PSTAT 126 Homework 4

Celeste Herrera

5/15/2020

1. This problem uses the water data set in the alr4 package. For this problem, consider the regression problem with response BSAAM, and three predictors as regressors given by OPBPC, OPRC, and OPSLAKE.

(a) Examine the scatterplot matrix drawn for these three regressors and the response. What should the correlation matrix look like (i.e., which correlations are large and positive, which are large and negative, and which are small)? Compute the correlation matrix to verify your results. (Hint: the R function cor() can be used to compute a correlation matrix.)

```
#install.package(alr4)
library(alr4)
```

```
## Loading required package: car
```

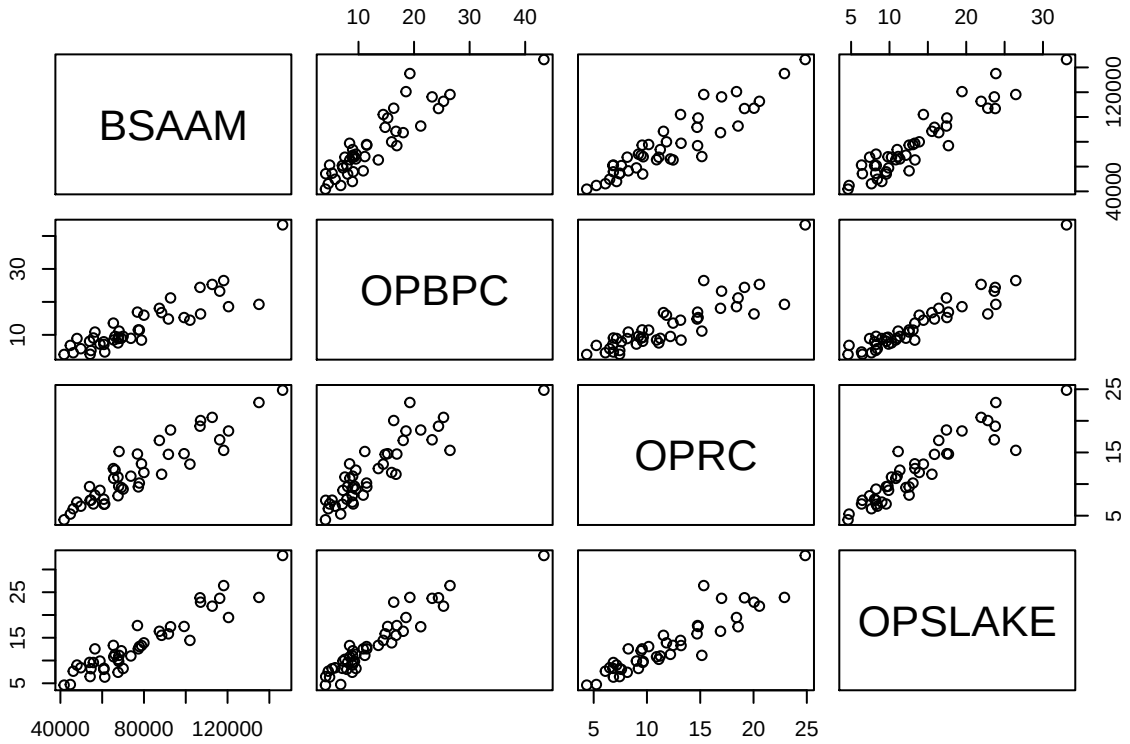
```
## Loading required package: carData
```

```
## Loading required package: effects
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
data("water")
OPBPC<- water$OPBPC
OPRC<- water$OPRC
OPSLAKE<- water$OPSLAKE
BSAAM<- water$BSAAM
pairs(BSAAM ~ OPBPC + OPRC + OPSLAKE, data=water)
```




```
cor_df = as.data.frame(cor(water))
cor_df
```

```
##                 Year         APMAM      APSAB    APSLAKE      OPBPC       OPRC
## Year     1.0000000000 -0.0007590557 0.05182523 0.17014669 0.11859943 0.02246824
## APMAM   -0.0007590557  1.0000000000 0.82768637 0.81607595 0.12238567 0.15441549
## APSAB    0.0518252272  0.8276863704 1.00000000 0.90030474 0.03954211 0.10563959
## APSLAKE  0.1701466883  0.8160759519 0.90030474 1.00000000 0.09344773 0.10638359
## OPBPC    0.1185994341  0.1223856707 0.03954211 0.09344773 1.00000000 0.86470733
## OPRC     0.0224682441  0.1544154918 0.10563959 0.10638359 0.86470733 1.00000000
## OPSLAKE  0.1380333978  0.1075421167 0.02961175 0.10058669 0.94334741 0.91914467
## BSAAM    0.1699631973  0.2385695382 0.18329499 0.24934094 0.88574778 0.91962700
##             OPSLAKE     BSAAM
## Year     0.13803340 0.1699632
## APMAM    0.10754212 0.2385695
## APSAB    0.02961175 0.1832950
## APSLAKE  0.10058669 0.2493409
## OPBPC    0.94334741 0.8857478
## OPRC     0.91914467 0.9196270
## OPSLAKE  1.00000000 0.9384360
## BSAAM    0.93843604 1.0000000
```

