STAT 516 Lec 05

One-way analysis of variance (ANOVA)

Karl Gregory

2025-02-25

Rust inhibitors example

Data from Kutner et al. (2005).

Ten experimental units assigned to each of four brands of rust inhibitors.

```r
link <- url("https://people.stat.sc.edu/gregorkb/data/KNNLrust.txt")
rust <- read.csv(link,col.names=c("score","brand","rep"),sep = "", header = FALSE)
head(rust)
```

```
  score brand rep
1  43.9     1   1
2  39.0     1   2
3  46.7     1   3
4  43.8     1   4
5  44.2     1   5
6  47.7     1   6
```

Do the brands differ in effectiveness? Is there a best brand?

Randomized experiments comparing treatments

Rust data: $N = 40$

$a = 4$

total # EUs

Start with N experimental units (EUs), e.g. subjects, mice, etc.

Randomly assign each EU to one of a treatment groups. # treatment groups

Measure on each EU after treatment a response Y.

Compute the average of the responses in each treatment group...

Questions we'd like to answer:

- Is the response mean the same in all treatment groups?
- If not, then which pairs of means are different?

One-way ANOVA setup

Consider the model

$$Y_{ij} = \mu + \overset{\text{"tau"}}{\tau_i} + \overset{\text{Error term}}{\varepsilon_{ij}}, \quad j = 1, \dots, \overset{\text{\# EUs in treatment group } i}{n_i}, \quad i = 1, \dots, a,$$

where

- Y_{ij} is the response for EU j in treatment group i.
- μ represents ~~an overall or~~ a baseline mean.
- τ_i is the treatment effect for treatment i.
- The ε_{ij} are independent Normal$(0, \sigma^2)$ error terms.
- The n_i are the numbers of replicates in the treatment groups.

RUST: $N = 40$

$a = 4$

$$\left.\begin{array}{l} n_1 = 10 \\ n_2 = 10 \\ n_3 = 10 \\ n_4 = 10 \end{array}\right\} \text{"balanced"}$$

Of central interest are the hypotheses

H_0: $\tau_i = 0$ for all i versus H_1: At least one τ_i is nonzero.

If we reject H_0, we may wish to sort/compare the treatments.

Identifiability constraint in the treatment effects model

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \qquad i = 1, \ldots, a$$

$$j = 1, \ldots, n_i \quad ,$$

We want to estimate $\underbrace{\mu, \tau_1, \ldots, \tau_a}_{a+1}$ and σ^2.

The model has $a + 1$ parameters to describe a treatment means.

To identify μ, $\tau_1, \ldots, \tau_a$ uniquely, we typically set $\boxed{\tau_1 = 0.}$

↑

This is what most software does.

Alternative “cell means model” setup

An alternate version of the model is “cell means”

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, a,$$

where

Estimate as $\hat{\mu}_i = \bar{Y}_{i.}$

- Y_{ij} is the response for EU j in treatment group i.
- μ_i represents the mean of treatment group i.
- The ε_{ij} are error terms distributed as Normal$(0, \sigma^2)$.

In this version of the model the central hypotheses become

$$H_0\colon \mu_1 = \cdots = \mu_a \quad \text{versus} \quad H_1\colon \mu_i \neq \mu_i' \text{ for some } i \neq i'.$$

Goals in one-way ANOVA

Under the one-way ANOVA setup

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, a,$$

where $\varepsilon_{ij} \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2)$, we wish to

1. ✓ Visualize the data.
2. ✓ Estimate the parameters $\mu, \tau_1, \dots, \tau_a$.
3. ✓ Estimate the error term variance σ^2.
4. ✓ Decompose the variation in the Y_{ij} as signal plus noise.
5. ✓ Test whether there is any difference in treatment group means.
6. Sort/compare the treatment means if there is any difference.
7. Check whether the model assumptions are satisfied.

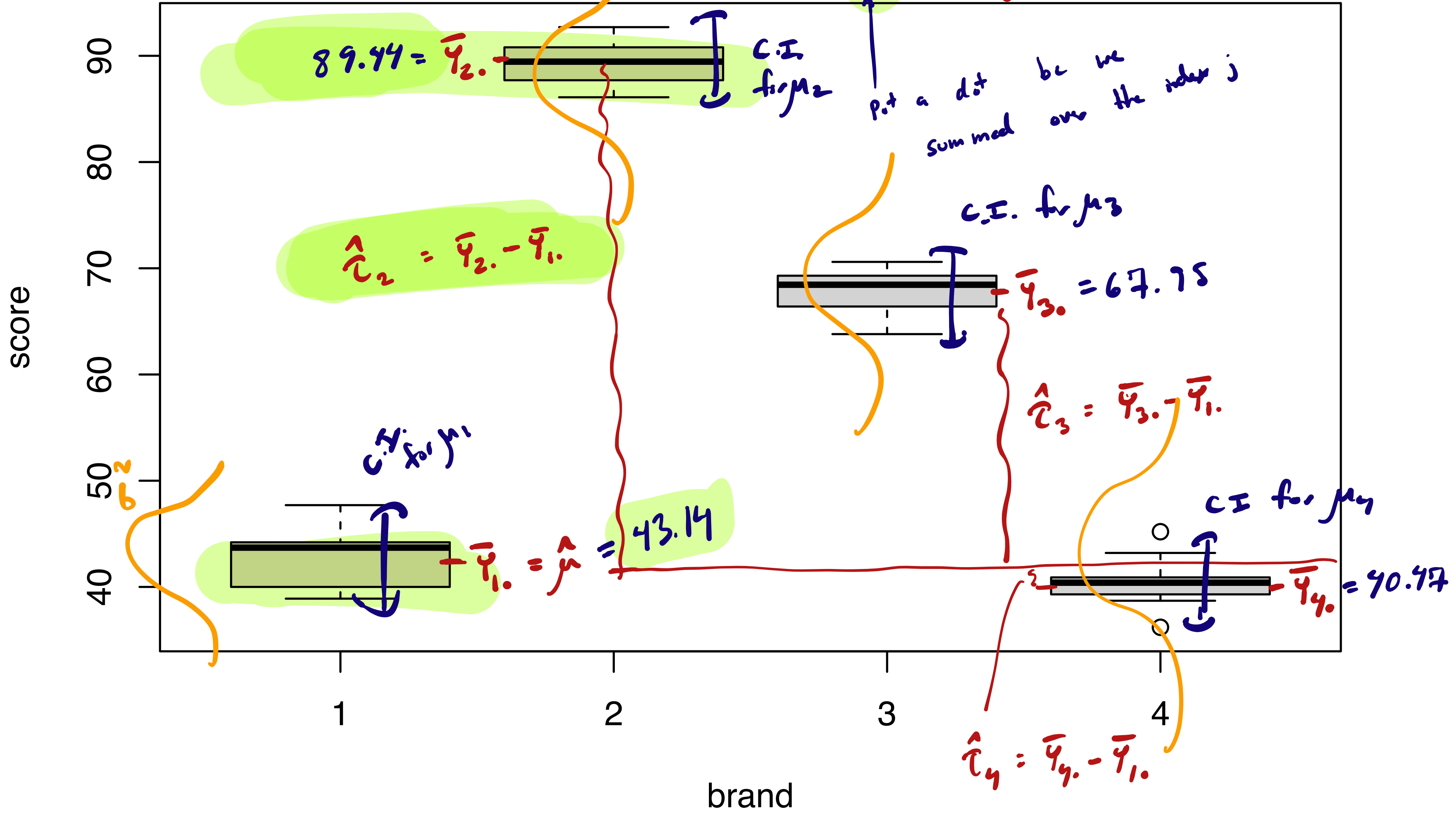
Rust inhibitors example (cont)

Visually compare the means of several treatment groups with boxplots.

```
boxplot(score ~ brand, data = rust)
```

$Y_{ij} = \mu_i + \varepsilon_{ij}$

$\bar{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$



Treatment effect estimation in one-way ANOVA

$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$

- For each $i = 1, \dots, a$ define the observed treatment group mean as

Uses $\bar{Y}_{1.}$ as baseline

$$\bar{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}.$$

- Then, setting $\tau_1 = 0$, estimate μ and $\tau_2, \dots, \tau_a$ as

$$\boxed{\hat{\mu} = \bar{Y}_{1.}} \quad \text{and} \quad \hat{\tau}_i = \bar{Y}_{i.} - \bar{Y}_{1.} \quad \text{for } i = 2, \dots, a.$$

- So treatment group 1 is regarded as a baseline, where:
  1. The baseline has estimated mean $\hat{\mu}$.
  2. The estimates $\hat{\tau}_2, \dots, \hat{\tau}_a$ are deviations from the baseline.

- One obtains the fitted values $\hat{Y}_{ij} = \hat{\mu} + \hat{\tau}_i = \bar{Y}_{i.}$ for $i = 1, \dots, a$.

$\bar{Y}_{1.} + (\bar{Y}_{i.} - \bar{Y}_{1.}) = \bar{Y}_{i.}$

Fitted values are just the group means.

Rust inhibitors example (cont)

Use `lm()` with `as.factor()` to fit the one-way ANOVA model.

```r
# use as.factor() to designate brand as a "factor"
lm_out <- lm(score ~ as.factor(brand), data = rust)
lm_out
```

```
Call:
lm(formula = score ~ as.factor(brand), data = rust)

Coefficients:
      (Intercept)  as.factor(brand)2  as.factor(brand)3  as.factor(brand)4
            43.14              46.30              24.81              -2.67
```

$\hat{\mu} = \bar{Y}_{1.}$

$\hat{\tau}_2 = \bar{Y}_{2.} - \bar{Y}_{1.}$
$= 89.44 - 43.14$

$\hat{\tau}_3 = \bar{Y}_{3.} - \bar{Y}_{1.}$
$= 67.95 - 43.14$

$\hat{\tau}_4 = \bar{Y}_{4.} - \bar{Y}_{1.}$
$= 40.47 - 43.14.$

See how $\hat{\mu}, \hat{\tau}_2, \hat{\tau}_3, \hat{\tau}_4$ are related to $\bar{Y}_{1\cdot}, \bar{Y}_{2\cdot}, \bar{Y}_{3\cdot}, \bar{Y}_{4\cdot}$.

```r
# compute the group means
aggregate(rust$score, by = list(rust$brand), FUN = mean)
```

```
  Group.1     x
1       1 43.14
2       2 89.44
3       3 67.95
4       4 40.47
```

$\leftarrow \bar{Y}_{1\cdot}$

$\leftarrow \bar{Y}_{2\cdot}$

$\leftarrow \bar{Y}_{3\cdot}$

$\leftarrow \bar{Y}_{4\cdot}$

Estimation of the error term variance σ^2

As in linear regression, define the

- fitted values $\hat{Y}_{ij}$ as $\hat{Y}_{ij} = \bar{Y}_{i.}$ for $j = 1, \dots, n_i$, and the
- residuals $\hat{\varepsilon}_{ij}$ as $\hat{\varepsilon}_{ij} = Y_{ij} - \bar{Y}_{i.}$

$$\hat{\varepsilon}_{ij} = Y_{ij} - \hat{Y}_{ij}$$
$$= Y_{ij} - \bar{Y}_{i.}$$

for $j = 1, \dots, n_i$, $i = 1, \dots, a$.

