project update 2

Hongkai Wang

2022-11-07

Question 1

Group Number: 7

Group Name: Regression Heroes

Group Members: Ryan Wang, Stella Nam, Hongkai Wang

Question 2

Part a

Yes, we did a literature review of similar problems relating various cancer outcomes to socioeconomic factors. However, a lot of the studies that looked into the possible influence of socioeconomic factors on cancer outcomes are fairly recent and it appears that there needs to be more research into this relationship overall. Additionally, a lot of the research in this field was done using data from nations other than the US.

There was a **2018 paper published in JAMA** that looked at county-level median incomes and cancer death rates in the US. They used a different dataset, data from the Institute for Health Metrics and Evaluation from 2014 along with the income data from the 2012 US Census Bureau Small Area Income and Poverty Estimates. The data were stratified into low-, medium-, and high-income groups. They used a single- and multiple-mediator model to assess changes in parameter estimates after adding potential mediators other than household income, such as smoking. They also used a multivariate normal regression model for sensitivity analysis. Although the main questions being posed by this paper is very similar to our project, we will that we can more variables we can consider that potentially influence cancer death rates as well as individual data, which can hopefully confirm the results from this study also provide more insight into these complex relationships between the covariates. We think the multiple-mediator model might be interesting to include in our project because there are multiple factors and potential mediators that likely result in cancer deaths. We can further study these complex relationships in our data using this model.

Additionally, there was another paper published in **Frontiers in 2022** that looks at cancer-free life expectancy trends from 2006-2018 based on income inequalities using German Health Insurance Data. They used a proportional hazard regression model for their statistical analysis. This paper not only looked at cancer risk overall, but also looked into specific cancer outcomes, such as colon, stomach, and lung cancer based on gender. We feel that we can use the finding and potential factors from this paper such as gender and insurance coverage to assess cancer death rates in our dataset.

There was a [**2018 paper published in PLoS One**][https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5819792/] specifically looking at lung cancer and socioeconomic status. This study used a pooled analysis of different case-control studies. Here, they looked at data from Europe and Canada. They used an unconditional logistic regression model as well as random-effect meta-regression models to assess this association. The paper concludes that SES is a risk factor for lung cancer after adjustment for smoking behavior. However, a major limitation of this study was that they were not able to adequately control smoking behavior, which is one of the leading causes of lung cancer. Although the conclusions from the study are interesting as the authors are studying topics similar to our project topic but in different countries, the models and methods used in this study do not apply to our project.

Part b

Yes, we received our peers' and the Teaching Team's reviews of our 1st Project Check-in.

From our peers, we received a few comments about looking at more biologically-grounded factors such as health conditions and genetics on the risk of cancer deaths. Although we do not think these factors aligned with the purpose of our research questions since we are trying to assess the influence of SES factors, we think it would be interesting if we overlay our results from our analysis to the existing literature on cancer outcomes based on genetics or other biological markers at the end.

There were also a few concerns about the data being generalized to all cancers rather than specific cancers. Based on our prior education knowledge and reviewing the literature, we agree that a lot of cancer outcomes vary greatly based on the specific type of cancer. Some cancers are also more treatable than others, resulting in a lower death rate on average. However, there was also a comment from one of our other reviewers regarding our outcome variable. They correctly identified that the data was specific to lung cancer data rather than cancer generally. Going forward, we will take that into account in our analysis, specific to lung cancer.

The peer reviewers also mentioned that we should focus on key variables such as income since socioeconomic factors are very vague. Upon our EDA, we can on focusing on the variables that seem most correlated with cancer death rates overall, which include but are not limited to median income, median age, and public coverage.

From the teaching team, we had a comment about whether we wanted to predict cancer-related death or consider the association between cancer-related death and one or a limited number of exposures. Upon doing more research into lung cancer, we think it would be more appropriate to explore the latter. This is because there is already strong evidence in the literature suggesting that lung cancer has a strong association with environmental factors such as smoking. These factors are not included in our data, nor is it the purpose of our project. Additionally, socioeconomic influences are not well understood in lung cancer incidence and death. Thus, trying to understand the association between cancer deaths and certain socioeconomic factors is something we are trying to explore with our project.

Lastly, we did not contact any non-teaching team domain experts because we stated in our first project check-in that we thought it was not necessary for the scope of our project. Upon conducting a literature review and reviewing our peers' and Teaching Team's reviews, we stand by this decision.

Question 3

At the moment, we do not have too many changes we want to make to our analysis plan. Changes to our plan is shown in blue.

Exploratory Data Analysis:

- Better understand the structure and complexity of our data set through str and summary functions in R
- Check for missing data points categorized by different covariates (by employment status of different age groups)
- Explore data patterns and modeling – Check for correlation between different socioeconomic status determining variables and cancer prevalence through a pairs plot.

  - One of the variables of interest of our project is the effect of region/state on cancer prevalence. Graph this relationship to observe for any patterns.

Primary Inference Problem:

- Poisson regression (within a generalized linear framework) will be our main modeling approach for statistical inference due to our primary outcome being rate/count data on lung cancer death.
- With 33 covariates of interest, we will need to employ model selection techniques, potentially finding a good in-between for various automated methods and integrating domain information
- Once we have identified some covariates of interest, we will perform various smoothing methods in order to determine the relative correlations in data
  - Currently, based on our EDA and project checkins, median income, median age, and public health coverage seem like good potential factors we can include as SES. Additionally, we are continually updating our dataset with other factors we believe are revelant to our primary and secondary questions such as data on politcal preference and education attainment data.
- Data cleaning and computing additional predictors (e.g., regions, climate…)
  - Given that we have a lot of missing data, especially for the college education section, we will discard this data as it will be difficult to address and work with data where so much of it is missing. (more on this in question 4)
  - For insurance data where we have around 20 percent missing data, we will perform a sensitivity analysis by viewing models under complete cases, under imputation, and under other missing data methods such as inverse proportional weighting.
- As our data is on the county-level, we aim to add a dimension to our analysis by aggregating these data onto the state-level, which would allow us to study state-by-state differences in socioeconomic conditions and cancer outcomes
- Create comprehensive visualization of our demographic data to illustrate differences between US counties and states
- We will explore different imputation techniques to adjust for the 3046 incomplete observations in these data
- Identify and adjust for potential confounders (on the association of SES to cancer prevalence) such as, but not limited to, county and income.
- For any of these covariates which we are not interested in interpreting, we aim to adjust for them flexibly using splines/GAMs (and if not needed, we will keep using linear terms)

Secondary Inference Problem:

- We are potentially interested in looking at state-level and county-level differences in demographics and insurance status
- Here we will employ multinomial models in order to model the probability of some set of socioeconomic conditions and cancer outcomes being in a given county or state
- We are also interested in looking at the relationship between socioeconomic status and political preference of different demographic groups. We will find an additional data on voting outcomes from a recent US election and merge that with our current dataset. Since the US elections use a bipartisan system, we can use logistic linear regression to understand this relationship.

Predictive Problem:

- Aside from the inference problem, we are also interested in the prediction problem and if there is time, we will explore the performance of our generalized linear models on prediction metrics (accuracy, AUC, etc.)

- We will also explore how penalized models (LASSO, Ridge, Elastic Net) and traditional ML algorithms (random forests/basic nnets for primary analysis, RFs/KNN/nnets/etc. for secondary analysis tasks) compare with the inference models we previously built
- Here we are also interested in exploring how the number of predictors can correlate with generalizability in these 3 forms of models

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Exploratory Data Analysis

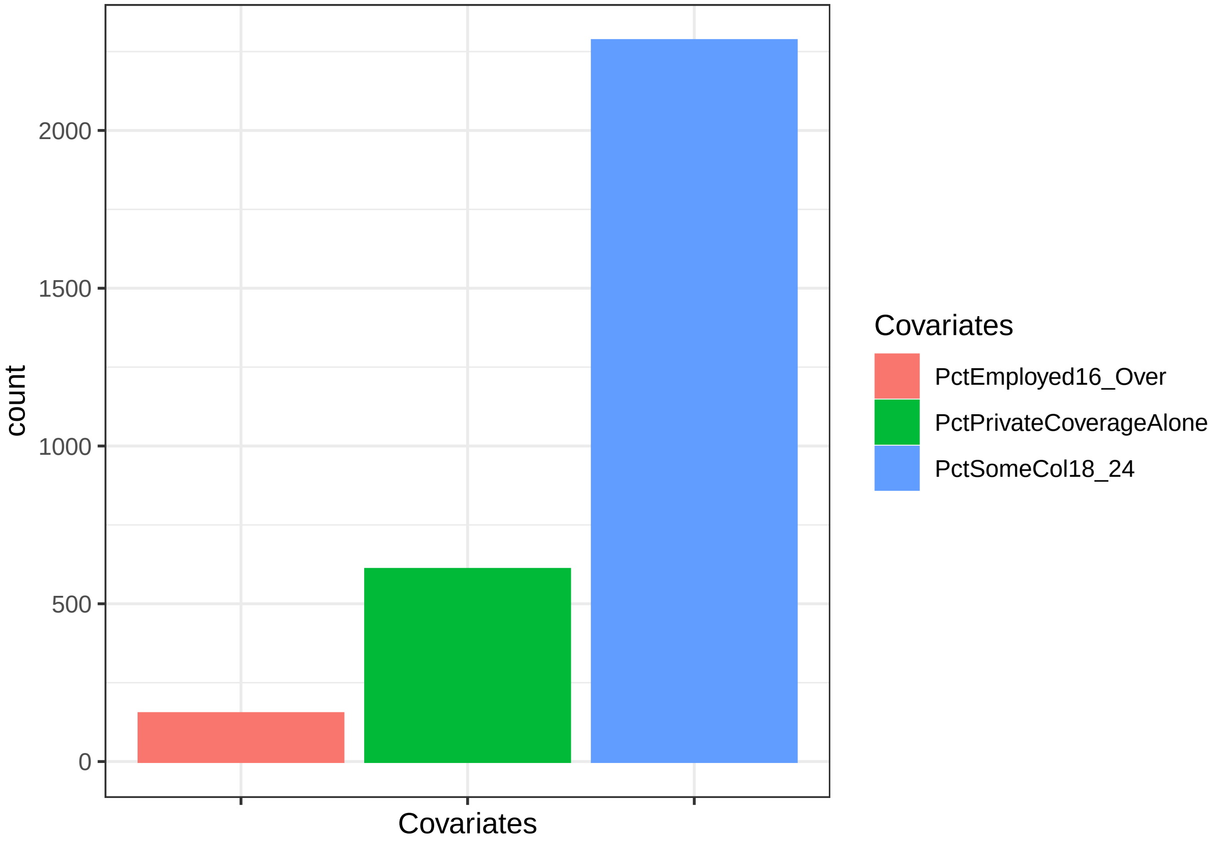
```
## Rows: 3047 Columns: 34
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): binnedInc, Geography
## dbl (32): avgAnnCount, avgDeathsPerYear, TARGET_deathRate, incidenceRate, me...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
## spec_tbl_df [3,047 x 34] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ avgAnnCount            : num [1:3047] 1397 173 102 427 57 ...
##  $ avgDeathsPerYear       : num [1:3047] 469 70 50 202 26 152 97 71 36 1380 ...
##  $ TARGET_deathRate       : num [1:3047] 165 161 175 195 144 ...
##  $ incidenceRate          : num [1:3047] 490 412 350 430 350 ...
##  $ medIncome              : num [1:3047] 61898 48127 49348 44243 49955 ...
##  $ popEst2015             : num [1:3047] 260131 43269 21026 75882 10321 ...
##  $ povertyPercent         : num [1:3047] 11.2 18.6 14.6 17.1 12.5 15.6 23.2 17.8 22.3 13.1 ...
##  $ studyPerCap            : num [1:3047] 499.7 23.1 47.6 342.6 0 ...
##  $ binnedInc              : chr [1:3047] "(61494.5, 125635]" "(48021.6, 51046.4]" "(48021.6, 51046.4]
##  $ MedianAge              : num [1:3047] 39.3 33 45 42.8 48.3 45.4 42.6 51.7 49.3 35.8 ...
##  $ MedianAgeMale          : num [1:3047] 36.9 32.2 44 42.2 47.8 43.5 42.2 50.8 48.4 34.7 ...
##  $ MedianAgeFemale        : num [1:3047] 41.7 33.7 45.8 43.4 48.9 48 43.5 52.5 49.8 37 ...
##  $ Geography              : chr [1:3047] "Kitsap County, Washington" "Kittitas County, Washington" "
##  $ AvgHouseholdSize       : num [1:3047] 2.54 2.34 2.62 2.52 2.34 2.58 2.42 2.24 2.38 2.65 ...
##  $ PercentMarried         : num [1:3047] 52.5 44.5 54.2 52.7 57.8 50.4 54.1 52.7 55.9 50 ...
##  $ PctNoHS18_24           : num [1:3047] 11.5 6.1 24 20.2 14.9 29.9 26.1 27.3 34.7 15.6 ...
##  $ PctHS18_24             : num [1:3047] 39.5 22.4 36.6 41.2 43 35.1 41.4 33.9 39.4 36.3 ...
##  $ PctSomeCol18_24        : num [1:3047] 42.1 64 NA 36.1 40 NA NA 36.5 NA NA ...
##  $ PctBachDeg18_24        : num [1:3047] 6.9 7.5 9.5 2.5 2 4.5 5.8 2.2 1.4 7.1 ...
##  $ PctHS25_Over           : num [1:3047] 23.2 26 29 31.6 33.4 30.4 29.8 31.6 32.2 28.8 ...
##  $ PctBachDeg25_Over      : num [1:3047] 19.6 22.7 16 9.3 15 11.9 11.9 11.3 12 16.2 ...
##  $ PctEmployed16_Over     : num [1:3047] 51.9 55.9 45.9 48.3 48.2 44.1 51.8 40.9 39.5 56.6 ...
##  $ PctUnemployed16_Over   : num [1:3047] 8 7.8 7 12.1 4.8 12.9 8.9 8.9 10.3 9.2 ...
##  $ PctPrivateCoverage     : num [1:3047] 75.1 70.2 63.7 58.4 61.6 60 49.5 55.8 55.5 69.9 ...
##  $ PctPrivateCoverageAlone: num [1:3047] NA 53.8 43.5 40.3 43.9 38.8 35 33.1 37.8 NA ...
##  $ PctEmpPrivCoverage     : num [1:3047] 41.6 43.6 34.9 35 35.1 32.6 28.3 25.9 29.9 44.4 ...
##  $ PctPublicCoverage      : num [1:3047] 32.9 31.1 42.1 45.3 44 43.2 46.4 50.9 48.1 31.4 ...
##  $ PctPublicCoverageAlone : num [1:3047] 14 15.3 21.1 25 22.7 20.2 28.7 24.1 26.6 16.5 ...
##  $ PctWhite               : num [1:3047] 81.8 89.2 90.9 91.7 94.1 ...
##  $ PctBlack               : num [1:3047] 2.595 0.969 0.74 0.783 0.27 ...
##  $ PctAsian               : num [1:3047] 4.822 2.246 0.466 1.161 0.666 ...
##  $ PctOtherRace           : num [1:3047] 1.843 3.741 2.747 1.363 0.492 ...
##  $ PctMarriedHouseholds   : num [1:3047] 52.9 45.4 54.4 51 54 ...
##  $ BirthRate              : num [1:3047] 6.12 4.33 3.73 4.6 6.8 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   avgAnnCount = col_double(),
##   ..   avgDeathsPerYear = col_double(),
##   ..   TARGET_deathRate = col_double(),
##   ..   incidenceRate = col_double(),
##   ..   medIncome = col_double(),
##   ..   popEst2015 = col_double(),
##   ..   povertyPercent = col_double(),
##   ..   studyPerCap = col_double(),
##   ..   binnedInc = col_character(),
##   ..   MedianAge = col_double(),
##   ..   MedianAgeMale = col_double(),
##   ..   MedianAgeFemale = col_double(),
##   ..   Geography = col_character(),
##   ..   AvgHouseholdSize = col_double(),
##   ..   PercentMarried = col_double(),
##   ..   PctNoHS18_24 = col_double(),
##   ..   PctHS18_24 = col_double(),
##   ..   PctSomeCol18_24 = col_double(),
##   ..   PctBachDeg18_24 = col_double(),
##   ..   PctHS25_Over = col_double(),
##   ..   PctBachDeg25_Over = col_double(),
##   ..   PctEmployed16_Over = col_double(),
##   ..   PctUnemployed16_Over = col_double(),
##   ..   PctPrivateCoverage = col_double(),
##   ..   PctPrivateCoverageAlone = col_double(),
##   ..   PctEmpPrivCoverage = col_double(),
##   ..   PctPublicCoverage = col_double(),
##   ..   PctPublicCoverageAlone = col_double(),
##   ..   PctWhite = col_double(),
##   ..   PctBlack = col_double(),
##   ..   PctAsian = col_double(),
##   ..   PctOtherRace = col_double(),
##   ..   PctMarriedHouseholds = col_double(),
##   ..   BirthRate = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

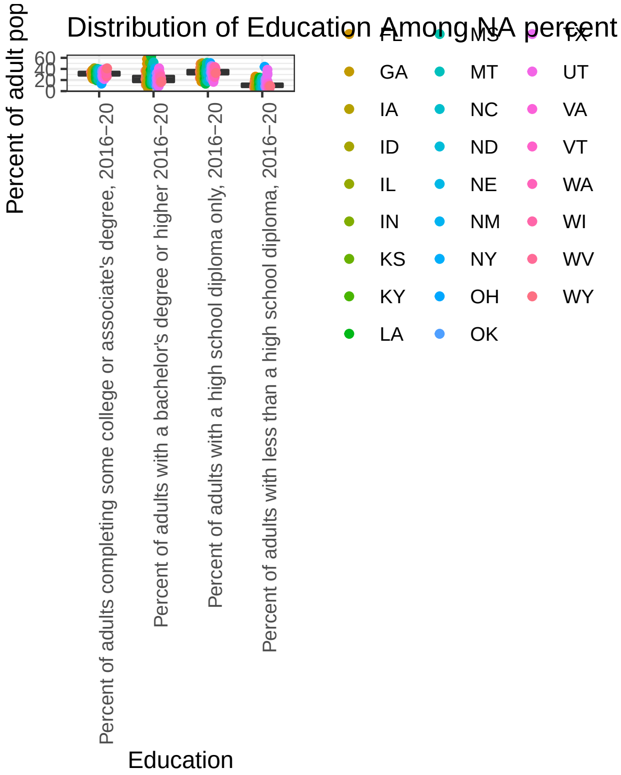
```
## Rows: 169235 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): State, Area name, Attribute
## dbl (2): FIPS Code, Value
##
```

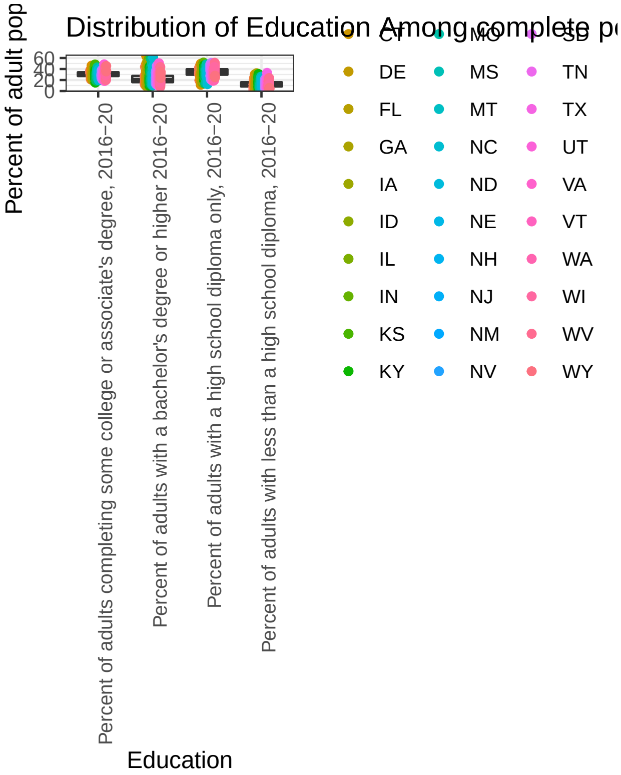
```
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 50 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): State, Abbr, State Capital, Region
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

