**Binary Logistic Regression using SPSS**

**Dataset:**

**Source -** <https://catalogue.data.govt.nz/dataset/wcc-tracks>

Wellington City Council provides information related to local roads, footpaths, parks, recreational area, sewage system etc. The dataset being used for analysis contains information related to location and details of Wellington City Council tracks within Wellington area.

**Variable Description:**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Type | Variable Label | Description |
| Type\_class | Independent variable | Typeclass | The type of track in consideration, categorised as 1=4WD, 2=Accessway, 3=Mountain Bike Track, 4=Road Section, 5=Shared Use |
| Walk\_class | Independent variable | Walkclass | The type of track for walking, 1=Path, 2=Short walk, 3=Tramping Track, 4=Walking Track |
| Shape\_length | Independent variable | Shape\_length | The size of the track |
| Vehicle\_accessible | Dependent variable | Vehicle | Is the vehicle accessible on the track? 1=Yes, 0=No. |

**Table 1. Description of Dependent and Independent variables.**

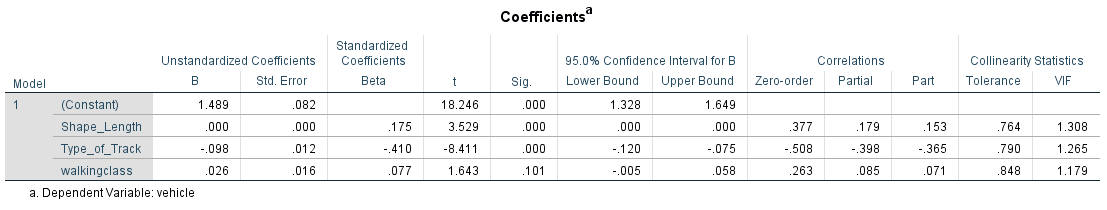
**Objective:**

For analysis we shall be predicting if the tracks within the Wellington City area is accessible for Vehicles or not based on the independent factors related to the tracks.

**Step 1: Checking the Assumptions**

**Sample size:** The sample size to be chosen for binary logistic regression must not be of large size as this will produce unexpected results for our analysis. The dataset contains 1362 rows, for our analysis the sample size chosen is 379 which is roughly equivalent to 25% of the rows.

**Multicollinearity:** To check for multicollinearity between the independent we shall be using the collinearity statistics test for multiple regression to check the tolerance value for different independent variables. A tolerance value less than 0.1 indicates high collinearity between the independent variables.



**Figure 1: Coefficients Statistic**

From the above table we can see that the tolerance values for each independent variable is higher than 0.1 meaning there is no collinearity between these variables.

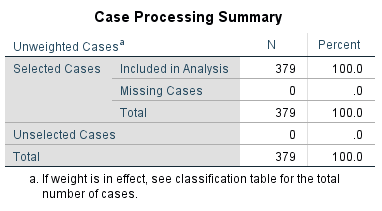
**Step 2: Data preparation**

In this dataset there were many columns that were irrelevant for our analysis hence those columns were dropped from the final dataset. In the columns selected, the dataset included rows which had N/A values which were eliminated for our analysis as they lead to improper results.

Variable coding was done to each of the categorical variables and our dichotomous dependent variable for which the description is given in Table 1. As detailed in the variable description in Table 1 our dichotomous variable was coded as 0 and 1. For a value of 1, the response to our objective would be Yes, the vehicle is accessible in the cycle tracks and for a value of 0, the response will be No, the vehicle is not accessible in the cycle tracks. Shape\_length is a continuous variable, higher the value of shape\_length indicates wider and longer track.

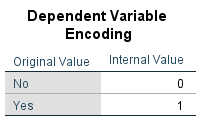
**Step 3: Analysis**

For our analysis we had taken 400 cases from our dataset. As you can see below from the Case Process Summary, all the 400 cases have been included for our analysis.



