

Student ID: 17169836

Multiple Regression






1. Objective

To build a multiple regression model that predict 'life expectancy at birth in years' using 'current health expenditure per capita in us dollars', 'percentage of population using least basic drinking water' and 'International Health Regulations monitoring framework's legislation score'.

2. Dataset

- Variables used in build this model are accessed from world health organisation website. Both dependent and independent variables in this model are continuous.
- 'Life expectancy at birth' in this model is accessed from <http://apps.who.int/gho/data/view.main.SDG2016LEXv?lang=en> which provided 'life expectancy rate' of countries for multiple years out of which data of 35 countries for the year 2015 is used in this model.
- 'Current health expenditure (CHE) per capita in us dollars' in this model is accessed from <http://apps.who.int/gho/data/view.main.GHEDCHEpcUSSHA2011v> which provided current health expenditure per capita of 35 countries for the year 2015.
- 'Percentage of population using least basic drinking water' in this model is accessed from <http://apps.who.int/gho/data/node.main.WSHWATER?lang=en> which provided percentage of population using least basic drinking water of 35 countries for the year 2015.
- 'International Health Regulations monitoring framework's legislation score' in this model is accessed from <http://apps.who.int/gho/data/view.main.GHEDCHEpcUSSHA2011v> which provided legislation score of 35 countries for the year 2015.

Sample rows:

	 Country	 Life expectancy at birth years	 Current health expenditure CHE per capita in US\$	 Population using at least basic drinking water service	 Legislation
1	Afghanistan	63.2000000000...	60.1000000000...	71.0	0
2	Australia	82.6000000000...	4934.00000000...	100.0	100
3	Austria	81.4000000000...	4536.10000000...	100.0	75
4	Bangladesh	72.2000000000...	31.8000000000...	97.5	100
5	Belgium	80.9000000000...	4228.30000000...	100.0	100
6	Bhutan	70.2000000000...	91.1000000000...	97.5	75
7	Brunei Darussalam	76.2000000000...	812.2000000000...	99.5	100
8	China	76.2000000000...	425.6000000000...	96.0	100
9	Estonia	77.7000000000...	1112.00000000...	99.5	25
10	France	82.7000000000...	4026.10000000...	100.0	100
11	Hungary	75.4000000000...	893.7000000000...	100.0	100
12	Israel	82.0000000000...	2756.10000000...	100.0	0
13	Jamaica	75.8000000000...	294.3000000000...	92.5	50
14	Latvia	74.8000000000...	783.8000000000...	98.5	100
15	Lithuania	74.4000000000...	923.3000000000...	96.5	100
16	Luxembourg	82.4000000000...	6236.00000000...	100.0	100
17	Malaysia	75.1000000000...	385.6000000000...	94.0	100
18	Maldives	78.1000000000...	943.9000000000...	98.0	75
19	Mauritius	74.6000000000...	506.1000000000...	100.0	75
20	Mexico	76.2000000000...	534.8000000000...	97.0	100

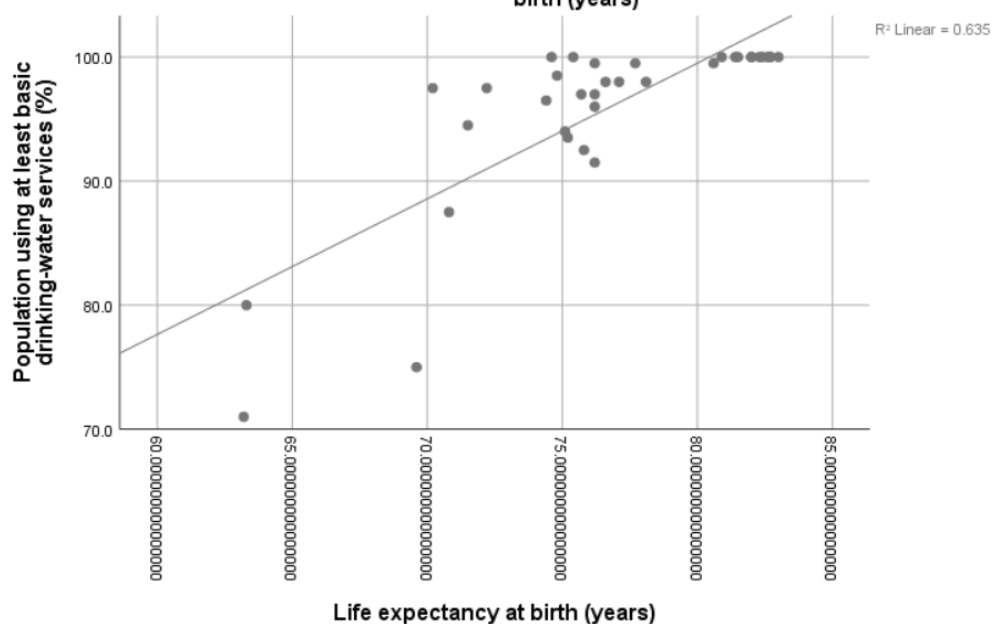
Variable type:

Variable	Type	Scale
Life expectancy at birth	Dependent	Continuous
Current health expenditure (CHE) per capita in us dollars	Independent	Continuous
Percentage of population using least basic drinking water	Independent	Continuous
International Health Regulations monitoring framework's legislation score'	Independent	Continuous

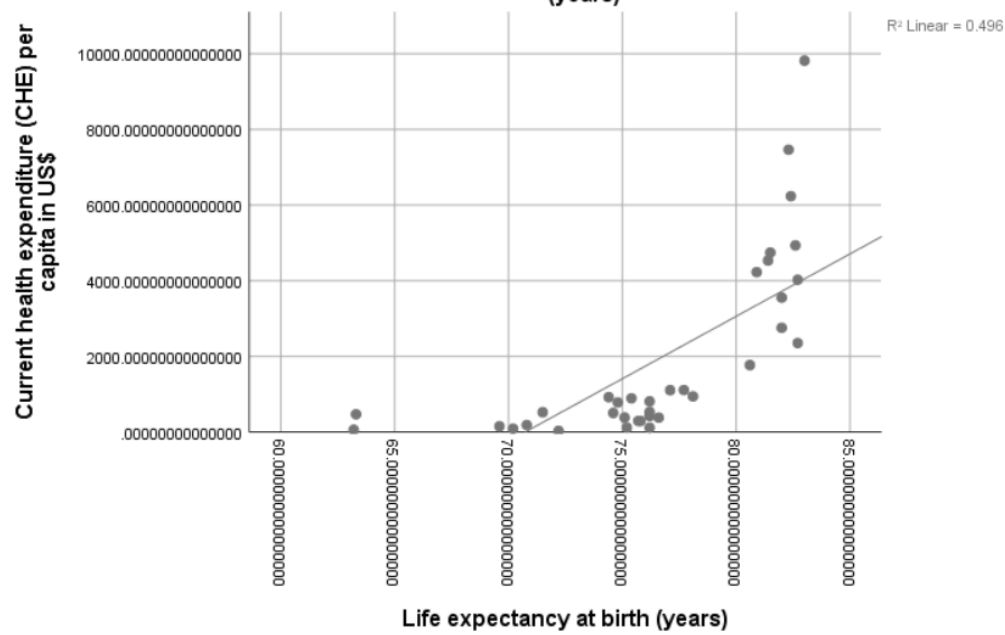
3. Linearity

Linear regression assumes that there is a straight-line relationship between predictors and outcome. So, below plots are to check linearity between the predictors and outcome and from the plots shows that there is linear relationship.

