Understanding the impact of socio-economic factors on cancer mortality rates in South Dakota using variable selection techniques

Ben Pond, Emma Spors, & Isaac Ofori
Department of Mathematics and Statistics, South Dakota State University

STAT 686 : Spring 2022

1 Introduction

Humans were created to live healthy over a long period of time before they die. However, some have a very short lifespan due to misfortunes such as cancer. Cancer mortality has continued to be an alarming issue in the United States over the years. According to the Centers for Disease Control and Prevention (CDC), cancer was the second leading cause of death after heart disease in 2020 [National Center for Health Statistics, 2020].

This research project sought to understand how several socio-economic and lifestyle factors contributed to cancer mortality rates at a county level in the state of South Dakota (SD). Previous literature suggests there is a link between cancer mortality with several human behaviors such as smoking, alcohol consumption, feeding habits, educational attainment, and other factors such as genetic/environmental disorders, obesity, poverty, uninsured status, *etc* [National Cancer Institute, 2015].

The project considered the dependent variable (cancer mortality rates per 100,000 for all cancer sites) and eight predictors $(p-1)$ including cancer incidence rate per 100,000, obesity percentage, drinking percentage, uninsured percentage, food index, educational attainment, poverty percentage, and smoking percentage. The dataset included 66 observations (n) for the 66 counties in SD.

1.1 Data Sources

The cancer incidence and mortality data was extracted from the South Dakota Department of Health. The data included rates for all diagnosed cancers and cancer-related deaths for SD residents by county. The data was presented as an age-adjusted rate per 100,000 standardized to the 2000 United States (US) Census for all reported cases from 2009 to 2018 [South Dakota Cancer Registry, 2022].

The obesity percentage, smoking percentage, food index, and excessive drinking percentage data were from County Health Rankings and Roadmaps [County Health Rankings and Roadmaps, 2022]. The obesity data focused on the percentage of adults that reported a BMI of 30 or greater. Smoking and drinking were the percentage of adults who were smoking or drinking excessively. The food index measured several factors that contributed to a healthy food environment on scale of zero (worst) to ten (best) [County Health Rankings and Roadmaps, 2022].

The American Community Survey, via the US Census Bureau, provided information on insurance status, educational attainment, and poverty. The uninsured data was the percentage of a county who were uninsured as estimated in 2018. This excluded those on Medicare and Medicaid. However, it included those who only had insurance through Indian Health Services (IHS), as IHS was not considered full coverage [American Community Survey, 2018c]. Educational attainment data measured the percentage of county (25 years and older) with a bachelor's degree or higher in 2018-2019 [American Community Survey, 2018a]. Poverty data was the percentage of county residents for whom poverty status was determined in 2018-2019. Those whose family's pre-tax

income (USD) fell below the poverty threshold were considered in poverty [American Community Survey, 2018b].

For our final prediction model, the cancer mortality rates of the year 2020 were predicted using data collected during that year. Only three variables were used in that final model; cancer incidence, educational attainment, and poverty. Educational attainment [American Community Survey, 2020a] and poverty [American Community Survey, 2020b] information for the year 2020 was attained from the American Community Survey for that year. As no information on cancer incidences or mortality was available for the year 2020, the average of cancer incidence rates [South Dakota Cancer Registry, 2022] from 2014-2018 from the South Dakota Cancer Registry was reused under the assumption that the average rate of incidence would not change drastically over two years.

2 Exploratory Data Analysis

In order to better understand the data, exploratory data analysis was completed on the independent and dependent variables. A five number summary can be seen in Table 1 and histograms can be seen in Figure 2. In addition, due to the geographical nature of the data, choropleth maps of all the variables are featured in Figure 3. From this analysis, smoking percentage clearly and strongly deviated from a normal distribution, Figure 2h, which is confirmed with the numerical summary. Some other variables were not strictly normal, but generally followed the distribution. From visual analysis of Figure 3, it was clear that some of the factors were correlated.

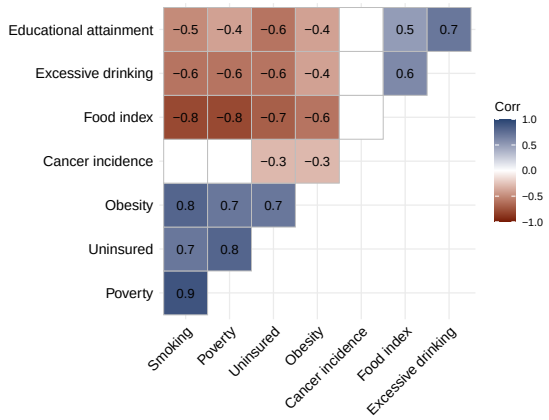
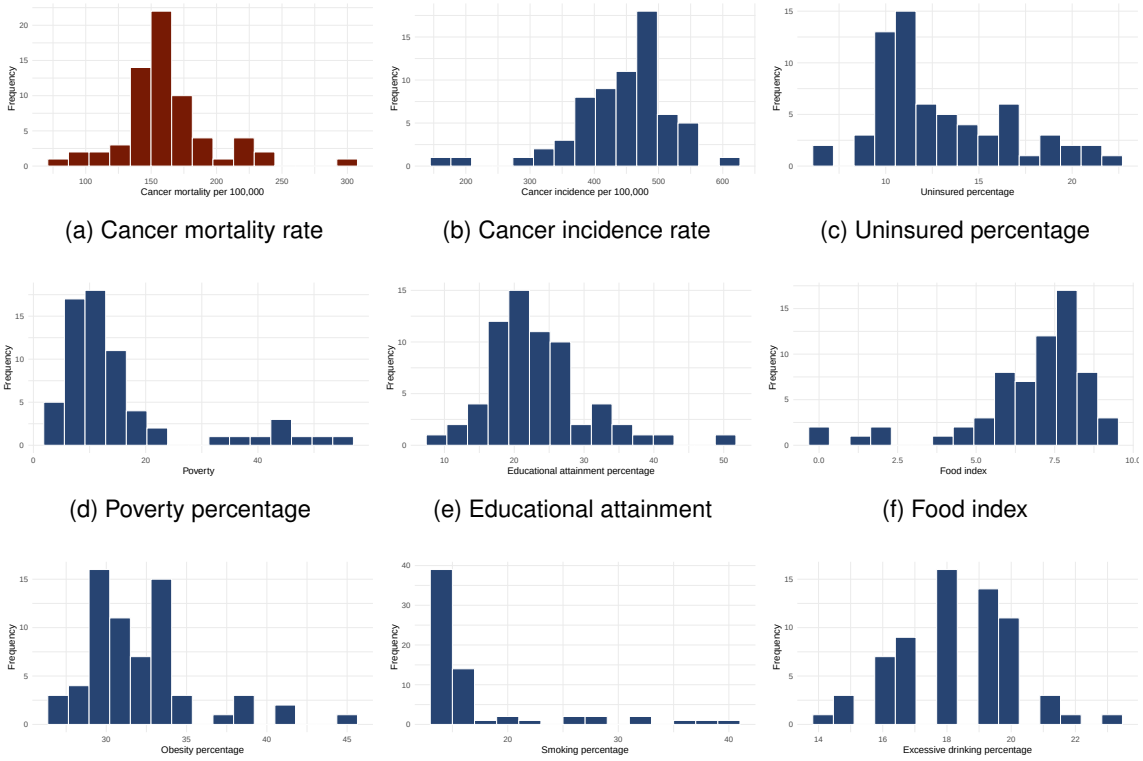


Figure 1: Correlation plot between predictor variables based on Pearson's correlation coefficient with only significant correlations shown.

