실험 계획 및 실습

Homework #(4)

2014-16757 김보창

---

1 Q1

6.5 y_{ijk}를 vibration이라 하면,

여기서 모형은 $y_{ijk} = \mu + A_i + B_j + AB_{ij} + \epsilon_{ijk}$, $i = 1, 2$, $j = 1, 2$, $k = 1, 2...n$, $\sum_{i=1}^{2} A_i = 0$, $\sum_{j=1}^{2} B_j = 0$, $\sum_{i=1}^{2} AB_{ij} = 0$ $\sum_{j=1}^{2} AB_{ij} = 0$ 으로 놓을 수 있다. ($\epsilon_{ijk} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$)

(effect 모형)

여기서, treatment A와 B의 level이 2가지뿐이므로, 2^2 factorial design을 이용하여 분석할 수 있다.

1.1 1-(a)

데이터를 분석하기위해, 먼저 각 effect에 따른 효과가 있는지를 알아보기 위해 다음과 같은 가설을 세워서 ANOVA test를 진행한다.

$$H_{0A} : A_i = 0, \forall i = 1, 2$$
$$H_{1A} : A_i \neq 0, \exists i$$

$$H_{0B} : B_j = 0, \forall j = 1, 2$$
$$H_{1B} : B_j \neq 0, \exists j$$

$$H_{0AB} : AB_{ij} = 0, \forall i = 1, 2, j = 1, 2$$
$$H_{1AB} : AB_{ij} \neq 0, \exists i, j$$

로 귀무가설과 대립가설을 세우고, ANOVA test를 진행하자.

이때, $SS_c = \frac{(contrast)^2}{n \sum_l c_l^2}$

이고, $contrast = \sum_l c_l S_l$ 과 같이 나타낼 수 있다.

이때, $contrast_a = (ab) - (b) + (a) - (1)$ $contrast_b = (ab) - (a) + (b) - (1)$ $contrast_{ab} = (ab) - (a) - (b) + (1)$ 임을 안다.

또한, 2^2 factorial design에서 SS_A, SS_B, SS_{AB}의 df는 모두 1임을 아므로 각각은 MS_A, MS_B, MS_{AB}와 같게 된다.

또한, 각각의 H_0하에서 $\frac{MS_c}{MS_E} \sim F_{1,4(n-1)}$임을 알 고 있으므로,

$F_0 = \frac{MS_C}{MS_E}$로 두고, 유의수준 α에서 $F_0 > F_{1,4(n-1)}(\alpha)$이면 귀무가설을 기각할 것이다.

이를 구하기 위해 F_0의 값을 구할것인데, 계산을 쉽게 하기 위해 R을 이용할 것이다.

다음 R코드를 이용하여 F_0의 값을 구한다.

```
trt_A <- as.factor(c(rep(c("-"), 4), rep(c("+"), 4), rep(c("-"), 4), rep(c("+"
    ), 4)))
trt_B <- as.factor(c(rep(c("-"), 4), rep(c("-"), 4), rep(c("+"), 4), rep(c("+"
    ), 4)))
y <- c(18.2, 18.9, 12.9, 14.4, 27.2, 24.0, 22.4, 22.5, 15.9, 14.5, 15.1, 14.2,
     41.0, 43.9, 36.3, 39.9)
df <- data.frame(trt_A, trt_B, y)
result <- aov(y ~ trt_A + trt_B + trt_A*trt_B, data = df)
summary(result)
```

실험 계획 및 실습

Homework #(4)

2014-16757 김보창

데이터를 입력해주고, trt_A, trt_B이라는 벡터에 각 데이터가 어떤 treatment에서 나왔는지 표기해준다.
그후, data.frame 함수를 이용해 data frame으로 만들어주고,
aov와 summary함수를 이용하여 결과를 출력하면 다음과 같은 값이 나온다.

```
> df <- data.frame(trt_A, trt_B, y)
> df
   trt_A trt_B    y
1      -     - 18.2
2      -     - 18.9
3      -     - 12.9
4      -     - 14.4
5      +     - 27.2
6      +     - 24.0
7      +     - 22.4
8      +     - 22.5
9      -     + 15.9
10     -     + 14.5
11     -     + 15.1
12     -     + 14.2
13     +     + 41.0
14     +     + 43.9
15     +     + 36.3
16     +     + 39.9
```

