

Teaching Regression using American Football Scores

Everson, Phil

Swarthmore College

Department of Mathematics and Statistics

500 College Avenue

Swarthmore, PA1908, USA

E-mail: peverso1@swarthmore.edu

1. Introduction and Summary

Extensive data is available from the National Football League (NFL) on American football games. I have compiled data on the 768 regular-season games from the 2004, 2005 and 2006 NFL seasons and linked the file to my web page: www.swarthmore.edu/NatSci/peverso1/SportsData.htm. Along with the home and road team ID's and the scores in each game, I have included the points scored by each team in each quarter of the game so that partial-game scores and point spreads can be computed. I define the point spread to be the home team's score minus the road team's score, so that negative values indicate that the home team has lost (or is losing). There is a point spread after each quarter, and these serve as predictors for the final point spread.

The Las Vegas betting spread is available before a game begins, and provides a nearly unbiased estimate of the final point spread in a game. It is not intended as an estimate of the outcome, but rather provides a baseline for even-money betting (minus a commission). If Dallas were favored by 7 points over Philadelphia, for example, a person betting on Dallas would win only if Dallas beat Philadelphia by more than 7 points. A person betting on Philadelphia would win if Philadelphia won, or if Dallas won by fewer than 7 points. If Dallas were to win by exactly 7 points, no money would change hands. The betting spread changes during the week before a game in reaction to the amount of money bet on each team, and different casinos occasionally list different betting spreads for the same game. I recorded the spreads reported in the Philadelphia Inquirer on the days of each game. A bookie collects a commission (typically 10%) from people who win their bets. So someone who bets \$100 would lose \$100 if they lose the bet, and would collect only \$90 if they win. The bookie will make a profit regardless of who wins as long as roughly the same amount is bet on each team. Then the winners can be paid off with money collected from the losers, with 10% of that amount kept as a profit. It is rather remarkable that this profit-driven process results in a number that is about right on average as an estimate of the actual margin of victory.

Linear models do a good job of describing the association between the final point spread and the Vegas spread and partial game spreads. Multiple regression models show that the Vegas spread remains important as a predictor even after learning the scores part way through a game, but that the first quarter spread, for example, become insignificant once the half-time spread is included. Replacing the final point spread with an indicator for whether or not the home team won leads to a logistic regression model to estimate the probability of winning from the Vegas spread and/or partial-game information.

2. Data Summaries

Figure 1 shows histograms of the points scored by the home team, by the road team, and the winning margin for the home team. While the distributions of each team's score are stacked up against 0 and hence skewed right, the difference in scores has a roughly symmetric distribution that is very close to Normal, as shown by the nearly linear Normal quantile plot. The overall mean point spread is about 2.3 points, meaning that on average the home teams scored 2.3 points more than the road teams, and therefore tended to win. The standard deviation of the point spreads was about 14.4 points, meaning the middle 95% of point spreads range from about -26 points to 31 points (roughly $2.3 \pm 1.96(14.4)$). The Normal approximation to the probability that the home team wins is $P(Z < (0-2.3)/14.4) \approx 0.56$, which almost perfectly matches the actual home-team winning percentage of $432/768 = 0.5625$. Note that I do not apply the continuity correction in any calculations in this paper, although that would certainly be appropriate. It might also be worthwhile to factor in the fact that a winning margin of 0 is essentially

