Predicting cancer mortality in the United States

Haoran Hu, Runqi Ma, Alyssa Vanderbeek, Zelos Zhu

17 December 2018

Abstract

Background Cancer is a leading cause of death in the United States [Wang]. Several factors have been linked to cancer risk on individual and ecological scales, including family history, lifestyle habits, median regional income, and others [NIH]. We seek to understand what population attributes can predict cancer mortality per capita (100,000 persons).

Data Data for this study is aggregated from United States Census Bureau database, ClinicalTrials.gov database, and National Cancer Institute (NIH) database. The final dataset, which consists of information for 3047 counties, includes demographic, cancer incidence, and cancer mortality data.

Methods We evaluate educational, economic, and other population characteristics (e.g. age) and their association with aggregate cancer outcomes on a state level and fit a linear model to fit and predict cancer mortality.

Results Our final model predicts cancer mortality rate based on cancer incidence, percent change in the number of new diagnoses from year to year, percent of the population in poverty, median age, percent of the population that is married, percent of the population with a college degree, percent of the population that is non-white, percent unemployment, and the percent of the population with health insurance.

Discussion Our final model is not the one with the lowest MSE, but it accounts for over 50% of the variability in the outcome and uses what we believe is the fewest number of significant and/or practically meaningful covariates. Additionally, limitations may exist in limiting prediction to linear models. Further exploration on this topic using alternative and non-linear methods may produce higher accuracy in prediction.

Introduction

Cancer is a leading cause of death in the United States [Wang]. Large amount of resources are poured into cancer research and treatment, in an effort to better understand the biological pathways, risk factors, and improve outcomes [NIH]. Several factors have been linked to cancer risk on individual and ecological scales, including family history, lifestyle habits, median regional income, and others [NIH]. In the last twenty years, cancer incidence has been shown to increase. However, it's unclear the extent to which this rise in incidence ought to be attributed to improved detection, poorer lifestyles compared to earlier populations, rampant exposure to risk factors, etc. [NIH]. And for many malignancies, prognosis has not much improved despite increased knowledge of biological mechanisms and technological/medical developments. For example, glioblastoma is one of the most aggressive cancers, and despite the large amount that is known about the biology, there are few effective therapies [2]. Here, we seek to understand what factors can predict cancer mortality per capita (100,000 persons) using linear regression methods.

Methods

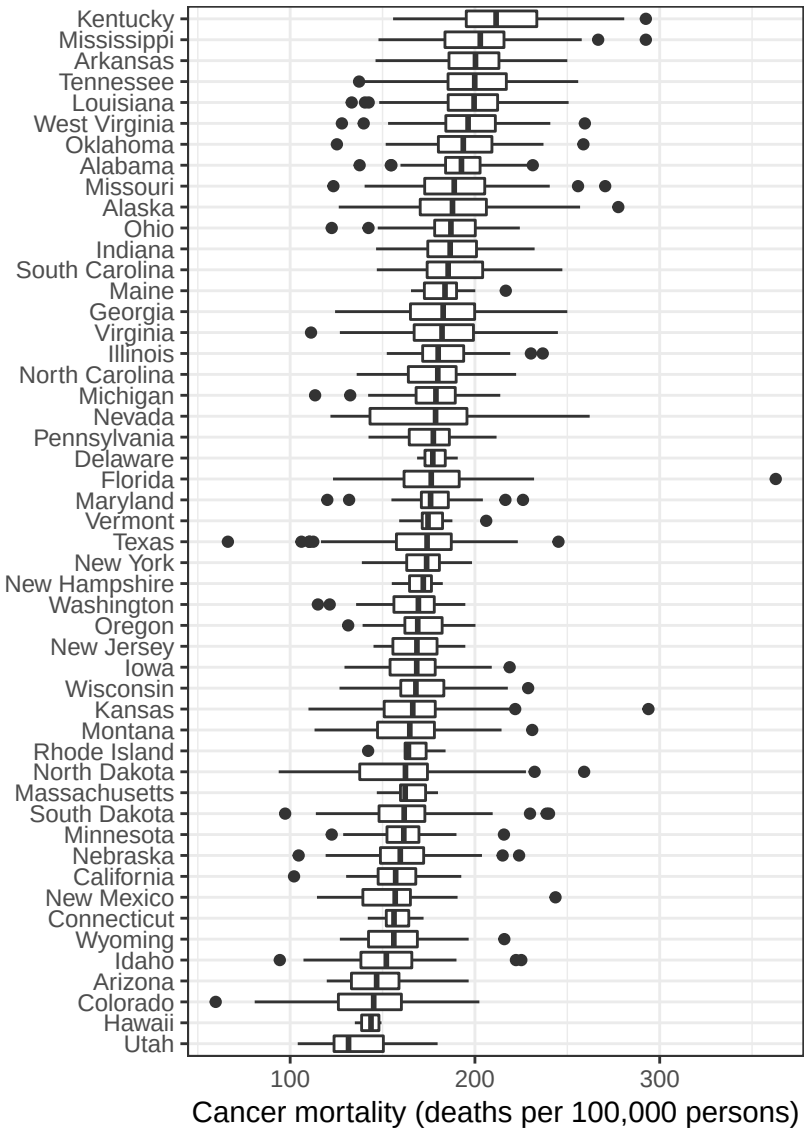
Data was aggregated across multiple sources - including the American Community Survey, US Census, ClinicalTrials.gov, and the National Institutes of Health (NIH).

