ST404 Assignment 2

Frank Or, Remos Gong, Sam Glanfield, Thomas Broadbent

2022-03-03

Contents

1 Findings

From the modelling approach taken in this report and our preliminary EDA we have determined that the major determinants in of high mortality rates in the US, as well as key predictors in modelling the mortality rate are: incidence rate, log of median income, poverty percent, percent employed and unemployed 16 and over, percent private, publci and employee provided private coverage, log of percent black, education levels and percent married households and percent married. From the list above we see that there are variables that measure similar or opposite quantities such as percent employed and percent unemployed 16 and over. Thus during the modelling process we will take caution with similar variables to ensure we keep the explanatory power of the model.

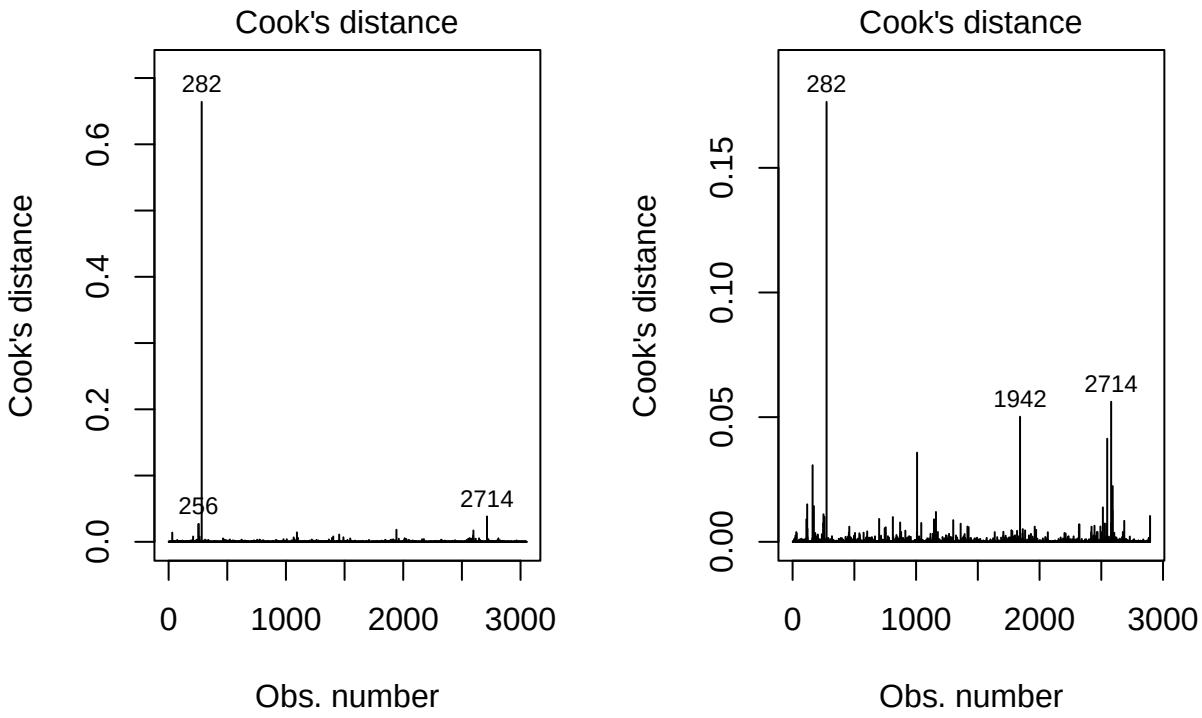
We do see examples of counties that do not conform to the general pattern with unusually high or low mortality rates. One example we have is 'Williamsburg City, Virginia' where we see a high incidence rate of 1014 yet a comparatively small death rate of 162, however we see that it's percent private coverage is 19% above the US average and the median male and female ages of 26 and 24 respectively are considerably lower than the median US ages, potentially being causes for this low death rate. Another county with an unsusually high mortality rate is 'Union County, Florida' where we see a death rate of 363 which is 168 above the upper quartile (75% of data points). Here we see a higher than average poverty percent and a lower than average education level being possible causes but not fully explaining the drastically high death rate level and similarly high incidence rate of 1207. We also see a slighlty less but still unusual value for 'Charlottseville City, Virginia' with a high incidence rate of 719, but again a comparatively low death rate of 178. In the other case we saw this was also in Virginia and again we see a higher than average percent private coverage, 4% higher than the upper quartile, again potentially being the cause for a high incidence rate but low death rate. Conversely, we see an unsuaully low death rate of 60 in 'Pitkin County, Colorado,' however this can be explained by a high median income and percent private coverage.

