HW2 Solution CS6220-Data Mining

1. JWHT problem 10, p. 56

(a)

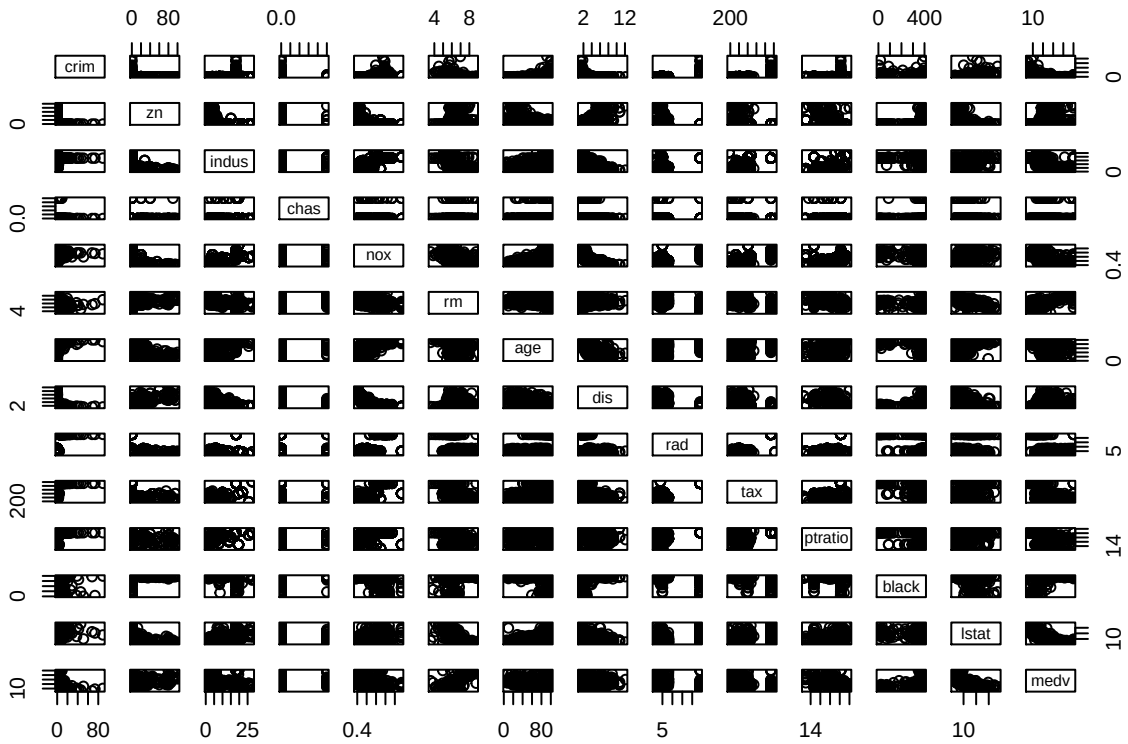
```
library(MASS)
?Boston
dim(Boston)
```

```
## [1] 506  14
```

The rows represent the individual neighborhoods, and the columns are the descriptors of the neighborhoods.

