**STATISTICS CA2.**

**Name : Atif Feroz Jeelani**

**Student’s ID : X17169992**

**Name Of College : NATIONAL COLLEGE OF IRELAND (NCI )**

**First piece of analysis : Multiple Regression**.

Multiple Regression : It is basically a set of techniques that can be used to explore the relationship between one continuous dependent variable and a number of independent variables. It is based on correlation. It explains how much variance in dependent variable (which can be continuous or interval/ratio) can be explained by independent variable (which can be either continuous or categorical). Different types of multiple regression depending on ones query are :

1. Standard or simultaneous .
2. Hierarchical or sequential.
3. Stepwise.

For my piece of analysis I will be using Standard multiple regression as it suits my nature of query.

**Objective of analysis :**

1. To analyse if there is a correlation between employers Total satisfaction level , a measure of the satisfaction expressed by employers who use their provider to train their employers to their low, high and medium organisation score as independent variable.
2. To analyse among which of the independent variables is the best predictor for total satisfaction.

**Data Source :**

As per the instruction the data set is taken from depository of official UK government statistics and databases : data.gov.uk

<https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/508138/FEPI_OpenData201415_FEChoices_Organisations_Provider_EmployerSatisfaction.csv/preview>

**Number and Levels of Measurement** : For my standard multiple regression, I am using interval/ratio levels of measurement for both dependent as well as independent variable i.e. both the variables are continuous . There are 3 number of independent variables ‘Orglower, Orgmiddle and Orghigher’ as predictor’s and one dependent variable final\_score as prediction variable.

**Null Hypothesis** : The independent variables defines variability on the dependent variable which is satisfaction level of employer

**Alt Hypothesis** : The independent variable has no affect on the satisfaction level of employer.

**PREMINARY TEST ASSUMPTIONS :**

1. **Small sample :** not recommended for small datasets where distribution is very skewed .
2. **Multicollinearity and singularity :** independent variables should not be highly correlated .Also one independent variable should not be combination of other .
3. **Normality:** It should follow a normal distribution.
4. **Homoscedasticity:** Should obey homogeneity of variances.
5. **Significance:** The significance level should be less than 0.05 i.e. p>0.05
6. **Outliers :** Multiple regression is very sensitive to outliers.
7. Dependent variable should be continuous where as independent can be categorical as well

**Procedure for analysis :**

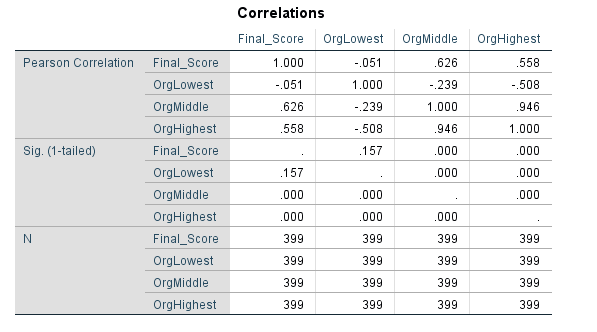
1. Click on analyze, then select regression, then linear.
2. In dependent box put your dependent variable that is to be predicted (Final\_score) and in independent box put your predictor variables (OrgLowest , OrgMiddle , OrgHigest ).
3. For Method make sure Enter is selected .(this will give us standard multiple regression)
4. From Statistics select Estimates ,Confidence Intervals ,Model fit, Descriptive , Part and partial correlations and collinearity diagnostics .
5. In plots put ZRESID into Y box and ZPRED is the X box and from Standardized Residual plots tick Normal probability plot. Now click OK.

**Interpretation of Output from Standard Multiple Regression :**

**Step 1 : Checking the assumptions**

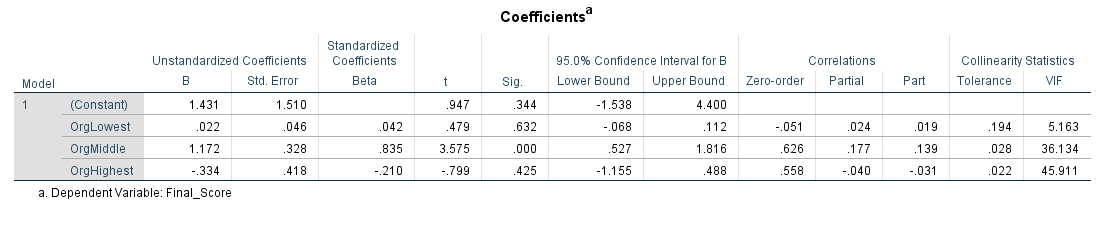
***Multicollinearity***

We need to check our independent samples show some relation with dependent variables .For this we check table Correlations.