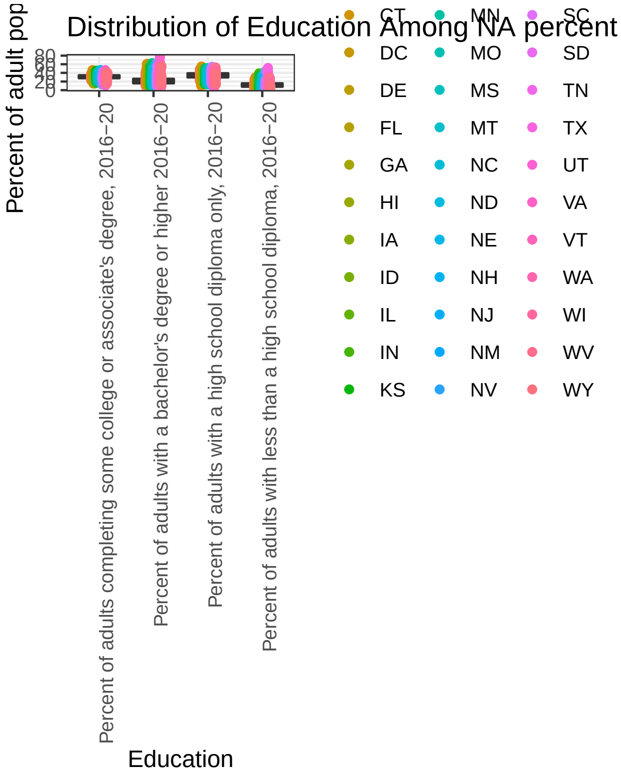
How much missing data is there?

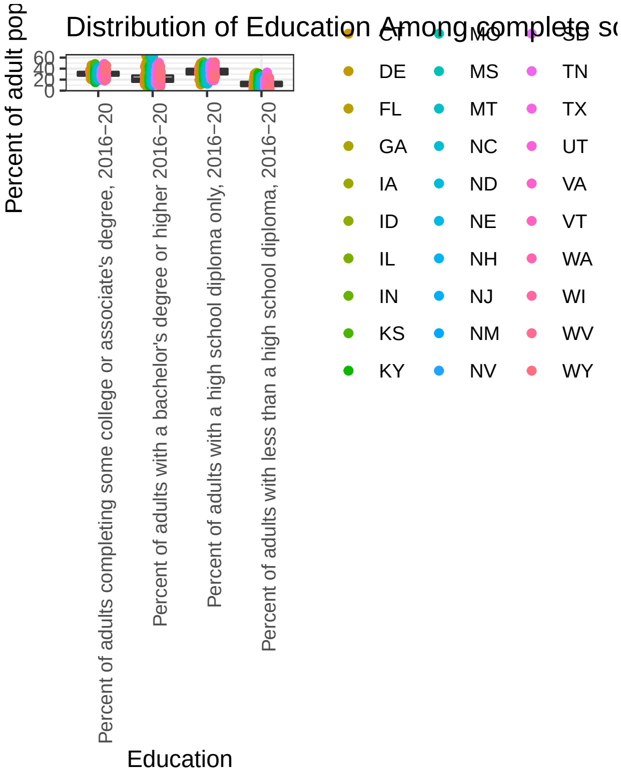


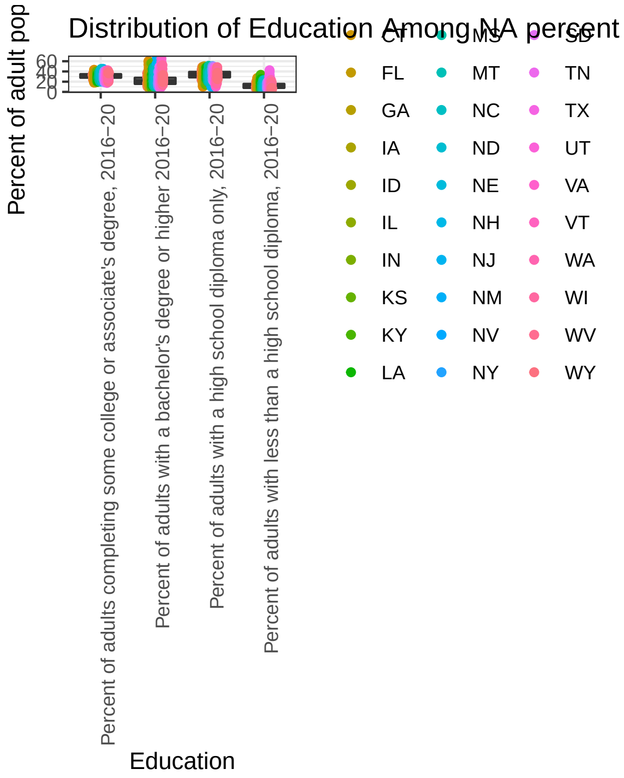
Interestingly, there is some missing data in the reporting of county employment for residents age 16 and over. We observe more missing data in percentages of county residents with only private healthcare coverage, and missing data in the majority of counties for reports of the percent of county residents between 18 and 24 years old with with some college as their highest attained education. One hypothesis for why there may be a large amount of missing data in educational reporting for 18 to 24 year olds with some college may be because some counties may not have high emphasis on higher education and many individuals in this age range that are pursuing college will probably be in more college-oriented counties.

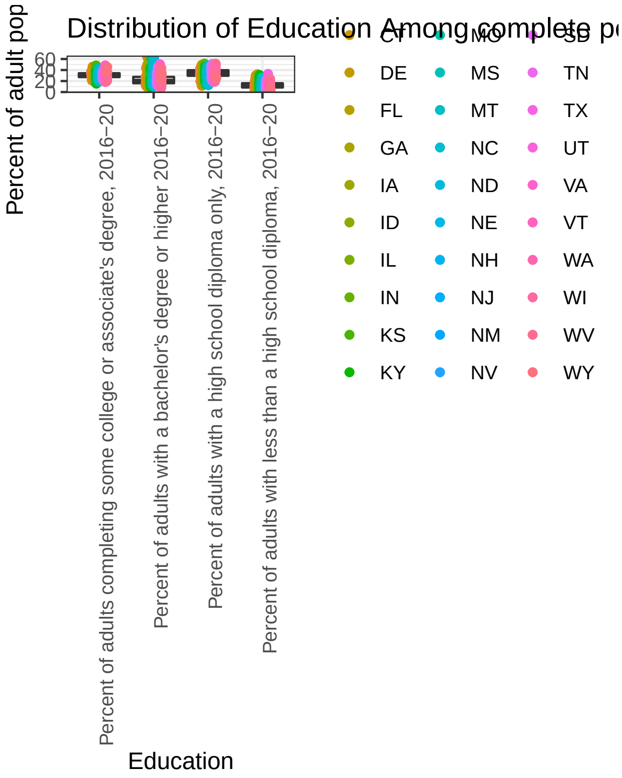

Distribution of Education Among NA percent
Percent of adult pop
60
40
20
0
Percent of adults completing some college or associate's degree, 2016−20
Percent of adults with a bachelor's degree or higher 2016−20
Percent of adults with a high school diploma only, 2016−20
Percent of adults with less than a high school diploma, 2016−20
Education
FL
GA
IA
ID
IL
IN
KS
KY
LA
MS
MT
NC
ND
NE
NM
NY
OH
OK
TX
UT
VA
VT
WA
WI
WV
WY


Distribution of Education Among complete p
Percent of adult pop
60
40
20
0
Percent of adults completing some college or associate's degree, 2016−20
Percent of adults with a bachelor's degree or higher 2016−20
Percent of adults with a high school diploma only, 2016−20
Percent of adults with less than a high school diploma, 2016−20
Education
CT
DE
FL
GA
IA
ID
IL
IN
KS
KY
MO
MS
MT
NC
ND
NE
NH
NJ
NM
NV
SD
TN
TX
UT
VA
VT
WA
WI
WV
WY


