Project 1

Nancie Kung, Calvin Raab, David Collier and Eitan Shimonovitz

4/21/2021

Load libraries and import the data.

```
library(readr)
library(dplyr)
library(leaps)
library(ggplot2)
library(reshape2)
library(scales)
library(corrplot)
library(car)
library(stargazer)
library(lmtest)
library(broom)
library(AER)
library(caret)

Fat_Supply_Quantity_Data <- read_csv("Fat_Supply_Quantity_Data.csv")
```

Data cleaning process.

```
#select columns that are not filled with zeros
Fat_Supply_data <- Fat_Supply_Quantity_Data %>% select(Country, `Animal Products`,
  `Animal fats`, `Cereals - Excluding Beer`, Eggs, `Fish, Seafood`, `Fruits - Excluding Wine`,
  Meat, `Milk - Excluding Butter`, Offals, Oilcrops, Pulses, Spices, `Starchy Roots`,
  Stimulants, Treenuts, `Vegetal Products`, `Vegetable Oils`, Vegetables, Obesity,
  Undernourished, Population, Confirmed, Deaths)

Fat_Supply_data <- Fat_Supply_data[Fat_Supply_data$Deaths > 0,]
Fat_Supply_data <- Fat_Supply_data[!is.na(Fat_Supply_data$Deaths),]


Fat_Supply_data$Undernourished[Fat_Supply_data$Undernourished == "<2.5"] <- 2.5
Fat_Supply_data$Undernourished <- as.numeric(Fat_Supply_data$Undernourished)

# replace NAs in Obesity and Undernourished with the median values
Fat_Supply_data$Obesity[is.na(Fat_Supply_data$Obesity)] <-
  median(Fat_Supply_data$Obesity, na.rm=TRUE)
Fat_Supply_data$Undernourished[is.na(Fat_Supply_data$Undernourished)] <-
  median(Fat_Supply_data$Undernourished, na.rm=TRUE)

data <- Fat_Supply_data
```

```r
# Here is a dataset that includes the parameters found in backAIC,
# along with: Country, Population, Confirmed, and Deaths (Using For Analysis)
backAICdata.plus <- data_frame(data$`Country`, data$`Animal fats`,
    data$`Cereals - Excluding Beer`, data$`Fruits - Excluding Wine`,
    data$`Oilcrops`, data$`Pulses`, data$`Spices`, data$`Starchy Roots`,
    data$Stimulants, data$Treenuts, data$`Vegetal Products`, data$`Vegetable Oils`,
    data$Vegetables, data$Obesity, data$Undernourished, data$Population,
    data$Confirmed, data$Deaths)

names(backAICdata.plus) <- c("Country", "Animal_Fats", "Cereals", "Fruits",
   "Oilcrops", "Pulses", "Spices", "Starchy_Roots", "Stimulants", "Treenuts",
   "Vegetal Products", "Vegetable_Oils", "Vegetables", "Obesity", "Undernourished",
   "Population", "Confirmed", "Deaths")
```

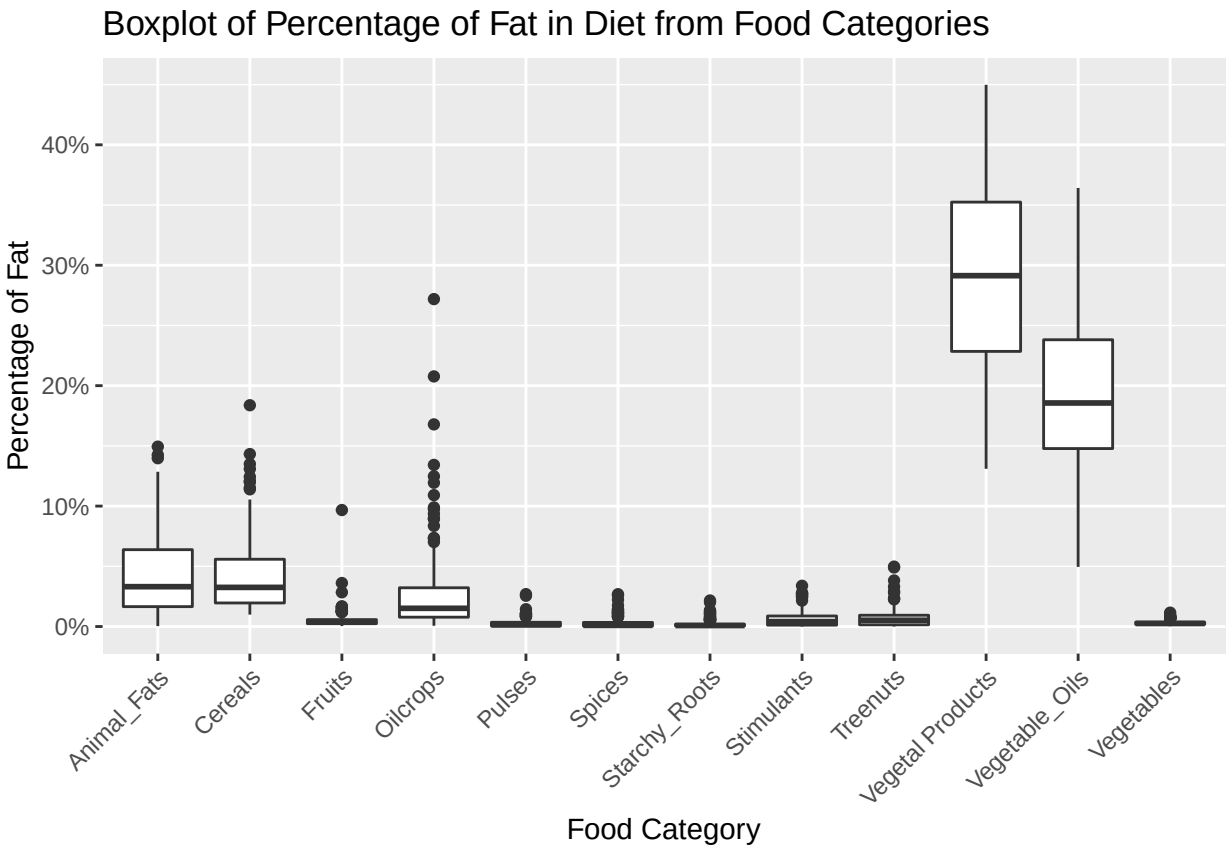
Part 1

Provide a descriptive analysis of your variables. This should include histograms and fitted distributions, correlation plot, boxplots, scatterplots, and statistical summaries (e.g., the five-number summary). All figures must include comments.

Columns in dataset:
* Fat Supply Measures - Average percentage (out of 100) of fat in diet that comes from each category of food - Categories included: Animal_Fats, Cereals, Fruits, Oilcrops, Pulses, Spices, Starchy_Roots, Stimulants, Treenuts, Vegetal Products, Vegetable_oils, and Vegetables

