ST404 Assignment 2

Frank Or, Remos Gong, Sam Glanfield, Thomas Broadbent

2022-03-09

Contents

1 Findings **2**
- 1.1 Key Findings . . . 2
- 1.2 Summary of EDA . . . 2
- 1.3 Major Determinants of Mortality Rates . . . 4
- 1.4 Modelling Approach . . . 4
- 1.5 Areas that do not conform the general pattern . . . 5

2 Statistical Methodology **6**
- 2.1 Outcomes of EDA . . . 6
- 2.2 Modelling Approach and Variable Selection . . . 8
- 2.3 Statistical Interpretation and Validation . . . 12

3 Author Contributions **13**

4 References **13**

5 Appendix **14**
- 5.1 Outcomes of EDA . . . 14
- 5.2 Modelling Approach and Variable Selection . . . 14
- 5.3 Statistical Interpretation and Validation . . . 17

1 Findings

1.1 Key Findings

Given our preferred model the most important predictors for cancer death rates for a given state were incidence rates, median income and private health care coverage. All predictors for the model are in the table below with the parameter estimate and a scaled model to show importance of predictors. This shows that `incidenceRate` and `log medianIncome` followed by `PctPrivateCoverage` have the largest influence over death rates.

Intercept	IncidRT	PctEmp	PctUnemp	PrivCov	EmpPrivCov	Edu	lPctBlack	lMedInc
470.47	0.226	-0.251	0.291	-0.315	0.212	-10.056	0.904	-32.388
0.00	0.426	-0.078	0.036	-0.130	0.088	-0.082	0.057	-0.284

Understandably cancer diagnoses are a key predictor for death rates, with an increase in the number of cancer diagnoses by 50 per 100,000 increasing cancer death rates by 10 per 100,000. This of course does not imply that there should be a policy to decrease cancer diagnoses. Increasing cancer diagnoses will lead to more people with cancer on death certificates even if cancer may not have been the primary cause of death. Wealthier north western states have similar mean incidence rates to poorer south western states but much lower death rates showing that diagnoses are a good baseline additional predictors for poverty need to be included.

Due to the strong collinearity of many variables and thus our choice of regression method, we can not give a strong interpretation for keeping all but one variable the same and giving the associated change in death rates. An increased Death Rate is more likely if a county has a low median income, low private coverage, these though will be dependent on low employment and high unemployment percentages which may be because of a low education rate. Due to variables explaining the same amount of information about death rates the LASSO model reduces the complexity, for example even though Percent Public Coverage is correlated well with death rates it can be left out as it is strongly correlated with both percent private coverage and employee provided coverage.

1.2 Summary of EDA

In the previous EDA report we have explored the data set and we will perform data cleaning and variable transformations before constructing a linear model.

—Might Merge in one if run out of space?

1.2.1 Missing/Incorrect Values and Outliers

We identified 152 missing values in **Percent Employed 16 and Over**. Since they have no pattern, we deduced that these values are missing at random (MCAR). We used other complete data entries to calculate what we would expect these values to be and impute them back to our data set. We also identified values in **Average Household Size** that unreasonably small. We believed this is an error in data entry and we scaled them by 100 in order to make it normal. For outliers, We removed **Williamsburg City, Virginia** because of its high **Incidence Rate** but low **Death Rate**, hence high influence to our model.

1.2.2 Transformations

In order to fulfill the model assumptions, namely **Linearity**, **Homoscedasticity**, **Normality**, we performed log transforms for two variables: **Median Income** and **Percent Black** which suffered the most among all variables. Although the transformations did not cure the problems, they result in improvements in linearity and normality, reducing heteroscedasticity for both said variables. For other variables, we deduced that they are good enough to fulfill the model assumptions and hence did not perform any transformations for model simplicity.

–Might add plots?

1.2.3 Multicollinearity

We have discovered a few pairs or clusters of predictor variables that are highly correlated with each other. The obvious ones are **Median Age Female** and **Median Age Male**, **Percent Married** and **Percent Married Households**, **Binned Income** and **Median Income**. These variables measure the same force and hence at least one in the pairs are expected to be excluded in our modelling.

1.2.4 Correlation with Death Rate

While most predictor variables have absolute correlation coefficient roughly 0.3 to 0.4 with death rate, Median Age Male, Median Age Female and Average Household Size have almost zero correlation coefficient. We expect our model to not include these uncorrelated variables.

1.3 Major Determinants of Mortality Rates

From our model building process we have identified that the major determinants in high mortality rates of cancer in the US are: **Incidence Rate, Percent Unemployed 16 and Over, Percent Employer Provided Private Coverage and Percent Black**. Whereas, the major determinants in low mortality rates are: **Percent Employed 16 and Over, Percent Private Coverage, Education Levels and Median Income**.

1.4 Modelling Approach

We used Stepwise Regression, RIDGE Regression and LASSO Regression to build our model. We compared the outputs and the variables selections of these models. We also analyse the goodness of fits of these models using Leave-one-out cross validation, R^2 statistics and residuals analysis. We did not include **Geography** and **Binned Income** and we were using the **Log Percent Black** and **Log Median Income** in the following analysis.

—I found the following sections very tricky to write up. Might need a better structure. NEED HELP!

