Paid to Play: An Analysis of NHL Defensemen Salary in Relation to Individual On-Ice Performance

By: Kimberly Woo

Executive Summary

Objective *How can one predict the salary of an NHL defensemen based upon their individual 2017-2018 season performance data?*

- Developed a log-linear regression model
- Individual on ice performance variables most influential to defensemens' salaries were identified.
- 10 performance variables were selected for assessing individual performance, scaled to 60 minutes of time on ice for even and uneven strength game statistics.
- The model shows that 50% of the data is explained.

Note: The model does not account for traded players and rookies

Key Findings *What are some of the trends and highlights that the data shows?*

- Defensemen are most valued for specializing in the game situations of power plays and penalty kills.
- Power plays tested to hold the most significant performance variables, specifically in blocks, hits, and penalties drawn, according to salary increase.
- Salaries show rewarding defensemen for "defensively aggressive" behavior (physical contact or non-puck possession performances).

Implications *How do we move forward with the information presented in this report? What are some possible recommendations?*

- This study provides a preliminary model with specific assumptions and a small data that demonstrates the need for continued research.
- Shows evidence opposing to the use of plus/minus statistics in evaluating the performance of defensemen (relies more on teammate and opponent strength as opposed to an individual defensemen's ability).
- Future research or models could examine performance over time instead of just a single season.
- Tracking total passes and completed passes could be more telling to puck turnover rates in addition to monitoring giveaways and takeaways.
- Once an appropriate model has been established, it should be implemented into a database or program which could identify an approximate salary of a player by inputting their hockey position and performance statistics.
- This could be further enhanced by creating an input for the potential team the player could belong to, as salary caps vary across teams.

Power Play

1 point increase/decrease reflected as a % increase/decrease in salary:

Corsi For: 1%	Prob. of Goal: 39%
Fenwick For: 0%	Goals Scored: -14%
Giveaways: 1%	Hits For: 9%
Takeaways: 4%	Blocks: 10%
Total Assists: 14%	Penalties Drawn: 30%

Trends

Roughly 60% of the players did not play in power plays and 40% did not play in penalty kills during the season. Players who do not play in power plays show a 11.6% decrease in salary and declines to 78.5% for players who do not play in penalty kills.

Consider

Expanding to create a more comprehensive model in addition to performance, including measuring factors such as intangibles, the effects of age, the value in physical features, or external revenue generating factors such as fan popularity. In order to help managers build their teams, a program could be written to match a position and certain salary value to the players who fit that criteria. This could become valuable for making trade decisions.

PAID TO PLAY: AN ANALYSIS OF NHL DEFENSEMEN SALARY IN RELATION TO INDIVIDUAL ON-ICE PERFORMANCE

By KIMBERLY WOO

As the highest competitive organization of hockey in North America, the National Hockey League (NHL) pays its players between $500,000 upwards to around $15,000,000 per year (Badenhausen, 2017). Existing studies in analyzing NHL player salaries lack in variables that emphasize an individual player's performance. The current measure of on-ice performance calculated for salary is based on the total point production statistic (commonly referred to as plus/minus points) which has been shown to be inaccurate in gauging a player's individual performance. This raises the economic question; how can one predict the salary of an NHL defensemen based upon their individual 2017-2018 season performance data? Using a log-linear regression model, individual performance variables most influential to salary valuation were identified, where the predictability power of model holds at ~50%. The results show that defensemen are most valued for specializing in the game situations of power plays and penalty kills. Specifically, blocks, hits, and penalties drawn during power plays were most influential in terms of salary increase. This result was interpreted as rewarding a defensemen's salary for "defensively aggressive" behavior which are physical contact performances as opposed to puck possession performances.

I. Introduction

Many kids who play hockey dream of making a professional career or playing in the National Hockey League (NHL) and while passion for the sport is key in driving such interests, the lucrative nature of the industry is also desirable. The NHL is the top performing professional hockey league within North America. It currently hosts 31 teams and it is each team's responsibility to dictate their players' salaries. The current range of salaries in the NHL start at a minimum of $500,000 and can go to upwards of $15,000,000 dollars (Badenhausen, 2017).

A problem observed by NHL teams is that many players are being overvalued or undervalued for their performance, because their contracts, all other factors aside, are mainly monetized by one statistic: their total production of points (commonly referred to

as plus/minus) for all games played. The plus/minus statistic is a number produced by one point given to the players on the ice for every goal scored by their team and a deduction of one point given to the players on the ice for every goal scored by their opposition. This is a problematic statistic. For example, a player could score lots of goals but still have a negative plus-minus value if the opposition scored more. The limitations of the plus/minus statistic are that it relies heavily on the performance of a player's teammates and opponents, which makes evaluating their performance based on their own abilities more challenging, especially for defensemen due to the nature of their role. While salaries are decided based on other factors that do not include performance, I am interested in the value that these players produce during a game and if their salary matches their level of production. The research question I will be answering is how to predict the salary of an NHL defensemen based upon their 2017-2018 season performance data.

