STATISTICAL ANALYSIS USING MULTIPLE LINEAR REGRESSION AND LOGISTIC REGRESSION IN SPSS:

Multiple linear regression:

Objectives of Analysis: Building a multiple regression model to be able to predict the value of one variable (outcome or DV) based on two or variables (predictors or IV).

This model is to see if the amount of cigarette smoking can be predicted based on drinking habits and symptoms of depression.

Please note, in my research I have labelled dependent variables as DV(outcome) and independent variables as IV(predictors).

Data Description:

The below datasets have been taken from Eurostat's (see Appendix)

1.Daily consumption of fruit and vegetables by sex, age and educational attainment level

2.Frequency of heavy episodic drinking by sex, age and educational attainment level

3. Current depressive symptoms by sex, age and educational attainment level.

These 3 measurements are for the year 2014 taken across 29 nations hence we have 29 measurements for each variable. From these I have selected my outcome and predictors as shown below:

DV-People smoking ""20 or more cigarettes a day". IV1-People who "Drink at least once a week". IV2-People having Depressive symptoms". The final dataset was cleaned and formed from these three datasets using pivot table and VLookup on Excel.

ASSUMPTIONS: To check if data can be analysed using multiple regression.

1: DV is measured on a continuous scale and we require a minimum two or more IV's to be either continuous or ordinal.

In our model we have number people smoking "20 or more cigarettes a day" as DV measured on a continuous scale and also two IV's which are count of people who "Drink at least once a week" and count of people who have "depressive symptoms" as continuous measurements.

2: The relationship between each of the IVs and the DV is linear.

The second assumption of Multiple Regression is that the relationship between each IV and the DV should be represented by a straight line. This check is done by producing scatterplots of the relationship between each of our IVs and our DV.

Dependent variable in Y axis against each IV in X axis to show each of the IV's have a linear relationship with the DV shown by a straight line. The simple scatterplots of our model is shown below:

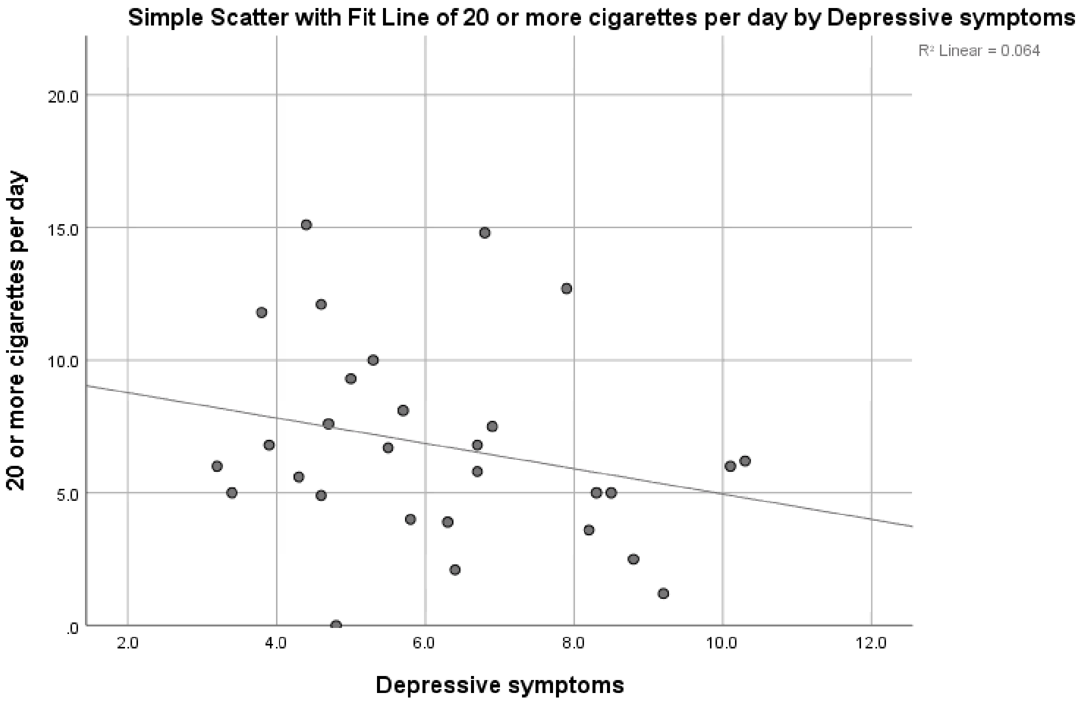


Fig: Scatter plot of DV against 1st IV.

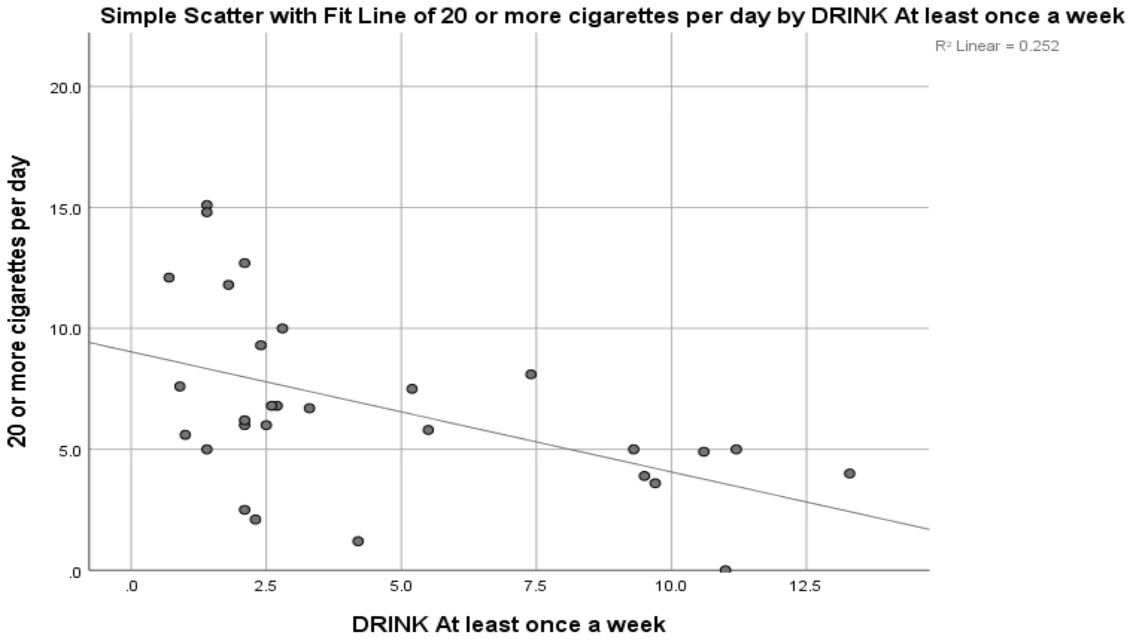
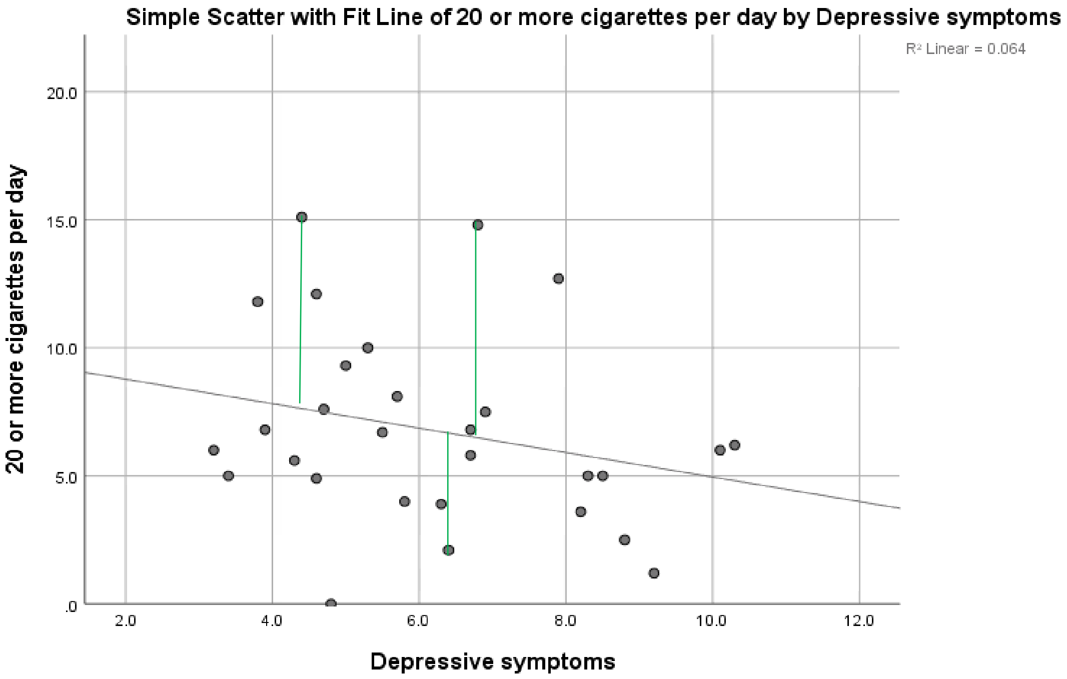


