PSTAT 126 HW 4

Tamjid Islam

5/13/2020

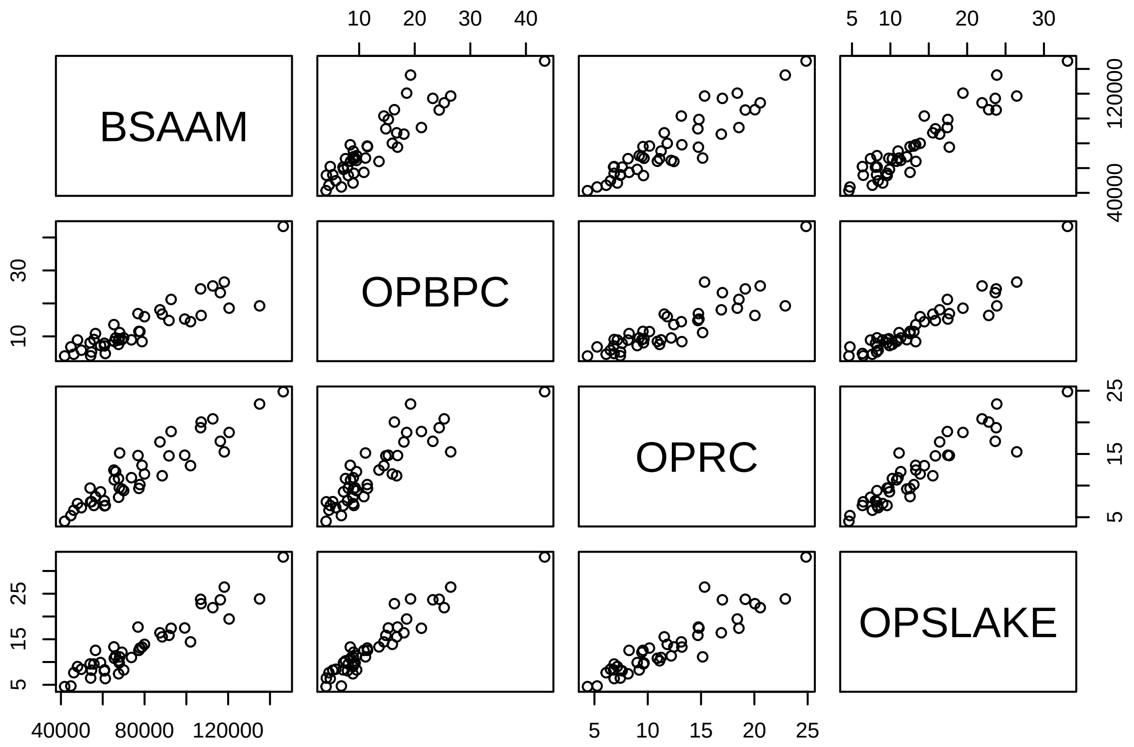
1. This problem uses the water data set in the alr4 package. For this problem, consider the regression problem with response BSAAM, and three predictors as regressors given by OPBPC, OPRC, and OPSLAKE.

```
# install.packages('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.
```

(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.)

```
library(alr4)
data(water)
pairs(BSAAM ~ OPBPC + OPRC + OPSLAKE, data = water)
```



All correlations are large and positive based on the scatterplot matrix.

```
data <- cbind(water$BSAAM, water$OPBPC, water$OPRC, water$OPSLAKE)
colnames(data) <- c('BSAAM', 'OPBPC', 'OPRC', 'OPSLAKE')
cor(data)
```

```
##             BSAAM     OPBPC      OPRC   OPSLAKE
## BSAAM   1.0000000 0.8857478 0.9196270 0.9384360
## OPBPC   0.8857478 1.0000000 0.8647073 0.9433474
## OPRC    0.9196270 0.8647073 1.0000000 0.9191447
## OPSLAKE 0.9384360 0.9433474 0.9191447 1.0000000
```

Based on the correlation matrix, the correlations seem to large and positive values. Thus, our reasoning based on the scatterplot matrix is correct.

