STAT-530/430 Project

Factors Influencing Cancer Death Rates in US Counties

BHAVYA

April 20, 2024

1. Introduction

Life expectancy in the USA has seen an improvement over the past two decades, increasing from 76.1 years in 2000 to 78.9 years in 2019. This is evidently due to the advancements in medical technology and improvements in living conditions. In contrast, the death rate has decreased significantly from 856.9.2 per 100,000 population in 2000 to 715.2 per 100,000 population in 2019. Although the death rate has decreased, it is important to note that the majority of the deaths are due to medical conditions, especially different kinds of Cancers. According to the American Cancer Society, it is estimated that there were 606,000 cancer deaths in the USA in 2019. Most of these deaths are dependent on age, gender, and clinical history of an individual. More importantly, research has shown that survival rate in Cancer patients differs from medication used and quality of treatment an individual received. In this project, we will be analyzing how the Cancer death rate in each US county in 2012 is influenced by living conditions, income, poverty rate and insurance coverage of the population. The death rate being the response of this research, the predictors chosen are percentage of population aged between 18 to 25 who are graduated, percentage of employed individuals over 16 years old, incidence rate of cancer, median income of household in a county, percentage of unemployed population over 16 years old, percentage of population with no coverage or at least one coverage of insurance. According to recent studies, there is evidence that there is a correlation between insurance coverage and Cancer deaths. Studies have shown that uninsured individuals are more likely to get diagnosed with Cancer at a later stage, when it is more difficult and expensive to treat. Unemployment, Poverty rate of a county and average income of households are also the factors that have a strong influence on Cancer deaths.

2. Data

The Cancer death rate and socio economic statistics of this project was retrieved from an open source dataset named “Health Outcomes and Socioeconomic Factors” from Kaggle. The multiple sources for the given dataset are The American Community Survey, clinicaltrails.gov and cancer.gov. The dataset consists of different statistics on a county level, including state-fips (two-digit code that identifies the state), countyfips (a three-digit code that identifies the county), median household income, population estimate for 2015, poverty percent, study per capita, binned income, average annual count of cancer cases, average deaths per year, target death rate, and demographic data like median age of male and female population percent married households adults over 25 years old without a high school diploma, adults with a high school diploma, percentage of adults with some college education bachelor's degree holders, temporary private coverage, available temporary private coverage, available public coverage, available public coverage, available alone percentages, married household percentage, birth rate and percentage of white black asian races in a county.

The death rate mentioned in the data is per 100k individuals from each county in the USA. The overall count of counties in the US is 3047. Percentage of all the variable factors are considered against the population from each county. Depending on the coverage of an individual, the data gives a brief information on the individuals with public or private or temporary coverage.

3. Summary Statistics and Data Visualization :

Based on the 34 attributes from the dataset, selection of appropriate variables for the design of efficient models is a crucial step to deal with. As this report is more focused towards predicting the death-rate, eliminating the variable with no significant correlation with the response is necessary. As the death-rates caused by cancer are mainly impacted due to larger factors like age, gender and more importantly the medical condition of the individual, the indirect effect of financial status, living conditions and insurance coverage can have less correlation with the response. So, taking a significance level of less than 0.5 but greater than 0.3 can be an agreeable point from the debate of choosing higher correlation variables. Fig 3.1 shows a correlation grid between all the variables based on the defined significance level. From the grid, it is evident that the correlation is considerable between poverty percentage and other socioeconomic variables of the county.

