Huiwen Chen ID: 02156341 uniquename: huiwenc

STAT 500 HW7

1.Use the prostate data with lpsa as the response and the other vari- ables as predictors. Implement the following variable selection meth- ods to determine the “best” model:

Comment on the models selected (similarities and/or differences). Compare the fits of the full model and those selected by the methods above.

(a) Backward Elimination

Code:

```
library(ElemStatLearn)
data("prostate")
attach(prostate)
## the full model
g = lm(lpsa ~ ., data = prostate)
summary(g)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.177306	1.338810	0.132	0.89495
lcavol	0.564417	0.088387	6.386	8.08e-09
lweight	0.622204	0.202179	3.077	0.00279
age	-0.021306	0.011383	-1.872	0.06460 .
lbph	0.096833	0.058441	1.657	0.10113
svi	0.761466	0.242697	3.138	0.00233
lcp	-0.105872	0.090661	-1.168	0.24609
gleason	0.049967	0.158955	0.314	0.75401
pgg45	0.004434	0.004485	0.989	0.32558
trainTRUE	0.004104	0.162772	0.025	0.97994

Residual standard error: 0.7035 on 87 degrees of freedom Multiple R-squared: 0.6634, Adjusted R-squared: 0.6286

F-statistic: 19.05 on 9 and 87 DF, p-value: < 2.2e-16

```
##method1: backward elimination
g1 = update(g,. ~ .-train)
summary(g)
##continue dropping
g1 = update(g1,. ~ .-gleason)
summary(g1)
g1 = update(g1,. ~ .-lcp)
summary(g1)
g1 = update(g1,. ~ .-pgg45)
summary(g1)
g1 = update(g1,. ~ .-age)
summary(g1)
g1 = update(g1,. ~ .-lbph)
summary(g1)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.77716	0.62300	-1.247	0.215367
lcavol	0.52585	0.07486	7.024	3.49e-10
lweight	0.66177	0.17564	3.768	0.000289
svi	0.66567	0.20709	3.214	0.001798

Residual standard error: 0.7076 on 93 degrees of freedom

Multiple R-squared: 0.6359 Adjusted R-squared: 0.6242

F-statistic: 54.15 on 3 and 93 DF, p-value: < 2.2e-16

other predictors can also made an optimal fit

summary(lm(lpsa ~ lcavol+ svi + lbph))

Coefficients:

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

(Intercept) 1.52816 0.11304 13.518 < 2e-16

lcavol 0.57359 0.07512 7.636 1.94e-11

svi 0.74352 0.21458 3.465 0.000804

lbph 0.14853 0.05160 2.878 0.004957

Residual standard error: 0.728 on 93 degrees of freedom Multiple R-squared: 0.6147, Adjusted R-squared: 0.6023 F-statistic: 49.46 on 3 and 93 DF, p-value: < 2.2e-16

1) After dropping points one by one, the selected model includes 3 predictors lcavol, lweight and svi. We notice that, the p-values of such three coefficients all largely decrease, which indicates they are related to the response.
2) But in another model the predictors lcavol, svi and llbph also show significant, so it's insufficient to conclude that the variables omitted are not related to the response. The model selected by this method may not be a optimal one for prediction or explanation.
3) The adjusted R-squared in the full model is reduced only by 0.004 in the full model. So the removal of the 6 predictors causes a slight reduction in fit.

