Homework 4

Qianbo Wang
uni: qw2180

Consider the Boston dataset, in R library MASS, on Housing Values in Suburbs of Boston

(a). Fit a multiple linear regression model to predict medv (median value of owner-occupied homes in $1000s) using the following set of predictors: crim, zn, indus, nox, rm, age, tax.

The linear regression result is as follows:

Multiple linear regression result

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

Coefficients:
(Intercept)         crim           zn        indus          nox
 -19.615259    -0.132538     0.022103    -0.014980     0.010643
         rm          age          tax
   7.606508    -0.023198    -0.009006
```

The multiple linear regression is:

$$medv = -19.62 - 0.13crim + 0.02zn - 0.01indus + 0.1nox + 7.61rm - 0.02age - 0.01tax$$

And the simple t-test results on the parameters are as follows:

Test on coefficients result

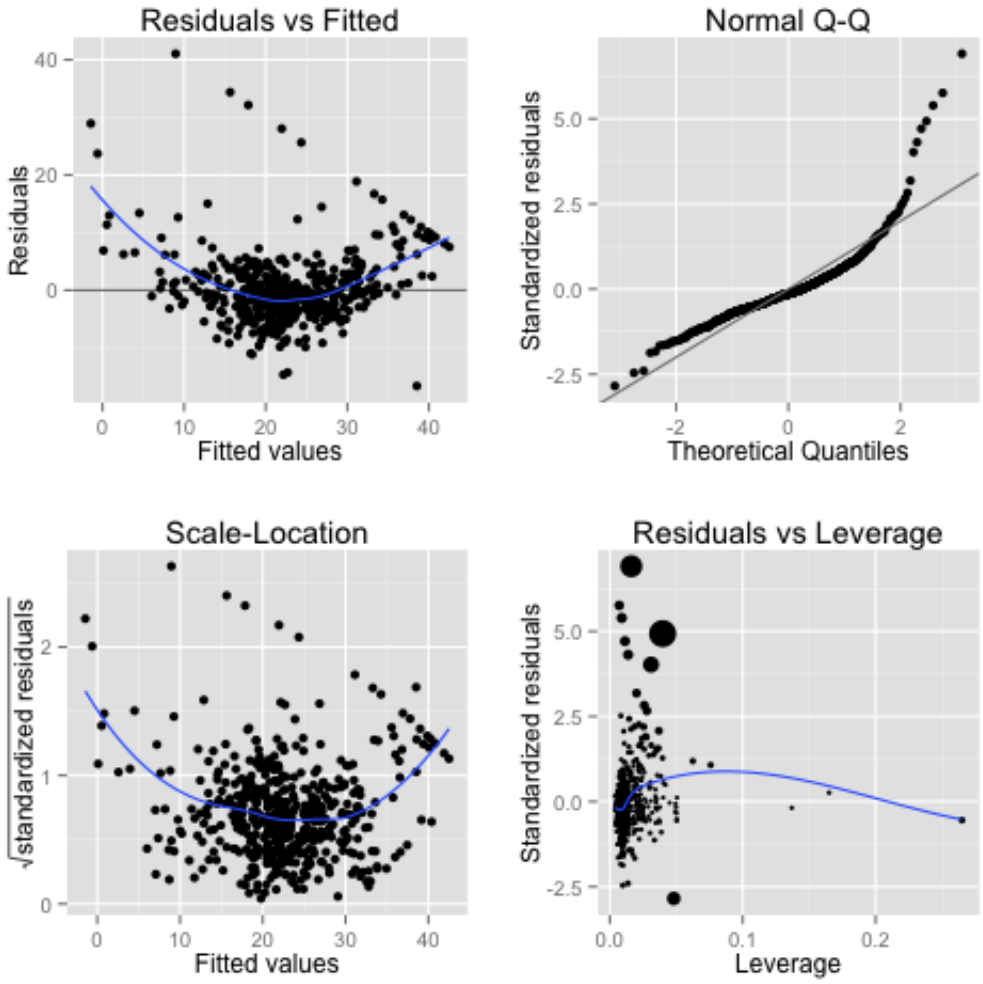
```
Call:
lm(formula = medv ~ crim + zn + indus + nox + rm + age + tax,
    data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-16.625  -3.161  -0.833   2.089  41.042

