실험계획법 과제 1

4월 16일 수요일 23시 59분까지 제출

1. 세 개의 회사(A, B, C)에서 생산된 건전지에 수명의 차이가 있는지 알아보고자 한다. 완전확률화설계에 의해 각 회사에서 6개씩 건전지를 선택하여 수명실험을 하여 다음의 데이터를 얻었고 이들은 정규분포를 따른다고 한다.

A	B	C
100	76	108
96	80	100
98	84	101
96	84	98
92	78	102
95	82	100

위의 데이터를 R을 이용해 만드는 과정은 아래와 같다.

```
y = c(100, 96, 98, 96, 92, 95, 76, 80, 84, 84, 78, 82, 108, 100, 101, 98, 102, 100)
company = rep(c("A", "B", "C"), each = 6)
data = data.frame(company, y)
head(data)
```

```
##   company   y
## 1       A 100
## 2       A  96
## 3       A  98
## 4       A  96
## 5       A  92
## 6       A  95
```

1. 실험설계에 맞는 통계적 모형을 쓰시오.

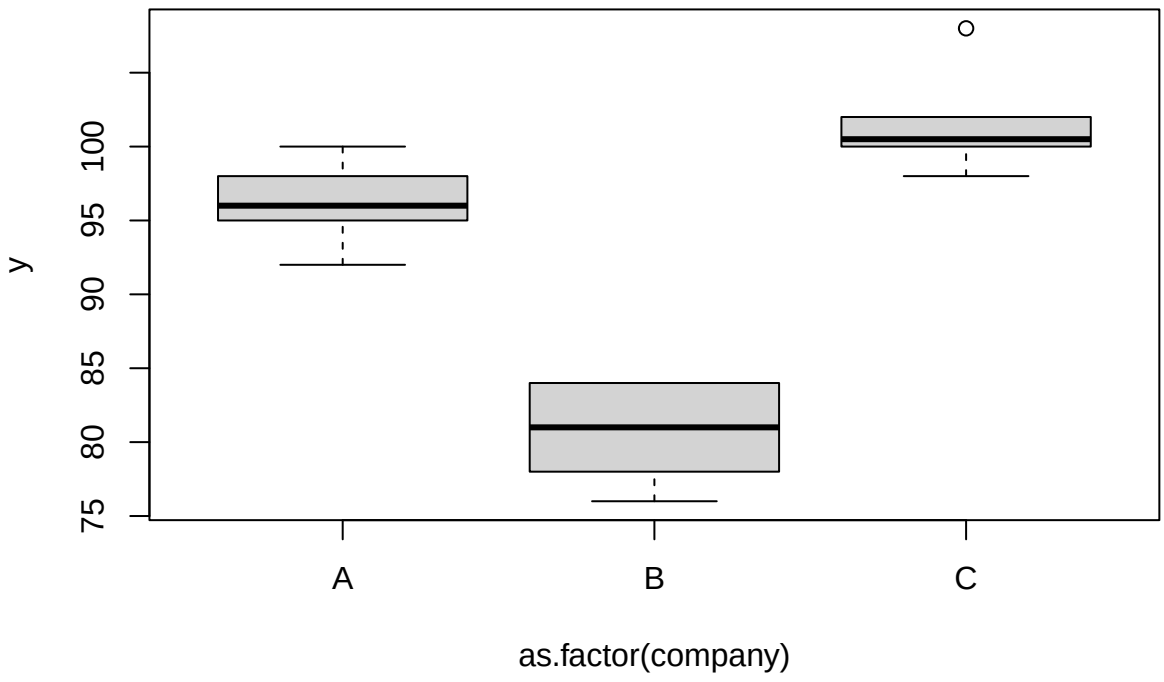
반응변수 Y_{ij}는 i번째 건전지 회사의 j번째 건전지의 수명을 나타내며, 다음과 같이 완전확률화실험에 대한 모형을 설정한다.

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad i = 1, 2, 3, \quad j = 1, 2, 3, 4, 5, 6$$

이 때, μ는 모집단 전체 평균, τ_i는 i번째 처리효과(건전지 회사에 따른 효과), ϵ_{ij}는 i번째 처리의 j번째 측정값에 대한 오차항이며 $\epsilon \overset{i.i.d.}{\sim} N(0, \sigma^2)$으로 표현. 즉 오차항에 대해서는 독립이고, 등분산성을 가지는 정규분포에서 나왔다고 가정.

2. 각 회사에 따라 건전지 수명이 다른지 확인하기 위해 상자그림과 각 회사에 따른 평균과 표준편차를 구하고, 결과를 해석하시오.

```
boxplot(y~as.factor(company))
```




```
tapply(y, company, mean);
```

```
##         A         B         C
##  96.16667  80.66667 101.50000
```

```
tapply(y, company, sd);
```

```
##        A        B        C
## 2.714160 3.265986 3.449638
```

위 상자그림과 건전지 회사에 따라 수명을 확인해보면, A회사와 C회사의 건전지 수명에 비해 B회사의 건전지 수명이 대체로 짧아보인다. 실제로 각 회사의 평균은 96.16667, 80.66667, 101.50000 로 주어져 B회사의 수명이 다른 회사의 건전지 수명에 비해 짧음이 확인된다. 표준편차의 경우 평균에 비해 크지 않은 값이고, 이는 각 회사들의 수명이 평균에 가까이 분포해있다는 뜻이다. 즉, 대체로 수명이 회사에 따라 다른 값을 나타냄을 확인할 수 있다.

3. 각 회사에 따라 건전지 수명이 다른지 확인하기 위한 적절한 통계적 가설을 설정하고 (1번 문제에서 정의한 통계적 모형을 이용) 분산분석표를 작성한 후, 설정한 가설을 검정하시오.

