Stat 151A Final Project Proposal

Kendall Kikkawa, Jonathan Luo, Andre Sha

11/7/2020

Intro:

Cancer has a large impact on society, affecting people from all different walks of life. We wanted to pursue a regression project on cancer to be able to see what sort of relationships and patterns can we draw from key characteristics in the United States at the county level. Our curiosity for pursuing this project lies within a common interest in Healthcare, in which a question we asked ourselves prior to choosing this dataset included "What sort of governmental interventions lead towards reduced cancer mortality rates?" For our project, we want to predict cancer mortality rate for counties across the United States. Through these predictions, we aim to create a threshold upon our predicted dependant variable, *target_deathrate*, and for values above this set threshold, we will flag the respective counties and think of ways policy makers can improve socioeconomic outcomes in these areas. The dataset was aggregated from various U.S. governmental sources, including census.gov, clinicaltrials.gov, and cancer.gov.

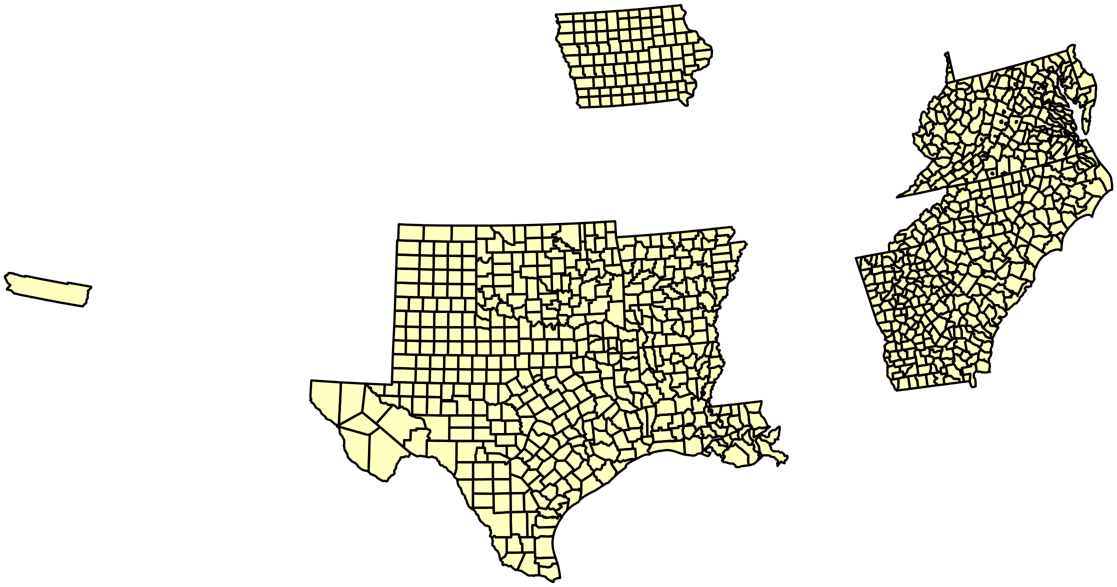
Regression Analysis Plan:

Exploratory Data Analysis

In order to get a better grasp of the data, it would be helpful to first create a histogram of our dependent variable to look at the distribution of our data. If it's not normally distributed, a box-cox transformation may need to be applied to our data. We can also see if our data has any clear outliers through this histogram. Some of our columns may need to be split into multiple columns. For example, the binned income column gives a lower and upper bound. This column should probably be split into two separate columns with one indicating a lower bound and the other an upper bound.

After looking at the data, there are a few columns with a large amount of null values where extrapolation to fill in these null values wouldn't make sense. As a result, we will drop these columns from our dataset. These columns are *PctSomeCol18_24*, *PctEmployed16_Over*, and *PctPrivateCoverageAlone*. A few columns also have some strange values. For example, counties in the states of Minnesota, Kansas, and Nevada all contain the same average annual count as well as incident rate and neither of these values appear to be any sort of a national mean or median. It's not clear where this value comes from and it's possible that this data could be the result of encoding errors. We may want to compare a model including these rows and model without these rows and see how they perform. If necessary, we may also look into the possibility of appending new data from other datasets to the original dataset. Because we have location data in our dataset, we can also map our data on a map and view trends by state and region. If there's a clear relationship between our features and regions of the map, we might want to add additional regionality features. Below is an example of a regional plot that we may create:

```r
#Example plot
usmap::plot_usmap("counties", fill = "yellow", alpha = 0.25,
                  # 06065 = Riverside County, CA
                  include = c(.south_region, "IA", "06065"),
                  # 12 = FL, 48141 = El Paso County, TX
                  exclude = c(.east_south_central, "12", "48141"))
```



We could then create a pairs plot to look at relationships between our variables and determine if any variables are collinear or need to be transformed to be more linear. With this pairs plot and along with intuition, we can probably determine a few different variables we may want to explore interactions between. In order to explore these interactions, we could create scatter plots isolated by variable for continuous:categorical interactions and boxplots isolated by variable for categorical:categorical interactions. If we notice any differences (i.e. a difference in slopes) between the plots of the different variables, then we may want to add that interaction to our model. We could also explore interaction terms after we've cut down on our model size to reduce the number of interactions we'd need to investigate. It's also possible that in the process of model selection, we'll come across features that seem to have interesting interactions that we may want to explore.

