THE MINIMIZATION PROCESS IN THE CORRELATION ESTIMATION SYSTEM (CES) COMPARED TO LEAST SQUARES IN LINEAR REGRESSION

RUDY A. GIDEON

Abstract. This presentation contains a new system of estimation, starting with correlation coefficients, that rivals least squares and for much data does better. One example of SAT and ACT data is used to illustrate minimization through the Correlation Estimation System (CES) in a two-variable linear regression; in this example the CES results appear to be a better representation of the meaning of the data. This result is completely typical; it was not cherry-picked. If you are a least squares-bible toting statistician then what you see in this presentation is blasphemy, but if you are a more secular statistician then you may appreciate a rival estimation system that should be widely used.

Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + error$, with b_i as the estimate of β_i, $i = 1, 2$.

In least squares regression procedures the derivatives involved in the minimization process lead to Pearson's correlation coefficient (CC) equations where the explanatory variables are made to have zero correlation with the residuals. Thus, the minimization of the sum of squares of the residuals and the normal equations are intimately connected. Solving the equations for the regression coefficients gives the "best" estimation. In CES the two processes (minimization and zero correlations) are only approximately equivalent. That is, the CC made to have zero correlation of the xs with the residuals is not exactly equal to the minimum of some function of the residuals. The estimates should be close and they were in the cases studied using the Greatest Deviation CC (GDCC). On the other hand, finding the coefficients for a minimum as described below gives values only approximately equal to those produced by the zero correlation method. This exposition emphasizes the new general minimization case. See the CES overview paper for the zero correlation case,"Obtaining Estimators from Correlation Coefficients: The Correlation Estimation and R," *Journal of Data Science* 10 (2012).

It is apparent in the two graphs that least squares and the CES minimization with Pearson's CC are essentially equivalent. However, the analyzes using GDCC are noticeably better. Because CES can use other CCs, it should be clear that CES is a very powerful and general method of estimation.

The CES Minimization for Regression

Let y be the response variable with, say, one or two predictor variables, x_1 (only x if just one variable) and x_2, using the model above. To estimate coefficients β_1 and β_2 from data in the vector form $(\underline{y}, \underline{x}_1, \underline{x}_2)$, let $\underline{y} - b_1\underline{x}_1 - b_2\underline{x}_2$ be the uncentered residuals in which b_1 and b_2 are to be determined by the CES minimization process.

Now let $r(\underline{x}, \underline{y})$ be a CC — Pearson, GDCC, Gini, Absolute Value, Median Absolute Deviation (MAD), Spearman, Kendall. See "The Correlation Coefficients," *Journal of Modern Applied Statistical Methods*, 6, no. 2 (2007). Any of these can be used to find a minimum slope as explained below. There is an R-code program available from the author that defines each of these CCs; it is called Cordef&reg, and an outline is included below. It is #18 on my website. The max-min procedure for tied values, as defined in several papers, is essential for the rank based correlation coefficients; it allows estimation in all cases and is included in Cordef&reg. For Pearson's CC the estimated betas are very close to the least squares values and in fact, simulations show that the distributions of b are the same. The only exception is for small sample sizes and low correlations. In these cases, the estimated βs can be somewhat different.

To start the actual procedure, let res^0 be the sorted residuals for selected b_1 and b_2 (or b if only one x). Plot res^0 versus $\underline{y}^0$ with $\underline{y}^0$ on the horizontal axis. Note that y is ordered independently of the residual ordering. The goal is to choose b_1 and b_2 so that the sorted residuals, although monotonically increasing, are small in absolute value and increase as little as possible by some criterion. Note that if $b_1 = b_2 = 0$, the plot becomes a $(\underline{y}, \underline{y})$ plot and the slope of a line fitted to the data is 1. On the other hand if $b_1 = \beta_1$ and $b_2 = \beta_2$ and there is no error then res^0 is a constant vector and the slope of a line fitted to this data is 0. (An intercept would make this vector the zero vector) The goal then is for non-zero error to determine b_1 and b_2 so that a linear regression line using $res^0 = (\underline{y} - b_1\underline{x}_1 - b_2\underline{x}_2)^0$ has the smallest possible slope, somewhere between 0 and 1. Recall that in least squares, $1 - SS(res)/SS(y) = (SS(y) - SS(res))/SS(y)$, the explained variation divided by the total variation, is the coefficient of determination. In the CES setting, by minimizing the slope of a line fitted to the $(\underline{y}^0, res^0)$ data, the ratio of the unexplained variation to the total variation, i.e. σ_{res}/σ_y, is estimated directly as s, the slope of the line. Thus, $1 - s^2$ can be considered to be the CES coefficient of determination.

