stat final project workplace

```
suppressMessages(library(tidyverse))
suppressMessages(library(usmap))
suppressMessages(library(scales))
suppressMessages(library(mice))
suppressMessages(library(glmnet))
suppressMessages(library(boot))
suppressMessages(library(grid))
suppressMessages(library(gridExtra))
```

```
propo = read.csv("cancer_reg.csv")
```

```
testpropo = propo
testpropo <- testpropo %>% mutate(Target_div_Income = TARGET_deathRate/medIncome)
```

```
testpropo1 = cbind(testpropo, str_match(testpropo$Geography,"(.+), (.+)")[ ,-1])
colnames(testpropo1)[37] ="State"
colnames(testpropo1)[36] = "County"
testpropo1[167,36] <- "Dona Ana County"
testpropo1[821,36] <- "La Salle Parish"
```

```
codes <- rep(NULL, length(testpropo1$County))

for (i in 1:length(testpropo1$avgAnnCount)){
 codes[i] = fips(state = testpropo1$State[i], county = testpropo1$County[i])
}
```

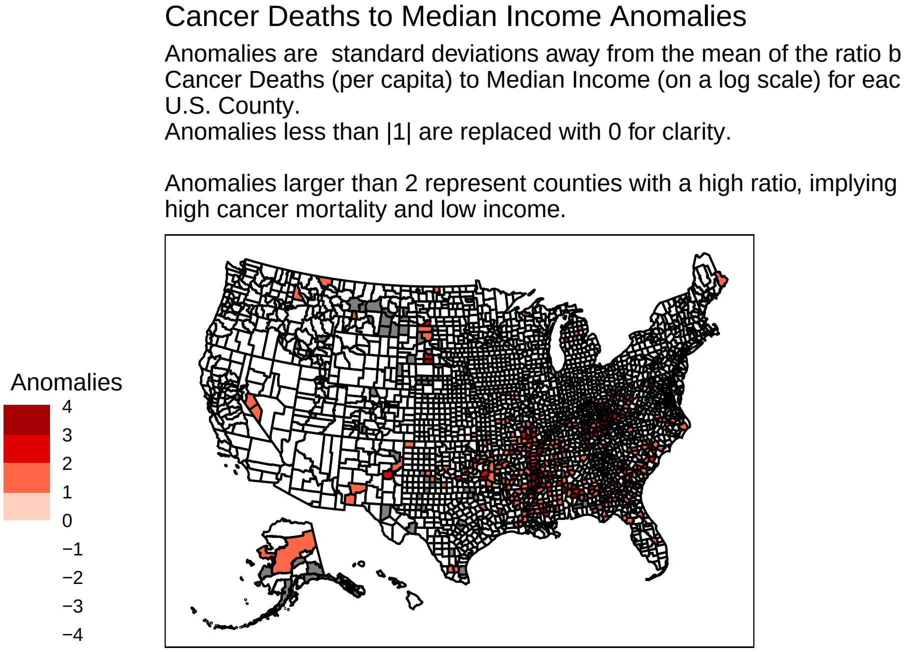
```
testpropo2 = cbind(testpropo1, fips = codes)
graphdata = data.frame(fips = testpropo2$fips, values = scale(testpropo2$Target_div_Income
newbie <- graphdata %>% mutate(anomalies = ifelse(abs(values) > 1, values, 0))
newbie <- newbie[,c(1,3)]
```

New attempt, log ratio things instead of using scale

```
testpropo3 <- testpropo2 %>% mutate(Target_div_LogIncome = TARGET_deathRate/log(medIncome)
testpropolog = cbind(testpropo3, fips = codes)
graphdatalog = data.frame(fips = testpropolog$fips, values = testpropo3$Target_div_LogInco
```

```
newbieLOG <- graphdata %>% mutate(anomalies = ifelse(abs(scale(values)) > 1, values, 0))
newbieLOG <- newbieLOG[,c(1,3)]
```

```
plot_usmap(data = newbieLOG, values = "anomalies") +
  scale_fill_stepsn(breaks= -4:4, limits = c(-4,4),
                    colors=c("white","white", "white","red","dark red"),
                    guide = guide_colorsteps(even.steps = FALSE), name = " Anomalies") +
  theme(panel.background = element_rect(color = "black")) +
  theme(legend.position = "left") + labs(title = "Cancer Deaths to Median Income Anomalies
       subtitle = "Anomalies are  standard deviations away from the mean of the ratio betw
       \nAnomalies larger than 2 represent counties with a high ratio, implying \nhigh can
```



