HOW DIFFERENCE IN TEAM HITS AFFECTS THE PROBABILITY OF WINNING IN MAJOR LEAGUE BASEBALL GAMES

RUDY A. GIDEON

Abstract. Four sets of major league baseball data and the generalized linear model on binary data are used to estimate the probability of winning a major league baseball (MLB) game. The statistical computing language R is used to implement CES, a new system of estimation based on correlation coefficients. The winner of a game is predicted based on the difference in hits by the two teams.

1. Introduction

In this paper, the generalized linear model is considered with the response variable y being a win or a loss and the independent variable being x or dh, the difference in hits by the two teams in a MLB game. 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 and the maximum likelihood estimates are derived from the scores by a weighted least squares iterative process. However, in this paper, the estimation of the parameters of the model is done by a new and very general method called CES or the Correlation Estimation System (Gideon, 2012).

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 wins, the probability of a win, and the link function at i, $i = 1, 2, \ldots, k$, respectively. The classical equation to be solved (for β) is

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

Key words and phrases. absolute value correlation, baseball data, generalized linear models, Logistic model, Greatest Deviation correlation.

In the above, capitals are matrices containing the data, $W = diag\{n_i(d\pi_i/d\eta_i)^2/\pi_i(1-\pi_i)\}$ and $z_i = \eta_i + \dfrac{y_i - n_i\pi_i}{n_i}\dfrac{d\eta_i}{d\pi_i}$. The link function η_i involves the covariates in the X matrix. 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 vector x. The matrix X is $k \times 2$ with the first column all 1s and the second column the x_i (dh_i).

It is expected that as the independent variable, difference in hits, dh, increases, the probability of winning must increase. Also when $dh = 0$ the probability of winning should be near $1/2$ and the distribution function should be nearly symmetric about $1/2$, depending on the team quality. 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. For the 2016 MLB season data every game has a home and visiting team. In this scenario, dh is the home team hits minus the visiting team hits. So the model that fits should estimate home field advantage.

The data is given in Tables 1 (2009 Seattle Mariners and 1992 Atlanta Braves), 2 (2012 MLB team summary statistics), and 3 (2016 home team versus visiting team); it is interesting by itself and including it here allows further analysis. 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. Note that where dh is below negative 7 or above positive 7 the probabilities of a win are 0 and 1, respectively, suggesting that a Logistic fit is appropriate. In the 2012 data of Table 2, note that Seattle, in the American League West, was 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- and 2016-season data were obtained from the Major League Baseball official web site. There was one game in 2016 that remained

an end of the season tie. Because of a mix-up, five games in June of 2016 were inadvertently missed; so the 2016 sample size is 2424 instead of 2430.

While other CCs could have been used, for brevity and quality of fit, in this paper mainly the Absolute Value CC (r_{av}) is used. However part of the generality of CES is that with a slight change in the computer code any other CC such as the Median Absolute Deviation (MAD) or the Greatest Deviation can be used and was used in areas needing robustness (see Gideon, 2007 for the definitions of these CCs and 2012 for the general usage). A few comparisons are made between classical approaches and CES methods.

The Absolute Value CC is defined by: for $SA_x = \sum |x_i - \bar{x}|$ and similarly for SA_y, let

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

Absolute value statistical techniques are more robust than least square methods and CES with r_{av} is an example of this.

In the initial examination of the data, it was clear that the probability of winning increases dramatically after dh exceeds five and decreases rapidly as dh decreases from negative five. So, three such models were investigated: Logistic, Probit, and Extreme (Log-Log). There was very little difference and so only the Logistic is presented. Recall that that the Logistic function is $f(x) = 1/(1 + exp(-(\beta_0 + \beta_1 x)))$. Also, several CCs were tried using all these models and r_{av} gave a very good fit. See Figures 1, 2, and 4.

2. The CES Logistic Estimation

Gideon (2012) shows how to use CES to estimate a continuous density function. That method is modified here to estimate the binary response variable probablities as related to the explanatory variable in a general linear model. This is a new minimization process on ordered data.

