**STATS 500 - Homework 7**

Use the prostate data with lpsa as the response and the other variables as predictors. Implement the following variable selection methods to determine the “best” model:

1. Backward Elimination

Start from the full model with all the other variables as predictors.

Call:

lm(formula = lpsa ~ ., data = prostate)

Residuals:

Min 1Q Median 3Q Max

-1.7331 -0.3713 -0.0170 0.4141 1.6381

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.669337 1.296387 0.516 0.60693

lcavol 0.587022 0.087920 6.677 2.11e-09 \*\*\*

lweight 0.454467 0.170012 2.673 0.00896 \*\*

age -0.019637 0.011173 -1.758 0.08229 .

lbph 0.107054 0.058449 1.832 0.07040 .

svi 0.766157 0.244309 3.136 0.00233 \*\*

lcp -0.105474 0.091013 -1.159 0.24964

gleason 0.045142 0.157465 0.287 0.77503

pgg45 0.004525 0.004421 1.024 0.30886

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7084 on 88 degrees of freedom

Multiple R-squared: 0.6548, Adjusted R-squared: 0.6234

F-statistic: 20.86 on 8 and 88 DF, p-value: < 2.2e-16

Drop largest p-value, which is gleason.

Call:

lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +

pgg45, data = prostate)

Residuals:

Min 1Q Median 3Q Max

-1.73117 -0.38137 -0.01728 0.43364 1.63513

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.953926 0.829439 1.150 0.25319

lcavol 0.591615 0.086001 6.879 8.07e-10 \*\*\*

lweight 0.448292 0.167771 2.672 0.00897 \*\*

age -0.019336 0.011066 -1.747 0.08402 .

lbph 0.107671 0.058108 1.853 0.06720 .

svi 0.757734 0.241282 3.140 0.00229 \*\*

lcp -0.104482 0.090478 -1.155 0.25127

pgg45 0.005318 0.003433 1.549 0.12488

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7048 on 89 degrees of freedom

Multiple R-squared: 0.6544, Adjusted R-squared: 0.6273

F-statistic: 24.08 on 7 and 89 DF, p-value: < 2.2e-16

Continue dropping lcp.

Call:

lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + pgg45,

data = prostate)

Residuals:

Min 1Q Median 3Q Max

-1.77711 -0.41708 0.00002 0.40676 1.59681

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.980085 0.830665 1.180 0.24116

lcavol 0.545770 0.076431 7.141 2.31e-10 \*\*\*

lweight 0.449450 0.168078 2.674 0.00890 \*\*

age -0.017470 0.010967 -1.593 0.11469

lbph 0.105755 0.058191 1.817 0.07249 .

svi 0.641666 0.219757 2.920 0.00442 \*\*

pgg45 0.003528 0.003068 1.150 0.25331

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7061 on 90 degrees of freedom

Multiple R-squared: 0.6493, Adjusted R-squared: 0.6259

F-statistic: 27.77 on 6 and 90 DF, p-value: < 2.2e-16

Continue dropping pgg45

Call:

lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)

Residuals:

Min 1Q Median 3Q Max

-1.83505 -0.39396 0.00414 0.46336 1.57888

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.95100 0.83175 1.143 0.255882

lcavol 0.56561 0.07459 7.583 2.77e-11 \*\*\*

lweight 0.42369 0.16687 2.539 0.012814 \*

age -0.01489 0.01075 -1.385 0.169528

lbph 0.11184 0.05805 1.927 0.057160 .

svi 0.72095 0.20902 3.449 0.000854 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7073 on 91 degrees of freedom

Multiple R-squared: 0.6441, Adjusted R-squared: 0.6245

F-statistic: 32.94 on 5 and 91 DF, p-value: < 2.2e-16

Continue dropping age

Call:

lm(formula = lpsa ~ lcavol + lweight + lbph + svi, data = prostate)

Residuals:

