Project 1

Nancie Kung, Calvin Raab, David Collier and Eitan Shimonovitz

4/21/2021

Load libraries and import the data.

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(leaps)
library(ggplot2)
library(reshape2)
library(scales)
```

```
##
## Attaching package: 'scales'

## The following object is masked from 'package:readr':
##
##     col_factor
```

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

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

```
##
## -- Column specification --------------------------------------------------------------------------
## cols(
##   .default = col_double(),
##   Country = col_character(),
##   Undernourished = col_character(),
##   `Unit (all except Population)` = col_character()
## )
## i Use `spec()` for the full column specifications.
```

Data cleaning process.

```
#select columns that are not filled with zeros
Fat_Supply_data <- Fat_Supply_Quantity_Data %>% select(Country, `Animal Products`, `Animal fats`, `Cerea

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$Undernourished)] <- median(Fat_Supply_data$Underno

data <- Fat_Supply_data

# Here is a dataset that includes the parameters found in backAIC, along with: Country, Population, Con
backAICdata.plus <- data_frame(data$`Country`, data$`Animal fats`, data$`Cereals - Excluding Beer`, data
```

```
## Warning: `data_frame()` is deprecated as of tibble 1.1.0.
## Please use `tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
names(backAICdata.plus) <- c("Country", "Animal_Fats", "Cereals", "Fruits", "Oilcrops", "Pulses", "Spic
```

Part 1

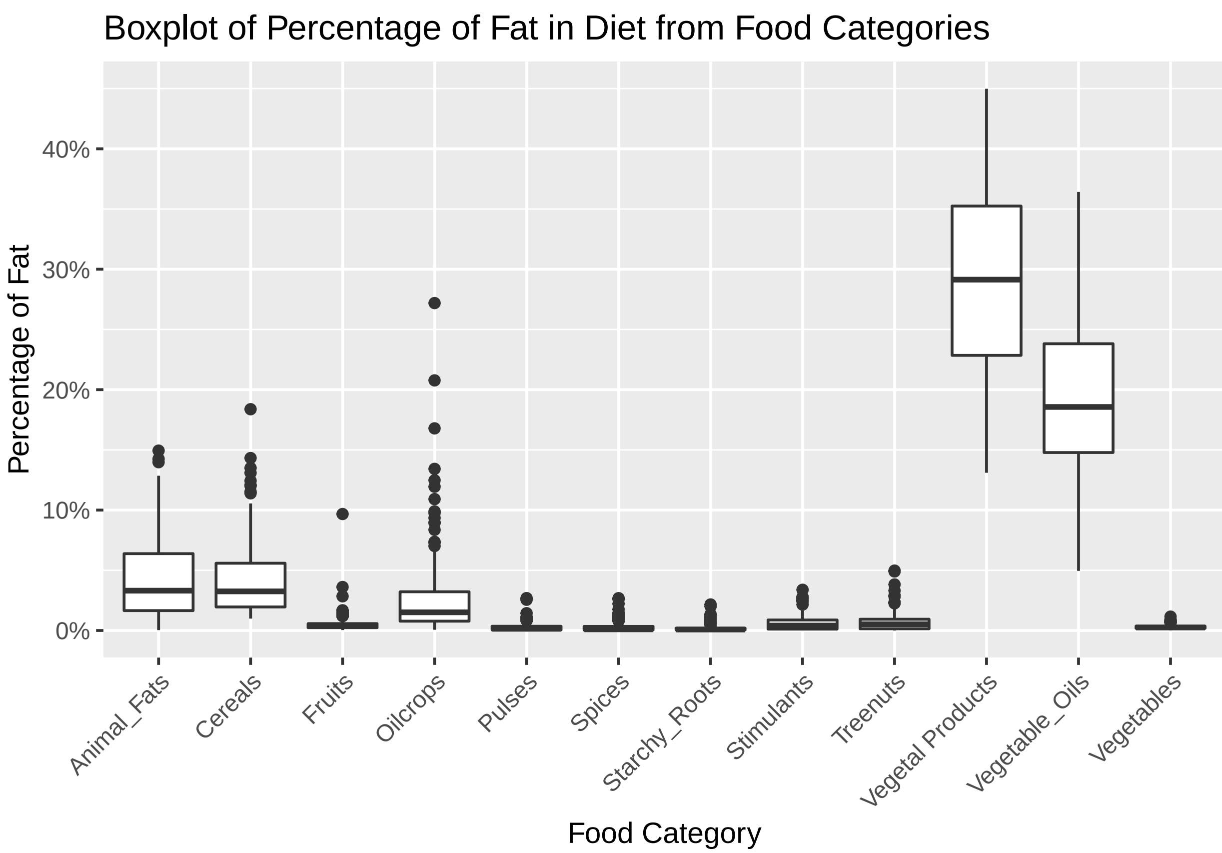
Provide a descriptive analysis of your variables. This should include histograms and fitted distributions, correlation plot, boxplots, scatterplots, and statistical summaries (e.g., the five-number summary). All figures must include comments.

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

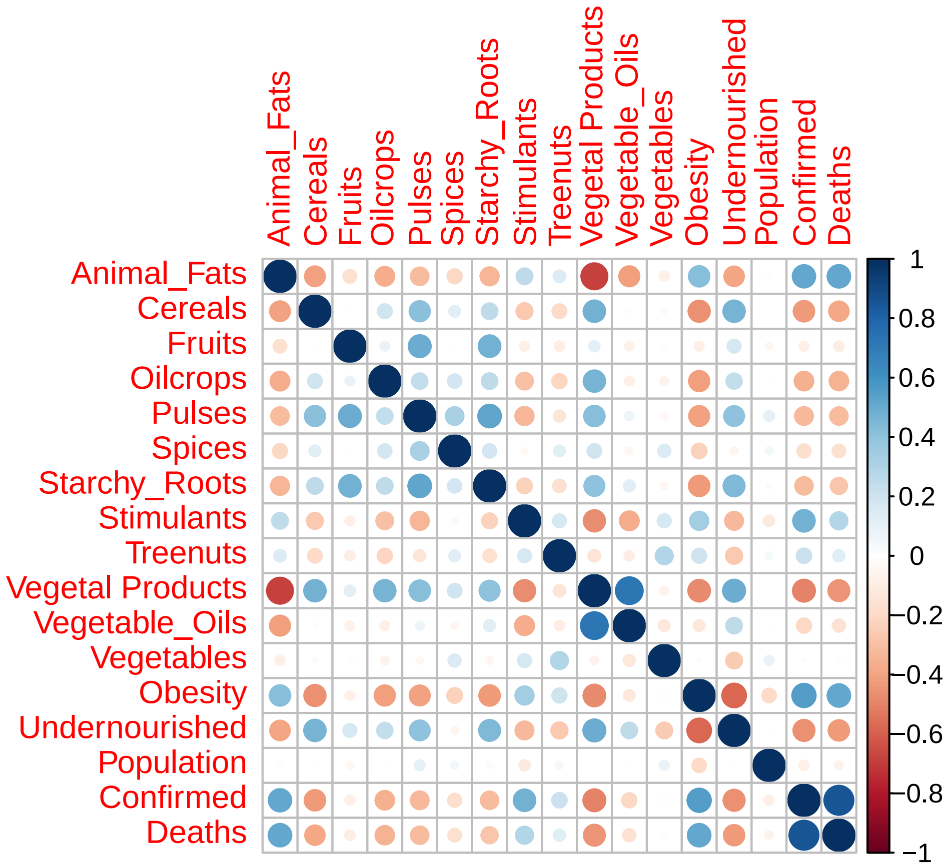
– Deaths - Percentage of population that died from COVID-19

