Week 5

Mengye Liu

4/27/2022

Guideline:

- The random effect Model/ Model II (section 3.9 in textbook)
- Residual Plots
- Fisher LSD comparison

1. Random Effect Model

1.1 Review: Model I:

So far, we considered fixed effects model or Model I of ANOVA:

- $y_{ij} = \mu_i + \epsilon_{ij} = \mu + \tau_i + \epsilon_{ij}$ where $i = 1, \ldots, t$ and $j = 1, \ldots, n_i$
- τ_i's are **fixed** unknown treatment effects
- We are testing $H_0 : \tau_1 = \ldots = \tau_t$

1.2 Model II

Sometimes, these treatments are random selection from possible treatments.

- Example 3.10 (textbook P104): Textile factory → Looms(random) → Fabric(random). We only select 4 Looms, then select fabric samples from these selected looms.

The corresponding model formula is

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

where $\tau_i \sim N(0, \sigma_\tau^2)$ and $\epsilon_{ij} \sim N(0, \sigma^2)$. Also, τ_i and ϵ_{ij} are independent. Then $\text{Var}(y_{ij}) = \sigma_\tau^2 + \sigma^2$.

- Model II is also called 'Variance components model', 'Hierarchical model', 'Random effect model'.

Note:

- $E(y_{ij}) = \mu$ under model I and II
- $\text{Var}(y_{ij}) = \begin{cases} \sigma^2, & \text{Model I} \\ \sigma_\tau^2 + \sigma^2, & \text{Model II} \end{cases}$
- For same i(same loom in example),

$$\text{Cov}(y_{ij}, y_{ik}) = \begin{cases} 0, & \text{Model I} \\ \text{Cov}(\mu + \tau_i + \epsilon_{ij}, \mu + \tau_i + \epsilon_{ik}) = \sigma_\tau^2, & \text{Model II} \end{cases}$$

- Intraclass correlation coefficient: $\text{Corr}(y_{ij}, y_{ik}) = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma^2}$

Parameters in Model II:

- μ, σ_τ^2(viability due to treatment), σ^2(random error component)

- $H_0 : \sigma_\tau^2 = 0$ for F test in ANOVA table. Under H_0, we have
$$\frac{MS_{treatment}}{MS_{error}} \sim F_{t-1,N-t}$$
Exactly the same F as in model I.
- Still having equation: $SS_{total} = SS_{treatment} + SS_{error}$
- Model I and Model II ANOVA tables exactly the same.

Estimates of Parameters in Model II:

- $\hat{\sigma}^2 = MS_{error}$
- $\hat{\mu} = \bar{y}_{..}$
- $\hat{\sigma}_\tau^2 = \frac{MS_{treatment}-MS_{error}}{n_0}$, where $n_0 = \frac{1}{t-1}\left(N - \frac{\sum_i n_i^2}{N}\right)$.
- Getting CI for intraclass correlation coefficient(P106 textbook):
$$L = \frac{1}{n}\left(\frac{MS_{treatment}}{MS_{error}}\frac{1}{F_{1-\alpha/2,t-1,N-t}} - 1\right)$$
$$U = \frac{1}{n}\left(\frac{MS_{treatment}}{MS_{error}}\frac{1}{F_{\alpha/2,t-1,N-t}} - 1\right)$$
Then the $(1-\alpha)\%$ CI for $\frac{\sigma_\tau^2}{\sigma_\tau^2+\sigma^2}$ is
$$\left[\frac{L}{1+L}, \frac{U}{1+U}\right]$$
Taking example 3.10 data:

```
library(knitr)
opts <- options(knitr.kable.NA = "")
obs = c(98,97,99,96,91,90,93,92,96,95,97,95,95,96,99,98)
looms = factor(rep(1:4, each = 4))
fit = aov(obs~looms)
summ = summary(fit)[[1]]
kable(summ)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
looms	3	89.1875	29.729167	15.68132	0.0001878
Residuals	12	22.7500	1.895833		

Then estimate all the parameters

