Homework 5

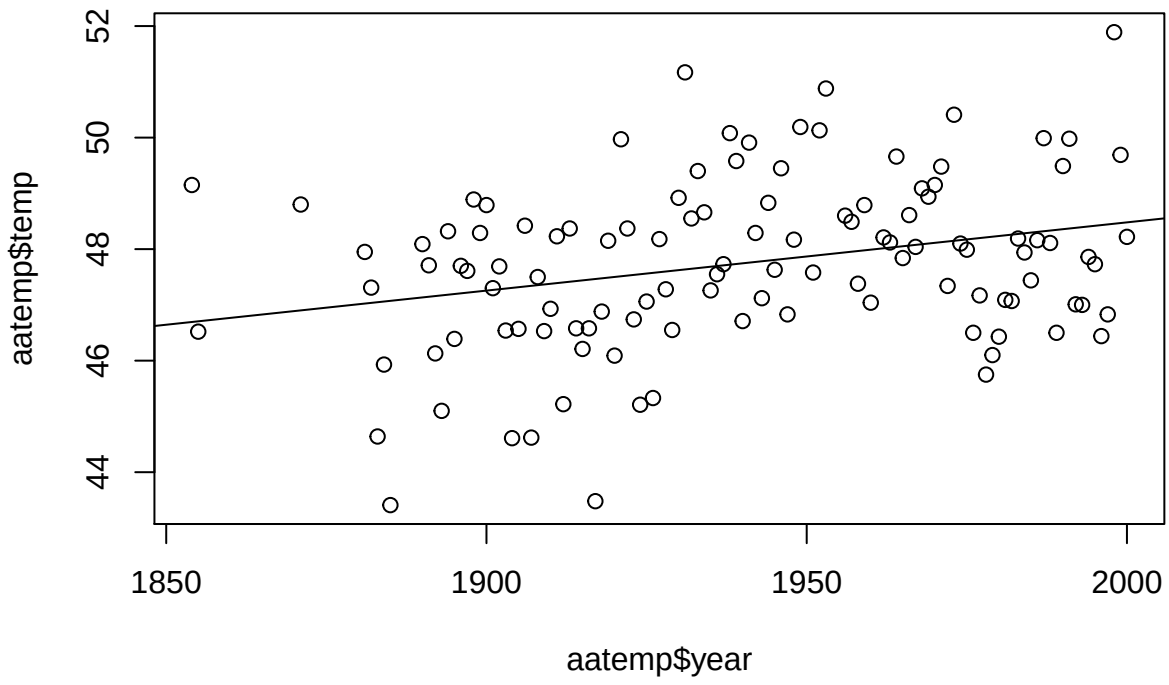
Hyunjoon Rhee

3/28/2021

Problem 1

a)

```
library(faraway)
data(aatemp)
prob1 = lm(temp ~ year, data = aatemp)
plot(aatemp$year, aatemp$temp)
abline(prob1)
```



```r
cor(aatemp$year, aatemp$temp)
```

```
## [1] 0.2921634
```

The graph shows that there is a weak linear relationship between year and the temperature. The correlation of 0.292 shows the linear trend.

b)

```r
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.0.2
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

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

```r
dwtest(prob1)
```

```
##
##  Durbin-Watson test
##
## data:  prob1
## DW = 1.6177, p-value = 0.01524
## alternative hypothesis: true autocorrelation is greater than 0
```

Because the dwtest shows that it has a p value that is smaller than 0.05, reject the null hypothesis meaning that it can be said that there is correlation in the error.

```r
library(nlme)
prob1_b = gls(temp ~ year, correlation = corAR1(form= ~year), data=aatemp)
summary(prob1_b)
```

```
## Generalized least squares fit by REML
##   Model: temp ~ year
##   Data: aatemp
##        AIC     BIC    logLik
##   426.5694 437.479 -209.2847
##
## Correlation Structure: ARMA(1,0)
##  Formula: ~year
##  Parameter estimate(s):
##      Phi1
```

```
## 0.2303887
##
## Coefficients:
##                Value Std.Error  t-value p-value
## (Intercept) 25.18407  8.971864 2.807006  0.0059
## year         0.01164  0.004626 2.516015  0.0133
##
##  Correlation:
##      (Intr)
## year -1
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.7230803 -0.6321970 -0.0520135  0.6645795  2.3775123
##
## Residual standard error: 1.475718
## Degrees of freedom: 115 total; 113 residual
```

Although the Phi Coefficient is not significant with a value of 0.2303, the residual standard error of the model that is fitted with autocorrelated error is 1.475, which means that there can be a possibility of a linear fit.

c)

```
prob1_c = lm(temp ~ I(year) + I(year^2) + I(year^3) + I(year^4) + I(year^5) + I(year^6) + I(year^7) +I(
prob1_back = step(prob1_c, direction = 'backward', trace=10)
```