Socioeconomic factors tend to be interrelated. Therefore, to understand how the factors were correlated, a correlation plot was created using Pearson's correlation coefficient, r, where only values with significant p-values ($\alpha = 0.05$ significance level) were shown (Figure 1). Upon analysis, it was noted that several of the variables were highly correlated ($|r| > 0.5$). For example, poverty was positively associated with uninsured percentage ($r = 0.8$) but negatively correlated with the food index. This indicated that these predictors would introduce the issue of multicollinearity in a linear model. Most of the correlation values matched expectation based on previous knowledge. However, the excessive drinking data did result in some unexpected values, such as a negative association with smoking and poverty.

Table 1: Numerical summary of all the variables considered in this study.

Variable	Min	1st Qu.	Median	3rd Qu.	Max
Mortality (rate)	75.0	143.9	155.2	173.3	295.9
Incidence (rate)	169.5	406.3	455.9	495.0	619.5
Uninsured (%)	6.40	10.43	11.65	15.50	21.90
Poverty (%)	4.00	8.53	11.15	14.7	55.5
Education (%)	8.5	18.73	22.00	26.23	49.90
Food index	0.00	6.13	7.30	8.00	9.20
Obesity (%)	27.00	30.00	31.00	33.75	45.00
Smoking (%)	13.00	14.00	15.00	16.75	41.99
Drinking (%)	14.00	17.00	18.00	19.00	23.00



(a) Cancer mortality rate (b) Cancer incidence rate (c) Uninsured percentage

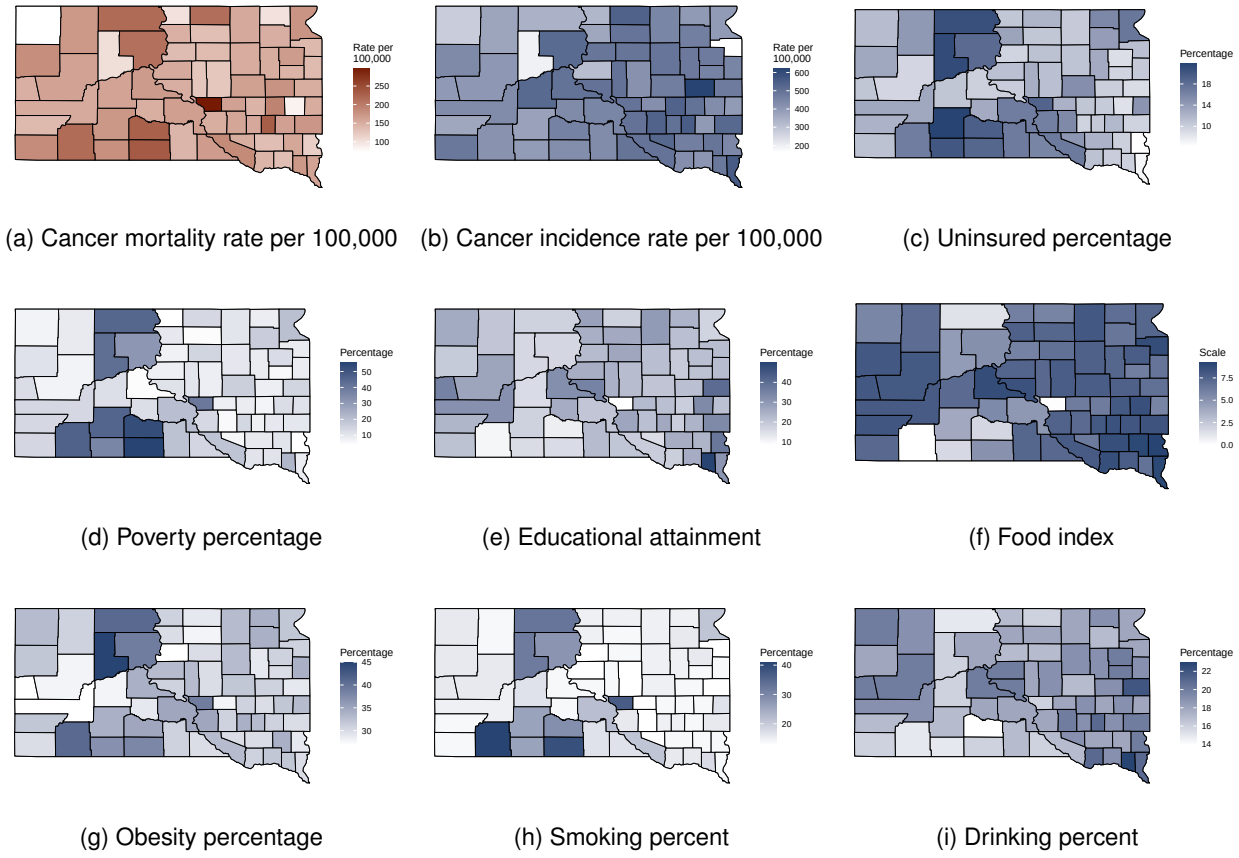
(d) Poverty percentage (e) Educational attainment (f) Food index

(g) Obesity percentage (h) Smoking percent (i) Drinking percent

Figure 2: Histograms of cancer mortality (red) and all eight of the predictor variables (blue).

3 Methods

All analysis was conducted with the statistical programming language R and the IDE RStudio [R Core Team, 2021, RStudio Team, 2020]. Several figures and general data wrangling were done with the sub-packages in the R meta-package *tidyverse* [Wickham et al., 2019]. Data and code can be found on the GitHub repository: espors/cancer_mortality_modeling.



(a) Cancer mortality rate per 100,000 (b) Cancer incidence rate per 100,000 (c) Uninsured percentage

(d) Poverty percentage (e) Educational attainment (f) Food index

(g) Obesity percentage (h) Smoking percent (i) Drinking percent

Figure 3: Choropleth maps of cancer mortality and all eight of the predictor variables.

3.1 Data Preparation and Adjustments

During EDA, it was confirmed that there was no missing data, so no imputation was needed. For penalized regression, it was necessary to standardize the independent data. This was achieved by calculating $x_{ij} = \frac{X_{ij} - \text{mean}(X_j)}{\text{sd}(X_j)}$, for $i = 1, ..., 66$ and $j = 1, ..., p-1$. This transformed data is represented in matrix notation as $\boldsymbol{x}$. No other transformations were completed.

