League Adjusted Salary Model using Local Polynomial Regression

Shinwoo Kang
Development
Seattle Humane Society
Bellevue, USA
e-mail: chrisk@seattlehumane.org

Abstract **- Since the 2012 National Hockey League (NHL) Lockout, there have been many economic trends in the league that one might argue inconsistent. While many players' salaries were significantly altered as results of buy-outs or extravagant contract signings, the salary cap has fluctuated dramatically in the following years due to these chaotic activities. To understand the seemingly contradicting NHL economic trends, in this paper, we discuss League Adjusted Salary Model (LASM) applying Local Polynomial Regression Modeling to properly gauge a player's monetary vs. production feasibility value. The League Adjusted Salary Model is a approach that is dependent on a player's League-Relative Salary Percentages and his Individual Production. The League Relativity is emphasized to account for the different payrolls of all 30 NHL teams and to understand the year-by-year economic trend. The Individual Production is a user flexible element of the individual level model that can be improved with utilizations of "Enhanced Statistics" such as Unblocked Shot Attempt Relative Percentage values. Combining these two data sets, we apply the Local Polynomial Regression Modeling to compute the feasibility of cost and production.**

Keywords-hockey; Local Polynomial Regression; Economics; Salary Cap.

I. Introduction

After the new Collective Bargaining Agreement (CBA) in 2012, National Hockey League (NHL) teams were granted opportunities to buy-out players under contract. A record number of 26 players were bought out since June 23, 2013. Of the 26 players, only 16 remain in NHL at reduced salary (with a notable exception of Christian Ehrhoff). Unfortunately, the rate of reduction in salary is seemingly random. The sudden decrease in salaries for these players impacts the overall economy of the game. The new cap space acquired by the decrease in salaries allows (1) teams to sign more players, or (2) teams to re-sign players with a bump in salary. These two scenarios present difficulties in projecting salaries of other players based on performance.

Once a player's decrease or increase in salary can be put into the context of whole league, then we may establish a regression model that projects a player's upcoming salary, which we'll call "League Adjusted Salary Model." League Adjusted Salary Model employs Local Polynomial Regression Modeling to account for random noises and possibly misunderstood NHL contracts. League Adjusted Salary Model is an improvement from the simple linear regression on salary vs. performance, which is the traditional school thought in hockey analytics community [1].

The rest of the paper is organized as follows: Section 2 explains the methodology behind League-Relative Salary Percentage, League-Relative Cap Percentage, and League-Adjusted Salary Model. Section 3 describes the application of the model on training set data from the 2010~2011 NHL season to 2013~2014 NHL season. Section 4 concludes the paper with final remarks on the potential of the proposed model and possible improvements to it.

II. The League Adjusted Salary Model

The League Adjusted Salary Model is a two-part process, where the League-Relative Percentages (Salary and Cap) must be computed first. Then, a comparison of Linear Regression and Local Polynomial Regression Modeling is performed to provide a method that better fits the Cap and Salary. With the League Adjusted Salary Model, one may apply it for various purposes such as for determining the Expected Future Salary or possible statistical areas of improvement to maximize the salary potential. For this research, only the data for forwards and defensemen were considered, as goalies have independent valuation processes.

A. League-Relative Percentages

In order to compensate for the uncertainty level of buyouts, we introduce "League-Relative Salary Percentage" and "League-Relative Cap Percentage." The League-Relative Percentages ignore the unpredictability of contract buy-outs and re-signing, as one player moves from one team to another, the relative worth changes in respect to the particular team. The League-Relative Salary Percentage allows for low-market teams that are bounded by internal payroll amount**.** The League-Relative Salary Percentage is essentially a proportion of a player's True Salary/Cap to a sum of all NHL team's payrolls. We make a note that True Salary and Cap will be treated separately as explained further later.