Several groups of covariates were observed to be associated with each other, or had skewed distributions. The estimates for the employment rate, percent of people with private insurance alone, and the percent of people between ages and 18-24 who attended some college, and so we exluded them from consideration. This was inconsequential since they showed an association with other variables that did not have any missing data.

Among correlated covariates, we chose the single possible predictor that is most associated with the outcome. Namely, we recoded counties as their home state; we looked across age groups at the percent of people with college degrees; we transformed the number of new cases of cancer to a logarithmic scale; and we looked at the percent of the population that was non-white, as opposed to looking at the makeup of specific races (Table 1). Ultimately, we examined the significance of number of diagnosed cancer, cancer incidence rate, percentage of the population in poverty, per capita number of cancer-related clinical trials, median age, average number of people in a household, percent of the population that is married, education level, and health care coverage in predicting cancer mortality rate across states.

We fit a linear model using a combination of stepwise and Lasso methods. We use AIC criterion in stepwise model selection. In the Lasso model, a variety of λ values ranging from 10^{-2} to 10^{5}, were tested to determine the ideal lambda for the Lasso model. Once a final model was determined, 10-fold cross validation was used to assess predictive ability based on the root mean squared error (RMSE). Additional detail on data cleaning, model building, and cross validation can be found in the Appendix.

Measure	Min	1st Q	Mean	Median	3rd Q	Max	Std Dev
Avg household size	0.022	2.37	2.48	2.5	2.63	3.97	0.43
Cancer incidence	201.3	420.3	448.27	453.55	480.85	1206.9	54.56
Percent change in number of new cases	1.79	4.33	5.32	5.14	6.25	10.55	1.43
Median age among men	22.4	36.35	39.57	39.6	42.5	64.7	5.23
Pop. percentage with college degree	2.7	13.2	19.44	17.7	23.75	86.3	8.88
Pop. percentage with insurance	65.4	96.25	100.61	101.3	105.8	131.7	7.39
Pop. percentage non-white	0	4.55	16.35	9.94	22.7	89.8	16.38
Unemployment rate	0.4	5.5	7.85	7.6	9.7	29.4	3.45
Pop. percentage married	23.1	47.75	51.77	52.4	56.4	72.5	6.9
Pop. percentage in poverty	3.2	12.15	16.88	15.9	20.4	47.4	6.41
Avg number of clinical studies	0	0	155.4	0	83.65	9762.31	529.63
Cancer mortality rate (per 100,000)	59.7	161.2	178.66	178.1	195.2	362.8	27.75


