The Utility and General Definition of Correlation Coefficients

Rudy A. Gideon
University of Montana
Missoula, MT 59812

Abstract

Previous attempts at defining other correlation measures mostly tried to generalize the inner product definition used in Pearson's correlation coefficient. This does not allow for certain useful correlation's, like the Greatest Deviation, or Gini's. In this work the idea in Gideon and Hollister (1987) of seeing correlation, as the difference between distance from perfect negative and perfect positive correlation will be used to bring together a general setting. Pearson, Spearman, and Kendall correlation coefficients are then seen as special cases where a linear restriction holds. It will also be seen how to define a wide variety of correlation coefficients. Simple linear regression with these correlations will be discussed in order to illustrate an introduction to statistical estimation with correlation coefficients. The general focus of this paper is simply to outline notation and concepts necessary for using correlation coefficients as estimating functions.

Keywords: correlation coefficients, Greatest Deviation Correlation, Kendall Correlation, regression, scale estimation, minimization

AMS Subject Classification: 62

Introduction: Kendall in Chapter 2 of "Rank Correlation Methods", tries to give a general theory of rank correlation but it is based on the inner product case and does not include other correlations. Instead we start with a definition of correlation that is much more general. Let us start with two bivariate random variables X and Y whose expectations are zero and variances are one. Then $(X+Y)^2=(X-(-Y))^2$ is to be viewed as squared distance from perfect negative correlation and $(X-Y)^2$ is to be viewed as squared distance from perfect positive correlation. It is easily seen that $E(X+Y)^2 = E(X^2)+2E(XY)+E(Y^2)=1+2r+1=2(1+r)$ where r is the correlation parameter. Similarly $E(X-Y)^2 = 2(1-r)$. Let $Dneg^2=(X+Y)^2$ and $Dpos^2=(X-Y)^2$; i.e. squared distances from negative and positive correlation. A little table will make this clear.

correlation r	$E(Dneg^2)=E(X-(-Y))^2$	$E(Dpos^2)=E(X-Y)^2$
-1	0	4
0	2	2
+1	4	0

In what follows, all other correlations, plus a few that are not widely known and a few new ones will be shown to be interpretable as Dneg – Dpos. Note that for zero correlation the expected difference is zero, 2 –2 =0. A correct rescaling of this difference is to divide by 4 to put the correlation between –1 and +1. Other correlation coefficients will have different scale factors. The correlation coefficient here would be defined as $r = \frac{Dneg^2 - Dpos^2}{4}$.

A Data Interpretation of Dneg - Dpos

Let $\{x_i, y_i\}, i = 1,2,3,\cdots n$ be a bivariate data set or in vector notation, let (x, y) be the data. Then $\|(x-(-y)\|^2$ can be viewed at the squared distance between x and $-y$, and $\|(x-y)\|^2$ can be viewed as the squared distance between x and y. Let $\bar{x}$ and $\bar{y}$ be the sample means and s_x and s_y the sample standard deviations. If X and Y are highly correlated random variables then $Dpos^2 = \|(x-y)\|^2$ is small whereas $Dneg^2 = \|(x-(-y))\|^2$ is large and the opposite is true for highly negative correlated variables. Because correlation is defined for standardized data let now $Dneg^2 = \left\| (\frac{x-\bar{x}}{s_x}) + (\frac{y-\bar{y}}{s_y}) \right\|^2$ which is norm (computer) notation for $\sum (\frac{x_i-\bar{x}}{s_x} + \frac{y_i-\bar{y}}{s_y})^2$, and $Dpos^2 = \left\| (\frac{x-\bar{x}}{s_x}) - (\frac{y-\bar{y}}{s_y}) \right\|^2 = \sum (\frac{x_i-\bar{x}}{s_x} - \frac{y_i-\bar{y}}{s_y})^2$.

Finally $\frac{Dneg^2 - Dpos^2}{4(n-1)} = \frac{1}{n-1}\sum (\frac{x_i-\bar{x}}{s_x})(\frac{y_i-\bar{y}}{s_y})$, the usual Pearson's correlation coefficient, $r(x, y)$. The reduction of this correlation to its classical form is due to the nice properties of Euclidean distance. Note also that $Dneg^2 + Dpos^2 = 4(n-1)$, a linear restriction. It is well-known that Spearman's correlation coefficient is defined by using Pearson's r on the ranks of the data. Thus, the same Dneg – Dpos concept holds for Spearman with $\bar{x} = \frac{n+1}{2}$ and $s_x^2 = s_y^2 = n(n+1)/12$. This is discussed below.

