STA303/1002 - Methods of Data Analysis II

(Week 04 lecture note)

Wei (Becky) Lin

Jan. 24/26, 2017

UNIVERSITY OF TORONTO

Review: Topics learned last week

One-way ANOVA: Comparing means for a factor with more than two levels

- Fixed effect model

$= \mu_i + \varepsilon_{ij}$, $\mu_i = \mu + \alpha_i$ = mean of level i

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, (i = 1, 2, \ldots, r, j = 1, 2, \ldots, n_i)$$

  - μ deterministic unknown parameter
  - α_i is the unknown main effect of treatment i. $\mu_i = \mu + \alpha_i$
  - $\epsilon_{ij} \sim_{iid} N(0, \sigma^2)$, unobserved random noise with equal variance across all groups i. → Assumption: ① independent ② Normality ③ common σ^2
  - To estimate unknown parameters in the fixed effect model, we need indentifiability constraints.
    - Common used constraints: $\alpha_1 = 0$ (default in R) or $\mu = 0$ or $\sum_i \alpha_i = 0$
  - Want to test treatment effect is equivalent to test equal means

$$H_0 : \mu_1 = \ldots = \mu_r, \qquad H_a : \exists i \neq j \text{ s.t. } \mu_i \neq \mu_j$$

This is the same as testing (because $\mu_i = \mu + \alpha_i$, equal $\mu_i \Rightarrow \alpha_i = 0$)

$$H_0 : \alpha_1 = \ldots = \alpha_r = 0, \qquad H_a : \exists i \text{ s.t. } \alpha_i \neq 0$$

Review: Topics learned last week

- One-way ANOVA Table

Source	df	SS	Mean Square	F
Treatments	r-1	SS_T	MST=SS_T/(r-1)	F=MST/MSE
Residual	n-r	SS_E	MSE	
Total	n-1	SS		

(*)

- Want to test whether we have equal population means or not

$$H_0 : \mu_1 = \ldots = \mu_r, \qquad H_a : \exists i \neq j \text{ s.t. } \mu_i \neq \mu_j$$

  - **Test statistic**

treatment = 0

E(MST) = E(MSE)

$$F^* = \frac{MST}{MSE} = \frac{\text{treatment+error}}{\text{error}} = \frac{SS_T/(r-1)}{SS_E/(n-r)} \sim F_{r-1,n-r} \text{ under } H_0$$

  - **Decision:** rejecting H_0 at significance level α if F^* is too big, that is, we reject H_0 if

$$F^* > F_{1-\alpha,r-1,n-r}$$

Review Topics Learned Last Week

- One-way ANOVA: Bartlett Test of Homogeneity of Variances. In R: *bartlett.test()*. If levels=2, *var.test()*
- One-way ANOVA: Post-hoc multiple comparison
  - Adjust P-value: it takes into account that multiple tests are being conducted so that the familywise error rate is maintained at the desired level α. Good reference: paper by S.P Wright (1992).
    - Bonferroni Procedure: given a collection of hypotheses $(H_1, \ldots, H_n)$ and $\alpha_{FW} = \alpha$. Each individual hypothesis H_i is tested at a reduced signficance level α_i s.t. $\sum_i \alpha_i = \alpha$,typically, $\alpha_i = \alpha/n$. Let p_i be the unadjusted p-value for H_i, then H_i is rejected at $p_i \leq \alpha/n$ that is $p_{i,Bonf} = np_i \leq \alpha$. $P(\cup_i H_i) \leq \sum_i P(H_i) = n\frac{\alpha}{n} = \alpha$
    - Tukey HSD: using the Studentized range distribution for pairwise comparisons in a balanced design.
    - Dunnett's Method:(comparing one control group (Y_c) to all other k-1 groups.) using the Dunnett's t distribution.
    - There are many other methods (e.g. Holm, Benjamini and Hochberg)
- Data example: Spock Conspiracy Trial.

Review Topics learned Last Week

Two-way Crossed ANOVA

- Factor Effect model

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

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

  - μ deterministic unknown parameter
  - α_i is the unknown main effect of treatment A at level $i, i = 1, \ldots, a$.
  - β_j is the unknown main effect of treatment B at level $j, j = 1, \ldots, b$.
  - $k = 1, \ldots, n_{ij} = r$: replicates for treatment combination (i,j). Assume balanced design $n_{ij} = r$. This implies that the total sample size $n = abr = n_{11} + \ldots + n_{ab}$
  - $\epsilon_{ijk} \sim_{iid} N(0, \sigma^2)$, unobserved random noise with equal variance across all .
  - Constraints required to keep model from being over-parameterized. e.g. $\sum_i^a \alpha_i = \sum_j^b \beta_j = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0$
- A model without the interaction term, i.e.$\mu_{ij} = \mu + \alpha_i + \beta_j$, is called an additive model or main effect model: $Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$
- Advantage of factor effect model: Effects can be analyzed separately. This is the model we used.
- Using indicator variables, this is still a multiple regression problem

Review Topics Learned Last Week

- Two-way Crossed ANOVA Table

n = total sample size

Source	SS	df	Mean Square	F
Factor A	SSA	(a-1)	MSA$=\frac{SSA}{(a-1)}$	F=MSA/MSE
Factor B	SSB	(b-1)	MSB$=\frac{SSB}{(b-1)}$	F=MSB/MSE
Interaction AB	SSAB	(a-1)(b-1)	MSAB$=\frac{SSAB}{(a-1)(b-1)}$	F=MSAB/MSE
Error	SSE	ab(r-1)	MSE$=\frac{SSE}{df(SSE)}$	

$\Sigma df = n-1 = abr-1$

- Single Two-way ANOVA Table (if interaction is non-significant)

Source	SS	df	Mean Square	F
Factor A	SSA	(a-1)	MSA$=\frac{SSA}{(a-1)}$	F=MSA/MSE
Factor B	SSB	(b-1)	MSB$=\frac{SSB}{(b-1)}$	F=MSB/MSE
Error	SSE	abr-a-b+1	MSE$=\frac{SSE}{df(SSE)}$	

$\Sigma df = n-1 = abr-1$

Review Topics Learned Last Week

F-tests for two-way ANOVA

- Interaction: $H_0:$ all $(\alpha\beta)_{ij} = 0, \quad H_a:$ not all $(\alpha\beta)_{ij} = 0$
  - Test statistic: $F^*_{AB} = MSAB/MSE \sim F_{(a-1)(b-1),ab(r-1)}$ under H_0
  - Reject H_0 if $F^*_{AB} > F_{\alpha,(a-1)(b-1),ab(r-1)}$
- Main effect A: $H_0:$ all $\alpha_1 = \ldots = \alpha_a = 0, \quad H_a:$ not all α_i equal 0
  - Test statistic: $F^*_A = MSA/MSE \sim F_{(a-1),df(SSE)}$ under H_0
  - Reject H_0 if $F^*_A > F_{\alpha,(a-1),df(SSE)}$
- Main effect B: $H_0:$ all $\beta_1 = \ldots = \beta_b = 0, \quad H_a:$ not all β_j equal 0
  - Test statistic: $F^*_B = MSB/MSE \sim F_{(b-1),df(SSE)}$ under H_0
  - Reject H_0 if $F^*_B > F_{\alpha,(b-1),df(SSE)}$

Data example: paint data analysis.

Learning Objective This Week

- Identifiability problem
- Understand and interpret interaction effect in Two-way ANOVA
- Data Example

Identifiability Issue in ANOVA

Identifiability Issues

We use one-way ANOVA as example to examine the identifiability issues in ANOVA analysis.

- The one-way ANOVA model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

  - μ unknown parameter.
  - α_i unknown effect of treatment i
  - $\epsilon_{ij} \sim_{iid} N(0, 1)$ unobserved random noise.

- The identifiability issue
  - Model $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ is equivalent to $Y_{ij} = \mu' + \alpha'_i + \epsilon_{ij}$, with $\mu' = \mu + \delta, \alpha'_i = \alpha_i - \delta$.
  - There are infinitely many parameter combinations providing the same data.
  - Parameters are meaningless without additional constraints.

Identifiability Issues - reparameterization method

- If design matrix $X_{n\times p}$ has rank $r < p$, there is not a unique solution $\hat{\beta}$ to the normal equations. We need to reducing the model to of full rank to ensure that matrix of $X'X$ is invertible.
- Example: One-way ANOVA with a foctor of 2 levels (subgroups)

$\rightarrow$ rank(X)=2 since X[,1]=X[,2]+X[,3]

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \end{pmatrix} = X\beta + \epsilon = \begin{pmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \cdots \\ \epsilon_{2n_2} \end{pmatrix} = \begin{pmatrix} \mu + \alpha_1 \\ \vdots \\ \mu + \alpha_1 \\ \mu + \alpha_2 \\ \vdots \\ \mu + \alpha_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \cdots \\ \epsilon_{2n_2} \end{pmatrix}$$

Let X_1 consist of the first and third columns of X. Then

$$X = X_1 \begin{pmatrix} 1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}, \alpha = \begin{pmatrix} 1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} \mu + \alpha_1 \\ \alpha_2 - \alpha_1 \end{pmatrix}$$

We have $X\beta = X_1\alpha$.

Identifiability Issues (be contd)

- Example (contd): One-way ANOVA with a factor of 2 levels
  - Since $Y = X\beta + \epsilon = X_1\alpha + \epsilon$, so $\hat{\alpha} = (X_1'X_1)^{-1}X_1'Y$

$$\hat{\alpha} = \begin{pmatrix} n & n_2 \\ n_2 & n_2 \end{pmatrix} \begin{pmatrix} n_1\bar{Y}_{1.} + n_2\bar{Y}_{2.} \\ n_2\bar{Y}_{2.} \end{pmatrix} = \begin{pmatrix} \bar{Y}_{1.} \\ \bar{Y}_{2.} - \bar{Y}_{1.} \end{pmatrix}$$

  - Hence

$$\hat{Y} = X_1\hat{\alpha} = \begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_1 \\ \hat{\mu}_2 \\ \vdots \\ \hat{\mu}_2 \end{pmatrix} = \begin{pmatrix} \bar{Y}_{1.} \\ \vdots \\ \bar{Y}_{1.} \\ \bar{Y}_{2.} \\ \vdots \\ \bar{Y}_{2.} \end{pmatrix}$$

where $\mu_1 = \mu + \alpha_1, \mu_2 = \mu + \alpha_2$

Identifiability Issues - Imposing constraints

Another way to ensure there is a unique solution $\hat{\beta}$ to the normal equations would be imposing identifiability constraints.

- Impose $s = p - r$ constraints on β to make β uniquely determined (identifiable), i.e. such that for any $\theta \in R$, there is a unique β satisfying

$$\underset{n\times p}{X}\underset{p\times 1}{\beta} = \theta, \text{ and } \underset{s\times p}{H}\underset{p\times 1}{\beta} = 0$$

$\left\{ \begin{array}{l} p = \text{full rank of } X \\ r < p \end{array} \right.$

so that

$$\begin{pmatrix} \theta \\ 0 \end{pmatrix} = \underset{(n+s)\times p}{\begin{pmatrix} X \\ H \end{pmatrix}} \underset{p\times 1}{\beta} = G\beta \rightarrow (n+s)\times 1$$

- To estimate β, we solve $\hat{Y} = X\hat{\beta}$ and $H\hat{\beta} = 0$,i.e., we solve the augmented normal equations $X'X\hat{\beta} = X'Y$ and $H'H\hat{\beta} = 0$, that is $(X'X + H'H)\hat{\beta} = (\underset{p\times(n+s)}{G'}\underset{(n+s)\times p}{G})\hat{\beta} = X'Y$. therefore

$$\hat{\beta} = (\underbrace{G'G}_{p\times p})^{-1}\underset{p\times n}{X'}\underset{n\times 1}{Y}$$

- A unique solution exists if and only if G has rank p and the rows of H are linearly independent of the rows of X.

