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

Benjamin Pond, Emma Spors, and Isaac Ofori

Mathematics and Statistics, South Dakota State University

May 4, 2022

Outline

Introduction

- According the CDC, cancer was the second leading cause of death in 2020 [National Center for Health Statistics, 2020].
- There is a growing body of evidence that socio-economic (poverty, insurace status, *etc*,...) factors contribute the cancer mortality rates [National Cancer Institute, 2015].

Introduction

- According the CDC, cancer was the second leading cause of death in 2020 [National Center for Health Statistics, 2020].
- There is a growing body of evidence that socio-economic (poverty, insurace status, *etc*,...) factors contribute the cancer mortality rates [National Cancer Institute, 2015].
- This study sought to understand how eight socio-economic and lifestyle factors impacted cancer mortality rates per 100,000 on a county level in South Dakota (SD).
- To do this, we considered several variable selection techniques to create a parsimonious model.

Data Sources

- **Cancer incidence** and **mortality** age-adjusted rates per 100,000 (2009-2018) were taken from the South Dakota Department of Health [South Dakota Cancer Registry, 2022].

Data Sources

- **Cancer incidence** and **mortality** age-adjusted rates per 100,000 (2009-2018) were taken from the South Dakota Department of Health [South Dakota Cancer Registry, 2022].
- **Uninsured** percentage, **poverty** percentage, and **educational attainment** (percentage of adults with bachelors degree or higher) were collected by the American Community Survey via the US Census Bureau [American Community Survey, 2018c,
  American Community Survey, 2018b,
  American Community Survey, 2018a].

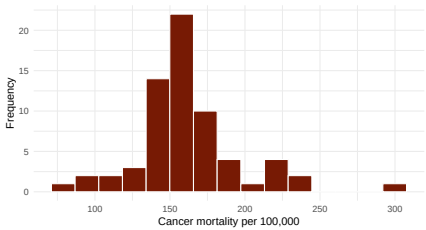
Data Sources

- **Cancer incidence** and **mortality** age-adjusted rates per 100,000 (2009-2018) were taken from the South Dakota Department of Health [South Dakota Cancer Registry, 2022].
- **Uninsured** percentage, **poverty** percentage, and **educational attainment** (percentage of adults with bachelors degree or higher) were collected by the American Community Survey via the US Census Bureau [American Community Survey, 2018c,
American Community Survey, 2018b,
American Community Survey, 2018a].
- **Obesity** percentage, **food index**, **smoking** percentage, and excessive **drinking** percentage were provided by the Counties Ranking and Roadmap site [County Health Rankings and Roadmaps, 2022].

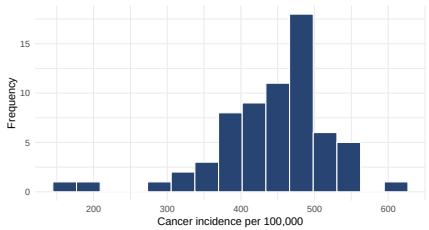
Data Sources

- **Cancer incidence** and **mortality** age-adjusted rates per 100,000 (2009-2018) were taken from the South Dakota Department of Health [South Dakota Cancer Registry, 2022].
- **Uninsured** percentage, **poverty** percentage, and **educational attainment** (percentage of adults with bachelors degree or higher) were collected by the American Community Survey via the US Census Bureau [American Community Survey, 2018c,
American Community Survey, 2018b,
American Community Survey, 2018a].
- **Obesity** percentage, **food index**, **smoking** percentage, and excessive **drinking** percentage were provided by the Counties Ranking and Roadmap site [County Health Rankings and Roadmaps, 2022].
- All code and full data set are available at https://github.com/espors/cancer_mortality_modeling.

