Models for NHL Team Statistics

by

Eric Foote

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Bachelor of Science Mathematics and Statistics

In the Graduate Academic Unit of

Supervisor(s):	Connie Stewart,Department of Mathematics and Statistics
	Tim Alderson,Department of Mathematics and Statistics
	Orla Murphy,Department of Mathematics and Statistics
Examining Board:	
External Examiner:	

This thesis is accepted

Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

November, 2018

Abstract

In this discussion I use multiple linear and ridge regression techniques to predict National Hockey League(NHL) team points for the previous season of 2017-2018 as a method to predict overall team success. I also use logistic regression to predict the probability that a team makes the playoffs; this method can also be used to predict overall team success. For this exploration, 24 variables were used, and for the AIC prediction model, goals against per game was an influential variable. For the p-value model, out-shooting the opponent regardless of a win or loss increases the accuracy of the prediction. Finally, the logistic regression model predicted nearly every playoff team for the previous season.

Table of Contents

Abstract **i**

Table of Contents **ii**

List of Tables **iii**

List of Figures **1**

1 Introduction **2**

2 Data **4**

3 Methods **8**

4 Results **13**

5 Conclusion **28**

List of Tables

2.1 Table of Variables 6

4.2 AIC Prediction 15

4.4 p value Prediction 17

4.6 Ridge Regression Prediction 23

List of Figures

4.1 AIC Error Plot 18

4.2 P Value Error 18

4.3 Ridge Trace 21

4.4 Ridge Prediction Error 24

1 42

2 Residuals vs Fitted Values for AIC 49

3 QQ Plot for AIC 49

4 Scale-Location for AIC 50
5 Residuals vs Leverage for AIC 50
6 Residuals vs Fitted Values for R Squared 51
7 QQ Plot for R Squared 51
8 Scale-Location for R Squared 52
9 Residuals vs Leverage for R Squared 52

Chapter 1

Introduction

The NHL as a league is over 100 years old[4]; however, "hockey statistics is in its infancy".[1] Every year, publications, news outlets and blogs[3] publish articles about projected team standings. While most publications strictly consider roster makeup and coaching systems,[5] an ever-growing group of people are considering, for example, the "individual impact on NHL 5v5 Shot Rates"[6] and how this would affect a game played, and are then running a simulation of the game.[7] My work is not going to be undertaking the same level of detail; however, I am going to be predicting NHL team points for the season of 2017-2018 and, when appropriate, whether a team will make the playoffs. This serves the purpose of predicting overall team success. My reasoning for these distinctions will be expanded upon during the discussion on how the methods for model building were applied. The models that will be explored are linear and ridge regression, a logistic model will then be

built. The discussion will begin with an examination of the data collection technique, and an exploration of the variables. Following that we look at the theory behind the methods listed earlier as well as the implementation. Finally, we discuss our predictions.

Chapter 2

Data

Data Collection

The data that we used for this analysis comes from an application programmer interface, (API) https://statsapi.web.nhl.com/api/v1, that has a variety of modifiers used to collect our desired data. The data is stored in javascript object notation (JSON). Therefore, in order to collect the data we have to make calls to the API, which requires the R JSON package in R. Rather then showing the entire R code for this process, I show how I obtained the data for New Jersey Devils. This code appears in Appendix A. This data collection technique has several steps. First, we use the function fromJSON and we send to it the website with the modifier /teams/i/stats?season=20072008, where i is an unique number corresponding to the team and 20072008 corresponds to a specific season; this modifier can be changed to specify any team and season that we wish to examine. This process creates a data frame

for each individual year. The second step is to extract the raw data. To do so we use the command $stats[[1]]$splits[[1]]. This allows us to extract the variables that we are going to be using. In Appendix B there are examples of the uncleaned data. In Appendix C is the code for cleaning the data, and what the cleaned data looks like. This process was repeated for all 31[1] teams in the NHL. We are now going to explain the process of cleaning the data set.

Cleaning of the Dataset

One year of data, the 2012-2013 season was removed from the overall dataset. This was due to that season having been shortened to 48 games because of a lockout between the National Hockey League Players Association(NHLPA) and team owners. Our next step was to remove variables that are highly correlated. To do so we look at a heat map (see Appendix D). We see that the variable *NHL team points* is strongly correlated (≥ 0.90) with losses, wins, point percentage, and overtime losses; these latter four variables are therefore removed. Lost face offs are also removed due to there already being face offs won in the data set. We now have our cleaned dataset of 270 observations. The 24 variables are listed in a table below.

[1]The thirty first team Vegas Golden Knights only needed data from the inaugural season of 2017-2018

Variable	Description
points	Points are calculated as follows $pts = 2 * win + 1 * ot$.
goalsPerGame	average number of goals a team scores per game
goalsAgainstPerGame	average number of goals a team gives up per game
EvGAARatio	goals the goalie gives up per "even strength regulation time".[8]
PowerPlayPercentage	percentage of power play goals a team scores
PowerPlayGoals	the number of power play goals a team scores
Penalty Kill Percentage	percentage of power plays that a team does not get scored on
Power Play Goals Against	number of power play goals against
Power Play Opportunities	number of power plays a team got during the season
ShotsPerGame	average number of shots on goal
ShotsAllowed	average number of shots that a team gives up
WinScoreFirst	number of games won when a team scored first
WinOppScoreFirst	number of games that a team won when it did not score first
winLeadFirstPer	number of wins when a team had the lead after first period
winLeadSecondPer	number of wins when a team had the lead after second period
winOutshootOpp	number of wins when a team had led in shots
winOutshotByOpp	number of wins when a team did not lead in shots
ShootingPctg	the percentage of shots that resulted in goals scored
savePctg	percentage of shots against that did not result in goals
faceoffsTaken	number of faceoffs taken
faceoffsWon	number of faceoffs won
faceoffWinPercentage	percentage of faceoffs won
madePlayoffs	1 if a team made playoffs and 0 if a team missed

Table 2.1: Table of Variables

Additional Variable Definition

We require one further variable that does not come from the API; this variable is *made playoffs*, a quantitative variable that takes on the values of 1 or 0. If a team makes the playoffs, then *made playoffs* is 1, otherwise, *made playoffs* is 0. This variable was hard coded into R; the website https://www.hockey-reference.com/playoffs/ was used to determine which teams made the playoffs in a given year. Our next step was to establish the theory behind the models we created.

Chapter 3

Methods

This section provides the theoretical basis for the models we built. We begin with multiple linear regression, discuss the method used for variable selection for one of the models, ridge regression and logistic regression.

Multiple Linear Regression

The multiple linear regression model expresses the mean of response variable Y as a function of one or more distinct predictor variables $x_1, ..., x_k$. It takes the form:

$$\mu_{Y|_{x_1 ... x_k}} = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k,$$

which can be rewritten as:

$$Y|_{x_1 ... x_k} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_i + \epsilon_i \text{ where } i = 1, 2, 3, ..., n.$$

Here, the β_k's are our model parameters, and ϵ_i is the *residual*, the difference between the predicted value and the mean value.

It is common to express the second equation in vector form; which allows us to use matrix algebra to estimate the β_k values.

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ ... \\ Y_n \end{bmatrix} \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ ... \\ \beta_k \end{bmatrix} \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ ... \\ \epsilon_k \end{bmatrix}$$

We define an additional matrix, x, which is an $n \times (k+1)$ matrix with the first column containing all 1's and the other k columns consisting of the values of the predictor variables.

$$x = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{31} & ...x_{k1} \\ 1 & x_{12} & x_{22} & x_{32} & ...x_{k2} \\ ... & ... & ... & ... & ... \\ 1 & x_{1n} & x_{2n} & x_{3n} & ...x_{kn} \end{bmatrix}$$

Now we consider the assumptions for multiple linear regression, which are:

1. $\mathrm{E}(\epsilon) = \hat{0}$ and $var(\epsilon) = E(\epsilon * \epsilon^T) = \sigma^2 I$.

