Obtaining Estimators from Correlation Coefficients: The Correlation Estimation System and R

Abstract

Correlation coefficients (CCs) are generally viewed as summaries, causing them to be underutilized. Viewing them as functions leads to their use in diverse areas of statistics. Because there are many correlation coefficients (see, for example, Gideon (2007)) this extension makes possible a very broad range of statistical estimators that rivals least squares. The whole area could be called a "Correlation Estimation System" (CES). This paper concentrates on outlining the numerous possibilities for using the CES without thorough explanation but with some illustrative examples. It gives the formulae to make possible both the estimation and the computer coding to implement it. This approach has been taken in hopes that this condensed version of the work will make the ideas accessible, show their practicality, and promote further developments.

1. Introduction

After a professional lifetime of teaching and research the author realized that virtually everything that one does with least squares or normal theory could be done with any of a multitude of correlation coefficients; moreover, it could be done in a coherent fashion, with essentially one basic equation. Both continuous and rank based CCs use the same formulae without change of notation. This means that the same computer code could be written to encompass all estimation involving all the CCs. A user could designate which CC was desired and then all computer calculations thereafter would be based on that choice with minimal change to the rest of the computer code. This includes least squares through Pearson's CC. However, since the degree of robustness of all estimations emanating from a particular CC depends on the degree of robustness of the CC itself, CCs other than Pearson's are usually more desirable.

One focus of this paper is to show how to use any correlation coefficient to estimate location, scale and slope coefficients in simple and multiple linear regression. Once these procedures are developed, the CES is extended into nonlinear regression and estimation of parameters for a particular density type. Some of the results are illustrated with a continuous and with a rank based CC using absolute values. Although not done in this paper the CES can be easily extended into time series and general linear models. Many of these areas have been tested using various CCs over 25 years and all results lead one to believe in the value of the approach.

2. Simple Linear Regression

For a random sample (x, y) in a simple linear regression model and for CC r, let b be the estimate of b, i.e. b is the slope of the regression line, which is the line that makes the residuals $y - bx$ uncorrelated with x. In other words, by analogy with the population correlation of the independent random variable with the residual random variable being

zero, the estimate of b is found by setting the sample equivalent to zero, that is, by solving the regression equation

$$r(x, y - bx) = 0 . \qquad (1)$$

The function $r(x, y - bx)$ is non-increasing as b increases, which makes equation (1) easy to solve numerically. Of course, for least squares, solving this equation with Pearson's r is equivalent to the more familiar minimization process. The median of the uncentered residuals provides a robust estimate of the intercept. However, it is not necessarily the case that the solution to equation (1) corresponds to the solution to a particular minimization for every correlation coefficient. As an example, two CCs are now introduced – one continuous and one rank based. The general framework for them is found in Gideon (2007). First a continuous absolute value CC is given.

Let $SA_x = \sum |x_i - \bar{x}|$ and similarly for y, and define

$$r_{av} = \frac{1}{2}\left(\sum \left| \frac{x_i - \bar{x}}{SA_x} + \frac{y_i - \bar{y}}{SA_y} \right| - \sum \left| \frac{x_i - \bar{x}}{SA_x} - \frac{y_i - \bar{y}}{SA_y} \right|\right).$$

For a bivariate normal distribution with CC r the population value of r_{av} is $r_{av} = \frac{\sqrt{1+r} - \sqrt{1-r}}{\sqrt{2}}$. This is one of the examples in which solving $r_{av}(x, y - bx) = 0$ does not necessarily minimize the L-one norm of $y - (a + bx)$.

In the same way Spearman's CC is found by substituting ranks in place of the original data in Pearson's CC, substituting ranks in r_{av} gives Gini's CC. However, the formula is simplified by ordering the original data by the x-values so that the data replacement is

$$\left.\begin{array}{l} x_1, y_1 \\ x_2, y_2 \\ \vdots \\ x_i, y_i \\ \vdots \\ x_n, y_n \end{array}\right\} \rightarrow \begin{array}{ll} 1, & q_1 \\ 2, & q_2 \\ \vdots & \\ i, & q_i \\ \vdots & \\ n, & q_n \end{array}$$

where q_i is the y point that corresponds to the i^{th} smallest x value. Gini's CC is

$$r_{mf} = \frac{\sum |n + 1 - q_i - i| - \sum |q_i - i|}{\left\lfloor \frac{n^2}{2} \right\rfloor};$$ the subscript *mf* stands for *modified footrule* of Spearman. In 1906 Spearman attempted to formulate a CC based on just $\sum |q_i - i|$, but Gini's valid version was not formulated until 1914. The notation in the denominator is that of greatest integer. Tied value concerns must always be addressed when using rank

based CCs. It has been found that producing a unique value for any nonparametric CC using the max-min tied value procedure outlined in Gideon and Hollister (1987) handles this issue. It can be used on all nonparametric CCs, but in the case of GDCC must be used, as it is the only known viable procedure. While the CES technique is completely general, the population value is not always known. But, for the bivariate normal, the population value of r_{mf} can be given explicitly as $\frac{2}{p}\left[\sin^{-1}(\frac{1+r}{2})-\sin^{-1}(\frac{1-r}{2})\right]$. The regression equation for Gini is $r_{mf}(x,y-bx)=0$ and again there is no known equivalent minimization technique.

