STAT407 - Statistical Design and Analysis of Experiments Project

Louis Dulana, Qing Gong

Research Question 1: *Do the machine type, wearing a mask, and subjects affect the heart rate?*

We will input the data as follows:

A = machine, B = mask, C = subject, where

-1 = Louis, Apple, and without mask and

1 = Qing, Fitbit, with mask.

```
data2 <- read.csv("raw.csv")
#head(data)
#covert  A, B, C into factors
nms <- c("A", "B", "C")
data2[,nms] <- lapply(data2[,nms], factor)
```

subject	before	after	mask	device	diff	A	B	C
Qing	91	94	1	treadmill	3	1	1	1
Qing	91	124	0	elliptical	33	-1	-1	1
Qing	92	121	1	elliptical	29	-1	1	1
Qing	97	111	0	treadmill	14	1	-1	1
Louis	72	86	1	elliptical	14	-1	1	-1
Louis	72	92	0	elliptical	20	-1	-1	-1
Louis	71	79	1	treadmill	8	1	1	-1
Louis	76	79	0	treadmill	3	1	-1	-1

Let's begin with fitting a full model.

Since it is impossible to obtain information from the full model, that is, checking which factors have significant effect on the heart rate, we will instead check the half-normal probability plot.

```
m1 <- lm(diff ~A*C*B, data = data2)
anova(m1)
```

```
## Warning in anova.lm(m1): ANOVA F-tests on an essentially perfect fit are
## unreliable
```

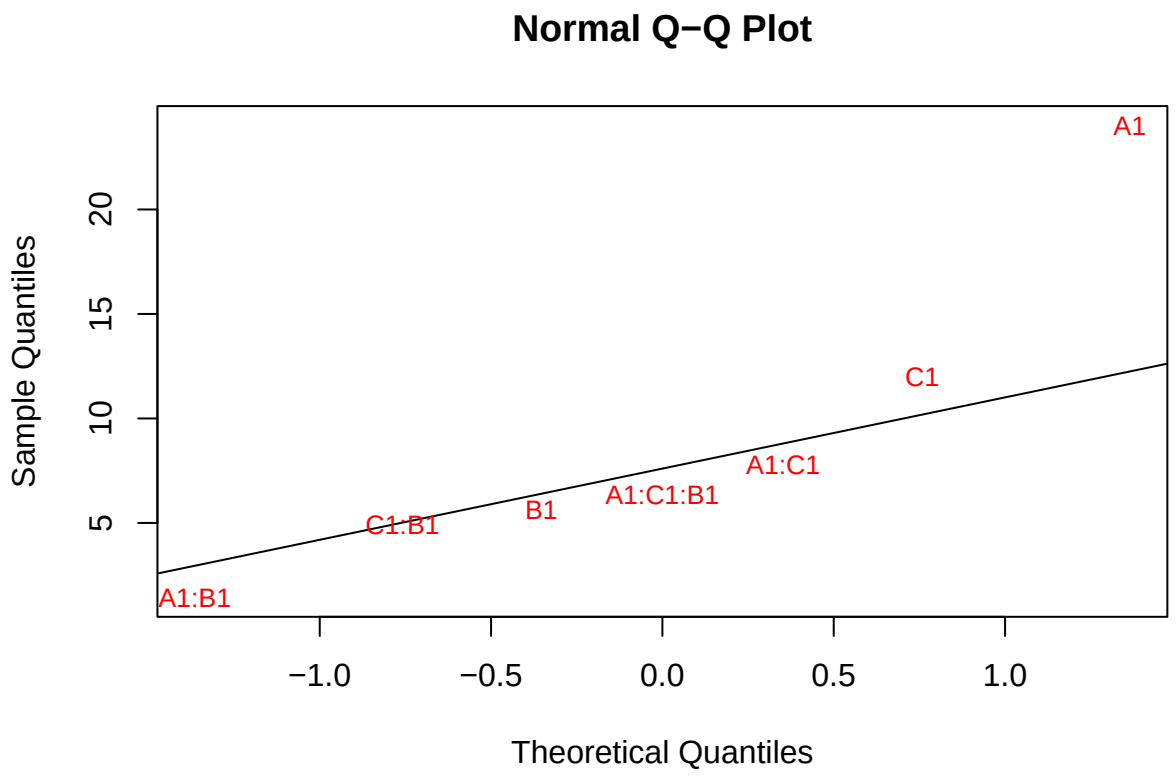
```
## Analysis of Variance Table
##
## Response: diff
##           Df Sum Sq Mean Sq F value Pr(>F)
## A          1  578.0   578.0     NaN    NaN
## C          1  144.5   144.5     NaN    NaN
## B          1   32.0    32.0     NaN    NaN
```

```
## A:C        1   60.5    60.5     NaN    NaN
## A:B        1    2.0     2.0     NaN    NaN
## C:B        1   24.5    24.5     NaN    NaN
## A:C:B      1   40.5    40.5     NaN    NaN
## Residuals  0    0.0     NaN
```

```
effect <- abs(m1$effects)[-1]
effect
```

```
##        A1        C1        B1     A1:C1     A1:B1     C1:B1  A1:C1:B1
## 24.041631 12.020815  5.656854  7.778175  1.414214  4.949747  6.363961
```

```
half <- qqnorm(effect, type = "n")
qqline(effect)
text(half$x, half$y, labels = names(effect), cex = 0.8, col="red")
```



