Predicting NHL Salaries

Scott Lundquist

December 22, 2017

Contents

1. Introduction

Throughout the world of professional sports there have always been elite players who has earned significantly more than others throughout their careers. In professional hockey some examples of these high value players are Bobby Orr, the first player in history to sign a million dollar contract, or Wayne Gretzky, the highest scoring player in NHL (National Hockey League) history, recording a record 894 goals in his career. But what makes a player more valuable than others? This report will focus on answering this question, more specifically which player statistics or characteristics can be used to optimally predict the level of salary a player will recieve.

The National Hockey League is a professional hockey league which was formed in Montreal, Canada in 1927. The League initially started with 6 teams and has now expanded to 31 teams playing in most major cities across Canada and United States [1]. Each team consists of 18 players and 2 goaltenders, and every team in the leagues is subject to a salary cap [2]. This is the maximum amount of money allowed to be paid out to its players during a single season, currently the cap is 70.2 million (USD) [3].

This report will focus on a dataset provided by "Hockey Abstract" which includes statistics and salary information on 898 NHL players who played during the 2015-2016 season [4]. This data does not include information on any goalies, as their recorded statistics are significantly different from a forward or defensemen. The source dataset included approximately 100 statistics on each player, in order to focus on key contributers I have chosen a subset of 11 statistics which will be used as explanatory variables for the salary in millions (USD) of a given player.

2. Variables of Interest

2.1 Goals, Points, and Assists

Due to the number of statistics on each player it was neccessary to narrow down the data set to include a subset of these. In order to decide which variables would be used as potential indicators for salary prediction, I had to further investigate the relationship particular variables had with a players salary. The first explanatory variables to investigate was the total goals, assists, and points a player earned through the season. In order to avoid the having linear combinations present in the design matrix atmost two of these variables can be selected as predictors.

Figure 2.1.1-2.1.3 indicate a positive relationship between the number of points, goals, and assists earned in a season with respect to the log of a players salary. Although there does appear to be a larger variance in the distribution for smaller values of each explanatory variables which could be problematic for prediction purposes. From the graphs it appears that the total points a player earns throught the season has the strongest effect on their salary as the positive trend appears most evident. I have chosen to map an aesthetic to the rookie players as they have a salary cap of approximately 925,000 (USD)[5], so even though these players may be earning many points throughout the season this will not be a significant indicator of their salary relative to the effect on player not classified as a rookie.

Figure 2.1.1: Players Salary vs. Total Points

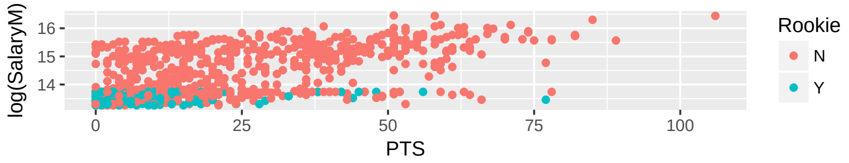


Figure 2.1.2: Players Salary with respect to Assists Recieved

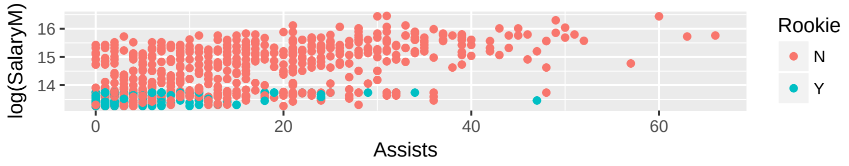
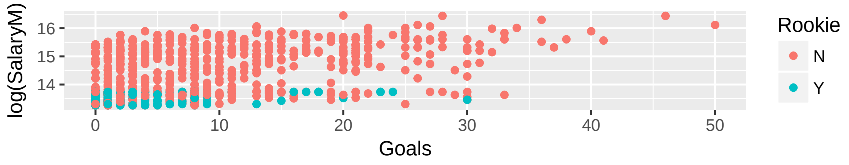


Figure 2.1.3: Players Salary with respect to Goals



2.2 Time on Ice and Total Penalties in Minutes

In figure 2.2.1 we can see that the players total time on the ice per game appears to have a positive relationship with respect to their salary. This makes intuitive sense as the coach will tend to play his most expensive players more frequently as they are suppose to have the greatest contribution to the teams success. I have chosen to include total ice time per game as appose to total ice time, because a high value player who was injured for a part of the season may not necessarily have a high value of total ice time.