df에 저장된 형태.

```
> result <- aov(y ~ trt_A + trt_B + trt_A*trt_B, data = df)
> result
Call:
   aov(formula = y ~ trt_A + trt_B + trt_A * trt_B, data = df)

Terms:
                    trt_A     trt_B trt_A:trt_B Residuals
Sum of Squares  1107.2256  227.2556    303.6306   71.7225
Deg. of Freedom         1         1           1        12

Residual standard error: 2.444765
Estimated effects may be unbalanced
> summary(result)
            Df Sum Sq Mean Sq F value   Pr(>F)
trt_A        1 1107.2  1107.2  185.25 1.17e-08 ***
trt_B        1  227.3   227.3   38.02 4.83e-05 ***
trt_A:trt_B  1  303.6   303.6   50.80 1.20e-05 ***
Residuals   12   71.7     6.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

anova 분석 결과.

이를 통해, 여기서의 MS_E의 값은 6.0, MS_A, MS_B, MS_{AB}의 값은 각각 1107.2, 227.3, 303.6임을 알 수 있고,
각각에서 $F_{0A} = 185.25$, $F_{0B} = 38.02$, $F_{0AB} = 50.80$임을 알 수 있다.
또한 이때의 P-value가 각각 $1.17 * 10^{-8}$, $4.83 * 10^{-5}$, $1.20 * 10^{-5}$으로,
따라서, $\alpha = 0.05$로 택했을때, P-value가 0.05보다 작으므로 모든 귀무가설을 기각할 수 있다.
즉, A의 effect와 , B의 effect와, AB의 interaction effect가 모두 존재한다는 결론을 내리게 된다.

1.2 1-(b)

residual의 normal probability plot과 residual vs predicted vibration level plot을 그리기 위해,
먼저 residual과 prediected value를 다음과 같이 구한다.

실험 계획 및 실습

Homework #(4)

2014-16757 김보창

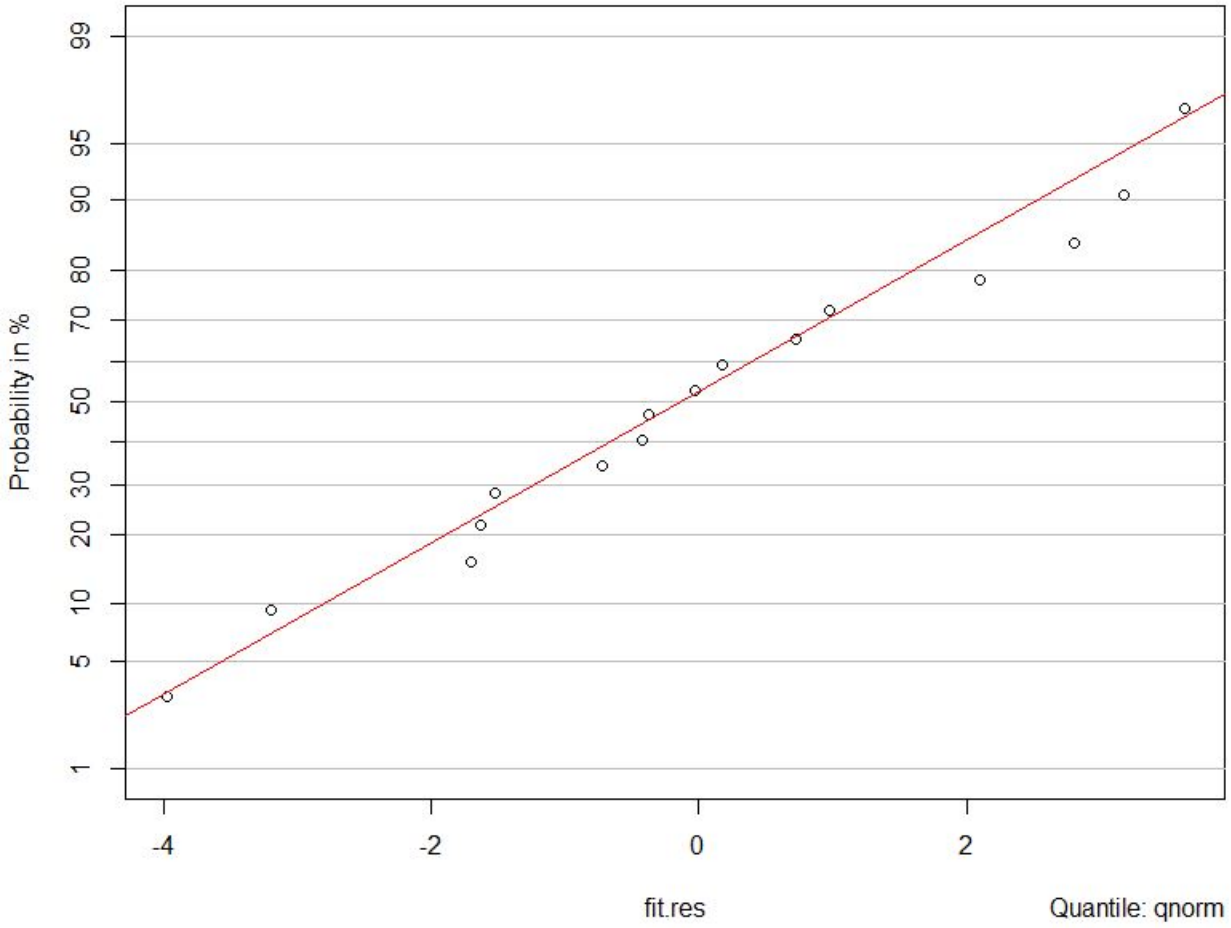
```r
library(e1071) # install.packages("e1071")

