BIS623 Homework 7

Due on 11/14/2019 before the lecture

Joanna Chen

13 November, 2019

Please use the data `Leinhardt` in the R library `carData`.

```
if(!require("carData")){
  install.packages("carData")
}
library(carData)

if(!require("data.table")){
  install.packages("data.table")
}
library(data.table)

if(!require("ggfortify")){
  install.packages("ggfortify")
}
library(ggfortify)

# Read in the data and set it as data.table
dt = Leinhardt
setDT(dt)
```

The dataset four variables:

- income: Per-capita income in U. S. dollars
- infant: Infant-mortality rate per 1000 live births.
- region: A factor with levels
- oil: Oil-exporting country; yes or no

There are missing data in the original dataset. Please use `na.omit` to remove the missing data which results in 101 observation.

```
dt = na.omit(dt)
```

Question 1:

Fit a MLR to assess the impact of income and oil exporting on infant mortality. What is the adjusted R^2 for this model?

```
myfit=lm(infant ~ income + oil, data=dt)
summary(myfit)
```

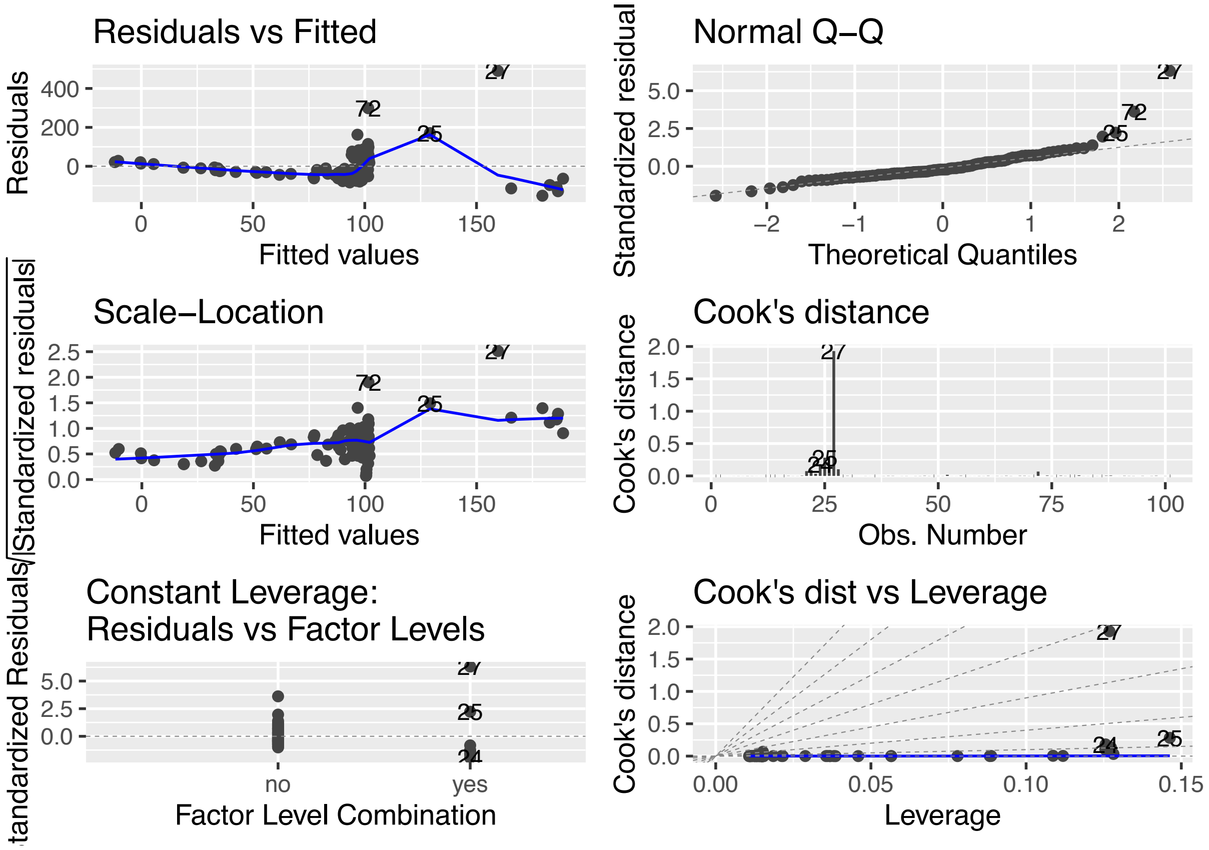
```
##
## Call:
## lm(formula = infant ~ income + oil, data = dt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -151.39  -45.94  -17.78   28.88  490.43
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 103.083399  10.490381   9.826 2.88e-16 ***
## income       -0.020540   0.005793  -3.546 0.000602 ***
## oilyes       87.910994  30.642453   2.869 0.005045 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 83.15 on 98 degrees of freedom
## Multiple R-squared:  0.1783, Adjusted R-squared:  0.1615
## F-statistic: 10.63 on 2 and 98 DF,  p-value: 6.631e-05
```

