Maryna Khatnyuk

Regression Analysis

Contents

[a. Build and estimate 1](#_Toc110153432)

[Creating dummy variables for categorical predictors. 1](#_Toc110153433)

[Checking the Assumptions for regression model. 2](#_Toc110153434)

[Building Linear Regression Model. 3](#_Toc110153435)

[Conclusions. 5](#_Toc110153436)

[b. Assess the distribution of the residuals and how they may affect the suitability of the model. 5](#_Toc110153437)

[c.Prediction 7](#_Toc110153438)

[d. Collinearity of Age, Seasoninleague and Time 8](#_Toc110153439)

# Build and estimate

## Creating dummy variables for categorical predictors.

Data File: HockeyPlayerSurveyTransformed.sav

H1: Players born in the **first quarter of the year** (January, February and March) **earn more** than players born in the last quarter (October, November and December);

H2: Earnings have increased over the years covered by the data;

H3: Earnings increase as players grow older;

H4: Earnings are positively related to a player’s height;

H5: Earnings are positively related to a player’s weight;

H6: Earnings are positively related to years of experience;

H7: Captains earn more than players who are not captains;

H8: Canadian-born players earn more than players born in other all other countries

H9: In the same model, determine the position associated with the highest earnings

For H1 H8 H9 dummy variable are created

H1: Q1, Q2, Q3, Q4 variable for birth month

There are studies suggesting that players born earlier in the year earn more.

Q1 is a baseline,

Q2, Q3, Q4 are included in the model.

H8:new var Canadian indicate Canada born players (=1)

H9: created variable to indicate if player falls into one of the categories Center

Right Wing

Left Wing

Defense

Goalie

Forward

Winger

From my first assignment, Forward may be the highest paid position, it is going to be our reference category. All the other will be included into the model

Other positions will be our indicator variables to see if they add or subtract from the salary.

## Checking the Assumptions for regression model.

* Relationships between dependent and independent variables are linear.
* Normality of data of dependent variable SalaryAdj to assume normality of residuals.
* Errors are normally distributed (see section b)
* Homoscedasticity of all errors of all values of independent variables(see section b)

Chart, line chart

Description automatically generatedGraphical user interface, text, application

Description automatically generated

A picture containing chart

Description automatically generated

Right skewed distribution with multiple outliers and not anywhere close to normal distribution.

This variable requires transformation.

## Building Linear Regression Model.

Dependent variable: SalaryAdj

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***Model Summaryb*** | | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Durbin-Watson |
| 1 | 0.632a | 0.399 | 0.398 | 1233911.087103143 | 1.820 |
| a. Predictors: (Constant), Forward, Q3, Canadian, Winger, Height Height, Captain Captain, Left Left Wing, Time Year of survey, Right Right Wing, Q4, Age Player's Age, Center, Q2, Weight Weight, Seasoninleague Years of Experience, Defense | | | | | |
| b. Dependent Variable: SalaryAdj Salary Adjusted | | | | | |

Adjusted R square = 39.8% indicate the 60.2% of the NHL Players salary can be explained by other factors than our stated (independent variables).

