Linear Models Assignment 2

Subhrajyoty Roy (BS - 1613)

October 7, 2018

Question 1

Consider data on Table 15.1. One step in the manufacture of large engines requires that holes of very precise dimensions be drilled. The tools that do the drilling are regularly examined and are adjusted to ensure that the holes meet the required specifications. Part of the examination involves measurement of the diameter of the drilling tool. A team studying the variation in the sizes of the drilled holes selected this measurement procedure as a possible cause of variation in the drilled holes. They decided to use an experiment as one part of this examination. The diameters in millimeters (mm) of five tools were measured by the same operator at three times (8:00 a.m., 11:00 a.m., and 3:00 p.m.). Three measurements were taken on each tool at each time. The person taking the measurements could not tell which tool was being measured, and the measurements were taken in random order.

```
tooldata <- read.csv('tool-diameter.csv')
head(tooldata)
```

```
  Tool Time Diameter.1 Diameter.2 Diameter.3
1    1    1     25.030     25.030     25.032
2    1    2     25.028     25.028     25.028
3    1    3     25.026     25.026     25.026
4    2    1     25.016     25.018     25.016
5    2    2     25.022     25.020     25.018
6    2    3     25.016     25.016     25.016
```

```
sapply(tooldata, class)
```

```
      Tool       Time Diameter.1 Diameter.2 Diameter.3
 "integer"  "integer"  "numeric"  "numeric"  "numeric"
```

We find that *Tool* and *Time* are of integer data type, and we need to convert them as factors.

```
tooldata$Tool <- as.factor(tooldata$Tool)
tooldata$Time <- as.factor(tooldata$Time)
```

However, to perform two way ANOVA model, we need to stack the diameters in a single column. We are using *melt* function from *reshape2* package for this.

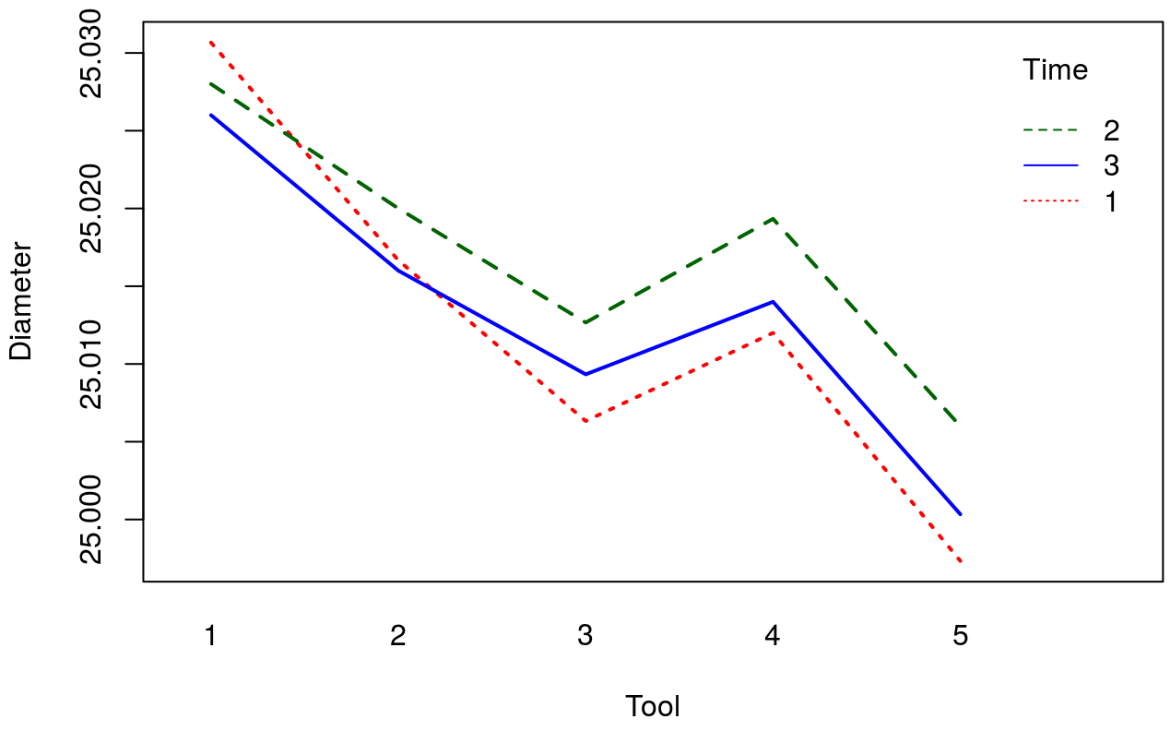
```r
library(reshape2)
tooldata = melt(tooldata)
knitr::kable(head(tooldata))
```

