FMPH221 Final Exam

Kevin Nguyen

2022-11-29

```
library(alr4)
```

```
## Loading required package: car
## Loading required package: carData
## Loading required package: effects
## Warning: package 'effects' was built under R version 4.1.1
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
waterOP = read.table("/Users/kevinnguyen/Downloads/waterOP.txt")
```

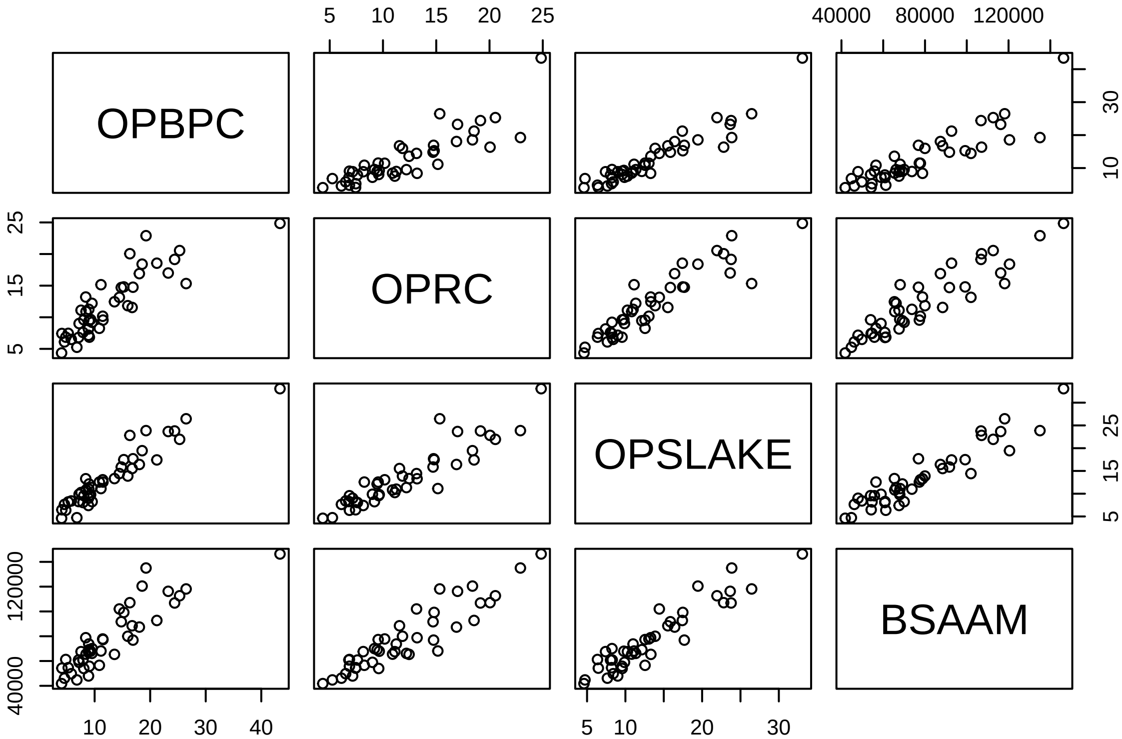
Problem 1

A)

```
model1 = lm(BSAAM ~ OPBPC + OPRC + OPSLAKE, data = waterOP)
summary(model1)
```

```
##
## Call:
## lm(formula = BSAAM ~ OPBPC + OPRC + OPSLAKE, data = waterOP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15964.1  -6491.8   -404.4   4741.9  19921.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22991.85    3545.32   6.485  1.1e-07 ***
## OPBPC          40.61     502.40   0.081  0.93599
## OPRC         1867.46     647.04   2.886  0.00633 **
## OPSLAKE      2353.96     771.71   3.050  0.00410 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8304 on 39 degrees of freedom
## Multiple R-squared:  0.9017, Adjusted R-squared:  0.8941
## F-statistic: 119.2 on 3 and 39 DF,  p-value: < 2.2e-16
```

```
pairs(waterOP[-1])
```



In the pairs plot, we are able to see that all of the variables OPBPC, OPRC, OPSLAKE, and BSAAM have positive linear correlations with each other. This gives basis for collinearity.

B)

```
confint(model1)
```

```
##                   2.5 %    97.5 %
## (Intercept) 15820.7771 30162.930
## OPBPC        -975.5885  1056.807
## OPRC          558.6962  3176.216
## OPSLAKE       793.0327  3914.880
```

The confidence interval for OPBPC is [-975.5885, 1056.807]. It does not make sense for the lower bound coefficient to be negative because the variable represents precipitation measurements in inches and the lowest it could possibly go is 0 which would represent no precipitation in the area.

