# RANDOM VARIABLES, REGRESSION, AND THE GREATEST DEVIATION CORRELATION COEFFICIENT

RUDY A. GIDEON
*University of Montana, Missoula*

Summary. For a bivariate random variable (X, Y), a general definition of a slope parameter is given for the simple linear regression model. A regression equation is defined for a nonparametric correlation coefficient based on greatest deviations such that the slope parameter is obtained when the equation is solved. This result is explored geometrically and then is used to relate the asymptotic distribution of the sampling estimate of the slope to the asymptotic distribution of the correlation coefficient. The bivariate normal and Cauchy distributions are used to illustrate the principles and the general concepts apply to any correlation coefficient.

## 1. INTRODUCTION

If (X,Y) is a jointly continuous random variable, then if the appropriate moments exist, E(Y|X=x) is defined as the regression line of Y on X. The bivariate normal distribution with parameters $(m_1, m_2, s_1^2, s_2^2, r)$ where the subscripts 1 and 2 are for X and Y, respectively, has the form $E(Y|X = x) = m_2 + (s_2/s_1)r(x - m_1)$. Textbooks, such as Ross(1988) use the bivariate normal distribution to show that the least squares criterion recovers this regression equation. In fact, Ross states that "the best linear predictor" in cases where the means, variances, and correlation are known is given by choosing "a" and "b" to minimize $E(Y - (a + bX))^2$. It is the purpose of this article to show that there are other "best linear predictors" such as the method in this article using the Greatest Deviation Correlation Coefficient, $r_g$ (Gideon and Hollister (1987)), which recover this regression equation for the bivariate normal distribution. Even further, the method presented here can recover the theoretical regression line for the bivariate Cauchy distribution after a more general definition of a slope parameter is presented. A normal distribution example will be given to illustrate the general nature of the $r_g$ method, and then a more general definition is given for a slope parameter so that the technique can be used for the Cauchy distribution. A general framework of regression based on random sampling with $r_g$ has been given in Gideon, Rummel, Li (1994) and Gideon et al. (1993) and this paper gives some theoretical justification for such work. The underlying basis, in using nonparametric correlation coefficient $r_g$, seems to be an equalizing of the "distance" away from perfect positive and negative correlation and this will be seen by studying the distribution of (X,Y-bX) where b is the regression coefficient.

## 2. THE BIVARIATE NORMAL AND THE REGRESSION CRITERIA (MSSV)

In order to make the concepts clear, the standardized bivariate normal distribution is used first; thus, X and Y have N(0,1) marginal distributions and $r$ is the correlation parameter. The theoretical regression line is $E(Y|X = x) = r x$ and this is pictured in Figure 1 when $r$ >0. Let $f(x,y)$ be the bivariate density. Let Q$_i$ refer to the set of points in quadrant i, i = I, II, III, IV respectively, and P(Q$_i$) = P((X,Y) is in Q$_i$). It is known that P(Q$_1$) = P(Q$_3$) = $\frac{1}{4} + \frac{\sin^{-1} r}{2p}$ and that P(Q$_2$) = P(Q$_4$) = $\frac{1}{4} - \frac{\sin^{-1} r}{2p}$. Let $A = \{(x,y): x \ge 0, 0 \le y \le rx\}$, so that P(A) = $\int_0^{\infty}\int_0^{rx} f(x,y)\,dy\,dx = (\sin^{-1} r)/2p$ represents the probability of an observation falling in region A. This is shown in Rummel's Dissertation (1991). In Figure 1, r$_g$ is written as r(GD) and its relationship to these bivariate normal probabilities is shown. In Q$_1$ and Q$_3$ the volumes are split into two pieces about the dotted regression line. The dashed line is the y = x line, and the correlation is 0.6.

Consider the following criterion to fit a regression line (y=bx) to the standardized bivariate normal; pick b to equalize the "distances" which are really the volumes or probabilities above and below the line. Clearly, from Figure 1 with the contour lines for $f(x,y)$ shown, it can be seen that the probability of landing in the regions above or below each line are equal and some restriction must be made. Let V$_1$ and V$_4$ be the probabilities above the line and below the line within quadrants Q$_1$ and Q$_4$. First note that for $b = r$, V$_1$=P(Q$_1$) - P(A) = 1/4 and V$_4$=P(Q$_4$) + P(A) = 1/4 so that this choice of b would satisfy the regression line criterion.

Now consider $b = r + d$ where $d$ may be a plus or minus increment, and let $V_i(d)$ be the probabilities, i=1and 4 as a function of this increment. Let v be the probability of the region $\{(x,y): x > 0, bx \le y \le rx\}$ if $d$ <0 and similarly, for $d$ >0 where $bx > rx$. Now for $d$ <0, V$_1$=1/4 +v and V$_4$=1/4-v and likewise for $d$ >0, V$_1$=1/4 -v and V$_4$=1/4+v. For example, if $d = -r, b = 0,$ v=P(A), and V$_1$=P(Q$_1$)>P(Q$_4$) =V$_4$ since we are considering the case where $r$ >0. If $d = 1 - r, b = 1,$ V$_1$=P(Q$_1$)/2 = $1/8 + \sin^{-1} r/4p < V_4 = 3/8 - \sin^{-1} r/4p$. For arbitrary, $d(v),$ the sum of squares is $V_1^2 + V_4^2 = (1/4 + v)^2 + (1/4 - v)^2$, and it is easy to see that this sum of squares is minimized when v=0, or $d$ =0. Thus, if this criterion of minimizing the sum of squares of probabilities (MSSV, v for volume) to the right of a vertical line through the center of the distribution is used to determine a regression line, the correct line is chosen. Because of the symmetry of the bivariate normal distribution, if MSSV is used restricted to Q$_2$ and Q$_3$, the same result occurs. This result also holds for the Cauchy distribution as seen in a later section.

