Comparison Between MLR and Lasso Regression in Predicting U.S. Cancer Death Rates by County

Group Members: Albert Li, Peize Zhang, Yue Zhang, Yukuan Zou, Fangzhou Yu

Course: STA302 Methods of Data Analysis I

Instructor: Yaoming Zhen

August 14, 2024

Table of contents

Introduction

In 2022, cancer was the second leading cause of death in the United States, accounting for 608,371 deaths or 142.3 deaths per 100,000 people, representing 18.5 percent of all deaths that year (Centers for Disease Control and Prevention, 2024). While overall cancer incidence and death rates have declined by about one-eighth and one-third from 1992 to 2021, decreasing cancer incidence and mortality remains a significant public health and economic interest (National Cancer Institute). This report aims to identify what demographic and social factors are related to cancer mortality, using and comparing lasso and multiple linear regression to identify the strength, type, and robustness of any relationships.

The data used is available on data. world (https://data.world/exercises/linear-regression-exercise-1) and compiles U.S. Census Bureau (census.gov) and the National Institutes of Health (cancer.gov, clinicaltrials.gov) data. It contains cancer death rates for 3,047 U.S. counties and county-equivalents, as well as 32 other demographic and socioeconomic variables providing extensive county-level information on educational attainment, poverty level, racial identity, healthcare coverage type (public and private), and more. However, we were left with 591 observations and 31 total variables after removing missing values and curating the variables.

We developed regression models using the training data for both the lasso and multiple linear regression approaches, subsequently applying these models to the test data. Our methodology incorporates STA302 techniques, including variable selection, cross-validation, data sorting, and model comparison procedures. Specifically, we evaluated the predictive ability of the models using MSE, while variable selection and reduction of significant multicollinearity were guided by VIF and AIC metrics. By comparing the MSE of the two models, we found that the multiple linear regression model exhibited superior performance.

The coefficients of the models varied, but the variables included were the same between models. We found that the rate of cancer incidence, adults over the age of 25 with a high school diploma, the unemployment rate for those aged 16 and over, and the proportion of a county's population identifying as Black influenced the prediction estimate for a county's cancer mortality rate, while other factors were excluded or statistically insignificant. For instance, we predict that a county with a high school graduate rate 10 percentage points higher than an otherwise similar county will have about 13 fewer cancer deaths per 100,000 people per annum. We also predict that a county with 100 additional cancer diagnoses will experience approximately 17 additional cancer deaths (See Figure 7, Results 1 for full results). Overall, we found the positive or negative influence of the predictor factors on cancer deaths unsurprising, and a clear positive numerical relationship between incidence and mortality is intuitive.

Due to the use of VIF to immediately rule out many variables, we believe that the impact of variables such as public or private health coverage was overlooked. We discuss this further in the conclusion of our paper (limitations). Notwithstanding this exclusion, the complexity of cancer and the limited ability of the data to capture potential significant factors such as local health spending and infrastructure restrain our chosen model's explanatory breadth. Additionally, the presence of missing values greatly reduced the number of observations used. Nevertheless, we hope to capture some meaningful insight to better inform policymakers for healthcare administration and funding.

Methodology

Generally, we try to compare two models and randomly divide the overall data into 80% training data and 20% testing data based on seed 1111 to ensure the accuracy of the final model.

1. Multiple linear regression model

1.1 Preparation and Assumptions

We try to build the multiple linear regression model, which includes variables such as 'incidencerate', 'medianage', 'pcths25_over', as predictors. For this model, we need to check 2 conditions first:

Condition 1: the conditional mean response is a single function of a linear combination of the predictors

Condition 2: the conditional mean of each predictor is a linear function with another predictor.

If one of the two conditions does not hold, then any pattern we might see in the residual plots can only tell us that an incorrect model has been fit. We could verify Condition 1 by plotting a scatterplot between target_death_rate and the predicted values, and Condition 2 by plotting a scatterplot of the relationship between all numerical predictors. If both conditions are met, we will proceed to construct residual plots to examine the fitted target_death_rate, numerical predictors, and their normal QQ plots to verify the four assumptions of linear regression. Specifically, if the data distribution in the residual plot is concentrated around 0, with no obvious clusters and evenly distributed (no tendency to spread), the assumptions of linearity, uncorrelatedness, and constant variance can be assumed to be satisfied. If the residuals are clustered around the standard line in the normal QQ plot, the residuals satisfy normality. If any of the above conditions are not met, we can apply specific transformations to the data to improve model performance. However, if all conditions are satisfied, we proceed to the model selection.

