Stat500(Section002): Homework #8

Due on Dec.01, 2021 at 3:00pm

Instructor:Naisyin Wang

Tiejin Chen **tiejin@umich.edu**

Problem 1

Part a

We using the following code to get the model and calculate the AIC from models.

```
library(faraway)
library(quantreg)
library(MASS)
library(nlme)
library(splines)
data(aatemp)
year<-aatemp$year[-(1:3)]; temp<-aatemp$temp[-(1:3)]
out4<-lm(temp ~ poly(year, 4))
out5<-lm(temp ~ poly(year, 5))
out6<-lm(temp ~ poly(year, 6))
choice<-seq(min(year), max(year), length =5);
knots <- c(rep(choice[1],4), choice[2:4], rep(choice[5],4))
bx <- splineDesign(knots, year, outer.ok = TRUE)
gs <- lm(temp ~ bx-1)
```

```
Aic4 = 112*log(sum(out4$residuals^2)/112) + 2*5
Aic5 = 112*log(sum(out5$residuals^2)/112) + 2*6
Aic6 = 112*log(sum(out6$residuals^2)/112) + 2*7
Aic_spilne = 112*log(sum(gs$residuals^2)/112) + 2*7
```

Here, we know the spline model does not have a constant, Hence we only use 7 to mupltply 2 instead (7+1). And we can get the result aic for all the models:

1. Orthogonal Polynomials to the 4-th degree AIC:82.98333
2. Orthogonal Polynomials to the 5-th degree AIC:78.39366
3. Orthogonal Polynomials to the 6-th degree AIC:79.61436
4. Cubic spline with three internally evenly spaced knots AIC:78.37617

Using the following code to compute BIC with AIC we get:

```
Bic4 = Aic4 + (log(112)-2)*5
Bic5 = Aic5 + (log(112)-2)*6
Bic6 = Aic6 + (log(112)-2)*7
Bic_spline = Aic_spilne + (log(112)-2)*7
```

We can get the result:

1. Orthogonal Polynomials to the 4-th degree BIC:96.57582
2. Orthogonal Polynomials to the 5-th degree BIC:94.70465
3. Orthogonal Polynomials to the 6-th degree BIC:98.64385
4. Cubic spline with three internally evenly spaced knots BIC:97.40566

For BIC, we can know the minimal value comes from model with 5 degree polynomials. Hence, I will recommend Orthogonal Polynomials to the 5-th degree model according to BIC. And for AIC, the mimimal value comes from model with cubic spline. Hence I will recommend Cubic spline with three internally evenly spaced knots model.

Problem 2

Part a

AIC:

```
library(leaps)
data(prostate)
step(lm(lpsa~., prostate), k=2)
```

We can get the result:

```
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
    0.95100      0.56561      0.42369     -0.01489
       lbph          svi
    0.11184      0.72095
```

From the result we can see that with backward step, the best model is using lcavol,lweight,age,lbph,svi as predictors with $AIC = -61.37$. We delete gleason,lcp and pgg45 3 predictors. And the AIC decrease from -58.32 to -61.37. And we can see that lcavol is always the most influentional predictor here, however, with the delete of other predictors, the influence of lcavol might decrease.

