**Assignment 2**

**Independent Samples t-test and Regression**

**Submitted by**

**Mazaharuddin Mohammad**

**N01464978**

**Question 1**

**Background:**

A courier company operates a number of depots in the GTA. A depot is the building from which the courier routes are dispatched with the day’s deliveries. There is one depot on the west side of the GTA (DEPOT\_15) and another depot on the east side of the GTA (DEPOT\_18) that preliminary analysis (taking averages) indicates may be experiencing differences in a number of key performance indicators (KPIs). The two KPIs of particular interest to upper management are the percent of packages delivered on-time (on-time percentage, or OTP) and the number of delivery stops the couriers make per hour (stops per hour, or SPH).

**Your task:**

Your task is to determine, where possible, if there is a statistically significant difference in these two KPIs between the two depots.

**Data given:**

You have been given the daily OTP and SPH performance data for both depots for the past 18 years (assume no changes in technology or management over that timeframe). Each record in the “A TALE OF TWO DEPOTS.xlsx file >> 18\_YR\_DATA tab” represents a day at the particular depot listed under DEPOT\_ALIAS.

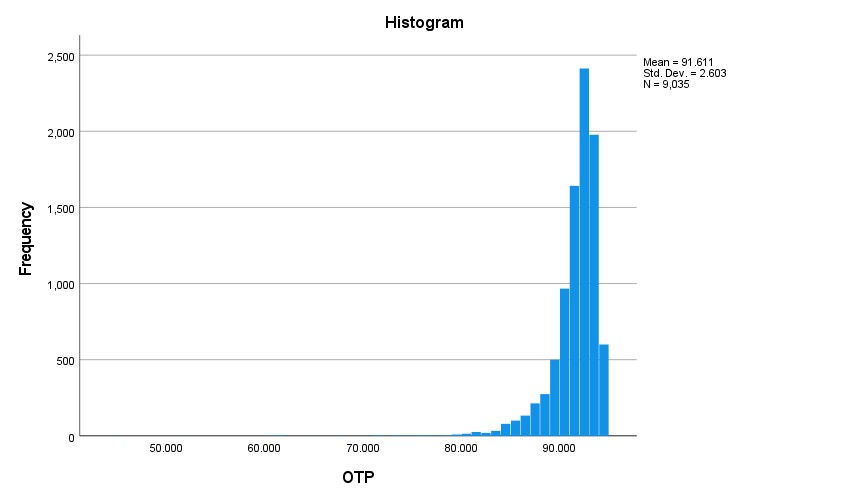
1. **OTP:**
   1. **Conduct tests to determine if an independent samples t-test can be conducted on the OTP data. That is, does the data violate any assumptions which prevent the independent sample t-test from being conducted? Provide evidence (graphs, any test results) to support your decision.**

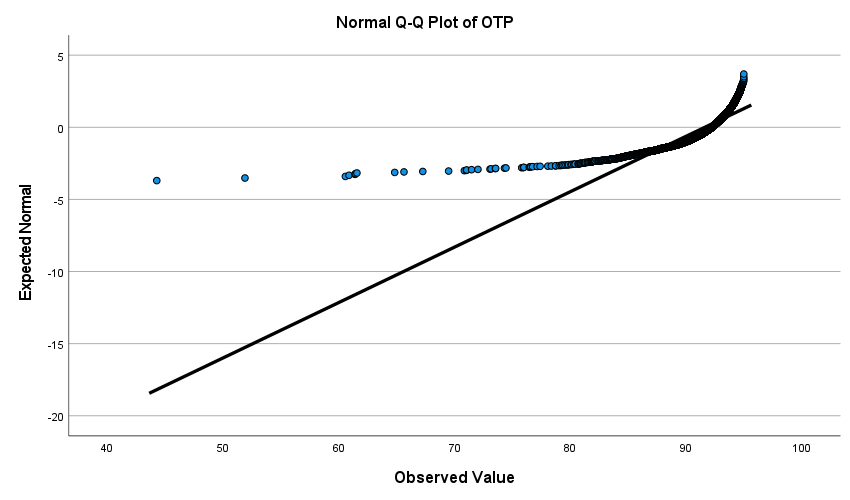
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Case Processing Summary** | | | | | | |
|  | Cases | | | | | |
| Valid | | Missing | | Total | |
| N | Percent | N | Percent | N | Percent |
| OTP | 9035 | 100.0% | 0 | 0.0% | 9035 | 100.0% |

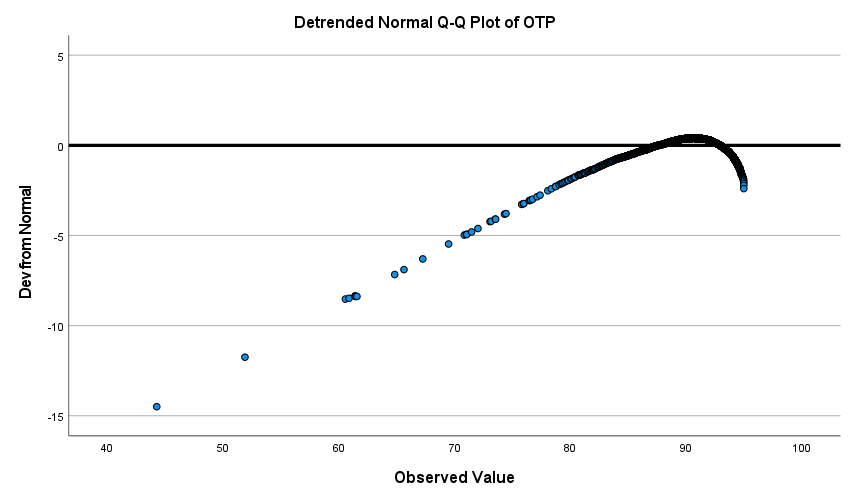
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Descriptives** | | | | |
|  | | | Statistic | Std. Error |
| OTP | Mean | | 91.61142 | .027387 |
| 95% Confidence Interval for Mean | Lower Bound | 91.55773 |  |
| Upper Bound | 91.66510 |  |
| 5% Trimmed Mean | | 91.89626 |  |
| Median | | 92.21798 |  |
| Variance | | 6.777 |  |
| Std. Deviation | | 2.603192 |  |
| Minimum | | 44.265 |  |
| Maximum | | 94.983 |  |
| Range | | 50.718 |  |
| Interquartile Range | | 2.247 |  |
| Skewness | | -4.093 | .026 |
| Kurtosis | | 37.960 | .052 |

