05

데이터분석방법론(1)

Analysis of Variance

통계·데이터과학과 장영재 교수

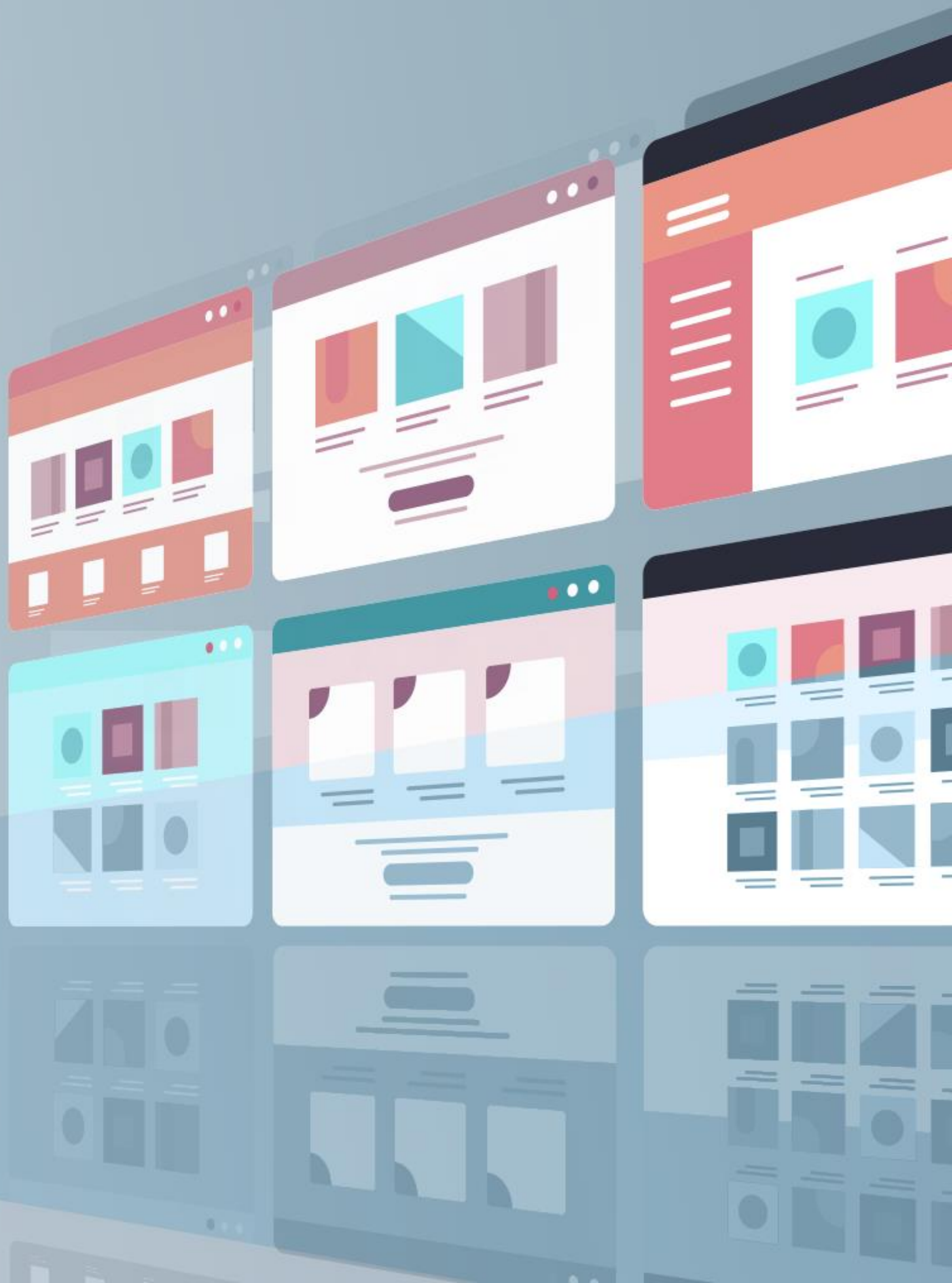
한국방송통신대학교 대학원

학습목차

1. One-way analysis of variance
2. Pairwise comparisons and multiple testing
3. Two-way analysis of variance

01

One-way analysis of variance

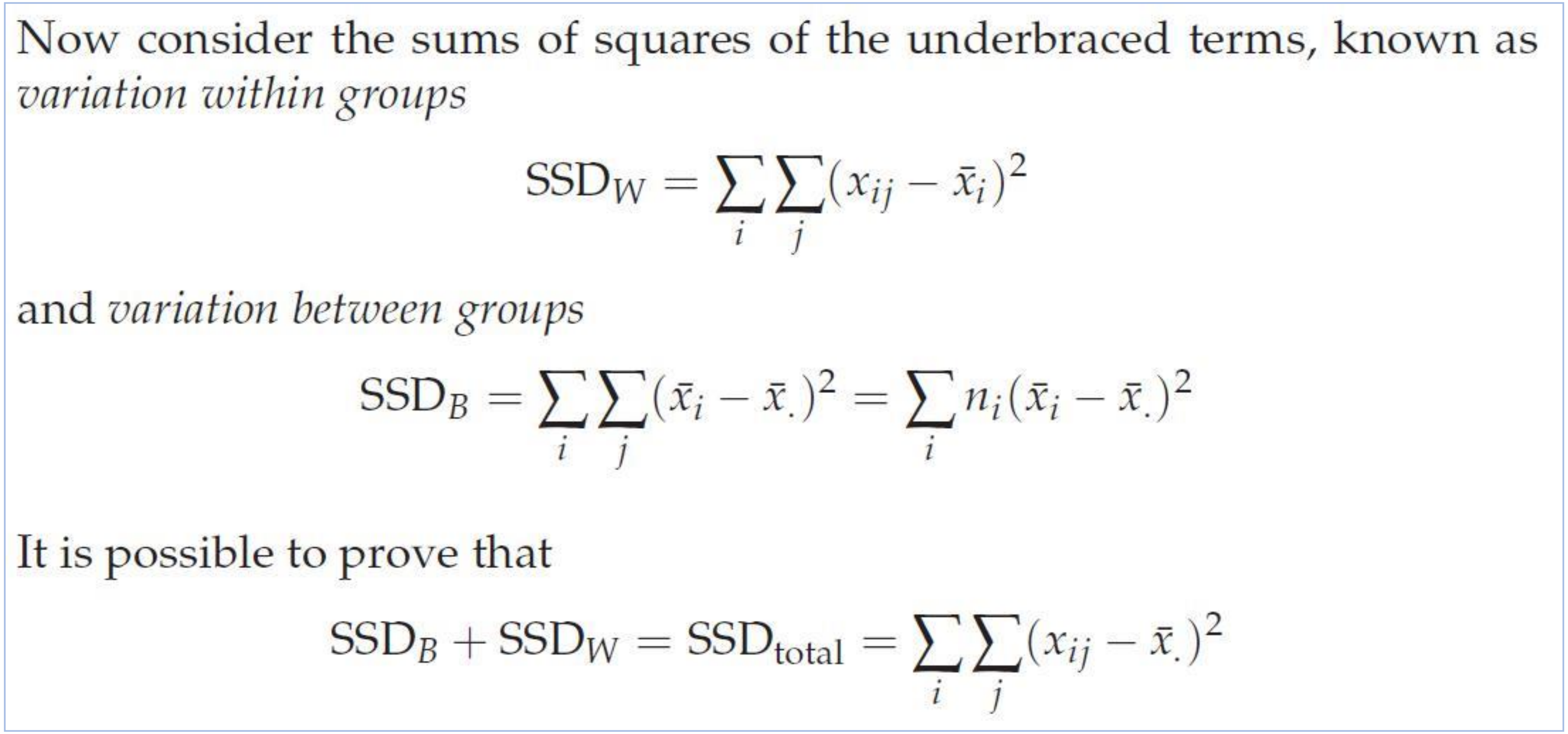


1. Data Structure

그룹	1	⋯	i	⋯	k
1	x_{11}		x_{i1}	⋯	x_{k1}
⋯					
J	x_{1j}	⋯	x_{ij}	⋯	x_{kj}
⋯					
n	x_{1n}	⋯	x_{in}	⋯	x_{kn}
평균	$\bar{\mathrm{x}}_{1}$		$\bar{\mathrm{x}}_{i}$		$\bar{\mathrm{x}}_{k}$

- Let x_{ij} denote observation no. j in group i, so that x_{35} is the fifth observation in group 3; $\bar{\mathrm{x}}_{i}$ is the mean for group i, and $\bar{x}_{.}$ is the grand mean (average of all observations).
- Decompose $x_{ij} - \bar{x}_{.} = (x_{ij} - \bar{x}_{i}) + (\bar{x}_{i} - \bar{x}_{.})$
- Model $X_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \varepsilon_{ij} \sim N\,(0, \sigma^2)$

2. Decomposition of Variation



Now consider the sums of squares of the underbraced terms, known as *variation within groups*

$$\mathrm{SSD}_W = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2$$

and *variation between groups*

$$\mathrm{SSD}_B = \sum_i \sum_j (\bar{x}_i - \bar{x}_.)^2 = \sum_i n_i (\bar{x}_i - \bar{x}_.)^2$$

It is possible to prove that

$$\mathrm{SSD}_B + \mathrm{SSD}_W = \mathrm{SSD}_{\text{total}} = \sum_i \sum_j (x_{ij} - \bar{x}_.)^2$$

3. ANOVA table and test

- 분산분석표 (Analysis of Variance Table)

요인	제곱합	자유도	평균제곱	F값
그룹간	SSD_B	k-1	MS_B	MS_B/MS_w
그룹내	SSD_W	N-k	MS_w	
합	SSD_{total}	N-1		

- 검정 $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_k vs H_1 : n\,ot H_0$