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.615259   3.221482  -6.089 2.27e-09 ***
crim         -0.132538   0.038482  -3.444 0.000621 ***
zn            0.022103   0.014823   1.491 0.136547
indus        -0.014980   0.072282  -0.207 0.835909
nox           0.010643   4.230468   0.003 0.997994
rm            7.606508   0.418424  18.179  < 2e-16 ***
age          -0.023198   0.014893  -1.558 0.119964
tax          -0.009006   0.002662  -3.384 0.000772 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Residual standard error: 5.989 on 498 degrees of freedom
Multiple R-squared:  0.5818,Adjusted R-squared:  0.576
F-statistic: 98.99 on 7 and 498 DF,  p-value: < 2.2e-16
```

Since from the result we can find that use t-test on whether the coefficient is 0, there are 4 variables that are not significant under 5% significance level, including zn, indus, nox, and age. The $R^2 = 0.5818$ and the adjusted $R^2 = 0.576$, which is not very good. The following is the plot of the residuals and fitted values:

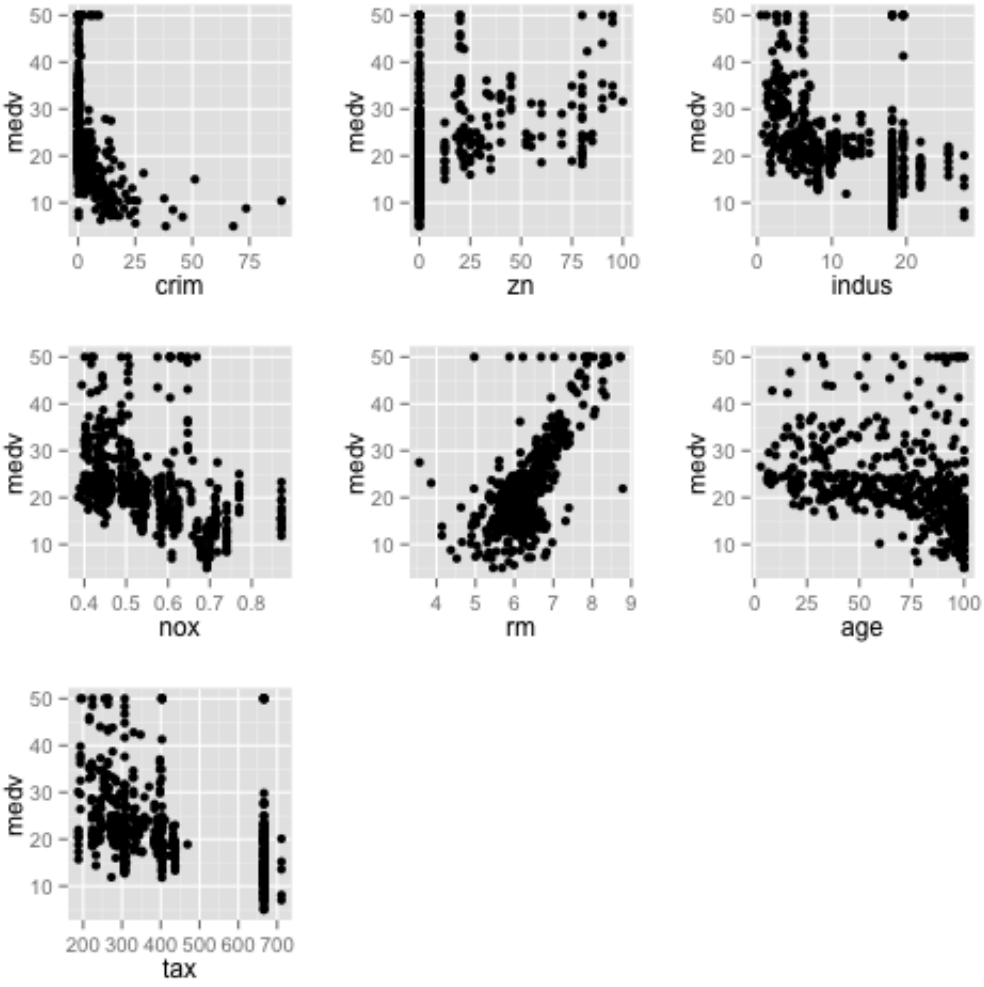


(b). State and assess the validity of the underlying assumptions, and suggest remedial measures in case of violations of any of the underlying assumptions.

- Linearity/functional form
- Normality
- Homoscedasticity
- Uncorrelated error

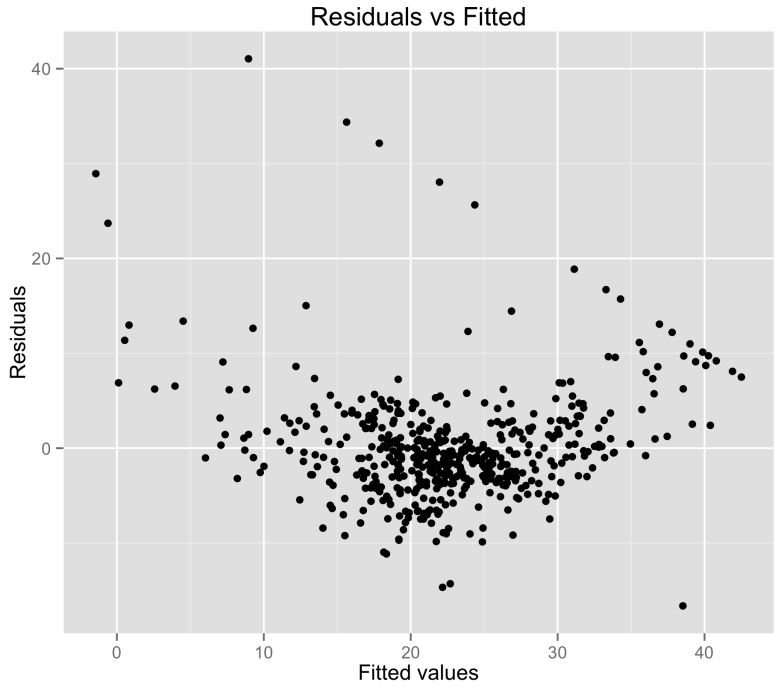
1. Linearity/functional form

(1). Plot the scatter plot of the response vs. explanatory variables.
From the scatter plot we can find that for the explanatory variable zn, indus, nox and tax, the scatter plot is a mess. Since the regression is not a straight line and is not monotonic, and the variablity of response is about the same at all values of x, then we should either try quadratic form or exclude the explanatory variables because the linear form doesn't make sense.



(2). Compute R^2.
