Predictive model for cancer mortality at county level: A multiple linear regression.

Members: Joy Hsu, Xinyi Lin, Matt Parker, Apoorva Srinivasan, Jiawei Ye

P8130: Biostatistical Methods I

Dec 17, 2018

I. Abstract

In order build a predictive model for cancer mortality rate, we used data from American Community Survey, census.gov, clinicaltrials.gov, and cancer.gov. We found that mortality rate is positively associated with cancer incidence rate, percentage of county residents aged 18-24 with the highest education being a high school diploma, percentage of government-provided health coverage. In addition, mortality rate is negatively associated with median age of male in the county, percentage of county residents aged 25 and over with the highest degree as bachelor's degree, percentage of minority residents (non-white, non-Asian, non-black), and percentage of married households. Based on our findings, improving higher education attainment and quality of health coverage may lower cancer mortality risk at the county level.

II. Introduction

In the United States, an estimated 38.4% of the population will be diagnosed with cancer during their lifetime ("Cancer Statistics", 2018). Cancer mortality has economic consequences at the population level, resulting in productivity loss and increased health expenditure. Although overall cancer death rates have declined from 1999 to 2015 ("Annual Report to the Nation 2018: National Cancer Statistics", 2018), some counties have experienced an increase (O'Connor, 2018). Research has shown socioeconomic gradients in cancer mortality rates. Several factors associated with cancer disparities at the county level include physical inactivity, smoking, obesity, unaffordable care, low-quality care, food insecurity, state smoke-free laws, and Medicaid payment rates (O'Connor, 2018). The aim of this investigation is to develop a regression model to predict cancer mortality rates at the county level, using a collection of socioeconomic, demographics, and population level parameters. Projecting cancer mortality at the geographic level and within demographic subgroups is key to resource planning and reduction efforts.

III. Materials and Methods

Data

The dataset was collated from registries including the American Community Survey, census.gov, clinicaltrials.gov, and cancer.gov. Data were available on 35 parameters for 3047 counties in the US. Parameters on cancer mortality rate (per 100,000) and cancer incidence rate (per 100,000) were calculated from 2010-2016 estimates. The remaining parameters on socioeconomic status, health coverage, education attainment, demographics were obtained from the 2013 Census.

Variable Selection

Parameters with high missing values and multicollinearity ($r > 0.7$) within a related group were excluded. To preserve sample size for regression modeling, three parameters with missing values were excluded: employed, aged 16 and over (5% NAs); college as highest education, aged 18-24 (75% NAs); and private health coverage (20% NAs). Variables removed due to multicollinearity were poverty percentage, binned income, percent married, aged ≥25 with high school as highest education. Of the health coverage variables, we retained "public health coverage alone". See **Figure 1. Variables used in Model Selection - Descriptive Statistics.**

Candidate Model Selection

Automatic search procedures, criterion based selection, and ridge regression were used for candidate model selection. Forward selection, backward selection, and stepwise regression procedures using the AIC criterion yielded identical 11-predictor candidate models (AIC=18159.34). Based on the criterion based procedures, we identified the most parsimonious subset that satisfied Mallow's Cp Criterion, a 9-predictor model (10 parameters, cp = 9.13). Optimizing for the adjusted R^2 criterion, we identified a 7-predictor model (adj. $R^2 = 0.493$). To evaluate model assumptions and detect outliers, we used Normal Q-Q plots to assess normality of residuals, residuals vs. fitted plots to assess homoscedasticity, and VIF scores to assess multicollinearity. All three models satisfied assumptions for normality of residuals, constant variance of residuals, and had VIF scores under 5.0. Ridge regression analysis on the 7 predictor model showed no substantial change in the coefficients.

Outliers & Leverage Values

Outliers in Y, leverage points and influential points were evaluated using studentized residuals, h_{ii} leverage values, and Cook's distance, respectively. On the 7 predictor model, we used studentized residuals to quantify the magnitude of residuals in standard deviation units. Sixty-eight counties had $|r_i| > 2.5$. Three counties were extreme outliers with $|r_i| > 5$, Williamsburg County, VA ($r_i = -5.31$); Madison County, MS ($r_i = 7.09$); and Woodson County, KS ($r_i = 5.54$). In our dataset, Williamsburg County has the second highest cancer incidence rate of 1014.2 per 100,000, which contributed to overprediction of cancer mortality by our model. In contradiction, NIH estimated incidence rate for Williamsburg County was substantially lower at 415.9 per 100,000 persons, for years 2011-2015 (NIH NCI, 2017). Since removal of this county did not alter any model coefficients by more than 2%, we opted to retain the observation. Madison County and Woodson County had the 4th and 2nd highest mortality rates, which suggests that our model under predicts for locations with high mortality rate. Of all observations, Union County, FL had the highest leverage value ($h_{ii} = 0.06$) while falling under the threshold for an outlier in X. Since Union County follows the trend of remaining data, inclusion of the observation improved R^2 (adj.) from 0.485 to 0.493. Further investigation of the Y outliers using Cook's distance measures confirmed that none were influential, all observations had D_i value under 0.5.

