Stats500 Homework 1

Di Lu, Sept.27, 2017

1. Chapter 2, problems 1 and 2 (page 30)

Clean data

Note that the experience variable has some negative value, which most likely indicate missing data. We remove the 33 observations with negative exper values before running linear regression. On the other hand, education data has no negative values.

Fit the linear regression model

1. *Fit a regression model with weekly wages as the response and years of education and experience as predictors. Present the output*

   Put weekly wages as the response; Put years of education and experience as predictors. We can get a linear regression model as follows:

   wage = -239.1146 + 51.8654 x educ + 9.3287 x exper
   (50.7111) (3.3423) (0.7602)

   R-square = 0.1348

   We can tell from the model both educ and exper are significant at 1% significance level.

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

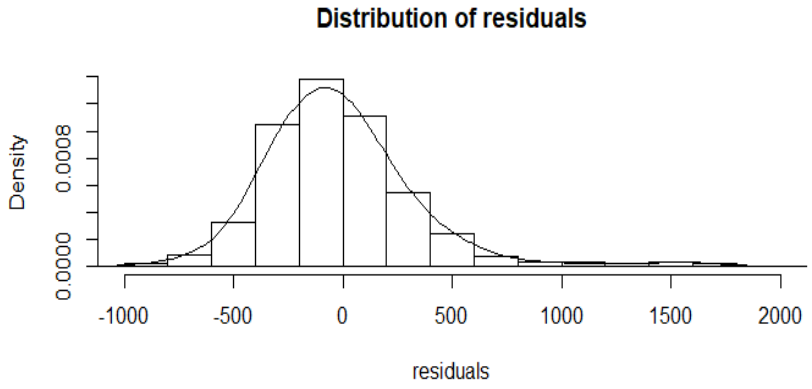
   R-square = 0.1348 shows that 13.48% of variation in the response is explained by the predictors.

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

   Apply which.max(model$residuals), we find that the largest residual is index 1576 (and 1550). They both have residual 7249.2.

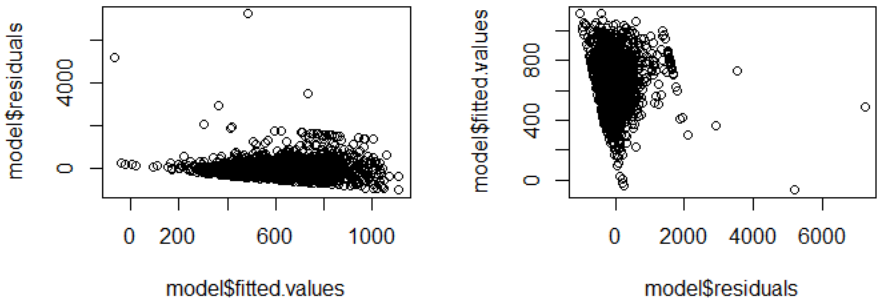
4. *Compute the mean and median of the residuals. Explain what the difference between the mean and the median indicates.*

   Mean of the residual is -1.38e-15, median of the residual is -52.14, which means the distribution of residual is skewed towards left as shown in the graph.



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

   The correlation of the residuals with the fitted values is 6.35678e-17, the p-value is 1, which means at 1% siganificance level we cannot reject that residuals and the fitted values have 0 correlation. But look at the scatter plot(right), we can tell that residuals increase as fitted values go larger, which is contradictary with our homoscedasticity assumption of residuals.



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

   wage = -239.1146 + 51.8654 x educ + 9.3287 x exper

   with Δ(exper) = 1, the predicted weekly wages will be 9.3287 higher.

Fit the linear regression model after transformation

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?

Put lwage = log(wage) as the response variable, We get the linear model:

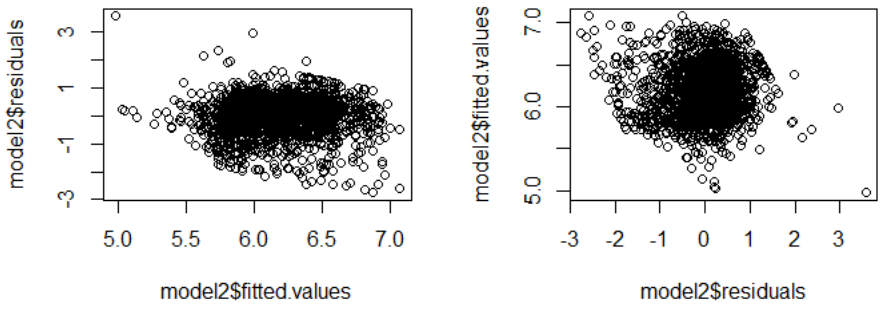
$$\text{lwage} = 4.650319 + 0.090506 \times \text{educ} + 0.018079 \times \text{exper}$$

$$(0.078354) \quad (0.005167) \qquad (0.001160)$$

R-Square = 0.1749

We can see this model has improve R-Square from 13.48% to 17.49% and its interpretation is more natural. Worker with +1 education years will enjoy 9.05% increase in weekly wage, while with +1 experience years will enjoy 1.80% increase in weekly wage.

By this way, the correlation of the residuals with the fitted values in this model is 2.479713e-17, the p-value is 1, which means at 1% significance level we cannot reject that residuals and the fitted values have 0 correlation. Here we can see in the scatter plot: residuals are more randomly distributed around zero and seems equally variance at varied fitted value level.



Appendix:

```
# load the "faraway" package and load the dataset "uswages"
rm(list=ls())
library(faraway)
library(ggplot2)
attach(uswages)

# clean data
# Note that the experience variable has some negative values which most likely indicate missing data.
# Those observations should be removed from the analysis.
summary(educ)
summary(exper)
exper[exper<0]=NA #33 numbers are negative

#1. Fit a regression model with weekly wages as the response and years of education
# and experience as predictors. Present the output
model <- lm(wage ~ educ+exper, na.rm = TRUE)
model3 <- lm(wage ~ educ+exper)
summary(model)
summary(model3)
model$residuals

#2. What percentage of variation in the response is explained by these predictors?
# R-square

#3. Which observation has the largest (positive) residual? Give the case number.
which.max(model$residuals)
which(abs(model$residuals - 7249.2)<0.1)
#4. Compute the mean and median of the residuals. Explain what the di
erence between the
# mean and the median indicates.
mn = mean(model$residuals)
md = median(model$residuals)
hist(model$residuals,breaks = 30, xlim = range(-1000,2000), xlab = "residuals",main="Distribution of
residuals",probability = TRUE)
lines(density(model$residual,adjust=2))

#5. Compute the correlation of the residuals with the fitted values. Plot residuals against fitted values.
cor(model$residuals,model$fitted.values)
par(mfrow=c(1,2))
```

```
plot(model$fitted.values,model$residuals)
plot(model$residuals,model$fitted.values)
```

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

```
lwage = log(wage)
model2 <- lm(lwage ~ educ+exper, na.rm = TRUE)
summary(model2)
cor.test(model2$residuals,model2$fitted.values)
plot(model2$fitted.values,model2$residuals)
plot(model2$residuals,model2$fitted.values)
```