1번 문제에서 정의한 표현을 사용하면 건전지의 수명이 회사에 따라 다르다는 말은 τ_i들이 서로 다르다는 뜻이며 0이 아니라는 말이다. 따라서

$$H_0 : \tau_1 = \tau_2 = \tau_3 = 0, \quad vs. \quad H_1 : \text{ not } H_0$$

즉, 이때 대립가설은 적어도 한 회사의 건전지 수명은 다른 회사의 건전지 수명과 다르다는 형태로 표현될 수 있다.

표현한 가설을 검정하기 위해 분산분석을 실행하고 분산분석표를 얻어보면 아래와 같다.

```
aov = aov(y~as.factor(company))
summary(aov)
```

```
##                    Df Sum Sq Mean Sq F value   Pr(>F)
## as.factor(company)  2 1405.4   702.7   70.43 2.37e-08 ***
## Residuals          15  149.7    10.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

위의 분산분석표를 이용해 유의확률이 매우 작은 값을 나타냄을 알 수 있으므로, 귀무가설을 기각할 수 있다. 따라서 3개의 건전지 종류간 평균이 모두 동일하다고 할 수 없다.

4. R에서 `pairwise.t.test()` 함수를 이용하여 가능한 A, B, C 회사에 따른 건전지 수명을 비교하시오.

`pairwise.t.test()` 함수를 사용하여 모든 경우에 대한 비교를 해보면, $\tau_1 = \tau_2$, $\tau_1 = \tau_3$, $\tau_2 = \tau_3$에 대한 검정을 진행할 수 있다. 그리고 이는 $\mu_1 = \mu_2$, $\mu_1 = \mu_3$, $\mu_2 = \mu_3$에 대한 검정으로 표현된다. 이 때, 오차항에 대한 가정으로 모든 분산이 동일하다고 했으므로, option에서 `pool.sd = T`를 사용한다. (실제로는 바틀렛 검정을 통해 등분산성을 만족하는지 확인하여 등분산성을 만족한다면 같은 분산을 나타내는 공통분산(pooled standard deviation 사용))

```r
bartlett.test(y~as.factor(company))
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  y by as.factor(company)
## Bartlett's K-squared = 0.27759, df = 2, p-value = 0.8704
```

```r
pairwise.t.test(y, as.factor(company), pool.sd=T, p.adjust="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  y and as.factor(company)
##
##   A       B
## B 4.1e-07 -
## C 0.01    8.4e-09
##
## P value adjustment method: none
```

위 결과를 통해 모든 경우의 귀무가설 $\mu_1 = \mu_2$, $\mu_1 = \mu_3$, $\mu_2 = \mu_3$을 기각할 수 있음을 알 수 있다. 즉 세 회사의 건전지 수명이 모두 다르다고 판단할 수 있다. 또한 Bonferroni 방법으로 p-value를 보정해준 경우(여기서는 원래의 p-value의 3배로 얻어짐.)에도 모든 귀무가설이 기각됨을 알 수 있다. 따라서 세 회사의 건전지 수명이 모두 다르다고 판단할 수 있다.

```r
pairwise.t.test(y, as.factor(company), pool.sd=T, p.adjust="bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  y and as.factor(company)
##
##   A       B
## B 1.2e-06 -
## C 0.031   2.5e-08
##
## P value adjustment method: bonferroni
```

5. 잔차 $\hat{\epsilon}_{ij} = Y_{ij} - \hat{Y}_{ij}$를 계산하여 출력하고, x-축에 $\hat{Y}_{ij}$, y-축에 $\hat{\epsilon}_{ij}$를 사용한 산점도를 그리고 결과를 해석하시오. 또한 그 외의 방법을 이용하여 오차항에 대한 가정을 확인하시오.

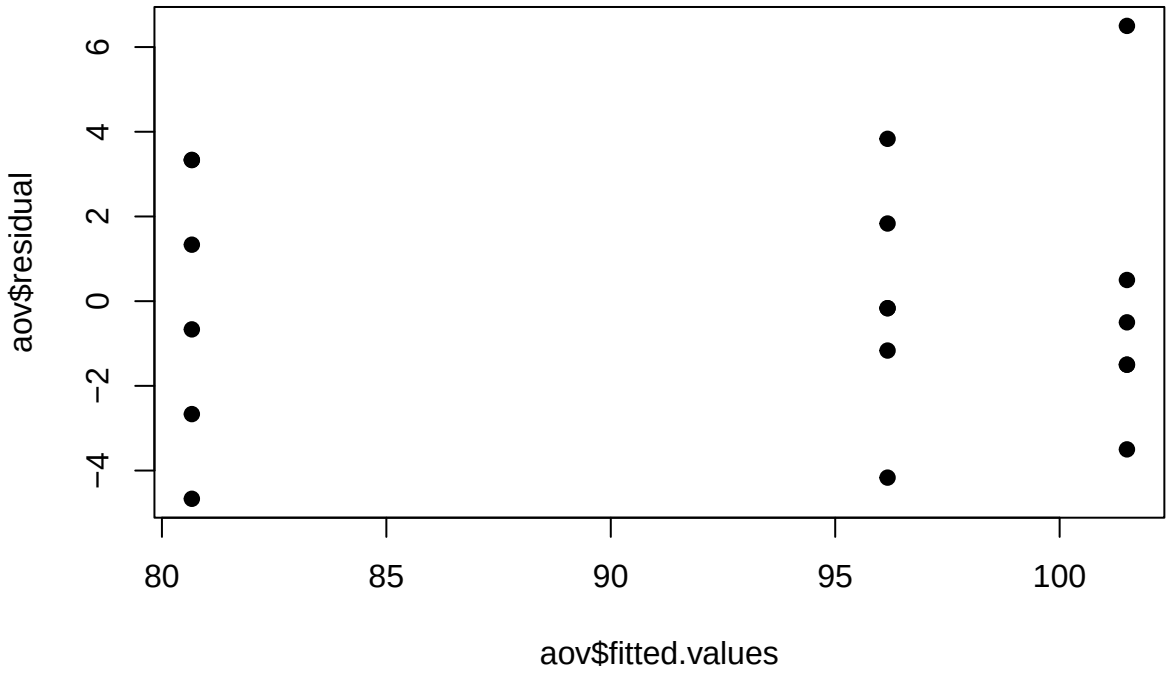
잔차는 분산분석을 실행한 결과에 residual로 저장되어 있으며 이를 다음과 같이 출력할 수 있다.