Please note that if the model assumption is the bivariate normal or the bivariate t class of distributions then X and $Y-bX$ are independent and therefore r is zero; both r_{av} and r_{mf} are also zero for these random variables.

3. Scale Equations

When working with least squares, finding the variation in the regression residuals via least squares is natural; the two pieces (slope and scale) are naturally connected. When working with these general CCs, we seek the same kind of natural connectivity and so it is inappropriate to use least squares of normal theory for the scale estimate. For example, in the case of r_{av} the measure of variation of x or the variation of the residuals from the regression should be based on absolute values, because the original slope calculations were. In general, the same ideas must be employed for both slope and scale to attain the connectivity we desire. It is the author's experience, that this natural connectivity requirement is necessary to retain desirable qualities such as the degree of robustness.

With this in mind, the variation in the estimate of residuals is found by solving for s in

$$r(y^0,(y-bx)^0-sy^0)=0. \qquad (2)$$

Here the superscript means the data are ordered. s estimates the ratio of standard deviations, s_{res}/s_y and it can also be viewed as the slope, not of the standard regression line, but of a specialized regression, discussed in Section 5. Note that this is essentially the same correlation regression equation (1) applied to ordered data: a numerical routine that solves (1) will also solve (2). Moreover, equation (2) can be used in an entirely different way. Instead of solving (1) for b and utilizing equation (2) to get s (called the Regression Equation Technique or RET), (2) could be used in a minimization: find the b that minimizes s. In this way, equation (2) could subsume equation (1). This minimization is called the Optimization Technique (OT); the idea is exploited in Section 5. Note too that the ordering on the residuals is independent of the ordering on the y variable. This is key because the set of residuals en masse are being measured relative to the set of y values; CES views these globally. The residual for any particular x may be reordered into a different position by the iterative technique used in seeking a minimum.

For the normal distribution, the quantity *s* is estimating $\frac{s_{res}}{s_y} = \sqrt{1 - r^2}$, where r is the population correlation coefficient. The scale equation (2) is examined in detail in Gideon and Rothan (under review). This paper can be found on the Web site as well.

4. Multiple Linear Regression

For matrix X_{nxp} (n rows of independent data in which each column is a regressor variable) and *y* (the dependent variable), the associated regression equations are

$$r(x_i, y - X_{nxp}\hat{b}) = 0, i = 1, 2, \ldots, p. \quad (3)$$

$\hat{b}$ is a solution vector whose i^{th} component , $\hat{b}_i$, is the coefficient of variable x_i. Rummel (1991) shows how to solve these using Gauss-Seidel for the case when *r* is GDCC, and again, ties are not a problem using the max-min procedure. Every CC has its own set of regression equations; these correspond to the normal equations in the case of Pearson's r_p. These regression equations would have solutions $\hat{b}_t$ had Kendall's Tau been chosen as the CC. Note that these regression equations give a way to incorporate Kendall's Tau into the realm of multiple regression; this is desirable as Tau is moderately robust and has several remarkable features. The regression equations for Tau can be solved by iterations involving the medians of the elementary slopes. (See Sen, 1968 or Gideon and Rummel, 1992, for the simple linear regression case and for an illustrated look at this work specialized to Kendall's Tau see *Correlation and Regression without Sum of Squares* on the Web site). Some CCs do not have the same linearity properties as does Pearson's r_p and so it is not necessarily true that for *l* a p-dimensional column vector of constants, $r(X_{nxp}l, y - X_{nxp}\hat{b})$ is exactly zero; however, computer simulations have shown that, at least for GDCC, and probably in general, it is zero or very close to zero.

The variation in the estimate of residuals, *s*, which is the slope of a regression line on ordered data (superscript ^{0}) is found as the solution to

$$r(y^0, (y - X_{nxp}\hat{b})^0 - sy^0) = 0. \quad (4)$$

This is, of course, the multiple regression version of equation (2); it is another instance of the Regression Equation Technique (RET). Again note that a minimization could be used instead: find the $\hat{b}$ that minimizes *s*; in other words the Optimization Technique (OT) could be employed.

A multiple CC is defined to be

$$\sqrt{1 - s^2} \ . \quad (5)$$

5. A Correlation Coefficient Approach to Minimization of the SD ratio, s_{res}/s_y

If one wants a minimal variation estimate *s*, again equation (4) is used but in the reverse way. Now the coefficients $\hat{b}_i, i = 1,2,\ldots, p$ are chosen to minimize *s* in

$$r(y^0, (y - X_{nxp}\hat{b})^0 - sy^0) = 0. \quad (6)$$

Again, for the normal distribution, *s* is estimating $\frac{s_{res}}{s_y} = \sqrt{1 - r^2}$ where now r is the multiple correlation coefficient and so minimizing *s* is equivalent to maximizing r.
It is not true that the results from RET and OT must be the same, but as expected it has been found that they are similar. The dearth of linearity properties of some CCs is the cause of this difference. A few examples are given using the two CCs defined above. All results for each method and all the examples were high-quality, leading us to even greater conviction of the value of the methods.

