Insert Real Title

A Thesis
Presented to
Department of Statistical Science
Duke University

Nathaniel Brown

May 2018

Approved for the
Bachelor of Science in Statistical Science

Mike West

Merlise Clyde

Cliburn Chan

Mine Cetinkaya-Rundel, DUS

Acknowledgements

Thank you to Kevin Cullen, Basketball Director of Information Technology for the Duke Men's Basketball team, for providing data for this research.

Preface

This is an example of a thesis setup to use the reed thesis document class (for LaTeX) and the R bookdown package, in general.

Table of Contents

List of Tables

List of Figures

Abstract

Abstract

This study is an investigation of Bayesian statistical models and analyses for problems arising in shooting a basketball. The dataset is from the Duke Men's Basketball team's player-tracking data, which is recorded on the SportVU cameras from STATS, LLC. Goals are to explore, develop, and apply Bayesian models to existing and new data on shooting outcomes. In addition, we want to understand and evaluate questions of inherent random variation, changes over time in shooting performance, and issues related to the "Hot Hand" concept in sports.

The models we use to investigate this data are a Bayesian logistic generalized linear model, a hierarchical model with mixed effects on the shooter identity, and a discounted likelihood model that reduces the influence of shots as their time difference from the current shot increases.

Our results so far show that the best-fitting model is ...

Dedication

Insert a real dedication. (optional)

Introduction

In the sport of basketball, points are awarded by the binary event of shooting the ball into the goal. Some factors we consider that may affect the success rate include the location of the shooter, the individual skill of the shooter, whether the shooter is playing on his home court or on an away court, and the shooting success in recent surrounding games. There have been previous studies on how much recent shooting success affects current shooting success, and the results vary. For example, *insert examples from literature review.*

The purpose of this paper is to investigate Bayesian modelling techniques shooting data, and to learn more about time-dependency in shooting data.

Chapter 1

Abstract

This study is an investigation of Bayesian statistical models and analyses for problems arising in shooting a basketball. The dataset is from the Duke Men's Basketball team's player-tracking data, which is recorded on the SportVU cameras from STATS, LLC. Goals are to explore, develop, and apply Bayesian models to existing and new data on shooting outcomes. In addition, we want to understand and evaluate questions of inherent random variation, changes over time in shooting performance, and issues related to the "Hot Hand" concept in sports.

The models we use to investigate this data are a Bayesian logistic generalized linear model, a hierarchical model with mixed effects on the shooter identity, and a discounted likelihood model that reduces the influence of shots as their time difference from the current shot increases.

Our results so far show that the best-fitting model is …

Chapter 2

Literature Review

2.0.1 how do I get the full citations to show up and not just last name and year?

Gilovich, Vallone, & Tversky (1985)

In this research paper from *Cognitive Psychology*, Thomas Gilovich, Robert Vallone, and Amos Tversky investigate peoples' belief in the Hot Hand in Basketball. The Hot Hand is the concept that the probability of a success increases for trials that follow a success in a binary sequence; in basketball, these binary events are shot attempts. The methods in this paper include an analysis of shot attempts from the Philadelphia 76ers of the National Basketball Association (NBA) in the 1981 season, analysis of free-throw attempts from the Boston Celtics in the 1981 and 1982 seasons, and a controlled shooting drill using male and female varsity basketball players at Cornell University. Statistical techniques they used to attempt to detect streakiness in the data included Walf-Wolfowitz run tests, autocorrelation tests on consecutive shot attempts, goodness-of-fit tests for the distribution of succesess, and paired t-tests comparing the mean of makes following a make to that of makes following a miss. In addition to this analysis of shooting, this research also contained a survey of basketball fans, that gauged how much people believed success probabilities changed given a success or a failure. The statistical tests did not detect significant evidence supporting the Hot Hand in basketball. The lack of statistical power in Gilovich, Vallone, and Tversky's frequentist tests motivates the use of Bayesian models in this thesis. Strengths of this paper include the fact that it was one of the first research papers to analyze streakiness in basketball data, and many future papers build off of it. Some weaknesses in this paper are the assumptions it makes in its analysis, such as all shots being independent of each other, and not accounting for shot location. Albert & Williamson (1999)

In this paper, Jim Albert attempts to improve upon the low-powered tests of Gilovich, Vallone, and Tverky's 1985 paper on the Hot Hand. Albert formally defines "streakiness" as the presence of nonstationarity (nonconstant probability between trials) or autocorrelation (sequential dependency). Albert uses Gibbs sampling to approximate posterior densities and to simulate data, then fits two types of models on binary data from baseball and basketball to try to characterize streakiness. He fits an overdispersion model to detect nonstationarity, and a markov switching model to detect sequential dependencies. While he did not uncover strong evidence for the hot hand, one of his takeaways was that overdispersion decreases as time goes on in basketball free throw shooting data. A weakness of this paper is that Albert does not show the results of both the Markov model and the overdispersion model on the same data. We use Albert's formal definitions of streakiness as well as his motivation for Bayesian models over frequentist tests.

Bar-Eli, Avugos, & Raab (2006)

This paper is a review of previous hot hand research. It reviews several papers investigating the concept of the "hot hand" in several sports such as basketball, baseball, volleyball, and horeshoe, and other fields such as cognitive science and economics. Bar-Eli, Avugos, and Raab evaluate the datasets, the tests and statistics used, and the conclusions of each study. Overall, the authors summarize 13 papers that oppose the hot hand phenomenon, and 11 that support it; they also acknowledge that the scientific evidence for the hot hand is weaker than the evidence against it, and it is typically more controversial. Instead of just looking to answer whether the hot hand exists, Bar-Eli, Avugos, and Raab also examine how people define a "hot hand", and the psychological factors behind the belief in it, such as the gambling and game strategy. The strengths of this paper are that it evaluates the strengths and weaknesses of many competing claims, and concisely summarizes the information into a table. A weakness is that they do not make any claim of their own. This paper is useful in this thesis because it describes several data analysis techniques to detect streaks in a binary sequence.

Ryan Wetzels (2016)

In this research paper, Wetzels conducts a simulation study to investigate the Hot Hand Phenomenon. His analysis consists of calculating Bayes Factors to compare evidence between a Hidden Markov Model with two states and a binomial model with one state. He applies this method to data from basketball foul shots and from visual discernment tests. In the basketball data, he found that Shaquille O'Neal's free-throws show evidence for a two-state Markov model, while Kobe Bryant's show more evidence for a one-state binomial model. In the data from the visual discernment tests, he found no strong evidence supporting one model over the other. A strength of this paper is Wetzel's formal comparison of a Bayesian Markov model to a binomial model. A weakness is that the Bayes Factors only compare evidence between the two models; it does not mean that either model is "good". We use this paper for the specification of the Hidden Markov Model.

Albert (1993)

In this paper, Albert uses a Markov switching model to analyze streakiness in baseball pitching data. He concludes that a few players exhibit streakiness, but not enough to reject the null hypothesis. An exploratory technique that we take from this paper is to examine the peaks and valleys in a moving average plot to observe streakiness. A strength of this paper is that Albert controls for situational variables such as home field advantage, the handedness of the pitcher, and the runners on the bases.

Albert (2013)

In this paper, Albert analyzes streakiness in baseball hitting data. His analysis techniques include using Bayes Factors to compare models of the form $f(y_j|p_j) = p_j(1-p_j)^{y_j}, y_j = 0, 1, 2, ...$; a consistent model with a constant p_j, and a streaky model with a varying p_j from a beta distribution. A useful insight that we apply to this paper is the concept that the existence of streakiness depends on the definition of "success" in binary outcome data. He found substantially more evidence for streakiness for when a success was coded as "not a strikeout" instead of a "hit". Likewise, in this paper we *blank*.

West, Harrison, & Migon (1985)

This textbook provides theory, applications, and examples of time series models such Dynamic Generalized Linear Models (DGLMs). More specifically, section 14.4 provides an example of a DGLM for a binomial response variable, which we apply in chapter *blank* of this research paper.

Chapter 3

Data

3.1 Description of Dataset

The data for this analysis comes from SportVU, a player-tracking system from STATS, LLC. that provides precise coordinates for all ten players and the ball at a rate of 25 times per second. The Duke University Men's Basketball team permitted us to use their SportVU data from the 2014 to 2017 basketball seasons for this project. Since the ability to record this data depends on specialized tracking cameras, Duke does not have this data for every game they play—only home games, and a few road games in arenas that had the techology installed. Therefore, there is a substantial amount of missing data between games. More specifically, between the 2014 and 2017 seasons, the Duke Men's Basketball team played 147 games; this dataset contains 94 games, with 82 at Duke and 12 on another court.

For our analysis, we use the following files for each game:

- Final Sequence Play-by-Play Optical:

This dataset comes in an a semi-structured Extensible Markup Language (XML) file, where there is a unique element for each "event" (an event is a basketball action such as a dribble, pass, shot, foul, etc.). Each event element has attributes describing the type of event, the time of the event, and the player who completed the action. We use these files to uncover when a shot is attempted in a game, who attempted the shot, and the result of the shot attempt.

- Box Score Optical:

We use this dataset to match the names and IDs of players who are in the game. This is also an XML file, with elements corresponding to individual players. These elements contain attributes describing information about the player (e.g. team name, jersey number) and various statistics for the game (e.g. points, assists, distance run).

- Final Sequence Optical:

These XML files contain the locations of all ten players and the ball during precise time intervals within the game. Each timeunit has a unique element, and these elements have attributes describing the locations. We merge this with the Final Sequence Play-by-Play Optical data on the time attribute to obtain the shooter's location at the moment of a shot attempt.