2. The ϵ_i are independent $\forall i$.

3. Multicollinearity, which means that the predictor variables are not highly correlated with each other.

4. All of the variables are normally distributed.

AIC Variable Selection

The method selected to determine which of the 24 variables are included in the model was "The Akaike Information Criterion (AIC)"[10]. This method attempts to select a model that has a small number of variables, but fits the dataset. The equation to determine this is:

$$AIC_p = nln(\frac{SSE_p}{n}) + 2p$$

We desire to minimize AIC_p. An advantage of AIC is that it allows us to compare models that do not contain the same variables. To use AIC we cannot have any missing values; we are fortunate in this regard because our data set has no missing values. Our next step is to discuss ridge regression techniques.

Ridge Regression

Ridge regression is another technique for predicting values; it is most effective when the predictor variables are highly collinear. We first need to define the ridge estimators, $\hat{\theta}(k)$, (Hoerl and Kennard[10]) $k \geq 0$:

$$\hat{\theta}(k) = (Z^T Z + kI)^{-1} Z^T Y$$

These estimators are biased, and usually have a small mean-squared error. If we have the standard form of a regression equation.

$$\hat{Y} = \theta_1 \hat{X}_1 + \theta_2 \hat{X}_2 + ... + \theta_p \hat{X}_p,$$

then the system of equations used for estimating ridge regression coefficients is

$$(1+k)\theta_1 + r_{12}\theta_2 + ... + r_{1p}\theta_p = r_{1y}$$
$$r_{21}\theta_1 + (1+k)\theta_2 + ... + r_{2p}\theta_p = r_{2y}$$
$$r_{p1}\theta_1 + r_{p2}\theta_2 + ... + (1+k) = r_{py}.$$

Here, r_{ij} is the correlation between predictors i and j and r_{iy} is the correlation between predictor i and response variable $\hat{Y}$. As k increases, the bias increases. If k increases dramatically, then the regression estimates tend to zero. We want to pick a value of k for which the variance is greater than the bias. Usually k is chosen by computing $\hat{\theta}_1, ..., \hat{\theta}_p$ for k between 0 and 1. We then take these values for $\hat{\theta}_1, ..., \hat{\theta}_p$ and plot them against k. This graph is known as the ridge trace, and is used to select k. Now that we have our basis for ridge regression; our next step is to discuss logistic regression.

Logistic Regression

Logistic regression is used when the response variable is qualitative. The logistic model can be expressed in the general form as:

$$P(Y = 1|X_1 = x_1, ..., X_p = x_p) = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k)}}{1+e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k)}}$$

This equation is *the logistic regression equation*, estimating $P(Y = 1)$. For our investigation we are going to be predicting the probability that a team makes the playoffs. The assumptions for logistic regression are: a qualitative

predictor; a large sample size; very little multicollinearity; independent observations, and a linear relationship between the predictors and the natural log of the probability. Now that we have the theory behind our models, we can discuss the results of our implementation.

Chapter 4

Results

This section is broken up into the linear models which predict points, and the logistic regression which predicts whether or not a team makes the playoffs.

Multiple Linear and Ridge Regression

Multiple Linear Regression

We begin with the linear regression models. We employed two different techniques for variable selection. The first was to use AIC on the model containing every variable to reduce it. The second model was to utilize the p values from the full model to create a smaller model(see Appendix G for the plots for these models). The AIC model is:

$$points = 11.8895(goalsPerGame) - 14.2963(goalsAgainstPerGame) +$$
$$0.1168(powerPlayPercentage) + 24.9296(winScoreFirst) +$$

$$26.3916(winOppScoreFirst) - 10.1651(winLeadSecondPer) +$$
$$24.6731(winOutshootOpp) + 23.2772(winOutshotByOpp) +$$
$$0.1465(faceOffWinPercentage) + 48.08239$$

Using the AIC variable selection method to predict the 2017-2018 season generated the following values:

Team	Actual Points	Fit	Lower Bound	Upper Bound
ANA	101	100.75284	99.97375	101.53194
ARI	70	69.95098	68.84575	71.05622
BOS	112	113.22077	112.29306	114.14848
BUF	62	61.34851	60.21012	62.48689
CGY	84	83.89612	83.32665	84.46558
CAR	83	78.64918	77.28295	80.01540
CHI	76	76.55805	75.56548	77.55063
COL	95	92.92405	91.32146	94.52664
CBJ	97	97.54118	96.82266	98.25970
DAL	92	95.47584	94.64515	96.30652
DET	73	70.79009	69.92879	71.65140
EDM	78	78.45955	77.35247	79.56662
FLA	96	92.36763	91.51401	93.22125
LAK	98	103.80054	102.81034	104.79075
MIN	101	100.14622	99.46543	100.82701
MON	71	70.25493	69.04528	71.46458

NAS	117	112.61700	111.66045	113.57356
NJD	97	94.20401	93.11517	95.29285
NYR	80	77.95396	76.49314	79.41479
NYI	77	77.25902	76.01372	78.50432
OTT	67	68.38637	67.32877	69.44398
PHI	98	95.86227	94.90189	96.82265
PIT	100	100.11463	98.75588	101.47339
SJS	100	99.20700	98.50750	99.90650
STL	94	92.63031	91.82807	93.43255
TBL	113	114.26413	112.97819	115.55006
TOR	105	104.10971	102.95802	105.26140
VAN	73	74.11990	73.02802	75.21179
VGK	105	105.72077	104.52086	106.92069
WIN	114	111.38760	110.38272	112.39248
WSH	109	109.24877	108.09746	110.40008

Table 4.2: AIC Prediction

The p value model is:

$$points = 35.603(winScoreFirst) + 36.087(winOppScoreFirst) +$$
$$44.795(winOutshootOpp) + 42.294(winOutshotByOpp) + 12.525$$

For the 2017-2018 season, the p value model provided the following values:

Team	Actual Value	Fit	Lower Bound	Upper Bound
ANA	101	99.39386	98.76494	100.02278
ARI	70	66.88440	65.62004	68.14876
BOS	112	112.15517	111.09914	113.21120
BUF	62	61.61019	60.34175	62.87862
CGY	84	86.50087	85.89007	87.11167
CAR	83	76.55718	75.21138	77.90297
CHI	76	74.48080	73.50540	75.45621
COL	95	91.48654	90.27831	92.69478
CBJ	97	99.37009	98.56669	100.17348
DAL	92	95.55320	94.54216	96.56423
DET	73	68.86493	67.86799	69.86188
EDM	78	81.65953	80.83764	82.48141
FLA	96	94.57227	93.80385	95.34068
LAK	98	102.03421	100.72097	103.34744
MIN	101	100.17189	99.49410	100.84968
MON	71	69.64649	68.37613	70.91685
NAS	117	110.43911	109.43556	111.44266
NJD	97	97.26480	96.42495	98.10466
NYR	80	79.71466	78.88915	80.54018
NYI	77	79.78500	78.96401	80.60599
OTT	67	70.26269	69.18237	71.34301

PHI	98	94.04029	93.18841	94.89217
PIT	100	101.05045	100.16146	101.93943
SJS	100	99.11771	98.37639	99.85902
STL	94	94.61598	93.76003	95.47192
TBL	113	116.02619	114.91506	117.13731
TOR	105	101.56422	100.61952	102.50893
VAN	73	72.16710	70.95580	73.37839
VGK	105	110.72874	109.69436	111.76313
WIN	114	110.03455	109.12397	110.94514
WSH	109	109.70151	108.43704	110.96598

Table 4.4: p value Prediction

Now that we have our predictions, we compare them to the actual points. Below are box plots for each model's error (see Appendix E for a table of error values; see figure 4.1 for the AIC prediction error; 4.2 for the p value error).