If a player was bought-out, we create a rule to apply weighted average of salary/cap as the adjusted predictor in

 ISBN: 978-1-61208-423-7

respect to performance before and after the buy-out. Since it is difficult to measure if a player was initially overpaid and/or still overpaid after the buy-out. A striking example is of Scott Gomez who received the cap and salary of $7,357,143 and $7,500,00 in the 2011-2012 season, while receiving $700,00 for cap and salary, after the buy-out. He had .289 Points per Game (PPG) in 2011~2012 and .385 PPG in 2012~2013 season. By having the weighted average on production for the years a particular player was bough-out, it relaxes the noise it would be created in the ratio of "bought-out" cap/salary vs. production. The formulas for League Relative Cap and Salary (shortened for Sal) are,

$$LeagueRelCap\%_i = \frac{Cap_i}{\sum_{Team \in NHL} \sum_{player \in Team} Cap_{player}} \quad (1)$$

$$LeagueRelSal\%_i = \frac{Sal_i}{\sum_{Team \in NHL} \sum_{player \in Team} Sal_{player}} \quad (2)$$

where i indicates a particular player on a team.

The advantage of League-Relative Salary Percentage is that each NHL season is treated as an independent economy as a whole.

B. League Adjusted Salary Model

With League-Relative Percentages, we compare two methods: Linear Regression and Local Polynomial Regression Modeling on Production vs. League-Relative Percentages. The results of the comparison in Section III will show why linear regression is insufficient for modeling Salaries and Cap, and need a more flexible methods that is capable of modeling general nonlinear relationship [2].

For the predictors, we utilize Points Per 60 Minutes (P60), Offensive Zone Start Relative % (OZS%), Unblocked Shot Attempt Relative % (USAT Rel%), and Time on Ice (TOI), as they are the modern day go-to-metrics for evaluating a player's game, in addition to the two traditional statistics, Goals and Assists. The number of different metrics we compare may not be limited to these six. The general model for the linear regression may be represented as follows:

$$y = \alpha + \beta_i x_i + \varepsilon \quad (3)$$

where y is desired expected League-Relative Percentages. x_i is the training set of above predictors. For the Local Polynomial Regression, we use the traditional tri-cube kernel weights [3]:

$$\omega(x) = \begin{cases} (1 - |x^3|)^3 & for\ |x| < 1 \\ 0 & for\ |x| \geq 1 \end{cases} \quad (4)$$

III. CASE STUDY

In this section, we discuss the procedure of obtaining the proper NHL data, and correctly modeling it, by separating fixed and random effects.

A. Data Sources

The data sources for the two components of the proposed model are [4]–[9]. During the research, many data sources had to be aggregated and cross validated into a single database, since the industry leading [9] ceased its operation in 2014. For the League Relative Salary Percentage, we utilize statistics beginning with 2010~2011 season.

It must be noted that for the purpose of the analysis, we make a distinction between Cap and Salary, as they are indicators of their monetary compensation, but hold different meanings. These two numbers will be treated differently, as Cap Space, due to its nature, is uniform through the duration of the contract, while the true Salary usually changes from year to year and it may trend upwards or downwards, depending on age, and whether a player is entering his prime or not.

For production, we gather data exclusively from [4] and [5]. In addition to conventional statistics, such as Goals/60 and Assists/60, we utilize advanced shot metrics to compare across different linear regressions and Local Polynomial Regression. As previously stated, we utilize Points Per 60 Minutes (P60), Offensive Zone Start Relative % (OZS%), Shot Attempt Relative % (SAT%), and Time on Ice (TOI), as initial predictors because they give contextual clues to a player's game.

Combining the six data sources, we create one data frame to help compute League Adjusted Salary Model. The final data frame will include two extra columns of predictors in League Adjusted Cap Percentage and League Adjusted Salary Percentage. The initial plot of the two League Adjusted Percentages against USAT Rel% (Figures 1 and 2) shows that Salary and Cap have different spreads.