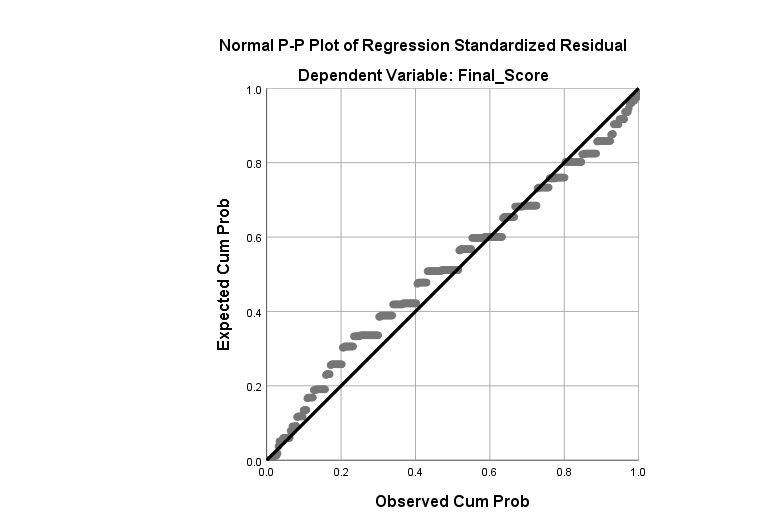


From this we conclude that OrgMiddle and OrgHighest correlate substantially with Final\_score where as OrgLowest does not . As both scores .626 and .558 is greater than 0.3 whereas - .051 < 0.3 . Also , among the 3 independent variables OrgMiddle and OrgHighest show bivariate correlation as their score 0.946 > 0.7 (highly correlated ) hence a violation.

Multicollinearity is more evident from Coefficients table.

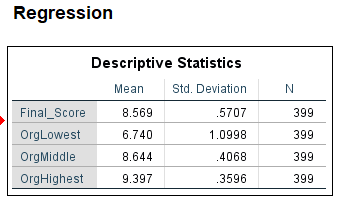
Tolerance is an indicator of amount of variability of the specified independent which is not explained by other independent variable calculated as 1-R square for every variable. Here clearly, 2 independent variables with tolerance value of .028 and .022 is less than .10 indicating multiple correlation. Also the VIF value of 2 variables are above 10 (36.13 and 45.911) ,hence showing multicollinearity .

***Normality and Outliers:***



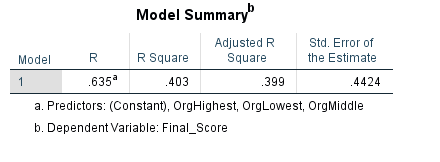
From normal Probability Plot (P-P) of the regression Standardized Residual we nearly get to see a reasonably straight diagonal line from bottom left to top right suggesting no major deviations from normality. Also depicts that model is linear with few outliers which are outside the linear line.

**Step 2 : Evaluating the Model.**

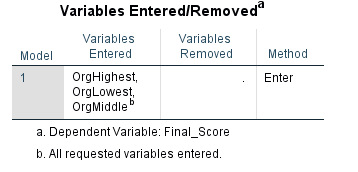


Here, Final\_score is my dependent variable or prediction which is continuous while as OrgLowest , OrgMiddle and OrgHighest are my predictors (independent variable) which are in scale measure.

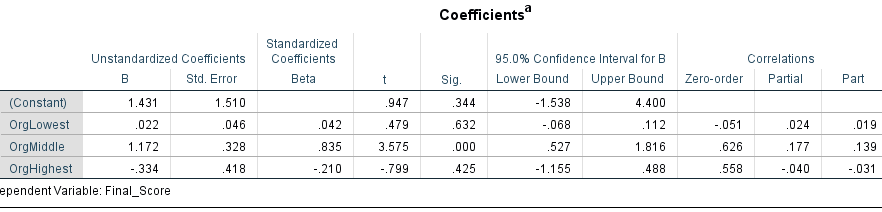
**Descriptive Statistics** : In this table we get to know the mean of all the variables and also their standard deviation from the mean. The greater the data points vary from each other the greater the standard deviation will be. Similarly the standard deviation would be low if great significance between data points mean.



