Correlation and Regression without Sums of Squares

(Kendall's Tau)

Rudy A. Gideon

ABSTRACT

This short piece provides an introduction to the use of Kendall's t in correlation and simple linear regression. The error estimate also uses Kendall's t so that sums of squares are avoided. A population or random variable approach using elementary slopes demonstrates a natural way to introduce Kendall's t . This paper uses population and sample concepts to fuse correlation and regression together into a correlation estimation system (CES) that allows regression to follow directly from correlation. A slightly different formulation of t is needed to do this. In Hollander and Wolfe (1999) as well as other publications there are separate areas on correlation and regression and it is not easy to see any connection and this paper hopes to correct that. It was pointed out by Huber (1981) that classical regression loses its optimality if only one percent of the data is questionable. Since this is almost always the case, regression with robust properties (implying that a few errant data points do not severely affect the inference) such as presented here are preferable.

Key words: rank statistics, nonparametric, elementary slopes, concordance

1. INTRODUCTION

A few short sections set the stage. Let (X,Y) be a continuous bivariate random variable and (X_1,Y_1) and (X_2,Y_2) two independent outcomes. Assume interest lies in both the correlation and the simple linear relationship between X and Y . The random variable

$\frac{Y_1 - Y_2}{X_1 - X_2}$ is known as an *elementary slope* (ELS) as it is the slope of the line between the two points (X_1, Y_1) and (X_2, Y_2). An elementary slope is called concordant if it is positive and discordant otherwise. The population version of Kendall's t can be formulated as the parameter

$t(X,Y) = P(\frac{Y_1 - Y_2}{X_1 - X_2} > 0) - P(\frac{Y_1 - Y_2}{X_1 - X_2} < 0) = 2P(\frac{Y_1 - Y_2}{X_1 - X_2} > 0) - 1$. Thus t is simply the excess of concordant over discordant slopes, or vice-versa. Let $p_c = P(\frac{Y_1 - Y_2}{X_1 - X_2} > 0)$ and $p_d = 1 - p_c$. Note that t lies between -1 and +1, and if $t = 0$, then $P(\frac{Y_1 - Y_2}{X_1 - X_2} > 0) = 1/2$.

2. POPULATION VALUE OF t FOR THE BIVARIATE NORMAL

Assume now that (X,Y) is a bivariate normal random variable with means m_x, m_y, variances s_x^2, s_y^2 and covariance $s_x s_y r$. Then the ELS random variable $\frac{Y_1 - Y_2}{X_1 - X_2}$ has a distribution equal to $r\frac{s_y}{s_x} + \frac{s_y}{s_x}\sqrt{1 - r^2}R_o$ where R_o has the standard Cauchy distribution. The density and cumulative distribution function of R_o are $1/(p(1 + x^2))$ and $\frac{1}{2} + \frac{\arctan(x)}{p}$ for x over the real line. The elementary slopes and correlation coefficient r can now be related.

$p_c = P(\frac{Y_1 - Y_2}{X_1 - X_2} > 0) = P(r\frac{s_y}{s_x} + \frac{s_y}{s_x}\sqrt{1-r^2}R_o > 0) = \frac{1}{2} + \frac{\arctan(r/\sqrt{1-r^2})}{p}$, and

$t(X,Y) = \frac{\arctan(r/\sqrt{1-r^2})}{p}$ This can be solved for r in terms of $P(\frac{Y_1 - Y_2}{X_1 - X_2} > 0) = p_c$.

After a wee bit of algebra the equation becomes $r = \sin(p(p_c - 0.5))$. Now because $t = 2p_c - 1$, this latter equation can also be written as $r = \sin(pt/2)$. Because of the symmetry of R_o about 0, the random variable $\frac{Y_1 - Y_2}{X_1 - X_2}$ is symmetric about the slope $r\frac{s_y}{s_x}$. This is used to discuss a natural estimate of the slope in simple linear regression.

3. THE REGRESSION SETTING

Two more theoretical relationships are needed. First the simple linear regression model for the bivariate normal is $E(Y|X = x) = m_y + r\frac{s_y}{s_x}(x - m_x)$; that is, the slope is $b = r\frac{s_y}{s_x}$ and the intercept is $a = m_y - r\frac{s_y}{s_x}m_x$. Second is needed the distribution of $Y - r\frac{s_y}{s_x}X$. From properties of the bivariate normal and standard variable transformation theory, this distribution is normal with mean $m_y - r\frac{s_y}{s_x}m_x$ and variance $s_y^2(1-r^2)$. Note that this distribution is also symmetric about its mean, which is the intercept in the simple linear regression.

