Predicting points in the NHL

Nathan Esau, Fernando Villasenor and Steve Kane

December 1, 2015

Outline

Background

- Trying to predict PTS scored by an NHL team at any time in the season using other variables
- Two points for a win, one point for overtime or shootout loss, no points for regulation loss
- 2007–2008 season to 2014–2015 season data
- 2012–2013 was a lockout season (48 of 82 games played)

Data

- Scraped from `hockey-reference.com`
  - Figure out which URLs to parse

    `http://www.hockey-reference.com/leagues/NHL_2014.html`
    `http://www.hockey-reference.com/leagues/NHL_2015.html`
    ⋮

  - In R, can use `XML::readHTMLTable(URL)`
  - Deal with data types (convert to `numeric`, `factor`, etc)
  - Merge each `data.frame` and keep common columns
- 2007–2008 season to 2014–2015 season (advanced metrics not available before this, and earlier seasons had higher scoring)

Variables

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

Summary statistics

- Some variables are a sum (i.e. GF, GA, BLK) whereas others are average (i.e. AvAge, PK%)
- When doing predictions, some coefficients needed to be scaled

	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

Table: 82 games played

	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: 48 games played

$\text{PTS} = \beta_0 + \beta_1\ \text{GF} + \beta_2\ \text{GA}$

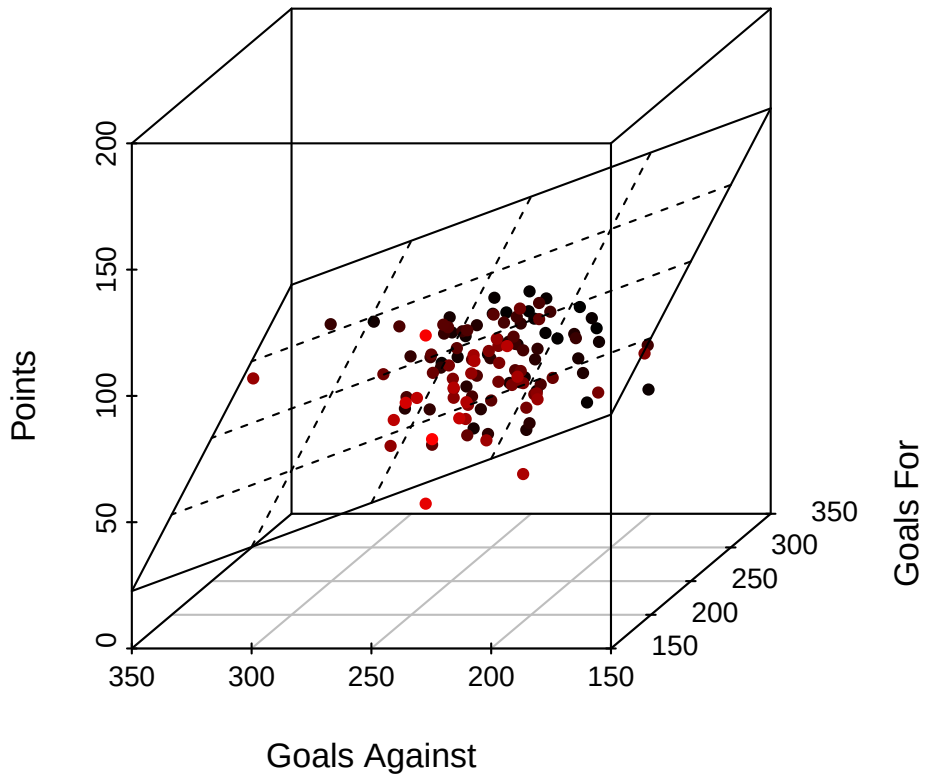

Figure: The model $\text{PTS} = \beta_0 + \beta_1\text{GF} + \beta_2\text{GA}$ has $R^2_{Adj} = 0.85$

Other models

Model									
(1)	β_0	GF	GA						
β	17.86	0.51	-0.2						
p	0	0	0						
(2)	β_0	GF	GA	S	SA				
β	8.57	0.35	-0.32	0.02	0.01				
p	2e-04	0	0	0	0				
(3)	β_0	Age	GF	GA	SRS	S	SA		
β	-15.58	0.86	0.21	-0.17	11.43	0.02	0.01		
p	0.0749	0.0051	0	8e-04	0.004	0	0		
(4)	β_0	Age	GF	GA	SA	FF%	BLK	SOS	PK%
β	-116.11	0.94	0.37	-0.3	0.02	1.39	0.01	18.89	0.33
p	0	0.0028	0	0	0	0	0.0041	0.0017	0.0189