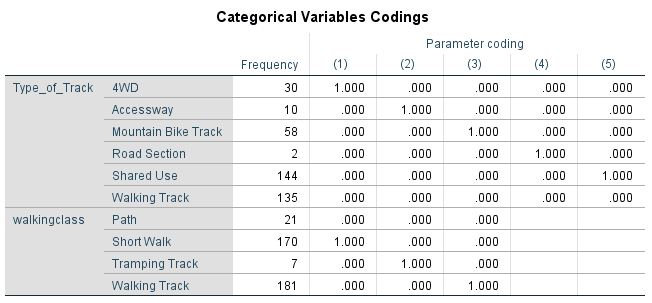
**Figure 2: Case Processing Summary.**

SPSS has a default encoding format of 0 and 1 for the dependent variable. For this purpose, our dependent variable vehicle\_accesible (label-vehicle) was encoded as 1=yes the vehicle is accessible and 0=No the vehicle is not accessible. This encoding helps us understanding the result easily.



**Figure 3: Dependent Variable Encoding.**

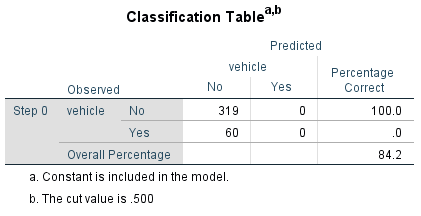
In case of our independent variables, for our analysis we have two categorical variables walkingclass and type\_of\_track and one continuous variable shape\_length. The categorical variables are encoded as below.



**Figure 4: Categorical Variable Encoding.**

**Analysis of Block 0**

In next step of our analysis, we look at Block 0 of the output. Block 0 output is based on analysis without taking into consideration the independent variables. This serves as a baseline for comparing our results when the independent variables are included in the analysis. For initial analysis, SPSS takes into consideration that none of the tracks are accessible for vehicle. From the below figure we can see that overall percentage of accuracy of predicting no as output was 84.2%. In our analysis with independent variables included, we should be expecting an accuracy of over 84.2%.



**Figure 5: Block 0 Classification Table.**

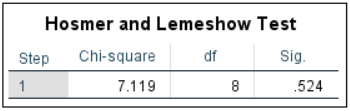
**Analysis of Block 1**

In Block 1, the Omnibus Tests of Model Coefficients is an indication of performance of our model with no independent variables included in the model. This test is known as ‘Goodness of fit’ of the model. The significant value for Omnibus Tests must be less than the p value of 0.0005 which indicates the method considered in Block 1 is better than Block 0. In our analysis we can see from the below table we have a significance value of 0.000 indicating our model performed better than the model SPSS considered in block 0. The Chi-Square value we obtained is 176.660 with 9 degrees of freedom.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Omnibus Tests of Model Coefficients** | | | | |
|  | | Chi-square | df | Sig. |
| Step 1 | Step | 176.660 | 9 | .000 |
| Block | 176.660 | 9 | .000 |
| Model | 176.660 | 9 | .000 |

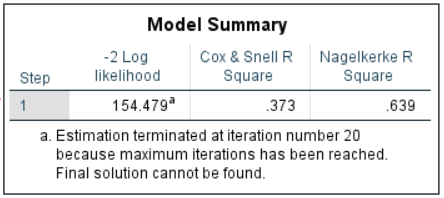
**Figure 6: Omnibus Tests of Model Coefficients.**

According to the definition mentioned by Wikipedia (2019), Hosmer-Lemeshow test is a statistical test for goodness of fit for logistic regression models. It is used frequently in risk prediction models. The test assesses whether or not the observed event rates match expected event rates in subgroups of model population. For this test, the significance value must be less than 0.05. From the results obtained, we can see that the significance value for Hosmer-Lemeshow test is 0.524 and Chi-Square value to be 7.119. This indicates that the logistic regression model is good fit for our analysis.



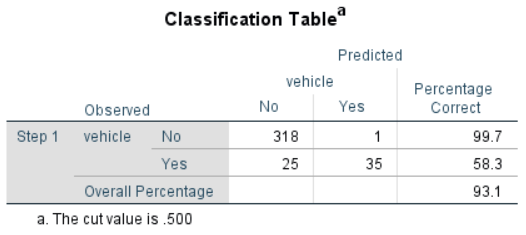
**Figure 7: Hosmer and Lemeshow Test.**

The amount of variation of the dependent variable in the model is explained by the Cox and Snell R Square and Nagelkerke R Square values. In our analysis, the values are 0.373 and 0.639 indicating 37.3% and 63.9% of variation respectively.



