HW3

Hanao Li

September 25, 2019

Question 1

a)

```
if(!require("pacman")) install.packages("pacman")
```

```
## Loading required package: pacman
```

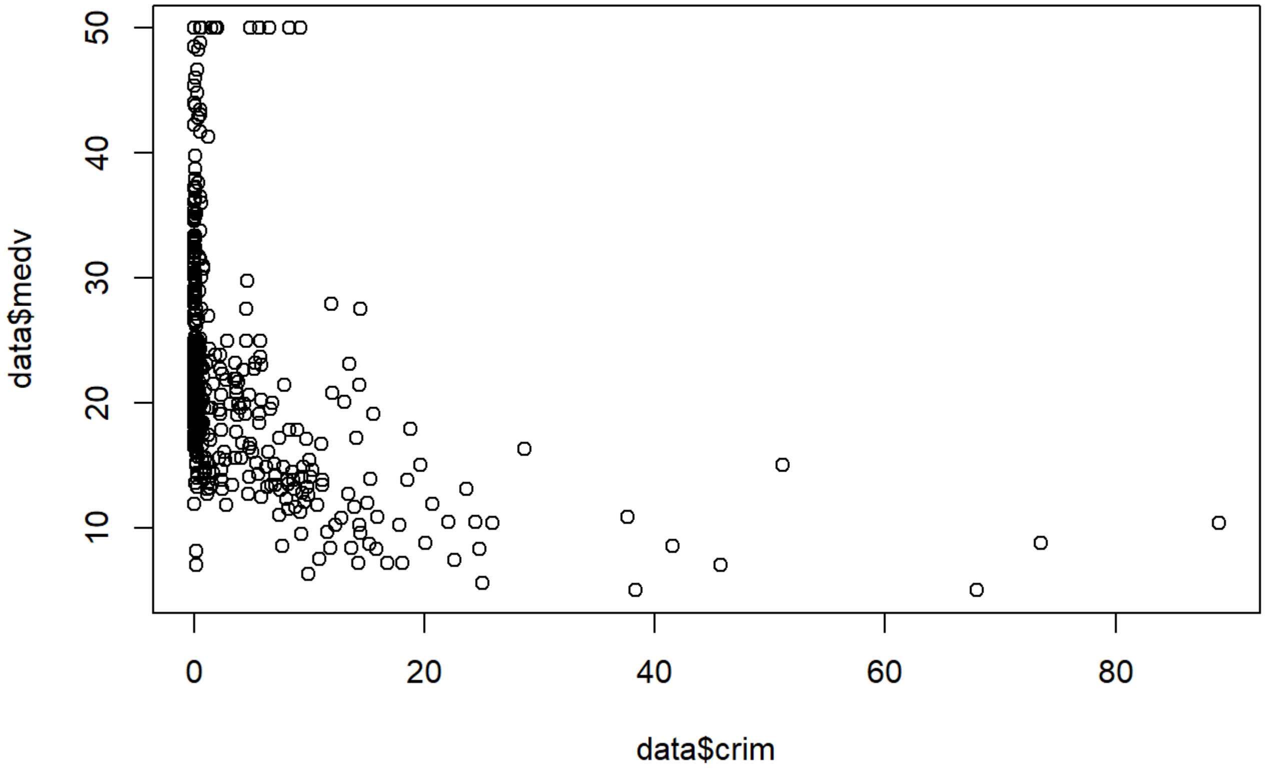
```
p_load(MASS, car)

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

