Statistics for Data Analytics CA1

Multiple Linear Regression Analysis

Harshitha Poolakanda Somanna
Student ID: x22150366
Statistics for Data Analytics
National College of Ireland, Dublin
x22150366@student.ncirl.ie

***Abstract*— Over the last few decades, there is a prevalent increase in cancer around the world. In this report, we estimate the best suited Multiple Linear Regression model that can be used to investigate the significant relationship of socio economic factors such as median income, poverty percentage, average household size, private insurance coverage and others influencing the death rate for Cancer patients in the US. The dataset sourced has data for over 3000 cancer patients in the US counties with about 25 attributes. SPSS has been used for descriptive statistics, visualizations, modeling and diagnostics.**

Keywords—independent variables; dependent variable; linear regression; normality test; linearity test; multicollinearity; Durbin-Watson test; Gauss-Markov assumptions.

I. INTRODUCTION

The aim of this analysis is to perform Multiple Linear Regression on the cancer dataset and to investigate socio-economic factors influencing the cancer mortality in the US using the best suited regression model. Multiple regression is a statistical technique used to examine the association between a dependent variable and multiple independent variables simultaneously. The independent variables whose values are known are used to predict the dependent variable in this technique and the formula is as follows:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta pXp + \varepsilon \qquad [1]$$

where:

- •Y is the dependent variable that we aim to estimate
- •$\beta 0$ is the intercept
- •$\beta 1, \beta 2, \ldots, \beta p$ are the regression coefficients corresponding to the independent variables X1, X2, ..., Xp
- •ε denotes the amount of variation in the dependent variable that cannot be accounted for by the independent variables.

The general foundations of multiple linear regression include the following:

- Assumptions: Ensure that the relationship between the variables is linear, errors are independent and normally distributed, and there is no multicollinearity.
- Model fitting: Estimate the coefficients of the independent variables using a statistical method that minimizes the sum of squared errors.
- Interpretation of coefficients: Coefficients indicate the change in the dependent variable for a one-unit change in the independent variable, holding all other variables constant.
- Model evaluation: Use metrics like R-squared, adjusted R-squared to evaluate the model's fit and ability to generalize to new data.

This report includes descriptive statistics and suitable visualizations to facilitate comprehension of the variables in the dataset. Also, the steps taken to construct the regression model leading to the final selection will be discussed along with details on the reasons for choosing the independent variables, the data transformation, outlier treatment, etc. Also, the procedures used to confirm the Gauss Markov assumptions for multiple regression, such as assessing normality, identifying outliers, and detecting multicollinearity among variables that may impact the precision of the model, are discussed for each model.

A. Data Description

The cancer.csv file with 3047 records and 25 attributes has been used as the source dataset. To predict the mortality rate in the cancer dataset, we can use the incidence rate and death rate columns as key factors in our regression model. In this context, the variables have been categorized into two groups: independent variables and dependent variables. The variable that we aim to examine and understand is the death rate, which is considered as the dependent variable. To investigate the death rate, we need to consider several independent variables. The socio economic variables and medical variables used in the dataset is shown in Fig.1.

deathRate
Population
incidenceRate
medIncome
povertyPercent
MedianAge
MedianAgeMale
MedianAgeFemale
AvgHouseholdSize
PctMarriedHouseholds
PctBachDeg18_24
PctBachDeg25_Over
PctUnemployed16_Over
PctPrivateCoverage
PctEmpPrivCoverage
PctPublicCoverage
PctPublicCoverageAlone

Table.1. List of Variables in the Cancer dataset

II. EXPLORATIVE DATA ANALYSIS

Explorative Data Analysis (EDA) is an essential first step to understand and summarize the key characteristics of data. It helps us to identify any potential outliers or anomalies, detecting missing values and to develop hypothesis for further testing. It typically involves various visual and statistical techniques, such as scatter plots, histograms, and correlation matrices.

A. Descriptive Statistics

Descriptive statistics is carried to summarize and describe the key features of a dataset such as the mean, median, mode, range, standard deviation, and variance. It also identifies outliers, patterns and trends in data and also helps us to explore the relationship between different variables in the data.

The initial stage of data preprocessing involves verifying the absence of any missing, empty, or null values, which were not identified in the dataset. When loading the dataset into SPSS, any preceding or succeeding spaces in string values were eliminated.

Table 2. shows the descriptive statistics of all possible variables in the cancer dataset where the characteristics of a dataset, such as its central tendency, variability, and distribution are described.

