**National College of Ireland  
Project Submission Sheet – 2019/2020**

**School of Computing**

|  |  |  |  |
| --- | --- | --- | --- |
| **Student Name:** | SANKARA SUBRAMANIAN V | | |
| **Student ID:** | X18179541 | | |
| **Programme:** | MSc Data Analytics – Group B | **Year:** | 2019-2020 |
| **Module:** | Statistics for Data Analytics | | |
| **Lecturer:** | Tony Delaney | | |
| **Submission Due Date:** | 25/11/2019 | | |
| **Project Title:** | A Study on Analysis of Multiple and Binary Logistic Regression | | |
| **Word Count:** | 2607 | | |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

|  |  |
| --- | --- |
| **Signature:** | Sankara Subramanian V |
| **Date:** | 25/11/2019 |

**PLEASE READ THE FOLLOWING INSTRUCTIONS:**

1. Please attach a completed copy of this sheet to each project (including multiple copies).

2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.

|  |  |
| --- | --- |
| **Office Use Only** | |
| Signature: |  |
| Date: |  |
| Penalty Applied (if applicable): |  |

1. **Objectives of the analysis:**

The main objective of this section is to perform multiple regression and analyze the influence of various predictor variables to predict a dependent variable of continuous value. The predictor or independent variables can either be continuous or categorical values.

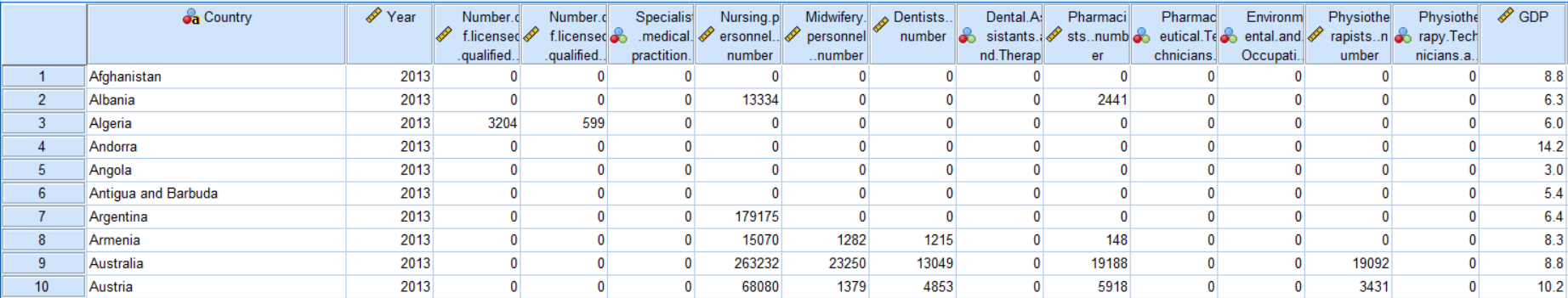
**Context of the data:**

The dataset is sourced from (WHO, 2019). In this dataset, multiple regression is applied to predict the percentage of gross domestic product (**GDP**) for the financial year 2013. The dataset contains data from 194 countries. The following are the independent variables:

1. Number.of.licensed.qualified.surgeons.actively.working
2. Number.of.licensed.qualified.obstetricians.actively.working
3. Specialist.medical.practitioners..number.
4. Nursing.personnel..number.
5. Midwifery.personnel..number.
6. Dentists..number.
7. Dental.Assistants.and.Therapists..number.
8. Pharmacists..number.
9. Pharmaceutical.Technicians.and.Assistants..number.
10. Environmental.and.Occupational.Health.and.Hygiene. Professionals...number.
11. Physiotherapists..number.
12. Physiotherapy.Technicians.and.Assistants..number.

The dataset contains 12 predictor variables and all the values are continuous. Multiple regression will help us predict the best predictors from the regression analysis using SPSS, **Analyze** **→** **Regression →** **Linear.**

**Dataset Sample View:**



Generally, datasets can be analyzed in 5 different methods namely, Enter, Stepwise, Forward, Backward and Remove. In this analysis, the Enter method is analyzed with casewise diagnostics, and the results are explained in the following sections.

**Assumptions of the Regression Model:**

1. The model is linear
2. The error terms have constant variances (Homoscedastic).
3. The error terms are independent
4. The error terms are normally distributed (Carlos & Ken, 2012).
5. **Data Cleaning/Transformation of the dataset:**

* As discussed, the dataset is prepared by combining 8 different CSV files based on the year and the country's attributes. The datasets have data from the year 1952 to 2018 and contain 3655 records.
* The dependent variable is joined into the dataset by using a **melt** function from the library (data.table) of R programming. The dataset which has extra spaces between the numbers for certain predictor variables is fixed using **stri\_replace\_all\_fixed** function from the library(stringi).
* All the null and no data values are replaced by 0's and the data type is cast to an integer type. Finally, the datasets are merged and written in a CSV file by filtering the data for the financial year 2013.