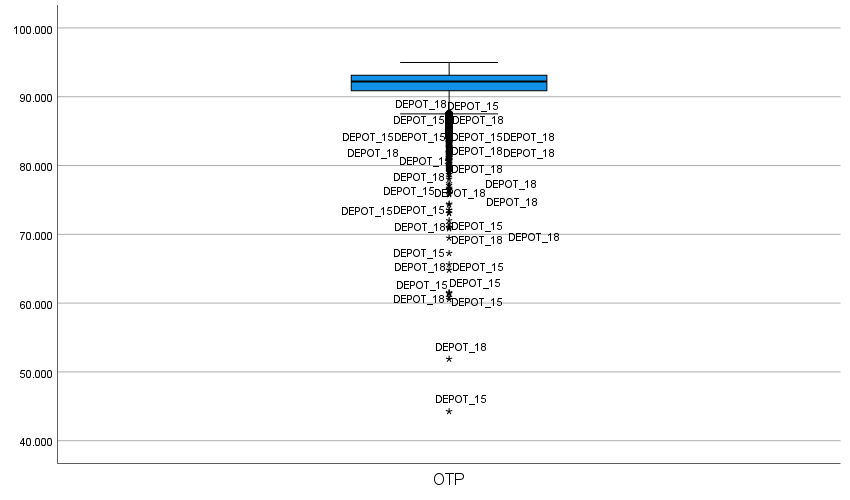
|  |  |  |  |
| --- | --- | --- | --- |
| **Tests of Normality** | | | |
|  | Kolmogorov-Smirnova | | |
| Statistic | df | Sig. |
| OTP | .145 | 9035 | .000 |

|  |
| --- |
| a. Lilliefors Significance Correction |









1. **SPH:**
2. **Conduct tests to determine if an independent samples t-test can be conducted on the SPH data. Does it violate any assumptions which prevent the independent sample t-test from being conducted? Provide evidence (graphs, any test results) to support your decision.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Case Processing Summary** | | | | | | |
|  | Cases | | | | | |
| Valid | | Missing | | Total | |
| N | Percent | N | Percent | N | Percent |
| SPH | 9035 | 100.0% | 0 | 0.0% | 9035 | 100.0% |

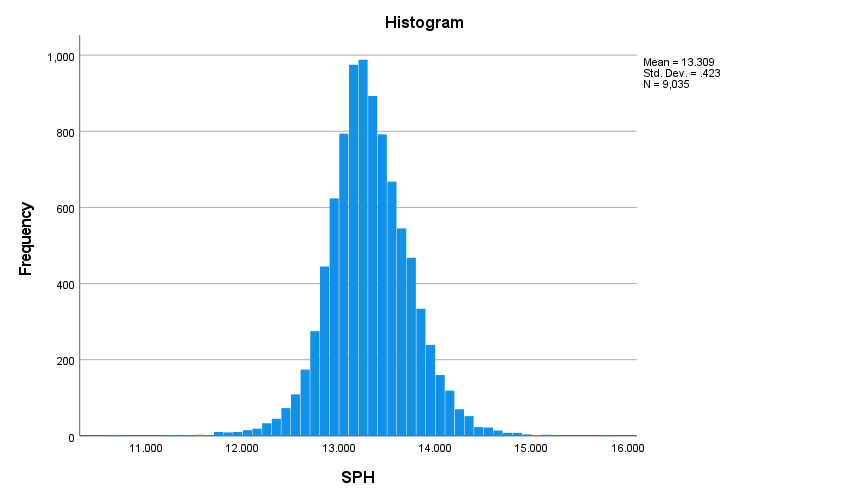
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Descriptives** | | | | |
|  | | | Statistic | Std. Error |
| SPH | Mean | | 13.30936 | .004447 |
| 95% Confidence Interval for Mean | Lower Bound | 13.30064 |  |
| Upper Bound | 13.31807 |  |
| 5% Trimmed Mean | | 13.30724 |  |
| Median | | 13.29000 |  |
| Variance | | .179 |  |
| Std. Deviation | | .422732 |  |
| Minimum | | 10.680 |  |
| Maximum | | 15.710 |  |
| Range | | 5.030 |  |
| Interquartile Range | | .520 |  |
| Skewness | | .079 | .026 |
| Kurtosis | | 1.452 | .052 |

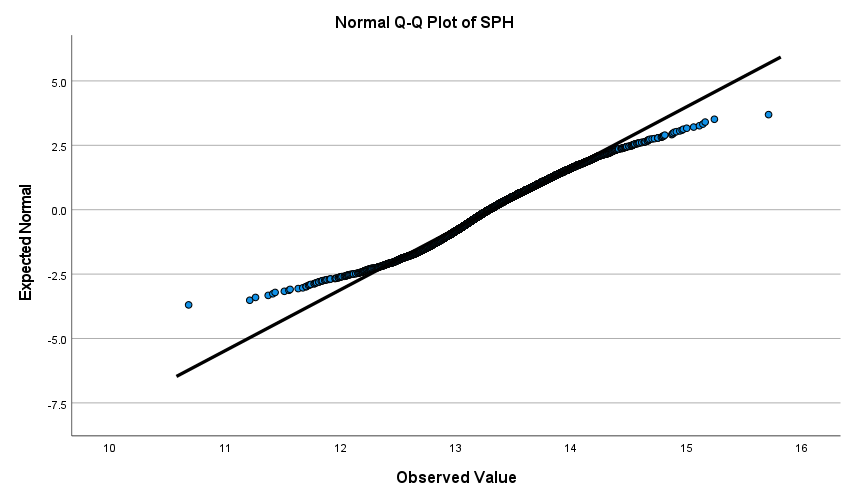
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Extreme Values** | | | | | |
|  | | | Case Number | DEPOT\_ALIAS | Value |
| SPH | Highest | 1 | 530 | DEPOT\_18 | 15.710 |
| 2 | 529 | DEPOT\_18 | 15.240 |
| 3 | 463 | DEPOT\_18 | 15.160 |
| 4 | 542 | DEPOT\_18 | 15.140 |
| 5 | 538 | DEPOT\_18 | 15.110 |
| Lowest | 1 | 309 | DEPOT\_18 | 10.680 |
| 2 | 348 | DEPOT\_18 | 11.210 |
| 3 | 305 | DEPOT\_18 | 11.260 |
| 4 | 12 | DEPOT\_15 | 11.370 |
| 5 | 301 | DEPOT\_18 | 11.410 |

