PSTAT 126 HW6

Mujie Wang

06/08/2018

Problem 1

```
data(state)
state.x77 = as.data.frame(state.x77)
dim(state.x77)
```

```
## [1] 50  8
```

```
colnames(state.x77)[4] = 'Life.Exp'
```

(a)

```
mod0 = lm(Life.Exp ~1, data = state.x77)
mod1 = lm(Life.Exp ~., data = state.x77)
step(mod0, scope = list(upper = mod1))
```

```
## Start:  AIC=30.44
## Life.Exp ~ 1
##
##              Df Sum of Sq    RSS     AIC
## + Murder      1    53.838 34.461 -14.609
## + Illiteracy  1    30.578 57.721  11.179
## + `HS Grad`   1    29.931 58.368  11.737
## + Income      1    10.223 78.076  26.283
## + Frost       1     6.064 82.235  28.878
## <none>                    88.299  30.435
## + Area        1     1.017 87.282  31.856
## + Population  1     0.409 87.890  32.203
##
## Step:  AIC=-14.61
## Life.Exp ~ Murder
##
##              Df Sum of Sq    RSS     AIC
## + `HS Grad`   1     4.691 29.770 -19.925
## + Population  1     4.016 30.445 -18.805
## + Frost       1     3.135 31.327 -17.378
## + Income      1     2.405 32.057 -16.226
## <none>                    34.461 -14.609
## + Area        1     0.470 33.992 -13.295
## + Illiteracy  1     0.273 34.188 -13.007
## - Murder      1    53.838 88.299  30.435
##
## Step:  AIC=-19.93
## Life.Exp ~ Murder + `HS Grad`
##
##              Df Sum of Sq    RSS     AIC
## + Frost       1    4.3987 25.372 -25.920
## + Population  1    3.3405 26.430 -23.877
## <none>                    29.770 -19.925
## + Illiteracy  1    0.4419 29.328 -18.673
## + Area        1    0.2775 29.493 -18.394
## + Income      1    0.1022 29.668 -18.097
## - `HS Grad`   1    4.6910 34.461 -14.609
## - Murder      1   28.5974 58.368  11.737
##
## Step:  AIC=-25.92
## Life.Exp ~ Murder + `HS Grad` + Frost
##
##              Df Sum of Sq    RSS     AIC
## + Population  1     2.064 23.308 -28.161
## <none>                    25.372 -25.920
## + Income      1     0.182 25.189 -24.280
## + Illiteracy  1     0.172 25.200 -24.259
## + Area        1     0.026 25.346 -23.970
## - Frost       1     4.399 29.770 -19.925
## - `HS Grad`   1     5.955 31.327 -17.378
## - Murder      1    32.756 58.128  13.531
##
## Step:  AIC=-28.16
## Life.Exp ~ Murder + `HS Grad` + Frost + Population
##
##              Df Sum of Sq    RSS     AIC
## <none>                    23.308 -28.161
## + Income      1     0.006 23.302 -26.174
## + Illiteracy  1     0.004 23.304 -26.170
## + Area        1     0.001 23.307 -26.163
## - Population  1     2.064 25.372 -25.920
## - Frost       1     3.122 26.430 -23.877
## - `HS Grad`   1     5.112 28.420 -20.246
## - Murder      1    34.816 58.124  15.528
```

```
##
## Call:
## lm(formula = Life.Exp ~ Murder + `HS Grad` + Frost + Population,
##     data = state.x77)
##
## Coefficients:
## (Intercept)       Murder    `HS Grad`        Frost   Population
##   7.103e+01   -3.001e-01    4.658e-02   -5.943e-03    5.014e-05
```

According to AIC, the “best” model has the four predictors: Murder, HS Grad, Frost and Population.

(b)

