Predicting NHL Player Salary using Multiple Linear Regression

By Ian Keller and Max Bauer

Introduction:

Over recent years, organizations across all sports have taken a deeper look into statistics to help evaluate players. This has resulted in shifts in how teams evaluate, sign, and compensate players, and the National Hockey League (NHL) is no different. As the league operates within the salary cap era, player salaries have become a primary focus for both management and fans. Our research takes a deeper dive into this world of NHL player salaries.

Firstly, the goal of this paper is to identify which game and player statistics have the greatest impact in determining the salary of an NHL player. Hockey is a sport that includes many factors for player performance and understanding which factors are most influential is key to understanding a player's earning potential. Furthermore, we want to determine the best model in relation to predicting player salary. This will help us understand what factors are significant in predicting salary. Specifically, we hypothesize that time on ice per game, draft position, years since draft, skater position, and points per game will be significant for predicting player salaries.

We believe that these statistics are indicative of player performance and therefore will relate strongly to salary. Time on ice per game and points per game represent a player's value as better players will be more trusted with more ice time and by no accident score more points each game, providing their team with value. Looking at draft positions we feel that earlier draft picks will produce more points, and therefore make more in salary, since they were highly scouted and teams selected these individuals with valuable draft picks.

We will examine if compensation is indicative of performance, and conversely, if lower-performing players are paid less. These results can be used to help explore techniques reminiscent of Moneyball, seeking to identify and acquire undervalued players whose performance may surpass their salary-based expectations. This can be done through a residual analysis of our final model.

The overall purpose of the study is to provide valuable insight into player salaries and the constraints of the salary cap era. The model could not only help aid in predicting salaries but also give teams a tool to optimize their roster. In order to construct a winning team, one must have a deep understanding of player compensation. Our study can help teams make data-driven decisions and potentially revolutionize the way players are valued in the league.

Data Collection (Methodology)

The dataset was collected as part of an observational study that focuses on the 2022-2023 NHL regular season. The data includes various performance metrics of players, specifically skaters. Goalies were excluded from the dataset. The sample collected includes NHL players who played at least one game during the season. However, to be included in the study, players must have played a minimum of 41 games during the 2022-2023 NHL season. This ensures players being observed have a meaningful presence in the NHL. Also, we only took a look at even strength statistics (5v5), excluding man up or man down. We decided to do this to standardize our player performance metrics (i.e. points per game, Corsi for percentage, et cetera.) and reduce the bias that special teams can have on player performance.

Publicly available NHL-related statistics were obtained from the websites NaturalStatTrick, FrozenPool, and Hockey-Reference, for the 2022-2023 NHL season. All of these websites incorporate data from the NHL database. For each player's salary during this season, their AAV (Average Annual Value), was obtained from FrozenPool.

The data consisted of three different datasets that needed to be merged based on player name. Originally there were 955 observations with 104 variables, but after the restrictions and cleaning of the data, this lessened to 788 observations with 55 variables. The data cleaning included removing duplicate variables and dealing with multiple observations for the same player. Some variables were excluded from the analysis, such as height, weight, nationality, etc. Additionally, we implemented feature engineering during the creation of the dataset. This was done to control for games played throughout the season–converting count statistics into rate statistics for equivalent analysis. Additionally, we engineered the feature years since the draft as we believe that entry-level contracts (the first 3 years of a player's career) would skew the data. Lastly, we performed a log transformation of the Salary variable to help normalize the variable and reduce homoscedasticity. We considered the following 41 variables included in the data analysis procedure:

- Point Per Game (PPG) - Number of points score per game (Points scored / Game Played)
- Corsi For (CF) - Quantifies the total number of shot attempts (on goal, missed, or blocked) a player generates while on the ice.
- Fenwick For (FF) - Fenwick for also measures shot attempts, but excludes blocked shots.

