Forecasting Cancer Death Rates Across the United States

Kendall Kikkawa, Jonathan Luo, Andre Sha

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 , and we used additional geographic data from the U.S. Department of Agriculture and census.gov. **Figure 1** shows the distribution of cancer death rates across the U.S by state. Through building a regression model and analyzing the predictions of our model, perhaps we can glean some insight into the characteristics of cancer mortality rates in the United States.

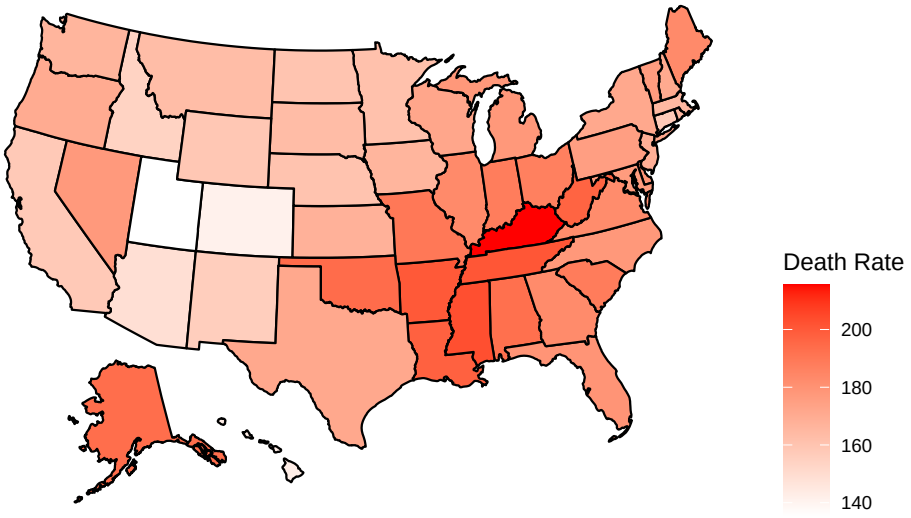

Figure 1: Distribution of mean per capita (100,000) cancer mortalities

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 (a full description of the raw features is listed in **Appendix A**:

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, dummy variables generated by our state and division levels, and interaction terms. The model's adjusted R^2 value is 0.584 which is an improvement from approximately 0.54 in our full data model after EDA.