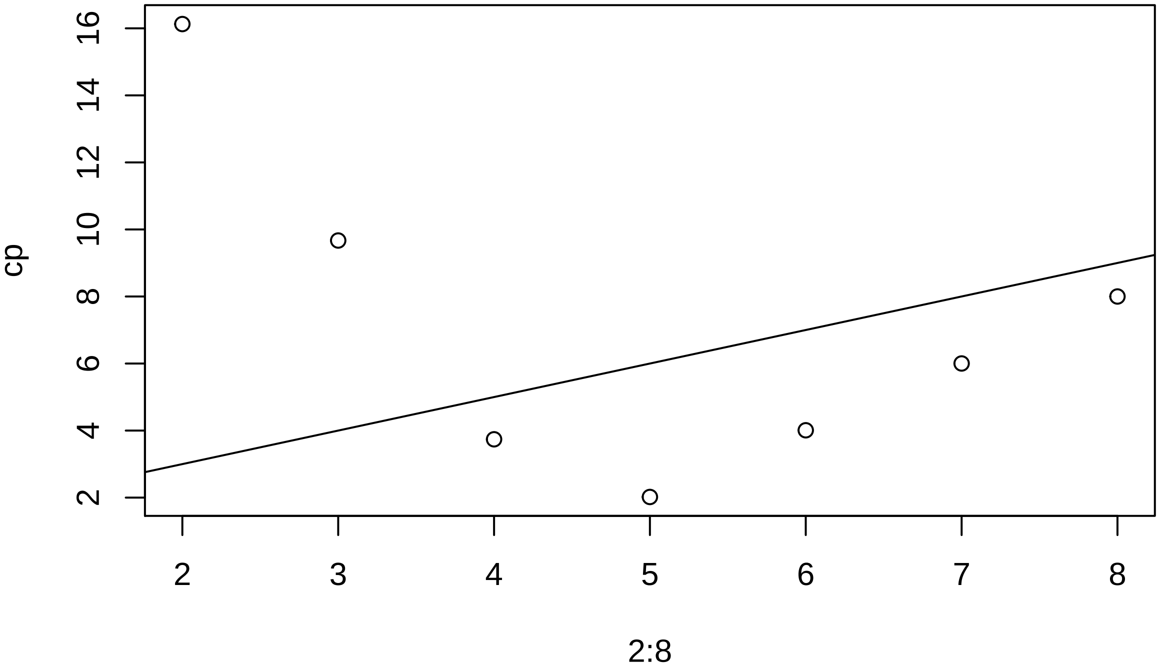
```
library(leaps)
mod = regsubsets(state.x77[,-4], state.x77[,4])
sum.mod = summary(mod)
sum.mod$which
```

```
##   (Intercept) Population Income Illiteracy Murder HS Grad Frost  Area
## 1        TRUE      FALSE  FALSE      FALSE   TRUE   FALSE FALSE FALSE
## 2        TRUE      FALSE  FALSE      FALSE   TRUE    TRUE FALSE FALSE
## 3        TRUE      FALSE  FALSE      FALSE   TRUE    TRUE  TRUE FALSE
## 4        TRUE       TRUE  FALSE      FALSE   TRUE    TRUE  TRUE FALSE
## 5        TRUE       TRUE   TRUE      FALSE   TRUE    TRUE  TRUE FALSE
## 6        TRUE       TRUE   TRUE       TRUE   TRUE    TRUE  TRUE FALSE
## 7        TRUE       TRUE   TRUE       TRUE   TRUE    TRUE  TRUE  TRUE
```

```
sum.mod$cp
```

```
## [1] 16.126760  9.669894  3.739878  2.019659  4.008737  6.001959  8.000000
```

```
cp = sum.mod$cp
plot(2:8, cp)
abline(1,1)
```

