Q3

Question 1

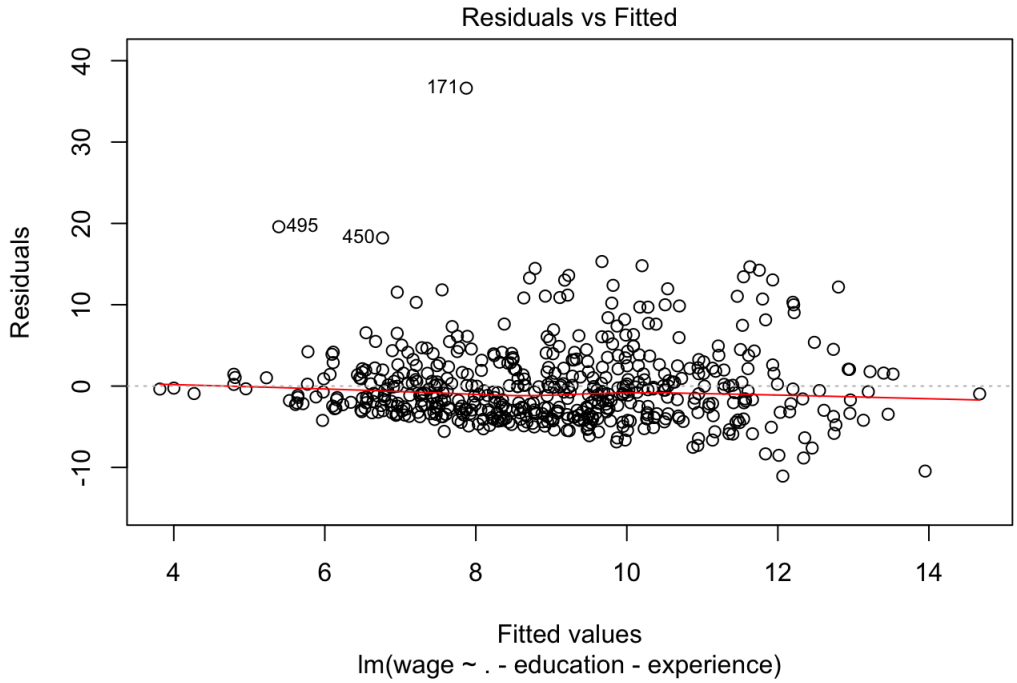
We could use linear model to fit this dataset. In this dataset, age, education and experience are all time related features. They all have strict linear relation with year. In other words they are highly correlated with each other. If we include them in the same time the model would become colinearity.

