Statistics for Data Analytics

CA2

Multiple Regression and Logistic Regression

MSc Data Analytics B

X17165849

Kaushik Rajan

Multiple Regression

Equation of multiple regression:

$$\hat{Y} = a + b_1X_1 + b_2X_2 \ldots\ldots + b_kX_k$$

Were Y is the independent variable and X1, X2,…Xk are the dependent variables.

Data Source:

URL :
https://webarchive.nationalarchives.gov.uk/20161021160913/http://www.apho.org.uk/default.aspx?QN=HP_DATATABLES

For this Project, multiple datasets were downloaded from the website mentioned above. The datasets that were downloaded were Lifeexpectancymale, Lifeexpectancyfemale, Adultswhosmoke, Bingedrinkingadults, Healthyeatingadults and Physicallyactiveadults. These datasets were then combined using the area name as common column. The data cleaning part was done on excel and SPSS.

The motive of this project is to check if life expectancy can be predicted using the Independent variables Adultswhosmoke, Bingedrinkingadults, Healthyeatingadults and Physicallyactiveadults.

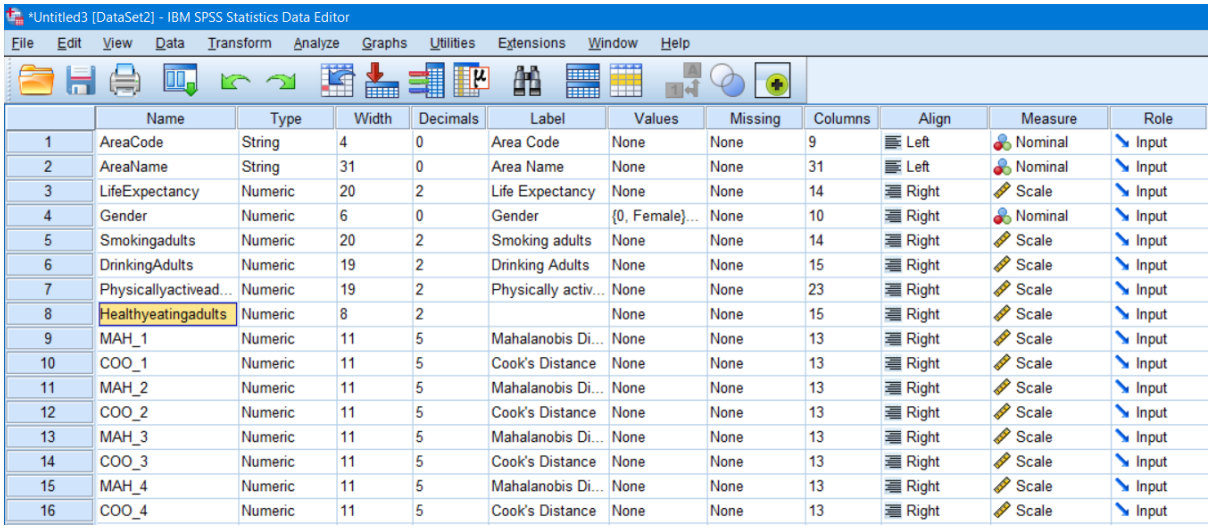
Requirements of a multiple regression: 1 dependent variable which is the life expectancy and 2 or more independent variables which are Adultswhosmoke, Bingedrinkingadults, Healthyeatingadults and Physicallyactiveadults.

Assumptions of Multiple Regression:

- Linear Relationship: There must be a linear relationship between the Dependent and independent variables.
- Normality: Multiple Regression assumes that the residuals are normally distributed
- Homoscedasticity: Multiple Regression assumes that the variance of error terms are similar across the values of the independent variables
- Multicollinearity: The model assumes that there is no high correlation between the independent variables
- Sample Size: More the number of sample size, better the prediction will be.

Explanation of the Dataset:

Columns of the datasets are AreaCode – contains the code of each area(Type – String), AreaName – Contains the area name(Type – String), LifeExpectancy – contains the average life expectancy of each area(Type – Numeric), Gender – 0 for Female and 1 for male (Type – Numeric), Smokingadults – Average of adults who smoke (Type Numeric), DrinkingAdults – Average of adults who drink (Type – Numeric), Physicallyactiveadults – Average of adults who are physically active (Type – Numeric), Healthyeatingadults – Average of adults who eat healthy (Type – Numeric).



