Nikita Anistratov

PART B: NHL SALARY RESEARCH PROJECT

Topic: <u>Does an increase in NHL player salary result in higher performance, or in other words an increase in points per game?</u>

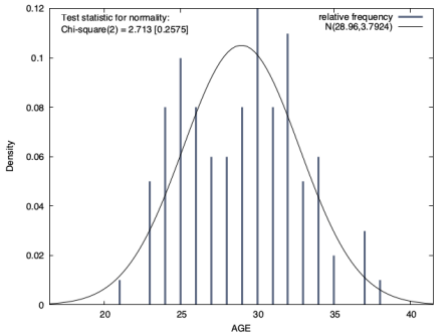
Dependent Variable: Points Per Game (PPG)

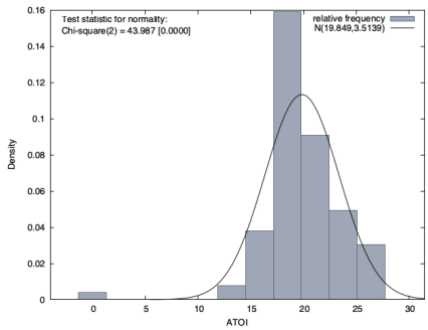
Independent Variable (*Variable of interest*): Player Salary (SALARY)

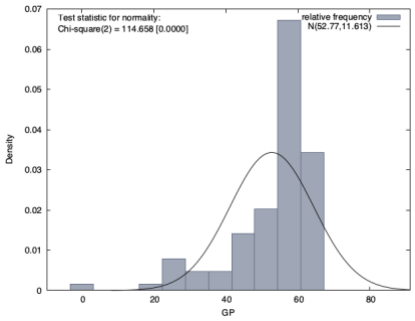
Other Independent Variables: Average Time On Ice - (ATOI), Games Played - (GP), Shots On Goal - (SOG), Offense/Defense - (POS), Age - (AGE)

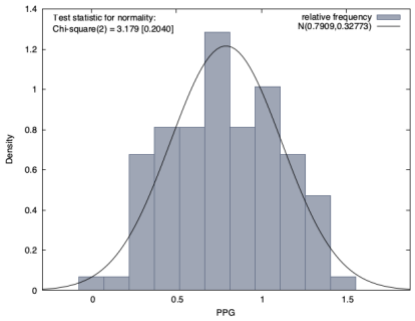
The dependent variable in this data set represents the overall performance of a player. The performance is measured as a ratio of points per games played (points divided by games played). “PPG” will represent this variable in our data set. The variable of interest is player salary which will be represented by “SALARY”, the measurement is in dollars and is scaled in millions. Average time on ice is represented by “ATOI” and will represent the average minutes a player spent on ice in all games played, this variable is measured in minutes. Games played represented by “GP”, is the total number of games played in this season, we are evaluating the 2021-2022 season. Shots on goal represented by “SOG”, is the total successful shots a player made on the net this season, this does not include shots made that missed the net. Position represented by “POS”, is a categorical variable that we have converted into a binary variable, this variable represents the position played by the player. Since many players play multiple positions, we have converted this variable into two categories: offense represented by 0 and defense represented by 1. It is not important for our model that a player plays center and left wing, so we have simplified this to two categories. The final independent variable is age represented by “AGE”, this is measured by the birthdate up to today’s date: 03/10/22.

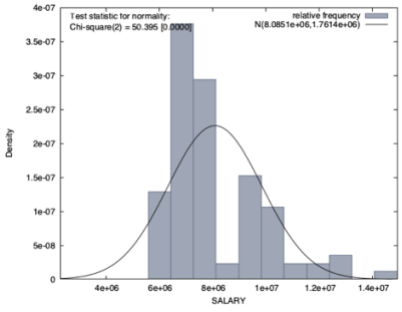
The histogram for AGE shows us an asymmetrical representation, the plot does not really give us much valuable information, and not much to comment on the skewness or kurtosis. The histogram for ATOI shows a symmetrical graph that is slightly positively skewed with S > 0 and the kurtosis is positive. The histogram for GP is showing a negatively skewed asymmetrical graph with S < 0 and the kurtosis is positive. The histogram for PPG is symmetrical and not skewed much, the kurtosis is close to 0. The histogram for SALARY shows us a very positively skewed graph S > 0 the kurtosis is also positive. Lastly the histogram for SOG is symmetric and kurtosis is close to 0.