Then an unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{N-a} \sum_{i=1}^{a} \sum_{j=1}^{n_i} \hat{\varepsilon}_{ij}^2 = \frac{1}{N-a} \sum_{i=1}^{a} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2.$$

Divide by $N - a$ since the N residuals depend on a estimated quantities...

number of EUs, so $N = n_1 + \dots + n_a$

Rust inhibitors example (cont)

```r
tab <- cbind(rust$brand,rust$score,lm_out$fitted.values,lm_out$residuals)
colnames(tab) <- c("brand","score","Fitted value","Residual")
head(tab,n = 13)
```

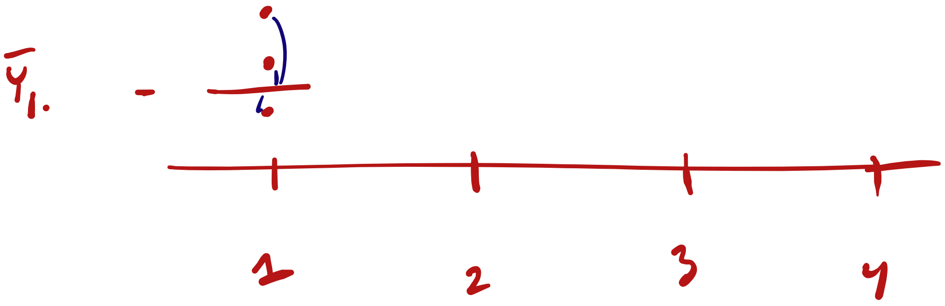
i Y_{ij} $\hat{Y}_{ij} = \bar{Y}_{i.}$ $Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \bar{Y}_{i.}$

```
   brand score Fitted value Residual
1      1  43.9        43.14     0.76
2      1  39.0        43.14    -4.14
3      1  46.7        43.14     3.56
4      1  43.8        43.14     0.66
5      1  44.2        43.14     1.06
6      1  47.7        43.14     4.56
7      1  43.6        43.14     0.46
8      1  38.9        43.14    -4.24
9      1  43.6        43.14     0.46
10     1  40.0        43.14    -3.14
11     2  89.8        89.44     0.36
12     2  87.1        89.44    -2.34
13     2  92.7        89.44     3.26
```

$= 43.90 - 43.14$

$= 39.00 - 43.14$

mean is 43.14



```r
sgsqhat <- sum(lm_out$residuals^2) / (nrow(rust) - 4)
sgsqhat
```

```
[1] 6.139833
```

$\hat{\sigma}^2$

The value of $\hat{\sigma}$ is printed in the summary() output:

```r
summary(lm_out)
```

```
Call:
lm(formula = score ~ as.factor(brand), data = rust)

Residuals:
   Min     1Q Median     3Q    Max 
-4.270 -1.597  0.395  1.275  4.730 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        43.1400     0.7836  55.056   <2e-16 ***
as.factor(brand)2  46.3000     1.1081  41.782   <2e-16 ***
as.factor(brand)3  24.8100     1.1081  22.389   <2e-16 ***
as.factor(brand)4  -2.6700     1.1081  -2.409   0.0212 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.478 on 36 degrees of freedom
Multiple R-squared:  0.9863,    Adjusted R-squared:  0.9852 
F-statistic: 866.1 on 3 and 36 DF,  p-value: < 2.2e-16
```

$N = 40$

$a = 4$

$\bar{y}_{1\cdot}$

$\hat{\tau}_2, \hat{\tau}_3, \hat{\tau}_4$

Irrelevant for us.

$N - a = 36$

$\hat{\sigma}$

so $\hat{\sigma}^2 = (2.478)^2$

Important

Sums of squares in the one-way ANOVA model

As in linear regression we decompose the variation in the Y_{ij} by defining:

- Total sum of squares: $\mathrm{SS}_{\mathrm{Tot}} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2$

(Basically just "Regression SS")

- Treatment sum of squares: $\mathrm{SS}_{\mathrm{Trt}} = \sum_{i=1}^{a} n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2 = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (\hat{Y}_{ij} - \bar{Y}_{..})^2$

$\bar{Y}_{i.}$

$= \sum_{i=1}^{a} \sum_{j=1}^{n_i} (\bar{Y}_{i.} - \bar{Y}_{..})^2$

$= \sum_{i=1}^{a} n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2$

- Error sum of squares: $\mathrm{SS}_{\mathrm{Error}} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2$

$= \sum_{i=1}^{a} \sum_{j=1}^{n_i} (Y_{ij} - \hat{Y}_{ij})^2$

$\bar{Y}_{i.}$

In the above, $\bar{Y}_{..}$ denotes the overall mean, defined as

$$\bar{Y}_{..} = N^{-1} \sum_{i=1}^{a} \sum_{j=1}^{n_i} Y_{ij}, \quad \text{where } N = n_1, \dots, n_a.$$

The mean of all the response values.

We have $\mathrm{SS}_{\mathrm{Tot}} = \mathrm{SS}_{\mathrm{Trt}} + \mathrm{SS}_{\mathrm{Error}}$.

Note that $\mathrm{SS}_{\mathrm{Trt}}$ is computed just like $\mathrm{SS}_{\mathrm{Reg}}$ in linear regression.

We again define $R^2 = \dfrac{\mathrm{SS}_{\mathrm{Trt}}}{\mathrm{SS}_{\mathrm{Tot}}}$.

Sampling distributions of our sums of squares

The SS, appropriately scaled, follow chi-square distributions:

- $\mathrm{SS}_{\mathrm{Tot}}/\sigma^2 \sim \chi^2_{N-1}(\phi_{\mathrm{Tot}})$
- $\mathrm{SS}_{\mathrm{Trt}}/\sigma^2 \sim \chi^2_{a-1}(\phi_{\mathrm{Trt}})$
- $\mathrm{SS}_{\mathrm{Error}}/\sigma^2 \sim \chi^2_{N-a}$,

SS_{Tot} has $df = N-1$

SS_{Trt} has $df = a-1$

SS_{Error} has $df = N-a$

where ϕ_{Tot} and ϕ_{Trt} are noncentrality parameters.

The mean squares in the one-way ANOVA model

Dividing $\mathrm{SS}_{\mathrm{Trt}}$ and $\mathrm{SS}_{\mathrm{Error}}$ by their dfs, we define:

- Treatment mean square: $\mathrm{MS}_{\mathrm{Trt}} = \dfrac{\mathrm{SS}_{\mathrm{Trt}}}{a-1}$
- Error mean square: $\mathrm{MS}_{\mathrm{Error}} = \dfrac{\mathrm{SS}_{\mathrm{Error}}}{N-a}$

The ratio $F_{\mathrm{stat}} = \dfrac{\mathrm{MS}_{\mathrm{Trt}}}{\mathrm{MS}_{\mathrm{Error}}}$ has an F distribution.

The Analysis of Variance (ANOVA) table

We often present the SS, df, and MS values in a table like this:

Source	Df	SS	MS	F value	p-value
Treatment	$a-1$	SS_{Trt}	MS_{Trt}	F_{stat}	$P(F > F_{\text{stat}})$
Error	$N-a$	SS_{Error}	MS_{Error}		
Total	$N-1$	SS_{Tot}			

In the table $F_{\text{stat}} = \dfrac{\text{MS}_{\text{Trt}}}{\text{MS}_{\text{Error}}}$.

The p-value is based on $F \sim F_{a-1,N-a}$.

Under $H_0: \mu_1 = \cdots = \mu_a$ we have $F_{stat} \sim F_{a-1, N-a}$

Rust inhibitors example (cont)

