Determining an NHL Center's Value: Salary Prediction Based on Performance Data

By AUSTIN TONACK

Using performance and salary data from the 2017-2018 NHL season, I build a multiple linear regression to predict salaries for NHL centers based on their on-ice performance. The predictive model looks to answer the question of an NHL center's value. Given the results, I find that average time on ice per game has a high positive correlation with salary, and I attempt to remedy the possibility of reverse causality by providing theoretical justifications. I also find that the overall model provides reasonable salary predictions, which can be utilized as a comparison tool to identify inefficiencies in salary allocation with the most underpaid or overpaid centers.

I. Introduction

In employment, an individual's value to a firm is generally defined by their salary. That is no different for hockey players in the National Hockey League. However, NHL teams are constrained by a salary cap, and must sign a certain number of players with a pre-determined amount of money. The imposition of a salary cap creates two competing sources of pressure. Much like any other firm, it is in the best interest of NHL team management to allocate their limited resources efficiently in order to build the most effective roster of players possible. This in turn creates an optimal on-ice product that helps generate more revenue for the team and raises the franchise valuation. On the other side, the players want to receive the financial compensation they deserve for their high level of skills and services, especially since a career in the NHL does not last long. I focus on centers

specifically because their role is unique and arguably the most important on a hockey team. Widely considered the core of a team, it is virtually impossible to have long term success in the NHL without a strong group of centers. To illustrate the motivation behind this question, and the importance of centers, I will refer to the 2017-2018 Stanley Cup Champion Washington Capitals. The Washington Capitals have been a dominant team for the better part of a decade and won the Stanley Cup this past season. Their success can be attributed to their solid group of centers and overall efficient use of salary dollars to build the best team. As a result of their on-ice success, the franchise valuation increased $100 million from $625 million to $725 million after a single season ("The Business of Hockey", 2018).

Given these motivating factors, my paper looks to answer the question of an NHL center's value based on their on-ice performance. Using a combination of player performance and salary data sets for the 2017-2018 NHL season provided by CKM Sports Management, I utilize a multiple linear regression as a salary prediction model consisting of a set of 17 independent variables, each representing different performance metrics I deem important for a center. The results of the model will assign a predicted salary value to a center based on their performance during the 2017-2018 season, which can be compared to their actual salary to identify the most underpaid and overpaid centers.

The remainder of my paper is divided into six sections. The first section will provide a brief look at the NHL and the salary cap, as well as previous literature related to the topic of NHL player valuation. The second section outlines the data I utilized and the process I took to transform the data into a useful resource. The third section breaks down the empirical methodology I utilized to build the predictive model along with some theoretical justifications. The fourth section displays the regression output and interpretations of the results of the model. The fifth section discusses the robustness of the model, sources of limitations, and an extension of

the results on rookies under entry-level contracts. The sixth and final section is a conclusion that summarizes the findings of the paper and its implications.

II. Background

The National Hockey League is a hockey league that consists of 31 teams situated in 31 cities across Canada and the United States. Compared to other hockey leagues, the NHL is the largest in terms of both popularity and revenue. However, the NHL is not just a sports league, it is a billion-dollar entertainment industry. During the 2017-2018 season, the NHL generated $4.86 billion in league-wide revenue, and that is projected to steadily increase year after year. The league-wide revenue determines each team's salary cap, or the upper limit of the payroll range for their players. This means that teams must build and sign a roster of 20 players while remaining under the salary cap, which was $75 million for the 2017-2018 season. The NHL began enforcing a salary cap in 2005-2006 as a parity mechanism that prevents large discrepancies in financial resources amongst teams from affecting the competitiveness of the league. For example, the Toronto Maple Leafs have a franchise valuation of $1.45 billion, and the Arizona Coyotes have a franchise valuation of just $290 million ("The Business of Hockey", 2018). If there was no salary cap, the Toronto Maple Leafs could have utilized their deeper financial resources to sign better players by outbidding other franchises like the Arizona Coyotes, therefore building a far superior team. Because the salary cap exists, all 31 teams have the same financial ability to sign players, regardless of their actual financial resources. As a result, all 31 teams can remain competitive on the ice.

The topic of NHL player value is not a prominent area of discussion in academia, but there is still some previously published literature regarding NHL salary determination. Before the implementation of the salary cap in 2005-2006, it

was argued that “the most fundamental issue in sports literature is the extent to which competitive balance among the teams is affected by institutional arrangements” (Richardson, 2000). In the paper, Richardson discusses the importance of institutions within the NHL prior to the introduction of the salary cap. Although his paper primarily focuses on other institutions like free agency and the ability of a player to bargain a higher salary through skillset shortages, it highlights the role of institutions within the league to maintain competitiveness amongst all teams. It emphasizes the significance of an institution like the salary cap on player values, as the restriction on teams forces them to assign value to players sparingly. That being said, the focus of my paper is centered more around an individual player’s performance in value determination, rather than institutional roles. A result from a paper by Jones and Walsh suggests that my focus on player performance is important as “skills appear to be the prime determinant of player salaries in the NHL” (Jones & Walsh, 1988). Even though they came to that conclusion using data from the pre-salary cap era 1977-1978 season, the result still suggests that player performance is the key component in player value. However, a paper by Kahane argues that “players with the same level of skill may in fact have differing salaries if they play on different teams, due to team specific effects” (Kahane, 2001). The result of Kahane’s paper supports the idea that different franchises have fixed effects that impact player salaries. Much like the previous two papers mentioned, his data comes from the pre-salary cap era. My paper looks to assess the presence of these arguments in the modern NHL as I answer the question of a center’s value in the modern NHL based on their performance.

