Advanced Regression Techniques

Vincent Ditmore, Cory Swift, Bobby Mullins

NHL Salaries Predicted… Good or Bad Values

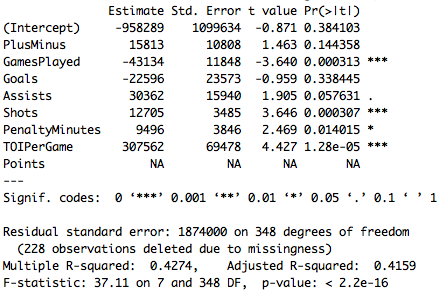
Final Project Paper

For our Final Project we decided to do a 2-part Project where we looked to predict NHL Forward’s salaries by using performance based statistics. We then compared the predicted NHL salaries to their actual salaries to determine whether or not the specific players were good or bad value for their respective team. We decided to do our project on this as we are dedicated hockey players who have grown up watching and playing with current NHL players. We are all very interested in the business of sports and why NHL players are getting paid the way they do.

In our model we used Salary as our response variable. Salary is determined by the predictor variables: Age, Plus/minus, Games Played, Goals, Assists, Shots, Penalty minutes, Ice-time per game and Points. We had a few options when it came to choosing our response variable as we had the option of choosing between Total Money Earned and Salary. We decided to use Salary as our Response Variable as it does not include bonuses which actually count towards the next year’s salary cap. As one might notice some of the predictor variables included are not performance based statistics. We used these predictor variables to eliminate some players from our data set. We did this with the predictor variables, Age and Games Played. In order for our model is be the most accurate in predicting NHL Salaries we thought to eliminate Players who have played less than 45 games, and players currently on entry level contracts in the NHL. We did this by eliminating all players that are 24 years old and younger.

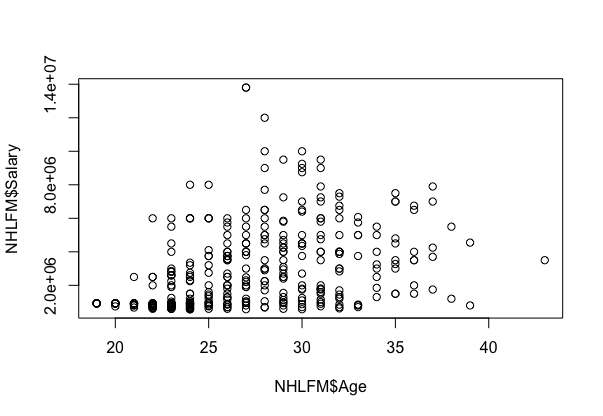
We excluded defenseman and goalies from our salary predicting model since they are valued by different statistics than forwards. A goalie would be valued by the amount of wins, saves and save percentage he has and a defenseman would be valued by how many goals he prevents, not necessarily by how many he scores. Therefore, to have a more accurate model we only included Forwards which the majority are valued on many of the same performance based statistics.

**Full Model= *Salary ~ PlusMinus + GamesPlayed + Goals + Assists + Shots + PenaltyMinutes + TOIPerGame + Points***

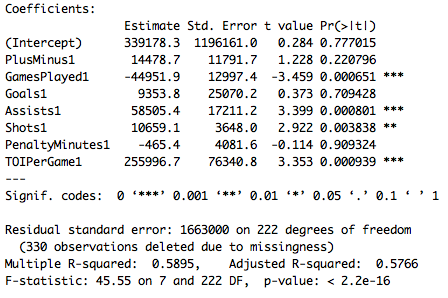


This full model we thought was a good starting point. In this model we included all NHL Forwards who have played 45 games or more. Obviously, by looking at the estimated coefficients for the predictor variables there were some surprises that we needed to look into. For example, we found it interesting that Games Played and Goals both have a negative coefficient which we will explore later. Also in the summary the predictor variable, Points, did not have any output and upon further examination we found that Points is something called a “Aliased Coefficient.” This means that Points is perfectly collinear with other predictor variables and therefor has no impact on our model. Also, note that our y-intercept is -958,289 which means that a player with absolutely no statistics at all should actually be paying the team he is rostered on which we found to be a bit humorous. All jokes aside, we know from our y- intercept that we have a long way to go in creating an accurate model. With an R^2 value of .4274 we are currently not satisfied with only 42.74% of the variability in our response variable that can be explained by our predictor variables.