Kentucky
Mississippi
Arkansas
Tennessee
Louisiana
West Virginia
Oklahoma
Alabama
Missouri
Alaska
Ohio
Indiana
South Carolina
Maine
Georgia
Virginia
Illinois
North Carolina
Michigan
Nevada
Pennsylvania
Delaware
Florida
Maryland
Vermont
Texas
New York
New Hampshire
Washington
Oregon
New Jersey
Iowa
Wisconsin
Kansas
Montana
Rhode Island
North Dakota
Massachusetts
South Dakota
Minnesota
Nebraska
California
New Mexico
Connecticut
Wyoming
Idaho
Arizona
Colorado
Hawaii
Utah
100
200
300
Cancer mortality (deaths per 100,000 persons)

Results

Our final model uses cancer incidence, percent change in the number of new diagnoses from year to year, percent of the population in poverty, median age, percent of the population that is married, percent of the population with a college degree, percent of the population that is non-white, percent unemployment, and the percent of the population with health insurance to predict the average cancer death rate per 100,000 across states. Chosen predictors and their effects are given in Table 2; estimates for individual states are not shown, but differences can be seen in Figure 1. There are positive associations with death rate in incidence, age, poverty, unemployment and proportion of minorities. Negative associations were seen with marriage rate, proportion of college graduates, and health insurance status.

Table 2: Predictors of cancer mortality (number of cancer deaths per 100,000 persons) across states

Variable	Estimate (SE)	p-value
(Intercept)	94.71 (11.317)	0.000
Cancer incidence	0.195 (0.007)	0.000
Pop. percentage in poverty	0.675 (0.109)	0.000
Median age among men	0.234 (0.095)	0.014
Pop. percentage married	-0.151 (0.096)	0.113
Pop. percentage with college degree	-0.75 (0.052)	0.000
Unemployment rate	0.685 (0.164)	0.000
Pop. percentage non-white	0.027 (0.038)	0.469
Pop. percentage with insurance	-0.056 (0.085)	0.506

We saw positive associations with death rate in incidence, age, poverty, unemployment and proportion of minorities; holding all else constant, each one unit increase of cancer incidence per capita per year, mean cancer mortality is expected to increase by 0.2; with each one percent increase in the percent of population in poverty, mean cancer mortality per capita is expected to increase by 0.67; with each one year increase in median age among men, mean cancer mortalities per capita is expected to increase by 0.23; with each one percent increase in the minority population, mean cancer mortality per capita is expected to increase by 0.06.

Negative associations were seen with marriage, proportion of college graduates, and health insurance status. Adjusting for all other covariates in model, with each one percent increase in the percentage of married people, mean cancer mortality per capita is expected to decrease by 0.17; with one percent increase in the percent of people who earned bachelor's degree, mean cancer mortality per capita is expected to decrease by 0.79; and with one percent increase health insurance coverage, mean cancer mortality per capita is expected to decrease by 0.09.

Percent married, percent of minority, and percent of health insurance coverage are not significant ($P > 0.05$). Expected cancer mortality rate varies by state, as expected. The adjusted R-squared of the model is 0.55, which is acceptable.

We checked model assumptions, and found that the distribution has long tails, suggesting that there are a substantial number of counties with extremely large or extremely small cancer mortality rates (Figure 2). Only a few of these outliers in Virginia and Nevada showed evidence of influencing the model, and upon inspection proved not to significantly alter the results or improve the diagnostics.

Cross validation of our model showed modest ability to predict cancer mortality, with an RMSE of ~18.5.

