Beating Vegas: Creating a Dynamic Sports Betting Model

Daniel A. Levine

April 19, 2019

Introduction

Welcome to the *R Markdown* thesis template. This template is based on (and in many places copied directly from) the Reed College LaTeX template, but hopefully it will provide a nicer interface for those that have never used TeX or LaTeX before. Using *R Markdown* will also allow you to easily keep track of your analyses in **R** chunks of code, with the resulting plots and output included as well. The hope is this *R Markdown* template gets you in the habit of doing reproducible research, which benefits you long-term as a researcher, but also will greatly help anyone that is trying to reproduce or build onto your results down the road.

Hopefully, you won’t have much of a learning period to go through and you will reap the benefits of a nicely formatted thesis. The use of LaTeX in combination with *Markdown* is more consistent than the output of a word processor, much less prone to corruption or crashing, and the resulting file is smaller than a Word file. While you may have never had problems using Word in the past, your thesis is likely going to be about twice as large and complex as anything you’ve written before, taxing Word’s capabilities. After working with *Markdown* and **R** together for a few weeks, we are confident this will be your reporting style of choice going forward.

Why use it?

R Markdown creates a simple and straightforward way to interface with the beauty of LaTeX. Packages have been written in **R** to work directly with LaTeX to produce nicely formatting tables and paragraphs. In addition to creating a user friendly interface to LaTeX, *R Markdown* also allows you to read in your data, to analyze it and to visualize it using **R** functions, and also to provide the documentation and commentary on the results of your project. Further, it allows for **R** results to be passed inline to the commentary of your results. You’ll see more on this later.

Having your code and commentary all together in one place has a plethora of benefits!

Who should use it?

Anyone who needs to use data analysis, math, tables, a lot of figures, complex cross-references, or who just cares about the final appearance of their document should use *R Markdown*. Of particular use should be anyone in the sciences, but the user-friendly nature of *Markdown* and its ability to keep track of and easily include figures, automatically generate a table of contents, index, references, table of figures, etc. should make it of great benefit to nearly anyone writing a thesis project.

Chapter 1 Introduction

Leading up to each NFL game, “bookmakers”, or casino staff, place a “spread” on which casino patrons can bet. The bookmakers create a spread for the game because most NFL games feature unbalanced teams, and the result of the game is not in much question; however, the amount by which the superior team will win creates a more fair proposition. The spread is the amount by which the bookmakers think the superior team will win. For example, the spread for the Super Bowl was the New England Patriots (-2.0) versus the Los Angeles Rams. This means the bookmakers expected the Patriots to win by 2 points. For the Patriots to beat the spread, they needed to win the game by greater than 2 points – which they did, as they won the game 13 - 3.

For the casino, the goal of creating this supposed fair value proposition is to give its patrons the opportunity to bet on the result of an NFL game with what should be a 50/50 proposition. However, the casino uses unfair odds to create an edge for itself. For a bet against the spread, a bettor must place 11 units in order to win 10 units.

This means that if there is an equal amount of money on both teams, the casino wins money. For example, if both teams have 11 units placed on them to beat the spread, the casino is guaranteed to make money. This is because one team will beat the spread and win 10 units for its bettor, while the other team will fail to cover the spread and instead lose 11 units for its bettor. Thus, for 22 units bet on the game, the casino is guaranteed to win 1 unit. So, their return is a guaranteed 122=4.5%122=4.5%. In most cases, the casino looks to place the spread at a point that will generate equal amount of money on both sides – not the true number of points by which they think a team will win. As a result, there is value in this market in finding the instances where the true result differs greatly from the spread. These points of value often come from betting against popular or trendy picks, as the market (or the bettors) tend to overreact to recent performance, as well as big-name players. If there are unequal amounts of money on each side leading up to the game, the casino will adjust the spread throughout the week. This means there are certain points in the week where it is more advantageous to bet on a certain team.

In addition, there are also times when the casino fails to move the spread even with unequal amounts of money on both sides. This means that the casino is essentially placing a bet that the side the public bets more on will lose. The cliche states, “the casino always wins”. It is important to identify the times when a casino is placing a bet in order to bet on the same side as the casino.

The goal of this project is to create a betting model that provides a statistical basis for choosing the timing, team and amount to bet on a certain game. To do this, I first create a model to forecast the spread throughout the week, as to determine when exactly in the week is the most advantageous point to bet. Treating the spread as a time series object is a good method to achieve this goal. Second, I create a model that can predict the difference in score between the two teams playing and provide a probability point estimate for each team “beating the spread”. Using this probability and the fact that a bettor must bet 11 units to win 10 units, I generate an expected value for betting on each team. In order to have a positive expected value to bet on a game, the probability point estimate must be greater than 52.8%. In addition, the forecasted spreads helps determine whether it would be more advantageous to wait to bet on the game. Finally, after generating expected value for all the games, I simulate how my models perform using a variety of different betting strategies. I examine the distribution of winnings for each betting strategy after running the simulations numerous times. Each betting strategy has different rules and parameters that determine the stake and timing of each bet. One key note is that there will not be a bet placed on every game, as if the model predicts a 50% probability of a team beating the spread, the expected value for betting on this game for either team is negative. Thus, it is not always advantageous to bet on the games.

The remainder of this thesis is organized as followed: in Section 2, I discuss how I gathered my data and the techniques I used for organizing these data into a usable format. In Section 3, I discuss my two different modeling techniques – starting first with the model to forecast the spread throughout the week before moving into the to predict the score of the game. Section 4 discusses my nine different betting strategies and evaluates the distribution of outcomes resulting from these betting strategies. Finally, Section 5 wraps up this thesis with a discussion of the process, feasibility and next steps for this project.

Chapter 2 Data

There were multiple data sources needed for this project. First, to create the dynamic linear model to forecast the spread, I gathered data on all the spread movements throughout the week leading up to the game for as many NFL games as possible. Through web scraping from the https://pregame.com/game-center website, I was able to gather these data on all NFL games from the past two seasons. These data needed a significant amount of manipulating and cleaning to be put in a usable format. Through using “stingr” manipulations, each game contained a data frame of approximately 100-200 observations of the variables listed in Table 2.1.

Table 2.1: Betting Statistics

Statistic	Description
Time and Date	The time and date of the observation; The first observation nearly always occurred after bot teams had finished playing their previous game, so usually the first observation was Sunday evening one week prior to the game with the final observation seconds before game time (usually the following Sunday)
Spread	The spread for the away team

Table 2.1: Betting Statistics

Statistic	Description
Away Cash Percentage	The percent of the money bet on the game that is bet on the away team
Away Cash Bet	The amount of money that is bet on the away team
Away Ticket Percentage	The percent of bets on this game that are on the away team
Away Ticket Number	The number of bets on this game that are on the away team

With these data, it is easy to calculate the same dataset for the home team through the simple formulas listed Equations (2.1) — (2.4)

$$\text{Home Cash Bet}=\frac{\text{Away Cash Bet}}{\text{Away Cash Percentage}}-\text{Away Cash Bet} \tag{2.1}$$

$$\text{Home Cash Percentage}=\frac{\text{Home Cash Bet}}{\text{Home Cash Bet}+\text{Away Cash Bet}} \tag{2.2}$$

$$\text{Home Ticket Number}=\frac{\text{Away Ticket Number}}{\text{Away Ticket Percentage}}-\text{Away Ticket Number} \tag{2.3}$$

$$\text{Home Ticket Percentage}=\frac{\text{Home Ticket Number}}{\text{Home Ticket Number}+\text{Away Ticket Number}} \tag{2.4}$$

(2.1)Home Cash Bet=Away Cash BetAway Cash Percentage−Away Cash Bet(2.2)Home Cash Percentage=Home Cash BetHome Cash Bet+Away Cash Bet(2.3)Home Ticket Number=Away Ticket NumberAway Ticket Percentage−Away Ticket Number(2.4)Home Ticket Percentage=Home Ticket NumberHome Ticket Number+Away Ticket Number

These variables are important because they influence the spread. This is because a casino wants to manipulate the spread so the percent of money on each team is 50%, as this will generate 4.5% of money for the casino, guaranteed. However, other times, the casino is essentially gambling by allowing for uneven money percentages. They take a position in a certain outcome that, according to their models, can raise their expected value.

The timing for each of the data points from these series are irregular. To start, each week, one game is played on Thursday night, one on Monday night, with the rest of the games played on Sunday. The series starts when the casinos first open the game for betting. This is usually occurs the Sunday one week prior to the start of the game. But, since not all games are played on Sunday, some games are open for betting for shorter or longer periods of time. In addition, the casinos open up the Week 1 games for betting weeks in advance. This is the first irregularity that causes for different length series'. This is why the shortest series has only 57 data points while the longest series has 265 data points. However, the 25% - 75% of data points is 130 to 170, and nearly all series fall between 100 - 200 data points.

Within each series, the data is not captured in standardized time intervals, but instead, each data point is captured when there is a shift in the percentage of money or the percentage of tickets that is bet on each team. Some data points can be spaced minutes apart while others can be spaced out 12 hours apart. It is the timing of bets that trigger a data point. For modeling purposes, I treat these irregular intervals as evenly timed data points. When I forecast the spread, I forecast for hh number of future points with hh chosen through a separate model used to predict how many more data points this series based on the week in the NFL season, the hours until the game starts, the number of bets and amount of cash bet based on the game. This is further discussed in the modeling section.
The next dataset is a CSV that is updated weekly that contains information, such as the teams, week, year and location, about all NFL games dating back to 2006, in addition to the opening and closing spreads for these games. This aspect is useful for cross-referencing. But, more importantly, this dataset contains the results of all the games, in addition to a list of all games. This list can be iterated through for all 414 separate DLMs.

Finally, in order to find team-level statistics, I needed to find separate data sets for each NFL season. This is because leading up to a Week 5, 2018 matchup, the only information that bettors have is all the season (and all previous seasons) data leading up to Week 5 in 2018. Football Outsiders has webpages with week by week statistics for a certain type of statistic. This is called Defense-adjusted Value Over Average (DVOA). DVOA measures a team’s efficiency by comparing success on every single play to a league average based on situation and opponent. In addition, the “Weighted DVOA” is another metric provided, and this statistic weights the team’s DVOA with a preseason projection that the website, Football Outsider, created. This is because after 1 week, a team’s DVOA will be very extreme, but weighting it with a projection ensures that the metric will not overreact to an extremely limited sample size. This is essentially similar to putting a prior on DVOA and updating the posterior with the data from the games played. The scale for these statistics is a percentage, and this indicates the percent above or below average that a team is. Table 2.2 describes all these statistics.

Table 2.2: Team-specific Statistics

Statistic	**Explanation**
Total DVOA	Measures a team’s efficiency by comparing success on every single play to a league ave based on situation and opponent
Weighted DVOA	Weights the DVOA with a preseason projection
Offense DVOA	Measures a team’s offensive efficiency
Defense DVOA	Measures a team’s defensive efficiency
Special Teams DVOA	Measures a team’s efficiency on Special Teams plays (field goals, punts, kickoffs)

Table 2.2: Team-specific Statistics

Statistic	Explanation
Record	A teams record of their wins, losses and ties

I needed to merge and match these data sets. I did so by matching the week and year of each game to the correct dataset for the statistics, and then matching the team to their statistics up to that certain week. This completed data set is used for modeling. Section @ref{appen1} of the Appendix shows one line of this data set.

Chapter 3 Modeling

I have two separate aspects to modeling: forecasting the point spread throughout the week and predicting the score of the game. For both processes, I tried different approaches to modeling and chose the best performing model based on performance on test datasets.

3.1 Point Spread Forecasting Model

I forecasted the point spread throughout the week by treating this object as a time-series. I explored the data with an aim to find the best approach to modeling, before then moving into the modeling procedure. The best performing model was a time-varying Bayesian Dynamic Linear Regression model that used ARIMA (Autoregressive Integrated Moving Average) methods to forecast the time-varying parameters that are used to forecast the point spread in the Dynamic Linear Regression Model. In addition, for utilizing the model, I needed to determine how many data points will be in the series. I used a mixed linear regression model for this purpose.

3.1.1 Exploratory Data Analysis

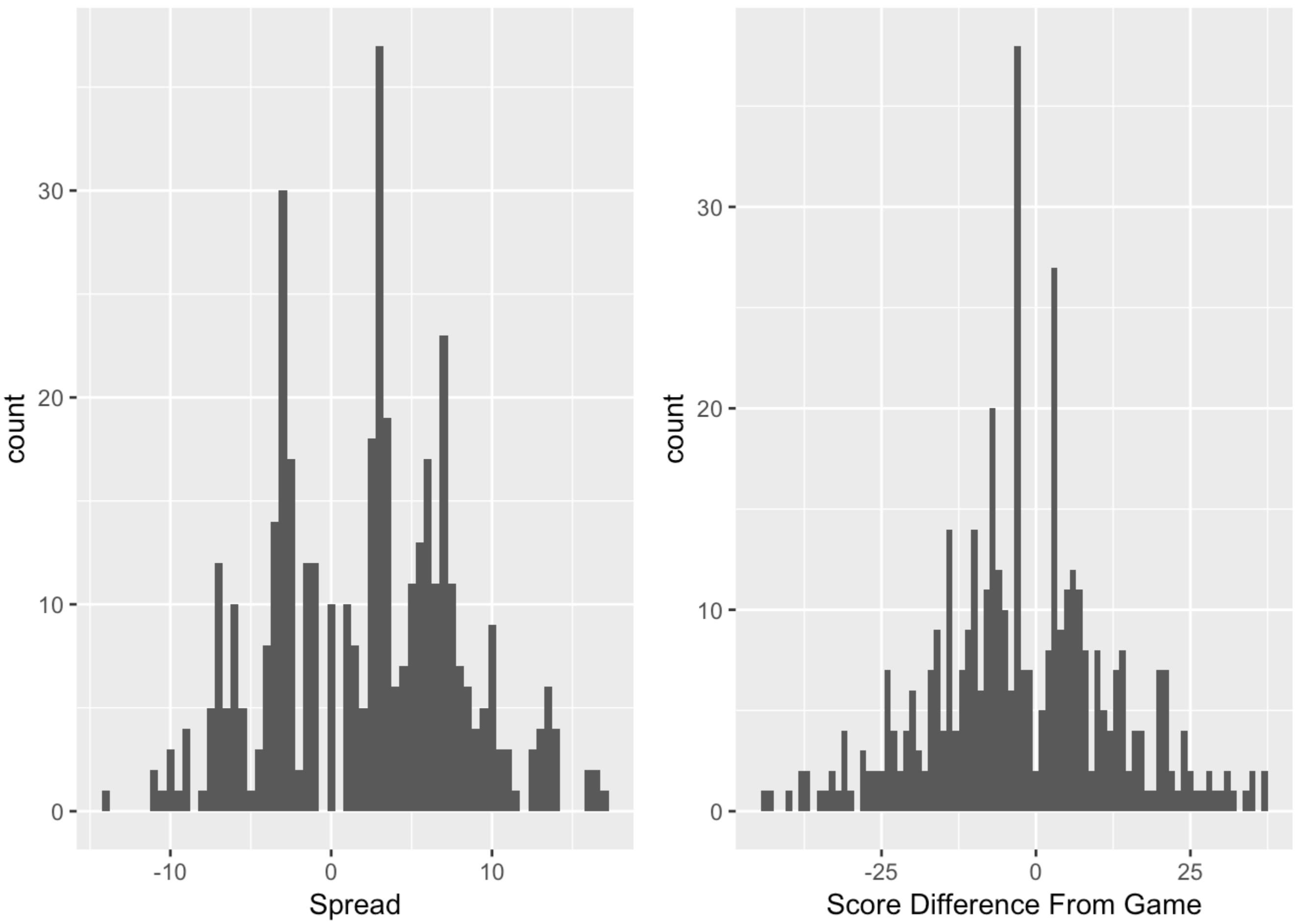

Figure 3.1: Histograms of Point Spreads and Score Differences from Games

The first aspect to examine when forecasting the spreads is the distribution of spreads. It is also important to look at the distribution of game outcomes that these spreads model. Figure 3.1 shows both of these distributions. Both have multiple peaks. These multiple peaks arise because in football, nearly all scores are worth 33 or 77 points. When predicting the difference between two teams, many games will end up with a forecasted spread near these key numbers, and the results of these games will fall at these numbers often. In addition, there are a few dead zones – mainly in between 00 and 33. The results of the games mirror the distribution of the forecast spreads, however, with a much wider distribution. It is difficult to forecast a blowout game, but they do occur, which is why there are much longer tails for the true score differences.

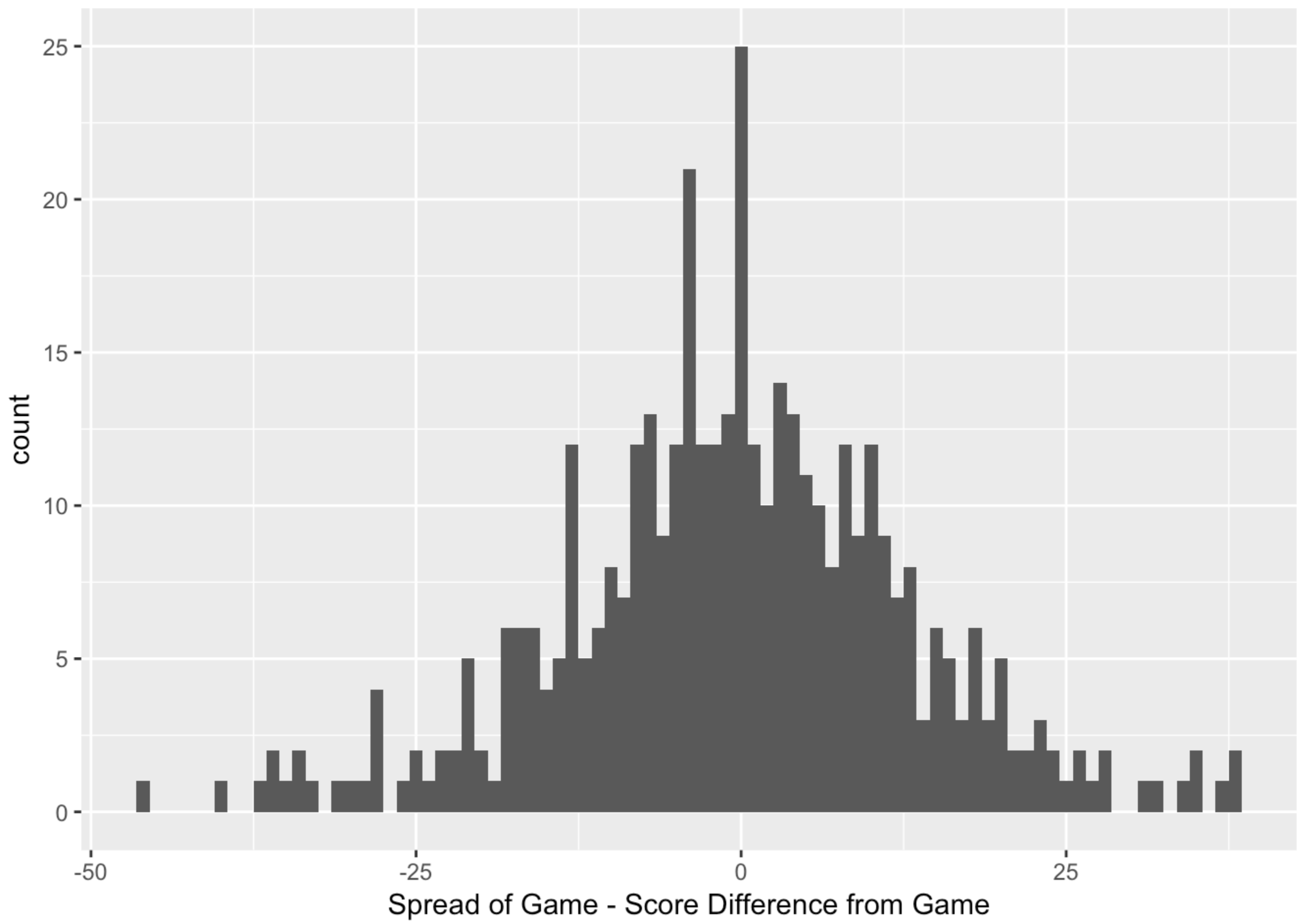

Figure 3.2: Histogram of Game Results against Spread

Figure 3.2 shows a distribution of the result of the game against the spread. A result of 00 would indicate that the game ended with the same result as the spread, and the result of the game would be a push, meaning that nobody wins and the bettor's stake is returned to the better. To demonstrate the accuracy of the bookmaker's, it is evident that the distribution is relatively normally distributed around 00, with a second peak at −3−3 indicating that many of the games resulted in the home team beating the spread by 33 points.

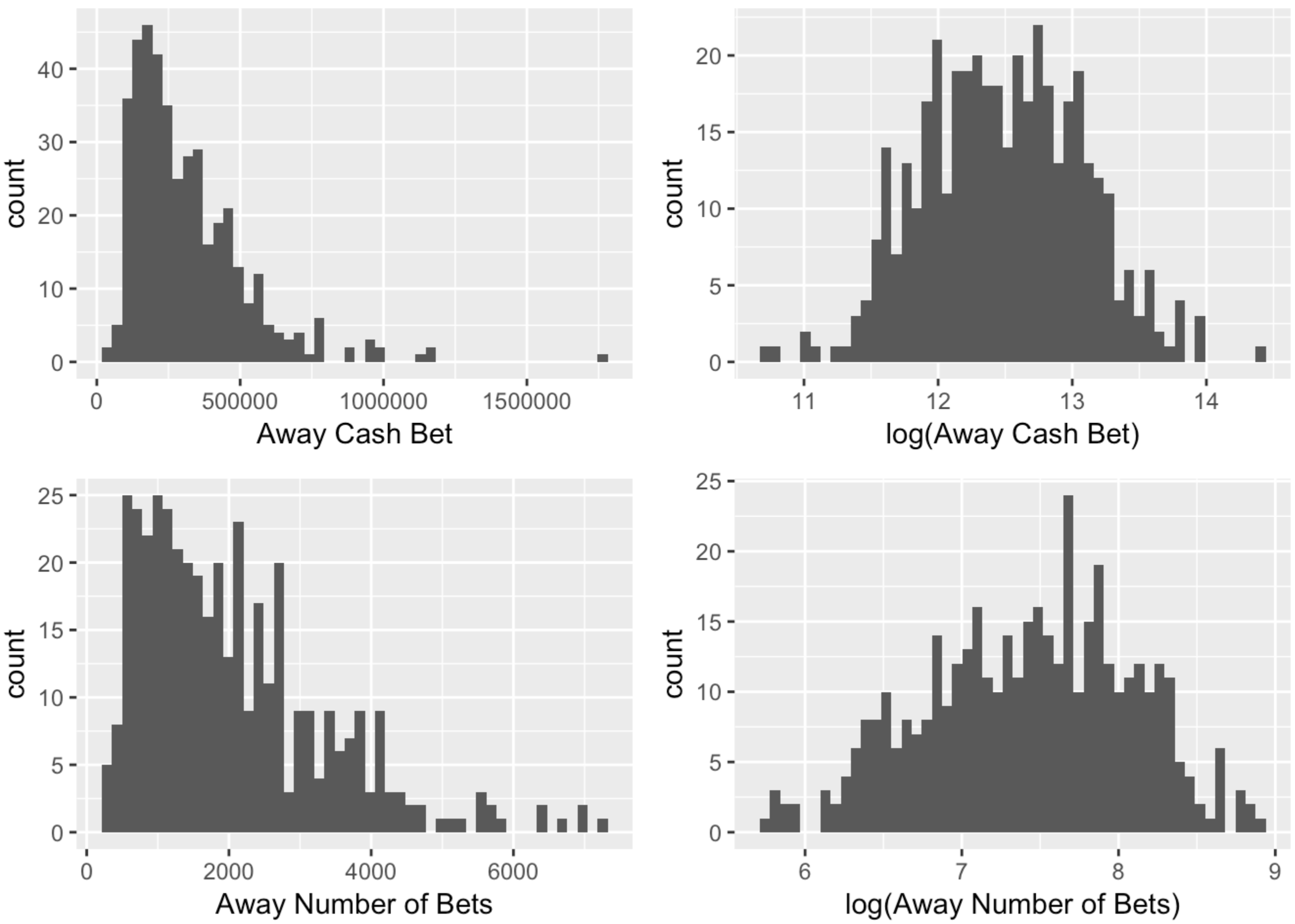

Figure 3.3: Transformations of Key Betting-Statistic Variables

Research[1] suggests that casinos adjust the line based on the amount of cash bet on each side, so that they can even out the amount of money bet on each time and guarantee themselves a return. Examining the cash variables can help evaluate this research. Figure 3.3 demonstrates the skewness of the cash and ticket number variables, as well as updated distributions after transformations. The cash variable is very right-skewed. For modeling and interpretability, it is integral to transform this variable into the log of the cash bet. The number of bets on each side is also right skewed. The $\log(\text{Away Cash Bet})$log(Away Cash Bet) and $\log(\text{Away Number of Bets})$log(Away Number of Bets) are both significantly closer to normally distributed.