1. **Results of Diagnosis/Tests of Multiple Regression:**
2. **Sample Size:**

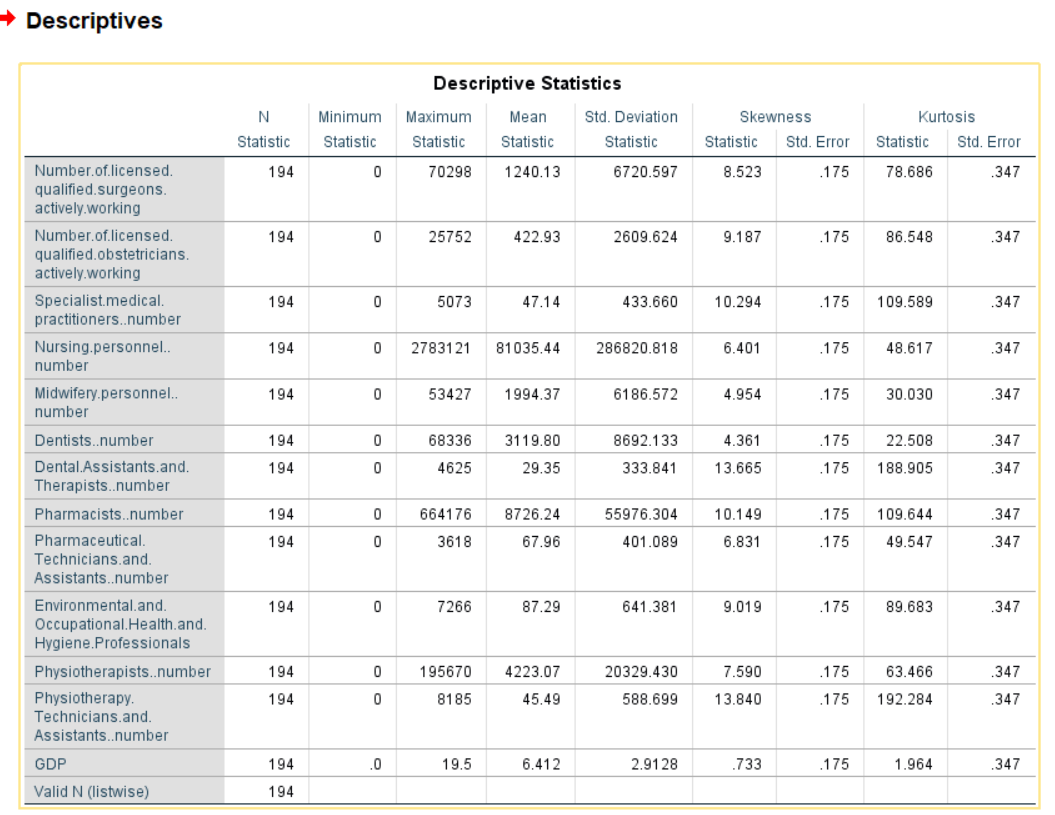
According to Fidell (Fidell, 2013), the minimum sample size requirement should satisfy the condition N> 50+8m.

m – the number of predictor variables.

50+8(12) = 146

In this dataset **N = 194> 146**, which is satisfactory.

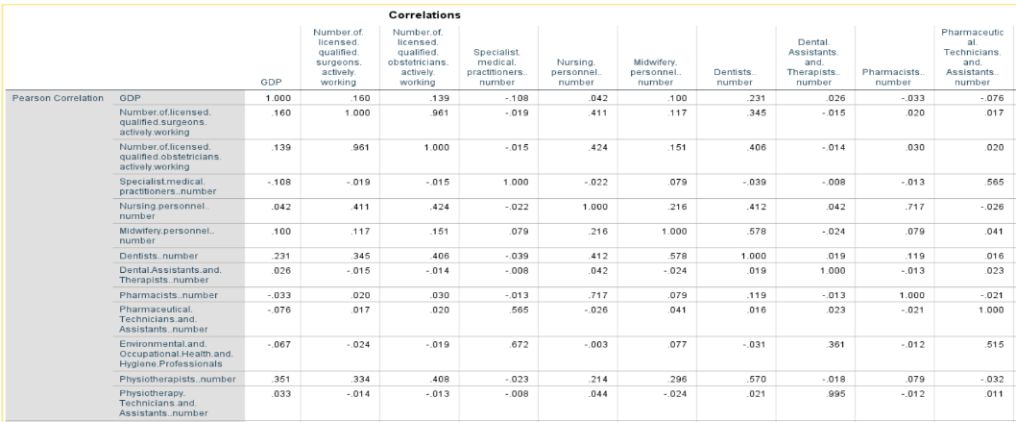
1. **Descriptive Statistics:**

Descriptive Statistics provides information on continuous variables (skewness and kurtosis). Descriptive Statistics also explains the distribution of scores. The symmetry of the distribution is explained by skewness and the peakedness is provided by kurtosis. A perfect normal distribution has skewness and kurtosis value of 0, which is rare. In this case, positive skewness indicates the scores are clustered to the left. Positive kurtosis explains the distribution is clustered at the center (Pallant, 2005).

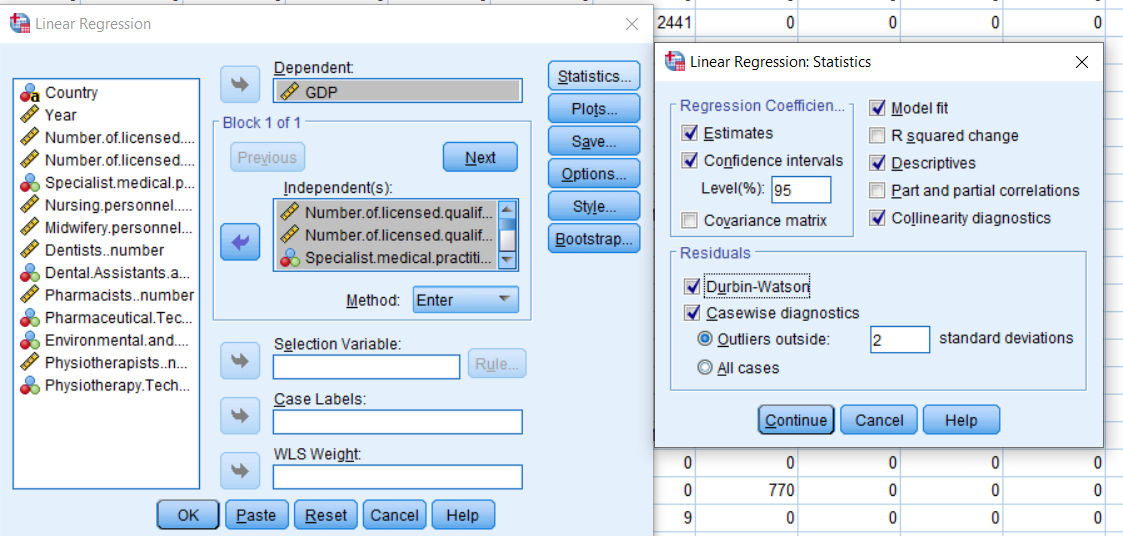
1. **Linearity Test:**

(The column names are boldened.)

