Predicting points in the NHL

by Nathan Esau, Fernando Villasenor and Steve Kane http://github.com/stat350/nhlmodel

Introduction and Motivation

The point system of the NHL currently awards two points to the winner, one point to a team that loses either in overtime or in a shoot-out, and zero points to a team that loses in regulation time. Our analysis is an attempt to explain the number of points that a team will get in a season using linear regression. The variables in our model include how many goals a team has scored or allowed, how many shots they have taken or allowed, how old their players are, and perhaps other variables. It is important to note that we do not need to know the outcome of any particular game to do this and we are using aggregate quantities to determine a teams current season standing.

Our data ranges from the 2007–2008 season to the 2014–2015 season. Before the 2007–2008 season many of our advanced metrics were not available. A peculiarity of our data is that the 2012–2013 season was affected by a lockout which led to only 48 out of a usual 82 games played. We still use all of this season in our analysis and deal with this problem by scaling certain covariates by the proportion of the season a team has played.

Exploratory analysis

Many of the variables in hockey data are redundant. For instance, points summarizes the number of wins, losses and overtime or shootout losses a team has and faceoff percentage summarizes the number of faceoffs a team has won and lost. We tried to choose variables that conveyed different information to avoid colinearity. These variables are shown below.

	Variable	Description		Variable	Description
AvAge	Average Age	Age weighted by time on ice	PTS	Points	A measure of a teams success
BLK	Blocks	Blocked shots	S	Shots	Shots on goal
FF%	Fenwick for percentage	A measure of puck possession	SA	Shots Against	Shots allowed on goal
GA	Goals Against	Goals allowed	SH	Short-handed goals	Goals scored while on penalty kill
GF	Goals For	Goals scored	SHA	Short-handed goals allowed	Goals allowed while on power play
PK%	Penalty kill perrcentage	Percentage other team scores on power play	SOS	Strength of schedule	Looks at whether team is good due to weak opponents
PP%	Power play percentage	Percentage team scores on power play	SRS	Simple rating system	Takes into account average goal differential

Table 1: Hockey statistics used in our analysis

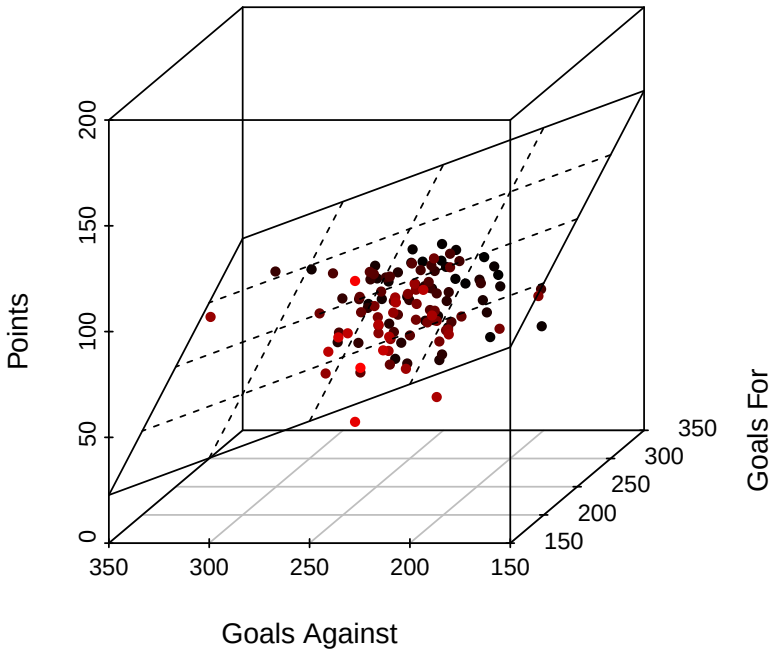

Figure 1: The model PTS = $\beta_0 + \beta_1$ GF + β_2 GA

82 GP	Mean	SD
AvAge	27.84	1.16
BLK	1119.59	151.79
GF, GA	228.88	23.5
PK%	81.86	2.87
PTS	91.83	13.26
S, SA	2456.1	94.89
SH, SHA	6.75	3.1

48 GP	Mean	SD
AvAge	27.64	1.08
BLK	686.57	83.63
GF, GA	50.04	3.09
PK%	81.75	3.43
PTS	53.4	9.64
S, SA	1398.8	94.89
SH, SHA	3.1	2.26

Table 2: Summary statistics from 2008–2015 for 82 game and 48 game seasons.

