Lighting the Lamp: Calculating Individual Contribution in Hockey

Senior Integrative Exercise

Daniel Clipper, Carleton College '21

November 9, 2020

How do hockey teams evaluate a player's overall contribution to their team? It's not quite as easy as other sports, like baseball, which is tailor-made for data analytics. Hockey is a fast and fluid game with a lot of moving parts, little scoring, and almost no player tracking data available. As a result, teams often face the difficult challenge of evaluating players with traditional metrics alone. One statistic that is commonly used to judge a player's overall contribution is plus-minus, which measures the difference between the total number of goals scored and the total number of goals conceded by a player's team while that player was on the ice. The problem with this metric, however, is that it only gives teams a marginal effect; it doesn't take into account the influence of all the other players on the ice. So it doesn't really tell us a whole lot about a player, other than that he or she happened to be on the ice when a goal was scored. Rather than simply using plus-minus, a better way to evaluate individual contributions would be to jointly model the effects that every player has on the scoreline, along with confounding information within various game scenarios, team quality, coaching, etc. This method will instead give us a partial effect for

individual player contribution and a better understanding of how one player impacts the scoreline regardless of who else is on the ice.

This paper will detail an algorithm for calculating these partial effects. The first step to doing this would be using a logistic regression model to predict, for every goal, which team has scored given which players were on the ice. Ordinarly, this method would require standard estimation techniques (ie. maximum likelihood estimation, fisher scores, etc.). However, this won't work as the model is too high dimensional. A way to solve this would be through regularization. Regularization would apply a penalty to our model and makes it easier to handle large and imbalanced data.

Regularized logistic regression will provide a better way to evaluate and compare player contributions. After discussing the problems teams currently face when evaluating players using traditional metrics, this paper will detail the regression model and regularization techniques used to obtain these partial effects. The paper will also analyze 11 seasons worth of play-by-play game data from the National Hockey League (NHL). First, a model will be fit based on goals, then a second model will be fit based on shots, and a comparison between the two will be explored, resulting in partial versions of traditional metrics that control for other players and confounding variables, so as to offer a deeper understanding of overall player quality.

The code and empirical work for this paper uses open source libraries for R [7] and the 'gamlr' package [2].

Introduction

Hockey is an exciting, fast action sport. But a challenge for viewers, and a possible reason why hockey isn't as popular as other sports in the United States, is that it's so fast paced that it can be difficult to follow along. Statisticians face the same challenge when it comes to evaluating players. The speed and unpredictability of player and puck movement has resulted in the NHL struggling to roll out a viable player and puck tracking system, something every other major American sport has done successfully, as well as soccer (sorry, *ahem*, football) overseas in Europe's top leagues. There was even some serious talk about a tracking system being rolled out for the 2020 Stanley Cup Playoffs, but it didn't materialize, and has been pushed to the start of the 2021 season (but we'll see how that pans out). What has taken the NHL so long? If every other sport can do it, why can't the NHL? The explanation is fairly simple: there's too much going on in this game. One thing that separates hockey from similar sports is the frequency of substitutions, or line-changes as they call it, within games. Hockey teams typically have four offensive lines, and three defensive lines, which change as frequently as every 30-60 seconds throughout games. This inevitably results in an enormous number of different combinations of players, between both teams, that are on the ice during a game and over the course of a season. Not to mention how fast the line-changes happen; teams are given between 5 and 8 seconds to make these changes, and play doesn't stop between them. Teams often even dump the puck deep into the offensive zone in order to make line changes. These factors, and the general fast pace of the game, make tracking this information extremely difficult, more so than other sports.

So what have hockey teams done to address the lack of metrics other sports derive from tracking data? They instead have to evaluate players using traditional metrics. One traditional metric that is used to judge a player's overall contribution is plus-minus (PM). This is the

difference in goals scored and goals conceded by a player's team while that player was one the ice. For example, if New York Islander's Mat Barzal (my favorite player) was on the ice in a game for four Islander's goals and two goals from the opposing team, his PM for that game would be +2.

What does this tell us about Barzal's performance? Statistically, not a whole lot. That number by itself doesn't give any information about what impact Barzal had on that scoreline. This is because PM is a *marginal effect*, which can be defined as "the average change in some response (goals for-vs-against) with change in some covariate (a player being on the ice) *without accounting for whatever else changes at the same time*" [1]. In other words, a player's PM only reflects a player's individual contributions given the impact of the other players on the ice. If the Islanders put Lebron James on Barzal's line, surely Barzal's PM would be worse. This would be because Lebron doesn't play hockey; he plays basketball. The Islanders will likely concede more goals with Lebron in the team, which obviously wouldn't be fair to Barzal. However, it also goes the opposite way: Lebron's PM would be much better than it should because he's out on the ice with a superstar, and, well, players who actually play hockey. So what is missing? PM isn't taking into account the ability of a player's teammates or the quality of their opponents, nor is it taking into account playing time, coaching, goalies, game situations, and salaries – all things that have a considerable impact on the scoreline.

