Analysis of NHL Team Statistics

Kevin Yan
Department of Statistics
University of Toronto
kevin@utstat.utoronto.edu

1 Introduction

In all sports, especially more recently, the idea of statistical analysis is becoming more prevalent. More and more statistics are being kept as time goes on. Most of this probably originated from the moneyball era in baseball, which was documented in the book and movie, Moneyball. So the question that I want to answer in this report is whether not all these stats have any real predictive power, and which statistics should I really be paying attention to, or be concerned about. This idea will be explored throughout the rest of this paper.

2 Purpose of the study

The main goal of the study is to see how informative statistics which are kept on each individual team are of their success. We measure success by the number of wins each team has at the end of an 82 game season. Intuitively we know how to interpret these statistics, like goals are good, and penalty minutes are bad, however, it's not certain which of these statistics, are really the most telling of a team's success. Because we use the total number of wins at the end of a season as the response variable, we're aiming to study the relationship between a team's overall success and their performance statistics; we can then, based on this relationship, predict the standings at the end of the season given the present value of these statistics. We will run models which will give us some indication of which statistics are significant in their ability to predict/affect the win-loss totals for teams. Note that we will not be able to determine or predict the outcome of the next Toronto Maple Leafs vs Detroit Red Wings game, however the model should offer insight into which team will be higher on the standings come playoff time.

3 Data

The data was collected from www.nhl.com, which is the official site of the National Hockey League. Specifically, I wanted statistics thats were representative of a team's performance, and could possibly predict/have some indication of their wins/losses record; thus, the cases were teams, and statistics were gathered for each team. Overall, I chose to collect 3 seasons worth of data, summing up to 90 data points. The data was collected over the 2013-2014, 2011-2012, and 2010-2011 seasons. The 2012-2013 was omitted because it was a shortened season due to labour negotiations, and thus was not fully representative of a full 82 game season. Furthermore, data further back is not as indicative of the game as it stands today because of the many rule changes put in place through this period, mainly to ensure player safety leading to a general shift towards a more skill based game.

The statistics that I chose were also per game statistics, meaning they were not accumulations or sums, but sums that were averaged by the number of games played, thus allowing these statistics to be relevant regardless of how far into a season we're in. This idea is important because this allows us

the ability to make predictions/check our model with results of teams in the current NHL season, which is roughly complete at the time this report was written. Additionally, we'd like to be able to use these statistics to make predictions and infer the success of a team at any time of the season, therefore it's important that the statistics be on a "per game" basis.

The specific statistics that were available for collection, were: goals per game, goals allowed per game, 5 on 5 goals for/allowed ratio, power-play percentage, penalty kill percentage, shots on goal per game, shots allowed per game, face off win percentage, and statistics regarding the teams winning rate when they were either trailing after 1 period, in the lead after 1 period, scored 1st, and when they were outshot, or when they outshot the other team. Each team had one value for each of the statistics mentioned above.

However, given that these statistics were collected at the end of the year, I had to select only a subset of these to use in the study because of some particular issues related to this time sensitivity. As an example, having the winning percentage trailing after the 1st period and the winning percentage after leading the 1st period, in some way almost gives you a good idea of the final win total for a team. Therefore, the subset that I selected includes: goals per game, shots on goal per game, shots allowed per game, face off percentage, penalty kill percentage, and power play percentage; all of which are useful performance statistics for a team throughout the season. Additionally, if you watch hockey broadcasts, these are the most common statistics that you'd hear quoted by panelists.

4 Findings

4.1 Initial Model

After fitting the model, it was quite convincing which statistics seemed to be most important. The two statistics that appeared to have significant impact on the total number of wins were goals scored per game, and goals allowed per game. Further analysis shows that by scoring an extra goal per game, the odds of winning a game increase by a factor of 2.23. While reducing the number of goals allowed per game by will increase the odds of a team winning a game by a factor of 1.92. This relationship can be seen in the following plot (Figure 1), where the lighter,larger circles represent teams with the most wins, and the small, darker circles represent teams with the fewest wins. There appears to be an apparent structure, where the best teams are in the upper left hand quadrant, while the weaker teams reside more towards the bottom right. Additionally, the coefficients in the model seem to suggest that offense is slightly more important than defense.