III. Data

Since the primary purpose of this model is to predict an NHL center’s salary based on various performance metrics, I utilize two different types of data; NHL

player salary data for the 2017-2018 season, and on-ice player performance data for the 2017-2018 season, both of which are from the most recent completed NHL season and were supplied by CKM Sports Management. The provided player salary data set includes the salary cap hit for every player that played a game during that season, the player's age, and the type of contract they are under, whether it was a standard contract, entry-level contract, or 35+ contract. As for the performance data, it is divided into three different data sets, which were delineated by the three different in-game situations, five-on-five, powerplay, and penalty kill. Each of the three data sets had 170 variables, all of which represent various on-ice performance metrics.

The 170 variables in the data set consisted of different types of performance metrics. The largest subtypes of performance variables were total production, individual production, production per 60 minutes played, and relative production. Of the four variable subtypes, I chose to focus on two. The first of which is individual production variables, because every player is personally responsible for their performance. Secondly, I focused on production per 60 minutes played variables to smooth the effects on players that missed games due to injury or other circumstances. I find the other two variable subtypes to be too reliant on aspects of production that are out of the individual player's control, such as teammate quality. Total production included linemate contributions, and relative production was defined as the difference in team performance with the player on the ice versus off the ice. Therefore, I did not utilize them in the model.

In order to get the full scope of a center's on-ice production, I merged the three performance data sets together, but kept the variables from each data set separate by labelling the five-on-five variables '5v5', the powerplay variables 'PP', and the penalty kill variables 'PK'. Once I had the combined performance data set, I appended the salary data set to it, which was not as straightforward because the player names were formatted differently between the two data sets. After I had

bridged the formatting differences in player names, I had my completed working data set consisting of approximately 500 variables, including all performance data and the accompanying salary data for the 800 players that played in the 2017-2018 season.

TABLE 1 – SUMMARY STATISTICS OF KEY VARIABLES

Variable	Obs	Mean	Std. Dev.	Min	Max
Cap Hit (in $)	150	3,591,350	2,495,109	625,000	10,500,000
Age (years)	150	28.1933	3.9658	21	41
G/60 (5v5)	150	0.5985	0.2421	0.13	1.4
A/60 (5v5)	150	0.8789	0.3671	0.09	1.99
iCF/60 (5v5)	150	11.3150	2.6054	4.64	18.85
iSF/60 (5v5)	150	6.6101	1.5626	2.87	11.07
iPent/60 (5v5)	150	0.7005	0.4830	0	4.82
iPend/60 (5v5)	150	0.7054	0.3798	0.08	3.25
iTKA/60 (5v5)	150	1.7889	0.6824	0.43	3.65
iGVA/60 (5v5)	150	1.5899	0.6665	0.54	3.8
iHF/60 (5v5)	150	5.0233	3.4345	0.77	19.97
iFOW/60 (5v5)	150	18.2237	11.1389	0.0915	45.2470
G/60 (PP)	150	1.0939	1.1440	0	4.38
A/60 (PP)	150	1.4918	1.5454	0	5.43
iFOW/60 (PP)	150	11.2536	12.3773	0	43.8678
iTKA/60 (PK)	150	1.6748	1.8110	0	7.45
iBLK/60 (PK)	150	2.6511	2.7808	0	12.83
iFOW/60 (PK)	150	14.1886	15.1209	0	55.6970
ATOI (mins)	150	14.7070	3.0003	7.6364	20.5698

Since the model is focused on centers in particular, I dropped all players that played different positions. Some players were labelled under multiple positions, so I checked team line-ups from the 2017-2018 season to confirm if a player played more games as a center or a winger. I substantiated my position adjustment decisions by checking how many face-offs the player took before changing them to center or winger. It was necessary that I did both as it is common for wingers to take a face-off and then return to playing as a winger for the remainder of their shift. Once I had narrowed my data set to centers, I removed players with missing data. This included players with no cap hit information or no games played. I extended this missing data removal by dropping all players that played less than 30 games, as their sample size was too small and could have skewed performance data. Additionally, it could be argued that a majority of those

players with less than 30 games played are not fully in the NHL anyway. The next step was to drop all rookie players, or players under an entry-level contract. This was necessary because entry-level contracts are standardized and limited to $925,000 regardless of the player's performance. If I did not drop them, they would have negatively impacted the predictive performance of the model as their salary has no relation to their performance. The final adjustment to the data consisted of fixing the team of players that played for multiple-teams due to mid-season trades. My decision to choose which team the player played for was based on which team offered the player the current contract they are under, as the team they were traded to have no say in the contract and subsequent cap hit.

