

General Factorial Designs

Yibi Huang

- 8.7 General Factorial Designs
 - Definition of 3-way and k-way interactions
 - 3-way interaction plots
 - Parameter estimates
 - Sum of Squares, dfs, and the ANOVA table
- 8.11 Hierarchy
- 8.9 Single Replicates

3-Way Interaction Contrast

Based on the means model $y_{ijkl} = \mu_{ijk} + \varepsilon_{ijkl}$ of a 3-way design, a 3-way interaction contrast between level (i_1, i_2) for factor A, level (j_1, j_2) for factor B, and level (k_1, k_2) for factor C is defined to be

$$\mu_{i_1 j_1 k_1} - \mu_{i_2 j_1 k_1} - \mu_{i_1 j_2 k_1} - \mu_{i_1 j_1 k_2} + \mu_{i_2 j_2 k_1} + \mu_{i_2 j_1 k_2} + \mu_{i_1 j_2 k_2} - \mu_{i_2 j_2 k_2}$$

Observe that any two μ_{ijk} 's in the contrast have

opposite signs if they differ by an **odd** number of indexes.
identical signs if they differ by an **even** number of indexes.

The 3-way interaction contrast above has 3 interpretations:

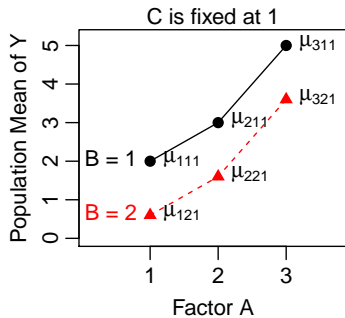
$$\begin{aligned} & \mu_{i_1 j_1 k_1} - \mu_{i_2 j_1 k_1} - \mu_{i_1 j_2 k_1} - \mu_{i_1 j_1 k_2} + \mu_{i_2 j_2 k_1} + \mu_{i_2 j_1 k_2} + \mu_{i_1 j_2 k_2} - \mu_{i_2 j_2 k_2} \\ = & \underbrace{(\mu_{i_1 j_1 k_1} - \mu_{i_2 j_1 k_1} - \mu_{i_1 j_2 k_1} + \mu_{i_2 j_2 k_1})}_{\text{AB interaction contrast when } C = k_1} - \underbrace{(\mu_{i_1 j_1 k_2} - \mu_{i_2 j_1 k_2} - \mu_{i_1 j_2 k_2} + \mu_{i_2 j_2 k_2})}_{\text{AB interaction contrast when } C = k_2} \\ = & \underbrace{(\mu_{i_1 j_1 k_1} - \mu_{i_1 j_2 k_1} - \mu_{i_1 j_1 k_2} + \mu_{i_1 j_2 k_2})}_{\text{BC interaction contrast when } A = i_1} - \underbrace{(\mu_{i_2 j_1 k_1} - \mu_{i_2 j_2 k_1} - \mu_{i_2 j_1 k_2} + \mu_{i_2 j_2 k_2})}_{\text{BC interaction contrast when } A = i_2} \\ = & \underbrace{(\mu_{i_1 j_1 k_1} - \mu_{i_2 j_1 k_1} - \mu_{i_1 j_1 k_2} + \mu_{i_2 j_1 k_2})}_{\text{AC interaction contrast when } B = j_1} - \underbrace{(\mu_{i_1 j_2 k_1} - \mu_{i_2 j_2 k_1} - \mu_{i_1 j_2 k_2} + \mu_{i_2 j_2 k_2})}_{\text{AC interaction contrast when } B = j_2} \end{aligned}$$

Three-Way Interactions

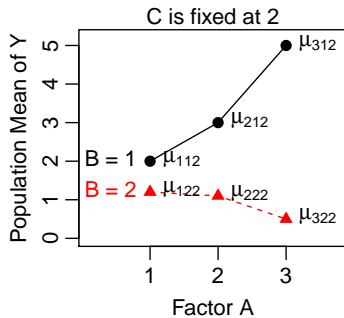
We say factor A, B, and C have **three-way interactions** if

- ▶ an AB interaction contrast changes with the levels of C, or
- ▶ a BC interaction contrast changes with the levels of A, or
- ▶ an AC interaction contrast changes with the levels of B.

E.g.,



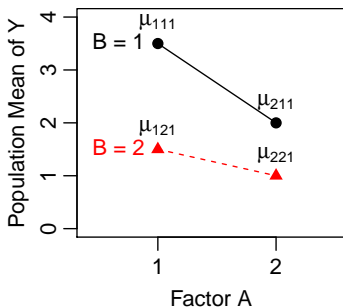
No AB interactions
when C is fixed at 1



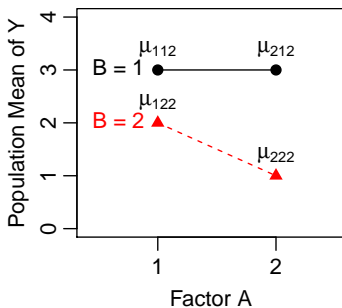
AB have interactions
when C is fixed at 2

Example 2: Three-Way Interactions

AB have interactions
when C is fixed at 1



AB have interactions
when C is fixed at 2

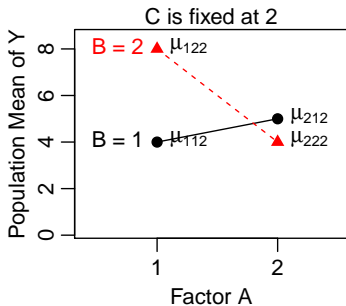
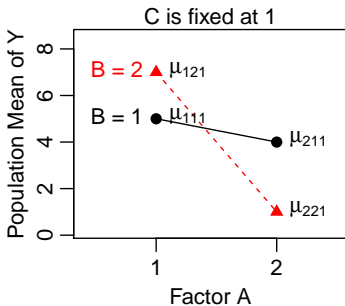


$$(\mu_{111} - \mu_{211}) - (\mu_{121} - \mu_{221}) > 0$$

$$\underbrace{(\mu_{112} - \mu_{212})}_{=0} - \underbrace{(\mu_{122} - \mu_{222})}_{>0} < 0$$

The AB interaction contrast $(\mu_{11k} - \mu_{21k}) - (\mu_{12k} - \mu_{22k})$ depends on the level k of factor C . Hence there exist ABC 3-way interactions.