- Shots For (SF) - Represents the total number of shots on goal taken by the team while the player is on the ice.
- Goals For (GF) - GF represents goals scored by a team when the player is on the ice.
- Expected Goals For (xGF) - Represents the probability of a shot resulting in a goal. This takes into account factors such as shot location, shot type, and other elements.
- Scoring Chances For (SCF) - Number of scoring chances generated by a team while a player is on the ice.
- High-Danger Goals For (HDGF) - Goals scored by a team when a player is on the ice in high-danger situations.
- Medium-Danger Goals For (MDGF) - Goals scored by a team when a player is on the ice in medium-danger situations.
- Low-Danger Goals For (LDGF) - Goals scored by a team when a player is on the ice in low-danger situations.
- PDO - Sum of OnIceSV and OnIceSH.
- Individual Points Percentage (IPP) - Percentage of goals and assists where a player is directly responsible.
- Shots - Number of shots on goal by a player.
- Rush Attempts - The number of offensive rushes initiated by a player.
- Rebounds Created - Number of scoring chances created by a player through rebounds.
- Penalties in Minutes (PIM) - Total number of penalty minutes by a player.
- Penalties Drawn - Number of penalties player draws from opponents.
- Giveaways - Instances where a player loses possession of puck to the opponent.
- Takeaways - Number of times a player steals the puck from an opponent.
- Hits - Number of checks or body contact by player.
- Shots Blocked - Number of opposing shots a player blocks.
- Goals Per Game (GPG) - Average number of goals per game by a player.
- Time on Ice Per Game (TOIpg) - Average time a player spends on the ice per game
- Draft Round - Round number that player was drafted.
- Overall Draft Position - The overall draft position a player was selected.
- Position (Offense_Defense) - Players position broken down by defense (1) and offense (0).

Descriptive Statistics:

First we looked at our response variable, salary. Figure 1 shows the five-number summary of salary. The data had a min of $730,000, a lower quartile of $860,000, a median of $2,975,000, a mean of $3,602,565, an upper quartile of $5,500,000, and a max of $12,500,000. Furthermore, visualizing the data we can see the right-skewed distribution as shown by the histogram in Figure 2. The skewness can also be seen by the difference in median and mean. The data was then transformed using a log transformation. We decided to go with this transformation to help mitigate the right-skewness and stabilize the variance. The transformation resulted in a more symmetric distribution, as shown in Figure 3. The log-transformed data had a min of 13.5, a lower quartile of 13.66, a median of 14.29, a mean value of 14.50, an upper quartile of 15.34, and a max of 16.34.

Figure 1: Five Number Summary Statistics of the Salary Variable

Min	***Q1***	***Median***	***Mean***	***Q3***	***Max***
$730,000	$1,200,000	$2,975,000	$3,602,565	$5,500,000	$12,500,000

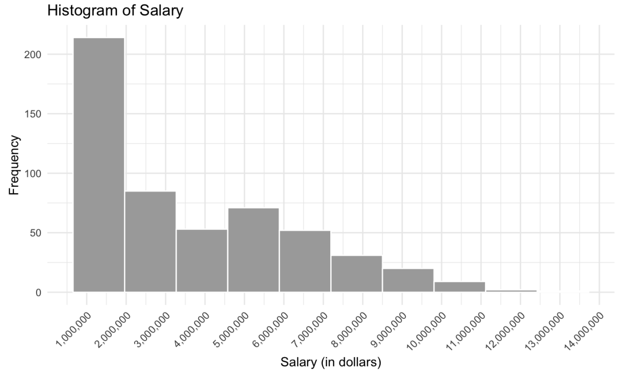


Figure 2: Histogram of the Salary response variable.

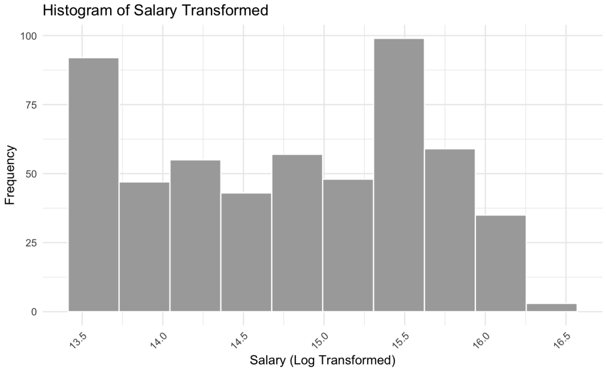


Figure 3: Histogram of the Salary response after log transform.

