PSTAT 126 Homework 4

Shaiyon Hariri

5/18/2020

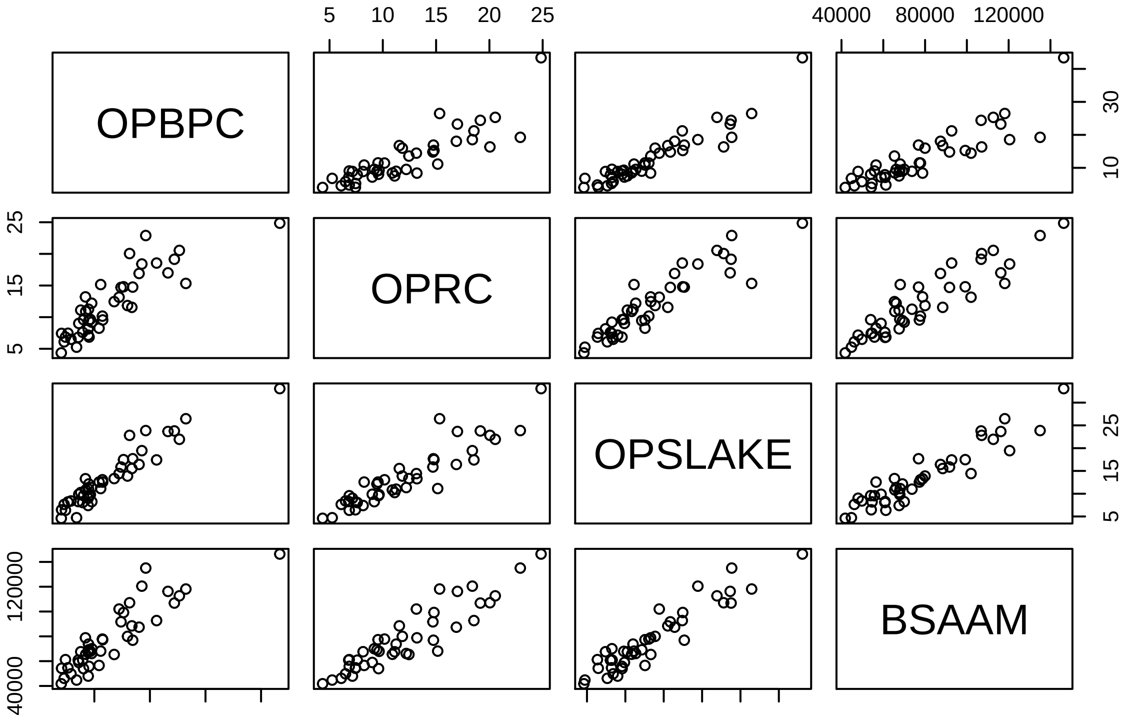
1. This problem uses the water data set in the alr4 package. For this problem, consider the regression problem with response BSAAM, and three predictors as regressors given by OPBPC, OPRC, and OPSLAKE.

```
library(alr4)
library(dplyr)
```

(a) Examine the scatterplot matrix drawn for these three regressors and the response. What should the correlation matrix look like? Compute the correlation matrix to verify your results.

```
data <- select(water, OPBPC, OPRC, OPSLAKE, BSAAM)

pairs(data)
```



It appears from the scatterplot matrix that the response variables have a positive linear relationship with the response BSAAM. There is also a single peculiar outlier present in most of the plots.

```
cor(data)
```

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

The correlation matrix supports the observation made prior, and displays a strong positive linear relationship between the variables.

(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=data)

summary(model)
```

```
##
## Call:
## lm(formula = BSAAM ~ OPBPC + OPRC + OPSLAKE, data = data)
##
## 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 regression summary contains the probability of observing a value larger than the absolute value of the test statistic t, which measures how many standard deviations away from 0 the coefficient is. A value smaller than the chosen significance level in this column means that we can reject the null hypothesis that the coefficient in question's true value is 0, or that there is no linear relationship between the variable and the response.