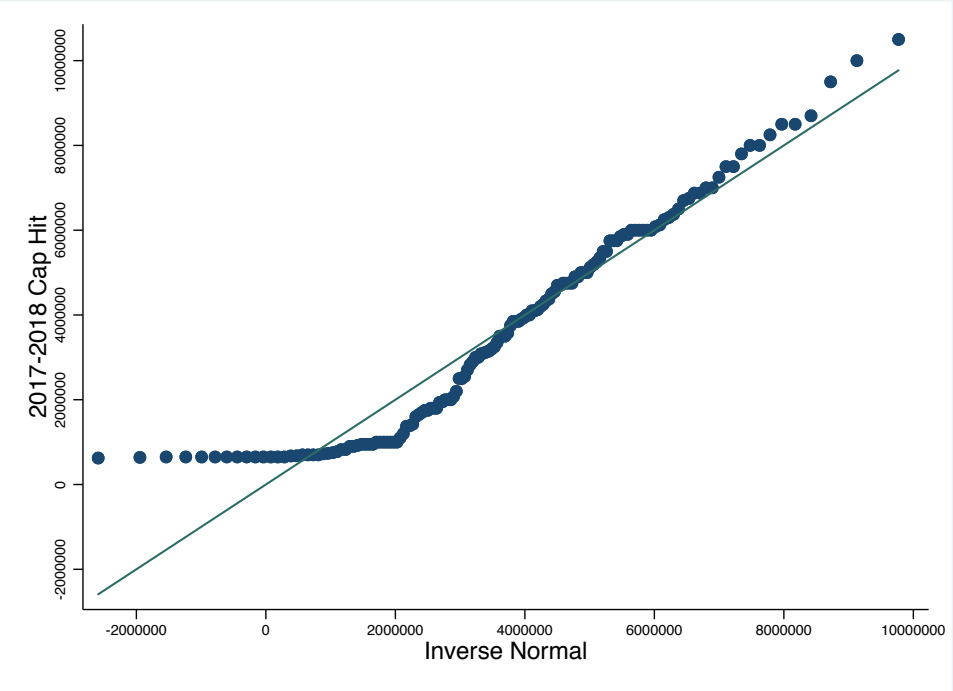

FIGURE 1. SALARY CAP DISTRIBUTIONS OF CENTERS IN 2017-2018

The reason behind adjusting the team of the players was to generate dummy variables for each of the 31 teams in order to assess potential team fixed effects, as

mentioned in Kahane (2001). Age dummy variables were also generated to establish an age control. In addition, some players did not appear on the powerplay or penalty kill, so I generated dummy variables for those players that had missing powerplay or penalty kill data to prevent them from being dropped when utilizing powerplay or penalty kill variables in the regression. Since I focused on metrics measured in 60 minutes played, I had to generate face-offs won per 60 minutes played, as they were the only variables in the data set that did not have per 60-minute measures. Additionally, the use of 60 minutes played variables led to the need for a measure of average time on ice per game, which was not present in the data set. I created the average time on ice variable by adding the player's total time on ice at five-on-five, powerplay, and penalty kill, and then dividing that sum by the number of games they played.

