2010

Using Correlation Coefficients to Estimate Slopes in Multiple Linear Regression

Rudy Gideon
University of Montana, Missoula

Sankhyā : The Indian Journal of Statistics
2010, Volume 72-B, Part 1, pp. 96-106

Using Correlation Coefficients to Estimate Slopes in Multiple Linear Regression

Rudy A. Gideon
University of Montana, Missoula, USA

Abstract

This short note takes correlation coefficients as the starting point to obtain inferential results in linear regression. Under certain conditions, the population correlation coefficient and the sampling correlation coefficient can be related via a Taylor series expansion to allow inference on the coefficients in simple and multiple regression. This general method includes nonparametric correlation coefficients and so gives a universal way to develop regression methods. This work is part of a correlation estimation system that uses correlation coefficients to perform estimation in many settings, for example, time series, nonlinear and generalized linear models, and individual distributions.

AMS (2000) *subject classification.* Primary 62J05, 62G05, 62G08.
Keywords and phrases. Correlation, rank statistics, linear regression, nonparametric, Kendall, greatest deviation correlation coefficient

1 Introduction

In this work, a linear multivariate model is assumed for random variables, (X, Y), where $X' = (X_1, X_2, \ldots, X_p)$ and Y is the dependent variable. The notation for the covariance matrix is

$$\Sigma_{p+1,p+1} = \begin{pmatrix} \sigma_1^2 & \sigma_{12}^{'} \\ \sigma_{12} & \Sigma_{22} \end{pmatrix}$$

where σ_1^2 is the variance of the dependent variable, σ_{12} is the column vector of covariances of the dependent variable with the independent variables, and Σ_{22} is the p by p covariance matrix of the pairs of independent variables. The conditional distribution assumption is that $E(Y|X = x) = \mu + (x - \mu_x)'\beta_0$ where $\beta_0 = \Sigma_{22}^{-1}\sigma_{12}$ is the vector of population regression parameters.

Let $\rho_{X_i,Y}$ be the correlation coefficient between X_i and Y. If σ_{ii} is the i^{th} diagonal element of Σ_{22}, that is, the variance of X_i and $\sigma_{12}^{'} = (\sigma_{X_1,Y}, \sigma_{X_2,Y}, \cdots, \sigma_{X_p,Y})$ then

$$\rho_{X_i,Y} = \frac{\sigma_{X_i,Y}}{\sqrt{\sigma_{ii}\sigma_1^2}}, \quad i = 1, 2, \ldots p.$$

For the bivariate model, with $\sigma_Y = \sigma_1$, this simplifies to

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \quad \text{and} \quad \beta_0 = \rho_{X,Y}\frac{\sigma_Y}{\sigma_X}.$$

If $l' = (l_1, l_2, \ldots, l_p)$is a vector of constants then $X'l$ and $Y - X'\beta = Y - \sum_{i=1}^{p} \beta_i X_i$ are both univariate random variables. The correlation coefficient between $X'l$ and $Y - X'\beta$ is a function of β, which, for emphasis, is here usually written as $f(\beta)$ instead of $\rho_{X'l,Y-X'\beta}$. This observation is the foundation of the ensuing work, in which an expression for this correlation coefficient is found and then expanded into a truncated Taylor series, thus connecting the correlation coefficient with the regression coefficients. After this, sample counterparts are substituted, further approximating this connecting relationship. Next the asymptotic distribution of the sample correlation coefficient is used to approximate the asymptotic distribution of the regression coefficients. This method is important because it allows inference on linear sums of the regression coefficients for any correlation coefficient, even those that themselves are not linear, such as the Greatest Deviation Correlation Coefficient (GDCC) or the Median Absolute Deviation (MAD). In fact, all rank based correlation coefficients, as well as continuous ones, are amenable to this method. An R program to calculate GDCC is given on the Website and consult Gideon (2007); MAD is also defined there.

Straightforward methods using the expectation of random variables show that

$$f(\beta) = \rho_{X'l,Y-X'\beta} = \frac{l'\sigma_{12} - l'\Sigma_{22}\beta}{\sqrt{l'\Sigma_{22}l}\,\sqrt{\sigma_1^2 + \beta'\Sigma_{22}\beta - 2\beta'\sigma_{12}}},$$

and that

$$V(Y - X'\beta_0) = \sigma_1^2 + \beta_0'\Sigma_{22}\beta_0 - 2\beta_0'\sigma_{12} = \sigma_1^2 - \sigma_{12}'\Sigma_{22}^{-1}\sigma_{12} = \sigma_{res}^2,$$

the variance of the conditional distribution. In this latter equation, *res* stands for residuals. Also note that $f(\beta_o)$ is zero because $l'\sigma_{12} - l'\Sigma_{22}\beta_0 = l'\sigma_{12} - l'\Sigma_{22}\Sigma_{22}^{-1}\sigma_{12} = 0$.

