STA 5820 Chapter 3 lab: Linear Regression

魏上傑

2023-04-22

Some supplemental codes for linear models.

Boston Data set

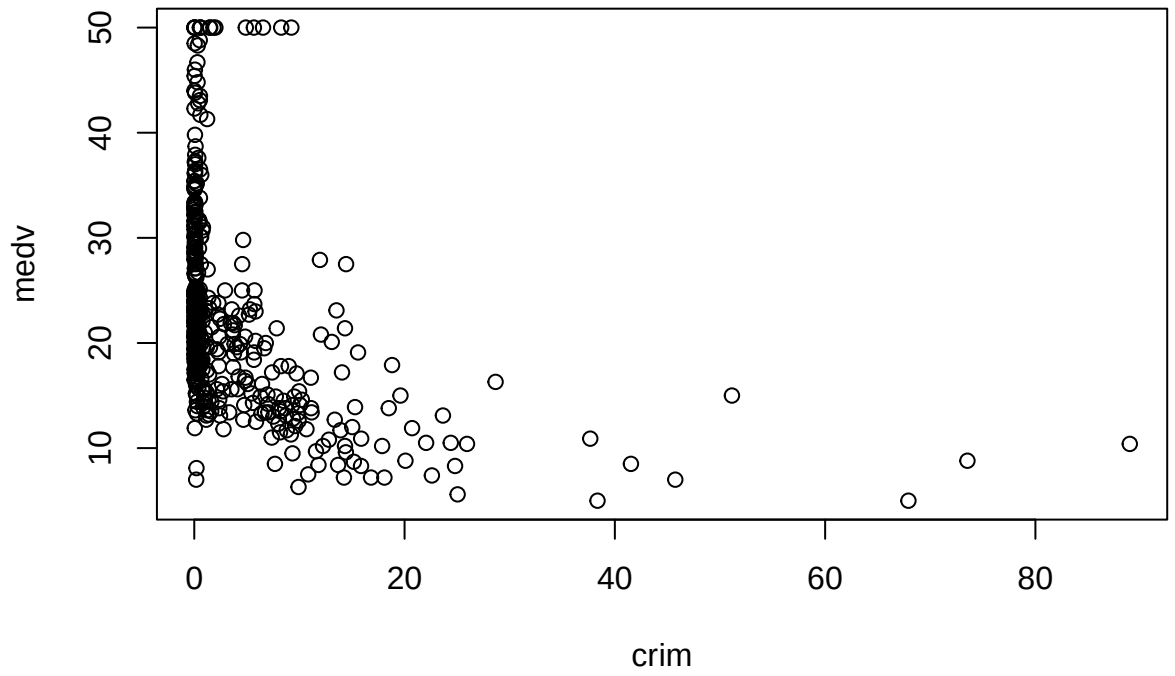
```
library(MASS) # Boston data
head(Boston) # median value of homes by town