1.2 Model Selection

The first step in model selection is to check the multicollinearity of the full model. In our data, variables with VIF > 5 will be disregarded, and we use variables with VIF less than 5 as predictors. After the VIF screening, we will form a new simplified model, and based on this, we will use the backward AIC stepwise-chosen method to delete some variables to achieve smaller AIC, higher adjusted R-squared, and smaller BIC. We will comprehensively evaluate all metrics in these models and ultimately select the model that best fits our data.

1.3 Criteria to examine model and variables

After determining the preferred model, we will examine the influential points (leverage points and outliers). They will be examined through three thresholds: the Cook's Distance, DFFITS, and DFBETAS. We will focus primarily on the first two metrics. If there is no valid contextual reason to remove these problematic observations, we will keep them and proceed to the model validation. During the model validation process, we will apply the model developed from the training data to the test data and calculate the MSE to assess its predictive performance. Additionally, by plotting the observed response values against the predicted response values from the test data, we can visually evaluate the model's fit.

Specifically, if the performance is satisfactory, as indicated by a strong linear relationship in the plot, this would suggest that the model is robust and reliable as a predictor.

2. Lasso regression model

2.1 Preparation and Assumptions

Lasso regression is a penalized regression model assuming a linear relationship between dependent and independent variables. By adding penalties on coefficients, the lasso can shrink some coefficients to zero, helping to select more significant predictors and potentially providing a simpler model.

Our lasso regression model uses the same training and testing dataset as the one used in multiple linear regression which ensures that our subsequent comparison of the models is valid. Based on the fact that the dataset already meets the four assumptions verification for linear regression, we need two additional assumptions required for lasso regression:

i) Sparsity: only a subset of all covariates has non-zero coefficients, which implies lasso performs automatic variable selection.

ii) Feature scaling: To ensure each predictor is equally penalized, we standardized the data before fitting the model. The glmnet( ) function performs this automatically.

In our multiple linear regressions, variables with VIF larger than 5 were disregarded. Although lasso regression can mitigate the effect of multicollinearity, the choice of which variable to keep and which to shrink to zero can be unstable in cases of high multicollinearity. Thus, we only kept variables with VIF less than 5 in our training and testing dataset.

2.2 Model Construction

We first need to determine the optimal value of the regularization parameter lambda, which controls the degree of penalization applied to coefficients. Using the training dataset, we apply K-fold cross-validation to plot the relationship between the log value of lambda and mean square error. After choosing the optimal lambda, we could use glmnet() function to construct a lasso regression model to fit the training data and get the coefficients of each predictor. If a predictor's coefficient is zero, it indicates that the predictor is not significant, while the remaining predictors are considered effective.

After constructing the model, we should apply it to the test data. By plotting a graph with the observed and predicted y-values (the cancer death rate), we can evaluate the model's performance. If the data points concentrate closely around the diagonal line where the predicted and observed y-values are equal, it indicates that our Lasso model is a strong predictor and vice versa. Additionally, we will assess the mean squared error of the model in test data to determine if its performance is satisfactory.

After analyzing the two models separately, we compare their MSEs to conclude.

Results

The full model fitted with all 30 predictors using target_deathrate as the response variable satisfies Condition 1 by Figure 1, and most predictors satisfy Condition 2. The model also satisfied the linearity, constant variance, uncorrelated errors, and normality assumptions in Figure 2 and Figure 3. The right tail of the QQ plot has some data that deviates slightly from expectations, but most of the data follows a normal distribution. Thus, there is no need to transform the data.

