Statistics for Data Analytics

Continuous Assessment 1 – Semester 1, 2020-21

Multiple Regression

Analysis of Alcohol Consumption using Multiple Regression.

Submitted by: Mayuresh Mohan Londhe

X20137265

I. MULTIPLE REGRESSION

Multiple Regression is similar to the Linear Regression Model instead of finding the relationship between one Dependent variable and Independent Variable here we use Multiple Independent Variables. It involves the use of multiple independent variables and one dependent variable which attempts to form a relationship [1].

Regression Equation of a model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Where Y is a dependent Variable,

$X_1, X_2, \ldots .. X_p$ are independent (explanatory) variables,

β_0 is Y – Intercept (Constant Term)

$\beta_1, \ldots\ldots, \beta_p$ are regression (Slope) coefficients for each independent variable.

ε is the Error Term (Residuals).

II. STATEMENT OF OBJECTIVES AND DESCRIPTION OF VARIABLES

To analyze the Alcohol Consumption in 120 different countries in the year 2016 based on multiple independent variables such as Average price 500 mls Beer in US$, Average price 750 mls Wine in US$, Average price 500 mls Spirits in US$, 15-years old first drink at 13 years or younger (%), Unemployment Rate, Advertising restrictions, Literacy Rate, Country Type.

Research Question:

Does Alcohol Consumption among different countries can be explained by the following factors. Average Price of Beer, Wine, Spirit, 15-years old first drink at 13 years or younger, Unemployment Rate, Advertising restrictions, Literacy Rate, Country Type?

The Dataset is taken from the World Health Organization Website. For analysis purposes, I have clustered 9 different Datasets into one Dataset. Initial Dataset was containing missing values which are removed before starting the regression process.

Dependent Variable: Total alcohol Consumption per capita

Independent Variables:

Continuous Variables:
1) Average price 500 mls Beer in US$,
2) Average price 750 mls Wine in US$,
3) Average price 500 mls Spirits in US$,
4) 15-years old first drink at 13 years or younger (%),
5) Unemployment Rate,
6) Literacy Rate.

Categorical Variables:
7) Type of Country (Developing 0, Developed 1)
8) Advertising restrictions (Yes 1, No 0)

Data source: https://apps.who.int/gho/data/node.imr

III. ASSUMPTIONS UNDERTAKEN

- **Assumption 1: Sample Size:**

This assumption is considered to verify whether we have sufficient observations to perform Multiple Regression. If we are having small samples result of the regression will not generalize to other samples [2]. We use the following formula to calculate the Sample Size.

$$N > 50 + 8m$$

$$120 > 114$$

Here, N is the Sample Size and m is the number of Independent Variables.
For 8 independent Variables, we should have a sample size greater than 114. Descriptive Statistics is used in SPSS to verify whether each Variable has observations greater than 114.

Descriptive Statistics

	Mean	Std. Deviation	N
Total alcohol consumption	15.069	7.0246	120
Average price 500 mls Beer in US$	2.1309	1.77838	120
Average price 750 mls Wine in US$	8.9112	4.46421	120
Average price 500 mls Spirits in US$	15.3859	5.35320	120
15-years old first drink at 13 years or younger %	19.7457	12.21676	120
Unemployment Rate	13.7137	7.60500	120
Literacy Rate	84.810	17.5004	120
Country Type	.32	.467	120
Advertising restrictions	.55	.500	120

- **Assumption 2: Absence for Multicollinearity:**

The relationship among independent Variables is referred as Multicollinearity [2]. Multicollinearity occurs when two or more independent variables (predictors) are functions of each other. When there is a presence of Multicollinearity among independent variables standard error will be larger than it should be. To test the Multicollinearity, Correlation Matrix is generated to check this assumption that there is no Multicollinearity between independent variables.