In this paper, I will introduce the background context and literature of hockey and how the majority of NHL teams currently evaluate performance based on salary. Then, I will explain the how the data set was created and the source, as well as the restrictions and modifications made. I will demonstrate how a log-linear model is justified and used in answering the economic question at hand. Finally, I will outline my results and provide a further discussion as well as concluding thoughts on the implications of my study.

II. Background

Put simply, the objective of any game of hockey is to hit a small puck across the ice with your stick into the opponent's net to score a goal. The team with the most goals

at the end of the game wins. The puck can be passed around the players on the ice and the role of the opposition is to try and prevent the other team from scoring. Each team can have a maximum of 20 players and of these 20 players, only six may be on the ice at any one time. The rest will be used as substitutes but can come and go from the game as often as required. These six players on ice include a goal tender and 5 skating players consisting of 3 forwards (centre, left wing, right wing) and 2 defensemen and each game lasts for three 20 minute periods.

To give a better understanding of the dynamics of hockey in terms of salary, I will analyze hockey players as well as the NHL through an economic lens by making the NHL analogous to a market. The NHL can be viewed as a market that contains firms made by each team. The objective of each firm is to produce as many wins in the games played within a season as possible by using its players as a form of a capital good. By investing or paying their players based on the skills they possess in the game of hockey, each team will want to maximize their utility and profit by being efficient. In other words, teams want to pay as little money as possible for the most amount of output, similar to the concept of economies of scale.

While there are many studies that provide models on determining NHL players' salary based on performance, they do not come to a consensus on what method or variables are best to use in an analysis such as the one conducted in this paper. I will present two studies most relevant in providing further context and support in my choice of methodology and variables for building my specification model.

Depken and Lureman find “evidence that salary inequality reduces team performance primarily through reduced defensive production (more goals allowed) than through offensive production (fewer goals scored)” (2017). This suggests that salary inequalities in the NHL could be justified through efficiency wages. That is, teams need to be willing to pay higher salaries for their most productive players in order to incentivize them to continue performing at a high level. Yet, the study bases their data on team performance rather than individual player performance. It would be important to identify if their model could support the use of efficiency wages within a team based on the performance of individual players.

A regularized logistic regression model created by Gremacy et al. examined how telling the plus/minus statistic is in terms of events that lead to a goal scored. By estimating the credit or blame players incur during the event of a goal scored, they identify how each player contributes on ice, “beyond their aggregate team performance aggregate team performance and other factors, to the odds that a given goal was scored by their team” (2013). Their results show that the plus/minus points yield a marginal effect which is related more to the performance of a player’s teammate and opponents’ strength as opposed to their individual contribution. Gremacy et al. also notes that “plus-minus does not control for sample size, such that players with limited ice-time will have high variance scores that soar or sink depending on a few chance plays” (2013). However, Gremacy et al.’s model did not account for performances in uneven strength events (power play and penalty kills), and only accounted for performance data during 5 on 5 game situations.

III. Data

A cross-sectional data set is created to reflect 456 performance variables and the salaries of 733 players who played in the NHL 2017-2018 season. Within this data set, 256 players are defensemen. There are 4 subsets of data used to create the final data set, provided by CKM Hockey Management. These 4 sets consist of 3 performance statistics set, one representing each of the 3 possible game situations, and the final data set contained the salary and contract details of the players. The 4 data sets were first merged to the names of each player. All players who do not play in the defensemen position and players traded within the season were dropped from the data set. The exclusion of traded players was done in order to rule out external factors that could influence performance. A restriction was placed on "entry level" contract players as these players usually do not play in very many games or receive a lot of ice time and receive different performance expectations from players who are under "standard or 35+" contracts. This leaves 195 defensemen remaining in the data set.

The variables selected for my model are what I expect to be attributes of a good defensemen in the overall game of hockey. The variables will also include statistics linked to power plays and penalty kills, as some defensemen are utilized more than others for certain game situations. This was also uncommon in other studies and a suggestion made by Gremacy et al. to further improve future models. As a defenseman, their role specializes in protecting their goal from being scored on by the opposing team. Most of their ice time is spent in the defensive/neutral zone in which they may see less puck possession and scoring opportunities.