Since $R^2 = 0.5818$, and adjust $R^2 = 0.576$, which is small, then this indicates that the model doesn't fit well on the data, i.e. the response data and explanatory data are not well-fitted linear association.

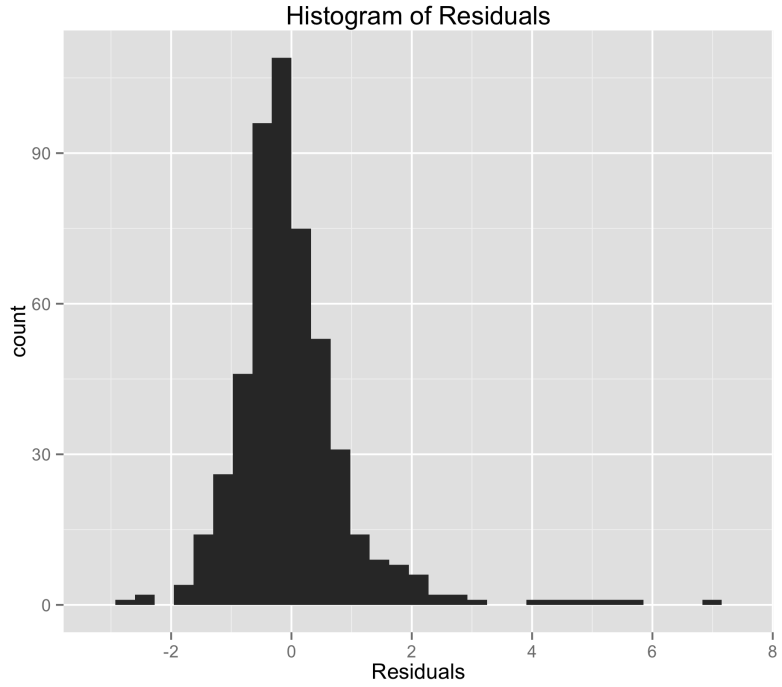
(3). Plot the residuals vs. fitted values.
Since from the plot we can see that most of the residuals are around x-line, except several outliers, and the outliers are extremely large, which affect much on the regression line. And this indicates that the regression line is not strictly a linearity/functional form. And the residuals are horn-shaped, then we should do some transform on y, such as logarithm and take reciprocal.



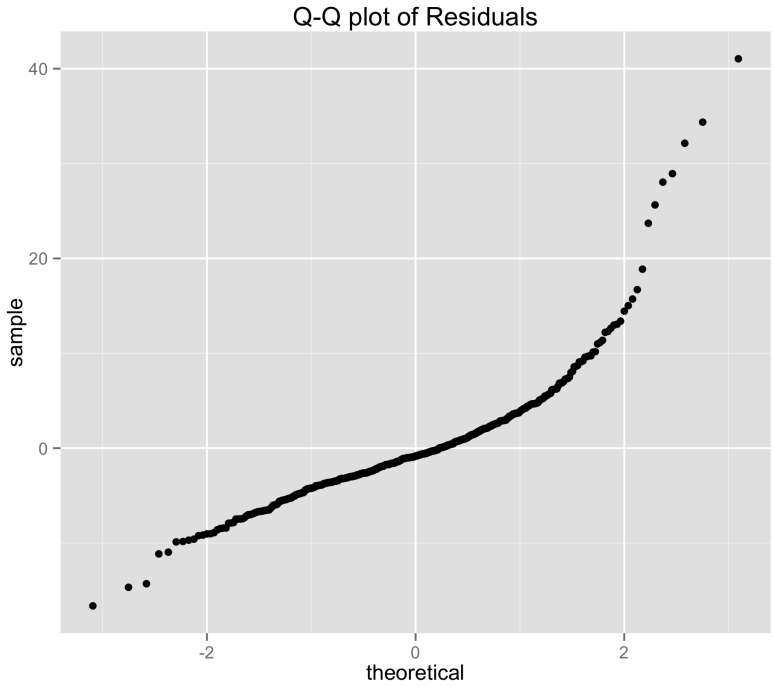
(4). Remedies
Since the three results conclude that the response and explanatory variables are not linearly form relation. So we should transform the data and exclude some variables, and we can also try some non-linear form on the data or try some new predictors.

2. Normality

(1). Plot the histogram of residuals