Correlations

		Total alcohol consumption	Average price 500 mls Beer in US$	Average price 750 mls Wine in US$	Average price 500 mls Spirits in US$	15-years old first drink at 13 years or younger %	Unemployment Rate	Literacy Rate	Country Type	Advertising restrictions
Pearson Correlation	Total alcohol consumption	1.000	-.080	-.587	-.441	.814	.552	.307	.636	-.603
	Average price 500 mls Beer in US$	-.080	1.000	.362	.445	-.006	-.346	.111	.202	.006
	Average price 750 mls Wine in US$	-.587	.362	1.000	.777	-.498	-.511	-.060	-.277	.309
	Average price 500 mls Spirits in US$	-.441	.445	.777	1.000	-.336	-.462	.006	-.141	.189
	15-years old first drink at 13 years or younger %	.814	-.006	-.498	-.336	1.000	.424	.233	.690	-.661
	Unemployment Rate	.552	-.346	-.511	-.462	.424	1.000	.064	.189	-.347
	Literacy Rate	.307	.111	-.060	.006	.233	.064	1.000	.363	-.015
	Country Type	.636	.202	-.277	-.141	.690	.189	.363	1.000	-.429
	Advertising restrictions	-.603	.006	.309	.189	-.661	-.347	-.015	-.429	1.000
Sig. (1-tailed)	Total alcohol consumption	.	.193	.000	.000	.000	.000	.000	.000	.000
	Average price 500 mls Beer in US$	.193	.	.000	.000	.472	.000	.114	.013	.474
	Average price 750 mls Wine in US$	.000	.000	.	.000	.000	.000	.257	.001	.000
	Average price 500 mls Spirits in US$	.000	.000	.000	.	.000	.000	.474	.063	.019
	15-years old first drink at 13 years or younger %	.000	.472	.000	.000	.	.000	.005	.000	.000
	Unemployment Rate	.000	.000	.000	.000	.000	.	.243	.019	.000
	Literacy Rate	.000	.114	.257	.474	.005	.243	.	.000	.435
	Country Type	.000	.013	.001	.063	.000	.019	.000	.	.000
	Advertising restrictions	.000	.474	.000	.019	.000	.000	.435	.000	.
N	Total alcohol	120	120	120	120	120	120	120	120	120

Total Alcohol Consumption is highly related to 15-years old first drink at 13 years or younger % and Moderately related to the country type and Unemployment Rate and Less dependent on Literacy Rate. It is inversely proportional to Average Beer, Wine, Spirit Price and Advertisement restriction.

The ideal value for correlation should be less than |0.8|, which is not the case here as per the above table. Variance Inflation Factor Value is taken into consideration for further multicollinearity checks.

VIF is a measure of how much common variance is present among independent variables. It is used to check the multicollinearity which is not evident in the correlation Matrix. We need to give attention to a predictor who has VIF greater than 10 as it could be collinear with other predictors. Tolerance is the inverse of the Variance Inflation Factor. As per the below table assumption of Multicollinearity is met. There are no variables which are having VIF greater than 10.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	9.001	2.801		3.214	.002		
	Average price 500 mls Beer in US$	.028	.227	.007	.123	.903	.684	1.463
	Average price 750 mls Wine in US$	-.239	.133	-.152	-1.800	.075	.315	3.177
	Average price 500 mls Spirits in US$	-.091	.103	-.069	-.882	.380	.364	2.745
	15-years old first drink at 13 years or younger %	.255	.050	.443	5.117	.000	.300	3.333
	Unemployment Rate	.116	.054	.124	2.154	.033	.676	1.480
	Advertising restrictions	-1.796	.899	-.128	-1.998	.048	.550	1.819
	Literacy Rate	.037	.024	.080	1.563	.121	.858	1.165
	Country Type	2.478	1.044	.165	2.373	.019	.466	2.146

a. Dependent Variable: Total alcohol consumption

- **Assumption 3: Outliers (No influential Data Points)**

Observations that are not predicted well by the model are called Outliers. These are also called as High Leverage Observations. High-leverage observations may or may not be influential observations. Model Parameter values affect inappropriately due to the presence of influential observations. Removing such observation can result in a significant change in the Model.

To check this assumption, we use Cook's Distance. Cooks distance shouldn't be greater than or equal to 1. As per the Residual Statistics, the Min and Max Cook's distance is 0.000 and 0.084 respectively hence proving that there are no outliers.