**Model Summary :** The R square value gives the amount of variance in the dependent variable explained by the model (which includes all independent variables). Here clearly R square value is (0.403) which means 40.3% of variance in dependent variable (Final\_score) which is a quite acceptable result. Value (0.635) indicates that the dependent (satisfaction score) is above average. The Adjusted R square value is considered if the sample would have been small ,it helps in countering this effect (.403) to (.399)



All the independent variables were entered simultaneously using the Forced entry method. As done in the case for Standard multiple regression



**Coefficients :** In the standardized coefficients Beta (.835) which is OrgMiddle value is the largest of all which means it makes the most strongest unique contribution in explaining the variance of the satisfaction score. Followed by (-.210) and (.042) indicating the other two variable made less contribution. Also (.835) has Sig. of (.000), meaning it makes a statistically significant unique contribution while the other two are not .

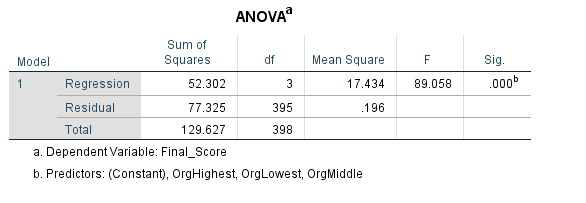
Also OrgMiddle has part correlation of (0.14 ) i.e. it explains satisfaction score of 1.96 % uniquely . Followed by OrgHighest and OrgLowest which explains .096 % and .0361% of dependent variable respectively.

From , this we conclude :

1. With unit change in OrgMiddle there will be increase in satisfaction level.
2. With unit change in OrgLowest there will be decrease in satisfaction level.
3. With increase in unit change of OrgHighest there will be decrease in satisfaction level.

To check the significance we go through Sig.

1. Sig. OrgLow = 0.632 > 0.05 which is not significant
2. Sig. OrgMid = 0.000 < 0.05 which is significant
3. Sig. OrgHigh = 0.425 > 0.05 which is not significant

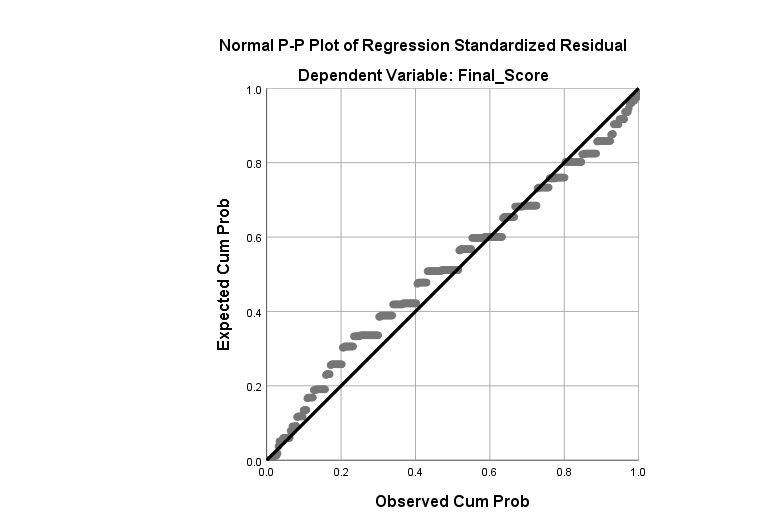


**ANOVA :** the table depicts whether the model is a good fit or not .

F-ratio at (3,395) = 89.058 . The model reaches statistical significance as Sig. = .000 ; this means p < 0.0005

From this we obtain that independent variables are significant in obtaining the total satisfaction (which is our dependent variable).

**Charts :**



The plot illustrates that there are few outliers in the model also that the model is linear. Since nearly all the points follow the straight line therefore normality assumption is also met .

**Conclusion :** The standard multiple regression was used to analyse if there exists a correlation between employers Total satisfaction score , a measure of the satisfaction expressed by employers who use their provider to train their employers to their low, high and medium organisation score as independent variable. Preliminary analyses were conducted to ensure no violation of the assumptions of normality, linearity and multicollinearity . Model followed a normal distribution . However, it violated the assumption of multicollinearity. The analysis concluded that the independent variable do show correlation to the dependent variable. From three independent predictors only OrgMiddle class is explaining variability to the prediction of final score (Satisfaction level) with p-value < .0005. The other two variables are not significant in predicting the Satisfaction level.