```
# 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")
```



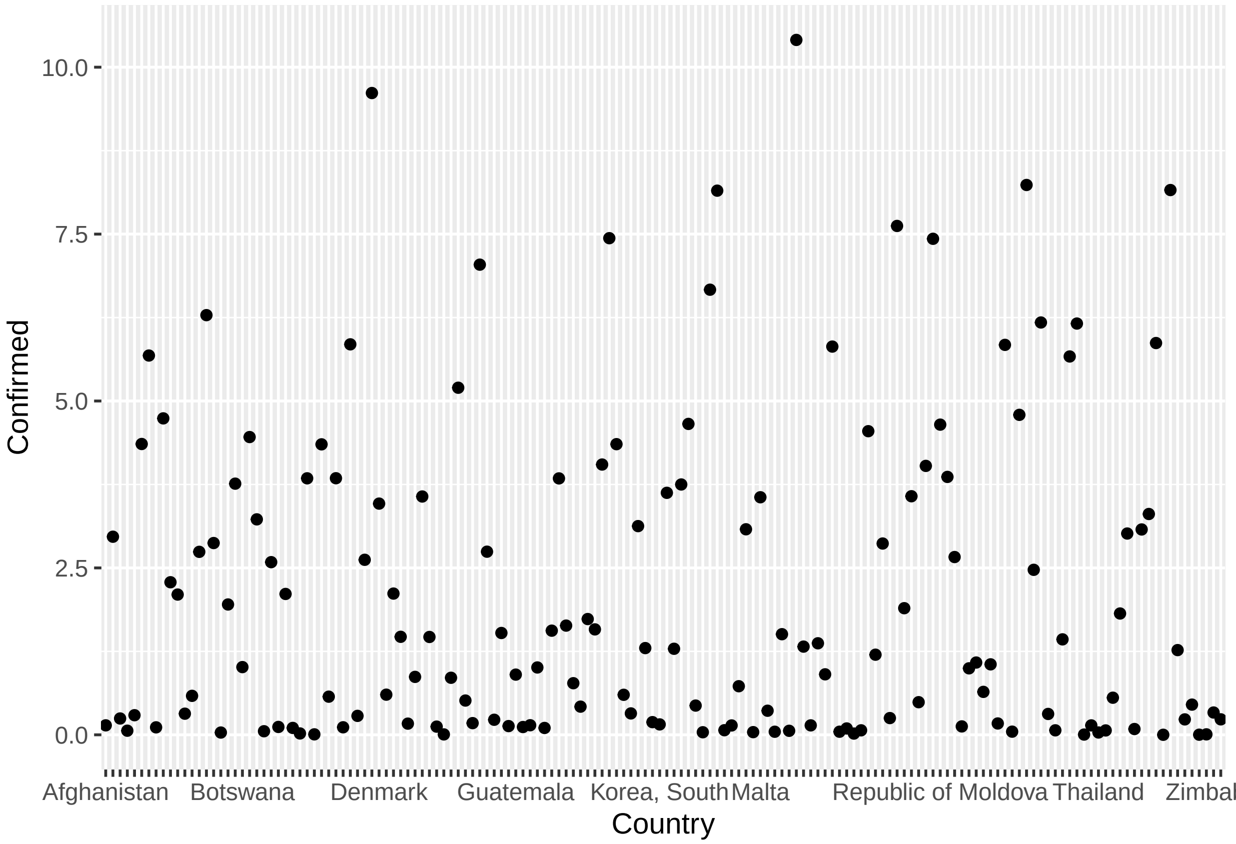
From the above correlation plot, we can see some interesting correlations between some food groups, such as between Vetegal Procucts and Animal Fats. We also see that Obesity and Undernourished are strongly negatively correlated, which makes sense, and that there is a very high correlation between Confirmed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

```r
library(ggplot2)
# Percentage of confirmed cases by country - I want to only post the labels of countries that have over
ggplot(data = backAICdata.plus, aes(x=Country, y=Confirmed, label= Country)) + geom_point() + scale_x_d
```



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.

```r
# 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 Be

#summary(model_all)

n <- length(data$Deaths)