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Population	3047	10169465	827	10170292	102637.37	329059.221	1.083E+11	14.290	.044	337.437	.089
deathRate	3047	303.1	59.7	362.8	178.664	27.7515	770.146	.275	.044	1.355	.089
incidenceRate	3047	1005.6	201.3	1206.9	445.654	57.4566	3301.259	.751	.044	13.794	.089
medIncome	3047	102995	22640	125635	47063.28	12040.091	144963787.33	1.408	.044	3.713	.089
povertyPercent	3047	44.2	3.2	47.4	16.878	6.4091	41.076	.931	.044	1.276	.089
MedianAge	3047	601.7	22.3	624.0	45.272	45.3045	2052.496	9.990	.044	100.910	.089
MedianAgeMale	3047	42.3	22.4	64.7	39.571	5.2260	27.311	.132	.044	.676	.089
MedianAgeFemale	3047	43.4	22.3	65.7	42.145	5.2928	28.014	-.208	.044	.577	.089
AvgHouseholdSize	3047	2.11	1.86	3.97	2.5297	.24845	.062	1.297	.044	3.874	.089
PctMarriedHouseholds	3047	55.08290694	22.99248989	78.07539683	51.243872141	6.5728137943	43.202	-.522	.044	1.414	.089
PctNoHS18_24	3047	64.1	.0	64.1	18.224	8.0931	65.498	.973	.044	2.211	.089
PctHS18_24	3047	72.5	.0	72.5	35.002	9.0697	82.260	.179	.044	.534	.089
PctBachDeg18_24	3047	51.8	.0	51.8	6.158	4.5291	20.512	1.956	.044	9.139	.089
PctHS25_Over	3047	47.3	7.5	54.8	34.805	7.0349	49.490	-.334	.044	.119	.089
PctBachDeg25_Over	3047	39.7	2.5	42.2	13.282	5.3948	29.103	1.095	.044	1.737	.089
PctUnemployed16_Over	3047	29.0	.4	29.4	7.852	3.4524	11.919	.891	.044	2.297	.089
PctPrivateCoverage	3047	70.0	22.3	92.3	64.355	10.6471	113.360	-.394	.044	-.004	.089
PctEmpPrivCoverage	3047	57.2	13.5	70.7	41.196	9.4477	89.259	.089	.044	-.302	.089
PctPublicCoverage	3047	53.9	11.2	65.1	36.253	7.8417	61.493	-.005	.044	-.089	.089
PctPublicCoverageAlone	3047	44.0	2.6	46.6	19.240	6.1130	37.369	.471	.044	.362	.089
PctWhite	3047	89.80084490	10.19915510	100.00000000	83.645286235	16.380025229	268.305	-1.681	.044	2.691	.089
PctBlack	3047	85.947798580	.000000000	85.947798580	9.1079776146	14.534537922	211.253	2.258	.044	5.039	.089
PctAsian	3047	42.619424540	.000000000	42.619424540	1.2539649642	2.6102763927	6.814	7.418	.044	78.397	.089
PctOtherRace	3047	41.930251420	.000000000	41.930251420	1.9835230038	3.5177101375	12.374	4.952	.044	35.537	.089

Table. 2. Descriptive Statistics

B. Visualisation and discussion of out of range values

1) Residuals

Residual statistics as shown in Table 3. summarizes the difference between actual and predicted values. They help evaluate the model's goodness-of-fit, including mean, standard deviation, minimum and maximum, skewness, kurtosis, Durbin-Watson, and Cook's Distance. These statistics assist in identifying outliers, influential cases, and assumptions of the model.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	110.506	354.491	178.747	19.2679	2827
Std. Predicted Value	-3.542	9.121	.000	1.000	2827
Standard Error of Predicted Value	.465	6.119	1.059	.513	2827
Adjusted Predicted Value	110.771	353.714	178.749	19.2738	2827
Residual	-51.4206	55.7293	.0000	15.6062	2827
Std. Residual	-3.286	3.561	.000	.997	2827
Stud. Residual	-3.292	3.566	.000	1.000	2827
Deleted Residual	-51.5896	55.8610	-.0021	15.6865	2827
Stud. Deleted Residual	-3.297	3.573	.000	1.000	2827
Mahal. Distance	1.495	431.091	14.995	25.483	2827
Cook's Distance	.000	.009	.000	.001	2827
Centered Leverage Value	.001	.153	.005	.009	2827

a. Dependent Variable: deathRate

Table 3. Residual Statistics

2) Scatter Plot

A scatter plot as shown in Fig1 for multiple linear regression helps visualize the relationship between variables and detect outliers and influential points that can impact the regression results. Clusters around a straight line indicate a strong linear relationship.

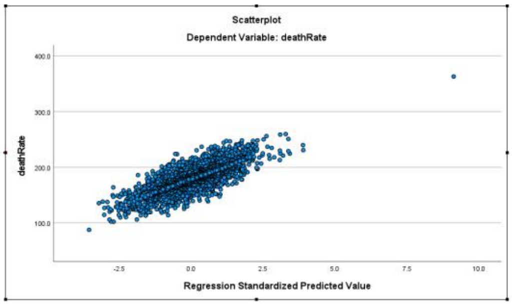


Fig 1. Scatter Plot

3) Histogram

Fig2. Shows a histogram plot that can identify distribution issues of independent variables in multiple linear regression by showing frequency distribution through grouped bins and count observations.

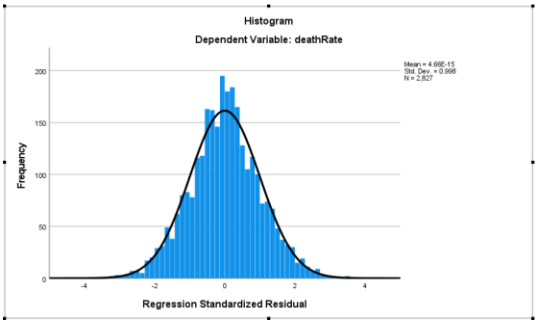


Fig 2. Histogram

4) Normal Probability plot

Normal plot as shown in Fig3. is useful for verifying the validity of model assumptions and the reliability of regression estimates in multiple linear regression by assessing the normality of the residuals[2]. A non-linear pattern in the normal plot of residuals indicates that the model assumptions have been violated and the regression estimates may be unreliable, necessitating a review of the model specification or alternative regression techniques.

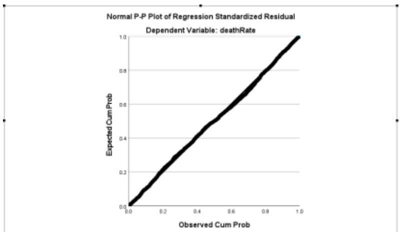


Fig 3. Normal P-P plot

5) Multicollinearity

Correlation matrix as displayed in Table4. depicts the correlation coefficients between variables in a dataset. In multiple linear regression, it helps identify if independent variables are highly correlated, indicating potential multicollinearity issues, which can cause unstable parameter estimates and inflated standard errors. Removing correlated variables or combining them can solve the issue and increase validity of results.