Prior to making our next model, Reduced Model 1, we wanted to examine one the nonperformance based statistics, Age. We did this by plotting Age of the X axis and Salary on the Y Axis.

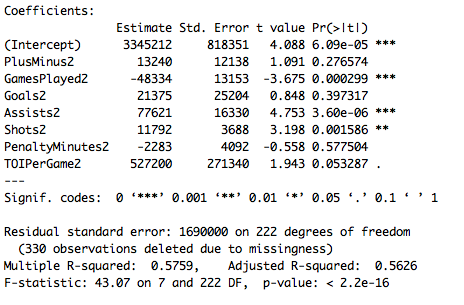


As we can see in the graph players 24 and under tend to have darker circles lower on the Y- Axis. This is due to the fact that players 24 and younger entering the league are all placed on entry level contracts. We thought by attempting to target the linear relationship on this plot (where variance is constant) it would be best for our model to only include players 25- 37 years old.

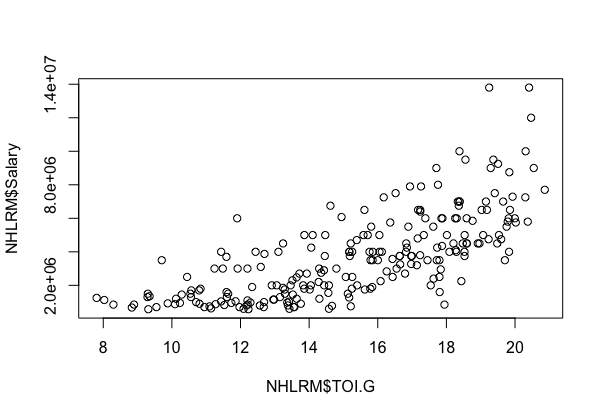


Just by looking at the summary of our Reduced Model 1, we were very satisfied with our decision to take out players both ,24 and younger, and 38 and older. Note that our R^2 value improved drastically from our full model, and that our Y-intercept is now a positive value. We previously were surprised that our predictor variable, Goals, had a negative coefficient, however, in Reduced Model 1 it now has a positive coefficient. We are still surprised that the coefficient is still very low compared to the other predictor variables. With our previous knowledge we still believe that goals should have one of the higher coefficients in our model because gifted goal scorers are very hard to come by in today’s NHL.

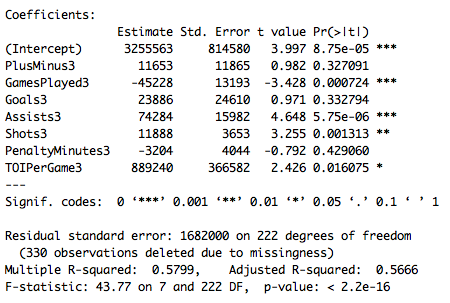
In our next 2 models, Reduced Model 2 and 3, we decided to create a Boolean variable for the predictor variable, TOI/G. We thought this would better represent how we value predicting salary through TOI/G. Note that the coefficient for TOI/G in Reduced Model 1 is the greatest coefficient with a value of 255,996.7. This means that for every minute played a player earns 255,996.7 dollars extra. We found this number to be pretty high so in creating model 2 we decided to try to fix this problem. In Reduced Model 2, we gave players who played 13 minutes and under a 0(4th line guys), players who played 14-17 minutes a 1(2nd/3rd line guys) and players who played 18-22 minutes a 2(1st line guys). We then ran the model.



We liked the Reduced Model 2’s estimated coefficients better than the previous model. If I was to interpret the coefficient for the predictor variable, TOI/G, we would say that 1st line guys make 527,200 dollars more than 2nd/3rd line guys just based on TOI/G. It is also true if we said 2nd/3rd line guys make 527,200 dollars more than 4th line guys. We found this to be a cool way to show that the more money a NHL forward makes the more time on ice he usually plays as illustrated by the plot.

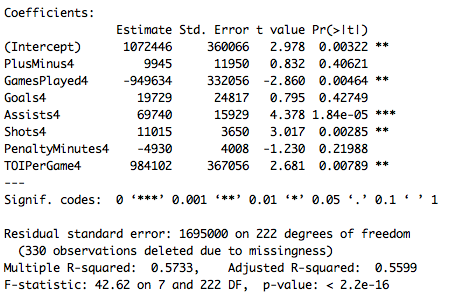


As one may have noticed the estimated coefficient for the predictor variable, Goals, increased the most significantly which we liked a lot and we thought that the slight dip in the R^2 value was fine because the weighted distribution of our predictor variables is better in our eyes. However, in Reduced Model 3 we were looking to better our R^2 value by using a different Boolean variable. This time, we were only looking to distinguish between bottom 6 and top 6 forwards. Bottom 6 forwards received a 0 and top 6 received a 1. We then ran the model.

