STAT5703 HW3 Ex3

Chao Huang (ch3474), Wancheng Chen (wc2687), Chengchao Jin (cj2628)

##Exercise 3.

Question 1.

- We use linear model to use sex, age, race, marr, occupation, sector, south and union as parameters to predict wage. As a person's age get older, his number of years of education and number of work experience always gets larger. It means these three variables are correlated to each other. So if we include age, education and experience at the same time, it will lead to colinearity.

QUestion 2.

```
library(readr)
library(MASS)
CpsWages <- read_table2("CpsWages.txt")[,c(-1,-4)]
```

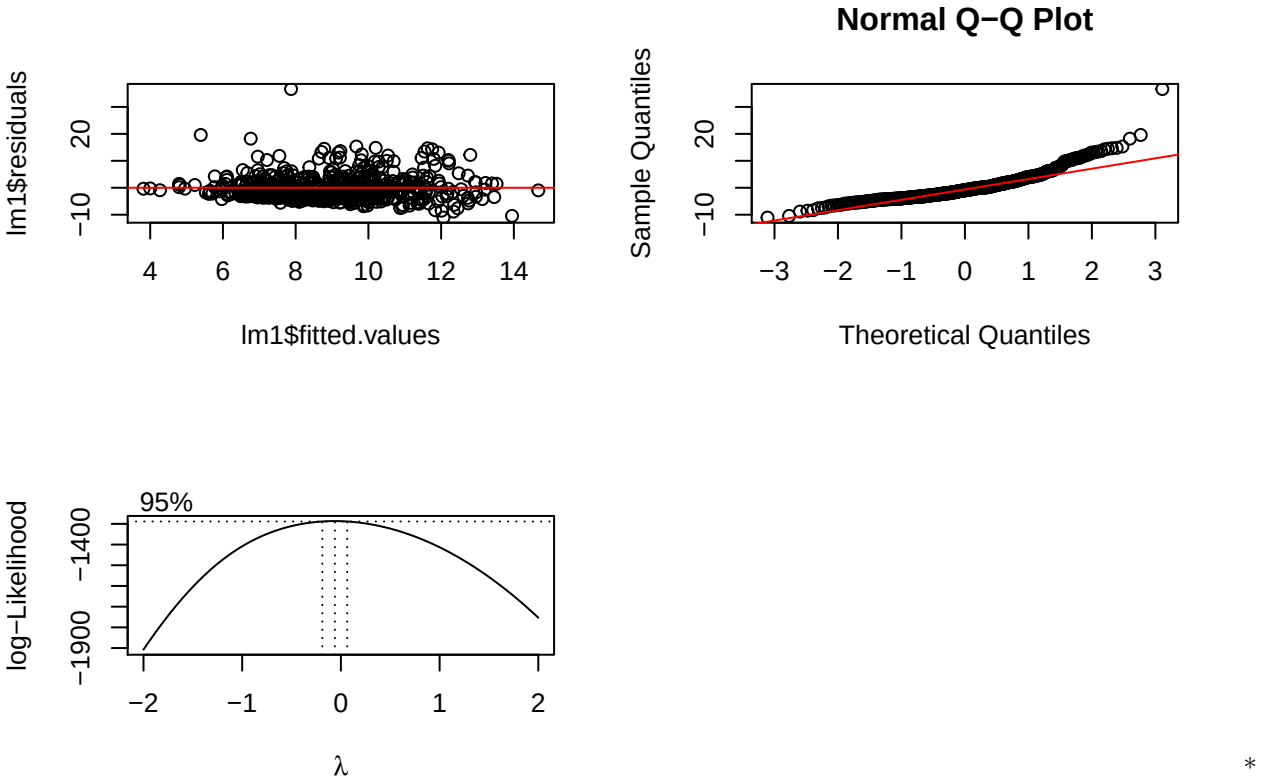
```
## Parsed with column specification:
## cols(
##   education = col_double(),
##   south = col_double(),
##   sex = col_double(),
##   experience = col_double(),
##   union = col_double(),
##   wage = col_double(),
##   age = col_double(),
##   race = col_double(),
##   occupation = col_double(),
##   sector = col_double(),
##   marr = col_double()
## )
```

```
#CpsWages<-data.frame(CpsWages[,c(4,5)],sapply(CpsWages[,c(1,2,3,6,7,8,9)],function(x) as.factor(x)))
lm1<-lm(CpsWages$wage~CpsWages$age+CpsWages$south+CpsWages$sex+CpsWages$union+CpsWages$race+CpsWages$oc
summary(lm1)
```

```
##
## Call:
## lm(formula = CpsWages$wage ~ CpsWages$age + CpsWages$south +
##     CpsWages$sex + CpsWages$union + CpsWages$race + CpsWages$occupation +
##     CpsWages$sector + CpsWages$marr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.065  -3.222  -0.931   1.970  36.626
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           6.94163    1.27517   5.444 8.02e-08 ***
## CpsWages$age          0.06699    0.01890   3.544 0.000429 ***
## CpsWages$south       -1.30714    0.46527  -2.809 0.005148 **
## CpsWages$sex         -2.33191    0.43661  -5.341 1.38e-07 ***
```

```
## CpsWages$union       1.72837    0.57516   3.005 0.002782 **
## CpsWages$race        0.75057    0.31166   2.408 0.016370 *
## CpsWages$occupation -0.39447    0.14134  -2.791 0.005448 **
## CpsWages$sector      0.25381    0.42207   0.601 0.547869
## CpsWages$marr        0.45287    0.45957   0.985 0.324872
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.825 on 525 degrees of freedom
## Multiple R-squared:  0.1318, Adjusted R-squared:  0.1186
## F-statistic: 9.966 on 8 and 525 DF,  p-value: 5.824e-13
```

```
par(mfrow=c(2,2))
plot(lm1$residuals~lm1$fitted.values)
abline(h=0,col="red")
qqnorm(lm1$residuals)
qqline(lm1$residuals,col="red")
boxcox(lm1)
```