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

```
## Start:  AIC=-994.59
## 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
##
##                                  Df Sum of Sq     RSS     AIC
## - data$Population                 1 0.0000067 0.20034 -996.59
## - data$Offals                     1 0.0001184 0.20045 -996.50
## - data$`Fish, Seafood`            1 0.0001461 0.20048 -996.48
## - data$Meat                       1 0.0001525 0.20048 -996.47
## - data$`Milk - Excluding Butter`  1 0.0001541 0.20048 -996.47
## - data$Eggs                       1 0.0001595 0.20049 -996.47
## - data$`Animal fats`              1 0.0001596 0.20049 -996.47
## - data$Undernourished             1 0.0015238 0.20185 -995.41
## - data$`Vegetal Products`         1 0.0020488 0.20238 -995.01
## - data$`Animal Products`          1 0.0023360 0.20267 -994.78
## <none>                                        0.20033 -994.59
## - data$Obesity                    1 0.0053382 0.20567 -992.49
## - data$Vegetables                 1 0.0063722 0.20670 -991.71
## - data$Pulses                     1 0.0065954 0.20692 -991.54
## - data$Spices                     1 0.0078568 0.20819 -990.59
## - data$`Fruits - Excluding Wine`  1 0.0081124 0.20844 -990.40
## - data$`Cereals - Excluding Beer` 1 0.0081288 0.20846 -990.39
## - data$Oilcrops                   1 0.0082301 0.20856 -990.31
## - data$`Vegetable Oils`           1 0.0082944 0.20862 -990.26
## - data$Treenuts                   1 0.0083540 0.20868 -990.22
## - data$Stimulants                 1 0.0089213 0.20925 -989.80
## - data$`Starchy Roots`            1 0.0097331 0.21006 -989.19
##
## Step:  AIC=-996.59
## data$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
##
##                                  Df Sum of Sq     RSS     AIC
## - data$Offals                     1 0.0001196 0.20046 -998.49
## - data$`Fish, Seafood`            1 0.0001480 0.20048 -998.47
## - data$Meat                       1 0.0001544 0.20049 -998.47
## - data$`Milk - Excluding Butter`  1 0.0001560 0.20049 -998.47
## - data$Eggs                       1 0.0001612 0.20050 -998.46
## - data$`Animal fats`              1 0.0001615 0.20050 -998.46
## - data$Undernourished             1 0.0015172 0.20185 -997.41
## - data$`Vegetal Products`         1 0.0020499 0.20239 -997.00
## - data$`Animal Products`          1 0.0023412 0.20268 -996.77
```

```
## <none>                                     0.20034 -996.59
## - data$Obesity                       1 0.0058070 0.20614 -994.13
## - data$Vegetables                    1 0.0064055 0.20674 -993.68
## - data$Pulses                        1 0.0066129 0.20695 -993.52
## - data$Spices                        1 0.0079276 0.20826 -992.53
## - data$`Fruits - Excluding Wine`     1 0.0081913 0.20853 -992.34
## - data$`Cereals - Excluding Beer`    1 0.0082018 0.20854 -992.33
## - data$Oilcrops                      1 0.0083006 0.20864 -992.25
## - data$`Vegetable Oils`              1 0.0083641 0.20870 -992.21
## - data$Treenuts                      1 0.0084189 0.20876 -992.17
## - data$Stimulants                    1 0.0090024 0.20934 -991.73
## - data$`Starchy Roots`               1 0.0098493 0.21019 -991.10
##
## Step:  AIC=-998.49
## data$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$Oilcrops +
##     data$Pulses + data$Spices + data$`Starchy Roots` + data$Stimulants +
##     data$Treenuts + data$`Vegetal Products` + data$`Vegetable Oils` +
##     data$Vegetables + data$Obesity + data$Undernourished
##
##                                     Df Sum of Sq     RSS      AIC
## - data$`Fish, Seafood`               1 0.0004077 0.20086 -1000.18
## - data$Meat                          1 0.0005939 0.20105 -1000.03
## - data$`Milk - Excluding Butter`     1 0.0006507 0.20111  -999.99
## - data$Eggs                          1 0.0008307 0.20129  -999.85
## - data$`Animal fats`                 1 0.0008604 0.20132  -999.83
## - data$Undernourished                1 0.0014433 0.20190  -999.38
## - data$`Animal Products`             1 0.0022260 0.20268  -998.77
## <none>                                         0.20046  -998.49
## - data$`Vegetal Products`            1 0.0027938 0.20325  -998.34
## - data$Obesity                       1 0.0060271 0.20648  -995.87
## - data$Vegetables                    1 0.0063792 0.20684  -995.61
## - data$Pulses                        1 0.0066106 0.20707  -995.43
## - data$Spices                        1 0.0079825 0.20844  -994.40
## - data$`Fruits - Excluding Wine`     1 0.0082223 0.20868  -994.22
## - data$`Cereals - Excluding Beer`    1 0.0082521 0.20871  -994.20
## - data$Oilcrops                      1 0.0083397 0.20880  -994.14
## - data$`Vegetable Oils`              1 0.0084021 0.20886  -994.09
## - data$Treenuts                      1 0.0084770 0.20893  -994.03
## - data$Stimulants                    1 0.0090332 0.20949  -993.62
## - data$`Starchy Roots`               1 0.0099274 0.21038  -992.95
##
## Step:  AIC=-1000.18
## data$Deaths ~ data$`Animal Products` + data$`Animal fats` + data$`Cereals - Excluding Beer` +
##     data$Eggs + data$`Fruits - Excluding Wine` + data$Meat +
##     data$`Milk - Excluding Butter` + data$Oilcrops + data$Pulses +
##     data$Spices + data$`Starchy Roots` + data$Stimulants + data$Treenuts +
##     data$`Vegetal Products` + data$`Vegetable Oils` + data$Vegetables +
##     data$Obesity + data$Undernourished
##
##                                     Df Sum of Sq     RSS      AIC
## - data$Meat                          1 0.0006941 0.20156 -1001.64
## - data$Eggs                          1 0.0009675 0.20183 -1001.43
```

```
## - data$`Milk - Excluding Butter`   1 0.0010820 0.20195 -1001.34
## - data$Undernourished             1 0.0017490 0.20261 -1000.82
## - data$`Animal Products`          1 0.0023369 0.20320 -1000.37
## <none>                                        0.20086 -1000.18
## - data$`Animal fats`              1 0.0028628 0.20373  -999.97
## - data$`Vegetal Products`         1 0.0029940 0.20386  -999.87
## - data$Vegetables                 1 0.0060011 0.20686  -997.58
## - data$Obesity                    1 0.0062896 0.20715  -997.37
## - data$Pulses                     1 0.0062992 0.20716  -997.36
## - data$Spices                     1 0.0076276 0.20849  -996.36
## - data$`Fruits - Excluding Wine`   1 0.0078459 0.20871  -996.20
## - data$`Cereals - Excluding Beer`  1 0.0078729 0.20874  -996.18
## - data$Oilcrops                   1 0.0079641 0.20883  -996.11
## - data$`Vegetable Oils`           1 0.0080274 0.20889  -996.06
## - data$Treenuts                   1 0.0081026 0.20897  -996.01
## - data$Stimulants                 1 0.0086587 0.20952  -995.59
## - data$`Starchy Roots`            1 0.0095243 0.21039  -994.95
##
## Step:  AIC=-1001.64
## data$Deaths ~ data$`Animal Products` + data$`Animal fats` + data$`Cereals - Excluding Beer` +
##     data$Eggs + data$`Fruits - Excluding Wine` + data$`Milk - Excluding Butter` +
##     data$Oilcrops + data$Pulses + data$Spices + data$`Starchy Roots` +
##     data$Stimulants + data$Treenuts + data$`Vegetal Products` +
##     data$`Vegetable Oils` + data$Vegetables + data$Obesity +
##     data$Undernourished
##
##                                  Df Sum of Sq     RSS      AIC
## - data$Eggs                       1 0.0002775 0.20184 -1003.42
## - data$`Milk - Excluding Butter`   1 0.0009096 0.20247 -1002.94
## - data$Undernourished             1 0.0015988 0.20316 -1002.41
## - data$`Animal Products`          1 0.0021904 0.20375 -1001.95
## <none>                                        0.20156 -1001.64
## - data$`Vegetal Products`         1 0.0029423 0.20450 -1001.38
## - data$Obesity                    1 0.0088600 0.21042  -996.93
## - data$Vegetables                 1 0.0113012 0.21286  -995.13
## - data$Pulses                     1 0.0124912 0.21405  -994.26
## - data$Spices                     1 0.0131713 0.21473  -993.76
## - data$`Animal fats`              1 0.0137087 0.21527  -993.37
## - data$`Fruits - Excluding Wine`   1 0.0138384 0.21540  -993.28
## - data$Treenuts                   1 0.0139802 0.21554  -993.18
## - data$`Cereals - Excluding Beer`  1 0.0141361 0.21569  -993.06
## - data$Oilcrops                   1 0.0141706 0.21573  -993.04
## - data$`Vegetable Oils`           1 0.0143908 0.21595  -992.88
## - data$Stimulants                 1 0.0150986 0.21666  -992.37
## - data$`Starchy Roots`            1 0.0155012 0.21706  -992.08
##
## Step:  AIC=-1003.42
## data$Deaths ~ data$`Animal Products` + data$`Animal fats` + data$`Cereals - Excluding Beer` +
##     data$`Fruits - Excluding Wine` + data$`Milk - Excluding Butter` +
##     data$Oilcrops + data$Pulses + data$Spices + data$`Starchy Roots` +
##     data$Stimulants + data$Treenuts + data$`Vegetal Products` +
##     data$`Vegetable Oils` + data$Vegetables + data$Obesity +
##     data$Undernourished
##
```

```
##                                   Df Sum of Sq     RSS      AIC
## - data$`Milk - Excluding Butter`   1 0.0007578 0.20259 -1004.84
## - data$Undernourished              1 0.0019305 0.20377 -1003.94
## - data$`Animal Products`           1 0.0023932 0.20423 -1003.59
## <none>                                         0.20184 -1003.42
## - data$`Vegetal Products`          1 0.0031963 0.20503 -1002.97
## - data$Obesity                     1 0.0088040 0.21064  -998.76
## - data$Vegetables                  1 0.0117353 0.21357  -996.61
## - data$Pulses                      1 0.0127129 0.21455  -995.90
## - data$`Animal fats`               1 0.0134312 0.21527  -995.37
## - data$Spices                      1 0.0136367 0.21547  -995.23
## - data$`Fruits - Excluding Wine`   1 0.0143418 0.21618  -994.72
## - data$Treenuts                    1 0.0144823 0.21632  -994.61
## - data$`Cereals - Excluding Beer`  1 0.0145691 0.21640  -994.55
## - data$Oilcrops                    1 0.0145859 0.21642  -994.54
## - data$`Vegetable Oils`            1 0.0148276 0.21666  -994.37
## - data$Stimulants                  1 0.0155281 0.21736  -993.86
## - data$`Starchy Roots`             1 0.0156675 0.21750  -993.76
##
## Step:  AIC=-1004.84
## data$Deaths ~ data$`Animal Products` + 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
##
##                                   Df Sum of Sq     RSS      AIC
## - data$Undernourished              1 0.0018897 0.20448 -1005.39
## - data$`Animal Products`           1 0.0022214 0.20481 -1005.14
## <none>                                         0.20259 -1004.84
## - data$`Vegetal Products`          1 0.0030295 0.20562 -1004.52
## - data$Obesity                     1 0.0088838 0.21148 -1000.15
## - data$`Animal fats`               1 0.0127277 0.21532  -997.34
## - data$Vegetables                  1 0.0138261 0.21642  -996.54
## - data$Pulses                      1 0.0149749 0.21757  -995.72
## - data$Spices                      1 0.0152391 0.21783  -995.53
## - data$`Fruits - Excluding Wine`   1 0.0160741 0.21867  -994.93
## - data$Treenuts                    1 0.0161826 0.21878  -994.85
## - data$Oilcrops                    1 0.0163106 0.21890  -994.76
## - data$`Cereals - Excluding Beer`  1 0.0163158 0.21891  -994.76
## - data$`Vegetable Oils`            1 0.0165892 0.21918  -994.56
## - data$`Starchy Roots`             1 0.0166486 0.21924  -994.52
## - data$Stimulants                  1 0.0176333 0.22023  -993.82
##
## Step:  AIC=-1005.39
## data$Deaths ~ data$`Animal Products` + data$`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
##
##                                   Df Sum of Sq     RSS      AIC
## - data$`Animal Products`           1 0.0024907 0.20697 -1005.50
## <none>                                         0.20448 -1005.39
```

```
## - data$`Vegetal Products`          1 0.0033666 0.20785 -1004.84
## - data$`Animal fats`               1 0.0133715 0.21785  -997.51
## - data$Obesity                     1 0.0139897 0.21847  -997.07
## - data$Pulses                      1 0.0154100 0.21989  -996.06
## - data$Vegetables                  1 0.0158057 0.22029  -995.78
## - data$Spices                      1 0.0165120 0.22099  -995.28
## - data$`Fruits - Excluding Wine`   1 0.0168259 0.22131  -995.06
## - data$`Starchy Roots`             1 0.0168303 0.22131  -995.05
## - data$Treenuts                    1 0.0169964 0.22148  -994.94
## - data$`Cereals - Excluding Beer`  1 0.0170345 0.22152  -994.91
## - data$Oilcrops                    1 0.0171182 0.22160  -994.85
## - data$`Vegetable Oils`            1 0.0173780 0.22186  -994.67
## - data$Stimulants                  1 0.0184431 0.22293  -993.92
##
## Step:  AIC=-1005.5
## data$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$Obesity
##
##                                   Df Sum of Sq     RSS      AIC
## <none>                                         0.20697 -1005.50
## - data$`Animal fats`               1  0.013220 0.22019  -997.84
## - data$Obesity                     1  0.014472 0.22145  -996.96
## - data$Vegetables                  1  0.015969 0.22294  -995.91
## - data$Pulses                      1  0.016152 0.22313  -995.78
## - data$Spices                      1  0.017684 0.22466  -994.71
## - data$Treenuts                    1  0.017716 0.22469  -994.69
## - data$`Cereals - Excluding Beer`  1  0.017785 0.22476  -994.64
## - data$`Starchy Roots`             1  0.017835 0.22481  -994.61
## - data$Oilcrops                    1  0.017873 0.22485  -994.58
## - data$`Fruits - Excluding Wine`   1  0.017895 0.22487  -994.57
## - data$`Vegetal Products`          1  0.018057 0.22503  -994.45
## - data$`Vegetable Oils`            1  0.018120 0.22509  -994.41
## - data$Stimulants                  1  0.019152 0.22613  -993.70
```

```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
```

```
# backBIC <- step(model_all ,direction="backward", data=data, k = log(n))
```

