Project Report

Topic: Predicting cancer mortality rate per capita based on socio-economic factors

By

Aswathi Jayakumar	AXJ210107
Dalia Debbarma	DXD210063
Muhammad Shahab Memon	MXM200032
Shagun Gupta	SXG210168
Shantanu Singh	SXS2200258
Vinay Bhanushali	VDB220000

Introduction

In the realm of healthcare and public health research, understanding the intricate interplay between socio-economic factors and cancer mortality is of paramount importance. In pursuit of this knowledge, the present project delves into a comprehensive real-life dataset named cancer-reg [32], a repository of information on cancer trials in the United States of America (U.S.). This dataset, meticulously compiled, encompasses a sample size of n=3047 individuals, each characterized by 32 predictor variables. The dataset serves as a rich tapestry of information, offering a nuanced perspective on the multifaceted dimensions of cancer and its correlation with socio-economic status.

Objectives

The primary objective of this endeavour is to scrutinize the intricate relationships existing between the socio-economic status of individuals in the U.S. and the average per capita cancer mortality rates. By elucidating these connections, our study aspires to provide valuable insights that can prove instrumental for various stakeholders, including government health agencies, cancer research institutions, and life insurance agencies. The implications of this research extend beyond academia, promising real-world applications that stand to inform policy decisions and enhance public health initiatives.

Hypothesis

In formulating our research hypothesis, we posit that several key factors, including but not limited to a person's income, age, family size, and race, wield a substantial influence on the cancer death rate. Through a meticulous examination of these variables, we aim to unravel patterns and correlations that can shed light on the complex web of socioeconomic determinants impacting cancer mortality. The chosen dependent variable, 'Target_deathRate,' serves as a crucial metric, while an array of independent variables encompassing socio-economic indicators adds depth and granularity to our exploration.

As we embark on this analytical journey through the cancer-reg [32] dataset, our pursuit is not merely academic but driven by a broader vision to contribute substantively to the discourse surrounding cancer epidemiology and its intersection with socio-economic dynamics. This report encapsulates our commitment to unravelling the intricacies of cancer mortality, offering meaningful implications for both research and practical applications in the broader landscape of public health.

Source of Data

As we embark on this analytical journey through the cancer-reg [32] dataset, it is imperative to acknowledge the sources from which this data emanates. The dataset is hosted on the website https://data.world/nrippner/ols-regression-challenge, and its origins can be traced back to data aggregation from prominent sources, including the American Community Survey (https://www.census.gov), https://www.clinicaltrials.gov, and https://www.cancer.gov. This collaborative amalgamation of data from diverse sources amplifies the richness and reliability of our dataset, reinforcing the robustness of our exploration into the complex dynamics of cancer mortality and socio-economic factors.

Variable Selection

Our dependent variable, Target_deathRate, is defined as the mean per capita (100,000) cancer mortalities (a).

Independent variables and Column Description

TARGET_deathRate: Dependent variable. Mean per capita (100,000) cancer mortalities(a)
avgAnnCount: Mean number of reported cases of cancer diagnosed annually(a)
avgDeathsPer Year: Mean number of reported mortalities due to cancer(a)
incidenceRate: Mean per capita (100,000) cancer diagnoses(a)
medianIncome: Median income per county (b)
popEst2015: Population of county (b)

povertyPercent: Percent of populace in poverty (b)
studyPerCap: Per capita number of cancer-related clinical trials per county (a)
binnedInc: Median income per capita binned by decile (b)
MedianAge: Median age of county residents (b)
MedianAgeMale: Median age of male county residents (b)
MedianAgeFemale: Median age of female county residents (b)
Geography: County name (b)
AvgHouseholdSize: Mean household size of county (b)
PercentMarried: Percent of county residents who are married (b)
PctNoHS18_24: Percent of county residents ages 18-24 highest education attained: less than high school (b)
PctHS18_24: Percent of county residents ages 18-24 highest education attained: high school diploma (b)
PctSomeCol18_24: Percent of county residents ages 18-24 highest education attained: some college (b)
PctBachDeg18_24: Percent of county residents ages 18-24 highest
education attained: bachelor's degree (b)
PctHS25_Over: Percent of county residents ages 25 and over highest
education attained: high school diploma (b)
PctBachDeg25_Over: Percent of county residents ages 25 and over highest education attained: bachelor's degree (b)
PctEmployed 16 Over: Percent of county residents ages 16 and over employed (b)
PctUnemployed16_Over: Percent of county residents ages 16 and over unemployed (b)
PctPrivateCoverage: Percent of county residents with private health coverage (b)
PctPrivateCoverageAlone: Percent of county residents with private health coverage alone (no public assistance) (b)
PctEmpPrivCoverage: Percent of county residents with employee-provided private health coverage (b)
PctPublicCoverage: Percent of county residents with government-provided
provided health coverage alone (b)
PctWhite: Percent of county residents who identify as White (b)
PctBlack: Percent of county residents who identify as Black (b) PctAsian: Percent of county residents who identify as Asian (b)
PctOtherRace: Percent of county residents who identify in a category which is not White, Black, or Asian (b)
PctMarried Households: Percent of married households (b)
BirthRate: Number of live births relative to number of women in county (b)
(a): years 2010-2016
(b): 2013 Census Estimates