2 Statistical Methodology

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

2.1 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.



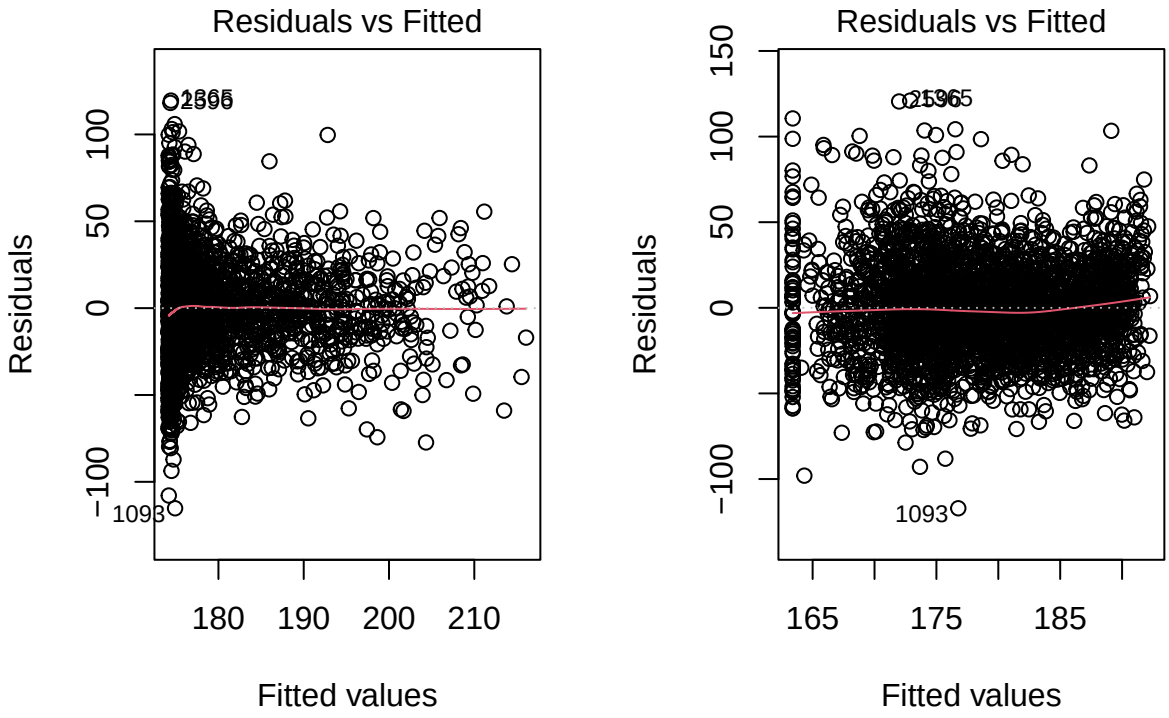
For counties with Average Household Size less than one we took the decision to scale the transformations by 100 and keep them in the dataset. This fixed the normality of AvgHouseholdSize as shown in the histogram.