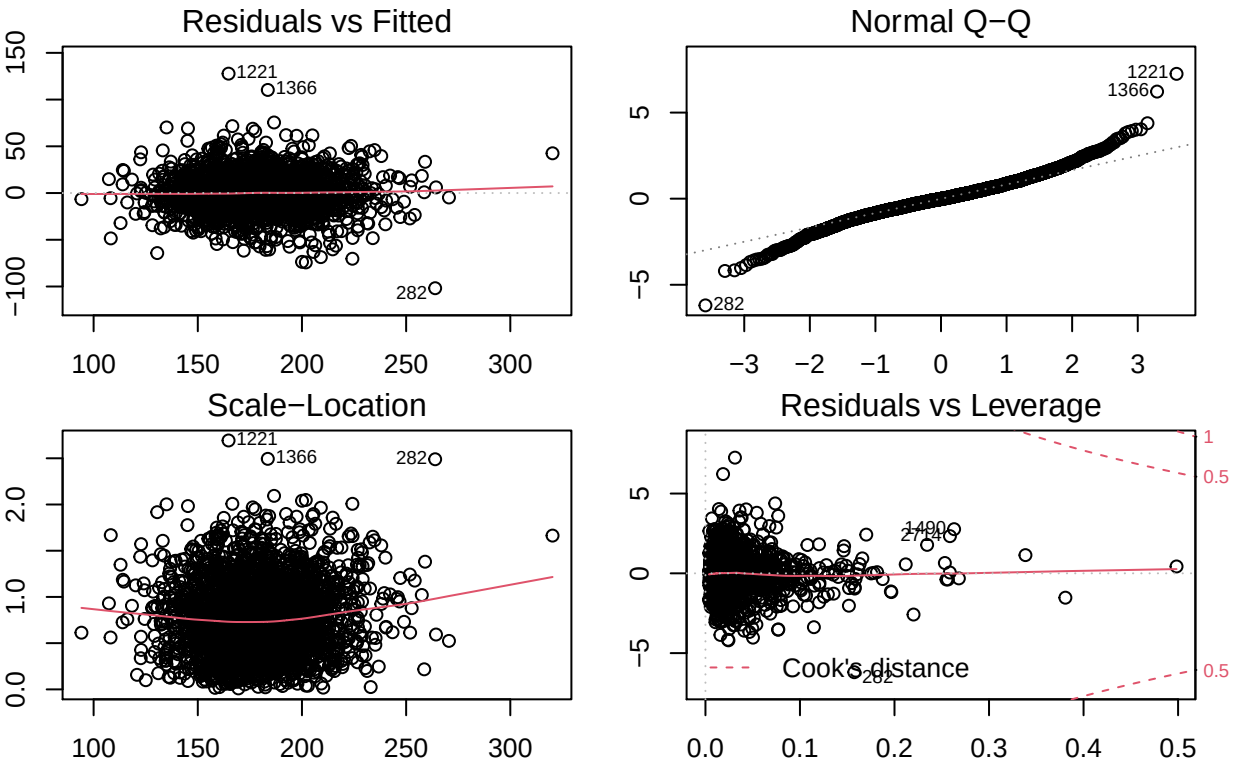

Figure 2: Model diagnostics plots of the final model.

The assumptions of linear regression are largely met as well. As seen in **Figure 2**, the residuals show no sign of patterns when plotted against the fitted values indicating that our data is fairly linear. The Normal Q-Q plot is almost linear as well, although there are slight deviations at the ends. This indicates that the model follows the normal distribution fairly well. The scale-location plot shows that the residuals are spread fairly equally along the fitted values which means that our final model has a relatively constant error variance. The residuals vs leverage plot reveals that while there are a few points with high leverage, no points exceed the Cook's distance boundary as demarked by the red dotted line. This leads us to interpret the data as not having any points that are too out of the ordinary

Discussion and Future Improvements

If we were to widen our project scope, we would like to further develop this model by fitting it to specific types of cancer. In those types of scenarios, perhaps we can analyze common statistically significant variables and see their association with the cancer's death rate for a county. Another idea we briefly discussed would be to work with government institutions to see if we can collect more data for feature engineering. For example, say if we were somehow able to aggregate another set of demographic features at the county level, how might we be able to fit them into our model, and account for interactions with existing variables. Since our dataset was based on aggregated data from 2010-2015 estimates, working with future data to forecast county death rates for counties would be serve as an excellent addendum to our project. With the addition

of data over time, we would be able to build time series models which would perhaps be more suited at predicting cancer death rates in the future.

While our model was built for the purpose of prediction, this model can serve as a stepping stone into causal inference. Future studies could be done using some of the coefficients we deemed important as treatment variables and perhaps further investigating these factors in a controlled environment as opposed to the observational data our dataset is comprised of.

Conclusion

Our goal with this analysis was to find the best possible model for predicting the cancer death rate of a county given various characteristics of that county. Prediction of cancer death rates is important for several reasons. If a county's actual cancer death rate is significantly different from its predicted cancer death rate, then that county may need to be investigated. This difference between predicted and actual death rate could stem from anything including a drastic change in county characteristics, the rise of a new explanatory variable that has a significant influence on cancer death rates, or perhaps errors when reporting cause of death. No matter the reason, and whether the predicted death rate is higher or lower than the actual death rate, a significant deviation from the prediction is cause for investigation. In addition, our model can serve as an initial tool to inform government institutions about the need to brainstorm and develop ideas into establishing support systems for regions who have high predicted death rates.

