the square of a population generalized correlation coefficient is then $r_{gcc}^2 = 1 - s_e^2/s_y^2$.

5. A symmetric version, r_{sgcc}

In all of the above it was assumed that the interest was in y as the response or dependent variable and x as the predictor or independent variable. However, in the case of the bivariate normal, x and y should be treated equally. This means that interchanging x and y should give the same correlation value. This last section shows one way to symmetrize the definition of the generalized correlation coefficient. Essentially the computation interchanges x and y and recomputes the correlation. To

make a single value, the geometric mean of the two parts is used. Let *sy* be the sum of squares in the numerator of the earlier definition: $sy = \frac{1}{2}\sum_{i=1}^{n2}(y_{2i-1} - y_{2i})^2$. Recall that this depended on the y's corresponding to the ordered x's. Now do so in reverse, order the y's and let the x's correspond. That is, it is ordered like the graph of x on y where y is now the horizontal axis. With this in mind let $sx = \frac{1}{2}\sum_{i=1}^{n2}(x_{2i-1} - x_{2i})^2$. Also let var($x$) and var($y$) denote the usual sample variances. Then the definition of the symmetrized correlation coefficient is

$$r_{sgcc}^2 = 1 - \frac{2}{n}\sqrt{\frac{sx * sy}{\text{var}(x) * \text{var}(y)}}.$$

The square root is needed as there are now two terms both estimating $1 - r^2$. It is clear that $r_{sgcc}^2(x, y) = r_{sgcc}^2(y, x)$. It is perhaps somewhat surprising that this definition works well even for fairly small sample sizes as seen in the following simulations. In the chemical example above, this symmetric-in-its-arguments correlation coefficient, r_{sgcc} = 0.9363. Note that this is very close both to the R of the model and to r_{gcc}.

This symmetrized verison should be best for normal and other bivariate models with elliptical contours for the joint density but may be useful in a variety of situations. Further study is needed. A table of comparison is given for some sample sizes for the standardized bivariate normal. The correlations range from 0.5 to 0.9 and zeros have been deleted from the averages.

Correlation coefficients for bivariate normal, 500 simulations							
		Normal parameters					
corr coef	statistic	n=20 $r = 0.5$	n=40 $r = 0.5$	n=40 $r = 0.80$	n=60 $r = 0.80$	n=80 $r = 0.90$	n=10 $r = 0.90$
Pearson	mean	0.4936	0.4973	0.7976	0.7997	0.8974	0.8910
	SD	0.1680	0.1208	0.0591	0.0497	0.0222	0.0826
r_{gcc}	mean	0.5469	0.5043	0.7765	0.7917	0.8915	0.8086
	SD	0.1910	0.1663	0.0935	0.0646	0.0303	0.1444
r_{sgcc}	mean	0.5202	0.4961	0.7808	0.7919	0.8925	0.8130
	SD	0.1785	0.1563	0.0801	0.0591	0.0277	0.1316
$r_{gd} - rob$	mean	0.6751	0.6515	0.8221	0.8419	0.9260	0.8570
	SD	0.1906	0.1691	0.1223	0.0883	0.0347	0.1615
# zeros	r_{gcc}	124	60	1	0	0	7
	r_{sgcc}	79	43	0	0	0	6
	$r_{gd} - rob$	146	80	6	3	0	59

Table 4: Demonstration of the symmetric version of correlation. It is compared to three other correlations for various cases.

The main purpose of this table is to compare r_{gcc} to r_{sgcc}. From the values in the table it seems to be that symmetrizing the correlation has slightly improved its standard deviation. The two versions of the generalized correlation coefficient give very similar results. For the correlation coefficients r_{gcc}, r_{sgcc}, and $r_{gd} - rob$ the number of zeros, i.e. the number of zero correlation estimates for each, is given at the bottom of the table. They occur, as suggested, for smaller sample sizes and/or low correlation. They have been deleted from the calculation of the averages and standard deviations.

6. Conclusion and summary

From the work above it is now clear that the famous horseshoe type model, whose Pearson correlation coefficient is zero, does not have correlation zero, but depends on the variation about the horseshoe curve. It is hoped that this quadratic example which appears in many textbooks showing that the correlation is zero will be deleted and replaced by the following. The correlation is not zero and its value depends on how the x data is generated and the local variation. Let the model be $Y = X^2 + Z$ where Z is normal with mean zero and variance s^2 and interest is in x values between -2 and +2. The manner in which the correlation is computed follows the same ideas given in the cubic model. First if Y is sampled an equal number of times

at the x values of $-2, -1, 0, 1, 2$, then the COD is $1 - \frac{s^2}{(14/5) + s^2}$. Second, if X is Uniform over the interval $(-2, 2)$ then the COD is $1 - \frac{s^2}{(64/45) + s^2}$. The correlation is the square root of COD. If $s = 0.5, 1, 2$, the two correlations become 0.958, 0.858, 0.642, and 0.922, 0.766, 0.512, respectively. The current way this example is used negates much of the value of the correlation especially for a beginning student. COD is not widely known, especially as used in this paper, and indeed, some elementary books still relate the COD to least squares.

Each student should be taught the following simple method to estimate correlation from a scattergram. First distinguish global variation for the variable plotted on the vertical axis from the local variation; that is, fix an x and look at the corresponding y values. Estimate by ruler or any other crude method the range of both the global and local variation. Take the ratio of local to global, square, subtract from one, and take the square root. This is an estimate of the correlation between the variables on the two axes. A large general education freshman class was taught this concept and for the quadratic curve above, they estimated several different cases very well. A second lecture to undergraduate mathematics students confirmed the viability of the technique. An instructor should, of course, tell students to use reason, to not include some extreme points in their estimate, and to look for appropriate x-values for the local variation. In some cases this would be sufficient but in other classes it may be appropriate to continue with the formula of the new correlation in this paper.