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

```
aov.result <- anova(model)
format(aov.result, scientific=FALSE)
```

```
##           Df      Sum Sq     Mean Sq    F value                      Pr(>F)
## OPBPC      1 21458217726 21458217726 311.161046 0.0000000000000000003471238
## OPRC       1  2561637374  2561637374  37.145758 0.0000038251219974488991733
## OPSLAKE    1   641654049   641654049   9.304489 0.0040973305039808246719041
## Residuals 39  2689509185    68961774         NA                          NA
```

```
model2 <- model <- lm(BSAAM ~ OPBPC + OPRC, data=data)

aov.resultwo <- anova(model2)
format(aov.resultwo, scientific=FALSE)
```

```
##           Df      Sum Sq     Mean Sq   F value                      Pr(>F)
## OPBPC      1 21458217726 21458217726 257.66636 0.0000000000000000004949317
## OPRC       1  2561637374  2561637374  30.75967 0.0000020511819479627289623
## Residuals 40  3331163233    83279081        NA                          NA
```

```
cat("SSR(OPSLAKE|OPBPC, OPRC) =", aov.result$`Sum Sq`[3],
    "   SSE(OPBPC,OPRC) =", aov.resultwo$`Sum Sq`[3])
```

```
## SSR(OPSLAKE|OPBPC, OPRC) = 641654049    SSE(OPBPC,OPRC) = 3331163233
```

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.

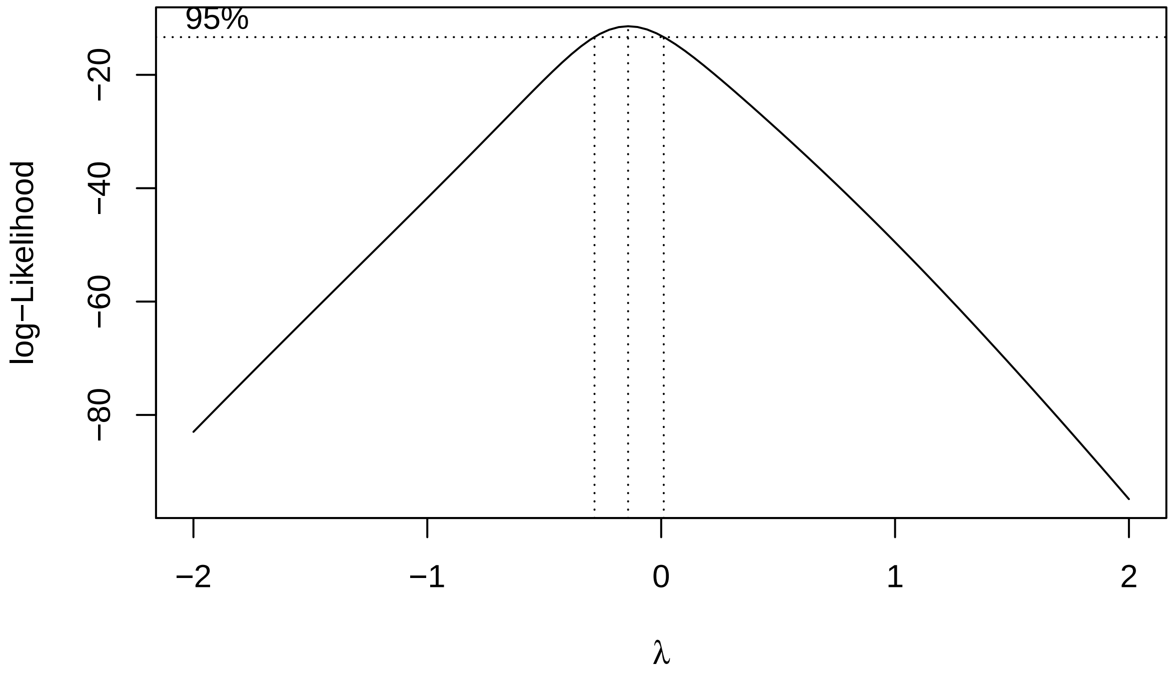
```r
library(MASS)
```

(a) Starting with the full second-order model

E(Life | Speed, Feed) = B0 + B1*Speed* + *B2*Feed + B11*Speed*ˆ*2* + *B22*Feedˆ2 + B12*Speed*Feed

Use the Box-Cox method to show that an appropriate scale for the response is the logarithmic scale.

```r
model <- lm(Life ~ Speed + Feed +Speed^2 + Feed^2 + Speed*Feed, data=lathe1)
boxcox(model)
```



The value of lambda is near 0, therefore a log(y) transformation is necessary.

