Multiple Linear Regression Model to Predict Death Rate Of Cancer Patients

Debmalya Deb
Msc in Data Analytics(School of Computing)
National College of Ireland, Dublin 1
x21242101@student.ncirl.ie

***Abstract*—The disease of Cancer is a state where the body cells in a human body develop and multiply uncontrollably. Cells can intrude and kill surrounding healthy tissue and organs. The Multiple Linear Regression(MLR) Model is a statistical method utilised to analyse and predict the death rate of cancer patients based on multiple factors. For this analysis, the dependent variable is the death rate, and the independent variables include various factors that can affect the death rate, such as incidence rate, poverty percent, avghouseholdSize, PctPrivateCoverage, PctPrivateCoverage etc. In this purpose, a Multiple Linear Regression (MLR) model has been studied using a delivered dataset of around 3000 patients. The model is developed by analysing a dataset of cancer patients and their corresponding death rates and then using this data to create a predictive model.**

***Index Terms*—death rate, multiple regression study, multiple linear regression, statistical methodology, predictive model, explanatory statistics**

I. Introduction

Since before medical documents were saved, cancer as a disorder has been defined in the past treatment. The most premature known reports of cancer occur in seven papyri, found and solved late in 19th century. They delivered the foremost natural understanding of Egyptian medical training. Two of them were known as the ”Edwin Smith” and ”George Ebers” papyri, which include definitions of cancer registered around 1600 B.C. and are accepted till date from authorities as earlier as 2500 B.C. [1].

Cancer study is the most useful means to develop new wisdom to advance oncology practice. Largely, in two types of clinical investigations: practical and observational. Observational investigations are launched without a detailed intervention. This can be the future. A vital study guide should be reprehensibility with increased reality to mark residents of welfare, and transfer ability to clinical trial. When initial idea expecting is desired, it is desirable for a clinician to fast employ shared bio statistician associates to undervalue bias, enhance statistical dominion, and deliver robust measures of product size and other ideal parameters [2].

II. LITERATURE REVIEW

A calculated 44,000 individuals in the Republic Of Ireland get cancer in each year. This picture is composed of both cancers that can flare above the area they formed and cancers that do not spread. It retains non-melanoma skin cancers (a typical type of cancer in ROI). The most current research displays that there are now larger than 24,000 cases out of them 13,027 are men and 11,299 are women, diagnosed per year. This does not enclose non-invasive cancers like non-melanoma skin cancers. The most current news from the National Cancer Registry of Ireland demonstrates that the Covid-19 pandemic resulted in a 10 percent decline in cancer diagnoses. [3].

Each year, the American Cancer Society calculates the number of the latest cancer patients and demises in the United States and gathers the most current data on population-based cancer events and consequences. Data gathered in 2018 were collected by the Management, Epidemiology; the National Program of Cancer Registries; and the North American Association of Central Cancer Registries. Death rate data for 2019 were compiled by the National Center for Health Statistics. In the year 2022, 1,918,030 new cancer cases and 609,360 cancer deaths were launched to appear in the United States, including about 350 deaths per day from lung cancer. [4].

In this project, I have used **Multiple Linear Regression** (MLR) to investigate the independent variables that may affect the death rate caused by cancer. This report is being investigated what variables(dependent/independent) have to be taken to start the model analysis and to what extent are they connected with each other. I have implemented 6 statistical assumptions and methodologies to find the best-fit model.

III. METHODOLOGY

I have been provided with a dataset to analyse the Death-Rate caused by cancer in various counties in the country of the United States Of America in a .csv format, cancer.csv. Which consists of 25 variables out of which 24 are independent variables and 1 dependent variable. In this dataset, the death rate caused by cancer depends upon various factors, just as follows,

1. **Population** - Population of the county
2. **Incidence Rate** - Mean of cancer diagnoses per capita 100000
3. **Med Income** - Per county median earning
4. **Poverty Percent** - In the county percentage of the poverty
5. **Median Age** - Median age of citizens