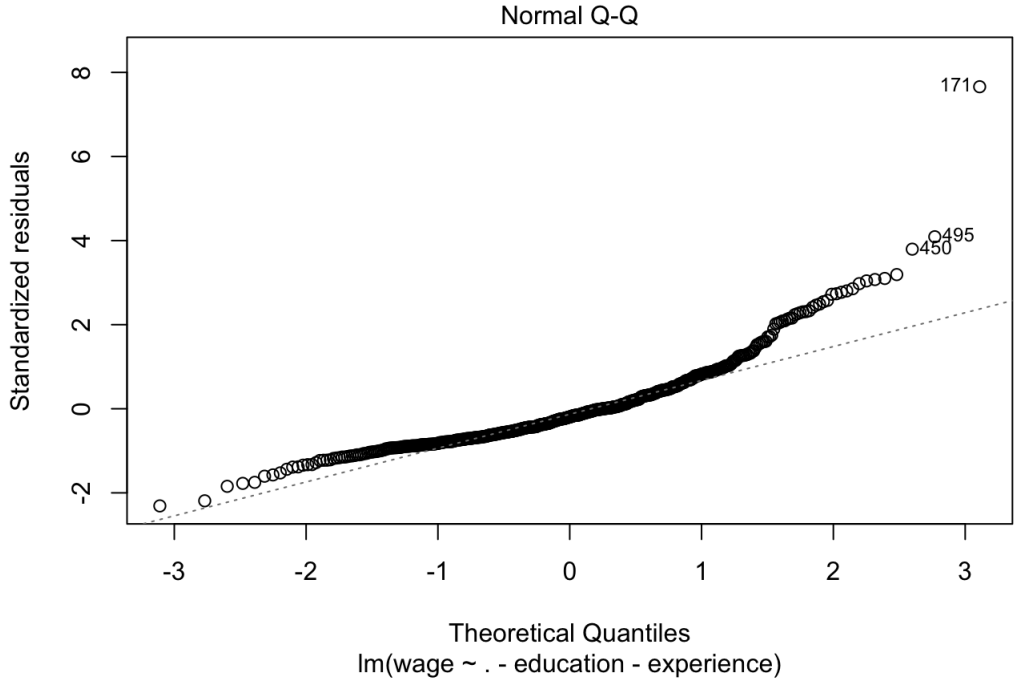
Question2

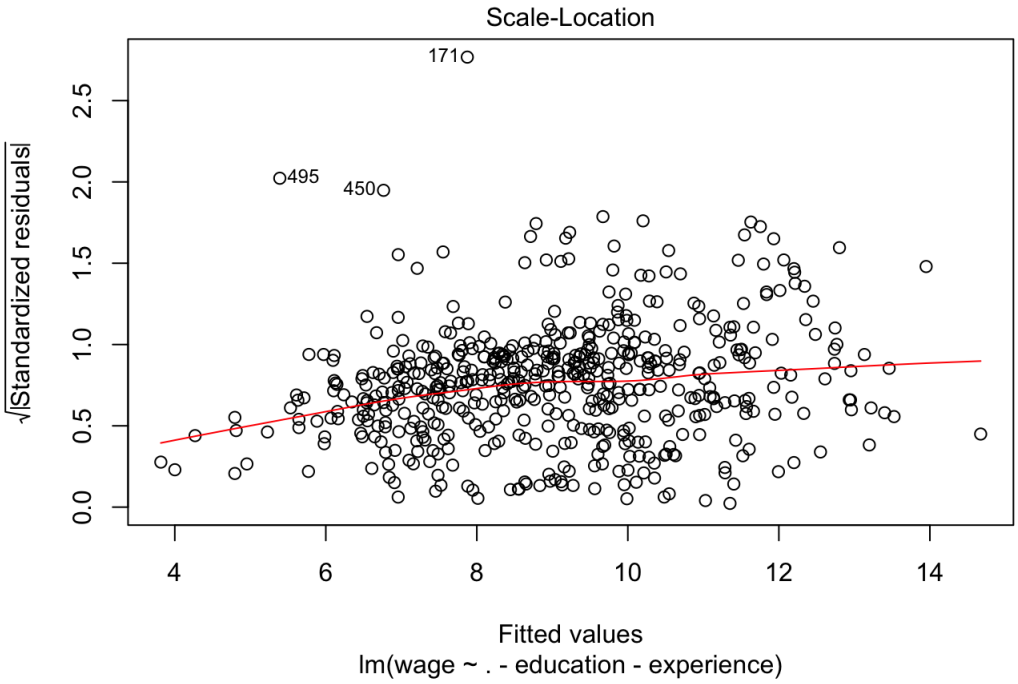
```
df<- read.table("CpsWages.txt",header = T)
```

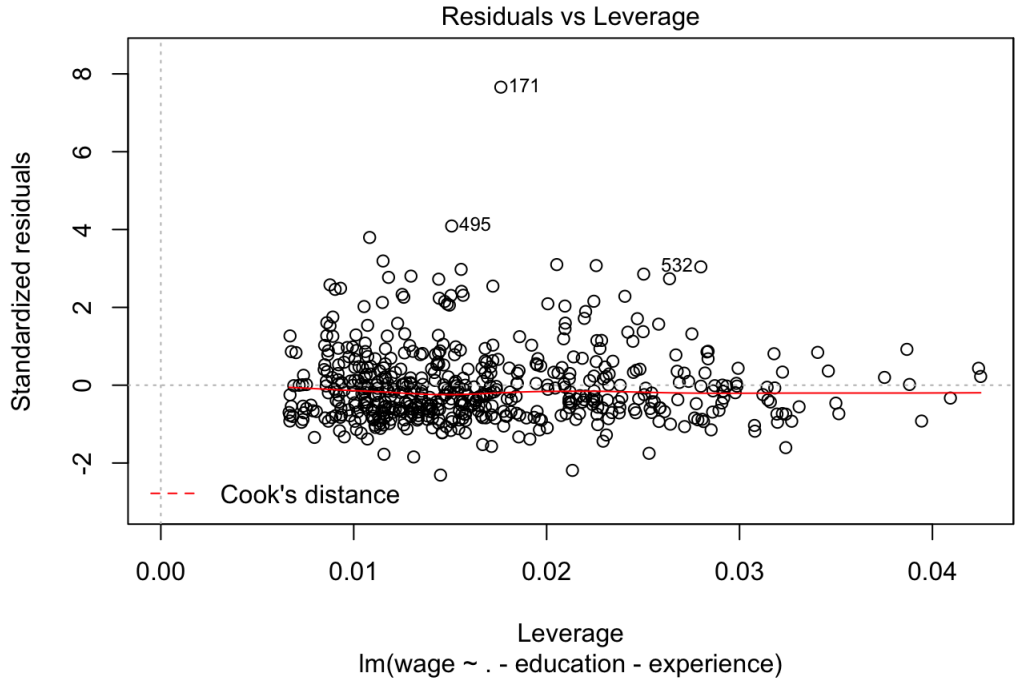
```
lm1 <- lm(wage~.-education-experience,data = df)
```

```
plot(lm1)
```









From the plot of residual vs fitted we cannot see any distict pattern so the hypothesis is satisfied, but from the QQ plot we can see much departure form the right tail so the residules may not follow a normal distribution.