* The attributes such as **Physiotherapists.Number**., **Dentists.Number**, **Number.Of.licensed. qualified.surgeons.actively.working** and **Number.Of.licensed.qualified.obstetricians.actively.working** have a better correlation with dependent variable **GDP**.
* **Pharmaceutical.Technicians.and.Assistants.number** and **Environmental.and.Occupational.Health.and.Hygiene.Professionals**. have negative coefficients, which shows weak correlation.
* The correlation between **Numbers.Of.licensed.qualified.surgeons.actively.working** and **Number.Of.licensed.qualified.obstetricians.actively.working** is high, and the value is >0.7. So, we should reject one of the attributes.

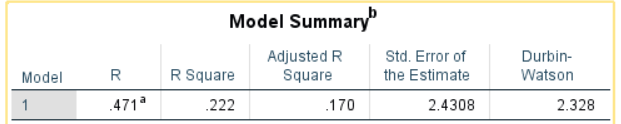
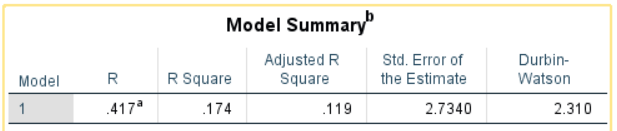


1. **SPSS Inputs and Outputs:**

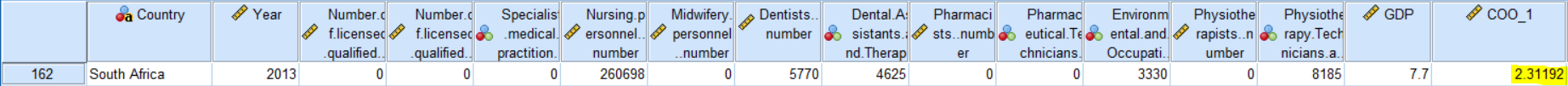


The above figure depicts the dependent and independent variables chosen for regression. In casewise diagnostics, outliers outside 3 standard deviations are neglected and the regression is rerun with 190 records.

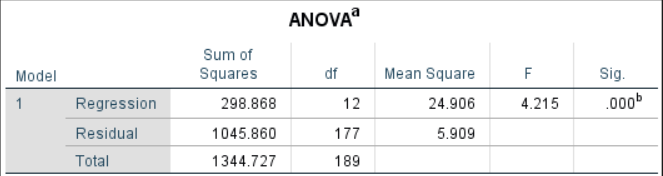
**Before Removing 3SD outliers After Removing Outliers**

****

**Outlier Check based on Cook’s distance:**

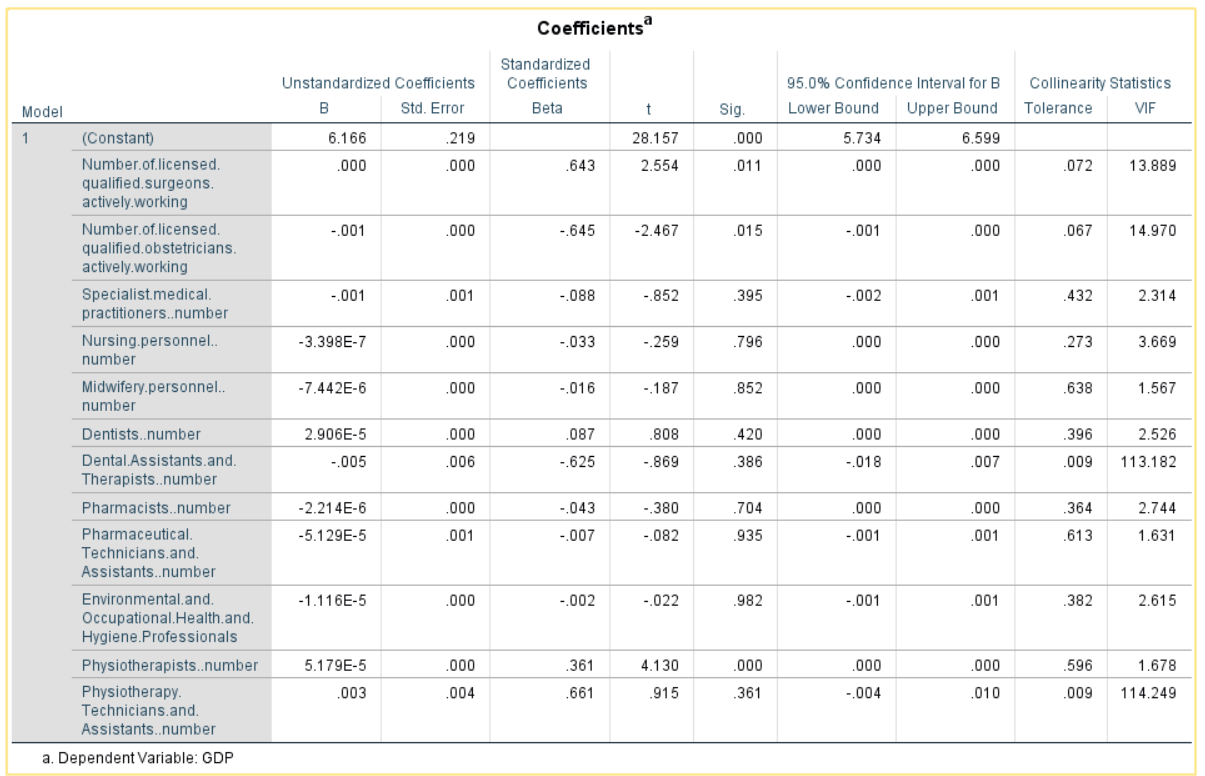


* The model summary explains the improvement of R-square value from **17.4 %** to **22.2 %** after removing **3 outliers** which are beyond 3 standard deviations from the mean.
* Based on the cook’s distance value out of 194 records, only **1 record** has a cook’s distance greater than 1. Filtering the cook’s distance value along with the other 3 records from casewise diagnostics we attained an acceptable Durbin-Watson value of **2.328**, which usually ranges from 0 to 4.
* A value of 2 explains the **non-autocorrelation**. The adjusted R-square is the better estimate of the true population. it provides the additional information if a new predictor variable is added, and its degree of freedom is increased.



