Homework 8 Key

Subrata Paul

11/5/2020

10.1

(a) The best model from the backward selection approach is the model including the predictors `lcavol`, `lweight`, and, `svi`.

```
library(faraway)
data(prostate, package = 'faraway')
lmod <- lm(lpsa ~., data = prostate)
sumary(lmod)
```

```
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  0.6693367  1.2963875  0.5163  0.606934
## lcavol       0.5870218  0.0879203  6.6767 2.111e-09
## lweight      0.4544674  0.1700124  2.6731  0.008955
## age         -0.0196372  0.0111727 -1.7576  0.082293
## lbph         0.1070540  0.0584492  1.8316  0.070398
## svi          0.7661573  0.2443091  3.1360  0.002329
## lcp         -0.1054743  0.0910135 -1.1589  0.249638
## gleason      0.0451416  0.1574645  0.2867  0.775033
## pgg45        0.0045252  0.0044212  1.0235  0.308860
##
## n = 97, p = 9, Residual SE = 0.70842, R-Squared = 0.65
```

```
lmod <- update(lmod, .~. -gleason)
sumary(lmod)
```

```
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  0.9539260  0.8294393  1.1501  0.253190
## lcavol       0.5916145  0.0860015  6.8791 8.069e-10
## lweight      0.4482924  0.1677706  2.6721  0.008965
## age         -0.0193365  0.0110659 -1.7474  0.084018
## lbph         0.1076711  0.0581076  1.8530  0.067202
## svi          0.7577335  0.2412818  3.1405  0.002290
## lcp         -0.1044823  0.0904775 -1.1548  0.251269
## pgg45        0.0053177  0.0034326  1.5492  0.124884
##
## n = 97, p = 8, Residual SE = 0.70475, R-Squared = 0.65
```

```
lmod <- update(lmod, .~. -lcp)
sumary(lmod)
```

```
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  0.9800848  0.8306648  1.1799  0.241157
## lcavol       0.5457697  0.0764313  7.1407 2.312e-10
## lweight      0.4494499  0.1680782  2.6741  0.008900
```

```
## age         -0.0174699  0.0109674 -1.5929  0.114692
## lbph         0.1057551  0.0581914  1.8174  0.072489
## svi          0.6416661  0.2197567  2.9199  0.004424
## pgg45        0.0035276  0.0030683  1.1497  0.253309
##
## n = 97, p = 7, Residual SE = 0.70606, R-Squared = 0.65
```

```
lmod <- update(lmod, .~. -age)
sumary(lmod)
```

```
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 0.0664614  0.6059095  0.1097  0.912898
## lcavol      0.5333978  0.0766752  6.9566 5.198e-10
## lweight     0.4052673  0.1671680  2.4243  0.017314
## lbph        0.0830338  0.0568906  1.4595  0.147862
## svi         0.6537607  0.2214728  2.9519  0.004017
## pgg45       0.0025285  0.0030288  0.8348  0.405992
##
## n = 97, p = 6, Residual SE = 0.71200, R-Squared = 0.64
```

```
lmod <- update(lmod, .~. -pgg45)
sumary(lmod)
```