(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.

Ho : The fit of the reduced model and the full model are equal

Ha : The fit of the reduced model is significantly reduced compared to the full model

```
full <- lm(log(Life) ~ Speed + Feed +Speed^2 + Feed^2 + Speed*Feed, data=lathe1)
reduced <- lm(log(Life) ~ 1, data=lathe1)

format(anova(full, reduced), scientific=FALSE)
```

```
##   Res.Df       RSS Df Sum of Sq        F            Pr(>F)
## 1     16  3.700648 NA        NA       NA                NA
## 2     19 41.532613 -3 -37.83196 54.52301 0.00000001272721
```

The p-value from the F test is extremely small, therefore we reject the null hypothesis that the fit of the reduced intercept-only model and the full model are equal.

(c) Explain the practical meaning of the hypothesis H0 : B1 = B11 = B12 = 0 in the context of the above model.

This null hypothesis implies that the Speed predictor has no linear relationship with the response, and adds no predictive power.

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

Ho : B1 = B11 = B12 = 0

Ha : B1 or B11 or B12 != 0

```
reduced <- lm(log(Life) ~ Feed + Feed^2, data=lathe1)

format(anova(full, reduced), scientific=FALSE)
```

```
##   Res.Df       RSS Df Sum of Sq        F            Pr(>F)
## 1     16  3.700648 NA        NA       NA                NA
## 2     18 34.039850 -2  -30.3392 65.58678 0.00000001951288
```

The p-value from the F test is extremely small, therefore we reject the null hypothesis that the fit of the reduced model and the full model are equal.

3. Consider the following model and the corresponding ANOVA table: Y = B0 + B1*X1* + *B2*X2 + epsilon, where r squared = 0.637, epsilon is the random error and the Yi's are independent.

The ANOVA Table

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

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

DF for the model is is 2, as there are two independent variables (X1, X2). Thus, the C total DF = 119, and n = 120.

R^2 = 1 - SSE/SST. We know that SSE is 17.90761, and R^2 is 0.637, therefore plugging in the numbers, SST = 49.33226.

SS model = SS - SSE, thus, SS for model = 31.42465.

MS = SSM/DF, thus, MS for model = 15.712325

F = model MS / MSE, thus, F statistic = 102.6537

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

Ho : B1 = B2 = 0

Ha : B1 or B2 != 0

(c) What is the estimated value of sigma squared based on then results shown in the table?

The estimated population variance is equivalent to the MSE, so 0.15306

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:

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

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

```
model <- lm(y ~ x)
summary(model)
```

```
##
## 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
```

y = 485.67 + 103.13*x

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

```
able <- 485.6667 + 1*103.1333
able
```

```
## [1] 588.8
```

```
unable <- 485.6667 + 0*103.1333
unable
```

```
## [1] 485.6667
```

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

```
summary(model)
```

```
##
## 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
```

The p-value is greater than the significance level alpha, therefore we fail to reject the null hypothesis that there is no lienar 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, fall, summer) effects on sales of a certain line of shoes: E(Y) = B0 + B1*X1 + B2*X2 + B3*X3. The X's are one hot encoded categorical variables (in the order listed) for the seasons.

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

As the X's are binary variables, when one is active the function will be just the intercepet and the respective coefficient.

Summer : y = B0

Winter : y = B0 + B1

Spring : y = B0 + B2

Fall : y = B0 + B3

(b) Interpret each of the following quantities: (i) B0 (ii) B1 (iii) B2 (iv) B3

(i) This is the amount of sales during the summer.

(ii) This is the estimated difference in sales between the winter and summer (positive if winter is more, negative if it's less)

(iii) This is the estimated difference in sales between the spring and summer (positive if spring is more, negative if it's less)

(iv) This is the estimated difference in sales between the fall and summer (positive if fall is more, negative if it's less)