$N = 40$

$a = 4$

Obtain the ANOVA table with the `anova()` function on the `lm()` output.

```r
anova(lm_out)
```

```
Analysis of Variance Table

Response: score
                   Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(brand)    3  15954  5317.8  866.12 < 2.2e-16 ***
Residuals          36    221     6.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Treatment

Error

$a-1 = 4-1$

$15954/3$

$221/36$

$\frac{5317.8}{6.1}$

$N-a$

Source	Df	SS	MS	F value	p-value
Treatment	$a-1$	SS_{Trt}	MS_{Trt}	F_{stat}	$P(F > F_{\mathrm{stat}})$
Error	$N-a$	SS_{Error}	MS_{Error}		
~~Total~~	~~$N-1$~~	~~SS_{Tot}~~			

Testing whether there is any difference in treatment means

Cell means model:

mean of group i

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad , \; i = 1, \ldots, a$$

Treatment effects model

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

baseline

effect of treatment i

In the one-way ANOVA model we wish to test

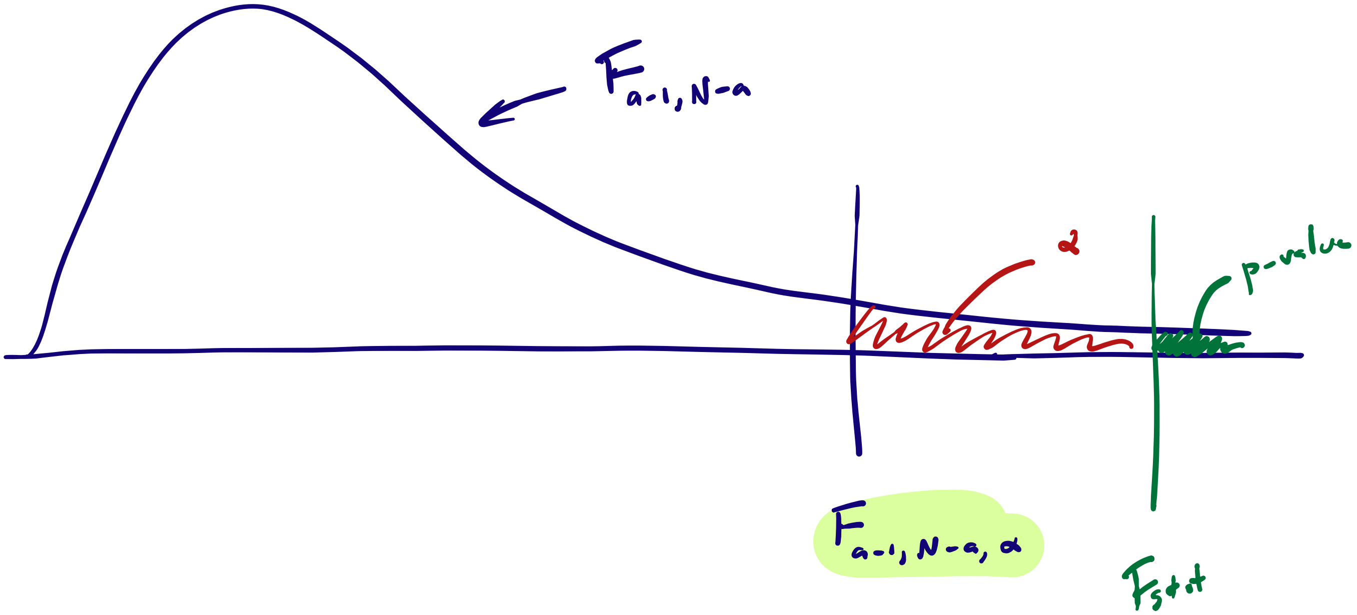
Treatment effects: H_0: $\tau_i = 0$ for all i versus H_1: At least one τ_i is nonzero.

Cell means: $H_0 : \mu_1 = \cdots = \mu_a$ versus $H_1 : \mu_i \neq \mu_j$ for some $i \neq j$ (Not all μ_i are equal).

We use the overall F test of significance:

1. Compute $F_{\text{stat}} = \dfrac{\text{MS}_{\text{Trt}}}{\text{MS}_{\text{Error}}}$
2. Reject H_0 at α if $F_{\text{stat}} > F_{a-1,N-a,\alpha}$.
3. Obtain p-value as $P(F > F_{\text{stat}})$, where $F \sim F_{a-1,N-a}$.

The value of F_{stat} and the p-value are printed in the `summary()` output.



Reject $H_0: \mu_1 = \cdots = \mu_a$ when F_{stat} is large.... why?

$$F_{stat} = \frac{MS_{Trt}}{MS_{Error}} = \frac{SS_{TRT}/(a-1)}{SS_{Error}/(N-a)}$$

$$= \frac{\sum_{i=1}^{a} n_i \left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)^2 / (a-1)}{\sum_{i=1}^{a}\sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2 / (N-a)}$$

$$= \frac{\text{Variation among group means } \bar{Y}_{1\cdot}, \dots, \bar{Y}_{a\cdot}}{\text{Variation in individual responses around the group means}}$$

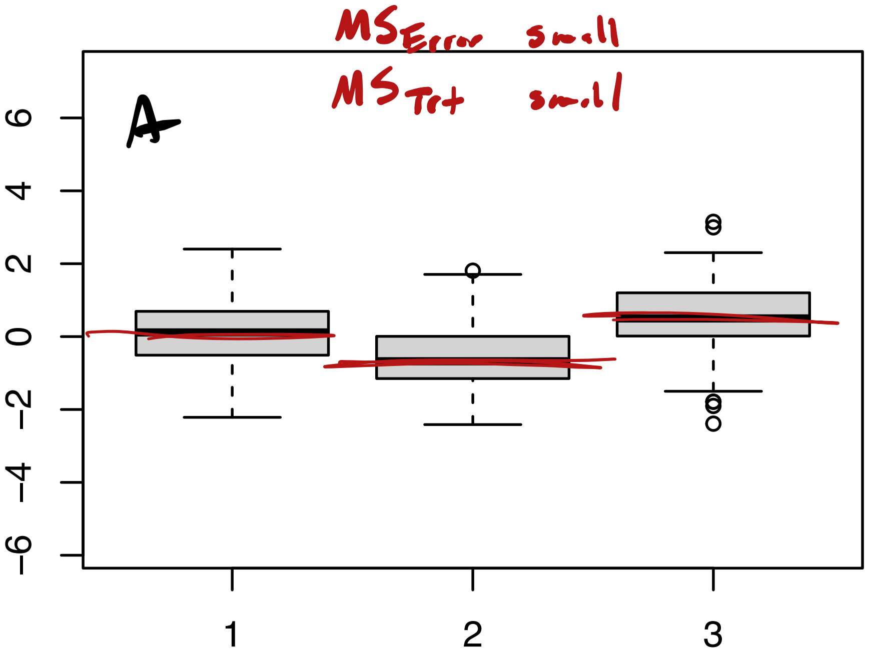
$$= \frac{\text{Between-group variation}}{\text{Within-group variation}}$$

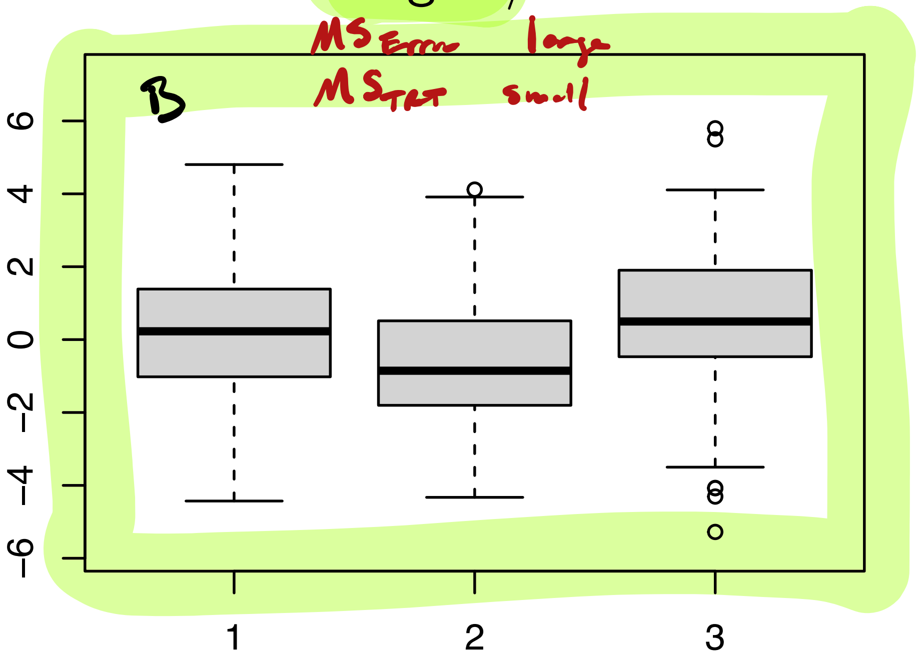
Interpretation of F statistic

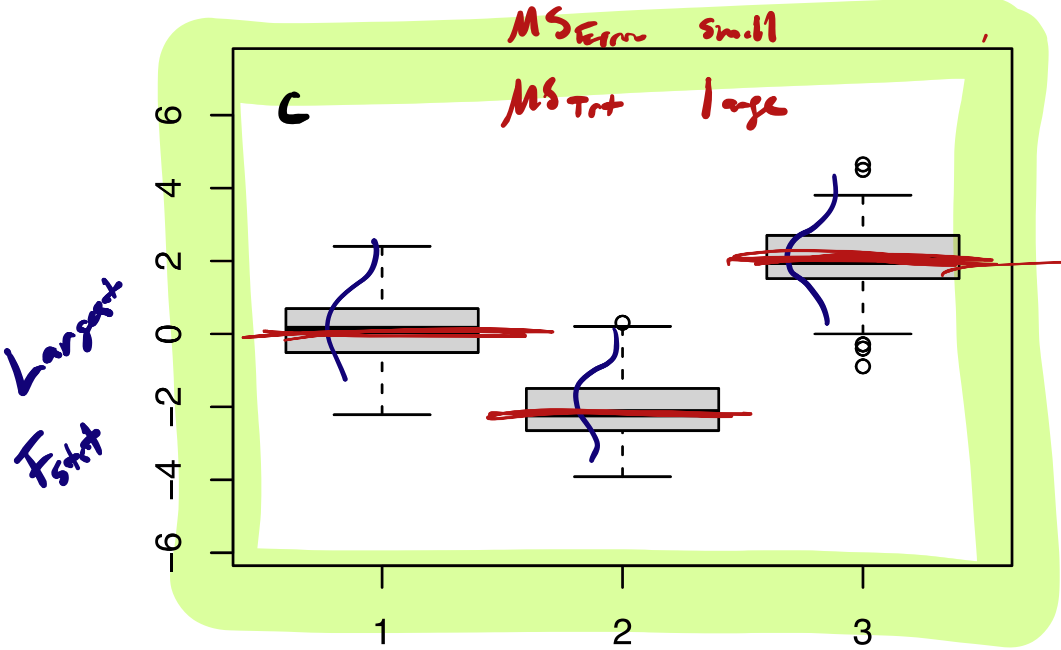
Note that F_{stat} is a ratio of the form $\dfrac{\text{Between treatment variation}}{\text{Within treatment variation}}$.

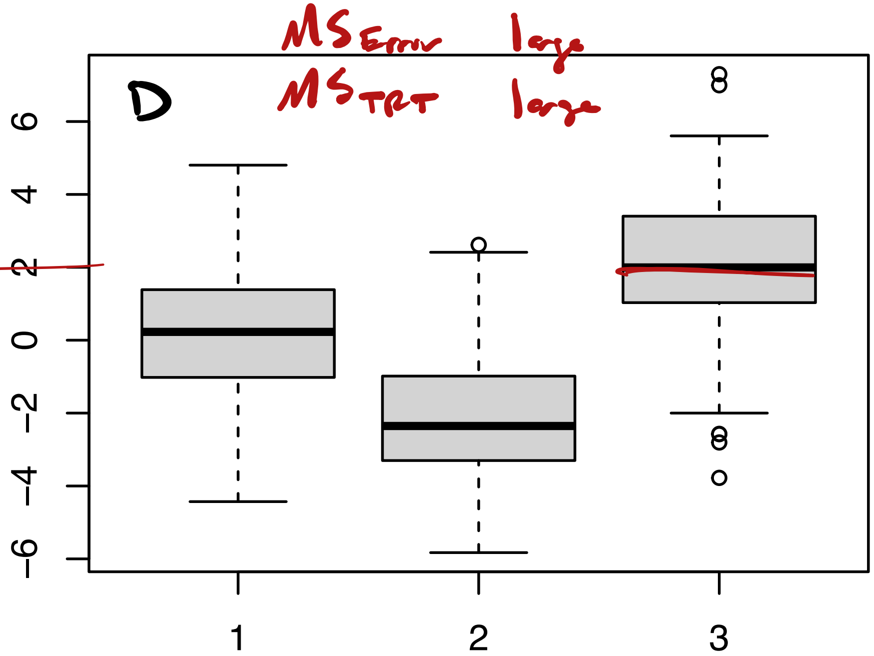
Exercise: For which data set will the F-statistic be largest/smallest?

Smallest F









Exercise: Compute F_{stat} for the rust data using the summary info:

$\bar{Y}_{1.}$

$a=4$

group	replicates	mean	standard deviation
1	$n_1=$ 10	43.14	3.00
2	$n_2=$ 10	89.44	2.22
3	$n_3=$ 10	67.95	2.17
4	$n_4=$ 10	40.47	2.44

$N=40$

$(3.00)^2 = S_1^2$, $9(3.00)^2 = \sum_{j=1}^{10} (Y_{1j} - \bar{Y}_{1.})^2$

p-value

Hint: $\text{SS}_{\text{Error}} = \sum_{i=1}^{a} \overset{9}{\boxed{(n_i - 1)}} S_i^2$, where $S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2$

$$F_{\text{stat}} = \frac{MS_{\text{Trt}}}{MS_{\text{Error}}} = \frac{SS_{\text{Trt}}/(a-1)}{SS_{\text{Error}}/(N-a)} = \frac{\sum_{i=1}^{a} n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2 / (a-1)}{\sum_{i=1}^{a} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 / (N-a)}$$

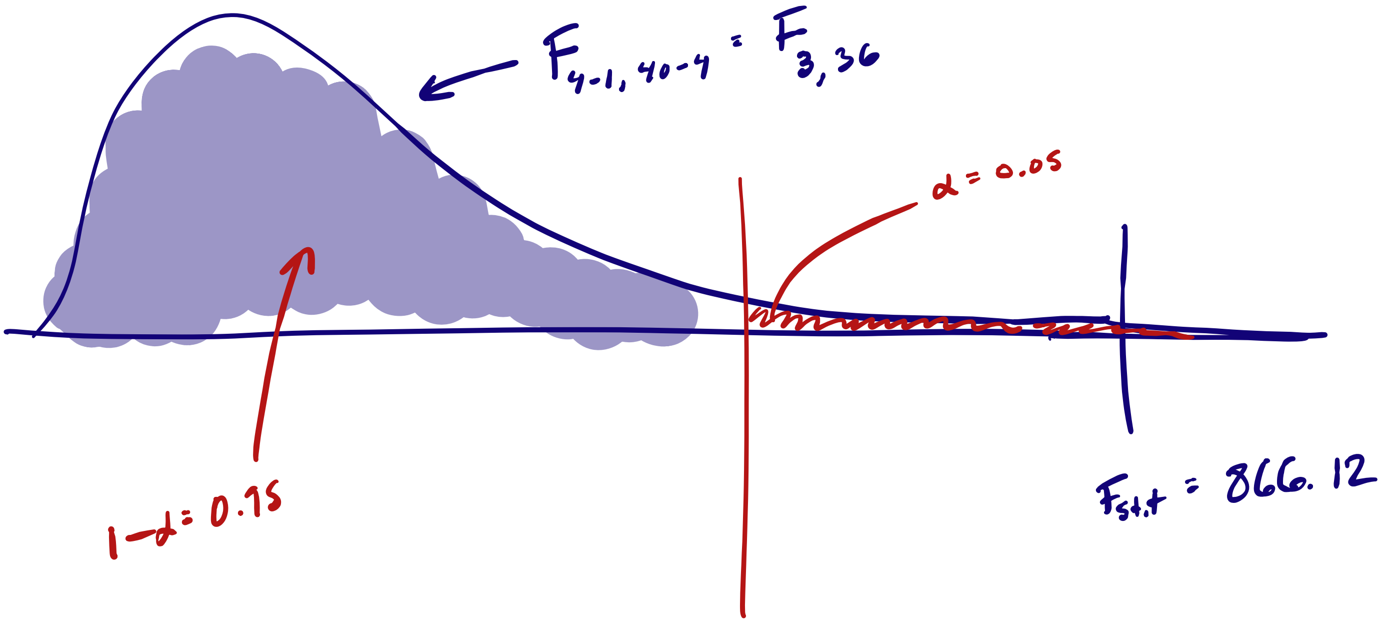
$$SS_{Error} = 9\cdot(3.00)^2 + 9\cdot(2.22)^2 + 9(2.17)^2 + 9(2.44)^2 = 221$$

$$\bar{y}_{..} = (43.14 + 89.44 + 67.75 + 40.47)/4 = 60.25$$

$$SS_{TRT} = \sum_{i=1}^{a} n_i(\bar{y}_{i.} - \bar{y}_{..})^2 = 10(43.14 - 60.25)^2$$
$$+ 10(89.44 - 60.25)^2$$
$$+ 10(67.75 - 60.25)^2$$
$$+ 10(40.47 - 60.25)^2$$

$$= 15{,}954$$

$$F_{stat} = \frac{15{,}954/(4-1)}{221/(40-4)} = \underline{866.12}$$



$$F_{3,36,0.05} = qf(0.95, 3, 36)$$
$$= 2.87$$

Some CI formulas (without familywise adjustment)

In the cell-means formulation of the model

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, a,$$

where $\mu_i = \mu + \tau_i$, we have the following CI formulas:

$$\hat{\sigma}^2 = \frac{1}{N-a} \sum_{i=1}^{a} \sum_{j=1}^{n_i} \hat{\varepsilon}_{ij}^2$$

Target	$(1-\alpha)100\%$ confidence interval
μ_i	$\bar{Y}_{i.} \pm t_{N-a,\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n_i}}$
$\mu_i - \mu_{i'}$	$\bar{Y}_{i.} - \bar{Y}_{i'.} \pm t_{N-a,\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}}$

Resemble C.I. for $\mu_1 - \mu_2$ in 509/515

Compare to Lecture 1:

$$\bar{X}_n \pm t_{n-1,\alpha/2} \frac{s_n}{\sqrt{n}}$$

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

Build C.I. (95%) for μ_1:

$\hat{\sigma}^2 = 6.14$ $\qquad$ $n_1 = 10$

$\bar{Y}_{1.} = 43.14$ $\qquad$ $N - a = 40 - 4 = 36$

$\alpha = 0.05$

$\alpha/2 = 0.025$

$$\bar{Y}_{1.} \pm \underbrace{t_{N-a,\,\alpha/2}} \frac{\hat{\sigma}}{\sqrt{n_1}}$$

qt(0.975, 36)

"

qt(0.025, 36, lower.tail = F) = 2.028

$$\text{C.I.:} \quad 43.14 \pm 2.028 \sqrt{\frac{6.14}{10}} = (41.55, 44.73)$$

C.I. for $\mu_2 - \mu_1$:

$$\bar{Y}_{2.} - \bar{Y}_{1.} \pm 2.028 \sqrt{6.14} \sqrt{\frac{1}{10} + \frac{1}{10}}$$

$$= (44.05, 48.55)$$

Rust inhibitors example (cont)

Compute 95% CIs for $\underline{\mu_1}$ and $\mu_2 - \mu_1$.