C)

```
m1_predict = predict(model1, newdata=data.frame(OPBPC = 0, OPRC = 0, OPSLAKE = 0),
                     interval = 'prediction', level = 0.95)
m1_predict = as.data.frame(m1_predict)
paste0("lower prediction coefficient: ",
       round(m1_predict$lwr, 2),
       " | upper prediction coefficient: ",
       round(m1_predict$upr, 2))
```

```
## [1] "lower prediction coefficient: 4728.06 | upper prediction coefficient: 41255.65"
```

The 95% prediction interval for the run-off when there is no precipitation at neither of the three sites is [4728.06, 41255.65]. This indicates that in a year with no precipitation at the three sites, there would be stream runoff of about x acre-feet with x being some value within the prediction band.

D)

```r
round(vif(model1),3)
```

```
##   OPBPC    OPRC OPSLAKE
##   9.086   6.447  14.772
```

Each of these numbers indicates the rate at which these predictors increase betahat or the estimated cofeeicients of each predictor because of their correlation. In this case, because of OPBPC's correlation with the other predictors, its betahat or estimate coefficient is increased by a factor of 9.086 Out of the three variables, OPBPC is the most predictable from the others because it has the middle VIF value which indicates that it can more closely be predicted using the two other variables.

E)

```r
m2 = lm(OPBPC ~ OPRC + OPSLAKE, data = waterOP)
round(summary(m2)$coefficients, 3)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.479      1.045  -2.373    0.023
## OPRC          -0.023      0.204  -0.114    0.909
## OPSLAKE        1.153      0.160   7.190    0.000
```

```r
m3 = lm(OPBPC ~ OPSLAKE + OPRC, data = waterOP)
round(summary(m3)$coefficients, 3)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.479      1.045  -2.373    0.023
## OPSLAKE        1.153      0.160   7.190    0.000
## OPRC          -0.023      0.204  -0.114    0.909
```

```r
anova(m2, m3)
```

```
## Analysis of Variance Table
##
## Model 1: OPBPC ~ OPRC + OPSLAKE
## Model 2: OPBPC ~ OPSLAKE + OPRC
##   Res.Df    RSS Df  Sum of Sq F Pr(>F)
## 1     40 273.22
## 2     40 273.22  0 1.1369e-13
```

In this case, we see that there are no differences between the two models when taking a look at the ANOVA tests and individual summaries of each model.

F)

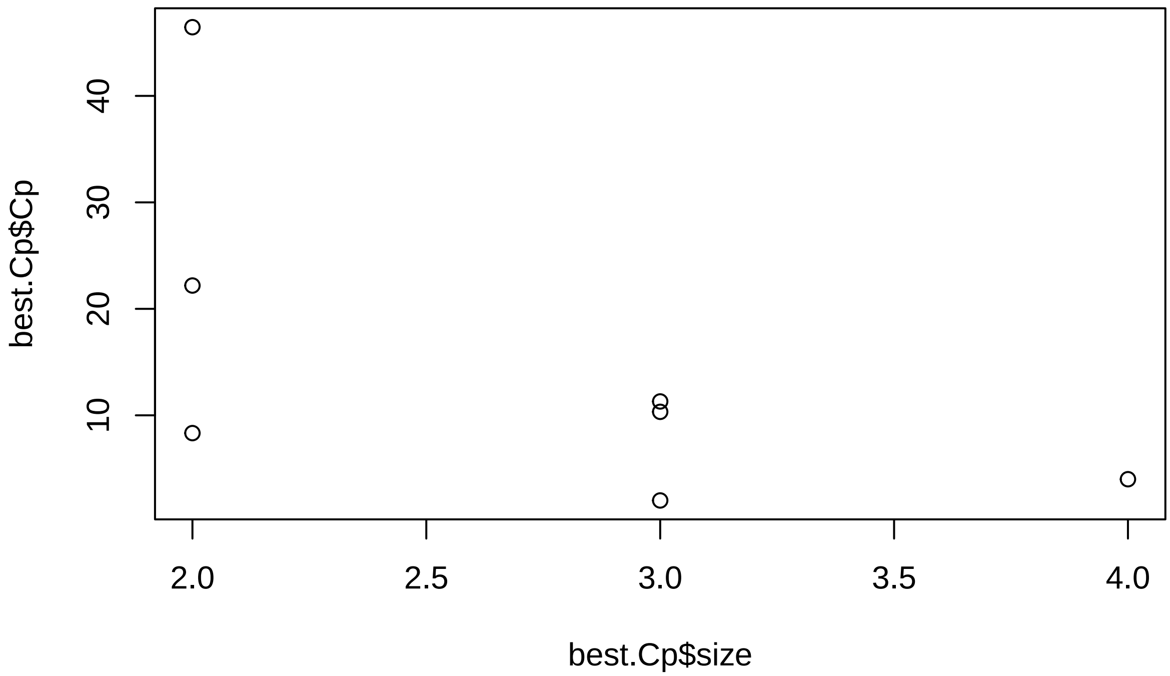
```r
library(leaps)
smaller.dat = waterOP[c("BSAAM", "OPBPC", "OPRC", "OPSLAKE")]
best.Cp <- leaps(as.matrix(smaller.dat[,2:4]), smaller.dat[,1], method="Cp")
(best.Cp.ind = which.min(best.Cp$Cp))
```

```
## [1] 4
```

```r
best.Cp$which[best.Cp.ind,]
```

```
##     1     2     3
## FALSE  TRUE  TRUE
```

```r
plot(best.Cp$size, best.Cp$Cp)
```