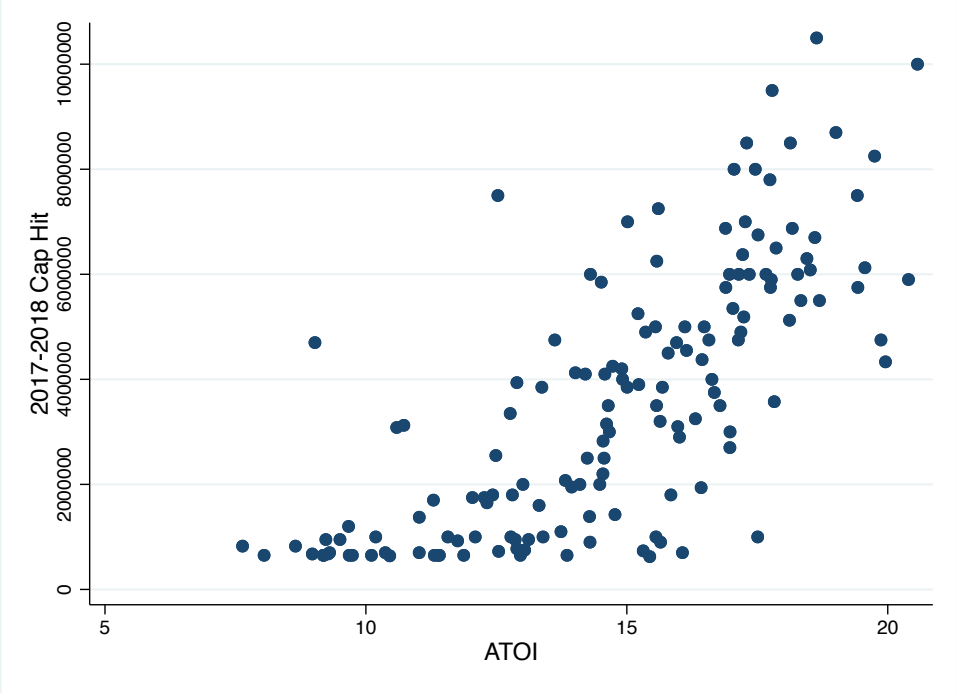

FIGURE 2. SCATTERPLOT OF SALARY CAP HIT AND ATOI

IV. Model

The empirical modelling technique I utilize as a salary prediction tool is a multiple linear regression model consisting of 17 independent variables, with the dependent variable being predicted cap hit. The baseline multiple linear regression equation is as follows:

$$(1) \quad y_i = \beta_0 + \beta_1 G/60_{5v5} + \beta_2 A/60_{5v5} + \beta_3 iCF/60_{5v5} + \beta_4 iSF/60_{5v5} + \beta_5 iPenT/60_{5v5} + \beta_6 iPenD/60_{5v5} + \beta_7 iTKA/60_{5v5} + \beta_8 iGVA/60_{5v5} + \beta_9 iHF/60_{5v5} + \beta_{10} iFOW/60_{5v5} + \beta_{11} G/60_{PP} + \beta_{12} A/60_{PP} + \beta_{13} iFOW/60_{PP} + \beta_{14} iTKA/60_{PK} + \beta_{15} iBLK/60_{PK} + \beta_{16} iFOW/60_{PK} + \beta_{17} ATOI + \varepsilon_i$$

Before settling with a multiple linear regression, I explored different modifications like changing independent variables into quadratic terms to see if the relationship between the independent terms and the dependent variable were better estimated as a quadratic relationship rather than a simple linear relationship. Upon doing that, I found no significant changes in coefficients, and the R-squared decreased, leading me to believe that I was better off keeping the model linear. Additionally, I looked at a log dependent variable as an alternate model specification but could not reasonably justify the log model being an improvement over the linear model, especially with the terms being expressed as percentage effects on predicted salary. On top of that, the R-squared fell slightly with the log model. The multiple linear regression model solves the question of an NHL center's value by predicting salary based on various on-ice performance metrics by taking the respective coefficient values of each independent variable in the regression and combining them to determine the deserved salary based on the player's statistics.

As mentioned before, the data set had over 500 variables, and I have narrowed it down to 17 variables that I consider important performance measures that lead to an accurate salary prediction for an NHL center. The 17 variables can be further broken down into 3 categories. The first category consists of offensive capabilities, and it includes goals scored per 60 minutes at 5v5 ($G/60_{5v5}$), goals scored per 60 minutes on the powerplay ($G/60_{PP}$), assists per 60 minutes at 5v5 ($A/60_{5v5}$), assists per 60 minutes on the powerplay ($A/60_{PP}$), and shots on goal per minutes at 5v5 ($iSF/60_{5v5}$). These five variables are among the most common measures for offensive production and are widely considered the most important statistics when determining a center's offensive effectiveness in the NHL.

However, a center must also provide defensive contributions in order to be successful in the NHL. That is why the second category consists of defensive capabilities, which includes takeaways at 5v5 ($iTKA/60_{5v5}$), takeaways on the penalty kill ($iTKA/60_{PK}$), hits at 5v5 ($HF/60_{5v5}$), and blocked shots on the penalty kill ($iBLK/60_{PK}$). Defensive performance metrics are not as popular as offensive ones, but in order to measure a center's defensive prowess, it is necessary that they are included in the model. The most common defensive measures are takeaways, hits, and blocked shots, which is the reason I included them in the model.

The third and final category is not as straightforward as the previous two, but it is equally important. It consists of various puck possession metrics, some of which are derived and supported by theoretical assumptions due to the fact that the NHL does not have a proper method for tracking a player's actual puck possession statistics. One of the more popular possession metrics included is individual corsi ($iCF/60_{5v5}$), which is the closest measure for puck possession currently available. Also included is faceoff wins in the three different situations ($iFOW/60_{5v5}$,

$iFOW/60_{PP}$, $iFOW/60_{PK}$), as faceoffs are a responsibility unique to centers. I included penalties taken at 5v5 ($iPenT/60_{5v5}$) as a proxy measure for instances in which a player did not have the puck, as a player generally takes a penalty when they do not have possession of the puck. I also included penalties drawn at 5v5 ($iPenD/60_{5v5}$) as a proxy measure for instances in which a player did have the puck, as a player draws a penalty when they either have the puck, or are involved in the play in some capacity that led to the opponent illegally checking them. Another variable I included under theoretical assumptions is giveaways at 5v5 ($iGVA/60_{5v5}$). Although giveaways are a negative performance measure, I justify the inclusion of this variable through two assumptions. The first assumption is that a player that has the puck more often, will give the puck away more often simply because they possess it more, and therefore are more likely to give it away. The second assumption is that good players with many giveaways will have higher tolerance from their coaches and will continue to play since their other contributions outweigh their mistakes, and bad players will receive less opportunities to play after only a few giveaways. With these assumptions, I can consider giveaways a proxy measure for puck possession. The more giveaways the player has, the more often they possess the puck.

