Cauchy Regression and Confidence Intervals for the Slope

Rudy Gideon
University of Montana
Missoula, MT 59812

and

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

This paper uses computer simulations to verify several features of the Greatest Deviation (*GD*) nonparametric correlation coefficient. First, its asymptotic distribution is used in a simple linear regression setting where both variables are bivariate. Second, the distribution free property of *GD* is demonstrated using both the bivariate normal and bivariate Cauchy distributions. Third, the robustness of the method is shown by estimating parameters in the Cauchy case. Fourth, a general geometric method is used to estimate a ratio of scale factors used in the confidence interval. The methods in this paper are an outgrowth of general research on the use of nonparametric correlation coefficients in statistical estimations. The results in this paper are not specific to *GD* and are appropriate for other rank based correlation coefficients.

Key words: bivariate normal, bivariate Cauchy, Greatest Deviation correlation coefficient, asymptotic distribution

This work depends in part on earlier unpublished work of Gideon and is available on his web site: www.math.umt.edu/gideon.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. The Bivariate Cauchy

The bivariate Cauchy is, of course, the same as the bivariate t with 1 degree of freedom (Cornish 1954, Dunnett and Sobel 1954). Since this distribution is seldom used in simulations and estimation, it is discussed here. The joint, marginal, and conditional distributions are given; the procedure that was used to generate bivariate data is then explained.

Let (t_1, t_2) be an outcome for a (T_1, T_2) bivariate Cauchy random variable. The joint distribution is given in two forms, one to generate random observations and the other to see the elliptical nature of the contours. The elliptical nature of the contours, as with the bivariate normal, allows *GD* to be distribution free over the whole class of bivariate t distributions.

$$f(t_1,t_2) = \frac{(1+t_1^2)^{-\frac{3}{2}}}{2p\sqrt{(1-r^2)}}\left(1+\frac{(t_2-rt_1)^2}{(1-r^2)(1+t_1^2)}\right)^{-\frac{3}{2}}$$

$$= \frac{1}{2p\sqrt{(1-r^2)}}\left(1+\frac{t_1^2-2rt_1t_2+t_2^2}{1-r^2}\right)^{-\frac{3}{2}}, \qquad \text{for } -\infty < t_1, t_2 < \infty .$$

Both marginal distributions have the same standard Cauchy:

$f(t) = \dfrac{1}{p(1+t^2)}$ for $-\infty < t < \infty$.

The conditional distribution of T_2 is given for T_1. For a fixed t_1,

$$f(t_2|t_1) = \frac{1}{2\sqrt{1-r^2}\sqrt{1+t_1^2}}\left(1+\frac{(t_2 - rt_1)^2}{(1-r^2)(1+t_1^2)}\right)^{-\frac{3}{2}}.$$

This density is related to a standard Student t distribution with 2 degrees of freedom by the following transformation.

Let $w = \dfrac{\sqrt{2}(t_2 - rt_1)}{\sqrt{1-r^2}\sqrt{1+t_1^2}}$, so that $dw = \dfrac{\sqrt{2}dt_2}{\sqrt{1-r^2}\sqrt{1+t_1^2}}$, and

$$f(w) = \frac{1}{2\sqrt{2}}\left(1+\frac{w^2}{2}\right)^{-\frac{3}{2}} \text{ for } -\infty < w < \infty.$$

These distributions are now used to generate bivariate Cauchy data. Let t_1 equal the outcome of a standard Cauchy random variable, and let *w* be an independent outcome of a Student t with 2 degrees of freedom. The computer package *S-Plus* was used in these simulations. The *S-Plus* commands: $t_1 <- rcauchy(n)$ and $w <- rt(n,2)$ were used to generate random samples of size *n*. After randomly generating t_1 and *w*, let $t_2 = r\,t_1 + \left(\dfrac{(1-r^2)(1+t_1^2)}{2}\right)^{\frac{1}{2}} w$. A standardized bivariate Cauchy random variable (t_1, t_2) has been generated. This latter form also shows the regression equation. The slope is *r* and the scale factor is $\sqrt{\dfrac{(1-r^2)(1+t_1^2)}{2}}$. The random quantity *w* plays the role of the error variable in the regression. It will be shown in the simulations that *GD* regression estimates *r* and gives a nearly correct asymptotic confidence interval for this slope. In addition, *GD* regression residuals are used to estimate an error scale factor, which is the above scale factor. Because the same type of estimation and confidence intervals are used for the bivariate normal, these simulations show the distribution free property and robustness of the *GD* simple linear regression method.