(b)

```
pairs(Boston)
```

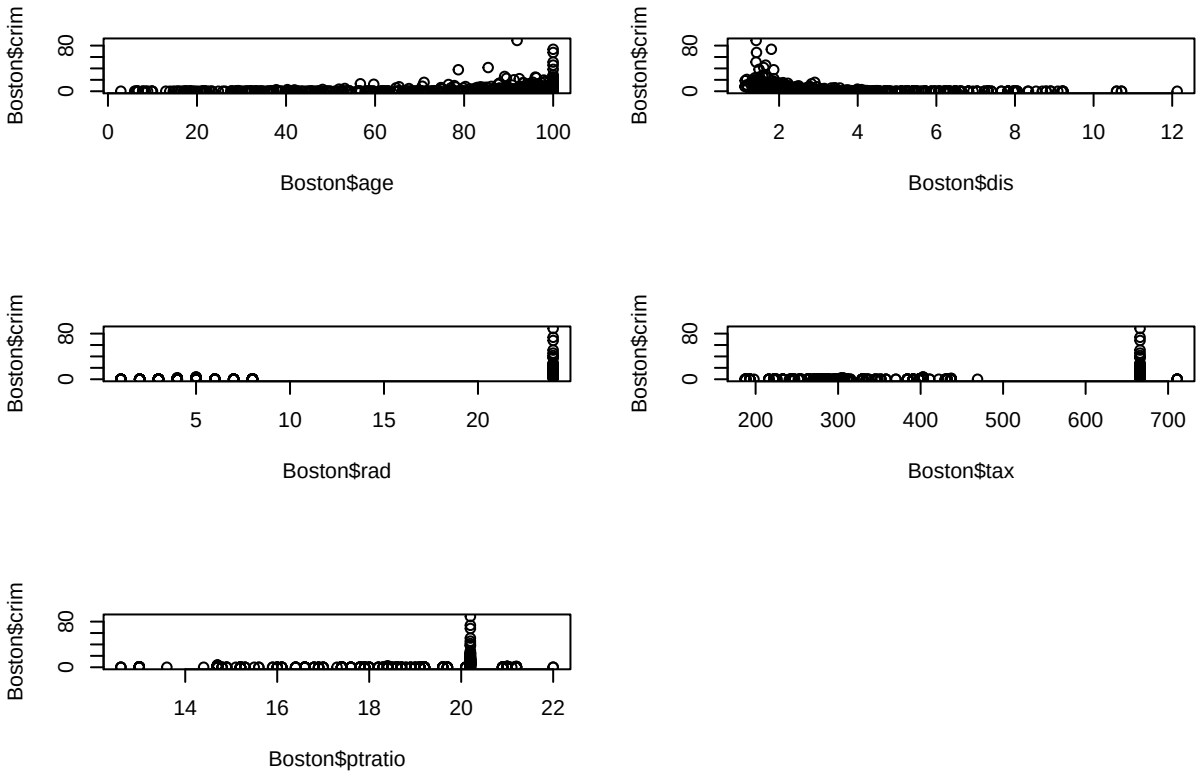



```
# X correlates with: a, b, c
# crim correlates with age, dis, rad, tax, ptratio
# zn correlates with indus, nox, age, lstat
# indus correlates with age, dis
# nox correlates with age, dis
# dis correlates with lstat
# lstat correlates with medv
```

This said, not all the associaions are linear; some associations can be difficult to see from the plots.

(c)

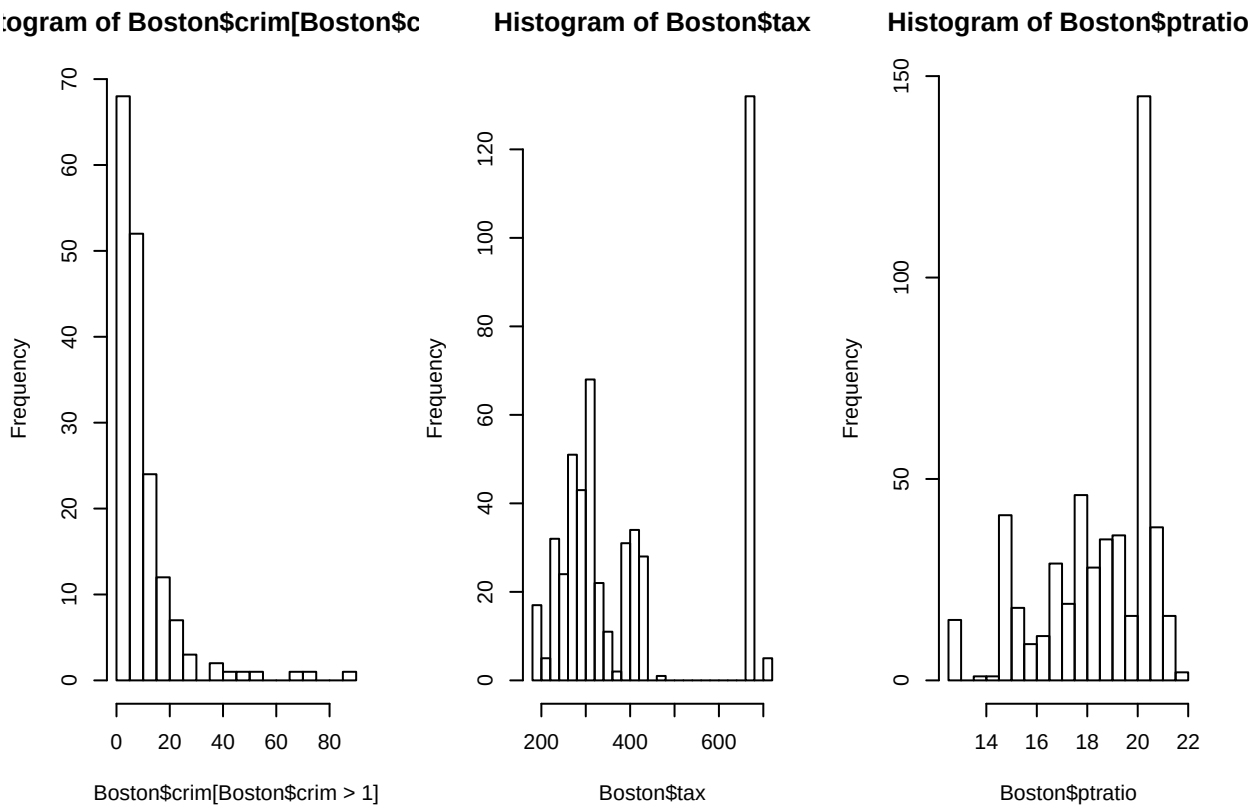
```
par(mfrow=c(3,2))
# Older homes, more crime
plot(Boston$age, Boston$crim)
# Closer to work-area, more crime
plot(Boston$dis, Boston$crim)
# Higher index of accessibility to radial highways, more crime
plot(Boston$rad, Boston$crim)
# Higher tax rate, more crime
plot(Boston$tax, Boston$crim)
# Higher pupil:teacher ratio, more crime
plot(Boston$ptratio, Boston$crim)
```