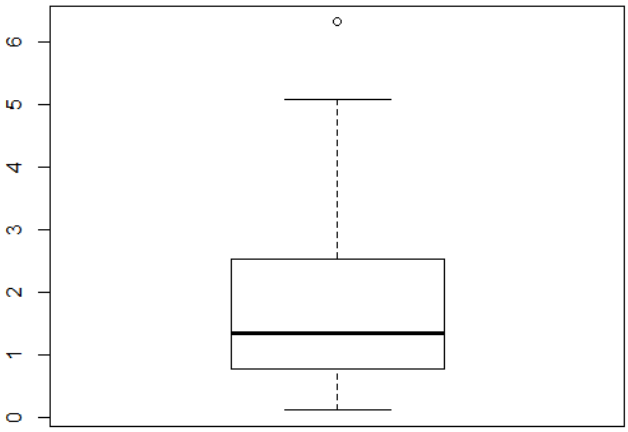

Figure 4.1: AIC Error Plot

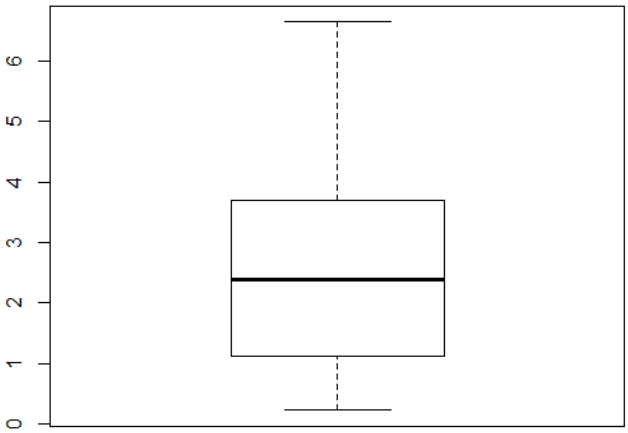

Figure 4.2: P Value Error

We now look at the model plots for the two methods; (see Appendix G for the plots). For the AIC models residual versus fitted value plot we can see that a linear relationship is a reasonable assumption, and that there are two outliers. For the normal $q-q$ plot, we can see that most of the points are on the line, so normality is assumed. The scale-location plot tells us that a linear relationship is a reasonable assumption. The residual versus leverage plot shows no Cook's Distance anywhere; therefore, everything was approximated well, with no significantly influential points. The p value residual versus fitted value plot shows us that a linear relationship is a reasonable assumption, but we have 3 outliers. The normal $q-q$ plot tells us that normality is assumed. The scale-location plot shows us that a linear relationship is an reasonable assumption. However, the residual versus leverage plot tells us that we have an influential point. The majority of the predictions made using the AIC model were one to two points off the actual value. The most extreme error occurred in the case of the Los Angeles Kings: the prediction was five points off what the team actually achieved. One reason for this is that LA had the lowest goals against per game in the NHL during this season at 2.890 goals against. In the model we have $-14.2963(goalsAgainstPerGame)$, which would have less weight for Los Angeles thus raising the predicted value. Another reason is that LA was ninth in wins when the opponent scored first at 0.4 or 40%, which is another variable with a heavy weight in our prediction function. The second most extreme error was in the case of the Nashville Predators. The prediction was four points off the actual value; one reason

for this is that, similar to LA, the Predators were second in goals against per game at 2.488. Other statistics of LA lend to the conclusion that good defensive teams, that limit their goals against, tend to have more points. According to this model. LA was 16th in goals per game; 18th in power play percentage; 7th in wins when the team scored first; 9th when the opponent scored first; 11th in wins when leading after the second period; 7th in wins when out shooting the opponent; 21st in wins when out shot and finally 17th in face off win percentage. For the p value model, most estimates were off significantly; up to almost 4 points. This is really due to the model lacking no deduction for poor defensive play (i.e. $-14.2963(goalsAgainstPerGame)$ and $-10.1651(winLeadSecondPer)$). The model also rewards teams for having lots of shots. For example, the Boston Bruins were first in wins when the opponent scored first, and third in wins when you have more shots then your opponent. This increased the value of their prediction. Therefore, this model is a better exploratory model then a predictor model. This model can be used to explore the effect of scoring and shooting more frequently on overall points. The next step is to talk about the ridge regression models.

Ridge Regression

For the ridge regression model R calculated 99 different k values(or λ as R calls it) and fit all 24 variables to it. This created the ridge trace below.

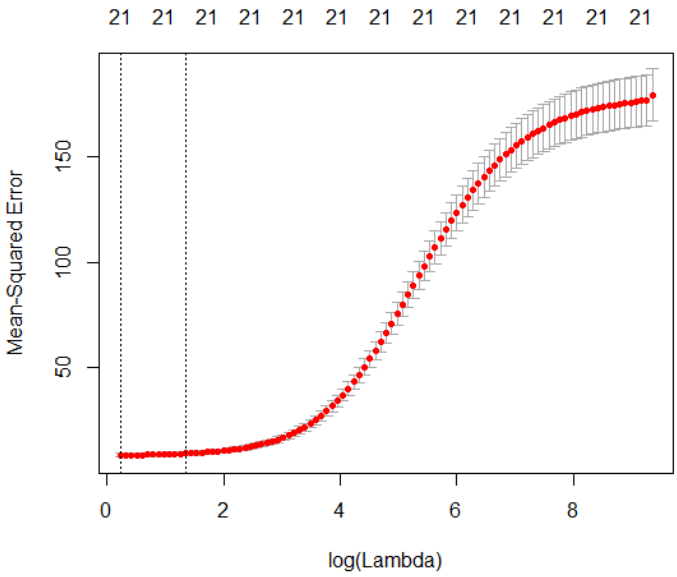

Figure 4.3: Ridge Trace

This method had the predictions:

Team	Actual Value	Points
ANA	101	99.06189
ARI	70	71.28012
BOS	112	112.35062
BUF	62	63.93565
CGY	84	84.82366
CAR	83	79.65816
CHI	76	77.87249
COL	95	94.57622
CBJ	97	96.59796
DAL	92	95.01542
DET	73	72.94090
EDM	78	80.39590
FLA	96	93.44379
LAK	98	103.62926
MIN	101	99.55909
MON	71	72.73987
NAS	117	110.75469
NJD	97	94.49269
NYR	80	80.68079
NYI	77	80.14972
OTT	67	70.15000

PHI	98	95.42240
PIT	100	99.72449
SJS	100	99.38607
STL	94	93.23966
TBL	113	112.73388
TOR	105	104.18809
VAN	73	76.84935
VGK	105	105.25785
WIN	114	111.00241
WSH	109	106.87331

Table 4.6: Ridge Regression Prediction

Next we need to compare the predictions to the actual points. Below this are box plots for the model error (see Appendix E for a table of error values).

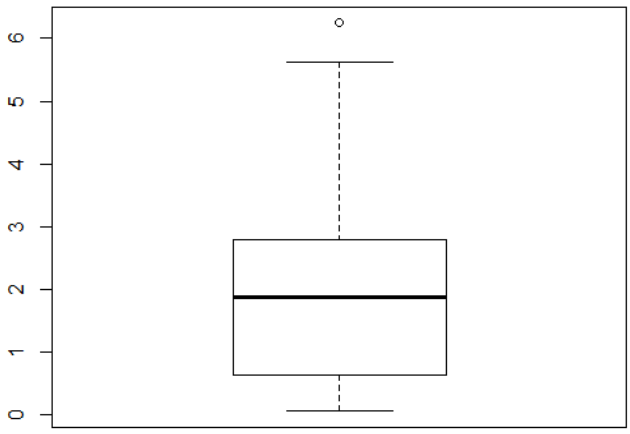

Figure 4.4: Ridge Prediction Error

The most extreme error in this model is for the Nashville Predators, 6 points off the actual point value. Since the ridge regression model takes every variable in our data set, the error is largely due to them being in the top 15 for almost every variable in our set. The ridge regression model weighs every statistic evenly. The next step is to talk about our logistic regression model.