  검정통계량 $F_0 = MS_B/MS_W \sim F\,(k-1, N-k)$

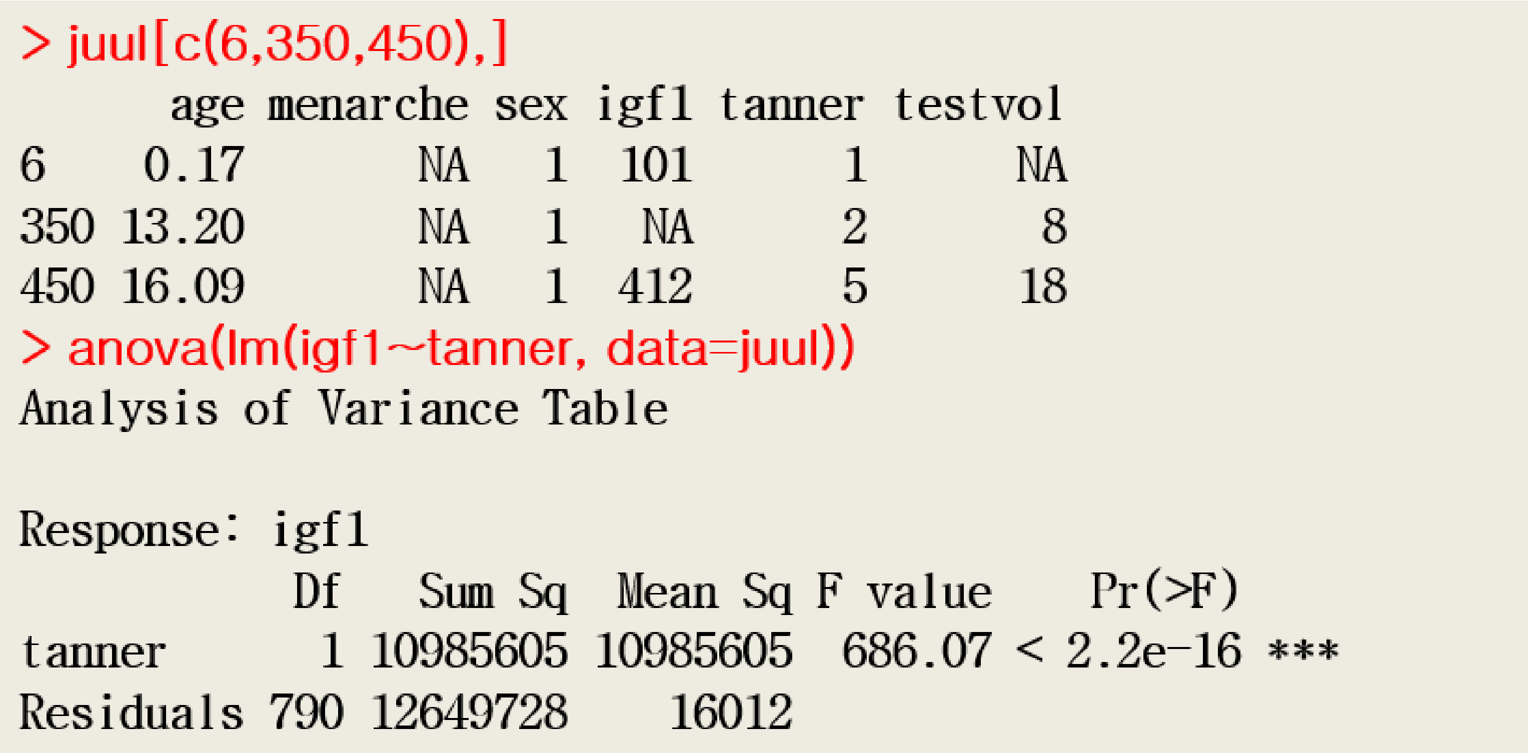
4. One-way analysis of variance example

- “red cell folate” data from Altman (1991, p. 208).

```
> library(ISwR)
> head(red.cell.folate, 2)
  folate ventilation
1    243  N20+02,24h
2    251  N20+02,24h
> attach(red.cell.folate)
> summary(red.cell.folate)
     folate          ventilation
 Min.   :206.0   N20+02,24h:8
 1st Qu.:249.5   N20+02,op :9
 Median :274.0   02,24h    :5
 Mean   :283.2
 3rd Qu.:305.5
 Max.   :392.0
> anova(lm(folate~ventilation))
Analysis of Variance Table
Response: folate
            Df Sum Sq Mean Sq F value   Pr(>F)
ventilation  2  15516  7757.9  3.7113 0.04359 *
Residuals   19  39716  2090.3
```

5. One-way analysis of variance example (Common mistakes)

- (Recall) the data set “juul”

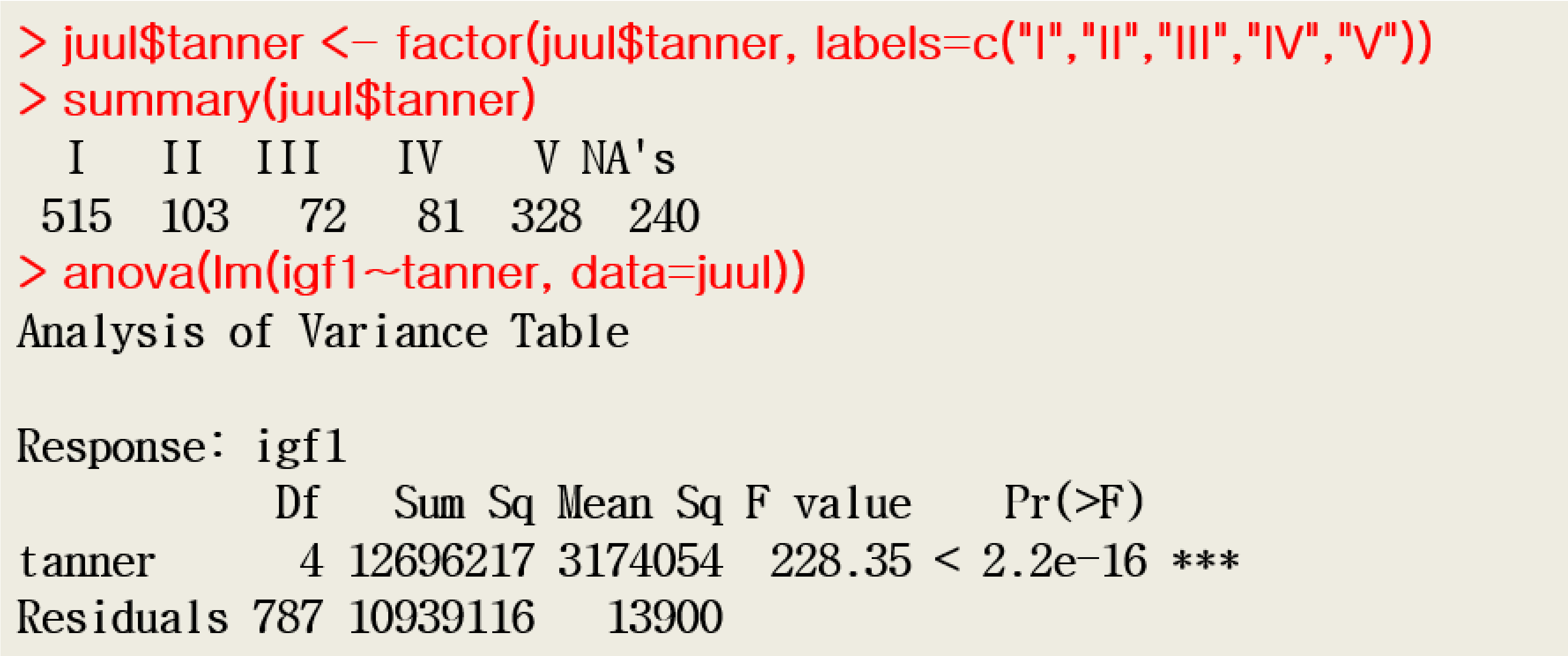


```
> juul[c(6,350,450),]
      age menarche sex igf1 tanner testvol
6    0.17       NA   1  101      1      NA
350 13.20       NA   1   NA      2       8
450 16.09       NA   1  412      5      18
> anova(lm(igf1~tanner, data=juul))
Analysis of Variance Table

Response: igf1
           Df   Sum Sq  Mean Sq F value    Pr(>F)    
tanner      1 10985605 10985605  686.07 < 2.2e-16 ***
Residuals 790 12649728    16012
```

- Wrong !!! Since tanner is not a factor variable.
- This does not describe a grouping of data but a linear regression on the group number! Notice “1 DF” for the effect of tanner in the table.

5. One-way analysis of variance example (Correction)

- (Recall) the data set “juul” : fix



```
> juul$tanner <- factor(juul$tanner, labels=c("I","II","III","IV","V"))
> summary(juul$tanner)
  I   II  III   IV    V NA's 
 515  103   72   81  328  240 
> anova(lm(igf1~tanner, data=juul))
Analysis of Variance Table

Response: igf1
           Df   Sum Sq Mean Sq F value    Pr(>F)    
tanner      4 12696217 3174054  228.35 < 2.2e-16 ***
Residuals 787 10939116   13900
```