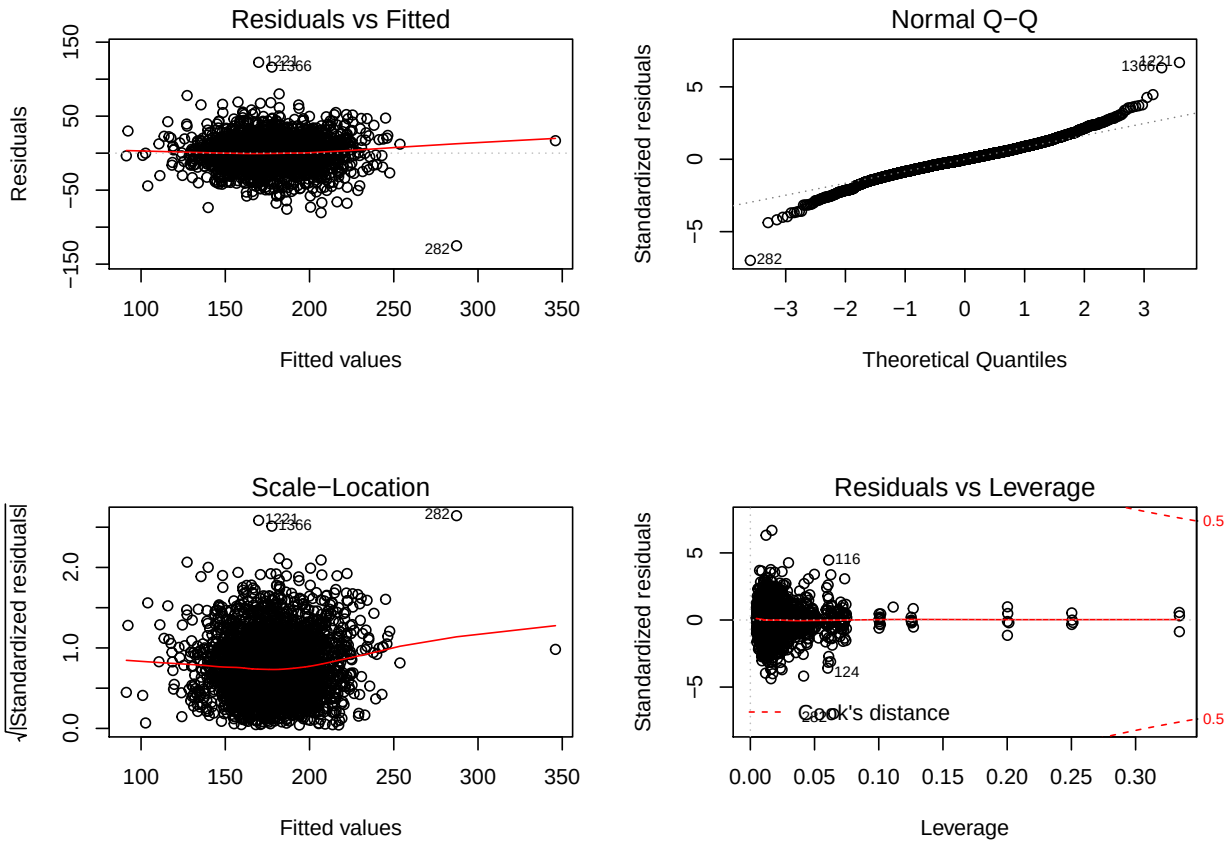


Figure 1: Model diagnostics

Discussion

The associations seen in our model results are not all too surprising considering the variables with the positive association are commonly deemed as disadvantages in society while those with the negative associations can be deemed as measures of success. It was interesting to note, however, that some known predictors of cancer outcomes were not statistically significant at the 5% significance level. For example, race (the proportion of the population that is non-white), is not significant here despite known associations with cancer outcomes [Katz]; the same can be said for health insurance coverage [Niu] and marital status [Aizer].

Our model does a modest job of predicting cancer mortality rates. Skewness in the data is a hurdle that we cannot seem to overcome using linear methods; non-linear alternatives or machine learning processes may better handle this issue and predict the outcome. That being said, our final model is not the one with the lowest MSE, but it accounts for over 50% of the variability in the outcome and uses what we believe is the fewest number of significant and/or practically meaningful covariates.

References

Abdelsattar ZM, Hendren S, Wong SL. The impact of health insurance on cancer care in disadvantaged communities. Cancer. 2016;123(7):1219-1227.

Aizer AA, Chen MH, McCarthy EP, et al. Marital status and survival in patients with cancer.J Clin Oncol.2013;31:3869-3876.