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.

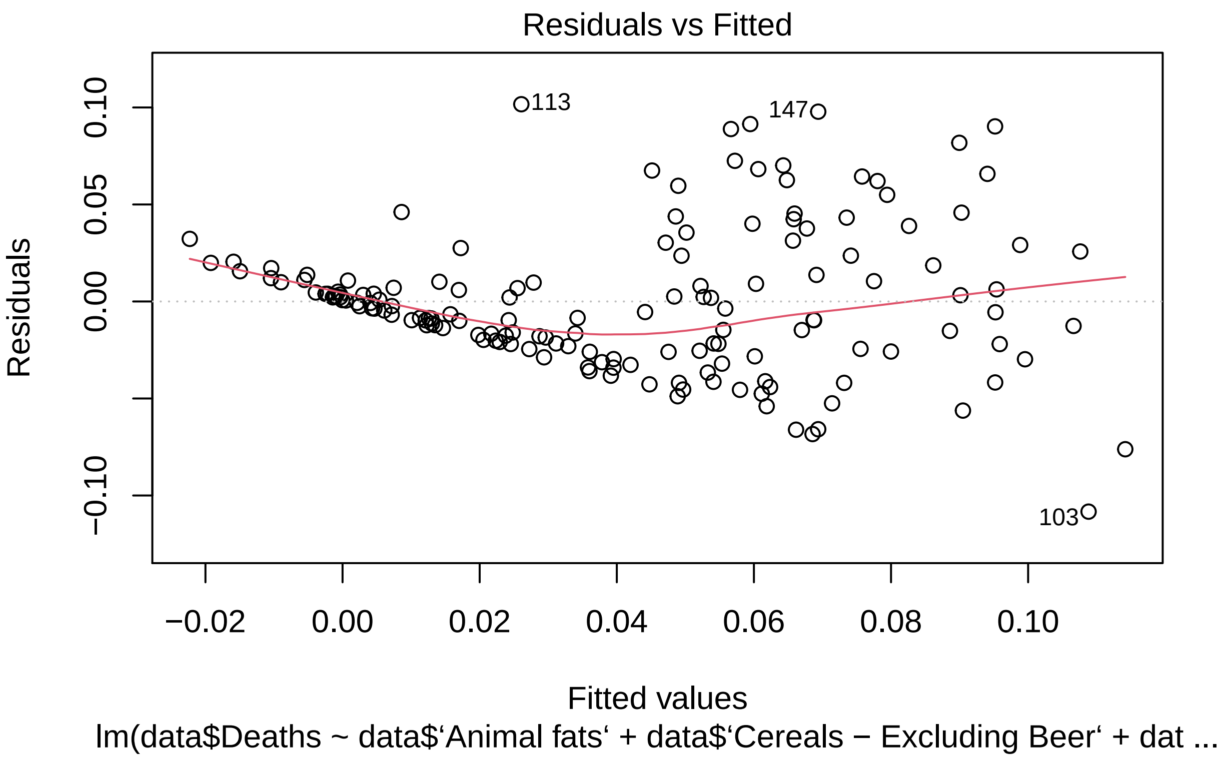
Part 3

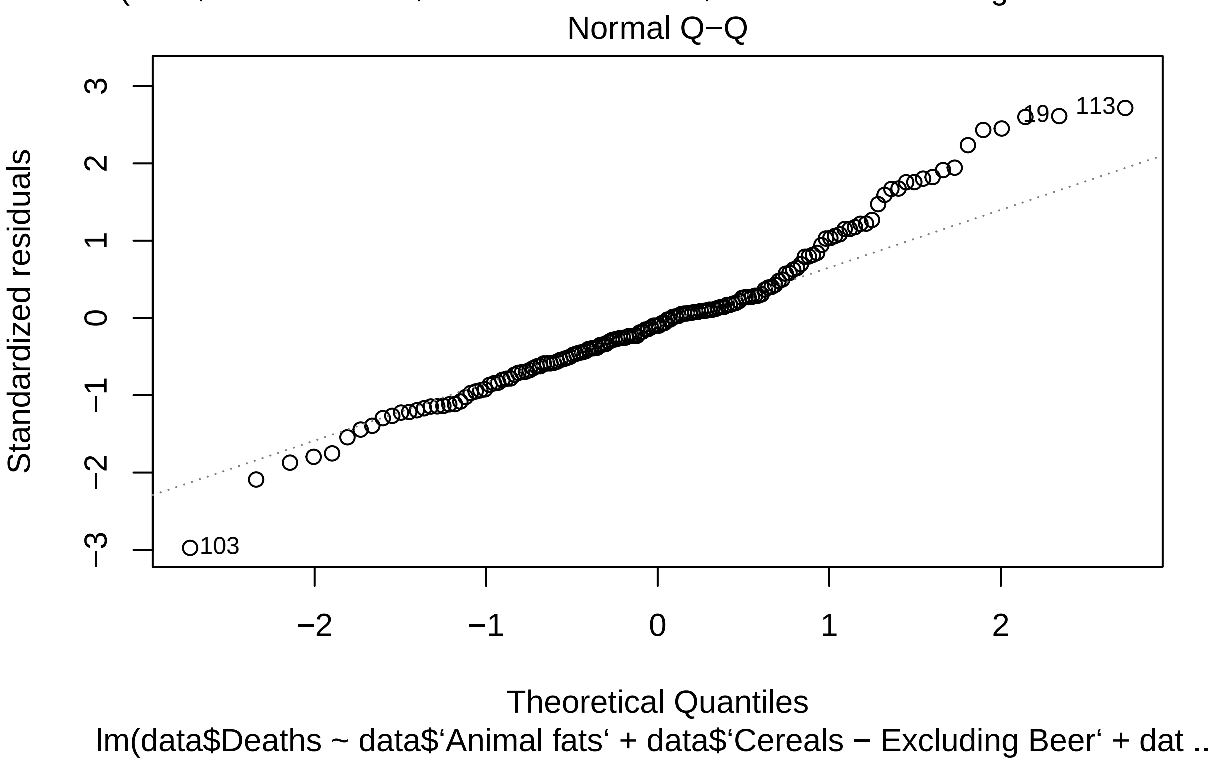
Identify if there are any outliers, high leverage, and or influential observations worth removing. If so, remove them but justify your reason for doing so and re-estimate your model.