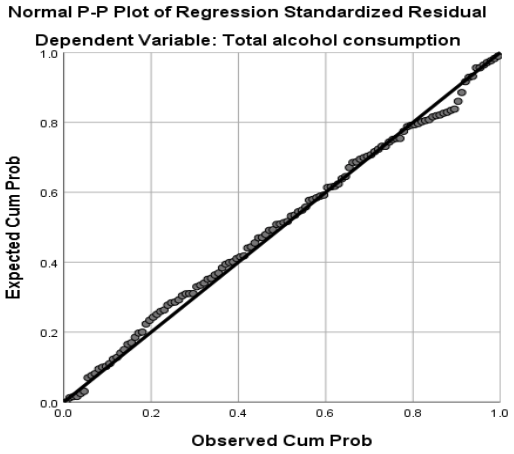
Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	3.936	26.325	15.069	6.1860	120
Std. Predicted Value	-1.800	1.820	.000	1.000	120
Standard Error of Predicted Value	.487	1.568	.923	.200	120
Adjusted Predicted Value	4.098	26.167	15.046	6.1801	120
Residual	-10.3422	8.1501	.0000	3.3285	120
Std. Residual	-3.001	2.365	.000	.966	120
Stud. Residual	-3.116	2.485	.003	1.003	120
Deleted Residual	-11.1489	8.9961	.0229	3.5937	120
Stud. Deleted Residual	-3.247	2.545	.002	1.015	120
Mahal. Distance	1.383	23.654	7.933	3.913	120
Cook's Distance	.000	.084	.009	.014	120
Centered Leverage Value	.012	.199	.067	.033	120

a. Dependent Variable: Total alcohol consumption

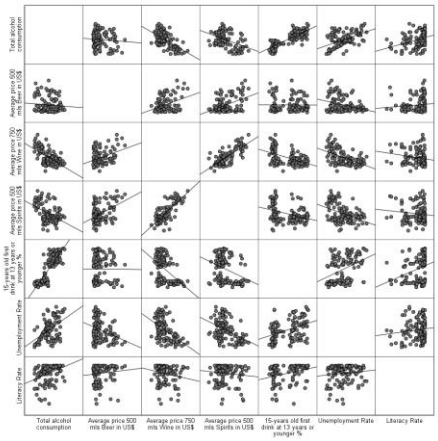
- **Assumption 4: Normality (Errors are normally distributed):**

P-P plot of Regression Standardized Residual is used to check the assumption of normality. The Errors should be Normally distributed. In which the datapoints (Dots) should be close to the reference line. This is used to check whether Residual Values are normally distributed. As per the below Plot Normality assumption is not violated.



- **Assumption 5: The linear relationship between the independent and the dependent variable**

As per the first point of a Gauss–Markov Assumptions assumption. We should have the correct functional form of a model. In that, we first check Linearity between Predictors and the Response.

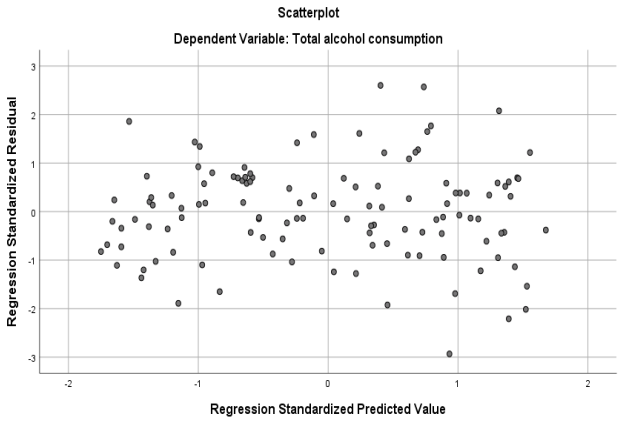


H_0: Slope is Equal to Zero.
H_1: Slope is not Equal to Zero.

