502_HW4

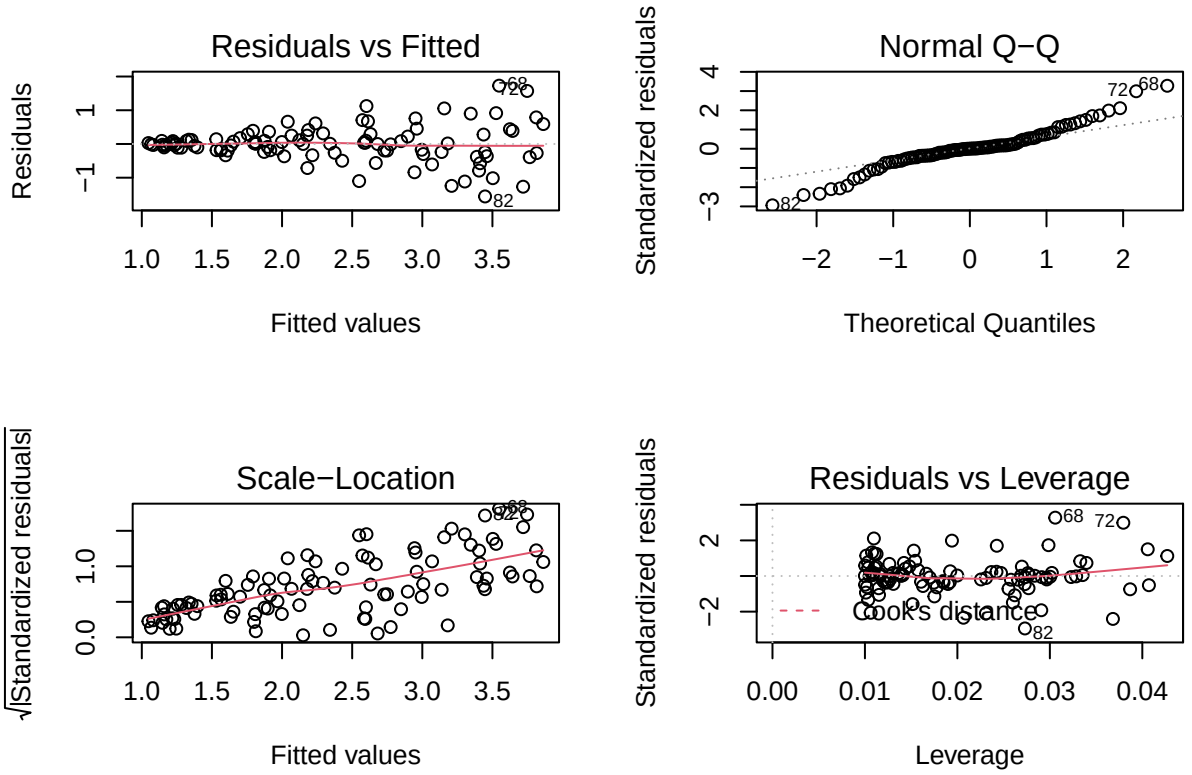
Shengbo Jin

2/2/2022

Q1

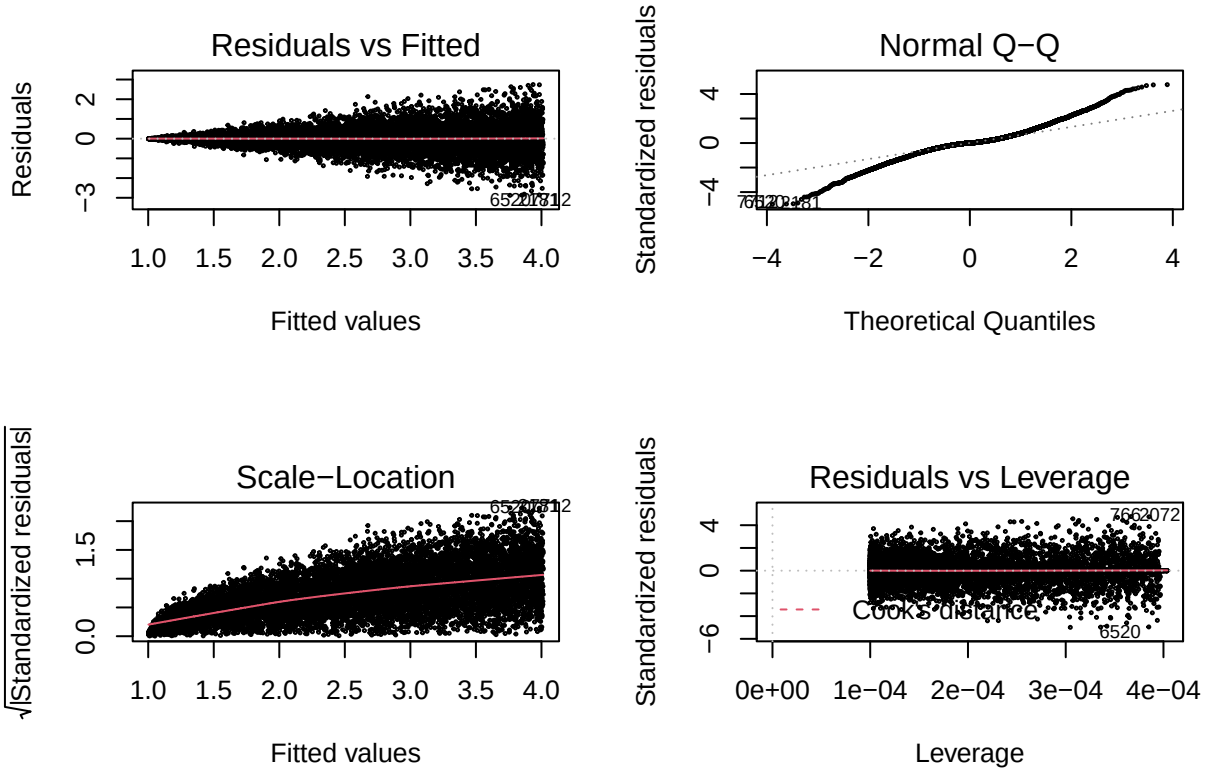
(a)

```
set.seed(446)
n <- 100
x1 <- runif(n)
eps <- rnorm(n)
y <- 1 + 3*x1 + x1*eps

model.a1 <- lm(y~x1)
par(mfrow=c(2,2))
plot(model.a1)
```