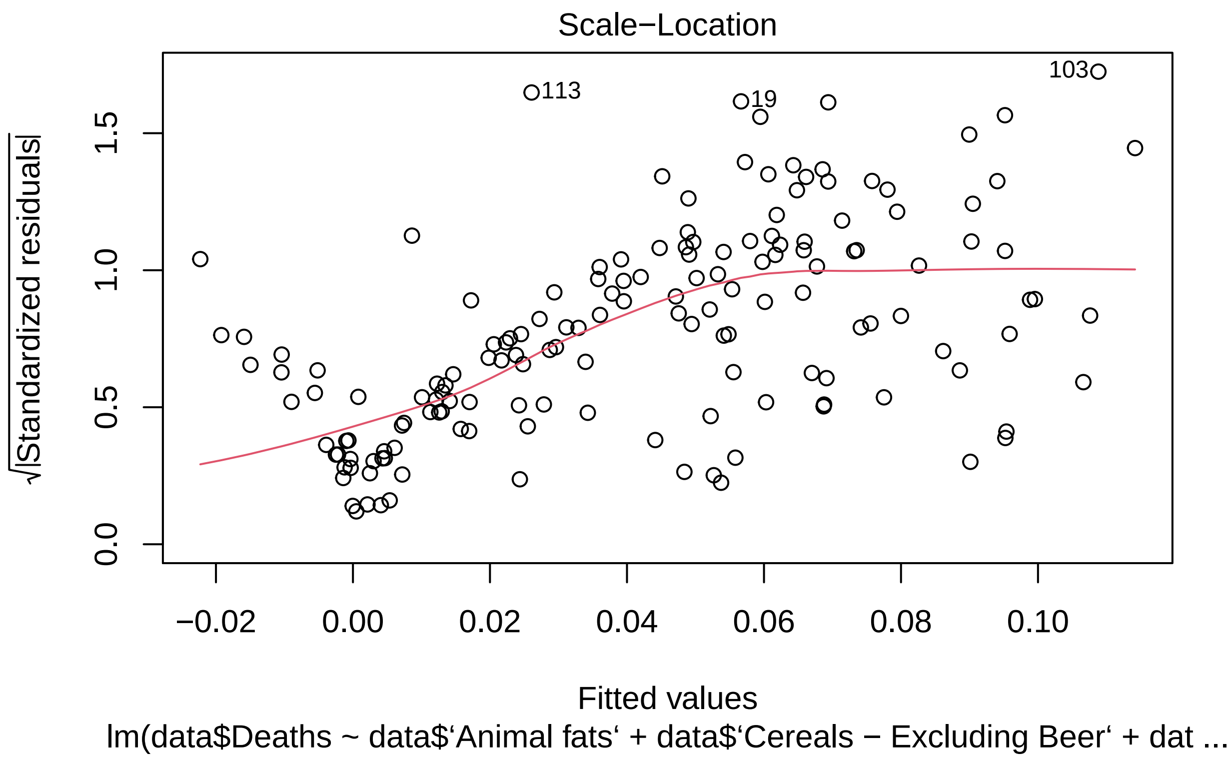
```
library(olsrr)
```

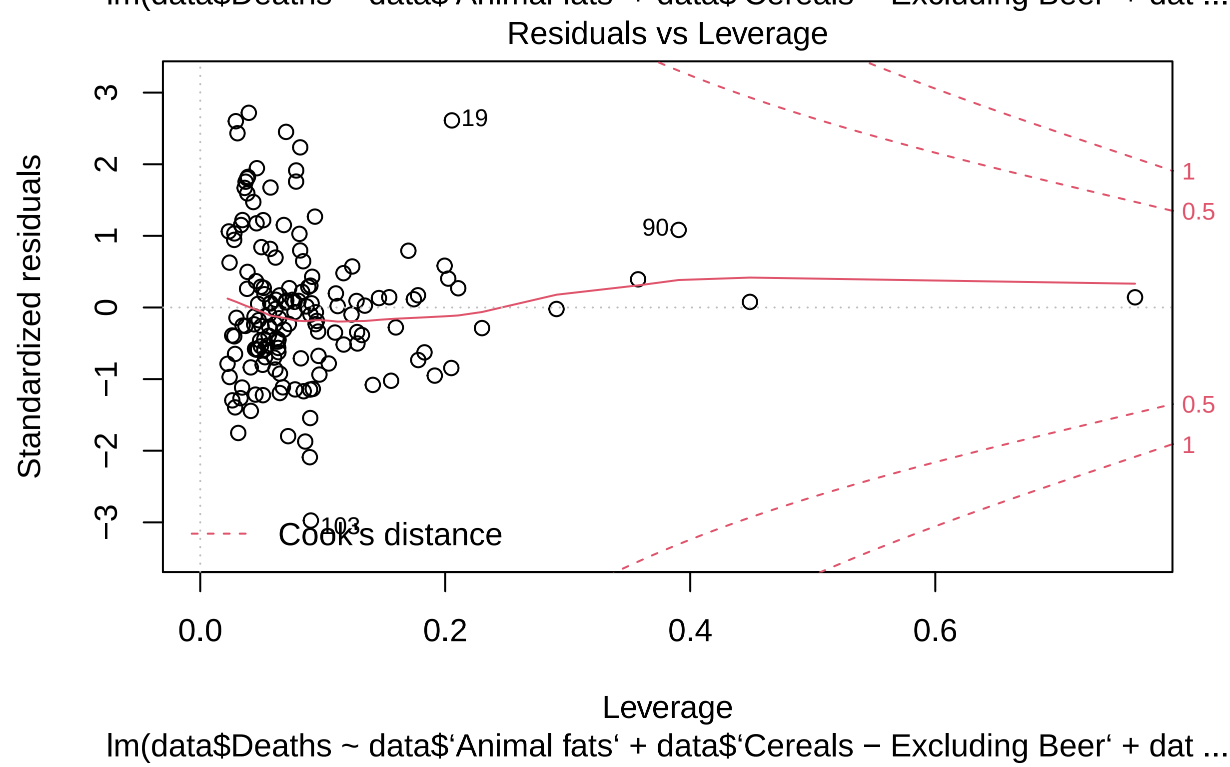
```
##
## Attaching package: 'olsrr'

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