**2 .Second piece of analysis : Logistic Regression**.

**Logistic regression** : It is used instead of multiple regression when the dependent variable is of nature dichotomous i.e it is categorical variable. The dependent variable can be in the form of 0 and 1 . Dependent variable has to be binary . It is used to test the predictive power of a set of variables also to check the contribution of each independent variable distinctly.

**Objective of Analysis** : The objective of my logistic analysis is to check the variance caused by country and gender which are predictor variables on the adults who have quit smoking and on adults who still smoke , which here is my dependent variable .

**Data Source** : As per the instruction the data set is taken from depository of official UK government statistics and databases : data.gov.uk

<https://catalog.data.gov/dataset/behavioral-risk-factor-data-tobacco-use-2011-to-present-e0ad1/resource/d2851494-0d6b-4bbb-b7a5-62b7c091c750?inner_span=True>

**Number and Levels of Measurement :** For my logistic regression I am using 2 independent variables and one categorical variable. Here location desk represents country (New York and Washinton) and gender represents male female which are to be taken as predictors and topicDesc which represents people who smoke and stopped smoking as dependent variable. Here adults who smoke was taken as 1 and non-smoker adults as 0 .

**Null Hypothesis** : The Null hypothesis states that the country and gender does affect the variability of adult people who smoke and quit smoking.

**Alt Hypothesis** : The alternate hypothesis states that the predictor variables has no affect on the variability of the adult people who smoke and quit smoking.

**PREMINARY TEST ASSUMPTIONS :**

1. Small Sample size : If we have small sample with large number of predictors it is a problem. Not recommended for small data samples where distribution is skewed .
2. Multicollinearity : The predictor variable should be strongly related to dependent variable but not strongly related to each other . Here in this case there is no violation
3. No outliers
4. Independence of Errors
5. Dependent variable should be categorical where as independent can be continuous as well . Here in this case the dependent variable is categorical

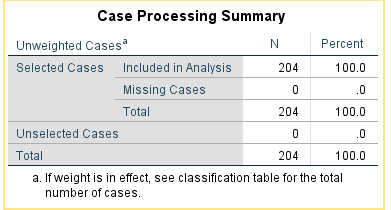
**Procedure for analysis :**

1. Click on analyze, then select regression, then binary logistic
2. Choose your dependent and independent variables. In dependent box put your dependent variable that is to be predicted (topicDesc) and in independent box put your predictor variables (country and gender ).
3. Click on options .Select Classification plots, Hosmer-Lemeshow ,goodness-of-fit ,Casewise listings of residuals and CI for Exp
4. Click continue.

**Interpretation of Output from Standard Multiple Regression :**

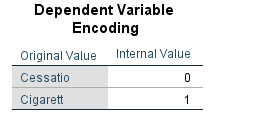
* The topicDesc is the dependent variable which is to be predicted which is a categorical variable .
* LocationDesc and Gender are our predictor variable .

The independent variable were added simultaneously while forced entry method is used to enter the predictor variable.



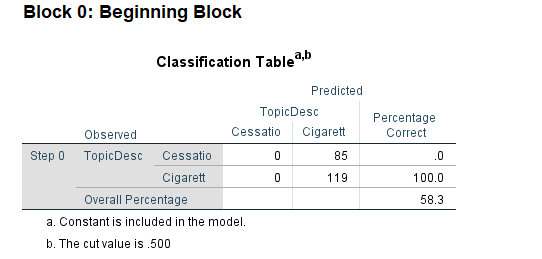
**Case Process Summary :**

The above table is the case processing summary . N is the number of cases denoted in the model. Also confirms that we have enough case for our analysis. Here missing values is 0 and total cases is 204.



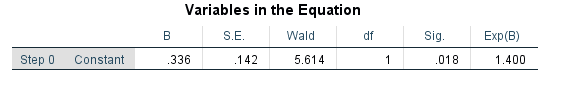
**Dependent Variable Encoding :**

To analyse the variability caused due to independent variable the dependent value needs to be encoded to binary . Here cessation is coded to 0 (quit smoking) and cigarette is coded to 1 (smokers) to execute the process.



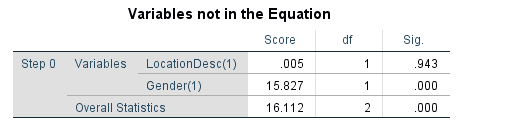
**Classification Table (Block 0) :**

It gives us the result of variance without the affect of the independent variable . Here overall percentage of correctly classified cases is 58.3 %. This means the data is significant for analysis to be performed .