From the model, we can see that the adjusted R-squared = 0.1615

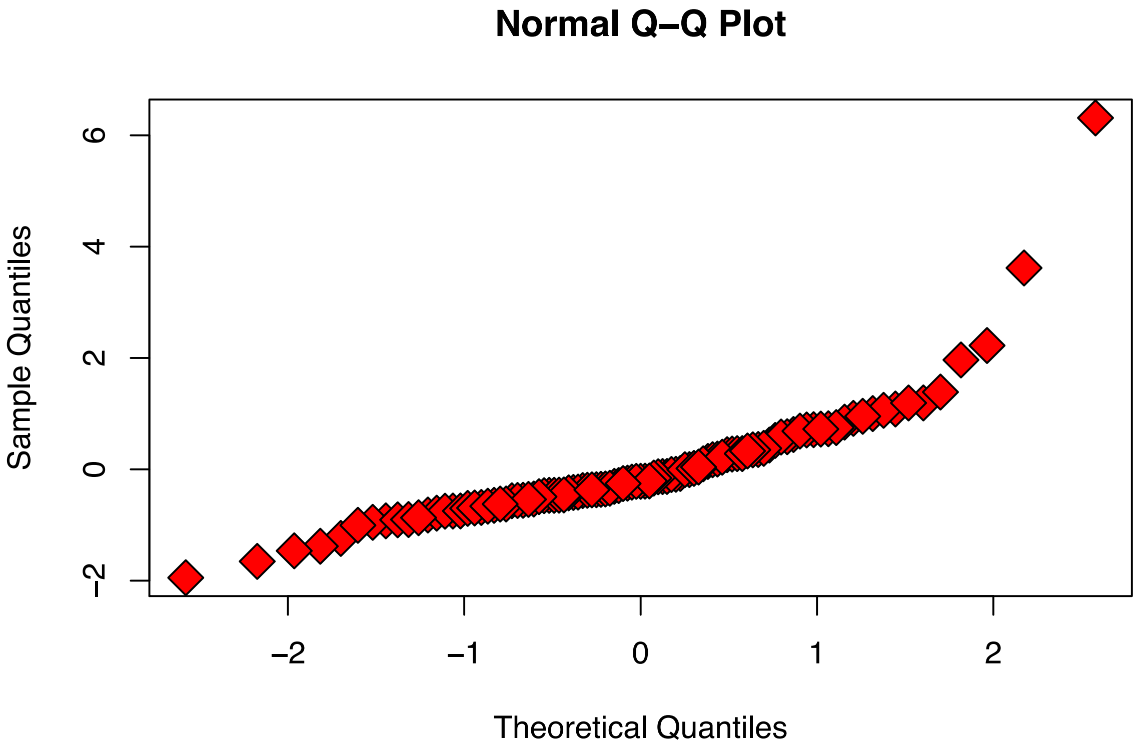
Question 2:

Plot the diagnosis plots. Do you see any problem in the model fitting and how you want to modify the model?

```r
autoplot(myfit,which=1:6,label.size = 3)
```



```r
qqnorm(rstandard(myfit), pch=23, bg='red', cex=2)
```