All variables in the model represent continuous, quantitative individual performance statistics scaled over 60 minutes of continuous ice-time, for all 3 types of game situations: 5 on 5 (5v5), power plays (PP), and penalty kills (PK), with the exception of two dummy variables. Figure 1 shows the summary statistics of the variables used within the model.

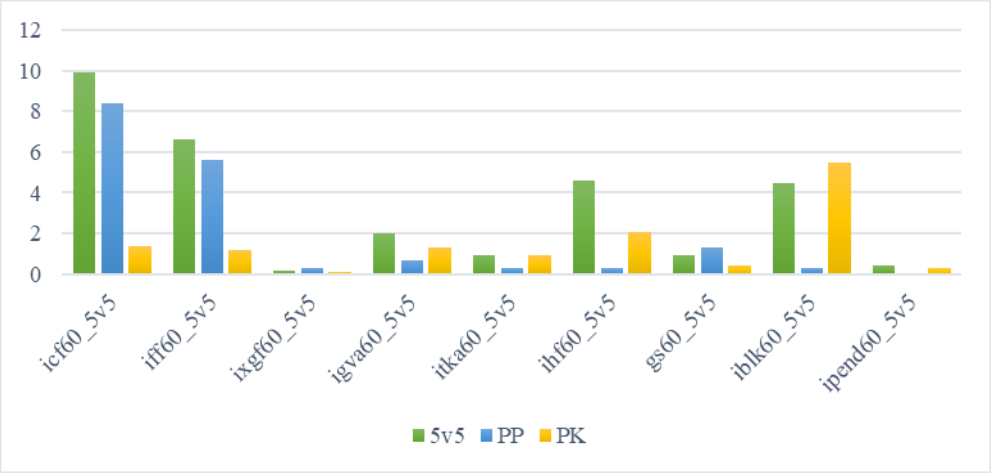
FIGURE 1. TABLE OF SUMMARY STATISTICS

	5v5			*PP*			*PK*		
Variable	N	Mean	Std. Dev.	N	Mean	Std. Dev.	N	Mean	Std. Dev.
icf60	195	9.9	2.7	195	8.4	12.0	195	1.4	1.4
iff60	195	6.6	1.9	195	5.6	8.0	195	1.2	1.3
ixgf60	195	0.2	0.1	195	0.3	0.5	195	0.1	0.1
igva60	195	2.0	0.8	195	0.7	1.1	195	1.3	1.3
itka60	195	0.9	0.5	195	0.3	0.5	195	0.9	1.1
ihf60	195	4.6	2.8	195	0.3	0.6	195	2.1	2.3
gs60	195	0.9	0.6	195	1.3	1.8	195	0.4	0.4
iblk60	195	4.5	1.1	195	0.3	0.5	195	5.5	4.7
ipend60	195	0.4	0.3	195	0.0	0.1	195	0.3	0.4
a60	195	0.5	0.3	195	1.3	1.9	195	0.1	0.3
logcaphit	195	14.61	0.9						
noPP	195	0.6							
noPK	195	0.3							

A total of 33 variables were selected for the specification model. Non-puck possession variables such as hits for (ihf60_), blocks (iblk60_), and penalties drawn (ipend60_). The puck possession variables for the probability of a goal scored (ixgf60_), goals scored (gs60_), assists (a60_), giveaways (igv60_), takeaways (itka60_), Corsi for factor (icf60_), and Fenwick for factor (iff60_). The Corsi factor is a blended statistic of the number of all shots (goals, saves, misses, and blocks) the team creates with a player on the ice while the Fenwick factor accounts for all unblocked shots (goals, saves, and

misses) the team creates with a player on the ice. Figure 2 shows the average statistics for all performance variables in 5 on 5 play. 4 variables, the Corsi factor, Fenwick factor, hits, and blocks have the highest performance statistics average among the potential 10 performance variables.

FIGURE 2. AVERAGE COEFFICIENTS FOR 5V5, PP, AND PK PERFORMANCE STATISTICS



Two dummy variables were created to represent players who may not play in power plays or penalty kills. In doing so, this could signify the difference in salary for a player who specializes in playing a certain game situation. Roughly 60% of the players in the data set do not play in power plays while 40% of the payers do not play in penalty kills.

A new variable, “logcaphit”, was created to take the natural logarithm of salary. This was done in order to eliminate the outliers within the data in order to improve the fit of the model, given the possible range of salary. A fixed effect placed on the team that a player belongs was used. Every team has different budgets or salary caps and therefore have different values for defensemen. The fixed effect shows the individual salary increase or decrease belonging to a particular team compared to the Anaheim Ducks.