The first step in the method is to develop a two term Taylor polynomial for $f(\beta)$. The goal is to write

$$f(\beta) \approx f(\beta_0) + (\beta - \beta_0) \left. \frac{\partial f(\beta)}{\partial \beta} \right|_{\beta=\beta_0} .$$

After ordinary vector differentiation,

$$\rho_{X'l,Y-X'\beta} = f(\beta) \approx \rho_{X'l,Y-X'\beta_0} - \frac{l'\Sigma_{22}}{\sqrt{l'\Sigma_{22}l}\sigma_{res}}(\beta - \beta_0) . \qquad (1.1)$$

Even though $f(\beta_0) = \rho_{X'l,Y-X'\beta_0}$ is zero, it is not left out because once sample values are employed this term becomes a random variable whose distribution is centered at zero, and is central to the argument.

2 Derivation of a Universal Method of Multiple Linear Regression Through Correlation

Nothing so far has depended on a particular correlation coefficient, but to continue, one must be chosen. While nearly any correlation coefficient (CC) can be used, a full derivation is given just for GDCC; see Gideon and Hollister (1987). Familiarity with GDCC is not necessary to follow the arguments, but the fact that a nonparametric correlation coefficient (NPCC) — moreover, one based on maxima and not linearity — can be implemented in a cohesive fashion in multiple regression is the key point. Using Pearson's correlation coefficient would derive the classical least squares result.

Gideon and Hollister (1987) show that for joint normal random variables Z_1, Z_2 the population value of GDCC on (Z_1, Z_2) denoted by $\rho_{gd}(Z_1,\ Z_2)$ is $\frac{2}{\pi}\sin^{-1}\rho_{Z_1,\ Z_2}$ (Kendall's Tau is the same) where $\rho_{Z_1,\ Z_2}$ is the bivariate normal correlation parameter between Z_1 and Z_2. Incidentally, this implies that $\sin(\frac{\pi}{2}\rho_{gd})$, not ρ_{gd}, estimates $\rho_{Z_1,\ Z_2}$. Note the enhanced notation, $\rho_{gd}(Z_1, Z_2)$, to reference a population correlation coefficient other than $\rho_{Z_1,\ Z_2}$, whose meaning has not changed. Note also that $\rho_{gd}(Z_1,\ Z_2)$ is written instead of ρ_{gd,Z_1,Z_2}. For random variables $X'l$ and $Y - X'\beta$, set $g(\beta) = \rho_{gd}(X'l, Y - X'\beta) = \frac{2}{\pi}\sin^{-1} f(\beta)$, where $f(\beta) = \rho_{X'l,Y-X'\beta}$ as above. The truncated Taylor series for $g(\beta)$ is

$$\begin{aligned} g(\beta) &= \rho_{gd}(X'l, Y - X'\beta) \\ &\approx \rho_{gd}(X'l, Y - X'\beta_0) + \left. \frac{\partial}{\partial\beta}\rho_{gd}(X'l, Y - X'\beta) \right|_{\beta=\beta_0} (\beta - \beta_0) . \end{aligned}$$

The partial derivative is

$$\frac{\partial}{\partial\beta}\rho_{gd}(X'l, Y - X'\beta) = \frac{2}{\pi}\frac{1}{\sqrt{1-\rho^2_{X'l,Y-X'\beta}}}\,\frac{\partial}{\partial\beta}\rho_{X'l,Y-X'l}\bigg|_{\beta=\beta_0}.$$

At $\beta = \beta_0$, $X'l$ and $Y - X'\beta_0$ are independent random variables, so $\rho_{X'l,Y-X'\beta_0} = 0$, and the latter partial derivative is

$$\frac{-l'\Sigma_{22}}{\sqrt{l'\Sigma_{22}l}\,\sigma_{res}}.$$

The truncated Taylor series becomes

$$\begin{aligned} g(\beta) &= \rho_{gd}(X'l, Y - X'\beta) \\ &\approx \rho_{gd}(X'l, Y - X'\beta_0) - \frac{2}{\pi}\frac{l'\Sigma_{22}}{\sqrt{l'\Sigma_{22}l}\,\sigma_{res}}(\beta - \beta_0)\,. \end{aligned} \tag{2.1}$$