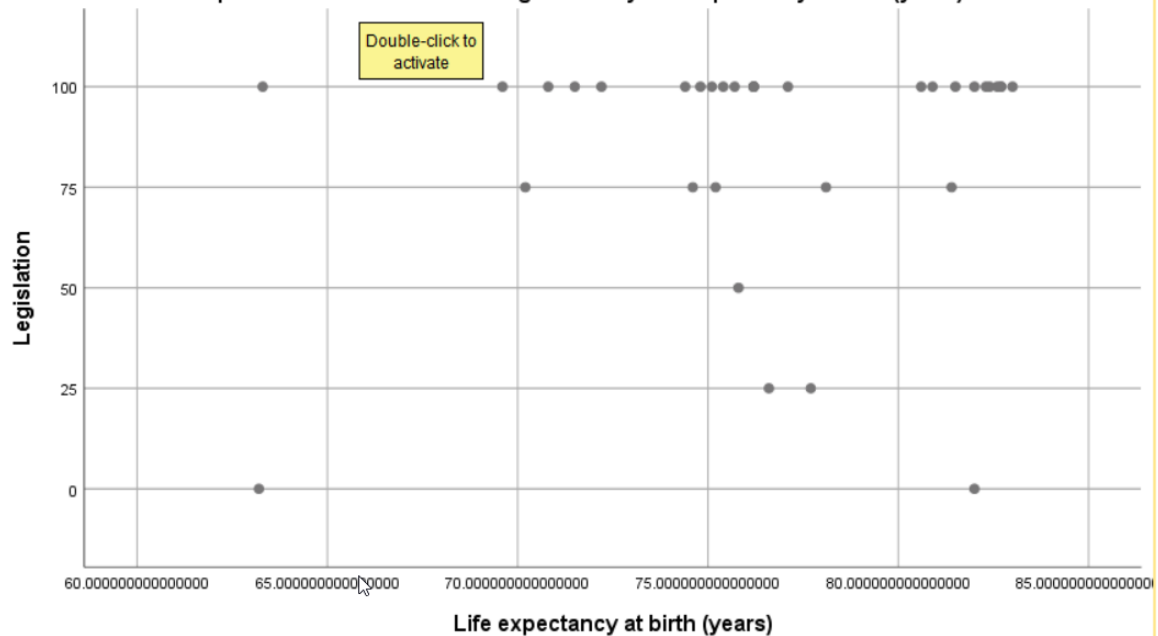
Simple Scatter with Fit Line of Population using at least basic drinking-water services (%) by Life expectancy at birth (years)



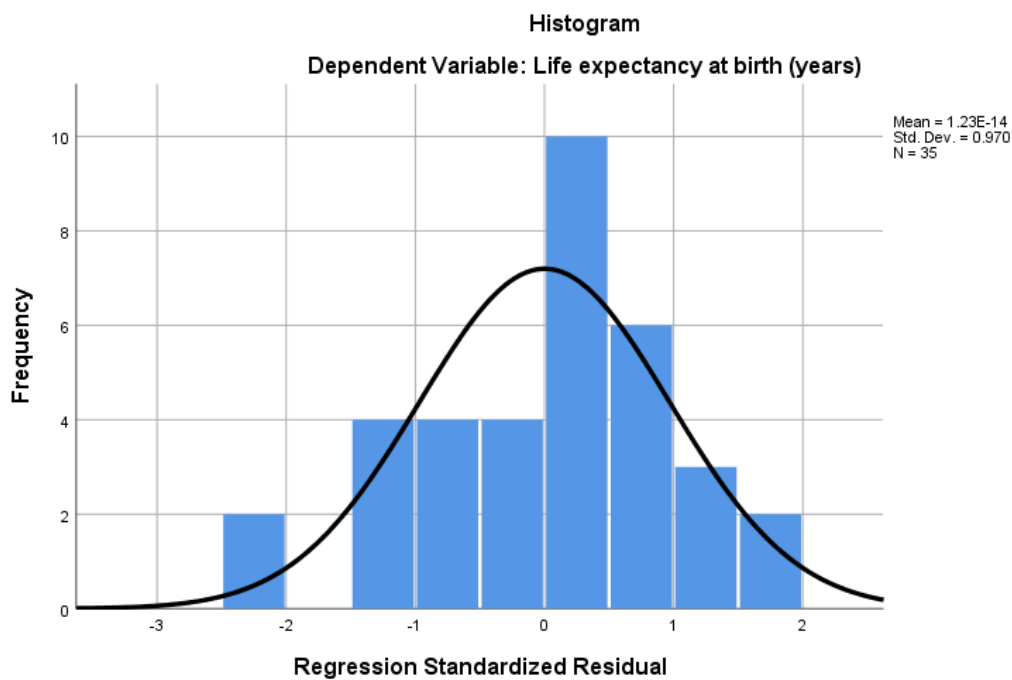
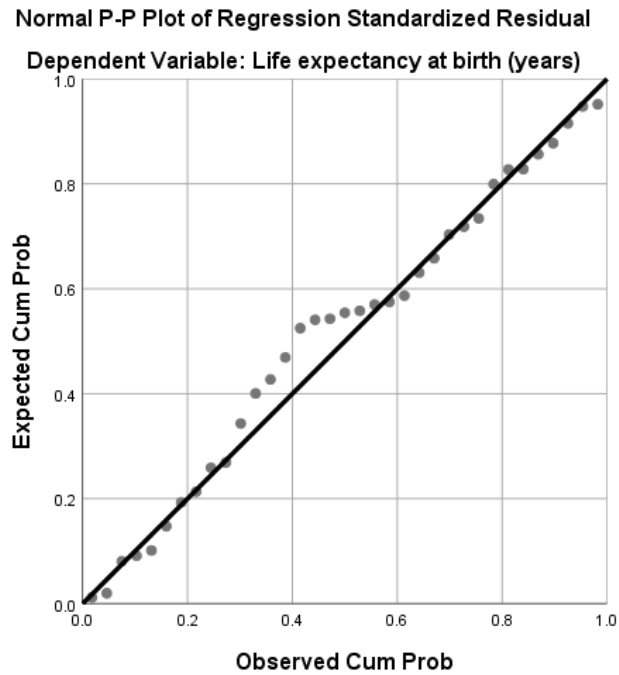
Simple Scatter with Fit Line of Current health expenditure (CHE) per capita in US\$ by Life expectancy at birth (years)



