International Journal of Science Culture and Sport
June 2018 : 6(2)
ISSN : 2148-1148
Doi : 10.14486/IntJSCS750

A Statistical Analysis of Team Defense and Performance in the National Hockey League

Andy W. CHEN
University of British Columbia, 2053 Main Mall, Vancouver, BC, V6T 1Z2, CANADA
Email: andywchenca99@gmail.com

Type: *Research Article (**Received:** 23.03.2018 – **Corrected:** 21.05.2018 - **Accepted:** 26.05.2018)*

Abstract

In this paper, I explore the relationship between goalkeeper and team performance in professional hockey using statistical models. The model is evaluated by checking for outliers and major assumptions of the ordinary least squares regression model. I find a negative relationship between goals allowed per game and total team points per season. In particular, I find that the intercept of the model to be 184.0 and coefficient for average goals allowed per game to be -33.819. This means that when goals allowed per game increases by 0.1, the total points for a team in a season will decrease by 3.38. The assumptions of the linear regression model are satisfied.

Keywords: Hockey, Sports Analytics, Statistics, Linear Regression, Ordinary Least Squares

Introduction

The performance of professional sports teams can be attributed to a number of factors. One of the most significant factors may be the defensive ability of the team. In professional hockey, the last line of defense is the goaltender. The ability of the goaltender to stop shots can dictate the outcome of a game. In this paper, I propose a statistical model that captures the relationship between goaltender and team performance. I evaluate the model by checking for outliers and major assumptions.

Related work in this area includes a paper by Seaton and Campos (2017), who find that in soccer, a goalkeeper's performance is highly dependent on goalkeeper's location and type. Ofoghi et al. (2013) build a model to capture performance analysis, data mining, and characteristics of various techniques. Lames and McGarry (2007) find the significance of including dynamic interactions in-game sports as key features of sport performance. Travassos et al. (2013) utilize ecological dynamics to study performance of professional sports teams. Reilly (2001) utilize notation analysis and motion analysis to study the performance of professional athletes.

Methods

I collect a dataset comprised of goals against average (GAA) and total team points in the 2009-2010 season. Goals against average are the average number of goals scored per game, while points is an indicator of team performance. A win is worth two points, while an overtime loss is worth one point. A loss without overtime is worth no point. There are 30 teams in total. The mean GAA is 2.72 and mean total points in the season is 92 points.

Table 1. Goals Against Average (GAA) and Total Team Points

Team	GAA	Points
New Jersey Devils	2.24	103
Boston Bruins	2.28	91
Phoenix Coyotes	2.34	107
Buffalo Sabres	2.42	100
Chicago Blackhawks	2.43	112
Calgary Flames	2.45	90
Detroit Red Wings	2.48	102
San Jose Sharks	2.51	113
Los Angeles Kings	2.52	101
New York Rangers	2.58	87
St. Louis Blues	2.61	90
Montreal Canadiens	2.61	88
Vancouver Canucks	2.63	103
Nashville Predators	2.66	100
Philadelphia Flyers	2.68	88
Washington Capitals	2.72	121
Colorado Avalanche	2.74	95

Florida Panthers	2.80	77
Ottawa Senators	2.80	94
Pittsburgh Penguins	2.82	101
Minnesota Wild	2.87	84
Anaheim Ducks	2.92	89
Dallas Stars	2.92	88
Columbus Blue Jackets	2.99	79
Atlanta Thrashers	3.01	83
Carolina Hurricanes	3.02	80
Tampa Bay Lightning	3.03	80
New York Islanders	3.09	79
Toronto Maple Leafs	3.15	74
Edmonton Oilers	3.34	62

I use a linear regression model to model the relationship between GAA and total points. The coefficient of GAA should indicate the correlation between these two variables and the extent to which GAA, an indicator of team defense, can affect team performance. I then conduct so robustness checks such as checking for outliers and checking that major assumptions are satisfied.

Results and Discussion

I start with some exploratory data analysis. Figure 1 shows the density histogram of GAA. It appears that the GAA average is unimodal with a slight skew to the left. Figure 2 shows the boxplot of GAA which allows a visual interpretation of the summary statistics and can show us the potential existence of some outliers. Figure 3 shows the density histogram of the average total points per season, which is unimodal with a slight skew to the left as well. Figure 4 shows the boxplot of average total points per season. Figure 5 shows the scatterplot of GAA and total points season. There appears to be a negative, linear relationship between the two variables, suggesting that a linear regression would be suitable.