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	AreaCode	String	4	0	Area Code	None	None	9	Left	Nominal	Input
2	AreaName	String	31	0	Area Name	None	None	31	Left	Nominal	Input
3	LifeExpectancy	Numeric	20	2	Life Expectancy	None	None	14	Right	Scale	Input
4	Gender	Numeric	6	0	Gender	{0, Female}...	None	10	Right	Nominal	Input
5	Smokingadults	Numeric	20	2	Smoking adults	None	None	14	Right	Scale	Input
6	DrinkingAdults	Numeric	19	2	Drinking Adults	None	None	15	Right	Scale	Input
7	Physicallyactivead...	Numeric	19	2	Physically activ...	None	None	23	Right	Scale	Input
8	Healthyeatingadults	Numeric	8	2		None	None	15	Right	Scale	Input
9	MAH_1	Numeric	11	5	Mahalanobis Di...	None	None	13	Right	Scale	Input
10	COO_1	Numeric	11	5	Cook's Distance	None	None	13	Right	Scale	Input
11	MAH_2	Numeric	11	5	Mahalanobis Di...	None	None	13	Right	Scale	Input
12	COO_2	Numeric	11	5	Cook's Distance	None	None	13	Right	Scale	Input
13	MAH_3	Numeric	11	5	Mahalanobis Di...	None	None	13	Right	Scale	Input
14	COO_3	Numeric	11	5	Cook's Distance	None	None	13	Right	Scale	Input
15	MAH_4	Numeric	11	5	Mahalanobis Di...	None	None	13	Right	Scale	Input
16	COO_4	Numeric	11	5	Cook's Distance	None	None	13	Right	Scale	Input

Questions to answer:

1) How well the independent variables predict the Dependent variable?
2) Which is the best predictor of the Life expectancy?

Steps to do Multiple Regression on SPSS:

Analyze -> Regression -> Linear -> Select dependent variable and move it to the dependent box -> Move the independent variables to the independent box -> Click Statistics box and select Estimates, Confidence interval, Model fit, Descriptives, Part and partial correlatons and collinearity diagnostics -> tick casewise diagnostics and outliers outside 3 Standard Deviation in the residual section -> click ok -> click on plots and move ZRESID and move it to the Y Box and ZPRED to the X Box and clock normal probability plot option in Standardized Residual plots and click on continue -> Click Ok

Interpretation of the output: Step 1: Checking the assumptions

Multicollinearity:

Correlations

		Life Expectancy	Gender	Smoking adults	Drinking Adults	Physically active adults	Healthyeating adults
Pearson Correlation	Life Expectancy	1.000	.857	-.483	-.269	.430	.331
	Gender	.857	1.000	-.240	-.111	.309	.106
	Smoking adults	-.483	-.240	1.000	.343	-.416	-.594
	Drinking Adults	-.269	-.111	.343	1.000	.109	-.441
	Physically active adults	.430	.309	-.416	.109	1.000	.309
	Healthyeatingadults	.331	.106	-.594	-.441	.309	1.000
Sig. (1-tailed)	Life Expectancy	.	.000	.000	.000	.000	.000
	Gender	.000	.	.000	.017	.000	.022
	Smoking adults	.000	.000	.	.000	.000	.000
	Drinking Adults	.000	.017	.000	.	.019	.000
	Physically active adults	.000	.000	.000	.019	.	.000
	Healthyeatingadults	.000	.022	.000	.000	.000	.
N	Life Expectancy	363	363	363	363	363	363
	Gender	363	363	363	363	363	363
	Smoking adults	363	363	363	363	363	363
	Drinking Adults	363	363	363	363	363	363
	Physically active adults	363	363	363	363	363	363
	Healthyeatingadults	363	363	363	363	363	363

For Multicollinearity, Correlations tab is checked. From the output, it can be inferred that all the Independent variables except Drinking adults show some relationship with the Dependent variable Life Expectancy. This is inferred by looking at the Pearson correlation tab. The values are .857, -.483, -.269, .430 and .331 for the independent variables Gender, smoking adults, drinking adults, physically active adults and healthy eating adults respectively. The correlation between drinking adults and life expectancy is at -.269 which is very less and which means that drinking adults does not have much effect on the values of life expectancy.