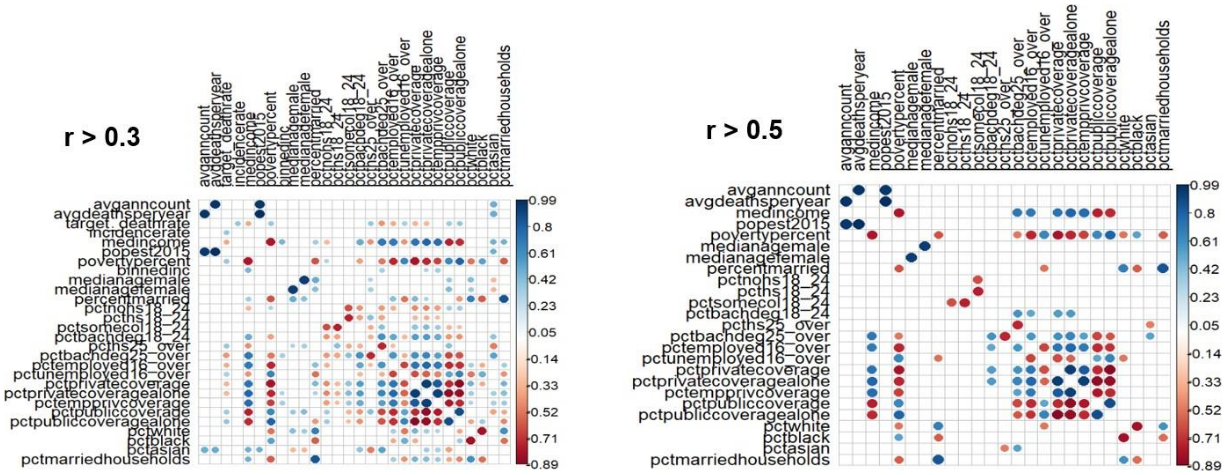


Fig 3.1: correlation grid of variable with defined correlation significance.

We collected the target deathrate , percentage of people aged 18-24 who graduated high school (pcths18_24) , percentage of people aged 16 and above who are employed(pctemployed16_over) , percentage of married people (percentmarried) , incidencerate , median household income (medincome) , percentage of people with private health insurance (pctprivatecoverage) , percentage of people with public health insurance (pctpubliccoverage) , percentage of people 16 and above who are unemployed (pctunemployed16_over) , percentage of people in poverty (povertypercent) , no.of clinical trials per capita in given county (studypercap) , percentage of people with private health insurance only (pctprivatecoveragealone) , percentage of people with both private health insurance and public health insurance (pctnocoverage) for 3047 counties across different states of U.S . Table 3.1 shows the statistical summaries of the response and the predictor variables .

Table 3.1 Statistical Summaries

	target_deathrate	pcths18_24	pctemployed16_over	percentmarried	incidencerate	medincome	pctprivatecoverage	pctpubliccoverage	pctunemployed16_over	povertypercent	studypercap	pctprivatecoveragealone	pctnocoverage
Min	59.7	0.00	17.60	23.10	201.3	22640	22.30	11.20	0.700	3.70	0.00	15.70	-13.55

1st Qu	161.9	29.07	48.77	47.67	420.6	38483	56.90	30.80	5.700	12.28	0.00	40.88	11.20
Median	179.7	34.70	54.55	52.30	453.5	44768	65.15	36.50	7.600	16.20	0.00	48.60	14.85
Mean	179.1	35.03	54.17	51.63	447.8	46868	64.28	36.27	7.824	16.92	158.44	48.35	15.29
3rd Qu	195.3	40.90	60.10	56.10	479.1	52202	72.00	41.80	9.600	20.60	78.41	55.70	18.60
Max	293.9	72.10	80.10	72.50	1014.2	122641	88.90	62.70	27.00	47.40	9762.31	78.90	43.10
S.D.	27.140	9.237	8.022	6.749	52.496	11939.244	10.512	7.869	3.344	6.189	565.276	10.087	5.821

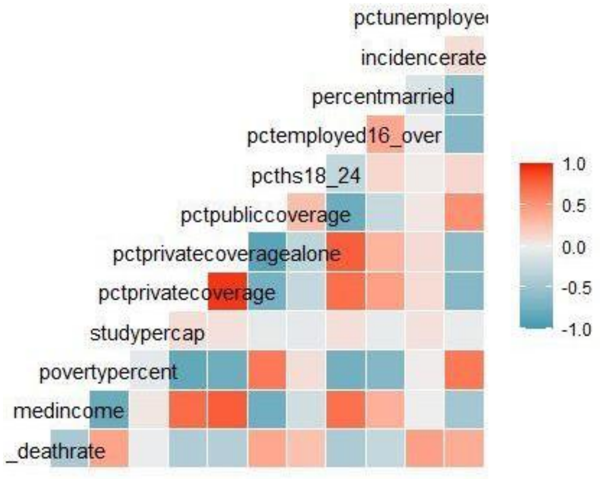
For the taken predictor and response variables a correlation matrix is drawn as shown in the graph 3.1 . We can interpret from the below diagram that poverty percent, incidence rate have good correlation with the target_deathrate and are good predictors while medincome , pctemployed16_over have less correlation and are weak predictors of the target_deathrate . pctemployed16_over , pctprivatecoverage , medincome have moderate negative correlation and pctpubliccoverage , incidencerate , pctunemployed16_over have moderate positive correlation . The predictor variables among themselves are also correlated moderately positive as well as negative .