FIGURE 3. TEAM FIXED EFFECT, COMPARED TO THE ANAHEIM DUCKS EXPRESSED AS A COEFFICIENT $100(e^{B1} - 1)\%$ CHANGE IN SALARY

Team	*Fixed Effect*	*Team*	*Fixed Effect*
Arizona Coyotes	0.3340	*Nashville Predators*	0.7388
Boston Bruins	1.0535	*New Jersey Devils*	0.2592
Buffalo Sabres	0.4385	*New York Islanders*	0.3819
Calgary Flames	0.3735	*New York Rangers*	0.7161
Carolina Hurricanes	0.7310	*Ottawa Senators*	-0.0237
Chicago Blackhawks	0.7474	*Philadelphia Flyers*	0.8906
Colorado Avalanche	0.1861	*Pittsburgh Penguins*	0.7169
Columbus Blue Jackets	0.3769	*San Jose Sharks*	-0.2587
Dallas Stars	0.0782	*St Louis Blues*	0.3347
Detroit Red Wings	0.7009	*Tampa Bay Lightning*	0.9008
Edmonton Oilers	1.1305	*Toronto Maple Leafs*	0.4079
Florida Panthers	0.4436	*Vancouver Canucks*	0.3711
Los Angeles Kings	-0.0162	*Vegas Golden Knights*	0.1324
Minnesota Wild	0.3258	*Washington Capitals*	0.7759
Montreal Canadiens	0.6226	*Winnipeg Jets*	0.3383

IV. Model

The regression methodology I will be using for my project is a log-linear model (as shown in Figure 4). Linear prediction theory aims to identify the optimal least-squares predictor in which the model, on average, yields a state with the smallest (squared) prediction error to get a B.L.U.E. or best linear unbiased estimator (Smith, 2015). Since the data is cross-sectional and I am trying to predict or estimate the salary of a defensemen, the log-linear model allows the quality of their performance to be independent variables.

FIGURE 4. LOG LINEAR REGRESSION EQUATION OF THE MODEL

$$Log(Y) = \beta_1 + \beta_2 itka60\_5v5 + \beta_3 igva60\_5v5 + \beta_4 icf60\_5v5 + \beta_5 iff60\_5v5 + \beta_6 GS60\_5v5 + \beta_7 ixgf60\_5v5 + \beta_8 a60\_5v5 + \beta_9 itka60\_PK + \beta_{10} igva60\_PK + \beta_{11} icf60\_PK + \beta_{12} iff60\_PK + \beta_{13} GS60\_PK + \beta_{14} ixgf60\_PK + \beta_{15} a60\_PK + \beta_{16} itka60\_PP + \beta_{17} igva60\_PP + \beta_{18} icf60\_PP + \beta_{19} iff60\_PP + \beta_{20} GS60\_PP + \beta_{21} ixgf60\_PP + \beta_{22} a60\_PP + \beta_{23} ihf60\_5v5 + \beta_{24} iblk60\_5v5 + \beta_{25} ipend60\_5v5 + \beta_{26} ihf60\_PK + \beta_{27} iblk60\_PK + \beta_{28} ipend60\_PK + \beta_{29} ihf60\_PP + \beta_{30} iblk60\_PP + \beta_{31} ipend60\_PP + \beta_{32}\, noPP + \beta_{33} noPK + \varepsilon$$

This is important in analyzing the salary for defensemen because the salary will vary based on differences in performance despite them all playing the same position. For a one unit change in the X variable or the performance variable, it will reflect the coefficient percentage change of the Y variable or salary variable. That is, the coefficient of X variables will either reward or demerit the value Y should be and the reason behind taking the logarithm of the salary is to reduce the outliers and improve the fit of the model given the large range in possible salary.

There are two advantages of using a log-linear model to analyze data. “The first is the ability to determine the relative influence of one or more predictor variables to the criterion value. The second advantage is the ability to identify outliers, or anomalies” (Weedmark, 2018). If we examine the R-squared value, it’s telling of how well the model fits the data; conceptually, it measures variation in the response measure explained by the model as a percentage of the total variation in the response measure.