3.2 Data Cleaning

Steps taken to clean the merged shooter IDs with shot locations include standardizing the locations onto a half-court setting (the teams switch sides of the court halfway through every game, which means that we have to flip the coordinates across the middle of the court for half of the data in every game), converting the x-y coordinates to polar coordinates (in the units of feet and radians), and including an indicator for home games. The final dataset had 5467 observations from 31 shooters over 94 games. A summary of the cleaned dataset is in Table 3.1:

Table 3.1: Summary of Dataset

Name	Type	Values	Extra Details
season	categorical	{2014, . . . , 2017}	
gameid	categorical	NA	94 unique values
time	continuous	NA	13-digit timestamp in milliseconds
globalplayerid	categorical	NA	31 unique values
r	continuous	$[0, \infty)$	Distance of shot from hoop (feet)
theta	continuous	$[-\pi, \pi]$	Angle of shot (radians)
home	categorical	{0,1}	1 if shot occured during a home game
result	categorical	{0,1}	1 if shot was made(response)

and a small subset of the cleaned data is displayed below in Table 3.2:

Table 3.2: Sample of Dataset

season	gameid	time	globalplayerid	r	theta	home	result
2014	201401070173	1389141733839	603106	4.2076	1.0746	1	1
2014	201401070173	1389141844712	601140	16.6537	1.2973	1	0
2014	201401070173	1389143172185	696289	18.7901	-0.0581	1	1
2014	201401070173	1389143196303	601140	23.4629	0.9539	1	1
2014	201401070173	1389143220261	756880	6.5365	0.0696	1	0

Figure 3.1 shows the location of all the shots in the dataset, excluding heaves from beyond half court. The variable θ has a range of 2π radians, but this plot shows that most of the attempts occur within the interval $(-\frac{\pi}{2}, \frac{\pi}{2})$. This figure also shows the

bimodal distribution of shot distance over all players.

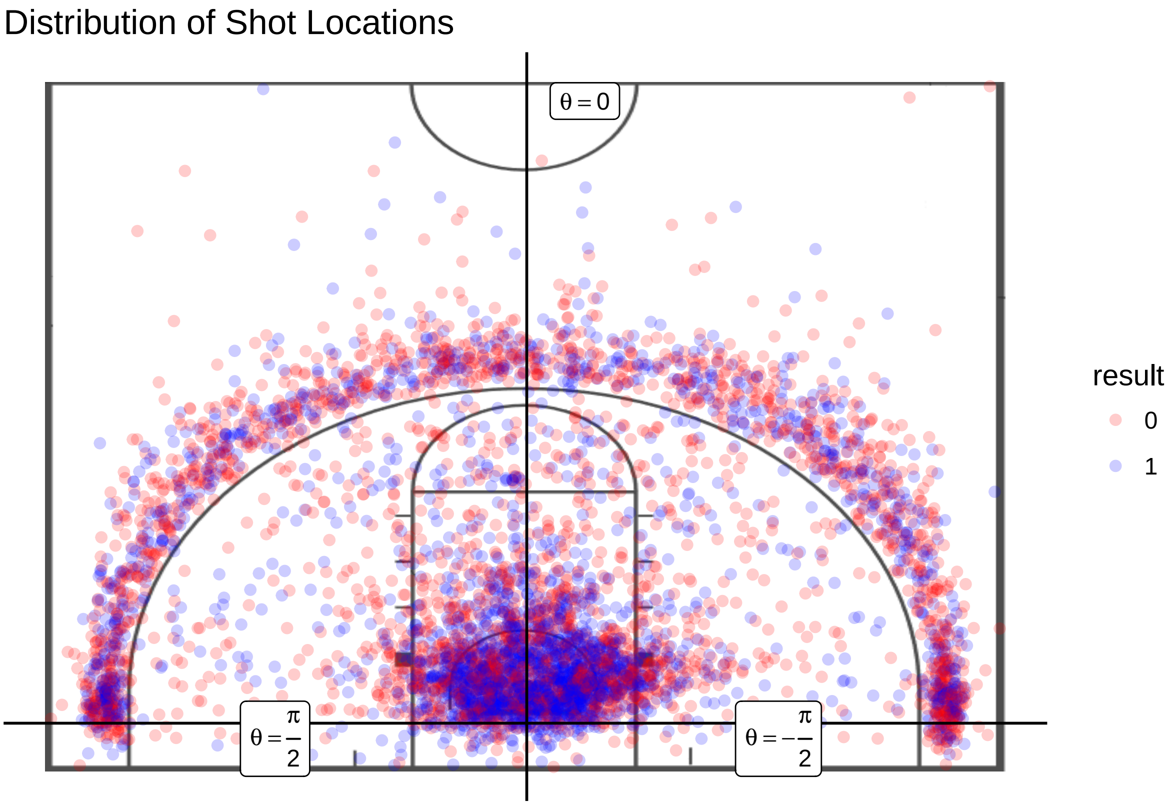
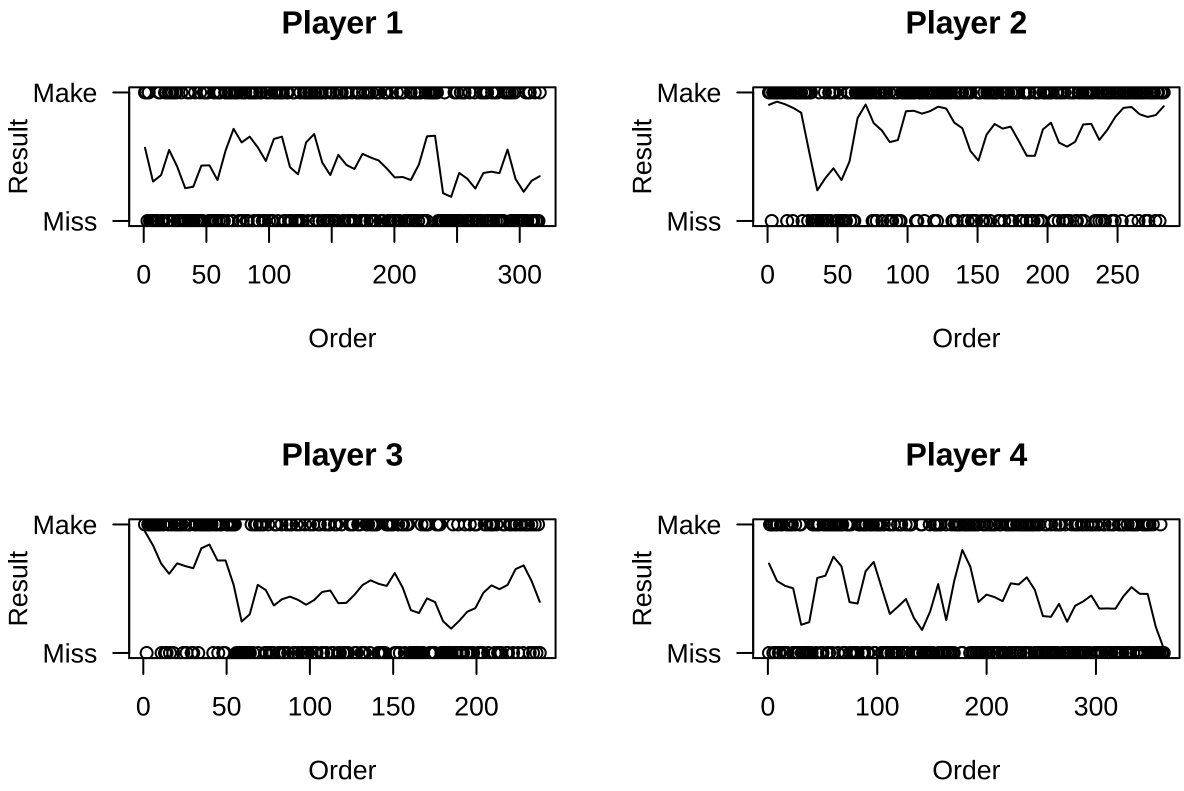


Figure 3.1: Locations and Results of All Shots

3.3 Exploratory Data Analysis

The following exploratory plots in Figure 3.2 examine how consistent the probability of a made shot is, using a loess smooth curve on the binary outcomes. We present these smoothed plots for four high-usage basketball players at Duke University between the 2013-2014 and 2016-2017 seasons, and we leave the others in Appendix **number**. Each plot represents a single player's ordered shooting outcomes for a single season. These plots do not account for the amount of time in between shots, but simply shot order and outcome.



We can see that the plots vary in the consistency of their made shots, since they all contain spikes and trends. For example, the third plot initially has a very high success rate, which quickly falls to the middle after about thirty shot attempts, and the second plot has a noticeable upward trend in shot success beginning around shot number one hundred fifty.

We investigate the shooting outcomes using Bayesian models, and present the results in the next section.