3.2 Moran's I

When dealing with data related to area, distances, or overall geological location, we need to test whether the data is geospatially correlated. If the data is geospatially correlated, it would mean that a data point is positively or negatively influenced by the points around or near it. This makes the data non-independent, dependent on each other, which contradicts the assumption of random distribution. If this is the case, using a standard linear model would not be appropriate.

There are two kinds of correlation when discussing geospatial data. The first is spatial clustering, or positive correlation. If the response variable has a strong value for a particular data point, then the

surrounding data points nearest to it will also be strong. Likewise, if a response has a weak value for a data point, the points around it will also have weak responses. In short – like responses are by like responses. The second type of geospatial correlation is dispersed, or negative, correlation. The opposite of clustering, a data point with a strong response will repel other strong responses and weak responses will repel weak responses. The sign of the Moran's I statistic will determine the type of correlation [NKU, nd].

Correlation Matrix, W: In order to calculate Moran's I, a correlation matrix, *W*, is needed. The matrix *W* is a square $N \times N$ matrix, where N is the number of observations in the data. The entries of W, w_{ij}, are either $0, 1$, indicating if the i^{th} row entry is a neighbor of the j^{th} column entry. As an entry cannot be a neighbor of itself, when $i = j$, then $w_{ij} = 0$. This means the diagonal of W will be all zeroes. Lastly, the response variable y is needed in calculating the statistic. In this research, the cancer mortality rate per $100,000$ people of each county is the response variable.

Calculating Moran's I: Moran's I is interpreted similar to the Pearson correlation coefficient. The value falls within $[-1, 1]$, with values near -1 indicating strong negative correlation, values near 1 indicating strong positive correlation, and values near 0 indicating no correlation. The formal function is defined as

$$I = \frac{N}{\mathbf{W}} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

where **W** is the sum of all rows of the matrix W [NKU, nd].

Variance of Moran's I: The initial statistic will give an indicator of the strength of correlation. To determine its significance, a z-score is calculated which implies that the expected value and variance of the I statistic is needed.

The expected value is simple enough. Under the null hypothesis of no spatial correlation, $E[I] = -\frac{1}{N}$. Here, $N = 66$, making $E[I] = -\frac{1}{66} \approx -0.015$ [NKU, nd].

The variance is slightly more complicated to calculate,

$$V = \frac{(N * S4 - S3 * S5)}{(N-1)(N-2)(N-3)\mathbf{W}^2} - E[I].$$

Where

$$S1 = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (w_{ij} + w_{ji})^2,$$

$$S2 = \sum_{i=1}^{N} \left(\sum_{j=1}^{N} w_{ij} + \sum_{j=1}^{N} w_{ji} \right)^2,$$

$$S3 = \frac{\frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y})^4}{\left(\frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y})^2 \right)^2},$$

$$S4 = (N^2 - 3 * N - 3)S1 - N * S2 + 3\mathbf{W}^2, \text{ and}$$

$$S5 = (N^2 - N)S1 - 2N * S2 + 6\mathbf{W}^2.$$

Once Moran's I statistic is calculated, the z-score is calculated with $Z = \frac{I - E[I]}{\sqrt{V}}$ and is used to determine the I's significance [NKU, nd].

3.3 Multiple Linear Regression Model

To gain insights into how the selected socio-economic predictor variables impacted cancer mortality in SD's counties, a traditional multiple linear regression (MLR) model was considered. For this, the Y observations, cancer mortality rate per county, were assumed to have the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i,p-1} + \epsilon_i, \text{ for } i = 1, 2, ..., 66,$$

where $X_1, ..., X_{p-1}$ were the data for the $p - 1$ predictor variables and $\epsilon_i \sim N(0, \sigma^2)$ [Kutner et al., 2004]. This model is represented in matrix form as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where

$$\boldsymbol{Y}_{66\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ ... \\ Y_{66} \end{bmatrix}, \boldsymbol{X}_{66\times p} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & ... & X_{1,p-1} \\ 1 & X_{2,1} & X_{2,2} & ... & X_{2,p-1} \\ ... & & & ... & \\ 1 & X_{66,1} & X_{66,2} & ... & X_{66,p-1} \end{bmatrix}, \boldsymbol{\beta}_{p\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ ... \\ \beta_{p-1} \end{bmatrix}, \text{ and } \boldsymbol{\epsilon}_{66\times 1} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ ... \\ \epsilon_{66} \end{bmatrix}.$$

Here, we consider cancer mortality as a representative sample of cancer mortality rates in South Dakota for both past and future years, rather than population data for a specific time period. Thus $\boldsymbol{b}$ will be estimated, instead of finding the true population parameters, $\boldsymbol{\beta}$.

Since this form is often unknown, the associated estimated MLR model has the form

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{b}.$$

The $\boldsymbol{b}$ vector was estimated with the least squares criterion, which minimized the sum of squares, $SSE = \sum_{i=1}^{66} e_i^2$, or,

$$\boldsymbol{b} = \arg\min_{\beta} \left\{ \sum_{i=1}^{66} \left(Y_i - \beta_0 - \sum_{j=1}^{p-1} \beta_j x_{ij} \right)^2 \right\} \tag{1}$$

and can be found with the closed form solution

$$\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{Y}.$$

In addition, while it is used extensively for inference, σ^2 is unknown. Instead σ^2 is estimated with mean square error (MSE), which is calculated as $MSE = \frac{SSE}{n-p}$, where $E(MSE) = \sigma^2$. For

additional inference, the variance-covariance of $\boldsymbol{b}$ was necessary [Kutner et al., 2004]. Formally, $Var(\boldsymbol{b}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, which can be estimated with

$$\hat{\text{Var}}(\boldsymbol{b}) = MSE \cdot (\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{2}$$

The underlying assumptions for this MLR models include linearity of the true regression model, independence of the residuals ($\boldsymbol{e}$), normality of residuals, and constancy of variance of the residuals. In addition, multicollinerity of the predictor variables would result in imprecise information on the true regression coefficients [Kutner et al., 2004].

3.4 Variable Selection Techniques

A parsimonious model - one that achieves a desired level of fit with the fewest predictor variables - is a gold standard in statistical modeling to optimize interpret-ability. Three techniques to achieve such a model through variable selection include stepwise regression, least absolute shrinkage & selection operator (lasso) regression, and elastic net regression. The former is an iterative models that adds and drops predictor variables to minimize a specified statistic. The latter two are part of a family of models called penalized regression models. For these a penalty is incorporated in Equation 1 to control the parameter estimates. The theory of these models are described below, but the technical application was accomplished with the *MASS* (stepwise regression) and *glmnet* (penalized regression) R packages [Friedman et al., 2010, Venables and Ripley, 2002]. Note that for the penalized models described, the model expected that the independent data was standardized.