Before proceeding to the examples, equation (6) is expressed in a more general form. Let the response variable *y* be modeled by some function, *f* with argument *x*, which relies on vector parameter b so the equation becomes

$$r(y^0, (y - f(x, \hat{b}))^0 - sy^0) = 0. \quad (7)$$

In other words, this form of regression is very general and relies on routines in numerical packages to determine $\hat{b}$ for minimizing *s* (in OT), the estimate of s_{res}/s_y. It is helpful to view this equation geometrically. The ordered residuals are plotted against the ordered *y*s and a simple linear regression is performed whose slope *s* is used to ascertain the closeness of the fit. This is done not by focusing on the sum of individual vertical deviations but by forcing the residuals overall to be relatively small as measured against the *y*s. Thus equation (7) plays the role in CES that the residual sum of squares does in classical least squares analysis. Theoretically *s* will vary between 0 and 1; a value of 0, or a correlation of 1, denotes an exact fit whereas 1, or a correlation of 0, means there is no information in the model under discussion.

6. Examples and Comparisons of the RET and OT Methods

In this section, some examples are given to illustrate the use of the CES method. The examples include simple and multiple linear regression as well as a nonlinear model, all done with the continuous absolute value correlation coefficient and compared to least squares and the Pearson correlation coefficient method using equations (1) through (7). Some of the examples use data from Chatterjee and Price [C/P], some from Draper and Smith [D/S], and some use simulation data. The SLR examples used many correlation coefficients. Section 6 introduces a specialized version of equation (7) that allows estimation of parameters for univariate distributions.

The computations to solve equations (1) through (7) are accomplished in software package R via routines uniroot, nlm, and optimize. Examples of the interaction of these R routines with CES is illustrated in the Appendix, Section 8 and commented on in the Conclusion.

6.1 Simple Linear Regression

This first example illustrates the broadness of the method by performing a simple linear regression on some data in C/P with several correlation coefficients. The R programs to do this depend on defining the correlation coefficient and utilizing it in a way that the R-routine uniroot can accept. The set of commands that calls uniroot can easily be given a new correlation coefficient argument so a new slope is obtained. It is interesting that both continuous and rank based correlation coefficients work equally well. Table 1 is given to allow comparison of the slope estimates for each correlation coefficient. Gideon (2007) should be consulted for the definitions of the various CCs and illustrative information.

The data from C/P, page 21, illustrates the results of using formulas (1) and (2) for some correlation coefficients. This data had four outliers, two on each side of the bulk of the data that made the regression line steeper than it would otherwise have been. Also equation (6) is illustrated for this simple linear regression with just the absolute value correlation coefficient r_{av} and MAD. MAD is a median deviation correlation coefficient which is compatible with the existing MAD scale estimator. It also appears in Gideon (2007). See Section (a) of the Appendix for an outline of the setup of the relevant R program.

Table 1: C/P data fitted using various correlation coefficients

Method	Slope from (1)	Intercept	Relative *s* from (2)
r_{av}	0.558	2.329	0.813
Gini	0.571	2.264	0.910
MAD	0.268	3.532	0.908
GDCC	0.383	3.060	0.923
Pces*	0.665	1.642	0.954
OT method (r_{av}) from (6)	0.571	2.264	0.812
OT method (MAD) from (6)	0.283	3.456	0.871
To compare, the LS values are slope = 0.665, intercept = 1.706, and *s* = 0.791, and LS with 4 outliers deleted gives slope = 0.260, intercept = 3.713, and *s* = 0.935			

* Here "Pces" means using equations (1) and (2) with Pearson's correlation coefficient; Pces stands for Pearson's with CES. Used below also.

For the slope calculation, Pces gives the same result as LS, but the median rather than the mean of uncentered residuals was used for all the intercept estimations, including the Pces calculation. Further, equation (2) was used for relative *s* for Pces whereas LS uses two separate estimates of s_{res} and s_y to compute *s*.

The slope estimates in Table 1make it apparent that MAD and GDCC are by far the most robust methods for this data. Note that the last column shows that almost all the correlational methods give a better estimate of the *s* ratio than LS assuming that the LS

results with the outliers deleted actually gives the best estimate. See C/P for their discussion. The two OT rows of Table 1 are produced by minimizing *s* in equation (6). Note the result is close to that of the RET method using MAD, but now of course the slope value of the OT only approximates the solution to equation (1).

6.2 Multiple Linear Regression

In this section some data from C/P page 59 is used as well as some multivariate normal that is generated randomly with a random correlation structure. For the C/P data there are six regressor variables to fit to the response variable. All six are fit and then the two most important, variables one and three, as determined in C/P, are fit. For the RET, the correlation coefficients MAD, GDCC, r_{av}, and r_{mf} are computed for variables one and three, as well as for all six. The Least Squares results are also given. The R-instructions using the R-routine nlm for the solutions are sketched in Section (d) of the Appendix.

Table 2: RET method for 4 CCs, C/P data, variables 1 and 3

Method	number iterations	intercept	coefficient of x_1	coefficient of x_3	relative *s*
MAD	9	10.362	0.602	0.265	0.531
r_{av}	7	10.320	0.640	0.218	0.531
r_{mf}	5	7.917	0.659	0.243	0.545
GDCC	5	11.141	0.489	0.372	0.618
LS		9.871	0.642	0.211	0.560*

*LS was computed by using $s_{res} = 6.817$ and $s_y = 12.173$ so the ratio is 0.560. For comparison, using the OT on r_{av} took 18 iterations and gave intercept 10.82, coefficients 0.618 and 0.258, and minimum s of 0.527.