It can be hard to judge graphically whether ABC interaction is present when AB interactions exist at both levels of C.



	C = 1		C = 2	
	B = 1	B = 2	B = 1	B = 2
A = 1	$\mu_{111} = 5$	$\mu_{121} = 7$	$\mu_{112} = 4$	$\mu_{122} = 8$
A = 2	$\mu_{211} = 4$	$\mu_{221} = 1$	$\mu_{212} = 5$	$\mu_{222} = 4$

The AB interactions at the two levels of C are equal and hence there is no ABC interaction.

$$(\mu_{111} - \mu_{211}) - (\mu_{121} - \mu_{221}) = (5 - 4) - (7 - 1) = -5$$

$$(\mu_{112} - \mu_{212}) - (\mu_{122} - \mu_{222}) = (4 - 5) - (8 - 4) = -5$$

Higher Order Interactions

- ▶ An ABCD 4-way interaction contrast is
 - ▶ the difference of some ABC 3-way interaction contrast at two different levels of D
 - ▶ the difference of some ABD 3-way interaction contrast at two different levels of C
 - ▶ the difference of some ACD 3-way interaction contrast at two different levels of B
 - ▶ the difference of some BCD 3-way interaction contrast at two different levels of A
- ▶ We say ABCD have **4-way interactions** if any of the ABCD 4-way interaction contrast is non-zero or if any 3-way interaction contrast between any 3 of the 4 factors changes with the levels of a 4th factor.
 - ▶ e.g., if some ACD 3-way interaction contrast changes with the levels of factor B, then there exist ABCD 4-way interaction
- ▶ We say k factors have **k -way interactions** means the $(k - 1)$ -way interaction of any $(k - 1)$ of the k factors changes with the levels of a k th factor.

General Factorial Models

The model and analysis of multi-way factorial data are generalization of those for two-way factorial data. E.g., consider a 4-way factorial design with factors A, B, C, and D.

$$\text{means model: } y_{ijklm} = \mu_{ijkl} + \varepsilon_{ijklm} \quad \text{for } \begin{cases} i = 1, \dots, a, j = 1, \dots, b, \\ k = 1, \dots, c, \ell = 1, \dots, d, \\ m = 1, \dots, n. \end{cases}$$

$$\begin{aligned} \text{effects model: } y_{ijklm} = & \underbrace{\mu}_{\text{grand mean}} + \underbrace{\alpha_i + \beta_j + \gamma_k + \delta_\ell}_{\text{main effects}} \\ & + \underbrace{\alpha\beta_{ij} + \alpha\gamma_{ik} + \alpha\delta_{il} + \beta\gamma_{jk} + \beta\delta_{jl} + \gamma\delta_{kl}}_{\text{2-way interactions}} \\ & + \underbrace{\alpha\beta\gamma_{ijk} + \alpha\beta\delta_{ijl} + \alpha\gamma\delta_{ikl} + \beta\gamma\delta_{jkl}}_{\text{3-way interactions}} \\ & + \underbrace{\alpha\beta\gamma\delta_{ijkl}}_{\text{4-way interaction}} + \underbrace{\varepsilon_{ijklm}}_{\text{error}} \end{aligned}$$

Zero-Sum Constraints for General Factorial Models

$$\begin{aligned}y_{ijklm} = & \mu + \alpha_i + \beta_j + \gamma_k + \delta_\ell \\ & + \alpha\beta_{ij} + \alpha\gamma_{ik} + \alpha\delta_{i\ell} + \beta\gamma_{jk} + \beta\delta_{j\ell} + \gamma\delta_{k\ell} \\ & + \alpha\beta\gamma_{ijk} + \alpha\beta\delta_{ij\ell} + \alpha\gamma\delta_{ik\ell} + \beta\gamma\delta_{jkl} \\ & + \alpha\beta\gamma\delta_{ijkl} + \varepsilon_{ijklm}\end{aligned}$$

All the effects have zero-sum constraints that they add to 0 when summing over any subscript, e.g.,

- ▶ $\sum_i \alpha_i = \sum_j \beta_j = \sum_k \gamma_k = \sum_\ell \delta_\ell = 0$
- ▶ $\sum_i \alpha\gamma_{ik} = \sum_k \alpha\gamma_{ik} = 0$, for all i, k ,
so do other 2-way interactions
- ▶ $\sum_i \alpha\gamma\delta_{ik\ell} = \sum_k \alpha\gamma\delta_{ik\ell} = \sum_\ell \alpha\gamma\delta_{ik\ell} = 0$, for all i, k, ℓ ,
so do other 3-way interactions
- ▶ $\sum_i \alpha\beta\gamma\delta_{ijkl} = \sum_j \alpha\beta\gamma\delta_{ijkl} = \sum_k \alpha\beta\gamma\delta_{ijkl} = \sum_\ell \alpha\beta\gamma\delta_{ijkl} = 0$,
for all i, j, k, ℓ .

