Homework 2

Lydia Strebe

June 27, 2019

Problem 1

```
library(alr4)

data(oldfaith)
attach(oldfaith)
```

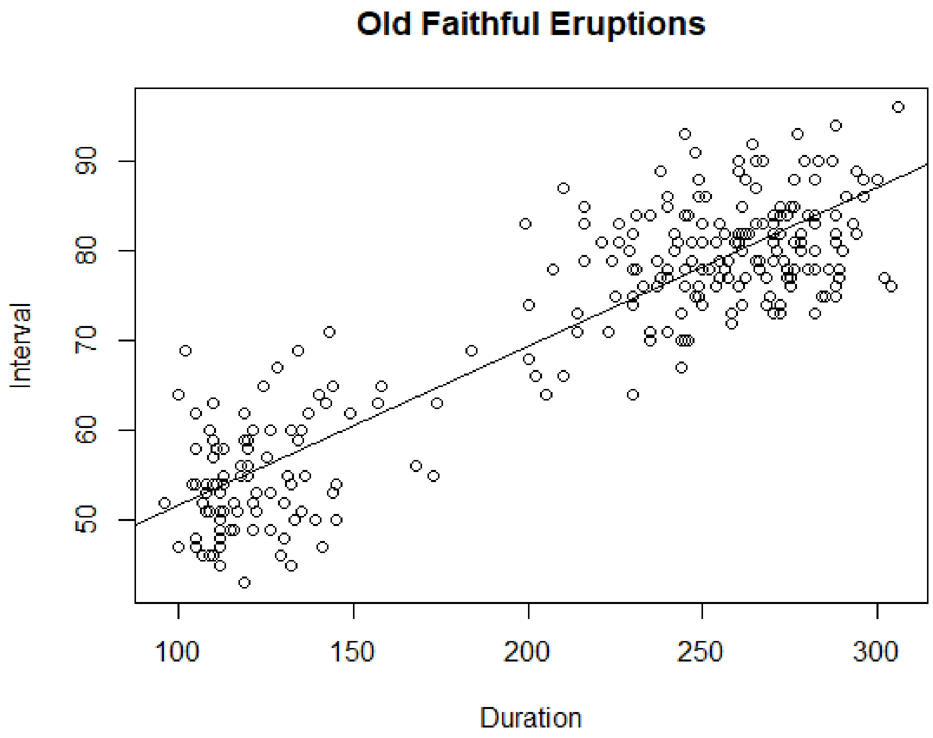
a) The explanatory variable is the duration of the last eruption (“Duration”) and the response variable is the time between the last eruption and the next eruption (“Interval”). We want to use Duration to predict Interval.

b)

