Modeling Match Results in Soccer using a Hierarchical Bayesian Poisson Model

Rasmus Bååth, Lund University Cognitive Science, rasmus.baath@gmail.com

Abstract

This technical report is a slightly modified version of my submission to the UseR 2013 Data Analysis Contest which I had the fortune of winning. The purpose of the contest was to do something interesting with a dataset consisting of the match results from the last five seasons of La Liga, the premium Spanish soccer league. In total there were 1900 rows in the dataset each with information regarding which was the home and away team, what these teams scored and what season it was. I decided to develop a Bayesian model of the distribution of the match end scores.

Ok, first I should come clean and admit that I know nothing about football. Sure, I've watched Sweden loose to Germany in the World Cup a couple of times, but that's it. Never the less, here I will show an attempt to to model the goal outcomes in the La Liga data set provided as part of the UseR 2013 data analysis contest. My goal is not only to model the outcomes of matches in the data set but also to be able to (a) calculate the odds for possible goal outcomes of future matches and (b) to produce a credible ranking of the teams. The model I will be developing is a Bayesian hierarchical model where the goal outcomes will be assumed to be distributed according to a Poisson distribution. I will focus more on showing all the cool things you can *easily* calculate in R when you have a fully specified Bayesian Model and focus less on model comparison and trying to find the model with highest predictive accuracy (even though I believe my model is pretty good). I really would like to see somebody *try* to do a similar analysis in SPSS (which most people uses in my field, psychology). It would be a pain!

This report[1] assumes some familiarity with Bayesian modeling and Markov chain Monte Carlo. If you're not into Bayesian statistics you're missing out on something really great and a good way to get started is by reading the excellent Doing Bayesian Data Analysis by John Kruschke. The tools I will be using is R (of course) with the model implemented in JAGS called from R using the rjags package. For plotting the result of the MCMC samples generated by JAGS I'll use the coda package, the mcmcplots package, and the `plotPost` function courtesy of John Kruschke. For data manipulation I used the plyr and stringr packages and for general plotting I used ggplot2. This report was written in Rstudio using knitr and xtable. The full R code presented in this report can be downloaded here: http://www.sumsar.net/files/academia/useR2013_modeling_contest_baath.zip .

I start by loading libraries, reading in the data and preprocessing it for JAGS. The last 50 matches have unknown outcomes and I create a new data frame `d` holding only matches with known outcomes. I will come back to the unknown outcomes later when it is time to use my model for prediction.

```
library(rjags)
library(coda)
library(mcmcplots)
library(stringr)
library(plyr)
```

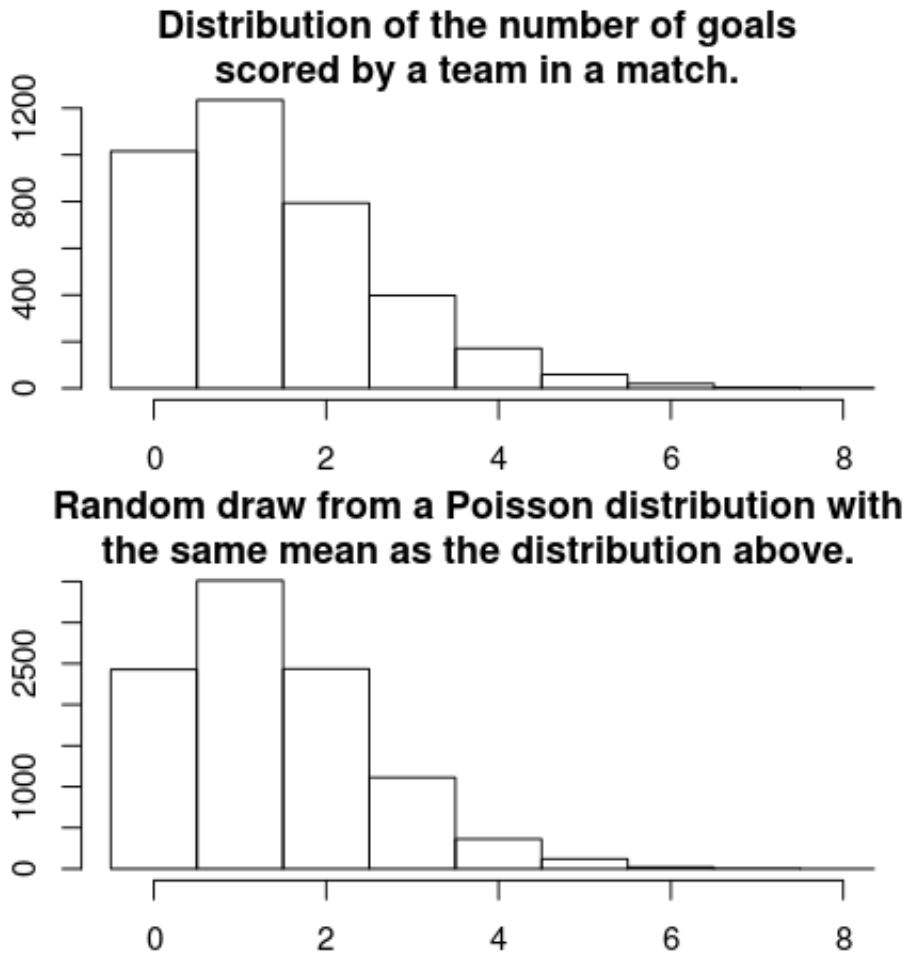
[1] Published as:
Bååth, R. (2015) Modeling Match Results in Soccer using a Hierarchical Bayesian Poisson Model. *LUCS Minor*, 18. (ISSN 1104-1609)
~
Bibtex:
@techreport{baathsoccer2015,
author = {Rasmus Baath},
title = {Modeling Match Results in Soccer using a Hierarchical Bayesian Poisson Model},
institution = {Lund University Cognitive Science},
year = 2015,
number = {LUCS minor 18}
}

```r
library(xtable)
source("plotPost.R")
set.seed(12345)  # for reproducibility

load("laliga.RData")

# -1 = Away win, 0 = Draw, 1 = Home win
laliga$MatchResult <- sign(laliga$HomeGoals - laliga$AwayGoals)

# Creating a data frame d with only the complete match results
d <- na.omit(laliga)
teams <- unique(c(d$HomeTeam, d$AwayTeam))
seasons <- unique(d$Season)

# A list for JAGS with the data from d where the strings are coded as
# integers
data_list <- list(HomeGoals = d$HomeGoals, AwayGoals = d$AwayGoals,
    HomeTeam = as.numeric(factor(d$HomeTeam, levels = teams)),
    AwayTeam = as.numeric(factor(d$AwayTeam, levels = teams)),
    Season = as.numeric(factor(d$Season, levels = seasons)), n_teams = length(teams),
    n_games = nrow(d), n_seasons = length(seasons))

# Convenience function to generate the type of column names Jags outputs.
col_name <- function(name, ...) {
    paste0(name, "[", paste(..., sep = ","), "]")
}
```

Modeling Match Results: Iteration 1

How are the number of goals for each team in a football match distributed? Well, let's start by assuming that all football matches are roughly equally long, that both teams have many chances at making a goal and that each team have the same probability of making a goal each goal chance. Given these assumptions the distribution of the number of goals for each team should be well captured by a Poisson distribution. A quick and dirty comparison between the actual distribution of the number of scored goals and a Poisson distribution having the same mean number of scored goals support this notion.

```r
par(mfcol = c(2, 1), mar = rep(2.2, 4))
hist(c(d$AwayGoals, d$HomeGoals), xlim = c(-0.5, 8), breaks = -1:9 + 0.5,
  main = "Distribution of the number of goals\nscored by a team in a match.")
mean_goals <- mean(c(d$AwayGoals, d$HomeGoals))
hist(rpois(9999, mean_goals), xlim = c(-0.5, 8), breaks = -1:9 + 0.5,
  main = "Random draw from a Poisson distribution with\nthe same mean as the distribution above.")
```



All teams aren't equally good (otherwise Sweden would actually win the world cup now and then) and it will be assumed that all teams have a latent skill variable and the skill of a team *minus* the skill of the opposing team defines the predicted outcome of a game. As the number of goals are assumed to be Poisson distributed it is natural that the skills of the teams are on the log scale of the mean of the distribution. The distribution of the number of goals for team i when facing team j is then

$$Goals \sim \text{Poisson}(\lambda)$$
$$\log(\lambda) = \text{baseline} + \text{skill}_i - \text{skill}_j$$

where baseline is the log average number of goals when both teams are equally good. The goal outcome of a match between home team i and away team j is modeled as:

$$HomeGoals_{i,j} \sim \text{Poison}(\lambda_{\text{home},i,j})$$
$$AwayGoals_{i,j} \sim \text{Poison}(\lambda_{\text{away},i,j})$$
$$\log(\lambda_{\text{home},i,j}) = \text{baseline} + \text{skill}_i - \text{skill}_j$$
$$\log(\lambda_{\text{away},i,j}) = \text{baseline} + \text{skill}_j - \text{skill}_i$$

Add some priors to that and you've got a Bayesian model going! I set the prior distributions over the baseline and the skill of all n teams to:

$$\text{baseline} \sim \text{Normal}(0, 4^2)$$
$$\text{skill}_{1...n} \sim \text{Normal}(\mu_{\text{teams}}, \sigma^2_{\text{teams}})$$
$$\mu_{\text{teams}} \sim \text{Normal}(0, 4^2)$$
$$\sigma_{\text{teams}} \sim \text{Uniform}(0, 3)$$

Since I know nothing about football these priors are made very vague. For example, the prior on the baseline have a SD of 4 but since this is on the log scale of the mean number of goals it corresponds to one SD from the mean 0 covering the range of $[0.02, 54.6]$ goals. A very wide prior indeed.