Parameter Estimates

For a 4-way model, the parameter estimates under the zero-sum constraints are

grand mean	$\hat{\mu} = \bar{y}_{\bullet\bullet\bullet\bullet}$
main effects	$\hat{\alpha}_i = \bar{y}_{i\bullet\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet\bullet}, \quad \hat{\beta}_j = \bar{y}_{\bullet j\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet\bullet},$ $\hat{\gamma}_k = \bar{y}_{\bullet\bullet k\bullet} - \bar{y}_{\bullet\bullet\bullet\bullet}, \quad \hat{\delta}_l = \bar{y}_{\bullet\bullet\bullet l} - \bar{y}_{\bullet\bullet\bullet\bullet}$
2-way	$\hat{\alpha\beta}_{ij} = \bar{y}_{ij\bullet\bullet} - \bar{y}_{i\bullet\bullet\bullet} - \bar{y}_{\bullet j\bullet\bullet} + \bar{y}_{\bullet\bullet\bullet\bullet}$ $\hat{\beta\gamma}_{jk} = \bar{y}_{\bullet jk\bullet} - \bar{y}_{\bullet j\bullet\bullet} - \bar{y}_{\bullet\bullet k\bullet} + \bar{y}_{\bullet\bullet\bullet\bullet}$ \vdots
3-way	$\hat{\alpha\beta\delta}_{ijl} = \bar{y}_{ij\bullet l} - \bar{y}_{ij\bullet\bullet} - \bar{y}_{i\bullet\bullet l} - \bar{y}_{\bullet j\bullet l}$ $\quad + \bar{y}_{i\bullet\bullet\bullet} + \bar{y}_{\bullet j\bullet\bullet} + \bar{y}_{\bullet\bullet\bullet l} - \bar{y}_{\bullet\bullet\bullet\bullet}$ $\hat{\alpha\gamma\delta}_{ikl} = \dots$
4-way	$\hat{\alpha\beta\gamma\delta}_{ijkl} = (16 \text{ terms, see the next page})$

$$\begin{aligned}
\widehat{\alpha\beta\gamma\delta}_{ijkl} &= \bar{y}_{ijkl\bullet} \\
&\quad - \bar{y}_{ijk\bullet\bullet} - \bar{y}_{ij\bullet l\bullet} - \bar{y}_{i\bullet kl\bullet} - \bar{y}_{\bullet jkl\bullet} \\
&\quad + \bar{y}_{ij\bullet\bullet\bullet} + \bar{y}_{i\bullet k\bullet\bullet} + \bar{y}_{i\bullet\bullet l\bullet} + \bar{y}_{\bullet jk\bullet\bullet} + \bar{y}_{\bullet j\bullet l\bullet} + \bar{y}_{\bullet\bullet kl\bullet} \\
&\quad - \bar{y}_{i\bullet\bullet\bullet\bullet} - \bar{y}_{\bullet j\bullet\bullet\bullet} - \bar{y}_{\bullet\bullet k\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet l\bullet} \\
&\quad + \bar{y}_{\bullet\bullet\bullet\bullet\bullet} \\
&= (\text{terms that average over 1 index}) \\
&\quad - (\text{terms that average over 2 indexes}) \\
&\quad + (\text{terms that average over 3 indexes}) \\
&\quad - (\text{terms that average over 4 indexes}) \\
&\quad + (\text{terms that average over 5 indexes})
\end{aligned}$$

Sum of Squares

SST can be decomposed into SS of main effects and interactions of all orders, e.g., in an $a \times b \times c \times d$ design with n replicates:

$$\begin{aligned}SST &= SS_A + SS_B + SS_C + SS_D \\ &+ SS_{AB} + SS_{AC} + SS_{AD} + SS_{BC} + SS_{BD} + SS_{CD} \\ &+ SS_{ABC} + SS_{ACD} + SS_{ABD} + SS_{BCD} \\ &+ SS_{ABCD} \\ &+ SSE\end{aligned}$$

where $SST = \sum_{ijklm} (y_{ijklm} - \bar{y}_{\dots\dots})^2$, $SSE = \sum_{ijklm} (y_{ijklm} - \bar{y}_{ijkl\bullet})^2$, and the SS for all other terms are the **sum of squares of corresponding parameter estimates under the zero sum constraints**, e.g.,

$$\begin{aligned}SS_C &= \sum_{ijklm} (\hat{\gamma}_k)^2 = abdn \sum_k (\hat{\gamma}_k)^2 \\ SS_{BC} &= \sum_{ijklm} (\hat{\beta}\hat{\gamma}_{jk})^2 = adn \sum_{jk} (\hat{\beta}\hat{\gamma}_{jk})^2 \\ SS_{ACD} &= \sum_{ijklm} (\hat{\alpha}\hat{\gamma}\hat{\delta}_{ikl})^2 = bn \sum_{ikl} (\hat{\alpha}\hat{\gamma}\hat{\delta}_{ikl})^2 \\ SS_{ABCD} &= \sum_{ijklm} (\hat{\alpha}\hat{\beta}\hat{\gamma}\hat{\delta}_{ijkl})^2 = n \sum_{ijkl} (\hat{\alpha}\hat{\beta}\hat{\gamma}\hat{\delta}_{ijkl})^2\end{aligned}$$

Degrees of Freedom

Say factor A, B, C, and D have respectively a , b , c , d levels, and there are n replicates.

