Sam Thomas

Analysis of Age and Time on Ice vs. NHL Goalscoring

Introduction

The National Hockey League, or NHL, is the premier professional hockey league in the world. It hosts 31 teams from Canada and the United States. Each team in the NHL shares one main objective every season, to win the Stanley Cup. The aim for each team's general manager is to create the best team they can with limited resources. What separates the good general managers from great general managers is the ability to accurately predict player performance for the upcoming year. This is the main goal of my project.

For my project, I will primarily be analyzing the age at which NHL players "peak" or achieve their best performance. To determine player performance, I will use the response variable Goals. I will investigate whether it is possible to fit a linear model between Age and Goals, thus seeing if there exists a correlation between age and peak play.

My other goal in this project is to determine whether Time on Ice, a different covariate, more accurately predicts player performance. Once again, I will use Goals to reflect player performance.

Additionally, I will investigate any possible confounding variables such as position. It is generally known that forwards score more goals than defensemen. Is it possible that different models will be necessary for forwards versus defensemen?

I will be using hockey-reference.com to get my data. Hockey Reference is a source that contains tons of hockey data for current and former NHL players, coaches, and teams. It is a part of Sports Reference LLC which tracks data for all the big American sports leagues. I will analyze three different NHL seasons: the 2016-2017 season, 2017-2018 season, 2018-2019 season. A link to the data for each season is listed below.

2016-2017 NHL Season: https://www.hockey-reference.com/leagues/NHL_2017_skaters.html

2017-2018 NHL Season: https://www.hockey-reference.com/leagues/NHL_2018_skaters.html

2018-2019 NHL Season: https://www.hockey-reference.com/leagues/NHL_2019_skaters.html

Hypothesis

I predict that Age will be able to accurately predict player performance. I expect there will be a strong correlation between Goals and Age. I believe that there will be a steady increase in the number of goals until the age of 26, and then a steady decline in those metrics. The reason is that players' physical attributes decline at the age of 26 and their goal scoring with this.

Time on Ice will be another way to predict goals via a linear model. The rationale is that the more time a player has playing, the more opportunities they have to score.

Initial Steps

Because hockey is a high contact sport, many injuries occur in the NHL. Players will frequently miss some games during a season due to injury. There are other players within an organization that only play a few games per season in the NHL that fill in for the injured players that regularly play on a team. Due to fewer games, these players do not score many goals over a full season. Additionally, players that are injured may skew the data because they are not playing as many games and their regular totals will be down. For this reason, I must establish a minimum game played mark to avoid a significant skew in the data.

I have decided on a requirement of at least 72 games played in the season to be considered part of the data analysis. Because there are 82 games in an NHL season, this translates to a maximum of 10 games missed per season for a player. This mark was selected because according to NHL rules, players put onto Long Term Injured Reserve must miss at least 10 games. Long Term Injured Reserve is a system the NHL has put into place to allow teams to facilitate a player that has a significant injury. Obviously, injuries can skew the data, and this is the reason I have selected a 72-game minimum.

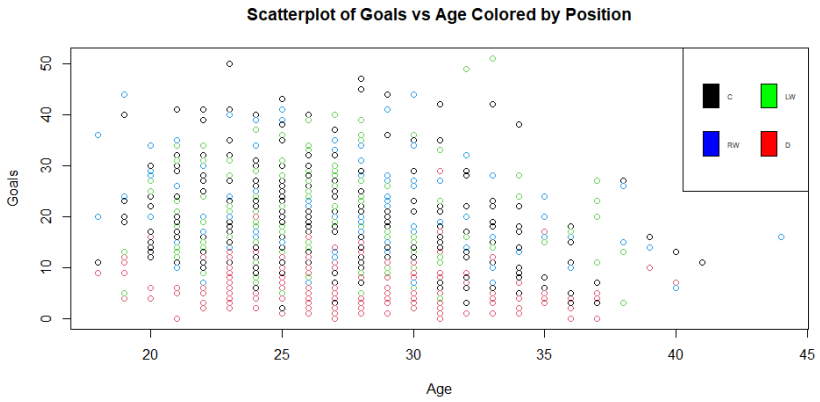
Originally, over the 2016-2017, 2017-2018, and 2018-2019 seasons combined, there were 2,684 players who played at least one game in the NHL. After filtering players who played a minimum of 72 games, there are 977 skaters who meet these requirements.

Next, I will be combining the data from each season into one master dataset. The reason I am doing this is for both ease on my part and a more representative dataset. Of course, if I have data over multiple seasons, then the results will be more accurate for any future season. Also, I am selecting three consecutive seasons to avoid any significant change in how the game is played that may affect player statistics.

Lastly, I will consider an alpha value of 0.05 for every test, and assume that every simple linear model is of the following form: $Y = \beta 0 + \beta 1X + \varepsilon$.

Goals vs Age

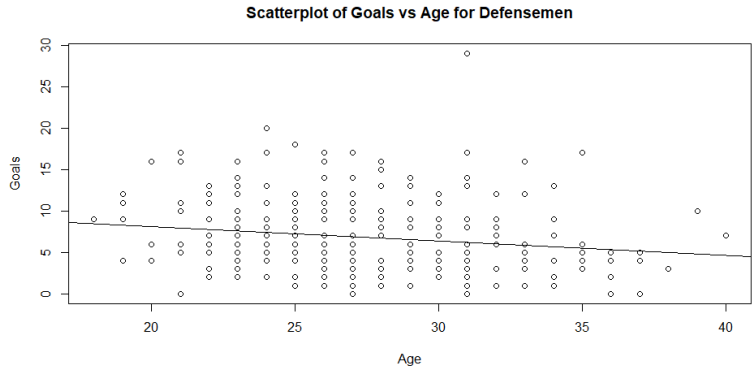
First, I will analyze the relationship between Goals and Age. Here is the initial plot between Goals and Age with colors partitioning position.



Already, it appears that defensemen (red) as a whole score a lot less goals than forwards. For this reason, I will split the analysis of Goals vs Age into two different analyses: one for Defensemen and one for Forwards. Additionally, it is already easy to see that there is a possible negative correlation between Defensemen and Forwards.