```
## Start:  AIC=83.25
## temp ~ I(year) + I(year^2) + I(year^3) + I(year^4) + I(year^5) +
##     I(year^6) + I(year^7) + I(year^8) + I(year^9) + I(year^10)
##
##
## Step:  AIC=83.25
## temp ~ I(year) + I(year^2) + I(year^3) + I(year^4) + I(year^5) +
##     I(year^6) + I(year^7) + I(year^8) + I(year^9)
##
##
## Step:  AIC=83.25
## temp ~ I(year) + I(year^2) + I(year^3) + I(year^4) + I(year^5) +
##     I(year^6) + I(year^7) + I(year^8)
##
##
## Step:  AIC=83.25
## temp ~ I(year) + I(year^2) + I(year^3) + I(year^4) + I(year^5) +
##     I(year^6) + I(year^8)
##
##
## Step:  AIC=83.25
## temp ~ I(year) + I(year^2) + I(year^3) + I(year^4) + I(year^5) +
##     I(year^8)
##
##
```

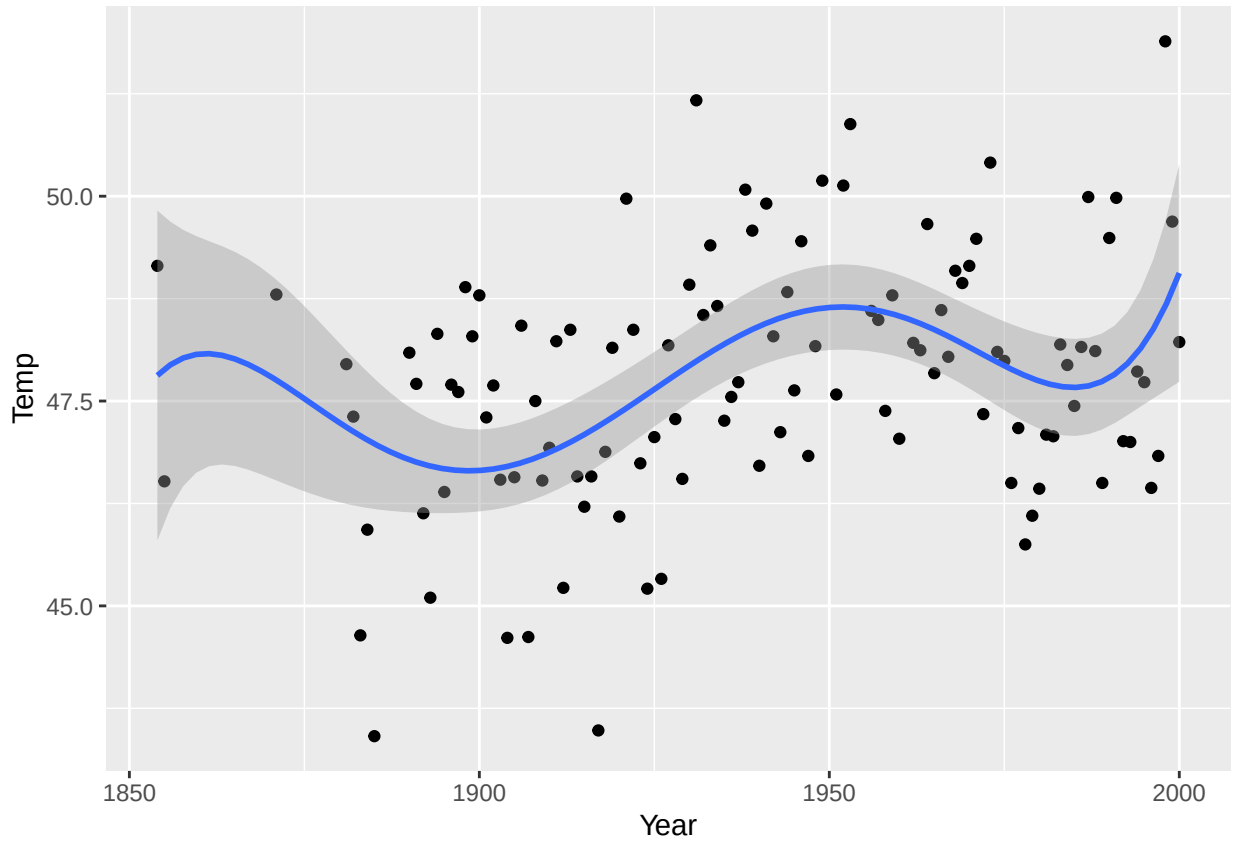
```
## Step:  AIC=83.25
## temp ~ I(year) + I(year^2) + I(year^3) + I(year^4) + I(year^8)
##
##             Df Sum of Sq    RSS    AIC
## <none>                    213.68 83.249
## - I(year)    1    11.093 224.78 87.069
## - I(year^2)  1    11.155 224.84 87.101
## - I(year^3)  1    11.217 224.90 87.133
## - I(year^4)  1    11.278 224.96 87.164
## - I(year^8)  1    11.512 225.19 87.284
```

```
summary(prob1_back)
```

```
##
## Call:
## lm(formula = temp ~ I(year) + I(year^2) + I(year^3) + I(year^4) +
##     I(year^8), data = aatemp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7126 -0.9175 -0.1441  0.9905  3.2313
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.049e+07  2.972e+07  -2.372   0.0194 *
## I(year)      1.676e+05  7.047e+04   2.379   0.0191 *
## I(year^2)   -1.526e+02  6.396e+01  -2.385   0.0188 *
## I(year^3)    6.347e-02  2.653e-02   2.392   0.0185 *
## I(year^4)   -1.031e-05  4.299e-06  -2.399   0.0182 *
## I(year^8)    1.074e-20  4.432e-21   2.423   0.0170 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.4 on 109 degrees of freedom
## Multiple R-squared:  0.1955, Adjusted R-squared:  0.1586
## F-statistic: 5.298 on 5 and 109 DF,  p-value: 0.0002141
```

```
prob1_c2 = lm(temp ~ I(year) + I(year^2) + I(year^3) + I(year^4) + I(year^8), data=aatemp)
```