Turning this into a JAGS model requires some minor adjustments. The model have to loop over all the match results, which adds some for-loops. JAGS parameterizes the normal distribution with precision (the reciprocal of the variance) instead of variance so the hyper priors have to be converted. Finally I have to "anchor" the skill of one team to a constant otherwise the mean skill can drift away freely. Doing these adjustments results in the following model description:

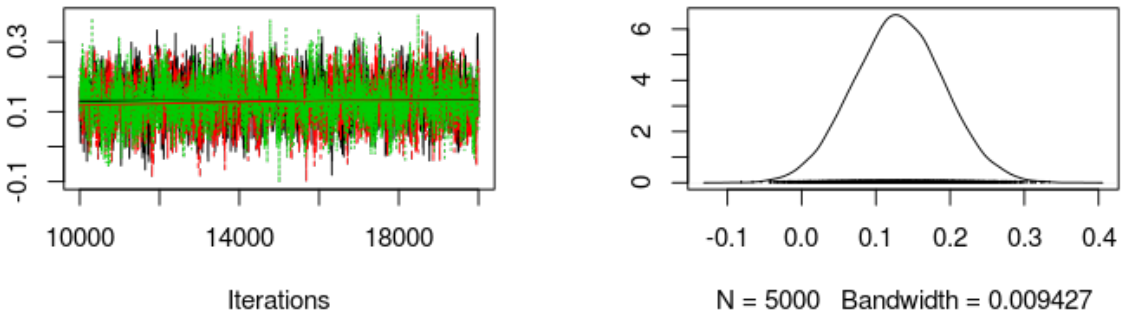
```
m1_string <- "model {
for(i in 1:n_games) {
  HomeGoals[i] ~ dpois(lambda_home[HomeTeam[i],AwayTeam[i]])
  AwayGoals[i] ~ dpois(lambda_away[HomeTeam[i],AwayTeam[i]])
}

for(home_i in 1:n_teams) {
  for(away_i in 1:n_teams) {
    lambda_home[home_i, away_i] <- exp(baseline + skill[home_i] - skill[away_i])
    lambda_away[home_i, away_i] <- exp(baseline + skill[away_i] - skill[home_i])
  }
}

skill[1] <- 0
for(j in 2:n_teams) {
  skill[j] ~ dnorm(group_skill, group_tau)
}

group_skill ~ dnorm(0, 0.0625)
group_tau <- 1 / pow(group_sigma, 2)
group_sigma ~ dunif(0, 3)
baseline ~ dnorm(0, 0.0625)
}"
```

I can then run this model directly from R using rjags and the handy `textConnection` function. This takes a couple of minutes on my computer, roughly enough for a coffee break.

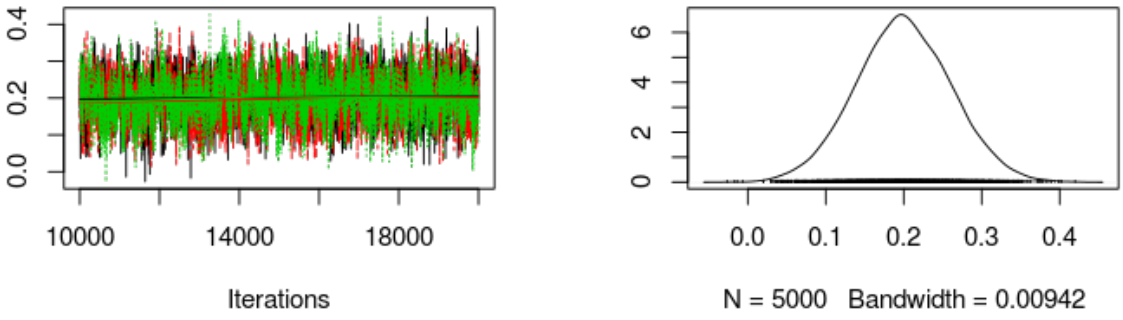
```
# Compiling model 1
m1 <- jags.model(textConnection(m1_string), data = data_list, n.chains = 3,
    n.adapt = 5000)
# Burning some samples on the altar of the MCMC god
update(m1, 5000)
# Generating MCMC samples
s1 <- coda.samples(m1, variable.names = c("baseline", "skill", "group_skill",
    "group_sigma"), n.iter = 10000, thin = 2)
# Merging the three MCMC chains into one matrix
ms1 <- as.matrix(s1)
```

Using the generated MCMC samples I can now look at the credible skill values of any team. Let's look at the trace plot and the distribution of the skill parameters for FC Sevilla and FC Valencia.

```
plot(s1[, col_name("skill", which(teams == "FC Sevilla"))])
```




```
plot(s1[, col_name("skill", which(teams == "FC Valencia"))])
```



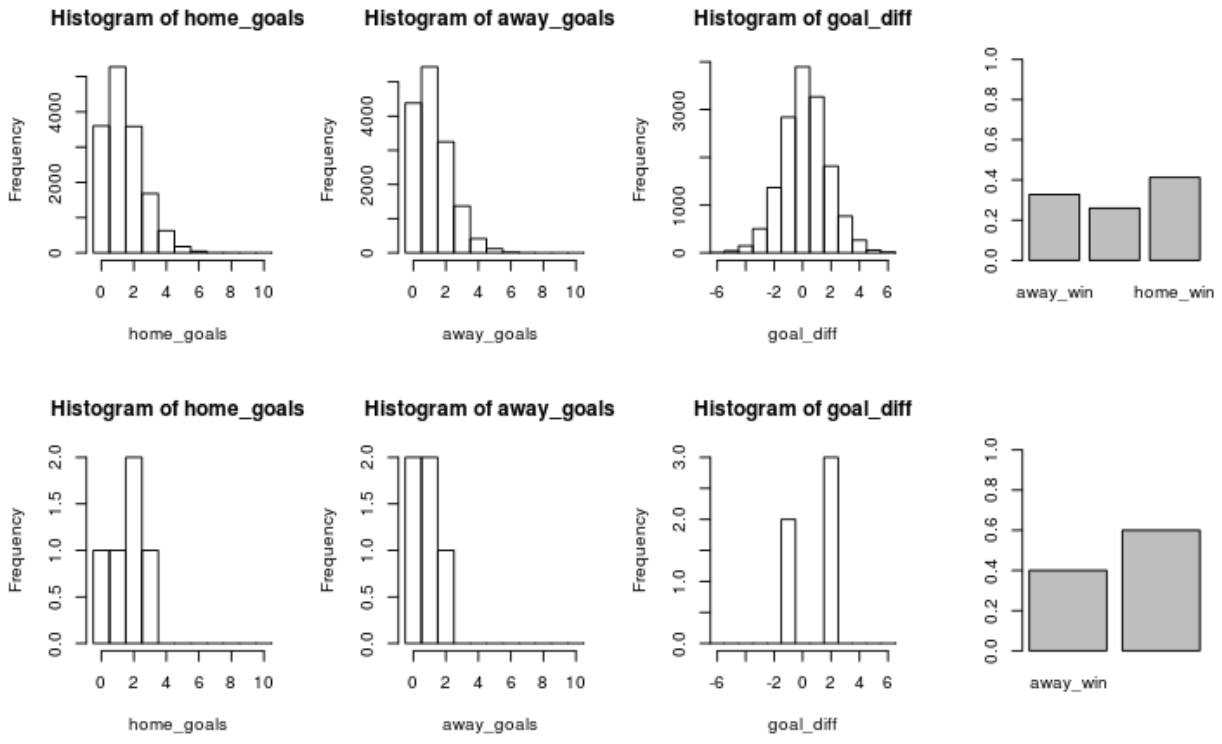
Seems like Sevilla and Valencia have similar skill with Valencia being slightly better. Using the MCMC samples it is not only possible to look at the distribution of parameter values but it is also straight forward to simulate matches between teams and look at the credible distribution of number of goals scored and the probability of a win for the home team, a win for the away team or a draw. The following functions simulates matches with one team as home team and one team as away team and plots the predicted result together with the actual outcomes of any matches in the `laliga` data set.

```
# Plots histograms over home_goals, away_goals, the difference in goals
# and a barplot over match results.
plot_goals <- function(home_goals, away_goals) {
    n_matches <- length(home_goals)
    goal_diff <- home_goals - away_goals
    match_result <- ifelse(goal_diff < 0, "away_win", ifelse(goal_diff > 0,
        "home_win", "equal"))
    hist(home_goals, xlim = c(-0.5, 10), breaks = (0:100) - 0.5)
```

```
    hist(away_goals, xlim = c(-0.5, 10), breaks = (0:100) - 0.5)
    hist(goal_diff, xlim = c(-6, 6), breaks = (-100:100) - 0.5)
    barplot(table(match_result)/n_matches, ylim = c(0, 1))
}


plot_pred_comp1 <- function(home_team, away_team, ms) {
    # Simulates and plots game goals scores using the MCMC samples from the m1
    # model.
    par(mfrow = c(2, 4))
    baseline <- ms[, "baseline"]
    home_skill <- ms[, col_name("skill", which(teams == home_team))]
    away_skill <- ms[, col_name("skill", which(teams == away_team))]
    home_goals <- rpois(nrow(ms), exp(baseline + home_skill - away_skill))
    away_goals <- rpois(nrow(ms), exp(baseline + away_skill - home_skill))
    plot_goals(home_goals, away_goals)
    # Plots the actual distribution of goals between the two teams
    home_goals <- d$HomeGoals[d$HomeTeam == home_team & d$AwayTeam == away_team]
    away_goals <- d$AwayGoals[d$HomeTeam == home_team & d$AwayTeam == away_team]
    plot_goals(home_goals, away_goals)
}
```

Let's look at Valencia (home team) vs. Sevilla (away team). The graph below shows the simulation on the first row and the historical data on the second row.

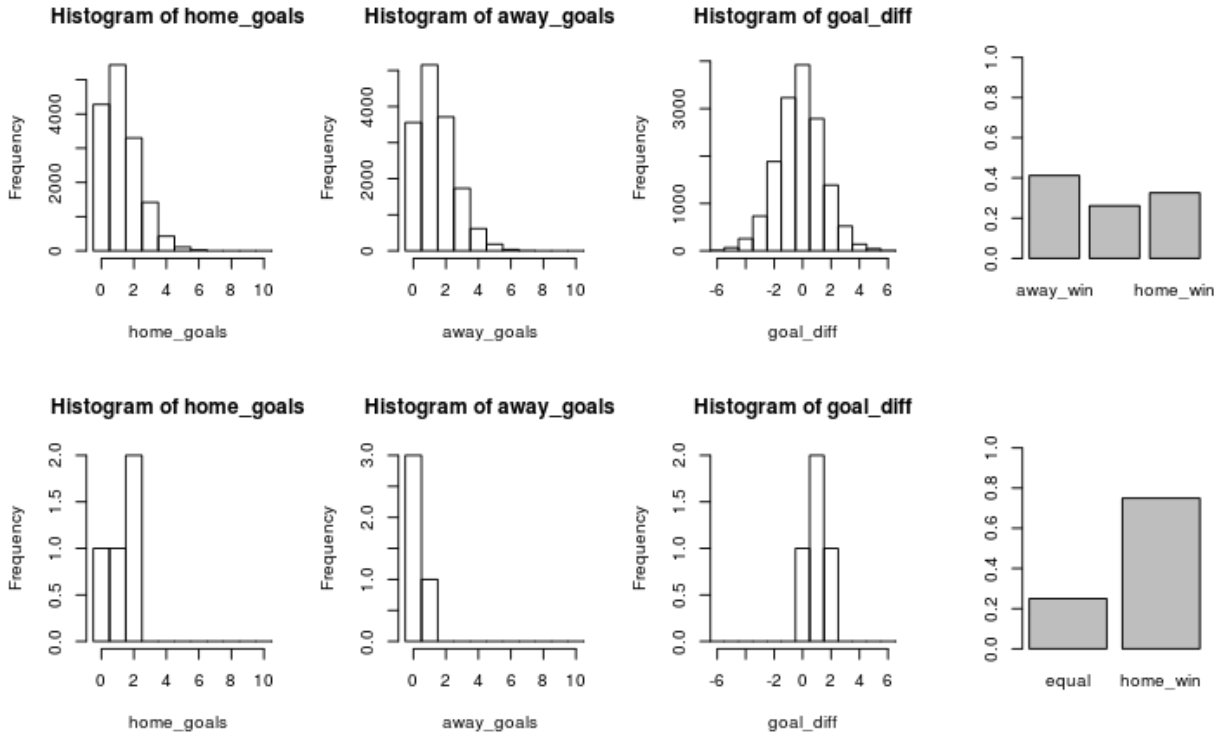
```
plot_pred_comp1("FC Valencia", "FC Sevilla", ms1)
```



The simulated data fits the historical data reasonably well and both the historical data and the simulation

show that Valencia would win with a slightly higher probability that Sevilla. Let's swap places and let Sevilla be the home team and Valencia be the away team.

```
plot_pred_comp1("FC Sevilla", "FC Valencia", ms1)
```



Here we discover a problem with the current model. While the simulated data looks the same, except that the home team and the away team swapped places, the historical data now shows that Sevilla often wins against Valencia when being the home team. Our model doesn't predict this because it doesn't considers the advantage of being the home team. This is fortunately easy to fix as I will show in the next iteration of the model.