Player ability is what should be the main interest when evaluating...a player's ability. PM alone doesn't come close to doing that. However, what can get us there is a *partial effect*, which we can define as the "change in the expected response that can be accounted for by change in your variable of interest *after removing the change due to other influential variables*" [1]. With a partial effect, it wouldn't matter that Mat Barzal was playing with LeBron James; their partial

effects wouldn't change. This information can provide a better way for teams to evaluate not only how their own players perform, but also players from other teams as potential trade or signing targets. If another team were interested in signing Mat Barzal, a partial effect tells them what kind of quality player he is, despite him playing with LeBron James. Conversely, teams would know not to sign LeBron James to play hockey, or perhaps learn that he's not only an excellent hockey player, but truly the GOAT.

These partial effects can be calculated using *regression*: modeling a response (whatever variable is sought to be measured) as function of some other variables (the players). Hockey may not have tracking data just yet, but observational game data can be used to calculate these partial effects, and when dealing with a smaller number of variables, it's fairly easy. However, hockey is not dealing with a small number of variables, so standard regression simply won't work. Since hockey data is so high dimensional, using these standard techniques will result in major *over-fitting*, which can result in larger effects for players who get significantly less playing time. Players also tend to play on the same line, and line matchups tend to be consistent with opponents (meaning coaches tend to match lines throughout the game), so the data will likely have clusters of individual players. Therefore, methods that won't fail when confronted with these factors should be considered.

These issues are overcomable, specifically through a technique called *regularization*. By adding a penalty term to our model, regularization minimizes the effects of individual players when they don't contribute, and puts a microscope on the players that do. Regularization also allows our model to fall in line with the assumption that many players are replacement level, meaning their contributions are neutral enough that if they were to be replaced with an average player, the effect would not change. The penalty we ultimately choose is one that is going to lead

to our model performing the best it can out of sample. These techniques will allow us to approach a complex problem such as this as if it were a standard regression problem.

Methods

The aim now is to create a regression model that associates a player's presence on the ice with some sort of observable impact. For now we can focus on how such presence impacts goal production. This seems as the most logical first place to go, given that in order to win games, a team needs to score goals. We could think about using a simple linear model to estimate partial effects for goal production, however because scoring in hockey is so infrequent, the data will likely be too disaggregated to calculate a binary response, such as a goal being scored by either the home or away team.

A better way to estimate such binary outcomes would be with a simple logistic regression model. This would allow us to estimate the log odds of a goal being scored as a function of player presence. The values estimated can be converted from the log scale into an interpretable quantity, which provides a true partial effect for goal production. This is a basic framework that can be enhanced if need be.

Given *n* goals in the NHL, we can set y_i to $+1$ if the *home* team scores a goal, and -1 if the *away* team scores a goal. We can set $q_i = p(y_i = 1) = p(\textit{home team scores goal } i)$. With player presence as a response, we can set up a logistic regression model for the log odds the home team has scored a goal as

$$log\left[\frac{q_i}{1-q_i}\right] = \alpha + \beta_{hi_1} + ... + \beta_{hi_6} - \beta_{ai_1} - ... - \beta_{ai_6},$$

where the subscripts on the coefficients β represents the six players on the ice for both the home team *h* and away team *a*, and α represents team-specific intercepts, for which there are none (for now). Theoretically, adding on new covariates to this model would control for other situations, something that will be discussed a bit further in this paper, but this only controls simply for the presence of players when a goal is scored.

Unfortunately, fitting a model this simplistic won't do much good. Hockey data is just too high dimensional for a standard logistic regression approach to handle, and won't render us any relevant estimates of player effects. And while it can be tempting to fit such a simplistic model, we have to be careful to avoid *over-fitting*, where parameters are fit to statistical noise instead of the relationship we actually want to observe. Simple logistic regression also doesn't allow for control over the fact that players are generally grouped into lines, where they spend most of their ice time with the same group of players. Because of this, a design that is too simplistic will result in too much *multicollinearity*, where groups of covariates become correlated. This makes putting a microscope on individual players virtually impossible.

One solution to avoid these issues would be a method called *regularization*, where some penalty λ can be applied to the model to bias player coefficients towards zero.