The Mallows statistic is telling us to use OPRC and OPSLAKE to build the model but to not use OPBPC. This is compatible with parts d and e because they were stating that the OPBPC variable was heavily influenced by the other variables.

Problem 2

A)

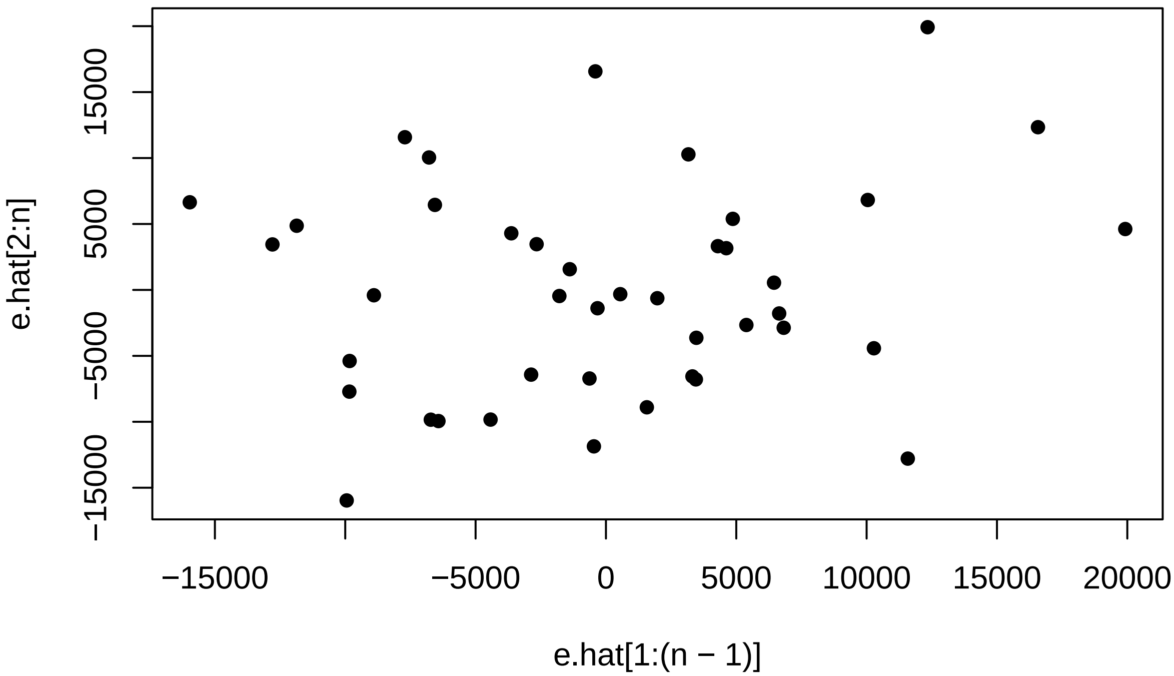
```
waterOP$AVG = (waterOP$OPBPC + waterOP$OPRC + waterOP$OPSLAKE)/3
waterOP$DELTA1 = waterOP$OPRC - waterOP$AVG
waterOP$DELTA2 = waterOP$OPSLAKE - waterOP$AVG
AVG_value = (0 + 0 + 0) / 3
new_model = predict(lm(BSAAM ~ AVG, data = waterOP),
                    newdata=data.frame(AVG = AVG_value),
                    interval = 'prediction', level = 0.95)
new_model = as.data.frame(new_model)
paste0("lower prediction coefficient: ",
       round(new_model$lwr, 2),
       " | upper prediction coefficient: ",
       round(new_model$upr, 2))
```

```
## [1] "lower prediction coefficient: 9233.78 | upper prediction coefficient: 46985.1"
```

