CORRELATION ESTIMATION SYSTEM ON BINARY DATA

RUDY A. GIDEON

Abstract. The Correlation Estimation System (CES) is expanded into the generalized linear model binary data area on three sets of major league baseball data. Because of the versatility of CES three different methods are given. There are some comparisons to the classical method. The statistical computing language R is used.

1. Introduction

In this paper, the generalized linear model is considered with the response variable y being binary. The classical method to fit a model is to substitute the chosen model into the log-likelihood function and then derive scores (partial derivatives). The information matrix is derived from the scores. Finally, the maximum likelihood estimates are derived from the scores and the information matrix in a weighted least squares iterative process. Alternative Correlation Estimation System (CES) estimation methods (Gideon, 2012) are developed from this setting.

The notation of this paper is based on that in McCullagh and Nelder (1991). Let n_i, y_i, π_i, and η_i be the sample size, the number of successes, the probability of a win, and the link function at i, $i = 1, 2, \ldots, k$. The equation to be solved (for β) is

$$X^T W X \beta = X^T W Z \tag{1}$$

where
$W = diag\{n_i(d\pi_i/d\eta_i)^2/\pi_i(1-\pi_i)\}$ and $z_i = \eta_i + \dfrac{y_i - n_i\pi_i}{n_i}\dfrac{d\eta_i}{d\pi_i}$. Generally the solution is found iteratively.

Key words and phrases. absolute value correlation, baseball data, generalized linear models, logistic, probit, log-log, median absolute deviation correlation.

This paper will use the link function $\eta_i = \beta_0 + \beta_1 x_i$ where x_i is a component of the independent variable x. The matrix X is $k \times 2$ with the first column all 1s and the second column the x_i. The last section gives an extension to two explanatory variables.

One of the CES methods is to rewrite equation (1) and use a correlation coefficient (CC) to solve it. Because W is a diagonal matrix, the equation can be rewritten as

$$(X^T W^{1/2})(W^{1/2} X)\beta = (X^T W^{1/2})(W^{1/2} Z). \tag{2}$$

This form makes it clear that this is a multiple linear regression problem. Starting values for β are used to compute the η_i and π_i (for the chosen model) and then the technique in Gideon (2010, 2012) is used to obtain new estimates of β. This is what Sheng (2002) did for the beetle mortality data in Dobson (1990). Sheng used the Greatest Deviation CC (GDCC) from Gideon (2007) with good results; his dissertation also shows that GDCC is robust. Sheng's work relied on the work in Hollister (1984), Gideon and Hollister (1987), and Rummel (1991). It is not necessary to review all of this to understand the technique.

Another CES technique uses routines in R to solve for β when the score functions are set to zero. The score functions involve (i) centering the data using the median or mean and (ii) using a CC to make the independent and the residual vectors have zero correlation.

The third method, called the scale method, uses the CDF (cumulative distribution function) and modifies the CES density estimation technique illustrated in Gideon (2012). The original idea is found in Gideon and Rothan (2011) where CES estimates scale variables. The model CDF is estimated directly by the CES scale technique which employs ordered residual data, $(\hat{\pi}_i - y_i/n_i)^o$, related to the x-data. Throughout, the superscript o indicates ordered data.

Three general linear models (GLM) were compared: Logistic, Extreme Value or Log-Log, and Probit. The models were fit by classical and CES methods for the Atlanta Braves and Seattle Mariners baseball data. The variables are dh, which is the difference in hits (team hits minus opponents' hits) for each game, or dw, which is the difference in walks (team walks minus opponents' walks) for each game and y, which is either the number of wins (for the analysis of the one season statistics for the Braves and the Mariners) or is the winning ratio (for the analysis involving the 2012 team statistics). The second independent variable, dw, is first introduced in Section 6 for the 2012 data to show the extension to the multiple variable case.

It is expected that as the independent variable dh increases, the probability of winning must increase. Also when $dh = 0$ the probability of winning should be near 1/2 and the CDF should be symmetric about 1/2, so the Probit model might be preferred. In the case of the 2012 baseball season, the average winning ratio for each team is used along with the mean value of "total hits for" minus "total hits against" for the 162 games of the season. This data was used so as to compare single team results with those for the thirty teams combined. Also note that dh only assumes integer values for the Braves and Mariners, whereas for the 2012 season, averages for the 30 major league baseball teams are used and hence are not integers.

The 2009 Seattle Mariners had a record of 85-77. They were third in the American League West, with 640 runs and 1430 hits. The 1992 Atlanta Braves had a record of 98-64. They were first in the National League West with 682 runs and 1391 hits. They beat Pittsburgh 4 games to 3 in the National League playoffs and lost to Toronto 4 games to 2 in the World Series. Thus, this data set has 175 games. The one season data for Atlanta and Seattle appear in Table 1. Note that below negative 7 and above positive 7 the probabilities of a win are 0 and 1 respectively; this suggests that GLM is appropriate. Also, the fitted data was grouped accordingly in these regions. The 2012-season summary statistics for the thirty major league baseball teams are given in Table 2; note that Seattle is in the American League West in 4^{th} place, ALW4. The 2009 and 1992 data were obtained on a daily basis by the author from box scores from a newspaper. The 2012-season data were obtained from the Major League Baseball official web site.