Simple Scatter with Fit Line of Legislation by Life expectancy at birth (years)



Below plot plots show that model is linear.



Above histogram shows that distribution of residual is normally distributed.

4. Correlation Matrix

In this model as per the correlation matrix following observations are made:

1. All three independent variables are having positive significant effect on the dependent variable.

2. 'Percentage of population using least basic drinking water' has most significant effect on 'life expectancy rate' ($r = 0.797$, $p = .000$) and IHR's Legislation has least positive effect on 'life expectancy rate' ($r = 0.162$, $p = 0.176$).
3. Highest correlation between the independent variables is between 'current health expenditure' and 'population using least basic drinking water' ($r = 0.418$, $p = 0.006$).

Correlations

		Life expectancy at birth (years)	Current health expenditure (CHE) per capita in US\$	Population using at least basic drinking-water services (%)	Legislation
Pearson Correlation	Life expectancy at birth (years)	1.000	.704	.797	.162
	Current health expenditure (CHE) per capita in US\$.704	1.000	.418	.187
	Population using at least basic drinking-water services (%)	.797	.418	1.000	.216
	Legislation	.162	.187	.216	1.000
Sig. (1-tailed)	Life expectancy at birth (years)	.	.000	.000	.176
	Current health expenditure (CHE) per capita in US\$.000	.	.006	.140
	Population using at least basic drinking-water services (%)	.000	.006	.	.107
	Legislation	.176	.140	.107	.
N	Life expectancy at birth (years)	35	35	35	35
	Current health expenditure (CHE) per capita in US\$	35	35	35	35
	Population using at least basic drinking-water services (%)	35	35	35	35
	Legislation	35	35	35	35

5. Model Summary

From the model summary following observations are made:

Model Summary ^c										
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	Change Statistics			Sig. F Change	Durbin-Watson
						F Change	df1	df2		
1	.797 ^a	.635	.623	3.168242242	.635	57.289	1	33	.000	
2	.895 ^b	.801	.789	2.371592577	.167	26.894	1	32	.000	2.317

a. Predictors: (Constant), Population using at least basic drinking-water services (%)
b. Predictors: (Constant), Population using at least basic drinking-water services (%), Current health expenditure (CHE) per capita in US\$
c. Dependent Variable: Life expectancy at birth (years)

1. R value in the first row represents the correlation between the outcome and the predictor 'population using at least basic drinking water services' (0.797). R value in second row represents the correlation between the outcome and predictors (0.895).
2. R square represents the amount of variability in the outcome by the predictors. R square value in the first row (0.635) means that predictor 'population using at least basic drinking water services' accounts for 63.5% of the variations in the 'life expectancy rate'. When other predictors are also included R square values increases to .801 or 80.1% of the variance in 'life expectancy rate'.
3. The adjusted R square value represents how well the model generalises and the difference for the final model means that if model were derived from the population rather than sample then it would account for 0.12% 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 2.317 which is close to 2 and in between 1-3 represents it is tenable.

6. ANOVA

ANOVA tells us whether regression model is significantly better at predicting values of the outcome than using the mean and Sig value in the ANOVA table for this model is 0 which means the model is significant. Zero sig value means that R square value is higher which indeed represents that model is significant.

