Statistics CA2

Name: Yogesh Maruti Patil

Student Id: x17169828

Multiple Regression:

1. Objective:

In this project we have used regression model that pertains to a summary estimate to how strongly crude death rate is dependent on other factors such as current health expenditure, food safety, physicians density and legislation score and predict the crude death rate accordingly.

2. Data:

In this project variables have been used from world health organization. These include

1. Crude death rate per 1000 population data for 43 countries for the year 2013 from http://apps.who.int/gho/data/view.main.CBDR2040 as dependent variable.

2. Current health expenditure (CHE) per capita in US$ data for 43 countries for the year 2013 from http://apps.who.int/gho/data/view.main.GHEDCHEpcUSSHA2011v as an independent variable.

3. Food safety data for 43 countries for the year 2013 from http://www.who.int/gho/ihr/monitoring/food_safety/en/ as an independent variable.

4. Physicians density per 1000 population data for 43 countries for the year 2013 from http://apps.who.int/gho/data/node.main.A1444 as an independent variable.

5. Legislation score data for 43 countries for the year 2013 from http://apps.who.int/gho//data/view.main.IHRCTRY01v?lang=en as an independent variable.

Variable types:

Variable	Type
Crude death rate per 1000 population	Dependent
Current health expenditure (CHE) per capita in US$	Independent
Food safety	Independent
Legislation score	Independent
Physicians density per 1000 population	Independent

File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help

1 :

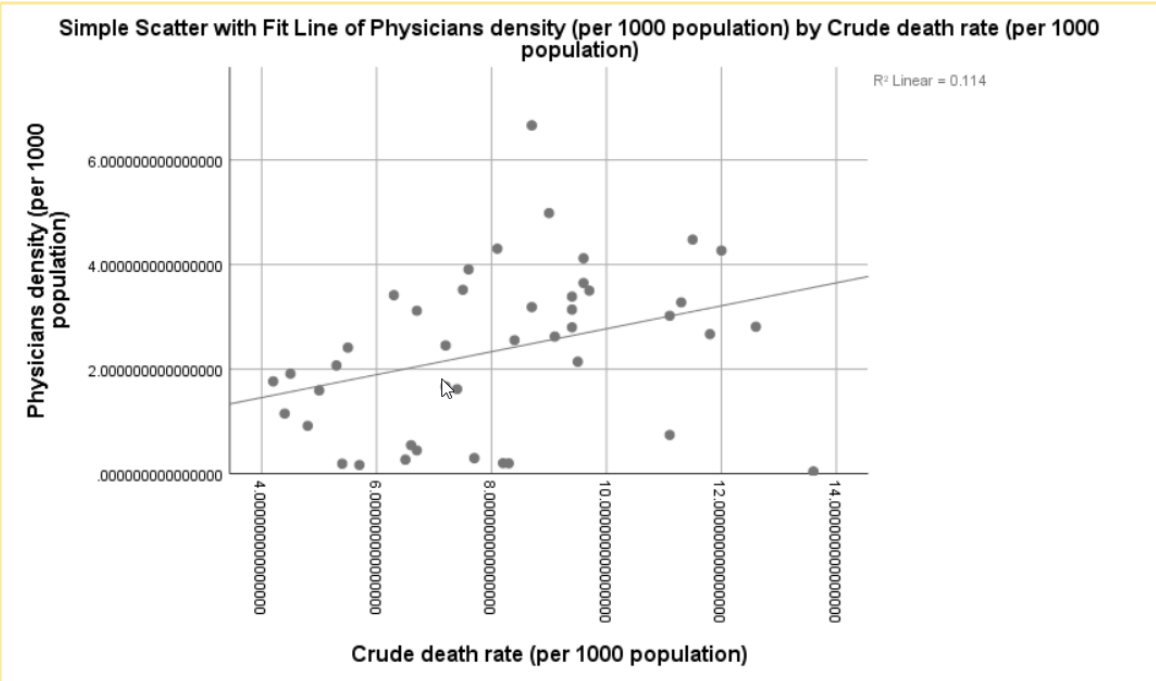
	Country	Crudedeathrateper1000population	CHE	Foodsafety	Physiciansdensityper1000population	Legislation	var
1	Afghanistan	7.70000000000000	38.106807440000000	20	.299000000000000	0	
2	Argentina	7.60000000000000	238.201684520000000	60	3.90700000000000	50	
3	Armenia	12.6000000000000	4.072009140000000	93	2.81100000000000	75	
4	Austria	9.00000000000000	3186.896790719999400	93	4.98400000000000	100	
5	Azerbaijan	6.30000000000000	132.568514980000000	93	3.41500000000000	50	
6	Bhutan	6.50000000000000	19.983669900000000	27	.269000000000000	50	
7	Cambodia	5.70000000000000	19.149137140000000	47	.168000000000000	50	
8	Canada	7.20000000000000	8572.535169550000000	100	2.45300000000000	100	
9	Chad	13.6000000000000	91.885767180000000	27	.044000000000000	25	
10	China	7.20000000000000	42539.798538140010000	100	1.66800000000000	100	
11	Colombia	4.20000000000000	3143.526488230000300	67	1.76600000000000	100	
12	Costa Rica	4.40000000000000	99.524991970000000	100	1.15000000000000	100	
13	Croatia	11.1000000000000	107.918710580000010	100	3.02100000000000	100	
14	Cyprus	5.50000000000000	34.519383960000000	100	2.41200000000000	75	
15	Denmark	9.60000000000000	2435.762115779999500	100	3.64800000000000	100	
16	Estonia	11.3000000000000	116.150772940000000	100	3.27700000000000	25	
17	Finland	9.40000000000000	1124.179622100000000	87	3.13900000000000	100	
18	France	8.70000000000000	17211.005909770000000	100	3.18700000000000	100	
19	Georgia	11.5000000000000	46.891908140000000	67	4.47700000000000	100	
20	Kazakhstan	9.70000000000000	1700.403341370000400	100	3.50100000000000	100	
21	Kenya	8.30000000000000	224.273635410000050	73	.199000000000000	50	
22	Kiribati	8.20000000000000	4.146516980000000	67	.203000000000000	100	
23	Lao People's Democratic Republic	6.70000000000000	10.080915319999999	87	.447000000000000	100	
24	Lithuania	12.0000000000000	136.749867230000000	100	4.26800000000000	100	
25	Malta	7.50000000000000	44.772968670000010	100	3.51800000000000	100	
26	Mauritius	7.40000000000000	37.848488250000000	60	1.61900000000000	75	

