Lukas Ronkin
Daniel Halteh
Victor Shih

Stats 130 Group Project Report

In professional Ice Hockey, or the NHL, points are generally used as the best quantitative identifier of a player's performance and overall value to the team. A player's total points is simply calculated by summing the number of goals and assists tallied by a player over the course of a season. A player's total points per season typically depends on a variety of other statistical variables, many of which do not correlate in an obvious manner. Through the use of the SPSS statistical software, our objective is to determine which individual player statistic variables (other than simply goals and assists) are significant in predicting a player's total points in one season. With a successful linear regression model, being able to predict a player's points would be valuable for a team trying to evaluate a player's skillset.

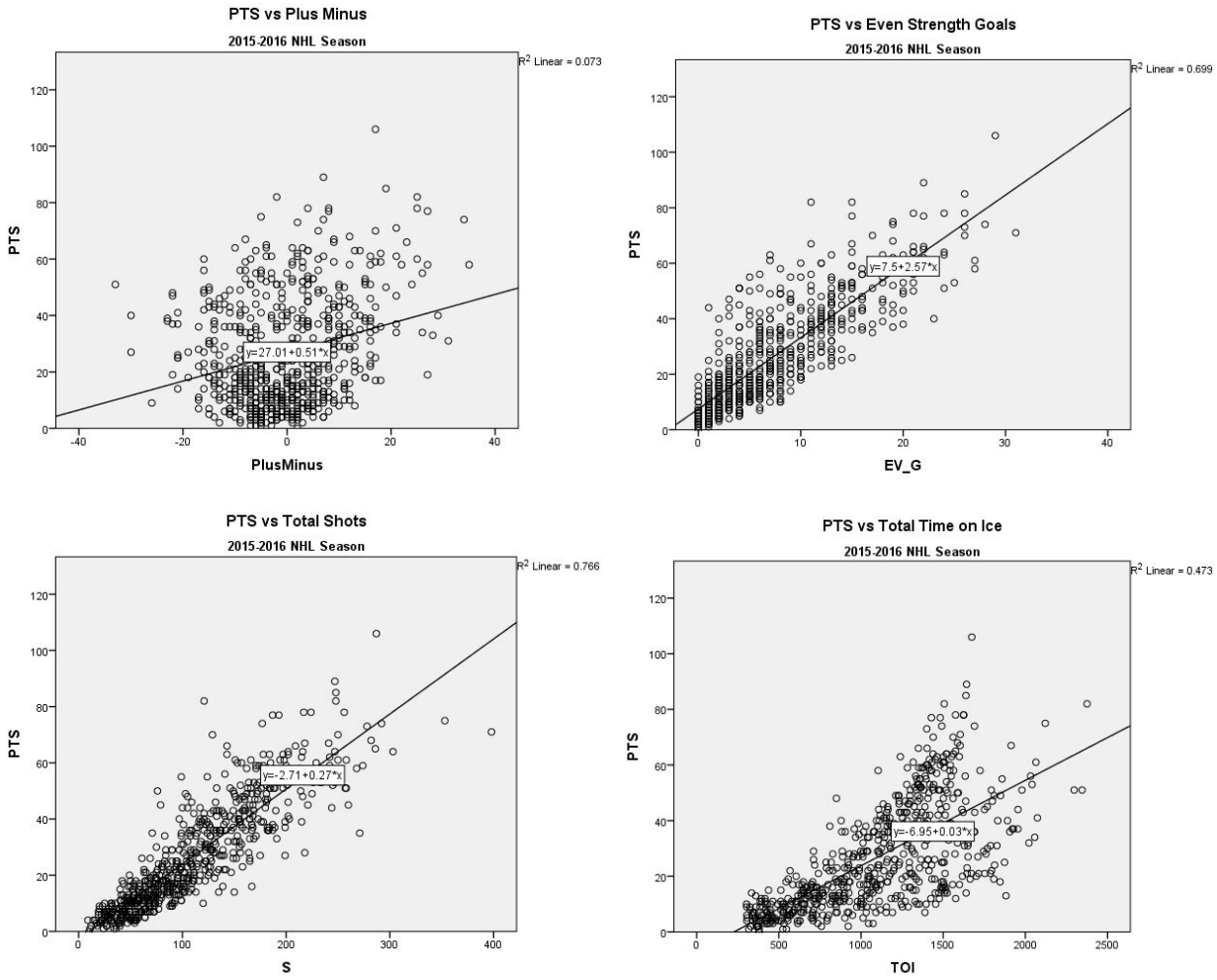
We wanted to use data from the most recent season so found 2015-16 NHL data from www.hockey-reference.com, which is a reliable sports statistics website. The original dataset consisted of statistics from all skaters over the course of the entire season. The dataset had 898 total observations (players) with 27 different variables. There is 24 scalar variables and 3 character variables in the dataset. In order to eliminate any players that did not contribute significant playing time during the season, we chose only include players who were on the ice for more than 300 minutes over the course of the entire season. Individuals who could not meet this requirement would not be able to provide significant data to our model because of possible chance of skewness. Following similar logic, we decided to include those individual players who played in at least 20 games during the season. The NHL season is 82 games, so we wanted players that players who played approximately 75% of the season or more. After successfully cleaning our data in SPSS, the dataset we used to perform a linear regression model had 639 observations.