Checking for nonlinearity, the first plot does not show a systematic nonlinear trend, which is consistent with the true model; Checking for nonnormality, the second plot shows the residuals are not normally distributed (heavy-tailed), which is consistent with the true model; Checking for nonconstant variance, the third plot shows a systematic trend due to the variability of residuals changing over Yi, which is consistent with the true model.

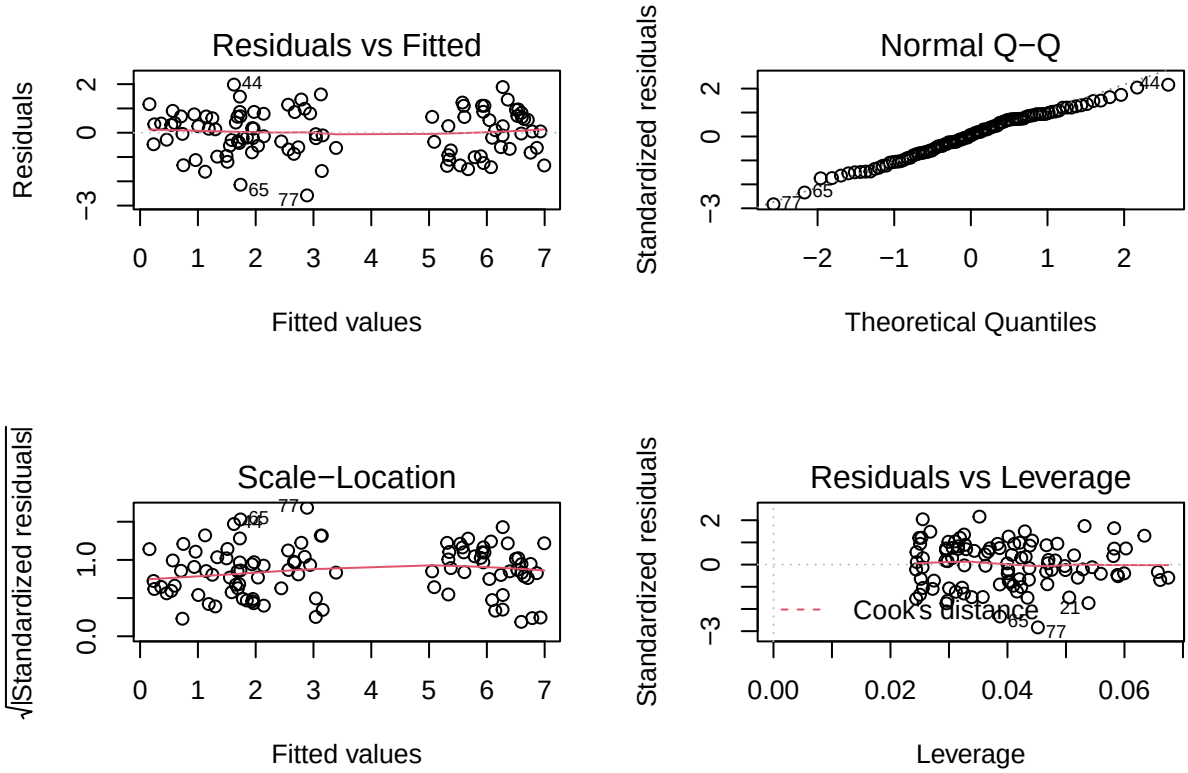
```
set.seed(446)
n <- 10000
x1 <- runif(n)
eps <- rnorm(n)
y <- 1 + 3*x1 + x1*eps

model.a2 <- lm(y~x1)
par(mfrow=c(2,2))
plot(model.a2, cex=0.25)
```



(b)