While other CCs could have been used, for brevity, in this paper only the Absolute Value CC (ABS) and Median Absolute Deviation CC (MAD) are discussed. However, comparisons with Pearson's CC are made. The Absolute Value CC is defined by:
for $SA_x = \sum |x_i - \bar{x}|$ and similarly for SA_y, let

$$r_{av} = \frac{1}{2}\Big(\sum \left|\frac{x_i - \bar{x}}{SA_x} + \frac{y_i - \bar{y}}{SA_y}\right| - \sum \left|\frac{x_i - \bar{x}}{SA_x} - \frac{y_i - \bar{y}}{SA_y}\right|\Big). \tag{3}$$

The MAD CC is defined by:

$$r_{mad} = \frac{1}{2}\Big(med\left|\frac{x_i - med(x_i)}{MAD_x} + \frac{y_i - med(y_i)}{MAD_y}\right| - med\left|\frac{x_i - med(x_i)}{MAD_x} - \frac{y_i - med(y_i)}{MAD_y}\right|\Big), \qquad (4)$$

where $MAD_x = med\,|x_i - med(x_i)|$ and similarly for MAD_y. The connection between this CC and the MAD scale estimate is given in Gideon (2007). In Gideon (2012) both of these CCs are shown to be good estimation tools. Further, both of these are more robust than least squares and r_{mad} (or MAD) is more robust than r_{av} (or ABS).

2. The Semi-Classical Method

As explained in the Introduction, equation (1) is rewritten as (2) so that this method mimics the classical maximum likelihood method. Letting $X^* = W^{1/2}X$ and $Z^* = W^{1/2}Z$, the equation to solve is

$$X^{*T}X^*\beta = X^{*T}Z^*, \qquad (5)$$

which is linear for the β vector. This equation is solved by choosing initial values for β_0 and β_1, computing π_i and η_i (see Section 3 for the CDFs of the three models) and then using a Gauss-Seidel iterative process. Of the three CES methods this one was the most sensitive to initial values. In the cases where there was convergence, the final β_0 and β_1 values were similar to the ones computed in Sections 3 and 4. Some results are given in Table 3.

3. The Score Function Method

The log-likelihood function for binary data is

$$l = \sum_{i=1}^{k}\left(y_i log(\frac{\pi_i}{1-\pi_i}) + n_i log(1-\pi_i) + log\begin{pmatrix} n_i \\ y_i \end{pmatrix}\right).$$

This function is to be minimized for the three models.

The CDFs are:

for the Logisitic model, $\pi_i = 1/(1 + exp(-(\beta_0 + \beta_1 x_i)))$;

for the Probit model, $\pi_i = \Phi(\beta_0 + \beta_1 x_i)$, where Φ is the cumulative distribution function for the standard normal distribution;

for the Log-Log model, $\pi_i = 1 - exp(-exp(\beta_0 + \beta_1 x_i))$.

For each model, the minimization is accomplished by setting the partial derivatives of the log-likelihood function with respect to β_0 and β_1 to zero. For all three models, $\frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \pi_i}\frac{\partial \pi_i}{\partial \beta_j}$, $j = 0, 1$, or $\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{k} \frac{\partial \pi_i}{\partial \beta_j}\frac{(y_i - n_i\pi_i)}{\pi_i(1-\pi_i)}, j = 0, 1$.

For the Logistic model, $\frac{\partial l}{\partial \beta_0} = \sum_{i=1}^{k}(y_i - n_i\pi_i)$ and $\frac{\partial l}{\partial \beta_1} = \sum_{i=1}^{k} x_i(y_i - n_i\pi_i)$.

For the Probit model, $\frac{\partial l}{\partial \beta_0} = \sum_{i=1}^{k} \frac{exp(-(\beta_0 + \beta_1 x_i)^2/2)}{\sqrt{2\pi}}\frac{(y_i - n_i\pi_i)}{\pi_i(1-\pi_i)}$, and $\frac{\partial l}{\partial \beta_1} = \sum_{i=1}^{k} \frac{exp(-(\beta_0 + \beta_1 x_i)^2/2)}{\sqrt{2\pi}} x_i \frac{(y_i - n_i\pi_i)}{\pi_i(1-\pi_i)}$.

For the Log-Log model,

let $C(\beta_0, \beta_1) = exp(-exp(\beta_0 + \beta_1 x_i) + (\beta_0 + \beta_1 x_i))$. Then $\frac{\partial l}{\partial \beta_0} = \sum_{i=1}^{k} C(\beta_0, \beta_1)\frac{(y_i - n_i\pi_i)}{\pi_i(1-\pi_i)}$ and $\frac{\partial l}{\partial \beta_1} = \sum_{i=1}^{k} C(\beta_0, \beta_1) x_i \frac{(y_i - n_i\pi_i)}{\pi_i(1-\pi_i)}$. These are not solved in the classical way, but their meaning must be reformulated for interpretation by CES. Setting $\frac{\partial l}{\partial \beta_0}$ to zero is equivalent to selecting the π_i to center the data around the model, while setting $\frac{\partial l}{\partial \beta_1}$ to zero is equivalent to choosing the π_i so that the model residuals are orthogonal to the independent variables. Note that for two of the models a weight factor appears and must be included in the analysis. This orthogonality requirement is then tantamount to Pearson's CC between x and the residuals being zero. For the centering procedure, the median is used to preserve robustness.

