

***Child Wellbeing Index - SPSS***

***Mohammed Abou Hassan***

ALL Data have been downloaded from Data.Gov.uk:

<https://data.gov.uk/dataset/road-safety>

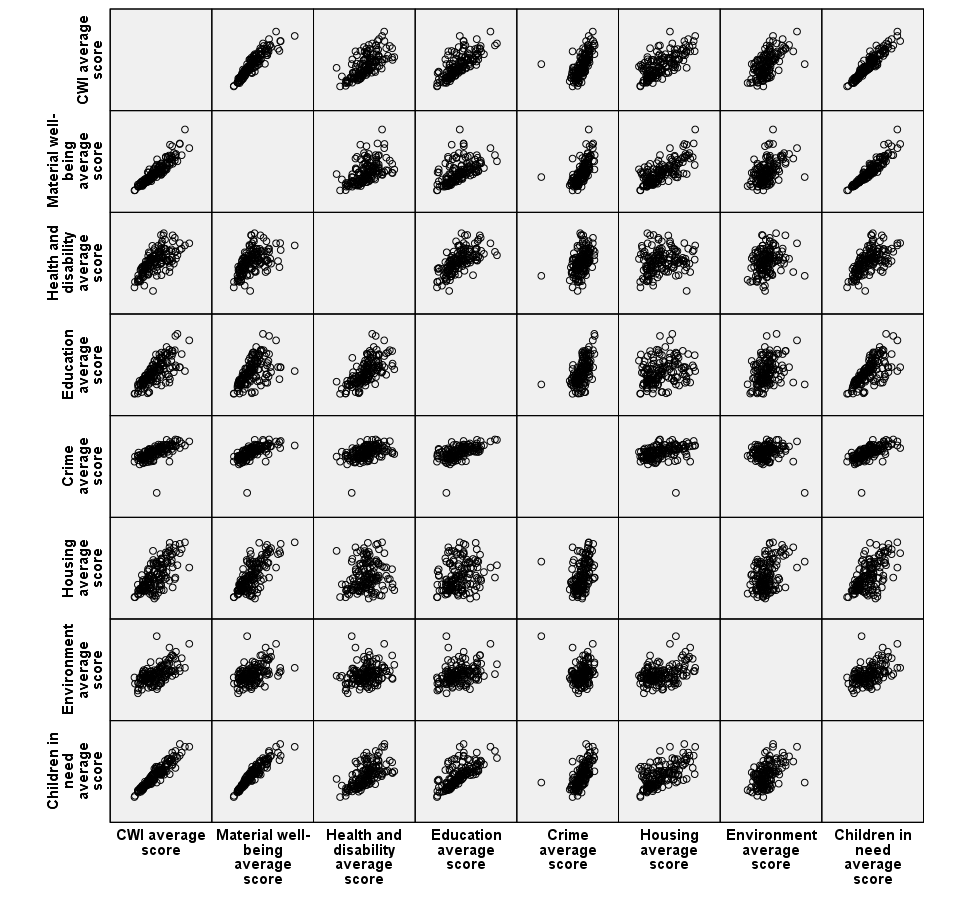
<https://data.gov.uk/dataset/bee-and-flower-abundance-and-diversity-and-bee-pollen-foraging-data-from-farms-in-england>

<https://data.gov.uk/dataset/child_well-being_index_cwi_2009/resource/b9804c8e-9ee1-40e8-ae57-9ba1868a4ff6>

**Multi Regression model:**

For this analysis, a data set called CWI – child wellbeing index, is going to be our source of data, where we are going to study the effect of different factors on a child’s wellbeing.

The factors or independent variables (Material wellbeing, Education average score, Crime average score, Housing average score, Environment average score, Children in need average score, Health and disability average score) will be tested to see if they have any correlation with the dependent variable CWI.



We start our study of our multi-regression model by presenting the correlation between all the independent variables and one dependent which is the CWI score.