```
plot(backAIC)
```

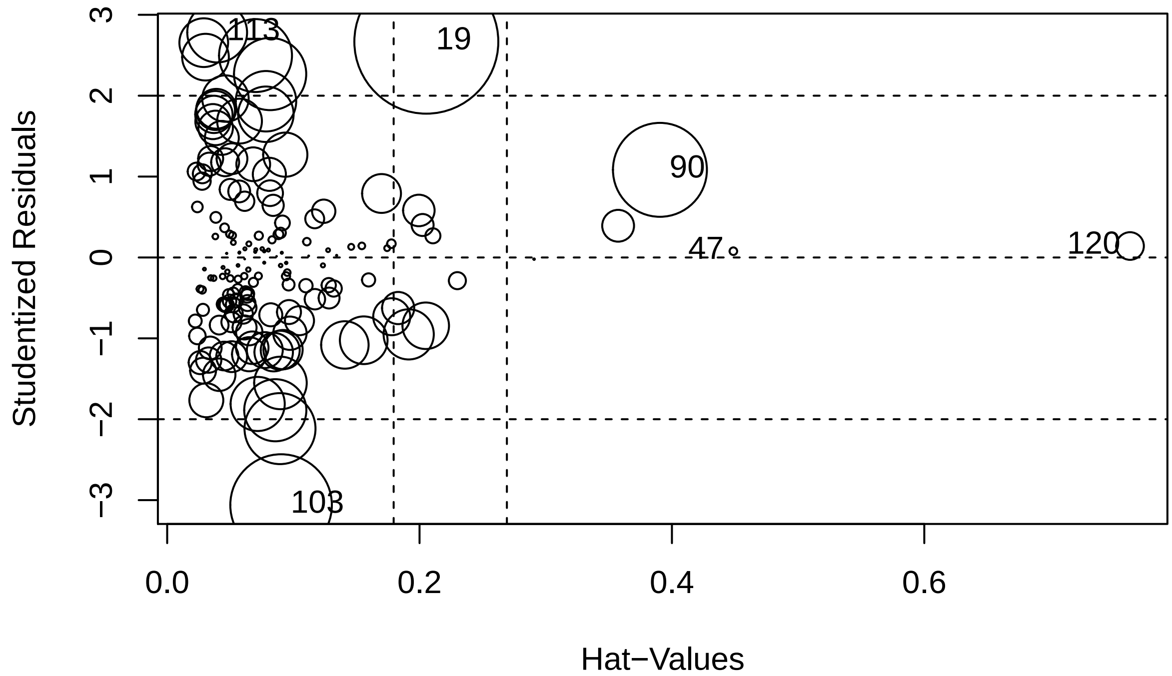

Residuals vs Fitted
Residuals
0.10
0.05
0.00
−0.10
113
147
103
−0.02
0.00
0.02
0.04
0.06
0.08
0.10
Fitted values
lm(data$Deaths ~ data$‘Animal fats‘ + data$‘Cereals − Excluding Beer‘ + dat ...


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
19
113
103
−2
−1
0
1
2
Theoretical Quantiles
lm(data$Deaths ~ data$‘Animal fats‘ + data$‘Cereals − Excluding Beer‘ + dat ...



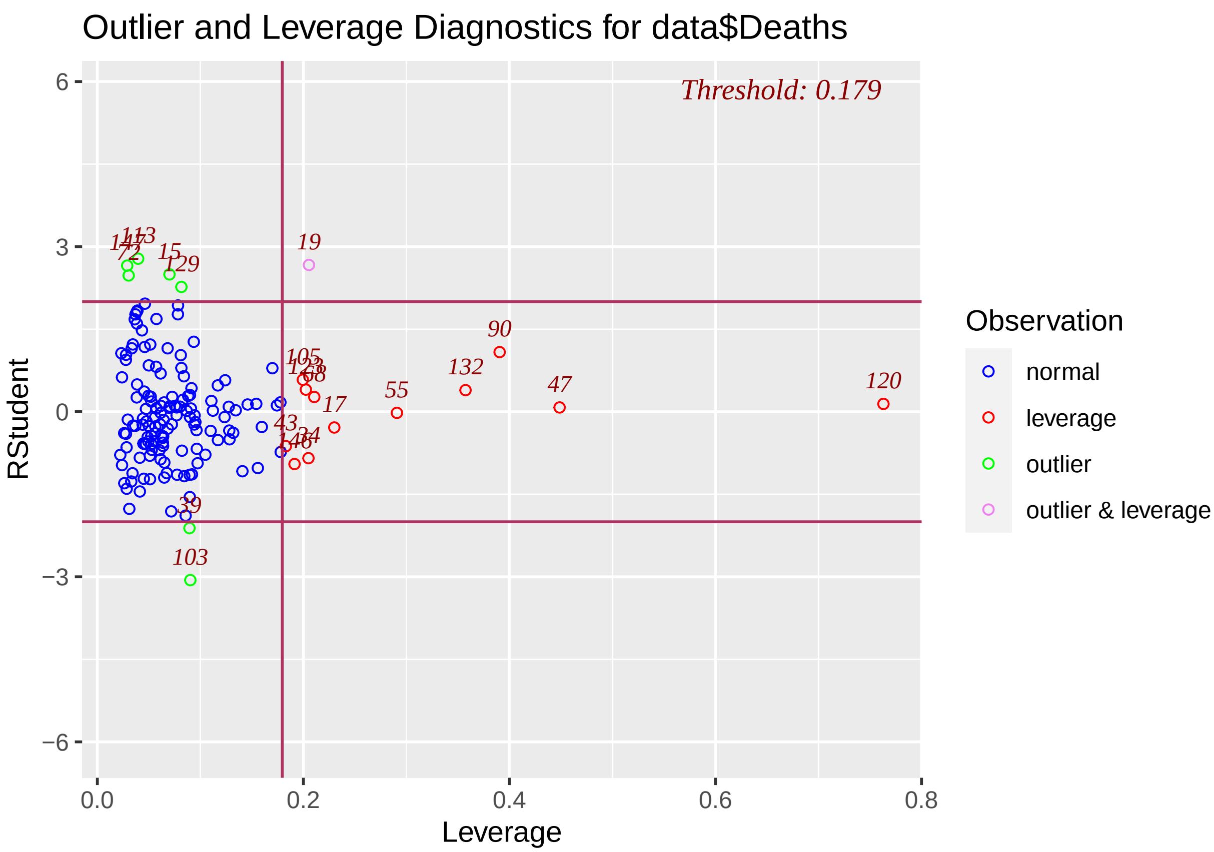



```
influencePlot(backAIC, id=list(n=3))
```




```
##         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)

# 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+`
mod1 <- lm(Deaths~Animal_Fats+Cereals+Fruits+Oilcrops+Pulses+Spices+Starchy_Roots+Stimulants+Treenuts+`
mod2 <- lm(Deaths~Animal_Fats+Cereals+Fruits+Oilcrops+Pulses+Spices+Starchy_Roots+Stimulants+Treenuts+`
mod3 <- lm(Deaths~Animal_Fats+Cereals+Fruits+Oilcrops+Pulses+Spices+Starchy_Roots+Stimulants+Treenuts+`
mod4 <- lm(Deaths~Animal_Fats+Cereals+Fruits+Oilcrops+Pulses+Spices+Starchy_Roots+Stimulants+Treenuts+`

stargazer(mod0, mod1, mod2, mod3, mod4, object.names = TRUE, title = "Regression Model Results", column
```

```
# New model
new_model_1 <- lm(Deaths~Animal_Fats+Cereals+Fruits+Oilcrops+Pulses+Spices+Starchy_Roots+Stimulants+Tre
summary(new_model_1)
```

```
##
## Call:
## lm(formula = Deaths ~ Animal_Fats + Cereals + Fruits + Oilcrops +
##     Pulses + Spices + Starchy_Roots + Stimulants + Treenuts +
```

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

```
##     `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.

```r
# 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
```

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

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

```r
# install.packages("Boruta")
library(Boruta)

Bor.res <- Boruta(Deaths~Animal_Fats+Cereals+Fruits+Oilcrops+Pulses+Spices+Starchy_Roots+Stimulants+Tre
# plot(Bor.res, xlab = "", 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))
#plot(Bor.res, xlab = "Attributes", main = "Boruta Algorithim")
# Fix labels
```

```r
# Testing to see which variables we want to remove
attStats(Bor.res)
```

```
##                        meanImp    medianImp      minImp    maxImp  normHits
## Animal_Fats        12.47296022 12.450433496 11.01201484 14.217951 1.0000000
## Cereals             8.25234307  8.321486801  5.84639011 10.463144 1.0000000
## Fruits              0.40674736 -0.006092799 -0.97208262  2.887568 0.0000000
## Oilcrops            9.69553090  9.694820296  7.28161167 11.805414 1.0000000
## Pulses              4.86158916  4.887645879  2.68668147  6.521805 0.9318182
## Spices              0.58189585  0.479447547 -1.26675825  2.749577 0.0000000
## Starchy_Roots       1.49404164  1.447256861 -0.82745698  4.236812 0.2500000
## Stimulants          1.91996161  1.867334726 -0.08305459  3.950449 0.2500000
## Treenuts            0.95802791  0.911825789 -0.27980842  2.901689 0.0000000
## `Vegetal Products` 10.71358597 10.606772305  8.54709820 12.892801 1.0000000
## Vegetable_Oils      3.81064914  3.745050779  1.74034256  6.818625 0.7727273
## Vegetables          0.01572744  0.300061321 -1.95030788  1.545737 0.0000000
## Obesity            17.22173698 17.465020891 15.12959794 20.274515 1.0000000
##                     decision
## Animal_Fats        Confirmed
## Cereals            Confirmed
## Fruits              Rejected
## Oilcrops           Confirmed
## Pulses             Confirmed
## Spices              Rejected
## Starchy_Roots       Rejected
## Stimulants          Rejected
## Treenuts            Rejected
## `Vegetal Products` Confirmed
## Vegetable_Oils     Confirmed
## Vegetables          Rejected
## Obesity            Confirmed
```

```r
sorted_vars <- attStats(Bor.res)[order(-attStats(Bor.res)$meanImp),]
print(sorted_vars)
```

```
##                        meanImp    medianImp      minImp    maxImp  normHits
## Obesity            17.22173698 17.465020891 15.12959794 20.274515 1.0000000
## Animal_Fats        12.47296022 12.450433496 11.01201484 14.217951 1.0000000
## `Vegetal Products` 10.71358597 10.606772305  8.54709820 12.892801 1.0000000
## Oilcrops            9.69553090  9.694820296  7.28161167 11.805414 1.0000000
## Cereals             8.25234307  8.321486801  5.84639011 10.463144 1.0000000
## Pulses              4.86158916  4.887645879  2.68668147  6.521805 0.9318182
## Vegetable_Oils      3.81064914  3.745050779  1.74034256  6.818625 0.7727273
## Stimulants          1.91996161  1.867334726 -0.08305459  3.950449 0.2500000
## Starchy_Roots       1.49404164  1.447256861 -0.82745698  4.236812 0.2500000
## Treenuts            0.95802791  0.911825789 -0.27980842  2.901689 0.0000000
## Spices              0.58189585  0.479447547 -1.26675825  2.749577 0.0000000
## Fruits              0.40674736 -0.006092799 -0.97208262  2.887568 0.0000000
## Vegetables          0.01572744  0.300061321 -1.95030788  1.545737 0.0000000
##                     decision
## Obesity            Confirmed
## Animal_Fats        Confirmed
## `Vegetal Products` Confirmed
## Oilcrops           Confirmed
```

```
## Cereals             Confirmed
## Pulses              Confirmed
## Vegetable_Oils      Confirmed
## Stimulants           Rejected
## Starchy_Roots        Rejected
## Treenuts             Rejected
## Spices               Rejected
## Fruits               Rejected
## Vegetables           Rejected
```

```r
# 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,
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.

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