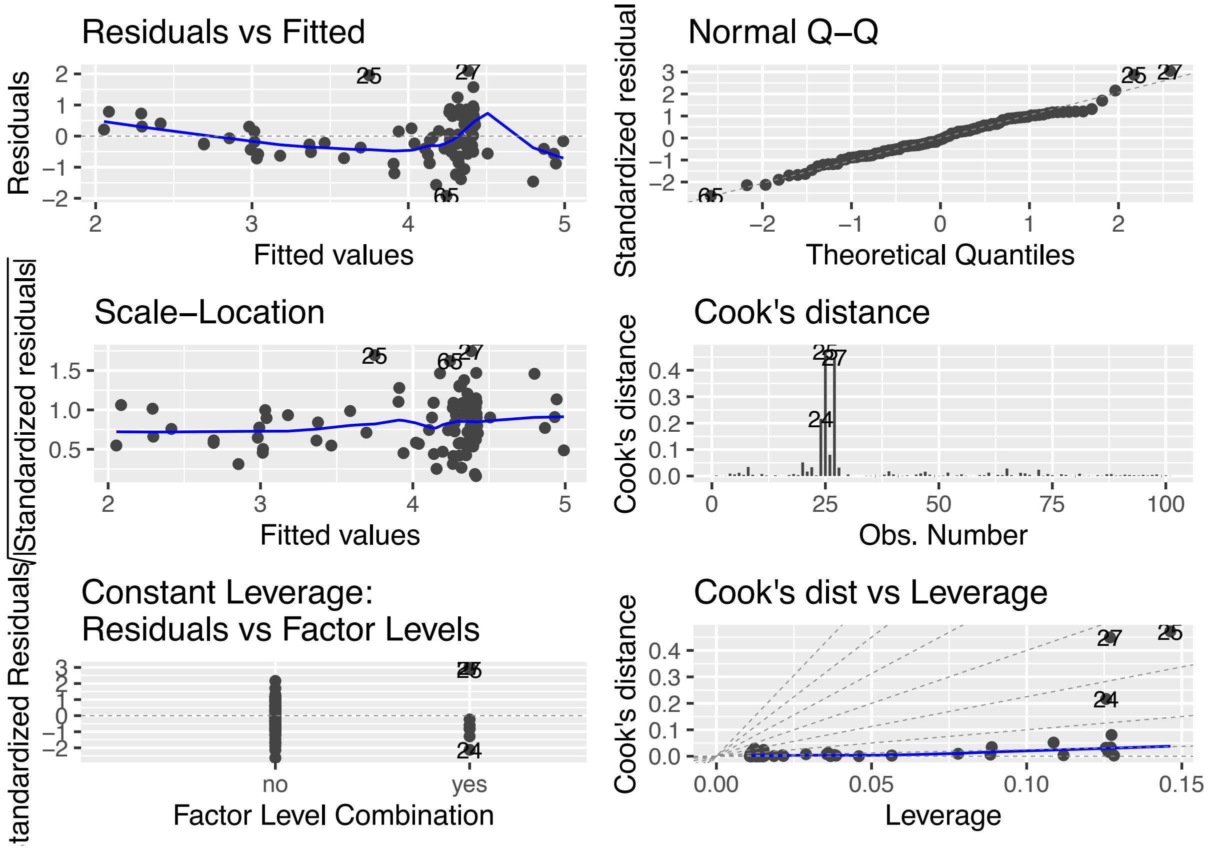
The residual is not normal which we can tell from the Normal QQ-plot. If the residuals were really normal, we'd expect this plot to be roughly on the diagonal. I would like to use log transformation on the data.

Question 3:

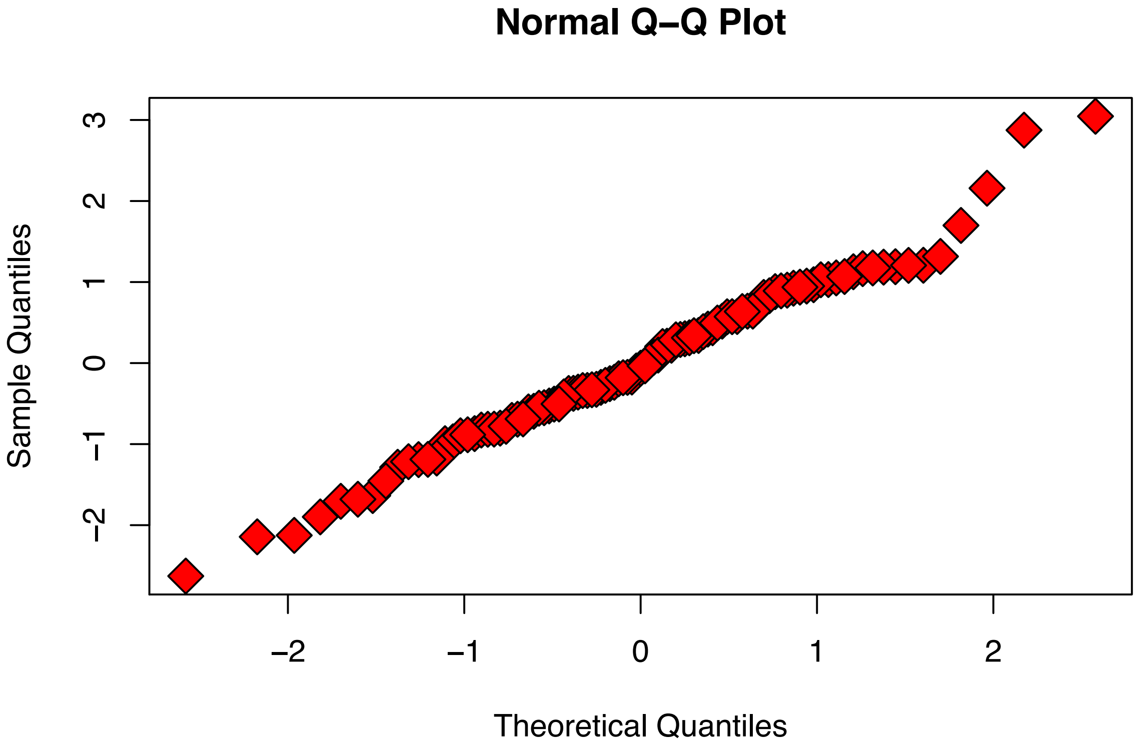
Re-fit the model based on the revised model and re-plot the diagnosis plots to make sure the model fitting issue has been solved.

I have tried taking log on the infant, income, on both, and weight variance. It turns out that taking log on the infant variable is the best. The new QQ plot is be roughly on the diagonal and that means the residual is close to normal.

```
dt$loginfant = log(dt$infant)
#Re-fit the model
myfit1=lm(loginfant ~ income + oil, data=dt)
#plot(myfit1)
#Re-plot the diagnosis plots
autoplot(myfit1,which=1:6,label.size = 3)
```



```r
qqnorm(rstandard(myfit1), pch=23, bg='red', cex=2)
```