Fig. Sample view of data

3. Linearity

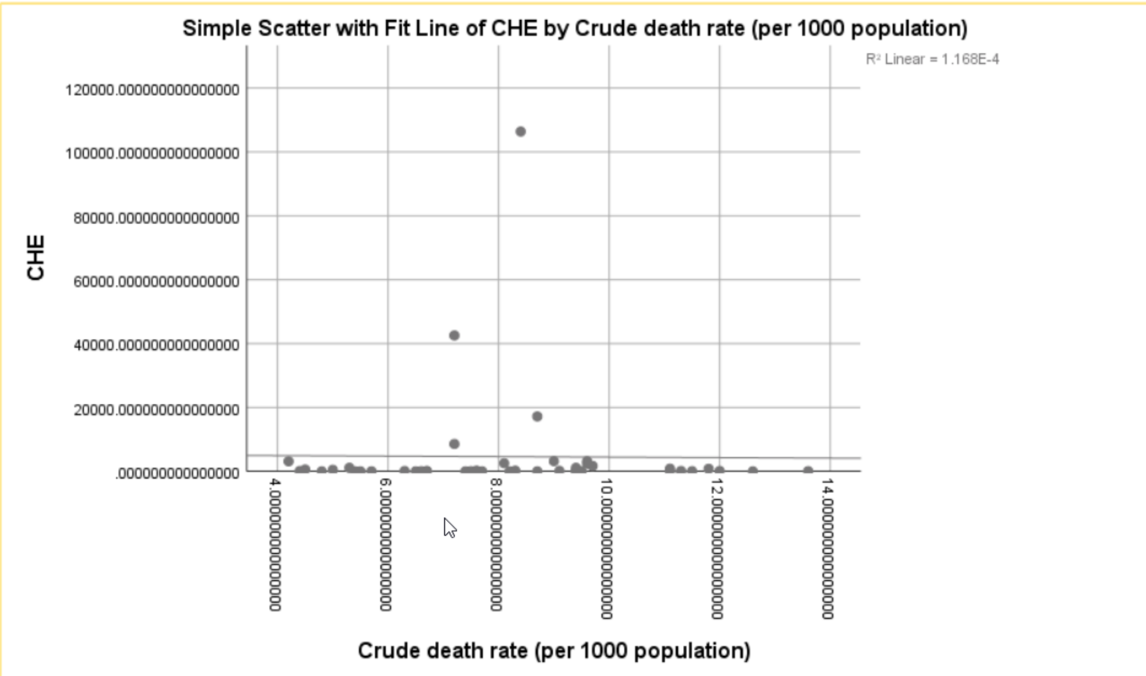
Linear regression requires the relationship between the independent and dependent variables to be linear. Scatterplots can show whether there is a linear or curvilinear relationship.