Min 1Q Median 3Q Max

-1.82653 -0.42270 0.04362 0.47041 1.48530

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.14554 0.59747 0.244 0.80809

lcavol 0.54960 0.07406 7.422 5.64e-11 \*\*\*

lweight 0.39088 0.16600 2.355 0.02067 \*

lbph 0.09009 0.05617 1.604 0.11213

svi 0.71174 0.20996 3.390 0.00103 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7108 on 92 degrees of freedom

Multiple R-squared: 0.6366, Adjusted R-squared: 0.6208

F-statistic: 40.29 on 4 and 92 DF, p-value: < 2.2e-16

Continue dropping lbph

Call:

lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)

Residuals:

Min 1Q Median 3Q Max

-1.72964 -0.45764 0.02812 0.46403 1.57013

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.26809 0.54350 -0.493 0.62298

lcavol 0.55164 0.07467 7.388 6.3e-11 \*\*\*

lweight 0.50854 0.15017 3.386 0.00104 \*\*

svi 0.66616 0.20978 3.176 0.00203 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7168 on 93 degrees of freedom

Multiple R-squared: 0.6264, Adjusted R-squared: 0.6144

F-statistic: 51.99 on 3 and 93 DF, p-value: < 2.2e-16

Final model：

1. AIC

Start: AIC=-58.32

lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +

pgg45

Df Sum of Sq RSS AIC

- gleason 1 0.0412 44.204 -60.231

- pgg45 1 0.5258 44.689 -59.174

- lcp 1 0.6740 44.837 -58.853

<none> 44.163 -58.322

- age 1 1.5503 45.713 -56.975

- lbph 1 1.6835 45.847 -56.693

- lweight 1 3.5861 47.749 -52.749

- svi 1 4.9355 49.099 -50.046

- lcavol 1 22.3721 66.535 -20.567

Step: AIC=-60.23

lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45

Df Sum of Sq RSS AIC

- lcp 1 0.6623 44.867 -60.789

<none> 44.204 -60.231

- pgg45 1 1.1920 45.396 -59.650

- age 1 1.5166 45.721 -58.959

- lbph 1 1.7053 45.910 -58.560

- lweight 1 3.5462 47.750 -54.746

- svi 1 4.8984 49.103 -52.037

- lcavol 1 23.5039 67.708 -20.872

Step: AIC=-60.79

lpsa ~ lcavol + lweight + age + lbph + svi + pgg45

Df Sum of Sq RSS AIC

- pgg45 1 0.6590 45.526 -61.374

<none> 44.867 -60.789

- age 1 1.2649 46.131 -60.092

- lbph 1 1.6465 46.513 -59.293

- lweight 1 3.5647 48.431 -55.373

- svi 1 4.2503 49.117 -54.009

- lcavol 1 25.4189 70.285 -19.248

Step: AIC=-61.37

lpsa ~ lcavol + lweight + age + lbph + svi

Df Sum of Sq RSS AIC

<none> 45.526 -61.374

- age 1 0.9592 46.485 -61.352

- lbph 1 1.8568 47.382 -59.497

- lweight 1 3.2251 48.751 -56.735

- svi 1 5.9517 51.477 -51.456

- lcavol 1 28.7665 74.292 -15.871

Call:

lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)

Coefficients:

(Intercept) lcavol lweight age lbph svi

0.95100 0.56561 0.42369 -0.01489 0.11184 0.72095

Final model：

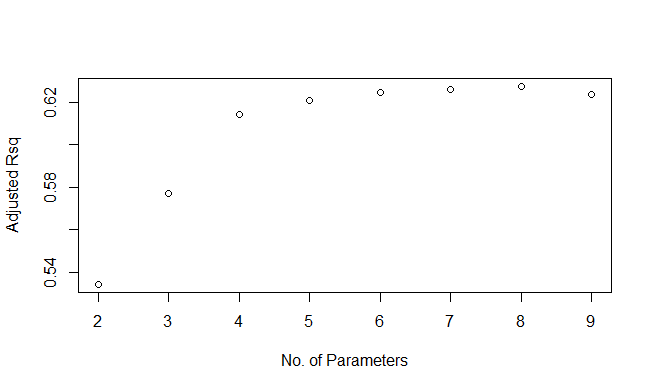
1. Adjusted R2

Subset selection object

Call: regsubsets.formula(lpsa ~ ., data = prostate)