1.4.1 Stepwise Regression

We used **Bayesian Information Criteria(BIC)** in our analysis. This penalises additional parameters harder than **Akaike information criterion(AIC)** which agrees with parsimony in our modelling. We performed forward, backward and hybrid stepwise regression. In all of the models that were generated we notice that we have groups of parameters that are similar and hence induce multicollinearity in the models. For example, we see both **Percent Married** and **Percent Married Households** in the models and as stepwise methods do not account for multicollinearity this causes increases in the variance of our coefficient estimates and makes the model sensitive to changes, thus reducing drastically the predictive power of the model. In order to address the issue of multicollinearity we apply Ridge and Lasso techniques which account for this.

1.4.2 RIDGE Regression

1.4.3 LASSO Regression

The suggested model contains eight variables which is the simplest model among our approaches. Since the purpose of this report is to reveal patterns in the mortality rate, we believed it is best to choose a simple model for stronger explanatory power.

1.4.4 Final Model Choice and Diagnostics

We used residuals plot and QQ plot to diagnose the stepwise regression model. Both plots are satisfactory, agreeing with model assumptions.

We also performed Leave-one-out cross-validation for this stepwise model, which gives R^2 value of 0.4621. We did the same cross-validation for our Ridge and Lasso model. The Ridge model gives R^2 value of 0.4436. The Lasso model gives R^2 value of 0.4441.

The stepwise model fits the observed data the best among all our models. The Lasso and the Ridge model is similar in this sense. However, the Lasso model contains the least number of predictor variables and we believed this is the most important criterion. Thus from our analysis briefly outlined above and in the next section we recommend the following model, as seen below, due to its mix of predictive and explanatory power as well as being a simple model.

$$DeathRate = 470.47 + 0.23IncidenceRate - 0.25PctEmployed16\_Over + 0.029PctUnemployed16\_Over - 0.32PctPrivateCoverage + 0.21PctEmpPrivCoverage - 10.06Edu18\_24 + 0.90log(PctBlack) - 32.34log(MedianIncome)$$

1.5 Areas that do not conform the general pattern

Similarly, we see from the average residual map that in particular the 3 states Utah (Average of -22), Idaho (Average of -19) and Colorado (Average of -18) all have high residual averages and as they are negative this implies that the model is overfitting for these states and thus these states don't conform to the general pattern. However, as the model if overfitting and we are estimating death rate this isnt a cause for concern as we know the average death rates for these states is highly likely to be lower than what the model predicts. It is also important to note that these all lie in a similar region and in general we see that the majority of states on the west coast have an average negative residual, so the model overfits for these states, whereas on the east coast and central we see positive average residuals, the model is underfitting these states. With the main states of concern for overfitting with an average residual of 14 each for Oklahoma and Arkansas and an average of 13 in the District of Columbia.

2 Statistical Methodology

We first combined our results and findings from our preliminary EDA which is mainly discussed in Outliers and Transformations.

2.1 Outcomes of EDA

2.1.1 Missing Or Incorrect Values

We also see counties with missing values in Percentage Employed 16 and Over and we conclude that the data is Missing Completely at Random. In order to rectify this we impute this data by fitting a linear regression model of Percentage Employed 16 and Over on the remaining variables to estimate what these values would be. (See Appendix 2.1.1).

For counties with Average Household Size less than one we took the decision to scale the transformations by 100 and keep them in the dataset. This fixed the normality of AvgHouseholdSize as shown in the histogram.

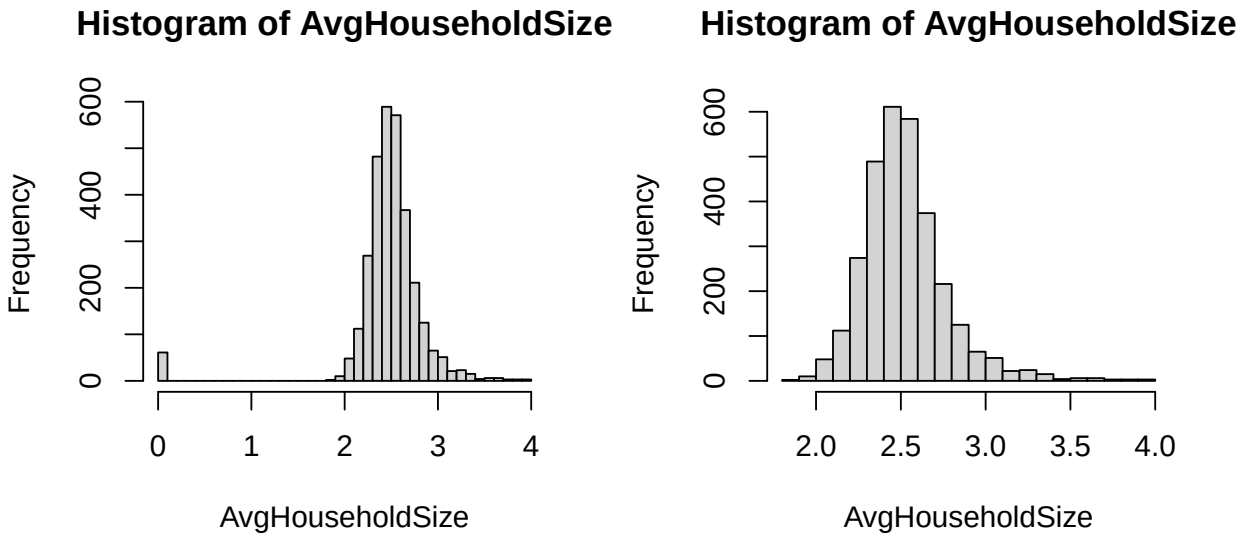


Figure 1: Histograms before(Left) and after(Right)