```
aov$residual
```

```
##          1          2          3          4          5          6          7
##  3.8333333 -0.1666667  1.8333333 -0.1666667 -4.1666667 -1.1666667 -4.6666667
##          8          9         10         11         12         13         14
## -0.6666667  3.3333333  3.3333333 -2.6666667  1.3333333  6.5000000 -1.5000000
##         15         16         17         18
## -0.5000000 -3.5000000  0.5000000 -1.5000000
```

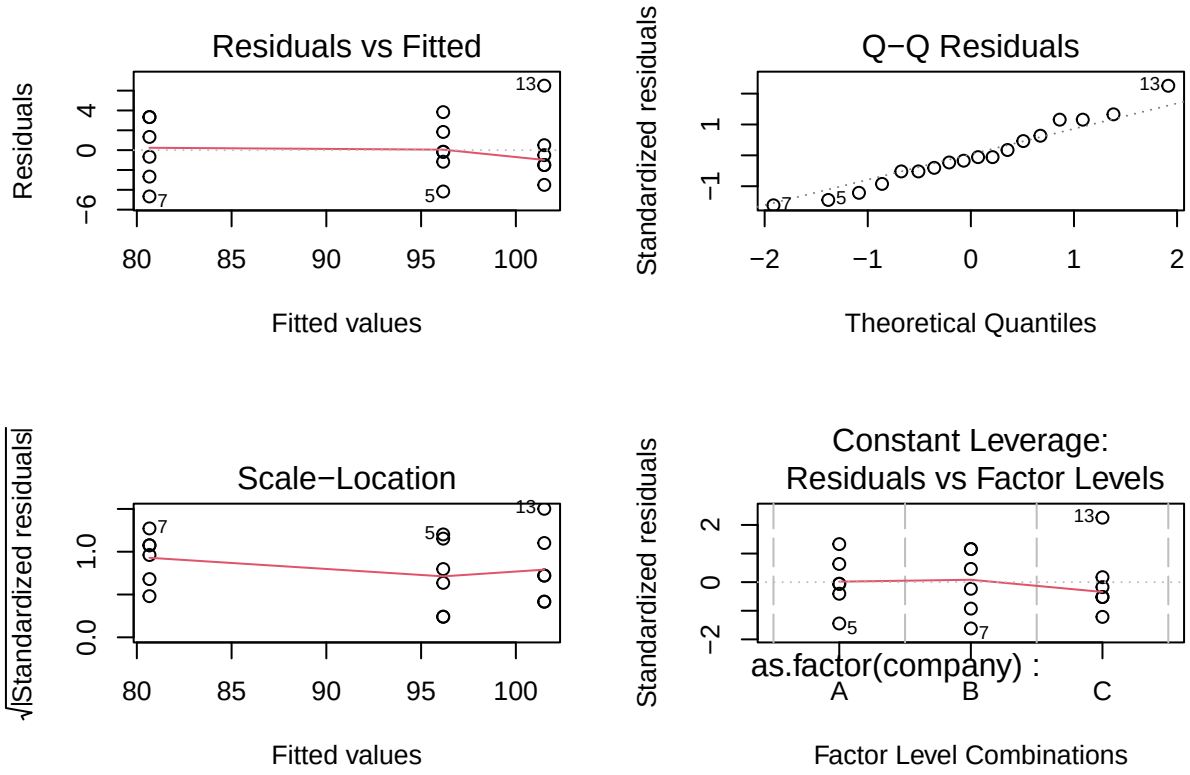
또한 적합값을 나태는 $\hat{Y}_{ij}$의 경우는 분산분석 결과에서 fitted.values로 저장되어 있다. 이를 이용하여 산점도를 그리면,

```
plot(x = aov$fitted.values, y = aov$residual, pch=19)
```



그외의 방법을 이용하기 위해 분산분석에서 제공하는 QQplot과 잔차와 관련한 그림들을 확인해보면,