Now the full set of C/P data is fit by the various techniques and the results are in Table 3.

Table 3: RET method for 4 CCs, C/P data, all six variables

Method	number of iterations	intercept	coefficient of x_1	coefficient of x_2	coefficient of x_3	coefficient of x_4	coefficient of x_5	coefficient of x_6	relative *s*
MAD	250*	31.41	0.667	-0.084	0.300	0.080	-0.329	-0.085	0.650
r_{av}	7	18.82	0.555	-0.029	0.255	0.208	-0.104	-0.210	0.523
r_{mf}	13	24.42	0.500	0.010	0.254	0.218	-0.180	-0.177	0.563
GDCC	26	33.01	0.320	0.113	0.393	0.250	-0.374	-0.131	0.556
LS		10.79	0.613	-0.073	0.320	0.082	0.038	-0.217	0.581**

* 250 iterations was set as the upper limit and so MAD did not converge. The reason may be that median methods have intervals for the solutions rather than specific points. Even

if convergence were near, the solution interval may be just big enough to contain the various iterates, not allowing convergence.

**LS was computed by using $s_{res} = 7.068$ and $s_y = 12.173$ so the ratio is 0.581.

For Table 4, the criterion to judge convergence was the smallness of either the sum of the absolute value of the changes in the coefficients or the sum of the absolute values of the correlations in equation (3).

Table 4: OT method for 2 CCs, C/P data, all six variables

Method	number of iterations	intercept	coefficient of x_1	coefficient of x_2	coefficient of x_3	coefficient of x_4	coefficient of x_5	coefficient of x_6	relative s
r_{av}	35	10.43	0.639	-0.053	0.282	0.065	0.024	-0.157	0.506
Pces*	39	8.19	0.594	-0.099	0.350	0.118	0.012	-0.133	0.502

*Pces is Pearson's correlation coefficient but used with the OT in equation (6). Note that from equation (5), the estimate of the multiple correlation coefficient is, for the Pces method, 0.865 and for r_{av} is 0.862.

The RET for multiple regression was explored using a random generation of seven normal variates with a random correlation structure; one variable was regressed on the other six. Least squares was compared to the Pces and the r_{av} methods. The LS method does not seem to be better and many times is worse even for strictly normal data. It is worth noting that this observation does not contradict the Gauss Markov theorem since the criterion for success in the CES is not that the standard residual variance is a minimum, but rather that the relative ratio (s_{res}/s_y) is a minimum.

Recall that the OT method relies on a geometric approach in which the slope of a simple linear regression estimates directly the relative ratio. In the RET method, this relative ratio is calculated after estimating the bs. In LS theory, the two approaches (minimizing the standard variance or using RET with Pearson's correlation coefficient or Pces) are identical, whereas in general in CES, the two approaches (RET and OT) give reasonably close results, but are not usually identical. Since any method would win a comparison within its own measurement technique, to give a valid comparison, a rank counting procedure was used. To compare LS, Pces, and r_{av}, 16 runs were made of the 7-variate normal and the closest to the true regression coefficient was recorded by ranks. For each of the 16 random correlation structures, a data set was generated with now known population values. The comparison is shown in Table 5; rank 1 was closest to the true parameter, etc. So 32 = (16)(2) is the expected total sum of the ranks for each column if all three methods are equally good. Note that the r_{av} method was best because estimates for all six coefficients were under the expected 32.

Table 5: Total ranks of 16 simulations of a 7-variate normal distribution

coefficient	1	2	3	4	5	6

LS	38.5	36	35.5	32	33.5	32.5
Pces	32.5	33	34.5	33	34.5	32.5
r_{av}	25	27	26	31	28	31

6.3 Nonlinear Regression

This section gives two examples of estimating the parameters in a nonlinear situation. The illustrations are kept simple by using the exponential distribution but generally any nonlinear model could be considered. Equation (7) is used first with $f(x,a,b) = a\exp(-bx)$ where parameters a, $b > 0$; data was randomly generated by adding normally distributed error to the model. An example from D/S uses an actual data set with the model $f(x,a,b) = a + (0.49 - a)\exp(-b(x-8))$.

In both examples, a and b are varied in order to minimize s; in other words the OT is being employed. Theoretically r can be any correlation coefficient, but for computational purposes the nlm routine in R works only on continuous functions in its minimization technique so only continuous correlation coefficients could be tried. Thus only r_{av} was employed for the randomization example; r_{av} and Pces were used on the D/S example.

6.3.1 The Randomization Example

Many simulations were run, but only one result is given; the sample size is 45, $a = 1$ and $b = 0.5$. The graphs show the two most basic concepts: first, the ordered residuals plotted against the ordered response variable with a regression line, Figure 1, i.e. results from equation (7) and second, the actual fit, Figure 2, with estimated values of 1.012 for a and 0.495 for b. Any curve fitting method is good only when there are sufficient data points throughout the essential range of the model; this was certainly observed in these simulations. With this understanding of having adequate data, very good fits were obtained as illustrated in Figures 1 and 2, showing again the viability of the CES and the usefulness of r_{av}.