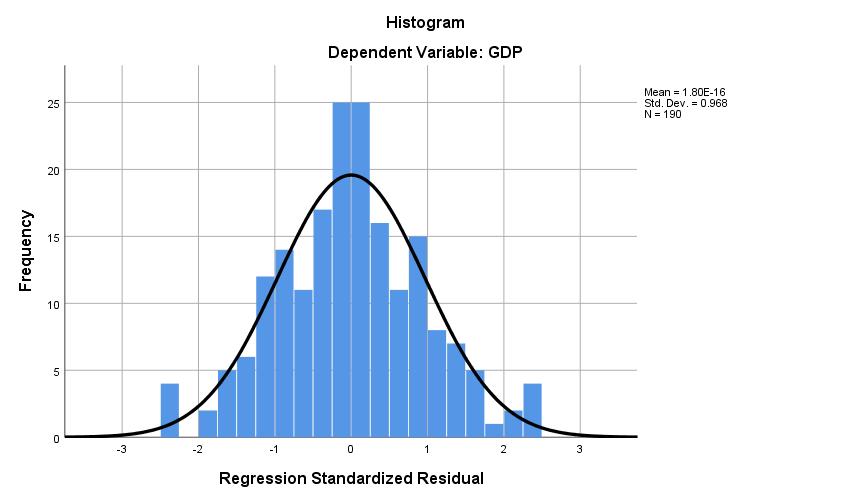
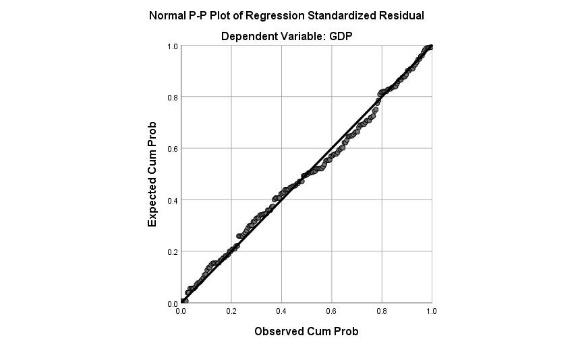
* The ANOVA table gives the information of Sum of Square of Regression (SSR) which has a degree of freedom of 12 (k - number of independent variables).
* Sum of Square of Errors (SSE) which has a degree of freedom of 177 (n-k-1 = 190-12-1 =177). Mean Sum of Square (MS) is obtained by dividing the respective Sums of Square by their degrees of freedom.
* F-statistic is the ratio of the mean square of regression to the mean square of residuals.   
  Sig or p-value is .000, which says the model is significant, p<0.05 and the model is a good fit.

1. **Coefficients table:**



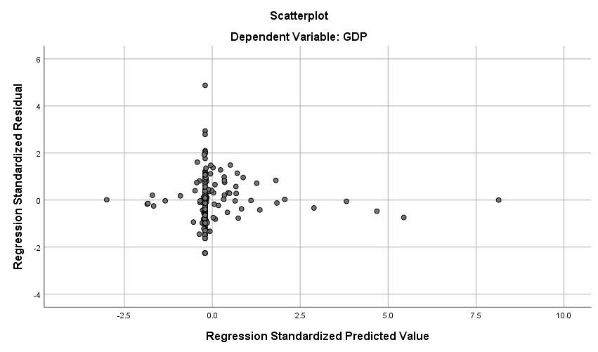
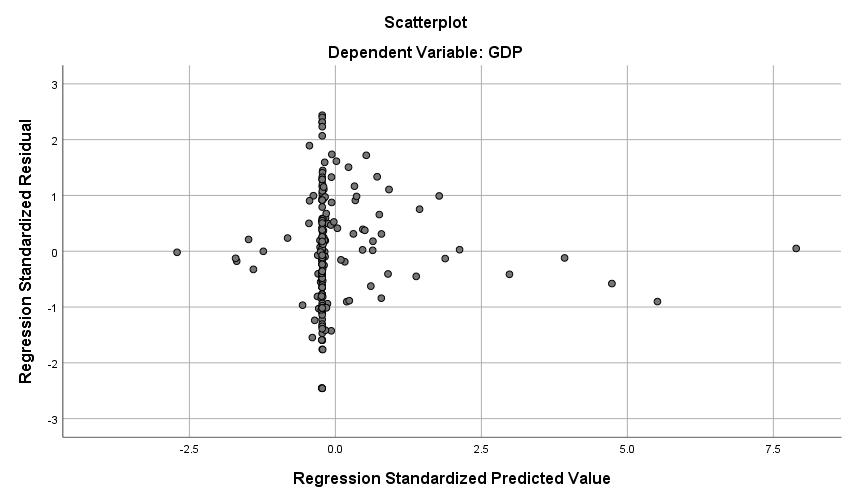
* The above table explains the significance value, VIF, and tolerance value of each predictor variable.
* Low tolerance and Variance Inflation Factor (VIF) are used to explain about multicollinearity.
* In general, the VIF value should be less than 10 and the tolerance value should be approximately nearly 1.
* If VIF is greater than 10, then it highly correlates with another predictor variable and it should be removed.
* In the above coefficients table, **Number.Of.licensed.qualified.surgeons.actively.working** and **Number.Of.licensed.qualified.obstetricians.actively.working** are highly correlated.
* Checking VIF and tolerance, **Physiotherapists.Number** is significant and (p<0.05). VIF and tolerance values are approximately equal to 1. So, it is the effective predictor of the dependent variable **GDP** which satisfies the assumptions.

1. **Normality of Errors:**



ss

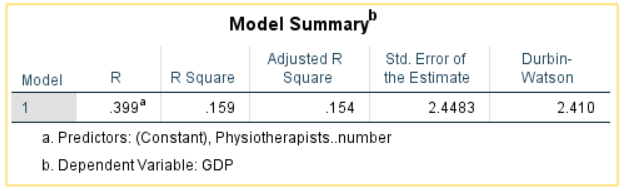
The regression standardized residual histogram displays the errors are normally distributed and the graph of Expected Cum Prob Vs Observed Cum Prob explains the linearity.