(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.

```
fit <- lm(BSAAM ~ OPBPC + OPRC + OPSLAKE, data = water)
summary(fit)
```

```
##
## 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 the output means the p-values of each variable. Something that is noticed is the higher the t value the lower the p value. It tells us which predictors have and do not have an effect on the response. Thus, telling us and gives us evidences to either reject or fail to reject our null hypothesis.

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

```r
anova(fit)
```

```
## 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
```

```r
fit2 <- lm(BSAAM ~ OPBPC + OPRC, data = water)
anova(fit2)
```

```
## Analysis of Variance Table
##
## Response: BSAAM
##           Df     Sum Sq    Mean Sq F value    Pr(>F)    
## OPBPC      1 2.1458e+10 2.1458e+10  257.67 < 2.2e-16 ***
## OPRC       1 2.5616e+09 2.5616e+09   30.76 2.051e-06 ***
## Residuals 40 3.3312e+09 8.3279e+07                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SSR(OPSLAKE|OPBPC, OPRC) = 6.4165e+08 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

$$Speed = \frac{Actual\ speed\ in\ feet\ per\ minute\ -\ 900}{300}$$

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

```r
library(alr4)
data(lathe1)
```

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

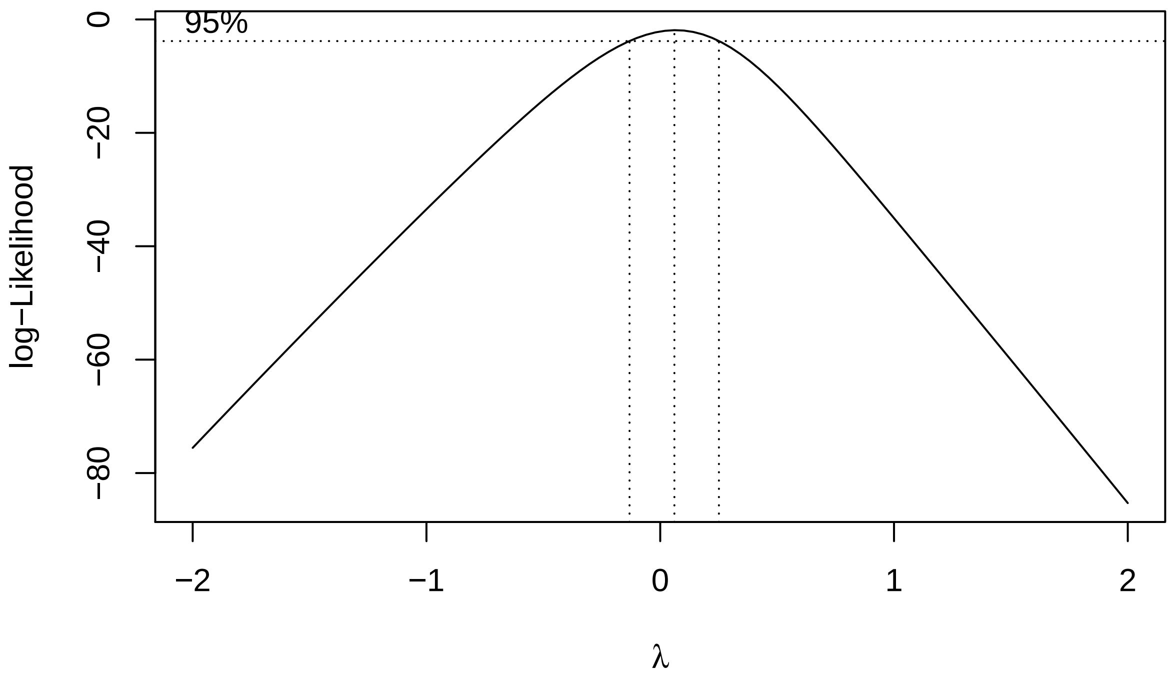
```r
library(alr4)
data(lathe1)
fit <- lm(Life ~  Speed + Feed + I(Speed^2) + I(Feed^2) + I(Speed * Feed), data = lathe1)
summary(fit)
```

```
##
## Call:
## lm(formula = Life ~ Speed + Feed + I(Speed^2) + I(Feed^2) + I(Speed *
##     Feed), data = lathe1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6601  -0.9607  -0.1383   0.7062  17.9193
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.338      2.733   1.222 0.241998
## Speed            -21.548      2.231  -9.657 1.44e-07 ***
## Feed             -10.494      2.231  -4.703 0.000339 ***
## I(Speed^2)        17.392      2.617   6.647 1.10e-05 ***
## I(Feed^2)          1.412      2.617   0.540 0.597837
## I(Speed * Feed)   10.975      2.733   4.016 0.001274 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.729 on 14 degrees of freedom
## Multiple R-squared:  0.9267, Adjusted R-squared:  0.9005
## F-statistic:  35.4 on 5 and 14 DF,  p-value: 1.831e-07
```

```r
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: Life
##                 Df Sum Sq Mean Sq F value    Pr(>F)
## Speed            1 5571.0  5571.0 93.2643 1.436e-07 ***
## Feed             1 1321.5  1321.5 22.1226  0.000339 ***
## I(Speed^2)       1 2700.4  2700.4 45.2075 9.725e-06 ***
## I(Feed^2)        1   17.4    17.4  0.2914  0.597837
## I(Speed * Feed)  1  963.6   963.6 16.1316  0.001274 **
## Residuals       14  836.3    59.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
library(MASS)
bc <- boxcox(fit)
```



```
lambda <- bc$x[which(bc$y == max(bc$y))]
lambda
```

```
## [1] 0.06060606
```