6. **Median Age Male** - Median of Male citizens in the county
7. **Median Age Female** - Median age of female county residents
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** - Citizens percentage of ages between 18-24 attended highest education, smaller than high school
11. **Pct HS 18-24** - Citizens percentage of ages between 18-24 attended the highest education, in their high school diploma
12. **Pct Batch Deg 18-24** - Citizens percentage of ages between 18-24 attended the highest education, in their bachelor's degree
13. **PctBachDeg25-Over** - Citizens percentage of ages 25 and over with the highest education, bachelor's degree
14. **Pct HS 25 Over** - Citizens percentage of ages 25 and over with the highest education in the high school diploma
15. **Pct Unemployed-16 Over** - Citizens percentage of ages 16 and above those jobless
16. **Pct Private Coverage** - Citizens percentage those are having private health insurance
17. **Pct Emp Priv Coverage** - Citizens percentage of ages those employer-provided health insurance
18. **Pct Public Coverage** - Citizens percentage of ages those are have government-provided health insurance
19. **Pct Public Coverage Alone** - Citizens percentage of ages those are have government-provided health insurance only
20. **Pct White** - White citizen percentage
21. **Pct Black** - Black citizen percentage
22. **Pct Asian** - Asian citizen percentage
23. **Pct Other Race** - Citizen percentage other than white, black and asian
24. **County** - Name of the counties in the states

Using these many independent variables I have to analyse the **Deathrate** in the States which is my **Dependent Variable** in this study. With these 24 independent variables, I have to build a statistical model which will analyse the death rate (dependent variable) in various counties in the States. As the dataset consists of multiple variables that's why I am selecting **Multiple Linear Regression** models to perform this study.

First, the process has been started by checking the **Normality** of the provided data by several tests which come under **Descriptive Statistics**. I have found the given cancer data are not normally distributed that's why data transformation is needed to get the best-fit model. To check the relationships between the dependent variable with the independent variables **Pearson's Correlations** check has been performed and checked if the datasets are fallen under the proper range or not. Further to check the proper significant level the **Coefficient** test has been gone through. To check if any independent variable has a proper significant level **F-test(ANOVA) and t-test** are conducted. To find the Best fit model multiple independent variables have been eliminated and the elimination is followed at the proper significant level under coefficient check. **Gauss–Markov** assumption is taken out to provide the dependability and proof of the model to the entire given dataset. These assumptions are not in presence of multicollinearity, homoscedasticity, independent errors, and normally distributed residuals. After testing multiple tests the best-fit model is not coming up with the normally distributed residuals. Hence again I have transformed the data and restarted the analysis. The Multiple Linear Regression model shows,

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} 2 + ... + \beta_p x_{ip} + \epsilon$

where i = n number of observations

y_i = predictive of dependent variable

x_i = independent variables

β_0 = intercept in y-axis (constant value)

β_p = Independent variable's slope coefficient

ϵ = residuals also knows as error of the model

IV. DISCUSSION AND RESULTS

A. *Normal Distribution*

At the beginning of the Descriptive Statistics, the first Normality test has been performed. In Fig 1, the **Kolmogorov-Smirnov(KS) test and Shapiro-Wilk** normality test has been performed in each variable. The KS test is a non-parametric fit test and it is used to determine if two distributions vary and also whether an underlying probability distribution differs from a hypothesized distribution [5].On the other hand, the Shapiro-Wilk testing is carried out hypothesis test which is used for a model whose null hypothesis is that the model has been developed from a normal distribution [6]. When the P-value is down, then and there we can reject the null hypothesis and clearly state that the model has not been developed from a normal distribution.

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

Fig. 1. Test of Normality

B. *Descriptive Statistics*

In the initial Descriptive Statistics analysis, from Fig 2 below, I find the **Skewness and Kurtosis** values are not normally distributed. Further, those variables are either positively

or negatively skewed. In statistics, skewness measurement is not the symmetry of the probability distribution of a random variable about its mean. In additional words, skewness means you the quantity and demand of skew (out from flat symmetry). The skewness significance could be both positive or negative even faint. If the skewness value is 0, that means the data are completely proportional, even though it is rather unlikely for the actual data. Whereas Kurtosis explains the height and sharpness of the major or central peak. [7].

