Name: Jiahong Hu
UNI: jh3561
HW4
STAT 4201

Problem a

```
> fit1<-lm(medv~crim+zn+indus+nox+rm+age+tax,data=data)
> summary(fit1)

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

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
```

Figure 1

I denote X1 as crim, X2 as zn, X3 as indus, X4 as nox, X5 as rm, X6 as age, X7 as tax, Y as medv.

The model:

$$Y = -19.615259 - 0.132538X1 + 0.022103X2 - 0.01498X3 + 0.010643X4 + 7.606508X5 - 0.023198X6 - 0.009006X7$$

Problem b

Assumption and validation

1. Linearity / Function Forms
    a. R^2

R^2 is used to measure linear association of the Y and X. As shown in the Figure of problem a, R^2 = 0.5815, which indicate some degree of linear association between respond variable and 7 explanatory variables, though not very strong.

    b. Plot of residuals VS fitted values

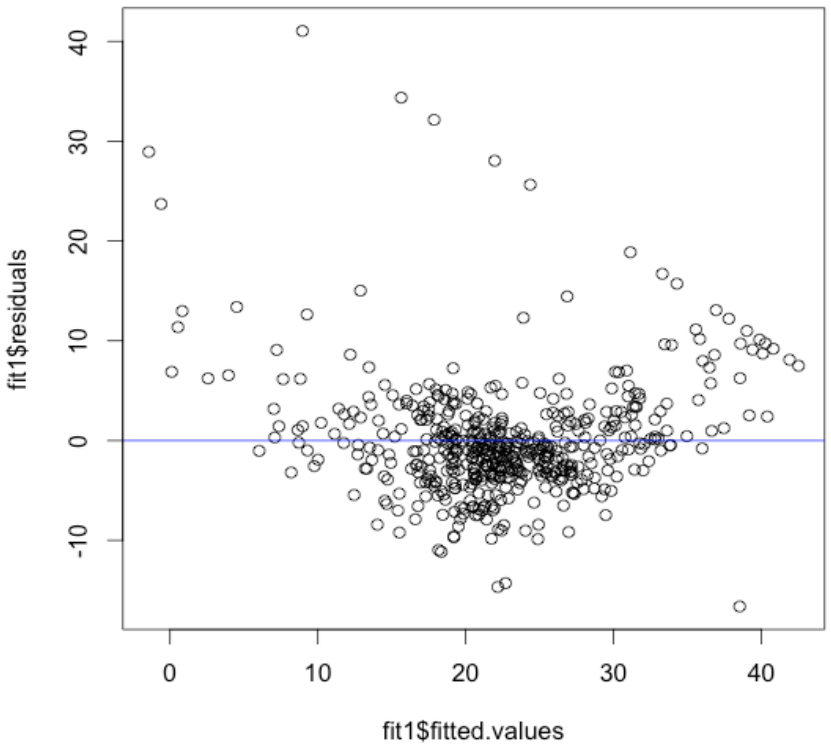
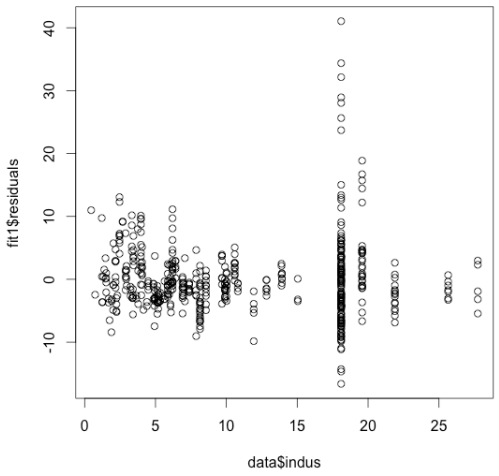