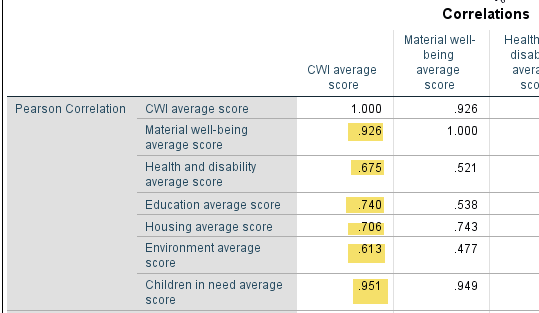
It is clear from the matrix above that there is a good correlation between CWI and most of the independent variables except Crime average.

Procedure:

1. From the menu -> Analyze -> Regression -> Linear
2. Choose CWI score as dependent variable and have the rest of the variable as independent
3. Method: Choose Enter and that will ask SPSS to give you standard multiple regression
4. Tick the box marked Estimates, Confidence Intervals, Model fit, Descriptives, Part and partial correlations and Collinearity diagnostics

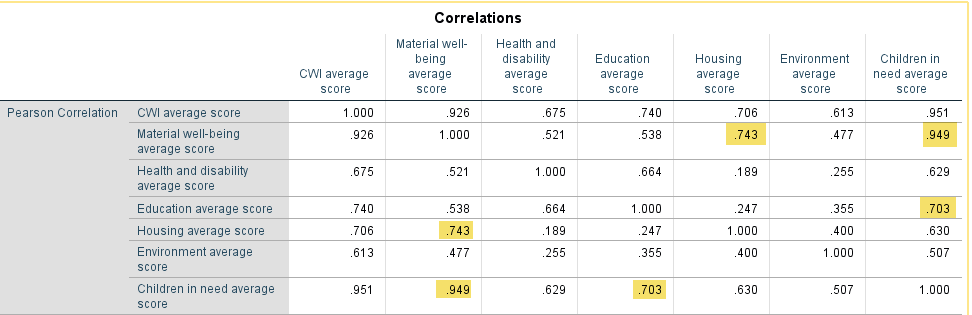
Step 1: Checking the assumptions

**Multicollinearity:**



The correlations between the variables in my model are provided in the table labelled Correlations. We must check if our independent variable shows at least some relationship with our dependent variables. We can see from the numbers highlighted above (.926, .675, .740, .706, .613, .951) that the dependent variable is correlated to each of the independent ones.

Also, we need to check if the intercorrelation between each of the dependent variables is high. In general, when two independent variables are strongly intercorrelated (Correlation > 0.7), we either must omit one (or more) of the correlated variables or combine the two of them to form one variable.

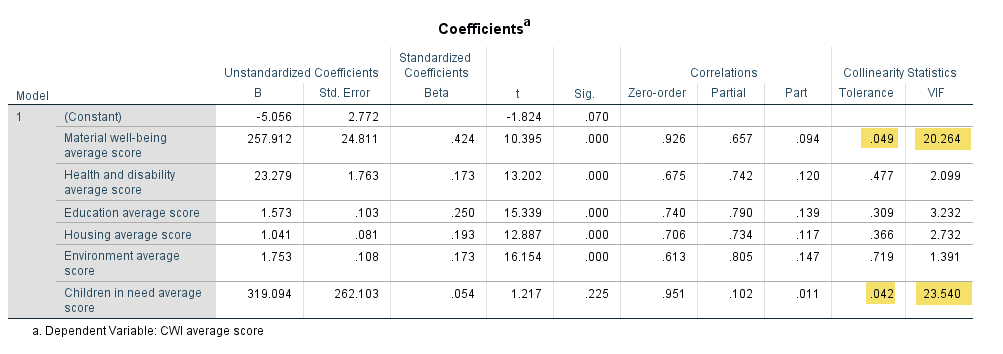
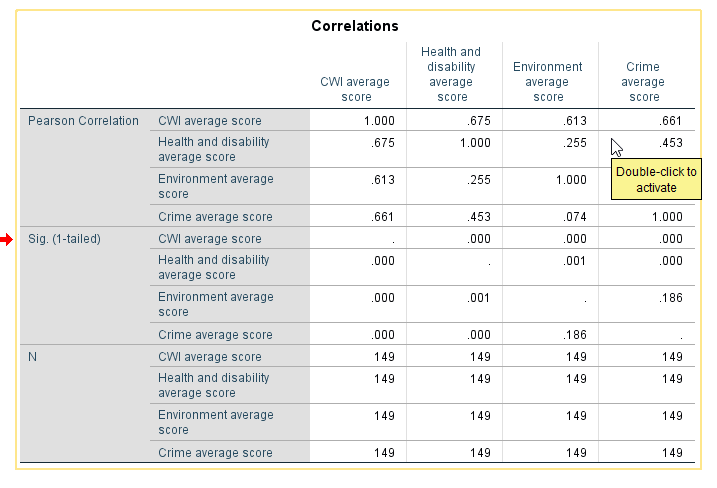