In a hockey game a player who commits a foul is penalized with a minimum of 2 minutes spent in the penalty box, and their team must play for the duration of that time with one less player. We would expect that a players salary would be negatively effected by their time spent in the penalty box as they are negatively influence the teams change of success. In Figure 2.2.2 we see that this relationship does not appear to be as strong as anticipated and appears that it has little effect on the players salary.

Figure 2.2.1: Players Salary with respect to Time on Ice per Game

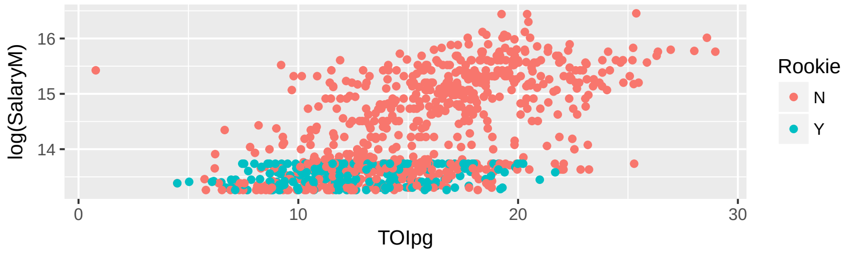
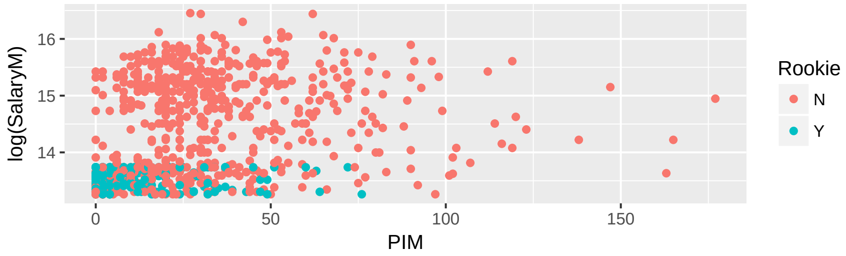


Figure 2.2.2: Players Salary with respect to Penalty Minutes



2.3 Defensive Shots Blocked and Total Faceoff Wins

In a game of hockey each time the referee blows his whistle and stops play, both teams set up for a faceoff in which two players battle for the puck when the referee drops it. A player winning these faceoffs is contributing to the success of their team, which should increase the value of that player. This variable may lack in predictive ability due to the fact that the faceoffs are taken by the centerman, so a defensemens ability to take faceoff should not have any relationship between his salary.

Another important statistic used to judge the effectiveness of a player to their team is the total shots blocked, but this statistic suffers from a similar problem to Faceoff wins. A high number of shots blocked will indicate a quality defensive player but will not give sufficent indidcation of the quality of a forward player. We can see from Figure 2.3.1-2, using statistics unique to a particular position will not serve adavntageous for prediction purposes.

Figure 2.3.1: Players Salary with repsect to Total Faceoff Wins

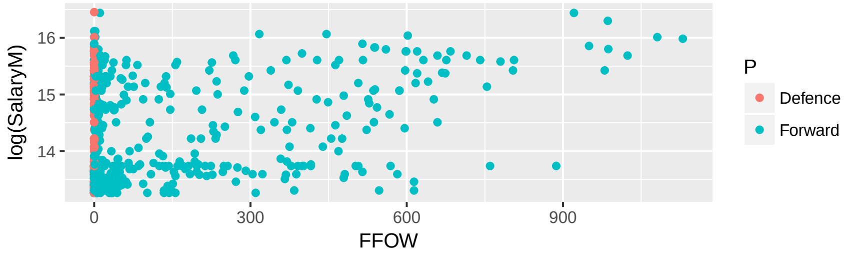
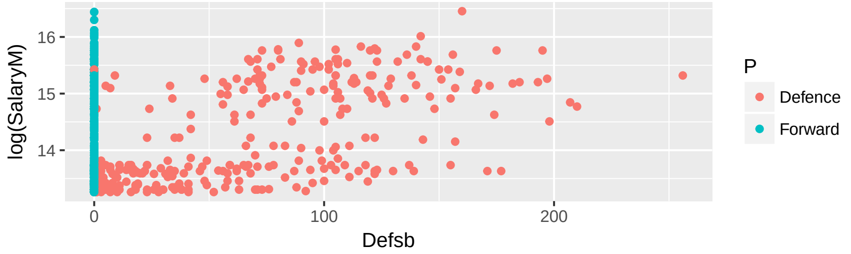