Taking this approach, we can set $\eta_i = log[q_i/1 - q_i]$ as our linear equation. A *regularized* regression algorithm would aim to minimize the negative of some penalized object such as $\iota(\eta; y) + n\lambda \sum_{j=1}^{p} \beta_j{}^2$, with $\lambda > 0$, *n* total goals, and *p* total players. This type of regularization is called an L_2 penalty, or *ridge* regression. With a sufficiently large λ, $\beta_j{}^2$ will shrink asymptotically close to 0. However, since individual player effects is desired, it's not adequate

enough for many β_j to be 'almost' zero. We want them to be exactly zero so that statistical noise does not at all interfere with the estimates for *individual* players by themselves. The way we can get these coefficients to zero is by applying the L_1 penalty, or the LASSO. This method is preferred as it allows for shrinkage of players that don't have large effects. Estimation can be done by optimizing $\iota(\eta; y) + n\lambda \sum_{j=1}^{p} |\beta_j|$. It is by taking the absolute value of β_j as opposed to squaring it that is going to produce the minimization that is needed. This is consistent with the assumption that many players do not have a significant effect, and will this shrink towards zero. These LASSO estimates can take the form

$$\hat{\beta}_j^L = \begin{cases} y_j - \lambda/2 & \text{if } y_j > \lambda/2; \\ y_j + \lambda/2 & \text{if } y_j < -\lambda/2; \\ 0 & \text{if } |y_j| \leq \lambda/2. \end{cases} \qquad [3]$$

where the least squares coefficients shrink by a constant amount, $\lambda/2$, and those that are less than $\lambda/2$ shrink entirely to zero, Figures 1 and 2 [3] show visual interpretations of the difference between how L_2 and L_1 penalties minimize coefficient.

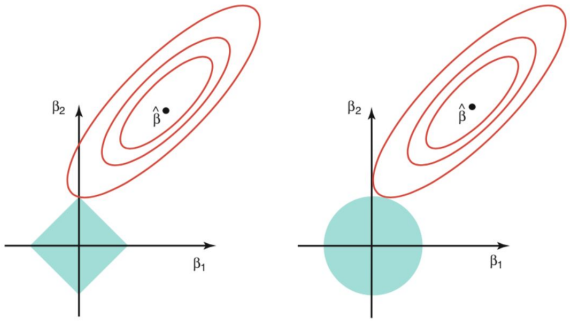

Figure 1: Ridge (right) vs. LASSO (left).

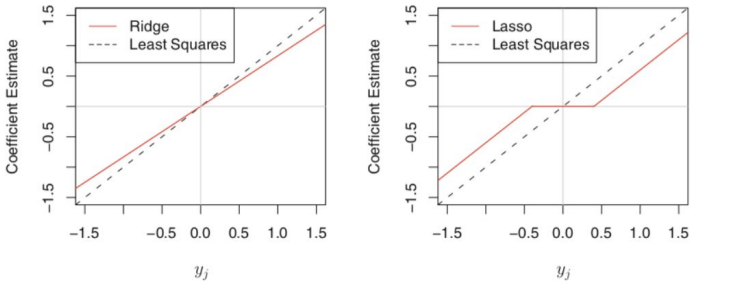

Figure 2: Ridge (left) shows coefficients minimized proportionally towards zero relative the the least squares estimates, while LASSO (right) shows coefficients are minimized to exactly zero give a certain threshold set by the least squares estimates.

Other variables can be added to control for other various effects and situations, however those would not be penalized under this model. It's only the player effects that need to shrink, and leaving the rest unpenalized would see to it that the player effects are not *polluted* by any confounding variables. Those specific additional effects in this model will be discussed in more detail later.

The size chosen for our penalty λ will cancel unwanted noise, increasing the focus on players individually. Large enough λ will shrink β_j to exactly zero. However, because we don't know λ, obtaining an estimation requires a *regularization path*, which we can define as "a $p \times T$ field of $\widehat{\beta}$ estimates while moving from high to low penalization along $\lambda^1 > \lambda^2 ... > \lambda^T$" [1].

A λ value can be chosen through a process called cross validation (CV). This is where a path of coefficients is fit over subsamples of our data repeatedly, and used as a predictor for the responses of the data that were left out of the subsample. Whichever λ value provides the least error is the one chosen. However, as outlined in [1] and [4], a computationally more efficient method would be a corrected Akaike Information Criteria (AICc), which is defined as

$$AICc = 2\sum_{i=1}^{n} \iota(\widehat{\eta}_\lambda; y) + \frac{2kn}{n-k-1}\ , \quad [4]$$

where $\widehat{\eta}_\lambda$ are the estimated log odds under penalty λ and k is the number of non-zero estimated coefficients at this penalty. We can assume $k \leq K$ total parameters given that most players are about replacement level, meaning their effects would hover around zero. This alternative to CV

will bear a model that performs just as well, and will also always yield the same result with the same data.