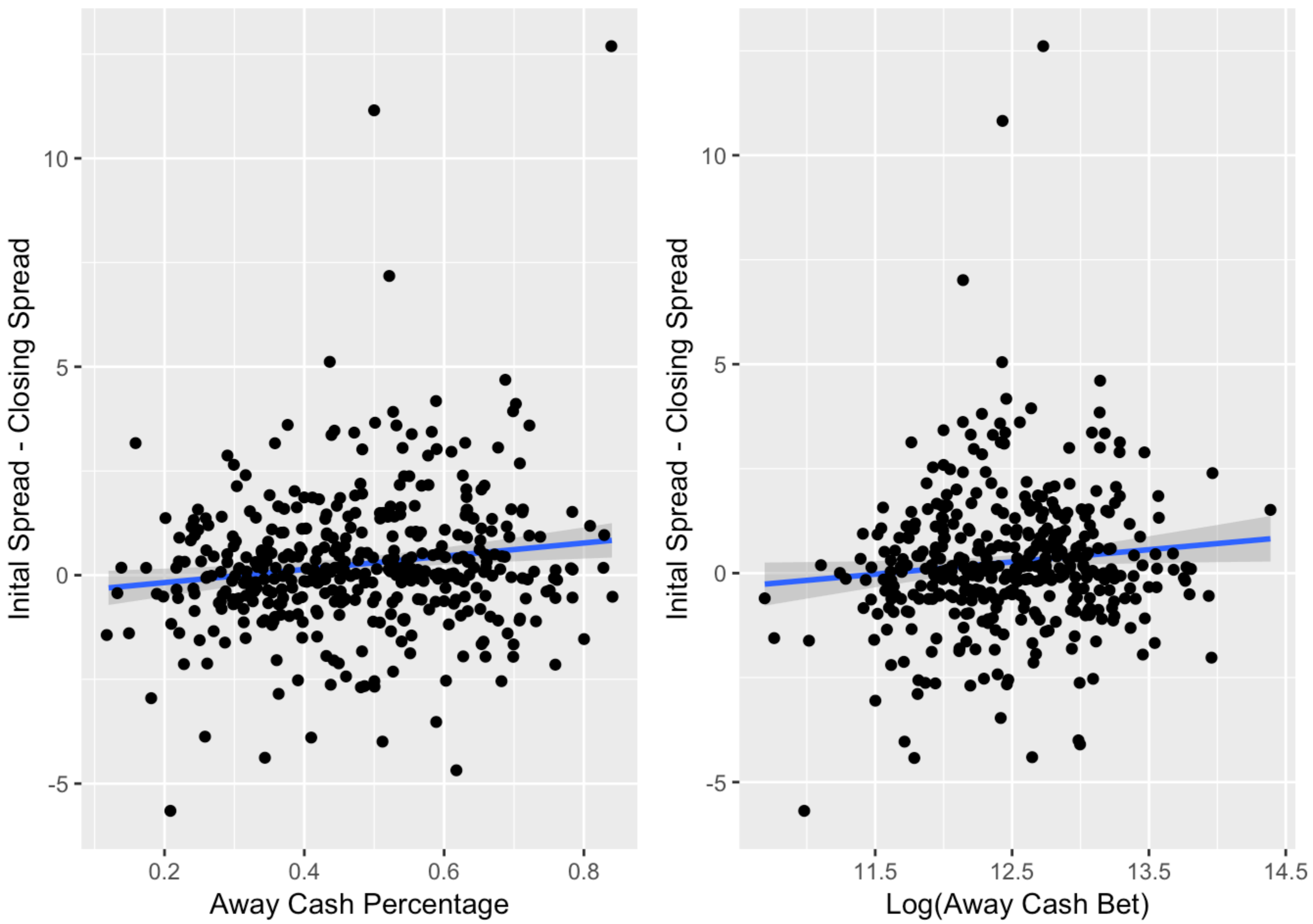

Figure 3.4: Line Difference versus Key Variables

Figure 3.4 shows the the line difference from when the casino first listed the spread to the spread when the game started compared with the cash percentage and the $\log(\text{Away Cash Bet})$log(Away Cash Bet). Research[2] suggests that with more cash bet on the away team, the casino would want to make the spread less favorable for the away team, in an effort to get more money placed on the home team and achieve a 50/50 split.
Here, while the effect is not major, for both the away cash percentage and away cash amount variables, as they increase, the line difference for the away tends to become more negative. This means that when more cash is on the away team, the spread tends to become more favorable to the home team. For example, a line difference of -2 means that the initial spread could have been the away team is favored by 6 points (-6), but then the spread moved to make the away team favored by 8 points (-8). The away team must now win by more than 8 points to cover the spread, opposed to the previous point where the away team only needed to win by more than 6 points.

There are a few outliers where the line difference is greater than 5 points. This sort of extreme movement only can occur due to big player news. For example, if there is news on the Friday leading up to the game that Tom Brady is injured and cannot play, this would cause a massive swing in the line that would not be related to the cash and ticket percentages.

3.1.2 Modeling the Point Spread

After exploratory data analysis, the next step is finding the best model to forecast the future spread. The model needs to forecast what the spread will be from a certain decision point. This first decision point is the first point when bets will be placed. The chosen decision point is after two-thirds of the observations in each time-series. The data frame containing the observations for each game is cut off at the two-thirds mark, and the model then forecasts the point spread for the final one-third of observations, using only the information up to this two-thirds point. I consider a Bayesian and frequentist approach to modeling the point spread. After forecasting the point spread for the final one-third observations, I calculate the error for each model by finding the difference between the forecasted point spread and true point spread for each observation. I use the forecasts from the model with the lowest average error across all my time-series’ in my betting strategies.

The Bayesian approach to modeling is a time-series random walk plus noise regression model. The process starts by placing a prior for the parameters in my model before updating these parameters with the posterior mean through finding the MLE of the parameters of this regression model. The regressors in the model are the $\log(\text{Away Cash Bet})$, Away Number of Bets (Away Ticket Number), $\log(\text{Home Cash Bet})$ and Home Number of Bets (Home Ticket Number).
The full process of creating the dynamic linear model is demonstrated through the example of the Week 2, 2018 game between the Minnesota Vikings and the Green Bay Packers.

Equations (3.1) and (3.2) express a dynamic linear regression model with time-varying parameters.

$$y_t = X_t' \theta_t + v_t \quad v_t \sim N(0, V_t); \text{ (observation equation)} \tag{3.1}$$
$$\theta_t = G_t \theta_{t-1} + \omega_t \quad \omega_t \sim N(0, W_t); \text{ (evolution equation).} \tag{3.2}$$

The vector of observations up to time t is $y_t=(y_1,\dots,y_t)$. The observation equation (Equation **??**) describes the vector of observations y_t (the spread at time t) through its State vector θ_t (the predictor variables at time t) and the vector of noise from the observations v_t. The evolution equation (Equation **??**) describes the evolution of the state vector over time with a Markov structure. θ_t is the state vector of the time-varying regression parameters (of number p); $\theta_t=(\alpha_t\ ;\beta_t)'$ with dimension $p\times 1$. α_t and β_t are the regression coefficients X_t' is the row vector of covariates at time t of dimension $1\times p$. w_t is the variance of the state-space vectors. G_t is an

evolution matrix of $p \times p$ dimension. This is the evolution matrix because it allows for the evolution of the state space vector by matching up the dimensions the parameters. G_t is typically, and in this model, an identity matrix.
This is the general setup for a dynamic linear regression model.
Equations (3.3) — (3.5) show the expansion of equation (3.1).

$$\begin{aligned} y_t &= \alpha_t + \beta_t X_t' + v_t & v_t &\sim N(0, V_t) & (3.3)\\ \alpha_t &= \alpha_{t-1} + \epsilon_{t\alpha} & \epsilon_{t\alpha} &\sim N(0, \sigma^2_\alpha) & (3.4)\\ \beta_t &= \beta_{t-1} + \epsilon_{t\beta} & \epsilon_{t\beta} &\sim N(0, \sigma^2_\beta) & (3.5) \end{aligned}$$

There are three parameters that need to be set, and that is the variance of the observations V_t, and then the variances of the regression coefficients for the state-space vector – σ^2_α and σ^2_β.
This can be done through a Bayesian method, where the initial parameter start values are set, and then through finding the MLE of the DLM using the *dlmMLE*, these parameters are updated with the posterior mean. I used sample observational variance of the spread up to the first decision point as the starting value of the observational variance V. I used a flat prior for the variances of the regression parameters have a flat prior. Table 3.1 shows the values for the prior and posterior means of the variance parameters.

Table 3.1: Prior and Posterior Values for Variance Parameters

	Prior Parameters	Posterior Paran
V	0.5261619	0.00
σ^2_α	0.0000000	0.09
σ^2_β	0.0000000	0.00

The posterior mean for the σ^2_α and σ^2_β values are used diagonally in the ω_t matrix. Looking back at equations (3.1) and (3.2), θ_t for each observation is found through using α_t and β_t values, which are drawn through σ^2_α and σ^2_β. The values of the design vector X_t' comes directly from the predictors and the variance for V is set. Thus, all the parameters needed for modeling are set, and I use a dynamic linear regression model through the function *dlmModReg* to calculate my values for the observational values (y_t) and the state-space parameters (θ_t). This is done through the filtering method.
The filtering distribution takes in the DLM, and returns a series of one-step forecasts for the observations. These one-step forecasts are created from filtering all the information up to time t. The first step of the filtering distribution has a starting

value $\theta_0 \sim N(m_0,C_0)$θ0∼N(m0,C0). m_0m0 and C_0C0 are the pre-sample means and variances for θθ.

1. $\theta_0 \sim N(m_0,C_0)$θ0∼N(m0,C0).
2. One-step forecast for the state:$\theta_t | y_{t-1} \sim N(a_t,R_t)$θt|yt−1∼N(at,Rt)

   for $a_t=G_t \cdot m_{t-1}$at=Gt·mt−1 and $R_t=G_t \cdot C_{t-1} \cdot G_t'+W_t$Rt=Gt·Ct−1·Gt′+Wt.

3. One-step forecast for the observation:$y_t | y_{t-1} \sim N(f_t,Q_t)$yt|yt−1∼N(ft,Qt)

   for $f_t=F_t \cdot a_t$ft=Ft·at and $Q_t=F_t \cdot R_{t-1} \cdot F_t'+V_t$Qt=Ft·Rt−1·Ft′+Vt.

Creating a filtered distribution with the *dlmFilter* function returns a series of one-step forecasts and variances for the observations, as well as the same information for the state-space vector.

For a time-invariant dynamic linear model, there would be no extra work for finding a forecast for the observations after a given point tt. But, for a time-varying model, such as this, the X_t'Xt′ values are also unknown past the given point tt. The Kalman filtering method extends the time-series with new future predictor values, but does not input future values for the observational values. Once the future predictor values are entered, I create a filtered distribution with this new set – using the filtered values of the extended observational values as my forecast.

There are a few common methods for finding new methods for the predictor values, such as inputting the last known observation, the mean or the median. However, since my predictor values continue to grow, these methods do not apply to this model. So, at the decision point, I fit ARIMA models for each of my new predictor values. I used the *auto ARIMA* method to generate these new values for each of my predictor variables. Using the ARIMA method is a frequentist approach to a time-series forecast. I used this approach because for two reasons: it is unrealistic to build a separate Bayesian DLM for each parameter and these parameters simply grow without fluctuation (unlike the point spread), so it is not as necessary to build as complex of a model. There are three parameters that go into that ARIMA method: *p* is the number of lag observations in the model, *d* is the degree of differencing and *q* is the order of the moving average.

The *auto.arima* function automatically chooses the best *p, d* and *q* values that will minimize the AIC and BIC of the model. However, by setting the seasonal parameter to “false”, I ensured that no model that incorporated a seasonal trend is chosen because that would not fit these data. Figure 3.5 is the forecasted number of tickets versus the true number of tickets for the Green Bay Packers versus Minnesota Vikings game. While this forecast is certainly not perfect, it generally follows a similar path to the true value. This is certainly an imperfect method and one area for improvement in this facet of the model.

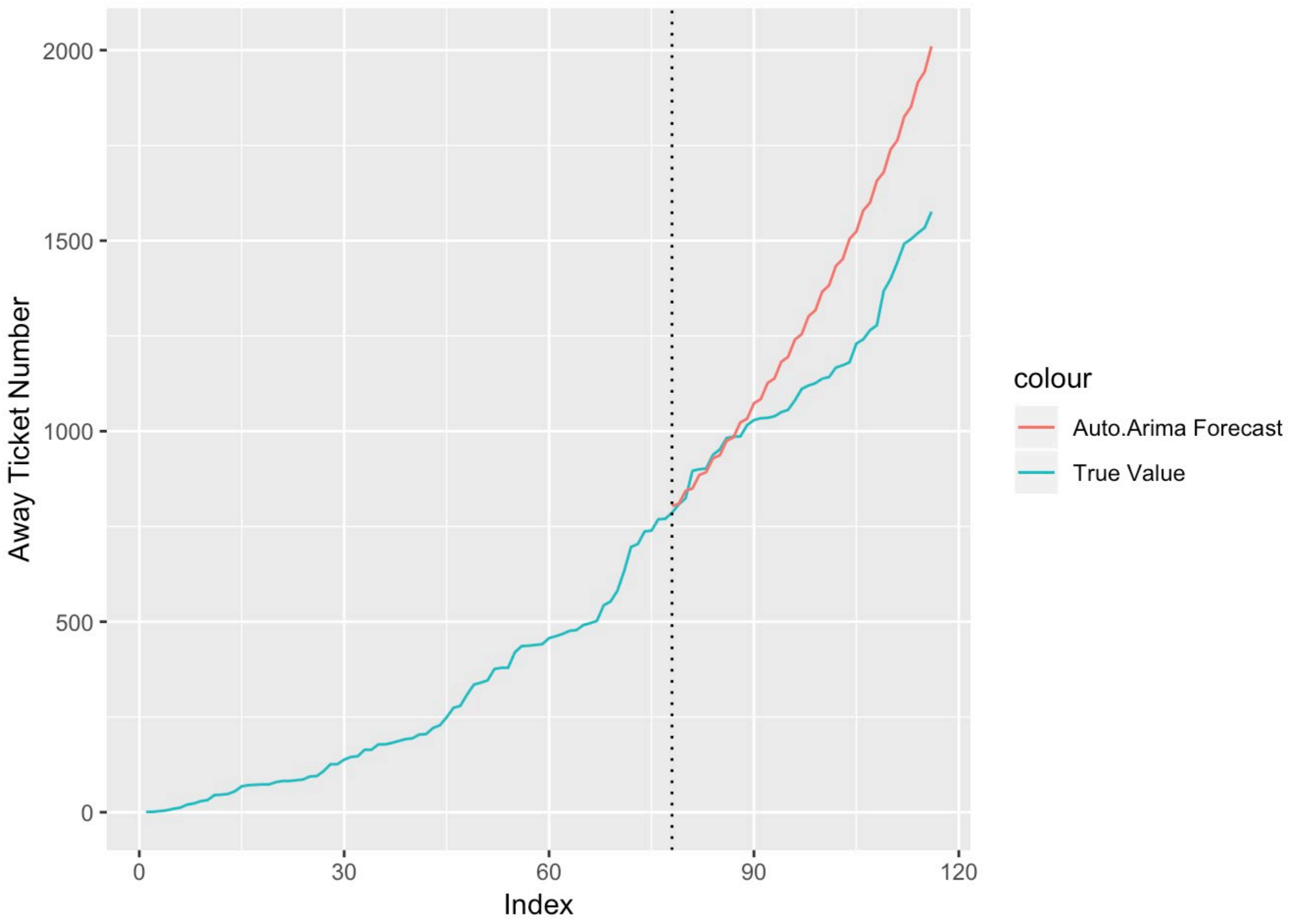

Figure 3.5: Forecasted versus True Away Ticket Number

This forecast model for the number of tickets is an ARIMA(1, 2, 2) model that is expressed in Equations (3.6) and (3.7).

$$\hat{Y}_t = \hat{y}_t + 2Y_{t-1} - Y_{t-2} \quad (3.6)$$

$$\hat{y}_t = \mu + AR1 \cdot y_{t-1} - MA1 \cdot e_{t-1} - MA2 \cdot e_{t-2} \quad (3.7)$$

(3.6)Yt^=yt^+2Yt−1−Yt−2(3.7)yt^=μ+AR1·yt−1−MA1·et−1−MA2·et−2

Table 3.2 displays the coefficients to the ARIMA(1, 2, 2) model.

Table 3.2: Coefficients of ARIMA(1, 2, 2) Model for Away Ticket Number

	Coef
AR1	-0.98
MA1	0.17

Table 3.2: Coefficients of ARIMA(1, 2, 2) Model for Away Ticket Number

	Coef
MA2	-0.42

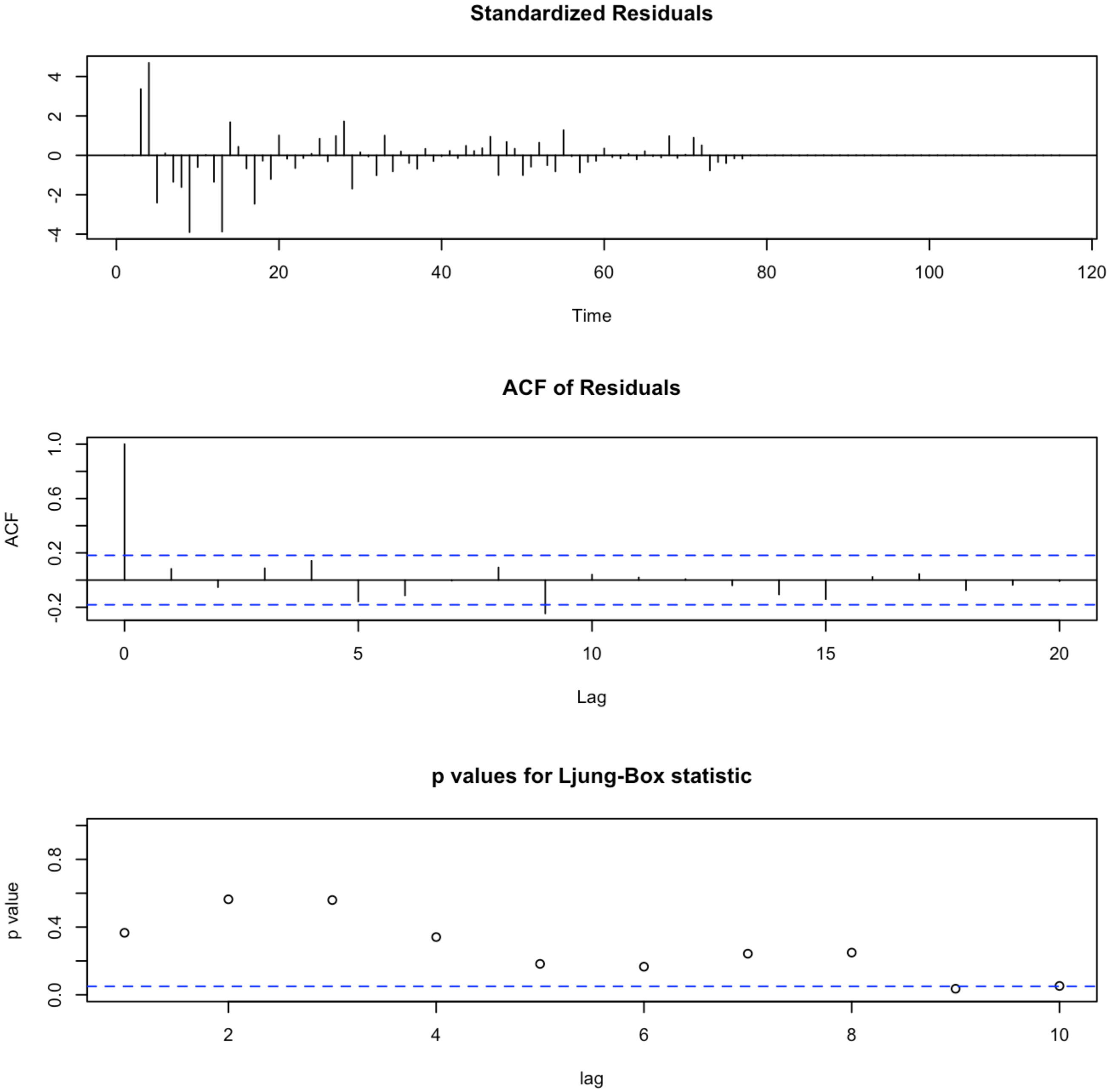

Figure 3.6: Diagnostic Plots for ARIMA(1, 2, 2) Model for Away Ticket Number

Figure 3.6 is the diagnostic plots for the *auto.arima* method for forecasting the number of tickets for the away team. The plots show that this model is a pretty good fit for the data, as the standardized residuals generally look like white noise, though the p-values for autocorrelation become significant when the lag factor reaches high values such as 9. As these models are automatically fitted to best describe the data at hand, they generally fit the data pretty well.

It is important to note that the automatic ARIMA is fit for each different new variable from each time-series (opposed to using the same ARIMA model for the cash bet for all series) because the trends are not the same across all series. While bookmakers generally look to obtain 50/50 amount of cash on each game, this is certainly not always the case, as bookmakers will take a position on many of the games. Thus, the automatic ARIMA model will fit the model best to the data for each of the predictor variables.

Finally, after generating new values for the predictor variables in my DLM, the Kalman filtering method can be used to find predictions for the spread. This method follows the exact same approach as above, however, the one-step forecasts for the last third of observations will replace the NAs.

In addition, for comparison, the spread is also modeled with the *auto.arima* forecast, using the same predictor variables as the Bayesian DLM as regressors. This is a frequentist approach for modeling each time-series. The accuracy of each approach is determined by looking at the average error in the predicted spread values versus the true spread values.

For this example game between the Green Bay Packers and the Minnesota Vikings, the *auto.arima* method fit an ARIMA(1, 0, 0) model, which is a first-order autoregressive model.

Equation (3.8) expresses this model.

$$Y_t = c + \phi_p Y_{t-1} + \epsilon_t \qquad \epsilon_t \sim N(0, \sigma^2_\epsilon) \tag{3.8}$$

c is the intercept or the constant in the equation and ϕ_p is the vector of coefficients for the autoregressive term (AR), as well as all the predictors. Table 3.3 shows the coefficients of this model and the variance parameter $\sigma^2_\epsilon = 0.00595$.