An Outline of the R Optimization Instructions for Simple Linear Regression

Let $rfcn = function(b, x, y) \quad r(x, y - b * x)$. This function is monotonically decreasing in b.

For the zero correlation process *uniroot* is used to find b such that $r(\underline{x}, \underline{y} - b\underline{x}) = 0$; that is, b is the zero solution. Set $rslp = uniroot(rfcn, a, x = x, y = y)\$root$, where a is an interval in which the solution, b, should be found.

For the minimization process using *uniroot*, the y assignment in $rslp$ is $y = res^0$ and the x assignment is $x = \underline{y}^0$, the response data. To summarize, $res^0 = (\underline{y} - b\underline{x})^0$ and b is unknown so it is estimated by choosing b such that $(\underline{y} - b\underline{x})^0$ is as small as possible with the selected CC, r. This involves finding the minimum value of s below.

The following R-function sets up the use of the *optimize* command.

$rtest = function(b, x, y) \quad \{y1 = sort(y - b * x)$
$s = uniroot(rfcn, a, x = y^0, y = y1)\$root$
$return(s)\}$
Here is a quick summary of the procedure, minimize s by choosing b in the equation $r(y, (y - bx)^o - sy) = 0$ solving for s. This CES procedure was used several places in the analysis. Note that the root s is the slope of the line through $(\underline{y}, (\underline{y} - b\underline{x})^0)$. The iteration process for b converges when s is minimized. Finally, this R-code produces the minimization:
$out = optimize(rtest, a, x = x, y = y)$.

An Outline of the R Optimization Instructions for a Two Variable Linear Regression

The dependent variable y is either ACT or SAT averages for the 50 states and District of Columbia. So $n = 51$. The two explanatory variables are state averages for Graduate Rate (GR) and Pupils per Teacher (PPT). The confounding variable, participation rate (PR), was eliminated by simple linear regression of ACT and SAT on PR. CES implemented simple linear regression with GDCC as explained above. So the response variables became the residuals from these two regressions, labeled resACT and resSATAV.

R-code for the two-variable CES linear regression is given. In order to check the process, two functions, one for each x, were run with *optimize* one at a time to monitor the convergence rate. A more general R-code utilizing the Gauss-Seidel iterative process is necessary for a procedure with two or more explanatory variables.

$g2sat = function(b2, x1, b1, x2, y) \quad \{y2 = sort(y - b1 * x1 - b2 * x2)$
$s = uniroot(rfcn, a, x = sort(y), y = y2)\$root$
$return(s)\}$
$g1sat = function(b1, x1, x2, b2, y) \quad \{y3 = sort(y - b1 * x1 - b2 * x2)$
$s = uniroot(rfcn, a, x = sort(y), y = y3)\$root$
$return(s)\}$

Now iterate through the following R-code until b1it and b2it stabilize as solutions for $b1$ and $b2$.
b1it = initial value # change manually each iteration
$out2s = optimize(g2sat, a, x1 = GR, b1 = b1it, x2 = PPT, y = resSATAV)$.
out2s # gives new b2
b2it = initial values # change manually each iteration
$out1s = optimize(g1sat, a, x1 = GR, b2 = b2it, x2 = PPT, y = resSATAV)$.
out1s # gives new b1

The intercepts for all of the regressions were calculated by taking the GDCC measure of centrality on the uncentered residuals. This general procedure is currently being written up. For a sample of size 51 the measure is the average of the 17, 18, 34, and 35 order statistics. This is essentially the average of the 1/3 and 2/3 quantiles. This intercept and the material in the section Mathematical Tools completes the analysis without resorting to classical methods.

What are the implications of the graph comparing LS and CES? First, CES with Pearson's CC appears equivalent to LS. My conjecture is that in all areas of regression CES minimization with Pearson's CC is equivalent to least squares minimization. Second, the comparison of the two minimization processes (LS and CES with GDCC) by plotting the ordered residuals versus the y (response) variable and the line determined by CES minimization, suggests a larger relationship between the explanatory variables and y. The Coefficient of Determination $(1 - s^2)$ is 0.64 for ACT and 0.59 for SAT for CES against 0.50 and 0.26 for LS. The reason for this is that there are 7-10 data points (states) that are somewhat different from the other 44-41 states. The analysis with CES using the robust GDCC gives equal weight to all of the states as opposed to LS that gives too much weight to the 7-10 outlier states. This type of graph may be useful in finding multi-dimensional outliers.