Application

This section will describe a model from [1] to calculate partial effects of NHL players. The data being used consists of 11 seasons worth of play-by-play data from the NHL between the 2002-03 and 2013-14 seasons. This data was obtained from nhl.com, and consists of significant events from each game, such as goals, shots, blocked shots, missed shots, etc. There are $p = 2439$ players and $n = 69449$ goals.

We can quantify individual performance for goal i in season s with the model

$$log\left[\frac{q_i}{1-q_i}\right] = \alpha + u_i'\gamma + v_i'\phi + x_i'\beta_0 + (x_i \circ s_i)'(\beta_s + p_i\beta_p), \quad [1]$$

where

- Vector u_i gives a team-season indicator (ie. New York Islanders 2013-14), which would be set to $u_{it} = +1$ if team-season t was the home team for goal i, $u_{it} = -1$ for the away team, and $u_{it} = 0$ if not on the ice (meaning the team was not involved in the game in which goal i occurred).
- Vector v_i gives a special-teams indicator, taking into account any scenario in which the game is not being played 5v5 (excluding the goalie), a situation that is typically the result of a penalty. This can also include a “pulled goalie” scenario, where one team has 6

skaters. We set $v_{ik} = +1$ if the home team was in special-team scenario k for goal i, $v_{ik} = -1$ for the away team, and $v_{ik} = 0$ if neither team was in a special-teams scenario. This is important because 35% of goals are scored in these scenarios.

- Vector x_i gives the player-presence indicator, where $x_{ij} = +1$ if player j was on the home team and on the ice for goal i, $x_{ij} = -1$ for the away team, and $x_{ij} = 0$ if not on the ice. We can use en element-wise product ($\circ$) to interact this vector with
    - Season vector s_i where $s_{it} = +1$ if goal i was scored in season t.
    - Postseason indicator p_i where $p_i = 1$ if goal i was scored in the Stanley Cup Playoffs, and $p_i = 0$ if scored during the regular season.

Our estimation comes from minimizing the penalized deviance which, again, only applies to the player-presence indicator, while the other parameters remain unpenalized. $\beta_0 + \beta_{sj}$ represents the player effects for season s, which estimates that log odds that, if a goal is scored, it was scored by player j's team. We can estimate these effects by calc a partial "for-%" with

$$PFP_{sj} = \frac{\beta_{sj}}{1-\beta_{sj}}$$

However this can be quite difficult to interpret as is, so it is necessary to transform this from the log-scale into a probability, where

$$PFP_{sj} = \left(1 + exp\left[-\beta_{0j} - \beta_{sj}\right]\right)^{-1} \quad [1]$$

This provides a partial version of another standard metric (also a marginal effect) for-% (FP), which is the total number of events by a player's team divided by the total number of events for both teams, while that player was on the ice. It's important that we translate PFP to a scale of probabilities, as we ultimately want our partial player effect, *partial plus-minus* (PPM), to live on the same scale and the traditional PM. This way, each can be interpreted the same way and thus easily comparable. So, if player j was on the ice for g_{sj} total goals during season s, we can define his PPM as

$$PPM_{sj} = g_{sj}PFP_{sj} - g_{sj}(1 - PFP_{sj}) = g_{sj}(1 - 2PFP_{sj}) \quad [1]$$

Table 1 ranks the top 20 and bottom 20 players ranked by PPM for the regular season. Table 2 does the same, but ranked by PM. You'll see players repeated, as these player effects were calculated per season for each player. For example, Sidney Crosby appears four times in the top ten, certifying him a bonafide star in the league during this time frame. The table only shows results from the regular season because there was no evidence that any player in the data set performed differently between the regular season and the playoffs. This means that the penalty minimized all $\widehat{\beta}_{pj}$, the playoff coefficients, to zero.

It is clear that there are some major differences between the PPM and PM rankings. For example, while Peter Forsberg's 2002-03 season with the Colorado Avalanche ranks first for PPM, Alex Ovechkin's 2009-10 season with the Washington Capitals ranks first for PM. That season for Ovechkin did not rank anywhere in the top 20 for PPM, and Forsberg's PPM was higher in 2002-03 than his PM. Also notably, while Sidney Crosby occupies many spots in the rankings for PPM, his presence in the PM rankings is limited, only appearing once in 18th place.

This suggests that PM is not doing star players like Crosby justice. Similarly with the bottom on the list, there are no duplicate player-seasons between the rankings for PPM and PM. Notably, Patrick Lalime and Jack Johnson both appear several times at the bottom for PPM, but nowhere for PM.