As per the above image, we can see that there is a Linear Relationship between Dependent Variable and Independent Variables. Hence, we reject the Null hypothesis. Also

- **Assumption 6: Checking for homoscedasticity**

This an important assumption of a multiple regression that checks whether error terms (Residuals) have constant variance. In this assumption the Null Hypothesis would be Errors are not having constant Variance i.e. heteroscedasticity is present Alternative Hypothesis would be Errors are having Constant Variance i.e. homoscedasticity. To confirm the homoscedasticity, we use a scatterplot. As per the below plot, residuals are equally scattered (distributed). So, we can confirm that assumption for homoscedasticity is met and there is no heteroscedasticity.



- **Assumption 7: Autocorrelation check between Errors (Independence of Errors)**

As per this assumption error terms (Residuals) shouldn't be correlated. To test this assumption, we use Durbin–Watson statistics. Durbin–Watson statistics value should be in the range of 1 to 3. Better if it is close to 2. Durbin Watson Statistics value is 2.175 hence the assumption of Independence of Residuals is held.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.866[a]	.750	.732	3.6333	2.175

a. Predictors: (Constant), Country Type, Average price 500 mls Spirits in US$, Literacy Rate, Advertising restrictions , Unemployment Rate, Average price 500 mls Beer in US$, Average price 750 mls Wine in US$, 15-years old first drink at 13 years or younger %

b. Dependent Variable: Total alcohol consumption

IV. UNDERSTANDING AND BUILDING THE MODEL:

Analyze option from SPSS is used to define the Dependent Variable and Independent Variables.

- **Variables Entered/Removed:**

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Country Type, Average price 500 mls Spirits in US$, Literacy Rate, Advertising restrictions , Unemployment Rate, Average price 500 mls Beer in US$, Average price 750 mls Wine in US$, 15-years old first drink at 13 years or younger %[b]	.	Enter

a. Dependent Variable: Total alcohol consumption

b. All requested variables entered.

Model 1:

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	9.001	2.801		3.214	.002		
	Average price 500 mls Beer in US$	.028	.227	.007	.123	.903	.684	1.463
	Average price 750 mls Wine in US$	-.239	.133	-.152	-1.800	.075	.315	3.177
	Average price 500 mls Spirits in US$	-.091	.103	-.069	-.882	.380	.364	2.745
	15-years old first drink at 13 years or younger %	.255	.050	.443	5.117	.000	.300	3.333
	Unemployment Rate	.116	.054	.124	2.154	.033	.676	1.480
	Advertising restrictions	-1.796	.899	-.128	-1.998	.048	.550	1.819
	Literacy Rate	.037	.024	.080	1.563	.121	.858	1.165
	Country Type	2.478	1.044	.165	2.373	.019	.466	2.146

a. Dependent Variable: Total alcohol consumption

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.866[a]	.750	.732	3.6333	2.175

a. Predictors: (Constant), Country Type, Average price 500 mls Spirits in US$, Literacy Rate, Advertising restrictions , Unemployment Rate, Average price 500 mls Beer in US$, Average price 750 mls Wine in US$, 15-years old first drink at 13 years or younger %

b. Dependent Variable: Total alcohol consumption

As per the above two images, we can find out that the model is having R Square value of 0.750 with Std. Error of 3.633. Model is describing 75% of the variance in Total Alcohol Consumption but there are Independent Variables that are not significant.

1) AveragePrice500mlsSpiritsinUS$
2) AveragePrice500mlsBeerinUS$
3) AveragePrice500mlsWinerinUS$
4) Literacy Rate

So, removed the first two variables from the above list to check whether there is an improvement in the model.

Model 2:

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.880[a]	.774	.762	3.4272	2.172

a. Predictors: (Constant), Advertising restrictions , Literacy Rate, Average price 750 mls Wine in US$, Unemployment Rate, Country Type, 15-years old first drink at 13 years or younger %

b. Dependent Variable: Total alcohol consumption

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	6.636	2.240		2.963	.004		
	Average price 750 mls Wine in US$	-.299	.088	-.190	-3.391	.001	.635	1.574
	15-years old first drink at 13 years or younger %	.230	.047	.401	4.859	.000	.294	3.400
	Unemployment Rate	.183	.050	.198	3.644	.000	.679	1.472
	Literacy Rate	.051	.020	.128	2.606	.010	.835	1.198
	Country Type	2.460	.983	.164	2.503	.014	.468	2.136
	Advertising restrictions	-1.954	.861	-.139	-2.270	.025	.534	1.873

a. Dependent Variable: Total alcohol consumption

Now we can see that all the Variables are significant below 0.05. The R square and adjusted R square values have improved by 0.024 and 0.030 i.e. 2.4%,3.0% respectively. There is a decrease in Durbin Watson Value by 0.003.and Std. Error of Estimate value by 0.205