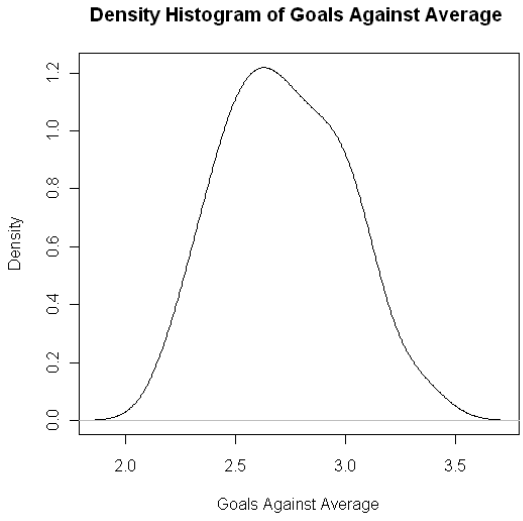

Figure 1. Density Histogram of Goals against Average (GAA)

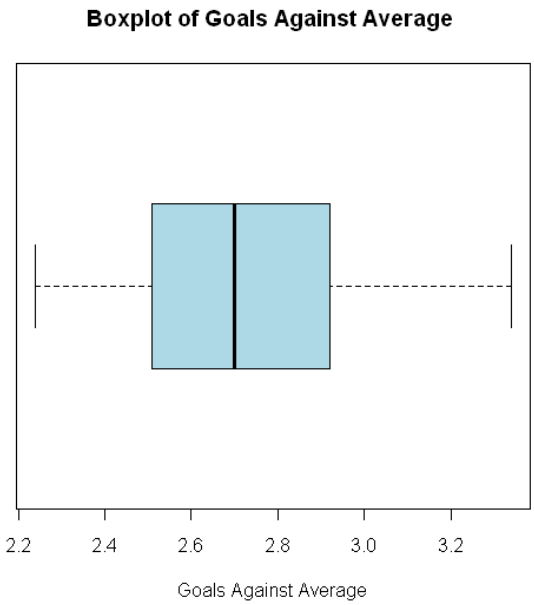

Figure 2. Boxplot of Goals against Average (GAA)