Table 1: Goal-Based Performance (by PPM)

Rank	Player	Team	Season	PPM	PM	PFP	FP	Beta
1	Peter Forsberg	COL	20022003	55.52	85	0.68	0.77	0.74
2	Sidney Crosby	PIT	20092010	43.47	60	0.60	0.64	0.41
3	Dominik Hasek	OTT	20052006	42.45	80	0.59	0.67	0.35
4	Sidney Crosby	PIT	20082009	42.26	48	0.60	0.61	0.41
5	Sidney Crosby	PIT	20052006	41.86	52	0.60	0.62	0.41
6	Peter Forsberg	PHI	20052006	40.67	61	0.68	0.77	0.74
7	Pavel Datsyuk	DET	20072008	39.49	87	0.60	0.72	0.40
8	Pavel Datsyuk	DET	20082009	39.49	69	0.60	0.67	0.40
9	Sidney Crosby	PIT	20062007	35.62	79	0.60	0.72	0.41
10	Mark Streit	NYI	20082009	35.08	24	0.59	0.56	0.35
11	Matt Moulson	NSH	20112012	34.92	37	0.60	0.61	0.41
12	Lubomir Visnovsky	ANA	20102011	34.52	70	0.58	0.66	0.31
13	Alex Ovechkin	WAS	20082009	34.46	80	0.57	0.66	0.28
14	Joe Thornton	SJS	20092010	33.91	52	0.60	0.65	0.39
15	Joe Thornton	SJS	20102011	33.91	48	0.60	0.64	0.39
16	Ondrej Palat	STL	20132014	32.75	37	0.64	0.66	0.56
17	Pavel Datsyuk	DET	20062007	32.61	70	0.60	0.71	0.40
18	Joe Thornton	BOS	20022003	32.17	47	0.60	0.64	0.39
19	Joe Thornton	SJS	20072008	32.17	69	0.60	0.71	0.39
20	Andrei Markov	MTL	20072008	31.90	47	0.57	0.60	0.28
10184	Martin Skoula	COL	20022003	-15.65	21	0.42	0.61	-0.33
10185	Patrick Lalime	BUF	20082009	-15.79	-15	0.43	0.44	-0.27
10186	Jack Johnson	LAK	20072008	-15.82	-34	0.45	0.39	-0.21
10187	Brett Clark	STL	20112012	-16.93	-47	0.44	0.35	-0.22
10188	Niclas Havelid	ATL	20082009	-16.97	-40	0.45	0.39	-0.19
10189	Jack Johnson	LAK	20102011	-17.21	9	0.45	0.53	-0.21
10190	Jack Johnson	FLA	20112012	-17.21	-1	0.45	0.50	-0.21
10191	Bryan Allen	FLA	20062007	-17.90	-17	0.45	0.45	-0.20
10192	Jack Johnson	LAK	20092010	-19.46	-4	0.45	0.49	-0.21
10193	Patrick Lalime	STL	20052006	-19.77	-29	0.43	0.40	-0.27
10194	Alexander Edler	TOR	20132014	-20.49	-35	0.37	0.27	-0.55
10195	Patrick Lalime	CHI	20072008	-22.29	-4	0.43	0.49	-0.27
10196	Tim Thomas	BOS	20092010	-24.22	-16	0.43	0.46	-0.26
10197	Andrej Meszaros	OTT	20062007	-27.32	-6	0.42	0.48	-0.33
10198	Bryce Salvador	NJD	20082009	-34.40	-31	0.35	0.37	-0.62
10199	Patrick Lalime	OTT	20022003	-37.81	47	0.43	0.58	-0.27
10200	Patrick Lalime	OTT	20032004	-37.81	37	0.43	0.56	-0.27
10201	Niclas Havelid	ATL	20062007	-62.64	-22	0.34	0.44	-0.67
10202	Niclas Havelid	ATL	20052006	-65.94	-41	0.33	0.40	-0.70
10203	Jay Bouwmeester	FLA	20052006	-69.62	-32	0.33	0.42	-0.69

Table 2: Goal-Based Performance (by PM)