Modeling Match Results: Iteration 2

The only change to the model needed to account for the home advantage is to split the baseline into two components, a home baseline and an away baseline. The following JAGS model implements this change by splitting `baseline` into `home_baseline` and `away_baseline`.

```
# model 2
m2_string <- "model {
for(i in 1:n_games) {
  HomeGoals[i] ~ dpois(lambda_home[HomeTeam[i],AwayTeam[i]])
  AwayGoals[i] ~ dpois(lambda_away[HomeTeam[i],AwayTeam[i]])
}

for(home_i in 1:n_teams) {
  for(away_i in 1:n_teams) {
    lambda_home[home_i, away_i] <-
      exp( home_baseline + skill[home_i] - skill[away_i])
```

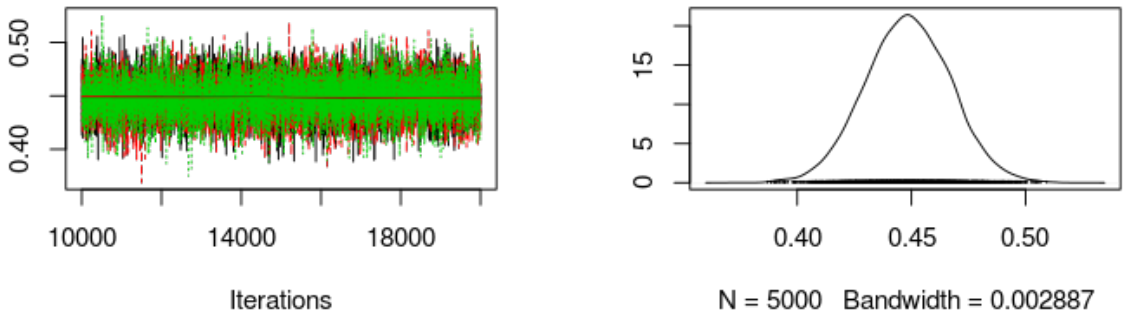
```
    lambda_away[home_i, away_i] <-
      exp( away_baseline + skill[away_i] - skill[home_i])
  }
}

skill[1] <- 0
for(j in 2:n_teams) {
  skill[j] ~ dnorm(group_skill, group_tau)
}

group_skill ~ dnorm(0, 0.0625)
group_tau <- 1/pow(group_sigma, 2)
group_sigma ~ dunif(0, 3)

home_baseline ~ dnorm(0, 0.0625)
away_baseline ~ dnorm(0, 0.0625)
}"
```

Another cup of coffee while we run the MCMC simulation...

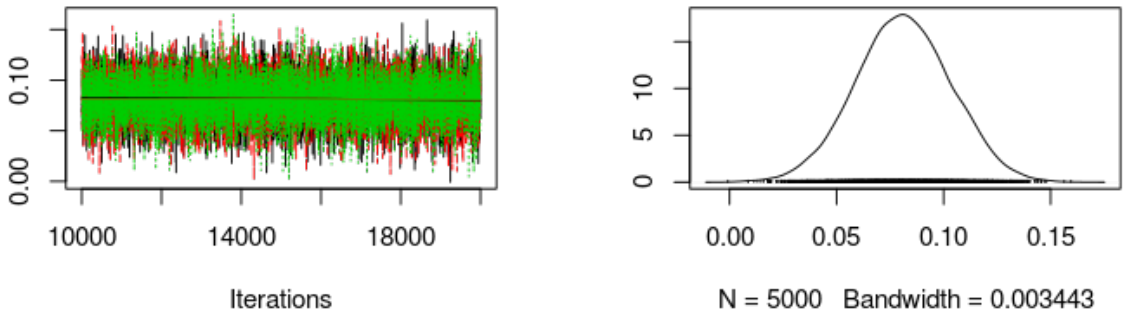
```
m2 <- jags.model(textConnection(m2_string), data = data_list, n.chains = 3,
    n.adapt = 5000)
update(m2, 5000)
s2 <- coda.samples(m2, variable.names = c("home_baseline", "away_baseline",
    "skill", "group_sigma", "group_skill"), n.iter = 10000, thin = 2)
ms2 <- as.matrix(s2)
```

Looking at the trace plots and distributions of `home_baseline` and `away_baseline` shows that there is a considerable home advantage.

```
plot(s2[, "home_baseline"])
```

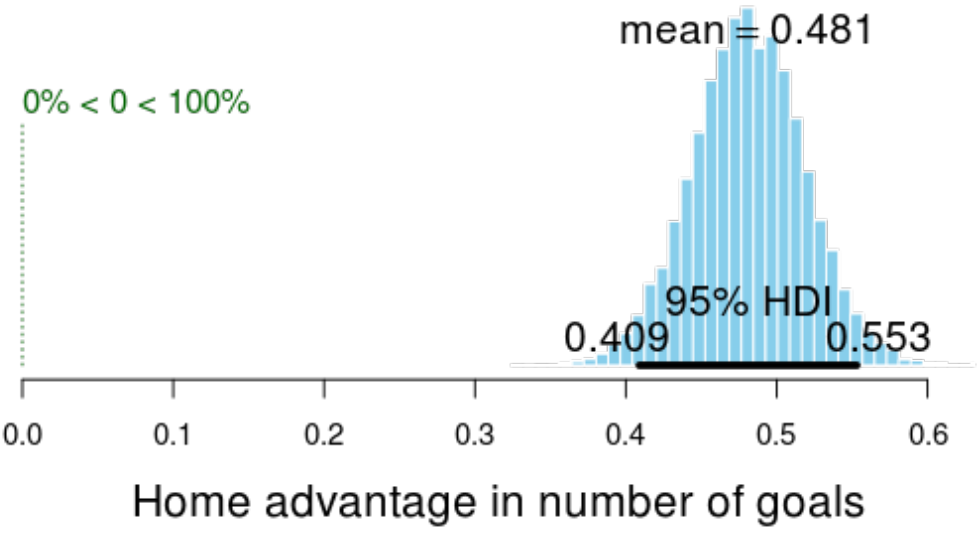



```
plot(s2[, "away_baseline"])
```



Looking at the difference between `exp(home_baseline)` and `exp(away_baseline)` shows that the home advantage is realized as roughly 0.5 more goals for the home team.

```
plotPost(exp(ms2[, "home_baseline"]) - exp(ms2[, "away_baseline"]), compVal = 0,
    xlab = "Home advantage in number of goals")
```



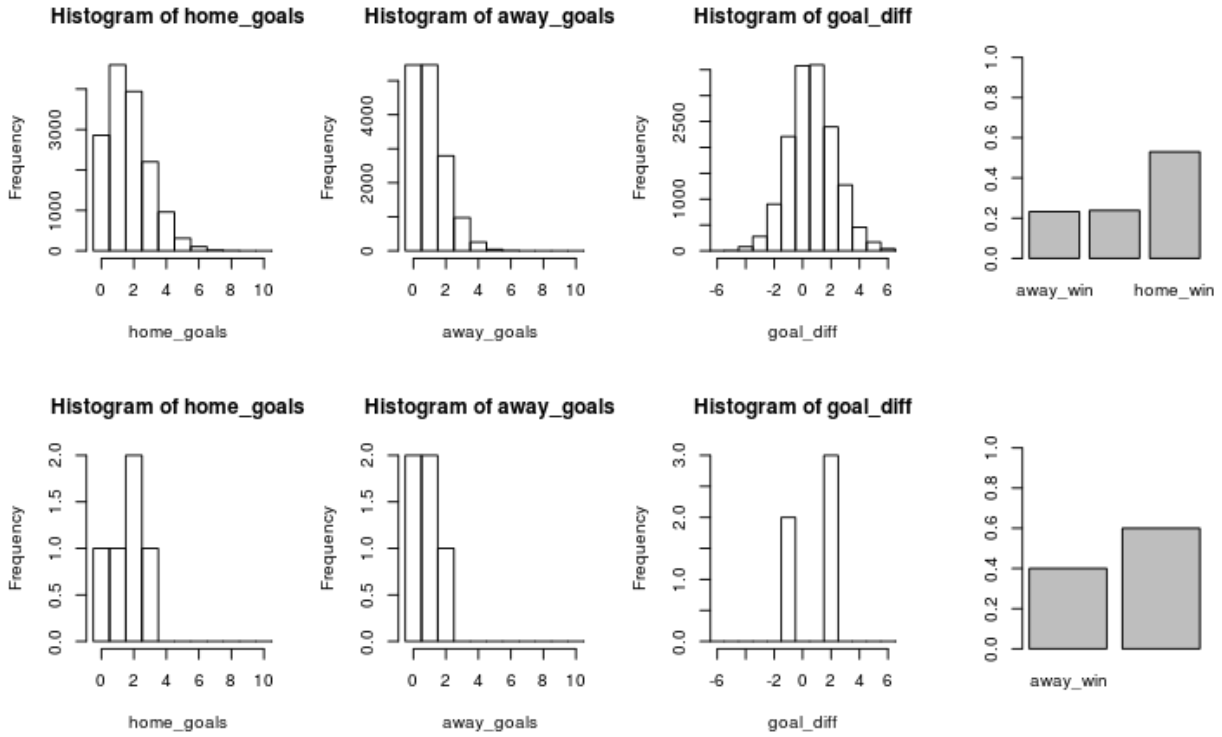
Comparing the DIC of the of the two models also indicates that the new model is better.

```
dic_m1 <- dic.samples(m1, 10000, "pD")
dic_m2 <- dic.samples(m2, 10000, "pD")
diffdic(dic_m1, dic_m2)
```

```
## Difference: 167
## Sample standard error: 26.02
```