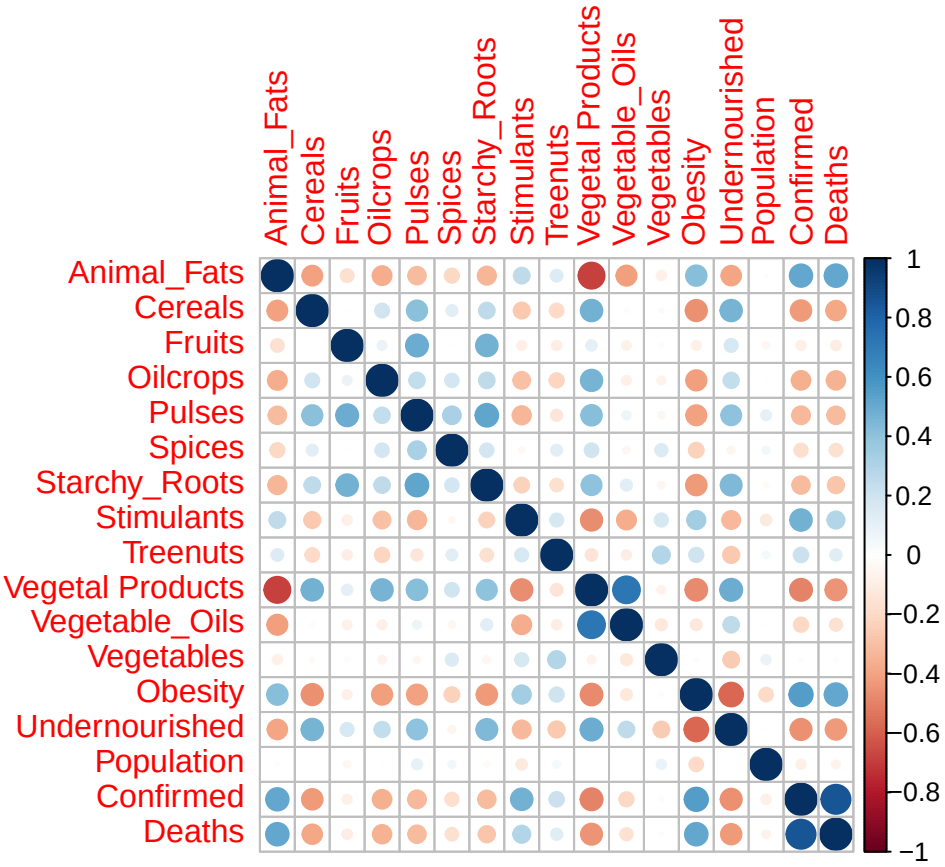
- Population Health Measures - Percentage of the population that falls into each category
  - Obesity and Undernourished
- Population and COVID Measures
  - Population - Population of country
  - Confirmed - Percentage of population with a confirmed positive test for COVID-19
  - Deaths - Percentage of population that died from COVID-19

```r
# create a boxplot of food categories


# melt the data into long form
fat_data <- melt(backAICdata.plus[,1:13], id = "Country")

# create boxplots
ggplot(fat_data, aes(x = variable, y = value)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_y_continuous(labels = label_percent(scale = 1)) +
  xlab("Food Category") +
  ylab("Percentage of Fat") +
  ggtitle("Boxplot of Percentage of Fat in Diet from Food Categories")
```



From the above boxplots, we can see that Vegetal Products and Vegetable Oils are major sources of fat for all countries, while the average values for other categories are low. We can also see that Oilcrops has a relatively large amount of high outliers compared to other groups.

```
# correlation plot of all variables
library(corrplot)
corrplot(cor(backAICdata.plus[,-1]), method = "circle")
```



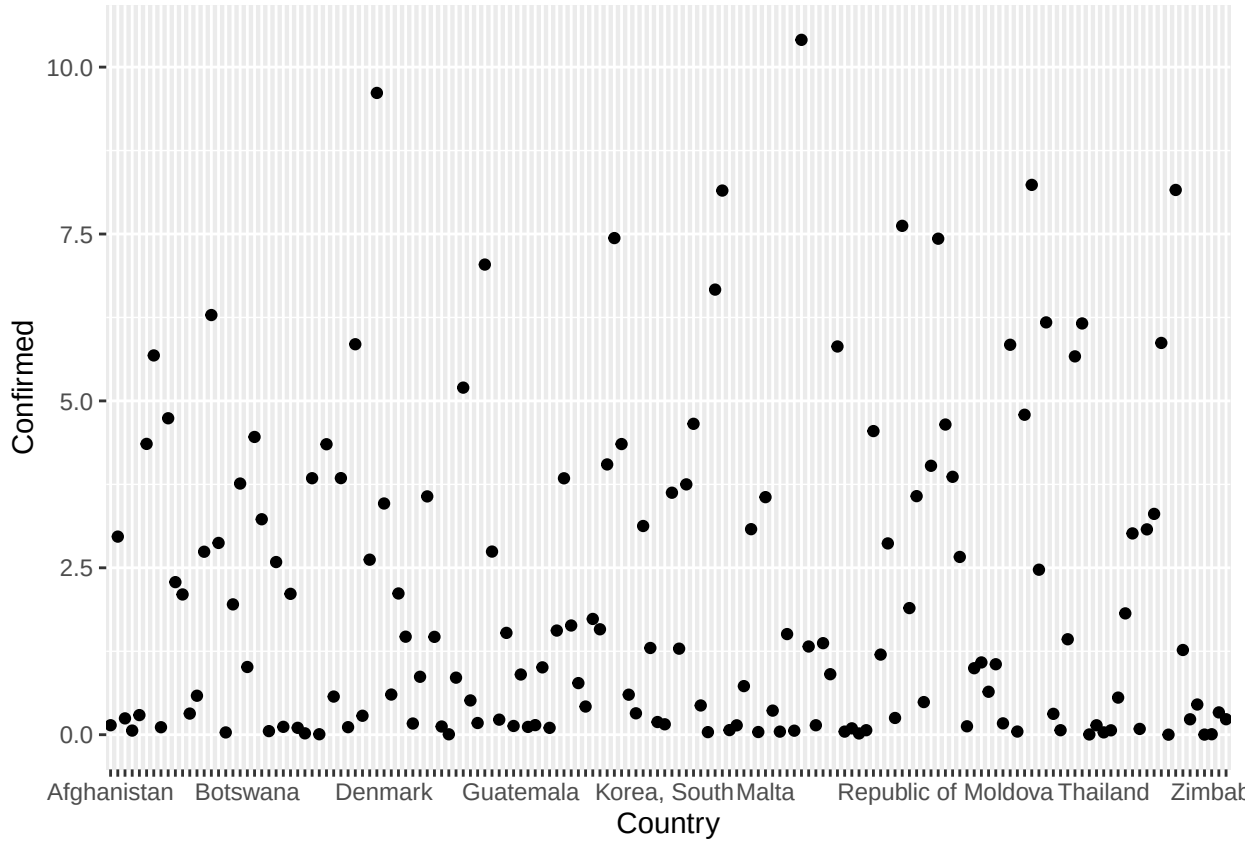
From the above correlation plot, we can see some interesting correlations between some food groups, such as between Vetegal Procucts and Animal Fats. We also see that Obesity and Undernourished are strongly negatively correlated, which makes sense, and that there is a very high correlation between Confirmed Cases and Deaths, which is also to be expected.

```
# five number summaries for each numeric column

apply(backAICdata.plus[,-1], 2, summary)
```