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
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

C. *Box Plot Analysis*

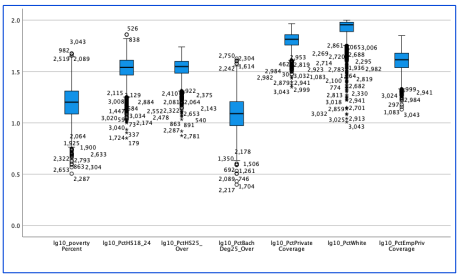

Fig. 3. Box Plot 1

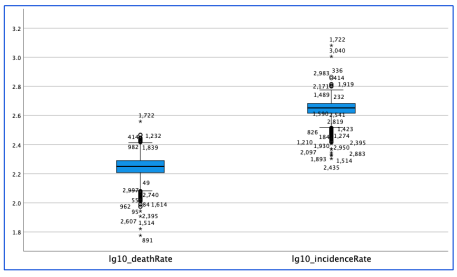

Fig. 4. Box Plot 2

To further strengthen the study **Box-plot** analysis has been performed to check the **outliers**. And in the beginning of the analysis, I find many outliers in the raw data. For that, I have

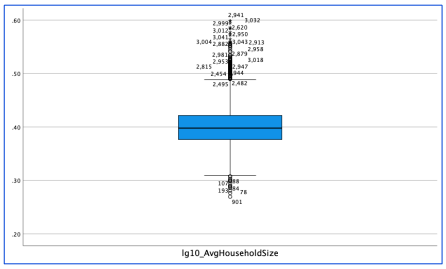

Fig. 5. Box Plot 3

transformed the data using the log transform method. The selected variables' values are more than 0. Therefore **log10** transformation has been carried out. After the transformation of the variables, to each and every converted data, the Kolmogorov Smirnov test and Shapiro-Wilk normality test, Boxplot and normal Q- Q plot are checked and corresponded with the original data provided in the actual .csv. The test and checksum have been carried out until the distribution requirements are met. The Multiple Regression Model has been started after successfully transforming the dataset by eliminating the proper independent variables and after checking the proper normal distribution.

D. *Multiple Linear Regression*

*1) **Coefficient**:* At the beginning of the analysis, a **significant level/P-value** from the coefficient table has been taken out. The below table clearly shows that many independent variables have a p-value which is greater than 0.05. Variables having a significance value of more than 0.05 have been eliminated from this analysis to get the best model. The variables which I have removed from the analysis as the p-value are greater than 0.05 are Population, medincome, medianAge, medianAgemale, medianAgeFemale, PctMarriedHouseholds, PctNoHS18-24, PctBachDeg18-24, PctUnemployed16-Over, PctPublicCoverage, PctPublicCoverageAlone, PctBlack, PctAsian.

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. 6. Coefficient Table Before Elimination and Transformation

Furthermore, I have seen in the Descriptive Statistic model that the independent variable PctOtherRace has a **kurtosis**

value of 35.537 which is very large and as it has many outliers along with that it has some **ZERO** value that's why eliminating this variable.

PctOtherRace	Mean		1.98352300	.063727051
	95% Confidence Interval for Mean	Lower Bound	1.85857063	
		Upper Bound	2.10847538	
	5% Trimmed Mean		1.43609581	
	Median		.826185211	
	Variance		12.374	
	Std. Deviation		3.51771014	
	Minimum		.000000000	
	Maximum		41.9302514	
	Range		41.9302514	
	Interquartile Range		1.88472912	
	Skewness		4.952	.044
	Kurtosis		35.537	.089

Fig. 7. Skewness and Kurtosis Check from Descriptive Statistics

Now the following variables incedenceRate, povertyPercent, avghouseholdSize, PctHS18-24, PctHS25-Over, PctBachDeg25-over, PctPrivateCoverage, PctEmpPrivCoverage,PctWhite can be considered as the group of independent variables which can reliably predict the deathrate as the p-value after log10 transformation is less than 0.05.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	.791	.073		10.774	.000		
	lg10_incidenceRate	.554	.016	.460	34.786	.000	.900	1.111
	lg10_povertyPercent	.075	.010	.179	7.311	.000	.262	3.810
	lg10_AvgHouseholdSize	-.100	.027	-.059	-3.647	.000	.600	1.666
	lg10_PctHS18_24	.040	.008	.072	4.853	.000	.722	1.385
	lg10_PctHS25_Over	.104	.015	.148	7.202	.000	.375	2.668
	lg10_PctBachDeg25_Over	-.073	.010	-.182	-7.350	.000	.257	3.891
	lg10_PctPrivateCoverage	-.148	.031	-.166	-4.800	.000	.132	7.598
	lg10_PctWhite	-.044	.009	-.073	-4.684	.000	.645	1.550
	lg10_PctEmpPrivCoverage	.092	.017	.140	5.513	.000	.245	4.087