2. The Bivariate Normal

A method of generating bivariate normal values that parallels the above method is to generate two independent *N*(*0*, *1*) outcomes, z_1 and z_3; then with the correlation parameter *r*, let $z_2 = r\,z_1 + \sqrt{1-r^2}\,z_3$. The bivariate random variable (Z_1, Z_2) has zero means, variances of one, and correlation of *r*.

For a more direct normal model in simple linear regression, let *X* be $N(m_1, s_1)$ where the second parameter is the standard deviation as this notation is commonly used in

computer packages. Let e be independent of X with a $N(0, s)$ distribution. Construct the standard model $Y = a + b\,x + e$. Thus, the conditional distribution is constructed given the *x* and a random error. In *S-Plus*, this becomes: $x < -rnorm(n, m_1, s_1)$ and $y < -a + b\,x + rnorm(n, o, s)$ where *n* is the sample size.

3. The *GD S-Plus* regression routines

Gideon and Hollister (1987) define the *GD* correlation coefficient. It is restated here for convenience as all work depends on it.

The *GD* correlation coefficient is defined as follows: let the bivariate data (x, y) be ordered by the *x*-data. Each data vector is then replaced by its ranks. The data is now (e, p) where e is the identity vector, 1,2,3,….,*n*, and p contains the corresponding ranks of the y-data. *I*, an indicator variable, is 0 or 1 depending whether its expression is false or true, respectively. Then $GD(x, y) = GD(e, p)$ and in the definition below the p_j are the components of p.

$$GD(x, y) = \frac{\max\limits_{1 \le i \le n} \sum\limits_{j=1}^{i} I(n + 1 - p_j > i) - \max\limits_{1 \le i \le n} \sum\limits_{j=1}^{i} I(p_j > i)}{\left[\frac{n}{2}\right]}$$

There is only one *GD S-Plus* function that needs to be directly called to perform the work. It is written in C code and called in a *S-Plus* calling sequence. Work over a long time period has made the calculation fairly efficient. Given a paired data vector (x, y), the *GD* simple linear regression function is $rgrgc(x, y, 0)$. This returns the intercept and slope, say *a* and *b*. The regression predicted values $\hat{y} = a + bx$ and the residuals $y - \hat{y}$ are now formed. A second regression is performed on the sorted x and sorted residuals $y - \hat{y}$. In *S-Plus*, this done by $rgrgc(sort(x), sort(y - \hat{y}), 0)$. The slope of this regression estimates $\sqrt{n}$ times the estimated standard error of the slope estimate in classical least squares; that is, $\sqrt{n}\, s_{\hat{b}} = \dfrac{\sqrt{n}\, \hat{s}_{res}}{\sqrt{\sum (x_i - \bar{x})^2}}$. This is demonstrated in the simulations and is explained fully in Gideon and Rothan (2004). The method is very general in that any correlation coefficient can be used in the same manner. *GD* is being used because it is very robust and also gives good results on data without outliers. It must be emphasized that this method of measuring the variation with the slope is necessary to make the method distribution free. One uses the actual observed sorted *x*-data as the standard to obtain a relative measure of the variability in the residuals. The expectation is that if the residuals have the same type of distribution, then slope of the regression line of the ordered residuals on the

ordered *x*'s (t_1 for the Cauchy) will provide a reasonable estimate of $\frac{\hat{s}_{res}}{s_x}$. In the simulations that follow, the data is bivariate normal or bivariate Cauchy; and apparently this geometrical method of estimating $\frac{\hat{s}_{res}}{s_x}$ does allow the distribution property to work. For an alpha significance level, nearly correct $1-a$ confidence interval levels are obtained.