One explanation for this right skewness can be based on entry-level NHL contracts which are signed at the beginning of an NHL player's career and last for three seasons. To control for this we created a “years since drafted” statistic (see more in the “Inferential Statistics” section.)

Next, we decided on the most common player performance metric of Points per game as it relates to a player's salary. The scatterplot shown in Figure 4 shows a moderately strong correlation between the two variables. Running a regression on these two variables confirms our hypothesis that Points Per Game (PPG) has a significant impact on player's salary.

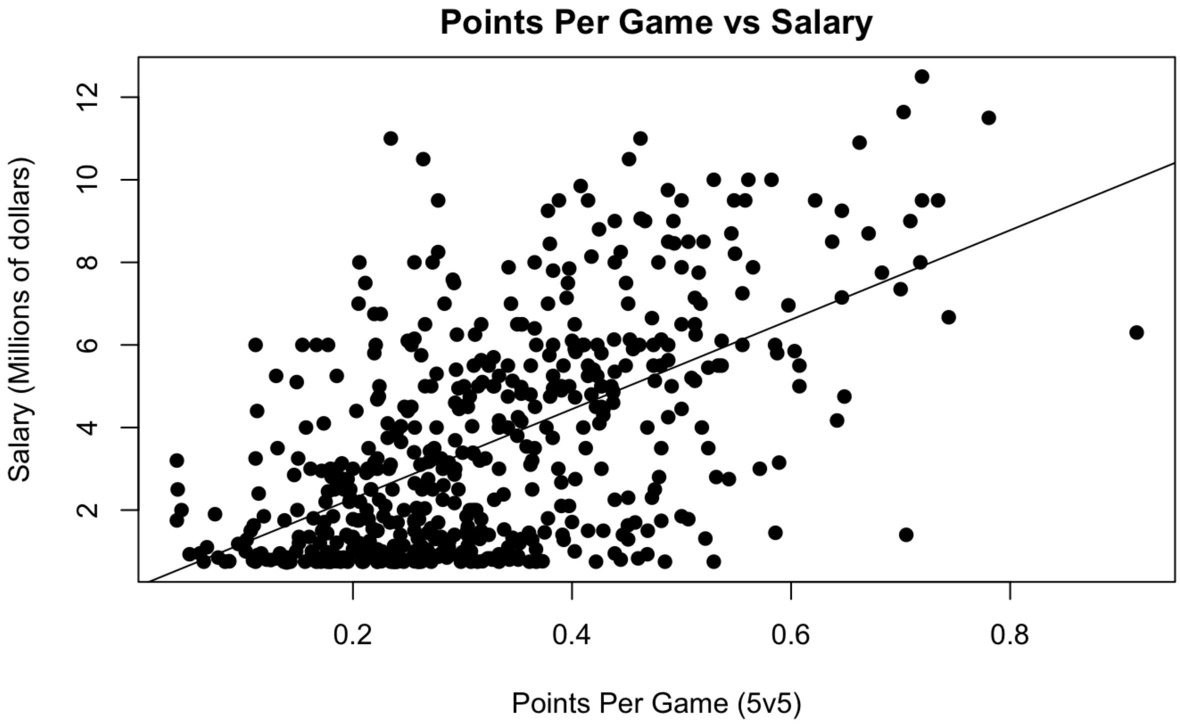


Figure 2: Scatterplot comparing Points Per Game at 5v5 and Salary in millions. This regression resulted in a p-value of <0.0001 and multiple r-squared of 0.332.

In this regression, it is important to note that the slope coefficient is 10.8. This means that for every additional PPG a player has, their salary has an increase of 10.8 million. Again, this shows that PPG is positively and significantly associated with player salaries.

Below, in Figure 3, we decided to take a look at the correlation matrix for our hypothesis variables (Salary, TOIpg, Overall Draft Position, PPG, Offense/Defense, and Years Since Draft). There was only one strong correlation, which was TOIpg and Offense/Defense (r = 0.73). There were three moderately strong correlations, two positive and one negative. They were TOIpg and Salary (r = 0.44), PPG and Salary (0.58), and PPG and Offense/Defense (r = -0.39).