Alexander B, Cloughesy T. Adult glioblastoma. J Clin Oncol. 2017;35(21):2402–2409.

NIH. Age and cancer risk. https://www.cancer.gov/about-cancer/causes-prevention/risk/age.

NIH. Risk factors for cancer. https://www.cancer.gov/about-cancer/causes-prevention/risk.

NIH. Research funding. https://www.cancer.gov/grants-training/grants-funding/funding-opportunities.

Ellis, Canchola AJ, Spiegel D, Ladabaum U, Haile R, Gomez SL.Trends in Cancer Survival by Health Insurance Status in California From 1997 to 2014.JAMA Oncol. 2018 Mar 1;4(3):317-323.

Katz M, Parrish ME, Li E, et al. The Effect of Race/Ethnicity on the Age of Colon Cancer Diagnosis. J Health Dispar Res Pract. 2013;6(1):62-69.

Wang H. et al. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. GBD 2015 Mortality and Causes of Death Collaborators. Lancet. 2016 Oct 8; 388(10053):1459-1544.

Niu X, Roche L, Pawlish K, Henry A. Cancer survival disparities by health insurance status. Cancer Med. 2013 Jun; 2(3): 403-411.

Appendix

Data exploration and cleaning

We first examined variables for any skewness, and then tested associations between covariates to identify potential sources of collinearity. We found that several sets of variables were correlated. For those groups of correlated covariates, we generally chose to include the single covariate that was most highly associated with the outcome, cancer mortality rate.

We saw that median age was highly skewed, to the point that we think there is some sort of data error. Keep this in mind for later; for now we remove from the list of selected parameters. We opt instead to use the median age in males as a proxy, since the distribution (discarding the outliers in median age) is comparable to median age overall and median age in females. It's also interesting to note that median age between males and females are highly correlated eith each other, but not much correlated with target death rate. In fact, median age in males is negatively correlated with death rate, while median age for females is positively correlated.

We looked again at the variables for the percent of the population of a certain race and found that log transformation was bimodal. It might make sense to make this binary (i.e. >5% of a certain race or not). In the end, we chose to make a single variable measuring the percent of the population that is non-white (considered minority). This variable was skewed, but we did not perform a transformation for the sake of interpretability.

We combined several variables to create aggregate measures. We initially thought to group counties by region (Northeast, South, Midwest, West), but found that state-level groupings provided better fit and predictive capability. We designated the District of Columbia as being part of Maryland, since it had a single observation. We also created single measures for the percent of the population with a bachelor's degree and percent of the population with health insurance.

Median income and the percent of the population in poverty are correlated with each other as well as the percent of the population with a college degree. Percent in poverty had a slightly higher association with the death rate, and a slightly lower correlation with education (as measured), so we opted to include it over median income.

Model selection

We used a combination of stepwise selection (using AIC) and Lasso to choose our final model. When we tested our original selected variables under both processes, we find the same parameters are downweighted/removed.

We also tried fitting Lasso on the full original dataset (with our transformed/grouped variables) and found that many of the covariates appeared meaningful. However, this model seemed to overfit the data, especially since it chose to include some predictors with known correlation (e.g. median income and the percent of the population in poverty), suggesting presence of multicollinearity.

We performed stepwise regression using both AIC and BIC criterion, and found that using BIC resulted in several fewer variables being selected. But due to known associations between the eliminated variables and cancer mortality, we choose to include them in our model. For example, stepwise regression based on both criteria eliminated the percent of the population with health insurance as a predictor of cancer mortality per capita. However, we suspect this may be a practically meaningful covariate, since it has been shown that those with health insurance tend to have improved cancer outcomes over those without health insurance. This was done on an individual level [Ellis] and by relating individual insurance coverage with communities [Niu].

