MSc/PGDip in Data Analytics Statistics for Data Analytics CA1

Akimuddin Aslam Shaikh

School of Computing, National College of Ireland

x22123245 and x22123245@student.ncirl.ie

Abstract

The goal of this research is to find the best combination of multiple linear regression models to estimate the predicted cancer-related mortality rate per county using the county's cancer-increased incidence and accessible macroeconomic data. The dataset includes medical information such as the incidence rate and mortality rate, as well as socioeconomic information such as income, age, family size, education level, healthcare insurance, and race-related information. Using the provided data, the project will create and test different linear regression models before selecting the optimal model according to numerical criteria. By creating a total of three predictive models using regression approach, it is found that the regression model that includes all independent set of variables is best suitable for prediction of the deathRate. This model exhibits coefficient of determination value of 0.986.

Introduction

Understanding the factors contributing to cancer incidence and increased mortality is crucial for optimal health scheduling and resource allocation. Prior study has found a substantial positive relationship between cancer incidence and death rates. In other words, places with greater cancer mortality rates also have higher death rates [2.]. This association has been reported in a variety of cancer types, particularly lung, breast, and colorectal cancer.

The association between cancer mortality and socioeconomic determinants, on the other hand, is more complicated. Individuals from low socioeconomic origins are more likely to receive a diagnosis of cancer and have poorer cancer outcomes than their more developed counterparts, according to research.

According to one study, cancer patients in the poorest parts of US have a 30% greater chance of dying from malignancy than those in the most impoverished areas.

Low income, a low education level, and a lack of transportation to medical facilities are socioeconomic characteristics that have been linked to worse cancer outcomes. These variables can affect cancer prognosis through several methods, including delaying the beginning and end of treatment, decreasing adherence to prescribed regimens, and raising the chance of comorbid diseases, which can complicate cancer therapy.

In conclusion, past research has proven a strong positive association between cancer mortality and death rates, as well as the importance of socioeconomic variables in predicting cancer outcomes. The current study seeks to extend previous research by employing a multiple-linear regression

approach to investigate the association between cancer-related death rates and sociodemographic characteristics to find the most relevant determinants of cancer consequences.

Data Description

The cancer dataset that includes the key attributes that are related to the cancer death is collected form the Kaggle. The name of the key attributes in the dataset is shown in the below picture

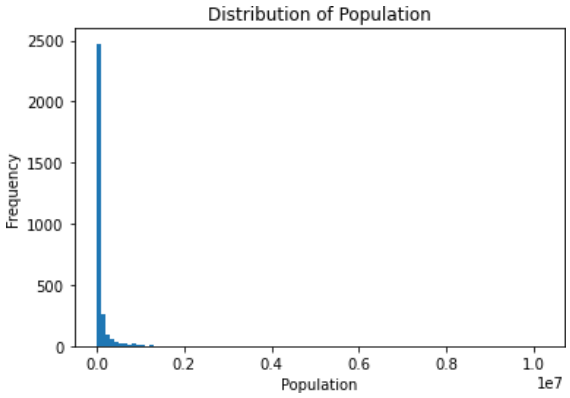
```
Index(['County', 'Population', 'deathRate', 'incidenceRate', 'medIncome',
       'povertyPercent', 'MedianAge', 'MedianAgeMale', 'MedianAgeFemale',
       'AvgHouseholdSize', 'PctMarriedHouseholds', 'PctNoHS18_24',
       'PctHS18_24', 'PctBachDeg18_24', 'PctHS25_Over', 'PctBachDeg25_Over',
       'PctUnemployed16_Over', 'PctPrivateCoverage', 'PctEmpPrivCoverage',
       'PctPublicCoverage', 'PctPublicCoverageAlone', 'PctWhite', 'PctBlack',
       'PctAsian', 'PctOtherRace'],
      dtype='object')
```

The proposed dataset has a total of 25 attributes and a total of 3,047 rows. The dataset's properties show the medical and demographic aspects of several towns throughout the country. The dataset contains essential parameters such as incidence rate, population size, mortality rate, population age, poverty percentage, the proportion of married families, average residential size, and many more.