A Population Interpretation for Kendall's Tau

A alternative probabilistic definition of Kendall's Tau will be given. The reason for this change follows later. Let the continuous bivariate random variable (X, Y) with cumulative distribution function $F(x, y)$ have two independent observations, (X_1, Y_1) and (X_2, Y_2). Then the elementary slope random variable is $\frac{Y_2 - Y_1}{X_2 - X_1}$. Let the

concordance probability $P = P(\frac{Y_2 - Y_1}{X_2 - X_1} > 0) = P((Y_2 - Y_1)(X_2 - X_1) > 0)$ =Dneg. It is seen that at $r = +1$, P=1; $r = 0$, P=1/2; $r = -1$, P=0. Let the discordant probability be $Q = P(\frac{Y_2 - Y_1}{X_2 - X_1} < 0) = P((Y_2 - Y_1)(X_2 - X_1) < 0)$ = Dpos. At $r = +1$, Q=0; $r = 0$, Q=1/2; $r = -1$, Q=1. Then the definition of Kendall's tau is

$t = Dneg - Dpos = P - Q = 2P - 1$ because $P + Q = 1$.

Let F be a bivariate normal random variable with means and variances, $m's$ and $s's$ and correlation r. Then from Gideon and Rothan (2004b) it follows with little effort that $median(\frac{Y_2 - Y_1}{X_2 - X_1}) = r\frac{s_y}{s_x}$, the population slope parameter in simple linear regression.

Another way to say this is that $P = P(\frac{Y_2 - Y_1}{X_2 - X_1} > r\frac{s_y}{s_x}) = \frac{1}{2}$. This can also be written as $P\left(\frac{(Y_2 - r\frac{s_y}{s_x}X_2) - (Y_1 - r\frac{s_y}{s_x}X_1)}{X_2 - X_1} > 0\right) = \frac{1}{2}$ or as $t(X, Y - r\frac{s_y}{s_x}X) = 0$. This nice result is the reason for the alternative definition.

Three New Continuous Correlation Coefficients

The results stated here are from Gideon:Web site paper #1. Based on the idea of Dneg – Dpos and the form of Pearson's correlation expressed this way, it is easy to use distance measures involving absolute values and medians to define other correlation coefficients. Let $SA_x = \sum|x_i - \bar{x}|$, $median(x) = m_x$, $SA_{mx} = \sum|x_i - m_x|$, and finally $MAD_x = med|x_i - m_x|$, the usual MAD definition of variation. Similar notation is used for the y variable. Now the three new correlations are defined. First an absolute value one is given,

$$r_{av} = \frac{1}{2}\left(\sum\left|\frac{x_i - \bar{x}}{SA_x} + \frac{y_i - \bar{y}}{SA_y}\right| - \sum\left|\frac{x_i - \bar{x}}{SA_x} - \frac{y_i - \bar{y}}{SA_y}\right|\right)$$

This correlation is the continuous analogue of Gini's rank correlation which will be given below. A second absolute value correlation using medians is as follows,

$$r_{avm} = \frac{1}{2}\left(\sum\left|\frac{x_i - m_x}{SA_{mx}} + \frac{y_i - m_y}{SA_{my}}\right| - \sum\left|\frac{x_i - m_x}{SA_{mx}} - \frac{y_i - m_y}{SA_{my}}\right|\right)$$

This correlation is a slight modification of r_{av} in which means are replaced by medians. This correlation unlike r_{av} is not always between –1 and +1. It nevertheless is a

reasonable correlation coefficient. Finally a median and absolute value correlation which ties in nicely with the MAD variation definition is defined. Substitute y for x and one gets the variation estimate.

$$r_{mad} = \frac{1}{2}\left(med\left| \frac{x_i - med(x_i)}{MAD_x} + \frac{y_i - med(y_i)}{MAD_y} \right| - med\left| \frac{x_i - med(x_i)}{MAD_x} - \frac{y_i - med(y_i)}{MAD_y} \right| \right).$$

None of these correlation have the linear nicety in that Dneg+Dpos is not a constant. For each correlation coefficient the left hand side of the minus sign is Dneg and to the right of the minus sign is Dpos. For example, for r_{mad} if r =+1, Dneg = +1 and Dpos = 0. However, r = -1, Dneg = 0 and Dpos = 1. These correlations are described in more detail on the Gideon WEB site, paper #1.

