Residual-Analysis.R

Shraddha Somani

```
require(faraway)
```

```
## Loading required package: faraway
```

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
require(lmtest)
```

```
## Loading required package: lmtest
```

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

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

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

```
require(car)
```

```
## Loading required package: car
```

```
##
## Attaching package: 'car'
```

```
## The following objects are masked from 'package:faraway':
##
##     logit, vif
```

```
require(gridExtra)
```

```
## Loading required package: gridExtra
```

```
library(scatterplot3d)
head(uswages)
```

```
##         wage educ exper race smsa ne mw so we pt
## 6085  771.60   18    18    0    1  1  0  0  0  0
## 23701 617.28   15    20    0    1  0  0  0  1  0
## 16208 957.83   16     9    0    1  0  0  1  0  0
## 2720  617.28   12    24    0    1  1  0  0  0  0
## 9723  902.18   14    12    0    1  0  1  0  0  0
## 22239 299.15   12    33    0    1  0  0  0  1  0
```

```
summary(uswages)
```

```
##       wage              educ           exper            race
##  Min.   :  50.39   Min.   : 0.00   Min.   :-2.00   Min.   :0.000
##  1st Qu.: 308.64   1st Qu.:12.00   1st Qu.: 8.00   1st Qu.:0.000
##  Median : 522.32   Median :12.00   Median :15.00   Median :0.000
##  Mean   : 608.12   Mean   :13.11   Mean   :18.41   Mean   :0.078
##  3rd Qu.: 783.48   3rd Qu.:16.00   3rd Qu.:27.00   3rd Qu.:0.000
##  Max.   :7716.05   Max.   :18.00   Max.   :59.00   Max.   :1.000
##       smsa             ne              mw              so
##  Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.000   Median :0.000   Median :0.0000   Median :0.0000
##  Mean   :0.756   Mean   :0.229   Mean   :0.2485   Mean   :0.3125
##  3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :1.000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000
##        we              pt
##  Min.   :0.00   Min.   :0.0000
##  1st Qu.:0.00   1st Qu.:0.0000
##  Median :0.00   Median :0.0000
##  Mean   :0.21   Mean   :0.0925
##  3rd Qu.:0.00   3rd Qu.:0.0000
##  Max.   :1.00   Max.   :1.0000
```

