A SHOT QUALITY ADJUSTED PLUS-MINUS FOR THE NHL

by

Gerald Smith
B.Sc., University of Toronto, 2012

A Project submitted in partial fulfillment
of the requirements for the degree of

Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Fall 2016

APPROVAL

Name: Gerald Smith

Degree: Master of Science

Title of Project: A Shot Quality Adjusted Plus-Minus for the NHL

Examining Committee:

Barbara Sanders
Assistant Professor
Chair

Dr. Tim Swartz
Professor
Senior Supervisor

Dr. Qian (Michelle) Zhou
Assistant Professor
Supervisor

Dr. Brad McNeney
Professor
Internal Examiner

Date Approved: December 19, 2016

Abstract

We explore two regression models for creating an adjusted plus-minus statistic for the NHL. We compare an OLS regression models and a penalized gamma-lasso regression model. The traditional plus-minus metric is a simple marginal statistic that allocates a +1 to players for scoring a goal and a -1 for allowing a goal according to whether they were on the ice. This is a very noisy and uninformative statistic since it does not take into account the quality of the other players on the ice with an individual. We build off of previous research to create a more informative statistic that takes into account all of the players on the ice. This previous research has focused on goals to build an adjusted plus-minus, which is information deficient due to the fact that there are only approximately 5 goals scored per game. We improve upon this by instead using shots which provides us with ten times as much information per game. We use shot location data from 2007 to 2013 to create a smoothed probability map for the probability of scoring a goal from all locations in the offensive zone. We then model the shots from 2014-2015 season to get player estimates. Two models are compared, an OLS regression and a penalized regression (lasso). Finally, we compare our adjusted plus-minus to the traditional plus-minus and complete a salary analysis to determine if teams are properly valuing players for the quality of shots they are taking and allowing.

Acknowledgements

I would like to thank my parents for never stifling my curiosity. I am forever indebted to my Grandma Smith for contributing to my passion for hockey by graciously playing goalie in the living room when I was three years old, and for being the loudest person in the stands at nearly everyone of my organized hockey games for 15 years straight. I am also forever grateful to Dr. Tim Swartz and Dr. Qian (Michelle) Zhou for their patience, encouragement and aid while I completed this project.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Assessing individual player contributions in team sports can be challenging, especially in a free flowing game like ice hockey where players are changing on the fly and linemates are steadily being shuffled throughout a game and a season. One of the oldest measures of individual performance in hockey is the plus-minus statistic. It is a simple aggregate statistic that measures a player's goal differential. The plus-minus statistic for a player is calculated as the number of goals that were scored by the player's team while he was on the ice minus the number of goals that were scored against his team while he was on the ice. The intuition behind the plus-minus statistic is appealing: a large plus-minus statistic indicates that a player is making a contribution in that more goals are being scored for his team than against his team while he is playing. The plus-minus statistic is routinely reported on major sports websites such as www.espn.go.com/nhl/statistics.

There is a caveat to the calculation of the plus-minus statistic in the National Hockey League (NHL) that we return to later. The plus-minus statistic treats uneven-strength situations (i.e. powerplays) differently. When a team is on a powerplay, the plus-minus statistic is not credited if the team scores. However, the short-handed team does receive credit should they score.

As Gramacy et al. (2013) aptly stated: in statistical terms, plus-minus is a marginal effect. The most obvious problem with using the traditional plus-minus to measure individual performance is that a player's plus-minus is highly influenced by the quality of their linemates and the opponents they face while on the ice. Therefore the marginal plus-minus for individual players are muddled. For example Alex Ovechkin, one of the most sought after players in the NHL, had a plus-minus of -8 which ranked him as the 735th best player in the NHL in the 2011-2012 season. One can debate how high he should be ranked, but almost certainly it should be higher than 735th. Alex Ovechkin has regularly been voted to NHL all-star teams. An improved measure of performance would be the partial effect of an individual player, having controlled for the contributions of linemates and opponents.

Another difficulty with the plus-minus statistic is that it does not account for the amount of ice time that a player receives. For example, if you double a player's ice time, you would expect his plus-minus statistic to double. Yet, the quality of the player has not changed. Returning to the case of Alex Ovechkin, a reason why Ovechkin performed poorly in terms of plus-minus during the 2011-2012 season is that he played many minutes during the powerplay on a middle-of-the-pack NHL team. He was rarely on the ice killing penalties.

With the introduction of big data in sports analytics, more complex measures of player performance have been proposed in an attempt to account for game events beyond goal scoring, such as face-offs, hits, and raw shot attempts. Many of the modern approaches take a regression approach of the form

$$y = \beta x + \epsilon \tag{1.1}$$

where the left hand side of the equation represents data y corresponding to events that indicate whether something good or bad has occurred such as a goal. On the right hand side of the equation are explanatory variables x such as the composition of players on the ice at the time of an event. The objective in regression modeling is to determine the coefficients β which relate the explanatory variables x to events y. We now review some proposed player evaluation methods that are intended to be improvements upon the traditional plus-minus statistic.

Perhaps the most technical of the proposed alternative statistics is due to Gramacy et al. (2013). In the regression context (1), Gramacy considered only goals as the dependent variable and hence utilized logistic regression methods. Their methods were based on regularization where they considered various penalty terms in a classical statistical framework. They also used these penalty terms to carry out full Bayesian analyses. For their covariates x, they introduced indicator variables for the players on the ice and they also specified team effects. The advantage of regularization is that many players are estimated as having no effect. This essentially reduces the parametrization of the problem and permits more accurate estimation of the remaining extreme players (i.e. those who are both really good and really bad). A drawback with the Gramacy et al. (2013) approach is that teams do not score many goals (roughly 5.5 total goals per match). Consequently, there was a sparsity in their dataset which was their motivation for using four seasons worth of data. However, it is unlikely that a player remains at the same level of performance over a four year period. Most rookies improve from their first year in the NHL to their fourth year, and the performance of a 30 year old player will likely decline by the time they are 34. We also question the logic of the team effect variable which ”improved” estimation. From our point of view, once the 10 skaters on the ice have been identified, there is no need for a team effect.