The main idea for this method is that the cumulative distribution function 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 CDF of the Logistic distribution where the link function is

$$\eta_i = g(\pi_i) = \beta_0 + \beta_1 * x_i \text{ and } \pi_i = F(\beta_0 + \beta_1 * x_i) = g^{-1}(\eta_i) = 1/(1 + exp(-(\beta_0 + \beta_1 * x_i))).$$

Let F_k be the "empirical" distribution function for the binary data; i.e., if y_i is the number of wins 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.

Using the notation mimicking that of R-code, the CES method of estimating β_0 and β_1 by b_0 and b_1 is to use simple linear regression on the ordered residuals, res^0, in the following manner. 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 the estimates b_0 and b_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 $y_i/n_i = wr_i$, the win ratio, and $res_i = wr_i - F(b_0 + b_1 * x_i), \quad i = 1, 2, \ldots, k$. 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 Logistic model is as follows:

$g \leftarrow function(b)\{s \leftarrow uniroot(CCslp, c(-5, 5), x = x, y = sort(wr - f(x, b[0], b[1])))\$root$
$return(s)\}$
$out \leftarrow nlm(g, c(-0.09, 0.26))$
$b0 \leftarrow out\$estimate[1]$; $b1 \leftarrow out\$estimate[2]$; $slp \leftarrow out\$minimum$

Note that the Logistic function is, in R-code,

$$f \leftarrow function(x, b0, b1) \quad 1/(1 + exp(-(b_0 + b_1 * x_i))).$$

$CCslp$ is a function which gives the slope of a simple linear regression of y on x for a particular CC, by using *uniroot* to determine the slope, s, with the 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 residual deviance is

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

where l is the log-likelihood function. For the case under consideration,

$$D = 2\sum_{i=1}^{k} [y_i log(y_i/\hat{y}_i) + (n_i - y_i) log((n_i - y_i)/(n_i - \hat{y}_i))]$$

where $\hat{y}_i$ denotes the fitted value.

There are three columns in Table 4 that evaluate the residuals of the logistic fits. First is the residual deviance, D, second is the column labeled $sum|res|$ which is the sum of the absolute values of $y_i - \hat{y}_i$, and third is the variable s that is a minimum slope in a simple linear regression of the sorted residuals on dh. The D formula does not work for the extreme tails of the data – all losses or all wins at some values of dh. Table 4 deletes these points in the calculation of D. Some effort was made to find out what R used to include these points but without success. Thus, the R computer output of D could not be compared to CES. In Table 4 note that CES with r_{av} has s values all less than ML. Even for D CES has one smaller value. For $sum|res|$ CES had one smaller value. The overall conclusion in these particular data sets is that ML and CES give reasonably similar results.

To emphasize, the s in Table 4 and in the R-code on page 4 is a minimum slope in a simple linear regression of sorted residuals, $res^0 = (wr - f(x, b_0, b_1))^0$ on the independent variable x, or dh, the difference in hits. The function f is the value of Logistic function for the current (b_0, b_1). The slope s is the value such that $CCslp(x, res^0 - sx) = 0$ and this is obtained via the

uniroot command. For a good fit, the residuals are close to zero and so the ordered residuals should increase slowly or the slope s should be as close as possible to zero. The nlm R- command iterates on (b_0, b_1) until the slope s is a minimum. The argument $CCslp$ contains the R-code for the absolute value correlation coefficient for the output given in Table 4.