```
# Manipulating data. We see that exper has negative values
uswages$exper[uswages$exper <0] = NA
# Convert race, smsa, and pt to factor variables
uswages$race = factor(uswages$race)
levels(uswages$race) = c("White","Black")
uswages$smsa = factor(uswages$smsa)
levels(uswages$smsa) = c("No","Yes")
uswages$pt = factor(uswages$pt)
levels(uswages$pt) = c("No","Yes")
# Create region, a factor variable based on the four regions ne, mw, so, we
uswages = data.frame(uswages,
                     region =
                       1*uswages$ne +
                       2*uswages$mw +
                       3*uswages$so +
                       4*uswages$we)
uswages$region = factor(uswages$region)
levels(uswages$region) = c("ne","mw","so","we")
# Delete the four regions ne, mw, so, we
uswages = subset(uswages,select=-c(ne:we))
# Take care of NAs
uswages = na.omit(uswages)
summary(uswages)
```

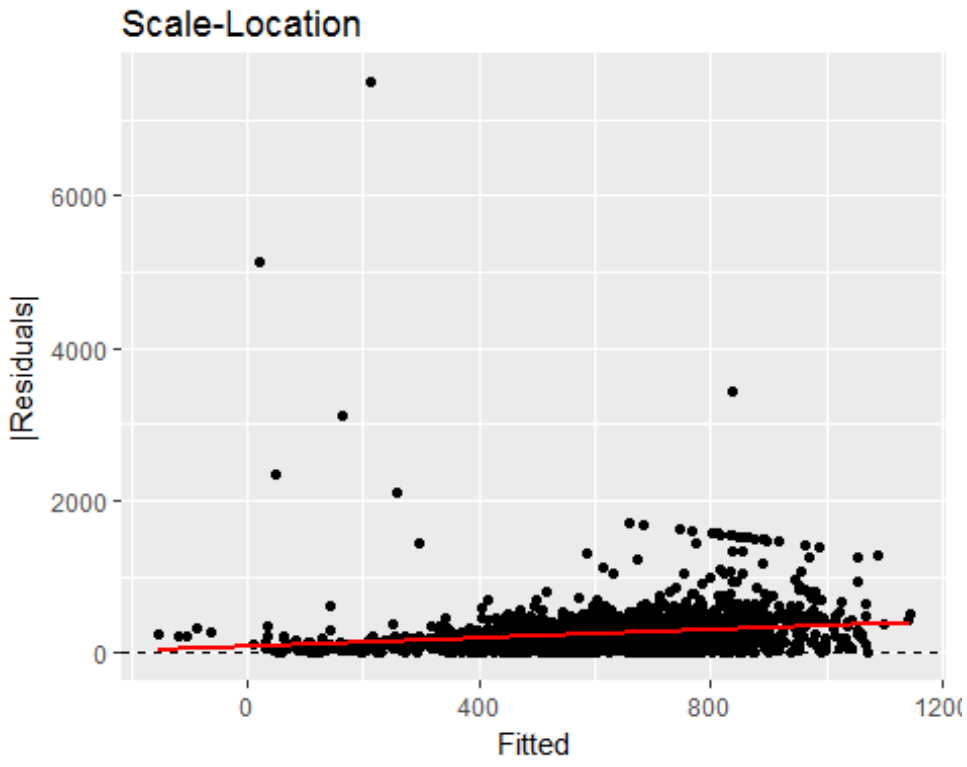
```
##       wage              educ           exper          race       smsa
##  Min.   :  50.39   Min.   : 0.00   Min.   : 0.00   White:1812   No : 483
```

```
##  1st Qu.: 314.69   1st Qu.:12.00   1st Qu.: 8.00   Black: 155   Yes:1484
##  Median : 522.32   Median :12.00   Median :16.00
##  Mean   : 613.99   Mean   :13.08   Mean   :18.74
##  3rd Qu.: 783.48   3rd Qu.:16.00   3rd Qu.:27.00
##  Max.   :7716.05   Max.   :18.00   Max.   :59.00
##    pt       region
##  No :1802   ne:448
##  Yes: 165   mw:488
##             so:616
##             we:415
##
##
```

```r
# Exercise 1 - Nonconstance variance
# 1(a) Using the uswage data, fit the model (m): wage ~ educ + exper + race + smsa + pt + region
m = lm(wage ~ educ + exper + race + smsa + pt + region, data = uswages)

# 1(b) Produce the Residuals vs Fitted plot, and discuss if there may be heteroskedasticiy in the error variance.
mod = fortify(m)
ggplot(mod, aes(.fitted, .resid)) + geom_point() + geom_hline(yintercept=0, color="red", linetype="dashed") + ggtitle("Residual Plot vs Fitted Plot") + geom_smooth(color = "red", se = F)
```

```
## `geom_smooth()` using method = 'gam'
```

```
# Answer:
# - The red line is slightly curved and the residuals seem to increase as the
fitted Y values increase.
# - So, the inference here is, heteroscedasticity exists.

# Statistical Heteroskedasticiy test
bptest(mod)

##
##  studentized Breusch-Pagan test
##
## data:  mod
## BP = 29.642, df = 8, p-value = 0.0002445

# Answer:
# - The test have a p-value less that a significance level of 0.05.
# - Therefore we can reject the null hypothesis that the variance of the
residuals is constant and infer that heteroskedasticiy is indeed present.