V. MODEL SUMMARY

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.880[a]	.774	.762	3.4272	2.172

a. Predictors: (Constant), Advertising restrictions , Literacy Rate, Average price 750 mls Wine in US$, Unemployment Rate, Country Type, 15-years old first drink at 13 years or younger %

b. Dependent Variable: Total alcohol consumption

R-Value:

To measure the strength of the model and the dependent variable we check the R-value, which is called a correlation coefficient.

R Square:

The proportion of variation in the dependent variable that is explained by the independent variables is called R^2. Here all independent variables show 77.4% variance in Alcohol Consumption.

Adjusted R Square:

R^2 Increases as the number of independent variables increases even though variables are not well associated with Response This could help us fitting Test Data well but not the Training Data. So, to balance this effect of Variables we calculate Adjusted R^2. Adjusted R^2 is the R^2 value computed considering the penalty given for using several independent variables.

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

Where **n** is the Sample Size and **p** in Number of Independent Variables.

Std. Error of the Estimate:

Standard Error of the Estimate is the amount Error model will have while making the prediction. The unit of the Dependent Variable and Standard Error of the Estimate is identical. There is no standard value which is considered good or bad while evaluating the model, it depends on what we are predicting.

As per the Model Summary, there will be a 3.4272 error while predicting the Alcohol Consumption of a country.

ANOVA:

To check whether the model is significantly better at predicting values of the outcome variable we use the F Test. Here we have a null Hypothesis as all **β** coefficients are equal to zero and an Alternative Hypothesis as not all **β** are equal to zero.

H_0: $\beta 1 = \beta 2 = \ldots. = \beta_K = 0$

H_1: Not all βs are equal to 0

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4544.850	6	757.475	64.491	.000[b]
	Residual	1327.226	113	11.745		
	Total	5872.076	119			

a. Dependent Variable: Total alcohol consumption

b. Predictors: (Constant), Advertising restrictions , Literacy Rate, Average price 750 mls Wine in US$, Unemployment Rate, Country Type, 15-years old first drink at 13 years or younger %

Sig value is 0.000 which is interpreted as all the variables are contributing to the prediction of Alcohol Consumption. At the 0.05 significance level, the critical value is 2.18. F(6,133) =64.491 is greater than the critical value which indicates that the regression model is fitting the data strongly and we reject the Null Hypothesis.

Coefficients:

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	6.636	2.240		2.963	.004		
	Average price 750 mls Wine in US$	-.299	.088	-.190	-3.391	.001	.635	1.574
	15-years old first drink at 13 years or younger %	.230	.047	.401	4.859	.000	.294	3.400
	Unemployment Rate	.183	.050	.198	3.644	.000	.679	1.472
	Literacy Rate	.051	.020	.128	2.606	.010	.835	1.198
	Country Type	2.460	.983	.164	2.503	.014	.468	2.136
	Advertising restrictions	-1.954	.861	-.139	-2.270	.025	.534	1.873

a. Dependent Variable: Total alcohol consumption

β value indicates that by what degree each Independent Variable affects the model if the outcome of the other variables are constant.

β_0 (Constant) = **6.636**
β_1 **= - 0.299** for every 1 US$ increase in wine price alcohol consumption is decreased by 0.299 per capita.
β_2 **= 0.230** for every 1 % increase in (15-years old first drink at 13 years or younger (%)), alcohol consumption increases by 0.230 per capita.
β_3 **= 0.183** for every 1 unit increase in the unemployment rate the alcohol consumption increases by 0.183 per capita.
β_4 **= - 1.954** if the country has advertisement restrictions the alcohol consumption is decreased by 1.954 per capita.
β_5 **= 0.051** for every 1unit increase in Literacy rate the alcohol consumption increases by 0.051s per capita.
β_6 **= 2.460** if the Country Type is Developed the alcohol consumption is increased by 2.460 per capita.

