final

Introduction

Cancer has a large impact on society, affecting people from all different walks of life across the United States. In this report, we created a regression model that can be used to predict cancer mortality rate for counties across the U.S. Our motivation for pursuing this project stems from a common interest in Healthcare and a curiosity about the types of factors that are related to cancer death rates.

The specific research objective is twofold: (1) Identify key characteristics that are associated with cancer death rates, and (2) Build a model to predict the cancer death rate in each county, normalized according to population. The dataset we used was aggregated from various U.S. governmental sources, including census.gov, clinicaltrials.gov, and cancer.gov[a]*nnotation* , and we used additional geographic data from the U.S. Department of Agriculture[a]*nnotation* and census.gov[a]*nnotation*.

Data Description

To predict our dependent variable *TARGET_deathRate* (the mean per capita cancer mortalities), We grouped each of our independent variables into seven main categories:

1. Cancer-related Demographics - These include *avgAnnCount*, *avgDeathsPerYear*, *incidenceRate*, and *studyPerCap*.
2. General Demographics - These include *MedianAge*, *popEst2015*, *MedianAgeFemale*, and *BirthRate*.
3. Racial Demographics - These include *PctWhite*, *PctBlack*, and *PctOtherRace*.
4. Education and Employment Demographics - These include *PctBachDeg18-24*, *PctHS25_Over*, and *PctBachDeg25_Over*.
5. Insurance Coverage Demographics - These include *PctPrivateCoverage* and *PctEmpPrivCoverage*.
6. Income and Household Demographics - These include *medIncome*, *povertyPercent*, *AvgHouseholdSize*, and *PctMarried*.
7. Geographic Features: These include *Division* and one hot encoded state variables.

Final Model

Our final model ends up regressing on 78 variables. This number includes the variables described in the data description section as well as the dummy variables generated by our categorical variables and interaction terms. The model's adjusted R^2 value is 0.584 which was approved from approximately 0.54 in our initial modelling.

Discussion and Future Improvements

< insert section here>

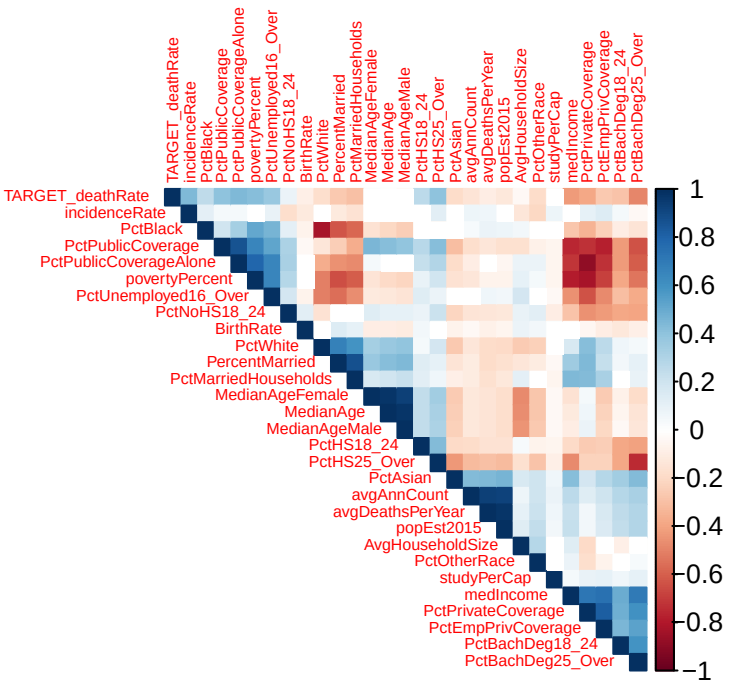
Additional Work

Data Cleaning and Exploratory Data Analysis

We started the EDA process by checking to see if there are any columns that contained substantial missing values. We removed columns *PctSomeCol18_24* (2285 N/A), *PctEmployed16_Over* (152 N/A), *PctPrivateCoverageAlone* (609 N/A). We then took out binnedInc as once we changed it into a numeric vector and taking means of every row's lower and upper decile, we decided that it would not be appropriate as it categorizes income into 10 splits and provides similar information to medIncome.

Two important issues that we would need to address in order for our model to yield substantial information would be linearity, in which there must be a linear relationship between the independent and dependent variables, and multicollinearity, where our independent variables are too highly correlated with one another. We have addressed the former by interpreting the model diagnostics during the discussion portion, and we will be dealing with multicollinearity in this section.

Since we have multiple variables for regression, we would want to detect multicollinearity within our regressors to avoid. Having multicollinearity affects the variance of our model's prediction, which reduces the quality of interpreting our independent variables. We tackle this issue through two means: construction of a correlation plot and removing variables that yield a relatively high variance inflation factor (VIF). We used a cutoff VIF of $10^annotation$to remove 8 features from our dataset: *avgDeathsPerYear*, *popEst2015*, *MedianAgeFemale*, *MedianAgeMale*, *MedianAge*, *PctPrivateCoverage*, *PctPublicCoverage*, and *PctPublicCoverageAlone*.