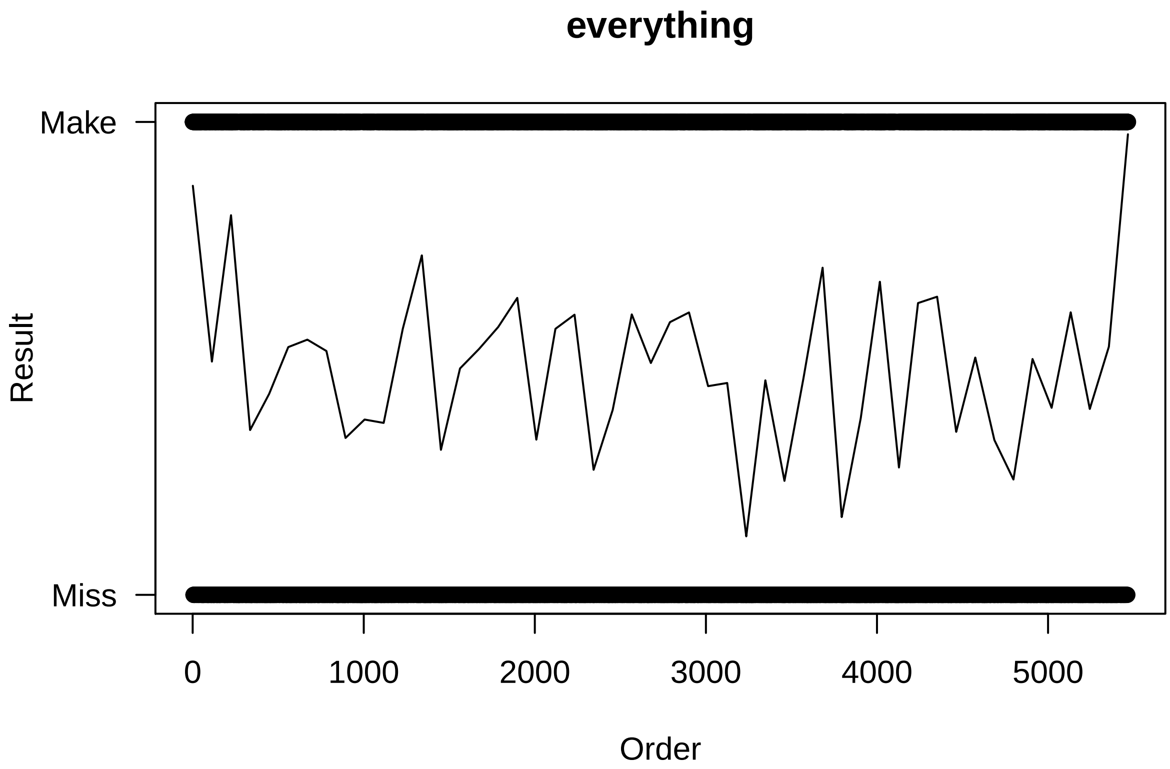


Figure 3.2: Moving Average of Shot Success Rate

Chapter 4

Models & Analysis

4.1 Description of Models

For our models, we consider the shot location, the shooter identity, a home court indicator, and a player's shooting success rate in nearby games as factors that can affect a shot outcome. In every model, we use the Just Another Gibbs Sampler library in R (`R2jags`). Each model is based off of a logistic regression model that provides the posterior distribution of the shot location parameters (distance and angle). The models do not account for covariance between these predictors. We expand upon this model by adding mixed effects and discounted likelihood models to control for shooter identity and game identity, respectively. These models will show us how consistent the shot location parameters are between shooters and between games. In our Gibbs Samplers, we estimate the posterior distributions using 10,000 simulations and a burn-in of 500. Our priors are made from the corresponding Maximum Likelihood Estimates for the first four games in the dataset, and we initialize our Monte Carlo Markov Chains using values of 0 for all means, and 1 for all variances.

The `R2jags` code used to build these models, as well as some diagnostic plots, can be found in Appendices **number** and **number**, respectively.

4.1.1 Generalized Linear Model

First, we build a logistic regression model of the following form:

$$\log\left(\frac{p}{1-p}\right) = \beta_{\text{int}} + x_{\text{r}}\beta_{\text{r}} + x_{\theta}\beta_{\theta} + x_{\text{H}}\beta_{\text{H}}.$$

In this model, the x refers to the data, and the βs are the pareters from the model. The subscripts *int*, r, θ, and H respectively refer to the intercept, the log-distance of the shot, the angle of the shot, and whether shot was taken on Duke's home court or another gym.

4.1.2 Hierarchical Generalized Linear Model

Our second model is a hierarchical model, with random effects on the j players in the dataset. These random effects occur for each of the four parameters of interest—the intercept, the home effect, the distance effect, and the angle effect. Each individual player's parameter values are sampled from a Normal distribution centered at the population values. The parameters for players without a lot of shot attemps are shrunk towards the population means.

$$\log\left(\frac{p}{1-p}\right) = \beta_{\text{int, j}} + x_{\text{r}}\beta_{\text{r, j}} + x_{\theta}\beta_{\theta,j} + x_{\text{H}}\beta_{\text{H, j}},$$

$$\begin{aligned}\beta_{\text{int, j}} &\sim N(\beta_{\text{int}}, \tau^2_{\text{int}}),\\ \beta_{\text{r, j}} &\sim N(\beta_{\text{r}}, \tau^2_{\text{r}}),\\ \beta_{\theta,j} &\sim N(\beta_{\theta}, \tau^2_{\theta}),\\ \beta_{\text{H, j}} &\sim N(\beta_{\text{H}}, \tau^2_{\text{H}}).\end{aligned}$$

4.1.3 Discounted Likelihood Hierarchical Model

In the discounted likelihood models, the likelihood of a single observation at time t is more heavily influenced by the observations close to it than the observations far away from it. We measure the "distance" between observations by the number of games between them; a shot attempt that occurs in the next or the preceding game will influence the likelihood of the observation more than a shot that occurs two games away. To conceptualize this model, we first start with a Forward Filtering model that uses a discounted Bayes Theorem. The discounted Bayes Theorem, like the traditional Bayes Theorem, states that the posterior is proportional to the product of the likelihood and the prior, but it weights the most recent data higher than the earlier data, which is weighted higher than the prior.

In other words, given the parameters:

$$\Theta = (\beta_{\text{int}}, \beta_{\text{r}}, \beta_{\theta}, \beta_{\text{H}}, \tau^2_{\text{int}}, \tau^2_{\text{r}}, \tau^2_{\theta})$$

the joint posterior distribution of the parameters given the data up to time unit t, X_t, is:

$$P_{g_t}(\Theta|X_t) \propto P(\Theta)^{\delta^t} P(X_1|\Theta)^{\delta^{t-1}} P(X_2|\Theta)^{\delta^{t-2}} ... P(X_{t-2}|\Theta)^{\delta^2} P(X_{t-1}|\Theta)^{\delta} P(X_t|\Theta)$$

$$0 < \delta < 1$$

This equation shows that as distance from time t increases, the effect on the posterior distribution decreases. The Reverse Updating extension of this concept allows the effect to take place as the distance from t increases in either the positive or negative direction.

$$P_{g_t}(\Theta|X_t) \propto P(\Theta)^{\delta^t}...P(X_{t-2}|\Theta)^{\delta^2}P(X_{t-1}|\Theta)^{\delta}P(X_t|\Theta)P(X_{t+1}|\Theta)^{\delta}P(X_{t+2}|\Theta)^{\delta^2}...$$

Specifically for our model, given an observed shot i in game g_i, attempted by player j, we use the above concept to discount the probability of a made shot in the hierarchical model.

$$\log\left(\frac{p}{1-p}\right) = \beta_{\mathrm{int,\ j}} + x_{\mathrm{r}}\beta_{\mathrm{r,\ j}} + x_{\theta}\beta_{\theta,j} + x_{\mathrm{H}}\beta_{\mathrm{H,\ j}},$$

$$L(p) = \Pi_{\mathrm{i=1}}^{\mathrm{n}} p^{\mathrm{y}_i}(1-p)^{1-\mathrm{y}_i},$$

$$\delta_i = \Delta^{|g_i - g_t|},$$

$$\pi = L(p)^{\delta_i},$$

$$\beta_{\mathrm{int,\ j}} \sim N(\beta_{\mathrm{int}}, \tau^2_{\mathrm{int}}),$$

$$\beta_{\mathrm{r,\ j}} \sim N(\beta_{\mathrm{r}}, \tau^2_{\mathrm{r}}),$$

$$\beta_{\theta,j} \sim N(\beta_{\theta}, \tau^2_{\theta}),$$

$$\beta_{\mathrm{H,\ j}} \sim N(\beta_{\mathrm{H}}, \tau^2_{\mathrm{H}}).$$

In this model, p represents the binomial likelihood, and π is the discounted likelihood. Both of these quantities are bounded in the interval [0,1]. The contribution of observed shot outcomes (in the "anchor game" g_t) to the likelihood of the current shot outcome (in game g_i) decreases as the distance between the observations increases, and as Δ decreases. In a model with $\Delta = 0$, only shots taken in the same game as g_i can contribute to the likelihood, while $\Delta = 1$ is equivalent to a model with no discounting. This model specification results in an MCMC chain for each combination of g_0 and Δ. Figure 4.1 illustrates how the likelihood weight of δ depends on the selected value of Δ and the distance from the anchor game ($|g_i - g_t|$). To build these discounted likelihood models in the `R2jags` library, we apply the "ones trick". This technique allows us to do a sampling distribution that do not exist in the library by modifying a common distribution—in this case, the Binomial. The probability p is estimated using the Bayesian hierarchical model. We discount this probability to estimate π, and then specify that artificial data that consists only of ones comes from a Binomial distribution with the discounted probability; this is equivalent to sampling from a distribution with discounted outcomes. **elaborate on the ones trick**. See Appendix **number** for the `R` code.

