Multiple Linear Regression Model to Predict Death Rate Caused by Cancer

Saheli Dutta
x21246513student@ncirl.ie
National College of Ireland
Masters in Data Analytics

***Abstract*—Cancer is one of the major causes of death all over the world, accounting for millions of deaths every year. Therefore, understanding the factors associated with cancer mortality is of paramount importance in developing effective prevention and treatment strategies. Cancer mortality rates vary by population, poverty percentage, race/ethnicity, incidence rate, etc. The multiple Linear Regression (MLR) model was implemented on a dataset that had around 3000 records. Several analyses, tests, and assumptions were done to comprehend the data and the relation between the variables. Results show an influential correlation between death rate and incidence rate, poverty percentage, average household size, ethnicity, etc. Although, on the other hand, the correlation between death rate with population, percentage of married households, a certain ethnicity, age, etc is not significant at all. This study can be used further for research purposes to predict the amount of death rate happened by cancer.**

***Index Terms*—Cancer, Death Rate, Multiple Linear Regression, Descriptive Analysis, Regression, Predictor and Predicted variables, Statistical Study.**

I. Introduction

The American Cancer Society evaluates the number of new cancer cases each year and the majority of the deaths in the United States happened due to cancer, These results are analyzing incidence data which is collected by central cancer registries, and mortality data gathered by the National Center for Health Statistics [1].

Generally, The major threat factors for cancer in western (host) countries are smoking, dietary habits, and reproductive behaviors. Transmissible agents are also one of the reasons in economically growing nations. Nevertheless, these practices are transforming very fast. While the percentage of smoking is dropping in economically developed countries. unfortunately, it is increasing in Asia and Africa continents [2].

The number of people who get a particular type of cancer in a specified time period is labeled under the Incidence of cancer. Also written as the number of cancer patients per 100,000 individuals in the population. Rates of cancer incidents rarely change from time to time, Although in a decade the changes are noticeable. Mortality statistics suggest the number of people who have passed away from a certain type of cancer in a year [3].

II. Literature Review

Previous studies have been conducted to deal with several kinds of cancer prediction. Magdy M. Fadel, Nadia G. Elseddeq, Zainab H. Ali and Ali I. Eldesouky utilized the WOA algorithm in classification accuracy. The framework is comprised of a Neural Network, and the input is the optimal set of feature selection layers [4].

Similarly, several researchers utilize various kinds of Machine Learning (ML) algorithms such as k-nearest neighbor (KNN), logistic regression (LR), decision trees (DT), random forest (RF), and support vector machine (SVM).

In this analysis, the Multiple Linear Regression (MLR) model was used to examine the factors that may influence the death rate caused by cancer. This analysis explored the variables and decided which of them should be involved to create the predictive model. On top of that, how likely dependent and independent variables are internally related to each other and check the normal distribution of data, Also, to find the best-fit model some statistical assumptions were made.

III. Methodology

A CSV formatted dataset was provided which consisted of 3048 records of death rate with 23 independent variables(Predictor Variable) and 1 dependent variable(Predicted Variable), This particular predicted variable will be predicted through those predictors.

In this analysis, the independent variables were -

1. **Population** - Population of the county
2. **Incidence Rate** - Rate of incidents which is happening due to cancer.
3. **Med Income** - Median income.
4. **Poverty Percent** - Percentage of poverty
5. **Median Age** - Median age of county residents
6. **Median Age Male** - Median age of male inhabitants.
7. **Median Age Female** - Median age of female inhabitants.
8. **Avg House Hold Size** - Mean household size of a county
9. **Pct Married Households** - Percent of married households
10. **Pct No HS 18-24** - Percent of county citizens aged between 18-24, the highest education is less than high school.
11. **Pct HS 18-24** - Percent of county citizens aged between 18-24, the highest education is not more than a high school diploma.
12. **Pct Batch Deg 18-24** - Percent of county citizens aged between 18-24, the highest education is a bachelor's degree
13. **Pct BachDeg25 Over**: Percent of county citizens aged between 25 and over, highest education is a bachelor's degree
14. **Pct HS 25 Over** - Percent of county citizens ages 25 and over highest education isn't more than a high school diploma
15. **Pct Unemployed-16 Over** - Percent of county citizens

aged 16 and over, the fall under unemployment.
16. **Pct Private Coverage** - Percent of county citizens with a private health coverage
17. **Pct Emp Priv Coverage** - Percent of county citizens with private health coverage (provided by an employee).
18. **Pct Public Coverage** - Percent of county citizens with government-provided health coverage.
19. **Pct Public Coverage Alone** - Percent of county citizens with health coverage(provided by government alone).
20. **Pct White** - Percent of county citizens who recognize as White
21. **Pct Black** - Percent of county citizens who recognize as Black
22. **Pct Asian** - Percent of county residents who recognize as Asian.
23. **Pct Other Race** - Percent of county citizens who recognize in a category that doesn't fall under White, Black, or Asian
24. **County** - Name of the enlisted county

The predicted variable is -
1. **DeathRate**: Dependent or Predicted variable. The rate of death happened due to cancer.

