Statistical learning assignment 3 - ch 2∼3

孫浩哲　　M072040002

October 4, 2018

2-6.

Parametric approach：Assume a form for f by estimating a set of parameters.

Non-parametric approach：Requires a very large number of observations to accurately estimate f.

Advantages：simplifying of modeling f to a few parameters and it doesn't need large number of observations.

Disadvantages：The assumption may be incorrect, overfitting if we use more flexible models.

2-7.

(a)

obs. 1：$\sqrt{0^2+3^2+0^2}=3$

obs. 2：$\sqrt{2^2+0^2+0^2}=2$

obs. 3：$\sqrt{0^2+1^2+3^2}=\sqrt{10}$

obs. 4：$\sqrt{0^+1^2+2^2}=\sqrt{5}$

obs. 5：$\sqrt{(-1)^2+0^2+1^2}=\sqrt{2}$

obs. 6：$\sqrt{1^2+1^2+1^2}=\sqrt{3}$

(b)

Green, because the obs. 5 is the closest neighbor when $K = 1$.

(c)

Red, because the most closest neighbors when $K = 3$ are obs. 5, obs. 6 and obs. 2, and they are Green, Red and Red.

(d)

Small, A small K would be flexible for a non-linear decision boundary.
Large K would try to fit a more linear boundary.

3-1.

predictors	null hypothesis	T/F
TV	TV ads have no effect on sales in the presence of radio and newspaper.	False
radio	radio ads have no effect on sales in the presence of TV and newspaper.	False
newspaper	newspaper ads have no effect on sales in the presence of TV and radio.	True

Table 1: null hypothesis for each predictor and True or False

3-3.

Model that predict the starting salary：

$$salary = 50+20(GPA)+0.07(IQ)+35(gender)+0.01(GPA\times IQ)-10(GPA\times gender)$$

(a.)
Assume that GPA $= x_0$, IQ $= x_1$

male(gender $= 0$)：

$$salary = 50 + 20x_0 + 0.07x_1 + 0.01(x_0 \times x_1)$$

female(gender $= 1$)：

$$salary = 85 + 20x_0 + 0.07x_1 + 0.01(x_0 \times x_1) - 10x_0$$

It can't conclude which gender's salary is higher than another one unless the GPA > 3.5.So we choose(iii.)

(b.)

$$50 + 20 \times 4.0 + 0.07 \times 110 + 0.01 \times (4.0 \times 110) - 10 \times 4.0 = 137.1$$

(c.)

False, because we haven't examined the p-value of the null hypothesis correspond to the coefficient.

3-8.

```
library(ISLR)
auto=Auto
auto=na.omit(Auto)
summary(auto)
attach(auto)
model=lm(mpg~horsepower)
summary(model)
```

```
> summary(model)

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
```

(a.)

(i)

Yes, because the coefficient corresponds to the predictor is not 0, and by hypothesis test, we can see that p-value is extremely small, it means that there is significant relationship between horsepower and mpg.

(ii)

The R^2 is 0.6049, meaning 60.49% of the variance in mpg is explained by horsepower.

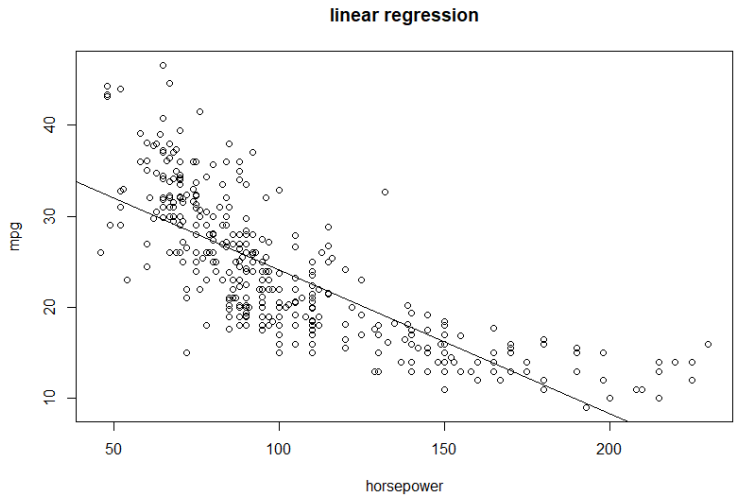
(iii)

Negative, because the coefficient corresponds to the predictor is lower than zero.

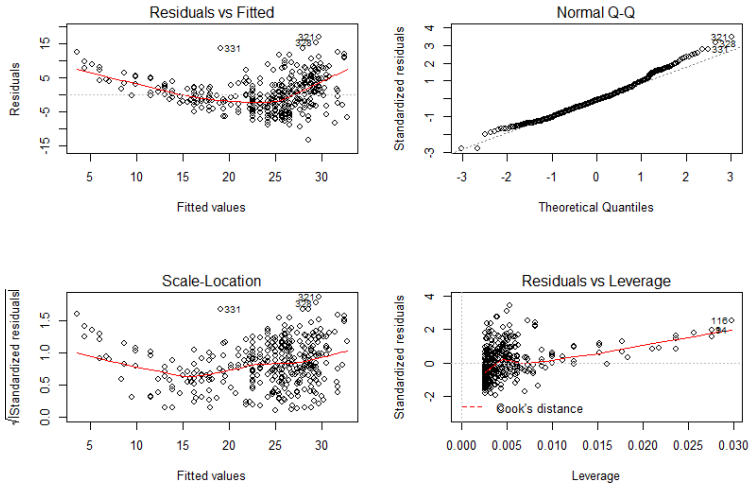
(iv)

```
> predict(model,data.frame(horsepower=c(98)),interval="confidence")
       fit      lwr      upr
1 24.46708 23.97308 24.96108
> predict(model,data.frame(horsepower=c(98)),interval="prediction")
       fit     lwr      upr
1 24.46708 14.8094 34.12476
```

(b.)



(c.)



After standardization, and according to its Q-Q plot, we can discover that the mpg we predict seems Normally distributed.