Nonparametric or Rank Based Correlations

This section gives rank based correlations all defined in the Dneg – Dpos manner, and also on Web site paper #1. Let the (x, y) data be ranked by the x data and then replace the data by ranks. Then the data is in the form (e, p) where $e' = (1,2,3,\cdots n)$ and $p' = (p_1, p_2, p_3, \cdots, p_n)$ are the corresponding ranks of the y-data. Absolutely continuous data is assumed. Also the greatest integer notation is used, $\left[\frac{7}{2}\right] = 3$, for example. In what follows, the correlation coefficients will be written in forms that will show their similarities. Let $q' = (n+1-p_1, n+1-p_2, n+1-p_3, \cdots, n+1-p_n)$. All of them will be functions defined on e, p, and q. Note that $p + q = n+1$. (n+1 here is really a constant vector) Spearman's correlation coefficient is defined from Pearson's replacing data by their ranks in the e, p, q form. From above,

$$r_s = \frac{Dneg - Dpos}{4(n-1)} = \frac{1}{4(n-1)}\left\{ \sum \frac{(i - \frac{n+1}{2} + p_i - \frac{n+1}{2})^2}{s_x^2} - \sum \frac{(i - \frac{n+1}{2} - (p_i - \frac{n+1}{2}))^2}{s_y^2} \right\}$$

where $s_x^2 = s_y^2 = \frac{n(n+1)}{12}$. By a little algebra Spearman's can be written as

$$r_s = \frac{3}{n(n^2-1)}\left\{ \sum (n+1-p_i - i)^2 - \sum (p_i - i)^2 \right\}.$$

Note that this can be written in norm form as $\frac{3}{n(n^2-1)}\left\{ \|q-e\|^2 - \|p-e\|^2 \right\}$. It is easily shown that $\frac{3}{n(n^2-1)}\left\{ \|q-e\|^2 + \|p-e\|^2 \right\} = 1$. This allows r_s to be written in one of its standard forms $1 - \frac{6}{n(n^2-1)} \sum (p_i - i)^2$.

In Kendall's correlation book, there was a Spearman footrule correlation that only involved the absolute values of the $p-e$ vectors. The modification of this to a real correlation coefficient leads to Gini's, but we label it mf for modified footrule. It is obtained by substituting ranks in the definition of r_{av}.

$$r_{mf} = \frac{\sum|n+1-p_i-i| - \sum|p_i-i|}{\left\lceil \frac{n^2}{2} \right\rceil \left\lfloor \frac{n^2}{2} \right\rfloor}$$

In comparing this to Spearman's it is seen that absolute values has replaced the squares. Again if there is perfect positve correlation, Dneg, the left side numerator equals the denominator, and Dpos = 0. If there is perfect negative correlation, Dpos, the right side numerator equals the denominator, and of course, Dneg = 0. Dneg + Dpos is not a constant

It is now time to introduce the Greatest Deviation correlation coefficient. It is defined as follows:

$$r_{gd} = \frac{\max\limits_{1\le i\le n}\sum\limits_{j=1}^{i} I(n+1-p_j > i) - \max\limits_{1\le i\le n}\sum\limits_{j=1}^{i} I(p_j > i)}{\left\lceil \frac{n}{2} \right\rceil \left\lfloor \frac{n}{2} \right\rfloor}$$

In this definition I is the indicator function,

$$I(x) = \begin{cases} 1 & \text{if true} \\ 0 & \text{otherwise} \end{cases}$$

It is seen again that the definition involves, the three vectors e, p, q and that Dneg the left hand side of numerator is maximum when there is perfect positive correlation and equals the denominator. These results are explained in detail in the Gideon and Hollister (1987) paper or on the Gideon WEB site papers, #1,2,3.

For the Kendall correlation statistic, the adjusted slope version that corresponds to the population definition is used.

Dneg = P, the number of concordances =

$$\sum_{i=2}^{n}\sum_{j=1}^{i-1} I\left(\frac{y_i - y_j}{x_i - x_j} > 0\right) = \sum_{i=2}^{n}\sum_{j=1}^{i-1} I\left(\frac{p_i - p_j}{i-j} > 0\right) = \sum_{i=2}^{n}\sum_{j=1}^{i-1} I\left(p_i - p_j > 0\right).$$

Dpos = Q, the number of discordances or the number of concordances on $q = (n+1-p)$ =

$$\sum_{i=2}^{n}\sum_{j=1}^{i-1} I\left(\frac{y_i - y_j}{x_i - x_j} < 0\right) = \sum_{i=2}^{n}\sum_{j=1}^{i-1} I\left(\frac{p_i - p_j}{i - j} < 0\right) = \sum_{i=2}^{n}\sum_{j=1}^{i-1} I\left(p_i - p_j < 0\right) =$$

$$\sum_{i=2}^{n}\sum_{j=1}^{i-1} I\left((n+1-p_i)-(n+1-p_j) > 0\right).$$

Now max(Dneg) = max(Dpos) = $\binom{n}{2}$ and Dneg + Dpos = P+ Q = $\binom{n}{2}$.