Using the mentioned data, a model having a significant relationship between the dependent variable and independent variables will be designed.
The Multiple Linear Regression was selected to perform statistical analysis because the dataset comprised lots of variables. First of all, The process was initiated by testing the normal distribution of data. If the data is not normally distributed(gaussian distribution), data modification was needed to deal with skewed and peaked data.
Now to check correlation, I used Pearson's r to check information about the dataset, and along with that, it helped to determine the strength of a relationship and access the significant level.
Secondly, F-test and t-test were conducted to check the correlation between the model and dependent variable is statistically significant. Also, to check if independent variables had significant coefficients. Particularly, the calculation of t-statistics and its p-value comes under the assumption that the sample comes from a normal distribution. That's why skewed data was already checked and transformed my data accordingly in the initial phase.
Before reaching to final conclusion several assumptions were made to confirm the dependability and verification of the sample model. Assumptions were the linearity between variables, no multicollinearity, values of residuals are independent, homoscedasticity and no influential cases biased my model. To meet these assumptions I had to recheck and rerun my model multiple times to reach the objective.

Finally, the Multiple linear regression model was shown as in:

$$y_i = \beta_0 + \beta_{i1}x_1 + \beta_2 x_{i2} + \beta_p\ x_{ip} + \epsilon$$

Where (for i = 1,2, ... n.),
y_i = = predicted/dependent variable
x_i = predictor/independent variables
β_0 = y-intercept
β_p = slope coefficients for each independent variable
ϵ = the model's error term (also known as the residuals)

IV. Results and Discussion

A. *Descriptive Statistics*

First of all, To check the normal distribution of variables I did one sample Kolmogorov-Smirnov test and Shapiro - Wilk test [5].
The null hypothesis of these tests is - the normal distribution of variables in the population. A various way to say the exact is that a variable's values are a simple arbitrary selection from a normal distribution. Generally, the null hypothesis gets rejected if p less than 0.05. The initial results showed that all of the variables except Pct Public Coverage were not normally distributed because of the p-value(it was less than 0.05) **(Fig. 1)**.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Population	.379	3047	.000	.257	3047	.000
deathRate	.028	3047	.000	.990	3047	.000
incidenceRate	.043	3047	.000	.939	3047	.000
medIncome	.079	3047	.000	.917	3047	.000
povertyPercent	.066	3047	.000	.954	3047	.000
MedianAge	.414	3047	.000	.141	3047	.000
MedianAgeMale	.041	3047	.000	.994	3047	.000
MedianAgeFemale	.041	3047	.000	.994	3047	.000
AvgHouseholdSize	.091	3047	.000	.928	3047	.000
PctMarriedHouseholds	.055	3047	.000	.979	3047	.000
PctNoHS18_24	.061	3047	.000	.957	3047	.000
PctHS18_24	.028	3047	.000	.995	3047	.000
PctBachDeg18_24	.093	3047	.000	.879	3047	.000
PctHS25_Over	.035	3047	.000	.993	3047	.000
PctBachDeg25_Over	.075	3047	.000	.938	3047	.000
PctUnemployed16_Over	.050	3047	.000	.963	3047	.000
PctPrivateCoverage	.038	3047	.000	.989	3047	.000
PctEmpPrivCoverage	.021	3047	.005	.998	3047	.000
PctPublicCoverage	.013	3047	.200*	.999	3047	.517
PctPublicCoverageAlone	.034	3047	.000	.987	3047	.000
PctWhite	.171	3047	.000	.802	3047	.000
PctBlack	.265	3047	.000	.658	3047	.000
PctAsian	.315	3047	.000	.405	3047	.000
PctOtherRace	.286	3047	.000	.524	3047	.000

*. This is a lower bound of the true significance.
a. Lilliefors Significance Correction

Fig. 1: KS & Shapiro-Wilk Test

SPSS produced a Quantile-Quantile (or QQ) plot to satisfy my practical on matching for normality. The normal Q-Q plots also showed distinct curvature. On top of that, I did the skewness(what extent to a variable is asymmetrically distributed) and Kurtosis (peakedness of the distribution for data) of data [6]. Unfortunately, our data was not properly skewed, the majority of them were either positively skewed or negatively skewed. **(Fig. 2)** is showing the significant kurtosis and skewness chart.
Before moving forward to other analyses, data transformation was needed in order to normalize the data. The minimum values of variables were greater than 0 except for the Pct Other Race variable. I have performed Hyperbolic arcsine on that particular variable(Because there was no limitation on performance in hyperbolic Arcsine)and Logarithmic transformation was done on the rest of the variables [7]. Boxplots of transformed data are displayed in **Fig. 3**. After doing normalization, the significance level of each variable was quickly checked so that I can discard the non-significant variables. This particular dataset consisted of more than 20

