실험 계획 및 실습

Homework #(**2**)
2014-16757 김보창

---

1 Q1

3.26
여기서 모형은 $y_{ij} = \mu + \tau_i + \epsilon_{ij} = \mu_i + \epsilon_{ij}, i = 1...3, j = 1, 2...5, \sum_{i=1}^{3} \tau_i = 0$으로 놓을 수 있다.
$(\epsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2))$

1.1 1-(a)

$$H_0 : \tau_i = 0, \forall i = 1, 2, 3$$
$$H_1 : \tau_i \neq 0, \exists i$$

로 귀무가설과 대립가설을 세우고, single factor ANOVA test를 진행하면 된다.
우리는 H_0하에서 $\frac{MS_{Trt}}{MS_E} \sim F_{a-1,N-a}$임을 알고 있으므로,
$F_0 = \frac{MS_{Trt}}{MS_E}$ 로 두고, 유의수준 α에서 $F_0 > F_{a-1,N-a}(\alpha)$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.

```
y1 <- c(100, 96, 92, 96, 92)
y2 <- c(76, 80, 75, 84, 82)
y3 <- c(108, 100, 96, 98, 100)
y <- c(y1, y2, y3)

group <- rep(1:3, c(5,5,5))
group_df <- data.frame(y, group)
group_df <- transform(group_df, group = as.factor(group))

result <- summary(aov(y ~ group, data = group_df))
```

데이터를 입력해주고, group이라는 벡터에 각 데이터가 어떤 그룹의 데이터인지 명시해준다.
그후, data.frame 함수를 이용해 data frame으로 만들어주고, transform함수를 통해 group을 factor형으로 바꿔준다.
이제, aov와 summary함수를 이용하여 결과를 출력하면 다음과 같은 값이 나온다.

```
> result
            Df Sum Sq Mean Sq F value   Pr(>F)    
group        2 1196.1   598.1   38.34 6.14e-06 ***
Residuals   12  187.2    15.6                     
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

이를통해, 여기서의 MS_E의 값은 15.6, MS_{Trt}의 값은 598.1임을 알 수 있고,
$F_0 = 38.34$임을 알 수 있다. 또한 이때의 P-value가 $6.14 * 10^{-6}$임을 알 수 있다.
따라서, $\alpha = 0.05$로 택했을때, P-value가 0.05보다 작으므로 귀무가설을 기각할 수 있다.
즉, 평균 배터리 수명이 다르다는 결론을 내리게 된다.

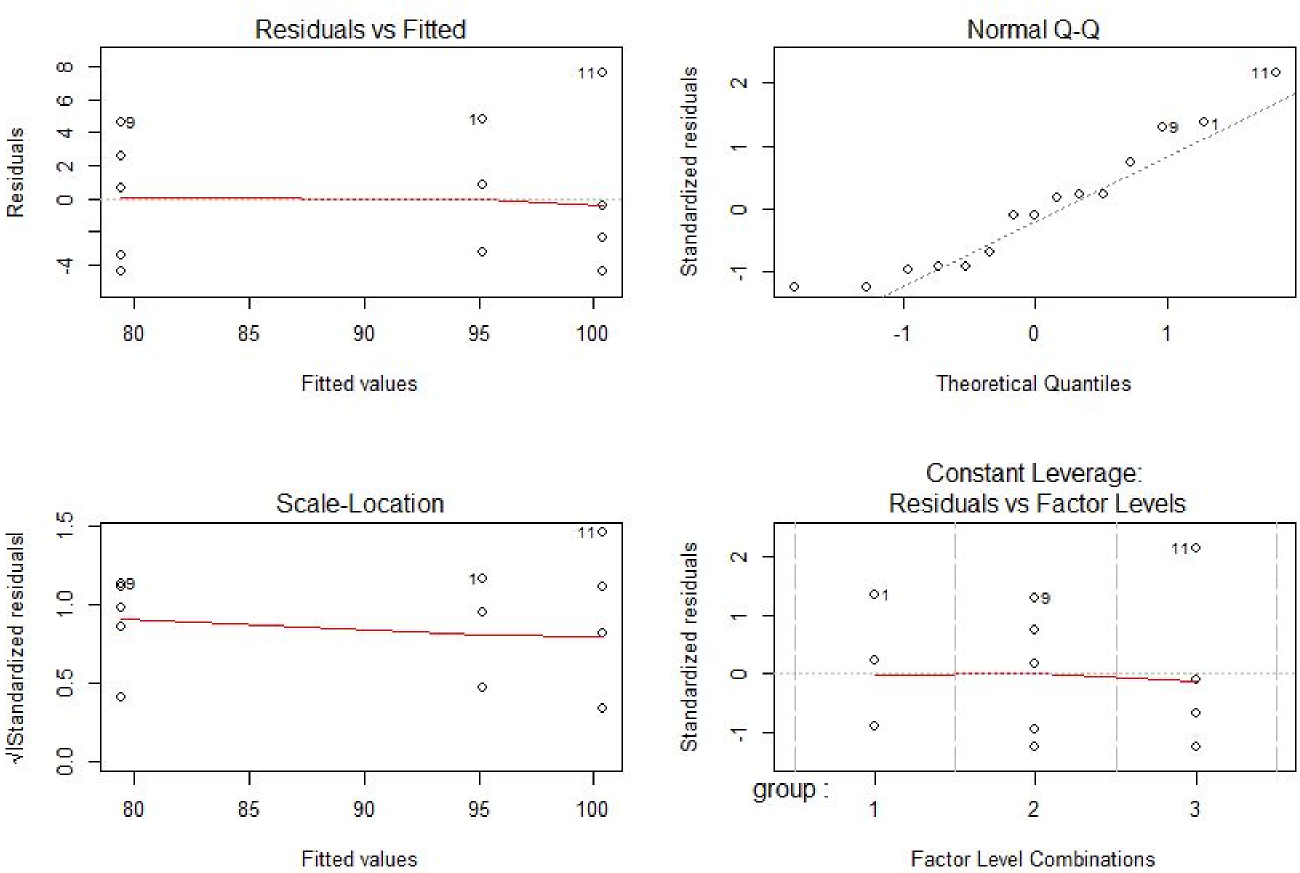
Homework #(**2**)
2014-16757 김보창

1.2 1-(b)

residual에 대한 분석을 하기 위해, 실습시간에 배운 R함수를 이용한다.