Identifiability Issues (be contd)

- Example (revisit): One-way ANOVA with a foctor of 2 levels

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \end{pmatrix} = X\beta + \epsilon = \begin{pmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \cdots \\ \epsilon_{2n_2} \end{pmatrix} = \begin{pmatrix} \mu + \alpha_1 \\ \vdots \\ \mu + \alpha_1 \\ \mu + \alpha_2 \\ \vdots \\ \mu + \alpha_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \cdots \\ \epsilon_{2n_2} \end{pmatrix}$$

- Constraint: set $\alpha_1 = 0$, i.e.

$$H\beta = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} = 0 \quad = 0\mu + 1\alpha_1 + 0\alpha_2$$

- It can be shown that under $\alpha_1 = 0$

$$\hat{\beta} = \begin{pmatrix} \bar{Y}_{1.} & 0 & \bar{Y}_{2.} - \bar{Y}_{1.} \end{pmatrix}'$$

Clearly, it satisfies the normal equations and $\alpha_1 = 0$, we have as before $\hat{Y} = X\hat{\beta} = (\bar{Y}_{1.}, \ldots, \bar{Y}_{1.}, \bar{Y}_{2.}, \ldots, \bar{Y}_{2.})'$

Identifiability Issues (be contd)

$0\mu + 1\alpha_1 + 1\alpha_2 = 0$

- Constraint: set $\alpha_1 + \alpha_2 = 0$, i.e.

$$H\beta = \begin{pmatrix} 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} = 0$$

- Assume $n_1 = n_2 = m$. It can be shown that under $\alpha_1 + \alpha_2 = 0$

$$\hat{\beta} = \begin{pmatrix} \bar{Y}_{..} \\ \frac{1}{2}(\bar{Y}_{1.} - \bar{Y}_{2.}) \\ \frac{1}{2}(\bar{Y}_{2.} - \bar{Y}_{1.}) \end{pmatrix}$$

Clearly, it satisfies the normal equations and $\alpha_1 + \alpha_2 = 0$.
Furthermore, we have as before
$\hat{Y} = X\hat{\beta} = (\bar{Y}_{1.}, \ldots, \bar{Y}_{1.}, \bar{Y}_{2.}, \ldots, \bar{Y}_{2.})'$

Common identifiability constraints

Under one-way ANOVA

- Model:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, i = 1, \ldots, r; j = 1, \ldots, n_i$$

  - $\alpha_1 = 0$ (default in R: *contr.treatment()*) or $\alpha_r = 0$ (default in SAS): α_i is the departure from mean value under treatment 1 (or r) induced by treatment i.
  - $\mu = 0$: α_i is the mean value under treatment i
  - $\sum \alpha_i = 0$: α_i is the departure from the grand mean induced by treatment i. (In R: *contr.sum()*)
- In R, the default constraint is $\alpha_1 = 0$. The parameter μ is the average response under treatment 1.
- More about reassigning different constraint and contrasts: STA305,STA442.

Paint Data Example: one-way ANOVA

```r
adhesion=c(4.0,4.5,4.3,5.6,4.9,5.4,3.8,3.7,
           4.0,5.4,4.9,5.6,5.8,6.1,6.3,5.5,5,5)
primer=factor(rep(rep(1:3,each=3),2))
application=factor(rep(1:2,each=9),labels=c("D","S"))
paint=data.frame(adhesion,primer,application)
with(paint,tapply(adhesion,primer,mean))
```

```
##        1        2        3
## 4.783333 5.683333 4.500000
```

← level mean of factor primer
"
marginal mean

```r
mean(adhesion) # find grand mean
```

```
## [1] 4.988889
```

$= \frac{\sum_{i=1}^{3}\sum_{j=1}^{2}\sum_{k=1}^{3} y_{ijk}}{n}$, n = total sample size

$n_{ij} = r = 3 \Rightarrow$ balanced design.

Paint Data Example: under constraint $\sum \alpha_i = 0$

- intercept: grand mean of response. $\bar{Y}$
- b_1 (primer1): $\bar{Y}_1 - \bar{Y}$
- b_2 (primer2): $\bar{Y}_2 - \bar{Y}$

```r
summary(lm(adhesion~primer,data=paint, contrasts=list(primer="contr.sum")))
```

```
##
## Call:
## lm(formula = adhesion ~ primer, data = paint, contrasts = list(primer = "con
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80000 -0.49583  0.01667  0.50000  1.00000
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.9889     0.1508  33.092 1.95e-15 ***
## primer1      -0.2056     0.2132  -0.964   0.3503
## primer2       0.6944     0.2132   3.257   0.0053 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6396 on 15 degrees of freedom
## Multiple R-squared:  0.4274, Adjusted R-squared:  0.3511
## F-statistic: 5.599 on 2 and 15 DF,  p-value: 0.01527
```

$\bar{Y}_{..} = 4.988889$

$\hat{\beta}_1 = 4.783333 - 4.988889$

$\hat{\beta}_2 = 4.5 - 4.988889$

$H_0: \beta_1 = \beta_2 = 0$

H_a: at least one $\neq 0$

$F = \frac{MST}{MSE}$

Paint Data Example: under constraint $\alpha_1 = 0$

- intercept: mean of the factor at reference level. $\bar{Y}_1 = \hat{\beta}_0$
- b_1 (primer2): mean difference between level 2 and level 1. $\bar{Y}_2 - \bar{Y}_1 = \hat{\beta}_1$
- b_2 (primer3): mean difference between level 3 and level 1. $\bar{Y}_3 - \bar{Y}_1 = \hat{\beta}_2$

```r
summary(lm(adhesion~primer,data=paint, contrasts=list(primer="contr.treatment")
```

```
##
## Call:
## lm(formula = adhesion ~ primer, data = paint, contrasts = list(primer = "con
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80000 -0.49583  0.01667  0.50000  1.00000
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.7833     0.2611  18.318 1.12e-11 ***
## primer2       0.9000     0.3693   2.437   0.0277 *
## primer3      -0.2833     0.3693  -0.767   0.4548
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6396 on 15 degrees of freedom
## Multiple R-squared:  0.4274, Adjusted R-squared:  0.3511
## F-statistic: 5.599 on 2 and 15 DF,  p-value: 0.01527
```

$$\begin{cases} \hat{\beta}_0 = \bar{Y}_1 = 4.7833 \\ \hat{\beta}_1 = \bar{Y}_2 - \bar{Y}_1 = 5.6833 - 4.7833 \\ \hat{\beta}_2 = \bar{Y}_3 - \bar{Y}_1 = 4.5 - 4.7833 \end{cases}$$

$$F = \frac{MST}{MSE} \leftarrow \begin{cases} H_0: \beta_1 = \beta_2 = 0 \\ H_a: \exists i, \beta_i \neq 0 \end{cases}$$

Paint Data Example

ANOVA output are the same

```r
anova(lm(adhesion~primer,data=paint, contrasts=list(primer="contr.sum")))
```

```
## Analysis of Variance Table
##
## Response: adhesion
##           Df Sum Sq Mean Sq F value  Pr(>F)
## primer     2 4.5811 2.29056  5.5989 0.01527 *
## Residuals 15 6.1367 0.40911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Paint Data Example

ANOVA output are the same

```
anova(lm(adhesion~primer,data=paint, contrasts=list(primer="contr.treatment")))
```

```
## Analysis of Variance Table
##
## Response: adhesion
##           Df Sum Sq Mean Sq F value  Pr(>F)
## primer     2 4.5811 2.29056  5.5989 0.01527 *
## Residuals 15 6.1367 0.40911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

* different constraints → different $\hat{\beta}_i$ in summary output.
→ ANOVA output remains unchanged.

Common identifiability constraints

Under two-way ANOVA (additive) model

- Model without interaction term

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$

  - Y_{ijk} is the k-th observation having received level i of factor 1 and level j of factor 2.
  - μ: unobserved mean value
  - α_i: unobserved deterministic effect of level i of factor 1
  - β_j: unobserved deterministic effect of level j of factor 2
  - $\epsilon_{ijk} \sim_{iid} N(0, \sigma^2)$: unobserved random noise
- Model above subject to identifiability issue as before.
- Common constraints under above additive model
  - $\mu = 0$ and $\alpha_1 = 0$
  - $\alpha_1 = 0$ and $\beta_1 = 0$ (default in R)
  - $\sum_i \alpha_i = 0$ and $\sum_j \beta_j = 0$

Paint Data Example: two-way ANOVA

R: find all mean
① aggregate()
② tapply

```
adhesion=c(4.0,4.5,4.3,5.6,4.9,5.4,3.8,3.7,
           4.0,5.4,4.9,5.6,5.8,6.1,6.3,5.5,5,5)
primer=factor(rep(rep(1:3,each=3),2))
application=factor(rep(1:2,each=9),labels=c("D","S"))
paint=data.frame(adhesion,primer,application)
aggregate(adhesion~primer+application, data = paint, FUN= "mean" )
```

①

all treatment combination = 3×2 = 6

```
##   primer application adhesion
## 1      1           D 4.266667
## 2      2           D 5.300000
## 3      3           D 3.833333
## 4      1           S 5.300000
## 5      2           S 6.066667
## 6      3           S 5.166667
```

Application \ Primer	1	2	3	
D	$\bar{Y}_{1D\cdot}$	$\bar{Y}_{2D\cdot}$	$\bar{Y}_{3D\cdot}$	$\bar{Y}_{\cdot D\cdot}$
S	$\bar{Y}_{1S\cdot}$	$\bar{Y}_{2S\cdot}$	$\bar{Y}_{3S\cdot}$	$\bar{Y}_{\cdot S\cdot}$
	$\bar{Y}_{1\cdot\cdot}$	$\bar{Y}_{2\cdot\cdot}$	$\bar{Y}_{3\cdot\cdot}$	

②

```
tapply(adhesion,list(primer,application),mean)
```

in blue: marginal mean for primer / application.

```
##          D        S
## 1 4.266667 5.300000
## 2 5.300000 6.066667
## 3 3.833333 5.166667
```

Paint Data Example: two-way ANOVA

```
adhesion=c(4.0,4.5,4.3,5.6,4.9,5.4,3.8,3.7,
           4.0,5.4,4.9,5.6,5.8,6.1,6.3,5.5,5,5)
primer=factor(rep(rep(1:3,each=3),2))
application=factor(rep(1:2,each=9),labels=c("D","S"))
paint=data.frame(adhesion,primer,application)
with(paint,tapply(adhesion,primer,mean))
```

```
##        1        2        3
## 4.783333 5.683333 4.500000
```

← marginal means of primer factor

```
with(paint,tapply(adhesion,application,mean))
```

```
##        D        S
## 4.466667 5.511111
```

← marginal means of application factor

```
with(paint,table(primer,application))
```

```
##       application
## primer D S
##      1 3 3
##      2 3 3
##      3 3 3
```

← find replicates for each treatment combination, i.e. find n_{ij}

← $n_{ij} = r = 3 \Rightarrow$ balanced design

Paint Data Example: two-way ANOVA

```
##
## Call:
## lm(formula = adhesion ~ primer + application, data = paint)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4056 -0.2403  0.0000  0.2028  0.4778
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.2611     0.1396  30.523 3.29e-14 ***
## primer2        0.9000     0.1710   5.264  0.00012 ***
## primer3       -0.2833     0.1710  -1.657  0.11972
## applicationS   1.0444     0.1396   7.482 2.96e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2961 on 14 degrees of freedom
## Multiple R-squared:  0.8854, Adjusted R-squared:  0.8609
## F-statistic: 36.07 on 3 and 14 DF,  p-value: 7.716e-07
```

$$\begin{cases} \hat{\beta}_0 = \bar{Y}_{1D\cdot} \\ \hat{\beta}_1 = \bar{Y}_{2\cdot\cdot} - \bar{Y}_{1\cdot\cdot} \\ \hat{\beta}_2 = \bar{Y}_{3\cdot\cdot} - \bar{Y}_{1\cdot\cdot} \\ \hat{\beta}_3 = \bar{Y}_{\cdot S\cdot} - \bar{Y}_{\cdot D\cdot} \end{cases}$$

$H_0: \beta_1 = \beta_2 = \beta_3 = 0 \quad H_a: \exists \text{ one } i \text{ s.t. } \beta_i \neq 0 \qquad F = \dfrac{(SS_A + SS_B)/(\overset{2}{df_A} + \overset{1}{df_B})}{MSE}$

Paint Data Example: two-way ANOVA

default in R

Under constraints $\alpha_1 = 0, \beta_1 = 0$ for Two-way ANOVA without interaction term.