To prepare for the random sample approximation solve

$$r_{gd}(x_i, y - X_{n\times p}\hat{\beta}_{gd}) = 0, \quad i = 1, 2, \cdots, p \tag{2.2}$$

for $\hat{\beta}_{gd}$ with data $X_{n\times p}$ and y and sample correlation coefficient r_{gd}, which is the sample counterpart of ρ_{gd}. Rummel (1991) shows how to solve equations (2.2) using Gauss-Seidel iterations. If Pearson's r_p is used, the set of equations (2.2) are equivalent to the usual least squares normal equations (without the intercept) and the solution vector is the standard least squares result. Motivation for both this and the GDCC formulation is found in publication #8 on the author's Website. As the correlation coefficient is varied, the set of equations changes and different solutions are obtained. The equations are valid for any correlation coefficient and are called "regression equations." When the correlation coefficient is not Pearson's, they generalize the normal equations of the classical case; they are used extensively in the author's correlation estimation system (CES). Incidentally, correlation coefficients are invariant with respect to location parameters, so this paper is solely concerned with inference on the regression coefficients. The intercept estimation comes afterward and is dealt with in Gideon and Rothan (2010).

Next substitute data and $\hat{\beta}_{gd}$ for β into the Taylor polynomial, obtaining

$$\begin{aligned} 0 \approx g(\hat{\beta}_{gd}) &= r_{gd}(X_{n\times p}l, y - X_{n\times p}\hat{\beta}_{gd}) \\ &\approx r_{gd}(X_{n\times p}l, y - X_{n\times p}\beta_0) - \frac{2}{\pi}\frac{l'\Sigma_{22}}{\sqrt{l'\Sigma_{22}l}\,\sigma_{res}}(\hat{\beta}_{gd} - \beta_0)\,. \end{aligned} \tag{2.3}$$

The GDCC does not have the same linearity properties that Pearson's r_p has and so it is not necessarily true that $r_{gd}(X_{n\times p}l, y - X_{n\times p}\hat{\beta}_{gd})$ is exactly zero; however, computer simulations have shown that $r_{gd}(X_{n\times p}l, y - X_{n\times p}\hat{\beta}_{gd})$ is zero or very close to zero. In the case of only one $l_i \neq 0$ this last equation becomes one of the equations in (2.2) and hence is exactly zero. Again $\rho_{gd}(X'l, Y - X'\beta_0)$ is zero and its sample equivalent multiplied by $\sqrt{n}$, $\sqrt{n}r_{gd}(X_{n\times p}l, y - X_{n\times p}\hat{\beta}_{gd})$, has an approximate N(0,1) distribution as shown in Gideon, Prentice, Pyke (1989). It now follows that

$$\frac{2}{\pi}\sqrt{n}\frac{l'\Sigma_{22}}{\sqrt{l'\Sigma_{22}l}\,\sigma_{res}}(\hat{\beta}_{gd} - \beta_0)$$

has an approximate N(0,1) distribution. Consequently, $l'\Sigma_{22}(\hat{\beta}_{gd} - \beta_0)$ is

$$N\left(0, \frac{\pi^2 l'\Sigma_{22} l \sigma^2_{res}}{4n}\right).$$

To connect to more common notation, let $l'\Sigma_{22} = k'$, so

$$k'(\hat{\beta}_{gd} - \beta_0) \text{ is approximately } N\left(0, \frac{\pi^2(k'\Sigma_{22}^{-1}k)\sigma^2_{res}}{4n}\right) \tag{2.4}$$

where $\hat{\beta}_{gd}$ solves the regression equations (2.2).

As a special case let k be a vector of 0s except for a 1 in the i^{th} position. The above result gives the asymptotic distribution of $\hat{\beta}_{i,gd} - \beta_0$ as

$$N\left(0, \frac{\pi^2\sigma^{ii}\sigma^2_{res}}{4n}\right)$$

where $\hat{\beta}_{i,gd}$ is the i^{th} component of $\hat{\beta}_{gd}$ and σ^{ii} is the (i,i) element of Σ_{22}^{-1}.