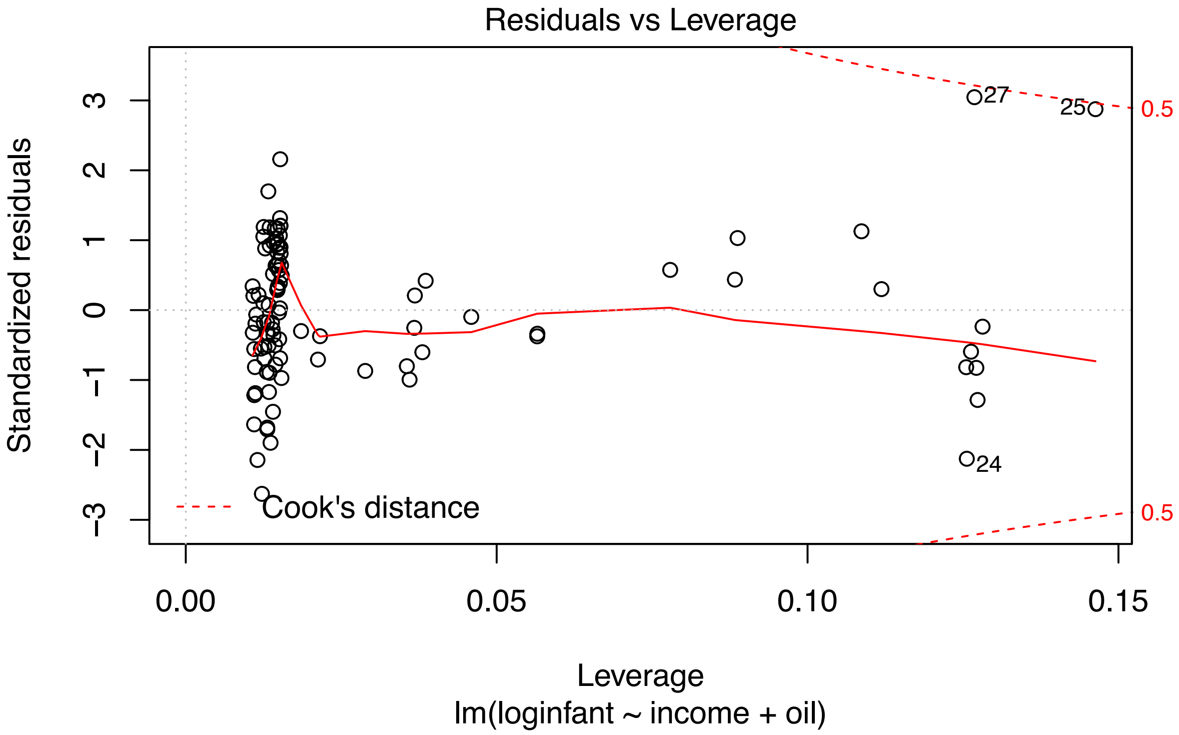


Question 4:

Based on the new diagnosis plots, do see you still see any leverage point/outliers/influential points? If so, remove them and re-fit the data. What is the adjuated R^2 for the current model? Give your conclusion on the impact of income and oil exporting on infact mortality.

i. Leverage points

```
plot(myfit1, 5)
```



The plot above highlights the top 3 most extreme points (#24, 25, 27).

Theoretically, typical rules of thumb are 2(p+1)/n or 3(p+1)/n, where n is the number of observations and p the number of predictor variables.

```
HighLeverage = cooks.distance(myfit1) > (3*(2+1)/nrow(dt))
```

We can see the #24,25,27 return false.

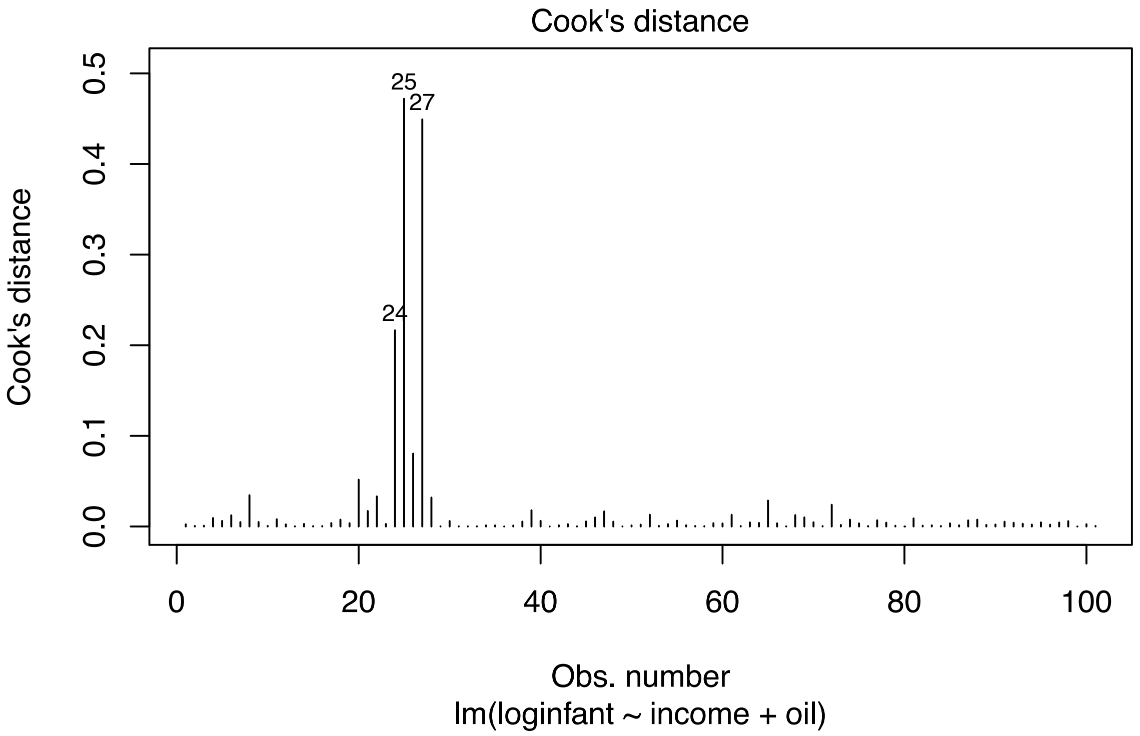
ii. Outlier
From the previous plot, we can see that #27 exceed 3 standard deviations which is considered as outlier. Theoretically,

```
LargeResiduals <- rstudent(myfit1) > 3
```