a. Dependent Variable: lg10_deathRate

Fig. 8. Coefficient Table After Elimination and Transformation

- **Variance Inflation Factor (VIF)**

Variance inflation factor or VIF is a measurement of the portion of **Multicollinearity** in a multiple regression analysis. Multicollinearity lives when there is inter-correlation between numerous independent variables in the multiple regression model. A big VIF noticed on the independent variable reveals a positively collinear connection to the further variables that should be evaluated or adapted for in the design of the regression model. [8]. At the moment the Variance inflation factor is more elevated than 10 and the tolerance is inferior than 0.1, there is meaningful multicollinearity that ought to be fixed [9]. In the shown **Fig 8**, the VIF score is less than 10 and collinearity tolerance is less than 0.01 for all of the independent variables. This means the variables can be taken to predict the deathrate.

- **Intercept and Slope**

In further analysis, the coefficients table(**Fig. 8**) provides the most fascinating facts about the regression model. Firstly, the **regression intercept which is constant** brings the data value 0.791 and is the indicated value of the death rate when all of the independent variables take the value 0. Just for, The **regression slope**, or unstandardised coefficient takes a value of 0.554 and is the amount by which the death rate can be predicted for the increase of 1 unit when all other predictors remain constant.

To analyse the significance level of the coefficients I have to form a statistical test which is documented under column t. These are simply B divided by Standard Error(**B/std. error**). Now the slope of the Incidence Rate and the t value is 34.786 and this value can be approximated with the t distribution to check the null hypothesis for that the slope is zero. The final p-value of this test is under the significance column. The p-value is 0.000 which is less than 0.05. Therefore the null hypothesis has been **rejected** so that the slope coefficient on the Incidence Rate is zero.
However the intercept is different from zero. The t-value for the intercept is 10.774 and the p-value is 0.000 (written p-value less than 0.001) which is less than 0.05. Lastly, I can say that with proof to reject the null hypothesis where the y-axis intercept is zero [10].

*2) **R Square**:* The R square also known as the coefficient of determination is a standard that delivers knowledge regarding how a model fits in a better manner. According to the context of regression analysis, it is a statistical measurement of how nicely the regression line matches the original data. Therefore it is very crucial for a statistical model to be used either to predict forthcoming results in the hypothesis testing [11]. In this study, the R square value is 0.523 which is 52.3%. This means a 52.3% deathrate can be predicted from the predictor variables, incedenceRate, povertyPercent, avghouseholdSize, PctHS18-24, PctHS25-Over, PctBachDeg25-over, PctPrivateCoverage, PctEmpPrivCoverage, PctWhite.

Adding one more point in this study, the Standard Estimator Error also known as Residual Standard Error is 0.04777. This means that the average by which the real deathrate of cancer in the supplied dataset varies from what the analysis forecasts. In another way on an average, the analysed model reaches its prediction wrong by a value of 0.04777 deathrate.

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	.723[a]	.523	.521	.04777	.523	369.365	9	3036	.000	1.989

a. Predictors: (Constant), lg10_PctEmpPrivCoverage, lg10_AvgHouseholdSize, lg10_incidenceRate, lg10_PctHS18_24, lg10_PctWhite, lg10_PctHS25_Over, lg10_povertyPercent, lg10_PctBachDeg25_Over, lg10_PctPrivateCoverage

b. Dependent Variable: lg10_deathRate

Fig. 9. Model Summary Elimination and Transformation

*3) **ANOVA Test**:* Analysis of Variance or ANOVA is a typical framework which comprises the ground for trials of significance and delivers facts regarding the levels of variability in a linear regression model. This is identical to linear regression but the significant distinction observed is that the regression model is used to forecast a continuous output on the basis of one or more continuous independent variables. In other words, ANOVA is being utilised to predict a successive product on the basis of one or more categorical dependent

variables [12]. In my study and in the below figure the sum of squares value is 7.587 and the total Sum of Squares value is 14.516. It means that the regression model presented around (**SSR/SST**) 63% of the variability in the dataset.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7.587	9	.843	369.365	.000[b]
	Residual	6.929	3036	.002		
	Total	14.516	3045			

a. Dependent Variable: lg10_deathRate

b. Predictors: (Constant), lg10_PctEmpPrivCoverage, lg10_AvgHouseholdSize, lg10_incidenceRate, lg10_PctHS18_24, lg10_PctWhite, lg10_PctHS25_Over, lg10_povertyPercent, lg10_PctBachDeg25_Over, lg10_PctPrivateCoverage

Fig. 10. Anova Table

F-Test in statistical analysis is a hypothesis-testing approach which believes that two variances from two samples. The F-Test is employed when the distinction between two variances must be particularly evaluated, that is deciding whether two samples can be accepted as samples of the normal population with the same variance or not [13]. In my analysis of the death rate data, the F score is equivalent to 369.365. The allocation is F(9, 3036), and the probability of keeping a significance greater than or equal to 369.365 is less than 0.001. Thus, there is a powerful linear connection exists between the conditional variable and independent variables.