In what follows, it will be shown that a method of choosing a regression line with $r_g$, the Greatest Deviation correlation coefficient, involves equating probabilities so that essentially $r_g$ is a modified MSSV estimation procedure for the slope parameter. The correlation coefficient $r_g$ is computed by taking the difference between two supremums involving probabilities; the first measures probability away from perfect negative correlation and the second measures probability away from perfect positive correlation (see equation (3), and this concept will be illustrated after (3)). The $r_g$ slope parameter is the one that equalizes this "distance". In the bivariate Normal and Cauchy cases this method will agree with the MSSV idea because of their ellipitcal symmetry. However, for distributions with non-homogeneous variance the $r_g$ method does not necessarily agree with MSSV but is close enough to be called a modified MSSV. An example not presented here illustrated the fact that if the distribution of $(X,Y-bX)$ is symmetric with respect to a horizontal line in the case where $E(Y|X=x)=a+bx,$ then the $r_g$ selected the correct regression line. This suggests that $r_g$ gives the correct regression line in all cases in which the distribution of $(X,Y-bX)$ is symmetric about a horizontal line if $b$ is the true parameter. These concepts will now be developed for $r_g$ and the normal distribution.

## 3. THE REGRESSION EQUATION FOR $r_g$

Let (X,Y) be a bivariate random variable and $r(X,Y)=E(X-EX)(Y-EY)/(s_1 s_2)$, the correlation parameter. With a similar notation, $r_g$(X,Y) is the $r_g$ correlation parameter. For the standardized bivariate normal, let $r(X,Y)=r,$ and it is shown in Gideon and Hollister (1987) that

$$r_g(X,Y)=2(\sin^{-1} r)/p \qquad (1)$$

It is straight-forward to show that for any bivariate normal random variable $r(X,Y-bX)=(r s_2-b s_1)/s_{res}$ where $s_{res}=(Var(Y-bX))^{1/2}=(s_2^2-2br s_1 s_2+b^2 s_1^2)^{1/2}.$ For the standardized case, $r(X,Y-bX)=(r-b)/(1-2br+b^2)^{1/2}=r_b,$ say. Thus, $r_g(X,Y-bX)=2(\sin^{-1} r_b)/p$ .

This paper gives a theoretical justification of fitting a regression line with $r_g$ and one of the main ideas can now be shown. If (*x*,*y*) is an nx2 random sample from (X,Y), then solving the equation $r_g(x,y-xb)=0$ for $b$ gives the $r_g$ estimate of the slope (Gideon et al, 1994). As the sample goes to infinity, in the limit the equation becomes $r_g(X,Y-bX)=0.$ It is now shown that the unique solution is, for the standardized bivariate normal, $b=r.$ The derivative of $r_g(X,Y-bX)$ with respect to $b$ is $-2\sqrt{1-r^2}/(ps_{res}^2)$ which is less than zero for all

$r^2 \neq 1$. Therefore, $r_g(X, Y - bX)$ is monotonic decreasing in $b$ and the regression equation has a unique solution. The solution to

$$r_g(X, Y - bX) = 0 \qquad (2)$$

gives the equation $2\sin^{-1} r_b / p = 0$ which implies that $b = r$. The next section will use geometrical considerations to relate this result to the MSSV criterion.

## 4. MSSV, $r_g$, AND THE BIVARIATE NORMAL

Let the jointly continuous random variable (X,Y) have marginal cumulative distribution functions (cdf) F(x) and G(y) and joint cdf H(x,y). Let U=F(X) and V=G(Y), the probability integral transformations, and $C(u,v) = P(U \le u, V \le v) = P(X \le F^{-1}(u), Y \le G^{-1}(v))$, for $0 \le u, v \le 1$, the Copula function . In Gideon and Hollister (1987), it is shown that

$$r_g(X,Y) = 2\sup_{0 \le t \le 1} C(t, 1-t) - 2\sup_{0 \le t \le 1}(t - C(t,t)) \qquad (3)$$

