STATS 500 Homework 2 YUAN YIN

Problem 1

1. Fit the regression model with R, also we omit the missing values at first, the result is as follows:

```
Call:
lm(formula = wage ~ educ + exper, data = uswages)

Residuals:
   Min     1Q  Median     3Q    Max
-1014.7  -235.2   -52.1   150.1  7249.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -239.1146    50.7111  -4.715 2.58e-06 ***
educ          51.8654     3.3423  15.518  < 2e-16 ***
exper          9.3287     0.7602  12.271  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 426.8 on 1964 degrees of freedom
Multiple R-squared:  0.1348,    Adjusted R-squared:  0.1339
F-statistic:   153 on 2 and 1964 DF,  p-value: < 2.2e-16
```

The final regression model is as follows:

$$y = -239.1146 + 51.8654\text{educ} + 9.3287\text{exper}$$

2. From the result above we can see that Multiple R - squared is 0.1348, which means the percentage of variation in the response explained by these predictors is 13.48%.

3. We use which.max() function to find the maximum residual of our regression, the case number is 15387, it’s the 1550th data of uswages and the max residual value is 7249.17.

Attention: 1550th represents the number in data which has been omitted the missing values, so it’s not the same in original data “uswages”.

4. Use mean() and median() function finding that the mean of residuals is -1.381535e-15 and the median of residuals is -52.14337, we found that these two results have a large difference, this is because mean is influenced by the variance of every residuals but median is only influenced by the middle residuals after sorted. As we found that the mean is almost zero but median is much less than zero, it means that of all residuals, the number of negative residuals are more than positive residuals but the absolute value of positive residuals are commonly larger than negative ones.

5. The correlation of residuals and fitted values is 6.35678e-17, the plot is as follows:

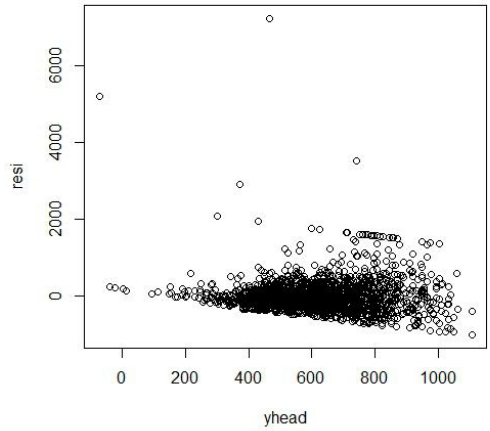

Fig 1 residuals against fitted values

6. The regression model is as follows:

$$y = -239.1146 + 51.8654\text{educ} + 9.3287\text{exper}$$

It means, for two people with the same education but one year difference in experience, their wages may have $9.3287 difference.

7. The new model is as follows:

```
Call:
lm(formula = ystar ~ educ + exper, data = uswages)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7527 -0.3383  0.1002  0.4297  3.5728

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.675518   0.077085   60.65   <2e-16 ***
educ        0.091940   0.005080   18.10   <2e-16 ***
exper       0.016516   0.001156   14.29   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6488 on 1964 degrees of freedom
Multiple R-squared:  0.1747,    Adjusted R-squared:  0.1739
F-statistic: 207.9 on 2 and 1964 DF,  p-value: < 2.2e-16
```

The final regression model is:

$$ystar = 4.675518 + 0.091940\text{educ} + 0.016516\text{exper}$$

Here “ystar” equals to log(wages)

We notice that in this model, R-squared has a little higher than before, but still, they are both far from 1, now we analyze this model from the plot of log(wages), we can found that when the year of education adds 1 unit, log(wages) will add 0.09, this means with the higher original wages, it will also increase more when year of education increases. Also we can see that the

changes of experience has little influence on wages, which also fits with the plot of wages and year of experiences. So we conclude that the second model is more natural.

Appendix

```
## read in the data
library(faraway)
data(uswages)
## get the X matrix
dim(uswages)
uswages$exper[uswages$exper < 0] = NA
uswages = na.omit(uswages)
## use the lm() function
temp = lm(wage ~ educ + exper, data = uswages)
summary(temp)
##residual sum of squares
resi = residuals(temp)
summary(resi)
which.max(resi)
## compute mean and median
mean(resi)
median(resi)
## fit the model
yhead = fitted(temp)
cor(resi,yhead)
plot(yhead,resi)
## log wage as response
ystar = log(uswages$wage)
tempstar = lm(ystar ~ educ + exper, data = uswages)
summary(tempstar)
```