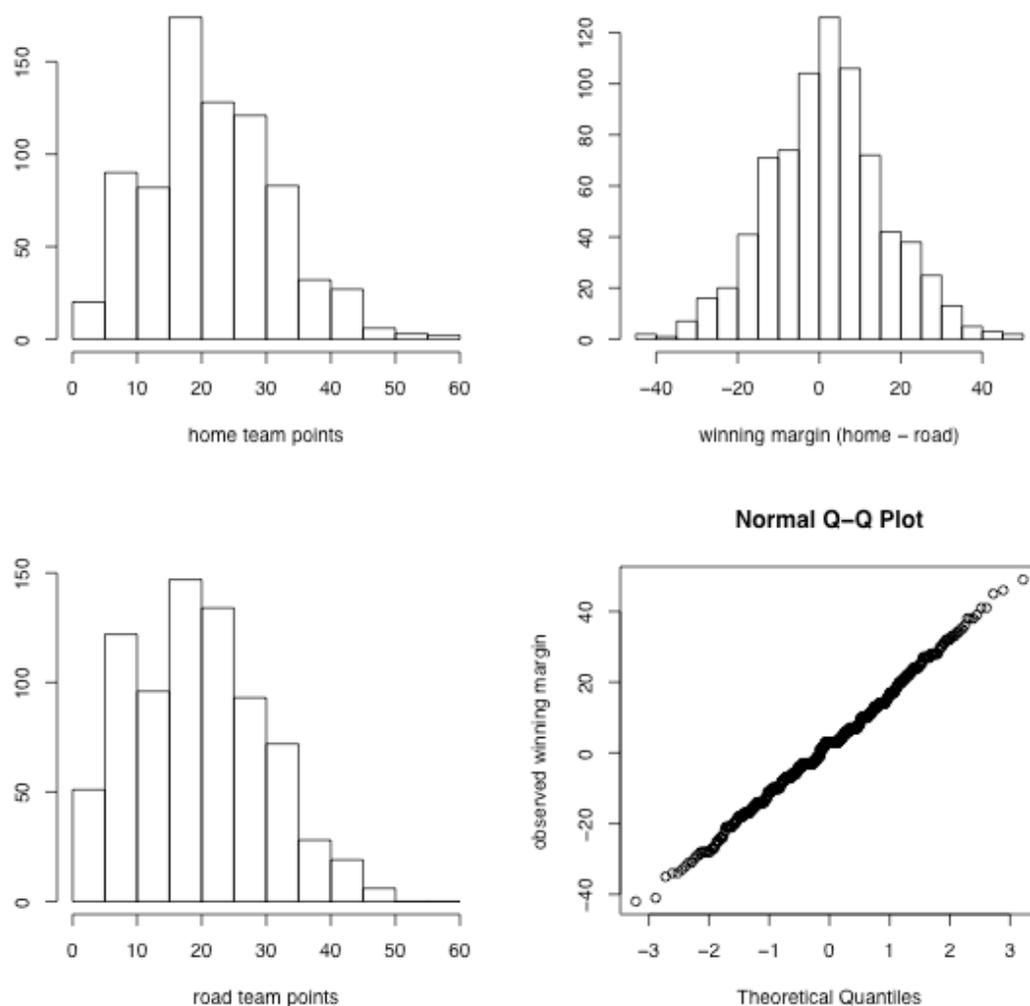


Figure 1: Point totals and winning margins for the home team for the football games played during the 2004-2006 NFL seasons.

ruled out (with sudden-death overtime, no ties have occurred in an NFL game since 2002). These issues could make for interesting discussions even with classes where the students would not actually be expected to carry out such computations.

Figure 2 shows that the distribution of the difference between the actual spread (home – road) and the Vegas spread is approximately Normal. The average deviation over the past three seasons was -0.34 points, with a standard deviation of 13.15 and standard error ($n=768$) of 0.47 points. Zero is squarely in the 95% confidence interval for the mean deviation, but this alone doesn't imply that any particular Vegas spread is unbiased for the actual spread. It would be possible, for example, that Vegas underestimates the high spreads and overestimates the low spreads in such a way that they average out to have a mean deviation of 0. To see that this doesn't happen, we can fit a linear model to predict the actual margin from the betting spread.