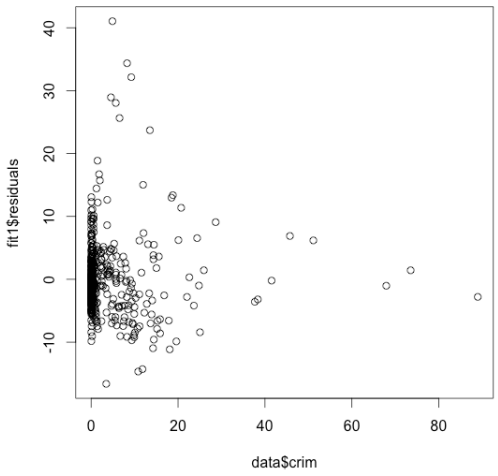
Figure 2

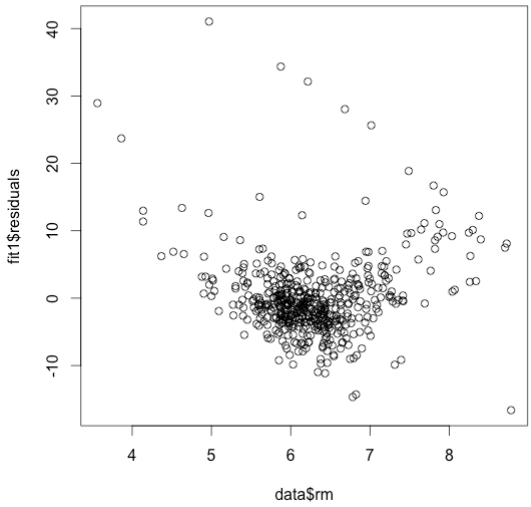
Figure 2 shows that most residuals is randomly distributed around zero and is dependent with fitted values but there are some residuals in the top of the figure are very far away from zero and has a abnormal decreasing linear pattern, which is a indicator that the function may not be appropriate.

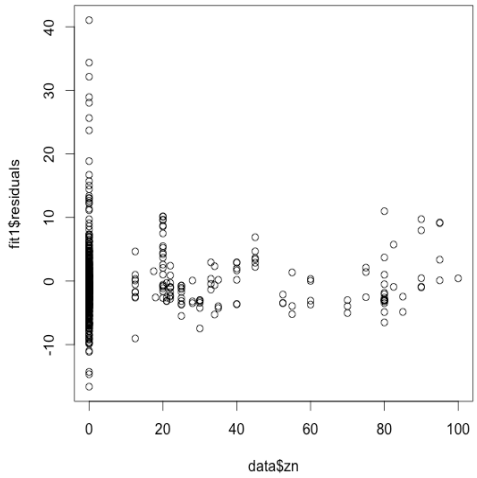
c. Plot of residual VS every explanatory variable.

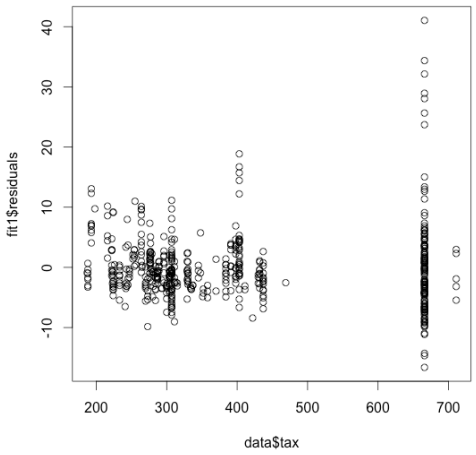
Those plots provide further information about the adequacy of the regression function with respect to the predictor variables. (e.g. whether a curvature effect is required for that variable)

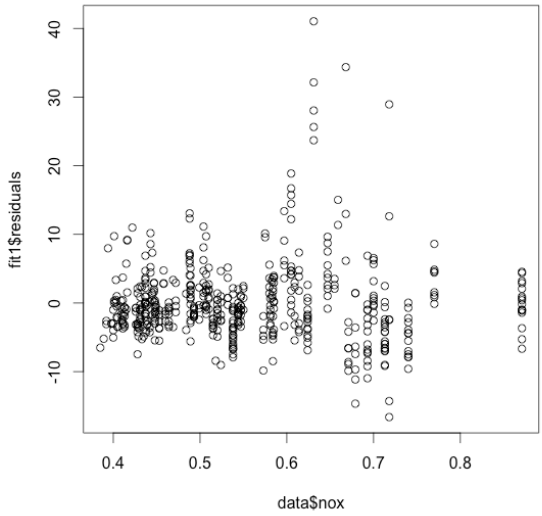


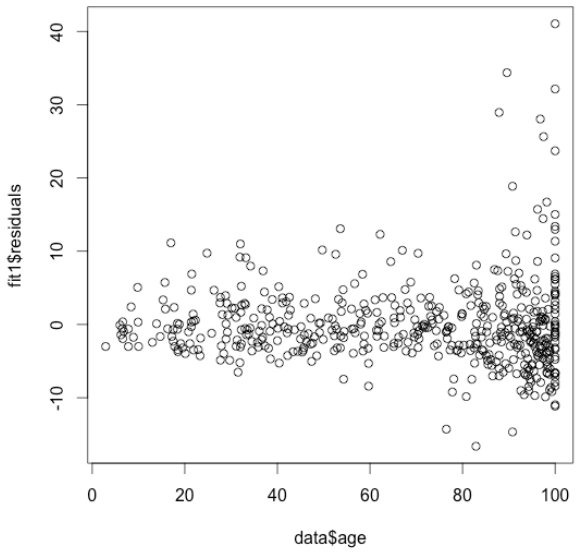












The seven figures show that linear function is suitable for the seven predictions because most residuals seem random though with several large residuals far away from zero. This might because there exit outliers or the variance of Y is not full explained by those 7 predictors.