Stepwise Regression: Stepwise regression is an iterative search process that adds and drops predictor variables in a sequence of linear models and selects the model that minimizes Akaike Information Criterion (AIC) (see Appendix A, Section 6) [Kutner et al., 2004]. Thus, the resulting subset of selected variables creates model that is the "best" in terms of AIC. However, the method has been shown to return a sub-optimal model. Instead another model may be better based on prior knowledge, residuals, or other diagnostic criteria. Thus, it is common practice to use the selected subset as a starting point for further feature selection [Kutner et al., 2004].

Ridge Regression: While not used in this research, ridge regression was a foundation for the other penalized regression models that were used in this study [Hoerl and Kennard, 1970]. Instead of estimating coefficients as in Equation 1, the penalty term $\lambda \sum_{j=1}^{p} \beta_j^2$, known as "$L_2$", was added such that $\boldsymbol{b}$ was now found with

$$\boldsymbol{b} = \underset{\beta}{\arg\min} \left\{ \sum_{i=1}^{66} \left(Y_i - \beta_0 - \sum_{j=1}^{p-1} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p-1} \beta_j^2 \right\}. \tag{3}$$

Here, λ is the parameter that controls the size of the coefficients. When $\lambda = 0$, there is no effect and the problem returns to an ordinary least squares approach. As $\lambda \to \infty$, the penalty becomes large, and all of the coefficients are forced close to zero. In the middle of these extremes,

for $0 < \lambda < \infty$, some of the coefficients are set closed to zero, decreasing their effect size, but are still retained in the model. However, now λ must be estimated as well, and there is no longer a closed form solution for finding $\boldsymbol{b}$. Instead, cross-validation (CV) (described in Section 6, Appendix A), is used to estimate λ by determining which value minimizes MSE. Ridge regression is very effective at handling multicollinearity, but does not perform the task of feature selection, a goal of this study [Boehmke, 2018].

Lasso Regression: Lasso regression was developed to improve certain shortcomings from stepwise regression and ridge regression [Tibshirani, 1996]. Lasso regression performs feature selection (sub-setting predictor variables that significantly contribute to the model) and has been shown to improve prediction accuracy [Tibshirani, 1996]. Similar to ridge regression, a penalty term $\lambda \sum_{j=1}^{p} |\beta_j|$, known as "$L_1$", is added to Equation 1 such that,

$$\boldsymbol{b} = \arg\min_{\beta} \left\{ \sum_{i=1}^{66} \left(Y_i - \beta_0 - \sum_{j=1}^{p-1} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p-1} |\beta_j| \right\}. \tag{4}$$

Again, λ is the tuning parameter estimated with CV, but in contrast to the L_2 penalty, the L_1 penalty will force coefficients to zero, effectively dropping them from the model. Lasso regression therefore acts a feature selection technique, but it does not handle multicollinearity as well as ridge regression [Boehmke, 2018].

Elastic Net Regression: Elastic net regression combines the best attributes of ridge and lasso regression [Zou and Hastie, 2005]. To find the estimated coefficients, the minimization function essentially combines Equations 3 and 4 to handle both multicollinearity and feature selection. For elastic net, $\boldsymbol{b}$ is found with

$$\boldsymbol{b} = \arg\min_{\beta} \left\{ \sum_{i=1}^{66} \left(Y_i - \beta_0 - \sum_{j=1}^{p-1} \beta_j x_{ij} \right)^2 + \lambda_1 \sum_{j=1}^{p-1} \beta_j^2 + \lambda_2 \sum_{j=1}^{p-1} |\beta_j| \right\}.$$

Alternatively, this can be written as,

$$\boldsymbol{b} = \arg\min_{\beta} \left\{ \sum_{i=1}^{66} \left(Y_i - \beta_0 - \sum_{j=1}^{p-1} \beta_j x_{ij} \right)^2 + \alpha \sum_{j=1}^{p-1} \beta_j^2 + (1-\alpha) \sum_{j=1}^{p-1} |\beta_j| \right\}, \tag{5}$$

where $\alpha = \frac{\lambda_2}{\lambda_2 + \lambda_1} \in [0, 1)$. In Equation 5, α acts as a "mixing" proportion that determines how much of the L_1 or L_2 penalty is used. As $\alpha \to 0$, the model will more resemble ridge regression. Alternatively, as $\alpha \to 1$, the model will act more like lasso regression. In practice, estimating λ_1 and α through CV is sufficient for finding $\boldsymbol{b}$.

3.5 Model Diagnostics

To assess the validity of the models, several model diagnostic techniques were used. This included t-tests on coefficients, residual analysis, and calculation of variance inflation factors.

Coefficient t-Tests: Coefficient t-tests were used to determine if an individual correlation coefficient was significantly different than zero. It can be shown that $t = \frac{b_k - \beta_k}{s_{b_k}} \sim t_{n-p}$, where $\beta_k = 0$ for this test and s_{b_k} was the square root of the k^{th} element along the main diagonal of the variance-covariance matrix of $\boldsymbol{b}$ in Equation 2. Typically, for p-values > 0.1, the associated coefficient is not considered to be significantly different from zero [Kutner et al., 2004].

Residual Analysis: For linear regression models, the main assumptions were that error terms of the estimated regression model are independent, normally distributed, and had constancy of variance [Kutner et al., 2004]. To check if the normality and constancy of variance assumptions were met, we used diagnostic residual plots. Normal Q-Q plot showed the normality (or lack thereof) of the error terms. In practice, points which closely follow the given diagonal line indicate normality. The Residual vs Fitted plots visualized the pattern of the variance of the error terms. The plot should show the error terms fluctuating randomly around the red base line.

Multicollinearity: As high correlation between several of the predictor variables was noted during EDA, variance inflation factors (VIF) values were used to assess multicollinearity in the models. The VIF of each predictor variable in the model was calculated using the *car* package in R [Fox and Weisberg, 2019].

This can also be calculated with

$$(\text{VIF})_k = \frac{1}{1 - R_k^2}, k = 1, 2, ...p - 1,$$

where R_k^2 was the coefficient of multiple regression (R-squared) (see Appendix A, Section 6) when a predictor was regressed on the other $p - 2$ predictor variables. The higher the VIF value of a predictor, the greater the correlation of that predictor with other predictor variables. Values of more than 4 or 5 were regarded as being moderate. However, values of 10 or more were being regarded as very high. A mean VIF value for a model greater than 1 indicates multicollinearity in the model [Kutner et al., 2004].