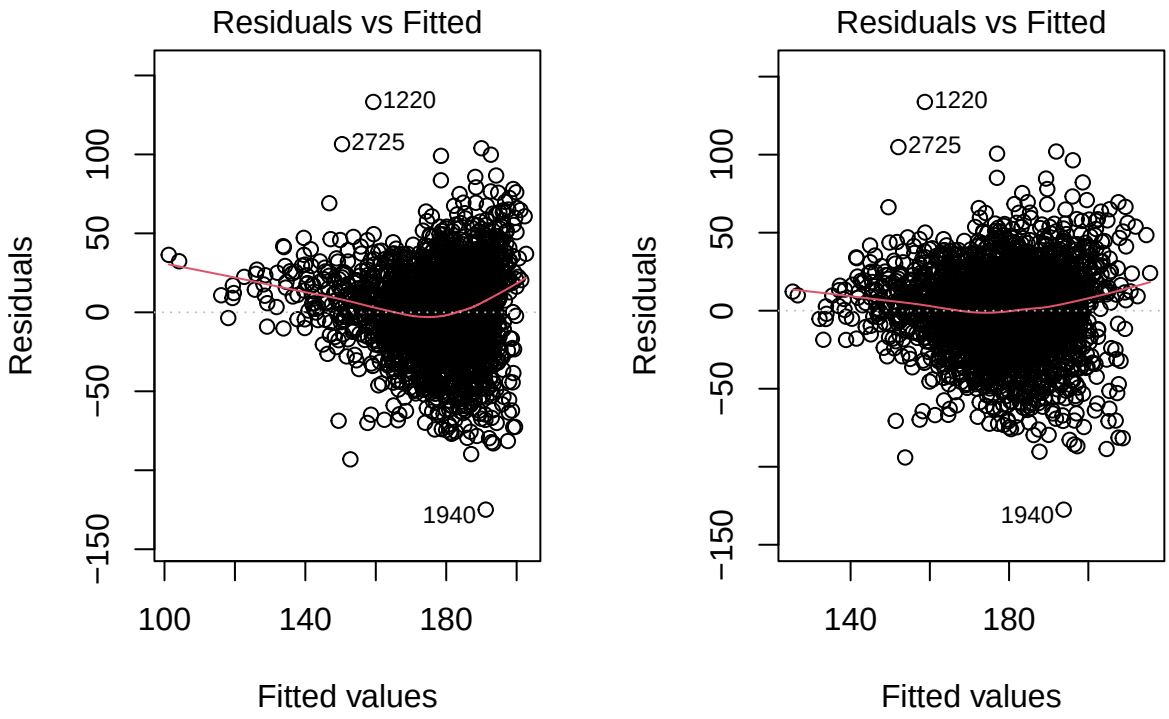
We also see counties with missing values in Percentage Employed 16 and Over and we conclude that the data is Missing Completely at Random. In order to rectify this we impute this data by fitting a linear regression model of Percentage Employed 16 and Over on the remaining variables to estimate what these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

2.2 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 trasformation. We do these transformations to ensure the data is not heavily skewed and allow for a more accurate model.