Cross Validation

To compare the predictive capability of the three models, N-Fold Cross Validation and K-Fold Cross Validation were performed. N-Fold Cross Validation was performed with 100 folds and an 80% training, 20% test split. K-fold Cross Validation was performed with both 5-folds and 10-folds. Across CV methods, RMSE values centered around 19, indicating that the 3 models have similar predictive capability. Based on principle of parsimony, we selected the 7 predictor model (adj. R^2: 0.49, p.value < 0.05). See **Figure 2. CV RMSE Values.**

*All analyses were performed using R Studio Version 1.1.456. A p <0.05 was considered significant.

IV. Results

Comparison of candidate models from automatic search procedures, Mallow Cp's criterion, and adjusted R-squared criterion on cross validation measures demonstrated equivalent predictive capacity. Based on parsimony, we selected a 7-predictor final model (adj. R-squared: 0.4927, p.value < 0.05) that includes incidence rate, median age male, PctBachDeg25_Over [1], PctHS18_24 [2], percent with only public health coverage, percent other race [3], and percent married households. Ridge regression analysis generated similar coefficients for the 7-predictor model. Our model satisfied assumptions for normality of residuals, constant variance of residuals, and had VIF scores under 5.0. While outliers in Y and high leverage values were identified, none were influential. See **Figure 3. Model Coefficients.**

[1] PctBachDeg25_Over - percent of county residents aged 25 and over with bachelor degree as the highest level of education attainment
[2] PctHS18_24 - percent of county residents aged 18-24 with high school diploma as the highest level of education attainment
[3] percent other race refers to non-white, non-asian, and non-black ethnicity

V. Discussion

Our predictive model shows that incidence rate, public health coverage, and percentage other race are associated with increases in cancer mortality at the county level, adjusting for other covariates. In contrast, median age male, PctBachDeg25_Over 1, PctHS18_24 2, and percent married households are associated with decreases in cancer mortality, adjusting for other covariates.

Interestingly, increases in public health coverage predicted higher cancer mortality rates. Further examination of our data revealed that public health coverage is correlated with increases in poverty levels (r = 0.80, p.value < 0.05) and inversely correlated with private health insurance coverage (r = -0.89, p.value < 0.05). Our findings suggest that disparity in health outcomes may exist between groups with private health insurance versus government-provided health coverage. Bittoni et al. found that inadequate health coverage in the U.S. is strongly associated with cancer mortality (Bittoni, Wexler et al. 2015). Furthermore, high school diploma as highest education is inversely correlated with bachelor degree attainment (r = -0.74); we suspect that higher education may be requisite for skilled jobs that improve socioeconomic status and accessibility to private health coverage. Consistent with our model's prediction that percentage of married households decrease risk for cancer mortality, Aizer et al. found that married cancer patients had better survival outcomes and were less likely to present with metastatic late-stage cancer (Aizer, Chen et al. 2013).

Several limitations limit the predictive accuracy and generalizability of our model. Since cancer mortality is aggregated across cancer types, we cannot assess population risk factors for specific cancers. Secondly, we cannot make individual level inferences since our data consisted of county-level parameters. Furthermore, our model coefficients may be biased due to data quality, our dataset contained parameters averaged from different time points and registries with different collection standards. Lastly, our investigation was limited to linear regression analyses. Future analyses should explore measures of association with nonlinear regression models and Cox Hazard Ratios, an appropriate method for survival analysis.

Prediction models for cancer mortality provide an important approach for assessing population-level risk and forecasting of healthcare resources. By understanding the geographic distribution of cancer burden, we can promote planning of clinical trial sites and research initiatives.

VI. References

Aizer, A. A., M. H. Chen, E. P. McCarthy, M. L. Mendu, S. Koo, T. J. Wilhite, P. L. Graham, T. K. Choueiri, K. E. Hoffman, N. E. Martin, J. C. Hu and P. L. Nguyen (2013). "Marital status and survival in patients with cancer." J Clin Oncol **31**(31): 3869-3876.