# 1(c) Produce the Scale-Location plot, and discuss if there may be
heteroskedasticiy in the error variance.
qplot(.fitted, abs(.resid), data = mod) + geom_hline(yintercept = 0, linetype
= "dashed") + labs(title = "Scale-Location", x = "Fitted", y = "|Residuals|")
+ geom_smooth(method = "gam", color = "red", se = F)
```



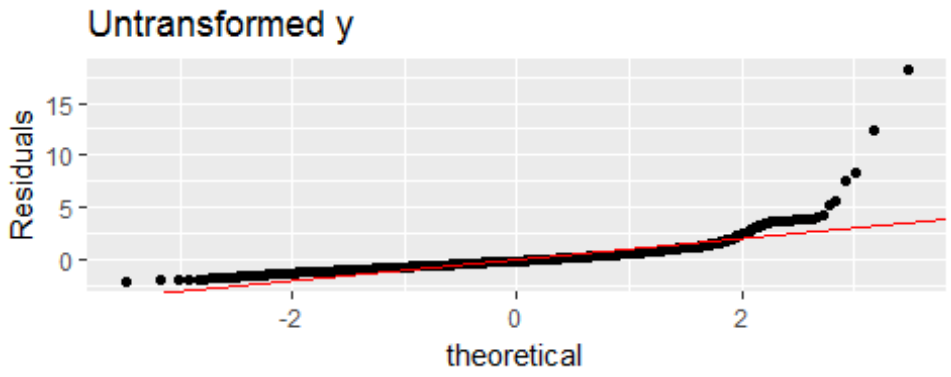

```
# Answer:
# - Heteroskedasticiy is not present, if the red line is a straight line.
```

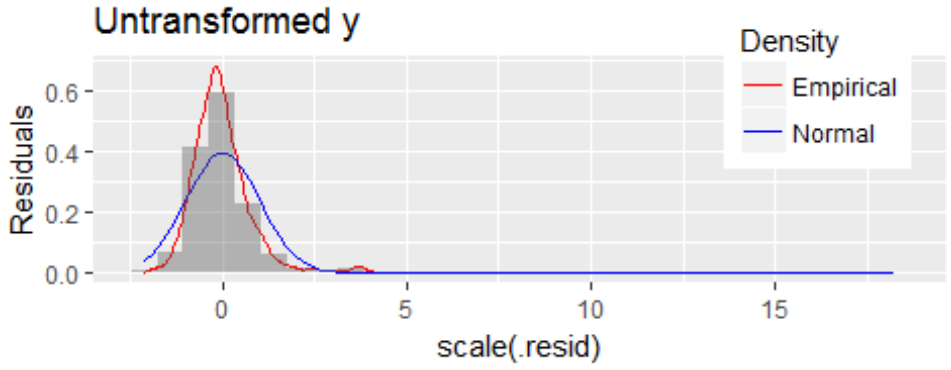
```r
# - But in our case, the red line is not a straight line so the inference
here is, heteroscedasticity exists.

# 1(d) Perform the approximate test of noncontant error variance.
summary(lm(abs(residuals(m)) ~ fitted(m)))
```

```
##
## Call:
## lm(formula = abs(residuals(m)) ~ fitted(m))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -370.9 -152.6  -51.9   71.8 7367.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 80.43904   23.36854   3.442 0.000589 ***
## fitted(m)    0.27303    0.03615   7.552 6.53e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 323.9 on 1965 degrees of freedom
## Multiple R-squared:  0.02821,    Adjusted R-squared:  0.02771
## F-statistic: 57.03 on 1 and 1965 DF,  p-value: 6.528e-14
```

```r
# Answer:
# - We look at the t-test for the slope coefficient with null hypothesis that
the slope is zero.
# - At the 10% level of significance, we conclude that the slope is not zero
since the p-value, 6.528e-14, is less than 0.10
# - Therefore, we conclude that there is nonconstant error variance.