Model Selection

Because our model is focused on prediction rather than causal inference, we decided to undergo a rigorous variable selection process. After removing variables deemed too severely multicollinear, we're left with a "full" model consisting of 21 explanatory variables.

The reason we chose to screen our variables with VIF beforehand is that removing explanatory variables that are collinear not only helps with the assumptions of linear regression, but also helps computationally, as we have less variables to search through when searching for the best model.

Screening With LASSO

Before performing best subsets regression, we decided to run LASSO on our model in order to get a sense of variable importance in a predictive context.

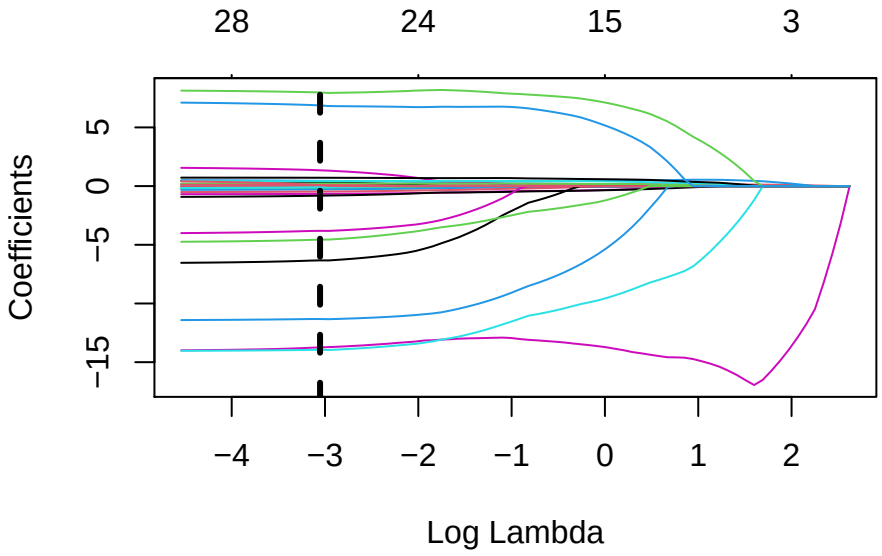


Figure 1: Lasso coefficient trails. The dotted line marks the optimal lambda.

Taking a look at the results of LASSO, we see that none of our coefficients have been zeroed out, meaning that we will need to take a look at other variable selection methods if we want to shrink our model. As a result, we explore a different method of model shrinkage: best subsets regression.

Best Subsets Regression

Best subsets regression exhaustively searches every combination of variables for every possible model size and selects the best models for each model size according to different criteria. The criteria we considered were Adjusted R^2, Mallow's Cp, and BIC. We chose these three criteria since they're supported by the R function regsubsets, and we wanted to use the same library for the sake of consistency in model selection.

	Adjusted R^2	Mallow's Cp	BIC
Number of Variables	23	21	16

Table 1: The number of variables in the "best" model as chosen by various criteria. Note that the number of variables has increased due to dummy variables being added to the model.

As seen in **Table 1**, Adjusted R^2 as our criterion resulted in the largest model, while BIC as our criterion resulted in the smallest model. Taking a closer look at the actual models that were selected, we see that some of our dummy variables for our only categorical variable, region, ended up being dropped by best subset regression. Because it's not possible to write a formula that drops some of these dummy variables as well as the fact that the majority of dummy variables were kept for all 3 models, we chose to keep Division in all 3 of our models even if some of the dummy variables ended being dropped. This isn't too consequential as in the Adjusted R^2 model and the Mallow's Cp Model, only the dummy variable associated with the South

Atlantic division is dropped, while in the BIC model, only the dummy variables associated with the South Atlantic division and New England division are dropped.

Cross Validation

After creating our models (2 distinct ones in this case), it's clear the the criteria don't agree on which model is the best. In order to assess the performance of our models, we need to evaluate the predictive ability of our models on data they have never seen before. Rather than using a train-test split of our data, we decided to use cross validation since cross validation tends to smooth out noise or randomness, and also provides more precision while reducing bias as we have more data for fitting the models. Leave-one-out CV is too computationally expensive due to the large number of rows, so we went with k-fold CV instead with a fold size of 10. We also computed the MSE from CV for the full model as well as the LASSO model to serve as comparisons.

	Full Model	LASSO Model	Adjusted R^2 Model	Mallow's Cp Model	BIC Model
MSE	359	359	358	358	359

Table 2: The MSE from k-fold CV of our various models. Note that the MSE of our full model and LASSO model are the same since the two models are the same (albeit it's definitely possible for two different models to have the same MSE).

As seen in **Table 2**, the models with the lowest MSE ended up being our Mallow's Cp and Adjuted R^2 models. Because the Mallow's Cp model is smaller (2 fewer features), we'll choose that model as our "final" model for this step.