Thus, $t = \frac{Dneg - Dpos}{\binom{n}{2}} = \frac{P - Q}{\binom{n}{2}} =$

$$\frac{\sum\sum I(p_i - p_j > 0) - \sum\sum I((n+1-p_i)-(n+1-p_j) > 0)}{\binom{n}{2}}. = \frac{2Dneg}{\binom{n}{2}} - 1.$$

Again Kendall's Tau is based on the difference between Dneg, the distance from perfect negative correlation which is the concordance count, and Dpos, the distance from perfect positive correlation which is the discordant count. It is also defined on the vectors p and q.

Simple Linear Regression and Correlation Coefficients

Correlation coefficients seem to be undervalued. This made be because statisticians believed there was no natural way for them to be involved in the many areas of statistics related to model evaluation through regression. In a long term study of the Greatest Deviation Correlation Coefficient, it does seem that there is a straight-forward way to perform regression of all sorts. This is illustrated in many of the papers appearing on the Web site. Here is given mainly the notation used that makes the use of any correlation in regression a very simple manner. For rank based correlations coefficients the maximum and minimum tie breaking method given in Gideon and Hollister (1987) is essential.

Introduction

A little background notation and concept development are necessary. Both the slope in a simple linear regression and the residual standard deviation need to be presented. This development comes mainly from the papers on the Web site. First for the regression slope, let us take the Bivariate Normal with means zero and variances one in order to concentrate on the bare essentials. For any correlation coefficient $r(X, Y - bX)$ has a Null distribution; i.e., X and $Y - bX$ are independent, and $b = r$ for the standardized case; this was seen earlier for Kendall's Tau. Series expansion on the parameters and substitution for the random variables with data can be used to draw "r" inferences on the slope. Both the exact and asymptotic null distributions can be used.

This is demonstrated on WEB site papers 4 through 8 and the Baseball example with game time as the dependent variable.

For a data set (x, y), the slope estimate comes by solving for b in

$$r(x, y - bx) = 0 . \qquad (1)$$

As an example, with Pearson's r, solving this equation leads to $\hat{b} = \frac{\sum xy}{\sum x^2}$, the classical least squares estimate. Similarly for Kendall's tau, solving this equation leads to $\hat{b} = median\left\{\frac{y_j - y_i}{x_j - x_i}\right\}$. Note that from the population interpretation , $\hat{b}$ must converge to the true slope. With other correlations this equation must be solved iteratively.

Now for the residual standard deviation estimate, we start with a univariate example. Again use normal data to illustrate the concepts. Let $Y \overset{D}{=} N(m, s)$ and $Z \overset{D}{=} N(0,1)$. For the order statistics $Z_{(i)}, i = 1,2,\ldots,n$ let $E(Z_{(i)}) = k_i$, $i = 1,2,\cdots,n$. The vector k will denote the expected values of the standard normal order statistics. Now for a random sample of Y, the order statistics $y_{(i)} = m + sk_i + error$. This linear equation can be used in a simple linear regression to obtain an estimate of s which will be the slope in the regression. Let y^o be the vector of order statistics of the random variable. For any correlation coefficient r, an estimate of the standard deviation is obtained by solving for s in the equation

$$r(k, y^o - sk) = 0 .$$

For the residuals in a regression, let res^o be the ordered uncentered residual vector $\{y_i - \hat{b}x_i\}$. Then solve for s in the equation

$$r(k, res^o - sk) = 0 . \qquad (2)$$

For non-normal data a relative estimate of the residual standard deviation is obtained by solving for s in

$$r(y^o, res^o - sy^o) = 0 . \qquad (2a)$$

Here one obtains how large the standard deviation of the residuals is relative to the variation in the y-data. A more lengthy presentation of this appears in a Power Point presentation on the Web site, paper #8. In the theoretical case with r =0 (also the slope is zero) we have $res^o = y^o$ because $\hat{b} = b = 0$. Recall we are using standardized data. Equation (2a) becomes

$$0 = r(y^o, res^o - sy^o) = r(y^o, y^o - sy^o) = r(y^o, (1-s)y^o)$$

and this implies that s =1; i.e. the residuals and Y have the same standard deviation. In this case the max-min tie breaking method in Gideon and Hollister is essential because for any nonparametric correlation coefficient $r(y^o, 0) = 0$

where here 0 is a vector. For the case r =1, $\hat{b} = b = 1$ and $res^o = y^o - y^o = 0$;thus, equation (2a) becomes $0 = r(y^o, res^o - sy^o) = r(y^o, 0 - sy^o)$ and this implies that s=0 as any other s would give –1 as the value of the correlation. It follows that equation (2a)

plays the role in classical normal theory of $s^2_{y.x} = (1 - r^2)s^2_y$, but here a relative value is assumed with $s = \frac{s_{y.x}}{s_y} = \sqrt{1 - r^2}$. The quantity s is bounded between 0 and 1 from the regression and this is seen from the analogy with the classical case. If $s = 1$, $r = 0$ and if $s = 0$ $r = 1$. This definition easily generalizes to the multiple regression problem as well as equation (1) easily generalizes to the multiple regression case!