It seems like it may be quite obvious that if a team scores a lot of goals per game, while limiting the opponents goals, then they should be winning a lot of games., which appears to be the case under this analysis. However, the reason that both these statistics were considered was that these were averages over multiple games; this means that it is quite possible for a team to have a high amount of goals per game by scoring a lot versus select opponents, and not scoring in any of their other games. The same argument follows for goals allowed per game. We will now compare the actual top 8 teams in the league (in the 2014-2015 season) and their number of wins, to those predicted by our model:

Table 1: Comparisons between the truth and model predictions

Actual wins are in bold, while the predicted wins are italicized

Ranks:	Actual Ranks:	Predicted Ranks:
1	New York Rangers **(52)**	New York Rangers *(52)*
2	Anaheim Ducks **(50)** *(46)*	Tampa Bay Lightning *(51)*

3	St.Louis Blues **(49)**	St. Louis Blues *(48)*
4	Montreal Canadiens **(48)** *(45)*	Washington Capitals *(48)*
5	Tampa Bay Lightning **(48)**	Chicago Blackhawks *(47)*
6	Chicago Blackhawks **(48)**	Nashville Predators *(46)*
7	Nashville Predators **(47)**	Minnesota Wild **(45)** *(46)*
8	New York Islanders **(46)** *(45)*	Calgary Flames **(44)** *(45)*

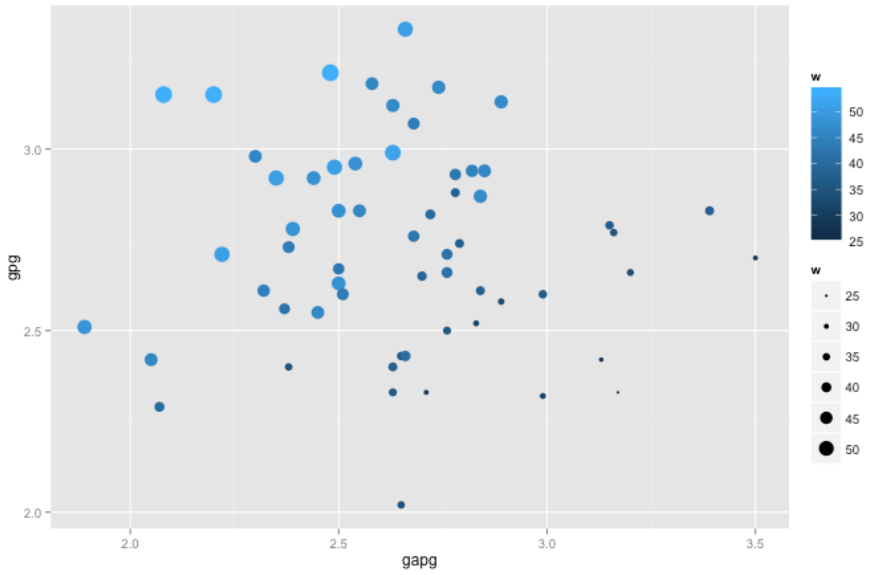


Figure 1: Wins based on goals allowed per game and goals scored per game

Overall, we were able to get half out the top 8 teams in the league with our model. However, the predictions of win totals were actually rather close, the largest difference being 6 wins. So we see that although the model including these two statistics does a rather decent job, the results show us that we do in fact need to incorporate other factors in estimating the success of teams, whether it be considering other statistics, or factors like morale and such, which are not quantitatively catalogued.

4.2 Modifications to the model

Because these two statistics were so significant, overpowering the effect of the others, it is worth considering what results we will get by removing one or even both from the model. After dropping goals allowed per game, the variables that were significant were goals per game, face off percentage, and penalty kill percentage. Scoring one more goal per game, improved the odds of winning a game by a factor of 2.26, verys similar to the model before; winning 1% more of their faceoffs would improve the odds of winning a game by 3.3%, and succeeding 1% more in their penalty kills would have a 5.5% improvement in the odds of winning a game, all of which surmount to a teams success at the end of the season. If instead, we dropped goals per game, the remaining variables that significant are goals allowed per game, power play percentage, and shots on goal per game.

It seems that regardless of which one we drop, the other one still remains significant. The other statistics in the new models then become replacements for the variable that was dropped. For example, GAPG, which is a defensive statistic, was replaced by faceoff percentage, and penalty kill percentage. These can be viewed as defensive statistics, where if you win more faceoffs, the other team has fewer opportunities to score, and killing penalties amounts to reducing the quality chances that the other team has to score. Equivalently GPG was replaced by power-play percentage and shots on goal per game, both of which lead to more opportunities for a team to score more goals. Provided this, we see that winning teams must perform well on both ends of the ice (ie: defensively and offensively), offering evidence that a team cannot just be built with a heavy bias towards one or the other.