**Figure 8: Model Summary.**

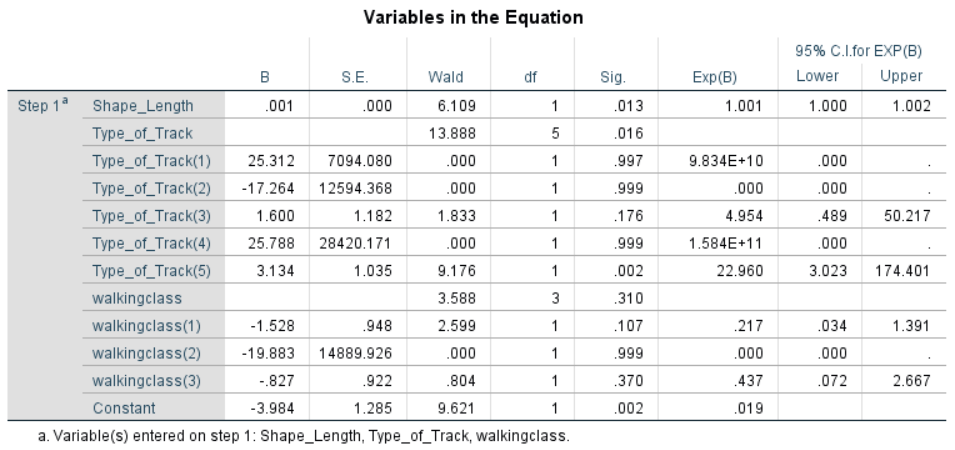
The classification table in Block 1 indicates how much increase there is in the prediction of our model when the independent variables are included. From the table below we can see that our model is 93.1% correct in prediction. There is a significant increase in prediction of our model.



**Figure 9: Block 1 Classification Table.**

Pallant (2005) says that the positive predictive value is the percentage of cases that the model classifies as having the characteristics that is actually observed in the group. In our analysis, this value will be equal to number of values predicted correstly as yes by total predicted yes. This gives us 35/36 equal to 99.7 percent. Pallant (2005) also says that the negative predictive value is the percentage of cases not to have the characteristic that is actually observed not to have the characteristic. In our analysis, this value is given by 318/343 which is equal to 92.7%.

The table variables in the equations provides information related to each of our independent variables and their significance towards predicting the dependent variable. We will be using the Wald test to find the significance of prediction of each variable towards the model. The significance value for Wald test must be less than 0.05, from the table below we can see that two variables contribute significantly towards prediction. They include type\_of\_track=0.002 and shape\_length = 0.013.

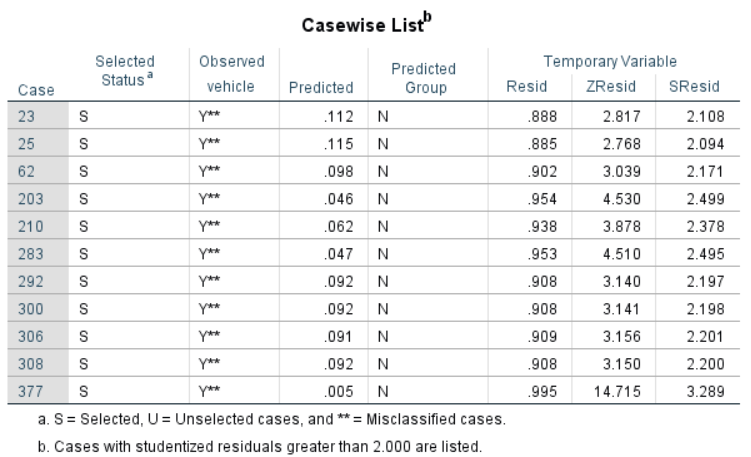


**Figure 10: Variables in the Equation.**

The B values in the table are those which we can use in the equation to find the probability of a case falling into one of the categories of predicting variable. The sign of the B value indicates the direction of relationship which shall increase the outcome to be yes and which shall decrease it. A negative value of B value in our case indicates that the outcome will be that the vehicle is not accessible in that track.