2.1.2 Outliers

Counties with high Incidence Rates, namely 'Union County, Florida' and 'Williamsburg City, Virginia.' We looked into the cook's distance plots and noticed only 'Williamsburg City, Virginia'

has large cook's distance and hence influential. The first cook's distance plot used a linear model with only incidenceRate as the predictor variable. The second used all the numerical variables. We concluded although 'Union County, Florida' has high leverage, it is not influential and hence should be kept in our data set. (See Appendix 2.1.2)

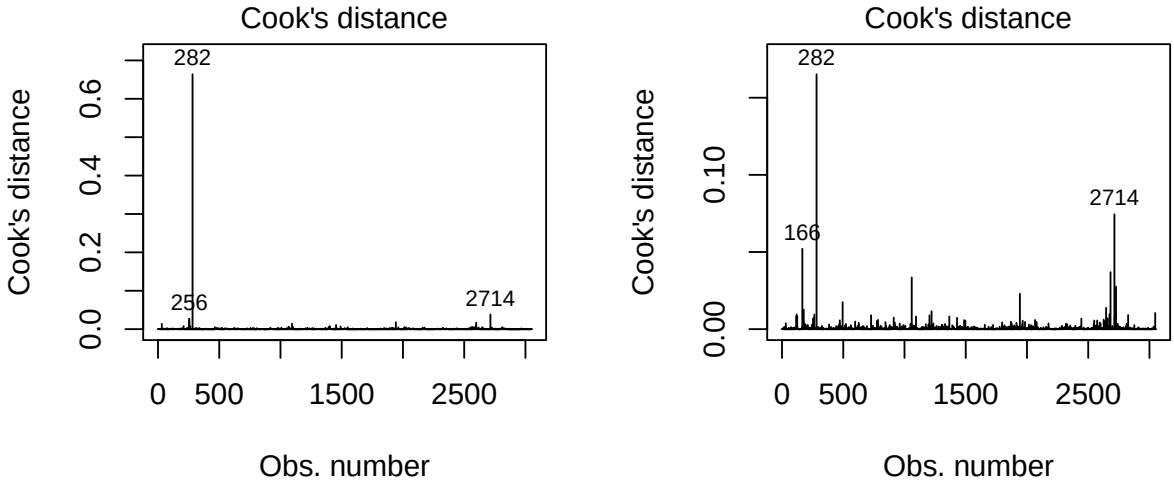


Figure 2: Cook's distance plots of models with only incidence Rate(Left) and all variables(Right)

2.1.3 Transformations

We transform Percent Black by first shifting the values upwards by 0.05, to ensure we have no zero values, then take a log transform. We also transform the Median Income by again taking a log transformation. We do these transformations to ensure the data is not heavily skewed and allow for a more accurate model. (See Appendix 2.1.3)

The following residual plots show the improvements in homoscedasticity in PctBlack and medIncome after log-transform respectively.

—Might add other improvements

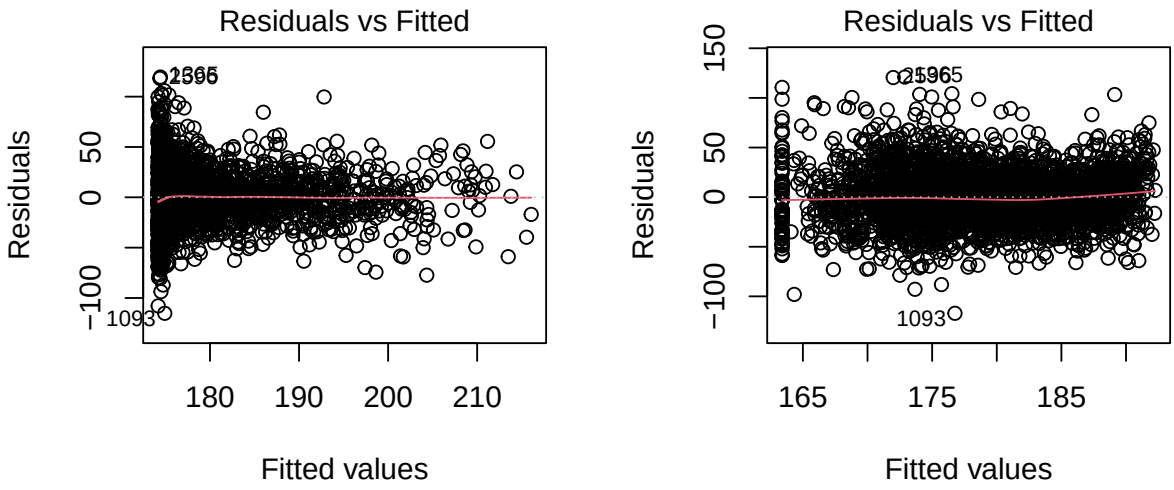


Figure 3: Residuals plot Percent Black(Left) and Log Percent Black(Right

2.2 Modelling Approach and Variable Selection

2.2.1 AIC and BIC Forward and Backward Variable Selection

We perform forward, backward and hybrid stepwise regression according to BIC to ensure a more parsimonious model. We see that the models generated for BIC are all the same. We see that in the BIC model has 11 variables and in Table 1 we see the estimates for these variables.