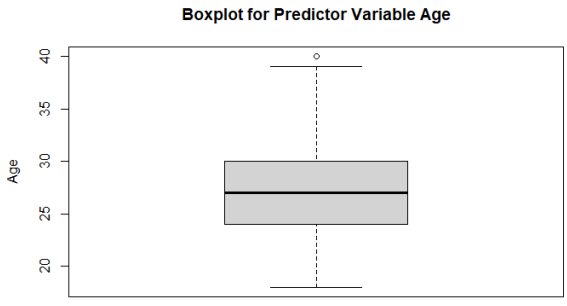
Defensemen

After fitting a linear model in R, we see the following regression line visually on top of the Scatterplot of Goals vs Age for Defensemen. Here, the regression line is E(Goals) = b0 + b1*Age



The line of best fit confirms our assumption that there is a potential negative relationship between Goals and Age. After consulting the Summary table, it appears that b1 = -0.17281. Additionally, the p-value of b1 is 0.00372, confirming a significant linear relationship between Age and Goals. In summary, we have b0 = 11.53 with Var(b0) = 2.66 and b1 = -0.17281 with Var(b1) = 0.0035. Since it is confirmed that there exists a relationship between Goals and Age, it is necessary to look at the diagnostic plots.

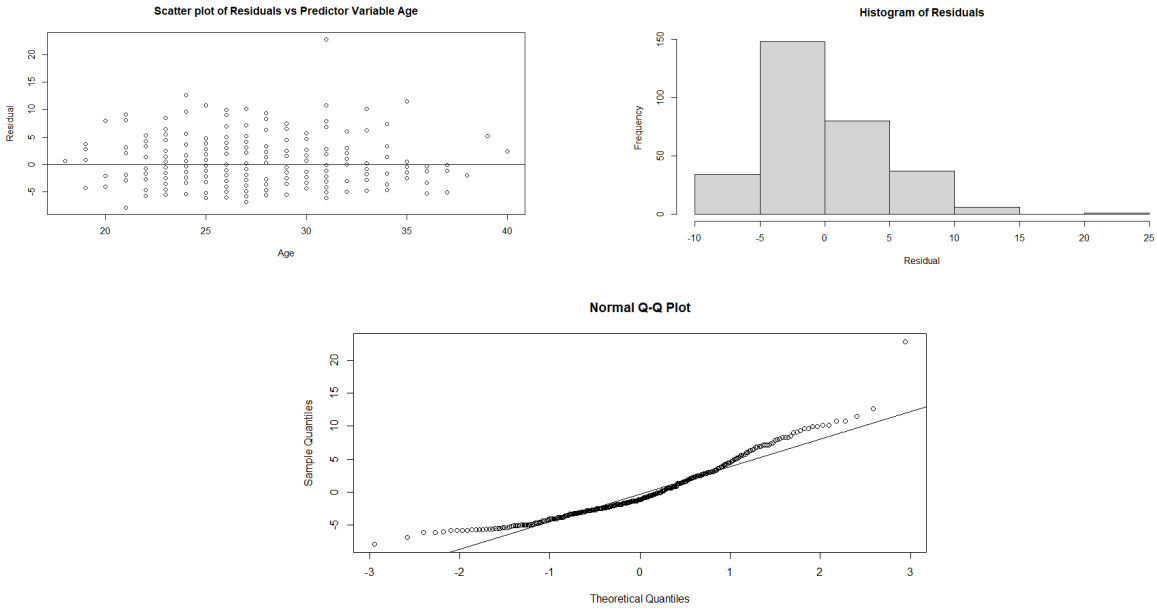
i. Diagnostic Plots for the Predictor Variable Age



It appears that there is an outlier in the predictor variable Age at 40 years old. The name of the player is Zdeno Chara, and everything about him is significant. Interestingly, at 6' 9", he is the tallest player to ever play in the NHL. He is still playing in the league today at age 44 and considered by most hockey critics to be a sure-fire Hall of Famer. I believe because of how unique he is, it is fair to eliminate him from the data.

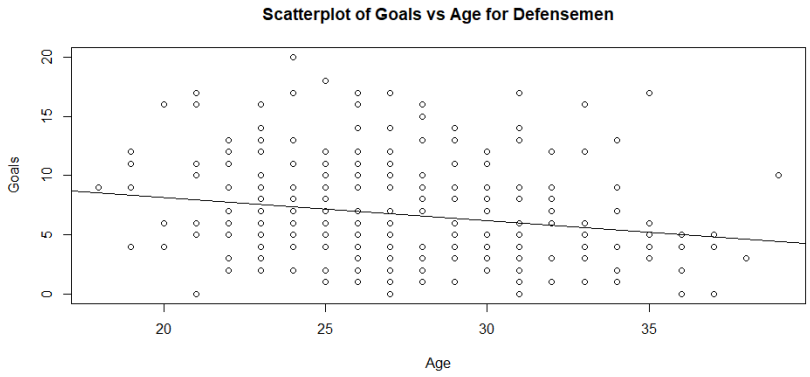
Note: The Sequence Plot for Predictor Variable is not necessary because all data was collected at the same time. The same is true for the Sequence Plot for Residuals.

ii. Diagnostic Plots for Residuals



It appears that there is no pattern between residuals and the predictor variable, confirming a linear relationship between Goals and Age. Also, from the Residuals vs Age plot, the residuals generally appear to have constant variance. One thing to note is the very large residual that is Brent Burns' 29 goal total from the 2016-2017. This is another extreme outlier, even for this player, he has not even reached 17 goals since this season. Additionally, the next closest tally in this dataset is 20 goals. Because of this, I will eliminate this data point from the dataset. The last thing we can see from the scatterplot is that the error terms are independent with Age. Next, from both the Histogram and QQ-Plot, it appears that the error terms are skewed to the right. However, this departure does not appear to be too serious when looking at the QQ-Plot. Once again, the QQ-Plot highlights how much of an outlier Brent Burns' 2016-2017 season was. We can conclude that our assumptions hold, and a linear regression model is a good fit for this data.