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

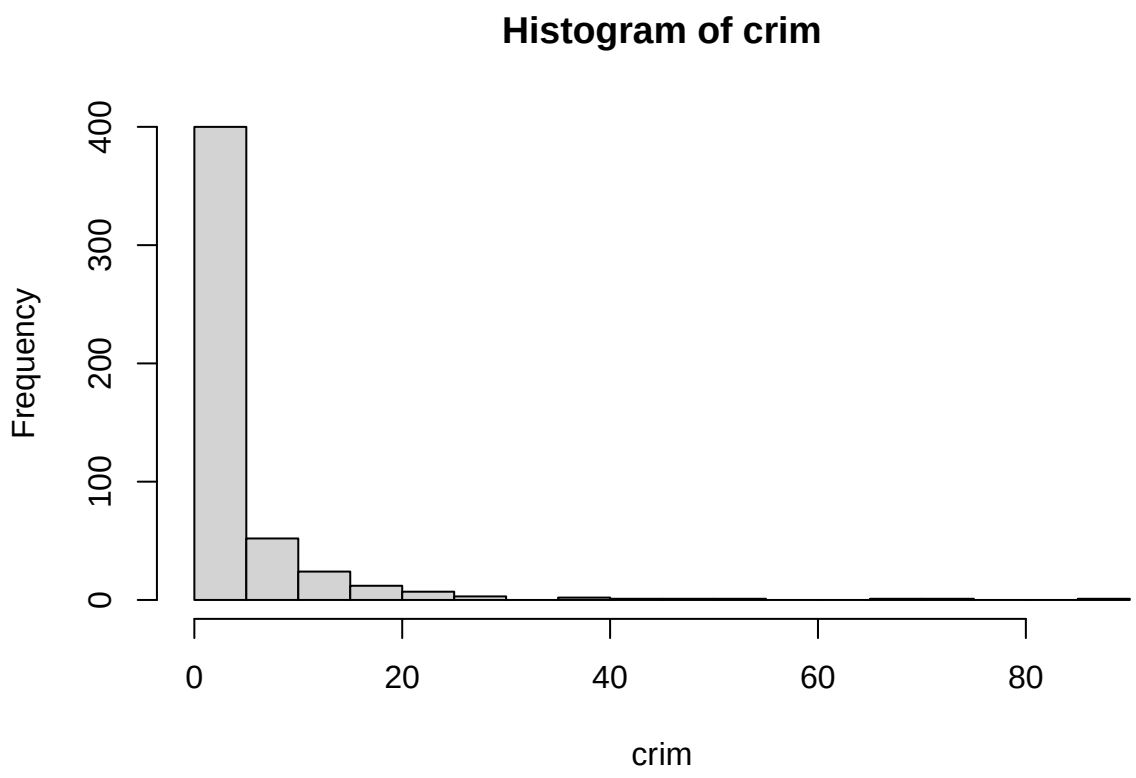
```
# see help for definition of variables
attach(Boston)
```

Scatter plot for crime rate vs. median house value

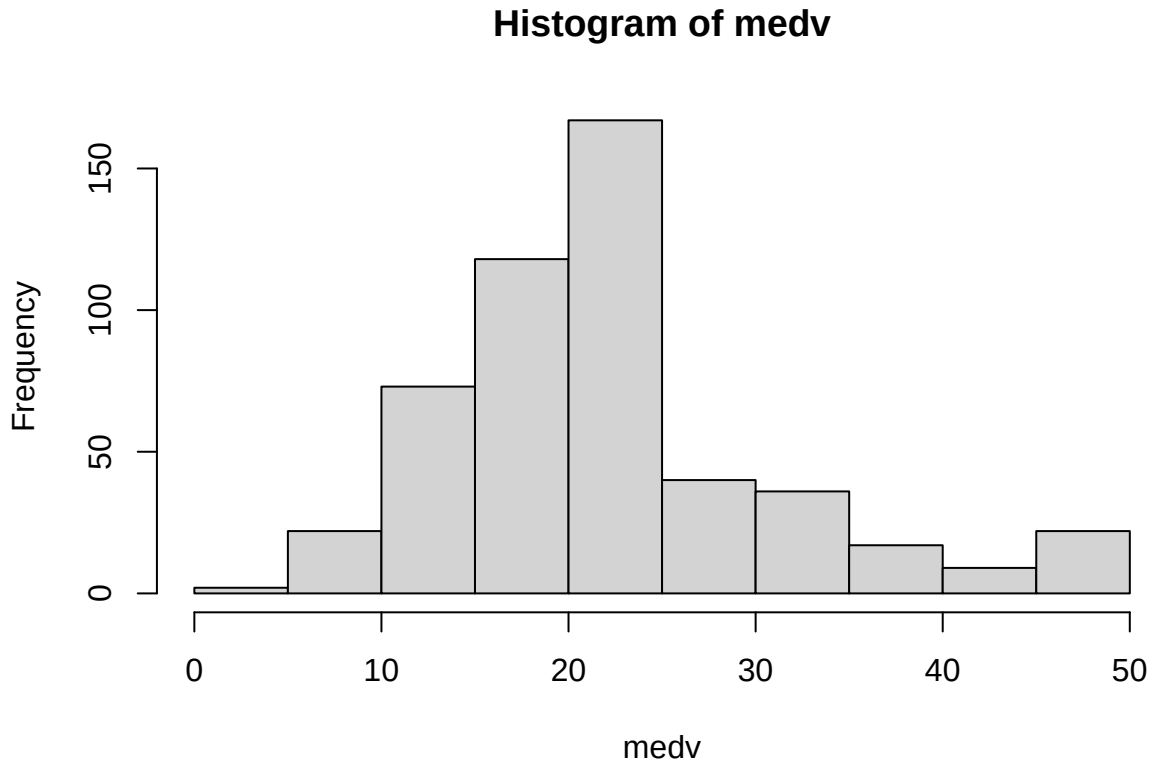
```
plot(crim, medv)
```




