Elementary Slopes in Simple Linear Regression

Rudy Gideon
University of Montana
Missoula, MT 59812

and

Adele Marie Rothan, CSJ
College of St. Catherine
St. Paul, MN 55105

In a bivariate data plot, every two points determine an “elementary slope.” For n points with distinct x-values, there are $n(n – 1)/2$ elementary slopes. These elementary slopes are examined under the two classical regression assumptions: (1) the regressor variable values are fixed and the error is independent and normal, and (2) the data is bivariate normal. For case (1), it is demonstrated that a weighted average of the elementary slopes gives the standard least squares estimate. In case (2), it is shown that the elementary slopes have a rescaled Cauchy distribution; this Cauchy distribution is then used to estimate bivariate normal parameters. Two nonparametric correlation coefficients, Kendall’s t and the Greatest Deviation correlation coefficient (*GD*), are used with elementary slopes in regression estimation. Simulations show the robustness of the nonparametric method of estimation using Kendall’s t and *GD*.

Keywords: bivariate normal, Cauchy distribution, Kendall’s t , Greatest Deviation correlation coefficient

This work depends in part on earlier unpublished work of Gideon and is available on his web site: www.math.umt.edu/gideon. Some of the references will refer to papers posted at this web site.

1. Simple Linear Regression with fixed regressor variable data

Let the regression equation model be $y = a + bx + e$, errors independent with $V(e) = s^2$ and $E(e) = 0$. Let $\{(x_i, y_i) \mid i = 1, 2, 3, \ldots, n\}$ be the data with distinct x-values. Then the set of $n(n$-1)/2 elementary slopes are $\left\{\frac{y_j - y_i}{x_j - x_i}\right\}, i \neq j$. According to the model,

$$\frac{y_j - y_i}{x_j - x_i} = \frac{(a + bx_j + e_j) - (a + bx_i + e_i)}{x_j - x_i} = b + \frac{e_j - e_i}{x_j - x_i}.$$

Because $E(e_j - e_i) = 0$, each slope is unbiased for b . Also $V\left(\frac{Y_j - Y_i}{x_j - x_i}\right) = \frac{V(e_j - e_i)}{(x_j - x_i)^2} = \frac{2s^2}{(x_j - x_i)^2}$. In *U*-Statistics methods (Randles and Wolfe 1979), it would be suggested that the $\binom{n}{2}$ elementary slopes be averaged to obtain an unbiased estimate. However, this is slightly changed here by taking a weighted average with the weights being the reciprocals of the variances of the elementary slopes.

Lemma 1: The weighted average of the elementary slopes gives the usual least squares estimate of the slope.

Proof: Let A be the $\begin{pmatrix} n \\ 2 \end{pmatrix}$ set of indices (*i*, *j*), $i \neq j$. Then the sum of the weights *W* is given by: $W = \frac{1}{2s^{2}} \sum_{(i,j)eA} (x_j - x_i)^2 = \frac{n}{2s^{2}} \sum_{i=1}^{n} (x_i - \bar{x})^2$. (1.1)

Result (1.1) is related to one-sample *U*-statistics (Randles and Wolfe 1979, pp. 61-63) and follows from the demonstration of the equality (1.2), shown below, when *y* is taken as *x*.

For distinct *x*-values and $i \neq j$, the weighted-average estimate is

$$\hat{b} = \frac{1}{W} \sum_{(i,j)eA} \frac{y_j - y_i}{x_j - x_i} \frac{(x_j - x_i)^2}{2s^{2}}$$

$$= \frac{1}{W2s^{2}} \sum_{(i,j)eA} (y_j - y_i)(x_j - x_i) = \frac{1}{W} \frac{n}{2s^{2}} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) . \qquad (1.2)$$

Finally, substituting for *W* gives:

$$\hat{b} = \frac{2s^{2}}{n \sum_{i=1}^{n} (x \ - \bar{x})^2} \frac{n}{2s^{2}} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^{2`}} . \qquad \blacklozenge$$

To demonstrate equality (1.2), an example is shown for *n* =4. Then the set *A* has 6 points; $A = \{(1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 4)\}$.

$$\sum_{(i,j)eA} (y_j - y_i)(x_j - x_i) = 3\sum_{i=1}^{4} x_i y_i - \sum_{i=1}^{3} \sum_{j=i+1}^{4} x_i y_j - \sum_{i=1}^{3} \sum_{j=i+1}^{4} y_i x_j$$