```
alpha <- 0.05
lo1 <- y1bar - qt(1-alpha/2,N-a) * sqrt(sgsqhat) / sqrt(n1)
up1 <- y1bar + qt(1-alpha/2,N-a) * sqrt(sgsqhat) / sqrt(n1)
c(lo1,up1)
```

```
[1] 41.55084 44.72916
```

```
lo21 <- y2bar - y1bar - qt(1-alpha/2,N-a) * sqrt(sgsqhat) * sqrt(1/n1 + 1/n2)
up21 <- y2bar - y1bar + qt(1-alpha/2,N-a) * sqrt(sgsqhat) * sqrt(1/n1 + 1/n2)
c(lo21,up21)
```

```
[1] 44.05259 48.54741
```

Post-hoc comparisons of means

- If we reject H_0: $\mu_1 = \cdots = \mu_a$, then we may wish to compare means.
- Call such comparisons post-hoc as we do them *after* the F-test.
- We may wish to compare several pairs of means, which is like testing several hypotheses at once.
- When several hypotheses are tested at once, the familywise Type I error rate is the probability that *any* Type I error is committed.
- We discuss two methods for post-hoc comparisons of means which control the familywise Type I error rate.

Comparing all pairs of means

- We want to build a CI for $\mu_i - \mu_{i'}$ for all pairs $i \neq i'$.
- Suppose the design is <u>balanced</u>, i.e. $n_i = n$ for all $i = 1, \dots, a$.
- If we build for all $i \neq i'$ the ordinary $(1-\alpha) \times 100\%$ CIs

  $$\bar{Y}_{i.} - \bar{Y}_{i'.} \pm t_{a(n-1),\alpha/2} \hat{\sigma} \sqrt{2/n},$$

  each one will cover its target with probability $1 - \alpha$.
- But now we want *simultaneous* coverage with probability $1 - \alpha$, i.e.

  $$P(\cap_{i \neq i'} \{\text{CI for } \mu_i - \mu_{i'} \text{ captures target}\}) = 1 - \alpha.$$

- Above probability is called the <u>familywise coverage</u>.

The venerable John Tukey

Figure 1: John Tukey, 1915 – 2000

Multiple comparisons of means with Tukey's HSD

- Suppose the design is balanced, i.e. $n_i = n$ for all $i = 1, \ldots, a$.
- Suppose we could find the value $q_{a,a(n-1),\alpha}$ such that

$$P\left(\max_{i \neq i'}\left\{\frac{|(\bar{Y}_{i.} - \bar{Y}_{i'.}) - (\mu_i - \mu_{i'})|}{\hat{\sigma}/\sqrt{n}}\right\} \leq q_{a,a(n-1),\alpha}\right) = 1 - \alpha.$$

- Then with probability $1 - \alpha$ the CIs

$$\bar{Y}_{i.} - \bar{Y}_{i'.} \pm q_{a,a(n-1),\alpha}\hat{\sigma}/\sqrt{n}$$

  will simultaneously cover the targets $\mu_i - \mu_{i'}$ for all $i \neq i'$. **Show!**
- Tukey made tables of the values $q_{a,a(n-1),\alpha}$.
- Can use the simultaneous intervals to sort/compare the means.

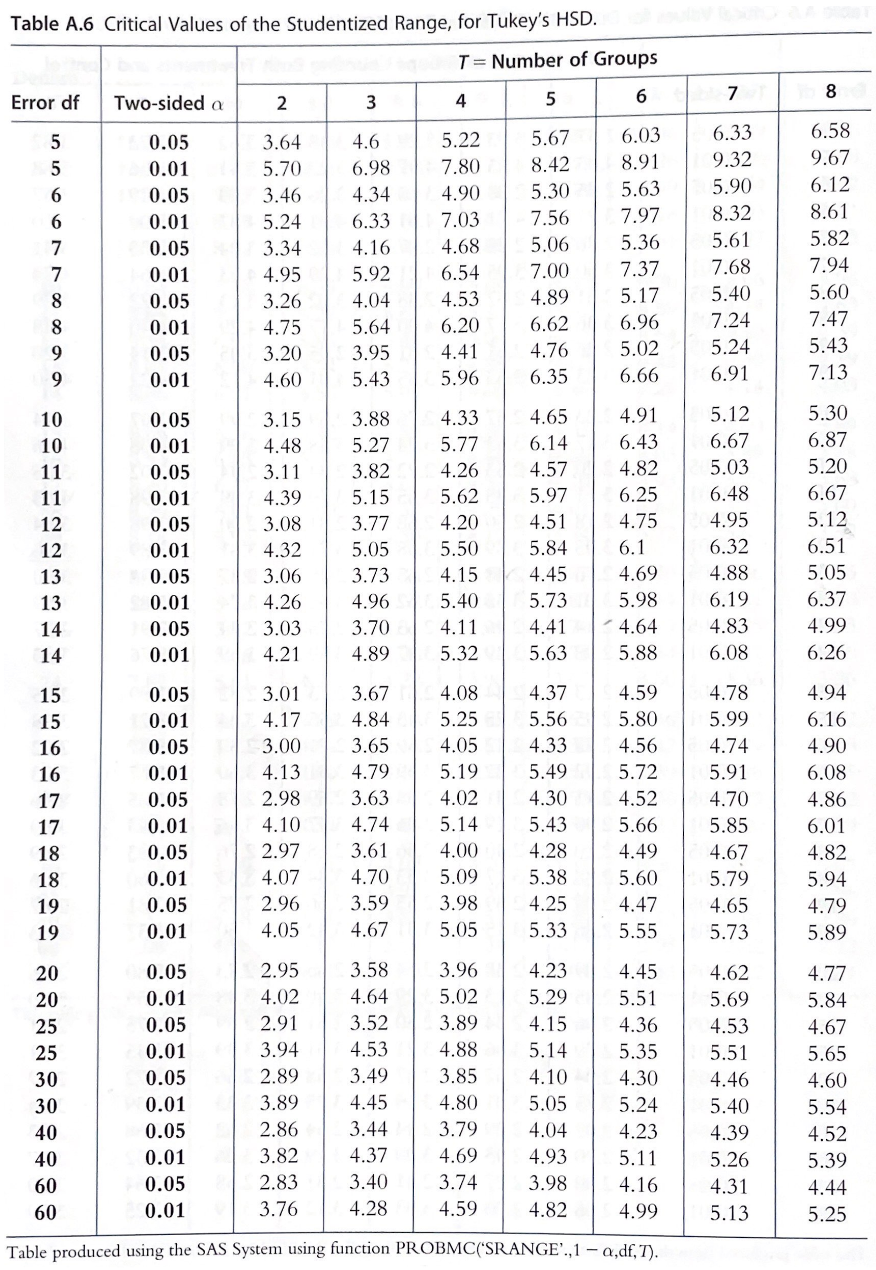


Table A.6 Critical Values of the Studentized Range, for Tukey's HSD.

Error df	Two-sided α	T = Number of Groups						
		2	3	4	5	6	7	8
5	**0.05**	3.64	4.6	5.22	5.67	6.03	6.33	6.58
5	**0.01**	5.70	6.98	7.80	8.42	8.91	9.32	9.67
6	**0.05**	3.46	4.34	4.90	5.30	5.63	5.90	6.12
6	**0.01**	5.24	6.33	7.03	7.56	7.97	8.32	8.61
7	**0.05**	3.34	4.16	4.68	5.06	5.36	5.61	5.82
7	**0.01**	4.95	5.92	6.54	7.00	7.37	7.68	7.94
8	**0.05**	3.26	4.04	4.53	4.89	5.17	5.40	5.60
8	**0.01**	4.75	5.64	6.20	6.62	6.96	7.24	7.47
9	**0.05**	3.20	3.95	4.41	4.76	5.02	5.24	5.43
9	**0.01**	4.60	5.43	5.96	6.35	6.66	6.91	7.13
10	**0.05**	3.15	3.88	4.33	4.65	4.91	5.12	5.30
10	**0.01**	4.48	5.27	5.77	6.14	6.43	6.67	6.87
11	**0.05**	3.11	3.82	4.26	4.57	4.82	5.03	5.20
11	**0.01**	4.39	5.15	5.62	5.97	6.25	6.48	6.67
12	**0.05**	3.08	3.77	4.20	4.51	4.75	4.95	5.12
12	**0.01**	4.32	5.05	5.50	5.84	6.1	6.32	6.51
13	**0.05**	3.06	3.73	4.15	4.45	4.69	4.88	5.05
13	**0.01**	4.26	4.96	5.40	5.73	5.98	6.19	6.37
14	**0.05**	3.03	3.70	4.11	4.41	4.64	4.83	4.99
14	**0.01**	4.21	4.89	5.32	5.63	5.88	6.08	6.26
15	**0.05**	3.01	3.67	4.08	4.37	4.59	4.78	4.94
15	**0.01**	4.17	4.84	5.25	5.56	5.80	5.99	6.16
16	**0.05**	3.00	3.65	4.05	4.33	4.56	4.74	4.90
16	**0.01**	4.13	4.79	5.19	5.49	5.72	5.91	6.08
17	**0.05**	2.98	3.63	4.02	4.30	4.52	4.70	4.86
17	**0.01**	4.10	4.74	5.14	5.43	5.66	5.85	6.01
18	**0.05**	2.97	3.61	4.00	4.28	4.49	4.67	4.82
18	**0.01**	4.07	4.70	5.09	5.38	5.60	5.79	5.94
19	**0.05**	2.96	3.59	3.98	4.25	4.47	4.65	4.79
19	**0.01**	4.05	4.67	5.05	5.33	5.55	5.73	5.89
20	**0.05**	2.95	3.58	3.96	4.23	4.45	4.62	4.77
20	**0.01**	4.02	4.64	5.02	5.29	5.51	5.69	5.84
25	**0.05**	2.91	3.52	3.89	4.15	4.36	4.53	4.67
25	**0.01**	3.94	4.53	4.88	5.14	5.35	5.51	5.65
30	**0.05**	2.89	3.49	3.85	4.10	4.30	4.46	4.60
30	**0.01**	3.89	4.45	4.80	5.05	5.24	5.40	5.54
40	**0.05**	2.86	3.44	3.79	4.04	4.23	4.39	4.52
40	**0.01**	3.82	4.37	4.69	4.93	5.11	5.26	5.39
60	**0.05**	2.83	3.40	3.74	3.98	4.16	4.31	4.44
60	**0.01**	3.76	4.28	4.59	4.82	4.99	5.13	5.25

Table produced using the SAS System using function PROBMC('SRANGE',,1 − α,df,T).

Figure 2: Table A.6 from Mohr, Wilson, and Freund (2021)

Rust inhibitors example (cont)

For the rust data we have $n = 10$ and $a = 4$.

At $\alpha = 0.05$ we have $q_{a,a(n-1),\alpha} = q_{4,36,0.05} \approx 3.85$ from table.

Obtain exact value with `qtukey(.95,4,36)` = 3.8087984.

Build the Tukey HSD CI for $\mu_2 - \mu_1$.

