CA 1 -Multiple -Linear Regression Model on Cancer Data

Sakshi Kishor Khanvilkar
Master of Science – Data Analytics
National College of Ireland
Dublin, Ireland.
x22117776@student.ncirl.ie

Abstract— Linear Regression refers to the mathematical technique of fitting given data to a function of a certain type. This paper provides a brief description of implementing Multilinear Regression which analysis death rates based on given cancer data. The unified mechanism is divided into three field a.) Identifying the descriptive statistic and visualization. b) Curing the outliers and transforming the data c) Feature selection d) Implementing regression model. Furthermore, regression models are checked using Gauss Markov and multilinear regression parameters in order to achieve an appropriate model that meets all of the parameters.

Keywords—Linear Regression, Multi-linear Regression, Gauss Markov.

I. INTRODUCTION

Researchers used several techniques, like early-phase screening, to find different types of cancerous development before they affect health. Also, they have developed novel methods for predicting disease treatment outcomes in advance. With medical advancements, a quantity of data on malignant growth has been accumulated and is available to the diagnostic test team.

In every field, algorithms for machine learning have become quite effective. It can enable logical decisions and provide insightful facts from statistics. Linear regression analysis is a predictive modelling technique used in machine learning that assesses the association between a dataset target or dependent variable and its independent variable.

When the target and independent variables exhibit a linear connection with one another, various regression analysis techniques are applied. There are several ways to perform regression analysis, and the use of each method depends on a number of parameters. In this article, I look at building a simple multilinear regression model using a dataset provided by institute (NCI) as a CSV file of Cancer variable's.

I started in section I. as introduction II by dataset description. In section III, I go descriptive statistics and analysis, IV with dataset subset and feature transformation V implementing the regression model. I have

used different python library like a. pandas for reading the cancer data, b. NumPy for the numerical operation c. Matplotlib and seaborn for visualization of data d. Sci-kit learn for using linear regression model.

II. DATASET DESCRIPTION

The given data file contains 3047 rows and 25 columns. The important steps before moving further for the analysis are checking the missing value or Nan value and duplication of data, there were no duplicate data and no missing values. The data holds a feature 'county' which is categorical in nature has no used in prediction so we drop it, but rest feature are numerical.

The synopsis of data record in tabular forms looks as follows.

Columns	Data_Types
County	Categorical
Population	Continous
deathRate	Continous
incidenceRate	Continous
medIncome	Continous
povertyPercent	Continous
MedianAge	Continous
MedianAgeMale	Continous
MedianAgeFemale	Continous
AvgHouseholdSize	Continous
PctMarriedHouseholds	Continous
PctNoHS18_24	Continous
PctHS18_24	Continous
PctBachDeg18_24	Continous
PctHS25_Over	Continous
PctBachDeg25_Over	Continous
PctUnemployed16_Over	Continous
PctPrivateCoverage	Continous
PctEmpPrivCoverage	Continous
PctPublicCoverage	Continous
PctPublicCoverageAlone	Continous
PctWhite	Continous
PctBlack	Continous
PctAsian	Continous
PctOtherRace	Continous

Fig. 01. Data Field Summary

The variables in the table above are independent and contain different information about the target. We need to determine a distinct set of variables to predict the outcome

III. DESCRIPTIVE STATISTICS AND ANALYSIS

A. Estimation of Standard Descriptive Statistical Parameters.

Each field's descriptive statistical analysis are computed. Since the county field is a Categorical feature, it has not been included here. Before implementing the regression model, an insight of basic statistical parameters is necessary. Figure represents the computation of mean, maximum, count, standard deviation, and percentile of the field.

