A Model to Predict Cancer Mortalities of Each County in the United States

Jun Lu (jl5297), Qi Shao (qs2200), Kangkang Zhang (kz2334), Yizhi Ma (ym2715)

Abstract

Cancer is among the leading causes of death globally. There are many factors related to cancer mortality, including socioeconomic status, age, race and so on. In this paper, we aimed to build a multiple linear regression model to predict cancer mortalities of each county in the United States. In the final model, we chose six variables mainly related to education level, race, employment status, income or incidence rate as our predictors. As a result, our final model has a certain predictive ability.

Introduction

Cancer is one of the most important contributors to loss of life worldwide (Bray et al. 2018). In the United States, it is the second major cause of death. Approximately 38.4% of people will be diagnosed with cancer at some point during their lifetimes (National Cancer Institute., n.d.).

Cancer is a collection of diseases that amount of the body's cells divide abnormally and spread into other parts of the body (World Health Organization., n.d.; National Cancer Institute, n.d.). The causes of cancer are complex and vary between individuals. Epidemiology studies showed that cancer mortality trends differ depending on risk factors, including socioeconomic status, age, smoking status, family history and so on.

Age is a risk factor for many common types of cancer. The incidence of most cancers increases with age (White et al. 2014), like colon cancer (Shi et al. 2013) and lung cancer (Brown et al. 1996). Older adults present not only with physiological declines associated with aging but also with other impairments and social factors that might prevent them from undergoing cancer therapies (Marosi and Köller 2016).

Socioeconomic inequality exists in cancer incidence, mortality, and survival (Wells and Horm 1992; Clegg et al. 2009). Education, poverty, unemployment, health insurance all affect cancer diagnosis and access to treatment. Higher income, employment, and insurance coverage indicate better socioeconomic status, which is significantly associated with major cancers such as lung cancer, female breast, etc. (Clegg et al. 2009). Many behavioral factors associated with socioeconomic status may influence cancer risk including diet, physical activity, cigarette-smoking and so on (Danaei et al. 2005). Furthermore, factors correlated with socioeconomic status may affect cancer survival. People with high socioeconomic status tend to participate in cancer screening programs and receive timely treatments. Additionally, people with different socioeconomic statuses may choose different types of cancer treatment (Hussain et al. 2008; Liu et al. 2017).

Cancer mortality also varies between different race groups. A research shows that breast cancer mortality is higher in non-Hispanic black women than in non-Hispanic black women in every state in the United States (DeSantis et al. 2017). The race-based disparity results from a complex interaction of biologic and nonbiologic factors.

Although statistical trends could not be applied to patients directly, having an accurate estimate of cancer mortality is conducive to a rational allocation of resources and effective cancer control. In this paper, we used data from the National Cancer Institute and the United States Census Bureau to fit a multiple linear regression model to predict cancer mortalities of each county in the United States.

Method

Data description

The data for this project were aggregated from multiple sources including American Community Survey census.gov, clinicaltrials.gov, and cancer.gov. The final dataset contains data for mean per capita(100,000) cancer mortalities and related demographic information from 3047 counties. After the

previous literature review and missing data examination, we focused on the 11 predictors which are mainly related to education level, age, race, income, employment status, health coverage or cancer incidence rate (Table 1).

Descriptive analyses

Descriptive analyses were performed to summarize all variables of interest. The mean, mode, range, interquartile range and standard deviation of all variables of interest were calculated.

Model building

Multiple linear regression models were built to predict the mean per capita (100,000) cancer mortalities of each county. We used two approaches, including stepwise approaches and criterion-based procedures, to select important predictors. Adjusted R squared, Mallows's Cp criterion, Akaike information criterion (AIC) and Bayesian information criterion (BIC) were calculated to decide the best predictive model.

Model diagnostics

We used Cook's distance to detect influential points. Model assumptions were checked by diagnostic plots (residuals versus fitted values plot, QQ-plot, scale-location plot, and Cook's distance plot).

Model validation

Model predictive accuracy was evaluated by the leave-one-out cross-validation and the bootstrap method. In the leave-one-out cross-validation, the PRESS criterion was calculated and was compared with the sum square error to assess the predictive ability of the model. In the bootstrap method, the root mean square error was calculated (repeat 100, 1000, and 10000 times respectively).