For the model generated we see from the summary output, with strong evidence, that all the coefficients are different from zero. From observing the output we see that we would expect there to be multicollinearity in these models due to variables that measure similar or opposite quantities, for example Percent Married and Percent Married Households.

We compute the Variance Inflation Factors (VIF) for the model and see that there are numerous values that are at least 5 indicating that we have multicollinearity. Specifically, we see a VIF of at least 10 in Percent Married and a VIF of 8 in Percent Married Households. As well as in log(Median Income), Percent Employed 16 and Over, Percent Employers Private Coverage and Percent Private Coverage. Due to this frequent multicollinearity between the predictors in the stepwise model this suggests that the stepwise model is not a suitable approach as multicollinearity is not taken into account and thus we proceed below with Ridge Regression and Lasso to reduce the effects of multicollinearity on the model.

To complete the section on stepwise regression we perform leave one out cross validation (in order to compute the R^2 of the stepwise models and also the Root Mean Squared Error (See Appendix 2.2.1). This allows us to compare these models with the Ridge and Lasso models we generate in the enxt sections. Performing this we see in the stepwise model we have an R^2 value of 0.4623.

2.2.2 RIDGE Regression

To address the multi-collinearity present in the data we assess the viability of a RIDGE Regression Model. For this we fit all continuous variables as predictors using the `glmnet` library.

The trace shows that the ridge penalisation method reduces many variables close to 0 but does not remove any variables from the model itself. (See Apendix 2.2.2)

Fitting a ridge model and using the value of λ one standard error further away from the minimum does not remove any of the terms from the model and gives the coefficients as seen in Table 1. We also include the parameter estimates when the parameters are scaled to highlight how significant the parameter is.

Performing a leave one out cross validation gives an R^2 statistic of 0.4438. Which is similar to the previous step wise regression method and is harder to diagnose therefore we conclude that it it is not a suitable model.

2.2.3 LASSO Regression

We used cross-validation from the glmnet library which leaves out a 10th of the data every time. We produced the following plot of mean-squared error against log(lambda). We decided to use the 1-standard-error-lambda because this likely to shrink some predictor variables to zero, performing variable selection. We prefer a simpler model. (See Apendix 2.2.3)

```
## [1] 0.6378303
```

We produced a plot that shows a trace of each parameter estimate for different values of log(lambda). (See Fig. 4 and Apendix 2.2.3) We scaled the model for a better visual interpretation. (See Fig. 5 and Apendix 2.2.3)

The Lasso regression using 1-standard-error-lambda produces the following parameter estimates. Alongside we see the parameter estimates for Ridge regression and stepwise regression to allow for easier comparison of the models.

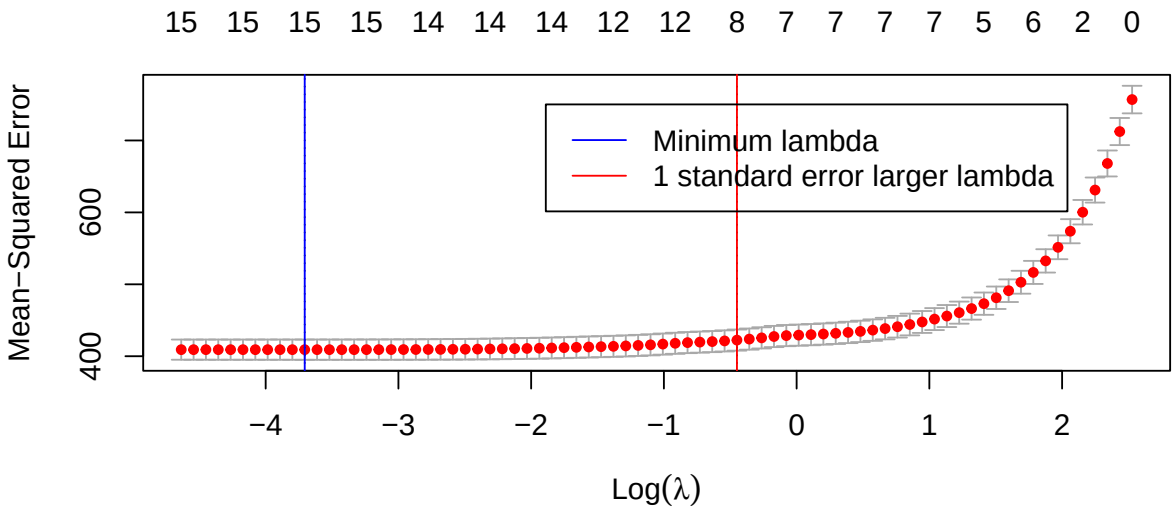


Figure 4: Mean-squared error against log(lambda)
(#fig:Lasso cv)