The value Exp(B) is called the odds ratio. According to Tabachnick and Fidell (2001), the odds ratio is ‘the increase (or decrease if the ratio is less than one) in odds of being in one outcome category when the value of predictor increases by one unit’. In our analysis the vehicle being accessible in the track is 22.960 times higher for every 3.134 unit increase of value of B. In our analysis we also have one continuous variable shape\_length, we can say that 0.001 unit in increase of shape\_length the odds will increase by 1 unit. Also, in the above table we have confidence interval (CI) for each variable. For example, the CI for walking3 is .072 to 2.667 meaning we can be sure that the odds ratio will be within this interval which is true (OR=.437).

The last table in the output labelled Casewise List, gives us information about the cases for which the model does not fit well, the outcome which were supposed to be yes were predicted to be no and vice-versa. They also indicate that they are outliers for our analysis and must be examined closely.



**Figure 11: Casewise List.**

**Results:**

Based on the above analysis

1. Independent variables type\_of\_track and shape\_length contributed significantly towards predicting if a vehicle is accessible on a track or not.
2. We can also say that Binary Logistic Regression model was a good fit as per Hosmer-Lemeshow test.

**References:**

Tabachnich, B.G. and Fidell, L.S. (2001) *Multivariate Statistics* (4th edn). Boston: Allyn and Bacon. Chapter 12.

Pallant, J.F. (2005) *SPSS survival manual* (2nd edn). NSW: Allen and Unwin. Chapter 14.

**Multiple Linear Regression using SPSS**

**Abstract:**

River water quality is considered as an important factor for many reasons from its role in supporting people and industrial needs, underwater habitat and ecological functions. River water quality depends on many factors namely nutrients such as phosphorous, nitrogen etc, macroinvertebrate content and several other factors. The nutrients are essential for plant life and underwater animals but too much content of these nutrients can be toxic, similarly increased content of macroinvertebrate can lead to illness of underwater plants. In my research we shall be using a multiple linear regression model to predict the water quality that is the dependant variable based on several factor which act as the independent variable.

**Objective:**

1. How well our independent variables can predict the river water clarity? How much variance in river water clarity can be explained by our independent variables?
2. Which independent factor is highly influential in predicting the river water quality/clarity?

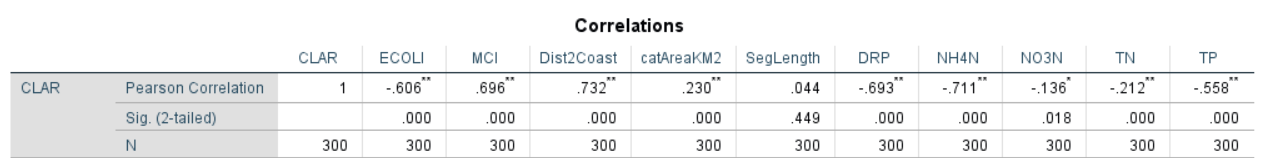
**Dataset:**

Source - <https://catalogue.data.govt.nz/dataset/predicted-river-water-quality-200913>

The variable included in the dataset are median value of visual clarity, median value of nitrogen content, median value of phosphorous content, median value of macroinvertebrate community index, distance of river segment to the coast, area of upstream catchment, coordinates of river segment, source, region and length of segment.