Schuckers et al. (2011) also considered a regression approach to player evaluation where the right hand side of (1) included indicator variables corresponding to the players on the ice. We prefer this choice of covariates over the Gramacy et al. (2013) approach which also specified a team effect. For the response variable, Schuckers et al. (2011) used $Y - \mathrm{E}(Y)$ corresponding to an event. Events were defined as actions that lead to goals and included shots, hits, takeaways, etc. One of the nice features in Schuckers et al. (2011) was that they developed timeframes for which the influence of an event can lead to a goal. For example, it was determined that only within 10 seconds of a hit can the result of a hit lead directly to a goal. Then, tabulating occurrences of these events, they estimated the probability that a goal would result within the timeframe. Thus the probabilities were the terms $\mathrm{E}(Y)$ in the dependent variable and $Y = 1(0)$ corresponding to whether a goal was actually scored. Our view is that this is not entirely reasonable since a player is punished according to the metric when they do a good thing (e.g. a hit) but the hit does not lead to a goal. We see the hit as a positive contribution whether or not a goal was scored. Another criticism of the Schuckers et al. (2011) approach is that the distribution of $Y - \mathrm{E}(Y)$ is distinctly non-normal (possibly bi-modal) and therefore inference procedures based on normal theory should not be entertained.

The Corsi rating or Corsi number is probably the most discussed player evaluation statistic apart from the plus-minus statistic in hockey. For example, Corsi statistics for the 2014-2015 season can be found at www.stats.hockeyanalysis.com. It is very similar to plus-minus but considers shots rather than goals. In other words, a player is credited one point when he is on the ice when his team generates a shot. Similarly, he is deducted one point when he is on the ice when a shot is generated against his team. The attraction of Corsi over plus-minus is that shots are more common than goals and are an indicator of positive play. Of course, the Corsi statistic also suffers from the fact that it is a marginal statistic. Strangely, a quick internet search suggests that Corsi is an advanced analytics statistic.

Similar to the Corsi rating is the Fenwick rating (www.battleofalberta.blogspot.ca) which excludes blocked shots. (i.e. it considers shots on net and missed shots).

Macdonald's (2012) approach is related to the previous regression approaches. However, his modelling records an observation for each shift. The dependent variable is a rate, a rate that takes into account the length of the shift and is scaled to represent a performance measure based on 60-minutes (i.e. the length of a game). He considers measures based on goals, Corsi and Fenwick. For covariates, he also uses indicator variables to denote the players on the ice. In addition, he also uses covariates to take into account whether a shift began with a faceoff in a team's offensive or defensive zone. In addition, Macdonald (2012) uses ridge regression for estimation purposes. Whereas there are many nice features of the Macdonald (2012) approach, there may be a concern with huge rates which can occur during shifts of short duration. Also, treating all shots of equivalent value does not seem to be ideal.

With the advent of widely accessible data and increasing interest in sports analytics, there is now a considerable and growing literature on player evaluation in hockey. In addition to the papers previously described, we also mention the contributions due to Thomas et al. (2013), Stair et al. (2011) and Mason and Foster (2007).

This project also considers a regression based approach to player evalution which avoids some of the problems previously discussed. For our covariates, we likewise consider indicator variables for the players on the ice and do not introduce a team effect. For our response variable, we only focus on shots (shots on net, missed shots, unfortunately blocked shots have no distance variable in the dataset so they cannot be incorporated). The rationale is that goals are always preceded by shots, and therefore are the definitive measure of performance. Whereas there are few goals in a game (5.5 on average), there are many more shots (> 50 on average), and hence data sparsity issues are less severe. Focusing solely on shots also avoids possible confounding. When a hit or a takeaway leads to a shot, then the Schuckers et al. (2011) approach involves double counting.

Some might argue that missed and blocked shots should not be incorporated, or considered good, since they are attempted shots that never reached the goal. However, missed and blocked shots are still a positive event since the player is attempting to put the puck in the net. Furthermore it means the player has possession of the puck and is likely in the offensive zone which are beneficial contributions. An additional benefit of utilizing shots, is that collinearity will be reduced, and hence so will the size of the errors associated with the estimates. For example, Daniel and Henrik Sedin of the Vancouver Canucks are often on the ice together. By considering shots rather than goals, there are more opportunities when they may not be on the ice at the same time. Therefore, with shots, we have many more unique combinations of players than we would be available by just using goals. Another benefit of our method is that the amount of ice-time a player receives does not affect our metric. The exception is with players who play in very few games. Due to randomness they can be on the ice for nearly all positive or negative events and hence their estimates will be very large in the positive or negative direction. These players can be easily excluded from any analysis.

A final feature of our approach is that we do not consider all shots to be of the same quality. Clearly a shot directly in front of a net has a greater probability of a goal (and hence should be assigned greater value) than a shot taken from outside the blue line. We determine the value of shots using historic data where probabilities of goals are estimated from various locations on the ice.

The remainder of the project is organized as follows. Chapter 2 outlines the methodology used in our approach for player evaluation. We describe data and the regression models with specific emphasis on the dependent variable. The dependent variable is a score which is assigned to a shot. The value of the score is based on historical data which takes into account the success rates of goals from various locations on the ice. Chapter 3 describes the data and we fit the model. We then provide comparisons of our player evaluation metric with the plus/minus statistic and salary data. Chapter 4 contains a conclusion including limitations of our work and ideas for future research directions.