We now combine the two uses of any correlation coefficient into one equation in order to estimate both the slope and the standard deviation of the residuals. In the following equation, determine the b that minimizes s;

$$r(y^o, (y - bx)^o - sy^o) = 0. \qquad (3)$$

Note that if b were known then this is the equation for the variation estimate, s. This equation wants b chosen so that the residual variation (the slope of a regression on the x-order statistics) is as small as possible. As small as possible, of course, differs depending on the correlation criteria used. So for any correlation coefficient, equation (3) is an implicit equation in which b is chosen to minimize the solution s, the residual standard deviation of the regression error. Three examples are now given.

Example 1: Pearson's r

Equation (3) becomes $\min_b s = \min_b \frac{\sum y_{(i)}(y_{i_1} - bx_{i_1})_{(i)}}{\sum y^2_{(i)}} = \min_b r(y^o, (y - bx)^o)$;

i.e., determine b so that the correlation between the ordered residuals and y is as small as possible. The classical b comes from solving $r(x, y - bx) = 0$. One would hope that these two estimates are close. In any case to solve for b in equation (3) use the classical estimate of b as the initial estimate in an iterative process.

Example 2: Kendall's Tau

Equation (3) becomes $\min_b s = \min_b median\left\{\frac{(y_{i_1} - bx_{i_1})_{(j)} - (y_{i_2} - bx_{i_2})_{(i)}}{y_{(j)} - y_{(i)}}\right\}$. It can be seen that this minimization tries make the residuals as small as possible relative to the y-data by means of a median measurement;i.e., in the "Kendall Tau" sense. The slope of the residual line relative to the y-data is minimized. A "classical Tau" estimate of the slope would be $\hat{b} = median\left\{\frac{y_{(j)} - y_{(i)}}{x_{(i)} - x_{(i)}}\right\}$. It would be nice if these two ways of finding b were the same. However, the classical can be used as the initial estimate in an iterative process to determine b from equation (3).

Example 3: the Greatest Deviation Correlation Coefficient

In this example there is no partial solution to equation (3). However, let us take a random sample from a bivariate Cauchy distribution , see Gideon and Rothan (2004c). Let initial estimate, $\hat{b}_o$, be from solving (r = GD), $r(x, y - bx) = 0$. Now let $res^o = (y - \hat{b}_o x)^o$ and solve for s in equation (2a), $r(y^o, res^o - sy^o) = 0$. We determine if this s gives a minimum by determining the s values in the neighborhood of $\hat{b}_o$.

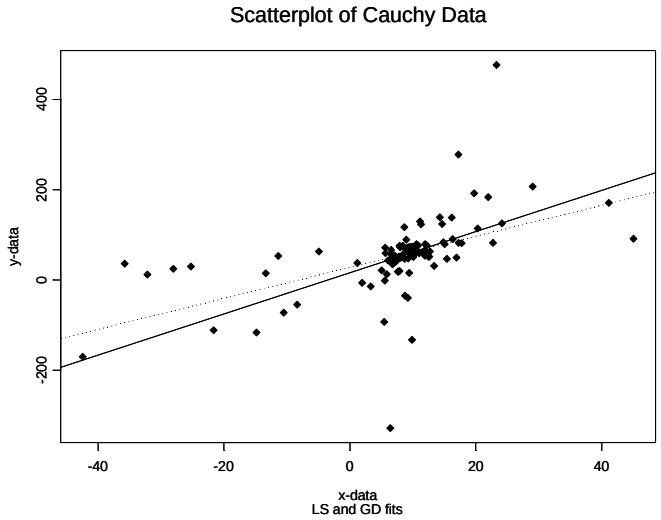
For our example the sample size is 100. The least squares or Pearson's correlation method and the Greatest Deviation, GD, are compared. Summarizing statistics for the simple linear regression are given in the table. Equation (2) is used with standardized Cauchy quantiles to compute the GD error estimate s. Following s in parenthesis is the error estimate using classical methods; i.e. the square root of the squared residuals divided by n-2 = 98. The large difference in the error estimate values means that with the deletion of some outliers the classical estimate can be as low as the first value, with suspected outliers presen, given by the GD method. Recall that there are no outliers with this data, but classical methods probability would show there were some.