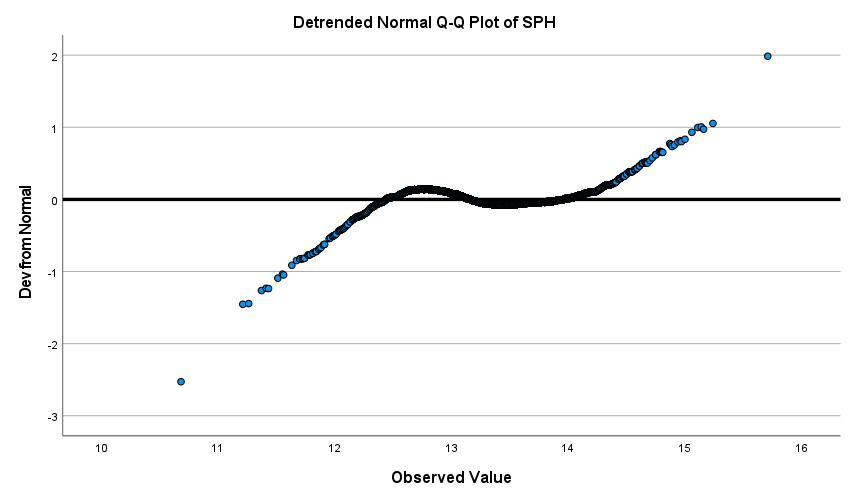
|  |  |  |  |
| --- | --- | --- | --- |
| **Tests of Normality** | | | |
|  | Kolmogorov-Smirnova | | |
| Statistic | df | Sig. |
| SPH | .034 | 9035 | .000 |

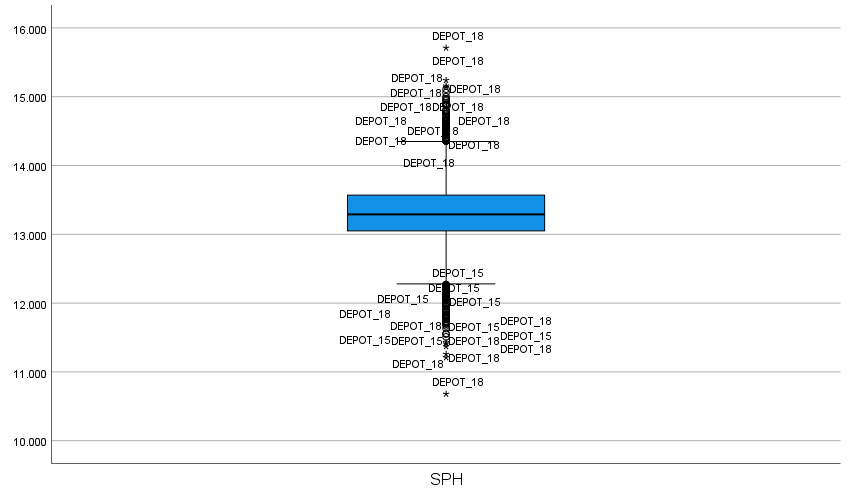
|  |
| --- |
| a. Lilliefors Significance Correction |