```
par(mfrow=c(2,2))
plot(aov)
```




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

이 때, qqplot의 경우 대체로 정규성을 만족하는 것처럼 보인다. 세 번째 그림에서 표준화 잔차 역시 0을 주변으로 특별한 패턴없이 0과 2사이에 위치하고 있음을 알 수 있다. 또한 마지막 그림을 통해 각 회사별로 얻을 수 있는 데이터들이 퍼져있는 정도가 대체로 비슷해보이고 특별한 패턴도 없음을 알 수 있다. 또한 첫 번째 잔차에 대한 그림에서 전체적으로도 특별한 패턴이 보이지 않으므로 독립성을 가정하기에 무리가 없어보인다. 즉 오차항에 대해서 독립이고 동일한 정규분포를 따른다는 가정에서 크게 벗어나 보이지 않음을 알 수 있다.

6. $H_0 : \frac{\mu_A+\mu_B}{2} = \mu_C$에 대해 $H_0 : \frac{\mu_A+\mu_B}{2} \neq \mu_C$를 유의수준 0.05에서 검정하시오.

위의 가설을 다시표현하면, $H_0 : \mu_A + \mu_B - 2\mu_c = 0$이며 계수값들의 합이 1+1+(-2)=0이므로 이를 대비를 이용해 표현한 경우로 볼 수 있다. 즉 이를 이용해서, 대비를 검정하는 방법을 적용하면,

$$t_L^2 = \left(\frac{\sum_{i=1}^{k} c_i \bar{Y}_{i\cdot}}{\sqrt{MSe \sum_{i=1}^{k} \frac{c_i^2}{n_i}}} \right)^2 \overset{H_0}{\sim} F_{1,N-k}$$

를 이용하여 검정을 진행할 수 있다. 구체적으로는 위의 통계량을 이용하여 유의확률을 계산할 수 있으며, 자유도 1, N-3을 가지는 F분포에서 t_L^2보다 큰 영역을 구하면 된다. 즉,

$$P\left(F_{1,N-k} > t_l^2\right) = p - value$$

```
N = length(y)
k = 3
m = tapply(y, as.factor(company), mean)
ni = tapply(y, as.factor(company), length)
c1 = c(1,1,-2)
```

```r
summ.aov = summary(aov)
mse = summ.aov[[1]]$`Mean Sq`[2]
f = sum(c1*m)^2 / (mse*sum(c1^2/ni))
1 - pf(f, 1, N-k)
```

```
## [1] 5.590339e-07
```

유의확률이 매우 작은값으로 얻어졌으므로 첫 번째 회사와 두 번째 회사의 평균이 세번째 회사의 건전지 수명과 다르다고 판단할 수 있다.

2. 다음은 24명에 대해서 얻은 같은 종류의 라면에 대한 100점 만점으로 계산한 시식 평가이다.

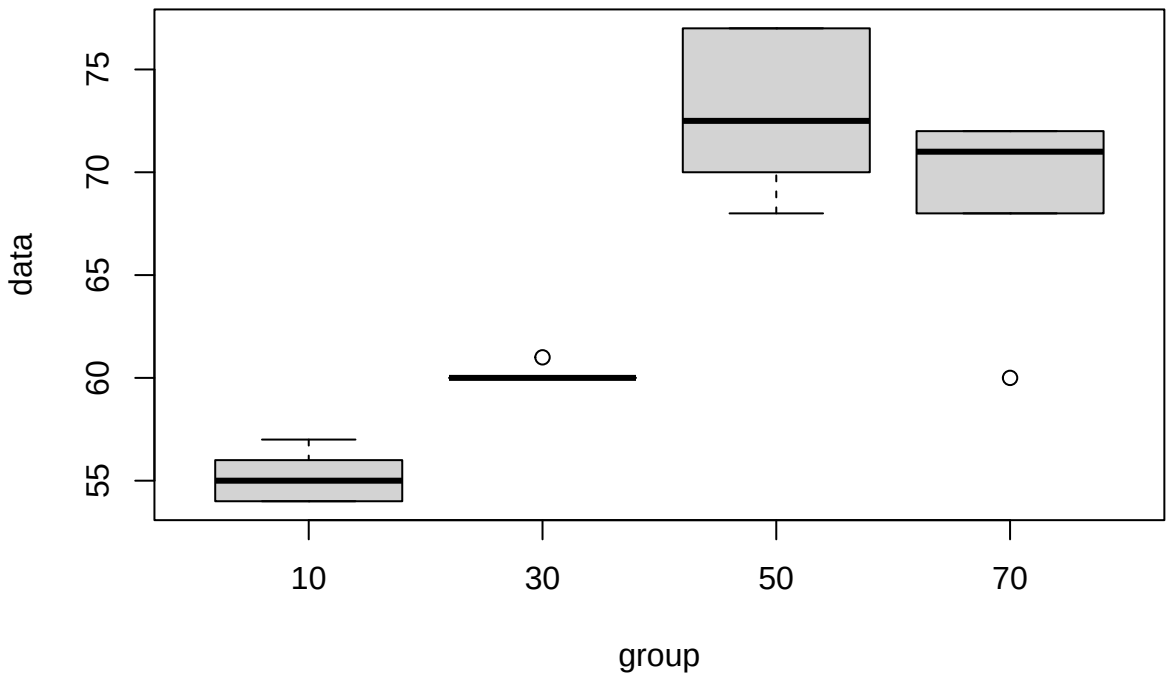
```r
data = c(55, 55, 57, 54, 54, 56, 60, 61, 60, 60, 60, 60,
         70, 72, 73, 68, 77, 77, 72, 72, 72, 70, 68, 60)
```

추가적인 조사를 통해 위의 24개 맛 평가 점수가 4개의 연령대로부터 얻어진 완전확률화설계로부터 얻어진 데이터임을 알게되었다. 각 맛평가 점수와 연령대는 다음과 같다.

10	30	50	70
55	60	70	72
55	61	72	72
57	60	73	72
54	60	68	70
54	60	77	68
56	60	77	60

(1) 각 연령에 따른 맛 평가 점수의 상자그림을 그리고 연령에 따라 라면 평가 점수에 차이가 보이는지 확인하여라.