The correlation between the independent variables are not too high, all are under 0.7 as observed from the correlation table and therefore all the independent variables are retained. Another way to check for multicollinearity is to check the Coefficients tab.

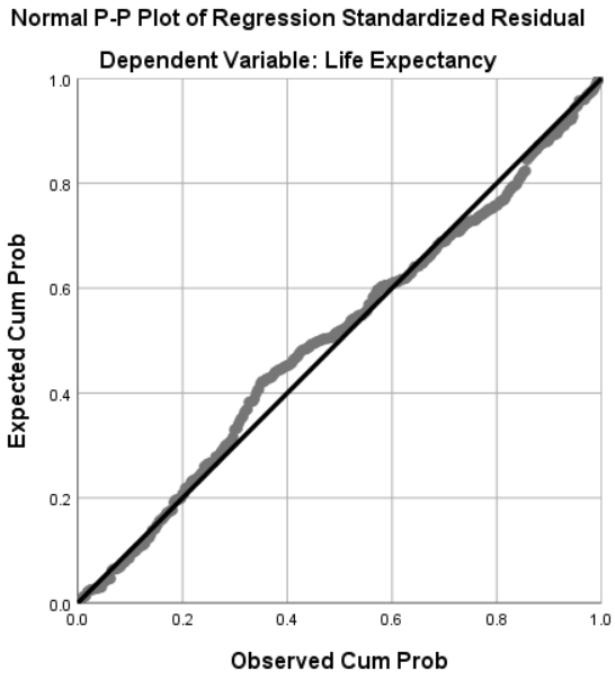
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	78.783	.928		84.908	.000	76.959	80.608					
	Gender	4.354	.131	.761	33.133	.000	4.095	4.612	.857	.869	.708	.867	1.154
	Smoking adults	-.111	.018	-.179	-6.211	.000	-.146	-.076	-.483	-.312	-.133	.552	1.811
	Drinking Adults	-.070	.017	-.107	-4.173	.000	-.103	-.037	-.269	-.216	-.089	.689	1.451
	Physically active adults	.162	.038	.113	4.314	.000	.088	.236	.430	.223	.092	.671	1.490
	Healthyeatingadults	.039	.018	.063	2.187	.029	.004	.074	.331	.115	.047	.554	1.806

a. Dependent Variable: Life Expectancy

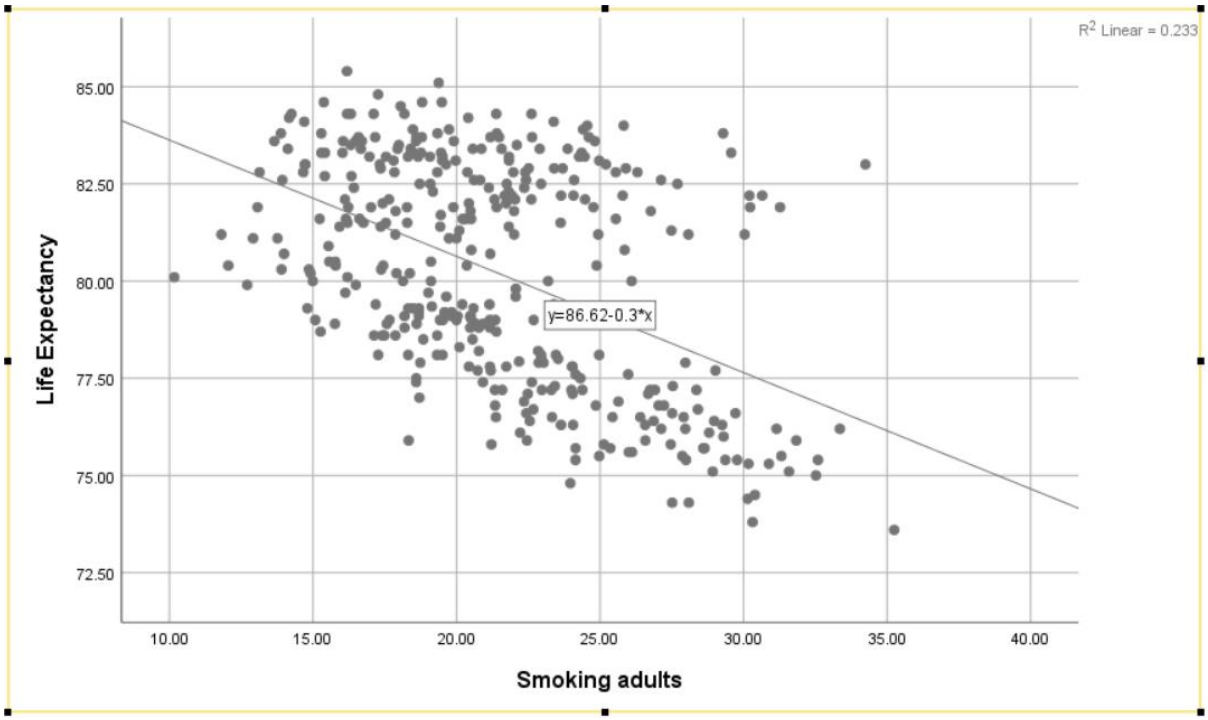
The Tolerance value and VIF are the 2 other indicators of multicollinearity. From the table, it can be observed that all the tolerance values are greater than 0.1 and all the VIF values are less than 10 and hence, it can be inferred that there is no violation of multicollinearity assumption.