Data Cleaning Fundamentals -

Data cleaning is a crucial step in the data analysis process, involving the identification and correction of errors and inconsistencies to improve the quality of data. This step is crucial because obtaining reliable analysis insights and avoiding erroneous judgments depends on the accuracy of the dataset. By cleaning data, we ensure the integrity of analytics outcomes, which supports informed decision-making. Data cleaning includes various tasks from correcting typographical errors to addressing issues like missing or outlier values.

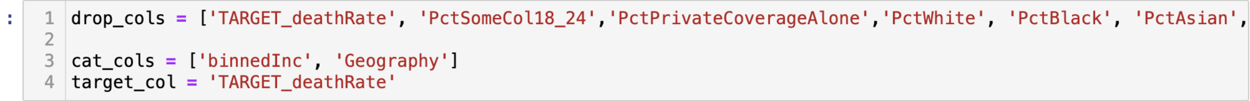
Identified Data Anomalies-

Upon inspecting the cancer mortality dataset, several anomalies have been identified that could potentially skew the results. These include missing or null values, numerical outliers and unwanted data observations. Each of these anomalies can introduce bias or inaccuracies into subsequent analyses if not properly addressed, hence their identification is the first critical step in the data cleaning process.

Data Cleaning and Validation on Model

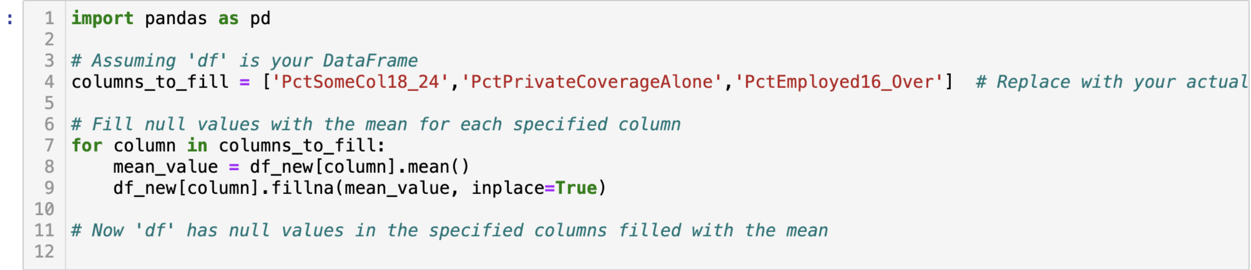
To ensure good quality data, specific data cleaning actions are undertaken for cancer mortality regression model. We began data cleaning by creating a fresh copy of the data to preserve the original. We listed the columns to be excluded from the independent variables under two variables: 'drop_cols' and 'cat_cols'. We removed the 'TARGET_deathRate' column along with other less relevant columns like 'PctSomeCol18_24', 'PctPrivateCoverageAlone', 'PctWhite', 'PctBlack', 'PctAsian', and 'PctOtherRace', 'binnedInc', 'Geography' to narrow our focus to the most impactful variables. Next, we identified the columns which contained null values or missing values in the dataset and stored them as ‘columns_to_fill’. Using the mean of the columns, we filled in any missing numbers to ensure completeness and consistency.
The "Target_deathRate" column was designated as the dependent variable, and numerical null values were replaced with mean values to handle missing data. This approach aids in maintaining statistical integrity while preparing the data for model validation.

This step was crucial to prepare the dataset for accurate modeling, as gaps or missing data can lead to incorrect analysis. Finally, we set up our variables for the regression model, identifying 'TARGET_deathRate' as the outcome we wanted to predict and organizing the remaining data for the analysis.



```
drop_cols = ['TARGET_deathRate', 'PctSomeCol18_24','PctPrivateCoverageAlone','PctWhite', 'PctBlack', 'PctAsian',

cat_cols = ['binnedInc', 'Geography']
target_col = 'TARGET_deathRate'
```



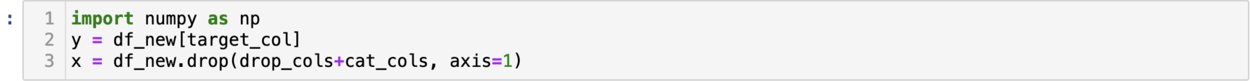
```
import pandas as pd

# Assuming 'df' is your DataFrame
columns_to_fill = ['PctSomeCol18_24','PctPrivateCoverageAlone','PctEmployed16_Over']  # Replace with your actual

# Fill null values with the mean for each specified column
for column in columns_to_fill:
    mean_value = df_new[column].mean()
    df_new[column].fillna(mean_value, inplace=True)

# Now 'df' has null values in the specified columns filled with the mean

```



```
import numpy as np
y = df_new[target_col]
x = df_new.drop(drop_cols+cat_cols, axis=1)
```

Testing Models:

Model 1:

Ordinary Least Squares regression (OLS) is a common technique for estimating coefficients of linear regression equations which describe the relationship between one or more independent quantitative variables and a dependent variable (simple or multiple linear regression).