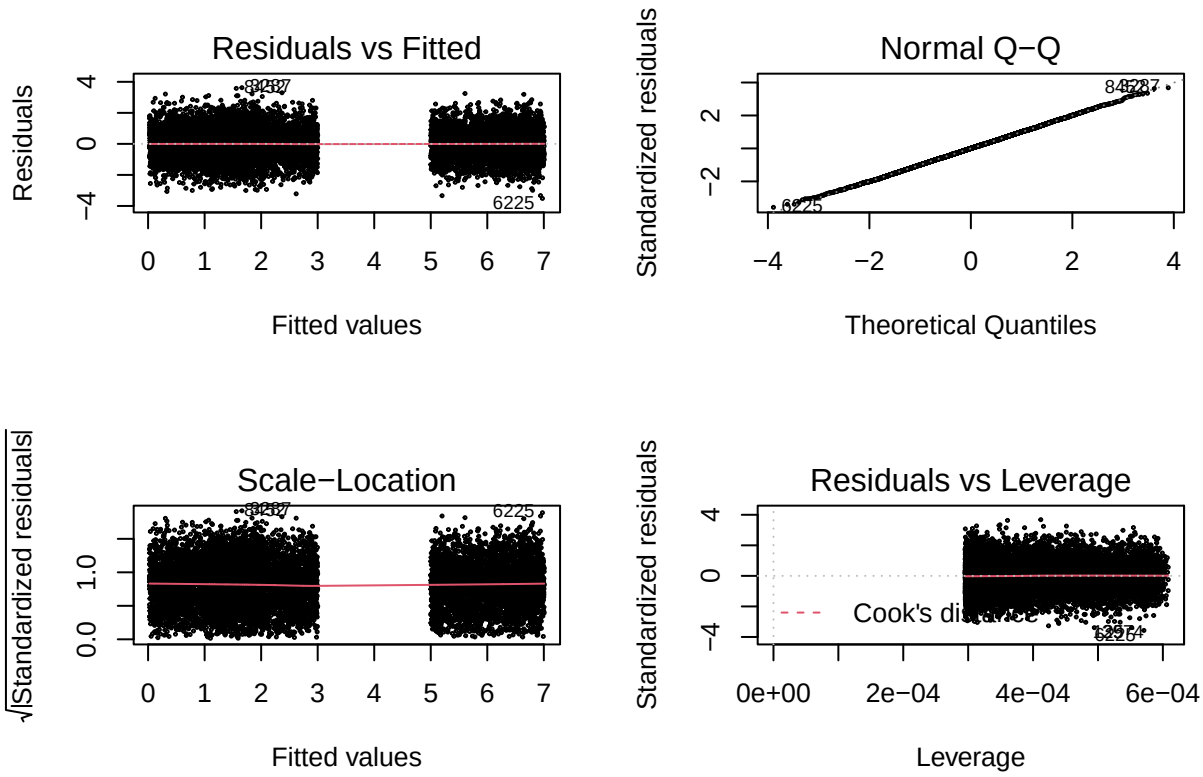
```
set.seed(576)
n <- 100
x1 <- runif(n)
x2 <- sample(factor(c("A","B","C")), replace=TRUE, size=n)
eps <- rnorm(n)
y <- 1 + 2*x1 + -1*(x2=="B") + 4*(x2=="C") + eps
```

```
model.b1 <- lm(y~x1+x2)
par(mfrow=c(2,2))
plot(model.b1)
```



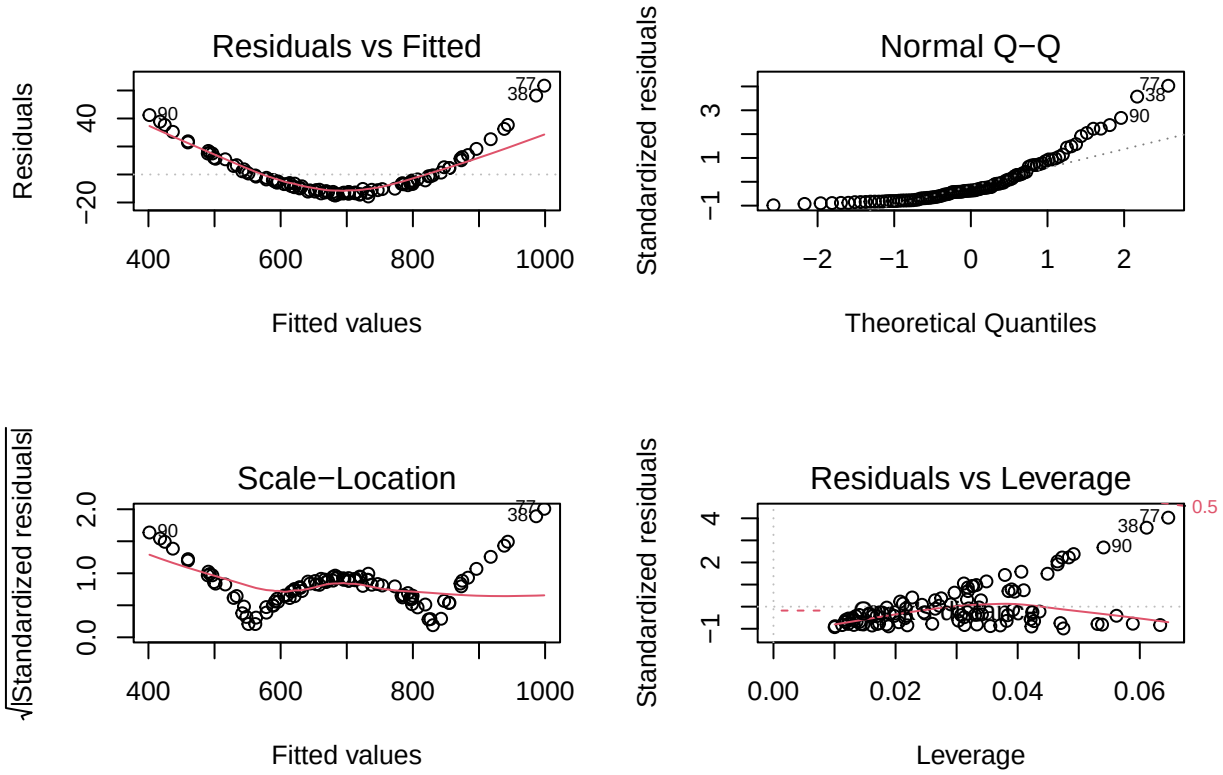
Checking for nonlinearity, the first plot does not show a systematic nonlinear trend, which is consistent with the true model; Checking for nonnormality, the second plot shows the residuals are normally distributed, which is consistent with the true model; Checking for nonconstant variance, the third plot does not shows a systematic trend, which is consistent with the true model.

```
set.seed(576)
n <- 10000
x1 <- runif(n)
x2 <- sample(factor(c("A","B","C")), replace=TRUE, size=n)
eps <- rnorm(n)
y <- 1 + 2*x1 + -1*(x2=="B") + 4*(x2=="C") + eps

model.b2 <- lm(y~x1+x2)
par(mfrow=c(2,2))
plot(model.b2, cex=0.25)
```