**SPH**

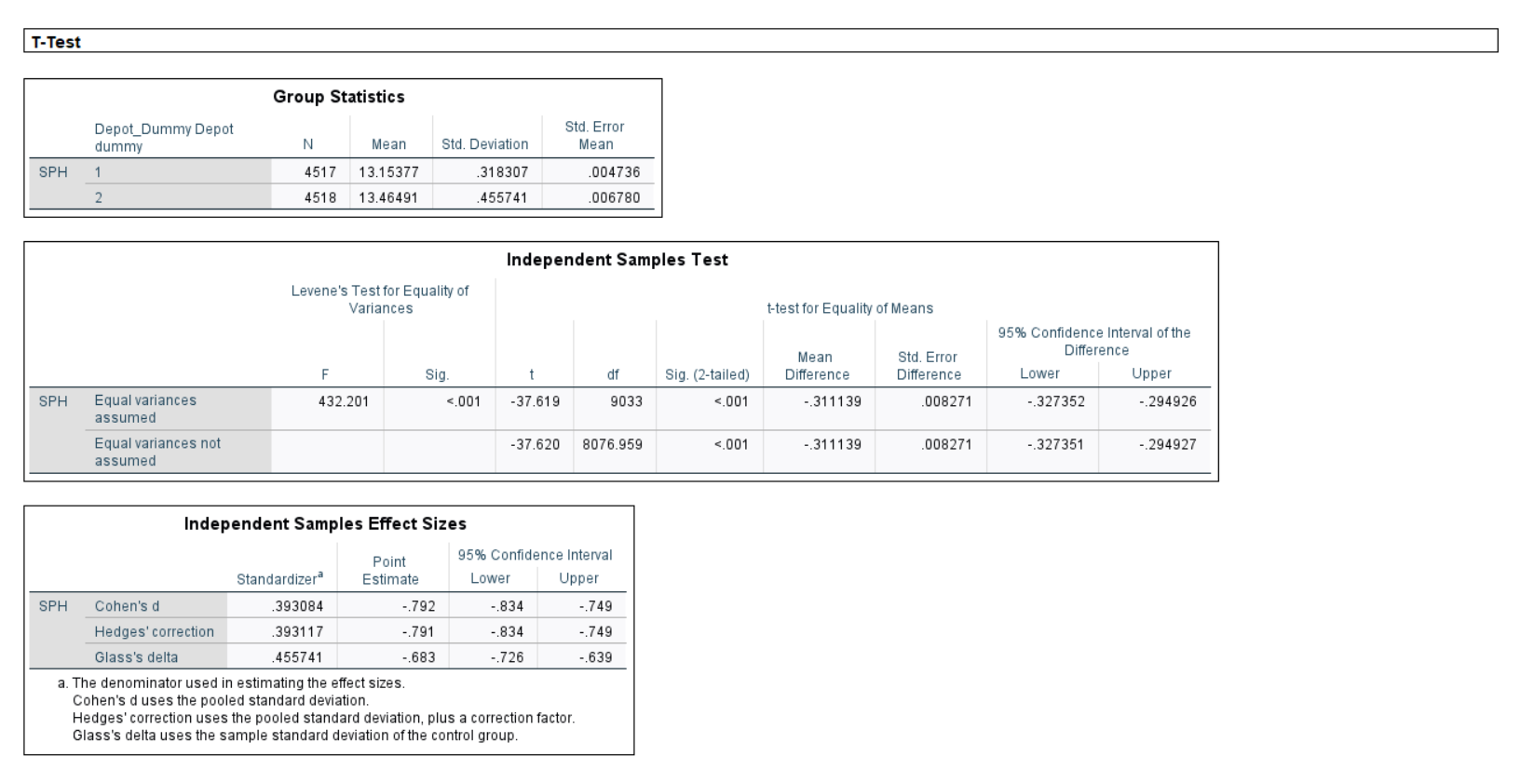








1. **Assuming that 2a allows for the independent samples t-test to be conducted, determine if there is a statistically significant difference in SPH for the two depots. Note: You may need to create a dummy variable (for example DEPOT\_15=1 and DEPOT\_18=2) in order to conduct the t-test.**



An independent samples t-test was conducted to compare the SPH (Stops per hour) scores for DEPOT\_15 and DEPOT\_18.

There is a significant difference in scores for DEPOT\_15 (M = 13.15377, SD = 0.318307) and DEPOT\_18 (M = 13.46491, SD = 0.455741)

If Sig value under independent samples Test table is

> 5% then we should consider t-value of Equal variances assumed and

<=5% then we should consider t-value of Equal variances not assumed.

Hence t value is represented as t (9033) = -37.619 and p value is less than 0.001, two-tailed).

To identify the difference between groups we should refer to Sig (2-tailed) column under Independent samples Test table.

If Sig (2-tailed) value under independent samples Test table is

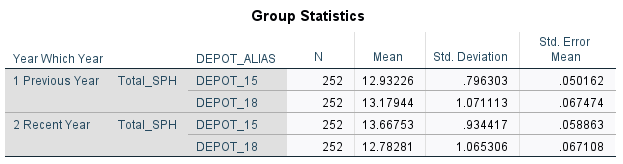
<= 5% then we can conclude that there is a significant difference between two depots.

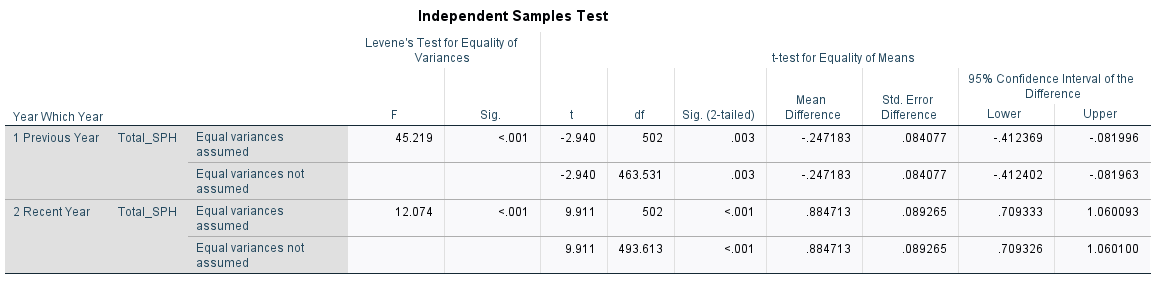
> 5% then we can conclude that there is no significant difference between two depots.

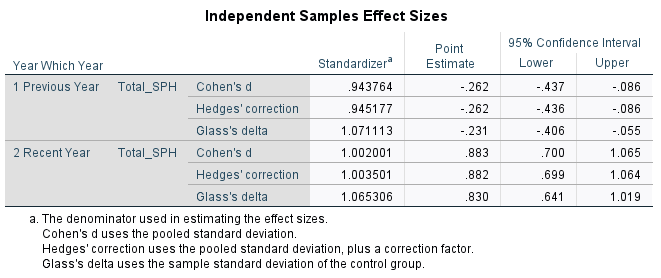
Hence, from the above figures we can conclude that there is a significant difference in SPH (Stops per hour) scores for DEPOT\_15 and DEPOT\_18.

1. **Fast forward one year: Based on the results of your analysis in 2b, the company conducted an intervention and swapped the Senior Managers between DEPOT\_15 and DEPOT\_18. They now have a year of productivity data under the new management structure. The latest year’s data with the new management structure is shown in the column ”STOPS\_PER\_HR\_2”, and the previous year’s data under the old management structure is shown in the column “STOPS\_PER\_HR” on the “AFTER SR MGR SWAP” tab. Assume that the data meets all the assumptions required to conduct an independent samples t-test.**
   * 1. **Conduct an independent-samples t-test for both depots comparing their SPH performance in the latest year to the previous year. Note: you will have to perform some data manipulation (put all the SPH data into one column) and create a YEAR variable.**