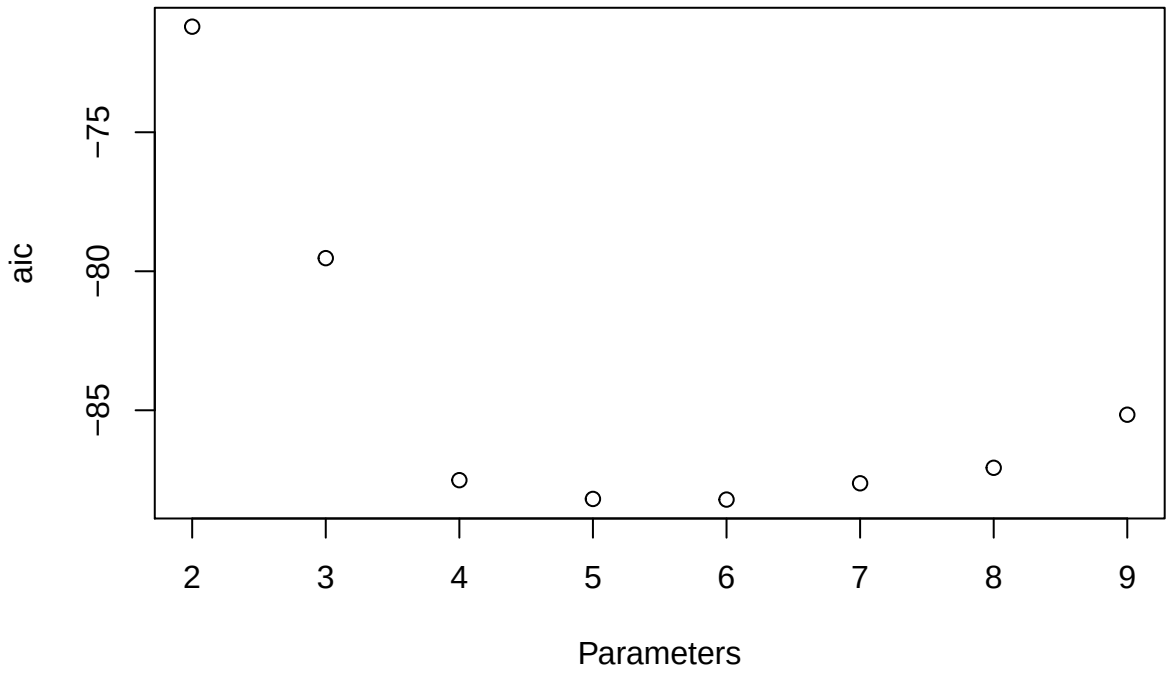
```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 0.145541   0.597473  0.2436  0.808088
## lcavol      0.549603   0.074055  7.4215 5.645e-11
## lweight     0.390876   0.166003  2.3546  0.020667
## lbph        0.090093   0.056166  1.6041  0.112130
## svi         0.711737   0.209957  3.3899  0.001031
##
## n = 97, p = 5, Residual SE = 0.71082, R-Squared = 0.64
```

```
lmod <- update(lmod, .~. -lbph)
sumary(lmod)
```

```
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -0.268093   0.543500 -0.4933  0.622984
## lcavol       0.551638   0.074668  7.3879 6.304e-11
## lweight      0.508541   0.150170  3.3864  0.001039
## svi          0.666158   0.209777  3.1756  0.002029
##
## n = 97, p = 4, Residual SE = 0.71681, R-Squared = 0.63
```

(b) According to the AIC criterion, the best model has the predictors `lcavol`, `lweight`, `age`, `lbph`, and `svi`.

```
library(leaps)
rc <- regsubsets(lpsa ~ ., data = prostate)
rcs <- summary(rc)
aic <- rcs$bic - 2:9*log(nobs(lmod)) + 2:9*2
plot(aic ~ I(2:9), xlab = 'Parameters')
```




```
rcs$which[which.min(aic),]
```

```
## (Intercept)      lcavol     lweight         age        lbph         svi
##        TRUE        TRUE        TRUE        TRUE        TRUE        TRUE
##         lcp     gleason       pgg45
##       FALSE       FALSE       FALSE
```

(c) According to the adjusted r-squared criterion, the best model includes all predictors except gleason.

```
rcs$which[which.max(rcs$adjr2),]
```

```
## (Intercept)      lcavol     lweight         age        lbph         svi
##        TRUE        TRUE        TRUE        TRUE        TRUE        TRUE
##         lcp     gleason       pgg45
##        TRUE       FALSE        TRUE
```

(d) According to Mallow's C_p criterion, the best model has the predictors `lcavol`, `lweight`, `age`, `lbph`, and `svi`.