Correlations

		deathRate	Population	incidenceRate	medIncome	povertyPercent	MedianAge	MedianAgeMale	MedianAgeFemale	AvgHouseholdSize
Pearson Correlation	deathRate	1.000	-.120	.512	-.429	.429	.004	-.022	.012	-.037
	Population	-.120	1.000	.032	.236	-.065	-.025	-.177	-.178	.165
	incidenceRate	.512	.032	1.000	-.036	.054	.016	-.019	-.014	-.118
	medIncome	-.429	.236	-.036	1.000	-.789	-.013	-.092	-.153	.151
	povertyPercent	.429	-.065	.054	-.789	1.000	-.029	-.214	-.148	.144
	MedianAge	.004	-.025	.016	-.013	-.029	1.000	.129	.125	-.074
	MedianAgeMale	-.022	-.177	-.019	-.092	-.214	.129	1.000	.934	-.585
	MedianAgeFemale	.012	-.178	-.014	-.153	-.148	.125	.934	1.000	-.631
	AvgHouseholdSize	-.037	.165	-.118	.151	.144	-.074	-.585	-.631	1.000
	PctMarriedHouseholds	-.293	-.128	-.181	.446	-.605	.015	.222	.162	.134
	PctNoHS18_24	.088	-.127	-.162	-.289	.288	.006	.100	.136	.155
	PctHS18_24	.262	-.152	.061	-.190	.094	.051	.241	.243	.038
	PctBachDeg18_24	-.288	.248	.015	.493	-.387	-.017	-.034	-.071	-.134
	PctHS25_Over	.405	-.312	.138	-.471	.194	.037	.318	.345	-.184
	PctBachDeg25_Over	-.485	.297	-.068	.705	-.532	-.020	-.132	-.181	-.044
	PctUnemployed16_Over	.378	.051	.149	-.453	.655	.019	-.143	-.111	.266
	PctPrivateCoverage	-.386	.053	.045	.724	-.823	.005	.082	.047	-.280
	PctEmpPrivCoverage	-.267	.159	.103	.747	-.683	-.037	-.209	-.252	.012
	PctPublicCoverage	.405	-.160	.078	-.755	.651	.049	.399	.455	-.208
	PctPublicCoverageAlone	.449	-.041	.083	-.720	.799	-.003	.002	.048	.131
	PctWhite	-.177	-.190	-.044	.167	-.509	.035	.398	.340	-.347
	PctBlack	.257	.073	.133	-.270	.512	-.017	-.243	-.157	.104
	PctAsian	-.186	.464	-.007	.426	-.157	-.038	-.238	-.259	.207
	PctOtherRace	-.190	.241	-.192	.084	.047	-.030	-.267	-.274	.359

Table 4. Pearson Correlation Matrix

III. Model Development

Model building requires modification to the selection of variables, resulting in a more accurate model. To get a best fit model, it is required to run different models by analyzing and modifying variables that gives a higher accuracy score.

A. Model 1

Model 1 in multiple regression analysis involves predicting the dependent variable (death rate) using all independent variables available in the dataset, based on a dataset with 3047 observations.

1) Model Summary

Model Summary[b]

					Change Statistics					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
1	.736[a]	.542	.538	18.8599	.542	155.312	23	3023	.000	1.978

a. Predictors: (Constant), PctOtherRace, PctBachDeg18_24, PctMarriedHouseholds, MedianAge, incidenceRate, MedianAgeFemale, Population, PctNoHS18_24, PctHS18_24, PctAsian, PctUnemployed16_Over, PctBlack, PctHS25_Over, PctEmpPrivCoverage, AvgHouseholdSize, PctPublicCoverageAlone, PctBachDeg25_Over, povertyPercent, medIncome, PctWhite, MedianAgeMale, PctPrivateCoverage, PctPublicCoverage

b. Dependent Variable: deathRate

Table 5. Model Summary for Model 1

The model summary in Table 5. gives a correlation(R value) of .736 between the predicted and observed values which is 74% correlated and is a reliable model as the direction and strength of the relationship can be confidently interpreted. The R Square value of 54.2% says that the model is moderately good but is not enough to make strong conclusions about relationship between variables. The adjusted R Square value 53.8% says that 54% of the data points fit the linear regression. The value of 1.978 for Durbin-Watson says that there is a good autocorrelation between the residuals. Considering all these points, the model is a moderately good fit for predicting death rate considering all socio economic variables in the dataset.

2) ANOVA

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1270601.342	23	55243.537	155.312	.000[b]
	Residual	1075264.533	3023	355.695		
	Total	2345865.875	3046			

Table 6. ANOVA for Model 1

The ANOVA in Table 6. summarizes the multiple regression equation by checking the variation of the dependent variable. The F-ratio of 155.3 and p value of 0 shows how the independent variables affect the death rate.

3) Multicolinearity

The correlation coefficient of 0.934 from Table 7. suggests a strong association between MedianAgeFemale and MedianAgeMale, necessitating the removal of one variable. Similarly, the variables PctPrivateCoverage and PctEmpPrivCoverage show a significant correlation with medIncome, indicating the need to drop one of these variables.

Table 8 indicates that several variables, including MedianAgeFemale, PctPrivateCoverage, PctPublicCoverage, and PctPublicCoverageAlone, exhibit VIF (variance inflation factor) values exceeding 10. This indicates a strong correlation between these variables.