ANOVA ^a						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	575.056	1	575.056	57.289	.000 ^b
	Residual	331.246	33	10.038		
	Total	906.302	34			

2	Regression	726.319	2	363.160	64.568	.000 ^c
	Residual	179.982	32	5.624		
	Total	906.302	34			

7. Coefficients:

Coefficients ^a											
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	20.993	7.363		2.851	.007					
	Population using at least basic drinking-water services (%)	.580	.077	.797	7.569	.000	.797	.797	.797	1.000	1.000
2	(Constant)	32.258	5.924		5.445	.000					
	Population using at least basic drinking-water services (%)	.443		.609	7.018	.000	.797	.779	.553	.825	1.212
	Current health expenditure (CHE) per capita in US\$.001	.000	.450	5.186	.000	.704	.676	.409	.825	1.212

a. Dependent Variable: Life expectancy at birth (years)

Equation for multiple regression as per the coefficient values in the coefficients table is:

$$\text{Life Expectancy Rate} = 32.423 + (0.001) X_1 + (0.450) X_2 + (-0.010) X_3$$

where X_1 = current health expenditure per capita, X_2 = population using at least basic drinking water and X_3 = International Health Regulations monitoring framework's legislation score.

So, for $x_1 = 350000$, $x_2 = 80$ and $x_3 = 75$

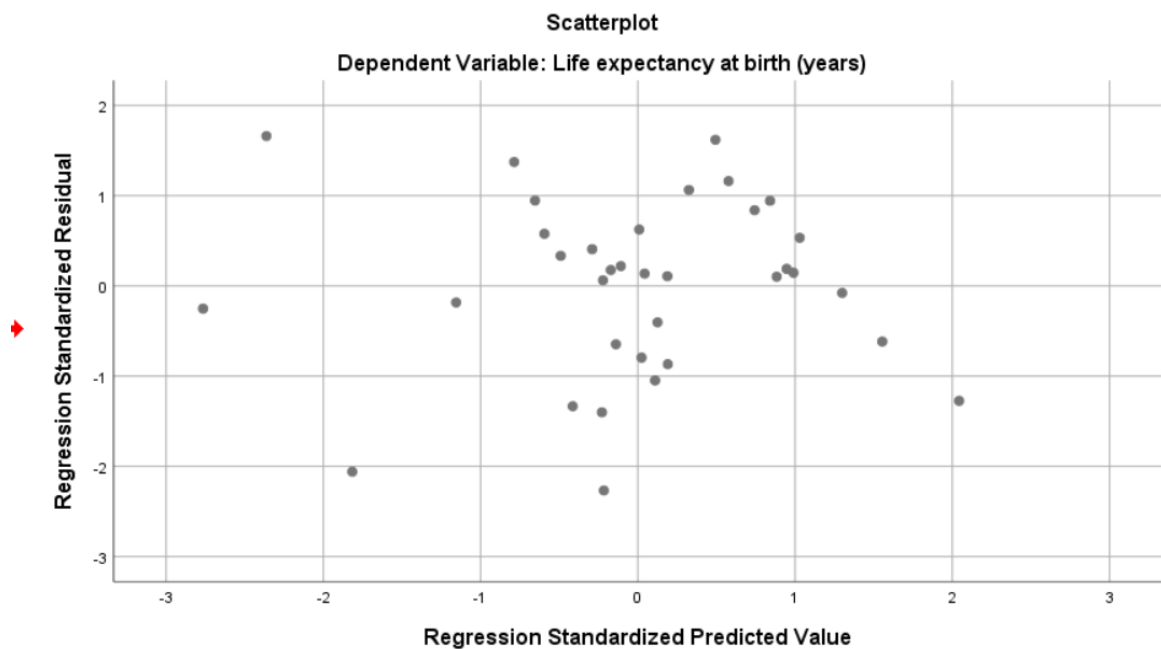
$$\text{life expectancy rate} = 32.423 + (0.001)3500 + (0.450)80 + (-0.010) 75 = 32.423 + 3.5 + 36 - 0.75 = 71.173.$$

Below observations are made with the help of coefficients table:

1. $B = 0.443$ represents that if population using at least basic drinking water increases by 1% then life expectancy rate will increase by 0.443 years.
2. $B = 0.001$ represents that if current health expenditure per capita increases by one us dollars then life expectancy rate will increase by 0.001 years.
3. Standardised beta value tells us the same thing but in terms of standard deviation so, If population using at least basic drinking water increases by 1 standard deviation then life expectancy rate will increase by 0.609 standard deviation. Standardised beta value provides better insight because standardised beta values are measured in standard deviation unit.
4. Sig value in t test represents the significance level of predictors to the model, smaller the sig value and larger the value of t the greater the contribution of predictors to the model.
8. Multicollinearity can be checked using the collinearity statics each value in the VIF values (1, 1.2) being less then 10 and average of all values (1.1) coming close to 1 represents there is no cause of concern in terms of multicollinearity.