```
##
## 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
```

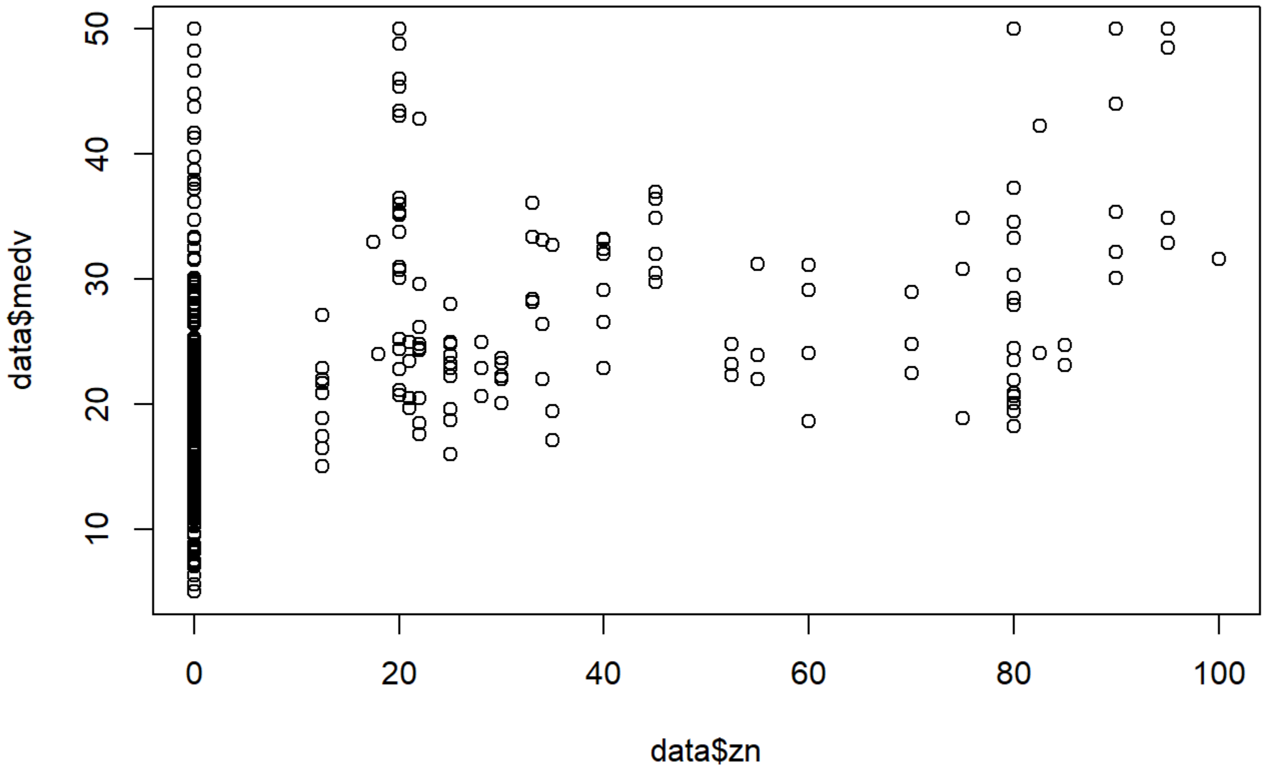
```
# Our regression model is medv = -19.615 -0.133crim + 0.022zn - 0.015indus + 0.011nox + 7.607rm
 - 0.023age - 0.009tax. Only crim, rm and tax are