Part b

BIC:

```
step(lm(lpsa~., prostate), k=log(97))
```

We can get the result:

```
Start:  AIC=-35.15
lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
    pgg45

          Df Sum of Sq    RSS     AIC
- gleason  1    0.0412 44.204 -39.634
- pgg45    1    0.5258 44.689 -38.576
- lcp      1    0.6740 44.837 -38.255
- age      1    1.5503 45.713 -36.377
- lbph     1    1.6835 45.847 -36.095
<none>                 44.163 -35.149
- lweight  1    3.5861 47.749 -32.151
- svi      1    4.9355 49.099 -29.448
- lcavol   1   22.3721 66.535   0.030

Step:   AIC=-39.63
lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45

          Df Sum of Sq    RSS     AIC
- lcp      1    0.6623 44.867 -42.766
```

```
- pgg45     1     1.1920 45.396 -41.627
- age       1     1.5166 45.721 -40.936
- lbph      1     1.7053 45.910 -40.537
<none>                   44.204 -39.634
- lweight   1     3.5462 47.750 -36.723
- svi       1     4.8984 49.103 -34.014
- lcavol    1    23.5039 67.708  -2.849

Step:  AIC=-42.77
lpsa ~ lcavol + lweight + age + lbph + svi + pgg45

          Df Sum of Sq    RSS     AIC
- pgg45     1     0.6590 45.526 -45.926
- age       1     1.2649 46.131 -44.644
- lbph      1     1.6465 46.513 -43.844
<none>                   44.867 -42.766
- lweight   1     3.5647 48.431 -39.925
- svi       1     4.2503 49.117 -38.561
- lcavol    1    25.4189 70.285  -3.800

Step:  AIC=-45.93
lpsa ~ lcavol + lweight + age + lbph + svi

          Df Sum of Sq    RSS     AIC
- age       1     0.9592 46.485 -48.478
- lbph      1     1.8568 47.382 -46.623
<none>                   45.526 -45.926
- lweight   1     3.2251 48.751 -43.862
- svi       1     5.9517 51.477 -38.583
- lcavol    1    28.7665 74.292  -2.997

Step:  AIC=-48.48
lpsa ~ lcavol + lweight + lbph + svi

          Df Sum of Sq    RSS     AIC
- lbph      1     1.3001 47.785 -50.377
<none>                   46.485 -48.478
- lweight   1     2.8014 49.286 -47.377
- svi       1     5.8063 52.291 -41.636
- lcavol    1    27.8298 74.315  -7.542

Step:  AIC=-50.38
lpsa ~ lcavol + lweight + svi

          Df Sum of Sq    RSS     AIC
<none>                   47.785 -50.377
- svi       1     5.1814 52.966 -44.966
- lweight   1     5.8924 53.677 -43.673
- lcavol    1    28.0445 75.829 -10.160
```

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

Coefficients:
(Intercept)      lcavol     lweight         svi
    -0.2681      0.5516      0.5085      0.6662
```

We can see the result is that with backward step, the best model is using lcavol,lweight,svi as predictors with $BIC = -50.38$, And we delete gleason,lcp,pgg45,age,lbph 5 predictors which is more than situaion in AIC due to the more strict requirement for BIC.

Part c

Adjusted R2: In this part, we will using regsubsets function to get the best model

```
leaps = regsubsets(lpsa~., prostate,method = "backward")
summary(leaps)$which
summary(leaps)$adjr2
```

We get the result