The thrust of this work is to introduce correlation and robust correlation to models that are not linear. Three types of comparisons are made. Comparisons for linear models show that the new correlations are valid in this case and the theory provides a way to make the definition plausible. A cubic model brings out the true worth of the new correlations in assessing the relationship between the variables. Also erratic error is introduced using the Cauchy distribution to show the value of the robust version. The final example is an exponential model with real data. It shows that the correlation analysis as developed in this paper gives useful information about the data before choosing a model. It also suggests what to expect before fitting the model. The most important point is to dissociate the idea of correlation from that of linear correlation so that the use of the word correlation actually relates to what someone sees intuitively when looking at nonlinear data plots. If enough data points have been selected to adequately follow the form of the model, then the new correlation coefficients provide good descriptive statistics about the association between variables related nonlinearly. Even beginning students can be taught to estimate this new correlation for any data.

7. References

Draper, N.R. & Smith, H. (1966, 1981, 1998), Applied Regression Analysis, Wiley & Sons, Inc.

Gideon,R.A. & Hollister,R.A. (1987), A rank correlation coefficient resistant to outliers, *Journal of the American Statistical Association*, **82**, 656-666.

Kendall, M.G. & Gibbons, J.D. (1990), Rank Correlation Methods, 5th ed. Oxford University Press, or also Kendall, M.G. (1962), Rank Correlation Methods, 3rd ed. Hafner Publ. Co.

Pearson, K. (1911), On the probability that two independent distributions of frequency are really samples from the same population. *Biometrika*, **8**, 250-254.

Spearman, C. (1904), The proof and measurement of association between two things. *Amer. J. Psychol*, 15, 72-101

Appendix: four S-PLUS or R programs for the robust estimate of general correlation

0. A main program to utilize the programs below.

1. rguniq: Computes GDCC for the unique ranks of y where y has been sorted relative to x.

2. GDave: For any set of bivariate data, this routine computes two values of GDCC and averages them for a unique result. ccp is the value of GDCC computed so that ties are broken to achieve maximum positive correlation. ccn is the value of GDCC computed so that ties are broken to achieve the least positive correlation. Both ccp and ccn call rguniq.

3. GDreg: This routine is iterative and computes the linear regression slope coefficient, b, of y on x. It calls GDave repeatedly. At the conclusion, $y - bx$ and x are GDCC uncorrelated.

```
0. # file for SLR with GD and adding the intercept,
x and y must have the data entered
slope <- GDrg(x,y)
res <- y - slope*x
int  <- median(res)
plot(x,y)
abline(int,slope)
ccar <- GDave(x,res)
out <- c(int,slope,ccar)
out
```

```
1. rguniq <-
function(rky)
{ n <- length(rky); n1 <- n-1
   dy <- NULL; dyn <- NULL
   ryr <- n + 1 - rky
   for(i in 1:n1){
      dy <- c(dy, sum(rky[1:i] - i > 0))
      dyn <- c(dyn, sum(ryr[1:i] - i > 0))}
   mdyr <- max(dyn)
   mdy <- max(dy)

   corrg <- (mdyr - mdy)/(n %/% 2)
   corrg  }


2. GDave<-function(x,y)
{      n <- length(x)
    xt<-x[order(y,x)] #x order by y with y ties ordered by x
        rky<-1:n
        rky<-rky[order(xt,rky)]    # ranks of y ordered by x
        ccp  <- rguniq(rky)   # GD positive
# GD negative below

   xrr <-  n +1 -rank(x)  #reverse ranks on the x
   xt <- x[order(y,xrr)]  #x ordered by y with y ties ordered by rev(x)
   rky <- order(xt,n:1)   #ranks of y ordered by x with y ties
   ccn <- rguniq(rky)            #ordered by rev(y)

         (ccp+ccn)/2  }


3. This function compute the GD-estimate of the simple regression

GDrg <-function(x, y)  {

#Compute B(i,j) for all different i,j, elementary slopes
   n <- length(x)
   z <- numeric()
   for(i in 1:(n - 1)) {
      for(j in (i + 1):n) {
         k <- n * (i - 1) - (i * (i - 1))/2 + (j - i)
         if(x[j]!=x[i])
            z[k] <- (y[j] - y[i])/(x[j] - x[i])
         else z[k] <- NA }  }

   z <- sort(z) #Delete the tie values
   k <- 1
   z1 <- numeric()
   z1[1] <- z[1]
   for(i in 2:length(z)) {
      if(z[i]!=z1[k]) {
         k <- k + 1
         z1[k] <- z[i] }  }

   z <- numeric()
   z <- z1   #Bisection Method to find the right and the left end point .
   ns <- length(z)
   jr <- ns
   jl <- ns
   ir <- 1
   il <- 1
```

```
   while(jl > il) {
      b1 <- (z[il] + z[jl])/2
      res <- y - x * b1
      a <- GDave(x, res)
      if(a <= 0) {
         while(z[jl] > b1) jl <- jl - 1 }

      else while(z[il] < b1) il <- il + 1 }

   while(ir < jr) {
      b1 <- (z[ir] + z[jr])/2
      res <- y - x * b1
      a <- GDave(x, res)
      if(a >= 0) {
         while(z[ir] < b1) ir <- ir + 1 }

      else while(z[jr] > b1) jr <- jr - 1 }
      (z[il] + z[jr])/2
}  # end  of function
```