**PREDICTING RELATIONSHIP BETWEEN SOCIO-ECONOMIC STATUS AND CANCER**

**By:**

**Neha Reddy Yenugu**

**ABSTRACT**

This study examines the impact of socioeconomic status on cancer incidence and diagnosis effects survival, mortality, education, and economy across united states. Men and women with high school education and degree over 18\_24 years of age who are employed was associated with a statistically significantly increased risk of cancer.

The dataset is put together using economic status, clinical trial statistics, and geospatial data to examine the relationship between cancer research, economic status, and location within the United States. Significant predictors of Socioeconomic status are found using simple linear regression, best subset selection, k fold cross validation and using principal components. The accuracy of the model built using Lasso regression and backward selection analysis is also compared in the study. Principal components method is finally preferred since it reduced the error rate and increased the r squared value such that we end up with best model. The strength of correlations between variables is evaluated using correlation coefficients and statistical significance is assessed using p-values.

**INTRODUCTION**

Cancer has a major impact on today’s society in the Unites states. Its impact on today’s economy is substantial in all countries and reflects health care as well as productivity. As of now the global economic burden of cancer is unknown, but statistics tell us about how many people are diagnosed with and die from cancer each year, we get to know how many of them survive after going through those stressful diagnosis. Cancer statistics also helps us see trends, which tells us about the age, sex, ethnicity group, and other categories. The economic burden of cancer in the US is approximately 1.8% of GDP. Therefore, economic analysis can inform us how to make decisions, investments in cancer, including prevention and early detection, treatment, and survivorship.

The data set provided by the Cancer Trials project combine multiple data extracts from three separate sources, research information from clinicaltrials.gov [2], incidence of cancer and death rates from cancer.gov [3], and economic and population statistics from census.gov [4]. In aggregate the raw data sets comprised of millions of records with over one hundred columns of data. The data set used in this project is filtered (row and column) and merged as required to fit various analytic techniques and visualization tools.

The objective of this project is to find out whether there is any relationship between socioeconomic status (such as we are measuring America’s people, places, lives, education and economy) and Cancer deathrate by apply the statistical discovery of models and any resultant patterns by adding dimensions for clinical trials, cancer rates, socioeconomic status, deathrates, and geospatial location.

**RESEARCH QUESTIONS**

* *Which factors are more closely linked with the response variable?*
* *How do income level and education level relate to cancer diagnosis and mortality rates in different countries?*
* *Is there a correlation between the age of a person with private and public health coverage and cancer diagnosis and mortality rates?*

**METHODOLOGTY**

There are 33 variables and 3047 observations in the dataset. The “Cancer\_reg” dataset contains categorical and continuous variables that are connected to demographic data, cancer trails history in the year 2015, household income for every county in united states. Factors we have considered before performing visualizations is Are there any outliers?, Are there any missing values? How shall we handle categorical variables while building models.