Tool	Time	variable	value
1	1	Diameter.1	25.030
1	2	Diameter.1	25.028
1	3	Diameter.1	25.026
2	1	Diameter.1	25.016
2	2	Diameter.1	25.022
2	3	Diameter.1	25.016

(a) Use different plots (to be decided by you) to summarize the main features of the data.

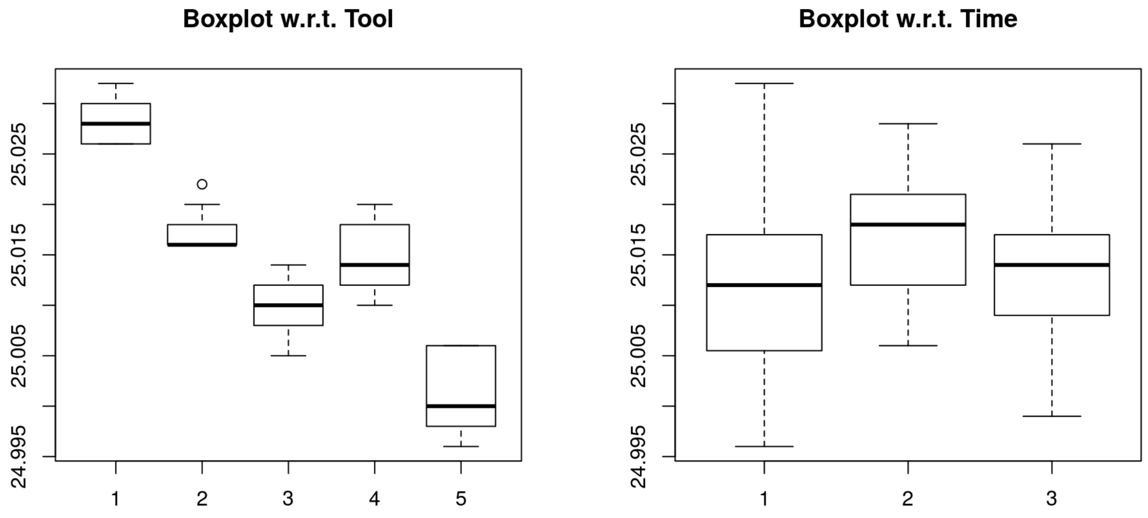
Let us first check whether there seems some interaction effect between Tool and Time factors. We can visualize this using an interaction plot.

```r
interaction.plot(tooldata$Tool, tooldata$Time, tooldata$value, lwd = 2, xl
ab = "Tool", ylab = "Diameter", trace.label = "Time", col = c("red","darkg
reen","blue"))
```



We can also look at the boxplots of the diameter measurements with respect to different levels of tools and different levels of time.

```
par(mfrow = c(1,2))  #creates the grid for subplot
boxplot(value ~ Tool, data = tooldata, main = "Boxplot w.r.t. Tool")
boxplot(value ~ Time, data = tooldata, main = "Boxplot w.r.t. Time")
```



It seems there are sufficient variation with respect to both the factors.

(b) Use a suitable two way model to explain the data and analyze it. Write a short report on your findings from the plot as well as the ANOVA results.

As it seems by the interaction plot, we consider a two way ANOVA model with interaction. The summary of the model is as follows:

```
fit <- aov(value ~ Tool * Time, data = tooldata)
summary(fit)
```

```
            Df   Sum Sq   Mean Sq F value   Pr(>F)    
Tool         4 0.003597 0.0008993 412.944  < 2e-16 ***
Time         2 0.000190 0.0000950  43.602 1.33e-09 ***
Tool:Time    8 0.000133 0.0000167   7.645 1.55e-05 ***
Residuals   30 0.000065 0.0000022                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that the p-value of the factor effects (i.e. of levels of *Tool* and levels of *Time*) and the interaction effect between the levels of factor *Tool* with levels of factor *Time* is considerably small. Therefore, under the significance level of 0.001, we reject the null hypothesis that these factors and interaction has no effect on the dimensions of holes. That means we accept the fact that the given data shows a strong indication of significant effects of *Tool* and *Time* and their interaction.

This also agrees with the intuition of significant effects after the visualization of the boxplots.

Question 2

Consider data on Table 15.2 which describe a study designed to determine how the frequency that a supermarket product is promoted at a discount and the size of the discount affect the price that customers expect to pay for the product. Each one of you should delete some observations to make this data set unbalanced such that for each of the 40%, 30%, 20% and 10% discounts there are in all 38 observations, spread unequally over the number of promotions. (e.g. for 40% discount there are 9, 10, 10, 9 observations across the number of promotions).