#Exercise 2 - Non-normal errors
#2(a) Plot the Normal QQ Plot and Histogram of the residuals from model m
Exercise 1. Do they indicate non-normal errors?
p1 = qplot(sample = scale(.resid), data = mod) + geom_abline(intercept = 0,
slope = 1, color = "red") + labs(title = "Untransformed y", y = "Residuals")
p2 = qplot(scale(.resid), data = mod, geom = "blank") + geom_line(aes(y =
..density.., colour = "Empirical"), stat = "density") + stat_function(fun =
dnorm, aes(colour = "Normal")) + geom_histogram(aes(y = ..density..), alpha =
0.4) + scale_colour_manual(name = "Density", values = c("red", "blue")) +
theme(legend.position = c(0.85, 0.85)) + labs(title = "Untransformed y", y =
"Residuals")
grid.arrange(p1, p2, nrow = 2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

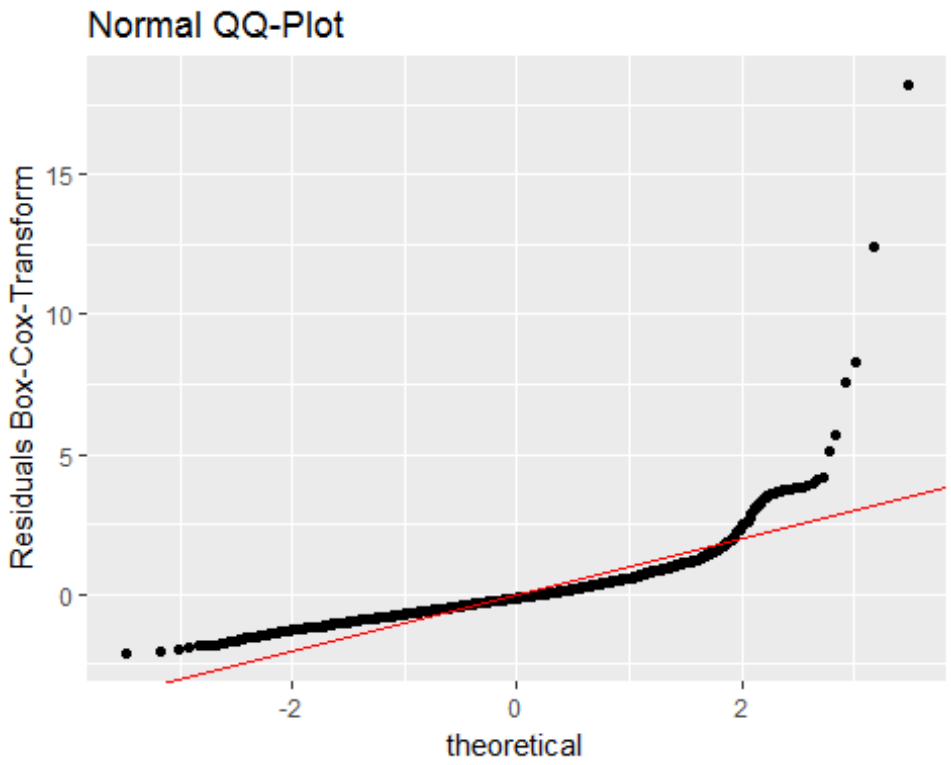





```
#Answer:
#Clearly the residuals of the model indicate non-normal error.

#2(b) Perfrom the Shapiro-Wilk test of normality for the residuals of model
m. What is the P-value and what does it say about normality?
shapiro.test(residuals(m))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(m)
## W = 0.71236, p-value < 2.2e-16

#Answer
#The null hypothesis is that the the residuals are normal.
#Since the p-value is smaller than the significant value (0.05), we reject
the null hypothesis.
#The residuals are not normal.

#2(c) Find the optimal Box-Cox power transform and apply it to wage, refit
model m, replot Normal Q-Q Plot and perform the Shapiro-Wilk test of
normality again. Did the Box-Cox Power Transform work?
lambda = powerTransform(m)
lambda
```

```
## Estimated transformation parameters
##        Y1
## 0.1034019
```