Logistic Regression

For the logistic regression models, we are predicting the probability that a team makes the playoffs. For this model we use every variable, rather than using AIC. The reason for this is AIC would default to the model containing

just the intercept every single time, regardless of whether we use forward selection, backward selection or stepwise selection. Therefore our model is:

$$madePlayoffs = -596.2016 + 0.6785186(pts) +$$
$$17.31825(goalsPerGame) - 6.797520(goalsAgainstPerGame) -$$
$$18.07236(evGAARatio) - 0.1268944(powerPlayPercentage) -$$
$$0.03467402(powerPlayGoals) - 0.1270233(powerPlayGoalsAgainst) -$$
$$0.004215109(powerPlayOpportunities) -$$
$$0.04239098(penaltyKillPercentage) - 0.3554432(shotsPerGame) -$$
$$16.05426(winLeadFirstPer) + 3.368749(winLeadSecondPer) +$$
$$3.367920(winOutshootOpp) + 12.48498(winOutshotByOpp) +$$
$$0.07043268(faceOffsTaken) - 0.1464181(faceOffsWon) +$$
$$7.594018(faceOffWinPercentage) - 1.656188(shootingPctg) +$$
$$214.9347(savePctg)$$

Below is the table of probabilities that a team makes the playoffs:

Team	Probabilities
ANA*	9.999420e-01
ARI	1.371664e-10
BOS*	1.000000e+00
BUF	8.768418e-13
CGY	3.308281e-04
CAR	3.893547e-04
CHI	6.972501e-07
COL*	4.071688e-01
CBJ*	9.659823e-01
DAL	8.537965e-01
DET	5.600266e-09
EDM	5.735255e-07
FLA	7.984136e-01
LAK*	9.981411e-01
MIN*	9.990949e-01
MON	1.113290e-10
NAS*	1.000000e+00
NJD*	9.873069e-01
NYR	4.818513e-06
NYI	4.298577e-06
OTT	3.435248e-12

PHI*	9.995894e-01
PIT*	9.995588e-01
SJS*	9.989304e-01
STL	5.896700e-01
TBL*	1.000000e+00
TOR*	9.999801e-01
VAN	3.011319e-10
VGK*	9.997866e-01
WIN*	1.000000e+00
WSH*	1.000000e+00

Here * indicates if a team made the playoffs. This model has Anaheim, Boston, Vegas, Winnipeg, Washington, Toronto, Tampa Bay, San Jose, Philadelphia , Pittsburgh, New Jersey, Nashville, Minnesota, LA, Columbus and Dallas all having a high probability of making the playoffs. The only team out of this list that did not make the playoffs during this season was Dallas, (instead, Colorado made it). One possible reason for this is that during the 2013-2014 season the NHL changed the playoff format; rather than the top 8 teams from each conference making it; the top 3 teams from each division made it and the two final teams representing each conference were wild card teams. No good packages were available to create the model plots for logistic regression. Now we go on to the conclusion.

Chapter 5

Conclusion

In conclusion, if I was going to continue work on these models, then the next step would be to build supplementary models to predict our predictor variables, extending the models to predicting unknown, or in progress years. I would also like to consider other variables that are commonly used in other areas such as Wins above replacement(WAR), and shots for percentage. I would also like to consider other machine learning algorithms that are commonly used in other non hockey related fields like Nave Bayes Classifier Algorithm[12]. Another thing I would like to investigate, is the prediction of the Stanley Cup winner (league champion). In the AIC multiple linear regression model, defensive statistics were weighted heavier than offensive ones. The win condition model weighed shots heavily due to variable selection. The ridge regression model weighted teams which were the top of every category heavily. Finally the logistic regression model predicted every playoff

team but one for the previous season.

Bibliography

[1] Joshua Weissbock Forecasting Success in the National Hockey League using In-Game Statistics and Textual Data School of Electrical Engineering and Computer Science Faculty of Engineering University of Ottawa, Ottawa, Canada, 2014

[2] Drew Hynes, nhlapi, January 9,2018, https://gitlab.com/dword4/nhlapi.git

[3] Micah Blake McCurdy,Hockey Viz 2018,http://hockeyviz.com/txt/preview1819,2018-2019 Regular Season Predictions

[4] https://www.nhl.com/fans/nhl-centennial

[5] NHL Standings Predictions: Offseason Edition,Larry Fisher , July 22,2018, https://thehockeywriters.com/nhl-standings-predictions-offseason-edition-2019/

[6] The Magnus Prediction Models,Estimating Individual Impact on NHL 5v5 Shot Rates, September 24, 2018, Micah Blake McCurdy,http://hockeyviz.com/txt/magnusEV

[7] The Magnus Prediction Models,September 30, 2018, Micah Blake McCurdy,http://hockeyviz.com/txt/magnus

[8] Seneca Labs Inc. 2014 - 2018,https://captaincalculator.com/sports/goals-against-average-calculator/

[9] Introduction to Probability and Statistics Principles and Applications for Engineering and the Computing Sciences,J.Susan Milton Jesse C. Arnold, McGraw Hill 2003.

[10] Regression Analysis by Example Fifth Edition,Samprit Chatterjee and Ali S. Hadi,Wiley Publication.

[11] Statistics Solutions 2018,https://www.statisticssolutions.com/assumptions-of-multiple-linear-regression

[12] Top 10 Machine Learning Algorithms,https://www.dezyre.com/article/top-10-machine-learning-algorithms/202,May 11, 2018

[13] The Analysis Factor. Karen-Grace Martin. What are nested models? 2017.

[14] Statistics Solutions 2018,https://www.statisticssolutions.com/assumptions-of-logistic-regression/

Appendix A

Code for API Call

```
NJ.2007.2008 <- fromJSON(file = "https://statsapi.web.
    nhl.com/api/v1/teams/1/stats?season=20072008")
```