$$= 2\sup_t H(F^{-1}(t), G^{-1}(1-t)) - 2\sup_t (t - H(F^{-1}(t), G^{-1}(t)).$$

The use of the population definition of $r_g$ in equation (3) will first be illustrated, as before, with the standardized bivariate normal with correlation coefficient $r$. Let $\Phi$ be the cdf of a N(0,1) random variable. Then for this case $F = G = \Phi$ and for a fixed t in the (0,1) interval, $(F^{-1}(t), G^{-1}(t)) = (\Phi^{-1}(t), \Phi^{-1}(t))$ and $(F^{-1}(t), G^{-1}(1-t)) = (\Phi^{-1}(t), \Phi^{-1}(1-t))$. Because of the symmetry of a N(0,1) random variable, these two sets of points for t in the interval (0,1) trace out lines through the origin with slopes +1 and -1, respectively. Figure 2 demonstrates geometrically the evaluation of the supremums in equation (3) by showing the evaluation for a fixed t<1/2. Figure 2a gives $a_1 = -a_2 = \Phi^{-1}(t)$ and $a_2 = \Phi^{-1}(1-t)$. Then Figures 2b and 2c show the regions where the cdf's are evaluated for two cases; $0 < b < r$ in 2c and $b = r > 0$ in 2b. The value of t was -1.28, the 10th percentile. Technically, X and $(Y - Xb)/s_{y-xb}$ must be used to keep the parametric equations above as lines with slopes $\pm 1$.

In this section, volume rather than probability of events is used because the geometrical view is being stressed. Let $W_2(t)$ be the volume over the infinite rectangle with corner at $(a_1, a_1)$ open towards the northwest in Figures 2b and 2c. Let $W_3(t) = H(F^{-1}(t), G^{-1}(1-t))$, the volume over the infinite rectangle with corner at $(a_1, a_2)$. At t=1/2, the corners of the rectangles are at (0,0) and $W_2(1/2) = P(Q_2)$ and $W_3(1/2) = P(Q_3)$. The geometrical view for a t>1/2 in shown in Figure 3. With this notation and for a fixed $b$, $r_g(X, Y - bX) = 2\sup_t W_3(t) - 2\sup_t W_2(t)$. Because of the unimodal and elliptically symmetric nature of the bivariate normal distribution and geometrical considerations, it is seen that the two functions $W_2(t)$ and $W_3(t)$ involved in the two supremums in the evaluation of $r_g(X,Y-bX)$ will have their maximums achieved at t=1/2. Since $\Phi^{-1}(1/2)=0$, the coordinates of points where the maximums are achieved are (0,0) in Figures 2 and 3. This means that in Figure 2b, $r_g(X,Y-rX)=2P(Q_3)-2P(Q_2)=0.$

For the case in Figure 2c $(0<b<r)$ it is clear that in comparing $W_2(t)$ to $W_3(t)$ that $W_3(t)$ will achieve a greater maximum. The result is that $r_g(X,Y-bX)=2P(Q_3)-2P(Q_2)=2\sin^{-1}r_b/p\,.$ Earlier work showed that $r_g(X,Y-bX)=0$ only for $b=r$ because in this case $X$ and $Y-rX$ are independent random variables so that equation (3) becomes (where G is the marginal cdf of $Y-rX$), $r_g(X,Y-rX)=2\sup_t t(1-t)-2\sup_t(t-t^2)=2(1/2*1/2)-2(1/2-1/4)=0$ .

At the earlier examples of $b$ =0 and $b$ =1, in $r_g(X,Y-bX)=2(\sin^{-1}r_b)/p$ , the value of $r_b$ is $r$ for $b=0$ and $-((1-r)/2)^{1/2}$ for $b=1$. These are now interpreted as the difference between two volumes. For $b=0,\ P(Q_1)-P(Q_4)=\sin^{-1}r/p=r_g(X,Y)/2$ and for $b=1, V_1-V_4=-1/4+(\sin^{-1}r)/2p=\sin^{-1}\left(-\sqrt{(1-r)/2}\right)/p=r_g(X,Y-X)/2<0.$ The latter result contains a trigonometric identity that can be proved by making the substitution $r=\cos t.$ The quantity $V_1$ is really measuring the "distance" (volume) of the proposed line away from perfect negative correlation and $V_4$ the "distance" away from perfect positive correlation. It should be made clear that $V_1$ and $V_4$ are volumes for (X,Y) whereas $r_b$ above refers to (X,Y - $b$ X). When (X,Y) is transformed to (X,Y - $b$ X) the region above the line y= $b$ x in $Q_1$ gets stretched into all of $Q_1$ and the region below the line y= $b$ x in $Q_1$and $Q_4$ (for $r$ >0 only) gets squeezed into $Q_4$. Thus, the volumes ($V_1$,$V_4$) for (X,Y) go to probabilities (P($Q_1$), P($Q_4$)) for (X,Y - $b$ X). Since $V_1$ -$V_4$ =$r_g$(X,Y)/2 < 0, the proposed slope gives a regression line that is still too far away from perfect positive correlation. The concept of "distance" away from perfect positive and negative correlation can be made clear by considering two absolutely continuous random variables (X,Y). First let X=Y so that there is perfect positive correlation. Then by using formula (3) it is easily seen that $\sup_{0\le t\le 1}C(t,1-t)=1/2$ and that C(t,t)=t so that C(t,1-t) which measures distance from perfect negative correlation is maximum at 1/2 whereas t - C(t,t) =0 for all t and distance from perfect positive correlation is 0, a minimum. It follows that $r_g$(X,X)=1. Second, let X = -Y so that there is perfect negative correlation. In this case C(t,1-t) = 0 for all t, and distance from perfect negative correlation is 0. It is easily shown that $\sup_t[t-C(t,t)]=\max\left[\sup_{t\le 1/2}t,\sup_{t>1/2}(1-t)\right]=\max(1/2,1/2)=1/2$ , and the distance from perfect positive correlation is a maximum. It now follows that $r_g$(X,-X)= 2(0 - 1/2) = -1.

