Chapter 14: One-Way ANOVA

Gunwoong Park

Lecture Note

University of Seoul

One-Way ANOVA

- **AN**alysis **O**f **VA**riance: partition the overall variance in the response into parts due to each of the factors (explained variance) and the error (unexplained variance)
- Often done by comparing within and between group variances but can also be done via a linear model
- Predictors are all categorical/qualitative and are called factors
- Parameters are called effects
- Only look at fixed-effects models

ANOVA as a Linear Model

Given a factor α with $i = 1, \ldots, I$ levels, and $j = 1, \ldots, J_i$ observations per level, we use the model

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \ldots, I; \quad j = 1, \ldots, J_i$$

- Same assumption on errors required:

$$\epsilon_{ij} \text{ i.i.d. } N(0, \sigma^2)$$

- Problem: model is **unidentifiable** without constraints

Possible Constraints

- Set $\alpha_1 = 0$ (R default) and use $I - 1$ dummy variables
- Set $\mu = 0$ and use I dummy variables
- Set $\sum_i \alpha_i = 0$
- Set $\sum_i J_i \alpha_i = 0$. Solution:

$$\hat{\mu} = \bar{y}_{..}, \quad \hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$$

All equivalent fits but first two are easier to implement.

Blood Coagulation Example

- Study of blood coagulation times
- Factor: 4 different diets
- $n = 24$
- $J_1 = 4, J_2 = 6, J_3 = 6, J_4 = 8$

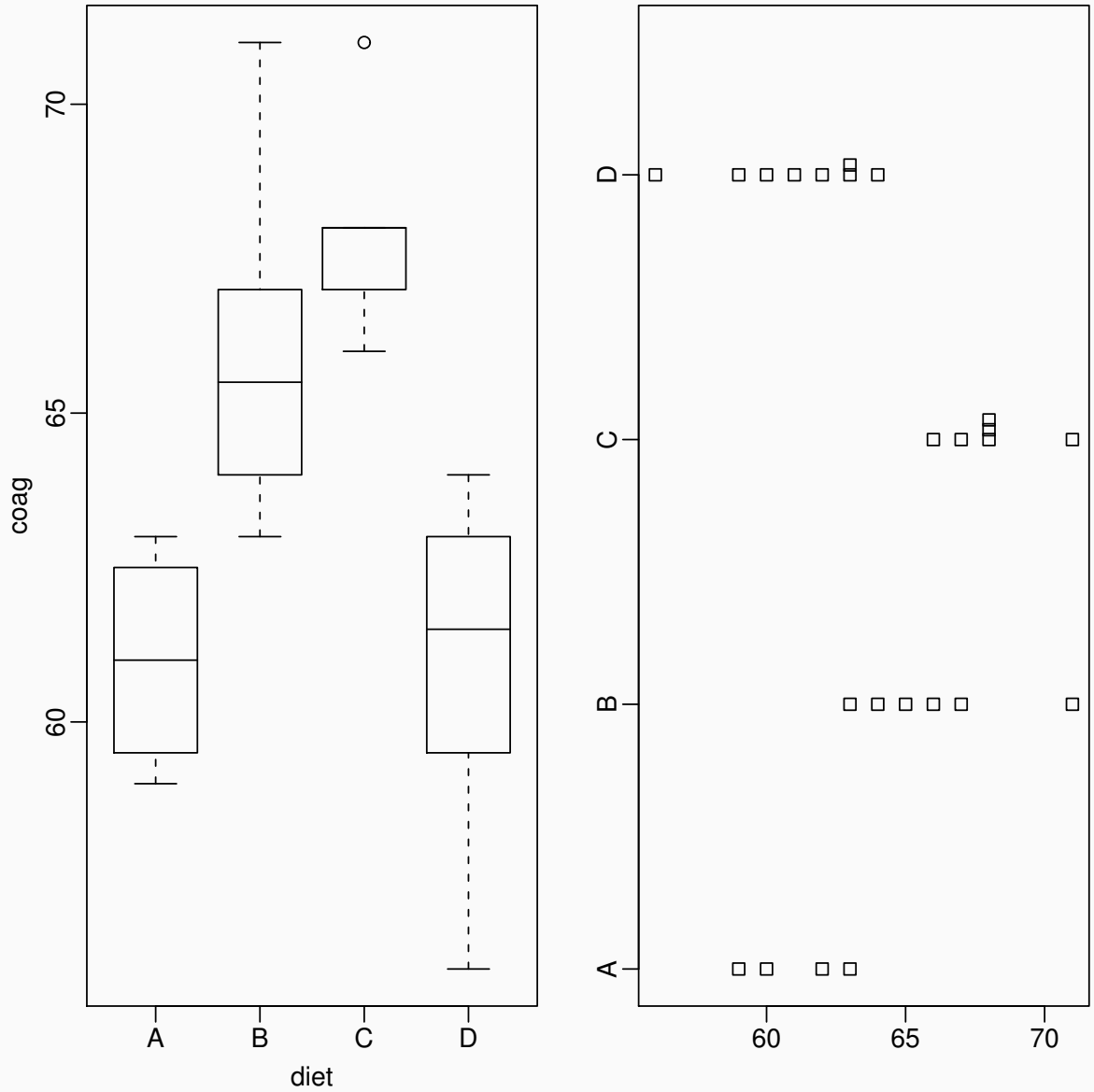
Blood Coagulation Example Continued

```
> library(faraway)
> data(coagulation)
> coagulation
coag diet
1    62    A
2    60    A
... ...
23   63    D
24   59    D
```