Standard Error of Estimate is very high probably due to skewness of the data

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Coefficientsa* | | | | | | | | |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Collinearity Statistics | |
| B | Std. Error | Beta | Tolerance | VIF |
| 1 | (Constant) | 1570116.78 | 534733.97 |  | 2.94 | 0.00 |  |  |
| Time Year of survey | 12903.57 | 2909.51 | 0.04 | 4.43 | 0.00 | 0.70 | 1.42 |
| Q2 | 3232.00 | 32849.26 | 0.00 | 0.10 | 0.92 | 0.73 | 1.38 |
| Q3 | 75139.52 | 34825.90 | 0.02 | 2.16 | 0.03 | 0.75 | 1.34 |
| Q4 | 26388.67 | 36226.81 | 0.01 | 0.73 | 0.47 | 0.76 | 1.32 |
| Player's Age | -23518.28 | 4139.43 | -0.06 | -5.68 | 0.00 | 0.47 | 2.11 |
| Height Height | -25660.68 | 8531.40 | -0.03 | -3.01 | 0.00 | 0.50 | 2.01 |
| Weight Weight | 6580.17 | 1173.00 | 0.06 | 5.61 | 0.00 | 0.46 | 2.15 |
| Seasoninleague Years of Experience | 274469.82 | 6206.09 | 0.55 | 44.23 | 0.00 | 0.39 | 2.54 |
| Captain Captain | 1851183.07 | 64844.51 | 0.23 | 28.55 | 0.00 | 0.94 | 1.06 |
| Canadian | -296048.48 | 26064.60 | -0.09 | -11.36 | 0.00 | 0.96 | 1.04 |
| Center | -288613.33 | 110032.50 | -0.08 | -2.62 | 0.01 | 0.07 | 13.74 |
| Right Right Wing | -477367.80 | 111072.06 | -0.11 | -4.30 | 0.00 | 0.09 | 10.79 |
| Left Left Wing | -545089.90 | 110745.89 | -0.13 | -4.92 | 0.00 | 0.09 | 11.53 |
| Defense | -484554.34 | 108783.77 | -0.14 | -4.45 | 0.00 | 0.06 | 16.40 |
| Winger | -502374.80 | 180347.00 | -0.03 | -2.79 | 0.01 | 0.65 | 1.55 |
| Goalie | -12453.40 | 113874.30 | 0.00 | -0.11 | 0.91 | 0.14 | 7.13 |
| a. Dependent Variable: SalaryAdj Salary Adjusted | | | | | | | | |

Linear regression equation to Estimate Salary is following:

**Salary estimated = 1570116.78**

**+ 12903.57\* Time\_year\_Of\_Survey**

**+ 3232.00\*Q2 + 75139.52\*Q3 + 26388.67\*Q4**

**- 23518.28 \* Age**

**-25660.68 \*Height +6580.17 \*Weight**

**+274469.82\* Seasoninleague**

**+1851183.07 \* Captain - 296048.48\* Canadian**

**-288613.33 \* Center**

**-477367.80 \*Right**

**-545089.90 \*Left**

**-484554.34 \*Denese**

**-502374.80 \*Winger**

**-12453.40 \*Goalie**

## Conclusions.

H1: Players born in the **first quarter of the year** (January, February and March) **earn more** than players born in the last quarter (October, November and December);

With all other variables being the same, players born in the first quarter earns on average less. Other quarters (Q2, Q3,Q4) have positive coefficients.

H2: Earnings have increased over the years covered by the data;

Yes, on the average, each season earnings increase by $**12,903.57**

H3: Earnings increase as players grow older;

According to the model, on the average earnings decrease by 23518.28 with each additional year of player’s age.

H4: Earnings are positively related to a player’s height;

On contrary, on average salary decreases for every additional inch of a player’s height by 25660.68

H5: Earnings are positively related to a player’s weight;

Yes, the model indicates that earnings are positively related to a player’s weight with coefficient 6580.17.

H6: Earnings are positively related to years of experience;

Earnings are positively related to years of experience on the average adding 274469.82 with each year.

H7: Captains earn more than players who are not captains;

The model supports this claim (positive coefficient = 1851183.07)

H8: Canadian-born players earn more than players born in other all other countries

No, that is not true. On the average Canada born player earns 296048.48 less.

H9: In the same model, determine the position associated with the highest earnings

Forward player position was my base category.

Taking in account that all other positions have negative coefficient, I can conclude that Forwards earn on average more. Coefficient for goalie is not significant, though.

# b. Assess the distribution of the residuals and how they may affect the suitability of the model.

Chart, histogram

Description automatically generated

Standardized residual plot is slightly right-skewed with positive kurtosis but normality of residuals is not severely violated.