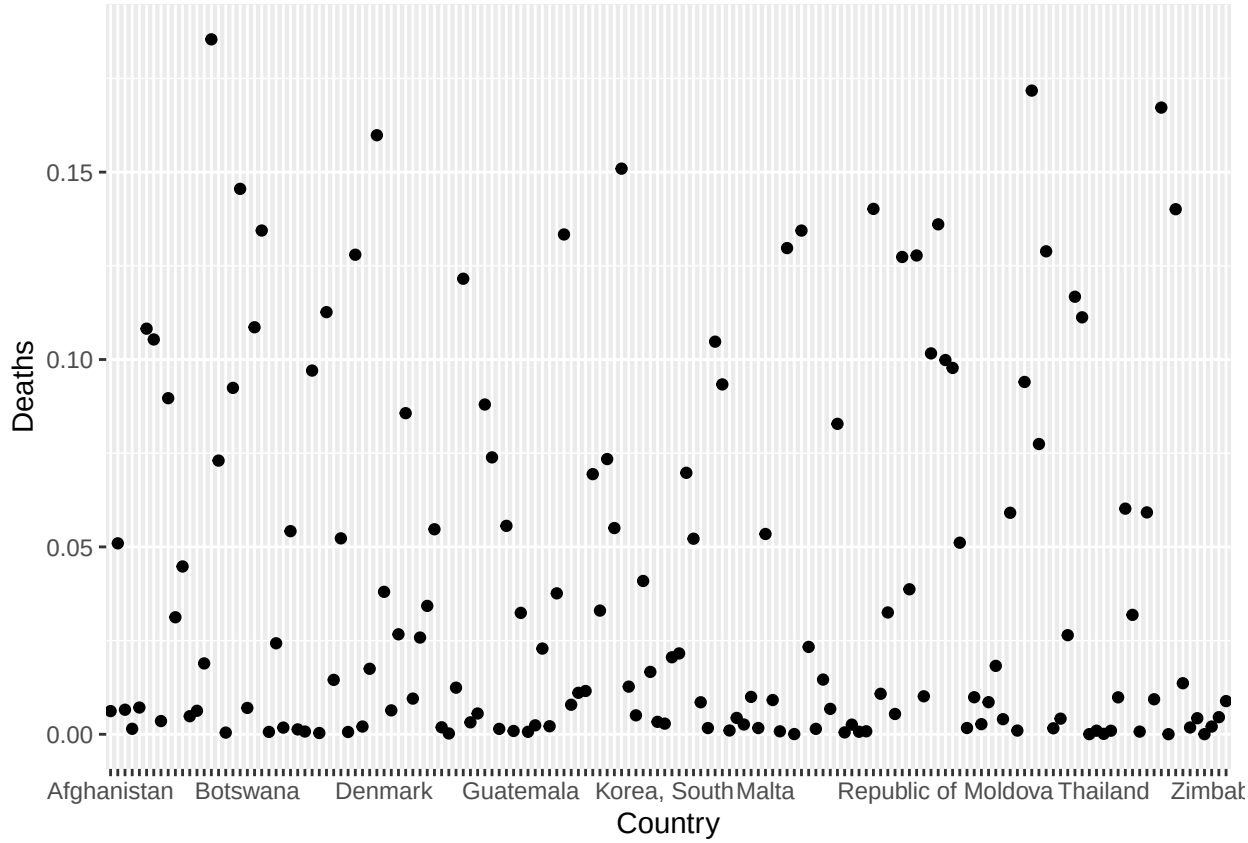
```
##         Animal_Fats   Cereals    Fruits  Oilcrops    Pulses    Spices
## Min.       0.034800  0.990800 0.0373000  0.064000 0.0000000 0.0000000
## 1st Qu.    1.652325  1.957100 0.2365750  0.778650 0.0466250 0.0377750
## Median     3.307600  3.250700 0.3660000  1.512000 0.1423000 0.1000500
## Mean       4.226423  4.346977 0.5428968  2.825626 0.2654897 0.2845878
## 3rd Qu.    6.381550  5.587000 0.5734500  3.218400 0.3518250 0.3418250
## Max.      14.937300 18.376300 9.6727000 27.189200 2.6909000 2.6851000
##         Starchy_Roots Stimulants  Treenuts Vegetal Products Vegetable_Oils
## Min.        0.0124000  0.0000000 0.0000000         13.09820        4.95490
## 1st Qu.     0.0472750  0.1171500 0.1451250         22.84708       14.77958
## Median      0.0846000  0.4103000 0.5204000         29.13450       18.56225
## Mean        0.2158314  0.6533244 0.7410333         29.29365       19.05175
## 3rd Qu.     0.1941250  0.8760750 0.9371250         35.24250       23.81273
## Max.        2.1636000  3.3838000 4.9756000         44.98180       36.41860
##         Vegetables  Obesity Undernourished Population    Confirmed       Deaths
## Min.     0.0263000  2.10000           2.50      98000  0.000852111 3.515586e-05
## 1st Qu.  0.1808000  8.92500           2.50    3534000  0.164875969 2.680607e-03
```

```
## Median   0.2521500 21.30000           6.40   10689500  1.234634653 1.455834e-02
## Mean     0.3090872 18.70449          10.95   47797346  2.124024943 4.138856e-02
## 3rd Qu.  0.3660250 25.70000          13.40   34390750  3.570670605 7.313225e-02
## Max.     1.1538000 37.30000          59.60 1402385000 10.408199357 1.854277e-01
```

```
library(ggplot2)
# Percentage of confirmed cases by country
ggplot(data = backAICdata.plus, aes(x=Country, y=Confirmed, label= Country)) +
  geom_point() + scale_x_discrete(guide = guide_axis(check.overlap = TRUE))
```




```
# Percentage of deaths by country
ggplot(data = backAICdata.plus, aes(x=Country, y=Deaths, label= Country)) +
  geom_point() + scale_x_discrete(guide = guide_axis(check.overlap = TRUE))
```



The above scatterplot details the percentage of confirmed cases in each country. Here it can be seen that the majority of cases lie between zero and 2.5%. From this graphic it can be seen that the highest percentage of covid cases is above 10%.

Part 2

Estimate a multiple linear regression model that includes all the main effects only (i.e., no interactions nor higher order terms). We will use this model as a baseline. Comment on the statistical and economic significance of your estimates. Also, make sure to provide an interpretation of your estimates.

```
# To reduce the number of columns in our dataset to a more workable amount,
# we used backward selection with AIC to pick the predictors we wanted to include.
# The dataset used for question one used only the selected columns.

model_all <- lm(data$Deaths ~ data$`Animal Products` + data$`Animal fats` +
  data$`Cereals - Excluding Beer` + data$Eggs + data$`Fish, Seafood` +
    data$`Fruits - Excluding Wine` + data$Meat + data$`Milk - Excluding Butter` +
    data$Offals + data$Oilcrops + data$Pulses + data$Spices + data$`Starchy Roots` +
    data$Stimulants + data$Treenuts + data$`Vegetal Products` +
    data$`Vegetable Oils` + data$Vegetables + data$Obesity + data$Undernourished + data$Population)

#summary(model_all)

n <- length(data$Deaths)
```

```r
backAIC <- step(model_all ,direction="backward", data=data)
```

```r
# Baseline Model
summary(backAIC)
```

```
##
## Call:
## lm(formula = data$Deaths ~ data$`Animal fats` + data$`Cereals - Excluding Beer` +
##     data$`Fruits - Excluding Wine` + data$Oilcrops + data$Pulses +
##     data$Spices + data$`Starchy Roots` + data$Stimulants + data$Treenuts +
##     data$`Vegetal Products` + data$`Vegetable Oils` + data$Vegetables +
##     data$Obesity)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.108313 -0.021927 -0.003573  0.014199  0.101676
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     0.0129970  0.0265602   0.489 0.625356
## data$`Animal fats`              0.0040506  0.0013450   3.012 0.003077 **
## data$`Cereals - Excluding Beer` 0.1691210  0.0484157   3.493 0.000637 ***
## data$`Fruits - Excluding Wine`  0.1715263  0.0489531   3.504 0.000614 ***
## data$Oilcrops                   0.1690407  0.0482729   3.502 0.000618 ***
## data$Pulses                     0.1612441  0.0484375   3.329 0.001111 **
## data$Spices                     0.1755959  0.0504123   3.483 0.000659 ***
## data$`Starchy Roots`            0.1681581  0.0480719   3.498 0.000626 ***
## data$Stimulants                 0.1798248  0.0496084   3.625 0.000402 ***
## data$Treenuts                   0.1701807  0.0488130   3.486 0.000652 ***
## data$`Vegetal Products`        -0.1702633  0.0483737  -3.520 0.000581 ***
## data$`Vegetable Oils`           0.1702400  0.0482837   3.526 0.000569 ***
## data$Vegetables                 0.1629736  0.0492373   3.310 0.001183 **
## data$Obesity                    0.0014200  0.0004507   3.151 0.001985 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03818 on 142 degrees of freedom
## Multiple R-squared:  0.4464, Adjusted R-squared:  0.3957
## F-statistic: 8.806 on 13 and 142 DF,  p-value: 5.175e-13
```

As can be seen in the model output above, all food categories are statistically significant at the $\alpha = .05$ level. Obesity is also a statistically significant predictor, though Undernourished is surprisingly not statistically significant. The magnitude of the estimates for the food categories is roughly the same, with a 1% increase in fat from each food category leading to a .17 - .19 percent change in expected death rate from COVID-19. What is interesting is that Vegetal Products is the only statistically significant predictor with a negative coefficient, while all other food categories are positive. An increase of 1% in population obesity leads to an increase in .001% of expected COVID-19 death rate.

Part 3

Identify if there are any outliers, high leverage, and or influential observations worth removing. If so, remove them but justify your reason for doing so and re-estimate your model.

We examined for unusual observations using the base residual plot as well as influence and residual plots from the olsrr package to get a comprehensive overview of which points may be unusual observations.