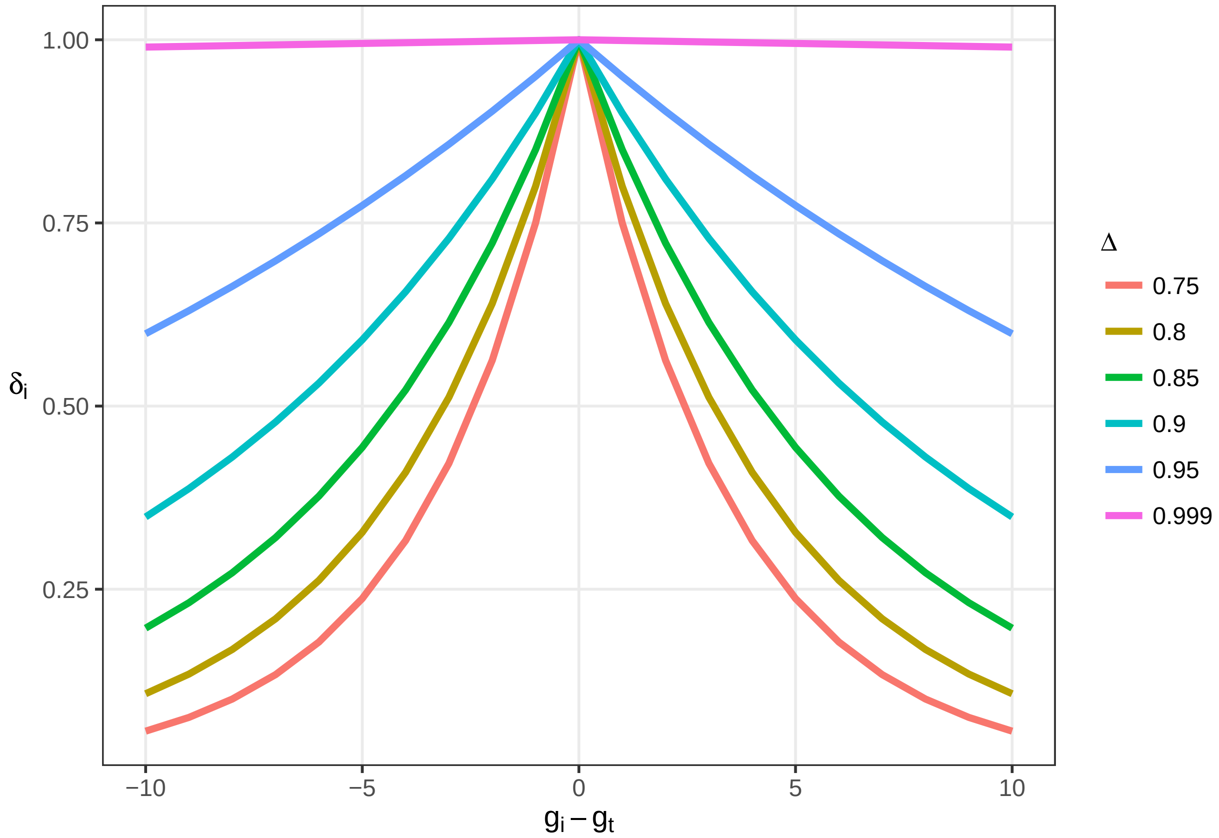
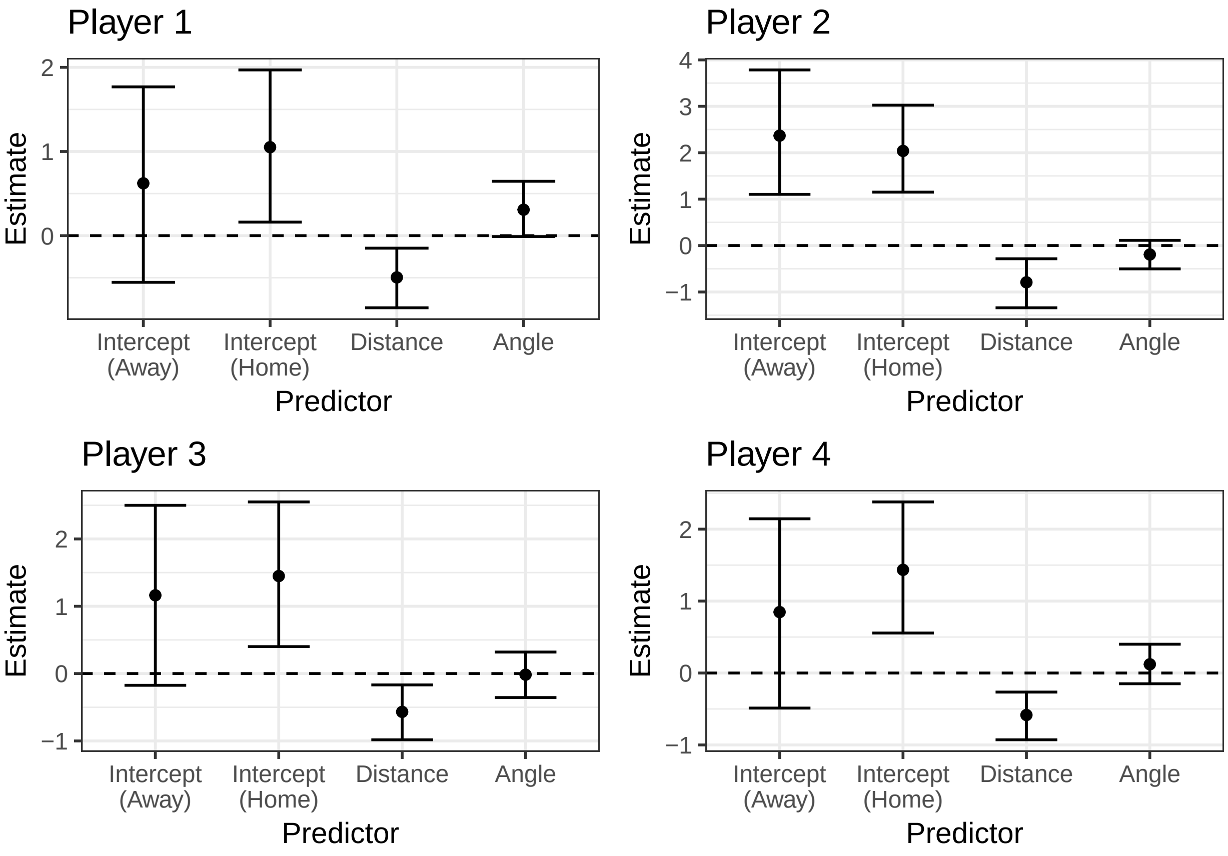


Figure 4.1: Illustration of Discounted Weighting

4.2 Analysis

4.2.1 Generalized Linear Model

In our logistic regression model, we only look at shot location and the home court indicator as predictors for the shot outcome. To look at these effects for particular players, we simply subset the dataset to shots attempted by that player before running the Gibbs Sampler. Below in Figure 4.2, The 95% credible intervals of the posterior parameters are reported for the same four players that were introduced in Figure 3.2.



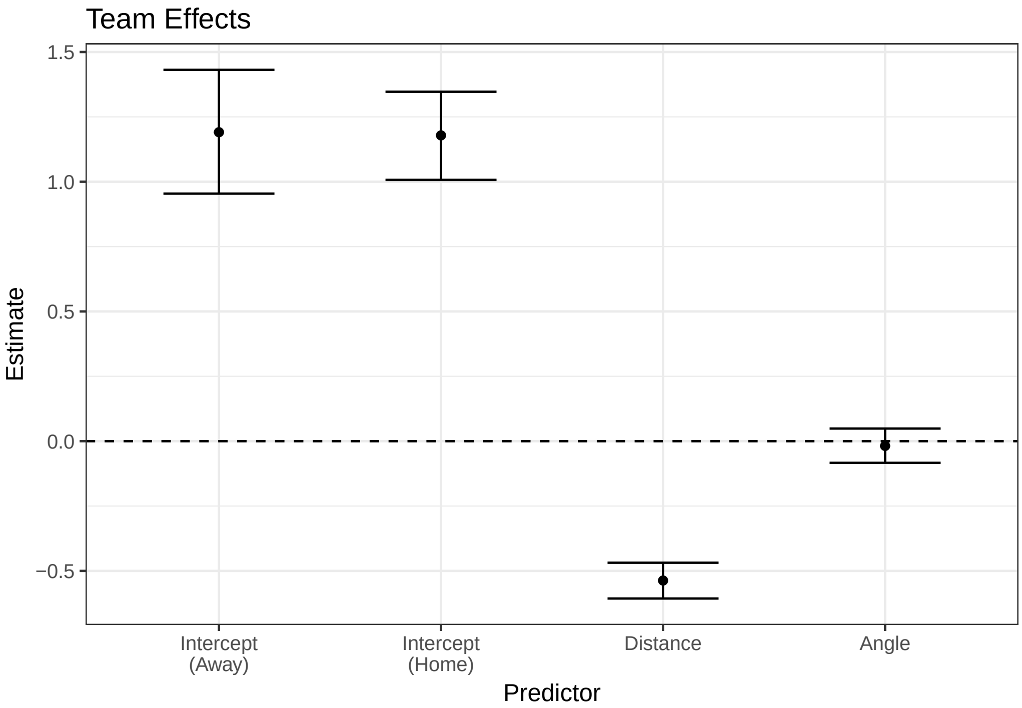


Figure 4.2: GLM Posterior Distributions for Four Players

In the generalized linear model, the intercepts correspond to the log-odds of making a shot when angle is zero (the middle of the court) and the log distance is zero (one foot away from the hoop).

From these plots, we see that the team-wide 95% credible interval of the angle effect contains zero, and it is therefore probably not predictive of a made shot. The average distance effect shows us that the log-odds of a made shot decrease by $\beta_r = 0.5372$ as the log distance increases by one unit and the other predictors remain constant. In the probability scale, this decrease is:

$$\frac{\mathrm{e}^{\beta_r}}{1+\mathrm{e}^{\beta_r}}$$

which is equal to 0.3689.

We also see that the 95% credible interval on the effect of distance is completely negative, which follows the intuitive idea that the probability of a made shot significantly decreases as distance from the basket increases. The intercepts show us that there is not a substantial diference in baseline shooting performance between home games and away games.

4.2.2 Hierarchical Generalized Linear Model

In this hierarchical model, we add random effects to allow the parameters to vary for each player in the dataset. We present the results below using densities of how the four players of interest compare to the population distribution, and by using contour plots that illustrate each player's probability of a made shot given their location on the court for a game at home. These hierarchical model results show us characteristics of our four high-usage players of interest compared to the population of players in the dataset. For example, the intercept plots show us that Player 2 is excellent at finishing under baseline conditions, but he also has a steeper-than-averge drop in his odds of scoring as his distance from the basket increases. This means that most of his scoring occurs close to the basket. We can also see that Player 1 strongly increases his odds of scoring when his angle is negative, which corresponds to the left side of the basket.