Outliers, Normality, independence of residuals:

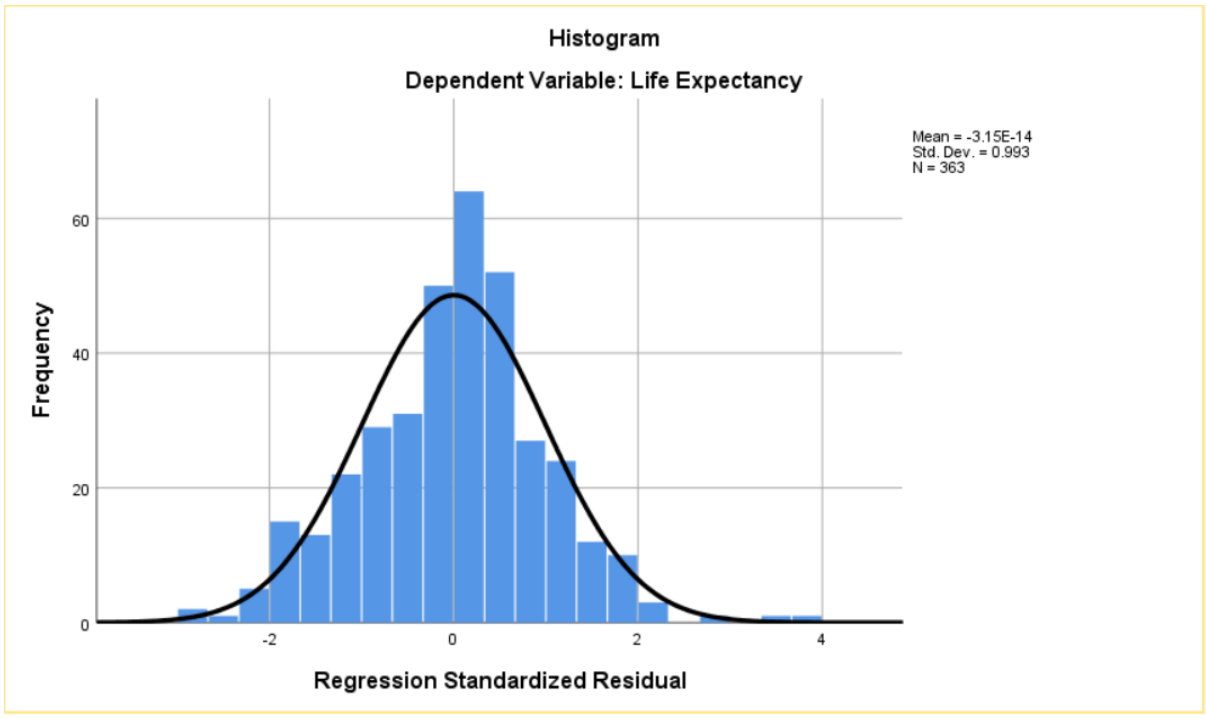


From the P-P plot, it can be inferred that the residual points lie in a reasonably straight diagonal line which suggests that there are no major deviations from the normality. From the scatterplot, the outliers can be viewed. There are very few outliers present and hence can be ignored.