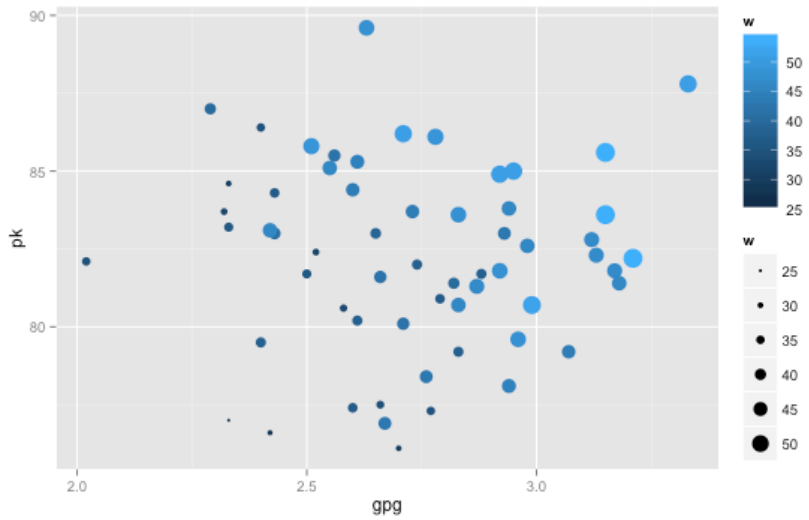


Figure 2: Model after dropping GAPG

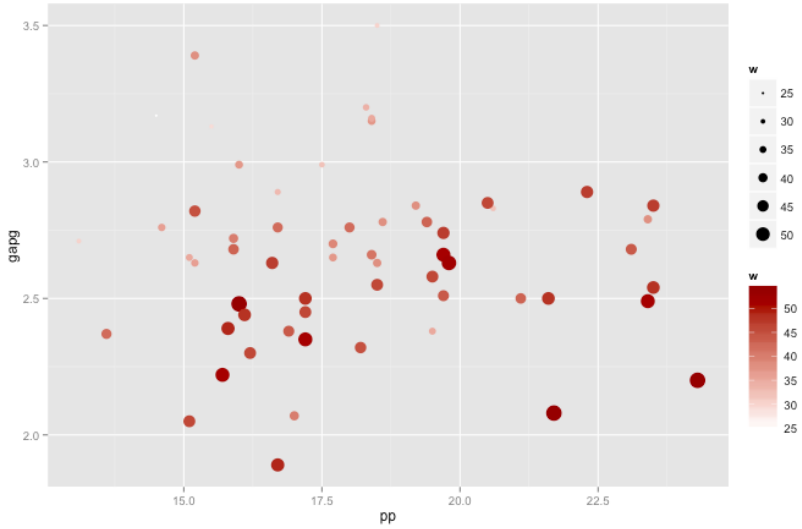


Figure 3: Model after dropping GPG

Lastly, we consider the effect of dropping both these "primary" statistics. Then the statistics that become significant are powerplay, penalty kill percentages, and shots on goal per game, and we can see that these features becomes much less informative as can be seen by the plot below (between the 2 most significant of the 3).

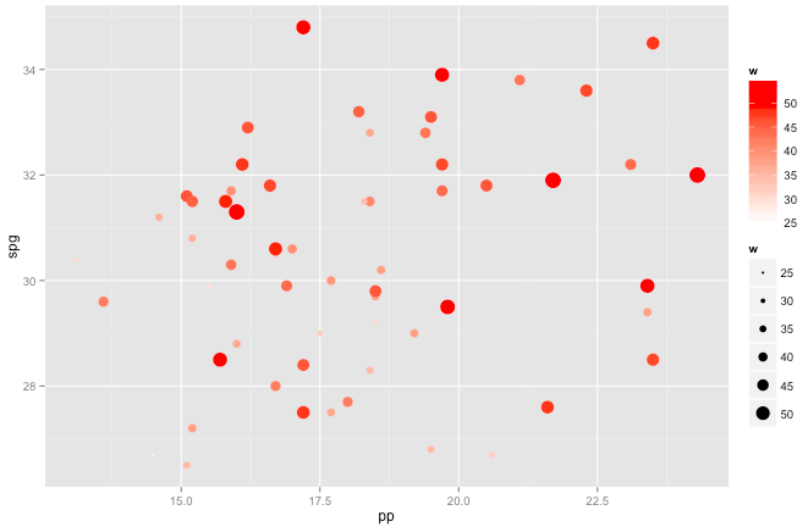


Figure 4: Model without GPG and GAPG

5 Conclusion

As we can see, depending on which statistics we include, the common denominator is that both offensive and defensive statistics are important in estimating a team's success. Though the one's that you should be keeping track of, if available, are a team's goals allowed per game, and goals scored per game. Moreover, we see that particular statistics do provide good insight into how good a team is, and should be considered in any evaluation of an NHL team. Although something to consider is that there are lots of other team statistics that were not included in the paper, or available on nhl.com; therefore we cannot say with certainty that the GPG and GAPG are the most important to evaluate team performance. Ultimately however, this all justifies the direction that sports are moving into, and that's the direction of statistical analysis and big data.