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Population	3047	827	10170292	102637.37	329059.221	14.290	.044	337.437	.089
deathRate	3047	59.7	362.8	178.664	27.7515	.275	.044	1.355	.089
incidenceRate	3047	201.3	1206.9	445.654	57.4566	.751	.044	13.794	.089
medIncome	3047	22640	125635	47063.28	12040.091	1.408	.044	3.713	.089
povertyPercent	3047	3.2	47.4	16.878	6.4091	.931	.044	1.276	.089
MedianAge	3047	22.3	624.0	45.272	45.3045	9.990	.044	100.910	.089
MedianAgeMale	3047	22.4	64.7	39.571	5.2260	.132	.044	.676	.089
MedianAgeFemale	3047	22.3	65.7	42.145	5.2928	-.208	.044	.577	.089
AvgHouseholdSize	3047	1.86	3.97	2.5297	.24845	1.297	.044	3.874	.089
PctMarriedHouseholds	3047	22.9924899	78.0753968	51.2438721	6.57281379	-.522	.044	1.414	.089
PctNoHS18_24	3047	.0	64.1	18.224	8.0931	.973	.044	2.211	.089
PctHS18_24	3047	.0	72.5	35.002	9.0697	.179	.044	.534	.089
PctBachDeg18_24	3047	.0	51.8	6.158	4.5291	1.956	.044	9.139	.089
PctHS25_Over	3047	7.5	54.8	34.805	7.0349	-.334	.044	.119	.089
PctBachDeg25_Over	3047	2.5	42.2	13.282	5.3948	1.095	.044	1.737	.089
PctUnemployed16_Over	3047	.4	29.4	7.852	3.4524	.891	.044	2.297	.089
PctPrivateCoverage	3047	22.3	92.3	64.355	10.6471	-.394	.044	-.004	.089
PctEmpPrivCoverage	3047	13.5	70.7	41.196	9.4477	.089	.044	-.302	.089
PctPublicCoverage	3047	11.2	65.1	36.253	7.8417	-.005	.044	-.089	.089
PctPublicCoverageAlone	3047	2.6	46.6	19.240	6.1130	.471	.044	.362	.089
PctWhite	3047	10.1991551	100.000000	83.6452862	16.3800252	-1.681	.044	2.691	.089
PctBlack	3047	.000000000	85.9477986	9.10797761	14.5345379	2.258	.044	5.039	.089
PctAsian	3047	.000000000	42.6194245	1.25396496	2.61027639	7.418	.044	78.397	.089
PctOtherRace	3047	.000000000	41.9302514	1.98352300	3.51771014	4.952	.044	35.537	.089
Valid N (listwise)	3047								

Fig. 2: Descriptive Statistics

independent variables and it was becoming more complex to perform normalization on each of the variables. To make it simpler non significant variables were removed.

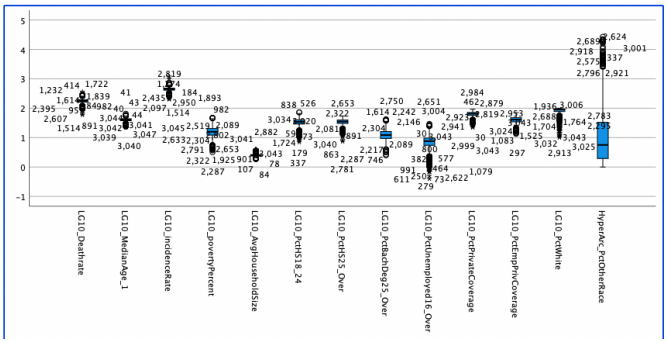

Fig. 3: Boxplot of transformed data

Some examples of Batch by Batch box plots are shown in **Fig. 4** . Because in Fig. 3 they weren't displaying in a proper way, that's why I did detailed box plots based on their range. Here, I have attached some of those boxplots, all of the box plots are attached in the code file.

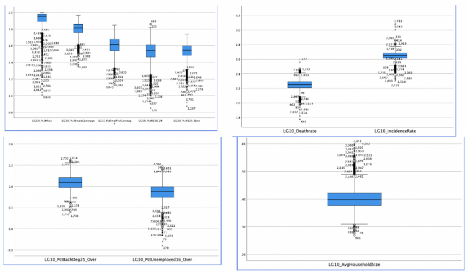

Fig. 4: Detailed boxplot of some of the transformed variables

There was a huge number of outliers in Data when trying to remove them I ended up with fewer data points, which was not advisable at all because I may end up losing crucial data. So, in some cases, I tried to remove unbalanced datapoint by replacing them with the mean of the particular variable.
For instance - Median Age, consisted of 30 plus outliers and the data were not make sense, because the value of the median age was 624, 619.2 etc).
Below is an example of adjusted data -
Median Age-Adjusted - 52.05 51.50 48.35 45.45
Median Age Given - 624.0 619.2 579.6 546.0

The Kolmogorov Smirnov test and Shapiro-Wilk normality test were done for every converted variable, and a normal Q-Q plot with Boxplot and skewness kurtosis were reviewed and made a comparison with the original data. Slight betterment was achieved after cleaning and processing the data. After the variables were successfully altered and reviewed for normal distribution a Multiple Linear Regression (MLR) model was built.