The 95% prediction interval for the run-off when there is no precipitation at neither of the three sites using AVG is [9233.78, 46985.1]. In this case, the lower and upper prediction coefficient in the model using AVG is higher than than in the model from 1c.

B)

```
e.hat = residuals(model1)
n = nrow(waterOP)
plot(e.hat[2:n] ~ e.hat[1:(n-1)], pch=16)
```




```
cor(e.hat[2:n], e.hat[1:(n-1)])
```

```
## [1] 0.1655335
```

```
cor(e.hat[3:n], e.hat[1:(n-2)])
```

```
## [1] 0.1480609
```

```
cor(e.hat[3:n], e.hat[2:(n-1)])
```

```
## [1] 0.1660651
```

Here, we are able to see that the predictors are correlated but not to the point of contributiong equally to run-off at BSAAM. In this case, we would reject the null hypothesis and say that the predictors do not contribute equally.

C)

```
model2 = lm(BSAAM ~ AVG + DELTA1 + DELTA2, data = waterOP)
round(summary(model2)$coefficients, 2)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22991.85    3545.32    6.49     0.00
## AVG          4262.02     276.44   15.42     0.00
## DELTA1       1826.85     811.97    2.25     0.03
## DELTA2       2313.35    1195.88    1.93     0.06
```

The coefficient of DELTA1 is 1826.85 which is calculated by OPRC - (OPBPC + OPRC + OPSLAKE)/3 saying that if we remove the average of all three regions of precipitation from the OPRC variable, it is the effect that the OPRC has on the model without worrying about correlations and collinearities with the other variables. In this case, for every unit increase in DELTA1, there would be an increase in 1826.85 acre-feet stream runoff.

D)

```
compareCoefs(model2, model1)
```

```
## Calls:
```

```
## 1: lm(formula = BSAAM ~ AVG + DELTA1 + DELTA2, data = waterOP)
## 2: lm(formula = BSAAM ~ OPBPC + OPRC + OPSLAKE, data = waterOP)
##
##              Model 1 Model 2
## (Intercept)   22992   22992
## SE             3545    3545
##
## AVG            4262
## SE              276
##
## DELTA1         1827
## SE              812
##
## DELTA2         2313
## SE             1196
##
## OPBPC                  40.6
## SE                    502.4
##
## OPRC                   1868
## SE                      647
##
## OPSLAKE                2354
## SE                      772
##
```

The two models have the same intercepts but different coefficients per variable since they are both using different variables. Although the variables of model2 are based off of model1, it is hard to compare them to each other since they are some calculation of model1.

E)

```r
ind = which(waterOP$Year == '1969' | waterOP$Year == '1983')
model2.1 <- lm(BSAAM ~ AVG + DELTA1 + DELTA2, data = waterOP[-ind,])
round(summary(model2.1)$coefficients, 3)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25311.571   3782.133   6.692    0.000
## AVG          4056.036    307.513  13.190    0.000
## DELTA1        891.855    996.404   0.895    0.377
## DELTA2       1424.310   1347.362   1.057    0.297
```

```r
compareCoefs(model2, model2.1)
```

```
## Calls:
## 1: lm(formula = BSAAM ~ AVG + DELTA1 + DELTA2, data = waterOP)
## 2: lm(formula = BSAAM ~ AVG + DELTA1 + DELTA2, data = waterOP[-ind, ])
##
##              Model 1 Model 2
## (Intercept)   22992   25312
## SE             3545    3782
##
## AVG            4262    4056
## SE              276     308
##
## DELTA1         1827     892
```

```
## SE              812      996
##
## DELTA2         2313     1424
## SE             1196     1347
##
```

The fitted coefficients changed drastically when we removed the two years from the data. Although the change in the standard errors are smaller than the change in coefficient values, it is still sizable. This shows that the influence of the two years are large.