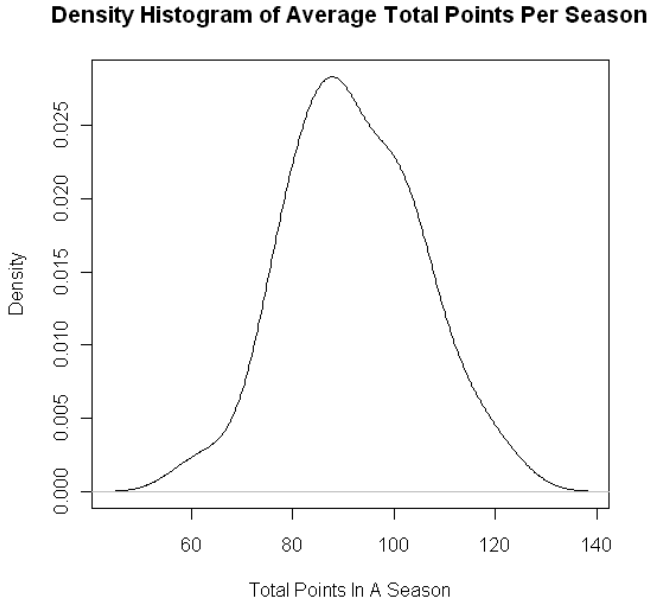

Figure 3. Density Histogram of Average Total Points per Season

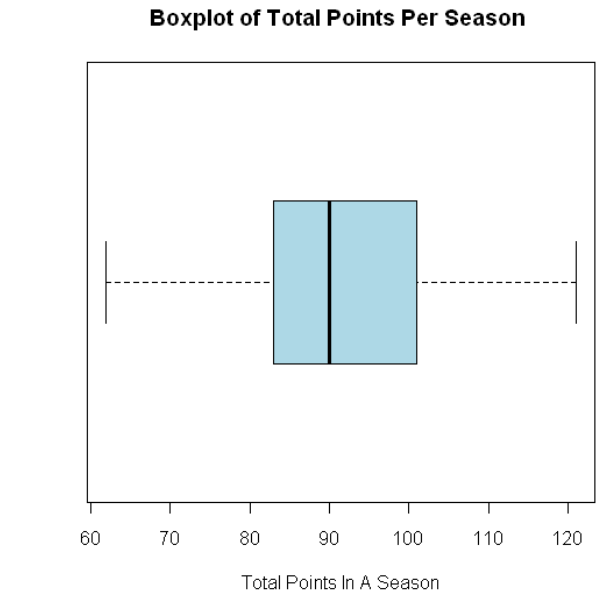

Figure 4. Total Points per Season

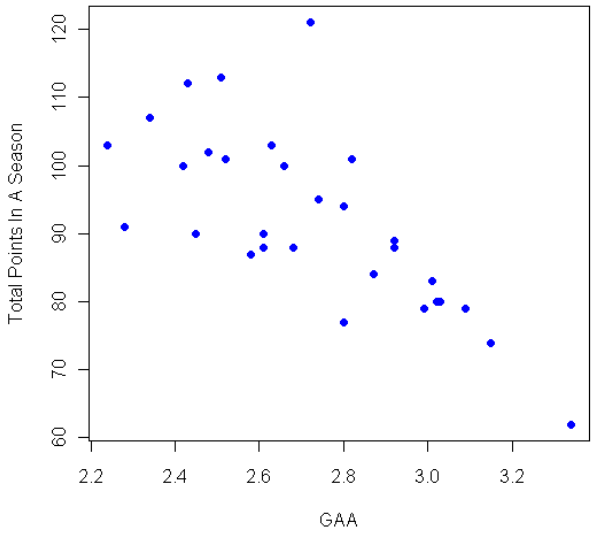

Figure 5. Scatterplot of GAA and Total Points per Season

I the estimate the particular least squares regression model below. Table 2 shows the results of the regression.

$$\text{Total Points Per Season} = \beta_0 + \beta_1 \text{GAA} + \epsilon$$

Table 2. Regression Results

	Estimate	St. Error	t-value	p-value
Intercept	184.088	16.971	10.847	1.55e-11
GAA	-33.819	6.204	-5.451	8.09e-6

The intercept indicates that we expect the average points for a team to be 184.0 when there is no goals against average (GAA=0). This is also not reasonable because it is impossible for a team to allow no goals per game during the whole season. The slope is more meaningful as the p-value indicates that it is significant at the 95% confidence level. The value of the coefficient of GAA, -33.819, suggests that with every increase in goals against average that the total points for a team in a season will decrease by 33.8, or decrease by 3.38 for every 0.1 increase in goals against average. This value confirms a negative correlation between GAA and total points per season.

I then use the regression results to construct the 95% confidence band for the regression line as in Figure 6. This is the band for which we can be 95% confident to contain the true linear regression line. Figure 7 shows the 95% pointwise prediction band, which is the interval we are 95% confident to contain the prediction at a given value for the GAA (predictor variable).