Result

The density plot of the dependent variable (cancer mortality) is well bell-shaped, assuming that the variable follows a normal distribution (Fig. 1). Descriptive statistics of variables of interest are shown in Table 2 and Fig. 2. Most of the variables of interest have small Pearson correlation coefficient with each other (Fig. 3).

By stepwise approaches, we generated a model with 7 predictors, which is exactly one of the recommended models in the all-subset analysis. Based on criterions and parsimony, we choose the model with 6 predictors which has the smallest BIC, comparatively larger adjusted R-Squared and smaller subset size (Table 3). Our final regression model contains 6 variables mainly related to education level, race, employment status, income or incidence rate (Table 4). 48.14% of the dependent variable variation is explained by this multiple linear regression model (R-squared 48.14%; Adjusted R-squared 48.04%).

Fig. 5 shows 4 diagnostic plots that we used to check the model assumptions of the final model. In residuals vs fitted value plot, most points are randomly distributed around 0, indicating equal error variance across the entire range of fitted values. However, we observed 3 potential outliers. This also happens in the other 3 plots. After removing these points and refitting the regression model, the coefficient estimates and diagnosis plots remains mostly the same (Fig. 6), which means they are not influential, and we decided to keep these points.

In the leave-one-out cross-validation, the raw estimate of the prediction error is 402, very close to the mean square error of the full-sample model, 400, indicating that our model has a certain predictive ability. In addition, we ran the residual sample bootstrapping method 100, 1000 and 10000 times, respectively.

The estimated RMSE is almost the same as the running time increases, showing that the model has a stable predictive ability.

Discussion

In this study, we fit a multilinear regression model to predict cancer mortality using several selected variables. Based on the results, higher education plays the most important role against cancer mortality, compared to other variables. A county tends to have lower cancer mortality when there has a higher percentage of people with at least a bachelor's degree. Also, poverty and unemployment could raise the cancer death rate and races that are not black or Asian have relatively low cancer death rate.

The adjusted R-squared of our model is 0.48, which is not very ideal, and there are several limitations of our model. First, there exists other risk factors can also affect our regression results. Different studies have proved that for most types of cancer, the risk is higher among people with the family history of the disease (Pearce et al. 2013), which could result from exposure to similar lifestyle or environmental factors, and these risk factors could vary by different states. For example, lung cancer has the largest variation by states, reflecting regional differences in smoking prevalence and environment quality (American Cancer Society., n.d.).Further study could include dataset related to these risk factors.

Second, the census data that we use has several limitations itself. Census surveys are relatively long, and it will affect the accuracy of the answers. Also, it appears that people tend to over-report some items in a census survey, especially age and income (Parkin, Wardman, and Page 2008).

Table 1. Variable Descriptions and Sources

Variables of interest	Description
target_death_rate (a)	Dependent variable. Mean per capita (100,000) cancer mortalities
incidence_rate (a)	Mean per capita (100,000) cancer diagnoses
poverty_percent (b)	Percent of population in poverty
median_age_female (b)	Median age of female county residents
pct_bach_deg25_over (b)	Percent of county residents ages 25 and over highest
pct_unemployed16_over (b)	Percent of county residents ages 16 and over unemployed
pct_public_coverage (b)	Percent of county residents with government-provided health coverage
pct_emp_priv_coverage (b)	Percent of county residents with employee-provided private health coverage
pct_white (b)	Percent of county residents who identify as White
pct_black (b)	Percent of county residents who identify as Black
pct_asian (b)	Percent of county residents who identify as Asian
pct_other_race (b)	Percent of county residents who identify in a category which is not White, Black, or Asian (b)