fit.lm <- lm(y ~ trt_A + trt_B + trt_A*trt_B, data = df)
aov(fit.lm)
fit.pred <- predict(fit.lm)
fit.res <- residuals(fit.lm)

probplot(fit.res)

plot(x = fit.pred, y = fit.res, xlab = "predicted vibration", ylab = "residual
    ", ylim = c(-5,5), main = "residual vs predicted vibration")
abline(h = 0)
```

probability plot을 그리기 위해 e1071 패키지를 이용하였다.

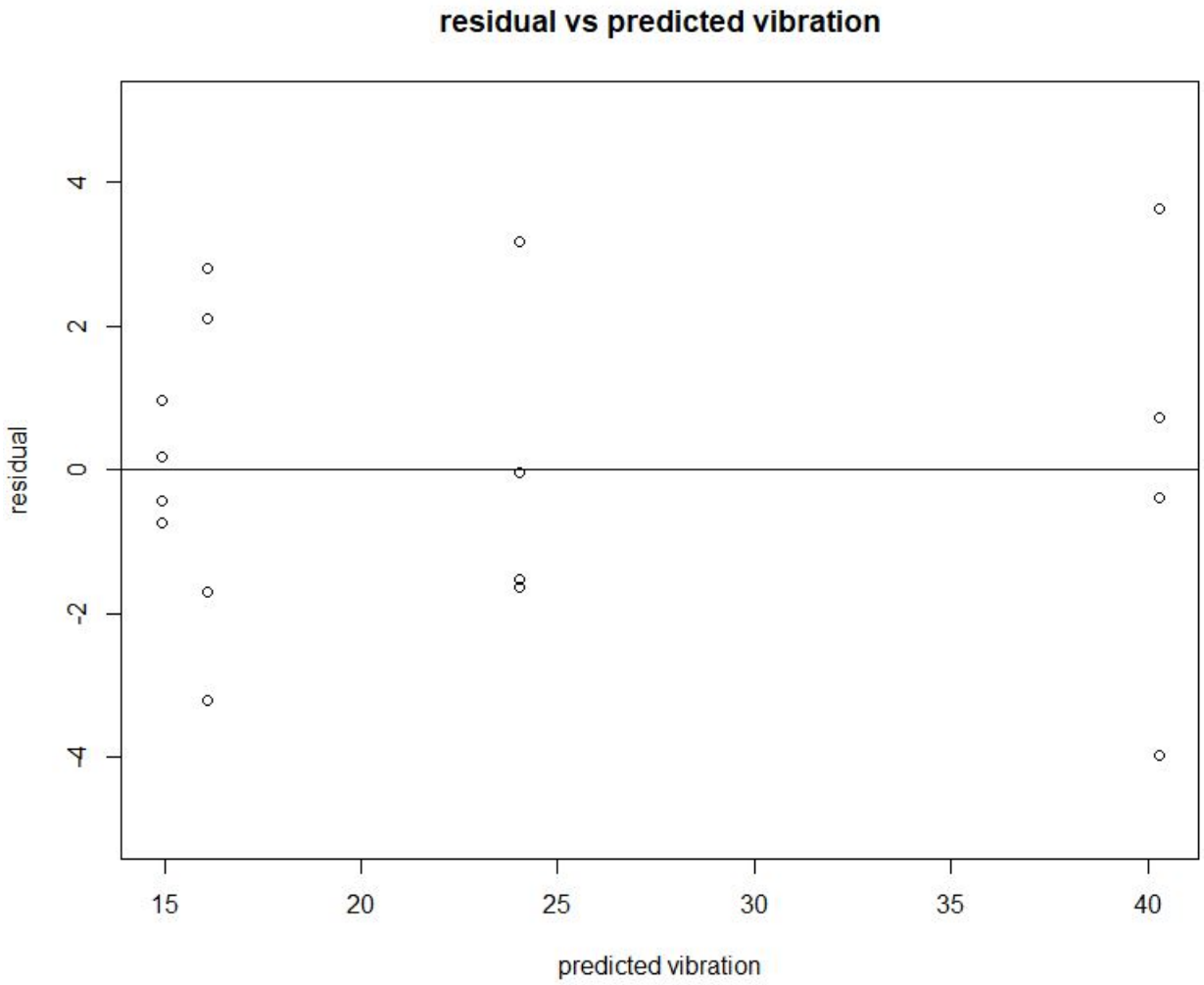
standard residual을 사용하는것이 권장되긴 하지만, (rstandard(fit.lm) 으로 사용 가능) 책에서는 residual에 대한 plot을 그릴것을 요구하였기 때문에 이를 통해 그린다.

그려진 normal probability plot과 residual vs predicted vibaration level plot은 다음과 같다.

Homework #(**4**)
2014-16757 김보창



normal probability plot