Add and subtract $\sum_{i=1}^{4} x_i y_i$ so that

$$\sum_{(i,j)eA} (y_j - y_i)(x_j - x_i) = 4\sum_{i=1}^{4} x_i y_i - \sum_{i=1}^{3} \sum_{j=i+1}^{4} x_i y_j - \sum_{i=1}^{3} \sum_{j=i+1}^{4} y_i x_j - \sum_{i=1}^{4} x_i y_i$$

$$= 4\sum_{i=1}^{4} x_i y_i - \left(\sum_{i=1}^{3} \sum_{j=i+1}^{4} x_i y_j + \sum_{i=1}^{3} \sum_{j=i+1}^{4} y_i x_j + \sum_{i=1}^{4} x_i y_i\right) \qquad (1.3)$$

Note that $(\sum_{i=1}^{4} x_i)(\sum_{j=1}^{4} y_j) = \sum_{i=1}^{4} x_i y_i + \sum_{i=1}^{3} \sum_{j=i+1}^{4} x_i y_j + \sum_{i=1}^{3} \sum_{j=i+1}^{4} y_i x_j$

So (1.3) becomes

$$\sum_{(i,j)eA} (y_j - y_i)(x_j - x_i) = 4\sum_{i=1}^{4} x_i y_i - (\sum_{i=1}^{4} x_i)(\sum_{j=1}^{4} y_j)$$

$$= \left[4\sum_{i=1}^{4} x_i y_i - (\sum_{i=1}^{4} x_i)(\sum_{j=1}^{4} y_j) \right] \frac{4}{4}$$

$$= 4\left[\sum_{i=1}^{4} x_i y_i - \frac{1}{4}\sum_{i=1}^{4} x_i \sum_{j=1}^{4} y_j\right]$$

Recognizing $[\bullet]$, above, as the computational form of $\sum_{i=1}^{4}(x_i - \bar{x})\,(y_i - \bar{y})$, yields

$$\sum_{(i,j)eA}(y_j - y_i)(x_j - x_i) = 4\sum_{i=1}^{4}(x_i - \bar{x})(y_i - \bar{y}).$$

Thus $\frac{1}{W2s^2}\sum_{(i,j)eA}(y_j - y_i)(x_j - x_i) = \frac{1}{W}\frac{4}{2s^2}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$.

Or for general positive integer *n*,

$$\frac{1}{W2s^2}\sum_{(i,j)eA}(y_j - y_i)(x_j - x_i) = \frac{1}{W}\frac{n}{2s^2}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

If not all x_i are distinct, the formulas and equations will hold if when $x_i = x_j$, $\frac{(x_j - x_i)^2}{x_j - x_i}$ is interpreted as zero since its limit as $x_j - x_i \to 0$ is zero. These terms must appear in the summations.

By extending the idea of a *U*-Statistic to a weighted average, it has been shown that the classical least squares estimate of slope is a weighted average of elementary slopes.

2. Simple Regression with Bivariate Normal Data

In this section, the data (X, Y) have a bivariate normal distribution. Again the elementary slopes are analyzed, but now both the numerator and denominator are random variables. It is shown that the elementary slopes for a bivariate normal distribution have a Cauchy distribution. It is then shown how to use the Cauchy distributed elementary slopes to estimate the regression parameters for the bivariate normal.

The equal in distribution notation $\overset{d}{=}$ defined in Randles and Wolfe (1979, p.13) is used.

Lemma 2: The elementary Slopes for a Bivariate Normal Distribution have a rescaled Cauchy Distribution.

Proof: First, let (X, Y) have a standardized bivariate normal distribution with correlation coefficient r. Then for two independent observations (X_1, Y_1) and (X_2, Y_2), let $U = Y_1 - Y_2$ and $V = X_1 - X_2$ so that *U*/*V* = *R* (*R* for ratio) is the elementary slope. The random variable (U, V) has a bivariate normal distribution with means 0, variances 2, and correlation coefficient r. In order to obtain the joint distribution of (*R*, *S*), let *R*=*U*/*V* and *S*=*U*. Obtain the joint distribution of (*R*, *S*) and integrate out *S* to obtain that the distribution of *R* is Cauchy with location parameter r and scale parameter $\sqrt{1-r^2}$. The density for *R* is

$$f(r)=\frac{1}{p\sqrt{1-r^2}\left[1+\frac{(r-r)^2}{1-r^2}\right]},\quad -\infty<r<\infty .$$

Let R_1 denote this Cauchy with location parameter r and scale parameter $\sqrt{1-r^2}$.