Question3

```
summary(lm1)
```

```
##
## Call:
## lm(formula = wage ~ . - education - experience, data = df)
##
## 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 ***
## south       -1.30714    0.46527  -2.809 0.005148 **
## sex         -2.33191    0.43661  -5.341 1.38e-07 ***
## union        1.72837    0.57516   3.005 0.002782 **
## age          0.06699    0.01890   3.544 0.000429 ***
## race         0.75057    0.31166   2.408 0.016370 *
## occupation  -0.39447    0.14134  -2.791 0.005448 **
## sector       0.25381    0.42207   0.601 0.547869
## 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
```

From the summary of lm1 we can see sector and marr are not significant since their p value are larger than 0.05.

Like all other features, 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.

Question4

```
library(MASS)
stepAIC(lm(wage~.,data = df),k = log(nrow(df)))
```

```
## Start:  AIC=1640.48
## wage ~ education + south + sex + experience + union + age + race +
##     occupation + sector + marr
##
##              Df Sum of Sq   RSS    AIC
## - age         1      2.89 10131 1634.3
## - experience  1      4.34 10133 1634.4
## - marr        1     19.89 10148 1635.2
## - occupation  1     26.26 10155 1635.6
## - education   1     27.74 10156 1635.7
## - south       1     51.32 10180 1636.9
## - race        1     54.42 10183 1637.1
## - sector      1     66.65 10195 1637.7
## <none>                    10128 1640.5
## - union       1    161.63 10290 1642.7
## - sex         1    558.49 10687 1662.9
##
## Step:  AIC=1634.35
## wage ~ education + south + sex + experience + union + race +
##     occupation + sector + marr
##
##              Df Sum of Sq   RSS    AIC
## - marr        1     19.25 10150 1629.1
## - occupation  1     25.55 10157 1629.4
## - south       1     51.87 10183 1630.8
## - race        1     54.70 10186 1631.0
## - sector      1     66.26 10198 1631.6
## <none>                    10131 1634.3
## - union       1    161.13 10292 1636.5
## - sex         1    555.78 10687 1656.6
## - experience  1    587.26 10718 1658.2
## - education   1   2330.91 12462 1738.7
##
## Step:  AIC=1629.09
## wage ~ education + south + sex + experience + union + race +
##     occupation + sector
```

```
##
##              Df Sum of Sq   RSS    AIC
## - occupation  1     26.48 10177 1624.2
## - south       1     49.96 10200 1625.4
## - race        1     58.52 10209 1625.9
## - sector      1     69.04 10220 1626.4
## <none>                    10150 1629.1
## - union       1    169.79 10320 1631.7
## - sex         1    555.42 10706 1651.3
## - experience  1    691.07 10842 1658.0
## - education   1   2367.16 12518 1734.7
##
## Step:  AIC=1624.2
## wage ~ education + south + sex + experience + union + race +
##     sector
##
##              Df Sum of Sq   RSS    AIC
## - sector      1     49.70 10227 1620.5
## - south       1     50.46 10227 1620.6
## - race        1     54.29 10231 1620.8
## <none>                    10177 1624.2
## - union       1    149.27 10326 1625.7
## - sex         1    532.35 10709 1645.2
## - experience  1    745.12 10922 1655.7
## - education   1   2569.80 12747 1738.2
##
## Step:  AIC=1620.52
## wage ~ education + south + sex + experience + union + race
##
##              Df Sum of Sq   RSS    AIC
## - south       1     52.38 10279 1617.0
## - race        1     57.25 10284 1617.2
## <none>                    10227 1620.5
## - union       1    160.36 10387 1622.5
## - sex         1    605.14 10832 1644.9
## - experience  1    768.45 10995 1652.9
## - education   1   2521.58 12748 1731.9
##
## Step:  AIC=1616.97
## wage ~ education + sex + experience + union + race
##
##              Df Sum of Sq   RSS    AIC
## - race        1     70.83 10350 1614.3
## <none>                    10279 1617.0
## - union       1    180.44 10460 1620.0
## - sex         1    595.08 10874 1640.7
## - experience  1    788.72 11068 1650.2
## - education   1   2677.57 12957 1734.3
##
## Step:  AIC=1614.35
## wage ~ education + sex + experience + union
##
##              Df Sum of Sq   RSS    AIC
## <none>                    10350 1614.3
## - union       1    163.22 10513 1616.4
## - sex         1    590.35 10940 1637.7
## - experience  1    798.23 11148 1647.8
## - education   1   2786.26 13136 1735.4
```

```
##
## Call:
## lm(formula = wage ~ education + sex + experience + union, data = df)
##
## Coefficients:
## (Intercept)    education          sex   experience        union
##     -4.3326       0.9350      -2.1477       0.1069       1.4711
```