This said, the nature of the relationship between the crime rate and the descriptors is likely more complex than just a pairwise association.

(d)

```
par(mfrow=c(1,3))
# The crime rates is low in most cities, but several suburbs appear
# to have a crime rate > 20
hist(Boston$crim[Boston$crim>1], breaks=25)
# there is a gap between suburbs with low and high tax rates
hist(Boston$tax, breaks=25)
# ptratio skew twoards high values, but no particularly high value
hist(Boston$ptratio, breaks=25)
```



(e)

```
nrow(subset(Boston, chas == 1))
```

```
## [1] 35
```

(f)

```
median(Boston$ptratio)
```

```
## [1] 19.05
```

(g)

```
t(subset(Boston, medv == min(Boston$medv)))
```

```
##             399      406
## crim    38.3518  67.9208
## zn       0.0000   0.0000
## indus   18.1000  18.1000
## chas     0.0000   0.0000
## nox      0.6930   0.6930
## rm       5.4530   5.6830
## age    100.0000 100.0000
## dis      1.4896   1.4254
```

```
## rad      24.0000  24.0000
## tax     666.0000 666.0000
## ptratio  20.2000  20.2000
## black   396.9000 384.9700
## lstat    30.5900  22.9800
## medv      5.0000   5.0000
```

```
#              399      406
# crim     38.3518  67.9208 above 3rd quartile
# zn        0.0000   0.0000 at min
# indus    18.1000  18.1000 at 3rd quartile
# chas      0.0000   0.0000 not bounded by river
# nox       0.6930   0.6930 above 3rd quartile
# rm        5.4530   5.6830 below 1st quartile
# age     100.0000 100.0000 at max
# dis       1.4896   1.4254 below 1st quartile
# rad      24.0000  24.0000 at max
# tax     666.0000 666.0000 at 3rd quartile
# ptratio  20.2000  20.2000 at 3rd quartile
# black   396.9000 384.9700 at max; above 1st quartile
# lstat    30.5900  22.9800 above 3rd quartile
# medv      5.0000   5.0000 at min
summary(Boston)
```

```
##       crim                zn              indus            chas
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000
##  1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000
##       nox               rm              age              dis
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127
##       rad              tax           ptratio          black
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90
##      lstat            medv
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

```
# Comparing to dataset summary, this city is somewhere in the middle of ranking of cities as a good pla
```

(h)