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 (**Figure 3**) and removing variables that yield a relatively high variance inflation factor (VIF). We used a cutoff VIF of $10^1$to remove 8 features from our dataset: `avgDeathsPerYear`, `popEst2015`, `MedianAgeFemale`, `MedianAgeMale`, `MedianAge`, `PctPrivateCoverage`, `PctPublicCoverage`, and `PctPublicCoverageAlone`.

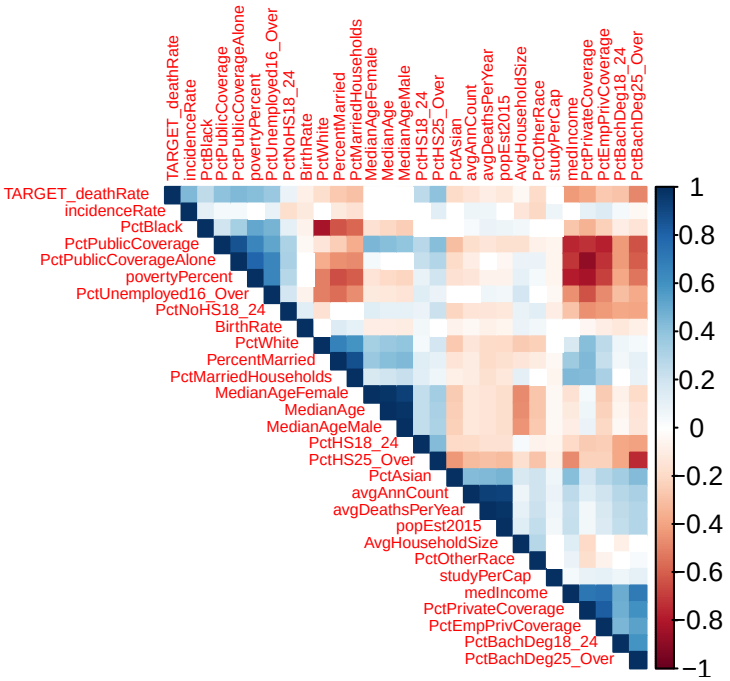

Figure 3: Correlation plot of numeric features.

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.

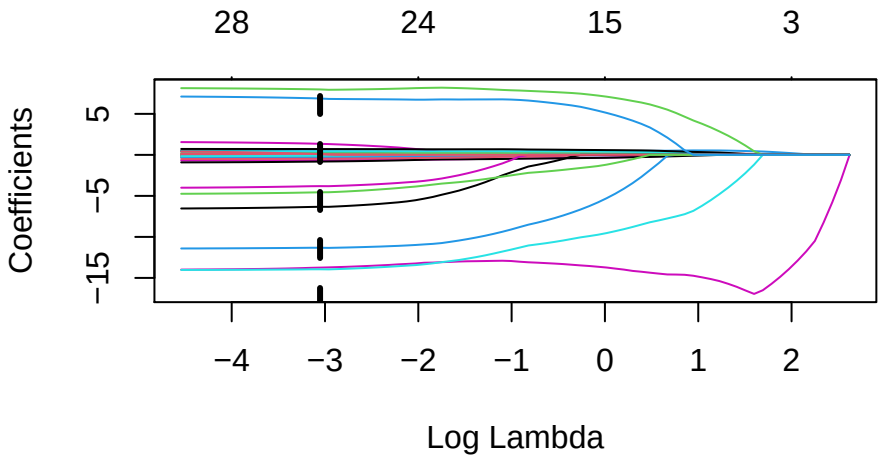

Figure 4: Lasso coefficient trails. The dotted line marks the optimal lambda.