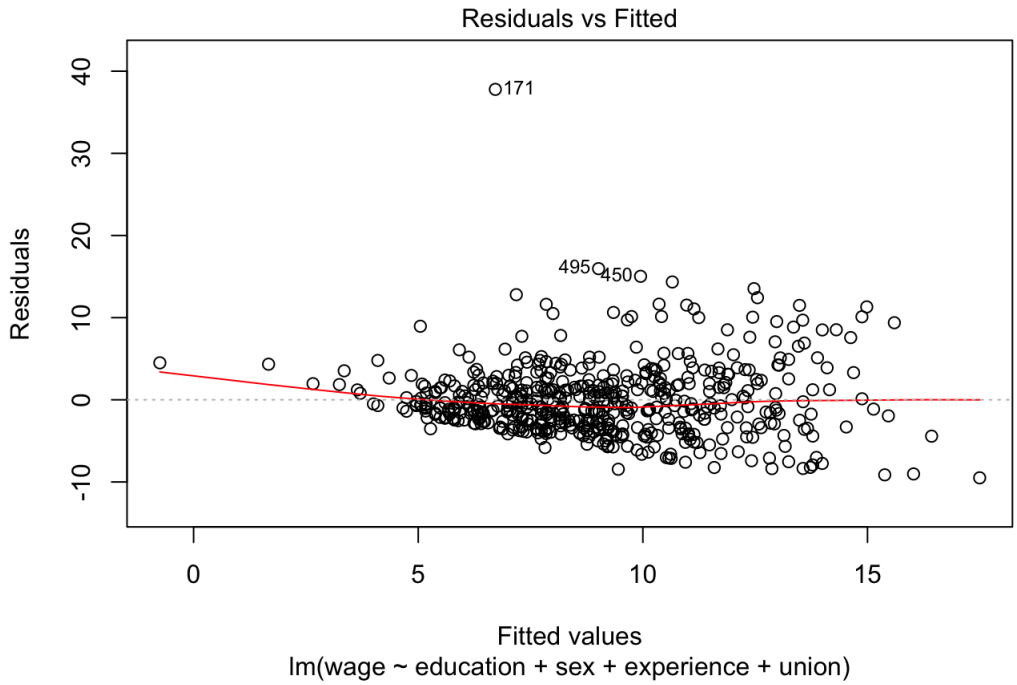
With the stepwise bic approch we find a simpler model as formula = wage ~ education + sex + experience + union. This model has BIC = 1614, which is much smaller than the original full model as BIC = 1640.48.