*4) **The Gauss Markov Assumptions**:*

- **Assumption 1: Linearity**

The foremost hypothesis of MLR is that the association between the independent variables and dependent variables must be **linear**. In the below figure shows the relationship between the dependent variable(deathrate) and independent variable(incedenceRate, povertyPercent, avghouseholdSize, PctHS18-24, PctHS25-Over, PctBachDeg25-over, PctPrivateCoverage, PctEmpPrivCoverage, PctWhite) should be a linear model.This leads to the fact that the connection between these variables is linear because it is required to complete the linearity hypothesis.

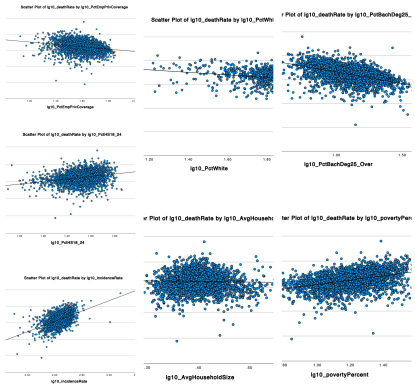

Fig. 11. Scatter Plot of Predictive Variable with the other Independent Variables after Log10 Transformation

- **Assumption 2: Non-Collinearity**

Non-collinearity refers to the absence of a high correlation between two or more predictor variables in a linear regression model. In other words, non-collinearity occurs when the predictor variables are not highly correlated with each other. In linear regression analysis, collinearity can lead to several problems such as instability of the regression coefficients, difficulties in interpreting the regression results, and reduced predictive accuracy. In this MLR study the multi-collinearity does not exist. The correlation data values in the below table are lower than 0.8. This indicated that the MLR prototype met the hypothesis and multicollinearity does not live.

Correlations

		lg10_deathRate	lg10_incidenceRate	lg10_povertyPercent	lg10_AvgHouseholdSize	lg10_PctHS18_24	lg10_PctHS25_Over	lg10_PctBachDeg25_Over	lg10_PctPrivateCoverage	lg10_PctWhite	lg10_PctEmpPrivCoverage
Pearson Correlation	lg10_deathRate	1.000	.521	.420	-.027	.258	.425	-.482	-.337	-.180	-.225
	lg10_incidenceRate	.521	1.000	.050	-.108	.069	.144	-.059	.070	-.041	.124
	lg10_povertyPercent	.420	.050	1.000	.090	.091	.260	-.604	-.807	-.436	-.694
	lg10_AvgHouseholdSize	-.027	-.108	.090	1.000	.060	-.176	-.074	-.303	-.334	-.007
	lg10_PctHS18_24	.258	.069	.091	.060	1.000	.431	-.396	-.234	.019	-.219
	lg10_PctHS25_Over	.425	.144	.260	-.176	.431	1.000	-.703	-.209	.135	-.211
	lg10_PctBachDeg25_Over	-.482	-.059	-.604	-.074	-.396	-.703	1.000	.610	.095	.507
	lg10_PctPrivateCoverage	-.337	.070	-.807	-.303	-.234	-.209	.610	1.000	.460	.825
	lg10_PctWhite	-.180	-.041	-.436	-.334	.019	.135	.095	.460	1.000	.308
	lg10_PctEmpPrivCoverage	-.225	.124	-.694	-.007	-.219	-.211	.507	.825	.308	1.000

Fig. 12. Correlations Table After log10 Transformation

- **Assumption 3: The values of the residuals are independent**

The **Durbin-Watson(DW)** test in statistics is used in statistical model to detect the presence of auto-correlation in the error of a linear regression analysis. The auto-correlation happens at a time when the residuals or errors of a regression sample are dependent, that means the residual spans are likened with each other over time period in other way across statements. [14]. The DW statistic ranges from 0 to 4, with a weight of 2 suggesting negative auto-correlation in the errors. A data value which is less than 2 points positive auto-correlation, means the residuals are associated with each other in a positive approach. A weight of better than 2 implies negative auto-correlation, indicating that the residuals are associated with each other in an averse direction. If you refer to the **Fig 9** of this MLR study the Durbin-Watson value has come to 1.989 which is close to 2, which shows that the residual data value is not dependent, and the analysis is fulfilled.