3. Results

In Table 7 basic location and scale statistics for three of the data sets are given. The GDCC location statistic is essentially the average of the 1/3 and 2/3 quantile statistics. In addition, the CES regression results in Table 4 are used to convert to the standardized form of the Logistic distribution with the mean and standard deviation which appears in Table 7. This allows an interpretation in terms of difference in hits. In Table 7, in the row labeled Location-Logistic under Home Team, the quantity -0.4468 is the mean and the point where the Logistic distribution function crosses the 50% probability level. In other words, the Home Team advantage is almost 1/2 of a hit. Likewise the Braves with -0.8795 had an almost even chance of winning a game with one fewer hits than their opponent. The Mariners, however, at 0.3361, needed about 1/3 of a hit more than their opponents to get an even probability of winning. Note that the ordinary mean does not give this information. The CES Logistic results in Table 4 are plotted in Figures 1, 2, and 4. It is clearly seen that the curves fit the data very well. The CES standard deviation estimates in Table 7 are explained in Gideon and Rothan (2011). Two standard deviations is about nine units in Figures 1, 2, and 4, which encompasses the mostly linear parts of the figures.

From the Logistic fit information in Table 4, the probabilities of winning a game when the number of hits is the same, that is, when $dh = 0$ are, for the Mariners 0.4592, for the Braves 0.5986, and for the Home Team 0.5506. Since the curves are close to linear around $dh = 0$ (see Figures 1, 2, 4) approximate slopes can be calculated. These slopes from dh = -1 to 0 are, for the Mariners 0.1160, for the Braves 0.1124, and for the Home Team 0.1123. Similarly the slopes from $dh = 0$ to +1 are, for the Mariners 0.1207, for the Braves 0.1029, and for the Home Team 0.1083. These slopes are the increases in the probability of winning a game for an increase in dh

of one hit. These numbers seem to confirm the use of the Logistic fit. In the area around $dh = 0$ all the fits are roughly parallel.

So when the visiting and home teams have the same number of hits, the home team wins 55 percent of the time, the so called home team advantage. Note that the first place 1992 Braves won about 60 percent of their games when $dh = 0$, and in contrast, the third place 2009 Mariners won only about 46 percent. Note that in all data sets the increase in winning percentage is 11 percent from dh = -1 to 0. The values from 0 to +1 were just slightly different. So an increase of one hit means at least a 10 percent increase in winning around $dh = 0$. From Figures 1, 2 and 4 the graphs of the distribution functions allow other deductions to be made further away from $dh = 0$. These three figures are very similar, but the interesting thing is that other unknown factors have small but significant changes on the winning probability in the Logistic fit and affect the value of b_0, which moves the curves left or right.

Table 5 contains correlative values for the observed winning probabilities and their predicted value, $(y, \hat{y})$. Note that the CES results in column 2 and the Pearson correlation coefficient are very close together. Table 6 contains the correlations of victories with difference in hits plus difference in walks, dhw, for the 2012 season data. Because of the discreteness of the data the Seattle and Atlanta values do not have enough points to include a two-variable regression of dh and dw, where dw is the difference in walks or bases on balls. So including them as one variable is an option. Here just the correlation coefficients are given to see if there would be improvement. However, the 2012 summary season statistics give an informative result in a two variable regression in Table 8. In Table 8, the row labeled CES(r_{av}) contains the coefficients that are used in Figure 3. The correlations were all transformed so as to estimate the correlation parameter ρ assuming the underlying distribution is from the class of bivariate t distributions.

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 sports analysis groups 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 factors can be weighted differently. 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. The results in Table 6 for dhw 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 for team evaluations is not any better than using just the hits and walks as in this paper. Table 8 gives regression results treating hits and walks as separate variables. Also in Table 6 are two rows that show Bill James' Pythagorean formula (Marchi and Albert, 2014) on runs, the main use, and on hits. Hits are one level back from runs in winning a game, so lower correlations would be expected. Note, however, that the dhw variable correlation with winning percentage is higher than the Pythagorean formula on hits. The fact that all the correlations within each row are very similar means the data is well-mannered.

Table 9 gives six measures of correlation on the 2012 data. The variable wr designates the winning ratio. All the (dh, dw) correlations are around 0.3 which means these are not strongly related variables. As may be expected, the correlation of (dh, wr), 0.80, is higher than the correlation of (dw, wr) which can be anywhere from about 0.3 to 0.55. The two robust correlation coefficients, MAD and GDCC, give the low value of around 0.3. The non-robust correlation coefficients giving larger values means there are a few teams whose (dw, wr) points are different from most of the teams. This motivates the work below.