Model	R^2_{Adj}	RMSPE
(1)	0.8516	3.1385
(2)	0.9093	2.2702
(3)	0.9144	2.3054
(4)	0.9179	2.4144

- Selected model PTS $= \beta_0 + \beta_1$ GF $+ \beta_2$ GA $+ \beta_3$ S $+ \beta_4$ SA

Residuals

- 2015–2016 season predictions produced right residual plot
- Largest residuals are Edmonton and Colorado (15 PTS)

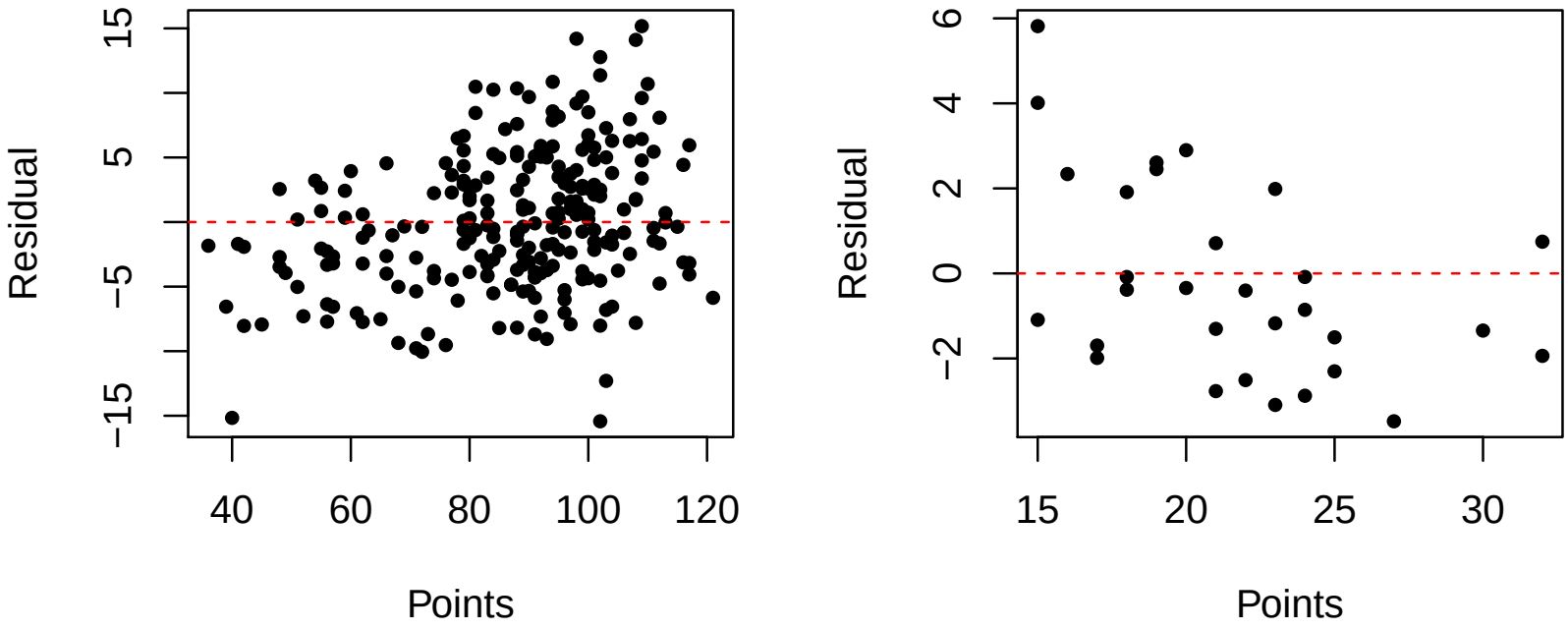

Figure: Fitted residuals (left); Out of sample residuals (right)

Remarks

- Main findings
  - Shots and shots against also significant (along with goals for and goals against)
- Points in hockey is relatively easy to predict (high R^2)
- Future work
  - Predicting end of season points
  - Time series model?