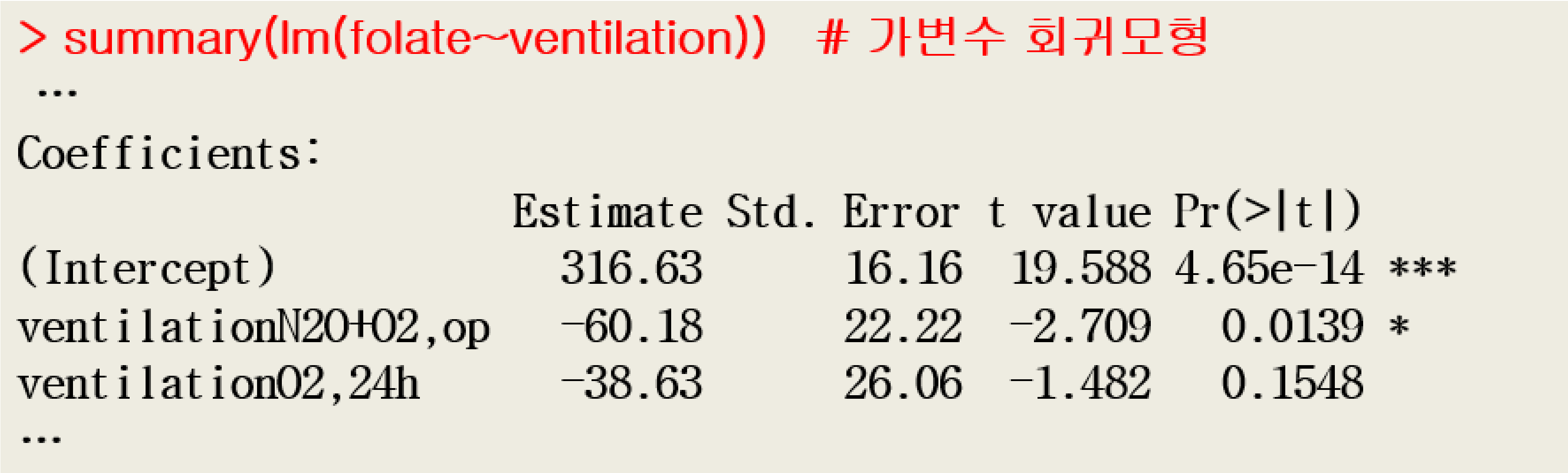
02

Pairwise comparisons and multiple testing

1. Detecting a difference between groups

Difference shown by the F test (ANOVA)

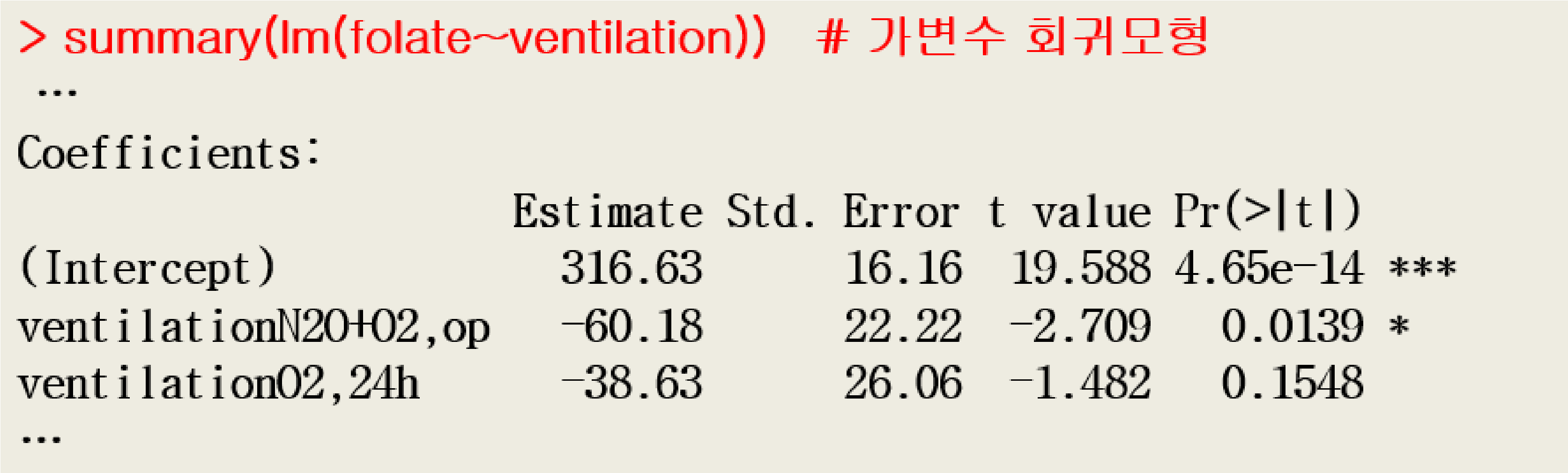
- If the F test shows that there is a difference between groups, the question quickly arises of where the difference lies. It becomes necessary to compare the individual groups.
- Part of this information can be found in the regression coefficients.



```
> summary(lm(folate~ventilation))   # 가변수 회귀모형
 ...
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            316.63      16.16  19.588 4.65e-14 ***
ventilationN2O+O2,op   -60.18      22.22  -2.709   0.0139 *
ventilationO2,24h      -38.63      26.06  -1.482   0.1548
...
```

- The interpretation of the estimates is that the intercept is the mean in the first group (N2O+O2,24h), whereas the two others describe the difference between the relevant group and the first one.

1. Detecting a difference between groups



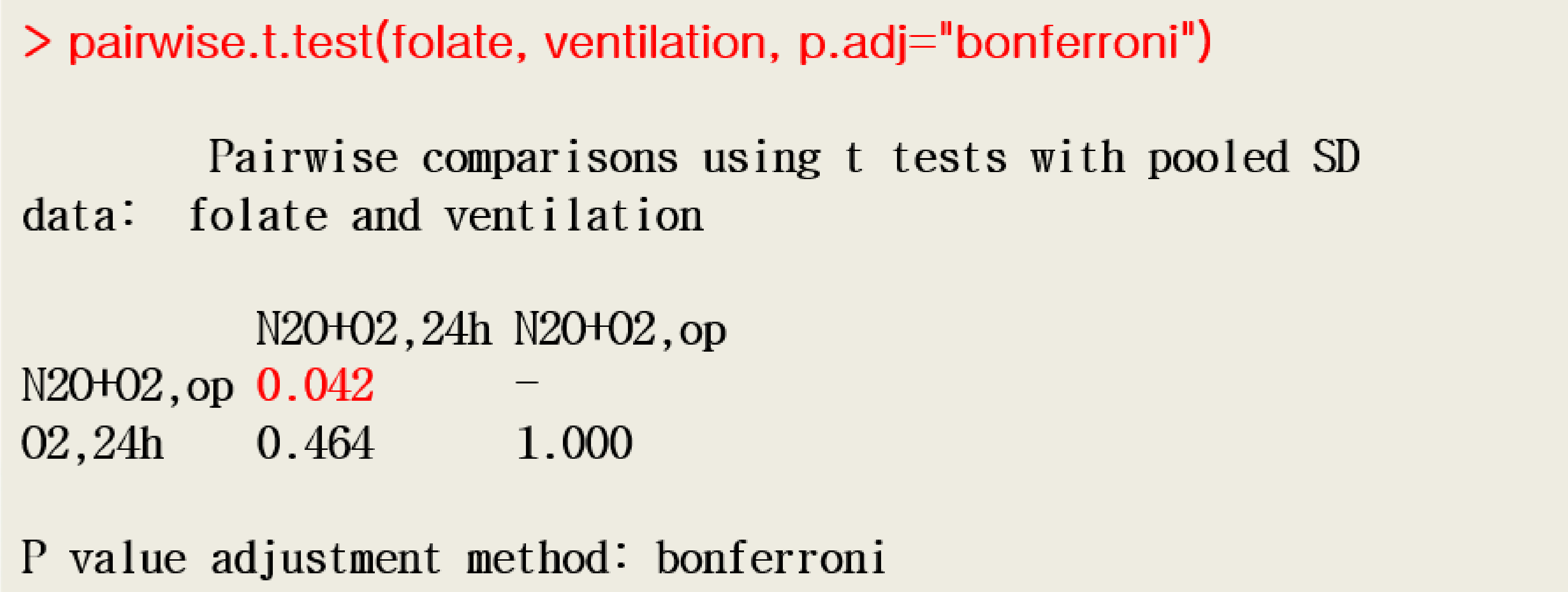
```
> summary(lm(folate~ventilation))  # 가변수 회귀모형
 ...
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            316.63      16.16  19.588 4.65e-14 ***
ventilationN2O+O2,op   -60.18      22.22  -2.709   0.0139 *
ventilationO2,24h      -38.63      26.06  -1.482   0.1548
...
```