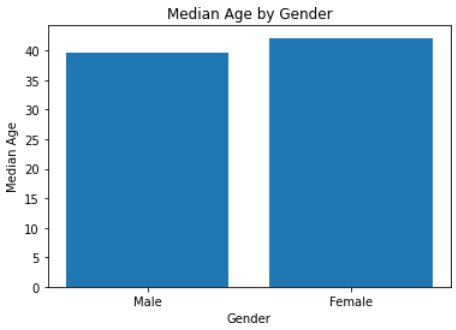
	count	mean	std	min	25%	50%	75%	max
Population	3047.0	102637.370528	329059.220504	827.000000	11684.000000	26643.000000	68671.000000	1.017029e+07
deathRate	3047.0	178.664063	27.751511	59.700000	161.200000	178.100000	195.200000	3.628000e+02
incidenceRate	3047.0	445.654447	57.456583	201.300000	413.150000	449.500000	482.050000	1.206900e+03
medIncome	3047.0	47063.281917	12040.090836	22640.000000	38882.500000	45207.000000	52492.000000	1.256350e+05
povertyPercent	3047.0	16.878175	6.409087	3.200000	12.150000	15.900000	20.400000	4.740000e+01
MedianAge	3047.0	45.272333	45.304480	22.300000	37.700000	41.000000	44.000000	6.240000e+02
MedianAgeMale	3047.0	39.570725	5.226017	22.400000	36.350000	39.600000	42.500000	6.470000e+01
MedianAgeFemale	3047.0	42.145323	5.292849	22.300000	39.100000	42.400000	45.300000	6.570000e+01
AvgHouseholdSize	3047.0	2.529682	0.248449	1.860000	2.380000	2.500000	2.640000	3.970000e+00
PctMarriedHouseholds	3047.0	51.243872	6.572814	22.992490	47.763063	51.669941	55.395132	7.807540e+01
PctNoHS18_24	3047.0	18.224450	8.093064	0.000000	12.800000	17.100000	22.700000	6.410000e+01
PctHS18_24	3047.0	35.002068	9.069722	0.000000	29.200000	34.700000	40.700000	7.250000e+01
PctBachDeg18_24	3047.0	6.158287	4.529059	0.000000	3.100000	5.400000	8.200000	5.180000e+01
PctHS25_Over	3047.0	34.804660	7.034924	7.500000	30.400000	35.300000	39.650000	5.480000e+01
PctBachDeg25_Over	3047.0	13.282015	5.394756	2.500000	9.400000	12.300000	16.100000	4.220000e+01
PctUnemployed16_Over	3047.0	7.852412	3.452371	0.400000	5.500000	7.600000	9.700000	2.940000e+01
PctPrivateCoverage	3047.0	64.354939	10.647057	22.300000	57.200000	65.100000	72.100000	9.230000e+01
PctEmpPrivCoverage	3047.0	41.196324	9.447687	13.500000	34.500000	41.100000	47.700000	7.070000e+01
PctPublicCoverage	3047.0	36.252642	7.841741	11.200000	30.900000	36.300000	41.550000	6.510000e+01
PctPublicCoverageAlone	3047.0	19.240072	6.113041	2.600000	14.850000	18.800000	23.100000	4.660000e+01

The municipalities in the dataset have populations ranging from 827 to 10,170,292, with a mean of 102,637. The death rate varies from 59.7 to 362.8 per 100,000 people, with a mean of 178.7, while the number of new cases ranges from 201.3 to 1,206.9 per 100,000 people, with a mean of 445.7. The median household income varies from $22,640 to $125,635, with a mean of $47,063, while the poverty rate is from 3.20 to 47.40 percent, with an average of 16.88 percent. The citizenry's median age spans from 22.30 to 64.70 years old, with a mean of 45.27, while the typical size of a household extends from 1.86 to 3.97 years old, with a mean of 2.53.