residual vs predicted value graph.

먼저, normal probablity plot을 보면, residual의 경향이 거의 직선으로 정렬된것을 볼 수 있다. 따라서, 이를 통해 error의 normality 가정을 위반하지 않음을 알 수 있다.

또한, prediected value vs residual graph를 보면, residual의 분포는 특정 경향성을 보이지 않고, predicted vibration이 14.925인 경우를 제외하고는 residual의 절대적인 크기도 모두 비슷함을 알 수 있다.

predicted vibartion이 14.925인 경우는 trt_A가 -, trt_B가 +일때인데,

이때 residual의 절대적인 크기가 작은 편이지만, 데이터의 개수가 적은 편이고, 또한 특정한 경향성은 보이지 않으므로 등분산 가정을 위반한다고 하기에는 애매하다.

결론적으로 scale에 따른 residual의 크기차이가 좀 있긴 하지만, 등분산 가정을 위반한다고 말하기에는 충분하지 않으므로 우리의 가정이 맞다고 할 수 있다.

1.3 1-(c)

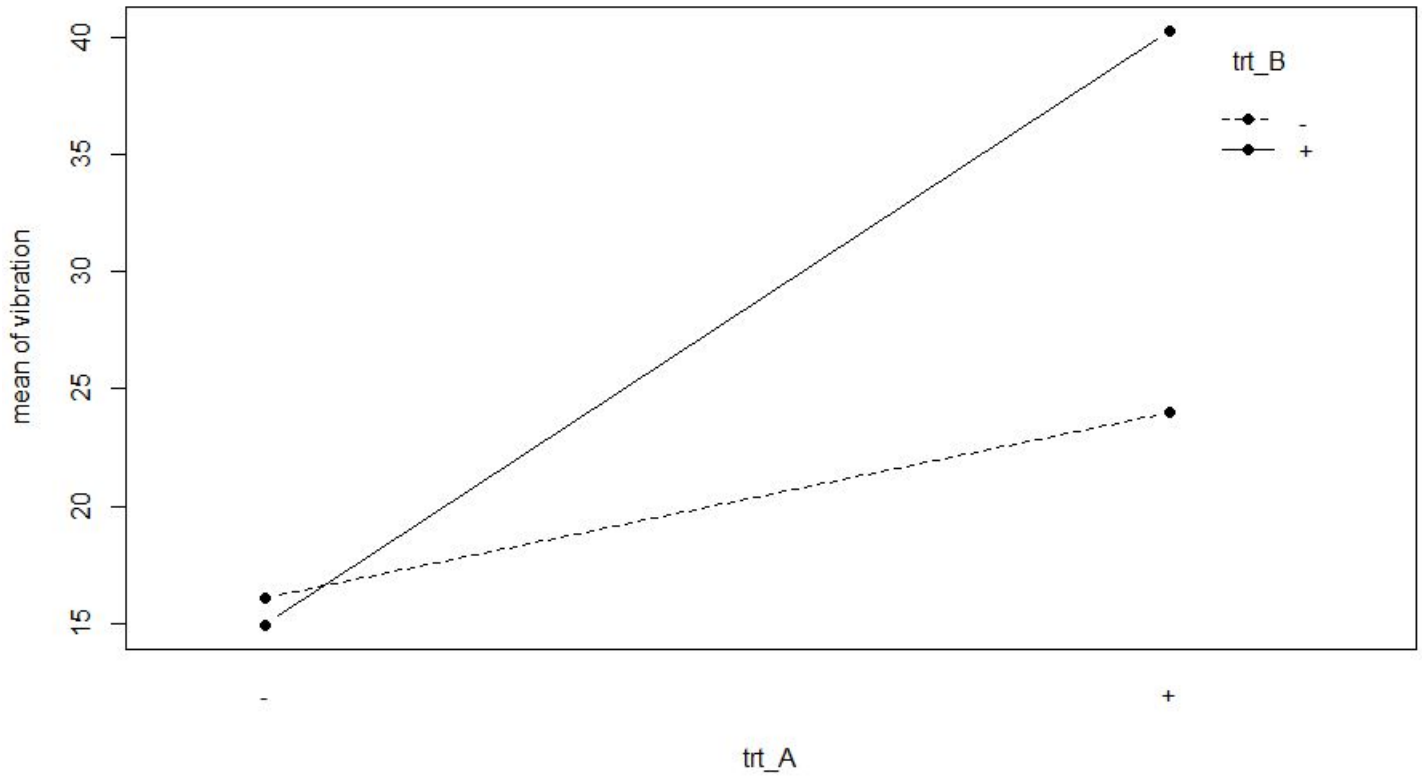
AB interaction plot은 다음과 같이 그릴 수 있다.

실험 계획 및 실습

Homework #(4)
2014-16757 김보창