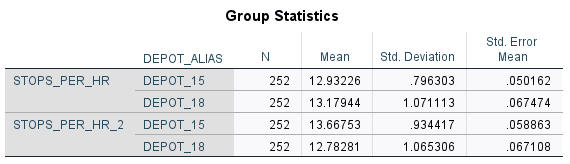
**T-Test**

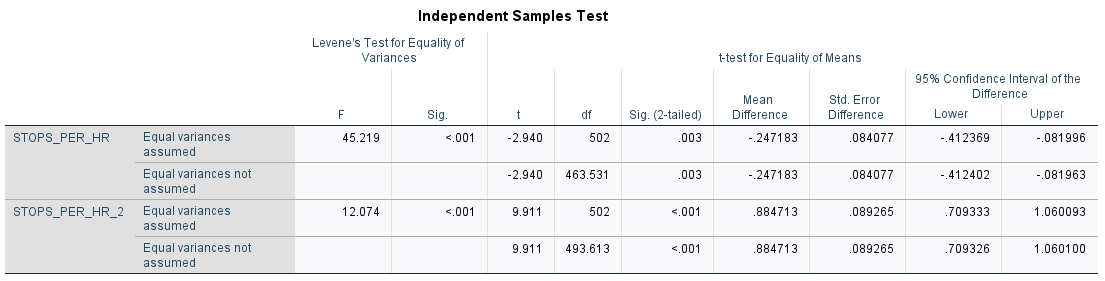


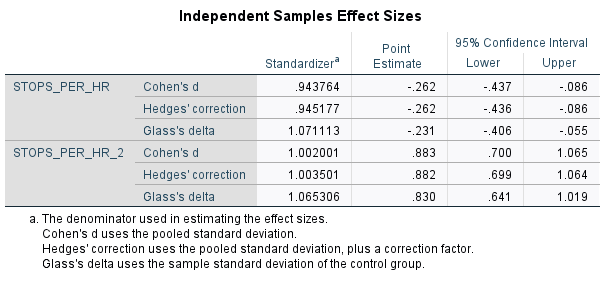




* + 1. **Did the intervention of swapping the Senior Managers between depots have a statistically significant impact on the stops per hour productivity at either depot?**







The significant values for SPH in both depots are 0.003 and 0.001 which is less than 0.05(5%). If the significant values are less than 5 % then we can conclude that there is a significant difference between two groups.

However, even after swapping the managers the significant value is less than 5%. So, we can conclude that intervention of swapping the managers has a significant impact between two groups.

* + 1. **What will you advise the company regarding the management situation at DEPOT\_15 and DEPOT\_18?**

We can compare mean values of both depots from previous year to latest year. From the Group statistics table we can see that DEPOT\_15 in previous year mean value is less than the recent year. So, there was a significant decrease from previous year to latest year in SPH score. However, there is a slight increase in mean value of DEPOT\_18 in latest year from previous year.

Hence we can conclude that intervention of swapping the managers has been slightly productive for DEPOT\_18 than DEPOT\_15.

**Question 2**

**The following analysis is to be done in SPSS. The data necessary for the analysis is in the SPSS file “Hockey Players – Salary Survey.sav”.**

a) Estimate a single linear regression model that will allow you to test all of the following hypotheses (NOTE: Do not transform any of the variables to render them normally distributed). Make sure that you screen the data and check all the necessary assumptions before proceeding with your analysis.

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;

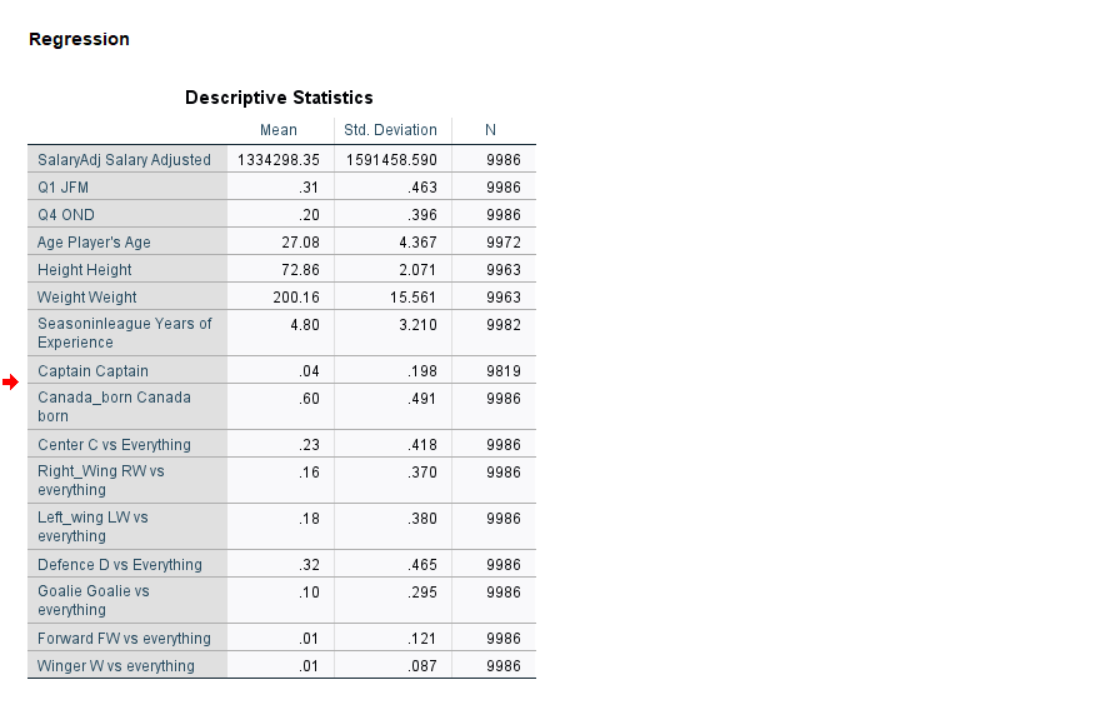
H5: Earnings are positively related to a player’s weight;