The one variable not included in any of the three categories is average time on ice per game (ATOI). Perhaps the most important of all variables, average time on ice per game can be considered a measure of a player's value to the coach. However, this variable also requires two assumptions. Firstly, it would be reasonable to assume that a coach would give better players more ice time than inferior players in order to improve team effectiveness and increase the probability of winning the game. The second assumption is that a coach will not give a player more ice time simply because the player is being paid more than others. The second assumption must hold to eliminate the potential presence of reverse causality. In most scenarios, the disconnect between the general manager's job to sign players

and the coach's job to optimally utilize the players he is given makes reverse causality unlikely.

On top of the baseline model, I added a control for age (regression 2), and a control for team (regression 3), along with a fourth regression that included both controls. An age control is necessary because the age of a player during contract negotiations has a significant impact on the term of the contract, which has a subsequent impact on the average annual value of the contract. An older player receives shorter contracts due to impending depreciation of their skills and health, therefore impacting their average annual value. On the other hand, younger players that sign longer deals will receive more per annum to compensate for the fact that they are locked up for more of their prime years and the salary cap generally rises annually. It controls for contract value differences in players that are entering their prime career years, currently in their prime years, or exiting their prime years. I introduce a team control in order to account for team fixed effects, which may arise due to differentiation in taxation across states and countries in which teams are situated, along with other individualistic components of teams such as management style, or a player's interest in that specific market.

V. Results

Table 2 displays the outputs for all four regressions:

TABLE 2 – REGRESSION OUTPUTS

VARIABLES	(1) CAP HIT	(2) CAP HIT	(3) CAP HIT	(4) CAP HIT
G/60 (5v5)	-635,201	-203,271	-818,224	151,228
	(651,568)	(655,022)	(788,863)	(816,894)
	0.331	0.757	0.302	0.854
A/60 (5v5)	3,425	257,177	24,272	282,871
	(432,628)	(466,796)	(520,172)	(571,451)
	0.994	0.583	0.963	0.622
iCF/60 (5v5)	-201,948	-128,807	-237,998	-103,861
	(150,485)	(159,573)	(178,424)	(189,371)
	0.182	0.421	0.185	0.585

iSF/60 (5v5)	344,245	228,468	373,058	144,097
	(257,745)	(276,700)	(311,693)	(336,556)
	0.184	0.411	0.234	0.670
iPent/60 (5v5)	1.012e+06***	770,930**	879,316**	230,869
	(357,173)	(385,008)	(443,176)	(505,648)
	0.00533	0.0476	0.0499	0.649
iPend/60 (5v5)	-696,941	-571,174	-663,025	-123,757
	(443,749)	(466,413)	(541,346)	(600,552)
	0.119	0.223	0.223	0.837
iTKA/60 (5v5)	-197,591	-178,886	-154,241	-158,026
	(221,458)	(227,230)	(275,901)	(295,405)
	0.374	0.433	0.577	0.594
iGVA/60 (5v5)	285,112	313,486	563,715*	654,341*
	(225,660)	(223,787)	(325,496)	(334,400)
	0.209	0.164	0.0863	0.0537
iHF/60 (5v5)	-12,808	25,955	15,420	65,250
	(48,721)	(52,193)	(59,392)	(64,177)
	0.793	0.620	0.796	0.312
iFOW/60 (5v5)	54,817***	44,347**	44,220**	38,638
	(18,388)	(19,500)	(20,899)	(23,625)
	0.00342	0.0248	0.0368	0.106
G/60 (PP)	64,854	62,043	95,004	33,230
	(158,062)	(167,031)	(188,016)	(203,551)
	0.682	0.711	0.614	0.871
A/60 (PP)	419,393***	278,604**	450,149***	254,553
	(133,157)	(139,374)	(154,759)	(166,084)
	0.00202	0.0480	0.00445	0.129
iFOW/60 (PP)	-21,918	-11,924	-18,361	-3,524
	(16,256)	(16,696)	(18,714)	(19,743)
	0.180	0.477	0.329	0.859
iTKA/60 (PK)	-136,769	-104,397	-137,602	-80,630
	(83,133)	(83,204)	(101,727)	(106,055)
	0.102	0.212	0.179	0.449
iBLK/60 (PK)	-86,080	-104,615*	-87,492	-97,207
	(54,877)	(57,883)	(65,953)	(71,020)
	0.119	0.0733	0.188	0.175
iFOW/60 (PK)	11,061	8,122	14,508	-239.3
	(13,540)	(13,821)	(15,285)	(16,912)
	0.415	0.558	0.345	0.989
ATOI	478,544***	483,220***	439,002***	461,526***
	(78,930)	(80,028)	(97,032)	(102,322)
	1.32e-08	2.00e-08	1.64e-05	2.08e-05
Constant	-4.461e+06***	-5.144e+06**	-4.081e+06**	-5.929e+06**
	(1.174e+06)	(1.973e+06)	(1.633e+06)	(2.561e+06)
	0.000219	0.0104	0.0140	0.0230
Age Control	No	Yes	No	Yes
Team Control	No	No	Yes	Yes
Observations	150	150	150	150
R-squared	0.701	0.766	0.746	0.809

