Predicting NHL Team Success

Eric Foote

UNBSJ

December 4 2018

Overview

- In this discussion I use multiple linear and ridge regression techniques to predict National Hockey Leauge(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

What we are going to Discuss

- We are first going to talk about the dataset used
- Then we are going to talk about the theory behind each of the models
- Then the implementation of the models
- Finally the results

Dataset

- How we got the data?
- What variables were removed?
- What variables were created?
- What are we going to consider for variables?

How we got Data?

- Comes from the NHL Stats application programmer interface(API)
- https://statsapi.web.nhl.com/api/v1
- Make API calls using the RJSON package in R
- /teams/i/stats?season=20072008 where i is the unique team id
- Each call creates a data frame that needs to be cleaned

Data Cleaning

- $stats[[1]]$splits[[1]] extracts the variables that we need
- The 2012-2013 season was removed due to the 48 game season
- Check for highly corrleated variables

Correlation Matrix

	wins	losses	ot	pts	ptPctg	goalsPerGame	goalsAgainstPerGame	evGGARatio
wins	1.00000000	-0.91338168	-0.341297748	0.97940092	0.97943941	0.704664112	-0.71164841	0.84647648
losses	-0.91338168	1.00000000	-0.070924704	-0.97677153	-0.97671844	-0.672865791	0.72828355	-0.83555919
ot	-0.34129775	-0.07092470	1.000000000	-0.14446655	-0.14468343	-0.172992065	0.06215191	-0.14481924
pts	0.97940092	-0.97677153	-0.144466552	1.00000000	0.99999393	0.704651090	-0.73581477	0.85999202
ptPctg	0.97943941	-0.97671844	-0.144683435	0.99999393	1.00000000	0.704513704	-0.73573369	0.85980309
goalsPerGame	0.70466411	-0.67286579	-0.172992065	0.70465109	0.70451370	1.000000000	-0.15609279	0.70354747
goalsAgainstPerGame	-0.71164841	0.72828355	0.062151915	-0.73581477	-0.73573369	-0.156092790	1.00000000	-0.67170634
evGGARatio	0.84647648	-0.83555919	-0.144819240	0.85999202	0.85980309	0.703547471	-0.67170634	1.00000000
powerPlayPercentage	0.41982235	-0.44541567	-0.000231642	0.44190562	0.44185060	0.544414139	-0.18107564	0.29573591
powerPlayGoals	0.32910785	-0.33981575	-0.021781765	0.34177916	0.34163665	0.565406475	-0.02088771	0.23541476
powerPlayGoalsAgainst	-0.34581768	0.36591564	0.002462457	-0.36352020	-0.36361698	0.043935434	0.61390918	-0.20476851
powerPlayOpportunities	0.04563803	-0.03685949	-0.026717561	0.04230449	0.04217171	0.251243070	0.14315771	0.04279225
penaltyKillPercentage	0.42958094	-0.43354817	-0.051542661	0.44115583	0.44125670	0.080174549	-0.63498703	0.26761526
shotsPerGame	0.46233472	-0.48632604	-0.009937758	0.48457414	0.48416396	0.534204191	-0.23986760	0.46359762
shotsAllowed	-0.43923605	0.41112171	0.126979525	-0.43511428	-0.43515702	-0.155592704	0.53532829	-0.39298290
winScoreFirst	0.79810571	-0.68683821	-0.369680813	0.76076539	0.76064391	0.500213491	-0.60829131	0.66087965
winOppScoreFirst	0.76179561	-0.70748125	-0.233051951	0.75189223	0.75191082	0.603717940	-0.44587325	0.63465239
winLeadFirstPer	0.64075934	-0.51220117	-0.387368877	0.59132389	0.59121951	0.428731019	-0.44209144	0.51247576
winLeadSecondPer	0.61379251	-0.45300265	-0.457977402	0.54776693	0.54778025	0.382089652	-0.45028997	0.50854446
winOutshootOpp	0.77975491	-0.72225018	-0.242955911	0.76867073	0.76877751	0.538012353	-0.56022198	0.66290709
winOutshotByOpp	0.78721280	-0.70581639	-0.299172796	0.76444504	0.76434258	0.564230257	-0.54101131	0.65287939
faceOffsTaken	0.01741349	-0.02766091	0.021199434	0.02288568	0.02288446	0.126097275	0.09393123	0.04656643
faceOffsWon	0.16497185	-0.19791268	0.052744883	0.18500006	0.18483375	0.177890260	-0.08796604	0.18645551
faceOffsLost	-0.15269533	0.17295833	-0.025208019	-0.16616073	-0.16598214	-0.002547341	0.23732638	-0.13197063
faceOffWinPercentage	0.23384489	-0.27283749	0.056985492	0.25841507	0.25815307	0.137194250	-0.23363303	0.23434261
shootingPctg	0.50180026	-0.44514085	-0.201727416	0.48491920	0.48503859	0.788922782	-0.02019992	0.49412681
savePctg	0.51216756	-0.55022334	0.015493709	0.54249747	0.54236586	0.059050043	-0.78652894	0.50077181

Figure: 1

	powerPlayPercentage	powerPlayGoals	powerPlayGoalsAgainst	powerPlayOpportunities	penaltyKillPercentage
wins	0.419822348	0.32910785	-0.345817684	0.045638032	0.42958094
losses	-0.445415674	-0.33981575	0.365915643	-0.036859486	-0.43354817
ot	-0.000231642	-0.02178177	0.002462457	-0.026717561	-0.05154266
pts	0.441905624	0.34177916	-0.363520199	0.042304485	0.44115583
ptPctg	0.441850604	0.34163665	-0.363616984	0.042171706	0.44125670
goalsPerGame	0.544414139	0.56540648	0.043935434	0.251243070	0.08017455
goalsAgainstPerGame	-0.181075641	-0.02088771	0.613909175	0.143157709	-0.63498703
evGGARatio	0.295735912	0.23541476	-0.204768514	0.042792252	0.26761526
powerPlayPercentage	1.000000000	0.70322283	-0.012569278	-0.002757976	0.04386814
powerPlayGoals	0.703222832	1.00000000	0.391540595	0.701800423	0.03636346
powerPlayGoalsAgainst	-0.012569278	0.39154059	1.000000000	0.564128384	-0.67333921
powerPlayOpportunities	-0.002757976	0.70180042	0.564128384	1.000000000	0.01716528
penaltyKillPercentage	0.043868141	0.03636346	-0.673339212	0.017165280	1.00000000
shotsPerGame	0.292268749	0.24663523	-0.156444594	0.049563577	0.10484035
shotsAllowed	-0.073394849	-0.11342427	0.320195622	-0.098203232	-0.37595366
winScoreFirst	0.311366761	0.25558023	-0.255997371	0.052046586	0.34137338

Figure: 2

winOppScoreFirst	0.289308244	0.22755706	-0.258737882	0.028981575	0.28569704
winLeadFirstPer	0.239702645	0.19546996	-0.186227246	0.037874966	0.21729705
winLeadSecondPer	0.161403929	0.16134846	-0.176323074	0.077309297	0.27651375
winOutshootOpp	0.314641637	0.33466882	-0.209337485	0.154644006	0.34719233
winOutshotByOpp	0.322037676	0.14427312	-0.346573072	-0.115526034	0.31951071
faceOffsTaken	0.137333283	-0.19789345	-0.147048342	-0.426817727	-0.16877184
faceOffsWon	0.197440432	-0.05160372	-0.164421053	-0.280083931	-0.02791413
faceOffsLost	-0.006787933	-0.24292305	-0.043713224	-0.340802697	-0.22463914
faceOffWinPercentage	0.153272903	0.13234617	-0.095760519	0.028906627	0.13980661
shootingPctg	0.432090378	0.49409953	0.162271800	0.269743843	0.02783772
savePctg	0.153303640	-0.06293843	-0.483110820	-0.240545460	0.46897920

Figure: 3

	shotsPerGame	shotsAllowed	winScoreFirst	winOppScoreFirst	winLeadFirstPer	winLeadSecondPer	winOutshootOpp
wins	0.462334723	-0.439236052	0.79810571	0.76179561	0.64075934	0.61379251	0.779754908
losses	-0.486326041	0.411121708	-0.68683821	-0.70748125	-0.51220117	-0.45300265	-0.722250176
ot	-0.009937758	0.126979525	-0.36968081	-0.23305195	-0.38736888	-0.45797740	-0.242955911
pts	0.484574140	-0.435114278	0.76076539	0.75189223	0.59132389	0.54776693	0.768670731
ptPctg	0.484163959	-0.435157021	0.76064391	0.75191082	0.59121951	0.54778025	0.768777508
goalsPerGame	0.534204191	-0.155592704	0.50021349	0.60371794	0.42873102	0.38208965	0.538012353
goalsAgainstPerGame	-0.239867598	0.535328292	-0.60829131	-0.44587325	-0.44209144	-0.45028997	-0.560221981
evGGARatio	0.463597618	-0.392982899	0.66087965	0.63465239	0.51247576	0.50854446	0.662907088
powerPlayPercentage	0.292268749	-0.073394849	0.31136676	0.28930824	0.23970264	0.16140393	0.314641637
powerPlayGoals	0.246635226	-0.113424266	0.25558023	0.22755706	0.19546996	0.16134846	0.334668825
powerPlayGoalsAgainst	-0.156444594	0.320195622	-0.25599737	-0.25873788	-0.18622725	-0.17632307	-0.209337485
powerPlayOpportunities	0.049563577	-0.098203232	0.05204659	0.02898157	0.03787497	0.07730930	0.154644006
penaltyKillPercentage	0.104840354	-0.375953657	0.34137338	0.28569704	0.21729705	0.27651375	0.347192332
shotsPerGame	1.000000000	-0.258853784	0.28678666	0.44957561	0.19238156	0.16389312	0.354487265
shotsAllowed	-0.258853784	1.000000000	-0.37700839	-0.28031934	-0.32406756	-0.28792336	-0.353536068
winScoreFirst	0.286786664	-0.377008391	1.00000000	0.31269667	0.77835728	0.65640692	0.581611214
winOppScoreFirst	0.449575607	-0.280319338	0.31269667	1.00000000	0.29383348	0.33157639	0.636969537
winLeadFirstPer	0.192381557	-0.324067561	0.77835728	0.29383348	1.00000000	0.64015184	0.447557008
winLeadSecondPer	0.163893120	-0.287923364	0.65640692	0.33157639	0.64015184	1.00000000	0.452923869
winOutshootOpp	0.354487265	-0.353536068	0.58161121	0.63696954	0.44755701	0.45292387	1.000000000
winOutshotByOpp	0.374159442	-0.289494864	0.64697234	0.58281202	0.52586969	0.51910622	0.314526639
faceOffsTaken	0.211414233	0.269862564	0.01216178	0.05907396	0.02305632	-0.07009931	0.006635249
faceOffsWon	0.391846582	-0.001910823	0.15965255	0.11801185	0.12085054	0.02394003	0.103217573
faceOffsLost	-0.105825024	0.409695451	-0.15485638	-0.03881538	-0.09629936	-0.13185910	-0.101970911
faceOffWinPercentage	0.371083292	-0.287962241	0.22821957	0.12086332	0.16001685	0.11001504	0.151884410
shootingPctg	-0.094068541	-0.004809620	0.39526761	0.38320788	0.37790444	0.34104090	0.381813878
savePctg	0.084244349	0.097655331	0.43998873	0.31435720	0.28301617	0.32236449	0.397208195

Figure: 4

	winOutshotByOpp	faceOffsTaken	faceOffsWon	faceOffsLost	faceOffWinPercentage	shootingPctg	savePctg	
wins	0.78721280	0.017413488	0.164971849	-0.152695333	0.23384489	0.50180026	0.51216756	
losses	-0.70581639	-0.027660908	-0.197912685	0.172958331	-0.27283749	-0.44514085	-0.55022334	
ot	-0.29917280	0.021199434	0.052744883	-0.025208019	0.05698549	-0.20172742	0.01549371	
pts	0.76444504	0.022885683	0.185000055	-0.166160730	0.25841507	0.48491920	0.54249747	
ptPctg	0.76434258	0.022884457	0.184833753	-0.165982140	0.25815307	0.48503859	0.54236586	
goalsPerGame	0.56423026	0.126097275	0.177890260	-0.002547341	0.13719425	0.78892278	0.05905004	
goalsAgainstPerGame	-0.54101131	0.093931233	-0.087966038	0.237326382	-0.23363303	-0.02019992	-0.78652894	