significant according to the summary and our a
djusted r suqared value is only 0.576 so our model is not a very good model.
```

b)

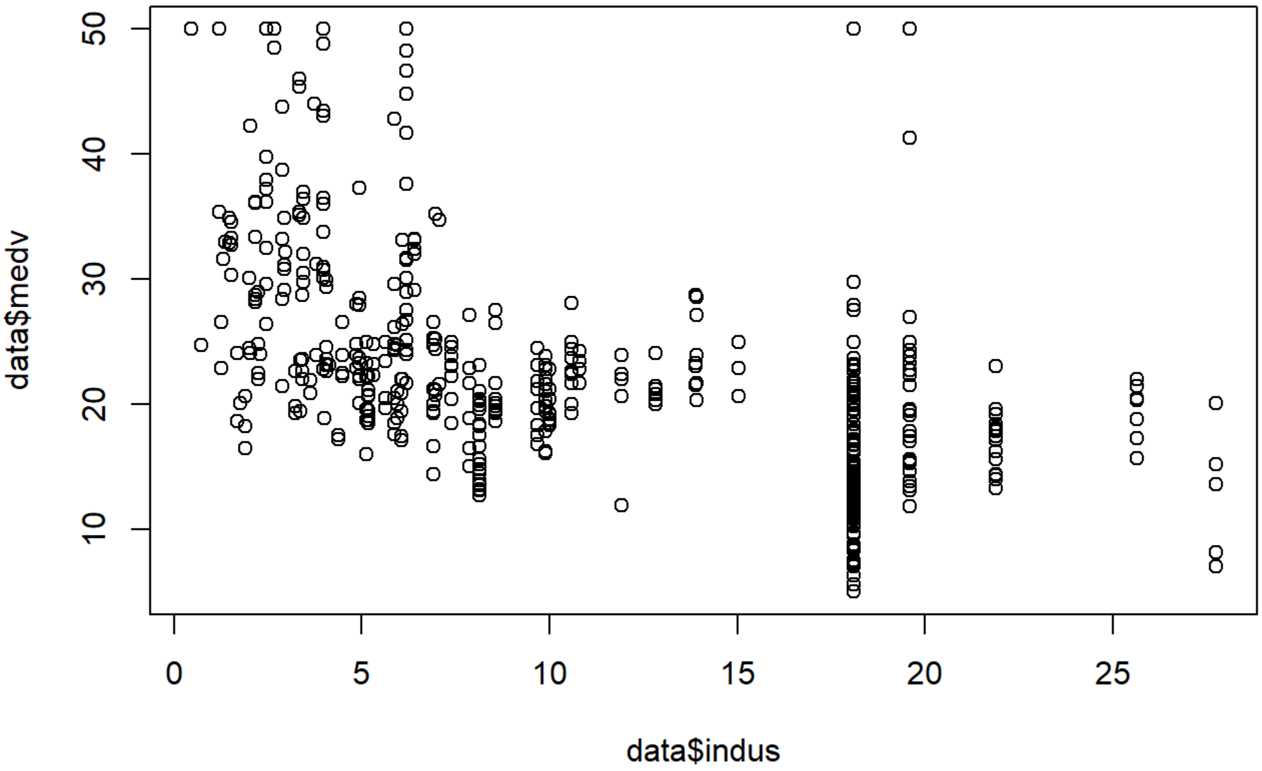
```
# Linearity / functional form
plot(data$crim, data$medv)
```



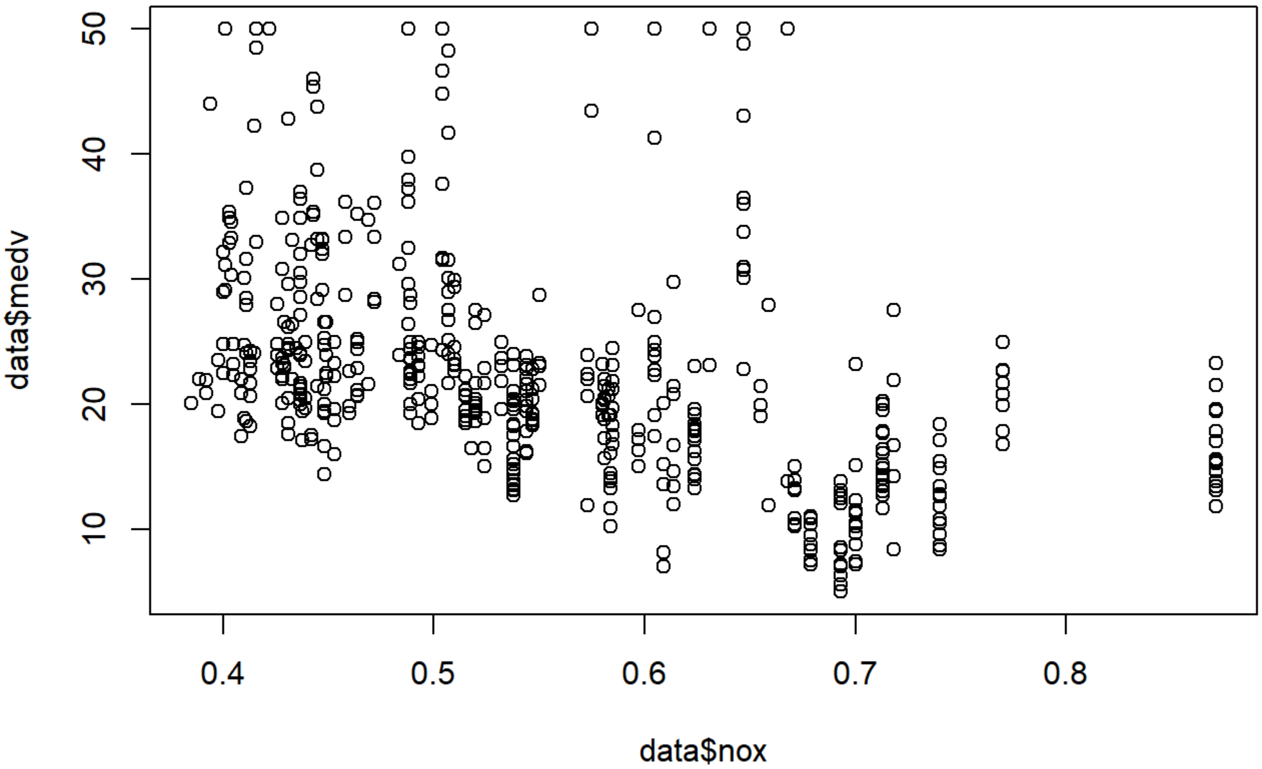