(b) AIC

Code:

method 2:AIC

>step(g)

Start: AIC=-58.78

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

Df Sum of Sq RSS AIC

- train 1 0.0003 43.058 -60.779

- gleason 1 0.0489 43.107 -60.669

- pgg45 1 0.4837 43.542 -59.696

- lcp 1 0.6749 43.733 -59.271

<none> 43.058 -58.780

- lbph 1 1.3588 44.417 -57.766

- age 1 1.7339 44.792 -56.950

- lweight 1 4.6874 47.745 -50.756

- svi 1 4.8720 47.930 -50.382

- lcavol 1 20.1817 63.240 -23.494

Step: AIC=-60.78

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

Df Sum of Sq RSS AIC

- gleason 1 0.0491 43.108 -62.668

- pgg45 1 0.5102 43.569 -61.636

- lcp 1 0.6814 43.740 -61.256

<none> 43.058 -60.779

- lbph 1 1.3646 44.423 -59.753

- age 1 1.7981 44.857 -58.810

- lweight 1 4.6907 47.749 -52.749

- svi 1 4.8803 47.939 -52.364

- lcavol 1 20.1994 63.258 -25.467

Step: AIC=-62.67

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

	Df	Sum of Sq	RSS	AIC
- lcp	1	0.6684	43.776	-63.176
<none>			43.108	-62.668
- pgg45	1	1.1987	44.306	-62.008
- lbph	1	1.3844	44.492	-61.602
- age	1	1.7579	44.865	-60.791
- lweight	1	4.6429	47.751	-54.746
- svi	1	4.8333	47.941	-54.360
- lcavol	1	21.3191	64.427	-25.691

Step: AIC=-63.18

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

	Df	Sum of Sq	RSS	AIC
- pgg45	1	0.6607	44.437	-63.723
<none>			43.776	-63.176
- lbph	1	1.3329	45.109	-62.266
- age	1	1.4878	45.264	-61.934
- svi	1	4.1766	47.953	-56.336
- lweight	1	4.6553	48.431	-55.373
- lcavol	1	22.7555	66.531	-24.572

Step: AIC=-63.72

lpsa ~ lcavol + lweight + age + lbph + svi

	Df	Sum of Sq	RSS	AIC
<none>			44.437	-63.723
- age	1	1.1588	45.595	-63.226
- lbph	1	1.5087	45.945	-62.484
- lweight	1	4.3140	48.751	-56.735
- svi	1	5.8509	50.288	-53.724
- lcavol	1	25.9427	70.379	-21.119

Call:

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

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.49473	0.87652	0.564	0.57385	
lcavol	0.54400	0.07463	7.289	1.11e-10	
lweight	0.58821	0.19790	2.972	0.00378	
age	-0.01644	0.01068	-1.540	0.12692	
lbph	0.10122	0.05759	1.758	0.08215	.
svi	0.71490	0.20653	3.461	0.00082	

Residual standard error: 0.6988 on 91 degrees of freedom Multiple R-squared: 0.6526, Adjusted R-squared: 0.6335 F-statistic: 34.19 on 5 and 91 DF, p-value: < 2.2e-16

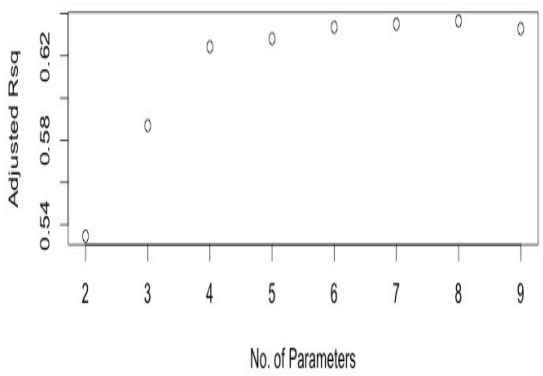
1) The AIC selected 5 predictors. It uses a search method to compare models sequentially and drop one predictor at a time, the process is more like that in the backward selection. And the sequence of variable removal was the same as in backward selection.
2) The differences are that AIC doesn't contain hypothesis testing, and it retained two more variables which are not significant in the final model. But the R-squared increased a little compared that in the full model. So AIC is desirable for prediction, and backward selection tends to pick models that are smaller, it cares more about whether the relationship between each predictor and the response are significant.