GGraph

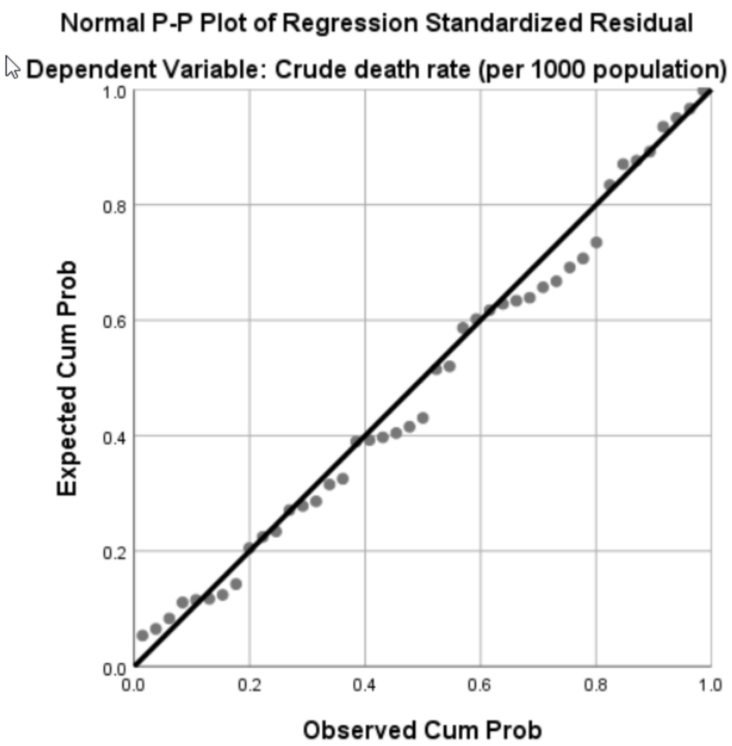


GGraph

GGraph

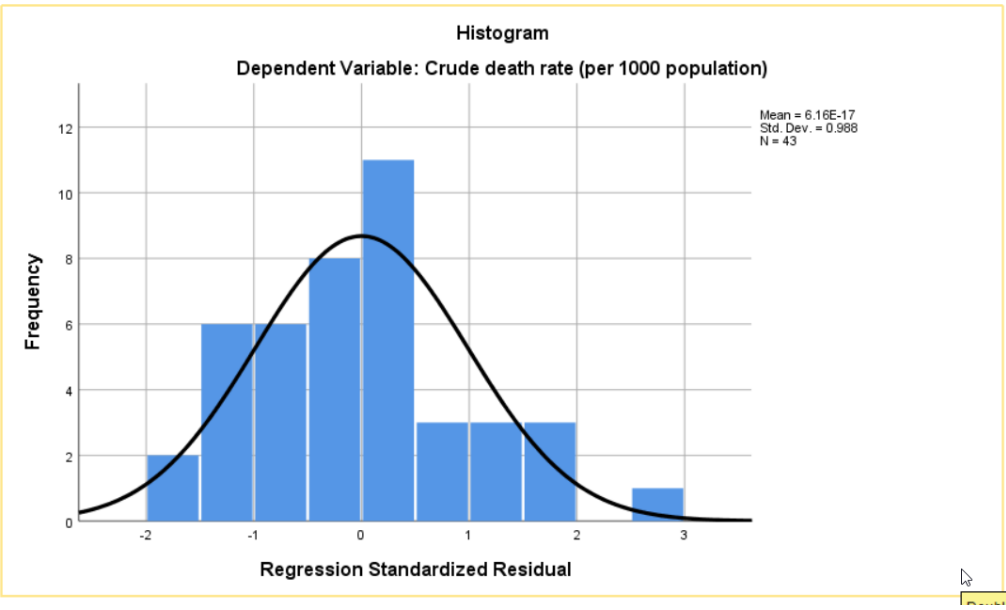


The plot shows the linearity in the model



The chart shows that the model is normally distributed.

Charts



4. Correlation Matrix

By referring to the correlation table we have come to some findings which are as follows:

1. From the analysis we can state that out of the four independent variables two that are food safety and physician density have a positive effect on the dependent variable and other two namely CHE and legislation Score have a negative effect on the dependent variable.
2. Physician density had the highest effect accounting to 0.338 r value and 0.013 p value.
3. We could observe the highest correlation between food safety and physician density with r value 0.595 and p value as 0.000.

Correlations

		Crude death rate (per 1000 population)	CHE	Food safety	Physicians density (per 1000 population)	Legislation
Pearson Correlation	Crude death rate (per 1000 population)	1.000	-.011	.086	.338	-.083
	CHE	-.011	1.000	.195	.021	.184
	Food safety	.086	.195	1.000	.595	.686
	Physicians density (per 1000 population)	.338	.021	.595	1.000	.380
	Legislation	-.083	.184	.686	.380	1.000
Sig. (1-tailed)	Crude death rate (per 1000 population)	.	.473	.292	.013	.298
	CHE	.473	.	.106	.446	.118
	Food safety	.292	.106	.	.000	.000
	Physicians density (per 1000 population)	.013	.446	.000	.	.006
	Legislation	.298	.118	.000	.006	.
N	Crude death rate (per 1000 population)	43	43	43	43	43
	CHE	43	43	43	43	43
	Food safety	43	43	43	43	43
	Physicians density (per 1000 population)	43	43	43	43	43
	Legislation	43	43	43	43	43

5. Model Summary