```
plot(data$zn, data$medv)
```



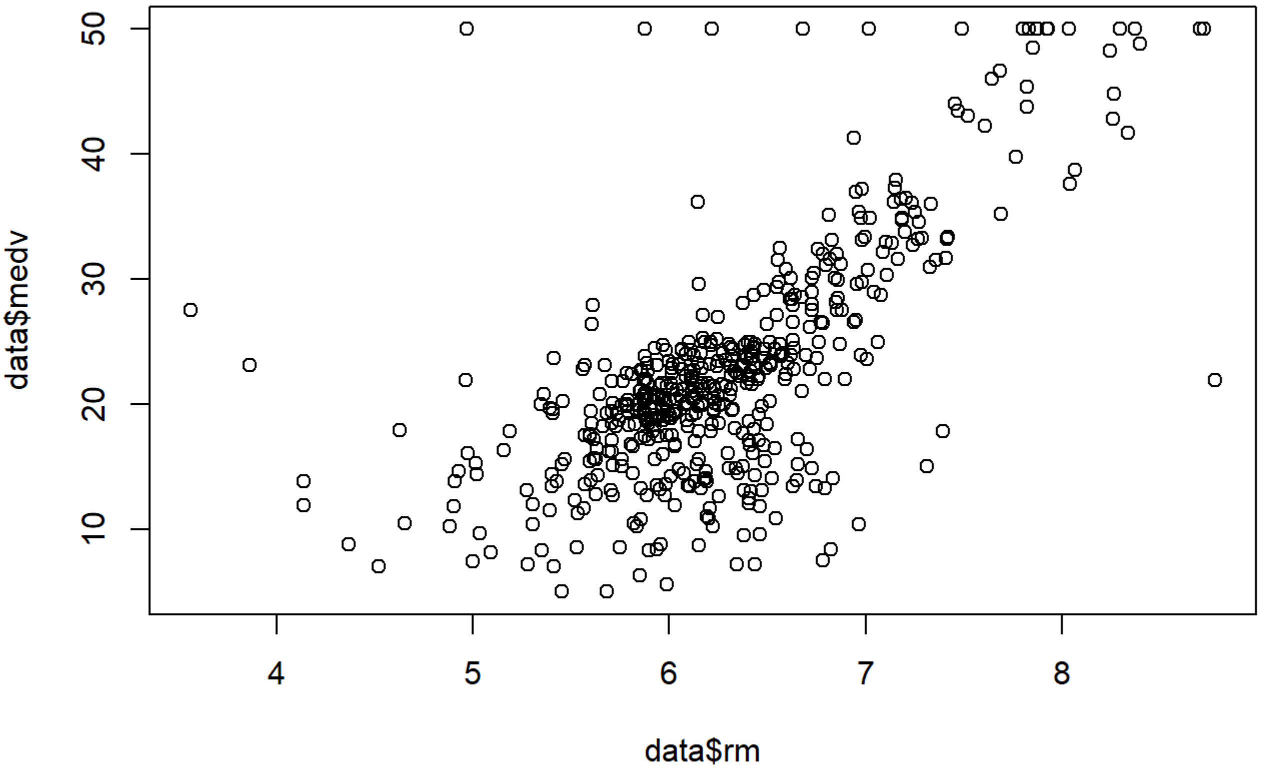
```
plot(data$indus, data$medv)
```



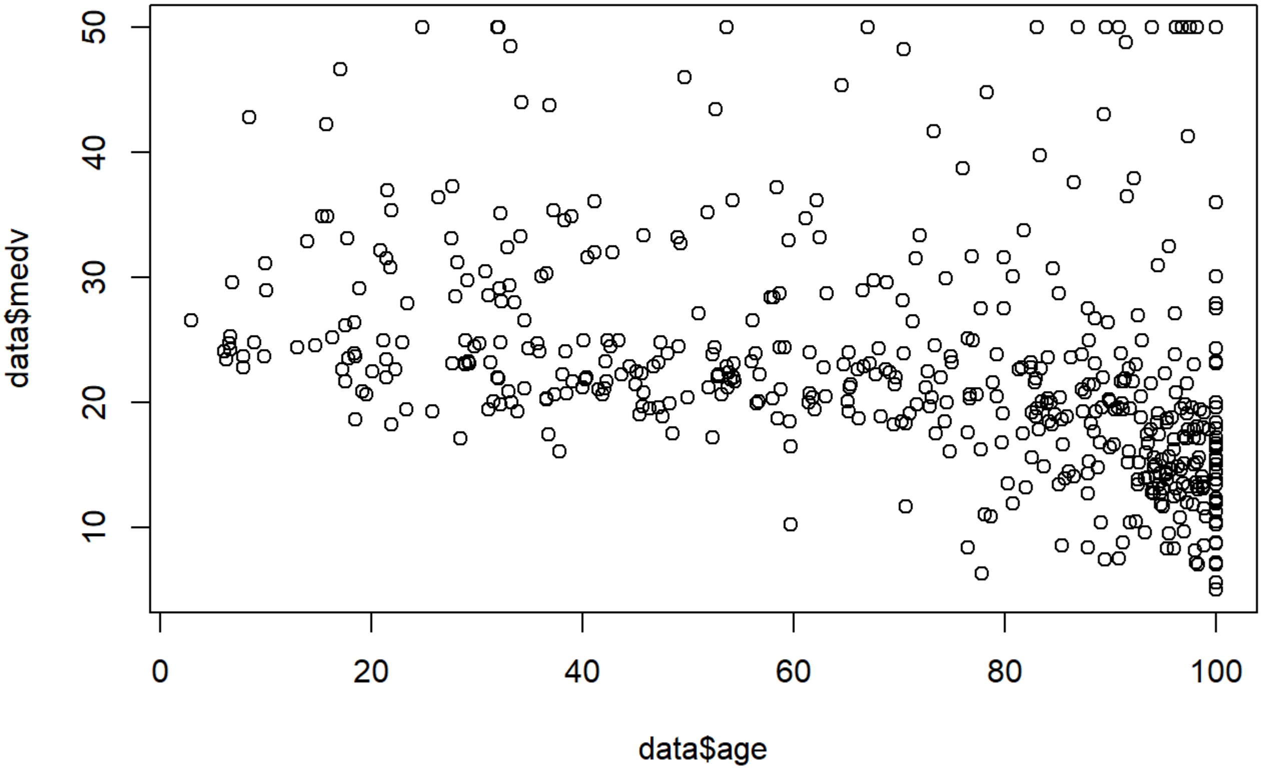
```
plot(data$nox, data$medv)
```



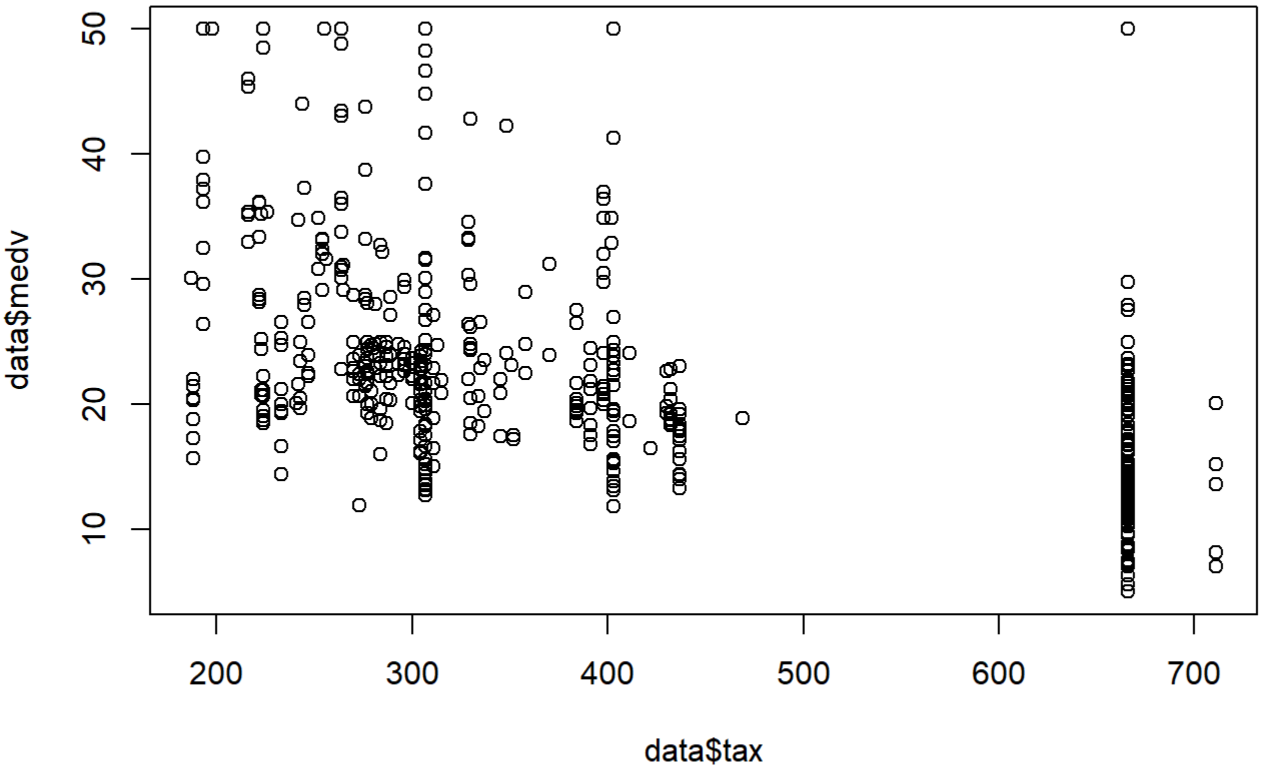
```
plot(data$rm, data$medv)