Figure: 5

evGGARatio	0.65287939	0.046566430	0.186455515	-0.131970627	0.23434261	0.49412681	0.50077181
powerPlayPercentage	0.32203768	0.137333283	0.197440432	-0.006787933	0.15327290	0.43209038	0.15330364
powerPlayGoals	0.14427312	-0.197893451	-0.051603720	-0.242923051	0.13234617	0.49409953	-0.06293843
powerPlayGoalsAgainst	-0.34657307	-0.147048342	-0.164421053	-0.043713224	-0.09576052	0.16227180	-0.48311082
powerPlayOpportunities	-0.11552603	-0.426817727	-0.280083931	-0.340802697	0.02890663	0.26974384	-0.24054546
penaltyKillPercentage	0.31951071	-0.168771839	-0.027914132	-0.224639143	0.13980661	0.02783772	0.46897920
shotsPerGame	0.37415944	0.211414233	0.391846582	-0.105825024	0.37108329	-0.09406854	0.08424435
shotsAllowed	-0.28949486	0.269862564	-0.001910823	0.409695451	-0.28796224	-0.00480962	0.09765533
winScoreFirst	0.64697234	0.012161780	0.159652550	-0.154856379	0.22821957	0.39526761	0.43998873
winOppScoreFirst	0.58281202	0.059073956	0.118011852	-0.038815378	0.12086332	0.38320788	0.31435720
winLeadFirstPer	0.52586969	0.023056320	0.120850539	-0.096299357	0.16001685	0.37790444	0.28301617
winLeadSecondPer	0.51910622	-0.070099308	0.023940027	-0.131859098	0.11001504	0.34104090	0.32236449
winOutshootOpp	0.31452664	0.006635249	0.103217573	-0.101970911	0.15188441	0.38181388	0.39720820
winOutshotByOpp	1.00000000	0.065346076	0.183977010	-0.100915103	0.21013820	0.39729891	0.42394340
faceOffsTaken	0.06534608	1.000000000	0.750142584	0.696559468	0.07855773	-0.01341218	0.08622892
faceOffsWon	0.18397701	0.750142584	1.000000000	0.048053924	0.71763281	-0.07546721	0.10169598
faceOffsLost	-0.10091510	0.696559468	0.048053924	1.000000000	-0.65998739	0.06162474	0.01990475
faceOffWinPercentage	0.21013820	0.078557727	0.717632810	-0.659987389	1.00000000	-0.09984663	0.06403039
shootingPctg	0.39729891	-0.013412180	-0.075467208	0.061624735	-0.09984663	1.00000000	0.01402797
savePctg	0.42394340	0.086228923	0.101695975	0.019904755	0.06403039	0.01402797	1.00000000

Figure: 6

What does this tell us?

- Our predictor points is nearly perfectly correlated with wins, losses and point percentage
- So we remove them
- over time losses are also removed since they are in points calculation
- faceOff losses are also removed since faceOff wins are in the set

Dataset Description

- 270 observations to build model
- 2007-2008 until 2016-2017
- 24 variables

Variables considered

- Points: a team gets 2 for a win and 1 for a overtime loss
- Goals per game: average number of goals a team scores per game
- Goals against per game: average number of goals a team gives up
- evGAARatio: goals the goalie gives up per even strength regulation time
- Power play percentage: percentage of power play goals a team scores
- Power play goals: the number of Power play goals a team scores

- Penalty kill percentage: percentage of power plays against that do not result in a goal
- Power play goals against: number of power play goals against
- Power play opportunities: number of power plays a team got during the season
- Shots per game: average number of shots on goal
- Shots allowed: average number of that a team gives up

- Win score first: percentage of wins that a team scored first
- Win opponent scored first: percentage of wins that the opponent scored first
- Win leading after first period: percentage of wins that a team had the lead after the first period
- Win leading after second period: percentage of wins that a team had the lead after the second period
- Win out shooting opponent: percentage of wins that a team out shot the opponent
- Win out shot by opponent: percentage of wins that the opponent out shot the team

- Shooting percentage: percentage of shots that resulted in goals
- Save percentage: percentage of shots that did not result in goals against
- Face offs taken: number of face offs taken
- Face offs won: number of face offs won
- Face off win percentage: percentage of face offs won

Additional Variable Definition

- made Playoffs
- We define this categorical variable ourselves
- Definition: this variable gets a 1 if a team makes the playoffs and a 0 if a team misses the playoffs

Methods

- Multiple Linear Regression
- AIC Variable Selection
- Ridge Regression
- 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 $x_1, ..., x_k$

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

- Which can be rewritten as:

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

- where β_k are our model parameters and ϵ_i is the random difference between $Y|_{x_i}$ and its mean $\mu|_{x_i}$

- We can express a series of the first equations in matrix form as

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

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

Assumptions

- $\mathrm{E}(\epsilon) = \hat{0}$ and $var(\epsilon) = E(\epsilon * \epsilon^T) = \sigma^2 I$
- ϵ_i are independent $\forall i$
- Predictor variables are not highly correlated with each other
- The variance of error terms are similar across the values of independent variables

AIC Variable Selection

- The Akaike Information Criterion
- Select a model to balance accuracy and simplicity

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

- An advantage to AIC is it allows us to compare non-nested models
- A model is nested if "the parameters in one model are a subset of another"

Ridge Regression

- Alternative estimation method that may be used to advantage when the predictor variables are highly collinear
- This helps us since some of our variables are computed from one another
- The first thing we need to define is the ridge estimators which have been described by Hoerl and Kennard in the 1970's as a class of estimators indexed by a parameter $K \geq 0$. The estimator is

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

- The idea of ridge regression is to pick a value of k for which the reduction in total variance is not exceeded by the increase in bias.
- Usually a value of k is chosen by computing $\hat{\theta}_1, ..., \hat{\theta}_p$ for a range of k values between 0 and 1 and plotting the results against k.
- This graph is known as the ridge trace and is used to select an appropriate value for k.
- In R this graph is plotted against $ln(k)$
- `https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Ridge_Regression.pdf`

- If we have the standard form of a regression equation:

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

- The equations used for estimating the ridge regression coefficients are

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

- This is the alternate method to find them
- where r_{ij} is the correlation between the ith and jth predictor variables and r_{iy} is the correlation between the ith predictor variable and the response variable $\hat{Y}$

Assumptions

- Same as Multiple Linear Regression
- $\mathrm{E}(\epsilon) = \hat{0}$ and $var(\epsilon) = E(\epsilon * \epsilon^T) = \sigma^2 I$
- ϵ_i are independent $\forall i$
- Predictor variables are not highly correlated with each other
- The variance of error terms are similar across the values of independent variables

Logistic Regression

- Response variable is qualitative
- logistic model can be expressed in the general form as:

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

Assumptions

- https://www.statisticssolutions.com/assumptions-of-logistic-regression/
- Does not require any of the previous assumptions
- Requires a binary dependent variable
- Observations need to be independent of each other
- Requires little or no multicollinearity
- Independent variables need to be linearly related to log odds
- Large sample size

Results

- We are going to break it up into 3 sections
- Multiple linear regression
- Ridge regression
- Logistic regression

Results for Multiple Linear Regression

- Two different techniques for variable selection
- AIC on the model containing every variable to reduce it to a better model
- R^2 values from the full model to create a smaller model

AIC Results

- the AIC model is:

$$\begin{aligned} points = {} & 11.8895(goalsPerGame) - 14.2963(goalsAgainstPerGame) + \\ & 0.1168(powerPlayPercentage) + 24.9296(winScoreFirst) + \\ & 26.3916(winOppScoreFirst) - 10.1651(winLeadSecondPer) + \\ & 24.6731(winOutshootOpp) + 23.2772(winOutshotByOpp) + \\ & 0.1465(faceOffWinPercentage) + 48.08239 \end{aligned}$$