```
n <- 10
a <- 4
MSE <- sum(lm_out$residuals^2) / ( a*(n-1))
y1bar <- mean(rust$score[rust$brand == 1])
y2bar <- mean(rust$score[rust$brand == 2])
me <- qtukey(.95,a,a*(n-1)) * sqrt(MSE) / sqrt(10)
lo21 <- y2bar - y1bar - me
up21 <- y2bar - y1bar + me
c(lo21,up21)
```

```
[1] 43.31554 49.28446
```

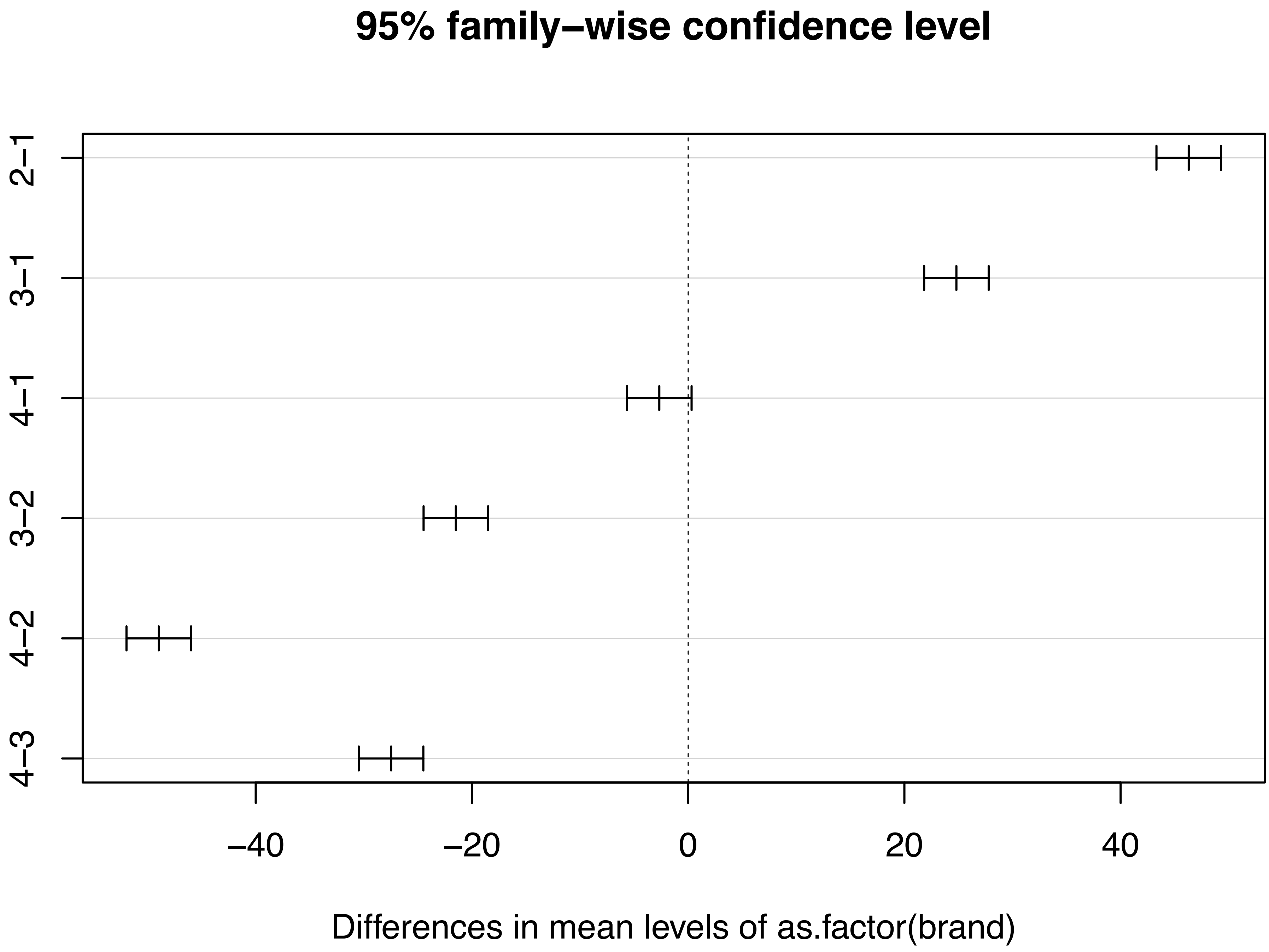
Rust inhibitors example (cont)

Use TukeyHSD() on aov() output to obtain the simultaneous CIs.

```
# must use the aov() function instead of the lm() function
aov_out <- aov(score ~ as.factor(brand), data = rust)
TukeyHSD(aov_out)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = score ~ as.factor(brand), data = rust)

$`as.factor(brand)`
      diff        lwr         upr     p adj
2-1  46.30  43.315536  49.2844635 0.0000000
3-1  24.81  21.825536  27.7944635 0.0000000
4-1  -2.67  -5.654464   0.3144635 0.0933303
3-2 -21.49 -24.474464 -18.5055365 0.0000000
4-2 -48.97 -51.954464 -45.9855365 0.0000000
4-3 -27.48 -30.464464 -24.4955365 0.0000000
```

```
plot(TukeyHSD(aov_out))
```



Comparison of treatments with a baseline treatment

- It may be that not all pairwise comparisons are of interest.
- Then Tukey's method is too conservative (CIs wider than necessary).
- Say we want to compare all treatments to a "baseline" treatment.
- Build CIs for $\mu_i - \mu_1$, $i = 2, \ldots, a$, 1 the baseline treatment.
- This makes $a - 1$ CIs instead of $\binom{a}{2}$ CIs.
- Can use Dunnett's method, Dunnett (1964).

The equally venerable Charles Dunnett

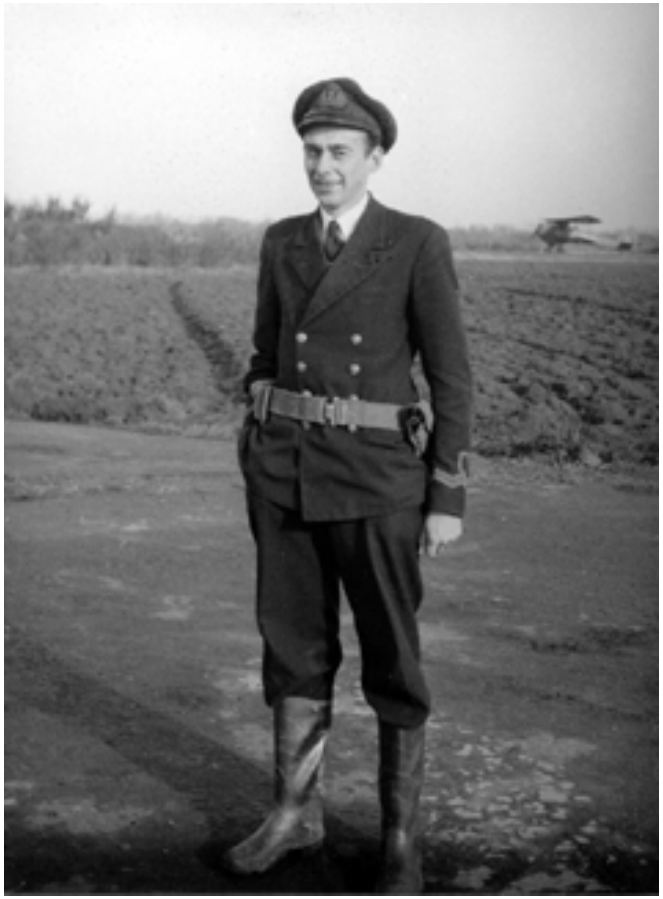


Figure 3: Charles Dunnett, 1921 – 2007 (Canadian, served in WWII, photo taken in Belgium)

Dunnett's method for comparisons with a baseline

- Assume $n_i = n$ for all i (balanced case).
- Given a value $d_{n,a(n-1),\alpha}$ such that

$$P\left(\max_{2\le i\le a}\left|\frac{(\bar{Y}_{i.}-\bar{Y}_{1.})-(\mu_i-\mu_1)}{\hat{\sigma}\sqrt{2/n}}\right|\le d_{n,a(n-1),\alpha}\right)=1-\alpha,$$

  with probability $1-\alpha$ the CIs

$$\bar{Y}_{i.}-\bar{Y}_{1.}\pm d_{n,a(n-1),\alpha}\hat{\sigma}\sqrt{2/n}$$

  will simultaneously cover the targets $\mu_i-\mu_1$ for all $i=2,\dots,a$.
- Dunnett made tables of the values $d_{n,a(n-1),\alpha}$.
- Cannot sort all the means after Dunnett's.

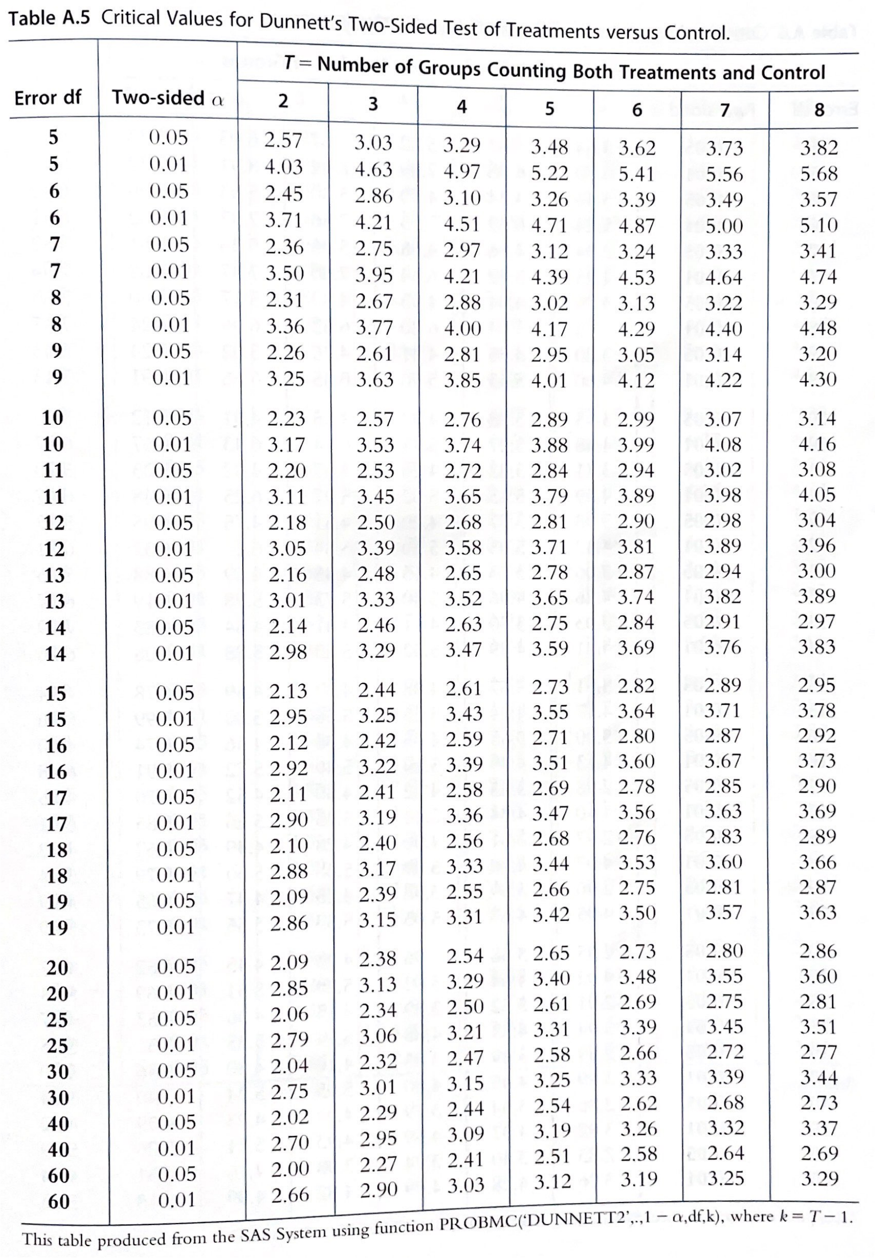


Table A.5 Critical Values for Dunnett's Two-Sided Test of Treatments versus Control.

Error df	Two-sided α	T = Number of Groups Counting Both Treatments and Control						
		2	3	4	5	6	7	8
5	0.05	2.57	3.03	3.29	3.48	3.62	3.73	3.82
5	0.01	4.03	4.63	4.97	5.22	5.41	5.56	5.68
6	0.05	2.45	2.86	3.10	3.26	3.39	3.49	3.57
6	0.01	3.71	4.21	4.51	4.71	4.87	5.00	5.10
7	0.05	2.36	2.75	2.97	3.12	3.24	3.33	3.41
7	0.01	3.50	3.95	4.21	4.39	4.53	4.64	4.74
8	0.05	2.31	2.67	2.88	3.02	3.13	3.22	3.29
8	0.01	3.36	3.77	4.00	4.17	4.29	4.40	4.48
9	0.05	2.26	2.61	2.81	2.95	3.05	3.14	3.20
9	0.01	3.25	3.63	3.85	4.01	4.12	4.22	4.30
10	0.05	2.23	2.57	2.76	2.89	2.99	3.07	3.14
10	0.01	3.17	3.53	3.74	3.88	3.99	4.08	4.16
11	0.05	2.20	2.53	2.72	2.84	2.94	3.02	3.08
11	0.01	3.11	3.45	3.65	3.79	3.89	3.98	4.05
12	0.05	2.18	2.50	2.68	2.81	2.90	2.98	3.04
12	0.01	3.05	3.39	3.58	3.71	3.81	3.89	3.96
13	0.05	2.16	2.48	2.65	2.78	2.87	2.94	3.00
13	0.01	3.01	3.33	3.52	3.65	3.74	3.82	3.89
14	0.05	2.14	2.46	2.63	2.75	2.84	2.91	2.97
14	0.01	2.98	3.29	3.47	3.59	3.69	3.76	3.83
15	0.05	2.13	2.44	2.61	2.73	2.82	2.89	2.95
15	0.01	2.95	3.25	3.43	3.55	3.64	3.71	3.78
16	0.05	2.12	2.42	2.59	2.71	2.80	2.87	2.92
16	0.01	2.92	3.22	3.39	3.51	3.60	3.67	3.73
17	0.05	2.11	2.41	2.58	2.69	2.78	2.85	2.90
17	0.01	2.90	3.19	3.36	3.47	3.56	3.63	3.69
18	0.05	2.10	2.40	2.56	2.68	2.76	2.83	2.89
18	0.01	2.88	3.17	3.33	3.44	3.53	3.60	3.66
19	0.05	2.09	2.39	2.55	2.66	2.75	2.81	2.87
19	0.01	2.86	3.15	3.31	3.42	3.50	3.57	3.63
20	0.05	2.09	2.38	2.54	2.65	2.73	2.80	2.86
20	0.01	2.85	3.13	3.29	3.40	3.48	3.55	3.60
25	0.05	2.06	2.34	2.50	2.61	2.69	2.75	2.81
25	0.01	2.79	3.06	3.21	3.31	3.39	3.45	3.51
30	0.05	2.04	2.32	2.47	2.58	2.66	2.72	2.77
30	0.01	2.75	3.01	3.15	3.25	3.33	3.39	3.44
40	0.05	2.02	2.29	2.44	2.54	2.62	2.68	2.73
40	0.01	2.70	2.95	3.09	3.19	3.26	3.32	3.37
60	0.05	2.00	2.27	2.41	2.51	2.58	2.64	2.69
60	0.01	2.66	2.90	3.03	3.12	3.19	3.25	3.29

This table produced from the SAS System using function PROBMC('DUNNETT2',,1 − α,df,k), where $k = T - 1$.

Figure 4: Table A.5 from Mohr, Wilson, and Freund (2021)

Rust inhibitor data (cont)

For the rust data we have $n = 10$ and $a = 4$.

At $\alpha = 0.05$ we have $d_{a,a(n-1),\alpha} = d_{4,36,0.05}$.

Use value 2.44 in the table (should be close).

Treat Brand 1 as the baseline and make comparisons with Dunnett's.