3.6 Predictions

To estimate future rates of cancer mortality by county, new X data can be gathered in the form $\boldsymbol{X}_{n,i} = \begin{bmatrix} 1 & X_{n,1} & ... & X_{n,p-1} \end{bmatrix}^T$ such that $\hat{\boldsymbol{Y}}_{n,i} = \boldsymbol{X}_{n,i}\boldsymbol{b}$ for $i = 1..., 66$. The standard error for these estimated cancer mortality rates are given by

$$s_i^2\{\hat{Y}_{n,i}\} = MSE\left(1 + \boldsymbol{X}_{n,i}^T(\boldsymbol{X}_{n,i}^T\boldsymbol{X}_{n,i})^{-1}\boldsymbol{X}_{n,i}\right).$$

Since 66 new predictions were estimated, a multiple comparisons adjustment was made for the confidence intervals using the Bonferroni procedure [Kutner et al., 2004]. The confidence interval was created with

$$\hat{Y}_{n,i} \pm B \cdot s\{\hat{Y}_{n,i}\}, \text{ where } B = t_{(1-\alpha/(2\cdot 66)), df=n-p)}.$$

and t is the quantile student's t distribution.

4 Results

The process for this study was as follows: compute Moran's I to determine if there was spatial auto-correlation for the cancer mortality rate (Section 4.1), create regression models from a full model to a reduced model (Section 4.2), and predict future cancer mortality rates for SD counties based on a regression model from 4.2 (Section 4.3).

4.1 Moran's I

From the results of Moran's I, Table 2, the I value is negative and slightly less than the expected value. The I value being very close to zero indicated very little spatial correlation, though the negative value would suggest slight negative correlation, if any.

Using S_i's to calculate the variance of I, the resulting z-score is a value of -0.35. This corresponds to a p-value of 0.36, which is much larger than 0.05. As a result, we fail to reject the null hypothesis of no spatial correlation. This suggests the data is not spatially correlated, so the assumption that the model residuals are independent is upheld.

Table 2: Numerical summary for Moran's I and the z-score.

Statistic	**Value**
Mean Y	160.4379
Moran's I	-0.0678
E[I]	-0.0152
S1	692
S2	256
S3	5.5864
S4	3221676
S5	3661512
Variance	0.0214
W	348
Z	-0.3584
p	0.3632

4.2 Models

Models summaries can be found in Table 3, model diagnostic plots in Figure 4, and summary VIF values in Table 4.

For Model 1, all eight predictor variables were used. Cancer incidence & smoking were positively associated with cancer mortality and obesity & educational attainment were negatively associated with mortality at the 0.1 significance level. The other coefficients were determined to be not significantly different from zero. For all models, coefficients with p-values less than 0.1 will not be discussed for brevity. The model attained an adjusted R-squared value of 0.57. In the diagnostic plots, Figures 4a & 4f, showed some deviation from normality in the residuals. Most concerning was the mean VIF value of 4.1 which indicated high multicollinearity in the predictor variables.

Given the non-normal distribution of the smoking percentage and its high correlation with other values, smoking was removed from the set of predictor variables to create Model 2. Cancer incidence & poverty percentage were positively associated with mortality and educational attainment was negatively associated with cancer mortality at the 0.01 significance level. The model attained an adjusted R-squared value of 0.56 and a mean VIF value of 2.7. There was an improvement in the residuals, Figures 4b and 4f, over Model 1.

Next, stepwise regression in the forwards and backwards direction was completed for Model 3. The variable subset that minimized AIC included positively associated cancer incidence & poverty, negatively associated educational attainment, and drinking percentage as insignificant. The model attained an adjusted R-squared value of 0.55. The Q-Q plot, Figure 4d suggests the residuals are generally normal, but the residual versus fitted plot, Figure 4i showed a potential parabolic pattern in the residuals. The mean VIF value did decrease to 2.34 compared to Models 1 & 2.

Next, lasso and elastic net regression were completed as alternative methods of feature selection. Through CV, $\hat{\lambda} = 2.26$ was selected as the parameter estimate which minimized MSE in the lasso model. The resulting linear model selected cancer incidence, food index, educational attainment, and poverty as predictors. For elastic net regression, the CV grid search returned $\hat{\alpha} = 0.9$ and $\hat{\lambda} = 2.32$ which meant the model tended toward a lasso model over a ridge regression model. Thus, the same predictor variables were selected as associated with cancer mortality. Note that since penalized regression required that the independent variables be standardized before modeling, the coefficients from these models are not in the units of the original predictor variables and thus are not comparable to the other models. Instead, the selected variables were used in the traditional MLR model, resulting in Model 4. Cancer incidence & poverty were positively associated with cancer mortality and educational attainment was negatively associated with mortality at the 0.015 significance level. Food index was also selected but the coefficient was not significantly different from zero. The model attained an adjusted R-squared value of 0.54 and had a mean VIF value of 2.77. The model residuals indicated some deviation from normal, particularly in the tails.

Since feature selection techniques result in a good starting point for further model development, the decision was made to further reduce Models 3 and 4. Model 5 was created using cancer incidence, educational attainment, and poverty as predictor variables as these were the three significant variables in both Model 3 and Model 4. Cancer incidence and poverty were positively associated with cancer mortality. Educational attainment was negatively associated with mortality. All were significant. The model attained an adjusted R-squared value of 0.55. The Q-Q plot, Figure 4e, showed some deviation from a normal distribution in the tails. There is also a slight pattern in the residuals in Figure 4j.

Table 3: Summary of the regression models including the estimated coefficients, t-value, p-value, and adjusted R-squared. The intercept is not included because it is not within the scope of this problem.

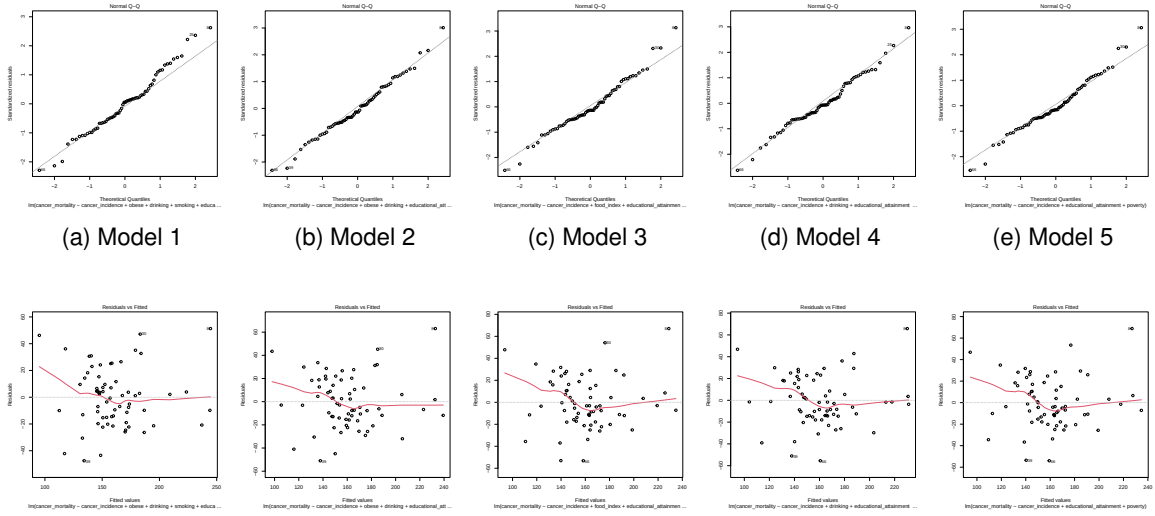
	Model 1 - Full			**Model 2 - No Smoking**		
Variables	Coeff	t-value	Pr($>\|t\|$)	Coeff	t-value	Pr($>\|t\|$)
Cancer Incidence (Rate)	0.188	4.675	$1.84 \cdot 10^{-5}$	0.188	4.555	$2.74 \cdot 10^{-5}$
Obesity (%)	-2.44	-1.719	0.0911	-1.32	-0.995	0.323
Education Attainment (%)	-1.709	-2.607	0.011	-2.058	-3.185	0.002
Poverty (%)	0.931	1.242	0.219	2.038	4.024	0.0002
Uninsured (%)	-0.623	-0.407	0.685	-1.178	-0.763	0.448
Food Index	0.237	0.093	0.926	-1.359	-0.549	0.585
Drinking (%)	3.44	1.303	0.197	4.23	1.586	0.118
Smoking (%)	2.90	1.964	0.054	–	–	–
	Adj. R-sqr: 0.567			Adj. R-sqr: 0.546		