```
nrow(subset(Boston, rm > 7))
```

```
## [1] 64
```

```
nrow(subset(Boston, rm > 8))
```

```
## [1] 13
```

```
summary(subset(Boston, rm > 8))
```

```
##       crim                zn             indus            chas        
##  Min.   :0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.0000  
##  1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.0000  
##  Median :0.52014   Median : 0.00   Median : 6.200   Median :0.0000  
##  Mean   :0.71879   Mean   :13.62   Mean   : 7.078   Mean   :0.1538  
##  3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200   3rd Qu.:0.0000  
##  Max.   :3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.0000  
##       nox               rm              age              dis        
##  Min.   :0.4161   Min.   :8.034   Min.   : 8.40   Min.   :1.801  
##  1st Qu.:0.5040   1st Qu.:8.247   1st Qu.:70.40   1st Qu.:2.288  
##  Median :0.5070   Median :8.297   Median :78.30   Median :2.894  
##  Mean   :0.5392   Mean   :8.349   Mean   :71.54   Mean   :3.430  
##  3rd Qu.:0.6050   3rd Qu.:8.398   3rd Qu.:86.50   3rd Qu.:3.652  
##  Max.   :0.7180   Max.   :8.780   Max.   :93.90   Max.   :8.907  
##       rad              tax           ptratio          black       
##  Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min.   :354.6  
##  1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st Qu.:384.5  
##  Median : 7.000   Median :307.0   Median :17.40   Median :386.9  
##  Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mean   :385.2  
##  3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd Qu.:389.7  
##  Max.   :24.000   Max.   :666.0   Max.   :20.20   Max.   :396.9  
##      lstat           medv      
##  Min.   :2.47   Min.   :21.9  
##  1st Qu.:3.32   1st Qu.:41.7  
##  Median :4.14   Median :48.3  
##  Mean   :4.31   Mean   :44.2  
##  3rd Qu.:5.12   3rd Qu.:50.0  
##  Max.   :7.44   Max.   :50.0
```

```
summary(Boston)
```

```
##       crim                 zn               indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917
```

```
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000
##       nox               rm              age              dis
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127
##       rad              tax           ptratio          black
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90
##      lstat            medv
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

```
# The crime and lstat are lower in suburbs with rm > 8
```

2. JWHT problem 8, p. 121 (a)

```
Auto = read.csv("/Users/ovitek/Dropbox/Olga/Teaching/CS6220/Fall15/Homeworks/Hw2/Auto.csv", header=T, na
Auto = na.omit(Auto)
summary(Auto)
```

```
##       mpg          cylinders      displacement     horsepower
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0
##
##      weight      acceleration        year           origin
##  Min.   :1613   Min.   : 8.00   Min.   :70.00   Min.   :1.000
##  1st Qu.:2225   1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000
##  Median :2804   Median :15.50   Median :76.00   Median :1.000
##  Mean   :2978   Mean   :15.54   Mean   :75.98   Mean   :1.577
##  3rd Qu.:3615   3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000
##  Max.   :5140   Max.   :24.80   Max.   :82.00   Max.   :3.000
##
##                  name
##  amc matador       :  5
##  ford pinto        :  5
##  toyota corolla    :  5
##  amc gremlin       :  4
```

```
##  amc hornet        :  4
##  chevrolet chevette:  4
##  (Other)           :365
```

```
attach(Auto)
lm.fit = lm(mpg ~ horsepower)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

i. Testing the null hypothesis of all regression coefficients equal to zero shows F-statistic is much larger than 1 and the p-value is close to zero. So, we can reject the null hypothesis and state that horsepower and mpg have some relationship.

ii. mean of mpg: 23.4459 RSE of the lm.fit: 4.906 (percentage error of 20.9248%.) R2 of the lm.fit: 0.6059 (60.5948% of the variance in mpg is explained by horsepower.)

iii. The more horsepower is, the less mpg will be. So, the relationship is negative.

iv.

