Biostatistical Methods I Final Project

Authors: Eleanor Zhang, Bingyu Sun, Annie Yu, Justin Hsie

Abstract

There are many factors that contribute to cancer mortality rate. In this study, we fit a linear model to cancer data to determine which factors best predict cancer mortality rate. In our data cleaning and selection step, we narrowed down the total number of variables available in the dataset from 34 variables to 15. We used model building and testing techniques to fit a model to the data and find a model that best predicts cancer mortality rate. The model we selected includes information about incidence rate, income level, age, employment status, race, marriage status, education level, and insurance coverage.

Introduction

Cancer is a disease of many different types that affect the body in different areas with varying severity and outcomes. There is no singular cause of cancer, but rather a combination of many different risk factors [1]. It has been shown that gender can have an impact on cancer susceptibility, with hematologic and most other cancers being more common in men than women [2]. Data is constantly being collected to help us better understand the disease and the factors that contribute to its development. A healthy lifestyle is known to affect cancer risk, where a healthy diet, weight management, less alcohol consumption, regular exercising, and ceasing to smoke all decrease the risk for cancer [3]. These factors stemming from a healthy lifestyle are associated with the variables included in our dataset, such as poverty and education level, and have an impact on the development of cancer.

In this study, we are using data from multiple sources including the American Community Survey Census Bureau, NIH Clinical trials, and NIH National Cancer Institute. The data contains many

factors across various demographics that could potentially contribute to cancer. Our goal is to use this data to fit a model that will determine the factors that best predict cancer mortality rate.

Methods

Our dataset contains many factors that are potentially associated with cancer mortality rate. These include a wide range of data on income status, social standing, age, demographics, geography, marriage status, education obtained, insurance coverage, and race. After our literature search, we chose the variables we thought were most important to test to include in our model. These variables were mean per capita (100,000) cancer diagnoses, percent of population in poverty, median age of county residents, median age of males, median age of females, average household size, percent 16 and over who are unemployed, percent who identify as White, Black, Asian, and Other race respectively, percent of married households, percent of 18-24 year olds whose highest education is up to high school, percent with insurance coverage, and income category.

Given that the median income has a narrow interquartile range ($38882, $52492) with significant outliers on the top end, we stratified the median incomes per county into two groups following the guidelines from US bureau income statistics and another published paper[4]: Below $35,000 and above $35,000. We also combined those with private insurance coverage with those with public health insurance alone and put them into a variable indicating insurance coverage. For education level variables, we created a new variable for age group 18-24 of education level to high school by adding no high school degree and maximum education level high school. We will not use the variables for age group 25 and above in the original dataset because their education information is

not intact. Therefore, we end up with a pool of 15 variable candidates to predict cancer mortality rate. The distributions for each variable we selected are in the supplement (fig.1, fig2).

For both low income and high-income groups, we started with the linear regression with all potential variables. At the county level, we used automatic stepwise model selection and criterion-based model selection for each dataset and selected the predictors that would fit the data well. We used the partial F test and cross-validation to determine the fitness of the model and how our model performs in predicting cancer mortality rate. In each income group, two model selection procedures end up with two different models. The AIC and BIC value for models selected by automation procedure and criterion-based procedure do not deviate from each other significantly. Therefore, we chose the model with the least predictors out of the principle of parsimony.

We found outliers and influential points using Cook's distance, quantile-quantile plot, and residuals plots to determine whether there were any outliers that would significantly impact our model. The coefficients of low-income and high-income regression are not seriously affected by influential points at either extreme end. Therefore, we will keep those influential points in the dataset for regression.

In order to test the model predictive ability, we employed cross-validation and leave one out cross validation to compare the training MSE and testing MSE.

Results

For counties with median income less than $35,000, the variables in our final model include incidence rate, income, age and gender, race, and insurance. The final model for high-income counties contains incidence rate, income, female age, minority race excluding Asian, education, and employment. For both income classes, the incidence rate is positively correlated to cancer mortality rate, and income is negatively correlated to cancer mortality rate. For every 10,000 dollars increase in median income, the mean cancer mortality rate of the county is expected to decrease by 24 per capita (100,000) in the low-income group and decrease by 6 per capita (100,000) in the high-income group. For the low-income group, all racial groups and male age are negatively associated with the cancer mortality rate, whereas insurance coverage and female age exhibit a positive relationship with cancer mortality rate. For the high-income group, female age, education and other race are negatively correlated with the cancer mortality rate, while unemployment and black are positively associated with increased cancer mortality rate.