Figure 2.3.2: Players Salary with respect to Total Shots Blocked



2.4 Categorical Variables

Aside from players statistics there are some other key categorical variables which can greatly effect a players salary. As mentioned previously, a rookie players salary is capped by a much lower maximum salary clause which was put into place to protect teams from investing in players that are unable to make the transition into professional play. This will cause rookies who preform as well as the best players in the league to earn significantly less. In Figure 2.4.1 as well as the plots in Section 2.1 and 2.2 we can see a significant difference in rookie salaries relative to average players. As a result I have chosen to remove these players from the dataset as their salaries are not dependent on their current playing preformance.

On each NHL team there is a designated captain as well as assistant or alternate captain who act as the leaders. It is common for these players who are selected as captains or alternate captains to be amongst the highest paid players on the team, although this is not always the case. We can see in Figure 2.4.2 that the players with a captain or assistant captain designation have a significantly higher median salary compared to those without the designation, though we do note what appears to be outliers in the population of captians.

Figure 2.4.1: Rookie Salary vs. Non−Rookie Salary

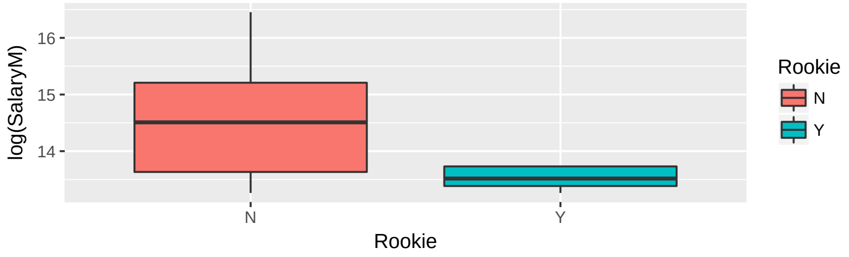
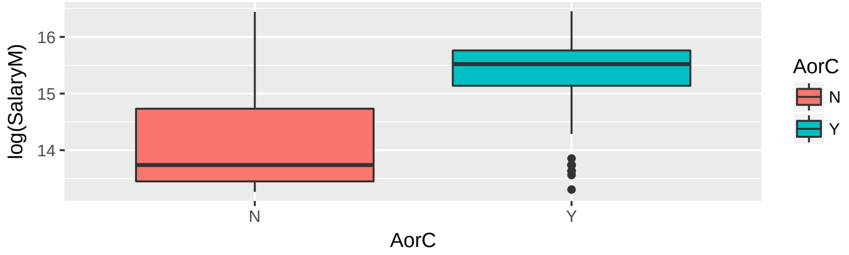