A well-known model for predicting points in hockey is PTS = $\beta_0 + \beta_1$ GF + β_2 GA, shown above in Figure 1 for non-lockout seasons. Notice that the dots are nicely scattered around the plane with the exception of a few outliers. Later on, we show that it is possible to obtain a model with a lower R^2_{Adj} and out of sample MSE by incorporating other variables into the model.

Not all of the variables increase as the number of games played increases. In particular, AvAge, FF%, PK%, PP%, SOS and SRS are relatively stable as the number of games played changes. This is evident by comparing the value of PK% in the 48 game season to the value of PK% in the 82 game season in Table 2. Later on in our analysis, when we predict points in the 2015–2016 season, we scale the coefficients of these variables by the proportion of the season which a team has played and obtain the MSE of these predictions.

Methods

To explain the number of points a team has we used multivariate regression. We used the R function `regsubsets` from the `leaps` package to determine the best model given a certain number of covariates. This function tried all possible combinations of the variables shown in Table 1 given a fixed model size and returned the variables which were most significant (had the smallest p-values). We then compared models of different sizes using R^2_{Adj} and the square root of out of sample MSE (RMSPE) which as calculated based on our current season points predictions as at November 21, 2015.

As explained previously, some variables are not proportional to the number of games a team plays. For this reason, some variables such as SRS and SOS did not work very well

when including the lockout season in the training set. We fit separate models on two training sets, one of which included the 48 game lockout season and one of which excluded the 48 game lockout season. The models shown in Table 3 and Table 4 use the variables suggested by the `regsubsets` function.

Model										R^2_{Adj}	*RMSPE*
	β_0	GF	GA								
coef	17.86	0.51	-0.2							0.8516	3.1385
p	0	0	0								
	β_0	GF	GA	S	SA						
coef	8.57	0.35	-0.32	0.02	0.01					0.9093	2.2702
p	2e-04	0	0	0	0						
	β_0	AvAge	GF	GA	SRS	S	SA				
coef	-15.58	0.86	0.21	-0.17	11.43	0.02	0.01			0.9144	2.3054
p	0.0749	0.0051	0	8e-04	0.004	0	0				
	β_0	AvAge	GF	GA	SA	FF%	BLK	SOS	PK%		
coef	-116.11	0.94	0.37	-0.3	0.02	1.39	0.01	18.89	0.33	0.9179	2.4144
p	0	0.0028	0	0	0	0	0.0041	0.0017	0.0189		

Table 3: Models fit to data with 2012–2013 lockout season included

When models were fit to the first training set, we can see that goals for and goals against were important variables followed by shots and shots against. Most of the coefficients have an intuitive explanation. For instance, if a team scores many goals they have a better chance of winning games and getting more points so the coefficient for goals for is positive. One interesting coefficient is that for average age, which is positive for all models. This would imply that teams with older players tend to do better than teams with younger players since they have more experience.

Model							R^2_{Adj}	*RMSPE*
	β_0	SRS	SOS	SH				
coef	93.83	28.66	-25.7	-0.29			0.8992	2.319
p	0	0	0	0.0023				
	β_0	GF	GA	AvAge	SH	SHA		
coef	80.5	0.34	-0.35	0.49	-0.27	0.2	0.9119	2.2933
p	0	0	0	0.0572	0.0052	0.0486		

Table 4: Models fit to data with 2012–2013 lockout season excluded

For the second training set, we can see that goals for and goals against are no longer the most significant variables. Now, the simple rating system and strength of schedule metrics are most significant. When models were fit to this training set we found that R^2_{Adj} values were smaller, most likely as a result of using less data when fitting the model. We ended up choosing the model PTS = $\beta_0 + \beta_1$GF + β_2GA + β_3S + β_4SA since it had the lowest RMSPE, comparable R^2_{Adj} to other models, and is a model which is easy to interpret.

