MATH3821 Statistical Modelling and Computing
Predicting the FIFA 2014 World Cup Winner

Nathan Wilson

August 13, 2016

Introduction

The topic for this assignment was "Could you have predicted the winner of the 2014 FIFA world cup final?" There are two ways to interpret this that will be examined in this report.

The first is: "If you were to try to predict the winner before it happened, would you have been successful?." This will be answered by building a regression model using only data that could have been obtained before the match, then seeing what it predicts and whether it is successful.

The second, more general question is: "Is it possible to successfully predict soccer games (and sports in general) over the long term using regression?". This will be answered by examining both the results of this analysis, as well as the state of the wider field of sports forecasting.

Approach

Before gathering data, the first consideration is what we are aiming to predict and how we might model outcomes. Three options were considered for the response:

- Binary response modeling a win or loss
- Ordered logistic regression or multinomial regression over the categories 'win' 'draw' or 'loss'
- Poisson regression modeling the number of goals scored or difference in number of goals scored, which can be used to determine the match result

The outcomes was modelled as a binary response, with a 0 indicating an away team win and a 1 indicating a home team win. This method was chosen primarily because ordered logistic regression, multinomial regression and poisson regression are not in the scope of the first three weeks of lectures. Tie games would be dropped from the dataset entirely as our model cannot handle them.

The dataset was limited to the 2014 FIFA season, and predictors would be constructed from various statistics about each team. There were several reasons for limiting the dataset to the most recent season:

- By using only FIFA games, predictors could be used that were specific to FIFA, such as the FIFA rankings.
- By only using the most recent season, teams were guaranteed the same players for all of the dataset - 'Australia' from 2014 is not the same as 'Australia' from 2010 or 2006.
- By keeping the set of teams small (32 in total) it was feasible to manually enter data from web pages in a reasonable amount of time. Each extra 'cup' included would add another 32 teams, making manual data entry impractical. As this course is focused on the statistical analysis and not necessarily on getting huge quantities of data, the 63 matches in the FIFA 2014 cup were deemed sufficient.

After formulating predictors, the necessary data was then gathered and transformed into clean data frames. Predictors were constructed and analysed for various assumptions, and a model was built out of the most promising predictors. Finally, this model would be tested against the prediction in question and analysed.

Choosing Predictors

In order to construct a dataset, predictors needed to be decided on. Also important was a symmetric way of representing team statistics, symmetric meaning that the model would output an equivalent response regardless of which team was home or away. To achieve this, predictors were constructed as the difference in statistics between two teams - 'distance_travelled_home - distance_travelled_away', for example. Constructing predictors in this manner resulted in each predictor being partitioned at zero such that a positive number would indicate an advantage for one team, while a negative number indicated an advantage for the other.

After conversations with friends and a review on the current literature, the following predictors were established:

Home team advantage: It has been shown and makes logical sense that a team will perform better when playing in their home stadium. This is a categorical variable which is either "HOME" (Home team has home team advantage), "NONE" (neither has home team advantage) or "AWAY" (away team has home team advantage).

Difference in distance team has travelled from their home capital: This is of similar relevancy to home/away but applies across even teams that are both playing 'away'

Recent possession difference: Possession is calculated by taking the average possession percentage of the home team in the games played during FIFA 2014 and subtracting the same stat for the away team.

Difference in FIFA world ranking: This takes into account the recent performance of the team and is a proxy for 'how well has this team been performing' in the last 4 years (recently). These stats were published by FIFA the month before the world cup took place.

Difference in change between teams ranking over past 12 months: That is, a team 'on the up' may be seen as more likely to win. The rankings in the previous predictor were compared to the rankings for the same teams in June 2013.

Difference in average player age: Age may be relevant as a proxy for player experience, and was calculated by averaging the ages of the players in each team.

Difference in average player height: May represent the 'vitality' of players in a similar way to height in basketball (although probably less relevant if you're not jumping for a hoop).

Difference in number of players nominated for a Ballon d'Or: The Ballon d'Or is an annual medal given by FIFA to the 'world's best male player.' There are 20 nominees each year and the number of players nominated for this award acts as a good proxy for how many 'star players' a team has.

Difference in average goals scored recently: The average number of goals scored across the first 63 matches in FIFA is a good measure for recent team performance.

Difference in average concessions in recent history: Similar to average goals, this also is a measure of recent team performance.