(a) Years 2010-2016 (b) 2013 Census Estimates

Table 2. Descriptive Statistics of Variables of Interest

Variable	Mean	Std	Min	Q1	Median	Q3	Max
target_death_rate	178.7	27.8	59.7	161.2	178.1	195.2	362.8
incidence_rate	448.3	54.6	201.3	420.3	453.6	480.9	1206.9
poverty_percent	16.9	6.4	3.2	12.2	15.9	20.4	47.4
median_age_female	42.2	5.3	22.3	39.1	42.4	45.3	65.7
pct_bach_deg25_over	13.3	5.4	2.5	9.4	12.3	16.1	42.2
pct_public_coverage	36.3	7.8	11.2	30.9	36.3	41.6	65.1
pct_white	83.7	16.4	10.2	77.3	90.1	95.5	100.0
pct_black	9.1	14.5	0.0	0.6	2.3	10.5	86.0
pct_asian	1.3	2.6	0.0	0.3	0.6	1.2	42.6
pct_other_race	2.0	3.5	0.0	0.3	0.8	2.2	41.9
pct_unemployed16_over	7.9	3.5	0.4	5.5	7.6	9.7	29.4
pct_emp_priv_coverage	41.2	9.5	13.5	34.5	41.1	47.7	70.7

Table 3. Cp, Adjusted R squared, AIC and BIC Criterion for Models in Each Size

No. of Parameters	2	3	4	5	6	7	8	9	10	11	12
Cp	1454.8	362.5	109.9	52.9	28.8	18.9	16.1	13.1	10.3	10.4	12.0
Adjusted R^2	0.235	0.421	0.464	0.474	0.479	0.480	0.481	0.482	0.482	0.482	0.482
AIC	28085	27238	27003	26947	26923	26913	26911	26908	26905	26905	26907
BIC	28103	27262	27033	26983	26966	26962	26965	26968	26971	26977	26985

Table 4. Model Coefficients

term	estimate	std.error	statistic	95% CI	p.value	VIF
(Intercept)	104.00	5.34	19.40	(93.27, 114.20)	< 0.001	
incidence_rate	0.21	0.01	29.90	(0.19, 0.22)	< 0.001	1.07
poverty_percent	0.74	0.09	8.29	(0.57, 0.92)	< 0.001	2.49
pct_bach_deg25_over	-1.78	0.08	-21.10	(-1.94, -1.61)	< 0.001	1.56
pct_unemployed16_over	0.57	0.15	3.90	(0.28, 0.86)	< 0.001	1.94
pct_white	-0.10	0.03	-3.45	(-0.16, -0.04)	< 0.001	1.72
pct_other_race	-0.92	0.11	-8.41	(-1.13, -0.70)	< 0.001	1.12

Table 5. Residual sample bootstrapping results

RMSE	N = 100	N = 1000	N = 10000
Min	18.94	18.95	18.40
Q1	19.78	19.69	19.71
Median	20.06	19.97	19.97
Mean	20.04	19.96	19.97
Q3	20.28	20.22	20.23
Max	21.10	21.06	21.59