```
hist(crim, breaks = seq(0, 90, 5))
```

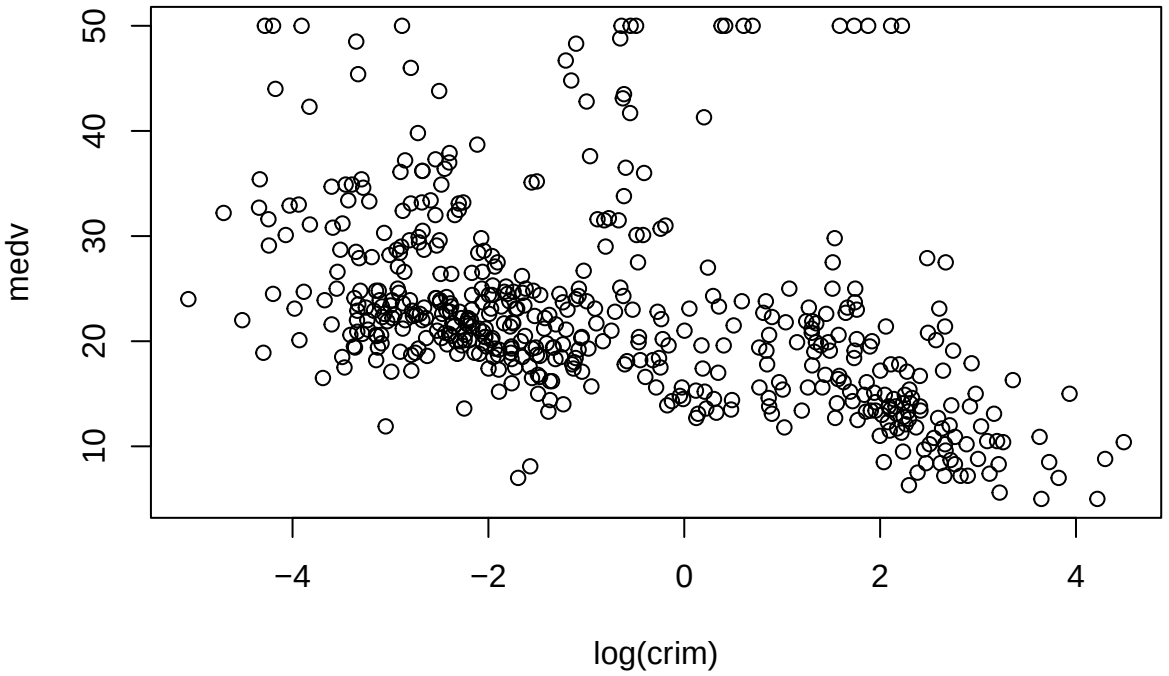



```
hist(medv, breaks = seq(0,50,5))
```



The crime rate is a skewed right, and needs a concave transformation to make the relationship between the two variables more linear.