FIGURE 5. TWO-WAY SCATTERPLOT OF PREDICTED LOGCAPHIT TO LOGCAPHIT

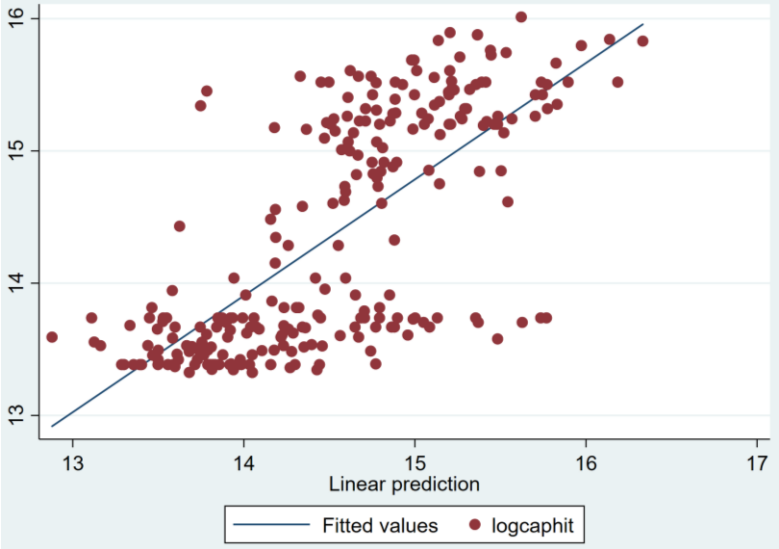


Figure 5 illustrates the predicted salaries of the model expressed as a logarithm, and how current salaries compare to the best fit line. The scatterplot shows that there are several plots that are very close or appear to be right on the line. This indicates that points above the line may be overestimates of salary while underestimates are represented below the line.

However, the limitations to this model is the nature of a predictive model. That is, there is a presence of omitted variable bias due to the fact that we cannot prove causality. While the R-squared is a source that can validate the model, it cannot determine whether the coefficient estimates and predictions are biased ("Albert.io", 2016). The problem of omitted variables arises because either the effect of the omitted variable on the dependent variable is unknown or because the data is not available ("Albert.io", 2016). By the Frisch-Waugh theorem, "coefficients become less accurate as they are more closely correlated with the omitted variables or more accurate when they are correlated with the

omitted variable, conditional on the other variables in the model" (Graves, 2018). This results in over-estimating or under-estimating the effect of one of more other explanatory variables. "Overfitting or overlearning is a condition in which the accuracy of a model is much higher on its training data set than on an independent data set." (Brownlee, 2016). To correct for overfitting the model, regularized regressions were executed and will be discussed more in depth in discussion section of this paper.

V. Results

The results of the log-linear regression produced in STATA are somewhat unexpected. For an increase in turnover performances such as giveaways and penalties drawn, I expect a decrease in the value of salary. Similarly, for an increase in defensive or goal scoring performances such as blocks or takeaways, I expect an increase in the value of salary. Figure 6 interprets the performance variables in relation to salary, expressed as percentages. The coefficients of the dummy variables noPP and noPK are interpreted as a $100(e^{\beta 1} - 1)\%$ change in salary.

FIGURE 6. EXPLANATORY VARIABLE COEFFICIENTS EXPRESSED AS A LOGARITHM TO SALARY WITH REPORTED STANDARD ERROR

Variable	*5v5*	*Std. Err.*	*PP*	*Std. Err.*	*PK*	*Std. Err.*		
icf60	0.0080	0.070	0.0414	0.040	-0.3686**	0.183		
iff60	-0.0448	0.112	-0.0873	0.071	0.3734%*	0.209		
ixgf60	-0.8520	1.613	0.7842	0.729	-0.3634	1.647		
igva60	0.0604	0.982	0.0336	0.093	0.0476	0.072		
itka60	-0.0955	0.144	0.0671	0.133	0.0297	0.074		
ihf60	-0.0231	0.030	0.1961	0.134	-0.0026	0.040		
gs60	-0.0451	0.155	0.1044	0.217	-0.2566	0.683		
iblk60	0.0394	0.057	0.1441	0.163	-0.0566	0.042		
ipend60	0.0159	0.252	0.1190	0.430	-0.0142	0.174		$R^2 = 0.6413$
a60	-0.0712	0.252	-0.0706	0.153	-0.1937	0.437		Adj. $R^2 = 0.4729$
noPP							-0.1233	
noPK							-1.5374	
constant							15.25742	

Note: p<0.10*, p<0.05**, p<0.01***

The model identifies the following variables most significant to the increase in players' salary by a percentage for every 1 unit increase in a particular performance: Fenwick for in penalty kills (37%), hits for (20%), blocks (14%), and penalties drawn (12%) in power play, and giveaways (6%) in 5 on 5 play. The Fenwick factor, hits, blocks, and penalties drawn are expected to have positive effects on defensemen's salaries. However, giveaways is considered a negative performance statistic, yet is shown to increase in the data. This could be due to omitted variables such as passes. Giveaways may be highly correlated with passes and the ability for teammates to make complete passes as opposed to just giving away the puck.