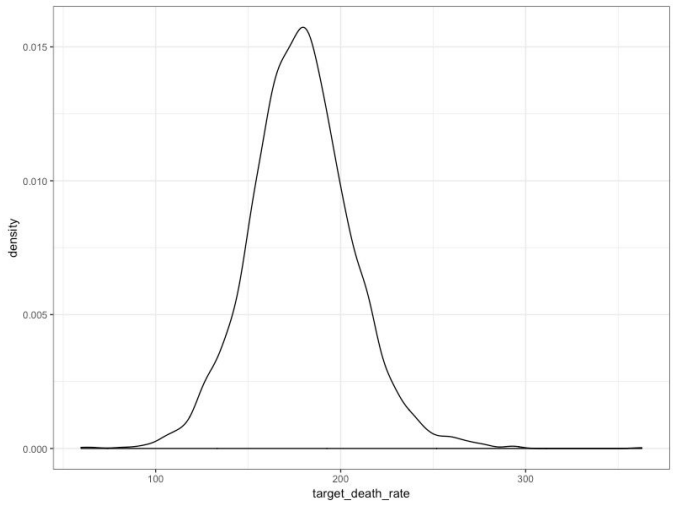

Figure 1. Density plot of target death rate

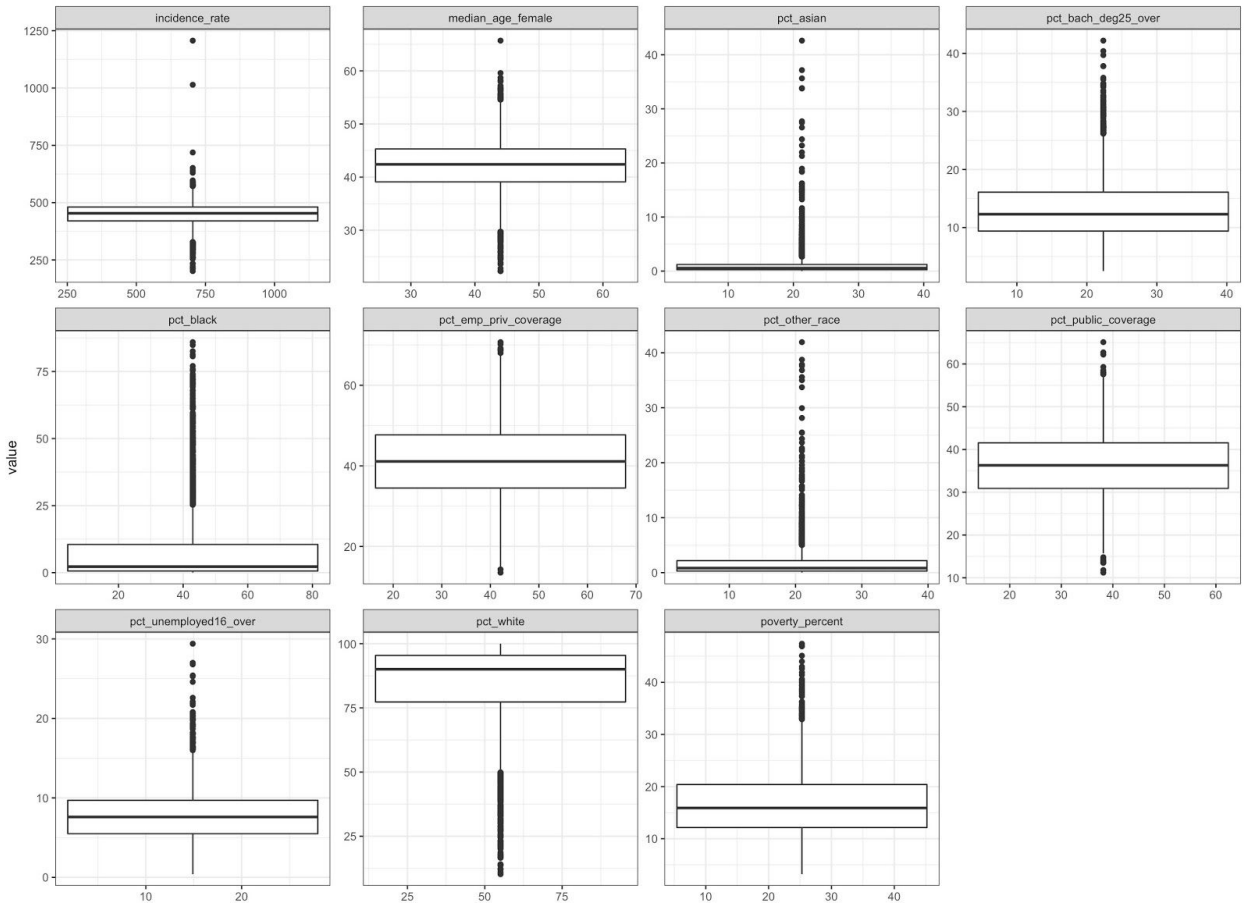

Figure 2. Box-plot of Variables of Interest

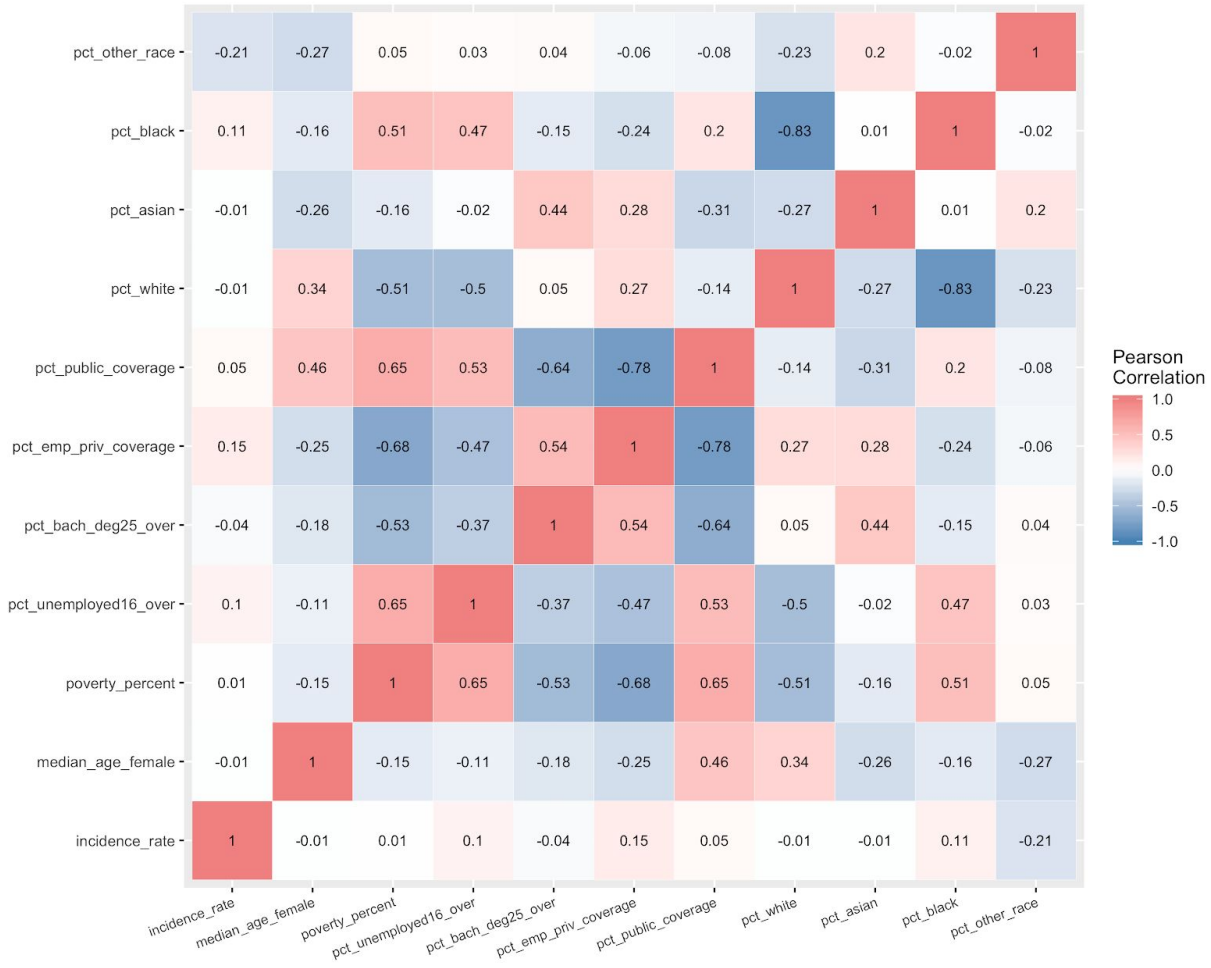

