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 [SOURCE]. 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. [SOURCE]. And for many malignancies, prognosis has not much improved despite increased knowledge of biological mechanisms and technological/medical developments [SOURCE]. 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. We cleaned the data to address these issues, and among correlated covariates, 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.

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

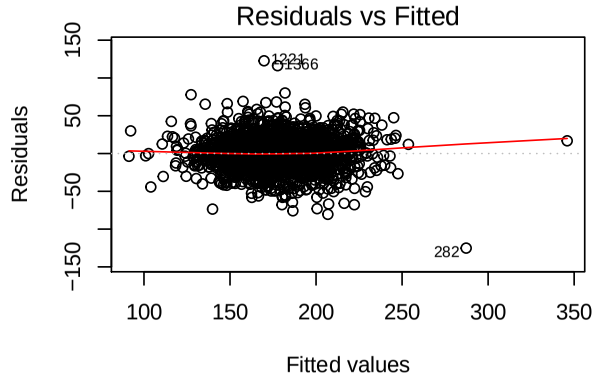
Results

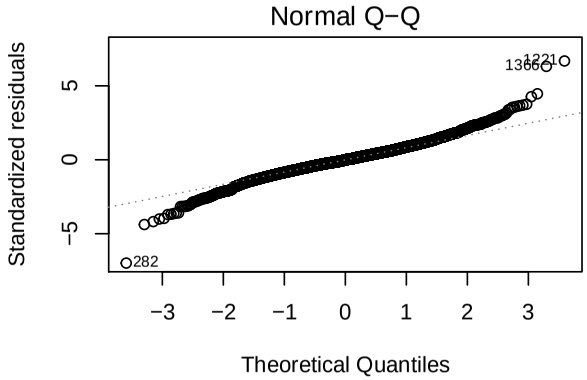
Our final model uses cancer incidence, percent change in the number of new diagnoses from year to year, percent of the population in poverty, median age, percent of the population that is married, percent of the population with a college degree, percent of the population that is non-white, percent unemployment,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

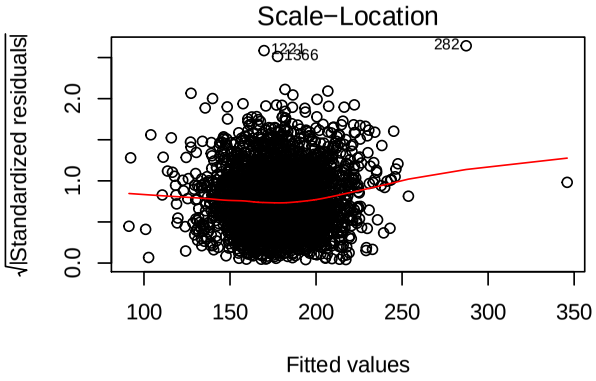
Table 2: Predictors of cancer mortality (number of cancer deaths per 100,000 persons) across states

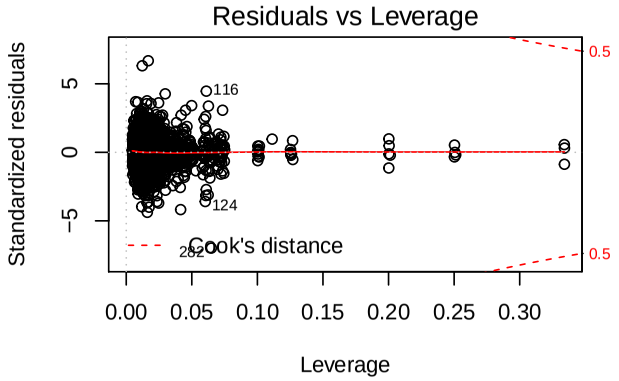
Variable	Estimate (SE)	p-value
(Intercept)	94.71 (11.317)	0.000
Cancer incidence	0.195 (0.007)	0.000
Pop. percentage in poverty	0.675 (0.109)	0.000
Median age among men	0.234 (0.095)	0.014
Pop. percentage married	-0.151 (0.096)	0.113
Pop. percentage with college degree	-0.75 (0.052)	0.000
Unemployment 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 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).


Residuals vs Fitted
Residuals
Fitted values
1221
1366
282
150
50
−50
−150
100
150
200
250
300
350


Normal Q−Q
Standardized residuals
Theoretical Quantiles
1221
1366
282
5
0
−5
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
Fitted values
1221
1366
282
2.0
1.0
0.0
100
150
200
250
300
350


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
116
124
282
0.5
0.5
5
0
−5
0.00
0.05
0.10
0.15
0.20
0.25
0.30

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 decent job of predicting cancer mortality rates, despite challenges. 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. 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.

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.

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.