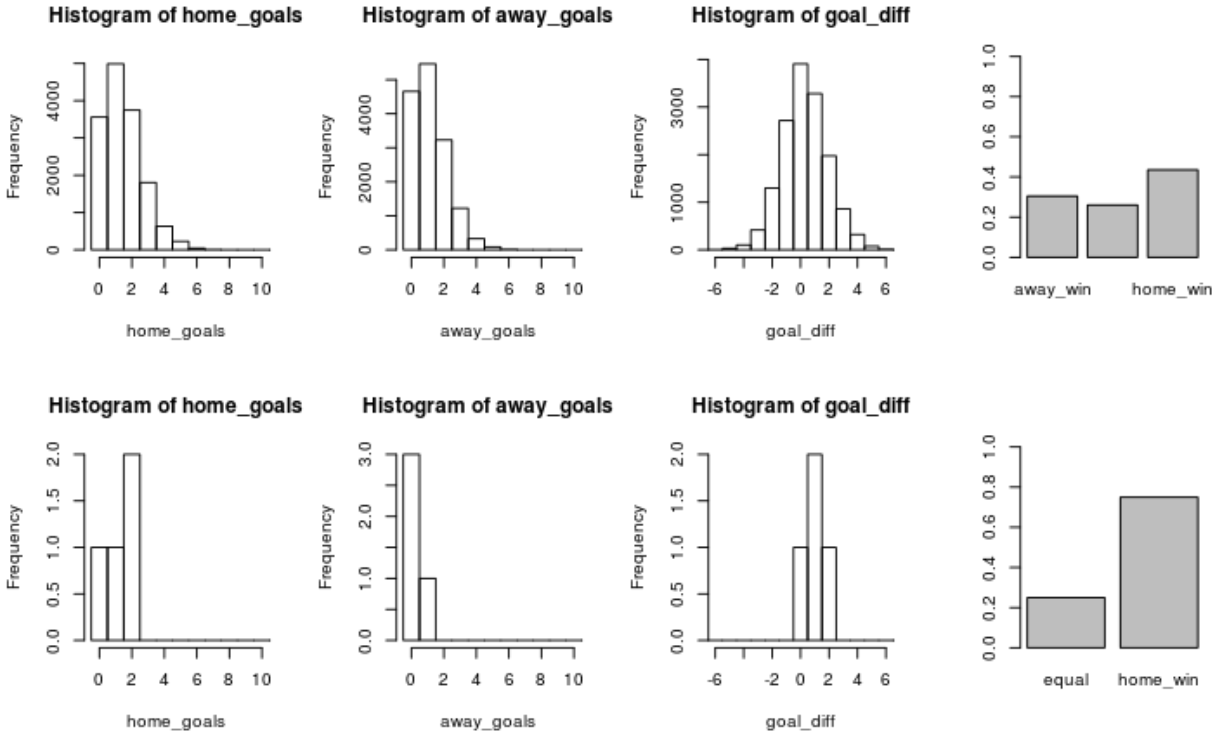
Finally we'll look at the simulated results for Valencia (home team) vs Sevilla (away team) using the estimates from the new model with the first row of the graph showing the predicted outcome and the second row showing the actual data.

```
plot_pred_comp2 <- function(home_team, away_team, ms) {
    par(mfrow = c(2, 4))
    home_baseline <- ms[, "home_baseline"]
    away_baseline <- ms[, "away_baseline"]
    home_skill <- ms[, col_name("skill", which(teams == home_team))]
    away_skill <- ms[, col_name("skill", which(teams == away_team))]
    home_goals <- rpois(nrow(ms), exp(home_baseline + home_skill - away_skill))
    away_goals <- rpois(nrow(ms), exp(away_baseline + away_skill - home_skill))
    plot_goals(home_goals, away_goals)
    home_goals <- d$HomeGoals[d$HomeTeam == home_team & d$AwayTeam == away_team]
    away_goals <- d$AwayGoals[d$HomeTeam == home_team & d$AwayTeam == away_team]
    plot_goals(home_goals, away_goals)
}

plot_pred_comp2("FC Valencia", "FC Sevilla", ms2)
```



And similarly Sevilla (home team) vs Valencia (away team).

```
plot_pred_comp2("FC Sevilla", "FC Valencia", ms2)
```



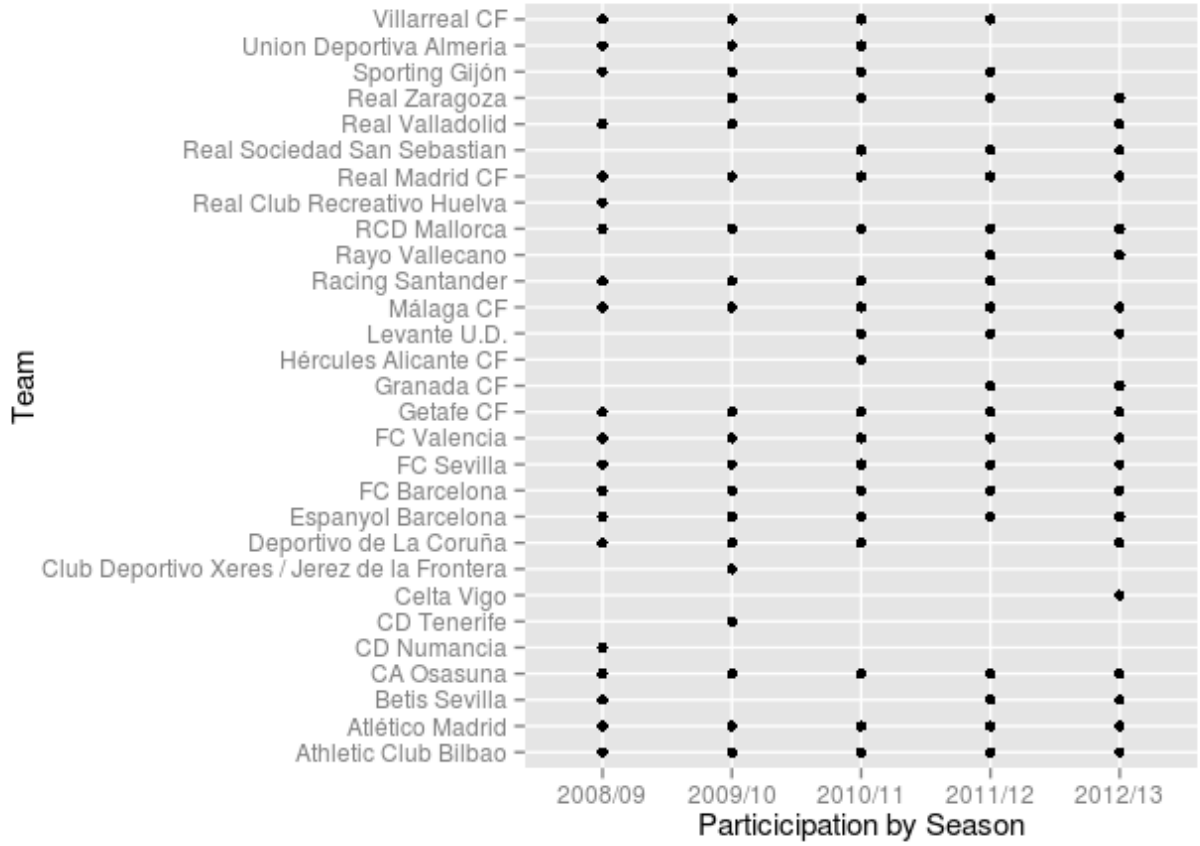
Now the results are closer to the historical data as both Sevilla and Valencia are more likely to win when playing as the home team.

At this point in the modeling process I decided to try to split the skill parameter into two components, offence skill and defense skill, thinking that some teams might be good at scoring goals but at the same time be bad at keeping the opponent from scoring. This didn't seem to result in any better fit however, perhaps because the offensive and defensive skill of a team tend to be highly related. There is however one more thing I would like to change with the model...

Modeling Match Results: Iteration 3

The data set `laliga` contains data from five different seasons and an assumption of the current model is that a team has the same skill during all seasons. This is probably not a realistic assumption, teams probably differ in their year-to-year performance. And what more, some teams do not even participate in all seasons in the `laliga` data set, as a result of dropping out of the first division, as the following diagram shows:

```
qplot(Season, HomeTeam, data = d, ylab = "Team", xlab = "Particicipation by Season")
```



The second iteration of the model was therefore modified to include the year-to-year variability in team skill. This was done by allowing each team to have one skill parameter per season but to connect the skill parameters by using a team's skill parameter for season t in the prior distribution for that team's skill parameter for season $t+1$ so that

$$\text{skill}_{t+1} \sim \text{Normal}(\text{skill}_t, \sigma^2_{\text{season}})$$

for all different t, except the first season which is given an vague prior. Here σ^2_{season} is a parameter estimated using the whole data set. The home and away baselines are given the same kind of priors and below is the resulting JAGS model.

```
m3_string <- "model {
for(i in 1:n_games) {
  HomeGoals[i] ~ dpois(lambda_home[Season[i], HomeTeam[i],AwayTeam[i]])
  AwayGoals[i] ~ dpois(lambda_away[Season[i], HomeTeam[i],AwayTeam[i]])
}

for(season_i in 1:n_seasons) {
  for(home_i in 1:n_teams) {
    for(away_i in 1:n_teams) {
      lambda_home[season_i, home_i, away_i] <- exp( home_baseline[season_i] +
        skill[season_i, home_i] - skill[season_i, away_i])
      lambda_away[season_i, home_i, away_i] <- exp( away_baseline[season_i] +
        skill[season_i, away_i] - skill[season_i, home_i])
    }
```

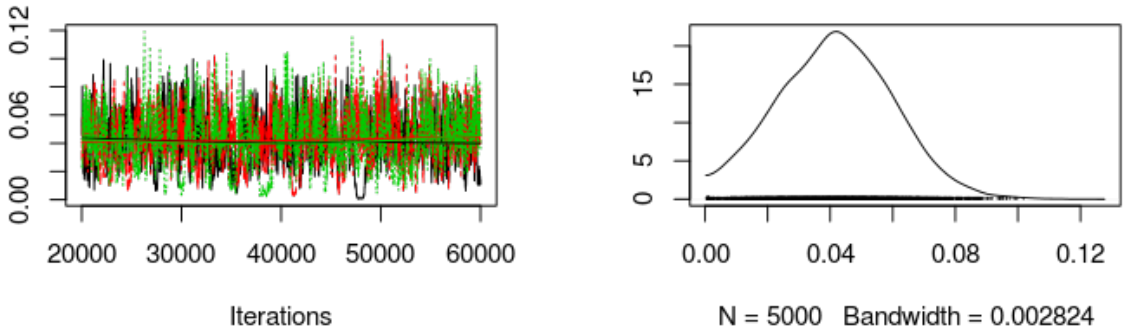
```
  }
}

skill[1, 1] <- 0
for(j in 2:n_teams) {
  skill[1, j] ~ dnorm(group_skill, group_tau)
}

group_skill ~ dnorm(0, 0.0625)
group_tau <- 1/pow(group_sigma, 2)
group_sigma ~ dunif(0, 3)

home_baseline[1] ~ dnorm(0, 0.0625)
away_baseline[1] ~ dnorm(0, 0.0625)

for(season_i in 2:n_seasons) {
  skill[season_i, 1] <- 0
  for(j in 2:n_teams) {
    skill[season_i, j] ~ dnorm(skill[season_i - 1, j], season_tau)
  }
  home_baseline[season_i] ~ dnorm(home_baseline[season_i - 1], season_tau)
  away_baseline[season_i] ~ dnorm(away_baseline[season_i - 1], season_tau)
}

season_tau <- 1/pow(season_sigma, 2)
season_sigma ~ dunif(0, 3)
}"
```