```
opar <- par(mfrow=c(2,2),cex=.8)
plot(aov(y ~ group, data = group_df))
```

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



왼쪽 위 그래프를 보면, fitted value와 residual로 그래프를 그렸을때, 특정 경향성이 나타나지 않는것을 알 수 있다.
따라서 우리 모형이 데이터를 잘 표현한다고 말할 수 있고,
오른쪽 아래 그래프를 봐도, factor level에 따라 residual의 차이가 거의 없으므로
등분산 가정을 위반하지 않음을 알 수 있다.
normal qq 그래프를 그려도, residual의 경향이 직선으로 정렬된 것을 보아 normality 가정을 위반하지 않음을 알 수 있다.
따라서 $\epsilon_{ij} \sim N(0, \sigma^2)$이라는 우리의 가정에 문제가 없음을 알 수 있다.

실험 계획 및 실습

Homework #(**2**)
2014-16757 김보창

1.3 1-(c)

먼저 배터리 2의 confidence interval을 구하자.

그전에, ANOVA모형으로 잠시 되돌아가서, $MS_E = \frac{SS_E}{N-a}$ 임을 알고 있고,

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \bar{y_{i.}})^2$$

,

$$\frac{SS_E}{\sigma^2} \sim \chi_{N-a}$$

임을 알고있다.

또한, 우리는 $\bar{y}$가 표본평균, S^2가 표본분산이라 할때, $\bar{y}$와 S^2가 independent함을 알고 있다.

때문에, $\bar{y_{i.}}$, S_i^2는 서로 indep하고, $SS_E = (n-1)\sum_{i=1}^{a} S_i^2$ 임을 안다.

이때, ϵ_{ij}가 서로 indep하므로, $i \neq j$일때 $\bar{y_{i.}}$와 S_j^2가 indep함은 자명하다.

($\bar{y_{i.}}$를 이루는 항들은 S_j^2를 계산하는데 전혀 사용되지 않음)

따라서, 모든 (j = 1, 2, ... a)에 대해 $\bar{y_{i.}}$와 S_j^2 가 indep 하므로, $\bar{y_{i.}}$ 와 SS_E는 indep하다.

그러므로 이제 u_i의 confidence interval을 구할 수 있다.

$\bar{y_{i.}} \sim N(\mu_i, \frac{\sigma^2}{n})$ 에서

$$\frac{\bar{y_{i.}} - \mu_i}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

,

$$SS_E/\sigma^2 \sim \chi_{N-a}$$

$$T = \frac{\bar{y_{i.}} - \mu_i}{\sqrt{\frac{MS_E}{n}}} \sim t_{N-a}$$

가 되므로,

따라서, 이를 이용해서 μ_i의 $100(1-\alpha)\%$ confidence interval을 구하면

$P(|T| \leq t_{N-a}(\frac{\alpha}{2})) = 1-\alpha$에서,

$$\bar{y_{i.}} \pm t_{N-a}(\frac{\alpha}{2})\sqrt{\frac{MS_E}{n}}$$

가 된다.

마찬가지로, $\bar{y_{i.}} - \bar{y_{j.}}$ 은 $\bar{y_{i.}}$, $\bar{y_{j.}}$가 각각 SS_E와 독립이므로,

SS_E와 독립이 되고,

$$\bar{y_{i.}} - \bar{y_{j.}}\ N(\mu_i - \mu_j, \frac{2\sigma^2}{)}n$$

에서,

같은 방법으로 $\mu_i - \mu_j$의 $100(1-\alpha)\%$ confidence interval을 구하면

$$\bar{y_{i.}} - \bar{y_{j.}} \pm t_{N-a}(\frac{\alpha}{2})\sqrt{\frac{2MS_E}{n}}$$

따라서, 위 계산과정을 이용하여 R 코드를 이용하여 답을 구하면,