Cancer Deaths to Median Income Anomalies

Anomalies are standard deviations away from the mean of the ratio b
Cancer Deaths (per capita) to Median Income (on a log scale) for eac
U.S. County.
Anomalies less than |1| are replaced with 0 for clarity.

Anomalies larger than 2 represent counties with a high ratio, implying
high cancer mortality and low income.

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

Data Pre processing - include everything up to testpropo3

```
moddat <- testpropo3

(colMeans(is.na(moddat)))*100
```

```
            avgAnnCount        avgDeathsPerYear        TARGET_deathRate 
               0.000000                0.000000                0.000000 
          incidenceRate               medIncome              popEst2015 
               0.000000                0.000000                0.000000 
         povertyPercent             studyPerCap               binnedInc 
               0.000000                0.000000                0.000000 
              MedianAge           MedianAgeMale         MedianAgeFemale 
               0.000000                0.000000                0.000000 
              Geography        AvgHouseholdSize          PercentMarried 
               0.000000                0.000000                0.000000 
```

```
           PctNoHS18_24              PctHS18_24         PctSomeCol18_24
               0.000000                0.000000               74.991795
        PctBachDeg18_24            PctHS25_Over       PctBachDeg25_Over
               0.000000                0.000000                0.000000
     PctEmployed16_Over    PctUnemployed16_Over      PctPrivateCoverage
               4.988513                0.000000                0.000000
PctPrivateCoverageAlone      PctEmpPrivCoverage       PctPublicCoverage
              19.986872                0.000000                0.000000
 PctPublicCoverageAlone                PctWhite                PctBlack
               0.000000                0.000000                0.000000
               PctAsian            PctOtherRace    PctMarriedHouseholds
               0.000000                0.000000                0.000000
              BirthRate       Target_div_Income                  County
               0.000000                0.000000                0.000000
                  State                    fips    Target_div_LogIncome
               0.000000                0.000000                0.000000
```

Since PctSomeCol18_24 has a NA rate of 74.99%, and represents the inbetween between high school diploma and bachelors, we can justify excluding it.

PctEmployed16_Over has only a 4.99% NA rate, and PctPublicCoverageAlone, which is the percentage of county residents with government-provided health coverage alone, has a 19.99% NA rate, but seems too important to ignore if we wish to consider the status of coverage as a variable(s).

Let us do MICE (Multiple Imputation by Chained Equations) to replace these NA values with very likely substitutions. MICE operates under the assumption that the data missing is MAR (Missing at Random).

Due to the data collection process (each row represents a county), the likely possible bias is that certain states refuse or fail to collect these variables in a systematic way, and thus the data is no longer MAR. We will check this assumption towards the end of the modelling by considering our finalized model on both the imputated and reduced dataset (rows including NA's will be removed), and assess their similarities. Regardless, modelling will be done using the imputed dataset, assuming MAR.

```
trim = moddat[,-18]
imp <- mice(trim, m = 5, maxit = 50, meth = "pmm")
```

```
Warning: Number of logged events: 505
```

```
complete(imp)
```

```
imputed <- complete(imp)
imputed_new <- imputed
```

Initial variable selection for our model will be informed by domain knowledge and insight gained from prior visualization of the data.