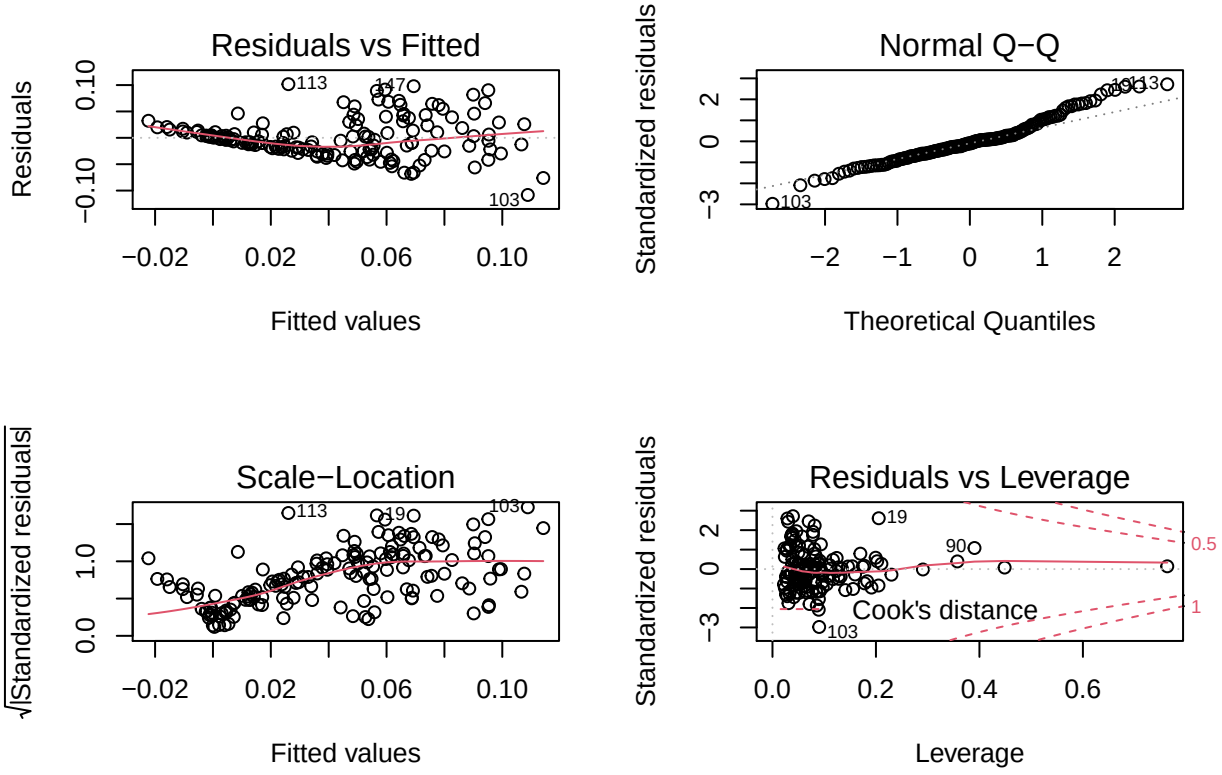
```
library(olsrr)
```

```
## Warning: package 'olsrr' was built under R version 4.0.5
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

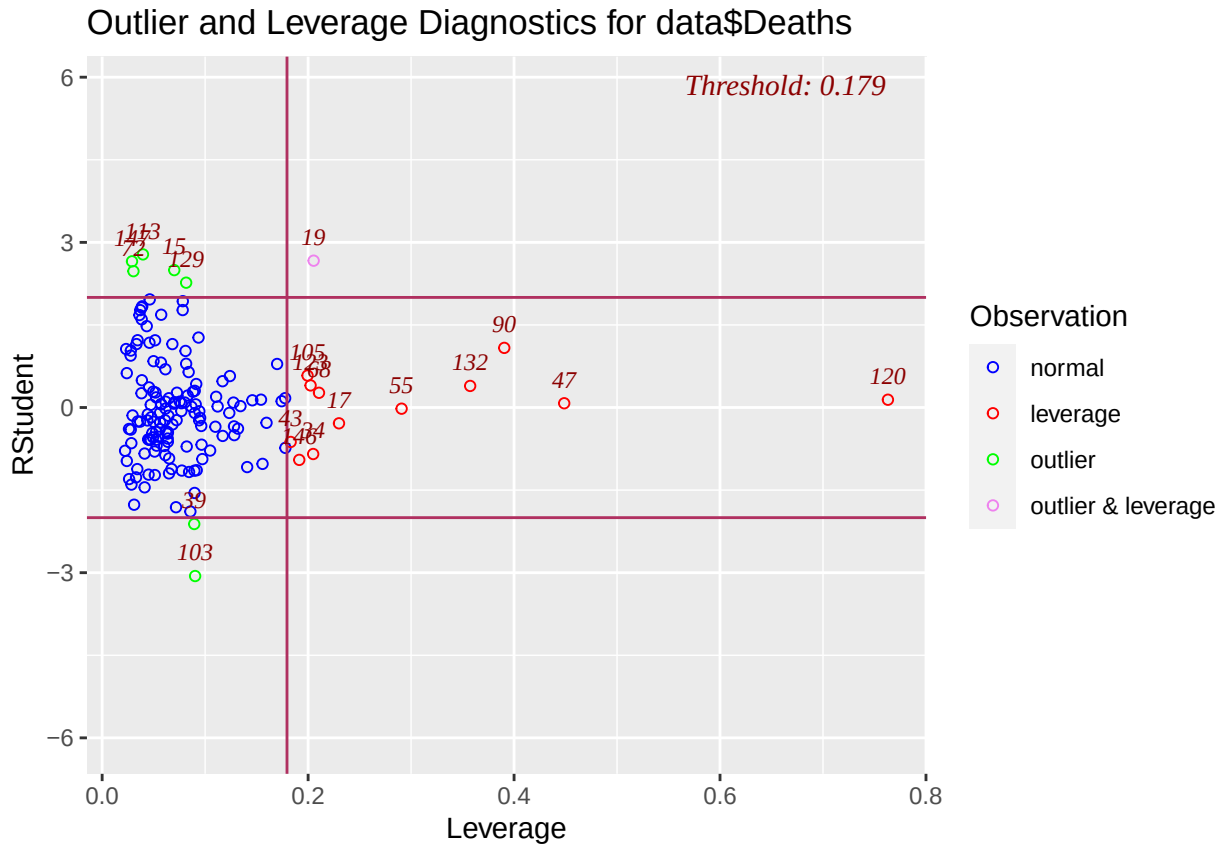
```
par(mfrow=c(2,2))
plot(backAIC)
```




```
influencePlot(backAIC, id=list(n=3))
```

```
##          StudRes        Hat        CookD
## 19    2.66751268 0.20538150 0.1259432991
## 47    0.07737851 0.44868938 0.0003505201
## 90    1.08331865 0.39049322 0.0536400049
## 103 -3.06088780 0.09027770 0.0627145547
## 113   2.78127866 0.03961905 0.0217618173
## 120   0.14167875 0.76304796 0.0046492212
```

```
ols_plot_resid_lev(backAIC)
```



The following observations were identified by both plots as unusual:
* High leverage = 120 (Rwanda), 90 (Maldives), 47 (Ethiopia)
* Outlier = 103 (New Zealand), 113 (Peru) * Influential = 19 (Bosnia/Herzegovina)

We next created new models that used reduced versions of the dataset, removing each type of unusual observation.