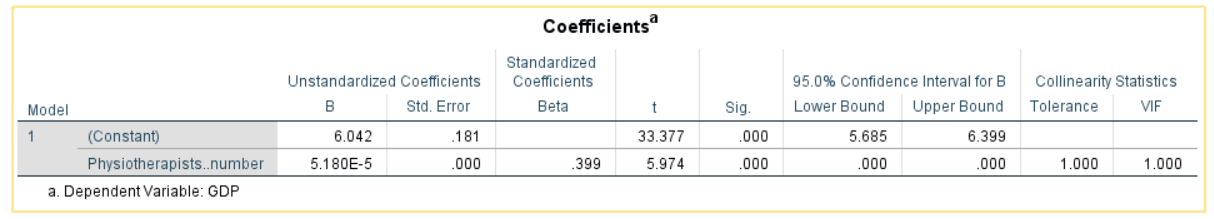


The residual standardized scatterplot on the left-side shows residuals before removing the outliers. The right-side scatterplot shows the residuals after removing the outliers greater than 3 SD, which satisfies the assumptions. But the scores are pileup at the center and residuals are symmetrically trailing off from the center, which satisfies the assumptions of homoscedastic (Jeremy Foster, 2006).

1. **Performance Tuning of the model:**

* The above observations show the evidence that **Number.of.licensed.qualified.surgeons.actively. working** and **Number.Of.licensed.qualified.obstetricians.actively.Working** attributes cannot be used for computing the model even though they have (p<0.05).
* Both variables have high VIF and highly correlated **(0.96)**. In this model, **Physiotherapists.Number** is the only attribute used to predict GDP which is the dependent variable and examine the model performance.





* From the above model summary and coefficients, it is evident that the physiotherapist number is the best predictor variable to calculate the GDP.
* It has both **tolerance and VIF 1** and **p<0.05**. It also has an R-square value of 15.9%. Even though the model has fewer R-square than the previous model, it has the best estimator for the dependent variable.

1. **Evaluation of the model:**

In general, multiple regression is expressed as:

c – constant

b1,b2,..,bn – coefficients of predictor variables.

Applying constant and coefficient of physiotherapist number to the above equation. We get

Suppose if there are 1000 physiotherapists in a country then their **GDP** is:

1. **Objectives of the analysis:**

The objective of this section is to perform binary logistic regression and analyze how the predictor variables are contributing to predict the dependent variable, which is a dichotomous value and how the model performance is improved after the introduction of an independent variable. The predictor variables can either be continuous or categorical values.

**Context of the data:**

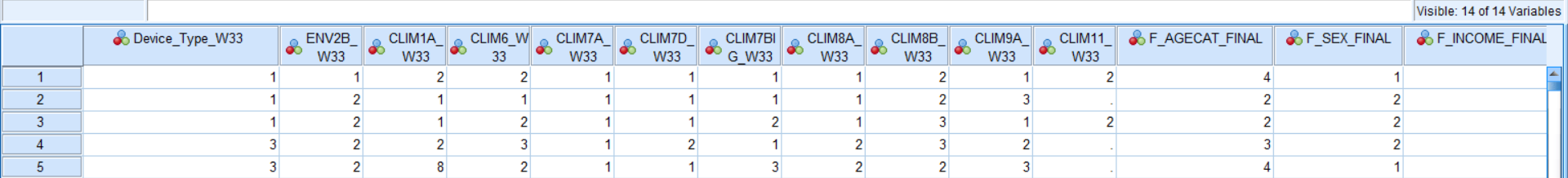
The dataset is sourced from (Center, 2019). In this dataset, logistic regression is performed to predict whether the effects of global climate change in your local community are impacting you, personally, or not **(CLIM11\_W33)**. The dataset contains 104 survey questions, only 13 predictor variables are considered in this analysis. The independent variables are:

1. Age Category - F\_AGECAT\_FINALs
2. Family Income - F\_INCOME\_FINAL
3. Which of these three statements about the Earth’s temperature comes closest to your view? - CLIM1A\_W33
4. Restrictions on power plant carbon emissions - CLIM7A\_W33
5. Do you favor or oppose EXPANDING each of the following sources of energy in our country? More nuclear power plants to generate electricity - ENV2B\_W33
6. Do you think policies aimed at reducing the effects of global climate change generally? - CLIM6\_W33.
7. Which ONE of the following proposals do you think would make the BIGGEST difference in reducing the effects of global climate change? - CLIM7BIG\_W33
8. Do you think the following would make a difference or not in reducing the effects of global climate change? Tax incentives for people to drive hybrid and electric automobiles - CLIM7D\_W33.
9. Sex - F\_SEX\_FINAL
10. Beyond the potential of these solar geoengineering techniques to reduce the effects of global climate change, overall, do you think using these techniques would? - CLIM8B\_W33
11. Do you think using these techniques would?- CLIM8A\_W33.
12. Wave 33 New Device Type - Device\_Type\_W33
13. How much, if at all, do you think global climate change is currently affecting…The United States? - CLIM9A\_W33

The dataset contains 2539 records, out of which 1491 has a valid response for the dependent variable. The dataset is filtered using case condition using SPSS, **Data** **→** **Select Cases .. →** **if condition is satisfied.**

|  |
| --- |
| **CLIM11\_W33=1 | CLIM11\_W33=2** |

**Dataset Sample View:**



The logistic regression is performed in SPSS, **Analyze** **→** **Regression →** **Binary Logistic**

The binary logistic is a more flexible technique than multiple regression because it doesn’t assume the nature of the relationship between independent and the predictor variables. Also, the independent variables need not be normally distributed (Jeremy Foster, 2006). In the above dataset, only **Age Category** is assumed to be categorical and the rest of the attributes are treated as continuous for the analysis.