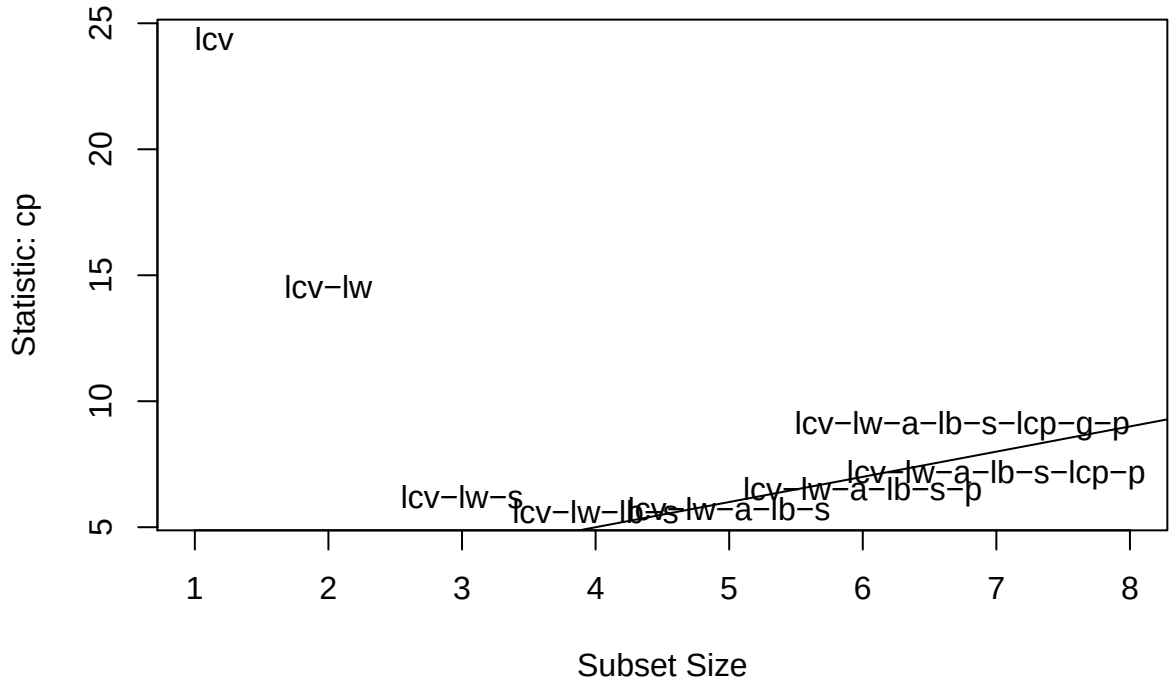
```
car::subsets(rc, statistic = 'cp', legend = F)
```

```
## Registered S3 methods overwritten by 'car':
##   method                           from
##   influence.merMod                 lme4
##   cooks.distance.influence.merMod lme4
##   dfbeta.influence.merMod          lme4
##   dfbetas.influence.merMod         lme4
```

```
##         Abbreviation
## lcavol           lcv
## lweight           lw
## age                a
## lbph              lb
## svi                s
## lcp              lcp
## gleason            g
```

```
## pgg45               p
```

```
abline(1,1)
```



10.4

Comparing the second-order model to the first-order model (using the F-test, AIC, BIC, and adjusted R^2 criteria), the second-order model is preferred. We shouldn't simplify the model.

```
data(trees)
lmod1 = lm(log(Volume) ~ Girth + Height, data = trees)
lmod2 = lm(log(Volume) ~ Girth + Height + I(Girth^2) + I(Height^2) + Girth:Height, data = trees)
anova(lmod1, lmod2)
```

```
## Analysis of Variance Table
##
## Model 1: log(Volume) ~ Girth + Height
## Model 2: log(Volume) ~ Girth + Height + I(Girth^2) + I(Height^2) + Girth:Height
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1     28 0.26214
## 2     25 0.17932  3  0.082817 3.8486 0.02156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
comp = data.frame(AIC = c(AIC(lmod1),AIC(lmod2)), BIC = c(BIC(lmod1),BIC(lmod2)),
                  AR2 = c(summary(lmod1)$adj, summary(lmod2)$adj))
row.names(comp) = c('Linear','Quadratic')
knitr::kable(comp)
```

	AIC	BIC	AR2
Linear	-51.98466	-46.24871	0.9661964
Quadratic	-57.75516	-47.71725	0.9741010