(c) Adjusted R2

Code:

```
> ## method 3:Adjusted R2

> library(leaps)

> g3 = regsubsets(lpsa ~., data = prostate)

> summary(g3)

lcavol lweight age lbph svi lcp gleason pgg45
trainTRUE

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

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

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

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

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

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

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

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

> ##plot adjusted R2 against p+1

> rs =  summary(g3)

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

> ##select model with largest adjusted R2

> which.max(rs$adjr2)

[1] 7
```

```
Coefficients:

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

(Intercept) 0.494155 0.873567 0.566  0.57304

lcavol       0.569546   0.085847   6.634 2.46e-09

lweight      0.614420   0.198449   3.096  0.00262

age         -0.020913   0.010978  -1.905  0.06000 .

lbph         0.097353   0.057584   1.691  0.09441 .

svi          0.752397   0.238180   3.159  0.00216

lcp         -0.104959   0.089347  -1.175  0.24323

pgg45        0.005324   0.003385   1.573  0.11923

Residual standard error: 0.696 on 89 degrees of
freedom  Multiple R-squared:  0.663,  Adjusted
R-squared:  0.6365  F-statistic: 25.01 on 7 and 89
DF,  p-value: < 2.2e-16
```



1) In this method, we see the selected model contains 7 variables lcavol, lweight, age, lbph, svi, lcp and pgg45, with which to achieve the largest adjusted R-squared. Surely the adjusted R-squared are larger than the former two models and the full model, but 4 of the predictors are not significant.

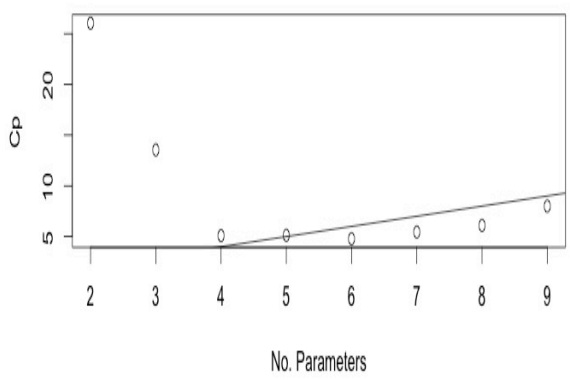
2) AIC and adjusted R-squared method both try to make a better fit with smaller RSS (the fit) when increasing complexity (p).

(d) Mallows' Cp

```
##method 4: Mallow Cp

which.min(rs$cp)

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

abline(0,1)
```

```
Coefficients:

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

(Intercept)	-0.3409	0.6936	-0.492	0.62420
lcavol	0.5285	0.0745	7.094	2.63e-10
lweight	0.5360	0.1964	2.729	0.00761
lbph	0.0786	0.0561	1.401	0.16454
svi	0.7055	0.2080	3.392	0.00102



Residual standard error: 0.704 on 92 degrees of freedom Multiple R-squared: 0.6436, Adjusted R-squared: 0.6281 F-statistic: 41.53 on 4 and 92 DF, p-value: < 2.2e-16

1) In this method, the selected model is a 5–parameter one, including lcavol, lweight, lbph and svi. The adjusted R-squared is 0.039 smaller than that in the full model, but is a little larger than that in backward elimination.

2) The model only contains one insignificant predictor than that in backward elimination. In order to minimize Cp, we choose to accept a larger model but with better fit. So Mallows' Cp, like AIC and adjusted R-squared all tradeoff fit against complexity, while backward elimination tends to decrease complexity.

2. In the above problem, which of the 4 criteria would you use to select your final model? Provide the reasoning behind your answer.

I prefer to choose the Mallow's Cp method.
1) Based on the comparison above, Cp, adjusted R-squared and AIC all try to achieve a better fit while increasing the complexity in different extent. Though backward elimination selects a smaller model, it may miss the optimal model, since dropped predictors sometimes also have a relationship with the response. So criterion- based methods are more desirable for prediction purposes.

2) The models in Cp, adjusted R-squared and AIC methods, have a similar conclusion, and their adjusted R-squared all don't have large differences with the full model. So we choose a simpler model with Cp method, which is easier to measure.