Gathering Data and Constructing Predictors

Raw data was gathered for the match results and player statistics for FIFA 2014, as well as possession statistics per game and the distance each team travelled to get to Brazil (sources in Appendix A). The rest of the data needed to construct the predictors - the FIFA team rankings for June 2014 and June 2013, and the number of Ballon d'Or nominees - was gathered by hand.

Some transformation and 'data massaging' was necessary in order to construct the final data frame. A few notes:

- The possession.csv file required transforming via regex before it was able to be read into R
- The match and player data did not join successfully by country at first. There were a few differences that needed correcting:
  - Colombia vs. Columbia
  - Bosnia and Herzegovina vs. Bosnia & Herzegovina
  - South Korea vs. Korea Republic
- The raw match data gathered included the final game, which of course wouldn't have been available to someone making a bet the night before the game. It was necessary to remove the last datapoint before constructing predictors and training data frames.

Once the raw match and player data had been collected, a 'teams' data frame was created that had for each team the statistics needed for predictors, such as player height, etc.

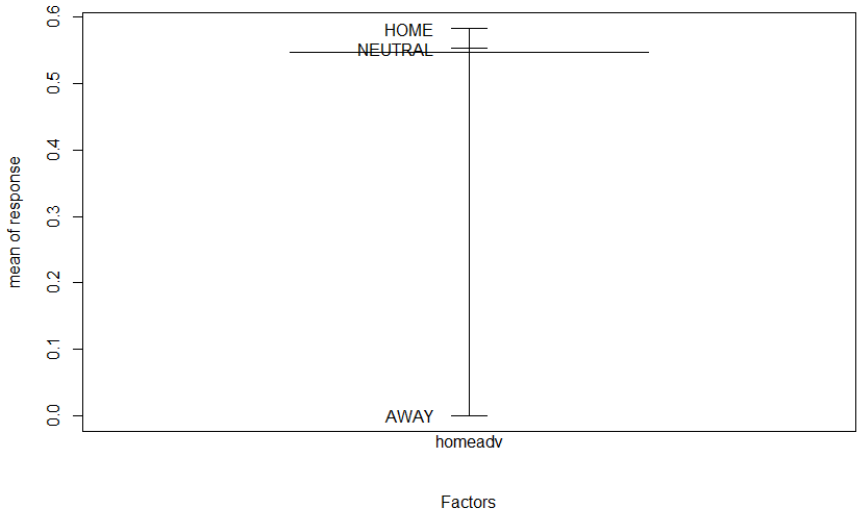
	Team	Home_adv	Distance_travelled	Possession	FIFA_rank	FIFA_rank_increase	Player_age	Player_height	Award_players	Recent_goals	Recent_concessions
1	Algeria	0	8074	41.25000	858	13	29.00000	184.5000	0	1.7500000	1.7500000
2	Argentina	0	2355	56.83333	1175	-2	28.66667	183.0000	2	1.3333333	0.5000000
3	Australia	0	15577	43.66667	526	-15	27.00000	179.2857	0	1.0000000	3.0000000
4	Belgium	0	9053	52.60000	1074	-1	25.25000	181.5833	2	1.2000000	0.6000000
5	Bosnia and Herzegovina	0	9594	53.66667	873	-6	33.00000	190.0000	0	1.3333333	1.3333333
6	Brazil	1	0	53.42857	1242	18	28.09091	182.8182	1	1.5714286	2.0000000
7	Cameroon	0	7275	45.66667	558	9	21.50000	176.5000	0	0.3333333	3.0000000
8	Chile	0	2844	55.25000	1026	11	30.20000	177.4000	0	1.5000000	1.0000000
9	Colombia	0	3218	47.60000	1137	-1	32.00000	182.5000	0	2.4000000	0.8000000
10	Costa Rica	0	4422	42.80000	762	20	28.90000	178.4000	0	1.0000000	0.4000000
11	Croatia	0	9510	49.33333	903	-14	23.16667	182.8333	0	2.0000000	2.0000000
12	Ecuador	0	3314	45.00000	791	-16	28.50000	182.1250	0	1.0000000	1.0000000
13	England	0	8849	54.33333	1090	-1	26.75630	182.4706	0	0.6666667	1.3333333
14	France	0	8807	53.00000	913	1	26.13043	178.5870	2	2.0000000	0.6000000
15	Germany	0	9682	56.33333	1300	0	25.83333	183.0769	5	2.8333333	0.6666667
16	Ghana	0	6114	48.66667	704	-16	24.00000	177.0000	0	1.3333333	2.0000000
17	Greece	0	9763	47.25000	1064	4	27.00000	183.1538	0	0.7500000	1.2500000
18	Honduras	0	4998	49.00000	731	19	27.72727	180.0909	0	0.3333333	2.6666667
19	Iran	0	12117	35.00000	641	24	28.42857	181.7857	0	0.3333333	1.3333333
20	Italy	0	9064	55.00000	1104	-1	27.39506	181.0864	0	0.6666667	1.0000000
21	Ivory Coast	0	5652	55.33333	809	-10	25.00000	190.0000	1	1.3333333	1.6666667
22	Japan	0	17361	55.66667	626	-14	27.66667	178.4000	0	0.6666667	2.0000000
23	Korea Republic	0	17417	52.33333	547	-17	26.85714	187.4286	0	1.0000000	2.0000000
24	Mexico	0	6378	48.75000	882	-3	27.96154	177.1923	0	1.2500000	0.7500000
25	Netherlands	0	9182	49.28571	981	-10	22.55000	178.7500	1	2.1428571	0.5714286
26	Nigeria	0	7045	50.75000	640	-13	25.33333	189.6667	0	0.7500000	1.2500000
27	Portugal	0	7376	50.66667	1189	2	25.65217	182.8696	0	1.3333333	2.3333333
28	Russia	0	11282	49.33333	893	-8	27.20588	182.3529	0	0.6666667	1.0000000
29	Spain	0	7849	56.00000	1485	0	27.06250	179.9219	3	1.3333333	2.3333333
30	Switzerland	0	9000	47.75000	1149	8	25.50000	179.8333	0	1.7500000	1.7500000
31	Uruguay	0	2334	45.75000	1147	12	23.00000	196.0000	0	1.0000000	1.5000000
32	USA	0	6458	44.75000	1035	15	28.82353	181.2353	0	1.2500000	1.5000000