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

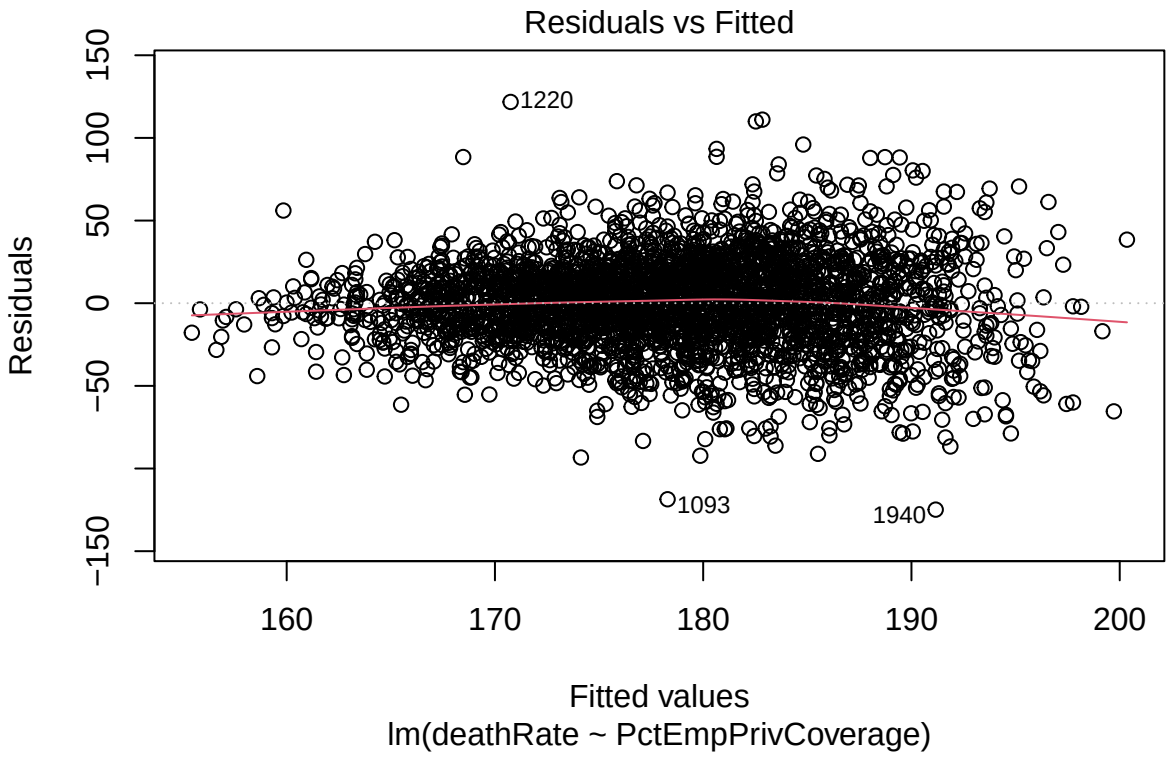




We also notice slight evidence of heteroscedasticity in PctEmpPrivCoverage as shown in the residual plot below. However, since it is not severe and the suggested power transformation using spread level plot makes it more heteroscedastic, we decided not to perform transformation on PctEmpPrivCoverage.

```
# Finding power transformation for PctEmpPrivCoverage
spreadLevelPlot(lm(deathRate~PctEmpPrivCoverage, data=cancer))
```

```
##
## Suggested power transformation:  -4.380067
```



2.3 Modelling Approach and Variable Selection

2.3.1 AIC Forward and Backward Variable Selection

We perform forward, backward and hybrid stepwise regression according to both AIC and BIC. We see that the models generated for AIC are all the same and ones generated for BIC are al the same. In the AIC case we have the model has 14 parameters, whereas the BIC model has 11 variables.

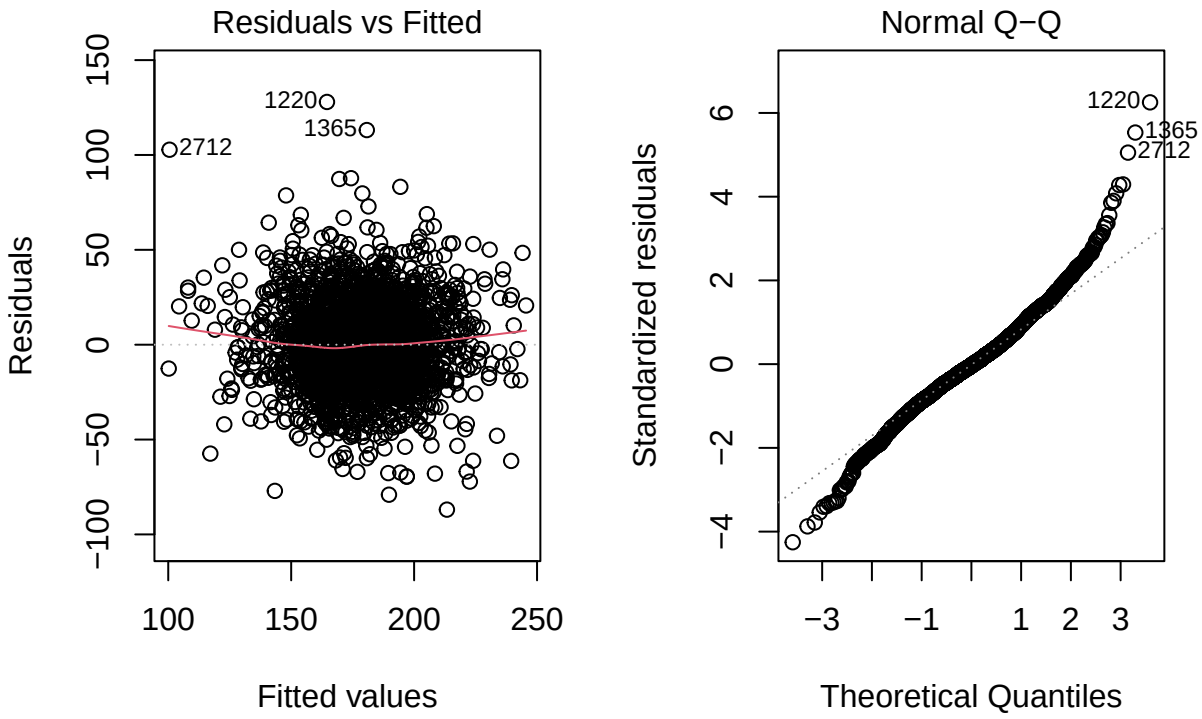
From the summary output we see that for the AIC model all coefficients have strong evidence that they are different from zero other than Percent Unemployed 16 and Over and Percent Public Coverage. For the BIC model these variables are not in the model and further from the summary output we see with strong 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 BIC 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.

Thus we remove the variables with the highest VIF out of the model, namely Percent Married and Percent Private Coverage. From the summary of this updated model Percent Married Households is no longer significant, so we remove it from the model and see in this further updated model that all coefficients are significant and have VIF less than 5.