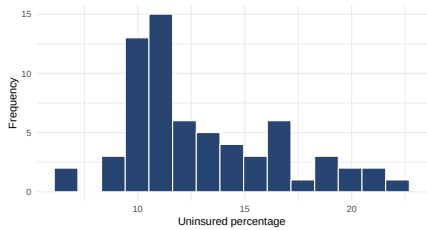
Exploratory Data Analysis I



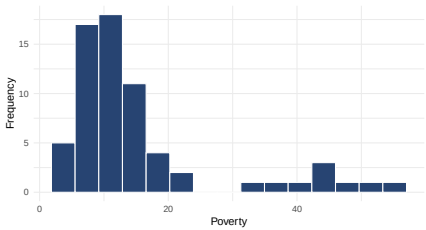
(a) Cancer mortality



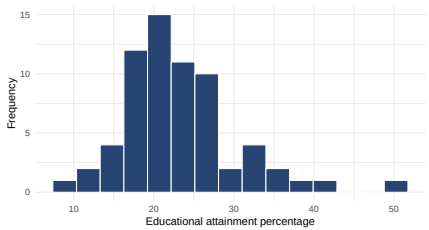
(b) Cancer incidence



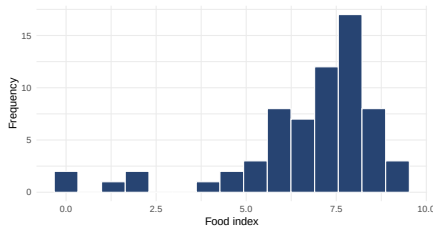
(c) Uninsured



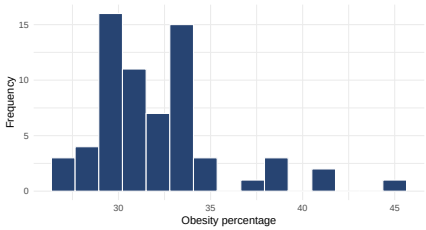
(d) Poverty percentage



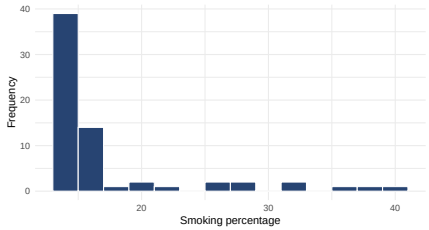
(e) Educational attainment



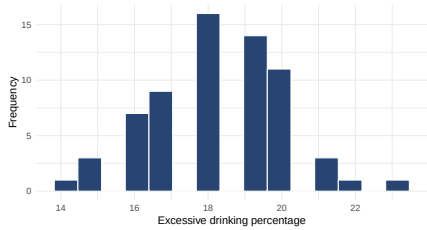
(f) Food index



(g) Obesity

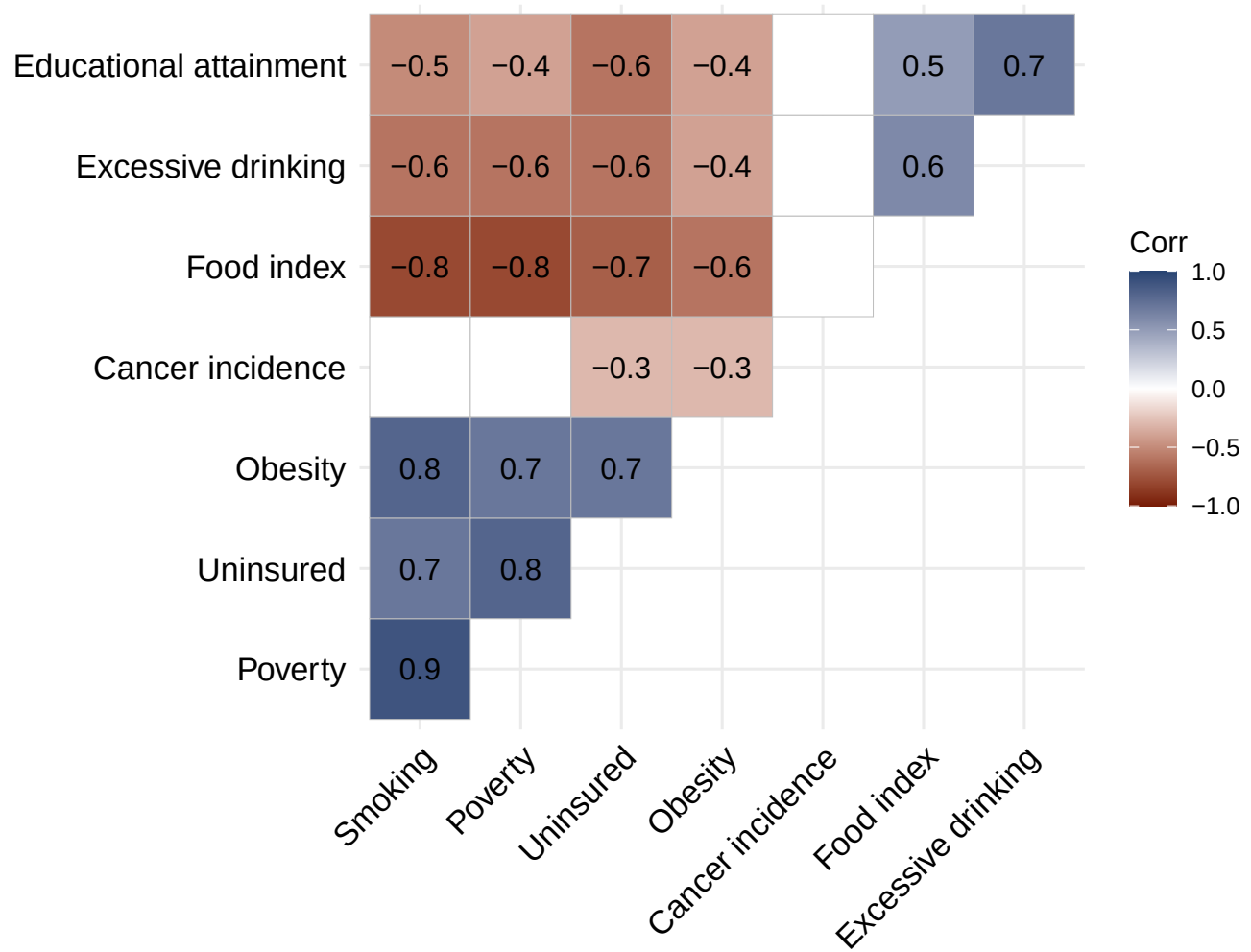


(h) Smoking



(i) Drinking

Exploratory Data Analysis II



Moran's I

- Moran's I was used to determine if the data was spatially uncorrelated, (preserve normality assumption)

$$I = \frac{N}{\mathbf{W}} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{1}$$

- $\mathbf{W}$ is sum of the weight matrix W
- $w_{ij} \in 0, 1$ indicates the value in W
- N is number of elements, x is response variable, $-1 \leq I \leq 1$

Multiple Linear Regression

- To gain insight into the selected socio-economic predictor variables impacted moralities in SD's counties, a traditional multiple linear regression model was considered.
  - The cancer mortality rates were considered as a representative sample for current and future rates, rather than a population statistic.

Multiple Linear Regression

- To gain insight into the selected socio-economic predictor variables impacted moralities in SD's counties, a traditional multiple linear regression model was considered.
  - The cancer mortality rates were considered as a representative sample for current and future rates, rather than a population statistic.
- In matrix notation, the estimated regression model has the form $\hat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{b}$ where $\boldsymbol{b}$ is found with

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

Variable Selection Techniques I