In Table 8 for the 2012 MLB data the formula $sum|res(LS)| - sum|res(CES)|$ is used for the column headed LS-CES res. When this expression is positive the LS residuals are bigger than those of CES, so LS is not as good as CES and vice-versa. Hence, the two-variable CES fit is better by 0.0359 while for the two one variable fits LS is slightly better. This Table compares LS to CES using the absolute value correlation coefficient. For simple linear regression CES gives a higher coefficient for both dh and dw. Note that for dh the regression coefficients are close to 10 percent as in the regressions for the other three data sets. The results for dw were not done for the other data sets but give a value near 10 percent for the 2012 data. It seems unnatural that the coefficients of dh and dw should be so close. Therefore, a two-variable regression was done to reconcile the importance of the two factors. It is seen that the coefficient of dh retains its value

but that of dw drops by about half. The CES result has the coefficients of dh and dw at 0.0906 and 0.0442 while LS gives 0.0828 and 0.0580, respectively. This shows the relative importance of the two factors. Note that CES gives a greater distance between the coefficients of dh and dw. The fact that this 2012 summary data for dh is very similar to the results on the other three data sets makes it very likely that the results for dw would also be very similar.

Because of the above correlation comments concerning Table 9, the robust correlation coefficient, GDCC, was used to produce three robust regressions to compare to the less robust methods that gave the values in Table 8. The regressions were $wr = 0.4934 + 0.0996dh$, $wr = 0.5006 + 0.1109dw$, and $wr = 0.4937 + 0.0963dh + 0.0380dw$. To do this, CES linear regression used the R-code function GDslp. As seen in Table 8 in the two-variable case, for the absolute value CC, r_{av}, the dh coefficient is about twice the size of the dw coefficient. For the GDCC regression the ratio of the these coefficients is about 2.5. So the robust regression may indicate a slightly greater importance of hits over walks for most of the teams.

4. Conclusion

The fact that four independent data sets gave consistent similar results relating winning percentage to difference in hits indicates that the results are valid. To enhance this conclusion some methods of this paper are compared to some comments in Marchi and Albert (2014). These authors say that in unlikely scenarios the Pythagorean formula gives more sensible winning percentage estimates because linear results can give values greater than one. One of their examples of an unlikely scenario is the 2001 Seattle Mariners with 116 wins to 46 losses. These Mariners had 1637 hits for and 1293 against producing an average of 10.10 hits for and 7.98 hits against, a difference of 2.12 hits. Although this paper does not study this team, the 1992 Atlanta Braves, which was studied, had a record somewhat close to these Mariners. Using the Logistic model in Table 4 or Figure 1, a winning percentage of 0.796 is predicted, compared with the actual Seattle value of 0.716. The 2012 season data can also be used for an estimate, namely 0.4944 + (0.1119)2.12 = 0.732 from Table 8 or Figure 3. An advantage of this approach is that the Logistic model always gives estimates between zero and one.

Marchi and Albert (2014) also cite the 2003 Detroit Tigers with a 42-119 record as another example of an extreme case requiring the Pythagorean formula. For this case, the 2009 Mariners can be used as they are the closest model studied. These Tigers had 1312 hits for and 1616 against resulting in an average of 8.10 hits for and 9.98 hits against, which gives a difference of -1.70 hits. The Logistic model in Table 4 or Figure 2, predicts a winning percentage of 0.271 compared with the actual value of 0.265. Again the 2012 data can also be used for an estimate, namely 0.4944 + 0.1119*(-1.7) = 0.304 from Table 8 or Figure 3.

CES does not need classical statistics to analyze this baseball data, but it was included because those seeing it for the first time need to relate to something they know. CES has been examined in other regression areas with excellent results. This paper originated as a means to evaluate CES methods in yet another area. CES proved exemplary and provided some very interesting baseball results.