The predictive ability of our models was validated with 80/20 cross-validation (100 repeats) and LOOCV. Since our testing MSEs are close to training MSEs, both methods indicate our models have good predictive ability.

Conclusion/Discussion

For the low-income dataset, although all racial groups (White, Asian, Black, and others) are negatively correlated with the cancer mortality rate, Asian demonstrates the greatest protective effect against cancer death. In addition, the estimate for income is close to zero, suggesting income differences among counties classified as low income has a limited effect on cancer mortality rate. In contrast, insurance coverage is positively associated with the cancer mortality rate, revealing an

increased cancer death rate in places with higher insurance coverage. Our model also implies age for males has a protective effect on cancer mortality, whereas aging causes more cancer deaths in females. Unlike the low-income model, black people are at increased risk for cancer death and female age is protective in high-income group. This direction change of coefficients of female age between low income and high-income group implies that elder women are more susceptible to cancer mortality with low-income level due to lack of care and socioeconomic status. Furthermore, income, education, and other race are negatively correlated with the cancer mortality rate, while unemployment is positively associated with increased cancer mortality rate.

Considering that our model did not fit the data very well, in the future we should further explore the functional form of predictors and experiment with polynomial or even nonlinear regression models.

Supplementary

Fig.1 Low-income group variables distribution

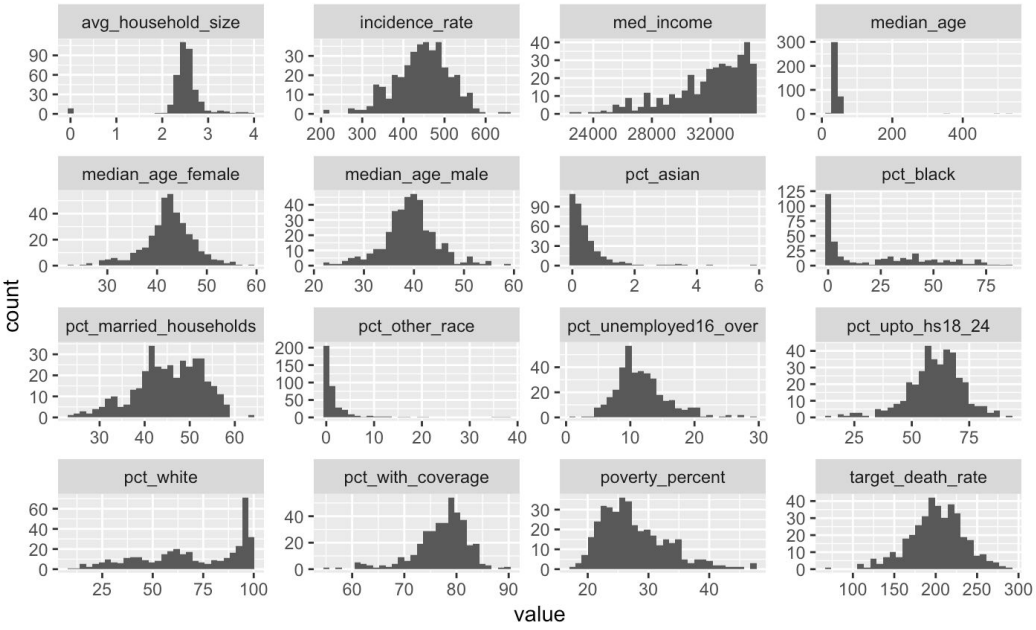

Fig.2 High-income group variables distribution

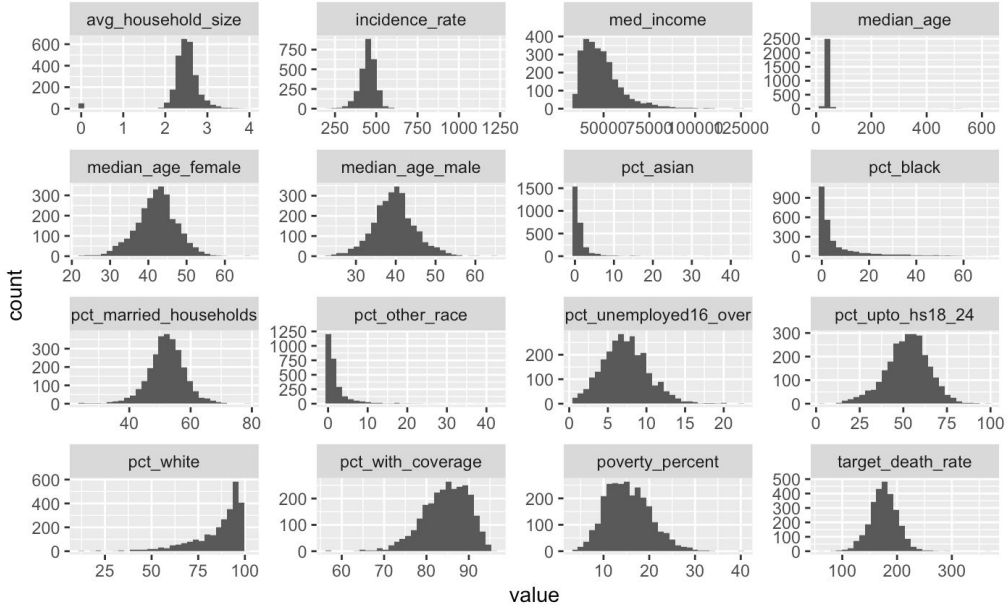

Table 1. Low-income group regression table