This data was then used to create a training data frame from the matches, by taking the result of each of the first 63 matches and joining it onto the team data by hometeam and awayteam. Each statistic for the away team was subtracted from that for the home team such that a score above zero would indicate an advantage for one team, and a score below zero for the opponent.

The R code used to do this, and all the other R code used in this assignment, is available in Appendix B.

Exploratory Analysis

Before fitting a model, exploratory analysis was done to determine a subset of predictors from the set gathered. First the factor, Home_adv was plotted against the response to check whether there was a between groups difference at all.



It appeared there was an effect, so it was included in the model and checked for significance.

A pairwise scatterplot was produced to check the assumption of multicollinearity (full plot in Appendix C). From this plot, it was apparent that there was a correlation between the three predictors 'recent goals', 'recent concessions' and 'FIFA ranking'. The relationship between goals and concessions makes sense intuitively, as goals and concessions are based on the same raw data and therefore represent the same information to an extent.

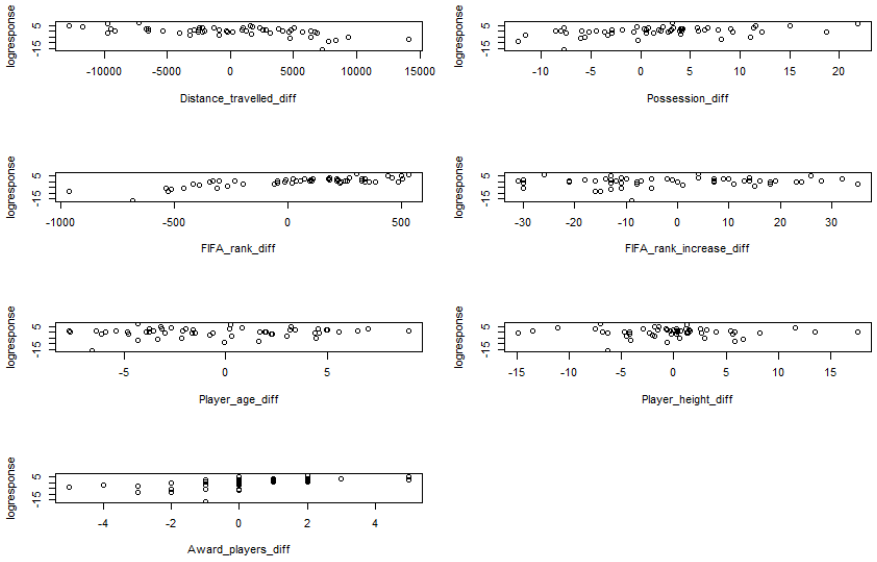
The 'FIFA ranking' stat is created by FIFA based on goals scored in games over the last 8 years, weighted by recency. The FIFA rankings came out before the cup and so are not based on the same data, but 'FIFA ranking' and 'Recent goals' stats are both proxies for the recent performance of a team. With this in mind I dropped Recent_goals_diff and Recent_concessions_diff from my model before fitting it.