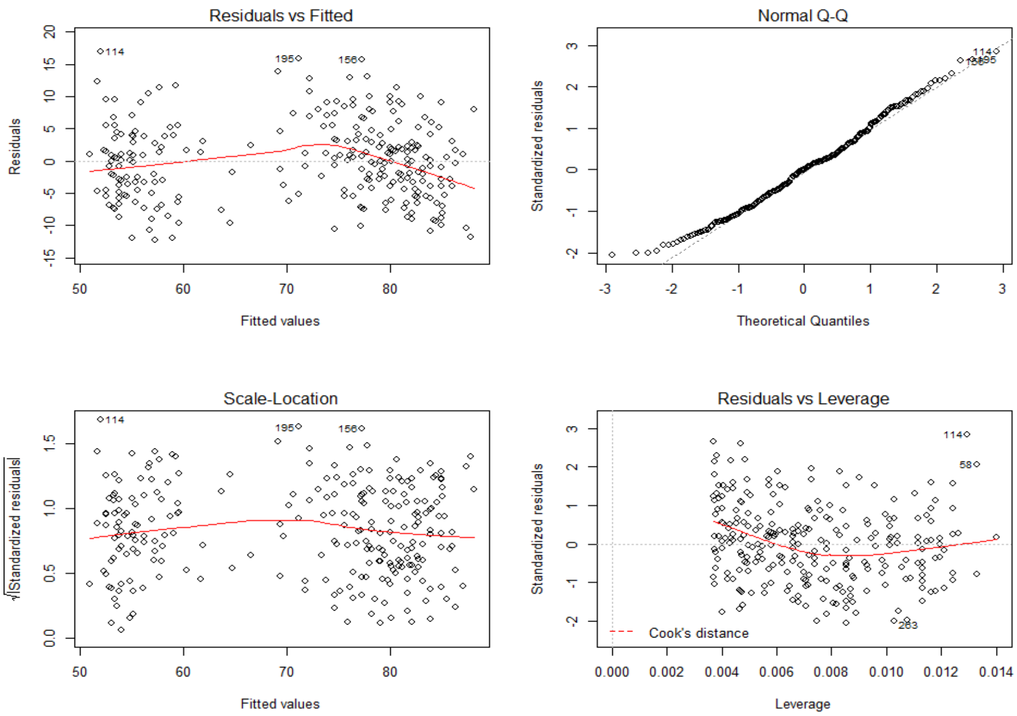
```
of_mod=lm(Interval~Duration)
summary(of_mod)

##
## Call:
## lm(formula = Interval ~ Duration)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.3337  -4.5250   0.0612   3.7683  16.9722
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.987808   1.181217   28.77   <2e-16 ***
## Duration     0.176863   0.005352   33.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.004 on 268 degrees of freedom
## Multiple R-squared:  0.8029, Adjusted R-squared:  0.8022
## F-statistic:  1092 on 1 and 268 DF,  p-value: < 2.2e-16

plot(Interval~Duration,main="Old Faithful Eruptions")
abline(of_mod)
```



```r
par(mfrow=c(2,2))
plot(of_mod)
```



```r
par(mfrow=c(1,1))
```

c) Based on the scatterplot of the data, the linearity assumption seems to hold, although there is a possible pattern in the residuals vs. fitted plot. This is also a sign that the equal variance assumption may be violated (slightly, if at all). However, based on the QQ plot, the normality assumption holds and we know from the description of the data collection that the independence assumption holds.

d) The intercept 33.99 represents the minimum amount of time in minutes between each eruption and 0.177, as the slope, is the additional time in minutes for every second the last eruption lasted. Therefore, the model estimates that the interval between the last eruption and the next eruption is 33.99 minutes plus 0.177 times the duration of the last eruption.

```
anova(of_mod)

## Analysis of Variance Table
##
## Response: Interval
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Duration    1  39358   39358    1092 < 2.2e-16 ***
## Residuals 268   9659      36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$

F-value: 1092

p-value: 2.2e-16

Conclusion: We reject the null hypothesis since the p-value is (much) less than 0.05.

Problem 2

a) $R^2 = 1\text{-}RSS/SST =$

```
1-(572.0136/731.961)

## [1] 0.218519
```

b) About 22% of the total sample variablility in weight is explained by the linear regression model.

Problem 3

$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} [(x_i-\bar{x})*(x_i-\bar{x})] = \sum_{i=1}^{n} [(x_i-\bar{x})x_i-(x_i-\bar{x})\,\bar{x}]$

$= \sum_{i=1}^{n} [(x_i-\bar{x})x_i] - \sum_{i=1}^{n} [(x_i-\bar{x})\,\bar{x}]$

$= \sum_{i=1}^{n} [(x_i-\bar{x})x_i] - \bar{x} *\sum_{i=1}^{n} [(x_i-\bar{x})]$

$= \sum_{i=1}^{n} [(x_i-\bar{x})x_i] - \bar{x} *[\sum_{i=1}^{n} (x_i) - n\bar{x}]$

$= \sum_{i=1}^{n} [(x_i-\bar{x})x_i] - \bar{x} *[\sum_{i=1}^{n} (x_i) - n*(\sum_{i=1}^{n} (x_i)/n)]$

$= \sum_{i=1}^{n} [(x_i-\bar{x})x_i] - \bar{x} *[\sum_{i=1}^{n} (x_i) - \sum_{i=1}^{n} (x_i)]$

$= \sum_{i=1}^{n} [(x_i-\bar{x})xi]$

Problem 4

a) The conclusion given by the business analyst is somewhat hasty. Before drawing these conclusions, the business analyst should check the assumptions of linearity, independence, normality, and equal variance. Additionally, it appears that there is at least one outlier in the data set and, since it is an extreme x-value, it could be a leverage point. A better model could probably be fit if this observation was removed.

b) The ordinary straight line regression model seems to fit the data well, although it could probably be improved by removing the influential outlier.

Problem 5