- ▶ d.f. of a main effect = number of levels -1 .
e.g., $df_A = a - 1$, $df_C = c - 1$.
- ▶ d.f. of an interaction = product of d.f.'s for the main effects of the involved factors, e.g.,
 - ▶ $df_{AD} = (a - 1)(d - 1)$,
 - ▶ $df_{BCD} = (b - 1)(c - 1)(d - 1)$,
 - ▶ $df_{ABCD} = (a - 1)(b - 1)(c - 1)(d - 1)$.
- ▶ d.f. of SST = total # of observation $-1 = abcdn - 1$
- ▶ d.f. of SSE = total # of observation $-$ total # of treatments
 $= abcdn - abcd = abcd(n - 1)$

Example 8.10 Amylase data (p.195)

- ▶ Goal: to study the amylase specific activity of sprouted maize
- ▶ An $8 \times 2 \times 2$ factorial design with 3 factors:
 - ▶ analysis temperature (40, 35, 30, 25, 20, 15, 13, or 10°C)
 - ▶ growth temperature of the sprouts (25 or 13°C)
 - ▶ variety of maize (B73 or Oh43)
- ▶ 3 replicates per treatment
- ▶ Response: the amylase specific activities (IU)
- ▶ You may load the data using the command

```
amyl = read.table(  
  "http://www.stat.uchicago.edu/~yibi/s222/amylaze.txt", h=T)
```

Example 8.10 Amylase data

Table 8.9: Amylase specific activity (IU), for two varieties of sprouted maize under different growth and analysis temperatures (degrees C).

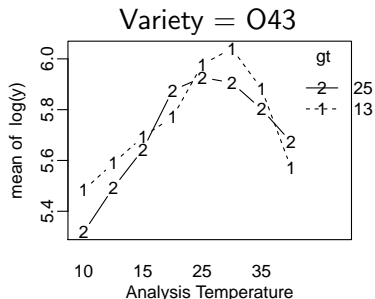
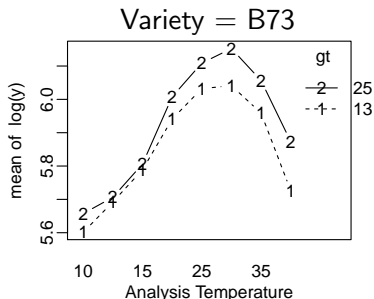
GT	Var.	Analysis Temperature							
		40	35	30	25	20	15	13	10
25	B73	391.8	427.7	486.6	469.2	383.1	338.9	283.7	269.3
		311.8	388.1	426.6	436.8	408.8	355.5	309.4	278.7
		367.4	468.1	499.8	444.0	429.0	304.5	309.9	313.0
	O43	301.3	352.9	376.3	373.6	377.5	308.8	234.3	197.1
		271.4	296.4	393.0	364.8	364.3	279.0	255.4	198.3
		300.3	346.7	334.7	386.6	329.2	261.3	239.4	216.7
13	B73	292.7	422.6	443.5	438.5	350.6	305.9	319.9	286.7
		283.3	359.5	431.2	398.9	383.9	342.8	283.2	266.5
		348.1	381.9	388.3	413.7	408.4	332.2	287.9	259.8
	O43	269.7	380.9	389.4	400.3	340.5	288.6	260.9	221.9
		284.0	357.1	420.2	412.8	309.5	271.8	253.6	254.4
		235.3	339.0	453.4	371.9	313.0	333.7	289.5	246.7

```
> amyl = read.table(
      "http://www.stat.uchicago.edu/~yibi/s222/amylaze.txt", h=T)
> str(amyl)
'data.frame': 96 obs. of 7 variables:
 $ atemp : int  40 35 30 25 20 15 13 10 40 35 ...
 $ gtemp : int  25 25 25 25 25 25 25 25 13 13 ...
 $ variety: Factor w/ 2 levels "B73","043": 1 1 1 1 1 1 1 1 1 1 ...
 $ y      : num  392 428 487 469 383 ...

> amyl$at = as.factor(amyl$atemp)
> amyl$gt = as.factor(amyl$gtemp)
> amyl$v  = as.factor(amyl$variety)
```

Three-Way Interaction Plots — Amylase data

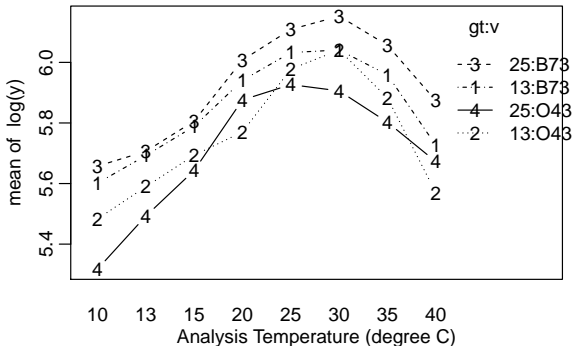
```
with(subset(amyl,v=="B73"),  
      interaction.plot(at,gt,log(y),type="b",xlab="Analysis Temperature"))  
with(subset(amyl,v=="O43"),  
      interaction.plot(at,gt,log(y),type="b",xlab="Analysis Temperature"))
```



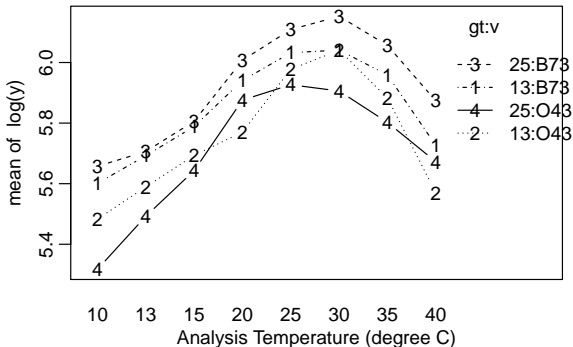
For both varieties, there appear little `at:gt` interaction as the 2 curves are pretty close in shape. Keep in mind that lines in an interaction plot may not be exactly parallel due to noise even if there is no interaction as we plot it using the sample means, not the population means. Hence, there's little signs of 3-way interactions.