- intercept: cell mean at the level 1 of factor A and level 1 of factor B. ($\bar{Y}_{1D.}$)
- b_1 primer2: mean difference between level 2 and level 1 of factor A. ($\bar{Y}_{2..} - \bar{Y}_{1..}$)
- b_2 primer2: mean difference between level 3 and level 1 of factor A. ($\bar{Y}_{3..} - \bar{Y}_{1..}$)
- b_3 application S: mean difference between level 2 and level 1 of factor B. ($\bar{Y}_{.S.} - \bar{Y}_{.D.}$)

Common identifiablity constraints

Under two-way crossed ANOVA model

- Model without interaction term

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

  - Y_{ijk} is the k-th observation having received level i of factor 1 and level j of factor 2.
  - μ: unobserved mean value
  - α_i: unobserved deterministic effect of level i of factor 1
  - β_j: unobserved deterministic effect of level j of factor 2
  - $(\alpha\beta)_{ij}$: unobserved interaction effect for level i of factor 1 and level j of factor 2.
  - $\epsilon_{ijk} \sim_{iid} N(0, \sigma^2)$: unobserved random noise
- Model above subject to identifiability issue as before.
- Common constraints under above additive model
  - $\mu = 0$, $\alpha_1 = 0$, $(\alpha\beta)_{1j} = 0\ \forall j$ and $(\alpha\beta)_{i1} = 0\ \forall i$
  - $\alpha_1 = 0$ and $\beta_1 = 0$, $(\alpha\beta)_{1j} = 0\ \forall j$ and $(\alpha\beta)_{i1} = 0\ \forall i$ (default in R)
  - $\sum_i \alpha_i = 0, \sum_j \beta_j = 0$, $\sum_i (\alpha\beta)_{ij} = 0, \forall j$, and $\sum_j (\alpha\beta)_{ij} = 0, \forall i$

Paint Data Example: two-way ANOVA

```
##
## Call:
## lm(formula = adhesion ~ primer * application, data = paint)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40000 -0.16667  0.03333  0.21667  0.33333
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.2667     0.1656  25.772  7.1e-12 ***
## primer2                 1.0333     0.2341   4.414 0.000845 ***
## primer3                -0.4333     0.2341  -1.851 0.088949 .
## applicationS            1.0333     0.2341   4.414 0.000845 ***
## primer2:applicationS   -0.2667     0.3311  -0.805 0.436265
## primer3:applicationS    0.3000     0.3311   0.906 0.382736
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2867 on 12 degrees of freedom
## Multiple R-squared:  0.9079, Adjusted R-squared:  0.8696
## F-statistic: 23.67 on 5 and 12 DF,  p-value: 7.89e-06
```

$$\begin{cases} \hat{\beta}_0 = \bar{Y}_{1D.} \\ \hat{\beta}_1 = \bar{Y}_{2D.} - \bar{Y}_{1D.} \\ \hat{\beta}_2 = \bar{Y}_{3D.} - \bar{Y}_{1D.} \end{cases}$$

$\hat{\beta}_3 = \bar{Y}_{1S.} - \bar{Y}_{1D.}$

$\hat{\beta}_4, \hat{\beta}_5$:

$(\widehat{\alpha\beta})_{ij} = \bar{Y}_{ij.} - \bar{Y}_{iD.} - \bar{Y}_{1j.} + \bar{Y}_{1D.}$

$$\begin{cases} H_0: \alpha_i = \beta_j = (\alpha\beta)_{ij} = 0, \ \forall i, j \\ H_a: \text{at least one is} \neq 0 \end{cases}$$

Want to know more about the contrasts, in R:

- **model.matrix(lm(adhesion~primer*application,data=paint))**

Paint Data Example: two-way ANOVA

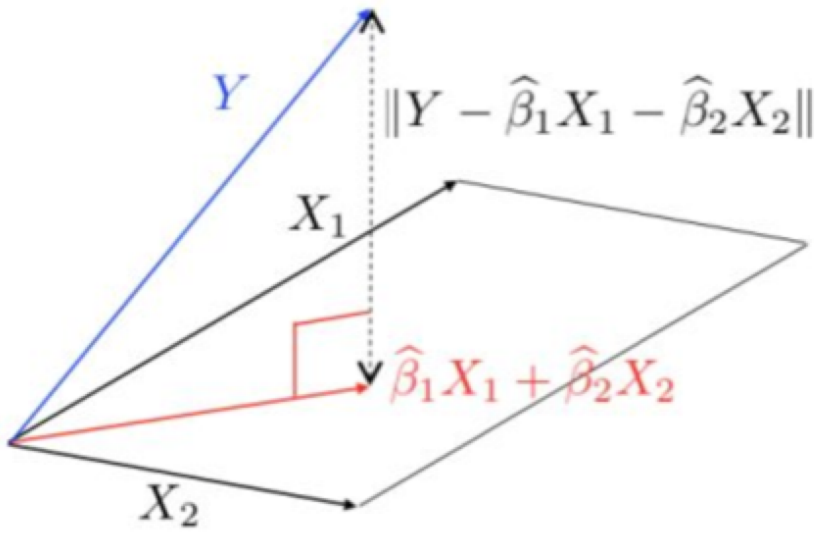
From two-way crossed model: using the default constraints setting in R
(Let P=primer (1,2,3), A=application(D,S))

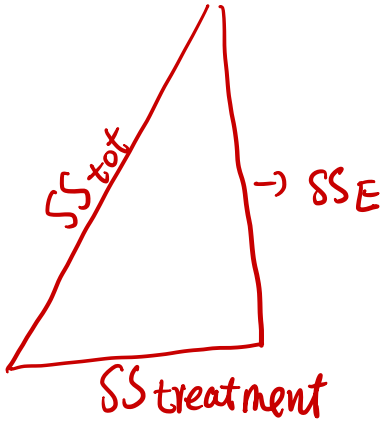
- intercept: mean outcome for P=1, A=D ,i.e. $\bar{Y}_{1D.}$
- primer2: mean difference in outcome between P=2 and P=1 when A=D.i.e. $\bar{Y}_{2D.} - \bar{Y}_{1D.}$
- primer3: mean difference in outcome between P=3 and P=1 when A=D.i.e. $\bar{Y}_{3D.} - \bar{Y}_{1D.}$
- applicationS: mean difference in outcome between A=S and A=D when P=1.i.e. $\bar{Y}_{1S.} - \bar{Y}_{1D.}$
- primer2:applicatonS: $\bar{Y}_{2S.} - \bar{Y}_{2D.} - \bar{Y}_{1S.} + \bar{Y}_{1D.}$
- primer3:applicatonS: $\bar{Y}_{3S.} - \bar{Y}_{3D.} - \bar{Y}_{1S.} + \bar{Y}_{1D.}$
- in general (in R): $\widehat{(\alpha\beta)}_{ij} = \bar{Y}_{ij.} - \bar{Y}_{i1.} - \bar{Y}_{1j.} + \bar{Y}_{11}$

Geometric view of ANOVA

With the matrix notation introduced earlier, $Y = \cancel{G}X\beta + \epsilon$, and under the least square estimation principle, we have:

- ANOVA model fit as an orthogonal projection
  - $\hat{Y} = \cancel{G}X\hat{\beta}$ is an element of **Span(X)** that minimizes $||Y - \hat{Y}||^2$. $= ||Y - X\hat{\beta}||^2$
  - $\hat{Y}$ is the orthogonal projection of Y on **Span(X)**





Understand and interpret interaction effect in Two-way ANOVA

Two-way factor effect model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \quad \epsilon_{ijk} \sim_{iid} N(0, \sigma^2)$$

- $(\alpha\beta)_{ij}$ is the interaction term, when it presents, it means that a comparison of the levels of α will depend on the level of β. Interpretation is not simple.
- Textbook constraint: $\sum_i \alpha_i = \sum_j \beta_j = \sum (\alpha\beta)_{ij} = 0$
  - Under this constraints everything gets compared to the GRAND MEAN
  - Grand mean: estimate $\hat{\mu} = \bar{Y}_{...}$
  - Main effect: $\hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...}, \hat{\beta}_i = \bar{Y}_{.j.} - \bar{Y}_{...}$
  - Interaction: $\widehat{(\alpha\beta)}_{ij} = \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}$
- R default constraint: $\alpha_1 = \beta_1 = (\alpha\beta)_{ij} = 0$ for i=1 and j=1
  - Under this constraints everything gets compared to the mean for the last level of each factor.
  - reference mean: estimate $\hat{\mu} = \bar{Y}_{11.}$
  - Main effect: $\hat{\alpha}_i = \bar{Y}_{i1.} - \bar{Y}_{11.}, \hat{\beta}_i = \bar{Y}_{1j.} - \bar{Y}_{11.}$
  - Interaction: $\widehat{(\alpha\beta)}_{ij} = \bar{Y}_{ij.} - \bar{Y}_{i1.} - \bar{Y}_{1j.} + \bar{Y}_{11.}$
- In either case, differences are identified and contrasts can be used with the **results being exactly the same**.

General Strategy for two-way ANOVA with interaction model

- Set up model with main effects and interaction(s), check assumptions, and examine interaction(s)
- If no significant interaction, examine main effects individually, using appropriate adjustments for multiple comparisons, main effects plots, etc. (Read 19.1 in KNNL about pooling)
- If interaction is significant, determine whether interactions are important. If not, can examine main effects as in Step 2.
- If interaction present & important, determine whether interaction is simple or complex.
- For simple interactions, can still talk about the main effects of A at each level of B
- For complex interaction, must simply consider all pairs of levels as separate treatments.

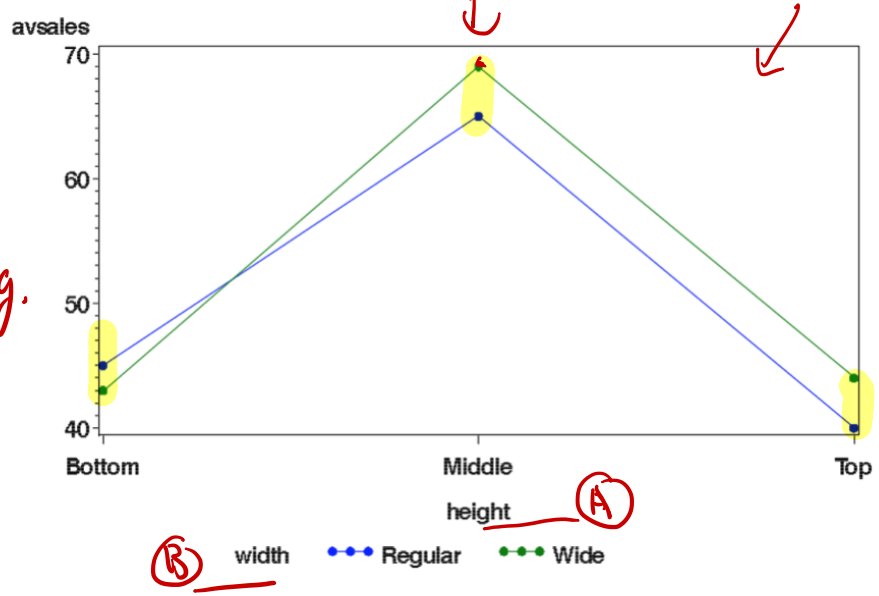
Unimportant interaction

Middle, Wide

In interaction plot, each point corresponds to a cell mean

main effect A: sig.

main effect B: non-sig.



① to see main effect A, check marginal mean changes

② to see main effect B, see gap between lines.

- If interaction effects are very small compared to main effects or only apparent in a small number of treatments, then they are probably unimportant.
- Lines will be not quite parallel, but close. → unimportant interaction.
- We can proceed by keeping interaction in the model, but using marginal means for each significant main effect individually
  - Marginal means: Averages over the levels of the other factor.