- Contrasts used by default are the so-called treatment contrasts, in which the first group is treated as a baseline and the other groups are given relative to that. Concretely, the analysis is performed as a multiple regression analysis by introducing two dummy variables, which are 1 for observations in the relevant group and 0 elsewhere.

- Among the t tests in the table, you can immediately find a test for the hypothesis that the first two groups have the same true mean (p = 0.0139, reject) and also whether the first and the third might be identical (p = 0.1548, not reject). However, a comparison of the last two groups cannot be found. This can be overcome by modifying the factor definition, but that gets tedious when there are more than a few groups.

2. Pairwise test (pooled SD) and Bonferroni correction

- Multiple testing (다중비교) : If we want to compare all groups, we ought to correct for multiple testing. Performing many tests will increase the probability of finding one of them to be significant; that is, the p-values tend to be exaggerated. A common adjustment method is the Bonferroni correction.
- Multiplying the p-values by the number of tests, we obtain a conservative test where the probability of a significant result is less than or equal to the formal significance level.

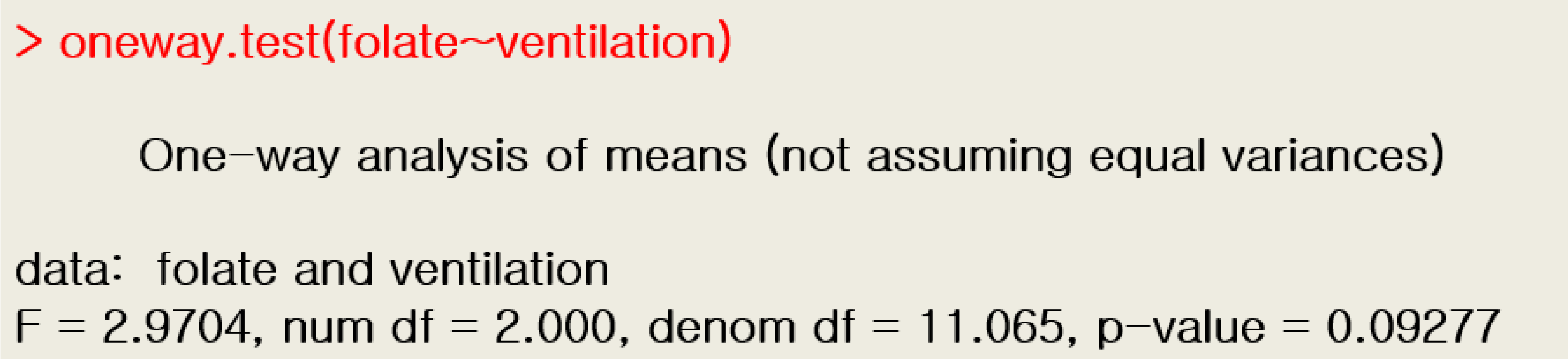


```
> pairwise.t.test(folate, ventilation, p.adj="bonferroni")

        Pairwise comparisons using t tests with pooled SD
data:   folate and ventilation

           N2O+O2,24h N2O+O2,op
N2O+O2,op  0.042      -
O2,24h     0.464      1.000

P value adjustment method: bonferroni
```

3. Not assuming equal variances

Relaxing the variance assumption

- The traditional one-way ANOVA requires an assumption of equal variances for all groups. There is, however, an alternative procedure that does not require that assumption. It is (due to Welch) similar to the unequal-variances t test. This has been implemented in the oneway.test function:

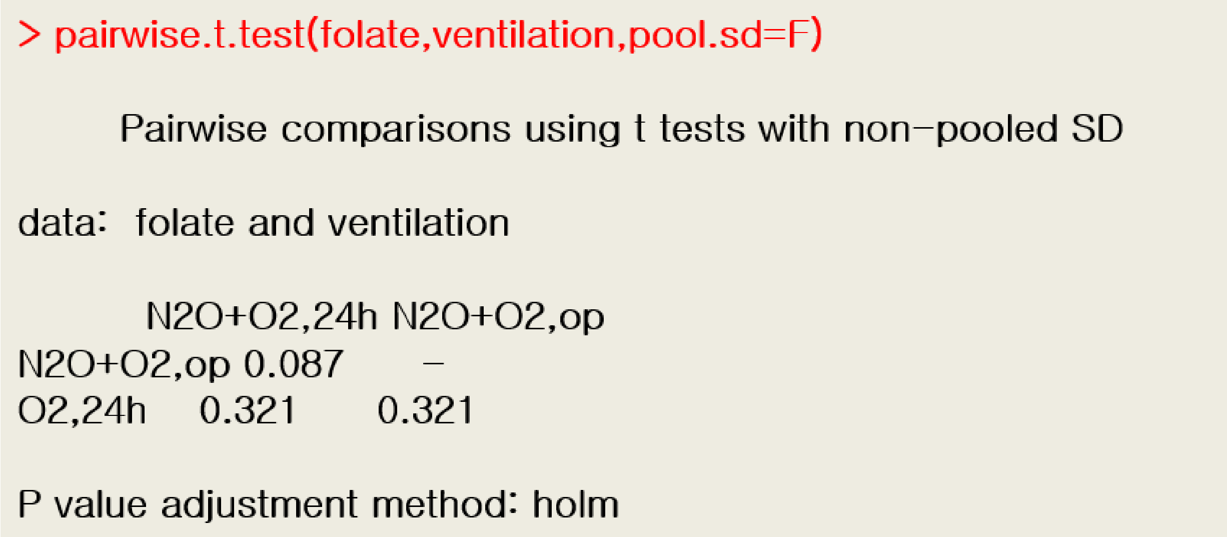


```
> oneway.test(folate~ventilation)

        One-way analysis of means (not assuming equal variances)

data:  folate and ventilation
F = 2.9704, num df = 2.000, denom df = 11.065, p-value = 0.09277
```

- In this case, the p-value increased to an insignificant value, presumably related to the fact that the group that seems to differ from the two others also has the largest variance.