Descriptive Statistics							
	mean	std	min	0.25	0.5	0.75	max
Population	34221.2	31255.6	1410	11745	23892	45549	153638
IncidenceRate	449.344	47.8681	310.1	419.5	452.3	482.9	579.7
medIncome	46063.8	8151.45	27382	40224	45368	51335	72648
povertyPercent	15.6483	4.64048	5.9	12	15.1	18.9	32.4
MedianAge	42.1911	4.07328	29.4	39.8	42.1	44.7	53.3
MedianAgeMale	40.9313	4.11385	27.2	38.4	40.7	43.5	51.7
MedianAgeFemale	43.4881	4.13027	30	41	43.4	46	54.5
AvgHouseholdSize	2.46474	0.17051	1.99	2.35	2.46	2.57	3.02
PctMarriedHouseholds	52.9291	4.34403	37.1637	50.0041	52.7966	55.8661	66.6213
PctNoHS18_24	17.6588	6.58081	0.8	13	17.1	21.7	37.3
PctHS18_24	36.1352	8.14951	12.1	31	36.2	41.4	57.4
PctBachDeg18_24	5.50059	3.05957	0	3.3	5.2	7.4	15.8
PctHS25_Over	37.2663	5.72632	20.2	33.4	37.1	41.3	52.7
PctBachDeg25_Over	12.2068	3.90073	4	9.3	11.7	14.6	25.9
PctUnemployed16_Over	7.17132	2.80337	0.4	5.2	7.1	9.1	15.6
PctPrivateCoverage	65.8645	8.73515	38.9	59.8	66.3	72.3	88.8
PctEmpPrivCoverage	41.7418	7.80619	18.6	36.3	42	47	66.3
PctPublicCoverage	36.8195	6.46629	18.5	32.2	36.8	41.1	54.6
PctPublicCoverageAlone	18.7627	5.06912	5	15.1	18.7	22	35.4
PctWhite	91.3912	7.26206	52.6415	88.7312	94.1482	96.4691	99.693
PctBlack	3.6147	5.25629	0	0.502	1.26735	4.15244	25.0701
PctAsian	0.57935	0.49903	0	0.24216	0.46425	0.75489	2.67051
PctOtherRace	0.97018	1.06212	0	0.22939	0.57528	1.31533	4.92915
deathRate	181.092	24.7426	93.8	165.2	180.5	196.7	280.8

Fig. 02. Descriptive Statistics.

B. Correlation of the data fields

Calculating the descriptive basic research, the correlation is calculated across all the data fields.

The goal is to use a regression model to forecast a desired death rate relying on all of the primary independent factors.

As a result, the Pearson's correlation statistic was calculated for all of the independent variables with 'deathRate.

The correlation matrix has so many numbers to be read, that is difficult to read. Ruther I analysis the correlation of independent variables using heatmap. Correlation heatmaps are graphical representations of the intensity of correlations between numeric values. Heatmaps may be employed to identify and assess potential correlations between variables. Outliers, linear and nonlinear correlations can also be observed utilizing correlation plots. The color coding of a cell simplifies the ability to see the connections between variables at a glance.

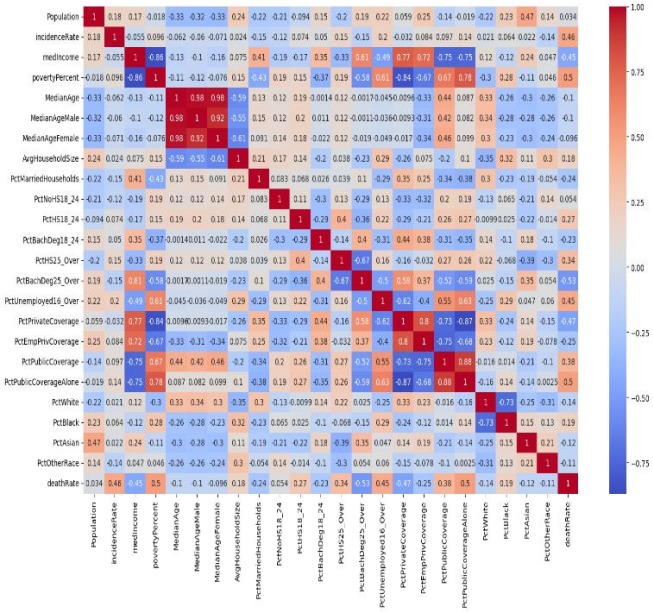

Fig. 03. Correlations Heatmap of variables

IV. DATA DISTRIBUTION AND OUTLIER HANDLING

A. Dist Plots of the Data Variables

I use the displot (also known as a distribution plot) to evaluate whether not all of the variables are normally distributed. It is a univariant set of collected data, which means the data distribution of one variable will be shown against another variable.

For visual analytics in Python, typically deploy the Seaborn package in conjunction with Matplotlib. It provides a user-friendly interface that enables users to exhibit data in statistical form in numerous informative and eye-catching images. With Seaborn, researchers utilize the seaborn.distplot() method to depict the proportion of one data variable versus another as a density distribution.