Distribution of Education Among NA percent
Percent of adult pop
80
60
40
20
0
Percent of adults completing some college or associate's degree, 2016−20
Percent of adults with a bachelor's degree or higher 2016−20
Percent of adults with a high school diploma only, 2016−20
Percent of adults with less than a high school diploma, 2016−20
Education
CT
DC
DE
FL
GA
HI
IA
ID
IL
IN
KS
MN
MO
MS
MT
NC
ND
NE
NH
NJ
NM
NV
SC
SD
TN
TX
UT
VA
VT
WA
WI
WV
WY


Distribution of Education Among complete sc
Percent of adult pop
60
40
20
0
Percent of adults completing some college or associate's degree, 2016−20
Percent of adults with a bachelor's degree or higher 2016−20
Percent of adults with a high school diploma only, 2016−20
Percent of adults with less than a high school diploma, 2016−20
Education
CT
DE
FL
GA
IA
ID
IL
IN
KS
KY
MO
MS
MT
NC
ND
NE
NH
NJ
NM
NV
SD
TN
TX
UT
VA
VT
WA
WI
WV
WY





```
#colMeans(is.na(no_geodat))*100
```

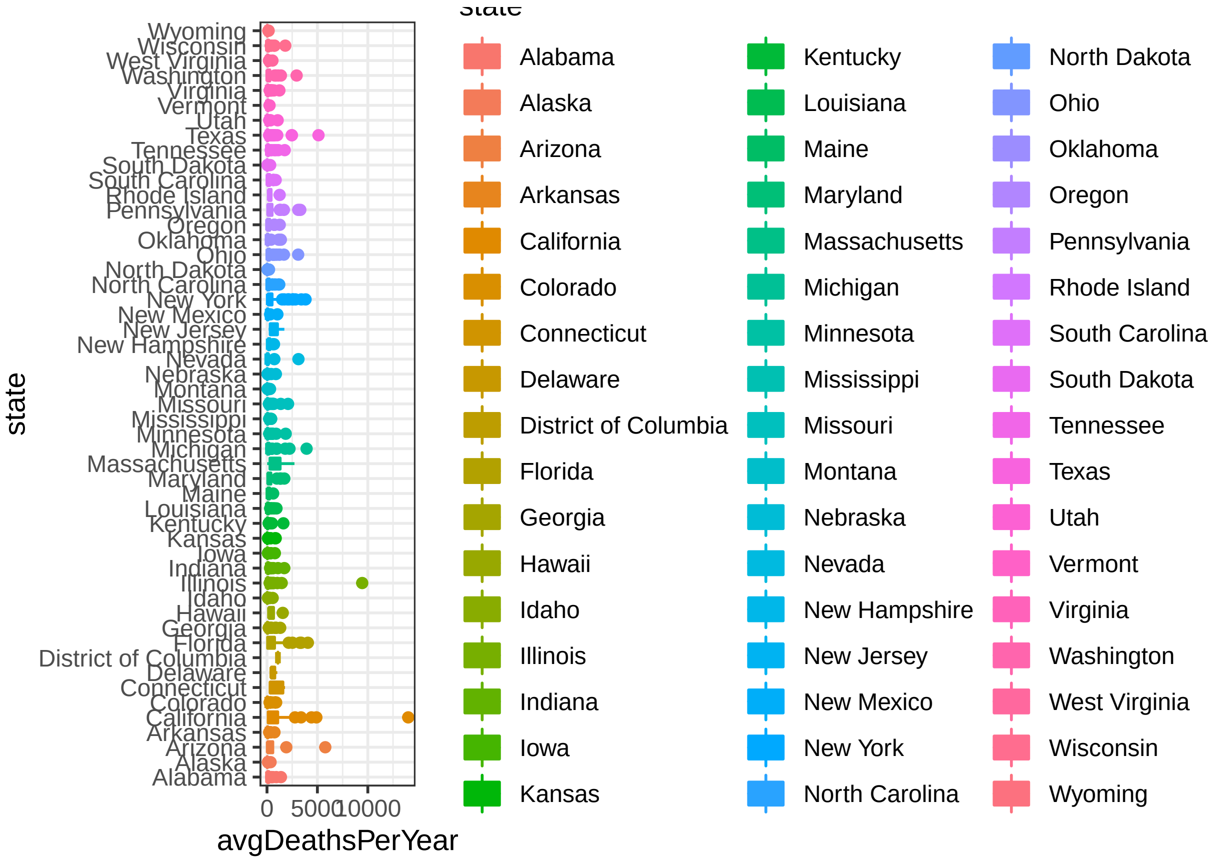
```
## # A tibble: 191 x 11
##    FIPS ~1 State count~2 Less ~3 High ~4 Some ~5 Bache~6 Perce~7 Perce~8 Perce~9
##      <dbl> <chr> <chr>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1       0 US    United~ 2.56e7  5.94e7  6.45e7  7.34e7   11.5     26.7    28.9
##  2    1000 AL    Alabama 4.39e5  1.01e6  1.01e6  8.77e5   13.1     30.3    30.3
##  3    1035 AL    Conecu~ 1.33e3  4.12e3  2.15e3  1.22e3   15.0     46.7    24.4
##  4    2000 AK    Alaska  3.32e4  1.37e5  1.68e5  1.45e5    6.86    28.4    34.7
##  5    2013 AK    Aleuti~ 3.54e2  1.08e3  7.77e2  4.38e2   13.4     40.7    29.4
##  6    2060 AK    Bristo~ 3.5 e1  1.68e2  2.13e2  1.45e2    6.24    29.9    38.0
##  7    2063 AK    Chugac~ 2.04e2  1.13e3  1.99e3  1.36e3    4.35    24.0    42.6
##  8    2066 AK    Copper~ 8.6 e1  6.5 e2  5.51e2  5.93e2    4.57    34.6    29.3
##  9    2068 AK    Denali~ 4.2 e1  5.63e2  5.31e2  7.3 e2    2.25    30.2    28.5
## 10    2105 AK    Hoonah~ 1.05e2  6.35e2  6.62e2  3.42e2    6.02    36.4    38.0
## # ... with 181 more rows, 1 more variable:
## #   ‘Percent of adults with a bachelor’s degree or higher 2016-20‘ <dbl>, and
## #   abbreviated variable names 1: ‘FIPS Code‘, 2: county_name,
## #   3: ‘Less than a high school diploma, 2016-20‘,
## #   4: ‘High school diploma only, 2016-20‘,
## #   5: ‘Some college or associate’s degree, 2016-20‘,
## #   6: ‘Bachelor’s degree or higher, 2016-20‘, ...

## # A tibble: 0 x 48
## # ... with 48 variables: FIPS Code <dbl>, State <chr>, county_name <chr>,
## #   Less than a high school diploma, 2016-20 <dbl>,
## #   High school diploma only, 2016-20 <dbl>,
## #   Some college or associate’s degree, 2016-20 <dbl>,
## #   Bachelor's degree or higher, 2016-20 <dbl>,
## #   Percent of adults with less than a high school diploma, 2016-20 <dbl>,
## #   Percent of adults with a high school diploma only, 2016-20 <dbl>, ...
```

Here we show some difference in county representation within our two integrated cancer trial and socioeconomic dataset with a dataset of education attainment by county. Notably, a large difference in the counties from both datasets is the inclusino of Puerto Rico. While the education dataset includes Puerto Rico, the cancer trial data set does not. This means this missing data is **MAR** for our primary inference since it depends on a covariate *State* (or **MNAR** for our secondary as county is an outcome), however we will consider our analysis without Puerto Rico as it is a unique situation and not localized to the North American land mass. Other missing cancer data are at the county level, not found in the education dataset, similarly, as we are focused on the cancer data, we will disregard these education data (as we have education data for all cancer-statistic counties we have).

Exploring data patterns and modeling

Look at the state-wide view



4. Missing data

a.

Overall, our data contains 19.9869% data missing from private health insurance coverage alone, 4.9885%% data missing from percent of individuals 16 or over that are employed, and 74.9918% data missing from percent of individuals who have some college education between ages of 18 to 24.

We believe the missing private health insurance data is *MAR* due to the fact that some other counties in the same state have this data, and the individual who created this data set did not include the processing scripts for insurance data. This leads me to believe the most probable reason for these missing data are due to not carefully checking for missing data. Due to it being 19.9869% missing, We don't believe we can simply disregard this data, as such we will need to formally address it via some missing data methods.

With similar reasoning, we believe the 4.9885% missing data from percentage of employed individuals 16 or over and 74.9918% missing data from percentage of individuals with some college education between ages 18 to 24 are also *MAR*. Especially with the college education, we found external census data on the county-level providing education data, reinforcing the belief that the original data was lost through data processing, perhaps due to issues like attributes not matching on data set joins.

b.

For the 4.9885% missing data from percentage of employed individuals 16 or over and 74.9918% missing data from percentage of individuals with some college education between ages 18 to 24, we will drop the missing data. This is because the employement data is still under the 5% threshold by rule-of-thumb, and should not have a huge impact on biasing our resulting estimates. Due to the large amount of missing data in the education attribute, we will simply drop that attribute, as we do not believe we would be able to get a meaningful analysis using it anyways (with or without missing data methods).

For the 19.9869% of missing insurance data, we will perform a sensitivity analysis. That is, viewing models under complete cases, under imputation (mean, EM, random forest), and under other missing data methods such as inverse proportional weighting. We will also explore these methods conditional on the states, as we believe that counties within a state are more likely to be similar to each other as they are geographically similar, and share state-level policies, which would likely influence the socioeconomic attributes within a county.

5. Modelling Approches

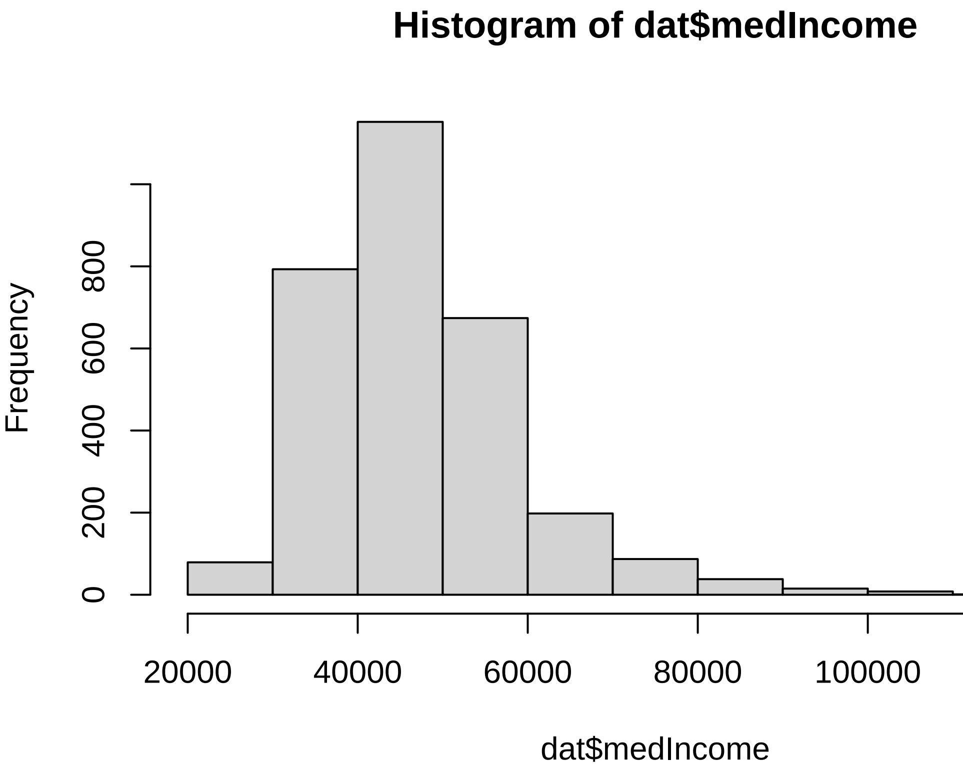
a. Fitting an linear model

Linear regression models hold a special position in the world of statistics, as it is one of the oldest, and the most tried-and-true method of models. For the purpose of this project, linear models, both GAM and multiple regression, serve as an valuable and important introduction to the exploration of the dataset. Our data set contains information on the socioeconomic status, lung cancer death rate, demographic, etc. of US counties. If we view cancer death rate as an continuous variable outcome, we can most certainly derive very useful information on the relationship between social factors and the cancer death rate within different US counties.

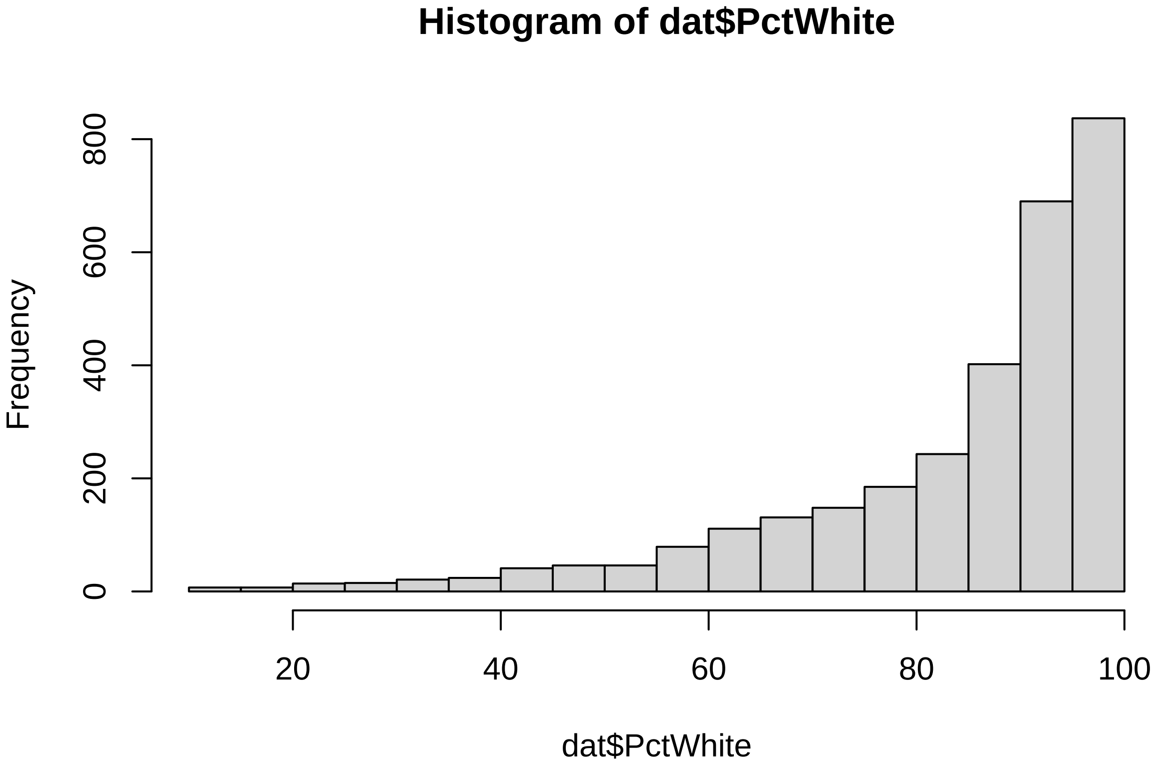
Our main interest in the data set is whether we can derive some useful relationship between the social-economic status of US counties, and their lung cancer-related death rate. In our data set, the lung cancer-related death rate is specifically defined to be the Mean per 100,000 citizen lung cancer mortality over the data collection period (2010 -2016, with exception of the additional education data that was collected in 2020). Before we formally think about the model structure, we should first define our set of social economic status indicators.