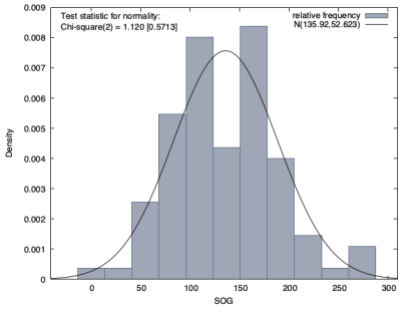












Below is the full summary statistics of the data. For AGE the skewness and kurtosis don't tell us much. The mean gives us an idea of the average age of the players, but the median is most likely the more valuable measure. The minimum and maximum give us a good age range for the data set. For POS since this is a binary variable most of this data isn't helpful, but the average essentially gives us the percent so 70 percent are forwards ands about 30 percent are defensemen. GP is a negatively skewed histogram the median is going to be more valuable than the mean, the standard deviation also shows the wide dispersion of the values. PPG the mean is going to give us a valuable value, this is very symmetrical data. SOG has a high standard deviation which shows us the wide dispersion and is positively skewed. The ATOI has a high kurtosis and is fairly skewed the median should be more valuable in this case. Lastly the salary has a very high standard deviation and is skewed, median will be more valuable as well as standard deviation and skewness.

Summary Statistics, using the observations 1 – 100

Variable	Mean	Median	Minimum	Maximum
AGE	28.960	29.000	21.000	38.000
POS	0.31000	0.0000	0.0000	1.0000
GP	52.770	57.500	0.0000	64.000
PPG	0.79090	0.79000	0.0000	1.4800
SOG	135.92	130.00	0.0000	274.00
ATOI	19.849	19.410	0.0000	26.380
SALARY	8.0851e+06	7.5000e+06	6.0000e+06	1.4500e+07

Variable	Std. Dev.	C.V.	Skewness	Ex. kurtosis
AGE	3.7924	0.13095	0.15055	-0.70615
POS	0.46482	1.4994	0.82163	-1.3249
GP	11.613	0.22007	-2.0114	4.3552
PPG	0.32773	0.41438	-0.0078257	-0.83671
SOG	52.623	0.38716	0.24308	-0.012251
ATOI	3.5139	0.17703	-1.5358	8.7580
SALARY	1.7614e+06	0.21786	1.3723	1.7381

Variable	5% Perc.	95% Perc.	IQ range	Missing obs.

AGE	23.000	35.000	6.0000	0
POS	0.0000	1.0000	1.0000	0
GP	27.000	62.000	10.750	0
PPG	0.29000	1.2895	0.51250	0
SOG	53.100	223.55	72.250	0
ATOI	15.580	25.708	4.0700	0
SALARY	6.1013e+06	1.2000e+07	2.1375e+06	0

Below we show the matrix of correlation coefficients. It is visible that SOG is the most important determinant of our dependent variable PPG. The statical significance of POS and AGE is low, but for the concept of our model this is very important, as position gives a different result for points per game and age influences salary. Time spent on ice is shown with low significance, but it is also a significant factor in both salary and PPG. All variables with low significance may seem unimportant, but conceptually they are important for understanding and proving our question at hand.