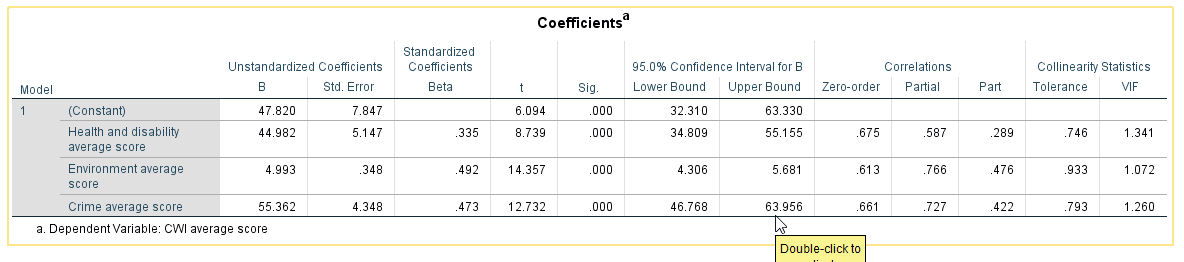
The values highlighted in the Correlations table above (.743, .949, .703) indicate there is strong correlation between the following independent variables:

Materials-wellbeing average < -- > Housing average score (.743)

Materials-wellbeing average < -- > Children in need average score (.949)

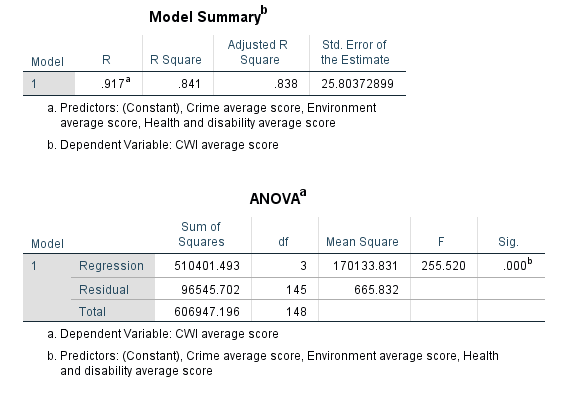
Children in need average score <-- > Education average score (.703). Also, when having a look at the Coefficients table below, we can see that the very high values of VIF highlighted (should be less than 10) confirm the multicollinearity concerns.

I omitted the variables with high multicollinearity and retested to find the best suitable combination of variables with suitable values of VIF. The new Correlation table is shown below:



VIF in the table above is lower than 10 and tolerance is lower than 1 which means multicollinearity is no longer a reason for concern anymore.

After having a model with acceptable levels of multicollinearity between variables, it is time to check other tables that analyse our model furthermore.

In the model summary table below, we can see R squared is 0.841 which means 84.1% of the variability in the outcome can be explained by our choice our predictors and our model. It is worth mentioning that the R square and adjusted R square have very close values (0.841 and 0.838) which is a pointer that our model can generalize in a good way. 

Also looking at the Anova table that test the null hypothesis we can see Sig = 0.000 which means the model reached statistical significance.