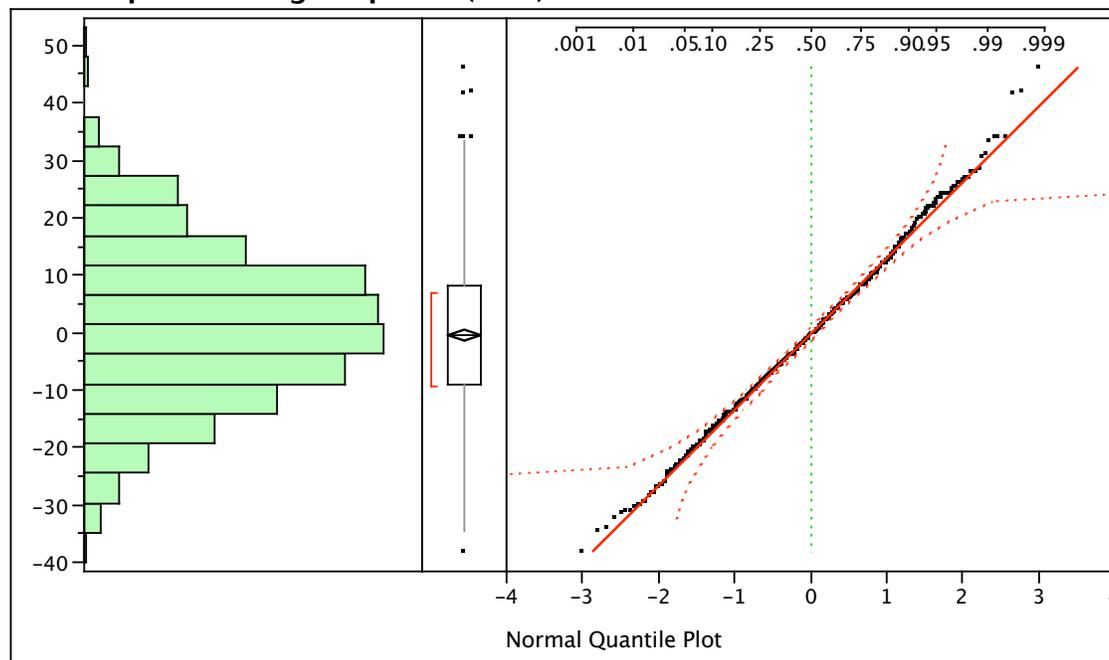
Actual Spread – Vegas Spread (H-R)

Figure 2: The deviation between the actual winning margin for the home team and the Vegas betting spread for 768 NFL games played between 2004 and 2006. The distribution is very close to Normal with a mean of 0 and a standard deviation of 13 points.

3. Simple Linear Regression Model

The least squares fit of the actual point spread (H-R) against the Vegas spread (Vegas H-R) yields a prediction equation that is very nearly the identity $y=x$. That is, the estimated mean winning margin is essentially the Vegas betting spread. From the JMP output in Figure 3 we see that the regression assumptions appear to be met, with the residuals showing symmetric variation about 0, and roughly the same standard deviation for all Vegas H-R values. The estimated residual standard deviation is given in the output as the Root Mean Square Error, which is about 13 points. Assuming Normal errors, we can use this to estimate game outcomes. If Dallas were favored by 7 points, for example, the probability of Dallas losing (according to the fitted model) is approximately the probability that a Normal variable would be more than $7/13$ standard deviations below its mean: $P(Z < (0-7)/13) \approx 0.3$. By symmetry, we can say that 7-point favorites in the NFL (whether the home or road team) win about 70% of the time.