```
# Remove the unusual observations from the data with slice
no_highleverage <- backAICdata.plus %>% slice(-c(120,90,47))
no_influential <- backAICdata.plus %>% slice(-19)
no_outlier <- backAICdata.plus %>% slice(-c(103,113))
no_unusual_observations <- backAICdata.plus %>% slice(-c(120,90,47,19,103,113))
```

```
# Create new models without the unusual observations
mod0 <- lm(Deaths~Animal_Fats+Cereals+Fruits+Oilcrops+Pulses+Spices+
             Starchy_Roots+Stimulants+Treenuts+`Vegetal Products`+
             Vegetable_Oils+Vegetables+Obesity, data=backAICdata.plus)
mod1 <- lm(Deaths~Animal_Fats+Cereals+Fruits+Oilcrops+Pulses+Spices+
             Starchy_Roots+Stimulants+Treenuts+`Vegetal Products`+
             Vegetable_Oils+Vegetables+Obesity, data=no_highleverage)
mod2 <- lm(Deaths~Animal_Fats+Cereals+Fruits+Oilcrops+Pulses+Spices+
             Starchy_Roots+Stimulants+Treenuts+`Vegetal Products`+
             Vegetable_Oils+Vegetables+Obesity, data=no_influential)
mod3 <- lm(Deaths~Animal_Fats+Cereals+Fruits+Oilcrops+Pulses+Spices+
```

```
            Starchy_Roots+Stimulants+Treenuts+`Vegetal Products`+
            Vegetable_Oils+Vegetables+Obesity, data=no_outlier)
mod4 <- lm(Deaths~Animal_Fats+Cereals+Fruits+Oilcrops+Pulses+Spices+
            Starchy_Roots+Stimulants+Treenuts+`Vegetal Products`+
            Vegetable_Oils+Vegetables+Obesity, data=no_unusual_observations)

stargazer(mod0, mod1, mod2, mod3, mod4, object.names = TRUE,
          title = "Regression Model Results", column.labels =
            c("Original", "No High Leverage", "No Influential",
              "No Oultlier", "No Unusual Observations"), type = 'latex',
          header = FALSE, no.space = TRUE, single.row = TRUE,
          font.size = "small", column.sep.width = "-15pt")
```

Table 1: Regression Model Results

	Dependent variable:				
			Deaths		
	Original	No High Leverage	No Influential	No Oultlier	No Unusual Observation
	(1)	(2)	(3)	(4)	(5)
	mod0	mod1	mod2	mod3	mod4
Animal_Fats	0.004*** (0.001)	0.004*** (0.001)	0.004*** (0.001)	0.005*** (0.001)	0.005*** (0.001)
Cereals	0.169*** (0.048)	0.193*** (0.054)	0.158*** (0.048)	0.170*** (0.046)	0.177*** (0.050)
Fruits	0.172*** (0.049)	0.194*** (0.055)	0.159*** (0.048)	0.172*** (0.046)	0.175*** (0.051)
Oilcrops	0.169*** (0.048)	0.193*** (0.054)	0.158*** (0.047)	0.170*** (0.046)	0.177*** (0.050)
Pulses	0.161*** (0.048)	0.185*** (0.057)	0.153*** (0.048)	0.161*** (0.046)	0.165*** (0.052)
Spices	0.176*** (0.050)	0.199*** (0.057)	0.156*** (0.050)	0.178*** (0.048)	0.175*** (0.052)
Starchy_Roots	0.168*** (0.048)	0.192*** (0.054)	0.159*** (0.047)	0.168*** (0.046)	0.176*** (0.050)
Stimulants	0.180*** (0.050)	0.204*** (0.056)	0.166*** (0.049)	0.179*** (0.047)	0.182*** (0.051)
Treenuts	0.170*** (0.049)	0.192*** (0.054)	0.160*** (0.048)	0.172*** (0.046)	0.178*** (0.050)
'Vegetal Products'	−0.170*** (0.048)	−0.194*** (0.054)	−0.159*** (0.048)	−0.171*** (0.046)	−0.178*** (0.050)
Vegetable_Oils	0.170*** (0.048)	0.194*** (0.054)	0.159*** (0.047)	0.171*** (0.046)	0.178*** (0.050)
Vegetables	0.163*** (0.049)	0.188*** (0.055)	0.146*** (0.049)	0.162*** (0.047)	0.166*** (0.051)
Obesity	0.001*** (0.0005)	0.002*** (0.0005)	0.001*** (0.0004)	0.001*** (0.0004)	0.002*** (0.0004)
Constant	0.013 (0.027)	0.013 (0.027)	0.015 (0.026)	0.009 (0.025)	0.012 (0.025)
Observations	156	153	155	154	150
R^2	0.446	0.445	0.457	0.498	0.514
Adjusted R^2	0.396	0.393	0.407	0.451	0.468
Residual Std. Error	0.038 (df = 142)	0.038 (df = 139)	0.037 (df = 141)	0.036 (df = 140)	0.035 (df = 136)
F Statistic	8.806*** (df = 13; 142)	8.565*** (df = 13; 139)	9.132*** (df = 13; 141)	10.667*** (df = 13; 140)	11.085*** (df = 13; 136)
Note:					*p<0.1; **p<0.05; ***p<0.0

As can be seen in the table above, the R-squared of the model that removes all outliers ahigh leverage points is a noticeable improvement over the base model including all data points. We noticed that the estimates for the coefficients did not change much between models, indicating that while the outliers were negatively affecting the accuracy metrics of the model, they were not changing the estimates for the parameters themselves. We decided to remove the outliers and leverage points because we wanted to be able to compare future models against each other fairly, without the outliers influencing the accuracy metrics of those models.

```
# New model
new_model_1 <- lm(Deaths~Animal_Fats+Cereals+Fruits+Oilcrops+Pulses+Spices+
                    Starchy_Roots+Stimulants+Treenuts+`Vegetal Products`+
                    Vegetable_Oils+Vegetables+Obesity, data=no_unusual_observations)
```

```
summary(new_model_1)
```

```
##
## Call:
## lm(formula = Deaths ~ Animal_Fats + Cereals + Fruits + Oilcrops +
##     Pulses + Spices + Starchy_Roots + Stimulants + Treenuts +
##     `Vegetal Products` + Vegetable_Oils + Vegetables + Obesity,
##     data = no_unusual_observations)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.080550 -0.019870 -0.003948  0.014479  0.100033
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.011575   0.024851   0.466 0.642119
## Animal_Fats         0.004995   0.001281   3.899 0.000151 ***
## Cereals             0.176951   0.050182   3.526 0.000575 ***
## Fruits              0.175030   0.050870   3.441 0.000771 ***
## Oilcrops            0.177386   0.050053   3.544 0.000541 ***
## Pulses              0.164786   0.052225   3.155 0.001974 **
## Spices              0.175371   0.052498   3.341 0.001080 **
## Starchy_Roots       0.176484   0.049783   3.545 0.000539 ***
## Stimulants          0.182433   0.051415   3.548 0.000533 ***
## Treenuts            0.177985   0.049779   3.576 0.000485 ***
## `Vegetal Products` -0.177986   0.050173  -3.547 0.000534 ***
## Vegetable_Oils      0.178008   0.050107   3.553 0.000525 ***
## Vegetables          0.166187   0.051183   3.247 0.001469 **
## Obesity             0.001604   0.000431   3.722 0.000288 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03533 on 136 degrees of freedom
## Multiple R-squared:  0.5145, Adjusted R-squared:  0.4681
## F-statistic: 11.09 on 13 and 136 DF,  p-value: 6.9e-16
```

Part 4

Use Mallows Cp for identifying which terms you will keep in the model (based on part 3) and also use the Boruta algorithm for variable selection. Based on the two results, determine which subset of predictors you will keep.

```
# Since Mallows CP has a lower number when testing our new_model_1, we will proceed with that model.
ols_mallows_cp(new_model_1, model_all)
```