Since from the histogram of the residuals we can see that the residuals are skew to the right and there are some outliers in the plot. So these conclude that the residuals are not strictly satisfying normal distribution.



(2). Plot the Q-Q plot of standardized residuals
From the Q-Q plot of the residuals we can see that the residuals are not located close around the $y = x$ line, in fact, most of the residuals depart far from that line. So we can conclude that the residuals are not normally distributed.



(3). Shapiro test on residuals
The hypothesis of shapiro test is :

$$H_0 : residuals \sim Normal$$

$$H_1 : residuals\ not\ \sim Normal$$

Since from the result of shapiro test, the $p-value < 2.2*10^-16$, then we should reject the null, and conclude that the residuals are not satisfying normal distributions.

Shapiro test on residuals result

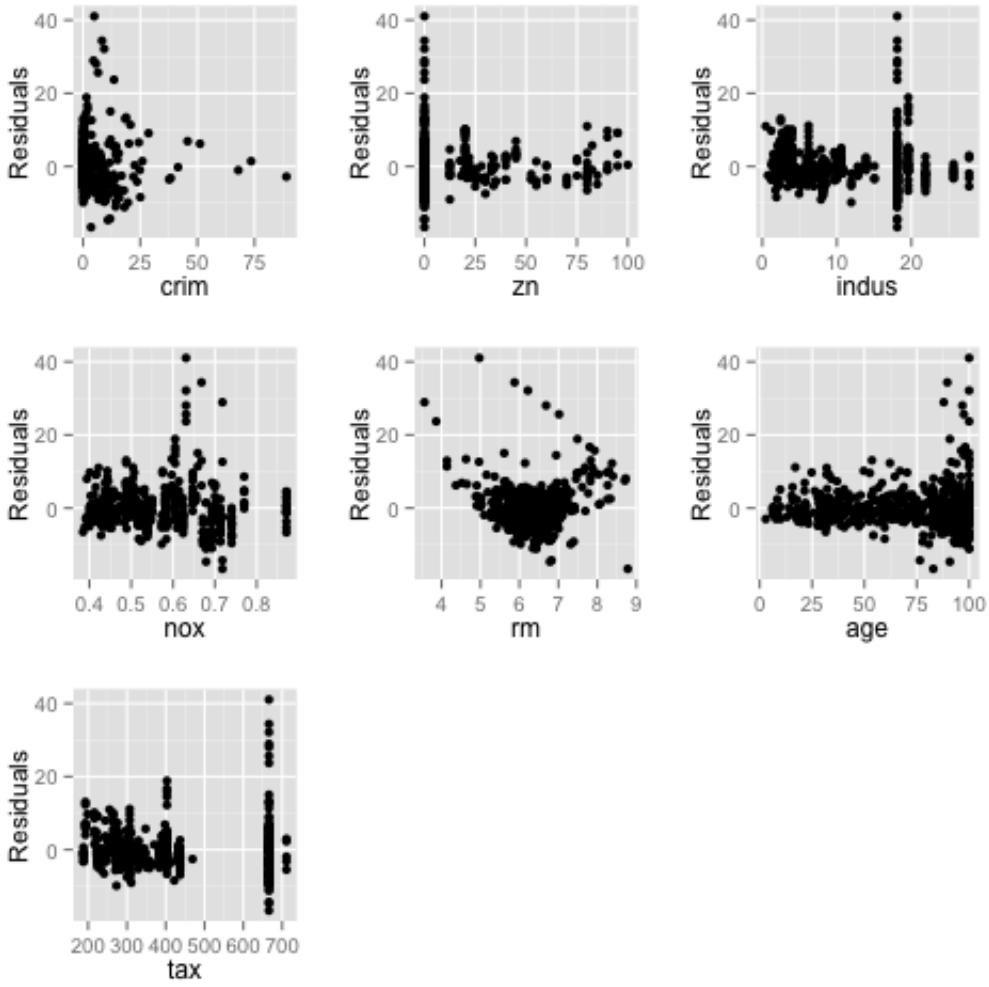
```
Shapiro-Wilk normality test