Rank	Player	Team	Season	PPM	PM	PFP	FP	Beta
1	Alex Ovechkin	WAS	20092010	26.97	101	0.57	0.76	0.28
2	Jaromir Jagr	NYR	20052006	20.16	93	0.55	0.75	0.21
3	Dany Heatley	OTT	20052006	10.45	90	0.53	0.72	0.10
4	Pavel Datsyuk	DET	20072008	39.49	87	0.60	0.72	0.40
5	Daniel Sedin	VAN	20102011	19.45	86	0.56	0.75	0.22
6	Peter Forsberg	COL	20022003	55.52	85	0.68	0.77	0.74
7	Milan Hejduk	COL	20022003	23.94	85	0.57	0.76	0.30
8	Nicklas Backstrom	WAS	20092010	12.45	84	0.53	0.70	0.12
9	Alex Ovechkin	WAS	20072008	30.64	83	0.57	0.69	0.28
10	Mike Green	WAS	20082009	6.34	83	0.52	0.70	0.06
11	Dominik Hasek	OTT	20052006	42.45	80	0.59	0.67	0.35
12	Alex Ovechkin	WAS	20082009	34.46	80	0.57	0.66	0.28
13	Henrik Sedin	VAN	20102011	26.10	80	0.57	0.72	0.29
14	Thomas Vanek	BUF	20062007	25.45	80	0.58	0.76	0.33
15	Teemu Selanne	ANA	20052006	24.45	80	0.58	0.75	0.31
16	Mike Green	WAS	20092010	7.42	80	0.52	0.67	0.06
17	Jason Spezza	OTT	20052006	5.71	80	0.52	0.77	0.08
18	Sidney Crosby	PIT	20062007	35.62	79	0.60	0.72	0.41
19	Alexander Semin	WAS	20092010	30.12	78	0.59	0.73	0.36
20	Daniel Alfredsson	OTT	20052006	23.43	78	0.56	0.70	0.24
10184	Josef Melichar	PIT	20032004	0.00	-50	0.50	0.32	0.00
10185	Brian Holzinger	PIT	20032004	0.00	-50	0.50	0.27	0.00
10186	Radoslav Suchy	CBJ	20052006	0.00	-50	0.50	0.31	0.00
10187	Brad Stuart	BOS	20062007	0.00	-50	0.50	0.37	0.00
10188	Zbynek Michale	PHX	20082009	0.00	-50	0.50	0.37	0.00
10189	Joel Kwiatkows	WAS	20032004	-0.55	-50	0.50	0.32	-0.01
10190	Richard Park	NYI	20092010	0.00	-51	0.50	0.32	0.00
10191	Bryce Salvador	STL	20052006	-0.73	-51	0.50	0.22	-0.02
10192	Todd Marchant	ANA	20102011	-1.69	-52	0.49	0.16	-0.04
10193	Brendan Witt	WAS	20032004	-9.29	-52	0.46	0.27	-0.16
10194	Nate Thompson	NYI	20092010	-3.17	-53	0.48	0.19	-0.07
10195	Jay McClement	STL	20102011	-11.47	-53	0.45	0.29	-0.18
10196	Todd Marchant	ANA	20092010	0.00	-55	0.50	0.26	0.00
10197	Johnny Oduya	ATL	20102011	0.00	-55	0.50	0.31	0.00
10198	Brian Elliot	OTT	20102011	-14.52	-57	0.47	0.40	-0.11
10199	Nikolai Khabibu	EDM	20102011	0.00	-58	0.50	0.38	0.00
10200	Scott Hannan	COL	20082009	-9.73	-58	0.47	0.31	-0.13
10201	Chris Phillips	OTT	20102011	-6.83	-60	0.48	0.30	-0.09
10202	Derian Hatcher	PHI	20062007	0.00	-61	0.50	0.33	0.00
10203	Brendan Witt	NYI	20082009	-11.73	-64	0.45	0.25	-0.19

This indicates that the traditional PM metric is awarding too much credit to these players, and likely many other players. It's also notable that many of the β values, particularly at the bottom, for players in the PM rankings are quite small, and often zero. This indicates the PM is rewarding or punishing players who don't get a lot of ice time, whereas PPM controls for that.

But what if we didn't want to only look at goal-based performance? In soccer, it's common to evaluate players based on how often they assist in creating "chances", essentially shots. By borrowing that idea from soccer, it is possible to calculate the same partial effects for shots using the same model. A commonly used metric for shots that can mimic "chance creation" is called Corsi, which is the aggregate of goals, shots on goal, blocked shots, and missed shots. Not only does this fill in nicely as a way of measuring chance creation, but Corsi could be an (imperfect) substitute possession statistics, which hockey doesn't have due to the lack of tracking data. The logic with that is that it's likely that the team that is taking more shots is holding onto the puck more.

The only change to the regression model using Corsi is that the response variable would be the log odds that, if a Corsi event happens, it was by the home team. The benefit of calculating this effect is that it gives us a much larger sample size to work with. The data used for this consists of $n_c = 1,329,679$ Corsi events, much larger than the sample of goals in the data set.