```
MSE <- result[[1]][2,3]
n <- 5
mean2 <- mean(y2)
mean3 <- mean(y3)
t1 <- c(-1, 1) * qt(0.975, 12)
t2 <- c(-1, 1) * qt(0.995, 12)
int1 <- mean2 + t1 * sqrt(MSE/n)
int2 <- mean2 - mean3 + t2 * sqrt(2*MSE/n)
```

```
> MSE <- result[[1]][2,3]
> n <- 5
> mean2 <- mean(y2)
> mean3 <- mean(y3)
> t1 <- c(-1, 1) * qt(0.975, 12)
> t2 <- c(-1, 1) * qt(0.995, 12)
> int1 <- mean2 + t1 * sqrt(MSE/n)
> int2 <- mean2 - mean3 + t2 * sqrt(2*MSE/n)
> int1
[1] 75.55145 83.24855
> int2
[1] -28.63024 -13.36976
> 
```

실행 결과

따라서 배터리 브랜드 2 평균 수명의 95% CI는 다음과 같다.

$$[75.55145, 83.24855]$$

마찬가지로, 브랜드 2 평균수명 - 브랜드 3 평균수명의 99% CI는 다믕과 같다.

$$[-28.63024, -13.36976]$$

1.4 1-(d)

각 브랜드의 평균은 다음과 같다.

```
> mean(y1)
[1] 95.2
> mean(y2)
[1] 79.4
> mean(y3)
[1] 100.4
> 
```

보다시피, 3번브랜드의 배터리 수명 표본 평균이 100.4로 가장 높으므로,3번 브랜드를 택할것이다.

이제, 3번브랜드의 배터리 수명 평균, 즉 ,$\mu_3 = 100.4$로 가정하고, 1-(a)에서 $MS_E = 15.6$임을 알 수 있으므로, $\sigma = \sqrt{MS_E} = \sqrt{15.6}$로 가정한다.

그리고 3번 브랜드의 배터리 수명이 normal분포를 따른다고 가정하면, 이제 배터리의 수명이 85보다 작을 확률을 구할 수 있다.

$y \sim N(100.4, 15.6)$에서, $P(y < 85)$를 구하면

$P(y < 85) = P(Z < \frac{(85-100.4)}{\sqrt{15.6}})$의 값을 구하면 되고, 이는 아래와 같이 구해진다.