From the model summary following observations are made

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.409[a]	.167	.079	2.295608806	.167	1.907	4	38	.129	2.228

a. Predictors: (Constant), Legislation, CHE, Physicians density (per 1000 population), Food safety

b. Dependent Variable: Crude death rate (per 1000 population)

1. In the model Summary R value represents the correlation between the outcome and the predictor i.e. Physician Density, CHE, Legislation and Food safety. Over here the R value is 0.409.
2. In the model summary R square represents the amount of variability in the outcome by the predictors. Over here the R square value is 0.167.
3. In the model summary adjusted R square value represents how well the model generalizes and the difference for the final model means that if model were derived from the population rather than sample then it would account for 0.088 less variance in the outcome.
4. The Durbin-Watson statistic informs us about whether the assumption of independent errors is tenable and Durbin-Watson value in this model is 2.228 which is close to 2 and in between 1-3 represents it is tenable.

5. ANOVA

1. In order to tell if a regression model is significantly better at predicting values of the outcome ANOVA is used.
2. As in the ANOVA table Sig value is 0.27 which means the model is significant. Values closer to zero represents that model is significant.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	27.469	1	27.469	5.288	.027[b]
	Residual	212.979	41	5.195		
	Total	240.448	42			

a. Dependent Variable: Crude death rate (per 1000 population)

b. Predictors: (Constant), Physicians density (per 1000 population)