```
plot(log(crim), medv)
```



While the median house values seem to be truncated at 50 ($50,000), the crime rate and the median house value has roughly a linear relationship.

Fitting a linear model

```
LM1 <- lm(medv~ log(crim))
summary(LM1)
```

```
##
## Call:
## lm(formula = medv ~ log(crim))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.303  -5.159  -2.427   2.666  33.271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.0246     0.3877   54.23   <2e-16 ***
```

```
## log(crim)    -1.9325     0.1688  -11.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.201 on 504 degrees of freedom
## Multiple R-squared:  0.2064, Adjusted R-squared:  0.2048
## F-statistic: 131.1 on 1 and 504 DF,  p-value: < 2.2e-16
```

$$medv = -1.93log(crim) + 21 + \epsilon$$

$$e^{medv} = crim \times e^{-1.93}$$

```
# confidence band, prediction band
confint(LM1, level = 0.99) # 99% confidence interval (default=95%)
```

```
##                 0.5 %    99.5 %
## (Intercept) 20.02222 22.026930
## log(crim)   -2.36900 -1.496093
```

```
CI <- predict(LM1, interval="confidence") # confidence interval
head(CI)
```

```
##        fit      lwr      upr
## 1 30.81106 29.22008 32.40205
## 2 27.98271 26.80467 29.16076
## 3 27.98413 26.80589 29.16237
## 4 27.65422 26.52041 28.78803
## 5 26.19013 25.23775 27.14250
## 6 27.81085 26.65608 28.96562
```

```
PI <- predict(LM1, interval = "prediction") #prediction interval
```

```
## Warning in predict.lm(LM1, interval = "prediction"): predictions on current data refer
```

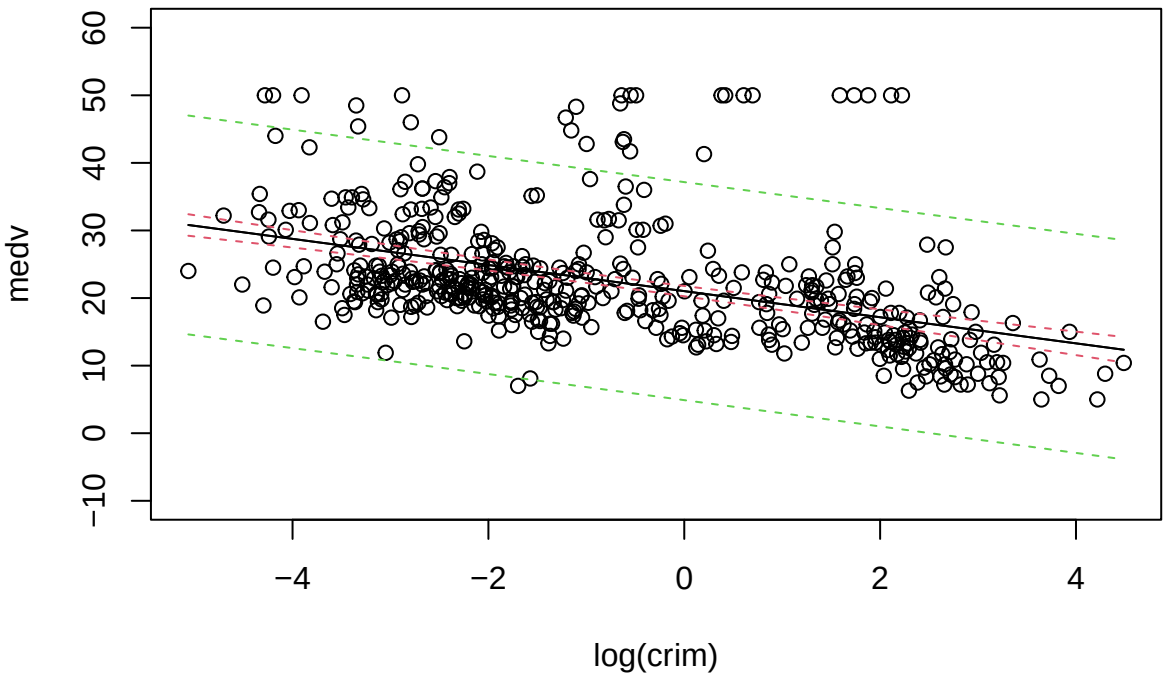
```
head(PI)
```