4. Greatest Deviation Asymptotics

The population parameters of *GD* for the bivariate normal and Cauchy distributions are found in Gideon and Hollister (2000). A Taylor Series expansion was used to relate the distribution of slope to the asymptotic distribution of r (Gideon, Prentice, and Pyke 1989). These results are used in forming confidence intervals. Let $GD(X,Y)$ be the population parameter of *GD* for the bivariate random variable (X,Y). Then for both the bivariate normal and Cauchy distributions $GD(X,Y)=\frac{2}{p}\sin^{-1} r$, where r is the correlation coefficient in the normal case, but can only be called the correlation parameter in the Cauchy case. The usual definition of r does not suffice for the Cauchy as moments do not exist. The Cauchy correlation parameter r does exist for *GD* and is derived from the population definition of *GD* in Gideon and Hollister (2000). Let F and G be the distribution functions of continuous random variables *X* and *Y*. Now let $U=F(X)$ and $V=G(Y)$, then the function $C(u,v)=P(U\le u, V\le v)$ is known as the Copula function. The population parameter of *GD* for (X,Y) is

$$GD(X,Y)=2\sup_{0\le t\le 1} C(t,1-t)-2\sup_{0\le t\le 1}(t-C(t,t)).$$

For the asymptotic distribution of *GD* when $r=0$, Gideon, Prentice, and Pyke (1989) showed that for data vectors (x,y), for a sample size n, that as n increases $\sqrt{n}GD(x,y)$ becomes $N(0,1)$. From the population parameter of *GD* and its asymptotic distribution, the asymptotic distribution of the estimated slope in either the bivariate Cauchy or normal is as follows:

$\hat{b}$ converges to a $N(r\frac{s_y}{s_x},\frac{p\,s_y\sqrt{1-r^2}}{2\sqrt{n}\,s_x})$ distribution.

In the computer simulations, $\frac{s_y\sqrt{1-r^2}}{s_x}=\frac{s_{res}}{s_x}$ is estimated directly as the slope in the *GD* regression where the residuals are regressed on the *x*-data

($rgrgc(sort(x), sort(y - \hat{y}), 0)$), that is, $\frac{\hat{s}_{res}}{s_x}$. The asymptotic $1 - a$ confidence interval is then this slope, $\frac{\hat{s}_{res}}{s_x}$, times the remaining terms in the asymptotic standard deviation above. So the distribution-free robust confidence interval for the slope in a simple linear regression is

$\hat{b} \pm \frac{p\ z_{a/2}}{2\sqrt{n}} \frac{\hat{s}_{res}}{s_x}$, where $z_{a/2}$ is the upper $a/2$ quantile of the standard normal.

For additional background material for this section, see Gideon (1992), Gideon and Rummel (1992), Gideon and Rothan (2004).

5. Computer Simulations for the Distribution-Free Confidence Interval

The computer simulations primarily demonstrate the confidence interval for the slope in the simple linear regression, but some correlation results are also given. Values of r used are 0, 0.5, 0.75, 0.9. For simplicity, only 0 is used for the location parameters and 1 for the scale parameters. The two distributions used are bivariate normal and bivariate Cauchy as explained earlier. For the standard univariate Cauchy, the 3rd quartile is 1 $(Q_1 = -1)$ so that the median of $1 + T_1^2$ is 2. Thus, the scale factor $\sqrt{\frac{1 - r^2}{2}(1 + t_1^2)}$ in the simulations should vary about $\sqrt{\frac{1 - r^2}{2} * 2} = \sqrt{1 - r^2}$. Note that this is the same as in the bivariate normal case (*i.e.*, the residual scale factor). Since the scale factor of the *x*-variable with standardized distributions is one, the ratio $s_{res} / s_x = \sqrt{1 - r^2}$ and $\hat{s}_{res} / s_x$ estimates $\sqrt{1 - r^2}$. The difference between the Cauchy and the Normal is that the scale factor for the Cauchy averages $\sqrt{1 - r^2}$; but for individual points, the scale factor depends on the given value of t_1. The scale factor is constant for the normal. The slopes and the correlations have the same value in the simulations since the standard distributions were used.