B. Model Building

The Multiple Linear regression model in **Fig. 5** analyzed the significance level for every variable, Population, med income, median age male, median age female, pct married household, pctnohs18-24, pct batch deg18-24, Pct public coverage, pct public coverage alone, pct black, pct Asian has been removed due to their significance level which is not less than 0.05. Also, I removed non-significant variables one by one, and all predictor variables were tested again.
For example, median age can play a significant role in predicting death rate that's why the median age of male and female was removed except the median-age column and again tested their significance level and it is less than 0.05. This suggested that these independent variables can be excluded from the model due to their less preference.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	175.945	15.509		11.344	.000		
	Population	-1.678E-6	.000	-.020	-1.356	.175	.705	1.419
	incidenceRate	.205	.007	.424	31.139	.000	.817	1.225
	medIncome	6.616E-5	.000	.029	.848	.397	.132	7.562
	povertyPercent	.367	.144	.085	2.551	.011	.137	7.287
	MedianAge	-.003	.008	-.005	-.387	.699	.978	1.022
	MedianAgeMale	-.220	.197	-.041	-1.115	.265	.110	9.059
	MedianAgeFemale	-.284	.216	-.054	-1.314	.189	.090	11.173
	AvgHouseholdSize	-16.033	2.710	-.144	-5.916	.000	.258	3.883
	PctMarriedHouseholds	.038	.098	.009	.392	.695	.282	3.549
	PctNoHS18_24	-.087	.054	-.025	-1.589	.112	.600	1.666
	PctHS18_24	.236	.048	.077	4.938	.000	.622	1.608
	PctBachDeg18_24	.014	.105	.002	.130	.896	.517	1.935
	PctHS25_Over	.323	.094	.082	3.448	.001	.269	3.712
	PctBachDeg25_Over	-1.246	.149	-.242	-8.366	.000	.181	5.533
	PctUnemployed16_Over	.416	.157	.052	2.652	.008	.398	2.512
	PctPrivateCoverage	-.675	.130	-.259	-5.197	.000	.061	16.391
	PctEmpPrivCoverage	.371	.098	.126	3.779	.000	.135	7.384
	PctPublicCoverage	-.091	.211	-.026	-.431	.666	.043	23.475
	PctPublicCoverageAlone	.226	.266	.050	.849	.396	.044	22.667
	PctWhite	-.162	.057	-.096	-2.848	.004	.134	7.442
	PctBlack	-.071	.054	-.037	-1.317	.188	.190	5.251
	PctAsian	-.027	.183	-.003	-.148	.882	.512	1.951
	PctOtherRace	-.879	.121	-.111	-7.279	.000	.648	1.544

a. Dependent Variable: deathRate

Fig. 5: Independent variables before transformation and their significance level(P)

Coming to the rest of the variables, the p-values for median-age, incidence-rate, poverty-per-cent, average-household-size, pct-hs-18-24, pct-hs-25-over, pct batch deg 25 over, pct-unemployed-16-over, pct-private-coverage, pct-emp-private-coverage, pct-white, pct-other-race were less than 0.05. This demonstrated a statistically significant relationship between

the death rate and this group of variables. **Fig. 6** shows their corresponding significance level.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	1.147	.095		12.031	.000		
	LG10_MedianAge_1	-.121	.025	-.099	-4.752	.000	.353	2.829
	LG10_incidenceRate	.524	.016	.435	32.173	.000	.842	1.188
	LG10_povertyPercent	.041	.012	.099	3.579	.000	.200	4.996
	LG10_AvgHouseholdSize	-.168	.035	-.099	-4.820	.000	.363	2.756
	LG10_PctHS18_24	.046	.008	.082	5.488	.000	.685	1.459
	LG10_PctHS25_Over	.098	.015	.139	6.377	.000	.325	3.082
	LG10_PctBachDeg25_Over	-.075	.010	-.189	-7.670	.000	.254	3.931
	LG10_PctUnemployed16_Over	.025	.005	.080	4.722	.000	.531	1.884
	LG10_PctPrivateCoverage	-.146	.031	-.163	-4.739	.000	.129	7.741
	LG10_PctEmpPrivCoverage	.061	.017	.093	3.515	.000	.219	4.573
	LG10_PctWhite	-.033	.009	-.056	-3.581	.000	.630	1.587
	HyperArc_PctOtherRace	-.007	.001	-.087	-5.801	.000	.684	1.462

a. Dependent Variable: LG10_Deathrate

Fig. 6: Transformed independent variables with their significance level(P) less than 0.05

- **VIF**

In regression analysis variance inflation factor (VIF) is used to estimate the amount of multicollinearity. When there is a correlation between numerous independent variables in a multiple regression model then this can adversely impact the regression outcomes. If an independent variable has a large VIF value then it reveals a highly collinear relationship with other variables. In that case, the selection of independent variables must be reconsidered. [9].