These changes to the model unfortunately introduces quit a lot of autocorrelation when running the MCMC sampler and thus I increase the number of samples and the amount of thinning. On my setup the following run takes the better half of a lunch break (and it wouldn't hurt to run it even a bit longer).

```
m3 <- jags.model(textConnection(m3_string), data = data_list, n.chains = 3,
    n.adapt = 10000)
update(m3, 10000)
s3 <- coda.samples(m3, variable.names = c("home_baseline", "away_baseline",
    "skill", "season_sigma", "group_sigma", "group_skill"), n.iter = 40000,
    thin = 8)
ms3 <- as.matrix(s3)
```

The following graph shows the trace plot and distribution of the `season_sigma` parameter.

```
plot(s3[, "season_sigma"])
```



Calculating and comparing the DIC of this model with the former model show no substantial difference.

```
dic_m3 <- dic.samples(m3, 40000, "pD")
diffdic(dic_m2, dic_m3)
```

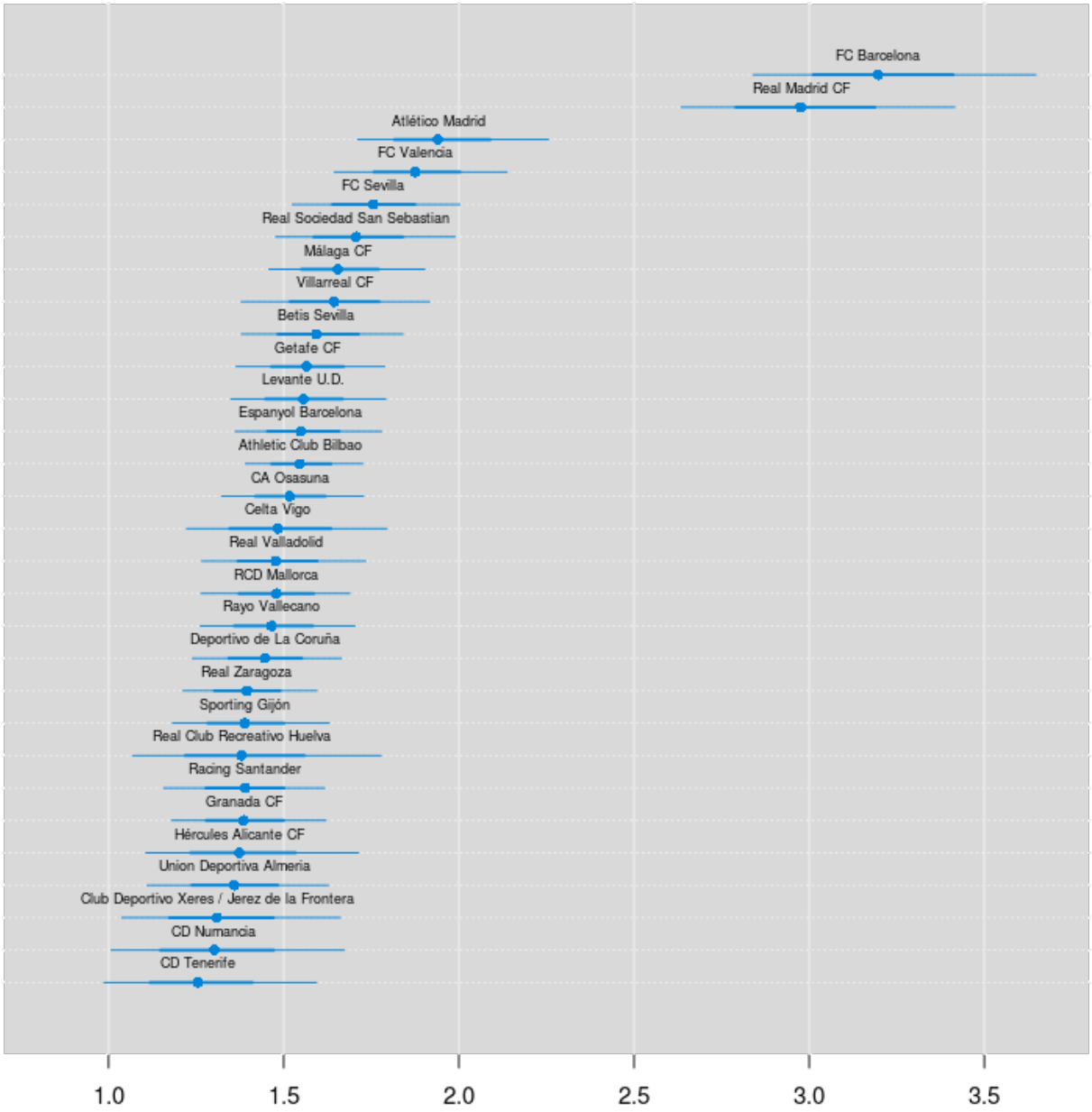
```
## Difference: 13.73
## Sample standard error: 9.575
```

However, I believe the assumptions of the current model (m3) are more reasonable so I'll stick with this model. Now it is time to complete the goals set out in the introduction. That is, ranking the teams in La Liga and predicting the outcome of future matches.

Ranking the Teams of La Liga

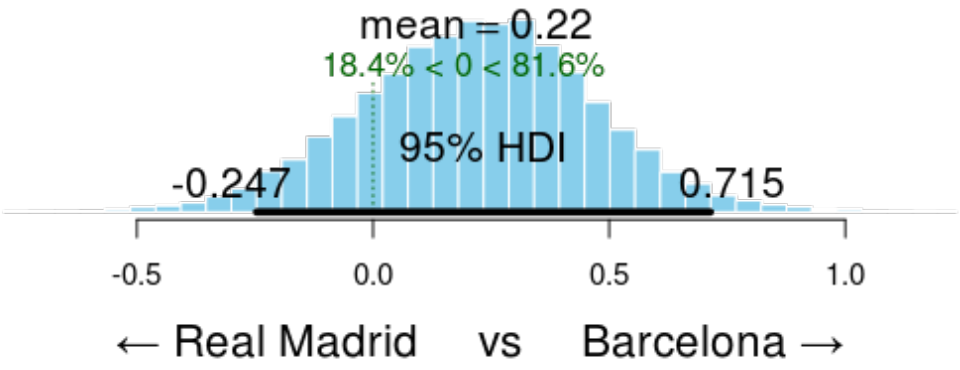
We'll start by ranking the teams of La Liga using the estimated skill parameters from the 2012/2013 season. The values of the skill parameters are difficult to interpret as they are relative to the skill of the team that had its skill parameter "anchored" at zero. To put them on a more interpretable scale I'll first zero center the skill parameters by subtracting the mean skill of all teams, I then add the home baseline and exponentiate the resulting values. These rescaled skill parameters are now on the scale of expected number of goals when playing home team. Below is a caterpillar plot of the median of the rescaled skill parameters together with the 68 % and 95 % credible intervals. The plot is ordered according to the median skill and thus also gives the ranking of the teams.

```
# The ranking of the teams for the 2012/13 season.
team_skill <- ms3[, str_detect(string = colnames(ms3), "skill\\[5,")]
team_skill <- (team_skill - rowMeans(team_skill)) + ms3[, "home_baseline[5]"]
team_skill <- exp(team_skill)
colnames(team_skill) <- teams
team_skill <- team_skill[, order(colMeans(team_skill), decreasing = T)]
par(mar = c(2, 0.7, 0.7, 0.7), xaxs = "i")
caterplot(team_skill, labels.loc = "above", val.lim = c(0.7, 3.8))
```



Two teams are clearly ahead of the rest, FC Barcelona and Real Madrid CF. Let's look at the credible difference between the two teams.

```
plotPost(team_skill[, "FC Barcelona"] - team_skill[, "Real Madrid CF"], compVal = 0,
    xlab = "← Real Madrid     vs     Barcelona →")
```



FC Barcelona is the better team with a probability of 82 % . Go Barcelona!

Predicting the End Game of La Liga 2012/2013

In the `laliga` data set the results of the 50 last games of the 2012/2013 season was missing. Using our model we can now both predict and simulate the outcomes of these 50 games. The R code below calculates a number of measures for each game (both the games with known and unknown outcomes):

- The mode of the simulated number of goals, that is, the *most likely* number of scored goals. If we were asked to bet on the number of goals in a game this is what we would use.
- The mean of the simulated number of goals, this is our best guess of the average number of goals in a game.
- The most likely match result for each game.
- A random sample from the distributions of credible home scores, away scores and match results. This is how La Liga actually could have played out in an alternative reality...

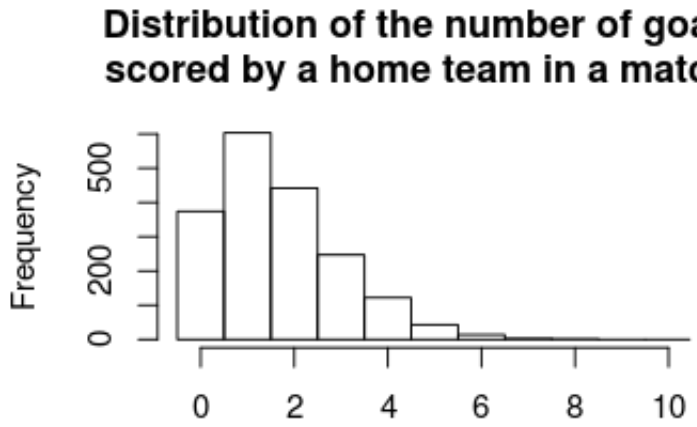
```
n <- nrow(ms3)
m3_pred <- sapply(1:nrow(laliga), function(i) {
  home_team <- which(teams == laliga$HomeTeam[i])
  away_team <- which(teams == laliga$AwayTeam[i])
  season <- which(seasons == laliga$Season[i])
  home_skill <- ms3[, col_name("skill", season, home_team)]
  away_skill <- ms3[, col_name("skill", season, away_team)]
  home_baseline <- ms3[, col_name("home_baseline", season)]
  away_baseline <- ms3[, col_name("away_baseline", season)]

  home_goals <- rpois(n, exp(home_baseline + home_skill - away_skill))
  away_goals <- rpois(n, exp(away_baseline + away_skill - home_skill))
  home_goals_table <- table(home_goals)
  away_goals_table <- table(away_goals)
  match_results <- sign(home_goals - away_goals)
  match_results_table <- table(match_results)

  mode_home_goal <- as.numeric(names(home_goals_table)[ which.max(home_goals_table)])
  mode_away_goal <- as.numeric(names(away_goals_table)[ which.max(away_goals_table)])
```

```
  match_result <-  as.numeric(names(match_results_table)[which.max(match_results_table)])
  rand_i <- sample(seq_along(home_goals), 1)

  c(mode_home_goal = mode_home_goal, mode_away_goal = mode_away_goal, match_result = match_result,
    mean_home_goal = mean(home_goals), mean_away_goal = mean(away_goals),
    rand_home_goal = home_goals[rand_i], rand_away_goal = away_goals[rand_i],
    rand_match_result = match_results[rand_i])
})
m3_pred <- t(m3_pred)
```

First let's compare the distribution of the number of goals in the data with the predicted mode, mean and randomized number of goals for all the games (focusing on the number of goals for the home team).