The tables are labeled by a Roman numeral in the upper left corner. Table I shows means of 1000 simulations each for a sample size of 68, except for the last two rows that are based on 5000 simulations. The categories are the ratio of the scale factors resulting from the simulations; the population scale factor; the intercept whose population parameter is 0; the slope whose population parameter is r; the proportion of the samples whose confidence interval contains the true slope, the *GD* estimate

of r , $\sin(p\, GD(x, y)/2)$; and the population correlation coefficient. It is known that the sine transformation of *GD* underestimates r except at zero. Except for the population parameters, all entries are the mean of the given number of simulations. Rows in the table alternate between Cauchy and Normal in order to view the distribution-free robustness property.

Table I	Ratio $\hat{s}_{res}/s_x$	$\sqrt{1-r^2}$	Intercept	Slope	Sample prop. 90% level	Corr.	r
Normal	1.0044	1	-0.0066	-0.0033	0.934	0.0012	0
Cauchy	0.9958	1	0.0127	0.0013	0.903	0.0032	0
Normal	0.8701	0.8660	-0.0013	0.5001	0.930	0.4671	0.50
Cauchy	0.8592	0.8660	0.0171	0.5015	0.898	0.4866	0.50
Normal	0.6650	0.6614	0.0012	0.7456	0.914	0.7076	0.75
Cauchy	0.6620	0.6614	0.0027	0.7490	0.922	0.7105	0.75
Normal	0.6644	0.6614	-0.0002	-0.7557	0.927	-0.7084	-0.75
Cauchy	0.6584	0.6614	-0.0082	-0.7473	0.901	-0.739	-0.75
Normal	0.4348	0.4358	-0.0004	0.9033	0.925	0.8664	0.90
Cauchy	0.4335	0.4358	0.0028	0.8986	0.890	0.8670	0.90
The following two rows are for 5000 simulations							
Normal	1.0010	1	-0.0037	-0.0002	0.9284	0.0001	0
Cauchy	0.9899	1	0.0024	-0.0006	0.9132	0.0000	0

Table II reflects changes made to the confidence coefficient, the correlation, and slope; the number of simulations is kept at 1000. The first two rows are for 90%, the next two are 80%, and the final two are for 50% confidence levels. It is becoming clear that for sample size 68, the confidence coefficient is a bit conservative. Note that the mean values for the ratios and those for the slope estimates are very close to their respective parameters. As expected, the direct estimate of the correlation via the *GD* sine transformation underestimates for $r > 0$.

Table II	Ratio $\hat{s}_{res}/s_x$	$\sqrt{1-r^2}$	Intercept	Slope	Sample prop. 90%, 80% 50% levels	Corr.	r
Normal	0.6654	0.6614	-0.0017	-0.7571	0.974	-0.7120	-0.75
Cauchy	0.6578	0.6614	0.0073	-0.7552	0.959	-0.7390	-0.75
Normal	0.4309	0.4358	-0.0005	0.9045	0.828	0.8687	0.90
Cauchy	0.4364	0.4358	0.0048	0.8984	0.830	0.8666	0.90
Normal	0.8014	0.8000	-0.0005	0.5957	0.554	0.5564	0.60
Cauchy	0.8102	0.8000	0.0001	0.6031	0.553	0.5676	0.60

In Table III sample sizes of 50, 75, 100, and 150 are used while the correlation remains at 0.80 and the confidence coefficient at 90%. The first two rows are from samples of size 50, the next two of 75, then 100, and finally 150. The number of simulations for each remains at 1000. It seems remarkable that the Cauchy confidence interval for the slope is more accurate than the normal.

