**Lab 5**

The data used in this assignment came from the France Soccer Championship of 2007-08. It was a round-robin tournament, where teams accumulated 3 tournament points for a win, 1 for a tie, and 0 for a loss. The data was difficult to summarize, as it was game by game tournament results for 394 games. However, to give a sense of what the data looks like, below is a *jittered* graph of the outcomes of games (whether or not the “team of interest” won) plotted against the cumulative tournament score difference between the teams participating in the match, the relationship this analysis will attempt to define. Note: *jittering* is a process that spreads points that are plotted on top of one another out to help give a feel of the density of how all the points fall. It is bad for precision, but good for relating how the data is distributed.



Figure 1: A plot of game outcomes (either a win for the team of interest, P1, or a not-win) against the score difference between the cumulative tournament scores, with respect to P1. Wins for P1 are represented as 2, and not-wins are represented by 1. As can be seen, wins rarely occur when P1 has a lower score than P2. Note: a “not-win” can be either a loss or a tie.

1. **Construct a GLM to predict the probability to win a game (response) using the current standing of both teams (explanatory). (Decide whether you need the quadratic terms in this problem; justify your choice).**

Simple and Multiple Logistic Regression are approaches to perform analysis using one or more independent variables, in our case, the difference between the tournament standing of a team of interest and their opponent, and different variations of this term, by raising it to various powers, etc. Simple logistic regression produces a GLM of the form:

Multiple logistic regression produces a GLM in the form of:

The model can be designed to account for individual levels of variables, as well as joint effect of the variables analyzed. A general model of the analysis is as follows:

and any combination of terms thereof.

However, in this analysis, it was found that the most effective model contained no joint effects, and was simply a regression of a single score difference term raised to the correct power.

A quadratic term was not necessary. The best model I found used only a cubed difference of cumulative tournament scores term:



Figure 2: A plot of the predicted probability of a tie occurring based on the cumulative tournament score difference between teams using the selected, where P1 is the team of interest

The specific data relating to the model is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Parameter/Coefficient | Estimate | Std. Error | P-Value |
| α | -0.2879 | 0.1058 | 0.0065 |
| βScore Difference3 | 4.688e-05 | 1.837e-05 | 0.0107 |
| AIC | 537.66 |

This model was chosen because it performed better than any other models found in two respects: (1) it performed slightly better than any other models considered in terms of AIC, and (2) more importantly, this model fared much better in terms of parameter significance, having highly significant parameters and no insignificant ones.

A quadratic score difference term was not used for two reasons: (1) models using a squared score difference term were not good fits, and (2) this would not result in an intuitive interpretation: the absolute value of a score difference in athletic games is usually not considered significant, there is valuable information in the sign or direction of the score difference (i.e. which team or athlete is better).

1. **Discuss the model results. What is the predicted dependence of the game result on the team standing? Does this supports/contradict the possibility of bribing? Why or why not?**

The chosen (best) model suggests that a team becomes increasingly more likely to win as their tournament score increases relative to their opponent. If a team has a lower tournament score than their opponent, they are unlikely to win; conversely, if a team has a higher tournament score than their opponent, they are likely to win. Because the score difference is cubic, the slope of the model prediction flattens as the score difference approaches zero, which makes intuitive sense, because if the score difference is small, the likelihood of a win for the team of interest should not change much.

This result contradicts the assertion that bribing was present. If the assumption that the cumulative tournament score is representative of a team’s ability to win (not a far stretch), then it would be expected that teams with more points than their opponents will win over their opponents. Because the model is valid, it suggests that the relationship predicted by the model is valid. Because the model suggests that a team is more likely to win if they have greater ability than their opponent, it makes intuitive sense that bribery was not involved.

1. **Find 95% CIs for all model parameters.**

|  |  |  |
| --- | --- | --- |
| Parameter/Coefficient | 2.5% | 97.5% |
| α | -0.49673 | -0.08164 |
| βScore Difference3 | 1.3679e-05 | 8.6858e-05 |

1. **Find a 95% CI for the probability to win when the team score is 50 and the opponent’s score is 10.**

|  |  |  |  |
| --- | --- | --- | --- |
| P1 Tournament Score | P2 Tournament Score | 2.5% | 97.5% |
| 50 | 10 | 0.806 | 1.070 |