Table 3.3: Coeffecients of ARIMA(1, 0, 0) Model for the Point Spread

	Coeff
AR1	0.85
Intercept	-2.41

Table 3.3: Coeffecients of ARIMA(1, 0, 0) Model for the Point Spread

	Coeff
Log Away Cash	-0.01
Log Home Cash	0.00
Away Ticket Number	-0.05
Home Ticket Number	0.00

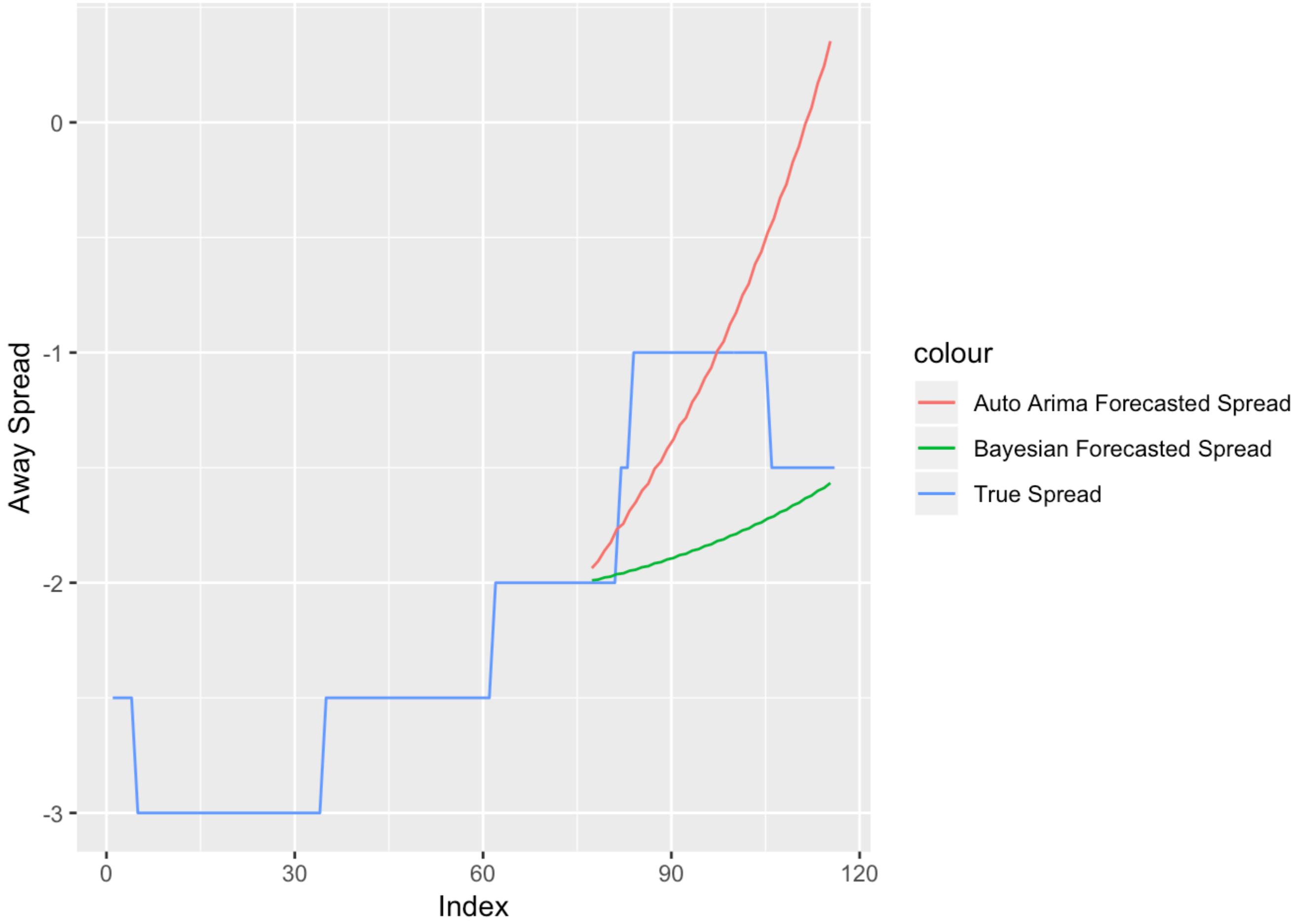

Figure 3.7: Spread versus Forecasts for Minnesota Vikings at Green Bay Packers Week 2, 2018

Figure 3.7 compares the Bayesian DLM and the frequentist ARIMA model's predictions with the true final spread values from the game between the Minnesota Vikings and the Green Bay Packers. The blue line represents the true spread, while the red and green lines represent the Bayesian and frequentist forecasts, respectively. Both forecasts correctly predict the spread to rise. However, the Bayesian approach does a better job, in this scenario, of being closer to the true spread values.

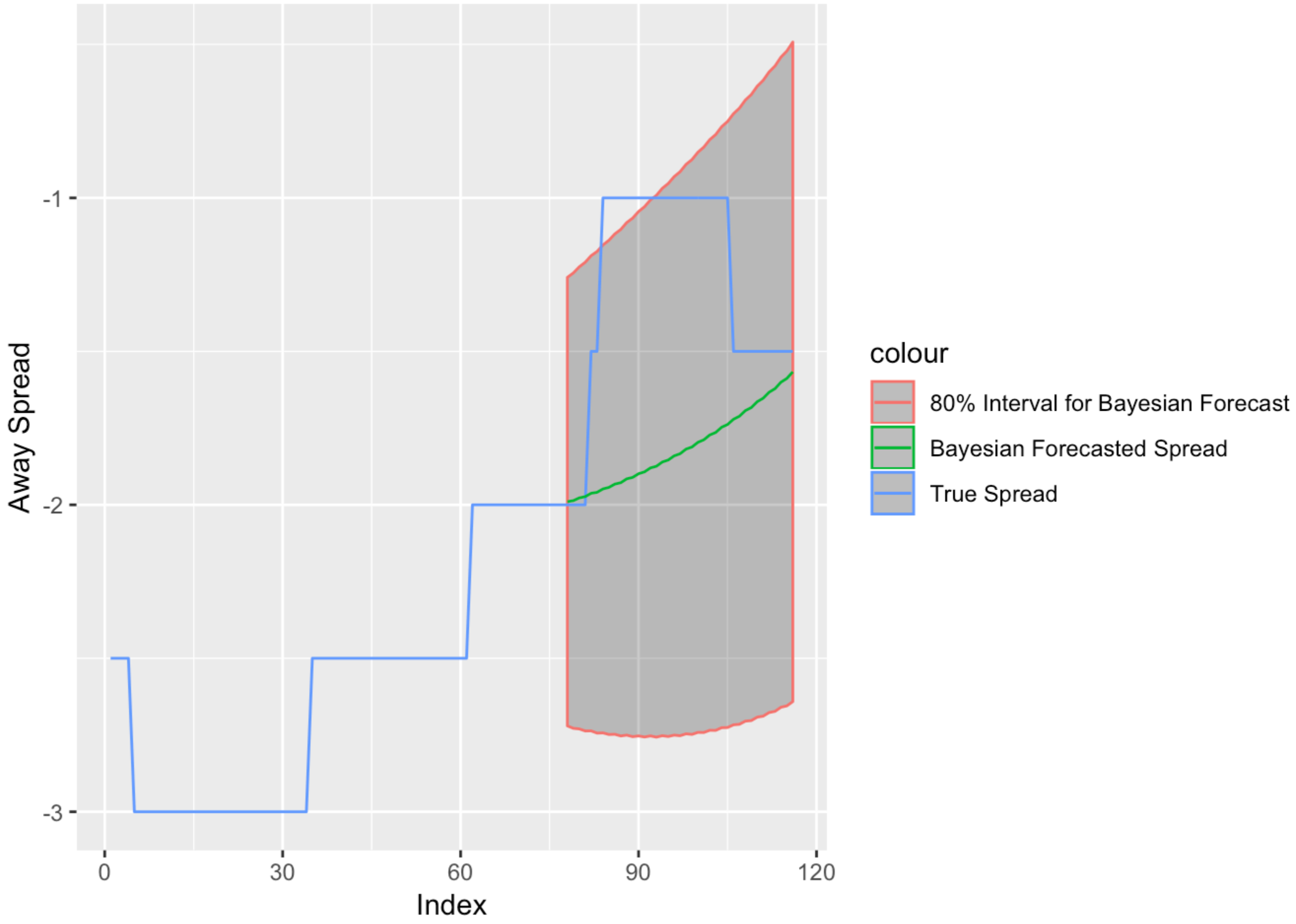

Figure 3.8: Spread versus Bayesian Forecast for Minnesota Vikings at Green Bay Packers Week 2, 2018 with 80% Confidence Interval

Figure 3.8 shows the Bayesian DLM forecast with a 80% confidence interval. I chose an 80% confidence based on trial and error. Here, while the spread at the decision point is within the 80% interval, there is a point when the spread reaches Vikings (-1) when the spread is out of the 80% interval. This will be a key distinction to make when it comes to betting strategies. Actually, this is why I chose an 80% confidence, as opposed to a more standard 95% confidence interval. With the wider 95% confidence interval, it is more rare for me to have a value outside of that interval. Since I make betting decisions based on whether the spread is within the selected interval, I use an interval that allows me to incorporate more instances of waiting to bet until the future spread moves to a more advantageous position. Also, while 95% confidence interval is more standard, the choice is as arbitrary as an 80% confidence interval.

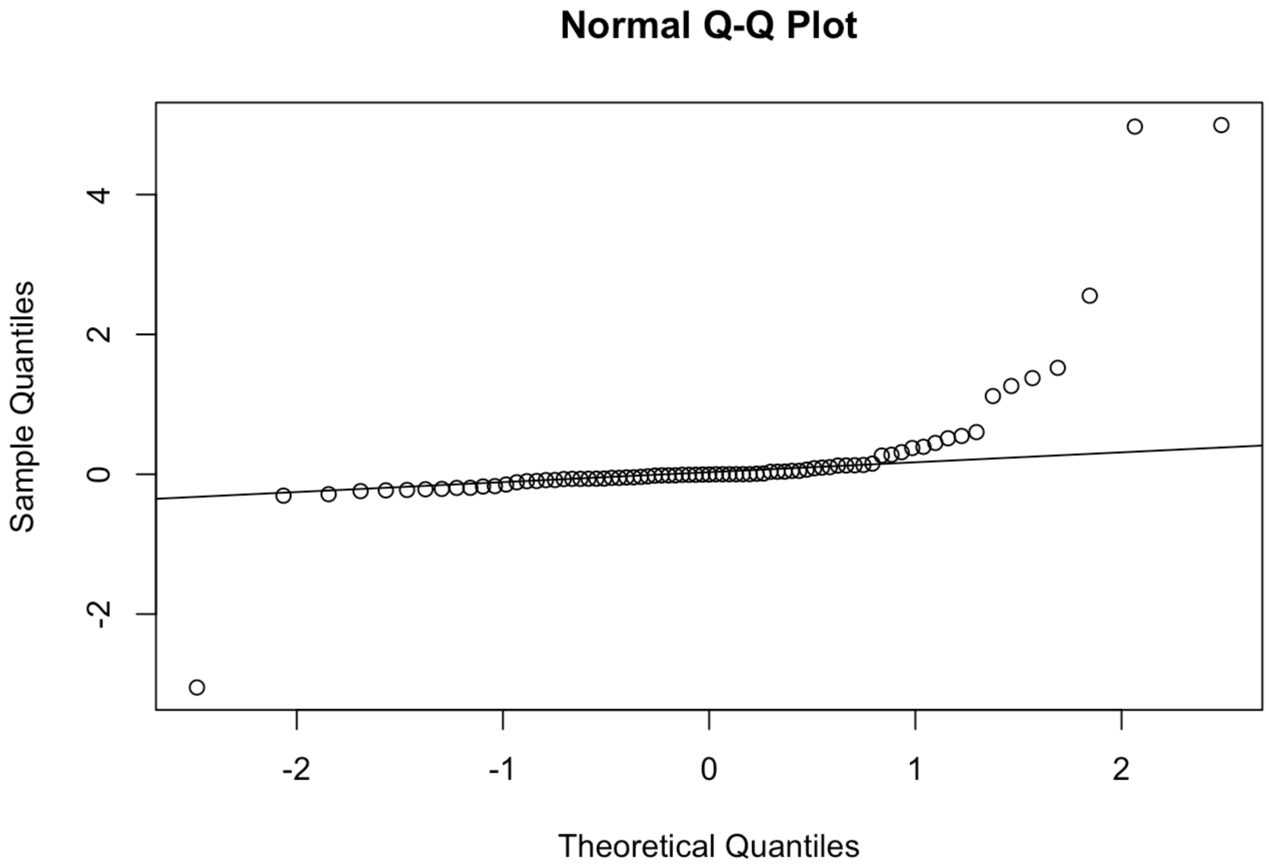

Figure 3.9: Normal QQ Plot for Residuals of the Forecasted Spread from the DLM

Figure 3.9 shows the residuals plot from the filtered distribution. The residuals do not seem to be completely normally distributed. This is due to the fact that the true spread can only move in increments of 0.5, which is a massive amount in terms of the jumps in the filtered values. When looking at the rounded values of the spread, however, the residuals are more likely to be normally distributed.

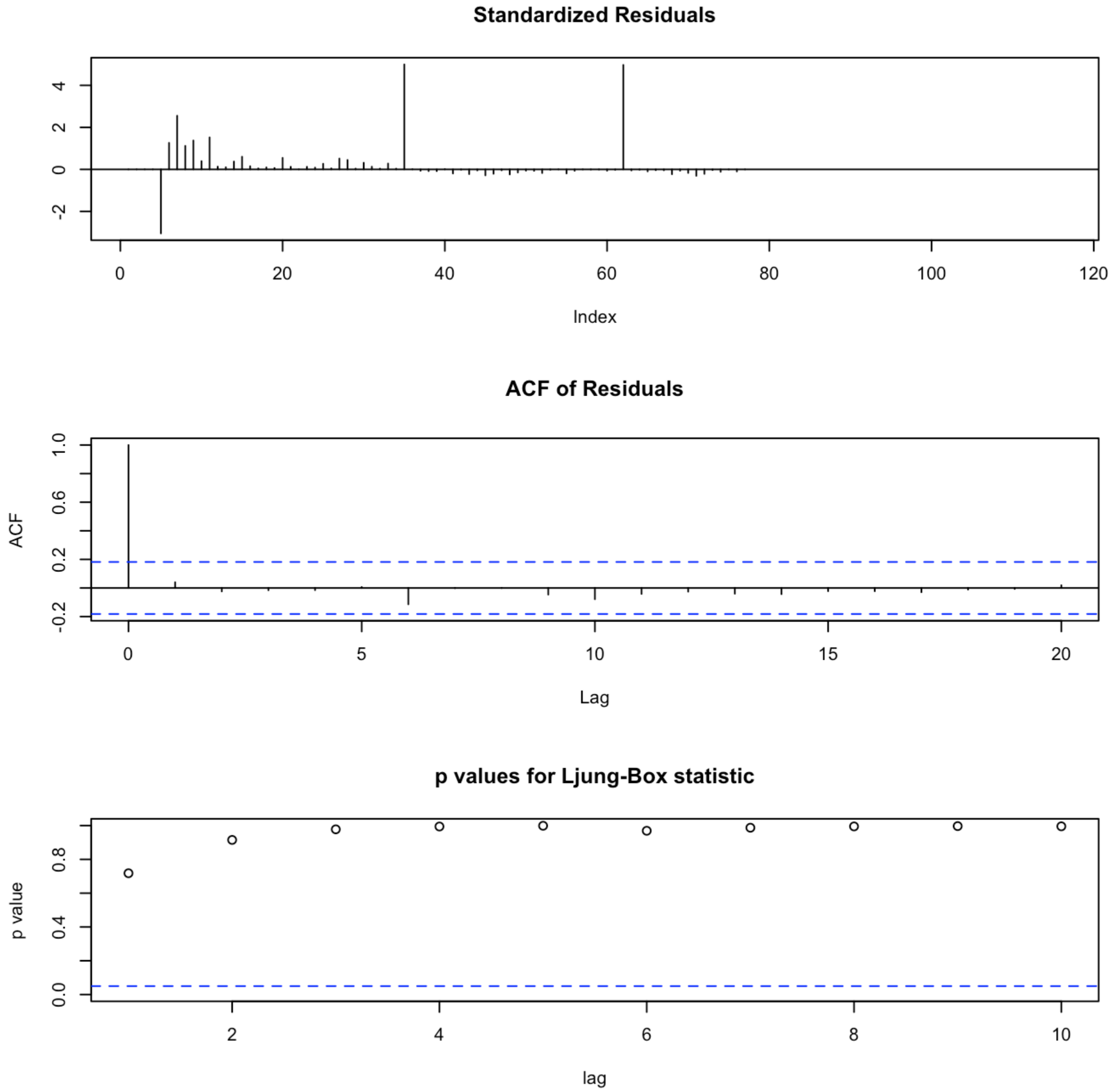

Figure 3.10: Diagnostic Plots for DLM of MIN at GB game

Figure 3.10 is the diagnostic plots for the Kalman-filtered model of the Minnesota Vikings at Green Bay Packers game. The p values for autocorrelation are all extremely high, indicating there is no autocorrelation. The residuals generally look like noise, with a few exceptions attributed to the nature of these data, and the ACF is within the bounds for all factors of the lag.

After building two models, I chose to use the forecasts from the best performing model. For each time-series, the error is the sum of the difference between each true spread and predicted spread. Each method had a vector of errors of 414 errors.

When looking at the error vectors, I removed 5 outliers where *each model* had error sums above 100 total points. It is interesting to note that both models had the same forecasts for some series' – especially those with the largest errors. These massive errors that both models found are likely due to games that were affected extraordinary circumstances for which my model cannot account. I did not use the time-series predictions for these 5 games for my simulations either.

Table 3.4: Sum of Error Greater than 100

gameID	Week
PHIvJAC	8
GBvDET	5
CARvATL	2
LARvTEN	16
LACvJAC	10

Table 3.4 shows the five games that were excluded. After taking a brief look at these games, it is noteworthy that the PHIvJAC game was played in London at 9:30AM ET (6:30AM PT) on a Sunday. The odd start time could have caused odd betting patterns where there were way fewer bets in the last third of observations than normal. Typically the amount of cash increases more linearly. However, with such an early start time on a Sunday morning, combined with the fact that people often have plans on Saturday nights, there may be a massive influx of money very close to the start of the game, as people wake up just before the game starts – opposed to having a few hours to place bets before the game starts.

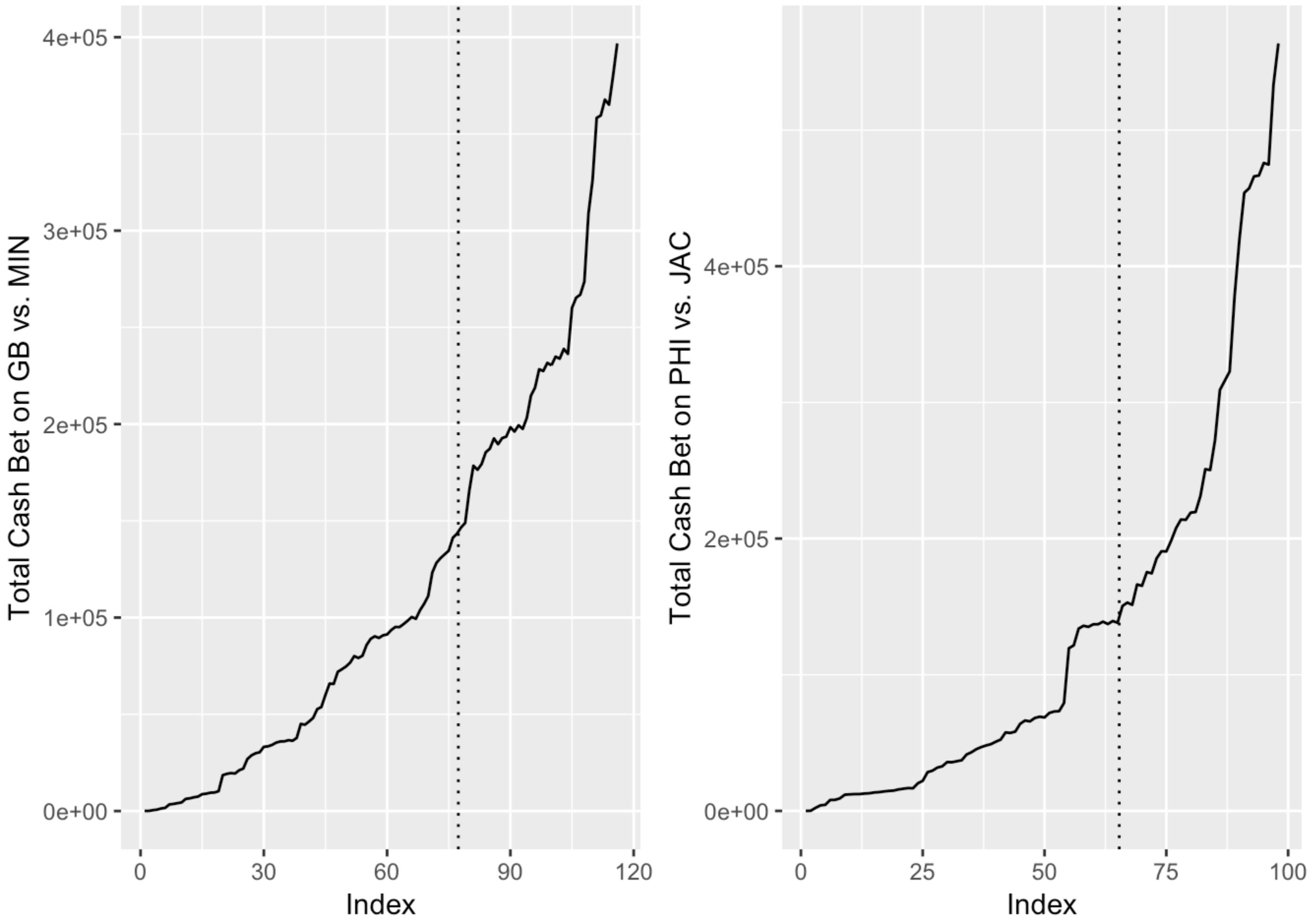

Figure 3.11: Total Cash Bet on GB vs. MIN (left) compared to PHI vs. JAC (right) Throughout the Week

Figure 3.11 represents the amount of cash bet throughout the week. The dotted line is the decision point. The charts show that the odd start time games have a significantly more massive exponential increase in the amount of money bet directly after the decision point. This makes these games tough to model. In addition, looking at the GB vs. DET game that was a massive outlier, star Green Bay Packers quarterback Aaron Rodgers was questionable to play throughout the week due to injury. He was finally announced as healthy late in the week. It is unclear the circumstances for the other three outliers.

Table 3.5: Summary Statistics for Errors

	Min.	**1st Q**	**Median**	**Mean**	**3rd Q**	
DLM Errors	-11.81761	-0.5605623	0	0.0519290	0.4935073	38.

Table 3.5: Summary Statistics for Errors

	Min.	1st Q	Median	Mean	3rd Q	
Auto ARIMA Errors	-11.81761	-0.5544576	0	0.0690394	0.4947444	38.

Table 3.6: Mean and Median Absolute Error

	DLM	Auto A
Median Abs. Error	0.5468115	0.76
Mean Abs. Error	1.4152842	3.40

Table 3.5 displays the summary statistics for my two vectors. these data shows that the DLM model has a lower mean error. In addition, when looking at simply absolute error, the Bayesian DLM approach provided a lower median absolute average error, as seen in Table 3.6, so I used this model's forecasts for incorporating the future values from my decision point.

3.1.3 Modeling Number of Observations

To predict how many future points to forecast from a certain time tt, I built a simple linear regression model. I gathered ten equally spaced data points from each of my data sets. Each data point contains information on the amount of total cash, total number of tickets and number of observations up to time tt, as well as the number of final data points in this series. One row of this data frame is shown in Section 6.2 of the Appendix. I then built a simple mixed linear regression model to forecast the number of total data points in the series, so I could find how many points hh I should use for forecasting at my decision point. While I considered using Poisson regression because the number of observations are a number of occurrences, the Poisson mixed linear and simple model did not fit the data as well as the linear mixed model, based on the diagnostics of the model. Equations (3.9) — (3.11) is the equation for this simple mixed model, with $n_i$ni representing the amount of final observations in the series, while $n_t$nt is the amount of observations up to time tt. Week is a factor and random effect (playoffs are treated here as week 0), as certain weeks attract more bettors than other weeks.

for j in 0 , ... , 17$\hat{n}_i=\beta_0+\beta_1\cdot\log(\text{Total Cash Bet})+\beta_2\cdot\log(\text{Total Tickets})+\beta_3\cdot n_t+\alpha_{\text{week}_j}+\epsilon_i \quad \epsilon_i\sim N(0,\sigma^2_{\text{residuals}})\alpha_{\text{week}_j}\sim N(0,\sigma^2_{\text{week}})$(3.9)(3.10)(3.11)

for j in 0 , ... , 17(3.9)ni^=β0+β1·log(Total Cash Bet)+β2·log(Total

$$\text{Tickets}) + \beta_3 \cdot n_t + \alpha_j^{week} + \epsilon_i \quad (3.10)$$

$$\epsilon_i \sim N(0, \sigma_{residuals}^2) \quad (3.11)$$

$$\alpha_j^{week} \sim N(0, \sigma_{week}^2)$$

The coefficients and diagnostics for this model are also shown in Section 6.2 of the Appendix, as this is a less essential part of the greater goal of this thesis.