Chapter 2

Data and Model

2.1 Data

Salary data was downloaded from spotrac.com. The play-by-play data was downloaded from nhl.com using A.C. Thomas' *nhlscrapr* package in R. It includes information about the teams playing, whether an event occurred in regulation time or overtime, which players are on the ice for an event, home and away indicators, x-y coordinates and a distance measure for where shots on goal, and goals originated from, as well as a distance measure for shots that missed the net. A.C. Thomas' code downloads and processes each individual game into a tab delimited text file. These text files can subsequently be read into one large data frame in R. Each row of the data frame is a play-by-play event during a game (ie. goal, shot, hit, blocked shot, etc). A shot on goal is defined in the data as a shot a player takes that is stopped by the goaltender, this includes shots that would have missed the net if the goaltender did not reach out to save it (shots that are stopped by players other than the goalie are not recorded as shots on goal, but instead are recorded as blocked shots. Unfortunately these do not have any distance measure or x-y coordinate associated with them). For the purposes of this study, an ”event” will be defined as either a goal, shot on goal, or a missed shot. The x coordinate represents the distance from the centre ice red line towards the attacking goal; the distance from centre ice to the goal line is 89 feet. The y coordinate represents the distance from the centre of the net, these values are positive if the shot originates from the left of the attacking goal, and negative otherwise. The distance from the centre of the net to the boards is 42.5 feet.

Since teams shoot at different ends of the rink, the x-y coordinates will vary depending on which team is shooting and which of the three periods the game is in. In order to create a map for the probability of scoring a goal we needed to reclassify the x-y coordinates so they all took place at one end of the ice (the offensive zone where the x coordinate is positive). Therefore, if x was negative both x and y needed to be flipped (ie. X becomes -X, and Y becomes -Y), if x is positive then everything remains the same. For example, a shot originating from (-25, -15), that is a shot in the defensive zone coming from the left of the net, became (25,15), the y coordinate must be flipped to ensure that the shot is still coming from the left of the net, therefore the proper zone and which wing the shot is originating from is maintained.

As stated above, x-y coordinates are only provided for shots on goal and goals. For missed shots only a distance measure is provided. Making use of code provided by A.C. Thomas on his nhlscrapr github website, the x-y coordinates were imputed for missed shots using the distance measure and x-y coordinates for shots on goal and goals. For each missed shot the distance measure is used and a random selection with replacement is taken from the sample of shots from the same distance to generate an x-y coordinate for the missed shot. Furthermore, since shot locations are recorded by hand it is known that some of the data points for distance are incorrect at some arenas (most notably the Madison Square Gardens) so these have been adjusted so they are more consistent with the entire league, this is done using a probability integral transform.

We restricted our analysis to when teams are at full-strength (5 skaters and a goalie for each team), but the analysis could easily be adopted to handle other situations. The probability of scoring from the neutral zone during 5 on 5 play is a minuscule 0.3% so we only included events occurring in the offensive zone. Goaltenders are not included due to the fact that they play a small part in where shots are taken from (except for rebounds that they allow).

To determine the probability of scoring from each location in the offensive zone, data from the 2007-2008 season up until the 2013-2014 season was used. A general additive logistic model (GAM) was employed to create a smoothed probability map. The bam function in the *mgcv* package in R was used, the benefit of the bam function is that it is meant for very large datasets and has a built in argument for computing in parallel. The model is formulated as:

$$logit(p_i) = \beta_0 + \beta_1 s(X_i)$$

where the response $Y_i \sim \text{Bernoulli}(p_i)$ is an indicator of whether the event resulted in a goal (1=Goal, 0=Shot on Goal or Missed Shot), and X_i is a vector containing the x-y coordinate for the ith event, the s indicates that the linear predictor depends on a smooth function of the predictor and ϵ_i is an error term.

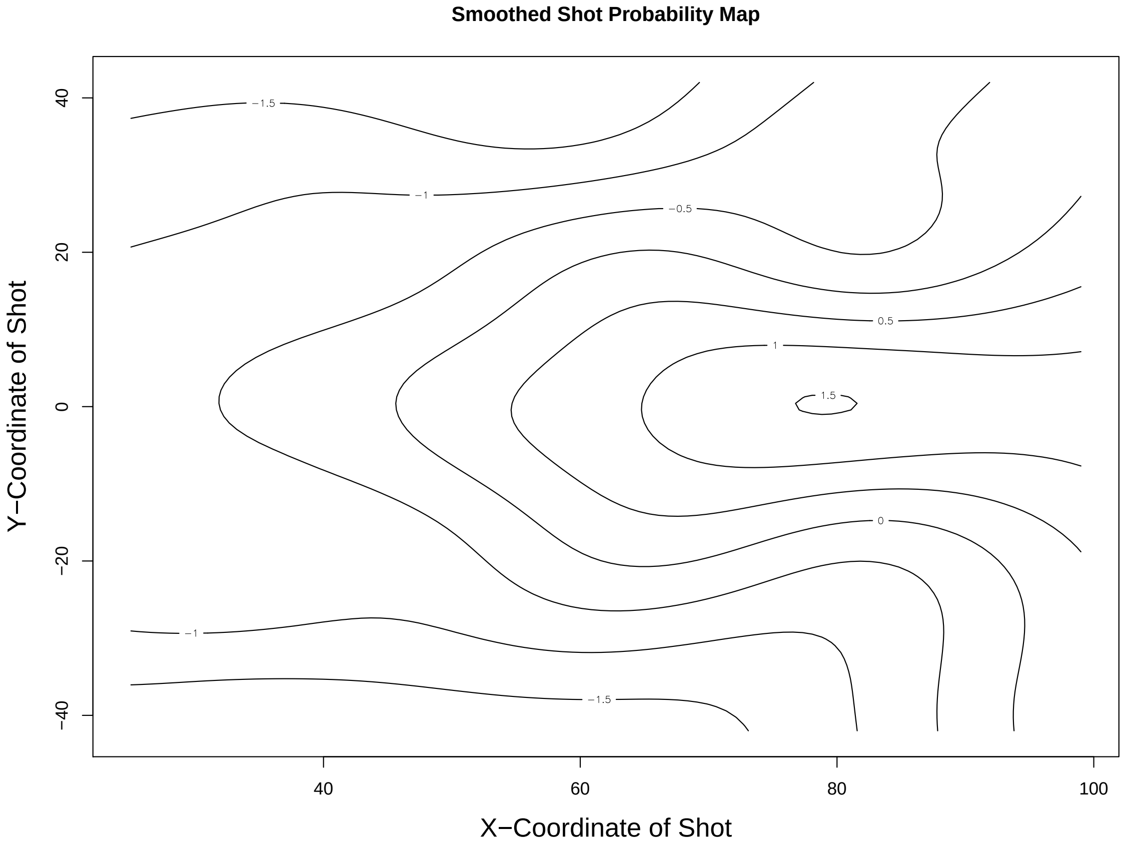


Figure 2.1: Probability of scoring from different locations where (89, 0) denotes the centre of the net. The values seen on the contour plots are not probabilities but are on the scale of the predictors (x-y coordinates).