Remedies:

- simple transformation
- non-linear model
- include other predictors

Because there is not strong evidence that the model should be non-linear. Hence, I decide to add more variables to the model.

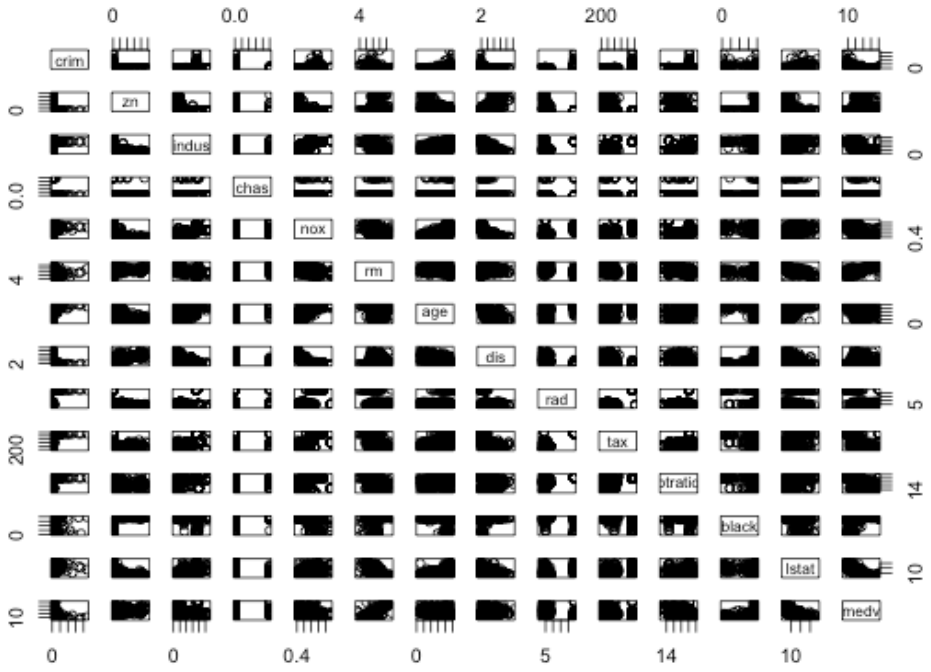
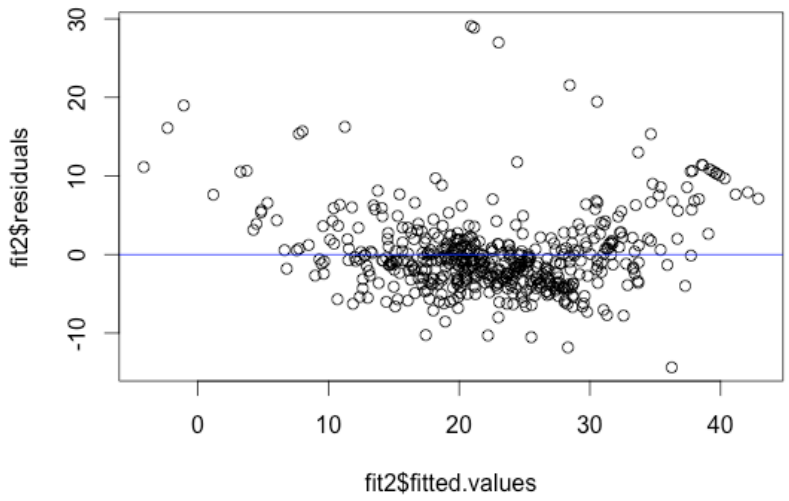


Figure 3

The Figure 3 shows that there might be relationship between “lstat”, “black”, “ptratio” and “medv” and I add those three variables into the model.