4. POPULATION REGRESSION WITH KENDALL'S TAU

The regression residuals are uncorrelated with the regressor variable for a correlation coefficient parameter q , if b is found such that

$$q(X, Y - bX) = 0. \quad (1)$$

In CES this equation is called the population regression equation. Solving for *b* gives the slope of the population linear model. It is now shown that the solution of $t(X, Y - bX) = 0$ is $b = b = r\frac{s_y}{s_x}$;that is t gives the correct slope in the population regression equation. To prove this, start with $t(X, Y - bX) = 0$ and from the definition of t , $P(\frac{(Y_1 - bX_1) - (Y_2 - bX_2)}{(X_1 - X_2)} > 0) = P(\frac{Y_1 - Y_2}{X_1 - X_2} - b > 0) = P(\frac{Y_1 - Y_2}{X_1 - X_2} > b) = 1/2$. Now by substituting the distribution of an ELS and for $b = b = r\frac{s_y}{s_x}$, $P(\frac{Y_1 - Y_2}{X_1 - X_2} > b) =$ $P(r\frac{s_y}{s_x} + \frac{s_y}{s_x}\sqrt{1 - r^2}R_0 > b) = P(\frac{s_y}{s_x}\sqrt{1 - r^2}R_0 > 0) = P(R_0 > 0) = 1/2.$ Again the last step follows from the symmetry of R_0 about 0. So indeed the population regression coefficient is obtained via t and the population regression equation.

5. CORRELATION AND SAMPLE REGRESSION WITH KENDALL'S TAU

5.1 Estimation of Kendall's t

For a random sample of size n from an absolutely continuous distribution with data in column vectors (x, y), form the elementary slopes — all $\binom{n}{2}$ of them. Estimate p_c, the probability of concordance, by $\hat{p}_c$, as the number of concordant elementary slopes divided by $\binom{n}{2}$. Because $p_d = 1 - p_c$, $\hat{p}_d = 1 - \hat{p}_c$. Then the sample value of Kendall's t is $\hat{t}(x, y) = \hat{p}_c - \hat{p}_d$, that is, the difference between the relative number of concordant and the relative number of discordant elementary slopes. It follows from the population ideas above that $\hat{r} = \sin(p(\hat{p}_c - .5)) = \sin(p\hat{t}/2)$.

5.2 Estimate of Slope

Note that for the bivariate normal the population median of ELS is $r\frac{s_y}{s_x}$ as given in Section 2. It is known that the sample median from a symmetric population has the population median as its point of symmetry. In the case of the set of elementary slopes, there is not total independence, but one would still expect the sample median to vary about the population median. In other words, a natural estimate of the slope, $r\frac{s_y}{s_x}$, is obtained as the median of the sample elementary slopes. It is now shown that t gives this estimate by using the sample equivalent of equation (1). To have the residuals and *x* uncorrelated, in equation (1) replace random variables by data and obtain the sample regression equation

$$t(x, y - bx) = 0. \qquad (2)$$

But also $t(x, y - bx) = p_c(b) - p_d(b)$ where $p_c(b)$ is the fraction of ELSs that are positive for a chosen b and $p_d(b)$ is the fraction of ELSs that are negative for a chosen b.

The ELSs for the sample with a given b are $\frac{(y_i - bx_i) - (y_j - bx_j)}{x_i - x_j} = \frac{y_i - y_j}{x_i - x_j} - b$ for $1 \le i < j \le n$. The sample regression equation is solved if there are an equal number of concordant and discordant ELSs; that is, when $p_c(b) = p_d(b)$ which means $\#\{\frac{y_i - y_j}{x_i - x_j} > b\} = \#\{\frac{y_i - y_j}{x_i - x_j} < b\}$ or that $b = median\{\frac{y_i - y_j}{x_i - x_j}\}$. Sen (1968) is a good source for a different approach to slope estimation via Kendall's Tau.