```r
data = c(55, 55, 57, 54, 54, 56, 60, 61, 60, 60, 60, 60,
         70, 72, 73, 68, 77, 77, 72, 72, 72, 70, 68, 60)
group = rep(c("10","30","50","70"),each=6)
group = as.factor(group)
plot(data ~ group)
```



(2) 실험설계에 맞는 통계적 모형을 쓰고 연령이 라면 맛 평가에 영향을 주는지 분산분석표를 작성한 후 유의수준 0.05에서 확인하여라.

반응변수 Y_{ij}는 i번째 연령대의 j번째 라면 맛 점수를 나타내며, 다음과 같이 완전확률화실험에 대한 모형을 설정한다.

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad i = 1,2,3,4, \quad j = 1,2,3,4,5,6$$

이 때, μ는 모집단 전체 평균, τ_i는 i번째 처리효과(건전지 회사에 따른 효과), ϵ_{ij}는 i번째 처리의 j번째 측정값에 대한 오차항이며 $\epsilon \overset{i.i.d.}{\sim} N(0,\sigma^2)$으로 표현. 즉 오차항에 대해서는 독립이고, 등분산성을 가지는 정규분포에서 나왔다고 가정.

```
aov1 = aov(data~group)
summary(aov1)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## group        3 1172.5   390.8   42.37 7.48e-09 ***
## Residuals   20  184.5     9.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(3) `model.tables()`함수를 사용하여 얻은 연령에 따른 효과 차이 결과가 `summary.lm()`함수를 사용하여 얻은 연령에 따른 평균 평가점수로 어떻게 표현되는지 설명하여라.

`model.tables()`를 통해 출력되는 결과는 아래와 같다.

```
model.tables(aov1)
```

```
## Tables of effects
```

```
##
##  group
## group
##     10     30     50     70
## -9.125 -4.125  8.542  4.708
```

즉 Tables of effects에서 group 10의 -9.125는 $\hat{\tau}_{10}$을 의미한다. 정리하면

$$\hat{\tau}_{10s} = -9.125, \quad \hat{\tau}_{30s} = -4.125, \quad \hat{\tau}_{50s} = 8.542, \quad \hat{\tau}_{70s} = 4.708$$

summary.lm을 통해 출력되는 결과는 아래와 같다.

```
summary.lm(aov1)
```

```
##
## Call:
## aov(formula = data ~ group)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0000 -0.8750 -0.1667  1.2083  4.1667
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   55.167      1.240  44.491  < 2e-16 ***
## group30        5.000      1.754   2.851  0.00987 **
## group50       17.667      1.754  10.075 2.79e-09 ***
## group70       13.833      1.754   7.889 1.45e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.037 on 20 degrees of freedom
## Multiple R-squared:  0.864,  Adjusted R-squared:  0.8436
## F-statistic: 42.37 on 3 and 20 DF,  p-value: 7.479e-09
```

분산분석 결과를 summary.lm을 통해 출력하게 되면, 다음과 같은 선형모형을 다룬 것으로 볼 수 있다.

$\hat{Y}_{ij} = \beta_0 + \beta_{30s} \times I(30s) + \beta_{50s} \times I(50s) + \beta_{70s} \times I(70s) = 55.167 + 5.000 \times I(30s) + 17.667 \times I(50s) + 13.833 \times I(70s)$

10대의 경우는 $I(30s) = I(50s) = I(70s) = 0$이며, 따라서 $\hat{\mu}_{10s} = 55.167$. 이를 이용하면 다음과 같이 표현된다.

$$\hat{\mu}_{30s} = 55.167 + 5.000 = 60.167$$
$$\hat{\mu}_{50s} = 55.167 + 17.667 = 72.834$$
$$\hat{\mu}_{70s} = 55.167 + 13.833 = 69.000$$

또한 models.tables의 결과와 비교하면,

$$\hat{\tau}_{10s} - \hat{\tau}_{30s} = (\hat{\mu}_{10s} - \hat{\mu}) - (\hat{\mu}_{30s} - \hat{\mu}) = \hat{\mu}_{10s} - \hat{\mu}_{30s}$$

즉

$$-9.125 - (-4.125) = 0 - (5.000)$$

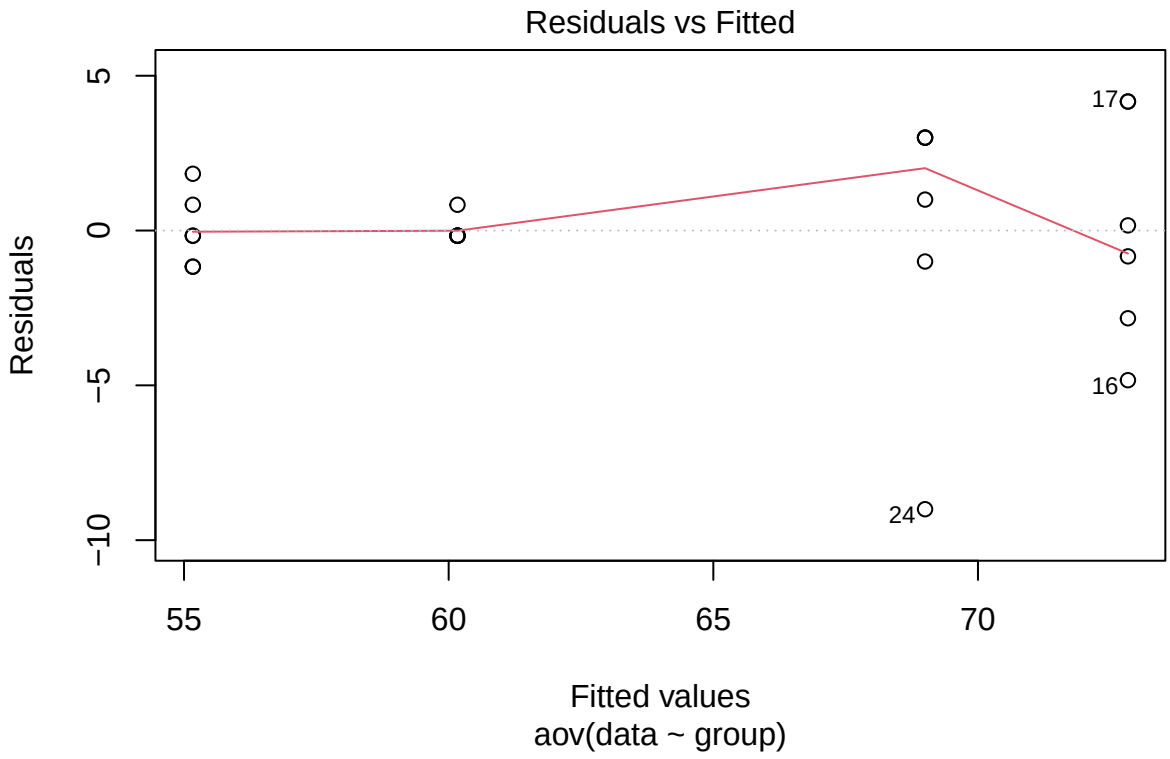
같은 방법으로 다른 연령대에 대한 비교도 가능하다.