After eliminating the two outliers in the data (Brent Burns' 29 goal season and 40-year-old Zdeno Chara), our final regression model appears as the following:



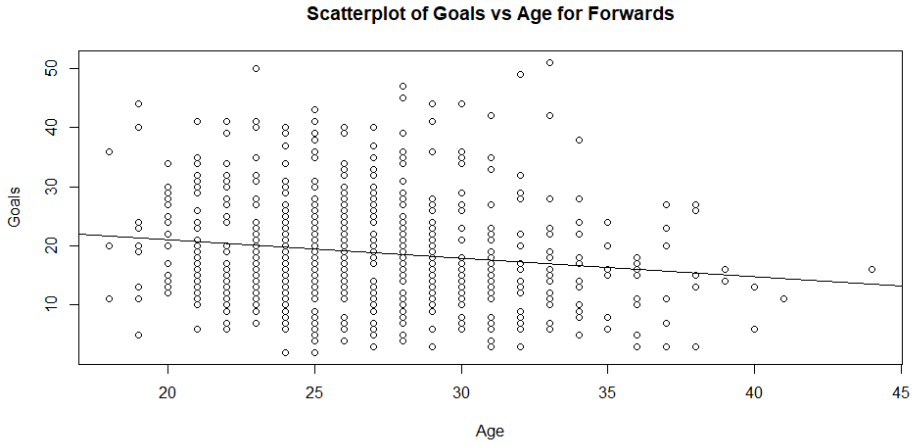
We have b0 = 12.03, Var(b0) = 2.51, b1 = -0.19, Var(b1) = 0.0033. Additionally, our final p-value for b1 is 0.000828. Additionally, we have R-Squared = 0.03639, and the following ANOVA table.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Age	1	206.0	206.048	11.406	0.0008279
Residuals	302	5455.4	18.064		

Our final equation is E(Goals) = 12.03 – 0.19*Age

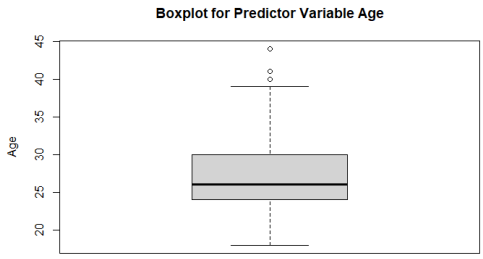
Forwards

After fitting a linear model in R, we see the following regression line visually on top of the Scatterplot of Goals vs Age for Forwards. Here, the regression line is E(Goals) = b0 + b1*Age



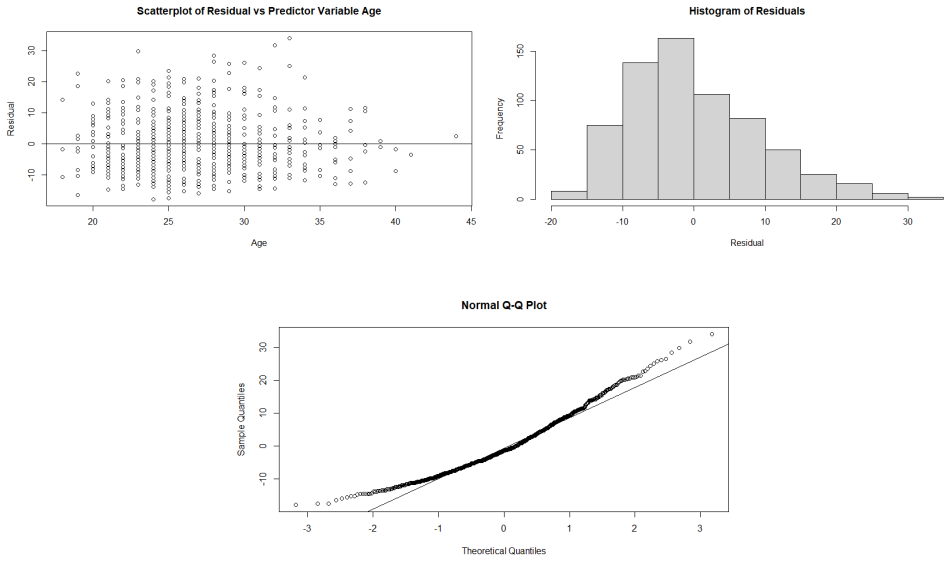
The line of best fit confirms our assumption that there is a potential negative relationship between Goals and Age. After consulting the Summary table, it appears that b1 = -0.317. Additionally, the p-value of b1 is 9.88x10^5, confirming a significant linear relationship between Age and Goals. In summary, we have b0 = 27.452 with Var(b0) = 4.83 and b1 = -0.317 with Var(b1) = 0.007. Since it is confirmed that there exists a relationship between Goals and Age, it is necessary to look at the diagnostic plots.

i. Diagnostic Plots for the Predictor Variable Age



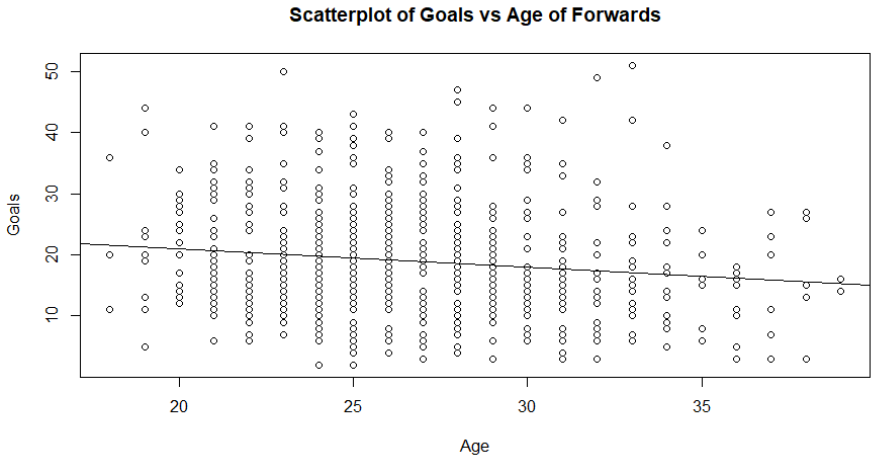
As with Defensemen, it appears that there are some Forwards who are so old that their age appears as an outlier. These players are 44-year-old Jaromir Jagr, 40/41-year-old Matt Cullen, and 40-year-old Shane Doan. It is incredibly rare for forwards to play into their 40's, and even more atypical that they play at least 72 games. Because of this, I will simply not include these players in the final model.