		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
Model		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	8.226	1.354		6.077	.000					
	CHE	3.778E-6	.000	.027	.180	.858	-.011	.029	.027	.944	1.059
	Food safety	-.001	.025	-.010	-.042	.967	.086	-.007	-.006	.391	2.559
	Physicians density (per 1000 population)	.674	.286	.437	2.355	.024	.338	.357	.349	.635	1.574
	Legislation	-.020	.016	-.247	-1.211	.233	-.083	-.193	-.179	.526	1.902

Coefficients[a]

a. Dependent Variable: Crude death rate (per 1000 population)

Below mentioned is the equation for multiple regression as per the coffecient values

Y=a + b1X1 + b2X2 + b3X3

In this case X1 is Current health expenditure (CHE) per capita in US$,

X2 is Food Safety,

X3 is Legislation score,

X4 is Physicians density per 1000 population

So after entering the values from the coefficients table

Crude death rate per 1000 population = 8.226 + (3.78) X1+ (-0.001) X2 + (.674) X3 + (-0.20) X4

So, for x1 = 3000, x2 = 90, x3 = 80, x4= 70

Crude death rate = = 8.226 + (3.78) 3000+ (-0.001) 90 + (.674) 80 + (-0.20)70

= 11388.056.

Below mentioned observations are being observed:

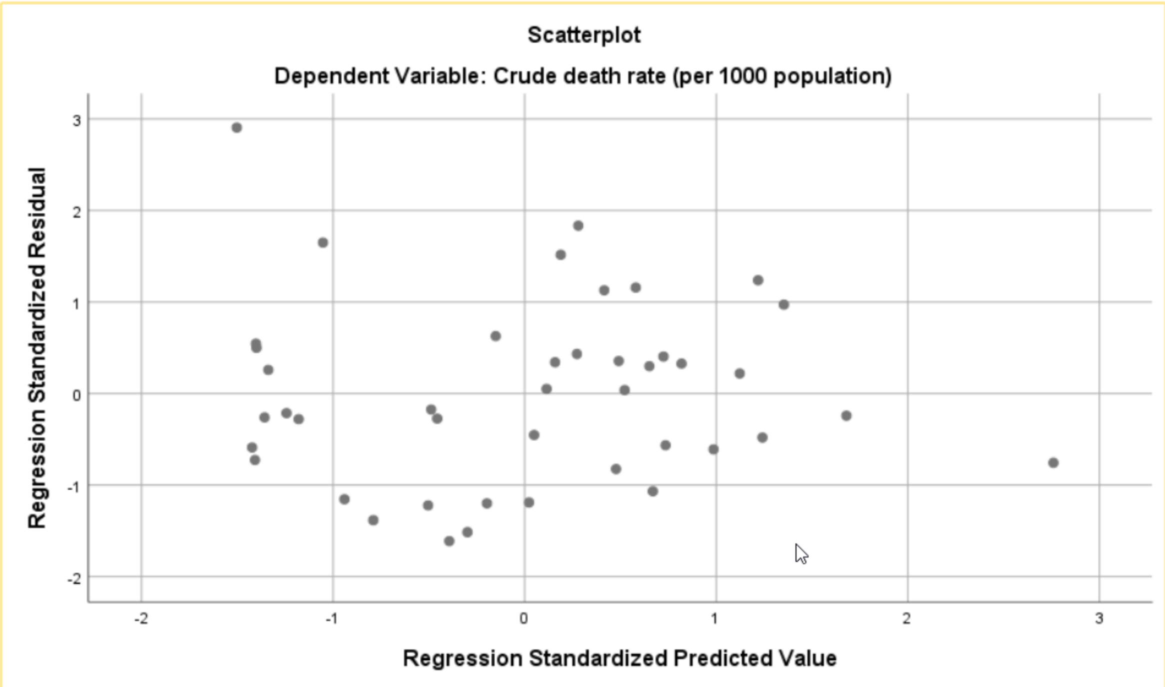
1. B=3.77 represents that if Current health expenditure increases by 1 % then the Crude death rate per 10000 will increase by 3.77
2. B=-0.01 represents that if Food safety decreases by 1 % then the Crude death rate per 10000 will decrease by 0.01
3. B=.674 represents that if Physician density per 1000 population increases by 1 % then the Crude death rate per 10000 will increase by .674
4. B=-0.20 represents that if Legislation Score decreases by 1 % then the Crude death rate per 10000 will decrease by 0.20
5. Significance level of predictors is being represented by Sig value. Larger the value of t the greater the contribution of predictors and vice versa.
6. Test of multicollinearity can be done using checking the VIF values i.e (1.059,2.559,1.574,1.902) are observed to be less than 10 and average of all values

is coming as 1.77 which represents that there is no cause of concern in multicollinearity.