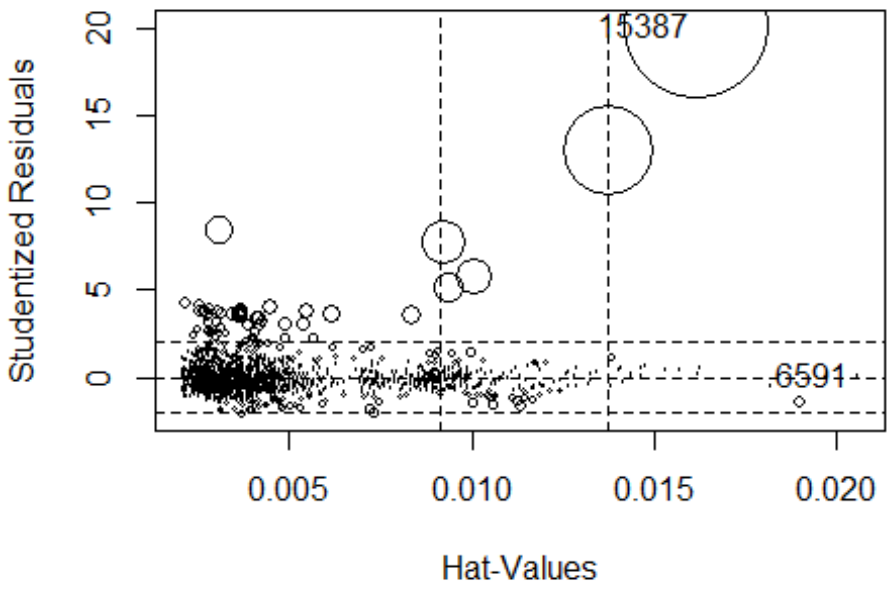
```
lam = lambda$lambda
mlam = lm(wage ~ educ + exper + race + smsa + pt + region, data = uswages)
modlam <- fortify(mlam)
qplot(sample = scale(.resid), data = modlam) + geom_abline(intercept = 0,
slope = 1, color = "red") + labs(title = "Normal QQ-Plot", y = "Residuals
Box-Cox-Transform")
```



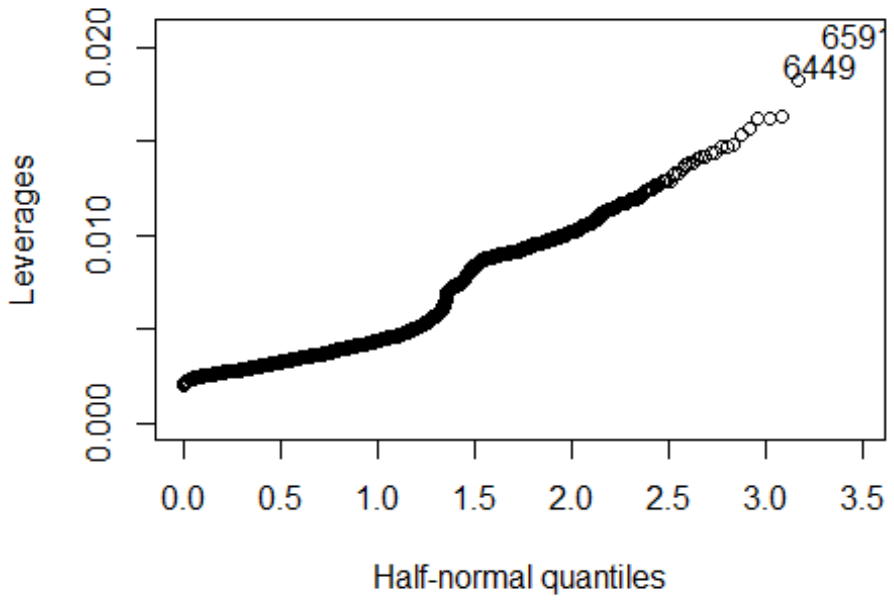

```
shapiro.test(residuals(mlam))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(mlam)
## W = 0.71236, p-value < 2.2e-16
```