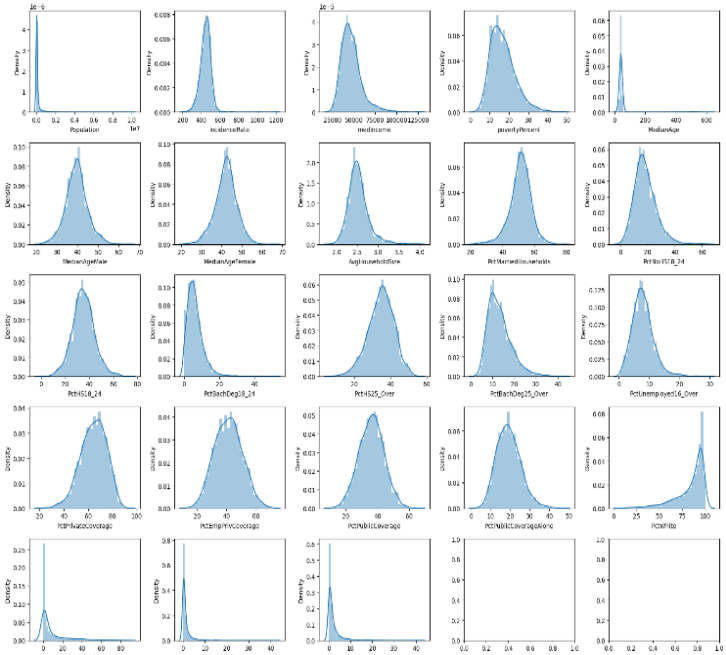

Fig. 04. Correlations Heatmap of variables

The Fig.05. depicts that for the field Population, incidencerate, medIncome, povertyPercent, medianAge, AvgHouseHoldsize, MedianAgeMale, PctNoHS18_24, PctBachDeg18_24, PctBachDeg25_over, PctUnemployed16_Over, PctEmpPrivCoverage, PctBlack, PctPublicCoerageAlone, PctAsian, PctOtherRaceit correspond to right skew which is also identified as positive skew distribution. Right-skew distribution illustrates that mean is greater than median.

The rest fields MedianAgeFemale, PctMarriedHouseholds, PctHS25_Over, PctPrivateCoverage, PctPublicCoverage, PctWhite correspond to left skew which is also identified as negative skew distribution. Left-skew distribution illustrates that mean is greater than median.

Fig. 04. shows the skewed value that describes the direction of outliers.

Population	14.289926
incidenceRate	0.750963
medIncome	1.408071
povertyPercent	0.930713
MedianAge	9.989944
MedianAgeMale	0.132041
MedianAgeFemale	-0.208384
AvgHouseholdSize	1.297096
PctMarriedHouseholds	-0.522362
PctNoHS18_24	0.973345
PctHS18_24	0.179209
PctBachDeg18_24	1.956201
PctHS25_Over	-0.333635
PctBachDeg25_Over	1.094837
PctUnemployed16_Over	0.891061
PctPrivateCoverage	-0.393537
PctEmpPrivCoverage	0.089416
PctPublicCoverage	-0.005436
PctPublicCoverageAlone	0.470949
PctWhite	-1.680904
PctBlack	2.258068
PctAsian	7.418041
PctOtherRace	4.952179

Fig. 05. Skewness values of data fields

B. *Box Plot*

The outlier is detected using the boxplot. The presence of data anomalies is one of the main drawbacks of every model's performance. Ideal outliers are the extreme values for a certain column that have an impact on the generalizability of the data and hypothesis. Outliers primarily have an impact on regression models since they substantially alter the equation.

The primary objective of outlier recognition and treatment is to guarantee that you are provided with the optimal platform possible for your data, given that your data is suitable for use with the proposed method. The data in this case is linear and thus compatible with the Linear Regression Algorithm. The box plot is a straight method of characterize outlier in data based on a five-number summary (minimum, first quartile (Q1), median, third quartile (Q3), and maximum). It is frequently used to validate data distribution and discover outliers.

The circles in Fig.06. display the outliers, of which there are many. Outliers can also be identified that use more than one parameter. One can change the code above to show outliers.