7. Residual Statistics

A residual plot is a graph that shows the residuals on the vertical axis and the independent variable i.e predicted values on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data.

In the below scatter plot all standardised values lies between -3 and +3 which means the model is suitable.



Influential Cases:

Cook's distance measures is a way to identify points that negatively affect your regression model. In this case the cooks distance value is 0.303 which is less than 1 so this is not a concern of worry in this case.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	6.311332703	11.12565994	8.193023256	.9782729544	43
Std. Predicted Value	-1.923	2.998	.000	1.000	43
Standard Error of Predicted Value	.462	2.115	.723	.303	43
Adjusted Predicted Value	6.028482914	11.92119217	8.183270309	1.068424578	43
Residual	-3.17669654	5.867140770	.0000000000	2.183559524	43
Std. Residual	-1.384	2.556	.000	.951	43
Stud. Residual	-1.454	2.793	.000	1.008	43
Deleted Residual	-3.69440389	7.006353855	.0097529468	2.463674023	43
Stud. Deleted Residual	-1.476	3.091	.008	1.037	43
Mahal. Distance	.721	34.677	3.907	5.490	43
Cook's Distance	.000	.303	.026	.050	43
Centered Leverage Value	.017	.826	.093	.131	43

a. Dependent Variable: Crude death rate (per 1000 population)

Logistic Regression:

1. Objective:

In this project we have used logistic regression model that pertains to a summary estimate in order to predict the type of education on the bases of sex and age.

2. Data:

In this project variables have been used from Europa. These include

1. Education Type data in European countries for the year 2014 from http://appsso.eurostat.ec.europa.eu/nui/submitViewTableAction.do as a dependent variable.

Education type consists of Upper secondary and post-secondary non-tertiary education (levels 3 and 4) and Tertiary education (levels 5-8).