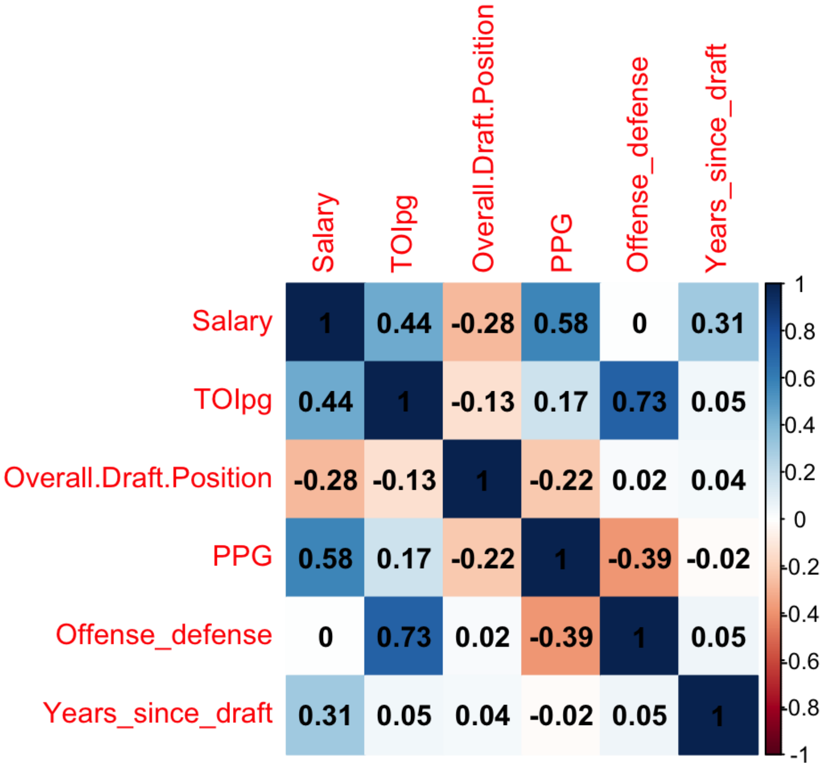
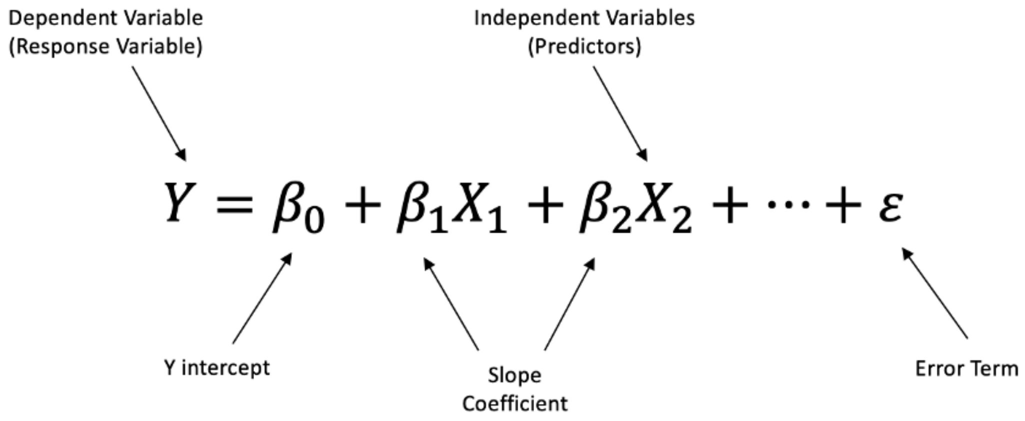


Figure 3: Correlation Matrix of our hypothesized significant predictor variables.

Inferential Statistics:

Next, we attempt to create a multiple linear regression model to predict player salary. The multiple linear regression model is modeled by the following equation:



Here, y is our response variable of salary, b0 is our intercept, b1x1+b2x2+...BiXi represents our predictor variables, and e is our error term. We began by fitting a model with all of our predictors.