For this updated model as all the coefficients are significant and the VIF for all variables are less than 5 we produce residual plots, as can be seen below, in Figure XXX. From the plots we can see that this model satisfies the assumptions of homeoscedasticity and normality. In the left plot we see a horizontal line about zero indicating the errors have zero mean and a constant variance and spread of points giving us constant variance in the errors. This is also supported through the non-constant variance test function from the car package which gives a p-value of 0.29215, hence we accept the null hypothesis that the errors have a constant variance. In the right plot we see a QQ plot and see that the majority of points appear to be either on or close to the dotted reference line. Thus, suggesting that the errors are normally distributed.

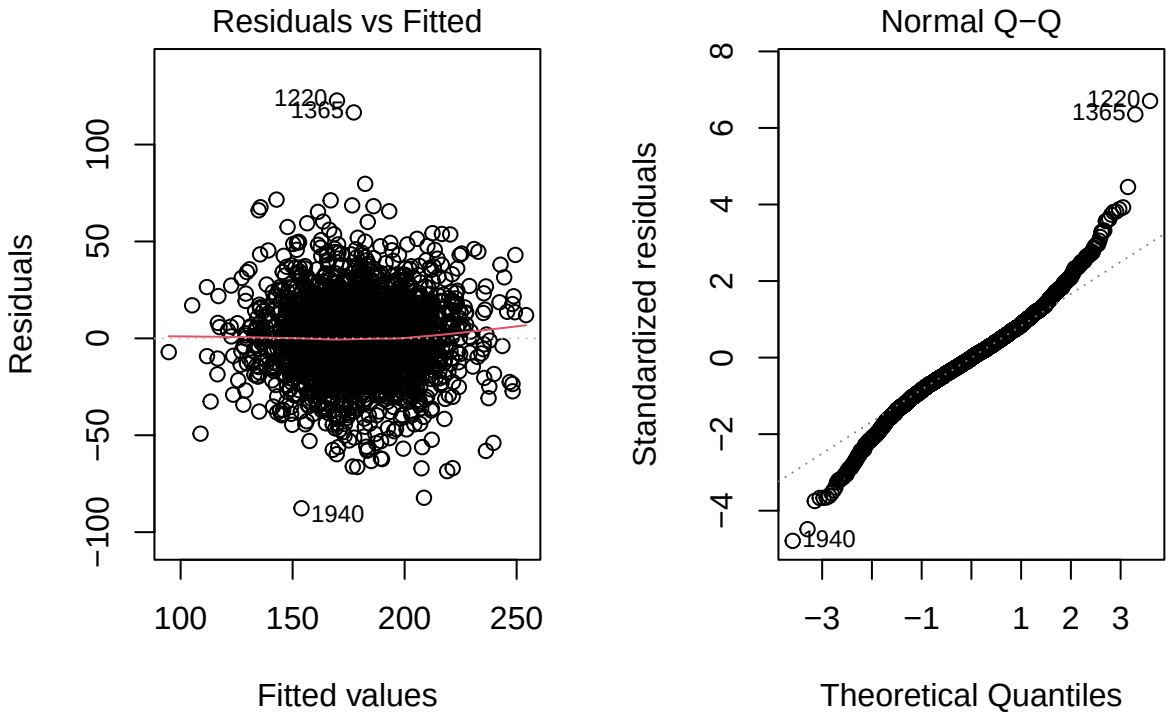


Now we perform stepwise with states included as a factor and as can be seen in both the AIC and BIC case the factor State is kept in the model and otherwise the other variables included are similar with the main different exclusion being log(Percent Black) which was included in the model without the addition of the states.

We see from the summary output that the majority of the coefficients are significant with the departures from this coming from the State factor variable for certain states. In the AIC model we see that Poverty Percent is not significant with a p value of 0.0628, however we see in the BIC model this variable is dropped and as Poverty Percent will have a high correlation with Percent Unemployed 16 and Over we adopt the BIC model.

Below in Figure XXX we can see the residual and QQplot for the BIC model. In the residual plot on the left hand side we see constant variance and a mean of zero with the majority of points near the horizontal red line. This is similar to what we saw when we didn't include the states in the selection process. Similarly, we see in the QQplot the majority of points on or near the reference line suggesting that the errors are indeed normally distributed, again as we see in the above case of

not including states in the selection process.



In order to test whether the addition of states in the model makes a significant difference on the model in the stepwise selection process we perform an F-Test using the anova function. Computing this we get a p-value of less than 2.2x10^-16, thus giving us strong evidence that we should reject the null hypothesis that the model not including states is better. Hence, the BIC model including the states as a factor variable is a better fit of the data according to the anova function.