```
> x = (85 - 100.4)/sqrt(15.6)
> pnorm(x)
[1] 4.82861e-05
> |
```

즉, $4.83 * 10^{-5}$ 정도의 확률을 가짐을 알 수 있다.

2 Q2

3.22에 있는 데이터는 circuit type에 따른 response time을 나타낸것으로, 다음과 같은 ANOVA 모델을 세울 수 있다.

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

, $i = 1...a, j = 1, 2...n$

이 성립하고, 3.22에서는 treatment 종류인 $a = 3$, 각 treatment별 $n = 5$임을 알 수 있다.

2.1 2-(a)

이제, least squares normal equation을 세워보면,

$$S = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \mu - \tau_i)^2$$

위와 같을때, 위를 최소화하는 μ, τ_i들을 찾으면 된다.

즉, 위의 S를 각 μ, τ_i로 편미분할때 0으로 만드는 값들을 $\hat{\mu}, \hat{\tau_i}$들로 놓았을때 이들에 대한 식이 normal eqaua-tion이 된다.

$\frac{\partial S}{\partial \mu} = 0$, $\frac{\partial S}{\partial \tau_i} = 0$에서,

$$N\hat{\mu} + n\hat{\tau_1} + ... + n\hat{\tau_a} = y_{..}$$

$$n\hat{\mu} + n\hat{\tau_i} = y_{i.}, \quad (i = 1, 2, ...a)$$

즉, 위의 a+1개의 식들이 normal equation이 되므로,

3.22의 데이터로 least square noraml equation을 세우면

$$15\hat{\mu} + 5\hat{\tau_1} + 5\hat{\tau_2} + 5\hat{\tau_3} = y_{..}$$

$$5\hat{\mu} + 5\hat{\tau_1} = y_{1.}$$

$$5\hat{\mu} + 5\hat{\tau_2} = y_{2.}$$

$$5\hat{\mu} + 5\hat{\tau_3} = y_{3.}$$

x 위와 같은 식들이 되고, $\sum_{i=1}^{3} \hat{\tau_i} = 0$인 constraint하에서, 이를 각 parameter에 대해 풀면

$$\hat{\mu} = \bar{y_{..}}$$

$$\hat{\tau_i} = \bar{y_{i.}} - \bar{y_{..}}, \quad (i = 1, 2, 3)$$

이 성립하게 된다.

따라서 이 데이터로 구한 $\tau_1 - \tau_2$의 추정값인 $\hat{\tau_1} - \hat{\tau_2}$의 값은 다음과 같다.

```r
y1 <- c(9, 12, 10, 8, 15)
y2 <- c(20, 21, 23, 17, 30)
y3 <- c(6, 5, 8, 16, 7)
y <- c(y1,y2,y3)
mu <- mean(y)
tau_1 <- mean(y1) - mean(y)
tau_2 <- mean(y2) - mean(y)
tau_3 <- mean(y3) - mean(y)
answer <- tau_1 - tau_2
mu
tau_1
tau_2
tau_3
answer
```

위 코드의 실행결과에서

```
> mu
[1] 13.8
> tau_1
[1] -3
> tau_2
[1] 8.4
> tau_3
[1] -5.4
> answer
[1] -11.4
```

-11.4가 답임을 알 수 있다. ($\hat{\mu} = 13.8, \hat{\tau_1} = -3, \hat{\tau_2} = 8.4, \hat{\tau_3} = -5.4$)

2.2 2-(b)

$$15\hat{\mu} + 5\hat{\tau_1} + 5\hat{\tau_2} + 5\hat{\tau_3} = y_{..}$$
$$5\hat{\mu} + 5\hat{\tau_1} = y_{1.}$$
$$5\hat{\mu} + 5\hat{\tau_2} = y_{2.}$$
$$5\hat{\mu} + 5\hat{\tau_3} = y_{3.}$$

constraint $\hat{\tau_3} = 0$에서 위 normal equation을 풀면

$$\hat{\mu} = \bar{y_{3.}}$$
$$\hat{\tau_1} = \bar{y_{1.}} - \bar{y_{3.}}$$
$$\hat{\tau_2} = \bar{y_{2.}} - \bar{y_{3.}}$$
$$\hat{\tau_3} = 0$$

위와 같아진다.
이제, 각 parameter의 값을 구하면 다음과 같이 구할 수 있다.

```
y1 <- c(9, 12, 10, 8, 15)
y2 <- c(20, 21, 23, 17, 30)
y3 <- c(6, 5, 8, 16, 7)
y <- c(y1,y2,y3)
alt_mu <- mean(y3)
alt_tau_1 <- mean(y1) - mean(y3)
alt_tau_2 <- mean(y2) - mean(y3)
alt_tau_3 <- 0
answer <- tau_1 - tau_2
alt_mu
alt_tau_1
alt_tau_2
alt_tau_3
answer
```

위 코드의 실행결과에서