ii. Diagnostic Plots for Residuals



Once again, it appears that there is not a particular pattern between residuals and the predictor variable, confirming a linear relationship between Goals and Age. Also, from the Residuals vs Age plot, even though there is a slight decrease in variance after age 34, the residuals generally appear to have constant variance. The last thing we can see from the scatterplot is that the error terms are independent with Age. Next, from both the Histogram and QQ-Plot, it appears that the error terms are skewed to the right. However, this departure does not appear to be too serious when looking at the QQ-Plot. Just to quickly comment on the largest residuals, the two largest residuals are from Alex Ovechkin in the 2017-2018 and 2018-2019 seasons. Alex Ovechkin is one the greatest goal scorers of all time, a first-ballot Hall of Famer, and currently the only player who has any chance at breaking Wayne Gretzky's career goals record. Even though these are such large outliers, I will keep this data in the model. We can conclude that our assumptions hold, and a linear regression model is a good fit for this data.

After eliminating the predictor variable outliers, the model appears as the following:

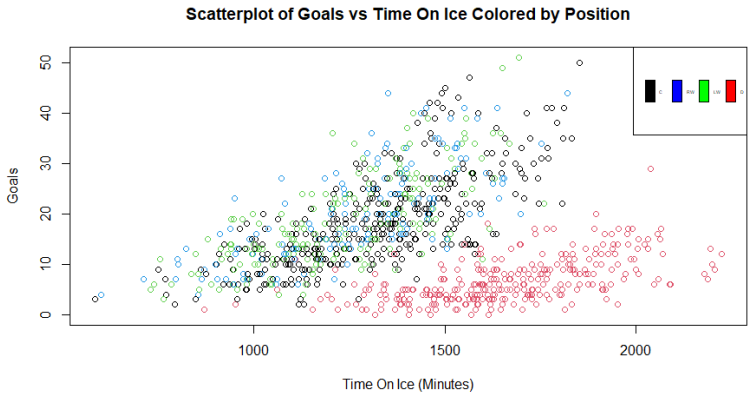


We have b0 = 27.15, Var(b0) = 5.15, b1 = -0.31, Var(b1) = 0.007. Additionally, our final p-value for b1 is 0.000299. Additionally, we have R-Squared = 0.01949, and the following ANOVA table.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Age	1	1132	1132.10	13.217	0.0002989
Residuals	665	56962	85.66		

Our final equation is E(Goals) = 27.15 – 0.31*Age

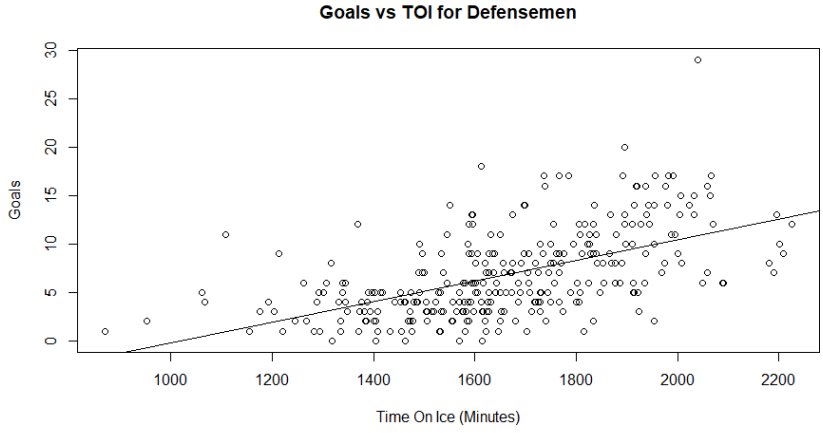
Goals vs Time on Ice



This is the initial plot of Goals vs Time on Ice with the colors partitioning position. Already, there are two distinct relationships: one with Defensemen (red) and one with Forwards. For this reason, Goals vs Time on Ice will be analyzed separately with different models being fitted for Defensemen and Forwards. Regardless, in both relationships it is already easy to see a possible positive correlation between Goals and Time on Ice.

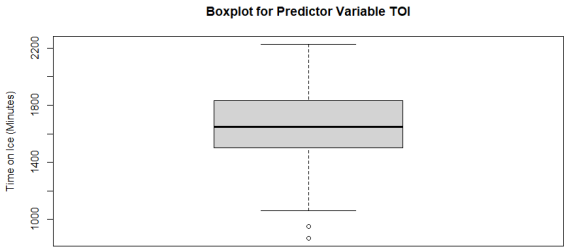
Defensemen

After fitting a linear model in R, we see the following regression line visually on top of the Scatterplot of Goals vs Time on Ice for Defensemen. Here, the regression line is E(Goals) = b0 + b1*TOI



The line of best fit confirms our assumption that there is a potential positive relationship between Goals and Time on Ice. After consulting the Summary table, it appears that b1 = 0.0106. Additionally, the p-value of b1 is less than 2x10^-16, confirming a significant linear relationship between Age and Time on Ice. In summary, we have b0 = -10.79 with Var(b0) = 2.500 and b1 = 0.01061 with Var(b1) = 8.2x10^-7. Since it is confirmed that there exists a relationship between Goals and Time on Ice, it is necessary to look at the diagnostic plots.

