CORRELATION IN SIMPLE LINEAR REGRESSION

Rudy A. Gideon
The University of Montana
Missoula, MT 59812

Many people who do data analysis take only a few classes in statistics and hence, in general, get introduced only to classical methods of statistics; i.e., least squares, normal theory, and possibly maximum likelihood methods. The motivation for most of these techniques is the maximizing of a function of the data with respect to some parameter—mean, variance, slope, e.g. Calculus is used by taking derivatives and solving an equation set to equal zero.

So many analyzers of data do not know about robust, nonparametric, or alternative statistical methods that are probably better suited to avoiding misinterpreting one's data. Many of these latter methods cannot be developed by Calculus! A possible way to avoid this dilemma is to offer a method of instruction that allows both classical and other estimation techniques to be developed simultaneously. Correlation coefficients offer a very general method of estimating parameters and hypothesis testing. The motivation for their use is based on n-dimensional geometry and the generalization of the concept of the parallelogram law and perpendicularity in a Hilbert Space. Classical statistical methods are represented by Pearson's Correlation Coefficient and the Cosine function. Other methods, such as median methods, are represented by Kendall's tau or by one or more absolute value correlation coefficients, still more techniques, equal area or volume, by the Greatest Deviation Correlation Coefficient . The regression approach is first shown using Pearson's and Kendall's correlation coefficients. Then n-dimensional geometry and orthogonality are used for motivation, a third correlation is introduced, the Greatest Deviation, and finally, some simple linear regression examples illustrate these ideas.

After the development of simple linear regression, these techniques can be broadened for location and scale estimation.

1. Introduction, Least Squares, Pearson's r, and Kendall's Tau

The stage for simple linear regression will be set by reviewing the relationship between least squares and Pearson's Correlation Coefficient, r_p. Let continuous bivariate data be defined as x-y vectors,

$$\{x_i, y_i\}_{i=1}^{n} = (x, y) = \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \bullet & \bullet \\ \bullet & \bullet \\ x_n & y_n \end{pmatrix}$$

Let the model be the usual $y = a + bx + e$ where a is the intercept, b is the slope, and the random error e has expectation zero. Thus, $E(Y|x) = a + bx$. Let X and Y-ßX be random variables. Since $Y - bX = a + e$ and the error random variable is assumed to be independent of random variable X, for any correlation coefficient, r, and a population model, $r_p(X, Y - bX)$ has a Null distribution with expectation zero. Because correlation coefficients are location invariant, the intercept parameter, a , is not involved in the estimation of the slope ß, and $\hat{b}$, the estimate, is a slope that makes the residuals $y - \hat{b}x$ uncorrelated with x. The estimation of ß is obtained by solving the sample equivalent of the expectation being zero,

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

The first example shows, as widely known, that Pearson's r_p is gives the same result as least squares in simple linear regresssion.

Example 1: Pearson's r_p

For this case let $s_x^2 = \dfrac{\sum(x_i - \bar{x})^2}{n-1}, s_y^2 = \dfrac{\sum(y_i - \bar{y})^2}{n-1}$, and $s_{xy} = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$, the sample covariance. Then $r_p(x, y) = \dfrac{s_{xy}}{s_x s_y}$.

To solve equation (1) using r_p we obtain

$r_p(x, y - bx) = \dfrac{s_{x,y-bx}}{s_x s_{y-bx}} = 0$ or $s_{x,y-bx} = s_{xy} - bs_x^2 = 0$. The final result is b = $\hat{b} = \dfrac{s_{xy}}{s_x^2} = r_p(x, y)\dfrac{s_y}{s_x}$. This is, of course, also known as the least squares solution.

The intercept estimate comes also from a population model. We want $E(Y - bX - a) = 0$ so we make the sum of the residuals zero. This leads to

$0 = \sum e_i = \sum(y_i - \hat{b}x_i - a) = \sum y_i - \hat{b}\sum x_i - na$. The solution for a is $\hat{a} = \bar{y} - \hat{b}\bar{x}$.