3.2 Game Result Prediction

3.2.1 Overview of Decisions

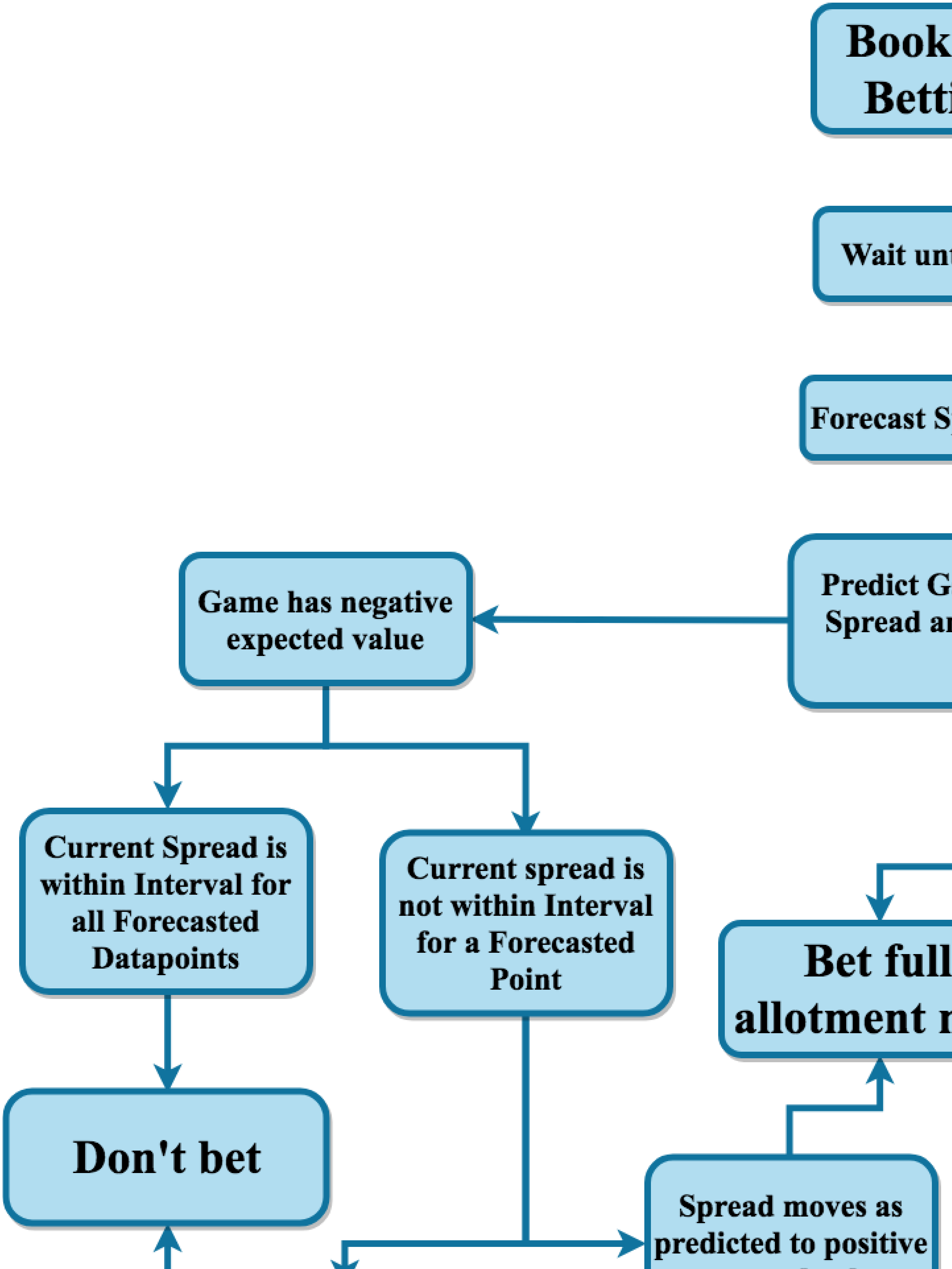
Book
Betti
Wait unt
Forecast S
Predict G
Spread ar
Game has negative expected value
Current Spread is within Interval for all Forecasted Datapoints
Current spread is not within Interval for a Forecasted Point
Bet full
allotment r
Don't bet
Spread moves as predicted to positive

Figure 3.12: Overview of Betting Decisions

Figure 3.12 shows a flow chart detailing the different possible scenarios and how much I would bet in each scenario. I use “allotment” to describe the betting amount because the many different betting strategies will bet different amounts for the same scenarios. The bookmakers open up betting on the game by placing an initial spread typically about a week before the game starts. I then wait until my decision point, forecast the spread for the rest of the week up until game time and provide a probability estimate for each team beating the spread. If betting on the game provides negative expected value based on the probability point estimate, I do not bet on the game, but I leave the opportunity open to bet later on in the week if a new, forecasted spread would make the advantageous to bet on. If the game has positive expected value, I place my bet on the game at the decision point. However, if the future forecasted spread projects a new spread that is even more advantageous to bet on, then I will only place a portion of my bet at the decision point and wait to place the rest of my bet. If the spread does in fact move as projected, I then place the rest of the bet the moment the spread hits my projections.

3.2.2 Exploratory Data Analysis

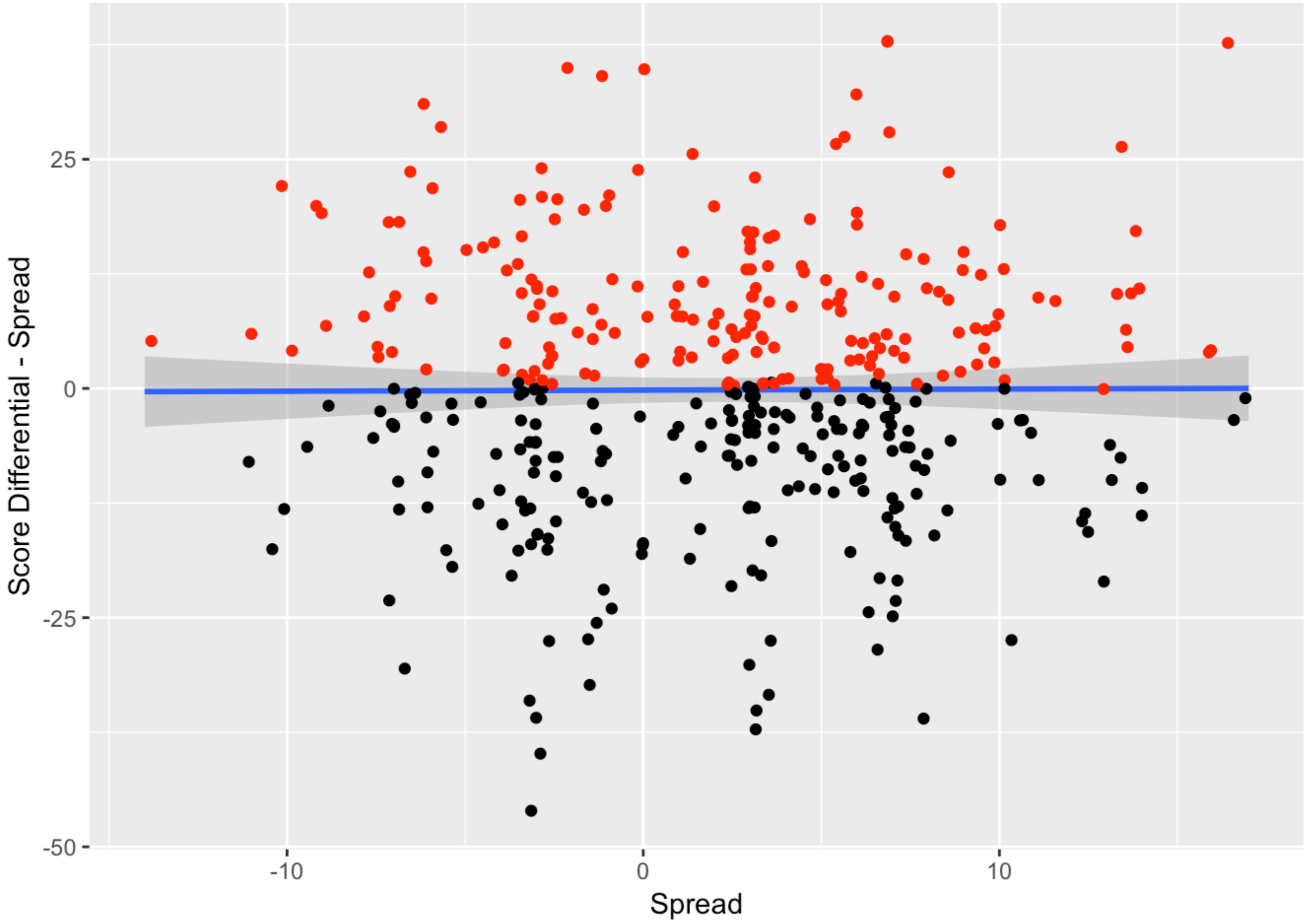

Figure 3.13: Game Result Against Spread vs. Spread – The red indicates that the team has beat the spread and the black indicates that the team has failed to beat the spread

Some key decisions determine whether the actual spread itself was a major factor in predicting team performance against the spread. In Figure 3.13, the y variable is the score differential during the game subtracted by the spread, in order to standardize the scores. For example, if the away team wins by 11 points, and the spread had the away team favored by 10 points, the y-variable in this scenario would be 1, as the away team performed one point better than the spread. The x variable is the spread. The red points are the observations where the away team covered the spread and the black points are the observations where the home team covered the spread.

The spread does not seem to have any impact on the team's performance against the spread. This means that bookmakers do not have any dead zones in making spreads where a certain team is much more likely to beat the spread at a certain point. There do not seem to be any biases (either making spreads too small or too large), with respect to the spread and the performance.

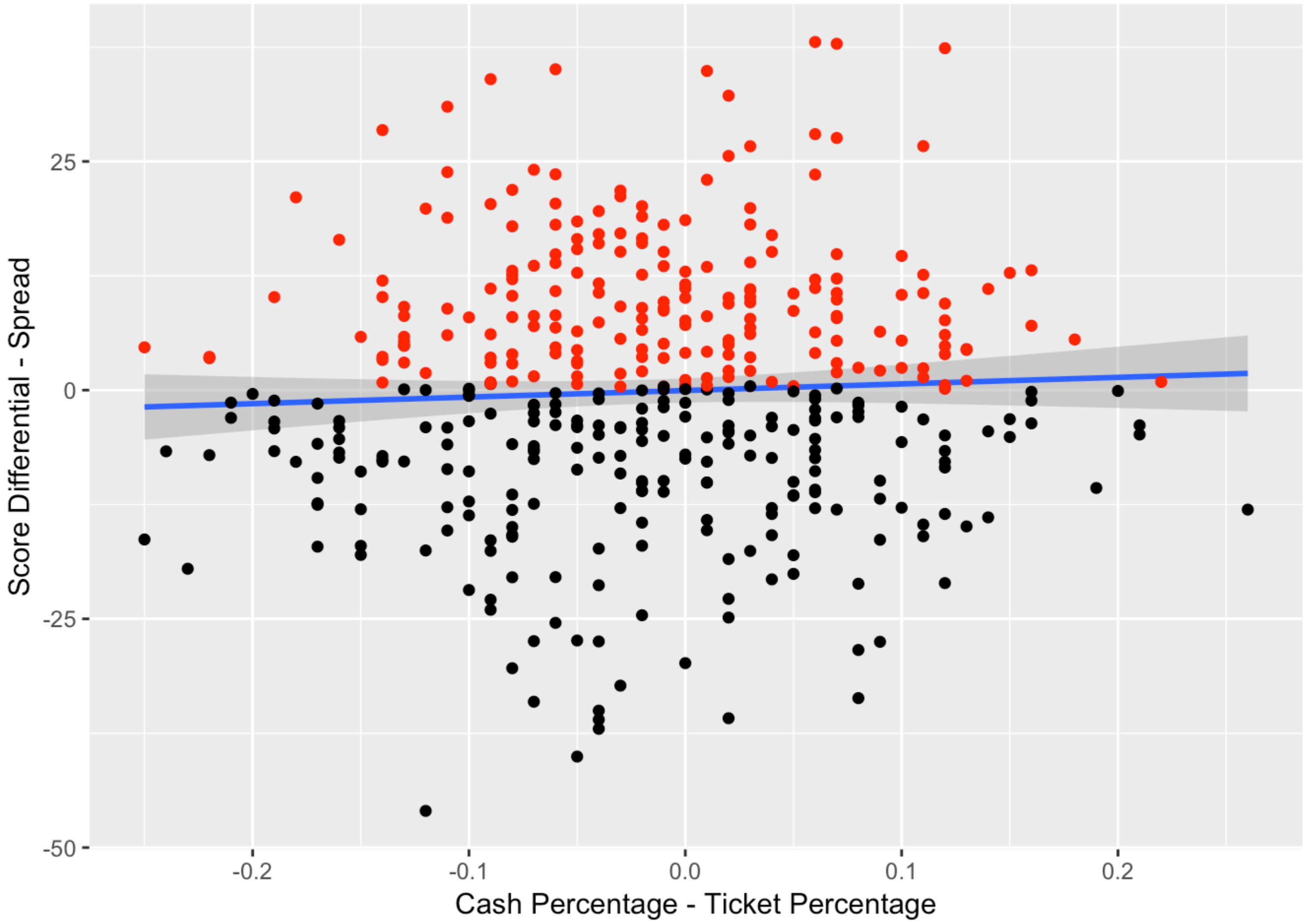

Figure 3.14: Result Against Spread vs. Cash-Ticket Percentage Difference – The red indicates that the team has beat the spread and the black indicates that the team has failed to beat the spread

Figure 3.14 examines the relationship between the cash and ticket percentages and the outcome against the spread. When there is a significantly higher percentage of cash bet on a team, in comparison to to the number of bets on a team, one of the teams is receiving larger bets. This is typically an indicator that professional bettors are betting on a team. Those who bet on sports for living tend to bet significantly more than those who bet recreationally, and the professional betters tend to be correct more often than the recreational betters.

From Figure 3.14, when the cash percentage rises, in comparison to the ticket percentage, the team tends to perform slightly better, with respect to the spread. This is an indication that the cash-ticket difference may be a useful indicator of performance.

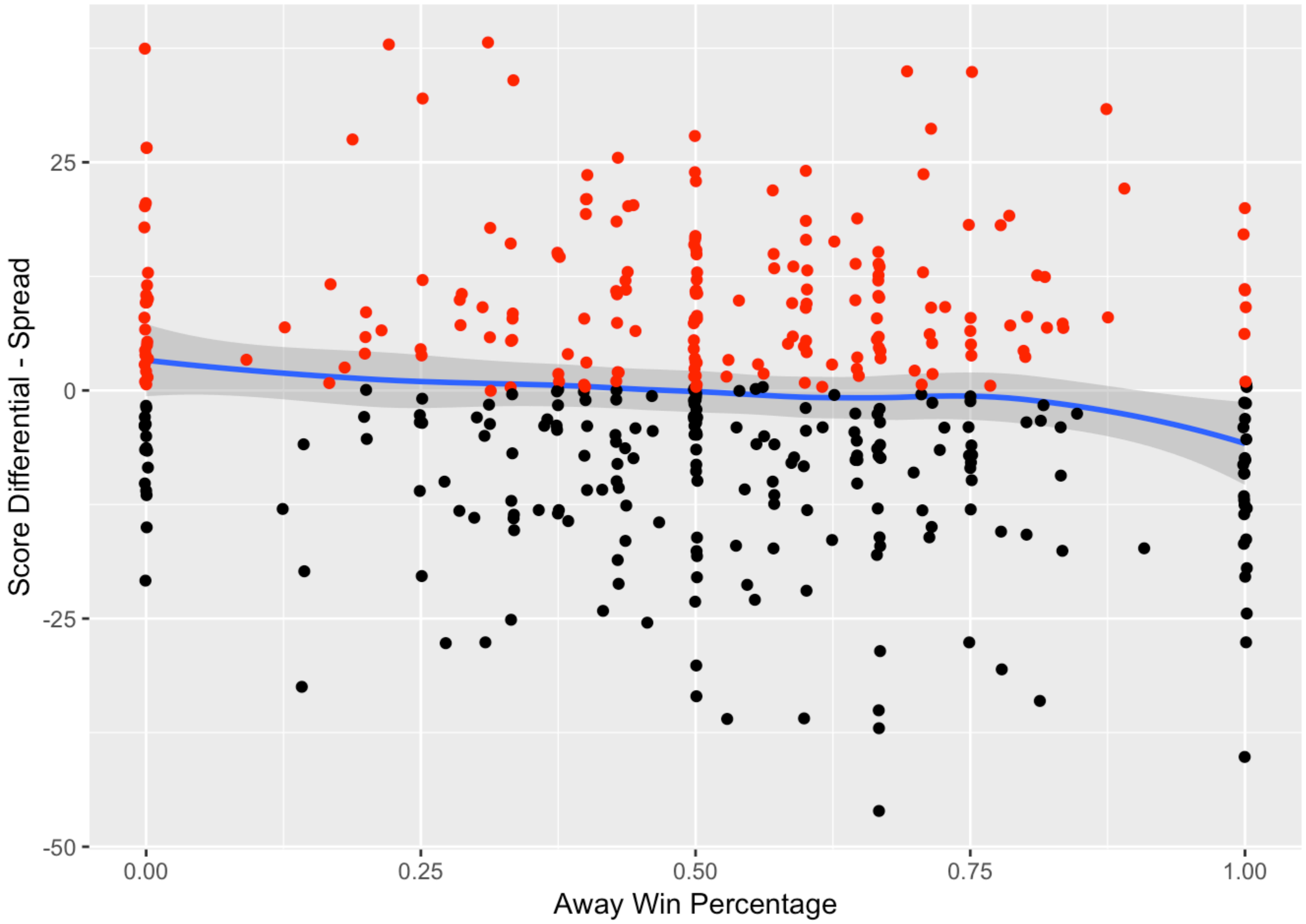

Figure 3.15: Game Result Against Spread vs. Away Win Percentage – The red indicates that the team has beat the spread and the black indicates that the team has failed to beat the spread

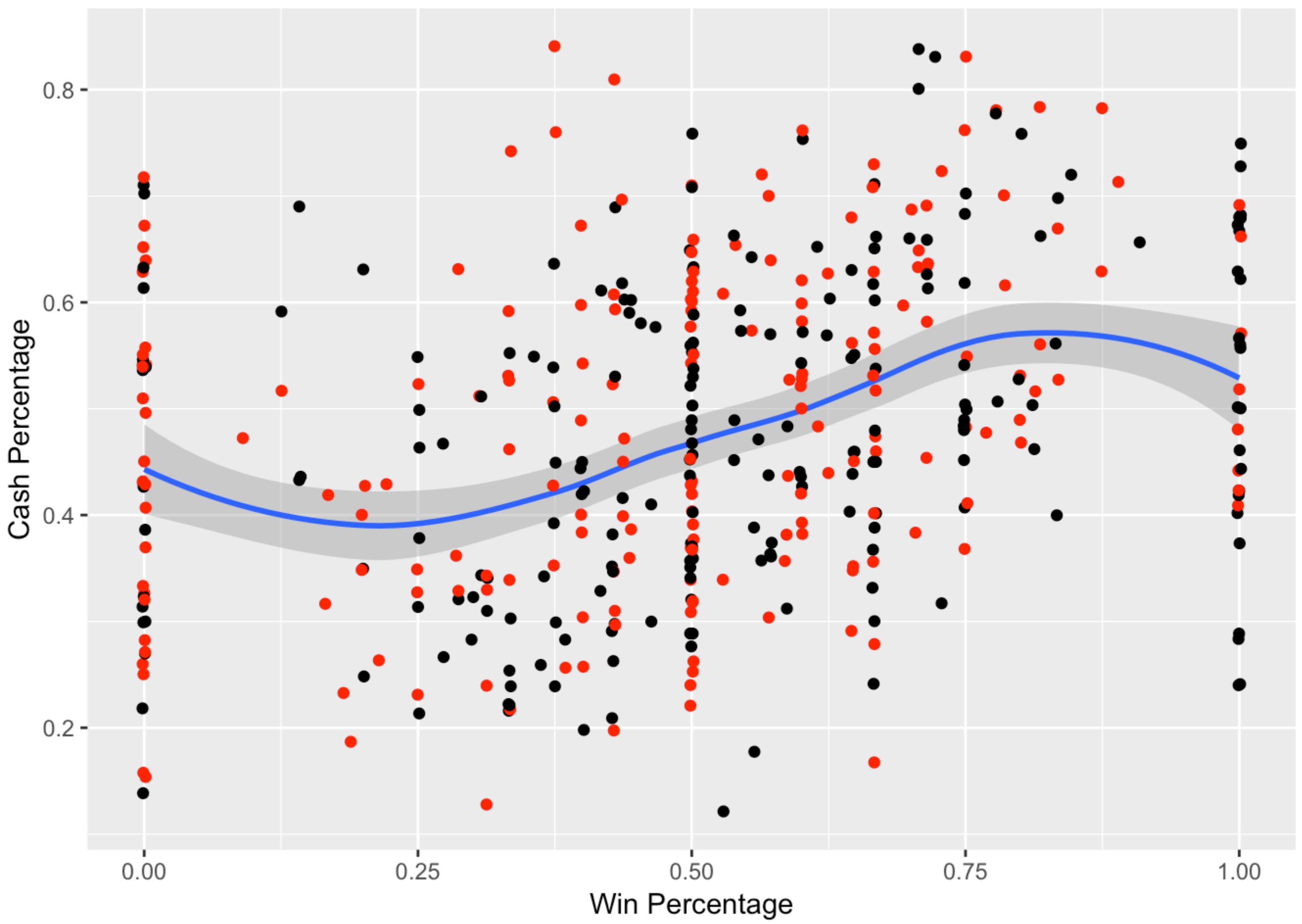

Figure 3.16: Cash Percentage vs. Win Percentage – The red indicates that the team has beat the spread and the black indicates that the team has failed to beat the spread

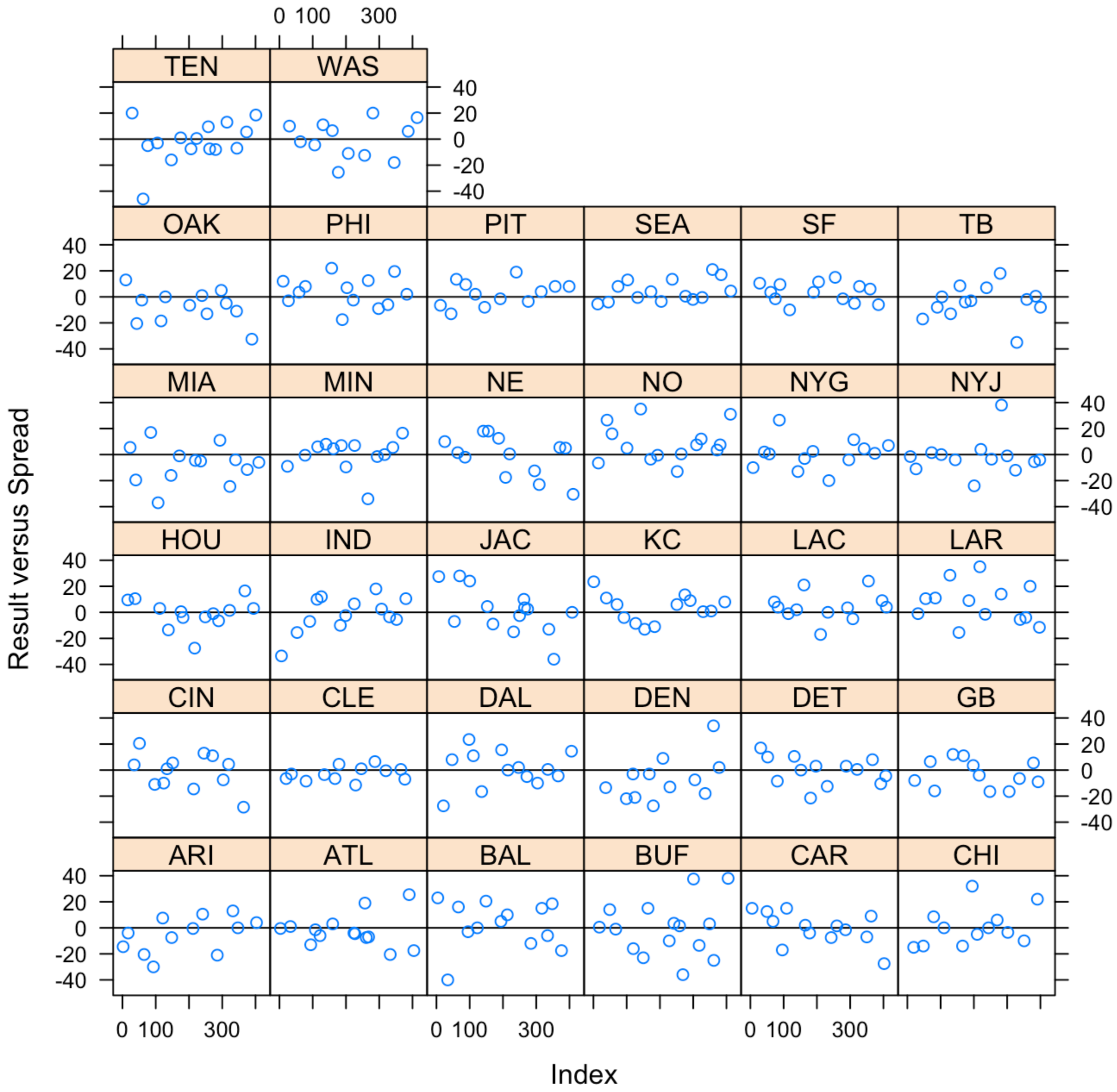

Figure 3.17: Result versus Spread by Away Team

Figure 3.15 shows a team's performance against the spread relative to its current win percentage. The data shows that as the win percent rises for a team, its performance against the spread gets worse. This is indicative of the fact that many bettors overreact to past performance – especially when it comes to undefeated or winless teams, so the bookmakers will "shade" the lines against the more popular team. For example, if a team is 2-0, many bettors will overreact to a small sample size, and in order for the bookmakers to achieve equal amount of money on each team to guarantee themselves a profit, the bookmakers will move the line against the undefeated team. The opposite phenomena occurs for winless teams.