```
predict(lm.fit, data.frame(horsepower=c(98)), interval="confidence")
```

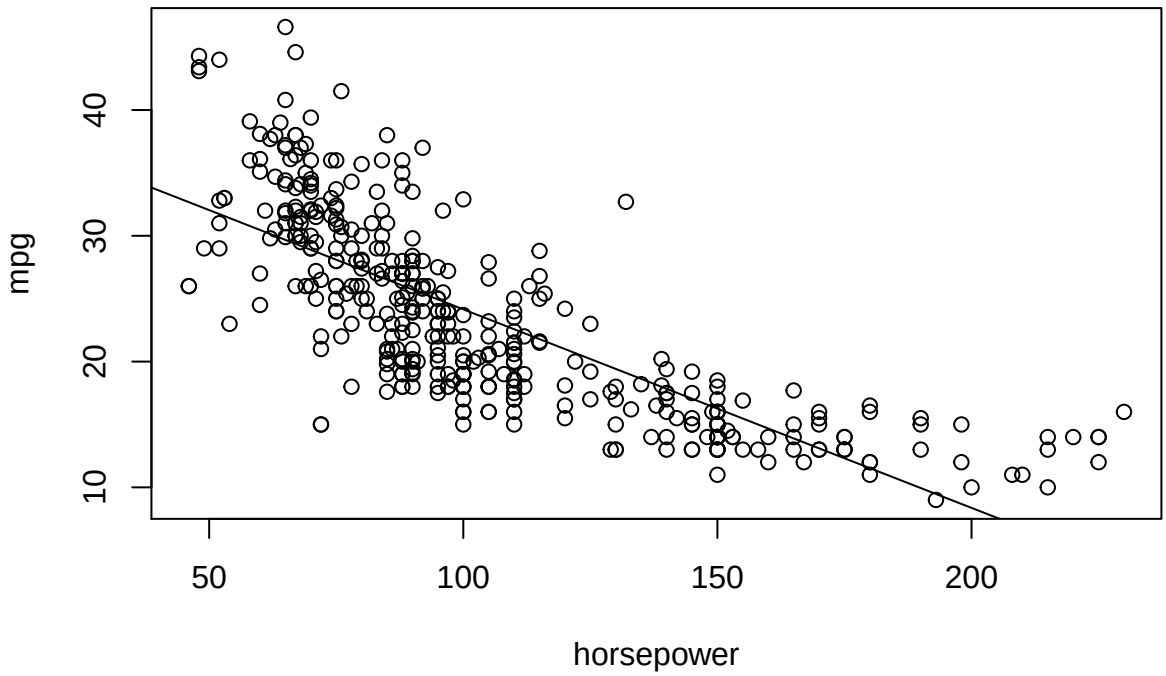
```
##        fit      lwr      upr
## 1 24.46708 23.97308 24.96108
```

```
predict(lm.fit, data.frame(horsepower=c(98)), interval="prediction")
```

```
##        fit     lwr      upr
## 1 24.46708 14.8094 34.12476
```

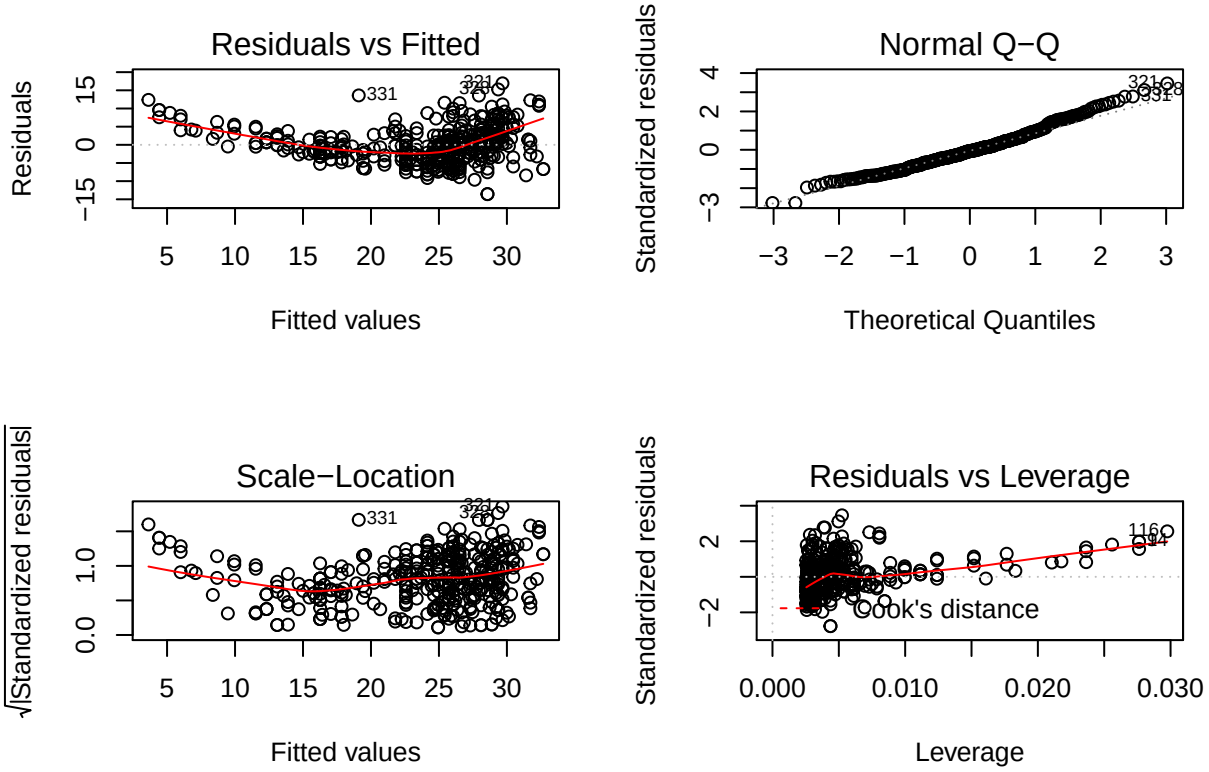
(b)