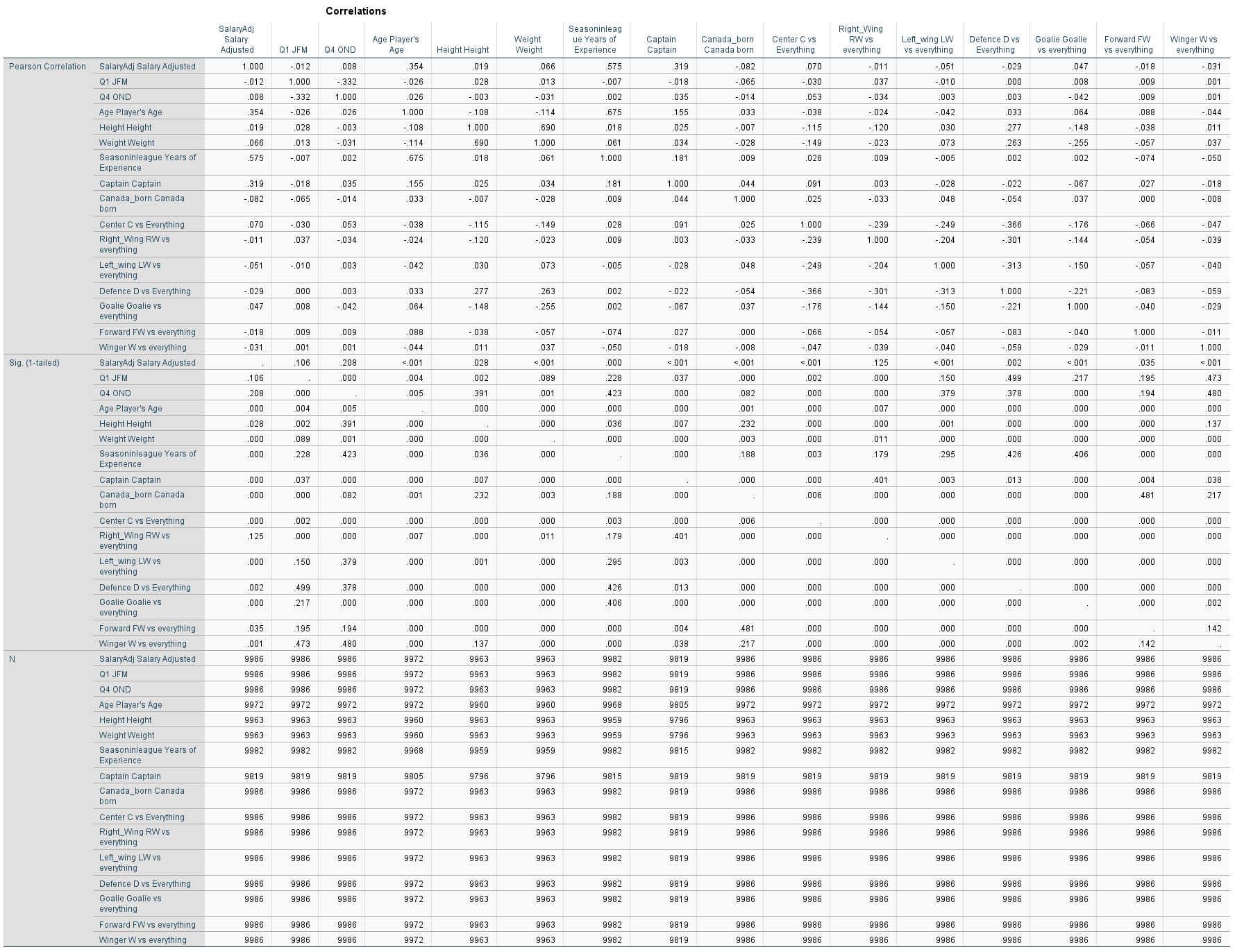
H6: Earnings are positively related to years of experience;

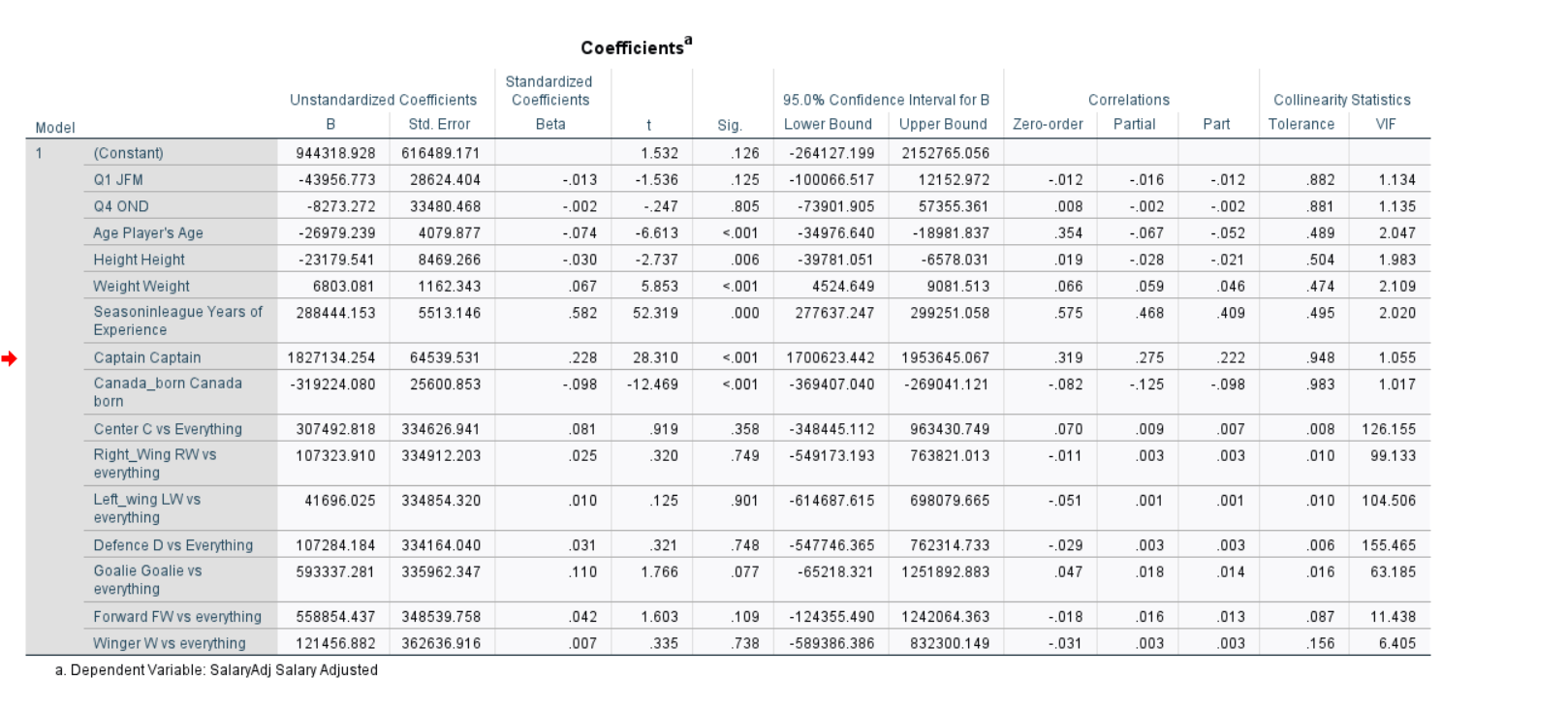
H7: Captains earn more than players who are not captains;