Figure 3.16 shows that as win percentage increases, the cash percentage tends to increase. At the edges with win percentages of 0% and 100%, this trend seems to slightly reverse. This is likely due to bookmakers shading the lines at such an extreme amount for these extreme win percentages, where they are able to achieve nearly equal action.

Figure 3.17 shows the result against the spread for each away team. There is great variation among all the teams, and while certain teams seemed to perform better against the spread, like the New Orleans Saints, treating the team as a random effect in modeling seems to suit the data.

3.2.3 Model Approach

There were a few different approaches to modeling that deserved consideration. Because scores are only in whole units, an ordinal regression model seemed as if it could have been appropriate. However, because there are an unbounded amount of levels, as well as the fact that there are so many levels – many of which have few data points – this approach would not have yielded appropriate results. A mixed linear model is a good approach to model these data with many different groups (the different teams). The downfall to this approach is that it does not give extra weight to the peaks in the score differences between games at 3 and 7, but still the score predictions would be more accurate than an inappropriately used ordinal regression model. Perhaps if there were tens of thousands of data points where each level would be represented numerous times, an ordinal regression would be more appropriate.

To first assess the best mixed linear models, the models were whittled down based on minimizing the BIC on the full dataset. After finding two models with similar BIC's but different predictors, the models were compared through k-fold validation. There were a few metrics in this used: error rate between predicted results for the test set and the actual results, and then betting (and bankroll) performance across each of the simulations. The k-fold validation used 100 simulations in order to get a large distribution of bankroll amounts. But, if this k-fold validation was performed as usual, this would leave the test data sets with only 4 data points. Instead, the data was randomly shuffled for each of the 100 iterations, and then broken up into 7 folds – with one fold used as a test data set and the rest as a training dataset.

3.2.4 Simulations

For generating the simulated probabilities of beating the spread for each game in the test dataset, I generated 500 draws from its posterior predictive distribution for each model.

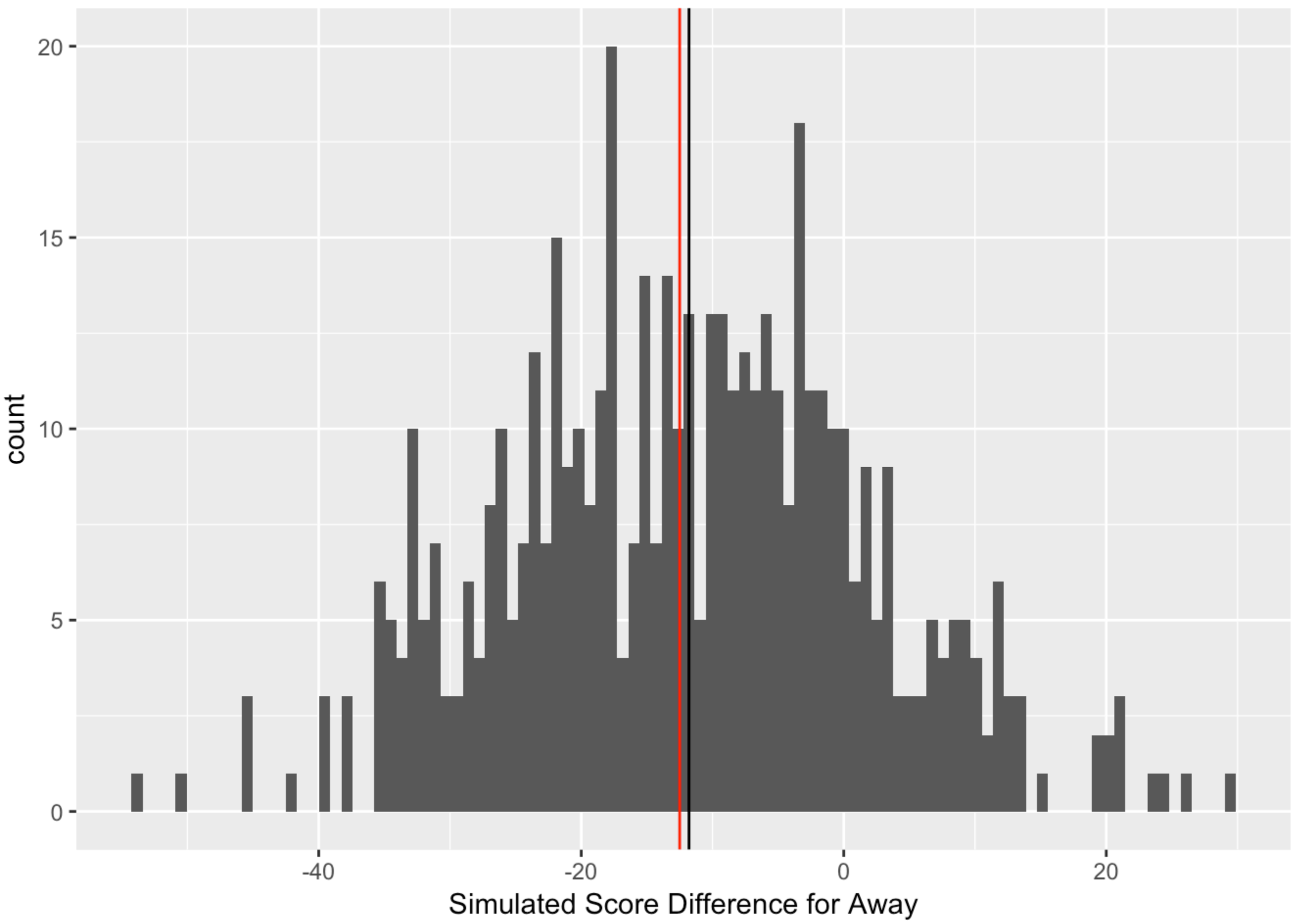

Figure 3.18: Simulated Outcomes for NYG @ Den Week 6, 2017

Figure 3.18 is a histogram representing the results of the 500 draws from the posterior predictive distribution from the best overall performing model (as will be discussed in Section 3.2.5) for an example game in a test dataset for the New York Giants at the Denver Broncos during Week 6, 2017. The vertical black line represents the median of the 500 draws from the posterior predictive distribution, and the vertical red line represents the actual point spread. The median of the simulated outcomes (the vertical black line) is placed at -11.8, meaning the away team, the Giants, are expected to lose this game by 11.8 points. However, the spread (the vertical red line) at our first decision point has the Giants +12.5 points, meaning to beat the spread, the Giants must lose by 12 points or fewer, or win. Thus, at first glance, there seems to be a slight edge on betting on the New York Giants +12.5 because the spread has the Giants losing by 12.5 points, but the model projects the Giants to only lose by 11.8 points.

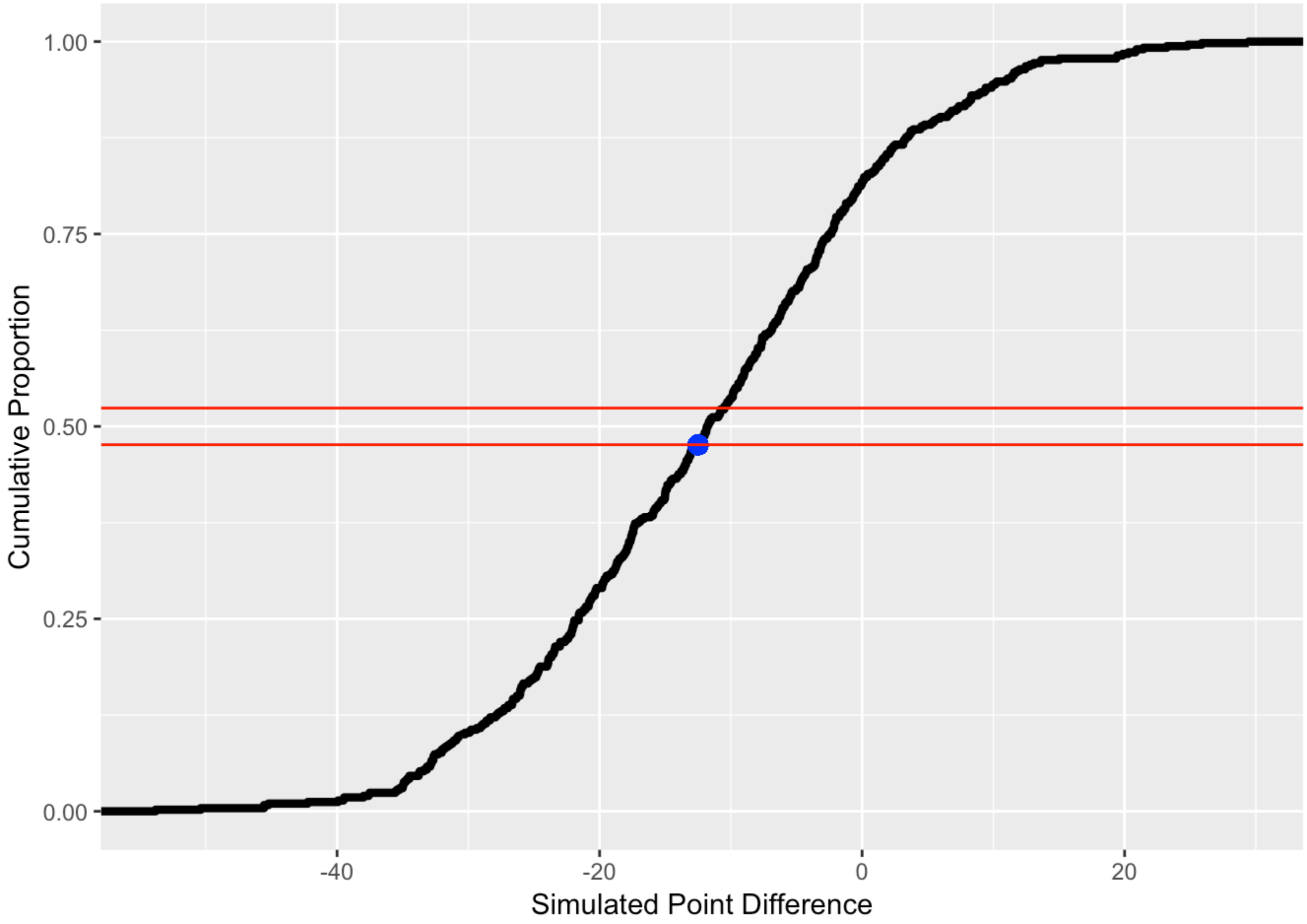

Figure 3.19: Empirical Cumulative Distribution of Simulated Outcomes for NYG @ DEN Week 6, 2017

Figure 3.19 is the empirical cumulative distribution (ECDF) of the 500 draws from the posterior predictive distribution. The blue represents where the point spread falls in the ECDF. Being either above or below the two redlines means that betting on this game will generate a positive expected value. If the point is below the lower redline, it is advantageous to bet on the away team, and if the point is above the top red line, then it is advantageous to bet on the home team. The interval of these red lines is (0.4762, 0.5238). If the ECDF is below 0.5, the probability of success is 1 - ECDF(Point Spread). Because the casino does not give fair odds, and offers -110 odds, where a bettor must stake 1.1 units to win 1 units, this interval of probabilities generates a negative expected value. The edges of the probability provide an expected value of 0. Expected value is calculated by adding the probability of failure multiplied by -1.1 (the amount of units lost if the bet loses) and the probability of success multiplied by 1 (the amount of units won if the bet wins). Equation (3.12) is the equation for expected value.

$$P(\text{Beating Spread})\cdot 1+(1-P(\text{Beating Spread}))\cdot -1.1. \quad (3.12)$$

(3.12)P(Beating Spread)·1+(1−P(Beating Spread))·−1.1.

Now, to find the probability of success for each game, I found where on the ECDF of the draws from the posterior predictive distribution the current spread falls. For example, the ECDF for this point spread of Giants (+12.5) is $0.478$0.478, so the probability of the Giants beating the spread is $1-0.478=0.522$1−0.478=0.522. The model expects the Giants to beat the spread with a proportion of $0.522$0.522. The model expects the Broncos to beat the spread with a proportion of 0.478. Since the spread, in this scenario, is 12.5, and not a full number, there is no probability of a push, or tying the spread.

After generating a probability of success, the expected value can be calculated. Since one must bet 1.1 units to win 1 unit, the expected value is $0.522-((1-0.522)\cdot 1.1)=-0.0038$0.522−((1−0.522)·1.1)=−0.0038. Betting on the Broncos is even more disadvantageous, as their expected value is $0.478-((1-0.478)\cdot 1.1)=-0.0962$0.478−((1−0.478)·1.1)=−0.0962.

In this scenario, the model suggests a negative expected value of betting on the Giants with this spread of -0.0038 units lost per unit bet. There is a negative expected value for betting on both teams! So, because of the negative expected value, there will be no bet on the game at this point. However, the forecasted spread impacts whether there may be a bet at a future time point.

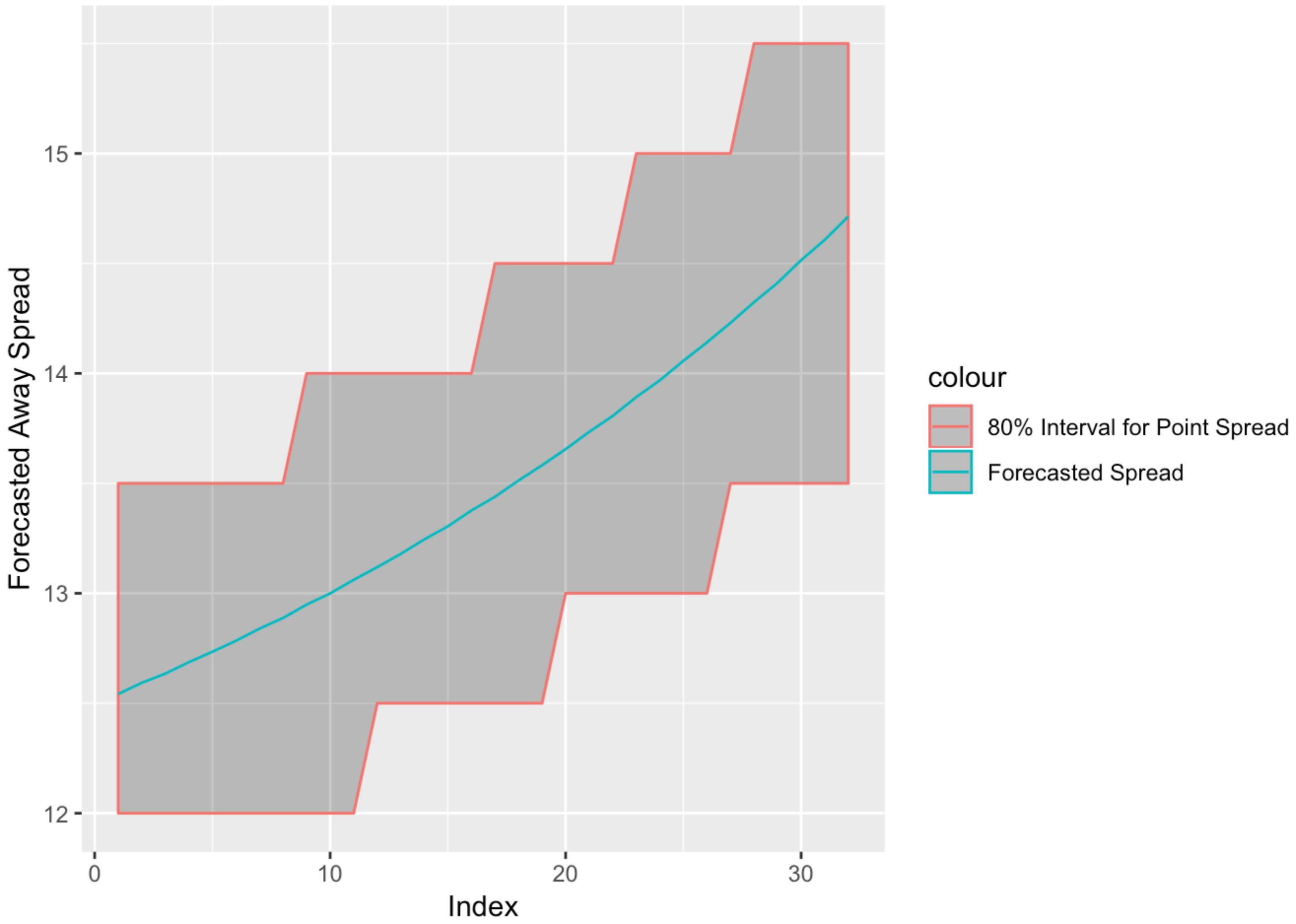

Figure 3.20: Forecasted Spread for the NYG @ DEN Week 6, 2017 with 80% Interval

Figure 3.20 shows the forecasted spread up until projected game time for the Giants and Broncos game. The 80% interval is using a rounded spread, to the nearest one-half, to calculate my interval. The forecasted spread predicts an 80% confidence interval of (13.0, 14.5) for this spread about 20 data points into the 32 point forecast. The current decision point spread of 12.5 is outside of this interval. The expected value changes once the spread enters my interval. The Empirical Cumulative Distribution for the Giants when the spread is Giants (+13) is 0.47 meaning the simulated probability is $1-0.47=0.53$1−0.47=0.53. The new expected value is $0.53-((1-0.53)\times1.1)=0.013$0.53−((1−0.53)×1.1)=0.013. Thus, if the spread does move within my interval at any point, I will bet.

This is an extremely small edge. However, the spread does actually move to 13.0, so there would be a bet on the Giants. However, the 80% interval later moves even further to (13.5, 15.5) about 30 index points into the forecast. The new ECDF of Giants (+13.5) is 0.45, meaning the new simulated probability is 0.55 and the new expected value is 0.055. The level of confidence that the spread will move to Giants (+13.5) is only 80%, but $0.055\times0.8>0.013$0.055×0.8>0.013, so at the first point of positive expected value, I choose that my bet is only one-third of the total allotment. For example, if the bet allotment for this game is 15 units, I would place a 5 unit bet on the Giants (+13). The other two-thirds of the allotment will be placed if the spread enters my interval and hits 13.5. In actuality, the spread does move to Giants (+13.5). So, two-thirds of the bet allotment – or 10 units – is placed at Giants (+13.5).

This ended up being extremely important because the Giants actually lost the game by 13 points, so a bet on the Giants at (+12.5) would have lost money, while the 5 unit bet on the Giants +13 is a push, meaning the money is returned, and the 10 unit bet at Giants +13.5 wins and returns a profit of $(10/1.1)=9.09$(10/1.1)=9.09 units!

This was the process I went through for each game in the test data set for each model, as there were different probabilities of beating the spread from the two different models. For comparison, I used a simple method from a simple multiple linear regression, where the point estimate was generated directly through utilizing the mean and variance of the predicted value from the formula to calculate the t-value of the point spread and the using the t-distribution to find a probability estimate. One row of my test data set with the probabilities included is displayed in the Section 6.3 of the Appendix.

3.2.5 Model Selection

There were two models that provided similar results of BIC on the full datasets. Both models used the away team as a random effect, and used the decision point spread as a predictor. But, the first model delves into more team-specific stats, such as win percentages, number of wins and the weighted DVOA, in order to best predict who will win. I will refer to this model as the “team-specific” model. The second model tends to look more at the betting trends, such as the log of the tickets and cash bet for both the away and home team, and the difference between the cash and ticket percentage (this model also uses the difference between the teams’ weighted DVOA).

for its predictors. The second model also uses the year as a random effect. This model will be referred to as the “betting-trends” model When looking to incorporate certain additional variables into the other models, the BIC rises.

The team-specific model is shown in Equations (3.13) — (3.15).

for i in 1 , ... , 414 and j in 1, ... , 32^Away Score - Home Score$_{i}$=α$_{\text{away team}j[i]}$+β$_{′}$·X$_{i}$+ϵ$_{i}$ϵ$_{i}$∼N(0,σ$_{2\text{residuals}}$)α$_{\text{away team}j}$∼N(0,σ$_{2\text{away team}}$)(3.13)(3.14)(3.15)for i in 1 , ... , 414 and j in 1, ... , 32(3.13)Away Score - Home Scorei^=αj[i]away team+β′·Xi+ϵi(3.14)ϵi∼N(0,σresiduals2)(3.15)αjaway team∼N(0,σaway team2)

Output of Team-Specific Model

	Dependent variable:
	Away Score - Home Score
poly(home.wins, 2)1	174.093
	(113.876)
poly(home.wins, 2)2	116.674
	(68.643)
poly(away.wins, 2)1	-90.551
	(100.298)
poly(away.wins, 2)2	-106.191
	(62.894)

first_decision_point_spread	-0.790
	(0.179)
home.winpercent	-5.878
	(5.919)
away.winpercent	-13.392
	(5.983)
away_WEI.dvoa	0.197
	(0.070)
poly(home.wins, 2)1:home.winpercent	-292.348
	(150.945)
poly(home.wins, 2)2:home.winpercent	-73.081
	(79.646)
poly(away.wins, 2)1:away.winpercent	182.016
	(141.167)
poly(away.wins, 2)2:away.winpercent	86.750
	(82.067)

Constant	9.993
	(3.839)