- **Stepwise regression** is a feature selection technique that iteratively adds and drops predictor variables in a succession of models and selects the model with predictors that has the smallest AIC value.

Variable Selection Techniques

- The following belong to a class of regression models referred to as "penalized regression models" because they add a penalty term to Equation 2.

Variable Selection Techniques

- The following belong to a class of regression models referred to as "penalized regression models" because they add a penalty term to Equation 2.
  - **Ridge regression** adds an 'L_2' penalty such that $\boldsymbol{b}$ is found with

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

Variable Selection Techniques

- The following belong to a class of regression models referred to as "penalized regression models" because they add a penalty term to Equation 2.
  - **Ridge regression** adds an 'L_2' penalty such that $\boldsymbol{b}$ is found with

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

  - **Lasso regression** (least operate shrinkage & selection operator) adds an 'L_1' penalty such that $\boldsymbol{b}$ is found with

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

Variable Selection Techniques

- **Elastic net regression** combines the best of ridge and lasso regression and $\boldsymbol{b}$ is found with

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

Variable Selection Techniques

- **Elastic net regression** combines the best of ridge and lasso regression and $\boldsymbol{b}$ is found with

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

or

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

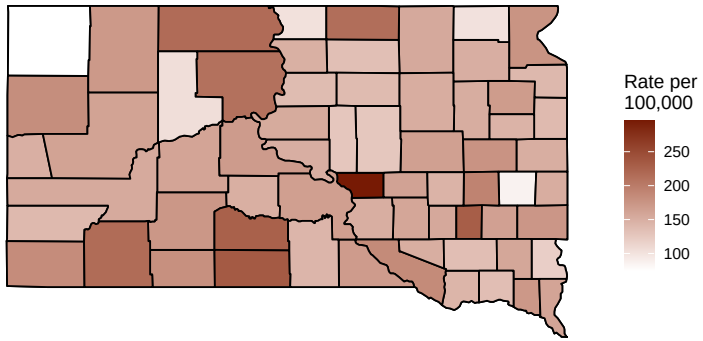
where $\alpha = \frac{\lambda_2}{\lambda_2 + \lambda_1} \in [0, 1)$.

- Predictions*

Moran's I

- Statistics calculated for Moran's I Mortality rate per 100,000.