Model Selection

After removing the aforementioned predictors, the model remaining was:

Response Home_adv + Distance_travelled_diff + Possession_diff
+ FIFA_rank_diff + FIFA_rank_increase_diff + Player_age_diff
+ Player_height_diff + Award_players_diff

I first plotted the predictors against the fitted log odds in order to visually examine for linearity in the predictors.



After fitting this initial model, the summary was:

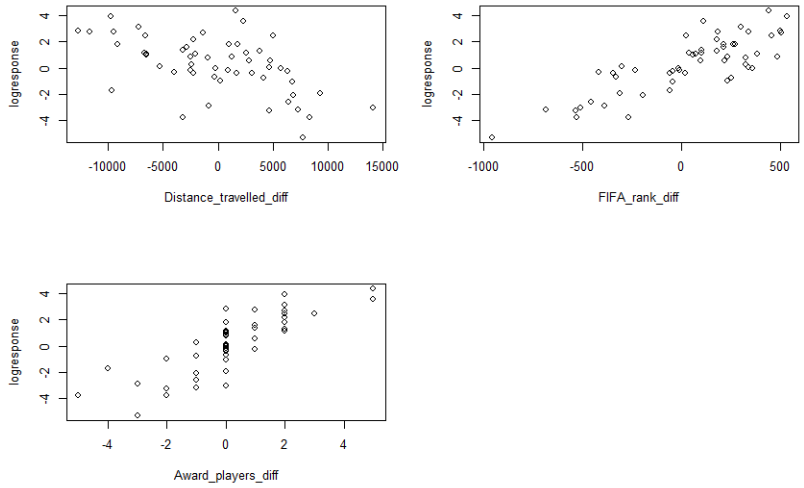
```
> summary(match_model)

Call:
glm(formula = Response ~ Home_adv + Distance_travelled_diff
    +
    Possession_diff + FIFA_rank_diff +
        FIFA_rank_increase_diff +
    Player_age_diff + Player_height_diff +
        Award_players_diff,
    family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5247  -0.4984   0.1105   0.6971   1.5867

Coefficients:
```

```
                            Estimate Std. Error z value Pr(>|z
                                |)
(Intercept)              -1.273e+01  2.400e+03  -0.005
    0.9958
Home_advHOME              1.152e+01  2.400e+03   0.005
    0.9962
Home_advNEUTRAL           1.259e+01  2.400e+03   0.005
    0.9958
Distance_travelled_diff -3.555e-04  1.701e-04  -2.090
    0.0366 *
Possession_diff          -9.736e-02  8.656e-02  -1.125
    0.2607
FIFA_rank_diff            5.563e-03  2.767e-03   2.010
    0.0444 *
FIFA_rank_increase_diff -3.789e-02  3.114e-02  -1.217
    0.2238
Player_age_diff          -1.415e-01  1.589e-01  -0.890
    0.3733
Player_height_diff       -2.004e-01  1.143e-01  -1.753
    0.0797 .
Award_players_diff        7.423e-01  3.733e-01   1.989
    0.0467 *
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 74.192  on 53  degrees of freedom
Residual deviance: 40.771  on 44  degrees of freedom
AIC: 60.771

Number of Fisher Scoring iterations: 15
```

A Wald test on the coefficients indicates the following were not significant:

- Home advantage: Distance travelled was a significant factor, and it is likely any information provided by the HOME_ADV factor was already accounted for in the distance travelled. Player age, player height were not shown to have predictive significance. It seems that any advantage given by experience is made up for by youth, and that size is not a factor in soccer as much as it is in other sports.
- Fifa_rank_increase and possession also showed no significance in relation to the response.
- The intercept was also not shown to be significant. Removing the intercept in a binomial regression model is equivalent to assuming log odds of 0 if all predictors are set to 0. In our model, we can interpret this as meaning that all else equal, the probability of the home team winning is 0.5. This makes intuitive sense and is to be expected.

After removing these predictors and ordering them from least to most significance (to analyse the model with drop1) the same analysis was run.