The following reference is to a study of various CCs to determine which worked best and found Pearson's to be best when data is normal, but GDCC to be far superior with non-normal data, particularly with outliers: Maturi, T.A. & Elsayiah, A., "A Comparison of Correlation Coefficients via Three-Step Bootstrap Approach," *Journal of Mathematics Research*, 2-2 (2010).

Mathematical Tools

The mathematical tools necessary for GDCC are listed for a two-variable explanatory case and used to calculated the SEs from the data and appear in the Power Point slides.

The standard errors of the regression coefficients are based on the papers Gideon, R.A. and Rothan, A.M., CSJ, "Location and Scale Estimation with Correlation Coefficients," *Communications in Statistics - Theory and Methods* 40 (2009) and Gideon, R.A., "Using Correlation Coefficients to Estimate Slopes in Multiple Linear Regression," *Sankhya* 72-B (2010).

$N\left(\beta, \frac{\pi^2 r^{ii}}{4n}\sigma_{res}^2\right)$is the asymptotic distribution of the estimate of β using GDCC regression. This holds over the class of multivariate t distributions.
$R = (r_{ii}) = \begin{pmatrix} S_1^2 & S_1 S_2 \hat{\rho} \\ S_1 S_2 \hat{\rho} & S_2^2 \end{pmatrix}$, $R^{-1} = (r^{ii})$
$S_1 = GDCC$ estimate of σ_1, $S_2 = GDCC$ estimate of σ_2
$\widehat{\left(\frac{\sigma_{res}}{\sigma_{x_i}}\right)}$, the slope of the GDCC regression line for the points (res^0, x_i^0), estimates $\frac{\sigma_{res}}{\sigma_{x_i}}$
$\hat{\rho} = \sin\left(\frac{\pi GDCC}{2}\right)$
For one x, the asymptotic distribution of $\hat{\beta} - \beta$ is $N\left(0, \frac{\pi^2 \sigma_{res}^2}{4n\sigma_x^2}\right)$.

Contents of R-program Cordef&reg

The program defines R-functions for the following correlation coefficients. These functions are for both the CC itself and the slope b in a simple linear regression for x-y data using the CC. The letter a denotes a region in which the slope should lie.
(1) The Greatest Deviation Correlation Coefficient (GDCC)

(a) CC: rank based, GDave(x,y)
(b) SLR: GDfcn(b,x,y)
(c) SLR example, GDslp = uniroot(GDfcn,a,x=x,y=y)$root.

(2) Kendall's Tau, rank based
(a) CC: KENtau(x,y)
(b) SLR: Kenfcn(b,x,y)
(c) SLR example, Kenslp = uniroot(Kenfcn,a,x=x,y=y)$root.

(3) Gini CC, rank based
(a) CC: Gini(x,y)
(b) SLR: Ginfcn(b,x,y)
(c) SLR example, Ginslp = uniroot(Ginfcn,a,x=x,y=y)$root.

(4) a continuous absolute CC using absolute values, the continuous version of Gini's CC
(a) CC: abscor(x,y)
(b) SLR: absfcn(b,x,y)
(c) SLR example, absslp = uniroot(absfcn,a,x=x,y=y)$root.

(5) a continuous absolute CC using absolute values and medians MADCC
(a) CC: MADcor(x,y)
(b) SLR: madfcn(b,x,y)
(c) SLR example, madslp = uniroot(madfcn,a,x=x,y=y)$root.

(6) a MAD covariance function and its corresponding CC, computed in the same way Pearson's CC is from its covariance function
(a) Covariance: CMAD(x,y)
(b) CC: CORMAD(x,y)=CMAD(x,y)/sqrt(CMAD(x,x)*CMAD(y,y))
(c) SLR: mad2fcn(b,x,y)
(d) SLR example, mad2slp = uniroot(mad2fcn,a,x=x,y=y)$root.

Note MAD CCs can be greater than one.

(7) Pearson's CC in SLR with the CES style
(a) CC: cor(x,y)
(b) SLR: Pfcn=function(b,x,y) cor(x,y-b*x) # as an example how the others are defined
(c) SLR example, Pslp = uniroot(Pfcn,a,x=x,y=y)$root.