6 Appendix

The response variable of the data was based on wins, and because there are 82 games in a season for all teams, it made sense to use a binomial regression model. Under this model, we would then assume that each of the 82 games is an independent event, for which the probability of winning was solely determined by the levels of certain statistics that we included in our model. One thing to note however is that there is some relationship between the amount of wins between teams; this is the case because it's a zero sum game, if your team wins a game, it means that some other team has lost a game. Thus in order to restore some independence in the data set, I randomly selected 60 out of the 90 total data points I collected. We can see the residual plot and the normal qq plot for the binomial model below. There doesn't appear to be any relationship in the residual plot, and the variance is more or less constant for all fitted values. Thus the data seems to satisfy assumptions made by the model. There are however some outliers discovered by the model, upon further examination 2 of these consist of teams with extremely high number of wins, and the other was a team with very few wins; outliers in this case are not that meaningful because any range from 0 to 82 wins is plausible, and should not be counted as outliers. In order to obtain the model I fit a binomial regression model with all the statistics mentioned in part 3, and ran the step function to find the best model with a combination of these statistics. The results are the following:

glm(formula = cbind(w, l + ot) ~ gpg + gapg, family = binomial, data = samp)

Coefficients:

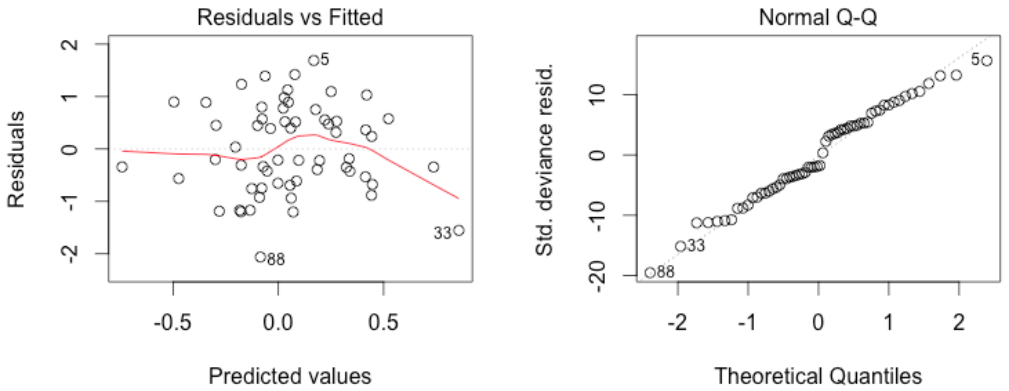
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.39785	0.38563	-1.032	0.302	
gpg	0.80461	0.10599	7.591	3.17e-14	***
gapg	-0.65557	0.09158	-7.158	8.17e-13	***

Null deviance: 138.509 on 59 degrees of freedom

Residual deviance: 22.114 on 57 degrees of freedom

AIC: 318.21

One thing to notice is that the residual deviance is less than the degrees of freedom, which seems to indicate there's no statistical difference in the ability for the current model to explain the observed data compared to the saturated model.



Alternatively, a linear model was considered. Although wins are actually discrete and range from 0 to 82, it was still a viable option. The reason the binomial model was chosen over the linear model was because wins were discrete and resembled the outcomes of a binomial distribution, where N=82. However good results were obtained using the linear model, and as we see below, the model seemed to be a good fit in terms of satisfying the assumptions made by the linear regression model.

Using all the statistics mentioned in part 3, I used the step function in order to find the best model.

lm(formula = w ~ gpg + gapg, data = samp)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	33.053	4.826	6.849	5.67e-09	***
gpg	16.155	1.314	12.293	< 2e-16	***
gapg	-13.178	1.138	-11.583	< 2e-16	***

Residual standard error: 2.798 on 57 degrees of freedom

Multiple R-squared: 0.841, Adjusted R-squared: 0.8355

F-statistic: 150.8 on 2 and 57 DF, p-value: < 2.2e-16

We can see that again the most significant statistics appear to be goals per game, and goals allowed per game. With a one unit increase in goals per game, the model estimates 16.155 more wins, and with each extra goal allowed per game, the amount of wins decrease by 13.178. We're also able to see that the model is able to explain around 84% of the total variability in the response variable (wins), which is pretty good.

Now observing the residual plot there doesn't appear to be any pattern in the residuals, and also the normal qq plot is relatively straight. Although the ends of the qq plot do seem to indicate longer/heavier tails, it seems close enough to justify the linear regression model for the data.