term	estimate	std.error	statistic	p.value
(Intercept)	93.5355565	24.7371687	3.781175	0.0001821
incidence_rate	0.2869252	0.0199783	14.361808	0.0000000
med_income	-0.0024180	0.0004974	-4.861421	0.0000017
median_age_male	-1.7544224	0.5841892	-3.003175	0.0028545
median_age_female	1.6709740	0.5763207	2.899382	0.0039633
pct_white	-0.2520327	0.1154077	-2.183847	0.0296044
pct_black	-0.4205350	0.1150302	-3.655864	0.0002936
pct_asian	-4.5604501	1.9050976	-2.393814	0.0171742
pct_other_race	-1.1299233	0.3164051	-3.571129	0.0004026
pct_with_coverage	1.1143671	0.2812032	3.962854	0.0000890

Table 2. High-income group regression table

term	estimate	std.error	statistic	p.value
(Intercept)	108.6499641	5.7261877	18.974223	0.0e+00
incidence_rate	0.1954631	0.0075315	25.952801	0.0e+00
med_income	-0.0005932	0.0000377	-15.726531	0.0e+00
median_age_female	-0.5124862	0.0815453	-6.284678	0.0e+00
pct_unemployed16_over	1.2027244	0.1489196	8.076336	0.0e+00
pct_black	0.1599914	0.0378132	4.231097	2.4e-05
pct_other_race	-0.9101158	0.1216118	-7.483777	0.0e+00
pct_upto_hs18_24	0.4188122	0.0335615	12.478941	0.0e+00

Table 3. Regression summary

dataset	model.mse	LOOCV.mse	CV.train.mse	CV.test.mse
low income	555.883	579.060	538.654	571.482
high income	385.128	387.517	384.248	384.179

References

[1] “Risk Factors.” National Cancer Institute, www.cancer.gov/about-cancer/causes-prevention/risk.

[2] Dorak, M Tevfik and Ebru Karpuzoglu. “Gender differences in cancer susceptibility: an inadequately addressed issue” Frontiers in genetics vol. 3 268. 28 Nov. 2012, doi:10.3389/fgene.2012.00268

[3] Khan, Naghma et al. “Lifestyle as risk factor for cancer: Evidence from human studies” Cancer letters vol. 293,2 (2010): 133-43.

[4] Choi SH, Terrell JE, et, al. Socioeconomic and Other Demographic Disparities Predicting Survival among Head and Neck Cancer Patients. PLoS One. 2016 Mar 1;11(3):e0149886. doi: 10.1371/journal.pone.0149886.