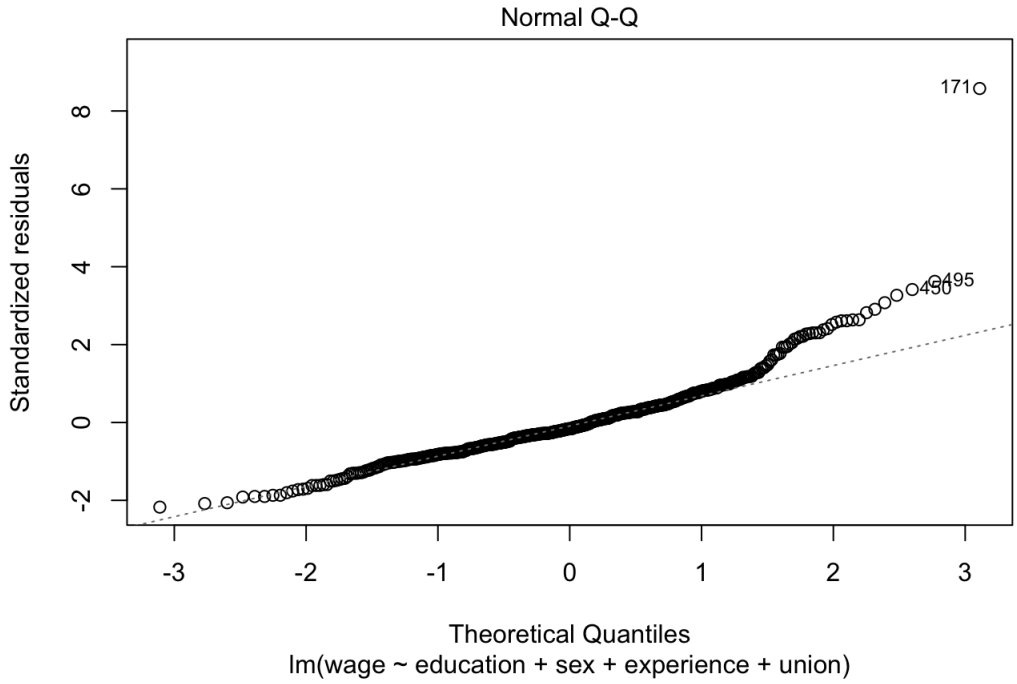
Question5

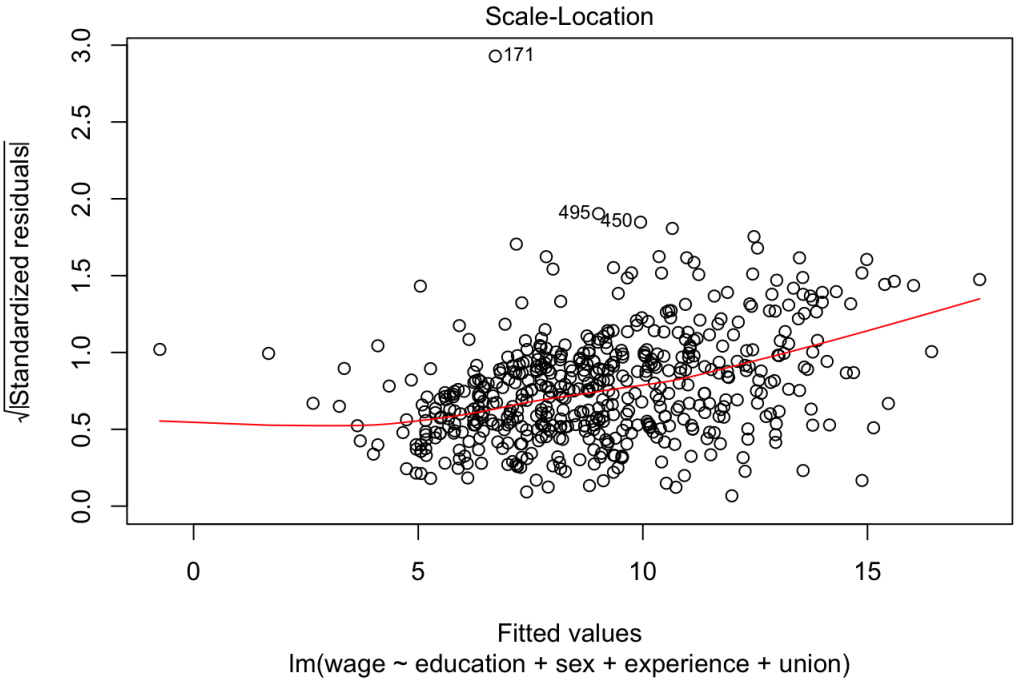
```
lm2 <-lm(formula = wage ~ education + sex + experience + union, data = df)
summary(lm2)
```

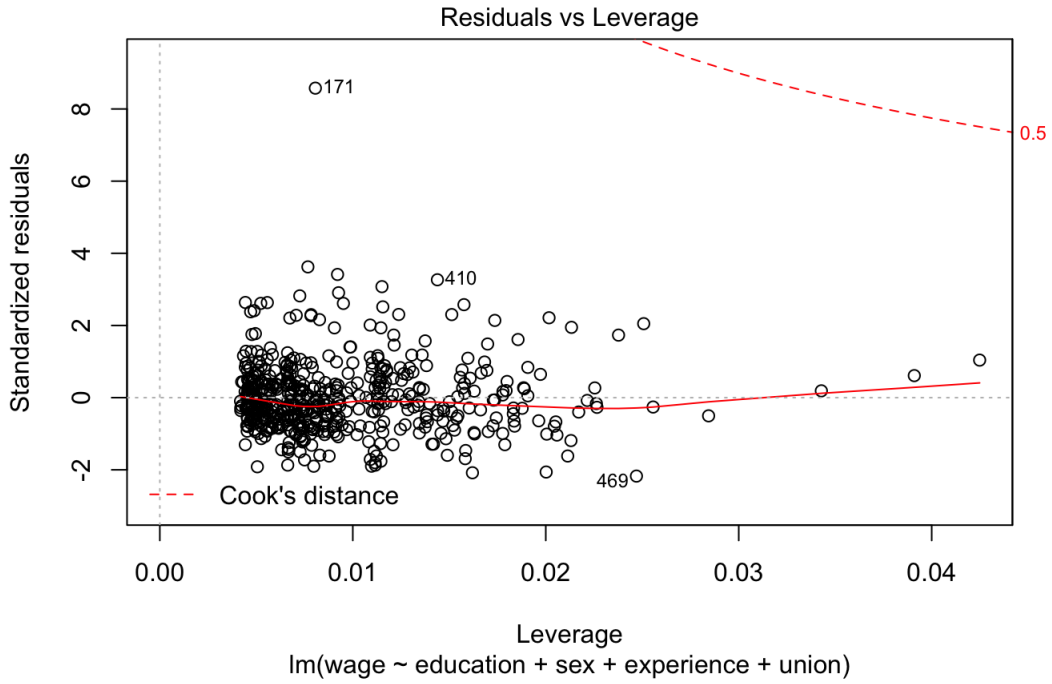
```
##
## Call:
## lm(formula = wage ~ education + sex + experience + union, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.496 -2.708 -0.712  1.909 37.784
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.33258    1.17993  -3.672 0.000265 ***
## education    0.93495    0.07835  11.934  < 2e-16 ***
## sex         -2.14765    0.39097  -5.493 6.14e-08 ***
## experience   0.10692    0.01674   6.387 3.70e-10 ***
## union        1.47111    0.50932   2.888 0.004031 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.423 on 529 degrees of freedom
## Multiple R-squared:  0.2648, Adjusted R-squared:  0.2592
## F-statistic: 47.62 on 4 and 529 DF,  p-value: < 2.2e-16
```

```
plot(lm2)
```

