90/100

Project Proposal: Analyzing Socioeconomic Factors on USA Cancer Mortality

Jasen Zhang, Alex Zhu, Christian Pascual

Introduction

Approximately one in every four deaths in the U.S. is due to cancer, making it the second leading cause of death after heart disease. Billions of dollars are spent annually on the research, treatment and prevention of cancer so that we understand the key factors determining the cancer mortality and from there, to have accurate estimates of cancer cases and deaths.

Each year, the American Cancer Society (ACS) estimates the number of new cancer cases and deaths with a focus on the death rates in various cancer sites. The ACS also briefly reported death rates by several socioeconomic factors like ethnicity and educational attainment respectively. In this project, we plan to analyze how socioeconomic factors relate to cancer mortality and create a model to predict cancer mortality based on projections from observed data between 2010-2016. These data were collected from multiple sources, spanning over 3,000 counties nationwide.

Mortality data is compiled based on the cause of death reported on death certificate and census report from ACS. The factors are categorized into five overarching themes: cancer-related figures, economic data, demographic data, education and county-specific information. Most variables in the dataset are continuous in nature, representing a some proportion of the county.

Exploratory Data Analysis

Our data is an aggregation of multiple sources including the American Community Survey, the Census, clinical trials and from cancer databases. The observational unit is a state county, and each observation contains various demographic, economic and educational factors. Our outcome is **cancer mortality rate**, measured as number of deaths per capita (100K people).

The data contains information on 3047 counties in the United States. The data is not fully descriptive of all counties in the US because 94 are missing. We surmise these counties are missing due to lack of data for them.

The average mortality per capita is 178.66 across the dataset. Cancer mortality has a nice bell shape, but there are a few outliers present (Figure 1).

We explored how other variables in the dataset correlated with the cancer mortality in the counties. The data also includes information on the incidence of cancer in each county, so we checked how it related to mortality. As expected, higher incidence correlates highly with higher mortality, so this should be accounted for in a model. There was no apparent trend between cancer mortality per capita and either age or education level.

We discovered some trends in the socioeconomic and demographic data. We found a relationship between insurance type and cancer mortality. Counties where a higher proportion of the population has a private insurance had lower cancer mortality. This trend was reversed for counties with a high proportion of public insurance. Race also seemed to have a relationship with cancer mortality (Figure 2). Counties with a higher proportion of African-Americans experienced higher rates of mortality, and the reverse trend was observed in white Americans. No obvious trend was seen in Asian-Americans or races defined as “other”. In terms of

education, higher rates of college education in a county was negatively correlated with mortality, and this was the only negative correlation seen in any of the education groups (no high school, high school graduate, bachelors).

It is commonly believed that higher education unlocks access to greater ecomonic opportunities later in life. We believe that there could be an interesting interaction between education and race that can be discerned from our data.

Questions Of Interest

Our exploratory data analyses lead us to two specific research questions:

1. Is there a significant interaction between any demographic and economic factors that contributes to increased cancer mortality in a county?
2. Can we use any significant findings from (1) that can help predict future cancer mortality in the counties?

Knowing that any interaction between education and race could possibly be beneficial to improving a subseqeunt prediciton model.

Initial Analyses

For our inference model, we plan to look at race proportions and education level proportions as the main effects in our model and include their interactions, our coefficients of interest. We also plan to include important confounders such as county size, cancer incidence, age, insurance type and sex to help adjust for these factors between counties.

We needed to perform some transformations on the data since some covariates such as average death rate had values that skewed extremely low or high (Figure 3). For these covariates with exterme values, we log-transformed them to give them a more even distribution. To help guide model selection later,we calculated the p-value and percent variance explained for each covariate (Figure 4). The percent variance explained was highest for percent of residents over 25 with a bachelor's degree (`PctBachDeg25_Over`) at 0.223 along with a p-value of 1.41e-129.

As an initial prediction model, we fit all 31 covariates individually with cancer death rate as the response. The resulting adjusted R^2 was 0.7664. We took note of the p-value for each covariate in this joint model, some of which significantly differedd from the p-value obtained when modeled alone (Figure 4). This likely happens because a few covariates are correlated with each other and explain similar underlying trends in the cancer rate. To gain a better understanding of collinearity of the covariates, we also plotted a heatmap of the pairwise correlations (Figure 5). This confirmed our initial belief about the redundancy in the data, so we can trim out redundant covariates from our model.

One approach we explored was principal component analysis (PCA) on all 31 covariates. We found that the first principal component explains nearly all the variance in the dataset. In response to this, we plan to see how removing some major confounding variables might help the principal components.

Future Directions

For our inference model, we plan to use the Wald Test to test the joint hypothesis that all the interaction coefficients are zero. We may also use bootstrap to understand these coefficients better.