Linearity:



Linearity is checked using the Scatterplot. The scatterplot above is to check the linearity between life expectancy and smoking adults. Similarly, all the combinations can be checked for linearity using the scatterplot. It iss observed that all the combinations are linear.



From the above image, it can be inferred that the residual is normally distributed.

Step 2: Evaluating the Model

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.915[a]	.837	.835	1.16526	.837	366.129	5	357	.000

a. Predictors: (Constant), Healthyeatingadults, Gender, Physically active adults, Drinking Adults, Smoking adults

b. Dependent Variable: Life Expectancy

The R Square tab in the model summary explains how much the variance in the dependent variable is explained by the model. Here, as inferred from the model summary, the model explains 83% of the variance of the dependent variable, i.e. the life expectancy. The adjusted R Square tab is to explain for the population.

ANOVA:

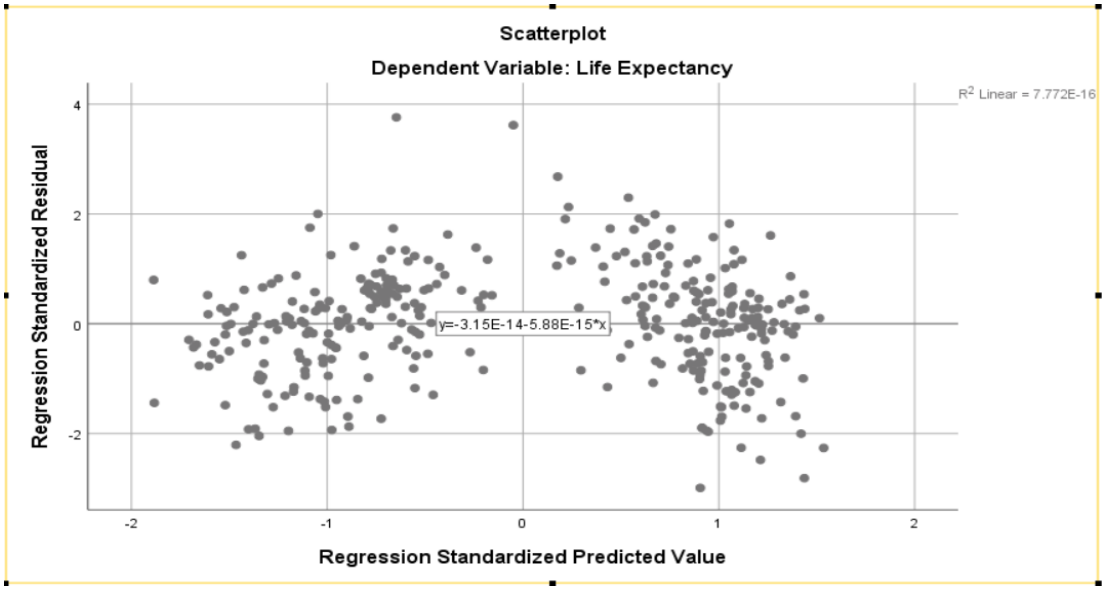
ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2485.705	5	497.141	366.129	.000[b]
	Residual	484.745	357	1.358		
	Total	2970.451	362			

a. Dependent Variable: Life Expectancy

b. Predictors: (Constant), Healthyeatingadults, Gender, Physically active adults, Drinking Adults, Smoking adults

The significance value of 0.000 shows that its statistically significant and the null hypothesis can be rejected. The F- test checks if the variance explained is significantly greater than the error within the model.



Healthy residual plot

Step 3: Evaluating the independent variables

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	78.783	.928		84.908	.000	76.959	80.608					
	Gender	4.354	.131	.761	33.133	.000	4.095	4.612	.857	.869	.708	.867	1.154
	Smoking adults	-.111	.018	-.179	-6.211	.000	-.146	-.076	-.483	-.312	-.133	.552	1.811
	Drinking Adults	-.070	.017	-.107	-4.173	.000	-.103	-.037	-.269	-.216	-.089	.689	1.451
	Physically active adults	.162	.038	.113	4.314	.000	.088	.236	.430	.223	.092	.671	1.490
	Healthyeatingadults	.039	.018	.063	2.187	.029	.004	.074	.331	.115	.047	.554	1.806

a. Dependent Variable: Life Expectancy

The B value under the unstandardized Coefficients explains the degree each independent variable affects the output. This means that when smoking adults decreases by 0.111 the life expectancy increases. Similarly, when Physically active adults increase by 0.162, the life expectancy increases.