```
N = length(obs)
ni = table(looms)
t = length(levels(looms))
n0 = 1/(t-1)*(N - sum(ni^2)/N)
# here is balanced case, so we can use n0 = n = 4 directly
mu = mean(obs)
sigma = summ['Residuals','Mean Sq']
sigma_tau = (summ['looms','Mean Sq'] - sigma)/n0
result1 = c(mu,sigma, sigma_tau)
names(result1) = c('$\\hat\\mu$', '$\\hat \\sigma^2$', '$\\hat\\sigma^2_\\tau$')
kable(result1)
```

	x
$\hat{\mu}$	95.437500
$\hat{\sigma}^2$	1.895833
$\hat{\sigma}_{\tau}^2$	6.958333

2. Residual Plots

(This part is included in textbook section 3.4.)

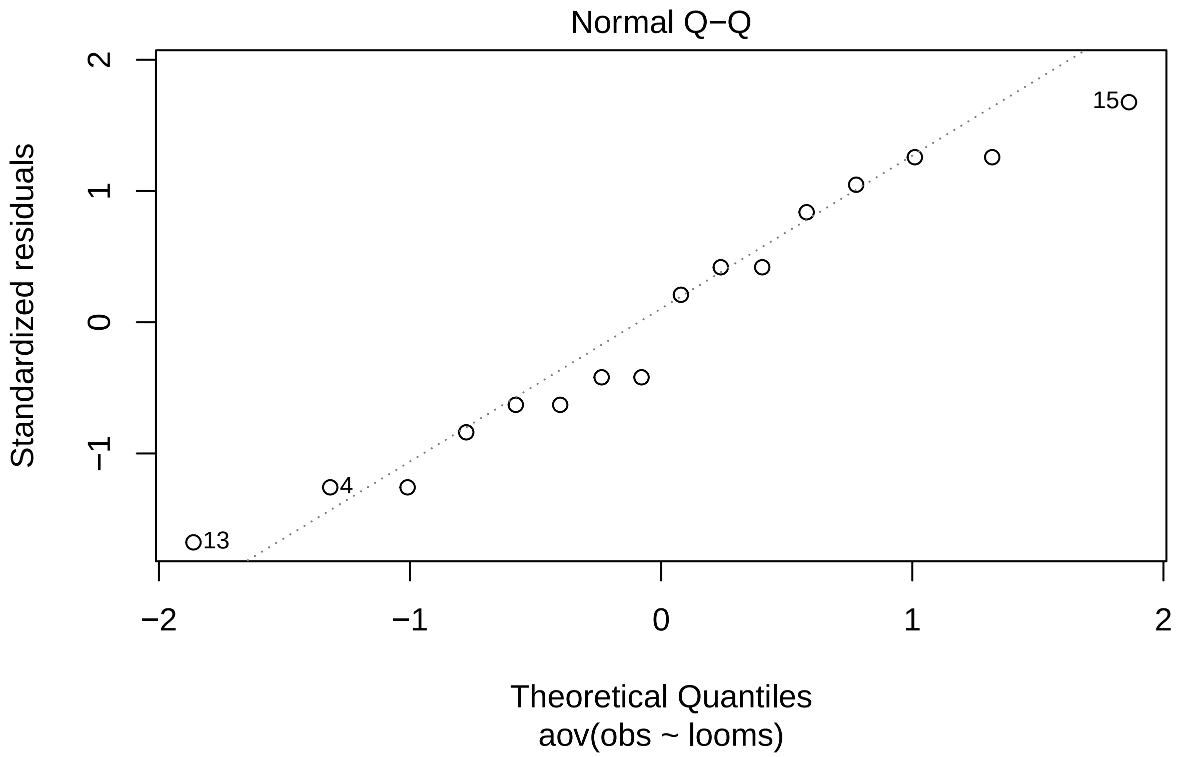
Recall:

- we have assumptions for ANOVA models:
  - Normality: ϵ_{ij} follows normal distribution
  - Constant varaince: $\text{Var}(\epsilon_{ij}) = \sigma^2$. σ^2 remains a constant for all i, j.
  - Independence: ϵ_{ij} are independent.
- Model adequacy: If the model is adequate, the residuals should be structureless. That is there is no obvious pattern in residual plots.

2.1 Q-Q plot of residuals

Normal Q-Q plot of residuals can help us check the normality assumption.