For the Logistic case, the two equations to solve, for β_0 and β_1, are (using just r_{av} to illustrate) $median(y - m) = 0$ and $r_{av}(x, y - m) = 0$, where x is the vector with i^{th} component x_i, y is the vector with i^{th} component y_i, and m is the vector with i^{th} component $n_i\pi_i$. In other words, m is the vector of expectations from the model.

For the Probit model, first observe that the term $\frac{exp(-(\beta_0 + \beta_1 x_i)^2/2)}{\sqrt{2\pi}}$ which appears in both partials is just $\phi(\beta_0 + \beta_1 x_i)$, the standard normal density evaluated at $\beta_0 + \beta_1 x_i$. So for this case, the two equations to solve, for β_0 and β_1, are $median\left(\phi(\beta_0+\beta_1 x_i)*\dfrac{y_i-n_i\pi_i}{\pi_i(1-\pi_i)}, i=1,2,...,k\right)=0$ and $r_{av}(u,v)=0$, where vector u has i^{th} component $\phi(\beta_0+\beta_1 x_i)*x_i$ and vector v has i^{th} component $\dfrac{y_i-n_i\pi_i}{\pi_i(1-\pi_i)}$, both for $i=1,2,...,k$. Note that the argument in the median expression is a type of weighted average as are both arguments in r_{av}.

For the Log-Log model the two equations to solve, again for β_0 and β_1, are $median\left(C(\beta_0,\beta_1)*\dfrac{y_i-n_i\pi_i}{\pi_i(1-\pi_i)}, i=1,2,...,k\right) = 0$ and $r_{av}(u,v) = 0$, where vector u has i^{th} component $C(\beta_0,\beta_1)x_i$ and vector v has i^{th} component $\dfrac{y_i-n_i\pi_i}{\pi_i(1-\pi_i)}$, both for $i=1,2,...,k$.

For the MAD CC, substitute r_{mad} for r_{av}. All three models fit by the score function method with Absolute Value CC are shown in Figure 1 for the Braves and Figure 2 for the Mariners and the β values are given in Table 4. Additionally, the 2012-season Probit fit is shown in Figure 3. This fit looks linear because season totals are being used and no tails are present.

Some of the R-coding is now given. The Probit Model with R-function

$$f \leftarrow function(x,b0,b1) \quad pnorm(b0+b1*x)$$

is used as an example. Note that in R, the CDF of the normal distribution, Φ, is denoted by *pnorm* and the normal density, ϕ, is denoted by *dnorm*. Also note that n is the vector whose i^{th} component is n_i, the sample size of the i^{th} group. The R-code is:

$fb1 \leftarrow function(x,y,n,b0,b1)\{$
$abscor(dnorm(b0+b1*x)*x,(y-n*f(x,b0,b1))/(f(x,b0,b1)*(1-f(x,b0,b1))))\}$
$fb0 \leftarrow function(x,y,n,b1,b0)\{$
$median((y-n*f(x,b0,b1))*dnorm(b0+b1*x)/(f(x,b0,b1)*(1-f(x,b0,b1))))\}$

In the above R-code, *abscor* is the author-generated R-code for the CC r_{av} as defined in (3). The Gauss-Seidel iterative technique is:

```
$iter \leftarrow 20; c1 \leftarrow 1; c2 \leftarrow 1; i \leftarrow 1$
$b0iter \leftarrow b0; \quad b1iter \leftarrow b1$
$while(i < iter \quad \& \quad (c1 > .01 \quad | \quad c2 > .01)) \quad \{$
$b0n \leftarrow uniroot(fb0, c(-.2, 0.1), x = x, y = y, n = n_i, b1 = b1iter)\$root$
$b2 \leftarrow uniroot(fb1, c(0.2, .35), x = x, y = y, n = n_i, b0 = b0iter)\$root$
$c1 \leftarrow abs(b2 - b1iter); c2 \leftarrow abs(b0n - b0iter)$
$b0iter \leftarrow (b0n + 2 * b0iter)/3; b1iter \leftarrow (b2 + 2 * b1iter)/3$
$i \leftarrow i + 1\}$
```

Note that a reduced step size was used after making many attempts at obtaining convergence.

4. The Scale Method

Gideon (2012) shows how to use CES to estimate a continuous density function. That method is modified here to estimate the binomial response variable probablities as related to the explanatory variable in a general linear model.

The main idea for the scale method is that the CDF of the proposed model and the "empirical" CDF can be easily used with the chosen CC to estimate β_0 and β_1. Theoretically, the proportion of wins increases monotonically as the difference in hits increases, but for actual data this is not necessarily true. This is the reason for the loose use of the term "empirical" distribution function.

Let $F(\beta_0 + \beta_1 * x)$ be the cumulative distribution function of one of the possible approximating distributions – Logistic, Probit, or Log-Log – where the link function is $\eta_i = g(\pi_i) = \beta_0 + \beta_1 * x_i$ and $\pi_i = F(\beta_0 + \beta_1 * x_i) = g^{-1}(\eta_i)$. The pertinent CDFs are given in Section 3. Let F_k be the "empirical" distribution function for the binary data; i.e., if y_i is the number of successes in n_i trials at x_i where $x_1 < x_2 < \cdots < x_k$, then let $F_k = (y_1/n_1, y_2/n_2, \ldots, y_k/n_k)$. The graph of $\{x_i, y_i/n_i\}, i = 1, 2, \ldots, k$ plots the "empirical" distribution function.

