Homework8

Yanyu Zheng yz2690

April 6, 2016

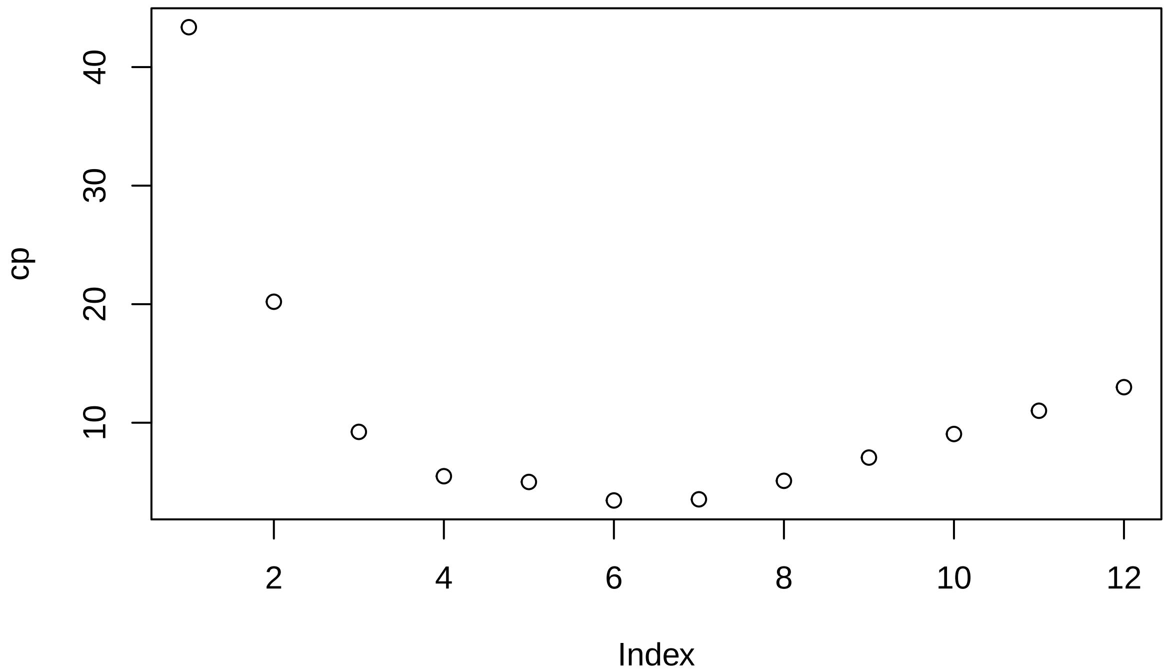
1. Chapter 12, problem 17

a

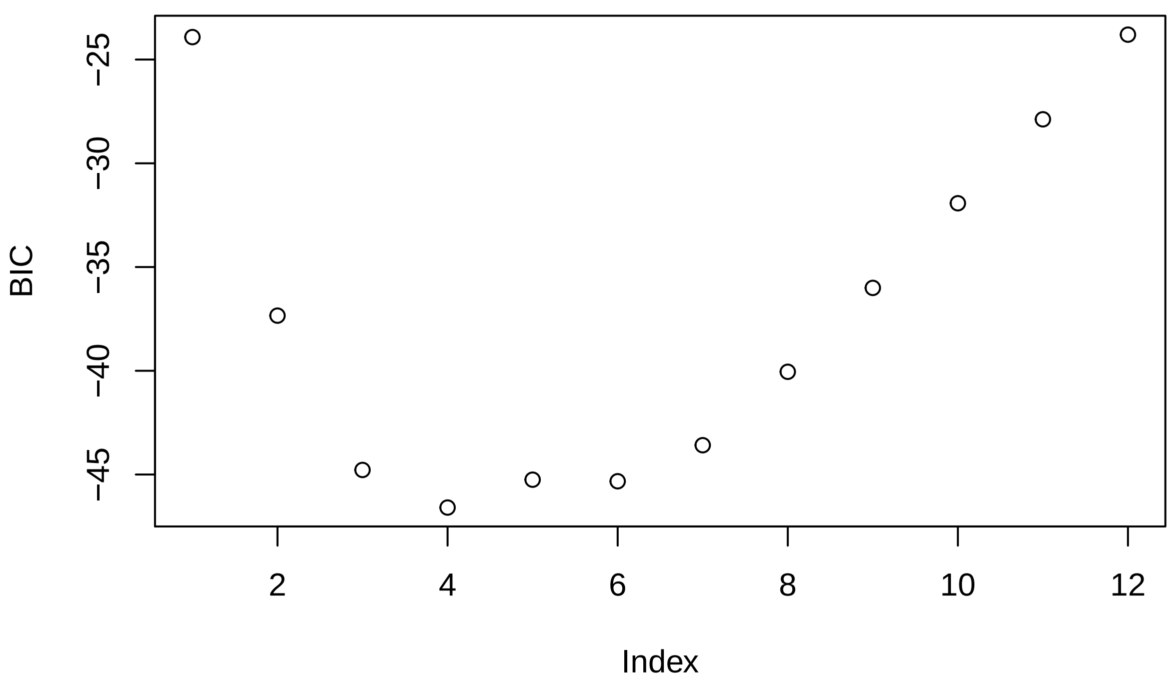
```
library("Sleuth3")
library(leaps)
library("dplyr")
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data("ex1217")

fit11 = regsubsets(Mortality~.,data=ex1217[,2:14],nvmax=13,nbest=1)
cp = summary(fit11)$cp
BIC = summary(fit11)$bic
plot(cp)
```




```
plot(BIC)
```




```
k = which.min(cp)
coef = names(coef(fit11,k)[-1])
coef = c("Mortality",coef)
data11 = ex1217[,coef]
fit11 = lm(Mortality~.,data11)
coef = c(coef,"HC","NOX","SO2")
data12 = ex1217[,coef]
data12 = mutate(data12,HC = log(HC),NOX = log(NOX), SO2 = log(SO2))
```