The fitted intercept and slope are within a standard error of 0 and 1, respectively. To test simultaneously that the slope is 1 and the intercept is 0, we can carry out an “extra sum of squares F test.” From the output in Figure 2 we see the Sum of Squares Error (SSE) for the simple regression of actual point spread (H-R pts) on the Vegas betting spread (Vegas H-R) is 132,633.9, with two regression parameters fit. This is computed by summing the squared deviations between the actual H-R pts and the fitted values based on the Vegas spreads:

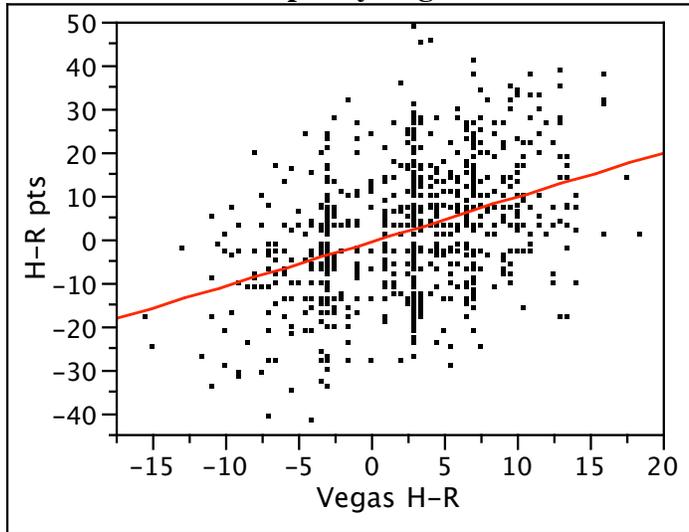
$$\text{Fitted Point Spread} = -0.402 + 1.025(\text{Vegas H-R}).$$

The reduced model assumes the mean H-R pts is exactly the Vegas H-R value:

$$\text{Fitted Point Spread} = 0.00 + 1.00(\text{Vegas H-R})$$

with no regression parameters fit. The sum of squares error for the reduced model is the sum of the squared

Bivariate Fit of H-R pts By Vegas H-R



Linear Fit

$$\text{H-R pts} = -0.402204 + 1.0250634 \text{ Vegas H-R}$$

Summary of Fit

RSquare	0.167986
RSquare Adj	0.166899
Root Mean Square Error	13.1587
Mean of Response	2.334635
Observations (or Sum Wgts)	768

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	26779.09	26779.1	154.6572
Error	766	132633.90	173.2	Prob > F
C. Total	767	159413.00		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	-0.402204	0.523344	-0.77	0.4424
Vegas H-R	1.0250634	0.082426	12.44	<.0001

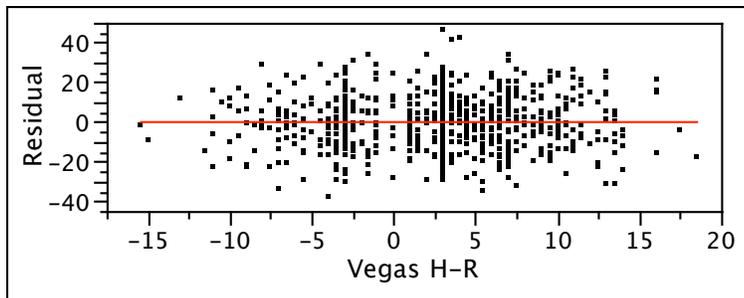


Figure 3: Simple regression of margin of victory for the home team (H-R pts) on the Vegas betting spread (Vegas H-R). The prediction equation is essentially $y=x$, meaning the Vegas spread is nearly unbiased for the actual winning margin.

deviations in Figure 2. This can be found from the sample mean and standard deviation of the difference between H-R pts and Vegas H-R pts for game i :

$$\sum_{i=1}^{768} y_i^2 = (n-1)s_y^2 + n\bar{y}^2 = (768-1)(13.16)^2 + 768(-0.335)^2 = 132,919.5.$$