Important interaction

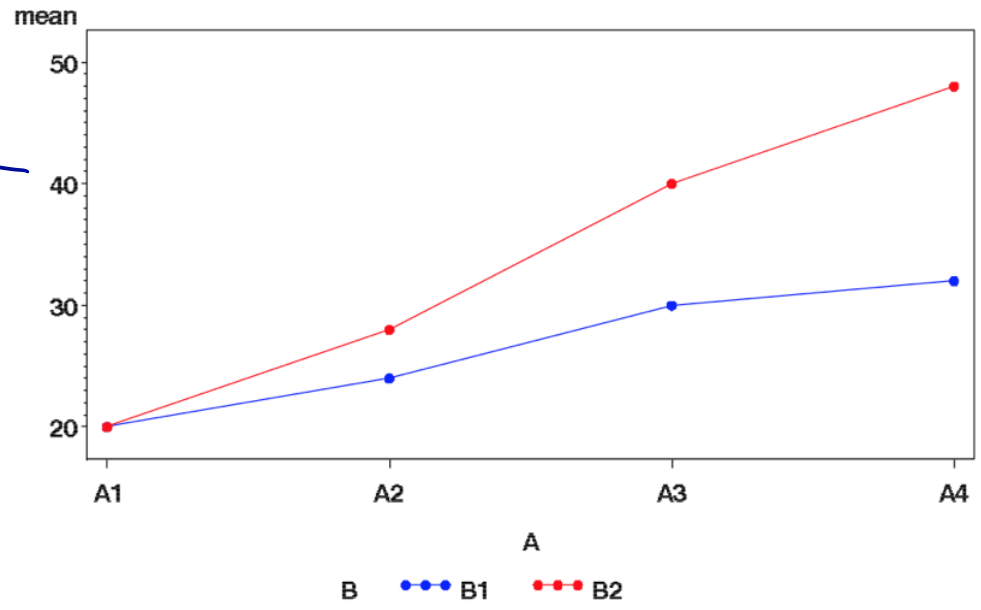
- The interaction effect is so **large and/or pervasive** that main effects cannot be interpreted on their own.
- In interaction plots, the lines will not be parallel. They may or may not criss-cross, but the differences between levels for one factor will **depend** on the level of the other factor
- Analyze interaction – Similar to interpreting as a one-way ANOVA with ab levels
- Report that the interaction is significant; plot the means and describe the pattern
- Discuss results for the levels of A for each level of B or vice versa

Simple VS Complex Interaction

- An interaction is considered **simple** if we can discuss trends for the main effect of one factor for each level of the other factor, and if the general trend is the same.
- An interaction is **complex** if it is difficult to discuss anything about the main effects. In this situation, one can only look at treatment combinations and cannot separate them into main effects easily.

Example 1: important interaction

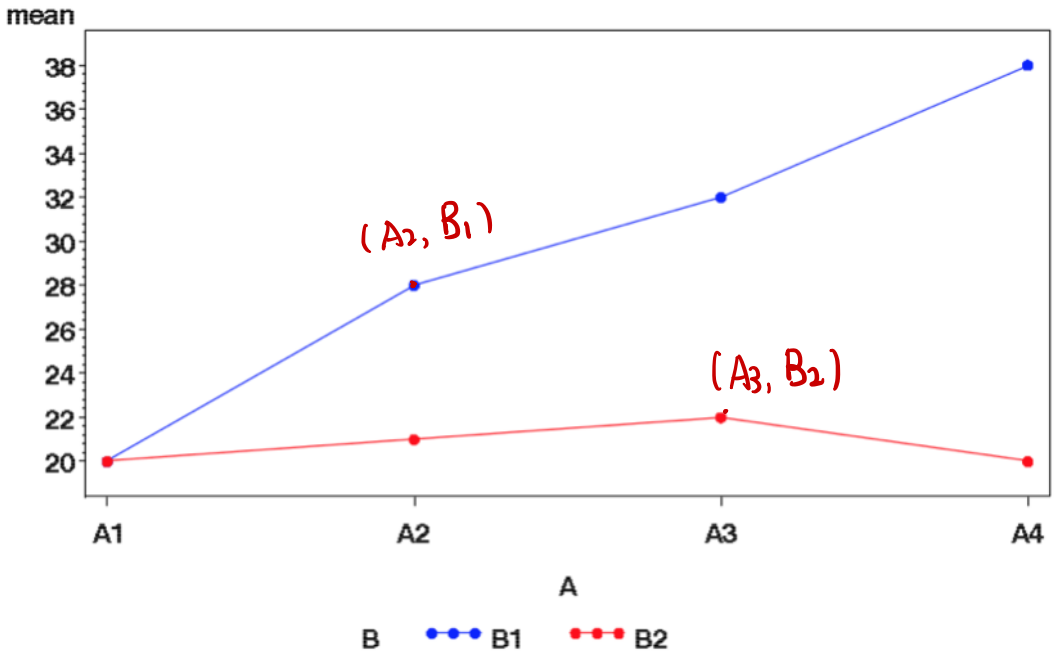
A: sig.
B: sig.
(AB): sig.



Simple.
- General trends are the same
- effect of of A at $B=B_2 > B=B_1$

- **Main effect of A is bigger for B = B2 than for B = B1**
- Both main effect and interaction effect exist
- When increasing one level of Factor A also increases the effect of Factor B, this is called a **reinforcement effect**. The other way is called an **interference effect**. We have reinforcement effect in this case.

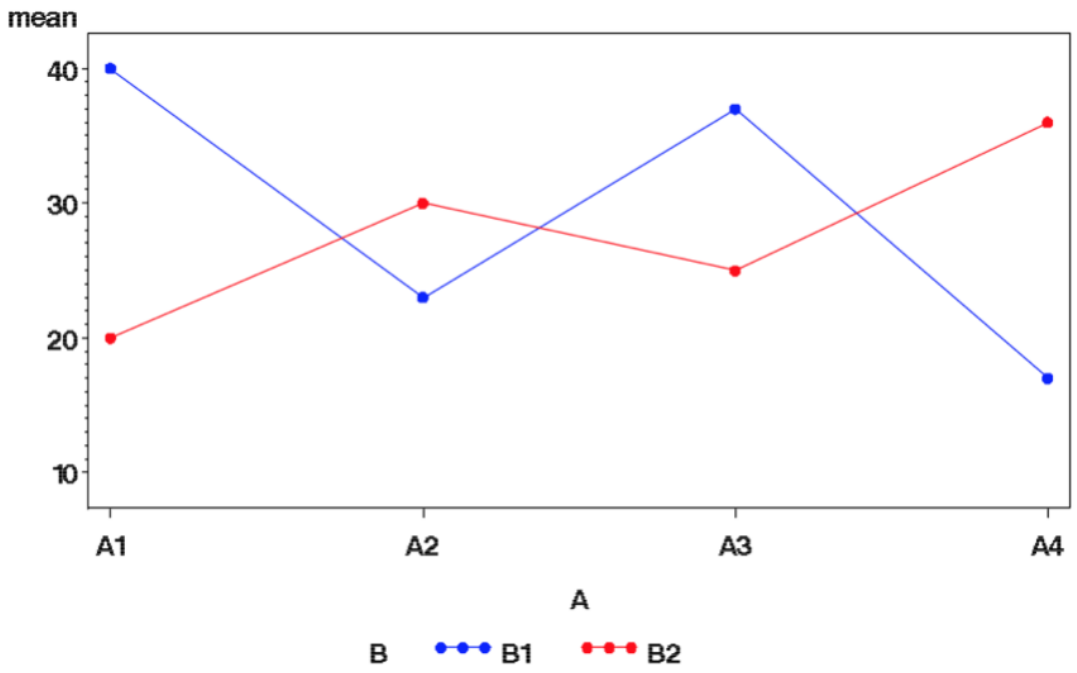
Example 2: important interaction



- **Suppression: Effect of A is suppressed if B = B2**
- Both main effect and interaction effect exist

Simple

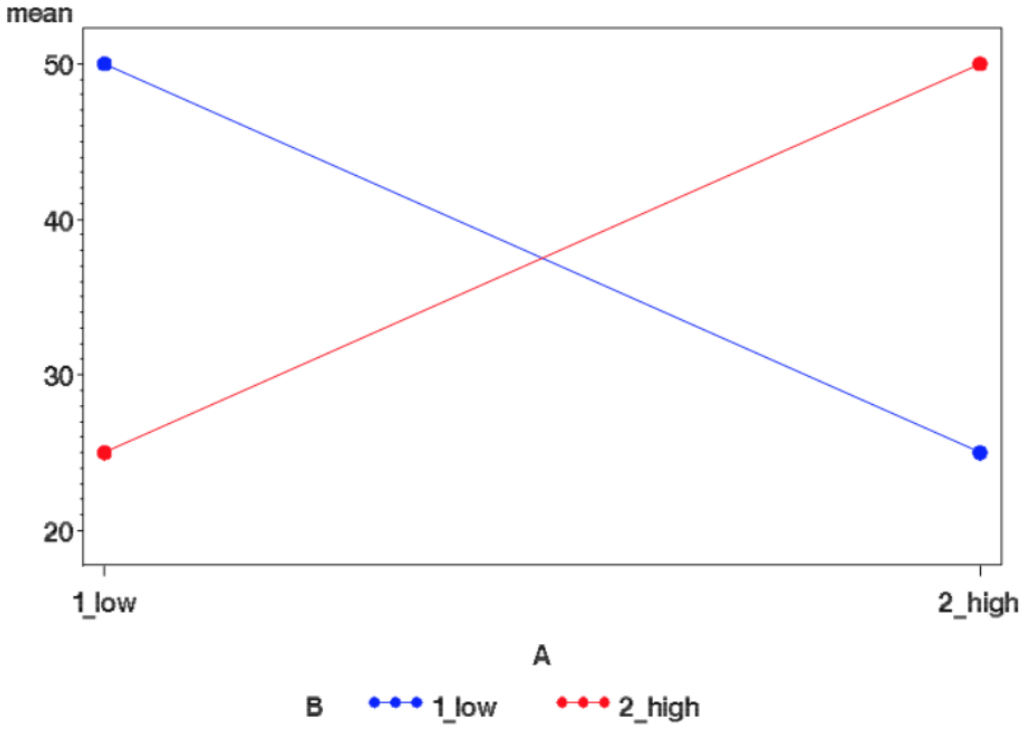
Example 3: important interaction



Complicated Interaction: Cannot separate effects

complex

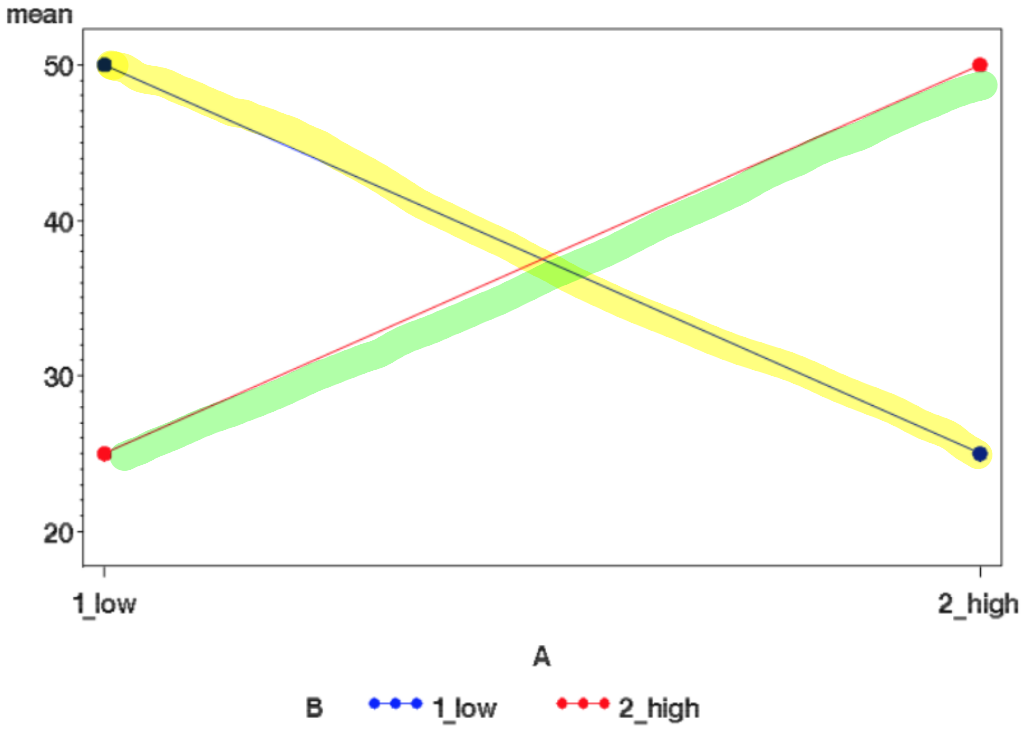
Example 4: important interaction



⇒ interaction exists and it's complex.