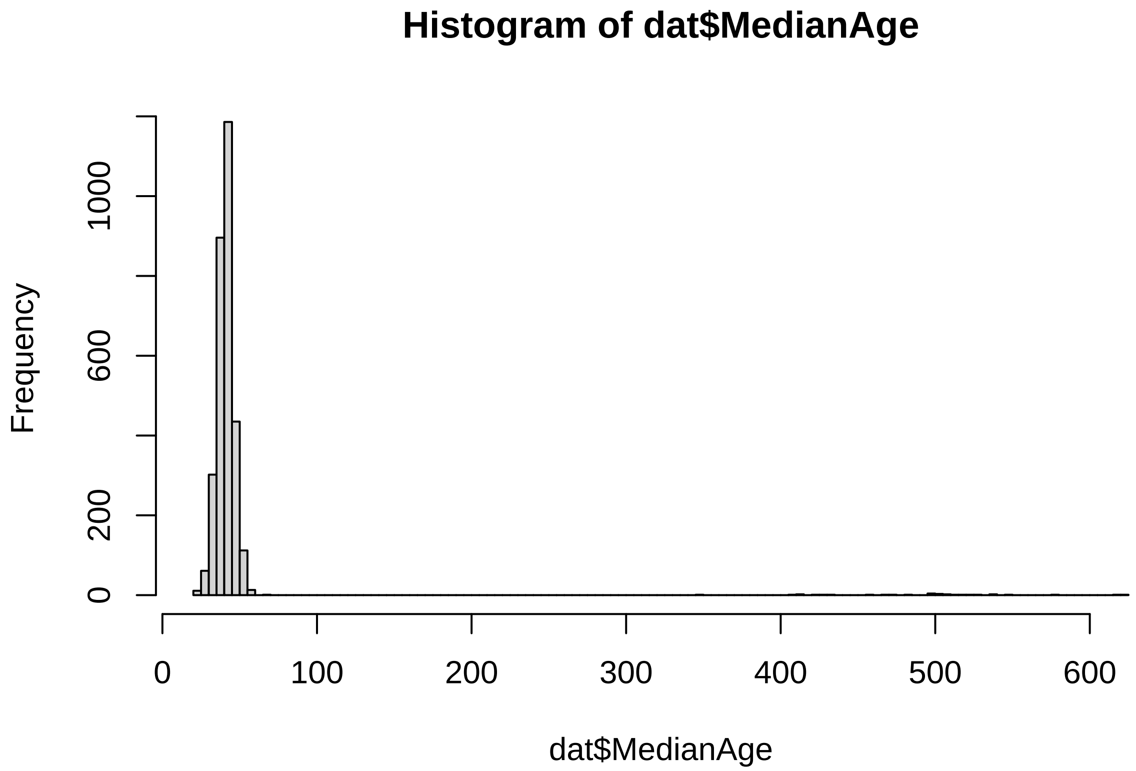
Several co-variates from our data is specifically useful in defining the social-economic status of a region. First, median income is a key predictor as it is often used by economists in their evaluation of the wealthiness of a region. Second, median age is an important demographic descriptor in two sense. Median age can describe both the likelihood to get lung cancer (older people are more susceptible to cancer) and the wealth level of the county (older people are more likely to be richer). Third, percentage of white people in each county could also be useful in defining the social-economic status of the region, as historically, white neighborhoods are more likely to be better funded, and thus result in better healthcare conditions. With these indicators in mind, let's first explore a multiple linear regression model.

Model assumptions: Since our data are sourced from census, and goverment sources, we have reason to believe that each predictor variables following the central limiting theorem (each covariate entry is the mean over multiple years samples), and the LINE assumptions should be met.

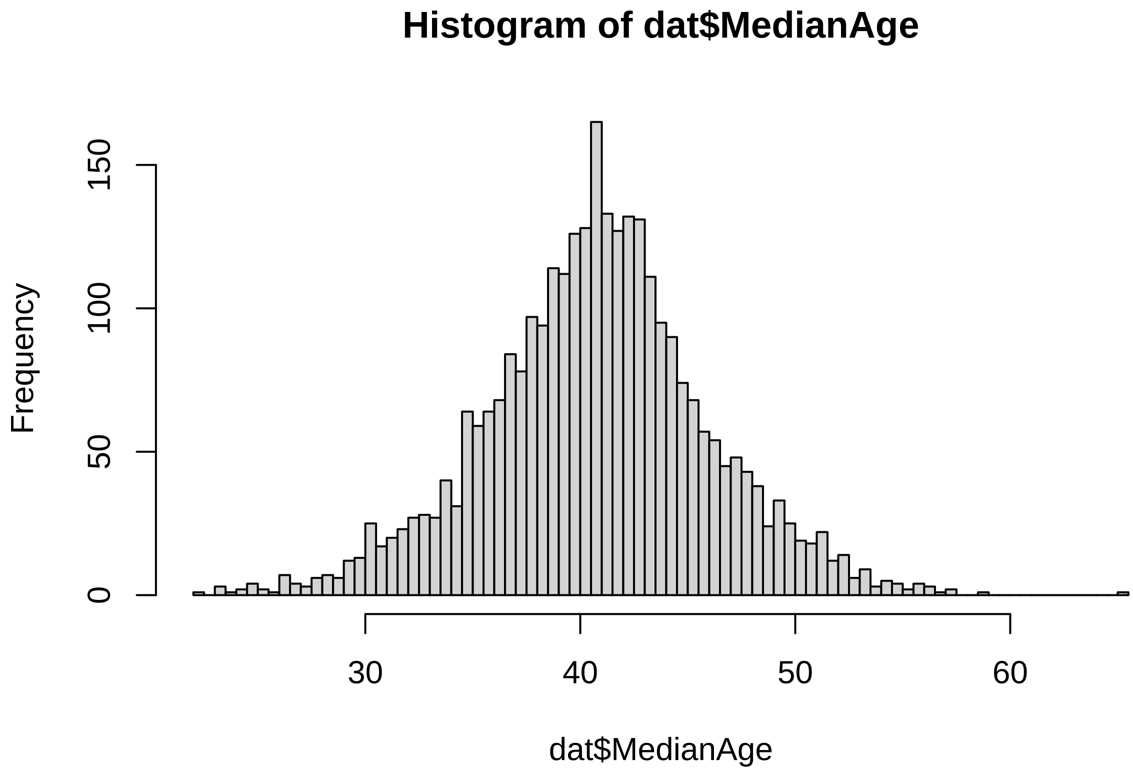


data transformation and cleaning:





we see that there are some insanely high median age here. These has to be errors in data collection or imputing process as it is improbable to have median ages that high. let's remove these rows of data for now.



The other two variables are skewed, but they should be workable. I am not normalizing the income data as the interpretability is a bit better this way.

model fitting:

```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome + MedianAge + PctWhite,
##     data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -123.117  -14.061    0.904   15.057  175.883
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.411e+02  4.250e+00  56.735  < 2e-16 ***
## medIncome   -9.492e-04  3.889e-05 -24.409  < 2e-16 ***
## MedianAge   -7.100e-02  9.591e-02  -0.740    0.459
## PctWhite    -1.785e-01  3.065e-02  -5.823  6.4e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.89 on 3013 degrees of freedom
## Multiple R-squared:  0.1954, Adjusted R-squared:  0.1946
## F-statistic:   244 on 3 and 3013 DF,  p-value: < 2.2e-16
```

Based on the model summary, we can see that median age is not a significant predictor of lung cancer death rate itself, which is a bit surprising based on previous assumptions. However, it could act still act as a confounder or effect modifier for our model. At the same time, it is good to see that both median income and the percentage of white population within a county is significantly correlated with lung cancer death rate. However, as previously illustrated, median age could also be a confounder to median income. Thus, let’s quickly do some tests on whether median age is a confounder here.

```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome + PctWhite, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -122.987  -14.167    0.874   15.145  175.870
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.387e+02  2.748e+00  86.894  < 2e-16 ***
## medIncome   -9.436e-04  3.813e-05 -24.746  < 2e-16 ***
## PctWhite    -1.876e-01  2.806e-02  -6.686 2.73e-11 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 24.89 on 3014 degrees of freedom
## Multiple R-squared:  0.1953, Adjusted R-squared:  0.1947
## F-statistic: 365.7 on 2 and 3014 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Model 1: TARGET_deathRate ~ medIncome + PctWhite
## Model 2: TARGET_deathRate ~ medIncome + MedianAge + PctWhite
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   3014 1867102
## 2   3013 1866763  1    339.57 0.5481 0.4592
```

We see that removing median age doesn’t affect the coefficients of the two other covariates at all. And LRT result supports this finding.

```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome + PctWhite + MedianAge +
##     medIncome * MedianAge, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -123.331  -13.843    0.929   14.955  175.902
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.012e+02  1.625e+01  12.379  < 2e-16 ***
## medIncome           -7.301e-05  3.464e-04  -0.211   0.8331
## PctWhite            -1.806e-01  3.064e-02  -5.894  4.2e-09 ***
## MedianAge            9.390e-01  4.082e-01   2.301   0.0215 *
## medIncome:MedianAge -2.213e-05  8.693e-06  -2.546   0.0110 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 24.87 on 3012 degrees of freedom
## Multiple R-squared:  0.1972, Adjusted R-squared:  0.1961
## F-statistic: 184.9 on 4 and 3012 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Model 1: TARGET_deathRate ~ medIncome + PctWhite + MedianAge + medIncome *
##     MedianAge
## Model 2: TARGET_deathRate ~ medIncome + MedianAge + PctWhite
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1   3012 1862754
## 2   3013 1866763 -1   -4008.3 6.4812 0.01095 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome + PctWhite + MedianAge +
##     PctWhite * MedianAge, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -123.122  -14.063    0.896   15.064  175.865
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.400e+02  1.824e+01  13.160   <2e-16 ***
## medIncome          -9.493e-04  3.890e-05 -24.401   <2e-16 ***
## PctWhite           -1.653e-01  2.123e-01  -0.779    0.436
## MedianAge          -4.166e-02  4.759e-01  -0.088    0.930
## PctWhite:MedianAge -3.438e-04  5.461e-03  -0.063    0.950
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 24.9 on 3012 degrees of freedom
## Multiple R-squared:  0.1954, Adjusted R-squared:  0.1944
## F-statistic: 182.9 on 4 and 3012 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Model 1: TARGET_deathRate ~ medIncome + PctWhite + MedianAge + PctWhite *
##     MedianAge
## Model 2: TARGET_deathRate ~ medIncome + MedianAge + PctWhite
##   Res.Df     RSS Df Sum of Sq     F Pr(>F)
## 1   3012 1866760
## 2   3013 1866763 -1   -2.4565 0.004 0.9498
```

Unfortunately, we can see that median age doesn’t really play any part as the effect modifier of median income or percentage of white population in the model. At this point, we can decide that for the linear model, we can clearly define the two core descriptor of social economic status is median income and percentage of white population. let’s expand a bit more here. First, I will include the precentage of white, black, asian, and other race percentage into the model to see if we can increase the predictive power of the model.

```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome + PctWhite + PctBlack +
##     PctAsian + PctOtherRace, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -118.383  -13.918    0.866   14.279  174.216
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.364e+02  5.931e+00  39.869  < 2e-16 ***
## medIncome    -8.450e-04  4.302e-05 -19.640  < 2e-16 ***
## PctWhite     -1.917e-01  6.064e-02  -3.161  0.00159 **
## PctBlack      1.204e-01  6.406e-02   1.879  0.06033 .
## PctAsian     -2.622e-01  2.130e-01  -1.231  0.21854
## PctOtherRace -1.395e+00  1.414e-01  -9.865  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 24.33 on 3011 degrees of freedom
## Multiple R-squared:  0.2317, Adjusted R-squared:  0.2305
## F-statistic: 181.7 on 5 and 3011 DF,  p-value: < 2.2e-16
```

We see that the adjusted R-squared value of model 1.4 is higher than model 1.3 and it is beneficial to include a fuller description of the racial distribution. However, we can also see that the Asian and Black percentages are not very significant. By definition, there should be some interactions between these percentages, and it is worth it to check it out.

```
## [1] -0.2658648
```

```
## [1] -0.8312116
```

```
## [1] -0.2331931
```