```r
interaction.plot(x.factor = df$trt_A, trace.factor = df$trt_B, response = df$y
    , type = "b", fixed = TRUE, pch = 19, xlab = "trt_A", trace.label = "trt_B
    ",ylab = "mean of vibration")
```

결과는 다음과 같다.



AB interaction graph

이를 해석하면, vibration level은 A가 -에서 +로 갈때 증가하고, A가 -일때는 B가 -에서 +로 갈때 감소하고, A가 +일때는 B가 -에서 +로 갈때 증가함을 알 수 있다.

즉, AB의 interaction effect가 존재해서 A와 B가 동시에 +가 될때, 경향성이 달라짐을 알 수 있다.

이제, routine operation에서 bit size(A)와 speed(B)에 따라 vibration이 달라지는데, notch의 dimensional variation이 vibration level이 클수록 커짐을 알고 있다.

일반적인 경우 notch의 dimensional varation이 작은편이 더 안정적인 공정에 도움이 되기 때문에, vibration level을 낮게 유지하는것이 좋고 따라서 A가 -, B가 +인 경우가 더 좋다고 할 수 있다.

이때, A가 -인것은 bit size가 $\frac{1}{16}$임을 의미하고, B가 +인 경우는 speed가 90 rpm임을 의미하므로, 이러한 setting을 사용하는것이 바람직 할 것이다.

실험 계획 및 실습

Homework #(4)
2014-16757 김보창

---

2 Q2

6.24

y_{ijkl}을 메일을 받은 1000명중 실제 상품을 주문한 사람의 수라 하면, 여기서 모형은

$y_{ijkl} = \mu + A_i + B_j + C_k + AB_{ij} + BC_{jk} + AC_{ik} + ABC_{ijk} + \epsilon_{ijkl}$, $i = 1, 2$, $j = 1, 2$,$k = 1, 2$, $l = 1, 2...n$, $\sum_{i=1}^{2} A_i = 0$, $\sum_{j=1}^{2} B_j = 0$, $\sum_{i=1}^{2} AB_{ij} = 0$ $\sum_{j=1}^{2} AB_{ij} = 0$, $\sum_{j=1}^{2} BC_{jk} = 0$ $\sum_{k=1}^{2} BC_{jk} = 0$, $\sum_{i=1}^{2} AC_{ik} = 0$ $\sum_{k=1}^{2} AC_{ik} = 0$, $\sum_{i=1}^{2} ABC_{ijk} = 0$ $\sum_{j=1}^{2} ABC_{ijk} = 0$, $\sum_{k=1}^{2} ABC_{ijk} = 0$ 으로 놓을 수 있다.

($\epsilon_{ijkl} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, n = 2)

(effect 모형)

여기서, treatment A,B,C의 level이 2가지뿐이므로, 2^3 factorial design을 이용하여 분석할 수 있다.

2.1 2-(a)

데이터를 분석하기위해, 먼저 각 effect에 따른 효과가 있는지를 알아보기 위해 각각의 effect에 따라 다음과 같은 가설을 세워서 ANOVA test를 진행한다.

$$H_{0A} : A_i = 0, \forall i = 1, 2$$
$$H_{1A} : A_i \neq 0, \exists i$$

$$H_{0B} : B_j = 0, \forall j = 1, 2$$
$$H_{1B} : B_j \neq 0, \exists j$$

$$H_{0C} : C_k = 0, \forall k = 1, 2$$
$$H_{1C} : C_k \neq 0, \exists k$$

$$H_{0AB} : AB_{ij} = 0, \forall i = 1, 2, j = 1, 2$$
$$H_{1AB} : AB_{ij} \neq 0, \exists i, j$$

$$H_{0BC} : BC_{jk} = 0, \forall j = 1, 2, k = 1, 2$$
$$H_{1BC} : BC_{jk} \neq 0, \exists j, k$$

$$H_{0AC} : AC_{ik} = 0, \forall i = 1, 2, k = 1, 2$$
$$H_{1AC} : AC_{ik} \neq 0, \exists i, k$$

$$H_{0ABC} : ABC_{ijk} = 0, \forall i = 1, 2, j = 1, 2, k = 1, 2$$
$$H_{1ABC} : ABC_{ijk} \neq 0, \exists i, j, k$$

로 귀무가설과 대립가설을 세우고, ANOVA test를 진행하자.

이때, $SS_D = \frac{(contrast)^2}{n \sum_l c_l^2}$

이고, $contrast = \sum_l c_l S_l$ 과 같이 나타낼 수 있다.

이때, $contrast_a = (abc) + (ab) + (ac) + (a) - (bc) - (b) - (c) - (1)$ $contrast_b = (abc) + (ab) + (bc) + (b) - (ac) - (a) - (c) - (1)$ $contrast_c = (abc) + (ac) + (bc) + (c) - (ab) - (a) - (b) - (1)$ $contrast_{ab} = (abc) + (ab) - (ac) - (bc) - (a) - (b) + (c) + (1)$ $contrast_{bc} = (abc) - (ab) - (ac) + (bc) + (a) - (b) - (c) + (1)$ $contrast_{ac} = (abc) - (ab) + (ac) - (bc) - (a) + (b) - (c) + (1)$ $contrast_{abc} = (abc) - (ab) - (ac) - (bc) + (a) + (b) + (c) - (1)$ 임을 안다.

또한, 2^3 factorial design에서 SS_A, SS_B, SS_C, SS_{AB}, SS_{BC}, SS_{AC}, SS_{ABC}의 df는 모두 1임을 아므로 각각의 SS_D는 MS_D와 같게된다.

또한, 각각의 H_0하에서 $\frac{MS_D}{MS_E} \sim F_{1,8(n-1)}$임을 알고 있으므로,

$F_0 = \frac{MS_D}{MS_E}$로 두고, 유의수준 α에서 $F_0 > F_{1,8(n-1)}(\alpha)$이면 귀무가설을 기각할 것이다.

이를 구하기 위해 F_0의 값을 구할것인데, 계산을 쉽게 하기 위해 R을 이용할 것이다.

다음 R코드를 이용하여 F_0의 값을 구한다.