Thus, in a regression interpretation, the magnitude of $r_g(X,Y-bX)$ measures twice the volume difference from the true regression line, and this difference gives the excess distance between perfect positive and negative correlation. If this difference is negative, $b$ must decrease but if the difference is positive $b$ must increase to obtain the true $b$. If $r_g(X,Y-bX)$=0 then $V_1 - V_4 = 0$ and a MSSV parameter has been obtained. This has a direct analogous interpretation in sampling using r$_g$ in simple linear regression as given in Gideon et al.(1994). For example, in Gideon and Hollister (1987), there was a YMCA data set of 16 points; thus, in a point process, each point has weight 1/16. Let (*x*,*y*) be this data, then r$_g$(*x*,*y*)=-3/8 = 2(3/16 - 6/16) = twice the "volume" difference from the true regression line. For this data, the r$_g$ slope and intercept estimates are -.6076923 and 13.77308, respectively, so that the sample regression line is $\hat{y} = 13.77308 - .6076923x$ and a $b$ =0 is -3/16 away, volume-wise, from the r$_g$ regression line. A scatterplot of this data appears in Gideon et al. (1989). For $b$ =0 the distance from a perfect negative regression is 3/16 whereas from perfect positive regression it is 6/16, and this indicates that a negative slope (b= -.6076923) must be used to balance the distances.

For a bivariate normal distribution with location and scale parameters $m_1, s_1 (m_2, s_2)$ for X(Y), the solution of $r_g(X,Y-bX)$=0 will be unique at $b = r s_2/s_1,$ the correct parameter. In order to complete the regression, the intercept of the regression is given and again this is based on the work in Gideon et al. (1992) and (1994). Now $E(Y|X = x) = (m_2 - r s_2 m_1/s_1) + r s_2 x/s_1$. A location estimate using r$_g$ is the average of the 1/3 and 2/3 quantiles; call this the r$_g$-mean. To complete the regression let "a" be the rg-$mean(Y - r s_2 x/s_1),$ and the r$_g$ theoretical regression equation is $a + r s_2 x/s_1$. For the bivariate normal X and $Y - r s_2 X/s_1$ are independent and $Y - r s_2 X/s_1$ is distributed as a $N(m_2 - r s_2 m_1/s_1, s^2_{res})$. This distribution is symmetric about its mean and hence, the average of the 1/3 and 2/3th quantiles will be the point of symmetry. In conclusion for the bivariate normal, the r$_g$ process recovers the correct regression equation. This is important because it is necessary to show that for the analogous sampling process the estimated regression line will in the limit converge to the correct equation.

## 5. SUMMARY OF MSSV AND r$_g$ METHOD

Before proceeding with the Cauchy example in which the more general features of the r$_g$ process become apparent, a summary is given which will set the stage for the other examples. The r$_g$ process for the slope $b$ involved determining $b$ so that two volumes were the same (the differences between two supremums). It was seen that equalizing the volumes on one side of a vertical line through $(m_1, m_2)$ ,say L, could be thought of as choosing $b$ so that the sum of squares of the volumes above and below the regression line was minimized. For the bivariate normal, for $0 \le t \le 1/2$ in the supremums, the comparison was on one side of L and for t>1/2 on the other side. By the symmetry of the distribution only one side needed to be considered, and then $r_g$ was a MSSV method. In other examples the joint distribution could have a symmetry property that allows the $r_g$ method to produce the true regression line, but it need not have the elliptical symmetry property like the normal about a line L.

In the limit, the population regression of least squares and $r_g$ are probably the same for many distributions, but in sampling estimation problems, volumes get replaced by relative frequency counts, and significant differences can occur. The overall minimization process remains the same for both least squares and $r_g$, but because a point can move to a distant outlier without changing the volume of a region, $r_g$ is a robust regression as opposed to least squares for which a distant outlier will destroy its good estimation properties. When this happens, the absolute value of $r_g$ can be large on $(X, Y - \hat{b}X)$ where $\hat{b}$ is the least squares estimate and Pearson's correlation coefficient is zero.

The $r_g$ value of the slope $b$ occurred for a $b$ for which

$$H(F^{-1}(t), G^{-1}(1-t)) = t - H(F^{-1}(t), G^{-1}(t)) \qquad \textbf{(4)}$$

was true for all $0 \le t \le 1.$ For this $b$, the distribution G is symmetric about $m_x = E(Y|X = x).$ For t< 1/2 let y>0 such that $m_x + y = G^{-1}(1-t)$ and $m_x - y = G^{-1}(t).$ Then the joint distribution H of (X,Y-$b$ X) where $E(Y|X = x) = a + bx = m_x$ satisfied $H(x, y + m_x) + H(x, m_x - y) = H(x, \infty),$ and dividing by the right-hand side the following condition on the conditional distribution is obtained. $H(x, y + m_x)/H(x,\infty) + H(x, m_x - y)/H(x,\infty) = 1.$ This says that the conditional distribution of$(Y - bX|X = x)$ is symmetric about $m_x$. Thus it appears that solving equation (2) will give the true $b$ if the distribution of (X,Y) has the symmetry property that is the basis of the definition of $r_g$, mainly that there exists a $b$ such that the distribution of $(X, Y - bX)$ ,with H, F, G, the cdf's, satisfies equation (5). This is formalized in the next section.