Or one can combine both plots into one.

```
with(amy1, interaction.plot(at, gt:v, log(y), type="b",  
                           xlab="Analysis Temperature (degree C)"))
```



- ▶ Little `at:gt` interactions for both varieties
⇒ little `at:v:gt` interactions
- ▶ Little `at:v` interactions for both `gt = 13` and `gt = 25`

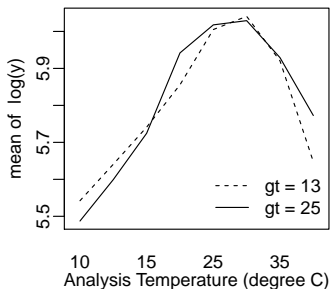


- ▶ Some evidence of $gt:v$ interactions since
 - ▶ the line $gt:v = 25:B73$ is always ABOVE $gt:v = 13:B73$
 - ▶ the line $gt:v = 25:O43$ is mostly BELOW $gt:v = 13:O43$
- ▶ Large at main effects as the lines are not horizontal
- ▶ Some v (variety) main effects as the two lines for $v = B73$ are always ABOVE the two lines for $v = O43$
- ▶ As noted earlier, the effect of growth temperature gt were different for the two varieties.

Two-Way Interaction Plots — Amylase data

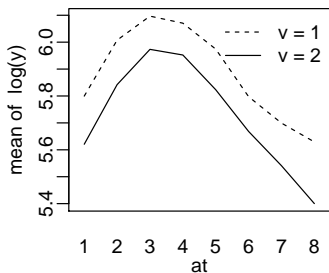
If one just check the two-way interaction plot between `at` and `gt`, the information of variety `v` would be ignored. The two curves below the curves for `gt = 13` and `gt = 25` averaged over the two varieties B73 and O43.

```
with(amy1,interaction.plot(at, gt, log(y),  
                           xlab="Analysis Temperature (degree C)"))
```

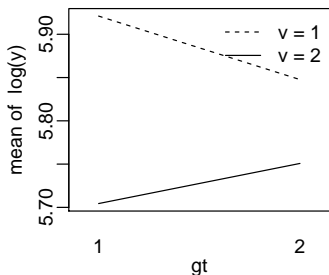


- ▶ Little evidence of `at:gt` interactions
- ▶ The `at` main effects appear significant as the lines are not horizontal
- ▶ There appear little `gt` main effects as there's little gap between the two lines

Example 8.10 Amylase data — Interaction Plots



- ▶ Does the main effect **at** appear significant?
- ▶ and the main effect **v**?
- ▶ **at:v** interaction?



- ▶ Does the main effect **gt** appear significant?
- ▶ and main effect **v**?
- ▶ **gt:v** interaction?

Parameter Estimates for General Factorial Models

The Box-Cox method suggests taking **log** of the amylase activity to make the variance closer to equal. The following are the sample group means $\bar{u}_{ijk\bullet}$, where $u_{ijkl} = \log y_{ijkl}$ is the log of amylase activity.

Variety		Growth Temperature			
		13°C ($k = 1$)		25°C ($k = 2$)	
		B73 ($j = 1$)	O43 ($j = 2$)	B73 ($j = 1$)	O43 ($j = 2$)
Analysis Temperature	10°C	$\bar{u}_{111\bullet} \approx 5.601$	$\bar{u}_{121\bullet} \approx 5.483$	$\bar{u}_{112\bullet} \approx 5.657$	$\bar{u}_{122\bullet} \approx 5.317$
	13°C	$\bar{u}_{211\bullet} \approx 5.692$	$\bar{u}_{221\bullet} \approx 5.589$	$\bar{u}_{212\bullet} \approx 5.706$	$\bar{u}_{222\bullet} \approx 5.492$
	15°C	$\bar{u}_{311\bullet} \approx 5.789$	$\bar{u}_{321\bullet} \approx 5.693$	$\bar{u}_{312\bullet} \approx 5.806$	$\bar{u}_{322\bullet} \approx 5.643$
	20°C	$\bar{u}_{411\bullet} \approx 5.941$	$\bar{u}_{421\bullet} \approx 5.771$	$\bar{u}_{412\bullet} \approx 6.008$	$\bar{u}_{422\bullet} \approx 5.876$
	25°C	$\bar{u}_{511\bullet} \approx 6.032$	$\bar{u}_{521\bullet} \approx 5.978$	$\bar{u}_{512\bullet} \approx 6.109$	$\bar{u}_{522\bullet} \approx 5.927$
	30°C	$\bar{u}_{611\bullet} \approx 6.041$	$\bar{u}_{621\bullet} \approx 6.041$	$\bar{u}_{612\bullet} \approx 6.153$	$\bar{u}_{622\bullet} \approx 5.906$
	35°C	$\bar{u}_{711\bullet} \approx 5.959$	$\bar{u}_{721\bullet} \approx 5.882$	$\bar{u}_{712\bullet} \approx 6.056$	$\bar{u}_{722\bullet} \approx 5.802$
	40°C	$\bar{u}_{811\bullet} \approx 5.726$	$\bar{u}_{821\bullet} \approx 5.569$	$\bar{u}_{812\bullet} \approx 5.873$	$\bar{u}_{822\bullet} \approx 5.672$