We can also represent Corsi with FP, and use that to calculate a Corsi-PFP (PFPc) and a Corsi-PPM (PPMc). Similarly to FP and PM, a Corsi-FP (FPc) and Corsi-PM (PMc) are marginal effects as well, and the problems with using them to evaluate a player's ability to create chances are the same as described for the same goals-based metrics. Table 3 ranks the top and bottom 20 players for PPMc for the regular season.

This is clearly a very different list than the rankings from Table 1. This time, it's Daniel Sedin who stands out. Sedin first appears in the rankings for goal-based PPM at 152nd. This tells us that Daniel Sedin was great at creating shots or opportunities for shots, but did not have as much of an effect on those shots becoming goals. It also tells us that it's possible that the players that occupy the top of Table 1, but not Table 3, are creating goals more efficiently (more goals with less shots), and that the quality of the opportunities they create is better than those players at the top of Table 3.

Table 3: Shot-Based Performance (by PPMc)

Rank	Player	Team	Season	PPMc	PMc	PFPc	FPc	Beta
1	Daniel Sedin	VAN	20102011	615.14	876	0.60	0.65	0.42
2	Eric Staal	CAR	20082009	605.41	619	0.58	0.59	0.34
3	Mikhail Grabovski	TOR	20102011	597.05	465	0.60	0.57	0.39
4	Joe Thornton	SJS	20112012	596.37	742	0.59	0.61	0.35
5	Alex Ovechkin	WAS	20092010	575.72	1047	0.59	0.66	0.35
6	Daniel Sedin	VAN	20072008	562.11	685	0.60	0.63	0.42
7	Daniel Sedin	VAN	20082009	547.83	680	0.60	0.62	0.40
8	Ryan Kesler	VAN	20102011	530.05	649	0.58	0.59	0.31
9	Sidney Crosby	PIT	20092010	517.86	815	0.57	0.62	0.30
10	Daniel Sedin	VAN	20112012	510.16	880	0.60	0.67	0.40
11	Henrik Zetterberg	DET	20112012	497.04	596	0.58	0.60	0.33
12	Claude Giroux	PHI	20102011	487.53	347	0.58	0.56	0.32
13	Zach Parise	NJD	20082009	486.45	843	0.58	0.64	0.33
14	Joe Thornton	SJS	20102011	482.72	647	0.58	0.60	0.31
15	Alex Steen	STL	20102011	475.50	561	0.59	0.61	0.38
16	Lubomir Visnovsky	ANA	20102011	474.91	446	0.56	0.56	0.26
17	Eric Staal	CAR	20102011	473.92	415	0.56	0.56	0.26
18	Justin Williams	LAK	20112012	471.53	717	0.59	0.63	0.34
19	Alex Ovechkin	WAS	20072008	463.14	1094	0.56	0.65	0.26
20	Patrik Elias	NDJ	20102011	461.75	461	0.60	0.60	0.39
10605	Mike Commodore	CBJ	20082009	-447.91	-537	0.43	0.42	-0.28
10606	Scott Hannan	CGY	20112012	-451.04	-591	0.42	0.40	-0.32
10607	Chris Phillips	OTT	20072008	-454.09	-644	0.43	0.40	-0.29
10608	Jay Bouwmeester	FLA	20052006	-457.01	-305	0.44	0.46	-0.23
10609	Karlis Skrastins	COL	20082009	-457.54	-754	0.43	0.38	-0.29
10610	Karlis Skrastins	DAL	20092010	-464.49	-655	0.42	0.39	-0.32
10611	Mattias Öhlund	VAN	20082009	-465.36	-212	0.42	0.47	-0.31
10612	Mattias Öhlund	VAN	20062007	-470.03	-147	0.43	0.48	-0.29
10613	Scott Hannan	COL	20082009	-478.83	-788	0.43	0.38	-0.30
10614	Douglas Murray	SJS	20092010	-486.16	-184	0.42	0.47	-0.32
10615	Scott Hannan	COL	20072008	-507.70	-504	0.42	0.42	-0.34
10616	Filip Kuba	OTT	20112012	-509.74	-77	0.42	0.49	-0.30
10617	Niclas Havelid	ATL	20072008	-516.86	-883	0.41	0.35	-0.36
10618	Johnny Oduya	NJD	20082009	-522.40	51	0.42	0.51	-0.34
10619	Douglas Murray	SJS	20102011	-540.83	-117	0.40	0.48	-0.40
10620	Dion Phaneuf	CGY	20062007	-552.69	-42	0.42	0.49	-0.32
10621	Niclas Havelid	ATL	20082009	-562.65	-604	0.40	0.40	-0.39
10622	Sergei Gonchar	PIT	20062007	-586.55	174	0.42	0.52	-0.31
10623	Paul Martin	NJD	20082009	-695.83	283	0.39	0.55	-0.45
10624	Bryce Salvador	NJD	20082009	-912.17	-407	0.32	0.42	-0.74

Conclusion

This paper has outlined a framework for applying regularization techniques, particularly the LASSO, to logistic regression to calculate individual player contributions in hockey as a partial effect. These partial effects provide us a better understanding of how players perform as individuals than traditional hockey metrics, and can be compared to such traditional metrics as a point of understanding whether or not players are better or worse than they may seem. The model detailed in this paper allows for the controlling of different aspects of the game (ie. special-teams scenarios, coaching, seasons, etc.).

