Andrew Curtin
CS 435
Final Project Report
December 7, 2016

Career Save Percentage in NHL Goalies (Measuring Goalie performance)

Introduction

The evaluation of goaltenders' performance faces significant challenges beyond similar quantification for forwards and defenders. To mark a forward's performance by scoring output is fairly simple, and, although more intricate analysis can become complicated, it is generally easy to connect scoring proficiency to future success. Defenders enjoy a similar simplicity of analysis. Although quantitatively depicting the skillfulness in negating scoring chances is difficult, their offensive output can be measured similarly to the forwards, and their overall impact consolidated from these figures. By virtue of the extreme difference in the position and certain aspects of hockey gameplay, measuring a goaltender's specific contribution statistically is considerably more difficult. A goaltender must rely on their teammates to score, and can't be singularly responsible for a winning outcome, barring a few extremely rare situations. Likewise, a goaltender can have a terrible performance, but be bailed out by the play of their teammates at the other end of the ice. A goaltender's statistical outcomes can be heavily influenced by the play of their teammates, disparities in skill between their team and the opposition, the style of play in any given game, and random factors common to hockey. Pucks in the enclosed space can deflect unpredictably and create situations that make saves more difficult. The speed at which the NHL game is played adds additional complexity to goaltending.

The difficulty of evaluation is apparent in the player selection patterns of NHL general managers. In the first round of the NHL Entry Draft, where teams select young players first eligible for the NHL, managers have the best opportunity to select talented players. In spite of this, with 30 total selections in each round, an average of 1.1 total goalies were selected from 2006 to 2016, with only 1 goalie chosen from 2013-2016 out of 120 selections. NHL general managers have not been willing to risk a first round selection on a goaltender, despite the necessity of strong goaltending in the modern game.

With these challenges in mind, a goaltender's save percentage offers the most convenient singular statistic for estimating an individual's contribution to a game. Single game save percentages can fluctuate wildly due to the factors listed above, but, because shot events are more frequent than goals scored, a goaltender's ability to stop shots expressed as a percentage of total shots faced tends to stabilize over time. Individual season save percentages, which are essentially just save percentages over a particular slice of games, can reflect a number of factors internal and external to individual performance. The career save percentage reflects these things as well, but in a large enough sample size over a period of time long enough to cover the best and worst performances of a goalie as they improve and later decline with age. For this reason, I've chosen to represent individual performance as career save percentage.

Accordingly, the goal of this study is to investigate career save percentage in terms of a number of arbitrary factors that are commonly presented as part of a goaltender's career statistics. The first step will be an exploratory analysis, to get a handle on broader properties of the data, followed by predictive analysis, to see if any other attributes have predictive value for projecting a goaltender's career save percentage.

Data and Preprocessing

All data used in this study was scrubbed from www.hockey-reference.com, from the goalie registry, found at www.hockey-reference.com/players/*/goalies.html, where * indicates a letter from "a-z." I wrote a Python script, using Requests and BeautifulSoup, to parse the HTML of the web page and gather appropriate data values into a CSV. The NHL began to officially track save percentage in the 1984 season. In order to qualify, a goalie must have played while save percentage was officially tracked and have appeared in at least 80 games. 80 was chosen as an arbitrary representation of the minimum amount of games played by a starting goalie over two seasons. This was chosen to ensure that each goalie in the study had a large enough body of work to be statistically significant.

Because there were goalies in the study who had played part or most of their career before 1984, for which their save percentage was not documented, I chose to have the percentage in their documented years be representative of their career output. With these years occurring toward the end of these goalies' careers, it is likely this measure underestimates their actual career save percentage, but this is only relevant to the extent that a large number of games played or minutes might be representative of save percentage.

227 goalies qualified according to the above criteria. For the 227 observations, 43 attributes were recorded. Beyond the ID and name of the goalie, there were 16 statistical attributes followed by signifiers (1 for "yes," 0 for "no") of whether a specific Hall of Fame defender had played with that player during any season. The list of Hockey Hall of Fame defenders was obtained from https://en.wikipedia.org/wiki/List_of_members_of_the_Hockey_Hall_of_Fame#Players, and included only defenders who had played during the years in which save percentage was recorded, in addition to defenders currently playing who were likely to be inducted upon retirement. There were 25 such players, and therefore 25 signifiers.

The 16 statistical attributes for each goalie were as follows:

Year of first season: range: [1969, 2014]

Year of final season: range: [1984, 2017], with 2017 representing currently playing goalies.

Age of first season in years: [18, 33]

Age of final season in years: [24, 43]

Height in cm: [165, 200]

Weight in kg: [68, 108]

Games played: [81, 1266]

Wins: [19, 691]

Losses: [28, 397]

Total shots faced: [160, 31709]

Save percentage as a frequency: [0.811, 0.925]

Best single season save percentage: [0.811, 0.940]

Minutes played: [4217, 74439]

There were three additional variables gathered: good starts, representing the number of games in which a goalie outperformed that season's mean save percentage, bad starts, games in which they underperformed the season mean, and the ratio of good to bad starts. Because this data was not gathered until 2009, very few goalies had measurements of these variables, so they were not included in any analysis.