```
trt_A <- as.factor(c(rep(c("-"), 2), rep(c("+"), 2), rep(c("-"), 2), rep(c("+"), 2),rep(c("-"), 2), rep(c("+"), 2), rep(c("-"), 2), rep(c("+"), 2)))
trt_B <- as.factor(c(rep(c("-"), 4), rep(c("+"), 4), rep(c("-"), 4), rep(c("+"), 4)))
trt_C <- as.factor(c(rep(c("-"), 8), rep(c("+"), 8)))
y <- c(50, 54, 44, 42, 46, 48 , 42, 43, 49, 46, 48, 45, 47, 48 ,56, 54)
df2 <- data.frame(trt_A, trt_B, trt_C, y)
result2 <- aov(y ~ trt_A + trt_B + trt_C + trt_A*trt_B + trt_B * trt_C + trt_A * trt_C + trt_A * trt_B * trt_C, data = df2)
summary(result2)
```

데이터를 입력해주고, trt_A, trt_B이라는 벡터에 각 데이터가 어떤 treatment에서 나왔는지 표기해준다.

그후, data.frame 함수를 이용해 data frame으로 만들어주고,

aov와 summary함수를 이용하여 결과를 출력하면 다음과 같은 값이 나온다.

```
> df2 <- data.frame(trt_A, trt_B, trt_C, y)
> df2
   trt_A trt_B trt_C  y
1      -     -     - 50
2      -     -     - 54
3      +     -     - 44
4      +     -     - 42
5      -     +     - 46
6      -     +     - 48
7      +     +     - 42
8      +     +     - 43
9      -     -     + 49
10     -     -     + 46
11     +     -     + 48
12     +     -     + 45
13     -     +     + 47
14     -     +     + 48
15     +     +     + 56
16     +     +     + 54
> |
```

df에 저장된 형태.

```
> summary(result2)
                  Df Sum Sq Mean Sq F value   Pr(>F)
trt_A              1  12.25   12.25   4.083 0.077971 .
trt_B              1   2.25    2.25   0.750 0.411694
trt_C              1  36.00   36.00  12.000 0.008516 **
trt_A:trt_B        1  42.25   42.25  14.083 0.005602 **
trt_B:trt_C        1  49.00   49.00  16.333 0.003728 **
trt_A:trt_C        1 100.00  100.00  33.333 0.000418 ***
trt_A:trt_B:trt_C  1   4.00    4.00   1.333 0.281537
Residuals          8  24.00    3.00
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

anova 분석 결과.

이를 통해, 각각의 MS_D의 값과 (Mean Sq 부분), 각각의 F value를 확인할 수 있다. (F value). 또한, 각 effect의 P-value역시 확인할 수 있다.

따라서, $\alpha = 0.05$로 택했을때, P-value가 0.05보다 작은 경우는,

C의 effect, A와 B의 interaction effect, A와 C의 interaction effect, B와 C의 interaction effect를 test하는 경우이므로,

즉, H_{0C}, H_{0AB}, H_{0AC} , H_{0BC}를 기각할 수 있고,

따라서 customer의 reponse rate는 1000명중 실제 주문을 한 사람의 비율이므로, 이러한 비율에 크게 영향을 주는 factor는 위와 같음을 알 수 있다. 즉, C의 effect, AB의 interaction effect, AC의 interaction effect, BC의 interaction effect가 존재한다고 말할 수 있다.