Literature on socioeconomic factors affecting cancer mortality point to poverty, education, and race as some of the most important factors. In the 2017 paper "Socioeconomic and Racial/Ethnic Disparities in Cancer Mortality, Incidence, and Survival in the United States, 1950–2014: Over Six Decades of Changing Patterns and Widening Inequalities," the authors concluded that individuals in lower income and education groups had significantly higher mortality and incidence rates. The authors also noted that Blacks had higher mortality and incidence rates than other races, likely due to the interconnection of race and income. In the 2021 paper "Leading cancers contributing to educational disparities in cancer mortality in the US, 2017," the authors concluded that there was a significant difference between the mortality rate between individuals with a bachelors degree and higher, and all education levels below that. Since both these studies use data exclusively from the U.S., and are within the the time frame of interest to us, we are comfortable using these conclusions to guide our variable selection.

The visualizations of our own data support these conclusions as well as suggest a categorical variable indicating whether a given county is in the Southwest region. To define which states belong to the Southeast, we will be using the regions specified by the Bureau of Economic Analysis, who divide the United States into 8 regions.

Additionally, the conclusions from the second paper suggest two new variables, **PctNoHS18_24** and **PctHS18_24**, which represent the percent of county residents ages 18-24 whose highest education attained is less than a high school degree, and then a high school degree, respectively. While there are several other variables related to educational goals, such as percentage of county residents ages 18-24 who have attained a bachelors, the literature above suggests that residents with lower educational achievements have a higher cancer mortality, while the opposite is not necessarily true.

The variables for the initial model will be povertyPercent (Percent of populace in poverty), PctBlack, and PctNoHS18 and PctHS18_24. For future investigation we will consider isSouthEast (a categorical variable created later), as well as the variables related to healthcare coverage **(PctPrivateCoverage, PctPrivateCoverageAlone, PctEmpPrivCoverage, PctPublicCoverage, PctPubliceCoverageAlone).**

```
mod1 <- lm(data = imputed_new, TARGET_deathRate ~ povertyPercent + PctBlack + PctNoHS18_24
```

```
summary(mod1)
```

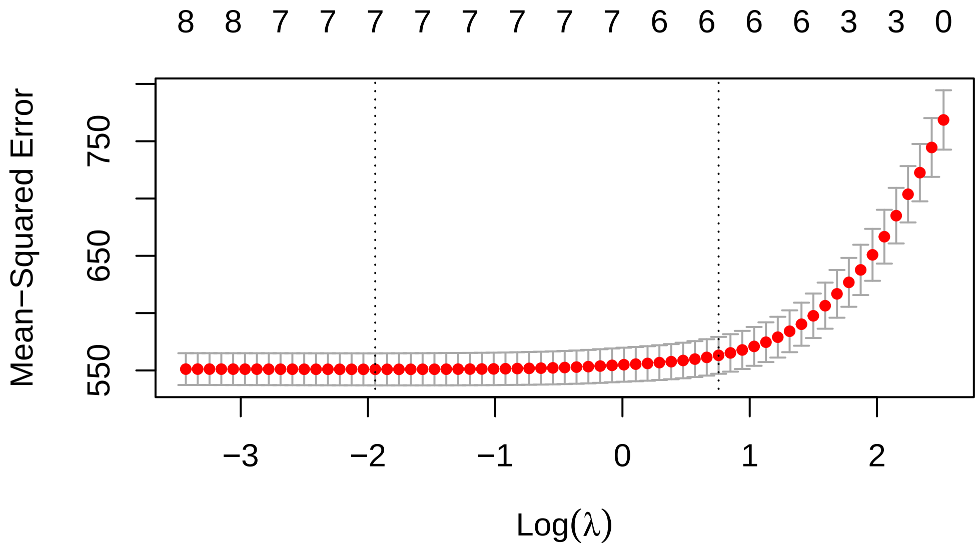
```
Call:
lm(formula = TARGET_deathRate ~ povertyPercent + PctBlack + PctNoHS18_24 +
    PctHS18_24, data = imputed_new)

Residuals:
     Min       1Q   Median       3Q      Max
-106.595  -13.332    1.245   14.515  164.404

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    127.58734    2.10954  60.481  < 2e-16 ***
povertyPercent   1.66957    0.08312  20.087  < 2e-16 ***
PctBlack         0.13644    0.03527   3.869 0.000112 ***
PctNoHS18_24    -0.17345    0.05673  -3.058 0.002251 **
PctHS18_24       0.70898    0.04883  14.518  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.21 on 3042 degrees of freedom
Multiple R-squared:  0.2402,    Adjusted R-squared:  0.2392
F-statistic: 240.4 on 4 and 3042 DF,  p-value: < 2.2e-16
```

The initial fit is rather weak, with a R-squared of 0.24.