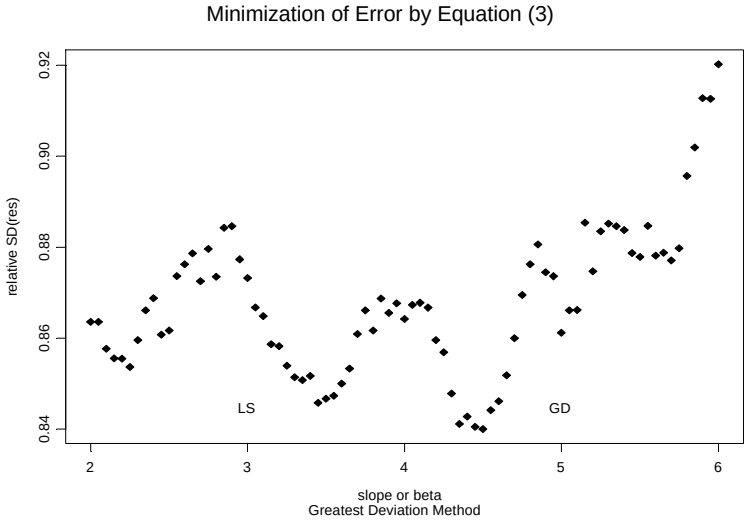
	Intercept	Slope	Standard error
Population	13	5	
LS, Pearson	28.01	3.44	74.36
GD	15.89	4.56	18.00 (75.88)

The sample standard deviation of the y-data is 86.46. Thus, the estimate of $\frac{s_{y.x}}{s_y}$ is $\frac{74.36}{86.46} = 0.8600$. Using nonparametric methods with GD, the same relative value is obtained. First the estimate of standard deviation of the y-data is by equation (2) with Cauchy quantiles is 21.67 (if standard normal quantiles were used this estimate would be 32.96). So from this and the GD residual standard error the ratio is $\frac{18.00}{21.67} = 0.8306$. If this estimate is calculated directly using equation (2a), the value is 0.8471. In the evaluation of the minima in equation (3) for the values of s used, the minima, the vertical scale, will range from 0.84 to 0.88. It is seen that the minimum s does not occur at the GD value from equation (1) given in the table, but the minimum is very close.

Four plots are given; first, the scatterplot and regression lines, second a plot demonstrating equation (3), third one showing the behavior of the correlation coefficient in the vicinity of the solution for the slope in equation (1), and finally a plot with the GD regression line for the estimate of the SD in equation (2). The first is the usual scatterplot of the data with the LS and GD regression lines drawn. GD is the solid line. Even though the data is Cauchy it looks probably not too unlike some actual experimental data. The second plot of (b, s) has the values of s coming from equation (3) for a sequence of slopes plotted along the horizontal axis. In viewing this graph, it is seen that there are local minimums. The plot runs from slopes of 2 to 6 by increments of 0.05. The two local minimums between 3 and 5 appear near the LS and GD estimates of

slopes, 3.44 and 4.56 respectively. A larger plot , 1<slope<6, revealed four local minima with the fourth one lying between 1 and 2, but its value was larger than the other three appearing in the graph below.





For slope 4.56, which solves equation (1), s = 0.846.

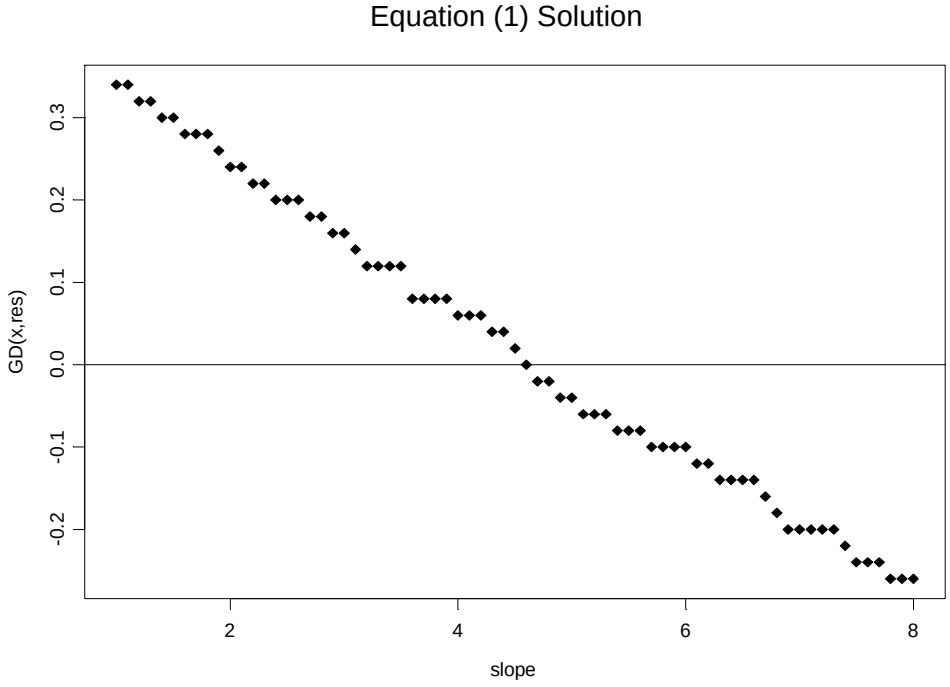
At slope 4.42, the minimum from equation (3), s = 0.839.