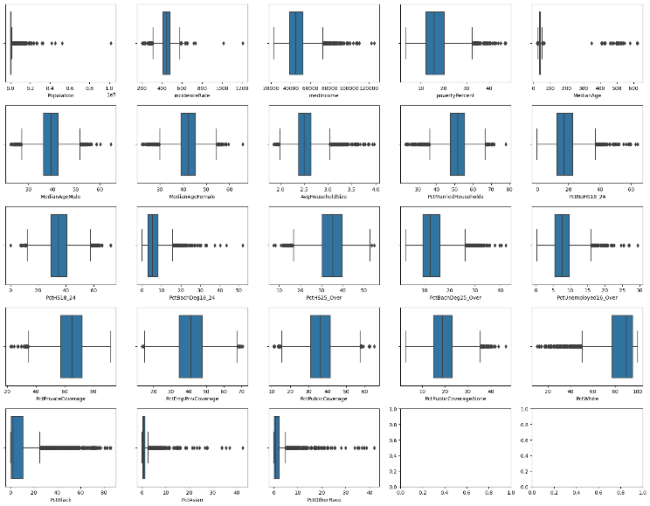

Fig. 06. Box Plot with Outliers

The box plots in Fig.06. show a large number of outliers,that Can potentially be contributing data points that effect the reg ression model.

To avoid this, I used the IQR rule stands for Inter-Quartile Range. It is defined as the difference between the 75th and 25th % of the data. To calculate the IQR

First, a threshold value is determined as the sum of 1.5 times the interquartile range plus the data field's third quantile value. It is theoretically expressed as follows:

Upper = 1.5*IQR (data field) + 3rd Quantile (data field)

Lower = 1.5*IQR (data field) - 3rd Quantile (data field)

Where, IQR= Q3-Q1

The calculated values 'Upper' and 'Lower' serve as a termination value beyond all data points in an individual data field.

are removed.

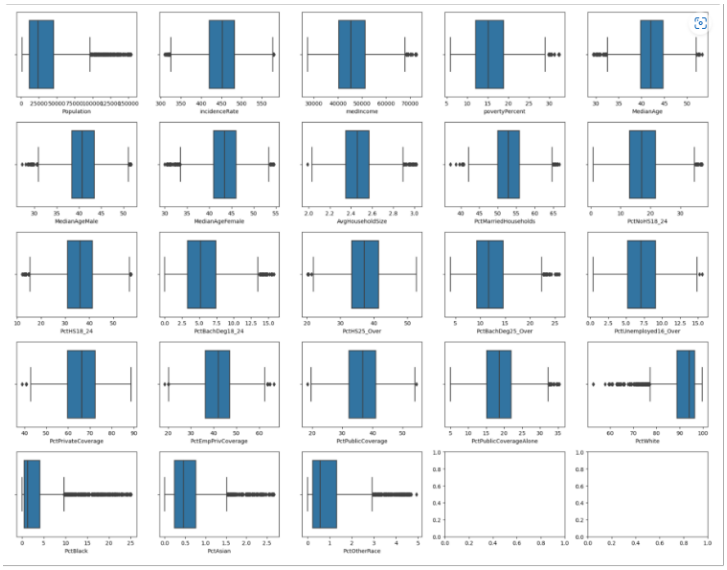

Fig. 07. Box Plot without Outliers

The Fig.07. demonstrates box plots using a scaled-down Y-axis, demonstrating how outliers are handled.

Population	1.573888
incidenceRate	-0.233523
medIncome	0.467059
povertyPercent	0.521794
MedianAge	-0.084635
MedianAgeMale	0.032772
MedianAgeFemale	-0.201785
AvgHouseholdSize	0.271484
PctMarriedHouseholds	0.049788
PctNoHS18_24	0.370201
PctHS18_24	-0.101152
PctBachDeg18_24	0.559955
PctHS25_Over	0.021837
PctBachDeg25_Over	0.563190
PctUnemployed16_Over	0.113435
PctPrivateCoverage	-0.212441
PctEmpPrivCoverage	-0.021370
PctPublicCoverage	0.093427
PctPublicCoverageAlone	0.290749
PctWhite	-1.692672
PctBlack	2.148722
PctAsian	1.580736
PctOtherRace	1.629122

Fig. 08. Box Plot without Outliers

Fig.08. Further, I checked the skewed value was also improved.