Considering the fact that insurance coverage variable i.e pctprivatecoverage, pctpubliccoverage are correlated with eachother, including these variables into the model can cause multicollinearity between the variables which can affect the output of the model. To reduce its effect yet to perform better research around this question, taking a variable which is obtained by combining the both correlated variables can produce a better output. Using set theory, the percentage of individuals with no public and private coverages and percentage of individuals with at least one of the coverage is obtained.

$$pctnocoverage = 100 - (pctprivatecoverage + pctpubliccoverage)$$

$$pctunioncoverage = pctprivatecoverage + pctpubliccoverage - pctnocoverage$$

Fig 3.1 Correlation Matrix of the variables



4. Methodology

4.1 Model Selection

From the data analysis, the variables which are more suitable for predicting the response are selected based on the linearity with response and normality of residuals when the variables are used as regressors. The variables presented in table 3.1 are satisfying the above two conditions. Analysis of variance is another step that says the significance of the predictor in the model. Below table 4.1 gives the Anova test values for all the variable. Here, the models are divided into two parts which contains “pctnocoverage” in one model and “pctunioncoverage” in other.

Table 4.1 Analysis of Variance Including pctbothcoverage

target_deathrate ~ pcths18_24 + log(pctemployed16_over) + percentmarried + incidencerate + log(medincome) + pctunemployed16_over + log(povertypercent) + studypercap + pctnocoverage

	Df	Sum of Sq	RSS
none			1222805
pcths18_24	1	148990	1371795
log(pctemployed16_over)	1	249299	1472104
percentmarried	1	53902	1276707

incidencerate	1	372900	1595705
log(medincome)	1	115011	1337816
pctunemployed16_over	1	10664	1233469
log(povertypercent)	1	1474	1224279
studypercap	1	989	1223794
pctnocoverage	1	6985	1229790

In the above hypothesis test using Anova, we used pctnocoverage. As seen the RSS value for the model is 1222805, yet the significance of the variables percentmarried, povertypercent, studycap are not <0.05 p-values. Below is the linear regression model for table 4.1.

Table 4.2 Regression Summary Including pctnocoverage

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	0.0488	0.5488	8.892	< 2e-16
pcths18_24	0.483	0.0464	10.415	< 2e-16
log(pctemployed16_over)	-0.1183	3.984	-2.97	0.003
percentmarried	-0.0303	0.083	-0.365	0.7151
incidencerate	0.2292	0.0077	29.595	< 2e -16
log(medincome)	-0.3586	4.134	-8.673	< 2e -16
pctunemployed16_over	0.638	0.1624	3.928	8.77E-05
log(povertypercent)	-0.1249	3.233	-0.039	0.9692
studypercap	-0.0008	0.0007	-1.217	0.2238
pctnocoverage	0.2765	0.0681	4.06	5.05E-05

Residual Standard Error: 20.59 on 2885 degrees of freedom

Multiple R-Squared: 0.4399	Adjusted R-Squared: 0.4381

F-statistic: 251.7 on 9 and 2885 DF	p-value: 0

Table 4.3 Regression Summary Including pctnocoverage

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	486.580	21.255	22.892	< 2e -16
pcths18_24	0.479	0.043	10.415	< 2e -16
log(pctemployed16_over)	-12.334	3.83	-3.212	0.0013
incidencerate	0.2291	0.0075	30.187	< 2e -16
log(medincome)	-35.7227	2.351	-15.191	< 2e -16
pctunemployed16_over	0.64	0.1534	4.2323	2.39e-05
pctnocoverage	0.287150	0.0581	4.873	1.16e-06

Residual Standard Error: 20.58 on 2885 degrees of freedom

Multiple R-Squared: 0.4396	Adjusted R-Squared: 0.4384
F-statistic: 377.5 on 6 and 2888 DF	p-value: <2.2e-16

From above two linear regression models, it is observed that the p-values are higher than 0.05 for povertypercent, studycap, percentmarried. Using these variables brings a less efficient model, so removing the variable from the model, we get a model with table 4.3. As p-values are less than 0.05 for all the variables, this model can predict the death-rate with higher precision if these variables are removed.