```
library(ggplot2)
ggplot(data = aatemp, aes(x = year, y = temp)) + geom_point() + geom_smooth(method = 'lm', formula = y
```



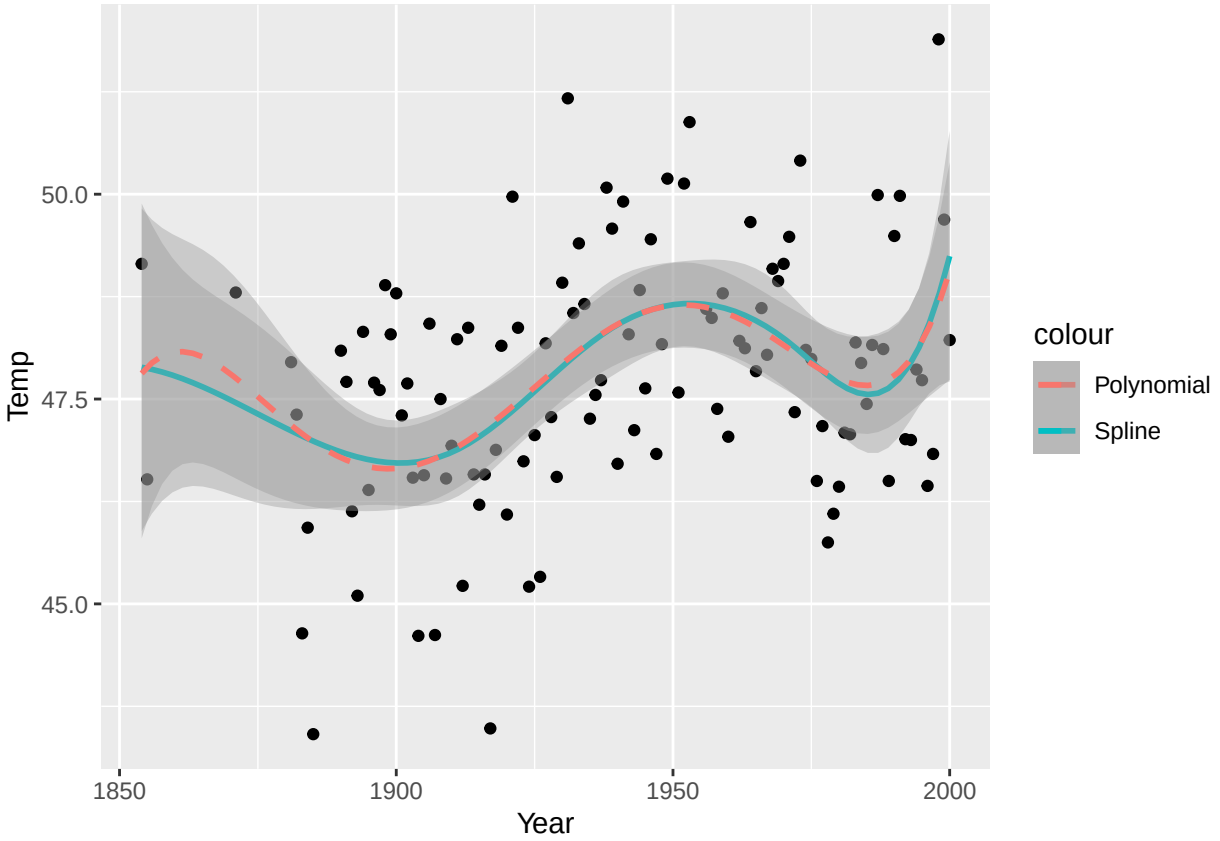

```
predict(prob1_back, newdata = data.frame(year=2001))
```

```
##        1
## 49.29922
```

The model fitted shows a pattern of a linear model. The predicted value at year 2001 is 49.29922

d)

```
library(splines)
prob1_d = lm(temp~bs(year, df = 6, intercept = TRUE), data = aatemp)
ggplot(data = aatemp, aes(x = year, y = temp)) + geom_point() + geom_smooth(method = 'lm', formula = y
```



The plot above shows that the two models are almost equal, which means that there is no much better fitted model. But just by looking at the data points, it still seems to have a linear relationship.

Problem 2

a)

```
data(infmort)
head(infmort)
```

```
##                       region income mortality           oil
## Australia               Asia   3426      26.7 no oil exports
## Austria               Europe   3350      23.7 no oil exports
## Belgium               Europe   3346      17.0 no oil exports
## Canada              Americas   4751      16.8 no oil exports
## Denmark               Europe   5029      13.5 no oil exports
## Finland               Europe   3312      10.1 no oil exports
```

Income and mortality are numerical variables and region and oil are categorical variables.

b)