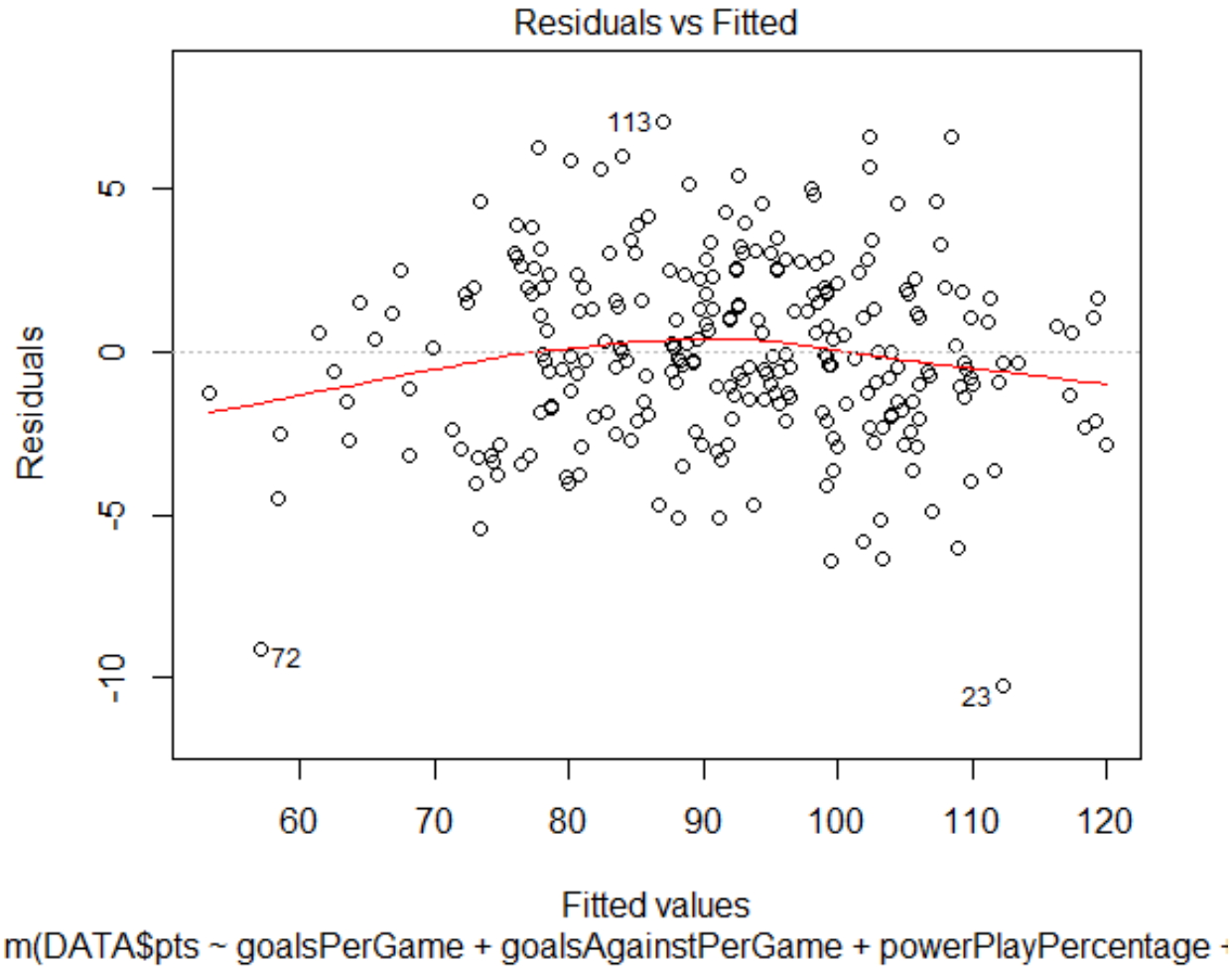

Figure: Residuals vs Fitted Values for AIC

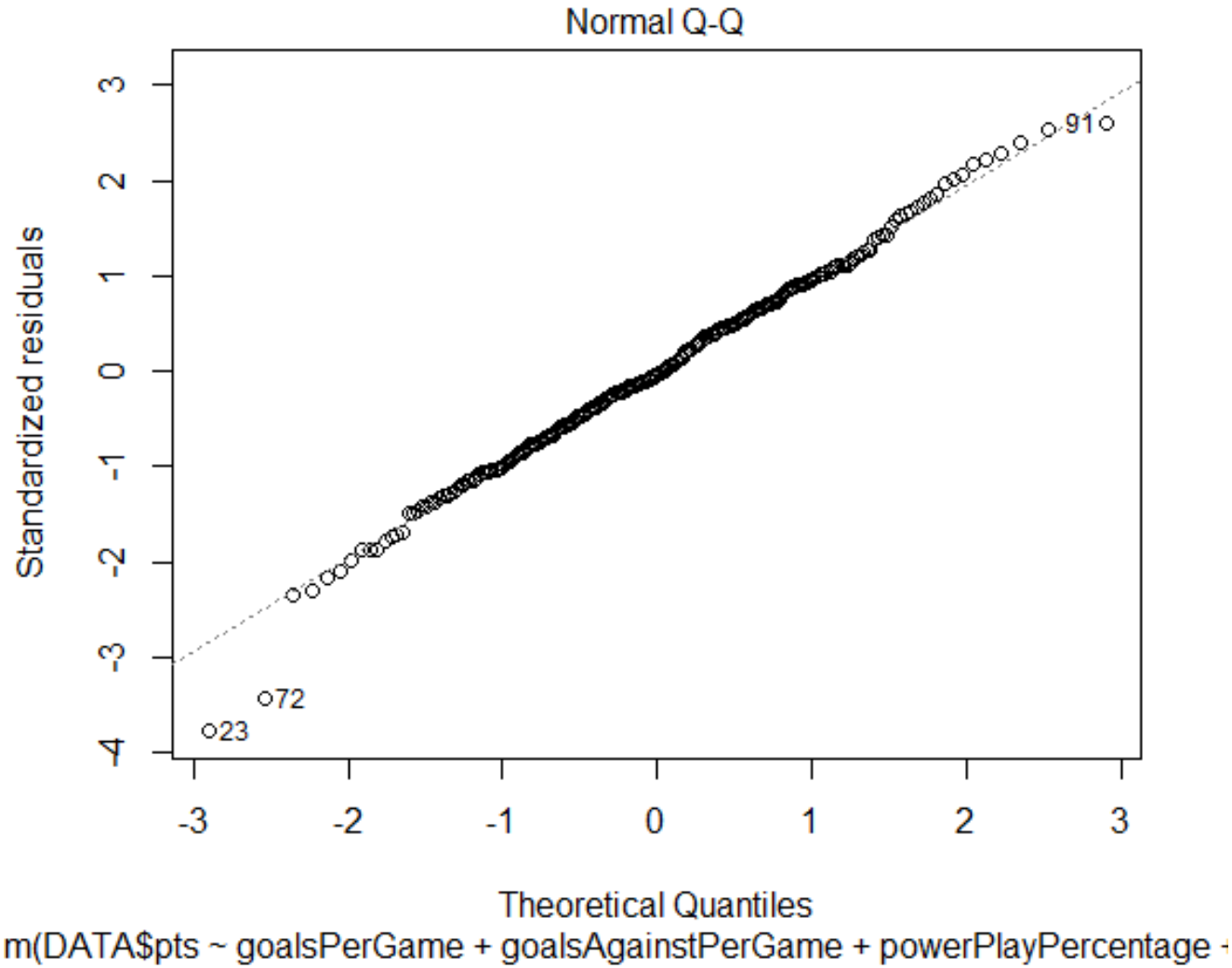

Figure: QQ Plot for AIC

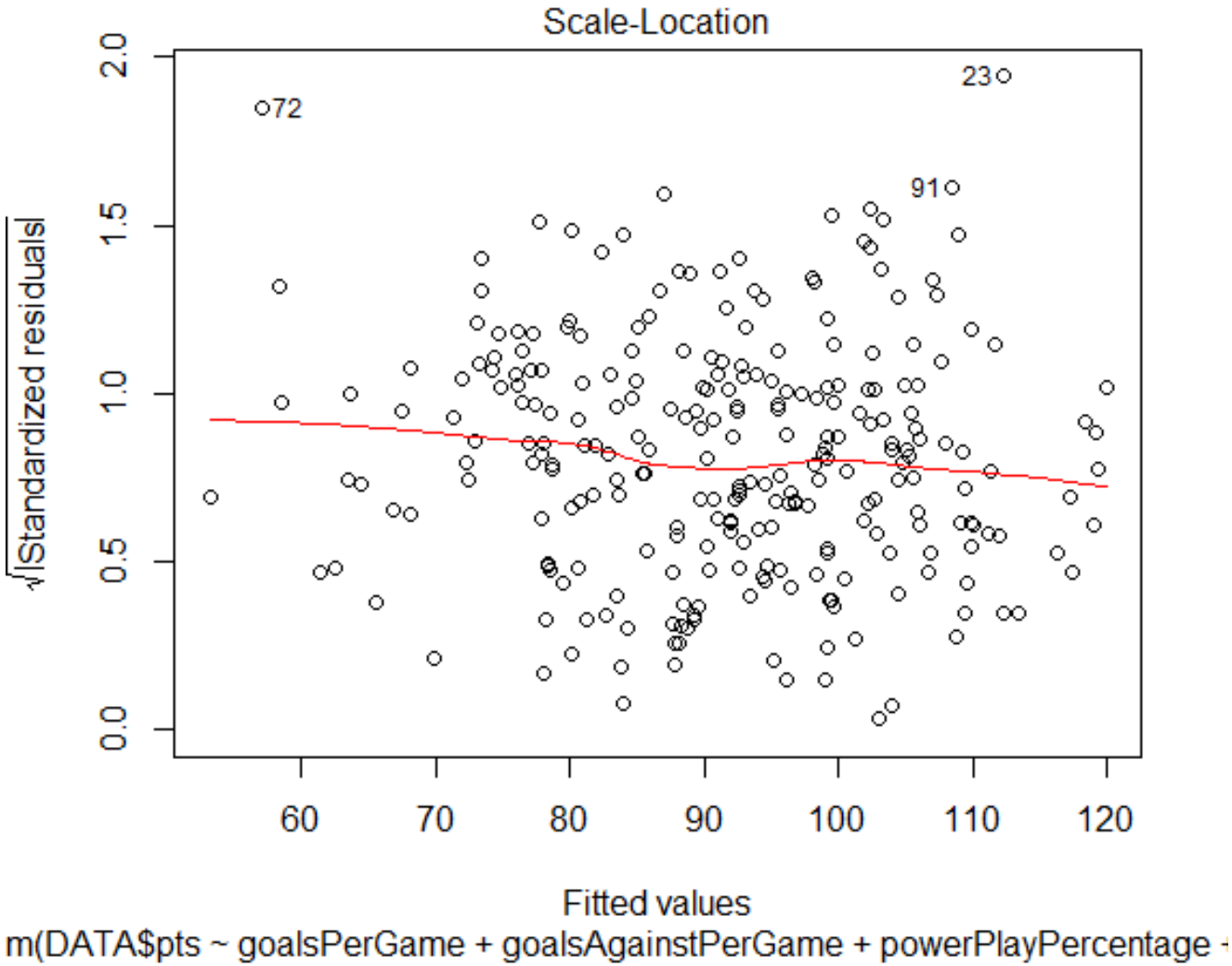

Figure: Scale-Location for AIC

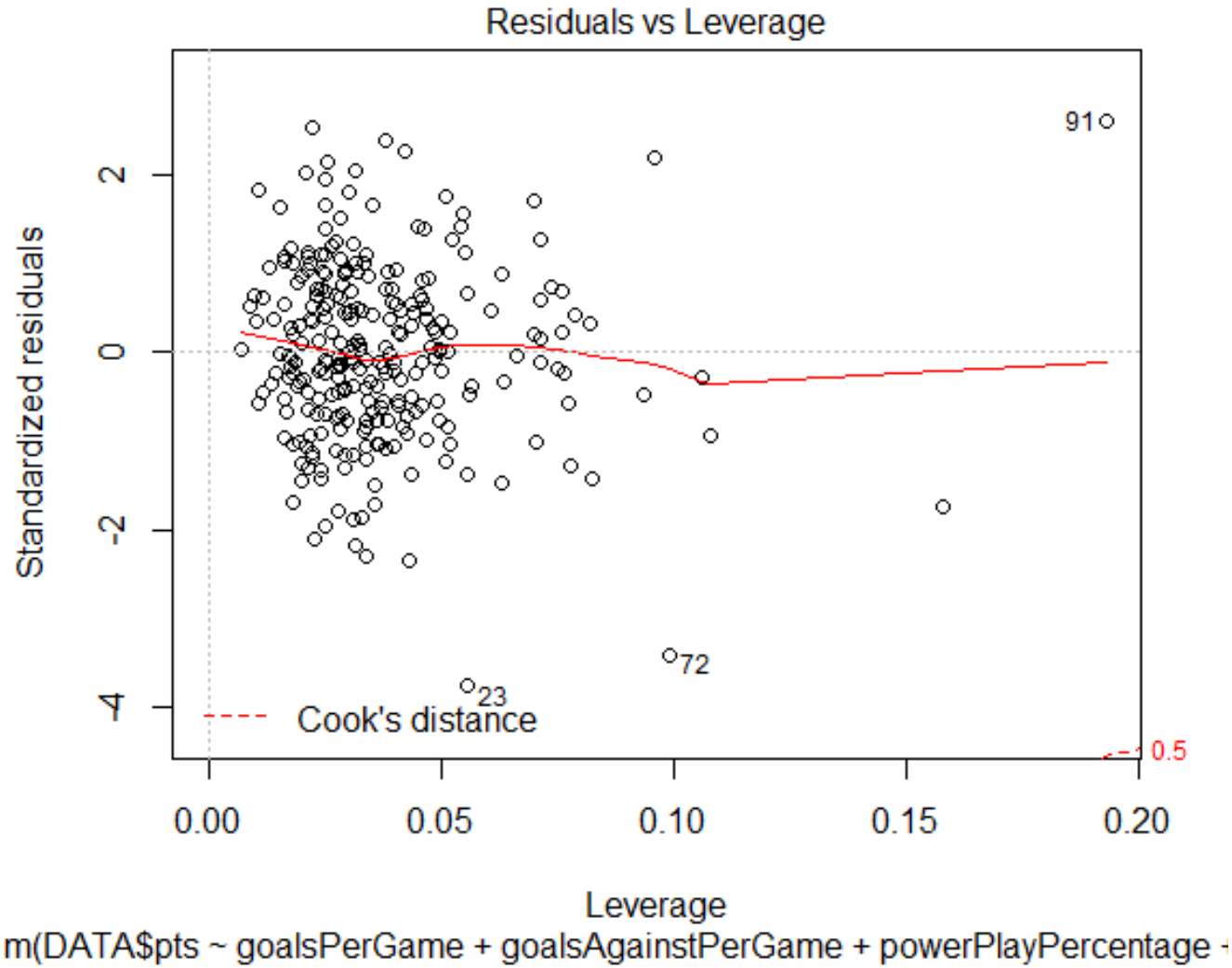

Figure: Residuals vs Leverage for AIC

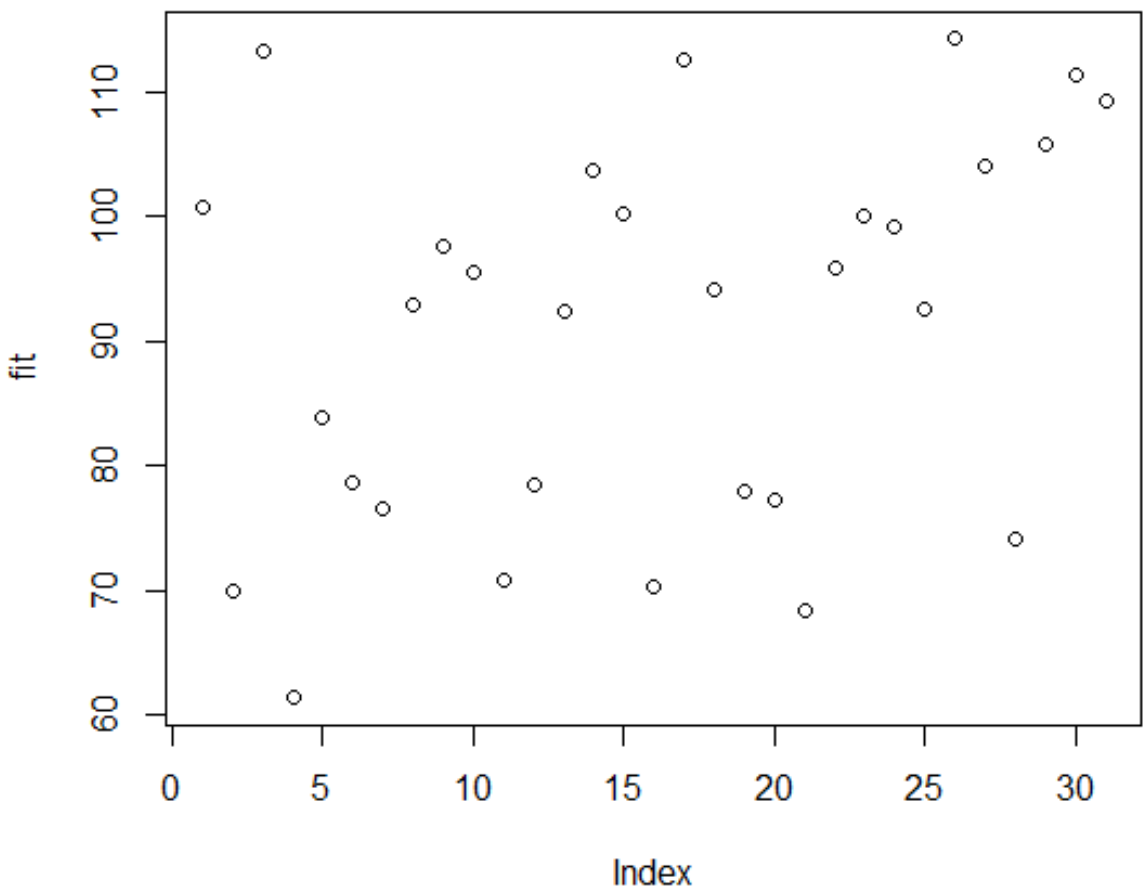

Figure: AIC Prediction

R^2 Results

- R^2 Model is:

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

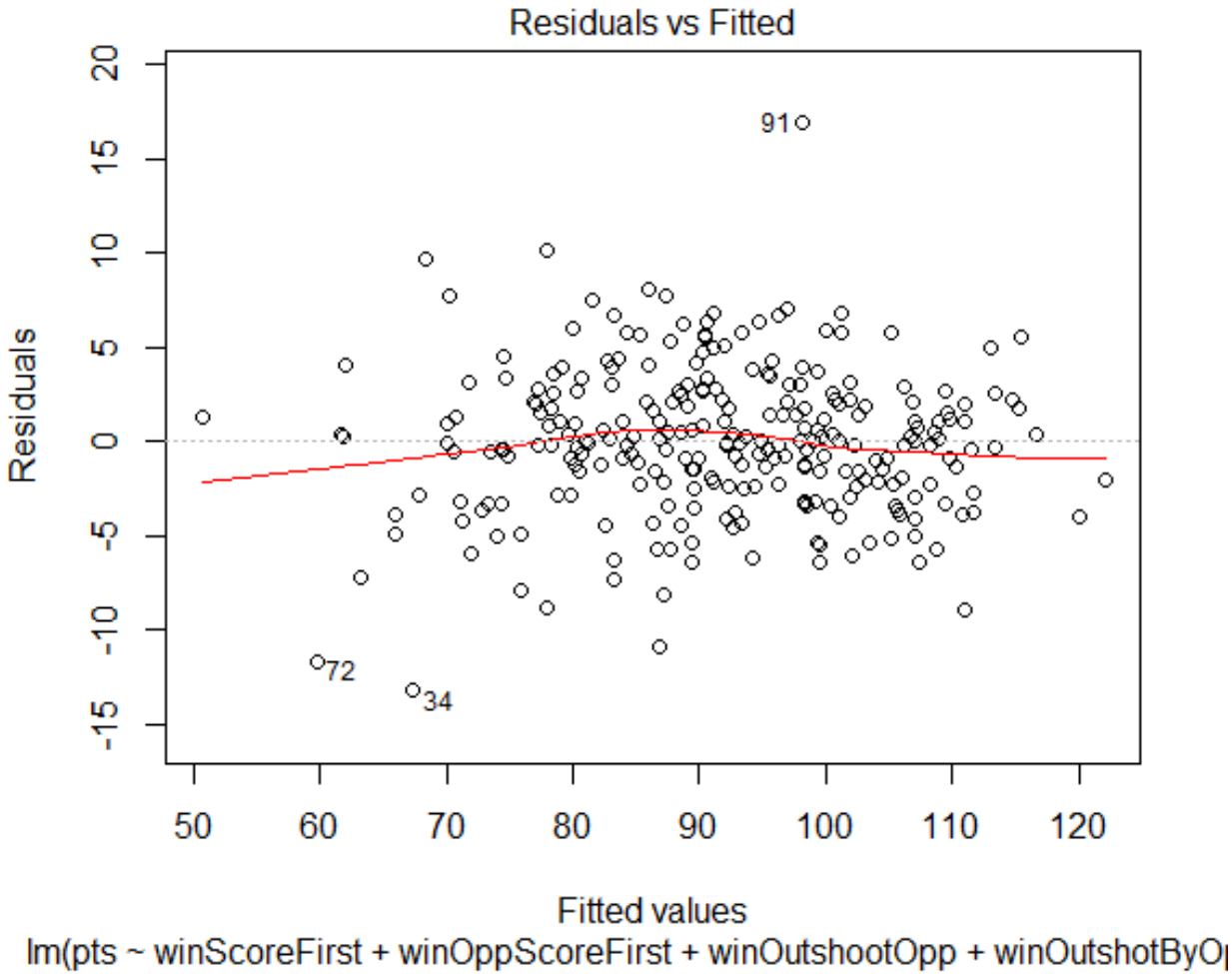

Figure: Residuals vs Fitted Values for R Squared

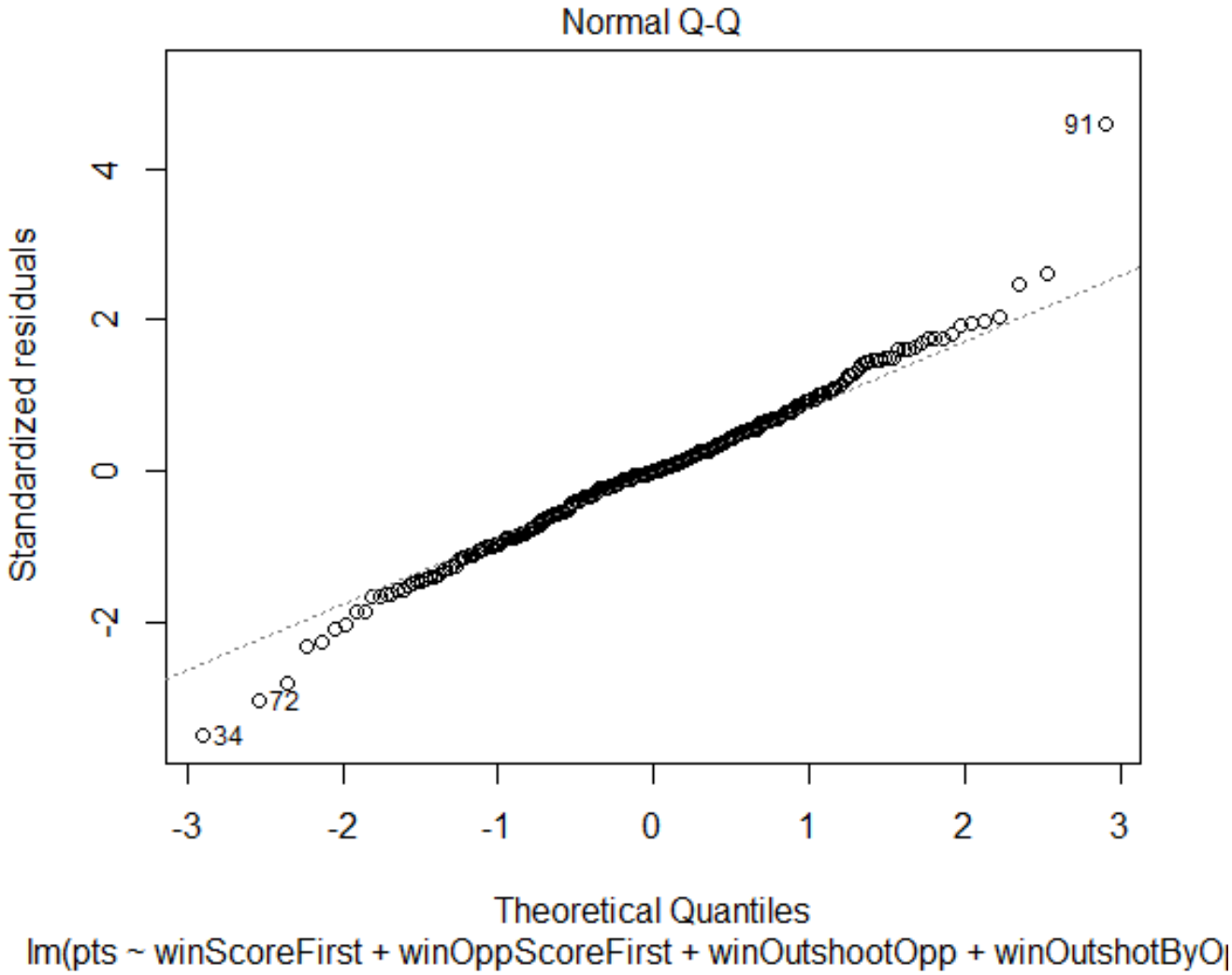

Figure: QQ Plot for R Squared

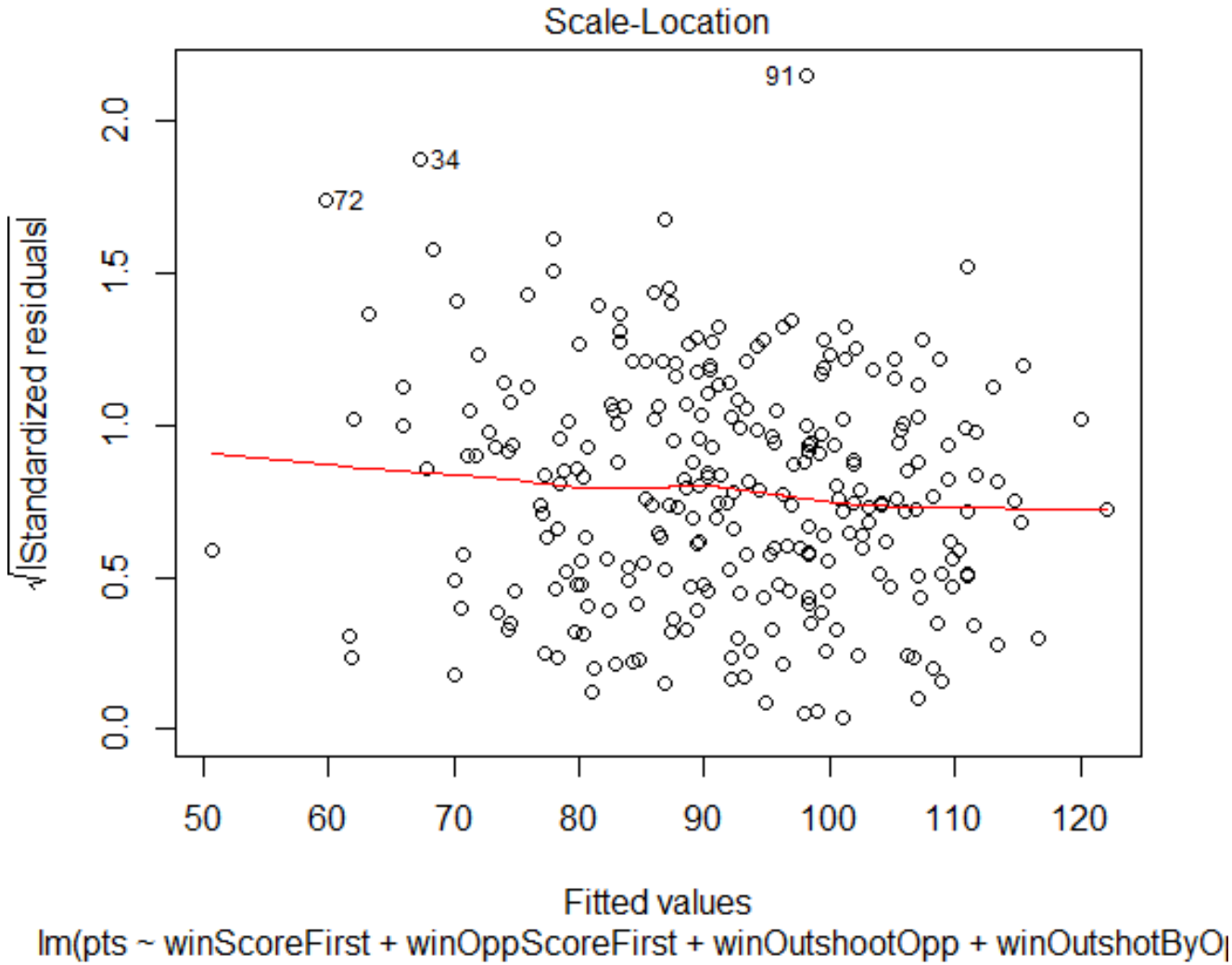

Figure: Scale-Location for R Squared

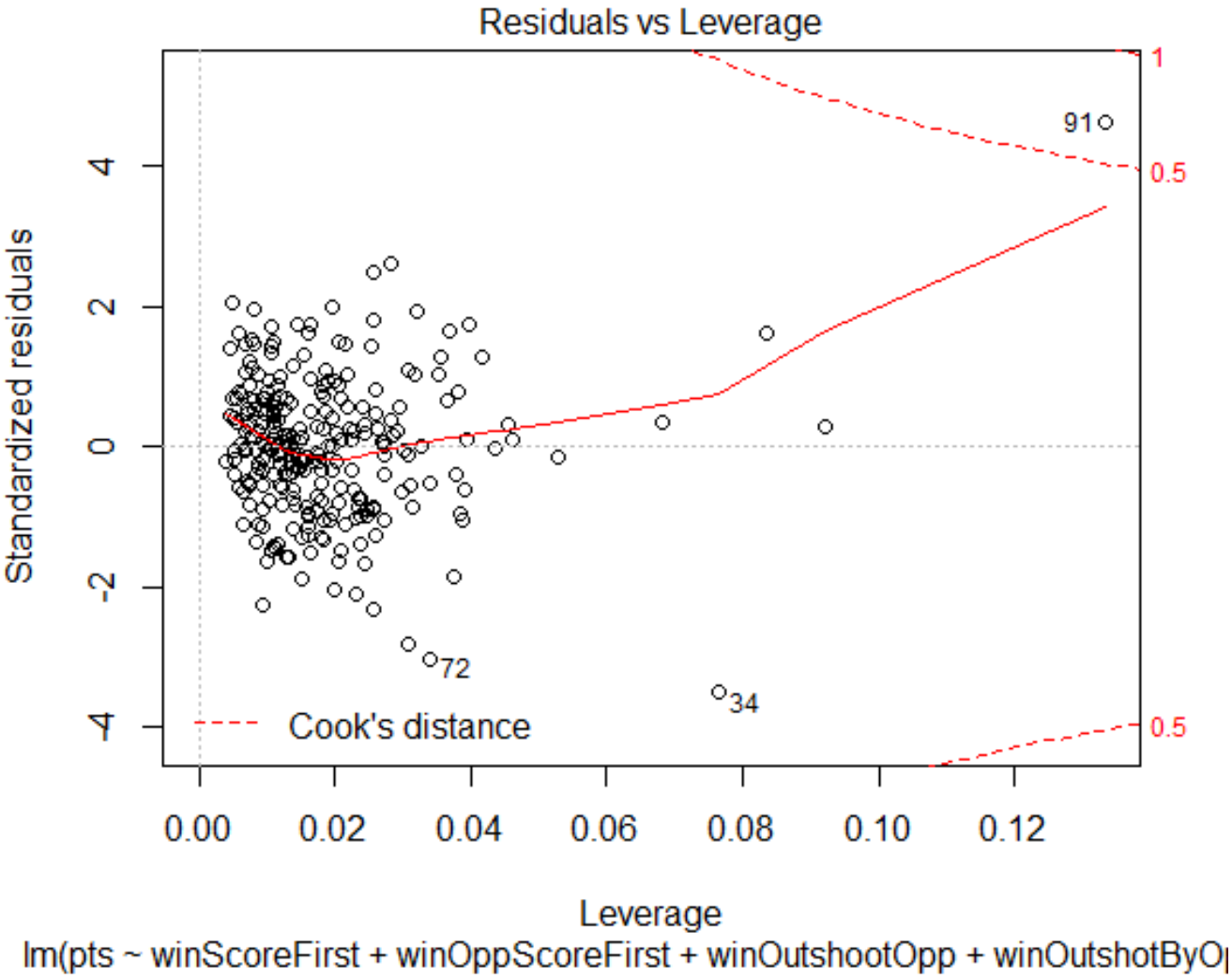

Figure: Residuals vs Leverage for R Squared

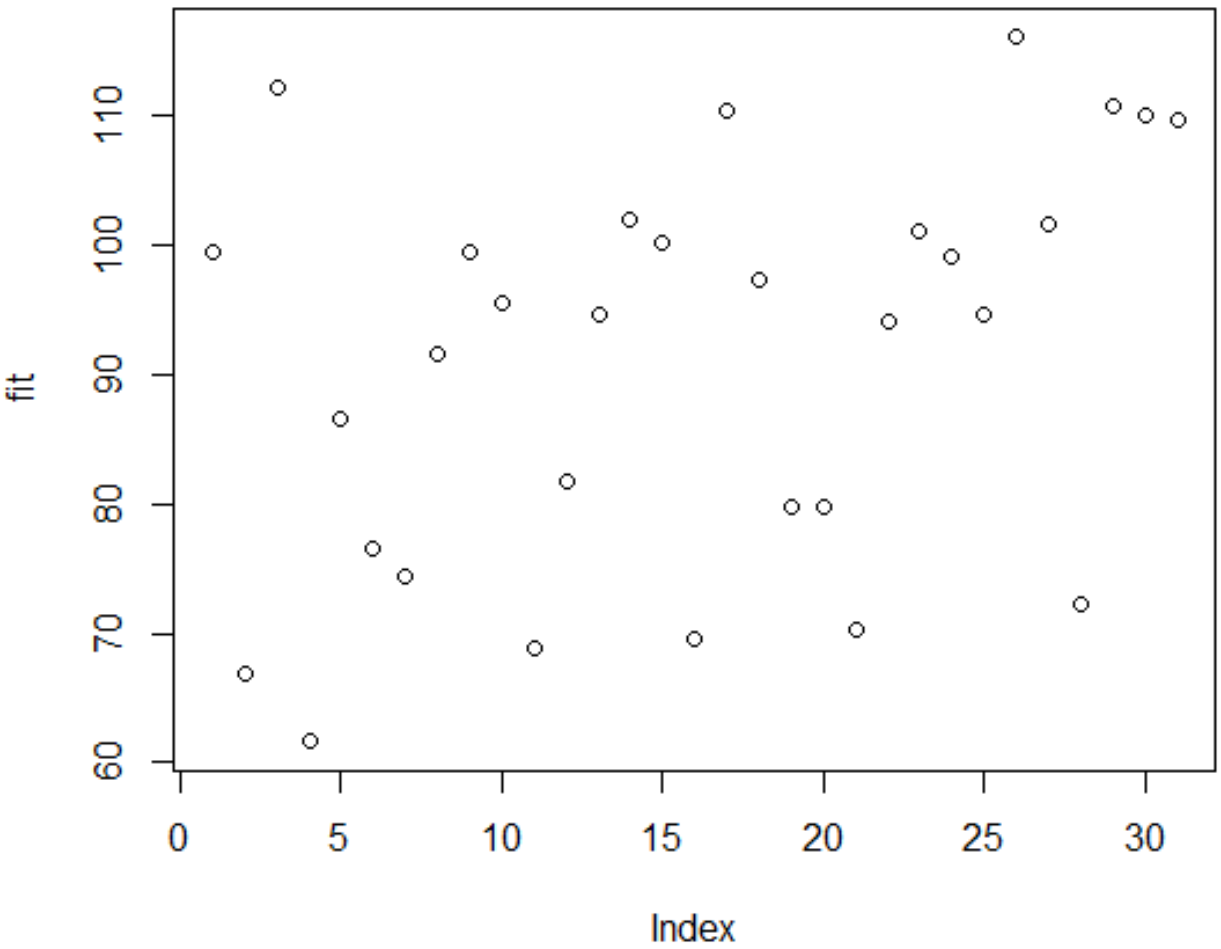

Figure: R Prediction

Errors

- What is the Error in our estimation?