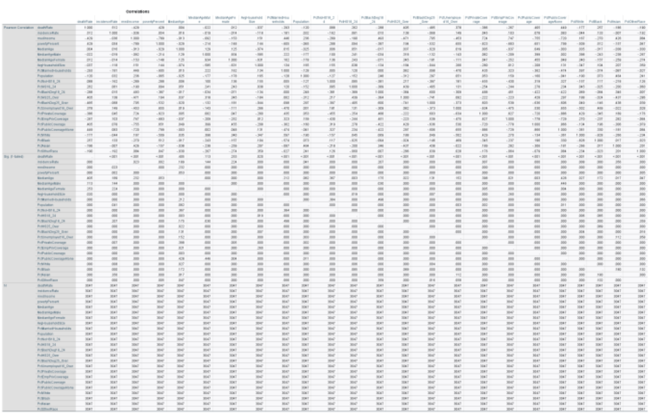


Table 7. Correlation Matrix for Model 1

Coefficients[a]

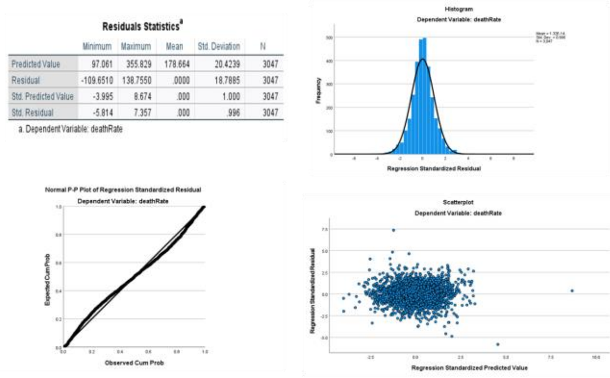
Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	VIF
1	(Constant)	175.945	15.509		11.344	<.001		
	incidenceRate	.205	.007	.424	31.139	<.001	.817	1.225
	medIncome	6.616E-5	.000	.029	.848	.397	.132	7.562
	povertyPercent	.367	.144	.085	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	<.001	.258	3.883
	PctMarriedHouseholds	.038	.098	.009	.392	.695	.282	3.549
	Population	-1.678E-6	.000	-.020	-1.356	.175	.705	1.419
	PctNoHS18_24	-.087	.054	-.025	-1.589	.112	.600	1.666
	PctHS18_24	.236	.048	.077	4.938	<.001	.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	<.001	.181	5.533
	PctUnemployed16_Over	.416	.157	.052	2.652	.008	.398	2.512
	PctPrivateCoverage	-.675	.130	-.259	-5.197	<.001	.061	16.391
	PctEmpPrivCoverage	.371	.098	.126	3.779	<.001	.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	<.001	.648	1.544

a. Dependent Variable: deathRate

Table 8. Coefficients for Model 1

4) Residuals

From Fig 4., it is observed that the standardized residual value of -5.814 indicates that the corresponding observation has a much lower value than what is predicted by the model and may need to be further examined for any errors or anomalies. The histogram plot, scatter plot and Normal P-P plot reveals the distribution of the variables here and it can be summarized that the dataset has outliers which must be further removed.



	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	97.061	355.829	178.664	20.4239	3047
Residual	-109.6510	138.7550	.0000	18.7885	3047
Std. Predicted Value	-3.995	8.674	.000	1.000	3047
Std. Residual	-5.814	7.357	.000	.996	3047

a. Dependent Variable: deathRate

Fig 4. Resuidual and Normality graphs for Model 1

B. *Model 2*

For Model 2, it was necessary to eliminate variables that exhibited high correlation with each other, as well as those that had a Variance Inflation Factor (VIF) greater than 10 for all the 3047 rows of data.

1) Model Summary

The Table 9 shows that the model's R value of .729 indicates a strong positive correlation between the dependent variable and the independent variables. The R Square value of .532 suggests that the model can explain 53% of the variance in the outcome variable based on the selected independent variables. The adjusted R Square value of .529 indicates that 53% of the data points closely fit the linear regression curve. Additionally, the Durbin-Watson statistic of 1.961 suggests that the residuals exhibit good autocorrelation. Overall, considering the 19 socio-economic variables in the dataset, the model can be considered a moderately good fit for predicting deathrate.

Model Summary[b]

					Change Statistics					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
1	.729[a]	.532	.529	19.0522	.532	180.824	19	3027	.000	1.961

a. Predictors: (Constant), PctEmpPrivCoverage, AvgHouseholdSize, MedianAge, incidenceRate, Population, PctHS18_24, PctBlack, PctOtherRace, PctNoHS18_24, PctAsian, PctBachDeg18_24, PctHS25_Over, PctUnemployed16_Over, MedianAgeMale, PctMarriedHouseholds, povertyPercent, PctBachDeg25_Over, medIncome, PctWhite

b. Dependent Variable: deathRate

Table 9. Model Summary for Model 2

2) ANOVA

The ANOVA analysis performed on Model 2, as shown in Table 10, evaluates how much the independent variables explain the variation in the dependent variable. The F-ratio of 180.8 and the corresponding p-value of 0 indicate that the independent variables have a statistically significant impact on the death rate outcome variable, with a much higher significance compared to the previous model.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1247100.980	19	65636.894	180.824	.000[b]
	Residual	1098764.895	3027	362.988		
	Total	2345865.875	3046			

Table 10. ANOVA for Model 2

3) Multicolinearity

Table 11 shows the high correlation coefficient of 0.7 between PctBachDeg25_Over, PctEmpPrivCoverage, and medIncome indicates a strong association among these variables, which could lead to redundancy. To avoid such redundancy, it is necessary to eliminate one of these variables. Additionally, the variables PctUnemployed16_Over and povertyPercent exhibit a correlation value of 0.66, which is indicative of a strong association. Furthermore, from Table 12, it is clear that PctBachDeg18_24 shows a correlation of 0.6 with PctBachDeg25_Over, which could suggest collinearity. Therefore, it is necessary to eliminate one of these variables to prevent collinearity issues.