4. Pairwise test not using pooled SD

- It is also possible to perform the pairwise t tests so that they do not use a common pooled standard deviation. This is controlled by the argument pool.sd.



```
> pairwise.t.test(folate,ventilation,pool.sd=F)

        Pairwise comparisons using t tests with non-pooled SD

data:  folate and ventilation

          N2O+O2,24h N2O+O2,op
N2O+O2,op 0.087      -
O2,24h    0.321      0.321

P value adjustment method: holm
```

- Again, it is seen that the significance disappears as we remove the constraint on the variances.

5. Test for same variance - Bartlett's test

Test for same variance in all groups

- Testing whether the distribution of a variable has the same variance in all groups can be done using Bartlett's test, although like the F test for comparing two variances, it is rather non-robust against departures from the assumption of normal distributions.

```
> bartlett.test(folate~ventilation)

        Bartlett test of homogeneity of variances

data:  folate by ventilation
Bartlett's K-squared = 2.0951, df = 2, p-value = 0.3508
```

$$\chi^2 = \frac{(N-k)\ln(S_p^2) - \sum_{i=1}^{k}(n_i-1)\ln(S_i^2)}{1+\frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\left(\frac{1}{n_i-1}\right)-\frac{1}{N-k}\right)}$$

where $N = \sum_{i=1}^{k} n_i$ and $S_p^2 = \frac{1}{N-k}\sum_i (n_i-1)S_i^2$

Reject $H_0\ if\, \chi^2 > \chi^2_{\alpha,k-1}$

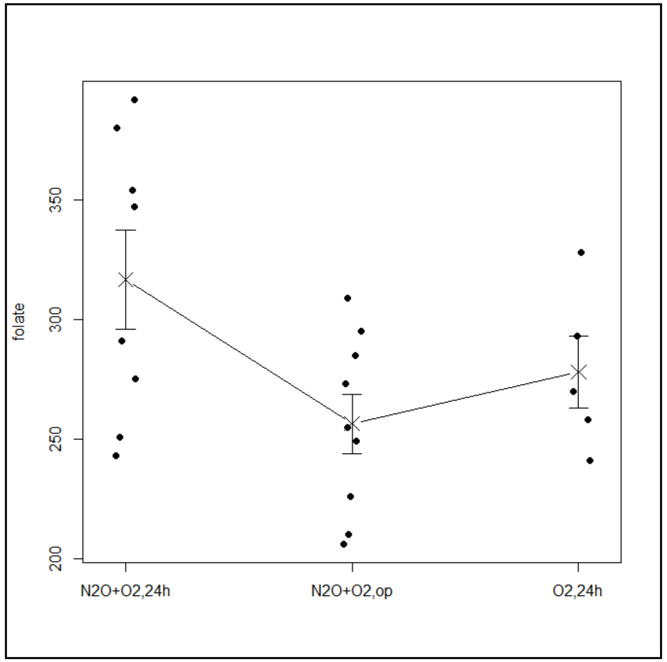
- nothing in the data contradicts the assumption of equal variances in the three groups.

6. Graphical presentation

Stripchart

- Plot where the raw data are plotted as a stripchart and overlaid with an indication of means and SEMs.

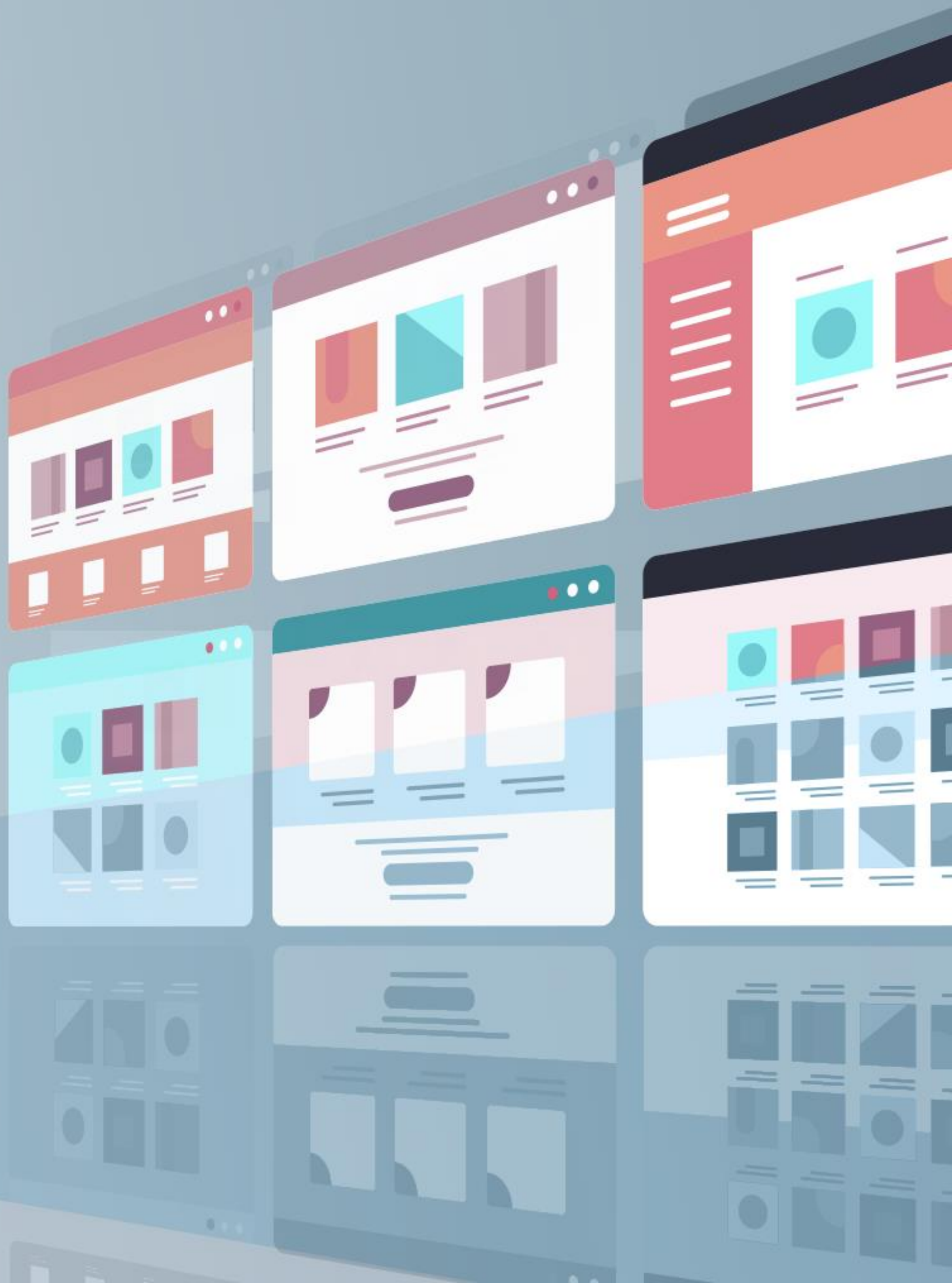
```
> xbar <- tapply(folate, ventilation, mean)
> s <- tapply(folate, ventilation, sd)
> sem <- s/sqrt(n)
> stripchart(folate~ventilation,
method="jitter",
+ jitter=0.05, pch=16, vert=T)
> arrows(1:3,xbar+sem,1:3,xbar-sem,
+            angle=90,code=3, length=.1)
> lines(1:3,xbar,pch=4,type="b",cex=2)
```