Since λ is close to zero, there is a log transformation.

(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.

$H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_{11} = \beta_2 2 = \beta_{12} = 0$
$H_1 : at\ least\ one\ \beta \neq 0$

```
fit <- lm(log(Life) ~  Speed + Feed + I(Speed^2) + I(Feed^2) + I(Speed * Feed), data = lathe1)
summary(fit)
```

```
##
## Call:
## lm(formula = log(Life) ~ Speed + Feed + I(Speed^2) + I(Feed^2) +
##     I(Speed * Feed), data = lathe1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43349 -0.14576 -0.02494  0.16748  0.47992
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.18809    0.10508  11.307 2.00e-08 ***
## Speed           -1.58902    0.08580 -18.520 3.04e-11 ***
## Feed            -0.79023    0.08580  -9.210 2.56e-07 ***
## I(Speed^2)       0.28808    0.10063   2.863 0.012529 *
## I(Feed^2)        0.41851    0.10063   4.159 0.000964 ***
## I(Speed * Feed) -0.07286    0.10508  -0.693 0.499426
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.2972 on 14 degrees of freedom
## Multiple R-squared:  0.9702, Adjusted R-squared:  0.9596
## F-statistic: 91.24 on 5 and 14 DF,  p-value: 3.551e-10
```

```r
fit1 <- lm(log(Life) ~ 1, data = lathe1)
summary(fit1)
```

```
##
## Call:
## lm(formula = log(Life) ~ 1, data = lathe1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5283 -0.5561 -0.3181  0.8988  2.8481
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.6120     0.3306   4.876 0.000105 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.478 on 19 degrees of freedom
```

```r
anova(fit1, fit)
```

```
## Analysis of Variance Table
##
## Model 1: log(Life) ~ 1
## Model 2: log(Life) ~ Speed + Feed + I(Speed^2) + I(Feed^2) + I(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
```

Since p-value is 3.551e-10 $< \alpha$, there is sufficient evidence that at least one of the slope paramters is not equal to 0. (F = 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.

The practical meaning of the hypothesis is to test whether or not the slope paramaters are equal to 0. Thus this shows us if any of the predictor variables if they are significant which then tells us if they have any relationship to the response. This is testing to see if the Life of a drill bit is significantly related to Speed, Speed^2 and Speed * Feed.