The half-normal probability plot shows that factors A and C (machine and subject) may have statistically significant effect on the heart rate. Thus, we will fit another model with just these two factors.

```
m2 <- lm(diff ~ A+C , data = data2)
anova(m2)
```

```
## Analysis of Variance Table
##
## Response: diff
```

```
##           Df Sum Sq Mean Sq F value  Pr(>F)
## A          1  578.0   578.0 18.1191 0.00804 **
## C          1  144.5   144.5  4.5298 0.08659 .
## Residuals  5  159.5    31.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Clearly, we could see that machine type (A) is significant at 5%. Subject (C) is not significant at 5%, but is significant at 10%. Let's now check the normality and homoscedasticity assumption for the reduced model (m2).

```
shapiro.test(m2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m2$residuals
## W = 0.89614, p-value = 0.2666
```

```
bartlett.test(m2$residuals, data2$A)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  m2$residuals and data2$A
## Bartlett's K-squared = 0.23357, df = 1, p-value = 0.6289
```

```
bartlett.test(m2$residuals, data2$C)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  m2$residuals and data2$C
## Bartlett's K-squared = 0.15026, df = 1, p-value = 0.6983
```

Based on the shapiro and bartlett's test, normality and equal variance assumptions are verified. We will now move on to Research Question 2.

Research Question 2: *Do watch brand, wearing a mask, and subject affect the reported heart rate while using the treadmill?*

We will input the data as follows:

A = watch, B = mask, C = subject, where

-1 = Louis, Apple, and without mask and

1 = Qing, Fitbit, with mask.

```
proj.data <- read.csv("project_data.csv")
#head(proj.data,16)
```

subject	before	after	mask	watch	difference	A	B	C
Louis	71	124	1	fitbit	53	1	1	-1
Louis	73	109	0	fitbit	36	1	-1	-1
Louis	76	143	0	apple	67	-1	-1	-1
Louis	71	113	1	apple	42	-1	1	-1
Louis	71	106	1	fitbit	35	1	1	-1
Louis	71	135	0	fitbit	64	1	-1	-1
Louis	69	111	0	apple	42	-1	-1	-1
Louis	70	101	1	apple	31	-1	1	-1
Qing	91	122	1	fitbit	31	1	1	1
Qing	85	112	0	apple	27	-1	-1	1
Qing	97	119	0	fitbit	22	1	-1	1
Qing	86	108	1	apple	22	-1	1	1
Qing	80	99	1	fitbit	19	1	1	1
Qing	84	108	1	apple	24	-1	1	1
Qing	78	97	0	fitbit	19	1	-1	1
Qing	86	108	0	apple	22	-1	-1	1

Let's convert variables into factors before running anova.

```
proj.data$A <- as.factor(proj.data$A)
proj.data$B <- as.factor(proj.data$B)
proj.data$C <- as.factor(proj.data$C)


m2 <- lm(diff ~ (A+B+C)^3, data = proj.data)
anova(m2)
```

```
## Analysis of Variance Table
##
## Response: diff
##           Df Sum Sq Mean Sq F value Pr(>F)
## A          1   1.00    1.00  0.0118 0.9162
## B          1   6.25    6.25  0.0737 0.7928
## C          1  42.25   42.25  0.4985 0.5002
## A:B        1  12.25   12.25  0.1445 0.7137
## A:C        1  30.25   30.25  0.3569 0.5667
## B:C        1  64.00   64.00  0.7552 0.4102
## A:B:C      1  36.00   36.00  0.4248 0.5328
## Residuals  8 678.00   84.75
```

We could see from the full model that only factor C, which is subject had a significant effect on the reported heart rate. Also, notice that factor A does not affect the reported heart rate given the high p-value (0.9657). So, let's fit a reduced model with just B and C as factors.