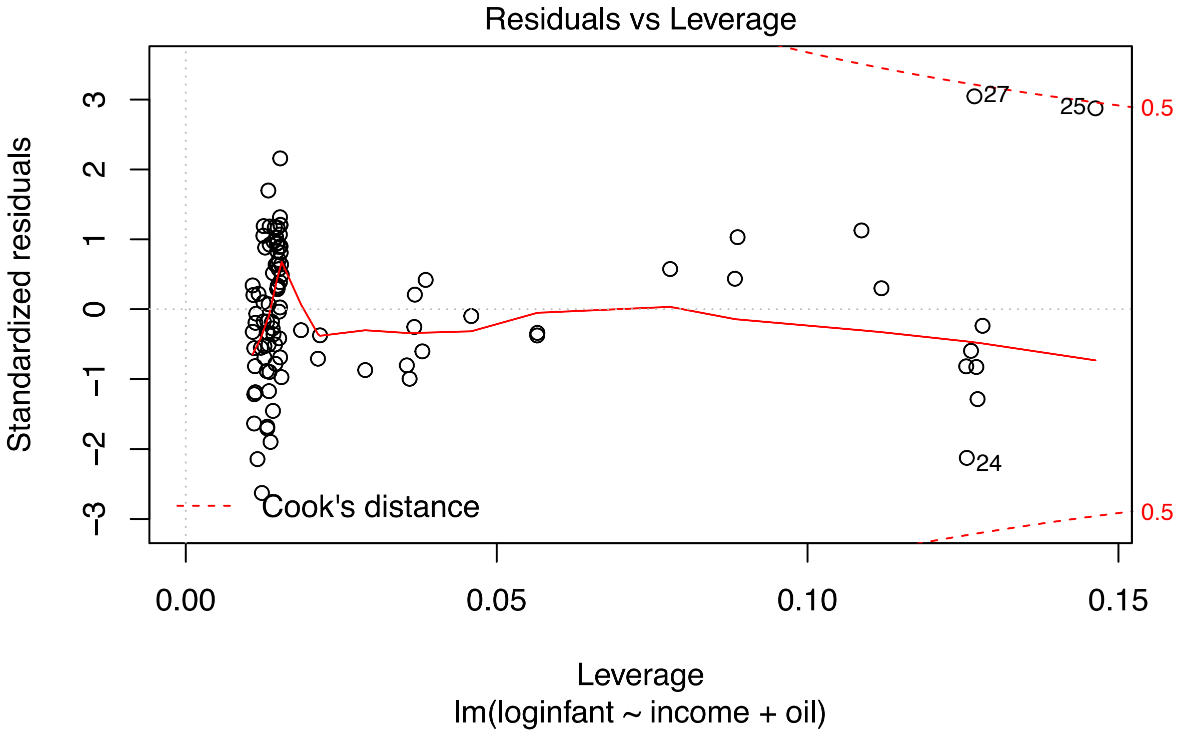
We can see the #27 returns false.

iii. Influential values
Use Cook's distance to determine the influence of a value.

```
# Cook's distance
plot(myfit1, 4)
```



```
# Residuals vs Leverage
plot(myfit1, 5)
```