data:  residuals
W = 0.83945, p-value < 2.2e-16
```

(4). Remedies
Since the three results conclude that the data are not satisfying normal distribution. Then we should take transform on data and also try some robust regression method.

3. Homoscedasticity

(1). Plot the scatter plot of residuals vs. explanatory variables.
From the plot we can see that none of the 7 explanatory variables are satisfying homoscedasticity, all of them are not constant variance around the line.



(2). Remedies
Since the plot result conclude that the data are not satisfying homoscedasticity. So we should transform the data. And also, we can try weighted least sum of squares since the data residuals are non-constancy variance, we can assemble different weight on them and solve the problem.

4. Uncorrelated error

(1). Durbin-Watson Test
The hypothsis of dw test are:

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

From the dw test result, we can see that $p - value = 0$, which indicates that there are correlated errors in this regression.

Durbin-Watson Test Result

```
 lag Autocorrelation D-W Statistic p-value
   1       0.6326847     0.7288349       0
 Alternative hypothesis: rho != 0
```

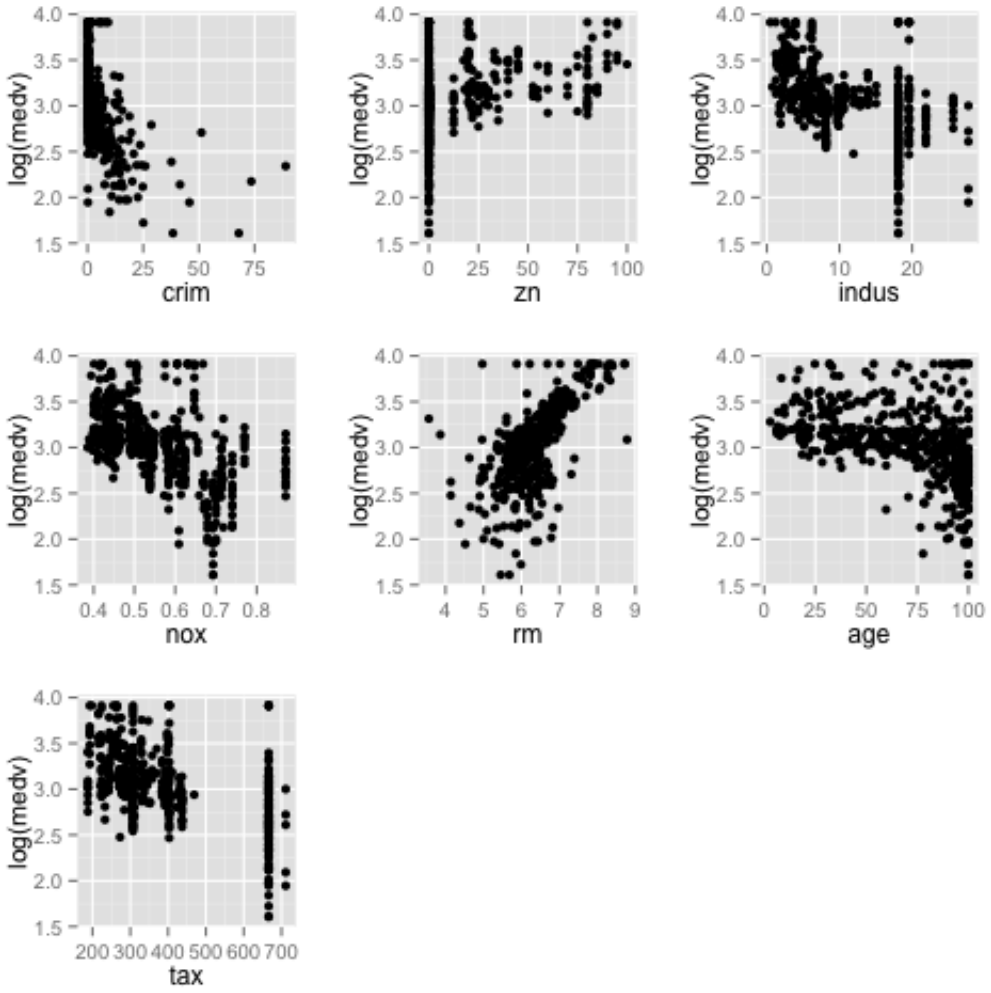
(2). Remedies

Since there are correlated errors, we should use Cochrane-Orcutt Procedure to do transform on data, and also use some models that incorporate the correlation structure such as Generalized Estimating Equations, to avoid the correlated errors.

5. Remedies to take

- non-linearility

For response, a horn-shaped residuals vs fitted plot indicates that we should transform on response, use logarithm. Plot new scatter plots on transformed response and explanatory variables. From the plot we can see some remedies on x.
For crim we should try reciprocal of y, because the regression is not a straight line and is monotonic, though some outliers and the variablity of response is decreasing with x increasing, then we don't need transform data
For zn, since there is no need to transform the data, because the regression is not a straight line and is not monotonic, and the variablity of response is not monotic with x increasing. Then we can try some quadratic form on zn.
For indus, since the regression is not a straight line and is monotonic decreasing, though some outliers and the variablity of response is not monotonic with x increasing, then we don't need transform data.
For nox, since the regression is not a straight line and is monotonic, though some outliers and the variablity of response is increasing with x increasing, then we can take the logarithm on x
For age, since the regression is not a straight line and is not monotonic, and the variablity of response is increasing with x increasing, then we can take some transform on x, take the reciprocal.
For tax, since the regression is not a straight line and is not monotonic, it is a mess and the variablity of response is about the same with x increasing, then we don't need transform on tax, i.e. exclude it from the model.



- non-normality

Since from the first part, we have took logrithm on the data, and this time we should try some robust regression method, such as least median of squares, not the ordinary least of sum of squares method.

- non-homoscedasticity

Since from the first part, we have took logrithm on the data, and the result is still not good.Then we should try weighted method on the data, since non-constancy-variance, then we just assign weights to make sure the weighted variance equal.

- correlated-errors

Use models that incorporate the correlation structure such as Generalized Estimating Equations.

(c). Repeat (a) using Least Median of Squares Regression and compare the results with those obtained in (a).

The Least Median of Squares regression result is as follows:

Multiple linear regression result