To add variables of interest, we will be using Lasso regression using a Lambda 1 standard deviation away from the minimum residual deviance Lambda value, which will be obtained using cross-validation. Here is the graph of Lambda values, with the minimum Lambda at 0.1434, and the 1 SE Lambda value at 2.1288.

```
new_england <- c("Connecticut", "Maine", "Massachusetts", "New Hampshire", "Rhode Island",
mideast <- c("Delaware", "District of Columbia", "Maryland", "New Jersey", "New York", "Pe
great_lakes <- c("Illinois", "Indiana", "Michigan", "Ohio", "Wisconsin")
plains <- c("Iowa", "Kansas", "Minnesota", "Missouri", "Nebraska", "North Dakota", "South
southeast <- c("Alabama", "Arkansas", "Florida", "Georgia", "Kentucky", "Louisiana", "Miss
southwest <- c("Arizona", "New Mexico", "Oklahoma", "Texas")
rocky_mountain <- c("Colorado", "Idaho", "Montana", "Utah", "Wyoming")
far_west <- c("Alaska", "California", "Hawaii", "Nevada", "Oregon", "Washington")

get_region <- function(state) {
  if (state %in% new_england) {
    return("New England")
  } else if (state %in% mideast) {
    return("Mideast")
  } else if (state %in% great_lakes) {
    return("Great Lakes")
  } else if (state %in% plains) {
    return("Plains")
  } else if (state %in% southeast) {
    return("Southeast")
  } else if (state %in% southwest) {
    return("Southwest")
  } else if (state %in% rocky_mountain) {
    return("Rocky Mountain")
  } else if (state %in% far_west) {
    return("Far West")
  } else {
    return(NA)
  }
}

imputed_new$Region <- sapply(imputed_new$State, get_region)

imputed_new$isSoutheast <- ifelse(imputed_new$Region == "Southeast", "Yes", "No")
```

```
set.seed(5)
y = imputed_new$TARGET_deathRate
x = data.matrix(imputed_new[, c('povertyPercent', 'PctBlack', 'PctHS18_24','PctNoHS18_24',
cv_model <- cv.glmnet(x, y, alpha = 1)
plot(cv_model)
```



For a more parsimonious model, we will be using the 1 SE Lambda value, to hopefully alleviate some co linearity. Here are the coefficients selected after using this Lambda value:

```r
min_lambda <- cv_model$lambda.min
se_lambda <- cv_model$lambda.1se
best_model <- glmnet(x, y, alpha = 1, lambda = se_lambda)
coef(best_model)
```

```
9 x 1 sparse Matrix of class "dgCMatrix"
                               s0
(Intercept)           127.0455984
povertyPercent          0.4272104
PctBlack                .
PctHS18_24              0.3092784
PctNoHS18_24            .
isSoutheast             9.1235082
PctPublicCoverage       0.1413345
PctPublicCoverageAlone  0.7151673
PctUnemployed16_Over    0.3115501
```

According to the Lasso results, PctBlack and PctNoHS18_24 were eliminated, while PctPublicCoverage, PctPublicCoverageAlone, and PctUnemployed16_Over were admitted. Thus, our final model after variable selection is as follows:

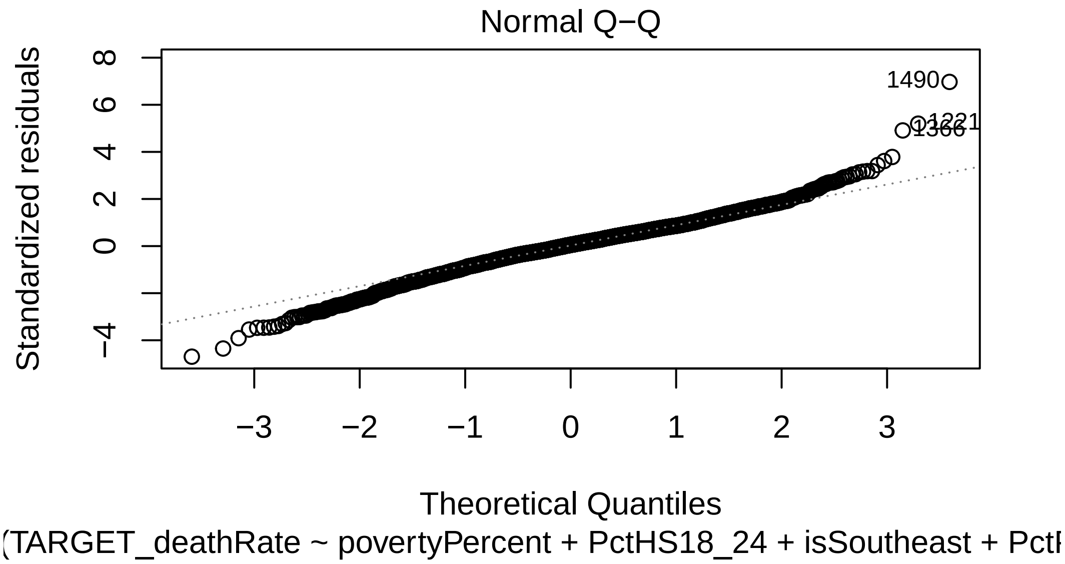
```
finmod <- lm(data = imputed_new, TARGET_deathRate ~ povertyPercent + PctHS18_24 + isSouthe
summary(finmod)
```

```
Call:
lm(formula = TARGET_deathRate ~ povertyPercent + PctHS18_24 +
    isSoutheast + PctPublicCoverage + PctPublicCoverageAlone +
    PctUnemployed16_Over, data = imputed_new)

Residuals:
     Min       1Q   Median       3Q      Max
-110.107  -13.052    1.293   14.243  163.409

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            122.3032     2.4841  49.234  < 2e-16 ***
povertyPercent           0.5638     0.1207   4.672 3.12e-06 ***
PctHS18_24               0.4931     0.0497   9.922  < 2e-16 ***
isSoutheastYes          11.6651     1.0157  11.485  < 2e-16 ***
PctPublicCoverage        0.2811     0.1109   2.535  0.01128 *
PctPublicCoverageAlone   0.5704     0.1827   3.123  0.00181 **
PctUnemployed16_Over     0.5630     0.1730   3.253  0.00115 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.46 on 3040 degrees of freedom
Multiple R-squared:  0.2866,	Adjusted R-squared:  0.2852
F-statistic: 203.5 on 6 and 3040 DF,  p-value: < 2.2e-16
```

The fit has improved somewhat, with an R-squared of 0.287.

For inference on these variables, let us check normality assumptions using a Q-Q plot:

```
plot(finmod, which =2)
```



While the line appears to fit well towards the center of the plot, the line veers off heavily towards the ends. For greater inference, we will use a Shapiro-Wilk test on the residual values to test for normality:

```
shapiro.test(finmod$residuals)
```