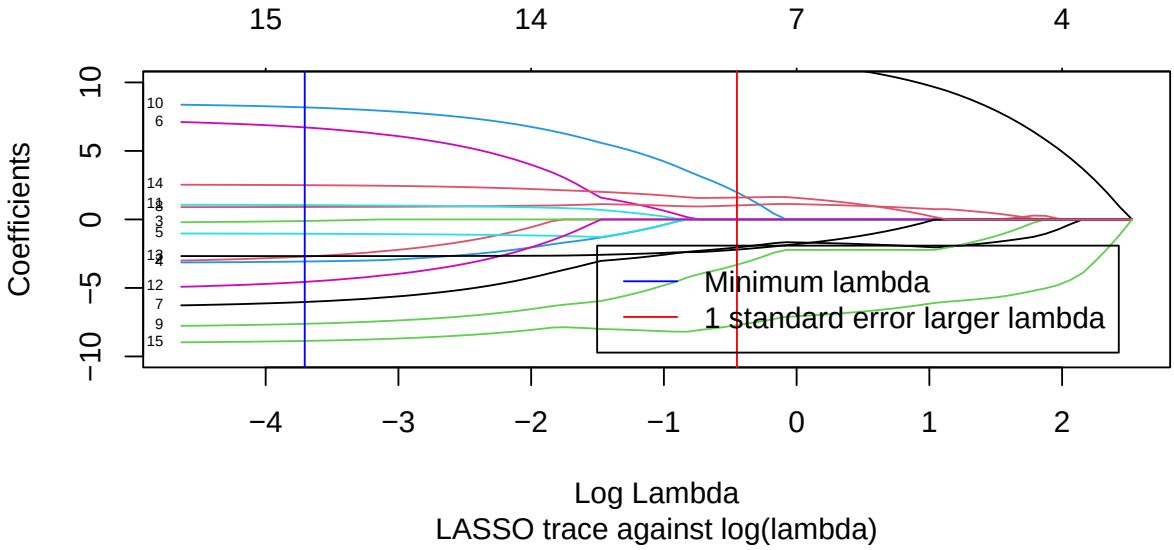


Figure 5: Trace of parameter estimates for different lambda
(#fig:Lasso trace)

Table 1: Parameter Estimates of the Models

	Ridge	LASSO	Stepwise
(Intercept)	353.941	470.470	507.558
incidenceRate	0.184	0.226	0.227
povertyPercent	0.228	0.000	0.000
MedianAgeMale	-0.125	0.000	0.000
MedianAgeFemale	-0.214	0.000	-0.421
AvgHouseholdSize	-5.098	0.000	0.000
PercentMarried	0.177	0.000	1.285
PctEmployed16_Over	-0.299	-0.251	-0.821
PctUnemployed16_Over	0.458	0.291	0.000
PctPrivateCoverage	-0.258	-0.315	-0.685
PctEmpPrivCoverage	0.258	0.212	0.899
PctPublicCoverage	0.282	0.000	0.000
PctMarriedHouseholds	-0.093	0.000	-0.973
Edu18_24	-11.649	-10.056	-11.839
logpctblack	1.060	0.904	1.405
logmedincome	-19.070	-32.388	-32.750

2.2.4 Interpretation

From the estimates output above, the Lasso regression eliminates a considerable number of predictor variables. The Lasso regression model has the following non-zero predictor estimates: **Incidence Rate**, **Percent Employed 16 and Over**, **Percent Private Coverage**, **Percent Employer Provided Private Coverage**, **Education Levels**, **Log of Percent Black**, **Log of Median Income**.

$$DeathRate = 470.47 + 0.23\, IncidenceRate - 0.25\, PecentEmployed\_Over16 + 0.29\, PercentUnemployed\_Over16 -$$

The two age variables and Average Household Size are removed from the model. This agrees with our EDA which showed that they have very close to zero correlation coefficient with Death Rate. Both employment variables and two out of three healthcare coverage variables are included in this model. We believe this is reasonable because although they showed evidence of collinearity/multicollinearity, we did not have strong arguments to remove any of them. Therefore, we agreed with the variable selection suggested by the Lasso Regression.

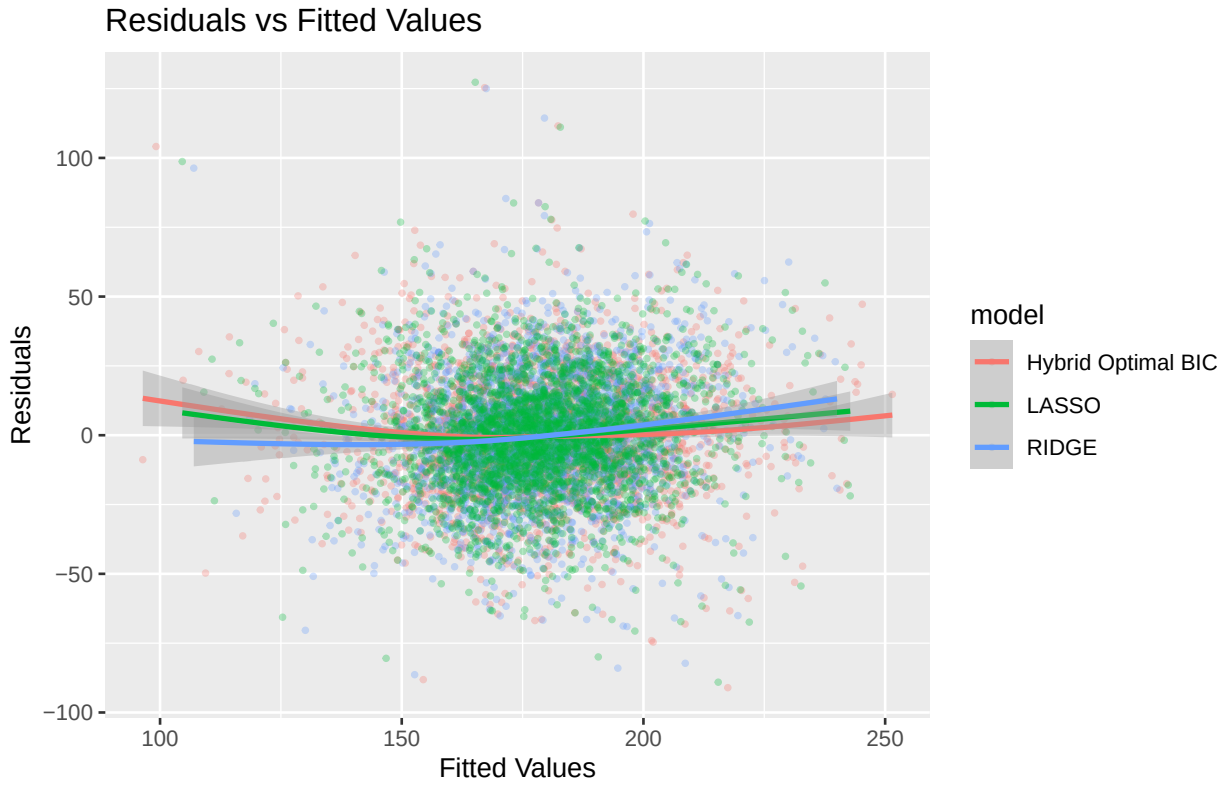
Most predictor estimates are smaller than one. However the estimates for Education Levels and