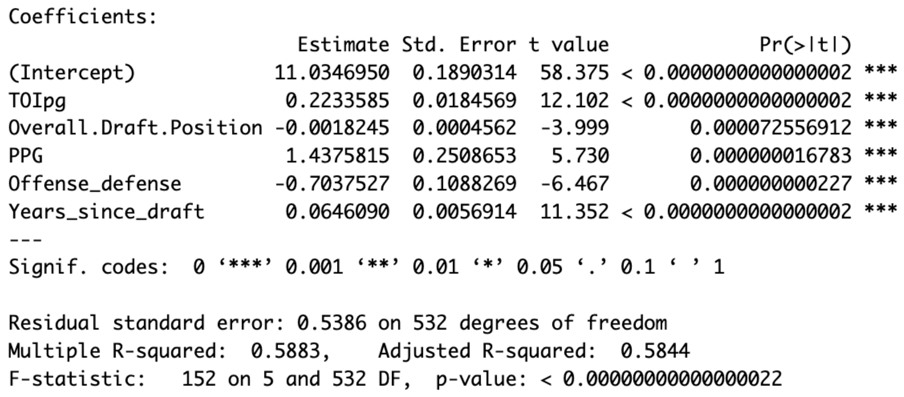
The summary of the full model is given below:

	Estimate	**Std. Error**	**t-value**	**P-value**
(Intercept)	18.503	8.180	2.262	0.024
PPG	2.1626549	0.5118167	4.225	2.83e-05
CF	-0.0480720	0.0330589	-1.454	0.146525
FF	0.0452693	0.0397783	1.138	0.255639
SF	-0.0246659	0.0359223	-0.687	0.492619
xGF	-0.0157515	0.0179546	-0.877	0.380740
SCF	0.0253065	0.0193030	1.311	0.190443
PDO	-7.6952564	8.0129418	-0.960	0.337332
Draft Round	0.2788991	0.0803055	3.473	0.000558
Overall Draft Position	-0.0103793	0.0025870	-4.012	6.92e-05
IPP	-0.0027764	0.0032821	-0.846	0.398002

Shots	0.0042071	0.0012495	3.367	0.000817
Rush Attempts	-0.0052881	0.0039461	-1.340	0.180816
Rebounds Created	-0.0026353	0.0051311	-0.514	0.607763
PIM	-0.0004852	0.0015147	-0.320	0.748847
Penalties Drawn	-0.0021967	0.0047550	-0.462	0.644289
Giveaways	0.0039110	0.0024618	1.589	0.112750
Takeaways	-0.0010196	0.0026029	-0.392	0.695442
Hits	-0.0006090	0.0005481	-1.111	-1.111
Shots Blocked	-0.0004256	0.0015566	-0.273	0.784646
GPG	-2.2346904	0.5906065	-3.784	0.000173
TOIpg	0.1726732	0.0231585	7.456	3.85e-13
Years Since Draft	0.0631072	0.0057030	11.066	< 2e-16
Offense/Defense	-0.7079339	0.1438934	-4.920	1.17e-06
GF	0.0297098	0.0357962	0.830	0.406943
HDGF	-0.0092486	0.0115856	-0.798	0.425077
MDGF	-0.0029591	0.0064721	-0.457	0,647710
LDGF	-0.0007670	0.0029895	-0.257	0.797620

Looking at the table we can determine that there are 8 significant variables. The variables included: PPG (p-value of 2.83e-05), Draft Round (p-value of .000558), Overall Draft Position (p-value of 6.92e-05), Shots (p-value of .000817), TOIpg (p-value of 3.85e-13), GPG (p-value of 0.000173), Years Since Draft (p-value of 2e-16), and Offense/Defense (p-value of 1.17e-06). Additionally, this model has an adjusted R-squared of 0.6099 and a p-value of 2.2 e-16. The residual standard error is 0.5218, which does not have a meaningful interpretation due to our log transformation.