In Figure 2.1 we can see that the probability of scoring a goal decreases the further from the net a shot occurs from, both in the north-south, and east-west directions and there is a high probability spot straight out from the net at about 10 feet, this is called the ”low slot” or the ”scoring area”. The probabilities range from approximately 0.2 right in front of the net, to about 0.01 when a shot occurs closer to the blue line and from the side boards. This model could be extended to include the type of shot (wrist, slap, backhand etc.) and whether the shot occurred as a rebound off the goalie.

In Figure 2.1 we generally observe that the probability of scoring decreases as the distance from the net increases and the angle from the net increases. Although we did not implement such constraints here, it would seem to be a good idea.

2.2 Regression Models

This analysis focuses on the current 2014-2015 NHL season. There were $n_p = 940$ players involved in $n_e = 571,562$ events. The data are arranged into a response vector Y and a design matrix X of indicator variables corresponding to individual player identities. For each event i the response vector contains

$$y_i = \pm z_i$$

where z is the probability of scoring from the location of the ith event, z is positive if the event is registered for the home team and z is negative if the event is registered for the away team. The corresponding i$^{\text{th}}$ row of X indicates the players on the ice when the event occurred, with X_{ij} equal to 1 for each home player on the ice and -1 for each away player on the ice. All other X_{ij} are equal to zero.

The design matrix X is extremely sparse: overall dimensions are $n_p * n_e = 940 * 571,562 = 53,7268,280$, but every row contains 930 zeros for over 98.7% sparsity. Although we have an increased number of events since we are using shots instead of just goals the sparsity is reduced but we will see in the Analysis section that it still negatively effects the results in OLS regression models. The first regression model we attempted is the standard OLS model and can be formulated as:

$$y_i = \alpha_0 + \beta' x_i + \epsilon_i$$

where x_i is the i$^{\text{th}}$ row of the design matrix $\mathbf{X}$, $x_{ij} = 1$ if player j was on the ice for the home team when a shot was taken i, $x_{ij} =$ -1 for away player j on the ice for shot i. Coefficient $\boldsymbol{\beta}'$ is the length-n_p vector of player effects, and α_0 is an intercept term which represents home team advantage, and ϵ is the Normal with constant variance error term. This model could be extended to incorporate additional covariates.

The second model we fit is similar to Gramacy et. al (2013) except we exclude the team effect. It is a gamma-lasso regression which utilizes an adaptive L1 penalization estimation and is formulated as:

$$y_i = \alpha_0 + \beta^{'} x_i + \epsilon_i.$$

Here all of the variables are defined the same as the OLS model above except the error term is more complex. Penalized regression is useful when the columns of $\mathbf{X}$ are highly correlated. Instead of maximizing the log-likelihood function, $\log(L(\theta \mid x))$, a penalized estimator is the solution to

$$\underset{\alpha, \beta \in \mathbb{R}}{\operatorname{argmin}} \left\{ \ell(\alpha, \beta + n\lambda \sum_{j=1}^{p} c(\beta_j) \right\}$$

where $\lambda > 0$ controls overall penalty magnitude and c() is the coefficient cost function. We use the lasso ℓ_1 cost function. The penalty size λ acts to suppress noise and focus on the true input signal. Since one doesn't know the optimal λ, in practical applications of penalized estimation a regularization path is required so as to obtain a field of $\hat{\beta}$ estimates while moving from low to high penalization. The gamma-lasso specification is based on the log penalty

$$c(\beta_j) = \gamma^{-1} log(1 + \gamma \mid \beta_j \mid)$$

where $\gamma > 0$ and $lim_{\gamma \to 0}\ c(b) = |b|$. Path behaviour is determined by γ and is referred to as the penalty scale. When $\gamma = 0$ the gamma-lasso is the usual lasso. We implement this model using the gamlr package in R, using the default settings for λ and γ (which gives us the usual lasso implementation). We set standardize to FALSE so to not standardize the coefficients to have a standard deviation of one. Setting this to true would be equivalent to scaling the coefficients by covariate standard deviations, which in our case has the drawback of favouring players who get little ice time. The code that we adapted can be found at https://github.com/TaddyLab/hockey

In the next section we will compare and discuss the results from these two models.

Chapter 3

Analysis

3.1 2014 - 2015 Season

In this section we will begin by looking at the results from the three models, attempting to determine if the results are sensible, where the models appear to fail, and whether or not the results match up with who we would expect to be the best and worst performing players in the league. Then we will look at how our shot quality statistic compares to the traditional plus-minus statistic. Finally, we will do a salary analysis to determine if teams are valuing players appropriately or not.

3.1.1 Discussion

The first model we attempted was

$$Y_i = \alpha_0 + \beta^{'} x_i + \epsilon_i$$

which is defined as the standard OLS regression model in the Model section of Chapter 2. This model gives some very peculiar results in that the vast majority of defencemen are in the bottom half of the estimates are nearly all forwards are in the top half of the estimates (when estimates are sorted in descending order). The reason for this is unclear since a player's position is not taken into consideration in the model. Furthermore, only 6% of the beta estimates are negative, while the remaining 94% are all positive. Despite having much more data at our disposal from using shots instead of goals it is possible that multicollinearity is still an issue since it is not uncommon for two players to be on the ice with each other for 70% of their even-strength ice-time.

The players with the largest and smallest estimates can be found in Table 3.1.