2.2 2-(b)

residual에 대한 분석을 하기 위해, 실습시간에 배운 R함수를 이용한다.

```
opar <- par(mfrow=c(2,2),cex=.8)
plot(aov(y ~ trt_A + trt_B + trt_C + trt_A*trt_B + trt_B * trt_C + trt_A * trt
    _C + trt_A * trt_B * trt_C, data = df2))
```

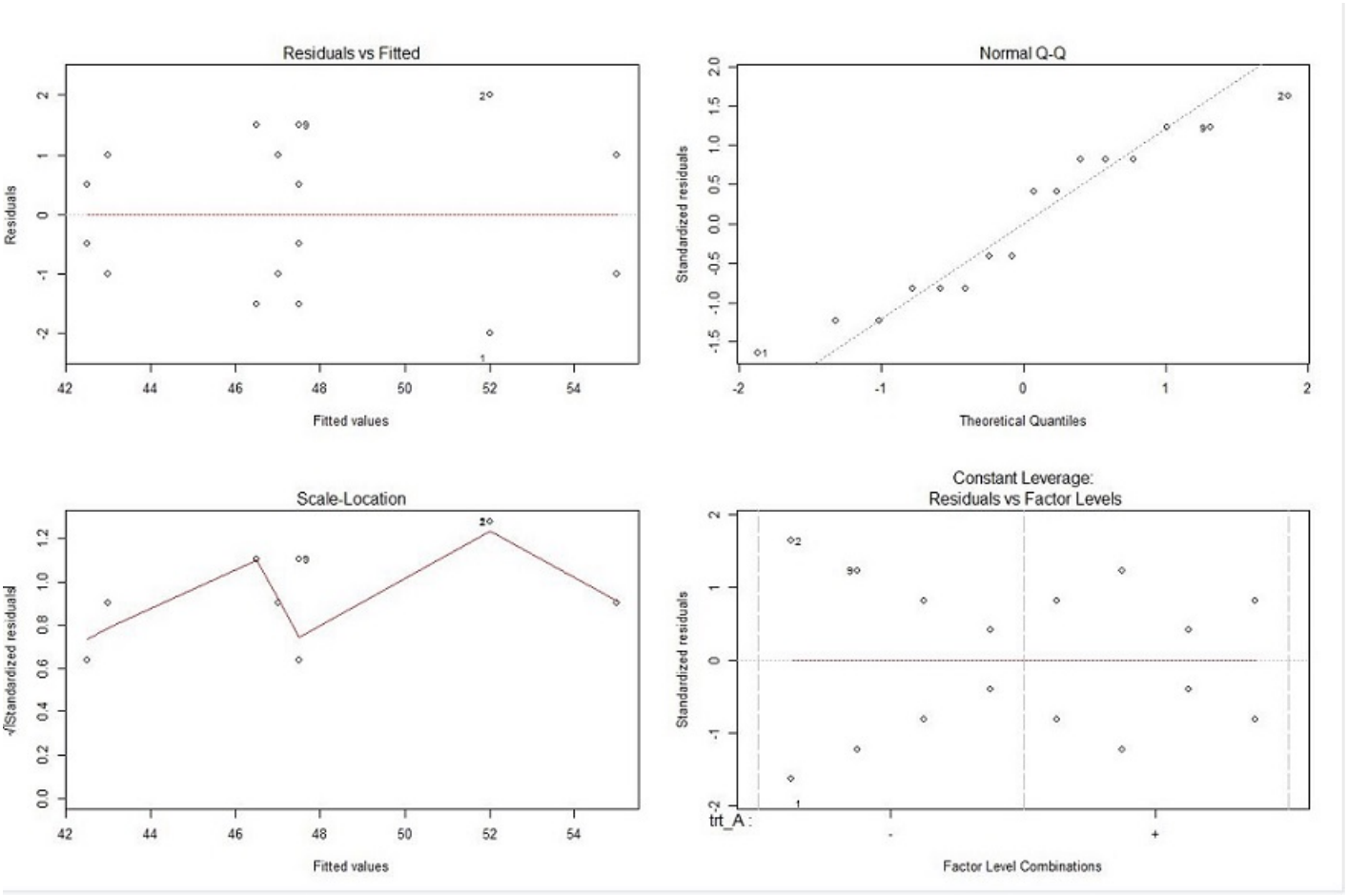
첫줄을 통해 그래프를 출력할 환경을 지정해주고, 두번째줄 plot을 이용하여 그래프를 출력하게 하였다.

결과는 다음과 같다.

실험 계획 및 실습

Homework #(4)

2014-16757 김보창



오른쪽 위 그래프, normal QQ 그래프를 살펴보면, residual의 경향이 비교적 직선과 비슷하게 정렬된것을 보아 normality 가정에 큰 문제가 없음을 알 수 있다.

또한, 왼쪽 위 그래프를 보면, fitted value와 residual로 그래프를 그렸을때, 특정 경향성이 나타나지 않는것을 알 수 있고, 따라서 우리 모형이 데이터를 잘 표현한다고 말할 수 있다. 또한, 각 fitted value에 따른 residual의 크기가 비슷한 편이므로, 등분산 가정 역시 알맞다고 할 수 있다.