	Model 3 - Stepwise			**Model 4 - Penalized**		
Variables	Coeff	t-value	Pr($>\|t\|$)	Coeff	t-value	Pr($>\|t\|$)
Cancer Incidence (Rate)	0.204	5.205	$2.41 \cdot 10^{-6}$	0.210	5.298	$1.7 \cdot 10^{-6}$
Education Attainment (%)	-1.799	-3.130	0.002	-1.268	-2.492	0.015
Poverty (%)	1.716	5.673	$4.11 \cdot 10^{-7}$	1.457	1.242	0.00034
Food Index	–	–	–	-0.599	-0.243	0.808
Drinking (%)	3.787	1.479	0.144	–	–	–
	Adj. R-sqr: 0.554			Adj. R-sqr: 0.538		

	Model 5 - Reduced		
Variables	Coeff	t-value	Pr($>\|t\|$)
Cancer Incidence (Rate)	0.21	5.333	$1.44 \cdot 10^{-6}$
Educational Attainment (%)	-1.310	-2.760	0.007
Poverty (%)	1.522	5.530	$6.82 \cdot 10^{-7}$
	Adj. R-sqr: 0.545		

Table 4: VIF mean value and maximum value for the regression models.

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
Mean VIF	4.07	2.30	1.71	1.93	1.19
Max VIF	10.76	3.70	2.34	2.77	1.30



(a) Model 1 (b) Model 2 (c) Model 3 (d) Model 4 (e) Model 5

(f) Model 1 (g) Model 2 (h) Model 3 (i) Model 4 (j) Model 5

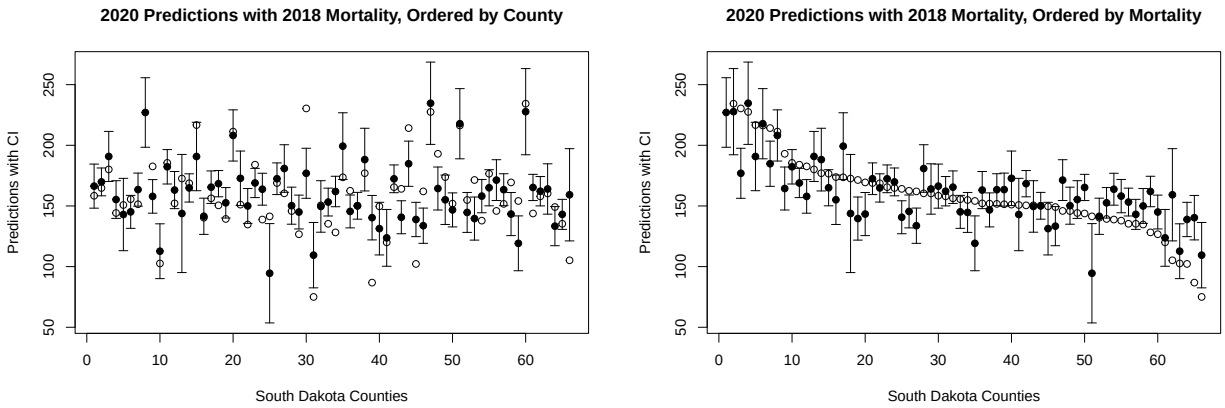
Figure 4: Quantile-quantile plots (top row) and residual versus fitted plots (bottom row) of Models 1 through 5.

4.3 Predictions

Among all models, Model 5 had the lowest mean VIF while maintaining a relatively high adjusted R-squared value. Thus, we chose this model to create cancer mortality predictions for 2020 for each county. To do this, the variables of educational attainment [American Community Survey, 2020a] and poverty [American Community Survey, 2020b] were obtained for the year 2020. The cancer incidence rate was not available at the county level after 2018. However, according to the National Cancer Institute, cancer incidence rates in SD have remained steady in recent years [National Cancer Institute, 2018]. As the rate used to build the model was an average for 2014 to 2018, the assumption was made that the average incidence rate would not change drastically in future years.

Figure 5a shows the prediction intervals for mortality rates in each county, along with the mortality rate from 2018. The mortality rate from 2018 should not necessarily coincide with the prediction intervals as they are of different years, but the mortality rates will be used as a reference for how the predictions do. We see that some counties may expect higher mortality rates than they currently have and others will have fewer. Figure 5b shows the 2018 mortality rates ordered in descending order with the corresponding county's prediction for 2020. When ordered like this, the predictions seem to do a better job as a whole of estimating the mortality rate than on their own for an individual county.

It could be interpreted from these plots that the predictions taken as a whole do well to estimate the trend of cancer mortality rates across the state of South Dakota, but for individual counties performance varies. Specifically, consider the current rates and predictions for the counties with the top ten most populous cities in SD, Figure 6. These counties include Beadle, Brookings, Brown,



(a) Alphabetical order of counties

(b) Ordered by predicted mortality

Figure 5: Scatter plot of 2020 mortality predictions (solid) and confidence bands with the 2018 mortality rates (circle).

Codington, Davison, Hughes, Lawrence, Minnehaha, Pennington, and Yankton. There, all of the past mortality rates (red stars) are within the confidence bands of the future predictions. This implies, that in the coming years, none of these counties will have significantly higher or lower cancer mortality rates than they have in previous years.

5 Discussion

The issue of multicollinearity was confirmed as the mean VIF for Model 1 was very high at 4.07. Removing smoking as a predictor in Model 2 helped to improve this value, but the feature selection techniques better solved this issue. Stepwise regression, Model 3, had a mean VIF value of 2.34 and penalized regression, Model 4, had a value of 2.77. Surprisingly, it was expected that Model 4 would have a smaller VIF over Model 3 as elastic net was specifically designed to handle multicollinearity in the data. However, since $\hat{\alpha} = 0.9$, the model was behaving more like lasso model which does not deal with multicollinearity as well. However, the decision to keep only the three "core" predictor variables in Model 5 further decreased mean VIF to 1.19 which is a more acceptable level.