At slope 3.44, the second local minimum near the LS estimate, s = 0.847.

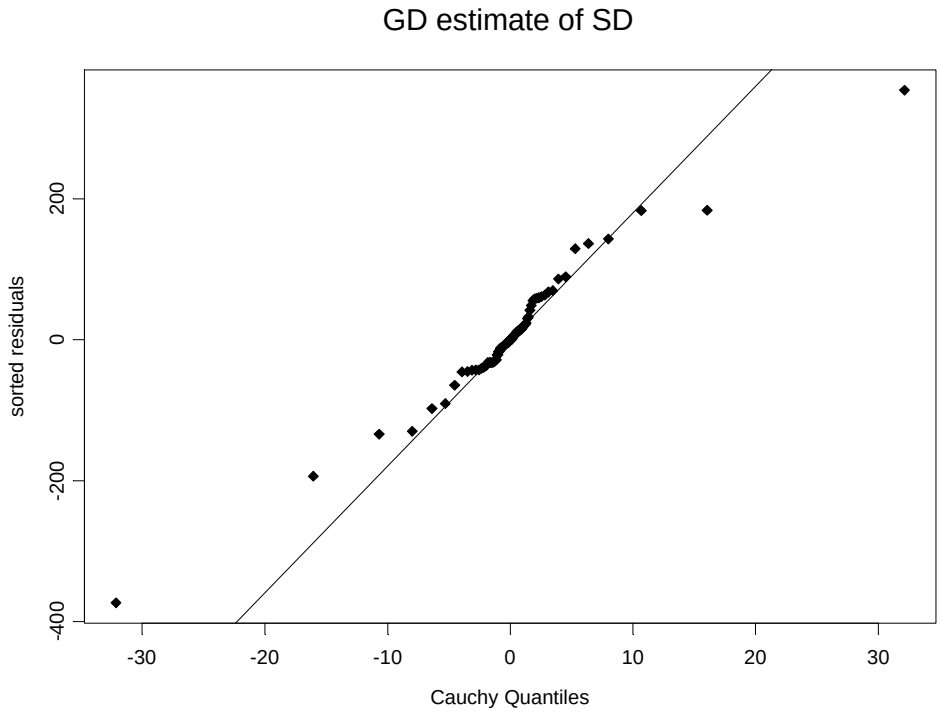
At slope 4.56 with of course, r = GD, $r(x, res(4.56)) = 0$, but at slope 4.42, $r(x, res(4.42)) = -0.04$. The estimate of correlation r using $\frac{s_{y.x}}{s_y} = \sqrt{1 - r^2} = 0.84$

gives an estimate of r of 0.54. For comparison only the GD estimate of r is sin(GD*pi/2) = sin(.5*pi/2) = 0.707. The true value is 0.75.

A plot is given related to equation (1). Slopes in the neighborhood of the GD solution,4.56, are the x axis. For each slope the GD correlation of the x-data with the residuals (y-slope*x) is plotted. This allows a visual look at the behavior of equation (1) in the vicinity of the solution.



The fourth graph shows the slope of 18.00 as the GD estimate of SD when the ordered residuals of the regression are plotted against the standard Cauchy quantiles. The solution to equation equation (2) for s is the 18 and the median of the uncentered residuals is the intercept in the straight line plot in the graph below.


GD estimate of SD
sorted residuals
200
0
-200
-400
-30
-20
-10
0
10
20
30
Cauchy Quantiles

Chambers,J.M., Cleveland,W.S., Kleiner, B., and Tudey,P.A., (1983), Graphical Methods for Data Analysis, Wadsworth and Duxbury Press, Belmont, California and Boston.