As performed with pctnocoverage, we replaced it with pctunioncoverage and performed the Anova test and linear regression, the observations are similar in case of p-values and RSS values. But the intercepts in the two models are changed as expected. If all the variables are held constant in a model with pctunioncoverage, a unit increase in percentage of individuals with atleast one coverage, the deathrate is decreasing by a factor of 0.14 times in the county. If the percentage of population with no coverage increased by 1%, then the death rate seems to increase by 0.28 factor.

Table 4.4 Variance Inflation Factors Including pctnocoverage

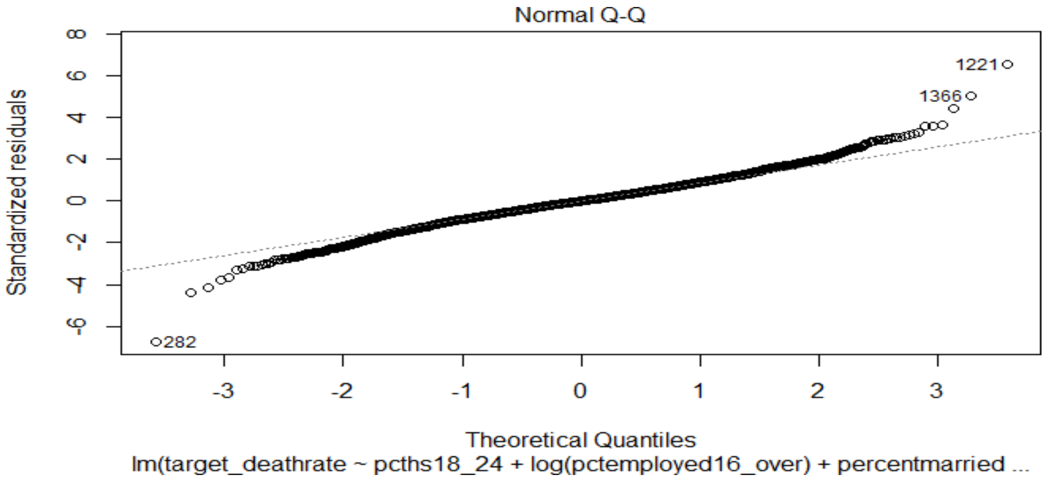
	GVIF	Df	GVIF^(1/2*Df))
pcths18_24	1.0809	1	1.03967
log(pctemployed16_over)	2.7306	1	1.65246
incidencerate	1.0953	1	1.04657
log(medincome)	2.1469	1	1.4652
pctunemployed16_over	1.919	1	1.3854
pctnocoverage	1.2862	1	1.134108

From the VIF values of the variables it is evident that every variable has a small multicollinearity with other variables but it is at a value of 1 to 2. In practice it is possible to have values less than 5 yet to get a better model.