- In this case, main effects would "appear to be zero". But this is misleading and inaccurate.
  - If you averaged over either factor, you would find "no change" when going from one level to the other.
  - In fact there is a change when going from one level to the next, and the type of change depends on the level of the second factor. (This is a good "definition" for interaction.)

Example 4: important interaction



Conclusions:

- At the low level of factor B, increasing A from low to high decreases the mean response.
- At the high level of factor B, increasing A from low to high increases the mean response.
- You cannot make statements here about Factor A alone or Factor B alone.

main effects A and B.

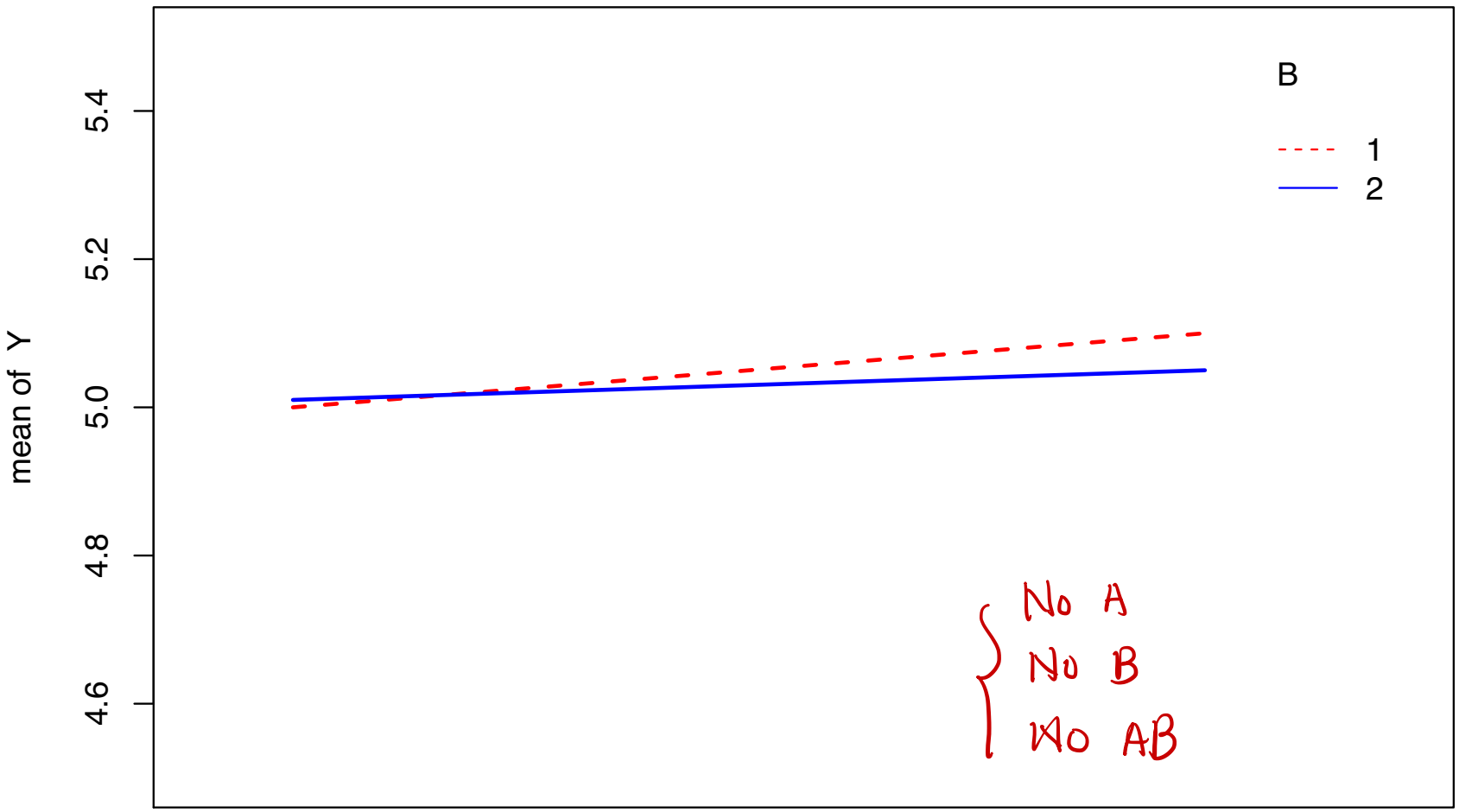
When interaction effect is significant, we usually skip the interpretation of

Take a break, and see you on Thursday

ANOVA with interaction effect (cooked example)

No significant effect

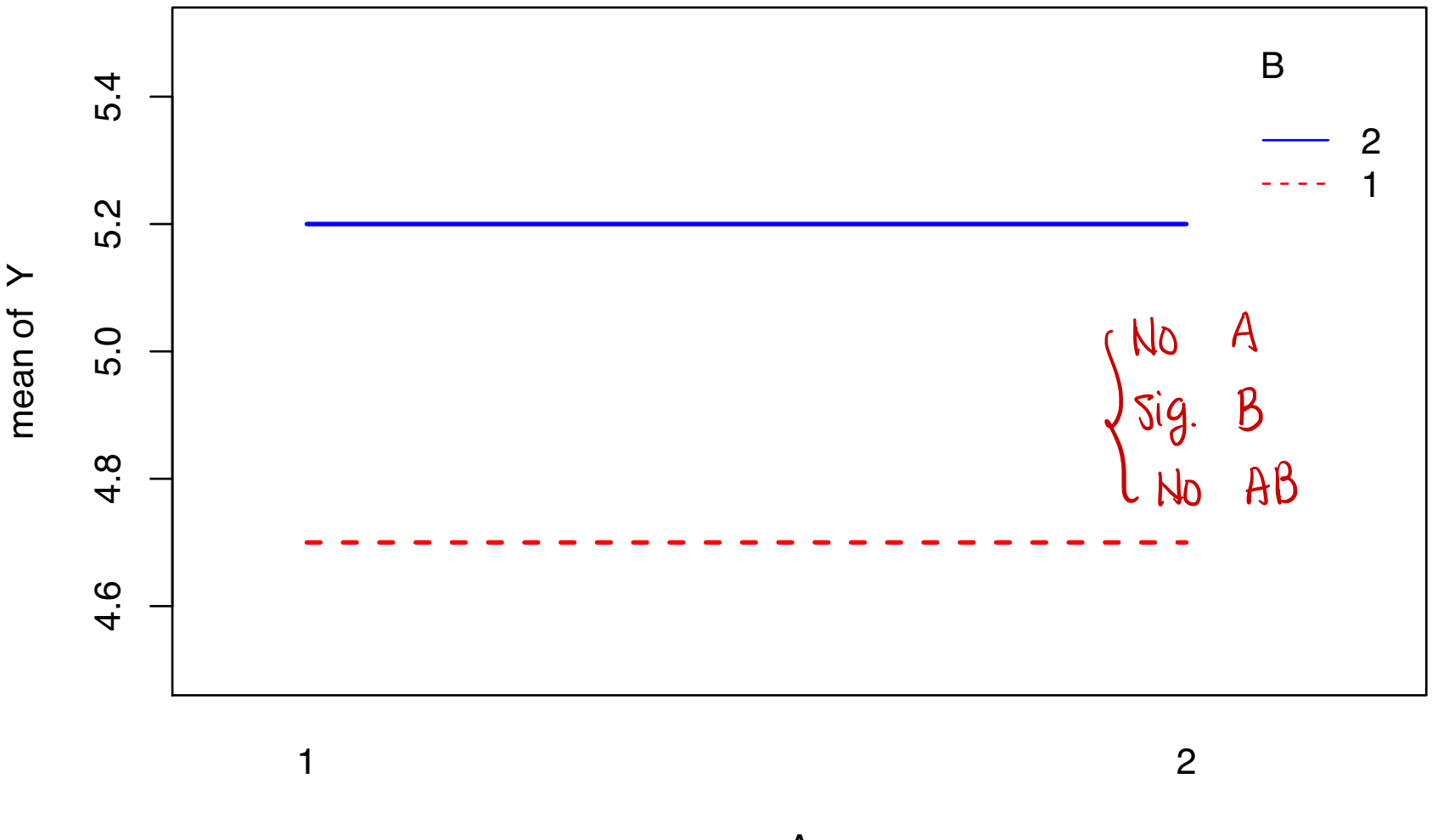
```
A=as.factor(c(1,1,2,2))
B=as.factor(c(1,2,1,2))
Y=c(5,5.01,5.1,5.05)
interaction.plot(A,B,Y,col=c("red","blue"),lwd=2,ylim=c(4.5,5.5))
```



ANOVA with interaction effect (cooked example)

Significant B effect, but no other effect

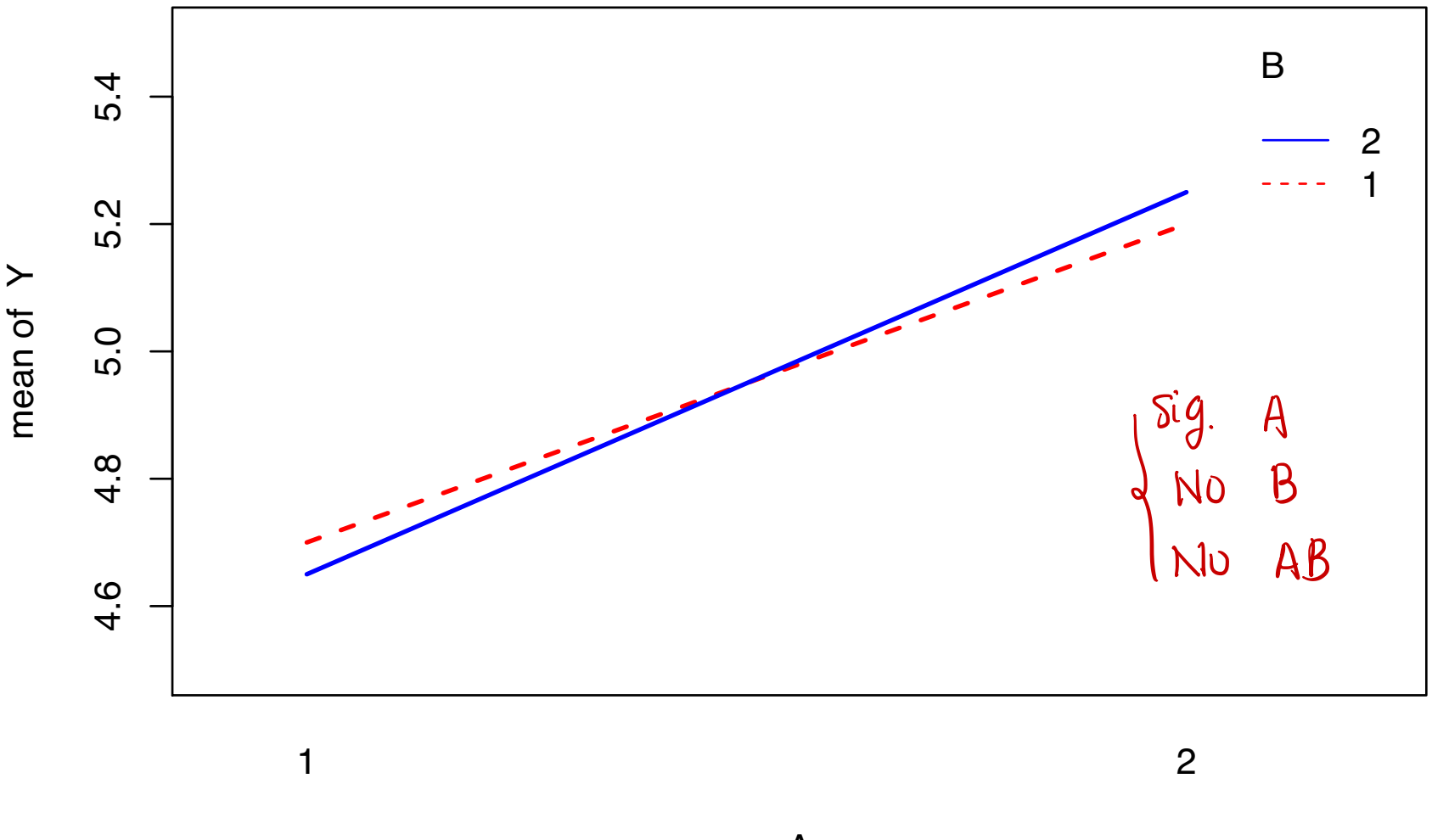
```
A=as.factor(c(1,1,2,2))
B=as.factor(c(1,2,1,2))
Y=c(4.7,5.2,4.7,5.2)
interaction.plot(A,B,Y,col=c("red","blue"),lwd=2,ylim=c(4.5,5.5))
```



