Statistical Research using Multiple Regression Model

Shubham Siras Katmusare
School of Computing
National College of Ireland
x19195117@student.ncirl.ie

I. RESEARCH QUESTION:

The objective is to implement the statistical analysis for the "total net official development assistance to medical research and basic health sectors per capita (US$), by recipient country" data for year 2017-18 using multiple regression model. This model will evaluate the most effective predictors among the selected features for the target variable i.e. "total net official development assistance" in this dataset.

II. DATA CONTEXT AND SOURCES :

The data for this analysis is extracted from the World Health organization Data[1]. The dataset contains 187 countries in total and each country possess the single instance for year 2018-19. It is featured with five variables collected from the different datasets are namely **"Country"**[2]**, "Total Net Official Development Assistance**"[2] , "**Total Alcohol Consumption Per Capita**" [3], "**Communication Disease By Country**"[4] and "**Estimate Of Tobacco Use Prevalence** " [5].

III. DATA VARIABLES :

The dataset contains below 5 columns and among them "Total Net Official Development Assistance" is the dependent variable and the rest i.e. "Total Alcohol Consumption Per Capita", "Communication Disease By Country" and "Estimate Of Tobacco Use Prevalence" league as independent variables.

1. Country :

   This refers to the nominal feature and differentiate each row for the data.

2. TotalNetOfficialDevelopmentAssistance :

   This is continuous feature and explains the total net official development assistance to medical research and basic health sectors per capita (US$), for year 2018-19.

3. TotalAlcoholConsumptionPerCapita :

   This is also a continuous feature and explicitly, contains data for the total (recorded + unrecorded) alcohol per capita consumption for population over the age of 15 for the recorded year.

4. CommunicationDiseaseByCountry :

   This is also a continuous feature and contains entries for the total new HIV infections per 1000 uninfected population for the recorded year.

5. EstimateOfTobaccoUsePrevalence :

   This is a continuous feature and contains data for the estimate of current tobacco use prevalence (%) for year 2018-19.

IV. DATA TRANSFORMATION :

The features in the dataset are collected from the different datasets and in order to merge all the features to form a single dataset and sustain the objective of analysis, below are the steps being followed :

1. Data for the year2018-19, for the features "Country"[2], "Total Net Official Development Assistance"[2] , "Total Alcohol Consumption Per Capita" [3], "Communication Disease By Country"[4] and "Estimate Of Tobacco Use Prevalence " [5] is collected from the data sources.

2. Grouped the data based on the countries using pandas Data frame.

3. Discarded null entries for the countries where more than 2 features were missing.

4. For feature "Communication Disease By Country", The data is converted to whole numbers from the existing percentage value in order to deduce the meaning out of it.

5. Eventually, followed by the steps above, the dataset is ready to analyze.

Country	TotalNetOfficialDevelopmentAssistance	TotalAlcoholConsumptionPerCapita	CommunicationDiseaseByCountry	EstimateOfTobaccoUsePrevelence
Uruguay	77.1	1	95	20
United_States_of_America	78.5	1	95	24
United_Republic_of_Tanzania	63.9	1	5	30
United_Kingdom_of_Great_Britain_and_Northern_Ireland	81.4	1	95	38
United_Arab_Emirates	77.2	1	95	20
Ukraine	72.5	1	95	30
Uganda	62.5	1	5	15
Tuvalu	68.8	1	43	5
Turkmenistan	68.2	1	95	28
Turkey	76.4	1	95	11
Tunisia	76.0	1	95	26
Trinidad_and_Tobago	71.8	1	95	23
Tonga	73.4	1	51	24
Togo	60.6	48	7	21
Timor-Leste	68.6	21	10	21
Thailand	75.5	0	78	22
Tajikistan	70.8	0	78	25
Syrian_Arab_Republic	63.8	1	95	24
Switzerland	83.3	0	95	9
Sweden	82.4	0	95	12
Suriname	71.8	3	92	6
Sudan	65.1	24	46	24

Fig1 : Transformed dataset

V. ASSUMPTIONS:

The following list of assumptions were considered during the multiple regression analysis.

A. Homoscedasticity test :

Check for each value of the predictors variance of the error term should be constant.