The work for Kendall's Tau is essentially the same because its population value is the same as for GDCC. The regression equations for Tau can be solved by Gauss-Seidel iterations involving the medians of elementary slopes and the solution denoted by $\hat{\beta}_\tau$. For large n, $\frac{3}{2}\sqrt{n-1}\,\tau$ has an approximate N(0,1) distribution, and so $\frac{3}{2}\sqrt{n-1}$ is the necessary multiplier; i.e. $k'(\hat{\beta}_\tau - \beta_0)$ has an asymptotic

$$N\left(0, \frac{\pi^2(k'\Sigma_{22}^{-1}k)\sigma^2_{res}}{9(n-1)}\right)$$

distribution. For simple linear regression $\hat{\beta}_\tau - \beta_0$ has an asymptotic

$$N\left(0, \frac{\pi^2 \sigma_{res}^2}{9(n-1)\sigma_x^2}\right)$$

distribution. (See Sen (1968) for the simple linear regression case, and publication #5 on the author's Website for an illustrated look at this procedure specialized to Kendall's Tau).

The CES includes the work of Jaeckel (1972) that is summarized in Hettmansperger (1984). In Hettmansperger's notation, let a be a score function and R represent ranks. For bivariate random variable (X, Y), consider the correlation coefficient $r_p(x, a[R(y)])$, where r_p is the Pearson correlation coefficient. Then for the multiple regression model, the system of equations in (5.2.8) is equivalent to $r_p(x_i, a[R(y_i - x_i'\beta)]) = 0,\ i = 1, 2, \ldots p$. To use CES, the population parameter needs to be known and also the asymptotic distribution of the sample version. Moreover, because the GDCC of the bivariate Cauchy is also $\frac{2}{\pi}\sin^{-1}\rho$, statistical analysis is possible for the Cauchy distribution where moments do not exist.

3 An Example of Multiple Regression with the 1992 Atlanta Braves Statistics

Far more regressions were run than appear here to illustrate the concepts of this paper. The example chosen shows how the correlation estimation technique parallels the standard multiple regression analysis, thus demonstrating that it is as viable as classical regression analysis. Though estimation from correlation is not the standard approach, it gives a cohesiveness to the analysis as it is valid for every correlation coefficient and rivals other robust techniques. Besides the distribution technique for the slopes, the estimates of the variation structure are given as shown in Gideon and Rothan (2010), including residual error, standard deviations of the slopes, and multiple correlation coefficient. Partial correlation coefficients can also be computed. Rank based correlations devalue extreme values; this allows GDCC to be far more robust than classical least squares as is seen in this example.

The generality of the technique includes dealing with tied values, so a baseball example was chosen because the data have numerous tied values; the data set (bb92) is on the Website. The max-min global method of dealing with ties as shown in Gideon and Hollister (1987), rather than the

local averaging technique allows NPCCs to be used on all data. This data has extreme values but none can be considered as outliers, so techniques of removing or reweighting are not appropriate. Also, one expects that hits and runs are fairly highly correlated, so this type of example tests the convergence of the numerical technique for solving the regression equations. A four variable regression was run on the baseball data with the response variable y being the length of a game in hours; 175 games were played. The regressor variables are:

x_1, the total number of runs by both teams in a game,

x_2, the total number of hits by both teams in a game,

x_3, the total number of runners by both teams left on base in a game,

x_4, the total number of pitchers used in a game by both teams.

Interest is in determining how various conditions in a game affect the length of the game. The main purpose is to use the asymptotic distributions of the slopes to compare least squares (LS) to the correlation estimation system (CES). This is accomplished by using the Pearson correlation coefficient for LS regression and the Greatest Deviation Correlation Coefficient (GDCC) for the CES. Other nonparametric correlation coefficients, as discussed in Gideon (2007), would be feasible as well. The residual standard deviations are compared and surprisingly the one derived from GDCC is less than that of LS. Also the multiple CCs are computed and quantile plots on the residuals are discussed.

Although time is a continuous random variable all the regressor variables are discrete; so at best for the classical analysis only an approximate multivariate normal distribution would model the data. All classical inference is based on the normal distribution or central limit theorems that give asymptotic results. Although the CES is based on limit theorems on continuous data, the results with the Taylor series appear good even though all the regressor variables are discrete. However, more work needs to be done on the asymptotics for discrete data. Also needed is more study of the merits of the max-min tied value procedure versus the standard local averaging technique.