Figure 1: y v.s. y_hat scatter plot

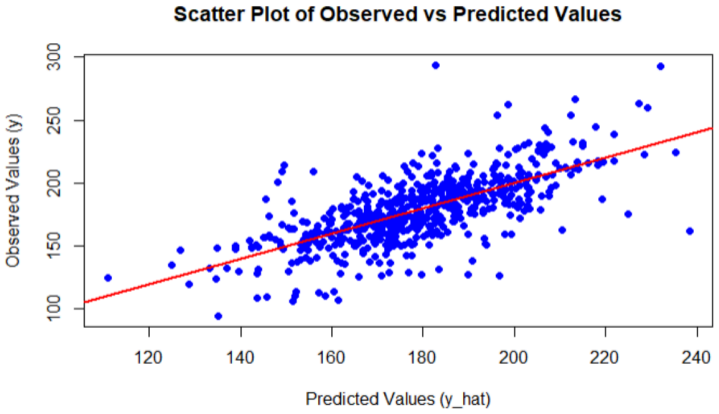


Figure 2: residual plot

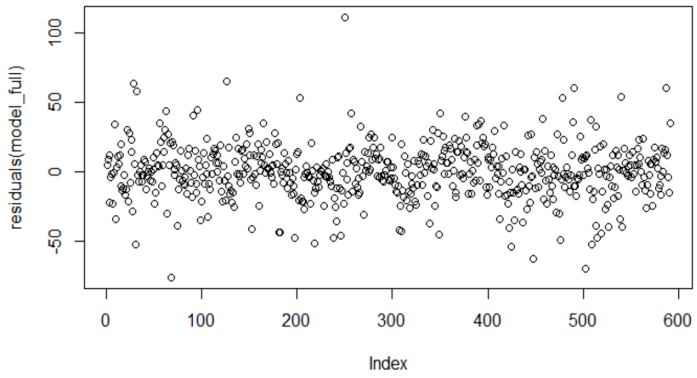
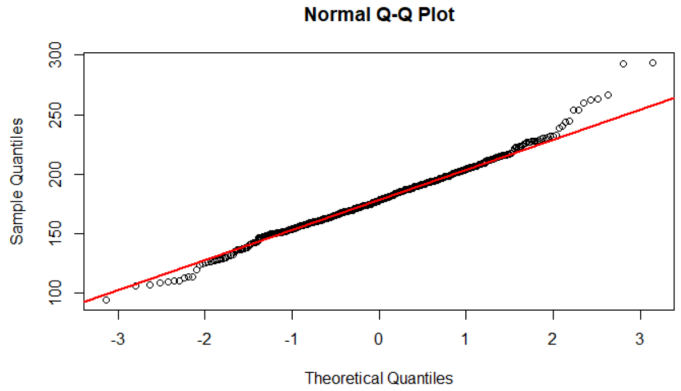


Figure 3: Q-Q plot



For both models, we select 9 predictors (Figure 4) with VIF less than 5 first to generate a new full model first, called Model 1.

Figure 4: Predictors with VIF less than 5

```
 [1] "incidencerate"      "1.24141194758994"  "studypercap"        "1.09612969438346"    "medianage"
 [6] "1.03599644712017"   "pcths25_over"      "4.10579915115973"   "pctunemployed16_over" "2.89591877094477"
[11] "pctblack"           "4.73220786532853"  "pctasian"           "2.30838735706969"    "pctotherrace"
[16] "1.66919154612755"   "birthrate"         "1.20475713210676"
```

1. Result for the Multiple Linear Regression Model

Based on our previous selection, we proceed with backward elimination on Model 1 to reduce the AIC. This results in a reduced model with a smaller AIC. The final reduced model does not violate any assumptions, as confirmed by Figure 5 and Figure 6.

Figure 5: Residual Plot of Reduced MLR Model

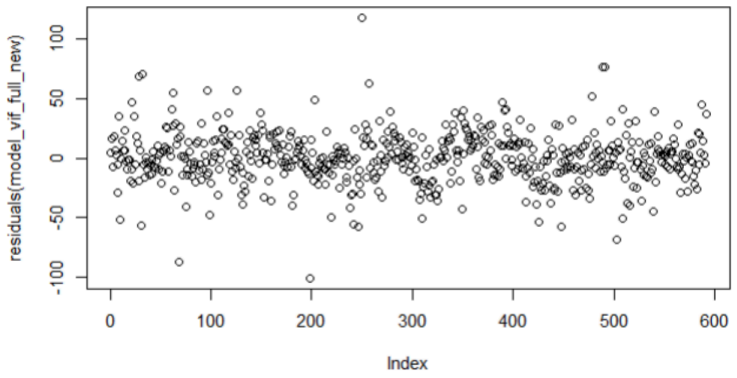
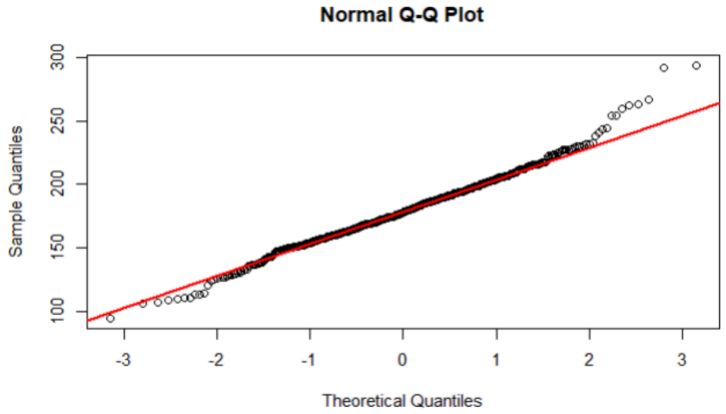