Upon inspecting the results from the OLS Regression, the variable with most significance is: PercentMarried with the coefficient of 1.3030. The R-squared value suggests that the model explains 52.3% of the variation in the death rate. The F-statistic is highly significant, with a p-value of 0.000, which means that the model is statistically significant as a whole.

The Durbin-Watson statistic is 1.781, which suggests that autocorrelation is not a major problem in the model. However, the Jarque-Berá statistic is highly significant, which suggests that the residuals are not normally distributed. This is not necessarily a major problem, but it is something to be aware of when interpreting the results.

However, The condition number (cond. no.) is 1.58e+07, which is very high. This indicates that the predictor variables in the model are highly collinear. Multicollinearity can make the regression model unstable and difficult to interpret. It is important to address multicollinearity before interpreting the results of the regression model.

To mitigate these problems, Model 2 and Model 3 are deployed.

Model 2:

Dealing with Multicollinearity:

Interpreting the OLS regression results upon the Model 2, the variables avgDeathsPerYear, popEst2015, MedianAgeFemale, PercentMarried, PctPrivateCoverage, PctPublicCoverage, PctPublicCoverageAlone posed as High VIF variables. Variance inflation factor (VIF) is a measure of the amount of Multicollinearity in a set of multiple regression variables. Hence, these variables were removed.

Dealing with Heteroscedasticity:

Heteroscedasticity occurs when the standard deviations of a predicted variable, monitored over different values of an independent variable or as related to prior time periods, are non-constant.
The Breush-Pagan test (This is a formal test for heteroscedasticity) was conducted to test for Heteroscedasticity. To deal with Heteroscedasticity, the variables and their squared terms were regressed. This resulted in Homoscedasticity, hence satisfying the conditions of OLS regression.

Model 2 + adding back Geography (states) as a dummy variable:

Adding the Geography (states) as a dummy variable to the Model 2 increased the variable count to 102. Regressing the model with Geography as a dummy variable increased the R-Squared value from 52.3% to 60.6%, improving the explanation of variation in the dependent variable: Death rate.

Conclusion

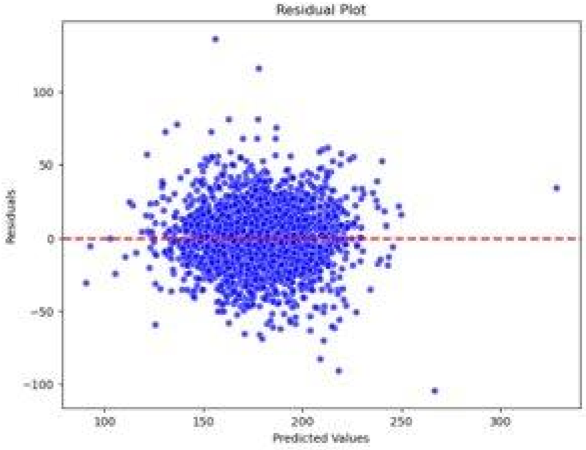
After carefully improving the model to the best of our ability, we now move on to reaffirm the OLS conditions of the model. These conditions are imperative to establish a causal relationship between the dependent and independent variables.

Condition 1 Linear in parameters: Condition met by taking in account squared terms

Condition 2 No perfect collinearity: Condition met by checking for collinearity using VIF and removing highly collinear columns

Condition 3 Random Sample: We assume the sample was taken randomly.

Condition 4 Zero expected mean value: Below graph affirms errors are zero with mean.



Condition 5 Homoscedastic: The data had heteroscedastic as suggested by the Breusch-Pagan test. For which our p-value came out to be 1.8e-21. This was taken care of by scaling independent variables.

Since all the conditions of an OLS are met, we can assume the estimators are **Best Linear Unbiased Estimates** in other words **ceteris paribus**. Now we are in a good position to find top factors which affect the cancer mortality rate.

Top 5 factors that affect cancer mortality rate the most:

1. Median Age (7.241)
2. Percent Married Households (-5.2)
3. Percent Married (5.185)
4. Percent Household 25 and over (2.0984)
5. Percent employed 16 and over (-1.617)

Top 5 States with that affect cancer mortality rate the most:

1. Hawaii (-35.686)
2. Kansas (26.200)
3. Nevada (21.439)
4. Minnesota (17.162)
5. Utah (-17.343)

In conclusion, socio economic factors do affect cancer mortality rate. As we can see Marital status, Age and employment of 16 and above individuals are major factors in cancer mortality rate. Other than that, geographic location also plays a major role in cancer mortality rate.