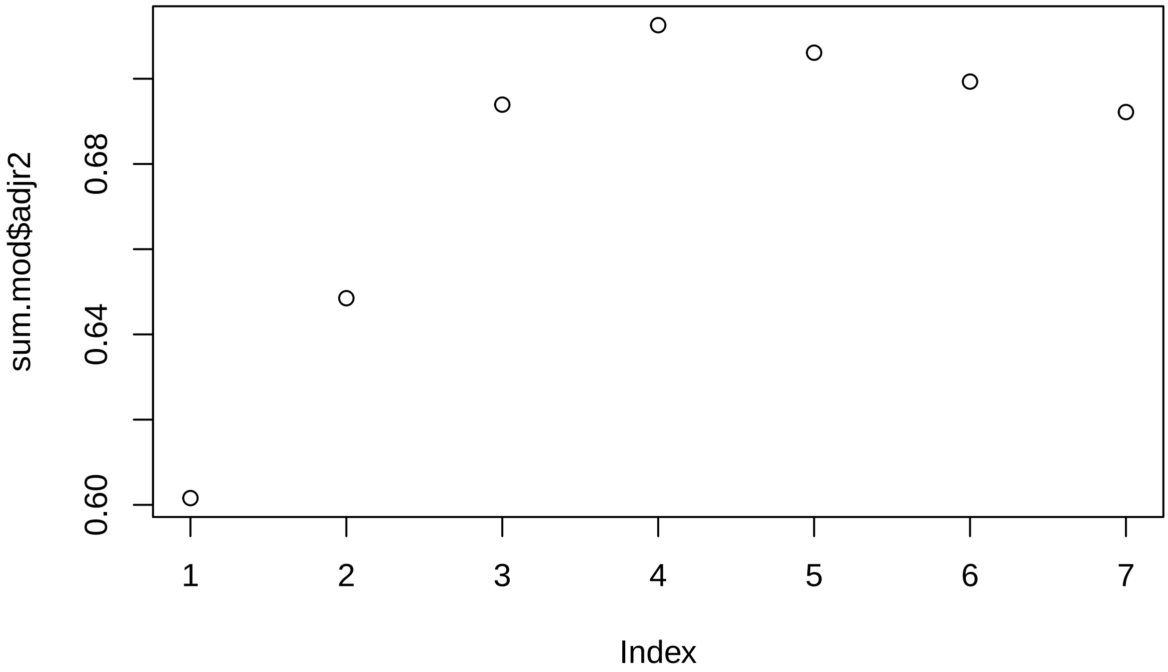


The Cp criterion suggests the ???best??? model with four predictors: Murder, HS Grad, Frost and Population. Cp is close to p when p=4, and the model with p = 4 is simpler than the model with p = 5 or p = 6. (c)

```
sum.mod$adjr2
```

```
## [1] 0.6015893 0.6484991 0.6939230 0.7125690 0.7061129 0.6993268 0.6921823
```

```
plot(sum.mod$adjr2)
```



The adjusted R2 criterion suggests the ???best??? model with four predictors: Murder, HS Grad, Frost and Population. This model has the largest adjusted R2 value.

(d)

```
mod.adjr2 = lm(Life.Exp ~ Population + Murder + `HS Grad` + Frost, data = state.x77)
hv = hatvalues(mod.adjr2)
p=5
n = nrow(state.x77)
which(hv > 3*p/n)
```

```
## California
##          5
```

```
rs = rstudent(mod.adjr2)
which(abs(rs) == max(abs(rs)))
```

```
## Hawaii
##     11
```

```
dfs = dffits(mod.adjr2)
which(abs(dfs) == max(abs(dfs)))
```

```
## Hawaii
##     11
```

California has the largest leverage value. Hawaii has the largest externally studentized residual and DIFFITS absolue value.

```
mod.delete = lm(Life.Exp ~ Population + Murder + `HS Grad` + Frost, data = state.x77[-11,])
summary(mod.delete)
```

```
##
## Call:
## lm(formula = Life.Exp ~ Population + Murder + `HS Grad` + Frost,
##     data = state.x77[-11, ])
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.48967 -0.50158  0.01999  0.54355  1.11810
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.106e+01  8.998e-01  78.966  < 2e-16 ***
## Population   6.363e-05  2.431e-05   2.618   0.0121 *
## Murder      -2.906e-01  3.477e-02  -8.357 1.24e-10 ***
## `HS Grad`    3.728e-02  1.447e-02   2.576   0.0134 *
## Frost       -3.099e-03  2.545e-03  -1.218   0.2297
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6796 on 44 degrees of freedom
## Multiple R-squared:  0.7483, Adjusted R-squared:  0.7254
## F-statistic: 32.71 on 4 and 44 DF,  p-value: 1.15e-12
```

Problem 2

```r
library(MASS)
library(alr4)
```

```
## Loading required package: car
```

```
## Warning: package 'car' was built under R version 3.4.4
```

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

```
## Warning: package 'carData' was built under R version 3.4.4
```

```
## Loading required package: effects
```