- In many fields it appears to have become the tradition to use 1 SEM "because they are the smallest"; that is, it makes differences look more dramatic.

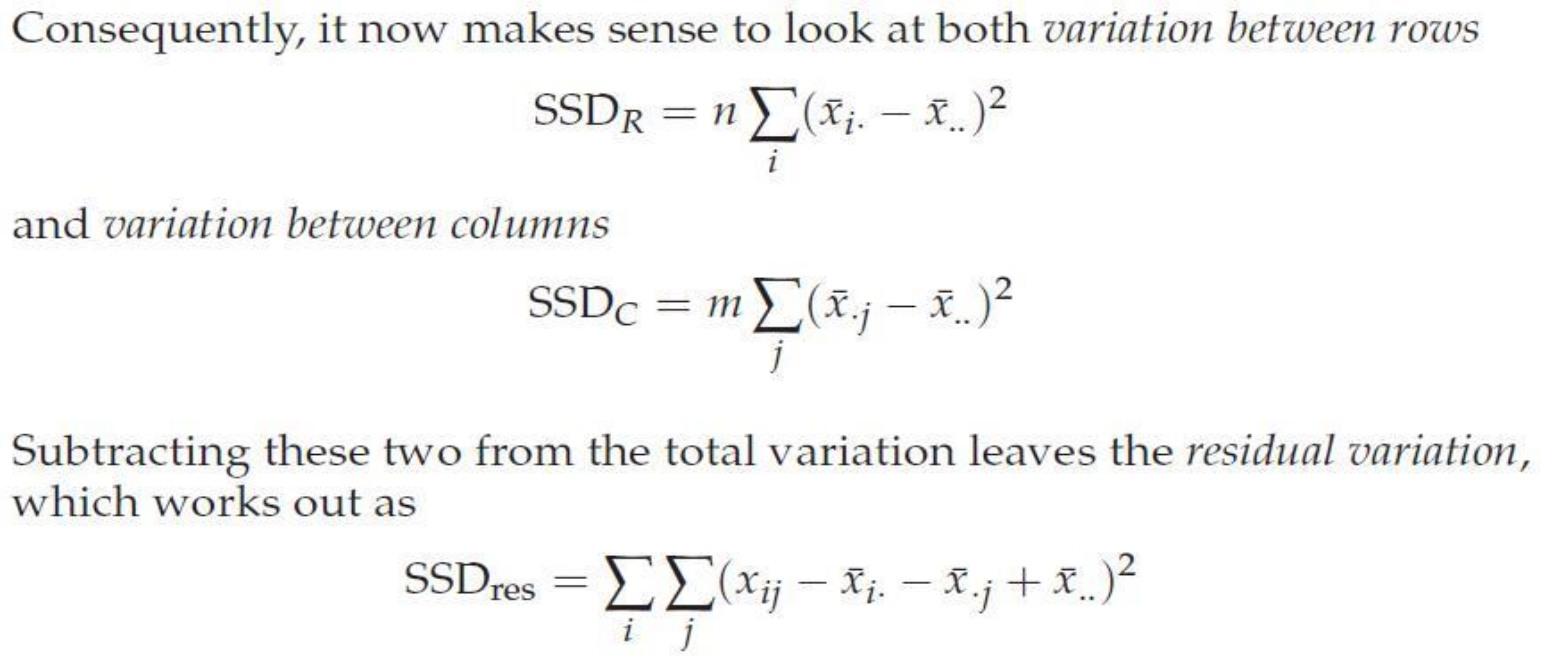
03

Two-way analysis of variance



1. Decomposition of Variation

- Let x_{ij} denote the observation in row i and column j of the m×n table. This is similar to the notation used for one-way analysis of variance, but notice that there is now a connection between observations with the same j, so that it makes sense to look at both row averages $\bar{x}_{i.}$ and column averages $\bar{x}_{.j}$. Here we restrict ourselves to the case of a single observation per cell.
- Decompose $x_{ij} - \bar{x}_{..} = (\bar{x}_{i.} - \bar{x}_{..}) + (\bar{x}_{.j} - \bar{x}_{..}) + (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})$



Consequently, it now makes sense to look at both *variation between rows*

$$\mathrm{SSD}_R = n\sum_i (\bar{x}_{i.} - \bar{x}_{..})^2$$

and *variation between columns*

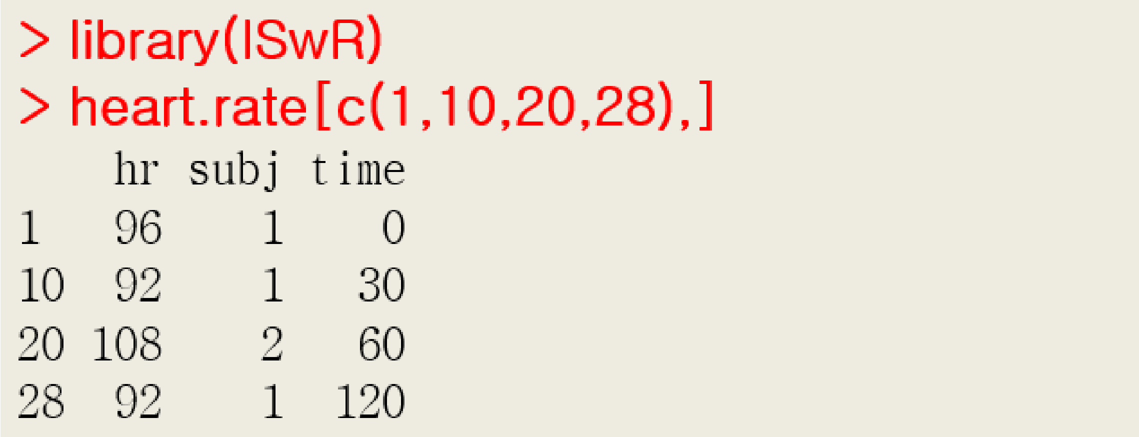
$$\mathrm{SSD}_C = m\sum_j (\bar{x}_{.j} - \bar{x}_{..})^2$$