Normally, if the VIF value is on the higher side, then there is more possibility of multicollinearity exists. There is considerable multicollinearity that must be fixed when VIF is greater than 10. From **Fig. 6** it can be concluded that none of the variable's **VIF was more than 10**, which means the variables are not highly correlated also the **tolerance scores are not lower than 0.10**, which means there is **no multicollinearity** in my data.

- **Correlations table**

In this particular table, Pearson's r Correlation of more than 0.8 may indicate a multicollinearity problem. For my data (**Fig. 7**), this was not an issue, as the corresponding analysis were reviewed to make sure of their collinearity.

Correlations

		LG10_Deathrate	LG10_MedianAge_1	LG10_incidenceRate	LG10_povertyPercent	LG10_AvgHouseholdSize	LG10_PctHS18_24	LG10_PctHS25_Over	LG10_PctBachDeg25_Over	LG10_PctUnemployed16_Over	LG10_PctPrivateCoverage	LG10_PctEmpPrivCoverage	LG10_PctWhite	HyperArc_PctOtherRace
Pearson Correlation	LG10_Deathrate	1.000	.015	.521	.420	-.027	.258	.425	-.482	.369	-.337	-.225	-.180	-.223
	LG10_MedianAge_1	.015	1.000	-.001	-.153	-.823	.272	.362	-.148	-.149	.092	-.187	.348	-.359
	LG10_incidenceRate	.521	-.001	1.000	.050	-.108	.069	.144	-.059	.179	.070	.124	-.041	-.191
	LG10_povertyPercent	.420	-.153	.050	1.000	.090	.091	.260	-.604	.578	-.807	-.694	-.436	.030
	LG10_AvgHouseholdSize	-.027	.823	-.108	.090	1.000	.060	-.176	-.074	.279	-.303	-.007	-.334	.414
	LG10_PctHS18_24	.258	.272	.069	.091	.060	1.000	.431	-.396	.160	-.234	-.219	.019	-.069
	LG10_PctHS25_Over	.425	.362	.144	.260	-.176	.431	1.000	-.703	.076	-.209	-.211	.135	-.401
	LG10_PctBachDeg25_Over	-.482	-.148	-.059	-.604	-.074	-.396	-.703	1.000	-.373	.610	.507	.095	.137
	LG10_PctUnemployed16_Over	.369	-.149	.179	.578	.279	.160	.076	-.373	1.000	-.564	-.385	-.404	.078
	LG10_PctPrivateCoverage	-.337	.092	.070	-.807	-.303	-.234	-.209	.610	-.564	1.000	.825	.460	-.145
	LG10_PctEmpPrivCoverage	-.225	-.187	.124	-.694	-.007	-.219	-.211	.507	-.385	.825	1.000	.308	-.035
	LG10_PctWhite	-.180	.348	-.041	-.436	-.334	.019	.135	.095	-.404	.460	.308	1.000	-.177
	HyperArc_PctOtherRace	-.223	-.358	-.191	.030	.414	-.069	-.401	.137	.078	-.145	-.035	-.177	1.000

Fig. 7: Correlations between variables

- **Model Summary Table**

Coming to the Model Summary table, I utilized the **Durbin-Watson** statistic which helped to make the assumption that residuals are not related. This value can range from 0 to 4, Ideally, for Independent residuals, the Durbin-Watson value must be close to 2 to meet my assumption. Values below 1 and above 3 are reasons for invalid analysis. In this case,

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics R Square Change	Change Statistics F Change	Change Statistics df1	Change Statistics df2	Change Statistics Sig. F Change	Durbin-Watson
1	.731[a]	.534	.532	.04724	.534	289.311	12	3033	.000	1.975

a. Predictors: (Constant), HyperArc_PctOtherRace, LG10_povertyPercent, LG10_PctHS18_24, LG10_incidenceRate, LG10_AvgHouseholdSize, LG10_PctWhite, LG10_PctHS25_Over, LG10_PctUnemployed16_Over, LG10_PctEmpPrivCoverage, LG10_MedianAge_1, LG10_PctBachDeg25_Over, LG10_PctPrivateCoverage

b. Dependent Variable: LG10_Deathrate

Fig. 8: Model Summary to describe R square and Durbin-Watson Value

the value is **1.975** which is close to **2 (Fig. 8)**. R-Square is a statistical measure of fit that denotes the proportion of variance in a dependent variable(in the regression model) which is explained by independent variables. Adjusted R-squared changes the statistic based on the independent variables' number in the model. The r-Square value on the higher indicates a better fit of the model because more of the variation in the dependent variable is explained by the independent variables. Adjusted R-Square penalizes adding more variables that do not enhance the accuracy of an existing model.