While this analysis does not provide an all-encompassing statistic to understand the overall quality of players, it does lay a solid foundation for more accurately evaluating players. The methods provided in this paper pave a better way to identify impactful players on a team's own roster, as well as identifying potential acquisition targets from other teams that may be impactful to a team seeking players to enhance their roster.

This paper has laid out comparisons between goal-based and shot-based partial effects. While it is not definitively clear that one metric is better than another, what matters at the end of the day is which team scored the most goals. Because goals are an indicator of wins, the goal-based metrics described might be more telling, at least on the surface, than the shot-based metrics. Again, it doesn't matter who shoots the most, it matters who scores the most.

A limitation I encountered involved Corsi-based effects for the playoffs. While goal-based effects saw no difference between the regular season and playoffs, many playoff coefficients $\widehat{\beta}_{pj}$ for Corsi-based effects were non-zero values, indicating that there was a difference between how many players performed in the regular season and the playoffs. However these values were quite difficult to interpret as many of them ended up actually being the same as

the values for the regular season. So while values were calculated, they did not provide significant information on how "clutch" players are in the playoffs versus the regular season. The idea of a player being "clutch" is hard to define, but it would certainly be interesting to evaluate this in future research.

Another point of further research to consider on shot metrics could be looking at how teams, rather than individual players, generally perform in relation to their player's PPMc. It's difficult to know from the information described in this analysis if the teams that create the most goal-scoring chances end up fairing better over a long regular season. Another direction this research can take is with a stronger focus on defensive performance. While the analysis in this paper certainly provides a base-understanding of when players are defensive liabilities, it doesn't paint the entire picture about how a player directly impacts *only* how many goals they save their own team from conceding.

Future research could also be done on determining individual player contribution through some economic concepts of Game Theory, more specifically Shapley Values discussed in [8]. This is a technique used to fairly distribute payoffs to a set of players in some coalitional game, and determine which players in game contribute the most and the least to goal scoring and conceding, but with a more team-based approach than the methods outlined in this paper. Shapley Values may also help in exploring positional differences in production.

One last important topic of future research would be studying player chemistry. Knowing how certain interactions between players or groups of players impact the scoreline would be incredibly useful for determining player lines. This is a topic that has been studied by soccer analysts in [5], which may be applicable similarly to hockey. What is nice about this soccer research is that it doesn't require player tracking data (though it may in future), something the

NHL is currently lacking. Partial effects similar to those outlined in this paper can instead be calculated on player lines, and provide a better understanding of which combinations of players contribute the most (and fewest) goals.

I believe there is much more work to do in the field of hockey analytics, and hockey is well behind other sports at the moment. The lack of player and puck tacking severely limits the depth of research that can be done. However, I am optimistic that tracking data will soon be more polished and widely available, and teams will continue to make strides in evaluating players to build winning teams and in-game strategies. It sometimes only takes simple models (like this paper outlines) to handle complex data; models that are easy to implement and interpret, which can be reproducible for research in other areas of hockey that are yet to be explored. The future of analytics is promising, and hockey is just getting started.

References

[1] Gramacy, R. B., Taddy, M., & Tian, S. (n.d.). Hockey Player Performance via Regularized Logistic Regression. *Handbook of Statistical Methods and Analyses in Sports*.

[2] Matt Taddy. gamlr: *Gamma Lasso Regression*, 2013. R package version 1.11-2.

[3] James, G., Witten, D., Hastie, T., & Tibshirani, R. (n.d.). *An Introduction to Statistical Learning with Applications in R*. Springer.

[4] Clifford M. Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 1989.

[5] Bransen, L., & Van Haaren, J. (2020). Player Chemistry: Striving for a Perfectly Balanced Soccer Team. *SciSports*.

[6] Matt Taddy. One-step estimator paths for concave regularization. arXiv:1308.5623, 2015.

[7] R Development Core Team. R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[8] Fletcher-Hill, P. (2014). Computing Shapley Values in the English Premier League.