Top 5

Player	**Est**	**Pos**	**Tm**	**GP**	**PM**	**Corsi %**	**ED**
Colin Wilson	0.0776	C	NSH	77	19	55.5	155
Christian Thomas	0.0763	L	MTL	18	-2	56.5	16
Filip Forsberg	0.0751	L	NSH	82	15	57.2	230
Brendan Gallagher	0.0735	R	MTL	82	18	54	98
Matt Cullen	0.0716	C	NSH	62	8	53	91
		Bottom 5					
Cody Ceci	-0.0499	D	OTT	81	-4	49.4	-64
Jared Cowen	-0.0507	D	OTT	54	-11	48.7	-88
Eric Gryba	-0.0525	D	OTT	75	11	46.8	-133
Mark Borowiecki	-0.0533	D	OTT	63	15	46	-120
Chris Phillips	-0.0581	D	OTT	36	0	43.9	-114

Table 3.1: Top 5 and Bottom 5 Player Estimates From Model 1. *Est* is the OLS estimate, *Pos* is the player's position, *Tm* is team, *GP* is games played, *PM* is plus-minus and *ED* is event differential

.

There are a couple ways of checking our results to make sure they are reasonable. One, we would expect our results to be somewhat consistent with a player's Corsi percentage number. Second, we check a players overall event differential throughout the entire dataset, which is just the total of positive events (shot attempts for) subtract negative events (shot attempts against). These two numbers should correlate very strongly as Corsi% is simply a ratio of the event differential. In this model, our top player is Colin Wilson who has a Corsi% of 55.5 and an event differential of 155 for the 2014-2015 season. The second best player is Christian Thomas who has a Corsi% of of 56.5 and an event differential of 16. Rounding out the top three is Filip Forsberg who has a Corsi% of 57.2 and an event differential of 230. So these are pretty consistent with who we would expect to be near the top, although it is somewhat concerning that only players from Nashville and Montreal are in the top 5. Furthermore, the bottom 5 players are all defensemen from the Ottawa Senators, so it appears the model is having a problem dissociating between players on the same team.

Even though we have increased the overall number of events by using shots instead of just goals, it appears that OLS regression is still not performing adequately. It's likely that the sparsity of our matrix, recall every row has approximately 850 columns and only 10 of those columns have a 1 or -1 and the rest are 0's. Furthermore, OLS regression does not perform well with multicollinearity and this is still a big problem since many players have the same line mates throughout the season, many players are on the ice together 50-75% of the time, switching from only goals to shots will give us slightly more variety in lines but not a significant amount.

We also see from the residual versus fitted plot in Figure 3.1 that the constant variance assumption of the model is being violated. We notice that the plot appears to be broken up into two separate groups which is likely the result of all home team players being assigned a 1 and the away team players assigned a -1. For the lower grouping in the plot we see that the residuals increase as the probability of scoring increases, the same holds for the upper group except there the residuals are increasing as probability of scoring increases for the away team. All of this points towards the fact that OLS regression is not suitable for our type of dataset.

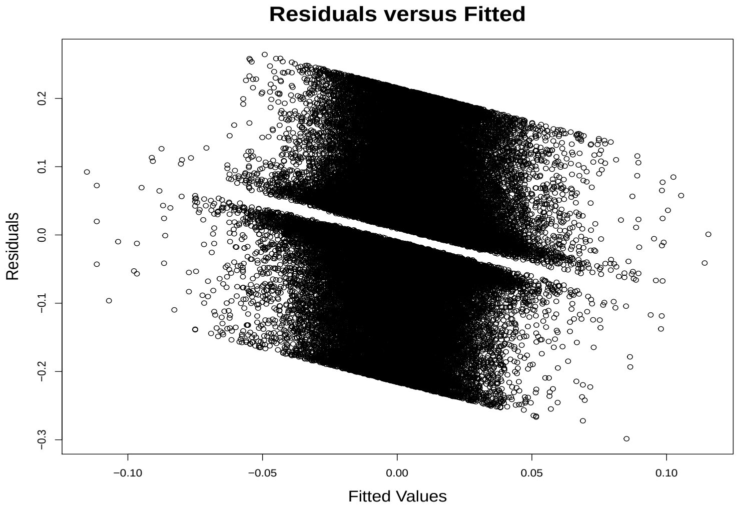


Figure 3.1: Residuals vs fitted in OLS regression

The second model we will look at is the gamma-lasso regression which utilizes an adaptive L1 penalization estimation. Similar to above we have

$$Y_i = \alpha_0 + \beta^{'} x_i + \epsilon_i$$

here we have just a regular intercept term and all the other variables are defined the same as the OLS models. This model gives much more sensible results when compared to common beliefs.

			Top 10				
Player	**Est**	**Pos**	**Tm**	**GP**	**PM**	**Corsi %**	**ED**
Patrice Bergeron	0.0174	C	BOS	81	2	58.9	204
David Pastrnak	0.0110	R	BOS	46	12	55.7	67
Colin Wilson	0.0103	C	NSH	77	19	55.5	155
Ryan Getzlaf	0.0095	C	ANA	77	15	52	140
Blake Wheeler	0.0094	R	WPG	79	26	54.5	130
Vladimir Tarasenko	0.0091	R	STL	77	27	55.4	189
Mathieu Perreault	0.0089	C	WPG	62	7	57.3	126
John Tavares	0.0086	C	NYI	82	5	55	200
Jordan Schroeder	0.0086	C	MIN	25	9	53.5	43
Dmitrij Jaskin	0.0086	R	STL	54	7	55.3	129
			Bottom 5				
Nicolas Deslauriers	-0.0122	L	BUF	82	-24	33.8	-339
Nathan Guenin	-0.0161	D	COL	76	-1	39.2	-290
Jared Boll	-0.0174	R	CBJ	72	-13	37.1	-158
Manny Malhotra	-0.0194	C	MTL	58	-6	34.6	-141
Adam Burish	-0.0234	R	SJS	20	-6	35.6	-79