"Annual Report to the Nation 2018: National Cancer Statistics." National Cancer Institute, 2018, seer.cancer.gov/report_to_nation/statistics.html.

Bittoni, M. A., R. Wexler, C. K. Spees, S. K. Clinton and C. A. Taylor (2015). "Lack of private health insurance is associated with higher mortality from cancer and other chronic diseases, poor diet quality, and inflammatory biomarkers in the United States." Prev Med **81**: 420-426.

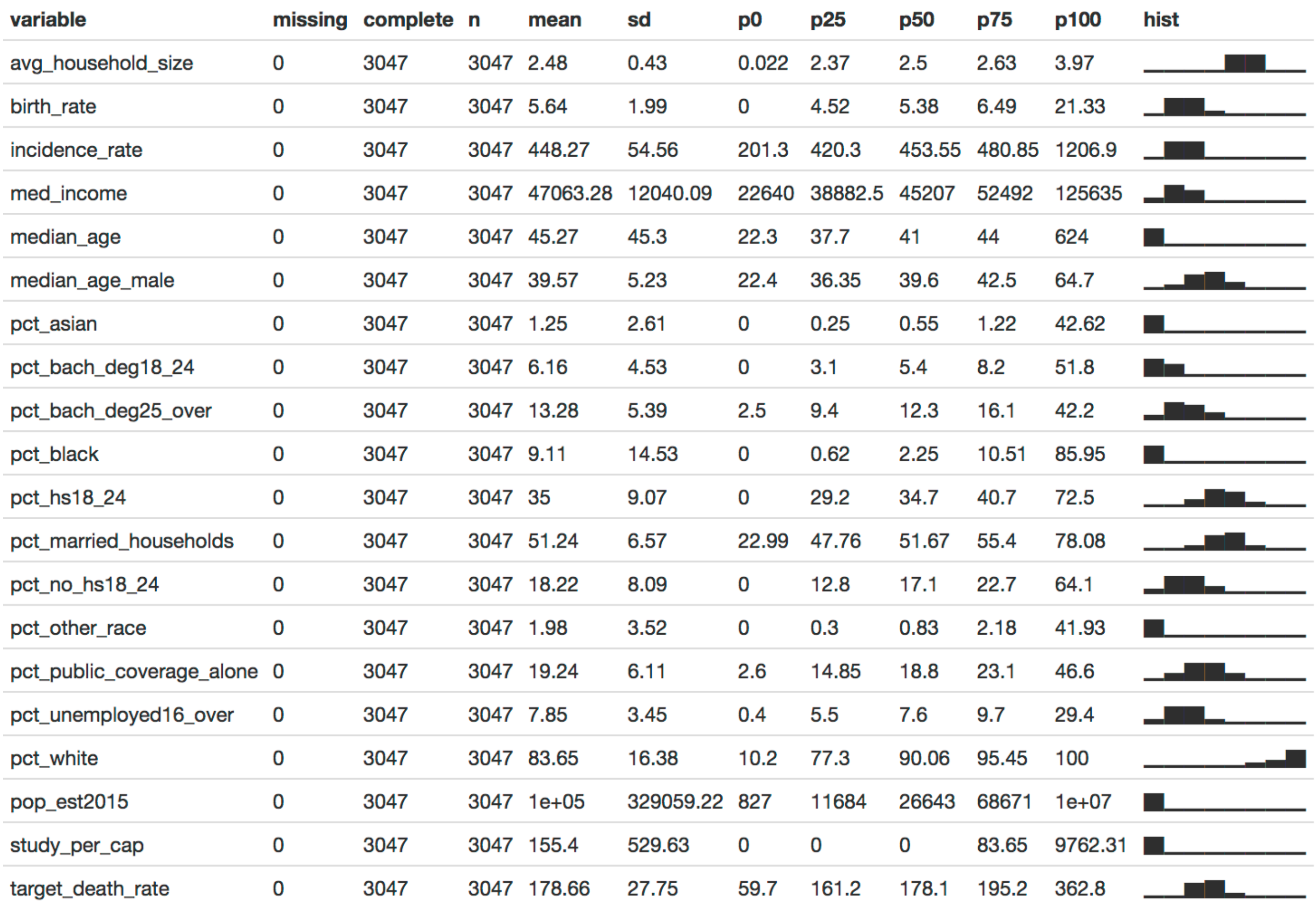
"Cancer Statistics." National Cancer Institute, 27 Apr. 2018, www.cancer.gov/about-cancer/understanding/statistics.