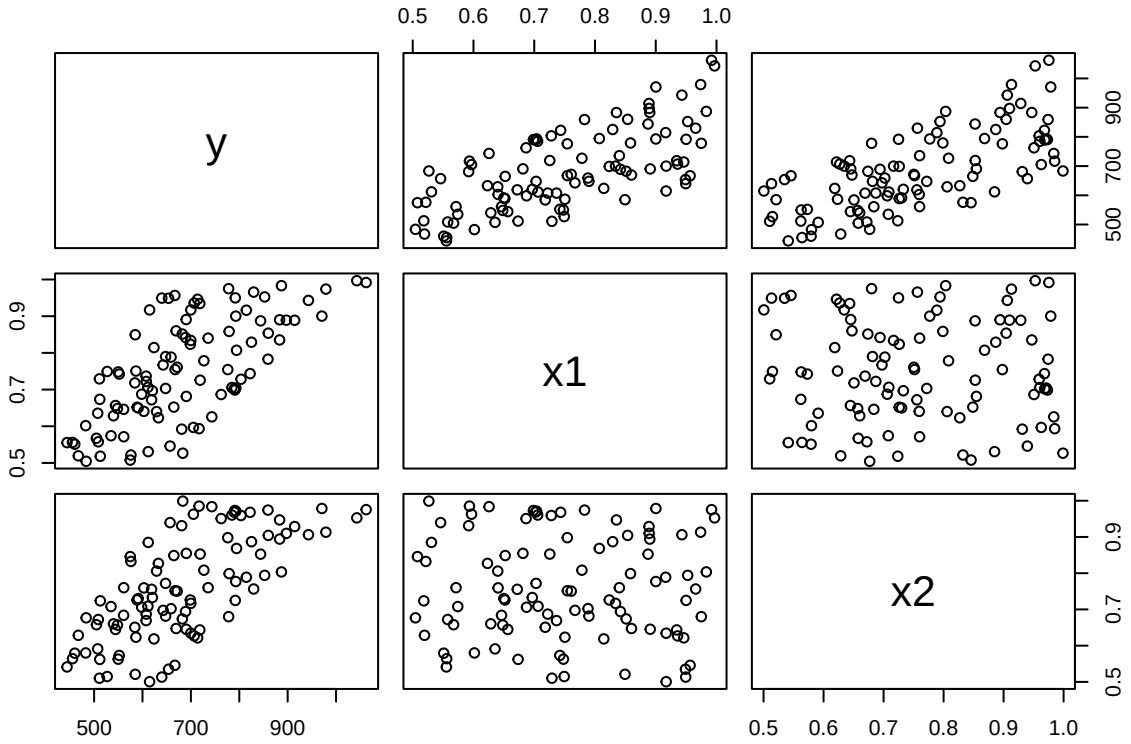
(c)

```
set.seed(308)
n <- 100
x1 <- runif(n, min=0.5, max=1)
x2 <- runif(n, min=0.5, max=1)
eps <- rnorm(n)
y <- exp(5+x1+x2) + eps

model.c1 <- lm(y~x1+x2)
par(mfrow=c(2,2))
plot(model.c1)
```



Checking for nonlinearity, the first plot shows a systematic nonlinear trend, which is consistent with the true model; Checking for nonnormality, the second plot shows the residuals are not normally distributed, which is not consistent with the true model; Checking for nonconstant variance, the third plot shows a systematic trend due to the variability of residuals changing over Yi, which is not consistent with the true model.

```
sim_reg <- data.frame(y, x1, x2)
pairs(sim_reg)
```



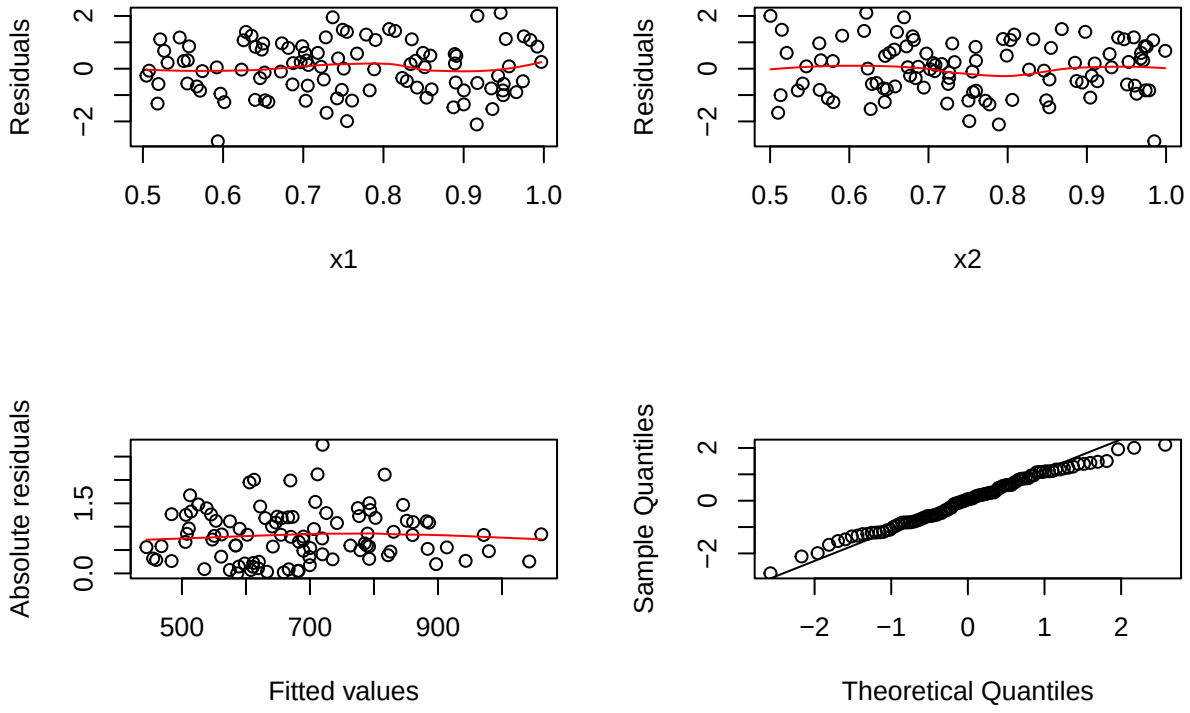
There is nonlinearity between x1 and x2 from the scatterplot matrix.

```
model.c2 <- nls(y~exp(b1+b2*x1+b3*x2), start=list(b1=5,b2=1,b3=1))
summary(model.c2)
```

```
##
## Formula: y ~ exp(b1 + b2 * x1 + b3 * x2)
##
## Parameters:
##    Estimate Std. Error t value Pr(>|t|)
## b1 5.001006   0.001174  4259.6   <2e-16 ***
## b2 0.999338   0.001039   961.7   <2e-16 ***
## b3 0.999652   0.001024   976.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9989 on 97 degrees of freedom
##
## Number of iterations to convergence: 2
## Achieved convergence tolerance: 1.795e-08
```