Figure 2.4.2: Captain Salary vs. Non−Captain Salary



3. Model Building

3.1 Variable Selection

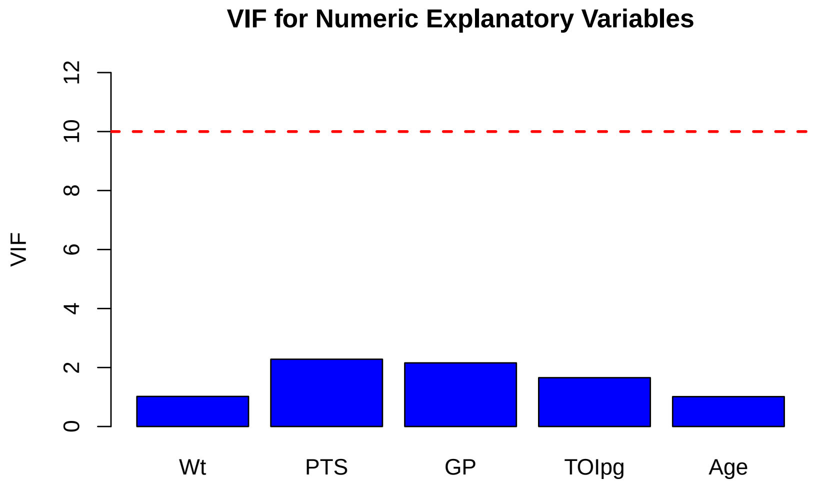
As mentioned earlier, we are interested in modelling a players salary for prediction purposes. In order to develop the model I have begun by first splitting the data into a training data set, approximately 70% of the data, and a test data set, approximately 30% of the data. I have chosen 70/30 split to reduce the risk of overfitting the data to the training data set. I have implemented the backward stagewise regression method using the AIC criterion to select the model which best fits the data. Using the players Age, Height, Weight, Games Played, Time on Ice Per game, Penalty in Minutes, Forward Faceoff wins, Percentage of Team Goals, Position, and Captain status. After running the algorithm the model which minimized the AIC included Position, Weight, Games Played, Total Points, Age, Time on Ice per game, Percentage of Total Goals, and Captain Indicator. Although the model appeared to be quite parsimonious with respect to the number of variables included in the data, the R^2 value was low at 0.5994 which could be improved for the purposes of accurate prediction.

3.2 Box Cox Transformation

Due to the large deviation in players salaries for star players relative to the average player the residual errors in the model were very large. In order to reduce the residual error I have applied a log transform technique to the response variable, so the final model will be predicting the log of the players salary in USD. The variance reduction not only reduced the magnitude of the residual errors but also improved the model in terms of R^2 as the value rose to 0.6282. Also, the distribution of the players salary apeared to be severly right skewed as there are significantly more players earning lower salaries than those with high salaries. As a result of the box-cox transformation, specifically the log transformation, the distribution of the players salary became approximately normal.

3.3 Multicolinearity

In order to ensure the effect of multicolinearity is not present in any of the variables I have used in the model, I have referred to the Variance Inflation Factor(VIF). After measuring the VIF for each quantitative variable in the data there does not appear to be any variables with VIF > 10. Thus, we can conclude that none of the variable involved in model possess significant multicolinearity with repsect to any other explanatory variable.



4. Final Model

In order to create a more parsimonious model for prediciton purposes, I have looked at the summary statistics for the proposed model derived using backward stagewise regression, and have found that the Total Games Played and Percentage of teams goals does not appear to be a significance influence in the model. With respect to the null hypotheses $H_0 : \beta_{GP} = 0$ and $H_0 : \beta_{GP} = 0$ there does not appear to be significant evidence to suggest that its slope parameters are different from zero, hence I have chosen to remove them. Therefore the final model will be of the form, $\hat{Y} = \hat{\beta_0} + \hat{\beta_P} X_P + \hat{\beta_W} X_W + \hat{\beta_{PT}} X_{PT} + \hat{\beta_T} X_T + \hat{\beta_A} X_A + \hat{\beta_C} X_C$ Here, the coefficients are as follows $\hat{Y}$: Logarithm of Players Salary in USD, X_P: Position (Forward = 1, Defence = 0), X_W: Player Weight, X_{PT}: Total Points, X_T: Time on Ice per Game, X_A: Age of Player, X_C: Captain Status (captain/Assistant = 1, Otherwise = 0).