ANOVA with interaction effect (cooked example)

Significant A effect, but no other effect

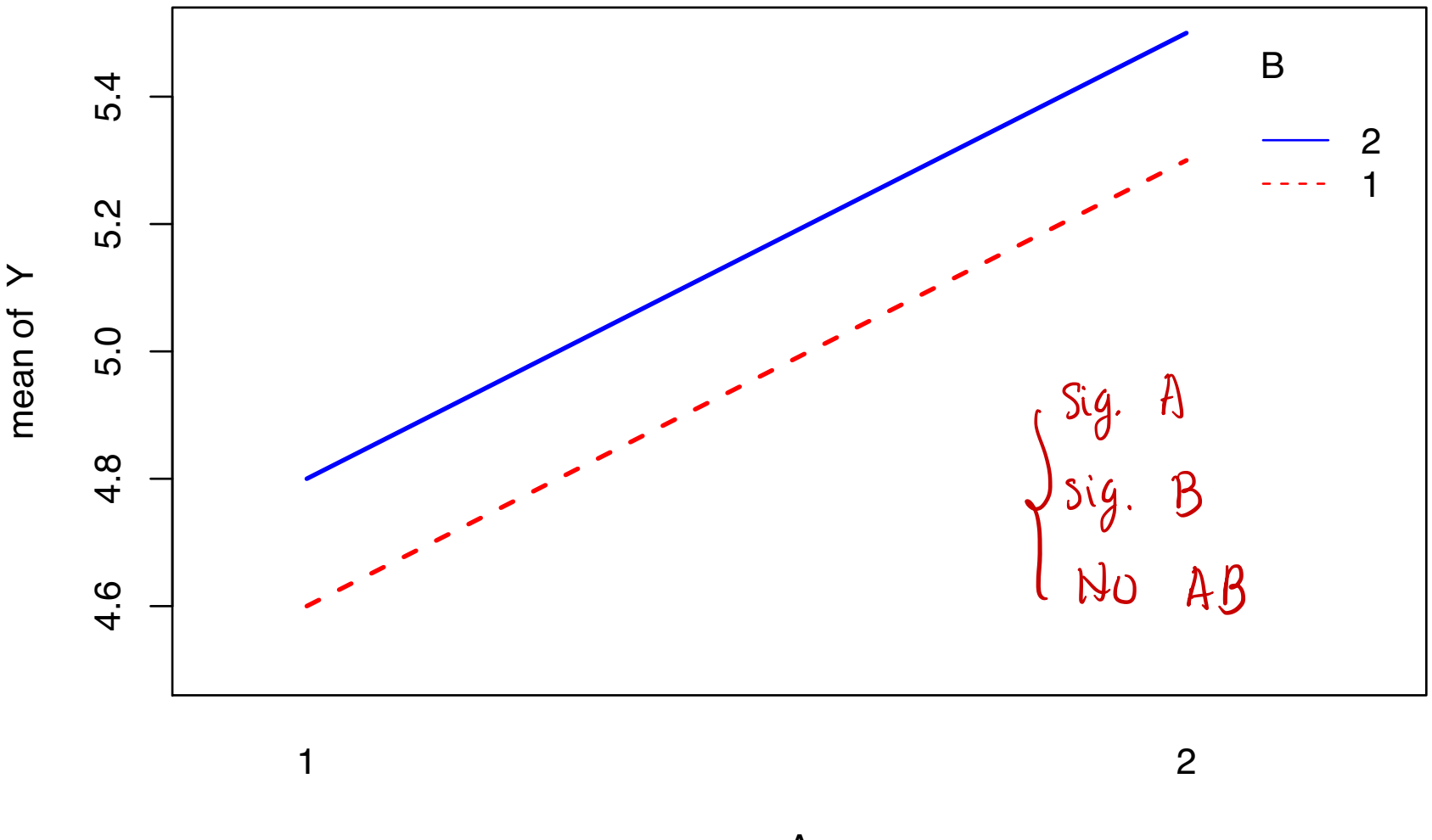
```
A=as.factor(c(1,1,2,2))
B=as.factor(c(1,2,1,2))
Y=c(4.70,4.65,5.2,5.25)
interaction.plot(A,B,Y,col=c("red","blue"),lwd=2,ylim=c(4.5,5.5))
```



ANOVA with interaction effect (cooked example)

Significant A and B effect, but no interaction effect

```
A=as.factor(c(1,1,2,2))
B=as.factor(c(1,2,1,2))
Y=c(4.6,4.8,5.3,5.5)
interaction.plot(A,B,Y,col=c("red","blue"),lwd=2,ylim=c(4.5,5.5))
```



Interaction study: Battery Example

- Study the effects of A = type of material and B = temperature on the lifetime of a battery (in hours).
- Factor A: three material types: Nickel- Cadmium, Nickel-Metal Hydride, and Lithium-Ion
- Factor B: Three temperatures: 15, 70, and 125 degrees Fahrenheit

```
battery=read.table("/Users/Wei/TA/Teaching/battery.txt",sep="",header=T)
battery$type=as.factor(battery$type)
battery$temp=as.factor(battery$temp)
str(battery)
```

```
## 'data.frame':    36 obs. of  3 variables:
##  $ time: int  130 155 74 180 34 40 80 75 20 70 ...
##  $ type: Factor w/ 3 levels "Lith","NiCd",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ temp: Factor w/ 3 levels "15","70","125": 1 1 1 1 2 2 2 2 3 3 ...
```

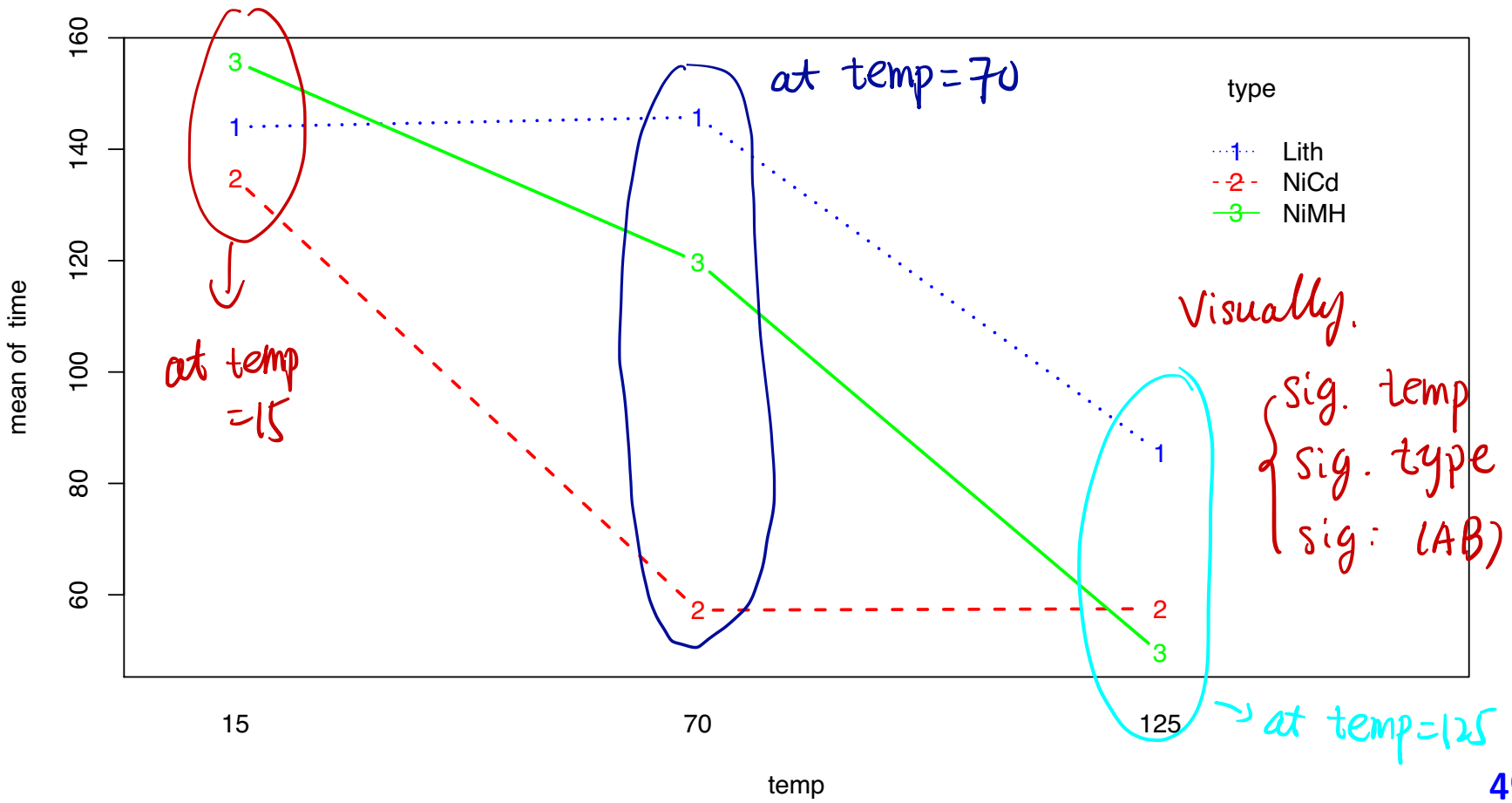
Battery Example

- Four observations per cell $n_{ij} = 4$ $\Rightarrow$ balanced design.
- Main aim is to examine the effects and hopefully find a material that will help the battery have a uniformly long life in the field.
- Steps in analysis:
  1. Check interaction plot
  2. Review ANOVA results / assumptions
  3. Check main effects if appropriate
  4. Draw conclusion

Battery Example - Interaction plot

- Interaction here is complicated but quite informative
- It appears the Ni-Cd battery is "worst"
- The plot seems to suggest that the Lithium ion battery is superior.

```
with(battery,interaction.plot(temp,type,time,
                              col=c("blue","red","green"),lwd=2,type="b"))
```



