Anastasia Shpurik

HW CH 5, 7, 8

Chapter 5

**5) For the prostate data, fit a model with lpsa as the response and the other variables as predictors.**

library(faraway)

data(prostate)

g ← lm(lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45, prostate)

a) Compute and comment on the condition numbers.

X ← model.matrix(g)[,1]

(xtx ← t(x) %\*% x)

e ← eigen(xtx)

sqrt(e$val[1]/e$val)

Output:

[1] 1.00000 2.78186 47.66094 52.22787 85.98499 103.73114 153.85414

[8] 243.30248

We see that kappa = 243.302, which is much greater than 30, so there must be a problem. We also see that most of the condition numbers are large, so this indicates that there is more than one independent linear combination. All of this means that collinearity exists, which means we have imprecise estimates of beta, so the effect of the predictors may not be accurate in our model.

b) Compute and comment on the correlations between the predictors.

round(cor(prostate),3)

Output:

lcavol lweight age lbph svi lcp gleason pgg45 lpsa

lcavol 1.000 0.194 0.225 0.027 0.539 0.675 0.432 0.434 0.734

lweight 0.194 1.000 0.308 0.435 0.109 0.100 -0.001 0.051 0.354

age 0.225 0.308 1.000 0.350 0.118 0.128 0.269 0.276 0.170

lbph 0.027 0.435 0.350 1.000 -0.086 -0.007 0.078 0.078 0.180

svi 0.539 0.109 0.118 -0.086 1.000 0.673 0.320 0.458 0.566

lcp 0.675 0.100 0.128 -0.007 0.673 1.000 0.515 0.632 0.549

gleason 0.432 -0.001 0.269 0.078 0.320 0.515 1.000 0.752 0.369

pgg45 0.434 0.051 0.276 0.078 0.458 0.632 0.752 1.000 0.422

lpsa 0.734 0.354 0.170 0.180 0.566 0.549 0.369 0.422 1.000

lcavol & svi = 0.539, lcavol & lcp = 0.675, **lcavol & lpsa = 0.734**

svi & lcavol = 0.539, svi & lcp = 0.673, **svi & lpsa = 0.566**

lcp & gleason = 0.515, lcp & pgg45 = 0.632, **lcp & lpsa = 0.549**

gleason & pgg45 = 0.752

I have listed all of the pairwise correlations that appear to be fairly large. We see that there are some large correlations both between predictors and between predictors and our response, lpsa; these are lcavol, svi, and lcp (bolded).

c) Compute the variance inflation factors.

vif(x)

Output:

lcavol lweight age lbph svi lcp gleason pgg45

2.054115 1.363704 1.323599 1.375534 1.956881 3.097954 2.473411 2.974361

We can interpret these values by taking the square root of each one:

sqrt(vif(x))

Output:

lcavol lweight age lbph svi lcp gleason pgg45

1.433218 1.167777 1.150478 1.172832 1.398886 1.760100 1.572708 1.724634

This shows that the standard error for each predictor is m times as large as it would have been without collinearity, where m is the squared value. So the standard error for lcp is 1.76 times as large as it would have been without collinearity. This gives us a nice sense of the size of the effect, and we see that none of these are very large.

Chapter 7

**3) Using the ozone data, fit a model with O3 as the response and temp, humidity, and ibh as predictors. Use the Box-Cox method to determine the best transformation on the response.**

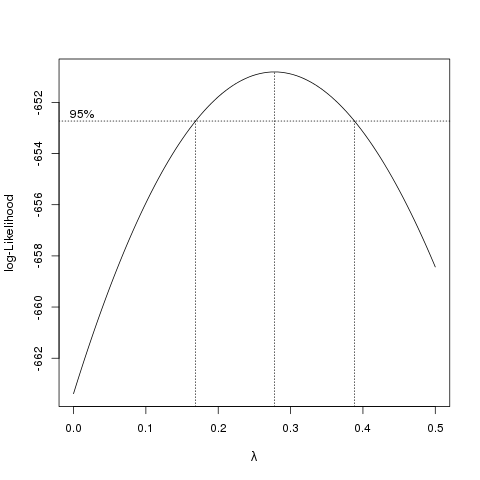
library(MASS)

data(ozone)

g ← lm(O3 ~ temp + humidity + ibh, ozone)

boxcox(g, plotit=T, lambda=seq(0, 0.5, by=0.05))

Output:



The above figure shows the box-cox transformation of the response. We see that the confidence interval for lambda runs from about 0.17 to about 0.39. It looks like lambda ~ 0.25, so a fourth- root transformation may be best here.

