Linear Regression Hockey Analysis

Bradley Behan, Derek Lasker, Abraham Rueckert, Zachery Velthoven

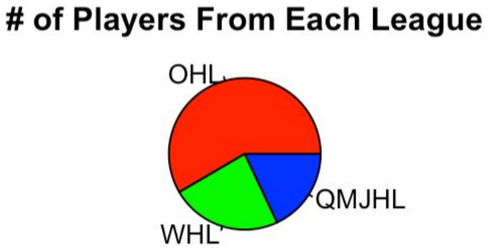
Special Thanks to: Dr. Albert Cohen, Trevor Nill, and Nickolas Gatt

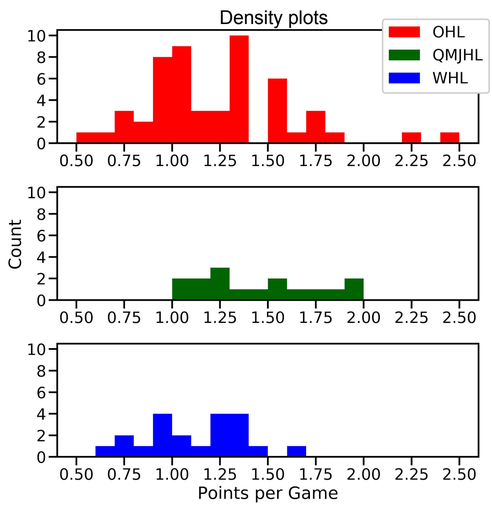
Introduction

Predicting player performance is quickly becoming an important function in hockey all over the world. In this poster we will use linear regression to predict the average points per game over a players first three NHL seasons. Our analysis includes forwards from the Canadian Hockey League, a conglomerate of 3 Canadian leagues (OHL, QMJHL, and WHL).

Data

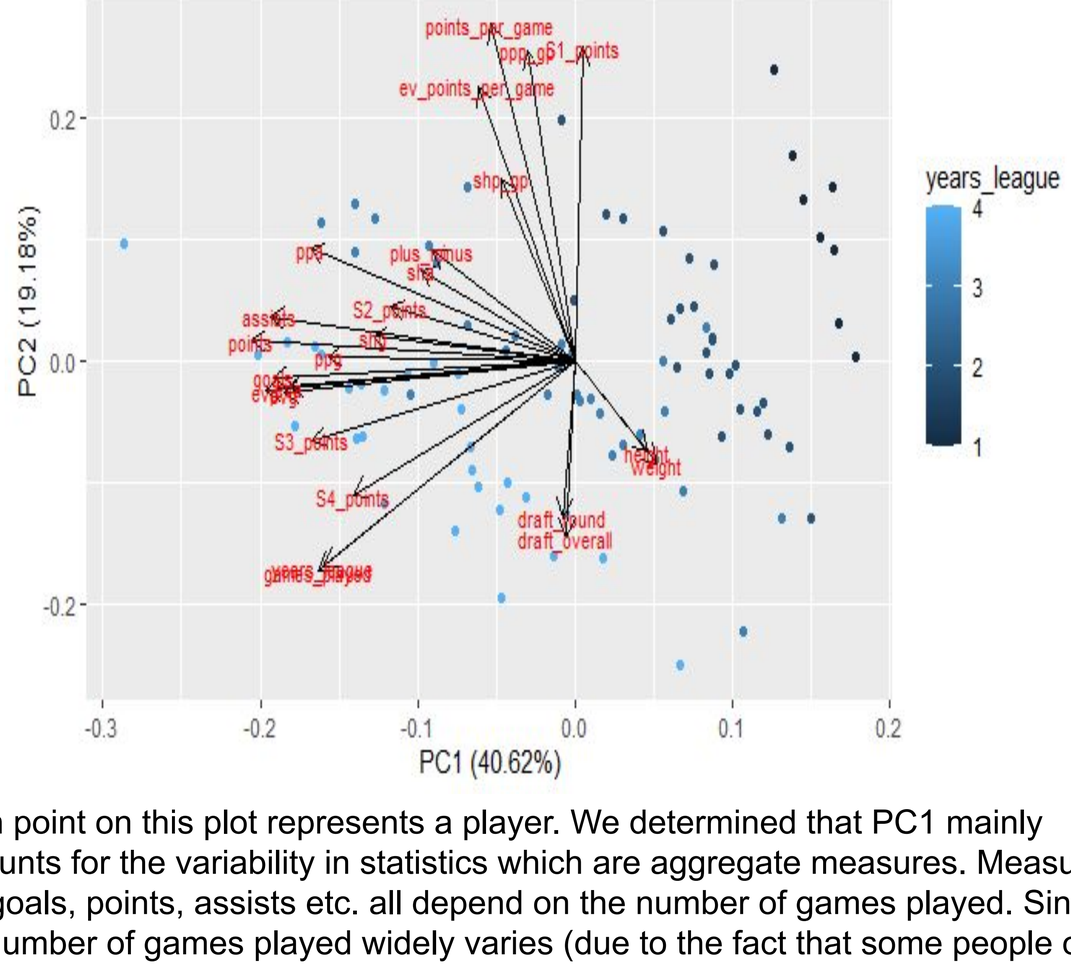
- Possible players included in this study are limited to those who completed the 2018/2019 NHL season
- The player must have only played in the CHL for any number of years before transferring permanently into the NHL/AHL
- To reflect the average NHL rookie contract, the player must have participated and completed 3 seasons in the NHL with at most 1 complete season (≥20 games) with one split season in the AHL before entering the NHL
- The 2020 season can count as a 3rd season completed, if necessary, because of early season termination due to COVID-19
- 2004-05 NHL Lockout season was disregarded
- Other generic rules also applied