## 6. SLOPE PARAMETER AND $r_g$

In this section a general definition of a slope parameter for a class of bivariate symmetric distributions is given and then equation (2) for the $r_g$ regression method is shown to yield this parameter.

Definition 1. For certain classes of symmetric continuous bivariate distributions, let H(x,y) be the cumulative distribution function (cdf). For values of $b$, consider the family of distributions (X, Y- $b$ X). If a $b$ can be found such that the distribution of (X,Y - $b$ X) is elliptically symmetric about a line of zero slope, then the value of this $b$, say $b$ (H), is said to be the slope parameter, and H belongs to the class $\mathrm{H}$ of continuous regression symmetric distributions.

Note that this defintion gives a slope parameter for the bivariate regression models for both the Normal and Cauchy distributions; it agrees with the expectation definition for the normal case and extends the regression model definition to the Cauchy distribution where expectations do not exist. This defintion could be extended to include distributions that are not elliptically symmetric, but, instead, are only symmetric for each point on a horizontal line. This then would include nonhomogenous variance models and the classical regression model as explained below.

For the standard univariate regression model where x is assumed fixed, say, Y = $a + bx + e$ and $e$ has a symmetric distribution about zero, then for each fixed x, the distribution of $Y - bx = a + e$ is symmetric about $a$. Thus, for all x, the conditional distributions $(x, Y - bx)$ are symmetric about the zero slope line centered at $a$ and $b$ is the slope parameter even if the expectation of $e$ does not exist.

It is now shown for class $\mathrm{H}$ that the rg MSSV method gives the true slope parameter. Let H* be in $\mathrm{H}$ for the bivariate random variable (X,Y), and let F* and G* be the marginal cdf's of X and Y, respectively. Let H, F, and G be the corresponding cdf's for the transformed variables $(X, Y - b\,(H^*) * X)$, where G has a point of symmetry at its median and without loss of generality, let this point be zero. From the definition of $b(H^*)$ and $\mathrm{H}$, following properties hold

(a) $H(u,-v) = H(u,\infty) - H(u,v) \equiv F(u) - H(u,v)$,

(b) $G(v) = 1 - G(-v)$, which implies for 0<t<1, $G^{-1}(1-t) = -G^{-1}(t)$.

These properties (a) and (b) imply that equation (4) is satisfied at $b(H^*)$ which means that the greatest deviation correlation coefficient regression method has the correct population value for the slope. The proof is as follows: from (a) and (b) and for all 0<t<1,

$$H(F^{-1}(t), G^{-1}(1-t)) = H(F^{-1}(t), -G^{-1}(t))$$
$$= H(F^{-1}(t), \infty) - H(F^{-1}(t), G^{-1}(t))$$
$$= FF^{-1}(t) - H(F^{-1}(t), G^{-1}(t))$$
$$= t - H(F^{-1}(t), G^{-1}(t)).$$

Thus the supremums in equation (3) are identical at $b = b(H^*)$ and
$r_g(X, Y - b\,(H^*) * X) = 0.$

An application of this result would be that the $r_g$ method would give the correct slope parameter for all bivariate t distributions including the case of one degree of freedom, the bivariate Cauchy.

## 7. THE BIVARIATE CAUCHY

The following example with the bivariate Cauchy distribution shows that the $r_g$ method is applicable for all distributions; not just those with finite first and second moments. The recent book by Hutchinson and Lai (1992) states that the bivariate Cauchy "$\cdots$ is of limited interest as it has no correlation parameter"; it may be that since the usual conditional expectation $E(Y|X=x)$ does not exist, no one has known how to recover the regression line $y = rx$ which does exist if definition 1. is applied. Hutchinson and Lai define the bivariate t and this density with one degree of freedom should be what is defined to be the bivariate Cauchy with correlation parameter $r$ ;

$$h(x,y) = \left[1 + (x^2 - 2rxy + y^2)/(1 - r^2)\right]^{-3/2} \Big/ (2p\sqrt{1 - r^2}) \qquad (5)$$

They defined the bivariate Cauchy to be this density with $r$ =0, but it should be defined as in (5) and the following will show that the $r_g$ method recovers the correct regression line in exactly the same geometrical manner as the bivariate normal case.

