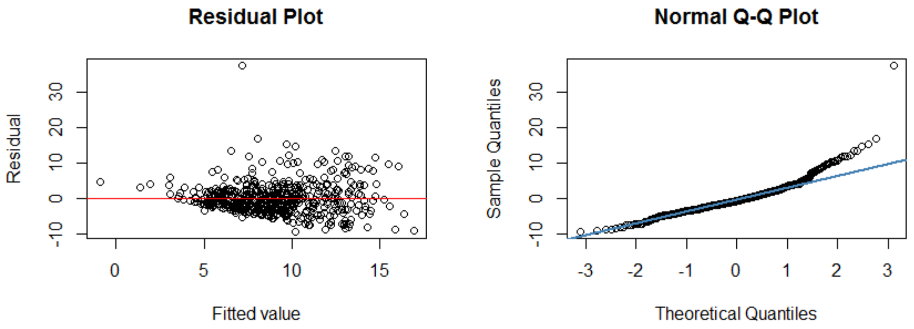
Group: Zining Fan, Mutian Wang, Siyuan Wang
UNI: zf2234, mw3386, sw3418

Graded Homework 3- Exercise 3

1. We can use linear regression model for this dataset.

   It's not a good idea because there's a colinearity problem. Elder people tend to have more education and work experience.

2. Yes, there's some departure from the hypotheses. The residuals are not normally distributed. In the Q-Q Plot, points on the right are not on the line.



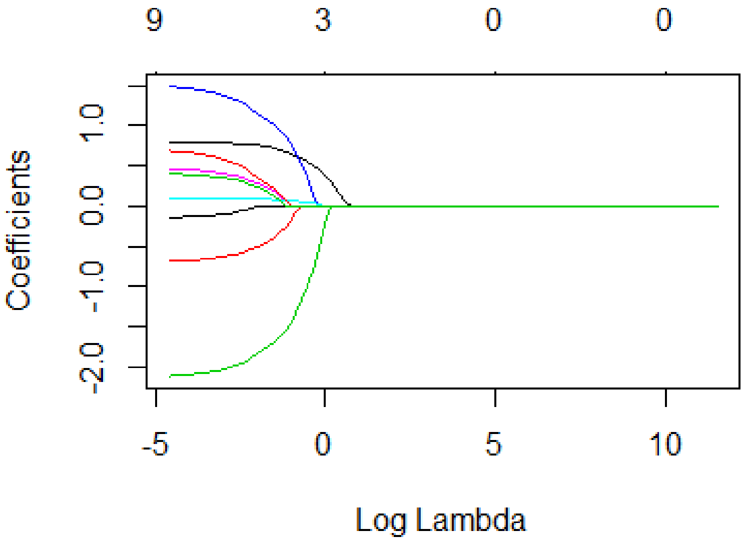
3. Coefficients:

```
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.0402     6.8790  -0.297  0.76690
education     1.3262     1.1082   1.197  0.23194
south        -0.6975     0.4285  -1.628  0.10414
sex          -2.1443     0.3993  -5.370 1.19e-07 ***
experience    0.5246     1.1086   0.473  0.63625
union         1.5168     0.5250   2.889  0.00403 **
age          -0.4282     1.1079  -0.386  0.69931
race          0.4786     0.2855   1.676  0.09426 .
occupation   -0.1527     0.1312  -1.165  0.24475
sector        0.7190     0.3876   1.855  0.06414 .
marr          0.4252     0.4195   1.013  0.31131
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

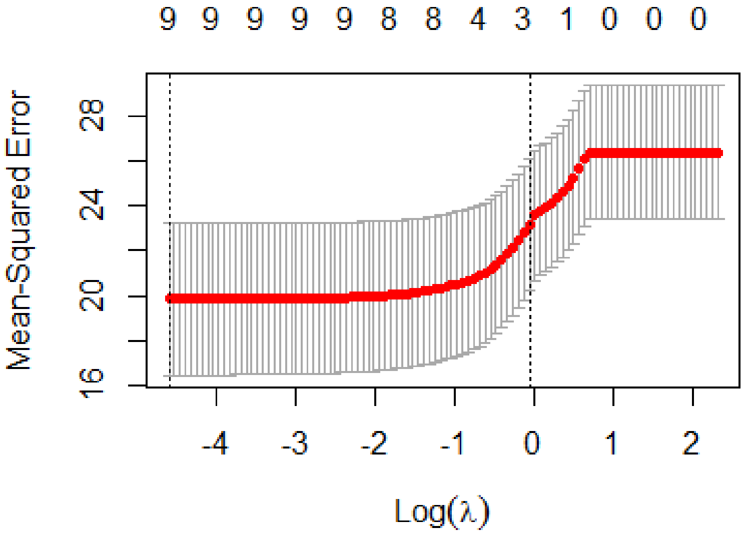
   Not all the parameters are significant. In fact, only sex and union are significant under 5% level.

   To test the significance of the sector variable, we can use a t-test. According to the two-sided p-value, it's significant under 10% level.

4. We can use Lasso regularization to find a simpler model. The below graph shows how coefficients change over the regularization λ. The top bar shows the number of non-zero coefficients according to $\log \lambda$.



5. We use cross validation to find the best regulariztion term λ. The result is shown in the following graph. The top bar shows the number of non-zero coefficients according to $\log \lambda$.



From the graph, we can see the optimal λ is between 0.1 and 10. The cross validation tells us the optimal $\lambda = 0.8$.

Now we use $\lambda = 0.8$ to fit Lasso again, and check what features are left.

```
education   0.46243181
south        .
sex        -0.56512373
experience  .
union       .
age         0.02664511
race        .
occupation  .
sector      .
marr        .
```

The selected features are 'education', 'sex' and 'age'. Now we fit a new linear model on these three features. The results are as follows.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.84275    1.24425  -3.892 0.000112 ***
df$education  0.82744    0.07462  11.089  < 2e-16 ***
df$sex       -2.33542    0.38807  -6.018 3.30e-09 ***
df$age        0.11310    0.01669   6.775 3.32e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Clearly, all variables are significant.

6. Point 2: If 171st and 200th rows are dropped, we can still observe some departure from the hypotheses. In fact, the two diagnostic plots are almost the same. The only difference is that the outlier is removed.

   Point 3: As for significance, the conclusion remains the same.

   Point 5: The optimal λ now becomes 0.5, and the accroding non-zero features are 'education', 'sex', 'union' and 'age'.