Then the distribution of $R_0=\frac{R_1-r}{\sqrt{1-r^2}}$ is the standard Cauchy. For the general bivariate normal, let $X\overset{d}{=}N(m_1,s_1^2)$, $Y\overset{d}{=}N(m_2,s_2^2)$ with correlation coefficient r. It follows that

$$\frac{(Y_1-Y_2)\big/\sqrt{2}s_2}{(X_1-X_2)\big/\sqrt{2}s_1}\overset{d}{=}R_1$$

and that the elementary slope $\frac{Y_1-Y_2}{X_1-X_2}\overset{d}{=}\frac{s_2}{s_1}R_1$. Lastly, a sample of elementary slopes is related to a standard Cauchy by

$$\frac{Y_1-Y_2}{X_1-X_2}\overset{d}{=}r\frac{s_2}{s_1}+\frac{s_2}{s_1}\sqrt{1-r^2}R_0 \qquad (2.1)\ \blacklozenge$$

In the case where (X,Y) has a bivariate normal distribution, the regression model is

$$Y=m_y+r\frac{s_2}{s_1}\left(x-m_x\right). \qquad (2.2)$$

It is now easy to estimate r and the slope parameter, $r\frac{s_2}{s_1}$, by nonparametric methods. The slope parameter is estimated as an intercept in a regression that uses the elementary slopes as the dependent variable; see equation (2.1). The correlation coefficient is estimated using both the slope and intercept in this regression.

For a complete development of the work that follows, the reader is referred to problem 1.2.14 in Randles and Wolfe (1979, p. 12) and papers 1 through 7 at the web site. A synopsis of necessary material follows.

Let $r_p(x,y)$ be the notation for the calculation of Pearson's correlation coefficient on a set of data (x,y). Let *GD* be the Greatest Deviation correlation coefficient (Gideon and Hollister 1987) and $GD(x,y)$ its value on a set of data. For simple linear regression, the least squares estimate of slope is obtained by solving for *b* in the equation $r_p(x,y-bx)=0$; that is, by making the correlation between the independent variable and the uncentered residuals zero. The *GD* slope estimate is similarly obtained by solving $GD(x, y - bx) = 0$. In fact any correlation coefficient can be used in this manner as explained in Gideon (1992) and Gideon and Rummel (1992). That is, for any correlation coefficient *r*, solve for *b* in

$$r(x, y - bx) = 0 . \tag{2.3}$$

This same type of correlation coefficient equation is used for location and scale estimation in (2.1) (Gideon and Rothan 2004). The form of the equation remains the same; only the arguments change. In this correlation method of estimation, scale must be estimated first and then location. Whereas the original sample size is *n*, the sample size of the elementary slopes is $\binom{n}{2}$. Let *m* = $\binom{n}{2}$ and let *q* be the ordered quantiles corresponding to equally spaced probabilities from the assumed distribution: the integers 1 through *m* each divided by *m* + 1. Here *q* comes from the standard Cauchy, R_0, distribution. These quantiles are paired with ordered sample data.

Let vector *v* be the ordered set of elementary slopes $\left\{\frac{y_j - y_i}{x_j - x_i}\right\}, i \neq j$. The sample size *m*, defining *q* above, is $n(n-1)/2$. Then an estimate of scale using r_p is found by solving for *s* in $r_p(q, v - sq) = 0$, where v equals the vector of ordered slopes. The location estimate comes from taking the mean of $v - sq$. Classical methods are not valid for the Cauchy distribution; the method used here is similar to that in Randles and Wolfe (1979, problem 1.2.14, p. 12) except that the work is defined through correlation coefficients and uses the ordered data. For a robust and valid estimate of scale based on *GD*, solve for *s* in $GD(q, v - sq) = 0$. The location estimate comes from taking the median of $v - sq$. The general scale equation for any correlation coefficient *r* using ordered quantiles corresponding to the ordered data is

$$r(q, ordered(data) - sq) = 0 \tag{2.4}$$

The same numerical routines used in (2.3) suffice to solve for *s* in equation (2.4).

The estimate of the slope parameter, $\frac{s_2}{s_1}\sqrt{1 - r^2}$, in (2.1) is the solution *s* in (2.4).

The estimate of the location parameter, $r\frac{s_2}{s_1}$, in (2.1) is the median of $v - sq$; call this estimate *c*. Alternatively stated, the estimate of the regression slope parameter $r\frac{s_2}{s_1}$ in (2.2) for the original data comes from the intercept estimate *c* in (2.1) where the elementary slopes are the dependent variable. Since *s* estimates $\frac{s_2}{s_1}\sqrt{1-r^2}$ and *c* estimates $r\frac{s_2}{s_1}$, the ratio *u* of *s* to *c* estimates $\frac{\sqrt{1-r^2}}{r}$. The equation, $\frac{\sqrt{1-r^2}}{r}$ = *u*, is now solved for r ; so the estimate of r is $\frac{1}{\sqrt{1+u^2}}sign(c)$. By dividing *c* by this estimate of r , one has an estimate of the ratio of the standard deviations ${}^{s_2}\!/\!{}_{s_1}$.