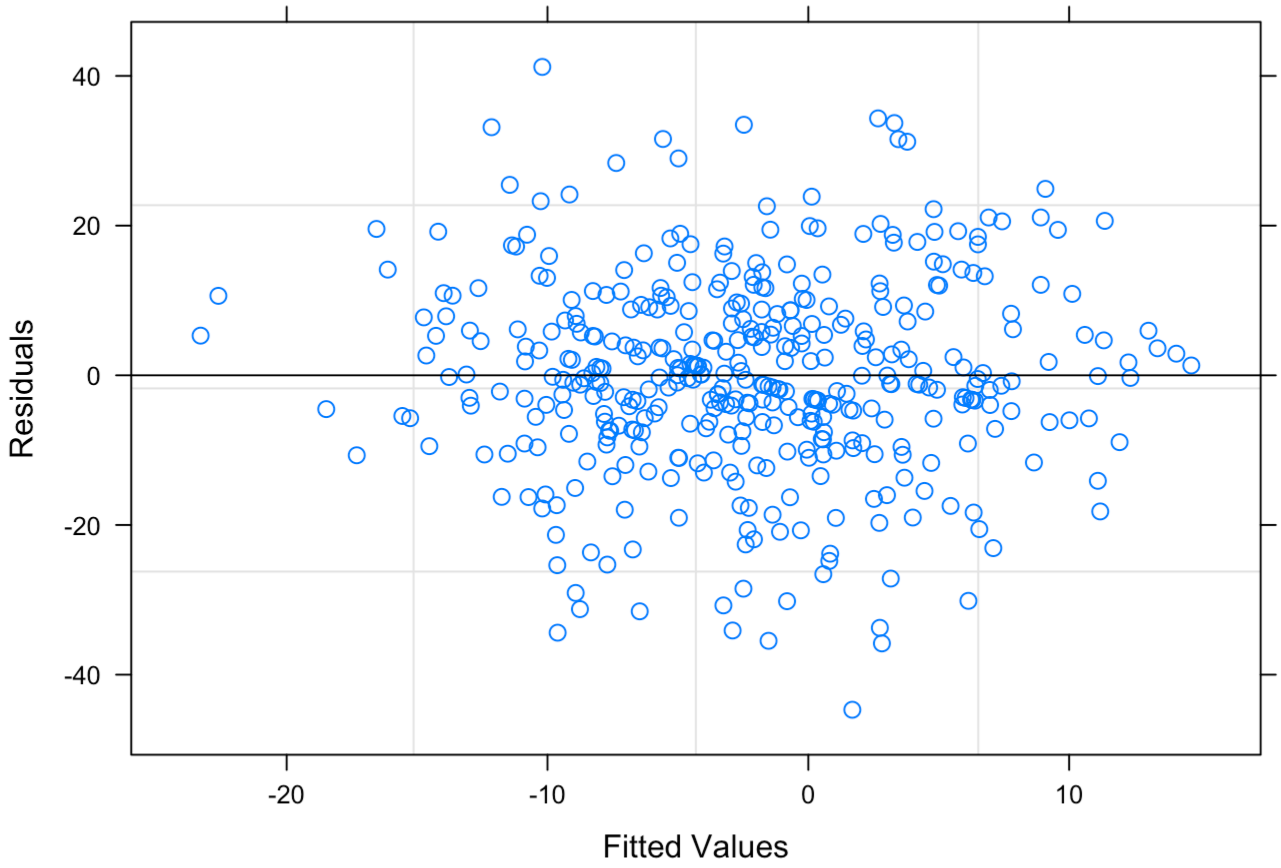

Figure 3.21: Residual Plots for Team-Specific Model

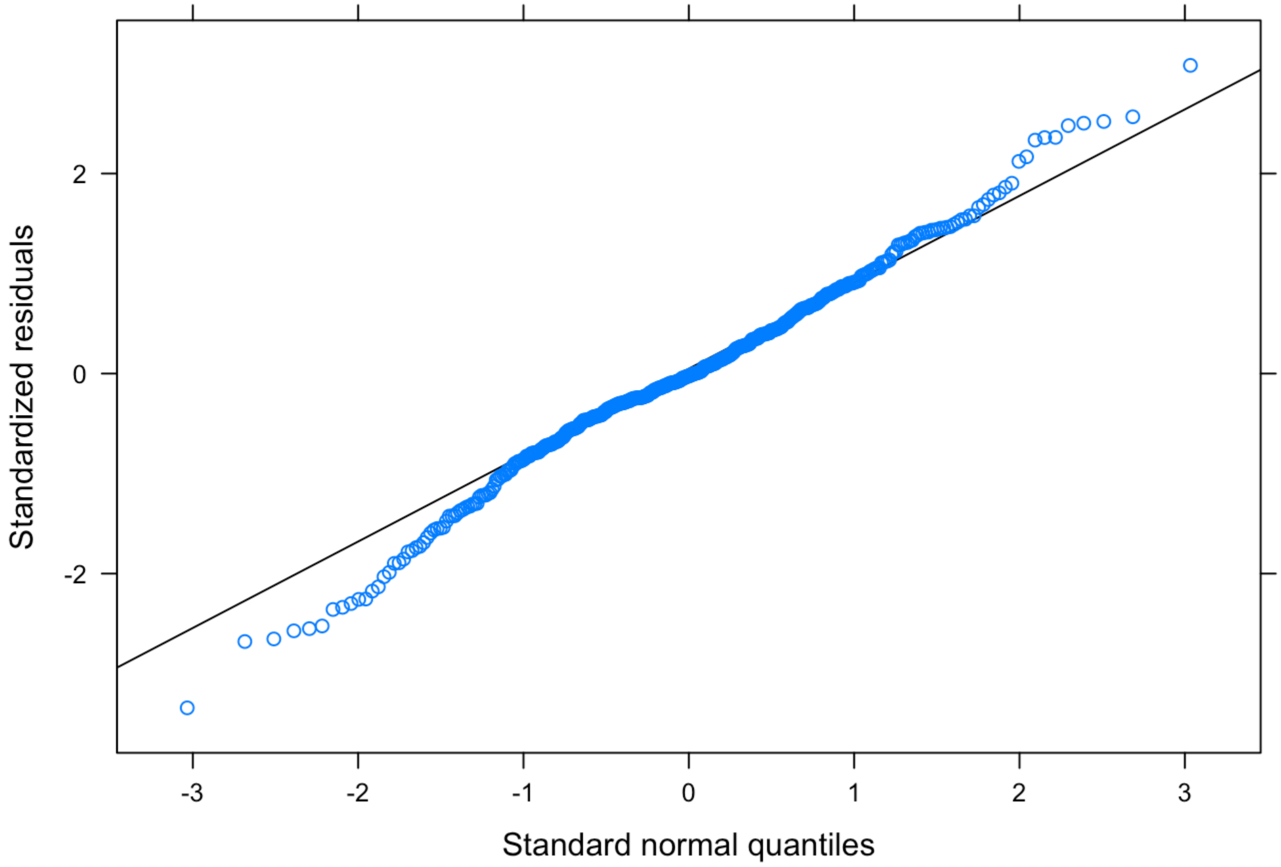

Figure 3.22: Residual Plots for Team-Specific Model

Table **??** is the output parameters for the Team-Specific model.
Figures 3.21 and 3.22 is the diagnostic plots for this first model, and the model seems to pass all the diagnostic tests. The residuals tend to be random and non-correlated; the residual plots based on the groups are shown in Section 6.4 of the Appendix, but there are no egregious errors.

The model focusing on betting trends is shown in Equations (3.16) — (3.19):

for i in 1 , ... , 414; j in 1, ... , 32 and m in 2017,2018^Away Score - Home Score$_{i}$=α$_{\text{away team}j[i]}$+α$_{\text{Year}m[i]}$+β$_{'}$·X$_{i}$+ϵ$_{i}$ϵ$_{i}$∼N(0,σ$_{\text{2residuals}}$)α$_{\text{away team}j}$∼N(0,σ$_{\text{2away team}}$)α$_{\text{Year}m}$∼N(0,σ$_{\text{2Year}}$)(3.16)(3.17)(3.18)(3.19)for i in 1 , ... , 414; j in 1, ... , 32 and m in 2017,2018(3.16)Away Score - Home Scorei^=αj[i]away team+αm[i]Year+β′·Xi+ϵi(3.17)ϵi∼N(0,σresiduals2)(3.18)αjaway team∼N(0,σaway team2)(3.19)αmYear∼N(0,σYear2)

Output of Betting-Stats Model

	Dependent variable:
	Away Score - Home Score
log_away_cash_bet	17.360
	(11.854)
log_home_cash_bet	-21.420
	(11.946)
log_away_tic_num	-13.576
	(11.148)
log_home_tic_num	19.662
	(11.332)
WEI_away.diff	0.105
	(0.054)
first_decision_point_spread	-0.719
	(0.276)
away_total.dvoa	-0.069
	(0.036)

cash_tic_diff	-63.509
	(50.684)
Constant	5.371
	(22.890)

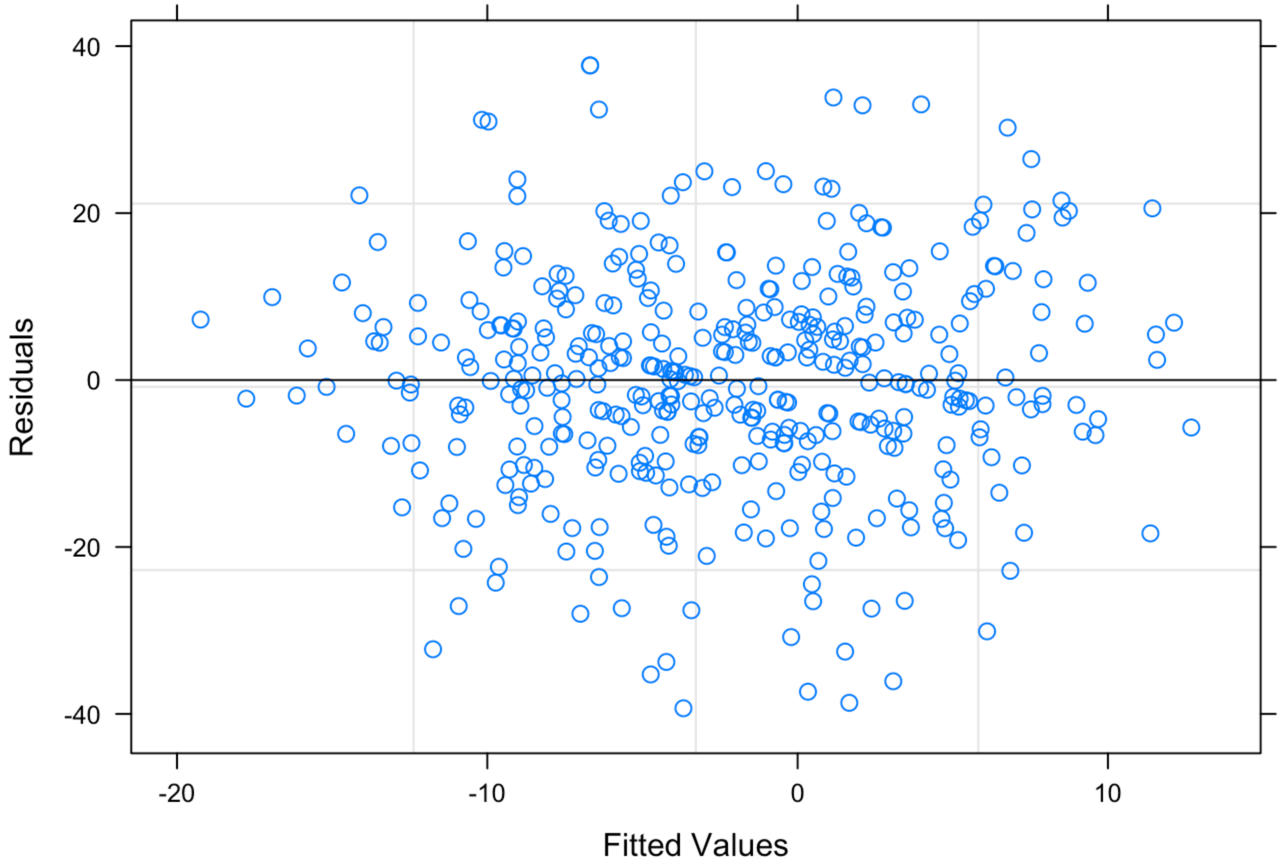

Figure 3.23: Residual Plots for Betting-Trends Model

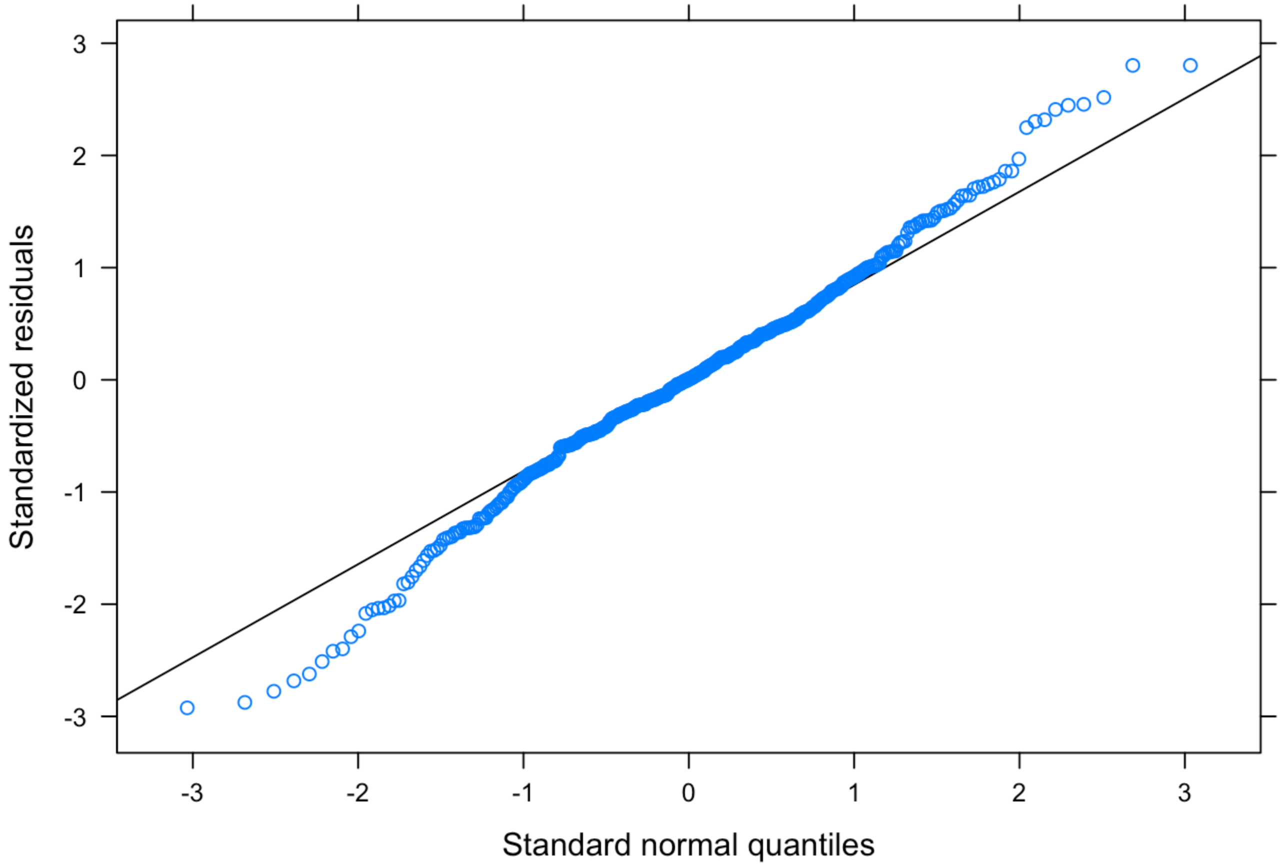

Figure 3.24: Residual Plots for Betting-Trends Model

Table **??** displays the parameters for the Betting-Trend model. This model also seems to pass all the diagnostic tests, shown in Figures 3.23 and 3.24, as the residuals tend to be random and non-correlated. The residual plots based on the groups are shown in Section 6.4 of the Appendix, but there are no egregious errors.

The mixed-linear models are appropriate for modeling these data, and k-fold validation using 100 test data sets is used to evaluate the models. It is possible that the models have different strengths and weaknesses, in terms of risk and reward, and this can be examined through looking at the distribution of winnings.

1. https://www.thesportsgeek.com/sports-betting/strategy/point-spread/↩
2. https://www.thesportsgeek.com/sports-betting/strategy/point-spread/↩

Chapter 4 Betting Strategy

Research[3] shows the most successful betting strategy is the Kelly Criterion. The Kelly Criterion states that each bet should be equivalent to the percent edge. For example, with an expected value of 0.109, or a 10.9% edge, the bet should be equivalent to 10.9% of the bankroll. However, this strategy is contingent on having an extremely accurate model where the expected edge is the true edge. Due to limited data of just two seasons right now, many of the expected values are above 0.2, which would be a ludicrous amount of the bankroll to bet on one game. So, all expected values are capped at 0.2 with this method.

Starting with a bankroll of 100 units, the absolute maximum bet is capped at 40 because it became too easy to lose any sort of winnings while using the Kelly criterion.

Another system used modifications to the Kelly criterion system. In this system, the expected values essentially received a square root transformation; however, since the absolute value of the expected value must be less than one, this is actually a squared transformation. The new Kelly proportion is the expected value squared.

The Martingale betting system is another popular method. This betting system starts by betting a certain amount of units (in this case, 5 units). If the bet loses, the subsequent bet doubles, so that winning the bet leaves the bettor ahead by 5 units. If the second bet instead loses again, the third bet once again double – now to 20 units. This keeps on going until either the bankroll is empty or the bettor is ahead 5 units. This is a risky strategy.

One strategy simply bets 15 units any time there is a positive expected value. Another strategy only bets when there is an extreme edge; the bet is 10 units if the expected value is above 0.1 and 20 units if the expected values is above 0.2.

Finally, another betting system relied on the agreement of the two linear mixed models. If both models suggested betting on the same team, and each expected value is positive, then the system bets on this team. Betting amounts are determined with the Kelly Criterion, and the Kelly proportion is the average of the two different expected values generated from the different models.

The different betting systems are listed in Table 4.1.

Table 4.1: Description of Betting Strategies

Strategy.Name	**Description.of.Strategy**
K1	Kelly Criterion using team-specific stats mixed linear model. Best performing model by error. M bet of 40 units. Capped expected value at 0.2. Utilize future betting strategy.

Table 4.1: Description of Betting Strategies

Strategy.Name	Description.of.Strategy
K2	Kelly Criterion using betting-specific model. Max bet of 40 units. Capped expected value at 0.2 Utilize future betting strategy.
K3	Kelly Criterion using simple linear regression model. Max bet of 40 units. Capped expected val 0.2. Utilize future betting strategy.
K2 - Transform	Utilize same strategy as K1, but the Kelly proportion is the expected value squared.
Simple Bets	If positive expected value at first decision point from team-specific model, bet 15 units. If nega expected value, do not bet. Do not incorporate future betting strategy
Extreme Edge (Simple)	If edge at first decision point from team-specific model is greater than 0.1, bet 10 units. If grea than 0.2, bet 20 units.
Extreme Edge (Transform)	Only bet if the expected edge from team-specific model is greater than 0.1 and use Kelly Crite with square transformation for betting amounts. Utilize future betting strategy.
Agreement	If both mixed linear models suggest betting on the same team, bet on that team. Use Kelly Cri with a mean of the two expected values. Utilize future betting strategy.
Martingale	Use Martingale strategy if positive expected value from team-specific model. Use 5 units for ir bet. Do not incorporate future betting strategy.

4.1 Results Using K-Fold Validation

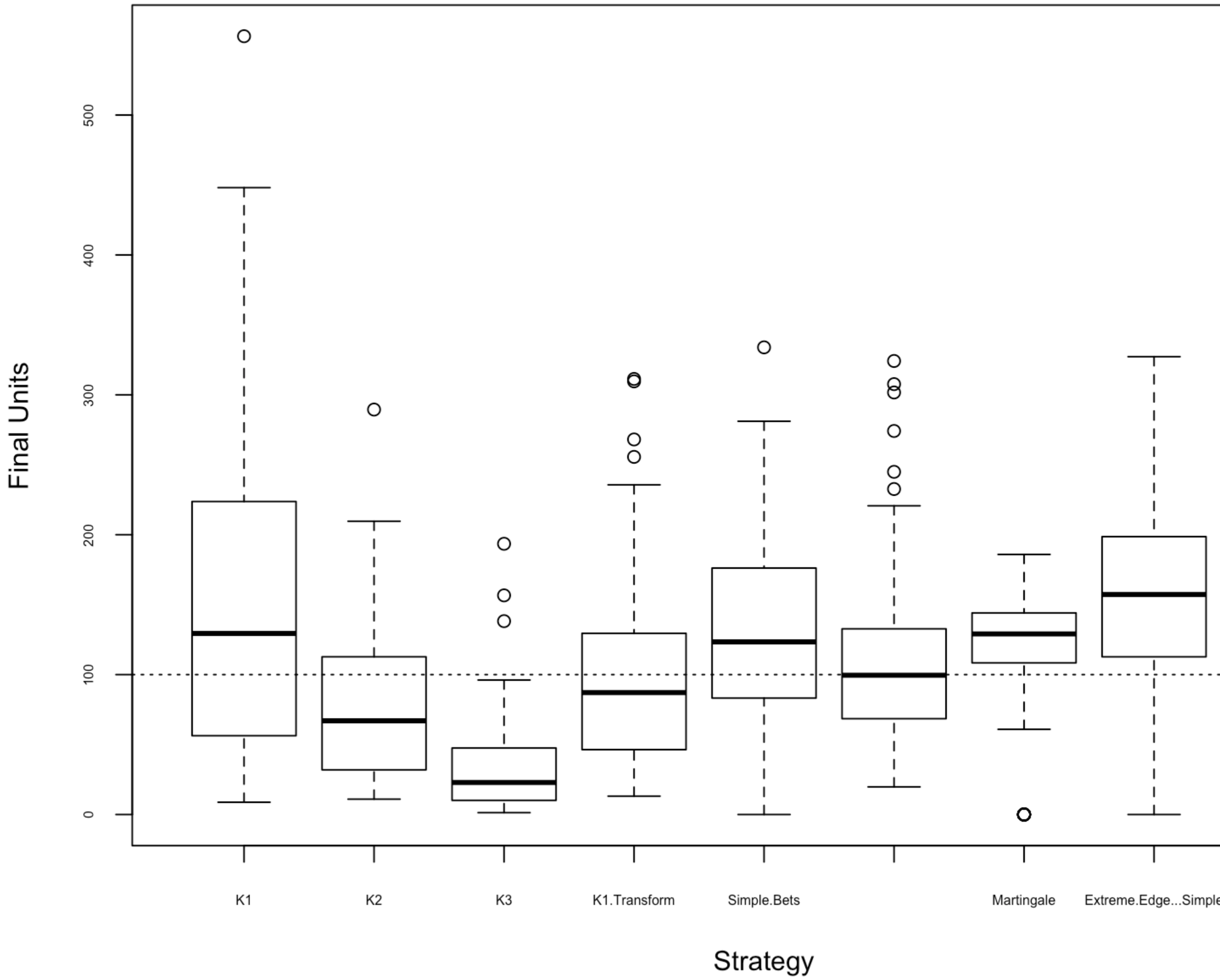

Figure 4.1: Final Bankroll Box Plots by Strategy

For each game in each test data set, the model went through the process outlined above for each betting strategy. After iterating through all the games in the test data set, I was left with the final bankroll from each model for each test data set in the k-fold. With 100 test data sets, I had a vector of 100 final bankrolls for each of the nine betting strategies.

Figure 4.1 is a boxplot looking at the distribution of the final bankrolls for each of the betting strategies over the 100 simulations. Most of the strategies have medians above 100, meaning they generated positive median returns. In addition, these distributions have a boundary on the low end of 0, as the strategy cannot end with a bankroll below 0, but with no maximum, the distributions tend to be right-skewed, leaving the means higher than the medians.

Table 4.2 shows the mean and median final bankrolls over the 100 simulations for the different methods. Four of the methods stand out; the K1 method, Martingale, Extreme Edge - Simple Bets and the LMER Agreement method. Outside of the Martingale method, which is interesting because this is the only strategy where the mean is significantly *lower* than the median, the other listed methods are noteworthy because of the huge returns – each of those three methods have mean returns above 35%.

Table 4.2: Mean Final Bankroll vs. Strategy

	Mean Final Bankroll	Median Final Ba
K1	153.61	1
K2	80.88	
K3	33.55	
K1.Transform	102.01	
Simple.Bets	131.82	1
Extreme.Edge...Transformed	109.19	
Martingale	113.62	1
Extreme.Edge...Simple	150.25	1
LMER.Agreement	137.62	1

Table 4.3 displays the summary statistics for the 4 top methods; the team-specific model using Kelly Criterion for betting amounts, the Martingale method, the Extreme Edge with Simple Bets and the Agreement method. The highest median is this K1 method. All these methods used the team-specific model to generate probabilities of

beating the spread. The returns are very high in each of the quartiles, with positive returns for over 75% of simulations for the Extreme Edge with Simple Bets and Martingale methods. However, these are both the riskiest methods, as in 5% and 15%, respectively for each method, the bettor would have lost the entire bankroll using these methods. This is a very high chance relative to most other more standard investments in the stock market.

Table 4.3: Summary Statistics for the Best Performing Methods

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	
Kelly Criterion	8.78	56.61	129.42	153.61	221.56	5
Extreme Edge - Simple	0.00	113.64	157.27	150.25	198.41	3
Model Agreement	20.31	71.41	121.40	137.62	179.07	4
Martingale	0.00	108.52	129.09	113.62	144.09	1

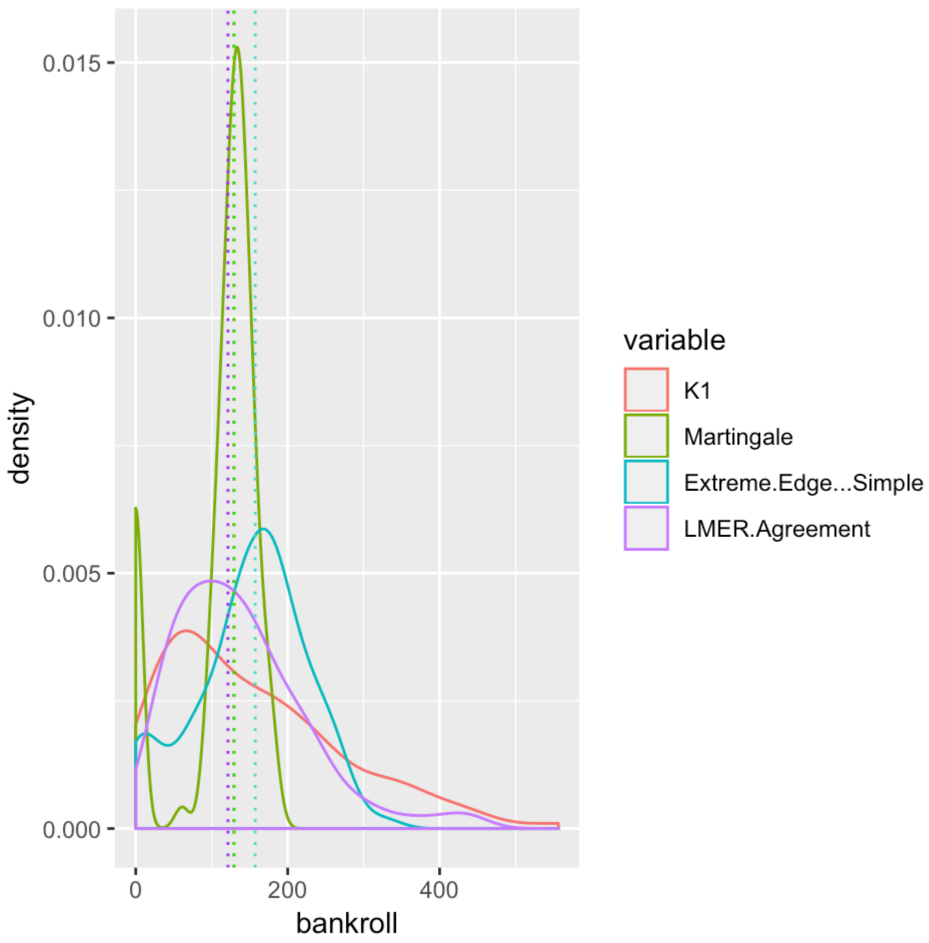

Figure 4.2: Density Plots comparing Four Successful Betting Strategies

Figure 4.2 shows the density plots for the 100 final bankrolls from the top four methods. The dotted lines represent the medians of each distribution. There seems to only be 3 dotted lines, but the medians for both the K1 and Martingale methods are both 129, so these lines are overlaid. The LMER-Agreement method seems to provide the safest betting strategy, as the density plot does not have as long of a tail, but a larger peak close to 175 units. It makes sense that this method is safest because this method only chooses to bet on games that both models agree on for whom to bet. Thus, a more selective group of games is chosen for bets. The Martingale method has 15% of observations at 0, but based on the nature of the bets, which continue to double if the bettor is losing, if the bettor does not lose the entire bankroll, it is a near guarantee to make money. The K1 and Extreme Edge - Simple methods both have extremely long tails and provide riskier investments than the LMER agreement methods, but also higher mean and median returns.