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.10

The only two variables that maintain relatively consistent coefficients across all four regressions is face-offs won per 60 minutes at 5v5 ($iFOW/60_{5v5}$) and average time on ice (ATOI). As mentioned before, I considered average time on ice to be one of the more important variables as it would be reasonable to assume that the better a player is, the more ice time they will be allotted by their coach. The regression output appears to substantiate my claim as the variable was statistically significant at a 0.01 level across all four regressions. As for face-offs won per 60 minutes at 5v5, it held statistical significance for three of the four regressions. However, there is a high likelihood that this variable is strongly correlated with average time on ice, since the more time a center spends on the ice, the more often they will take face-offs. Another interesting result is the negative coefficient on goals per 60 minutes at 5v5, which insinuates a negative correlation between the amount of goals a player scores and their salary. That being said, it is possible that the positive aspects of scoring a goal is contained within the positive coefficient for shots taken, as those two variables are highly correlated. Due to the high probability that all of these variables are correlated with one another, it is difficult to draw conclusions by singling out specific coefficients.

The baseline model has an R-squared of 0.701, and rises to 0.809 with the addition of controls. Although the R-squared is relatively high, it could be the result of simply having a large number of independent variables in the regression, since adding more variables mechanically increases the R-squared of a model. Table 3 and figure 3 display the distribution of actual cap hits with the best-fit line representing the predicted cap hits from the baseline regression.

TABLE 3 – SUMMARY STATISTICS OF BASELINE REGRESSION PREDICTION DIFFERENCE

Variable	Obs	Mean	Std. Dev.	Min	Max
B. Prediction Diff	150	0.0058	1,363,533	-4,023,677	4,231,788

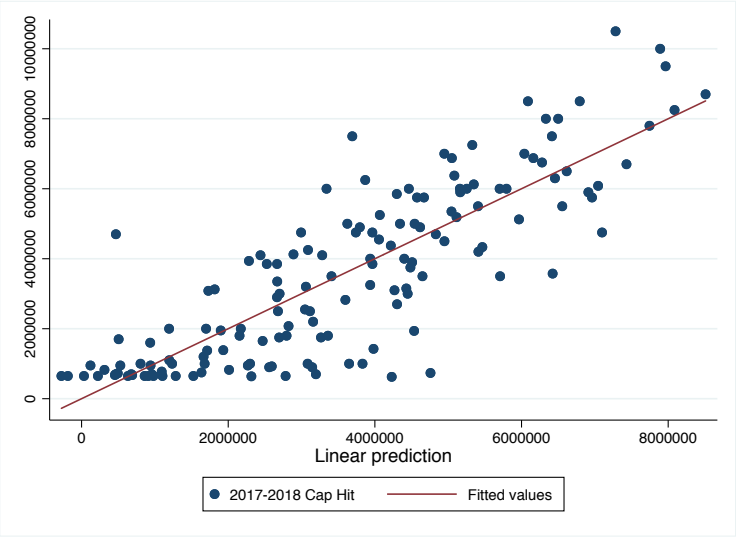

FIGURE 3. BASELINE REGRESSION BEST-FIT LINE

Table 4 and figure 4 display the distribution of actual cap hits with the best-fit line representing the predicted cap hits from the fully controlled regression.

TABLE 4 – SUMMARY STATISTICS OF CONTROLLED REGRESSION PREDICTION DIFFERENCE

Variable	Obs	Mean	Std. Dev.	Min	Max
C. Prediction Diff	150	-0.0092	1,091,215	-2,953,199	2,842,667