```
##        fit      lwr      upr
## 1 30.81106 14.61967 47.00245
## 2 27.98271 11.82668 44.13875
## 3 27.98413 11.82808 44.14018
## 4 27.65422 11.50135 43.80710
## 5 26.19013 10.04897 42.33128
## 6 27.81085 11.65649 43.96521
```

```
ORD <- order(crim) # index for crim in increasing order
head(ORD)
```

```
## [1]   1 285 286 342  56  55
```

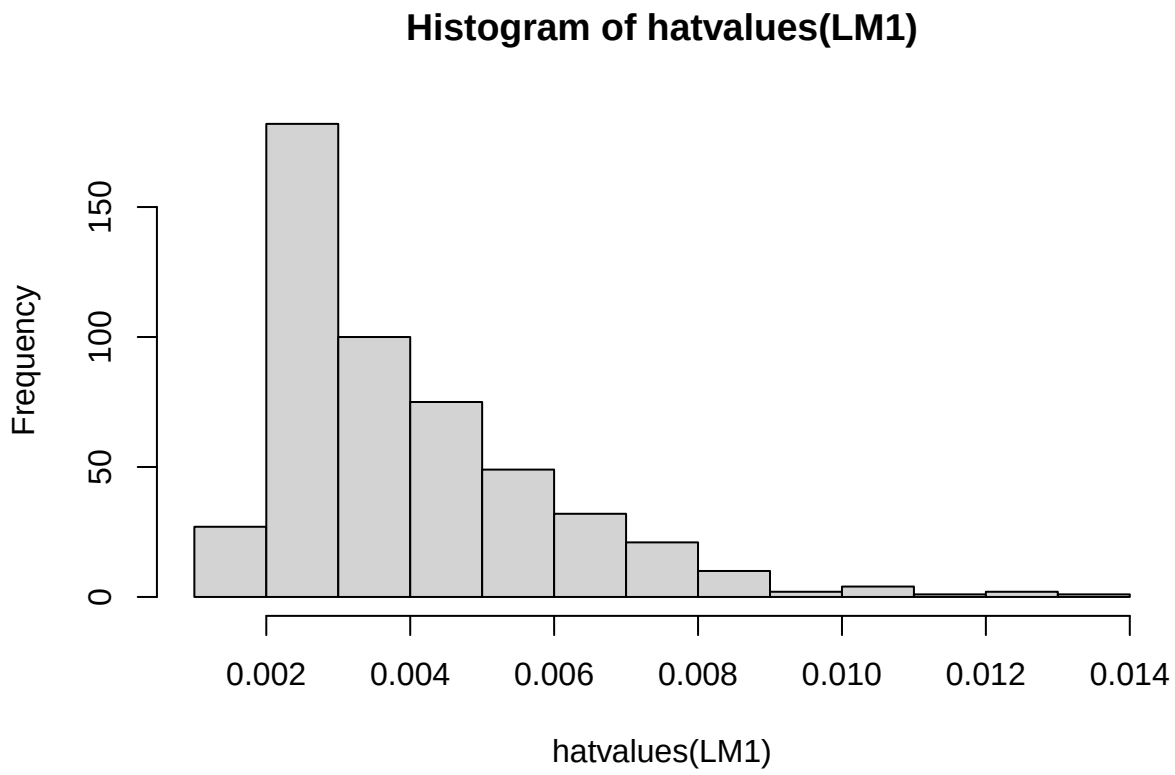
```
plot(log(crim), medv, ylim = c(-10, 60))
matlines(log(crim)[ORD], CI[ORD,], type="l", col=c(1,2,2), lty=c(1,2,2))
matlines(log(crim)[ORD], PI[ORD,], type="l", col=c(1,3,3), lty=c(1,2,2))
```



Roughly 95% of the observations are within the prediction bands (green). As there are many (506) observations, the confidence band (red) is narrow, indicating that the estimated regression line (black) is reasonably accurate.

leverage