```
pricedata <- read.csv('promotion-price.csv') #read the data
pricedata$promotions <- as.factor(pricedata$promotions) #change into facto
r
pricedata$discount <- as.factor(pricedata$discount)
knitr::kable(head(pricedata, 3))  #check top 3 data
```

promotions	discount	price.1	price.2	price.3	price.4	price.5	price.6	price.7	price.8	price
1	40	4.10	4.50	4.47	4.42	4.56	4.69	4.42	4.17	4.3
1	30	3.57	3.77	3.90	4.49	4.00	4.66	4.48	4.64	4.3
1	20	4.94	4.59	4.58	4.48	4.55	4.53	4.59	4.66	4.7

Before proceeding, we should order this dataset according to *discount* prices.

```
pricedata <- pricedata[order(pricedata$discount),]  # sort the data w.r.t.
discount variable
rownames(pricedata) <- 1:nrow(pricedata)  #reorder rownames
knitr::kable(head(pricedata,3))
```

promotions	discount	price.1	price.2	price.3	price.4	price.5	price.6	price.7	price.8	price
1	10	5.19	4.88	4.78	4.89	4.69	4.96	5.00	4.93	5.1
3	10	4.90	5.15	4.68	4.98	4.66	4.46	4.70	4.37	4.6
5	10	4.31	4.36	4.75	4.62	3.74	4.34	4.52	4.37	4.4

Now, we should delete some observation. For this, we choose 2 random numbers from 1 to 4, similarly from 5 to 8 and so on. After that, we choose one random number between 1 to 10 to remove the corresponding price.

```
set.seed(1613)  #that is my roll number, set the seed for reproducibility
a = list(1:4, 5:8, 9:12, 13:16)  #we should sample from each element of th
is list
a = sort(sapply(a, sample, size = 2)) #performs the sampling
b = sample(3:12, size = 8, replace = TRUE) #since 3:12 contains the prices
for (i in 1:8){
  pricedata[a[i], b[i]] <- NA  #set those values as NA
}
knitr::kable(head(pricedata))  #check top few data
```

promotions	discount	price.1	price.2	price.3	price.4	price.5	price.6	price.7	price.8	price
1	10	5.19	4.88	4.78	4.89	4.69	4.96	5.00	4.93	5.1
3	10	4.90	5.15	4.68	NA	4.66	4.46	4.70	4.37	4.6
5	10	4.31	NA	4.75	4.62	3.74	4.34	4.52	4.37	4.4
7	10	4.04	4.22	4.89	3.89	4.26	4.41	4.39	4.52	3.8
1	20	4.94	4.59	4.58	4.48	4.55	NA	4.59	4.66	4.7
3	20	4.88	4.80	4.46	4.73	3.96	4.42	4.30	4.68	N

Again, we should *melt* the dataset so that the prices comes in a single column.

```
pricedata <- melt(pricedata)
pricedata <- pricedata[complete.cases(pricedata), ]  #removes the NA value
s
knitr::kable(head(pricedata))
```

promotions	discount	variable	value
1	10	price.1	5.19
3	10	price.1	4.90
5	10	price.1	4.31
7	10	price.1	4.04
1	20	price.1	4.94
3	20	price.1	4.88

To check whether there are unequal number of observations, consider the following.