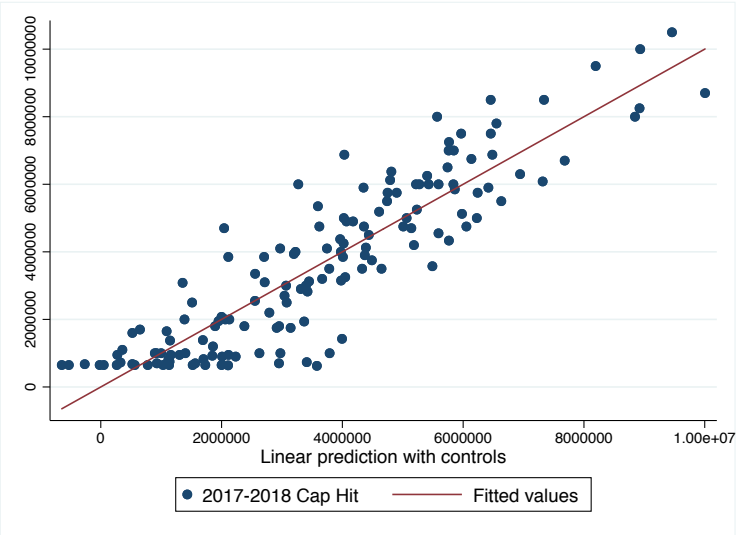

FIGURE 4. CONTROLLED REGRESSION BEST-FIT LINE

If we compare the two graphs and the summary statistic tables, it is the clear that the fully controlled regression has less variation, and less outliers. Because of this, I argue that the controls do have a positive impact on the predictive performance of the model.

To provide concrete results of the predictive ability of the model, I include a list of the ten most underpaid (Table 5) and overpaid players (Table 6) according to the baseline model.

TABLE 5 – MOST UNDERPAID PLAYERS (BASELINE RESULTS)

Player Name	2017-2018 Cap Hit	Predicted Cap Hit	Difference
Chris Tierney*	$735,000	$4,758,677	-4,023,677
Colton Sissons*	$625,000	$4,229,823	-3,604,823
Mikael Backlund*	$3,575,000	$6,424,591	-2,849,591
David Desharnais*	$1,000,000	$3,831,380	-2,831,380
William Karlsson*	$1,000,000	$3,648,983	-2,648,983
Vladislav Namestnikov	$1,937,500	$4,535,604	-2,598,104
Derek Ryan*	$1,425,000	$3,981,658	-2,556,658
Mattias Janmark*	$700,000	$3,195,681	-2,495,681
Vincent Trocheck	$4,750,000	$7,098,415	-2,348,415
Markus Granlund	$900,000	$3,145,315	-2,245,315

TABLE 6 – MOST OVERPAID PLAYERS (BASELINE RESULTS)

Player Name	2017-2018 Cap Hit	Predicted Cap Hit	Difference
Jori Lehtera*	$4,700,000	$468,212	4,231,788
Jason Spezza*	$7,500,000	$3,691,923	3,808,077
Jonathan Toews	$10,500,000	$7,283,001	3,216,999
Logan Couture*	$6,000,000	$3,340,468	2,659,532
Leon Draisaitl	$8,500,000	$6,087,620	2,412,380
Patrick Marleau	$6,250,000	$3,868,348	2,381,652
Anze Kopitar	$10,000,000	$7,890,907	2,109,093
Paul Stastny	$7,000,000	$4,947,251	2,052,749
David Krejci	$7,250,000	$5,328,518	1,921,482
Ryan Kesler*	$6,875,000	$5,050,368	1,824,632

Notes: * denotes a re-appearance in the list for most underpaid or overpaid in the fully controlled regression results

With prior knowledge of NHL centers and their performance in the 2017-2018 season, arguments in favour of the appearance of each player on either the underpaid or overpaid list can be made. Therefore, I believe the regression equation answers the question of an NHL center's value based on their on-ice performance with reasonable accuracy.

VI. Discussion

Although the results between the baseline regression model and the fully controlled regression model vary slightly, that can attest to the robustness of the baseline model. The addition of two major controls decreased the variability in predicted salaries, but many of the players in the underpaid or overpaid list according to the baseline model re-appear in the list for the fully controlled model. So despite the change in variation, each model provided similar results. As another robustness check, I added fenwick per 60 minutes at 5v5 as an extra variable into the baseline model to show that the model does not completely collapse with the introduction of another variable.

TABLE 7 – SUMMARY STATISTICS OF REGRESSION DIFFERENCE WITH ADDED VARIABLE

Variable	Obs	Mean	Std. Dev.	Min	Max
Robust Difference	150	-0.0093	8,187.992	-22,533.5	22,330.63

Table 7 shows the summary statistics of the difference between the original baseline regression prediction values and the prediction values of the new baseline regression with the added variable. Given that we are dealing with salary figures in the millions of dollars, a standard deviation of just $8188 and a minimum and maximum difference of approximately $22,500 is negligible. Therefore, it is safe to say that the added variable had practically zero effect on the original baseline model's prediction ability.

As previously mentioned, rookies were dropped from the data set because they are under standardized entry-level contracts that are limited to $925,000, regardless of their actual performance. The fairness of the limited entry-level contract that rookies must sign has been widely debated amongst the NHLPA, the NHL, agents, and fans. Although it greatly benefits team management to underpay

rookies and utilize the salary cap savings on other players, it could lead to significant losses in career earnings for the rookies. As an extension of the model, I ran the baseline regression on a separate data set of just rookie centers to determine the magnitude in which rookies are being underpaid.