2.3.2 RIDGE Regression

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

```
library(glmnet)
```

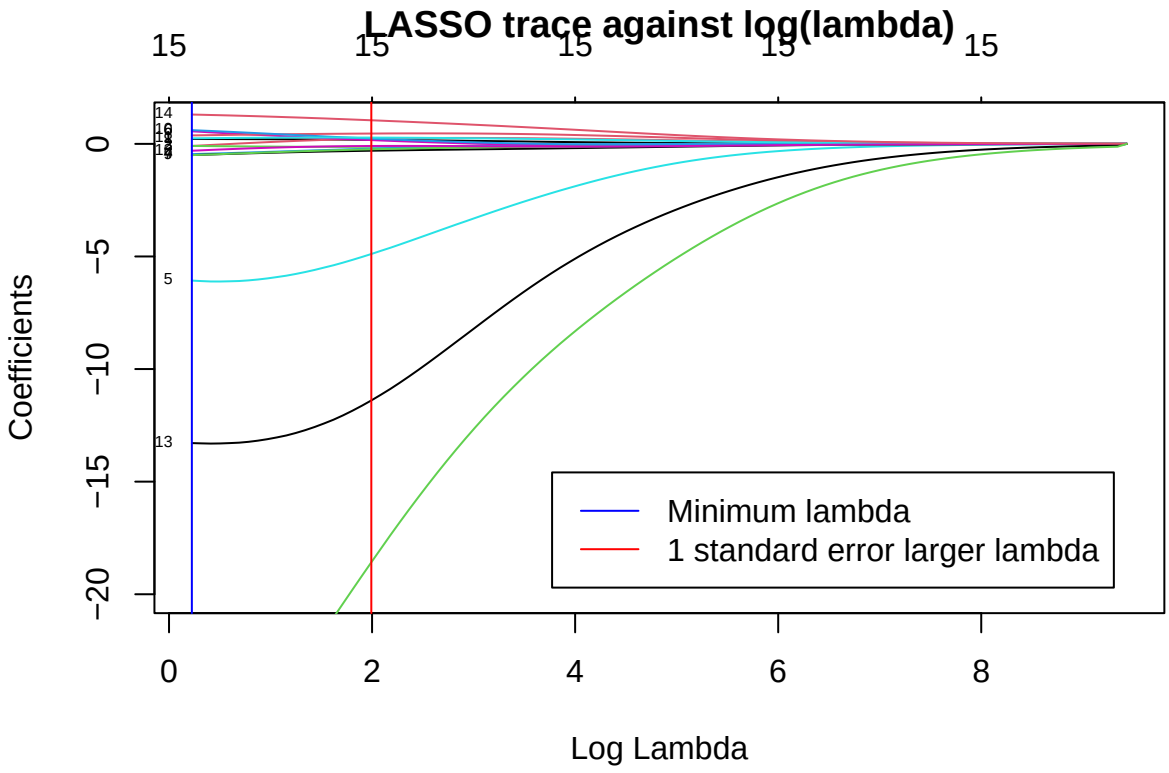
```
## Loading required package: Matrix
```

```
##
```