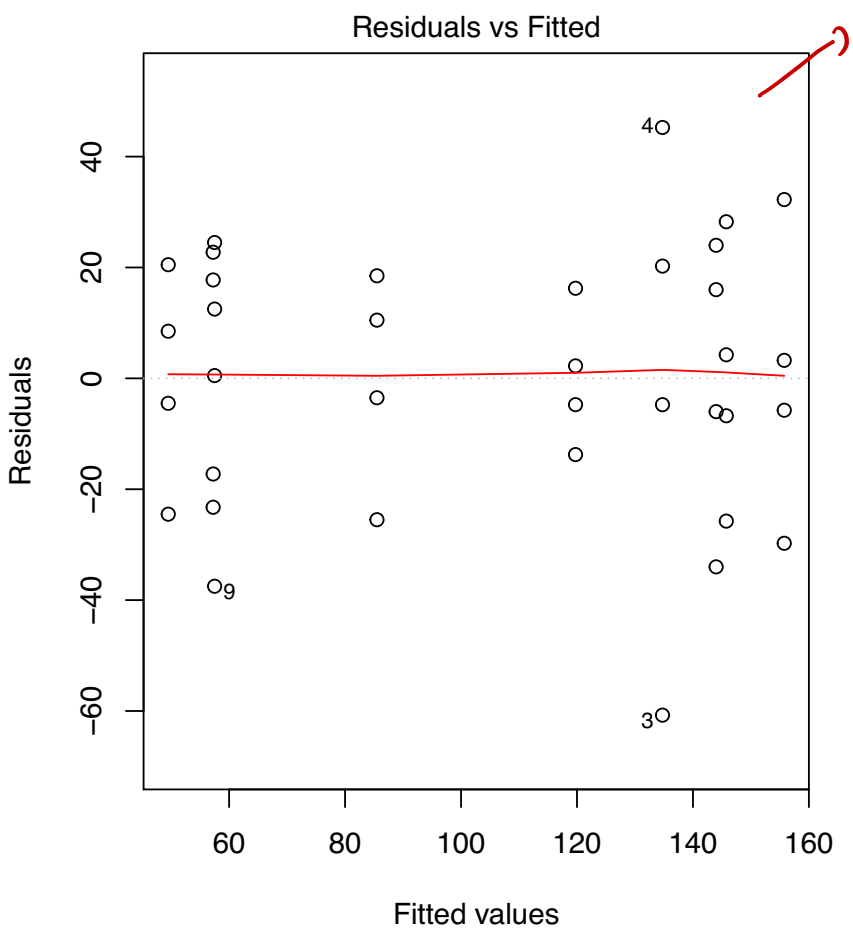
Battery Example - Checking assumption

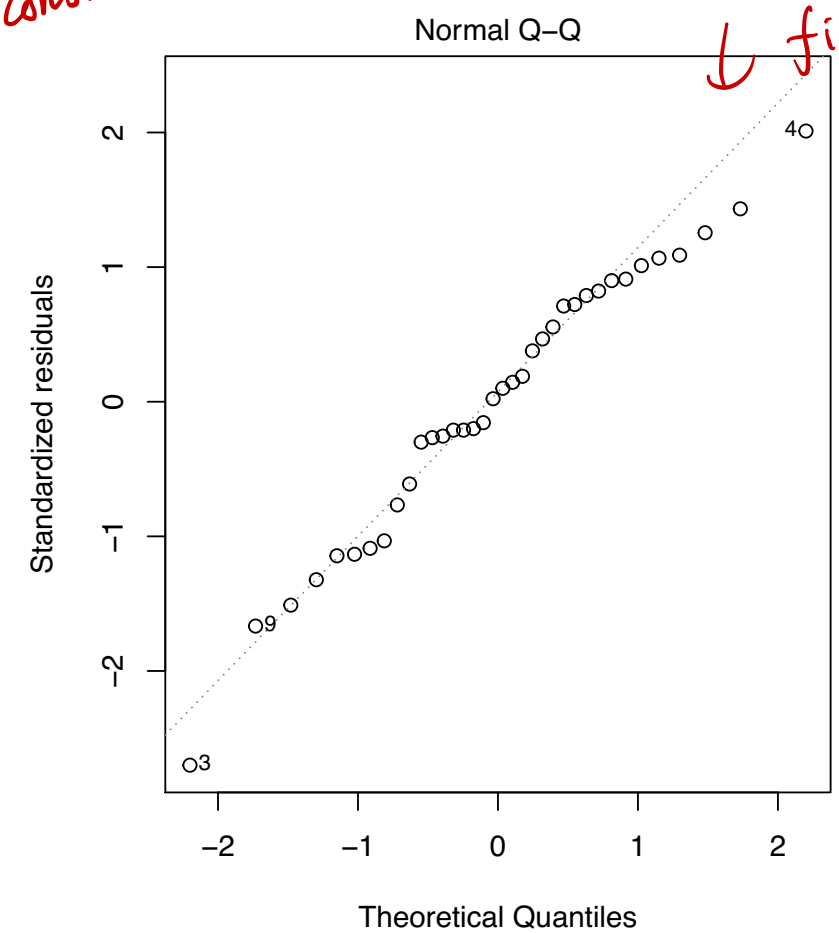
- No major violations of the assumptions are evident from reviewing the plots

```
fit=lm(time~type*temp,data=battery)
par(mfrow=c(1,2))
plot(fit,which=1:2)
```

constant σ^2 ✓

Normality: fine.





Battery Example - ANOVA

- All effects are significant, can look at multiple comparisons

```r
fit=lm(time~type*temp,data=battery)
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: time
##           Df Sum Sq Mean Sq F value    Pr(>F)
## type       2  10684  5341.9  7.9114  0.001976 **
## temp       2  39119 19559.4 28.9677 1.909e-07 ***
## type:temp  4   9614  2403.4  3.5595  0.018611 *
## Residuals 27  18231   675.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

→ type: sig

→ temp: sig

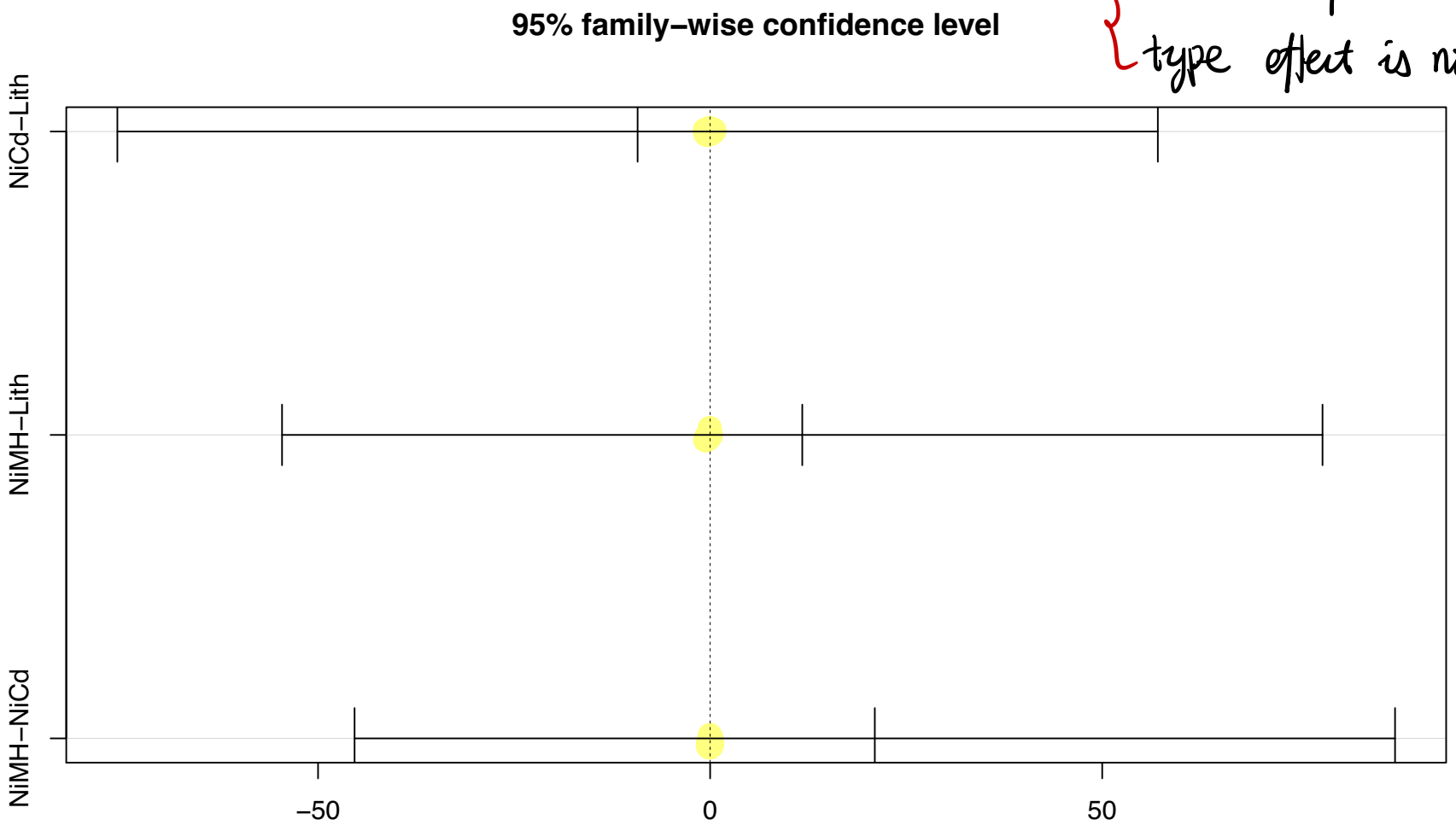
→ (AB): sig.

Battery Example - multiple comparison

- Conclusion 1: At 15 degrees there is no significant difference among the material type.