```
  (Intercept) lcavol lweight   age  lbph   svi   lcp gleason pgg45
1        TRUE   TRUE   FALSE FALSE FALSE FALSE FALSE   FALSE FALSE
2        TRUE   TRUE    TRUE FALSE FALSE FALSE FALSE   FALSE FALSE
3        TRUE   TRUE    TRUE FALSE FALSE  TRUE FALSE   FALSE FALSE
4        TRUE   TRUE    TRUE FALSE  TRUE  TRUE FALSE   FALSE FALSE
5        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE FALSE   FALSE FALSE
6        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE FALSE   FALSE  TRUE
7        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE  TRUE   FALSE  TRUE
8        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE  TRUE    TRUE  TRUE
[1] 0.5345838 0.5771246 0.6143899 0.6208036 0.6245476 0.6258707 0.6272521 0.6233681
```

We can see that the best model is using lcavol,lweight,age,lbph,svi,lcp,pgg45 as predictors with $AdjR^2 = 0.62725$. We only delete gleason here.

And We can know that with all criteria, best model includes lcavol,lweight,svi as predictors. BIC is the most strict criterion for the number of predictors, AIC is the second, And $AdjR^2$ is the most lenient one with the most predictors stay.

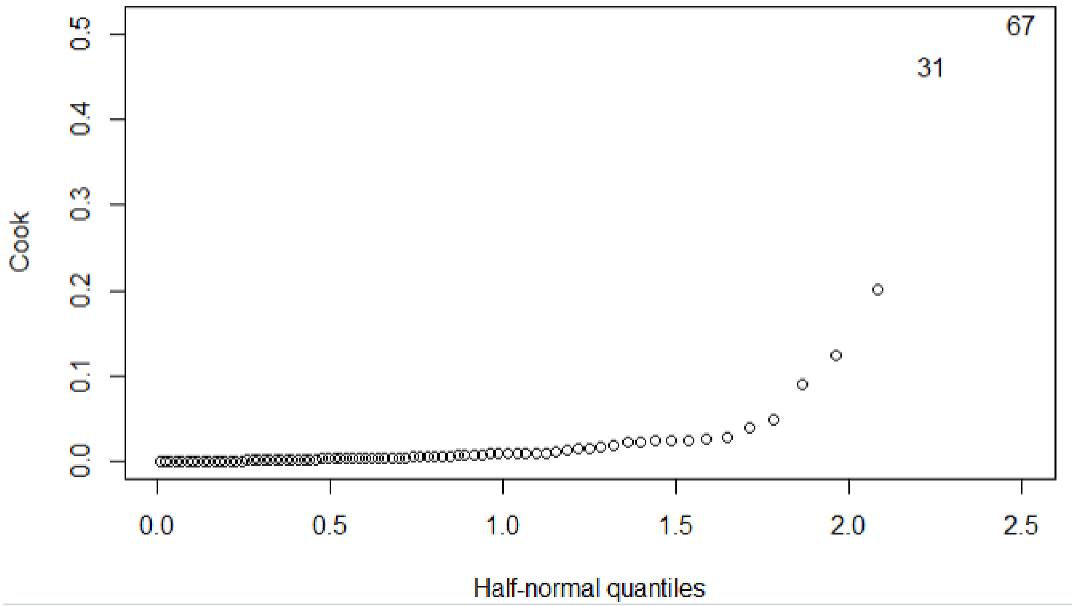
Problem 3

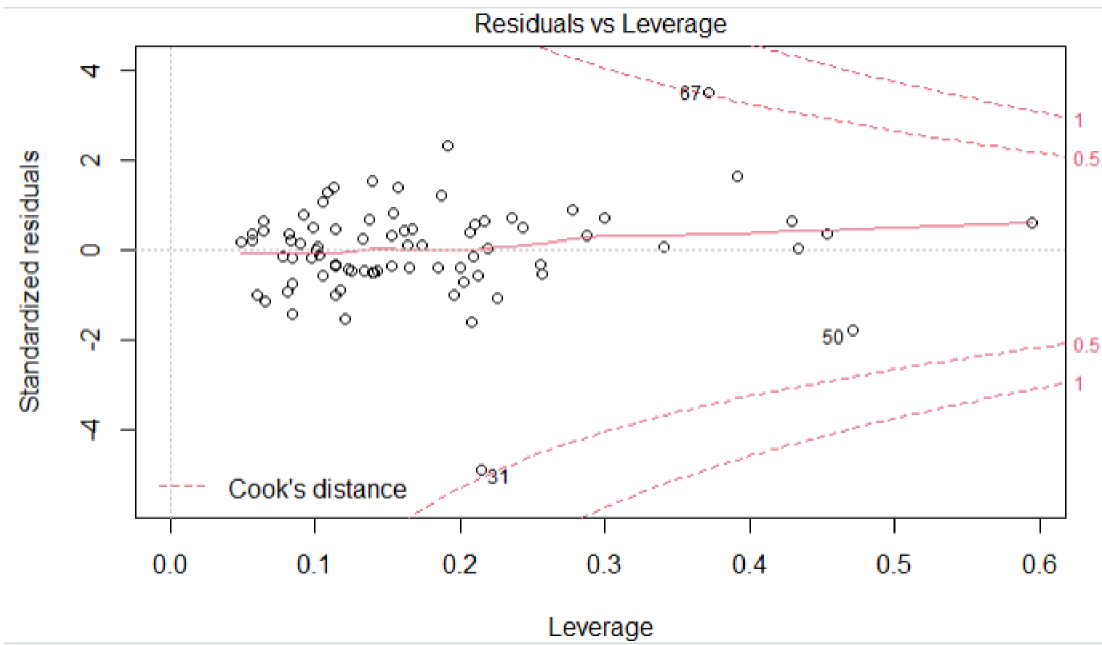
model1: $price\ bath * size + bed + halfbath + poly(age, 2) + garage + elem$ **Part a**

```
model1 = lm(price~bath*size+bed+halfbath+poly(age,2)+garage+elem, house.dat)
cook = cooks.distance(model1)
halfnorm(cook, ylab="Cook")
plot(model1)
```

We can the result:

Problem 3 continued on next page. . .





We can know that 67,31 is the number of two points which have the largest cook's distances. And from the plot, 67 is greater than 0.5, and should been seen as an influential point for sure. 31 is very close to 0.5 but does not reach it, Hence maybe we should not see it as an influential point. **Part b**
With No.67 31, We will use AIC as our criterion, we using:

```
step(model1)
```

To get result:

```
Start:  AIC=602.5
price ~ bath * size + bed + halfbath + poly(age, 2) + garage +
    elem
```

```
                Df Sum of Sq    RSS    AIC
- bath:size      2     953.6 106112 599.22
- halfbath       1     802.9 105961 601.10
- poly(age, 2)   2    3505.5 108664 601.12
<none>                       105158 602.50
- bed            1   11834.5 116993 609.03
- garage         1   20431.9 125590 614.70
- elem           3   28408.9 133567 615.63

Step:  AIC=599.22
price ~ bath + size + bed + halfbath + poly(age, 2) + garage +
    elem

                Df Sum of Sq    RSS    AIC
- bath           2    2002.4 108114 596.71
- halfbath       1     697.0 106809 597.74
- poly(age, 2)   2    3839.9 109952 598.06
<none>                       106112 599.22
- bed            1   12877.6 118989 606.38
- garage         1   19523.5 125635 610.73
- elem           3   29184.7 135296 612.66
- size           1   23169.9 129282 613.02

Step:  AIC=596.71
price ~ size + bed + halfbath + poly(age, 2) + garage + elem

                Df Sum of Sq    RSS    AIC
- halfbath       1     338.2 108452 594.96
- poly(age, 2)   2    5436.1 113550 596.64
<none>                       108114 596.71
- bed            1   16127.2 124241 605.84
- garage         1   19288.4 127402 607.85
- elem           3   30640.6 138755 610.67
- size           1   23892.8 132007 610.69

Step:  AIC=594.96
price ~ size + bed + poly(age, 2) + garage + elem

                Df Sum of Sq    RSS    AIC
- poly(age, 2)   2    5104.3 113557 594.64
<none>                       108452 594.96
- bed            1   15819.5 124272 603.86
- garage         1   19783.4 128236 606.37
- size           1   23667.8 132120 608.76
- elem           3   30624.4 139077 608.86

Step:  AIC=594.64
price ~ size + bed + garage + elem
```

Problem 3 continued on next page. . .

```
          Df Sum of Sq    RSS    AIC
<none>                 113557 594.64
- bed      1     14899 128456 602.51
- garage   1     16102 129658 603.25
- size     1     25001 138557 608.56
- elem     3     52229 165786 618.91

Call:
lm(formula = price ~ size + bed + garage + elem, data = house.dat)

Coefficients:
(Intercept)         size          bed       garage
     194.59        86.24       -21.35        21.81
      elemB        elemC        elemD
     -16.83       -64.48       -31.58
```

We can see that the best model is using size,bed,garage,elem as the predictors with AIC = 594.64.
Now we delete two points

```
model2 = lm(price~bath*size+bed+halfbath+poly(age,2)+garage+elem,
house.dat,subset = -c(31,67))
step(model2)
```

We get

```
Start:  AIC=533.95
price ~ bath * size + bed + halfbath + poly(age, 2) + garage +
    elem

                 Df Sum of Sq   RSS    AIC
- poly(age, 2)  2      1090 52279 531.60
- halfbath      1       221 51410 532.29
<none>                      51189 533.95
- bath:size     2      2823 54013 534.14
- bed           1      5969 57159 540.56
- garage        1     17366 68556 554.74
- elem          3     38318 89507 571.54

Step:  AIC=531.6
price ~ bath + size + bed + halfbath + garage + elem + bath:size

            Df Sum of Sq    RSS    AIC
- halfbath   1       105  52385 529.75
- bath:size  2      2591  54871 531.37
<none>                    52279 531.60
- bed        1      5800  58079 537.80
- garage     1     17185  69464 551.76
- elem       3     58783 111062 584.37

Step:  AIC=529.75
```

```
price ~ bath + size + bed + garage + elem + bath:size

            Df Sum of Sq    RSS    AIC
<none>                    52385 529.75
- bath:size  2      2875  55260 529.92
- bed        1      5718  58103 535.83
- garage     1     17546  69931 550.29
- elem       3     58693 111078 582.38