Figure 6: Q-Q Plot of Reduced MLR Model



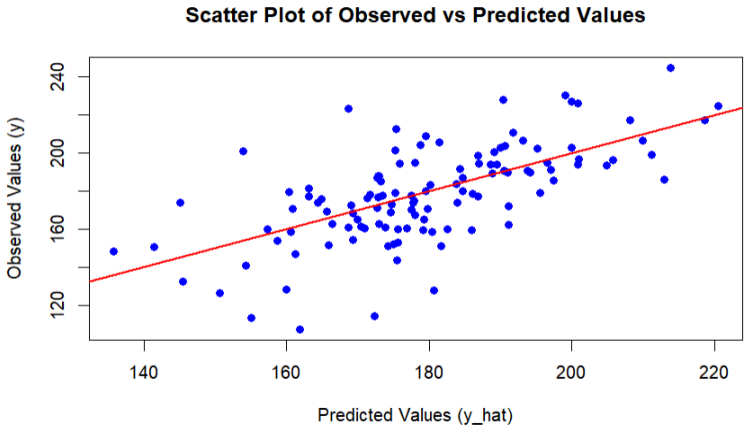
In the training set, we identified 32 leverage points, 3 outliers, and 26 influential points. For the testing set, there are 8 leverage points, 3 outliers, and 7 influential points. Based on our analysis, Cook's distance threshold is 0 for both sets, while the DFFITS is 0.2601 for the training set and 0.5208 for the testing set. Given these metrics, there are no compelling contextual reasons to remove the influential points. Figure 7 shows the coefficients of the multiple linear regression model on the training set. We found the 'medianage', 'pctasian', and 'pctotherrace' variables to be statistically insignificant at a 5% level. The detailed realistic interpretation is shown in Appendix 2.

Figure 7: Coefficients of the Multiple Linear Regression Model

```
Coefficients:
                        Estimate
(Intercept)            44.46021
incidencerate           0.16713
medianage              -0.02750
pcths25_over            1.28300
pctunemployed16_over    1.97807
pctblack                0.21197
pctasian               -0.93245
pctotherrace           -0.53336
```

Additionally, the MSE of the training model for test data is 366.2165. By Figure 8, since the data as a whole presents a linear relationship and is closer to the red line but not perfectly, the model is moderately strong as a predictor.

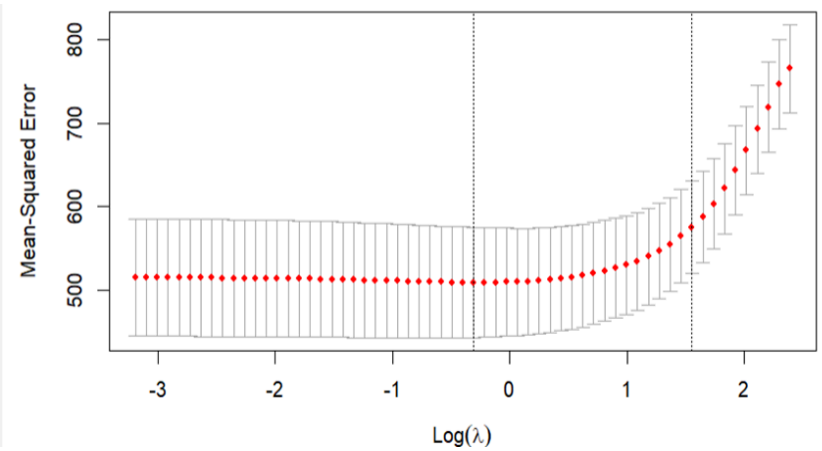
Figure 8: Observed y_test v.s. Predicted y_test



2. Result for Lasso regression model

For Lasso regression, we started with the 9 selected predictors. To optimize our lasso regression model for prediction, we will choose the lambda corresponding to the minimum mean cross-validated error. We used 10-fold cross-validation to plot. As Figure 9 shows, this is a convex curvature graph. By coding, our best lambda equals 0.7353472.