Fig: Scatter plot of DV against 2nd IV.

Also we check for residuals to show the errors in our model. In real life scenarios the data has errors. The residual is the vertical distance shown as green coloured lines between the points and the straight line. The closer the points are the more accurate the model is and vice versa.



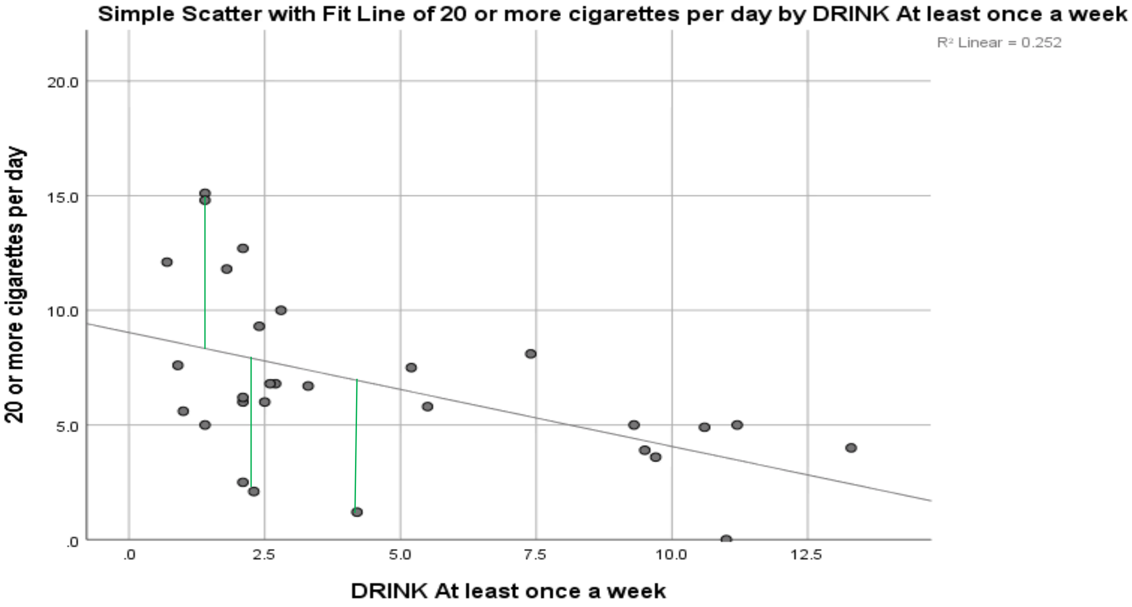
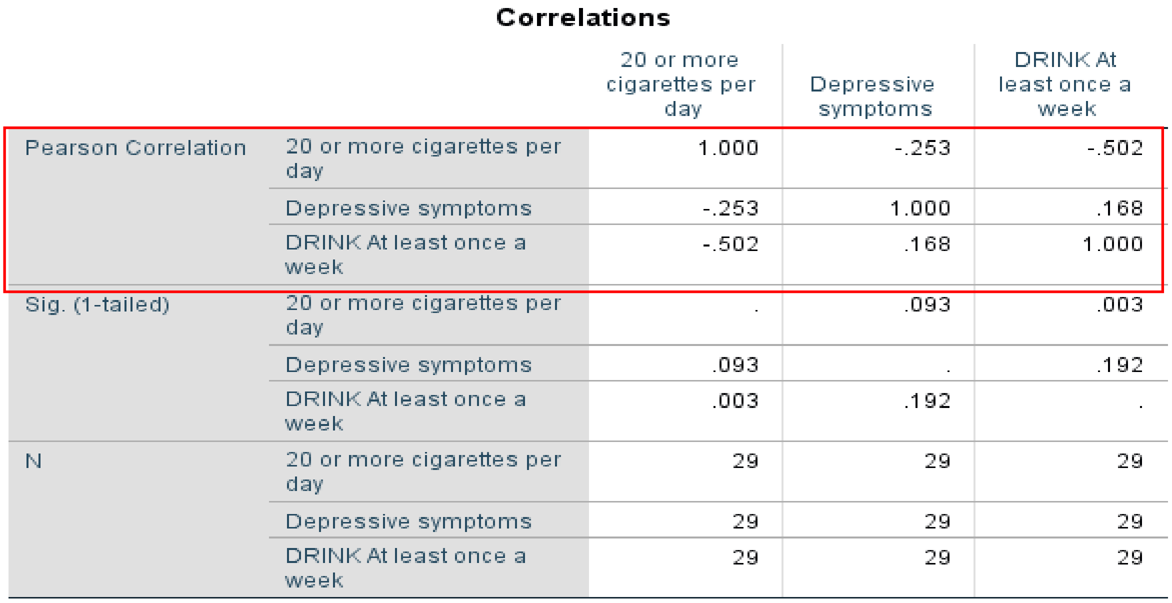


Fig: Two scatterplots of DV vs each IV showing the residuals.

The rest of the assumptions are performed under linear regression analysis.

3: Check for multicollinearity:

The predictors or IV's should not highly correlate with each other. For this check, we perform collinearity diagnostics which shows the various correlations.



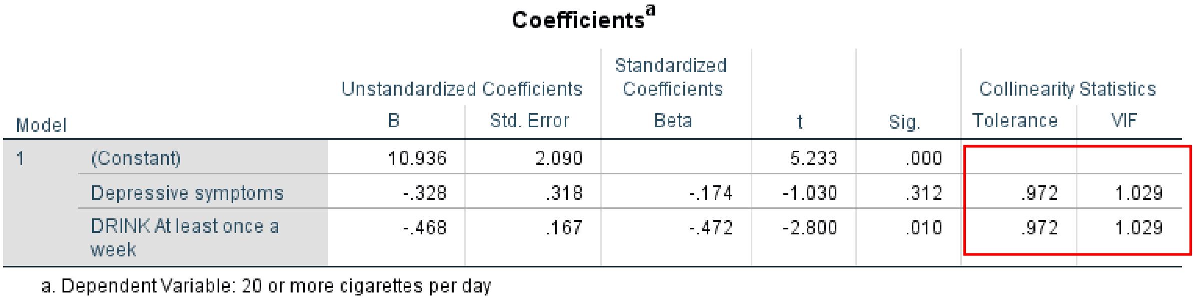
Correlations

		20 or more cigarettes per day	Depressive symptoms	DRINK At least once a week
Pearson Correlation	20 or more cigarettes per day	1.000	-.253	-.502
	Depressive symptoms	-.253	1.000	.168
	DRINK At least once a week	-.502	.168	1.000
Sig. (1-tailed)	20 or more cigarettes per day	.	.093	.003
	Depressive symptoms	.093	.	.192
	DRINK At least once a week	.003	.192	.
N	20 or more cigarettes per day	29	29	29
	Depressive symptoms	29	29	29
	DRINK At least once a week	29	29	29

Fig: Correlation Table with correlation coefficients.