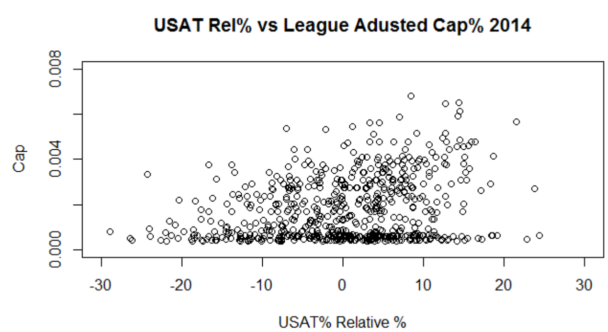

Figure 1. Disbrituion of League Adjusted Cap Percentage over USAT%

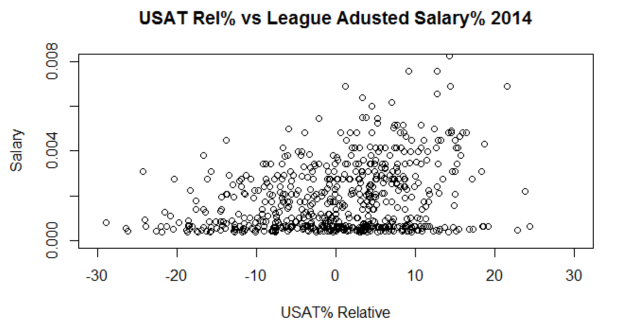

Figure 2. Disbrituion of League Adjusted Salary Percentage over USAT%

For this particular example in showing the difference in spreads, we use the 2014~2015 season data due to its availability in salary/cap information, but incompleteness in games played. The rest of the paper utilizes only 2010~2014 season data as training set for the model.

B. Applications of the Model and its Results

With the new data frame including League Adjusted Salary Percentage, we proceed with Linear Regression and Local Polynomial Regression on the proposed Enhanced Statistics. Utilizing R package, 'loess,' we compute the following results.

TABLE I. SCALED LEAGUE ADJUSTED CAP DISTRIBUTION

Min	*1Q*	*Median*	*3Q*	*Max*
.03752	.06099	.1365	.2739	.6823

TABLE II. SCALED LEAGUED ADJUSTED SALARY DISTRIBUTION

Min	*1Q*	*Median*	*3Q*	*Max*
.03785	.06194	.13760	.27530	.96350

TABLE III. LINEAR REGRESSION VS LOESS CAP

	LR Coeff Estimate	*LR Std. Error.*	*LRR^2*	*Loess Std Error*
G60	4.297e-04	8.680e-05	.03625	1.344e-04
A60	3.973e-04	6.033e-05	.06498	1.3e-04
P60	3.457e-04	4.486e-05	.08693	1.312e-04
OZS%	1.854e-05	3.600e-05	.04076	1.35e-04
USAT Rel%	4.232e-05	5.862e-06	.07707	1.335e-04
TOI	1.836e-04	1.029e-05	.3379	1.09e-04

TABLE IV. LINEAR REGRESSION VS LOESS SALARY

	LR Coeff Estimate	*LR Std. Error.*	*LRR^2*	*Loess Std Error*
G60	4.529e-04	9.450e-05	.03351	1.468e-04
A60	4.278e-04	6.565e-05	.06371	1.42e-04
P60	3.695e-04	4.886e-05	.08397	1.432e-04
OZS%	1.967e-05	3.918e-05	.03883	1.471e-04
USAT Rel%	4.477e-05	6.389e-06	.07294	1.454e-04
TOI	1.959e-04	1.129e-05	.3255	1.213e-04

Tables 1 and 2 display the feature scaled distribution of the League Adjusted Cap and Salary Models, respectively. Numbers suggest that the Salary Model has wider ranges of residuals than the Cap Model. This can be attributed to the fact that the cap numbers of a contract are uniform through out the duration of the contract, and salaries are often front or back-loaded by age, resulting in little changes despite a possible improvement or a decline in a player's performance. In accordance to the Residuals and Variances in the tables, the plots of the League Adjusted Cap Model (Figure 3) and League Adjusted Cap Model (Figure 4) display smooth lines with a concave dip in the center. The concavity of the plot is the result of players who possess large contracts with high variability in statistics across G60, A60, P60, OZS%, SAT, and TOI.