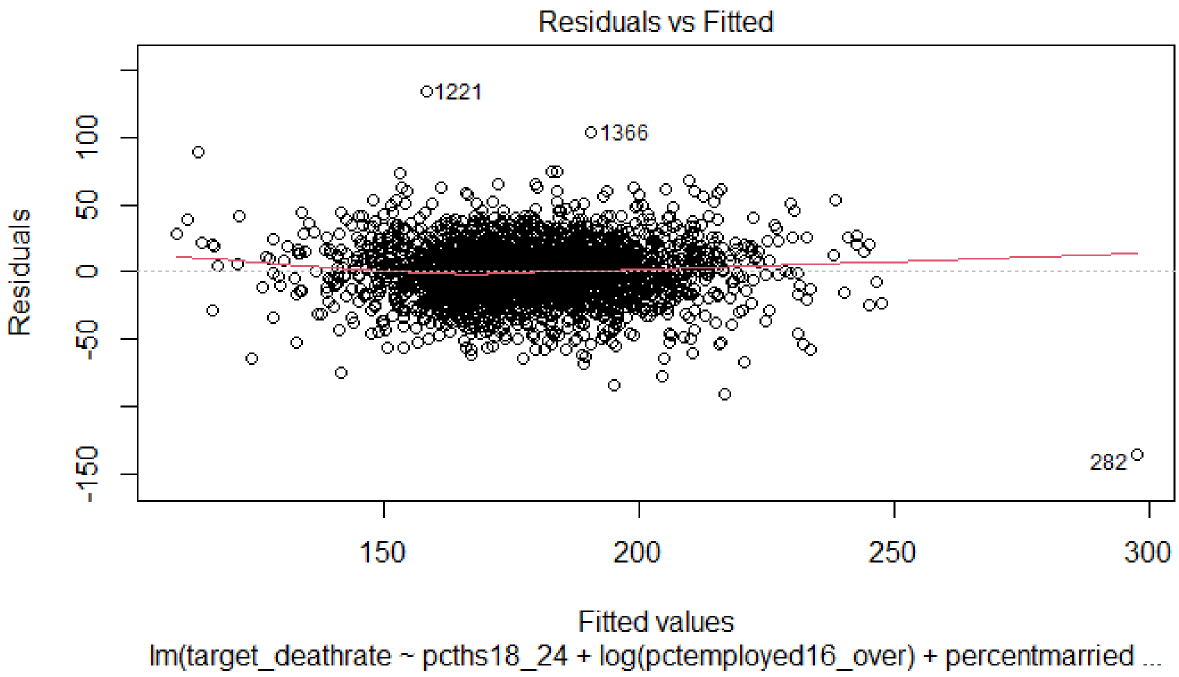
4.2 Model Diagnosis

The plots presented below are helpful in understanding the performance of the designed model. The residuals are normally distributed in the space and when we plot the residuals with fitted values, we are getting a constant variance plot which suggests that the sum of residuals are nearly equal to zero. These two points mentioned above are the principal components of a linear regression model to be a perfect fit for the data.

Graph 4-1 Normal Q-Q Plot Including pctnocoverage



Graph 4.2 Scatterplot of Residuals and Fitted Values Including pctnocoverage



When the designed model is tested against the test dataset which is obtained from selecting a random sample from original dataset, the RMSE value of the predicted values is 20.8. This says the model will give accuracy with deviation of 20 points from the actual value. This seems acceptable as the statistics of death rate range are max = 293 and min =59 and mean ~179.

Conclusion

From the designed model, it is clear that the target_death-rate due to cancer is dependant on factor of percentage of population aged between 18 to 25 who are graduated, percentage of employed individuals over 16 years old, incidence rate of cancer, median income of household in a county, percentage of unemployed population over 16 years old, percentage of population with no coverage or atleast one coverage of insurance.

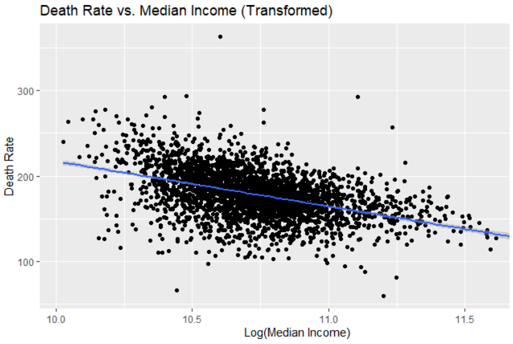
When talking about people with no insurance coverage in a county, a higher percentage of individuals with a high school diploma or graduate degree, lower unemployment rate, and higher percentage of individuals with health insurance coverage are associated with lower deathrate. On the other hand, higher disease incidence and lower median income are associated with higher death rate. The stats show that if the unemployment percent increases by 1%, the death-rate will increase by a factor of 0.64. Also, if the population with no health insurance increases by 1%, then the death rate of a county will increase by 0.28 times. The numbers and relations between independent and dependent variables are vise versa in the case of the population with at least one health insurance i.e the death rate is going down. Among all factors, income of a household and option of insurance are main variables judging the response variable.

The one important question to look into from this research is, whether the wealth of an individual, judges how long he/she can survive in case of cancer? This seems plausible if we are looking at the regressors, they are inclined towards the financial status of the individual. As we speak of wealth, when we consider a range of counties, each case in the dataset talks indirectly whether a county is rice or not. This suggests a relation between “death rate due to cancer differs between richer countries to poor countries” . But this hypothesis needs additional data to do research on.