Figure 3. Correlation Heatmap of Variables of Interest

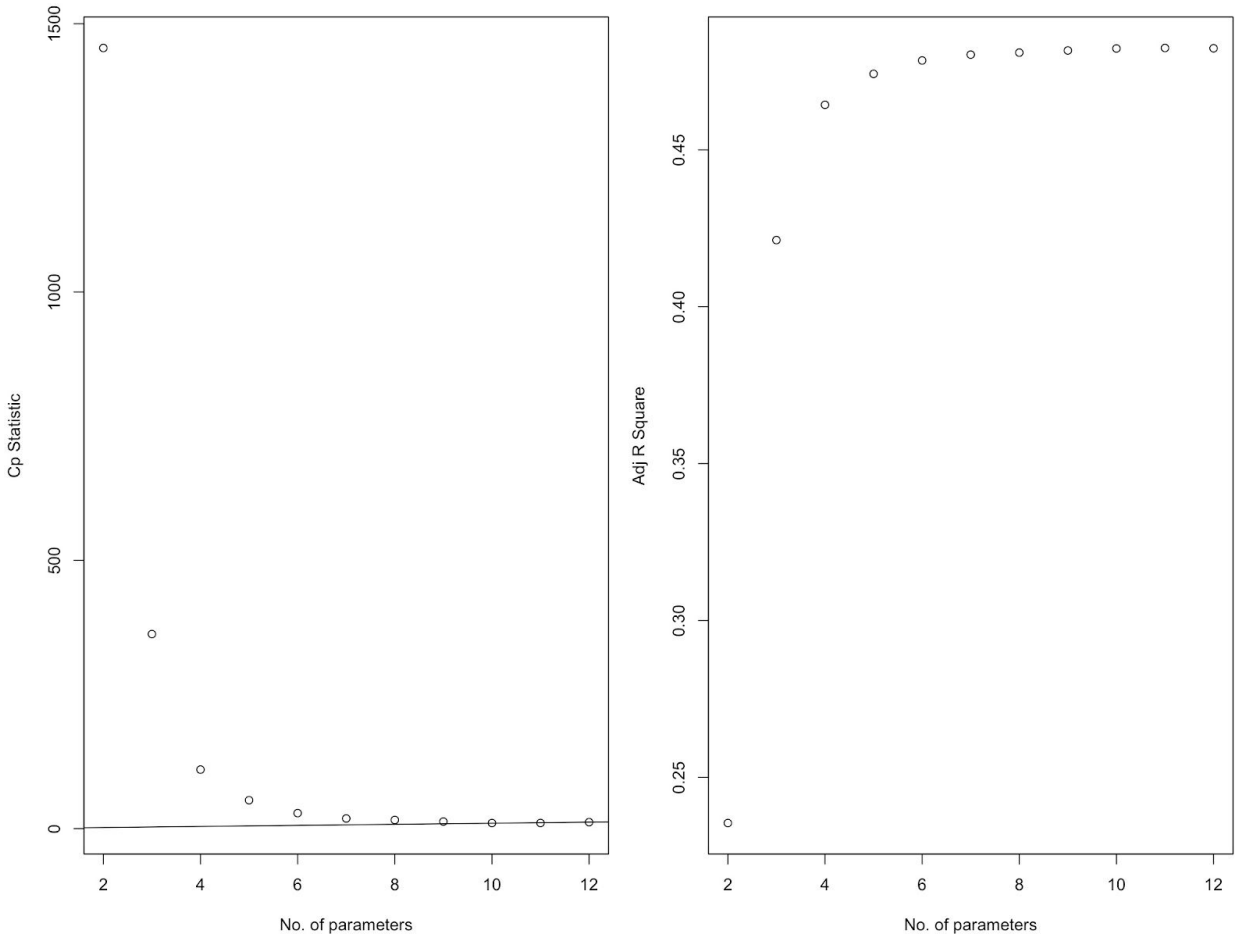

Figure 4. Cp, Adjusted R Squared VS Number of Parameters

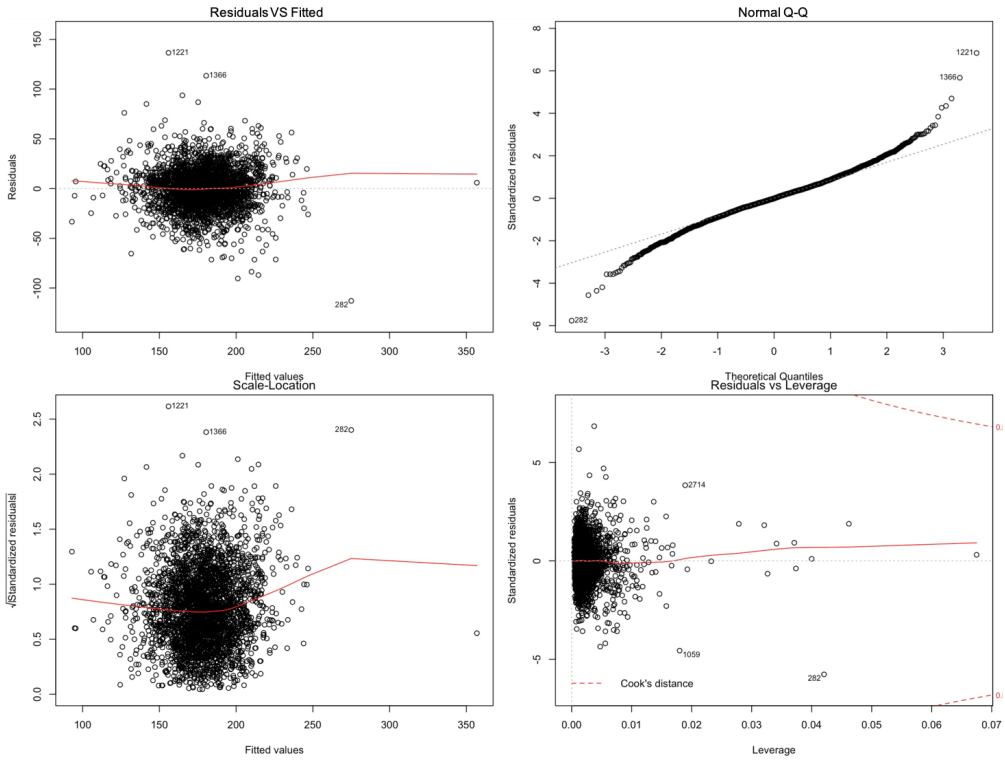

Figure 5. Diagnostic Plots of Regression Model