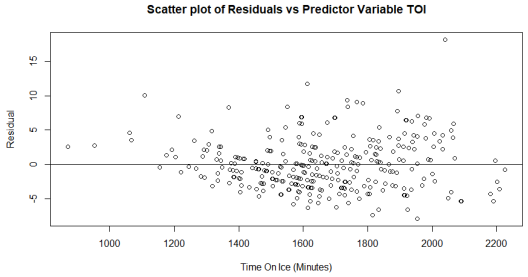
i. Diagnostic Plots for the Predictor Variable Time on Ice

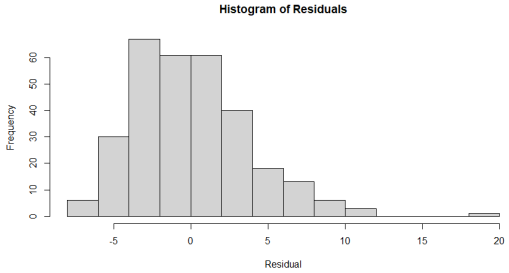


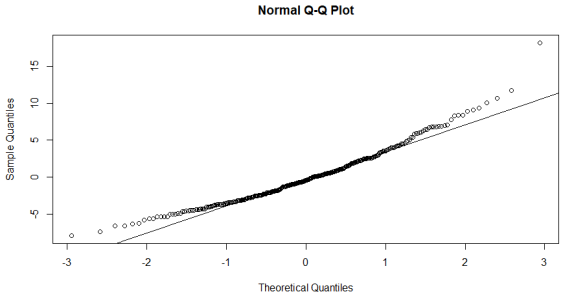
We see that there are few outliers of the predictor variable Time on Ice at 870 and 952 minutes. These two players are Yannick Weber and Scott Harrington, respectively. Because their minutes are so much lower than most defensemen that play at least 72 games in a season, I will not include their data in the final model.

Note: As with the Goals vs Age diagnostic plots, sequence plots are not necessary.

ii. Diagnostic Plots for the Residuals



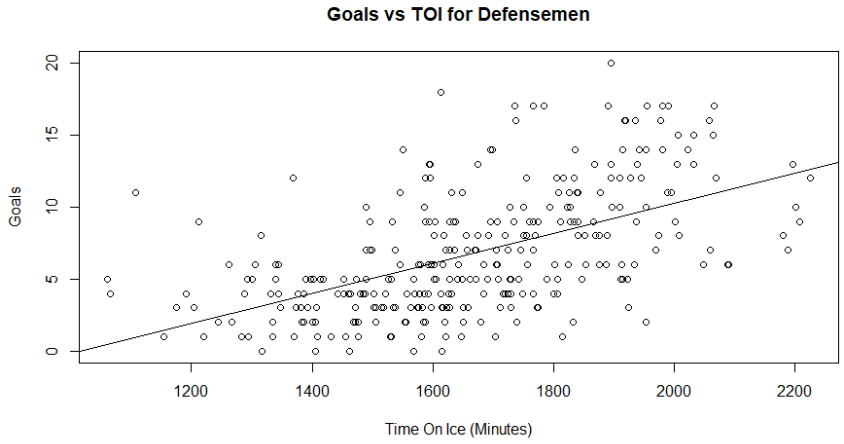




It appears that there is no particular pattern between residuals and the predictor variable, confirming a linear relationship between Goals and Time on Ice. Also, from the Residuals vs Age plot, the residuals generally appear to have constant variance. Like in the Goals vs Time on Ice residual plot, one thing to note is the large residual that is Brent Burns' 29 goal total from the 2016-2017. As explained previously, I will eliminate this data point from the dataset. The last thing we can see from the scatterplot is that the error terms are independent with Time on Ice. Next, from both the Histogram and QQ-Plot, it appears that the error terms are approximately

normal. We can conclude that our assumptions hold, and a linear regression model is a good fit for this data.

After correcting for the outliers in the predictor variable and Burns' goal total, we see the following model:



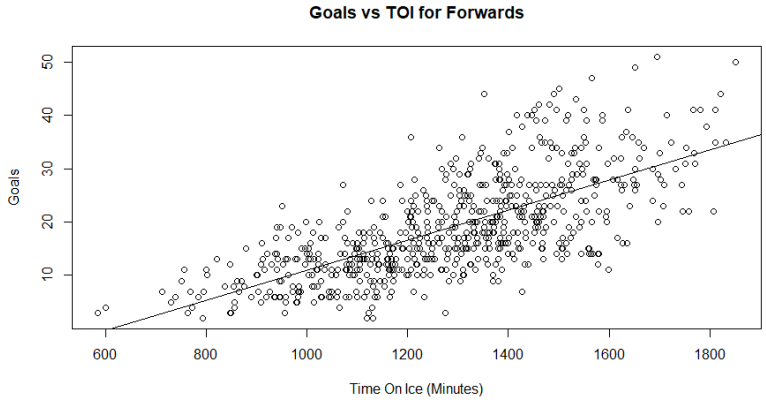
We have b0 = -10.57, Var(b0) = 2.32, b1 = 0.0104, Var(b1) = 8.21x10^-7. Additionally, our final p-value for b1 is less than 2x10^-16. Additionally, we have R-Squared = 0.3056, and the following ANOVA table:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
TOI	1	1712.9	1712.89	132.45	< 2.2e-16
Residuals	301	3892.7	12.93		

Our final model is E(Goals) = -10.57 + 0.0104 * TOI

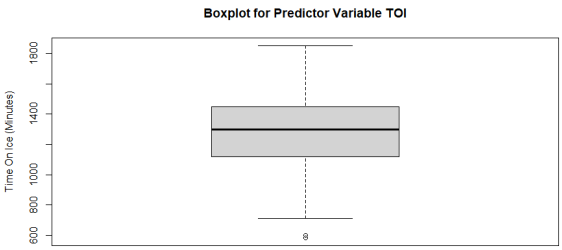
Forwards

After fitting a linear model in R, we see the following regression line visually on top of the Scatterplot of Goals vs Time on Ice for Forwards. Here, the regression line is E(Goals) = b0 + b1*TOI