Under Pearson Correlation we can see model has all values below the threshold of 0.8 and we can say that there is no multicollinearity in our data, the highest correlation is 0.168.

Another way to test for this assumption is by looking at VIF and Tolerance under coefficients a table. VIF values should be well below 10 and the tolerance scores should be above 0.2 .



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	10.936	2.090		5.233	.000		
	Depressive symptoms	-.328	.318	-.174	-1.030	.312	.972	1.029
	DRINK At least once a week	-.468	.167	-.472	-2.800	.010	.972	1.029

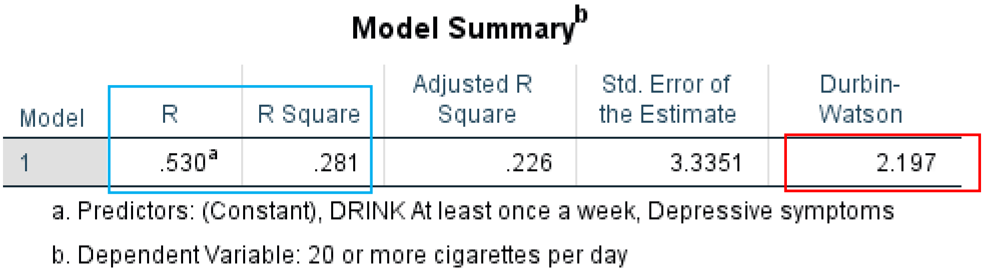
a. Dependent Variable: 20 or more cigarettes per day

Fig: Coefficients Table showing Tolerance and VIF.

In our model we have VIF as 1.029 and tolerance as 0.972 which shows no high correlation among the IV's.

4: The values of the residuals are Independent or have independence of observation.

This residual test can be checked using Durbin-Watson statistic. To meet the criteria for this assumption we require the Durbin-Watson statistic to be close to 2 whereas this generally varies between 0 and 4. Table is shown below:



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.530[a]	.281	.226	3.3351	2.197

a. Predictors: (Constant), DRINK At least once a week, Depressive symptoms

b. Dependent Variable: 20 or more cigarettes per day

Fig: MODEL SUMMARY TABLE.

As per out table we have the Durbin-Watson to be 2.197 which is pretty close to 2 and we can understand this assumption is met.

5: Data should show homoscedascity i.e. the variance of the residuals is constant.

This it to say that the spread of the residuals should be fairly constant across the linear model. We produce a scatterplot that includes the entire model and not just individual IV's. That is by plotting predicted values against residual values. ZPRED-X axis and ZRESID –Y axis using PLOT.

The assumption states that as predicted values increase in the X axis, variations in the residuals should also be similar. The scatterplot is to look like a random distribution of dots and not like a funnel shaped distribution of dots.

The output of the scatterplot is shown below and we can see that this assumption has been met.

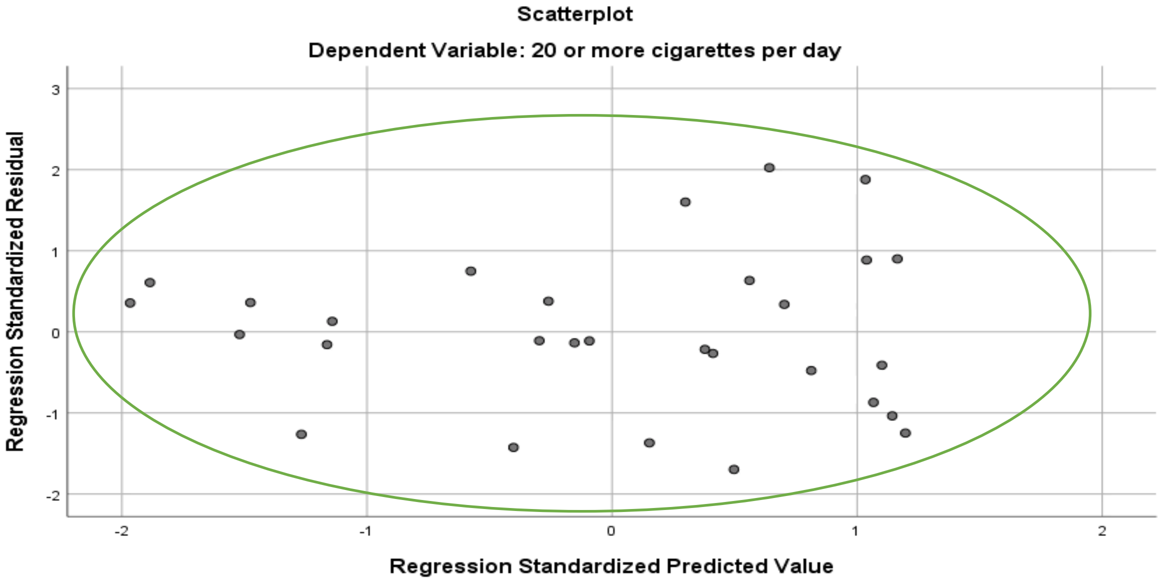


Fig: Scatterplot of entire model.

6: The residuals should be approximately normally distributed:

Which can be done by checking the normal probability plot under Standardized residual plots.

The P-P plot for the model is shown below:

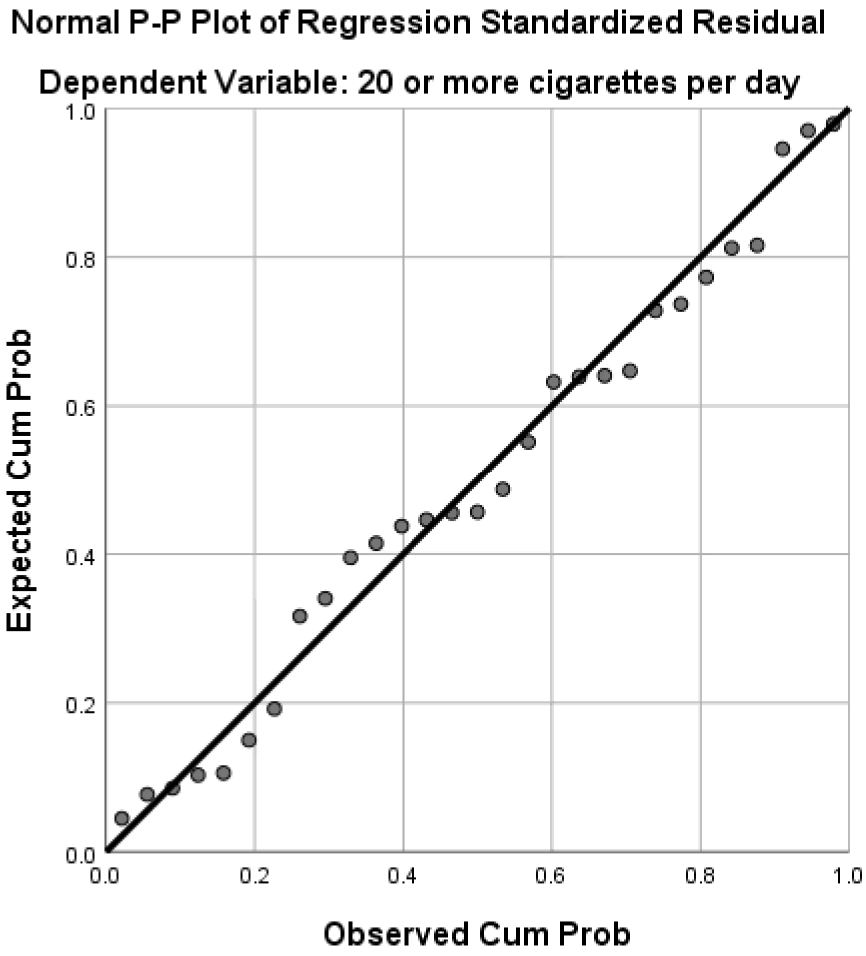


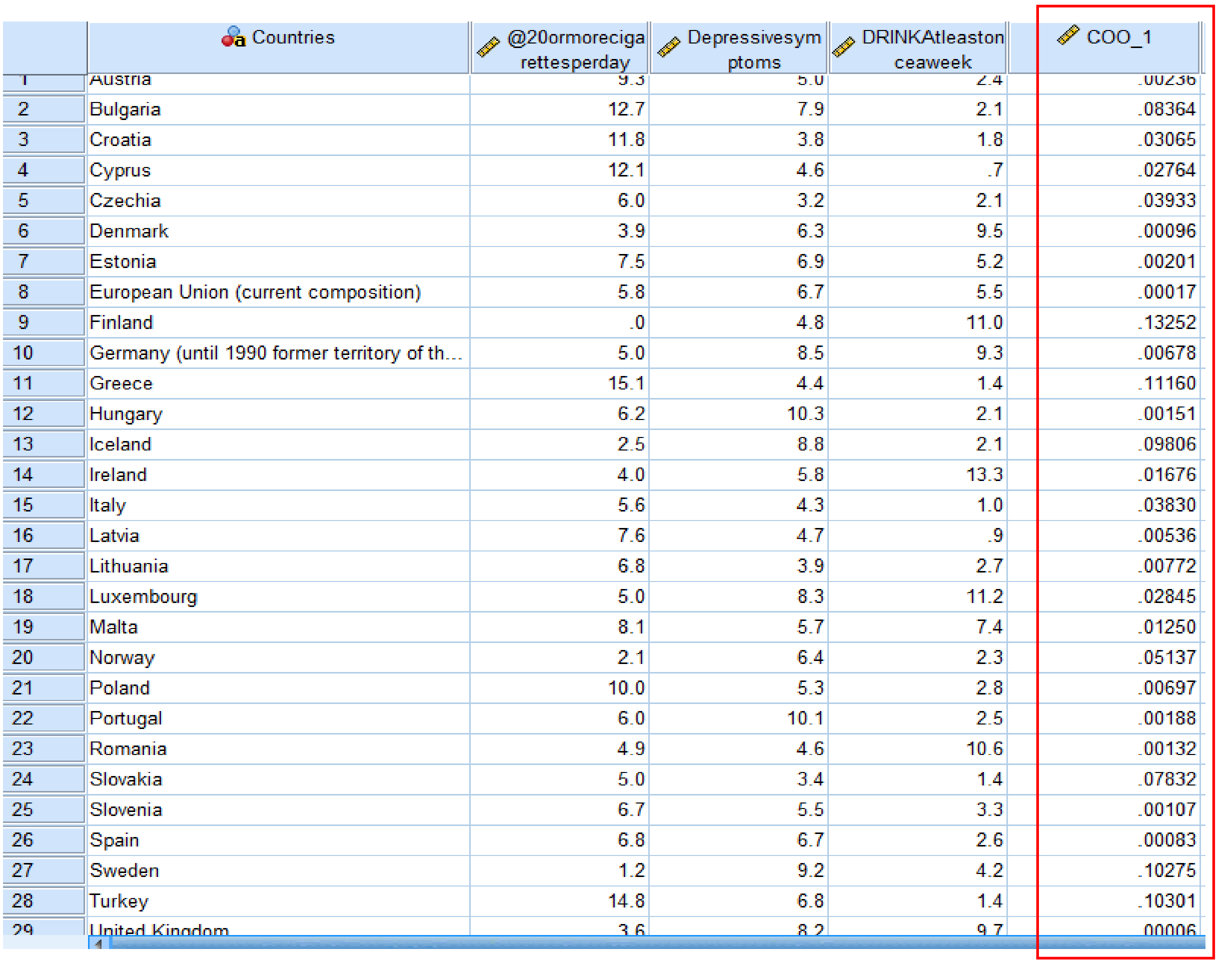
FIG: P-P plot of the model.

To fulfil this assumption, we require the dots to lie closely to the diagonal line and closer the dots are, the more normally distributed is the residuals. Looking at our P-P plot we can say this assumption has been met as most of the dots remain fairly close to the line except for a few points.

7: Checking for significant outliers, high leverage points or influential points which can bias the model.

For this we check the Cooks Distance. This generates a new column in our data labelled as COO_1 containing Cooks Distance Statistic for each participant. For this criteria to be met we check for values in COO_1 over 1 which highlight the significant outliers influencing the model.

Attached below is the screenshot of COO_1 suggesting no such outliers have been found.