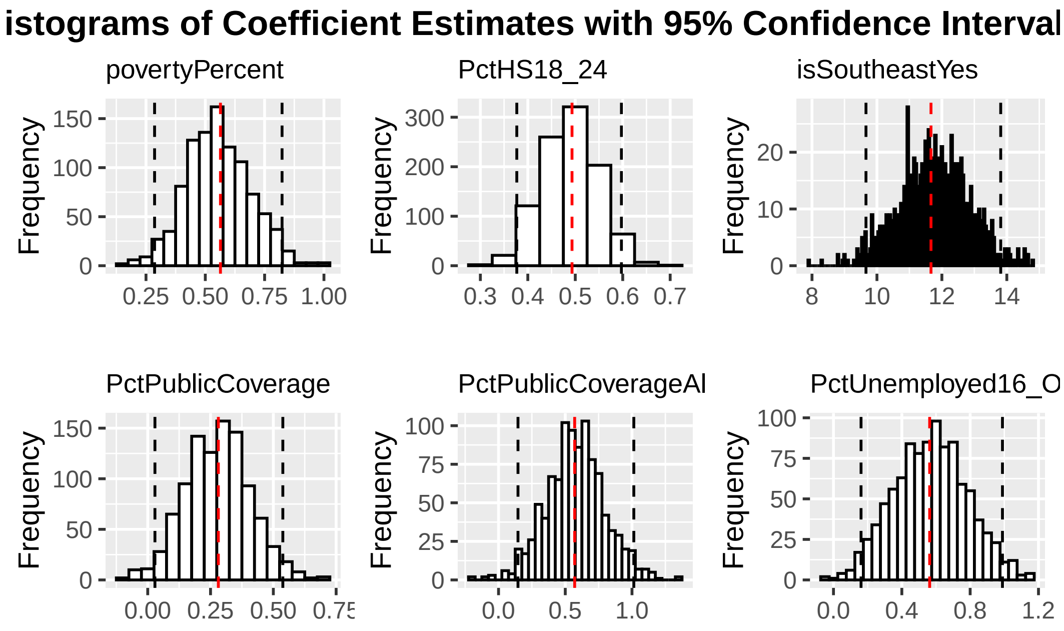
```

    Shapiro-Wilk normality test

data:  finmod$residuals
W = 0.98178, p-value < 2.2e-16
```

The p-value comes out to <2e-16, thus rejecting the null hypothesis that our fitted results are normal.

For inference, we will thus be using a Bootstrap algorithm to understand the sampling distribution of our coefficients. Using a percentile confidence interval from these sampling distribution, we will then test the null hypothesis that each variable's value is actually 0.

```
nboot <- 1000

# Create a function to calculate the coefficients using the bootstrap
```

```r
coef.boot <- function(data, indices) {
  model <- lm(TARGET_deathRate ~ povertyPercent + PctHS18_24 + isSoutheast + PctPublicCove
  return(coef(model)[-1]) # exclude intercept column
}

# Perform the bootstrap using the defined function
boot.results <- boot(data = imputed_new, statistic = coef.boot, R = nboot)

# Convert bootstrap results to a data frame
boot.df <- as.data.frame(boot.results$t)
colnames(boot.df) <- c("povertyPercent", "PctHS18_24", "isSoutheastYes","PctPublicCoverage

# Get coefficient estimates from original model
finmod <- lm(TARGET_deathRate ~ povertyPercent + PctHS18_24 + isSoutheast + PctPublicCover
coef.estimates <- coef(finmod)[-1 , drop = TRUE]

# Create a function to plot histograms with quantile and coefficient lines and a title
plot.hist <- function(x, coef.est, varname) {
  p <- ggplot(data.frame(x), aes(x = x)) +
    geom_histogram(binwidth = 0.05, color = "black", fill = "white") +
    geom_vline(xintercept = quantile(x, probs = c(0.025, 0.975)), linetype = "dashed") +
    geom_vline(xintercept = coef.est, color = "red", linetype = "dashed") +
    xlab("") + ylab("Frequency") +
    ggtitle(varname)+
    theme(plot.title = element_text(size = 9.5))
  return(p)
}

# Create a list of plots for each column in boot.df with titles
plot.list <- mapply(plot.hist, x = boot.df, coef.est = coef.estimates, varname = names(boo

# Combine the plots into a single figure with a title
grid.arrange(grobs = plot.list, ncol = 3, top = textGrob('Histograms of Coefficient Estima
)
```



As seen on the histograms, none of the 95% percentile intervals contain 0, rejecting our null hypothesis that our variables have no effect.