For our prediction model, we plan to explore model selection from multiple angles, such as PCA, regularization, or hand-picking non-correlated covariates that individually explain the response variable well. For the first two approaches, we will use cross-validation to pick an optimal hyperparameter (shrinkage factor or number of principal components respectively) to tune the linear model in preparation for cross-validation so we do

not overfit the data. Different error metrics such as absolute error or squared error may also be explored. We intend to use an 80-10-10 split of the data between training, cross-validation and testing.

Moving forward, we plan to distribute all analytic and writing work evenly. Christian will be responsible for the inference model, Jasen and Alex will work on prediction.

Figures

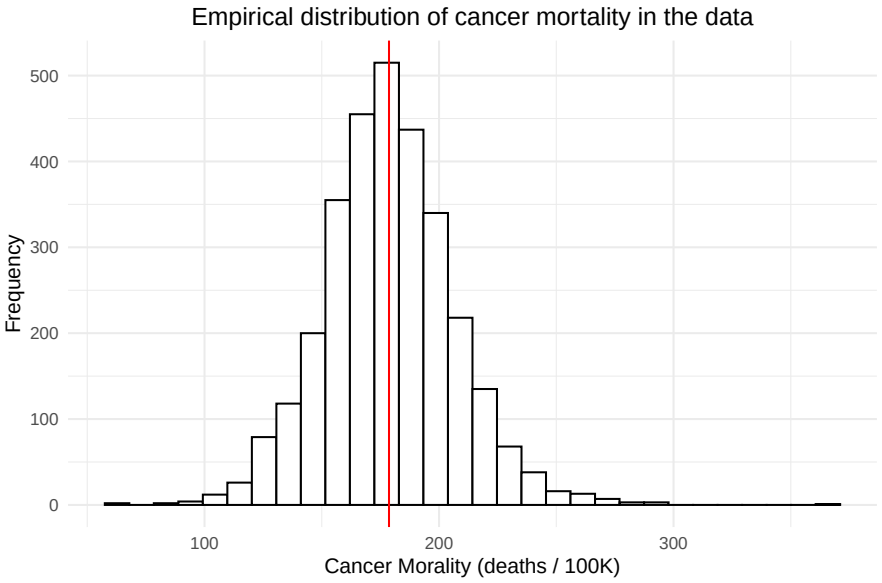

Figure 1

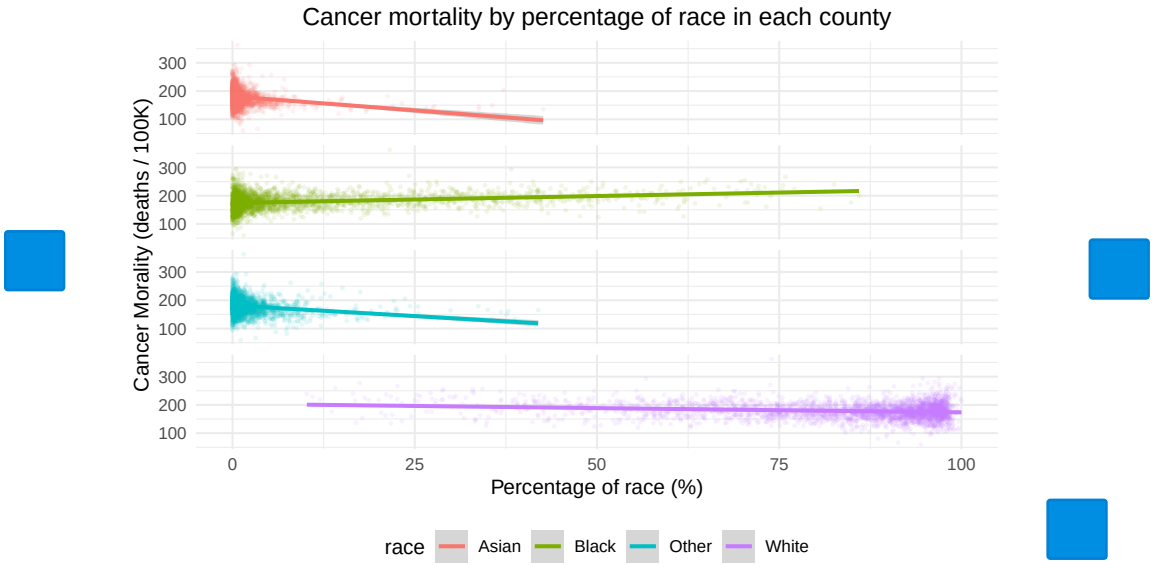

Figure 2

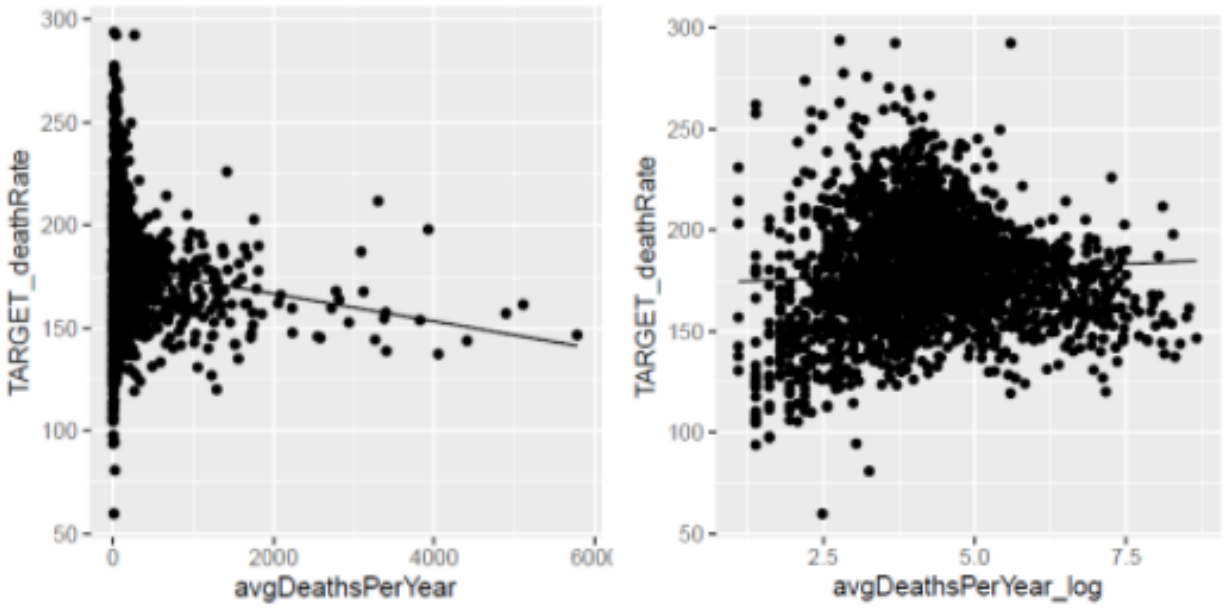

Figure 3: Log-transformation of deaths per year covariate to achieve better data spread.

	all_vars	combined_p_values	individual_var_explained	individual_p_values
1	PctBachDeg25_Over	7.025893e-02	2.288100e-01	1.412050e-129
2	PctPublicCoverageAlone	4.796630e-08	1.877945e-01	3.982831e-104
3	incidenceRate	2.191343e-33	1.774775e-01	6.329863e-98
4	povertyPercent	9.444326e-03	1.717786e-01	1.563116e-94
5	medIncome	1.894721e-10	1.700981e-01	1.548632e-93
6	PctEmployed16_Over	1.411380e-11	1.657457e-01	5.758225e-91
7	PctHS25_Over	6.868778e-01	1.589224e-01	5.799185e-87