```
# plot and leverage
hist(hatvalues(LM1))
```




```
sort(hatvalues(LM1))[501:506] # 6 largest hat values
```

```
##         405         415         411         406         419         381
## 0.01058067 0.01095391 0.01139357 0.01256163 0.01290060 0.01373612
```

There are no extreme hatvalues, compared to the others.

```
which.max(hatvalues(LM1)) # point with largest leverage
```

```
## 381
## 381
```

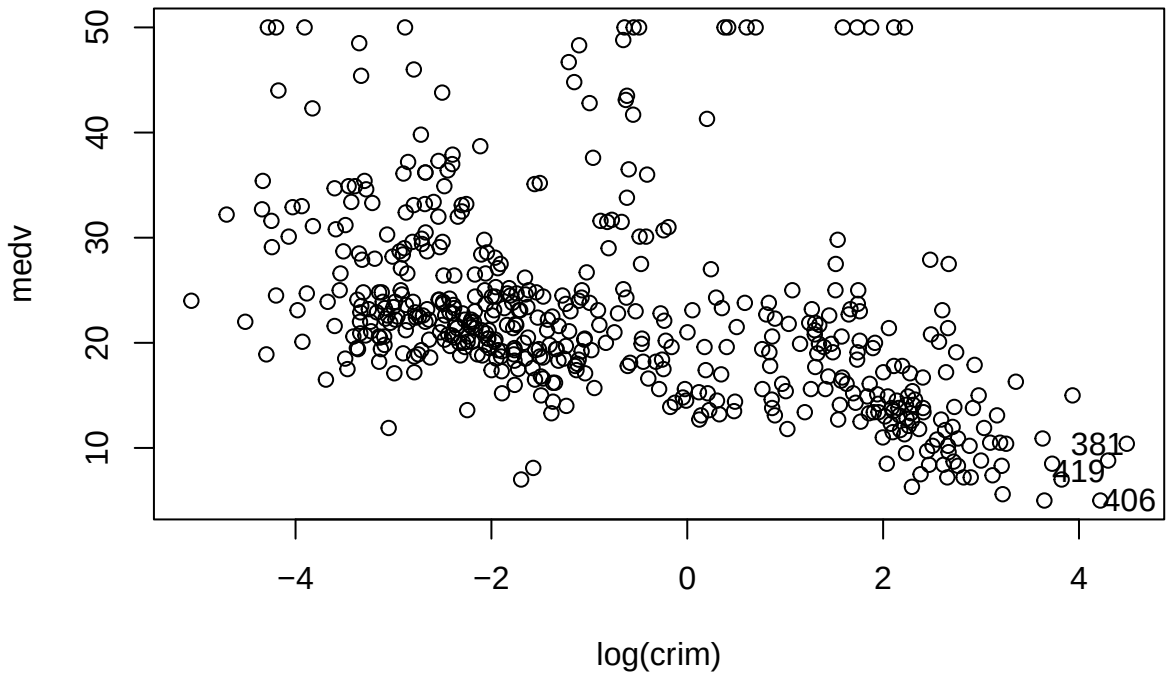
```
MAX3 <- order(hatvalues(LM1))[504:506] # points with 3 largest leverages
MAX3
```