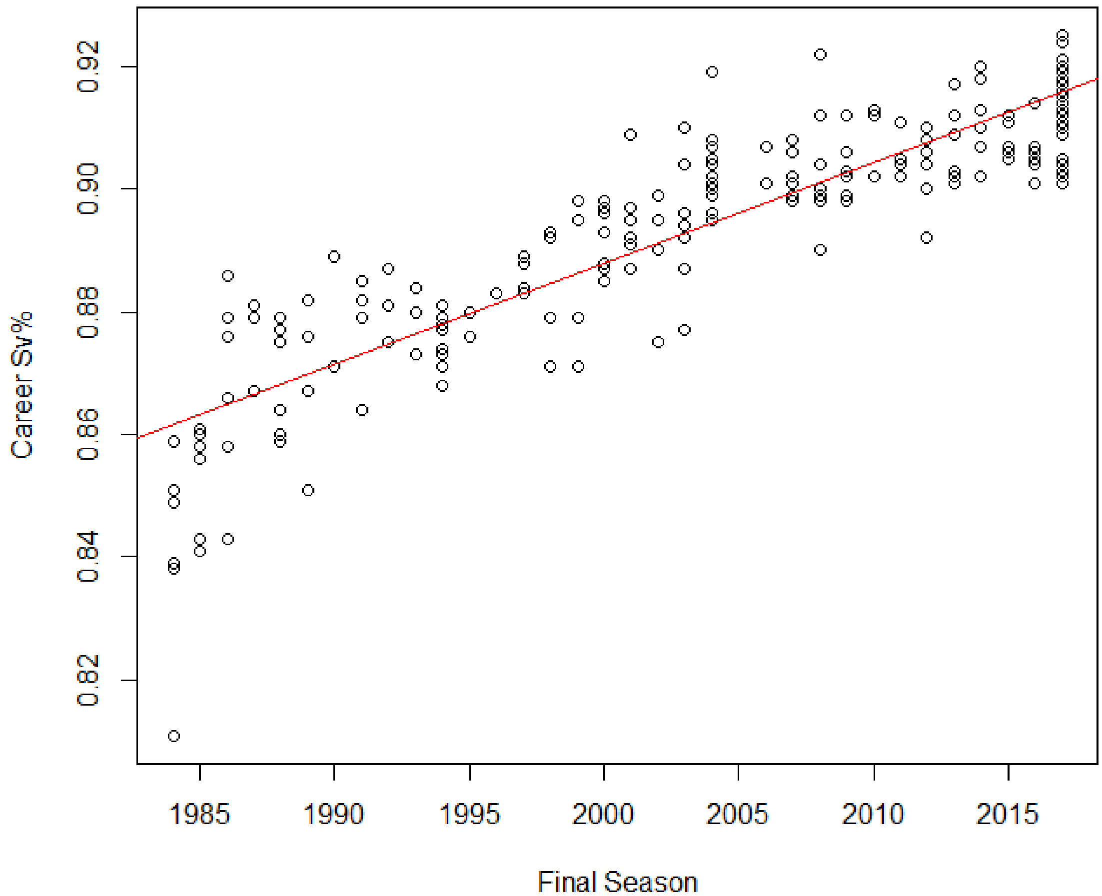
The 25 defender matrix was aggregated into a single total for each goalie, of the number of unique elite defenders they'd played with during their career. Later, clustering was performed to group goalies by the years in which their career took place. From the clusters, z-scores for save percentage were calculated. This will be discussed in more depth in the proceeding section.

Exploratory Analysis

Below is a plot of career save percentage by final season, with a regression line overlaid (Fig. 1). The R readout for the regression line is as follows:

```
 Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.392e+00  1.085e-01  -22.06   <2e-16 ***
goalies$to   1.640e-03  5.412e-05   30.31   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.008835 on 225 degrees of freedom
Multiple R-squared:  0.8033,	Adjusted R-squared:  0.8024
F-statistic: 918.6 on 1 and 225 DF,  p-value: < 2.2e-16
```

The linear fit of final season to career save percentage was immediately noticeable, with an R^2 value of 0.8033 with a simple linear regression, as well as residual standard error of roughly one tenth of a percent.



(Fig 1.)