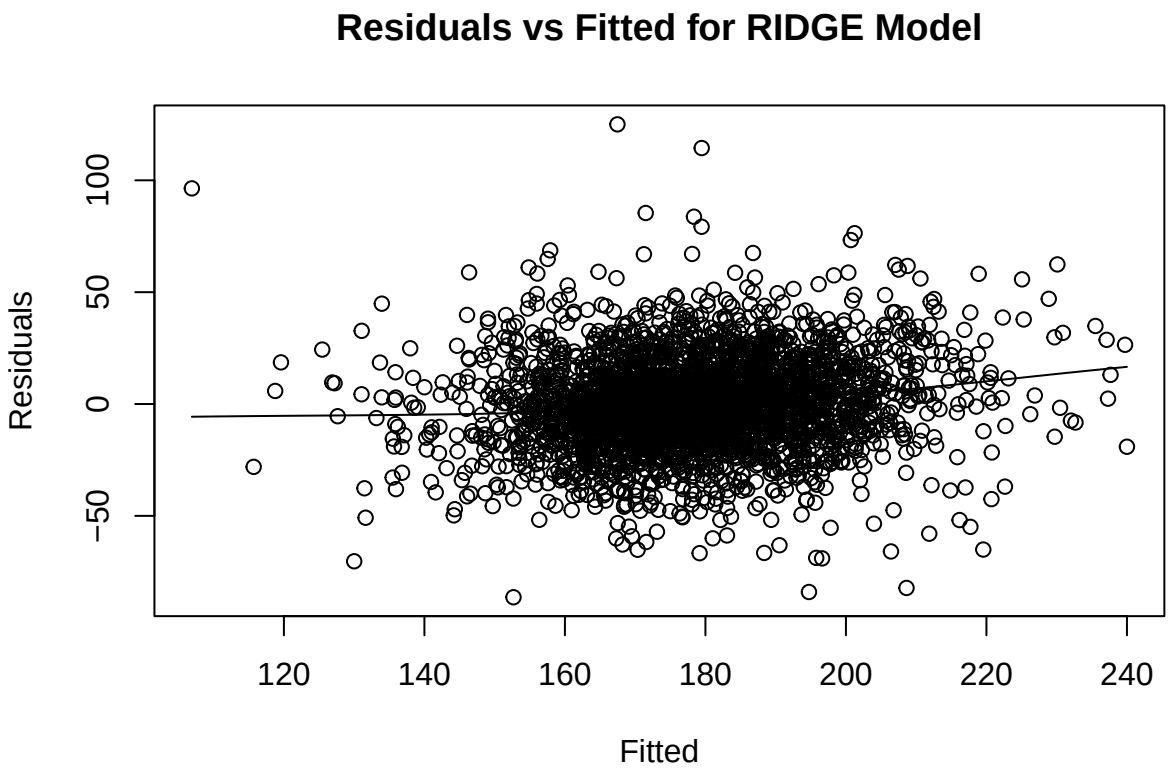
```
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loaded glmnet 4.1-3
```

```
# Create Data Matrix
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) {
  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)
  title(main="LASSO trace against log(lambda)")

}

traceLogLambda(lm.ridge, lm.ridge.cv, ylim=c(-20, 1))
```




```
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",
                main="Residuals vs Fitted for RIDGE Model")
```



Fitting a ridge model does not remove any of the terms from the model and gives the following coefficients. We also include the parameter estimates when the parameters are scaled to highlight how significant the parameter is.

	Parameter Estimates	Scaled Parameter Estimates
(Intercept)	360.246	0.000
incidenceRate	0.188	0.334
povertyPercent	0.217	0.055
MedianAgeMale	-0.126	-0.024
MedianAgeFemale	-0.227	-0.038
AvgHouseholdSize	-5.197	-0.043
PercentMarried	0.196	0.037
PctEmployed16_Over	-0.314	-0.089
PctUnemployed16_Over	0.452	0.058
PctPrivateCoverage	-0.269	-0.090
PctEmpPrivCoverage	0.278	0.074
PctPublicCoverage	0.281	0.078
PctMarriedHouseholds	-0.098	-0.022
Edu18_24	-11.836	-0.088
logpctblack	1.071	0.066
logmedincome	-19.646	-0.155