```r
prob2_b = lm(mortality ~ ., data = infmort)
summary(prob2_b)
```

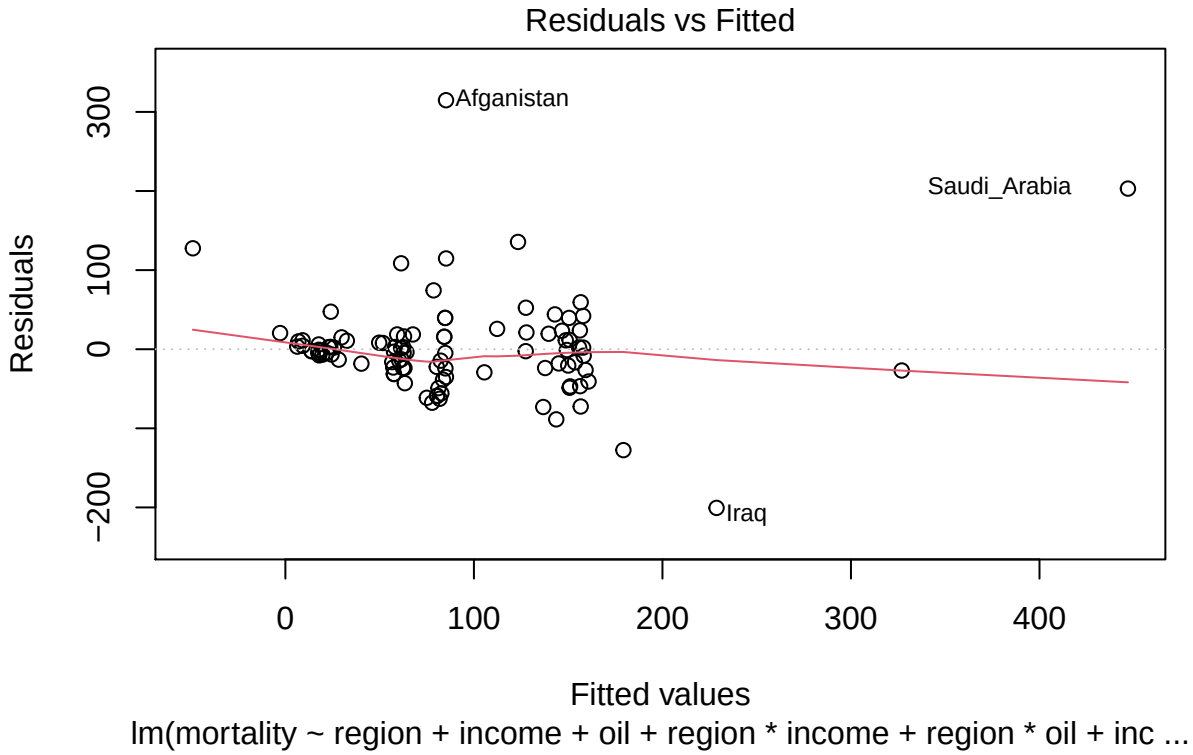
```
##
## Call:
## lm(formula = mortality ~ ., data = infmort)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -156.00  -32.20   -4.44   13.65  488.82
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.152e+02  2.974e+01   7.234 1.19e-10 ***
## regionEurope      -1.015e+02  3.073e+01  -3.303 0.001351 **
## regionAsia        -4.589e+01  2.014e+01  -2.278 0.024977 *
## regionAmericas    -8.365e+01  2.180e+01  -3.837 0.000224 ***
## income            -5.290e-03  7.404e-03  -0.714 0.476685
## oilno oil exports -7.834e+01  2.891e+01  -2.710 0.007992 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77.36 on 95 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.3105, Adjusted R-squared:  0.2742
## F-statistic: 8.556 on 5 and 95 DF,  p-value: 1.015e-06
```

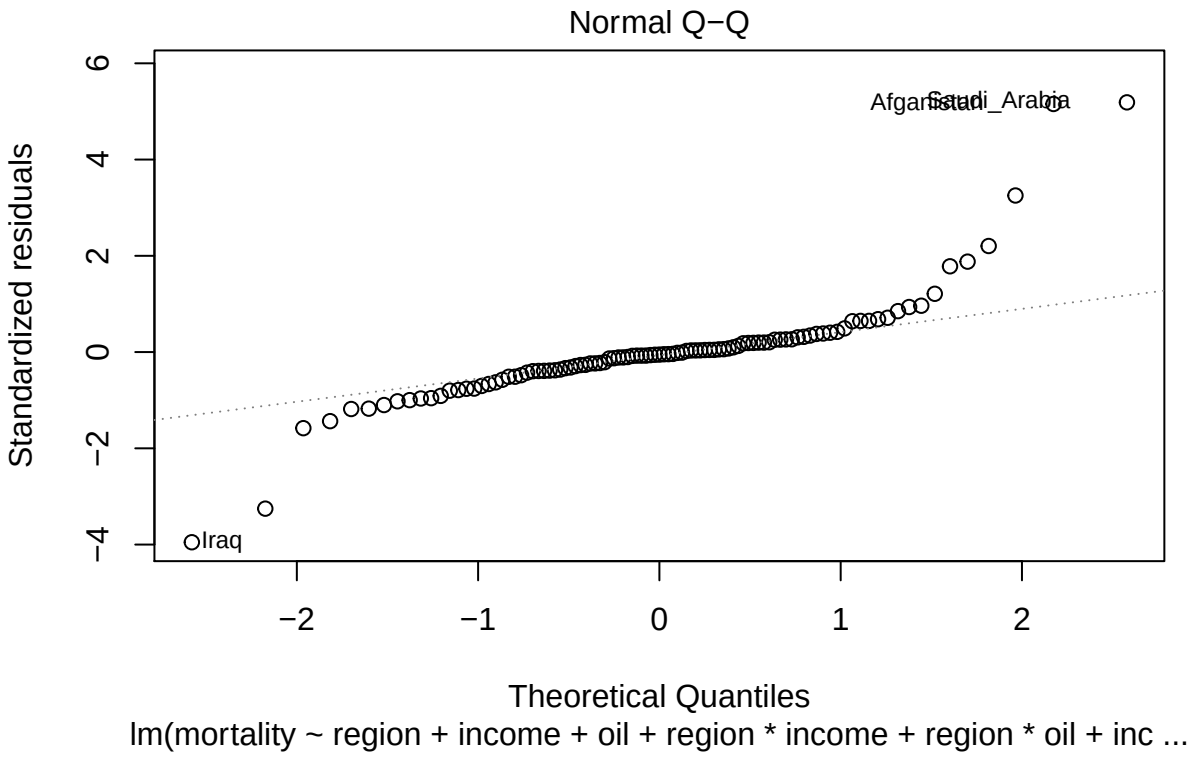
```r
prob2_b2 = lm(mortality ~ region + income + oil + region*income + region*oil + income*oil, data = infmo
summary(prob2_b2)
```

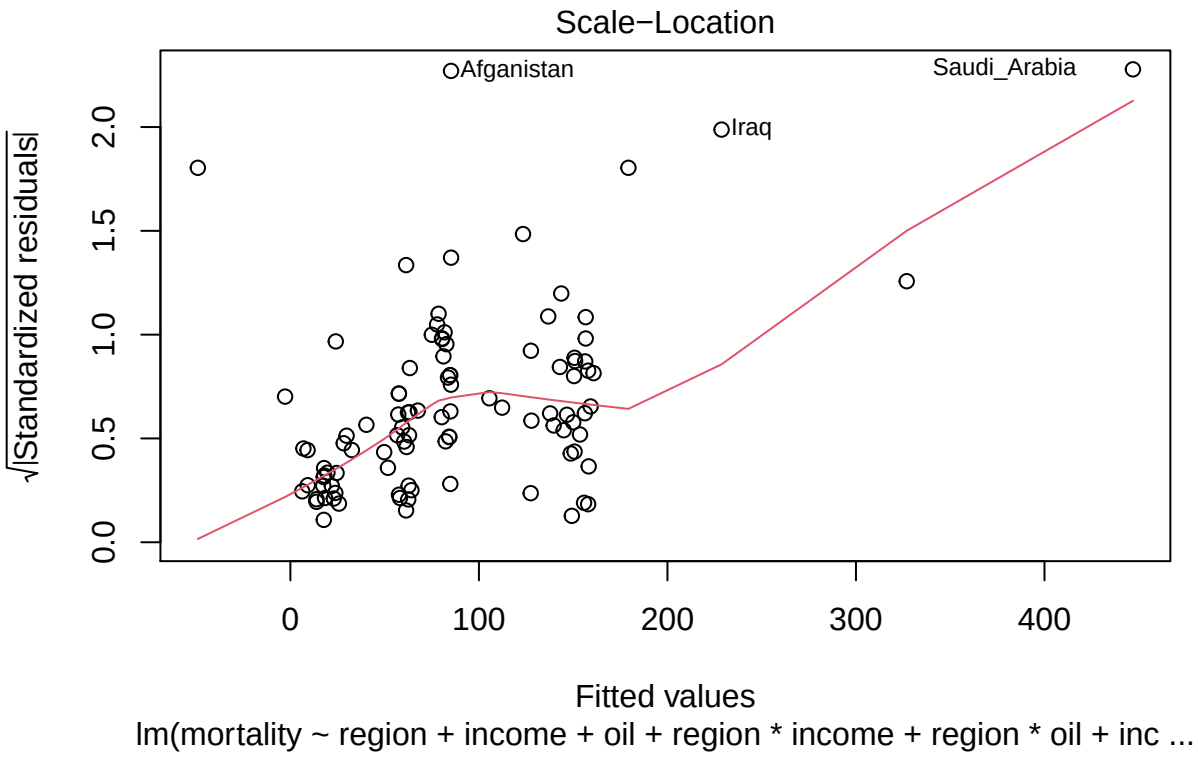
```
##
## Call:
## lm(formula = mortality ~ region + income + oil + region * income +
##     region * oil + income * oil, data = infmort)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -200.606  -23.858   -2.578   15.676  314.797
##
## Coefficients: (1 not defined because of singularities)
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       27.88469   48.70131   0.573 0.568383