B. Independence of errors

C. Independence of each data point for the dataset

D. Predictor variables must be independent of the error

E. Regression residuals follow Normally distribution curve.

F. Absence of multicollinearity between independent variables.

G. Absence of influential data points.

H. There is linear relationship between predictors and target variable.

VI. ANALYSYS AND RESULTS:

Post data transformation, the multiple regression Model analysis is generated on the SPSS utility software tool.

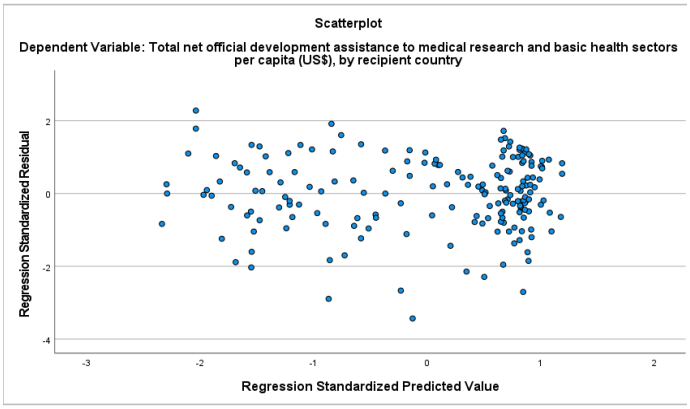

Fig2 : Scatter plot of standardized residual Vs standardized predicted value

From the figure 2, the scatterplot between the standardized residual Vs standardized predicted value do not follow any relationship and looks noisy so it can be concluded that the assumption of homoscedasticity is followed.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.772[a]	.596	.590	4.8340	.596	90.141	3	183	.000	2.246

a. Predictors: (Constant), Estimate of current tobacco use prevalence (%), New HIV infections (per 1000 uninfected population), Total (recorded+unrecorded) alcohol per capita (15+) consumption

b. Dependent Variable: Total net official development assistance to medical research and basic health sectors per capita (US$), by recipient country

Fig3 : Model Summary

The Durbin-Watson value in our analysis is 2.246 from the Figure 3, which is closer to 2 and hence it can be concluded that there is no autocorrelation between the errors.

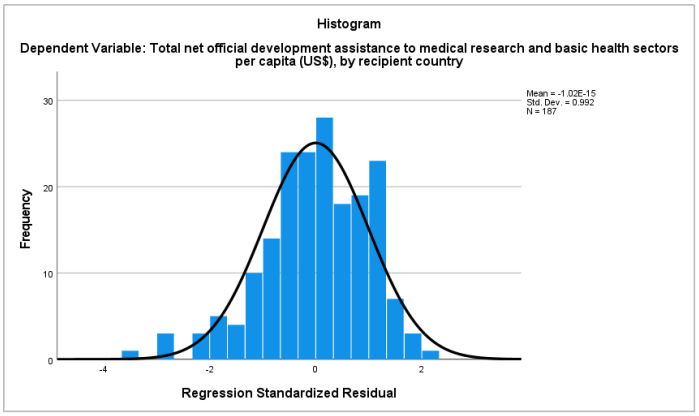

Fig 4 : Frequency Vs Regression residual

The above figure 4, it represents the relation between the frequency of the data points and regression standardized residual which follow the normal distribution curve. This is deduction of assumption of regression residuals following the normal distribution curve.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	62.748	1.257		49.927	.000					
	Total (recorded+unrecorded) alcohol per capita (15+) consumption	-.088	.029	-.177	-2.996	.003	-.571	-.216	-.141	.635	1.574
	New HIV infections (per 1000 uninfected population)	.132	.012	.634	10.826	.000	.751	.625	.508	.644	1.553
	Estimate of current tobacco use prevalence (%)	.068	.035	.094	1.974	.050	.209	.144	.093	.968	1.033

a. Dependent Variable: Total net official development assistance to medical research and basic health sectors per capita (US$), by recipient country

Fig 5 : Coefficients

The figure 5, the statistical VIF value sustains around 1, precisely is do not exceed 10 which represents that there is absence of multicollinearity between predictor variables.