If the contour lines for this density are drawn for various $r$ , they take the same elliptical pattern as do the contour lines for the standardized bivariate normal. For example, let u=x and $v = (y - rx)\big/\sqrt{1 - r^2}$ , then $h(u,v) = \left(1 + u^2 + v^2\right)^{-3/2}\big/2p$ , the bivariate Cauchy with $r$ =0, and the contours are circular and centered at the origin. The cdf of this standardized density is $H(u,v) = 1/4 + \left[\tan^{-1}u + \tan^{-1}v + \tan^{-1}\left(uv\big/\sqrt{1 + u^2 + v^2}\right)\right]\big/2p$ . The marginal or univariate Cauchy is given by $H(u,\infty) = H(\infty,u) = 1/2 + (\tan^{-1}u)/p$. Then the inverse of the univariate Cauchy is $F^{-1}(t) = u = \tan(p(t - 1/2))$, and using this it is relatively easy to show that the cdf of (U,V) satisfies the symmetry condition of $r_g$ in equation (4) and hence, $r_g(U,V) = 0.$ It seems to be clear that just like the bivariate normal if $r \neq 0,\ r_g(X,Y) \neq 0$ but $r_g(X,(Y - rX)\big/\sqrt{1 - r^2}) = 0$ and the $r_g$ method recovers the regression equation for the (generalized, $r \neq 0$) bivariate Cauchy distribution. This also means that the $r_g$ method can estimate the slope in a reasonable manner when sampling is done with the bivariate Cauchy density in equation (5) using all the data. Previously existing estimation techniques would be unstable as they depend on moments existing, and current robust methods would try to delete "outliers" to stabilize the process. Some simulations were run to verify this and indeed the $r_g$ method was up 10,000 times more efficient than least squares.

The intercept in this example is zero so no calulation for it is necessary. If, however, the bivariate Cauchy was shifted, the $r_g$-mean intercept method would recover the center of the distribution by averaging the one-third and two-thirds quantiles because of the symmetry of the Cauchy distribution.

In exactly the same manner as for the bivariate normal distribution, the value of $r_g(X,Y)$ can be obtained from equation (3) for any $r$ because the bivariate Cauchy has the same elliptical symmetry properties and the supremums are achieved at t = 1/2. Thus, $r_g(X,Y) = 2[P(Q_3) - P(Q_2)] = 2[P(Q_1) - P(Q_4)]]$. The probabilities $P(Q_1)$ and $P(Q_4)$ are obtained by integrating density (6) over the first and fourth quadrants. The double integral for $P(Q_1)$ is reduced to a single integral by changing to polar coordinates $(r,q)$ and integrating out the r to get $P(Q_1) = \left(\sqrt{1-r^2}/2p\right)\int_0^{p/2}(1-2\sin q\cos q)^{-1}\,dq$. The integral over the fourth quadrant, $P(Q_4)$, is the same except the limits of integration are from $-p/2$ to 0. The substitution of $\sin 2q$ for $2\sin q\cos q$ allows the use of a standard integration formula (#195 page 89 in Burington (1973)) involving the arctan function. In evaluating the result of the integration for $P(Q_4)$ at $-p/2$ care must be taken because the principal value is not the correct result; use $p - \tan^{-1}(\sqrt{(1+r)/(1-r)}\tan p/4)$. The final results with another identity are

$$P(Q_1) = p^{-1}\tan^{-1}\sqrt{(1+r)/(1-r)} = \sin^{-1} r/2p + 1/4,$$

$$P(Q_4) = (1/2) - p^{-1}\tan^{-1}\sqrt{(1+r)/(1-r)} = -\sin^{-1} r/2p + 1/4.$$

For the bivariate Cauchy distribution $r_g(X,Y) = 2(P(Q_1) - P(Q_4)) = (4/p)\tan^{-1}\sqrt{(1+r)/(1-r)} - 1 = (2/p)\sin^{-1} r$, and $r$ enters into the r$_g$ correlation of the bivariate Cauchy identical to the bivariate normal. It is also true that region A as defined for the bivariate normal has the same value and interpretation.

$$P(A) = \left(\sqrt{1-r^2}/2p\right)\int_0^{\tan^{-1} r}(1-2r\sin q\cos q)^{-1}dq = r_g(X,Y)/4,$$

and $(1 + r_g(X,Y))/4 = (1 + ((2\sin^{-1} r)/p))/\ 4.$

## 8. DISTRIBUTION FREE

Let (X,Y) be bivariate normal and let S have the distribution of the square root of an independent Chi-square random vaiable with $n$ degrees of freedom divided by its degrees of freedom. Then (X/S,Y/S) = (X*,Y*) is a bivariate t and if $n$=1, it is bivariate Cauchy. The division by S only changes the scale factor of (X,Y) and not the distribution of correlation r$_g$. Thus r$_g$(X, Y-$b$ X) = 0 implies r$_g$(X*, Y*-$b$ X*) = 0 and the r$_g$ method gives the same results for $b$ over the class (X,Y), (X*,Y*), $n$=1,2,$\cdots$,$\infty$. This means that the r$_g$ confidence interval for $b$ in Gideon et al. (1994) is valid for data from any of these bivariate distributions.