Table III	Ratio $\hat{s}_{res}/s_x$	$\sqrt{1-r^2}$	Intercept	Slope	Sample proportion 90% level	Corr.	r
Normal-50	0.6014	0.6000	0.0076	0.7938	0.928	0.7475	0.80
Cauchy-50	0.5987	0.6000	-0.0016	0.8015	0.889	0.7705	0.80
Normal-75	0.6046	0.6000	-0.0014	0.8058	0.943	0.7712	0.80
Cauchy-75	0.5922	0.6000	-0.0056	0.7992	0.916	0.7780	0.80
Normal-100	0.6030	0.6000	-0.0002	0.8042	0.937	0.7699	0.80
Cauchy-100	0.5997	0.6000	0.0033	0.8030	0.909	0.7705	0.80
Normal-150	0.5976	0.6000	0.0016	0.8008	0.921	0.7796	0.80
Cauchy-150	0.6009	0.6000	0.0016	0.7969	0.917	0.7763	0.80

In Table I, the Cauchy simulations usually gave values closer to the stated 90% confidence level. The means of the simulated scale ratio estimates for both the Normal and Cauchy were very close to the population parameter, $\sqrt{1-r^2}$. All intercept and slope means were close to the population parameters. From Table II, it appears that the proportion of confidence intervals containing the true slope is higher

than the stated confidence level. Tables I - III indicate that the Cauchy generated data gives results closer to the stated confidence level than the normal. The results do not seem to vary much between the moderate sample sizes of 50 to 150.

6. An Example of the Confidence Interval for the Cauchy Case

One more run of 1000 Cauchy simulations was made. The confidence coefficient was 90%, the sample size 168, $r = b = 0.8$, and $a = 0$. The simulation results for the last sample remain in *S-Plus* and are used for an illustrative example. The analysis of this last sample demonstrates that no special selection process was used to make this system perform well. The summary statistics from the run for the 1000 simulations follow. The mean intercept was –0.0013 and the mean slope was 0.7983. The mean of the sine transformation of *GD* was 0.7789. The proportion of the 90% confidence intervals containing the true slope was 0.918. The term $\sqrt{1-r^2} = 0.6$, and the mean value of the 1000 simulations for this term was 0.6001. The median of these was also always calculated and for this run it was 0.5955. Because the median and mean were always close, the median was not included in the above tables.

For this last simulation of sample size 168, confidence intervals and scatterplots with regression lines are given for *GD* and least squares. The axes are labeled (t_1, t_2) because bivariate Cauchy data is used. For *GD* estimation system, the slope of the regression used to estimate relative variation was 0.5940, *i.e.* $\frac{\hat{s}_{res}}{s_x} = 0.5940$. This is the slope as explained on page 3 for a regression on sorted data. The upper 95th percentile of the standard normal 1.645 is used in the term $\frac{p\ 1.645}{2\sqrt{168}} = 0.1993$. This number times the 0.5940 gives the distance of the confidence interval around the slope estimate. Table IV lists the intercepts, slopes and the confidence intervals for the *GD* general method and for least squares.

Table IV	Comparison of LS and *GD* on one sample where n = 168, Cauchy data			
	Intercept	Slope	CI lower	CI upper
GD	-0.0643	0.8187	0.7003	0.9371
LS	0.3487	0.6376	0.5534	0.7198

It is readily seen that *GD* is, as expected, much better than least squares (LS). The LS method does not include the true slope in the confidence interval. The standard error of LS slope estimate, 4.114, and $\sum (t_1 - \bar{t})^2$, 6858.56, are given so that the confidence interval for LS can be checked. Although this paper is not a comparison of *GD* to other robust procedures, the L-one and rreg (robust regression in *S-Plus*) results are given. The results are in Table V, but this author does not know the distributional properties of these estimators to obtain a confidence interval.