- **Assumption 4: Homoscedasticity**

Homoscedasticity also known as constant variance, is an assumption in statistical analysis, particularly in linear regression, that the variance of the errors or residuals is invariant throughout all decks of the independent variables. However, homoscedasticity thinks that the stretch of the residuals or errors is identical at each and every matter of the independent variables. The scatter plot below in the figure demonstrates that the data points in the plot picture aimlessly without creating any of the patterns such as fan-in, fan-out or funnel structure. This indicates that the variance of residuals is always uncorrelated. Thus, it can be concluded that the hypothesis is completed.

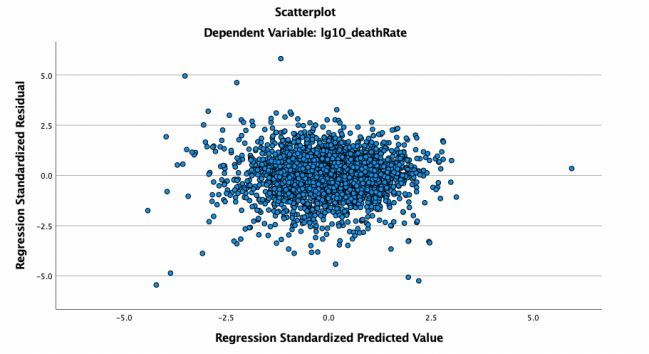

Fig. 13. Scatter Plot of Residuals

- **Assumption 5: Errors should be normally distributed**

A P-P plot, also known as a **probability-probability plot**, is a pictorial tool utilised in statistical research to correspond the distribution of a model to a hypothetical probability distribution, just like the normal distribution.

This p-p plot works by plotting the cumulative allocation function (CDF) of the sample data on the y-axis against the CDF of the theoretical distribution on the x-axis. If the sample data follows the theoretical distribution, the points on the P-P plot will fall along a linear line. However, if the sample data does not follow the visionary distribution, the points on the P-P plot will contrast from the linear line, showing a distinction between the experimental distribution and the standard distribution.

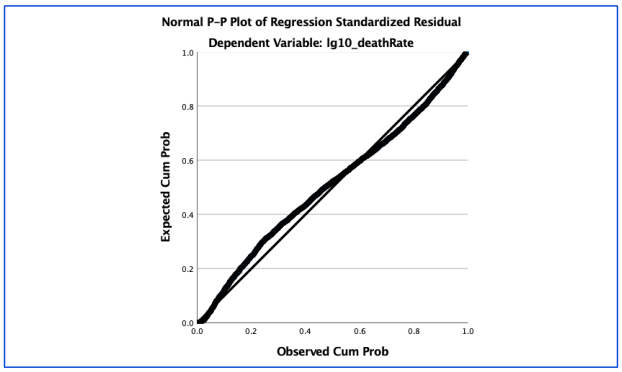

Fig. 14. P-P Plot

The P-P plot in the overhead figure illustrates that the hypothesis of the normality of the data points may have been barely disregarded. But, only the outer variations from normality are likely to have a substantial effect on the sample. The results are always exact and there are no notable results displayed in this analysis due to that.

- **Assumption 6: There are no influential cases biasing the model**

The **Cook's distance**, also known as Di is utilised in Regression Model to discover significant outliers in a dependent variable. In the other way, it is a practice to recognise issues that negatively impact a regression sample. It is a combination to measure each and every observation leverage and error values. The larger the leverage and residuals, the more elevated the Cook's distance value in the statistical model [15]. Many other writers recommend that any big considerable Cook's distance value must be examined. The agreement appears to be that a data value of greater than 1 displays an effective result, but you may like to examine values more than 0.5. Any significant value which attaches to the other should also be examined The lowest and highest value of the Cook's Distance in the below table does not exceed 1, which means that it does not seem to have any problem with significant observations in this statistical research.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	2.0260	2.5430	2.2467	.04992	3046
Std. Predicted Value	-4.422	5.937	.000	1.000	3046
Standard Error of Predicted Value	.001	.011	.003	.001	3046
Adjusted Predicted Value	2.0267	2.5426	2.2467	.04992	3046
Residual	-.26038	.27788	.00000	.04770	3046
Std. Residual	-5.450	5.817	.000	.999	3046
Stud. Residual	-5.484	5.832	.000	1.001	3046
Deleted Residual	-.26361	.27933	.00000	.04795	3046
Stud. Deleted Residual	-5.510	5.864	.000	1.002	3046
Mahal. Distance	.632	150.819	8.997	9.768	3046
Cook's Distance	.000	.072	.001	.002	3046
Centered Leverage Value	.000	.050	.003	.003	3046