Evaluating each of the independent variables: To see which of the independent variables included in the model contributed to the prediction of the dependent variable we are going to check the table above labelled Coefficients. Checking Beta Standardized coefficients will reveal that Environment average score has the biggest contribution to explaining the dependent variable (it is statistically significant too Sign 0.000).

**ANOVA:** One-way between groups analysis of variance was conducted to explore the impact of treatment (4 different levels are “Meja+ (different than Meja+) , Meja+, Cut+, Control) on percentage senesced which is our dependent variable.

Changes in agricultural practice across Europe and North America have been associated

with range contractions and a decline in the abundance of wild bees

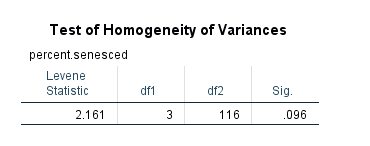
Interpretation of output from one-way between groups ANOVA with post-hoc tests:

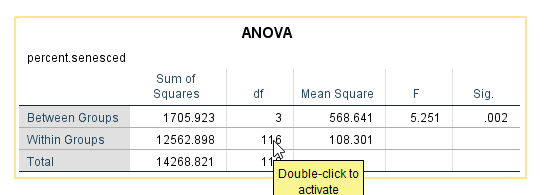
Test of homogeneity of variances: The homogeneity of variance option gives us Lavene’s test, which tests whether the variance in “percentage senesced” is the same for each of the groups tested (Treatment levels).

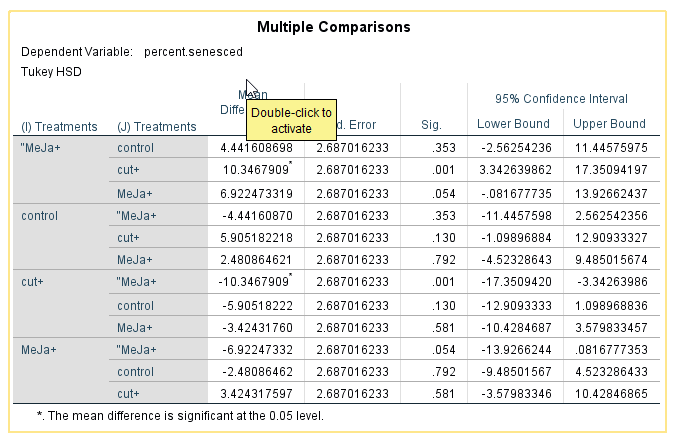
Treatment was not accepted by SPSS as a factor, so I had to use the Automatic recoding suggested by IBM because SPSS only sees the factors if they were numeric.

The significance value for Levene’s test is 0.096 which is greater than 0.05 and then we have not violated the assumption of homogeneity of variance.

ANOVA: The table here gives the between -groups and within groups sums of squares in addition to degrees of freedom.



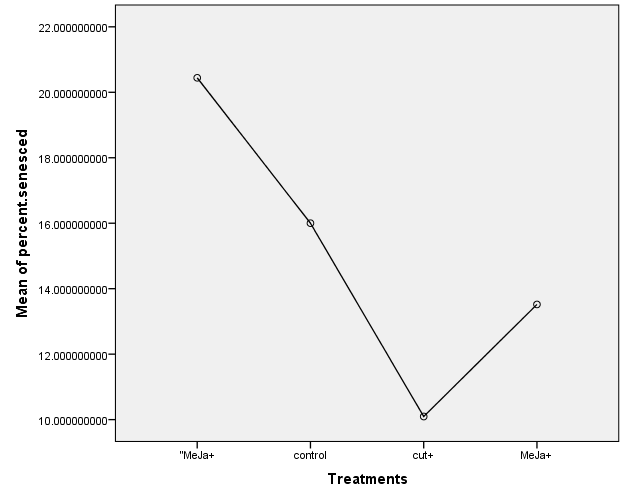


We are interested in the column labelled Sig. which is equal 0.002 < 0.05 in our case which means there is a significant difference somewhere among the mean scores on the Treatment variable over its groups. We check the Multiple Comparison table below to see the differences between each pair of the groups. 