NIH. (2017). "State Cancer Profiles."   Retrieved Dec 15, 2018, 2018, from https://statecancerprofiles.cancer.gov/incidencerates/index.php?stateFIPS=51&cancer=001&race=00&sex=0&age=001&type=incd&sortVariableName=rate&sortOrder=default#results.

O'Connor, Jeremy M., et al. "Factors Associated With Cancer Disparities Among Low-, Medium-, and High-Income US Counties." JAMA Network Open, vol. 1, no. 6, 2018, doi:10.1001/jamanetworkopen.2018.3146.

VII. Supplementary Material

Figure 1. Variables used in Model Selection - Descriptive Statistics



variable	missing	complete	n	mean	sd	p0	p25	p50	p75	p100	hist
avg_household_size	0	3047	3047	2.48	0.43	0.022	2.37	2.5	2.63	3.97	
birth_rate	0	3047	3047	5.64	1.99	0	4.52	5.38	6.49	21.33	
incidence_rate	0	3047	3047	448.27	54.56	201.3	420.3	453.55	480.85	1206.9	
med_income	0	3047	3047	47063.28	12040.09	22640	38882.5	45207	52492	125635	
median_age	0	3047	3047	45.27	45.3	22.3	37.7	41	44	624	
median_age_male	0	3047	3047	39.57	5.23	22.4	36.35	39.6	42.5	64.7	
pct_asian	0	3047	3047	1.25	2.61	0	0.25	0.55	1.22	42.62	
pct_bach_deg18_24	0	3047	3047	6.16	4.53	0	3.1	5.4	8.2	51.8	
pct_bach_deg25_over	0	3047	3047	13.28	5.39	2.5	9.4	12.3	16.1	42.2	
pct_black	0	3047	3047	9.11	14.53	0	0.62	2.25	10.51	85.95	
pct_hs18_24	0	3047	3047	35	9.07	0	29.2	34.7	40.7	72.5	
pct_married_households	0	3047	3047	51.24	6.57	22.99	47.76	51.67	55.4	78.08	
pct_no_hs18_24	0	3047	3047	18.22	8.09	0	12.8	17.1	22.7	64.1	
pct_other_race	0	3047	3047	1.98	3.52	0	0.3	0.83	2.18	41.93	
pct_public_coverage_alone	0	3047	3047	19.24	6.11	2.6	14.85	18.8	23.1	46.6	
pct_unemployed16_over	0	3047	3047	7.85	3.45	0.4	5.5	7.6	9.7	29.4	
pct_white	0	3047	3047	83.65	16.38	10.2	77.3	90.06	95.45	100	
pop_est2015	0	3047	3047	1e+05	329059.22	827	11684	26643	68671	1e+07	
study_per_cap	0	3047	3047	155.4	529.63	0	0	0	83.65	9762.31	
target_death_rate	0	3047	3047	178.66	27.75	59.7	161.2	178.1	195.2	362.8	

1. target_death_rate – main response. Mean per capita (100,000) cancer mortalities
2. avg_household_size - Mean household size of county
3. birth_rate - Number of live births relative to number of women in county
4. incidence rate - Mean per capita (100,000) cancer diagnoses
5. med_income - Median income per county
6. median_age - Median age of county residents
7. median_age_male - Median age of male county residents
8. pct_asian - Percent of county residents who identify as Asian
9. pct_white - Percent of county residents who identify as White
10. pct_black - Percent of county residents who identify as Black
11. pct_asian - Percent of county residents who identify as Asian
12. pct_other - Percent of county residents who identify in a category which is not White, Black, or Asian
13. pct_no_hs18_24 - Percent of county residents ages 18-24 highest education attained: less than high school
14. pct_married_households - Percent of married households
15. pct_public_coverage_alone - Percent of county residents with government-provided health coverage alone
16. pct_unemployed16_over - Percent of county residents ages 16 and over unemployed
17. pop_est2015 - Population of county
18. study_per_cap - Per capita number of cancer-related clinical trials per county

Figure 2. Cross Validation RMSE

Table 2.1 K-Fold Cross Validation, 5-Fold

Model	RMSE	R^2	MAE
11-predictor	19.736	0.494	14.698
7-predictor	19.825	0.489	14.847
9-predictor	19.694	0.496	14.663

Table 2.1 K-Fold Cross Validation, 10-Fold

Model	RMSE	R^2	MAE
11-predictor	19.684	0.493	14.676
7-predictor	19.794	0.490	14.826
9-predictor	19.681	0.497	14.665