Acknowledgement: Carol Ulsafer was invaluable in the editing.

5. 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.

Marchi, M. and Albert, J. (2014). *Analyzing Baseball Data with R. The R Series.* Boca Raton, FL: CRC Press, Taylor & Francis Group.

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

	1992 Braves			2009 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 By	Walks By	Hits Against	Walks Against	Hit Diff 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. Home Team 2016 Baseball Data

dh	wins	games	ratio
-19	0	1	0
-18	0	1	0
-16	0	1	0
-14	0	5	0
-13		6	0
-12	0	8	0
-11	0	18	0
-10	0	13	0
-9	0	29	0
-8	0	51	0
-7	6	73	0.082
-6	4	86	0.046
-5	13	108	0.120
-4	30	175	0.171
-3	37	167	0.222
-2	72	204	0.353
-1	98	191	0.513
0	106	195	0.544
1	149	207	0.720
2	140	185	0.757
3	147	276	0.835
4	137	154	0.890
5	102	114	0.895
6	87	95	0.916
7	57	61	0.934
8	37	37	1
9	20	20	1
10	7	7	1
11	17	17	1
12	9	9	1
13	6	6	1
14	3	3	1
15	1	1	1
totals	1285	2424	0.530

TABLE 4. Logistic Model with Absolute Value CC versus Maximum Likelihood

	b_0	b_1	D	$sum\|res\|$	s
Atlanta, CES	0.4000	0.4548	8.834	1.2484	0.00712
Atlanta, ML	0.5028	0.4719	8.950	1.2229	0.00730
Seattle, CES	-0.1633	0.4858	4.969	1.1494	0.00642
Seattle, ML	-0.1479	0.4346	3.500	1.1802	0.00675
Home Team, CES	0.2034	0.4552	16.240	0.5114	0.00163
Home Team, ML	0.2661	0.4506	14.210	0.4229	0.00173

TABLE 5. $(y, \hat{y})$ Correlation, Logistic Model using Absolute Value CC

	Pearson	r_{av}, Transformed	r_{av}, Actual
Atlanta	0.9738	0.9872	0.9167
Seattle	0.9868	0.9921	0.9351
Home Team	0.9977	0.9989	0.9764

TABLE 6. Correlation of Estimated and Actual Winning Percentage in Major League Baseball in 2012

	Pearson	GDCC	Gini	Kendall	Abs Value	MAD
dhw	0.8511	0.8090	0.8406	0.8396	0.8747	0.9307
Pythagorean on runs	0.9496	0.9335	0.9566	0.9533	0.9707	0.9589
Pythagorean on hits	0.7781	0.7771	0.7650	0.7760	0.7945	0.8517

Non-Pearson values have been transformed to be comparable to Pearson's CC.

TABLE 7. Basic Statistics on Difference in Hits

		Home Team	Mariners	Braves
SD	Classical	4.629	4.777	4.468
	CES(r_{av})	4.605	4.550	4.383
	Logistic	3.98	3.733	3.9888
Location	Mean	-0.1638	0.4938	0.3829
	Median	0	1	0
	GDCC	0	0.5	0.5
	Logistic	-0.4468	0.3361	-0.8795

TABLE 8. Regressions on the Summary Data of the 2012 MLB Season

Method	b_0	b_1 coef of dh	b_2 coef of dw	Multiple CC	LS-CES res
LS MLR	0.5000	0.0828	0.0582	0.857	
CES(r_{av})	0.4957	0.0906	0.0442	0.886*	0.0359
LS SLR	0.5000	0.0974			
CES(r_{av})	0.4944	0.1119			-0.0282
LS SLR	0.5000		0.0902		
CES(r_{av})	0.5006		0.1109		-0.0024

* r_{av} has been transformed by $r_{av}\sqrt{2 - r_{av}^2}$ to be comparable to Pearson's CC.