Call:
lm(formula = price ~ bath + size + bed + garage + elem + bath:size,
    data = house.dat, subset = -c(31, 67))

Coefficients:
(Intercept)        bath2        bath3         size
     428.99      -257.11      -276.74       -40.55
        bed       garage        elemB        elemC
     -14.26        23.78       -21.49       -71.62
      elemD   bath2:size   bath3:size
     -65.62       127.16       137.44
```

After removing the two points, we get the best model is $price\ bath + size + bed + garage + elem + bath : size$ with $AIC = 529.75$. We can know that removing 2 points make AIC become lower and save 2 more predictors.

model2:$price\ bed + halfbath + age + elem + bath * size$

Part a

```
model_res = lm(price~bed+halfbath+age+elem+bath*size,house.dat)
which(model_res$residuals == max(model_res$residuals))
which(model_res$residuals == min(model_res$residuals))
```

We get the No.67 has the largest residual and No.31 has the samllest residual.

```
predict(model_res,house.dat[31,],level = 0.95,interval = "prediction")
predict(model_res,house.dat[67,],level = 0.95,interval = "prediction")
house.dat[c(67, 31),]$price
```

Get the result

```
        fit      lwr      upr
31 315.2421 221.0477 409.4365
        fit      lwr      upr
67 340.1118 239.2524 440.9713
[1] 441.8 125.9
```

We find that the price's real value of two points both not fall into the 95% confidence interval of the prediction. Hence, prices for these two houses do not follow the general regression model that applies to this data set.

Part b

We change age to poly(age,2) here, And do the F-test compare new model and previous model to get:

```
model_res2 = lm(price~bed+halfbath+poly(age,2)+elem+bath*size,house.dat)
```

```
anova(model_res2, model_res)
```

We get:

```
Model 1: price ~ bed + halfbath + poly(age, 2) + elem + bath * size
Model 2: price ~ bed + halfbath + age + elem + bath * size
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     67 125590
2     68 127028 -1   -1438.3 0.7673 0.3842
```

Since the p-value is 0.3842 which is greater than 0.1, We reject to consider the quadratic function of "age".

Part c

In thie part, we change bath2 into reference, and for the difference to bath2 and bath1 in size, we have the the hypothesis test: $H_0 : \beta_{bath1:size} = 0, H_A : \beta bath1 : size \neq 0$. And we use the following code to get the 95% CI for this parameter. And we will drop No.67 and No.31 here and in next question.

```
new_data = house.dat
new_data$bath = relevel(house.dat$bath, ref="2")
model_new = lm(price~bed + halfbath + age + elem + bath * size,
new_data, subset=-c(31,67))
confint(model_new)
```

We can the result:

```
                  2.5 %        97.5 %
(Intercept)  100.687315  331.71741389
bed          -36.981244  -13.58368931
halfbath     -23.545456   11.69156353
age           -2.680237    5.90137415
elemB        -50.917995   11.96063344
elemC        -92.950503  -42.43537344
elemD       -105.272804  -42.79304665
bath1         30.148424  601.01331500
bath3       -229.738972  122.57906926
size          45.480387  158.80671006
bath1:size  -320.807723   -0.01217422
bath3:size   -59.080744  114.32342012
```

We can see that 0 is not in the 95% confidence interval of $\beta_{bath1:size}$, And we will reject H_0, And acceptH_1, which means we think the regression slopes for "size" are different between houses with 1 bathroom and houses with 2-bathrooms.

And for bath2 and bath3, we have the similar hypothesis test: $H_0 : \beta_{bath3:size} = 0, H_A : \beta bath3 : size \neq 0$. And we can see 0 is in the 95% confidence interval of $\beta bath3 : size$ indeed. Therefore we will not reject H_0, which means we will think the regression slopes for "size" are the same for houses with 2 bathroomsand houses with 3 bathrooms.

Part d

```
predict(update(model_res, subset=-c(67,31)),
data.frame(size=2, age=0 ,elem= "A", bed=3, bath="2",halfbath=1),
level=0.95,interval="prediction")
```

Using the code, we can get the result:

```
      fit      lwr      upr
1 338.7151 268.5634 408.8668
```

Since the unit here is 1k dollars, we see that the average price of the house we are interested should be 338715.1 dollars. The 95% confidence interval (268563.408866.8) dollars is the reasonable price range for such a house.