From the residual plot, we can see that the dots are not randomly scattered around 0 and there is a pattern of hetroscedasticity. From the QQ-plot, we can see that there are many dots away from the qqline. So there is departure from the normal hypothesis in this model.

Question 3.

- According to p-value, not all parameters are statistically significant. We use t-test to test whether 'sector' is significant or not. From the table, we get the p-value from t-test for 'sector' is 0.547869 which is much larger than 0.05. So we fail to reject null hypothesis and conclude that the parameter 'sector' is not significant and should be eliminated from the model.

Question 4.

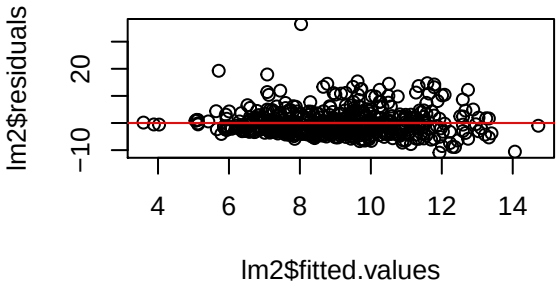
- From the summary table, we can see that the parameters 'sector' and 'marr' are not significant as their p-values are both larger than 0.05. So we use the leftover parameters to create a simplified model

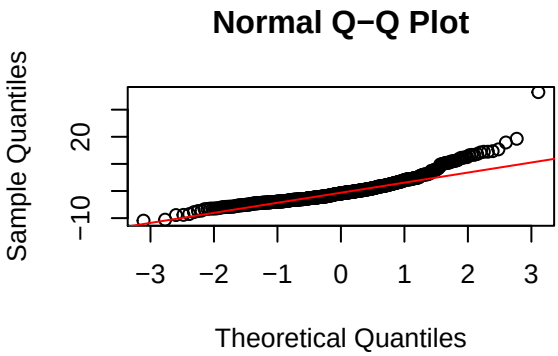
Question 5.

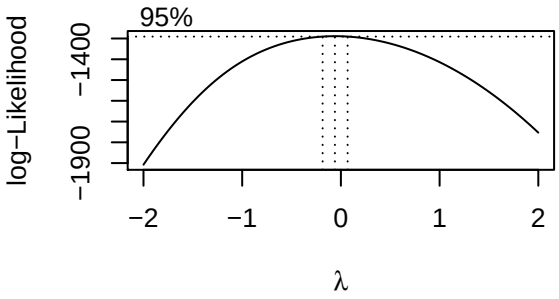
```
lm2<-lm(CpsWages$wage~CpsWages$age+CpsWages$south+CpsWages$sex+CpsWages$union+CpsWages$race+CpsWages$oc
summary(lm2)
```

```
##
## Call:
## lm(formula = CpsWages$wage ~ CpsWages$age + CpsWages$south +
##     CpsWages$sex + CpsWages$union + CpsWages$race + CpsWages$occupation)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.933  -3.133  -0.894   1.877  36.467
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            6.92054    1.26520   5.470 6.97e-08 ***
## CpsWages$age           0.07326    0.01810   4.049 5.93e-05 ***
## CpsWages$south        -1.29450    0.46478  -2.785  0.00554 **
## CpsWages$sex          -2.36143    0.43378  -5.444 8.00e-08 ***
## CpsWages$union         1.76223    0.57335   3.074  0.00222 **
## CpsWages$race          0.76740    0.31098   2.468  0.01392 *
## CpsWages$occupation -0.36670    0.13269  -2.764  0.00591 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.822 on 527 degrees of freedom
## Multiple R-squared:  0.1296, Adjusted R-squared:  0.1196
## F-statistic: 13.07 on 6 and 527 DF,  p-value: 8.203e-14
```

```
par(mfrow=c(2,2))
plot(lm2$residuals~lm2$fitted.values)
abline(h=0,col="red")
qqnorm(lm2$residuals)
qqline(lm2$residuals,col="red")
boxcox(lm2)
```