```
# just show the comparison of treatment 2 to the baseline
y1bar <- mean(rust$score[rust$brand == 1])
y2bar <- mean(rust$score[rust$brand == 2])

me <- 2.44 * sqrt(MSE) * sqrt(2/10) # margin of error for Dunnett's

lo21 <- y2bar - y1bar - me
up21 <- y2bar - y1bar + me

c(y2bar - y1bar,lo21,up21)
```

```
[1] 46.30000 43.59615 49.00385
```

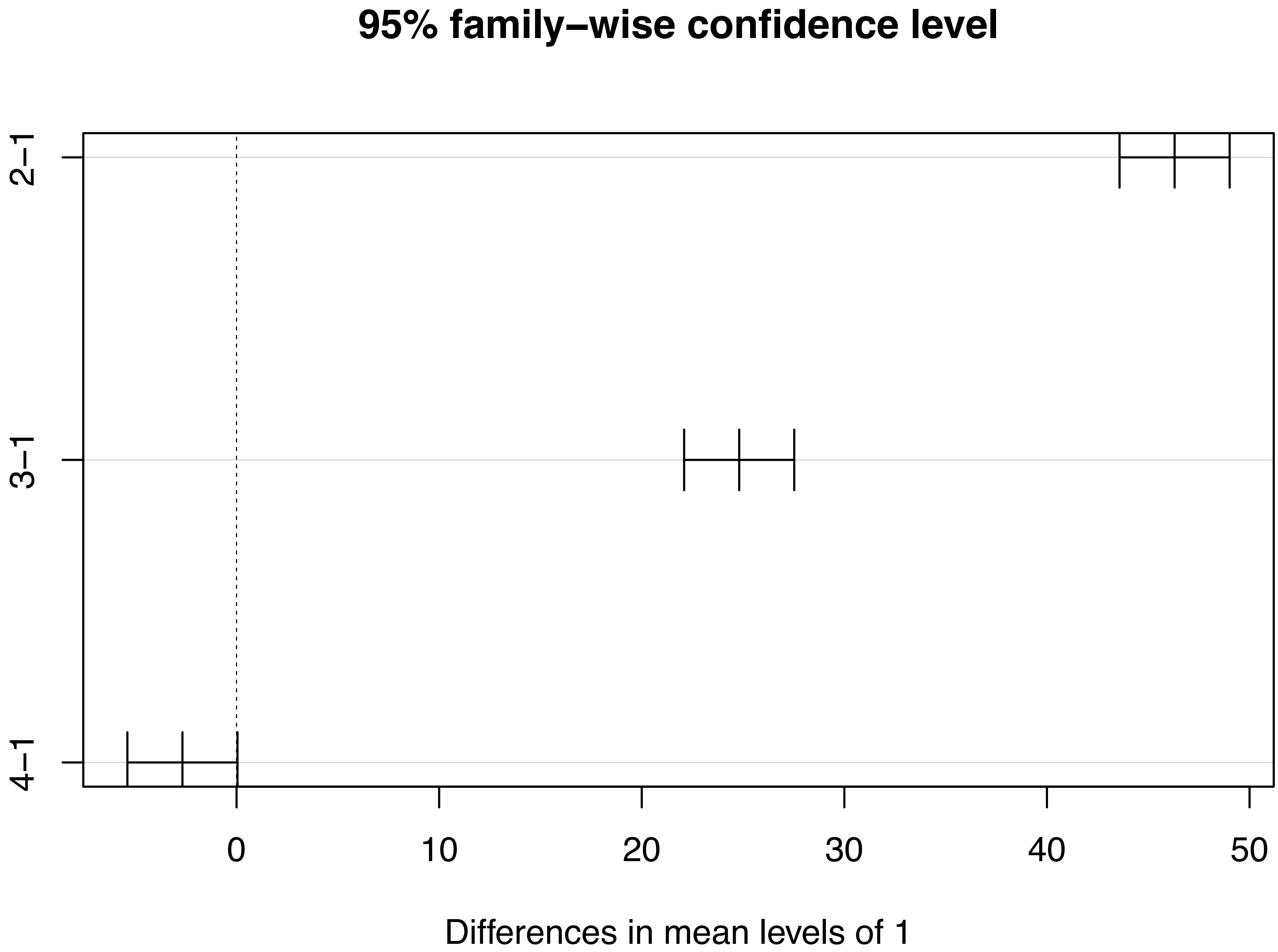
Rust inhibitor data (cont)

Use `DunnettTest()` from R package `DescTools`.

```
library(DescTools) # first time run install.packages("DescTools")
Dunnett_out <- DunnettTest(score ~ as.factor(brand), data = rust, control = "1")
Dunnett_out
```

```

  Dunnett's test for comparing several treatments with a control :
    95% family-wise confidence level

$`1`
     diff    lwr.ci    upr.ci   pval
2-1 46.30 43.582516 49.017484 <2e-16 ***
3-1 24.81 22.092516 27.527484 <2e-16 ***
4-1 -2.67 -5.387484  0.047484 0.0549 .

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(Dunnett_out)
```



Dunnett's vs Tukey's

- Tukey's is for comparisons between all pairs of means.
- Dunnett's is for comparison of means with a baseline.
- So Tukey's must make greater adjustments to control the familywise Type I error.
- Therefore Tukey intervals will be wider than Dunnett intervals.
- Tukey's allows you to sort the means, while Dunnett's does not.
- Both methods assume a balanced design, i.e. $n_i = n$ for all i. Modifications for unbalanced designs exist, but are not straightforward to implement in R.

Bonferroni correction

If building B CIs you can ALWAYS use the Bonferroni correction:

- Build each CI ordinarily, but use α/B instead of α.
- Ensures simultaneous coverage of all CIs with probability $\geq 1 - \alpha$.
- True prob of simultaneous coverage may be greater than $1 - \alpha$
- Bonferroni-corrected CIs will be wider than Dunnett's and wider than Tukey's if used for making those same comparisons.
- Use when we do not know how to adjust for multiple comparisons.

Rust inhibitor data (cont)

Compare Brand 3 to 4 and Brand 1 to 3, using the Bonferroni correction to control the familywise error rate.

```
y1bar <- mean(rust$score[rust$brand == 1])
y3bar <- mean(rust$score[rust$brand == 3])
y4bar <- mean(rust$score[rust$brand == 4])
alpha <- 0.05
B <- 2
me <- qt(1 - (alpha/B)/2,a*(n-1)) * sqrt(MSE) * sqrt(2/n)
tab <- rbind(c(y3bar - y4bar - me,y3bar - y4bar + me),
             c(y1bar - y3bar - me,y1bar - y3bar + me))
rownames(tab) <- c("3-4","1-3")
colnames(tab) <- c("lower","upper")
tab
```