## regionEurope                    -133.35340   37.99857  -3.509 0.000707 ***
## regionAsia                        74.83990   64.29758   1.164 0.247550
## regionAmericas                  -134.64863   69.44359  -1.939 0.055674 .
## income                            0.09935    0.02687   3.697 0.000376 ***
## oilno oil exports               140.12320   48.65760   2.880 0.004984 **
## regionEurope:income               0.13887    0.04451   3.120 0.002441 **
## regionAsia:income                 0.12561    0.04368   2.876 0.005041 **
## regionAmericas:income             0.13134    0.04397   2.987 0.003641 **
## regionEurope:oilno oil exports         NA         NA      NA       NA
## regionAsia:oilno oil exports    -156.27099   62.28212  -2.509 0.013915 *
```

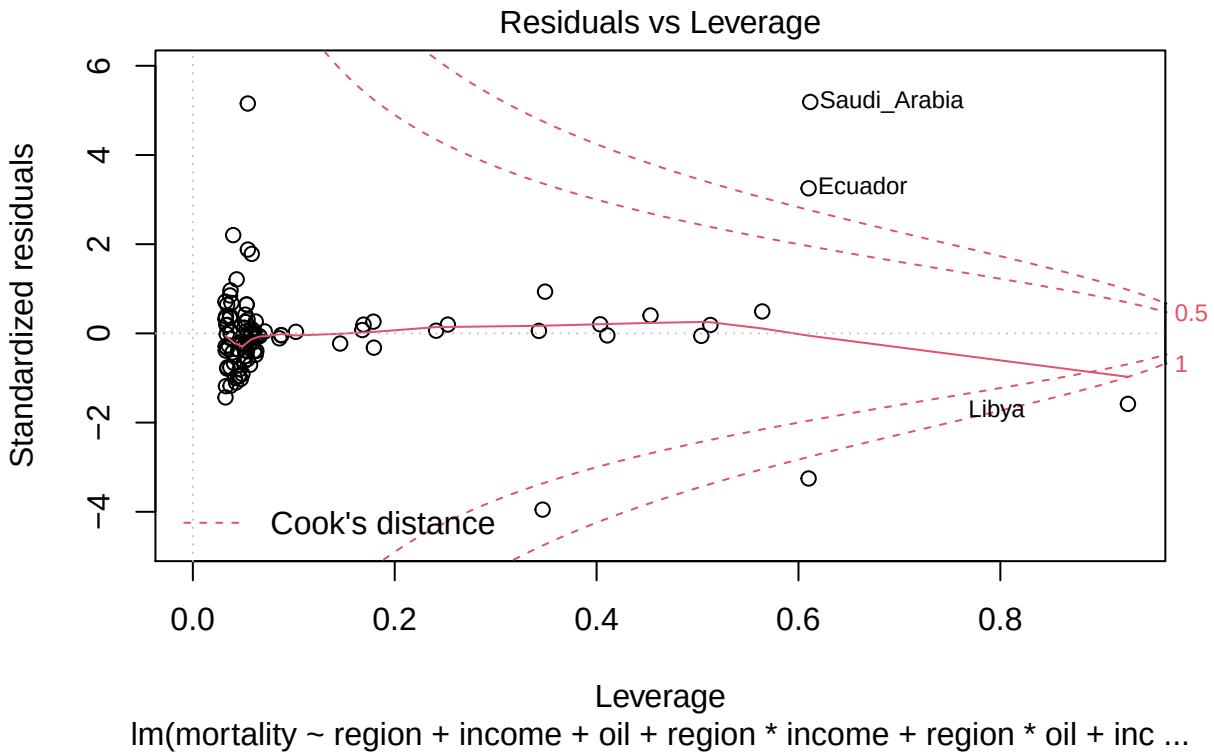
```
## regionAmericas:oilno oil exports   33.37454   67.81541   0.492 0.623834
## income:oilno oil exports           -0.24328    0.04140  -5.876 7.17e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.8 on 89 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.5742, Adjusted R-squared:  0.5216
## F-statistic: 10.91 on 11 and 89 DF,  p-value: 1.825e-12
```