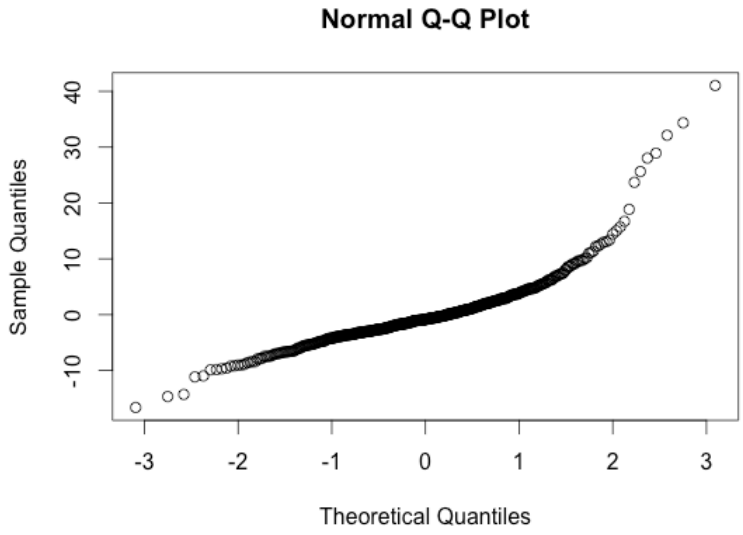
```
> fit2<-lm(medv~crim+zn+indus+nox+rm+age+tax+black+lstat+ptratio,data=data)
> summary(fit2)

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

Residuals:
    Min      1Q  Median      3Q     Max
-14.360  -3.095  -0.792   1.713  29.087

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.8274022  4.9209901   3.216  0.00138 **
crim        -0.0402022  0.0341022  -1.179  0.23902
zn          -0.0019530  0.0135372  -0.144  0.88535
indus        0.0544431  0.0624027   0.872  0.38339
nox         -6.3705903  3.9251218  -1.623  0.10522
rm           4.6196581  0.4430956  10.426  < 2e-16 ***
age          0.0270992  0.0136625   1.983  0.04787 *
tax          0.0006174  0.0025089   0.246  0.80572
black        0.0092797  0.0029046   3.195  0.00149 **
lstat       -0.5415893  0.0548906  -9.867  < 2e-16 ***
ptratio     -0.9644808  0.1384520  -6.966 1.04e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.151 on 495 degrees of freedom
Multiple R-squared:  0.6925,    Adjusted R-squared:  0.6863
F-statistic: 111.5 on 10 and 495 DF,  p-value: < 2.2e-16
```

Now, R ^2 = 0.6925, the linear relationship is stronger.



The residuals seem more random and are closer to zero.

2. Normality



It has a heavy right tail so may be skewed.

Ho : residuals are from normal distribution
H1: residuals are not from normal distribution

```
> ks=ks.test(fit1$residuals,"pnorm")
> ks

        One-sample Kolmogorov-Smirnov test

data:  fit1$residuals
D = 0.36196, p-value < 2.2e-16
alternative hypothesis: two-sided
```

P <0.05, therefore reject H0. The residuals are not normal

Remedies:

- Transformation
- Robust linear regression

I will apply robust linear regression in problem c

3. Homoscedastic (consistency of error variance)

From the plots of residuals vs explanatory variables above, the error variance seems not very consistent and it changes with X.

We use Bartlett Test.

H0: the variances of each group are the same.

H1: the variances of each group are not the same.

```
> v1=fitted(fit1)
> g=rep(1,506)
> g[201:506]=0
> bartlett.test(fitted(fit1),g)

        Bartlett test of homogeneity of variances

data:  fitted(fit1) and g
Bartlett's K-squared = 38.9243, df = 1, p-value = 4.406e-10
```

P-value is nearly 0 which indicates that the variances of each group are not the same.

Corrective measures

1. Transformation
2. Build variance structure into model

4. Uncorrelated Error

We run the durbin-watson test

H0: there is no correlation among residuals

H1: there is correlation among residuals

```
> durbinWatsonTest(fit1)
 lag Autocorrelation D-W Statistic p-value
   1       0.6326847     0.7288349       0
 Alternative hypothesis: rho != 0
>
```

The p-value is small, therefore , reject H0. There is correlation among residuals.

Corrective:

1. Transformation: Cochrane-Orcutt Procedure
2. Use models that incorporate the correlation structure.

Problem c

```
> library(MASS)
> fit3<-lmsreg(medv~crim+zn+indus+nox+rm+age+tax,data=data)
>
>
> fit3$coefficient
  (Intercept)          crim            zn         indus           nox            rm
-4.700874e+01 -1.237810e+00  1.019253e-02  7.226171e-02  1.593185e+01  1.030533e+01
          age           tax
-4.968688e-02  2.075212e-04
>
```

New

$$Y=-0.5879X1-0.02990X2-0.04771X3+8.7155X4+8.1355X5-0.078215X6-0.0075426X7-24.0147$$

old

$$Y=-19.615259-0.132538X1+0.022103X2-0.01498X3+0.010643X4+7.606508X5-0.023198X6-0.009006X7$$

The differences are large for both.