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

preds <- list(x1, x2)
labs <- c("x1", "x2")
```

```r
for (p in 1:length(preds)){
  x <- preds[[p]]
  plot(x, resid(model.c2), xlab = labs[p], ylab = "Residuals")
  smoother <- loess(resid(model.c2)~x)
  ord <- order(x)
  lines(x[ord], fitted(smoother)[ord], col="red")
}

plot(fitted(model.c2), abs(resid(model.c2)), xlab="Fitted values",
     ylab="Absolute residuals")
fit_loess <- loess(abs(resid(model.c2))~fitted(model.c2), span=1, deg=1)
ord_NLS <- order(fitted(model.c2))
lines(fitted(model.c2)[ord_NLS], fit_loess$fit[ord_NLS], col="red")

qqnorm(resid(model.c2), main="")
qqline(resid(model.c2))
```



Checking for nonlinearity, the first and second plot do not show a systematic nonlinear trend, which is consistent with the true model; Checking for nonconstant variance, the third plot does not shows a systematic trend, which is consistent with the true model; Checking for nonnormality, the last plot shows the residuals are normally distributed, which is consistent with the true model.

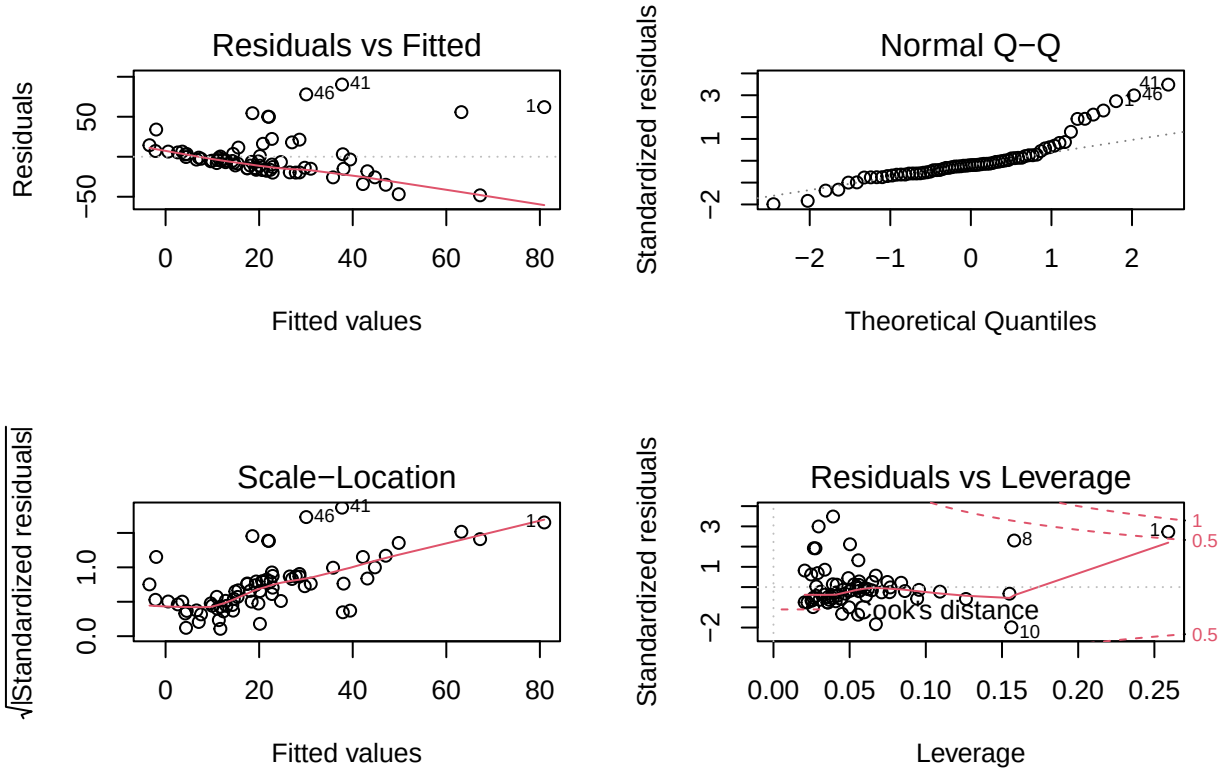
Q2