```



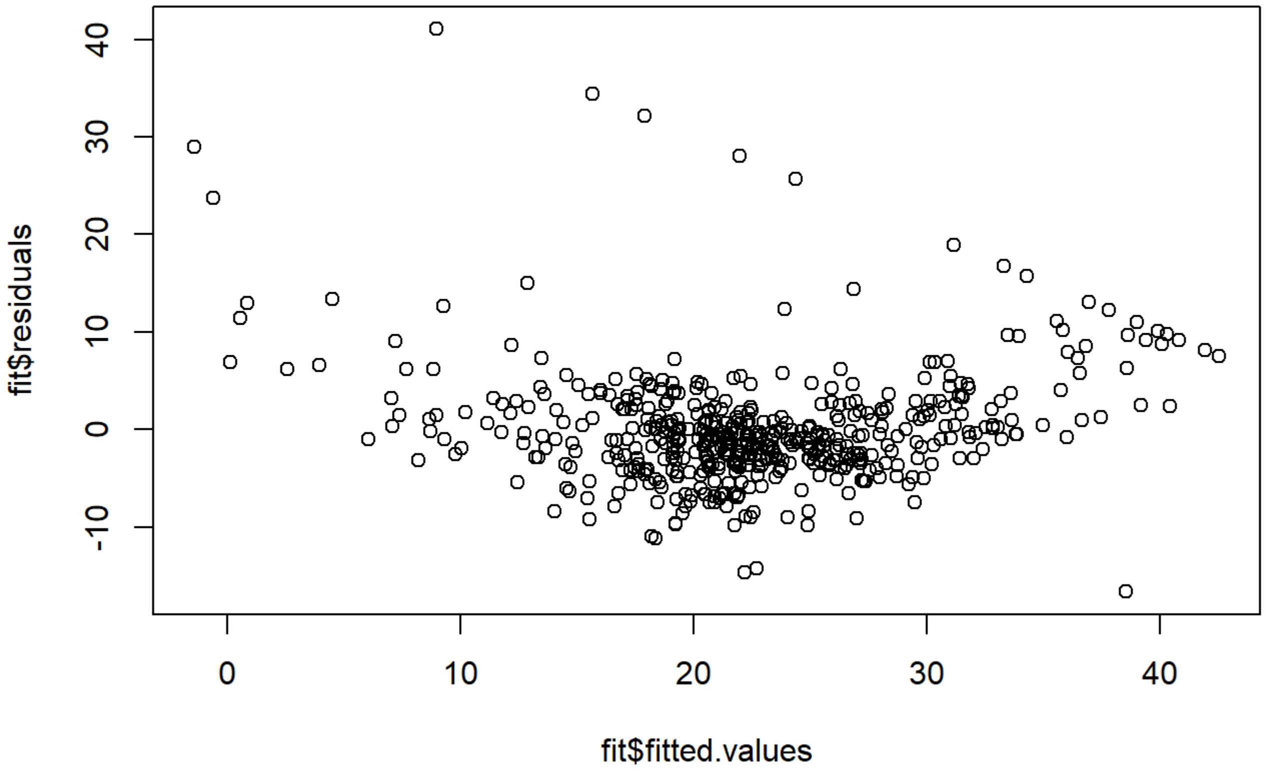
```
plot(data$age, data$medv)
```



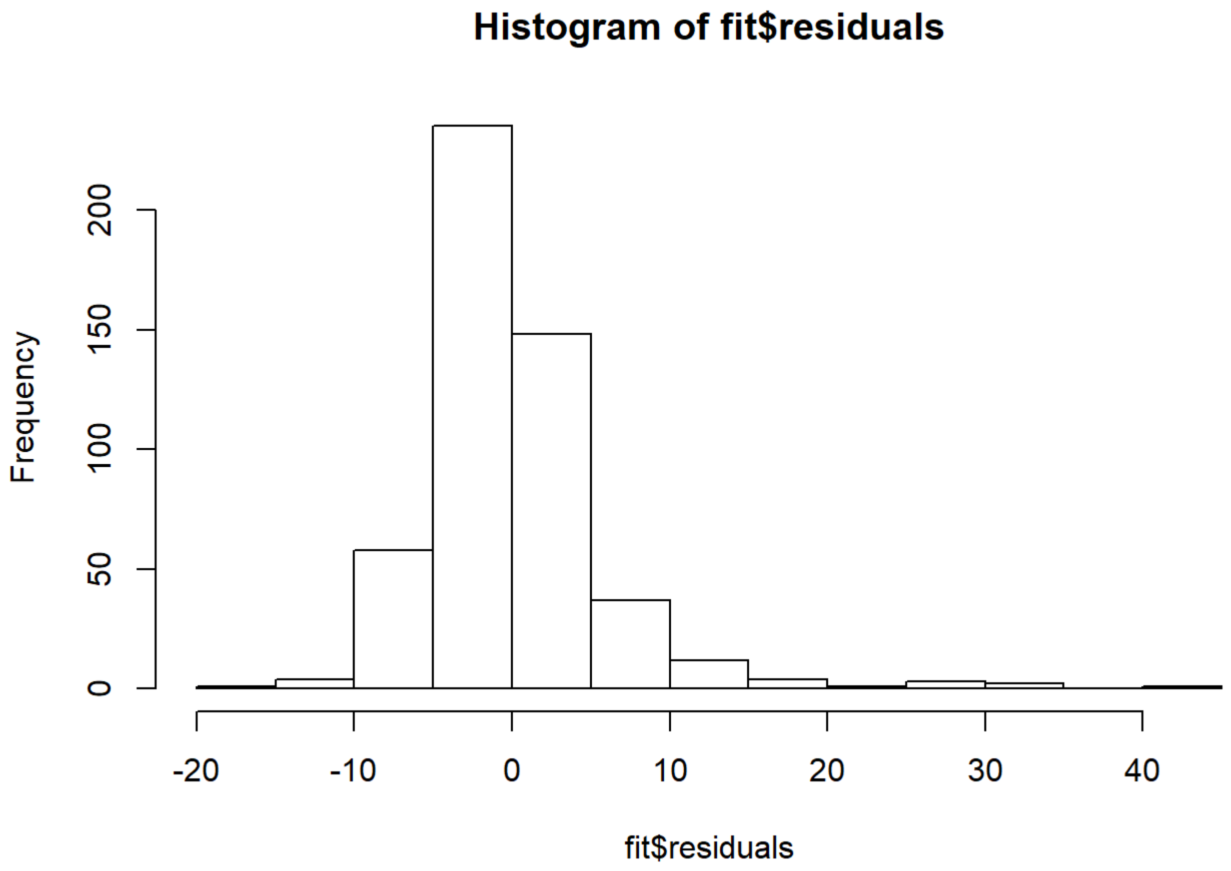
```
plot(data$tax, data$medv)
```



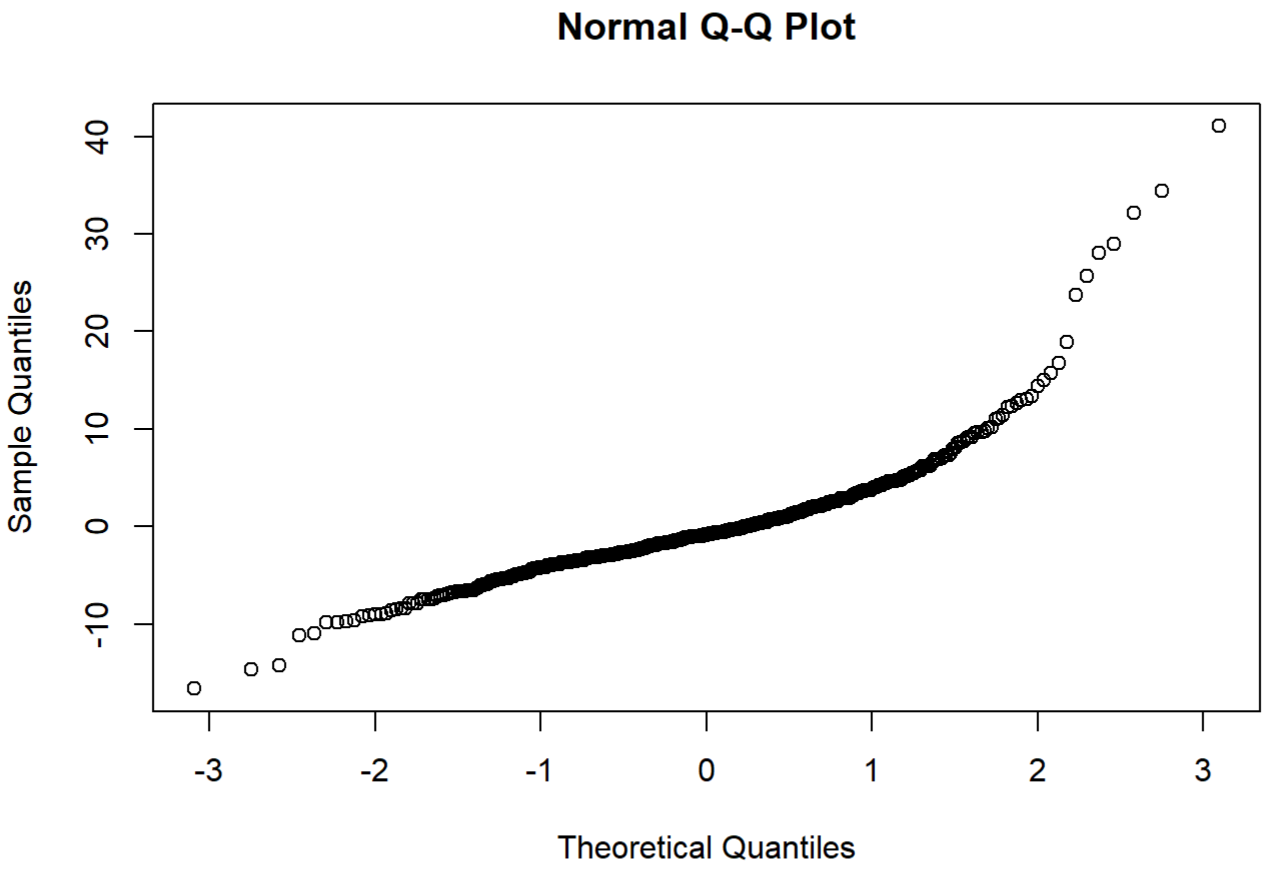
```
plot(fit$fitted.values, fit$residuals)
```