```
table(pricedata$promotions)
```

```
 1  3  5  7 
38 37 37 40 
```

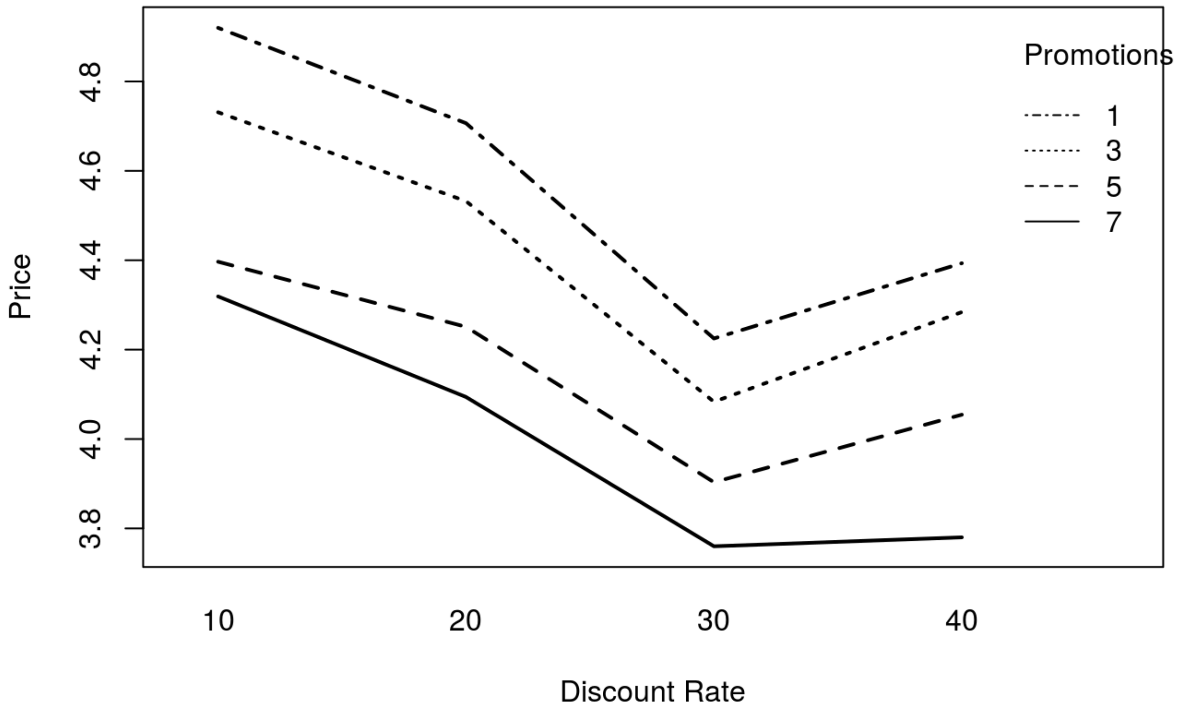
```
table(pricedata$discount)
```

```
10 20 30 40 
38 38 38 38 
```

Use different plots (to be decided by you) to summarize the main features of the data.

We again use interaction plot to see whether there is any interaction effect between discount and promotions.

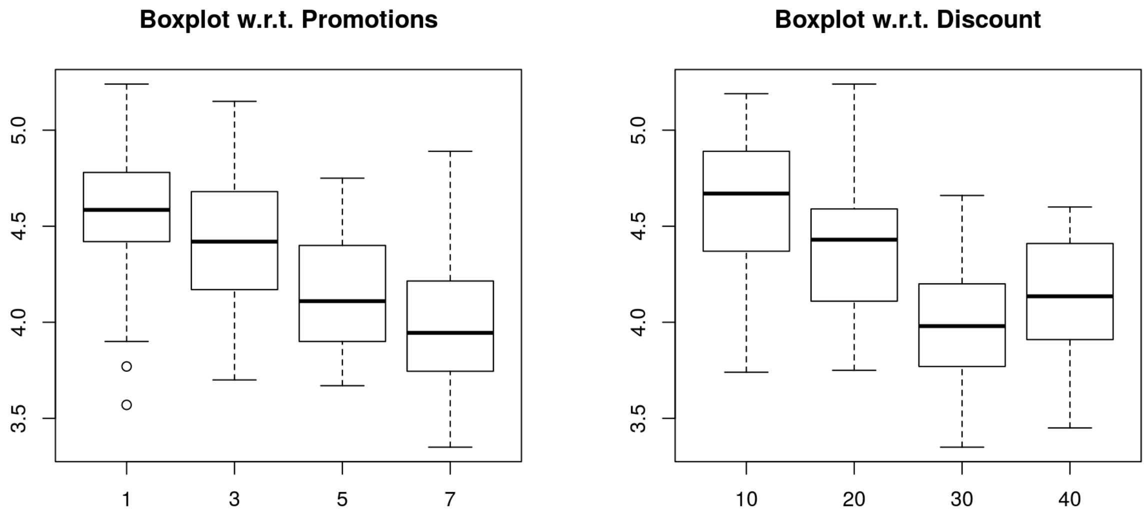
```
interaction.plot(pricedata$discount, pricedata$promotions, pricedata$value
, xlab = "Discount Rate", ylab = "Price", trace.label = "Promotions", lwd
= 2)
```



It seems that there is no strong interaction effect between these two factors in concern.

We can again see the boxplots of the prices with respect to different levels of Discount Rate and different levels of Promotion.