3. Regression Estimation using Kendall's t and Elementary Slopes

The basic regression estimation equation for Kendall's t is equation (2.3) with t replacing *r,* $t\,(x, y-bx)=0$, and solve for *b*. For fixed *b* the concordances and discordances are counted from the signs for all pairs of indices (*i, j*), *i*< *j*, of $\frac{y_j-bx_j-(y_i-bx_i)}{x_j-x_i} = \frac{y_j-y_i}{x_j-x_i}-b.$ For $t\,(x, y-bx)=0$, the number of concordant pairs must equal the number of discordant pairs and this implies *b* is determined so that $\#\left(\frac{y_j-y_i}{x_j-x_i}-b<0\right)= \#\left(\frac{y_j-y_i}{x_j-x_i}-b>0\right)$. This occurs when *b* is the median of the elementary slopes $\left\{\frac{y_j-y_i}{x_j-x_i}\right\}$. If a plot is made of *b* against $t\,(x, y-bx)$, it is a monotonic decreasing step function that only decreases at each elementary slope. The function $GD(x, y-bx)$ behaves in a similar manner except it decreases at some but not all of the elementary slopes thus assuming fewer values.

4. Illustration by Simulation of the Estimation in Regression by Correlation Coefficients and Elementary Slopes

In order to estimate $a = r\frac{s_2}{s_1}$ and $b = \frac{s_2}{s_1}\sqrt{1-r^2}$ in (2.1), the elementary slopes $\left\{\frac{y_j-y_i}{x_j-x_i}\right\}$ are the data to be ordered and regressed in equation (2.4) against the Cauchy quantiles, *q*. The elementary slopes for this data number $\binom{\binom{n}{2}}{2}$ where *n* is the original sample size. For $n=10, 30, 100$, the number of elementary slopes is 990, 94,395 and 12,248,775, respectively. The computer language *C* routine used in solving equation (2.1) or (2.4) for *GD* uses these elementary slopes as the data. The routine does a systematic search on these elementary slopes; and so for *n* greater than 30 or so, it is overwhelmed. However, a feature of equation (2.4) allows good estimation when the ordered data is truncated at each end.

Computer simulations demonstrate the practicality of these methods. Two types of data were considered: (1) bivariate normal (BN) and (2) bivariate normal with some outlier contamination in the *Y* variable to demonstrate the robustness of *GD* and Kendall's t methods. This outlier data was limited to a random amount; namely, the number of outliers in each sample was binomial with *n* equal to the sample size and probability of 0.20 for outliers. The outliers were generated from a N(0, 5) distribution. For each case (BN and BN with Outliers), sample sizes of *n* = 20 and *n* = 100 were used and 1000 simulations were run.

Estimates of the slope parameter in the BN, $r\frac{s_2}{s_1}$, were obtained using three methods: (1) the usual least squares or, in correlation language, Pearson's *r* estimate, (2) Kendall's t method, and (3) the *GD* method on the elementary slopes. Estimates of the correlation parameter r were obtained by: (1) the usual Pearson's *r*, (2) the *GD* regression using the elementary slopes and taking the ratio of the slope to the intercept and then using *u* as detailed in Section 2, and (3) the greatest deviation using the normal transformation $\sin(pGD/2)$, see Gideon and Hollister (1987). The means and standard deviations for the estimators in the simulations are given in Tables I – VIII below.

Table IX displays a summary of the *GD* estimate of the standard deviation of the original *x*-data found by using equation (2.4) and normal quantiles.

Following the tables, a short discussion of the principle results is given.

Table I	Estimation of Slope, no outliers, n=20 and 1000 simulations each				
parameter → Method ↓		0	1.0	1.5	1.8
t	mean	-0.0036	1.0163	1.5057	1.7926
	SD	0.5224	0.4625	0.3468	0.2388
LS or P	mean	-0.0051	1.0128	1.4997	1.7921
	SD	0.4746	0.4129	0.3191	0.2157
GD with the elem slopes	mean	-0.0038	1.0193	1.5061	1.7949
	SD	0.5526	0.4824	0.3705	0.2568

Table II	Estimation of Slope, with outliers, n=20 and 1000 simulations each				
parameter → Method ↓		0	1.0	1.5	1.8
t	mean	0.06197	0.9819	1.4942	1.7827
	SD	0.7969	0.6835	0.5039	0.3271
LS or P	mean	0.06221	0.9678	1.5209	1.7654
	SD	1.1451	1.0427	0.7406	0.5037
GD with the elem slopes	mean	0.06442	0.9624	1.5061	1.7779
	SD	1.1026	0.9406	0.6758	0.4441