Visualization of data

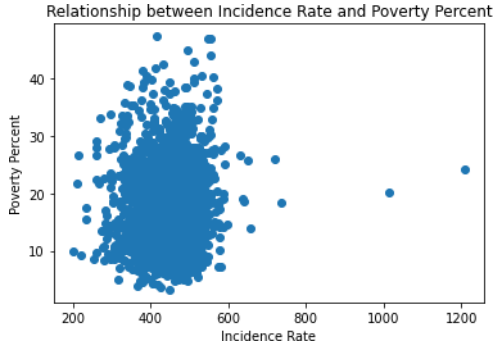


The above visual plot demonstrates the distribution of population. The distribution of population is clustered below $0.2 \times 1e^7$.

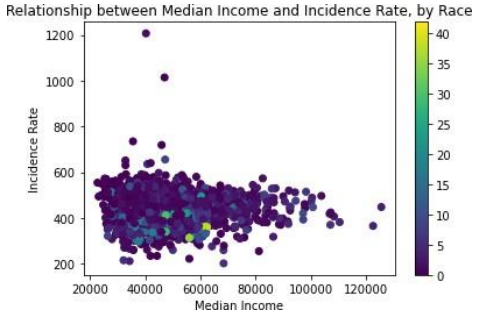


The bar plot above exhibits the median age of the “Male” and “Female” in the proposed sample dataset. The median ages of female are higher than the male.

The scatter plot below shows the linear relationship between "incidence rate" and "poverty percent" [3.]. With poverty percentage is increasing with increase in the incidence rate.

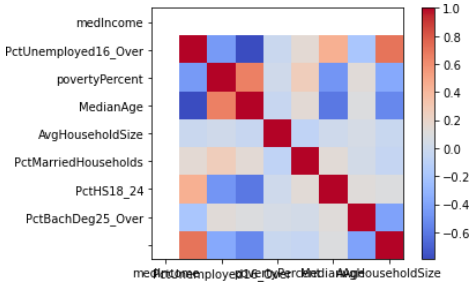


The accompanying scatter plot shows that the incidence rate is predictably and positively connected to the Median Income.



Plot of Correlation

Correlation plots are useful when looking for qualities that are highly connected to the target variable. The "death rate" variable is the goal variable, whereas the remaining set of features in the data are additional factors. The correlation plot displays the characteristics that are highly or weakly associated with the target variable. Some of the important factors that are positively connected with the "death rate" include the number of new cases, the percentage of examinations, and the poverty proportion.



Predictive Models

Linear Regression 1

The goal variable "death rate" is positively associated with the incidence rate. A regression model is created using the simple linear regression idea, using "death rate" as the goal variable and "incidence Rate" as the explanatory variables.

OLS Regression Results

Dep. Variable:	mortalityRate	R-squared (uncentered):	0.950
Model:	OLS	Adj. R-squared (uncentered):	0.950
Method:	Least Squares	F-statistic:	5.735e+04
Date:	Wed, 15 Mar 2023	Prob (F-statistic):	0.00
Time:	20:27:38	Log-Likelihood:	2958.7
No. Observations:	3047	AIC:	-5915.
Df Residuals:	3046	BIC:	-5909.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
incidenceRate	0.0009	3.69e-06	239.482	0.000	0.001	0.001

Omnibus:	654.587	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7474.521
Skew:	0.684	Prob(JB):	0.00
Kurtosis:	10.550	Cond. No.	1.00

The accompanying image depicts an overview of the multiple linear regression with an R-square value of 0.95. The set of independent variable can account for more than 95% of the variability in the output variables. Furthermore, the calculated coefficient of the explanatory variables is 0.009, indicating that a unit change there in the incidence rate raises the value of the death rate by 0.009. The model's predicted coefficient is significant since the obtained p-value is less than 0.05. Overall, the above-derived regression model has significant findings since the parameters AIC & BIC have the lowest value and the F-test p-value is less than the statistical significance. Generally, there is a considerable positive and linear relationship between "incidence Rate" and "death rate."