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

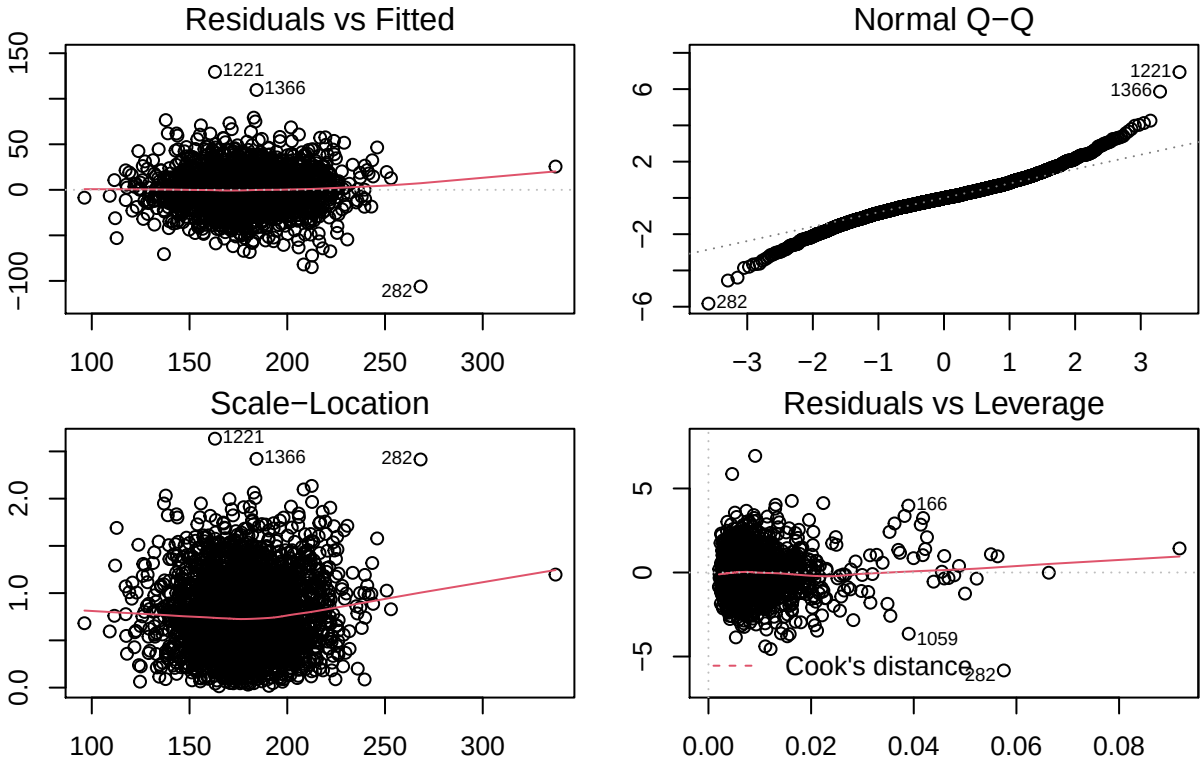

Figure 5: Diagnostic plots of the chosen model.

We notice an outlier in our residual plots in **Figure 5** 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 5** 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.

Interaction Selection

Because our model selection process yielded a design matrix with 38 features, it was computationally infeasible to assess the presence of 2^38 possible interactions. We hypothesized that the most informative, and likely most interpretable, interactions occur between a continuous value and a certain division of the U.S. Divisions are a finer categorization of region, broken down in **Appendix B**:

We acknowledge that there may be some state-level interactions as well, but again we decided it was infeasible to test all of these interactions. However, the forward selection above should identify "outlier" states, so the combination of division-level interactions and state-specific coefficients should capture state-level characteristics in our model. **Figure 6** displays two interactions that we assessed.