```
## [1] 406 419 381
```

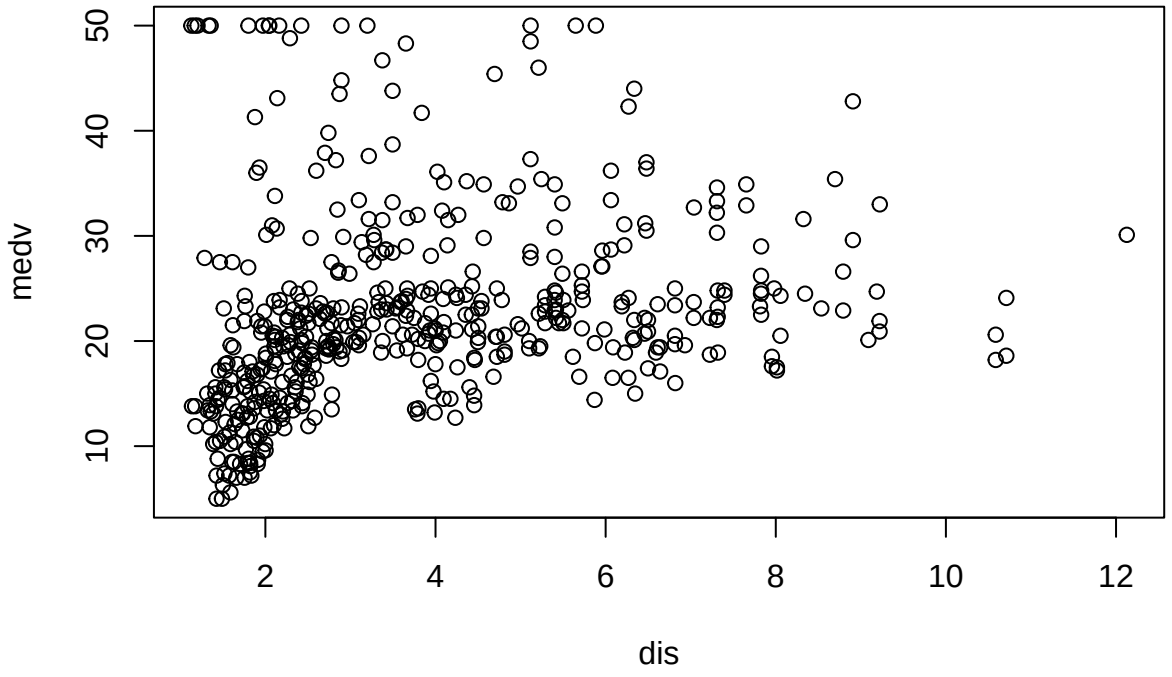
```
library(car) # for pointLabel
```

```
## Loading required package: carData
```

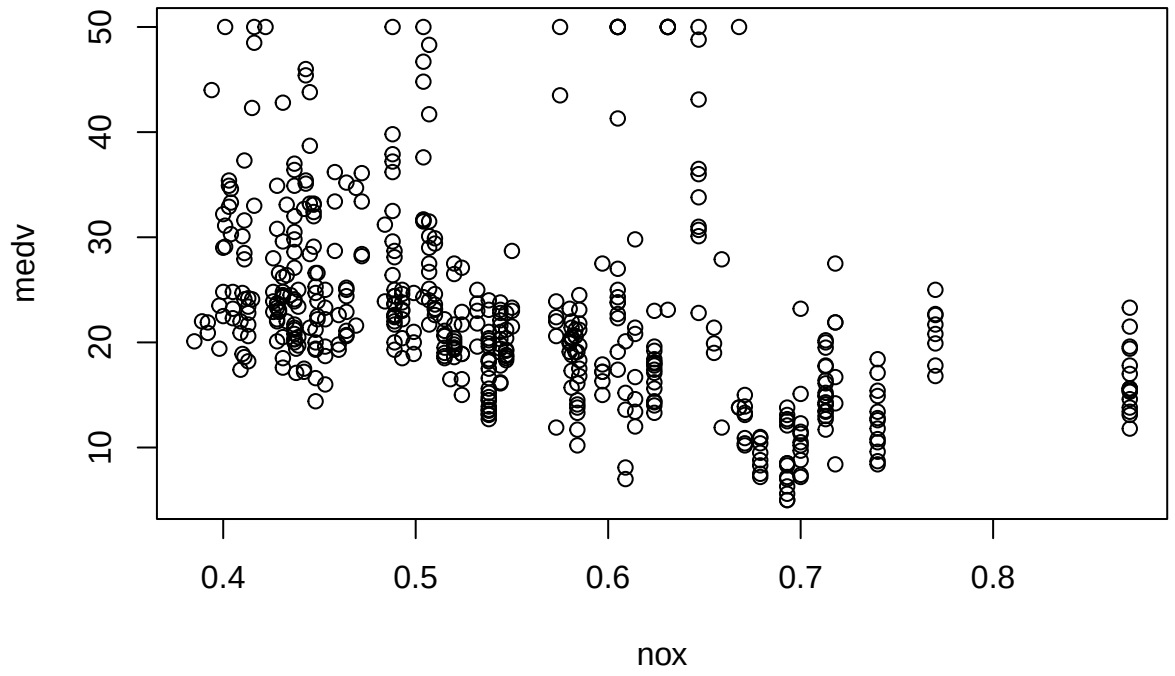
```
plot(log(crim), medv)
pointLabel(x=log(crim)[MAX3], y=medv[MAX3], labels = as.character(MAX3))
```