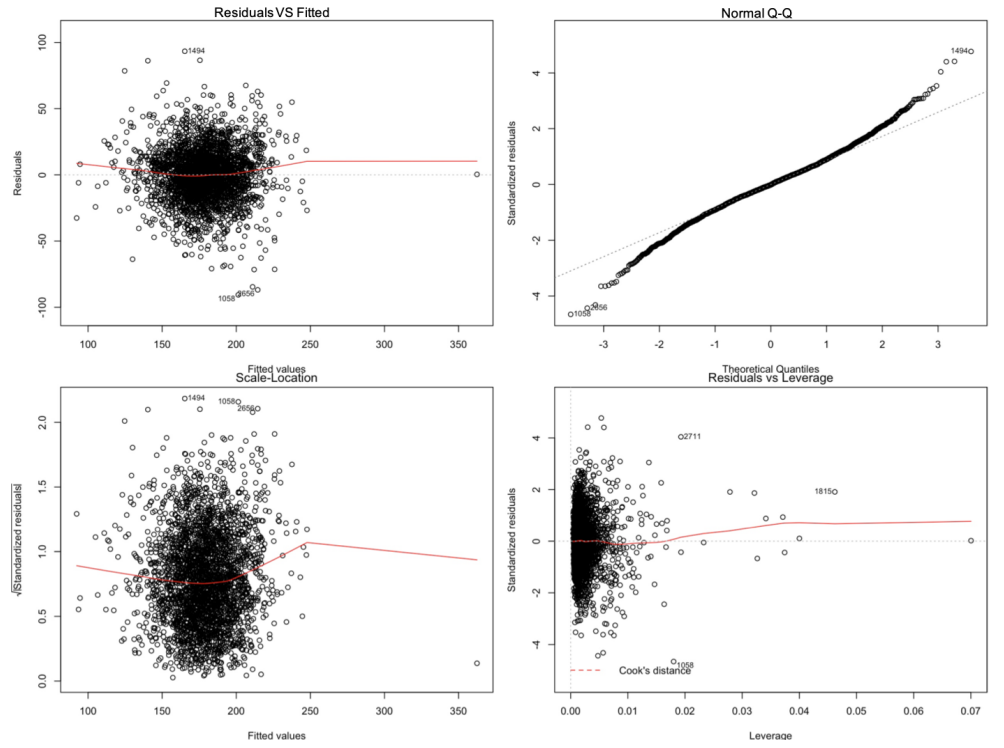

Figure 6. Diagnostic Plots of Regression Model (without potential outliers)

Reference

American Cancer Society. n.d. “Cancer Facts & Figures 2018. Atlanta: American Cancer Society; 2018.”

Bray, Freddie, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal. 2018. “Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.” *CA: A Cancer Journal for Clinicians* 68 (6): 394–424.