6.3.2 The Draper/Smith Example

In this section exponential data from D/S is used from their example illustrating nonlinear fitting. D/S as well as other practitioners show various sophisticated methods for dealing with the problem of non-linear curve fitting. The procedure indicated here gives a simple alternative way to get a feasible fit. When the Pces correlation coefficient was used (usually meaning results close to least squares), the methods of this paper gave essentially the same result that D/S obtained, as desired. The fit from D/S was very good, so CES passed its "stress test." After 12 iterations, the convergence criteria were satisfied giving final estimates of $a = 0.392$ and $b = 0.103$ for Pces and 0.391, 0.107 for r_{av}. For comparison, the D/S results were $a = 0.39$, and $b = 0.102$. Because the fit was so close no additional figures are shown.

7. Correlation Coefficient Estimation of the Parameters of Univariate Distributions

This last section shows the generality of the correlation coefficient method by adjusting equation (7) for use in the estimation of the parameters of univariate distributions. The response variable is replaced by some form of the empirical distribution function F_n and the estimating function F is the theoretical cumulative distribution function. The parameters b of the distribution function F are varied to find the minimum s. In addition, here the residuals en masse are minimized relative to the edf $F_n(x)$, so $F_n(x)$ appears in place of the earlier y. The $F_n(x)$ needs no superscript, of course, as it is intrinsically ordered. The adjusted equation is

$$r(F_n(x), (F_n(x) - F(x, \hat{b}))^0 - sF_n(x)) = 0. \qquad (8)$$

It has been shown that the solutions to equation (8) behave reasonably with respect to location and scale changes when a distribution that can be standardized, such as the normal, is used. For such distributions, however, the parameter estimation technique related to equation (2) is an alternative. A paper on this idea, Gideon and Rothan (2007), has been prepared and has been submitted for publication; it is also on the Internet. However, for distributions like the gamma, the proposed method of equation (8) is appropriate. One example is given for the gamma distribution with 25 randomly generated observations with parameters scale = 2 and shape = 3. The estimates were 1.19 for scale and 4.16 for shape. The results are summarized in Figures 3 and 4. The relevant R procedure is given in the Appendix. It is probably worth noting that nlm has some trouble staying in the appropriate solution space when working on certain non-linear problems. It seems that choosing a suitable starting value is critical. The example presented gave a good fit immediately.

8. Conclusion

The CES provides a very general method to estimate parameters in a number of different settings and with different estimation criteria. The CES has a multitude of possibilities; many have presented themselves just in putting together this paper. Certainly further study needs to be undertaken, but the area is so broad that definitive study by a single person is virtually impossible. A profitable study also needs better computational ability in R for the implicit equations of CES.

It is apparent that the R routine nlm needs to be fine-tuned (or a new routine created) for solving implicit equations involving non-linear functions, such as most distribution functions. The current form does not allow the CES method of estimation to work flawlessly when a location parameter is part of the minimization of equation (7). In running many simulations it was clear that a simple shift in location would have given the minimization technique a better solution. A work around is to include a constraint that allows the zero on the vertical axis of the residual plot to be centered within the residuals. Observe that this is the case for Figures 1 and 3. Also there were problems with nlm keeping the iterated values of the parameters within a feasible solution space; it is very

sensitive to initial values. No problems seem to occur with the R routines and the fitting of linear models when no location parameters were involved in the minimization.

A second improvement would be for the nlm to generalize its technique so that nonparametric correlation coefficients can be included as estimators. Estimation with GDCC was run for many years with a numerical system using a C program that obtained the centered point of a solution interval with never a problem of convergence. So the preferable nlm would also include centered solution points. This most likely would allow convergence of the MAD method as used in Table 3.

It is apparent that CES with just Pces rivals least squares, but the method can be used with all correlation coefficients (both continuous and nonparametric) yielding a unified general estimation system that can be used profitably in many diverse areas. Over the years GDCC, which displays robustness, was successfully incorporated into many areas of estimation, such as time series, general linear models, and of course nonlinear regression and estimation of parameters for a particular density type. See Sheng, 2002 for non-linear regression, time series and general linear models. This leads one to believe that any correlation coefficient could be similarly profitably employed as shown in this paper by r_{av} and the results in Table 1. Incidentally, all the necessary machinery involving R and r_{av} in the estimations of this paper are included so that anyone could reproduce this work. Further asymptotic inference on the RET method for multiple regression is given in the papers Gideon (2008) and Gideon, Prentice, and Pyke (1989).

9. Appendix: R-program outline

(a) Simple linear regression using *uniroot*

```
Let f = function(x,y)  {  … } # in curly brackets define a correlation coefficient on data
#(x,y)
let fslp = function(b,x,y)  f(x,y-b*x)  # R function to be used for regression, solve for b
# The next line estimates a slope, slp, with correlation coefficient f
slp= uniroot(fslp,c(l,u), x=x1, y=y1)$root  # using R function uniroot to find a root.
# (l,u) is a pair of lower and upper points so that fslp has opposite signs at l and at u
# (x1,y1) is the data for the regression
# Now for an intercept,
let int = median(y1-slp*x1)
# So the intercept and slope for correlation coefficient f is (int,slp)
```