Appendix B

Uncleaned Data

```
> NJ.2007.2008
$copyright
[1] "NHL and the NHL Shield are registered trademarks of the National Hockey League. NHI
$stats
$stats[[1]]
$stats[[1]]$type
$stats[[1]]$type$displayName
[1] "statsSingleSeason"
$stats[[1]]$splits
$stats[[1]]$splits[[1]]
$stats[[1]]$splits[[1]]$stat
$stats[[1]]$splits[[1]]$stat$gamesPlayed
[1] 82
$stats[[1]]$splits[[1]]$stat$wins
```

```
[1] 46
$stats[[1]]$splits[[1]]$stat$losses
[1] 29
$stats[[1]]$splits[[1]]$stat$ot
[1] 7
$stats[[1]]$splits[[1]]$stat$pts
[1] 99
$stats[[1]]$splits[[1]]$stat$ptPctg
[1] "60.4"
$stats[[1]]$splits[[1]]$stat$goalsPerGame
[1] 2.415
$stats[[1]]$splits[[1]]$stat$goalsAgainstPerGame
[1] 2.354
$stats[[1]]$splits[[1]]$stat$evGGARatio
[1] 1.0246
$stats[[1]]$splits[[1]]$stat$powerPlayPercentage
[1] "15.6"
$stats[[1]]$splits[[1]]$stat$powerPlayGoals
[1] 50
$stats[[1]]$splits[[1]]$stat$powerPlayGoalsAgainst
[1] 54
$stats[[1]]$splits[[1]]$stat$powerPlayOpportunities
[1] 320
$stats[[1]]$splits[[1]]$stat$penaltyKillPercentage
```

```
[1] "82.8"
$stats[[1]]$splits[[1]]$stat$shotsPerGame
[1] 28.8049
$stats[[1]]$splits[[1]]$stat$shotsAllowed
[1] 27.5244
$stats[[1]]$splits[[1]]$stat$winScoreFirst
[1] 0.725
$stats[[1]]$splits[[1]]$stat$winOppScoreFirst
[1] 0.405
$stats[[1]]$splits[[1]]$stat$winLeadFirstPer
[1] 0.806
$stats[[1]]$splits[[1]]$stat$winLeadSecondPer
[1] 0.879
$stats[[1]]$splits[[1]]$stat$winOutshootOpp
[1] 0.609
$stats[[1]]$splits[[1]]$stat$winOutshotByOpp
[1] 0.514
$stats[[1]]$splits[[1]]$stat$faceOffsTaken
[1] 4286
$stats[[1]]$splits[[1]]$stat$faceOffsWon
[1] 2160
$stats[[1]]$splits[[1]]$stat$faceOffsLost
[1] 2126
$stats[[1]]$splits[[1]]$stat$faceOffWinPercentage
```

```
[1] "50.4"
$stats[[1]]$splits[[1]]$stat$shootingPctg
[1] 8.4
$stats[[1]]$splits[[1]]$stat$savePctg
[1] 0.914
$stats[[1]]$splits[[1]]$team
$stats[[1]]$splits[[1]]$team$id
[1] 1
$stats[[1]]$splits[[1]]$team$name
[1] "New Jersey Devils"
$stats[[1]]$splits[[1]]$team$link
[1] "/api/v1/teams/1"
$stats[[2]]
$stats[[2]]$type
$stats[[2]]$type$displayName
[1] "regularSeasonStatRankings"
$stats[[2]]$splits
$stats[[2]]$splits[[1]]
$stats[[2]]$splits[[1]]$stat
$stats[[2]]$splits[[1]]$stat$wins
[1] "6th"
$stats[[2]]$splits[[1]]$stat$losses
[1] "8th"
$stats[[2]]$splits[[1]]$stat$ot
```

```
[1] "24th"
$stats[[2]]$splits[[1]]$stat$pts
[1] "6th"
$stats[[2]]$splits[[1]]$stat$ptPctg
[1] "6th"
$stats[[2]]$splits[[1]]$stat$goalsPerGame
[1] "27th"
$stats[[2]]$splits[[1]]$stat$goalsAgainstPerGame
[1] "5th"
$stats[[2]]$splits[[1]]$stat$evGGARatio
[1] "16th"
$stats[[2]]$splits[[1]]$stat$powerPlayPercentage
[1] "25th"
$stats[[2]]$splits[[1]]$stat$powerPlayGoals
[1] "27th"
$stats[[2]]$splits[[1]]$stat$powerPlayGoalsAgainst
[1] "6th"
$stats[[2]]$splits[[1]]$stat$powerPlayOpportunities
[1] "26th"
$stats[[2]]$splits[[1]]$stat$penaltyKillOpportunities
[1] "5th"
$stats[[2]]$splits[[1]]$stat$penaltyKillPercentage
[1] "13th"
$stats[[2]]$splits[[1]]$stat$shotsPerGame
```

```
[1] "15th"
$stats[[2]]$splits[[1]]$stat$shotsAllowed
[1] "8th"
$stats[[2]]$splits[[1]]$stat$winScoreFirst
[1] "13th"
$stats[[2]]$splits[[1]]$stat$winOppScoreFirst
[1] "4th"
$stats[[2]]$splits[[1]]$stat$winLeadFirstPer
[1] "5th"
$stats[[2]]$splits[[1]]$stat$winLeadSecondPer
[1] "14th"
$stats[[2]]$splits[[1]]$stat$winOutshootOpp
[1] "7th"
$stats[[2]]$splits[[1]]$stat$winOutshotByOpp
[1] "7th"
$stats[[2]]$splits[[1]]$stat$faceOffsTaken
[1] "29th"
$stats[[2]]$splits[[1]]$stat$faceOffsWon
[1] "29th"
$stats[[2]]$splits[[1]]$stat$faceOffsLost
[1] "4th"
$stats[[2]]$splits[[1]]$stat$faceOffWinPercentage
[1] "13th"
$stats[[2]]$splits[[1]]$stat$savePctRank
```

```
[1] "6th"

$stats[[2]]$splits[[1]]$stat$shootingPctRank

[1] "25th"

$stats[[2]]$splits[[1]]$team

$stats[[2]]$splits[[1]]$team$id

[1] 1

$stats[[2]]$splits[[1]]$team$name

[1] "New Jersey Devils"

$stats[[2]]$splits[[1]]$team$link

[1] "/api/v1/teams/1"
```

Appendix C

Code for Cleaning Data

```
NJ.2007.2008 <- NJ.2007.2008$stats[[1]]$splits[[1]]$stat
```

Cleaned Data

```
> NJ.2007.2008

$gamesPlayed

[1] 82

$wins

[1] 46

$losses

[1] 29
```

```
$ot
[1] 7
$pts
[1] 99
$ptPctg
[1] "60.4"
$goalsPerGame
[1] 2.415
$goalsAgainstPerGame
[1] 2.354
$evGGARatio
[1] 1.0246
$powerPlayPercentage
[1] "15.6"
$powerPlayGoals
[1] 50
$powerPlayGoalsAgainst
[1] 54
$powerPlayOpportunities
[1] 320
$penaltyKillPercentage
[1] "82.8"
$shotsPerGame
[1] 28.8049
```

```
$shotsAllowed

[1] 27.5244

$winScoreFirst

[1] 0.725

$winOppScoreFirst

[1] 0.405

$winLeadFirstPer

[1] 0.806

$winLeadSecondPer

[1] 0.879

$winOutshootOpp

[1] 0.609

$winOutshotByOpp

[1] 0.514

$faceOffsTaken

[1] 4286

$faceOffsWon

[1] 2160

$faceOffsLost

[1] 2126

$faceOffWinPercentage

[1] "50.4"

$shootingPctg

[1] 8.4
```

```
$savePctg

[1] 0.914
```

Appendix D

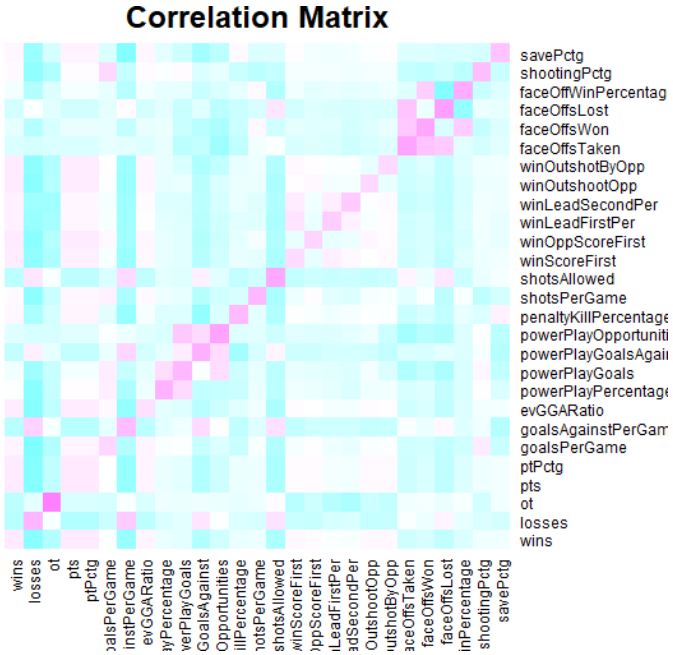

Figure 1

Heat Map

Appendix E

Error Table

Team	Actual Points	$\lvert AICError \rvert$	$\lvert R^2Error \rvert$	$\lvert RidgeError \rvert$
ANA	101	0.24715641	1.6061406	1.93811452
ARI	70	0.04901625	3.1156016	1.28011995
BOS	112	1.22076767	0.1551671	0.35062443
BUF	62	0.65149248	0.3898144	1.93565395
CGY	84	0.10388349	2.5008741	0.82366131