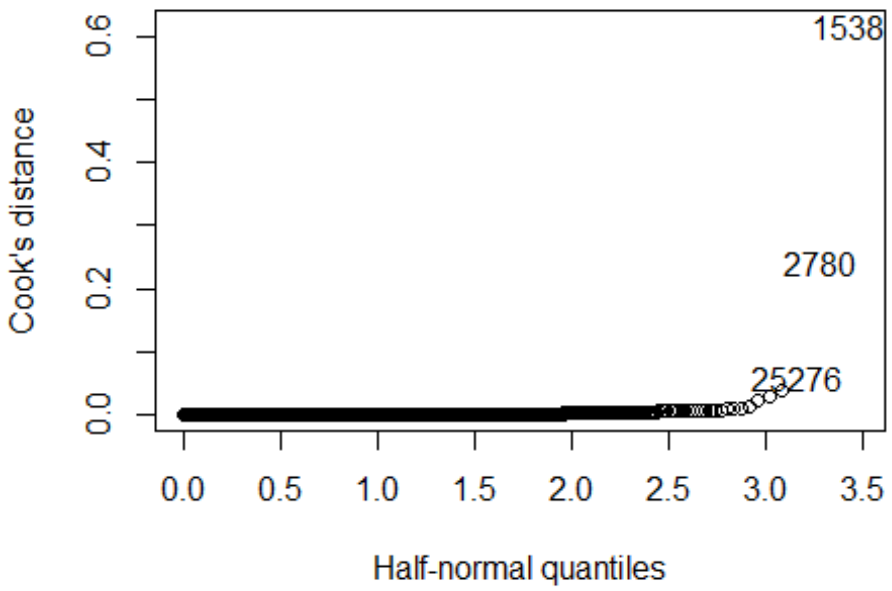
```
#Answer:
#The Box-Cox Power Transform did not work for our model.


#Exercise 3 -  Influential outliers
#3(a) Produce the influence plot for model m. Are there any really large
CookD values?
influencePlot(m)
```




```
##           StudRes        Hat        CookD
## 6591   0.1096679 0.02061630 0.0000281445
## 15387 20.1499155 0.01621022 0.6159365616

#Answer:
#There are large cookD values.

#3(b) Produce the half-normal plot of the leverage values. Are they any high
leverage data points?
islands <- row.names(uswages)
halfnorm(lm.influence(mlam)$hat, labs = islands, ylab = "Leverages")
```

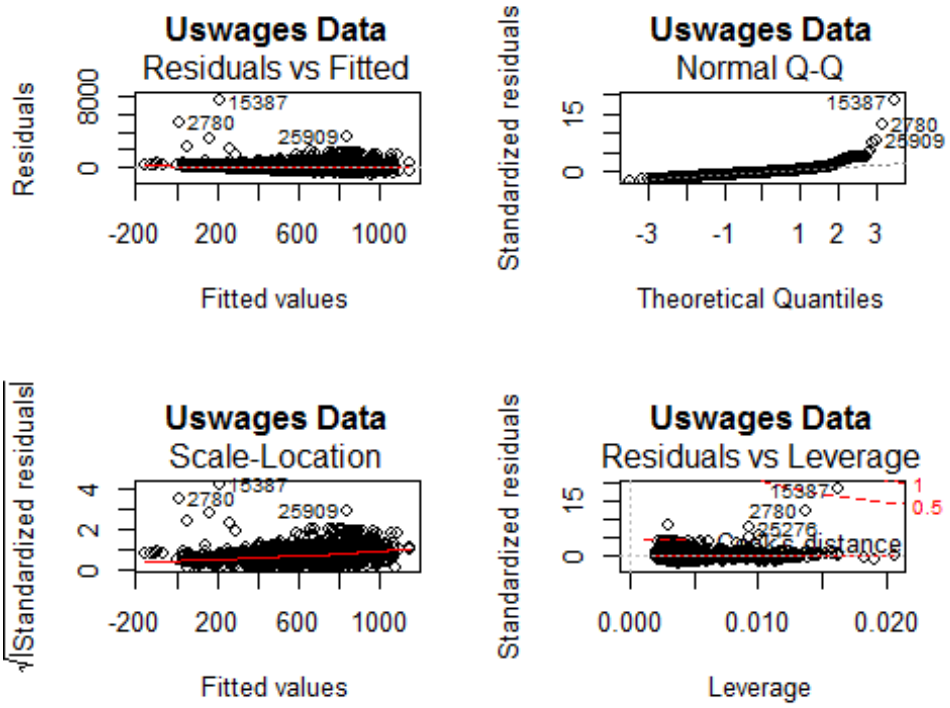



```
#Answer:
#Yes, there are high leverage data points

#3(c) Produce the half-normal plot of the Cook's distance. Are they any high
Cook's distance points?
cook <- cooks.distance(mlam)
halfnorm(cook, 3, labs = islands, ylab = "Cook's distance")
```