In my report, the R-Square is 0.534 which indicated that 53.4% of the variance in death rate was predicted from the independent variables and the adjusted R-Square value of 0.532 indicated that 53.2% of the variance in death rate was predicted from the independent variables. So R-Square and adjusted R-Square are almost similar in my model which indicates a good fit **(Fig. 8)**.

- **ANOVA**

The ANOVA table analyzes the variance in the model **(fig. 9)**.

Regression df(Degrees of Freedom) denotes the number of independent variables in a regression model. In my data, I considered 12 independent variables, which is why it is 12.

Residual df denotes the total number of rows(observations) of the dataset deducted by the number of variables being estimated. In my report, 3033 is the residual df.

Total df denotes the sum of both regression and residual degrees of freedom(i.e. dataset's size minus 1). In my report, 3045 is the total df [11].

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7.748	12	.646	289.311	.000[b]
	Residual	6.768	3033	.002		
	Total	14.516	3045			

a. Dependent Variable: LG10_Deathrate

b. Predictors: (Constant), HyperArc_PctOtherRace, LG10_povertyPercent, LG10_PctHS18_24, LG10_incidenceRate, LG10_AvgHouseholdSize, LG10_PctWhite, LG10_PctHS25_Over, LG10_PctUnemployed16_Over, LG10_PctEmpPrivCoverage, LG10_MedianAge_1, LG10_PctBachDeg25_Over, LG10_PctPrivateCoverage

Fig. 9: ANOVA table to describe F-test and its significance

Sum of Squares (SS)

Regression SS denotes the total variation in the dependent variable that is defined by the regression model.

From ANOVA results in **Fig. 9**, the regression Sum of Squares is 7.748 and the total Sum of Squares is 14.516. It represented the R-Square term is equivalent to 0.533, which means the regression model explained around 53.3 % of the variability in the predicted variable explained by the independent variable.
Residual SS(Error Sum of Squares) denotes the total variation in the dependent or predicted variable that is left unexplained by the regression model. **From the ANOVA table**, the residual SS is about **6.678**. Generally, the less the error, the more satisfactory the regression model explains the variation in the data set.

In the **ANOVA table** for the Death Rate data, the **F statistic** is equal to 289.311. The distribution is F(12, 3033), and the significance level is less than 0.001. That means the possibility of observing a value greater than or equal to 289.311 is less than 0.001. Hence, there is a significant linear relationship between the dependent variable and independent variables.

- **Coefficients Table**

Next look at the coefficients table(**Fig. 10**). The regression **intercept** (tagged Constant in SPSS) is equal to 1.147, which means it is the predicted value of the death rate when all of the predictors assume the 0 value.
For instance, The **regression slope**, or unstandardized coefficient, (B in SPSS) is equal to **0.041** which denotes the amount by which the death rate can be predicted for the increase of 1 unit in the poverty percent when all other predictors remain constant.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	1.147	.095		12.031	.000		
	LG10_MedianAge_1	-.121	.025	-.099	-4.752	.000	.353	2.829
	LG10_incidenceRate	.524	.016	.435	32.173	.000	.842	1.188
	LG10_povertyPercent	.041	.012	.099	3.579	.000	.200	4.996
	LG10_AvgHouseholdSize	-.168	.035	-.099	-4.820	.000	.363	2.756
	LG10_PctHS18_24	.046	.008	.082	5.488	.000	.685	1.459
	LG10_PctHS25_Over	.098	.015	.139	6.377	.000	.325	3.082
	LG10_PctBachDeg25_Over	-.075	.010	-.189	-7.670	.000	.254	3.931
	LG10_PctUnemployed16_Over	.025	.005	.080	4.722	.000	.531	1.884
	LG10_PctPrivateCoverage	-.146	.031	-.163	-4.739	.000	.129	7.741
	LG10_PctEmpPrivCoverage	.061	.017	.093	3.515	.000	.219	4.573
	LG10_PctWhite	-.033	.009	-.056	-3.581	.000	.630	1.587
	HyperArc_PctOtherRace	-.007	.001	-.087	-5.801	.000	.684	1.462

a. Dependent Variable: LG10_Deathrate

Fig. 10: Coefficients Table to describe the regression model

Test statistics is utilized for the test of the significance of the coefficients which are declared beneath the t column **B / Std. Error**.
For the **slope value on Incidence Rate**, the **t statistic** is 32.173. The null hypothesis is that the slope is 0. this value can be resembled with a t distribution to test the null hypothesis. The resulting p-value for the test under the Sig. column. The p-value (quoted under Sig.) is .000 (reported as p less than .001) which is less than 0.05. Therefore I have significant evidence to reject the null hypothesis that the slope coefficient on the Incidence Rate is zero.
Also checked if the intercept is different from zero. For the intercept here the **t statistic** is **12.031** and the p-value (quoted under Sig.) is .000 (reported as p less than .001) which is less than 0.05. Therefore, I have the proof to reject the null hypothesis(intercept is zero).