Table 3.2: Top 10 and Bottom 5 Player Estimates From Gamma-Lasso Model *Est* is the gamma-lasso beta estimate, *Pos* is the player's position, *Tm* is team, *GP* is games played, *PM* is plus-minus and *ED* is event differential

Here we have Patrice Bergeron as our top player, with a beta estimate of 0.0174 (Corsi % 56, event differential 204). As noted above he won the award for the best defensive forward for this season (20142015). This is encouraging since we would expect our method to find the best two-way players, that is players who are one the ice for high probability shots on offense and allow low probability shots on defense. Second on the list is another Boston Bruin player David Pastrnak with an estimate of 0.011 (Corsi % 55.7, event differential 67). Note that Pastrnak only played in 46 of 82 games so it's not surprising his event differential is a little low for being second on the list. The 2014-2015 season was his rookie campaign after being drafted in the 1st round (25th overall) so it appears the Bruins have made a great selection.

The top 10 also includes many players that you would expect to be near the top and are generally considered as great NHL players, such as Ryan Getzlaf, Blake Wheeler, Vladimir Tarasenko and John Tavares. All of these players have great Corsi percentages and event differentials so this is promising and a great sign that this method is working the way we would expect it to. There is one player near the top that sticks out and that is Tyson Barrie of the Colorado Avalanche who is ranked 16th with an estimate of 0.00787 but has a Corsi % of 46 and an event differential of -160. He is only 1 of 38 players with a positive estimate and a negative event differential and of all these players he has by far the worst event differential. The next closest is Dominic Moore with a -89. So what is going on here? Well one thing we know is that multicollinearity is an issue with all of these models, the gamma-lasso regression method deals with this much better than OLS regression but if two players are always on the ice together we know that one of the players will get a positive or negative estimate and the other will get an estimate of zero. This means that we have to be very careful to check which players are one the ice with and how often. Barrie is a defensemen so the most important players to check are other defensemen since that is who he will play with most often.

We find that about 53.5% of the time Barrie is playing with Nate Guenin, 12.4% with Jan Hejda, and 10% with Nick Holden. Nate Guenin has the 5th worst estimate in the league at -0.0161 (Corsi % 39.2, event differential -290), he is no longer in the league after being picked up by the Anaheim Ducks in 2016-2017 free agency and subsequently cut and sent to their AHL minor league affiliate. Jan Hejda has an estimate of -0.00324 (Corsi % 43.4, event differential -204) and is also no longer in the league with 2014-2015 being his final year. Hejda signed a try-out contract with the Columbus Blue Jackets' AHL affiliate the following season but was cut after appearing in 11 games. Nick Holden has an estimate of -0.00376 (Corsi % 42.6, event differential -253). Barrie's linemates do not inspire confidence and in fact the Colorado avalanche only have 3 players with positive estimates so they are a pretty weak team overall. It appears that Barrie's teammates are to blame for the poor performance and Barrie may be one of the few bright spots for the Avalanche, and in fact he was the only Avalanche player to receive a vote for the 2014-2015 NHL first and second All-Star Teams (the Professional Hockey Writers Association votes for the best performers at each position). Even though Barrie seems to be better relative to his teammates I'm still skeptical that he should be the highest rated defensemen. It seems like he is getting credit for nearly all the good shots that are produced while he is on the ice. Meanwhile his teammates are taking the brunt of the blame.

There are a couple other players we can look into who are noticeably missing from the top of our rankings, Drew Doughty who has the best event differential of any player and

Sidney Crosby who is widely considered to be the best player of his generation. We'll start with Doughty who plays for the Los Angeles Kings.

He has the top event differential at 269, he was 2nd in votes for the best defensemen in the league, 17th in voting for the most valuable player in the league, won the award for the best defensemen in the 2008 international World Junior Tournament and is currently regarded as one of the best, if not the best, defensemen in the league. Despite all of this he has an estimate of 0 in this model. This is a major shortcoming of the gamma-lasso method. For the 2014-2015 season 50% of the estimates are 0, and as stated previously this is because players tend to have the same linemates throughout the season. This is no different for Doughty, who was on the ice with fellow defensemen Jake Muzzin 56% of the time, and not surprisingly Muzzin is highly rated by this model with the 58th best estimate at 0.00511. Muzzin also has a very high event differential of 258 so even when he is not on the ice with Doughty he is performing very well.

The Kings are a great overall team with only 3 of their players having a negative event differential, the lowest of which is Jordan Nolan at -32, so it is likely difficult for the algorithm to decipher who should be getting the credit, the Kings have 11 players with an estimate of 0 which is one of the most for any team. Furthermore, due to the fact that Doughty is considered one of the top defensemen in the league he has the highest average ice-time for any player on the Kings which means he his likely playing with the 3rd and 4th forward lines in order to shore up their weaknesses and because of this he will be on the ice for more of the best quality shots by the opposing team compared to his teammates. If the players on those lines do not change much then the algorithm will see Doughty as the only player that changes and therefore would be unduly blamed for the quality of shot the opposing team is getting. To that end, he is paired on defense with Robin Regehr 23% of the time and Regehr is one of the five Kings players that has a negative estimate.

Again we see that looking at a players estimate on it's own is not sufficient and that if we dig deeper we start to see why a player may not have the estimate we expect. It is also important to remind the reader that in this analysis we are only looking at times when the game was played at full even strength (5 vs. 5), so our rankings do not take into account how a player impacts other situational aspects of the game such as the power play and penalty kill (this method can be expanded to these aspects of the game and will be discussed in the Further Research section). This is important because Doughty is a mainstay on both the power play and penalty kill for the Los Angeles Kings and as we will see it is a major factor when analysing our next player.