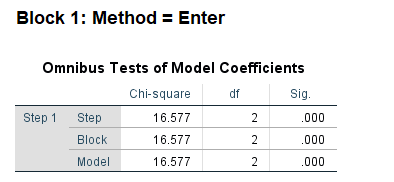


**Variables in the Equation :**

Here the value of exp(B) is obtained by dividing 85 by 119 i. e 1.400 . This means exponential of (B) . These values are odd ratios. It represents the change in odds of being in one of the categories of outcome when predictor increases by one unit .

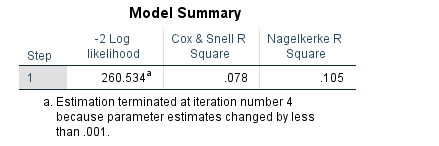


Here the overall statistics which is 16.112 is the value denoting all the forecasts inside the model.



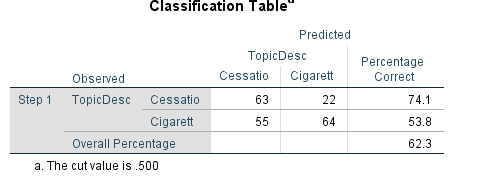
**Omnibus Test of Model Coefficients :**

The omnibus tests of model coefficients gives us an overall prediction of how well the model performed over and above Block 0. Here we want a highly significant level . The chi-square value is 16.577 at 2 degrees of freedom and with Sig. value .000 (p <.00005) which is less than p -value which means model is fit for logistic regression . Thus we fail to reject Null hypothesis .



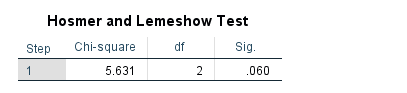
**Model Summary :**

Model Summary gives us the information about the usefulness of the model. The value obtained from Cox&Snell R Square and Nagelkerke R square explains the variability caused in the dependent variable by the model. Here the values are .078 and .105 which means 78 percent and 10.5 percent variability is explained in the dependent variable (adults who quit smoking and adults who smoke ) from the above set of values. Also the -2 log Likelihood can be observed has a large value which indicates that the model chosen has goodness of fit .



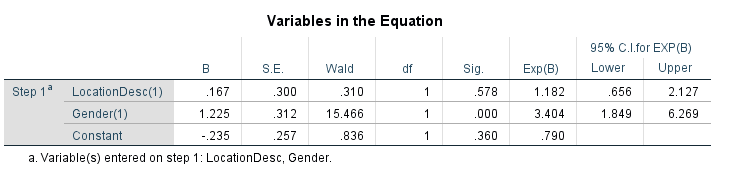
**Classification Table :**

Classification table gives an indication, how well the model predicts the correct category for distinct cases. We compare it with block 0 classification table to check the amount of variation when predictor variables are included in the analysis. Here the overall percentage is observed as 62.3 percent which is an improvement to the Block 0 overall percentage of 58.3 percent . This table suggests that specificity is 74.1 % adults quit smoking while we were able to correctly classify 53.8 % adults smoke .



**The Hosmer-Lemeshow Test:**

The Hosmer-Lemeshow Test is the most reliable test of model fit. The significance value less than .05 is a poor test . Here Sig. value is .060 , which is greater than p-value of .05 which supports the model. Hence, model is a good fit .



**Variables in the equation :**

Variables in the equation table gives us information related to the individual contribution of our predictor variable . The obtained values of Wald with respect to the Significance values needs to be checked if the Significance value is less than p-value of 0.05 than the variables contribute significantly .Here in this case Gender with Sig. value of .000 has shown significant variance to the adults who smoke and adults who quit smoking as compared to LocationDesc which does not show significant contribution to the dependent variable. Also here the B values show the positive and negative direction of the relationship . In this case, its positive.

**Conclusion** :

Logistic regression was performed to analyse the variance caused by country and gender on the dependent categorical variable which was cessation and adults smoking decoded as 0 and 1 respectively .Model 1 containing all predictors was noticed to be statistically significant , X ² (2, N = 204)  = 16.57, p< .001, depicting that the model was able to predict the difference. The model explained between as 78% Cox and Snell R Square and 10.5% as Nagelkerke R squared variance and correctly classified 62.3% of the cases . As illustrated only Gender predictor made a statistically significantly contribution to the model with p-value < 0.05.

Adjusted R square tells us that …% variance of dependent variable is explained by the set of independent variable.