Principal Component Analysis

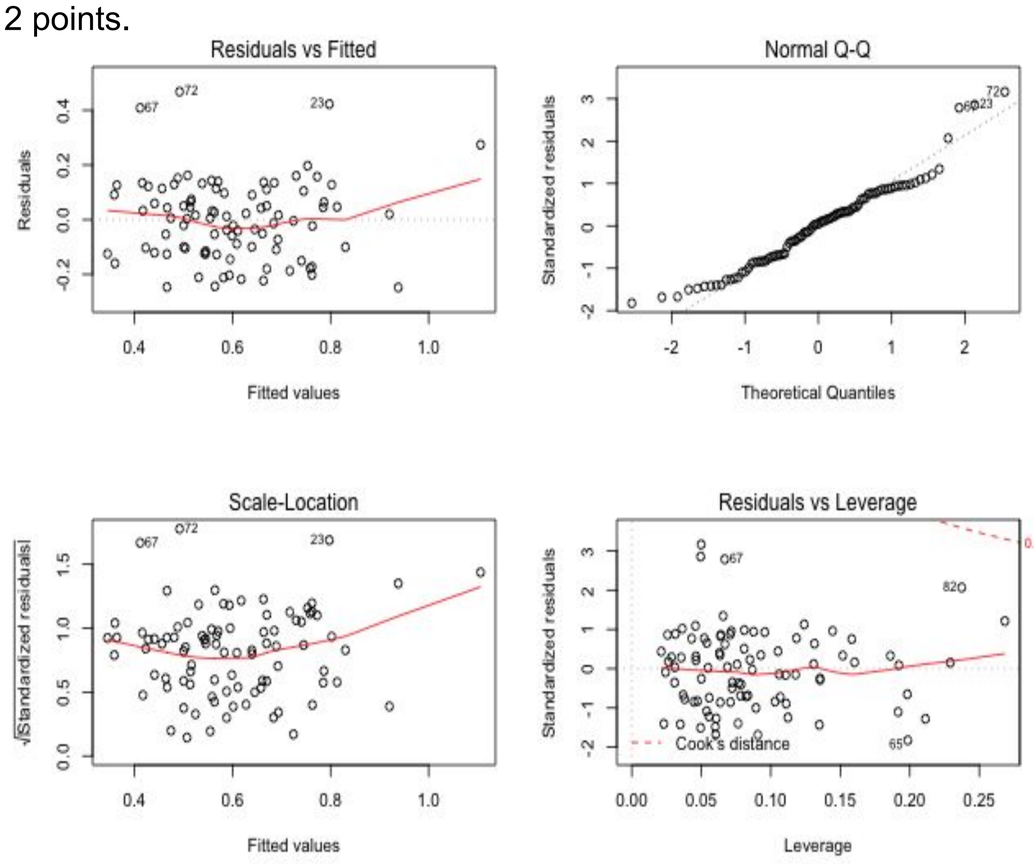
Principal component analysis is a technique which refines a data set with many variables. When dealing with a large dataset, it can be difficult to determine which variables matter most. Which observations best describe the data? Which ones produce the most variability? Are there groupings in the data and similarities among some of the data points?