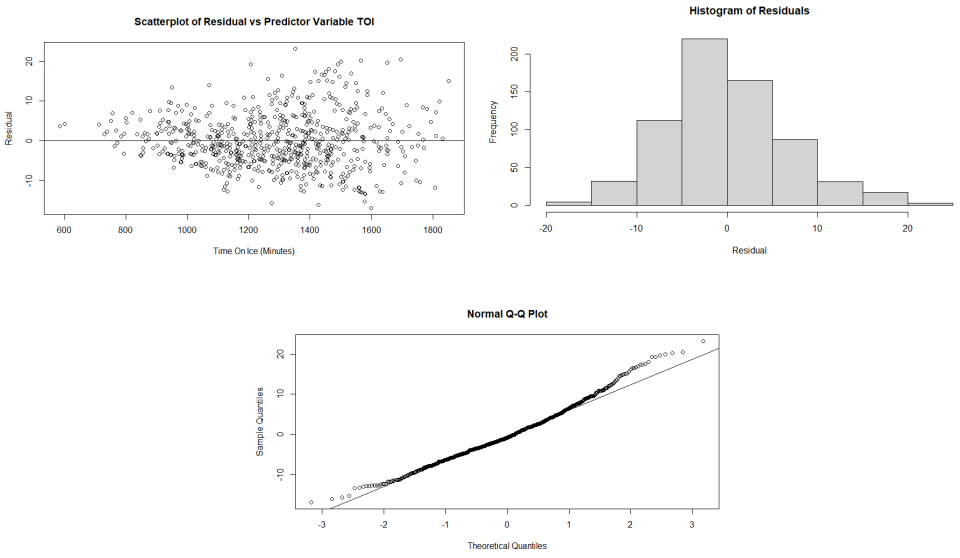
The line of best fit confirms our assumption that there is a potential positive relationship between Goals and Time on Ice. After consulting the Summary table, it appears that b1 = 0.028. Additionally, the p-value of b1 is less than 2x10^-16, confirming a significant linear relationship between Age and Time on Ice. In summary, we have b0 = -17.19 with Var(b0) = 2.13 and b1 = 0.028 with Var(b1) = 1.25x10^-6. Since it is confirmed that there exists a relationship between Goals and Time on Ice, we once again look at the diagnostic plots.

i. Diagnostic Plots for the Predictor Variable Time on Ice



Once again, there are a couple outliers for Time on Ice. These players are Michael Haley and Ryan Reaves who played 584 and 600 minutes, respectively. As with all the other players who have become outliers, these are two unique players. Both players are notable fighters, or players that are mainly used as assets for intimidation rather than skill. In fact, their penalty minutes are 212 and 94 minutes, respectively. Fighters are rarely in the NHL nowadays and for this reason it is acceptable to not include them in the data analysis.

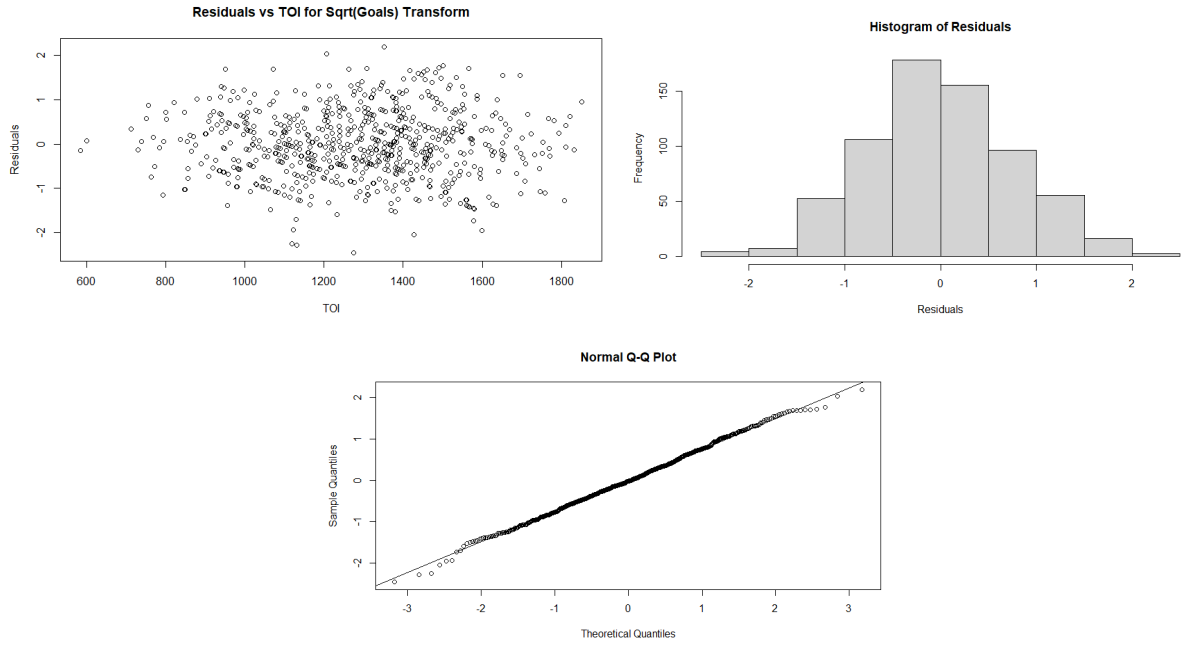
ii. Diagnostic Plots for the Residuals



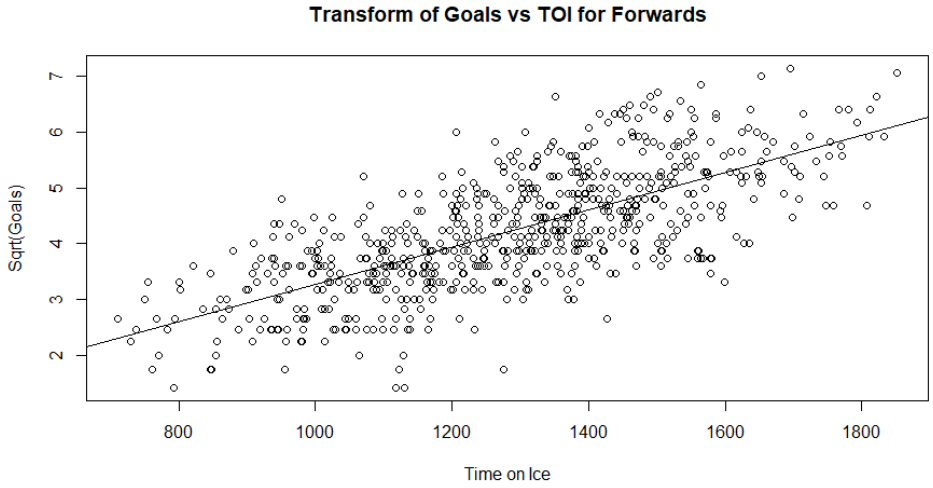
It appears that there is no pattern between residuals and the predictor variable, confirming a linear relationship between Goals and Time on Ice. We also see from the scatterplot that the error terms are independent with Time on Ice. Next, from both the Histogram and QQ-Plot, it appears that the error terms are approximately normal. However, from the Residual vs TOI plot we see that the variance of the error terms is not constant. Specifically, the variance of the residuals increases as TOI increases.