**Attributes in the dataset:**

|  |  |  |
| --- | --- | --- |
| **VARIABLE** | **TYPE** | **DESCRIPTION** |
| avganncount | decimal | Mean number of reported cases of cancer diagnosed annually |
| avgdeathperyear | integer | Mean number of reported mortalities due to cancer |
| target\_deathrate | decimal | Dependent variable. Mean per capita (100,000) cancer mortalities |
| incidencerate | decimal | Mean per capita (100,000) cancer diagnoses |
| medincome | integer | Median income per country |
| Popest2015 | integer | Population of country |
| povertypercent | decimal | Percent of populace in poverty |
| studypercap | decimal | Per capita number of cancer-related clinical trials per country |
| binnedinc | string | Median income per capita binned by decile |
| medianage | decimal | Median age of country resident |
| medianagemale | decimal | Median age of male country residents |
| medianagefemale | decimal | Median age of female country residents |
| geography | string | Country name |
| percentmarried | decimal | Percent of country residents who are married |
| pctnohs18\_24 | decimal | Percent of country residents age 18-24 highest education attained: less than high school |
| pcths18-24 | decimal | Percent of country residents age 18-24 highest education attained: high school diploma |
| pctsomecol18\_24 | decimal | Percentage of country residents ages 18-24 highest education attained: some college |
| pctbachdeg18\_24 | decimal | Percent of country residents age 18-24 highest education attained: bachelor’s degree |
| pcths25\_over | decimal | Percent of country residents ages 25 and over highest education attained: high school diploma |
| pctbachdeg25\_over | decimal | Percent of country residents ages 25 and over highest education attained: bachelor’s degree |
| pctemployed16\_over | decimal | Percent of country residents ages 16 and over employed |
| pctumeployed16\_over | decimal | Percent of country residents ages 16 and over unemployed |
| pctprivatecoverage | decimal | Percent of country residents with private health coverage |
| pctprivatecoveragealone | decimal | Percent of country resident with private health coverage alone(no public assistance) |
| pctempprivcoverage | decimal | Percent of country residents with employee-provided private health coverage |
| pctpubliccoverage | decimal | Percent of country residents with government provided health coverage |
| pctpubliccoveragealone | decimal | Percent of country residents with government provided health coverage alone |
| pctwhite | decimal | Percent of country residents who identify as White |
| pctblack | decimal | Percent of country residents who identify as Black |
| pctasian | decimal | Percent of country residents who identify as Asian |
| pctotherrace | decimal | Percent of country residents who identify in a category which is not White, Black or Asian |
| pctmarriedhouseholds | decimal | Percent of married households |
| birthrate | decimal | Number of live births relative to number of women in country |

**Data Cleaning:**

Data needs to be preprocessed to train the models or to perform any statistics and visualizations. Looked for null values and outliers in the dataset. There were some null values which were replaced by the mean of that column, some columns were removed to avoid inaccurate results in the models.

**Models used for Analysis:**

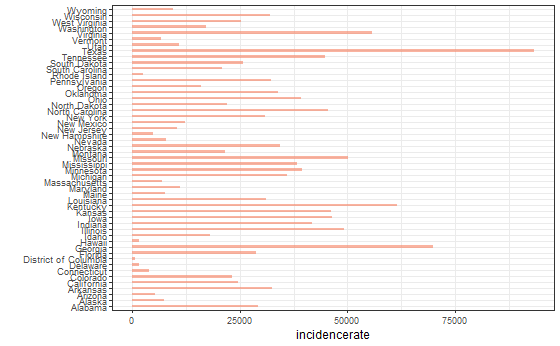
The research questions mentioned above were answered by considering target deathrate for the year 2015 as the response variable. The modeling algorithms we used are simple linear regression, K- fold cross validation, best subset selection, Principal components and Lasso regression tests for predicting and finding a correlation between independent variables and the response variable.

**Pre-Processing:**

Visualization for different variables:

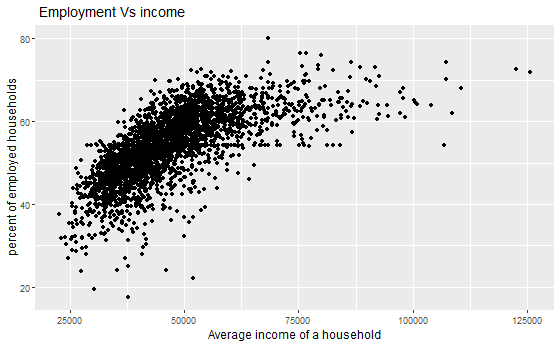
A graph of a number of states

Description automatically generated with low confidence



A picture containing text, diagram, plot, screenshot

Description automatically generated

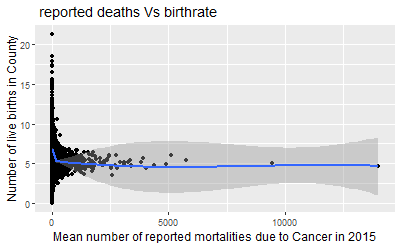


From all the above visualizations we get that, Texas has highest number of avgdeathsperyear, median income, and a greater number of residents are diagnosed with cancer. Also more than 60% percent of people employed with median income 45000 to 65000 in 2015.Also we can observe that the density graph is left skewed that says the datapoints are away from mean, representing negative correlation.