The asterisks in this table mean the two groups compared are significantly different from each other which are Meja+ and Cut+ with significance value Sig. = 0.001.

Only these two groups are statistically significantly different from one another in terms of percentage senesced.

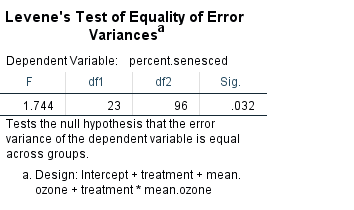
Mean plots:



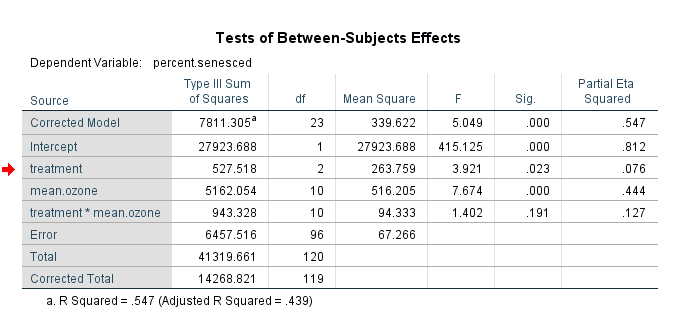
By looking at this graph we can see the differences in means between the different groups and we can confirm the Cut+ and “Meja+ have the biggest difference in mean of percentage.

Two-way Anova: Our aim here is to introduce a second independent variable and to study the effect of treatment and mean ozone on percentage senesced.

Levene’s Test gave us Sig 0.032 which is a significant value which means the variance of our dependent variable (percentage senesced) across the groups is not equal.



So I changed the significance level to 0.01 to make it more stringent. Looking at the table labelled Tests of Between-subjects effects, the column that has treatment\*meanOzone shows no significant interaction between the two independent variables (Sig = 0.191) hence we can conclude there is no significant difference in the effect of treatment on percentage senesced for a particular mean ozone.

Main effects: From the table below we can safely interpret the main effects because there was no significant interactions effect. We can clearly see that no significantmain effect for mean ozone. Only treatment with Sig. 0.023 has significant main effect. 

Logistic Regression:

Samples were taken from an English road safety study to analyse the factors affecting the severity of an injury in car accidents. The study is published on data.gov.uk in the form of three separate Excel files each of them highlighting a different area of the accidents (casualties, vehicles, area …). Two of these files were combined in R to using the accident ID to produce one file that has all predictors we need for our analysis. The severity of the crash was recoded in SPSS using 1 (severe) and 0 (not severe), creating a new dummy variable called dummy\_severity.

The independent variables used for this analysis are Age of driver, Weather conditions, light conditions, speed limit and number of vehicles. The dependent variable is the severity of the accident and it has two categories which are ”severe” and ”not severe”

The reason behind choosing one dummy variable is because we followed the rule that says number of dummies is k-1 (k = number of categories).

Few things we need to watch out for when doing this analysis like sample size, multicollinearity and outliers.

Options: we need to make sure the “No scientific notation for small numbers in tables” box has a tick next to it.

To start our analysis, we checked the number of cases mentioned in the Case Processing Summary table and we could see that numbers matched what we expected.

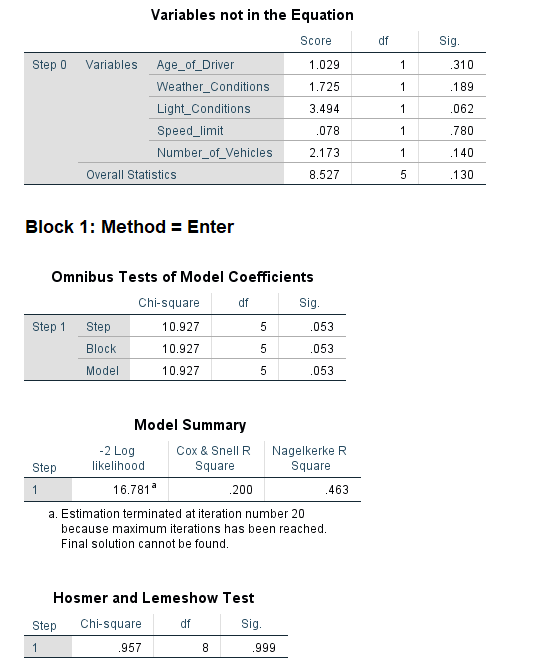
The next table we checked was the Dependent Variable Encoding table. We could see that SPSS dealt with the DV using 0 and 1 which was expected.