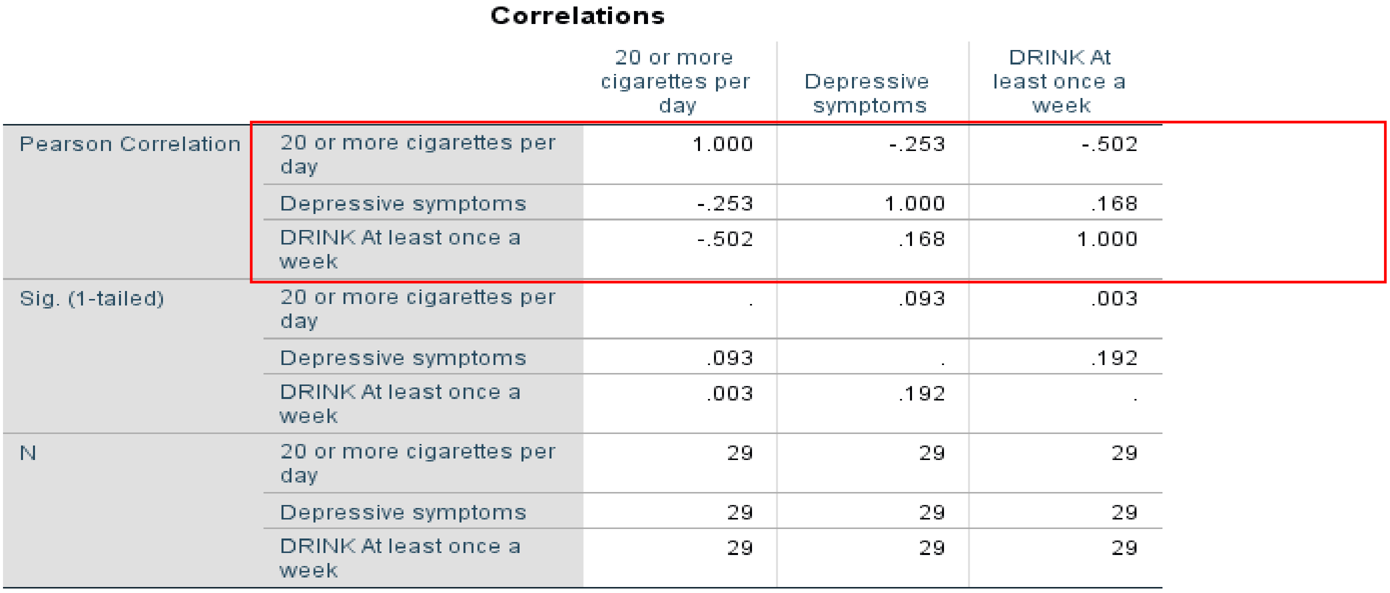


	Countries	@20ormorecigarettesperday	Depressivesymptoms	DRINKAtleastonceaweek	COO_1
1	Austria	9.3	5.0	2.4	.00236
2	Bulgaria	12.7	7.9	2.1	.08364
3	Croatia	11.8	3.8	1.8	.03065
4	Cyprus	12.1	4.6	.7	.02764
5	Czechia	6.0	3.2	2.1	.03933
6	Denmark	3.9	6.3	9.5	.00096
7	Estonia	7.5	6.9	5.2	.00201
8	European Union (current composition)	5.8	6.7	5.5	.00017
9	Finland	.0	4.8	11.0	.13252
10	Germany (until 1990 former territory of th...	5.0	8.5	9.3	.00678
11	Greece	15.1	4.4	1.4	.11160
12	Hungary	6.2	10.3	2.1	.00151
13	Iceland	2.5	8.8	2.1	.09806
14	Ireland	4.0	5.8	13.3	.01676
15	Italy	5.6	4.3	1.0	.03830
16	Latvia	7.6	4.7	.9	.00536
17	Lithuania	6.8	3.9	2.7	.00772
18	Luxembourg	5.0	8.3	11.2	.02845
19	Malta	8.1	5.7	7.4	.01250
20	Norway	2.1	6.4	2.3	.05137
21	Poland	10.0	5.3	2.8	.00697
22	Portugal	6.0	10.1	2.5	.00188
23	Romania	4.9	4.6	10.6	.00132
24	Slovakia	5.0	3.4	1.4	.07832
25	Slovenia	6.7	5.5	3.3	.00107
26	Spain	6.8	6.7	2.6	.00083
27	Sweden	1.2	9.2	4.2	.10275
28	Turkey	14.8	6.8	1.4	.10301
29	United Kingdom	3.6	8.2	9.7	.00006

FIG: Column COO_1 showing Cook's Distance.

INTERPRETATION OF RESULTS:

1. Correlations:



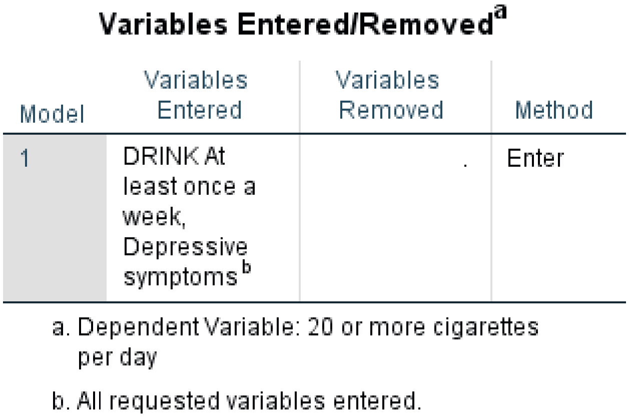
Correlations

		20 or more cigarettes per day	Depressive symptoms	DRINK At least once a week
Pearson Correlation	20 or more cigarettes per day	1.000	-.253	-.502
	Depressive symptoms	-.253	1.000	.168
	DRINK At least once a week	-.502	.168	1.000
Sig. (1-tailed)	20 or more cigarettes per day	.	.093	.003
	Depressive symptoms	.093	.	.192
	DRINK At least once a week	.003	.192	.
N	20 or more cigarettes per day	29	29	29
	Depressive symptoms	29	29	29
	DRINK At least once a week	29	29	29

Fig: Correlation Table with correlation coefficients.

The box shown in red shows various correlations between each of the variables. We have several correlations less than 0.8 hence we can suggest that multiple regression is possible and that the two IV's(predictors) are related to the DV(outcome).

2.**Variables entered/removed:**



Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	DRINK At least once a week, Depressive symptoms[b]	.	Enter

a. Dependent Variable: 20 or more cigarettes per day

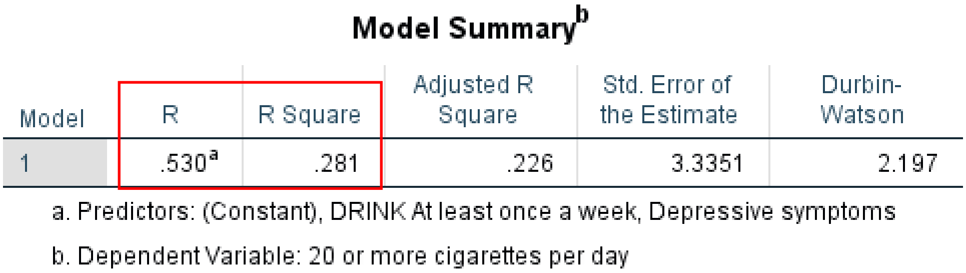
b. All requested variables entered.

Fig: Variables Entered/Removed Table.

This shows us the IV's which acts as predictors for our model which are:

1. Drink at least once a week.
2. Depressive Symptoms.

3. **Model Summary:**



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.530[a]	.281	.226	3.3351	2.197

a. Predictors: (Constant), DRINK At least once a week, Depressive symptoms

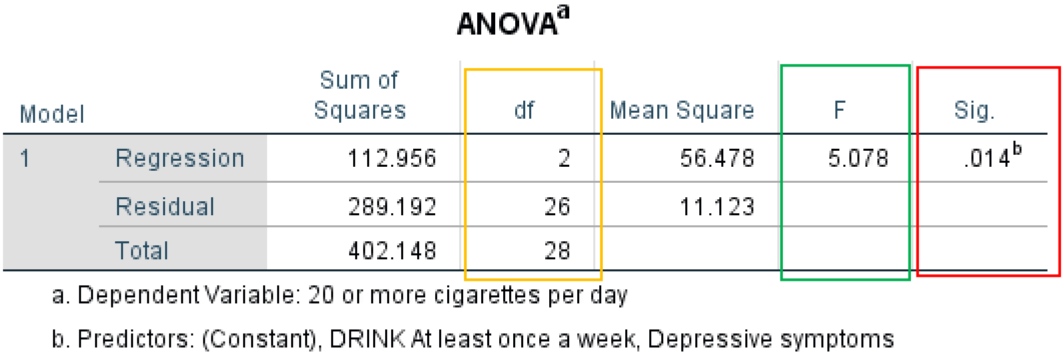
b. Dependent Variable: 20 or more cigarettes per day

Fig: Model Summary.

The Value under R column shows the strength of the relationship between DV and two IV's combined and can also be interpreted as a regular correlation coefficient. We have a **R value** of 0.530 which is a strong relationship and it suggests that our model is a good predictor of the DV.

R Square is an important value which indicates the proportion of variance in the DV that can be explained by the IV's. We have a R square of .281 or we can also say that 28.1 % of variance in data can be explained by the IV's.

4.**ANOVA:**



ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	112.956	2	56.478	5.078	.014[b]
	Residual	289.192	26	11.123		
	Total	402.148	28			

a. Dependent Variable: 20 or more cigarettes per day

b. Predictors: (Constant), DRINK At least once a week, Depressive symptoms

Fig: ANOVA Table.

ANOVA is used to predict if our model is a good predictor of the outcome (depressions symptoms) The significance level is less than p=0.05 hence we can say the multi linear regression model does a good job at predicting the DV.