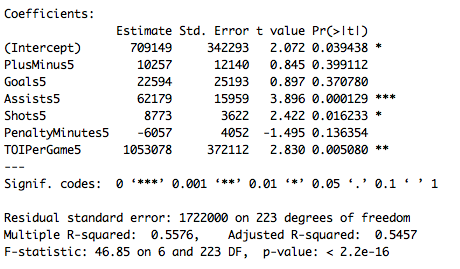


First, note that the R^2 value increased by a little which we liked a lot. Also we were very satisfied with how the coefficients changed from the previous model. The coefficient for Goals increased, while the coefficient for Assists decreased closing the gap between the two by a few thousand.

As we have previously mentioned, we do not understand why the coefficient for Games Played is negative and we have no reason to believe it should be negative. Trying to make sense of this predictor variable we attempted to create a positive coefficient. We attempted this by making Games Played a Boolean variable in Reduced Model 4. A player who played 59 games or less receives a 0 and players who played 60 or more games will receive a 1. We then ran the model.

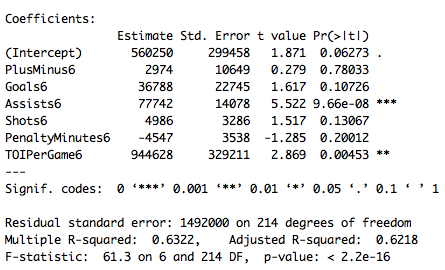


Not only did our R^2 value decrease, but the coefficient for Games Played is now a huge negative number. An interpretation of this coefficient would be that players who played in 60 or more games earn 949,634 dollars less than players who played in 59 or fewer games. After further investigation it makes sense that Games Played has a negative coefficient as the top tier players are on the ice more, and subject to more hits than the average player. As a result, top players tend to get injured. However, in our model we feel as if it is best to take the predictor variable, Games Played, out of the model due to the fact that no one is really positive how many games a player will play due to injury and other circumstances when given a contract. We then took the predictor variable, Games Played, out of our model and then ran it again.

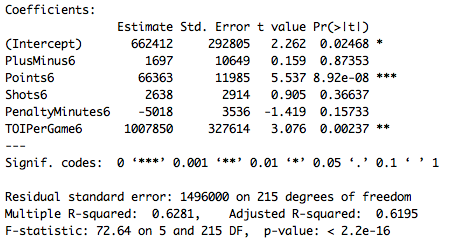


As one might have noticed, the R^2 value decreased a little as expected, however, our model coefficients have not significantly changed. We deemed it necessary to take out the predictor variable, Games Played, and stuck with Reduced Model 5 for the time being.

For our Reduced Model 6, we were set on taking out high leverage and high influence points expecting there to be a few, however, none were over the threshold. We took it upon ourselves to take a few players out of our data set in order to increase our R^2 and decrease our p values for certain predictor variables. In our Reduced Model 6, we took out Jonathon Toews, Rick Nash, Eric Staal, and Dustin Brown. Jonathon Toews the clearest pick to take out as he is not being paid 13.8 million dollars just to produce 60 points each year. He has a lot of other qualities that make him a top paid guy in the league such as defensive prowess, leadership, and face-offs. All of which are not included in our performance based model. The other guys excluded from the model either fall under defensive forwards which our model does not accurately account for or just clear under performers.