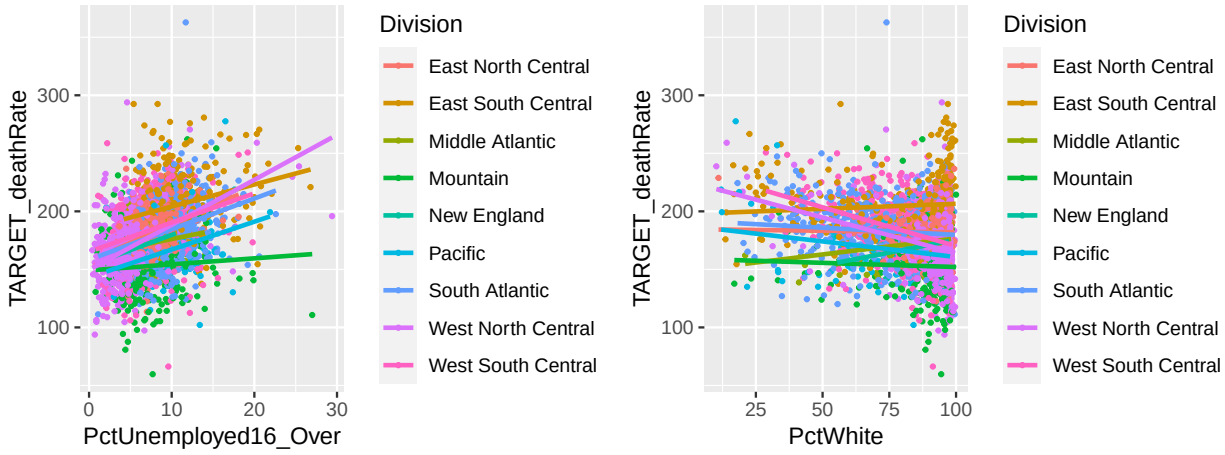

Figure 6: Division-level interactions.

After plotting all division-level interactions, we determined that there was visual evidence of 11 possible interactions: To ultimately decide whether or not these interactions improved our model, we used F-tests comparing the full model with all interactions to the model with each interaction removed. Because we conducted 11 tests, we applied two different correction factors to control the family wise error rate at $\alpha = 0.05$: Bonferroni and Benjamini-Hochberg. We then removed any interactions that were not significant at the corrected significance level.

The two methods yielded different results for significant interactions, so we again used 10-fold cross validation to determine which correction method produced a better model. Benjamini-Hochberg yielded a lower cross-validated MSE of 334, which was also the best MSE amongst all previous model iterations. The selected interactions are listed below:

1. Division:incidenceRate
2. Division:povertyPercent
3. Division:PctUnemployed16_Over
4. Division:PctEmpPrivCoverage
5. Division:PctWhite

Because this model with interactions produced the lowests MSE, we established this as our final model.

References

1. https://www.businessinsider.com/union-county-is-least-healthy-in-us-2017-10
2. https://www.cancer.org/latest-news/understanding-cancer-death-rates.html
3. https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2020.html
4. https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf
5. https://www.data.world/nrippner/ols-regression-challenge
6. Fox, John. Applied Regression Analysis and Generalized Linear Models. Sage, 2016.
7. Gordon, Rachel A. 2015. Regression Analysis for the Social Sciences. New York and London: Routledge.
8. https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/home/?cid=nrcs143_013697

Appendix

##Appendix A: Full Data Dictionary

- `TARGET_deathRate`: Dependent variable. Mean per capita (100,000) cancer mortalities(a)
- `avgAnnCount`: Mean number of reported cases of cancer diagnosed annually(a)
- `avgDeathsPerYear`: Mean number of reported mortalities due to cancer(a)
- `incidenceRate`: Mean per capita (100,000) cancer diagoses(a)
- `medianIncome`: Median income per county (b)
- `popEst2015`: Population of county (b)
- `povertyPercent`: 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)
- `AvgHouseholdSize`: Mean household size of county (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)

- `PctPubliceCoverageAlone`: 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)
- (a): years 2010-2016
- (b): 2013 Census Estimates

Appendix B: Descriptions of geographic division

Division	Region	Number of States
New England	Northeast	6
Middle Atlantic	Northeast	3
East North Central	Midwest	5
West North Central	Midwest	7
South Atlantic	South	9
East South Central	South	4
West South Central	South	4
Mountain	West	8
Pacific	West	5

Appendix C

Put all code here