```
plot(fit, which = 2)
```



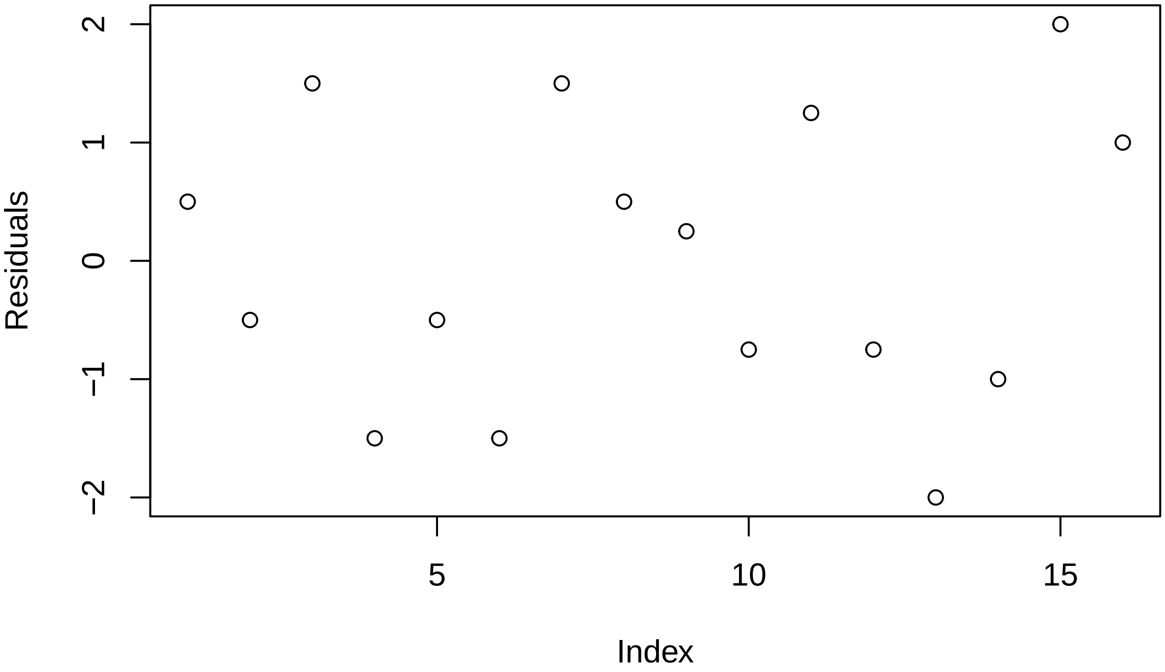
Thing to check: the points lie on (or close to) the dashed line.

In the plot we can see that almost all the points are closed to dashed straight line. Thus there is no severe violation of normal assumption. That is ϵ follows normal distribution.

2.2 Residual vs index

This plot can help us check independency assumption.