Correlations

		Total net official development assistance to medical research and basic health sectors per capita (US$), by recipient country	Total (recorded+unrecorded) alcohol per capita (15+) consumption	New HIV infections (per 1000 uninfected population)	Estimate of current tobacco use prevalence (%)
Pearson Correlation	Total net official development assistance to medical research and basic health sectors per capita (US$), by recipient country	1.000	-.571	.751	.209
	Total (recorded+unrecorded) alcohol per capita (15+) consumption	-.571	1.000	-.596	-.175
	New HIV infections (per 1000 uninfected population)	.751	-.596	1.000	.132
	Estimate of current tobacco use prevalence (%)	.209	-.175	.132	1.000

Fig 6 : Correlations

From the figure 6, the correlations between all the features can depicted.

Furthermore, the hook's distance values do not exceed 0.01 for any of the data entry points for the given dataset so it can

be concluded as there are no outliers in the dataset which may affect the prediction.

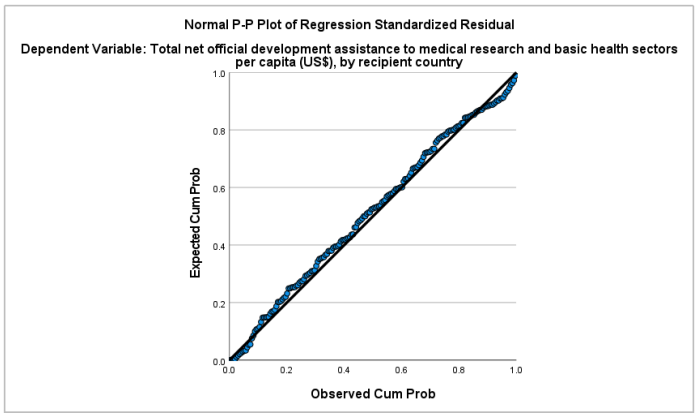

Fig 7 : P-P plot of regression residual

From the figure 7, it is demonstrated that there is linear relationship between the target variable and predictors and both are linearly dependent.

Above analysis deduce that the dataset follows all the assumptions to consider the multiple linear regression analysis to apply on the population data.

In figure 3, the R value is 0.772 which shows a good correlation between the predicted and the present values of the dependent or target variable. The co-efficient of determination (R Square) conclude that the 59.60 percent of the variance in target variable can be deduce using predictors.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6319.048	3	2106.349	90.141	.000[b]
	Residual	4276.199	183	23.367		
	Total	10595.247	186			

a. Dependent Variable: Total net official development assistance to medical research and basic health sectors per capita (US$), by recipient country

b. Predictors: (Constant), Estimate of current tobacco use prevalence (%), New HIV infections (per 1000 uninfected population), Total (recorded+unrecorded) alcohol per capita (15+) consumption

Fig7 : ANOVA table

From the ANOVA table, p value is 0.000 which is less than .05. further, it can be depicted that all the predictors are significantly influencing the target variable. Figure 5, coefficients also confirms that for "Total Alcohol Consumption Per Capita" , "Communication Disease By Country" and "Estimate Of Tobacco Use Prevalence" with significance value less than .05 is better fit and statistically significant to predict the target variable "Total Net Official Development Assistance". Interestingly, the predictor "Total Alcohol Consumption Per Capita " has the negative co-efficient of 0.088 which is inversely affecting the target variable.

VII. CONCLUSION :

The Multiple Regression Analysis carried out on the above dataset can be used to predict the "Total Net Official Development Assistance" for the different countries for the year 2018-19. The analysis confirms that "Total Alcohol Consumption Per Capita" , "Communication Disease By Country" affects slight significantly than "Estimate Of Tobacco Use Prevalence" to target variable "Total Net Official Development Assistance".

This model is affirmatively inclined towards all the assumptions considered for the dataset so this model can be implemented on the population data of this dataset.

VIII. REFERENCES :

1. [1] https://apps.who.int/gho/data/node.main
2. [2]https://apps.who.int/gho/data/node.main.SDG3B?lang=en
3. [3]https://apps.who.int/gho/data/node.main.SDG35?lang=en
4. [4]https://apps.who.int/gho/data/view.main.SDG33v?lang=en
5. [5]https://apps.who.int/gho/data/node.main.SDG3A?lang=en