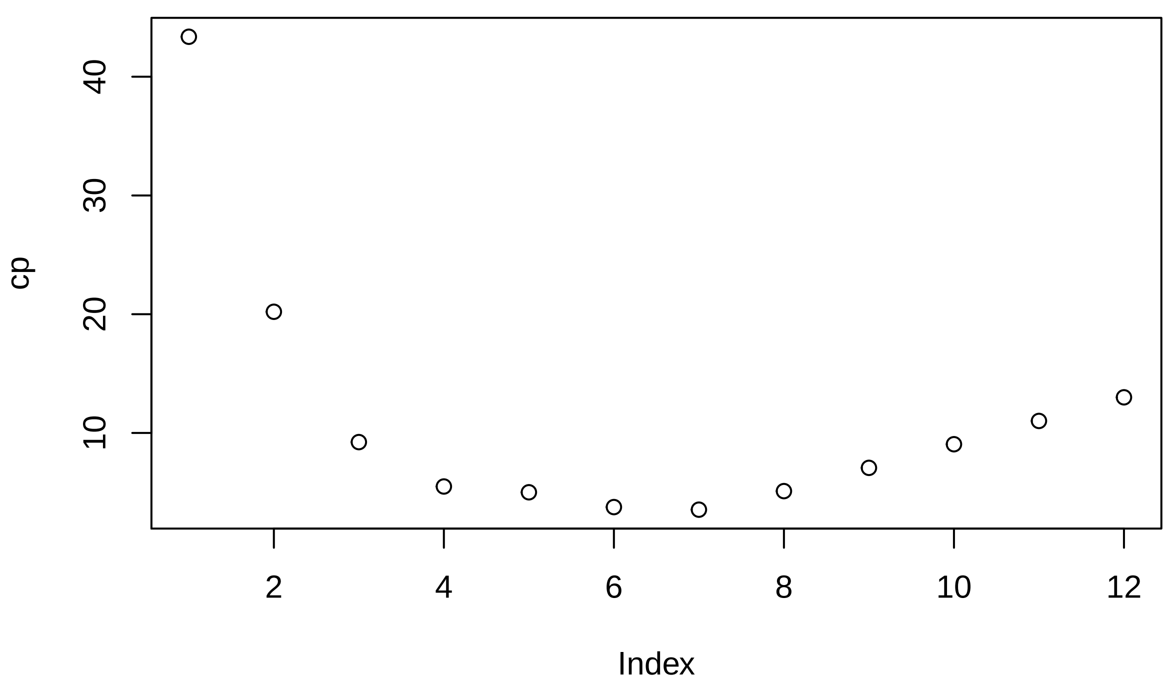
```r
fit12 = lm(Mortality~.,data12)
anova(fit11,fit12)
```

```
## Analysis of Variance Table
##
## Model 1: Mortality ~ Precip + JanTemp + JulyTemp + Educ + Density + NonWhite
## Model 2: Mortality ~ Precip + JanTemp + JulyTemp + Educ + Density + NonWhite +
##     HC + NOX + SO2
##   Res.Df   RSS Df Sum of Sq     F   Pr(>F)
## 1     53 66518
## 2     50 52712  3     13806 4.365 0.008313 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

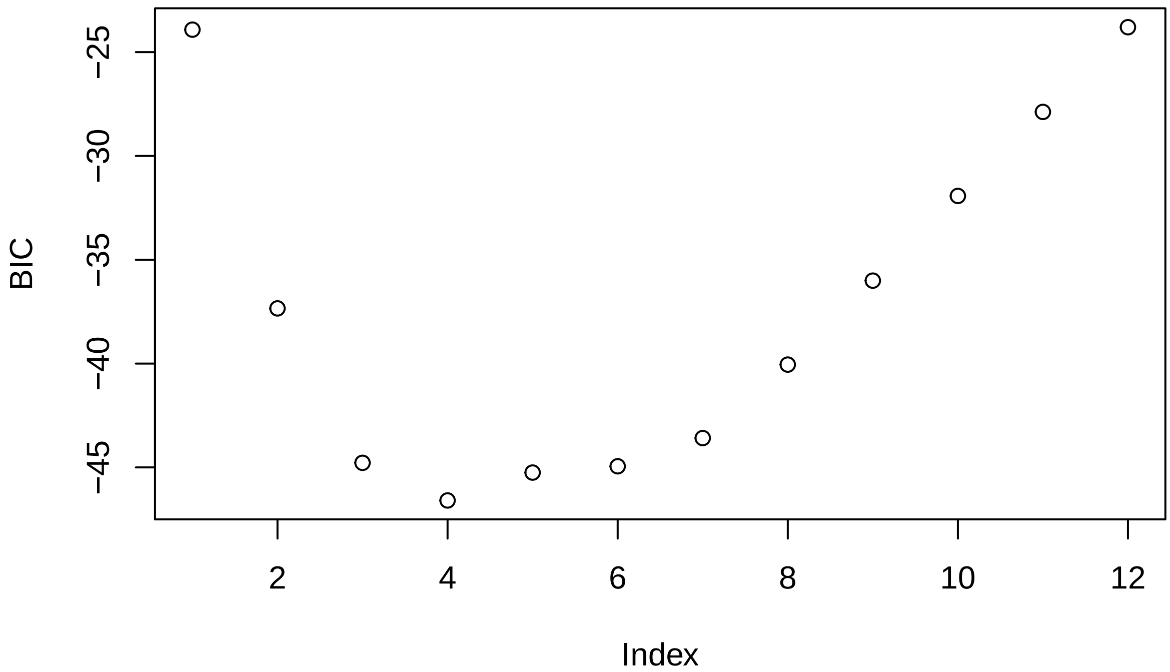
The p value is 0.008313

b

```r
fit11 = regsubsets(Mortality~.,data=ex1217[,2:14],nvmax=13,nbest=1,method = "forward")
cp = summary(fit11)$cp
BIC = summary(fit11)$bic
plot(cp)
```




```r
plot(BIC)
```




```r
k = which.min(cp)
coef = names(coef(fit11,k)[-1])
coef = c("Mortality",coef)
data11 = ex1217[,coef]
fit11 = lm(Mortality~.,data11)
coef = c(coef,"HC","NOX","SO2")
data12 = ex1217[,coef]
data12 = mutate(data12,HC = log(HC),NOX = log(NOX), SO2 = log(SO2))
fit12 = lm(Mortality~.,data12)
anova(fit11,fit12)
```

```
## Analysis of Variance Table
##
## Model 1: Mortality ~ Precip + JanTemp + JulyTemp + House + Educ + Density +
##     NonWhite
## Model 2: Mortality ~ Precip + JanTemp + JulyTemp + House + Educ + Density +
##     NonWhite + HC + NOX + SO2
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1     52 63955
## 2     49 50403  3     13552 4.3915 0.008162 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is 0.008162, which is almost the same.