The four-variate regression hyperplanes are

$$\text{LS } \hat{y} = 1.5753 + 0.0217x_1 - 0.0127x_2 + 0.0533x_3 + 0.0897x_4$$

with $r_p(y, \hat{y}) = \sqrt{0.6332} = .8205$ and $\hat{\sigma}_{res} = 0.2556$ on 170 degrees of freedom, and from software R, the P-values for variables 1, 2, 3, 4 are respectively, 0.0142, 0.1514, 0.0000, and 0.0000.

GDCC $\hat{y} = 1.7473 + 0.0457x_1 - 0.0310x_2 + 0.0567x_3 + 0.0718x_4$

with $r_{gd}(y, \hat{y}) = 0.5805$, $\sin(\pi r_{gd}/2) = 0.7906$, and $\hat{\sigma}_{res} = 0.2280$, and from the special case following result (2.4) and Gideon and Rothan (2010), the P-values for variables 1, 2, 3, 4 are respectively, 0.0001, 0.0045, 0.0000, 0.0000.

In this four-variable regression LS has a slightly higher multiple correlation coefficient, 0.8205 compared to 0.7906, but GDCC has a slightly smaller residual SE, somewhat at odds. The two slopes with the biggest difference between the two regressions are those for X_1 and X_2. The coefficient of X_2 for GDCC is over twice as large as for LS. If the value of the coefficient of X_2 for GDCC had been the LS coefficient, it would have been very significant, as the P-value (0.0003) is much lower than the 0.1514 given above (t-value -3.52). The coefficient of X_1 in the GDCC regression is also more than twice that of LS. So there is a real difference in these regressions.

The normal quantile plots of the residuals for the four-variable regression show 5 extremes for LS, and 4 to 6 (depending on visual judgement) for GDCC, with all the remaining residuals lying very close to the GDCC line. However, the distance from the GDCC line to the unusual residuals is much greater than that for LS. This explains one of the differences in this regression output. When the GDCC line is compared to the LS line on the normal quantile residual plot, the GDCC gives a better evaluation criterion. This is because the GDCC line goes through more of the sorted residuals and is not swayed by the extremes. So visually one can check more easily for normality. GDCC obtains a smaller residual SE by not weighting the very few unusual points as much as LS does. Whether or not the difference in the coefficients and the GDCC standard deviation of (0.2280)60=13.7 minutes is meaningful to a data analyst compared to (0.2556)60=15.3 minutes is entirely subjective. In general the model with the smaller variation estimate is preferable.

A small simulation study was done on the distribution of the GDCC on the variables time of game (continuous) and number of pitchers (discrete); i.e., variables Y and X_4 above. This study showed that indeed the asymptotic distribution is fairly normal even for small to moderate sample sizes

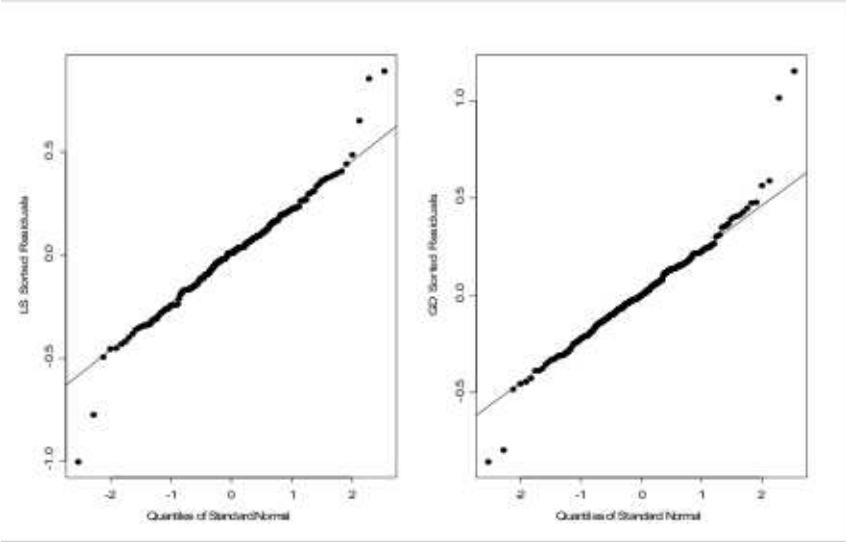

Figure 1: Normal Quantile Residual Plots for LS and GDCC

(> 10) which was surprising because GDCC itself converges very slowly to normality.