In order to determine which individual player statistic variables would result in being the most significant predictors of total points in a season, we performed a series of regression model tests utilizing each of the variables given in our cleaned data set. After comparing the various resulting linear models, we ultimately settled with a player's Plus/Minus, Even Strength Goals Scored (EV_G), Total Shots Attempted (S), and Total Time on Ice (TOI) as the best statistical predictors for Points (PTS) in a season. PTS, as defined before is a player's goals summed with their assists. The mean for PTS was 26.87 with a standard deviation of 19.227. A player's plus/minus (PlusMinus) is defined as how many goals their team scores or lets up when they are on the ice. For example, a player with a positive Plus/Minus means that their team scored more goals than the other team when that player was on the ice. If a player has a negative Plus/Minus, it means the opposite. Even strength goals can be defined as goals scored by a player when their team was at even strength, i.e. no player on their team was in the penalty box. The mean for even strength goals was 7.54 and had a standard deviation of 6.256. Total shots attempted is simply the amount of shots the player attempted over the course of the season. The mean for S was 110.77 and the standard deviation was 63.016. TOI can be defined as the total amount of time (in minutes) a player spends on the ice over the course of the entire season. The mean for TOI was 1100.35 with a standard deviation of 430.149. Below is a chart that details descriptive statistics for the variables used in our final model:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
PTS	639	0	106	26.87	19.227
PlusMinus	639	-33	35	-.28	10.178
EV_G	639	0	31	7.54	6.256
S	639	9	398	110.77	63.016
TOI	639	303	2375	1100.35	430.149
Valid N (listwise)	639				

After looking at the variable's descriptive statistics, we wanted to see how our dependent variables were related to the independent variable PTS. To see the relationships, we graphed scatter plots for each of the four variables compared to PTS. To get a better understanding of their relationships, we fit a regression line on top of the scatter plots. Each variable's scatter plot can be seen in the following graphs:



From the graphs, it is clear that EV_G and S have the strongest linear relationship with PTS. That makes sense, as a player scores an even strength goal, their PTS total will directly go up, and if a player takes more shots, that will likely mean they will score more goals resulting in more points. TOI seems to

have a fair linear relationship and can be understood as if a player spends more time on the ice, they will have more opportunities to add to their PTS total. The weakest association seems to be Plus/Minus. That might be explained by some star players with a lot of PTS are on bad teams and some worse players with low PTS are on good teams.

After trying a few different models to see which variables would best predict PTS, we created a model with the variables PlusMinus, EV_G, S, and TOI to predict PTS. After running the linear regression model in SPSS, the output proved that these four variables were a good predictor for PTS. The output shows that the model has an adjusted R square value of .853. That means 85.3% of the variation in PTS can be explained by the four variables. Furthermore, the ANOVA table shows our regression model is significant because the p value is .000. The coefficient table shows that all four variables have significant predictive value because all of their p values are less than 0.05. Our final regression model is as follows:

$$\text{Expected } PTS = -7.049 + 0.123PlusMinus + 1.357EG + .103S + 0.011TOI$$

The SPSS linear regression output can be seen below:

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	TOI, PlusMinus, EV_G, S[b]	.	Enter

a. Dependent Variable: PTS

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.924[a]	.854	.853	7.364

a. Predictors: (Constant), TOI, PlusMinus, EV_G, S

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	201472.138	4	50368.035	928.881	.000[b]
	Residual	34378.287	634	54.224		
	Total	235850.426	638			

a. Dependent Variable: PTS

b. Predictors: (Constant), TOI, PlusMinus, EV_G, S

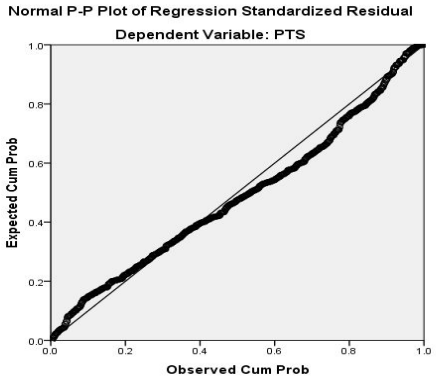
Coefficients[a]

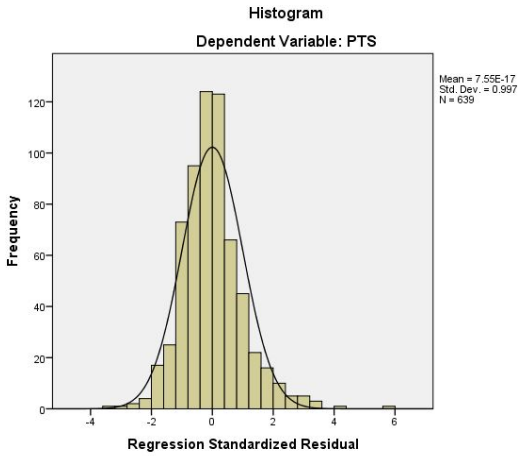
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-7.049	.825		-8.547	.000
	PlusMinus	.123	.031	.065	4.007	.000
	EV_G	1.357	.086	.442	15.802	.000
	S	.103	.011	.338	9.769	.000
	TOI	.011	.001	.250	10.973	.000

a. Dependent Variable: PTS

To understand the model, we need to interpret the coefficients. For clarification, we will hold all other variables constant when discussing change. For every one increase in a player's Plus/Minus, we expect player's PTS to increase by 0.123, on average. For every even strength goal a player scores, we expect player's PTS to increase by 1.357, on average. For shot a player attempts, we expect player's PTS to increase by 0.103, on average. Finally, for every minute spent on ice per game, we expect player's PTS to increase by 0.011, on average. Interpreting the intercept in this case is largely irrelevant, because it is impossible for a player to have less than 0 PTS in a season.

The Normal Residual Plot and Histogram of Residuals below show that the residuals are normally distributed as the 1st plot displays a bell curve and the 2nd plot displaying residuals close to the normality line. Thus, our model is adequate and has strong predictive value.





Applying our regression model to the context of our data, we calculated what the most "average" player's expected PTS would be. We used the four variable's averages found in the descriptive statistics and put them into our regression model: Predicted PTS = -7.049 + (0.123*-.28) + (1.357*7.54) + (.103*110.77) + (0.011*1100.35). As a result, we expect that the most "average" player would tally approximately 27 PTS over the course of a season. Another beneficial application of our model is we can calculate a real player's residual score for PTS from the 2015-16 season. For example, in 2015-16 Los Angeles Kings star Dustin Brown recorded a Plus/Minus of 18, scored 19 even strength goals, attempted 242, and spent 1416 minutes on ice. Based on our mode; : -7.049 + (0.123*18) + (1.357*19) + (.103*242) + (0.011*1416) we would predict Brown to record 61.45 PTS. In reality, Brown recorded 62 PTS during the season. Brown's residual score is 0.55 which shows our model is extremely accurate and reliable.

Using SPSS, we have constructed a regression model that successfully predicts PTS based on four linearly significant variables. Our model can be used by hockey coaches and executives to help aid their decision on how to properly evaluate players. The decision process of which players to acquire or not can be made easier thanks to our model making accurate predictions of a player's PTS. Our model could even be used as a starting point to see what other variables might be associated with each other or even used to try and create a new statistic.