Sidney Crosby is generally viewed as the best player in the league and one of the most decorated active players in the game, winning the Hart Trophy for being the leading scorer twice, the Lester B. Pearson/Ted Lindsay Award (Peer-voted best player) three times, the Hart Trophy (MVP of the league) twice, and the coveted Stanley Cup twice. However, the model only ranks Crosby as 39th overall. Again it's important to look at what the model is actually evaluating, which is the quality of the shots for and against during 5 on 5 play. Crosby is great goalscorer and an even better playmaker who sets his teammates up for easy goals. With this in mind it is significantly easier to score goals and make plays when there is more space available on the ice, which is what you find on power play opportunities. Therefore much of Crosby's value is attributable to how he performs on the power play. Looking at his basic statistics for 2014-2015 we see that his average power play time-on-ice is 3:36 per game and his average even strength time-on-ice is 15:59 per game, so he spends about 22% of the time on the power play compared to even strength. However, he scored 10/28 (35.7%) of his goals and accumulated 21/56 (37.5%) of his assists on the power play. He has a pretty mediocre plus-minus of 5, which puts him in the top 75% of all players in the league but is not very good for someone of his stature. I think we can conclude that Crosby is definitely one of the top players but teams do a fairly good job at limiting him during even strength and where he really excels is in the open space found on the power play.

3.1.2 Comparison to Plus-Minus

It appears that this method is a much better metric than just the basic plus-minus for ranking players at even strength, which is what we set out to accomplish in the first place. The players at the top of our list (Table 3.2) have widely varying plus-minus statistics, some are mediocre, some are the highest in the league and some are negative, which illustrates that plus-minus does not properly reflect the quality of scoring chances that individual players are generating. Our top ranked player Patrice Bergeron only has a plus-minus of 2 but it appears he his producing much higher quality shots than he is allowing. Our statistic is in agreement on Vladimir Tarasenko and Blake Wheeler, who are near the top in both. Max Pacioretty who had the highest plus-minus in the league at 38 is ranked 104th by us, his plus-minus is likely inflated from having the top goaltender in the league on his team (Carey Price who won the league MVP award for this season). In Figure 3.2 we see a few players

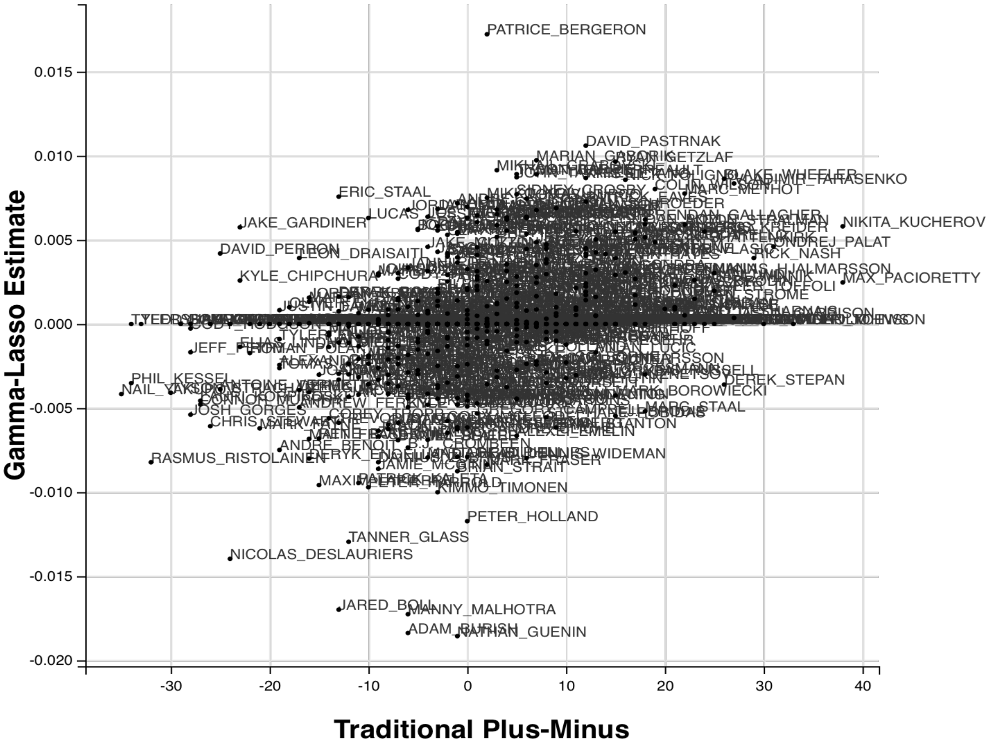


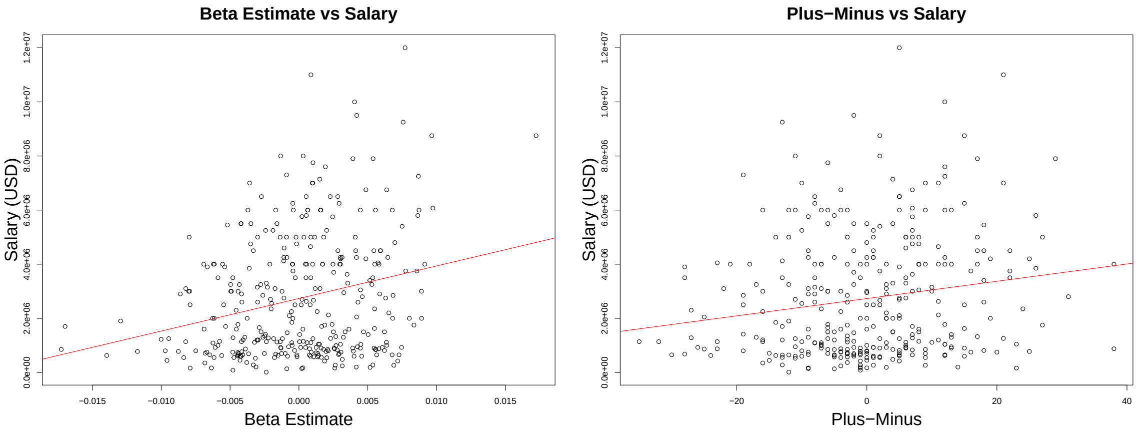
Figure 3.2: Plus-Minus vs Beta estimate

that stand out, especially those in the upper left who have a very negative plus-minus but a positive beta estimate. It appears that Eric Stall, Jake Gardiner, David Perron, and Kyle Chipchura were standout players on very poor teams. Near the bottom we have the poorest performing players from our model but who have only a slightly negative plus-minus. They include Nathan Guenin, Manny Malhotra, and Adam Burish. Illustrating the potential predictive capabilities of this model, none of these players were in the league following the 2014-2015 season.