```
load("Homework 4 Data.Rdata")
attach(centralbank)
head(centralbank)
```

```
##              Inflation Legal Turnover Industrial
## Argentina          143  0.40      1.0          D
## Australia            8  0.36      0.2          I
## Austria              4  0.61      0.1          I
## Bahamas, The         6  0.41      0.2          D
## Barbados             7  0.38      0.1          D
## Belgium              5  0.17      0.2          I
```

(a)

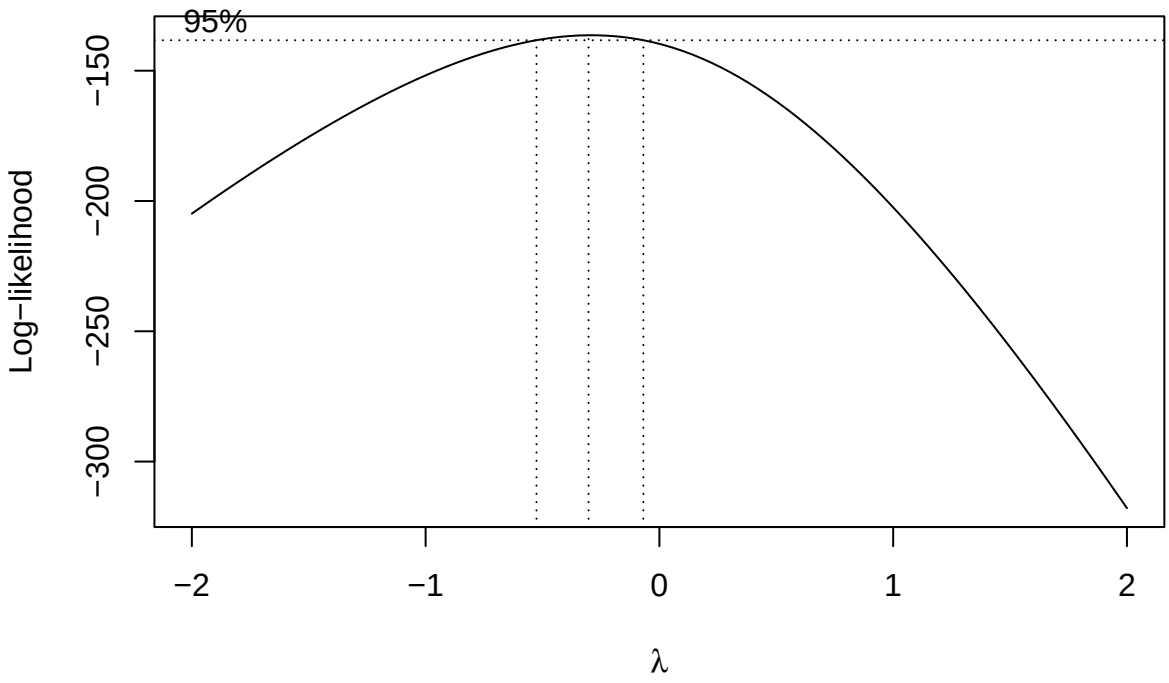
```
model.a <- lm(Inflation~Legal+Turnover+Industrial)
par(mfrow=c(2,2))
plot(model.a)
```



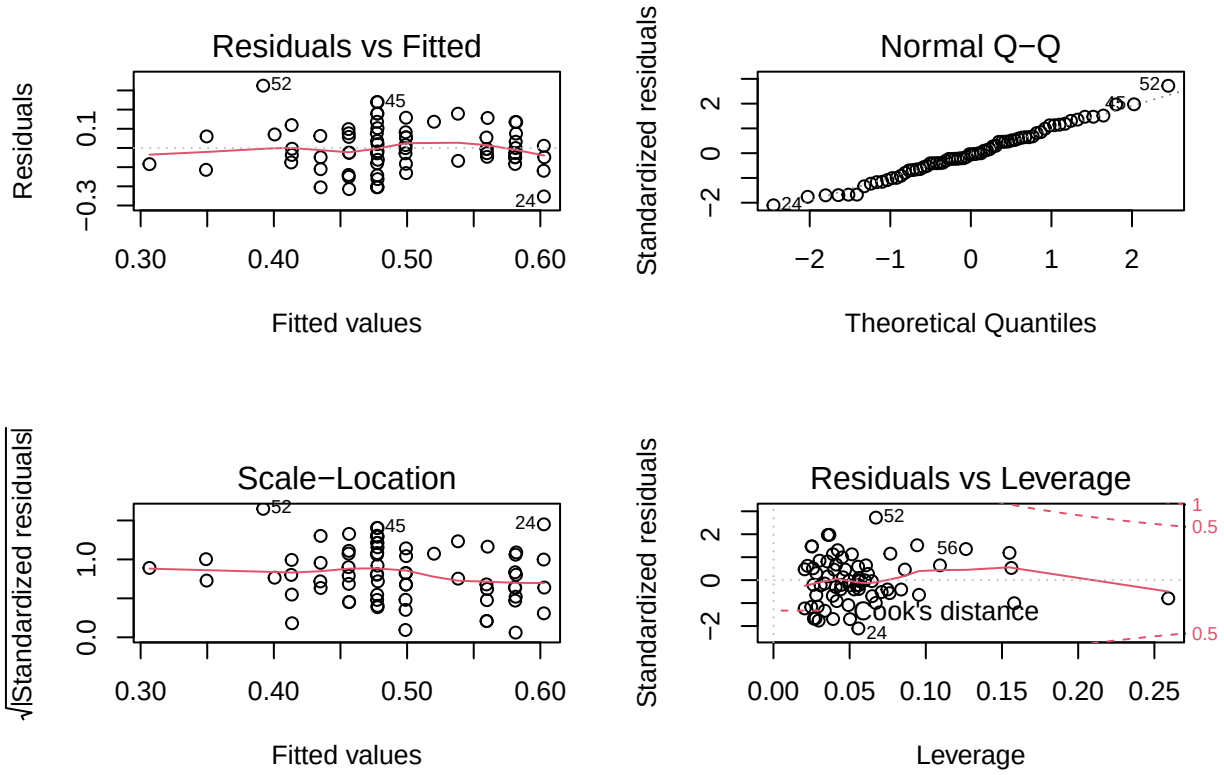
Checking for nonlinearity, the first plot shows a systematic nonlinear trend; Checking for nonnormality, the second plot shows the residuals are not normally distributed; Checking for nonconstant variance, the third plot shows a systematic trend due to the variability of residuals changing over Yi.

(b)