5.3 Intercept

After the slope b is estimated the intercept is estimated by the median of the set of residuals $y_i - bx_i$, call it a. This is because b is estimating $r\frac{s_y}{s_x}$ and $Y - r\frac{s_y}{s_x}X$ is symmetric about the population intercept $m_y - r\frac{s_y}{s_x}m_x$. Then the Kendall estimate of the linear relationship is $\hat{y}_i = a + bx_i$.

5.4 Inference on the Slope

For inference on $b = r\frac{s_y}{s_x}$, the asymptotic result in Gideon (2008) is used; it is shown there that $\hat{b}_t - b$ has an asymptotic $N(0, \frac{p^2 s_{res}^2}{9(n-1)s_x^2})$ distribution. (Sen 1968 also develops a different inference). The quantity s_{res}^2 is estimated from the residuals around the Kendall regression line, while s_x^2 is estimated by a sample variance of the regressor

variable *x*. In the spirit of using just Kendall's t to estimate all quantities, here is the CES approach. Let $z_i = \Phi^{-1}(i/(n+1))$ for $i = 1,2,\ldots,n$ where Φ is the distribution function of a N(0,1) random variable. Now the z_i are used with the order statistics for the residuals ($res_{(i)}$) and ($x_{(i)}$) to estimate the standard deviations. Solving the regression equations (with the same logic as above) but now with ordered data denoted by the superscript ^{o}, $t(z^o, x^o - \hat{s}_x z^o) = 0$ and $t(z^o, res^o - \hat{s}_{res} z^o) = 0$, leads to the following (Gideon and Rothan 2007):

$$\hat{s}_x = median\left(\frac{x_{(i)} - x_{(j)}}{z_{(i)} - z_{(j)}}\right) \text{ and } \hat{s}_{res} = median\left(\frac{res_{(i)} - res_{(j)}}{z_{(i)} - z_{(j)}}\right), \text{ for } 1 \le i < j \le n.$$

6. EXAMPLES AND SIMULATIONS

The above work is next carried out for the example in Sen (1968), *x*={1,2,3,4,10,12,18}, and *y*={9,15,19,20,45,55,78}. First, the t – estimated intercept and slope are 6 and 4. The two estimated standard deviations are $\hat{s}_x$ = 7.41 and $\hat{s}_{res}$ = 1.48; the classical results are 6.34 and 1.21, respectively. The ratio of the $\hat{s}_{res}/\hat{s}_x$ in the SD of the slope above can be computed in two ways. First by computing $\hat{s}_x$ and $\hat{s}_{res}$ and taking the ratio. A second way that avoids involving a distribution is possible with the CES by computing $ratio = median\left(\frac{res_{(i)} - res_{(j)}}{x_{(i)} - x_{(j)}}\right)$ directly; that is, solve $t(x^o, res^o - ratio * x^o) = 0$. Both give the same result, 0.200. The estimated SD is $\sqrt{\frac{p^2 \hat{s}_{res}^2}{9(n-1)\hat{s}_x^2}}$ which is $\frac{p}{3\sqrt{6}} * 0.200 =$ 0.0855. So the approximate 93% confidence interval (same level as in Sen) is $4 \pm 1.812 * 0.0855 = 4 \pm 0.155$ where 1.812 is the upper 0.965 quantile of a standard normal. The lower number is 3.85 compared to Sen's 3.71 and the upper number is 4.15

compared to 4.18; considering that the sample size is only 7, these numbers are very close. In his paper Sen points out why Tau (robust) should be used so it is not necessary to repeat the reasons here.

A number of simulations with various parameters were run in order to further authenticate the asymptotic distribution of the slope, but only one of each type is discussed. First a simple linear regression was constructed with standard normal variables for a sample of size 25 and a confidence coefficient of 50 %. The model was $y = rx + \sqrt{1 - r^2}e$ with x and e independent N(0,1), and so the correlation r is also the slope of the regression line. Of 1000 simulations, 48 % contained the true slope — just about what is expected.

A second simple linear regression was run on binomial variables. First x was generated as a binomial random variable based on 15 Bernoulli trials with the probability of success $p = 1/2$. Then y was generated by choosing a slope and adding error with the same Binomial but centered by subtracting $n*p$, so $y = slope * x + (B(15,1/2) - 7.5)$. Again 25 observations were taken and a 50 % confidence interval constructed. Of 1000 simulations, 63% of the confidence intervals included the true slope. This, of course, is far more than expected. There were many tied values on both the x and y data for the binomial and so the asymptotic distribution may be less accurate, but the fact that the confidence interval level was higher than expected is promising. A further motivation for giving this last example is to show that CES with Kendall's can be appropriately used on discrete data, just as normal theory techniques. In both cases, it is known that the results are only approximate.