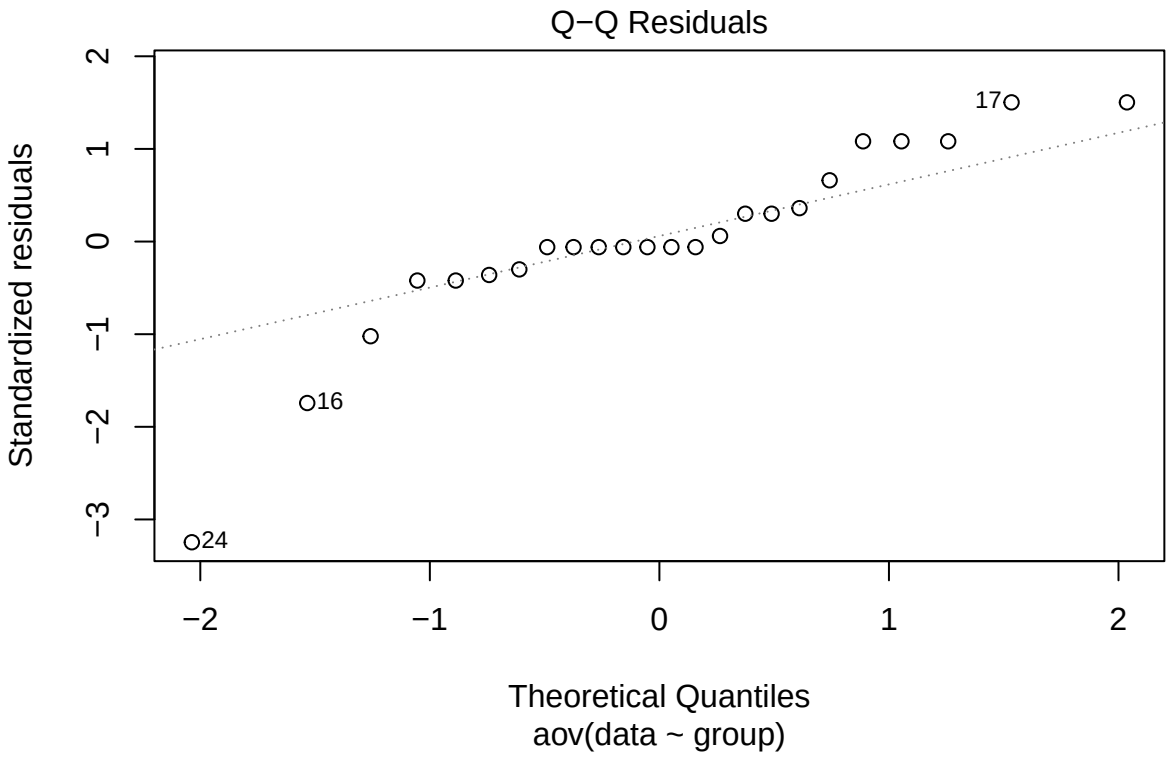
(4) R을 이용하여 잔차를 계산하여 출력하고, 데이터가 독립적이고 동일한 정규분포를 따른다는 가정을 확인하여라.