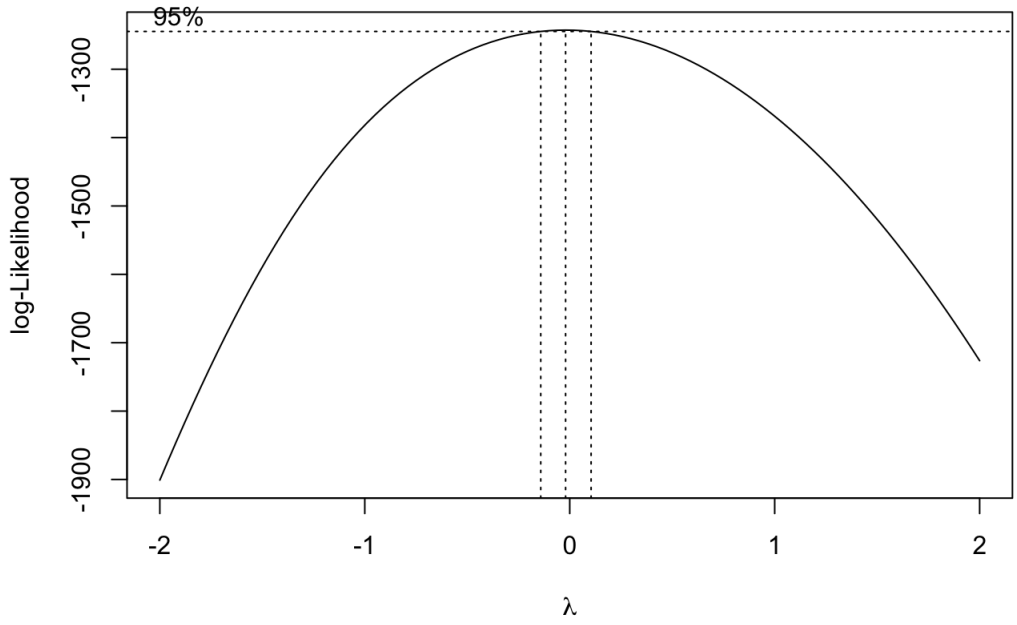



Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(wage ~ education + sex + experience + union)
171
495
450


Scale-Location
√|Standardized residuals|
Fitted values
lm(wage ~ education + sex + experience + union)
171
495
450




```
boxcox(wage ~ education + sex + experience + union, data = df)
```