```
> alt_mu
[1] 8.4
> alt_tau_1
[1] 2.4
> alt_tau_2
[1] 13.8
> alt_tau_3
[1] 0
> answer
[1] -11.4
```

로, $\tau_1 - \tau_2$의 추정값은 그대로 -11.4이고, $\hat{\mu} = 8.4, \hat{\tau_1} = 2.4, \hat{\tau_2} = 13.8, \hat{\tau_3} = 0$ 으로 달라졌음을 알 수 있다.

이 값이 달라진 이유로는, 아래와 같이 설명할 수 있다.

위 normal equation에서 constraint $\hat{\tau_3} = 0$ 으로 두는것은, treatment 3의 평균을 기준으로 다른 treatment의 평균을 측정하겠다는 이야기와 같다.

즉, $y_{ij} = \mu_i + \epsilon_{ij}$일때, 2-(a)에서의 constraint가 의미하는것은 $\mu = \frac{\sum i=1^3 \mu_i}{a}$를 기준으로 τ들의 값을 결정하겠다는 이야기이고, 2-(b)에서의 constraint가 의미하는것은 $\mu = \mu_3$을 기준으로 τ들의 값을 결정하겠다는 이야기이다.

즉, constraint의 의미는 어떤 treatment를 기준으로 effect를 관측할것인가에 대한 의미인것이다.

따라서, treatment 1과 treatment 2의 차이인 $\tau_1 - \tau_2$ 이 constraint가 바뀌어도 그대로인 이유도 위와같이 설명이 가능하다.

2.3 2-(c)

$\mu + \tau_1, 2\tau_1 - \tau_2 - \tau_3, \mu + \tau_1 + \tau_2$의 추정값은 각각
$\hat{\mu} + \hat{\tau_1}, 2\hat{\tau_1} - \hat{\tau_2} - \hat{\tau_3}, \hat{\mu} + \hat{\tau_1} + \hat{\tau_2}$로, 각 constraint에서 값은 다음과 같다.

```
> mu + tau_1
[1] 10.8
> 2*tau_1 - tau_2 - tau_3
[1] -9
> mu + tau_1 + tau_2
[1] 19.2
> alt_mu + alt_tau_1
[1] 10.8
> 2*alt_tau_1 - alt_tau_2 - alt_tau_3
[1] -9
> alt_mu + alt_tau_1 + alt_tau_2
[1] 24.6
```

$$\sum_{i=1}^{3} \hat{\tau}_i = 0$$

$\hat{\mu} + \hat{\tau}_1 = 10.8$ $2\hat{\tau}_1 - \hat{\tau}_2 - \hat{\tau}_3 = -9$ $\hat{\mu} + \hat{\tau}_1 + \hat{\tau}_2 = 19.2$

$$\hat{\tau}_3 = 0$$

$\hat{\mu} + \hat{\tau}_1 = 10.8$ $2\hat{\tau}_1 - \hat{\tau}_2 - \hat{\tau}_3 = -9$ $\hat{\mu} + \hat{\tau}_1 + \hat{\tau}_2 = 24.6$

$\hat{\mu}+\hat{\tau}_1$ 의 의미는 treatment 1의 평균으로, 두 경우 모두 같다. $2\hat{\tau}_1-\hat{\tau}_2-\hat{\tau}_3$ 의 의미는 treatment 1과 2, treatment 1과 3의 차이의 합으로, 두 경우 모두 같다. $\hat{\mu}+\hat{\tau}_1+\hat{\tau}_2$는 기준점에 treatment 1과 2의 효과를 더한것으로, 두 경우가 다르다. 기준점이 다르니 이는 당연한것이다.

3 Q3

example 5의 모델은 다음과 같다.

$y_{ij} = \mu + \tau_i + \epsilon_{ij}$ (i = 1,2,3,4, j = 1,2,3..6)

이 모델을 이용해서, general regression significance test를 하자.

$H_0 : \tau_i = 0 \forall i = 1, 2, 3, 4$, $H_1 : \tau_i \neq 0 \exists i$ 가 된다.

이를 test하기 위한 test statistic

$$F_0 = \frac{R(\tau|\mu)/(a-1)}{[\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - R(\mu,\tau)]/N-a}$$

이고, 여기서 $a = 4, n = 6, N = 24$가 된다.

H_0하에서 $F_0 \sim F_{a-1,N-a}$가 성립하므로, 유의수준 α에서 $F_0 > F_{a-1,N-a}(\alpha)$이면 H_0를 기각할것이다.

이제, test를 하기위해 $R(\tau|\mu)$, $R(\mu,\tau)$를 구해야한다.

$R(\tau|\mu) = R(\mu,\tau) - R(\mu)$이므로, 먼저 $R(\mu)$를 구하자.

general regression significance test를 위해, normal equation을 세우면,

$y_{ij} = \mu + \epsilon_{ij}$ 인 모델의 normal equation은

$$N\hat{\mu} = y_{..}$$

이므로, $R(\mu)$의 자유도는 normal equation중에 lin.indep한 식의 개수이므로 1이고,

$$\hat{\mu} = \frac{y_{..}}{N} = \bar{y}_{..}$$

에서

$$R(\mu) = \hat{\mu} * y_{..}$$

이 된다.

이제 $R(\mu, \tau)$를 구하면
$y_{ij} = \mu + \tau_i + \epsilon_{ij}$인 모델의 normal equation은

$$N\hat{\mu} + n\hat{\tau_1} + ... + n\hat{\tau_4} = y_{..}$$
$$N\hat{\mu} + n\hat{\tau_1} = y_{1.}$$
$$N\hat{\mu} + n\hat{\tau_2} = y_{2.}$$
$$N\hat{\mu} + n\hat{\tau_3} = y_{3.}$$
$$N\hat{\mu} + n\hat{\tau_1} = y_{4.}$$

이 되고, $R(\mu, \tau)$의 자유도는 normal equation중에 lin.indep한 식의 개수이므로 4가된다.
constraint $\sum_{i=1}^{4} \hat{\tau_i} = 0$에서,

$$\hat{\mu} = \frac{y_{..}}{N} = \bar{y_{..}}$$
$$\hat{\tau_i} = \bar{y_{i.}} - \bar{y_{..}}, (i = 1, 2, 3, 4)$$

이고,

$$R(\mu, \tau) = \hat{\mu} * y_{..} + \hat{\tau_1} * y_{1.} + \hat{\tau_2} * y_{2.} + \hat{\tau_3} * y_{3.} + \hat{\tau_4} * y_{4.}$$

가 된다.

이제, 데이터를 이용하여 실제 $R(\mu, \tau)$, $R(\mu)$를 구하면 다음과 같다.
R 코드를 이용하여 구하면