Model Selection

Given our data, our goal is to find the best model that gives the best prediction for cancer death rate in a county given their socioeconomic indicators. We've proposed a workflow in order to search for this particular model. Initially, we will perform an 80:20 split of our data into a training set and the testing set. Next, we perform Best Subset Regression on our training data to examine all possible models from our features, and we will track the top 5 models for each number of features (p = [1, 31]). In parallel, we will apply Lasso regression to see which of these explanatory variables get assigned non-zero coefficients to have a sense of feature importance. Afterwards, we'll assess the fit of the model using Mallow's Cp , AIC, and BIC to select a model from other contenders of differing variable sizes.

Now that we have a few different models of the same variable length, we'll use Stepwise selection to find a "Final model" based on ANOVA. If there are a few different models that converge to a single "Final model", we'd find a suitable local optimum (or possibly global optimum) and select those coefficients, but if there is no convergence during the stepwise selection process, then we may possibly try out an ensembling method or forwards/backwards selection to find a best performing model. At the end, we'll compare this best performing model with the lasso regression model.

Model Diagnostics

After we've identified a few top candidate models (according to our model selection criteria), we will test our linear modeling assumptions. First, we will compare the Adjusted-R2 values of each model to identify which ones achieve the linearity with the response variable. Next, we will analyze the outliers; we will construct fitted values vs residuals for our top models to identify any patterns (indicating non-linearity). If we observe a heteroscedastic pattern, we will consider making transformations to our response variable (number of deaths due to cancer) to fix the non-constant variance. We will also consider making a pairplot that compares the response variable to each of our explanatory variables to assess the individual linear relationships (correct for any relationships that were unaccounted for in EDA).

Additionally, we will create Q-Q plots to determine the normality of the residuals. We will also visualize influence vs the standardized residuals of each model to ensure that there are no counties that are dictating our model fit (and compute the corresponding Cook's Distances for each point). We have decided not to plot leverage because each data point is representative of a county, so we will be very hesitant to remove outliers, because doing so would restrict us from performing informed inference on those particular counties (i.e. we will only remove points if their influence is too large).

The last diagnostic we will use is to compute the test RMSE. After our top models have been selected, we use each model to make predictions on the test set. We will create a barplot that compares the training RMSE, CV RMSE, the Test RMSE of each model. An ideal model would yield similar values for all three of these metrics, which would indicate that the model would generalize well to future data.

Loaded Dataset Evidence

```r
cancer <- read.csv("Data/cancer_reg.csv") # load data
# remove columns with null values
cancer_cleaned <- subset(cancer, select=-c(PctSomeCol18_24, PctEmployed16_Over,
                                   PctPrivateCoverageAlone))
# PctSomeCol18_24 had 2285 missing values
# PctEmployed16_Over had 152 missing values
# PctPrivateCoverageAlone had 609 missing values
names(cancer_cleaned) # Feature names
```

```
##  [1] "avgAnnCount"             "avgDeathsPerYear"        "TARGET_deathRate"
##  [4] "incidenceRate"           "medIncome"               "popEst2015"
##  [7] "povertyPercent"          "studyPerCap"             "binnedInc"
## [10] "MedianAge"               "MedianAgeMale"           "MedianAgeFemale"
## [13] "Geography"               "AvgHouseholdSize"        "PercentMarried"
## [16] "PctNoHS18_24"            "PctHS18_24"              "PctBachDeg18_24"
## [19] "PctHS25_Over"            "PctBachDeg25_Over"       "PctUnemployed16_Over"
## [22] "PctPrivateCoverage"      "PctEmpPrivCoverage"      "PctPublicCoverage"
## [25] "PctPublicCoverageAlone"  "PctWhite"                "PctBlack"
## [28] "PctAsian"                "PctOtherRace"            "PctMarriedHouseholds"
## [31] "BirthRate"
```

```r
dim(cancer_cleaned) # Matrix dimensions
```

```
## [1] 3047   31
```

Sources

https://data.world/nrippner/ols-regression-challenge?fbclid=IwAR0BCbybr7WJbL3F561rSxWbDdYB8L3L2VfgT6yDQhgK-XgIDOaagi_gtyM

`https://www.cancer.org/latest-news/understanding-cancer-death-rates.html`

https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2020.html