While the mean VIF values were reduced drastically across the models, adjusted R-squared was not. In Model 1 with all the predictors, the adjusted R-squared value was 0.567, but after extensive feature selection was completed for Model 5, the adjusted R-squared had only decreased slightly to 0.545. This indicated that dropping several of the models increased interpret-ability while maintaining the model's ability to explain the mortality data.

Throughout all five models, the coefficients for cancer incidence rate, educational attainment, and poverty percentage were repeatedly different from zero. This indicated that these variables

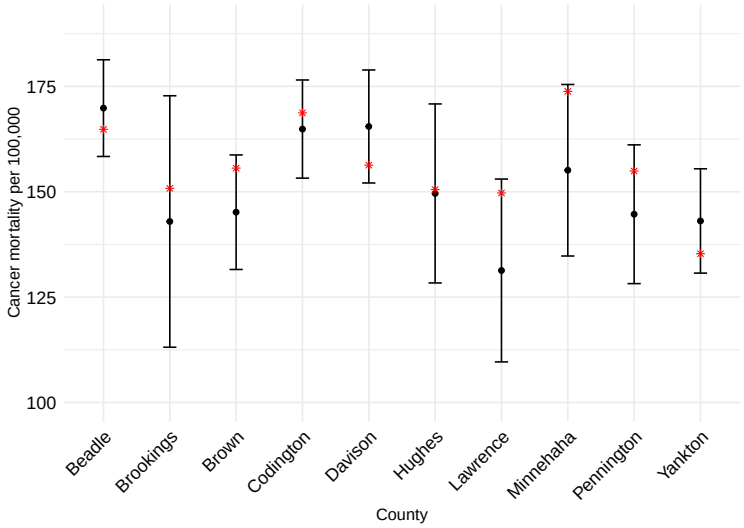


Figure 6: Plot of 2020 predictions (black circle) with confidence bands and 2018 mortality rates (red stars) for counties with the top ten most populous cities in SD.

had an association with cancer mortality rates in SD's counties. Morever, current literature agrees with these findings. In Model 5, the coefficient for incidence suggested that for an additional cancer diagnosis per 100,000 in a particular county, there was an additional 0.21 deaths per 100,000. Intrinsically, this makes sense as counties with more individuals diagnosed with cancer would therefore have more people likely to pass away from cancer. Fortunately, according to the CDC, from 2001 to 2020, cancer deaths have decreased 27 percent from 196.5 to 144.1 deaths per 100,000 while incidence rate remain relatively stagnant [Centers for Disease and Control and Prevention, 2022]. Thus, while cancer incidence was and will continue to be a large driver of cancer related deaths, we need to understand what other factors are contributing to mortality rates.

For example, even when controlling for cancer incidence and educational attainment (*i.e.* holding their values constant), Model 5 suggests that for a one percent increase in the poverty level of a county, there was an additional 1.522 cancer deaths per 100,000. That is, poorer counties in South Dakota tend to have higher mortality rates than more affluent counties. Poverty, particularly persistent poverty, has been linked to increased cancer deaths in the United States [Moss et al., 2020]. In SD, 11 counties have met the status of persistent poverty by having a poverty percentage of over 20 percent when measured by the 1980, 1990, and 2000 censuses. Unfortunately, community members from these areas typically are a minority population (American Indian/Alaskan Native in the case of SD), live in rural area away from healthcare services, and experience higher levels of chronic stress, among other disparities which lend to higher cancer deaths [Moss et al., 2020].

Finally, for a one percent increase in educational attainment in a county, there was a decrease of 1.31 deaths per 100,000. Counties whose residents have more years of education had fewer deaths than those who had fewer years of education. The link between cancer mortality and education is less clear than poverty's relationship with cancer, but one recent study did find that fewer years

of education was strongly associated with increased risk of cancer death. This is particularly true for cancers such as prostate, colorectal, and breast [Barcelo et al., 2021]. In a broader context, educational attainment is associated with overall better health [Zajacova and Lawrence, 2018]. In aggregate, those with college diplomas may have higher-paying stable jobs which afford them the ability to consistently receive routine healthcare, access to healthier food, and reduced chronic stress [Zajacova and Lawrence, 2018].

This study also included predictions on expected cancer mortality rates for 2020. Specifically, the counties with the ten most populous cities are not expected to have mortality rates that will be significantly different from historical numbers.

Limitations: Excluding cancer incidence, all predictor variables were from self-reported surveys at a sample level that were extrapolated to the county level. This may inject bias into the data from the potential unreliability of answers and uncertainty from the modeling. For example, individuals reporting weight may have under-report their actual numbers due to the stigma surrounding obesity. This, in turn, negatively impacts inferences made on the coefficients. Also, at a relatively small sample size $n = 66$, the residuals from our models indicate that not all the necessary assumptions were met. This indicates that our predictions are not as powerful as they could be. Moreover, we used the same incidence data in both building the model and creating the predictions, which is not considered good practice.

Future Work: To account for some of the underlying distributions of the data, modeling could be expanded to non-parametric models such as multivariate adaptive regression spline (MARS) models or clustering algorithms. Another option would be to complete a spatial linear model on the data. Additional parameters such as distance to healthcare centers and exposure to hazardous chemical could also be brought into the model to account for more variability in the data.

The identification of cancer incidence, poverty, and education as contributors to cancer mortality give insight into how to make public health decisions moving forward. Using this information to make predictions, health officials can identify counties who are expected to have higher rates of cancer-related deaths in the coming years. Empowered with this knowledge, the SD Department of Health can direct funding and resources to communities who need them the most.

6 Appendix A - Additional Theory

Akaike information criterion: AIC can be calculated with

$$\text{AIC} = 2p - 2\ln(\hat{L}),$$

where p is the number of estimated parameters and $\hat{L}$ is the maximum value of the likelihood function for the model [Kutner et al., 2004]. AIC rewards goodness of fit while penalizing the number of parameters in the model; a "good" model is one that minimizes AIC.