1. **Results of Diagnosis/Tests of Multiple Regression:**
2. **Sample Size:**

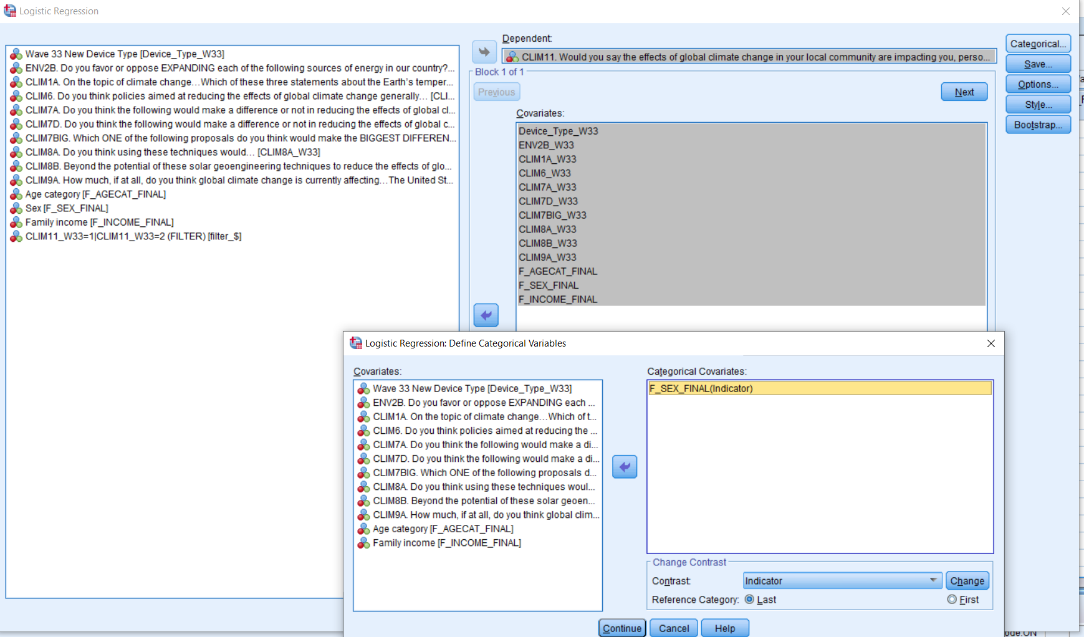
According to Fidell (Fidell, 2013), the minimum sample size requirement is N> 50+8m.

m – the number of predictor variables.

50+8(13) = 154

In this dataset **N = 1491> 154**, which is satisfactory.

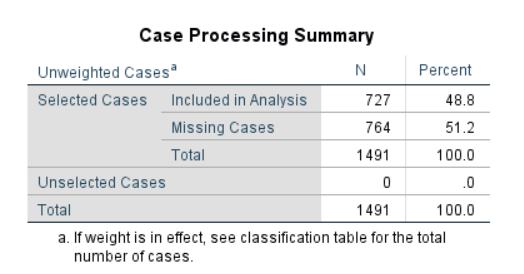
1. **SPSS Input:**



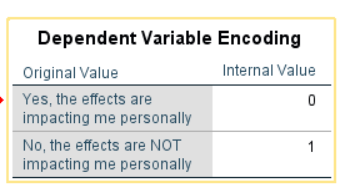
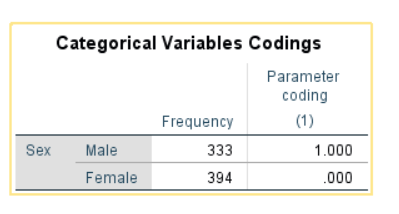
The above figure depicts the dependent and independent variables chosen for the analysis and the attribute of sex is chosen as categorical covariates. Generally, logistic regression there are 3 methods forward, backward and enter. In this analysis, the enter method is performed. Logistic regression we are choosing casewise residuals which are more than 2 standard deviations and eliminating the outliers.

1. **SPSS Outputs:**
2. **Casewise Summary:**

As discussed earlier after removing refused values, we are considering 1491 records for the analysis, out of which 727 records have values for all attributes which is 48.8 % of the total sample.

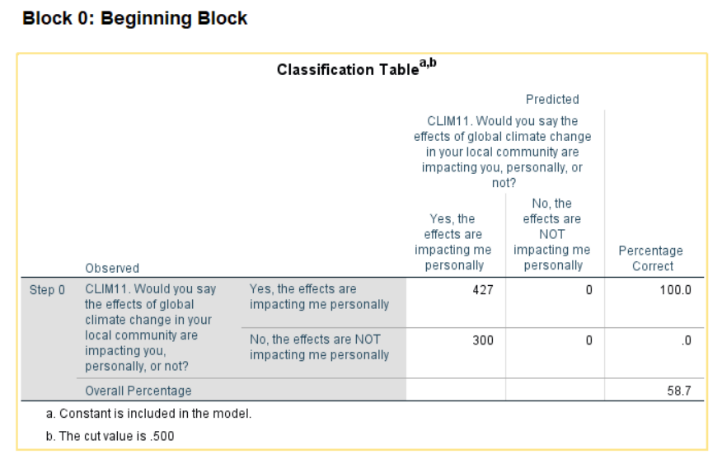


1. **Dependent and Categorical Variables:**

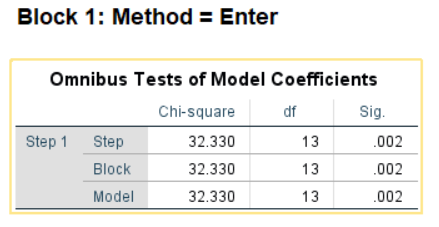
The dependent and categorical variables chosen in the inputs are discussed in the section. The dependent variable is the global climate change in your local community is impacting you, personally (0 – Yes, the effects are impacting me personally and 1 - No, the effects are impacting me personally). The categorical independent variable is sex (1-Male, 0-Female).

1. **Block 0:**

The baseline model of logistic regression is block 0 which does not include any independent variables. It considers only the constant value for the analysis and predicts all the true positive values correctly. In this analysis out of 727, 427 records are classified correctly as true positive but misclassify the remaining 300 records as a false negative, so the accuracy percentage is 58.7%.

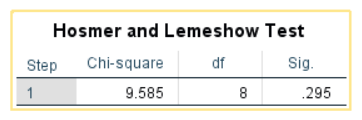


1. **Omnibus Tests of Block 1:**



* The enter method of logistic regression is that in one step it will include all the independent variables to the analysis.
* The goodness of fit of the model in the step, model, and block are considered by the Omnibus Tests of Model Coefficients.
* As we have placed all the independent variables in one block, the chi-square values are the same. The significant value (p<0.05) which says the chi-square value is significant and the predictor variables entered the model are the significant predictors of global climate change impacting personally or not (Jeremy Foster, 2006).