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
Model		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	122.790	12.474		9.844	<.001		
	PctEmpPrivCoverage	.041	.074	.014	.559	.576	.243	4.123
	AvgHouseholdSize	-7.417	2.423	-.066	-3.061	.002	.329	3.042
	MedianAge	-.004	.008	-.006	-.484	.628	.979	1.021
	incidenceRate	.205	.007	.425	31.218	<.001	.834	1.199
	Population	-7.955E-7	.000	-.009	-.642	.521	.716	1.396
	PctHS18_24	.292	.047	.095	6.168	<.001	.646	1.549
	PctBlack	-.116	.053	-.061	-2.177	.030	.197	5.065
	PctOtherRace	-.785	.121	-.099	-6.468	<.001	.655	1.528
	PctNoHS18_24	-.018	.053	-.005	-.337	.736	.644	1.553
	PctAsian	-.067	.184	-.006	-.364	.716	.514	1.946
	PctBachDeg18_24	.001	.106	.000	.009	.993	.521	1.918
	PctHS25_Over	.293	.094	.074	3.120	.002	.272	3.674
	PctUnemployed16_Over	.564	.151	.070	3.723	<.001	.436	2.295
	MedianAgeMale	-.422	.110	-.079	-3.849	<.001	.363	2.752
	PctMarriedHouseholds	-.161	.095	-.038	-1.703	.089	.307	3.256
	povertyPercent	.788	.134	.182	5.899	<.001	.162	6.158
	PctBachDeg25_Over	-1.414	.146	-.275	-9.697	<.001	.193	5.191
	medIncome	6.612E-5	.000	.029	.847	.397	.135	7.411
	PctWhite	-.154	.057	-.091	-2.681	.007	.135	7.428

a. Dependent Variable: deathRate

Table 11. Coefficient Matrix for Model 2

Correlations

		deathRate	Population	incidenceRate	medIncome	MedianAgeMale	PctWhite	PctBlack	PctAsian	PctOtherRace	povertyPercent	MedianAge	AvgHouseholdSize	PctMarriedHouseholds	PctNoHS18_24	PctHS18_24	PctHS25_Over
Pearson Correlation	deathRate	1.000	-.120	.512	-.429	-.022	-.177	.257	-.186	-.190	.429	.004	-.037	-.293	.088	.262	.405
	Population	-.120	1.000	.032	.236	-.177	-.190	.073	.464	.241	-.065	-.025	.165	-.129	-.127	-.152	-.312
	incidenceRate	.512	.032	1.000	-.036	-.019	-.044	.133	-.007	-.192	.054	.016	-.118	-.181	-.162	.061	.138
	medIncome	-.429	.236	-.036	1.000	-.092	.167	-.270	.426	.084	-.789	-.013	.151	.446	-.289	-.190	-.471
	MedianAgeMale	-.022	-.177	-.019	-.092	1.000	.398	-.243	-.238	-.267	-.214	.129	-.585	.222	.100	.241	.318
	PctWhite	-.177	-.190	-.044	.167	.398	1.000	-.828	-.266	-.234	-.509	.035	-.347	.597	-.157	.045	.188
	PctBlack	.257	.073	.133	-.270	-.243	-.828	1.000	.017	-.023	.512	-.017	.104	-.574	.117	-.025	-.024
	PctAsian	-.186	.464	-.007	.426	-.238	-.266	.017	1.000	.201	-.157	-.038	.207	-.087	-.218	-.200	-.437
	PctOtherRace	-.190	.241	-.192	.084	-.267	-.234	-.023	.201	1.000	.047	-.030	.359	-.027	.126	-.060	-.286
	povertyPercent	.429	-.065	.054	-.789	-.214	-.509	.512	-.157	.047	1.000	-.029	.144	-.605	.288	.094	.194
	MedianAge	.004	-.025	.016	-.013	.129	.035	-.017	-.038	-.030	-.029	1.000	-.074	.015	.006	.051	.037
	AvgHouseholdSize	-.037	.165	-.118	.151	-.585	-.347	.104	.207	.359	.144	-.074	1.000	.134	.155	.038	-.184
	PctMarriedHouseholds	-.293	-.129	-.181	.446	.222	.597	-.574	-.087	-.027	-.605	.015	.134	1.000	.005	.120	.062
	PctNoHS18_24	.088	-.127	-.162	-.289	.100	-.157	.117	-.218	.126	.288	.006	.155	.005	1.000	.085	.217
	PctHS18_24	.262	-.152	.061	-.190	.241	.045	-.025	-.200	-.060	.094	.051	.038	.120	.085	1.000	.439
	PctHS25_Over	.405	-.312	.138	-.471	.318	.188	-.024	-.437	-.286	.194	.037	-.184	.062	.217	.439	1.000
Sig. (1-tailed)	deathRate	.	<.001	<.001	<.001	.113	<.001	<.001	<.001	<.001	<.001	.405	.020	<.001	<.001	<.001	<.001
	Population	.000	.	.041	.000	.000	.000	.000	.000	.000	.000	.082	.000	.000	.000	.000	.000
	incidenceRate	.000	.041	.	.023	.144	.008	.000	.356	.000	.002	.189	.000	.000	.000	.000	.000
	medIncome	.000	.000	.023	.	.000	.000	.000	.000	.000	.000	.232	.000	.000	.000	.000	.000
	MedianAgeMale	.113	.000	.144	.000	.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	PctWhite	.000	.000	.008	.000	.000	.	.000	.000	.000	.000	.027	.000	.000	.000	.006	.000
	PctBlack	.000	.000	.000	.000	.000	.000	.	.180	.102	.000	.172	.000	.000	.000	.085	.089
	PctAsian	.000	.000	.356	.000	.000	.000	.180	.	.000	.000	.017	.000	.000	.000	.000	.000
	PctOtherRace	.000	.000	.000	.000	.000	.000	.102	.000	.	.005	.047	.000	.066	.000	.000	.000
	povertyPercent	.000	.000	.002	.000	.000	.000	.000	.000	.005	.	.053	.000	.000	.000	.000	.000
	MedianAge	.405	.082	.189	.232	.000	.027	.172	.017	.047	.053	.	.000	.212	.367	.003	.022
	AvgHouseholdSize	.020	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.	.000	.000	.019	.000
	PctMarriedHouseholds	.000	.000	.000	.000	.000	.000	.000	.000	.066	.000	.212	.000	.	.384	.000	.000
	PctNoHS18_24	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.367	.000	.384	.	.000	.000
	PctHS18_24	.000	.000	.000	.000	.000	.006	.085	.000	.000	.000	.003	.019	.000	.000	.	.000
	PctHS25_Over	.000	.000	.000	.000	.000	.000	.089	.000	.000	.000	.022	.000	.000	.000	.000	.
N	deathRate	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047
	Population	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047
	incidenceRate	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047
	medIncome	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047
	MedianAgeMale	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047
	PctWhite	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047
	PctBlack	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047
	PctAsian	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047
	PctOtherRace	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047
	povertyPercent	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047
	MedianAge	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047
	AvgHouseholdSize	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047
	PctMarriedHouseholds	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047
	PctNoHS18_24	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047
	PctHS18_24	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047
	PctHS25_Over	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047	3047