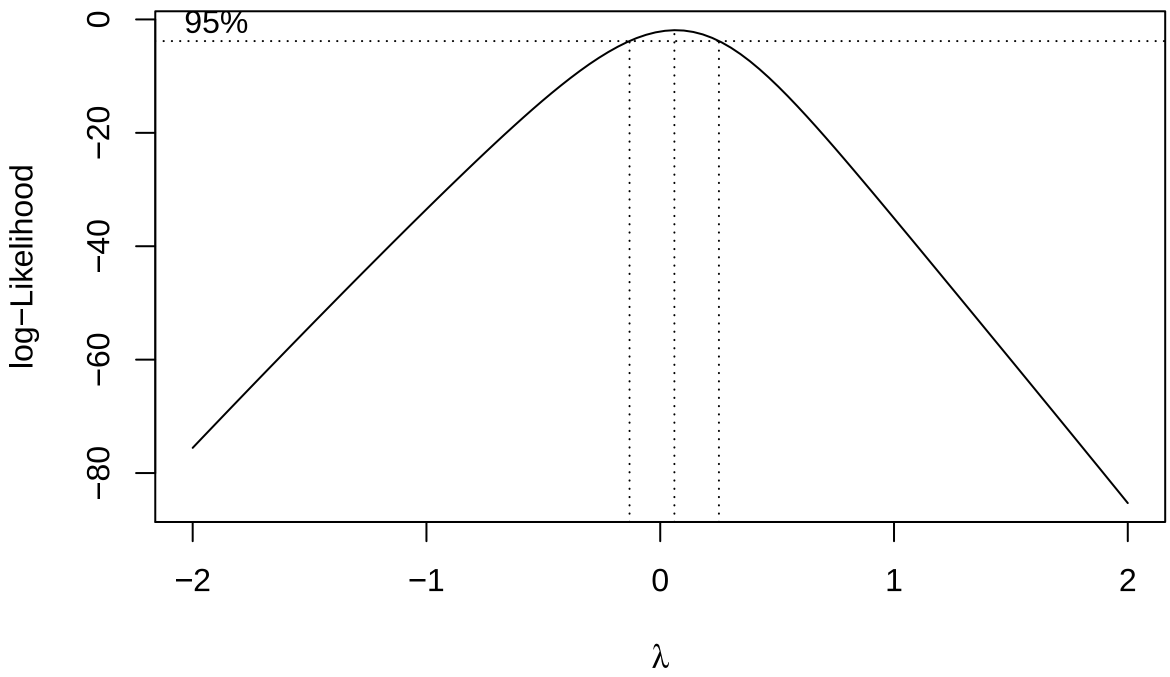
```
## Warning: package 'effects' was built under R version 3.4.4
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```r
data(lathe1)
```

(a)

```r
boxcox(Life ~ Speed + Feed + I(Speed^2) +I(Feed^2) + Speed*Feed, data = lathe1)
```



lamda = 0 is in the 95% confidence interval for lamda, and when lamda = 0, the plot suggests log-transforming the response.

(b)

```
mod.reduced = lm(log(Life)~1, data = lathe1)
mod.full = lm(log(Life)~ Speed + Feed + I(Speed^2) + I(Feed^2) + Speed * Feed, data = lathe1)
anova(mod.reduced, mod.full)
```

```
## Analysis of Variance Table
##
## Model 1: log(Life) ~ 1
## Model 2: log(Life) ~ Speed + Feed + I(Speed^2) + I(Feed^2) + Speed * Feed
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     19 41.533
## 2     14  1.237  5    40.296 91.236 3.551e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(c) The Null hypothesis means the response Life is not linearly related to Speed.

(d)

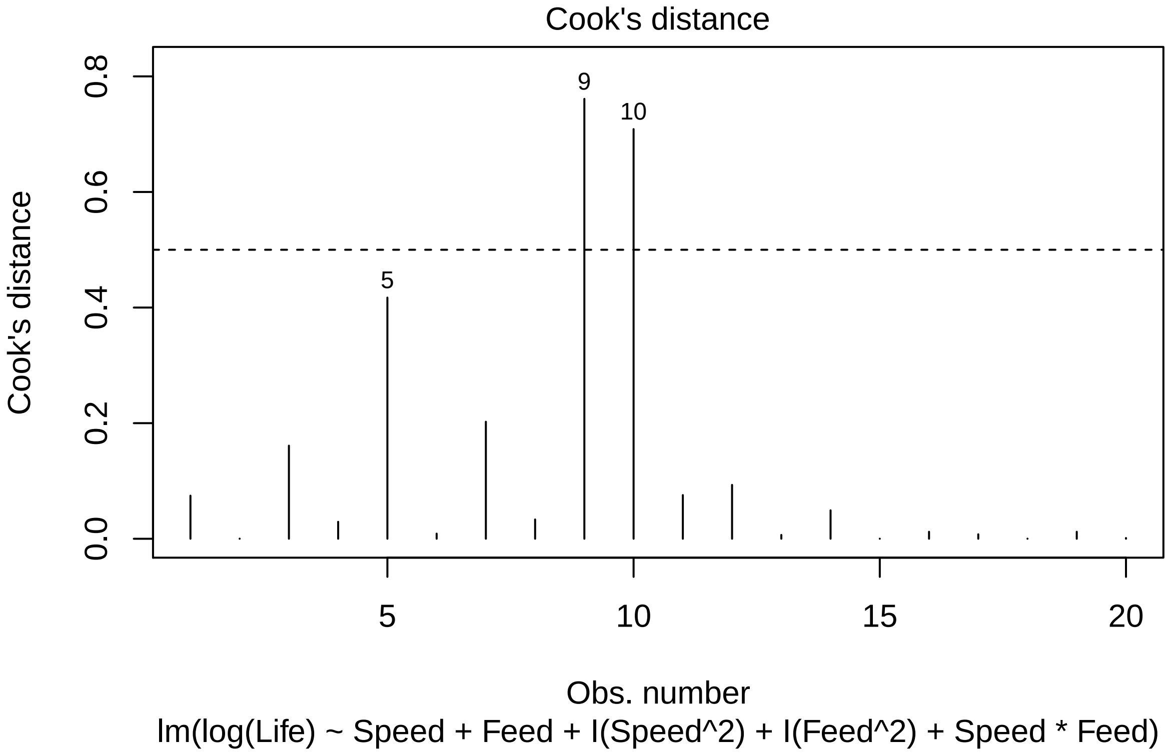
```
mod.reduced2 = lm(log(Life)~Feed + I(Feed^2), data = lathe1)
anova(mod.reduced2, mod.full)
```