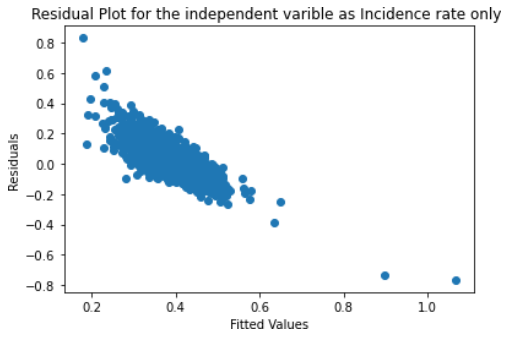


Figure 1: Residual Plot for the Regression Analysis Between Incidence Rate and DeathRate

Second Linear Regression Model

The regression model with many independent characteristics was developed using multiple linear regression techniques. The image below displays a summary of the regression model.

OLS Regression Results

Dep. Variable:	mortalityRate	R-squared:	0.402
Model:	OLS	Adj. R-squared:	0.400
Method:	Least Squares	F-statistic:	226.7
Date:	Wed, 15 Mar 2023	Prob (F-statistic):	0.00
Time:	20:28:05	Log-Likelihood:	5026.2
No. Observations:	3047	AIC:	-1.003e+04
Df Residuals:	3037	BIC:	-9972.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.6361	0.017	36.784	0.000	0.602	0.670
incidenceRate	-0.0004	1.54e-05	-27.904	0.000	-0.000	-0.000
medIncome	2.699e-07	1.59e-07	1.700	0.089	-4.14e-08	5.81e-07
PctUnemployed16_Over	0.0008	0.000	2.235	0.025	9.58e-05	0.001
povertyPercent	0.0018	0.000	5.952	0.000	0.001	0.002
MedianAge	-1.724e-05	1.88e-05	-0.919	0.358	-5.4e-05	1.95e-05
AvgHouseholdSize	-0.0076	0.004	-1.785	0.074	-0.016	0.001
PctMarriedHouseholds	-0.0008	0.000	-4.241	0.000	-0.001	-0.000
PctHS18_24	0.0007	0.000	6.385	0.000	0.000	0.001
PctBachDeg25_Over	-0.0040	0.000	-15.326	0.000	-0.004	-0.003

Omnibus:	685.327	Durbin-Watson:	1.960
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7218.370
Skew:	0.759	Prob(JB):	0.00
Kurtosis:	10.386	Cond. No.	9.97e+05

Overall, the resulting regression model is statistically significant, with an F-Statistics value of 226.7 and a p-value less than 0.05. The model's r-square value is 0.40. The set of exogenous variables may explain more than 40% of the fluctuation in the target variable. Several characteristics in the preceding regression model have estimated coefficients that are statistically meaningful at the 0.05 significance level.

Regression Model III

The third statistical method is created by combining the logarithmic conversions of "incidence Rate," "med Income," and "poverty Percent" with additional independent variables.

OLS Regression Results

Dep. Variable:	mortalityRate	R-squared (uncentered):	0.986
Model:	OLS	Adj. R-squared (uncentered):	0.986
Method:	Least Squares	F-statistic:	9053.
Date:	Wed, 15 Mar 2023	Prob (F-statistic):	0.00
Time:	20:28:29	Log-Likelihood:	4877.3
No. Observations:	3047	AIC:	-9709.
Df Residuals:	3024	BIC:	-9570.
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Population	-1.055e-08	3.21e-09	-3.287	0.001	-1.68e-08	-4.26e-09
incidenceRate	-0.0004	1.69e-05	-23.749	0.000	-0.000	-0.000
medIncome	2.943e-07	2.03e-07	1.452	0.147	-1.03e-07	6.92e-07
povertyPercent	0.0042	0.000	12.548	0.000	0.004	0.005
MedianAge	-2.509e-06	1.98e-05	-0.127	0.899	-4.14e-05	3.63e-05
MedianAgeMale	0.0006	0.001	1.100	0.271	-0.000	0.002
MedianAgeFemale	0.0017	0.001	3.016	0.003	0.001	0.003
AvgHouseholdSize	0.0591	0.005	10.776	0.000	0.048	0.070
PctMarriedHouseholds	-0.0006	0.000	-2.552	0.011	-0.001	-0.000
PctNoHS18_24	0.0003	0.000	1.835	0.067	-1.75e-05	0.001
PctHS18_24	0.0007	0.000	5.431	0.000	0.000	0.001
PctBachDeg18_24	-0.0002	0.000	-0.689	0.491	-0.001	0.000
PctHS25_Over	0.0016	0.000	6.549	0.000	0.001	0.002
PctBachDeg25_Over	-0.0010	0.000	-2.755	0.006	-0.002	-0.000
PctUnemployed16_Over	0.0002	0.000	0.594	0.552	-0.001	0.001
PctPrivateCoverage	0.0019	0.000	6.645	0.000	0.001	0.002
PctEmpPrivCoverage	0.0002	0.000	0.760	0.447	-0.000	0.001
PctPublicCoverage	-0.0010	0.001	-1.854	0.064	-0.002	5.83e-05
PctPublicCoverageAlone	0.0036	0.001	5.405	0.000	0.002	0.005
PctWhite	0.0006	0.000	4.465	0.000	0.000	0.001
PctBlack	0.0005	0.000	3.395	0.001	0.000	0.001
PctAsian	0.0013	0.000	2.821	0.005	0.000	0.002
PctOtherRace	-0.0007	0.000	-2.303	0.021	-0.001	-0.000