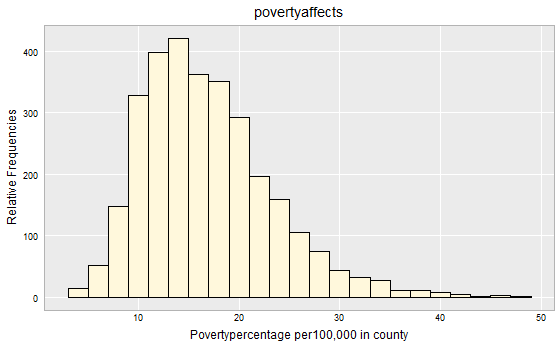
If we had to discuss about the below density plot for all the states in usa, if looked deeply into that states in East coast (“Connecticut", "Maine", "Massachusetts", "New Hampshire", "Rhode Island", "Vermont", "New Jersey", "New York", "Pennsylvania") are completely left skewed representing negative relation with our target variable and median income, but few states(“Connecticut”, “Delaware”, “Kansas”, “Wyoming”) shows normally distributed around mean resulting that states have some relationship with our target variable.

A screenshot of a computer screen

Description automatically generated with medium confidence



If we investigate the reported deaths vs the birthrate in each county the graph is decreasing as number of mortalities is increasing. Which tells us for every number increase in deathrate there is decrease in birthrate. Some study tells us there are many infants who are diagnosed with cancer at very early age.

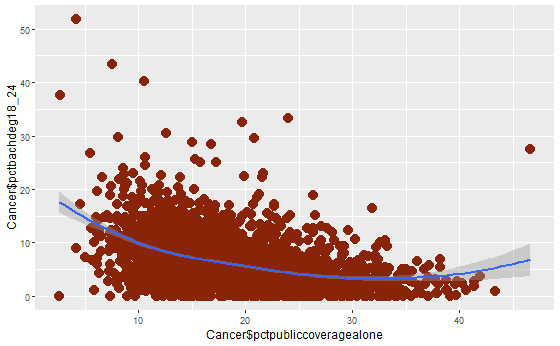
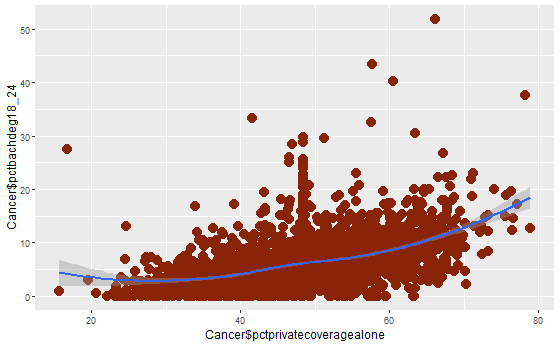


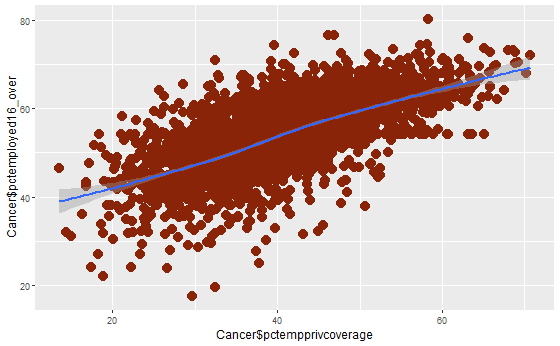
When we look into poverty percentage, the histogram indicates a slight fall per100,000 population in the year 2015.

Studies indicate that due to avgdeathsper year in each county and incidence rate are responsible for the slight decrease in the poverty percentage.