Variety	B73 ($j = 1$)	O43 ($j = 2$)	Mean	
Analysis Temperature	10	$\bar{u}_{11\bullet\bullet} \approx 5.629$	$\bar{u}_{12\bullet\bullet} \approx 5.400$	$\bar{u}_{1\bullet\bullet\bullet} \approx 5.515$
	13	$\bar{u}_{21\bullet\bullet} \approx 5.699$	$\bar{u}_{22\bullet\bullet} \approx 5.541$	$\bar{u}_{2\bullet\bullet\bullet} \approx 5.620$
	15	$\bar{u}_{31\bullet\bullet} \approx 5.797$	$\bar{u}_{32\bullet\bullet} \approx 5.668$	$\bar{u}_{3\bullet\bullet\bullet} \approx 5.733$
	20	$\bar{u}_{41\bullet\bullet} \approx 5.974$	$\bar{u}_{42\bullet\bullet} \approx 5.823$	$\bar{u}_{4\bullet\bullet\bullet} \approx 5.899$
	25	$\bar{u}_{51\bullet\bullet} \approx 6.071$	$\bar{u}_{52\bullet\bullet} \approx 5.952$	$\bar{u}_{5\bullet\bullet\bullet} \approx 6.011$
	30	$\bar{u}_{61\bullet\bullet} \approx 6.097$	$\bar{u}_{62\bullet\bullet} \approx 5.973$	$\bar{u}_{6\bullet\bullet\bullet} \approx 6.035$
	35	$\bar{u}_{71\bullet\bullet} \approx 6.007$	$\bar{u}_{72\bullet\bullet} \approx 5.842$	$\bar{u}_{7\bullet\bullet\bullet} \approx 5.925$
	40	$\bar{u}_{81\bullet\bullet} \approx 5.800$	$\bar{u}_{82\bullet\bullet} \approx 5.621$	$\bar{u}_{8\bullet\bullet\bullet} \approx 5.710$
Mean	$\bar{u}_{\bullet 1\bullet\bullet} \approx 5.884$	$\bar{u}_{\bullet 2\bullet\bullet} \approx 5.728$	$\bar{u}_{\bullet\bullet\bullet\bullet} \approx 5.806$	

GT	13°C ($k = 1$)	25°C ($k = 2$)	
Analysis Temperature	10	$\bar{u}_{1\bullet 1\bullet} \approx 5.542$	$\bar{u}_{1\bullet 2\bullet} \approx 5.487$
	13	$\bar{u}_{2\bullet 1\bullet} \approx 5.641$	$\bar{u}_{2\bullet 2\bullet} \approx 5.599$
	15	$\bar{u}_{3\bullet 1\bullet} \approx 5.741$	$\bar{u}_{3\bullet 2\bullet} \approx 5.725$
	20	$\bar{u}_{4\bullet 1\bullet} \approx 5.856$	$\bar{u}_{4\bullet 2\bullet} \approx 5.942$
	25	$\bar{u}_{5\bullet 1\bullet} \approx 6.005$	$\bar{u}_{5\bullet 2\bullet} \approx 6.018$
	30	$\bar{u}_{6\bullet 1\bullet} \approx 6.041$	$\bar{u}_{6\bullet 2\bullet} \approx 6.029$
	35	$\bar{u}_{7\bullet 1\bullet} \approx 5.920$	$\bar{u}_{7\bullet 2\bullet} \approx 5.929$
	40	$\bar{u}_{8\bullet 1\bullet} \approx 5.647$	$\bar{u}_{8\bullet 2\bullet} \approx 5.773$
Mean	$\bar{u}_{\bullet\bullet 1\bullet} \approx 5.799$	$\bar{u}_{\bullet\bullet 2\bullet} \approx 5.813$	

Variety	Growth Temperature	
	13°C ($k = 1$)	25°C ($k = 2$)
B73	$\bar{u}_{\bullet 11\bullet} \approx 5.848$	$\bar{u}_{\bullet 12\bullet} \approx 5.921$
O43	$\bar{u}_{\bullet 21\bullet} \approx 5.751$	$\bar{u}_{\bullet 22\bullet} \approx 5.704$

Parameter Estimates for General Factorial Models

For the full model with all 2-way and 3-way interactions:

$$u_{ijkl} = \log y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \beta\gamma_{jk} + \alpha\gamma_{ij} + \alpha\beta\gamma_{ijk} + \varepsilon_{ijkl}$$

The parameter estimates under the zero-sum constraints for a few parameters are

$$\hat{\mu} = \bar{u}_{\bullet\bullet\bullet\bullet} \approx 5.806$$

$$\hat{\alpha}_5 = \bar{u}_{5\bullet\bullet\bullet} - \bar{u}_{\bullet\bullet\bullet\bullet} \approx 6.011 - 5.806 = 0.205$$

$$\hat{\beta}_1 = \bar{u}_{\bullet 1 \bullet\bullet} - \bar{u}_{\bullet\bullet\bullet\bullet} \approx 5.884 - 5.806 = 0.078$$

$$\begin{aligned}\widehat{\beta\gamma}_{12} &= \bar{u}_{\bullet 12 \bullet} - \bar{u}_{\bullet 1 \bullet\bullet} - \bar{u}_{\bullet\bullet 2 \bullet} + \bar{u}_{\bullet\bullet\bullet\bullet} \\ &\approx 5.921 - 5.885 - 5.813 + 5.806 = 0.029\end{aligned}$$

$$\begin{aligned}\widehat{\alpha\beta\gamma}_{512} &= \bar{u}_{512\bullet} - \bar{u}_{51\bullet\bullet} - \bar{u}_{\bullet 12 \bullet} - \bar{u}_{5\bullet 2 \bullet} + \bar{u}_{5\bullet\bullet\bullet} + \bar{u}_{\bullet 1 \bullet\bullet} + \bar{u}_{\bullet\bullet 2 \bullet} - \bar{u}_{\bullet\bullet\bullet\bullet} \\ &\approx 6.109 - 6.071 - 5.921 - 6.018 + 6.011 + 5.884 + 5.813 - 5.806 \\ &= 0.001\end{aligned}$$

Other parameters can be estimated similarly.