There is a third method to motivate the least squares or Pearson method and it uses the idea of minimizing the distance from perfect negative correlation plus the distance from perfect positive correlation. This is important because this method alone generalizes to NPCC, not the minimization of residuals. Let

$f_n(b) = \sum_{i=1}^{n}(x_i + y_i - bx_i)^2$ (dpnc) and $f_p(b) = \sum_{i=1}^{n}(x_i - (y_i - bx_i))^2$ (dppc). Then

$\min_b \left(f_n(b) + f_p(b)\right)$ is easily shown to be equivalent to minimizing the usual residual sum of squares. It will be shown that setting Kendall's CC equal to zero and solving for β is equivalent to an analogous minimization.

Example 2: Kendall's tau, or r_K

To solve equation (1) for Kendall's tau, we must first review how to calculate r_K assuming no tied values. Let the x data be ordered $x_1 < x_2 < \cdots < x_n$, and relabel the y_j that corresponds to x_1 as y_1, etc. Then the data can be listed as it would be graphed from left to right

$$\begin{matrix} x_1 < x_2 < \cdots < x_n \\ y_1, \quad y_2, \qquad y_n \end{matrix}.$$

For any data pair $(x_i, y_i), (x_j, y_j)$ the slope of the line between them is $l_{ji} = \dfrac{y_j - y_i}{x_j - x_i}$ and the pair is said to be concordant if the slope is positive but discordant if negative. Note that if j>i, $x_j - x_i > 0$ and the concordance of the i j pair depends solely on $sign(y_j - y_i)$. For the $\binom{n}{2}$ pairs of data points, let

$$C = \#concordants = \sum_{i=1}^{n-1}\sum_{j>i} (sign(y_j - y_i) + 1)/2$$

$$D = \#discordants = -\sum_{i=1}^{n-1}\sum_{j>i} (sign(y_j - y_i) - 1)/2$$

By assumption there no ties , so C+D= $\binom{n}{2}$. Kendall's tau, r_K, is defined to be

$$\frac{C - D}{\binom{n}{2}} = 1 - \frac{2D}{\binom{n}{2}} = \frac{2C}{\binom{n}{2}} - 1.$$

We now solve equation (1) with r_K. The $\binom{n}{2}$ slopes l_{ji} are sometimes called elementary estimates of the slope ß. Let $ES = \{l_{ji}, j > i\}$ be this set of elementary slopes. To solve

$$r_k(x, y - bx) = 0$$

we need C=D. Now recall

$$\begin{Bmatrix} x_1 & < & x_2 & < \cdots < & x_n \\ y_1 - bx_1, & & y_2 - bx_2, & \cdots, & y_n - bx_n \end{Bmatrix}.$$

Note that for ß very negative or near $-\infty$ all the (i,j) pairs are concordant and $r_k(x, y - bx) = +1$. On the otherhand, if ß is near $+\infty$ all pairs are discordant and $r_k(x, y - bx) = -1$. Thus, as b increases continously from near $-\infty$,the (i,j) pair changes from C to D at $y_j - bx_j = y_i - bx_i$, or $l_{ji} = \dfrac{y_j - y_i}{x_j - x_i} = b$.

It follows that if b increases to median(ES), then C=D and equation (1) is satisfied. The solution to equation (1) then for Kendall's tau is b = $\hat{b} = median(l_{ji})$ because $r_k(x, y - \hat{b}x) = 0$.

For the intercept estimate, we choose median$(e_i) = 0$, where $e_i = y_i - \hat{b}x_i - a$ $, i = 1, 2, \cdots, n$. This implies that $\hat{a} = median(y_i - \hat{b}x_i)$.

The motivation for this location estimate comes from ideas contained in scale and location papers.

Just like least squares, Kendall's method minimizes the square of distances from C = dpnc plus D = dppc. C and D are the concordances and discordants between vectors x and y-βx. At β near minus infinity $C = \binom{n}{2}$ and D=0. As β increases, C increases by one at each elementary slope and D decreases by one. By a simple example it is easy to see that $C^2 + D^2$ is minimized when C=D or when $r_k(x, y - bx) = 0$.