The fix to this issue is to transform our Y which in this case is Goals. After three different transformations (sqrt(Goals), log(Goals), and 1/(Goals)), the best transformation to fit a linear

model is by far sqrt(Goals). Here are the residual diagnostic plots after taking the square root of Goals:



From the first plot, the residuals now have constant variance. Additionally, the residuals maintain no pattern with TOI and confirm a linear relationship between a sqrt(Goals) and TOI. Next, the residuals are independent from TOI. Lastly, the residuals are almost perfectly normally distributed. After removing the predictor variable outliers, our final model of the transformation Sqrt(Goals) vs TOI is the following:



We have b0 = -0.074, Var(b0) = 0.028, b1 = 0.0033, Var(b1) = 1.65x10^-8. Additionally, our final p-value for b1 is less than 2x10^-16. Additionally, we have R-Squared = 0.5039, and the following ANOVA table:

```
                  Df Sum Sq Mean Sq F value    Pr(>F)
datamasterO.TOI    1 389.24  389.24  677.57 < 2.2e-16
Residuals        667 383.17    0.57
```

Our final model is SquareRoot(Goals) = -0.074 + 0.0033*TOI

Goals vs Time on Ice and Age

Lastly, I will attempt to create a multiple linear regression model that will relate the predictor variables Age and Time on Ice to the response variable Goals. Here, I will be analyzing the model Goals = β0 + β1*TOI + β2*Age + ε.

As with the previous models, I will analyze Defensemen and Forwards separately. My reason is that if the simple models for the two predictor variables needed to be separated by Forwards and Defensemen, then so does the multiple linear regression model. Also, I will not be including the outliers as discussed in previous sections into these analyses as they may also skew these models.

Defensemen

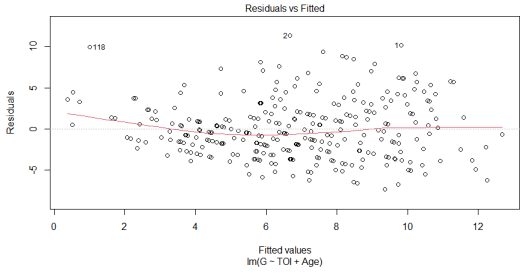
The following is the ANOVA Table obtained from R:

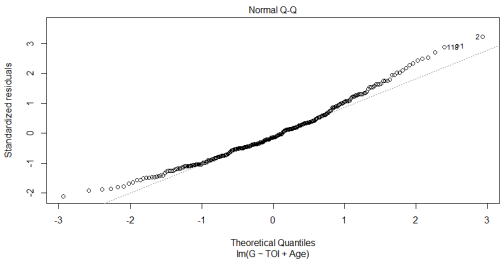
```
           Df Sum Sq Mean Sq F value    Pr(>F)
TOI         1 1712.9 1712.86 139.083 < 2.2e-16
Age         1  210.4  210.42  17.086 4.645e-05
Residuals 299 3682.3   12.32
```

Looking at TOI, we see that the p-value is significant, and TOI should be incorporated into the model. Of course, we already knew this from the previous sections. Next, we see that the p-value for Age is 4.6x10^-5. This means that we should also incorporate Age into the model. In other words, we have confirmed the Alternative Hypothesis that our model should be E(Goals) = b0 + b1*TOI + b2*Age.

From the Summary page, we get the following information: b0 = -5.24, b1 = 0.010, b2 = -0.196 with Var(b0) = 3.87, Var(b1) = 7.81x10^-7, and Var(b2) = 0.002.

We look at the Diagnostic plots and see the following:





From the Residuals vs Fitted plot, we see no pattern between the residuals and the fitted values and can say that the residuals are independent from the fitted values (and hence the prediction variables). Also, there appears to be a near perfect linear fit. Lastly, the variance of

residuals appears to slightly increase, but this increase is not drastic enough to warrant any type of transformation. Additionally, form the QQ-plot, the residuals are slightly skewed right, but this is not a serious skew. It should also be noted that I have already removed the predictor variable and Goal outliers as seen in both Age and TOI in the previous sections.

Therefore, this is a model that fits the Goals to Time on Ice and Age well. Specifically, we have E(Goals) = b0 + b1*TOI + b2*Age. Also, b0 = -5.238, b1 = 0.010, b2 = -0.196 where Var(b0) = 3.87, Var(b1) = 7.8x10^-7, and Var(b2) = 0.002.

Our final model is E(Goals) = -5.238 + 0.010*TOI - 0.196*Age

Forwards

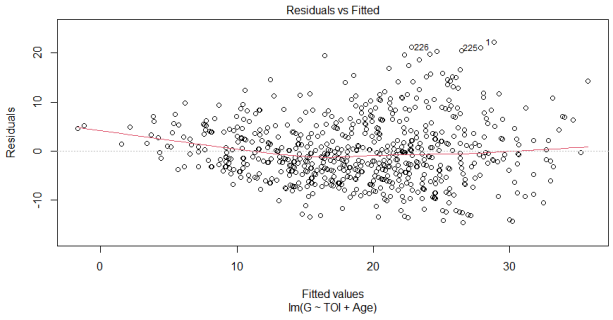
The following is the ANOVA Table obtained from R:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
TOI	1	28371.6	28371.6	649.796	< 2.2e-16
Age	1	832.8	832.8	19.074	1.457e-05
Residuals	668	29166.4	43.7		

Once again, we have a significant value for TOI, so Time on Ice should be incorporated into the model. Additionally, we have a significant p-value for Age, so Age should also be incorporated into the model. We have confirmed the Alternative Hypothesis that E(Goals) = b0 + b1*TOI + b2*Age

From the Summary page, we get the following information: b0 = -10.136, b1 = 0.028, b2 = -0.253 with Var(b0) = 4.68, Var(b1) = 1.22x10^-6, and Var(b2) = 0.003.