Our final model uses cancer incidence, percent change in the number of new diagnoses from year to year, percent of the population in poverty, median age, percent of the population that is married, percent of the population with a college degree, percent of the population that is non-white, percent unemployment, and

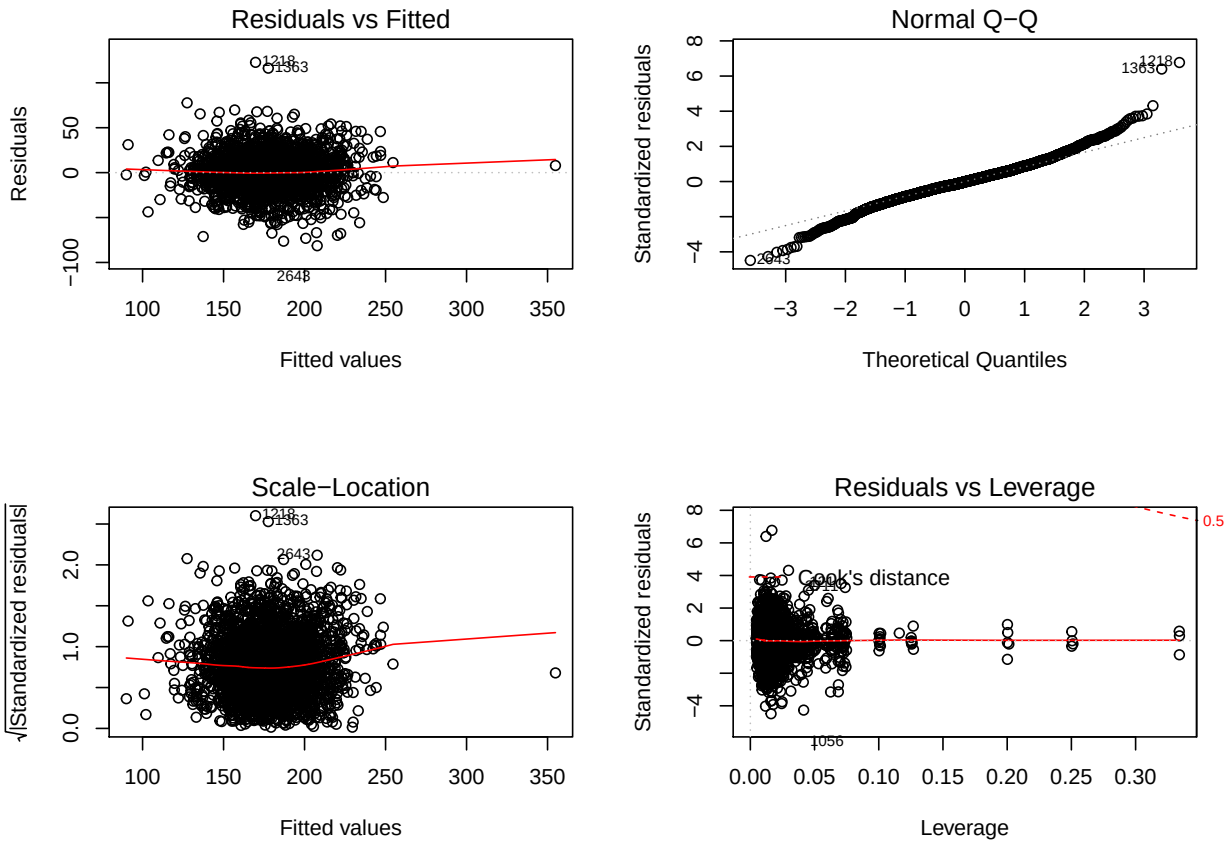


Figure 2: Diagnostics of model excluding potential influential points

the percent of the population with health insurance to predict the average cancer death rate per 100,000 across states.

Model diagnostics

We saw that there are long tails on both sides of the distribution (Figure 2). This may be due to outliers at either extreme, so next we examine presence of outliers and potentially influential points.

Skewness still exists without these influential points, but the model's R2 and Adj R2 both improve marginally (Figure 3). As a result, we opt to keep these observations in our model.

Model validation

Table 3: 10-fold Cross Validation

MSE	RMSE	SE
349.5709	18.69675	0.1162351

Table 4: Cross Validation across metrics

SSE	PRESS	RMSE	Adjusted Rsq
1025830	1068340	18.5257	0.5543684