Log Median Income are -10.05 and -32.39 respectively which are exceptionally large and negative. This means that they have much larger impact on the predicted death rates than other predictor variables. It makes sense that Median Income has the largest impact because patients in counties with higher median income tend to be able to afford better treatment which reduce mortality rates.

We calculated the R-squared using Leave-One-Out Cross Validation. The R-squared for this Lasso Regression model is 0.4441.

2.3 Statistical Interpretation and Validation

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

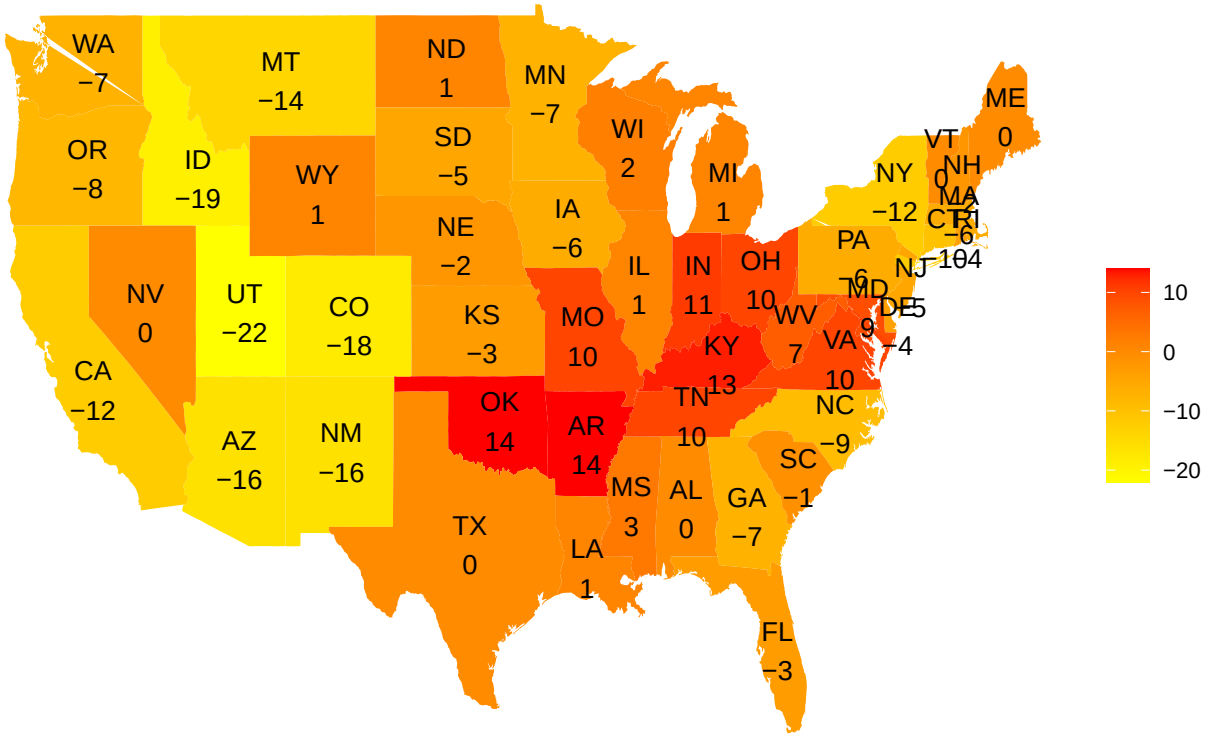


We see from the plot above that all three models appear to exhibit constant variance with an even spread of points as well as having a mean of zero with the distributions symmetrical about zero. Thus satisfying our key linear regression assumptions of homeoscedasticity and residuals having a mean of zero.