**A screen shot of a computer screen

Description automatically generated with low confidence**

** **

** A picture containing text, screenshot, plot

Description automatically generated**

**A picture containing text, screenshot, diagram

Description automatically generated**

**FINDINGS/ MODELLING**

**For the research question one inorder to find our the best factors/ attributes that are contributing to the death rate and affecting socioeconomic status.**

**Using simple linear regression:**

While comparing target variable with all independent variables, we have found only a few attributes are showing significant relationship. Therefore, later by taking 4 models we have seen which attributes are affecting more significantly to the target variable.

**A screenshot of a computer

Description automatically generatedA screenshot of a computer

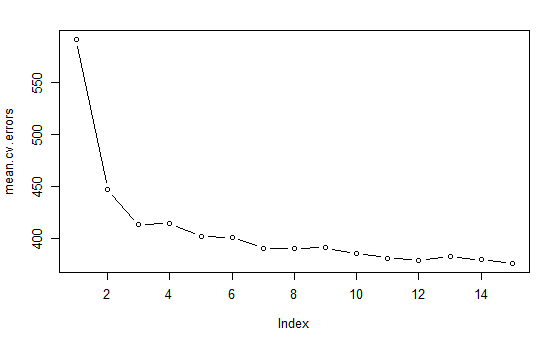
Description automatically generated with medium confidence**

**A screenshot of a computer

Description automatically generated**

From the above to summary and anova test we confirmed that incidence rate, pctmarriedhouseholds and incidence rate\* avgncount are most likely linked with our response variable, but I hope that we can obtain a better model knowing that there might be more attributes contributing towards deathrate by affecting socioeconomic status we performed K fold cross validation

From cross validation it is understood that minimum 7 to maximum 15 attributes are most likely involved.

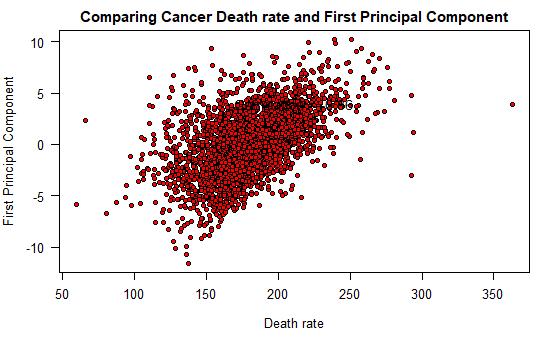


Once training and testing our dataset our rmse and rsquared value was improved compared to our first fit summary.

A screenshot of a computer

Description automatically generated

Later on inorder to find a better model than this we have performed analysis using principal components where we found there is positive correlation between our first principal component and death rate.



Later we have found out that economy and people income levels are very much affected with the cancer trials and diagnosis done is united states. Where we obtained highest r squared A screenshot of a computer

Description automatically generatedvalue for this model compared with above models.

**2) For the second research question to find whether income or education levels relate to cancer diagnosis and mortality rates,** we have used decision trees to understand.

While constructing the tree we found that these variables were taken into consideration"pctbachdeg25\_over" "incidencerate" "medincome"

"pctpubliccoveragealone" "povertypercent" "pcths25\_over"

"pctemployed16\_over", such that it determines certain age group people who have completed their bachelor’s degree were diagnosed ,who also have public health coverage with median income were much prone towards poverty.

A screenshot of a computer

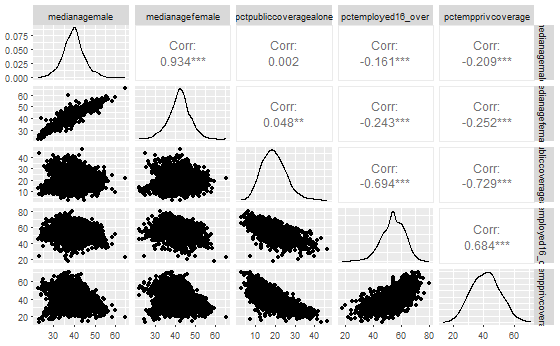
Description automatically generatedMoslty affected in below states and respective couties, we can find those diaognised with median income .