Table V	L-one and robust regression results from *S-Plus*	
	Intercept	Slope
L-one	-0.0748	0.7920
rreg	-0.0782	0.8421

Two scatterplots are presented. Figure 1 has all the data and in Figure 2 the data is restricted to those values between –5 and +5 in order to give a better view of the central area of the data. The graph of the restricted data demonstrates that, as is always the case, the *GD* regression line goes through the heart of the data. There are 139 of the 168 outcomes between –5 and +5.

Figure 3 shows the geometry of the *GD* estimate of $\frac{s_{res}}{s_{t_1}}$. The line fitted to the sorted data $\{sort(t_1), sort(residuals)\}$ by the *GD* system has slope 0.5940. The robustness of the *GD* fit to the line can be observed because the line appears uninfluenced by the large residuals.

7. Summary

Computer simulations have demonstrated the robustness and distribution-free properties of the rank based Greatest Deviation correlation coefficient in simple linear regression. This was accomplished by comparing the computer results for the bivariate normal and Cauchy distributions. The asymptotic properties of *GD* as a correlation coefficient and its extension to simple linear regression were used. One example was provided that demonstrated the *GD* geometrical method of obtaining an estimate of the variation in the slope estimator. This paper is based on the results in two published papers. Unfortunately, much of the background material is unpublished, but it is available on the web site. The generality of this work may not be apparent without viewing the papers on the web, but any correlation coefficient can be used in the manner demonstrated in this paper. The unpublished L-one correlation coefficient that should accompany all statistical L-one methods is a prime candidate. This paper is part of a system of estimation based on the use of correlation coefficients. The web site shows some of the work, but some further examples in times series, nonlinear estimation, and generalized linear models have not yet been posted.

References

Cornish, E. A. (1954), "The Multivariate t-Distribution Associated with a Set of Normal Sample Deviates," *Australian Journal of Physics*, 7, 531-542.

Dunnett, C. W., and Sobel, M. (1954), "A Bivariate Generalization of Students t-Distribution, with Tables for Certain Special Cases," *Biometrika*, 41, 153-169.

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.

----- (2000), "The Geometrical Definition of Greatest Deviation Correlation Coefficient and its Uniqueness," unpublished paper (URL: http://www.math.umt.edu/gideon/TheGeometricalDefinitionGDCC.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.

Gideon, R. A., Prentice, M. J., and Pyke, R (1989), "The Limiting Distribution of the Rank Correlation Coefficient, *GD*," *Contributions to Probability and Statistics, Essays in Honor of Ingram Olkin*, Gleser, L.J. et al. (eds.), New York: Springer Verlag, pp. 217-226

R. A. Gideon: Acknowledgments

Ron Pyke, University of Washington, my Masters supervisor, and for completing the asymptotic distribution derivation and write-up

John Gurland, University of Wisconsin, my Ph.D. advisor at Madison Wisconsin

Mike Prentice, University of Edinburgh, for asking the question, "What is it estimating?" and allowing me a Sabbatical in Scotland

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.