Block 0: This block has the results of the analysis without any of our independent variables used in the model. This serves like a baseline for comparing the model with our predictor variables included later on.

The percentage found in the table below (is 91.8%) means SPSS guessed all accidents have this percentage to be non-severe.

Block 1: This is where our model is tested.

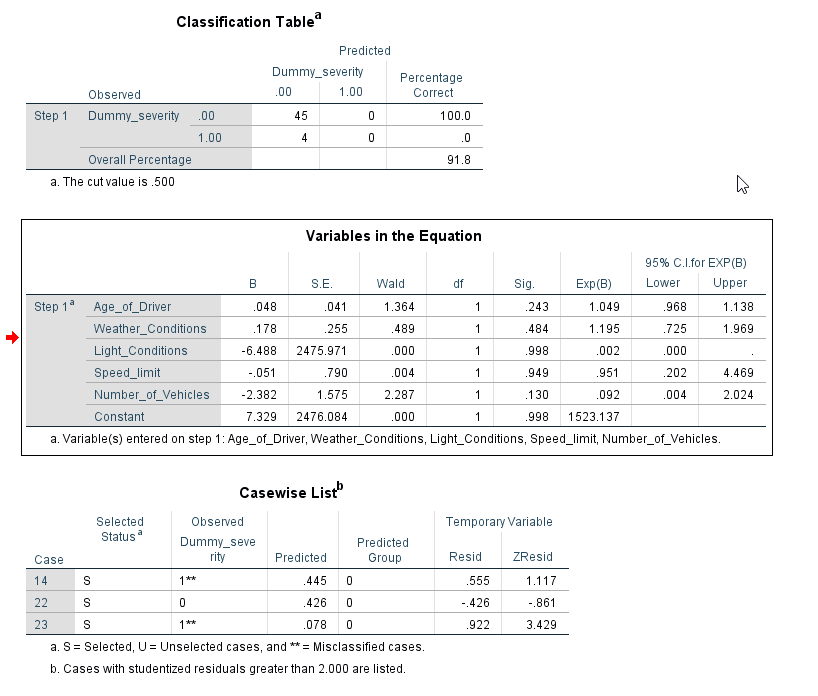
The Omnibus Tests of Model Coefficients gives us an indication of how well the model is doing over and above the results obtained for Block 0 with none of the predictors in the model. It is referred to as “goodness of fit” test. For our set of results Sig. is 0.053 which is considered as an acceptable answer (not ideal as the Sig. should be < 0.05) and therefore our model (including all predictors) is better than SPSS’s original guess shown in block 0.



Osmer-Lemeshow (called the most reliable test of model vaialable in SPSS) supports our model. In this test we found Chi-Square = 0.957, df =8 with significance of 0.999 ( > 0.05)

which means this test supports our model.

This finding is also supported by Cox & Snell R Square – Nagelkerke R Square gave a result of 0.2000 and 0.463 which suggests between 20% to 46.3% of variability is explained by our model and our set of variables used.



Wald test shown in the “Variables in the Equation” table, gives a good idea about the predictor that is contributing significantly to our model. By checking the Sign. Column it is we cannot find any predictor with p value less than 0.005 so none of them contribute significantly to the model. Exp(B) column shows that the odds for a severe injury is 1.19 times more for one category in weather condition than for another category, all other factors being equal.

In the classification table the PAC of 75.1 (Percentage accuracy in classification) indicates that the model correctly classified 91.8% of cases overall.