4 Conclusion

This article sets the framework for a very general method of multiple regression based on the distribution and population values of correlation coefficients. The quality of the GDCC results should eliminate lingering doubts as to the validity of this and other NP methods in linear regression. Some comments were given on the use of Kendall's Tau in the CES method. There are six other correlations in Gideon (2007) to which the process in this article can be applied. Which correlation to use on a particular data set is still a research question. The L-one correlation coefficient in Gideon (2007) could be profitably used whenever L-one methods are appropriate. A long-term goal would be to have a computer package that can select different correlation coefficients to use in performing the multiple regression analysis. Another important goal for this article was to reemphasize that the problem of tied values is apparently not an issue when the max-min method is used. Thus, the CES using rank based or continuous (including Pearson's which is equivalent to least squares) correlation coefficients in multiple linear regression estimation is not only a very viable technique, but also provides a

consistency not always found in other methods. These multiple regression results can segue into other estimation areas of statistics; some of these ideas can be pursued by consulting the Website. They include, for example, estimation in nonlinear regression, generalized linear models, and time series, as elucidated by Sheng (2002). Parameter estimation on individual distributions can also be done. While not all applications have been explored, enough has been done to be very optimistic about the direction, usefulness, and generality of this work. The CES provides a simple way (if computer programs have been written) to use robust methods in these latter areas without having to resort to data manipulation.

Acknowledgements. Special thanks to former students, Steven Rummel, Adele Rothan, Jacquelynn Miller, and especially to Carol Ulsafer, collaborator and editorial assistant. Also thanks to the referees and editor for their patience and valuable suggestions.

References

BURG, KARL V. (1975). *Statistical Models in Applied Sciences.* John Wiley and Sons, N.Y.

GIBBONS, J. D. and CHAKRABERTI, S. (1992). *Nonparametric Statistical Inference,* 3^{rd} ed. Marcel Dekker, Inc., N.Y.

GIDEON, R. A. (2008). Kendall's τ In Correlation and Regression, in progress.

GIDEON, R. A. (2007). The Correlation Coefficients, *Journal of Modern Applied Statistical Methods*, **6**, 517–529.

GIDEON, R. A., PRENTICE, M. J. and PYKE, R. (1989). The Limiting Distribution of the Rank Correlation Coefficient r_{gd}. In: *Contributions to Probability and Statistics (Essays in Honor of Ingram Olkin*) edited by Gleser, L. J., Perlman, M. D., Press, S. J., and Sampson, A. R. Springer-Verlag, N.Y., 217–226.

GIDEON, R. A. and HOLLISTER, R. A. (1987). A Rank Correlation Coefficient Resistant to Outliers, *J. Amer. Statist. Assoc.* **82**, 656–666.

GIDEON, R. A. and ROTHAN, A. M., CSJ (2010). Location and Scale Estimation with Correlation Coefficients. *Communications in Statistics–Theory and Methods*, accepted for publication.

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

JAECKEL, L. A. (1972). Estimating Regression Coefficients by Minimizing the Dispersion of the Residuals, *Ann. Math. Statist.*, **43**, 1449–1458.

MILLER, JACQUELYNN (1995). Multiple Regression Development with GDCC, Masters Thesis. University of Montana.

RUMMEL, STEVEN E. (1991). A Procedure for Obtaining a Robust Regression Employing the Greatest Deviation Correlation Coefficient, Unpublished Ph.D. Dissertation, University of Montana, Missoula, MT 59812, full text accessible through UMI ProQuest Digital Dissertations.

Sen, P.K. (1968). Estimates of the Regression Coefficient based on Kendall's Tau. *J. Amer. Statist. Assoc.*, **63**, 1379–1389.

Sheng, HuaiQing (2002). Estimation in Generalized Linear Models and Time Series Models with Nonparmetric Correlation Coefficients, Unpublished Ph.D. Dissertation, University of Montana, Missoula, MT 59812, full text accessible through http://wwwlib.umi.com/dissertations/fullcit/3041406.

Website: www.math.umt.edu/gideon.

Rudy A. Gideon
Department of Mathematical Sciences
University of Montana
Missoula, MT 59812, USA
E-mail: gideonr@mso.umt.edu

Paper received March 2008; revised March 2010.