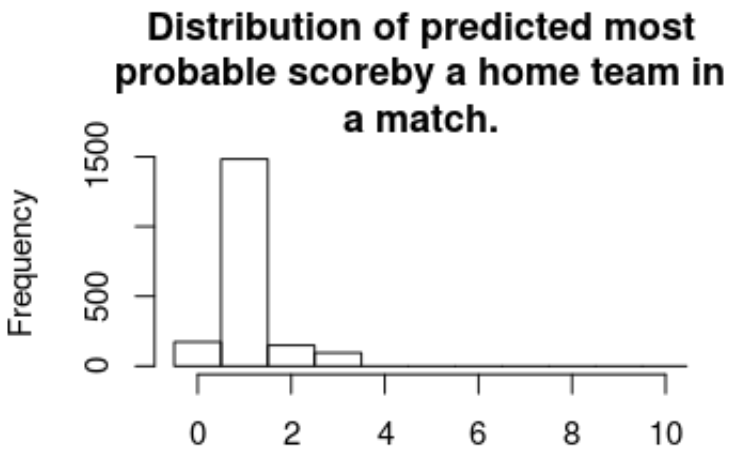
First the actual distribution of the number of goals for the home teams.

```
hist(laliga$HomeGoals, breaks = (-1:10) + 0.5, xlim = c(-0.5, 10),
     main = "Distribution of the number of goals\nscored by a home team in a match.",
    xlab = "")
```



This next plot shows the distribution of the modes from the predicted distribution of home goals from each game. That is, what is the most probable outcome, for the home team, in each game.

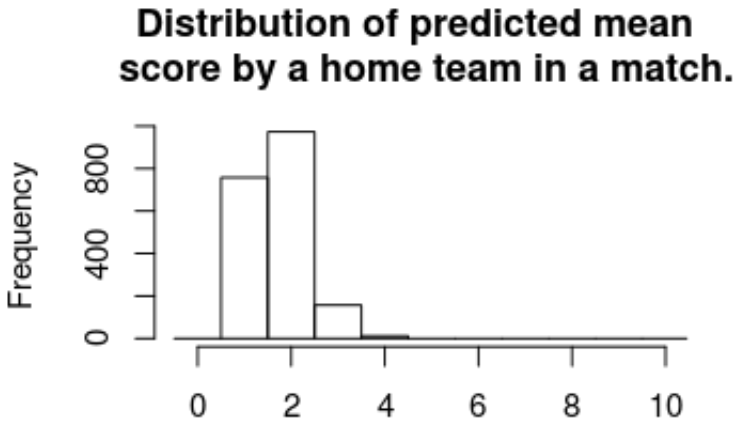
```
hist(m3_pred[, "mode_home_goal"], breaks = (-1:10) + 0.5, xlim = c(-0.5, 10),
    main = "Distribution of predicted most\nprobable scoreby a home team in\na match.",
    xlab = "")
```



For almost all games the single most likely number of goals is one. Actually, if you know nothing about a La Liga game betting on one goal for the home team is 78 % of the times the best bet.

Lest instead look at the distribution of the predicted mean number of home goals in each game.

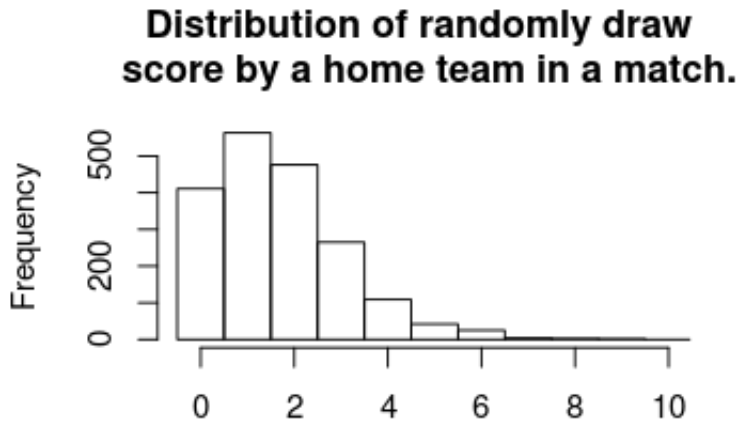
```
hist(m3_pred[, "mean_home_goal"], breaks = (-1:10) + 0.5, xlim = c(-0.5, 10),
    main = "Distribution of predicted mean \n score by a home team in a match.",
    xlab = "")
```



For most games the expected number of goals are 2. That is, even if your safest bet is one goal you would expect to see around two goals.

The distribution of the mode and the mean number of goals doesn't look remotely like the actual number of goals. This was not to be expected, we would however expect the distribution of randomized goals (where for each match the number of goals has been randomly drawn from that match's predicted home goal distribution) to look similar to the actual number of home goals. Looking at the histogram below, this seems to be the case.

```
hist(m3_pred[, "rand_home_goal"], breaks = (-1:10) + 0.5, xlim = c(-0.5, 10),
    main = "Distribution of randomly draw \n score by a home team in a match.",
    xlab = "")
```



We can also look at how well the model predicts the data. This should probably be done using cross validation, but as the number of effective parameters are much smaller than the number of data points a direct comparison should at least give an estimated prediction accuracy in the right ballpark.

```
mean(laliga$HomeGoals == m3_pred[, "mode_home_goal"], na.rm = T)
```

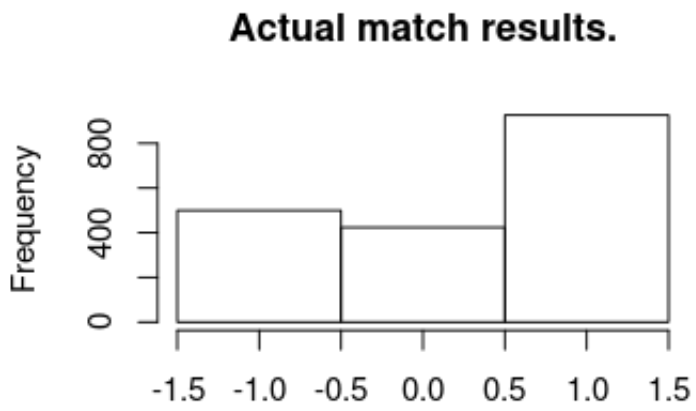
```
## [1] 0.3351
```

```
mean((laliga$HomeGoals - m3_pred[, "mean_home_goal"])^2, na.rm = T)
```

```
## [1] 1.452
```

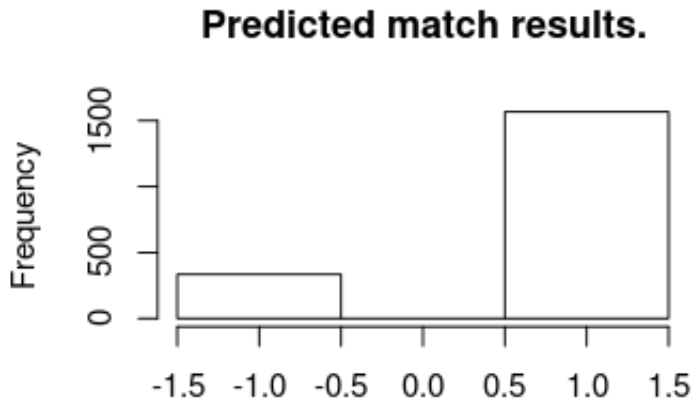
So on average the model predicts the correct number of home goals 34 % of the time and guesses the average number of goals with a mean squared error of 1.45 . Now we'll look at the actual and predicted match outcomes. The graph below shows the match outcomes in the data with 1 being a home win, 0 being a draw and -1 being a win for the away team.

```
hist(laliga$MatchResult, breaks = (-2:1) + 0.5, main = "Actual match results.",
    xlab = "")
```



Now looking at the most probable outcomes of the matches according to the model.

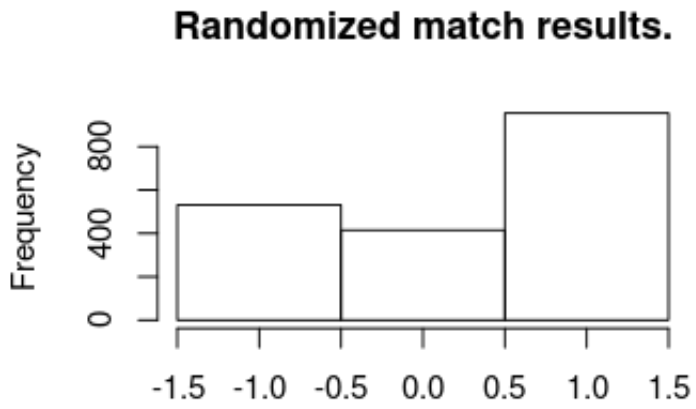
```
hist(m3_pred[, "match_result"], breaks = (-2:1) + 0.5, main = "Predicted match results.",
    xlab = "")
```



For almost all matches the safest bet is to bet on the home team. While draws are not uncommon it is *never* the safest bet.

As in the case with the number of home goals, the randomized match outcomes have a distribution similar to the actual match outcomes:

```
hist(m3_pred[, "rand_match_result"], breaks = (-2:1) + 0.5, main = "Randomized match results.",
    xlab = "")
```




```
mean(laliga$MatchResult == m3_pred[, "match_result"], na.rm = T)
```

```
## [1] 0.5627
```

The model predicts the correct match outcome 56 % of the time. Pretty good!