```
plot(prob2_b2)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(mortality ~ region + income + oil + region * income + region * oil + inc ...
Iraq
Afganistan
Saudi_Arabia
6
4
2
0
−2
−4
−2
−1
0
1
2


Scale−Location
Afganistan
Saudi_Arabia
Iraq
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
100
200
300
400
Fitted values
lm(mortality ~ region + income + oil + region * income + region * oil + inc ...



The model seems to show normal distribution and constant variance. Model could remove Afganistan and Saudi Arabia from the data to show better data, but it cannot be assumed that it has a linear relationship.

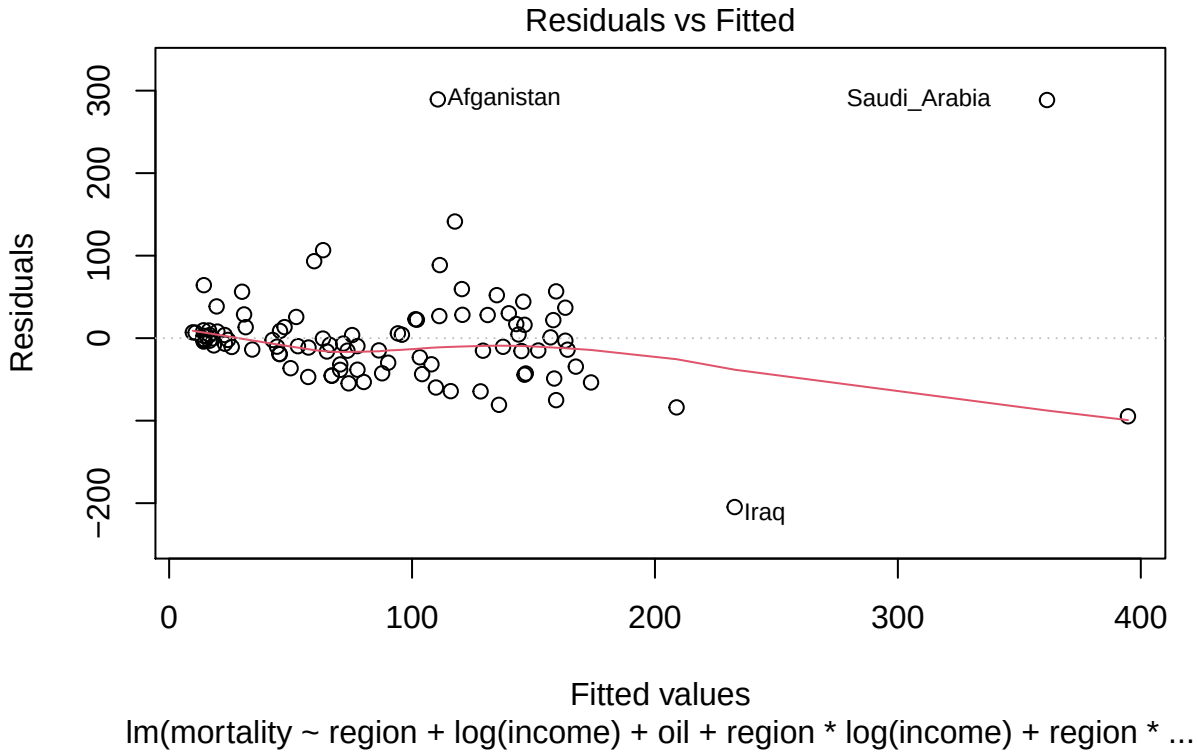
c)