```
m4 <- lm(diff ~ B+C, data = proj.data)
anova(m4)
```

```
## Analysis of Variance Table
##
## Response: diff
```

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## B          1   6.25   6.250  0.0989 0.7581
## C          1  42.25  42.250  0.6686 0.4283
## Residuals 13 821.50  63.192
```

Even with the reduced model, subject is still highly significant.

Let's also check the normality and homoscedasticity assumption.

```
shapiro.test(m4$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m4$residuals
## W = 0.92051, p-value = 0.1719
```

```
bartlett.test(m4$residuals, proj.data$B)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  m4$residuals and proj.data$B
## Bartlett's K-squared = 0.16569, df = 1, p-value = 0.684
```

```
bartlett.test(m4$residuals, proj.data$C)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  m4$residuals and proj.data$C
## Bartlett's K-squared = 2.5368, df = 1, p-value = 0.1112
```

Based on the shapiro and bartlett's test, normality and equal variance are verified except factor C violates homoscedasticiy assumption. So, as a remedy, let's try and take the log transformation of the response variable, difference.

```
proj.data$logdiff <- log(proj.data$diff)

m5 <- lm(logdiff ~ B+C, data = proj.data)
anova(m5)
```

```
## Analysis of Variance Table
##
## Response: logdiff
##           Df Sum Sq Mean Sq F value Pr(>F)
## B          1 0.0047 0.00467  0.0071 0.9342
## C          1 0.1217 0.12166  0.1845 0.6746
## Residuals 13 8.5730 0.65946
```

As a result, both normality and equal variance are now verified below.

```
shapiro.test(m5$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m5$residuals
## W = 0.90881, p-value = 0.1113
```

```
bartlett.test(m5$residuals, proj.data$B)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  m5$residuals and proj.data$B
## Bartlett's K-squared = 0.056527, df = 1, p-value = 0.8121
```

```
bartlett.test(m5$residuals, proj.data$C)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  m5$residuals and proj.data$C
## Bartlett's K-squared = 0.30849, df = 1, p-value = 0.5786
```

Additional Analysis: *Confounding 2^3 Factor Design with Two Blocks*

We will confound ABC with day as our blocks.