Multiple regression over all relevant factors, removing the least significant factor each time, produced this model:

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)             -1.153e+02  2.143e+01  -5.380 1.89e-07 ***
goalies$to               5.692e-02  1.063e-02   5.355 2.13e-07 ***
goalies$from             5.776e-02  1.083e-02   5.332 2.38e-07 ***
goalies$min              4.377e-07  7.038e-08   6.218 2.47e-09 ***
goalies$to:goalies$from -2.829e-05  5.370e-06  -5.269 3.25e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.007872 on 222 degrees of freedom
Multiple R-squared:  0.8459,	Adjusted R-squared:  0.8431 
F-statistic: 304.6 on 4 and 222 DF,  p-value: < 2.2e-16
```

An interaction term between first season and final season produced better results, as well as including minutes played. These were the variables found to be most statistically significant. The high correlation to the years in which a goalie played is interesting, but these are not variables that describe player performance. What they do indicate is that, over time, save percentages have risen overall in the NHL from 1984 to the current season.

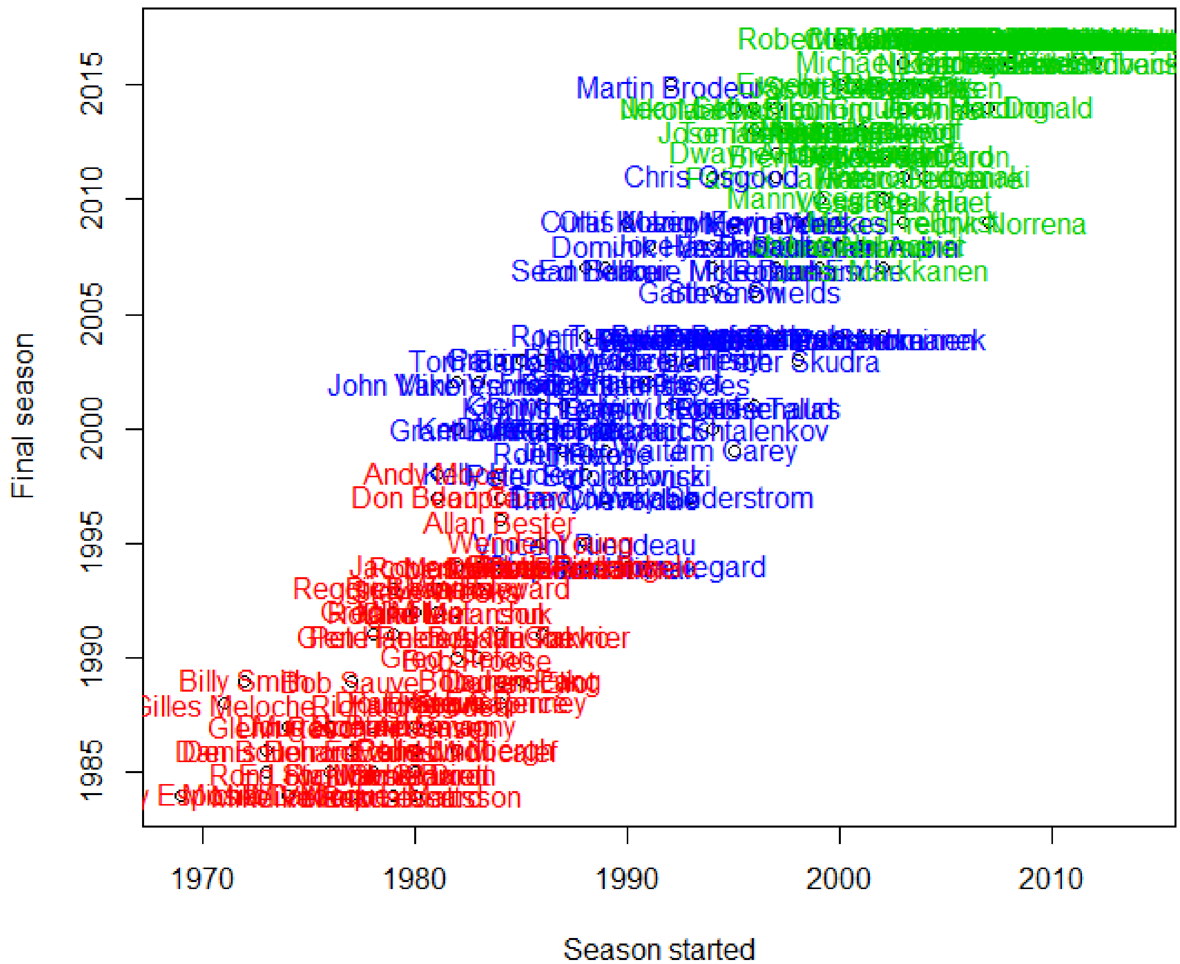
In order to look beyond this, I used k-means clustering in R on the first and final season attributes to see if the overall trend towards rising save percentages could be adjusted out of performance. I clustered on the season attributes using the following R code:

```
set.seed(1)
grpGoalies <- kmeans(g_subset[,c("from", "to")], centers = 3, nstart = 10)
```

The results of clustering:

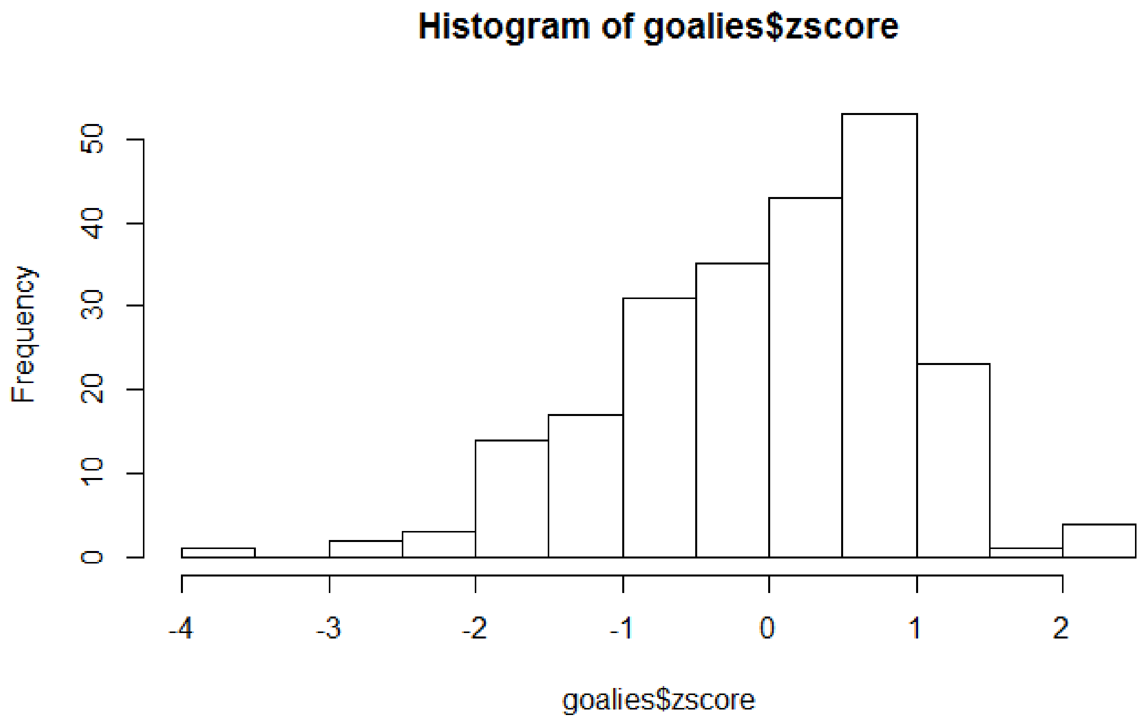
```
K-means clustering with 3 clusters of sizes 61, 93, 73

Cluster means:
      from       to
1 1980.230 1989.410
2 2005.323 2014.774
3 1991.466 2002.932
---