$$Error = |Actual - Theoritical|$$

AIC Error

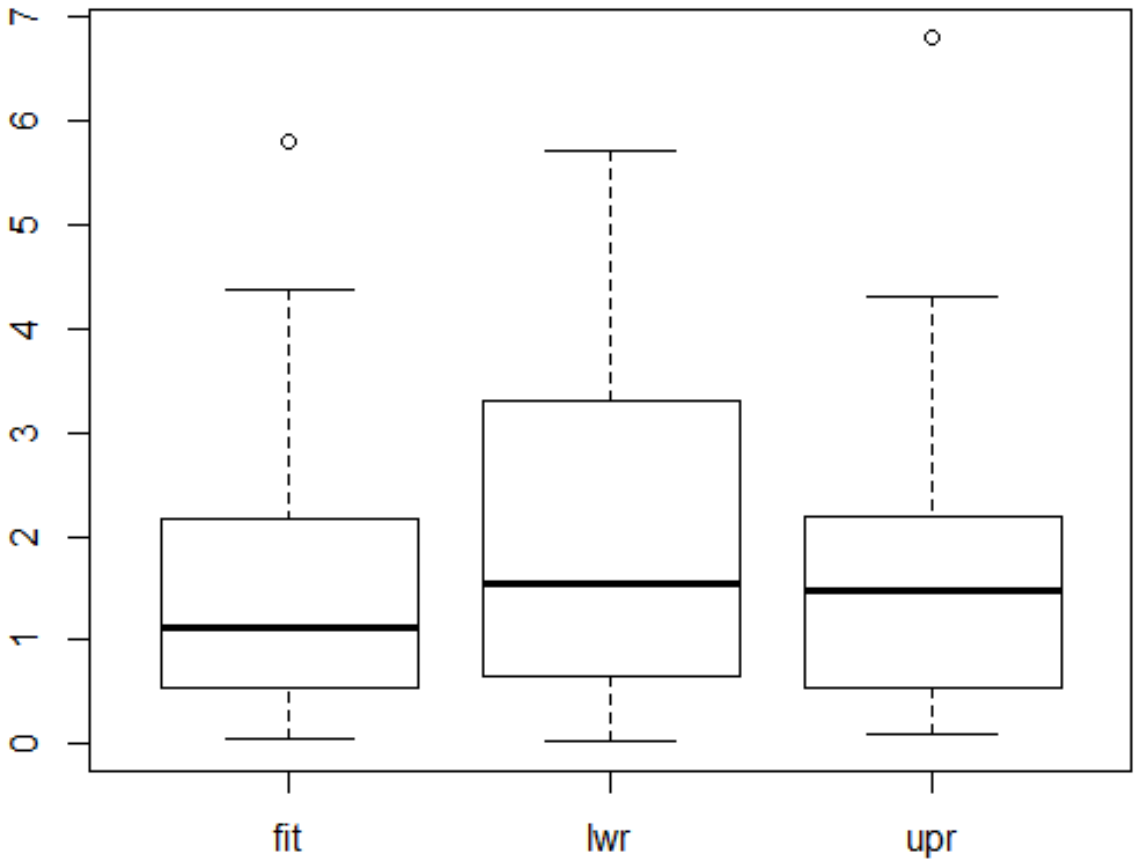

Figure: AIC Error Plot

R^2 Error

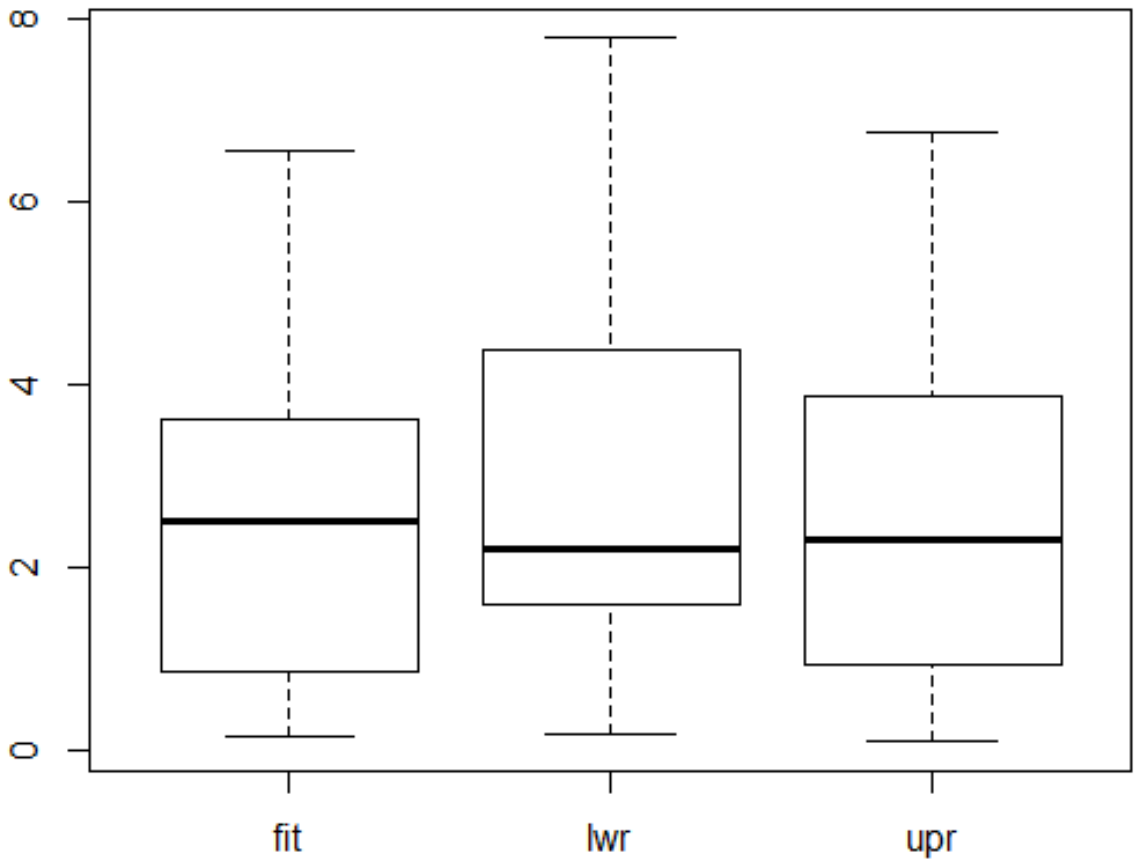

Figure: R^2 Error Plot

- For the AIC model a majority of the predictions were one to two points off the actual value
- Most extreme case is the Los Angeles Kings which were five points off what they actually had.
- 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 $-14.2963(goalsAgainstPerGame)$
- 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 regression equation.

- The second most extreme case is the Nashville Predators which were four points off their actual value
- One reason for this is like LA they were second in goals against per game at 2.488.
- 7th in wins when the team scored first
- We can draw the conclusion that good defensive teams that limit their goals against tend to have more points in this model.

- For the R^2 model a majority of estimates were off significantly
- Up to almost 4 points
- This 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 out shoot the opponent, this increased the value of their prediction.