CAR	83	4.35082331	6.4428237	3.34184350
CHI	76	0.55805277	1.5191952	1.87248740
COL	95	2.07595120	3.5134566	0.42377953
CBJ	97	0.54118090	2.3700857	0.40203955
DAL	92	3.47583721	3.5531989	3.01541779
DET	73	2.20990543	4.1350665	0.05910152
EDM	78	0.45954586	3.6595266	2.39590021
FLA	96	3.63236821	1.4277339	2.55620918
LAK	98	5.80054208	4.0342055	5.62925663
MIN	101	0.85378174	0.8281129	1.44090552
MON	71	0.74507075	1.3535107	1.73986889
NAS	117	4.38299534	6.5608927	6.24530799
NJD	97	2.79598870	0.2648046	2.50731256
NYR	80	2.04603762	0.2853365	0.68078794
NYI	77	0.25901951	2.7849972	3.14971673
OTT	67	1.38637441	3.2626904	3.14999867
PHI	98	2.13773068	3.9597064	2.57759596
PIT	100	0.11463457	1.0504457	0.27550942
SJS	100	0.79300290	0.8822942	0.61392558
STL	94	1.36968826	0.6159767	0.76033569
TBL	113	1.26412616	3.0261860	0.26612309
TOR	105	0.89029297	3.4357766	0.81191281
VAN	73	1.11990454	0.8329045	3.84934906
VGK	105	0.72077358	5.7287411	0.25785441

WIN	114	2.61239679	3.9654460	2.99759161
WSH	109	0.24877122	0.7015060	2.12668772

Appendix F

R Code

```
# This library allows us to run logistic regression models
library(glmnet)
# This line reads all of the csv files into R that we are going to use as our data to create the models
Dataset<-(lapply(list.files(), read.csv))
# This line reads in the 2017-2018 data
TrainingSet <- (lapply(list.files(), read.csv))
# These lines clean the dataset
for(i in 1:270){
Dataset[[i]]$X <- NULL
Dataset[[i]]$gamesPlayed <- NULL
}
for(i in 1:31){
TrainingSet[[i]]$X <- NULL
TrainingSet[[i]]$gamesPlayed <- NULL
}
wins <- NULL
losses <- NULL
ot <- NULL
pts <- NULL
ptPctg <- NULL
goalsPerGame <- NULL
goalsAgainstPerGame <- NULL
evGAARatio <- NULL
powerPlayPercentage <- NULL
powerPlayGoals <- NULL
powerPlayGoalsAgainst <- NULL
powerPlayOpportunities <- NULL
penaltyKillPercentage <- NULL
shotsPerGame <- NULL
shotsAllowed <- NULL
winScoreFirst <- NULL
winOppScoreFirst <- NULL
winLeadFirstPer <- NULL
winLeadSecondPer <- NULL
winOutshootOpp <- NULL
```

```
winOutshotByOpp <- NULL
faceOffsTaken <- NULL
faceOffsWon <- NULL
faceOffsLost <- NULL
faceOffWinPercentage <- NULL
shootingPctg <- NULL
savePctg <- NULL

for(i in 1:270){
wins[i] <- Dataset[[i]]$wins
losses[i] <- Dataset[[i]]$losses
ot[i] <- Dataset[[i]]$ot
pts[i] <- Dataset[[i]]$pts
ptPctg[i] <- Dataset[[i]]$ptPctg
goalsPerGame[i] <- Dataset[[i]]$goalsPerGame
goalsAgainstPerGame[i] <- Dataset[[i]]$goalsAgainstPerGame
evGAARatio[i] <- Dataset[[i]]$evGGARatio
powerPlayPercentage[i] <- Dataset[[i]]$powerPlayPercentage
powerPlayGoals[i] <- Dataset[[i]]$powerPlayGoals
powerPlayGoalsAgainst[i] <- Dataset[[i]]$powerPlayGoalsAgainst
powerPlayOpportunities[i] <- Dataset[[i]]$powerPlayOpportunities
penaltyKillPercentage[i] <- Dataset[[i]]$penaltyKillPercentage
shotsPerGame[i] <- Dataset[[i]]$shotsPerGame
shotsAllowed[i] <- Dataset[[i]]$shotsAllowed
winScoreFirst[i] <- Dataset[[i]]$winScoreFirst
winOppScoreFirst[i] <- Dataset[[i]]$winOppScoreFirst
winLeadFirstPer[i] <- Dataset[[i]]$winLeadFirstPer
winLeadSecondPer[i] <- Dataset[[i]]$winLeadSecondPer
winOutshootOpp[i] <-  Dataset[[i]]$winOutshootOpp
winOutshotByOpp[i] <- Dataset[[i]]$winOutshotByOpp
faceOffsTaken[i] <- Dataset[[i]]$faceOffsTaken
faceOffsWon[i] <- Dataset[[i]]$faceOffsWon
faceOffsLost[i] <- Dataset[[i]]$faceOffsLost
faceOffWinPercentage[i] <- Dataset[[i]]$faceOffWinPercentage
shootingPctg[i] <- Dataset[[i]]$shootingPctg
savePctg[i] <-  Dataset[[i]]$savePctg
}

Dataset <- cbind(wins,losses ,ot , pts , ptPctg ,goalsPerGame , goalsAgainstPerGame ,evGAARatio ,powerPlayPercentage , powerPlayGoals ,powerPlayGoalsAgainst ,

wins <- NULL
losses <- NULL
ot <- NULL
pts <- NULL
ptPctg <- NULL
goalsPerGame <- NULL
goalsAgainstPerGame <- NULL
```

```
evGAARatio <- NULL
powerPlayPercentage <- NULL
powerPlayGoals <- NULL
powerPlayGoalsAgainst <- NULL
powerPlayOpportunities <- NULL
penaltyKillPercentage <- NULL
shotsPerGame <- NULL
shotsAllowed <- NULL
winScoreFirst <- NULL
winOppScoreFirst <- NULL
winLeadFirstPer <- NULL
winLeadSecondPer <- NULL
winOutshootOpp <- NULL
winOutshotByOpp <- NULL
faceOffsTaken <- NULL
faceOffsWon <- NULL
faceOffsLost <- NULL
faceOffWinPercentage <- NULL
shootingPctg <- NULL
savePctg <- NULL

# These lines clean the prediction data
for(i in 1:31){