```
##              data$`Animal fats` data$`Cereals - Excluding Beer`
##                        2.194866                     2558.260164
##   data$`Fruits - Excluding Wine`                   data$Oilcrops
##                      192.491187                     3510.358228
##                     data$Pulses                     data$Spices
##                       35.561658                       57.333764
##            data$`Starchy Roots`                 data$Stimulants
```

```
##                     29.015962                          131.294727
##                 data$Treenuts            data$`Vegetal Products`
##                    176.857760                        16523.740166
##         data$`Vegetable Oils`                    data$Vegetables
##                  11059.282460                          10.146877
##                  data$Obesity
##                      1.790069
```

```r
model_vif <- lm(data$Deaths ~  data$`Cereals - Excluding Beer` + data$Eggs + data$`Fish, Seafood` + data

vif(model_vif)
```

```
## data$`Cereals - Excluding Beer`                          data$Eggs
##                        1.800261                           1.844473
##             data$`Fish, Seafood`  data$`Fruits - Excluding Wine`
##                        1.780198                           1.823960
##                       data$Meat  data$`Milk - Excluding Butter`
##                        1.599268                           1.712213
##                     data$Offals                      data$Oilcrops
##                        1.570224                           1.588263
##                     data$Pulses                        data$Spices
##                        2.705881                           1.421193
##           data$`Starchy Roots`                    data$Stimulants
##                        2.114050                           1.497971
##                   data$Treenuts                    data$Vegetables
##                        1.379212                           1.309479
##                    data$Obesity                data$Undernourished
##                        2.470468                           2.300126
##                 data$Population
##                        1.199551
```

```r
vif(new_model_2) # We will remove any variable with a VIF over 10 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
```

```r
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
```

```r
# Remove: Cereals, OilCrops, Vegetal Products, and Vegtable Oils
new_model_3 <- lm(Deaths~Animal_Fats+Pulses+Obesity, data=no_unusual_observations)

#New MOdel
summary(new_model_3)
```

```
##
## Call:
## lm(formula = Deaths ~ Animal_Fats + Pulses + Obesity, data = no_unusual_observations)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
```

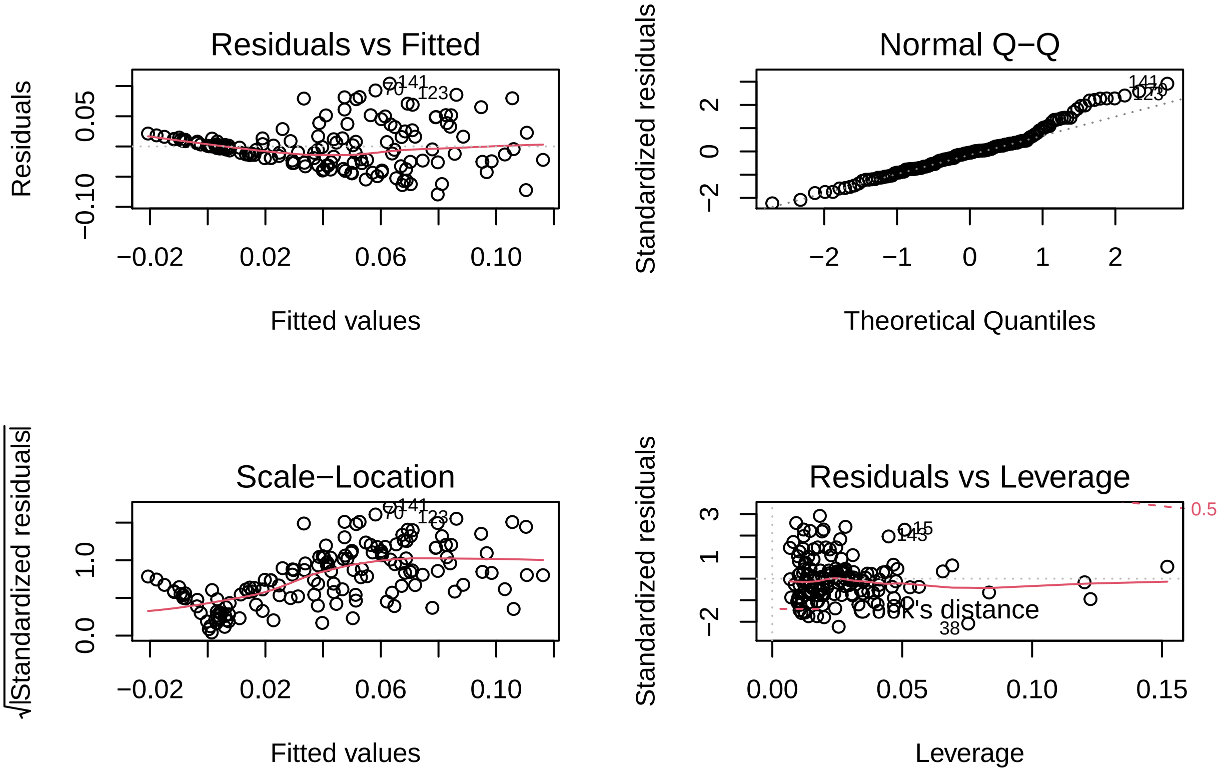
```
## -0.079491 -0.025211 -0.001937  0.014407  0.104142
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0118222  0.0092708  -1.275   0.2043
## Animal_Fats  0.0060356  0.0009975   6.051 1.16e-08 ***
## Pulses      -0.0208336  0.0124383  -1.675   0.0961 .
## Obesity      0.0017121  0.0003790   4.518 1.28e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03607 on 146 degrees of freedom
## Multiple R-squared:  0.4568, Adjusted R-squared:  0.4457
## F-statistic: 40.93 on 3 and 146 DF,  p-value: < 2.2e-16
```

Part 6

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

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.

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



Part 7

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

Here we tested our model by testing our model against a quadratic. Our result is a p-value of 0.7077 which means we should consider higher order powers.

```r
library(lmtest)
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

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

```
##
##  RESET test
##
## data:  new_model_3
## RESET = 0.2421, df1 = 3, df2 = 143, p-value = 0.8668
```

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.