Table 12. Pearson's Correlation for Model 2

4) Residuals

According to the results obtained in Fig 5, the standardized residual value of -6.069 suggests that the residual value deviates 6.069 standard deviations from the expected value, indicating the presence of an extreme outlier. Additionally, the analysis of the histogram, scatter plot, and Normality plot displayed in Figure 5 also indicates the existence of a significant outlier in the dataset, even after eliminating variables with a strong relationship and high VIF value.

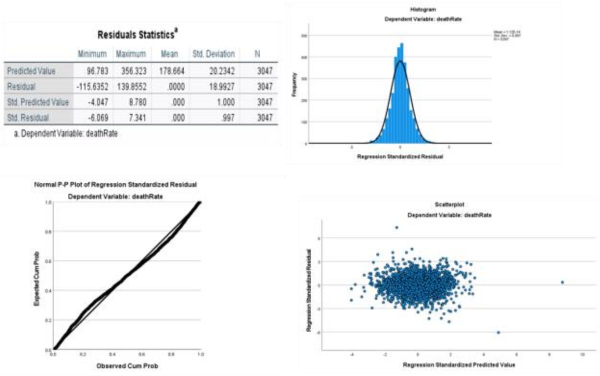


Fig 5. Residual and Normality graphs for Model 2

C. *Model 3*

In the current iteration, a variable from the previous model's set that showed strong correlation was removed, and outliers were eliminated using Cook's distance method.

1) *Model Summary*

The analysis utilizing Cook's method resulted in the identification of 220 rows of data as outliers, which were subsequently removed. The model was created on 2827 rows of data. Following this data cleansing, a new model was constructed using the remaining data.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics: R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
1	.777[a]	.604	.602	15.6478	.604	285.656	15	2811	.000	1.885

a. Predictors: (Constant), PctWhite, MedianAge, PctHS18_24, incidenceRate, PctNoHS18_24, Population, AvgHouseholdSize, PctOtherRace, medIncome, MedianAgeMale, PctAsian, PctHS25_Over, PctMarriedHouseholds, povertyPercent, PctBlack

b. Dependent Variable: deathRate

Table 13. Model Summary

The newly constructed model as summarized in Table 13 shows an R value of .775 and an R Square value of .604, indicating that the model can accurately predict the outcome by 60.4% based on the selected independent variables. Additionally, the adjusted R Square value of .602 suggests that the data points conform well to the linear regression curve. The Durbin-Watson value of 1.885 indicates a strong autocorrelation between the residuals.

2) *ANOVA*

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1049160.520	15	69944.035	285.656	.000[b]
	Residual	688283.853	2811	244.854		
	Total	1737444.373	2826			

a. Dependent Variable: deathRate

b. Predictors: (Constant), PctWhite, MedianAge, PctHS18_24, incidenceRate, PctNoHS18_24, Population, AvgHouseholdSize, PctOtherRace, medIncome, MedianAgeMale, PctAsian, PctHS25_Over, PctMarriedHouseholds, povertyPercent, PctBlack

Table 14. ANOVA for Model 3

In this case, from Table 14, the higher F-ratio value of 285.6 is better at explaining the variation in the dependent variable and is better than the other two models.

3) *Multicolinearity*

The model exhibited a strong degree of colinearity, with the majority of variables demonstrating linearity below 5 as shown in Table 15. Despite this, there was still a notable correlation between certain variables, such as a correlation of 0.592 between PctAsian and Population, 0.528 between Black and Population, and PctMarriedHousehold and PctWhite.