```
      lower   upper
3-4  24.888  30.072
1-3 -27.402 -22.218
```

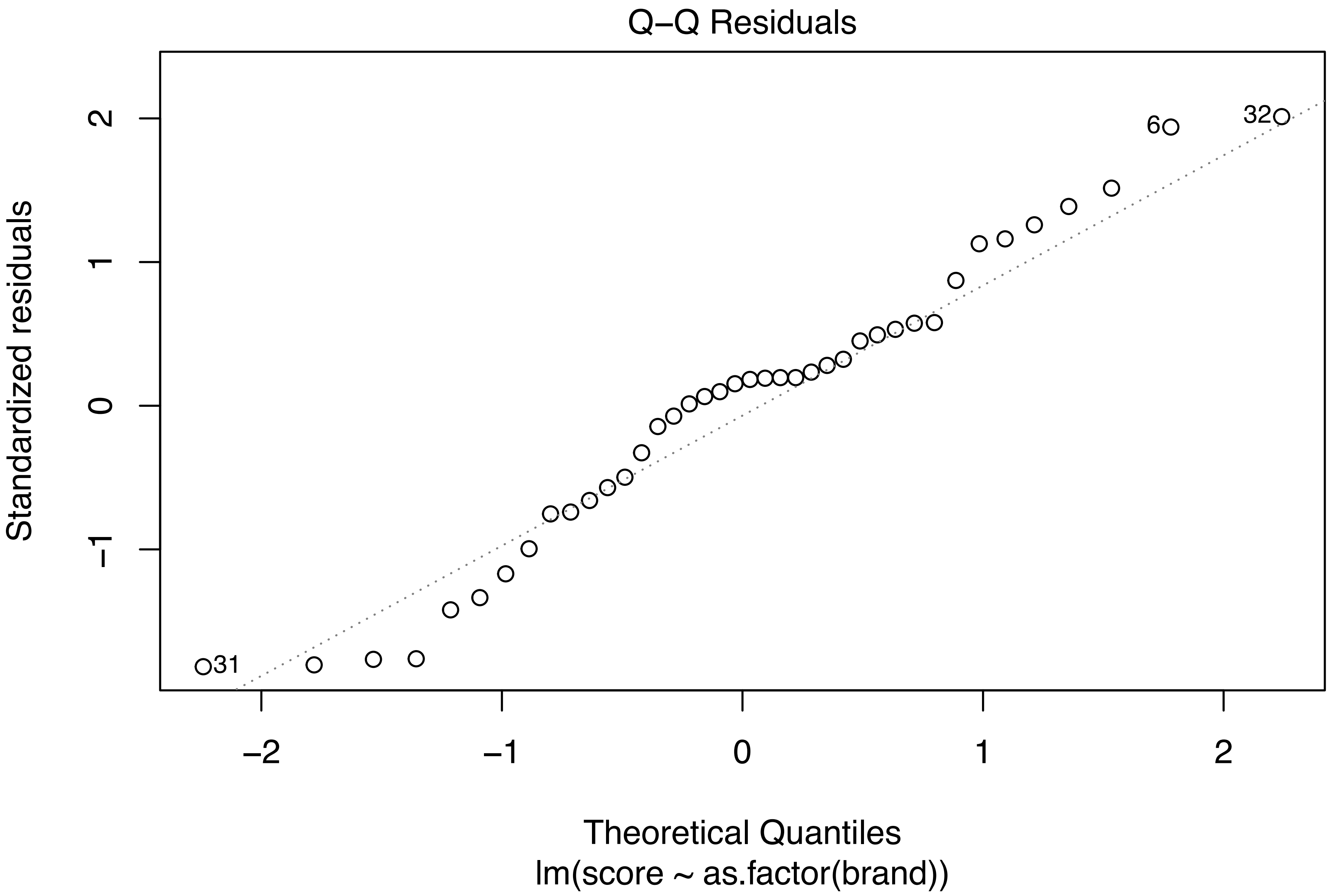
Checking model assumptions

Validity of the foregoing analyses depends on these assumptions:

1. The responses are normally distributed around the treatment means (Check QQ plot of residuals).
2. The response has the same variance in all treatment groups (Check residuals vs fitted values plot).
3. The response values are independent of each other (No way to check; must trust experimental design).

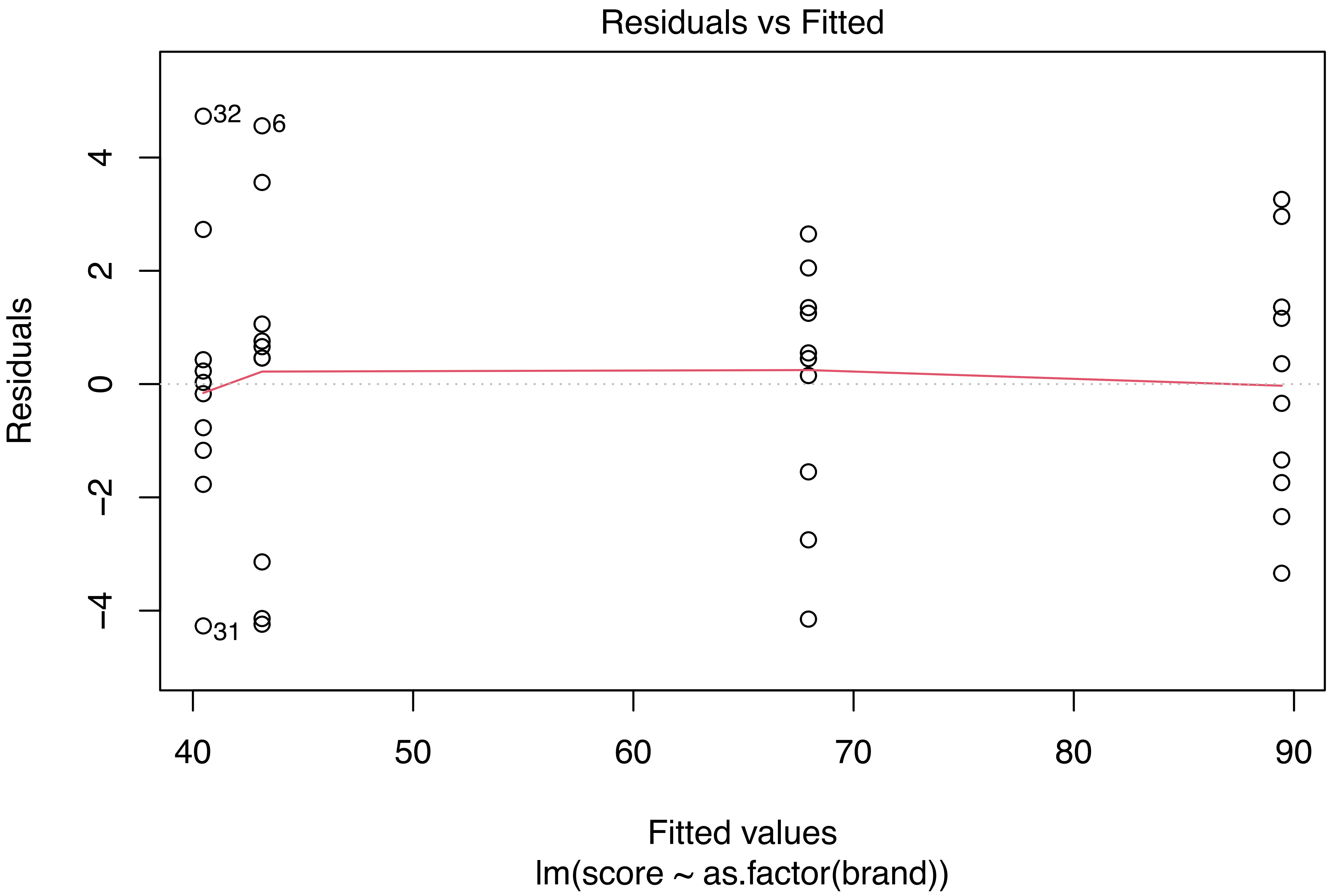
Rust inhibitors example (cont)

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



Rust inhibitors example (cont)

```
plot(lm_out,which = 1)
```



Perception of slope example

Do axis re-scalings affect how we perceive an x-y relationship?

For a single data set with data pairs (X_i, Y_i), with $X_i \sim$ Normal$(0, 1)$ and $Y_i =$ Normal$(X_i, 1)$ for $i = 1, \dots, 50$, three scatterplot treatments were constructed:

1. "Control" used x and y plotting limits given by the range of the data.
2. "X" extended the x-limits by 1.5 in each direction.
3. "Y" extended the y-limits by 1.5 in each direction.

Each student in a class was randomly assigned a scatterplot and told to draw with a ruler the best-fitting line through the data. The slope of each student-drawn line was measured and recorded as the response.

Is the response mean the same in the three treatment groups?

An artifact from each treatment group:

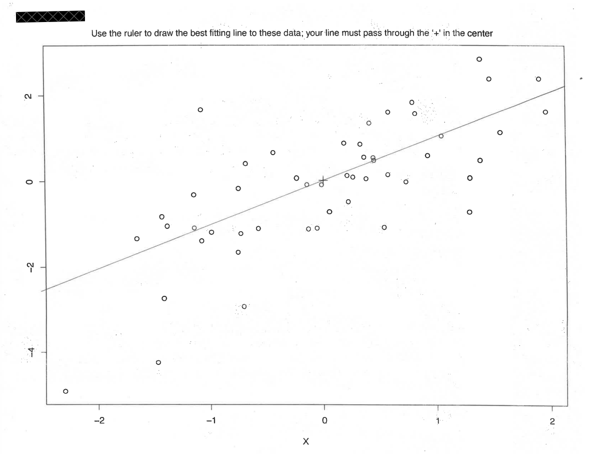


Figure 5: “Control”

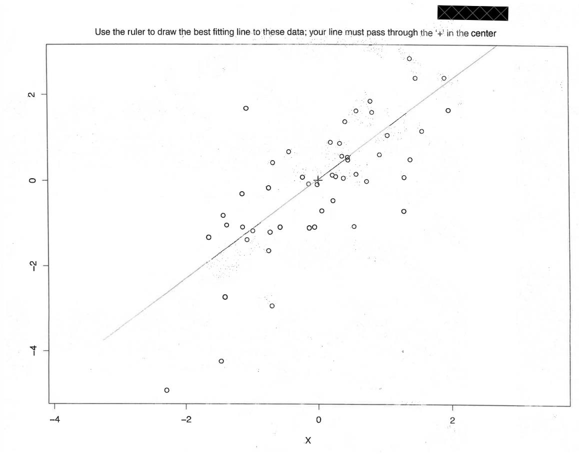


Figure 6: “X”

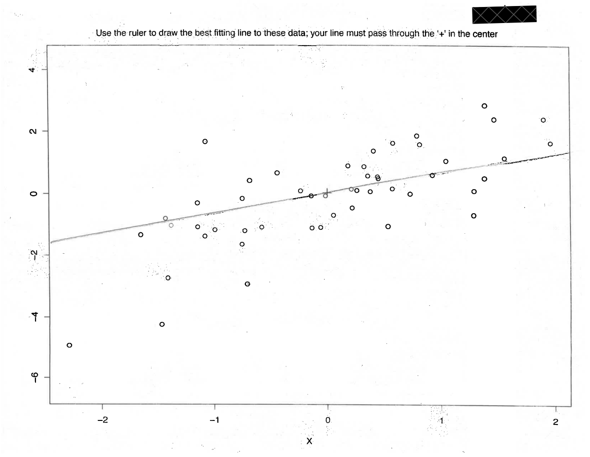
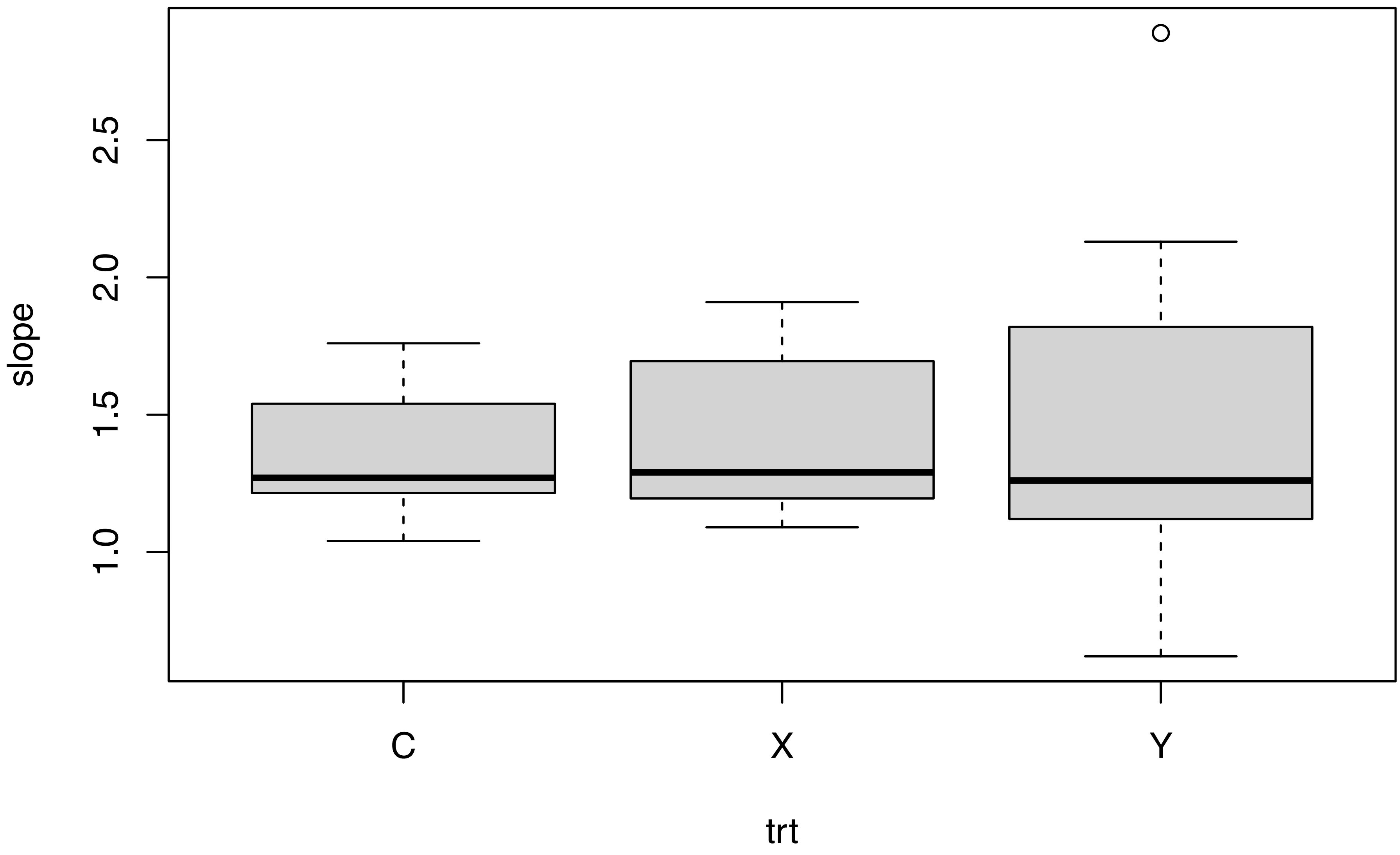


Figure 7: “Y”