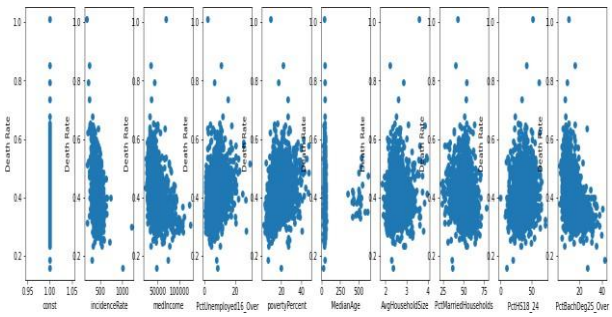
Omnibus:	562.063	Durbin-Watson:	1.970
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4499.965
Skew:	0.647	Prob(JB):	0.00
Kurtosis:	8.811	Cond. No.	2.14e+06

Overall, the resulting regression model is statistically significant, with an F-Statistics value of 9053 and a p-value less than 0.05. The model's r-square value is 0.986. The set of individual variables may explain more than 98% of the fluctuation in the target variable. Several characteristics in the preceding regression model have estimated coefficients that are clinically meaningful at the 0.05 level of significance.

Diagnostics of Regression Models

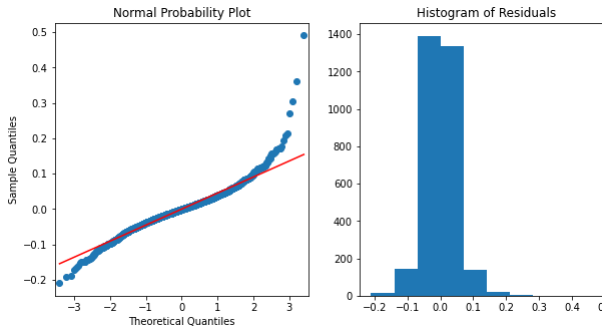
Regression with Multiple Regression Model 1

Linearity Check



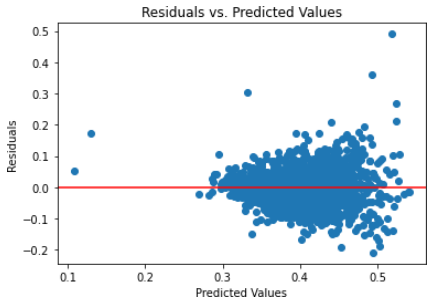
The above plot demonstrates the presence of linearity with the target variable for each of the independent variable in the derived regression model.

Normality Check



The data points are unevenly distributed and do not follow the QQ mean line. As a result, residuals are not regularly distributed.

Homoscedasticity

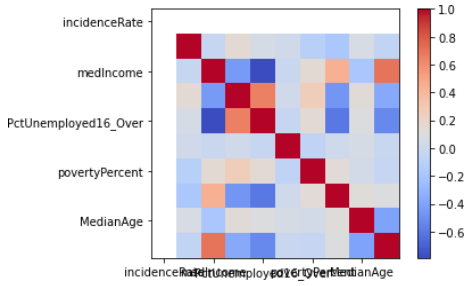


The scattering of residual pieces of information along the horizontal line demonstrates that the dispersion of residual is constant. The dots in the above scale-location figure for simple regression I are distributed randomly around a diagonal line at the median of the original number of the standardized residuals. This implies that the residual variability is constant over the expected value range [1.].