Correlations

		deathRate	PctHS25_Over	MedianAge	PctBlack	incidenceRate	AvgHouseholdSize	PctNoHS18_24	Population	PctHS18_24	PctOtherRace	povertyPercent	PctAsian	MedianAgeMale	PctMarriedHouseholds	medIncome	PctWhite
Pearson Correlation	deathRate	1.000	.454	-.017	.262	.539	-.019	.107	-.147	.309	-.223	.489	-.225	-.014	-.313	-.479	-.283
	PctHS25_Over	.454	1.000	.013	-.040	.540	-.183	.231	-.387	.452	-.323	.219	-.512	.314	.062	-.491	.289
	MedianAge	-.017	.013	1.000	-.013	.006	-.068	-.005	-.024	.030	-.033	-.049	-.034	.117	.024	.011	.017
	PctBlack	.262	-.040	-.013	1.000	.147	.150	.128	.091	-.036	-.015	.528	.032	-.256	-.576	-.270	-.884
	incidenceRate	.539	.540	.006	.147	1.000	-.079	-.134	.055	.076	-.282	.071	.017	-.012	-.183	-.043	-.076
	AvgHouseholdSize	-.019	-.183	-.068	.150	-.079	1.000	.129	.281	.063	.481	.084	.213	-.580	.185	.282	-.315
	PctNoHS18_24	.107	.231	-.005	.128	-.134	.129	1.000	-.161	.126	.119	.264	-.236	.128	.033	-.277	-.138
	Population	-.147	-.387	-.024	.091	.055	.281	-.161	1.000	-.182	.262	-.096	.592	-.218	-.148	.388	-.228
	PctHS18_24	.309	.452	.030	-.036	.076	.063	.126	-.182	1.000	-.090	.109	-.364	.248	.130	-.309	.062
	PctOtherRace	-.223	-.323	-.033	-.015	-.282	.481	.119	.262	-.090	1.000	.070	.276	-.381	-.057	.089	-.246
	povertyPercent	.489	.219	-.049	.528	.071	.084	.264	-.096	.109	.070	1.000	-.167	-.179	-.612	-.797	-.511
	PctAsian	-.225	-.512	-.034	.032	.017	.213	-.236	.592	-.364	.276	-.167	1.000	-.284	-.095	.469	-.251
	MedianAgeMale	-.014	.314	.117	-.256	-.012	-.580	.128	-.218	.248	-.381	-.179	-.284	1.000	.219	-.111	.386
	PctMarriedHouseholds	-.313	.062	.024	-.576	-.183	.185	.033	-.148	.130	-.057	-.612	-.095	.219	1.000	.455	.593
	medIncome	-.479	-.491	.011	-.270	-.043	.282	-.277	.388	-.309	.089	-.797	.469	-.111	.455	1.000	.169
	PctWhite	-.283	.289	.017	-.884	-.076	-.315	-.139	-.228	.062	-.246	-.511	-.251	.386	.593	.169	1.000
Sig. (1-tailed)	deathRate	.	<.001	.185	<.001	<.001	.167	<.001	<.001	<.001	<.001	<.001	<.001	.226	<.001	<.001	<.001
	PctHS25_Over	.000	.	.237	.017	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	MedianAge	.185	.237	.	.249	.377	.000	.395	.100	.054	.040	.004	.035	.000	.101	.279	.190
	PctBlack	.000	.017	.249	.	.000	.000	.000	.000	.001	.209	.000	.044	.000	.000	.000	.000
	incidenceRate	.000	.000	.377	.000	.	.000	.000	.002	.000	.000	.000	.185	.268	.000	.012	.000
	AvgHouseholdSize	.167	.000	.000	.000	.000	.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	PctNoHS18_24	.000	.000	.395	.000	.000	.000	.	.000	.000	.000	.000	.000	.000	.040	.000	.000
	Population	.000	.000	.100	.000	.002	.000	.000	.	.000	.000	.000	.000	.000	.000	.000	.000
	PctHS18_24	.000	.000	.054	.001	.000	.000	.000	.000	.	.000	.000	.000	.000	.000	.000	.001
	PctOtherRace	.000	.000	.040	.209	.000	.000	.000	.000	.000	.	.000	.000	.000	.001	.000	.000
	povertyPercent	.000	.000	.004	.000	.000	.000	.000	.000	.000	.000	.	.000	.000	.000	.000	.000
	PctAsian	.000	.000	.035	.044	.185	.000	.000	.000	.000	.000	.000	.	.000	.000	.000	.000
	MedianAgeMale	.226	.000	.000	.000	.268	.000	.000	.000	.000	.000	.000	.000	.	.000	.000	.000
	PctMarriedHouseholds	.000	.000	.101	.000	.000	.000	.040	.000	.000	.001	.000	.000	.000	.	.000	.000
	medIncome	.000	.000	.279	.000	.012	.000	.000	.000	.000	.000	.000	.000	.000	.000	.	.000
	PctWhite	.000	.000	.190	.000	.000	.000	.000	.000	.001	.000	.000	.000	.000	.000	.000	.
N	deathRate	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827
	PctHS25_Over	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827
	MedianAge	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827
	PctBlack	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827
	incidenceRate	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827
	AvgHouseholdSize	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827
	PctNoHS18_24	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827
	Population	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827
	PctHS18_24	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827
	PctOtherRace	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827
	povertyPercent	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827
	PctAsian	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827
	MedianAgeMale	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827
	PctMarriedHouseholds	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827
	medIncome	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827
	PctWhite	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827	2827

Table 15. Pearson's Correlation for Model 3

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	VIF
1	(Constant)	51.414	10.094		5.093	<.001		
	PctHS25_Over	.809	.062	.226	13.011	<.001	.466	2.144
	MedianAge	-.003	.009	-.005	-.392	.695	.981	1.020
	PctBlack	-.102	.058	-.058	-1.776	.076	.133	7.530
	incidenceRate	.209	.006	.448	34.997	<.001	.861	1.162
	AvgHouseholdSize	5.435	2.239	.049	2.427	.015	.350	2.857
	PctNoHS18_24	.034	.045	.010	.749	.454	.736	1.359
	Population	-1.733E-6	.000	-.018	-1.150	.250	.603	1.659
	PctHS18_24	.373	.040	.131	9.224	<.001	.700	1.429
	PctOtherRace	-.968	.126	-.118	-7.666	<.001	.595	1.681
	povertyPercent	1.204	.118	.288	10.161	<.001	.176	5.692
	PctAsian	-.084	.206	-.007	-.406	.685	.424	2.357
	MedianAgeMale	-.262	.087	-.053	-3.019	.003	.454	2.203
	PctMarriedHouseholds	-.174	.085	-.044	-2.047	.041	.305	3.275
	medIncome	.000	.000	-.067	-2.314	.021	.170	5.895
	PctWhite	-.145	.061	-.089	-2.396	.017	.103	9.721

a. Dependent Variable: deathRate

Table 16. Coefficient Matrix for Model 3

4) *Residual Statistcs*

According to Figure 6, the regression model with a standardized residual of -3.286 performed better than the other two models. The histogram and scatter plot also indicate a significant reduction in the number of outliers in this model. Furthermore, the normal P-P plot shows that this model is almost aligned with the linear line, indicating a good fit to the normal distribution.