R-squared and Adjusted R-squared: R-squared, or coefficient of multiple determination, is calculated with

$$R^2 = 1 - \frac{SSE}{SSTO} = 1 - \frac{\sum_{i=1}^{n} Y_i - \hat{Y}_i}{\sum_{i=1}^{n} Y_i - \bar{Y}}.$$

R-squared measures the proportion of the variability in the Y data as explained by the linear model [Kutner et al., 2004]. Typically a value closer to 1 is desirable. However, R-squared will increase as more predictors are added to the model, even if if they do not add much value. Thus, adjusted R-squared, is often computed instead which penalizes the addition of more parameters [Kutner et al., 2004]. This is found with

$$R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SSTO}.$$

k-Fold Cross Validation: k-Fold cross validation (CV) is a re-sampling technique that can be used to measure model performance and tune parameters. CV divides the original sample into k equal partitions, reserves the first "fold" as testing data and combines the rest of the data as training data to development a given model and calculate a specified measure of accuracy. Repeat this step k times, except treat the ith fold as the testing data each iteration. Finally, take the average of the test accuracy as the overall accuracy for the model. The standard value of k is 10. In the case, penalized regression CV is repeated over the range of potential λ and α values with MSE as the measure of accuracy. The values of $\hat{\lambda}$ and $\hat{\alpha}$ are from the model that result in the smallest MSE value.

7 Appendix B - Dataset Snippet

All code and full data set are available at https://github.com/espors/cancer_mortality_modeling.

Table 5: Data table containing some of the data used in this study. The original CSV file was analogous to a 67 by 11 matrix.

County	FIPS	Incidence	Morality	Uninsured	Obese	Drinking	Smoking	...
Aurora	46003	535.7	158.4	15.6	31	20	13	...
Beadle	46005	467.8	164.8	15.2	32	17	15	...
Bennett	46007	395	180.1	20.1	37	15	26	...
Bon Homme	46009	424.7	144.2	11.1	31	21	16	...
Brookings	46011	474.2	150.8	9.1	29	22	13	...
Brown	46013	447.1	155.6	10.5	33	18	15	...
Brule	46015	498.3	151.4	16.5	35	18	16	...
Buffalo	46017	495	295.9	18.9	39	16	37	...
Butte	46019	437.6	182.6	13.2	32	18	16	...
Campbell	46021	331	102.6	10.7	31	17	15	...
Charles Mix	46023	494.9	185.5	17	33	17	21	...
Clark	46025	479.2	152.1	13	29	19	15	...
Clay	46027	482.3	172.6	9.9	32	23	18	...
Codington	46029	474.3	168.7	8.9	31	20	16	...
Corson	46031	343.4	216.7	20.8	41	15	33	...
Custer	46033	369	140.3	12.3	32	17	14	...
Davison	46035	499	156.3	9.8	34	19	15	...
Day	46037	464.1	150.5	14.1	34	17	16	...
Deuel	46039	450.7	139.3	10.6	34	19	14	...
Dewey	46041	516.3	211.3	18.4	39	18	28	...
Douglas	46043	522.4	151.1	14.6	30	19	13	...
Edmunds	46045	492.9	134.7	10.7	28	18	14	...
...	...	...	...	...	...	...	...	...
Sully	46119	315	154.1	11.2	33	20	13	...
Todd	46121	428	234.3	18.7	38	16	39	...
Tripp	46123	433.3	143.8	16.9	31	17	17	...
Turner	46125	489.6	157.9	10	34	18	14	...
Union	46127	558.1	160.3	7	30	21	14	...
Walworth	46129	368.7	149.1	13.5	30	18	15	...
Yankton	46135	409.8	135.3	10.2	32	19	16	...
Ziebach	46137	201.1	105.2	21.2	45	16	32	...

References

[American Community Survey, 2018a] American Community Survey (2018a). *Educational Attainment, ACS 5-Year Estimates*. Retrieved from https://www.census.gov/topics/education/educational-attainment.html.

[American Community Survey, 2018b] American Community Survey (2018b). *Poverty Status in the Past 12 Months*. Retrieved from https://www.census.gov/topics/income-poverty/poverty/data/tables/acs.html.

[American Community Survey, 2018c] American Community Survey (2018c). *Small Area Health Insurance Estimates*. Retrieved from https://www.census.gov/programs-surveys/sahie.html.

[American Community Survey, 2020a] American Community Survey (2020a). *Educational Attainment, ACS 5-Year Estimates*. Retrieved from https://www.census.gov/topics/education/educational-attainment.html.

[American Community Survey, 2020b] American Community Survey (2020b). *Poverty Status in the Past 12 Months*. Retrieved from https://www.census.gov/topics/income-poverty/poverty/data/tables/acs.html.

[Barcelo et al., 2021] Barcelo, A., Duffett-Leger, L., Pastor-Valero, M., Pereira, J., Colugnati, F. A., and Trapido, E. (2021). The role of education on cancer amenable mortality among non-hispanic blacks & non-hispanic whites in the united states (1989- 2018). *BCM Cancer*, 21(907).

[Boehmke, 2018] Boehmke, B. (2018). Regularized regression. http://uc-r.github.io/regularized$_{r}egression$.

[Centers for Disease and Control and Prevention, 2022] Centers for Disease and Control and Prevention (2022). *An Update on Cancer Deaths in the United States*. https://www.cdc.gov/cancer/dcpc/research/update-on-cancer-deaths/index.htm.

[County Health Rankings and Roadmaps, 2022] County Health Rankings and Roadmaps (2022). *South Dakota Rankings Data, 2018*. Retrieved from https://www.countyhealthrankings.org/app/south-dakota/2021/overview.

[Fox and Weisberg, 2019] Fox, J. and Weisberg, S. (2019). *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, third edition.

[Friedman et al., 2010] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

[Hoerl and Kennard, 1970] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

[Kutner et al., 2004] Kutner, M. H., Nachtsheim, C. J., and Neter, J. (2004). *Applied Linear Regression Models*. McGraw-Hill Irwin, New York, NY, 4 edition.

[Moss et al., 2020] Moss, J. L., Pinto, C. N., Srinivasan, S., Cronin, K. A., and Croyle, R. T. (2020). Persistent Poverty and Cancer Mortality Rates: An Analysis of County-Level Poverty Designations. *Cancer Epidemiology, Biomarkers Prevention*, 29(10):1949–1954.

[National Cancer Institute, 2015] National Cancer Institute (2015). *Risk Factors for Cancer*. Retrieved from https://www.cancer.gov/about-cancer/causes-prevention/risk.

[National Cancer Institute, 2018] National Cancer Institute (2018). State cancer profiles - incidence rates. https://statecancerprofiles.cancer.gov/incidencerates/index.

[National Center for Health Statistics, 2020] National Center for Health Statistics (2020). *Leading Causes of Death*. Retrieved from https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm.

[NKU, nd] NKU (nd). *Global Moran's I and Global Geary's C*. .

[R Core Team, 2021] R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[RStudio Team, 2020] RStudio Team (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA.

[South Dakota Cancer Registry, 2022] South Dakota Cancer Registry (2022). *South Dakota Cancer Incidence and Mortality*. Retrieved from https://www.sdcancerstats.org/.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

[Venables and Ripley, 2002] Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

[Wickham et al., 2019] Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

[Zajacova and Lawrence, 2018] Zajacova, A. and Lawrence, E. M. (2018). The relationship between education and health: Reducing disparities through a contextual approach. *Annual Review of Public Health*, 39(1):273–289. PMID: 29328865.

[Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.