Blood Coagulation Example Continued

```
## Start from summary plots
> par(mfrow=c(1, 2))
> plot(coag ~ diet, coagulation)
> stripchart(coagulation$coag ~
coagulation$diet, method="stack")
## Stripchart plot is preferred over boxplots
## when there is little data
## Check for outliers, skewness or unequal variance
```

Blood Coagulation Summaries

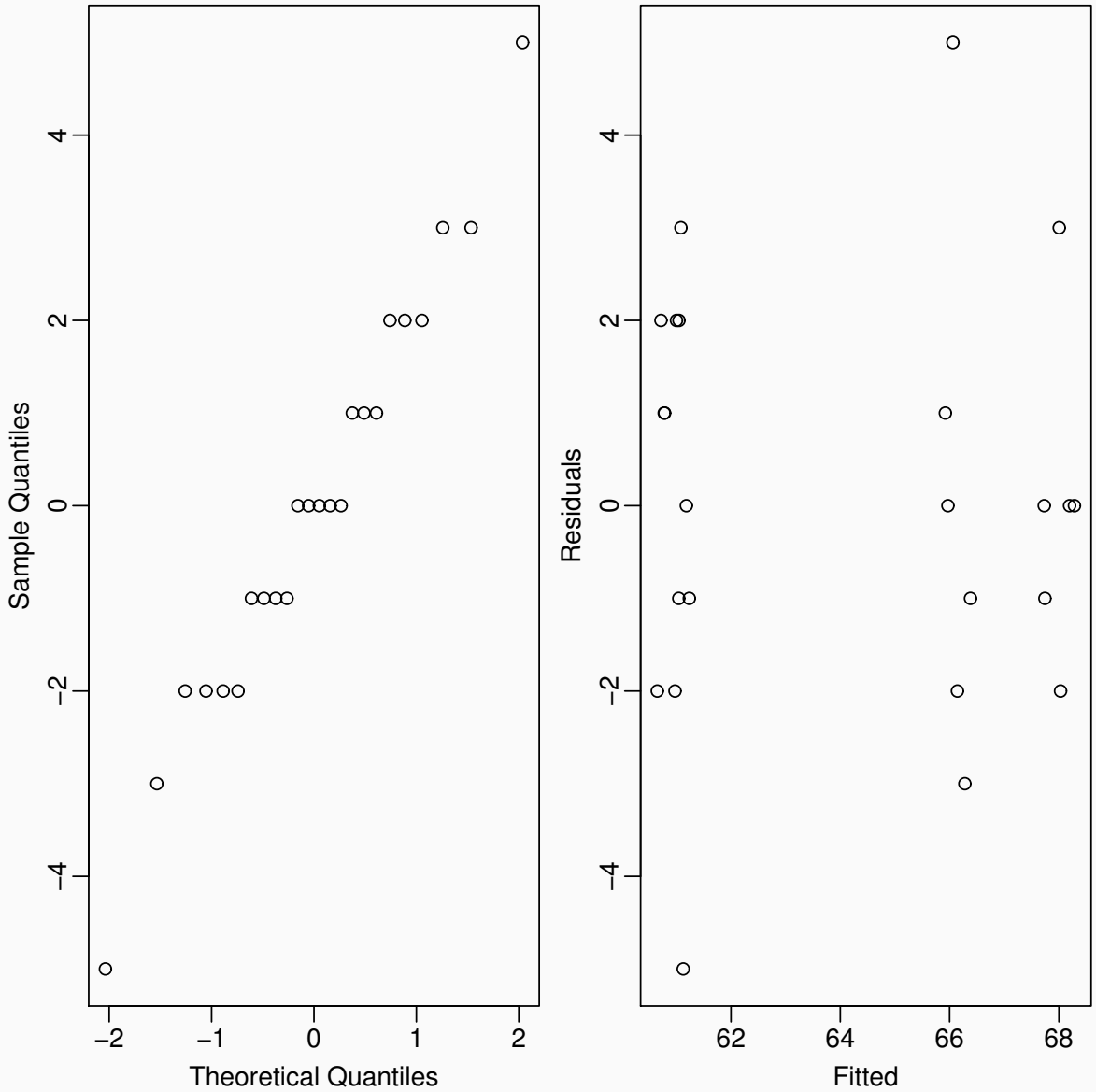


```
> g = lm(coag ~ diet, coagulation)
> summary(g)
Coefficients:
Estimate Std.Error t value Pr(>|t|)
Intercept 6.100e+01 1.183e+00  51.554  < 2e-16
dietB     5.000e+00 1.528e+00   3.273 0.003803
dietC     7.000e+00 1.528e+00   4.583 0.000181
dietD    -1.071e-14 1.449e+00  -7e-15 1.000000

Residual standard error: 2.366 on 20 degrees of freedom
Multiple R-Squared: 0.6706, Adjusted R-squared: 0.6212
F-statistic: 13.57 on 3 and 20 DF,  p-value: 4.658e-05

## Diagnostics
> qqnorm(residuals(g))
> plot(jitter(fitted(g)), residuals(g),
xlab="Fitted", ylab="Residuals")
```

Blood Coagulation Diagnostics



```
# Fit another contrast (drop the intercept)
> gi = lm(coag ~ diet - 1, coagulation)
> summary(gi)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
dietA  61.0000     1.1832   51.55   <2e-16
dietB  66.0000     0.9661   68.32   <2e-16
dietC  68.0000     0.9661   70.39   <2e-16
dietD  61.0000     0.8367   72.91   <2e-16
Residual standard error: 2.366 on 20 degrees of freedom
Multiple R-Squared: 0.9989, Adjusted R-squared: 0.9986
F-statistic:  4399 on 4 and 20 DF,  p-value: < 2.2e-16
## What do these p-values test?
```

Pairwise Comparisons

Exactly which levels or combinations of levels are different?