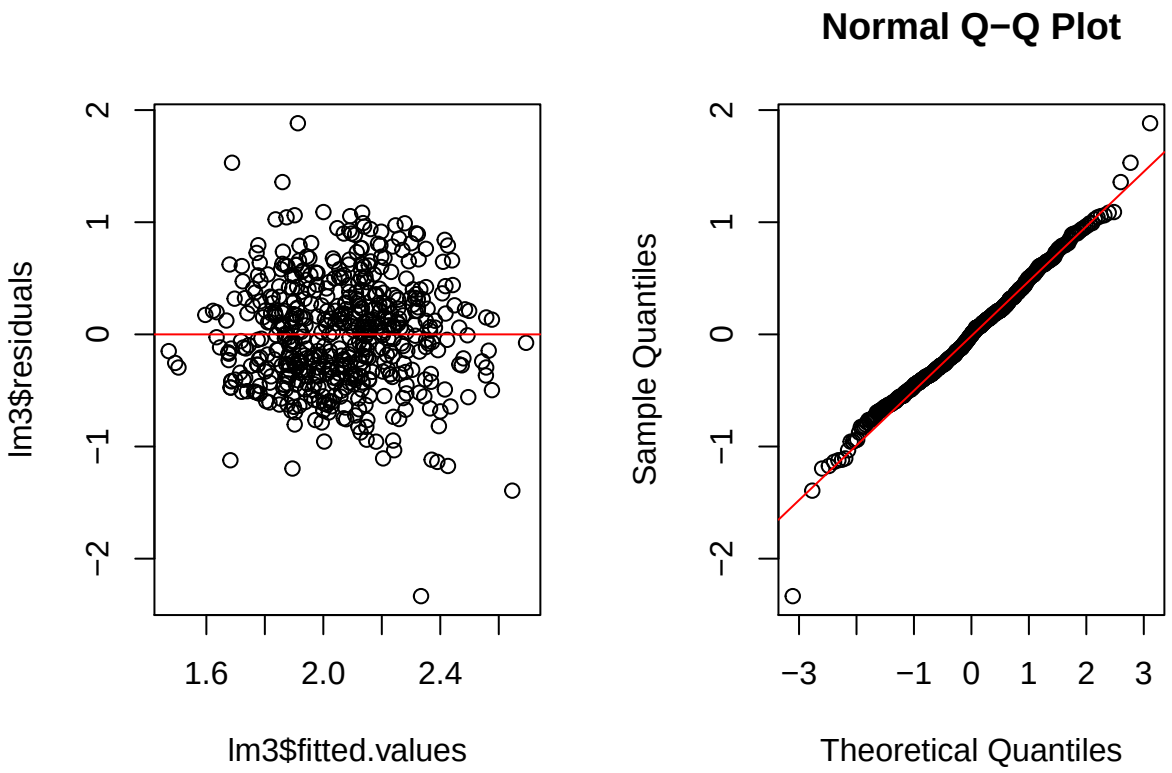


From box-cox plot, we can see that λ=0 lies inside the confidence interval. So we need to transform y to log(y) as response variable. As a result, we get the following new model:

```
lm3<-lm(log(CpsWages$wage)~CpsWages$age+CpsWages$south+CpsWages$sex+CpsWages$union+CpsWages$race+CpsWag
summary(lm3)
```

```
##
## Call:
## lm(formula = log(CpsWages$wage) ~ CpsWages$age + CpsWages$south +
##     CpsWages$sex + CpsWages$union + CpsWages$race + CpsWages$occupation)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3341 -0.3431 -0.0056  0.3153  1.8826
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.765407   0.127797  13.814  < 2e-16 ***
## CpsWages$age          0.008356   0.001828   4.572 6.04e-06 ***
## CpsWages$south       -0.164129   0.046947  -3.496 0.000512 ***
## CpsWages$sex         -0.245740   0.043816  -5.608 3.30e-08 ***
## CpsWages$union        0.228289   0.057913   3.942 9.17e-05 ***
## CpsWages$race         0.081842   0.031412   2.605 0.009434 **
## CpsWages$occupation -0.027776   0.013403  -2.072 0.038710 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4871 on 527 degrees of freedom
## Multiple R-squared:  0.1578, Adjusted R-squared:  0.1483
## F-statistic: 16.46 on 6 and 527 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(1,2))
plot(lm3$residuals~lm3$fitted.values)
abline(h=0,col="red")
qqnorm(lm3$residuals)
qqline(lm3$residuals,col="red")
```



Apparently, the residuals are now randomly scattered around 0 and the dots in qq-plot align with qq-line. So the normality assumption is met in this model. Also, from the summary table, all of parameters are statistically significant as all of their p-values are smaller than 0.05. So this simplied model is adequate.

Question 6.

No, deleting only two observations in a large data set will not significantly influence the conclusion.