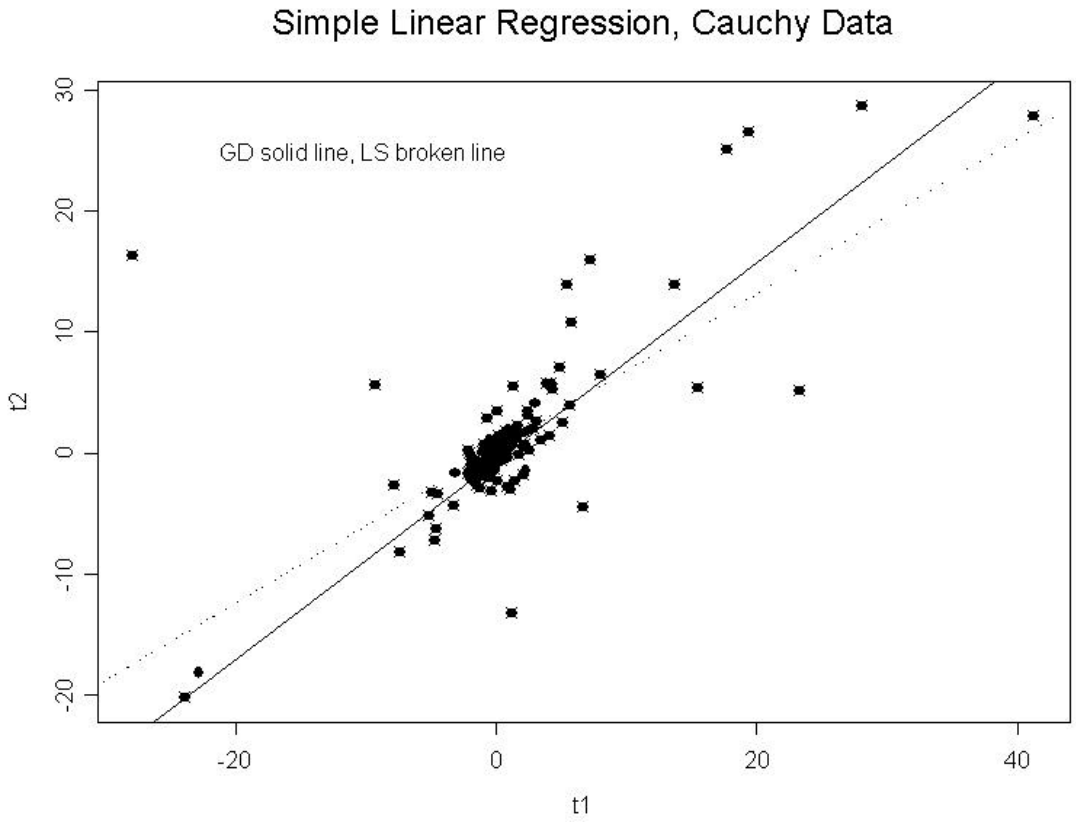
Simple Linear Regression, Cauchy Data
GD solid line, LS broken line
t2
t1
30
20
10
0
-10
-20
-20
0
20
40

Figure 1

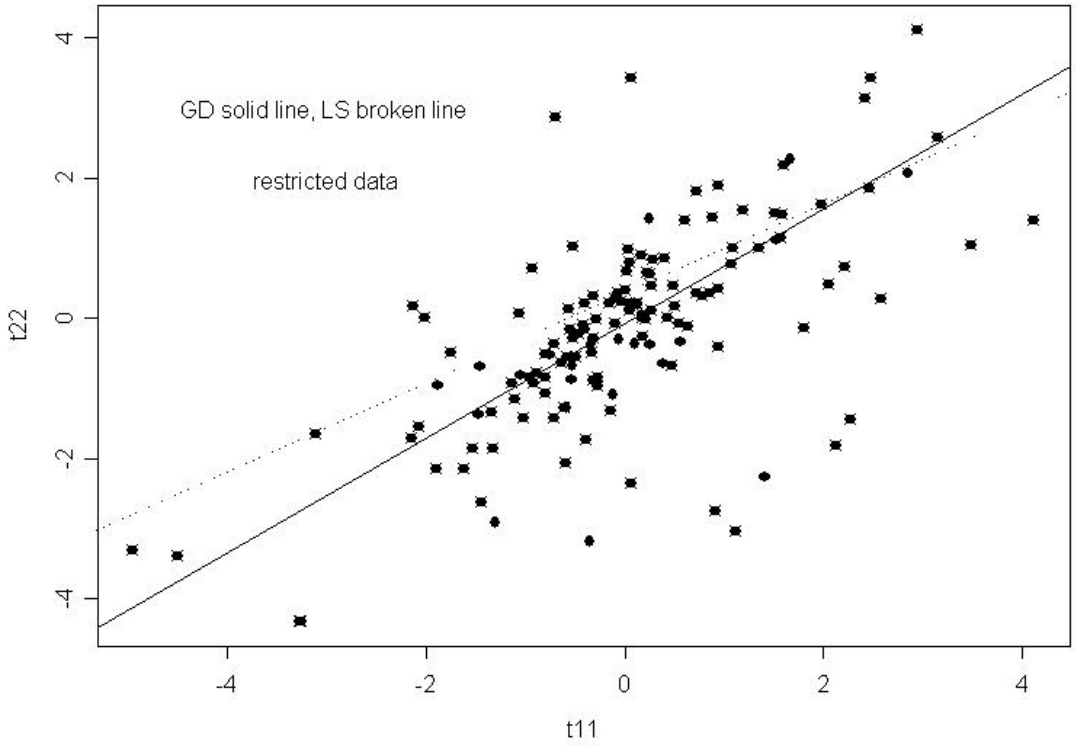
GD solid line, LS broken line
restricted data
t22
t11
4
2
0
-2
-4
-4
-2
0
2
4