a. Dependent Variable: lg10_deathRate

Fig. 15. Cook's Distance Analysis

V. Conclusion

In the above researched Multi Linear Regression model, the statistical summary along with the assumption demonstrates that there is a significant correlation between deathrate caused by cancer with incedenceRate, povertyPercent, avghouseholdSize, PctHS18-24, PctHS25-Over, PctBachDeg25-over, PctPrivateCoverage, PctEmpPrivCoverage, PctWhite. However, the correlation between the deathrate and Population, medincome, medianAge, medianAgemale, medianAgeFemale, PctMarriedHouseholds, PctNoHS18-24, PctBachDeg18-24, PctUnemployed16-Over, PctPublicCoverage, PctPublicCoverageAlone, PctBlack, PctAsian are not significant. Along with that, the final regression model indicates different assumptions from the dataset provided. To conclude the report I must say that this MLR model could help to predict the death rate due to cancer for a 3rd generation-leading country based on various factors just as age, race, income, employment etc.

VI. Acknowledgement

I would like to share my gratitude and thank fullness to my lecturer Professor Hicham Rifai for his guidance in class and extreme help towards my project

References

[1] National Institutes of Health, National Cancer Institute and USA.gov. Cancer: A Historic Perspective. [Online] Available: https://training.seer.cancer.gov/disease/history/

[2] Elsevier, Article from Clinical and Translational Radiation Oncology, Published online 2021 Jan 16. [Online] Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7829109/

[3] Cancer statistics, Cancer incidence in Ireland, "Cancer in Ireland 1994-2020: Annual Statistical Report 2022. National Cancer Registry of Ireland, 2022". [Online] Available: https://www.cancer.ie/cancer-information-and-support/cancer-information/about-cancer/cancer-statistics/

[4] Rebecca L. Siegel MPH, Kimberly D. Miller MPH, Nikita Sandeep Wagle MBBS, MHA, PhD, Ahmedin Jemal DVM, PhD, CA: A Cancer Journal for Clinicians. [Online] Available: https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21708/

[5] The Concise Encyclopedia of Statistics pp 283–287. [Online] Available: https://link.springer.com/referenceworkentry/10.1007/978-0-387-32833-1$_2$14/

[6] YOUR DATA TEACHER. [Online] Available: https://www.yourdatateacher.com/2022/11/07/a-practical-introduction-to-the-shapiro-wilk-test-for-normality/

[7] kristian.klima, Normality Testing - Skewness and Kurtosis [Online] Available: https://community.gooddata.com/metrics-and-maql-kb-articles-43/normality-testing-skewness-and-kurtosis-241/

[8] Variance Inflation Factor, investopedia [Online] Available: https://www.investopedia.com/terms/v/variance-inflation-factor.asp: :text=A

[9] Correction of Multicollinearity [Online] Available: https://corporatefinanceinstitute.com/resources/data-science/variance-inflation-factor-vif/

[10] The British Academy, Multiple Regression in SPSS [Online] Available: https://www.bristol.ac.uk/cmm/media/research/ba-teaching-ebooks/pdf/Multiple

[11] Coefficient of Determination, R-squared [Online] Available: https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html: :text=

[12] ANOVA for Regression [Online] Available: https://towardsdatascience.com/anova-for-regression-fdb49cf5d684/

[13] F-Test, Dheeraj Vaidya, CFA, FRM [Online] Available: https://www.wallstreetmojo.com/f-test/

[14] WILL KENTON, Durbin Watson Test: What It Is in Statistics, With Examples [Online] Available: https://www.investopedia.com/terms/d/durbin-watson-statistic.asp/

[15] Cook's Distance / Cook's D: Definition, Interpretation [Online] Available: https://www.statisticshowto.com/cooks-distance/