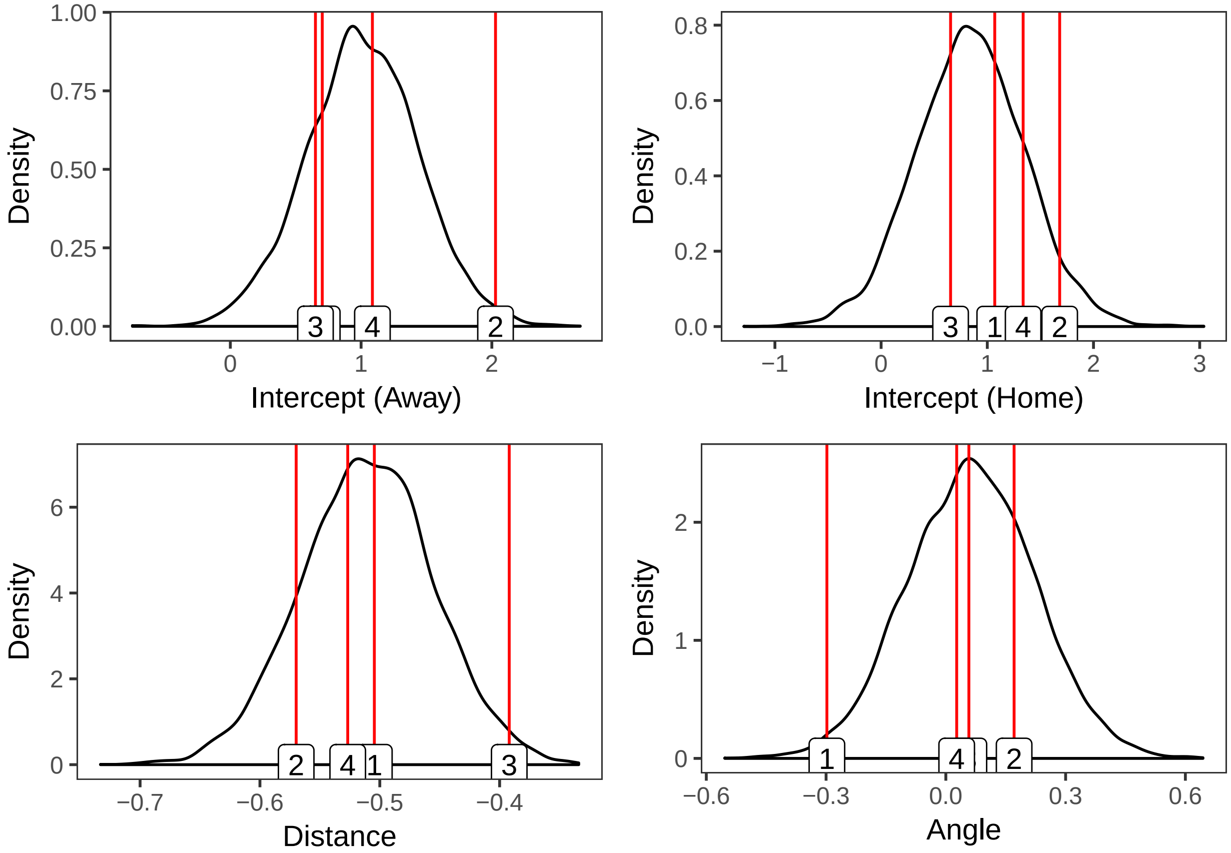
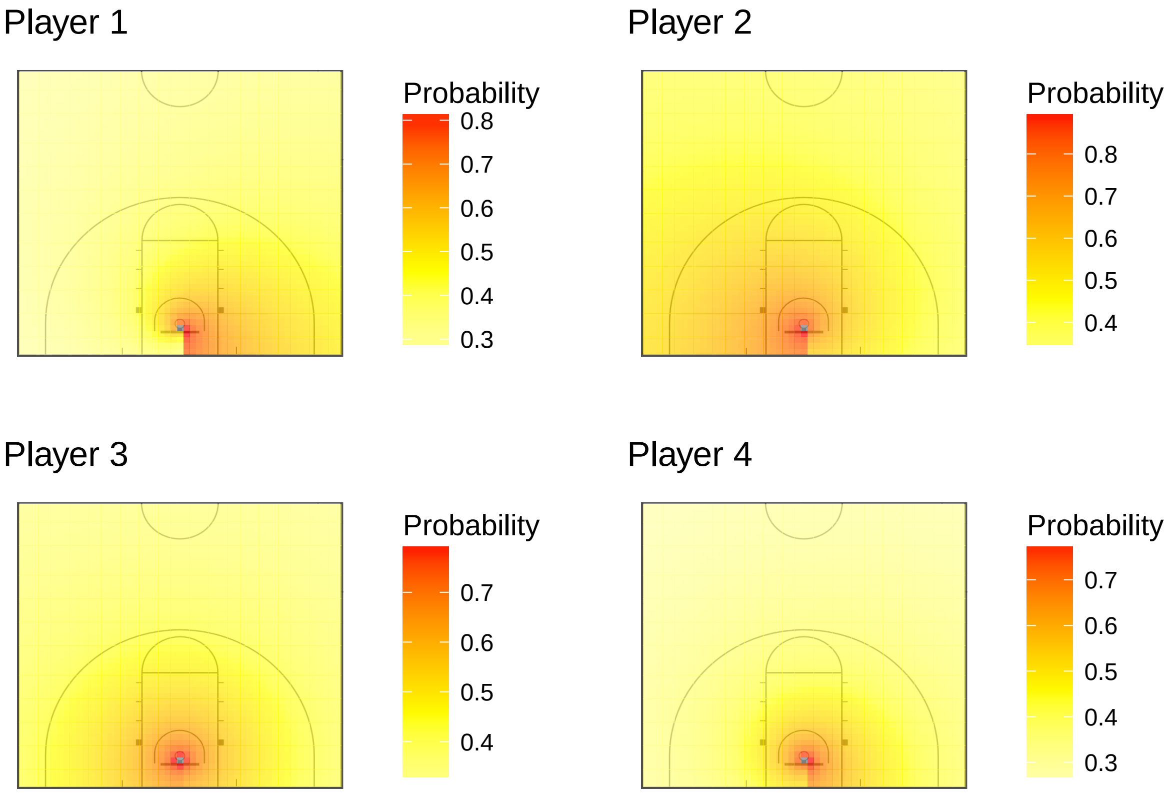


Figure 4.3: Population Distribution with Four Player Effects

In the figures below, we have contour plots showing players' expected field goal percentages at different locations on the court. Between our four players of interest, we can observe how Player 1 is more effective on the left side of the basket than the others, and how Player 2 has the darkest overall contour plot, suggesting he has the highest probability of scoring among those four.



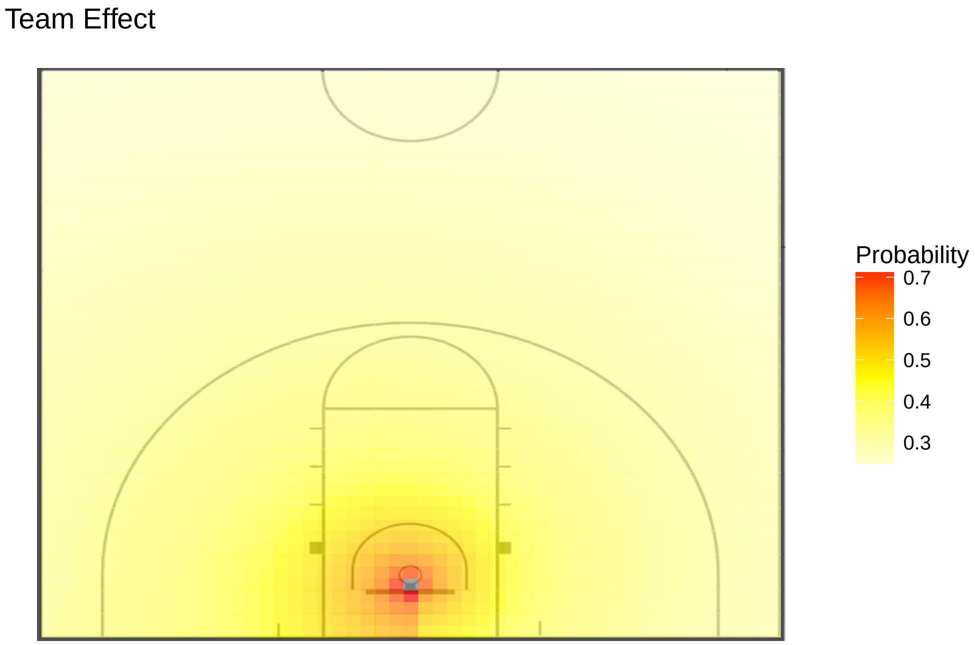


Figure 4.4: Contour Plots for Four Players and Population of Players

4.2.3 Discounted Likelihood Hierarchical Model

The values of Δ that we use to fit the discounted likelihood models are 0.750, 0.800, 0.850, 0.900, 0.950, and 0.999. We also build an MCMC chain to estimate the posterior using every game as g_0. We calculate predictions and fitted values for a particular shot in game g_i using the posterior median of the MCMC chain where g_i is g_0. If models with larger values of Δ best fit the data, this suggests that shooting success is consistent throughout a career. If smaller values of Δ are more likely in the data, however, then we can assume there is a substantial amount of time variation in the data on the game level.