The CES method of estimating β_0 and β_1 is to use simple linear regression in the following manner, with the notation mimicking that of R-code. Let $res = F_k - \underline{F}(b_0 + b_1 * x)$, where

$\underline{F}(b_0 + b_1 * x)$ is the vector with i^{th} component $F(b_0 + b_1 * x_i)$. The simple linear regression of res^0 on x (dh) is utilized to obtain estimates b_0 and b_1 of β_0 and β_1. The slope s of this simple linear regression is minimized using R-functions *uniroot* and *nlm*.

The slope s at the minimum represents the point at which $F(b_0 + b_1 * x)$ best estimates F_k. Keep in mind that "best" means measured by CES with the chosen CC. Since the residuals are possibly reordered for each iteration, as b_0 and b_1 are adjusted to estimate F_k, the slope becomes smaller. Ordering the residuals keeps the slope between 0 and 1. A zero slope means, of course, that the estimate is perfect.

To be specific, recall that $\pi_i = g^{-1}(\eta_i) = F(b_0 + b_1 * x_i)$ is the probability of a win for the i^{th} group. Let $y_i/n_i = pct_i$ and $res_i = pct_i - F(b_0 + b_1 * x_i), \quad i = 1, 2, \ldots, k$. Again note that res^o will indicate a vector of data ordered least to greatest. The iterative technique to choose b_0 and b_1 such that s gets minimized for the Probit model is as follows:

```
g ← function(b){s ← uniroot(CCslp, c(−5, 5), x = x, y = sort(pct − f(x, b[0], b[1])))$root
return(s)}
out ← nlm(g, c(−.09, .26))
b0 ← out$estimate[1]; b1 ← out$estimate[2]; slp ← out$minimum
```

Note that f is defined in the R-code in Section 3 for a Probit model. $CCslp$ is a function which gives the slope of a simple linear regression of y on x for some CC by using *uniroot* to determine the slope, s, with CC, r, in the equation $r(x, y - sx) = 0$. That is, for any CC, $CCslp \leftarrow function(s, x, y) \quad r(x, y - s * x)$ is the definition of a R-function giving the slope. Some statistics are now given to compare methods.

The log-likelihood ratio statistic or Deviance is

$$D = 2[l(b_{max}; y) - l(b; y)],$$

where l is the earlier defined log-likelihood function. For the case under consideration this becomes

$$\sum_{i=1}^{k} [y_i log(y_i/\hat{y}_i) + (n_i - y_i) log((n_i - y_i)/(n_i - \hat{y}_i))]$$

where $\hat{y}_i$ denotes the fitted value. In Tables 3, 5, and 6, the column labeled Dev is this statistic. The column labeled Sumdif is the sum of the absolute values of $y_i - \hat{y}_i$. This statistic is used because the classical method minimizes the Deviance using least squares, whereas CES is minimizing a slope in a simple linear regression with a chosen CC. Note that GLM stands for the topic generalized linear models, while *glm* represents an R-routine.

For the Probit fit for the 30 major league teams in which "total hits by" minus "total hits against" per game is compared to the winning ratio, $b_0 = -0.01$ and $b_1 = 0.28$. Compare this fit, shown in Figure 3, to the individual game fits for the Braves and the Mariners from Tables 5 and 6 in which for the Braves, $b_0 = 0.252$ and $b_1 = 0.250$ and for the Mariners, $b_0 = -0.89$ and $b_1 = 0.248$. The slopes of these fits have a nice interpretation and are easily calculated. Because the CDF of the normal distribution is the fitted curve, its derivative is the density function. The difference in hits for the Mariners and Braves is between -4 and +4 about 70% of the time. The Probit fit is nearly linear in this region. Thus we can compute the slope of the fits at a difference near zero to find the importance of a one hit difference. Using the normal density function and the computed estimates the slopes of the three fits are (1) 0.112 for the 30 sets of team statistics, (2) 0.0966 for the Braves, and (3) 0.0985 for the Mariners. So in all of these cases an increase of one hit, for differences between -4 and +4, leads to the probability of winning increasing approximately 10%. It is interesting that the individual teams and the yearly summary data more or less agree on this issue. Incidentally, a least squares linear regression fit to the 2012 data was nearly identical to the Probit fit.

In addition, Figures 1 and 2 show that the Braves had a better record than the Mariners which can be seen by the Braves' win ratio being to the left of the Mariners' (for values of dh in the range from -4 to +4) and not because of the slope parameter. Thus, it may be that the better teams have other factors that let them win more games for given dh values. This was seen but not shown here by plotting Atlanta and Seattle fits on the same graphs. The Braves' fit at $dh = -2$ was about 50% whereas Seattle's was no more than 40%. Looking at $dh = 0$, the Braves probability of winning a game is above 60% whereas the Mariners is only about 50%. These comments agree with the positions of Baltimore and Oakland to the left of the fitted curve in Figure 3, which is based on the score method.