C. *Feature Selection on bases of Correlation.*

Feature selection is a technique for minimizing the number of input variables to the model by utilizing only relevant data and removing noise. The performances of model dependence upon the set of features selected. Negatively related feature can impact the performances of the model. There are several reasons why feature selection is an important step before

constructing a model: Reduces overfitting, improves accuracy, and shortens training time. There are three types of Feature selection technique wrapper method, filter method and Embedded method. For this data, I used a filter method to select features based on correlation, which signifies how close two or more variables are to having a linear relationship between them. As shown in fig.03 Good correlation characteristics are primarily linearly dependent, which means they have almost the same influence as the dependent variable. As a result, when two characteristics have a strong connection, I drop using the correlation limit of 0. 9.

D. Train and split of data.

After selecting features using correlation matrix I split the data into 80:20 ratio that means 80% of train data and 20 of test data.

V. Multilinear Model Building Procedure

Linear regression is a method of statistics used to examine the connection between a number of independent variables with a set of dependent variables. A linear equation is employed to construct the relationship between variables in linear regression. Simple linear regression, multiple linear regression, and polynomial regression are the three types of linear regression. For our data we are using Multilinear regression model this model is used to build relation between multiple linear regression model and a dependent variable.

$$\mathbf{Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}$$

were,

Y: Dependent Variable., **x1-x2**: independent Variables.

β0: constant. **β1- β2**: coefficient of independent variables.

A. Model on preselected feature using correlation.

The multiple regression model was implemented further on test and train data. The R^2 gained for the model with no outliers and features preselected using correlation is 49%.Total 18 independent feature were used for the training dataset.

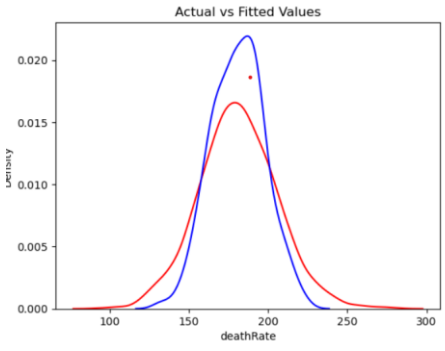

Fig. 09. Target output of features selected using correlation.

B. Fit Regression model using OLS.

OLS model is statistical model to fit multiple linear regression model. If every supposition is true, the Gauss-Markov theorem states, then the OLS eliminator has minimum variances among all the linear unbiased estimator. The below fig shows the OLS model.

OLS Regression Results

Dep. Variable:	deathRate	R-squared:	0.523
Model:	OLS	Adj. R-squared:	0.517
Method:	Least Squares	F-statistic:	80.29
Date:	Sun, 26 Mar 2023	Prob (F-statistic):	1.42e-250
Time:	08:38:28	Log-Likelihood:	-7257.4
No. Observations:	1705	AIC:	1.456e+04
Df Residuals:	1681	BIC:	1.469e+04
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	154.4438	24.653	6.265	0.000	106.089	202.798
Population	-2.041e-05	1.75e-05	-1.164	0.245	-5.48e-05	1.4e-05
incidenceRate	0.1821	0.009	19.243	0.000	0.164	0.201
medIncome	-3.373e-05	0.000	-0.242	0.809	-0.000	0.000
povertyPercent	0.8581	0.248	3.465	0.001	0.372	1.344
MedianAge	3.4468	1.416	2.433	0.015	0.669	6.225
MedianAgeMale	-2.0846	0.801	-2.604	0.009	-3.655	-0.514
MedianAgeFemale	-1.5373	0.741	-2.074	0.038	-2.991	-0.084
AvgHouseholdSize	-1.1219	4.747	-0.236	0.813	-10.433	8.189
PctMarriedHouseholds	-0.1503	0.141	-1.064	0.288	-0.427	0.127
PctNoHS18_24	-0.1533	0.075	-2.038	0.042	-0.301	-0.006
PctHS18_24	0.2637	0.063	4.183	0.000	0.140	0.387
PctBachDeg18_24	-0.1112	0.164	-0.677	0.498	-0.433	0.211
PctHS25_Over	0.0650	0.124	0.526	0.599	-0.178	0.308
PctBachDeg25_Over	-1.3944	0.216	-6.464	0.000	-1.817	-0.971
PctUnemployed16_Over	0.3729	0.236	1.583	0.114	-0.089	0.835
PctPrivateCoverage	-0.5869	0.194	-3.022	0.003	-0.968	-0.206
PctEmpPrivCoverage	0.4199	0.142	2.949	0.003	0.141	0.699
PctPublicCoverage	-0.3312	0.288	-1.152	0.249	-0.895	0.233
PctPublicCoverageAlone	0.5136	0.368	1.397	0.162	-0.207	1.235
PctWhite	-0.1871	0.101	-1.857	0.063	-0.385	0.011
PctBlack	0.0062	0.126	0.049	0.961	-0.240	0.252
PctAsian	-0.3605	1.068	-0.338	0.736	-2.455	1.734
PctOtherRace	-2.2696	0.465	-4.886	0.000	-3.181	-1.358