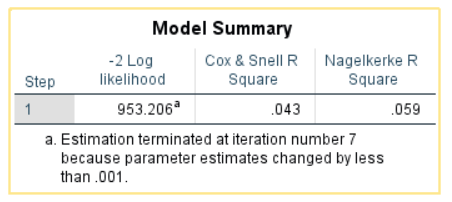
1. **Hosmer-Lemeshow Goodness of Fit Test:**



* In addition to the omnibus test of block 1, the goodness of the fit of the model is interpreted in a different way using Hosmer-Lemeshow fit test.
* In this case, the chi-square value is 9.585 and a Sig value is .295 for a degree of freedom 8.
* Unlike other tests, the Hosmer and Lemeshow test should denote Sig value (p>0.05). So, the model has Sig value is 0.295 and the model is a good fit.

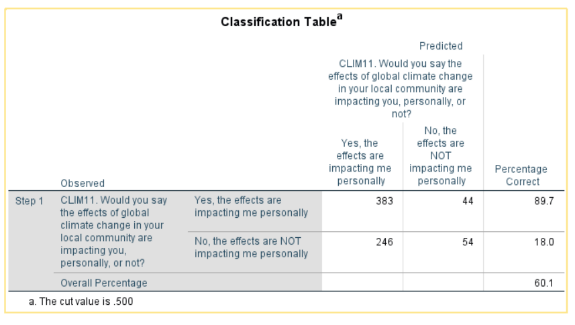
1. **Model Summary:**

The amount of variation in the dependent variable is explained by Cox & Snell R Square and the Nagelkerke R Square values. In this model, we get an R-square value of 4.3% to 5.9% which provides the usefulness of the model. This is also known as pseudo-R-square statistics. The third statistic is -2 Log-likelihood, which has a value of 953.206. It also decides the best fit for the line in logistic regression.