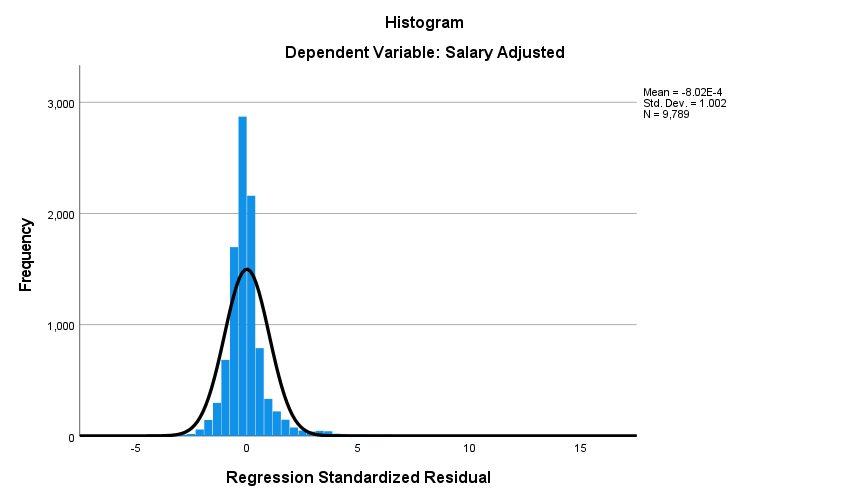
H8: Canadian-born players earn more than players born in other all other countries

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





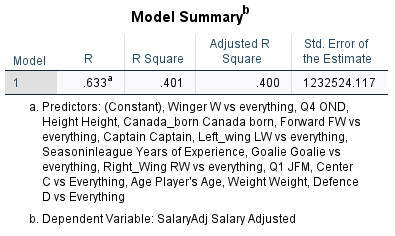




In the Co-relations table, if correlation value is

>= 0.7 then we should consider highly correlated variable and take out from the analysis

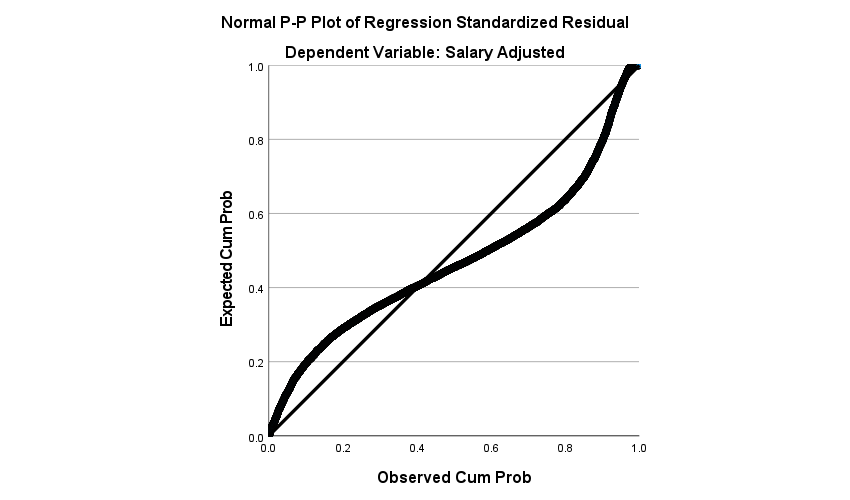
< 0.7 then we don’t have to eliminate any variable from the table.

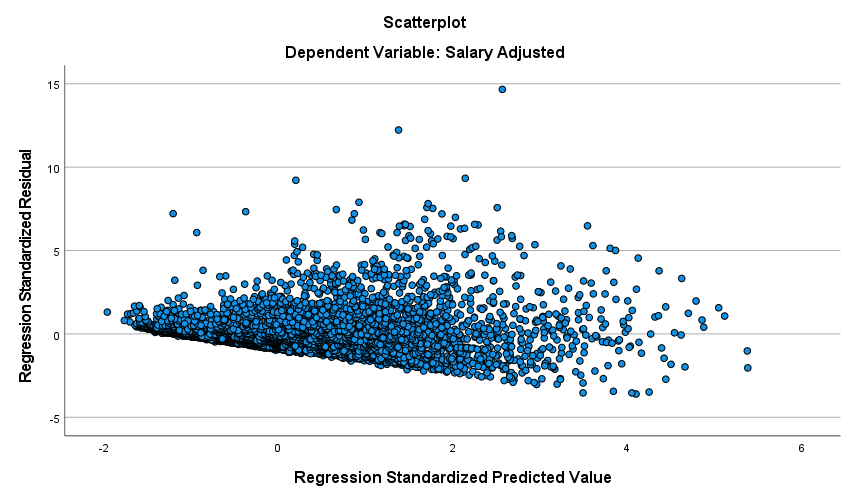


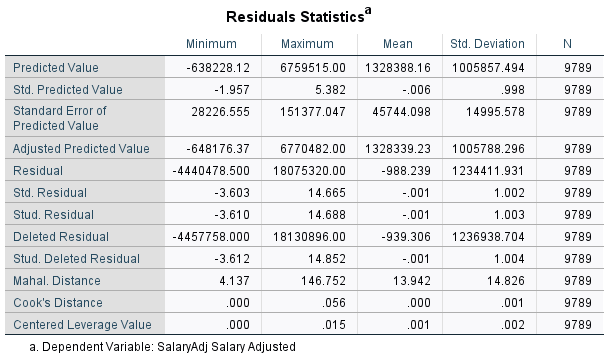
R square value tells the proportion of variance in the dependent variable which is predicted from independent variables.