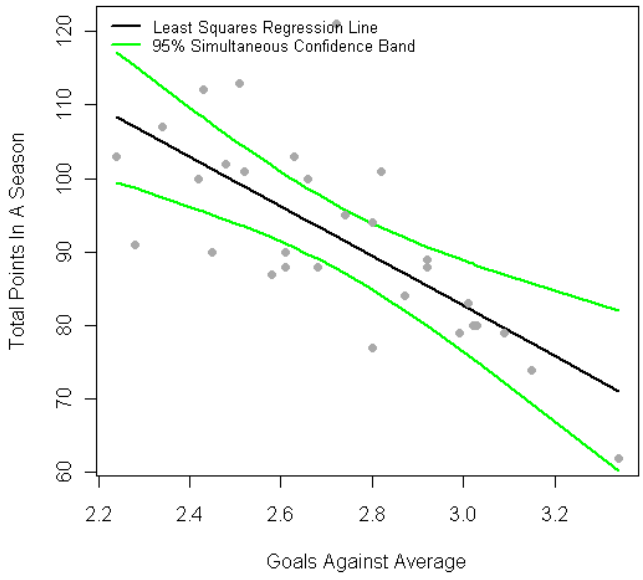

Figure 6. 95% Confidence Band of Least Squares Regression Line

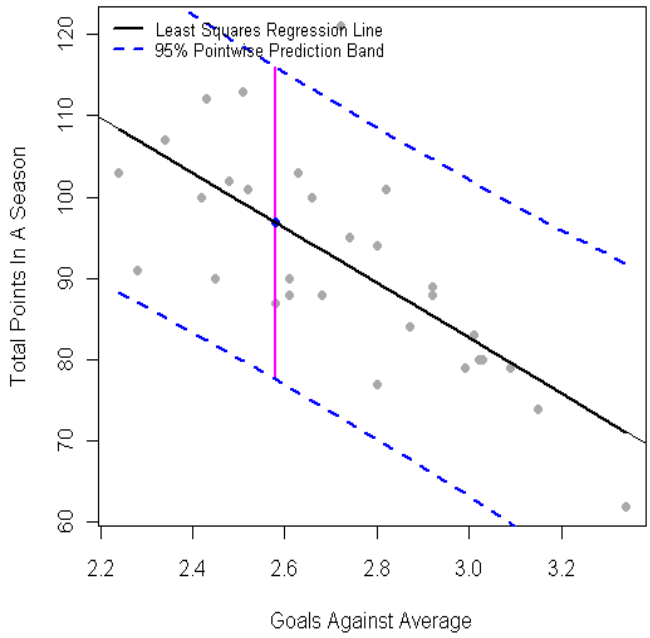

Figure 7. 95% Pointwise Prediction Band

The results suggest a negative linear relationship between goals against average and total points per season. This indicates that team performance in professional hockey highly depends on goaltending, which is a major part of a team's defense.

I check that the major assumptions of the linear regression model are satisfied. The scatterplot in Figure 5 shows that the linearity assumption is satisfied. That is, there appears to be a linear relationship between the predictor and response variable. Figure 8 shows that the variance of total points per season is similar across different values of GAA, which is the assumption of homoscedasticity. Figure 9 shows a random pattern of the residuals across different values of the response variable, total points per season. Figures 8 and 9 together suggest that the residuals are independent of the predictor and response variables. Regarding autocorrelation, it should be reasonable to assume that the errors are independent of one another because the teams played games independent of almost every other team (except the one they played against). Therefore the assumption of independent errors is satisfied. Figure 10 checks fort normality assumptions. The histogram and density histogram of residuals are approximately normal. There is one potential outlier with a residual near 30, but should not be significant enough to violate the normality assumption. The points in the QQ-plot line up along the straight line of agreement between the distribution of the residuals and a normal distribution, except again for the outlier at the end with residual near 30. Therefore the normality assumption is reasonably justified.

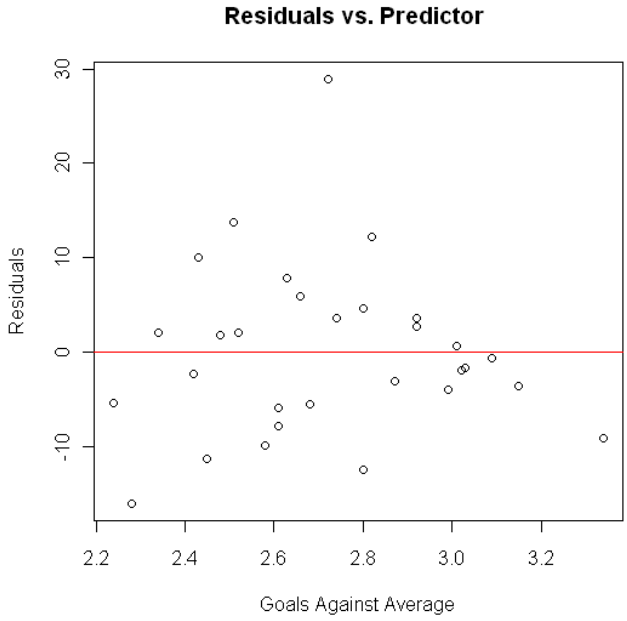

Figure 8. Residuals vs. Goals against Average (GAA)