Omnibus:	26.609	Durbin-Watson:	2.044
Prob(Omnibus):	0.000	Jarque-Bera (JB):	45.728
Skew:	-0.086	Prob(JB):	1.18e-10
Kurtosis:	3.784	Cond. No.	3.67e+06

Fig.10. OLS statistical Model.

C. Model on preselected feature using P-value.

Fig. 10. The P-value for some data fields is less than 0.05. P-values may also be employed to select features; it's a filter method. Features are eliminated based on the P-value threshold value. That is, 0.05, and if any variable's p value falls below or equitable to the threshold value, the column's strength is determined.

The multiple regression model was implemented further on test and train data. The R^2 gained for the model with no outliers and features preselected using P-value is 52.8%. Total 11 independent feature were used for the training dataset.

VI. Assumption for Linear Regression Model

A. *Non-Constant Variance Test (ncvTest)*

This test establishes if the residual's (or model error's) variance is fixed. The Residual versus Fitted plot may be used to determine if homoscedasticity (fan-in or fan-out pattern) and heteroscedasticity (linear or curved trend) are present.

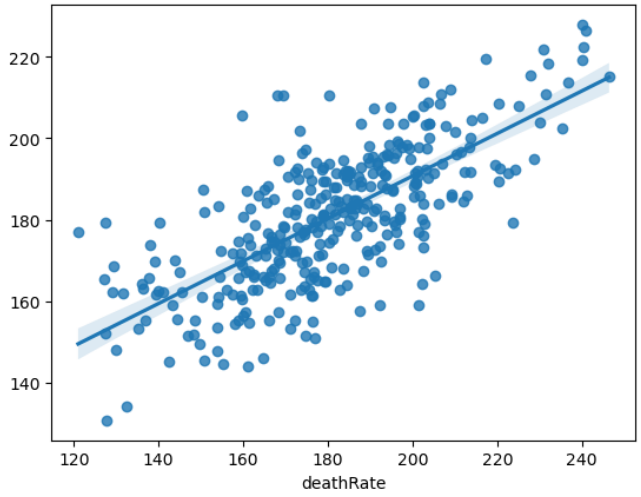

Fig.11. Homoscedasticity of error term.

B. *Normality of Error terms*

This test can determine whether the model's error terms are normally distributed. If the error term's p-value is less than 0.05, the error is normally distributed. If the error term is greater than 0.05, our confidence interval may be shrunk or inflated. The p-value of residual is 0.006.

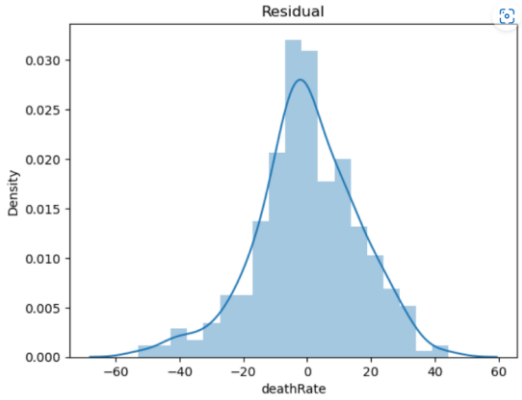

Fig.12. Normality of error term.

C. *Variance Inflation Factor (VIF) Test*

This study posits that the model's independent attributes are not correlated with one another. This leads in multicollinearity. The VIF statistic for predictors should preferably be close to 1. A satisfactory model, on the other hand, seems to have a VIF score of less than 5.

