PART C: 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 million dollars. 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. Our model looks at the effect of SALARY on the performance of a player in the NHL, in our case

PPG will gage the level of performance. We want to look at an incentivized system, can SALARY be used to incentivize better performance? Does an increase in NHL player salary result in higher performance, or in other words an increase in points per game?

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

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	0.557254	0.211264	2.638	0.0098	***
GP	−0.00950484	0.00235457	−4.037	0.0001	***
AGE	−0.00179332	0.00467046	−0.3840	0.7019	
POS	0.184783	0.0427527	4.322	<0.0001	***
SOG	0.00453088	0.000487964	9.285	<0.0001	***
ATOI	−0.00204756	0.00171062	−1.197	0.2344	
SALARY	−0.00233578	0.00435774	−0.5360	0.5932	

Mean dependent var	0.798889	S.D. dependent var	0.319463
Sum squared resid	3.262233	S.E. of regression	0.188306
R-squared	0.673828	Adjusted R-squared	0.652556
F(6, 92)	41.90205	P-value(F)	2.98e-24
Log-likelihood	28.45413	Akaike criterion	−42.90826
Schwarz criterion	−24.74243	Hannan-Quinn	−35.55834

As we see above the full regression output of PPG on all independent variables. Highlighted in red are the statistically significant coefficients according to a p-value of 10% or less. We will be interpreting the coefficients of these variables and the resulting impact of a one unit change for each, on the dependent variable PPG. Starting with GP (games played) measured in number of games, based on the coefficient in the regression output above a one unit increase in GP will result in a decrease of 0.9% in PPG which statistically makes sense, because this would increase the denominator of the ratio for PPG. Points per game may still increase, but the value of games has increased which is the denominator of the ratio, hence decreasing the value of the ratio PPG. I would expect this coefficient to have a positive sign, as realistically the more

games played the more opportunities for points, although like I explained above the ratio of the denominator increases decreasing the ratio. So, it is logical that there will be a small decrease in PPG with a one unit increase in GP, in this case 0.9%.

POS	0.184783	0.0427527	4.322	<0.0001	***

Next, we look at the coefficient of POS (position). This is a binary variable with 0 accounting for offense and 1 accounting for defense. So, if our coefficient POS is 1 which is defense, this will account for 18.48% increase in PPG compared to offense. We would expect this coefficient to be negative rather than positive, as we tend to expect offense to have higher PPG, this proves the opposite.

SOG	0.00453088	0.000487964	9.285	<0.0001	***

Lastly, we look at SOG (shots on goal). A one unit increase in this case is one shot, will result in a 4.5% increase in PPG, which is exactly what we would expect. The more shots on goal the more probability of them going in the net! So, the coefficient being positive is exactly what we expect.

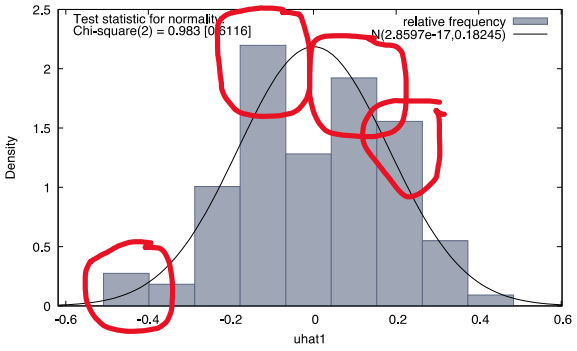
R-squared	0.673828	Adjusted R-squared	0.652556

Now, we look at how well our model fits the data. Here we see our R-Squared is 67.38%, this tells us our regressors (independent variables) explain 67.38% of the variance of our dependent variable PPG. We want to be as close to R-Squared = 100% or an R-Squared = 1.0 as we can indicating our model is a great fit, so 67.38% is an indicator that our model is pretty good at explaining the variation of PPG. Just looking at R-Squared can be deceiving so now we look at

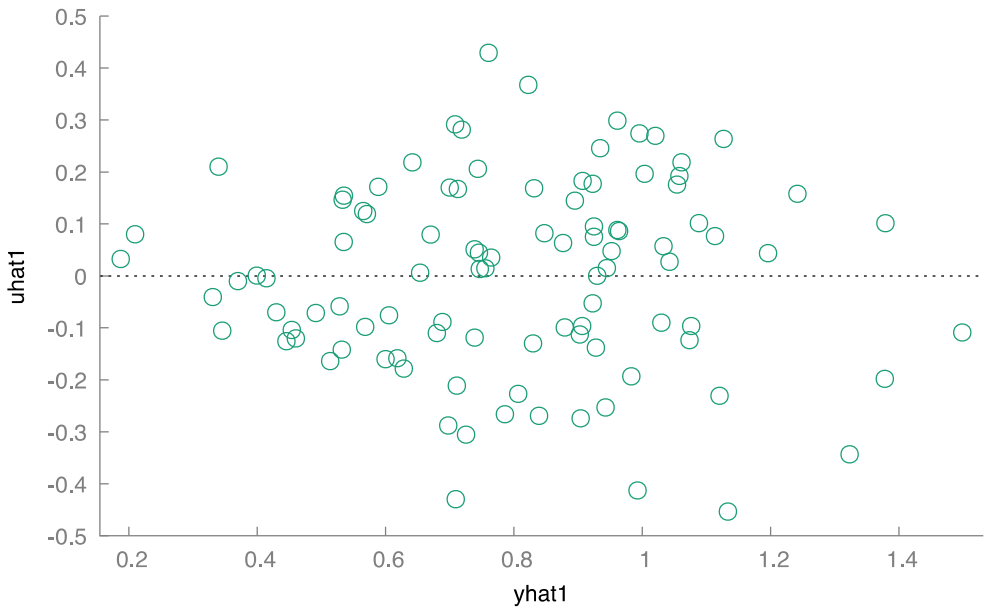
our S.E. of regression or SER which gives us another indicator of how well our model fits our data.