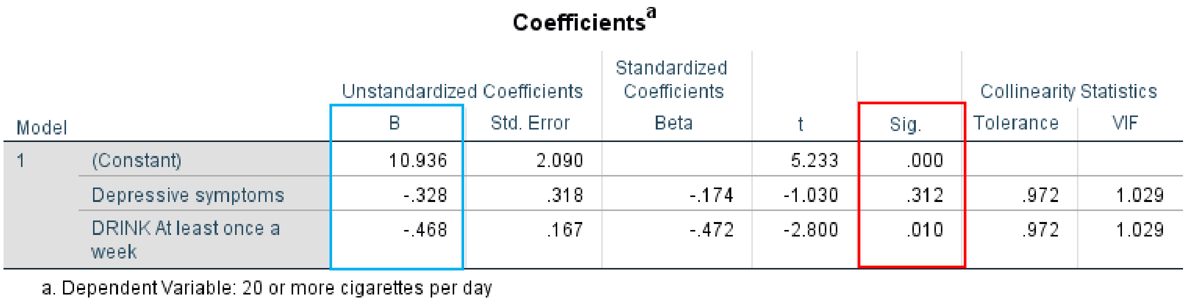
Findings:

Since we now know that the model is good we can report the results.

The results indicated that the model is a good predictor of Depression Symptoms:

F (Regression df, Residual df)= F- Ratio, p = Sig

F (2,26) = 5.078, p= .014.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	10.936	2.090		5.233	.000		
	Depressive symptoms	-.328	.318	-.174	-1.030	.312	.972	1.029
	DRINK At least once a week	-.468	.167	-.472	-2.800	.010	.972	1.029

a. Dependent Variable: 20 or more cigarettes per day

The Coefficients table gives input on the individual contributions of the predictor variables(IV's). The Sig column shows whether the IV's contributed to the model or not.

From our table we can see "Depressive Symptoms" significantly contributed to the model for p = 0.312 and "Drink at least once a week" as well at p = 0.010 has a better contribution.

The B values *unstandardized beta coefficients* shows relationship between DV and both IV's and in turn also gives insight into the effect of each IV on the DV *if the other IV is kept constant*. We have two negative values stating a negative relationship.

As per our *beta coefficients* Depressive Symptoms B0 = -.328, we can state that as Depressive Symptoms increases by 1 unit, cigarette smoking reduces by .328 units and the same relation is for "Drink at least once a week" having B1 as -.468 where "Drink at least once a week" increases by 1 unit, cigarette smoking reduces by .468.

In regression we have a statistical equation:

Y=B0 +B1X1+B2X2

Where Y= outcome variable

X1 =1st predictor variable

X2 = 2nd predictor variable

In our case we can say **20 or more cigarettes a day = B0 + B1Depressive Symptoms +**

B2Drink at least once a week.

Hence our predictive model as per B coefficients would be:

20 or more cigarettes a day = 10.936 + (-.328*Depressive Symptoms) +

(-.468*Drink at least once a week).

Final conclusion from the model:

1. We have a **R value of 0.530** suggesting a strong relationship between IV's and DV, and a **R square of .281** suggesting that 28.1 % of variance in data can be explained by the IV's stating significant levels of prediction for smoking.
2. **Significant prediction: F (2,26) = 5.078, p= .014 The significance level is less than p=0.05** hence the multiple linear regression model does a good job at predicting the DV.
3. **Final predictive model: 20 or more cigarettes a day = 10.936 + (-.328*Depressive Symptoms) + (-.468*Drink at least once a week**)- showing a negative relationship between DV and IV's, ***in other words the level of smoking among people decreases as people show more depression symptoms and increased drinking habits, with drinking habits showing a stronger connection as compared to depression symptoms.***

LOGISTIC REGRESSION:

Objectives of Analysis: To build a logistic regression model for the prediction of an observation belonging to a categorical dependent variables(DV) based on one or more independent variables(IV's). The IV can be either categorical or continuous. The model I will be using will be a binomial/binary logistic regression model as the DV has only two possible outcomes i.e. Male and Female(GENDER).

The model aims to predict if accidental deaths(IV1) and death by tuberculosis(IV2) effect on Males and Females. The DV is Sex in this case with only two possible outcomes hence we will select a binary regression approach.

Data Description: The below shown datasets have been taken from the World Health Organisation. (see appendix).

1. Accidental deaths for the year 2010 to 2012 for 41 countries based on sex.
2. Deaths by tuberculosis for the year 2010 to 2012 for same 41 countries based on sex.

The data was cleaned and transformed using pivot table and Vlookup on Excel. The final data to be studied has a total of 225 rows.

Before proceeding with the model we will check for assumptions to see if the data is compatible with logistic regression.

Assumptions:

1.**DV should be measured on a dichotomous scale i.e. the outcome variable should be mutually exclusive:**

Our dependent variable is categorical variable Sex and it has only two possible outcomes. i.e. either Male or female hence mutually exclusive.

2.**Two or more IV's which can be either continuous or categorical:** In this case we have two continuous IV's.

3.**Large sample size:**

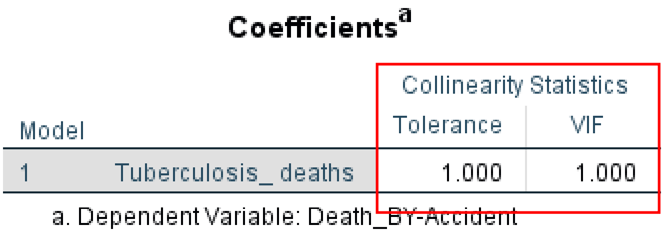
The dataset has 225 observations and is good enough a size to perform logistic regression.

4. **Absence of multi collinearity** i.e. the IV's should not be highly correlated to each other or independent of each other.

Since *logistic regression in SPSS does not provide collinearity diagnostics* for categorical variables we perform *collinearity check in linear regression* for the IV's.

The threshold is that the tolerance should not less than 0.1 and VIF values should not be greater than 10. As per the output in we have tolerance and VIF values of 1 hence suggesting no multi collinearity or to say that the IV's are not highly correlated to each other.

Page 914, Andy Field - Discovering Statistics using IBM Statistics (5th Edition).



Coefficients[a]

Model		Collinearity Statistics	
		Tolerance	VIF
1	Tuberculosis_ deaths	1.000	1.000

a. Dependent Variable: Death_BY-Accident

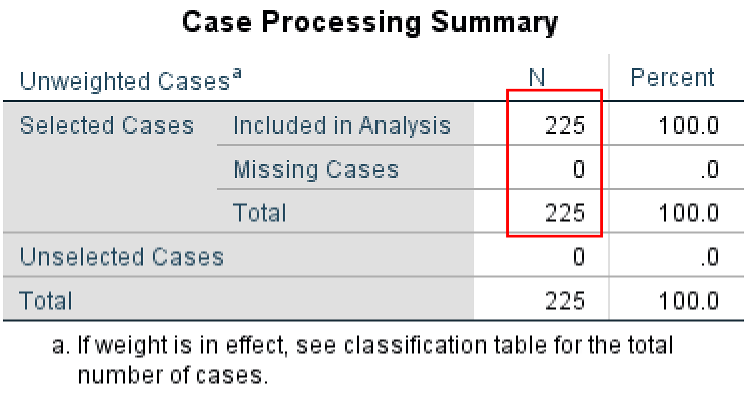
Fig: Absence of multi collinearity among IV's.

INTERPRETATION OF RESULTS

1.**Case processing summary:**

This shows the sample size N and the number of missing values in the data.

Below is the case processing summary showing sample size of 225 and 0 missing cases.This also shows the assumption is met for having large sample sizes.