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

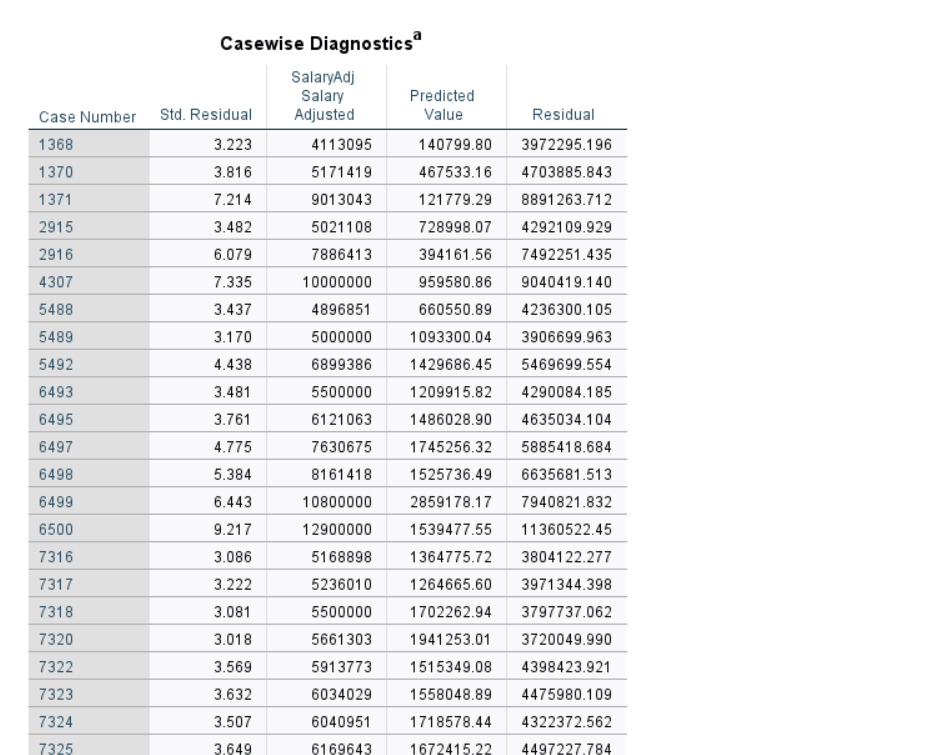
Residuals analysis assumptions can be performed by referring to Normal P – P plot and Scatterplot







The cook’s distance value should be always less than 1. From the above table we can conclude that cook’s distance value is 0.056 and so there shouldn’t be any concern.



The cases which are mentioned in Case-wise diagnostics table can be known as outlier residuals.

**c) 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?**

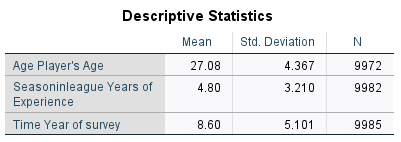
Based on results we can predict that the following player is earning around 2829007

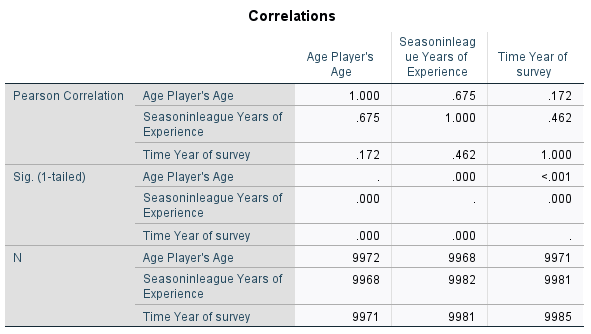
d) The variables Age, Seasoninleague and Time may be collinear.

* **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?).**

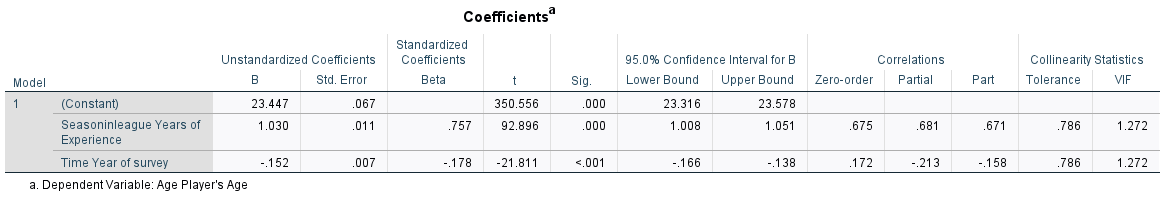
Multicollinearity exists when the independent variables are highly correlated.

**Regression**





* **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?**



From co-relations table we can conclude that if co-relation value is

>= 0.7 then we should consider highly correlated variable and take out from the analysis

< 0.7 then we don’t have to eliminate any variable from the table.

Also, under collinearity statistics from the coefficients table we can check Tolerance and VIF (Variance Inflation Factor) values for these variables.

If the Tolerance value is < 0.1 then we can conclude that the following variable is highly multiple correlated with other variables. From the above table we can that Tolerance value is 0.786 which is greater than 0.1 so we can say that following variables are normally correlated with other variables.

VIF is simply calculated by 1 / T.

If the VIF value is > 10 then there should be a concern with the following variable. From the above table we can say that VIF value is 1.272 which is less than 10 so there should not be any concern with other variables in the analysis.

**Notes:**

Variable Name Description

Time 0= 1990/91 season; 1= 1991/92 season ... to 17=2007/08 season

Birthmonth 1= January; 2= February ... to 12= December

SalaryAdj Inflation-adjusted annual earnings (in U.S. dollars)

Birth\_Country Country in which the player was born:

Canada Czech Republic

United States Slovakia

Russia Former Soviet Republic

Sweden Rest of Europe

Finland Rest of World

Age Age in years

Height The player’s height in inches

Weight The player’s weight in pounds

Seasoninleague Number of years of experience as a professional hockey player

Captain Dummy variable equal to 1 if player is the team captain; 0 otherwise

Position C = Centre;

R = Right Wing;

L = Left Wing;

D = Defence;

G = Goalie;

F = Forward (plays multiple positions including Centre, Right Wing and Left wing).

W=Winger (plays either wing)