**4) Using the pressure data, fit a model with pressure as the response and temperature as the predictor using transformations to obtain a good fit.**

data(pressure)

g ← lm(pressure ~ temperature, pressure)

summary(g)

Output:

Coefficients:

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

(Intercept) -147.8989 66.5529 -2.222 0.040124 \*

temperature 1.5124 0.3158 4.788 0.000171 \*\*\*

---

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

Residual standard error: 150.8 on 17 degrees of freedom

Multiple R-squared: 0.5742, Adjusted R-squared: 0.5492

F-statistic: 22.93 on 1 and 17 DF, p-value: 0.000171

For the linear model, temperature is significant but R^2 is fairly low.

We can try a quadratic model, using orthogonal polynomials.

g2 ← lm(pressure ~ poly(temperature, 2), pressure)

summary(g2)

Output:

Coefficients:

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

(Intercept) 124.34 17.07 7.283 1.83e-06 \*\*\*

poly(temperature, 2)1 722.17 74.42 9.704 4.16e-08 \*\*\*

poly(temperature, 2)2 545.95 74.42 7.336 1.67e-06 \*\*\*

---

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

Residual standard error: 74.42 on 16 degrees of freedom

Multiple R-squared: 0.9024, Adjusted R-squared: 0.8902

F-statistic: 74 on 2 and 16 DF, p-value: 8.209e-09

We see that both coefficients are significant and R^2 has increased a lot!

Let's try the cubic model.

g3 ← lm(pressure ~ poly(temperature,3), pressure)

summary(g3)

Output:

Coefficients:

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

(Intercept) 124.337 5.876 21.16 1.39e-12 \*\*\*

poly(temperature, 3)1 722.171 25.614 28.20 2.08e-14 \*\*\*

poly(temperature, 3)2 545.947 25.614 21.32 1.25e-12 \*\*\*

poly(temperature, 3)3 280.653 25.614 10.96 1.48e-08 \*\*\*

---

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

Residual standard error: 25.61 on 15 degrees of freedom

Multiple R-squared: 0.9892, Adjusted R-squared: 0.987

F-statistic: 456.4 on 3 and 15 DF, p-value: 5.889e-15

Again, all coefficients are still significant and R^2 is even closer to 1.

Here is the output of the 4th and 5th degrees:

Call:

lm(formula = pressure ~ poly(temperature, 4), data = pressure)

Residuals:

Min 1Q Median 3Q Max

-7.1989 -4.2112 0.2224 4.0172 7.0729

Coefficients:

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

(Intercept) 124.337 1.234 100.74 < 2e-16 \*\*\*

poly(temperature, 4)1 722.171 5.380 134.24 < 2e-16 \*\*\*

poly(temperature, 4)2 545.947 5.380 101.48 < 2e-16 \*\*\*

poly(temperature, 4)3 280.653 5.380 52.17 < 2e-16 \*\*\*

poly(temperature, 4)4 97.137 5.380 18.06 4.28e-11 \*\*\*

---

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

Residual standard error: 5.38 on 14 degrees of freedom

Multiple R-squared: 0.9996, Adjusted R-squared: 0.9994

F-statistic: 7841 on 4 and 14 DF, p-value: < 2.2e-16

Call:

lm(formula = pressure ~ poly(temperature, 5), data = pressure)

Residuals:

Min 1Q Median 3Q Max

-0.60843 -0.25856 0.03803 0.27201 0.42416

Coefficients:

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

(Intercept) 124.33671 0.09057 1372.78 < 2e-16 \*\*\*

poly(temperature, 5)1 722.17059 0.39480 1829.21 < 2e-16 \*\*\*

poly(temperature, 5)2 545.94688 0.39480 1382.85 < 2e-16 \*\*\*

poly(temperature, 5)3 280.65281 0.39480 710.88 < 2e-16 \*\*\*

poly(temperature, 5)4 97.13691 0.39480 246.04 < 2e-16 \*\*\*

poly(temperature, 5)5 20.07923 0.39480 50.86 2.41e-16 \*\*\*

---

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

Residual standard error: 0.3948 on 13 degrees of freedom

Multiple R-squared: 1, Adjusted R-squared: 1

F-statistic: 1.165e+06 on 5 and 13 DF, p-value: < 2.2e-16

We see that the 5-th degree polynomial is the best fit for the data since all of the coefficients are significant and R^2 is 1. However, we may have an issue of using too many degrees of freedom and choosing a more complicated model versus a simpler one. The difference between the 4th degree and 5th degree R^2 is 0.0004. This is a fairly insignificant change. Thus, we can conclude that the 4th degree polynomial model is the best fit.