Plots

Summary

```
> summary(match_model)

Call:
glm(formula = Response ~ FIFA_rank_diff +
    Distance_travelled_diff +
     Award_players_diff - 1, family = "binomial")

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.3544   -0.5640    0.2670    0.7643    1.5125

Coefficients:
                           Estimate Std. Error z value Pr(>|z
                              |)
FIFA_rank_diff            0.0021119  0.0013819   1.528
    0.1265
Distance_travelled_diff -0.0001386  0.0000755  -1.836
    0.0664 .
Award_players_diff        0.7289353  0.3144682   2.318
    0.0204 *
---

(Dispersion parameter for binomial family taken to be 1)

     Null deviance: 74.86  on 54  degrees of freedom
Residual deviance: 47.92  on 51  degrees of freedom
```

```
AIC: 53.92

Number of Fisher Scoring iterations: 5
```

The coefficients for this model indicate that a teams chances of winning goes up as:

- The distance the team has travelled to get to the venue goes down, as seen in the negative coefficient for distance_travelled_diff
- The relative FIFA rank goes up
- The team has more award winning players

Model Assessment

Coefficient Significance

The selected model with three predictors has summary output statistics:

```
glm(formula = Response ~ FIFA_rank_diff +
    Distance_travelled_diff +
     Award_players_diff - 1, family = "binomial")

Coefficients:
                            Estimate Std. Error z value Pr(>|z
                               |)
FIFA_rank_diff            0.0021119  0.0013819   1.528
    0.1265
Distance_travelled_diff -0.0001386  0.0000755  -1.836
    0.0664 .
Award_players_diff        0.7289353  0.3144682   2.318
    0.0204 *
---

    Null deviance: 74.86  on 54  degrees of freedom
Residual deviance: 47.92  on 51  degrees of freedom
AIC: 53.92
```

A Wald test on FIFA_rank_diff does not show significance at a 95% confidence level. However, it is close, and for the purposes of prediction this may be good enough. Examining plots of the predictors against the log odds shows a plausible relationship, and the reduction in deviance between the null model and our model is 26.94, which is greater than the upper 5% point of a chi-squared random variable with 3 degrees of freedom (7.814728). Therefore the deviance test indicates our model is significantly better than the null model.

Goodness of Fit

To test that the model with these three predictors is better than any subset, each predictor was dropped in the following R output.

```
> drop1(match_model)
Single term deletions

Model:
Response ~ FIFA_rank_diff + Distance_travelled_diff +
    Award_players_diff -
     1
                         Df Deviance    AIC
<none>                        47.920 53.920
FIFA_rank_diff            1   50.376 54.376
```

```
Distance_travelled_diff  1   51.826 55.826
Award_players_diff       1   55.926 59.926
```

The minimum values of both AIC and Deviance are at the full model - therefore it was kept as the best model found thus far.

Predictive Accuracy

There are several ways that the models produced can be assessed for predictive accuracy.

The PRESS Statistic was calculated for the model with 3 predictors and then removing one at a time:

```
Model                                PRESS
Response~<none>                      54.43639
Response~FIFA_rank_diff              54.19239
Response~Distance_travelled_diff     58.55649
Response~Award_players_diff          74.8599
```

The two larger models show roughly the same predictive residual error, and so are both equal best of the models evaluated.

We can also more rigorously test models by partitioning our data into test and training sets - a common split is 80% training data, 20% test data. Models created on the training data set are then tested for accuracy among the test data they were not trained on. This is a good way to test for over fitting and out of sample predictive error, but was not in the scope of this assignment and therefore not carried out.

Limitations

There are several limitations faced by this model:

- The primary limitation faced is the small training data set - after removing the tie games and final match we were left with only 54 data points. If data had been available in an easier format or more scope was given it may have been possible to collect data from multiple seasons, enabling stronger predictions and the ability to split out test data from the training data.
- This model also does not tolerate tie games, meaning 9/63 or 14% of the data available had to be discarded. If Ordinal Logistic Regression, Multinomial Regression or Poisson regression were utilised, these may have been included in the training data.

Results

A data frame with the values for the Germany vs. Argentina game was then created and a prediction run.

```
> newdata_dist_travelled = 2355-9682
> newdata_award_players = 2 - 5
> newdata_fif = 1175-1300
> newdataframe=data.frame(Distance_travelled_diff=
    newdata_dist_travelled, FIFA_rank_diff=newdata_fif,
    Award_players_diff=newdata_award_players)
> prediction = predict.glm(match_model, newdata=newdataframe
    , se.fit=T, level=0.95, interval="prediction")

> confidence_interval = c(prediction$fit-prediction$se.fit,
    prediction$fit+prediction$se.fit)
> prediction$fit
         1
-1.435332
> confidence_interval
          1          1
-2.4326826 -0.4379817
```

The model indicates with 95% confidence that the log odds will be below 0 - this is equivalent to predicting that the away team wins in our model. Thus the model successfully predicts the winner of the 2014 FIFA World Cup would be the away team, Germany.