Gideon,R.A. and Hollister,R.A. (1987),"A Rank Correlation Coefficient Resistant to Outliers, *Journal of the American Statistical Association*, 82, 656-666.

Gideon, R.A. and Rothan, A.M. (2004a), "Location and Scale Estimation with Correlation Coefficients," unpublished paper (URL: http://www.math.umt.edu/gideon/locscale.pdf), Web site paper #6, University of Montana, Dept. of Mathematical Sciences.

Gideon, R.A. and Rothan, A.M. (2004c), "Cauchy Regression and Confidence Intervals for the Slope," unpublished paper (URL: http://www.math.umt.edu/gideon/cauchyreg.pdf), Web site paper #11, University of Montana, Dept. of Mathematical Sciences.

Kendall,M.G. (1962), Rank Correlation Methods 3rd ed, Hafner Publ. Co., New York

Gideon, R.A. and Rothan, A.M. (2004b), "Elementary Slopes in Simple Linear Regression ," unpublished paper (URL: http://www.math.umt.edu/gideon/Elemslopes.pdf), Web site paper #10, University of Montana, Dept. of Mathematical Sciences.

Gideon,R.A. "A Generalized Interpretation of Pearson's r." (URL: http://www.math.umt.edu/gideon/GeneralizedCC.pdf), Web site paper #1

Gideon,R.A."The Correlation Coefficients." (URL: http://www.math.umt.edu/gideon/Corrcoef.pdf), Web site paper # 2.

Gideon,R.A. "The Geometrical Definition of GDCC and its Uniqueness." (URL: http://www.math.umt.edu/gideon/TheGeometricalDefinitionGDCC.pdf), Web site paper # 3.

Gideon,R.A. "The Correlation Principle." (URL: http://www.math.umt.edu/gideon/corrprinciple.htm), Web site paper # 8, A Power Point Presentation, use only Internet Explorer.

Sheng, HuaiQing (2002) "Estimation in Generalized Linear Models and Time Series Models with Nonparamatric Correlation Coefficients", (URL: http://wwwlib.umi.com/dissertations/fullcit/3041406), Web site Paper #12.

After Thoughts

This brief paper is meant to outline an introduction to a general method of defining correlation coefficients and to illustrate a straightforward manner to use them in simple linear regression. Equation (3) is an alternative to the standard theory of least squares. It can be used to minimize residual error for the correct choice of the slope statistic. This work was prompted by an extensive study of the use of the Greatest Deviation correlation coefficient as a statistical estimator in a wide variety of statistical settings. The Gideon WEB site contains enough of this work to demonstrate its feasibility. Every correlation coefficient can be used as a general statistical estimator in the same manner of GD. GD is a very robust estimator. Many of the existing least squares computer routines rely on normal equations that can be rewritten in terms of Pearson's correlation coefficient. Once this is done, any correlation coefficient can be substituted and new "normal equations" solved. For GD, the Gauss-Seidel method always seems to converge when the max-min tied breaking method is used so that all data can be accepted. This "Correlation" technique appears to equalize the "correlation methods" with the usual least squares method; i.e. wherever least squares goes so can any correlation coefficient. The permutation or randomization tests and bootstrapping techniques also make the correlation methods practical and easily comparing to classical least squares methods and exceeding them in any analysis needing a robust method. It is easier to use the GD estimation method, as in almost all cases, the robustness of the method does not require deletion of suspected outliers in order to have good results. In fact, classical methods usually give close to the GD method after deletion of outliers.

The statistical methods of this paper can be thought of as a merger of three areas, classical, classical nonparametric, and graphical methods. For example, the Chambers et al. book on Graphical Methods. There are many books on classical nonparamatrics by such authors as M. Hollander, D.A. Wolfe, R.H. Randles T.P. Hettmansperger, E.L. Lehmann, G.E. Noether, W.J. Conover, and W.W. Daniel. The references on the WEB site lists many of these books.

This correlation system of estimation is a system like least squares is an estimation system. The Web site shows some of its vast potential. Because of the fact that any correlation coefficient can be used, this practical research area has essentially only been examined with GD. The fact that GD has been examined in many areas of estimation, all with excellent results, demonstrates its potential. The Ph.D. Dissertation of Sheng, which is assessable on the WEB, demonstrates this method in advanced areas such as Time Series, nonlinear estimation, and general linear models.

The statistics r_{mad} and r_{av} should be paired with existing uses of MAD variation and L_1 techniques, respectively. Existing textbook descriptions of least squares techniques could be broaden to include correlation methods.