Ridge Regression

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

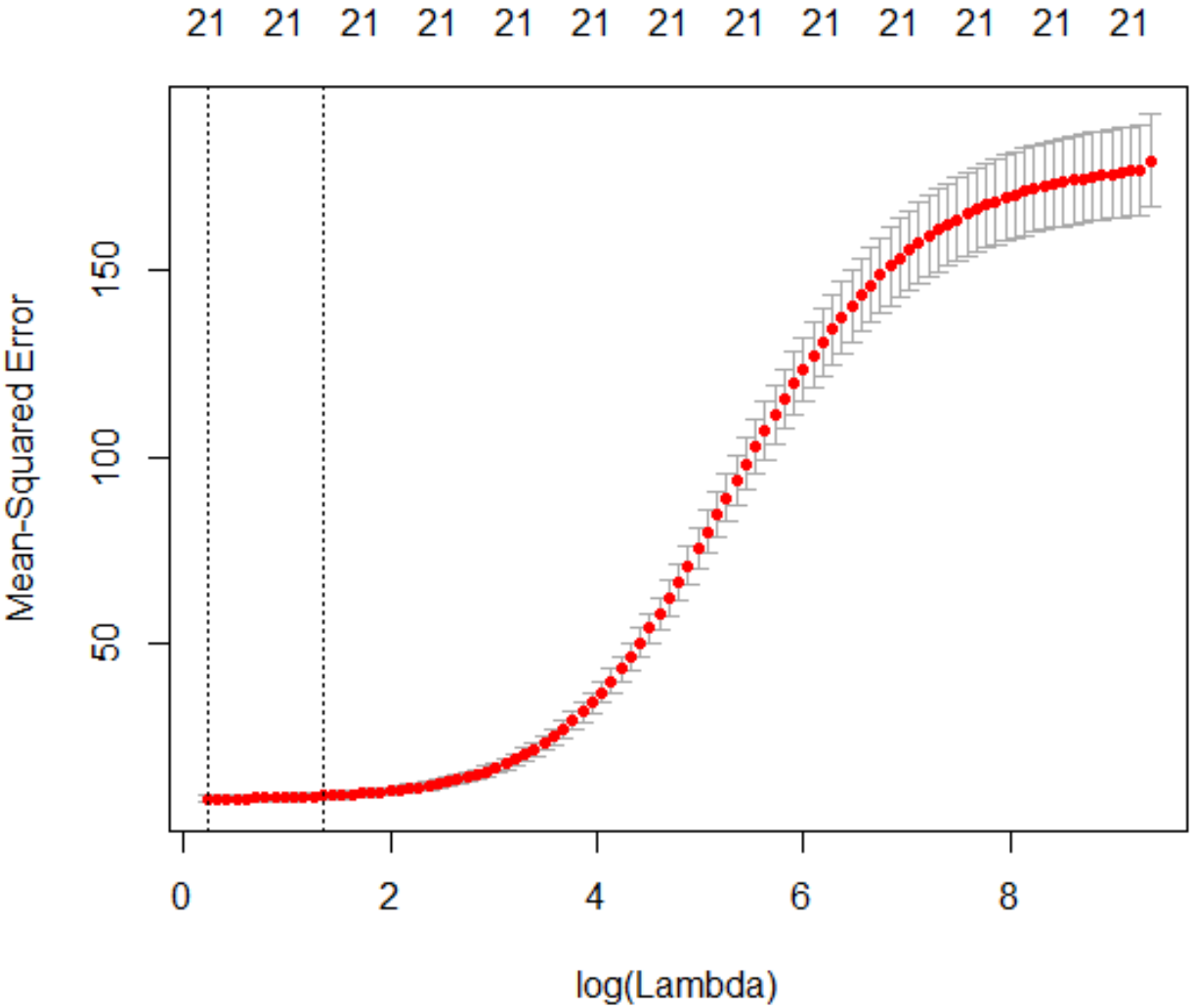

Figure: Ridge Trace

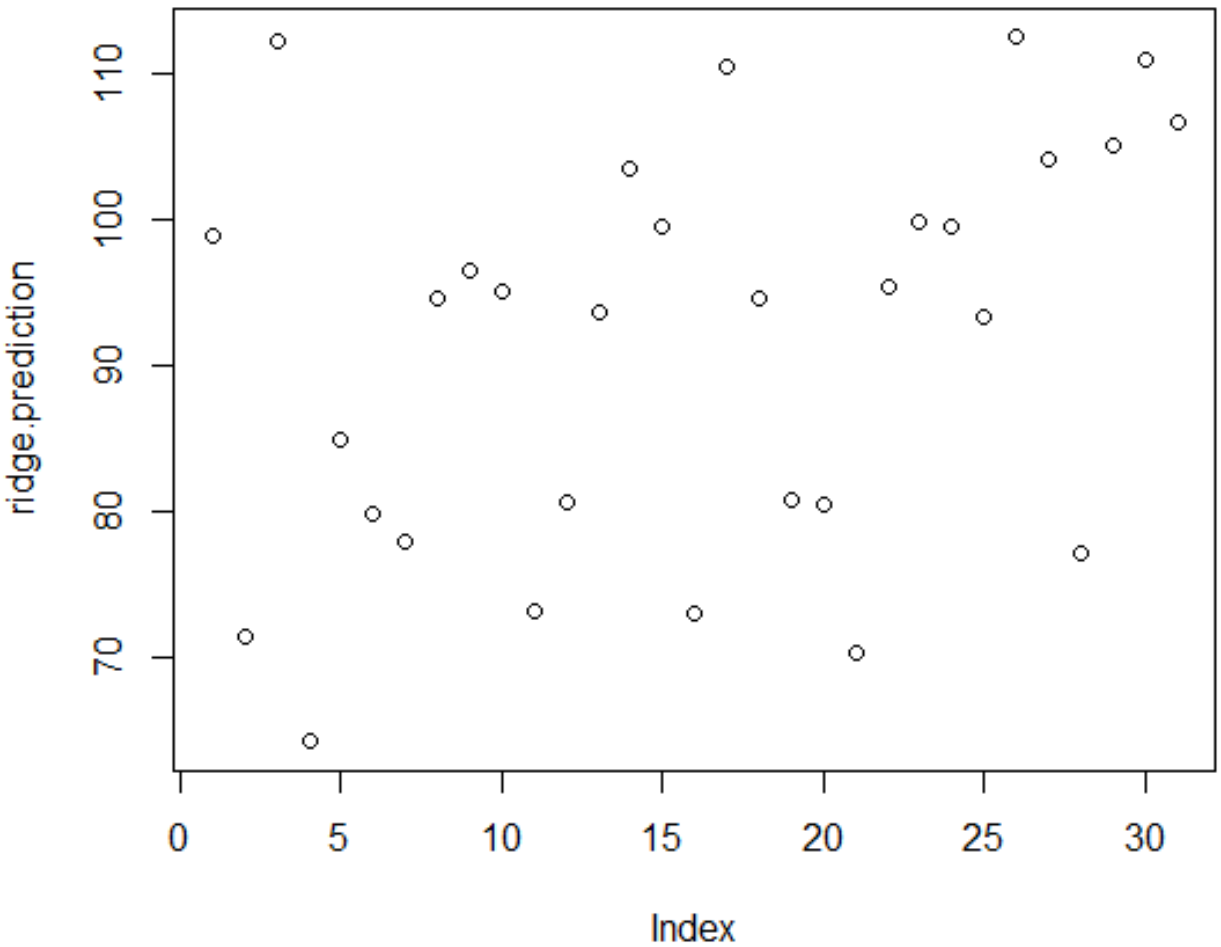

Figure: Ridge Prediction

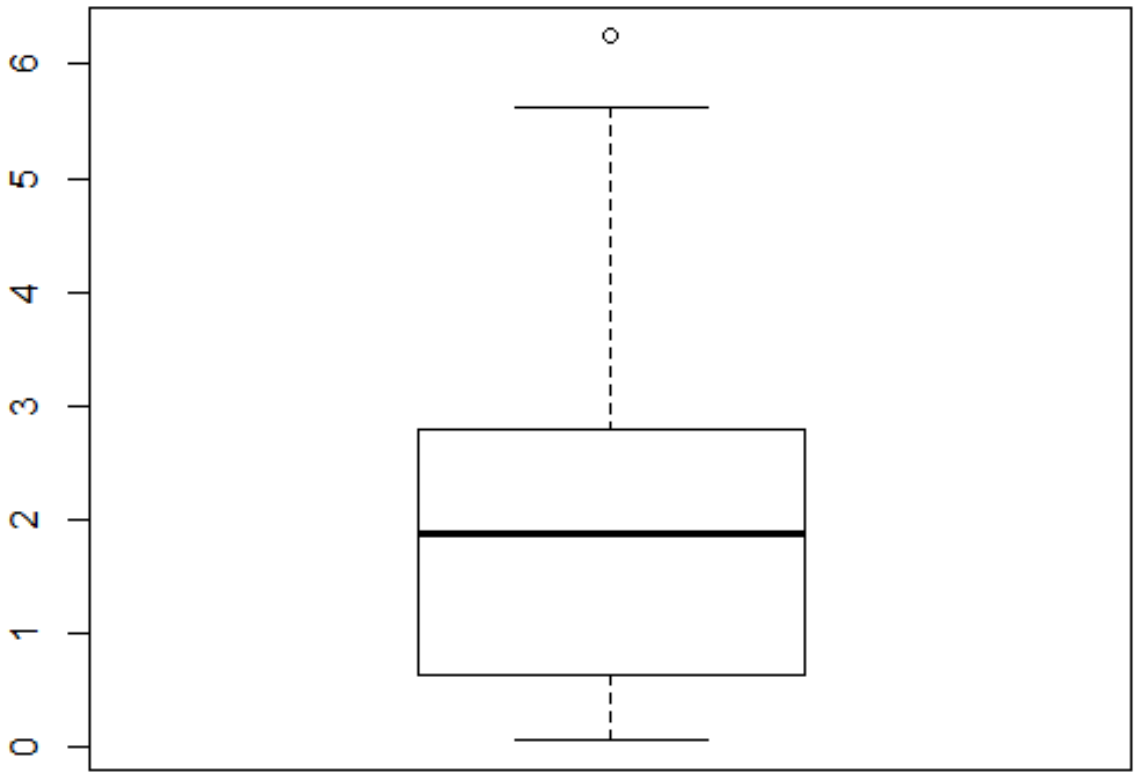

Figure: Ridge Prediction Error

- The most extreme error in this model is the Nashville Predators
- 6 points off their actual point value
- Since the ridge regression model takes every variable in our data set, it is simply due to them being in the top 15 for almost every variable in our set

Logistic Regression

- We are predicting the probability that a team makes the playoffs.
- For this model we are strictly using every variable, rather than using AIC.
- The model has these coefficients