Correlation coefficients, using the observations 1 - 100
5% critical value (two-tailed) = 0.1966 for n = 100

AGE	POS	GP	PPG	SOG	
1.0000	0.1274	-0.1578	-0.2648	-0.3268	AGE
	1.0000	0.1106	-0.4395	-0.1893	POS
		1.0000	0.1302	0.5881	GP
			1.0000	0.6630	PPG
				1.0000	SOG

ATOI	SALARY	
-0.1283	0.0868	AGE
0.6297	-0.0303	POS
0.3786	-0.1535	GP
0.2062	0.3290	PPG
0.3370	0.0940	SOG
1.0000	0.1399	ATOI
	1.0000	SALARY

Below is a simple regression model on our dependent variable PPG and our variable of interest SALARY. The p-value shows us that our variable is in fact statistically significant as it is less than 0.5 > 0.0020 and it is very small. Economically it is not so significant given how small the coefficient is, this does not give us confidence economically. The coefficient is negative and very small which is not what we expected given our hypothesis, as SALARY increases PPG should also increase, but this leads us to believe that as SALARY increases PPG will decrease. The PPG will decrease by 6.12190e-08 for every 1 dollar increase in SALARY, but since our salaries are primarily in millions, lets multiply the coefficient by 1 million to get a more significant understanding of 1 unit change of the salary. We now will consider 1-unit change a 1 million dollar increase in salary. This will result in a 0.061 decrease in PPG.

Although our R-squared value is very low at about 10 percent and this should tell us that our model does not fit our data, but since SALARY is statically significant as we have explained earlier, we can still deduce that our model could be significant, and our model could in fact fit our data. Statistically significant coefficients represent the mean change in the dependent variable given a one-unit shift in the independent variable. Our standard error of regression is very small at 0.311 which will indicate the values are not far from our regression line, so unlike R-squared this indicates that our model does fit our data. Since coefficient has a negative sign and this indicates to us that we have a negative correlation, which is not what our hypothesis states. It may be that our variables are not significant enough to give us an accurate prediction of our hypothesis model, or we can deduce that our hypothesis is in fact wrong and an increase in salary will decrease our PPG.

Model 1: OLS, using observations 1-100
Dependent variable: PPG
Heteroskedasticity-robust standard errors, variant HC1

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	0.295936	0.158495	1.867	0.0649	*
SALARY	6.12190e-08	1.92382e-08	3.182	0.0020	***

Mean dependent var	0.790900	S.D. dependent var	0.327732
Sum squared resid	9.482261	S.E. of regression	0.311059
R-squared	0.108259	Adjusted R-squared	0.099159
F(1, 98)	10.12611	P-value(F)	0.001959
Log-likelihood	−24.10648	Akaike criterion	52.21296
Schwarz criterion	57.42330	Hannan-Quinn	54.32168

Below is a regression model run on just the binary variable POS, the position of the player. When calculating the t-statistic the result was 0.048 which indicates we would reject our hypothesis, since $0.048 < 0.5$. This makes sense since our other indicators point in the same direction of rejecting our hypothesis. We need to reevaluate our data to get a better understanding of the accuracy of our model still. Our model conceptually makes sense, but our data has shown that our hypothesis was wrong, this could have to do with lack of proper independent variables or possible a few of the variables are not significant which may skew our data (even though conceptually these variables are important to the question at hand). With everything we described in the prior paragraphs, the analysis of the data shows that our model may not fit for this data. This t-statistic makes sense regarding the evidence of the standard error and the R-squared showing that our hypothesis fails for this data and the salary of a player will not increase his performance or in other words his points per game will not increase.

Model 2: OLS, using observations 1-100
Dependent variable: PPG
Heteroskedasticity-robust standard errors, variant HC1

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	0.886957	0.0367682	24.12	<0.0001	***
POS	−0.309860	0.0613546	−5.050	<0.0001	***

Mean dependent var	0.790900	S.D. dependent var	0.327732
Sum squared resid	8.579700	S.E. of regression	0.295885
R-squared	0.193138	Adjusted R-squared	0.184905
F(1, 98)	25.50559	P-value(F)	2.04e-06
Log-likelihood	−19.10529	Akaike criterion	42.21058
Schwarz criterion	47.42092	Hannan-Quinn	44.31930