We see that the value of PctBlack and PctWhite are strongly negatively correlated, and the other two percentages are weakly correlated. Thus, we should remove PctBlack from the model to reduce colinearity.

```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome + PctWhite + PctAsian +
##     PctOtherRace, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -118.382  -13.873    0.839   14.226  174.620
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.463e+02  2.763e+00  89.160   <2e-16 ***
## medIncome    -8.482e-04  4.301e-05 -19.723   <2e-16 ***
## PctWhite     -2.905e-01  3.025e-02  -9.604   <2e-16 ***
## PctAsian     -3.827e-01  2.032e-01  -1.883   0.0598 .
## PctOtherRace -1.495e+00  1.310e-01 -11.413   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.34 on 3012 degrees of freedom
## Multiple R-squared:  0.2308, Adjusted R-squared:  0.2298
## F-statistic:   226 on 4 and 3012 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Model 1: TARGET_deathRate ~ medIncome + PctWhite + PctAsian + PctOtherRace
## Model 2: TARGET_deathRate ~ medIncome + PctWhite
##   Res.Df     RSS Df Sum of Sq     F    Pr(>F)
## 1   3012 1784592
## 2   3014 1867102 -2    -82510 69.63 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After removing the PctBlack covariate, we see that there is a significant increase in adjusted R-squared value to mod1.1(baseline) and LRT results suggest the models are significantly different from each other. Let's take another look at the median income covariates again. My main interest here is that whether a quadratic term would benefit our model, as a income could have diminishing return effect on lung cancer prevention/death rate.

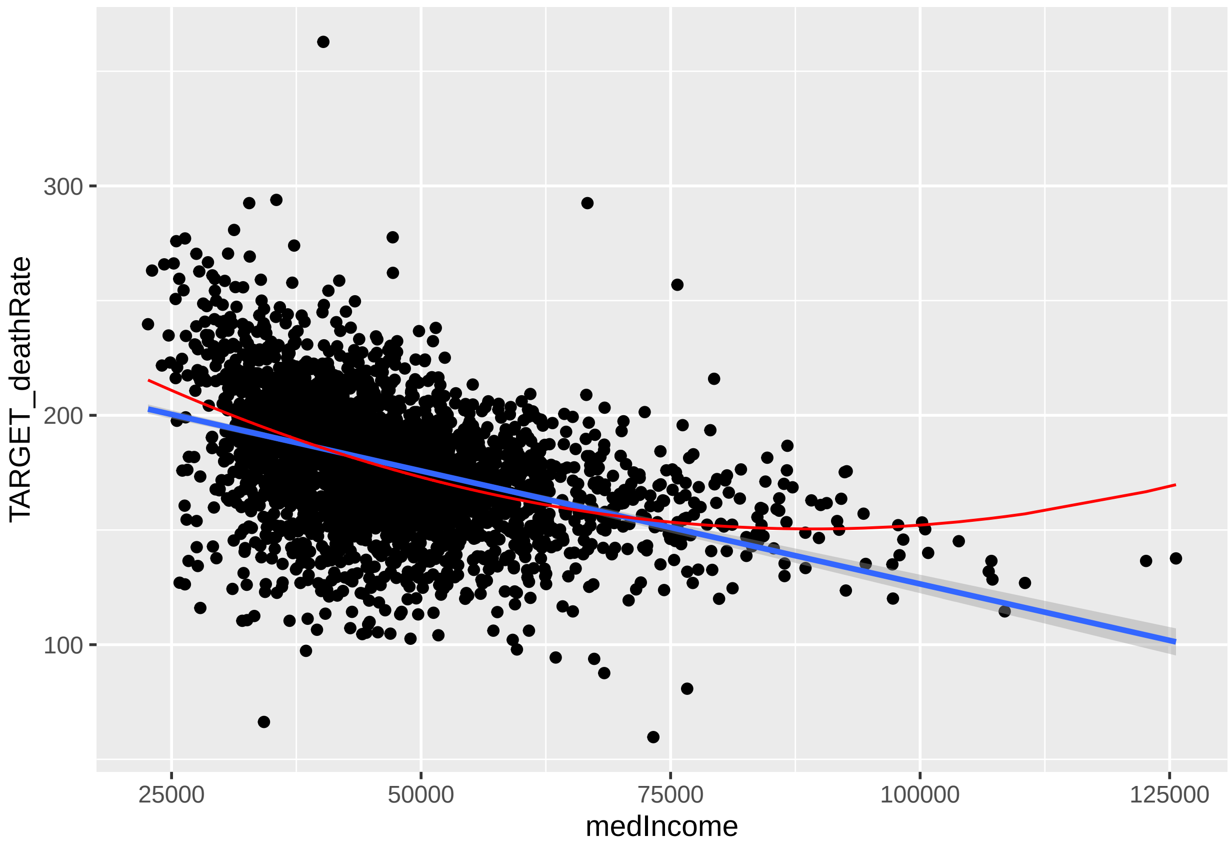
```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -124.962  -14.433    0.937   15.098  177.402
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.250e+02  1.840e+00  122.29   <2e-16 ***
## medIncome   -9.856e-04  3.788e-05  -26.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.07 on 3015 degrees of freedom
## Multiple R-squared:  0.1833, Adjusted R-squared:  0.1831
## F-statistic: 676.9 on 1 and 3015 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome + I(medIncome^2), data = dat)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max 
## -128.419  -13.923    1.128   14.799  177.132 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      2.670e+02  4.952e+00  53.913   <2e-16 ***
## medIncome       -2.609e-03  1.822e-04 -14.318   <2e-16 ***
## I(medIncome^2)  1.461e-08  1.605e-09   9.104   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 24.74 on 3014 degrees of freedom
## Multiple R-squared:  0.2052, Adjusted R-squared:  0.2047 
## F-statistic: 389.1 on 2 and 3014 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
## 
## Model 1: TARGET_deathRate ~ medIncome
## Model 2: TARGET_deathRate ~ medIncome + I(medIncome^2)
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)    
## 1   3015 1894793                                  
## 2   3014 1844082  1     50711 82.883 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## `geom_smooth()` using formula 'y ~ x'
```

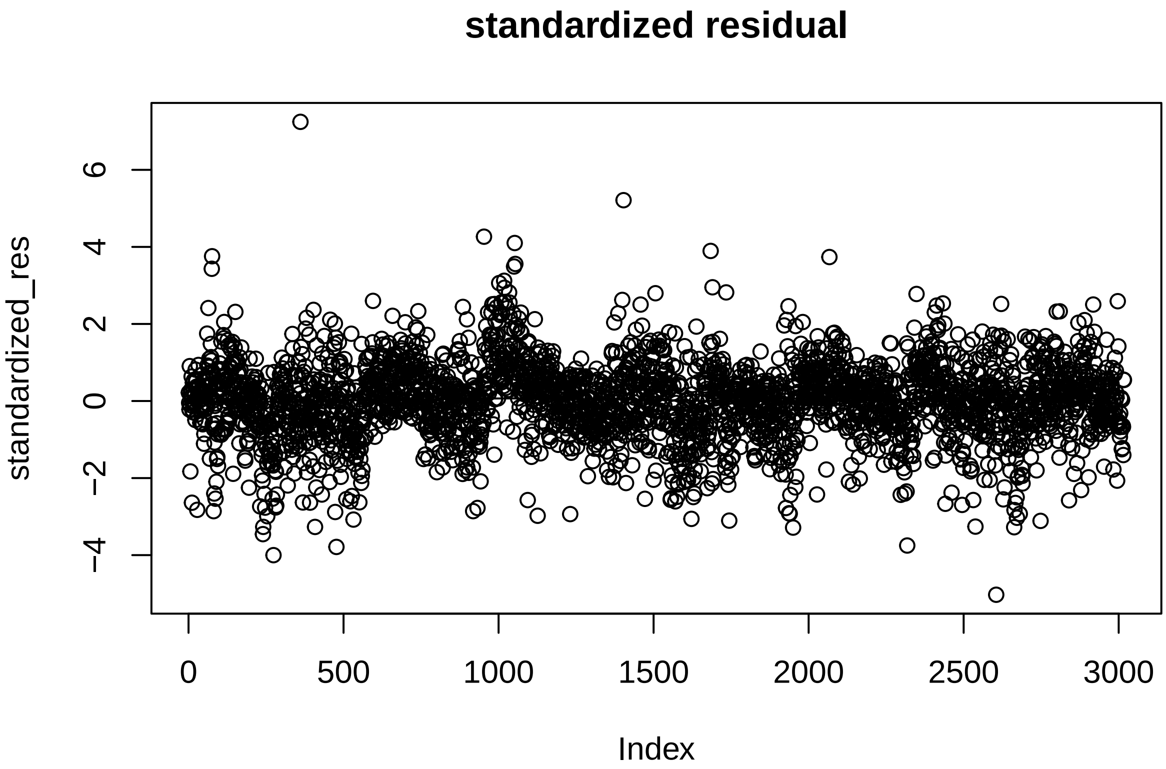


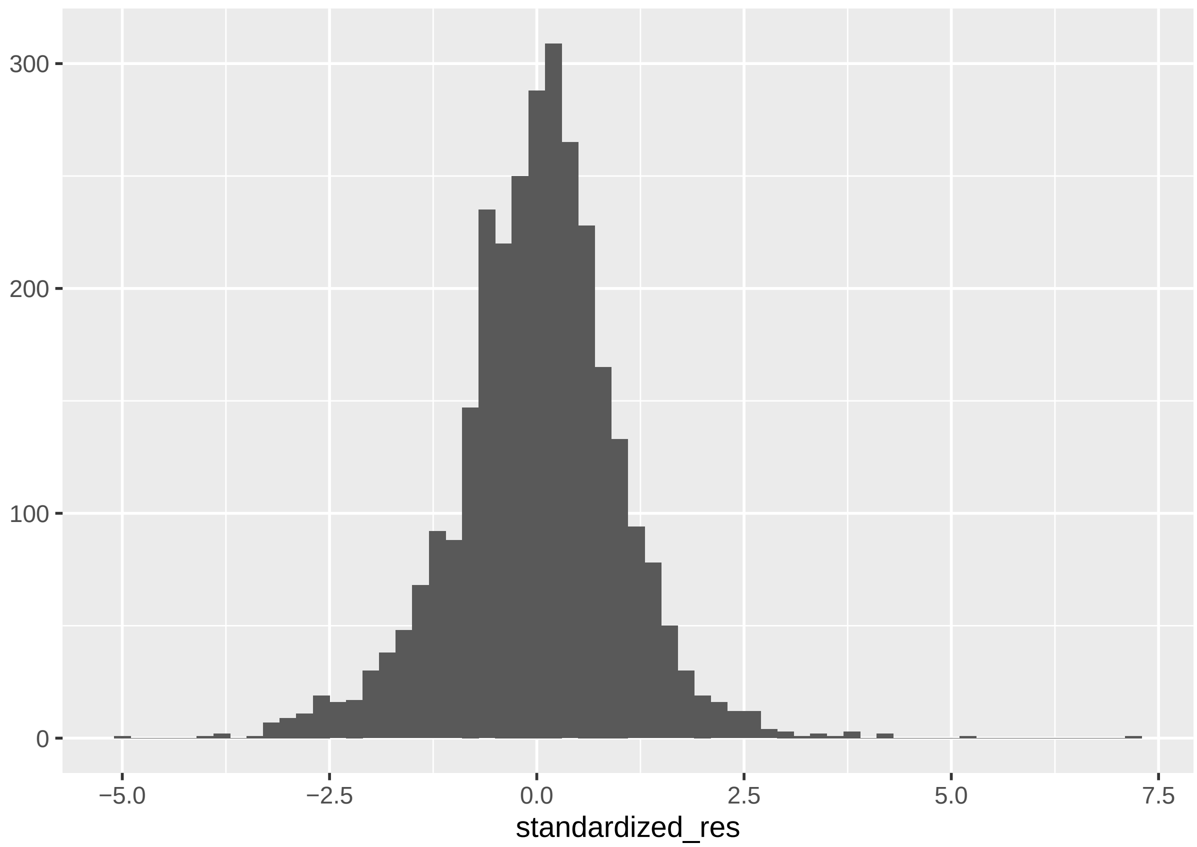
Judging from summary output, the quadratic model performs better than the purely linear model (based on adjusted R-squared value) and the two models are statistically different. However, a visual inspection of the data doesn't really show that the quadratic model is significantly better. At this point, we decide to keep the quadratic term due to the diagnostic stats.

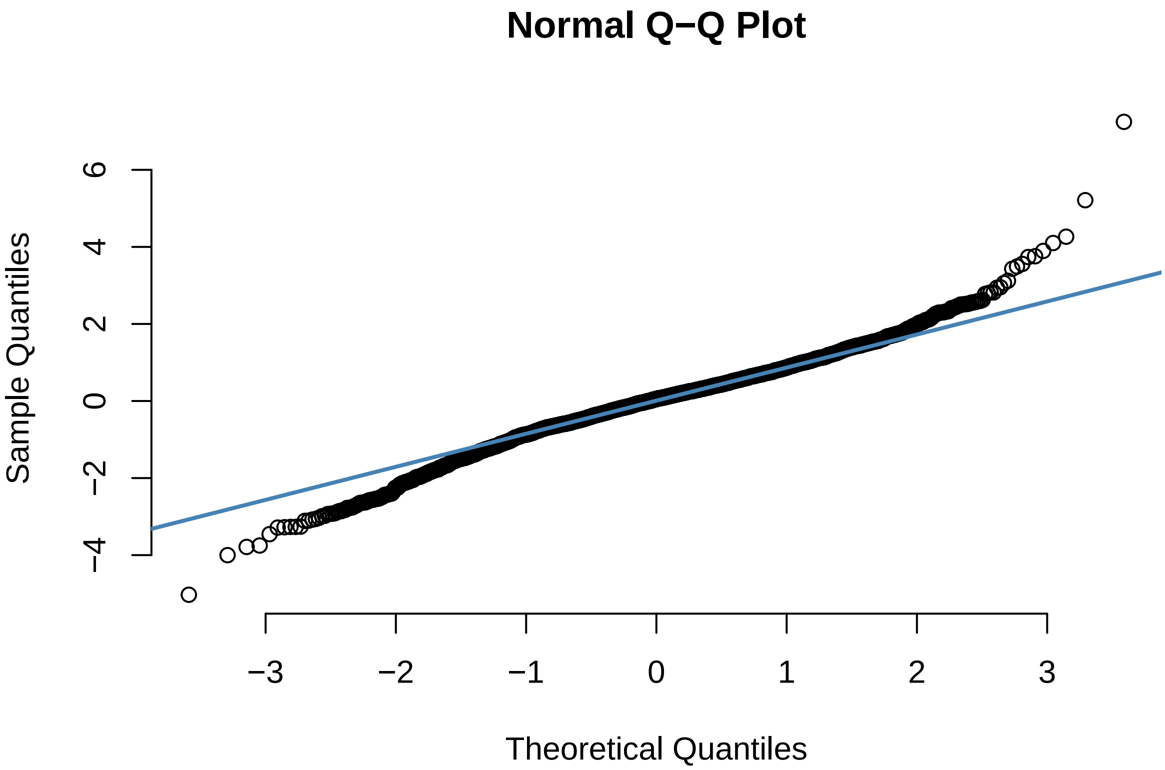
We now have this following core model:

```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome + I(medIncome^2) +
##     PctWhite + PctAsian + PctOtherRace, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -121.247  -13.750    1.302   14.195  174.871
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.756e+02  4.938e+00  55.821  < 2e-16 ***
## medIncome      -2.159e-03  1.885e-04 -11.456  < 2e-16 ***
## I(medIncome^2)  1.182e-08  1.656e-09   7.141 1.16e-12 ***
## PctWhite       -2.362e-01  3.095e-02  -7.632 3.08e-14 ***
## PctAsian       -5.664e-01  2.032e-01  -2.787  0.00535 **
## PctOtherRace   -1.419e+00  1.303e-01 -10.884  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 24.14 on 3011 degrees of freedom
## Multiple R-squared:  0.2437, Adjusted R-squared:  0.2424
## F-statistic:   194 on 5 and 3011 DF,  p-value: < 2.2e-16
```

let's evaluate the residual diagnostic to confirm the model's validity.


standardized residual
standardized_res
−4
−2
0
2
4
6
0
500
1000
1500
2000
2500
3000
Index


300
200
100
0
−5.0
−2.5
0.0
2.5
5.0
7.5
standardized_res



The model looks good, and the qqplot of the standardized residual plot only deviates from the straight line at the two edges. Thus, in conclusion, our core model is a good place to start for our future work.

future work and direction As mentioned in the missing data section, the education data is largely missing and we weren't able to directly add them to our analysis. We have acquired further information on the education level of each US county (data sourced from a 2020 study). However, the integration of the new data into our data set requires additional cleaning and this process will be conducted at a later time point. Furthermore, we are also looking to incorporate the insurance information of the US counties into the linear model.

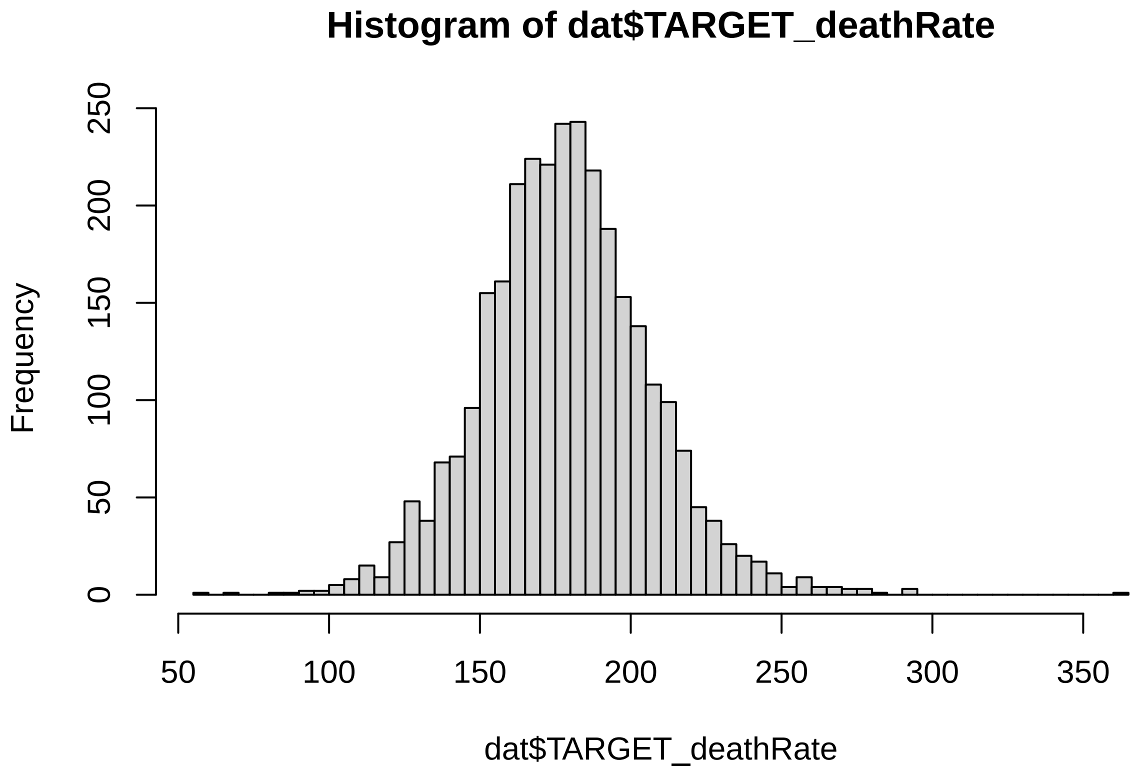
We are considering using GAM models in our future exploration as well. Using different smoothing methods could improve the performance of our models, and we are definitely interested in testing them out.

After we include all of the variables that we are interested in, we will preform Lasso, ridge, and elastic net models to reduce the overfitting and deciding on our final linear model. As we increase the number of predictor variables, we are expecting that some of them might be excluded from the final model during the either the lasso or the elastic net fitting process.

b. Logistic/multinomial/ordinal regression

Logistic/multinomial/ordinaln does not really fit into the exploration of our primary goals, as the lung-cancer death count can be mostly view as either a count statistics or a continuous outcome. However, there are several thing that we could do to incorporate Logistic/multinomial/ordinal regression into our project.

First, we can split the lung-cancer death rate into several categories to broadly access the healthcare system at each county. For example, we can artificially create three different categories in the death rate variable. Let's take a look at the death rate distribution.

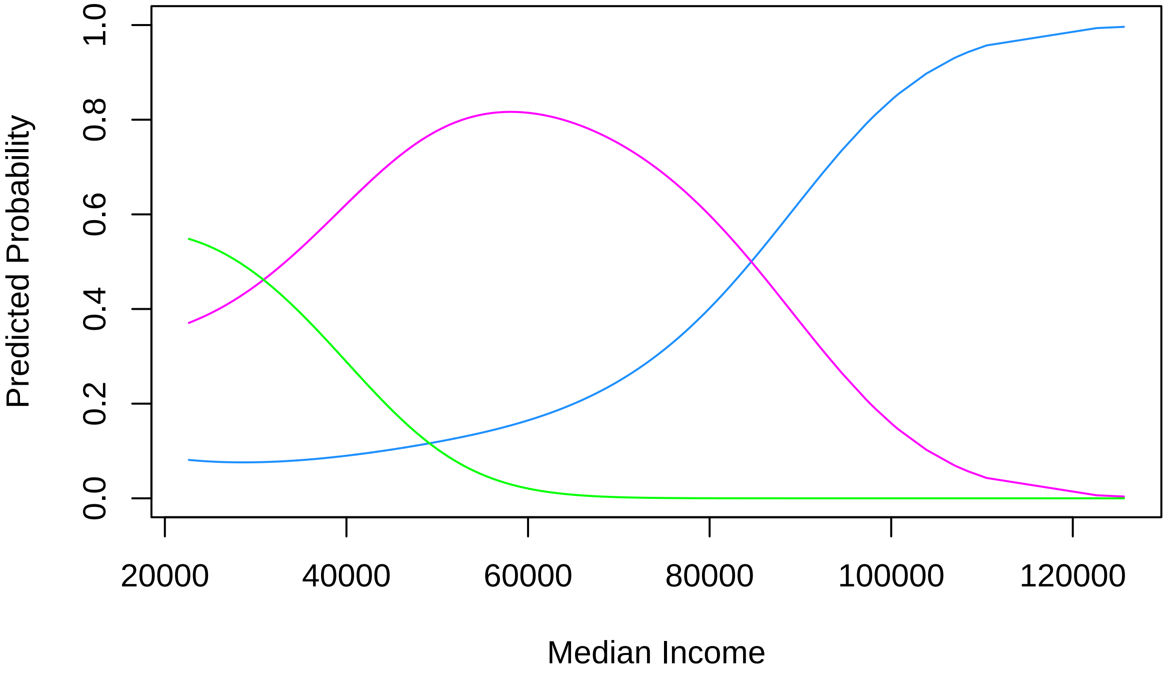


We can make three bins, deathrate < 150, 150 <= deathrate < 200, 200 <= deathrate, and use them as a proxy to the quality of lung cancer prevention and quality of cancer care for each US county. But in fairness, the use of logistic regression in this case wouldn't benefit us too much as it doesn't really bring any new insight to our analysis. The previous linear model already provides a fine baseline in assess the same information for each US county. The normal distribution of lung cancer death rate also suggests that dividing outcomes into bins doesn't really benefit us with our primary goal.

But let's do it anyway to test it out anyway.

```
## # weights:  12 (6 variable)
## initial  value 3314.513275
## iter  10 value 2320.489209
## final  value 2316.095848
## converged
```

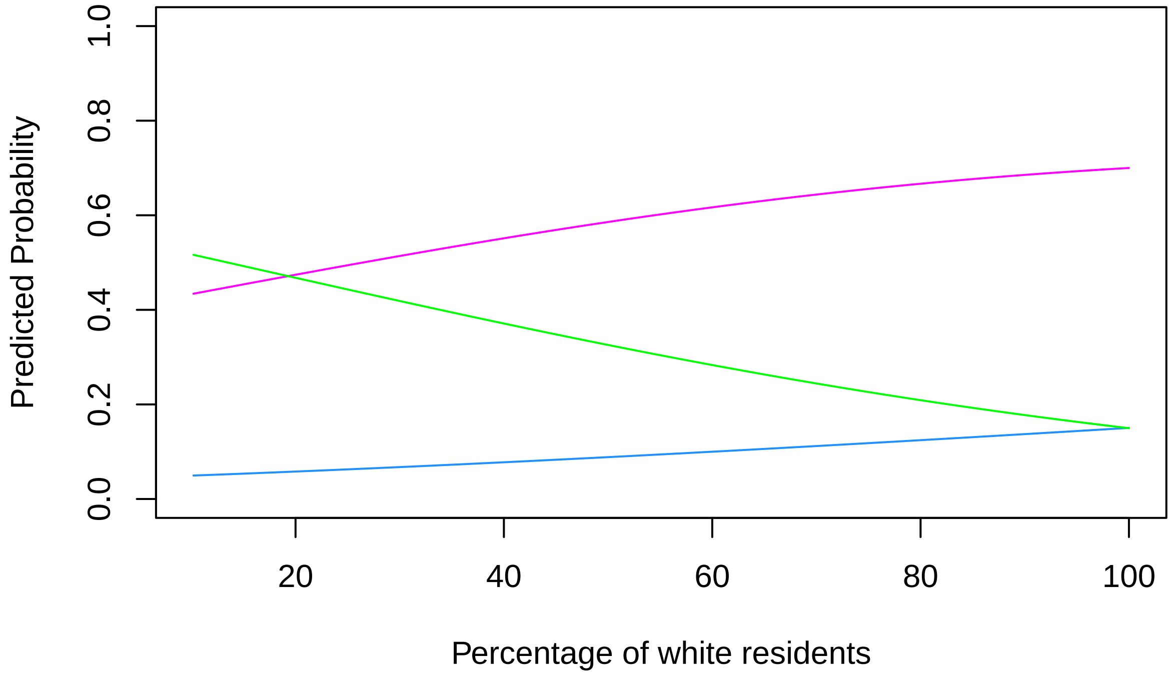
```
## Call:
## multinom(formula = multi ~ medIncome + I(medIncome^2), data = dat)
##
## Coefficients:
##    (Intercept)    medIncome I(medIncome^2)
## 2 2.417464e-09 9.165115e-05  -1.083448e-09
## 3 7.931352e-09 1.565121e-04  -3.185035e-09
##
## Std. Errors:
##    (Intercept)    medIncome I(medIncome^2)
## 2 6.249523e-21 2.523511e-16   1.799483e-11
## 3 1.659987e-20 7.128037e-16   3.143898e-11
##
## Residual Deviance: 4632.192
## AIC: 4644.192
```



we see that as median income increases, the county is much more likely to fall into the category 1, which is county with low lung cancer death rate. We can do the same with percentage of white citizens in the county as well.

