Introduction and Motivation

In the National Hockey League (NHL) there are 30 teams, but only the top 8 teams in each conference (those with the most points) make the playoffs. How can we predict points from the data? Is it good offence (many goals scored), or is it good defence (few goals against) that predicts points? Do factors such as the age of the team or the physicality of the team impact the number of points a team gets?

On a preliminary analysis, the most significant variables in predicting team point totals are goals scored and goals against, which together account for approximately 85% of the variance in team point totals. In addition to goals scored and goals against, we wish to further examine the effect of additional variables in the prediction of team point totals.

Description of Data

We have collected data from 8 NHL seasons (2007–2008 to 2014–2015) and all 30 NHL teams for a total of $n = 240$ rows and $p = 41$ columns. Our goal is to predict the number of points a team will get based on variables such as

- goals against
- goals for
- average age
- face-offs won
- hits
- blocked shots
- power plays
- power plays allowed
- shots taken

We believe that aggregating different seasons together is appropriate given the relative stability across our observation period with the lone exception being the 2012–2013 lockout season. Summary statistics comparing league averages (excluding the 2012–2013 lockout season) are shown below.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Pts	91.00	91.50	92.00	91.71	92.00	92.00
Goals allowed	224.00	224.50	228.00	228.90	231.00	239.00
Penalty kill %	81.05	81.56	81.98	81.88	82.18	82.69
Shots	2553.00	2620.00	2651.00	2636.00	2663.00	2683.00

Table 1: Comparison of league average for non-lockout seasons, 2008–2015

We plan to use a multiple linear regression model based on the seasons ranging from 2008–2015 and, if appropriate, analyze the effect of transformations and interaction terms in our model. To assess the quality of our model we will perform the requisite residual analyses while noting how the adjusted R^2 changes as we add different variables in our model. Lastly, we will examine if considerations for simplicity justify the additions introduced. The end goal is not a model that 'perfectly explains' past seasons, but can reliably predict the results of the upcoming 2015–2016 season.