Chart, line chart

Description automatically generated

P-P plot observed probability doesn’t exactly fit into expected probability which indicated violation of normality.

Chart, scatter chart

Description automatically generated

Residuals are not symmetrical and have non-constant value across Salary variable. Error increases as Salary increases. Which can be explained by many outliers in higher value salary range and supports violation of homoscedasticity. It suggests the transformation of the variables before regression analysis for better model.

# c.Prediction

Based on your results from part (a) what is your prediction for how much a 30 year old goalie born in Canada in the second quarter of the year, who is 72 inches tall, weighs 170 pounds, has played professional hockey for 10 years, and who is not a team captain, would have earned in the 2000/01 season?

Calculation (see Part\_C\_Calculation.xlsx)

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Coefficient** | **Predictor value** | **Variable impact** |
| **(Constant)** | 1,570,116.78 |  | 1,570,116.78 |
| **Time Year of survey** | 12,903.57 | 10 | 129,035.70 |
| **Q2** | 3,232.00 | 1 | 3,232.00 |
| **Q3** | 75,139.52 | 0 | 0.00 |
| **Q4** | 26,388.67 | 0 | 0.00 |
| **Player's Age** | -23,518.28 | 30 | -705,548.40 |
| **Height, inches** | -25,660.68 | 72 | -1,847,568.96 |
| **Weight, pounds** | 6,580.17 | 170 | 1,118,628.90 |
| **Years of Experience** | 274,469.82 | 10 | 2,744,698.20 |
| **Captain** | 1,851,183.07 | 0 | 0.00 |
| **Canadian** | -296,048.48 | 1 | -296,048.48 |
| **Center** | -288,613.33 | 0 | 0.00 |
| **Right Right Wing** | -477,367.80 | 0 | 0.00 |
| **Left Left Wing** | -545,089.90 | 0 | 0.00 |
| **Defense** | -484,554.34 | 0 | 0.00 |
| **Winger** | -502,374.80 | 0 | 0.00 |
| **Goalie** | -12,453.40 | 1 | -12,453.40 |
| **Total annual earning prediction** |  |  | **$2,704,092.34** |

# d. Collinearity of Age, Seasoninleague and Time

* Define what is meant by “multicollinearity” with particular reference to these three variables (i.e., why might a researcher be concerned about multicollinearity with respect to these three variables?).

The age of a player may correlate with years of experience. The older the player the more experience he has. Regarding time of survey, if the same players were surveyed, every new season they became older one year and get one year of experience more. So, this tree variable may correlate if survey used the same players for survey during all 9 seasons.

Multicollinearity reduces the precision of the estimated coefficients, which weakens the statistical power of the regression model.

* What evidence can you provide to assist in determining whether multicollinearity among these three variables might be a concern in the model that you have estimated?

The tolerance level for all tree variable is more than 0.10

VIF value is not ideal for two variables, slightly over 2. Cut-off level for rejecting is 10, therefore, we have not violated the multicollinearity assumption.

I generated correlation matrix to explore strength of correlation for this three variables.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Correlations* | | | | |
|  | | Time Year of survey | Player's Age | Years of Experience |
| Time Year of survey | Pearson Correlation | -- |  |  |
| N | 9985 |  |  |
| Player's Age | Pearson Correlation | **0.172\*\*** | -- |  |
| Sig. (2-tailed) | 0.000 |  |  |
| N | 9971 | 9972 |  |
| Years of Experience | Pearson Correlation | **0.462\*\*** | **0.675\*\*** | -- |
| Sig. (2-tailed) | 0.000 | 0.000 |  |
| N | 9981 | 9968 | 9982 |
| \*\*. Correlation is significant at the 0.01 level (2-tailed). | | | | |

There is weak positive correlation between variables Tame and age, Time, and Experience. The highest Pearson correlation coefficient 0.675 between Years of Experience and Age indicates moderate positive correlation.

I would remove player’s age from the model to see if the model is stronger without it.