```
y1 <- c(0.34, 0.12, 1.23, 0.70, 1.75, 0.12)
y2 <- c(0.91, 2.94, 2.14, 2.36, 2.86, 4.55)
y3 <- c(6.31, 8.37, 9.75, 6.09, 9.82, 7.24)
y4 <- c(17.15, 11.82, 10.95, 17.20, 14.35, 16.82)
y <- c(y1,y2,y3,y4)
y_dotdot <- sum(y)
y_1dot <- sum(y1)
y_2dot <- sum(y2)
y_3dot <- sum(y3)
y_4dot <- sum(y4)
mu_hat <- mean(y)
tau_1_hat <- mean(y1) - mean(y)
tau_2_hat <- mean(y2) - mean(y)
tau_3_hat <- mean(y3) - mean(y)
tau_4_hat <- mean(y4) - mean(y)

R_mu = mu_hat * y_dotdot
R_mu_tau = mu_hat * y_dotdot + tau_1_hat * y_1dot + tau_2_hat * y_2dot + tau_3_hat * y_3dot + tau_4_hat * y_4dot
R_tau_bar_mu = R_mu_tau - R_mu
sum_of_sqaure_y = sum(y**2)
```

```
F_0 =  (R_tau_bar_mu / 3) / ((sum_of_sqaure_y - R_mu_tau) / 20)

(R_tau_bar_mu / 3)
((sum_of_sqaure_y - R_mu_tau) / 20)
F_0
```

```
> (R_tau_bar_mu / 3)
[1] 236.1157
> ((sum_of_sqaure_y - R_mu_tau) / 20)
[1] 3.104054
> F_0
[1] 76.06688
> |
```

각각 책의 값과 비교해보면,
$R(\tau|\mu)/3$이 책의 MS_{Trt}인 236.1157, $[\sum_{i=1}^{4}\sum_{j=1}^{6} y_{ij}^2 - R(\mu,\tau)]/20)$ 이 책의 MS_E와 같은 3.1041과 거의 같음을 알 수 있고,
F_0이 책에 있는 F_0 인 76.07과 거의 같음을 알 수 있다.
유효숫자 표기에 의한 반올림을 고려하면, 사실상 책의 값과 똑같은 값이 나왔음을 알 수 있다.
귀무가설/대립가설이 같으므로 P-value도 같고, 따라서 example5와 똑같은 결론을 내릴 수 있다.
즉, usual한 ANOVA와 같은 결과가 나옴을 확인할 수 있었다.