The variables that are most significant to the decrease in players' salary by a percentage for every one unit increase in a particular performance are: the probability of a goal scored (-85%) and takeaways (-10%) in 5 on 5 play and goals scored (-26%), total assists (-19%), and Corsi factor (-37%) in penalty kills. For 5 on 5 play, the probability of goals scored, goals scored, and assists have a negative impact on salary. These results are telling of how current salaries for defensemen for 5 on 5 performance is rewarded for taking aggressive behaviour. However, if you are too offensively aggressive and focus too much on a role like goal scoring and less on protecting the net, you could be penalized for that action as illustrated by an 85% decline in salary as a player's probability of a goal being scored increases. Similar to the situation of giveaways, takeaways are expected to have a positive effect on salary since it is a gain in puck possession. However, it is possible that this is highly correlated with passing rates which is not included in this data set. Interestingly, for players who do not play in power plays,

there shows to be a 11.6% decrease in salary while players who do not play in penalty kills face a 78.5% decrease. This demonstrates that there is an increased value for players who specialize in playing uneven strength game situations, since all the players in the data play in 5 on 5 or an even strength game.

Defensive aggressiveness is rewarded in power play situations for defensemen, reflected in the positive salary influences for hits, blocks, and penalties drawn. For situations pertaining to penalty kills, the disadvantage of one less player on the ice turns the event to be highly defensive. Protecting the net is much harder and the chances of puck possession is less, this is reflected by the reward in the Fenwick factor and the demerit in the Corsi factor. For defensemen in a penalty kill situation, its observed that these players will clear the puck to the other side of the arena to run the penalty clock. Conducting offensive plays like scoring goals or making assists where there's a high risk of puck possession turnover is more likely is penalized in salary. The data shows the impact to salary as being negative when a player does not play in power plays or penalty kills. The R^2 at 64% is considered a good level of predictive power, but the adjusted R^2 is reported at 47%.

VI. Discussion

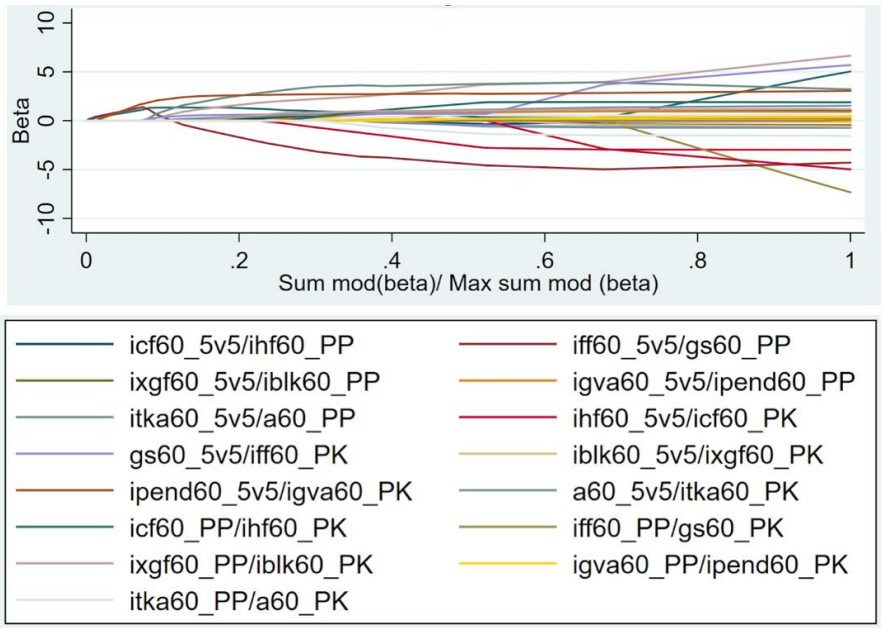
The model holds a relatively acceptable predictability level. The corrective measure to improve the model and seek a higher adjusted R^2 is to either add more variables or change the methodology. However, the risk of adding or eliminating variables could present omitted variable bias. A Ramsey RESET test was conducted to find that there was no evidence of omitted or non-liner variables present in the model,

yielding a F-statistic value of 0.000 with F(3, 129) =12.02, but this does not indicate that there are no omitted variables within the data. Omitted variables could be a result of a misspecification of a linear regression model where "the effect of the omitted variable on the dependent variable is unknown or because the data is not available. This forces the omission of that variable the regression which results in creating an upward or downward bias effect of one of more other explanatory variables" ("Albert.io", 2016). Furthermore, some the results in the model are difficult to interpret due to multicollinearity. Multicollinearity may be present in this model when independent variables are correlated and therefore "indicates that changes in one variable are associated with shifts in another variable" (Frost, 2017). Therefore, estimating the relationship between each performance variable individually to salary becomes difficult because the explanatory variables tend to change in unison (Frost, 2017). This may also be the reason for the low adjusted R^2 produced by the model.