```
plot(horsepower, mpg)
abline(lm.fit)
```



(c)

```
par(mfrow=c(2,2))
plot(lm.fit)
```



We will focus on the residual plot (top left). The plot shows that the linear model systematically over-estimates mpg for low and high predicted values, and under-estimates for the middle-range predicted values. This indicates that the linear fit is not appropriate. The scatterplot in part (b) confirms that.

3. JWHT problem 11, p. 123

```
set.seed(1)
x = rnorm(100)
y = 2*x + rnorm(100)
```

(a)

```
lm.fit = lm(y~x+0)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = y ~ x + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9154 -0.6472 -0.1771  0.5056  2.3109
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## x   1.9939     0.1065   18.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9586 on 99 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7776
## F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

The simulation mimics a situation where we have a population, with a linear relationship between x and y. In this population the slope is 2, and the variance of the residuals is 1. The simulation is useful, because it allows us to see how well the linear model uncovers the true 'population' parameter.

From the utput we see that the slope is 1.9939 (i.e., close to 2), and the residual mean square error is 0.9586 (i.e., close to 1). Therefore, the estimates of the linear model are close to the population values.

(b)

```
lm.fit = lm(x~y+0)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = x ~ y + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8699 -0.2368  0.1030  0.2858  0.8938
```

```
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## y  0.39111    0.02089   18.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4246 on 99 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7776
## F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

(c)

While the estimates of the slopes (and of their standard errors) differ between (a) and (b), the t statistics and the residual mean square errors are the same.

(d)

By definition $t = \frac{\hat{\beta}}{SE(\hat{\beta})}$

Now substitute $\hat{\beta} = \sum x_i y_i / \sum x_i^2$ and $SE(\hat{\beta}) = \sqrt{\frac{\sum (y_i - x_i\hat{\beta})^2}{(n-1)\sum x_i^2}}$

Therefore $t = \frac{\sum x_i y_i}{\sum x_i^2}\sqrt{\frac{(n-1)\sum x_i^2}{\sum (y_i - x_i\hat{\beta})^2}}$

This simplifies to $t = \frac{\sqrt{n-1}\sum x_i y_i}{\sqrt{\sum x_i^2 \sum (y_i - x_i\hat{\beta})^2}}$

Open the square in the denominator $t = \frac{\sqrt{n-1}\sum x_i y_i}{\sqrt{\sum x_i^2 \sum \left(y_i^2 - 2\beta x_i y_i + x_i^2\hat{\beta}^2\right)^2}}$

Multiply the two products in the denominator $t = \frac{\sqrt{n-1}\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2 - \sum x_i^2\beta\left(2\sum x_i y_i - \sum x_i^2\hat{\beta}\right)}}$

Replace $\hat{\beta}$ with its definition above $t = \frac{\sqrt{n-1}\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2 - \sum x_i y_i\left(2\sum x_i y_i - \sum x_i^2 \sum x_i y_i\right)}}$

Simplify the denominator $t = \frac{\sqrt{n-1}\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2 - \left(\sum x_i y_i\right)^2}}$

```
(sqrt(length(x)-1) * sum(x*y)) / (sqrt(sum(x*x) * sum(y*y) - (sum(x*y))^2))
```

```
## [1] 18.72593
```

(e)

The mathematical expression treats x and y symmetrically, and therefore the test statistic is the same for both regressions

(f)

```
lm.fit = lm(y~x)
lm.fit2 = lm(x~y)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8768 -0.6138 -0.1395  0.5394  2.3462
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.03769    0.09699  -0.389    0.698
## x            1.99894    0.10773  18.556   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9628 on 98 degrees of freedom
## Multiple R-squared:  0.7784, Adjusted R-squared:  0.7762
## F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = x ~ y)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90848 -0.28101  0.06274  0.24570  0.85736
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.03880    0.04266    0.91    0.365
## y            0.38942    0.02099   18.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4249 on 98 degrees of freedom
## Multiple R-squared:  0.7784, Adjusted R-squared:  0.7762
## F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

While the slopes are different, the test statistic (i.e., the 'signal-to-noise ratio') are the same for both regressions.

It is important to remember this fact in situations with no obvious choice of a response.