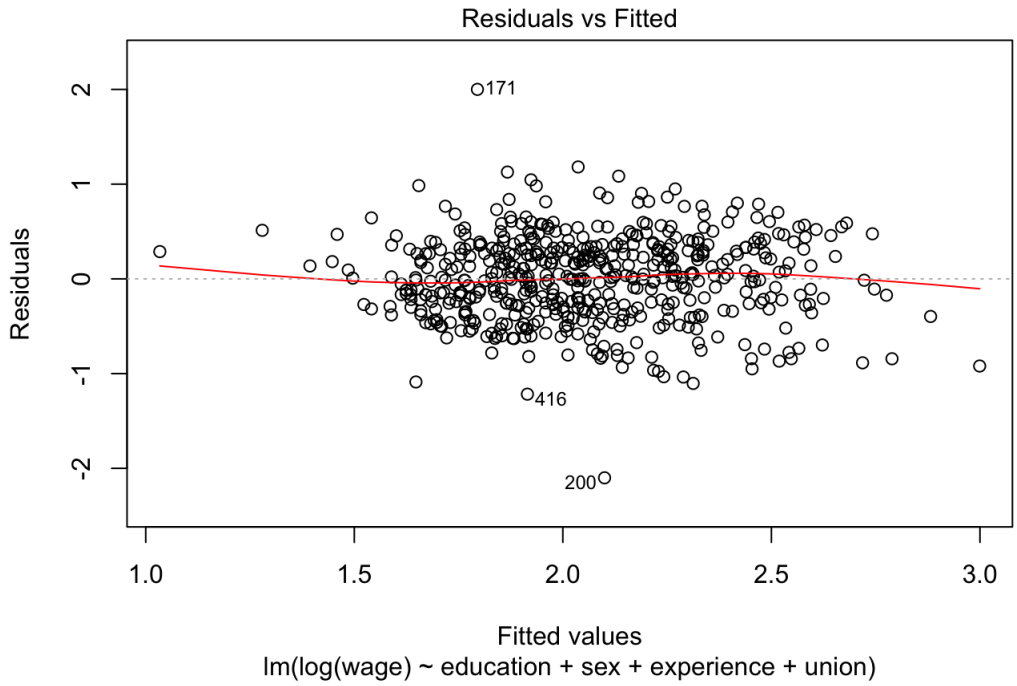


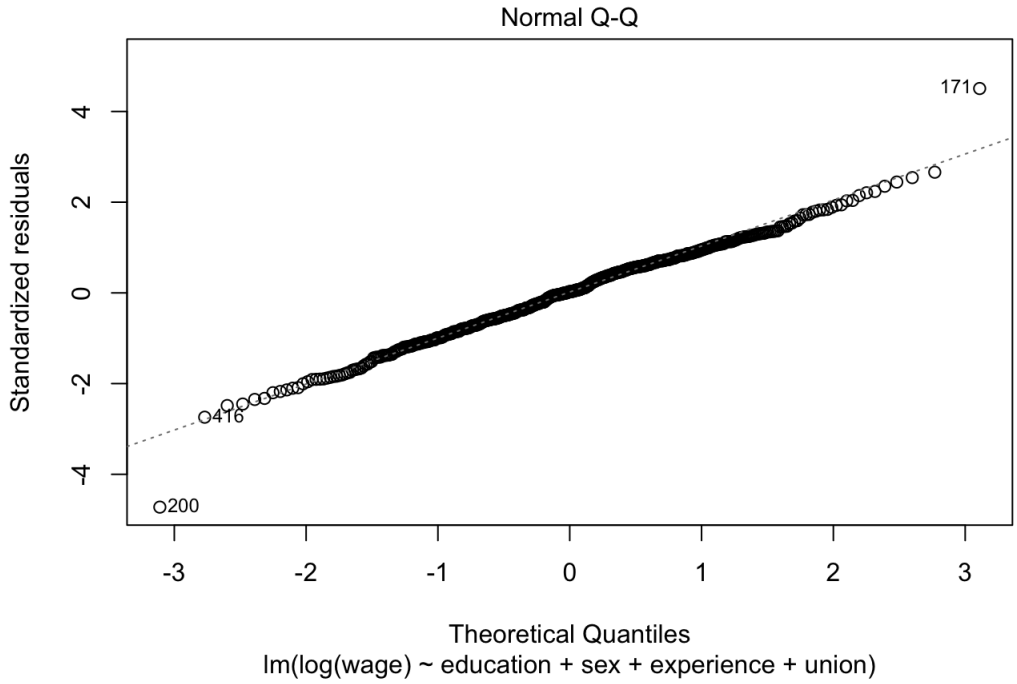
The QQ plot still have heavey tail, which indicate the abnormaluty of the residual. From box-cox plot, we can see that lambda=0 lies inside the confidence interval. So log transformation seems resonable.

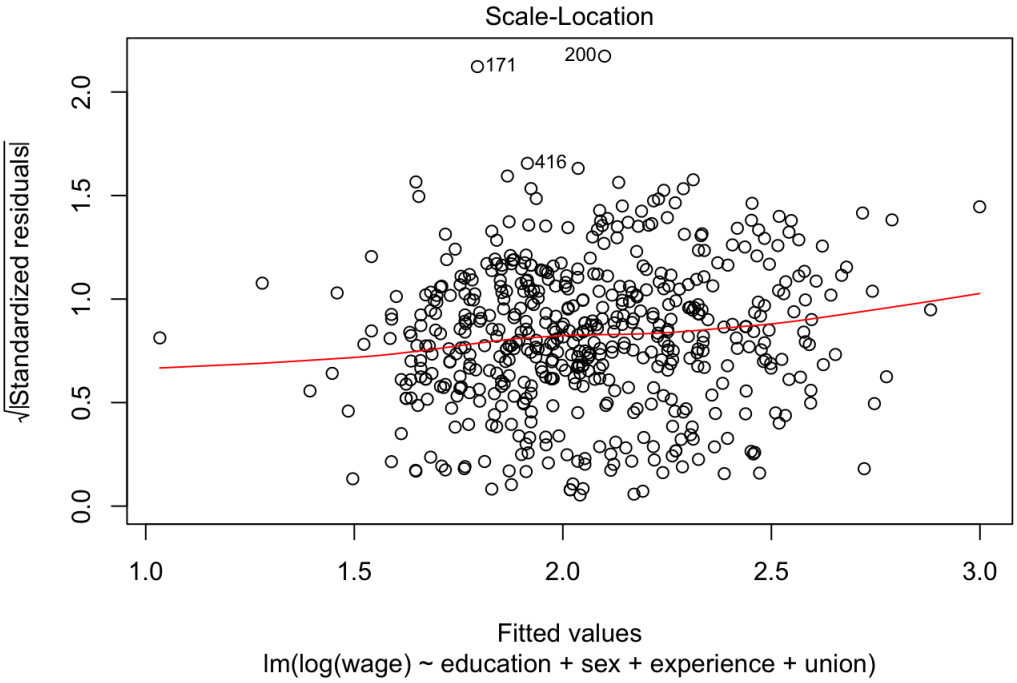
```
lm3 <-lm(log(wage) ~ education + sex + experience + union, data = df)
summary(lm3)
```