wins[i] <- TrainingSet[[i]]$wins
losses[i] <- TrainingSet[[i]]$losses
ot[i] <- TrainingSet[[i]]$ot
pts[i] <- TrainingSet[[i]]$pts
ptPctg[i] <- TrainingSet[[i]]$ptPctg
goalsPerGame[i] <- TrainingSet[[i]]$goalsPerGame
goalsAgainstPerGame[i] <- TrainingSet[[i]]$goalsAgainstPerGame
evGAARatio[i] <- TrainingSet[[i]]$evGGARatio
powerPlayPercentage[i] <- TrainingSet[[i]]$powerPlayPercentage
powerPlayGoals[i] <- TrainingSet[[i]]$powerPlayGoals
powerPlayGoalsAgainst[i] <- TrainingSet[[i]]$powerPlayGoalsAgainst
powerPlayOpportunities[i] <- TrainingSet[[i]]$powerPlayOpportunities
penaltyKillPercentage[i] <- TrainingSet[[i]]$penaltyKillPercentage
shotsPerGame[i] <- TrainingSet[[i]]$shotsPerGame
shotsAllowed[i] <- TrainingSet[[i]]$shotsAllowed
winScoreFirst[i] <- TrainingSet[[i]]$winScoreFirst
winOppScoreFirst[i] <- TrainingSet[[i]]$winOppScoreFirst
winLeadFirstPer[i] <- TrainingSet[[i]]$winLeadFirstPer
winLeadSecondPer[i] <- TrainingSet[[i]]$winLeadSecondPer
winOutshootOpp[i] <-  TrainingSet[[i]]$winOutshootOpp
winOutshotByOpp[i] <- TrainingSet[[i]]$winOutshotByOpp
faceOffsTaken[i] <- TrainingSet[[i]]$faceOffsTaken
faceOffsWon[i] <- TrainingSet[[i]]$faceOffsWon
faceOffsLost[i] <- TrainingSet[[i]]$faceOffsLost
```

```
faceOffWinPercentage[i] <- TrainingSet[[i]]$faceOffWinPercentage
shootingPctg[i] <- TrainingSet[[i]]$shootingPctg
savePctg[i] <-  TrainingSet[[i]]$savePctg
}
TrainingSet <- cbind(wins,losses ,ot , pts , ptPctg ,goalsPerGame , goalsAgainstPerGame ,evGAARatio ,powerPlayPercentage , powerPlayGoals ,powerPlayGoalsAgain

#This creates the correlation matrix
correlation.matrix <- cor(Dataset)
#This creates the heatmap
heatmap(correlation.matrix,Rowv = NA,Colv = NA, main = "Correlation Matrix",col = cm.colors(256))

Dataset <- data.frame(Dataset)
#This removes all the highly correlated variables
Dataset$wins <- NULL
Dataset$losses <- NULL
Dataset$ot <- NULL
Dataset$faceOffsLost <- NULL
Dataset$ptPctg <- NULL

#This creates a matrix out of our dataset to be used in the ridge regression
modelmatrix <- model.matrix(Dataset$pts ~. -1,data = Dataset)

#This cleans the training set
TrainingSet <- data.frame(TrainingSet)

TrainingSet$wins <- NULL
TrainingSet$losses <- NULL
TrainingSet$ot <- NULL
TrainingSet$faceOffsLost <- NULL
TrainingSet$ptPctg <- NULL

newmodelmatrix <- model.matrix(TrainingSet$pts ~. -1,data = TrainingSet)

y <- Dataset$pts
newy <- TrainingSet$pts

# Linear Regression Section

null.model <- lm(pts ~ 1, data = Dataset)
full.model <- lm(pts ~., data = Dataset)
#This runs our AIC variable selection
AIC.model <- step(full.model, scope = list(lower = null.model, upper = full.model), direction = "both")
#This creates the ridge regession model the key is the alpha=0
ridge.regression.model <- cv.glmnet(modelmatrix,y,alpha = 0)
#This function creates all of our plots
plot(AIC.model)
```

```
#This creates our predictions
ridge.prediction <- predict.cv.glmnet(ridge.regression.model, newx = newmodelmatrix)
AIC.prediction <- predict(AIC.model, TrainingSet)

plot(ridge.regression.model)

summary(full.model)

#These variables come from the p-values in the full model
pvalue.model <- lm(pts ~ winScoreFirst + winOppScoreFirst + winOutshootOpp + winOutshotByOpp, data = Dataset)

plot(pvalue.model)

pvalue.prediction <- predict(pvalue.model, TrainingSet)

AIC.error <- abs(newy - AIC.prediction)
pvalue.error <- abs(newy - pvalue.prediction)
Ridge.error <- abs(newy - ridge.prediction)

boxplot(AIC.error)
boxplot(pvalue.error)
boxplot(Ridge.error)

#Logistic Regression Section

#This variable was hardcoded

madePlayoffs <- c(1,1,0,1,0,1,1,1,1,0,0,1,1,1,0,0,0,0,1,1,1,1,1,1,0,0,1,0,0,1,1,0,0,0,0,1,1,0,0,0,0,1,0,1,0,1,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,0,1,0,0,1,0,

Dataset <- data.frame(cbind(Dataset,madePlayoffs))

logistic.full.model.playoffs <- glm(madePlayoffs ~ ., data = Dataset, family = "binomial")
logistic.null.model.playoffs <- glm(madePlayoffs ~ 1, data = Dataset, family = "binomial")

madePlayoffs <- c(1,0,1,0,0,0,0,1,1,0,0,0,0,1,1,0,1,1,0,0,0,1,1,1,0,1,1,0,1,1,1,1)

TrainingSet <- data.frame(cbind(TrainingSet,madePlayoffs))

logistic.full.model.playoff.prediction <- predict.glm(logistic.full.model.playoffs, TrainingSet, type = "response")
```