5. WAR and Major League Baseball

An article on major league baseball entitled "WAR is the Answer" by Sam Miller appears in the 4 March 2013 issue of ESPN Magazine. WAR stands for Wins Above Replacement and is a tool used to evaluate both individual players and teams. There are three enterprises that combine a large number of baseball statistics to produce one number, WAR, that is taken as the total worth of a player or team. There can be large disagreements among the three evaluations, as different weights are put on some of the factors. The Baseball Prospectus' version has the Pearson correlation coefficient between WAR and team victories as 0.86 and the writer seems to believe that proves WAR is very good. This number is compared to those produced in this paper.

This paper models the difference in hits as a way to introduce CES to GLM binary data problems. For the 2012 team summary the Pearson CC between the hit difference and winning ratio was 0.78, not much under the 0.86. The content of this paper could be easily adjusted to include bases on balls or walks. Introduce the variable dhw which is "total hits and walks for" minus "total hits and walks against". To evaluate this variable, the correlation of dhw with team victories was computed for all of the correlation coefficients in Gideon (2007)and is found in Table 7. This was done to be sure the nonrobust Pearson CC was not misleading. The correlations were all transformed so as to estimate the correlation parameter ρ assuming the underlying distribution is normal.

The results are all between 0.8090 and 0.9307, so that the Pearson value probably represents the data fairly well. Note that 0.8511 is nearly 0.86 so that WAR is not any better than using just the hits and walks as in this paper. Section 6 treats hits and walks as separate variables.

In addition, the results of the Probit fit with the absolute value CC for both Atlanta and Seattle show that the correlation of dh (difference in hits) and the fitted values $\widehat{dh}$ were all higher than the 0.86 correlation produced by WAR. See Table 8. Thus dh appears to offer good information about MLB.

The Miller article also talked about the Baltimore and Oakland teams in 2012 being atypical. Note that in Figure 3 both of these teams won far more games than the dh factor would predict. These teams are on the left of the prediction curve at the top; thus, it is suggested that the analysis in this paper may help to evaluate and improve the WAR methods for team performances.

6. Expansion of the Scale Method to Two Explanatory Variables

Only the scale method is done because the expansion for the other two methods is straightforward. The concept is possibly best seen by an analogy with classical multiple linear regression in which $SS(res) = \sum_{i=1}^{n}(y_i - a - b_1 * x_{1_i} - b_2 * x_{2_i})^2$, the sum of squares of the residuals, is the generic expression in a two variable multiple linear regression. The extension to more than two variables is obvious. To minimize $SS(res)$ the partial derivatives are taken with respect to the parameters β_1 and β_2, these partial derivatives are set to zero, and then solved for β_1 and β_2 to obtain the minimum.

In CES an iterative technique is used. The ordered residuals, res^0, are first regressed against the ordered x_1 to obtain an estimated β_1 which has minimized the slope of a regression line and then against the ordered x_2 to obtain an estimated β_2 which again has minimized the slope of a regression line. The new estimates are substituted in, the residuals are recalculated, and the Gauss-Seidel process is repeated until convergence, giving an overall minimum. Note that in the process above, the residuals are calculated by associating x_1, x_2, and y in the usual manner. However, in the regression the ordered res are run against the ordered x_1 and ordered x_2. To use this idea for other models, substitute the model expression for the multiple linear regression expression. In the case of the Probit fit, the CDF of the standard normal distribution, $\Phi(a + b_1 * x_1 + b_2 * x_2)$, is used; that is, $res^0 = sort(y - \Phi(\cdot))$. The essential R-code is given and the rest of the process can be completed by the reader. The notation dhs and dws denote the sorted dh and dw.

```
slp = function(b, x, y){abscor(x, y − b ∗ x)}
g1 = function(b){s = uniroot(slp, c(−5, 5), x = dhs,
y = sort(pct − pnorm(a + b ∗ dh + b2 ∗ dw)))$root     # solve for b1
return(s)}
g2 = function(b){s = uniroot(slp, c(−5, 5), x = dws,
y = sort(pct − pnorm(a + b1 ∗ dh + b ∗ dw))$root     # solve for b2
return(s) }
out1 = nlm(g1, c(−.09, .26))
out2 = nlm(g2, c(−.09, .26))
```

The iterative technique uses *out*1 and *out*2 based on functions $g1$ and $g2$ until the minimum for *nlm* becomes small enough or the changes in $b1$ and $b2$ are small enough. As in Section 3, *abscor* is the R-code for the Absolute Value CC as defined in equation (3). The function to be minimized is *slp*, which is defined in terms of *abscor*; the versatility of CES allows any CC to be used in the place of *abscor*. Nonparametric CCs might need different mathematical techniques to achieve the results.

To summarize all of the above, the classical and CES methods of multiple linear regression are compared to CES Probit model with r_{av}. The estimated coefficients and the resulting multiple CC are listed in Table 9.

Recall that the CDF of the Probit Model is $\Phi(a + b_1 * dh + b_2 * dw)$ so that at $dh = dw = 0$, $\Phi(a) = \Phi(0.0811) = 0.532$, which is slightly higher than the classical MLR values. In addition, to compare the slopes of the Probit model to the MLR models in the vicinity of $dh = dw = 0$, calculate the partial derivatives. Specifically, the partial derivative of the CDF for the Probit Model with respect to dh, using the usual ϕ for the standard normal density function, is $\phi(a + b_1 * dh + b_2 * dw) * b_1$ and its value at $dh = dw = 0$ is $\phi(a) * b_1 = 0.3976 * 0.2545 = 0.101$. This is quite comparable with the b_1 values in Table 9. A similar calculation shows that $\phi(a) * b_2 = 0.3976 * 0.793 = 0.0315$; compare this with the b_2 values.