- **Homoscedasticity**

Equal distribution of residuals is known as Homoscedasticity, which is simply whether the residuals lean to group together at some data points, On the other hand, some values spread apart. To visualize homoscedasticity it should look like a **shotgun blast** of **randomly dispersed** data [11].
The contrary of homoscedasticity is heteroscedasticity, if data looks cone-shaped or fan-shaped**(Fig. 11)** then this will be for sure heteroscedastic in nature. The below figure demonstrated that the variance of residuals was kind of fan-shaped which means residuals are not equally distributed.

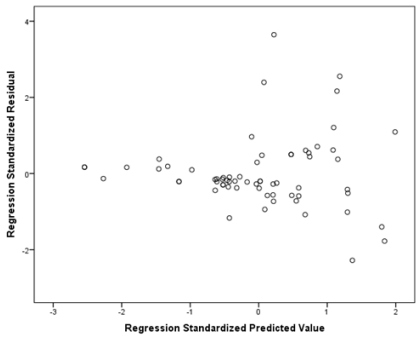

Fig. 11: Example of heteroscedasticity - The graph is funnel-shaped

The independent variable in the regression has a straight line which indicates a linear relationship with the predicted or outcome variable.
In this analysis, the scatter plot is not funnel-shaped which

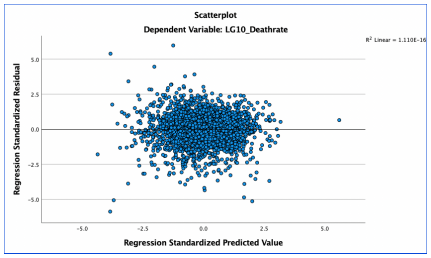

Fig. 12: Scatterplot of Residuals

means the assumption of homoscedasticity was fulfilled. **Fig. 12** is showing the significant graph of my analysis.

- **Linearity**

Linearity means the relationship between the predictor variables and the predicted variable should be linear. To meet linearity assumption the relationship between the dependent variable (death rate) and other independent variables must be defined by straight lines. **Fig. 13** shows some of the **significant curvatures** where the relationship between these variables was linear.

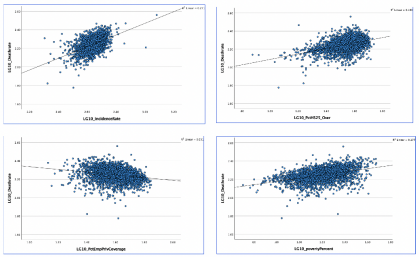

Fig. 13: Scatter Plot of transformed Dependent Variable and Independent Variables

- **Residuals should be normally distributed**

The probability plot (P-P plot) helps in statistical analysis to check the normal distribution of residuals. A perfect P-P Plot means Regression Standardised Residual should have data points near the diagonal line which is shown in **Fig. 14** Generally, the closer the data points lie to the diagonal line, the residuals will tend to be more normally distributed.

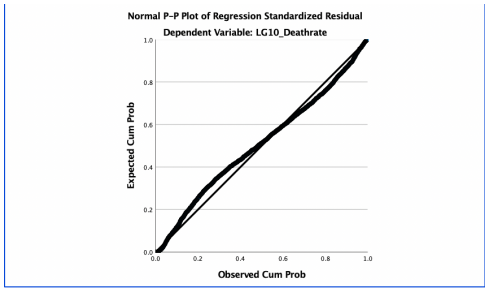

Fig. 14: Normal P-P Plot of Regression Standardised Residual

In **Fig. 14** shows that some of the data points lie on the slope while some of the data points are very much close to the slope. Nonetheless, looking at other diagnoses this result was still valid.

- **No influential case biasing the model**

The common measurement of a data point's influence on analysis is known as Cok's distance (Cook's D). If the value of the Cook's distance is 1 then it considers being on the higher side.

The data point is viewed as highly influential when the cook's distance flagged that particular data point [12].