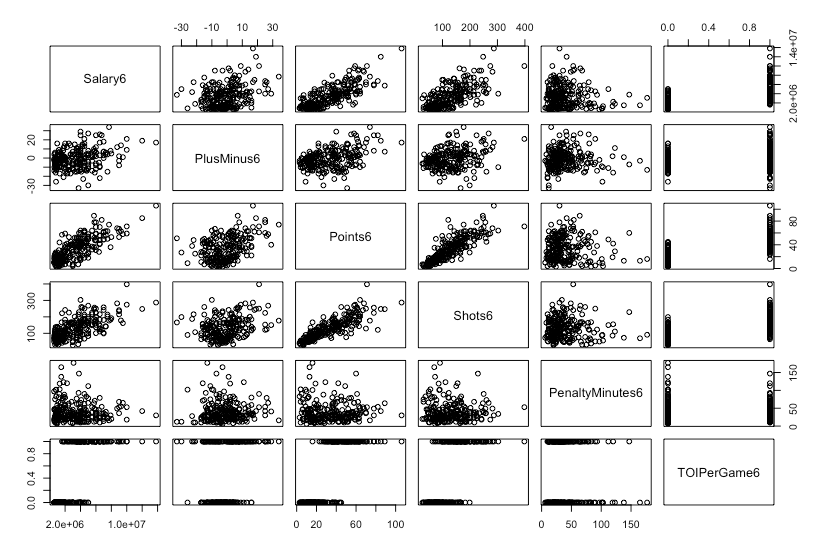


As expected, our R^2 value significantly increased. It is also important to note that the coefficient for Plus/Minus decreased significantly while its p value grew significantly. Made us realize that Plus/Minus is our least significant predictor variable for NHL forward’s salary. However, we are still confused why the coefficient for Assists is much greater than the coefficient for Goals. There is no explanation that we could come up with for that being the reason, hence why we switched back to using the predictor variable, Points, instead of the combination of Goals and Assists. By using Points, a goal and assists are weighted the same therefore, our dilemma is solved.

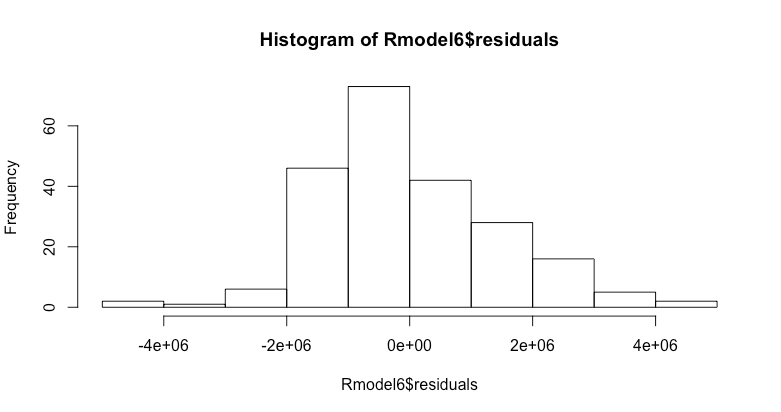


We are very satisfied with the summary of the Reduced Model 7(Final Model). Our R^2 did slightly decrease, however, the p value of points is very close to 0 and all of our coefficients make sense. It is important to note that the p value for Plus/Minus increased once again and so did the p value for shots.

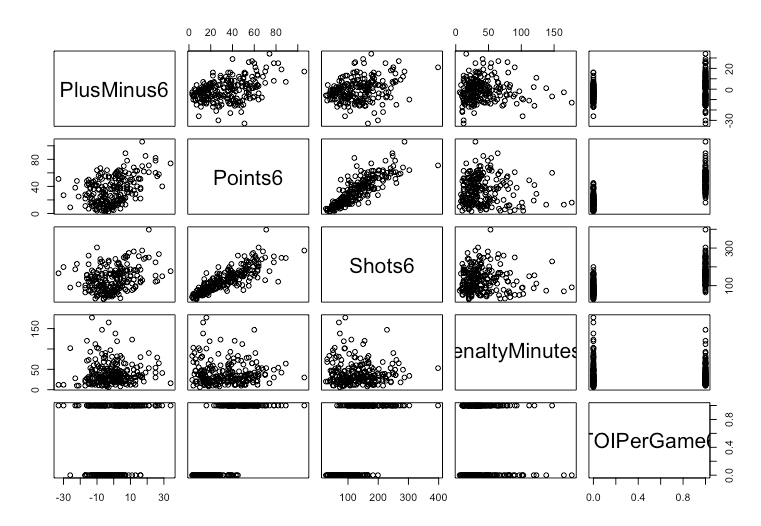
After the construction our final model for the project we needed to go through the four assumptions. When determining if the model is linear, it’s easy to tell by looking at the pairwise comparison. By looking at this graph one could see that salary has some linear relationships with the 5 predictor variables present in the final model. This graph also shows us which variables are our best predictors. In our final model, we can see Points, and TOI/G have the best linear relationship with salary. Plus/Minus and Penalty Minutes have a relationship with salary but doesn’t exactly resemble a linear relationship like Points, and TOI/G. Ultimately, we believed we had enough linear relationships with salary to confirm this assumption. If we were to go on and create another reduce model, we would first explore taking out the predictor variable, Plus/Minus.