하지만, 오른쪽 아래 그래프, factor level combination에 따른 residual의 경향을 보면, residual의 경향성이 없어야 하지만 약간의 경향성이 보이는것을 알 수 있고, 이점에서 model inadequacy가 약간 있다고 생각할 수 있을것이다.

결론적으로, 독립, 등분산, 정규분포 가정이 알맞다고 할 수 있어

따라서 $\epsilon_{ijkl} \sim N(0, \sigma^2)$이라는 우리의 가정에 문제가 없음을 알 수 있지만,

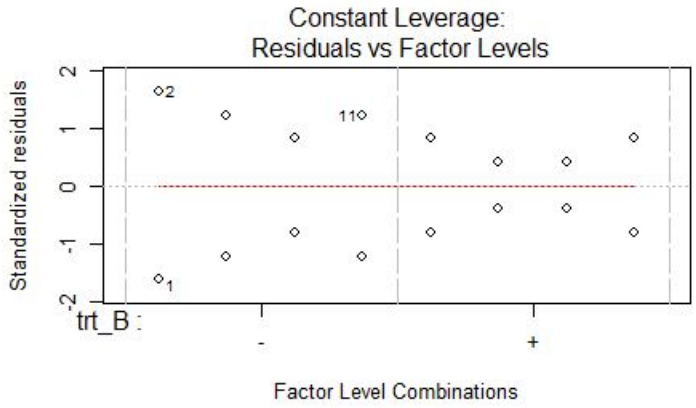
약간의 model inadequacy가 존재한다고 말할 수 있다.

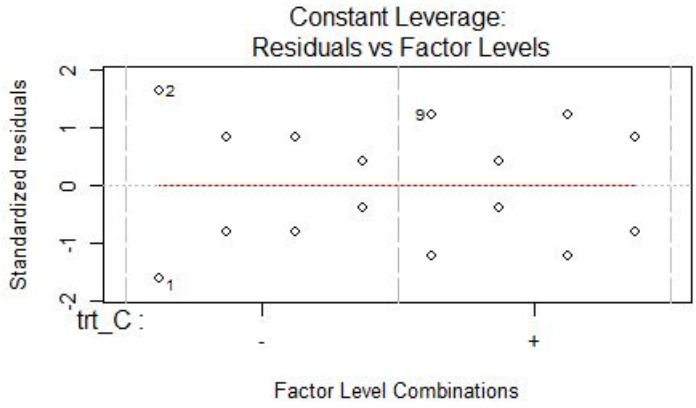
위에서 출력한 그래프는 treatment A에 의한 residual plot만 그리고 있는데, (우측 아래) treatment B, treatment C에 의한 residual plot은 다음과 같이 그릴 수 있다.

```
opar <- par(mfrow=c(2,2),cex=.8)
plot(aov(y ~ trt_B + trt_A + trt_C + trt_A*trt_B + trt_B * trt_C + trt_A * trt
    _C + trt_A * trt_B * trt_C, data = df2))
```

```
opar <- par(mfrow=c(2,2),cex=.8)
plot(aov(y ~ trt_C + trt_A + trt_B + trt_A*trt_B + trt_B * trt_C + trt_A * trt
    _C + trt_A * trt_B * trt_C, data = df2))
```

각각의 우측 아래 그래프만 따온것은 아래와 같다.





2.3 2-(c)

분석 결과에 따르면, 결국 number of order가 많은 쪽이 회사에 이득이 되는 부분이고, 따라서 최대한 number of order가 많은쪽을 택하면 될것이다.

위의 anova 모델을 사용하여 각각의 A,B,C combination에 대해 fitted 된 결과를 통해, prediction 된 값을 뽑아보자.

```
fit2.lm <- lm(y ~ trt_A + trt_B + trt_C + trt_A*trt_B + trt_B * trt_C + trt_A
    * trt_C + trt_A * trt_B * trt_C, data = df2)
predict(fit2.lm)
```

```
> predict(fit2.lm)
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
52.0 52.0 43.0 43.0 47.0 47.0 42.5 42.5 47.5 47.5 46.5 46.5 47.5 47.5 55.0 55.0
```

predicted value.

결과적으로, 15, 16번째의 predicted value가 가장 큰것을 알 수 있다.
이 부분은 A,B,C가 모두 적용되었을때의 predicted value, 즉, A가 +, B가 +, C가 +일때의 값이므로,
다시말해 1st class의 mail을 사용하고(A+), Color가 있는 mail을 사용하고, (B+), offered price가 24.95 달러일때 (C+) 기대되는 number of order가 가장 많다는 뜻이다.
따라서, 회사에는 이와 같은 조합을 사용할 것을 추천할 것이다.