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

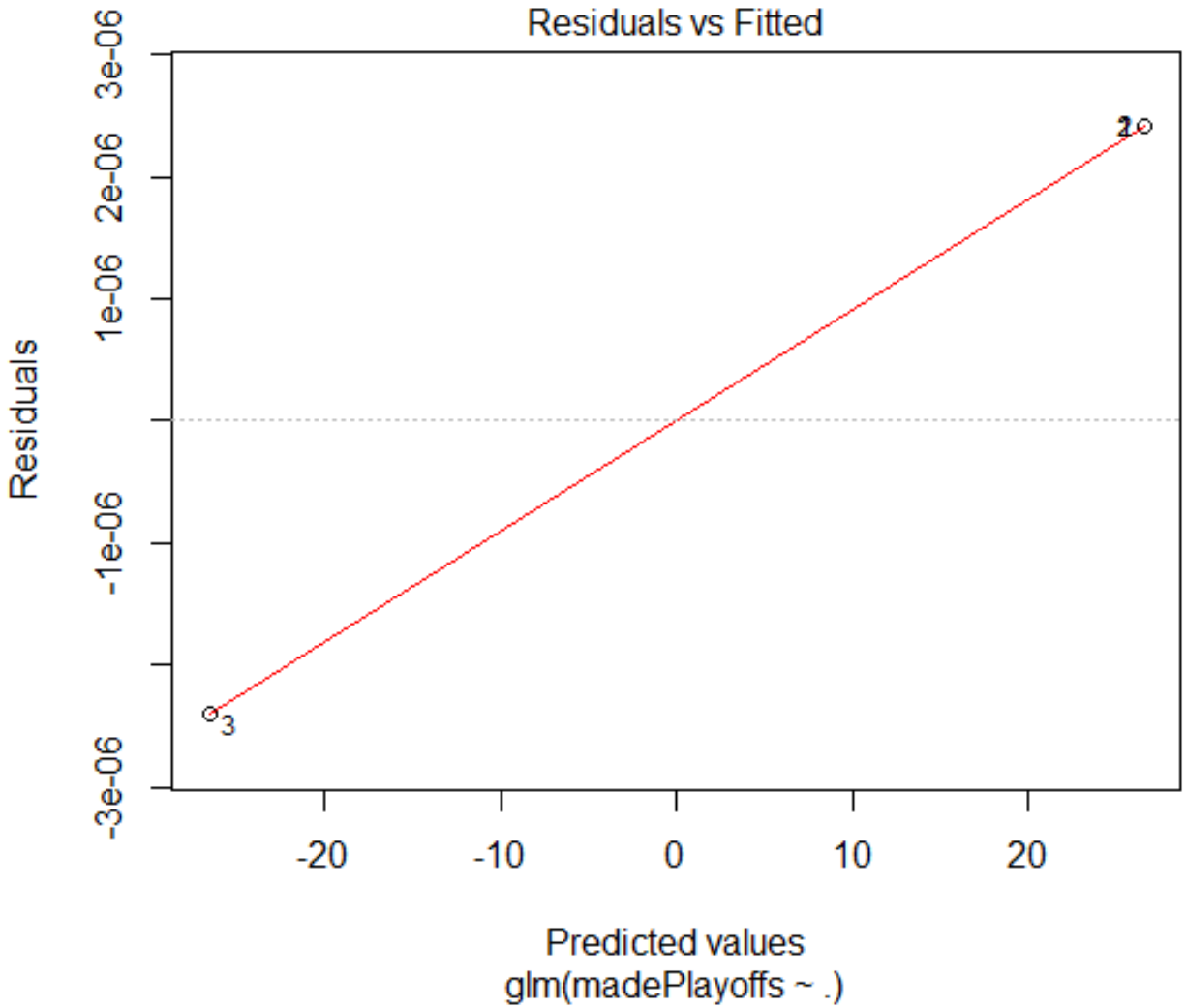

Figure: Residuals vs Fitted Values for Logistic Regression

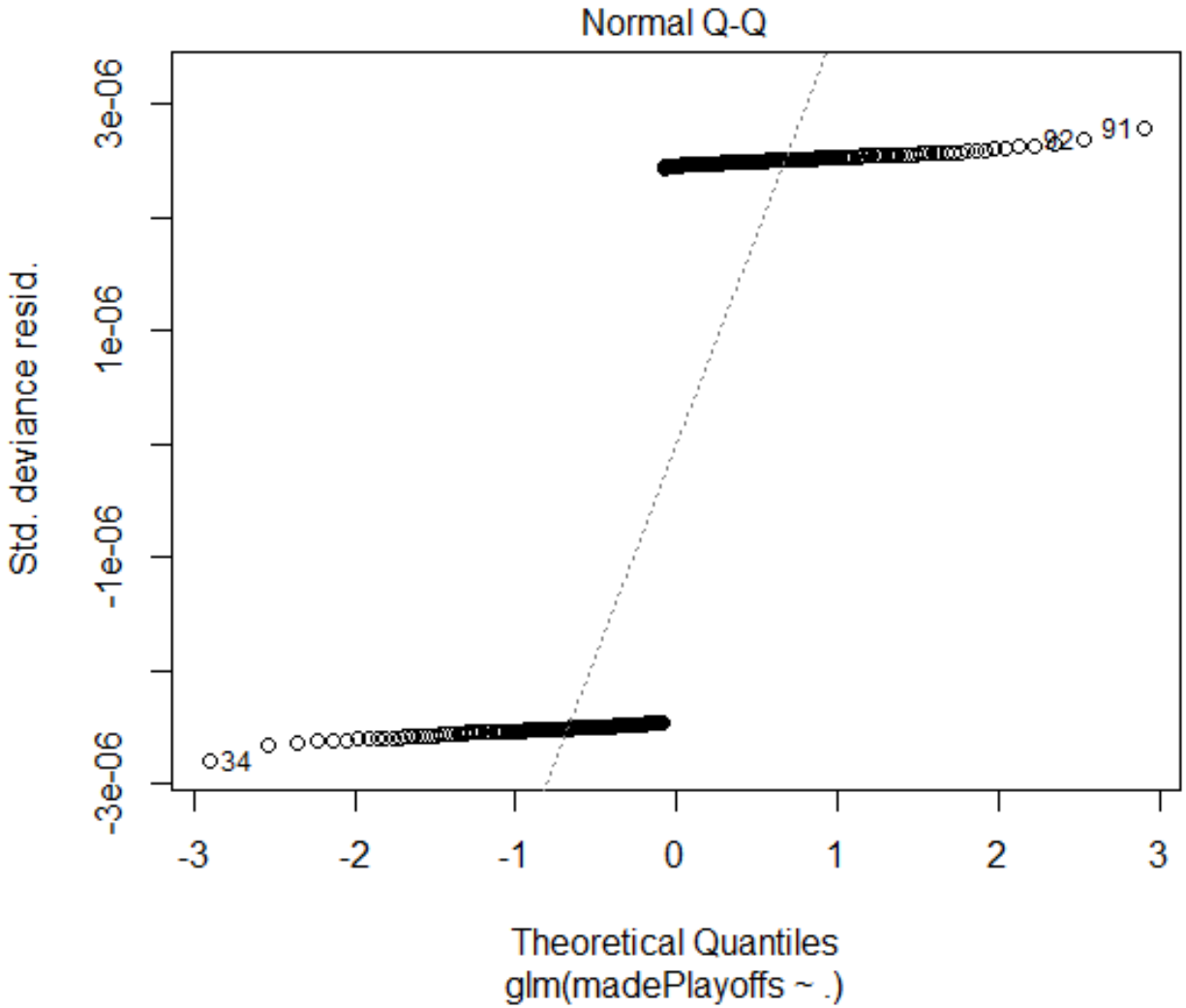

Figure: QQ Plot for Logistic Regression

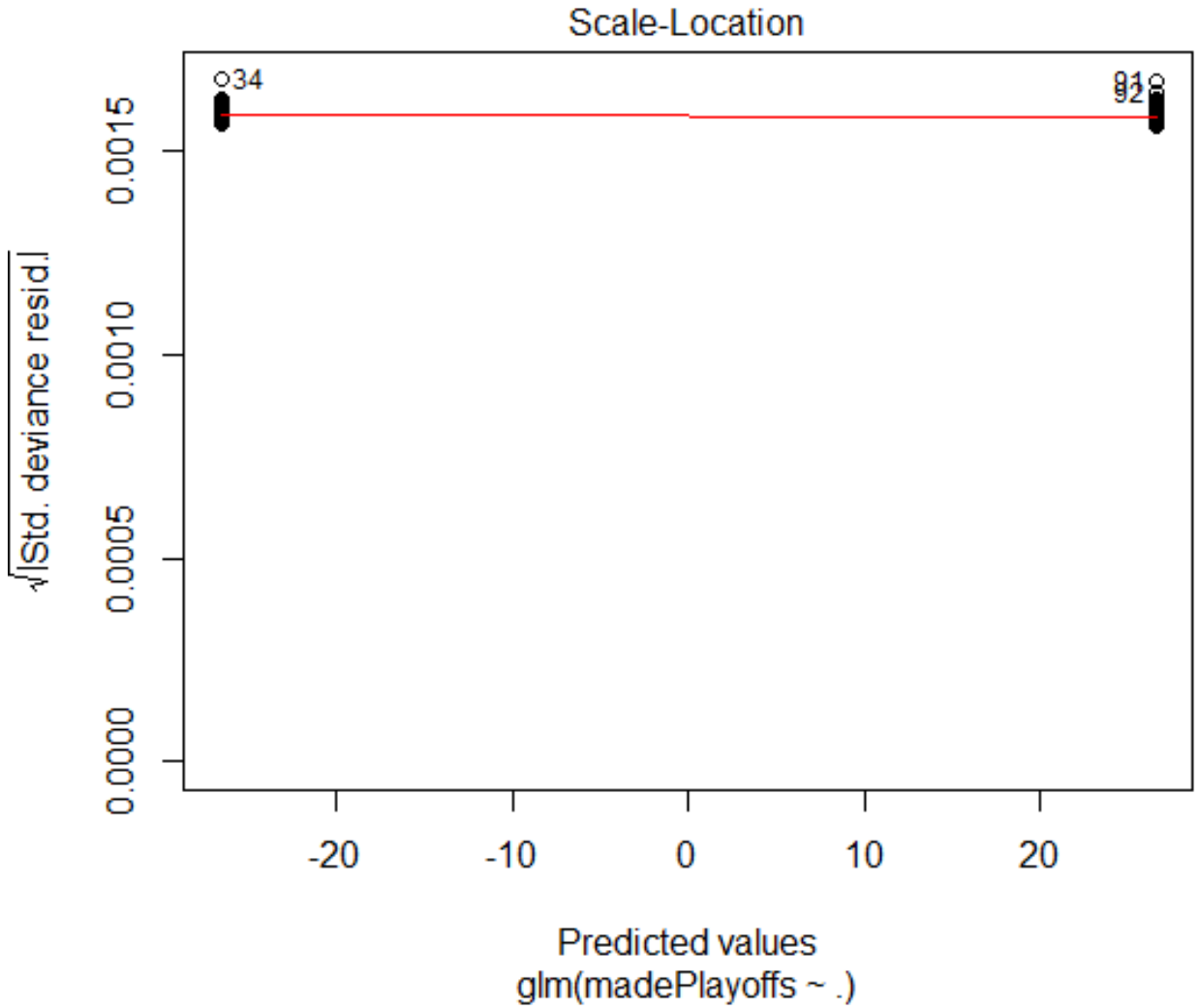

Figure: Scale-Location for Logistic Regression

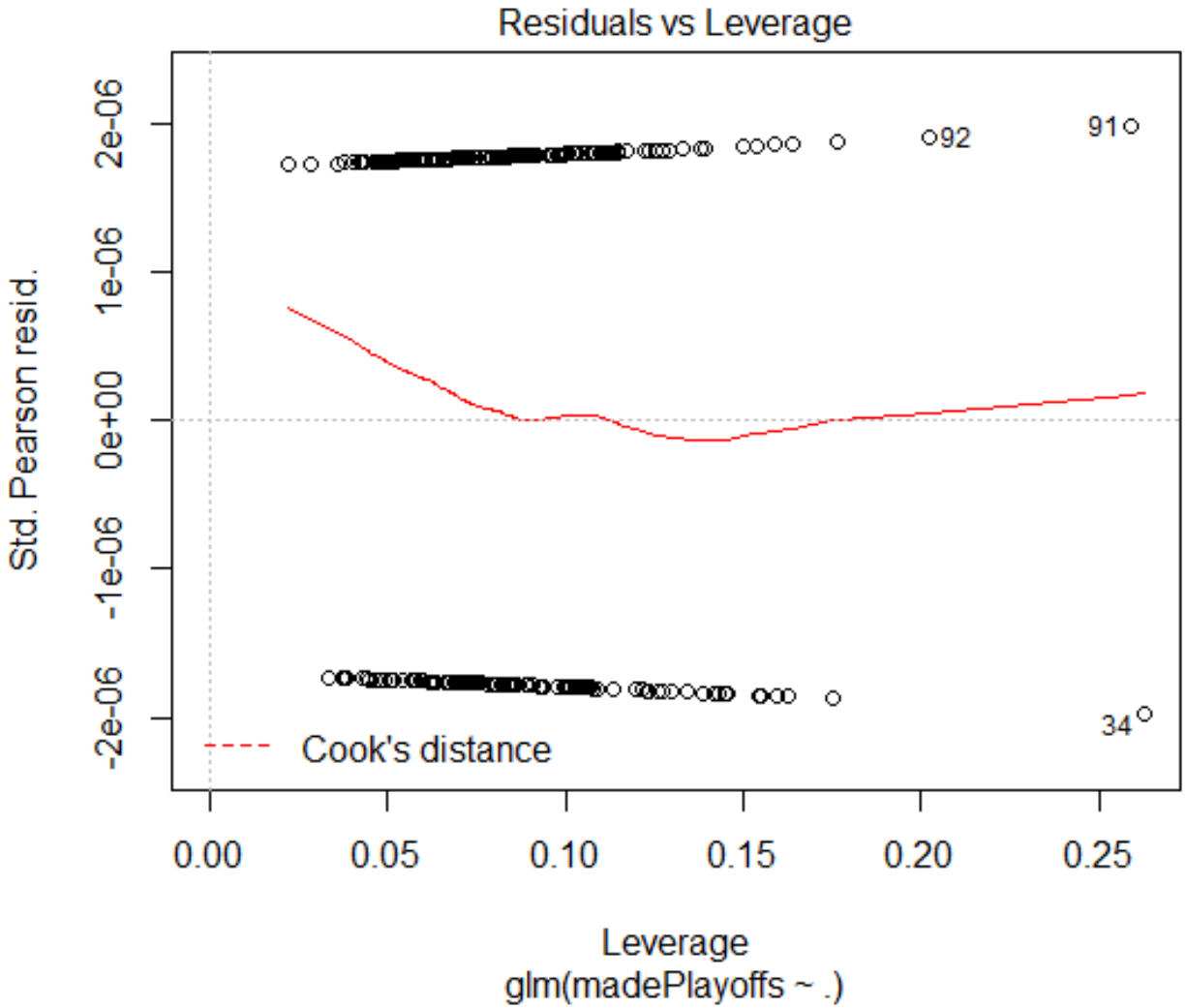

Figure: Residuals vs Leverage for Logistic Regression

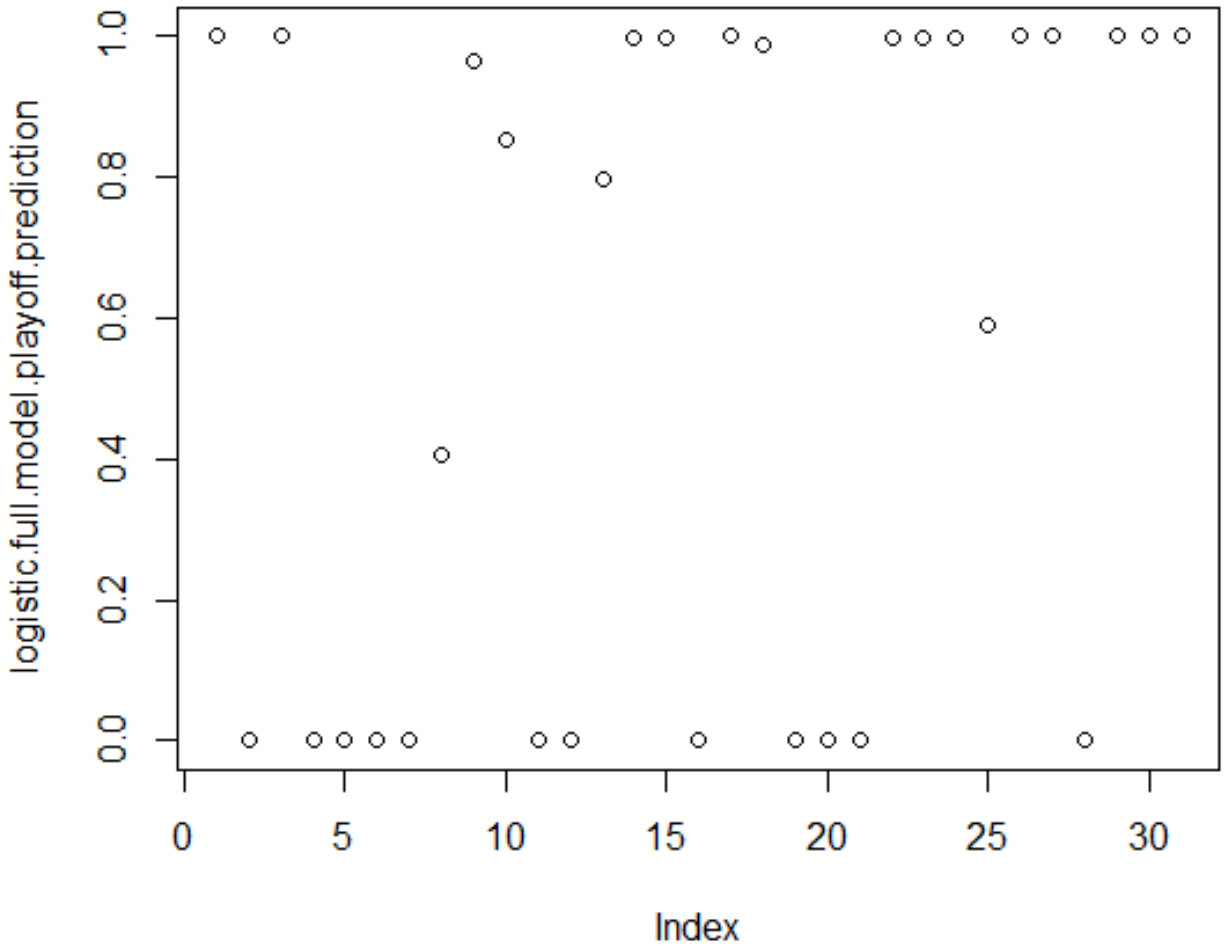

Figure: Logistic Prediction

- 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.
- Only team out of this list that did not make the playoffs during this season was the Dallas, instead Colorado made it.
- One possible reason for this is during the 2013-2014 season the NHL changed the playoff format; rather then it being the top 8 teams from each conference making it; the top 3 teams from each division makes it and the two final teams representing each conference are wild card teams.

Future Work

- The next major step would be extend 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 WAR, Corsi, etc...
- I would also like to consider other machine learning algorithms that are commonly used
- I would also like to finish up the graph based regression models

Conclusions

- In the AIC model, defensive statistics were weighted heavier then offensive ones.
- The R^2 model weighted shots very heavily due to variable selection.
- The ridge regression model weighted teams which were the top of every category very heavily.
- Finally the logistic regression model predicted every playoff team for the previous season, except for Colorado