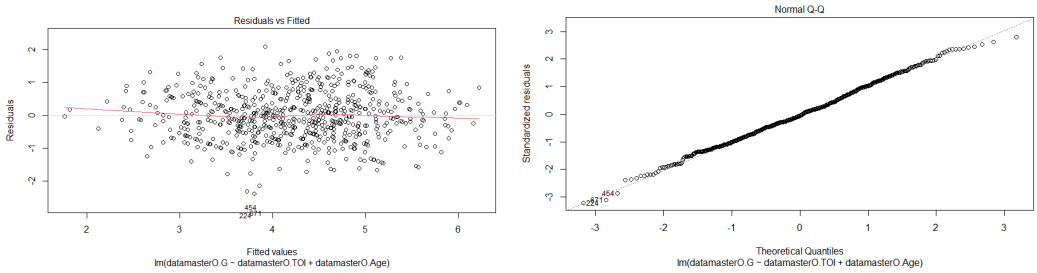
We look at the Residuals vs Fitted plot and see the following:



It looks like the variance of the residuals increases as the fitted value increases. Once again, it is necessary to perform a transformation on Y. The best transform, once again, is taking the square root of Goals. After performing the transform on Y, we see the following ANOVA table:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
datamaster0.TOI	1	400.34	400.34	723.500	< 2.2e-16
datamaster0.Age	1	13.56	13.56	24.504	9.402e-07
Residuals	668	369.63	0.55		

Again, both TOI and Age should be included in the model. As such, we have confirmed the Alternative Hypothesis E(sqrt(Goals)) = b0 + b1*Age + b2*TOI. We now look at the residual plots:



We observe that the variance of residuals is constant, the residuals are independent from the fitted values (and hence the prediction variables), and confirm a linear fit. Additionally, the QQ-plot shows that the residuals are almost normally distributed. It should also be noted that I have already removed the predictor variable outliers as seen in both Age and TOI in the previous sections.

Finally, we have a well-fitting model for sqrt(Goals) scored by forwards that takes into account both Age and Time on Ice. Specifically, we have E(sqrt(Goals)) = b0 + b1*TOI + b2*Age. Also, b0 = 0.824, b1 = 0.0033, b2 = -0.032 where Var(b0) = 0.059, Var(b1) = 1.5x10^-8, and Var(b2) = 4.26x10^-5.

Our final equation is E(Sqrt(Goals)) = 0.824 + 0.00333*TOI – 0.032*Age.

Conclusion

We have finally finished the analysis. In summary, we have fitted models for defensemen and forwards that predict Goals based on Age, Time on Ice, and both Age and Time on Ice. The final model equations are below:

Goals predicted by Age:

Defensemen: E(Goals) = 12.03 – 0.19*Age

Forwards: E(Goals) = 27.15 – 0.31*Age

Goals predicted by Time on Ice:

Defensemen: E(Goals) = -10.57 + 0.0104 * TOI

Forwards: E(Sqrt(Goals)) = -0.074 + 0.0033*TOI

Goals predicted by Time on Ice and Age:

Defensemen: E(Goals) = -5.238 + 0.010*TOI - 0.196*Age

Forwards: E(Sqrt(Goals)) = 0.824 + 0.00333*TOI – 0.032*Age

However, what have we really answered? Well, the first goal of this analysis was to predict when a player will “peak.” The conclusion is that players peak at their youngest age. However, I believe the question of the age at which players peak is extremely complicated. For

one, only using Goals as a metric of peak performance is not fair. If I were to continue this analysis, I would use more metrics such as Assists and Plus/Minus.

The next question is why does Goal scoring have such a simple, downward trend with Age? My personal belief is that the players who make it to the NHL at age 18, 19, or 20 are truly special talents. These players are incredibly gifted and do not need to spend years in the minors developing before they go to the NHL. Instead, they enter the NHL already dominating the sport. These players are incredibly rare, and in my dataset, there are only 45 out of 977 players that are under 21 years old.

Next, I asked whether Time on Ice more accurately predicts scoring. As shown above, scoring increases with player performance. I still believe my hypothesis that the more time on ice a player has, the more scoring chances he has holds. It is important to note that this is a correlation analysis, not a causation analysis. Hence, if I was a coach and wanted to boost a player's scoring, giving him more ice time may not necessarily lead to significantly more goals. However, the question of whether Age or Time on Ice more accurately predicts scoring still stands. This can be done by comparing Adjusted R-Squared. For Goals and Age, we have Defensemen R-Squared of 0.033 and Forward R-Squared of 0.018. For Goals and TOI, we have Defensemen R-Squared of 0.30 and Forward R-Squared of 0.50. For both Defensemen and Forwards, the Adjusted R-Squared is greater in the Time on Ice model. Therefore, Time on Ice more accurately predicts goal scoring.

Further, we created a model to predict Goals based upon Time on Ice and Age. Again, we can compare this model with our other models by looking at the Adjusted R-Squared. We have Adjusted R-Squared of 0.34 for Defensemen and Adjusted R-Squared of 0.53 for Forwards. These are the highest Adjusted R-Squared values we have seen, and therefore, considering both Time on Ice and Age best predicts goal scoring for Defensemen and Forwards.

Lastly, we asked whether position is a confounding variable that must be taken into account. The clear answer is yes. This is mostly influenced by the fact that in general, forwards score way more goals than defensemen. For this reason, separate models were needed for Forwards and Defensemen in each case.

Aside from predicting expected performance of a given player, another application of these models is using prediction intervals to analyze whether a player is overperforming or underperforming. For example, if there is a Defensemen whose Time on Ice last year was 1700 minutes, then a 95% prediction interval of my model predicts they will score between 0.06 and 14.24 Goals. If this player has actually scored 16 goals, then they are overachieving and have highlighted themselves as a valuable asset.

Overall, I really enjoyed doing this project and hope that my insights are clear. I apologize for this being lengthy, but I believe the length is necessary in order to be as thorough and accurate as possible. I believe models like these are incredibly useful for use in the NHL or really any sport. Thank you for reading!