Standardize β is easy to read as they are not dependent on the unit of measurement of the Variable.

It indicates the number of standard deviations by which outcome will change when there is one standard deviation change in the predictor.

E.g. As wine Price in US$ increases by 1 S.D, Consumption reduces by 0.299 of a Standard Deviation.

Collinearity Diagnostics:

This is used when the VIF factor for more than two variables is greater than 10. To figure out which of the variables are collinear. To check the Variance among multiple Independent Variables we check for the eigenvalues and Condition Index. Low eigenvalue and High condition index above 15 show the sign of collinearity. As per the matrix, the 9th dimension has eigenvalues above 15. So, we check the Variance Proportion of that dimension with the rest of the other variables. There shouldn't be more than 1 Variance Proportion greater than 0.9 in one single row where Condition Index is above 15. 9th Dimension has only One Variance Proportion above 0.9 i.e. for Constant.

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	Average price 500 mls Beer in US$	Average price 750 mls Wine in US$	Average price 500 mls Spirits in US$	15-years old first drink at 13 years or younger %	Unemployment Rate	Literacy Rate	Country Type	Advertising restrictions
1	1	6.806	1.000	.00	.00	.00	.00	.00	.00	.00	.00	.00
	2	1.091	2.497	.00	.00	.00	.00	.01	.00	.00	.11	.06
	3	.493	3.714	.00	.26	.00	.00	.00	.10	.00	.06	.01
	4	.277	4.955	.00	.05	.01	.01	.00	.02	.00	.34	.43
	5	.163	6.466	.00	.66	.11	.03	.00	.11	.00	.09	.07
	6	.087	8.858	.00	.01	.04	.00	.41	.59	.01	.29	.03
	7	.041	12.958	.02	.00	.14	.00	.41	.07	.46	.00	.32
	8	.028	15.522	.00	.02	.68	.88	.01	.00	.06	.01	.00
	9	.014	22.370	.97	.00	.01	.09	.14	.10	.46	.10	.07

a. Dependent Variable: Total alcohol consumption

Casewise Diagnostics:

This compute and compares the predicted and actual values of the dependent variable by using the dataset as input.
For 111th Country in Dataset, Actual Consumption is 10 per capita whereas the model has predicted the value for the same country as 20.501 which is having a 10.5011 Error.

Casewise Diagnostics[a]

Case Number	Std. Residual	Total alcohol consumption	Predicted Value	Residual
111	-3.064	10.0	20.501	-10.5011

a. Dependent Variable: Total alcohol consumption

Multiple Regression Equation:

Total Alcohol Consumption = **6.636** + (**- 0.299)** * Average Price 750 mls Wine in US$ + **0.230** * 15-years old first drink at 13 years or younger (%) + **0.183** * Unemployment Rate + (**-1.954)** * Advertising Restrictions + **0.051** * Literacy Rate + **2.460** * Country Type

VI. CONCLUSION

Multiple Regression was applied to validate whether Alcohol Consumption in different countries is dependent on several different factors. Out of selected factors Average Price 750 mls Wine in US$, 15-years old first drink at 13 years or younger (%), Unemployment Rate, Advertising Restrictions, Literacy Rate, Country Type were the factors which were significant enough to build the model. The model was able to explain 76.2 % of the Variance. It can significantly predict Alcohol Consumption F (6,133) =64.491, 0.0001

VII. REFERENCES

[1] S. Srikamdee and J. Onpans, "Forecasting Daily Air Quality in Northern Thailand Using Machine Learning Techniques," 2019 4th International Conference on Information Technology (InCIT), Bangkok, Thailand, 2019, pp. 259-263, doi: 10.1109/INCIT.2019.8912072.

[2] Pallant, Julie. Ebook: SPSS Survival Manual: a Step by Step Guide to Data Analysis Using IBM SPSS, McGraw-Hill Education, 2020.