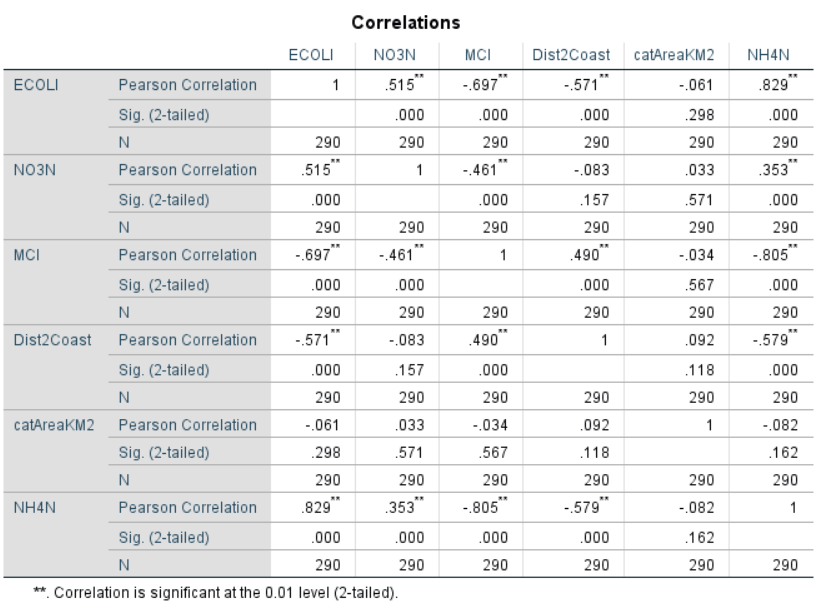
**Step 1: Checking the assumptions.**

**Multicollinearity**: Correlation coefficient is used to check the relation between dependent and independent variable with preferable value greater than 0.3 for Pearson Correlation.

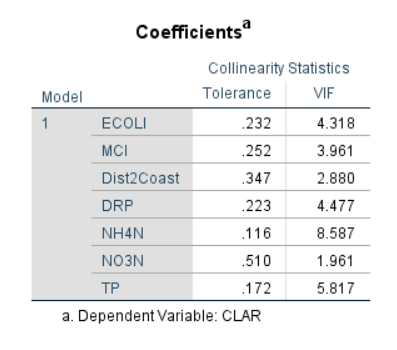


From the above table variables having value less than 0.3 have been eliminated namely catAreaKM2 (area of upstream catchment in m2), SegLength (length of river segment) NO3N (median value of oxidised nitrogen), TN (mean value of Total Nitrogen).

Correlation between two independent variables can be checked with the help of bivariate correlation. In a bivariate correlation, we must eliminate any variable which has a correlate value of 0.7 or more. From the below table we can see that NH4N ( median value of ammoniacal nitrogen) is more than the expected value, hence we have to eliminate this variable.



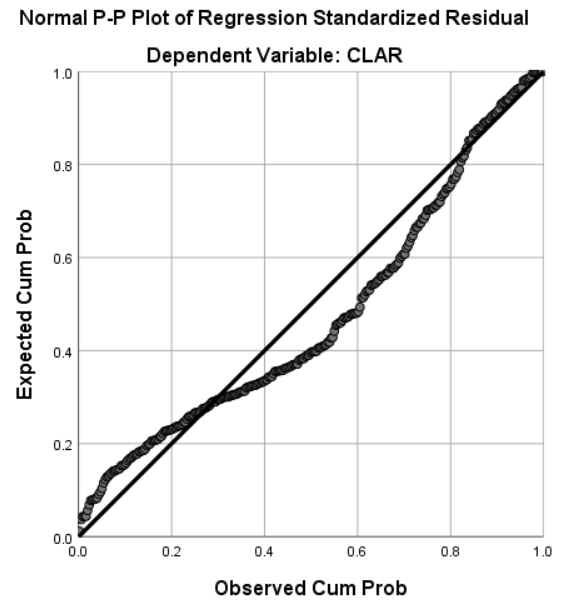
Next, we check the multicollinearity between the independent variables, for this we use the collinearity statistics. Two values are given for each independent variable Tolerance and VIF. Tolerance is an indicator of how much of the variability of one independent variable is not explained by other independent variable in the model and is calculated using the formula I-R2. VIF (Variance Inflation Factor) is the inverse of Tolerance Value. A tolerance value of 0.2 and above is preferably chosen for checking the collinearity between independent variables.