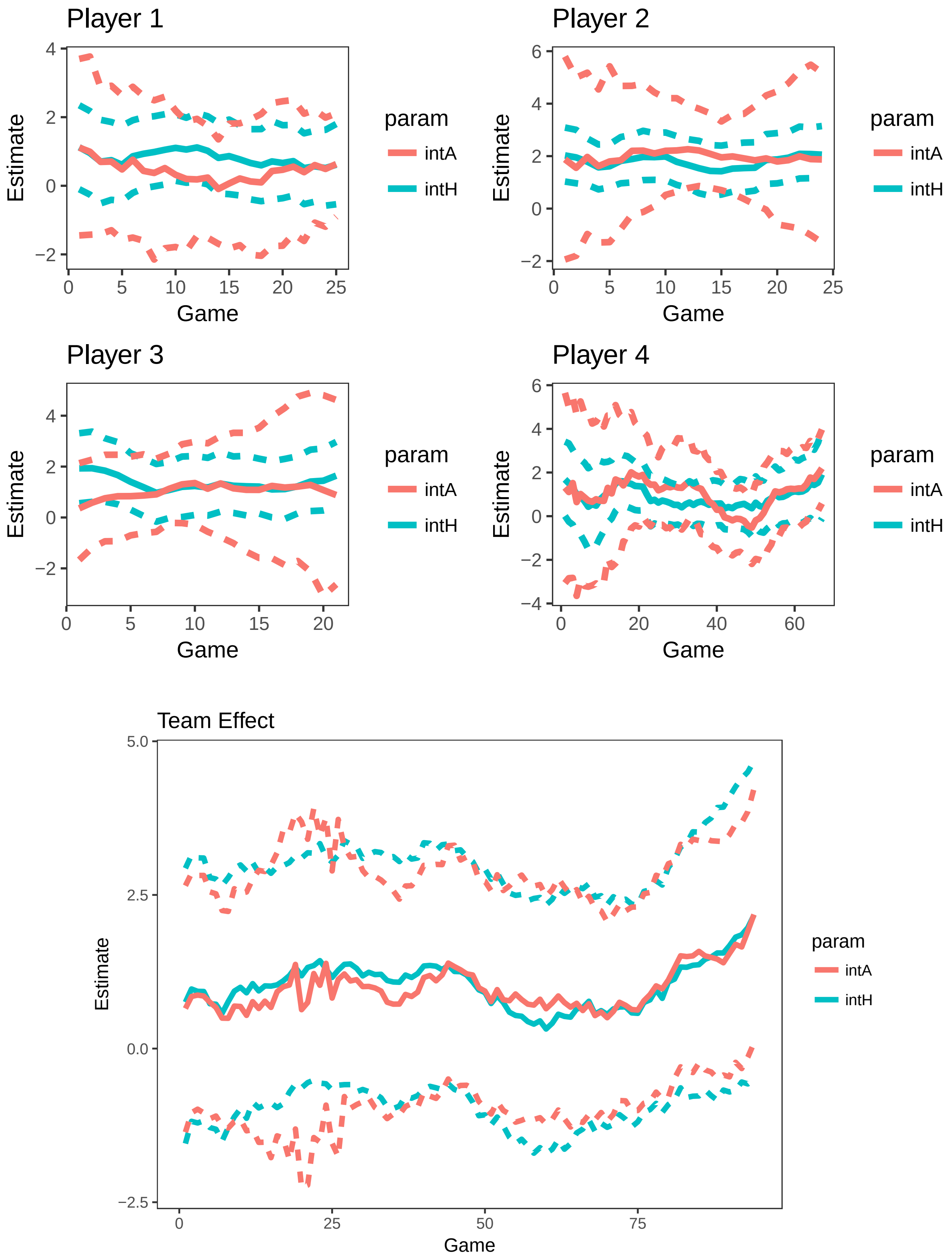


Figure 4.5: Intercepts for Four Players and Population over Time, $\Delta = 0.750$

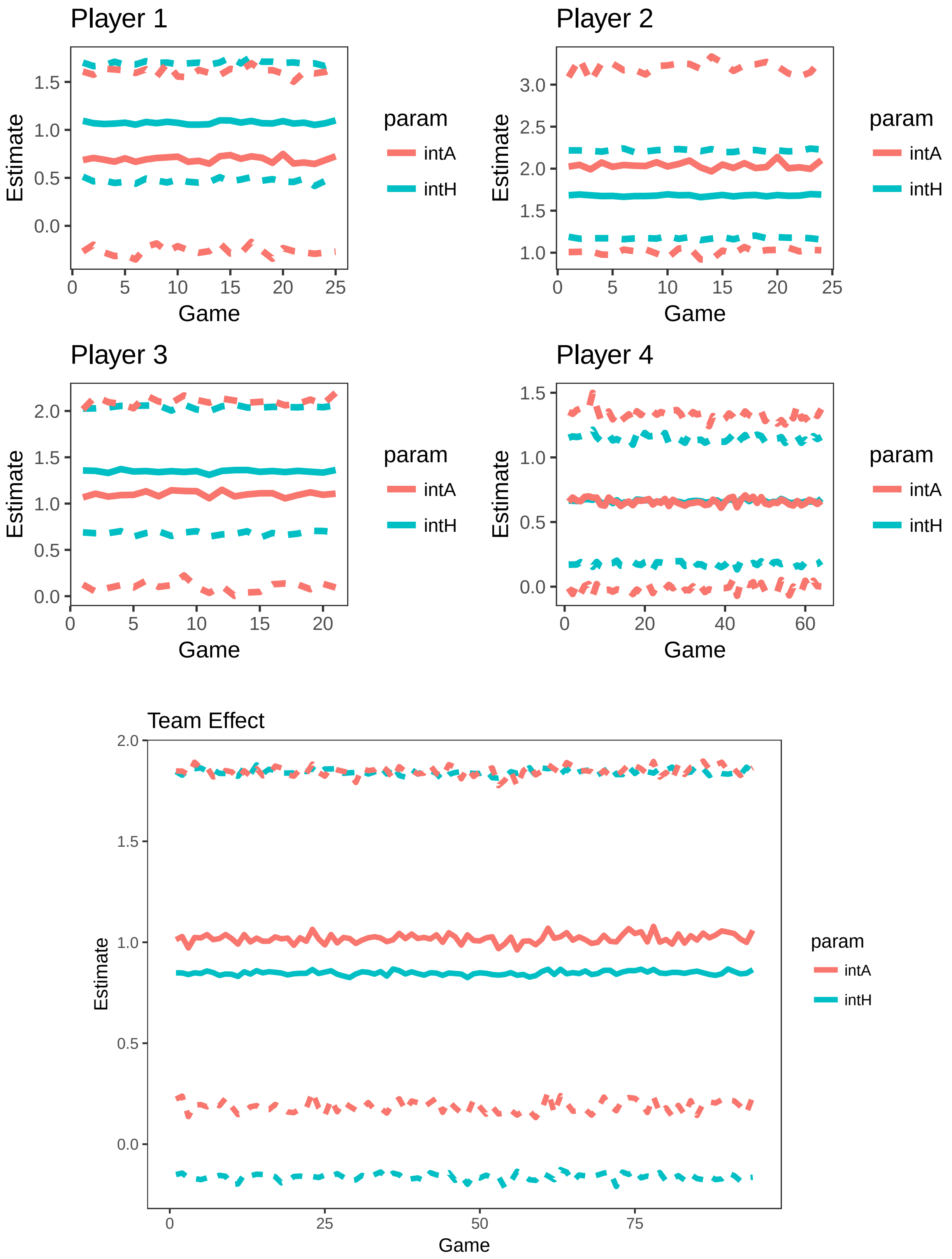


Figure 4.6: Intercepts for Four Players and Population over Time, $\Delta = 0.999$

The plots above illustrate how the intercepts change for players over time. These plots illustrate that smaller values of Δ allow for greater variation in parameter values over time.

Chapter 5

Discussion

Summarize and reflect upon results

5.1 Evaluation of Models

To evaluate these models, we use 5-fold cross-validation. This process used as many as 20 simultaneous RStudio Pro servers provided by the Duke University Statistical Science Department. In each train-test split, we evaluate a model's out-of-sample classification rate (using a cutoff probability of 0.5), Brier score (mean squared error), and log likelihood. The predictions and fitted values are obtained using MCMC averages; to calculate the probability for an individual shot, we calculate a response for each of the 9,500 posterior simulations, then take the average of those responses. The results are plotted below: From Figure 5.1, we can observe that the models have different strengths. The discounted likelihood model with the smallest value of Δ consistently has the highest likelihood. However, it does not test as well as the other models in areas of out-of-sample classification rate and Brier score. This suggests that models with smaller values of Δ, where the likelihood of an observed shot is more heavily influenced by shots closer to it, may overfit the model to the training data. The generalized linear models perform best in Brier score, but worst in log likelihood. The hierarchical models are about the same, but they have a better log likelihood performance than the GLMs. Therefore, a model that balances the trade-off between predictive accuracy and likelihood is a discounted likelihood model with Δ = 0.850. In addition, we can see that the variation in model performance is small. For example, most of the out-of-sample classification rates fall between 0.58 and 0.62. This is within the 95% confidence interval for a proportion of 0.6 using a sample size of 40 (for each combination of 5-fold cross-validation and 8 different models), which is (0.5225, 0.6775). Therefore, the evidence that certain models predict better than others is not particularly strong.

For the discounted likelihood model with $\Delta = 0.850$, we build calibration plots to

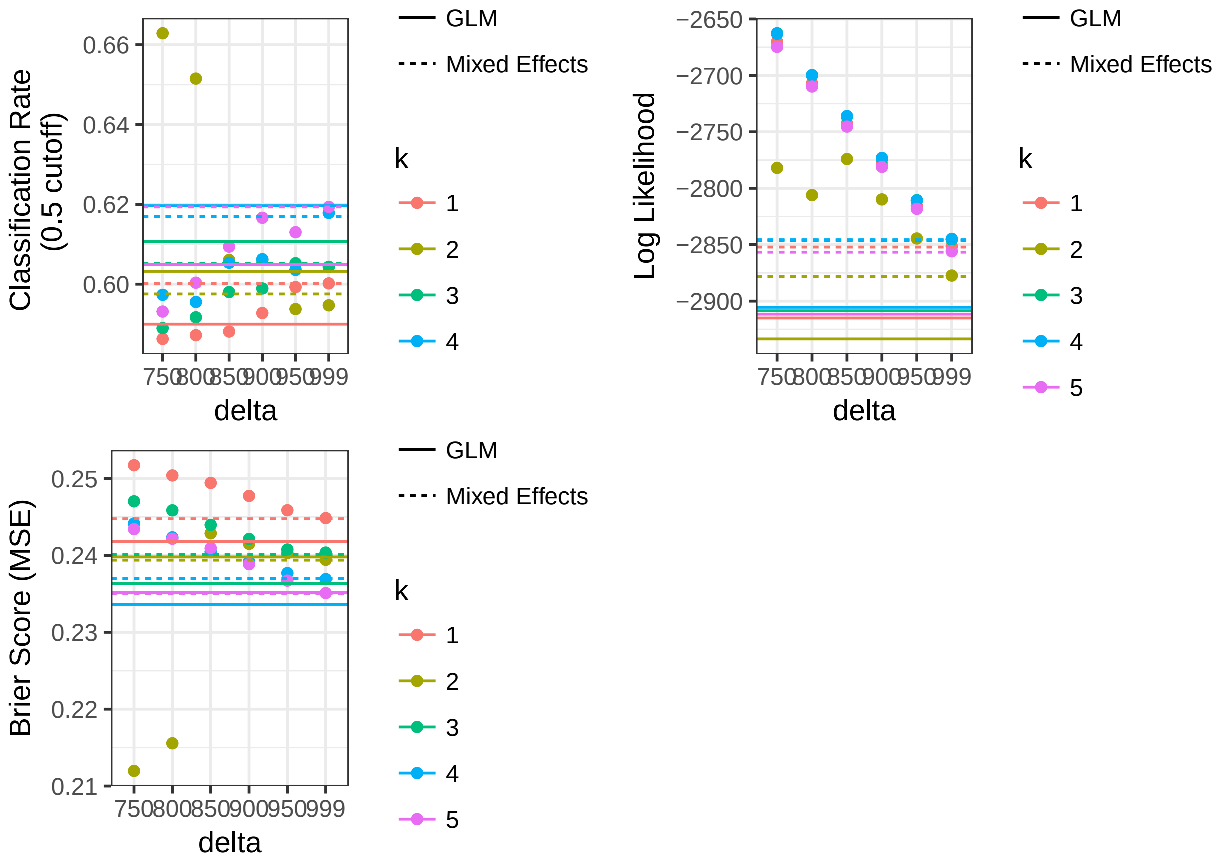


Figure 5.1: Model Evaluation

assess how well the estimated probabilities fit the actual proportions. In Figure **??**, we present these plots for the training set, the testing set, and for two random games.

in prev paragrpah, also include a point about how all the models are fairly similar. like ranges of classification rates, and expected standard deviations (MLEs)

To account for possible unexplained variation between seasons, and variation introduced from having such a small population of road games, I repeated this analysis on a subset of the data that only consisted of shots from available games in the 2015 season (25 games), and shots from home games (82 games). **results!** (show these results in Appendix?)