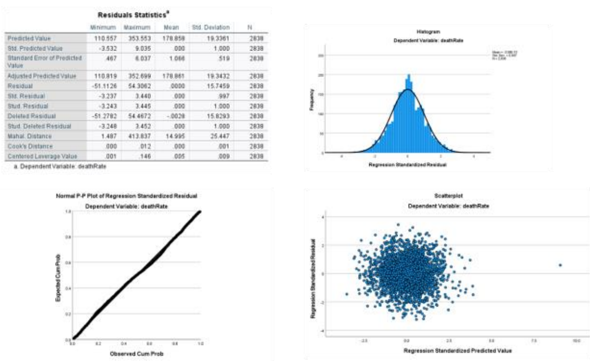


Fig 6. Resuidual and Normality graphs for Model 3

IV. MODEL SUMMARY

After analyzing Model 1, Model 2, and Model 3 in a regression analysis, it was determined that Model 3, which included multiple independent variables such as deathRate, Population, incidenceRate, medIncome, povertyPercent, MedianAge, MedianAgeMale, PctWhite, PctBlack, PctAsian, PctOtherRace, AvgHousseholdSize, PctMarriedHouseholds, PctNoHS18_24, PctHS18_24, and PctHS25_over, outperformed the other two models in predicting the death rate.

To improve the accuracy of the analysis, outliers in these variables were removed based on a threshold for Cook's distance of 0.00131, which resulted in the elimination of 220 rows. Cook's distance is a measure of the influence of individual observations on a regression model, and a threshold value can be set to identify influential observations.

The removal of these outliers allowed for a more precise analysis of the relationship between the independent variables and the death rate. Based on these results, it can be concluded that Model 3 is the most effective in predicting the death rate, and the inclusion of multiple independent variables can provide a more comprehensive understanding of the factors that contribute to the outcome variable.

A) Gauss Markov Theorem

The Gauss-Markov theorem asserts that if the errors in the multiple linear regression model are normally distributed, have zero mean, constant variance (homoscedasticity), and are uncorrelated with each other, then the ordinary least squares (OLS) estimator is the best linear unbiased estimator (BLUE) [3].

1. Multicolinearity

The analysis indicates that there is no high degree of correlation between the independent variables in all three models. However, after transforming the model and conducting further analysis, Model 3 is found to be the best fitting model, as it has a correlation coefficient less than 6 and a Variance Inflation Factor (VIF) value less than 10. These findings suggest that Model 3 has the lowest level of multicollinearity among the three models and is therefore the most appropriate for predicting the outcome variable.

2. Independence of Error

Independence of errors is one of the key assumptions of the Gauss-Markov theorem in multiple regression. It requires that the errors in the model are independent of one another. Durbin-Watson test was done on all three models and all of them exhibited the results < 1.978 indicating acceptable levels of autocorrelation where Model 3 outperformed the other models by having result 1.885.

3. Homoscedasticity

Homoscedasticity is when the variance of residuals in a regression model is constant across all independent variable values. It's tested using a scatter plot to check for patterns that may indicate heteroscedasticity.

Model 1, 2 and 3 has a clustered pattern in the scatter plot which is more desirable than a funnel shape pattern, as it suggests that there are clear relationships between the independent variables and the dependent variable. However, in Model 3, the cluster is lying in a line with very few outliers, it suggests that there is a strong linear relationship between the independent variable(s) and the dependent variable. This is a desirable pattern because it indicates that the model is able to explain a large proportion of the variation in the dependent variable based on the independent variable(s).

4. Normality

The normal probability plot (P-P plot) is a graphical tool that helps to assess the normality assumption of the Gauss-Markov theorem in multiple regression. It plots the standardized residuals against the expected values of a normal distribution, with a roughly straight line indicating that the residuals are normally distributed.

In this scenario, the analysis of all three models showed that the normal probability plot for Model 3 exhibited a roughly straight line with points evenly distributed along it. This suggests that the residuals for Model 3 are normally distributed and the normality assumption of the Gauss-Markov theorem is satisfied. Additionally, this finding indicates that Model 3 is the best fitting model among the three, as it satisfies the normality assumption.

5. Linearity

To satisfy linearity, the relationship between the dependent variable and the independent variables must be linear. In this scenario, it seems that all three models have shown a significant degree of linearity. However, Model 3 is the best fitting model in terms of linearity, which makes it a better choice for making predictions or drawing conclusions about the relationships between the variables.

V. CONCLUSION

The study employs a multiple regression analysis to examine how the incidence rate and a set of socio-economic variables impact the death rate of individuals diagnosed with cancer. My work has shown that the models have a respectable level of R square or adjusted R square of 53%. However, the adjusted R square value increases to 60% when outliers and a few correlated independent variables are removed from the dataset. The study identifies 11 variables that can be used to predict the death rate in the cancer dataset with relatively high accuracy. These variables include Population, incidenceRate, medIncome, povertyPercent, MedianAge, MedianAgeMale, PctWhite, PctOtherRace, AvgHouseholdSize, PctNoHS18_24, PctHS18_24, and PctHS25_over.

REFERENCES

[1] M.H. Kutner, C.J. Nachtsheim, J. Neter, and W. Li, Applied Linear Statistical Models, McGraw-Hill, 2005.

[2] G. Lind, W. Marchal, and S. Wathen, "Statistical Techniques in Business and Economics," 15th ed.

[3] J. G. Dai and Z. Niu, "Gauss-Markov Theorem," in Encyclopedia of Biophysics, G. C. K. Roberts, Ed. Berlin, Germany: Springer, 2013, pp. 893-895. DOI: 10.1007/978-3-642-16712-6_740.