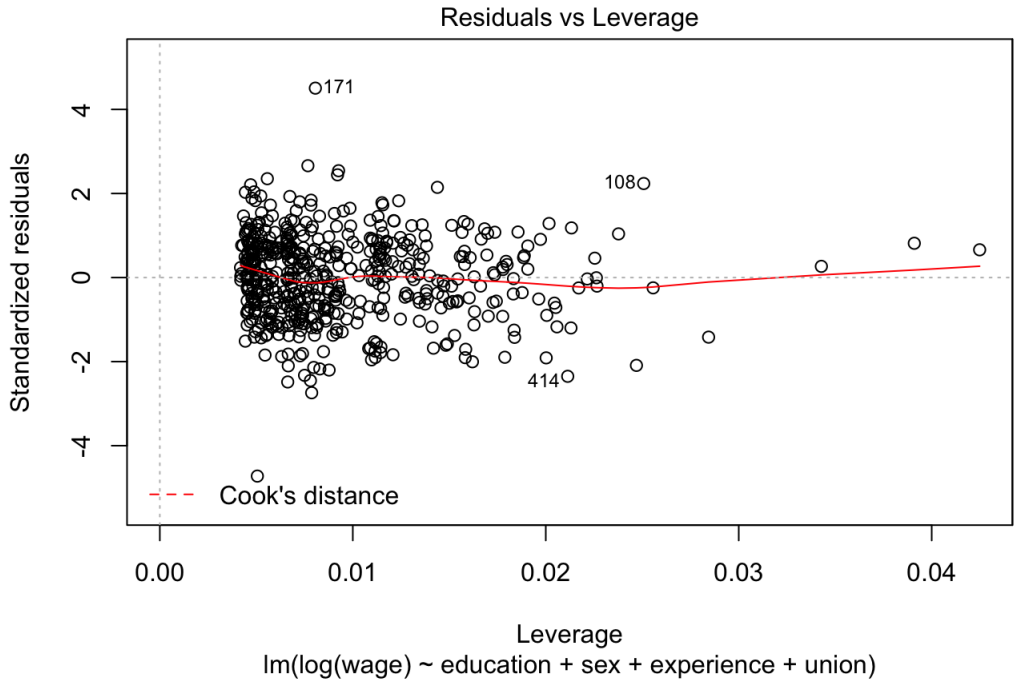
```
##
## Call:
## lm(formula = log(wage) ~ education + sex + experience + union,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1000 -0.2953  0.0120  0.3121  2.0003
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.651259   0.118898   5.477 6.68e-08 ***
## education    0.097213   0.007895  12.314  < 2e-16 ***
## sex         -0.228810   0.039397  -5.808 1.09e-08 ***
## experience   0.011758   0.001687   6.970 9.46e-12 ***
## union        0.210133   0.051323   4.094 4.90e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4457 on 529 degrees of freedom
## Multiple R-squared:  0.2921, Adjusted R-squared:  0.2867
## F-statistic: 54.56 on 4 and 529 DF,  p-value: < 2.2e-16
```

```
plot(lm3)
```




Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(log(wage) ~ education + sex + experience + union)
171
416
200


Scale-Location
√|Standardized residuals|
Fitted values
lm(log(wage) ~ education + sex + experience + union)
171
200
416



From the plot above we can see the residual seems uncorrelated , indipendent and follow a normal trend. 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.

Question6

```
df <- df[-c(171,200),]
lm3 <-lm(log(wage) ~ education + sex + experience + union, data = df)
summary(lm3)
```

```
##
## Call:
## lm(formula = log(wage) ~ education + sex + experience + union,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.21471 -0.28468  0.00972  0.30993  1.19811
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.649986   0.114258   5.689 2.12e-08 ***
## education    0.097143   0.007584  12.808  < 2e-16 ***
## sex         -0.246328   0.037934  -6.494 1.94e-10 ***
## experience   0.012404   0.001624   7.639 1.04e-13 ***
## union        0.203882   0.049325   4.133 4.16e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4282 on 527 degrees of freedom
## Multiple R-squared:  0.3157, Adjusted R-squared:  0.3105
## F-statistic: 60.79 on 4 and 527 DF,  p-value: < 2.2e-16
```

The result seems improved since the residual standard of error decreased and the R-square improved. From previous residual plot we can see the 171th and 200th point are outlier， which have adnormal large residual. Therefore remove these two points would improve our model performance.