Subtracting these two from the total variation leaves the *residual variation*, which works out as

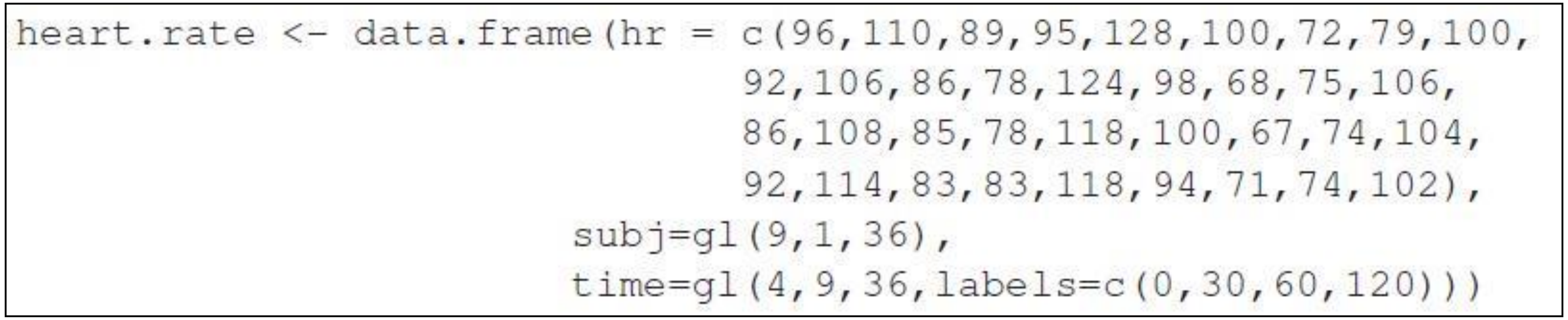
$$\mathrm{SSD}_{\mathrm{res}} = \sum_i \sum_j (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})^2$$

2. Two-way analysis of variance example

- Model $X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}\varepsilon_{ij} \sim N\,(0, \sigma^2)$
- ex) heart rate after administration of enalaprilate (Altman, 1991, p. 327)

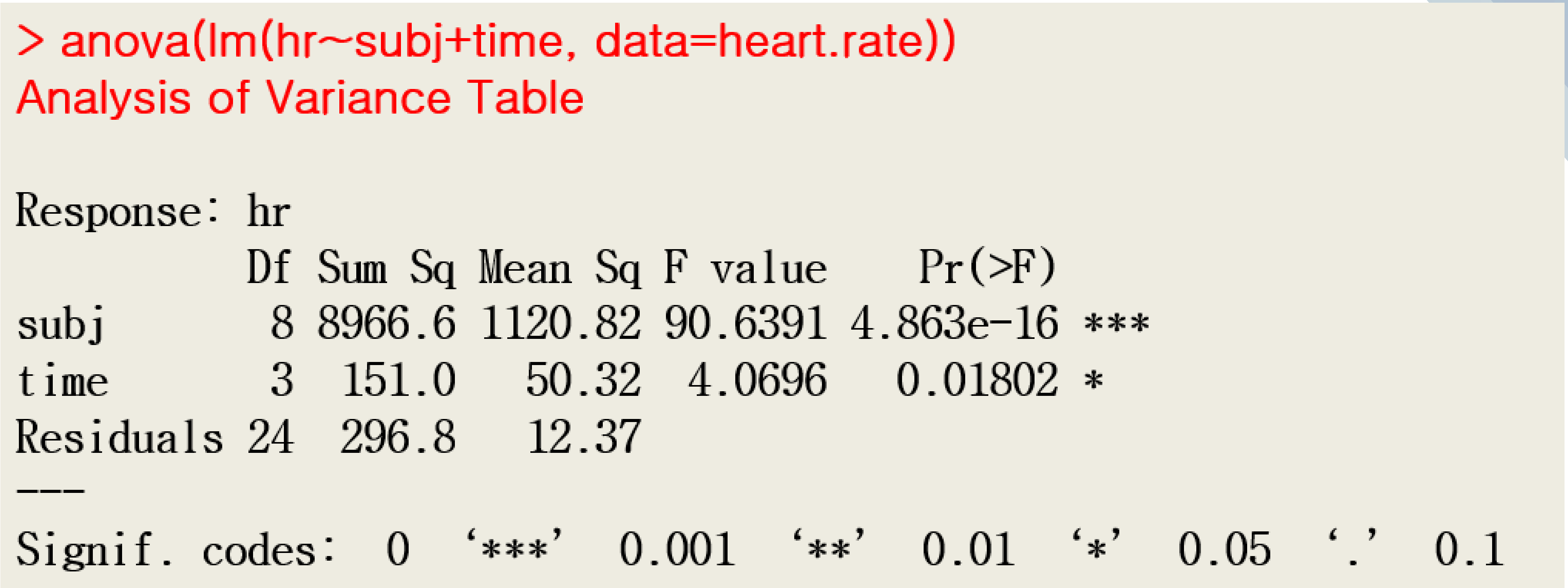


```
> library(ISwR)
> heart.rate[c(1,10,20,28),]
    hr  subj  time
1    96     1     0
10   92     1    30
20  108     2    60
28   92     1   120
```



```
heart.rate <- data.frame(hr = c(96,110,89,95,128,100,72,79,100,
                                92,106,86,78,124,98,68,75,106,
                                86,108,85,78,118,100,67,74,104,
                                92,114,83,83,118,94,71,74,102),
                       subj=gl(9,1,36),
                       time=gl(4,9,36,labels=c(0,30,60,120)))
```

2. Two-way analysis of variance example



```
> anova(lm(hr~subj+time, data=heart.rate))
Analysis of Variance Table

Response: hr
          Df Sum Sq Mean Sq F value    Pr(>F)
subj       8 8966.6 1120.82 90.6391 4.863e-16 ***
time       3  151.0   50.32  4.0696   0.01802 *
Residuals 24  296.8   12.37
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

다음시간 안내

06

Simple Linear Regression