Bibliography

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 Drew Hynes, nhlapi, January 9,2018, https://gitlab.com/dword4/nhlapi.git Micah Blake McCurdy,Hockey Viz 2018,http://hockeyviz.com/txt/preview1819,2018-2019 Regular Season Predictions https://www.nhl.com/fans/nhl-centennial NHL Standings Predictions: Offseason Edition,Larry Fisher , July 22,2018, https://thehockeywriters.com/nhl-standings-predictions-offseason-edition-2019/

The Magnus Prediction Models,Estimating Individual Impact on NHL 5v5 Shot Rates, September 24, 2018, Micah Blake McCurdy,http://hockeyviz.com/txt/magnusEV The Magnus Prediction Models,September 30, 2018, Micah Blake McCurdy,http://hockeyviz.com/txt/magnus Introduction to Probability and Statistics Principles and Applications for Engineering and the Computing Sciences,J.Susan Milton Jesse C. Arnold, McGraw Hill 2003. Regression Analysis by Example Fifth Edition,Samprit Chatterjee and Ali S. Hadi,Wiley Publication. Statistics Solutions 2018,https://www.statisticssolutions.com/assumptions-of-multiple-linear-regression Top 10 Machine Learning Algorithms,https://www.dezyre.com/article/top-10-machine-learning-algorithms/202,May 11, 2018 The Analysis Factor. Karen-Grace Martin. What are nested models? 2017.https://www.theanalysisfactor.com/what-are-nested-models/ https://www.statisticssolutions.com/multicollinearity/ https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Ridge Regression.pdf https://www.statisticssolutions.com/assumptions-of-logistic-regression/