How to find $\bar{u}_{ijk\bullet}$ in R?

```
library(mosaic)
mean(log(y) ~ at+v+gt, data=amyl)
```

Similarly,

- ▶ $\bar{u}_{ij\bullet\bullet}$: `mean(log(y) ~ at+v, data=amyl)`
- ▶ $\bar{u}_{i\bullet k\bullet}$: `mean(log(y) ~ at+gt, data=amyl)`
- ▶ $\bar{u}_{\bullet j k\bullet}$: `mean(log(y) ~ v+gt, data=amyl)`
- ▶ $\bar{u}_{i\bullet\bullet\bullet}$: `mean(log(y) ~ at, data=amyl)`
- ▶ $\bar{u}_{\bullet j\bullet\bullet}$: `mean(log(y) ~ v, data=amyl)`
- ▶ $\bar{u}_{\bullet\bullet k\bullet}$: `mean(log(y) ~ gt, data=amyl)`
- ▶ $\bar{u}_{\bullet\bullet\bullet\bullet}$: `mean(log(y) ~ 1, data=amyl)`

Model Formula in R

The R command for fitting the full 3-way model

$$\log y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \beta\gamma_{jk} + \alpha\gamma_{ij} + \alpha\beta\gamma_{ijk} + \varepsilon_{ijkl}$$

is

```
lm(log(y) ~ at+v+gt+at:v+at:gt+v:gt + at:v:gt, data=amyl)
```

A simpler syntax is

```
lm(log(y) ~ at*v*gt, data=amyl)
```

The term `at*v*gt` and `at:v:gt` both mean the 3-way interaction terms $\alpha\beta\gamma_{ijk}$, but

- ▶ `at*v*gt` will automatically include all relevant main effects and lower order interactions in the model.
- ▶ `at:v:gt` will not include the lower order terms

Hence, the R command below

```
lm(log(y) ~ -1 + at:v:gt, data=amyl)
```

will fit the means model $\log y_{ijkl} = \mu_{ijk} + \varepsilon_{ijkl}$.

Example 8.10 Amylase data — ANOVA Table

```
> logfit1 = lm(log(y) ~ at*v*gt, data=amyl)
> anova(logfit1)
Analysis of Variance Table

Response: log(y)
          Df Sum Sq Mean Sq  F value    Pr(>F)
at         7 3.01613  0.43088   78.8628 < 2.2e-16 ***
v          1 0.58957  0.58957  107.9085 2.305e-15 ***
gt         1 0.00438  0.00438    0.8016 0.3739757
at:v       7 0.02758  0.00394    0.7212 0.6543993
at:gt      7 0.08106  0.01158    2.1195 0.0539203 .
v:gt       1 0.08599  0.08599   15.7392 0.0001863 ***
at:v:gt    7 0.04764  0.00681    1.2457 0.2916176
Residuals 64 0.34967  0.00546
```

Only the analysis temperature (**at**), variety (**v**), and the variety by growth temperature interactions (**v:gt**) are highly significant.

Can I fit a model like $y_{ijkl} = \mu + \alpha_i + \beta_j + \beta\gamma_{jk} + \varepsilon_{ijkl}$?

```
logfit2 = lm(log(y) ~ at + v + v:gt, data=amyl)
```

8.11 Hierarchy

A model is *hierarchical* if any term in the model implies the presence of all the composite lower-order terms.

- ▶ $y_{ijkl} = \mu + \alpha_i + \beta_j + \beta\gamma_{jk} + \varepsilon_{ijkl}$ is not hierarchical because including the term $\beta\gamma_{jk}$ must include *both* β_j and γ_k as well.
- ▶ $y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$ is hierarchical.
- ▶ A hierarchical model with a term $\alpha\beta\gamma_{ijk}$ must also include:
 - ▶ the relevant main effects: $\alpha_i + \beta_j + \gamma_k$
 - ▶ and the included two-way effects: $\alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk}$.

8.11 Hierarchy

Unless having a specific reason, we should stick to hierarchical models.

- ▶ This is because a k -way interaction is defined upon $(k - 1)$ -way interactions. It is strange to consider a ABC interaction while claiming A and B have no 2-way interaction.
- ▶ E.g., when we say there are no AB interactions, we also imply that there are no higher order interactions that involve AB interactions, like ABD interactions, or ABCD interactions.

Why Maintaining Hierarchy?

Let's consider a model for a 2×2 factorial design.

$$\begin{aligned}y_{ijk} &= \mu_{ij} + \varepsilon_{ijk} \\ &= \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}\end{aligned}$$

If $\alpha_1 = \alpha_2 = 0$, but $\alpha\beta_{11} \neq 0$, can Factor A have any effect on the response? Consider the example below.

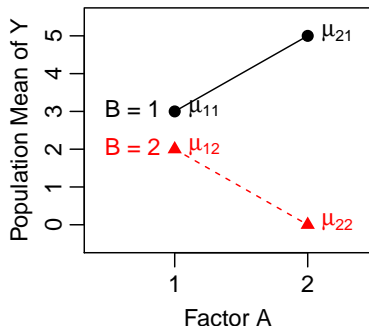
	B = 1	B = 2	Mean
A = 1	$\mu_{11} = 3$	$\mu_{12} = 2$	$\mu_{1\bullet} = 2.5$
A = 2	$\mu_{21} = 5$	$\mu_{22} = 0$	$\mu_{2\bullet} = 2.5$
Mean	$\mu_{\bullet 1} = 4$	$\mu_{\bullet 2} = 1$	$\mu_{\bullet\bullet} = 2.5$

Under the zero-sum constraint,

$$\alpha_i = \bar{\mu}_{i\bullet} - \bar{\mu}_{\bullet\bullet} = 2.5 - 2.5 = 0$$

for $i = 1, 2$.

Clearly $\alpha\beta_{ij} \neq 0$ as the lines are not parallel.