To test for overfitting, 2 penalized regressions were conducted, as well as a robust check to compare the results. The Lasso regression is a shrinkage and variable selection method for linear regression models. It aims to obtain the subset of predictors that minimizes prediction error for a quantitative response variable by imposing a constraint on the model parameters that causes regression coefficients for some variables to shrink toward zero ("Coursera", 2018). Variables with a coefficient of zero are excluded from the model, meaning that variables with non-zero regression coefficients variables are most strongly associated with the response variable ("Coursera", 2018). The Elastic Net regression is similar to the Lasso regression, but accounts for the possibility of the

variable selection if it is dependent on data and thus unstable ("Coursera", 2018). The Lasso regression is used as a comparison to the result of the log-linear model, with the exception that the dummies were not included and instead assigned as a "missing observation". It is visually represented by the graph and shows pairs of explanatory variables that hold significance. Figure 7 shows a stage-wise plot of the Lasso regression results, which pairs together coefficients that are identified to have significant coefficient values.

FIGURE 7. LASSO REGRESSION STAGE-WISE PLOT, PAIRED COEFFICIENTS WITH HIGH SIGNIFICANCE



The Elastic Net regression held similar results to the Lasso regression as one would expect. The R^2 value of the elastic net regression and robust check produced merely identical values, showing 50% predictive power and a 43% cross validation mean squared error. From the coefficients of the lasso and elastic net, it is once again emphasized that power play and penalty kill performance is most impactful on salary as

opposed to 5 on 5 performance. The coefficients and the values that were significant to the Elastic Net regression is presented in Figure 8.

FIGURE 8. MOST INFLUENTIAL EXPLANATORY VARIABLE COEFFICIENTS INTERPRETED AS CHANGES TO SALARY, AFTER REGULARIZED (ELASTIC NET) REGRESSION TESTING

Variable	*Coefficient Value*	
ihf60_5v5	-3.1%	R^2=50.4%
icf60_PP	0.5%	Adj. R^2=50.0%
ixgf60_PP	39.3%	Cross Validation MSE=42.6%
igva60_PP	1.3%	
itka60_PP	4.0%	
ihf60_PP	9.0%	
gs60_PP	-14.3%	
iblk60_PP	10.0%	
ipend60_PP	30.2%	
a60_PP	13.8%	
iff60_PK	3.6%	
igva60_PK	14.6%	
itka60_PK	5.1%	
ihf60_PK	2.7%	
iblk60_PK	1.5%	
a60_PK	-12.4%	

In addition, a regularized robust check was performed and confirms a 50% R^2 score. This model does not account for intangible assets of a player. Intangible factors could have impact on individual ice performance which are not monitored in performance statistics such as attitude, leadership, or how they treat their teammates. However, quantifying this asset is subjective and therefore difficult.

This model was also tested to examine if similar results would be produced if the sample consists of entry-level or "rookie" players. The sample stands at 64 players and yield a R^2 at 90% and only picked the expected probability of goals in power play (ixgf_60PP) to be significant. While this seems pleasantly optimistic, the adjusted R^2 is presented at -59%, yet passes the Ramsey RESET test. This shows that in the case of

rookies, this model may have unnecessary variables in the model or that the data faces issues with collinearity. Rookies have different performance opportunities than standard players, as well as a dissimilar salary evaluation. Because rookies are easily substituted, the data or the variables selected may not follow this economic intuition. Therefore, a new model should be created to examine the effects of performance to salary.

It is crucial to acknowledge that this model only looks at on-ice performance statistics, a single but important component for what contributes to the value of NHL salaries. Re-examining the theory of the firm for NHL teams, winning games is the main source of revenue for teams, but it could also be the player's popularity with fans which generates revenue in other ways like putting seats in the arena or selling jerseys (Sumo, 2013). The question this model tries to speak to for players is if the salary you are given proportional to the demand for your talent or the performance you execute on the ice. The current problem seems to be that "hockey executives often make mistakes by rewarding their own players with lucrative contract extensions incommensurate with their true value, or by paying a hefty premium on the free-agent market" (Dayal, 2018). As mentioned in Depken and Lureman's study, this may be rational behaviour to consider in terms of using efficiency wages as an incentive to motivate players to perform at their best. The model I have produced show that certain performances influence salary more than others for defensemen and that the most valuable monetary skills lie in a player's penalty kill abilities. Furthermore, performance statistics associated with defensively aggressive behaviour is most rewarded in salary while opposite affects occur for offensively aggressive behaviour. The data set containing in 195 defensemen is relatively

small, increasing the sample size as well as using time-series performance data could potentially strengthen the model.