7. Conclusion

This paper extends CES into GLMs; comparisons with classical methods are very favorable. Using Dev, Table 5 gives a comparison of two CCs with CES using the scale method with the classical method. It shows that CES MAD is minimum for Logistic, CES ABS is minimum for the Probit, and CES MAD is minimum for the Log-Log. Again using Dev, Table 6 shows that the classical *glm* gives all the minima. However, when Sumdif is used, the CES ABS method gives a smaller result for the Probit and Log-Log models. Returning to Table 5, Sumdif is smaller for CES MAD (Logistic), CES ABS and *glm* tied (Probit), and *glm* is least for the Log-Log model. Some of these values are close enough so that they do not indicate a statistical difference. However, it is clear that the classical least squares method may not always be the best for real data with its usual noise.

Tables 3 and 4 give the fitted values for the three GLM models for the Braves and Mariners data using the score and semi-classical methods with r_{av}. This is intended to contrast the three CES fitting styles. For the Braves and Mariners data the semi-classical and score methods gave estimates of β_1 reasonably close to that of the scale and quite close to each other. The estimates of β_0, however, vary more widely. The parameter β_1 is estimating the slope of the line through the data whereas β_0 is like an intercept parameter and moves the line right or left. It appears that the data makes the estimates of β_1 a bit more stable than the estimates of β_0. Note that the Dev and Sumdif values in Table 3 are quite comparable to those in Tables 7 and 8. Apparently, small changes in the β_0 estimates do not change the overall fit much. To better evaluate the changes in the estimates of β_0 in the Logistic model for the Braves, compare $\Phi(0.1707) = 0.57$ (scale method), $\Phi(0.2745) = 0.61$ (semi-classical method), and $\Phi(0.6662) = 0.74$ (score method) which are the estimates of the probability of the Braves winning when $dh = 0$.

These values indicate a fairly large change in the probability of a Braves win when they and their opponents have an equal number of hits. Note that all estimates are greater than 0.50 for a team that was first place in their League. This analysis points out the difficultly of fitting by three different techniques, but it may also point out that the usual unjustified belief in classical methods – that give just one numerical value – is suspect.

This paper extends the scale method of estimation into multiple linear regression and GLM. The CES score function method also introduces the use of weighted regression. Because of the high correlation values in Section 5, the GLM techniques of this paper could be used to improve WAR methods for major league baseball team comparisons. In Table 7 the two most robust CCs give the two extreme values which indicate that they may each measure different properties of unusual data. A new estimation technique such as CES provides ample room for further investigation.

Acknowledgement: Carol Ulsafer was invaluable in the editing.

8. References

Dobson, A. J. (1990). *An Introduction to Generalized Linear Models.* New York: Chapman & Hall.

Gideon, R. A. (2007). The Correlation Coefficients. *Journal of Modern Applied Statistical Methods* 6:517-529.

Gideon, R. A. (2010). The Relationship Between a Correlation Coefficient and Its Associated Slope Estimates in Multiple Linear Regression. *Sankhya* 72-B:96-106.

Gideon, R. A. (2012). Obtaining Estimators from Correlation Coefficients: The Correlation Estimation System and R. *Journal of Data Science* 10:597-617.

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 Rothan, A. M. (2011). Location and Scale Estimation With Correlation Coefficients. *Communications in Statistics—Theory and Methods* 40:1561-1572.

Hollister, R. A. (1984). A Correlation Coefficient Based on Maximum Deviation. Unpublished Ph.D. Dissertation, University of Montana, Missoula, MT 59812, full text accessible through UMI ProQuest Digital Dissertations.

McCullagh, P. and Nelder, J. A. (1991). *Generalized Linear Models, 2nd ed. Monographs on Statistics and Applied Probability, 37.* New York: Chapman & Hall.

R Development Core Team. (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rummel, S. 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 UMI ProQuest Digital Dissertations.

Table 1. Braves and Mariners Baseball Data

	Braves			Mariners		
dh	wins	games	ratio	wins	games	ratio
-13	-	-	-	0	1	0
-12	-	-	-	0	1	0
-11	-	-	-	0	2	0
-10	0	2	0	0	2	0
-9	0	2	0	0	2	0
-8	0	1	0	0	4	0
-7	0	5	0	0	4	0
-6	1	6	0.167	0	2	0
-5	1	5	0.200	0	4	0
-4	3	12	0.250	1	7	0.143
-3	5	13	0.385	2	6	0.333
-2	1	10	0.100	3	11	0.273
-1	9	20	0.450	6	12	0.500
0	14	18	0.778	6	16	0.375
1	11	15	0.773	12	20	0.600
2	9	13	0.692	9	15	0.600
3	11	13	0.846	11	14	0.786
4	10	11	0.909	10	13	0.769
5	9	9	1	4	4	1
6	5	5	1	5	6	0.833
7	5	5	1	6	6	1
8	4	4	1	4	4	1
9	2	2	1	1	1	1
10	-	-	-	3	3	1
11	1	1	1	1	1	1
12	1	1	1	-	-	-
13	1	1	1	1	1	1
16	1	1	1	-	-	-
totals	104	175		85	162	