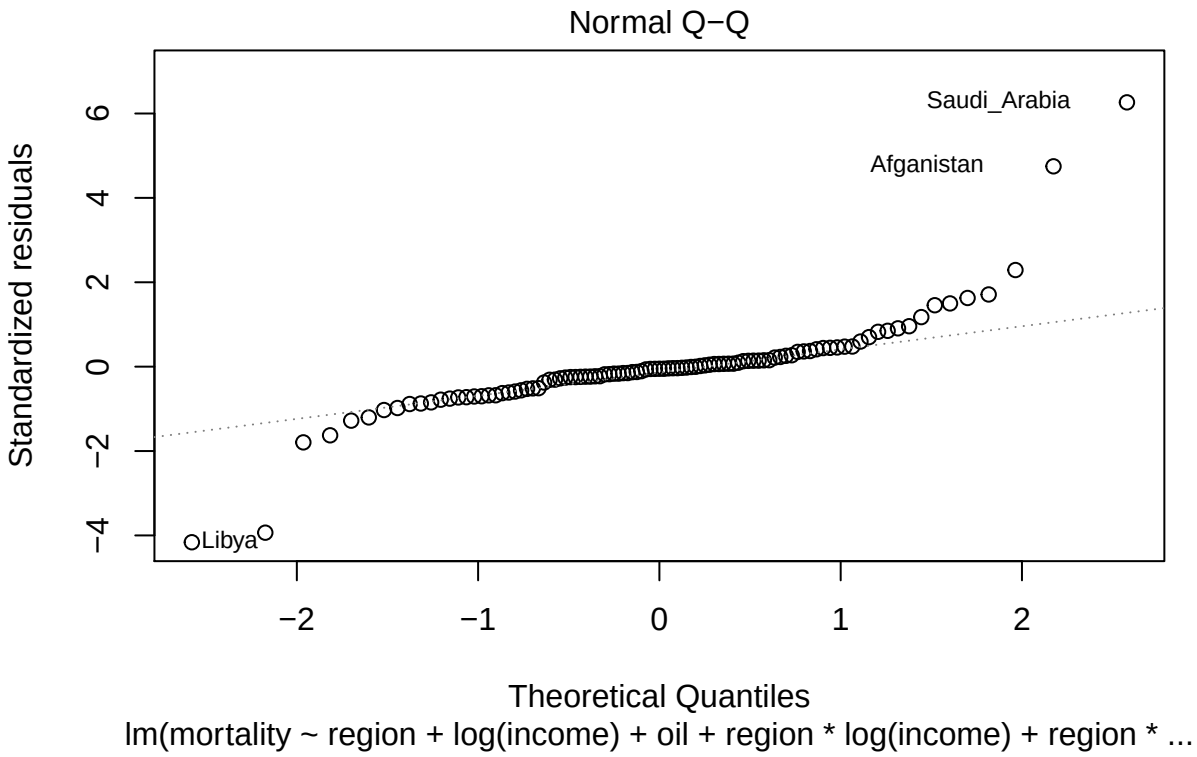
```
library(MASS)
prob2_b3 = lm(mortality ~ region + log(income) + oil + region*log(income) + region*oil + income*oil, da
summary(prob2_b3)
```

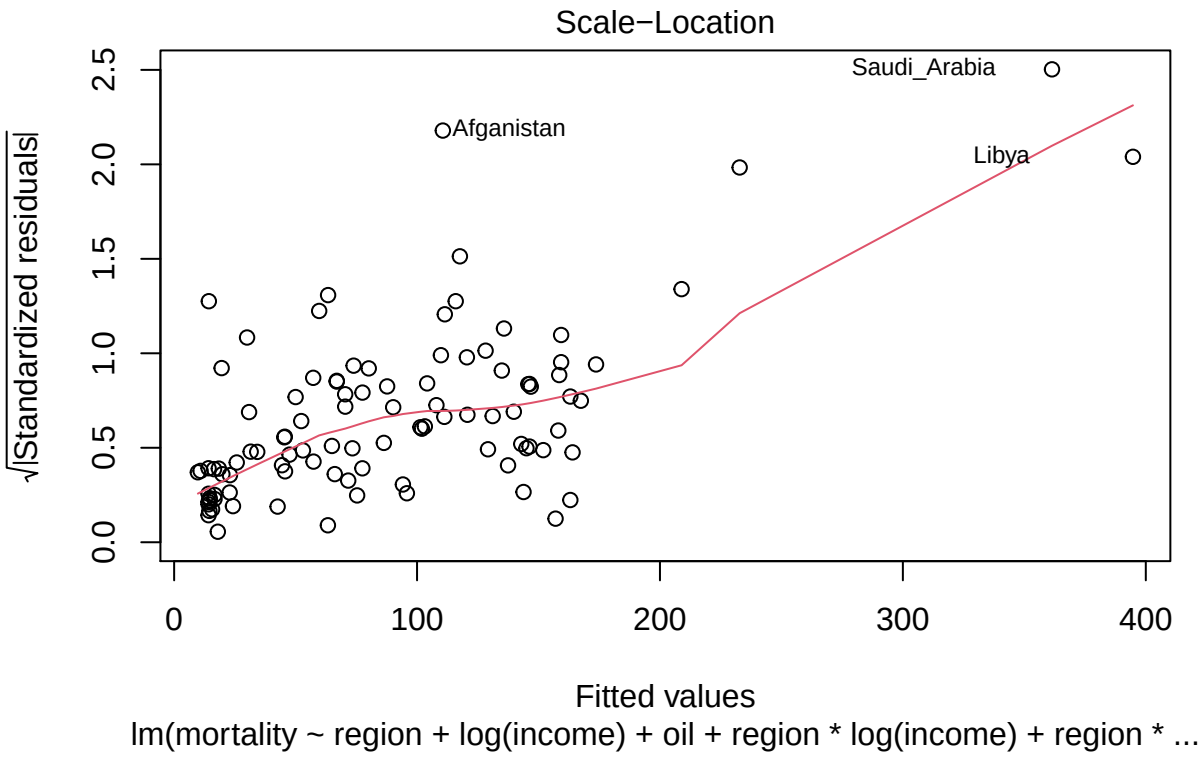
```
##
## Call:
## lm(formula = mortality ~ region + log(income) + oil + region *
##     log(income) + region * oil + income * oil, data = infmort)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -204.706  -29.775   -3.101   13.387  289.383
##
## Coefficients: (1 not defined because of singularities)
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    155.91464  109.90910   1.419 0.159554
## regionEurope                   -63.61927  266.11532  -0.239 0.811610
## regionAsia                     179.60007  138.45716   1.297 0.197969
## regionAmericas                  28.21563  192.85881   0.146 0.884018
## log(income)                    -31.99028   17.77066  -1.800 0.075260 .
```