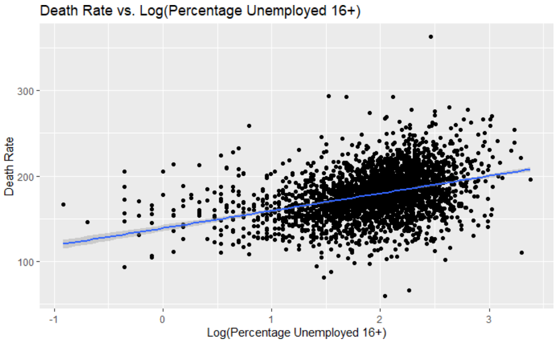
7. Appendix

A. Scatterplots between Death Rate and other numerical predictors

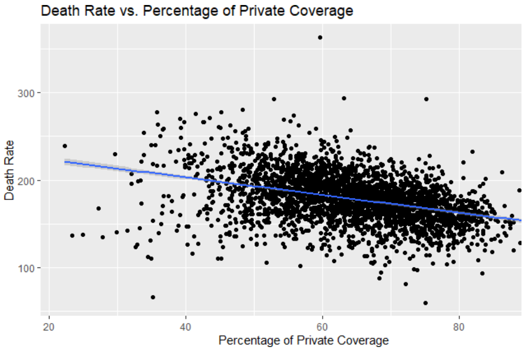
Graph A-1 Death Rate vs. Log(Median Income)



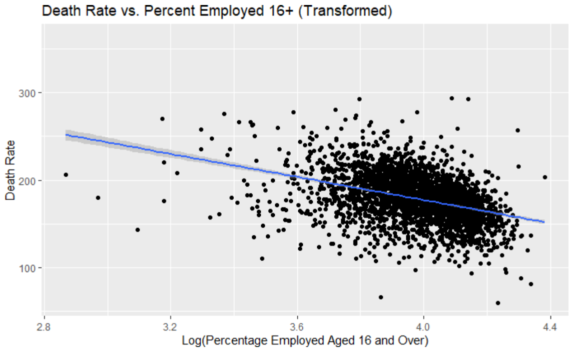
Graph A-2 Death Rate vs Log(percentage Unemployed 16+)



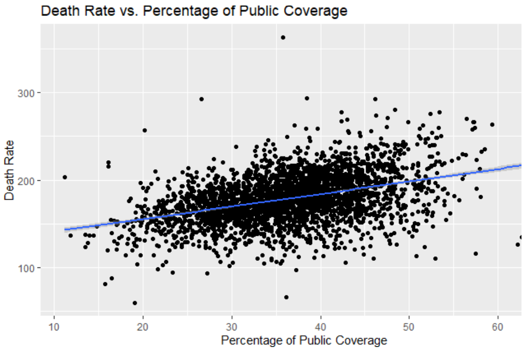
Graph A-3 Death Rate vs. Percentage of Private Coverage



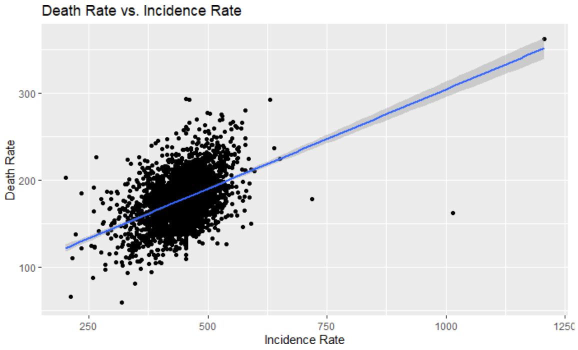
Graph A-6 Death Rate vs. Log(Percentage of People Employed Aged 16 and Over)