```
## Analysis of Variance Table
##
## Model 1: log(Life) ~ Feed + I(Feed^2)
## Model 2: log(Life) ~ Speed + Feed + I(Speed^2) + I(Feed^2) + Speed * Feed
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     17 32.300
## 2     14  1.237  3    31.063 117.22 3.726e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is very close to 0, the we reject Ho.

(e)

```
fit1 = lm(log(Life) ~ Speed + Feed + I(Speed^2) +I(Feed^2) + Speed*Feed, data = lathe1)
plot(fit1, which = 4)
abline(h = 0.5, lty = 2)
```



```
cds = cooks.distance(fit1)
sort(cds, decreasing = TRUE)
```

```
##            9           10            5            7            3
## 0.7611370235 0.7088115474 0.4172638143 0.2024479551 0.1611290980
##           12           11            1           14            8
## 0.0932562838 0.0755462115 0.0745581876 0.0491977930 0.0333705363
##            4           16           19            6           17
## 0.0293444172 0.0121013330 0.0121013330 0.0089104068 0.0077362334
##           13           20            2           15           18
## 0.0066483194 0.0012883357 0.0002358999 0.0001916341 0.0001916341
```

The 9 and 10 cases have the largest Cook???s distance Di. Di for both cases are larger than 0.5, which suggests that the two cases might be in uential.

```
fit2 = lm(log(Life) ~ Speed + Feed + I(Speed^2) +I(Feed^2) + Speed*Feed, data = lathe1[-c(9,10),])
summary(fit1)
```

```
## 
## Call:
## lm(formula = log(Life) ~ Speed + Feed + I(Speed^2) + I(Feed^2) + 
##     Speed * Feed, data = lathe1)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.43349 -0.14576 -0.02494  0.16748  0.47992
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.18809    0.10508  11.307 2.00e-08 ***
## Speed       -1.58902    0.08580 -18.520 3.04e-11 ***
## Feed        -0.79023    0.08580  -9.210 2.56e-07 ***
## I(Speed^2)   0.28808    0.10063   2.863 0.012529 *
## I(Feed^2)    0.41851    0.10063   4.159 0.000964 ***
## Speed:Feed  -0.07286    0.10508  -0.693 0.499426
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2972 on 14 degrees of freedom
## Multiple R-squared:  0.9702, Adjusted R-squared:  0.9596
## F-statistic: 91.24 on 5 and 14 DF,  p-value: 3.551e-10
```

```r
summary(fit2)
```

```
##
## Call:
## lm(formula = log(Life) ~ Speed + Feed + I(Speed^2) + I(Feed^2) +
##     Speed * Feed, data = lathe1[-c(9, 10), ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39963 -0.14660  0.00387  0.14917  0.32783
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.18809    0.08241  14.417 6.11e-09 ***
## Speed       -1.43300    0.08241 -17.388 7.10e-10 ***
## Feed        -0.79023    0.06729 -11.743 6.15e-08 ***
## I(Speed^2)   0.28022    0.12363   2.267 0.042700 *
## I(Feed^2)    0.42244    0.09217   4.583 0.000629 ***
## Speed:Feed  -0.07286    0.08241  -0.884 0.394025
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2331 on 12 degrees of freedom
## Multiple R-squared:  0.9759, Adjusted R-squared:  0.9658
## F-statistic: 97.07 on 5 and 12 DF,  p-value: 2.804e-09
```

The coeffcient most affected is the main effect for Speed, while the others stay mostly the same. Also, the standard errors are uniformly smaller using the reduced data set. The R^2 and adjusted R^2 are larger using the reduced data set.