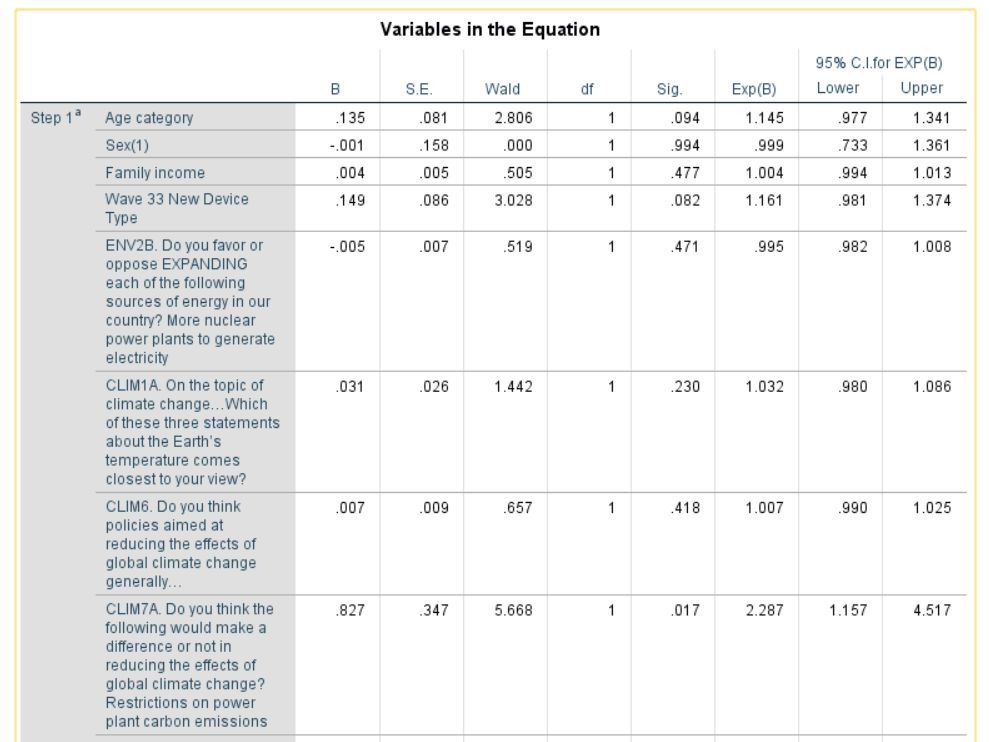


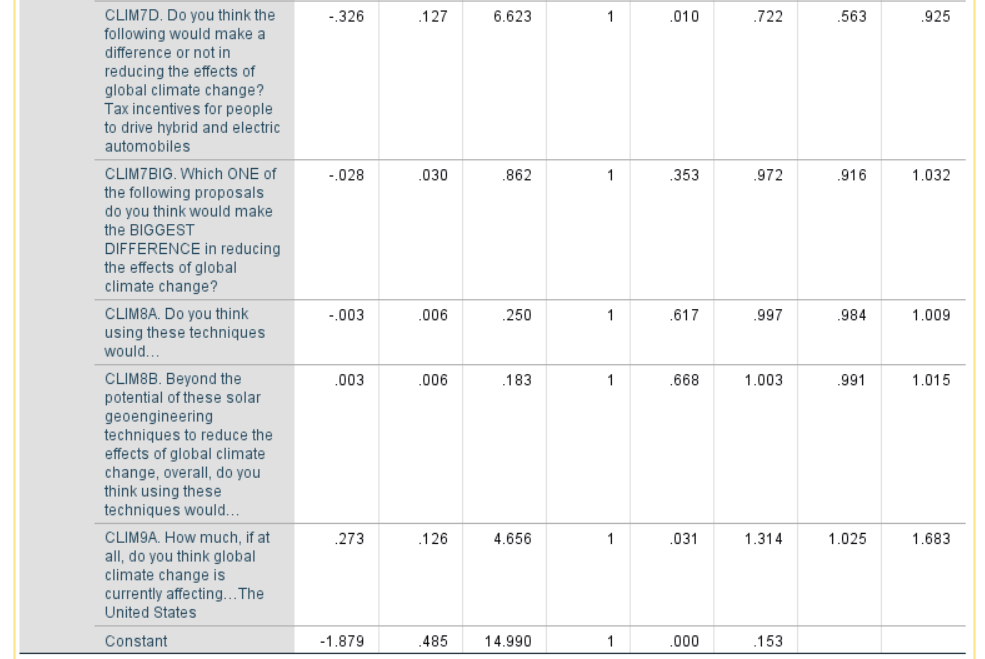
1. **Classification Table:**

* The number of participants correctly classified based on the dependent variable using this classification table. This can be compared with the classification table of model 0 and check the performance of predictor variables in the model.
* The **accuracy** of the model is 60.1 % which has an improvement of **1.4 %** from 58.7% in model 0. The sensitivity (classified true positive correctly) and specificity (classified true negative correctly) of the model is 18.0% and 89.7 % respectively.

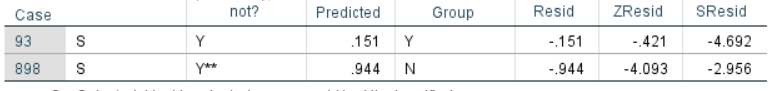


1. **Variables in the equation:**



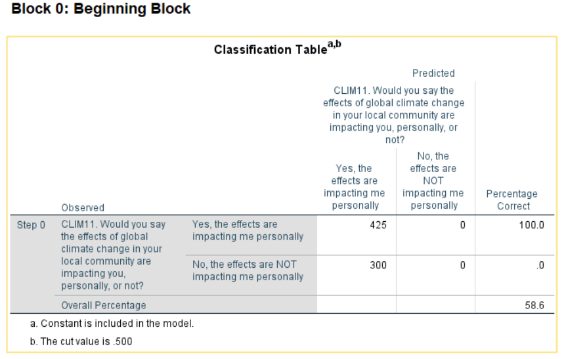
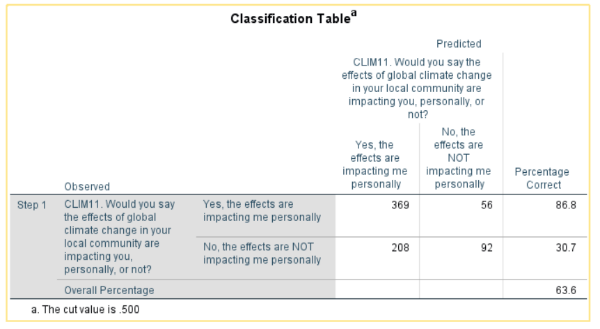


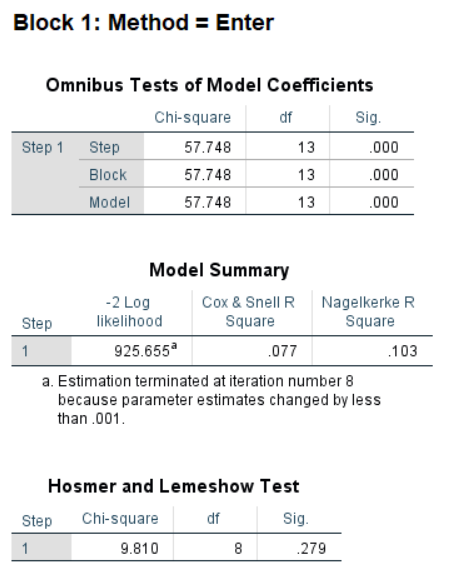
* The variables in the equation tell us the information about the importance of the predictor variables.
* Wald statistic is the test used to estimate the predictor variable based on sig value (p<0.05).
* From the above table, it is evident that attributes such as CLIM7A\_W33, CLIM7D\_W33, and CLIM9A have (p <0.05) and are significant.
* The above variables in the equation contain 2 outliers in the model.
* B value in logistic is similar to multiple regression, which tells about the likelihood of factor, increasing or decreasing the dependent variable.
* **Exp(B)** are odd ratios (OR) of each independent variable. In this case, the ratio of odds of Male saying yes, the global climate change in my local community is impacting me personally is 0.999 times higher than females.



1. **Performance Tuning and Evaluation of the model:**

* After removing the 2 outliers in the model, the dataset has 1489 records included in the analysis out of which 725 records don’t have missing values.
* Performing a logistic model with a new dataset, block 0 has an accuracy of 58.6 % in the classification table and there are no changes in the dependent and the categorical variables.
* After entering block 1, the omnibus test of model coefficients has a sig value of 0.000 and a chi-square value of 57.748.
* The model summary has Cox & Snell R-square value of 7.7 % and Nagelkerke R-square value is 10.3 which has improved from 4.3%-5.9% in the previous model.
* We also have sig (p>0.05) value in Hosmer and Lemeshow test which shows the model has a good fit.
* Finally, the classification table shows an accuracy of 63.6 %. Sensitivity is 30.7 % and Specificity is 86.8 % for the new model.
* The predictor variables in the equation which has sig (p<0.05) are Device\_Type\_W33, CLIM1A\_W33 and CLIM9A\_W33.









* The equation of logistic regression is

b1, b2, b3 are the coefficients of the predictors and b0 is the constant.

# References

Carlos, C. & Ken, B., 2012. Simple Regression and Correlation. In: *Statistics For Business and Economics.* s.l.: John Wiley & Sons, Ltd, pp. 502-503.

Center, P. R., 2019. *Pew Research Center.* [Online]   
Available at: https://www.pewresearch.org/science/dataset/american-trends-panel-wave-33/  
[Accessed November 2019].

Fidell, t., 2013. In: s.l.:s.n., p. 123.

Jeremy Foster, E. B. C. Y., 2006. *Understanding and Using Advanced Statistics.* First ed. s.l.: SAGE Publications.

Pallant, J. F., 2005. Descriptive statistics. In: *SPSS survival manual : a step by step guide to data.* s.l.:Allen & Unwin, pp. 51-52.

WHO, 2019. [Online]   
Available at: https://www.who.int/gho/en/