```
# From the scatter plots, we could see that only the variable rm has some kinds of linear relati
onship with the response variable. From the residual against fitted value plot, we could find th
at the there are some outliers with residuals larger than 20 which could affect our linear regre
ssion line. This suggests the line should not be a linear model. And also from the adjusted R sq
uared value we computed in the previous question, it is only 0.576, so the linear regression doe
s not seem to be a good model. What we can do is to transform the data such as taking the logari
thm or the response, take the reciprocal of it or we could remove some noisy variables or add so
me new variables.


# Normality
hist(fit$residuals)
```



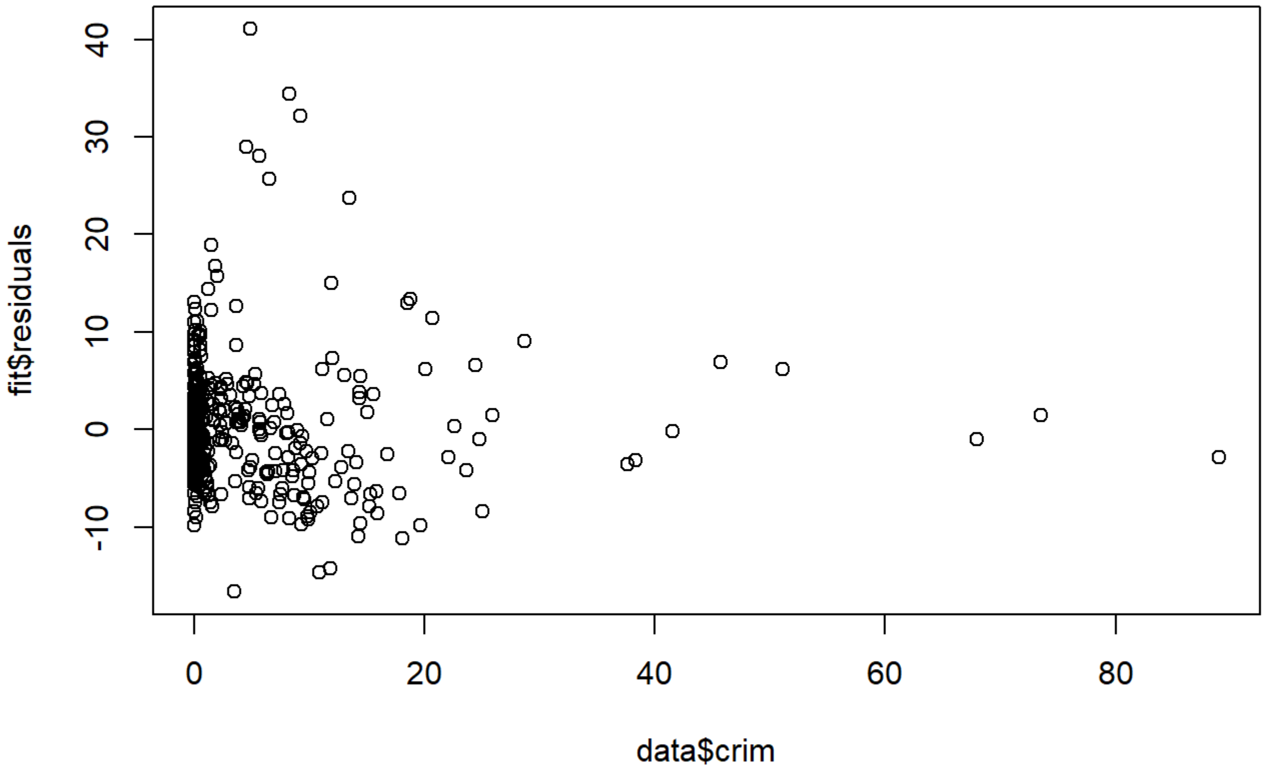
```
qqnorm(fit$residuals)
```



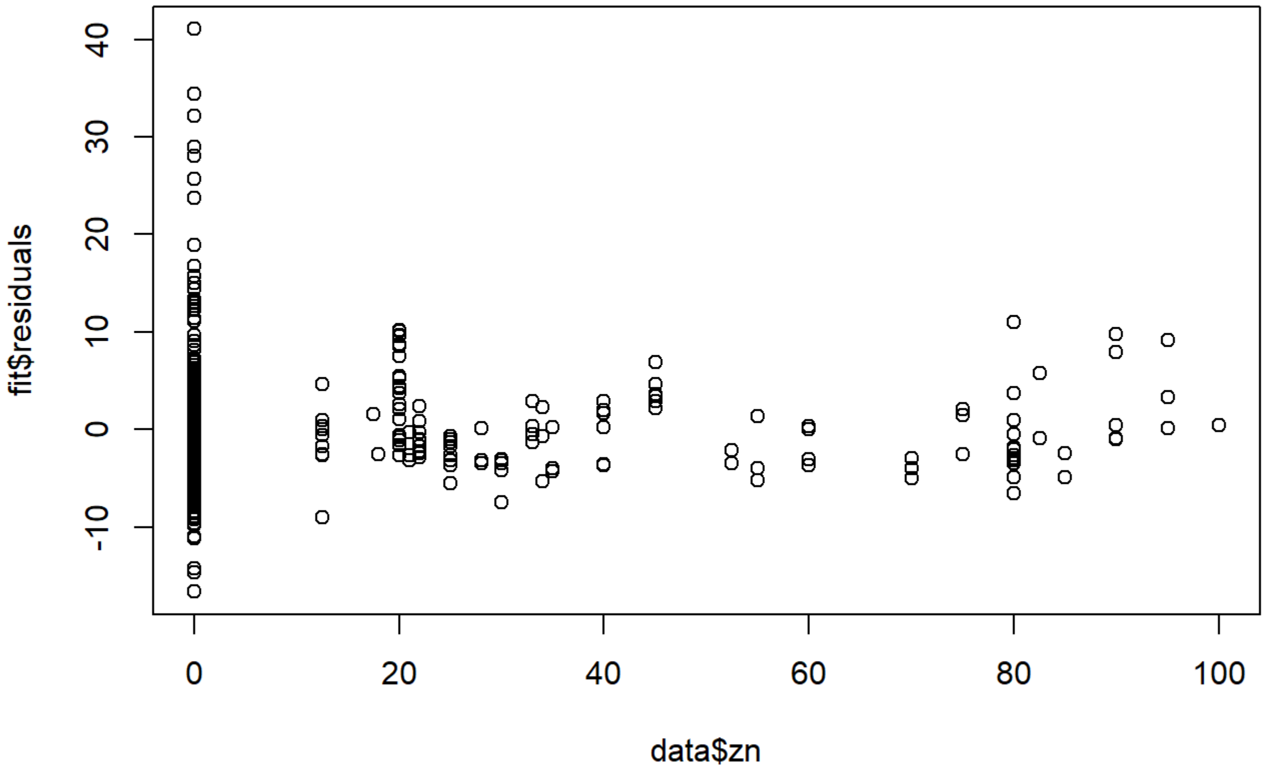
```
shapiro.test(fit$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit$residuals
## W = 0.83945, p-value < 2.2e-16
```

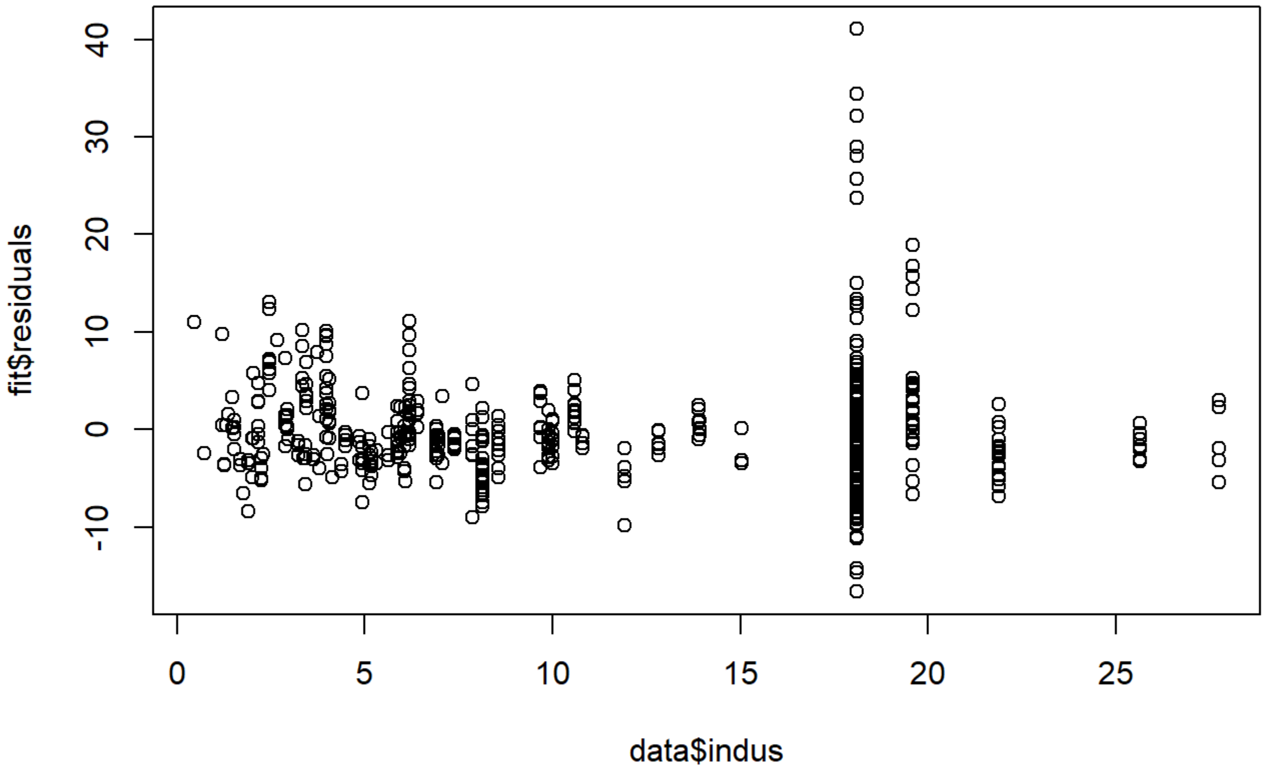
```
# From the histogram and qqplot of residuals, we could see that our data is not normally distrib
uted. And from the shapiro test, it showed p-value is less than 0.05, so we will reject the null
and conclude data is not normally distributed. What we can do is to transform the data to make i
t normally distributed or we could robust regression methods.