From the above table, independent variables NH4N (median value of ammoniacal nitrogen) and TP (median value of Total Phosphorous) are eliminated since the fall below the tolerance level of 0.2.

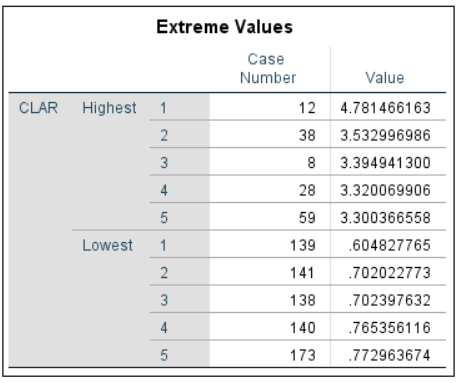
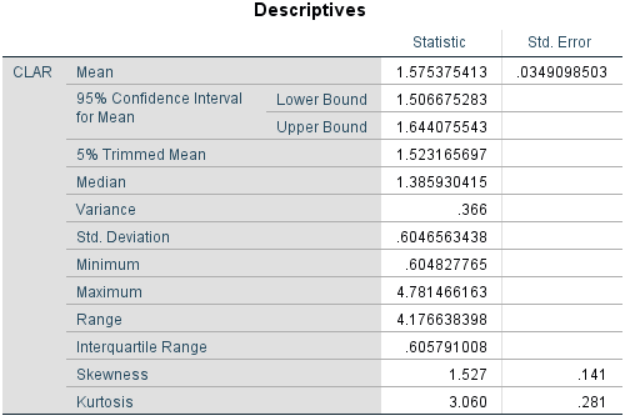
Finally, the selected independent variables include ECOLI (median value of Escherichia Coli), MCI (median value of macroinvertebrate community index), Dist2Coast (distance of river segment to coast), DRP (median value of dissolved reactive phosphorous) and NO3N (median value of oxidised nitrogen). All the independent variables are all continuous variables. The dependent variable is CLAR (median value of river clarity) which is also a continuous variable.

**Outliers, Linearity, Normality, Homoscedasticity, Independence of Residuals:** One of the ways that these assumptions can be checked is by inspecting the residuals scatterplot and normality probability plot of regression standardised residuals Pallant (2005). In Normal Probability plot we must check if the points lie along the diagonal line to see if there are no major diversions from normality.

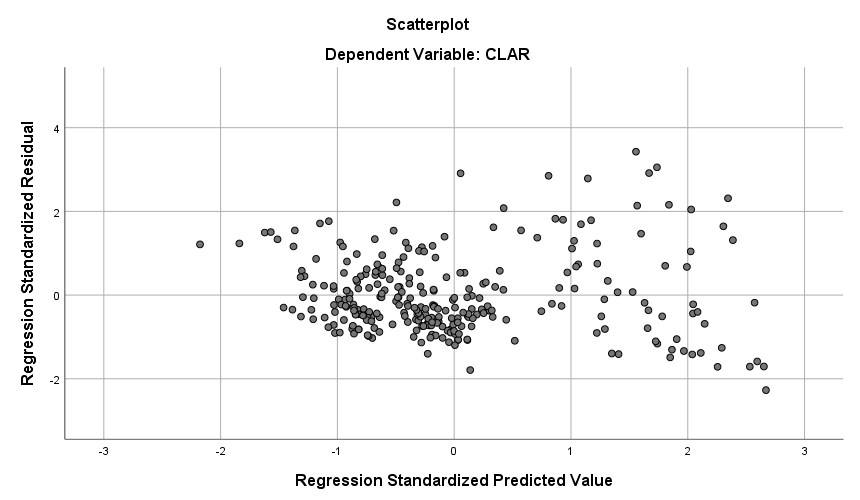


From the above figure we can see that points lie along the diagonal line suggesting normality in the dataset.