Conclusion

To answer the two questions from the Introduction:

If you were to try to predict the winner before it happened, would you have been successful?

The model constructed was successful in predicting the winner so yes, it would have been possible to construct a model a day before the match that successfully predicted the game winner.

Is it possible to successfully predict soccer games (and sports in general) over the long term using regression?

Review of the literature has shown that yes, it is also possible to do this in the long run. There are companies working in the field of sabermetrics (baseball analysis) that make profit solely from predicting outcomes better than the bookies, as well as numerous other professional gamblers out there. However, this requires a lot of time and constant updating to do correctly. A prediction made before the cup by Goldman Sachs picked Brazil as the winner, according to their own model. Had Brazil's best player not suffered a knee injury, they might have been right, but it is precisely these last minute events that statistical models often fail to incorporate well.

In professional settings, a mathematical model is usually augmented by an expert odds-setter, rather than being relied on alone. A model is first used to make a prediction, taking into account predictors such as historical records or player stats. The odds-setter then adjusts these predicted odds to account for, say, an injury to a star player. When dealing with uncertainty and prediction, chance will always be a factor, and as of today, the best results are obtained through the combination of mathematical insight and human insight together.

Appendix A

Data Sources and References

Data

players.csv

https://datahub.io/dataset/fifa-world-cup-2014-all-players

matches.csv

http://www.kdnuggets.com/2015/01/data-mining-text-analytics-world-cup-2014.html

possession.csv

http://resources.fifa.com/mm/document/footballdevelopment/ technicalsupport/02/42/15/40/2014fwc_tsg_report_15082014web_neutral.pdf

distances.csv

https://allagora.wordpress.com/2014/07/29/does-the-traveling-distance-affect-the-results-of-the-fifa-world-cup/

FIFA rankings

http://www.fifa.com/fifa-world-ranking/ranking-table/men/rank=239/index.html
http://www.fifa.com/fifa-world-ranking/ranking-table/men/rank=227/index.html

B'allon d'or nominees

https://en.wikipedia.org/wiki/2014_FIFA_Ballon_d%27Or

Other useful sources

http://eandt.theiet.org/magazine/2010/08/predicting-football.cfm

http://www.statisticssolutions.com/assumptions-of-logistic-regression/

http://www.mit.edu/ vgalle/files/WCPredictions.pdf

http://www.goldmansachs.com/our-thinking/outlook/world-cup-sections/world-cup-book-2014-statistical-model.html

Appendix B

R Code

```
#FIFA Regression
#Author - Nathan Wilson, z3287546
#2016-08-13

#====================================================
#==================== load data =====================
#====================================================

matches_2014_raw <- read.csv("matches.csv", header=T)
players_2014_raw <- read.csv("players.csv", header=T)
possession_2014_raw <- read.csv("possession.csv", header = F)

#must remove last match data for the purposes of our analysis
matches_2014_raw <- matches_2014_raw[-64,]
possession_2014_raw <- possession_2014_raw[-64,]


#====================================================
#========= calculate statistics for teams ==========
#====================================================

#teams vector
teams = levels(matches_2014_raw$home_team.country)

#1 if the country has home team advantage (ie is brazil)
home_team_advantage = as.numeric(teams == "Brazil")

#distance in km travelled to brazil
team_distance_from_home =
 c(8074,2355,15577,9053,9594,0,7275,2844,
 3218,4422,9510,3314,8849,8807,9682,6114,
 9763,4998,12117,9064,5652,17361,17417,6378,
 9182,7045,7376,11282,7849,9000,2334,6458)

#avg possession over this FIFA cup
all_teams_by_game = data.frame(unlist(list(
    matches_2014_raw$home_team.country,
    matches_2014_raw$away_team.country)))
all_possession_by_game = c(possession_2014_raw$V1,
    possession_2014_raw$V2)
avg_possession = aggregate(all_possession_by_game,
    all_teams_by_game, mean)[2]