Within cluster sum of squares by cluster:
[1] 2013.541 2828.581 2794.822
 (between_SS / total_SS =  86.2 %)
```

I ran further trials with more clusters, which increased the $between_{SS}$ to $total_{SS}$ ratio, but produced less intelligible clusters, with overlapping ranges and combinations of means that didn't make intuitive sense for separating the goalies by a range of time. The cluster means can be seen in Fig. 2 below. Goalies with long careers are the most difficult to classify, but, overall, this method produced intelligible slices of the data. Cluster 1, in red, are the earliest goalies that made the cutoff for the study. Cluster 2, in green, includes goalies playing through the present season. Cluster 3, in blue, represents the years in-between. The initial season mean of Cluster 2, 2005.323, is noteworthy. The NHL had lost a season to labor disputes the prior year, and returned with a different interpretation of the rule-set in an attempt to increase scoring. Having the impacts of these rule changes confined to a single cluster is a good outcome, in this case.

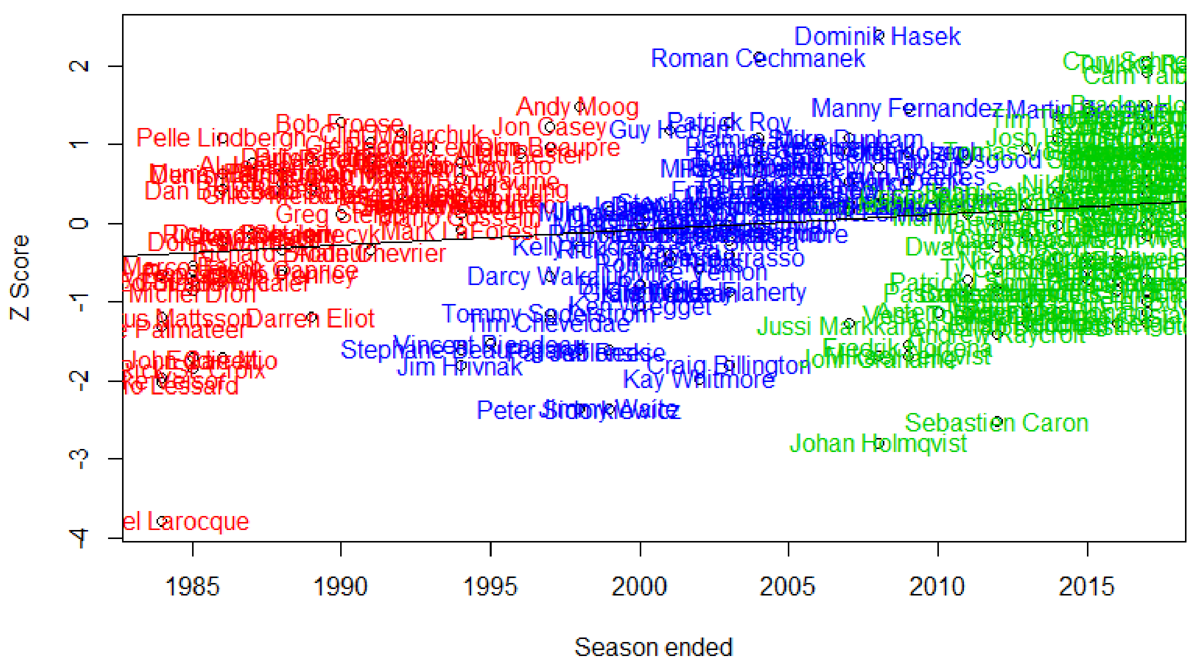


(Fig. 2)

With the goalies clustered, I used R to produce z-scores for career save performance within each cluster, to compare goalies based on how they performed within their peer groups. Plotting z-score with respect to season ended produced a less distinct relationship (Fig. 4).


Histogram of goalies$zscore
Frequency
0
10
20
30
40
50
-4
-3
-2
-1
0
1
2
goalies$zscore

(Fig. 3)


Z Score
Season ended
1985
1990
1995
2000
2005
2010
2015
2
1
0
-1
-2
-3
-4
Dominik Hasek
Roman Cechmanek
Andy Moog
Bob Froese
Pelle Lindbergh
Jon Casey
Patrick Roy
Manny Fernandez
Greg Stefan
Darcy Wakaluk
Darren Eliot
Michel Dion
Tim Cheveldae
Jim Hrivnak
Craig Billington
Kay Whitmore
Jussi Markkanen
Sebastien Caron
Johan Holmqvist

(Fig 4.)

Predictive Analysis

Having partitioned the goalies according to the years they played and calculated their performance relative to their peers, I then attempted to build a model using linear regression to predict a goalie’s z-score based on their non-save percentage, non-id attributes. These would be the other readily available statistical attributes a general manager could try to use from simple box score stats. I built a model to predict z-score using the age at start and end of career, height, weight, number of shuts faced, games played, wins, losses, and minutes played. The full data set was partitioned 50% training, 50% testing. The model statistics follow:

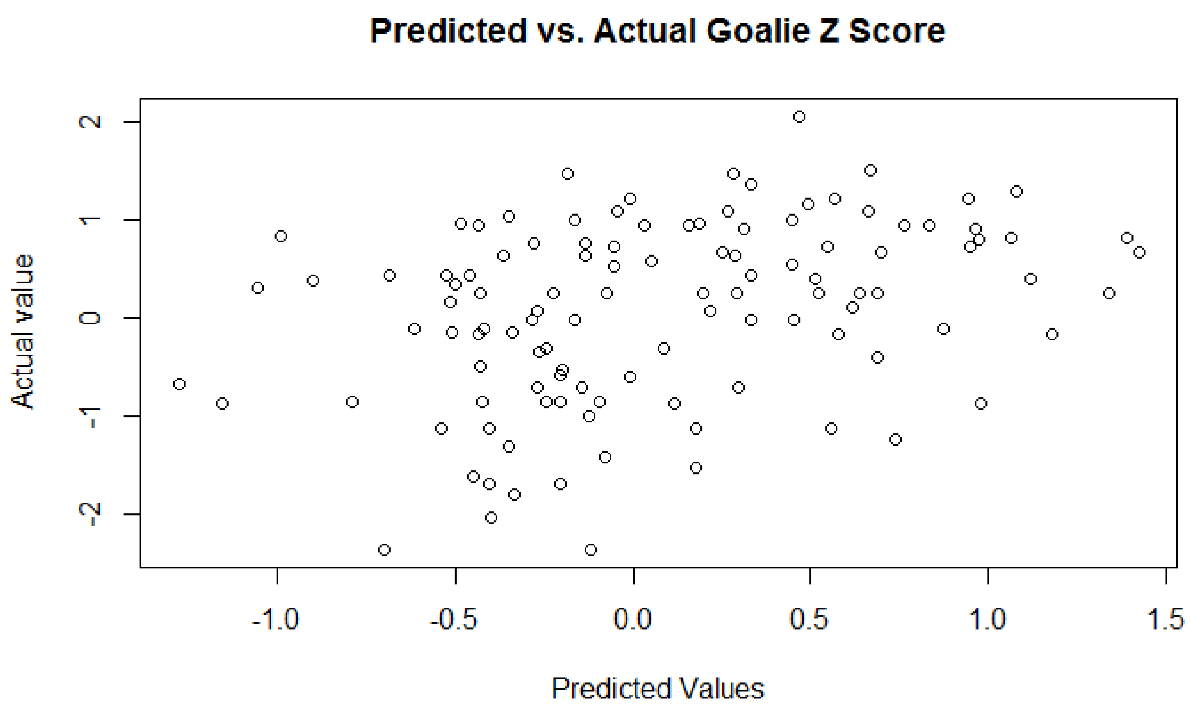
```
lm(formula = zscore ~ age_from + age_to + height_cm + weight_kg +
    gp + w + l + total_shots + min, data = training.set)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2971 -0.6846  0.1036  0.6836  1.6023