```
Call:
lqs.formula(formula = medv ~ crim + zn + indus + nox + rm + age +
    tax, data = Boston, method = "lms")
```

```
Coefficients:
(Intercept)         crim           zn        indus          nox
 -27.988943    -0.710493     0.025242     0.022510     3.699771
         rm          age          tax
   8.029181    -0.040069     0.001258

Scale estimates 3.623 3.344
```

The least median of squares regression is:

$$medv = -27.99 - 0.71crim + 0.03zn + 0.02indus + 3.70nox + 8.02rm - 0.04age - 0.001tax$$

Compare the two results from (a) and (c).

Table 1: The coefficients of least sum of square and least median of square

	intercept	crim	zn	indus	nox	rm	age	tax
sum of square	-19.62	-0.13	0.02	-0.01	0.1	7.61	-0.02	-0.01
median of square	-27.99	-0.71	0.03	0.02	3.70	8.02	-0.04	-0.001

From the table we can see that the coefficients are approximately close, but on intercept, crim, indus and tax, there are significantly differences between the two models.

R Code:

```
rm(list=ls())
library(MASS)
library(ggplot2)
#delete containing na value rows
Boston <- na.omit(Boston)

multilinear <- lm(medv ~ crim+zn+indus+nox+rm+age+tax, data=Boston)
sink("/Users/Qianbo/Downloads/HW4/lsresult.txt")
multilinear
sink()
sink("/Users/Qianbo/Downloads/HW4/multilinear.txt")
summary(multilinear)
sink()

#plot the mulilinear regression
plot1<-ggplot(multilinear,aes(.fitted,.resid))+
  geom_hline(yintercept=0,color="grey50",size=0.5)+
  geom_point()+geom_smooth(size=0.5,se=FALSE)+
  ggtitle("Residuals vs Fitted")+xlab("Fitted values")+ylab("Residuals")

plot2<-ggplot(multilinear,aes(sample=.stdresid))+
  stat_qq()+geom_abline(color="grey50")+
  ggtitle("Normal Q-Q")+
  ylab("Standardized residuals")+xlab("Theoretical Quantiles")

plot3<-ggplot(multilinear,aes(.fitted,sqrt(abs(.stdresid))))+
  geom_point()+geom_smooth(se=FALSE)+
  ggtitle("Scale-Location")+
  xlab("Fitted values")+ylab(expression(sqrt("standardized residuals")))

plot4<-ggplot(multilinear,aes(.hat,.stdresid,size=.cooksd))+
  geom_point()+geom_smooth(se=FALSE,size=0.5)+
  ggtitle("Residuals vs Leverage")+scale_size_continuous(guide=FALSE)+
  xlab("Leverage")+ylab("Standardized residuals")

library(grid)
library(gridExtra)

grid.arrange(plot1,plot2,plot3,plot4,ncol=2)
png("/Users/Qianbo/Downloads/HW4/multilinear.png")
grid.arrange(plot1,plot2,plot3,plot4,ncol=2)
dev.off()