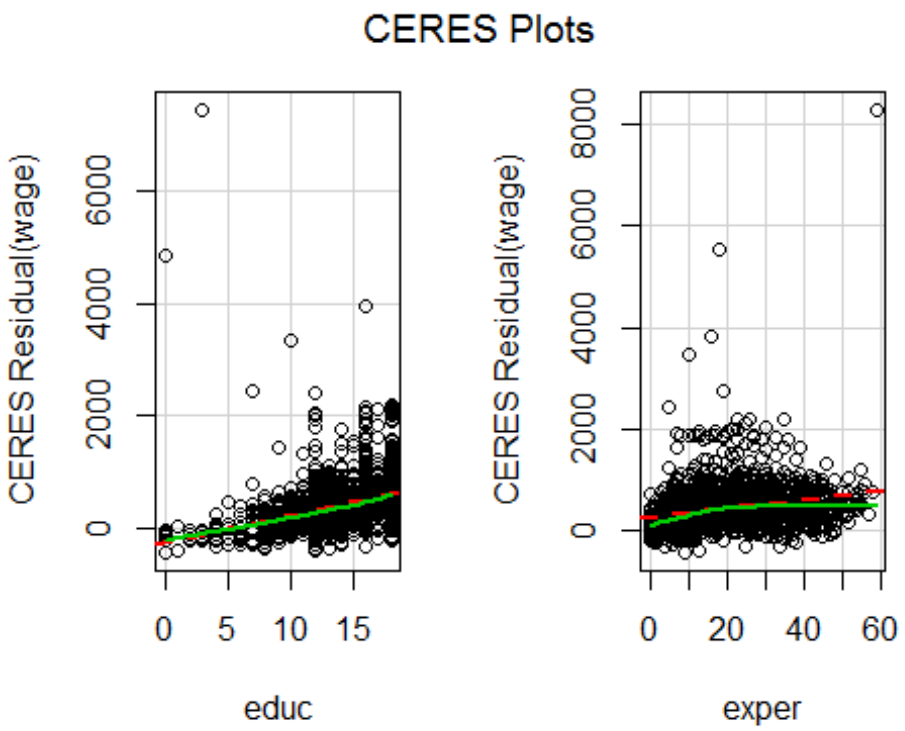



```
#Answer:
#Yes, there are high Cook's distance points

#3(d) Fit model excluding observation with largest Cook's Distance. Do the
coeficients change? Are there any coeficients with notable changes?
mlam1 = lm(wage ~ educ + exper + race + smsa + pt + region, data = uswages,
subset = (cook < max(cook)))
compareCoefs(mlam, mlam1)

##
## Call:
## 1: lm(formula = wage ~ educ + exper + race + smsa + pt + region, data =
##    uswages)
## 2: lm(formula = wage ~ educ + exper + race + smsa + pt + region, data =
##    uswages, subset = (cook < max(cook)))
##                Est. 1     SE 1   Est. 2     SE 2
## (Intercept) -259.070   55.176 -282.520   50.239
## educ          49.299    3.261   52.843    2.973
## exper          8.966    0.737    8.280    0.672
## raceBlack   -121.887   35.349 -107.065   32.186
## smsaYes      116.570   21.894  107.917   19.934
## ptYes       -326.294   33.623 -374.813   30.701
## regionmw      -6.541   27.199   -7.346   24.758
## regionso       2.804   26.079  -10.362   23.748
## regionwe      47.802   28.238   47.035   25.705
```

```
#Answer:
#Yes, the coefficients change.
#There are no notable changes.

#3(e) Produce the omnibus diagnotic plot for model m. Which observation
consistantly stands out as an outlier-influential point in all four plots?
oldpar = par(mfrow = c(2, 2))
plot(mlam, main = "Uswages Data")
```




```
#Answer:
#Observations that consistently stands out as an outlier-influential point in
all four plots are - 2780, 15387 and 25909


#Exercise 4 - Model structure
#4(a) Produce the CERES plots for model m. Do the factor varibles stop the
plots from printing?
ceresPlots(m, terms = ~.)

## Warning in ceresPlots(m, terms = ~.): Factors skipped in drawing CERES
## plots.
```




```
#Answer:
#Yes, the factor variables stop the plots from printing.

#4(b) How many plots are there? Why these?
#Answer:
#There are two plots - Educ and Exper. Only these two plots are plotted
because the rest of the variable are converted into factor variables.

#4(c) Do the plots indicate a polynomial model should be considered?
#Answer:
#Yes, a polynomial model should be considered.


#Exercise 5 - Interaction model
#5(a) Fit an interaction model using the region and the two numeric
variables. Is the model useful?
uswages$dummy = factor(uswages$exper < 18)
summary(uswages)