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.451e+00  2.887e+00  -0.849  0.39803
age_from     7.222e-02  4.424e-02   1.632  0.10564
age_to      -1.446e-01  4.344e-02  -3.328  0.00121 **
height_cm    3.068e-02  1.908e-02   1.608  0.11081
weight_kg   -1.430e-02  1.614e-02  -0.886  0.37776
gp          -1.011e-03  1.395e-02  -0.072  0.94237
w            1.050e-03  1.321e-02   0.079  0.93682
l           -8.209e-03  1.388e-02  -0.592  0.55548
total_shots  7.380e-05  4.091e-05   1.804  0.07412 .
min          9.339e-05  2.873e-04   0.325  0.74575
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9164 on 103 degrees of freedom
Multiple R-squared:  0.3225,  Adjusted R-squared:  0.2633
F-statistic: 5.449 on 9 and 103 DF,  p-value: 4.201e-06
```

Age at end of career and total shots faced were the most statistically significant factors. A more succint model with only age at end of career, total shots, and an interaction term between minutes and games played produced:

```
lm(formula = zscore ~ age_to + total_shots + min * gp, data = training.set)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2564 -0.6167  0.0957  0.5737  2.1701
```

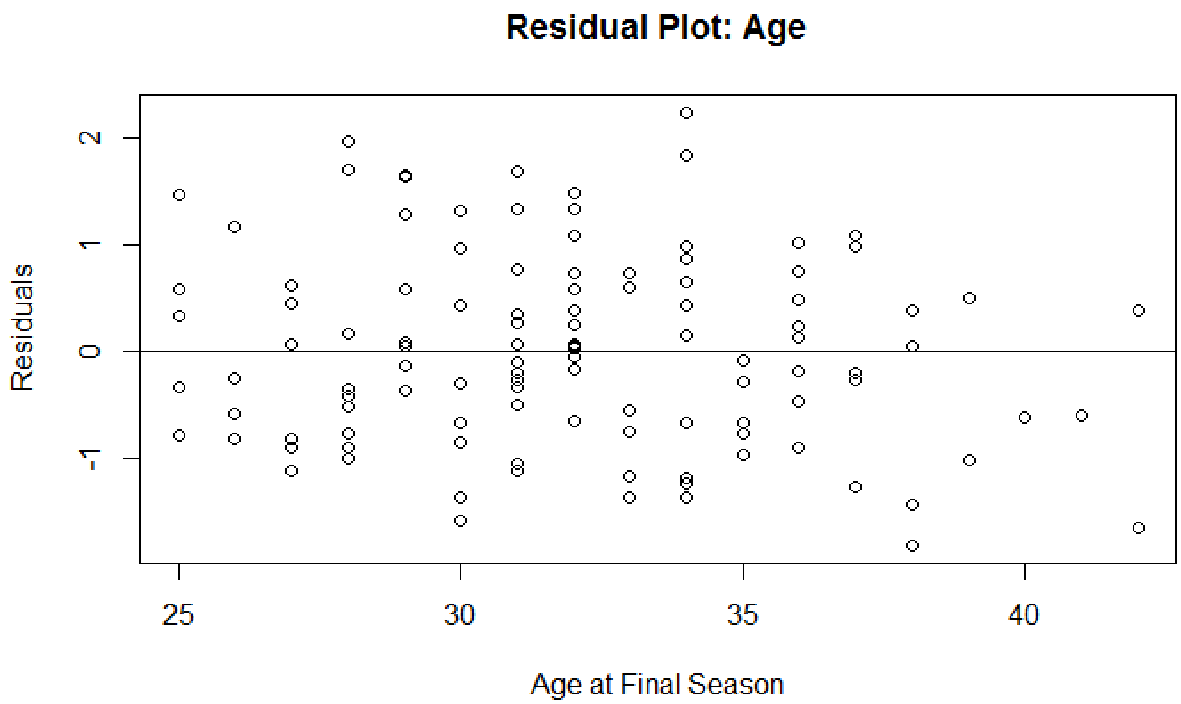
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.700e+00  9.639e-01   1.764  0.08062 .
age_to      -8.543e-02  3.355e-02  -2.547  0.01230 *
total_shots  1.100e-04  3.675e-05   2.992  0.00344 **
min          5.482e-04  1.959e-04   2.798  0.00610 **
gp          -2.845e-02  1.127e-02  -2.525  0.01302 *
min:gp      -6.975e-08  3.509e-08  -1.988  0.04938 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.923 on 107 degrees of freedom
Multiple R-squared:  0.286,	Adjusted R-squared:  0.2526
F-statistic: 8.572 on 5 and 107 DF,  p-value: 7.427e-07
```