The reduction in sum of squares error due to fitting the intercept and slope from the data is $132,919.5 - 132,633.9 = 285.6$. This is the “extra sum of squares” that represents the error in the reduced model, with the intercept and slope fixed at 0 and 1, that is not present in the full model, with an arbitrary intercept and slope. This reduction was achieved at the cost of fitting two additional parameters, so the mean square reduction (MSR) is $285.6/2 = 142.8$. If the true intercept and slope were in fact 0 and 1, the MSR would be an unbiased estimate of the residual variance σ^2 . If the true intercept and slope differed from these values, the MSR would typically overestimate σ^2 . The mean square error (MSE) from the full model is an unbiased estimate of σ^2 , regardless of the true parameter values (assuming a linear model is appropriate) and for this regression takes the value 173.2 (the square of the root mean square error, or the SSE divided by $768-2=766$, the degrees of freedom for error). As this estimate is larger than the MSR there is no reason to think that the MSR is overestimating σ^2 , meaning there is no evidence to reject 0 and 1 as the true intercept and slope.

Under the usual regression assumptions, we can do a formal test of the null hypothesis $H_0: \beta_0 = 0$ and $\beta_1 = 1$, against the alternative that either β_0 or β_1 (or both) differ from these values. With Normal errors, and assuming H_0 is true, the ratio $F = \text{MSR}/\text{MSE}$ follows an F distribution. In this case, the degrees of freedom in the numerator is 2, the difference in the number of parameters fit ($2-0 = 2$) and the denominator degrees of freedom is $768-2=766$, the number of observations minus the number of parameters fit in the full model. If H_0 were not true, then fitting an intercept and slope allows for a better fit and a greater reduction in the SSE, meaning the MSR and the ratio of MSR to MSE would tend to be larger than if H_0 were true. So the test rejects H_0 for large values of F . We observe $F = \text{MSR}/\text{MSE} = 142.8/173.2 \approx 0.82$. The mean of any F distribution is larger than 1.0, so an F statistic smaller than 1.0 will never provide significant evidence to reject H_0 . This does not mean we can claim the Vegas spread is in fact unbiased for the final point spread (we don't get to conclude H_0 is true), only that the observed point spreads have been quite consistent with this assumption over the past three seasons.

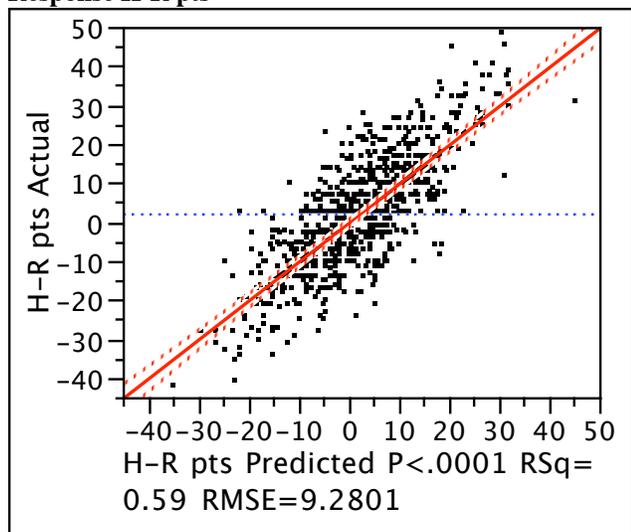
3. Multiple Regression Model

The partial-game point spreads allow for numerous multiple regression possibilities. Figure 4 shows JMP output for the multiple regression of the final point spreads (H-R pts) on the betting spreads (Vegas H-R) and the half-time spreads (Q2 H-R). I also have data on the scores after the first and third quarters, but these aren't as appropriate as predictors because the game doesn't re-start after these breaks as it does after half-time. American football is a game of field position, and a team's position on the field is preserved after the first and third quarters (although they change directions and move to the opposite end of the field). So if one team were very close to the opponent's end-zone at the end of the first quarter, they would start the second quarter with a higher probability of scoring quickly than at the beginning of the game or of the second half, when there are kick-offs. An extra sum of squares F test shows that using the half time points for both the home-team and road-team as predictors does not improve the fit significantly over using only the half-time point spread (Q2 H-R). The prediction equations for the simple regression model of Section 2 and for this multiple regression are:

$$\text{Fitted Point Spread} = -0.40 + 1.025(\text{Vegas H-R}).$$

$$\text{Fitted Point Spread} = -0.60 + 0.493(\text{Vegas H-R}) + 0.926(\text{Q2 H-R}).$$