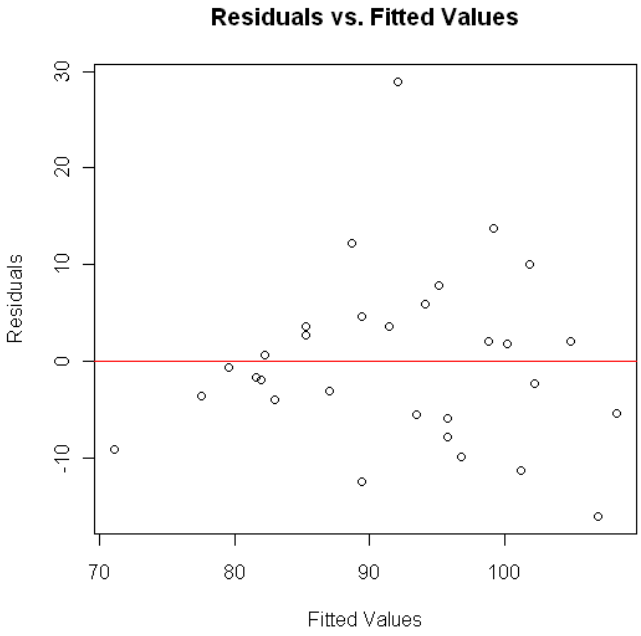

Figure 9. Residuals vs. Total Points per Season

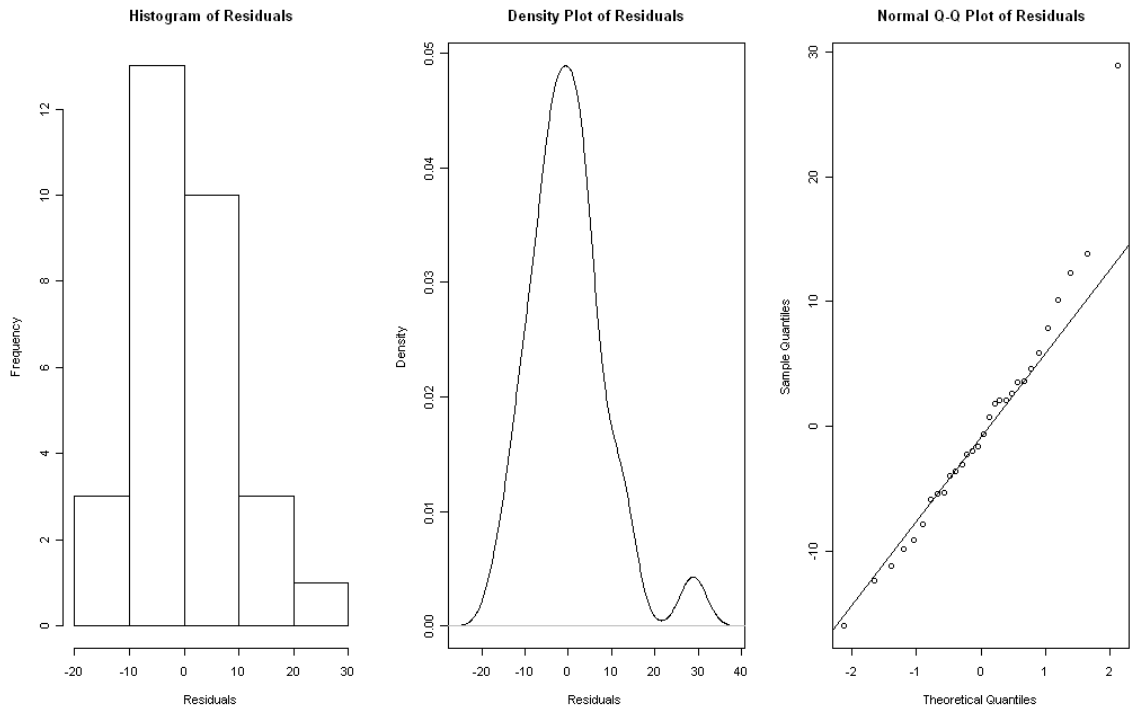

Figure 10. Normality Assumption Checks

Conclusion

In this paper, I present a least squares linear regression model that predicts team performance based on goals against average, a proxy for team defense. I train the model using data on professional hockey teams, provided error metrics such as confidence and prediction intervals, and checked the model's assumptions. Future work may include adding more variables to enrich the model to capture more information on what factors affect the performance of professional sports.

Conflict of Interest

The author has no conflicts of interest.

References

Lames M, McGarry T (2007). On the search for reliable performance indicators in game sports. International Journal of Performance Analysis in Sport. 7(1):62-79.

Ofoghi B, Zeleznikow J, MacMahon C, Raab M (2013). Data Mining in Elite Sports: Measurement in Physical Education and Exercise Science. 17(3):171-186.

Reilly T (2001). Assessment of sports performance with particular reference to field games. European Journal of Sport Science. 1(3):1-12.

Seaton M, Campos J (2017). Distribution competence of a football clubs goalkeepers. International Journal of Performance Analysis in Sport. 11(2):314-324.

Travassos B, Davids K, Araujo D, Esteves TP (2013). Performance analysis in team sports: Advances from an Ecological Dynamics approach. International Journal of Performance Analysis in Sport. 13(1):83-95.