## 9. THE ASYMPTOTIC RELATIONSHIP BETWEEN THE DISTRIBUTIONS OF THE CORRELATION COEFFICIENTS AND THE SLOPE ESTIMATES

The limiting distribution of the null distribution of any correlation coefficient can be used to determine the limiting distribution of the corresponding estimate of the slope or slopes in linear regression. As in the earlier sections let $b$ represent the slope and $b(H)$ the regression parameter for joint distribution H. Also let r be a correlation coefficient. The idea is to expand $r(X,Y-bX)$ as a Taylor series in $b$ about $b(H)$, replace the random variable (X,Y) by the sample vectors $(\underline{x},\underline{y})$, and for large n , evaluate this equation at the r estimate of the slope, say at $\hat{b}$; then, multiple by the square root of n and use the limiting distribution of r and the Taylor series to relate the limiting distribution of the correlation coefficient to the slope estimate. The truncated Taylor series is

$$r(X,Y-bX) \cong r(X,Y-b(H)X) + \frac{d}{db} r(X,Y-Xb)\Big|_{b=b(H)}(b-b(H)) \qquad (6)$$

In order to show the validity of the method, the case where r is Pearson's correlation and (X,Y) is bivariate normal will be used to show that the standard result is obtained. The notation of section 3 is used.

The derivative in equation (6) is obtained by differentiating the quantity that is given just below equation (1). The result is $-s_1 s_2^2(1-r^2)/s_{res}^3$ and this evaluated at $b =$ $b(H) \equiv s_2 r/s_1$ gives $-s_1/(s_2\sqrt{1-r^2})$. For the Bivariate Normal distribution X and $Y-b(H)X$ are independent random variables so that for a random sample the asymptotic null distribution of Pearson's r is needed. By theorem 4.2.6 in Anderson (1958) , for large n, $r\cdot\sqrt{n}$ is approximately N(0,1). The usual least squares estimate of the slope parameter, $\hat{b}$, is obtained by replacing the random variables X and $Y-bX$ by data vectors in the lefthand side of equation (6) , setting the result equal to 0, and solving for $b$. Equation (6) holds asymptotically when data vectors replace random variables and $b$ is evaluate at $\hat{b}$, and the asymptotic connection between the null distribution of r and $\hat{b}$ is obtained by multiplying the result by $\sqrt{n}$. For simplicity n rather than n-3 is used. The result is that $0 \cong \sqrt{n}\,r(x,y-x\,s_2 r/s_1) - \sqrt{n}\,s_1(\hat{b}-s_2 r/s_1)/(s_2\sqrt{1-r^2})$. Because the first term has an approximate N(0,1) distribution, so does the second term. This result can be written as $\hat{b}$ has an approximate $N(b(H), \dfrac{s_2^2(1-r^2)}{ns_1^2})$ distribution.

Note that in the classical simple linear regression case with x fixed and $s^2$ the residual variance, $s^2 = s_2^2(1-r^2)$ and $ns_1^2 \cong \sum(x_i-\bar{x})^2$. With these substitutions $\hat{b}$ has an approximate $N(b, \dfrac{s^2}{\sum(x_i-\bar{x})^2})$ distribution; that is, the standard result.

This technique can be used to connect the limiting null distribution of any correlation coefficient to the limiting distribution of the slope estimator. As shown above, the quantities necessary to complete the calculation are the population correlation coefficient of X with Y-$b$ X, its derivative with respect to $b$, and the asymptotic distribution øf the correlation coefficient for uncorrelated data; that is, a random sample of (X, Y-$b$ X). The correlation coefficient $r_g$ has the same population value for the Cauchy and Normal distributions, and so the only difference in the asymptotic distribution of the slope estimator would be if there were a difference in the limiting distribution of $r_g$ on uncorrelated data.

Since the limiting distributions of Spearman's and Kendall's correlation coefficients are known, this section shows a simple way to determine the asymptotic distributions of the corresponding slope estimators in a simple linear regression. Now, however, the above technique is used to compute the limiting distribution of the $r_g$ estimate of the slope for the Normal distribution and then to compare the asymptotic standard deviation to the classical case. It is done only for the standarized bivariate Normal. For this case, the derivative ( see just before equation (2)) evaluated at $b$ (H) is $-2/(p\sqrt{1-r^2})$. Since in Gideon et al. (1989), $\sqrt{n}r_g$ (null case) goes to a N(0, 1) distribution, the final result is that the $r_g$ estimate of the slope, $\hat{b}_g$, has an approximate (here, $b$ (H) = $r$ ) $N(r,p^2(1-r^2)/4n)$ distribution. Thus, the ratio of the asymptotic standard deviation between the r and $r_g$ slope estimators is

$$\frac{\dfrac{p\sqrt{1-r^2}}{2\sqrt{n}}}{\dfrac{\sqrt{1-r^2}}{\sqrt{n}}} = \frac{p}{2} = 1.57.$$

Since Kendall's $t$ has the same value of the population correlation coefficient on the Normal distribution as $r_g$ does, and it is known that $t\sqrt{n}$, null case, is approximately $N(0,4/9)$ distributed, by the method above, the $t$ estimate of the slope, $\hat{b}_t$, has a asymptotic variance of 4/9 of $p^2(1-r^2)/4n$. Hence, the ratio of the asymptotic standard deviations of $t$ to r for the Normal distribution is $\sqrt{4/9} = 2/3$ of $p/2$ or about 1.047. This result , of course, has been obtained by other means, but for $r_g$ there are no other ways to obtain asymptotic results. This shows the generality of the method. In addition, the general ideas of this section can be easily applied to all correlation coefficients to connect them to regression results.