TABLE 9. Correlations on 2012 MLB Variables

Corr	(dh, dw)	(wr, dh)	(wr, dw)
Pearson	0.3119	0.7797	0.5817
GDCC	0.2588	0.8090	0.3090
Kendall	0.3432	0.7782	0.5790
Gini	0.3400	0.7695	0.5392
MAD	0.3799	0.8126	0.3359
r_{av}	0.3280	0.7971	0.5565

Non-Pearson values have been transformed to be comparable to Pearson's CC.

Figure 1: Atlanta Braves 1992

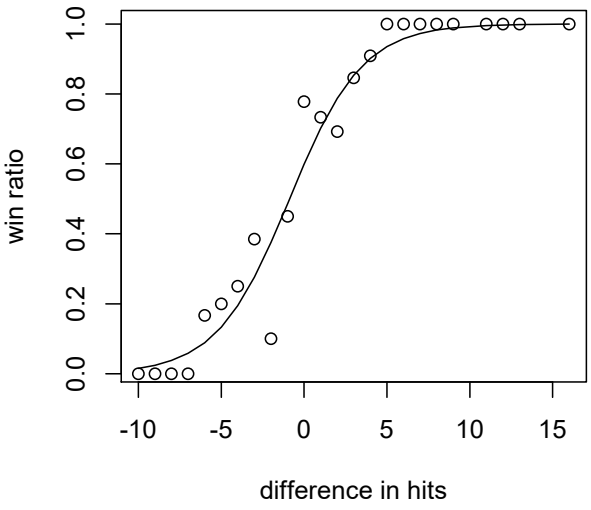

Figure 2: Seattle Mariners 2009

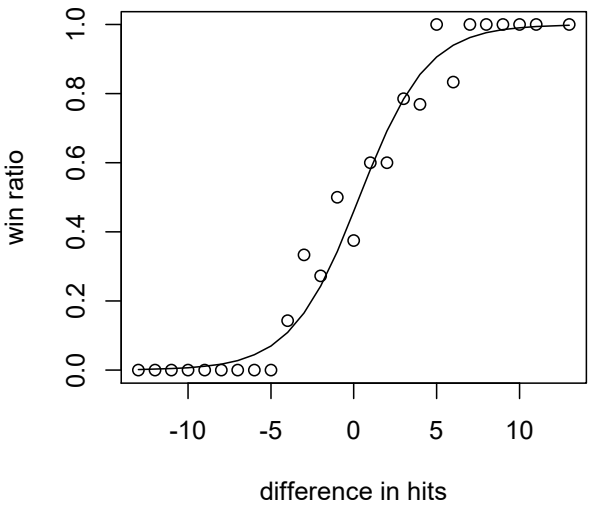

Figure 3: 2012 Major League Baseball

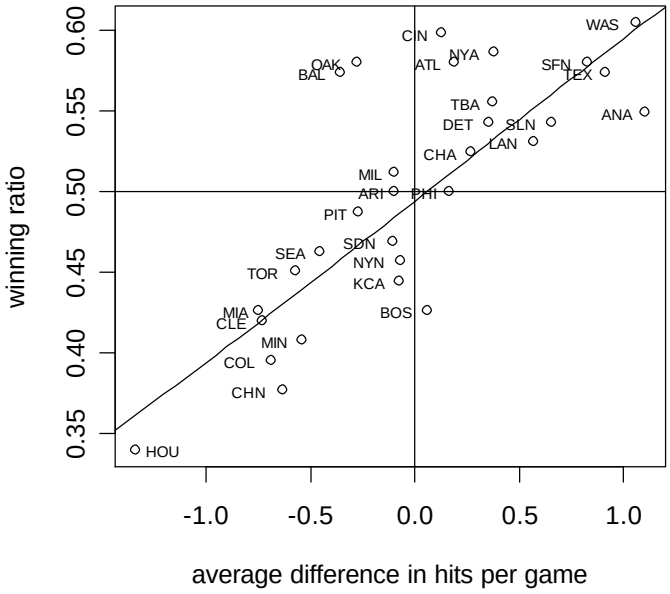

Figure 4: Home Team Advantage 2016