Brown, J. S., D. Eraut, C. Trask, and A. G. Davison. 1996. “Age and the Treatment of Lung Cancer.” *Thorax* 51 (6): 564–68.

Clegg, Limin X., Marsha E. Reichman, Barry A. Miller, Benjamin F. Hankey, Gopal K. Singh, Yi Dan Lin, Marc T. Goodman, et al. 2009. “Impact of Socioeconomic Status on Cancer Incidence and Stage at Diagnosis: Selected Findings from the Surveillance, Epidemiology, and End Results: National Longitudinal Mortality Study.” *Cancer Causes & Control: CCC* 20 (4): 417–35.

Danaei, Goodarz, Stephen Vander Hoorn, Alan D. Lopez, Christopher J. L. Murray, and Majid Ezzati. 2005. “Causes of Cancer in the World: Comparative Risk Assessment of Nine Behavioural and Environmental Risk Factors.” *The Lancet* 366 (9499): 1784–93.

DeSantis, Carol E., Jiemin Ma, Ann Goding Sauer, Lisa A. Newman, and Ahmedin Jemal. 2017. “Breast Cancer Statistics, 2017, Racial Disparity in Mortality by State.” *CA: A Cancer Journal for Clinicians* 67 (6): 439–48.

Hussain, Shehnaz K., Andrea Altieri, Jan Sundquist, and Kari Hemminki. 2008. “Influence of Education Level on Breast Cancer Risk and Survival in Sweden between 1990 and 2004.” *International Journal of Cancer. Journal International Du Cancer* 122 (1): 165–69.

Liu, Yang, Jian Zhang, Rong Huang, Wei-Liang Feng, Ya-Nan Kong, Feng Xu, Lin Zhao, et al. 2017. “Influence of Occupation and Education Level on Breast Cancer Stage at Diagnosis, and Treatment Options in China: A Nationwide, Multicenter 10-Year Epidemiological Study.” *Medicine* 96 (15): e6641.

Marosi, Christine, and Marcus Köller. 2016. “Challenge of Cancer in the Elderly.” *ESMO Open* 1 (3): e000020.

National Cancer Institute. n.d. “National Cancer Institute. (2018). *Cancer Statistics*. [online] Available at: Https://www.cancer.gov/about-Cancer/understanding/statistics [Accessed 17 Dec. 2018].”

National Cancer Institute. n.d. “National Cancer Institute. (2018). *What Is Cancer?*. [online] Available at: Https://www.cancer.gov/about-Cancer/understanding/what-Is-Cancer [Accessed 17 Dec. 2018].”

Parkin, John, Mark Wardman, and Matthew Page. 2008. “Estimation of the Determinants of Bicycle Mode Share for the Journey to Work Using Census Data.” *Transportation* 35 (1): 93–109.

Pearce, Celeste Leigh, Mary Anne Rossing, Alice W. Lee, Roberta B. Ness, Penelope M. Webb, for Australian Cancer Study (Ovarian Cancer), Australian Ovarian Cancer Study Group, et al. 2013. “Combined and Interactive Effects of Environmental and GWAS-Identified Risk Factors in Ovarian Cancer.” *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 22 (5): 880–90.

Shi, Runhua, Manga Devi Kodali, Stephani Chang Wang, Kalyana C. Lavu, Lihong Liu, Joseph Ryan Shows, and Glenn Morris Mills. 2013. “Mortality Risk Factors and Survival of Colon Cancer Patient.” *Journal of Clinical Orthodontics: JCO* 31 (15_suppl): e14653–e14653.

Wells, B. L., and J. W. Horm. 1992. “Stage at Diagnosis in Breast Cancer: Race and Socioeconomic Factors.” *American Journal of Public Health* 82 (10): 1383–85.

White, Mary C., Dawn M. Holman, Jennifer E. Boehm, Lucy A. Peipins, Melissa Grossman, and S. Jane

Henley. 2014. “Age and Cancer Risk: A Potentially Modifiable Relationship.” *American Journal of Preventive Medicine* 46 (3 Suppl 1): S7–15.
World Health Organization. n.d. “Who.int. (2018). *Cancer*. [online] Available at: Https://www.who.int/en/news-Room/fact-Sheets/detail/cancer [Accessed 17 Dec. 2018].”