7. CONCLUSION

Most practitioners of statistics focus only on least squares and sums of squares procedures and have the impression that these are the only easy to use and straightforward procedures for regression analyses. The compact synthesis (of population, sample, correlation, and regression) presented herein is meant to promote the use of Kendall's Tau and the CES to show that there are alternatives that are not only easy to use but are in addition inherently robust. Other correlation coefficients, both nonparametric and continuous, could be studied in a similar manner; the author has examined a few of these, but only Greatest Deviation Correlation Coefficient (Gideon and Hollister 1987) has been extensively studied. To summarize, both the population and sample versions of Kendall's Tau are used to examine correlation and the parameters in a simple linear regression. It should be noted that the invariance properties given in Sen (1968) are easily approached through the correlation notions in this paper. What, perhaps, has held up using Kendall's and Sen's concepts in statistical analysis is that their work in correlation and regression has not been viewed as related much less as a universal system; correcting this has been one of the goals of this paper. Additionally the advance of computing has made this approach very feasible. This approach also leads to a nice way to extend Kendall's Tau into multiple linear regression; see Gideon (2008).

Customized R or S-Plus programs are given in the appendix to obtain the Kendall estimates in simple linear regression. These use the max-min method for dealing with ties found in Gideon and Hollister (1987) so that estimates are available no matter how many ties there are. This method of tie breaking gives the same results as those in packaged S-

Plus routines but not as those in the R routines. However, this max-min tie breaking method is the only feasible method for the Greatest Deviation Correlation Coefficient and so is more general.

APPENDIX: R OR S-PLUS ROUTINES FOR CORRELATION AND SIMPLE LINEAR REGRESSION WITH KENDALL'S TAU

```
# computing   the two unique vectors with ties present: the function is tauuniq
tauuniq <- function(x,y) {
n <- length(x)
e <- 1:n
xrr <- n+1 -rank(x)
xtp <- x[order(y,x)]
xtn <- x[order(y,xrr)]
rkyp <- order(xtp,e)
rkyn <- order(xtn,n:1)
out <- cbind(rkyp,rkyn)
out  }

# calculation of Kendall's tau on unique max-min vectors
# the function is rtau
rtau <- function(x,y){
ot <- tauuniq(x,y)
rkyp <- ot[,1]
rkyn <- ot[,2]
dyp <- 0
dyn <- 0
n <- length(x)
n2 <- ((n*(n-1))/2)
n1 <- n-1
for(i in 1:n1) {j <- i+1
tempp <- rkyp[i]-rkyp[j:n]
tempn <- rkyn[i]-rkyn[j:n]
dyp <- dyp + sum( tempp<0)
dyn <- dyn + sum( tempn<0)
}
out <- (dyp + dyn)/n2 -1
out  }

# output is slope and intercept, function name tauslp in positions 1 and 2
tauslp <- function( x,y) {
rat <- c(outer(y,y,"-")/outer(x,x,"-"))
ratv <- rat[!is.na(rat)]
slp <- median(ratv)
```

```
res <- y - slp * x
aint <- median(res)
res <- res - aint
ck <- rtau(x,res)
ck1 <- sum(res)
ck2 <- median(res)
out <- c(slp,aint,ck,ck1,ck2)
out   }
```

REFERENCES

R.A. Gideon, The Correlation Coefficients, *Journal of Modern Applied Statistical Methods* **6** (2007) .

R.A. Gideon, The Relationship between a Correlation Coefficient and its Associated Slope Estimates in Multiple Linear Regression, *Sankhya* (2008) under review.

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

R.A. Gideon and A. M. Rothan CSJ, Location and Scale Estimation with Correlation Coefficients, *Communications in Statistics-Theory and Methods* (2008) under review.

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

M. Hollander and D.A. Wolfe, *Nonparametric Statistical Methods*, 2nd ed., John Wiley & Sons, 1999.

P.J. Huber, *Robust Statistics*, John Wiley & Sons, 1981.

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

Website www.math.umt.edu/gideon contains copies of the author's unpublished work.