```
library(MASS)
L.max <- boxcox(Inflation~Legal+Turnover+Industrial, ylab="Log-likelihood")
```



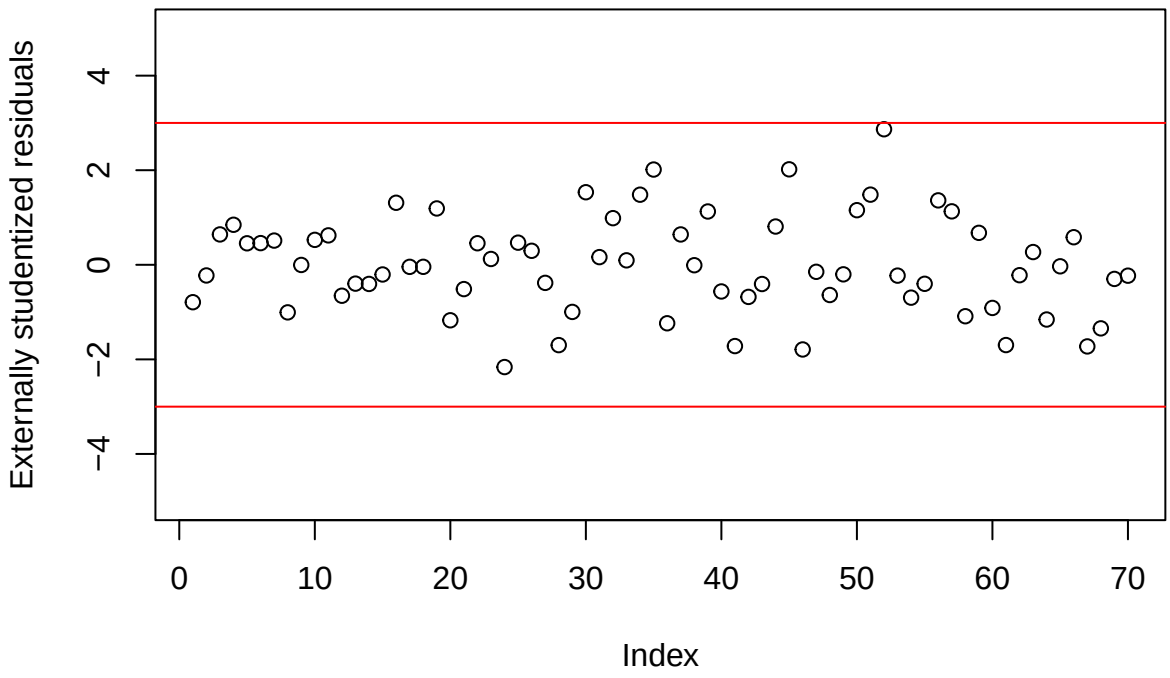

```
alpha <- L.max$x[which.max(L.max$y)]
Inflation.bc <- Inflation^alpha
model.bc <- lm(Inflation.bc~Legal+Turnover+Industrial)

par(mfrow=c(2,2))
plot(model.bc)
```



The residual plots now show that the residuals are more consistent with constant variance and normality. And there is no apparent nonlinearity.

(c)

```
minresid <- min(rstudent(model.bc))
maxresid <- max(rstudent(model.bc))
plot(rstudent(model.bc), ylim = c(min(-5, minresid),
     max(5, maxresid)), ylab="Externally studentized residuals")
abline(h=c(-3, 3), col="red")
```

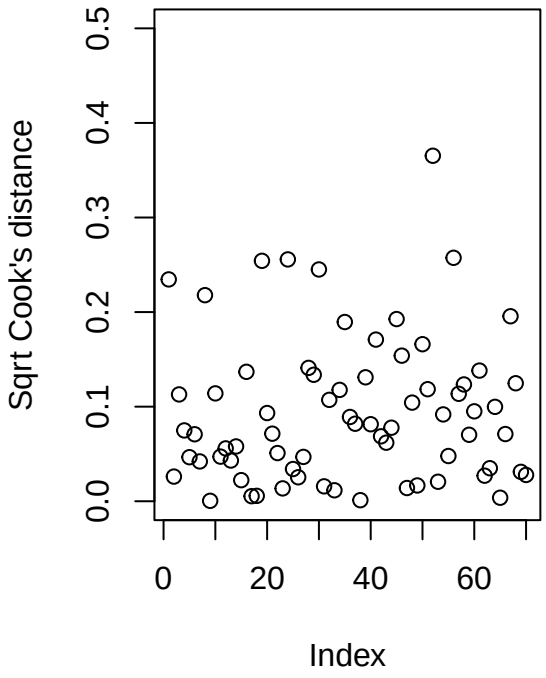


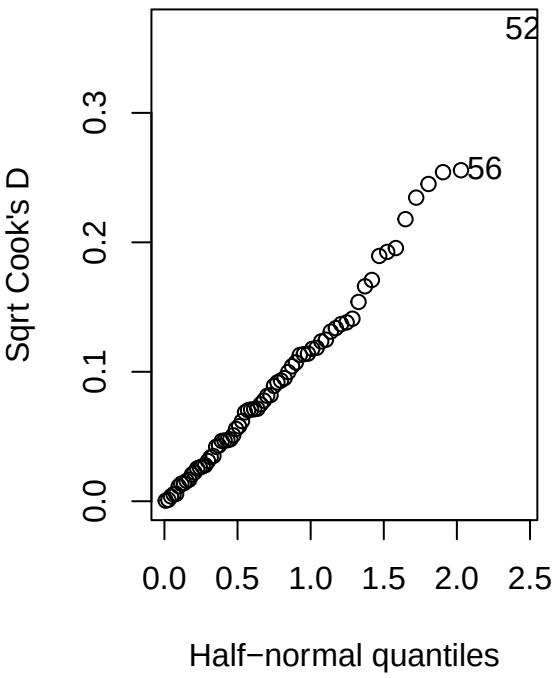