```

```
40 #team FIFA rating
41 team_rankings =
42  c(858,1175,526,1074,873,1242,558,1026,1137,
43  762,903,791,1090,913,1300,704,1064,731,641,
44  1104,809,626,547,882,981,640,1189,893,1485,1149,1147,1035)
45
46 #increase in team FIFA rank over last 12 months
47 team_rankings_increase_past_year =
48 c(13, -2, -15, -1,-6,18,9,11,-1,20,-14,-16,
49  -1,1,0,-16,4,19,24,-1,-10,-14,-17,-3,-10
50  ,-13,2,-8,0,8,12,15)
51
52 #average player age
53 avg_player_age = subset(aggregate(players_2014_raw$Age, list
       (Country=players_2014_raw$Club..country), mean), Country
       %in% teams)
54
55 #average player height
56 avg_player_height = subset(aggregate(players_2014_raw$Height
       ..cm, list(Country=players_2014_raw$Club..country), mean)
       , Country %in% teams)
57
58 #number of ballon dor nominees
59 team_ballon_nominees = c(0,2,0,2,0,1,0,0,0,0,0,0,0
60 ,2,5,0,0,0,0,0,1,0,0,0,1,0,0,0,3,0,0,0)
61
62 #avg goals
63 all_goals_by_game = c(matches_2014_raw$home_team.goals,
       matches_2014_raw$away_team.goals)
64 avg_goals = aggregate(all_goals_by_game, all_teams_by_game,
       mean)[2]
65
66 #avg concessions
67 all_concessions_by_game = c(matches_2014_raw$away_team.goals
       , matches_2014_raw$home_team.goals)
68 avg_concessions = aggregate(all_concessions_by_game,
       all_teams_by_game, mean)[2]
69
70 #team data
71 team_data = data.frame(Team=teams,
72                        Home_adv=home_team_advantage,
73                        Distance_travelled=
                               team_distance_from_home,
74                        Possession=avg_possession$x,
75                        FIFA_rank=team_rankings,
76                        FIFA_rank_increase=
                               team_rankings_increase_past_year,
77                        Player_age=avg_player_age$x,
78                        Player_height=avg_player_height$x,
79                        Award_players=team_ballon_nominees,
```

```
80                          Recent_goals=avg_goals$x,
81                          Recent_concessions=avg_concessions$x)
82
83
84  #=================================================
85  #= Create data frame with predictors and responses =
86  #=================================================
87
88  #response is 1 for home win, 0.5 for a tie and 0 for away
        win
89  home_wins = as.numeric(matches_2014_raw$winner_code == as.
        character(matches_2014_raw$home_team.code))
90  ties = as.numeric(matches_2014_raw$winner_code == "Draw") *
        0.5
91  responses = home_wins + ties
92
93  home_adv_fac = team_data$Home_adv[Team=
        matches_2014_raw$home_team.country] - team_data$Home_adv[
        Team=matches_2014_raw$away_team.country]
94  home_adv_fac = ifelse(home_adv_fac==1, "HOME", home_adv_fac)
95  home_adv_fac = ifelse(home_adv_fac=="-1", "AWAY",
        home_adv_fac)
96  home_adv_fac = ifelse(home_adv_fac=="0", "NEUTRAL",
        home_adv_fac)
97
98  match_data = data.frame(Response=responses,
99                          Home_adv=factor(home_adv_fac),
100                         Distance_travelled_diff=
                                team_data$Distance_travelled[Team
                                =matches_2014_raw$home_team.
                                country] -
                                team_data$Distance_travelled[Team
                                =matches_2014_raw$away_team.
                                country],
101                         Possession_diff=team_data$Possession
                                [Team=matches_2014_raw$home_team.
                                country] - team_data$Possession[
                                Team=matches_2014_raw$away_team.
                                country],
102                         FIFA_rank_diff=team_data$FIFA_rank[
                                Team=matches_2014_raw$home_team.
                                country] - team_data$FIFA_rank[
                                Team=matches_2014_raw$away_team.
                                country],
103                         FIFA_rank_increase_diff=
                                team_data$FIFA_rank_increase[Team
                                =matches_2014_raw$home_team.
                                country] -
                                team_data$FIFA_rank_increase[Team
                                =matches_2014_raw$away_team.
```

```
                            country],
104                         Player_age_diff=team_data$Player_age
                            [Team=matches_2014_raw$home_team.
                            country] - team_data$Player_age[
                            Team=matches_2014_raw$away_team.
                            country],
105                         Player_height_diff=
                            team_data$Player_height[Team=
                            matches_2014_raw$home_team.