Next, after fitting the full model we wanted to test if our hypothesized predictors were significant. Figure 4 represents the model's summary output.



```
Coefficients:
                        Estimate Std. Error t value            Pr(>|t|)    
(Intercept)           11.0346950  0.1890314  58.375 < 0.0000000000000002 ***
TOIpg                  0.2233585  0.0184569  12.102 < 0.0000000000000002 ***
Overall.Draft.Position -0.0018245  0.0004562  -3.999       0.000072556912 ***
PPG                    1.4375815  0.2508653   5.730       0.000000016783 ***
Offense_defense       -0.7037527  0.1088269  -6.467       0.000000000227 ***
Years_since_draft      0.0646090  0.0056914  11.352 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5386 on 532 degrees of freedom
Multiple R-squared:  0.5883,	Adjusted R-squared:  0.5844 
F-statistic:   152 on 5 and 532 DF,  p-value: < 0.00000000000000022
```

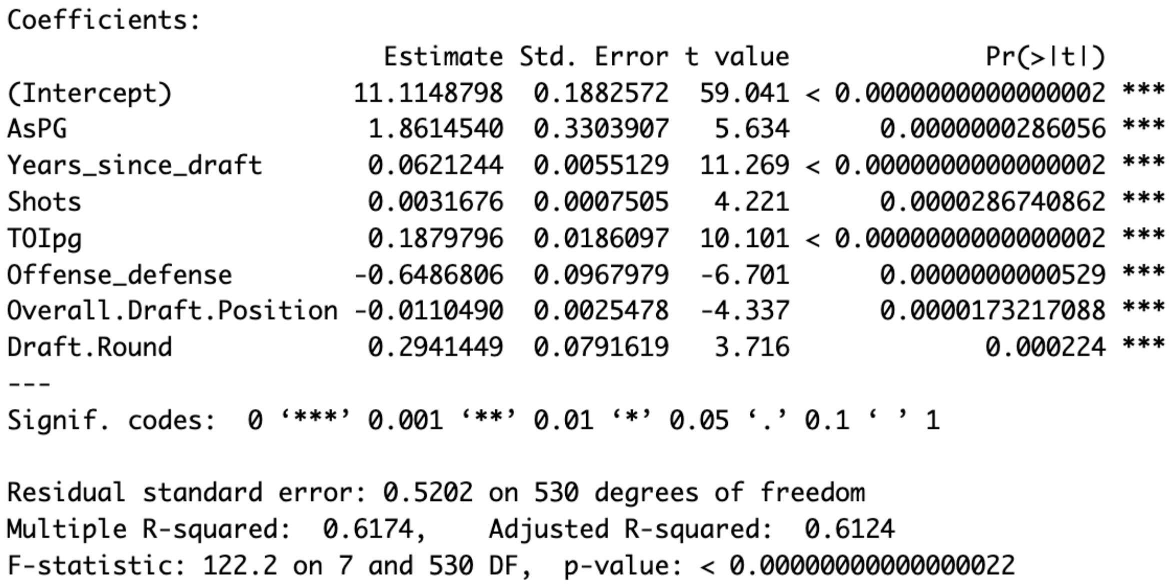
Figure 4: Our hypothesized best model regression output.

Here we confirm our hypothesis that the predictors' Time on ice per game, Draft position, Points per game, Player Position, and years since drafted were all significant in predicting salary. The model outputted an F-stat of 152 on 5,532 df which when tested at a 95% confidence level can be shown that the model is significant. The adjusted r-squared of 0.5844 is a slight decrease over the full model but since we only use 5 predictors this model may be more convenient.

Lastly, we wanted to find the best possible multiple regression model using stepwise model selection comparing the different model's Bayesian Information Criterion (BIC). BIC is a statistical measure used to evaluate and compare the fit of different statistical models, taking into account both the likelihood of the data given the model and the complexity of the model. BIC is important because it helps to balance model fit and complexity, aiming to prevent overfitting by penalizing models with too many parameters, making BIC a better criterion for model selection when compared to AIC.