Model Diagnostics

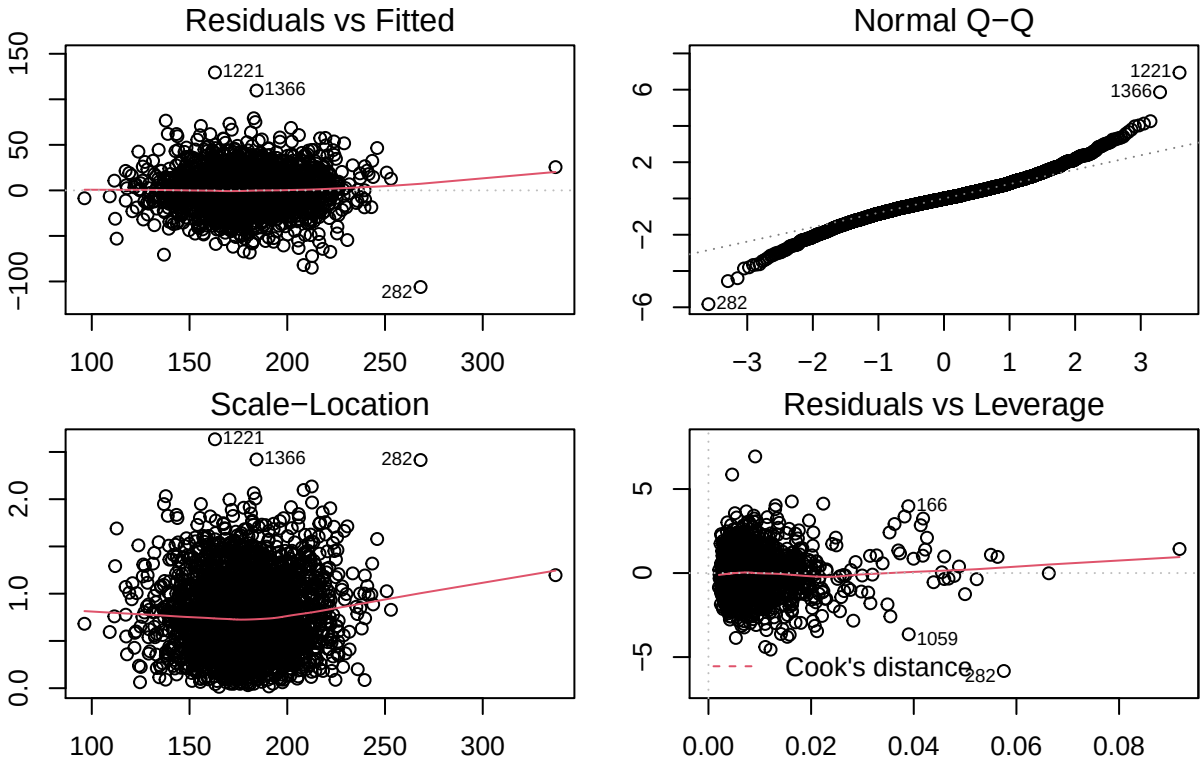


Figure 2: Diagnostic plots of the chosen model.

We notice an outlier in our residual plots in **Figure 2** that reveal a point with a somewhat high leverage. After investgating the possibility of an encoding error, we discovered that this point belonged to Union County, Florida which is known to have a disproportionately high cancer death rate compared to the rest of the United States, so we left that data point in. Something that was concerning during EDA was that a few of our explanatory variables didn't have normal distributions. When we applied a Box-Cox transformation, our model performance actually slightly decreased with a lower R^2 in our model as well as a higher MSE during cross validation. As a result, we decided to not pursue a transformation of our variables prior to variable selection. After variable selection, applying a Box-Tidwell transformation was considered, but due to powers being pushed to infinity and being unable to diagnose this issue as Box-Tidwell wasn't covered in class, we decided to not continue pursuing this particular transformation. **Figure 2** reveals that the assumptions of linear regression are mostly followed anyways, so a transformation wouldn't necessarily create a huge improvement.

Adding States to the Model

While we have region as one of the variables in our model, it's possible that certain states may go against the trend of the region. As a result, we will consider adding states as variables to our model. Doing so will allow the coefficient of a state to "counteract" the coefficient of its region in the event that a state is significantly different than its region. In order to decide which states to add to our model, we will use forward selection using AIC and BIC as our criteria. We chose to do forward selection rather than best subsets regression here due to the large number of additional columns we have added via one hot encoding the state variable. We chose AIC and BIC as our criteria since they're supported by the step function and we want to use the same library for the sake of consistency during model selection.

The results of our forward selection reveal that adding states does in fact add precision to our model - AIC adds 16 states to our model and BIC adds 7 states to our model. AIC adds significantly more variables than BIC, though that's not surprising considering that BIC penalizes model complexity more heavily. The states chosen by both BIC and AIC tend to be in the Southern and Midwest regions of the United States, perhaps revealing that these regions contain many outlier states. In order to determine which model fits the data better, we again ran k-fold cross validation (with a fold size of 10) on these two models and because the AIC model had a lower MSE, we chose the AIC model as our "final" model for this stage of the model selection process. Our model now has 38 total variables (counting the dummy variables for Division as separate variables) with the addition of the 16 state variables.