Table III	Estimation of Correlation r , no outliers, n=20 and 1000 simulations				
parameter → Method ↓		0	0.5	0.75	0.90
LS or P	mean	-0.0015	0.4910	0.7387	0.8940
	SD	0.2264	0.1666	0.1115	0.0487
GD with	mean	-0.0150	0.4856	0.7309	0.8883
elem Slopes	SD	0.2593	0.1826	0.1283	0.0579
GD with the sine transf	mean	0.0002	0.4284	0.6586	0.8125
	SD	0.2704	0.2283	0.1694	0.1129

Table IV	Estimation of Correlation r , with outliers n=20, and 1000 simulations				
parameter → Method ↓		0	0.5	0.75	0.90
LS or P	mean	0.0116	0.2553	0.4686	0.6692
	SD	0.2269	0.2394	0.2090	0.1798
GD with	mean	-0.0124	0.3518	0.5556	0.7579
elem Slopes	SD	0.2807	0.1977	0.1950	0.1441
GD with the sine transf	mean	0.0222	0.3496	0.5725	0.7237
	SD	0.2745	0.2397	0.1873	0.1486

The standard deviation parameter of the *X* variable was "3" and did not change over the simulations. Only the *Y* variable was contaminated by outliers as described earlier. For sample size 20, the sample mean of the 1000 simulations of the *GD* estimate of the standard deviation parameter of *X* was usually about 3.09 with a sample standard deviation of 0.64. There apparently is a slight upward bias. Stated another way, the estimate ± two standard errors is 3.09 ± 0.04.

Table V Estimation of Slope, no outliers n=100, and 1000 simulations each

parameter → Method ↓		0	1.0	1.5	1.8
t	mean	0.0003	1.0108	1.5032	1.8012
	SD	0.2149	0.1860	0.1428	0.0929
LS or P	mean	0.0016	1.0114	1.5008	1.7994
	SD	0.2051	0.1783	0.1369	0.0877
GD with the elem slopes	mean	0.0004	1.0106	1.5032	1.8011
	SD	0.2149	0.1858	0.1428	0.0929

Table VI Estimation of Slope, with outliers n=100, and 1000 simulations each

parameter → Method ↓		0	1.0	1.5	1.8
t	mean	-0.0023	0.9865	1.4986	1.8007
	SD	0.2889	0.2494	0.1905	0.1307
LS or P	mean	-0.0054	0.9697	1.4945	1.7986
	SD	0.4798	0.4109	0.3108	0.2115
GD with the elem slopes	mean	-0.0023	0.9862	1.4985	1.8007
	SD	0.2888	0.2493	0.1904	0.1307

Table VII Estimation of Correlation *r* , no outliers n=100, and 1000 simulations

parameter → Method ↓		0	0.5	0.75	0.90
LS or P	mean	0.0010	0.5019	0.7458	0.8992
	SD	0.1011	0.0769	0.0448	0.0188
GD with	mean	0.0043	0.5049	0.7466	0.8999
elem Slopes	SD	0.1068	0.0904	0.0623	0.0276
GD with the sine transf	mean	0.0033	0.4780	0.7181	0.8744
	SD	0.1327	0.1077	0.0736	0.0405

Table VIII Estimation of Correlation r with outliers, n=100 and 1000 simulations

parameter → Method ↓		0	0.5	0.75	0.90
LS or P	mean	-0.0010	0.2318	0.4346	0.6554
	SD	0.0990	0.0984	0.0960	0.0808
GD with	mean	0.0040	0.3886	0.6382	0.8318
elem Slopes	SD	0.1058	0.1024	0.0827	0.0510
GD with the sine transf	mean	0.0016	0.4014	0.6247	0.7909
	SD	0.1296	0.1113	0.0878	0.0593

For these runs of sample size 100, the sample mean of the 1000 simulations of the *GD* estimate of the standard deviation parameter of *X* was usually about 3.02 with a sample standard deviation of 0.266. Thus the increased sample size reduced the bias. Here the estimate ± two standard errors is 3.02 ± 0.02.

A general summary of the results now follows. For both outlier and no outlier data, two sample sizes are used, 20 and 100. For the sample size 100, the *GD* method on the elementary slopes (*GD*-ES) used the middle 100 elementary slopes for the regression in equation (2.4). These middle 100 elementary slopes (50 on each side of the median of the elementary slopes) are paired with the middle 100 Cauchy quantiles. It would seem that this data reduction would be detrimental to the estimation process, but the results seem little influenced by this reduction.

From Tables I and V, no outliers, all methods are nearly unbiased for the slope parameter. The *GD*-ES and t methods are very comparable and only marginally less effective than the least squares method in terms of slightly larger variation.

From Tables II and VI, with outliers, for n = 20, the best method for the slope is t , then comes *GD*-ES, and finally LS. For n = 100, *GD*-ES and t are nearly the same and quite superior to LS.