Note (7) can be used to define CES for Spearman's CC, but since it is rank based it would need to be defined using the min-max method of breaking ties contained in Cordef&reg

Justification of the CES minimization process in multiple linear regression

Let Y and $X_{p\times 1}$ have a multidimensional normal distribution with mean vector $\begin{pmatrix}\mu_0\\ \underline{\mu}\end{pmatrix}$ and covariance matrix $\begin{pmatrix}\sigma_y^2 & \sigma_{y\cdot x}\\ \sigma_{x\cdot y} & \Sigma\end{pmatrix}$. Let σ^2 be the conditional variance of Y on $\underline{X} = \underline{x}$, so $\sigma_{res}^2 = \sigma^2 = \sigma_y^2 - \sigma_{y\cdot x}\Sigma^{-1}\sigma_{x\cdot y}$, the theoretical residual variance. Also the multiple CC of Y and $\underline{X} = \underline{x}$ is $\rho = \sqrt{1-\sigma^2/\sigma_y^2}$. The quantity $\frac{Y-\mu_0}{\sigma_y}$ has a standard normal distribution, Z. For order statistics, $\frac{Y_{(i)}-\mu_0}{\sigma_y} = Z_{(i)}, i = 1, 2, \ldots, n$ where n is the sample size. For random variable $Y - (\beta_0 + \sum_{j=1}^{p} \beta_j X_j)$ the conditional distribution of $Y|\underline{X} = \underline{x}$ is $N(\mu_0 + \sigma_{y\cdot x}\Sigma^{-1}(\underline{x} - \underline{\mu}), \quad \sigma^2)$ where the regression coefficients $\underline{\beta}' = \sigma_{y\cdot x}\Sigma^{-1}$ and $\beta_0 = \mu_0 - \underline{\beta}'\underline{\mu}$. The square of the

conditional CC is $\rho^2 = 1 - \sigma^2/\sigma_y^2 = 1 - (\sigma_y^2 - \sigma_{y \cdot x}\Sigma^{-1}\sigma_{x \cdot y})/\sigma_y^2 = (\sigma_{y \cdot x}\Sigma^{-1}\sigma_{x \cdot y})/\sigma_y^2$.
$res = Y - (\mu_0 + \sigma_{y \cdot x}\Sigma^{-1}(\underline{x} - \underline{\mu})$ is $N(0, \sigma^2)$. For a random sample of size n let $res_{(i)} = i^{th}$ order statistic of res then $\frac{res_{(i)} - 0}{\sqrt{\sigma_y^2 - \sigma_{y \cdot x}\Sigma^{-1}\sigma_{x \cdot y}}} = Z_{(i)}$ and
$\frac{res_{(i)} - 0}{\sqrt{\sigma_y^2 - \sigma_{y \cdot x}\Sigma^{-1}\sigma_{x \cdot y}}} = \frac{Y_{(i)} - \mu_0}{\sigma_y}$ since Y is $N(\mu_0, \sigma_y^2)$. It follows that
$res_{(i)} = \frac{\sqrt{\sigma_y^2 - \sigma_{y \cdot x}\Sigma^{-1}\sigma_{x \cdot y}}}{\sigma_y}(Y_{(i)} - \mu_0)$.
Taking σ_y under the radical and separating terms gives $res_{(i)} = \sqrt{1 - \rho^2}(Y_{(i)} - \mu_0)$; res^0 is the vector of increasing order statistics.
So the regression line of res^0 on y^0 has a population slope of $S = \sqrt{1 - \rho^2}$ or $1 - S^2 = \rho^2$ which is the coefficient of determination.
Because $\frac{X - \mu_x}{\sigma_x}$ has a $N(0, 1)$ distribution, in a similar fashion $res_{(i)} = \frac{\sigma}{\sigma_x}(X_{(i)} - \mu_x)$ so the regression of res^0 on X^0 has a slope of σ/σ_x where $\sigma = \sigma_{res}$.

The email addresses for Professor Emeritus Rudy Gideon of the University of Montana, Missoula, MT are
gideon@mso.umt.edu and ragideon38@gmail.com.
The website is www.math.umt.edu/gideon
Any request for the R program which sets up the CCs and their simple linear regression use will be honored, and is also available as #18 on the website.