```r
plot(fit$residuals, ylab = 'Residuals')
```



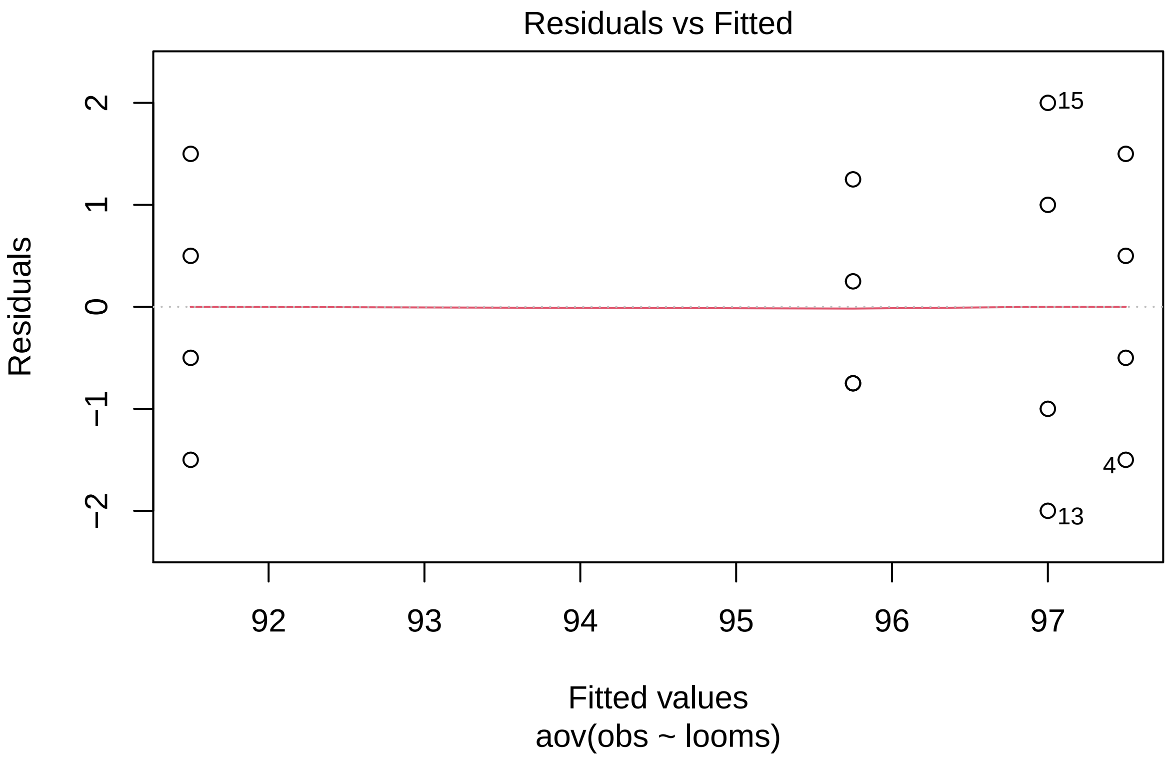
Thing to check: The residual points show any pattern and trends.

In plot above, points spread randomly. Thus there is no strong correlation between residuals. In other words, the independence assumption on the errors has not been violated.

2.3 Residual vs fitted values

This plot can help us check constant variance assumption and model adequacy.

```r
plot(fit, which = 1)
```



Things to check:

- Points are symmetric.
- Points spread in a constant band (Constant variance). There is no outward(inward)-opening funnel pattern. Outward(Inward)-opening funnel means increasing(decreasing) variance along fitted values or non-normal distribution.
- The red line is horizontal. No obvious curve or linear trend.

In above plot, residual points are symmetric around $y = 0$ and spread in a constant band. The red line is almost horizontal which means there no specific pattern exists. Thus the constant varince assumption fits and the model is adequate.

3. Fisher LSD comparison

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
Strength = c(3129, 3000, 2865, 2890,3200, 3300, 2975,
3150,2800, 2900, 2985, 3050, 2600, 2700, 2600, 2765)
Tech = factor(c(1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,4))
data = data.frame(strength = Strength, tech = Tech)
fit1 = aov(strength~tech, data = data)
```

Fisher LSD

To get the Fisher LSD criterion

$$|\bar{y}_{i.} - \bar{y}_{j.}| > t_{1-\alpha/2,N-t}\sqrt{MS_{error}(\frac{1}{n_i} + \frac{1}{n_j})}$$

Yes ⇔ reject $H_0 : \mu_i = \mu_j$ ⇔ difference between ith and jth level is significant.

```r
N = length(data$strength)
t = length(levels(data$tech))
#balanced case
n = N/t
MSE = summary(fit1)[[1]][2,3]
# For alpha = 0.05, balanced case
fish = qt(0.975,N-t)*sqrt(MSE*(2/n))
#getting yi._bar
g = data %>%
  group_by(tech) %>%
  summarise(m = mean(strength), .groups = 'drop')
#Getting pairwise differences
diff = as.numeric(dist(g$m))
#checking the criteria
diff>fish
```

```
## [1]  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
```