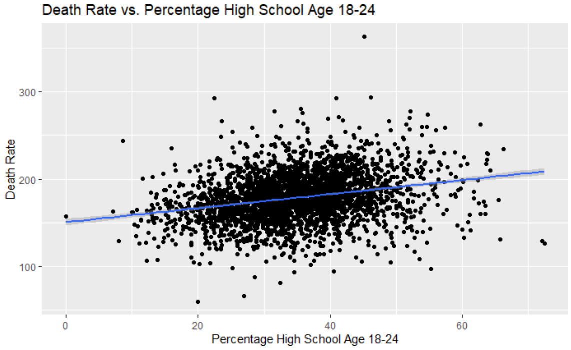
Graph A-4 Death Rate vs. Percentage of Public Coverage



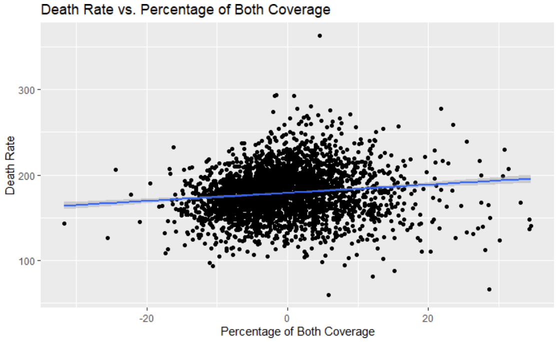
Graph A-8 Death Rate vs. Incidence Rate



Graph A-5 Death Rate vs. Percentage of People Aged 18-24 Who Completed High School



Graph A-11 Death Rate vs. Percentage of People With noCoverages



B. Regression Summaries between death rate and each of the predictors

Table B-1 Death Rate ~ Percentage of High School Ages 18-24

lm(target_deathrate ~ pcths18_24, data = cancer)

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	150.60664	1.93491	77.84	0
pcths18_24	0.80159	0.05351	14.98	0

Residual Standard Error: 26.79 on on 2893 degrees of freedom

Multiple R-Squared: 0.06863	Adjusted R-Squared: 0.06833
F-statistic: 224.4 on 1 and 3045 DF	p-value: 0

Table B-2 Death Rate ~ Log(Percentage of People Employed Aged 16 and Over)

lm(target_deathrate ~ log(pctemployed16_over), data = cancer)

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	440.779	11.347	38.85	<2e-16
log(pctemployed16_over)	-65.881	2.849	-23.12	<2e-16

Residual Standard Error: 25.24 on 2893 degrees of freedom

Multiple R-Squared: 0.156	Adjusted R-Squared: 0.1557
F-statistic: 534.7 on 1 and 2893 DF	p-value: <2e-16

Table B-3 Death Rate ~ Incidence Rate

lm(target_deathrate ~ incidencerate, data = cancer)

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	76.191299	3.718329	20.49	<2e-16
incidencerate	0.228597	0.008234	27.76	<2e-16

Residual Standard Error: 24.79 on 2893 degrees of freedom

Multiple R-squared: 0.202	Adjusted R-Squared: 0.2017
F-statistic: 770.1 on 1 and 3045 DF	p-value: <2e-16

Table B-4 Death Rate ~ Log(Median Income)

lm(target_deathrate ~ log(medincome), data = cancer)

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	744.8	20.3	36.80	<2e-16
log(medincome)	-52.7	1.886	-27.98	

Residual Standard Error: 24.76 on 2893 degrees of freedom

Multiple R-squared: 0.2046	Adjusted R-Squared: 0.2043
F-statistic: 783 on 1 and 3045 DF	p-value: <2e-16

Table B-7 Death Rate vs. Percentage of People Unemployed Aged 16 and Over

lm(target_deathrate ~ pctunemployed16_over, data = cancer)

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	154.7784	1.1566	133.82	<2e-16
pctunemployed 16_over	3.0418	0.1348	22.56	<2e-16

Residual Standard Error: 25.69 on 3045 degrees of freedom

Multiple R-Squared: 0.1432	Adjusted R-Squared: 0.1429
F-statistic: 508.9 on 1 and 3045 DF	p-value: <2e-16

Table B-8 Death Rate vs. Percentage of People With NoCoverages

lm(target_deathrate ~ pctnocoverage, data = cancer)

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	178.93726	0.50805	352.204	<2e16
pctnocoverage	0.48177	0.06876	7.006	3.03e-12

Residual Standard Error: 27.24 on 2893 degree of freedom

Multiple R-Squared: 0.01668	Adjusted R-Squared: 0.01634
F-statistic: 49.09	p-value: 3.033e-12