Two points should be made clear. One is that, because $t$ and $r_g$ have the same population values for the Normal distribution, the ratio of their asymptotic standard deviations in estimating the slope depends only on the asymptotic null distributions of these correlation coefficients. The second point can be illustrated with Spearman's correlation coefficient, say, $r_S$.

Because it has a different population value on the Normal distribution than $t$ and $r_g$ does, its derivative term in equation (6) needs to be computed. Then using that result with the asymptotic null distribution of $r_S$, its limiting distribution for its slope estimator can be obtained. Note that this asymptotic distribution of the slope estimate is obtained without an explicit expression for this estimator of the slope.

## 10. CONCLUSION

It has been shown that the Greatest Deviation correlation coefficient can be used to determine the true slope $b$ in a simple linear regression by making random variables X and Y - $b$ X $r_g$-uncorrelated as in equation (2). This result is apparently true for all distributions (X, Y- $b$ X) that have the symmetry property (4) which is the basis for the definition of $r_g$ in equation (3). Since the bivariate Cauchy distribution has this symmetry property, $r_g$ even works for distributions without any finite moments. The relationship of the correlation parameter $r$ of both the bivariate Cauchy and Normal distributions are related in the same manner to $r_g$; that is, $r = \sin(pr_g/2)$. Now $r$ is interpretable as the cosine of the angle between X and Y and $r = \cos(q) = \sin(.5p - q) = \sin(.5pr_g)$ implies that

$$q = .5p(1-r_g) \text{ or } r_g = 1-2q/p \qquad (7)$$

Rummel (1991) first noticed this result. Thus, $r_g$ is linearly related to the angle between X and Y. From Figure 1 it can be seen that (1-$r_g$)/8 is a statistical distance between X and Y and a perfect postive relationship; that is, (1-$r_g$)/8 is the volume in the first and fourth quadrants between the lines y=x and y= $r$ x. Now 4$p$ (1-$r_g$)/8 = .5$p$ (1-$r_g$)=$q$ so that multiplication by 4$p$ converts this statistical distance to the angle between X and Y. Hence, $r_g$ gives a simple connection between the angle between two jointly distributed random variables and a statistical distance in a linear fashion. Note that if $q = \left(0, p/4, p/2, 3p/4, p\right)$, $r$ =(1, .7071, 0, -.7071, -1), $r_g$=(1, .5, 0, -.5, -1), and (1-$r_g$)/8 is (0, 1/16, 1/8, 3/16, 1/4). In the sampling estimation of $b$, the calculation of both the $r_g$ estimate of $b$ and the value of (1-$r_g$)/8 on $\left(X, Y - \hat{b}X\right)$ where $\hat{b}$ is the classical least squares estimate gives valuable insight into the quality of the data with respect to elliptical symmetry and outliers.

The results of Section 9 allow a simple and general asymptotic connection between the null distribution of a correlation coefficient and its corresponding slope estimator.

Important equivariant properties are given in Rousseeuw and Leroy (1987) for regression techniques. Of these, it is easily seen that all correlation coefficient regression techniques are regression and scale equivariant because of standard properties of correlation coefficients. However, $r_g$ is not affine equivariant because it is a nonlinear technique.

BIBLIOGRAPHY

Anderson, T.W. (1958), An Introduction to Multivariate Statistical Analysis, Wiley and Sons, New York.

Burington, R.S.(1973). The Handbook of Mathematical Tables and Formulas, 5th ed. McGraw-Hill Book Co.

Gideon, R.A. and Hollister, R.A. (1987), A Rank Correlation Coefficient Resistant to Outliers, Journal of The American Statistical Association, Vol. 94, 14,656-666.

Gideon, R.A., Rummel, S. A., and Li, H. (1994), Nonparametric Correlation and Regression Methods, submitted to the Journal of Computational and Graphical Statistics, Spring 1994.

Gideon, R.A., Li, H., Rummel, S.A., Bruder, J., Lee, L.C., and Thiel, M. (1992). Robust Location and Scale Estimation with Nonparametric Correlation Coefficients, unpublished.

Gideon, R.A., Rummel, S., and Li, Hongzhe (1993), Multiple Linear Regression Using Nonparametric Correlation Coefficient $r_g$, To be submitted to Statistics when ready and if all goes well.

Hutchinson, T.P. and Lai, C.D. (1992). Continuous Bivariate Distributions, Emphasising Applications. Rumsby Scientific Pub., Adelaide, Australia 5000.

Gideon,RA.,Prentice,M.J.,and Pyke,R.(1989). The Limiting Distribution of the Rank Correlation Coefficient $r_g$, Appears 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-Verlang ,N.Y. pp 217-226.

Ross, Sheldon(1988). A First Course in Probability, 3rd ed. , Macmillan Publ. Co. Page 292.

Rousseeuw, P.J. and Leroy, A.M. (1987), Robust Regression and Outlier Detection, Wiley and Sons, New York.

Rummel, S.E. (1991). A Procedure for Obtaining a Robust Regression Employing the Greatest Deviation Correlation Coefficient. Unpublished Ph. D. Dissertation, University of Montana.