Produce map of each states average residual value for the LASSO model.

```
## Warning: Ignoring unknown aesthetics: inherit.aes, label.size

## Warning: Use of `statenames$lab` is discouraged. Use `lab` instead.
```



3 Author Contributions

4 References

5 Appendix

5.1 Outcomes of EDA

5.1.1 Missing Or Incorrect Values

```
# Impute the missing data seen in the data set
mod1=lm(PctEmployed16_Over~+incidenceRate+medIncome+binnedInc+povertyPercent+MedianAgeMale+MedianAgeFemale+AvgHouseholdSize
        +PercentMarried+PctUnemployed16_Over+PctPrivateCoverage+PctEmpPrivCoverage+PctPublicCoverage+PctBlack
        +PctMarriedHouseholds+Edu18_24,cancer)
missdf = cancer[which(is.na(cancer$PctEmployed16_Over)==TRUE),]
imputed = predict(mod1,missdf)
cancer[which(is.na(cancer$PctEmployed16_Over)==TRUE),"PctEmployed16_Over"] = imputed
# Scale average household sizes that are less than 1 by 100
cancer[which(cancer$AvgHouseholdSize < 1), ]$AvgHouseholdSize <-
  100*cancer[which(cancer$AvgHouseholdSize < 1), ]$AvgHouseholdSize
hist(cancer$AvgHouseholdSize, breaks=30, xlab="AvgHouseholdSize", main="Histogram of AvgHouseholdSize")
```

5.1.2 Outliers

```
# Cook's distance Plot
par(mfrow=c(1,2))
plot(lm(deathRate ~ incidenceRate,data=cancer),4)
plot(lm(deathRate ~ .,data=cancer[-c(1,4)]),4)
# Removing outlier incidence rate 'Williamsburg City, Virginia'
cancer <- filter(cancer, incidenceRate <= 850)
```

5.1.3 Transformations

```
# Log transforming the heavily skewed distributions of PctBlack and medIncome
cancer$logpctblack = log(cancer$PctBlack+0.05)
cancer$logmedincome = log(cancer$medIncome)
# Showing improvements in homoscedasticity in PctBlack
# Similar plots can be produced for medIncome
par(mfrow=c(1,2))
plot(lm(deathRate~PctBlack,data=cancer),1)
plot(lm(deathRate~logpctblack,data=cancer),1)
```

5.2 Modelling Approach and Variable Selection

5.2.1 AIC and BIC Forward and Backward Variable Selection

```
# We perform stepwise regression for BIC
cancermodel <- cancer %>% select(
```

```
  !c("Geography", "medIncome", "binnedInc", "PctBlack"))
c0=lm(deathRate~1,cancermodel)
cmax=lm(deathRate~.,cancermodel)
forwardoptimalBIC = step(c0,direction="forward",
scope=list("lower"=c0,"upper"=cmax),trace=0,k=log(3045))
backwardoptimalBIC = step(cmax,direction="backward",
scope=list("lower"=c0,"upper"=cmax),trace=0,k=log(3045))
hybridoptimalBIC = step(c0,direction="both",
scope=list("lower"=c0,"upper"=cmax),trace=0,k=log(3045))
# An example code for summary and coefficient of models
summary(hybridoptimalBIC)
coef(hybridoptimalBIC)
# Computing the VIF of the BIC stepwise regression model
vif(hybridoptimalBIC)
library(caret)
#specify the cross-validation method
#fit a regression model and use LOOCV to evaluate performance
loocv <- function(lm1, data=cancer) {
  ctrl <- trainControl(method = "LOOCV")
  xnam <- names(lm1$coefficients)[-1]
  fmla <- as.formula(paste("deathRate ~ ", paste(xnam, collapse= "+")))
  model <- train(fmla, data = data, method = "lm", trControl = ctrl)
  return(model)
}
hybridoptimalBIC.loocv <- loocv(hybridoptimalBIC,data=cancermodel)
```

5.2.2 RIDGE Regression

```
library(glmnet)
# Producing a 1-standard-error-lambda RIDGE regression model
# We also used similar codes for LASSO regression
cancer.dm <- cancer %>%
  select(!c("Geography", "medIncome", "binnedInc", "PctBlack", "deathRate")) %>%
  data.matrix()
lm.ridge <- glmnet(cancer.dm, cancer$deathRate, alpha=0)
lm.ridge.cv <- cv.glmnet(cancer.dm, cancer$deathRate, alpha=0)
# Create Model Plot Func
traceLogLambda <- function(lm1, lm1.cv, ylim=NULL, sub=NULL) {
  plot(lm1,"lambda",label = T, ylim=ylim)
  abline(v=log(lm1.cv$lambda.1se),col="red")
  abline(v=log(lm1.cv$lambda.min),col="blue")
  legend("bottomright",legend=c("Minimum lambda", "1 standard error larger lambda"),lty=c(1,1),col=c("blue","red"), ins=0.05)
  title(sub=sub)
}
traceLogLambda(lm.ridge, lm.ridge.cv, ylim=c(-20, 1), sub="RIDGE trace against log(lambda)")
```

```
lm.ridge.1se <- glmnet(cancer.dm, cancer$deathRate, alpha = 0,
                lambda = lm.ridge.cv$lambda.1se)
lm.ridgefitted <- predict(lm.ridge.1se, newx=cancer.dm)
scatter.smooth(cancer$deathRate - lm.ridgefitted, x=lm.ridgefitted,
                xlab="Fitted", ylab="Residuals",
                sub="Residuals vs Fitted for RIDGE Model")
# Leave-one-out cross-validation for RIDGE Regression
# Calculate R-squared statistics
# We also performed the same procedures for LASSO Regression
library(parallel)
library(foreach)
library(doParallel)
numCores <- detectCores()
registerDoParallel(numCores)
n <- dim(cancer)[1]
dev.ratios <- rep(NA, n)
dev.ratios1 <- rep(NA, n)
lm.ridge.cv <- cv.glmnet(cancer.dm, cancer$deathRate, alpha=0, nfolds=n, parallel=TRUE)
lm.ridge.cv$lambda.1se
for (i in 1:n) {
  lm.ridge.1se <- glmnet(cancer.dm[-i,], cancer$deathRate[-i], alpha = 0,
                lambda = lm.ridge.cv$lambda.1se)
  dev.ratios1[i] <- lm.ridge.1se$dev.ratio
}
mean(dev.ratios1)
##Produce column of table for coefficients table
library(knitr)
# This produces a table of parameter estimate for RIDGE model
# We also produced similar table for LASSO model and combined these with the stepwise coefficients into a 'kable'
a <- data.frame(sapply(round(coef(lm.ridge.1se), 3), FUN=identity))
colnames(a) <- "Parameter Estimates"
rownames(a) <- rownames(round(coef(lm.ridge.1se), 3))
```