Figure 9: Mean Square Error v.s. Log-Lambda



Then as Figure 10 shows, we used the predict() function to get y_hat in both training and testing data.

Figure 10: Lasso Regression Model

```
#lasso.mod
lasso.mod = glmnet(x_train, y_train, family = 'gaussian', alpha = 1, lambda = bestlam) #lasso model
train_pred = predict(lasso.mod, newx = x_train, s = bestlam) #predicted y_train
lasso.pred = predict(lasso.mod, s = bestlam, newx = x_test) #predicted y_test
lasso.coef = predict(lasso.mod, type = "coefficients", s = bestlam) #coefficients
```

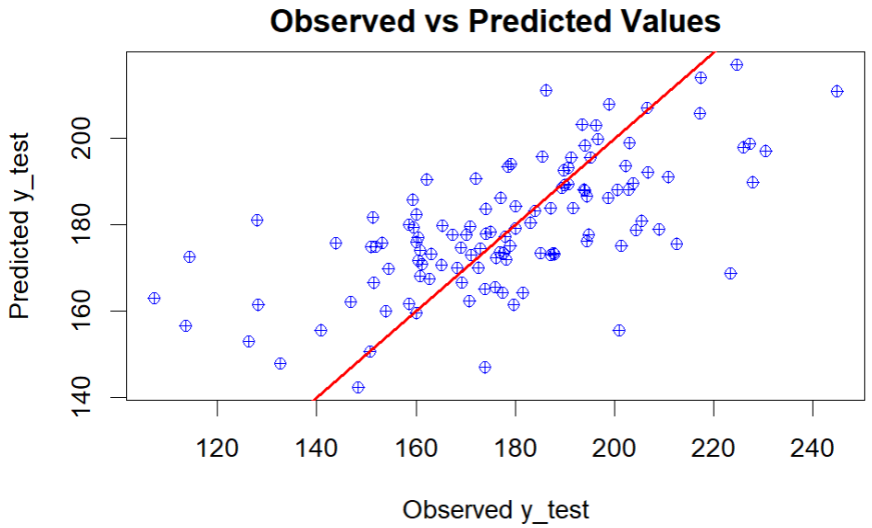
Figure 11 shows the coefficients of our lasso regression model. The model finally selects 7 predictors, which are 'incidencerate', 'medianage', 'pcths25_over', 'pctunemployed16_over', 'pctblack', 'pctasian', and 'pctotherrace'. Note that predictors 'studypercap' and 'birthrate' have zero coefficients. This means these two predictors are not important. So our lasso regression will eventually have 7 effective predictors.

Figure 11: Coefficients of the Lasso Regression Model

	s1
(Intercept)	51.74345830
incidencerate	0.15518343
studypercap	.
medianage	-0.00822254
pcths25_over	1.23855363
pctunemployed16_over	1.78322324
pctblack	0.18149491
pctasian	-0.63509629
pctotherrace	-0.40908717
birthrate	.

To evaluate the accuracy of model prediction, we plot the graph of observed and predicted y-values (the cancer death rate) in the test dataset, as Figure 12 shows. The red line is a 45-degree diagonal line that equates predicted and observed y-values. As shown in the figure, since most points are closer to the red line but not perfectly, this indicates that our lasso model is moderately strong as a predictor. Additionally, the MSE of the test model is 372.7964. Since this is higher than our multiple linear regression model, we did not interpret these regression estimates in our report.

Figure 12: Predicted y_test v.s. Observed y_test



Conclusion

1. **Comparison of MLR and Lasso Regression Models**

   Based on our analysis, the multiple linear regression (MLR) model demonstrated superior performance. The MSE of test data for the MLR model was 366.2165, notably lower than the 372.7964 observed for the Lasso regression model. This significant difference in MSE indicates that the MLR model provides more accurate predictions on our test dataset, perhaps due to the lasso regression underfitting the data. Consequently, we have determined that the MLR model is a more appropriate and reliable choice for analysis of overall results, as it better captures the underlying patterns and relationships within the data.