The best model is as follows:

Salary = 11.11 + 1.86(AsPG) + 0.06(Years_since_draft) + 0.003 (Shots) + 0.19 (TOIpg) - .65 (Offense_defense) - 0.01 (Overall.Draft.Position) + 0.29 (Draft.round) + Error



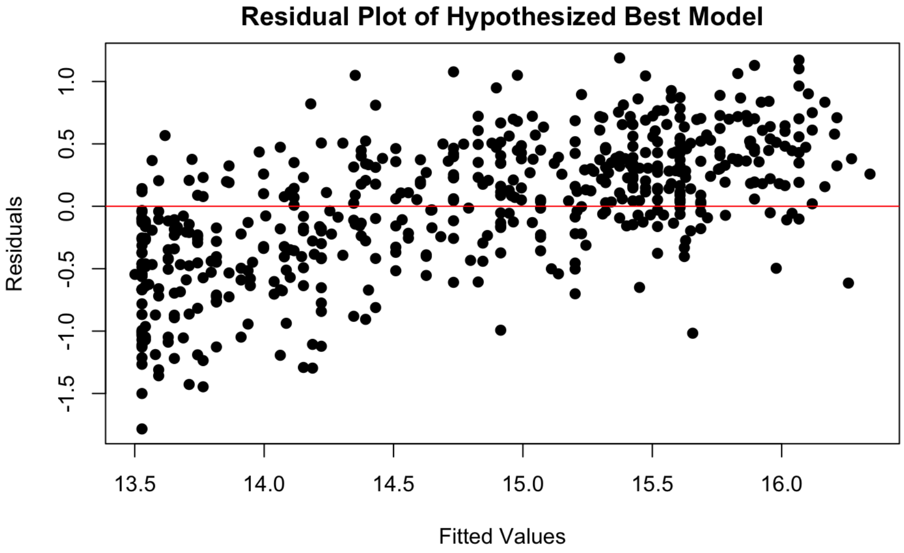
```
Coefficients:
                         Estimate Std. Error t value            Pr(>|t|)    
(Intercept)            11.1148798  0.1882572  59.041 < 0.0000000000000002 ***
AsPG                    1.8614540  0.3303907   5.634      0.0000000286056 ***
Years_since_draft       0.0621244  0.0055129  11.269 < 0.0000000000000002 ***
Shots                   0.0031676  0.0007505   4.221      0.0000286740862 ***
TOIpg                   0.1879796  0.0186097  10.101 < 0.0000000000000002 ***
Offense_defense        -0.6486806  0.0967979  -6.701      0.0000000000529 ***
Overall.Draft.Position -0.0110490  0.0025478  -4.337      0.0000173217088 ***
Draft.Round             0.2941449  0.0791619   3.716             0.000224 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5202 on 530 degrees of freedom
Multiple R-squared:  0.6174,	Adjusted R-squared:  0.6124 
F-statistic: 122.2 on 7 and 530 DF,  p-value: < 0.00000000000000022
```

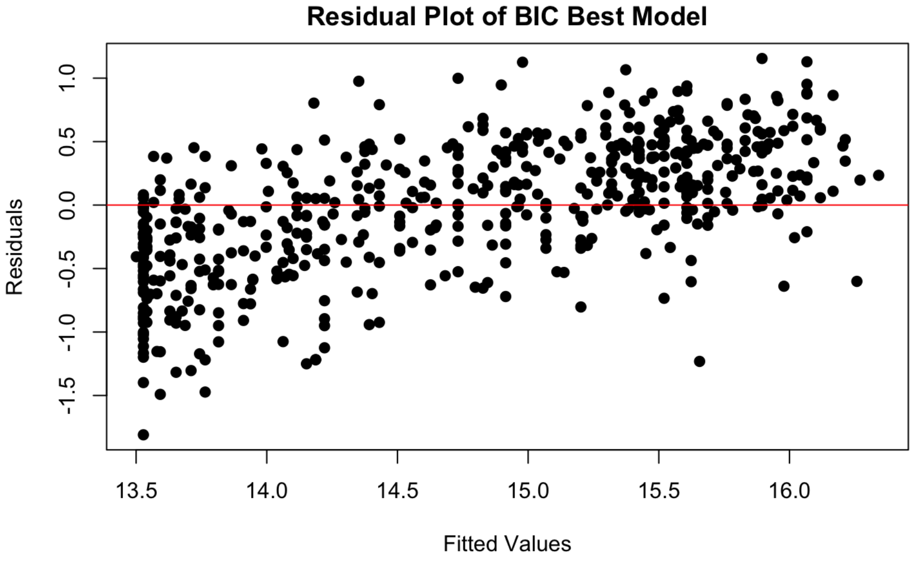
Figure 5: BIC Best Model Output Summary

The BIC for this model was -661.04, reduced from the min model’s BIC of AIC=-188.12 using stepwise model selection. Here we have a very concise model in which all predictors (and the model as a whole) are highly significant at the 99% confidence level. Additionally, this model has an adjusted R-squared of 0.6124 which is the highest of all of our models. Using Adjusted R-squared as a metric for model performance is important because it penalizes models for using too many predictors. Lastly, the model has a residual standard error of only 0.5202 which is lower than the max model, despite using fewer predictors.