Table 2. 2012 Major League Baseball Team Summary

Team	League	Wins	Hits	Walks	Hits	Walks	Hit Diff
			By	By	Against	Against	Ratio
ARI	NLW3	81	1416	539	1432	417	-0.0988
ATL	NLE2	94	1341	567	1310	464	0.1914
BAL	ALE2	93	1375	480	1433	481	-0.3580
BOS	ALE5	69	1459	428	1449	529	0.0617
CHA	ALC2	85	1409	461	1365	503	0.2716
CHN	NLC5	61	1297	447	1399	573	-0.6296
CIN	NLC1	97	1377	481	1356	427	0.1296
CLE	ALC4	68	1385	555	1503	543	-0.7284
COL	NLW5	64	1526	450	1637	566	-0.6852
DET	ALC1	88	1467	511	1409	438	0.3580
HOU	NLC6	55	1276	463	1493	540	-1.3395
KCA	ALC3	72	1492	404	1504	542	-0.0741
ANA	ALW3	89	1518	449	1339	483	1.1049
LAN	NLW2	86	1369	481	1277	539	0.5679
MIA	NLE5	69	1327	484	1448	495	-0.7469
MIL	NLC3	83	1442	466	1458	525	-0.0988
MIN	ALC5	66	1448	505	1536	465	-0.5432
NYN	NLE4	74	1357	503	1368	488	-0.0679
NYA	ALE1	95	1462	565	1401	431	0.3765
OAK	ALW1	94	1315	550	1360	462	-0.2778
PHI	NLE3	81	1414	454	1387	409	0.1667
PIT	NLC4	79	1313	444	1357	490	-0.2716
SDN	NLW4	76	1339	539	1356	539	-0.1049
SFN	NLW1	94	1495	483	1361	489	0.8272
SEA	ALW4	75	1285	466	1359	449	-0.4568
SLN	NLC2	88	1526	533	1420	436	0.6543
TBA	ALE3	90	1293	571	1233	469	0.3704
TEX	ALW2	93	1526	478	1378	446	0.9136
TOR	ALE4	73	1346	473	1439	574	-0.5741

Table 3. Three Models and the Semi-Classical Method with CES using ABS on Each Team

Braves	Model	b_0	b_1	Dev	Sumdif
	Logistic	0.2745	0.4464	12.16	13.31
	Probit	0.1931	0.2625	11.19	12.73
	Log-Log	-0.1801	0.2923	10.64	11.93
Mariners					
	Logistic	-0.1144	0.4390	6.44	8.80
	Probit	-0.0798	0.2589	5.67	8.50
	Log-Log	-0.3858	0.2715	9.53	10.24

Table 4. Three Models and the Score Function Method with CES using ABS on Each Team

Braves	Model	b_0	b_1
	Logistic	0.6662	0.4432
	Probit	0.4179	0.2744
	Log-Log	-0.0396	0.2899
Mariners			
	Logistic	-0.1331	0.4266
	Probit	-0.0996	0.2505
	Log-Log	-0.6720	0.2962

Table 5. Three Models and Classical Method Compared to Scale Method with CES on Mariners Data

Mariners	Model	b_0	b_1	Dev	Sumdif
CES MAD	Logistic	-0.1506	0.4319	6.36	8.90
	Probit	-0.1283	0.2308	6.04	8.54
	Log-Log	-0.5631	0.2725	7.75	9.65
CES ABS	Logistic	0.0071	0.4162	6.95	9.17
	Probit	-0.0889	0.2480	5.61	8.49
	Log-Log	-0.4189	0.3008	10.22	10.28
glm	Logistic	-0.1479	0.4346	6.37	9.60
	Probit	-0.0881	0.2579	5.65	8.49
	Log-Log	-0.5604	0.2807	7.78	9.60

Table 6. Three Models and Classical Method Compared to Scale Method with CES on Braves Data

Braves	Model	b_0	b_1	Dev	Sumdif
CES MAD	Logistic	-0.2167	0.4209	24.44	22.64
	Probit	-0.1031	0.2448	22.67	21.71
	Log-Log	-0.5602	0.2799	21.93	21.57
CES ABS	Logistic	0.1707	0.3935	14.22	14.73
	Probit	0.2521	0.2494	10.77	11.61
	Log-Log	-0.1978	0.2601	11.56	12.03
glm	Logistic	0.5028	0.4719	10.89	12.11
	Probit	0.2987	0.2770	10.39	12.27
	Log-Log	-0.1506	0.2995	10.54	12.11

Table 7. Correlation of dhw with Victories in Major League Baseball in 2012

Pearson	GDCC	Gini	Kendall	Absolute Value	MAD
0.8511	0.8090	0.8406	0.8396	0.8747	0.9307