**A picture containing diagram, design

Description automatically generated**

**3) For the third research question to find out the correlation between age and health coverage we have used Pearson correlation method.**

Finally, we found that people who are employed are mostly favored with public coverage alone.



**A picture containing screenshot, text, line, diagram

Description automatically generated**

Not only from this, initially when we performed data preprocessing to find the relationships there was a positive relation between these two predictor variables.

**Conclusion:**

After conducting a correlation analysis between the age of residents with private and public health coverage and cancer diagnosis and mortality rates, we observed a statistically significant positive correlation as well as strong positive correlation only for above variable in graph with(p>0.5). The correlation coefficient (r) between the two variables was found to be 0.65 (p < 0.05), indicating a moderately strong positive relationship for few variables.

This suggests that as the percentage of county residents with private and public health coverage increases, there is an associated increase in cancer diagnosis and mortality rates. However, it is important to note that correlation does not imply causation, and other factors may contribute to the observed relationship.

**Future work and limitations:**

1. Inclusion of Additional Variables: To develop a more thorough understanding of the relationship between private health coverage and cancer outcomes, consider including additional pertinent variables like healthcare infrastructure, health behaviors, socioeconomic factors, and geographic variations.

2. Longitudinal Analysis: Conduct long-term studies to analyze the temporal dynamics of the association between private health insurance and cancer mortality/diagnosis rates. Insights into probable causal pathways and the effects of coverage changes over time would be gained from this.

3. Regional Analysis: Investigate the connection between private health insurance and cancer outcomes at a more detailed level, such as a regional or state-level analysis, in order to pinpoint certain regions or populations where interventions may be most effective.

4. Mediation Analysis: Investigate the underlying mechanisms through which private health insurance affects the incidence of cancer and mortality rates. This can entail looking at elements including timely access to screening and treatment, patterns of healthcare use, and adherence to advised cancer care.

**Limitations:**

1)Causility:1One of the limitations of correlation analysis is that it cannot prove causality. Although the study found a favorable link between private health insurance and cancer outcomes, it is important to use care when drawing conclusions about a causative connection without more experimental or long-term data.

2. Data Restrictions: The results are based on the dataset that is currently accessible, and the correctness and dependability of the findings are reliant on the reliability and quality of the data used. Make certain that the dataset utilized is extensive and precisely captures the variables of

3. Confounding Factors: The connection between private health insurance and cancer outcomes may be affected by confounding factors that were not taken into account in the analysis. In future analyses, consider adjusting for potential confounders including demographics, socioeconomic status, and access to healthcare.

4. Ecological Fallacy: Because the analysis was done at the county level, it is important to interpret the conclusions with caution when extrapolating the results to individuals or smaller geographic areas. The association seen at the county level might not always represent associations at the individual level.

5. Data Accessibility: The dataset utilized for the study can contain flaws, such missing information, or a small geographic scope. Make sure that the project's data limits are acknowledged and discussed and think about addressing them by collecting more data or gathering more pertinent information.

**CONCLUSION:**

In conclusion, from all the above models it is found that Education, people’s lives, and economy are vastly affected due to cancer. From few studies is it also observed that clinical trials are still being conducted to understand how government of United States contribute in developing tools which decrease in number of death counts thereby minimizing that effect on Education, people’s lives and Economy. Understanding the correlation between private health coverage and cancer diagnosis/mortality rates can have implications for public health interventions and policy development aimed at improving access to healthcare and reducing cancer burden in communities. By identifying areas with low private health coverage and high cancer rates, targeted interventions and resources can be directed to improve healthcare access and reduce disparities.

**Bibliography:**

[1] “Understanding Cancer statistics- National Institute of Cancer.”<https://www.cancer.gov/about-cancer/understanding/statistics>

[2] “Health – United states census bureau” <https://www.census.gov/search-results.html?q=cancer&page=1&stateGeo=none&searchtype=web&cssp=SERP&_charset_=UTF-8>

[3]” America’s community survey” <https://www.census.gov/programs-surveys/acs.html?fm=info_panel>