```
airfare=read.table("http://gattonweb.uky.edu/sheather/book/docs/datasets/airfares.txt",header=TRUE)

attach(airfare)
air_mod=lm(Fare~Distance)

a)

sort(hatvalues(air_mod))

##         15          3          8          7         12         10
## 0.05882392 0.06213499 0.06908091 0.07351534 0.07665861 0.07745448
##          9          6          4          5          2         11
## 0.08588990 0.09116201 0.11267276 0.12959128 0.13071191 0.13095319
##         14          1         16         17         13
## 0.13209804 0.13416776 0.15398450 0.23783524 0.24326517
```

The observation with the largest leverage is the 13th observation. In this case 4/n =

```
4/17

## [1] 0.2352941
```

Which is less than the leverage for *both* the 13th and 17th observation. Therefore we would consider these to be leverage points.

b)

```
sort(rstandard(air_mod))

##           17           16            1            2            9
## -2.009328033 -1.539476397 -1.070782935 -0.729205219 -0.449554976
##           14            7           11           10           15
## -0.442860375 -0.006706541  0.062407480  0.102395666  0.230068107
##            5           12            4            6            8
##  0.235495724  0.274377543  0.276572788  0.573338687  0.701659153
##            3           13
##  0.835139842  2.919064802
```

The observation with the largest standardized residual is also the 13th data point. It is an outlier according to the rule in the textbook since the standardized residual is above 2 (i.e., outside of the -2 to 2 range).

```
sort(cooks.distance(air_mod))

##            7           11           10           15           12
## 1.784460e-06 2.934379e-04 4.401410e-04 1.654116e-03 3.125113e-03
##            5            4            9           14            6
## 4.128465e-03 4.856507e-03 9.494655e-03 1.492552e-02 1.648618e-02
##            8            3            2            1           16
## 1.826705e-02 2.310385e-02 3.997799e-02 8.883565e-02 2.156824e-01
##           17           13
## 6.299398e-01 1.369600e+00
```

The observation with the largest Cook's distance is, once again, the 13th observation. This is noteworthy given that it is quite a bit larger than

```
4/(17-2)

## [1] 0.2666667
```

which is the rough cut off according to Fox.

Problem 6

a) Without calculating the leverage of the two observations, we can see that the 11th observation has the higher leverage because the x-value is further from the mean.

b) Leverage of 11th observation:

```
h11=(1/100)+((30.5-20)^2)/400
h11

## [1] 0.285625
```

Leverage of the 69th observation:

```
h69=(1/100)+((26.6-20)^2)/400
h69

## [1] 0.1189
```

These results are consistent with the prediction in part a).

c) Residuals
11th observation:

```
e11=5-5.3
e11

## [1] -0.3
```

69th observation:

```
e69=4.5-4.3
e69

## [1] 0.2
```

d) Standardized residuals

```
S=sqrt(24.5/98)
```

11th observation:

```
r11=e11/(S*sqrt(1-h11))
r11

## [1] -0.7098852
```

69th observation:

```
r69=e69/(S*sqrt(1-h69))
r69

## [1] 0.4261352
```

e) Cook's distance 11th observation:

```
((r11^2)/2)*(h11/(1-h11))

## [1] 0.1007433
```

69th observation:

```
((r69^2)/2)*(h69/(1-h69))

## [1] 0.01225241
```