# Homoscedasticity
plot(data$crim, fit$residuals)
```



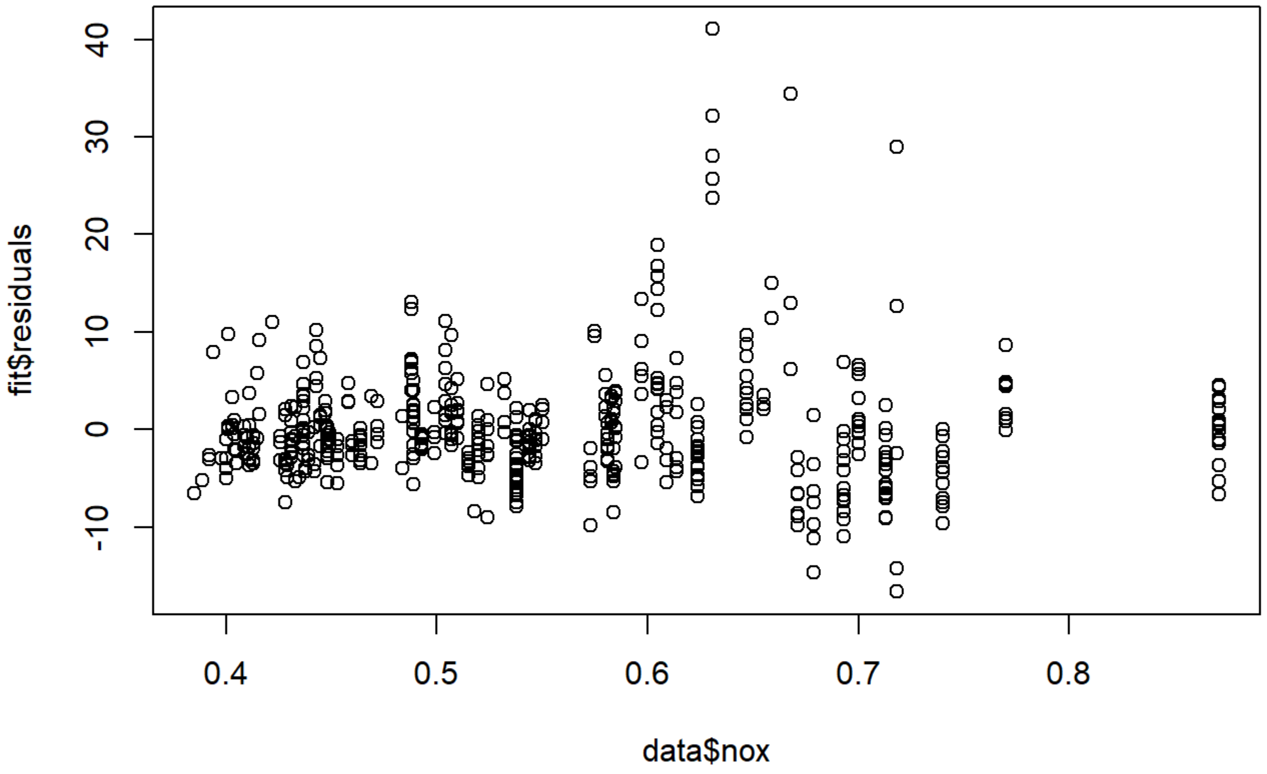
```
plot(data$zn, fit$residuals)
```



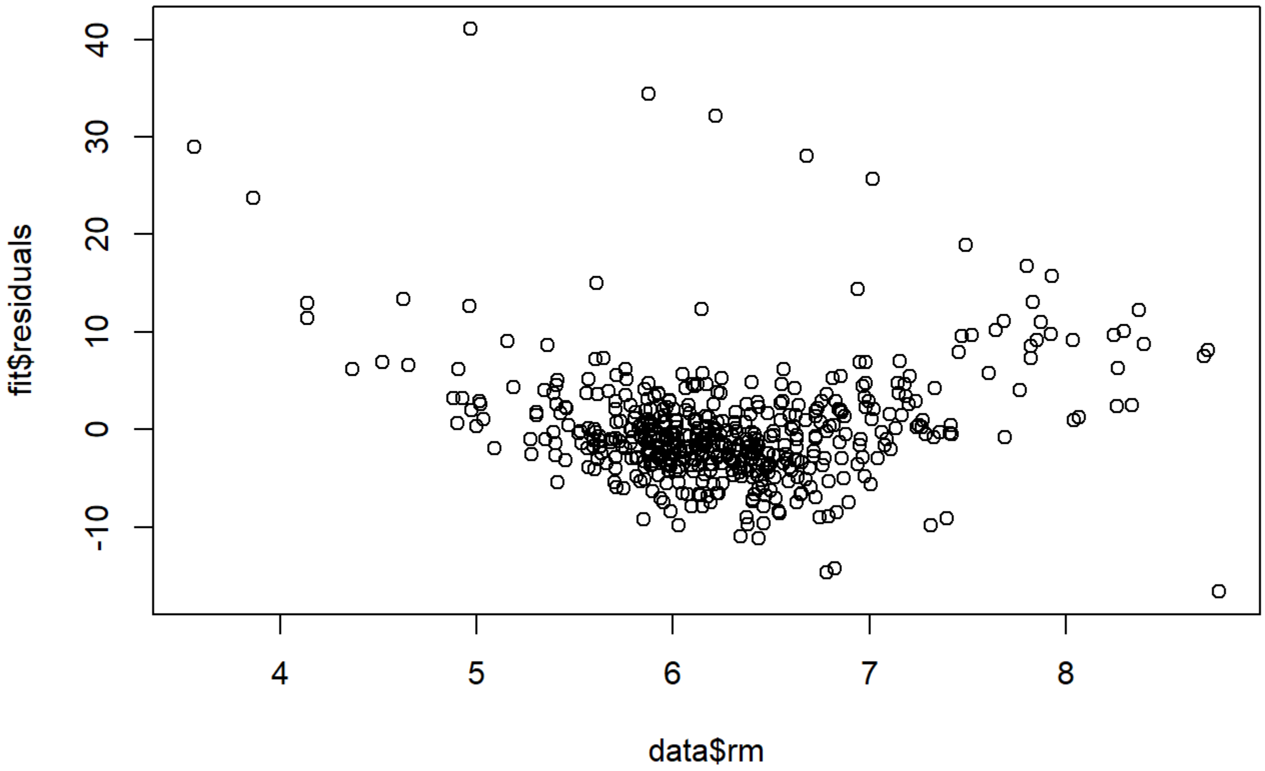
```
plot(data$indus, fit$residuals)
```



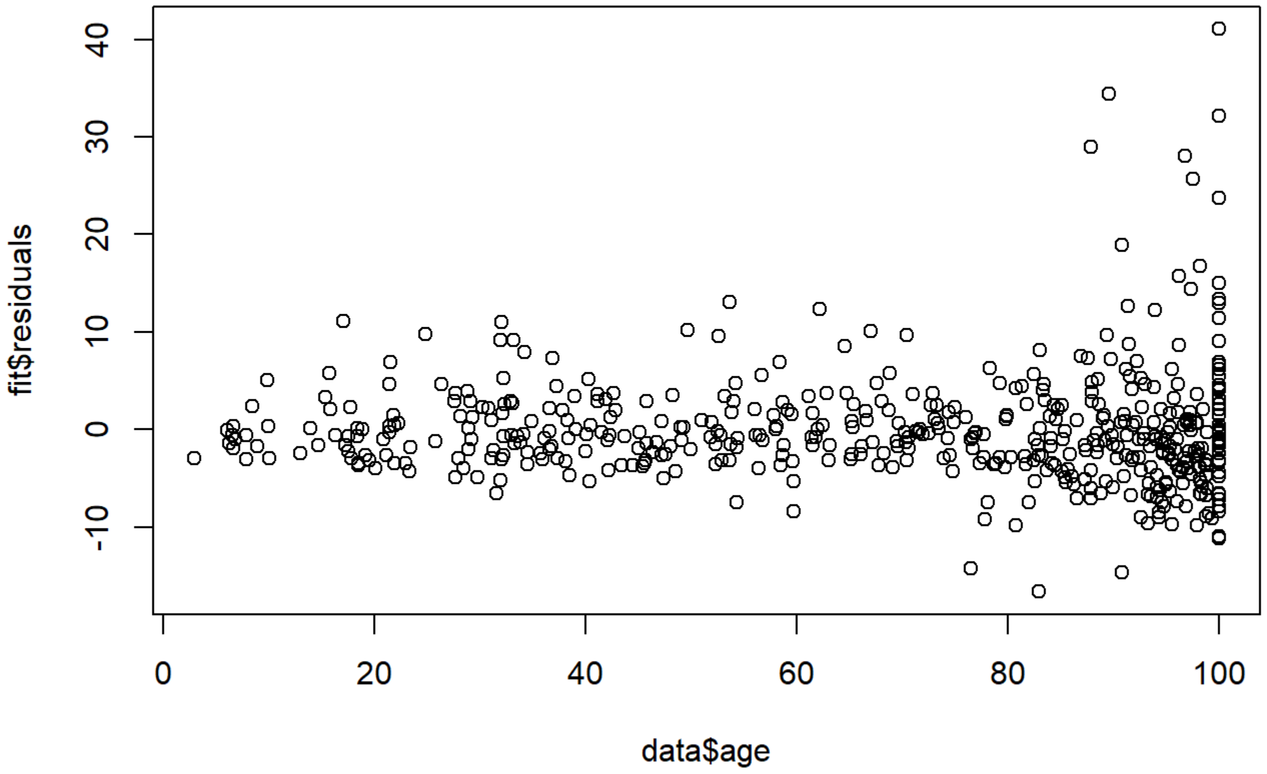
```
plot(data$nox, fit$residuals)
```



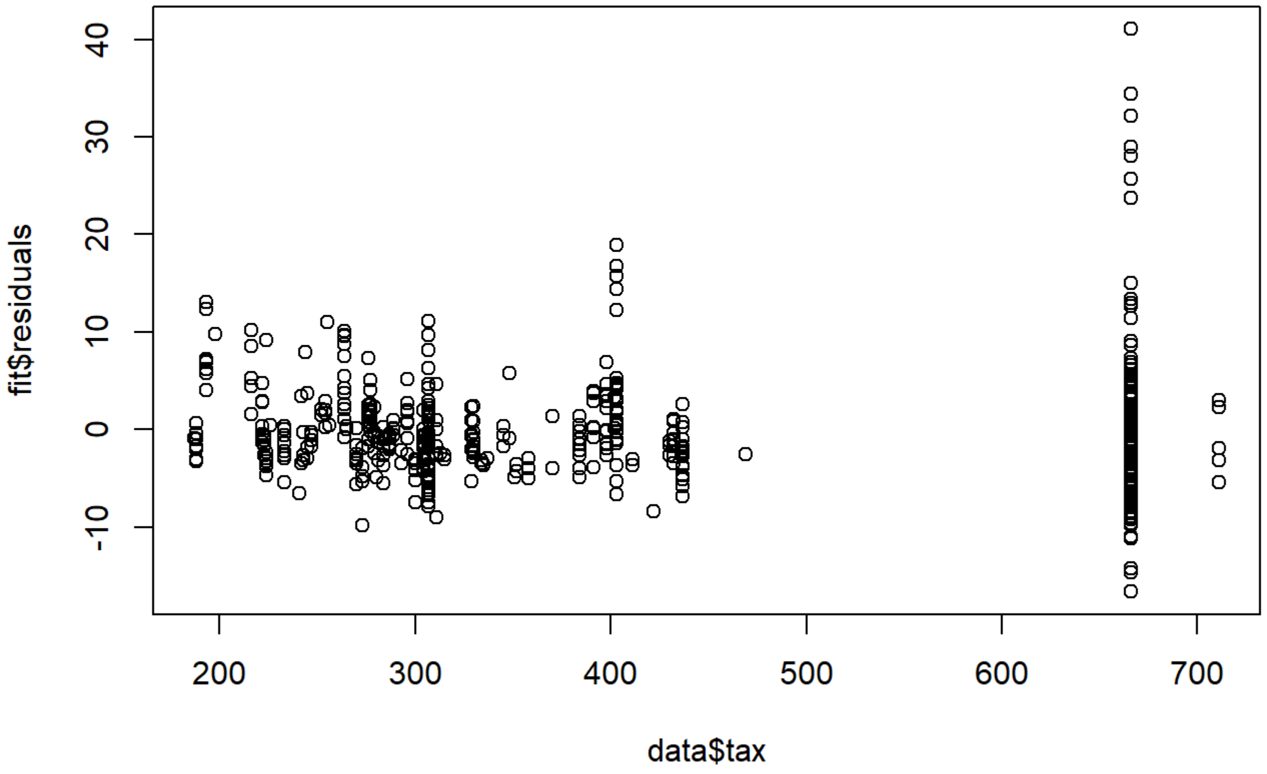
```
plot(data$rm, fit$residuals)
```