```
## [1] -8.420024
```

```
ols_mallows_cp(backAIC, model_all)
```

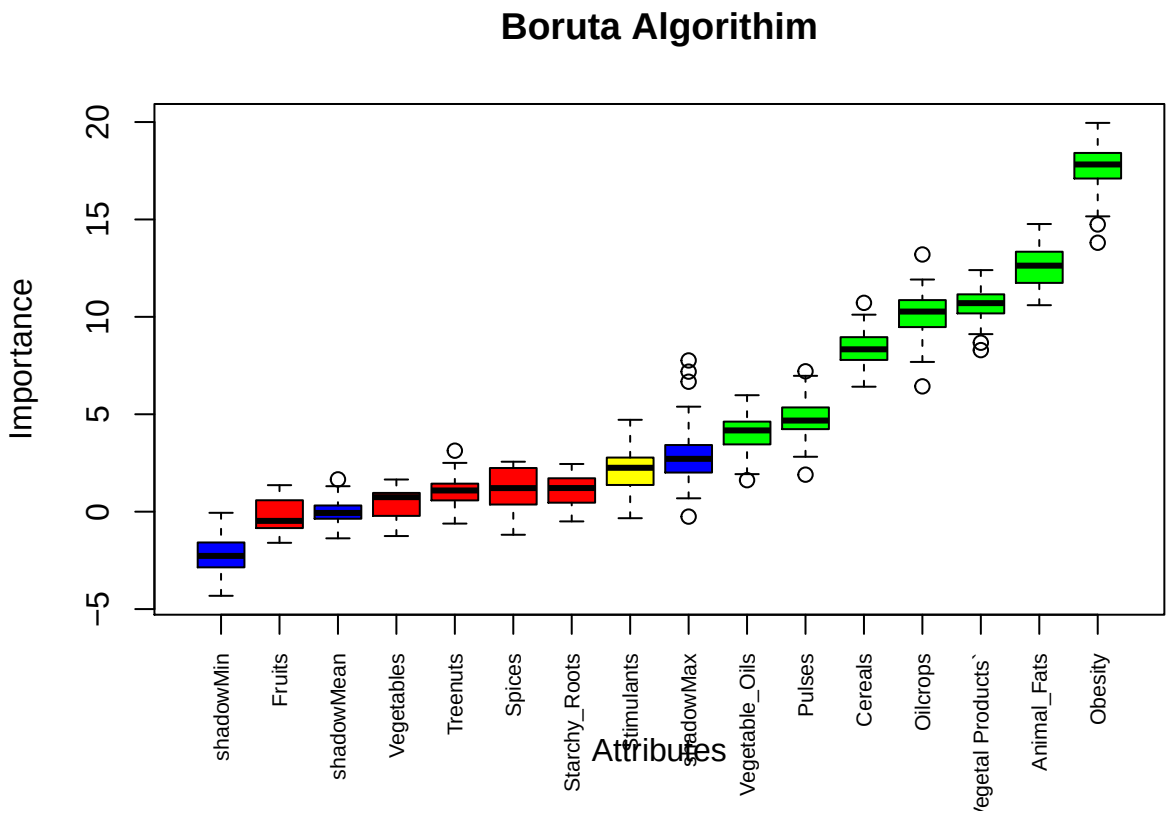
```
## [1] 10.44401
```

```
library(Boruta)
```

```
## Warning: package 'Boruta' was built under R version 4.0.5
```

```
Bor.res <- Boruta(Deaths~Animal_Fats+Cereals+Fruits+Oilcrops+Pulses+Spices+
                    Starchy_Roots+Stimulants+Treenuts+`Vegetal Products`+
                    Vegetable_Oils+Vegetables+Obesity, data = no_unusual_observations,
                  doTrace = 2 )
```

```
plot(Bor.res, xlab = "Attributes", xaxt = "n", main = "Boruta Algorithim")

lz<-lapply(1:ncol(Bor.res$ImpHistory),function(i)
Bor.res$ImpHistory[is.finite(Bor.res$ImpHistory[,i]),i])
names(lz) <- colnames(Bor.res$ImpHistory)
Labels <- sort(sapply(lz,median))
axis(side = 1,las=2,labels = names(Labels),
at = 1:ncol(Bor.res$ImpHistory), cex.axis = 0.7)
```




```
# Testing to see which variables we want to remove
sorted_vars <- attStats(Bor.res)[order(-attStats(Bor.res)$meanImp),]
print(sorted_vars)
```

```
##                       meanImp  medianImp     minImp    maxImp   normHits
## Obesity            17.6984206 17.8224609 13.8050425 19.955594 1.00000000
## Animal_Fats        12.5706143 12.6288479 10.5993562 14.765442 1.00000000
## `Vegetal Products` 10.6808348 10.7035321  8.2924748 12.400460 1.00000000
## Oilcrops           10.1462635 10.2712677  6.4325966 13.201918 1.00000000
## Cereals             8.3850288  8.3331003  6.4136872 10.719157 1.00000000
## Pulses              4.7879618  4.6775063  1.9003809  7.205380 0.92929293
## Vegetable_Oils      4.0938204  4.1704235  1.6164735  5.977203 0.82828283
## Stimulants          2.0954668  2.2532366 -0.3360883  4.715206 0.37373737
## Starchy_Roots       1.1142363  1.2124076 -0.4997432  2.447596 0.02020202
## Treenuts            1.0957088  1.0882064 -0.6119642  3.130963 0.01010101
## Spices              1.0415566  1.2102707 -1.1852769  2.563577 0.03030303
## Vegetables          0.4237355  0.7441404 -1.2517944  1.651195 0.02020202
## Fruits             -0.2073581 -0.4733795 -1.5983819  1.360835 0.00000000
##                       decision
## Obesity            Confirmed
## Animal_Fats        Confirmed
## `Vegetal Products` Confirmed
## Oilcrops           Confirmed
## Cereals            Confirmed
## Pulses             Confirmed
## Vegetable_Oils     Confirmed
## Stimulants         Tentative
## Starchy_Roots       Rejected
## Treenuts            Rejected
## Spices              Rejected
## Vegetables          Rejected
## Fruits              Rejected
```

```
# We will reject: Stimulants, Treenuts, Starchy Roots, Vegetables, Spices, Fruits

# Our New Model
new_model_2 <- lm(Deaths~Animal_Fats+Cereals+Oilcrops+Pulses+`Vegetal Products`+
                    Vegetable_Oils+Obesity, data=no_unusual_observations)
summary(new_model_2)
```

```
##
## Call:
## lm(formula = Deaths ~ Animal_Fats + Cereals + Oilcrops + Pulses +
##     `Vegetal Products` + Vegetable_Oils + Obesity, data = no_unusual_observations)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.080164 -0.024293 -0.003623  0.014763  0.103286
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.0089355  0.0235770  -0.379 0.705259
## Animal_Fats         0.0060034  0.0012583   4.771  4.5e-06 ***
## Cereals             0.0001003  0.0024696   0.041 0.967667
## Oilcrops            0.0007169  0.0024889   0.288 0.773735
## Pulses             -0.0156242  0.0140885  -1.109 0.269302
## `Vegetal Products` -0.0010518  0.0024610  -0.427 0.669725
## Vegetable_Oils      0.0013550  0.0023390   0.579 0.563285
```

```
## Obesity             0.0016209  0.0004172   3.885 0.000156 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03635 on 142 degrees of freedom
## Multiple R-squared:  0.4635, Adjusted R-squared:  0.437
## F-statistic: 17.52 on 7 and 142 DF,  p-value: < 2.2e-16
```

Part 5

Test for multicollinearity using VIF on the model from (4) . Based on the test, remove any appropriate variables, and estimate a new regression model based on these findings.

```
vif(new_model_2) # We will remove any variable with a VIF over 5 to satisfy collinearity assumption
```

```
##        Animal_Fats            Cereals           Oilcrops             Pulses
##           1.967723           7.059653          10.259789           1.596918
## `Vegetal Products`     Vegetable_Oils            Obesity
##          45.029589          27.253117           1.597495
```

```
# Vegetal products has the highest VIF, so we remove that and repeat
vif_1 <- lm(Deaths~Animal_Fats+Cereals+Oilcrops+Pulses + Vegetable_Oils+Obesity,
            data=no_unusual_observations)
vif(vif_1)
```

```
##    Animal_Fats        Cereals       Oilcrops         Pulses Vegetable_Oils
##       1.823054       1.614271       1.524979       1.478783       1.440639
##        Obesity
##       1.575614
```

After removing Vegetal Products, all other predictors have a VIF under 5, so we will keep all of those.