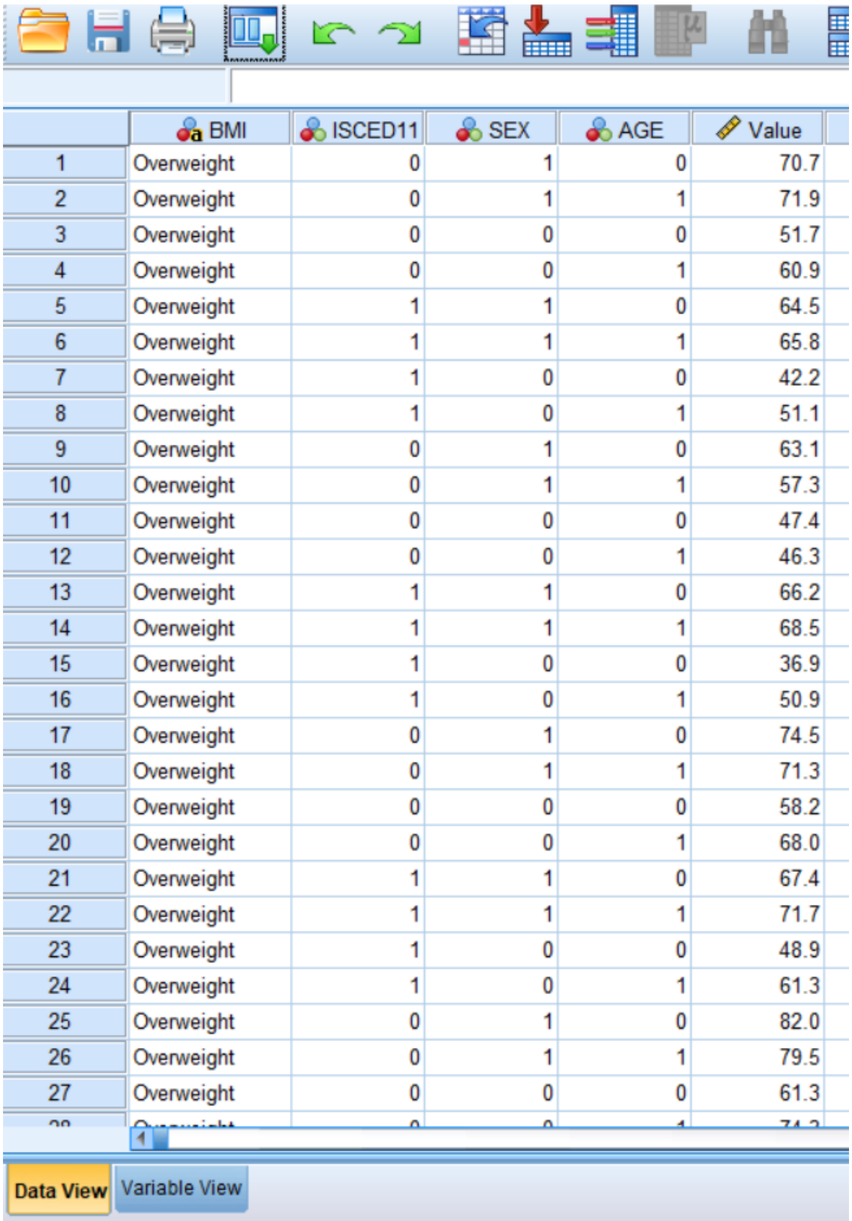
2. Sex Type data in European countries for the year 2014 from http://appsso.eurostat.ec.europa.eu/nui/submitViewTableAction.do as independent variable.

3. Age group data in European countries for the year 2014 from http://appsso.eurostat.ec.europa.eu/nui/submitViewTableAction.do as an independent variable.

Age group included in this project is From 45 to 64 years and From 65 to 74 years.

.

Variable	**Type**
Education Type	Dependent
Sex	Independent
Age Group	Independent



	BMI	ISCED11	SEX	AGE	Value
1	Overweight	0	1	0	70.7
2	Overweight	0	1	1	71.9
3	Overweight	0	0	0	51.7
4	Overweight	0	0	1	60.9
5	Overweight	1	1	0	64.5
6	Overweight	1	1	1	65.8
7	Overweight	1	0	0	42.2
8	Overweight	1	0	1	51.1
9	Overweight	0	1	0	63.1
10	Overweight	0	1	1	57.3
11	Overweight	0	0	0	47.4
12	Overweight	0	0	1	46.3
13	Overweight	1	1	0	66.2
14	Overweight	1	1	1	68.5
15	Overweight	1	0	0	36.9
16	Overweight	1	0	1	50.9
17	Overweight	0	1	0	74.5
18	Overweight	0	1	1	71.3
19	Overweight	0	0	0	58.2
20	Overweight	0	0	1	68.0
21	Overweight	1	1	0	67.4
22	Overweight	1	1	1	71.7
23	Overweight	1	0	0	48.9
24	Overweight	1	0	1	61.3
25	Overweight	0	1	0	82.0
26	Overweight	0	1	1	79.5
27	Overweight	0	0	0	61.3

Fig: SPSS View of data

3. Case processing Summary

The case processing summary tells us about the number of cases included in the analysis.

The second row tells that there are 3 data missing on some of the parameters and thereafter 253 are being used in the analysis.

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	253	98.8
	Missing Cases	3	1.2
	Total	256	100.0
Unselected Cases		0	.0
Total		256	100.0

a. If weight is in effect, see classification table for the total number of cases.

4. Dependent Variable Encoding

Dependent variable encoding tells us how our outcome variable is encoded. In our case its 0 or 1.