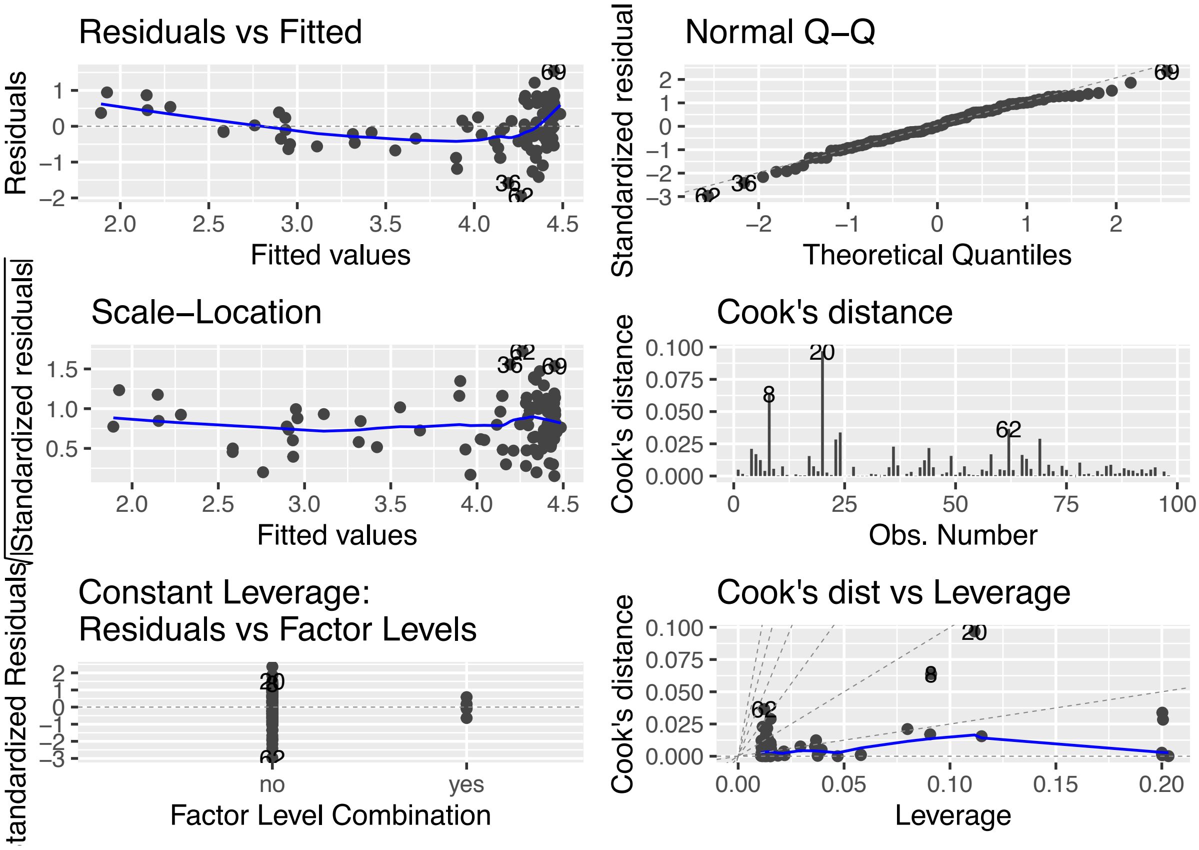
The top 3 most extreme values are labelled on the Cook's distance plot. But these data don't present any influential points because all points are well inside of the Cook's distance lines on the Residuals vs Leverage plot.

Remove the values with high leverage and large residuals and refit the model. We can see that the updated adjusted R-squared = 0.5037. The diagnostic plot has been shown as following.

```
dt <- dt[!HighLeverage & !LargeResiduals,]
myfit2 = lm(loginfant ~ income + oil, data=dt)
summary(myfit2)
```

```
##
## Call:
## lm(formula = loginfant ~ income + oil, data = dt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.94143 -0.40656  0.00528  0.47323  1.54132
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.485e+00  8.340e-02  53.779   <2e-16 ***
## income      -4.635e-04  4.651e-05  -9.965   <2e-16 ***
## oilyes       5.191e-02  3.032e-01   0.171    0.864
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6578 on 95 degrees of freedom
## Multiple R-squared:  0.5139, Adjusted R-squared:  0.5037
## F-statistic: 50.22 on 2 and 95 DF,  p-value: 1.312e-15
```

```
autoplot(myfit2,which=1:6,label.size = 3)
```



Conclusion:

Let's first convert the categorical variable oil into a dummy variable which takes values 0 for yes and 1 for no.

The infant-mortality rate is higher by exp(5.191e-02) units for oil-exporting country than for non-oil-exporting country while all other variables held constant. The expected mean change in infant-mortality rate for one unit of change in the per-capita income while holding other predictors in the model constant is exp(-4.635e-04).