```
new_model_3 <- lm(Deaths~Animal_Fats+Cereals+Oilcrops+Pulses + Vegetable_Oils+
                    Obesity, data=no_unusual_observations)

#New Model
summary(new_model_3)
```

```
##
## Call:
## lm(formula = Deaths ~ Animal_Fats + Cereals + Oilcrops + Pulses +
##     Vegetable_Oils + Obesity, data = no_unusual_observations)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.079008 -0.024487 -0.004085  0.015139  0.103496
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.0140714  0.0202269  -0.696 0.487759
## Animal_Fats     0.0061492  0.0012077   5.092  1.1e-06 ***
```

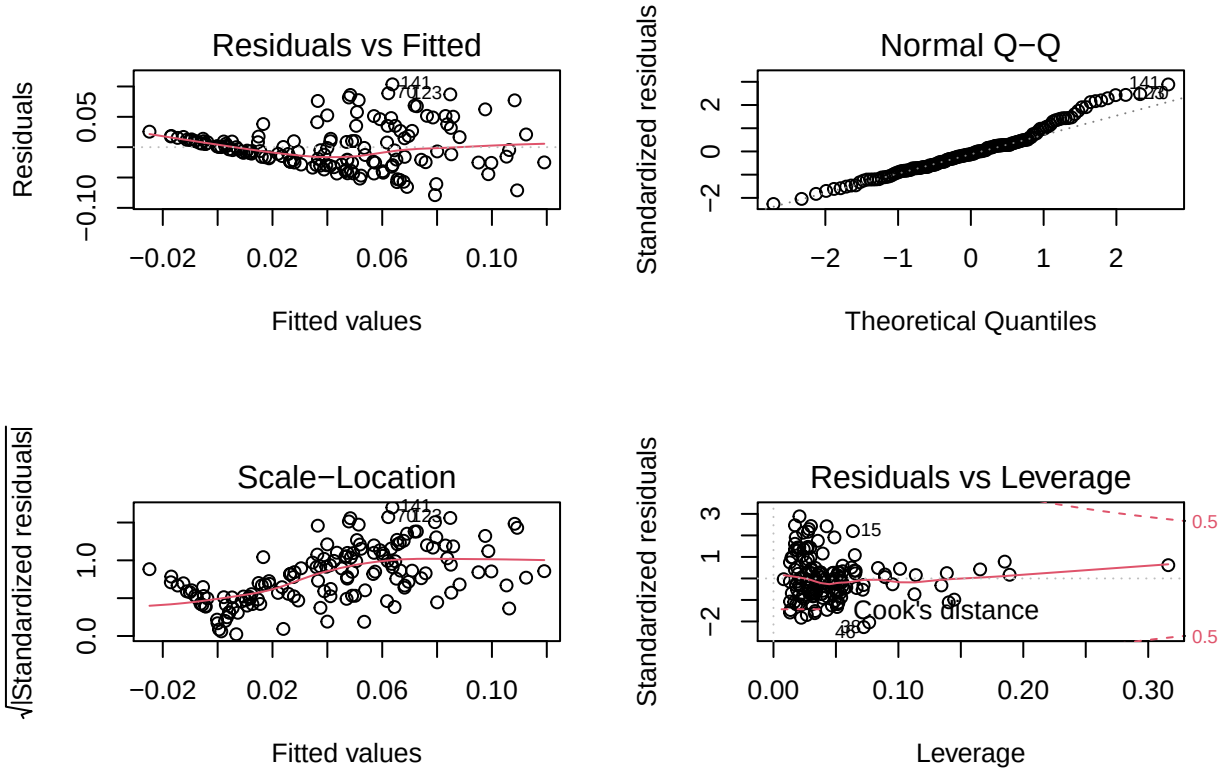
```
## Cereals        -0.0008268  0.0011776  -0.702 0.483758
## Oilcrops       -0.0002647  0.0009568  -0.277 0.782492
## Pulses         -0.0172620  0.0135186  -1.277 0.203702
## Vegetable_Oils  0.0003821  0.0005362   0.713 0.477266
## Obesity         0.0016001  0.0004131   3.873 0.000163 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03625 on 143 degrees of freedom
## Multiple R-squared:  0.4628, Adjusted R-squared:  0.4403
## F-statistic: 20.53 on 6 and 143 DF,  p-value: < 2.2e-16
```

In our new model, we can see that Animal Fats and Obesity are the only predictors that are statistically significant.

Part 6

For your model in part (5) plot the respective residuals vs. y_hat and comment on your results.

```
par(mfrow=c(2,2))
plot(new_model_3)
```



From the residuals vs fitted plot it can be seen that our residuals appear to spread out the greater our fitted value is. The red smoother runs close to zero which is a good thing.

Part 7

For your model in part (5) perform a RESET test and comment on your results.

```
resettest(new_model_3, power = 2, type = "regressor")
```

```
##
##  RESET test
##
## data:  new_model_3
## RESET = 1.092, df1 = 6, df2 = 137, p-value = 0.3703
```

Here we tested our model by testing our model against a quadratic. Our result is a p-value of 0.3703 which means that we fail to reject the null hypothesis of higher powers existing and should not consider higher powers in later version of our model.

Part 8

For your model in part (5) test for heteroskedasticity and comment on your results. If you identify heteroskedasticy, make sure to account for it before moving on to (9).

Below we will test for heteroskedacity using the ncvTest and bptest.

```
# Non-constant error variance: Ho: variance = constant
ncvTest(new_model_3) # Reject Ho
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 31.85058, Df = 1, p = 1.665e-08
```

```
# BP test
bptest(new_model_3) #Reject Ho
```

```
##
##  studentized Breusch-Pagan test
##
## data:  new_model_3
## BP = 31.9, df = 6, p-value = 1.705e-05
```

From the above tests it can be seen that heteroskedacity is present in our data, as the p-value for both tests is essentially zero. In order to account for that we will now run our model with robust white standard errors. Here our new standard errors can be found.

```
cov1 <- hccm(new_model_3, type = "hc1")
#Have our model account for those errors.
new_model_3_adjusted <- coeftest(new_model_3, vcov. = cov1)
tidy(new_model_3_adjusted)
```

```
## # A tibble: 7 x 5
##   term            estimate std.error statistic   p.value
##   <chr>              <dbl>     <dbl>     <dbl>     <dbl>
```

```
## 1 (Intercept)    -0.0141    0.0195      -0.722 0.472
## 2 Animal_Fats     0.00615   0.00141      4.35  0.0000259
## 3 Cereals        -0.000827  0.000928    -0.891 0.374
## 4 Oilcrops       -0.000265  0.000730    -0.362 0.718
## 5 Pulses         -0.0173    0.00837     -2.06  0.0410
## 6 Vegetable_Oils  0.000382  0.000502     0.761 0.448
## 7 Obesity         0.00160   0.000384     4.16  0.0000538
```

The above table has the adjusted standard errors for the estimates when heterskedasticity is accounted for. We can see that Pulses now becomes statistically significant when using the adjusted standard errors despite not being significant earlier.

Part 9

Estimate a model based on all your findings that also includes interaction terms (if appropriate) and if needed, any higher power terms. Comment on the performance of this model compared to your other models. Make sure to use AIC and BIC for model comparison.

```
# Our RESET test suggested there is no existence of higher power terms,
# so we will not test for those

# Testing for interaction
interaction_terms <- lm(Deaths~Animal_Fats+ Cereals + Oilcrops + Pulses +
  Vegetable_Oils + Obesity + Animal_Fats:Cereals + Animal_Fats:Oilcrops +
    Animal_Fats:Pulses + Animal_Fats:Vegetable_Oils + Animal_Fats:Obesity +
    Cereals:Oilcrops + Cereals:Pulses +
    Cereals:Vegetable_Oils + Cereals:Obesity + Oilcrops:Pulses +
    Oilcrops:Vegetable_Oils + Oilcrops:Obesity +
    Pulses:Vegetable_Oils + Pulses:Obesity +  Vegetable_Oils:Obesity,
  data=no_unusual_observations)

summary(interaction_terms) #Animal_Fats:Vegetable_oils is statistically significant, this will be added
```

```
##
## Call:
## lm(formula = Deaths ~ Animal_Fats + Cereals + Oilcrops + Pulses +
##     Vegetable_Oils + Obesity + Animal_Fats:Cereals + Animal_Fats:Oilcrops +
##     Animal_Fats:Pulses + Animal_Fats:Vegetable_Oils + Animal_Fats:Obesity +
##     Cereals:Oilcrops + Cereals:Pulses + Cereals:Vegetable_Oils +
##     Cereals:Obesity + Oilcrops:Pulses + Oilcrops:Vegetable_Oils +
##     Oilcrops:Obesity + Pulses:Vegetable_Oils + Pulses:Obesity +
##     Vegetable_Oils:Obesity, data = no_unusual_observations)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.085936 -0.020716 -0.001209  0.014357  0.097553
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -4.803e-02  6.128e-02  -0.784   0.4346
## Animal_Fats                 -3.614e-03  6.989e-03  -0.517   0.6060
## Cereals                      6.040e-03  6.080e-03   0.993   0.3224
## Oilcrops                     2.768e-03  3.314e-03   0.835   0.4051
```