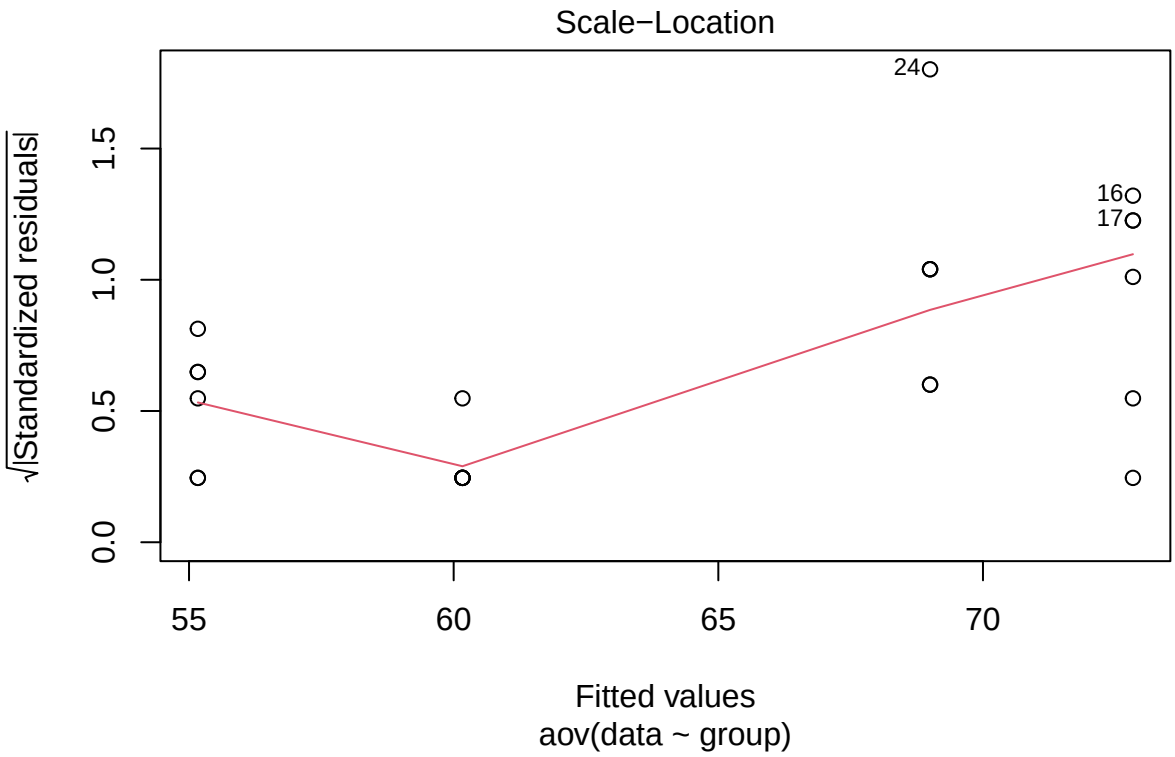
```
aov1$residual
```

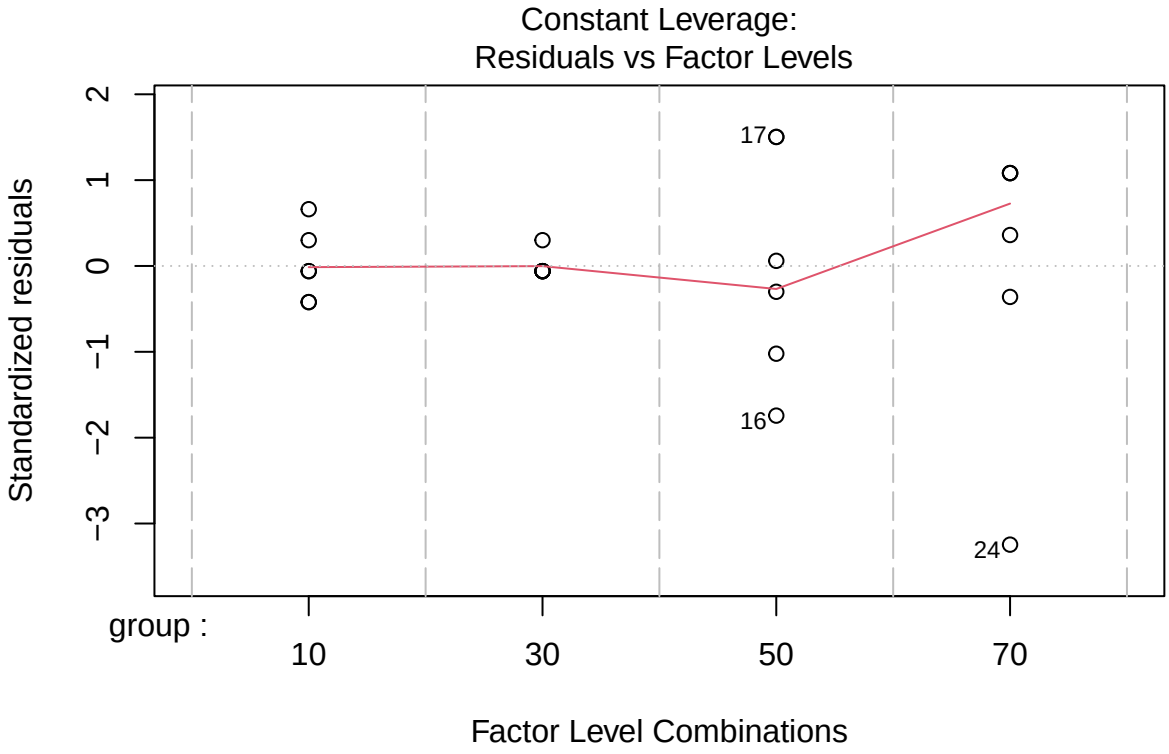
```
##          1          2          3          4          5          6          7
## -0.1666667 -0.1666667  1.8333333 -1.1666667 -1.1666667  0.8333333 -0.1666667
##          8          9         10         11         12         13         14
##  0.8333333 -0.1666667 -0.1666667 -0.1666667 -0.1666667 -2.8333333 -0.8333333
##         15         16         17         18         19         20         21
##  0.1666667 -4.8333333  4.1666667  4.1666667  3.0000000  3.0000000  3.0000000
##         22         23         24
##  1.0000000 -1.0000000 -9.0000000
```

```
plot(aov1)
```




Q−Q Residuals
Standardized residuals
2
1
0
−1
−2
−3
17
16
24
−2
−1
0
1
2
Theoretical Quantiles
aov(data ~ group)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
24
16
17
55
60
65
70
Fitted values
aov(data ~ group)