[3] www.investopedia.com/articles/trading/04/091504.asp↩

4.1 Results Using K-Fold Validation

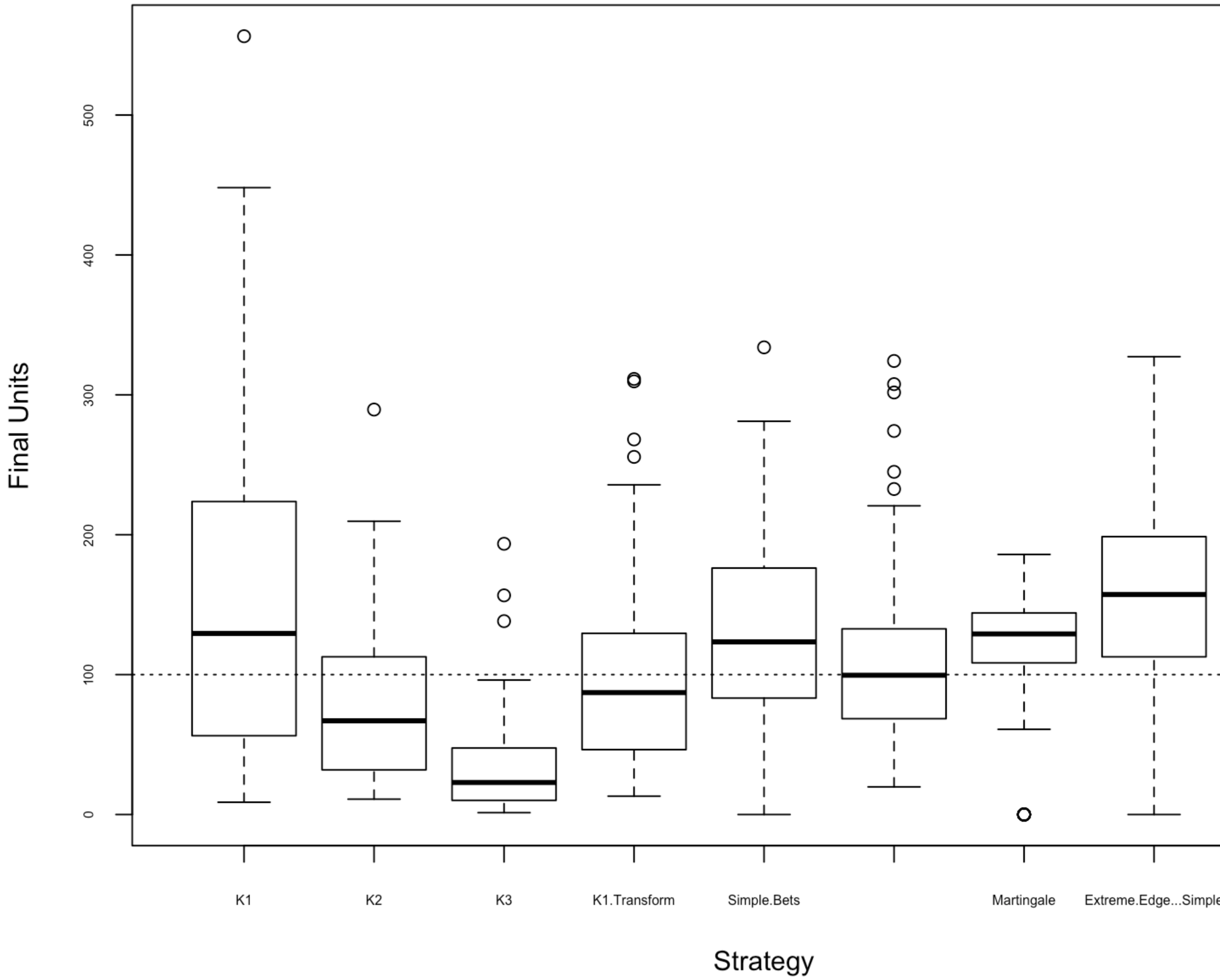

Figure 4.1: Final Bankroll Box Plots by Strategy

For each game in each test data set, the model went through the process outlined above for each betting strategy. After iterating through all the games in the test data set, I was left with the final bankroll from each model for each test data set in the k-fold. With 100 test data sets, I had a vector of 100 final bankrolls for each of the nine betting strategies.

Figure 4.1 is a boxplot looking at the distribution of the final bankrolls for each of the betting strategies over the 100 simulations. Most of the strategies have medians above 100, meaning they generated positive median returns. In addition, these distributions have a boundary on the low end of 0, as the strategy cannot end with a bankroll below 0, but with no maximum, the distributions tend to be right-skewed, leaving the means higher than the medians.

Table 4.2 shows the mean and median final bankrolls over the 100 simulations for the different methods. Four of the methods stand out; the K1 method, Martingale, Extreme Edge - Simple Bets and the LMER Agreement method. Outside of the Martingale method, which is interesting because this is the only strategy where the mean is significantly *lower* than the median, the other listed methods are noteworthy because of the huge returns – each of those three methods have mean returns above 35%.

Table 4.2: Mean Final Bankroll vs. Strategy

	Mean Final Bankroll	Median Final Ba
K1	153.61	1
K2	80.88	
K3	33.55	
K1.Transform	102.01	
Simple.Bets	131.82	1
Extreme.Edge...Transformed	109.19	
Martingale	113.62	1
Extreme.Edge...Simple	150.25	1
LMER.Agreement	137.62	1

Table 4.3 displays the summary statistics for the 4 top methods; the team-specific model using Kelly Criterion for betting amounts, the Martingale method, the Extreme Edge with Simple Bets and the Agreement method. The highest median is this K1 method. All these methods used the team-specific model to generate probabilities of

beating the spread. The returns are very high in each of the quartiles, with positive returns for over 75% of simulations for the Extreme Edge with Simple Bets and Martingale methods. However, these are both the riskiest methods, as in 5% and 15%, respectively for each method, the bettor would have lost the entire bankroll using these methods. This is a very high chance relative to most other more standard investments in the stock market.

Table 4.3: Summary Statistics for the Best Performing Methods

	Min.	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	
Kelly Criterion	8.78	56.61	129.42	153.61	221.56	5
Extreme Edge - Simple	0.00	113.64	157.27	150.25	198.41	3
Model Agreement	20.31	71.41	121.40	137.62	179.07	4
Martingale	0.00	108.52	129.09	113.62	144.09	1

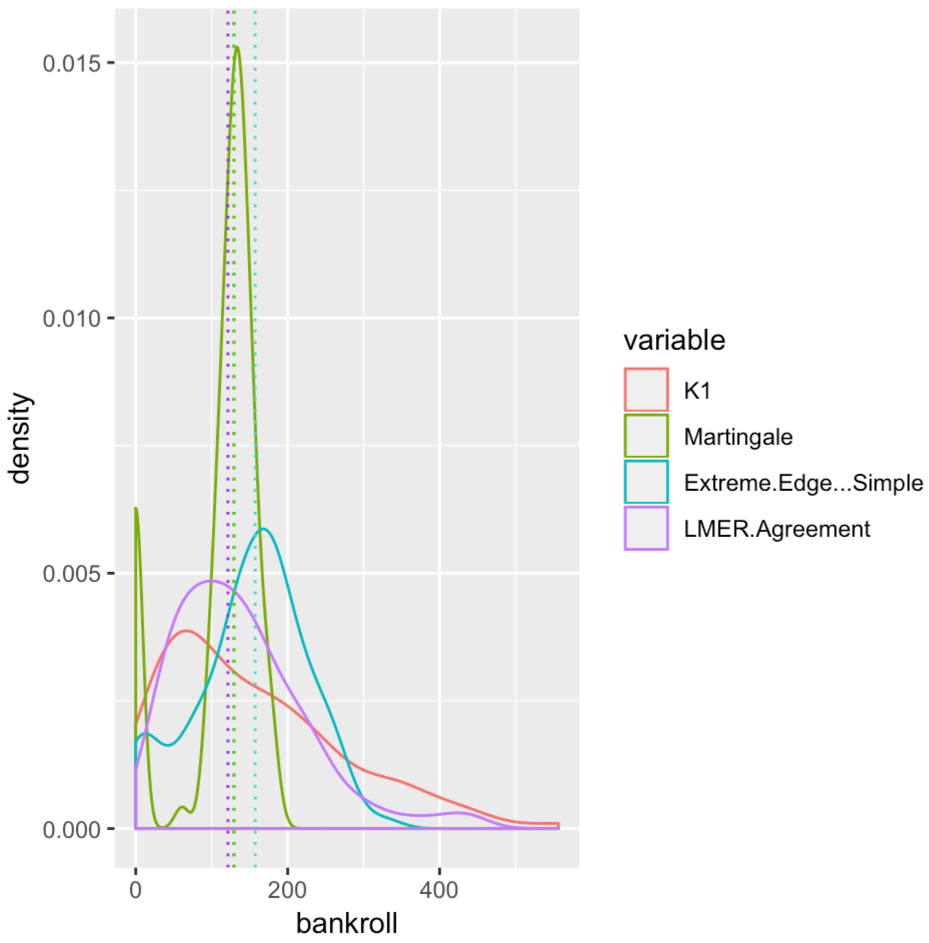

Figure 4.2: Density Plots comparing Four Successful Betting Strategies

Figure 4.2 shows the density plots for the 100 final bankrolls from the top four methods. The dotted lines represent the medians of each distribution. There seems to only be 3 dotted lines, but the medians for both the K1 and Martingale methods are both 129, so these lines are overlaid. The LMER-Agreement method seems to provide the safest betting strategy, as the density plot does not have as long of a tail, but a larger peak close to 175 units. It makes sense that this method is safest because this method only chooses to bet on games that both models agree on for whom to bet. Thus, a more selective group of games is chosen for bets. The Martingale method has 15% of observations at 0, but based on the nature of the bets, which continue to double if the bettor is losing, if the bettor does not lose the entire bankroll, it is a near guarantee to make money. The K1 and Extreme Edge - Simple methods both have extremely long tails and provide riskier investments than the LMER agreement methods, but also higher mean and median returns.

[3] www.investopedia.com/articles/trading/04/091504.asp↩

Chapter 5 Discussion

The goal of this thesis is to create a betting model that provides a statistical basis for choosing the timing, team and amount to bet on a certain game. Through combining a dynamic linear model for the point spread throughout the week with a mixed-linear model for predicting the score difference in the game, I found different betting strategies that generated astronomical average returns.

So, if I was able to create a model with such high average returns, why are there not large funds that specialize in sports gambling? Casinos are not forced to accept a bet, so if a customer keeps coming with massive bets and continues to win, the casino will not accept the bets of these customers. There are difficult and complex ways to circumvent some of the casino staff and place bets, but that is another factor that makes sports gambling quite difficult. For betting small amounts of money, however, the casino is less likely to notice the bets and utilizing this model can be a fun way to increase this cash. However, each of these betting methods – even the safest of them – are extremely risky, with a much larger chance of losing over 50% of your bankroll than just about any other type of typical investment. Human emotion plays a role, as well, as many people will quit when they are down, even though continuing to bet may be the statistically savvy decision. As sports gambling becomes more legalized, more recreational bettors will place bets, and these bets typically lose. Thus, casinos are more likely to accept bets from the “sharper” bettors. One leading prop trading firms, Susquehanna International Group (SIG), already has a “quantitative sports trader” role, and it is likely more firms will follow suit as sports gambling becomes legal in their states (SIG is based in Pennsylvania where sports gambling is legal) and sports gambling becomes destigmatized.

For my future work with this model, after optimizing my arbitrarily chosen parameters, such as the decision point at two-thirds of the way through the week, the 80% confidence interval, my future betting amounts and more, I would like to expand this model to hedge against risk. I want to make this model more applicable to a real person willing to invest their money, and even with advertising the massive average returns, few people would invest money in models that are so risky. Furthermore, this model only provides the basis on when to bet. In order to make this model usable, I need to create a computer program that tracks the current spreads (across multiple markets) and places bets at the first decision point, and then if the spread for any game reaches its key number. Finding ways to hedge risk and creating a computer program to carry out bets are both necessary to implement this model.

Chapter 6 Appendix

6.1 Example Row of Dataframe used for Modeling the Game Score

Table 6.1 shows one row of the dataset used for modeling. For size purposes, I included just general columns for the key variables used in modeling, but there are individual columns for both the home and away teams that show the statistics of each team. The DVOAs listed are percentages. My mixed-linear models utilized only the listed variables, and transformations of the listed variables.

Table 6.1: Example Row of D

Date	Game ID	Home Team	Away Team	Home Score	Away Score	Week	Year	Total DVOA
2017-09-10	INDvLAR	LAR	IND	46	9	1	2017	-5.1

Note:

These stats are available and specified for the home and away team.

6.2 Modeling Number of Observations

Table 6.2: Row of Dataframe used for Observation Point Model

Final Number of Obsv.	Obsv. at Time	Total Ticket Number	Total Cash
188	132	1467	210373

Output of Observation Number Model

Dependent variable:

	Number of Observations
Log(Total Ticket Number)	-9.824
	(1.105)
Log(Total Cash Bet)	-16.211
	(0.938)
Observations up to Point	0.971
	(0.012)
Constant	311.840
	(4.786)