Dependent Variable Encoding

Original Value	Internal Value
0	0
1	1

Block 0: Beginning Block

Classification Table[a,b]

Observed			Predicted ISCED11 0	1	Percentage Correct
Step 0	ISCED11	0	0	126	.0
		1	0	127	100.0
	Overall Percentage				50.2

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.008	.126	.004	1	.950	1.008

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	SEX	.004	1	.950
		AGE	.004	1	.949
		Value	16.617	1	.000
	Overall Statistics		29.357	3	.000

5. Beginning Block

- The first model is the model with no predictors and is also called null predictors.
- The constant in the second table named Variables in the Equation gives the unconditional log odds of type of education.
- The third table named labelled Variables not in the Equation provided the results of the score test. The column labelled Score gave the estimated change in model fit if the term is added to the model, the other two columns give the degrees of freedom, and p-value i.e. Sig. for the estimated change. Based on the table, all three of the predictors, age, sex and overweight value are expected to improve the fit of the model.

6. Omnibus Test
   Basically it is used to check if it's better than the baseline model.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	31.274	3	.000
	Block	31.274	3	.000
	Model	31.274	3	.000

- The table named Omi test for model Coefficients gives the overall test for the model that includes the predictors. Here the chi-square value of 31.274 with p value i.e Sig of less than 0.0005 signifies that this model fits significantly better than a model with no predictors. The model will be a good predictor as the predictive variable is going to do a good job of making a prediction.

7. Model Summary

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	319.455[a]	.116	.155

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

- Nagelkerke R^2 value is 0.155 which indicates that the model is descent but not that great.
- It signifies that 15 % of variation in the outcome is being predicted by the model
- With the help of Cox & Snell's R^2 value of .116 so we can interpret its value as 11 % probability of the type of education is being explained by the logistic model.

8. Hosmer and Lemeshow Test

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.797	8	.280

- Here we want the value to be greater than 0.05 in order to classify this model as a good model which in this case is .280 which is good.

9. Contingency Table

Contingency Table for Hosmer and Lemeshow Test

		ISCED11 = 0		ISCED11 = 1		
		Observed	Expected	Observed	Expected	Total
Step 1	1	17	19.797	8	5.203	25
	2	18	17.038	7	7.962	25
	3	15	15.681	10	9.319	25
	4	17	14.327	8	10.673	25
	5	12	13.101	13	11.899	25
	6	11	11.859	14	13.141	25
	7	15	10.796	10	14.204	25
	8	12	9.673	13	15.327	25
	9	6	7.987	19	17.013	25
	10	3	5.742	25	22.258	28

- This above table also tells us about how good our model is. It breaks the outcome into groups and progressively tries to fit our model to the actual outcomes.
- From the columns Observed and Expected values we can get a clear idea that the closer the numbers are in these columns the better is the model. In my case the variations is less so it’s a good model.

10. Classification Table

Classification Table[a]

Observed			Predicted ISCED11 0	Predicted ISCED11 1	Percentage Correct
Step 1	ISCED11	0	80	46	63.5
		1	47	80	63.0
	Overall Percentage				63.2

a. The cut value is .500

- The above tables tell us that the model is able to predict 63.2 % of the categories so we can say that 63.2 % of all the outcomes were correctly predicted by this model which is much better than the null hypothesis which was 50.2 %. So as we can get near the 65% correct threshold we can say that this model is a good model and we are doing well with the predictive ability.

11. Variables in the Equations

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	SEX	1.309	.371	12.414	1	.000	3.701
	AGE	.308	.273	1.279	1	.258	1.361
	Value	-.083	.016	25.956	1	.000	.920
	Constant	4.326	.883	23.981	1	.000	75.670

a. Variable(s) entered on step 1: SEX, AGE, Value.

- Here are the odd ratios related to each of these variables and so higher the odd ratio is over one the more likely it is to predict the type of education if they have high values. It also tells the direction of influence for each variable.

- So for example if the Exp(B) is 3.701 in case of Sex so we can say that its 3.7 times more likely to predict the type of education by sex .

- Also these values give the magnitude of the effect that each of these variables might have on predicting the outcome.

- As per the Wald test more the value more it is contributing to the prediction. So in this case sex is likely to predict the type of outcome.

- B tells us what effect the predictors will have on the dependent variables in terms of standard deviation.