Figure 2

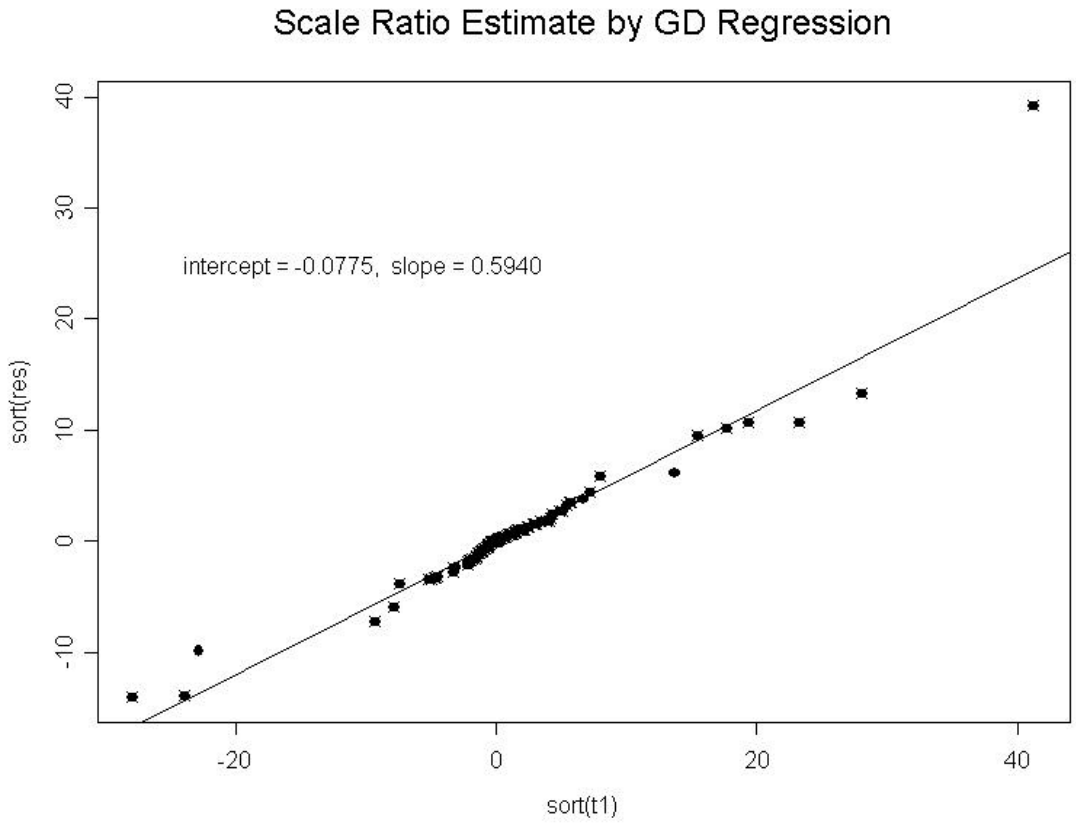
Scale Ratio Estimate by GD Regression
intercept = -0.0775, slope = 0.5940
sort(res)
sort(t1)
40
30
20
10
0
-10
-20
0
20
40

Figure 3