```
## Pulses                       3.023e-03  1.255e-01   0.024   0.9808
## Vegetable_Oils               1.327e-03  2.004e-03   0.662   0.5090
## Obesity                      4.814e-03  2.456e-03   1.960   0.0521 .
## Animal_Fats:Cereals         -2.291e-04  6.214e-04  -0.369   0.7129
## Animal_Fats:Oilcrops        -8.267e-04  7.070e-04  -1.169   0.2444
## Animal_Fats:Pulses          -1.117e-03  7.576e-03  -0.147   0.8830
## Animal_Fats:Vegetable_Oils   5.216e-04  2.165e-04   2.410   0.0174 *
## Animal_Fats:Obesity          4.998e-05  2.176e-04   0.230   0.8187
## Cereals:Oilcrops             1.051e-04  3.495e-04   0.301   0.7642
## Cereals:Pulses               1.262e-03  5.241e-03   0.241   0.8100
## Cereals:Vegetable_Oils      -3.031e-04  2.207e-04  -1.373   0.1721
## Cereals:Obesity             -1.282e-04  1.545e-04  -0.829   0.4084
## Oilcrops:Pulses             -1.676e-03  4.793e-03  -0.350   0.7271
## Oilcrops:Vegetable_Oils      1.233e-05  1.385e-04   0.089   0.9292
## Oilcrops:Obesity            -2.839e-04  2.254e-04  -1.260   0.2100
## Pulses:Vegetable_Oils        2.929e-04  3.133e-03   0.094   0.9256
## Pulses:Obesity              -1.431e-03  1.993e-03  -0.718   0.4740
## Vegetable_Oils:Obesity      -9.205e-05  7.552e-05  -1.219   0.2252
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03452 on 128 degrees of freedom
## Multiple R-squared:  0.564,  Adjusted R-squared:  0.4924
## F-statistic: 7.884 on 21 and 128 DF,  p-value: 1.105e-14
```

```
new_model_4 <- lm(Deaths~Animal_Fats+Cereals+Oilcrops+Pulses + Vegetable_Oils+
                    Obesity + Animal_Fats:Vegetable_Oils, data=no_unusual_observations)

summary(new_model_4)
```

```
##
## Call:
## lm(formula = Deaths ~ Animal_Fats + Cereals + Oilcrops + Pulses +
##     Vegetable_Oils + Obesity + Animal_Fats:Vegetable_Oils, data = no_unusual_observations)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.077977 -0.025445 -0.001541  0.016170  0.100535
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 0.0171752  0.0246193   0.698 0.486548
## Animal_Fats                -0.0010627  0.0035308  -0.301 0.763862
## Cereals                    -0.0010108  0.0011657  -0.867 0.387313
## Oilcrops                   -0.0006128  0.0009582  -0.640 0.523513
## Pulses                     -0.0196365  0.0133914  -1.466 0.144762
## Vegetable_Oils             -0.0008681  0.0007824  -1.109 0.269100
## Obesity                     0.0014326  0.0004151   3.451 0.000736 ***
## Animal_Fats:Vegetable_Oils  0.0003984  0.0001836   2.170 0.031663 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03579 on 142 degrees of freedom
## Multiple R-squared:   0.48,  Adjusted R-squared:  0.4544
```

```
## F-statistic: 18.73 on 7 and 142 DF,  p-value: < 2.2e-16
```

```
# Testing with AIC and BIC

AIC(new_model_1, new_model_2, new_model_3, new_model_4)
```

```
##             df       AIC
## new_model_1 15 -561.8820
## new_model_2  9 -558.9032
## new_model_3  8 -560.7103
## new_model_4  9 -563.6042
```

```
BIC(new_model_1, new_model_2, new_model_3, new_model_4)
```

```
##             df       BIC
## new_model_1 15 -516.7225
## new_model_2  9 -531.8074
## new_model_3  8 -536.6252
## new_model_4  9 -536.5085
```

```
# Adding Robust Standard Errors to this new model since we know heteroskedasticity is present

cov2 <- hccm(new_model_4, type = "hc1")
#Have our model account for those errors.
new_model_4_adjusted <- coeftest(new_model_4, vcov. = cov2)

tidy(new_model_4_adjusted)
```

```
## # A tibble: 8 x 5
##   term                        estimate std.error statistic  p.value
##   <chr>                          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                  0.0172    0.0219       0.783 0.435
## 2 Animal_Fats                 -0.00106   0.00337     -0.315 0.753
## 3 Cereals                     -0.00101   0.000937    -1.08  0.282
## 4 Oilcrops                    -0.000613  0.000712    -0.861 0.391
## 5 Pulses                      -0.0196    0.00888     -2.21  0.0286
## 6 Vegetable_Oils              -0.000868  0.000600    -1.45  0.150
## 7 Obesity                      0.00143   0.000389     3.68  0.000327
## 8 Animal_Fats:Vegetable_Oils  0.000398  0.000176     2.27  0.0250
```

As stated above, we did not test for higher powers because the RESET test suggested we should not test our model with quadratic variables. We only tested for interaction terms, and in our model that included all interaction terms we found that the interaction between Animal Fats and Vegetable Oils was statistically significant. This interaction was added to model 3, creating new_model_4. We then went and tested all of our models with AIC and BIC and it was confirmed that new_model_4 had the lowest AIC and was virtually tied in BIC with model 3. This lead us to believe that we had found the best model in model 4. In part 8 we learned that heteroskedacity is present in our data, we took this into cosideration and calculated the robust standard errors for new_model_4, which created new_model_4_adjusted. Again, we see that Pulses becomes statistically significant when accounting for the adjusted standard errors.

Part 10

Evaluate your model performance (from 9) using cross-validation, and also by dividing your data into the traditional 2/3 training and 1/3 testing samples, to evaluate your out-of-sample performance. Comment on your results.

```
# split data into 2/3 train 1/3 test
train <- sample(nrow(no_unusual_observations), nrow(no_unusual_observations) * 2/3)
data_train <- no_unusual_observations[train,]
data_test <- no_unusual_observations[-train,]

# do 5-fold cross validation on the training partition
# using model_vif below as placeholder
fitControl <- trainControl(method="cv", number = 5, savePredictions = T)
model_cv <- train(Deaths~Animal_Fats+Cereals+Oilcrops+Pulses + Vegetable_Oils+
                    Obesity + Animal_Fats:Vegetable_Oils, data=data_train,
                  trControl = fitControl, method = "glm")

model_cv
```

```
## Generalized Linear Model
##
## 100 samples
##   6 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 80, 80, 80, 80, 80
## Resampling results:
##
##   RMSE        Rsquared   MAE
##   0.03465109  0.4733077  0.0274295
```

From our cross validation, we see that the RMSE is .036 and the R-squared is .471 within the cross-validation partition.

```
# make predictions on the testing partition
pred <- predict(model_cv, data_test, vcov. = cov2)

# calculate RMSE
RMSE(pred, data_test$Deaths)
```

```
## [1] 0.03951344
```

The RMSE on our test partition from the model created with the training partition is .037, which is in line with the value from above. This is a good sign for the out-of-sample performance of the model.

Part 11

Provide a short (1 paragraph) summary of your overall conclusions/findings.

When examining the data overall, there were not any massive outliers that stood out. We found that many of the food categories that were included in tracking fat sources were not statistically or practically significant.

Our model was quickly trimmed down from the initial pool of variables that we considered. Through model comparison with AIC and BIC, we found that the best predictors to include were Animal Fats, Cereals, Oilcrops, Pulses, Vegetable Oils, Obesity, and the interaction between Animal Fats and Obesity. Unsurprisingly, Obesity was the most significant predictor in our final model, which is expected given the link between obesity and increased risk to COVID-19. As we went through the parts sequentially, the overall quality of the model improved. We also found that adjusting the errors for heteroskedasticity was a very important step, as Pulses went from insignificant to significant both times that we adjusted.