9. Residual Statistics:

Residual statistics assesses the accuracy of the model in the sample, i.e. how well the model fits the data and for that to happen the standardised values should lie between -3 and +3 and for this model as per the histogram and scatter plot all standardised values lie between -3 and +3 which means the model is fit.



Influential Cases:

Cook's distance measures the influence of each case on the model as a whole and the should be less than 1 and distance in this model is 0.529 which means it is out of concern.

Residuals Statistics^a

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	63.79798889	86.02124023	76.57714286	4.621940231	35
Std. Predicted Value	-2.765	2.043	.000	1.000	35
Standard Error of Predicted Value	.452	1.509	.640	.273	35
Adjusted Predicted Value	64.05215454	87.73584747	76.62860195	4.771569827	35
Residual	-5.37817574	3.939686060	.0000000000	2.300782752	35
Std. Residual	-2.268	1.661	.000	.970	35
Stud. Residual	-2.335	1.971	-.010	1.033	35
Deleted Residual	-5.94079638	5.547848225	-.051459089	2.631080474	35
Stud. Deleted Residual	-2.523	2.070	-.018	1.069	35
Mahal. Distance	.266	12.795	1.943	3.084	35
Cook's Distance	.000	.529	.053	.130	35
Centered Leverage Value	.008	.376	.057	.091	35

a. Dependent Variable: Life expectancy at birth (years)

Logistic Regression

1. Objective

To build a logistic regression model that predict the Level of education which is dichotomous using Gender, Age group and Frequency of drinking alcohol every month. Here binary logistic regression is used as dichotomous dependent variable is used, if there would have been more than two categories then multinomial logistic regression would have been used.

2. Dataset

- Variables used in build this model are accessed from Europa Eurostat website.
- 'Level of education' is accessed from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Alcohol_consumption_statistics#General_overview . Data for three level of education was available out of which two where selected namely 'Less than primary, primary and lower secondary education (levels 0-2)' which is encode as 1 and 'Upper secondary and post-secondary non-tertiary education (levels 3 and 4)' which is encoded as '0' so that the dependent variable is dichotomous.
- 'Age group' is accessed from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Alcohol_consumption_statistics#General_overview . Data for multiple age groups where available in the website out of which two where accessed namely 'From 15 to 24 years' which is encoded as 0 and 'From 18 to 44 years' which is encoded as '1'.
- 'Gender' is accessed from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Alcohol_consumption_statistics#General_overview . Female is encoded as '1' and male as '0'.
- 'Frequency of drinking alcohol every month' is accessed from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Alcohol_consumption_statistics#General_overview .

Variable Type:

Variable	Type	Scale
Level of education	Dependent	Dichotomous
Gender	Independent	Dichotomous
Age group	Independent	Dichotomous
Frequency of drinking alcohol every month	Independent	Continuous

Sample rows:

Level of Education	Gender	Age Group	Frequency of drinking alcohol every month
0	0	0	14.900000000000...
0	0	1	18.500000000000...
0	1	0	11.300000000000...
0	1	1	9.500000000000...
1	0	0	27.700000000000...
1	0	1	25.600000000000...
1	1	0	20.500000000000...
1	1	1	14.600000000000...
0	0	0	29.100000000000...
0	0	1	25.700000000000...
0	1	0	31.000000000000...
0	1	1	22.500000000000...
1	0	0	27.700000000000...
1	0	1	33.100000000000...
1	1	0	24.900000000000...
1	1	1	20.700000000000...
0	0	0	8.600000000000...
0	0	1	26.300000000000...
0	1	0	2.700000000000...
0	1	1	6.700000000000...
1	0	0	20.100000000000...

3. Case processing summary:

To build a logistic regression 240 cases were considered as shown in the image below.