In both of these examples there is an explicit solution to the regression equation (1). Before giving a third example, in which no explicit solution to equation (1) is known to exist, we will look at the n-dimensional view of this regression.

2. A General n-Dimensional Correlation Interpretation of Regression

In Figure 1, x and y, the data vectors are represented as vectors in Euclidean n-space. Pearson's r_p is the cosine of the angle "a" between vectors x and y or as shown in Gideon(1998) it is $\cos^2\frac{a}{2} - \sin^2\frac{a}{2} = \frac{\|x+y\|^2}{4} - \frac{\|x-y\|^2}{4}$. In the Figure, the angle a/2 is the angle between vectors x and $x + y$. To find the estimate of the slope, b, the usual interpretation is to project vector y onto x and this occurs at bx on x. This is equivalent to determining "b" so that length(x+y-bx) = length(x-(y-bx)); the corresponding vectors are shown in the Figure. Correlation is a function of standarized data and so without a notation change we ask the reader to think of all n-dimensional vectors to be standarized(centered at zero and length 1). With Pearson's r this is the usual "normalization". We interpret length(x+y) as the distance from perfect negative correlation (dpnc) with a maximum of 2 when y=x. Likewise, length(x-y) is distance form perfect positive correlation (dppc) with a maximum length of 2. This idea is elaborated in Gideon (1998). We now connect this to the so called Parallelogram Law in a Hilpert Space; with $\|x\|^2 = \|y\|^2 = 1$, the Law is $\|x+y\|^2 + \|x-y\|^2 = 2$. When y=x,

$\|x+y\|^2 = 2$ and $\|x-y\|^2 = 0$, so that dpnc= a maximum. Likewise when y=-x, $\|x-y\|^2 = 2$ and $\|x+y\|^2 = 0$, so that dppc = a maximum. Also in a Hilpert Space, x is orthogonal to y if $\|x+y\|^2 - \|x-y\|^2 = 0$. If x is not orthogonal to y, then determining "b" so that $\|x+y-bx\|^2 - \|x-(y-bx)\|^2 = 0$ is the regression equation in n-space. In general, any CC can be interpreted as “distance” from perfect negative correlation,dpnc, with “length”(x+y) minus “distance” from perfect positive correlation, dppc, with “length”(x-y). The regression equation is

$$corr(x, y - bx) = "length"(x + y - bx) - "length"(x - (y - bx)) = 0 \, . \qquad (1^*)$$

For Kendall’s Tau $"length"(x + y - bx) = C$ and $"length"(x - (y - bx)) = D$.

For Pearson’s CC, the location estimate which is the mean of the uncentered residuals, $\hat{a} = \bar{y} - \hat{b}\bar{x}$, is the n-dimensional projection of $y - \hat{b}x$ to the constant vector, 1. The cosine of the angle between $\hat{a}1$ and $y - \hat{b}x$ is $n\hat{a}^2 \big/ \sum (y_i - \hat{b}x_i)^2$.

3. The Greatest Deviation Correlation Coefficient, r_{gd}, GDCC

This correlation is introduced because it is a robust estimator that can be used to illustrate some general ideas and it does not have an explicit solution ot the regression equation (1) or (1*).

The calculation of r_{gd} can be defined by transforming to ranked data with the x-data ordered from smallest to largest as above.