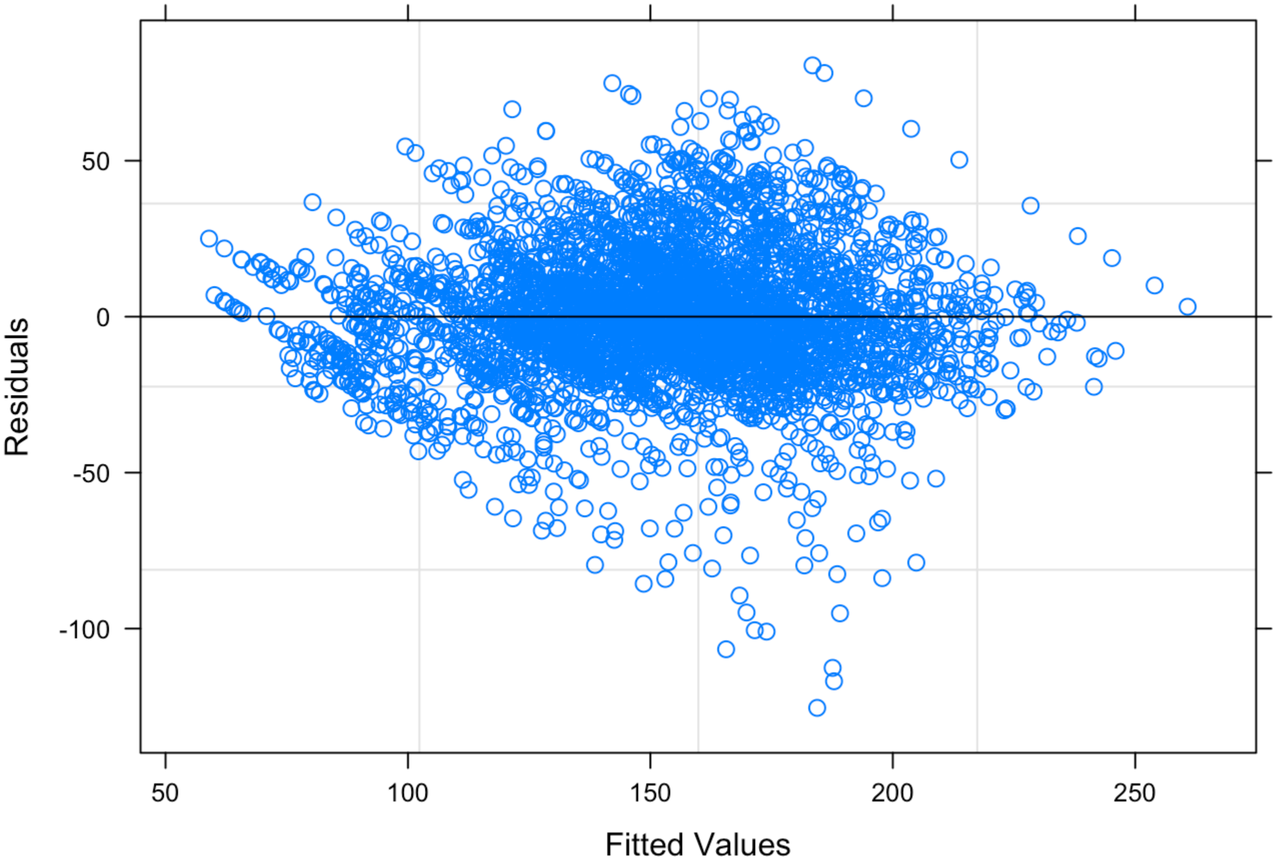

Figure 6.1: Residual Plots for Observation Number

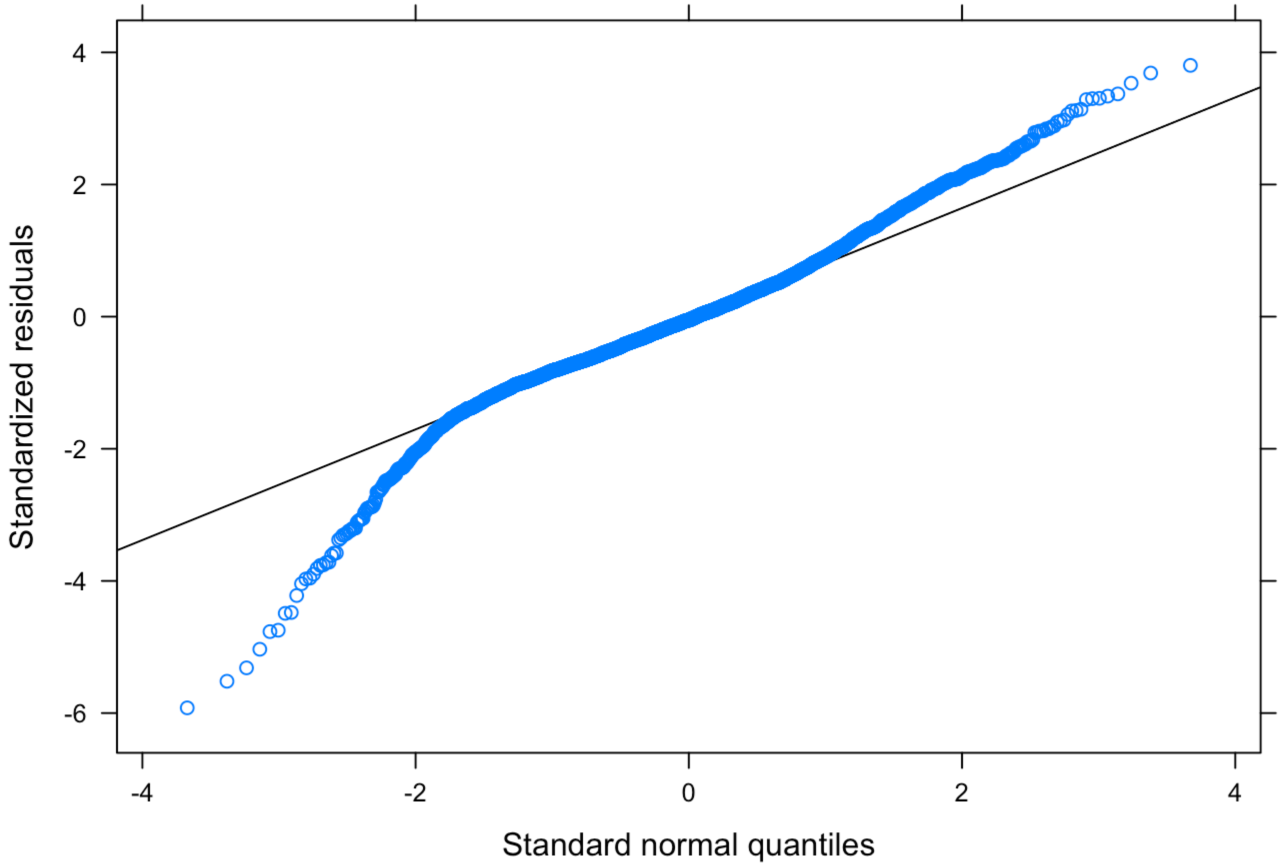

Figure 6.1: Residual Plots for Observation Number

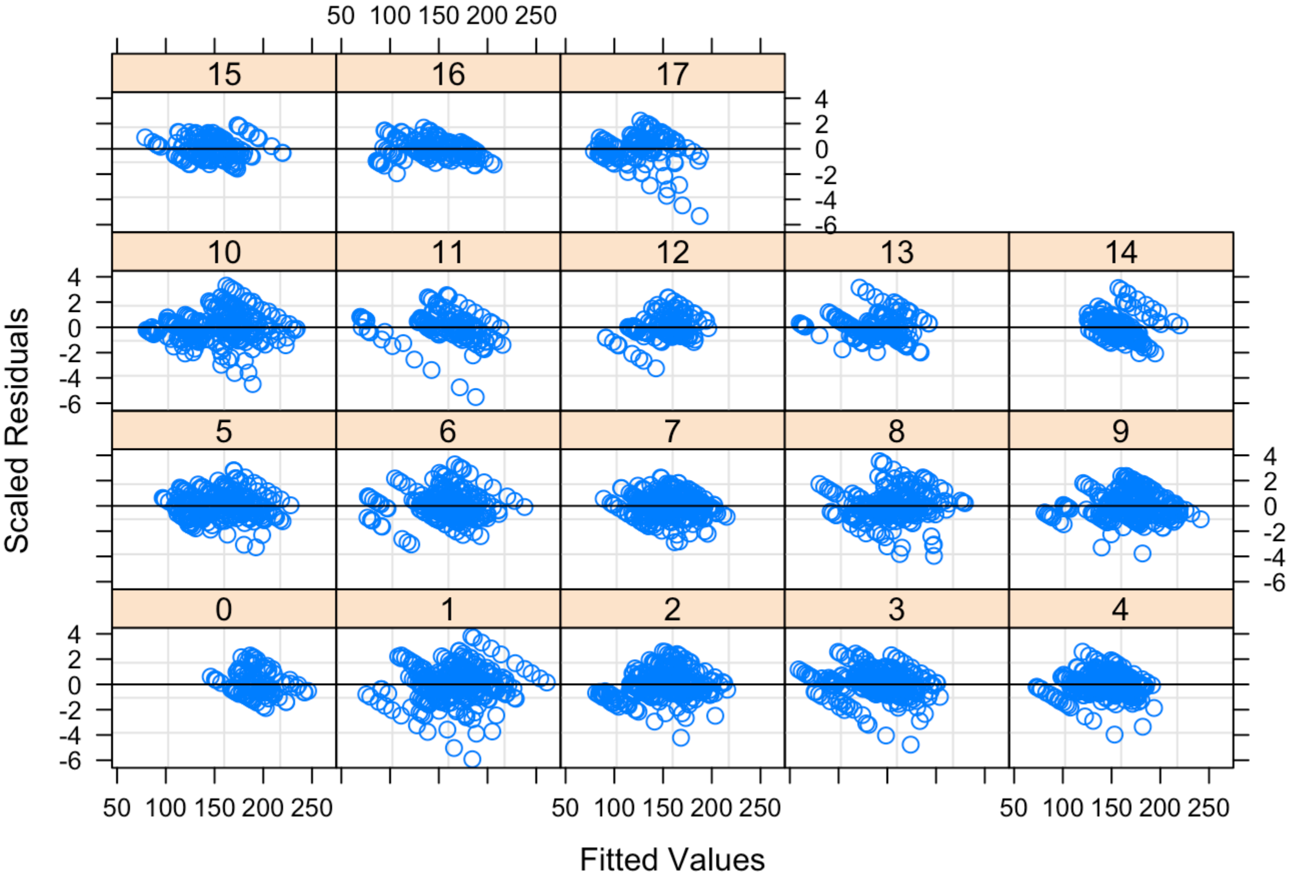

Figure 6.2: Residual Plots for Observation Number by Week

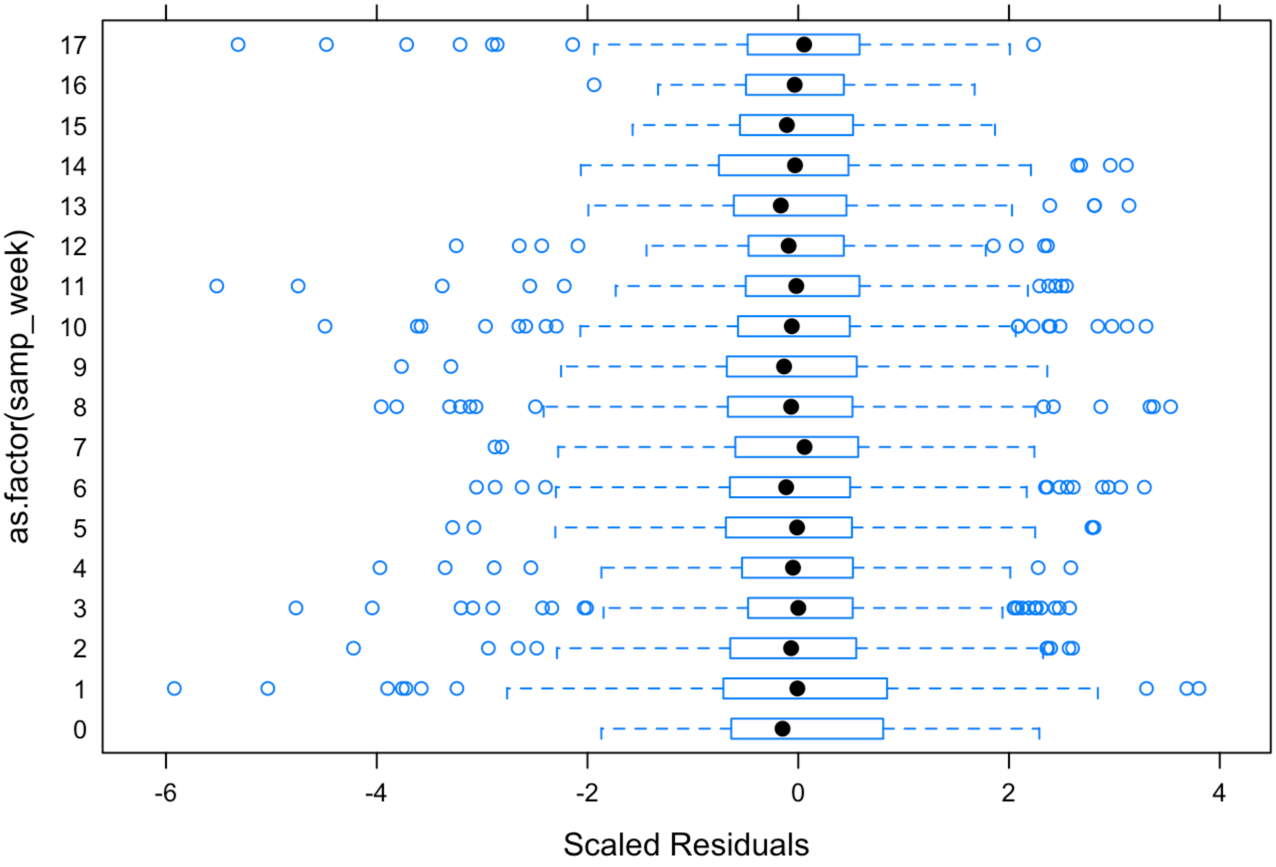

Figure 6.3: Residual Plots for Observation Number by Week

Table 6.2 is an example row of the dataframe used to model the number of observations that will be in a series based on the amount of cash, tickets and observations up to a time tt.
Table **??** displays the parameters for this model, while Figures 6.1 — 6.3 shows the diagnostic plots for this model. The residual plots show that a mixed linear model is a good approach. The residuals, while not perfect, seem to follow the normal distribution and the residuals are relatively evenly distributed for each week – the random effect.

6.3 Example Row of Test Data set with Probabilities

The test data set includes all the same columns shown in 6.1, in addition to the following key columns (and more for each of the different betting strategies). Using the K1 betting strategy for the game between the Indianapolis Colts and Jacksonville Jaguars on November 11, 2018, the spread at the first decision point was the away team, the Jaguars (+3). At this decision point spread, I expect the away team to beat the spread with a proportion of $1-0.444=0.556$1−0.444=0.556. If the Bet Team variable is equal to 1, it means the bet will be on the away team. With an expected value of betting on the Jaguars (+3) of 0.0676, the K1 betting strategy calls for me to bet on the Jaguars. My allotment for this bet is 6.76% of my current bankroll. If the future forecasted spread were to make the game more advantageous to bet on, in terms of expected value, then I would only be one-third of my allotment now. However, the forecasted future spread has the Jaguars (+2.5), which would be a much worse bet. The expected value of the Jaguars (+2.5) is not only lower than our expected value at the first decision point, but it is negative. Because the expected value would lower by betting on the game at the number my DLM forecasts for the spread, I bet my full allotment at Jaguars (+3), and there will be no future bet, regardless of where the spread actually does move. The Future Bet Team column refers to the fact that because the simulated probability is below 0.5, I would bet on the away team (the Jaguars), if I were to make a future bet.

6.4 Diagnostics for Random Effects in Mixed-Linear Models for the Score Difference

Figures 6.4 and 6.5 show the diagnostic plots for the random effects of the first mixed linear model, while Figures 6.6 — 6.8 show the diagnostic plots for the random effect of the second mixed linear model. The residuals seem relatively consistent throughout the groups, and the mixed linear models seem to fit both models appropriately.

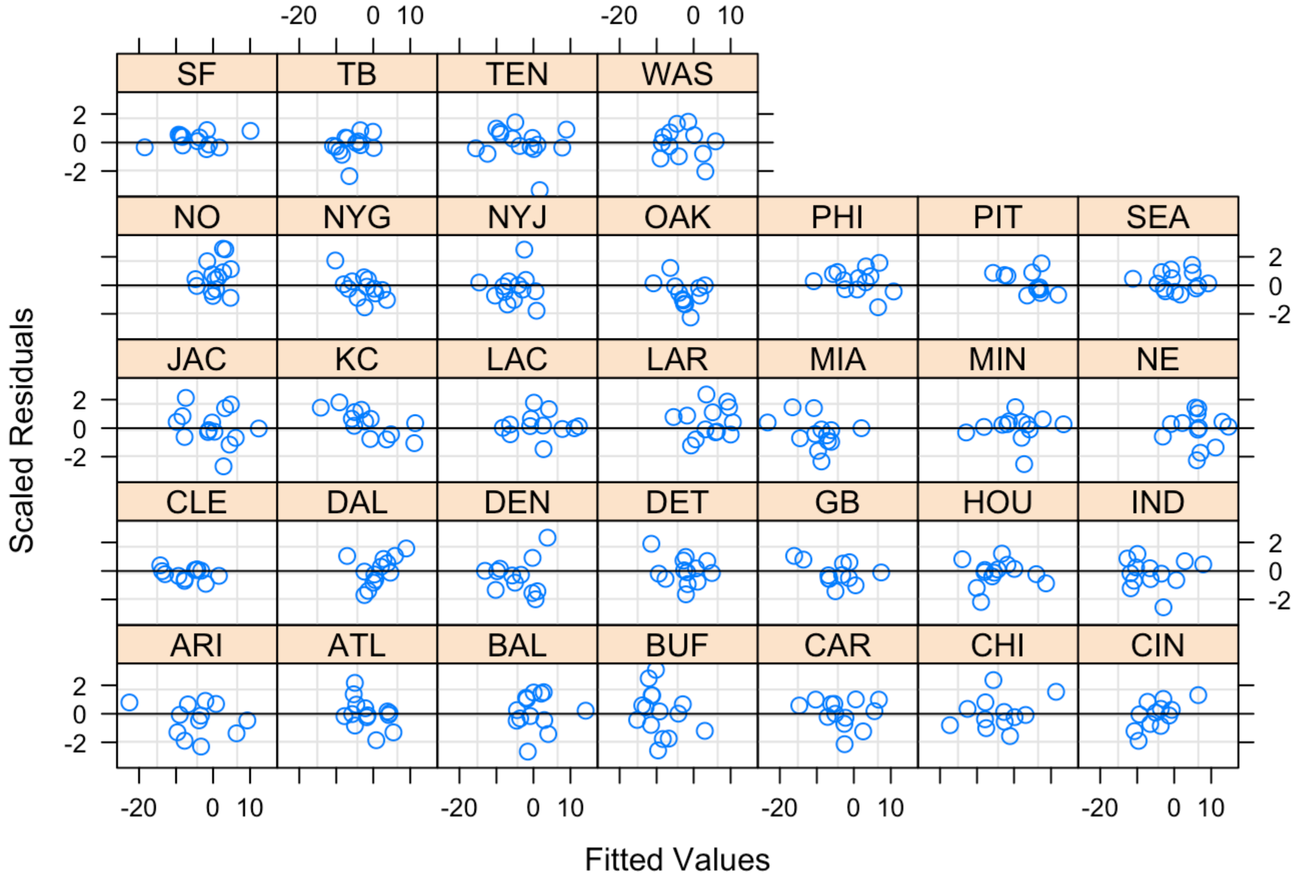

Figure 6.4: Team-Specific Model Diagnostics by Away Team Group

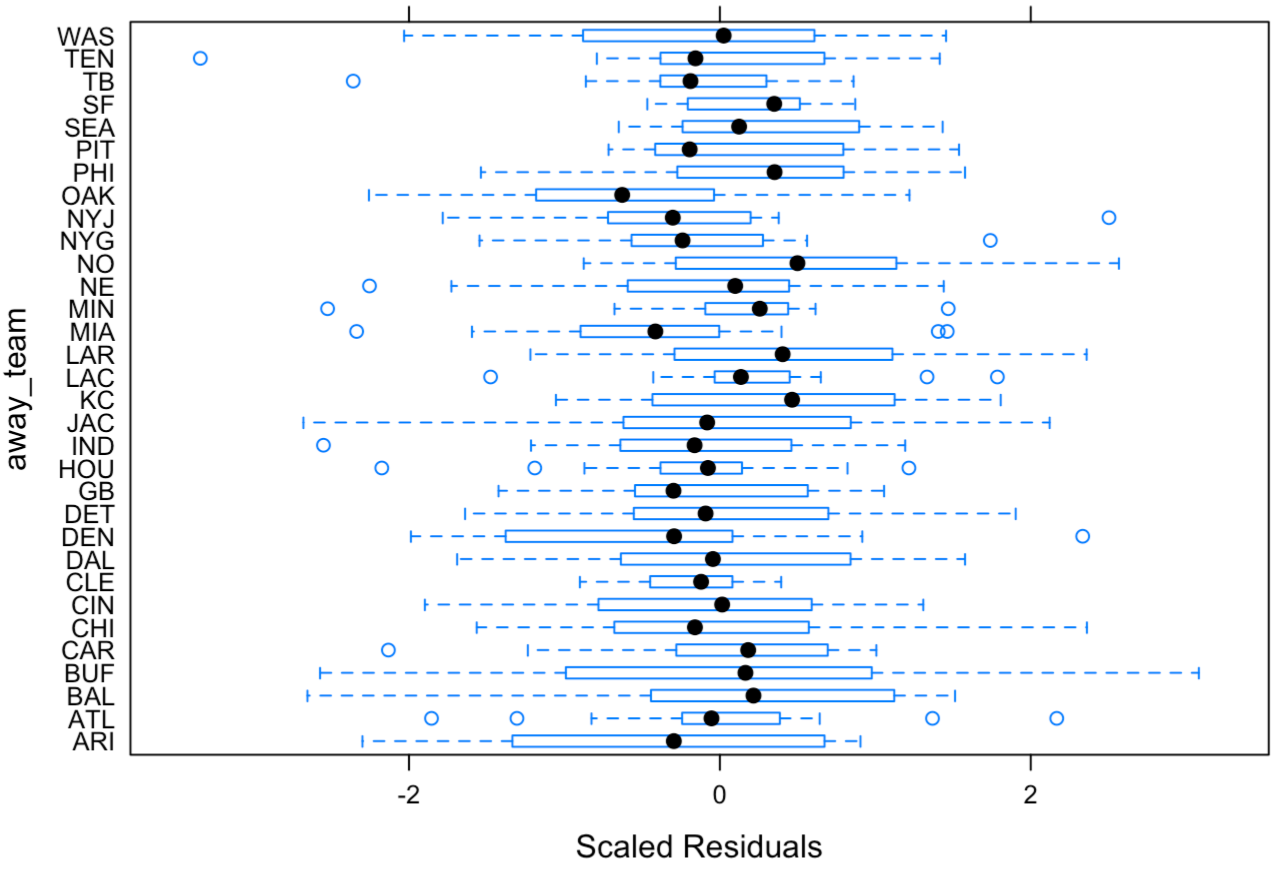

Figure 6.5: Team-Specific Model Diagnostics by Away Team Group

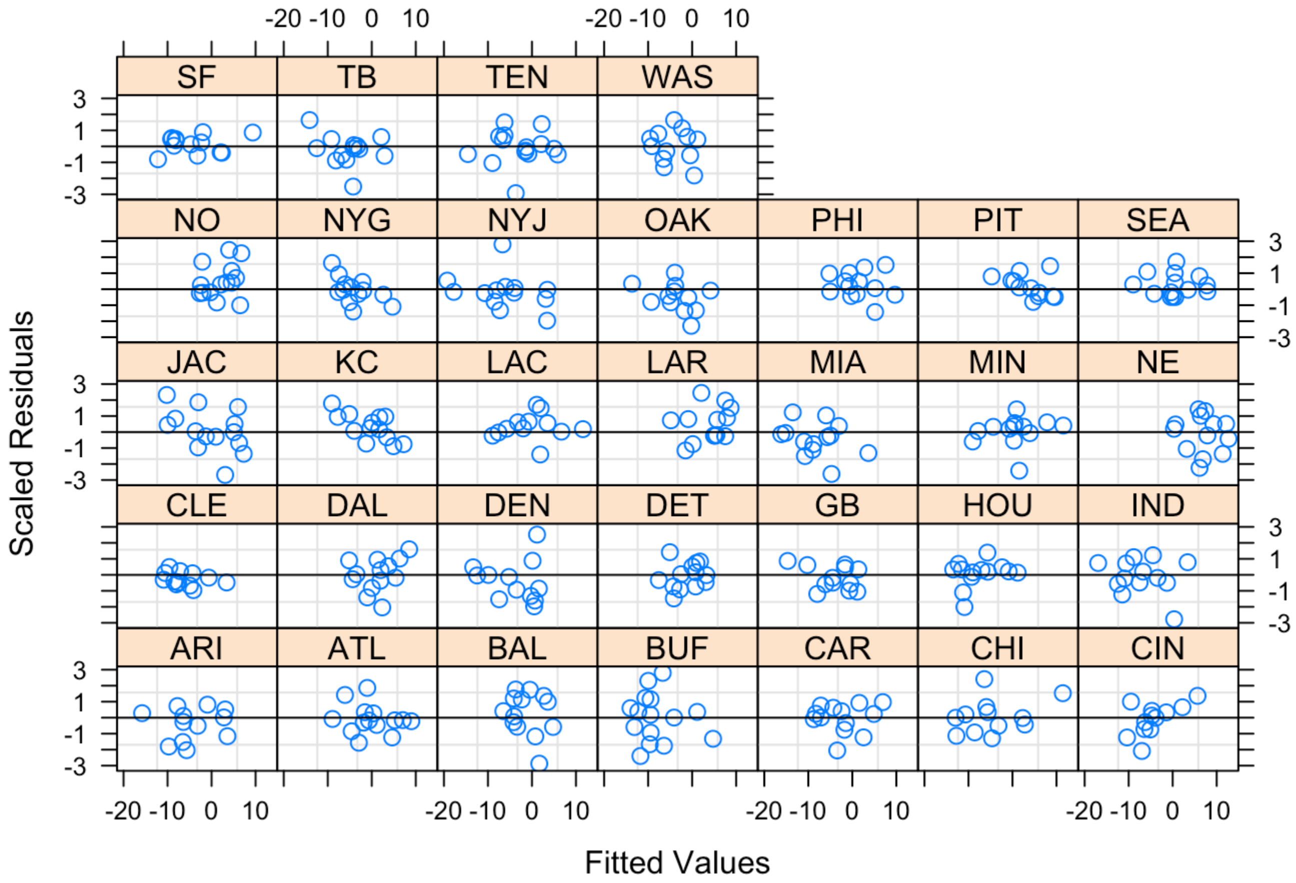

Figure 6.6: Betting-Trend Model Diagnostics by Away Team

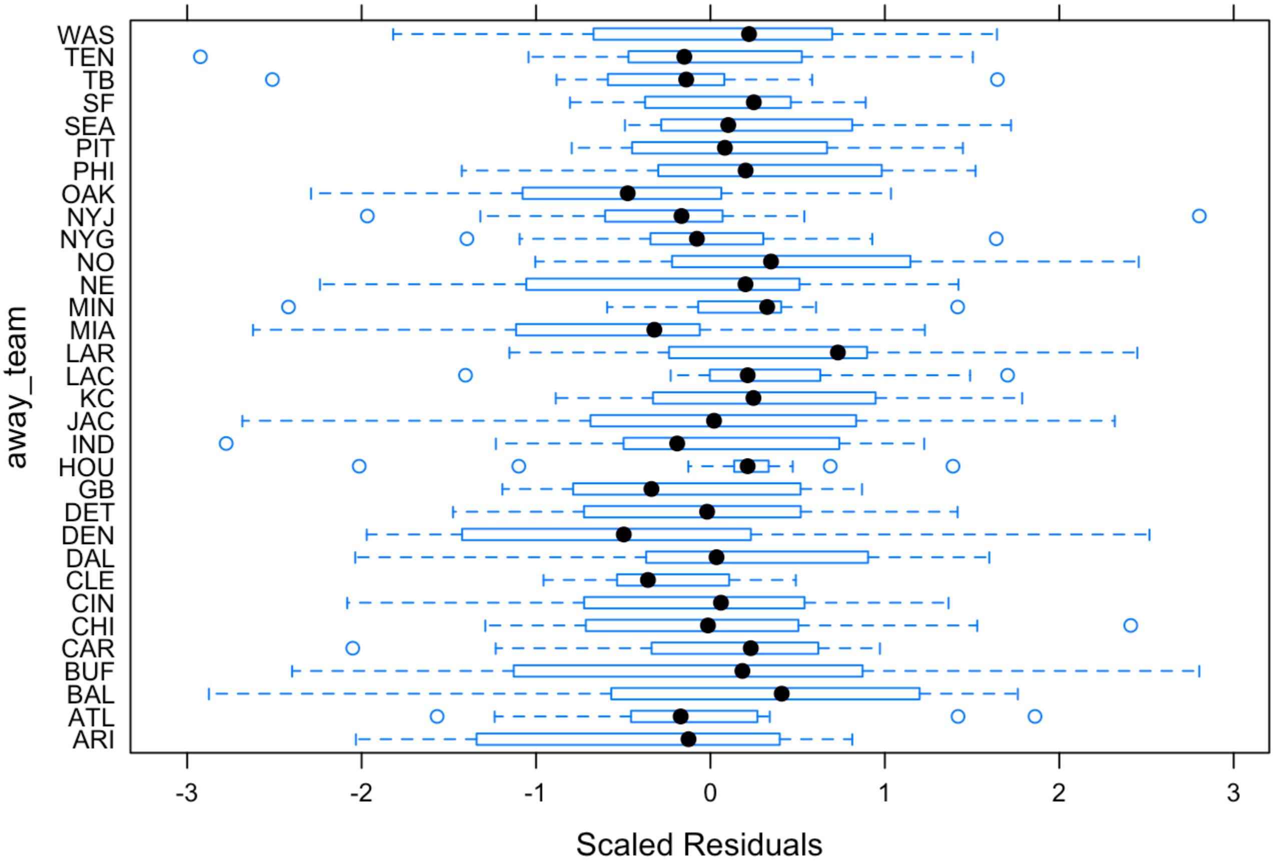

Figure 6.7: Betting-Trend Model Diagnostics by Away Team

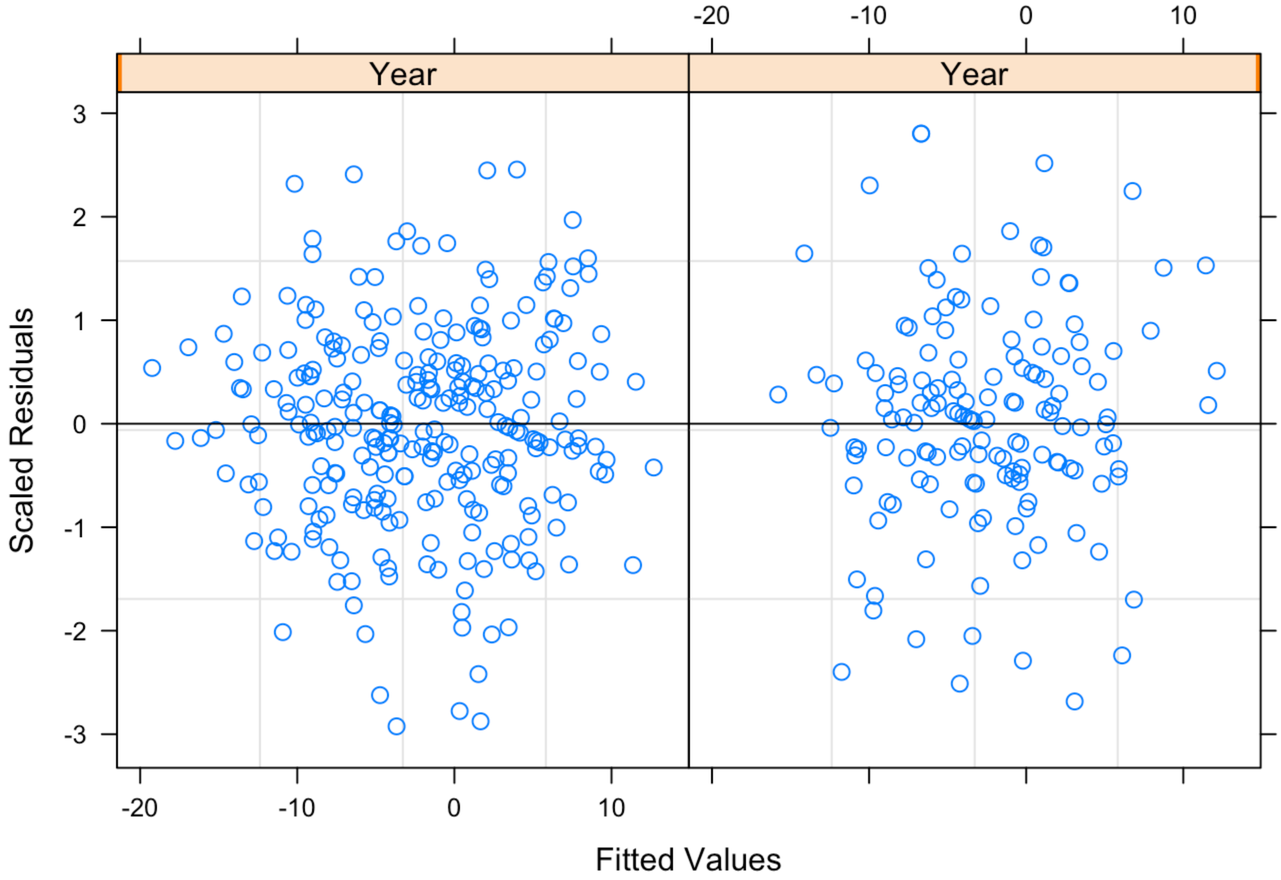

Figure 6.8: Betting-Trend Model Diagnostics by Year

Chapter 7 References

1. Kuepper, Justin. “Using the Kelly Criterion for Asset Allocation and Money Management.” *Investopedia*, 27 Mar. 2019, www.investopedia.com/articles/trading/04/091504.asp.
2. “Point Spread Betting Strategy - How Point Spread Bets Works.” *The Sports Geek*, www.thesportsgeek.com/sports-betting/strategy/point-spread/.