The Beta value under the standardized coefficients explains the degree each independent variable affects the output but expressed as standard deviations. It can be inferred from the Sig column that all the variables are statistically significant (sig < 0.05), which essentially means that all the independent variables make significant contributions. Healthyeatingadults contribute less since the B value is less.

Hypothesis test:

H0 = b1 = b2 = b3 = bk = 0

H1 = Not all b are equal to 0

Here, for this project, from the coefficients tab, it can be inferred that the significant values of gender, smoking adults, drinking adults, physically active adults, healthyeatingadults are 0.000,0.000,0.000,0.000 and 0.029 respectively, which are all statistically significant (p < 0.05) and hence we reject H0 and we can infer that not all b are equal to 0.

Standardized Residual Statistics: The Standardized Residual Statistics values are generated.

RES_1
.07625
.25074
-.99429
-.21559
.06257
-.50489
.76948
-.34175
.81238
.82507
.79022
.96554
1.20382

95% of the Standardized residuals lies between + or – 2 and 99% of the standardized residuals lie between + or – 2.5

The Multiple Regression equation is given below:

Y = 78.783 + 4.353X1 – 0.111X2 – 0.070X3 + 0.162X4 + 0.039X5

Where Y is the Dependent variable and X1, X2, X3, X4 and X5 are the independent variables.

Result – Multiple Regression was used to predict the value of dependent variable Life expectancy using multiple independent value and how much each independent value effects the change in value of dependent variable. Preliminary analysis of normality, Linearity, multicollinearity and homoscedasticity were made in order to check the assumptions initially made. Adjusted R Square value of 83% shows how good the model is. From the multiple Linear regression, it can be inferred that, when smoking adults decreases by 0.111 the life expectancy increases. Similarly, when Physically active adults increase by 0.162, the life expectancy increases.

Logistic Regression

For Logistic Regression, the same dataset has been used. The motive of the project is to predict the gender using the continuous variables LifeExpectancy, Smokingadults, DrinkingAdults, Physicallyactiveadults and Healthyeatingadults. Gender is a categorical variable where 0 is male and 1 is female. Binomial Logistic Regression is use since categorical variable is used as the dependent variable.

Equation for Logistic Regression

$$E(Y) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}}$$

Assumptions

- Sample size: Sample size must be on the higher side in order to process logistic regression.
- Multicollinearity: The model assumes that there is no high correlation between the independent variables.
- Outliers: Outliers should be identified using inspecting the residuals.

Dataset Information: Columns of the datasets are AreaCode – contains the code of each area(Type – String), AreaName – Contains the area name(Type – String), LifeExpectancy – contains the average life expectancy of each area(Type – Numeric), Gender – 0 for Female and 1 for male (Type – Numeric), Smokingadults – Average of adults who smoke (Type Numeric), DrinkingAdults – Average of adults who drink (Type – Numeric), Physicallyactiveadults – Average of adults who are physically active (Type – Numeric), Healthyeatingadults – Average of adults who eat healthy (Type – Numeric).

Out of the above-mentioned variables, Gender is the dependent variable which is predicted in the project. LifeExpectancy, Smokingadults, DrinkingAdults, Physicallyactiveadults, Healthyeatingadults are the independent variables which are used as the predictor variable, using which the model is built.
Gender has categorical values as 0 – Male and 1 – Female.

Procedure in SPSS: Analyze -> regression -> Binary Logistic -> Move the dependent into dependent box and the independent variables into the Covariates box -> Options -> Select Classification Plots, Hosmer-Lemeshow, goodness of fit, case wise listing of residuals and CI for Exp(B) -> Ok

Interpretation of the output:

Case Processing Summary

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	363	36.5
	Missing Cases	631	63.5
	Total	994	100.0
Unselected Cases		0	.0
Total		994	100.0

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

The above tables show the number of expected cases.