TABLE 8 – SUMMARY STATISTICS OF BASELINE REGRESSION ON ROOKIES

Variable	Obs	Mean	Std. Dev.	Min	Max
ELC Difference	36	-1,948,153	1,873,139	-5,332,489	1,300,106

With a mean of approximately -$2,000,000, that means rookies are underpaid by $2,000,000 per year on average. Of course, this underpayment is mitigated by the various performance bonuses of $212,500 contained in the entry-level contract, but the bonuses are capped at $850,000. So if a rookie were to max out their performance bonuses, the most they can earn in a year is $1,775,000. That still results in an underpayment of over $1,000,000 per year, and potentially much more for more prolific rookies. This information can be utilized by the NHLPA to make amendments to the entry-level contract under the current collective bargaining agreement (CBA) to make compensation more fair for productive rookies. The information can also be useful for agents that are negotiating future contracts for players coming off of their entry-level contract to reimburse lost earnings through the inclusion of back pay in their next contract.

Despite the decent results, the predictive performance of the model is hindered by several limitations. These limitations arise in two different ways. The first way is through inherent aspects of the game of hockey and the NHL as a league. This includes the inability to quantify the intangible aspects of a player, like their leadership, experience, grit, or affect on locker room morale. These aspects play a significant role in a player's contribution to a team, and therefore their financial value. Additionally, contract negotiations within the NHL place a lot of

weight on precedence of recent signings of comparable players. This influences salary values in a way that cannot be measured by statistics. Furthermore, the availability of certain types of players in free agency varies year by year, which impacts the bargaining power of either the player or the team depending on a market surplus or shortage of a specific player type. This establishes another influence on salary values that cannot be quantified. The second way limitations arise is from absences in the data set provided by CKM Sports Management. There was no data for four-on-four situations or three-on-three situations, which includes overtime. Although minor, that extra data would have made the model slightly more accurate. In addition, the data set provided only contained one season of data, which prevented me from being able to assess the impact of historical performance on current contracts. Lastly, there was no contract length data. As mentioned before, the term of a contract affects the average annual value of the contract. For example, the Toronto Maple Leafs signed Auston Matthews to a five-year deal with an average annual value of $11.634 million dollars. The Toronto Maple Leafs could have signed Auston Matthews for more years, but that would have required them to pay him more annually because they would have locked him up for more of his prime years and would have had to offset the annual rise in the salary cap. This mechanism can work inversely too, as non-elite players can negotiate a longer term in exchange for a lower average annual value in order to ensure job security. This can be exemplified by the six-year, $2 million dollar per year contract signed by Calle Jarnkrok and the Nashville Predators. Although not much can be done to overcome the inherent limitations, it would be easy to overcome the data limitations with further additions of the data that was absent.

With annual rises in the salary cap and a constant roster size of 20 players, that means the average annual value of salaries increases every year. Therefore, the predicted values determined by the model must be updated every year with current salary figures in order to account for the annual increase. Additionally, the

performance of certain centers can vary greatly year by year, so their predicted values can change drastically. Regardless, as long as the way hockey is played remains the same, the independent variables in the model should predict a center's value based on their on-ice performance with reasonable accuracy.

VII. Conclusion

In this paper, I identified two motivating factors behind answering the question of an NHL center's value based on their on-ice performance. One side being the need for management to efficiently allocate their limited salary dollars, and the other being the need for players to receive the proper financial compensation they deserve. These two pressures are created by the imposition of a salary cap, which was implemented as a mechanism for financial parity. I answer the question by using a multiple linear regression with 17 independent variables that represent different performance measures that I believe are important for a center. Due to the high correlation between variables, it is difficult to determine which particular variables are influential on a player's salary. However, I did find that average time on ice is the most significant determinant in salary, but I could not find a remedy for reverse causality besides basic theoretical assumptions. Despite this, the overall model produces predicted salaries that can be utilized as a comparison against actual player salaries to determine the most underpaid or overpaid players. That being said, the results are not perfect because of the various limitations that arise inherently within the game of hockey and minor data limitations. I believe great strides in salary prediction based on on-ice performance would be made with improved tracking of the amount of time each player possesses the puck per game, as puck possession is currently the name of the game.

REFERENCES

Jones, J., & Walsh, W. (1988). Salary Determination in the National Hockey League: The Effects of Skills, Franchise Characteristics, and Discrimination. *Industrial and Labor Relations Review, 41*(4), 592-604. doi:10.2307/2523593

Leo H. Kahane (2001) Team and player effects on NHL player salaries: a hierarchical linear model approach, Applied Economics Letters, 8:9, 629-632, DOI: 10.1080/13504850010028607

Richardson, D. (2000). Pay, Performance, and Competitive Balance in the National Hockey League. *Eastern Economic Journal, 26*(4), 393-417. Retrieved from http://www.jstor.org.ezproxy.library.ubc.ca/stable/40326440

The Business of Hockey. (2018). *Forbes*. Retrieved from https://www.forbes.com/nhl-valuations/list/