#check assumptions
#check Linear/functional form
#scatter plot of y vs x
s1<-ggplot(Boston,aes(crim,medv))+geom_point()
s2<-ggplot(Boston,aes(zn,medv))+geom_point()
s3<-ggplot(Boston,aes(indus,medv))+geom_point()
s4<-ggplot(Boston,aes(nox,medv))+geom_point()
s5<-ggplot(Boston,aes(rm,medv))+geom_point()
s6<-ggplot(Boston,aes(age,medv))+geom_point()
s7<-ggplot(Boston,aes(tax,medv))+geom_point()

png("/Users/Qianbo/Downloads/HW4/scatter1.png")
grid.arrange(s1,s2,s3,s4,s5,s6,s7,ncol=3)
dev.off()

#scatter plot of residuals vs fitted
residplot<-ggplot(multilinear,aes(.fitted,.resid))+
  geom_point()+ggtitle("Residuals vs Fitted")+
  xlab("Fitted values")+ylab("Residuals")
ggsave('/Users/Qianbo/Downloads/HW4/resid.png')

#check Normality
#histogram of residuals
```

```
hist<-ggplot(multilinear,aes(.stdresid))+
  geom_histogram()+ggtitle("Histogram of Residuals")+xlab("Residuals")
hist
ggsave( file ="/Users/Qianbo/Downloads/HW4/residhist.png")

#qq plot of standardized residuals
qqresid<-ggplot(multilinear,aes(sample=.stdresid))+
  stat_qq()+geom_abline(aes(intercept=0,scope=1),color="grey50",size=0.5)+
  ggtitle ("Q-Q plot of std Residuals")
qqresid
ggsave( file ="/Users/Qianbo/Downloads/HW4/residqq.png")

#shapiro test on residuals
residuals = resid( multilinear )
shapiro.test( residuals )
sink("/Users/Qianbo/Downloads/HW4/shapiro.txt")
shapiro.test( residuals )
sink()

#check Homoscedasticity
#scatter plot of residuals vs x
r1<-ggplot(multilinear,aes(crim,.resid))+geom_point()+ylab("Residuals")
r2<-ggplot(multilinear,aes(zn,.resid))+geom_point()+ylab("Residuals")
r3<-ggplot(multilinear,aes(indus,.resid))+geom_point()+ylab("Residuals")
r4<-ggplot(multilinear,aes(nox,.resid))+geom_point()+ylab("Residuals")
r5<-ggplot(multilinear,aes(rm,.resid))+geom_point()+ylab("Residuals")
r6<-ggplot(multilinear,aes(age,.resid))+geom_point()+ylab("Residuals")
r7<-ggplot(multilinear,aes(tax,.resid))+geom_point()+ylab("Residuals")

grid.arrange(r1,r2,r3,r4,r5,r6,r7,ncol=3)
png("/Users/Qianbo/Downloads/HW4/residx.png")
grid.arrange(r1,r2,r3,r4,r5,r6,r7,ncol=3)
dev.off()

#check Corrected errors
#dw test
library(car)
durbinWatsonTest(multilinear)
sink("/Users/Qianbo/Downloads/HW4/dwtest.txt")
durbinWatsonTest(multilinear)
sink()
#remedies
#transform on data
n1<-ggplot(Boston,aes(crim,log(medv)))+geom_point()
n2<-ggplot(Boston,aes(zn,log(medv)))+geom_point()
n3<-ggplot(Boston,aes(indus,log(medv)))+geom_point()
n4<-ggplot(Boston,aes(nox,log(medv)))+geom_point()
n5<-ggplot(Boston,aes(rm,log(medv)))+geom_point()
n6<-ggplot(Boston,aes(age,log(medv)))+geom_point()
n7<-ggplot(Boston,aes(tax,log(medv)))+geom_point()

png("/Users/Qianbo/Downloads/HW4/transform.png")
grid.arrange(n1,n2,n3,n4,n5,n6,n7,ncol=3)
dev.off()

# use least median square method do linear regression

lmslinear<-lmsreg(medv ~ crim+zn+indus+nox+rm+age+tax, data=Boston)
sink("/Users/Qianbo/Downloads/HW4/lmsresult.txt")
lmslinear
sink()
```