2.4 Statistical Interpretation and Validation

3 References

4 Appendix

```
library(dplyr)
library(car)
library(tidyr)


## Included Libraries
# For pipe operator and general mutation
load('cancer.rdata')
```

```r
# 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 rates 'Williamsburg City, Virgin850ia'
cancer <- filter(cancer, incidenceRate <= 850)
# 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$AvgHouseholdSize, breaks=30, xlab="AvgHouseholdSize", main="Histogram of AvgHouseh
# Impute the missing data seen in the dataset
mod1=lm(PctEmployed16_Over~+deathRate+incidenceRate+medIncome+binnedInc+povertyPercent+MedianAg
missdf = cancer[which(is.na(cancer$PctEmployed16_Over)==TRUE),]
imputed = predict(mod1,missdf)
cancer[which(is.na(cancer$PctEmployed16_Over)==TRUE),"PctEmployed16_Over"] = imputed

# 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 and medIncome
par(mfrow=c(1,2))
plot(lm(deathRate~PctBlack,data=cancer),1)
plot(lm(deathRate~logpctblack,data=cancer),1)
plot(lm(deathRate~medIncome,data=cancer),1)
plot(lm(deathRate~logmedincome,data=cancer),1)
# Finding power transformation for PctEmpPrivCoverage
spreadLevelPlot(lm(deathRate~PctEmpPrivCoverage, data=cancer))
plot(lm(deathRate~PctEmpPrivCoverage,data=cancer),1)
# Below we perform stepwise regression for both AIC and BIC
cancermodel = cancer[,-c(1,3,15)]
c0=lm(deathRate~1,cancermodel)
cmax=lm(deathRate~.,cancermodel)
```

```
forwardoptimalAIC = step(c0,direction="forward",
scope=list("lower"=c0,"upper"=cmax),trace=0)
backwardoptimalAIC = step(cmax,direction="backward",
scope=list("lower"=c0,"upper"=cmax),trace=0)
hybridoptimalAIC = step(c0,direction="both",
scope=list("lower"=c0,"upper"=cmax),trace=0)
forwardoptimalBIC = step(c0,direction="forward",
scope=list("lower"=c0,"upper"=cmax),trace=0,k=log(3044))
backwardoptimalBIC = step(cmax,direction="backward",
scope=list("lower"=c0,"upper"=cmax),trace=0,k=log(3044))
hybridoptimalBIC = step(c0,direction="both",
scope=list("lower"=c0,"upper"=cmax),trace=0,k=log(3044))

# We compute the summaries of the stepwise models
summary(hybridoptimalAIC)
summary(hybridoptimalBIC)

# Computing the VIF of the BIC stepwise regression model
vif(hybridoptimalBIC)

# Update the BIC model removing high VIF predictors
updateBIC = lm(deathRate~logmedincome+incidenceRate+PctEmpPrivCoverage+PctEmployed16_Over+Edu18
summary(updateBIC)

# Update further BIC model removing insignificant variable
updateBIC2 = lm(deathRate~logmedincome+incidenceRate+PctEmpPrivCoverage+PctEmployed16_Over+Edu
summary(updateBIC2)
vif(updateBIC2)

# Performing non-constant variance test
ncvTest(updateBIC2)
```

```r
# Producing residual plots for updateBIC2 model
par(mfrow=c(1,2))
plot(updateBIC2,which=c(1,2))

# Create stepwise regression models including the states
cancermodel2 = separate(cancer,"Geography", into=c("County","State"),sep=",")[-c(1,4,5,16)]
c0=lm(deathRate~1,cancermodel2)
cmax=lm(deathRate~.,cancermodel2)
hybridoptimalAIC2 = step(c0,direction="both",
scope=list("lower"=c0,"upper"=cmax),trace=0)
hybridoptimalBIC2 = step(c0,direction="both",
scope=list("lower"=c0,"upper"=cmax),trace=0,k=log(3044))

# Produce summary and coefficients of state stepwise regression models
coef(hybridoptimalAIC2)
coef(hybridoptimalBIC2)
summary(hybridoptimalAIC2)
summary(hybridoptimalBIC2)

# Residual plots of state stepwise regression
par(mfrow=c(1,2))
plot(hybridoptimalBIC2, which=c(1,2))

# Perform F-Test on the BIC models above
anova(updateBIC2,hybridoptimalBIC2)

summary(updateBIC2)

library(glmnet)

# Create Data Matrix
```

```r
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) {
  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)
  title(main="LASSO trace against log(lambda)")

}

traceLogLambda(lm.ridge, lm.ridge.cv, ylim=c(-20, 1))

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",
                main="Residuals vs Fitted for RIDGE Model")

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$dev.ratio
  # lm.residuals <- predict(lm.ridge.1se, newx=cancer.dm[-i, ]) - cancer$deathRate[-i]
  # dev.ratios[i] <- 1 - sum(lm.residuals^2)/sum((cancer$deathRate[-i] -mean(cancer$deathRate[

}
mean(dev.ratios1)

library(knitr)
library(Matrix)



a <- data.frame(sapply(round(coef(lm.ridge.1se), 3), FUN=identity))
colnames(a) <- "Parameter Estimates"

cancer.dm.scale <- cancer %>%
  select(!c("Geography", "medIncome", "binnedInc", "PctBlack", "deathRate")) %>%
```

```
  scale %>%
  data.matrix()
lm.ridge.scale.cv <- cv.glmnet(cancer.dm.scale, scale(cancer$deathRate), alpha = 0)
lm.ridge.1se.scale <- glmnet(cancer.dm.scale, scale(cancer$deathRate), alpha = 0, lambda = lm.r
a$"Scaled Parameter Estimates" <-sapply(round(coef(lm.ridge.1se.scale), 3), FUN=identity)


rownames(a) <- rownames(round(coef(lm.ridge.1se), 3))
kable(a)
```