Each point on this plot represents a player. We determined that PC1 mainly accounts for the variability in statistics which are aggregate measures. Measures like goals, points, assists etc. all depend on the number of games played. Since the number of games played widely varies (due to the fact that some people only play 1 year in juniors and some play 4) the aggregate measures will have the most variance. On the other hand, PC2 accounts for the variability in most of the per-game statistics. These are the main numbers which determine how good a player is and they do not depend on games played as much.

Linear Regression

The method we chose to make our predictions was linear regression. We ran several regressions with varying results. Our best model utilized Backward Stepwise regression. The model achieved a 0.478 multiple R-squared, our best result. With our next best result coming from a regression that composed of the highest and lowest outliers. The response variable was points per game over the first 3 NHL seasons. The predictors were assists, powerplay goals, even strength assists, plus minus, and season 2 points.

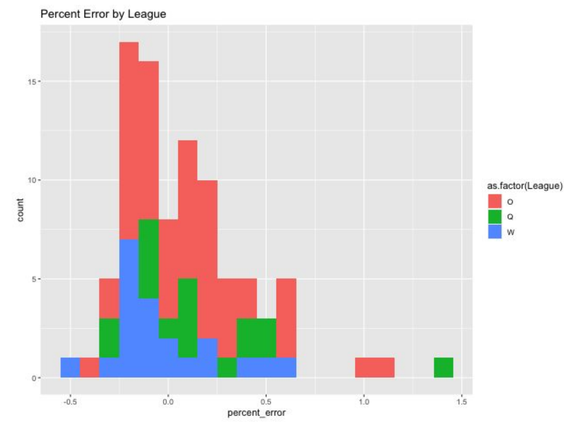


Below is a summary table of all the regressions we completed.

Regression	Significant Predictors (Level of Stat. Significance)	Multiple R-Squared
Original	games_played(***), goals(*)	0.3825
0.5 to 0.9 ppg	games_played(.)	0.1068
Low (1.0-1.4 points per game)	None	0.2163
High (1.0-1.4 points per game)	games_played(**), goals(*), assists(*)	0.4108
Extreme (1.0-1.4 points per game)	games_played(***), goals(*), assists(*)	0.4612
Backward Stepwise	assists(**), ppg(***), eva(**), ppa(*), plus_minus(.), shp(*), s2(*)	0.478

Conclusion / Future Topics

Before conducting our analysis of the regression, we noticed that players in the WHL had lower point-per-game rates on average in comparison to the other two leagues. After using the regression to predict the NHL points-per-game rates for those players and comparing those to the actual rates those players produced, we analyzed the percent errors across the three leagues. The percent errors showed that on average, WHL players were unpredicted, QMJHL players were overpredicted, and OHL players were accurately predicted.



GIven hockey's chaotic, random nature, we determined that the multiple R-squared value of 0.478 was strong, even though that may be considered weak in comparison to other statistical models. We also concluded that this model is not a definitive evaluator of overall player talent, rather an estimator of potential scoring talent once that player reaches the NHL. There were three main factors that our model could not account for due to data limitations: player age, team strength, and player ice time. This model, in conjunction with other statistics, can help teams make decisions on who to draft and where to play them in their lineup. A great deal of future topics could be explored, such as expanding the dataset criteria and analyzing past draft success for both junior and NHL teams.

Scan the code to see the full project on our website!