Pairwise comparison: a confidence interval for $\alpha_i - \alpha_{i'}$ is

$$(\hat{\alpha}_i - \hat{\alpha}_{i'}) \pm t_{df}^{\alpha/2} SE(\hat{\alpha}_i - \hat{\alpha}_{i'})$$

where

$$SE(\hat{\alpha}_i - \hat{\alpha}_{i'}) = \hat{\sigma}\sqrt{1/J_i + 1/J_{i'}}, \quad df = n - I$$

For $\alpha_i - \alpha_1$, the SE is in the output

Blood Coagulation Example – one CI

```
# CI for the difference between groups B and A
> qt(0.975, 20)
[1] 2.085963
> c(5 - 2.086 * 1.53, 5 + 2.086 * 1.53)
[1] 1.80842 8.19158
# CI for the difference between groups B and C
> SE =  2.366*sqrt(1/6+1/6)
> c((5-7) - 2.086 * SE, (5-7) + 2.086 * SE)
[1] -4.8494984  0.8494984
# or change the reference level to C
> coagulation$diet2 = relevel(coagulation$diet, ref='C')
```

Multiple Pairwise Comparisons

- If there are many pairwise comparisons, type I error can be much higher than α
- Bonferroni adjustment is OK for a few tests but very conservative
- For all pairwise comparisons, we use Tukey's Honest Significant Difference (HSD) CI:

$$(\hat{\alpha}_i - \hat{\alpha}_{i'}) \pm q(1-\alpha, I, df)\frac{1}{\sqrt{2}} \cdot se\,(\hat{\alpha}_i - \hat{\alpha}_{i'})$$

- q is a quantile of the studentized range distribution

Blood Coagulation: Tukey's CIs

```
> qtukey(0.95, 4, 20)/sqrt(2)
[1] 2.798936
> c(5 - 2.8 * 1.53, 5 + 2.8 * 1.53)
[1] 0.716 9.284
> ## Get all CIs
> TukeyHSD( aov(coag ~ diet, coagulation) )
Tukey multiple comparisons of means
95% family-wise confidence level
$diet
diff         lwr       upr
B-A  5.000000e+00   0.7245544  9.275446
C-A  7.000000e+00   2.7245544 11.275446
D-A -2.131628e-14  -4.0560438  4.056044
C-B  2.000000e+00  -1.8240748  5.824075
D-B -5.000000e+00  -8.5770944 -1.422906
D-C -7.000000e+00 -10.5770944 -3.422906
```

Multiple Contrast Comparisons

A contrast among the effects $\alpha_1, \ldots, \alpha_I$ is a linear combination $\sum_i c_i \alpha_i$ where c_i are known and $\sum_i c_i = 0$.

Examples:

- $\alpha_1 - \alpha_2$
- $\alpha_1 - (\alpha_2 + \alpha_3 + \alpha_4)/3$

Scheffé's CI for multiple constrast comparisons:

$$\sum_i c_i \hat{\alpha}_i \pm \sqrt{(I-1)F^{\alpha}_{I-1,n-I}}\hat{\sigma}\sqrt{\sum_i \frac{c_i^2}{J_i}}$$

Multiple Contrast Comparisons

```
> ## Scheffe's CI for (B+C)/2 - (A+D)/2
> sqrt( 3*qf(0.95, 3, 20) ) * 2.37 *
sqrt(1/4 + 1/6 + 1/6 + 1/8)/2
[1] 3.040647
> (5 + 7)/2 - (0 + 0)/2
[1] 6
> c(6 - 3.04, 6 + 3.04)
[1] 2.96 9.04
```

Further Extensions of ANOVA

- Two-way ANOVA

$$X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

  - ▷ Interactions interpreted as in analysis of covariance
  - ▷ Constraints are needed
- ANOVA for designed experiments
  - ▷ Not all combinations of levels can be observed - which should you choose in order to estimate all the effects of interest?
- Random-effects models
  - ▷ Fixed effect: a constant parameter α_j
  - ▷ Random effect: α_j is a random variable for which we want to estimate the distribution
  - ▷ Mixed-effects models: combination of fixed and random effects