```
plot(data$age, fit$residuals)
```



```
plot(data$tax, fit$residuals)
```



```
# From the scatter plots, we could tell that there is no constant variance for these variables s
o they do not satisfy homoscedasticity. What we can do is to do the transformation or build vari
ance structure into model: WLS

# Uncorrelated error
durbinWatsonTest(fit)
```

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

```
# From the durbin watson test, we could see that the p-value is 0,so we can reject the null and
 conclude that there are correlated erros in the regression. What we can do is to do a cochrane-
orcutt transformation procesure or we could use models that incorporate the correlation structur
e such as generalized estimating equations.
```

c)

```
fitlms <- lmsreg(medv ~ crim + zn + indus + nox + rm + age + tax, data)
fitlms
```

```
## Call:
## lqs.formula(formula = medv ~ crim + zn + indus + nox + rm + age +
##     tax, data = data, method = "lms")
##
## Coefficients:
## (Intercept)         crim           zn        indus          nox
##  -38.383180    -0.157378     0.028858    -0.044152     7.164195
##          rm          age          tax
##    9.732783    -0.070628     0.002251
##
## Scale estimates 3.914 3.919
```

```
# From the results, we could see that some of the coefficients are pretty cloase. Except for the
intercept, crim and nox increased a lot and indus changed from negative to positive.
```