```
slope <- c(1.23,1.80,1.81,1.29,2.89,1.58,0.99,1.24,
           1.26,1.57,1.27,1.19,1.82,1.76,1.91,1.25,
           1.09,1.29,1.12,1.51,2.13,1.16,0.62,1.04)
trt <- c("X","Y","X","X","Y","X","Y","C",
         "Y","C","C","C","Y","C","X","Y",
         "X","X","Y","C","Y","X","Y","C")
```

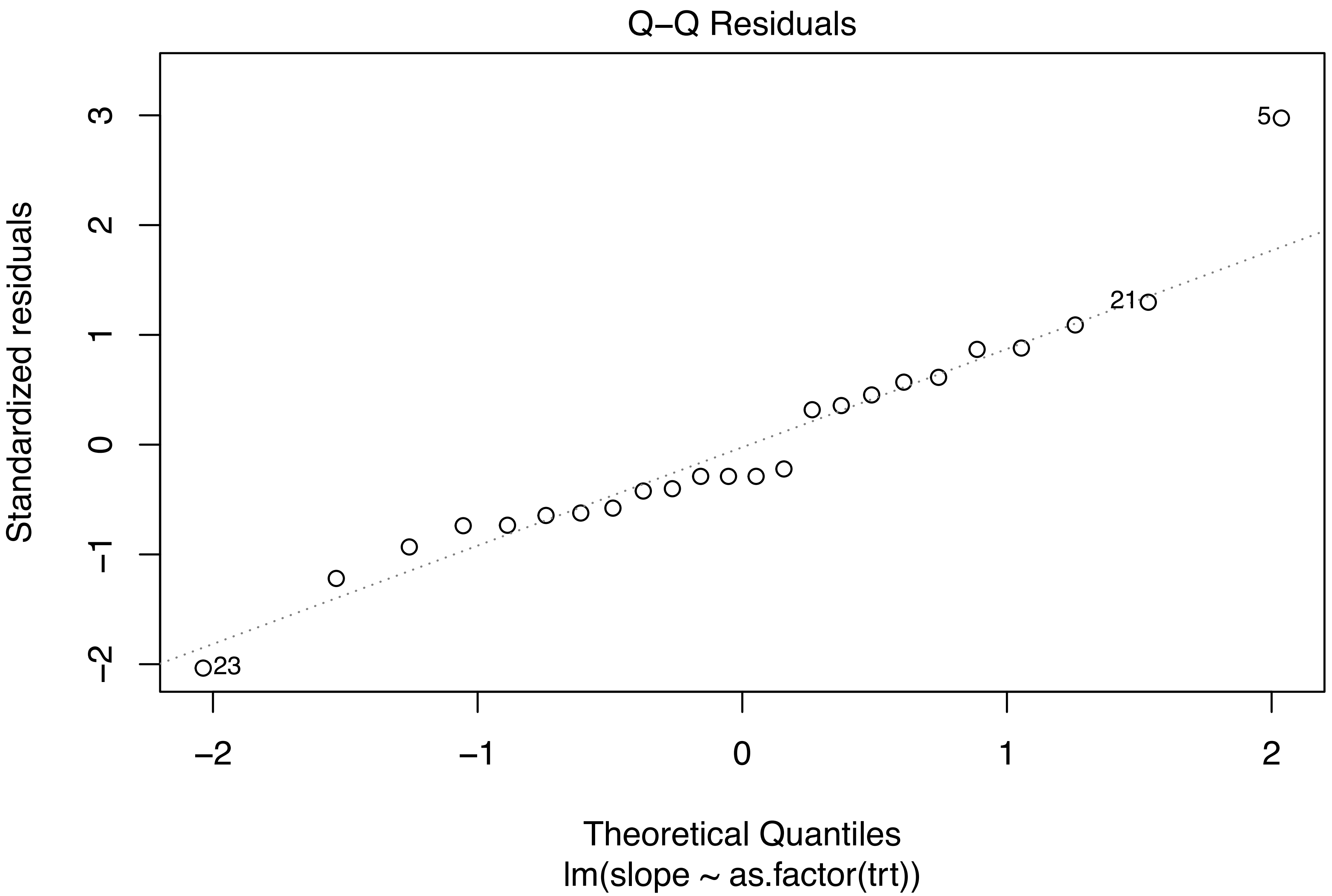
```r
boxplot(slope ~ trt)
```



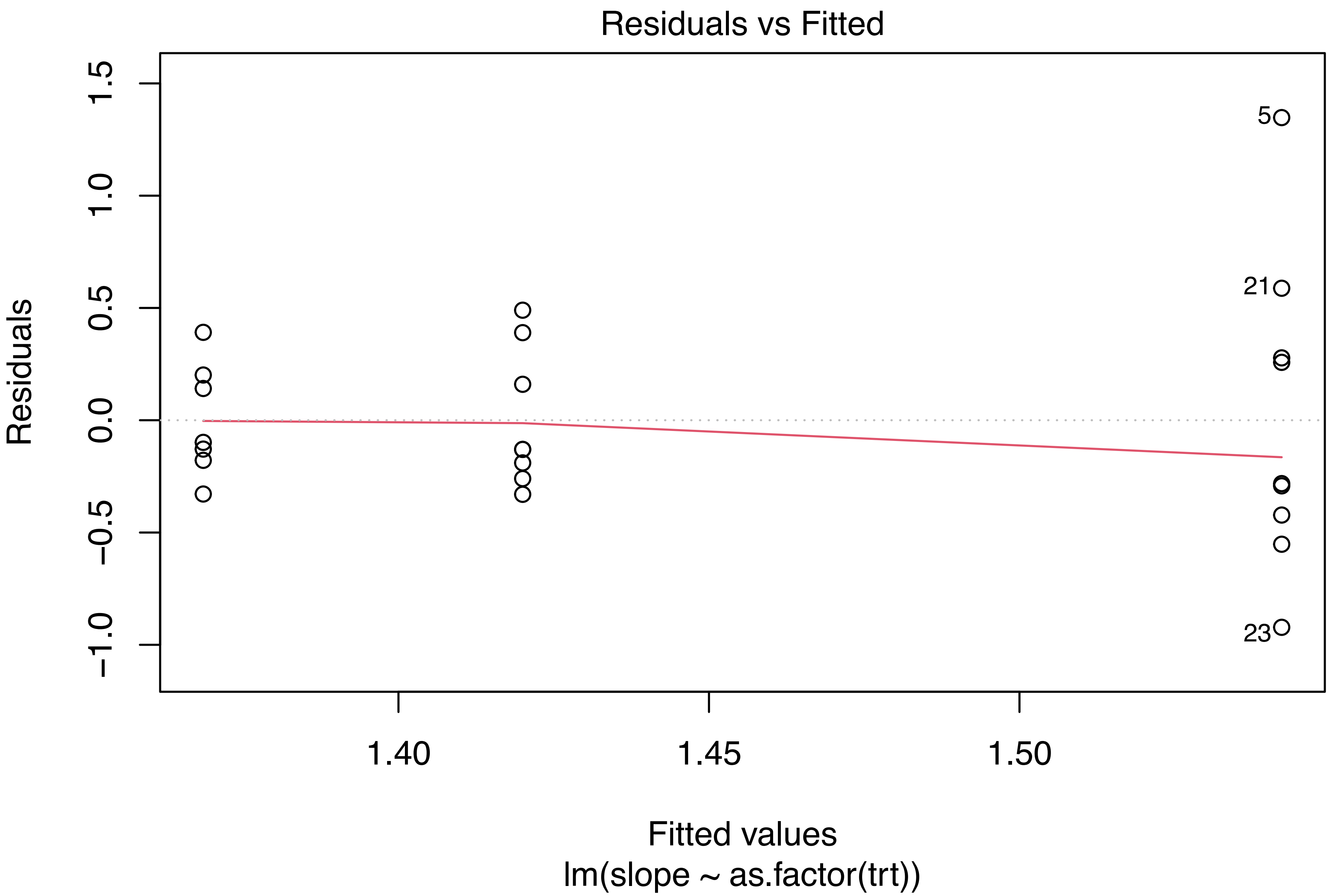
```r
lm_slope <- lm(slope ~ as.factor(trt))
summary(lm_slope)
```

```
Call:
lm(formula = slope ~ as.factor(trt))

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9222 -0.2847 -0.1293  0.2628  1.3478 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      1.36857    0.18161   7.536 2.12e-07 ***
as.factor(trt)X  0.05143    0.24868   0.207    0.838    
as.factor(trt)Y  0.17365    0.24215   0.717    0.481    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4805 on 21 degrees of freedom
Multiple R-squared:  0.02614,	Adjusted R-squared:  -0.06661 
F-statistic: 0.2818 on 2 and 21 DF,  p-value: 0.7572
```

```r
plot(lm_slope,which = 2)
```



```
plot(lm_slope,which = 1)
```



Levene's test for equality of variances

Checks if the mean magnitude of the residuals is equal across groups:

1. Obtain the residuals $\widehat{\varepsilon}_{ij}$ from the one-way ANOVA model.
2. Treat the absolute values $|\widehat{\varepsilon}_{ij}|$ of the residuals as *new* responses.
3. Test for equal means of the new responses with the F test.

So, do the ordinary F-test with the $|\widehat{\varepsilon}_{ij}|$ as the responses.

Perception of slope example (cont)

Perform Levene's test:

```
ehat <- lm_slope$residuals
lm_levene <- lm(abs(ehat) ~ as.factor(trt))
summary(lm_levene)
```

```

Call:
lm(formula = abs(ehat) ~ as.factor(trt))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.29136 -0.12769 -0.04980  0.08219  0.79864 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)  
(Intercept)      0.20980    0.09352   2.243   0.0358 *
as.factor(trt)X  0.05020    0.12805   0.392   0.6990  
as.factor(trt)Y  0.33934    0.12469   2.721   0.0128 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2474 on 21 degrees of freedom
Multiple R-squared:  0.303, Adjusted R-squared:  0.2367 
F-statistic: 4.565 on 2 and 21 DF,  p-value: 0.02258
```

Can also use the `leveneTest()` function in the R package `car`.

```
library(car)
leveneTest(slope~as.factor(trt),center = mean)
```

```
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value  Pr(>F)
group  2  4.5652 0.02258 *
      21
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We conclude that the variances are *not* equal across treatment groups.

References

Dunnett, Charles W. 1964. “New Tables for Multiple Comparisons with a Control.” *Biometrics* 20 (3): 482–91.

Kutner, Michael H, Christopher J Nachtsheim, John Neter, and William Li. 2005. *Applied Linear Statistical Models*. McGraw-hill.

Mohr, Donna L, William J Wilson, and Rudolf J Freund. 2021. *Statistical Methods*. Academic Press.