In SPSS, a separate column was created in the data file. Minimum and Maximum values can be observed in **Fig. 15** where the maximum value was **0.74 (didn't exceed 1)**. Because the cook's distance maximum value of more than 1 denotes significant outliers, it may have an unnecessary impact on the regression model.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	2.0269	2.5288	2.2467	.05044	3046
Std. Predicted Value	-4.357	5.593	.000	1.000	3046
Standard Error of Predicted Value	.001	.012	.003	.001	3046
Adjusted Predicted Value	2.0277	2.5280	2.2467	.05045	3046
Residual	-.27704	.28217	.00000	.04715	3046
Std. Residual	-5.865	5.973	.000	.998	3046
Stud. Residual	-5.909	5.992	.000	1.001	3046
Deleted Residual	-.28121	.28400	.00000	.04746	3046
Stud. Deleted Residual	-5.942	6.027	.000	1.002	3046
Mahal. Distance	1.176	190.021	11.996	12.241	3046
Cook's Distance	.000	.074	.001	.003	3046
Centered Leverage Value	.000	.062	.004	.004	3046

a. Dependent Variable: LG10_Deathrate

Fig. 15: Residual statistics

C. Conclusion

- **Assumption 1**: **Linear relationship must exist between the independent variables and the dependent variable.** ✓
  Scatterplots show this assumption had been met.
- **Assumption 2**: **Absence of multicollinearity in the data.** ✓
  By analyzing the coefficients box in SPSS, it was clear that this assumption was satisfied, Because the VIF scores were not more than 10, and tolerance scores were above 0.1. Also, Pearson's correlation was checked for the same.
- **Assumption 3**: **The values of the residuals are independent.**✓
  The Durbin-Watson value indicated that this assumption had been met Because the value of Durbin-Watson was close to 2 (Durbin-Watson = 1.97).
- **Assumption 4**: **The variance of the residuals is constant (Homoscedasticity)**. ✓
  The scatter plot between standardized residuals and standardized predicted values did not look funnel or cone-shaped, which means the assumption of homoscedasticity had been met.
- **Assumption 5**: **The values of the residuals are normally distributed**. ✓
  The P-P plot(Probability Plot) indicated that the assumption might have been narrowly violated for the model. Nonetheless, only disproportionate tangents from normality are likely to have a considerable influence on the results, the results are likely still valid.
- **Assumption 6**: **Absence of influential cases which can bias my model (Cook's Distance).** ✓
  Cook's Distance values were all under 1 which means individual data points did not disproportionately affect my model.

The derived statistical summary and assumption results clearly proved that there was enough significant correlation between the death rate and other independent variables. On top of that, my regression model satisfied all of the assumptions which are also known as Gauss-Markov Assumptions.

V. Acknowledgement

I would really like to express my gratitude towards lecturer Hicham Rifai for his valuable guidance also organized a class to clear our doubts which was very helpful for my concluded analysis.

References

[1] Rebecca L Siegel, Kimberly D Miller, Nikita Wagle, Ahmedin Jamal, "Cancer statistics, 2023", [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/36633525/

[2] Carol DeSantis, Elizabeth M. Ward, Melissa M. Center, Ahmedin Jamal, "Cancer statistics, 2023", [Online]. Available: https://aacrjournals.org/cebp/article/19/8/1893/68607/Global-Patterns-of-Cancer-Incidence-and-Mortality/

[3] National Cancer Control Programme [Online]. Available: https://www.hse.ie/eng/services/list/5/cancer/pubs/intelligence/registrydata.html/

[4] Magdy M. Fadel, Nadia G. Elseddeq, Zainab H. Ali, "A Fast Accurate Deep Learning Framework for Prediction of All Cancer Types" [Online]. Available: https://ieeexplore.ieee.org/document/9951602/authorsauthors/

[5] "SPSS tutorial by IBM" [Online]. Available: https://www.spss-tutorials.com/spss-shapiro-wilk-test-for-normality/ https://www.spss-tutorials.com/spss-kolmogorov-smirnov-test-for-normality//

[6] "SPSS tutorial by IBM" [Online]. Available: https://www.spss-tutorials.com/skewness/ https://www.spss-tutorials.com/kurtosis/ /

[7] "SPSS tutorial by IBM" [Online]. Available: https://www.spss-tutorials.com/normalizing-variable-transformations//

[8] "Article on VIF " [Online]. Available: https://www.investopedia.com/terms/v/variance-inflation-factor.asptoc-formula-and-calculation-of-vif/

[9] "How to Read a Regression Table " [Online]. Available: https://www.freecodecamp.org/news/https-medium-com-sharadvm-how-to-read-a-regression-table-661d391e9bd7-708e75efc560//

[10] "Testing Assumptions of Linear Regression in SPSS " [Online]. Available: https://www.statisticssolutions.com/testing-assumptions-of-linear-regression-in-spss//

[11] "Testing Assumptions of Linear Regression in SPSS " [Online]. Available: https://www.statisticssolutions.com/testing-assumptions-of-linear-regression-in-spss//

[12] "Testing Assumptions of Linear Regression in SPSS " [Online]. Available: https://lymielynn.medium.com/a-little-closer-to-cooks-distance-e8cc923a3250: :text=In/