Statistic	Value
Mean Response	160.4379
Moran's I	-0.0678
E[I]	-0.0152
Variance	0.0214
W	348
Z	-0.3584
p	0.3632



$$Z = \frac{E[I] - I}{\sqrt{v}} \approx -0.36 \implies p = 0.36 \geq 0.05 \tag{3}$$

- Fail to reject H_0 : no spatial correlation

Regression Models

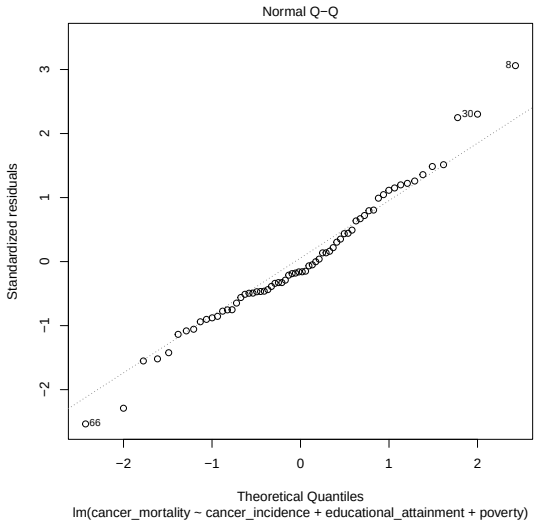
Summary of the regression models including the estimated coefficients, t-value, p-value, and adjusted R-squared.

	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, Mean VIF: 4.07			Adj. R-sqr: 0.546, Mean VIF: 2.30		

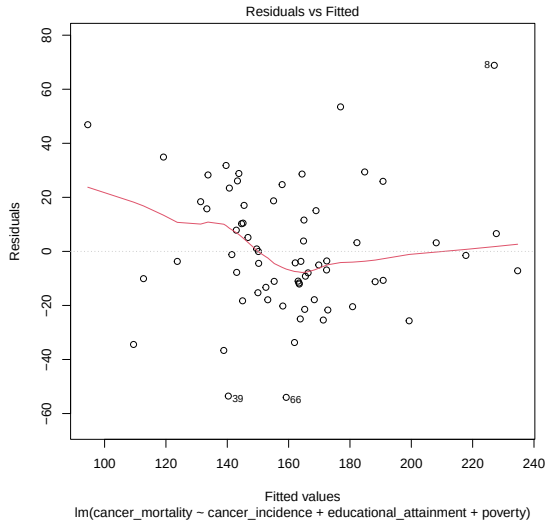
	Model 3 - Stepwise			**Model 4 - Penalized**		
Variables	Coeff	t-value	$\Pr(>\|t\|)$	Coeff	t-value	$\Pr(>\|t\|)$
Cancer Incidence (Rate)	0.204	5.205	$2.41 \cdot 10^{-6}$	0.210	5.298	$1.7 \cdot 10^{-6}$
Education Attainment (%)	−1.799	−3.130	0.002	−1.268	−2.492	0.015
Poverty (%)	1.716	5.673	$4.11 \cdot 10^{-7}$	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, Mean VIF: 1.71			Adj. R-sqr: 0.538, Mean VIF: 1.93		

Regression Models

	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, Mean VIF: 1.30		



(j) Q-Q Plot



(k) Residual vs Fitted

Discussion

- Feature selection techniques improved the mean VIF value, while maintaining a constant adjusted R-squared values.

Discussion

- Feature selection techniques improved the mean VIF value, while maintaining a constant adjusted R-squared values.
- Cancer incidence was positively associated with cancer mortality. In Model 5, for an additional case per 100,000, there was an additional 0.21 deaths per 100,000.

Discussion

- Feature selection techniques improved the mean VIF value, while maintaining a constant adjusted R-squared values.
- Cancer incidence was positively associated with cancer mortality. In Model 5, for an additional case per 100,000, there was an additional 0.21 deaths per 100,000.
- Poverty percentage was positively associated with cancer mortality. For an one percent increase in a county's poverty percentage, there was an additional 1.52 cancer-deaths per 100,000.
  - Persistent poverty is strongly associated with increased cancer mortality rates [Moss et al., 2020].
  - Eleven counties in SD are designated as experiencing persistent poverty.

Discussion

- Educational attainment was negatively associated with cancer mortality. For an additional percentage in educational attainment, there were a decrease of 1.31 deaths per 100,000.
  - There is less literature on education and cancer mortality, but one study did find a relationship between those who didn't finish high school and mortality [Barcelo et al., 2021].
  - There is emerging evidence on the relationship between education and general health [Zajacova and Lawrence, 2018].

Discussion

- Empowered with knowledge, health and government officials can make informed decisions on where to direct funding and education to counties that need them the most.

Discussion

- Empowered with knowledge, health and government officials can make informed decisions on where to direct funding and education to counties that need them the most.
- **Limitations:** Much of the data was based on self-reported surveys and the residuals indicate that not all of the linear assumptions may have been met.

Discussion

- Empowered with knowledge, health and government officials can make informed decisions on where to direct funding and education to counties that need them the most.
- **Limitations:** Much of the data was based on self-reported surveys and the residuals indicate that not all of the linear assumptions may have been met.
- **Future work:** The data analysis could be expanded to include non-parametric and clustering models.

References I


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).
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).
Small Area Health Insurance Estimates.
Retrieved from https://www.census.gov/programs-surveys/sahie.html.

References II


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).


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


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.

References III


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


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


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

References IV


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.