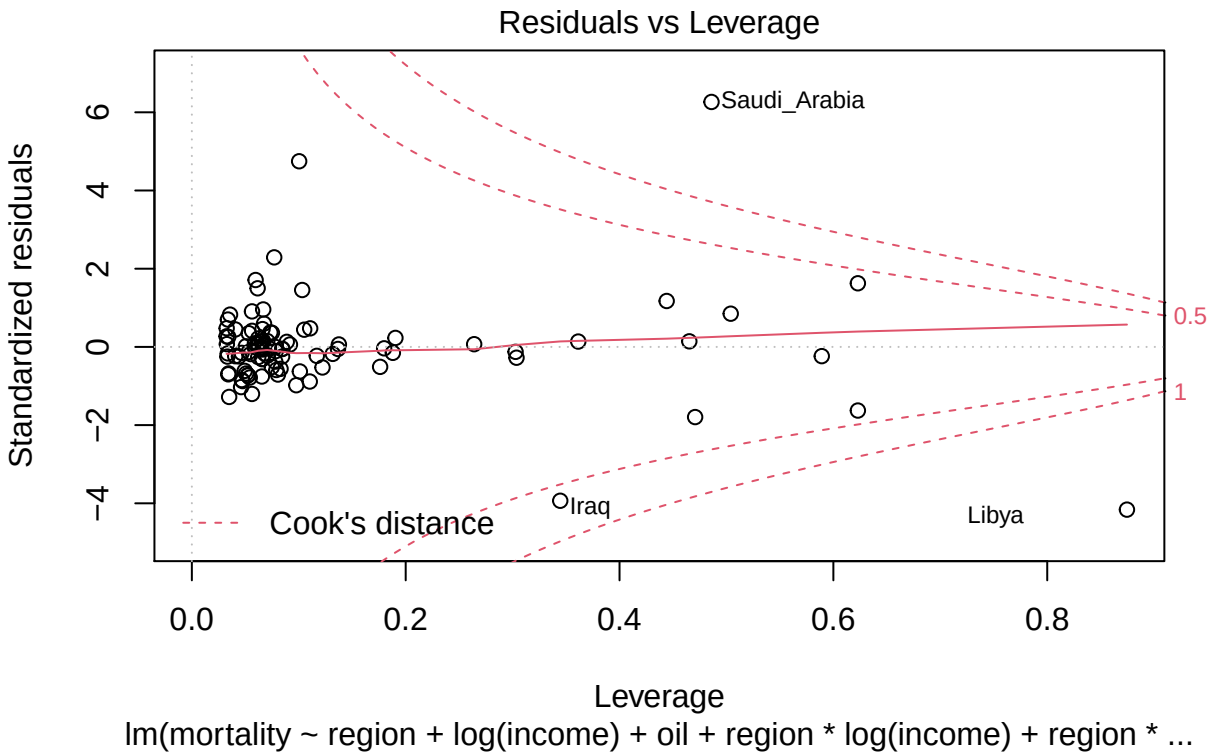
```
## oilno oil exports                 142.47222   51.14658   2.786 0.006543 **
## income                              0.16446    0.02951   5.573  2.7e-07 ***
## regionEurope:log(income)            1.12957   38.47905   0.029 0.976648
## regionAsia:log(income)              1.20496   20.53597   0.059 0.953343
## regionAmericas:log(income)         -6.21740   28.83767  -0.216 0.829798
## regionEurope:oilno oil exports           NA         NA      NA       NA
## regionAsia:oilno oil exports     -235.12398   61.07299  -3.850 0.000224 ***
## regionAmericas:oilno oil exports  -35.77635   66.24508  -0.540 0.590519
## oilno oil exports:income           -0.15554    0.02866  -5.427  5.0e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64.26 on 88 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.5592, Adjusted R-squared:  0.4991
## F-statistic: 9.304 on 12 and 88 DF,  p-value: 2.53e-11
```

```
plot(prob2_b3)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(mortality ~ region + log(income) + oil + region * log(income) + region * ...
Saudi_Arabia
Afganistan
Libya
−4
−2
0
2
4
6
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
0
100
200
300
400
Saudi_Arabia
Afganistan
Libya
Fitted values
lm(mortality ~ region + log(income) + oil + region * log(income) + region * ...



log transformation could be performed. After using log transformation the data seems to have better constant variance. But there is not much difference.

d)

In Asia, America, Asia*oil export, the mortality is decreased, but rest of them increases the mortality.