```
centralbank[rstudent(model.bc) > 2.8, ]
```

```
##           Inflation Legal Turnover Industrial
## Singapore         3  0.29      0.6          D
```

There is no data point with an externally studentized residual larger than 3. However, Singapore may be an outlier.

```
par(mfrow=c(1,2))
plot(sqrt(cooks.distance(model.bc)), ylim=c(0,0.5),
     ylab=("Sqrt Cook's distance"))

library(faraway)
halfnorm(sqrt(cooks.distance(model.bc)), ylab=("Sqrt Cook's D"))
```






```
centralbank[52, ]
```

```
##           Inflation Legal Turnover Industrial
## Singapore         3  0.29      0.6          D
```

Singapore is an influential point.

(d)

```
summary(model.bc)
```

```
##
## Call:
## lm(formula = Inflation.bc ~ Legal + Turnover + Industrial)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25279 -0.07890 -0.00521  0.07606  0.32483
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.51982    0.05057  10.279 2.48e-15 ***
## Legal        0.00186    0.12269   0.015   0.9879
```

```
## Turnover    -0.21392    0.08456  -2.530   0.0138 *
## IndustrialI  0.08220    0.03522   2.334   0.0227 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1236 on 66 degrees of freedom
## Multiple R-squared:  0.2287, Adjusted R-squared:  0.1937
## F-statistic: 6.525 on 3 and 66 DF,  p-value: 0.0006223
```

Legal is not statistically significant at a 5% significance level. The validity of the hypothesis test is strong.

(e)

If h is strictly increasing, then h^{-1} is also strictly increasing:

$$P(A < h(Y_0) < B) = P(h[h^{-1}(A)] < h(Y_0) < h[h^{-1}(B)] = P\big(h^{-1}(A) < Y_0 < h^{-1}(B)\big) = 0.90$$

If h is strictly decreasing, then h^{-1} is also strictly decreasing:

$$P(A < h(Y_0) < B) = P(h[h^{-1}(A)] < h(Y_0) < h[h^{-1}(B)] = P\big(h^{-1}(B) < Y_0 < h^{-1}(A)\big) = 0.90$$

```
CI <- predict(model.bc,
        newdata=centralbank[row.names(centralbank) == "United States", ],
        interval="prediction", level=0.90)^(1/alpha)
upr <- CI[2]
lwr <- CI[3]
CI[2] <- lwr; CI[3] <- upr
CI
```

```
##                    fit      lwr      upr
## United States 5.983207 2.140853 26.83252
```

Q3

```
attach(ibm)
head(ibm)
```

```
##    time duration trades
## 1 34370       96      7
## 2 34397       27      0
## 3 34418       21      1
## 4 34464       46      2
## 5 34491       27      1
## 6 34608      117      4
```

(a)

```r
y <- as.numeric(trades == 0)
x <- log(duration)

fit <- glm(y~x, family=binomial(link="logit"), data=data.frame(y, x))
summary(fit)
```

```
##
## Call:
## glm(formula = y ~ x, family = binomial(link = "logit"), data = data.frame(y,
##     x))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1049  -0.6270  -0.1826   0.6538   2.0677
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   6.8040     1.0154   6.701 2.07e-11 ***
## x            -1.7371     0.2483  -6.997 2.61e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 267.93  on 193  degrees of freedom
## Residual deviance: 157.53  on 192  degrees of freedom
## AIC: 161.53
##
## Number of Fisher Scoring iterations: 5
```

(b)

$$\hat{p}(x) = P(Y_i = 1|X_i = x) = H(\beta_0 + \beta_1 x) = \frac{1}{1 + e^{-\beta_0 - \beta_1 x}}, \qquad \beta_0 = 6.8040, \beta_1 = -1.7371$$

$$\begin{aligned} \hat{p}(x) &= 0.5 \\ e^{-\beta_0 - \beta_1 x} &= 1 \\ \beta_0 + \beta_1 x &= 0 \\ x &= -\frac{\beta_0}{\beta_1} \\ \log(d) &= -\frac{\beta_0}{\beta_1} \\ d &= e^{-\frac{\beta_0}{\beta_1}} \end{aligned}$$

```r
(d <- exp(-summary(fit)$coef[1,1]/summary(fit)$coef[2,1]))
```

```
## [1] 50.23812
```