(b) Estimate of scale or error of the regression, also using *uniroot*

```
# Next comes the estimate of  s_res/s_y  as an entity, labeled s,
# First compute residuals
res = y1- (int + slp*x1)  # and enter them in a regression with the y variable, y1
s = uniroot(fslp, c(l,u), x = sort(y1), y = sort(res) )$root
# So s is the slope of the regression of ordered data, and
```

$\sqrt{1-s^2}$ estimates the regression correlation coefficient

(c) The minimum SD program using *optimize* (selects b to find the minimum s for a simple linear regression)

```
ftest = function(b,x,y) {y3 = sort(y-b*x)
        s = uniroot(fslp,c(-1,2),x = sort(y), y = y3)$root
        return(s)  }
out1 = optimize(ftest, c(0,1), x = x1, y = y1)
out1$objective  #contains the minimum s and
out1$minimum  # contains the slope which gives the minimum s
```

(d) Multiple linear regression, using *uniroot* and *nlm*
Let y1 be the response data, and XM the n x k matrix of regressor variable data where there are k variables and the sample size is n. Again let f and fslp be as above in (a) and let b be the notation for the vector of regression coefficients not including intercept

(d1) Optimization Technique (OT) using *nlm*

```
# define a function g of the regression coefficients to be used with R routine nlm.

g = function(b)  {
        s = uniroot(fslp, c(l,u), x = sort(y1), y = sort(y1 - XM%*%b))$root
        return(s)  }
# Note: b was 6 dimensional in the simulations and 2 and 6 in the C/P analysis
# The output for the multiple regression is obtained by
out = nlm(g, initialb)  # initialb contains the initial values of the regression coefficients, b
# In the paper examples of f were Pearson's correlation coefficient and r_av.
out$estimate  #  contains the slopes b indexed by  [i],
out$minimum  # contains the minimum s
int = median(y1 - XM%*% b)
```

(d2) Regression Equation Technique (RET) using *uniroot*

```
# The RET method requires the R computer notation XM[,-i] to delete the i^th column of
# matrix XM. This is needed to obtain the Gauss-Seidel solution to the regression
# equations (3).  Again f is the correlation coefficient being used.
# can use least squares method to compute an initial b value

while(  de> 0.005 & ct <250 & ctcor > 0.01) { bp = b
        for(i in 1:k) { XMS = XM[,-i]
                bs = b[-i]
                ys = y1 - XMS%*%bs
                b[i] = uniroot(fslp,c(bl[i],bu[i]), x = XM[,i], y = ys)$root
                }
                de = sum(abs(bp-b))  # the total change in the coefficients
```

```
        ct = ct +1    # a counter that is initially zero
        yres = y1 - XM%*%b  # the updated resdiduals.
        for(i in 1:k)  bcor[i] = f( XM[,i],yres )
        ctcor = sum( abs(bcor))  }    # this is the total deviation of the regression
        # equations (3) from zero

#  The three convergence factors are ct, ctcor, and de; various choices can be made.
# Upon exiting the intercept is calculated using the median
int = median(y1 - XM%*%b)  # the intercept of the fit
yhat = int + XM%*%b            # the predicted values of the model
```

Thus, the fitted model estimates are in b and int.
Generally the regression equations (3) (as all numerical calculations) are only solvable to within some tolerance. The convergence measures used here are (1) ct, upper bound on total number of iterations, (2) de, the smallness of the sum total of the absolute value of the changes in the slopes, and (3) ctcor, the smallness of the sums of the absolute values of the correlations of the regressor variables with the residuals. The necessity of each of these has been observed; there may be some overarching convergence measure that is yet to be found.

(e) Nonlinear estimation using *uniroot* and *nlm*

Let the data be in x and y. Let ysort = sort(y). Now define a function, g2, for the estimation. Only r_{av} was used in our trials. Any non-linear function can be estimated with varying number of parameters, but given is the exponential outlined in the text with two parameters.

```
g2 = function(b)  {
        s = uniroot(fslp,c(0,1), x= ysort, y = sort(y-b[1]*exp(-b[2]*x)))$root
        return(s)  }
out = nlm(g2, c(μ,η) , steptol=1e-3)    # μ and η are the initial values of b[1] and b[2]
out$minimum  # contains the minimum s and
out$estimate[1] or [2] #contains the two parameter estimates
```

(f) Univariate distribution estimation of parameters using *uniroot* and *nlm*

```
fn = ecdf(x) # the empirical distribution function of data x, an R function
plot(fn)  # the plot of the ecdf; an addition to the plot is below
g1 = function(b)  {
        s = uniroot(fslp, c(0,1), x=fn(x), y = sort(fn(x)-F(x,b[1],b[2])))$root
        return(s)  }  # F is the theoretical d.f. under consideration assuming two
        # parameters and  fn(x) gives the i/n increments of the ecdf
out = nlm(g1,c(μ,η))  # μ and η are the initial values of b[1] and b[2]
out$minimum  #contains the minimum s
out$estimate  # contains the final values of b[1] and b[2]
```

lines(x, F(x,b[1],b[2]), type = "l") # puts the fitted cdf on the ecdf plot

Now we plot the outcome of the minimization, i.e.
sorted residuals $(fn - F)^o$, versus fn . The slope of the fit is the minimum s.

```
yres = sort(fn(x)-F(x,b[1],b[2]))
ss = out$minimum
int = median(yres - ss * fn(x))
plot(fn(x),yres) ; abline(int,ss)  # the final iteration plot
```