8	PctPublicCoverage	1.271489e-16	1.521016e-01	5.408219e-83
9	PctPrivateCoverage	4.107398e-01	1.456395e-01	2.918247e-79
10	PctUnemployed16_Over	1.040344e-07	1.389930e-01	1.883181e-75
11	PctPrivateCoverageAlone	8.430301e-01	1.340869e-01	1.171070e-72
12	PctMarriedHouseholds	2.863466e-01	8.360056e-02	9.028143e-45
13	PercentMarried	1.400362e-01	6.901610e-02	5.479133e-37
14	PctBlack	4.585135e-01	6.786217e-02	2.237119e-36
15	PctEmpPrivCoverage	2.537896e-01	6.597241e-02	2.232411e-35
16	PctHS18_24	9.639247e-04	5.949843e-02	5.723899e-32
17	PctOtherRace_log	4.311569e-07	3.785192e-02	1.026490e-20
18	PctWhite	3.747625e-01	2.917737e-02	2.932701e-16
19	PctAsian_log	4.350925e-02	2.609729e-02	1.106420e-14
20	PctBachDeg18_24_log	2.951633e-01	2.401898e-02	1.278132e-13
21	PctSomeCol18_24	9.191668e-01	1.973504e-02	1.972995e-11
22	BirthRate	1.738155e-04	8.539812e-03	1.081705e-05
23	studyPerCap_log	4.609396e-04	8.504366e-03	1.128405e-05
24	PctNoHS18_24	7.846877e-01	5.475922e-03	4.302856e-04
25	avgAnnCount_log	7.189643e-12	5.224525e-03	5.843177e-04
26	avgDeathsPerYear_log	0.000000e+00	4.194613e-03	2.066787e-03
27	popEst2015_log	0.000000e+00	2.638423e-03	1.459996e-02
28	MedianAgeMale	3.336970e-05	1.165561e-03	1.046778e-01
29	AvgHouseholdSize	1.311899e-01	1.051249e-03	1.233355e-01
30	MedianAge	9.915725e-06	3.325723e-04	3.861893e-01
31	MedianAgeFemale	1.374841e-02	2.532389e-06	9.397294e-01

Figure 4: Metrics to evaluate predictive power for each covariate, including p-value and percent variance explained.

