W4201 Advanced Data Analysis HW 5

Yao Chen (Uni: yc2798)

Due by 10/11/2013

Question (a)

Run

```
reg <- lm(medv~crim+zn+indus+nox+rm+age+tax+ptratio,data=Boston)
summary(reg)
```

obtaining,

```
Call:
lm(formula = medv ~ crim + zn + indus + nox + rm + age + tax +
    ptratio, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-14.034  -3.128  -0.689   1.929  40.201

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.176513   4.891698   1.263 0.207306
crim         -0.124580   0.036866  -3.379 0.000784 ***
zn           -0.011524   0.015032  -0.767 0.443667
indus         0.014053   0.069345   0.203 0.839488
nox         -10.606370   4.341827  -2.443 0.014920 *
rm            6.977955   0.411199  16.970  < 2e-16 ***
age          -0.017220   0.014288  -1.205 0.228704
tax          -0.001766   0.002762  -0.639 0.522947
ptratio      -1.045103   0.153838  -6.794 3.14e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.735 on 497 degrees of freedom
Multiple R-squared:  0.6174,Adjusted R-squared:  0.6112
F-statistic: 100.2 on 8 and 497 DF,  p-value: < 2.2e-16
```

The fitted linear model is:

$$\begin{aligned} medv &= 6.177 - 0.125 \times crim - 0.012 \times zn + 0.014 \times indus - 10.61 \times nox \\ &\quad +6.978 \times rm - 0.017 \times age - 0.0018 \times tax - 1.045 \times ptratio \end{aligned}$$

Question (b)

i. **linearity form**

The p-value from the F-test of the linear model is less than $2.2E-16$. So the linear assumption is invalid.

ii. **normality**

Based on the Q-Q plot for regression residuals, the normality assumption is invalid.

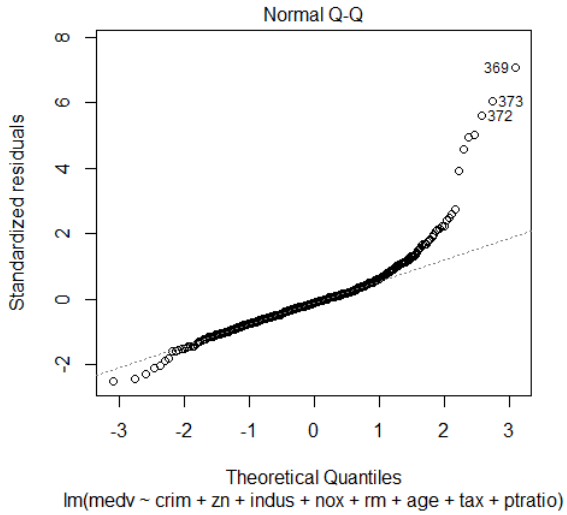

Figure 1: Normal Q-Q plot for regression residuals.

For further consideration, Shapiro-Wilk normality test is conducted:

```
> shapiro.test(e)

Shapiro-Wilk normality test

data:  e
W = 0.8299, p-value < 2.2e-16
```

The p-value from the test is less than $2.2E-16$, which also shows that the normality assumption is invalid.

iii. **Homoscedasticity**

From the residuals v.s. fitted values plot, we can see that the homoscedasticity assumption is invalid. The relationship between residuals and fitted values is more like a quadratic relationship.

For further information, the residuals are separated into two groups. Compare the variances of such two groups:

```
> e <- reg$residuals
> var.test(e[1:253],e[254:506])
```

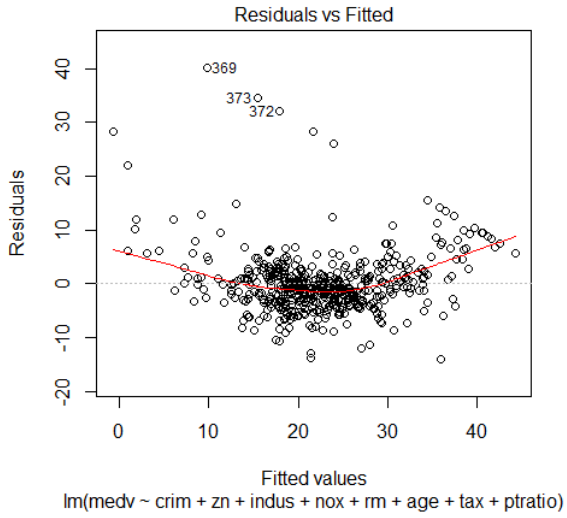

Figure 2: Regression residuals against fitted value.

```
F test to compare two variances

data:  e[1:253] and e[254:506]
F = 0.326, num df = 252, denom df = 252, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2545463 0.4175759
sample estimates:
ratio of variances
         0.3260252
```

Since the p-value from the two variance F-test is extremely small, the homoscedasticity is invalid

iv. **Uncorrelated error**

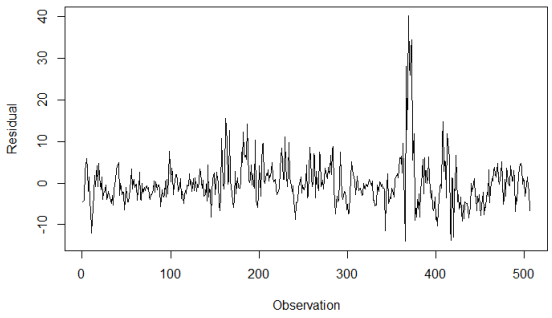

Figure 3: Regression residuals against observation number.

From the figure, we can see there is potential correlation. For further information, a Durbin-Watson test for 1 st order AR model is conducted:

```
> dwtest(medv~crim+zn+indus+nox+rm+age+tax+ptratio,data=Boston)

Durbin-Watson test

data:  medv ~ crim + zn + indus + nox + rm + age + tax + ptratio
DW = 0.7919, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

The p-value from DW test is extremely small, which indicates that there is autocorrelation greater than 0.

v. **Influential points and outliers**

Calculate the Studentized deleted residuals. The critical value is

$$t_{\frac{\alpha}{2n},n'-p-1} =$$

Run

```
lmi <- lm.influence(reg)
h <- lmi$hat
si <-lmi$sigma
student.resid <-  e/(si*(1-h)^0.5)
critical.value <- qt(1-0.05/2/506,506-8-1)
abline(h=critical.value)
abline(h=-critical.value)
which(abs(student.resid)>critical.value)
```

Obtaining,

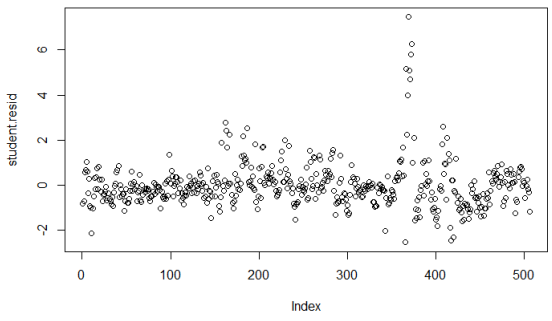

Figure 4: Studentized deleted residuals against observation number.

and the influential points and outliers are observation 366, 368, 369, 370, 371, 372, 373.

Question (c)

i. **linearity form**

Conduct EDA to check potential relationship between Y and X. Transform data properly. Or we can establish and fit a non-linear model.

ii. **normality**

We can still transform the data. Or we can use robust regression method.

iii. **Homoscedasticity**

We can use a WLS model instead of OLS model.

iv. **Uncorrelated error**

We can use a data transformation following Cochrane-Orcutt Procedure. Or we can use models that incorporate the correlation structure, such as Generalized Estimating Equations (GEE) method.