Appendix G

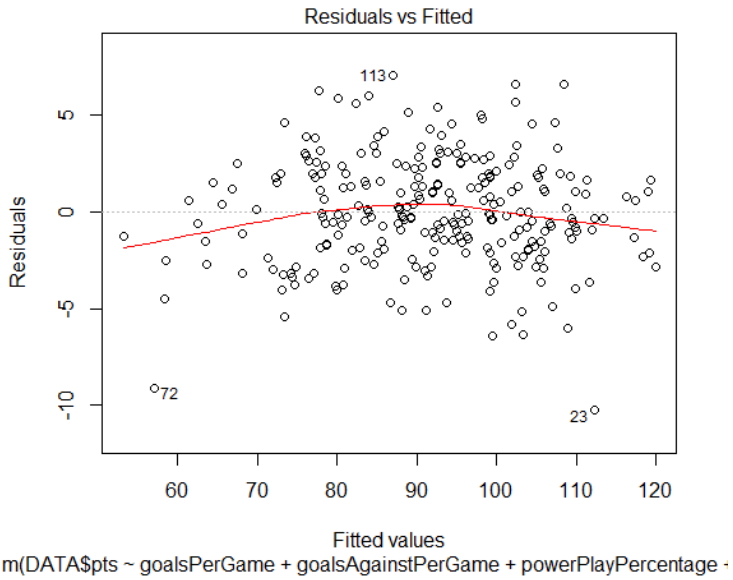

Figure 2: Residuals vs Fitted Values for AIC

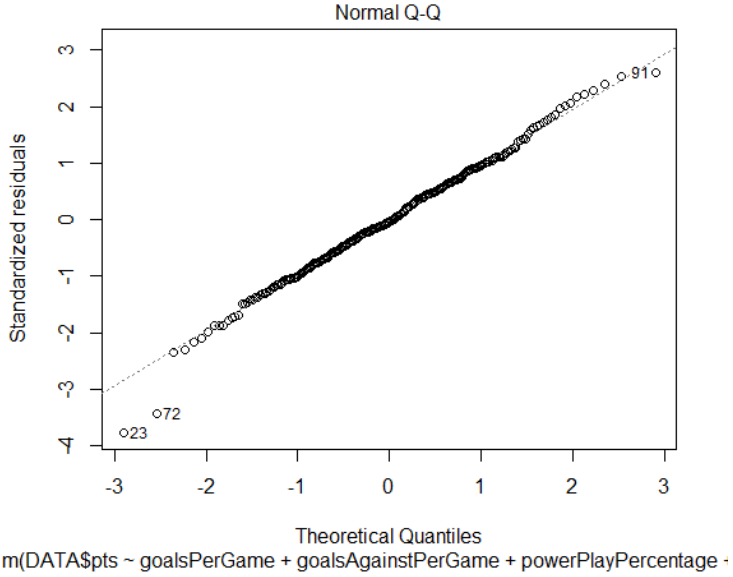

Figure 3: QQ Plot for AIC

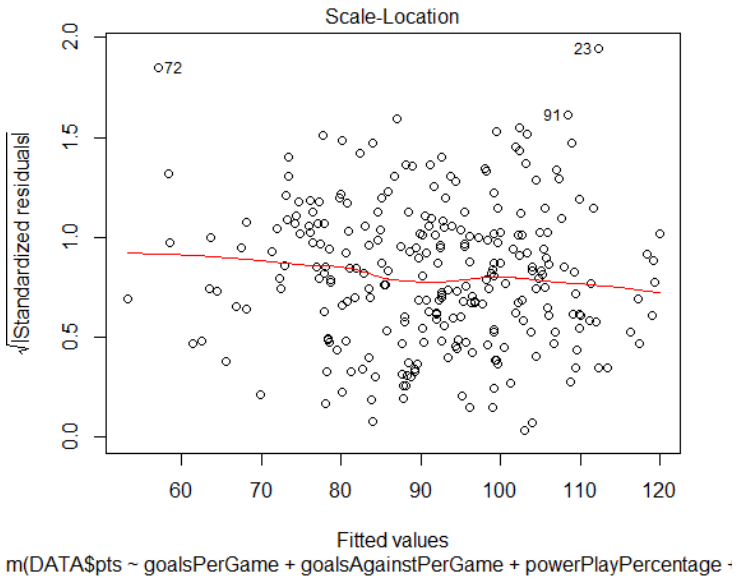

Figure 4: Scale-Location for AIC

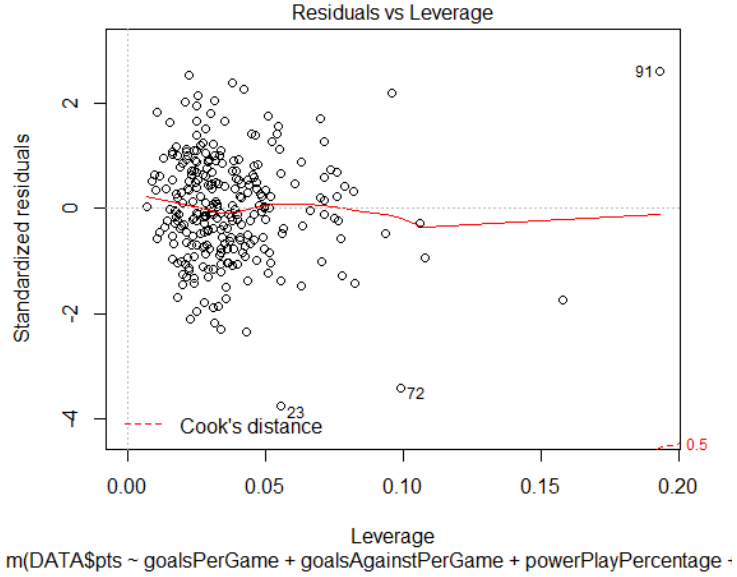

Figure 5: Residuals vs Leverage for AIC

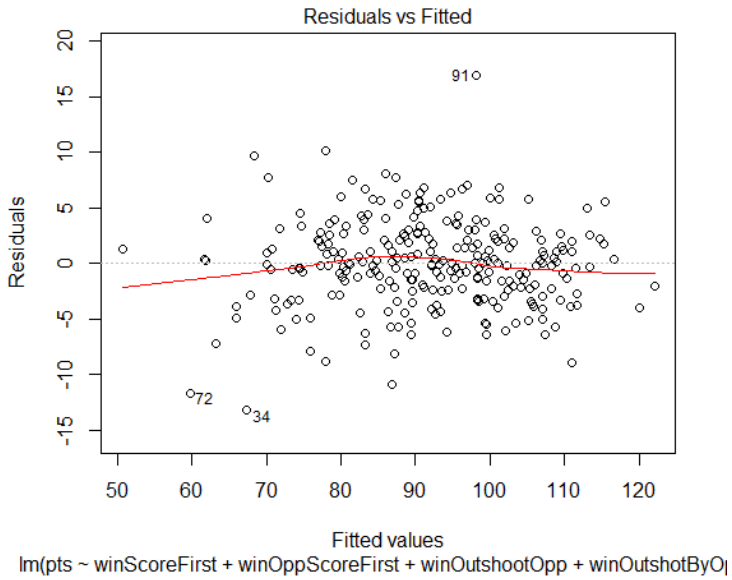

Figure 6: Residuals vs Fitted Values for R Squared

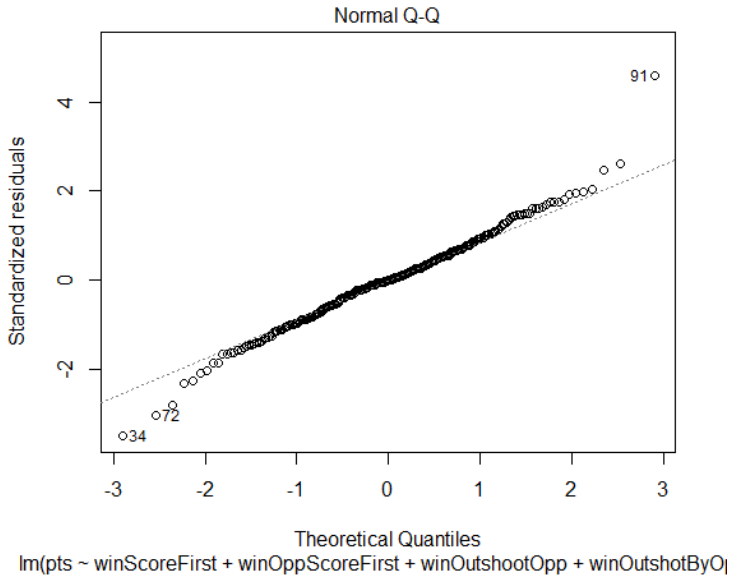

Figure 7: QQ Plot for R Squared

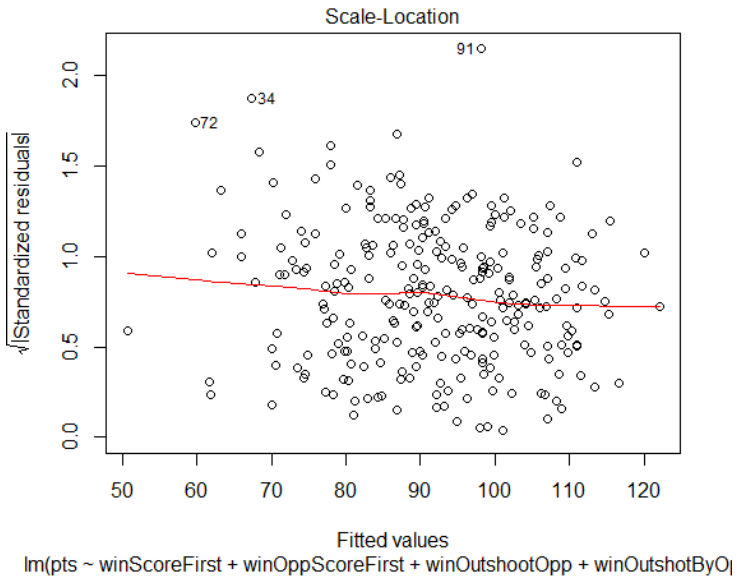

Figure 8: Scale-Location for R Squared

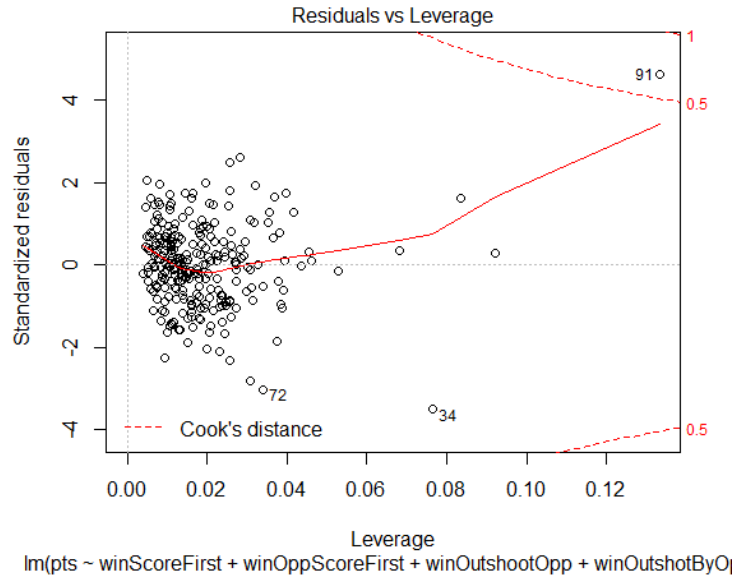

Figure 9: Residuals vs Leverage for R Squared