2. **Suggestions for Study Improvements**

   Our original dataset comprised 3,047 entries from U.S. counties, but after removing entries with missing values and irrelevant variables, only 591 observations remained. This substantial reduction may introduce bias, as the remaining dataset might not accurately represent U.S. counties as a whole. For instance, counties with missing data may be characteristically different, perhaps due to lower administrative capacity, potentially skewing the analysis towards counties with more complete datasets. Future studies should consider using imputation methods to handle missing data or acquiring more comprehensive data sources to ensure a representative sample.

   For MLR, although most predictors satisfied the assumption conditions, there are still a few predictors that did not fully satisfy Condition 2, which could impact the reliability of residual plots and model validity. The model's reliance on linearity and assumption checks means deviations could lead to suboptimal predictions. Though it showed strong predictive capability, the absence of regularization heightens the risk of overfitting, particularly given the small number of predictors, potentially reducing its effectiveness on new datasets.

   For Lasso, the model struggled with performance due to its strong penalization of coefficients. This resulted in an underfitting issue, where significant predictors were penalized too heavily, leading to a less effective model. Consequently, the multiple linear regression model performed better on our dataset, as it allowed for more flexibility in capturing complex relationships without imposing penalties.

   Given that cancer was the second leading cause of death in the United States in 2022, accounting for 18.5% of all deaths (Centers for Disease Control and Prevention, 2024), there is significant public health interest in employing advanced modeling techniques to better understand and address these dynamics. Policymakers should consider reallocating resources and adjusting policies to reduce cancer mortality in counties with higher Black populations, lower high school graduation rates, and higher unemployment rates. Methods such as Generalized Additive Models (GAM), Ridge regression, Principal Component Regression (PCR), and Partial Least Squares (PLS) regression can enhance analysis by addressing non-linearity, high dimensionality, and multicollinearity. These advanced approaches not only improve analytical rigor but also provide clearer insights into the factors influencing cancer mortality, enabling more effective and targeted interventions

within the context of US healthcare spending, which averaged 16-17% of GDP from 2009 to 2022.

3. Summary of Research Question and Further Questions

Based on Appendix 1, our model indicates that cancer mortality is significantly influenced by factors such as cancer diagnosis incidence, average age of the population, education level, unemployment rate, and racial composition. Specifically, higher cancer incidence rates, an older population, and elevated unemployment rates are associated with increased cancer mortality. Additionally, the racial composition of the population plays a crucial role in these outcomes.

Further research is warranted to explore several areas. First, examining the impact of socioeconomic factors, such as median income and poverty rates, on cancer mortality could provide valuable insights, especially when incorporating alternative models or investigating interaction effects with variables such as healthcare coverage. Second, a more detailed analysis of healthcare coverage — differentiating between private and public insurance — using methods such as hierarchical modeling or exploring interaction terms may yield additional insights. Third, investigating the relationship between educational attainment (e.g., high school diplomas and bachelor's degrees) and cancer outcomes, particularly in conjunction with other demographic factors, could offer new perspectives. Lastly, analyzing racial demographics (e.g., percentages of white, black, and Asian populations) alongside socioeconomic and healthcare access factors may reveal significant health disparities among different racial groups.

To address cancer mortality effectively, it is essential to focus on increased early screening, enhanced care for the elderly, and expanded education and employment opportunities. Additionally, ensuring equitable distribution of healthcare resources and addressing health disparities among different ethnic groups are critical steps toward improving overall cancer outcomes.