```
fit=aov(time~type,data=subset(battery,temp==15))
plot(TukeyHSD(fit, "type",conf.level=0.95))
```

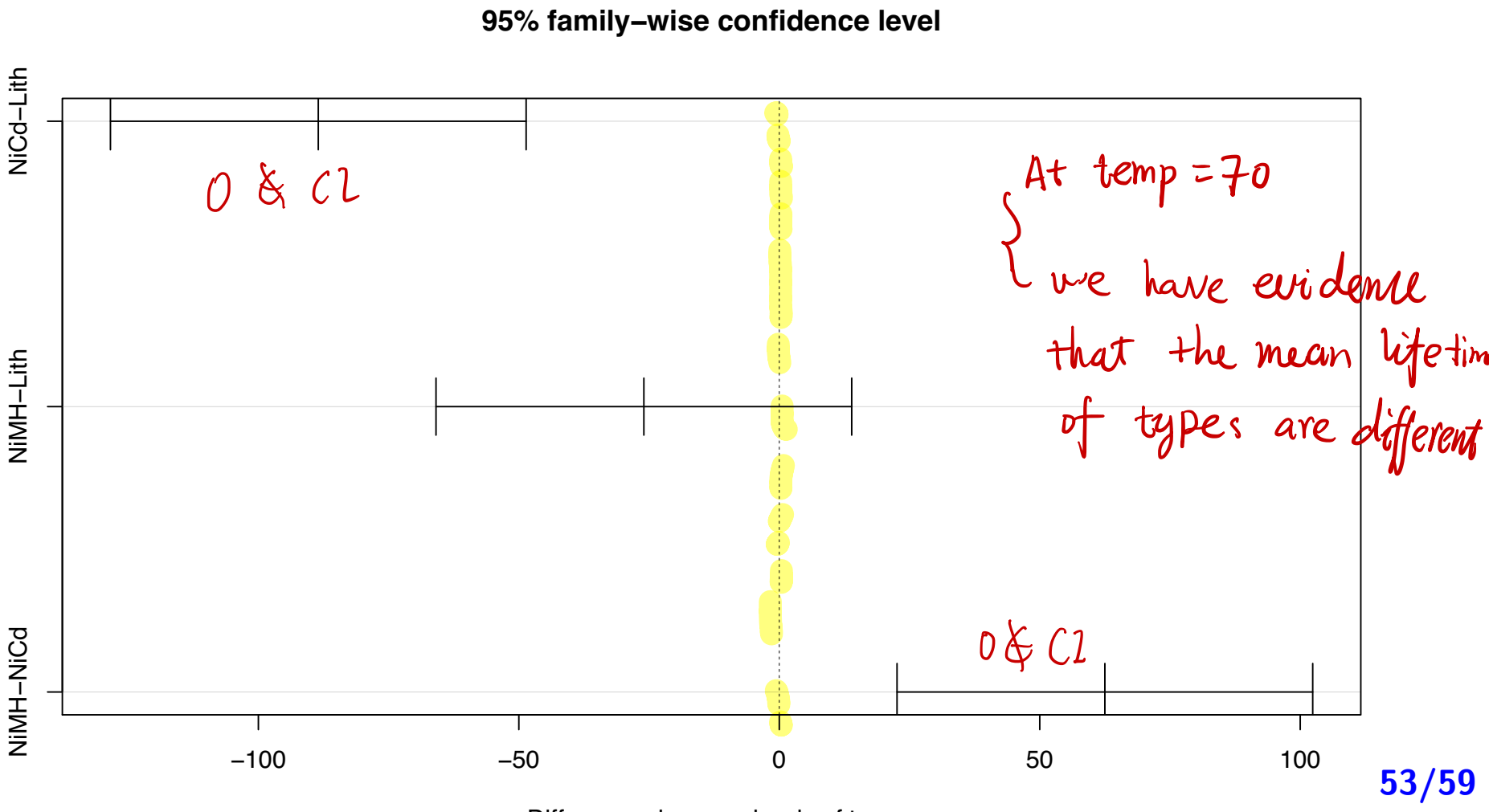
Fixed temp = 15
type effect is not sig.



Battery Example - multiple comparison

- Conclusion 2: At 70 degrees Lith or NiMH are significantly different from Ni-Cd.

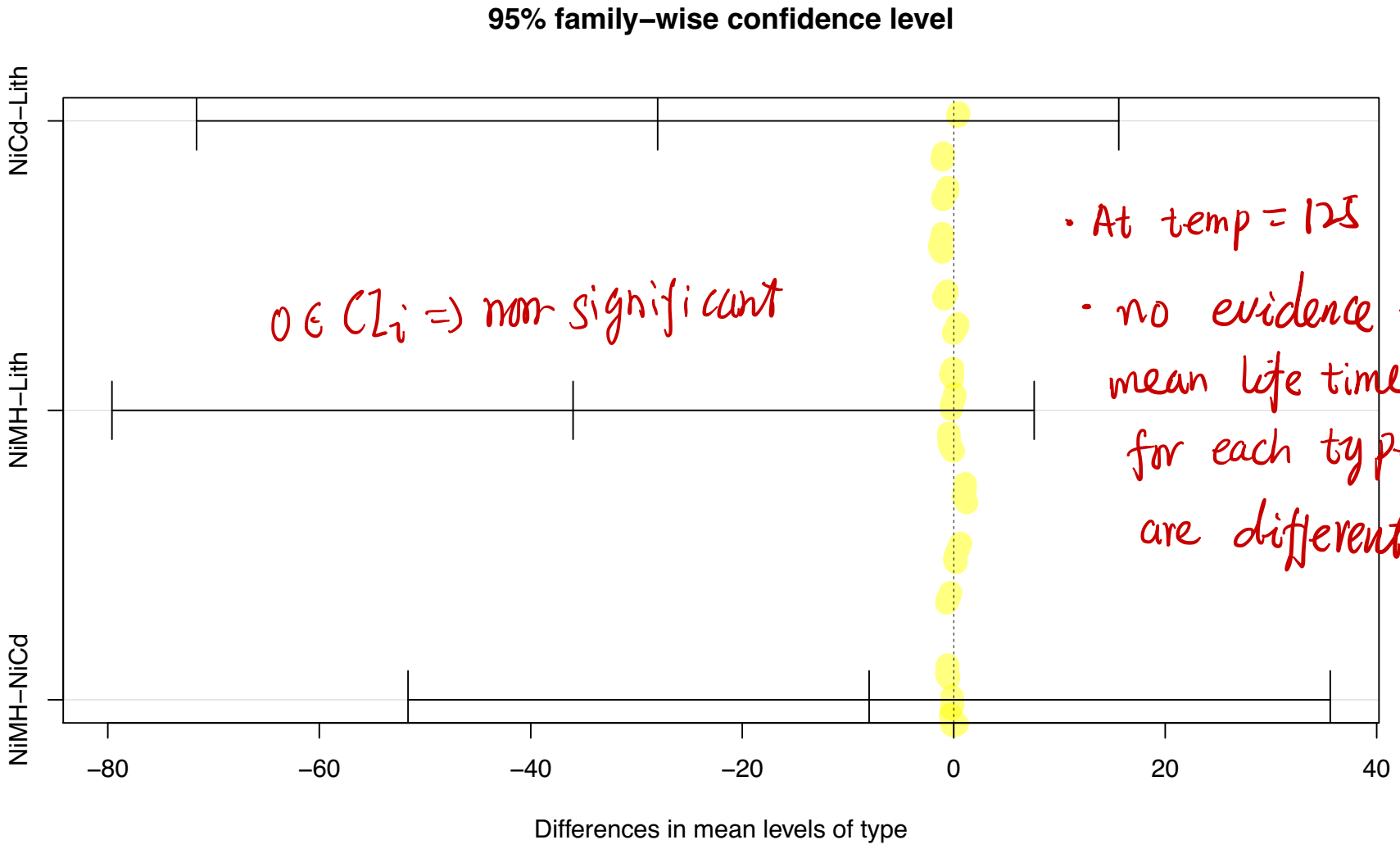
```
fit=aov(time~type,data=subset(battery,temp==70))
plot(TukeyHSD(fit, "type",conf.level=0.95))
```



Battery Example - multiple comparison

- Conclusion 2: At 125 degrees there was no significant difference.

```
fit=aov(time~type,data=subset(battery,temp==125))
plot(TukeyHSD(fit, "type",conf.level=0.95))
```



Understanding Hypothesis Testing

For any Hypothesis testing, it consists of the following component:

- Null and alternative hypothesis
- Test statistic
- The null distribution, i.e. the distribution for the test statistic when H_0 is true.
  - Given significance level α, we find the R.R. (rejection Region), $\alpha = P(R.R.|H_0)$
  - R.R. splits the whole real line into two exclusive set. $R.R. \cup R.R.^C$
- Under H_a, the test statistics following another distribution with $\theta \in H_a$, the type II error and power are found under this distribution.
  - $\beta = P(\text{Type II error}) = P(R.R.^C|H_a)$
  - Power=P(correctly reject $H_0|H_0$: false)= P(correctly reject $H_0|H_a$: True)=1-β=P($R.R.|H_a$)

Understanding Hypothesis Testing

Example: Assume $X_1, \cdots X_n \sim N(\mu, \sigma^2)$ and σ^2 is further assumed known.

① Want to test

$H_0: \mu = 0$ $\qquad$ $H_a: \mu > 0$

② Test statistic: $T = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ ← function of $\underbrace{\text{data}}_{X_1 \cdots X_n}$ & $\underbrace{\text{one unkown}}_{\mu}$

③ under H_0: $T_0 = \frac{\bar{X} - 0}{\sigma/\sqrt{n}} \sim N(0,1)$ ← null distribution

④ Given siginificance level $\alpha = 5\%$, find R.R.

R.R. = {Rejection region} = {area we reject H_0}

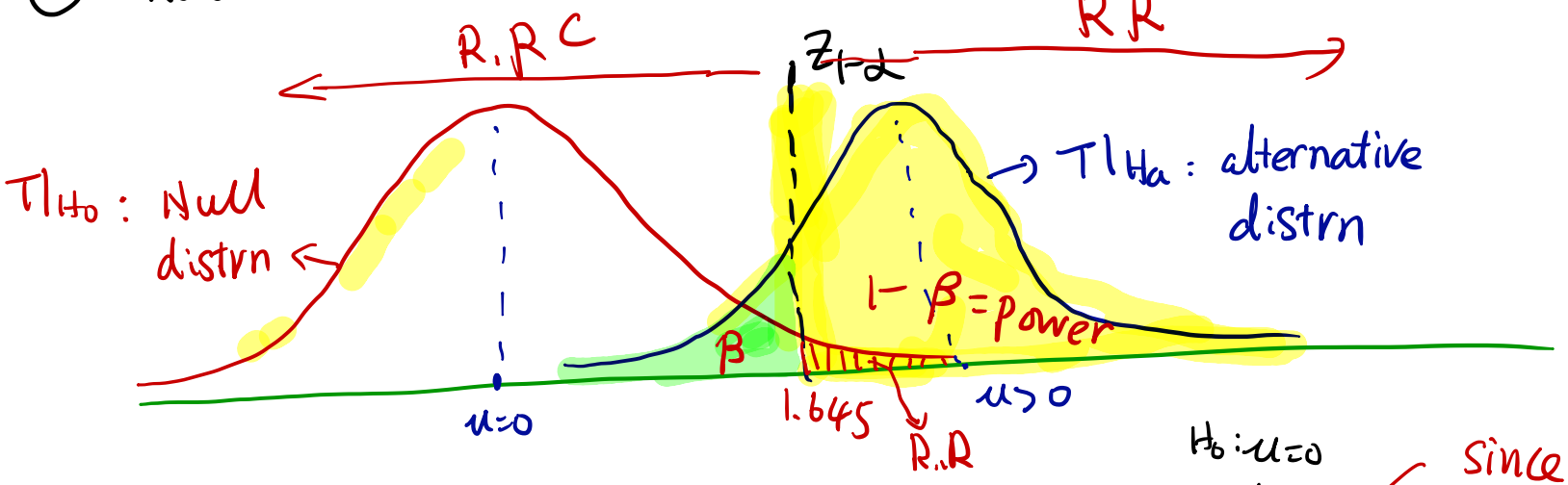
= {area we accept H_a}

Since $H_a: \mu > 0$, it is on the right of $\mu = 0$

$\Rightarrow$ R.R. = {Right tail of T_0 distribution}

Understanding Hypothesis Testing

④ Null distribution + α $\Longrightarrow$ R.R.



$H_0: \mu = 0$

since $H_a: \mu > 0$

$$\alpha = P(\text{type I error}) = P(R.R. \mid H_0) = P\left(\frac{\bar{x}-0}{\sigma/\sqrt{n}} \geq 1.64\right)$$

$$\Rightarrow R.R. = \left\{ \bar{x} \geq 1.64 \frac{\sigma}{\sqrt{n}} \right\}$$

$$\beta = P(\text{type II error}) = P(\text{fail to reject } H_0 \mid \overbrace{H_0 \text{ is false}}^{H_a: T})$$

$$= P(R.R.^c \mid H_a: T) = P(\text{Green area})$$

$$\Rightarrow \text{power} = 1-\beta = 1 - P(R.R \mid H_a) = P(\text{Yellow area})$$

Understanding Hypothesis Testing

⑤ Power = P(correctly reject H_0 when H_0 is false)

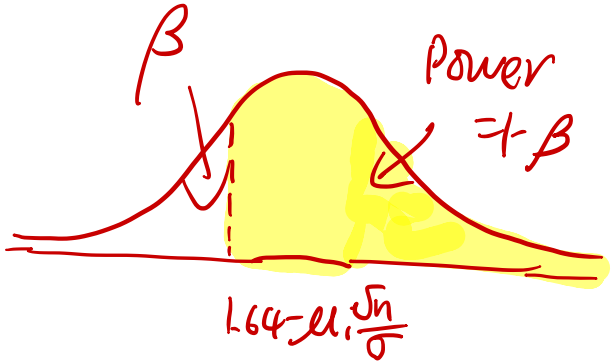
= P(correctly reject H_0 when H_a is true)

$= P(R.R. \mid H_a) \leftarrow = 1-\beta = 1 - P(R.R.^c \mid H_a)$

continue the example

$$\text{power} = P\left(\bar{X} \geq 1.64\frac{\sigma}{\sqrt{n}} \;\middle|\; H_a: \mu_1 > 0\right)$$

$$= P\left(\underbrace{\frac{\bar{X}-\mu_1}{\sigma/\sqrt{n}}}_{\sim N(0,1)} \geq \frac{1.64\frac{\sigma}{\sqrt{n}} - \mu_1}{\sigma/\sqrt{n}}\right)$$



$$= P\left(Z \geq 1.64 - \mu_1\frac{\sqrt{n}}{\sigma}\right) = \Phi\left(1.64 - \mu_1\frac{\sqrt{n}}{\sigma}\right)$$

After Lecture This Week

Practice problems

- Check portal this week, Assignment 1 will be posted.
- Try the posted practice problem on two-way ANOVA (solution is available too)
- Try all the R example in slides.
- Review all types of t-test and one-way ANOVA analysis.
- Background reading: KNNL Ch19.

Topics for next week:

- ANOCA
  - ANOVA with covariate analysis
- Weighted Least Squares and Ridge Regression