Variable	VIF
povertyPercent	3.891392
PctBachDeg25_Over	3.854447
PctEmpPrivCoverage	3.323969
PctWhite	2.986439
MedianAgeMale	2.715261
PctHS25_Over	2.715021
AvgHouseholdSize	2.703995
PctBlack	2.379421
PctUnemployed16_Over	2.283788
PctMarriedHouseholds	1.925322
Population	1.707939
PctAsian	1.607646
PctHS18_24	1.450552
PctBachDeg18_24	1.438082
PctOtherRace	1.383624
PctNoHS18_24	1.343336
incidenceRate	1.162450

Fig.13. VIF of independent Variables.

D. *Durbin Watson Test:*

The Durbin-Watson statistic tells us if the error terms are autocorrelated. Autocorrelation persists if the value is less than one or greater than three. The optimum Durbin-Watson stat is near two.

Durbin-Watson statistic: 1.9011626084838993

E. *Normal Q-Q Plot:*

This test generates a plot indicating the dataset's univariate normality. The Normal Q-Q plot must ideally be close to the normal distribution line and pass through the source. The discrepancy in the line represents that the plot is unsatisfactory.

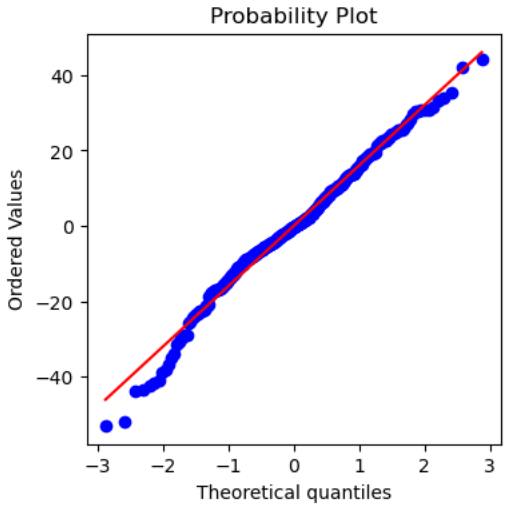

Fig.14. Normal Q-Q plot.

E.Cook's Distance Plot:

The Cook's Distance plot assists in identifying the key points of data. Each data point must meet Cook's distance less than one in order for this test to be successful. All the tests will pass easily with a great regression model. In contrast, erroneous data will always exist in the actual world, making it less likely that the regression model would pass all of these criteria.

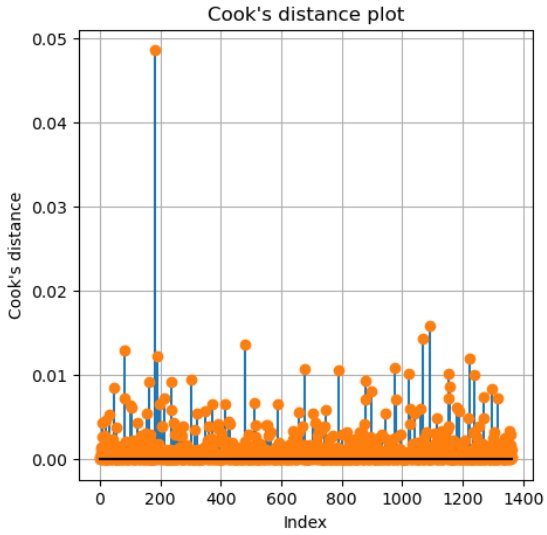

Fig.15. Cook's Distance plot.

CONCLUSION

In this article, we proposed the multi linear regression model to predict the death rate on base of data fields. Analysis the results of R^2 feature selection using p-value gave a descent R^2 value compared to feature selected by correlation matrix. Moreover, the model was evaluated utilizing various Gauss-Markov theorem assumptions, and the OLS model was fit, offering a brief description of the statistical values of the implemented regression model. In a realistic situation where the data is not clean, the acquired regression model needs to check most of the boxes for a good model.

VII. References

1. Assessing Assumptions of Multivariate Linear Regression Framework implemented for Directionality Analysis of fMRI

2. Correlation based Feature Selection impact on the classification of breast cancer patients response to neoadjuvant chemotherapy.