Huiwen Chen   Department of Statistics   UM-ID : 02156341

STAT 500 Homework 2

1. Out put of the regression model with wages as the response and years of education and experience as preditors:

Code:

```
>library (faraway)
##read in and check out the data
>data (uswages)
>attach (uswages)
>uswages
uswages$exper[uswages$exper < 0] <- NA
>dim ( uswages )
>n = dim ( uswages ) [1]
>p = dim ( uswages ) [2]
>x = cbind ( 1, as.matrix(uswages [,] ))
>reg = lm( wage ~ educ + exper, data = uswages)
>summary(reg)
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
   (33 observations deleted due to missingness)
Multiple R-squared:    0.1348,     Adjusted R-squared:    0.1339
F-statistic:      153 on 2 and 1964 DF,    p-value: < 2.2e-16
```

2. What percentage of variation in the response is explained by these predictors?

Code:

```
>regs <- summary(reg)
>names (regs)
## extract R2
>regs $ r.squared
  0.134793
```

Answer: Since the Multiple R-squared is 0.134793, it shows that nearly 13.48 % of the variation in the response is explained by these predictors.

3. Which observation has the largest (positive) residual? Give the case number.

Code:

```
>rsd= as.vector(regs$residuals)
>which.max (rsd)
 [1] 1550
```

Answer: The No. 1550 case has the largest positive residual.

4. Compute the mean and median of the residuals. Explain what the difference between the mean and the median indicates.

Code:

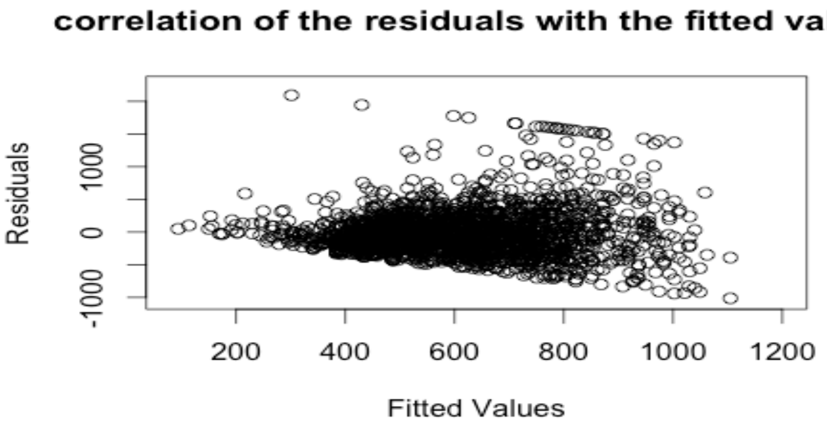
```
>mean(rsd)
[1] -1.381535e-15
> median(rsd)
[1] -52.14337
```

Answer: The mean of the residuals is -1.381535e-15, and the median is -52.14337. The median indicates that half of the residuals was less than -52.14337, while the mean indicates that the average residual is -1.381535e-15 ≈0,it may be effected by the extremely large or small data.

5. Compute the correlation of the residuals with the fitted values. Plot residuals against fitted values.

Code:

```
> cor ( rsd, reg$fitted.values)
[1] 6.35678e-17
> plot (rsd, reg$fitted.values,log = 'y',main="correlation of the residuals with the fitted values",ylim=c(80,1500), xlim=c(-1050,2250),ylab="Fitted Values",xlab="Residuals")
```



6. For two people with the same education and one year difference in experience, what would be the difference in predicted weekly wages?

Answer: See output in summary, the estimate coefficient of experience is 9.3287, that is the difference in predicted weekly wages.

7. Fit the same model but with log (weekly wages) as the response and interpret the regression coefficient for experience. Which model has a more natural interpretation?

```
Code:
>log_wage <-log (uswages$wage)
>log_reg <- lm ( log_wage ~ educ + exper, data = uswages)
>summary(log_reg)
Call:
lm(formula = log_wage ~ 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
  (33 observations deleted due to missingness)
Multiple R-squared:  0.1747,   Adjusted R-squared:  0.1739
F-statistic: 207.9 on 2 and 1964 DF,  p-value: < 2.2e-16
```

Answer:

The regression coefficient for experience 0.016516 is smaller than before, which will make it easy to calculate estimate response.

And since the R-squared 0.1747 in this case is more closer to 1 than the former R-squared 0.1348, the fit with log (weekly wages) as the response is better and has a more natural interpretation.