5.2 Conclusion

The results of this paper show that there is not a lot of time-dependency in shooting success rate in this dataset of player-tracking data from the Duke Men's Basketball team. The fact that the models with higher values of

5.3 Future Goals

Future goals for this research are to build a better-fitting model to predict basketball shots using more advanced factors that can be approximated from the dataset; possibilities for this include using the distance of the nearest defender as a proxy for defense quality, or using the amount of time a player has played without a substitution or timeout to approximate fatigue.

Appendix A

Appendix 1: Code

A.1 Generalized Linear Model

```r
Xtrainsub <-
  Xtrain %>%
  filter(as.integer(as.factor(gameid)) < 5)

priormod <- glm(formula = result ~ log(r) + theta,
                data = Xtrainsub,
                family = "binomial")

mu0r <- summary(priormod)[["coefficients"]]["log(r)","Estimate"]
mu0theta <- summary(priormod)[["coefficients"]]["theta","Estimate"]

fit_glm <- function(dat, S = 10000, B = 500){

  model.glm <- function(){

    # Liklihood function (for N observations)
    for(i in 1:N){

      logit(prob[i]) <- beta_int*int[i] +
        beta_home*home[i] +
        beta_r*logr[i] +
        beta_theta*theta[i]

      result[i] ~ dbern(prob[i])
    }

    # Priors
  # we expect less variation in the distance parameter,
  # because shot success rate should get worse
  # as distance increases under baseline circumstances.
  beta_int   ~ dnorm(0, 0.1)
  beta_home  ~ dnorm(0, 0.1)
  beta_r     ~ dnorm(mu0r, 0.01)
  beta_theta ~ dnorm(mu0theta, 0.1)
}

datlist.glm <-
  list(
    int = rep(1, nrow(dat)),
    logr = log(dat$r),
    theta = dat$theta,
    result = dat$result,
    home = dat$home,
    N = nrow(dat),
    mu0r = mu0r,
    mu0theta = mu0theta
  )

params.glm <- c("beta_int",
                "beta_home",
                "beta_r",
                "beta_theta")

initslist <- list(list("beta_int"=0,
                       "beta_r"=0,
                       "beta_theta"=0,
                       "beta_home"=0))

sim <-
  jags(data = datlist.glm,
       n.chains = 1, n.iter = S, n.burnin = B, n.thin = 1,
       inits=initslist,
       parameters.to.save = params.glm,
       model.file=model.glm
  )

sim.mcmc <- as.data.frame(as.mcmc(sim)[[1]])

# Changing from a baseline mean + a shift amount
# to two different means based on the type of game.
sim.mcmc <-
```

```
    sim.mcmc %>%
    mutate(beta_intA = beta_int,
           beta_intH = beta_int + beta_home) %>%
    select(beta_intA, beta_intH, beta_r, beta_theta)

  return(sim.mcmc)
}
```

A.2 Hierarchical Generalized Linear Model

```
Xtrainsub <-
  Xtrain %>%
  filter(as.integer(as.factor(gameid)) < 5)

priormod <- glm(formula = result ~ log(r) + theta,
                data = Xtrainsub,
                family = "binomial")

mu0r <- summary(priormod)[["coefficients"]]["log(r)","Estimate"]
mu0theta <- summary(priormod)[["coefficients"]]["theta","Estimate"]

fit_players <- function(dat = NA, S = 10000, B = 500){

  model.player <- function(){

    # Likelihood function for N observations
    for(i in 1:N){

      # the parameters now vary by player id.
      logit(prob[i]) <- beta_int[player[i]]*int[i] +
        beta_home[player[i]]*home[i] +
        beta_r[player[i]]*logr[i] +
        beta_theta[player[i]]*theta[i]

      result[i] ~ dbern(prob[i])
    }

    # Priors
    for(j in 1:M){
        beta_int[j] ~ dnorm(beta_int0,tau_int)
        beta_home[j]  ~ dnorm(beta_home0, tau_int)
        beta_r[j] ~ dnorm(beta_r0,tau_r)
      beta_theta[j] ~ dnorm(beta_theta0,tau_theta)
  }

  # Hyperpriors
  beta_int0   ~ dnorm(0, 0.1)
  beta_home0  ~ dnorm(0, 0.1)
  beta_r0     ~ dnorm(mu0r, 0.01)
  beta_theta0 ~ dnorm(mu0theta, 0.1)
  tau_int ~ dgamma(10, 100)
  tau_r ~ dgamma(10, 0.2)
  tau_theta ~ dgamma(10, 10)
}

datlist.player <-
  list(
    logr = log(dat$r),
    theta = dat$theta,
    home = dat$home,
    result = dat$result,
    player = as.integer(as.factor(dat$globalplayerid)),
    N = nrow(dat),
    int = rep(1, nrow(dat)),
    M = n_distinct(dat$globalplayerid),
    mu0r = mu0r,
    mu0theta = mu0theta
  )

# we want posteriors for the overall effects
# and for the individual player effects
params <- c("beta_int",
            "beta_home",
            "beta_r",
            "beta_theta",
            "beta_int0",
            "beta_home0",
            "beta_r0",
            "beta_theta0")

M <- datlist.player$M

initslist <- list(
  list("beta_int"=rep(0,M),
       "beta_home"=rep(0,M),
       "beta_r"=rep(0,M),
       "beta_theta"=rep(0,M),
       "beta_int0"=0,
       "beta_home0"=0,
       "beta_r0"=0,
       "beta_theta0"=0,
       "tau_int"=1,
       "tau_r"=1,
       "tau_theta"=1
))

sim.player <-
  jags(data = datlist.player,
       n.iter = S, n.chains = 1, n.burnin = B, n.thin = 1,
       inits=initslist,
       parameters.to.save = params,
       model.file=model.player
)

sim.mcmc.player <- as.data.frame(as.mcmc(sim.player)[[1]])


# Changing from a baseline mean + a shift amount
# to two different means based on the type of game.
hometext <- paste0("`beta_intH[",1:M,"]` =
                   `beta_int[",1:M,"]` +
                   `beta_home[",1:M,"]`",
                   collapse=",\n")

awaytext <- paste0("`beta_intA[",1:M,"]` =
                   `beta_int[",1:M,"]`",
                   collapse=",\n")

sim.mcmc.player <- eval(parse(text=
  paste0("sim.mcmc.player %>%
          mutate(",hometext,", beta_intH0 = beta_int0 + beta_home0)",
         " %>% rename(",awaytext,", beta_intA0 = beta_int0)"
         ))) %>%
  select(grep("(beta_int)|(beta_theta)|(beta_r)",names(.)))

sim.mcmc.player <- sim.mcmc.player[ ,order(colnames(sim.mcmc.player))]


# Renaming mixed effects columns from default factor levels (integers) to the
```

```
  factorids <- str_extract_all(names(sim.mcmc.player), "[[:digit:]]+") %>% as.numeric()
  fids <- data.frame(factorid = factorids, order = 1:length(factorids))

  datmap <- dat %>%
    mutate(factorid = as.integer(as.factor(globalplayerid))) %>%
    select(globalplayerid, factorid)

  gameids <- merge(datmap, fids, all.x=FALSE,all.y=TRUE) %>%
    unique() %>%
    mutate(globalplayerid = ifelse(is.na(globalplayerid),0,globalplayerid)) %>%
    arrange(order)

  names(sim.mcmc.player) <- str_replace_all(names(sim.mcmc.player), "[[:digit:]]+", as.
  -
  return(sim.mcmc.player)

}
```