From Tables III and VII, no outliers, in the estimation of r , the *GD*-ES method appears unbiased for all correlations. *GD*-sine is unbiased near 0, but underestimates for larger r . LS and *GD*-ES are fairly comparable with respect to variation.

From Tables IV and VIII, with outliers, in the estimation of r , *GD*-ES has substantially less bias for all r and for $r > 0$ less variation and is preferable to Pearson's *r*. Except near 0, *GD*-sine is also preferable over the classical method.

From Table IX the estimation of the standard deviation of *X* by the *GD* method via equation (2.4) is seen as slightly biased. Some results are taken from Fraser (1976) to aid in a comparison. Let *s* be the usual classical standard deviation and s_{gd} be the *GD* estimate. Then we use

$E(s) \cong s\ (1+\frac{1}{4(n-1)})^{-1}$ and $V(s) = E(s^2) - (E(s))^2$ to construct the following table.

Table IX	Comparison of the Standard Deviation of X	
s =3, n=20	E(s) = 2.961 E(s_{gd}) $\cong$ 3.09	SD(s) = 0.4821 SD(s_{gd}) $\cong$ 0.64
s =3, n=100	E(s) = 2.9924 E(s_{gd}) $\cong$ 3.02	SD(s) = 0.2134 SD(s_{gd}) $\cong$ 0.266

The values for *s* are from Fraser (1976). The values for s_{gd} are estimates from the 1000 simulations.

5. A Further Note on Estimation with Fixed Regressor Variable Data

Equation (2.3) can be used to estimate the slope and intercept in a simple linear regression with any correlation coefficient. This is detailed in the simple linear regression papers (Gideon 1992, Gideon and Rummel 1992). Simple linear regression has been carried out many times using *GD*. As in classical simple linear regression, the residuals can be computed. However, the regression standard error is computed by using equation (2.4) with *q* being normal quantiles and the ordered data the ordered residuals. The slope of the line is the standard error of the regression fit. In the same manner as normal theory or classical methods, the estimates of residual standard error and the standard deviation of *y* can be used to compute a "regression correlation coefficient,"
$\hat{r} = \sqrt{1-\frac{s_{y|x}^2}{s_y^2}}$. The "*s*" statistics all come from *GD* methods. This method is detailed in Gideon and Miller (1992).

If the error is Cauchy in the fixed-x model, the *GD* method using Cauchy quantiles will give regression parameter estimates as well as location and scale estimates of all the involved parameters. In addition, estimation of parameters with the bivariate Cauchy can also be carried out. In these situations, of course, classical methods cannot be used.

6. A Final Summary

This paper illustrates some interesting facts about the use of elementary slopes in simple linear regression. A rank-based correlation estimator, *GD*, was used on the elementary slopes to demonstrate that the Cauchy distribution can be profitability used in estimation. With outliers distributed symmetrically, both Kendall's t and *GD* operating on the original data and *GD* operating on the elementary slopes of the bivariate data are shown to be robust.

It appears that researchers in estimation have previously overlooked the Cauchy distribution, but this paper has shown that the Cauchy distribution is essential in the estimation procedure in bivariate normal situations.

This work is an extension of an entire system of estimation based on any correlation coefficient and, in particular, on nonparametric correlation coefficients. Several papers exploring this system are available on the web site. It is hoped that more of the web site material will be published and the web site material enlarged to show more of the methodology of the correlation coefficient system of estimation. The main computer programs are written in *C* code and are interfaced with the *S-Plus* statistical package.

References

Fraser, D. A. S. (1976), *Probability & Statistics, Theory and Applications*, problem 12, page 406, North Scituate, MA: Duxbury Press.

Gideon, R. A. (1992), “Random Variables, Regression, and the *GD*,” unpublished paper (URL: http://www.math.umt.edu/gideon/SLRtheory.pdf), University of Montana, Dept. of Mathematical Sciences.

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

Gideon, R. A., and Miller, J. M. (1995), “Multiple Regression Technique with Asymptotics,” unpublished paper (URL: http://www.math.umt.edu/gideon/C5-Miller-asymptotics.pdf), University of Montana, Dept. of Mathematical Sciences.

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

Gideon, R. A., and Rummel, S. E. (1992), “Correlation in Simple Linear Regression,” unpublished paper (URL: http://www.math.umt.edu/gideon/CORR-N-SPACE-REG.pdf), University of Montana, Dept. of Mathematical Sciences.

Randles, RH and Wolfe, D.A. (1979), *Introduction to the Theory of Nonparametric Statistics*, New York: John Wiley and Sons.

R. A. Gideon: Acknowledgments

This work has been in progress for many years with very few published papers available to acknowledge all the faculty and student help. These people are also listed at the web site, and I hope no one has been missed. I thank these people for all the help they have given me. They are the ones who have kept this research alive.