Now that we've checked that the model reasonably predicts the La Liga history let's predict the La Liga endgame! The code below displays the predicted mode and mean number of goals for the endgame and the predicted winner of each game.

```
laliga_forecast <- laliga[is.na(laliga$HomeGoals), c("Season", "Week", "HomeTeam",
    "AwayTeam")]
m3_forecast <- m3_pred[is.na(laliga$HomeGoals), ]
laliga_forecast$mean_home_goals <- round(m3_forecast[, "mean_home_goal"], 1)
laliga_forecast$mean_away_goals <- round(m3_forecast[, "mean_away_goal"], 1)
laliga_forecast$mode_home_goals <- m3_forecast[, "mode_home_goal"]
laliga_forecast$mode_away_goals <- m3_forecast[, "mode_away_goal"]
laliga_forecast$predicted_winner <- ifelse(m3_forecast[, "match_result"] ==
    1, laliga_forecast$HomeTeam, ifelse(m3_forecast[, "match_result"] == -1,
    laliga_forecast$AwayTeam, "Draw"))

rownames(laliga_forecast) <- NULL
print(xtable(laliga_forecast, align = "ccccccccccc"), type = "html")
```

	Season	Week	HomeTeam	AwayTeam	mean_home_goals	mean_away_goals	mode_home_goals	mode_away_goals	predicted_winner
1	2012/13	34	Celta Vigo	Athletic Club Bilbao	1.50	1.10	1.00	1.00	Celta Vigo
2	2012/13	34	Deportivo de La Coruña	Atlético Madrid	1.20	1.50	1.00	1.00	Atlético Madrid
3	2012/13	34	FC Barcelona	Betis Sevilla	3.20	0.50	3.00	0.00	FC Barcelona
4	2012/13	34	FC Sevilla	Espanyol Barcelona	1.80	1.00	1.00	0.00	FC Sevilla
5	2012/13	34	FC Valencia	CA Osasuna	2.00	0.90	1.00	0.00	FC Valencia
6	2012/13	34	Getafe CF	Real Sociedad San Sebastian	1.40	1.20	1.00	1.00	Getafe CF
7	2012/13	34	Granada CF	Málaga CF	1.30	1.30	1.00	1.00	Granada CF
8	2012/13	34	RCD Mallorca	Levante U.D.	1.50	1.10	1.00	1.00	RCD Mallorca
9	2012/13	34	Real Madrid CF	Real Valladolid	3.20	0.50	3.00	0.00	Real Madrid CF
10	2012/13	34	Real Zaragoza	Rayo Vallecano	1.50	1.10	1.00	1.00	Real Zaragoza
11	2012/13	35	Athletic Club Bilbao	RCD Mallorca	1.60	1.00	1.00	1.00	Athletic Club Bilbao
12	2012/13	35	Atlético Madrid	FC Barcelona	1.00	1.80	0.00	1.00	FC Barcelona
13	2012/13	35	Betis Sevilla	Celta Vigo	1.70	1.00	1.00	1.00	Betis Sevilla
14	2012/13	35	CA Osasuna	Getafe CF	1.50	1.10	1.00	1.00	CA Osasuna
15	2012/13	35	Espanyol Barcelona	Real Madrid CF	0.80	2.10	0.00	1.00	Real Madrid CF
16	2012/13	35	Levante U.D.	Real Zaragoza	1.80	1.00	1.00	0.00	Levante U.D.
17	2012/13	35	Málaga CF	FC Sevilla	1.50	1.10	1.00	1.00	Málaga CF
18	2012/13	35	Rayo Vallecano	FC Valencia	1.20	1.40	1.00	1.00	FC Valencia
19	2012/13	35	Real Sociedad San Sebastian	Granada CF	2.00	0.90	1.00	0.00	Real Sociedad San Sebastian
20	2012/13	35	Real Valladolid	Deportivo de La Coruña	1.60	1.10	1.00	1.00	Real Valladolid
21	2012/13	36	Celta Vigo	Atlético Madrid	1.20	1.40	1.00	1.00	Atlético Madrid
22	2012/13	36	Deportivo de La Coruña	Espanyol Barcelona	1.50	1.20	1.00	1.00	Deportivo de La Coruña
23	2012/13	36	FC Barcelona	Real Valladolid	3.40	0.50	3.00	0.00	FC Barcelona
24	2012/13	36	FC Sevilla	Real Sociedad San Sebastian	1.60	1.10	1.00	1.00	FC Sevilla
25	2012/13	36	Getafe CF	FC Valencia	1.30	1.30	1.00	1.00	Getafe CF
26	2012/13	36	Granada CF	CA Osasuna	1.50	1.20	1.00	1.00	Granada CF
27	2012/13	36	Levante U.D.	Rayo Vallecano	1.70	1.00	1.00	1.00	Levante U.D.
28	2012/13	36	RCD Mallorca	Betis Sevilla	1.50	1.20	1.00	1.00	RCD Mallorca
29	2012/13	36	Real Madrid CF	Málaga CF	2.90	0.60	2.00	0.00	Real Madrid CF
30	2012/13	36	Real Zaragoza	Athletic Club Bilbao	1.40	1.20	1.00	1.00	Real Zaragoza
31	2012/13	37	Athletic Club Bilbao	Levante U.D.	1.60	1.10	1.00	1.00	Athletic Club Bilbao
32	2012/13	37	Atlético Madrid	RCD Mallorca	2.10	0.80	2.00	0.00	Atlético Madrid
33	2012/13	37	Betis Sevilla	Real Zaragoza	1.80	1.00	1.00	0.00	Betis Sevilla
34	2012/13	37	CA Osasuna	FC Sevilla	1.40	1.30	1.00	1.00	CA Osasuna
35	2012/13	37	Espanyol Barcelona	FC Barcelona	0.80	2.30	0.00	2.00	FC Barcelona
36	2012/13	37	FC Valencia	Granada CF	2.10	0.80	2.00	0.00	FC Valencia
37	2012/13	37	Getafe CF	Rayo Vallecano	1.70	1.00	1.00	1.00	Getafe CF
38	2012/13	37	Málaga CF	Deportivo de La Coruña	1.80	1.00	1.00	0.00	Málaga CF
39	2012/13	37	Real Sociedad San Sebastian	Real Madrid CF	0.90	1.90	0.00	1.00	Real Madrid CF
40	2012/13	37	Real Valladolid	Celta Vigo	1.60	1.10	1.00	1.00	Real Valladolid
41	2012/13	38	Celta Vigo	Espanyol Barcelona	1.50	1.10	1.00	1.00	Celta Vigo
42	2012/13	38	Deportivo de La Coruña	Real Sociedad San Sebastian	1.30	1.30	1.00	1.00	Deportivo de La Coruña
43	2012/13	38	FC Barcelona	Málaga CF	3.00	0.60	3.00	0.00	FC Barcelona
44	2012/13	38	FC Sevilla	FC Valencia	1.50	1.20	1.00	1.00	FC Sevilla
45	2012/13	38	Granada CF	Getafe CF	1.40	1.20	1.00	1.00	Granada CF
46	2012/13	38	Levante U.D.	Betis Sevilla	1.60	1.10	1.00	1.00	Levante U.D.
47	2012/13	38	RCD Mallorca	Real Valladolid	1.60	1.10	1.00	1.00	RCD Mallorca
48	2012/13	38	Rayo Vallecano	Athletic Club Bilbao	1.50	1.10	1.00	1.00	Rayo Vallecano
49	2012/13	38	Real Madrid CF	CA Osasuna	3.10	0.60	3.00	0.00	Real Madrid CF
50	2012/13	38	Real Zaragoza	Atlético Madrid	1.10	1.50	1.00	1.00	Atlético Madrid

While these predictions are good if you want to bet on the likely winner they do not reflect how the actual endgame will play out, e.g., there is not a single draw in the `predicted_winner` column. So at last let's look at a *possible* version of the La Liga endgame by displaying the simulated match results calculated earlier.

```
laliga_sim <- laliga[is.na(laliga$HomeGoals), c("Season", "Week", "HomeTeam",
    "AwayTeam")]
laliga_sim$home_goals <- m3_forecast[, "rand_home_goal"]
laliga_sim$away_goals <- m3_forecast[, "rand_away_goal"]
laliga_sim$winner <- ifelse(m3_forecast[, "rand_match_result"] == 1, laliga_forecast$HomeTeam,
    ifelse(m3_forecast[, "rand_match_result"] == -1, laliga_forecast$AwayTeam,
        "Draw"))

rownames(laliga_sim) <- NULL
print(xtable(laliga_sim, align = "cccccccc"), type = "html")
```

	Season	Week	HomeTeam	AwayTeam	home_goals	away_goals	winner
1	2012/13	34	Celta Vigo	Athletic Club Bilbao	3.00	0.00	Celta Vigo
2	2012/13	34	Deportivo de La Coruña	Atlético Madrid	0.00	2.00	Atlético Madrid
3	2012/13	34	FC Barcelona	Betis Sevilla	3.00	0.00	FC Barcelona
4	2012/13	34	FC Sevilla	Espanyol Barcelona	2.00	3.00	Espanyol Barcelona
5	2012/13	34	FC Valencia	CA Osasuna	1.00	0.00	FC Valencia
6	2012/13	34	Getafe CF	Real Sociedad San Sebastian	4.00	0.00	Getafe CF
7	2012/13	34	Granada CF	Málaga CF	1.00	1.00	Draw
8	2012/13	34	RCD Mallorca	Levante U.D.	1.00	1.00	Draw
9	2012/13	34	Real Madrid CF	Real Valladolid	3.00	0.00	Real Madrid CF
10	2012/13	34	Real Zaragoza	Rayo Vallecano	2.00	2.00	Draw
11	2012/13	35	Athletic Club Bilbao	RCD Mallorca	1.00	1.00	Draw
12	2012/13	35	Atlético Madrid	FC Barcelona	0.00	2.00	FC Barcelona
13	2012/13	35	Betis Sevilla	Celta Vigo	2.00	1.00	Betis Sevilla
14	2012/13	35	CA Osasuna	Getafe CF	1.00	3.00	Getafe CF
15	2012/13	35	Espanyol Barcelona	Real Madrid CF	1.00	0.00	Espanyol Barcelona
16	2012/13	35	Levante U.D.	Real Zaragoza	5.00	1.00	Levante U.D.
17	2012/13	35	Málaga CF	FC Sevilla	1.00	2.00	FC Sevilla
18	2012/13	35	Rayo Vallecano	FC Valencia	1.00	2.00	FC Valencia
19	2012/13	35	Real Sociedad San Sebastian	Granada CF	0.00	1.00	Granada CF
20	2012/13	35	Real Valladolid	Deportivo de La Coruña	7.00	1.00	Real Valladolid
21	2012/13	36	Celta Vigo	Atlético Madrid	6.00	0.00	Celta Vigo
22	2012/13	36	Deportivo de La Coruña	Espanyol Barcelona	3.00	1.00	Deportivo de La Coruña
23	2012/13	36	FC Barcelona	Real Valladolid	3.00	0.00	FC Barcelona
24	2012/13	36	FC Sevilla	Real Sociedad San Sebastian	2.00	0.00	FC Sevilla
25	2012/13	36	Getafe CF	FC Valencia	1.00	1.00	Draw
26	2012/13	36	Granada CF	CA Osasuna	0.00	3.00	CA Osasuna
27	2012/13	36	Levante U.D.	Rayo Vallecano	1.00	0.00	Levante U.D.
28	2012/13	36	RCD Mallorca	Betis Sevilla	1.00	2.00	Betis Sevilla
29	2012/13	36	Real Madrid CF	Málaga CF	0.00	1.00	Málaga CF
30	2012/13	36	Real Zaragoza	Athletic Club Bilbao	0.00	1.00	Athletic Club Bilbao
31	2012/13	37	Athletic Club Bilbao	Levante U.D.	0.00	0.00	Draw
32	2012/13	37	Atlético Madrid	RCD Mallorca	4.00	0.00	Atlético Madrid
33	2012/13	37	Betis Sevilla	Real Zaragoza	2.00	0.00	Betis Sevilla
34	2012/13	37	CA Osasuna	FC Sevilla	1.00	3.00	FC Sevilla
35	2012/13	37	Espanyol Barcelona	FC Barcelona	0.00	1.00	FC Barcelona
36	2012/13	37	FC Valencia	Granada CF	3.00	0.00	FC Valencia
37	2012/13	37	Getafe CF	Rayo Vallecano	0.00	0.00	Draw
38	2012/13	37	Málaga CF	Deportivo de La Coruña	3.00	1.00	Málaga CF
39	2012/13	37	Real Sociedad San Sebastian	Real Madrid CF	1.00	1.00	Draw
40	2012/13	37	Real Valladolid	Celta Vigo	4.00	1.00	Real Valladolid
41	2012/13	38	Celta Vigo	Espanyol Barcelona	1.00	1.00	Draw
42	2012/13	38	Deportivo de La Coruña	Real Sociedad San Sebastian	1.00	0.00	Deportivo de La Coruña
43	2012/13	38	FC Barcelona	Málaga CF	4.00	1.00	FC Barcelona
44	2012/13	38	FC Sevilla	FC Valencia	1.00	0.00	FC Sevilla
45	2012/13	38	Granada CF	Getafe CF	0.00	2.00	Getafe CF
46	2012/13	38	Levante U.D.	Betis Sevilla	1.00	1.00	Draw
47	2012/13	38	RCD Mallorca	Real Valladolid	1.00	2.00	Real Valladolid
48	2012/13	38	Rayo Vallecano	Athletic Club Bilbao	1.00	2.00	Athletic Club Bilbao
49	2012/13	38	Real Madrid CF	CA Osasuna	1.00	0.00	Real Madrid CF
50	2012/13	38	Real Zaragoza	Atlético Madrid	2.00	1.00	Real Zaragoza