Results

Team	Actual	Predicted	Team	Actual	Predicted
Anaheim Ducks	18	17.62	Montreal Canadiens	32	32.75
Arizona Coyotes	21	19.70	Nashville Predators	25	22.70
Boston Bruins	19	21.61	New Jersey Devils	21	18.23
Buffalo Sabres	18	17.92	New York Islanders	23	24.99
Calgary Flames	17	15.01	New York Rangers	30	28.66
Carolina Hurricanes	15	13.91	Ottawa Senators	23	21.83
Chicago Blackhawks	24	23.92	Philadelphia Flyers	17	15.30
Colorado Avalanche	15	20.82	Pittsburgh Penguins	24	21.12
Columbus Blue Jackets	16	18.34	San Jose Sharks	22	21.60
Dallas Stars	32	30.06	St. Louis Blues	27	23.52
Detroit Red Wings	22	19.49	Tampa Bay Lightning	21	21.71
Edmonton Oilers	15	19.01	Toronto Maple Leafs	18	19.91
Florida Panthers	19	21.45	Vancouver Canucks	20	22.90
Los Angeles Kings	24	23.14	Washington Capitals	25	23.50
Minnesota Wild	23	19.91	Winnipeg Jets	20	19.66

Table 5: Predictions for 2015–2016 season at November 21, 2015 using the model PTS = β_0 + β_1GF + β_2GA + β_3S + β_4SA fit with the lockout season included

The largest residuals in Table 5 come from the predictions for the Colorado Avalanche and Edmonton Oilers, which both have 15 points. Unlike Carolina, these teams have a near zero goal differential but few points as a result of losing close games, which our model cannot account for. Residual plots for the fitted model and predicted points are shown in Figure 2.

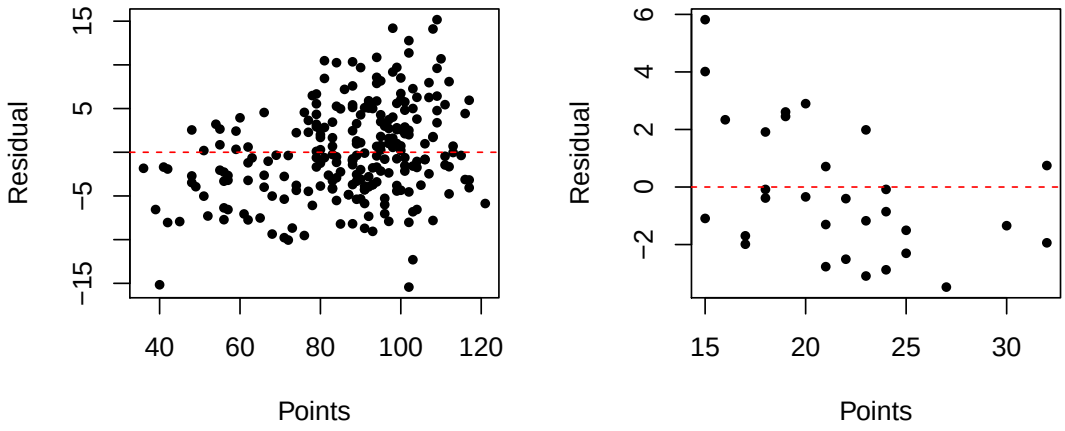

Figure 2: Residuals (fitted: left; predicted: right) for PTS = $\beta_0 + \beta_1$GF + β_2GA + β_3S + β_4SA

Conclusion

We compared models using R^2_{Adj} and out of sample MSPE. We chose the model PTS = $\beta_0 + \beta_1$GF + β_2GA + β_3S + β_4SA. Our model can estimate the points a team has any time during the season if these variables are known. An area for future work would be to project a teams end of season points, perhaps with the use of a time series model.