	VIF Factor	feature
0	420.125711	const
1	1.097300	incidenceRate
2	5.133082	medIncome
3	2.037800	PctUnemployed16_Over
4	5.193429	povertyPercent
5	1.013960	MedianAge
6	1.581678	AvgHouseholdSize
7	2.150909	PctMarriedHouseholds
8	1.255902	PctHS18_24
9	2.715702	PctBachDeg25_Over

The above image depicts the summary of the VIF test, which shows certain variables such as characteristics at the mean of the original number of the standardized residuals. This implies that the residual's coefficient of variation is constant over the expected value range.

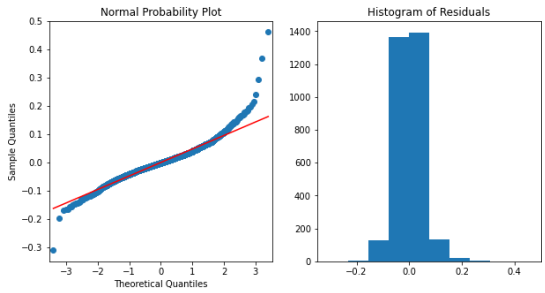
Test for Multicollinearity



The results of the VIF test are displayed in the above image, which shows that various variables, such as PctPrivateCoverage, PctPublicCoverage, and PctPublicCoverageAlone, have VIF values of more than 15. As a result, these characteristics are closely associated with other predictive factors. As a result, there is a cointegration relationship among variables.

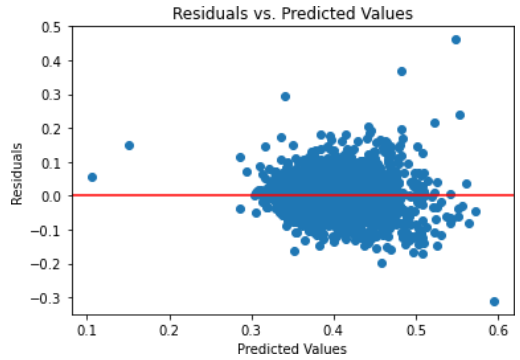
Multiple Regression Model II

Normality Test



The data elements are randomly distributed and do not follow the QQ mean line. As a result, residuals are not regularly distributed.

Homoscedasticity



The scattering of residual data points along the diagonal line demonstrates that the variability of residual is constant. The dots in the above scale-location figure for regression model II are distributed randomly around a diagonal line at the mean of the original number of the standardized residuals. This implies that the residual's variance is constant over the expected value range.

Conclusion

Multiple regression analysis models have been created in this study to determine the "death rate" caused by cancer. Each model's determination factor is unique. The third regression model created by considering "death rate" as the dependent variable and several independent variables with some proportional treatment of some features yields the greatest R-Square value. The r-square value of this model is 0.986. "incidence rate" and "povertyPercent" is the most important independent characteristics.

References

[1.] Battineni, G., Sagaro, G.G., Chinatalapudi, N. and Amenta, F., 2020. Applications of machine learning predictive models in the chronic disease diagnosis. Journal of personalized medicine, 10(2), p.21.

[2.] Chien, L.H., Chen, C.H., Chen, T.Y., Chang, G.C., Tsai, Y.H., Hsiao, C.F., Chen, K.Y., Su, W.C., Wang, W.C., Huang, M.S. and Chen, Y.M., 2020. Predicting lung cancer occurrence in never-smoking females in Asia: TNSF-SQ, a prediction model. Cancer Epidemiology, Biomarkers & Prevention, 29(2), pp.452-459.

[3.] Ijaz, M.F., Attique, M. and Son, Y., 2020. Data-driven cervical cancer prediction model with outlier detection and over-sampling methods. Sensors, 20(10), p.2809.