(b) Get the regression summary for the regression of BSAAM on these three regressors. Include OPBPC, OPRC, and OPSLAKE sequentially. Explain what the "Pr(> |t|)" column of your output means.

```
model = lm(BSAAM ~ OPBPC + OPRC + OPSLAKE, data = water)
summary(model)
```

```
##
## Call:
## lm(formula = BSAAM ~ OPBPC + OPRC + OPSLAKE, data = water)
```

```
##
## 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
```

The Pr(> |t|) column of of the output means the p-value for the certain values within that column.

(c) Use R to produce an ANOVA table for this regression fit. What is SSR(OPSLAKE|OPBPC, OPRC)? What is SSE(OPBPC, OPRC)?

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: BSAAM
##           Df     Sum Sq    Mean Sq  F value    Pr(>F)
## OPBPC      1 2.1458e+10 2.1458e+10 311.1610 < 2.2e-16 ***
## OPRC       1 2.5616e+09 2.5616e+09  37.1458 3.825e-07 ***
## OPSLAKE    1 6.4165e+08 6.4165e+08   9.3045  0.004097 **
## Residuals 39 2.6895e+09 6.8962e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model.full = lm(formula =BSAAM ~ 1)
model.reduced = lm(formula =BSAAM ~  OPBPC + OPRC)
anova(model.reduced,model.full)
```

```
## Analysis of Variance Table
##
## Model 1: BSAAM ~ OPBPC + OPRC
## Model 2: BSAAM ~ 1
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1     40 3.3312e+09
## 2     42 2.7351e+10 -2 -2.402e+10 144.21 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#SSE(OPBPC, OPRC) is 3.3312e+09

# In order to get SSR(OPSLAKE|OPBPC, OPRC) we need to compute a few things.
#SSR(OPBPC, OPRC)
2.1458e+10 + 2.5616e+09
```