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

```r
fit <- lm(log(Life) ~  Speed + Feed + I(Speed^2) + I(Feed^2) + Speed * Feed, data = lathe1)
fit1 <- lm(log(Life) ~ Speed + I(Speed^2) + I(Speed * Feed), data = lathe1)
anova(fit1, fit)
```

```
## Analysis of Variance Table
##
## Model 1: log(Life) ~ Speed + I(Speed^2) + I(Speed * Feed)
## Model 2: log(Life) ~ Speed + Feed + I(Speed^2) + I(Feed^2) + Speed * Feed
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     16 10.2574
## 2     14  1.2367  2    9.0207 51.061 3.703e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since p-value is 3.703e-07 $< \alpha$, there is sufficient evidence that that Speed, Speed^2 and Speed * Feed are significantly related to Life of a drill bit. (F = 51.061)

3. Consider the following model and the corresponding ANOVA table:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i,$$

where ϵ is the usual random error and $Y_i^{'}$s are independent.

The ANOVA Table

Analysis of Varience					
Source	DF	Sum of Square	Mean Square	F Stat	Prob > F
Model	*	*	*	*	*
Error	117	17.90761	0.15306		
C Total	*	*			

Further assume $R^2 = 0.637$ for the above model.

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

```
df <- 2
df_total <- 117 + 2
n <- 120
SSE <- 17.90761
MSE <- 0.15306
SST <- SSE / (1-0.637)
SST
```

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

```
SSM <- SST - SSE
SSM
```

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

```
MSM <- SSM / df
MSM
```

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

```
f_stat <- MSM / MSE
f_stat
```

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

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

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

Model df : 2
C Total df : 119
SSE for Model is : 31.42465
SSE for C Total : 49.33226
Mean Square Model : 15.71232

F-stat : 102.65467
Prob > F (p-value): 1.79849e-26

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

$H_0 : \beta_1 = \beta_2 = 0$
$H_1 : \textit{At least one } \beta_i \neq 0$

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

```
sigma2 <- SSE / (n-2)
sigma2
```

```
## [1] 0.1517594
```

The estimated value of σ^2 is 0.1517594 which is the same as the MSE.

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:

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

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

```
y <- c(474, 619, 584, 638, 399, 481, 624, 582)
x <- c(0, 1, 0, 1, 0, 1, 1, 1)
cbind(x,y)
```

```
##      x   y
## [1,] 0 474
## [2,] 1 619
## [3,] 0 584
## [4,] 1 638
## [5,] 0 399
## [6,] 1 481
## [7,] 1 624
## [8,] 1 582
```

```
fit <- lm(y ~ x)
fit
```

```
## 
## Call:
## lm(formula = y ~ x)
## 
## Coefficients:
## (Intercept)            x  
##       485.7        103.1
```

y = 485.7 + 103.1x

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

```
beta <- coef(fit)
yhat <- beta[1] + beta[2] * 0
yhat
```

```
## (Intercept)
##    485.6667
```

```
yhat1 <- beta[1] + beta[2] * 1
yhat1
```

```
## (Intercept)
##       588.8
```

y = 458.667 for unable
y = 588.8 for able

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

```
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## 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 ***
## x             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
```

Since the p value is 0.108 which is greater than ($\alpha = 0.05$), we fail to reject the H_0. Therefore, there is no 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: $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$. The $X^{'}$s are indicator variables defined as follows:

Season	X_1	X_2	X_3
Winter	1	0	0
Spring	0	1	0
Fall	0	0	1
Summer	0	0	0

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

Summer : y = b_0
Winter : y = $b_0 + b_1 x_1$ so $x_1 = 1$, thus y = $b_0 + b_1$
Spring : y = $b_0 + b_2 x_2$ so $x_2 = 1$, thus y = $b_0 + b_2$
Fall : y = $b_0 + b_3 x_3$ so $x_1 = 1$, thus y = $b_0 + b_3$

The response functions are relating to summer because summer doesn't have an indicator variable.

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

(i) β_0 = The summer effects in sales of a certain line of shoes.
(ii) β_1 = Estimated difference in sales between winter and summer.
(iii) β_2 = Estimated difference in sales between spring and summer
(iv) β_3 = Estimated difference in sales between fall and summer.