3.1.3 Salary Analysis

In Figure 3.3, the left figure is a plot of 2014-2015 player salaries versus non-zero $\hat{\beta}$ estimates, the right figure is a plot of salaries versus plus-minus.

The line overlaid in each plot is an ordinary least squares fit, it's fairly clear that the gamma-lasso estimates have a stronger linear relationship with salaries than the traditional plus-minus statistic does. The points in the plus-minus plot are much more scattered and



(a) Salary vs Beta estimate

(b) Salary vs PM

Figure 3.3: The plot on the left shows non-zero Beta estimates from the gamma-lasso model vs Salary. The plot on the right shows non-zero Beta estimates from the gamma-lasso model vs Plus-Minus, an OLS fit is overlaid in each plot

there is no clear trend, whereas there is a clearer trend going from the bottom left to the top right in the gamma-lasso plot. Correlation numbers back this up as well, the correlation between salary and the gamma-lasso estimates are 0.23 whereas it is only 0.14 between salary and plus-minus. I would expect the trend in the gamma-lasso plot to be even stronger if we included power play and penalty kill situations as I suspect this is where a lot of the higher paid players value really shows.

With that said teams could benefit from using our estimates as a guide to salary allocation as there are a number of salaries that are clearly unwarranted. For instance the two highest paid players in the plot, Sidney Crosby ($12 million) and Zach Parise ($11 million) appear to be overpaid when compared to our top estimated player Patrice Bergeron ($8.75 million), although all three are quite far from the fitted line. Two of our lowest estimated players Manny Malholtra ($850 000) and Jared Boll($1.7 million) are at least being paid appropriately for their poor performance as they are close to the fitted line.

Chapter 4

Conclusions and Future Research

We find that despite using shots instead of just goals for our events, hence increasing the number of events we have in our dataset by roughly 10x, OLS regression techniques are still inadequate for this type of problem. All of the extra data is still not enough to overcome the issues caused by sparsity and multicollinearity due to players being on the ice with the same linemates for a large percentage of their events. The gamma-lasso regularized regression technique described by Gramacy et. al (2013) works well at providing partial player-effects and produces generally what we find to be sensible results. The drawback with this method is that a large percentage (approximately 50%) of the player estimates are zero since this model is also unable to differentiate between players that are on the ice with each other for a high percentage of their events. However, with some in-depth analysis it is straightforward to find out which players this is effecting the most. Running the model over multiple seasons so as to further increase the variety of player linemates reduces this problem, but at the expense of being able to evaluate how a player has performed over a single season or less. Overall we think that the decisions of general managers, coaches and fantasy players could benefit from our methods. At the very least, our approach identifies unusual cases, and in such cases it may be worthwhile for managers to investigate the situation.

With regards to future research it could be beneficial to include power play and penalty kill situations. I would be hesitant to include this all in the same model since there are only a small subset of players that are used in these situations and the number of shots for or against are very unbalanced in these situations and hence these players estimates will disproportionality benefit/suffer compared to players not used. Players used for the penalty kill players will have mostly negative events since they spend the vast majority of the time in the defensive zone. Meanwhile players who play on the power play will have the benefit of playing almost solely in the offensive zone. Creating separate models for each situation would allow for equal comparison over similar situations. Similar to the NBA, the NHL now has a camera tracking system in place at each of it's arenas and the wealth of data that is gathered would likely improve our estimates since we would be able to incorporate variables such as shot speed, whether the goalie was being screened (which would increase the probability that shots from further away would be goals), and how far a goaltender had to move across his crease to make a save attempt (it is much more difficult for a goalie to make a save if he has to move a large distance across his crease). One other idea with regards to players that are on the ice with a teammate often should be noted. If one is interested in a specific player such as we saw with Drew Doughty, you could look at who he is paired with most often and remove those players before running the model so that you would get a clearer picture of the player of interest.

Finally, the regularization method has various tuning parameters that can be adjusted. We did not focus on this and it is possible that additional tuning could lead to more realistic results. For example, we set the γ parameter to zero, which is the usual lasso, but this parameter can be set higher. This would create a larger penalty and would likely lead to less non-zero estimates, the players with non-zero estimates would be the exceptionally good and bad players.

Bibliography

[1] Gramacy, R.B., Jensen, S.T. and Taddy, M. (2013) “Estimating Player Contribution in Hockey with Regularized Logistic Regression,” *Journal of Quantitative Analysis in Sports*, 9 , 97-111.

[2] Macdonald, B. (2012) “Adjusted Plus-Minus for NHL Players using Ridge Regression with Goals, Shots, Fenwick, and Corsi,” *Journal of Quantitative Analysis in Sports*, 8(3), pp-. doi:10.1515/1559-0410.1447.

[3] Mason, D.S., and Foster, W.M. (2007) “Putting Moneyball on Ice?,” *International Journal of Sport Finance*, 2, 206-213.

[4] Schuckers, M.E., Lock, D.F., Wells, C., Knickerbocker, C.J., and Lock, R.H. (2011) “National Hockey League Skater Ratings Based upon All On-Ice Events: An Adjusted Minus/Plus Probability (AMPP) Approach,” *Unpublished.*

[5] Stair, A., Neral, J., Thomas, L., and Mizak, D. (2011) “Team Performance Characteristics which Influence Wins in the National Hockey League,” *Journal of International Business Disciplines*, 6, 47-56.

[6] Thomas, A.C., Ventura, S.L., Jensen, S.T., and Ma, Stephen (2013) “Competing process hazard function models for player ratings in ice hockey,” *The Annals of Applied Statistics*, 7(3), 1497-1524.