```
## [1] 24019600000
```

```
#SSR(OPSLAKE, OPBPC, OPRC)
2.1458e+10 + 2.5616e+09 + 6.4165e+08
```

```
## [1] 24661250000
```

```
#SSR(OPSLAKE|OPBPC, OPRC) is 641650000
24661250000 - 24019600000
```

```
## [1] 641650000
```

So we now see that our SSR(OPSLAKE|OPBPC, OPRC) = 641650000 and SSE(OPBPC, OPRC) = 3.3312e+09

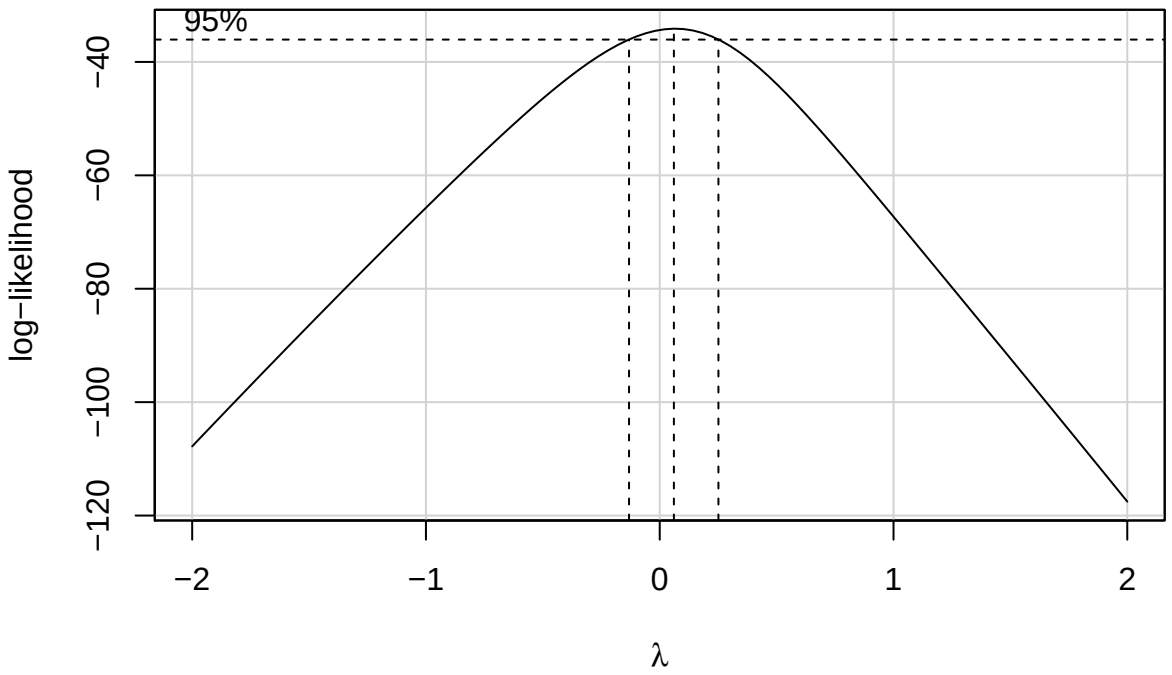
2. The lathe1 data set from the alr4 package contains the results of an experiment on characterizing the life of a drill bit in cutting steel on a lathe. Two factors were varied in the experiment, Speed and Feed rate. The response is Life, the total time until the drill bit fails, in minutes. The values of Speed and Feed in the data have been coded by computing

$$\text{Speed} = \frac{\text{Actual Speed -900}}{300}$$

$$\text{Feed} = \frac{\text{Actual feed rate in thousandths of an inch per revolution} - 13}{6}$$

(a) Starting with the full second-order model E(Life | Speed, Feed) = $\beta_0 + \beta_1$**Speed** $+ \beta_2$**Feed** $+ \beta_{11}$**Speed**$^2 + \beta_{22}$**Feed**$^2 + \beta_{12}$**Speed** $\times$ **Feed use the Box-Cox method to show that an appropriate scale for the response is the logarithmic scale.**

```
#library(alr4)
data("lathe1")
life<- lathe1$Life
speed <- lathe1$Speed
feed <- lathe1$Feed
speedsq <-speed^2
feedsq<- feed^2
fit<- lm(life~speed + feed +speedsq + feedsq +speed * feed)
box.cox = boxCox(fit)
```



(b) State the null and alternative hypotheses for the overall F-test for this model using log(Life) as the response. Perform the test and summarize results. $H_0 : \beta_1 = \beta_{11} = \beta_{12} = \beta_{22} = 0$ H_1: At least 1 $\beta_k \neq 0$

```
loglife <- log(life)
reduced<- lm(loglife~1)
full<- lm(loglife~ speed+feed+speedsq+feedsq+speed*feed)
anova(reduced,full)
```

```
## Analysis of Variance Table
##
## Model 1: loglife ~ 1
## Model 2: loglife ~ speed + feed + speedsq + feedsq + speed * feed
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     19 41.533
## 2     14  1.237  5    40.296 91.236 3.551e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results of loglife show to have a very small p-value which is 3.551e-10 with a relativley large F statistic of 91.236.

(c) Explain the practical meaning of the hypothesis $H_0 : \beta_1 = \beta_{11} = \beta_{12} = 0$ in the context of the above model.

H_0: The speed, the speed^2, and speed * feed do not have a relationship with life

(d) Perform a test for the hypothesis in part (c) and summarize your results.

```
newreduced<- lm(loglife~feed +feedsq)
anova(newreduced,full)
```

```
## Analysis of Variance Table
##
## Model 1: loglife ~ feed + feedsq
## Model 2: loglife ~ speed + feed + speedsq + feedsq + speed * feed
```

```
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     17 32.300
## 2     14  1.237  3    31.063 117.22 3.726e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value given which is 3.726e-10 is very small therefore we do not reject the null hypothesis

3. Consider the following model and the corresponding ANOVA table: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ where ϵ is the usual random error and Y_i 's are independent. Further assume $R^2 = 0.637$ for the above model.

(a) Fill in the missing values (denoted by star) in the ANOVA table.