Table 8. $(y, \hat{y})$ Correlation, Probit Model using r_{av}

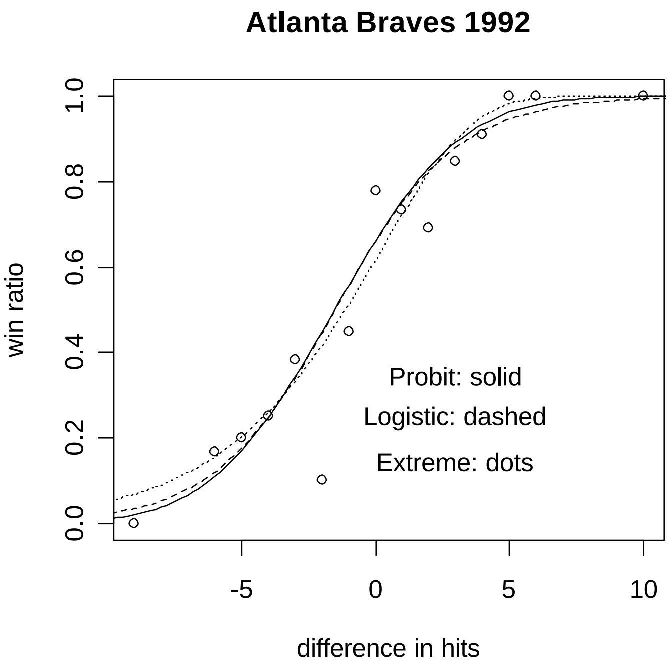
	Pearson	Absolute Value	Absolute Value
		Transformed	Actual
Atlanta	0.9662	0.9816	0.8995
Seattle	0.9452	0.9535	0.8358

Table 9. Two Variable MLR and the Probit Model using r_{av}

Method	a	b_1	b_2	Multiple CC
CES MLR w/ r_{av}	0.4930	0.0918	0.0469	0.886*
LS MLR	0.5000	0.0828	0.0582	0.857
Probit w/ r_{av}	0.0811	0.2550	0.0793	0.868*

* r_{av} has been transformed by $r_{av}\sqrt{2 - r_{av}^2}$

to be comparable to the classical case.


Atlanta Braves 1992
win ratio
0.0
0.2
0.4
0.6
0.8
1.0
-5
0
5
10
difference in hits
Probit: solid
Logistic: dashed
Extreme: dots

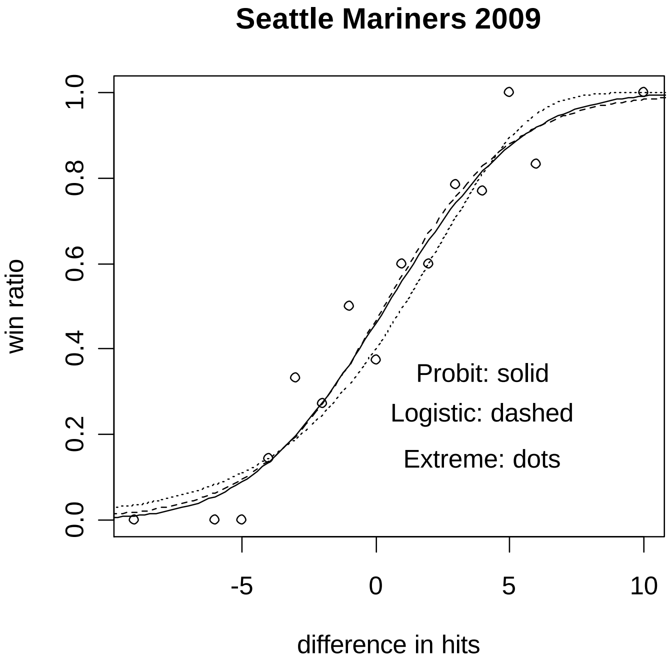
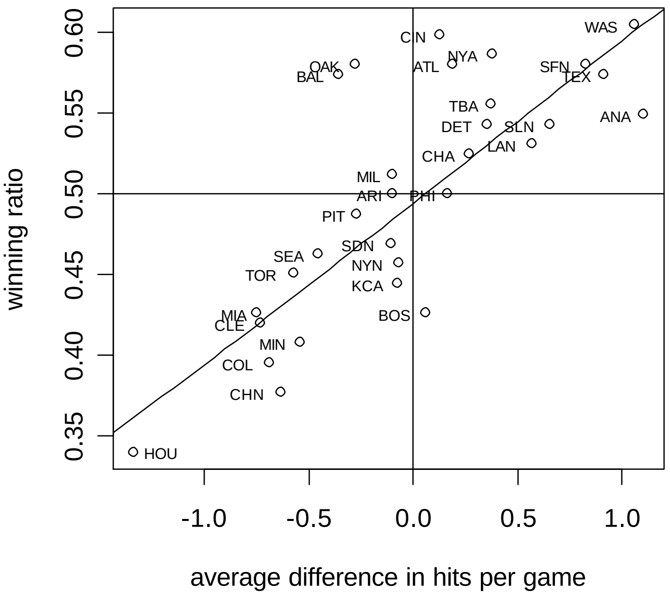

Seattle Mariners 2009
win ratio
0.0
0.2
0.4
0.6
0.8
1.0
-5
0
5
10
difference in hits
Probit: solid
Logistic: dashed
Extreme: dots

Figure 3: 2012 Major League Baseball


WAS
CIN
NYA
OAK
BAL
ATL
SFN
TEX
TBA
ANA
DET
SLN
LAN
CHA
MIL
ARI
PHI
PIT
SEA
SDN
TOR
NYN
KCA
MIA
CLE
BOS
MIN
COL
CHN
HOU
winning ratio
average difference in hits per game