8 Variables (and intercept)

Forced in Forced out

lcavol FALSE FALSE

lweight FALSE FALSE

age FALSE FALSE

lbph FALSE FALSE

svi FALSE FALSE

lcp FALSE FALSE

gleason FALSE FALSE

pgg45 FALSE FALSE

1 subsets of each size up to 8

Selection Algorithm: exhaustive

lcavol lweight age lbph svi lcp gleason pgg45

1 ( 1 ) "\*" " " " " " " " " " " " " " "

2 ( 1 ) "\*" "\*" " " " " " " " " " " " "

3 ( 1 ) "\*" "\*" " " " " "\*" " " " " " "

4 ( 1 ) "\*" "\*" " " "\*" "\*" " " " " " "

5 ( 1 ) "\*" "\*" "\*" "\*" "\*" " " " " " "

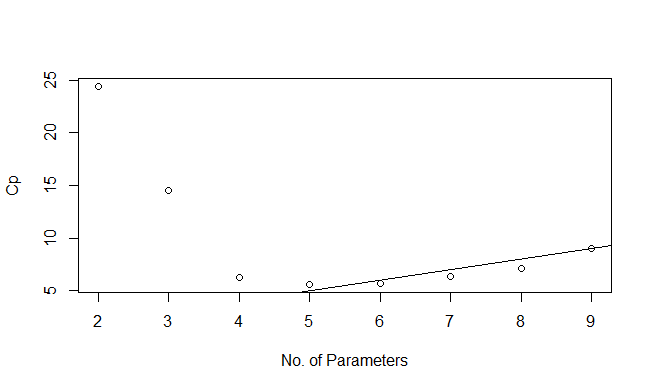
6 ( 1 ) "\*" "\*" "\*" "\*" "\*" " " " " "\*"

7 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" " " "\*"

8 ( 1 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*" "\*"

When 7 predictors are in the model, the adjusted R2 get maximum.

Final model:

1. Mallows’ Cp

When there are four parameters in the model. Cp reaches minimum.

Select p on or below the line, which is 5.

Final model：

Similarities:

All of the selected models have three predictors in common: lcavol, lweight and svi.

Differences:

The back elimination got only these three predictors, AIC got five (with age and lbph in addition), adjusted R2 got seven (only dropping gleason from full model), Mallow’s Cp got five (with age lbph in addition).

Compare the fits of the full model and those selected by the methods above:

The adjusted R2 selected the model that maximize the adjusted R2. The back elimination method selected the predictors with significance at the same time. The other two methods selected a balanced model. The Cp method got a good fit with Cp around 5.

2. In the above problem, I would use the AIC or Cp method to select my final model. As I have discussed in the answer to question 1, they did not lose too much fit(decrease in adjusted R2 from the full model and they got a relatively small model.

R-code used:

library(faraway)

data(prostate)

attach(prostate)

g=lm(lpsa~.,data=prostate)

summary(g)

g=update(g,.~.-gleason)

summary(g)

g=update(g,.~.-lcp)

summary(g)

g=update(g,.~.-pgg45)

summary(g)

g=update(g,.~.-age)

summary(g)

g=update(g,.~.-lbph)

summary(g)

g=lm(lpsa~.,data=prostate)

step(g)

library(leaps)

b=regsubsets(lpsa~.,data=prostate)

rs=summary(b)

plot(2:9,rs$adjr2,xlab="No. of Parameters",ylab="Adjusted Rsq")

which.max(rs$adjr2)

which.min(rs$cp)

plot(2:9,rs$cp,xlab="No. of Parameters",ylab="Cp")

abline(0,1)