```
par(mfrow = c(1,2))  #creates the grid for subplot
boxplot(value ~ promotions, data = pricedata, main = "Boxplot w.r.t. Promotions")
boxplot(value ~ discount, data = pricedata, main = "Boxplot w.r.t. Discount")
```



Analyze the data with a two-way ANOVA without interaction. (Justify from your plots if this is a reasonable assumption). Prepare a short report using your plots and ANOVA results, explaining how the expected price depends on the number of promotions and the percent of the discount.

We clearly observe that from the interaction plot itself, we do not find any reasonably strong interaction effect. However, we should perform a statistical test in order to justify our claim.

```
fit <- aov(value ~ discount * promotions, pricedata)
summary(fit)
```

```
                     Df Sum Sq Mean Sq F value Pr(>F)    
discount              3  8.203  2.7342  43.490 <2e-16 ***
promotions            3  7.622  2.5407  40.412 <2e-16 ***
discount:promotions   9  0.232  0.0257   0.409  0.928    
Residuals           136  8.550  0.0629                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observe that in two-way unbalanced ANOVA model, the sum of squares of the interaction does not depend on the order of the explanatory covariates. Hence, from the above table, we find that the F-value for the interaction effect is resonably low, as well as the p-value is quite high, nearly 1. This suggests that we should not reject the null hypothesis that there is no interaction effect based on the above data.
Now, we can simply ignore the effect of interaction and fit a two-way ANOVA model without

interaction. In such a case, there are two possible hypothesis to test:

- There is no effect of the variable discount.
- There is no effect of the variable promotions.

Starting with the first hypothesis, we require the adjusted sum of squares for the variable discount, after adjusting for the effect of promotions variable. Hence, we perform the following test.

```
fit <- aov(value ~ promotions + discount, pricedata)
summary(fit)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)    
promotions    3  7.575  2.5251   41.69 <2e-16 ***
discount      3  8.250  2.7499   45.40 <2e-16 ***
Residuals   145  8.782  0.0606                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We observe that the p-value of the F-statistic for adjusted discount variable is extremely close to 0, suggesting we should reject the null hypothesis that there is no effect of discount variable,
i.e. accepting that the variable discount plays a significant role in determination of price of the commodity.
To test the second hypothesis, we require to perform the same statistical test, but reversing the order of the covariates.

```
fit <- aov(value ~ discount + promotions, pricedata)
summary(fit)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)    
discount      3  8.203  2.7342   45.14 <2e-16 ***
promotions    3  7.622  2.5407   41.95 <2e-16 ***
Residuals   145  8.782  0.0606                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We find that the F-statistic does not differ much. Also, noting that the p-value corresponding to the promotion variable being reasonably low, we can reject the null hypothesis that promotion has no effect and conclude that the variable promotion has statistically significant effect on determination of price variable.

```
print(fit$coefficients)
```

```
(Intercept)  discount20  discount30  discount40 promotions3 promotions5 
  4.8771643  -0.1981544  -0.6002632  -0.4662469  -0.1526176  -0.4101336 
promotions7 
 -0.5727482 
```

From the above list of coefficients, we can explain the effect of different levels of factors in determination of the price. Starting with the base level, i.e. when promotion is at level 1 and discount is 10%, then the expected price is 4.87. Increasing promotion to level 3, 5 and 7 decreases the expected price by 0.15, 0.41 and 0.57 respectively. After eliminating the effect of the promotions, increasing discount rate to 20%, 30% or 40%, decreases the expected price by 0.19, 0.60 and 0.46 units respectively.

THANK YOU