Sum squared resid	3.262233	S.E. of regression	0.188306

Here we have an SER of .19 meaning that the standard deviation of the regression residuals is .19 and the units are points per game. SER tells us the distance of our actual data values away from our regression line, our predicated values. .19 tells us the distance is small and there isn't a very large spread on the scatter plot, so our actual values aren't too far off from our predicted values. We want SER to be small, since we want our values as close to the regression line as possible, since our spread is minimal this tells us our predictions of PPG will be fairly accurate with the chosen regressors.



Here we have performed a residual analysis to further interpret. First, we look at our histogram of residuals uhat1. We are checking here for normality and outliers. We see some obvious outliers here when the residual is negative at around .17 it is underpredicting our variable PPG and when the residual is positive at around .15 it is overpredicting our variable PPG. I have circled a few other obvious outliers; we can see over prediction and under prediction indicating flaws in our model that we should address to increase the accuracy of our model. It is very unlikely to have a perfect normal distributed histogram, but the normal distribution is the best guideline for where we want our model to fall.



For the scatter we are looking for an even distribution of the residuals around 0 on the y axis and for them to be spread out evenly across the x axis, predicated values. We are also hoping there are no clear patterns we want a good cluster in the middle of the plot. Let's recall that uhat1 is the difference between the actual and the predicted values u` = Y – Y`, so if we have a negative residual, it tells us our model is overpredicting, if there was a clear pattern this would tell us there is a linear relationship in the residuals which is not what we want. We can kind of see this pattern in the bottom left of the plot, but luckily the pattern is not clear. It looks like our values are fairly spread out evenly across the predicated values, and we have a decent cluster in the middle. So once again just like the SER indicated our scatter plot gives us indication that our predictions of PPG will be fairly accurate with the chosen regressors.

Test on Model 4:

Null hypothesis: the regression parameters are zero for the variables
AGE, ATOI, SALARY
Test statistic: Robust F(3, 92) = 3.49522, p-value 0.0187096
Omitting variables improved 1 of 3 information criteria.

Model 6: OLS, using observations 1-100 (n = 100)
Dependent variable: PPG
Heteroskedasticity-robust standard errors, variant HC1

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	0.448234	0.133797	3.350	0.0012	***
GP	−0.0119987	0.00241024	−4.978	<0.0001	***
POS	0.181703	0.0445180	4.082	<0.0001	***
SOG	0.00497995	0.000524219	9.500	<0.0001	***

Mean dependent var	0.798889	S.D. dependent var	0.319463
Sum squared resid	3.738782	S.E. of regression	0.198382
R-squared	0.626181	Adjusted R-squared	0.614376
F(3, 95)	59.05860	P-value(F)	1.23e-21
Log-likelihood	21.70491	Akaike criterion	−35.40982
Schwarz criterion	−25.02934	Hannan-Quinn	−31.20986

We now explore the possibility of multicollinearity by omitting all insignificant variables and testing the null hypothesis that AGE, ATOI, and SALARY are jointly equal to zero. Here we see highlighted in red our F-statistic is 3.49522, with 3 regressors at the 5% significance level our critical value is 2.60. Since our F is 3.495 > 2.60 this proves, we are forced to reject the null that AGE, ATOI, and SALARY are jointly equal to zero. So, we can't omit, and we must keep these variables to avoid omitted variable bias. To explore the possibility of multicollinearity further we will look at the matrix of correlation coefficients only on the omitted variables. We want to look at the correlation between these variables as a group. A high correlation could imply multicollinearity.

Correlation coefficients, using the observations 1 – 100

5% critical value (two-tailed) = 0.1966 for n = 100

AGE	ATOI	SALARY	
1.0000	-0.0741	-0.1153	AGE
	1.0000	0.8653	ATOI
		1.0000	SALARY

As we can see SALARY and AGE are highly correlated, this implies there is a linear function of the regressors that is highly corelated with another regressor, and in result implies multicollinearity is possible. Here we see AGE is not very correlated with SALARY or ATOI, so omitting AGE shouldn't cause any bias. Although, we decide to include AGE as a control variable, because often SALARY (our variable of interest) is affected by the age of a player, older players tend to have more experience and higher salary, so it is best we include this for the logic of our model. So, we have proved multicollinearity is possible and we must keep these variables in our model.

To the General Manager of NHL:

We have tried our best to create a model to test the incentives of salary on performance in the NHL. The idea was to create a model that is going to show the impact of Points Per Game based on an increase in salary. This would create an incentivized system for teams to get more out of their star players, or better yet motivate young talented players to push over the boundaries of the sport and take their skills to new heights. Our model turned out to be generally good at predicting our dependent variable PPG, which is used to represent performance of players, this was best described with our analysis of R-Squared and SER. Although our model was a pretty good fit for our data perhaps there are a lot more detailed measures of skill that could be used to formulate a more accurate depiction of performance rather than just points per game. It would be

interesting to modify this model to see the effect on salary as our dependent variable and how the other variables would affect salary which could be helpful for GMs of NHL teams to see how accurate their chosen salary for a player could be based on several parameters of skill, are they paying this player too much? Are they paying too little? If so, should we pay more to incentivize the players performance? I believe given we tweaked this model for a more suitable approach, this could be very useful for salary cap analysis and could be used to predict incentivized player performance.