Furthermore, when looking at the model’s performance it is essential to look at the residual charts. Residual charts are crucial for diagnosing potential issues with a statistical model. They help to identify patterns, heteroscedasticity, outliers, or other anomalies that might indicate violations of model assumptions.

Below are the residual charts of the best BIC model and our hypothesized best model. Both residual charts are very similar. We see a slight breach of homoscedasticity as the residuals slope up and to the right which could've been expected due to the skewness of our response variable.





Discussion

In our analysis, we found several variables to be significant in predicting player salary. These included PPG, Draft Round, Overall Draft Position, Shots, TOIpg, GPG, Years SInce Draft, and Offense/Defense. The significance of the draft round and overall position is expected, as this is a good representation of a player's potential. Shots, TOIpg, and GPG were expected, as these are indicative of player performance. Years Since Draft also is understandable as it recognizes a player's experience.

We were shocked to find out that stats such as xGF, GF, and CF were not significant, as these all represent a player's offensive performance. This may indicate some non-linear relationships.

The best model to predict player salaries was the model including AsPG, Years SInce Draft, Shots, TOIpg, Offense/Defense, Overall Draft Position, and Draft Round. The model had an F-stat of 122.2 and an Adjusted R-squared value of 0.6124. This indicates that 61.24% of the variance in player salaries is determined by the model and that the overall model is statistically significant. This model was selected on the criteria of BIC. We decided to choose BIC due to the complexity of our model.

Our model can help all fields in hockey help determine a player's worth when negotiating. Players can identify what aspects of their game are significant in earning a higher salary, while front offices can use these findings to help build competitive teams. It can potentially be used to help other sports by looking at draft rounds and positions, however, most stats are unable to transfer to other sports. Although, this model would be useful to other hockey leagues.

It is worth noting that we did encounter a few limitations to this study. One problem was struggling to transform the salary back to its original form. This would have helped us give a better interpretation of the salaries. Another issue we encountered was the all-subsets regression in R. We wanted to find the best model based on adjusted R-squared and Mallows Cp using the leaps package, however, our model included too many variables to be evaluated. Additionally, once we reduced our predictor variable count we could not get the code to function properly, due to numerous errors. Lastly, we encountered trouble when scraping data from capfriendly as the website would only load exactly 300 rows of data which did not give us a complete picture of off-season signings. This forced us to change our discussion question to be based on only one year's worth of data.

For future research, we could take a deeper look into the interaction of variables and take a deeper look into player stats such as career milestones or salary fluctuations. Another idea for future research would be to use Naive Bayes or Lasso regression. Ultimately, we believe this model lays solid groundwork for predicting player salaries in the future.

Sources:

“Natural Stat Trick.” *Player Season Totals - Natural Stat Trick*, www.naturalstattrick.com/playerteams.php?fromseason=20222023&thruseason=

20222023&stype=2&sit=5v5&score=all&stdoi=oi&rate=n&team=ALL&pos=S&loc=B&toi=0&gpfilt=none&fd=&td=&tgp=410&lines=single&draftteam=ALL. Accessed 30 Oct. 2023.

“NHL Stats, History, Scores, Standings, Playoffs, Schedule & Records.” *Hockey*, www.hockey-reference.com/. Accessed 30 Oct. 2023.

“Report Generator.” *Frozen Tools*, frozenpool.dobbersports.com/frozenpool_report.php?pos=Skaters&team=ALL&mingp=1&rookie=All&period=2022-2023%3AR&startdate=&enddate=&report=Big%2BBoard&reportdata_length=30. Accessed 30 Oct. 2023.