Table 1:

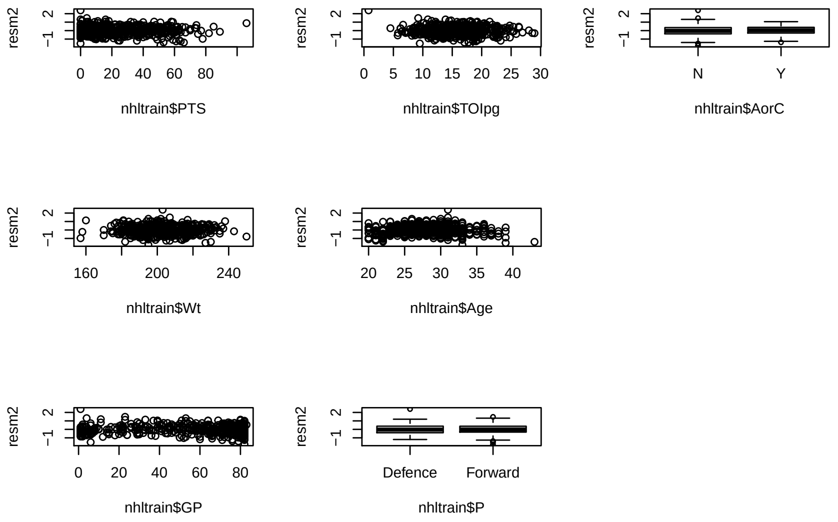
	Dependent variable:	
	log(SalaryM)	
	(1)	(2)
PForward	0.228^{**}	0.219^{**}
	(0.091)	(0.088)
Wt	0.004^{**}	0.004^{**}
	(0.002)	(0.002)
PTS	0.011^{***}	0.010^{***}
	(0.003)	(0.002)
GP	0.001	
	(0.001)	
TOIpg	0.084^{***}	0.087^{***}
	(0.012)	(0.011)
Age	0.086^{***}	0.087^{***}
	(0.006)	(0.006)
AorCY	0.306^{***}	0.296^{***}
	(0.077)	(0.074)
PercentTG	-0.992	
	(1.452)	
Constant	9.485^{***}	9.458^{***}
	(0.452)	(0.449)
Observations	470	470
R^2	0.628	0.628
Adjusted R^2	0.622	0.623
Residual Std. Error	0.537 (df = 461)	0.536 (df = 463)
F Statistic	97.344^{***} (df = 8; 461)	130.067^{***} (df = 6; 463)
Note:		*p<0.1; **p<0.05; ***p<0.01

5. Residual Analysis

5.1 Residuals with Respect to Explanatory Vairables

When analyzing the residual plots with respect to the explanatory variables we are interested in determining whether or not the variance of the residuals with respect to each explanatory variable remains constants across all values. This constant variance idea can be visualized by a horiztontal band of residuals and it should be centered about zero. After analyzing the residuals there does not appear to be any significant violations of the constant variance assumption. In the plot of residuals vs. Games Played there does appear to be less variance in residuals about he middle values but this is likely due to lack of data in this region as this would represent the players who were injured, or players who were called up as replacments for injured plyers for a large portion of the season. We can also see a slight fanning out of the residuals with respect to Games played but this does not appear to be a significant deviation from the assumption.

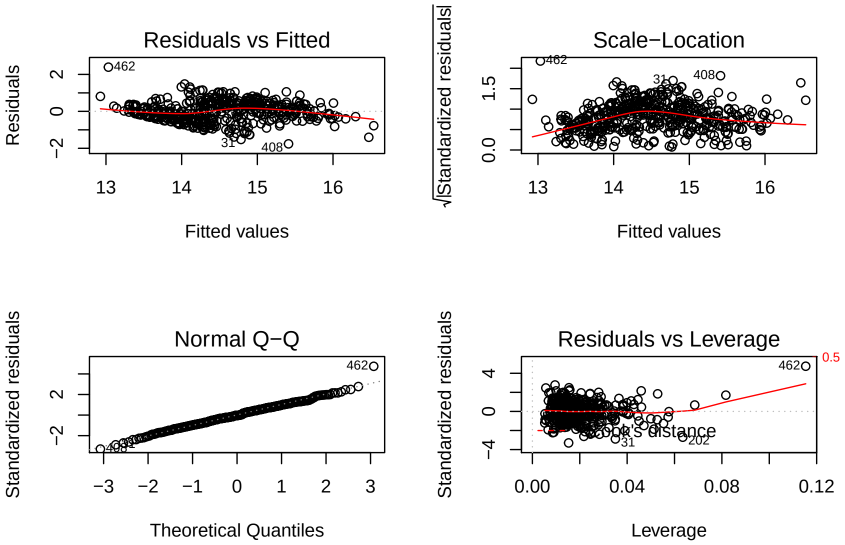
In the plots of the categorical variables the IQR of residuals for defensive players appears to be slightly larger than the IQR forwards but there does not appear to be any significance violations. We can also see that both forward and defensemen appear to a have a few outliers, but this is expected as there are few extremely high earning players relative to the rest of the league. Also looking at the residual plot for captain status we see that the players with captain designation appear to have a median residual value which is slightly above zero, indicating the prediction values for these players may be slightly lower on average but there does not appear to be a significant bias which would greatly effect prediciton.