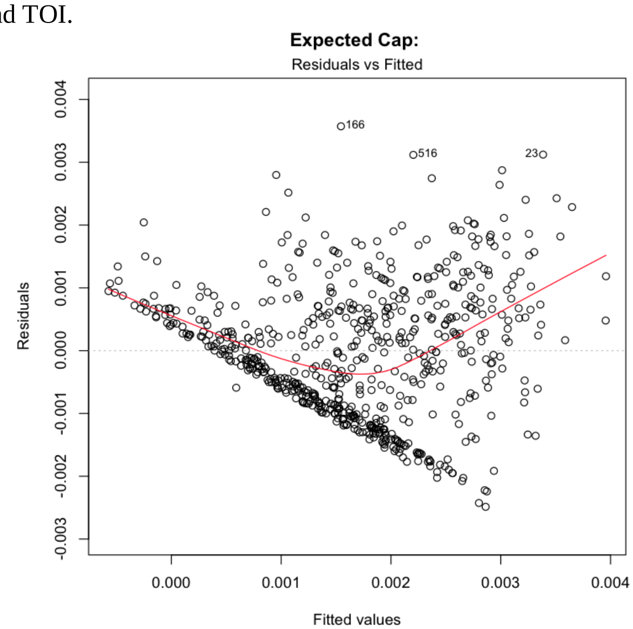

Figure 3. Plot of League Adjusted Cap Model

Figure 4. Plot of League Adjusted Salary Model

Tables 3 and 4 are the direct results (coefficient estimates, standard errors, R-squared) of the Linear Regression and Local Polynomial Regression on the Enhanced statistics vs. League Adjusted Cap and Salaries. Standard Errors of all the estimates are negligibly small. There exist many counterintuitive results from the League Adjusted Salary Model. As can be seen in Tables 5 and 6, Time on Ice has the strongest R-squared value at .3379 and .3325 for both Salary and Cap. This may suggest that despite any type of production, Time on Ice is the most likely determining factor contract signings. What may be surprising is that the next highest determining factor of salary is the P60. In modern Enhanced Statistics, USAT Rel % is generally accepted as better indication of a player's ability than P60. However, this result may shows that perhaps obvious numbers in production are more valued in contract signing than, possession numbers (USAT Rel %).

In addition, as evident by near-0 Residuals for the random effect (Teams), the team association has zero impact on the salary itself. In other words, given a set of on-ice production, you will not be paid higher or lower by playing on a certain team.

IV. CONCLUSION

The League Adjusted Salary Model proposed in this paper is not just a predictive model to gauge a player's potential salary. As discovered through this analysis, with the weighting the bought-out players, and by deriving the League-Relative Salary Percentage, we can create a meaningful training set for which a plethora of statistical models, not limited to Local Polynomial Regression Modeling, may be applied. While this model is at an early stage with comparisons of only six advanced statistics as dependent variables, with expanded parameters and caution, League Adjusted Salary Model has the potential to become a powerful tool in analyzing sports economics.

There are many possible areas of improvements to League Adjusted Salary Model. As is the case in most statistical analyses, it is possible to improve the underlying statistical model. While we incorporated Local Polynomial Regression Modeling to account for standard error in Linear Regression, a more advanced modeling technique could be applied to better fit the data and reduce errors. Another area of improvement could be within the data itself. There were assumptions made in the data and methodology that may be deemed unnecessary in the hockey analytics community. Incorporating more independent variables, such as age and nationality may result in a better training set for the League Adjusted Salary Model. Inclusion of goalies in a much more complicated model is due next. An examination of previous lockout years such as the 2004 NHL Lockout may be another relevant area of research. Finally, valuation of contract clauses, such as No Trade Clause (NTC) was ignored for this paper. The author believes that these issues could have a significant impact in salary models to come.

REFERENCES

[1] R. Vollman, I. Fyffe, and T. Awad, "Rob Vollman's Hockey Abstract," N.p.: CreateSpace Independent Platform, 2014.

[2] M. Kuhn, and K. Johnson. "Applied Predictive Modeling," Springer, 2013, pp. 464-465.

[3] T. Hastie, R. Tibshriani, and J. Friedman. "The Elements of Statistical Learning," Springer, 2013, pp. 197-198.

[4] NHL.com

[5] War-on-Ice.com

[6] HcokeyBuzz.com

[7] Spotrac.com

[8] NHLPA.com

[9] CapGeek.com