Figure 1: Exponential Inference

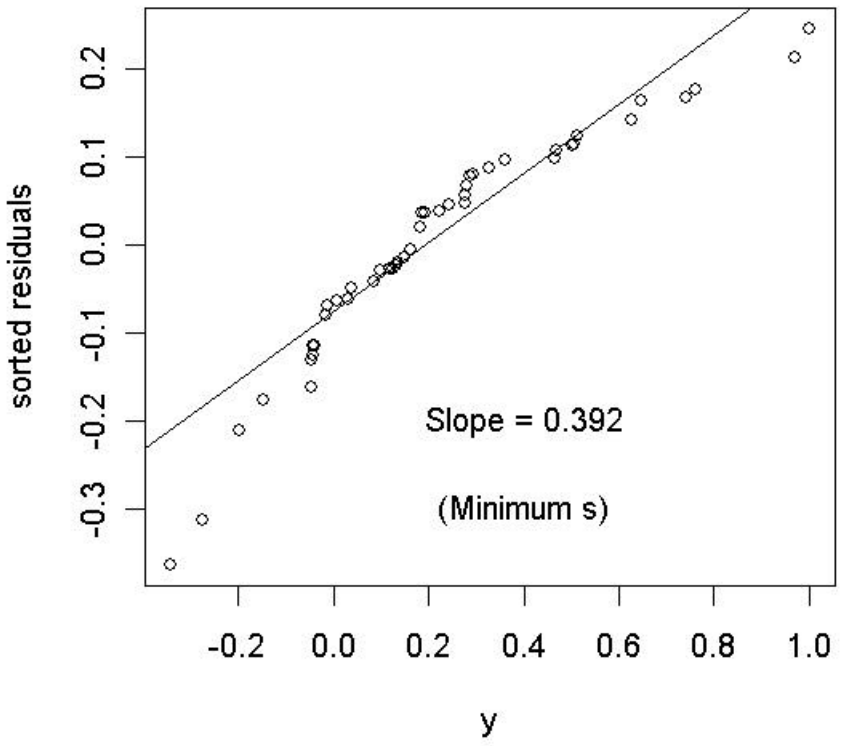


Figure 2: Exponential Curve Fit

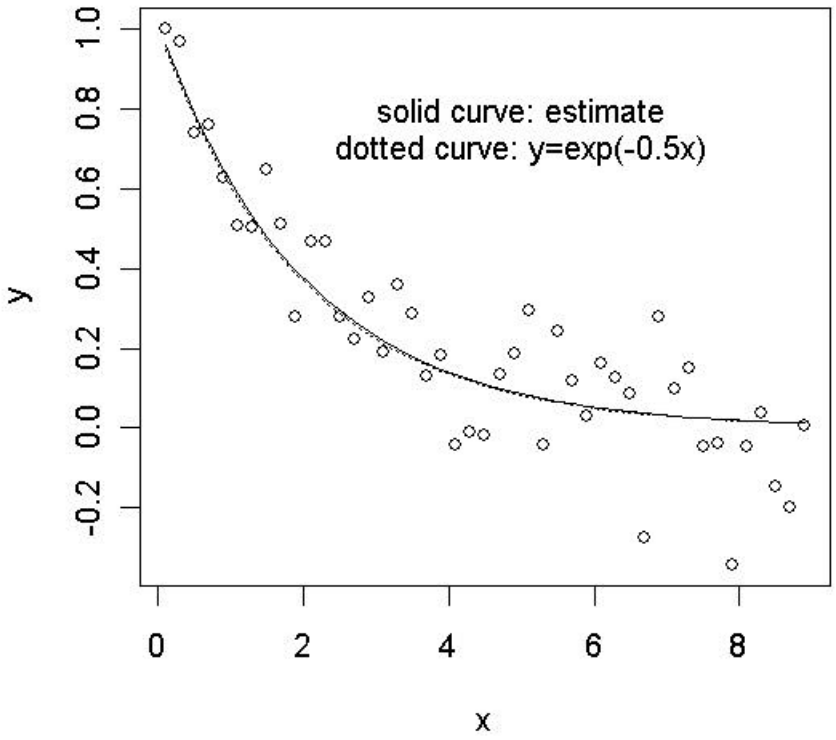


Figure 3: Gamma Inference

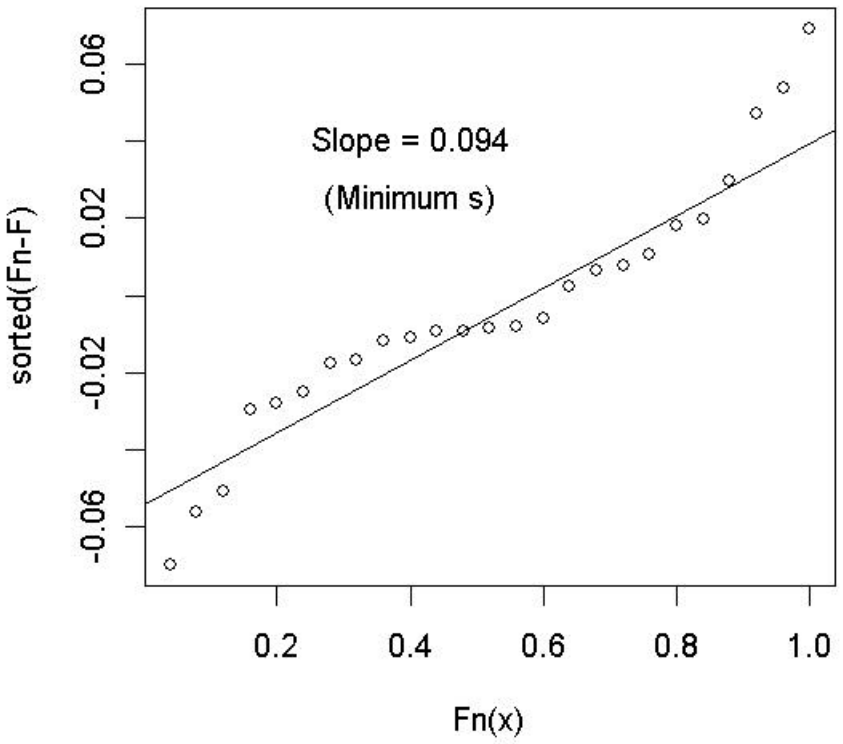
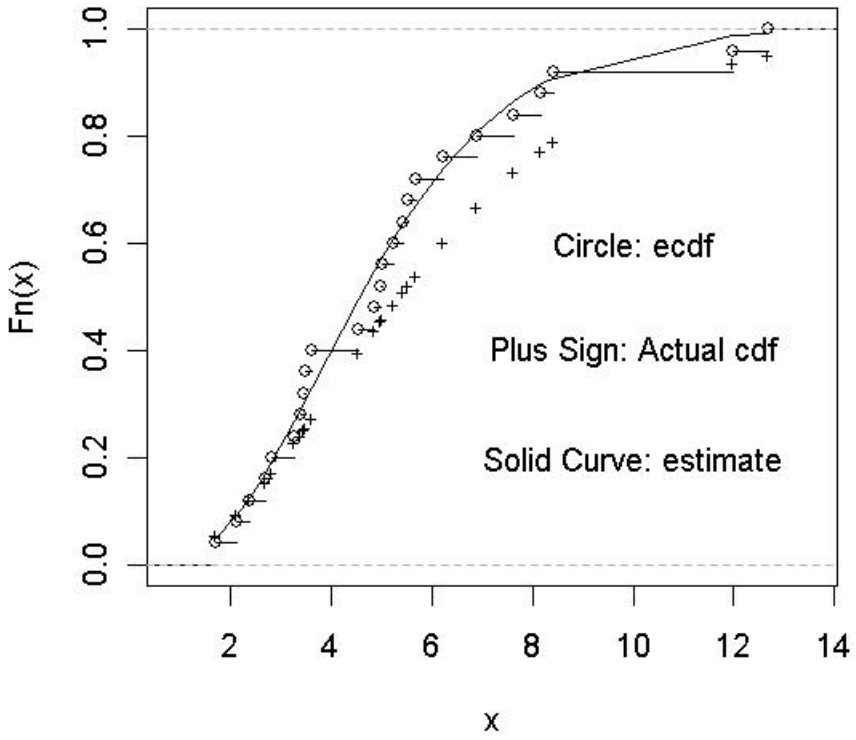


Figure 4: Gamma Estimation



10. References

Chatterjee, S. and Price, B. (1977), Regression Analysis by Example, Wiley & Sons Inc., N.Y.

Draper, N.R. & Smith, H. (1966, 1981, 1998), Applied Regression Analysis, Wiley & Sons Inc., N.Y.

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

Gideon, R.A., and Hollister, R.A. (1987), "A Rank Correlation Coefficient Resistant to Outliers," *Journal of the American Statistical Association*, **82**, no.398, 656-666.

Gideon, R. A. (2007), "The Correlation Coefficients," *Journal of Modern Applied Statistical Methods*, **6**, no.2, 517-529.

Gideon, R.A., Prentice, M.J., and Pyke, R.(1989), "The Limiting Distribution of the Rank Correlation Coefficient r_g", in *Contributions to Probability and Statistics* (Essays

in Honor of Ingram Olkin), ed. Gleser, L.,J., Perlman, M.D., Press, S.J., and Sampson, A.R., New York: Springer-Verlang, pp. 217-226.

Gideon, R. A. and Rothan, A. M., CSJ (2007). "Location and Scale Estimation with Correlation Coefficients." Communications in Statistics-Theory and Methods, under review.

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.

Gini, C. (1914). L'Ammontare c la Composizione della Ricchezza della Nazioni, Bocca, Torino.

Kendall, M.G., and Gibbons, J.D. (1990), *Rank Correlation Methods*, 5th ed. Oxford University Press, or also Kendall, M.G. (1962), *Rank Correlation Methods*, 3rd ed. Hafner Publishing Company.

Pearson, K. (1911), "On The Probability That Two Independent Distributions Of Frequency Are Really Samples From The Same Population", *Biometrika*, **8**, 250-254.

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

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

Spearman, C. (1904), "The Proof And Measurement Of Association Between Two Things," *American Journal of Psychology*, **15**, 72-101

Web site www.math.umt.edu/gideon.