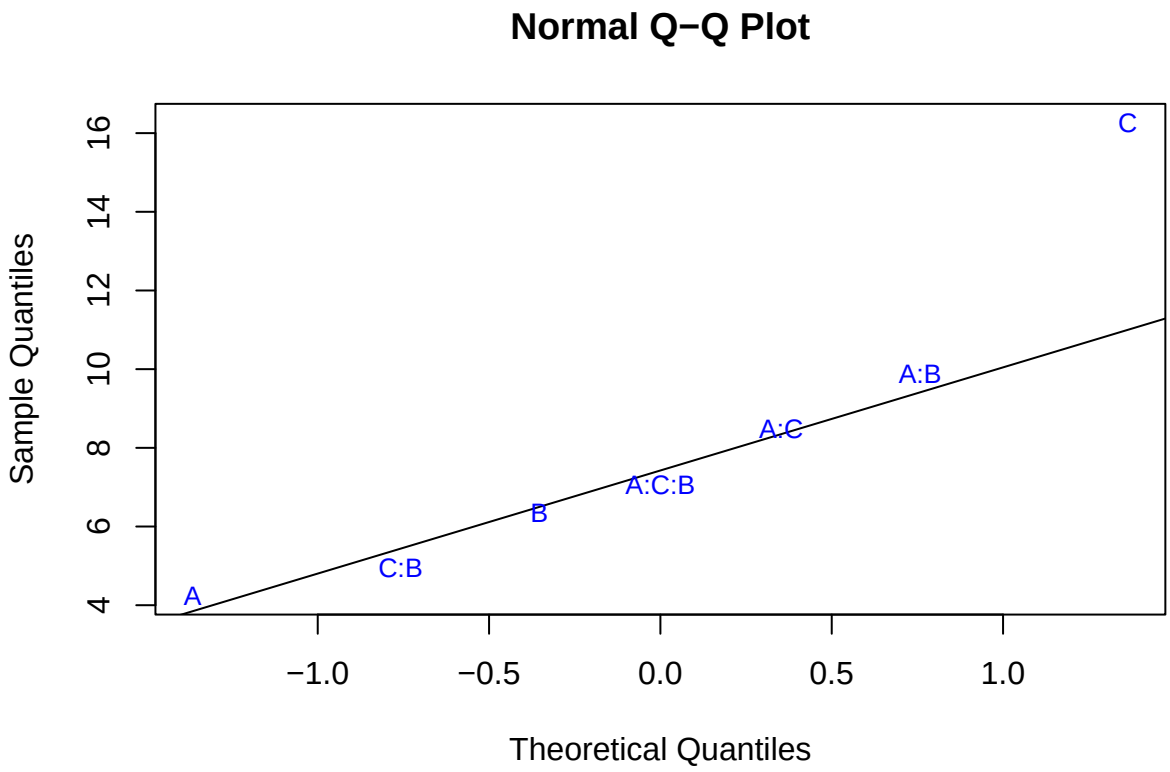
```
#ABC confounded as blocks
d2 <- read.csv("DAY2.csv")
ABC <- d2$A*d2$B*d2$C
mm1 <- lm(diff ~A*C*B, data = d2)
anova(mm1)
```

```
## Warning in anova.lm(mm1): ANOVA F-tests on an essentially perfect fit are
## unreliable
```

```
## Analysis of Variance Table
##
## Response: diff
##           Df Sum Sq Mean Sq F value Pr(>F)
## A          1   18.0    18.0     NaN    NaN
## C          1  264.5   264.5     NaN    NaN
## B          1   40.5    40.5     NaN    NaN
## A:C        1   72.0    72.0     NaN    NaN
## A:B        1   98.0    98.0     NaN    NaN
## C:B        1   24.5    24.5     NaN    NaN
## A:C:B      1   50.0    50.0     NaN    NaN
## Residuals  0    0.0     NaN
```

Since we only have one replicate, again, we will not be able to obtain useful information from the ANOVA table above. So, let's create a normal probability plot to check which factors are significant.

```r
effect <- abs(mm1$effects)[-1]
half <- qqnorm(effect, type = "n")
qqline(effect)
text(half$x, half$y, labels = names(effect), cex = 0.8, col="blue")
```



*Based on the half-normal probability plot, we could see that factor C is highly significant and interactions AB and AC could be significant as well.

```r
m1 <- lm(avg-before ~ A+C+A*B + A*C, data = d2)
anova(m1)
```

```
## Analysis of Variance Table
##
## Response: avg - before
##           Df Sum Sq Mean Sq F value Pr(>F)
## A          1   40.5   40.50  0.3208 0.6282
## C          1  968.0  968.00  7.6673 0.1094
## B          1  180.5  180.50  1.4297 0.3544
## A:B        1   50.0   50.00  0.3960 0.5934
## A:C        1  144.5  144.50  1.1446 0.3967
## Residuals  2  252.5  126.25
```

Based on the ANOVA table above, none of the factors and interactions AB and AC turned out to be significant. Thus, let's try and fit another reduced model, but with just with factor C, B, ABC.

```
m2 <- lm(avg-before ~ B+C + ABC, data =d2)
anova(m2)
```

```
## Analysis of Variance Table
##
## Response: avg - before
##           Df Sum Sq Mean Sq F value  Pr(>F)
## B          1  180.5  180.50  1.5851 0.27650
## C          1  968.0  968.00  8.5005 0.04344 *
## ABC        1   32.0   32.00  0.2810 0.62410
## Residuals  4  455.5  113.87
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again, based on the ANOVA table above, only subject (C) is significant.

Below also shows verification of the normality and homoscedasticity assumptions.

```
shapiro.test(m2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m2$residuals
## W = 0.89811, p-value = 0.2778
```

```
bartlett.test(m2$residuals, d2$B)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  m2$residuals and d2$B
## Bartlett's K-squared = 0.1024, df = 1, p-value = 0.749
```

```
bartlett.test(m2$residuals, d2$C)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  m2$residuals and d2$C
## Bartlett's K-squared = 0.30783, df = 1, p-value = 0.579
```