```
df<- 2
df_total = 117+2
df_total
```

```
## [1] 119
```

```
n <- 120
sse<- 17.90761
sst<- sse/(1-.637)
sst
```

```
## [1] 49.33226
```

```
r2<- 1-(sse/sst)#formula with value given of 0.637(given)
r2
```

```
## [1] 0.637
```

```
ssr<- sst - sse
ssr
```

```
## [1] 31.42465
```

```
ms_regression<- ssr/df
ms_regression
```

```
## [1] 15.71232
```

```
ms_error <- 0.15306  # given
f_stat<- ms_regression/ms_error
f_stat
```

```
## [1] 102.6547
```

```
pf(f_stat,2,117, lower.tail = FALSE)
```

```
## [1] 1.79849e-26
```

(b) State the null and alternative hypothesis for the "F-test" in the ANOVA table.

$H_0 : \beta_1 = \beta_2 = 0$ $H_1 :$ Any $\beta_k \neq 0, k = 1, 2$

(c) What is the estimated value of σ^2 based on then results shown in the table?

Based the results given the estimated value of σ^2 is 0.15306, since the MSE is also known to be a unibased estimator of the variance.

4. A psychologist made a small scale study to examine the nature of the relation between an employee's emotional stability (Y) and the employee's ability to perform in a task group (X). Emotional stability was measured by a written test and ability to perform in a task group (X = 1 if able, X = 0 if unable) was evaluated by the supervisor. The results were as follows:

$$\begin{pmatrix} i:1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ Y_i:474 & 619 & 584 & 638 & 399 & 481 & 624 & 582 \\ X_i:0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}$$

(a) Fit a linear regression and write down the fitted model.

```
yi<-c(474,619,584,638,399,481,624,582)
xi<-c(0,1,0,1,0,1,1,1)

#fit the linear regression
yi_xi_fit<- lm(yi~xi)
```

The fitted model for the dataset is $Y_i = \beta_0 + \beta_{1x_i}$. The Y_i is the emotional stability of an employee i was able to perform in a taske group, and 0 if employee i was unable to perform in a task group.

(b) Write down separate estimated regression equations for "able" employees and "unable" employees.

```
summary(yi_xi_fit)
```

```
##
## Call:
## lm(formula = yi ~ xi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -107.80  -30.42   11.70   38.70   98.33
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   485.67      43.18  11.248 2.95e-05 ***
## xi            103.13      54.61   1.888    0.108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74.78 on 6 degrees of freedom
## Multiple R-squared:  0.3728, Adjusted R-squared:  0.2682
## F-statistic: 3.566 on 1 and 6 DF,  p-value: 0.1079
```

For the summary of the dataset, we notice the estimate regression of the equation is $Y_i = 485.67 + 103.13x_i$, which is also known to be the same thing as Stability = 485.67 + 103.13(Performance)

The estimated regression for "unable" employees ($x_i = 0$) is $Y_i = 485.67$ or Stability = 485.67. The estimated regression equation for "able" employees ($x_i = 1$) is $Y_i = 588.8$ or Stability = 588.8

(c) Is there a linear relationship between X and Y ? Test at 5% level.

To see if there is a relationship on a linear relationship between X and Y, we will be testing the null hpothesis $H_0 : \beta - 1 = 0$ Vs. $H_A : \beta_1 \neq 0$. We can test this using the output from the summary function as shown in the previous question part (b). The p-value given is 0.1079, and the p-value is much larger than $\alpha = .05$ so we fail to reject the null hypothesis and conclude that there is not a linear relationship between X and Y.

5. A marketing research trainee in the national office of a chain of shoe stores used the following response function to study seasonal (winter, spring, summer, fall) effects on sales of a certain line of shoes: $\mathbf{E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3}$. The Xs are indicator variables defined as follows:

(a) State the response functions for the four types of seasons.

winter y = $\beta_0 + \beta_1 \times x_1$
spring y = $\beta_0 + \beta_2 \times x_2$
fall y = $\beta_0 + \beta_3 \times x_3$
summer y = 0

(b) Interpret each of the following quantities: (i) β_0 (ii) β_1 (iii) β_2 (iv) β_3

(i) β_0 : is the unchanged price

(ii) β_1 : Is the changed in price for the winter

(iii) β_2 : Is the changed price for the spring

(iv) β_3 : Is the changed price for the fall $$