Case Processing Summary

Unweighted Cases ^a		N	Percent
Selected Cases	Included in Analysis	240	100.0
	Missing Cases	0	.0
	Total	240	100.0
Unselected Cases		0	.0
Total		240	100.0

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

4. Variable encoding:

Dependent variable encoding, and independent variable encoding done to build the model is as shown in below figure:

Dependent Variable Encoding			
Original Value		Internal Value	
0		0	
1		1	

Categorical Variables Codings			
Frequency			Parameter coding (1)
Age Group	0	120	1.000
	1	120	.000
Gender	0	120	1.000
	1	120	.000

5. Beginning block:

- Beginning block is the first model in logistic regression, it describes the input variables before the regression is performed.
- Classification table in this block predicts the probability of the outcome based on the frequency of the predicted value.
- Sig value in the table Variables in the equation whether the predictors are significant enough to predict the outcome and the value 1.0 in this model represent strong significant.
- The constant in the second table named Variables in the Equation gives the unconditional log odds of type of education.
- Score of each predictor in the table named Variables not in the Equation gives the estimated change in model if the predictors are considered in the model.

Block 0: Beginning Block

Classification Table^{a,b}

Observed			Predicted		Percentage Correct
			Level of Education		
			0	1	
Step 0	Level of Education	0	0	120	.0
		1	0	120	100.0
	Overall Percentage				50.0

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	.000	.129	.000	1	1.000	1.000

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Gender(1)	.000	1	1.000
		Age Group(1)	.000	1	1.000
		Frequency of drinking alcohol every month	15.408	1	.000
		Overall Statistics	17.289	3	.001

6. Omnibus test:

- This test is done to check whether the predictors are significant enough to predict the outcome. As in this model sig value is 0 which is less than 0.0005 represents that predictors are significant, so null hypothesis is model has significant predictive capacity.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	18.282	3	.000
	Block	18.282	3	.000
	Model	18.282	3	.000

7. Model Summary:

- R square value in model summary assess the model, R square value less than 0 means that model is not significant and R square value 1 represents the model is perfect.
- Cox and Snell R square value of 0.073 indicates that 7.3% of the variations in the outcome can be predicted using this model.
- Nagelkerke R square value of 0.098 indicates that 9.8% of the variations in the outcome can be predicted using this model.
- The lower the value in -2 Log likelihood the better the fit. As -2 Log likelihood value is high also R square values are not good this model is not that great.

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	314.429 ^a	.073	.098
a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.			

8. Hosmer and Lemeshow Test:

- This test assess the model based on the Sig value, and value greater than 0.05 represents that model is significantly strong. Sig value of 0.591 indicates support for the model.

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
1	6.506	8	.591

9. Variables in the Equation:

- The p values (sig) for all the variables suggest whether they contribute significantly to the predictive ability of the model or not, and according to the values in this table for this model variables Frequency of drinking alcohol every month is significant and other two variables are not significant.

		Variables in the Equation					
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	Gender(1)	-.428	.290	2.168	1	.141	.652
	Age Group(1)	.016	.268	.004	1	.952	1.016
	Frequency of drinking alcohol every month	.047	.012	15.681	1	.000	1.048
	Constant	-.726	.291	6.215	1	.013	.484
a. Variable(s) entered on step 1: Gender, Age Group, Frequency of drinking alcohol every month.							

- Greater the Wald score higher the contribution of predictors to the model, so in this model Frequency of drinking alcohol every month has higher contribution to the model than Gender and Age group.
- B value tells the effect of individual predictors on the outcome in terms of the standard deviation and negative sign in the values implies that they have negative correlation on the outcome.
- Exp(B) indicates the change in odds resulting from a unit change in the predictor. Value greater than one implies that probability of outcome occurring increases and less than one implies probability of outcome occurring decreases. So, Frequency of drinking alcohol every month and Age group are significant, and gender is not significant.

10. Classification Table:

- Percentage correct value in this table represents the percentage accuracy in the classification which indicates what percentage of cases will the model classify the outcome correctly.
- 59.6 value represents that 59.6% cases of overall cases this model will predict the outcome correctly.
- Prediction percentage increased from 50 to 59.6% after inserting the predictors.

Classification Table^a

Observed			Predicted		Percentage Correct
			Level of Education 0	1	
Step 1	Level of Education	0	77	43	64.2
		1	54	66	55.0
Overall Percentage					59.6

a. The cut value is .500