Chapter 8

**1) Using 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.**

data(prostate)

g ← lm(lpsa ~ ., prostate)

a) Backward Elimination

summary(g)

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

We need to remove the variable with the highest p-value grater than alpha\_crit (5%).

We will remove gleason.

g1 ← lm(lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45, prostate)

summary(g1)

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

---

Now we will remove lcp.

g2 ← lm(lpsa ~ lcavol + lweight + age + lbph + svi + pgg45, prostate)

summary(g2)

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

---

Now we will remove pgg45.

g3 ← lm(lpsa ~ lcavol + lweight + age + lbph + svi, prostate)

summary(g3)

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 \*\*\*

---

Now we will remove age

g4 ← lm(lpsa ~ lcavol + lweight + lbph + svi, prostate)

summary(g4)

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 \*\*

---

Now we will remove lbph

g5 ← lm(lpsa ~ lcavol + lweight + svi, prostate)

summary(g5)

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 \*\*

---

This leaves us with three predictors; lcavol, lweight, and svi. These are all significant at the 5% level. So g5 is the best model.

b) AIC

step(g)

Output:

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

Here we see that variables were removed in a similar manner as to backward elimination, excepted based on the AIC. This retained two more variables than in part (a); age and lbph. According to this criterion, the best model is g6 = lcavol + lweight + age + lbph + svi.

c) Adjusted R^2

library(leaps)

b ← regsubsets(lpsa ~ ., prostate)

(rs ← summary(b))

Output:

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 ) "\*" "\*" "\*" "\*" "\*" "\*" "\*" "\*"

Here we see that the best one predictor model uses lcavol, the best two predictors model uses lcavol and lweight, the best three predictors model uses lcavol and lweight and svi, and etc.

rs$which[which.max(rs$adjr2),]

Output:

(Intercept) lcavol lweight age lbph svi

TRUE TRUE TRUE TRUE TRUE TRUE

lcp gleason pgg45

TRUE FALSE TRUE

We want to maximize R^2adj, so we select the best model with 7 predictors:

This can also be verified with a plot.

plot(2:9, rs$adjr2, xlab=”Number of Parameters”, ylab=”Adjusted R-square”)

![A description...](data:None;base64,)

The maximum occurs at 8 parameters, so 7 predictors.

The best model, according to Adjusted R^2 criterion is:

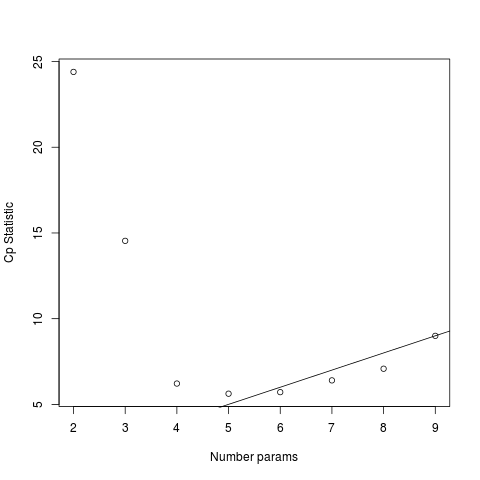
g7 = lcavol + lweight + age + lbph + svi + lcp + pgg45

d) Mallows C\_p

We want to minimize C\_p, so we can construct a plot

plot(2:9, rs$cp, xlab=”Number of Parameters”, ylab=”Cp Statistic”)

abline(0,1)



We are looking for models that are on or below the Cp=p line, which indicate good fits. Here, the smallest one is 6 parameters, which means 5 predictors.

Thus, the best model contains 5 parameters, according the Cp criterion:

g8 = lcavol + lweight + age + lbph + svi