Now we see a number of games resulting in a draw. We also see that Málaga manages to beat Real Madrid in week 36, against all odds, even though playing as the away team. An amazing day for all Málaga fans!

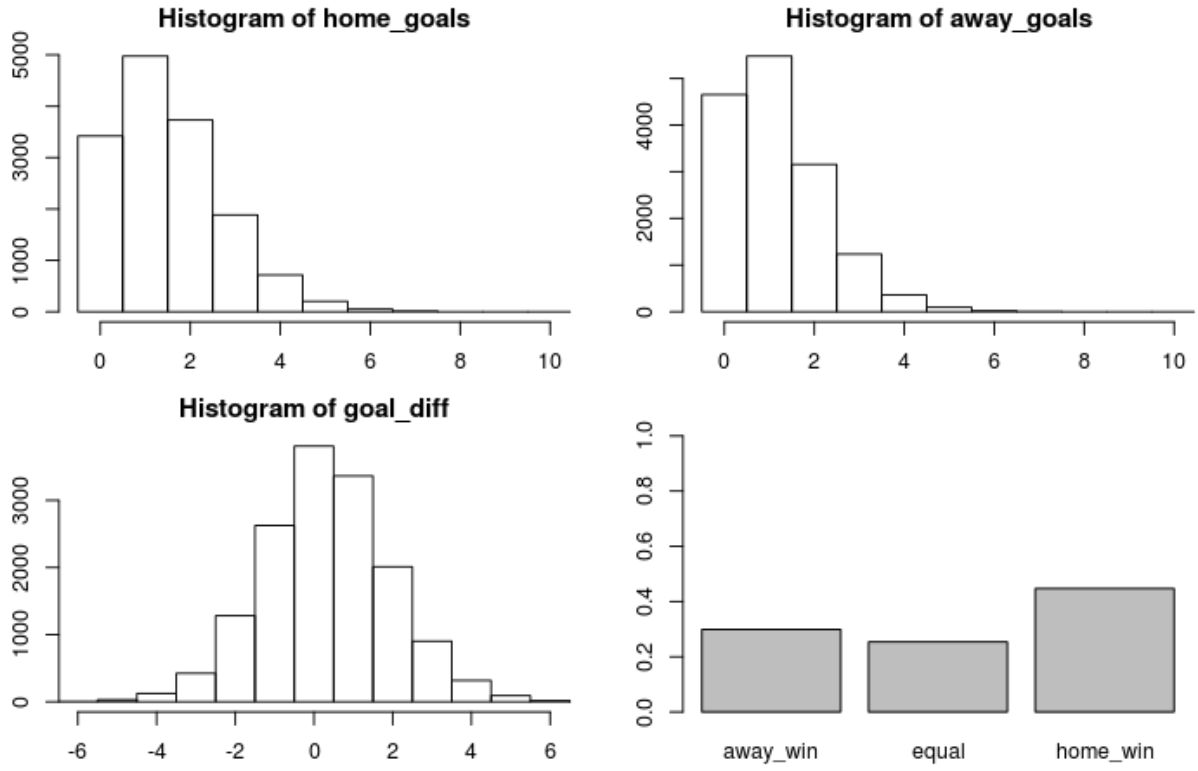
Calculating the Predicted Payout for Sevilla vs Valencia, 2013-06-01

At the time when I developed this model (2013-05-28) most of the matches in the 2012/2013 season had been played and Barcelona was already the winner (and the most skilled team as predicted by my model). There were however some matches left, for example, Sevilla (home team) vs Valencia (away team) at the 1st of June, 2013. One of the powers with using Bayesian modeling and MCMC sampling is that once you have the MCMC samples of the parameters it is straight forward to calculate any quantity resulting from these estimates while still retaining the uncertainty of the parameter estimates. So let's look at the predicted distribution of the number of goals for the Sevilla vs Valencia game and see if I can use my model to make some money. I'll start by using the MCMC samples to calculate the distribution of the number of goals for Sevilla and Valencia.

```
n <- nrow(ms3)
home_team <- which(teams == "FC Sevilla")
away_team <- which(teams == "FC Valencia")
season <- which(seasons == "2012/13")
home_skill <- ms3[, col_name("skill", season, home_team)]
away_skill <- ms3[, col_name("skill", season, away_team)]
home_baseline <- ms3[, col_name("home_baseline", season)]
away_baseline <- ms3[, col_name("away_baseline", season)]

home_goals <- rpois(n, exp(home_baseline + home_skill - away_skill))
away_goals <- rpois(n, exp(away_baseline + away_skill - home_skill))
```

Looking at summary of these two distributions shows that it will be a close game but with a slight advantage for the home team Sevilla.

```
par(mfrow = c(2, 2), mar = rep(2.2, 4))
plot_goals(home_goals, away_goals)
```



When developing the model (2013-05-28) I got the following payouts (that is, how much would I get back if my bet was successful) for betting on the outcome of this game on the betting site www.betsson.com:

```
 Sevilla  Draw  Valencia
   3.2    3.35    2.15
```

Using my simulated distribution of the number of goals I can calculate the predicted payouts of my model.

```
1/c(Sevilla = mean(home_goals > away_goals), Draw = mean(home_goals == away_goals),
    Valencia = mean(home_goals < away_goals))
```

```
##  Sevilla     Draw Valencia
##    2.237    3.940    3.343
```

I should clearly bet on Sevilla as my model predicts a payout of 2.24 (that is, a likely win for Sevilla) while betsson.com gives me the much higher payout of 3.2. It is also possible to bet on the final goal outcome so let's calculate what payouts my model predicts for different goal outcomes.

```
goals_payout <- laply(0:6, function(home_goal) {
    laply(0:6, function(away_goal) {
        1/mean(home_goals == home_goal & away_goals == away_goal)
    })
})

colnames(goals_payout) <- paste("Valencia", 0:6, sep = " - ")
rownames(goals_payout) <- paste("Sevilla", 0:6, sep = " - ")
goals_payout <- round(goals_payout, 1)
print(xtable(goals_payout, align = "cccccccc"), type = "html")
```

	Valencia - 0	Valencia - 1	Valencia - 2	Valencia - 3	Valencia - 4	Valencia - 5	Valencia - 6
Sevilla - 0	14.20	12.10	20.00	53.60	197.40	789.50	7500.00
Sevilla - 1	9.80	8.40	14.30	35.10	122.00	365.90	1666.70
Sevilla - 2	13.10	10.70	19.20	52.80	166.70	714.30	2142.90
Sevilla - 3	24.50	21.60	41.40	96.20	312.50	1153.80	Inf
Sevilla - 4	65.20	61.50	97.40	217.40	937.50	5000.00	Inf
Sevilla - 5	227.30	202.70	340.90	1071.40	3750.00	Inf	15000.00
Sevilla - 6	1500.00	600.00	1363.60	5000.00	7500.00	Inf	Inf

The most likely result is 1 - 1 with a predicted payout of 8.4 and betsson.com agrees with this also offering their lowest payout for this bet, 5.3. Not good enough! Looking at the payouts at bettson.com I can see that Sevilla - Valencia: 2 - 0 gives me a payout of 16.0, that's much better than my predicted payout of 13.1. I'll go for that!

Wrap-up

I believe my model has a lot things going for it. It is conceptually quite simple and easy to understand, implement and extend. It captures the patterns in and distribution of the data well. It allows me to *easily* calculate the probability of any outcome, from a game with whichever teams from any La Liga season. Want to calculate the probability that RCD Mallorca (home team) vs Málaga CF (away team) in the Season

2009/2010 would result in a draw? Easy! What's the probability of the total number of goals in Granada CF vs Athletic Club Bilbao being a prime number? No problemo! What if Real Madrid from 2008/2009 met Barcelona from 2012/2013 in 2010/2011 and *both* teams had the home advantage? Well, that's possible...

There are also a couple of things that could be improved (many which are not too hard to address). * Currently there is assumed to be *no* dependency between the goal distributions of the home and away teams, but this might not be realistic. Maybe if one team have scored more goals the other team "looses steam" (a negative correlation between the teams' scores) or instead maybe the other team tries harder (a positive correlation). Such dependencies could maybe be added to the model using copulas. * One of the *advantages* of Bayesian statistics is the possibility to used informative priors. As I have no knowledge of football I've been using vague priors but with the help of a more knowledgeable football fan the model could be given more informative priors. * The predictive performance of the model has not been as thoroughly examined and this could be remedied with a healthy dose of cross validation. * The graphs could be much prettier, but this submission was for the data *analysis* contest of UseR 2014 not the data *visualtization* contest, so...

So this has been a journey, like a pirate on the open sea I've been sailing on a sea of data salvaging and looting whatever I could with the power of JAGS and R (read ARRRHHH!). And still without knowing anything about football I can now log onto bettson.com and with confidence bet 100 Swedish kronas on Sevilla next week winning with 2 - 0 against Valencia. ¡Adelante Sevilla!

Edit: At time of writing the match between Sevilla and Valencia has been played and my bet was a partial success. I betted 50 sek on Sevilla winning the game and 50 sek on Sevilla winning with 2 - 0. Against the (betsson.com) odds Sevilla did win the game, which gave me $50 \cdot 3.2 = 160$ sek, but unfortunately for me not with 2-0 but with 4-3. In total I betted 100 sek and got 160 sek back so good enough for me :)