$$\left.\begin{array}{l} x_1, y_1 \\ x_2, y_2 \\ \vdots \\ x_i, y_i \\ \vdots \\ x_n, y_n \end{array}\right\} \rightarrow \begin{array}{l} 1, q_1 \\ 2, q_2 \\ \vdots \\ i, q_i \\ \vdots \\ n, q_n \end{array}$$

where $q_i = rank(y_i)$ within the y’s which is paired with the i$^{\text{th}}$ smallest x value, i=1,2,...,n. We now define dppc and dpnc. At x_i or i, let

$$d_i^+ = \#\{q_j; i < q_j, 1 \le j \le i\}, \; i = 1, 2, \cdots, n \, .$$

$$\max_{1 \le i \le n} d_i^+ = \text{dppc}$$

$$d_i^- = \#\{q_j; i < n + 1 - q_j, 1 \le j \le i\}, \; i = 1, 2, \cdots, n \, .$$

$$\max_{1 \le i \le n} d_i^- = \text{dpnc.}$$

It can be shown that both dppc and dpnc have maximum values of $\left[\frac{n}{2}\right]$, this is greatest integer notation. When dppc is max, dpnc is zero and vice-versa. So GDCC is

$$r_{gd}(x,y) = \frac{dpnc - dppc}{\left[\frac{n}{2}\right]}\ .$$

GDCC is a nonparametric rank CC and like Kendall's Tau, the value of $r_{gd}(x, y-bx)$ changes only at the elementary slopes, l_{ji}. However, r_{gd} may not change at each l_{ji} because it has fewer values and is an area counting CC. Let

$r_{gd}(x, y-bx)$ = $f(b)$. Just like Kendall's Tau, $f(b)$ is +1 when b is very negative and is -1 when b is very positive. This function $f(b)$ is pseudo monotonic decreasing(PMD) which means it is a step function and as b increases, $f(b)$ will decrease at n or n+1 of the l_{ji}. The PMD property allows the regression equation to be efficiently numerically solved; $r_{gd}(x, y-bx) = 0$.

4. The First Example

This first example was chosen to show the similarities of each of the three correlations in their method of estimation of the slope when the data has no tied values. The data comes from the 1989 Major League Baseball season. The predictor variable is the team's pitching earned run average (ERA) amd the response variable is the team's winning percentage. In 1989 there were 26 major league teams; so the sample size is 26.

A plot of the data with three regression lines is given at the end of the paper. ERA is the horizontal axis and winning percentage is the vertical axis. The Pearson, Kendall, and GD regression lines are drawn. Following the graph is a page giving the data and the intercepts and slopes used in the graphs.

Two figures are included which show the graph (r = one of our correlation coefficients) of b -vs- r(x,y-bx) for a set of slopes, b. The Figure 2 gives a wider range of b to indicate the common behavior of each of the three correlations, Figure 3 gives a narrower range of b in the vicinity of the solution to the regression equations. For each graph there are two points of interest, r(x,y), b=0, and the b for which r(x,y-bx)=0. These points are summarized in the table:

	Correlation	Coefficient	
	GDCC	Kendall	Pearson
slope	-0.0835	-0.1024	-0.1145
correlation	-0.3846	-0.4092	-0.6833

From the graphs, it is now shown how to obtain confidence intervals. For each correlation coefficient (CC), if b indicates the slope parameter, then for vector random variables and each season, r(X,Y-b X) has a Null distribution centered at zero, with n=26 (the number of teams has increased since 1989). A confidence interval can be obtained in a similar fashion for each of the CC's. Let $r_{a/2}$ be the upper $a/2$ quantile for a CC with the Null distribution.

Then

$$P(-r_{a/2} \le r(X,Y) \le r_{a/2}) = 1 - a\ ,$$

for Pearson’s CC with the normality assumption, but this can only be approximately obtained for the nonparametric CC’s GD and Kendall because of the discrete nature of the distributions. The confidence interval is obtained by projecting from the correlation axis at the points $\pm r_{a/2}$ on the graphs to the slope axis; ie, determine

b_l such that $\quad r(x, y - b_l x) = r_{a/2}$

b_u such that $\quad r(x, y - b_u x) = -r_{a/2}$.

The asymptotic distributions were used to approximate these points; $r_{a/2}$ for $1 - a = .80$.

For large n, it is approximatly true that

$\sqrt{n}GD$ is N(0,1), $\binom{n}{2} r_k$ is $N(0, \frac{(n-1)n(2n+5)}{18})$, and $\sqrt{n-3}\, r_p$ is $N(0,1)$.