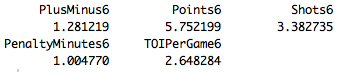


The next of the four assumptions is that the errors are independently and identically distributed. When determining if our model passed this assumption, we decided to plot residuals in a histogram. When examining this histogram, we want to see a normal curve. As you can see in our Final Model Residuals histogram the distribution of error is normal, confirming that the assumption holds for our model.



The 3rd assumption is that the predictor variables are independent of eachother. When deciding if our model met this assumption we decided to use the pairwise comparison again, but this time without plotting salary against the predictors. When looking at this comparison, we don’t want the predictor variables to have any linear relationship. We can see in our Final model that we have two variables that express a linear relationship, Points and Shots. We see that these two variables have collinearity, but when running the VIF Test for our Final Model we can see that none of the predictor variables have a VIF greater than 10. This means collinearity is present, however, it will not be a problem. Our final model passes the 3rd assumption.





The last assumption is that our data is reliable. We feel our data is very reliable since we trimmed down our data set tremendously from the start. We also looked into our outliers, but even though we didn’t have any leverage points greater than .2 or any with a Cook’s distance greater than .8. We took it upon ourselves to scan our data to find potential outliers and remove a few people who salaries are greatly influenced by non-performance based statistics. The actions we took confirm our final model passes the 4th and final assumption.

The Final stage of our project was to determine if the players were a good or bad value. We did this by dividing Predicted Salary by Actual Salary which gave us a percent. If the percent was greater than 100 then we knew the player was a good value for the 2015-16 NHL season, however, if the percent was less than 100 then we knew the player was a bad value and did not live up to expectations. The 3 best values in the NHL during the 2015/16 season were Lee Stempniak, Kyle Palmieri, and Colton Sceviour. The success of teams really depends on finding players like these who just over perform the contract and make the General Manager look like a genius. Lee Stempniak was the best value out of the 3, he was getting paid a salary of $850,000 while our model predicted his salary to be $5,256,027. His predicted salary is 618.36% greater than what he is getting paid, which is just ridiculous. In our Final Model we still had players 24 and younger out of the model, however, we thought it would be cool to go back and predict the Salary of Conner McDavid. This was Conner McDavid’s first year in the league therefor he was on a max rookie contract of $925,000. Although he only played in 45 games, he had a predicted salary of $5,040,655 which means his predicted salary is 544% greater than what his rookie contract is paying him. When finding out that a lot of the superstars were among the worst values in the NHL, according to our model, I was not really that surprised. Sidney Crosby predicted Salary was 65% of his actual 12 million dollar contract. Other notable bad values were Marian Hossa, Ryan Callahan, Evgeni Malkin, and Ryan Getzlaf with all having percentages below 65.

In conclusion, we are very satisfied with how we transformed the model. We started with a R^2 value of .4274, then by cutting down our data set and transforming some of the predictor variables we got our final model to have a R^2 value of .6281 which is very respectably for real world data. We have determined that the most significant predictor variables when predicting a NHL forward’s salary are Points, and TOI/G which to us makes complete sense. It is also confirmed by their relatively high estimated coefficients and low p values. On the other hand, the least significant predictor variables when predicting a NHL forward’s salary are Plus/Minus and PIMs. It is confirmed by their relatively low estimated coefficients (close to 0) and high p values. It is important to note that the predictor variable Plus/Minus has the least significant impact on predicting salary and with its p value of 0.87 we think that taking it out of the model will have little to no impact. The model created is best when predicting a NHL forward’s salary in the middle of the range. We noticed that when predicting NHL superstars’ salaries, the model was not as efficient. NHL forwards who consistently put up outrageous numbers are more likely to earn the max contract. After running a 95% confidence interval. We know that the average NHL forward salary is between 3,232,462 and 3,875,737 dollars. This just confirms why it is harder for our model to predict extreme NHL salaries high or low due to the fact that most of our NHL forward salaries fall around the average.

We had a lot of fun working with a real world data set that we were very interested by. Like we said earlier, we are very interested in careers that revolve around finance and sports, andthis project brought those two worlds together.