5.2 Basic Residual Plots

To determine the adaquacy of the final model we must also looks at some basic residual plots. Looking at the residual vs. fitted plot we do see a slight deviation from the costant variance assumption as the model appears to be predicting better for small values of the response variable. We can also see there are some outliers present in the data, most significantly are points 618, and 109. We can see the same pattern being exhibited in the scale-location plot about the centeral fitted values. The variance assumption doesnt appear to be seriously violated although this may have a slight adverse effect on prediction about values in that range.

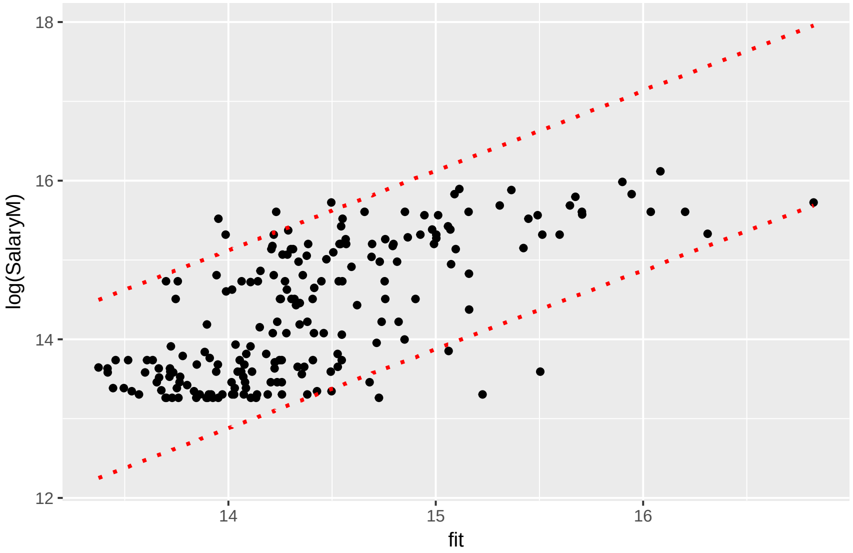
To test the normality assumption of the multiple regression model we look at the Q-Q plot of the standardized residuals vs. the standard normal values. According to the Q-Q plot there does not appear to be any significant violation to the normality assumption. There does appear to be one outlier in the upper end of the data, which deviates from the normal distribution quantile but this is likely caused by one player who is earning significantly more than the rest of the league.



6. Prediction

To investigate the predictive capability of the model I have created a plot of the players log(salary) with respect to their fitted (predicted) values and have included a 95% Predicition interval (red dashed line). We can see from the plot that the players true log(salary) appear to be well contained within the interval aside from a observations, indicating a strong predictive capability of the model, and no evidence of over training inherent in the model.

Figure 6.1: True Values vs Fitted Values with 95% Prediciton Interval



7. Conclusion

In conclusion, I have found the variables which were most influential in predicting players salaries to be their position (forward/defence), total points (assists plus goals), weight, time on ice per game, age, and captain status. For the purposes of normalizing the data and reducing the variance, modelling the log of a players salary was optimal. The overall model apears to satisfy the three primary assumptions of regression model (normality, constant variance, and independence), without serious violations indicating that multiple linear regression was adaquate for modelling salaries. Finally, we saw in figure 6.1 that the linear model was deemed useful for prediction purposes as the 95% prediction intervals contained nearly all data point in the test data set.