We give a table that summarizes the results and the reader should relate them to the graphs.

	Correlation	Coefficient	
	GD	Kendall	Pearson
b_l	-0.153	-0.142	-0.148
b_u	-0.047	-0.071	-0.081

Both this example and the next one were chosen to illustrate data for which each data point should have equal importance. The assumption of bivariate normality to make Pearson’s CC analysis valid is a very questional act. Many studies in Sociology, Medicine, and Psychology are examples in which all data points are of equal importance and the normality assumption to make classical analysis valid should only be done after appropriate investigation. However, many times this cannot be done because like baseball, the experiments are not really repeatable. Thus, if interest is in how Pitching ERA relates to winning percentage, the two nonparametric measure are probably most appropriate.

5. A Second Example

This example shows that nonparametric counting techniques are not only valid but easy to apply when there are many tied values. The data is from the 1992 baseball season and concerns the number of runs (y) and the number of hits (x) for each of the 175 games that the Atlanta Braves played. The data is split into two parts, hits and runs for the Braves is one set and likewise for their opponent in each game. There are many games with the same number of hits and runs and so many tied values. To give some idea of the data, the following table gives basic statistics:

			hits		n=175		runs			
per game	mean	SD	Q_1	med	Q_3	mean	SD	Q_1	med	Q_3
Bravs	8.52	3.25	6	8	10	4.21	2.86	2	4	5
Oppn	8.14	3.29	6	7	10	3.54	2.90	1	3	5

A second table summarizes the correlation between x=hits and y=runs

	Correlation Table	
	Braves	Opponent
Pearson	0.7770	0.7447
Kendall	0.4368	0.4942
GD	0.5536	0.5245

It should not be inferred from the correlation table that Pearson's correlation shows more of a relationship between x and y because it is substantially larger; each CC is estimating a different feature of the x-y relationship and in general, Kendall and GD are always less than Pearson. Their P-value could be more or less that the P-value for Pearson, depending upon what assumption one wants to make on the distribution.

The question of this example is "what is the average number of hits to produce a run?" a simple method would be to compute from the table above :

Braves, 8.52/4.21 = 2.02 Opponents, 8.14/3.54 = 2.30,
or alternatively 8/4 = 2.00 7/3 = 2.33.
A more refined method would be to compute the slope in a simple linear regression of y on x. There are many tied values and max-min procedure of Gideon and Hollister (1984) is used. For any nonparametric CC, for each set of (x,y-bx) vectors, the ranks are computed within the restriction of the tied values to produce the maximum and minimum correlation of x and y-bx. These are averaged to produce a unique CC. The graphs of b and y-bx then maintain their PMD feature necessary for slope determination. This was done for GD and Kendall.

It is this tied value procedure that allows the CC estimation techniques to be used in many other areas of statistics, e.g. location and scale estimation. The slopes and their reciprocals are now listed in a table:

slope= runs per	hit	reciprocal =	hits to produce	one run
CC	Braves	recip	Opponents	recip
Pearson	0.6828	1.46	0.6553	1.53
Kendall	0.6250	1.60	0.6666	1.50
GD	0.5000	2.00	0.6125	1.63

Note that GD indicates more hits to produce a run than Pearson or Kendall and each case (Braves, Opponents) is closer to the simple technique above. Thus, the graph of (b,GD(x,y-bx)) lies below that of Kendall and Pearson to the left of the slope estimate. Sports examples are great examples because arguments about which method is best depicts reality can be heated but in reality are only important to the entertainment industry. Medicine, Pschology, and Sociology problems, however, are usually important for Mankind and so should be investigated with at least two methods in today's computer world in order to avoid coming to an inappropriate conclusion.