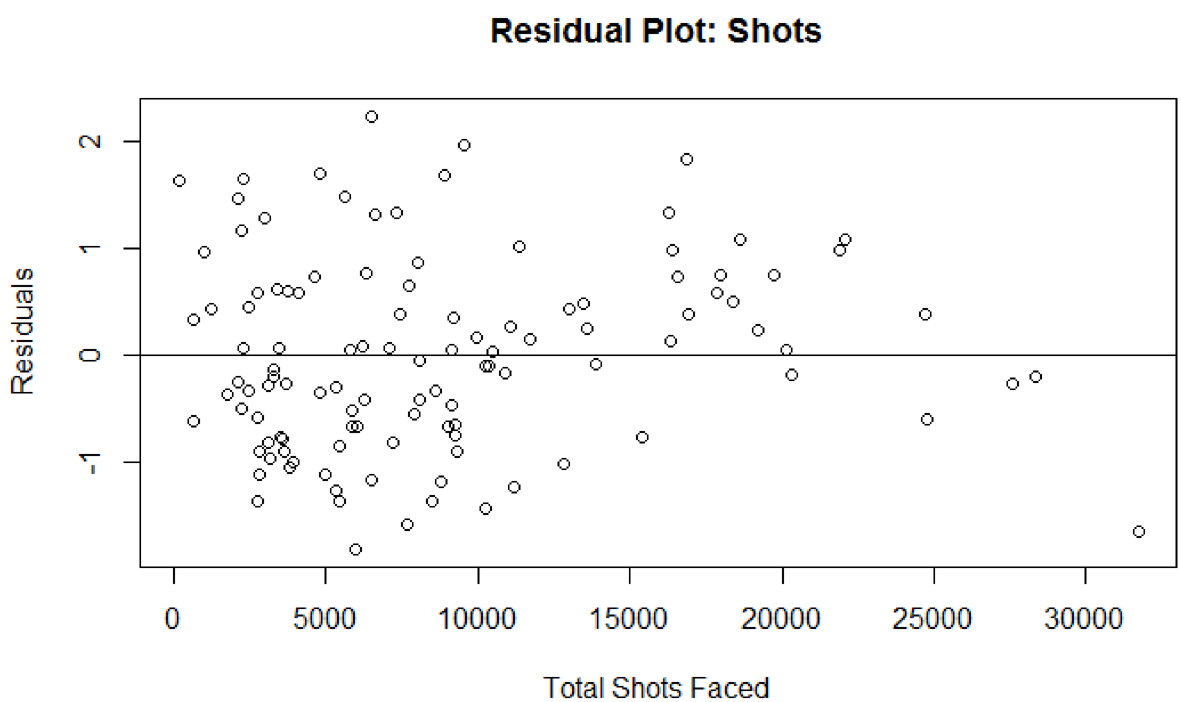
Using the second model to predict the values of the test set, the residual sum of squares over 114 predictions was quite large at 91.2. A plot of the predicted values and the actual values shows the discrepancies between the real and formula-generated values (Fig. 5).