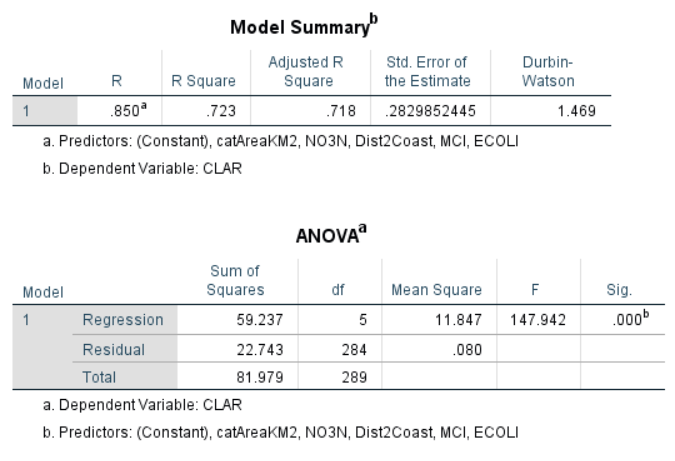
The presence of outliers can be found out by using descriptive analysis. The below table gives the ID of rows in the dataset which act as outliers having extreme mean values. These values have been removed from the dataset.

From the below scatterplot we can see that we have eliminated the outliers from the dataset.



**Step 2: Evaluating the Model.**

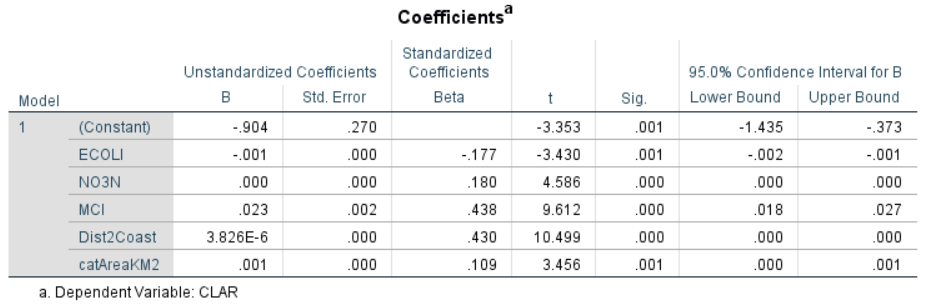
****

The variance of dependent variable in the data can be evaluated with the help of R Square value. The R Square value lies between 0 and 1, higher the value of R Square suggests that the model fits well with the data. By looking at the model summary we can see that the value of R Square is 0.723 which suggests that the variance of dependent variable is around 72.3% and the model fits the data well.

From the ANOVA table we can check for the statistical significance, the F value for our model is F (5,284) = 147.942. The Sig value of .000 means that there is very less probability that the F value of 147.942 is obtained in the dataset and the model is statistically significant.

**Step 3: Evaluating each of the Independent variable.**

In this step we evaluate which variable has highest effect on predicting the dependent variable. This information can be found in the coefficients table in the output box.



In this table, we interpret the value beta column under the Standardized coefficients. We make use of standardized coefficients because all the beta values will be standardized on a scale. From the table we can observe that both MCI (median value of macroinvertebrate community index) and Dist2Coast (distance of river segment to coast) have values of .438 and .430 respectively, suggesting that both variables have significant contribution towards predicting the dependent variable. The beta value of catAreaKM2 has the least value of .109 indicating it has the least significant influence on the predicting variable CLAR (median value of visual clarity/quality of river).

Next, we check for Sig value in the table. A value of .05 for Sig value indicates that the variable is making significant contribution towards predicting CLAR. From the Coefficients table we see that the Sig value for each of the variables is less the critical value of 0.05.

**Results:**

Based on the analysis above, we have come to the following conclusions

1. Our independent variables ECOLI, NO3N, MCI, Dist2Coast and catAreaKM2 explains for 72.3 percent of variance in our dependent variable CLAR.
2. We can see that from our independent variables, both MCI (median value of macroinvertebrate community index) and Dist2Coast (distance of river segment to coast) have high influence on predicting the dependent variable CLAR (median value of visual clarity/quality of river).

**References:**

1. Pallant (2005), Julian Pallant SPSS 6th Edition Survival Manual. New York, McGraw Hill Education.