Multiple Regression Model: Final Margin on Vegas Spread and Half-time Margin
Response H-R pts



Summary of Fit

RSquare	0.586722
RSquare Adj	0.585641
RMSE	9.280093
Mean of Response	2.334635
Observations	768

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	93531.10	46765.5	543.0269
Error	765	65881.90	86.1	Prob > F
C. Total	767	159413.00		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	-0.601116	0.369154	-1.63	0.1039
Vegas H-R	0.4933015	0.061188	8.06	<.0001
Q2 H-R	0.9263356	0.033273	27.84	<.0001

Residual by Predicted Plot

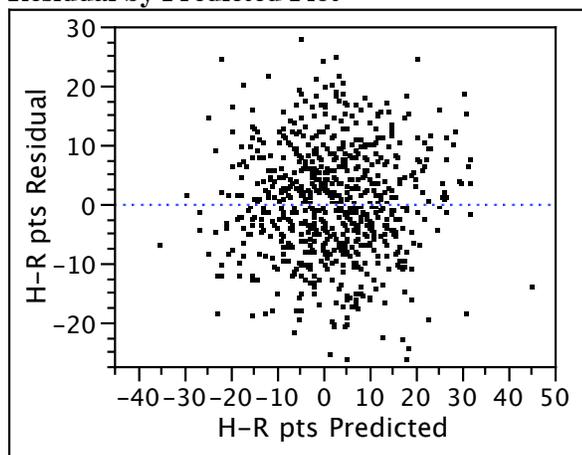


Figure 4: Multiple Regression of the margin of victory for the home team (H-R pts) on the Vegas betting spread (Vegas H-R) and the half-time winning margin for the home team (Q2 H-R).

The Vegas H-R coefficient is reduced by about 50% when the half-time (2nd quarter) point spread is included in the model, but it is still significantly larger than zero ($p < .0001$). This means the betting spread still provides useful information about the outcome even after observing the half-time score. However, knowledge of the half-time score increases the R-square value to 59% from 17% in the regression on Vegas spread only, showing that partial-game point spreads explain much more of the variability in actual point spreads than does the prior expert opinion. And there is still considerable uncertainty remaining (about 41% of the variability is unexplained) at the mid-way point in the game.

As an example, last December 25 the Philadelphia Eagles (my local team) played the Dallas Cowboys in Dallas, Texas. The Cowboys were favored by 7 points, but at half-time the Eagles were winning by 6 points (13-7). The linear models allow us to estimate the distributions of the final point spread before the game began and at the half-way point. The Vegas H-R value was 7 points for the Cowboys, and the residual standard deviation is about 13.16. So before the game began, the probability of the Cowboys winning was approximately $P(Z > (0.0 - 7.0)/13.16) = P(Z > 0.53) \approx 0.70$.

At half-time, however, the probability of winning sways slightly in favor of the Eagles, who were leading 13-7. The Cowboys were at home, so $Q2\ H-R = 7 - 13 = -6$. The residual standard deviation estimate for the multiple regression is about 9.28 points (the Root Mean Squared Error) and the fitted final winning margin for the Cowboys at half-time is

$$\text{Fitted point spread} = -0.60 + 0.493(7) + 0.926(-6) \approx -2.7 \text{ points.}$$