6. Historical Perspective

Sen's 1968 paper, "Estimates of the Regression Coefficient Based on Kendall's Tau" gives further history and shows the unbiasedness of the slope estimate for Tau. Because Tau is a discrete CC, the regression equation (1) or (1*) can have an interval for a solution. Sen gives a mathematical definition for using the midpoint of this interval for the solution. Further, he shows invariance under linear transformations, gives asymptotic properties of the estimate, and develops the confidence interval using the slopes $\{l_{ji}\}$. His method throws out tied value in the x's and so a numerical routine using his method may be slightly different from the general development of this paper. He does not formulate the regression equation directly as was done in this paper and does not present the graphical method of confidence intervals.

Noether's 1990 book, Introduction to Statistics, the Nonparametric Way, develops the elementary slope method of estimating the slope and shows how to do it with "minitab" He gives basic ideas behind the methods for beginning students.

Book's like Hettmansperger 1984 Statistical Inference Based on Ranks develop regression using the linearity property of many CC's, but by passes the direct use of correlation. Sometimes the x is left alone while the y-bx is transformed to ranks. Also "score" functions are used on either the x or y part of the data. Some of these score functions are taken to be a set of equally spaced , centered at zero, numbers and others use "normal scores", i. e., the cumulative distribution function of the standard normal. These methods are not general enough to include GD as a estimator of the slope in a simple linear regression because it is not a simple linear function as are the Pearson and Kendall CC's. Hettmansperger's book in Section 1.5 does illustrate the graphical idea of the confidence interval as is done in this paper but he only does it for the one-sample location problem.

The Randles and Wolfe book, Introduction to the Theory of Nonparametric Statistics also relies on linearity and makes regression seem very difficult. Rousseeuw and Leroy, Robust Regression & Outlier Detection, relies on the least median of squares technique which has a very high breakdown point. The recent (1997) book by Sheskin covers correlations such as Pearson, Spearman, and Kendall, and various used of them, but this appears at the end of the book and does not relate correlation to regression and other statistical methods.

7. Correlation as Estimating Functions

In Gideon (1998), there are other CC's which can be used to estimate the slope in a simple linear regression. The estimating technique would be as illustrated in this paper. Which CC is best for which sampling situation is an open question. For small sample sizes and assumptions on the X,Y distributions, the Null Distribution of any CC can be approximated by simulation; then the confidence interval for the slope could be obtained. The asymptotic Null Distribution has not been worked out for all of these CC's, so further work would have to been done for large samples.

In Gideon (1998), the use of CC's in elementary time series was indicated. Based on the work in this paper and the other 1998 paper, the estimation of location and scale parameters is the next step. In general, CC's can be used in many estimation situations and should be considered as a class of estimating functions rather than, as it currently is, just summary statistics.

It will be shown in the paper on scale and location estimation, the third paper in this correlation series, that for nonparametric correlation coefficients that the location estimate is more accurate when done after scale estimation. Thus, after that paper, a better estimate of the intercept of the regression can be obtained by first computing the residual estimate of the scale factor with nonparametric correlation coefficient. This current paper is the second in the series on estimation with correlation coefficients.

REFERENCES

Gideon, R.A. and Hollister, R.A. (1987), "A Rank Correlation Coefficient Resistant to Outliers," Journal of the American Statistical Association 82,no.398, 656-666.

Gideon, R. A., (1998), "A Generalized Interpretation of Pearson's r."

Hettmansperger, T.P. (1984), *Statistical Inference Based on Ranks*, John Wiley & Sons.

Kendall, M.G. and 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.

Noether, G.E. (1991), *Introduction to Statistics, The Nonparametric Way,* Springer-Verlag New York, Inc.

Randles, R.H. and Wolfe, D.A. (1979), *Introduction to the Theory of Nonparametric Statistics*, John Wiley & Sons.

Rousseeuw,P.J. and Leroy, A.M. (1987), *Robust Regression & Outlier Detection,* John Wiley & Sons.

Sen, P.K. (1968), "Estimates of the Regression Coefficient Based on Kendall's Tau," Journal of the American Statistical Association, 63, 1379-1389.

Sheskin, D.J. (1997), Handbook of Parametric and Nonparametric Statistical Procedures, CRC Press, New York.