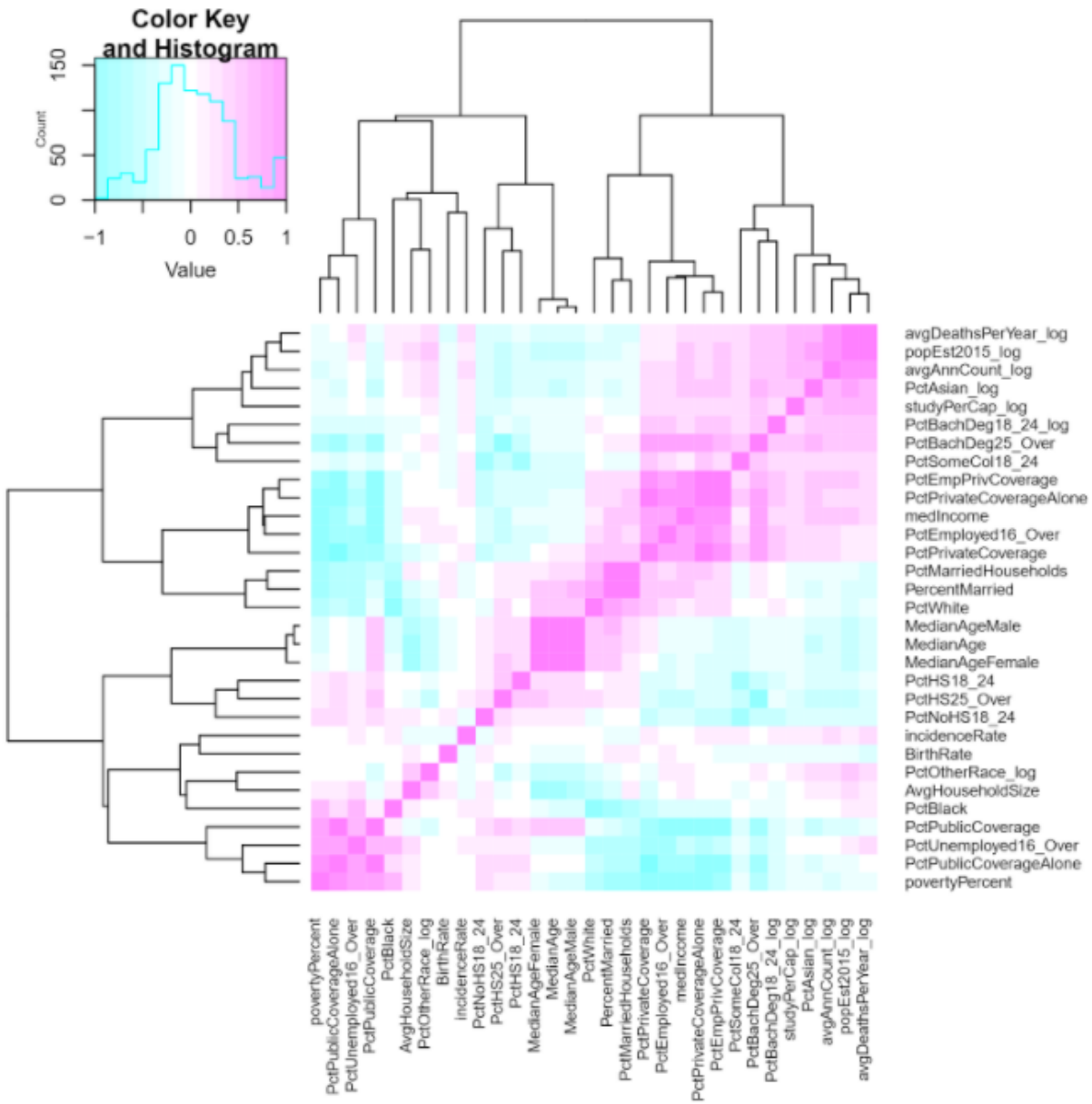

Figure 5: Heatmap of the pairwise correlation of all 31 covariates.