The negative fitted value indicates the Cowboys are more likely to lose than to win. Now the approximate probability of the Cowboys winning is

$$P(Z > (0 - (-2.7))/9.28) = P(Z > 0.29) \approx 0.39.$$

The probability that Dallas would go on to “cover the spread” (in this case, win by more than 7 points) so that people who had bet on Dallas would win money was approximately

$$P(Z > (7 - (-2.7))/9.28) = P(Z > 1.045) \approx 0.15.$$

In my first *A Statistician Reads the Sports Pages* column in *Chance* Magazine, I made estimates after the first quarter of this game when I had to turn it off and go to a holiday party. The first-quarter spread is a significant predictor with or without the Vegas spread, but when it is added to this model with the half-time score it is not at all significant. The output below shows that the parameter estimate for Q1 H-R is very close to 0, and far from significant ($p > 0.9$). This would be intuitive to most sports fans, as knowledge of the score later in the game makes any earlier scores irrelevant. So this provides an easy context for discussing the interpretation of multiple regression coefficients.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	-0.600528	0.369424	-1.63	0.1045
Vegas H-R	0.4935139	0.061252	8.06	<.0001
Q1 H-R	-0.007927	0.064906	-0.12	0.9028
Q2 H-R	0.929579	0.042588	21.83	<.0001

4. Logistic Regression Model

Throughout, I have used the linear regression models to estimate the probability of one of the teams winning the game. Logistic regression is a tool for predicting such probabilities directly, without considering the margin of victory. Below is JMP output for a logistic regression of “Winner” (Home or Road team) on the Vegas betting spread and the half-time point spread.

Nominal Logistic Fit for Winner

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-0.1703115	0.101806	2.80	0.0943
Vegas H-R	0.10846001	0.0172967	39.32	<.0001
Q2 H-R	0.14948959	0.01266	139.43	<.0001

For log odds of Home/Road

The linear fit for a logistic regression is an estimate of the log-odds of the home team winning vs. the road team winning. To find the fitted probability that the home team wins, first find the estimated log-odds, then solve the equation to find p. For example, with the Eagles-Cowboys game, Vegas H-R = 7 and Q2 H-R = -6. So the estimated log-odds is

$$\text{Fitted log-odds} = -0.170 + 0.108(7) + 0.149(-6) = -0.308 \approx \phi = \log(p/(1-p)).$$

The odds are computed as $p/(1-p)$, where p is the probability of the home team winning, and the logs are taken with base $e \approx 2.718$. So for log-odds $\phi = -0.308$, inverting this equation to estimate p yields

$$p = \frac{e^{\phi}}{1 + e^{\phi}} = \frac{1}{1 + e^{-\phi}} = \frac{1}{1 + e^{-(-0.308)}} \approx 0.42.$$

This estimate doesn't differ much from the multiple regression estimate of 0.39, despite ignoring information about the margins of victory. The NCAA has required that Bowl Championship Series (BCS) computer ranking models for college football recognize only who wins, and not by how much, when assessing the relative strengths of the teams. Logistic regression is the simplest model for making such estimates with one or more continuous or categorical predictors. However, the NCAA would certainly balk at the use of betting spreads to rank teams.

REFERENCES

Stern, H. S. (1991). “On the Probability of Winning a Football Game,” *The American Statistician*, Vol. 45, pp. 179-183.

Everson, P. (2007). A Statistician Reads the Sports Pages: “Describing Uncertainty in Games Already Played.” *Chance*, Volume 20, number 2, pp. 49-56.

Keywords: Extra Sum of Squares F Test; Regression t test; Normal Quantile Plot.

ABSTRACT

Scores in professional American football games follow a distribution that is noticeably skewed towards larger values. However, the difference between the home teams' scores and the road teams' scores (the point spreads) do follow a distribution that is very close to Normal. Furthermore, the residuals from linear regressions of the point spreads on the Las Vegas betting spreads and on partial game point spreads (e.g., the point spreads at half-time) are also very close to Normal, and suggest that a linear model is appropriate. These data are also suitable for logistic regression problems, such as estimating the probability that the home team wins given the betting spread, the half time score, or both. This presents an ideal setting for teaching simple and multiple linear regression and logistic regression procedures for any level of statistics course.