Case Processing Summary

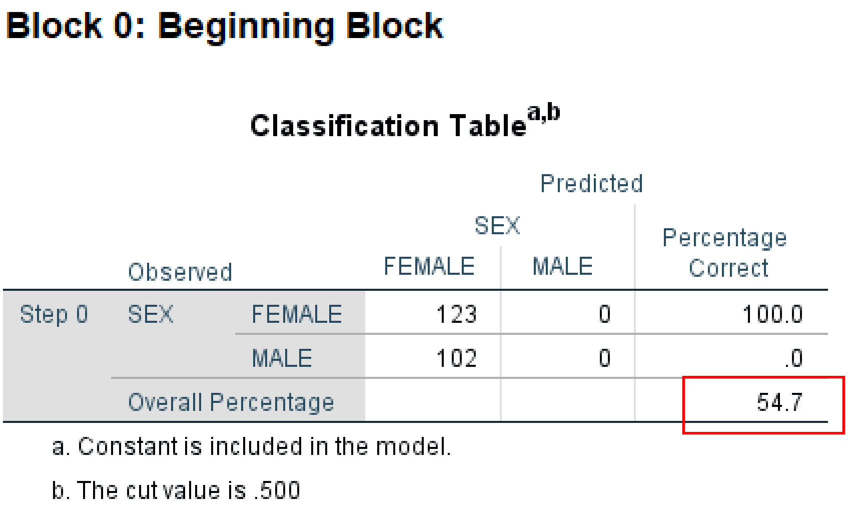
Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	225	100.0
	Missing Cases	0	.0
	Total	225	100.0
Unselected Cases		0	.0
Total		225	100.0

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

Fig: Case processing summary showing sample size and missing values.

2.T**o see of the model is a good fit or not:** Shown below are various ways to understand if the model is a good fit or not.

A. Block 0: Beginning block shown below is similar to a null hypothesis and shows that out of 225 observations 123 females and 102 males would be the ratio and overall model predictability will be a correct prediction 54.7% of the time **without any IV's involved**. A good model with IV's will be able to increase the above percentage.



Block 0: Beginning Block

Classification Table[a,b]

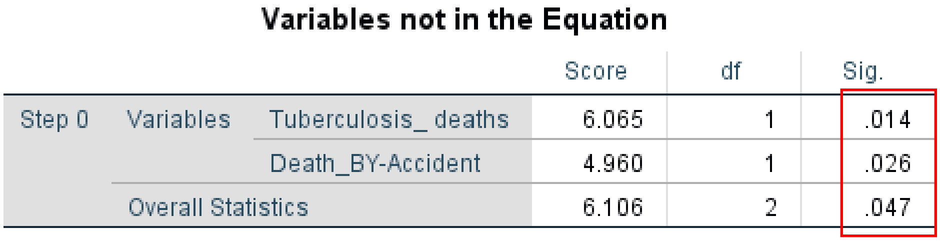
Observed			Predicted SEX FEMALE	Predicted SEX MALE	Predicted Percentage Correct
Step 0	SEX	FEMALE	123	0	100.0
		MALE	102	0	.0
	Overall Percentage				54.7

a. Constant is included in the model.

b. The cut value is .500

Fig: Block 0: Beginning block.

B. The IV's significance below 0.5 suggests a good p value and our model shows that the IV's individually are good predictors of the DV (i.e. p value < 0.5) as shown below.

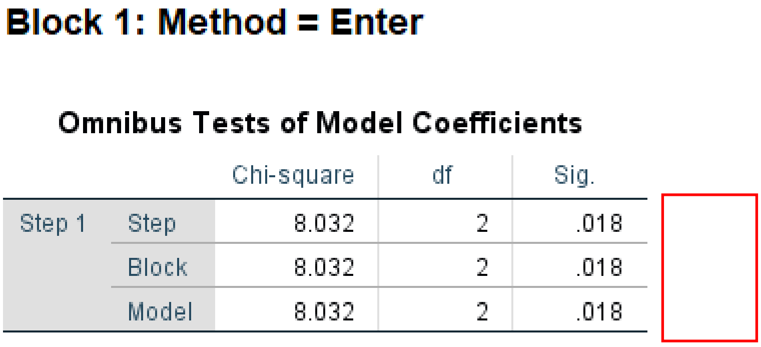


Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Tuberculosis_ deaths	6.065	1	.014
		Death_BY-Accident	4.960	1	.026
	Overall Statistics		6.106	2	.047

Fig: Significance of IV's

C. Omnibus test of model coefficients adds the predictor values to the model and compares it the null hypothesis from earlier and a significance less than 0.5 shows a good predictor model. **In this case p value is .018 as shown below.**



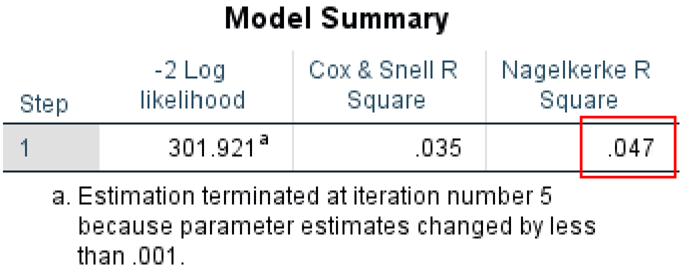
Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	8.032	2	.018
	Block	8.032	2	.018
	Model	8.032	2	.018

Fig: Omnibus test of model coefficients showing final p values.

D. Model summary shows below the Nagelkerke R Square similar to the one in linear regression showing amount of variance in DV explained by the IV's. In this case 4.7% of variance is predicted by IV's.



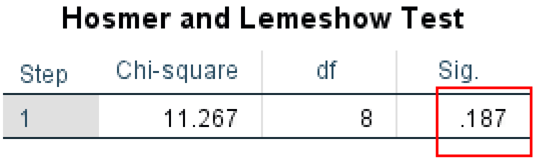
Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	301.921[a]	.035	.047

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Fig: Nagelkerke R square.

E. Hosmer and Lemeshow Test – Also gives an idea of how good the model is i.e the p value should be **greater than 0.5** to be significantly good model. In this case we **p value of .187** as shown below.

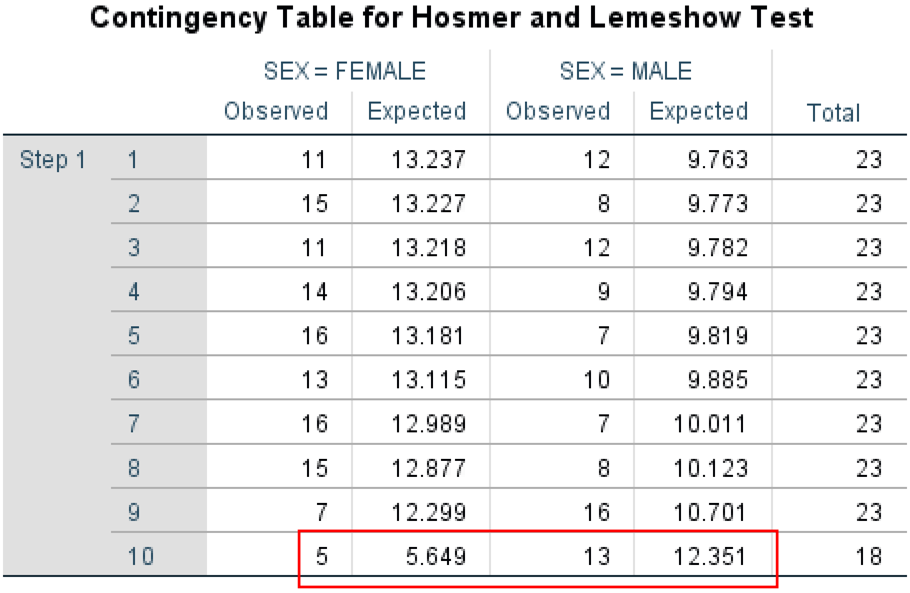


Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	11.267	8	.187

Fig: Hosmer and Lemeshow test.