##       wage              educ           exper          race       smsa
##  Min.   :  50.39   Min.   : 0.00   Min.   : 0.00   White:1812   No : 483
##  1st Qu.: 314.69   1st Qu.:12.00   1st Qu.: 8.00   Black: 155   Yes:1484
##  Median : 522.32   Median :12.00   Median :16.00
##  Mean   : 613.99   Mean   :13.08   Mean   :18.74
##  3rd Qu.: 783.48   3rd Qu.:16.00   3rd Qu.:27.00
##  Max.   :7716.05   Max.   :18.00   Max.   :59.00
```

```
##    pt       region     dummy
##  No :1802   ne:448   FALSE: 893
##  Yes: 165   mw:488   TRUE :1074
##             so:616
##             we:415
##
##
```

```
m_interaction = lm(wage ~ educ + exper * dummy + race + smsa + pt + region,
data = uswages)
m_1 = lm(wage ~ exper + region + educ, data = uswages)
anova(m_1, m_interaction)
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ exper + region + educ
## Model 2: wage ~ educ + exper * dummy + race + smsa + pt + region
##   Res.Df       RSS Df Sum of Sq      F    Pr(>F)
## 1   1961 356452814
## 2   1956 319107031  5  37345783 45.783 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Answer:
#The model is useful as p value is less than 0.05

#5(b) Test the interaction model versus model m. What is the p-value and
which model does it indicate?
anova(m, m_interaction)
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ educ + exper + race + smsa + pt + region
## Model 2: wage ~ educ + exper * dummy + race + smsa + pt + region
##   Res.Df       RSS Df Sum of Sq      F    Pr(>F)
## 1   1958 333222495
## 2   1956 319107031  2  14115464 43.261 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Answer:
#P-value = 2.2e-16
#It indicates model 2


#Exercise 6 - Collinearity
#6(a) Find the variance inflation factors for model m.
vif(m)
```

```
##            GVIF Df GVIF^(1/(2*Df))
## educ   1.114749  1        1.055817
```

```
## exper  1.100628  1        1.049108
## race   1.048380  1        1.023904
## smsa   1.026374  1        1.013101
## pt     1.004126  1        1.002061
## region 1.061190  3        1.009948

#6(b) Do they indicate collinerairty in the predictors?
a.df = data.frame(uswages)
b = subset(a.df, select = c(wage, educ, exper))
summary(b)

##       wage              educ           exper
##  Min.   :  50.39   Min.   : 0.00   Min.   : 0.00
##  1st Qu.: 314.69   1st Qu.:12.00   1st Qu.: 8.00
##  Median : 522.32   Median :12.00   Median :16.00
##  Mean   : 613.99   Mean   :13.08   Mean   :18.74
##  3rd Qu.: 783.48   3rd Qu.:16.00   3rd Qu.:27.00
##  Max.   :7716.05   Max.   :18.00   Max.   :59.00

round(cor(b),1)

##       wage educ exper
## wage   1.0  0.3   0.2
## educ   0.3  1.0  -0.3
## exper  0.2 -0.3   1.0
```