Appendix: An Example of the Process Using Bivariate Normal Data with Some Contamination in the *Y* variable.

In this example, a sample of size 20 was generated from a bivariate normal distribution with parameters $m_x = 5$, $s_x = 3$, $m_y = 6$, $s_y = 6$, and $r = 0.6$ with the *Y* variable subject to random outlier contamination; the number of outliers was binomial with *n* equal to 20 and the probability of 0.2 for outliers. In this run, five outliers were generated from a N(0, 5) distribution. Only three of these five outliers are apparent on the scatterplot of the data. The data and the middle 20 elementary slopes out of the 190, 20 choose 2, are given below.

Figure 1 displays the scatterplot and three regression lines: the classical least squares, the ordinary *GD* regression line from (2.3), and the *GD* elementary slope regression line from (2.4), the darkest line.

Figure 2 shows a plot of 101 ordered elementary slopes, order statistics 46 to 146, versus the corresponding quantiles for the standard Cauchy. These 101 order statistics of elementary slopes were used as the dependent variable and the corresponding quantiles for the standard Cauchy random variable as the regressor variable in (2.1). The slope of this regression line is the solution *s* to (2.4); here *s* = 2.0485. The intercept of this regression line is the median, *c*, of the uncentered residuals, namely *c* =1.5708. The *GD* elementary slope estimate of r is 0.6085.

Thus the *GD* elementary slope regression line, for the original data, shown in Graph 1 is $Y = -3.564 + 1.5708X$. The intercept, *c*, of the regression line in Graph 2 is the estimate of the slope and median of $(y - 1.5708x)$ using the original data is the intercept.

A summary of the three fits is now given in the form Method(intercept, slope):
LS (-10.76, 2.13); *GD* (-4.22, 1.73); *GD* on elementary slopes (-3.564, 1.5708). The two *GD* methods are very similar and preferable to the LS method.

The theoretical regression line, without outliers, was $E(Y) = 1.2x$ with the standard deviation of the residuals being 4.8. The correlation parameter was 0.6. The *GD* elementary slope method of estimating correlation was excellent in this example, and the slope estimate was superior to the classical least squares method and the ordinary *GD* regression.

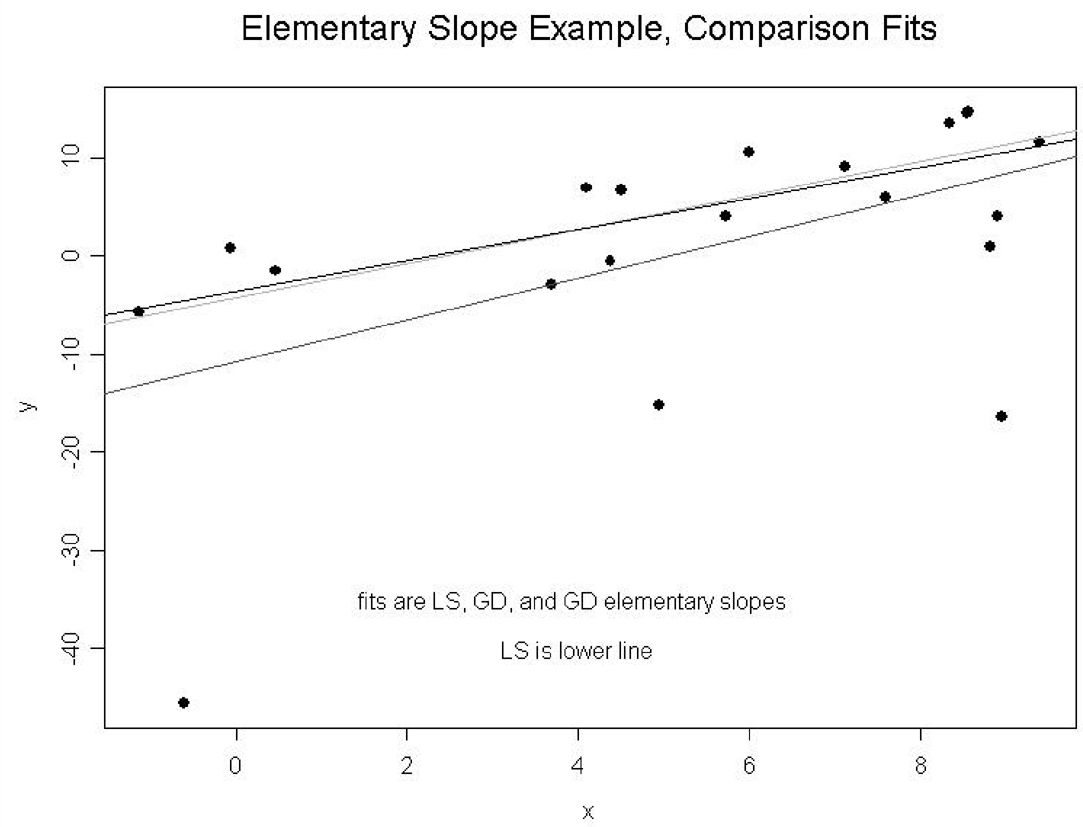

Figure 1

The ordered x-data:

-1.12515065	-0.06094178	-0.60768302	0.46794319	3.68813456
4.10270239	4.37505686	4.50201012	4.94726900	5.72811198
5.99845688	7.11044077	7.59002677	8.34436002	8.53905350
8.55759799	8.80890911	8.90513910	8.95517566	9.38513190

The corresponding y-data:

-5.7732964	-45.6719067	0.7431537	-1.5227881	-2.9523418
6.9241551	-0.5675762	6.6362350	-15.3101496	3.9947963
10.5330247	9.0918245	5.9374006	13.5577737	14.5609510
14.7592589	0.9640291	4.0345965	-16.3929683	11.6197997

The middle 20 elementary slopes:

1.291506	1.339262	1.343713	1.425320	1.473849
1.484517	1.524588	1.563921	1.585425	1.597985
1.606721	1.615651	1.626274	1.651427	1.654865
1.721414	1.758763	1.801382	1.804985	1.903659

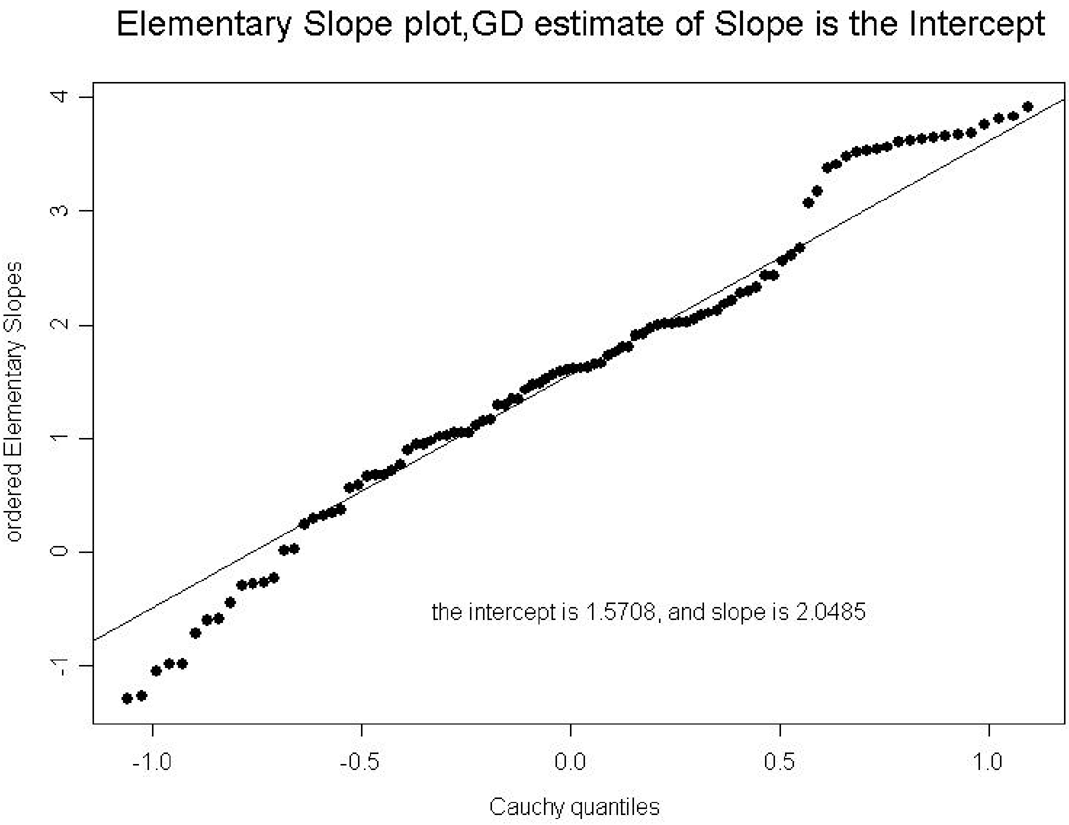
Elementary Slope plot,GD estimate of Slope is the Intercept
ordered Elementary Slopes
4
3
2
1
0
-1
the intercept is 1.5708, and slope is 2.0485
-1.0
-0.5
0.0
0.5
1.0
Cauchy quantiles

Figure 2