```
## # weights:  9 (4 variable)
## initial  value 3314.513275
## iter  10 value 2552.935350
## iter  10 value 2552.935350
## final  value 2552.935350
## converged
```

```
## Call:
## multinom(formula = multi ~ PctWhite, data = dat)
##
## Coefficients:
##   (Intercept)     PctWhite
## 2    2.241079 -0.007027921
## 3    2.609593 -0.026135639
##
## Std. Errors:
##   (Intercept)    PctWhite
## 2   0.3398884 0.003901638
## 3   0.3612030 0.004210873
##
## Residual Deviance: 5105.871
## AIC: 5113.871
```



We see that percentage of white resident has less impact on the categorization of the county.

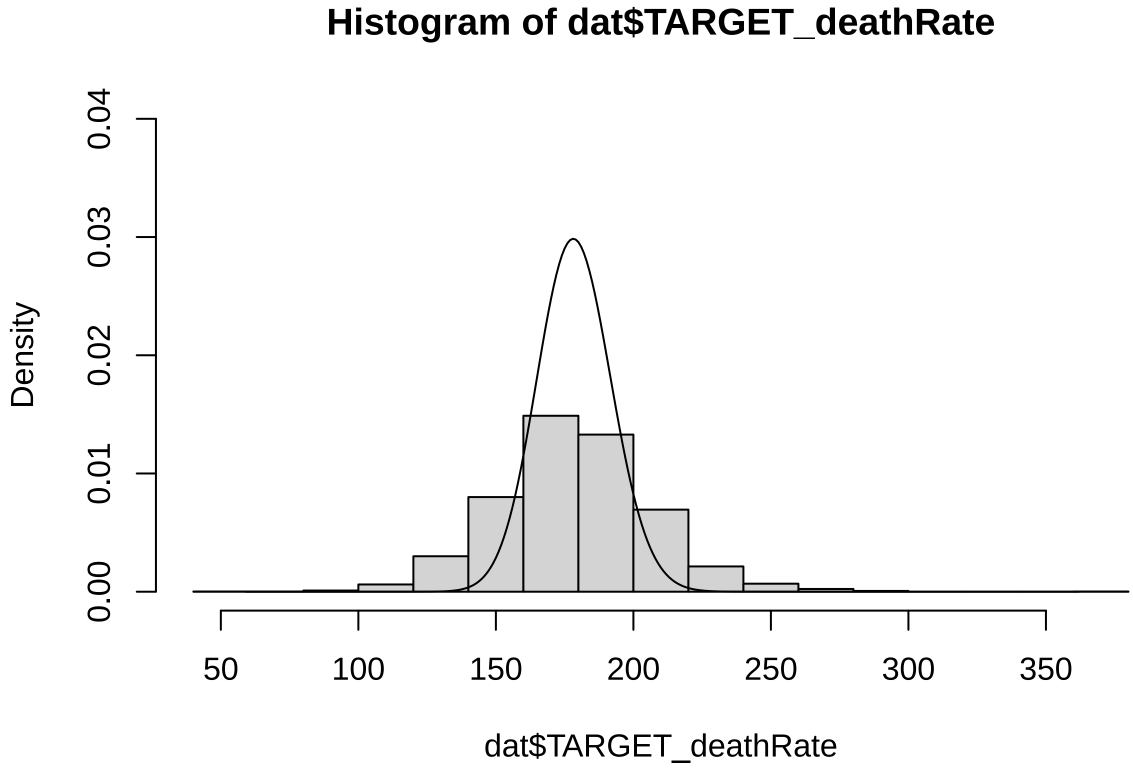
However, we can see that this type of multinomial regression is less interesting than the linear model, as we are grossly categorizing the outcome variables. Thus, we have decided to take the logistic regression to another direction. Logistic regression could serve as an excellent method for the exploration of our secondary goals.

One possible direction that we could take is to explore socioeconomic status and political tendencies. The US bipartisan system provides a excellent binary outcome for our logistic models. We could decide whether what would drive the county to vote for either candidates. To explore this topic, we would need to find additional data on the voting outcomes of 2016 on a county level and merge with our existing data set. This exploration will have to be conducted at a later time after we find a trustworthy and accurate data source. Currently, the politico site serves as the preliminary source for our data as it contains detailed county level voting outcomes. Source: https://www.politico.com/2016-election/results/map/president/.

Additional data wrangling is required before we move forward with the Logistic/multinomial/ordinal regression analysis. Our current plan for the regressions invovles using the covariates from the previous linear models and develop our model from there. We think that the logistic regression model would be helpful in showing the political preference of different demographic groups.

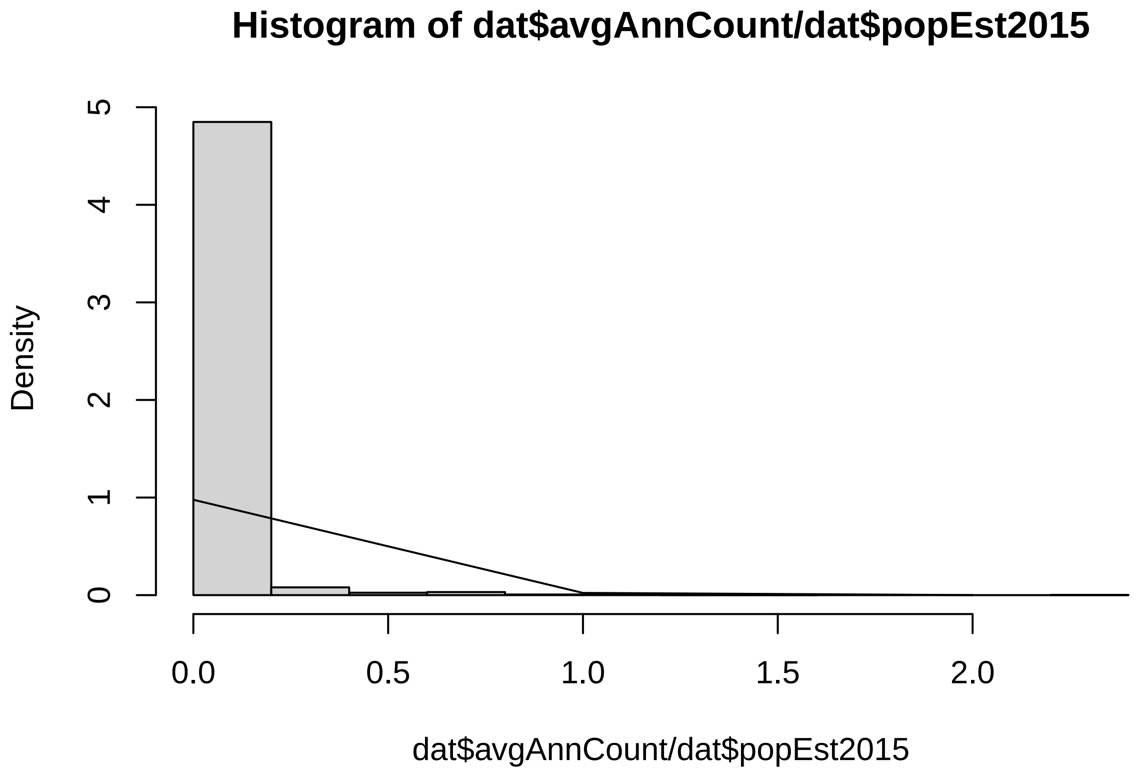
c. Poisson Regression

Over-dispersion



```
## [1] 178.6452
```

```
## [1] 769.2961
```

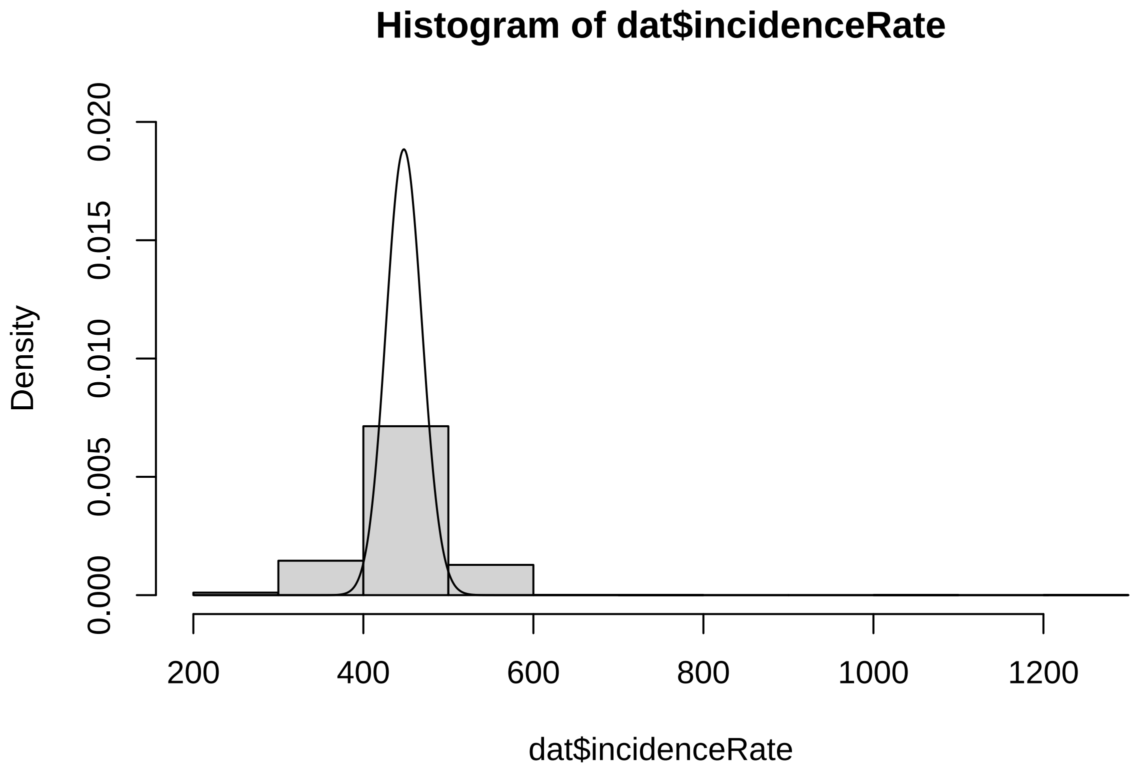


```
mean(dat$avgDeathsPerYear/dat$popEst2015)
```

```
## [1] 0.002287129
```

```
var(dat$avgDeathsPerYear/dat$popEst2015)
```

```
## [1] 3.729806e-07
```



```
## [1] 448.1764
```

```
## [1] 2982.145
```

We notice for each of our potential target variables, while the Poisson density roughly fits, they are overdispersed. This suggests the potential of modelling our outcome of interest using an extension of the Poisson model. Due to not having zero-inflated data, this narrows down potentially modelling approaches to the Negative Binomial regression, which should account for over-dispersion, to make sure we do not have overconfident estimates. For these data, we do not need to account for lag as they are all averaged over the same time period for each covariate and outcome respectively.

We would not need to modify our data at all to perform a Negative Binomial regression as our data is already in counts or rate form. Although one potential modification we might do is see how the regression turns out modelling the rate per 100 000 (our target of interest), and the pure count itself adjusting for population.

Model fits

```
##
## Call:
## glm(formula = TARGET_deathRate ~ medIncome + State + popEst2015,
##     family = poisson(), data = .)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -9.7899  -0.9173   0.0055   0.9062  11.6936
```

```
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.548e+00  1.957e-02 283.437  < 2e-16 ***
## medIncome   -4.502e-06  1.435e-07 -31.378  < 2e-16 ***
## StateAL     -1.111e-01  1.995e-02  -5.571 2.54e-08 ***
## StateAR     -7.953e-02  1.958e-02  -4.062 4.86e-05 ***
## StateAZ     -3.451e-01  2.765e-02 -12.482  < 2e-16 ***
## StateCA     -2.272e-01  2.063e-02 -11.015  < 2e-16 ***
## StateCO     -3.604e-01  2.065e-02 -17.454  < 2e-16 ***
## StateCT     -1.658e-01  3.318e-02  -4.998 5.78e-07 ***
## StateDC     -1.967e-02  5.528e-02  -0.356 0.722031
## StateDE     -9.943e-02  4.658e-02  -2.134 0.032813 *
## StateFL     -1.562e-01  1.995e-02  -7.829 4.90e-15 ***
## StateGA     -1.527e-01  1.871e-02  -8.160 3.35e-16 ***
## StateHI     -3.009e-01  4.533e-02  -6.637 3.19e-11 ***
## StateIA     -2.000e-01  1.917e-02 -10.433  < 2e-16 ***
## StateID     -3.036e-01  2.154e-02 -14.092  < 2e-16 ***
## StateIL     -1.067e-01  1.901e-02  -5.612 2.00e-08 ***
## StateIN     -8.772e-02  1.911e-02  -4.589 4.46e-06 ***
## StateKS     -2.028e-01  1.914e-02 -10.591  < 2e-16 ***
## StateKY      6.348e-04  1.881e-02   0.034 0.973082
## StateLA     -7.386e-02  1.984e-02  -3.723 0.000197 ***
## StateMA     -1.453e-01  2.723e-02  -5.335 9.54e-08 ***
## StateMD     -7.295e-02  2.326e-02  -3.136 0.001710 **
## StateME     -1.317e-01  2.552e-02  -5.161 2.45e-07 ***
## StateMI     -1.645e-01  1.948e-02  -8.443  < 2e-16 ***
## StateMN     -2.181e-01  1.941e-02 -11.237  < 2e-16 ***
## StateMO     -1.165e-01  1.896e-02  -6.144 8.04e-10 ***
## StateMS     -7.530e-02  1.948e-02  -3.864 0.000111 ***
## StateMT     -2.559e-01  2.101e-02 -12.182  < 2e-16 ***
## StateNC     -1.783e-01  1.928e-02  -9.248  < 2e-16 ***
## StateND     -2.243e-01  2.068e-02 -10.847  < 2e-16 ***
## StateNE     -2.474e-01  1.968e-02 -12.571  < 2e-16 ***
## StateNH     -1.354e-01  2.982e-02  -4.540 5.63e-06 ***
## StateNJ     -9.329e-02  2.430e-02  -3.839 0.000123 ***
## StateNM     -3.131e-01  2.282e-02 -13.720  < 2e-16 ***
## StateNV     -1.216e-01  2.525e-02  -4.815 1.47e-06 ***
## StateNY     -1.560e-01  2.004e-02  -7.781 7.17e-15 ***
## StateOH     -9.885e-02  1.924e-02  -5.138 2.77e-07 ***
## StateOK     -8.282e-02  1.945e-02  -4.259 2.05e-05 ***
## StateOR     -1.967e-01  2.184e-02  -9.009  < 2e-16 ***
## StatePA     -1.547e-01  1.993e-02  -7.761 8.40e-15 ***
## StateRI     -1.563e-01  4.256e-02  -3.673 0.000239 ***
## StateSC     -1.265e-01  2.073e-02  -6.100 1.06e-09 ***
## StateSD     -2.384e-01  2.035e-02 -11.720  < 2e-16 ***
## StateTN     -6.076e-02  1.916e-02  -3.172 0.001515 **
## StateTX     -1.964e-01  1.830e-02 -10.731  < 2e-16 ***
## StateUT     -3.846e-01  2.404e-02 -15.999  < 2e-16 ***
## StateVA     -1.029e-01  1.873e-02  -5.491 3.99e-08 ***
## StateVT     -1.459e-01  2.668e-02  -5.468 4.54e-08 ***
## StateWA     -2.024e-01  2.149e-02  -9.420  < 2e-16 ***
## StateWI     -1.687e-01  1.968e-02  -8.569  < 2e-16 ***
## StateWV     -8.860e-02  2.018e-02  -4.390 1.13e-05 ***
## StateWY     -2.220e-01  2.404e-02  -9.235  < 2e-16 ***
## popEst2015  -1.060e-08  4.808e-09  -2.205 0.027479 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 13026.8  on 3016  degrees of freedom
## Residual deviance:  7555.3  on 2964  degrees of freedom
## AIC: Inf
##
## Number of Fisher Scoring iterations: 4
```