2. Chapter 12, problem 20

```r
data(ex1220)
data = select(ex1220,-Island,-Native)
```

```
fit21 = regsubsets(Total~.,data,nvmax=7,nbest=1,force.in = "Area")
k = which.min(summary(fit21)$cp)
coef = names(coef(fit21,k)[-1])
```

For the total number of species, the variable selected were "Area" "Elev" "DistSc" "AreaNear".

```
data = select(ex1220,-Island,-Total)
fit21 = regsubsets(Native~.,data,nvmax=7,nbest=1,force.in = "Area")
k = which.min(summary(fit21)$cp)
coef = names(coef(fit21,k)[-1])
```

For the natives, the variable selected were "Area" "Elev" "DistSc" "AreaNear".

```
data = ex1220 %>%
   mutate(Native = Total-Native)%>%
   select(-Island,-Total)
fit21 = regsubsets(Native~.,data,nvmax=7,nbest=1,force.in = "Area")
k = which.min(summary(fit21)$cp)
coef = names(coef(fit21,k)[-1])
```

For the non-natives, the variable selected were "Area" "Elev" "AreaNear".

3. Chapter 20, problem 11

a

```
data(ex2011)
logit <- glm(Failure ~ Temperature, ex2011,family = "binomial")
summary(logit)
```

```
##
## Call:
## glm(formula = Failure ~ Temperature, family = "binomial", data = ex2011)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2125  -0.8253  -0.4706   0.5907   2.0512
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 10.87535    5.70291   1.907   0.0565 .
## Temperature -0.17132    0.08344  -2.053   0.0400 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28.975  on 23  degrees of freedom
## Residual deviance: 23.030  on 22  degrees of freedom
## AIC: 27.03
##
## Number of Fisher Scoring iterations: 4
```

Intercept: 10.87535 sd: 5.70291

Temperature: -0.17132 sd: 0.08344

b

```r
pnorm(-2.053)
```

```
## [1] 0.02003629
```

The one sided p value is 0.02.

c

```r
anova(logit,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Failure
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                           23     28.975
## Temperature  1   5.9441        22     23.030  0.01477 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is 0.01477.

d

```r
confint.default(logit)
```

```
##                  2.5 %       97.5 %
## (Intercept) -0.3021540 22.052852259
## Temperature -0.3348573 -0.007783746
```

Confidence interval for Temperature is (-0.3348573, -0.007783746)

e

```
prob = predict.glm(logit, data.frame(Temperature = 31), type="response")
problogit = predict.glm(logit, data.frame(Temperature = 31))
```

The logit of the probability is 5.564414. The prediction of the probability of failure is 0.9961828.

f

Because the standard deviation of the estimate is positively related to the squared distance of the new obs and X bar. So a new obs out of the range of the exsisting Xs would lead to a prediction with high variance.

4. Chapter 20, problem 15

a

```
attach(ex2015)
p = vector()
for(i in 1:7){
  p = c(p,t.test(ex2015[1:30,i+1],ex2015[31:60,i+1],alternative="less")$p.value)
}
```

For ring 1 to 7 the p value is $4.8124666 \times 10^{-4}, 0.011629, 3.447054 \times 10^{-5}, 0.0046197, 1.3555605 \times 10^{-4}, 2.6165661 \times 10^{-4}, 0.2481928$ **.**

b

```
logit = glm(Site~.,ex2015,family = "binomial")
summary(logit)
```

```
##
## Call:
## glm(formula = Site ~ ., family = "binomial", data = ex2015)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.39021  -0.74745  -0.01854   0.72308   1.90308
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  9.80304    3.38934   2.892  0.00382 **
## PctRing1    -0.05708    0.03713  -1.537  0.12422
## PctRing2     0.11730    0.04990   2.351  0.01873 *
## PctRing3    -0.12181    0.05199  -2.343  0.01913 *
## PctRing4     0.01694    0.04277   0.396  0.69201
## PctRing5    -0.03296    0.03905  -0.844  0.39875
## PctRing6    -0.10891    0.06631  -1.642  0.10051
## PctRing7     0.05157    0.03619   1.425  0.15415
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.178  on 59  degrees of freedom
## Residual deviance: 52.107  on 52  degrees of freedom
## AIC: 68.107
##
## Number of Fisher Scoring iterations: 6
```

As we can see, only ring2 and ring3 are significant, which means within 1.4km radius.