Table 2.2 N-Fold Cross Validation, 100-Fold

Model	RMSE
11-predictor	19.684
7-predictor	19.794
9-predictor	19.681

Figure 2.2.1 N-Fold Cross Validation, 100 Fold – Distribution of RMSE Values

*Model Auto – 11 predictor model, Model Cp – 9 predictor model, Model R^2 – 7 predictor model

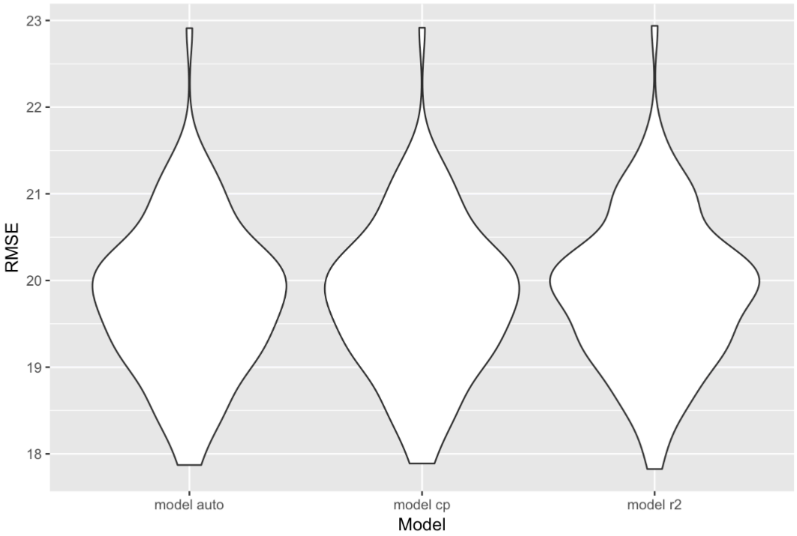
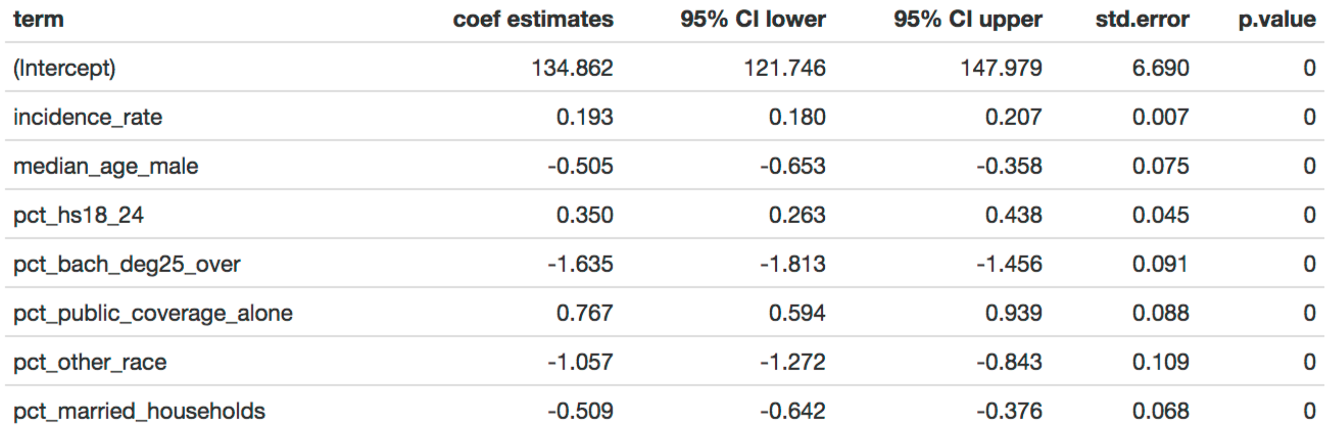


Figure 3. Final Model Coefficients



term	coef estimates	95% CI lower	95% CI upper	std.error	p.value
(Intercept)	134.862	121.746	147.979	6.690	0
incidence_rate	0.193	0.180	0.207	0.007	0
median_age_male	-0.505	-0.653	-0.358	0.075	0
pct_hs18_24	0.350	0.263	0.438	0.045	0
pct_bach_deg25_over	-1.635	-1.813	-1.456	0.091	0
pct_public_coverage_alone	0.767	0.594	0.939	0.088	0
pct_other_race	-1.057	-1.272	-0.843	0.109	0
pct_married_households	-0.509	-0.642	-0.376	0.068	0