A.3 Discounted Likelihood Hierarchical Model

```
fit_game <- function(dat = NA, g0 = NA, Delta = NA, S = 10000, B = 500){

  model.game <- function(){

    for(i in 1:N){
      delta[i] <- Delta^abs(games[i]-g0)
      # Delta = discount rate for game g relative to anchor game g0

      # player-level random effects
      logit(prob[i]) <- beta_int[player[i]]*int[i] +
                        beta_home[player[i]]*home[i] +
                        beta_r[player[i]]*logr[i] +
                        beta_theta[player[i]]*theta[i]

      # Likelihood function
      p1[i] <- prob[i]^result[i]
      p2[i] <- (1-prob[i])^(1-result[i])

      # Discounted likelihood function
      pi[i] <- (p1[i] * p2[i])^delta[i]

      # defines correct discounted likelihood function
    y[i] ~ dbern(pi[i])

    # result = actual outcome
    # prob   = actual likelihood
    # y      = artificial "ones trick" outcomes
    # pi     = discounted likelihood
  }

  # Priors
  for(j in 1:M){
    beta_int[j] ~ dnorm(beta_int0,tau_int)
    beta_home[j] ~ dnorm(beta_home0, tau_int)
    beta_r[j] ~ dnorm(beta_r0,tau_r)
    beta_theta[j] ~ dnorm(beta_theta0,tau_theta)
  }

  # Hyperoriors
  beta_int0   ~ dnorm(0, 0.1)
  beta_home0  ~ dnorm(0, 0.1)
  beta_r0     ~ dnorm(mu0r, 0.01)
  beta_theta0 ~ dnorm(mu0theta, 0.1)
  tau_int ~ dgamma(10, 100)
  tau_r ~ dgamma(10, 0.2)
  tau_theta ~ dgamma(10, 10)
}

datlist.game <-
  list(
    int = rep(1, nrow(dat)),
    logr = log(dat$r),
    theta = dat$theta,
    result = dat$result,
    home = dat$home,
    player = as.integer(as.factor(dat$globalplayerid)),
    N = nrow(dat),
    M = n_distinct(dat$globalplayerid),
    mu0r = mu0r,
    mu0theta = mu0theta,
    Delta = Delta,
    games = as.integer(as.factor(dat$gameid)),
    g0 = g0,
    y = rep(1, nrow(dat))
  )
```

```
params <- c("beta_int",
            "beta_r",
            "beta_home",
            "beta_theta",
            "beta_int0",
            "beta_home0",
            "beta_r0",
            "beta_theta0")

M <- n_distinct(dat$globalplayerid)

initslist <- list(list("beta_int"=rep(0,M),
                       "beta_r"=rep(0,M),
                       "beta_theta"=rep(0,M),
                       "beta_int0"=0,
                       "beta_r0"=0,
                       "beta_theta0"=0,
                       "tau_int"=1,
                       "tau_r"=1,
                       "tau_theta"=1
                       ))

sim.game <- jags(data = datlist.game,
            n.iter = S, n.chains = 1, n.burnin = B, n.thin = 1,
            inits = initslist,
            parameters.to.save = params,
            model.file=model.game
)
sim.mcmc.game <- as.data.frame(as.mcmc(sim.game)[[1]])

# Changing from a baseline mean + a shift amount
# to two different means based on the type of game.

hometext <- paste0("`beta_intH[",1:M,"]` = `beta_int[",1:M,"]` + `beta_home[",1:M,"]`"
awaytext <- paste0("`beta_intA[",1:M,"]` = `beta_int[",1:M,"]`", collapse=",\n")

sim.mcmc.game <- eval(parse(text=
  paste0("sim.mcmc.game %>%
          mutate(",hometext,", beta_intH0 = beta_int0 + beta_home0)",
         " %>% rename(",awaytext,", beta_intA0 = beta_int0)")
  )) %>% select(grep("(beta_int)|(beta_theta)|(beta_r)",names(.)))


# Renaming mixed effects columns from default factor levels (integers) to the corres
```

```
  sim.mcmc.game <- sim.mcmc.game[ , order(colnames(sim.mcmc.game))]

  factorids <- str_extract_all(names(sim.mcmc.game), "[[:digit:]]+") %>%
    as.numeric()
  fids <- data.frame(factorid = factorids, order = 1:length(factorids))

  datmap <- dat %>%
    mutate(factorid = as.integer(as.factor(globalplayerid))) %>%
    select(globalplayerid, factorid)

  gameids <- merge(datmap, fids, all.x=FALSE,all.y=TRUE) %>%
    unique() %>%
    mutate(globalplayerid = ifelse(is.na(globalplayerid),0,globalplayerid)) %>%
    arrange(order)

  names(sim.mcmc.game) <- str_replace_all(names(sim.mcmc.game), "[[:digit:]]+", as

  return(sim.mcmc.game)

}
```

Appendix B

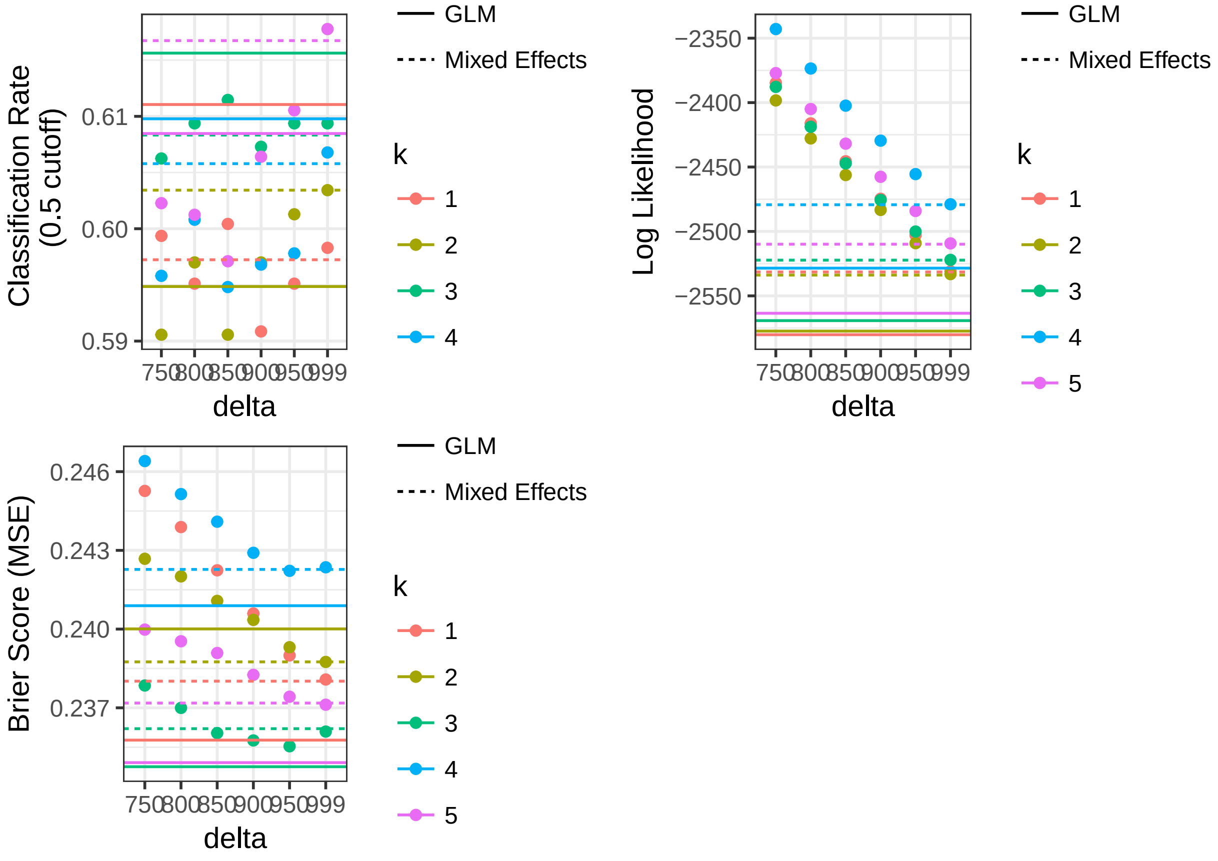
Appendix 2: Diagnostic Plots

B.1 Generalized Linear Model:

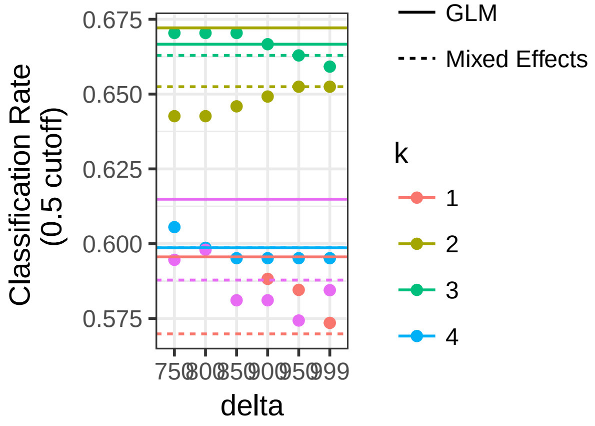
Appendix C

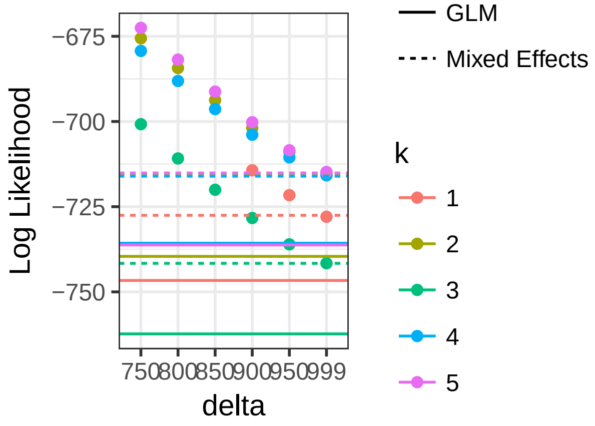
Appendix 3: Reproducing Models

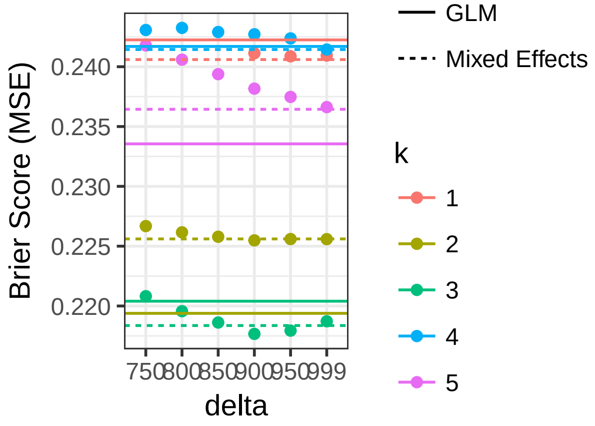
C.1 Home Games Only



C.2 2015 Season Only







References

Albert, J. (1993). Statistical analysis of hitting streaks in baseball: Comment. *Journal of the American Statistical Association*, *88*(424), 1184–1188.

Albert, J. (2013). Looking at spacings to assess streakiness. *Journal of Quantitative Analysis in Sports*, *9*(2), 1–13.

Albert, J., & Williamson, P. (1999). Using model/data simulations to detect streakiness. *The American Statistician*, *55*, 41–50.

Bar-Eli, M., Avugos, S., & Raab, M. (2006). Twenty years of "hot hand" research: Review and critique. *Psychology of Sport and Exercise*, *7*, 525–553.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, *17*, 295–314.

Miller, J. B., & Sanjurjo, A. (2016). Surprised by the gambler's and hot hand fallacies? A truth in the law of small numbers. *IGIER Working Paper No. 552*.

Ryan Wetzels, e. a. (2016). A bayesian test for the hot hand phenomenon. *Journal of Mathematical Psychology*, *72*, 200–209.

West, M., Harrison, P. J., & Migon, H. S. (1985). Dynamic generalised linear models and bayesian forecasting (with discussion). *Journal of the American Statistical Association*, *80*, 73–97.