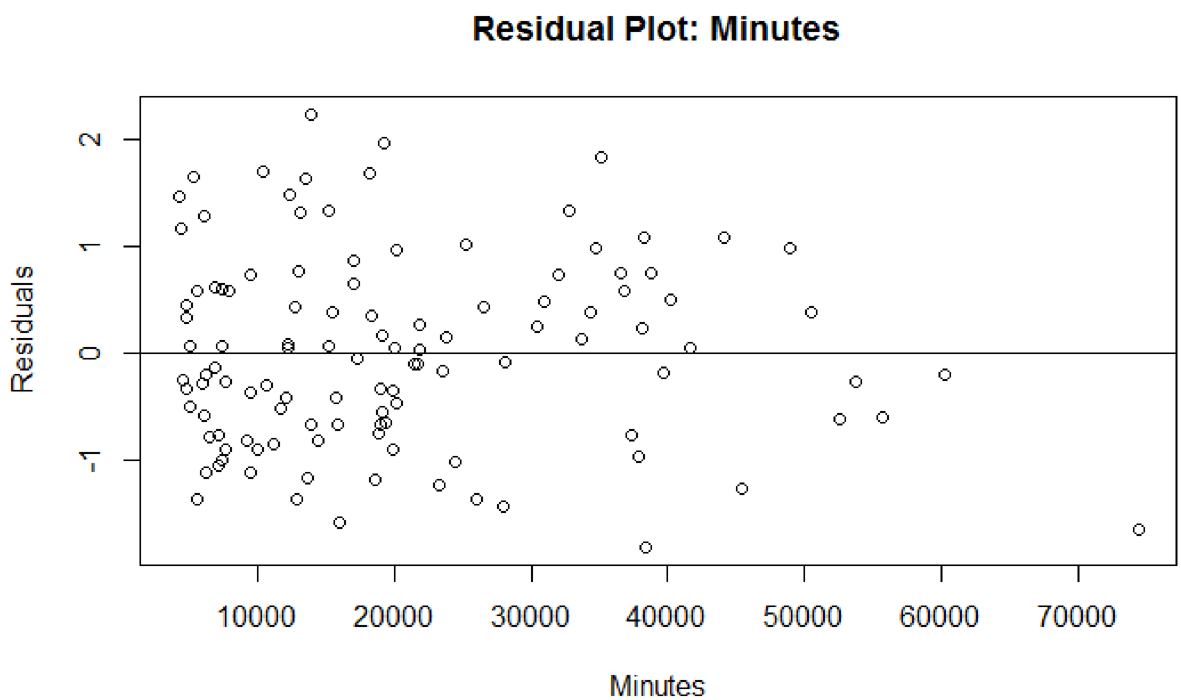


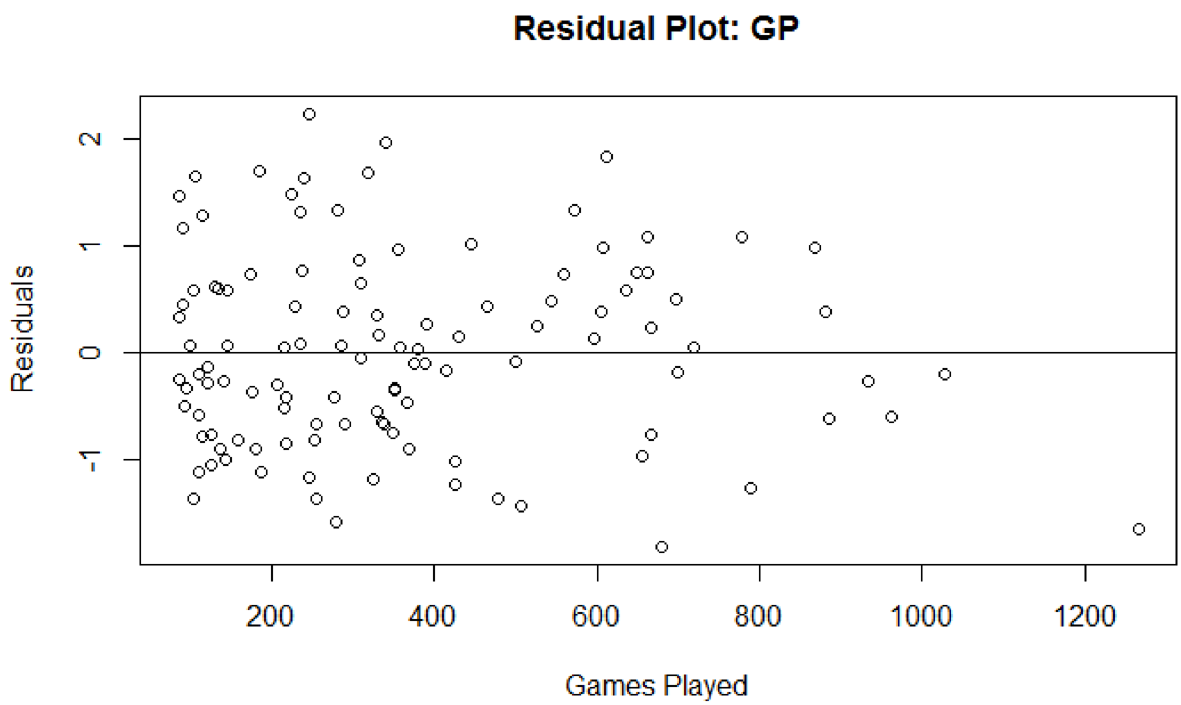
(Fig. 5)

Residual plots with predictor variables:









Conclusions

As can be seen, the most interpretable and effective of the predictive models generated only could explain 28.6% of the deviation values, with poor predictive results, after accounting for the effect of the year of play on performance. In addition, the most significant predictors don’t necessarily provide any information that wouldn’t already be obvious to a casual observer. As a goalie ages, their stats will be more exposed to the effects of aging, which will cause a decline in the average. The same is true of games played as a symbol of workload. Too many games and an overreliance on a single goalie can cause fatique, which affects performance. If a goalie faces a lot of shots, they’ll more chances to raise their overall save percentage, and if they don’t succeed in that time, it’s not likely they’ll be given a large enough sample size of games to make it into a study like this one.

What is more generally shown by this, by not fully explaining goaltender performance, is that more elaborate and descriptive metrics are needed to provide a statistically reasonable modelling of goaltender performance. Simple data like height and weight, for a top-level athlete, are not generally representative of broader performance for the population. It would be useful to find a way to represent techniques and between-game recovery, to find optimal use rates and frameworks that support good performance once it’s been identified, but these cannot be gleaned from simple box score stats.

It’s also possible that reducing the range of discrete values and binning save percentage performance into categories will increase the predictive value of these metrics. I attempted to do this with z-scores, but a better binning scheme might improve results.