8. References

[1] National Hockey League. (2017, December 19). Retrieved December 21, 2017, from https://en.wikipedia.org/wiki/National_Hockey_League
[2] Hockey Operations Guidelines. (n.d.). Retrieved December 21, 2017, from http://www.nhl.com/ice/page.htm?id=26377
[3] NHL salary cap. (2017, December 06). Retrieved December 21, 2017, from https://en.wikipedia.org/wiki/NHL_salary_cap
[4] Download Statistics. (n.d.). Retrieved December 21, 2017, from http://www.hockeyabstract.com/testimonials
[5] Collective Bargaining Agreement FAQs. (n.d.). Retrieved December 21, 2017, from http://www.nhl.com/ice/page.htm?id=26366

Appendices

A. Model Selection Technique: Backward Stagewise

```
step(nhlmodel, IC=AIC)
```

```
## Start:  AIC=13401.39
## SalaryM ~ Age + HT + Wt + GP + PTS + TOIpg + PIM + FFOW + PercentTG + 
##     P + Defsb + AorC
## 
##              Df  Sum of Sq        RSS   AIC
## - PIM        1 7.8392e+10 1.0750e+15 13399
## - FFOW       1 3.9403e+11 1.0753e+15 13400
## - HT         1 4.9340e+11 1.0754e+15 13400
## - Defsb      1 6.8670e+11 1.0756e+15 13400
## <none>                    1.0749e+15 13401
## - P          1 4.7530e+12 1.0797e+15 13402
## - PercentTG  1 5.7321e+12 1.0806e+15 13402
## - Wt         1 6.4790e+12 1.0814e+15 13402
## - GP         1 2.0045e+13 1.0950e+15 13408
## - TOIpg      1 5.0234e+13 1.1251e+15 13421
## - AorC       1 7.0916e+13 1.1458e+15 13429
## - PTS        1 8.1056e+13 1.1560e+15 13434
## - Age        1 3.2691e+14 1.4018e+15 13524
## 
## Step:  AIC=13399.43
## SalaryM ~ Age + HT + Wt + GP + PTS + TOIpg + FFOW + PercentTG + 
##     P + Defsb + AorC
##
##              Df  Sum of Sq        RSS   AIC
## - FFOW       1 3.7600e+11 1.0754e+15 13398
## - HT         1 5.0596e+11 1.0755e+15 13398
## - Defsb      1 6.8622e+11 1.0757e+15 13398
## <none>                    1.0750e+15 13399
## - P          1 4.7016e+12 1.0797e+15 13400
## - PercentTG  1 5.7569e+12 1.0807e+15 13400
## - Wt         1 6.8724e+12 1.0819e+15 13400
## - GP         1 2.3155e+13 1.0981e+15 13407
## - TOIpg      1 5.0272e+13 1.1253e+15 13419
## - AorC       1 7.2307e+13 1.1473e+15 13428
## - PTS        1 8.1063e+13 1.1560e+15 13432
## - Age        1 3.2782e+14 1.4028e+15 13522
##
## Step:  AIC=13397.59
## SalaryM ~ Age + HT + Wt + GP + PTS + TOIpg + PercentTG + P +
##     Defsb + AorC
##
##              Df  Sum of Sq        RSS   AIC
## - HT         1 4.4892e+11 1.0758e+15 13396
## - Defsb      1 5.0702e+11 1.0759e+15 13396
## <none>                    1.0754e+15 13398
## - P          1 5.1173e+12 1.0805e+15 13398
## - PercentTG  1 6.1125e+12 1.0815e+15 13398
## - Wt         1 6.8984e+12 1.0823e+15 13399
## - GP         1 2.2784e+13 1.0981e+15 13405
## - TOIpg      1 5.2291e+13 1.1277e+15 13418
## - AorC       1 7.4309e+13 1.1497e+15 13427
## - PTS        1 8.3187e+13 1.1586e+15 13431
## - Age        1 3.2821e+14 1.4036e+15 13521
##
## Step:  AIC=13395.79
## SalaryM ~ Age + Wt + GP + PTS + TOIpg + PercentTG + P + Defsb +
##     AorC
##
##              Df  Sum of Sq        RSS   AIC
## - Defsb      1 5.8457e+11 1.0764e+15 13394
## <none>                    1.0758e+15 13396
## - P          1 5.3508e+12 1.0812e+15 13396
## - PercentTG  1 5.9460e+12 1.0818e+15 13396
## - Wt         1 