```
plot(dis, medv)
```



```
plot(nox, medv)
```




```
# ANOVA (Analysis of Variance)
LM2 <- lm(medv~ log(crim)+dis+nox)
summary(LM2)
```

```
##
## Call:
## lm(formula = medv ~ log(crim) + dis + nox)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.784  -5.255  -2.090   2.773  31.970
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  40.6757     3.9767  10.228  < 2e-16 ***
## log(crim)    -1.5398     0.2733  -5.635 2.92e-08 ***
## dis          -1.1342     0.2700  -4.201 3.15e-05 ***
## nox         -27.1141     5.8364  -4.646 4.34e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 8.012 on 502 degrees of freedom
## Multiple R-squared:  0.2455, Adjusted R-squared:  0.241
## F-statistic: 54.45 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
anova(LM2)
```

```
## Analysis of Variance Table
##
## Response: medv
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## log(crim)   1   8816  8816.2 137.3245 < 2.2e-16 ***
## dis         1    286   286.1   4.4565   0.03526 *
## nox         1   1386  1385.6  21.5823 4.336e-06 ***
## Residuals 502  32228    64.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
LM2b <- lm(medv~ log(crim)+dis+nox)
# change of order doesn't matter here
summary(LM2b)
```

```
##
## Call:
## lm(formula = medv ~ log(crim) + dis + nox)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.784  -5.255  -2.090   2.773  31.970
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  40.6757     3.9767  10.228  < 2e-16 ***
## log(crim)    -1.5398     0.2733  -5.635 2.92e-08 ***
## dis          -1.1342     0.2700  -4.201 3.15e-05 ***
## nox         -27.1141     5.8364  -4.646 4.34e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.012 on 502 degrees of freedom
```

```
## Multiple R-squared:  0.2455, Adjusted R-squared:  0.241
## F-statistic: 54.45 on 3 and 502 DF,  p-value: < 2.2e-16
```

```
anova(LM2b) # change of order matters in ANOVA
```

```
## Analysis of Variance Table
##
## Response: medv
##             Df Sum Sq Mean Sq  F value    Pr(>F)
## log(crim)    1   8816  8816.2 137.3245 < 2.2e-16 ***
## dis          1    286   286.1   4.4565   0.03526 *
## nox          1   1386  1385.6  21.5823 4.336e-06 ***
## Residuals 502  32228    64.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(LM1, LM2) # see if LM2 is significantly better than LM1
```

```
## Analysis of Variance Table
##
## Model 1: medv ~ log(crim)
## Model 2: medv ~ log(crim) + dis + nox
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    504 33900
## 2    502 32228  2    1671.7 13.019 3.073e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```