F. The **contingency table** related to Hosmer and Lemeshow Test also again shows how well model is at predicting outcomes. This breaks the outcomes into groups and progressively tries to fit the predictors in the model. *We look at the last row of the table and closer the difference between the observed and expected is the better the model is*.



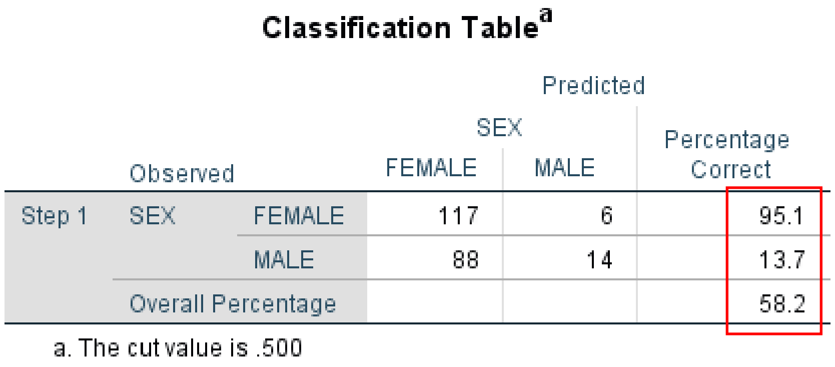
Contingency Table for Hosmer and Lemeshow Test

		SEX = FEMALE		SEX = MALE		
		Observed	Expected	Observed	Expected	Total
Step 1	1	11	13.237	12	9.763	23
	2	15	13.227	8	9.773	23
	3	11	13.218	12	9.782	23
	4	14	13.206	9	9.794	23
	5	16	13.181	7	9.819	23
	6	13	13.115	10	9.885	23
	7	16	12.989	7	10.011	23
	8	15	12.877	8	10.123	23
	9	7	12.299	16	10.701	23
	10	5	5.649	13	12.351	18

Fig: Contingency table showing prediction outcomes.

In this case we have that out of observed number of males in a group of subjects was 13 and the model was able to predict 12.351 of those being male showing it to be a good model.

G. Classification table shows how good model is at predicting actual outcomes. This model was able to correctly predict **58.2%** of the outcomes showing it does a decent job as compared to earlier null hypothesis of **54.7%.** It does a better job at predicting females compared to males**.** *A good model would be having more than 65% hence we can say our model does a decent job.*



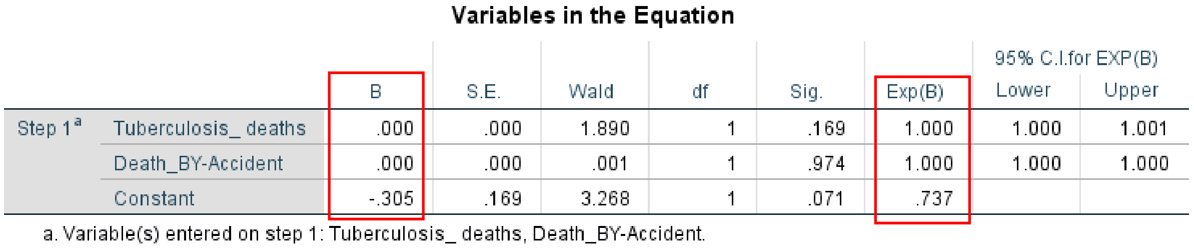
Classification Table[a]

Observed			Predicted SEX FEMALE	Predicted SEX MALE	Percentage Correct
Step 1	SEX	FEMALE	117	6	95.1
		MALE	88	14	13.7
	Overall Percentage				58.2

a. The cut value is .500

Fig: Classification table showing prediction percentage of complete model.

H. Variations in the Equation table is outcome of **Wald test** to shows the beta coefficients that can be plugged into regression equation and the odds ratios which can be interpreted as *"the higher the odds ratio over 1 the more the prediction of outcome would be".* In this case it shows as odds ratio as 1.00 which has been rounded off by SPSS.



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
Step 1[a]	Tuberculosis_ deaths	.000	.000	1.890	1	.169	1.000	1.000	1.001
	Death_BY-Accident	.000	.000	.001	1	.974	1.000	1.000	1.000
	Constant	-.305	.169	3.268	1	.071	.737		

a. Variable(s) entered on step 1: Tuberculosis_ deaths, Death_BY-Accident.

Fig: Table showing beta coefficients and odds ratio.

Final conclusion from the model:

A logistic regression was performed to predict the gender of a person dying due to tuberculosis and accidental death. The model was statistically significant We have a Nagelkerke **R square of .047** suggesting that **4.7 %** of variance in outcomes can be explained by the IV's stating significant levels of prediction for gender which is not a good score and concluding that gender cannot be predicted from death records due to tuberculosis and accidents. There is variability which cannot be explained well by the IV's.

The beta coefficients are rounded of to **.000** by SPSS and odds ratio is also rounded off to **1.000** by SPSS which are less than expected and again show the model does not do so well. Hosmer and Lemeshow test shows a p value of **.187** which is greater than the threshold of .05 showing it as a good model. This model was able to correctly predict **58.2%** of the outcomes as compared to prediction of 54.7% in the hypothesis model without predictors which is not much of a difference.

Appendix:

Multiple linear regression:
All data taken from Eurostat's:

https://ec.europa.eu/eurostat/data/database?p_p_id=NavTreeportletprod_WAR_NavTreeportletprod_INSTANCE_nPqeVbPXRmWQ&p_p_lifecycle=0&p_p_state=normal&p_p_mode=view&p_p_col_id=column-2&p_p_col_pos=1&p_p_col_count=2

http://appsso.eurostat.ec.europa.eu/nui/show.do?query=BOOKMARK_DS-735690_QID_6AD3CEA3_UID_-3F171EB0&layout=HLTH_PB,L,X,0;GEO,L,Y,0;UNIT,L,Z,0;ISCED11,L,Z,1;TIME,C,Z,2;SEX,L,Z,3;AGE,L,Z,4;INDICATORS,C,Z,5;&zSelection=DS-735690UNIT,PC;DS-735690INDICATORS,OBS_FLAG;DS-735690AGE,TOTAL;DS-735690ISCED11,TOTAL;DS-735690TIME,2014;DS-735690SEX,T;&rankName1=TIME_1_0_-1_2&rankName2=ISCED11_1_2_-1_2&rankName3=UNIT_1_2_-1_2&rankName4=AGE_1_2_-1_2&rankName5=INDICATORS_1_2_-1_2&rankName6=SEX_1_2_-1_2&rankName7=HLTH-PB_1_2_0_0&rankName8=GEO_1_2_0_1&rStp=&cStp=&rDCh=&cDCh=&rDM=true&cDM=true&footnes=false&empty=false&wai=false&time_mode=ROLLING&time_most_recent=false&lang=EN&cfo=%23%23%23%2C%23%23%23.%23%23%23&lang=en

Logistic regression:

All data taken from the World Health Organization”

https://gateway.euro.who.int/en/indicators/hfamdb_60-deaths-accidents/ - Death by accident.

https://gateway.euro.who.int/en/indicators/hfamdb_799-deaths-tuberculosis/ - Death by tuberculosis.

References:

www.open.ac.uk SPSS- Tutorials

URL- http://www.open.ac.uk/socialsciences/spsstutorial/files/tutorials/assumptions.pdf