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

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 34.27766, Df = 1, p = 4.7784e-09
```

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

```
##
##  studentized Breusch-Pagan test
##
## data:  new_model_3
## BP = 31.482, df = 3, p-value = 6.73e-07
```

From the above tests it can be seen that heteroskedacity is present in our data. 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.

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

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic     p.value
##   <chr>          <dbl>     <dbl>     <dbl>       <dbl>
## 1 (Intercept) -0.0118   0.00568      -2.08 0.0391
## 2 Animal_Fats  0.00604  0.00111       5.44 0.000000217
## 3 Pulses      -0.0208   0.00791      -2.64 0.00931
## 4 Obesity      0.00171  0.000357      4.80 0.00000390
```

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 may be an existence of higher power terms, which will be tested here. 

higher_power <- lm(Deaths~Animal_Fats+Pulses+Obesity+I(Animal_Fats^2)+I(Pulses^2)+I(Obesity^2), data=no
summary(higher_power) #None of the higher powers are statistically significant
```

```
##
## Call:
## lm(formula = Deaths ~ Animal_Fats + Pulses + Obesity + I(Animal_Fats^2) +
##     I(Pulses^2) + I(Obesity^2), data = no_unusual_observations)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.075159 -0.024384 -0.001036  0.013998  0.104419
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -1.406e-02  1.514e-02  -0.929   0.3545
## Animal_Fats       5.430e-03  3.083e-03   1.761   0.0803 .
## Pulses           -4.049e-02  3.197e-02  -1.266   0.2074
## Obesity           2.541e-03  1.624e-03   1.565   0.1198
## I(Animal_Fats^2)  2.641e-05  2.354e-04   0.112   0.9108
## I(Pulses^2)       2.113e-02  2.963e-02   0.713   0.4770
## I(Obesity^2)     -2.210e-05  4.300e-05  -0.514   0.6081
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03635 on 143 degrees of freedom
## Multiple R-squared:  0.4596, Adjusted R-squared:  0.4369
## F-statistic: 20.27 on 6 and 143 DF,  p-value: < 2.2e-16
```

```
interaction_terms <- lm(Deaths~Animal_Fats+Pulses+Obesity+(Animal_Fats*Pulses)+(Animal_Fats*Obesity)+(Pu

summary(interaction_terms) #Pulses:Obesity is statistically significant, this will be added to a new mo
```

```
##
## Call:
## lm(formula = Deaths ~ Animal_Fats + Pulses + Obesity + (Animal_Fats *
##     Pulses) + (Animal_Fats * Obesity) + (Pulses * Obesity), data = no_unusual_observations)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.075775 -0.020958 -0.001059  0.012111  0.097262
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.0152967  0.0143334  -1.067  0.28768
## Animal_Fats          0.0022704  0.0037480   0.606  0.54564
## Pulses               0.0217707  0.0234310   0.929  0.35438
## Obesity              0.0020372  0.0006898   2.953  0.00368 **
## Animal_Fats:Pulses  -0.0015543  0.0052331  -0.297  0.76689
```

```
## Animal_Fats:Obesity  0.0001612  0.0001542   1.045  0.29783
## Pulses:Obesity      -0.0030519  0.0013112  -2.328  0.02134 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03541 on 143 degrees of freedom
## Multiple R-squared:  0.4872, Adjusted R-squared:  0.4657
## F-statistic: 22.64 on 6 and 143 DF,  p-value: < 2.2e-16
```

```
new_model_4 <- lm(Deaths~Animal_Fats+Pulses+Obesity+(Pulses*Obesity), data=no_unusual_observations)

summary(new_model_4)
```

```
##
## Call:
## lm(formula = Deaths ~ Animal_Fats + Pulses + Obesity + (Pulses *
##     Obesity), data = no_unusual_observations)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.073850 -0.023112  0.001155  0.013670  0.096866
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.0238339  0.0102148  -2.333   0.0210 *
## Animal_Fats     0.0054686  0.0010030   5.452 2.09e-07 ***
## Pulses          0.0197345  0.0198853   0.992   0.3226
## Obesity         0.0025700  0.0004985   5.156 8.14e-07 ***
## Pulses:Obesity -0.0032891  0.0012729  -2.584   0.0108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03539 on 145 degrees of freedom
## Multiple R-squared:  0.4808, Adjusted R-squared:  0.4664
## F-statistic: 33.56 on 4 and 145 DF,  p-value: < 2.2e-16
```

```
# Testing with AIC and BIC
library(broom)
library(AER)
```

```
## Loading required package: sandwich
```

```
## Loading required package: survival
```

```
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  5 -565.0577
## new_model_4  6 -569.8100
```

```
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  5 -550.0045
## new_model_4  6 -551.7462
```

```r
# 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)
library(broom)
tidy(new_model_4_adjusted)
```

```
## # A tibble: 5 x 5
##   term           estimate std.error statistic     p.value
##   <chr>             <dbl>     <dbl>     <dbl>       <dbl>
## 1 (Intercept)    -0.0238   0.00584      -4.08 0.0000739
## 2 Animal_Fats     0.00547  0.00113       4.84 0.00000327
## 3 Pulses          0.0197   0.0100        1.97 0.0511
## 4 Obesity         0.00257  0.000477      5.38 0.000000286
## 5 Pulses:Obesity -0.00329  0.000927     -3.55 0.000524
```

Above we first tested for higher powers because the RESET test suggested we should test our model with quadratic variables. We found no statistically significant powers. After testing for higher powers we tested for interaction terms. The interaction between Pulses and Obesity was statistically significant so it was added to the model, 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 BIC, leading us to believe that we had found the best model. In part 8 we learned that heteroskadacity 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.

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.

```r
# install.packages("caret")
# install.packages("lattice")
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:survival':
##
##     cluster
```

```r
model_final <- new_model_4 # replace this with the model from 9 once we have it

# split data into 2/3 train 1/3 test
train <- sample(nrow(data), nrow(data) * 2/3)
data_train <- data[train,]
data_test <- data[-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 ~  `Cereals - Excluding Beer` + Eggs + `Fish, Seafood` + `Fruits - Excluding Wi
```

```r
model_cv
```

```
## Generalized Linear Model
##
## 104 samples
##  17 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 83, 84, 83, 83, 83
## Resampling results:
##
##   RMSE        Rsquared   MAE
##   0.04827098  0.2073397  0.038133
```

```r
summary(model_cv)
```

```
##
## Call:
## NULL
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.063427  -0.024961  -0.004214   0.022294   0.129318
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      7.026e-02  3.501e-02   2.007   0.0479 *
## `\\`Cereals - Excluding Beer\\`` -1.612e-03  1.595e-03  -1.011   0.3149
## Eggs                            -3.656e-04  8.385e-03  -0.044   0.9653
## `\\`Fish, Seafood\\``           -2.304e-02  9.970e-03  -2.311   0.0232 *
## `\\`Fruits - Excluding Wine\\``  -1.647e-03  6.507e-03  -0.253   0.8007
## Meat                            -7.696e-05  1.095e-03  -0.070   0.9441
## `\\`Milk - Excluding Butter\\``   1.104e-04  1.629e-03   0.068   0.9461
## Offals                          -6.845e-02  4.513e-02  -1.517   0.1330
## Oilcrops                        -7.850e-04  1.877e-03  -0.418   0.6769
## Pulses                          -8.192e-03  2.341e-02  -0.350   0.7273
## Spices                          -1.637e-03  1.097e-02  -0.149   0.8817
## `\\`Starchy Roots\\``             9.942e-03  1.558e-02   0.638   0.5252
## Stimulants                       1.045e-02  6.431e-03   1.624   0.1080
## Treenuts                        -1.247e-03  6.209e-03  -0.201   0.8413
## Vegetables                      -1.284e-02  2.323e-02  -0.553   0.5818
## Obesity                          7.527e-04  7.813e-04   0.963   0.3380
## Undernourished                  -6.908e-04  5.342e-04  -1.293   0.1994
## Population                       9.845e-12  3.336e-11   0.295   0.7686
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.001725167)
##
##     Null deviance: 0.24265  on 103  degrees of freedom
## Residual deviance: 0.14836  on  86  degrees of freedom
## AIC: -348.32
##
```

```
## Number of Fisher Scoring iterations: 2
```

```r
# make predictions on the testing partition
pred <- predict(model_cv, data_test)

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

```
## [1] 0.04707336
```

Part 11

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

Things we may want to say: - We learned very quickly that as far as predicting deaths, there were a very limited number of variables that were statistically signifcant, we were cutting down variabels fast. We started with 32 variables and in our first model that was cut to 14. Obesity and Animal_fats appear to have the best prediction power.