```
##
## Call:
## MASS::glm.nb(formula = TARGET_deathRate ~ medIncome + State +
##     popEst2015, data = ., init.theta = 118.3508188, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.7153  -0.5830   0.0045   0.5662   6.8103
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.542e+00  3.153e-02 175.761  < 2e-16 ***
## medIncome   -4.428e-06  2.224e-07 -19.913  < 2e-16 ***
## StateAL     -1.088e-01  3.236e-02  -3.361 0.000776 ***
## StateAR     -7.718e-02  3.182e-02  -2.426 0.015280 *
## StateAZ     -3.421e-01  4.296e-02  -7.963 1.68e-15 ***
## StateCA     -2.255e-01  3.298e-02  -6.837 8.11e-12 ***
## StateCO     -3.588e-01  3.273e-02 -10.964  < 2e-16 ***
## StateCT     -1.662e-01  5.159e-02  -3.221 0.001277 **
## StateDC     -1.972e-02  8.826e-02  -0.223 0.823198
## StateDE     -9.806e-02  7.408e-02  -1.324 0.185601
## StateFL     -1.549e-01  3.222e-02  -4.808 1.52e-06 ***
## StateGA     -1.501e-01  3.030e-02  -4.954 7.28e-07 ***
## StateHI     -3.001e-01  6.833e-02  -4.392 1.12e-05 ***
## StateIA     -1.985e-01  3.091e-02  -6.422 1.34e-10 ***
## StateID     -3.020e-01  3.423e-02  -8.822  < 2e-16 ***
## StateIL     -1.046e-01  3.078e-02  -3.399 0.000676 ***
## StateIN     -8.608e-02  3.099e-02  -2.778 0.005476 **
## StateKS     -2.014e-01  3.088e-02  -6.521 6.97e-11 ***
## StateKY      2.029e-03  3.063e-02   0.066 0.947177
## StateLA     -7.152e-02  3.223e-02  -2.219 0.026472 *
## StateMA     -1.436e-01  4.300e-02  -3.339 0.000842 ***
## StateMD     -7.193e-02  3.733e-02  -1.927 0.053977 .
## StateME     -1.299e-01  4.107e-02  -3.163 0.001559 **
## StateMI     -1.627e-01  3.147e-02  -5.169 2.36e-07 ***
## StateMN     -2.161e-01  3.120e-02  -6.925 4.35e-12 ***
## StateMO     -1.144e-01  3.073e-02  -3.723 0.000197 ***
## StateMS     -7.231e-02  3.167e-02  -2.283 0.022432 *
## StateMT     -2.536e-01  3.359e-02  -7.548 4.43e-14 ***
## StateNC     -1.760e-01  3.116e-02  -5.650 1.61e-08 ***
## StateND     -2.228e-01  3.306e-02  -6.739 1.60e-11 ***
## StateNE     -2.452e-01  3.160e-02  -7.760 8.50e-15 ***
## StateNH     -1.336e-01  4.725e-02  -2.828 0.004683 **
## StateNJ     -9.247e-02  3.869e-02  -2.390 0.016846 *
## StateNM     -3.105e-01  3.615e-02  -8.589  < 2e-16 ***
## StateNV     -1.222e-01  4.045e-02  -3.022 0.002511 **
## StateNY     -1.541e-01  3.227e-02  -4.774 1.81e-06 ***
## StateOH     -9.731e-02  3.117e-02  -3.122 0.001799 **
## StateOK     -8.144e-02  3.157e-02  -2.580 0.009875 **
## StateOR     -1.942e-01  3.500e-02  -5.549 2.88e-08 ***
## StatePA     -1.528e-01  3.214e-02  -4.752 2.01e-06 ***
## StateRI     -1.558e-01  6.651e-02  -2.342 0.019186 *
## StateSC     -1.248e-01  3.357e-02  -3.717 0.000202 ***
## StateSD     -2.368e-01  3.261e-02  -7.263 3.78e-13 ***
## StateTN     -5.883e-02  3.114e-02  -1.890 0.058816 .
## StateTX     -1.936e-01  2.963e-02  -6.535 6.36e-11 ***
## StateUT     -3.825e-01  3.729e-02 -10.257  < 2e-16 ***
## StateVA     -1.011e-01  3.034e-02  -3.331 0.000865 ***
## StateVT     -1.434e-01  4.263e-02  -3.364 0.000768 ***
## StateWA     -2.004e-01  3.437e-02  -5.830 5.53e-09 ***
## StateWI     -1.674e-01  3.174e-02  -5.276 1.32e-07 ***
## StateWV     -8.637e-02  3.278e-02  -2.635 0.008423 **
## StateWY     -2.196e-01  3.800e-02  -5.778 7.58e-09 ***
## popEst2015  -1.010e-08  7.324e-09  -1.379 0.168041
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(118.3508) family taken to be 1)
##
##     Null deviance: 5232.5  on 3016  degrees of freedom
## Residual deviance: 3051.8  on 2964  degrees of freedom
## AIC: 27080
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  118.35
##           Std. Err.:  5.12
##
##  2 x log-likelihood:  -26971.87
```

Here we find that the negative binomial regression has slightly higher standard errors and resulting test statistics compared to the standard Poisson regression. This is as we expected since it suggests that the standard Poisson regression may cause overconfidence in model evaluations, due to the property of $\mathbb{E}[Y] = Var(Y)$ being unsatisfied.

Furthermore, we can interpret the Negative Binomial regression supposing median income is our covariate of interest. For a 1000 dollar increase in median income, a county has an estimated `exp(coef(state_inc_pop_nb)[[2]])` times the incident rate compared to if a county did not have a 1000 dollar increase in median income, on average, holding all other covariates fixed. This was a significant relationship as well, suggesting that median income may be negatively associated with the incidence rate of lung cancer deaths, however this is only a very slight relationship at `exp(coef(state_inc_pop_nb)[[2]])` with these covariates in this model, according to these data.

d. Survival Analysis

We are not going to incorporate survival analysis into our project. This is because we do not have any time-to-event related relationships in our data. Furthermore, we cannot really reproduce these relationships using a modification of our variables since they are all aggregates over a time period, and so we cannot infer the individual-level observations to re-create time-to-event data.

Abstract

Lung cancer is one of the leading causes of global cancer incidence and mortality, accumulating over 1.8 million deaths each year. Lung cancer has become the second most common forms of cancer diagnosis for both men and women (Thandra et al). Although the incidence rate and mortality rate of lung cancer has been decreasing in the US due to public education and tobacco control policies (De groot et al), a better understanding of the disease within the socialeconomic and demographic context is still required to provide future guidance on combating the prevalent form of cancer. Data from US census, county level reports, and various government agencies were collected to create tabular data on lung cancer death rate, social economic status, demographic information, education level, and medical insurance information on the US county level. The primary goal of the project is to create a explanatory model between social economic status and the lung cancer death rate of each US county. A secondary goal of this project is to predict the political inclination of each county base on the existing data set using logistic regression models.

Our primary goal would be achieved using two regression models, linear regression and poisson regression. The two methods were chosen due to possible different interpretation of the interested outcome, the lung cancer related death rate of each US county. Comprehensive models from the two methods were constructed and the final models are shown below:

insert final model here

The comparison of the two models results in the selection of _____ as the final model. We have found several significant factors within the model that provide significant context and utility to increasing health care equity in the United states. There factors and their interpretation are listed below:

insert table of important factors here. Will include name of covariates, model coefficient, interpretation

We also found some additional conclusions without our secondary goals. And these are our findings.

conclusion section We will talk about the potential implication of our findings in a public health context, what future directions that we could take. And any potential shortcomings of our approach.

Citations:

1. Thandra, Krishna Chaitanya et al. "Epidemiology of lung cancer." Contemporary oncology (Poznan, Poland) vol. 25,1 (2021): 45-52. doi:10.5114/wo.2021.103829
2. de Groot, P. M., Wu, C. C., Carter, B. W., & Munden, R. F. (2018). The epidemiology of lung cancer. Translational lung cancer research, 7(3), 220–233. https://doi.org/10.21037/tlcr.2018.05.06

Question 6

Part b

Introduction/Background:

Lung cancer is one of the leading causes of global cancer incidence and mortality, accumulating over 1.8 million deaths each year. Lung cancer has become the second most common form of cancer diagnosis for both men and women (Thandra et al). The death rate of lung cancer is almost equal to the death rate of prostate, breast, and colon cancer combined in the United States (Dela Cruz and Tanoue).

One of the known leading causes of lung cancer is cigarette and tobacco smoking. Not only does exposure to smoking suggest a strong link with lung cancer onset, but first- and second-hand exposure to tobacco smoke can have a genetic influence on individuals. The susceptibility of genetic markers to lung carcinogens as well as acquired epigenetic polymorphisms can drastically increase one's chances of being diagnosed with lung cancer. A family history of cancer, especially lung cancer, can also increase the risk for lung cancer in both smokers and non-smokers (Dela Cruz and Tanoue).

Since the 1970s, there has been a drastic decrease in tobacco use in the US. Additionally, the exacerbated efforts to find genetic influences associated with lung cancer in recent years and the push for scientific insight into chemotherapy and pharmaceutical drugs to potential lung cancer treatments, have helped healthcare professionals find better ways to target and improve lung cancer prognosis (Lemjabbar-Alaoui et al). Therefore, the incidence rate and mortality rate of lung cancer have been decreasing in the US due to public education and tobacco control policies (De groot et al). However, a better understanding of the disease within the socioeconomically and demographic context is still required to provide future guidance on combating the prevalent form of cancer. With lower tobacco smoking rates in the US and an increasing prevalence of lung cancer in non-smokers, environmental factors as well as socioeconomic influences need to be further investigated.

Data description and motivation:

Data from the US census, county-level reports, and various government agencies were collected to create tabular data on lung cancer death rate, social economic status, demographic information, education level, and medical insurance information at the US county level. Specifically, the data used and analyzed is from "OLS Regression Challenge - dataset by nrippner | data.world", which aggregates socioeconomic and clinical data from census.gov, clinicaltrials.gov, and cancer.gov. These cancer outcomes data were aggregated from cancer trials from 01/01/2010 through 06/01/2016 and socioeconomic and demographic data were aggregated from 2013 U.S. census data. Additionally, data from the United States Department of Agriculture (USDA) was used to assess educational attainment as a potential factor associated with socioeconomic status.

Given the lack of research on socioeconomic factors on cancer prevalence in the United States specifically, the dataset seemed appropriate for providing insight into the potential links between socioeconomic status and cancer overall. The primary goal of this study is to create an explanatory model between social economic status and the lung cancer death rate of each US county. A secondary goal of this project is to predict the political inclination of each county base on the existing data set using logistic regression models. A separate data set from Politico was merged with the data from the US census to assess the secondary questions.

Citations

1. Thandra, Krishna Chaitanya et al. "Epidemiology of lung cancer." Contemporary oncology (Poznan, Poland) vol. 25,1 (2021): 45-52. doi:10.5114/wo.2021.103829

2. Dela Cruz CS, Tanoue LT, Matthay RA. Lung cancer: epidemiology, etiology, and prevention. Clin Chest Med. 2011 Dec;32(4):605-44. doi: 10.1016/j.ccm.2011.09.001. PMID: 22054876; PMCID: PMC3864624.

3. Lemjabbar-Alaoui H, Hassan OU, Yang YW, Buchanan P. Lung cancer: Biology and treatment options. Biochim Biophys Acta. 2015 Dec;1856(2):189-210. doi: 10.1016/j.bbcan.2015.08.002. Epub 2015 Aug 19. PMID: 26297204; PMCID: PMC4663145.

4. de Groot, P. M., Wu, C. C., Carter, B. W., & Munden, R. F. (2018). The epidemiology of lung cancer. Translational lung cancer research, 7(3), 220–233. https://doi.org/10.21037/tlcr.2018.05.06

Part c

Research and Analysis Methods:

The data employed 34 potential covariates of interest related to socioeconomic and demographic factors. For our primary analysis, a Poisson regression model is employed to estimate the socioeconomic status-related incidence and mortality rates of lung cancer from our data. Furthermore, the data were stratified by state level to understand the state-by-state differences in socioeconomic conditions and lung cancer outcomes. To account for overdispersion, an extension of the Poisson model, the Negative Binomial regression, was used to prevent overconfident estimates. No further modifications were made to the data set to perform a Negative Binomial regression because the data was already in counts or rate form.

Once we have identified some covariates of interest, various smoothing methods were utilized in order to determine the relative correlations in the data. (Please note, we have not finalized yet at this point in our project. However, one of the covariates we are considering is median income and the percentage of the white population within a county because it is significantly correlated with the lung cancer death rate. We also need to further study the potential confounders and effect modifiers within our potential covariates in our data set – median age was considered in this project check-in. For the covariates we are not interested in interpreting in our final project, we plan to adjust for them flexibly using splines/GAMs).

For our secondary analysis, multinomial regression models were employed to model the probability of some set of socioeconomic conditions and cancer outcomes being in a given county or state in order to better understand state-level and county-level differences in demographics and insurance status. To investigate whether the political tendencies of different demographic groups could impact lung cancer outcomes, the logistic linear regression model was used.

All the statistical analyses were carried out with R. The graphical visualization of the results was done in R using basic R and the package “tidyverse.”

Note: We did not feel that we were ready to outline any results or discussion yet in our project.

Question 8

At this moment in our project, we are not trying to reach a publication involving the work or results from this project. However, we agree that our research questions are data are important in the context of the research field of understanding connections between SES and cancer. That being said, if our project ends up resulting in interesting findings, we would be open to the idea of potentially publishing our results in some capacity.