                            country] -
                            team_data$Player_height[Team=
                            matches_2014_raw$away_team.
                            country],
106                         Award_players_diff=
                            team_data$Award_players[Team=
                            matches_2014_raw$home_team.
                            country] -
                            team_data$Award_players[Team=
                            matches_2014_raw$away_team.
                            country],
107                         Recent_goals_diff=
                            team_data$Recent_goals[Team=
                            matches_2014_raw$home_team.
                            country] - team_data$Recent_goals
                            [Team=matches_2014_raw$away_team.
                            country],
108                         Recent_concessions_diff=
                            team_data$Recent_concessions[Team
                            =matches_2014_raw$home_team.
                            country] -
                            team_data$Recent_concessions[Team
                            =matches_2014_raw$away_team.
                            country])
109
110 #drop tie games
111 match_data = subset(match_data, Response != 0.5)
112
113 #================================================
114 #=== Analyse for multicollinearity ==============
115 #================================================
116
117 #Plot predictors against log odds of response
118 par(mfrow = c(1,1))
119 attach(match_data)
120
121 #Plot boxplot of home adv vs response
122 boxplotframe = data.frame(response=Response, homeadv=
        Home_adv)
123 plot.design(boxplotframe)
124
```

```
125 #pairwise plots
126 pairs(match_data)
127
128 #remove recent goals and concessions after looking at
        pairwise plot
129 match_data = subset(match_data, select = -c(
        Recent_goals_diff, Recent_concessions_diff) )
130
131 #=================================================
132 #=== Fit model ====================================
133 #=================================================
134 match_model = glm(Response~Home_adv+Distance_travelled_diff+
        Possession_diff+FIFA_rank_diff+FIFA_rank_increase_diff
135                 +Player_age_diff+Player_height_diff+
                        Award_players_diff, family="binomial")
136
137 #=================================================
138 #=== Model Diagnostics and Tests===================
139 #=================================================
140
141 #plot predictors vs log odds
142 logresponse=log(match_model$fitted.values/(1-
        match_model$fitted.values))
143 par(mfrow=c(4,2))
144 plot(Distance_travelled_diff,logresponse)
145 plot(Possession_diff,logresponse)
146 plot(FIFA_rank_diff,logresponse)
147 plot(FIFA_rank_increase_diff,logresponse)
148 plot(Player_age_diff,logresponse)
149 plot(Player_height_diff,logresponse)
150 plot(Award_players_diff,logresponse)
151
152 #view summary
153 summary(match_model)
154
155 #Wald test indicates we should drop HOME_ADV, PLAYER_AGE,
        FIFA_RANK_INCREASE_DIFF, POSSESSION_DIFF,HEIGHT_DIFF
156 match_model = glm(Response~FIFA_rank_diff+
        Distance_travelled_diff+Award_players_diff-1, family="
        binomial")
157 summary(match_model)
158
159 #deviance test
160 qchisq(0.95, 3)
161 74.86 - 47.92
162
163 #drop1 test
164 drops =drop1(match_model)
165
166 #press statistics per model
```

```
167 pr <- (resid(match_model)/(1 - lm.influence(match_model)$hat
        ))
168 press1=sum(pr^2)
169 match_model_drop1 = glm(Response~Distance_travelled_diff+
        Award_players_diff-1, family="binomial")
170 pr <- (resid(match_model_drop1)/(1 - lm.influence(
        match_model_drop1)$hat))
171 press2=sum(pr^2)
172 match_model_drop2 = glm(Response~Award_players_diff-1,
        family="binomial")
173 pr <- (resid(match_model_drop2)/(1 - lm.influence(
        match_model_drop2)$hat))
174 press3=sum(pr^2)
175 match_model_drop3 = glm(Response~-1, family="binomial")
176 pr <- (resid(match_model_drop3)/(1 - lm.influence(
        match_model_drop3)$hat))
177 press4=sum(pr^2)
178
179 #====================================================
180 #=== Model Prediction Test ==========================
181 #====================================================
182 newdata_dist_travelled = 2355-9682
183 newdata_award_players = 2 - 5
184 newdata_fif = 1175-1300
185 newdataframe=data.frame(Distance_travelled_diff=
        newdata_dist_travelled, FIFA_rank_diff=newdata_fif,
        Award_players_diff=newdata_award_players)
186 prediction = predict.glm(match_model, newdata=newdataframe,
        se.fit=T, level=0.95, interval="prediction")
187 confidence_interval = c(prediction$fit-prediction$se.fit,
        prediction$fit+prediction$se.fit)
188 confidence_interval
```

Appendix C

Pairwise Plots of Independent Variables