(5) `pairwise.t.test()` 함수를 이용하여 연령에 따른 라면 평가 차이를 확인하여라. 이 때, LSD와 Bonferroni 방법으로 결과를 확인하고, 이를 바탕으로 두 방법에 대해서 각각 비슷한 평가 점수를 가지는 연령을 그룹으로 표현하여라. 두 방법의 결과가 다르다면 왜 그런지 이유를 작성하여라.

```
pairwise.t.test(data, group, p.adjust="none", pool.sd=T)
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  data and group
##
##    10      30      50
## 30 0.0099  -       -
## 50 2.8e-09 5.4e-07 -
## 70 1.4e-07 6.3e-05 0.0409
##
## P value adjustment method: none
```

```
pairwise.t.test(data, group, p.adjust="bonferroni", pool.sd=T)
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  data and group
##
##    10      30      50
## 30 0.05921 -       -
## 50 1.7e-08 3.3e-06 -
```

```
## 70 8.7e-07 0.00038 0.24525
##
## P value adjustment method: bonferroni
```

Bonferroni 방법을 이용하여 다중비교에 대한 보정을 해준결과, $H_0 : \mu_{10s} = \mu_{30s}$, $H_0 : \mu_{50s} = \mu_{70s}$ 두 경우를 기각하지 못하였다. Bonferroni 방법의 경우, 검정해야할 귀무가설의 개수를 이용하여 p-value를 보정하는 과정에서, LSD 방법에서 얻은 유의확률에 귀무가설의 개수인 6을 곱하는 형태로 유의확률이 얻어지게 되어 귀무가설에 대한 검정력이 낮아지게 된다.

(6) 구체적인 비교를 위해 아래와 같은 두 개의 귀무가설을 설정하였다. 아래의 두 가설을 표현하는 대비를 작성하여라. 이 때 두 대비는 서로 직교하는가? 그 이유를 직교대비의 정의를 이용해 작성하여라.

$$\begin{aligned} H_0 &: \frac{1}{4}\tau_{10} + \frac{1}{2}\tau_{30} + \frac{1}{4}\tau_{50} - \tau_{70} = 0 \\ H_0 &: 2\tau_{10} - \tau_{30} - \tau_{70} = 0 \end{aligned}$$

$$\left(\frac{1}{4}\right)(2) + \left(\frac{1}{2}\right)(-1) + \left(\frac{1}{4}\right)(0) + (-1)(-1) = 1 \neq 0$$

따라서 두 대비는 직교하지 않는다.

(7) 위의 두 귀무가설 중 두 번째 가설$H_0 : 2\tau_{10} - \tau_{30} - \tau_{70} = 0$에 대한 검정을 진행하고 유의수준 0.05에서 결과를 해석하여라.

```
tapply(data,group,mean)
```

```
##       10       30       50       70
## 55.16667 60.16667 72.83333 69.00000
```

```
contrastmatrix = c(2,-1,0,-1)
contrasts(group) = contrastmatrix
contrasts(group)
```

```
##    [,1]       [,2]        [,3]
## 10    2 -0.2867757  0.03306064
## 30   -1 -0.3677574 -0.66939360
## 50    0  0.8603272 -0.09918192
## 70   -1 -0.2057940  0.73551488
```

```
summary.lm(aov(data~group))
```

```
##
## Call:
## aov(formula = data ~ group)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0000 -0.8750 -0.1667  1.2083  4.1667
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  64.2917     0.6200 103.700  < 2e-16 ***
## group1       -3.1389     0.5062  -6.201 4.67e-06 ***
## group2       10.5135     1.2400   8.479 4.69e-08 ***
## group3        5.0754     1.2400   4.093 0.000566 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.037 on 20 degrees of freedom
## Multiple R-squared:  0.864,  Adjusted R-squared:  0.8436
## F-statistic: 42.37 on 3 and 20 DF,  p-value: 7.479e-09
```

(8) 이 데이터에 대해 처리수준이 4개인 경우 직교대비는 다음과 같다.

$$\begin{bmatrix} -3 & 1 & -1 \\ -1 & -1 & 3 \\ 1 & -1 & -3 \\ 3 & 1 & 1 \end{bmatrix}$$

위 직교 대비를 이용한 검정을 진행하고 결과를 해석하여라. 또 이 때, t 분포를 따르는 통계량들의 제곱의 합이 aov에서는 SSt와 일치함을 R 결과와 함께 보여라.

```
contrastmatrix = cbind( c(-3, -1, 1, 3),
                        c(1, -1, -1, 1),
                        c(-1, 3, -3, 1))
contrasts(group) = contrastmatrix
summary.lm(aov(data~group))
```

```
##
## Call:
## aov(formula = data ~ group)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0000 -0.8750 -0.1667  1.2083  4.1667
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  64.2917     0.6200 103.700  < 2e-16 ***
## group1        2.7083     0.2773   9.768 4.69e-09 ***
## group2       -2.2083     0.6200  -3.562 0.001954 **
## group3       -1.2083     0.2773  -4.358 0.000305 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.037 on 20 degrees of freedom
## Multiple R-squared:  0.864,  Adjusted R-squared:  0.8436
## F-statistic: 42.37 on 3 and 20 DF,  p-value: 7.479e-09
```

```
summary(aov(data~group))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## group        3 1172.5   390.8   42.37 7.48e-09 ***
## Residuals   20  184.5     9.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit = summary.lm(aov(data~group))
sum(fit$coefficients[,3][-1]^2)*(sum( fit$residuals^2)/20)
```

```
## [1] 1172.458
```

```r
c1 = c(-3, -1, 1, 3)
mean=tapply(data, group,mean)
ni = 6
a = sum(c1*mean)^2/(sum(c1^2/ni))

c2 = c(1, -1, -1, 1)
b = sum(c2*mean)^2/(sum(c2^2/ni))

c3 = c(-1, 3, -3, 1)
c = sum(c3*mean)^2/(sum(c3^2/ni))
a+b+c
```

```
## [1] 1172.458
```