VII. Conclusion

Given the nature of the sport, it is difficult to isolate individual performance in hockey. It is clear that the value of players based on plus/minus points can overvalue some players, while undervaluing others due to external factors such as opponent and team strengths as opposed to individual performance. The current literature pertaining to salary valuation based on individual performance statistics is very limited and mainly focuses on goal scoring statistics which don't necessarily highlight or value the role and skills of a defensemen in protecting the net. The Log-Linear model was chosen in order to predict the salary of a player based on their individual performance statistics, including even and uneven strength statistics. The model holds 50% predictive power, showing that performances related to power play, but more so to penalty kills, are quite significant to the value in salary. The challenge of this model is the presence of omitted variable bias, either the effect of the omitted variable on the dependent variable is unknown or because the data is not available. This results in over-estimating or under-estimating the effect of one of more other explanatory variables ("Albert.io", 2016). A Ramsey RESET test was conducted to find that there was no omitted variable bias present in the model. To test the model for outliers and robustness, a lasso regression and robust check was run which yielded similar results to the log linear model. Performances in power plays such as blocks, hits, and drawn penalties were most influential in increasing a player's salary. These results were interpreted to be associated with rewarding defensemen with

defensively aggressive behavior. However, it is important to remember the limitations of the model; not every attribute that quantifies salary was included in this model. One large factor that was not included is intangible factors such as a player's attitude or care towards his teammates which can influence the performance or outcome of a game. Regarding the model, improvements could be made by reducing omitted variable bias as well as acknowledging overfitting. For further investigation, different regression models could be utilized along with testing in different positions such as forwards in performance statistics for uneven and even strength game situations.

References

Albert.io. (2016, September 20). Omitted Variable Bias: A Comprehensive Econometrics Review. Retrieved December 15, 2018, from https://www.albert.io/blog/omitted-variable-bias-econometrics-review/

Badenhausen, K. (2017, December 5). The NHL's Highest-Paid Players 2017-18. Retrieved December 15, 2018, from https://www.forbes.com/sites/kurtbadenhausen/2017/12/05/the-nhls-highest-paid-players-2017-18/#7b9134eb2ac3

Brownlee, J. (2016, March 21). Overfitting and Underfitting With Machine Learning Algorithms. Retrieved December 15, 2018, from https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/

Coursera. (2018, March 13). What is Lasso Regression? Retrieved December 15, 2018, from https://www.coursera.org/lecture/machine-learning-data-analysis/what-is-lasso-regression-0KIy7

Dayal, H. (2018, September 19). These are the most overpaid NHL hockey players in 2018. Retrieved December 15, 2018, from http://dailyhive.com/vancouver/most-overpaid-nhl-players-2018?fbclid=IwAR0Xw86uhQ01rdH0DuY7QH7EJTyhl1dWDxFPRO_tHnoLkgSdOK-n07-MvuI

Depken, C. A., & Lureman, J. (2018). wage disparity, team performance, and the 2005 nhl collective bargaining agreement. *Contemporary Economic Policy*, *36*(1), 192-199. doi:10.1111/coep.12220

Frost, J. (2017, September 20). Multicollinearity in Regression Analysis: Problems, Detection, and Solutions. Retrieved December 15, 2018, from http://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

Gramacy, R. B., Jensen, S. T., & Taddy, M. (2013). Estimating player contribution in hockey with regularized logistic regression. *Journal of Quantitative Analysis in Sports*, *9*(1), 97-111. doi:10.1515/jqas-2012-0001

Graves, J. L. (2018, December 14). *Applied Economics - Lesson 4 - Causality, Residuals, and Omitted Variables.pptx*. Lecture presented at Additional Slides in University of British Columbia, Vancouver, BC. Retrieved December 15, 2018, from https://canvas.ubc.ca/courses/13252/files/2911772?module_item_id=839810

Smith, L. A. (2015). Data Assimilation and Predictability | Predictability and Chaos. Retrieved December 15, 2018, from https://www.sciencedirect.com/topics/earth-and-planetary-sciences/linear-prediction

Sumo, V. (2013, February 13). Are Hockey Players Overpaid? Retrieved December 15, 2018, from https://newschicagobooth.uchicago.edu/about/newsroom/news/2013/2013-02-13-hockey

Weedmark, D. (2018, March 13). The Advantages & Disadvantages of a Multiple Regression Model. Retrieved December 15, 2018, from https://sciencing.com/advantages-disadvantages-multiple-regression-model-12070171.html

Data Appendix

*Submitted as a separate file on Canvas