Dependent Variable Encoding

Original Value	Internal Value
Female	0
Male	1

Block 0 – Beginning Block

Classification Table[a,b]

			Predicted		
			Gender		Percentage Correct
	Observed		Female	Male	
Step 0	Gender	Female	0	180	.0
		Male	0	183	100.0
	Overall Percentage				50.4

a. Constant is included in the model.

b. The cut value is .500

Block 0 is the results produced without using any of the independent variables. This is used as the baseline for future reference. In the classification table, the total percentage of classified cases is at 50.4%. This shows that most of the gender will be Male. To improve the accuracy Block 1 is used.

Block 1:

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	450.685	5	.000
	Block	450.685	5	.000
	Model	450.685	5	.000

This is the block where the model is tested. To test how the model performs, Omnibus Tests of model Coefficients is used. In Block 1 all the independent variables are used. As observed, the Significance value is 0.000 (<0.05) which proves that the model is better.

Hosmer and Lemeshow Test

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	1.009	8	.998

To further prove the model, Hosmer and Lemeshow test is used. As per this test, if the Significance value is less than 0.05, it means that the model is poor fit. But the significance value observed for the test run for this model has significance level of 0.998 which is greater than 0.05 which further supports the model.

Model Summary

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	52.515[a]	.711	.948

a. Estimation terminated at iteration number 10 because parameter estimates changed by less than .001.

The Cox and Snell R square and Nagelkerke R Square are similar to R2 measure in multiple regression. They provide an indication of the amount of variation in the dependent variable explained by the model. The values are 0.711 and 0.948 which suggests that 71.1 percent and 94.8 percent of the variables of gender are explain by the model.

Classification Table

Classification Table[a]

Observed			Predicted		
			Gender		Percentage Correct
			Female	Male	
Step 1	Gender	Female	174	6	96.7
		Male	7	176	96.2
	Overall Percentage				96.4

a. The cut value is .500

This table provides the information on how well the model predicts and it can be inferred from the table that the model has predicted right 96.4% of the time. This table can be compared with the classification table of Block 0. On comparing, the overall percentage of the classification table in block 0 predicted right only 50.4 % of the time while, the classification table of Block 1 predicted 96.4 of the time.

Variables in the Equation

H0 -> There is no relation between the dependent variable Y and the independent variable X

H1 -> Dependent variable and independent variables are related.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Life Expectancy	3.444	.623	30.586	1	.000	31.307	9.238	106.093
	Smoking adults	-.081	.115	.493	1	.483	.923	.737	1.155
	Drinking Adults	.237	.117	4.062	1	.044	1.267	1.007	1.595
	Physically active adults	-.168	.243	.478	1	.489	.845	.525	1.361
	Healthyeatingadults	-.673	.166	16.548	1	.000	.510	.369	.705
	Constant	-257.341	46.292	30.903	1	.000	.000		

a. Variable(s) entered on step 1: Life Expectancy, Smoking adults, Drinking Adults, Physically active adults, Healthyeatingadults.

This depicts the contribution of each variable towards the dependent variable. The variables which has significance value of lesser than 0.05 contribute significantly to the predictive ability of the model. This is also known as the Wald test and the Wald value is similar to t-statistic in Regression. Here, in this model, the variables Life expectancy, Drinking Adults, and Healthyeatingadults contribute the most since they are statistically significant ($P < 0.05$). The B values are similar to B values in the multiple regression and this shows the direction of relationship for each variable.

Here, the H0 is rejected.

The values in Exp(B) are odd ratios for each independent variable. This shows that the odd of a male person having a high life expectancy will be 31 % more than the other odds.

Results

Logistic Regression was performed to predict the gender using multiple independent variables such as LifeExpectancy, Smokingadults, DrinkingAdults, Physicallyactiveadults, Healthyeatingadults. The chi square value was statistically significant and hence indicating that the model was able to distinguish. The model explained 71.1 percent (Cox and Snell R square) and 94.8 percent (Nagelkerke R Square) of the variance and classified 96.4% of the cases. The table above shows that only 3 of the 5 independent variables made significant contribution to the model.

References

https://myadm2014.files.wordpress.com/2017/02/spss-survival-manual-a-step-by-step-guide-to-data-analysis-using-spss-for-windows-3rd-edition-aug-2007-2.pdf

http://www.biostathandbook.com/multiplelogistic.html

https://www.youtube.com/watch?v=d2lSw1WUQ8I&list=PLJy0LHDLpgHHUxFGxrqdeCXPyTR2Q57pG