9.2165e+12 1.0850e+15 13398
## - GP         1 2.2630e+13 1.0984e+15 13404
## - TOIpg      1 5.1842e+13 1.1277e+15 13416
## - AorC       1 7.5001e+13 1.1508e+15 13426
## - PTS        1 8.3025e+13 1.1588e+15 13429
## - Age        1 3.3736e+14 1.4132e+15 13522
##
## Step:  AIC=13394.04
## SalaryM ~ Age + Wt + GP + PTS + TOIpg + PercentTG + P + AorC
##
##             Df  Sum of Sq        RSS   AIC
## <none>                    1.0764e+15 13394
## - P          1 4.9230e+12 1.0813e+15 13394
## - PercentTG  1 7.0932e+12 1.0835e+15 13395
## - Wt         1 9.5768e+12 1.0860e+15 13396
## - GP         1 2.2784e+13 1.0992e+15 13402
## - TOIpg      1 6.5826e+13 1.1422e+15 13420
## - AorC       1 7.5317e+13 1.1517e+15 13424
## - PTS        1 8.2440e+13 1.1588e+15 13427
## - Age        1 3.3762e+14 1.4140e+15 13520
```

```
##
## Call:
## lm(formula = SalaryM ~ Age + Wt + GP + PTS + TOIpg + PercentTG +
##     P + AorC, data = nhltrain)
##
## Coefficients:
## (Intercept)          Age           Wt           GP          PTS
##    -8685174       207358        10393       -12839        56890
##       TOIpg    PercentTG     PForward        AorCY
##      179907     -7202487       376735      1238990
```

```r
summary(nhlm1)
```

```
##
## Call:
## lm(formula = SalaryM ~ P + Wt + PTS + GP + TOIpg + Age + AorC +
##     PercentTG, data = nhltrain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4513851  -850562  -103466   645930  8185073
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -8685174    1286358  -6.752 4.41e-11 ***
## PForward      376735     259453   1.452   0.1472    
## Wt             10393       5132   2.025   0.0434 *  
## PTS            56890       9574   5.942 5.56e-09 ***
## GP            -12839       4110  -3.124   0.0019 ** 
## TOIpg         179907      33883   5.310 1.71e-07 ***
## Age           207358      17244  12.025  < 2e-16 ***
## AorCY        1238990     218152   5.679 2.40e-08 ***
## PercentTG   -7202487    4132337  -1.743   0.0820 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1528000 on 461 degrees of freedom
## Multiple R-squared:  0.5994, Adjusted R-squared:  0.5924 
## F-statistic: 86.22 on 8 and 461 DF,  p-value: < 2.2e-16
```

```
summary(nhlm2)
```

```
## 
## Call:
## lm(formula = log(SalaryM) ~ P + Wt + PTS + GP + TOIpg + Age + 
##     AorC + PercentTG, data = nhltrain)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.75936 -0.35811 -0.01604  0.37268  2.39280 
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  9.4851458  0.4518578  20.991  < 2e-16 ***
## PForward     0.2279631  0.0911376   2.501  0.01272 *  
## Wt           0.0039548  0.0018027   2.194  0.02875 *  
## PTS          0.0113627  0.0033631   3.379  0.00079 ***
## GP           0.0006808  0.0014437   0.472  0.63746    
## TOIpg        0.0844691  0.0119021   7.097 4.84e-12 ***
## Age          0.0863656  0.0060573  14.258  < 2e-16 ***
## AorCY        0.3062959  0.0766299   3.997 7.47e-05 ***
## PercentTG   -0.9915626  1.4515621  -0.683  0.49489    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
```

```
## Residual standard error: 0.5368 on 461 degrees of freedom
## Multiple R-squared:  0.6282, Adjusted R-squared:  0.6217
## F-statistic: 97.34 on 8 and 461 DF,  p-value: < 2.2e-16
```