Acknowledgment

Albert Li: Introduction, results interpretation, and significance

Peize Zhang: Coding, methodology, results of lasso regression model

Yukuan Zou: Coding, results of multiple linear regression, integration of overall results

Yue Zhang: Methodology of multiple linear regression, integration of overall methodology

Fangzhou Yu: Model comparison and limitation discussion, further research questions

Appendix 1: Table of Variable Interpretations

Variable name	Interpretation
avganncount	Mean number of reported cases of cancer diagnosed annually (a)
avgdeathsperyear	Mean number of reported mortalities due to cancer (a)
target_deathrate	**Dependent variable. Mean per capita (100,000) cancer mortalities (a)**
incidencerate	Mean per capita (100,000) cancer diagnoses (a)
medincome	Median income per county (b)
popest2015	Population of county (b)
Population of county (b)	Percent of populace in poverty (b)
studypercap	Per capita number of cancer-related clinical trials per county (a)
binnedinc	Median income per capita binned by decile (b)
medianage	Median age of county residents (b)
medianagemale	Median age of male county residents (b)
medianagefemale	Median age of female county residents (b)
geography	County name (b)
percentmarried	Percent of county residents who are married (b)
pctnohs18_24	Percent of county residents ages 18-24 highest education attained: less than high school (b)
pcths18_24	Percent of county residents ages 18-24 highest education attained: high school diploma (b)
pctsomecol18_24	Percent of county residents ages 18-24 highest education attained: some college (b)
pctbachdeg18_24	Percent of county residents ages 18-24 highest education attained: bachelor's degree (b)
pcths25_over	Percent of county residents ages 25 and over highest education attained: high school diploma (b)
pctbachdeg25_over	Percent of county residents ages 25 and over highest education attained: bachelor's degree (b)
pctemployed16_over	Percent of county residents ages 16 and over employed (b)
pctunemployed16_over	Percent of county residents ages 16 and over unemployed (b)
pctprivatecoverage	Percent of county residents with private health coverage (b)
pctprivatecoveragealone	Percent of county residents with private health coverage alone (no public assistance) (b)
pctempprivcoverage	Percent of county residents with employee-provided private health coverage (b)
pctpubliccoverage	Percent of county residents with government-provided health coverage (b)
pctpubliccoveragealone	Percent of county residents with government-provided health coverage alone (b)
pctwhite	Percent of county residents who identify as White (b)
pctblack	Percent of county residents who identify as Black (b)
pctasian	Percent of county residents who identify as Asian (b)
pctotherrace	Percent of county residents who identify in a category which is not White, Black, or Asian (b)
pctmarriedhouseholds	Percent of married households (b)
birthrate	Number of live births relative to number of women in county (b)

Note: Bolded variable is the predicted variable, other grey variables are predictors used in both the MLR and lasso models. The variables in white are excluded from both regression models.

Appendix 2: Estimated Coefficients of the training set of multiple linear regression model

```
Coefficients:
                      Estimate
(Intercept)           44.46021
incidencerate          0.16713
medianage             -0.02750
pcths25_over           1.28300
pctunemployed16_over   1.97807
pctblack               0.21197
pctasian              -0.93245
pctotherrace          -0.53336
```

While the intercept has no realistic interpretation, we can interpret the statistically significant estimated relationships as follows:

1. A county with 100 more cancer diagnoses than an otherwise similar county will experience an estimated 16-17 additional deaths from cancer.
2. A county with a 10 percentage point higher high school graduation rate among those 25 years old and over will experience an estimated 13 fewer cancer deaths per 100,000 people per annum (all other predictor variables equal).
3. A county with a 10 percentage point higher unemployment rate among those 16 years old and over will experience an estimated 20 more cancer deaths per 100,000 people per annum.
4. A county with a 10 percentage point proportion of its population identifying as Black will experience an estimated 2 additional cancer deaths per 100,000 per annum.

These results seem largely intuitive (although not necessarily their degree), as it makes sense that counties with lower unemployment rates likely have lower rates of healthcare coverage or fewer economic resources to allocate to local health services. The absence of variables such as the proportion of a county's population with any health insurance coverage is likely due to the variable selection process immediately excluding variables with high risk of multicollinearity, rather than iteratively removing those with significant risk.

References

1. Centers for Disease Control and Prevention. (2024, May 2). *Deaths and Mortality*. https://www.cdc.gov/nchs/fastats/deaths.htm.
2. Gunja, M. Z., Gumas, E. D., & Williams II, R. D. (2023, January 31). *U.S. Health Care from a Global Perspective, 2022: Accelerating Spending, Worsening Outcomes*. Global Perspective on U.S. Health Care | Commonwealth Fund. https://www.commonwealthfund.org/publications/issue-briefs/2023/jan/us-health-care-global-perspective-2022.
3. National Cancer Institute. (n.d.). *Cancer Stat Facts: Cancer of Any Site*. SEER. https://seer.cancer.gov/statfacts/html/all.html.