5.2.3 LASSO Regression

```
# Produce a LASSO Regression model
# Produce a plot of mean-squared error against log(lambda)
set.seed((934))
lm.lasso <-glmnet(cancer.lasso.dm,cancer$deathRate,alpha=1)
lm.lasso.cv <- cv.glmnet(cancer.lasso.dm,cancer$deathRate,alpha=1)
plot(lm.lasso.cv)
abline(v=log(lm.lasso.cv$lambda.1se),col="red")
abline(v=log(lm.lasso.cv$lambda.min),col="blue")
legend("topright",legend=c("Minimum lambda", "1 standard error larger lambda"),lty=c(1,1),col=c("blue","red"), ins=0.1)
lm.lasso.cv$lambda.1se
```

```r
lm.lasso.1se <- glmnet(cancer.lasso.dm,cancer$deathRate,lambda = lm.lasso.cv$lambda.1se,alpha=1)
lm.lasso.1se.fitted <- predict(lm.lasso.1se, newx=cancer.lasso.dm)
lm.lasso.1se.residuals <- cancer$deathRate - lm.lasso.1se.fitted
# Produce a plot that shows the trace of each parameter estimate for different lambda
lm.lasso.scaled<-glmnet(scale(cancer.lasso.dm),cancer$deathRate,alpha=1)
traceLogLambda(lm.lasso.scaled,lm.lasso.cv,ylim=c(-10,10), sub="LASSO trace against log(lambda)")
```

5.3 Statistical Interpretation and Validation

```r
# Create plot of Residual vs Fitted
library(ggplot2)
multResidualPlot <- function(residual.list, fitted.list, models) {
  a <- data.frame()
  for (i in 1:length(models)) {
    x <- fitted.list[[i]]
    y <- residual.list[[i]]
    model <- rep(models[i], times=length(x))
    df.temp <- data.frame(x, y, model)
    colnames(df.temp) <- c("x", "y", "model")
    a <- rbind(a, df.temp)
  }
  mrp <- ggplot(a, aes(x=x, y=y, colour=model)) +
    geom_point(alpha=0.3, size=0.75) + geom_smooth() +
    labs(x="Fitted Values", y="Residuals", title="Residuals vs Fitted Values", model="Models")
  return(mrp)
}
lm.ridge.residuals <- cancer$deathRate - lm.ridgefitted
residual.list <- list(hybridoptimalBIC$residuals, lm.ridge.residuals, lm.lasso.1se.residuals)
fitted.list <- list(hybridoptimalBIC$fitted.values, lm.ridgefitted, lm.lasso.1se.fitted)
mrp <- multResidualPlot(residual.list, fitted.list, c("Hybrid Optimal BIC", "RIDGE", "LASSO"))
```

5.3.1 Produce map of each states average residual value for the LASSO model.

```r
#average for each state
averageresidual=round(tapply(lm.lasso.1se.residuals,cancermodel2$State,mean)[-c(2,12)],0)
states <- map_data("state")
averagedf <- data.frame(region=unique(states$region), averageresidual)
mergedf <- merge(states, averagedf, by="region")
statenames <- data.frame(region=tolower(state.name), clong=state.center$x, clat=state.center$y)
statenames <- merge(statenames, averagedf, by="region")
statenames$lab <- paste(c(state.abb, 'DC')[match(statenames$region, c(tolower(state.name), 'District of Columbia'))], '\n', stat
#Produce heatmap of average residual
qplot(long, lat, data=mergedf, geom="polygon", fill=averageresidual, group=region) +
  scale_fill_gradient(averageresidual,low="yellow",high="red") +
```

```
  geom_text(data=statenames,aes(clong,clat,label=statenames$lab,inherit.aes = FALSE,label.size=0.001)) +
   theme_void() + theme(legend.title = element_blank(),plot.title = element_text(hjust = 0.5))
+ ggtitle("Heatmap of Average Residuals of US States")
```