Back to the Amylase Data

Here is a hierarchical model that leaves out all insignificant terms.

```
> logfit2 = lm(log(y) ~ at + v + gt + at:gt + v:gt, data=amyl)
> anova(logfit2)
Analysis of Variance Table
```

Response: log(y)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
at	7	3.01613	0.43088	79.0981	< 2.2e-16	***
v	1	0.58957	0.58957	108.2305	< 2.2e-16	***
gt	1	0.00438	0.00438	0.8040	0.3726670	
at:gt	7	0.08106	0.01158	2.1258	0.0503809	.
v:gt	1	0.08599	0.08599	15.7861	0.0001571	***
Residuals	78	0.42489	0.00545			

The insignificant `gt` main effect cannot be left out since the two-way interaction `v:gt` is significant.

The SS's and d.f.'s of the left-out terms are **pooled into the SSE and the df of error** while the SS's and d.f.'s of the remaining stay unchanged.

Goodness-of-Fit Test

Comparing the reduced model with the full 3-way model (equivalent to the means model), the large P -value indicates the adequacy of the reduced model.

```
> anova(logfit2,logfit1)
Analysis of Variance Table
```

```
Model 1: log(y) ~ at + v + gt + at:gt + v:gt
```

```
Model 2: log(y) ~ at * gt * v
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	78	0.42489				
2	64	0.34967	14	0.075223	0.9834	0.4801

More On Model Formula in R (1)

Instead of writing terms explicitly in the model formula

```
> logfit2 = lm(log(y) ~ at + v + gt + at:gt + v:gt, data=amyl)
```

Here is another simpler expression for the same model. R will automatically create the smallest hierarchical model that include both `at:gt` and `gt:v` interactions.

```
> logfit2a = lm(log(y) ~ at*gt + v*gt, data=amyl)
```

```
> anova(logfit2a)
```

Analysis of Variance Table

Response: log(y)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
at	7	3.01613	0.43088	79.0981	< 2.2e-16	***
gt	1	0.00438	0.00438	0.8040	0.3726670	
v	1	0.58957	0.58957	108.2305	< 2.2e-16	***
at:gt	7	0.08106	0.01158	2.1258	0.0503809	.
gt:v	1	0.08599	0.08599	15.7861	0.0001571	***
Residuals	78	0.42489	0.00545			

More On Model Formula in R (2)

If one wants a model with all two-way interactions but no 3-way interaction, one can explicitly write down every term

```
logfit3a = lm(log(y) ~ at + v + gt + at:v + v:gt + at:gt, data=amyl)
```

Here is another way to obtain everything up to the 2-way interactions

```
logfit3b = lm(log(y) ~ (at + v + gt)^2, data=amyl)
```

Here is another way to “leave out” the 3-way interaction

```
logfit3c = lm(log(y) ~ at*v*gt - at:v:gt, data=amyl)
```

You can verify the 3 model formulas are identical in R.

```
anova(logfit3a)  
anova(logfit3b)  
anova(logfit3c)
```

Example 8.10 Amylase data — Model Checking

Always check model assumptions FIRST!

Recall that we took log of the original response. If we didn't...

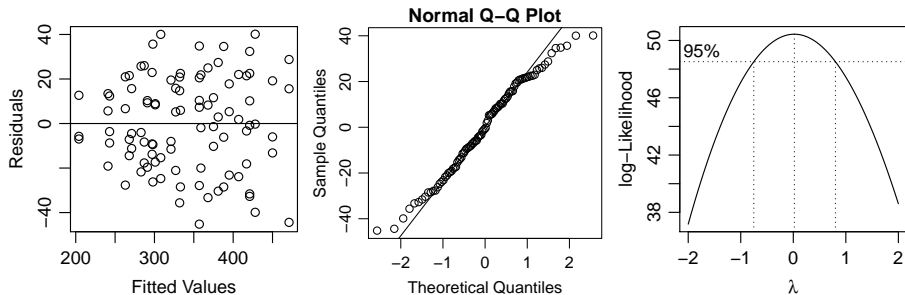
```
> fit1 = lm(y ~ at*v*gt, data=amyl); anova(fit1)
Analysis of Variance Table
```

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
at	7	327811	46830	72.9366	< 2.2e-16	***
v	1	63809	63809	99.3801	1.192e-14	***
gt	1	1155	1155	1.7988	0.1845968	
at:v	7	1174	168	0.2611	0.9665902	
at:gt	7	7158	1023	1.5925	0.1537663	
v:gt	1	10648	10648	16.5839	0.0001305	***
at:v:gt	7	6257	894	1.3922	0.2240596	
Residuals	64	41092	642			

Don't drop insignificant terms before checking model assumptions. If any assumption is violated, the ANOVA table is not reliable.

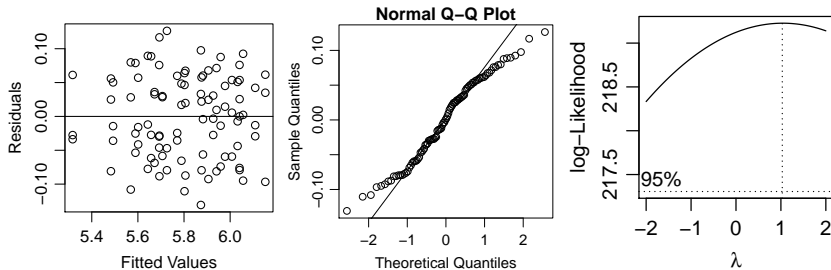
Example 8.10 Amylase data — Model Checking (2)



- ▶ The residual plot exhibits non-constant variability — the size of residuals increases with fitted values.
- ▶ The QQ plot looks symmetric but a bit short-tailed.
- ▶ The Box-Cox method suggests a log-transformation.

Check the model again after log-transformation but before dropping terms, i.e. check the model

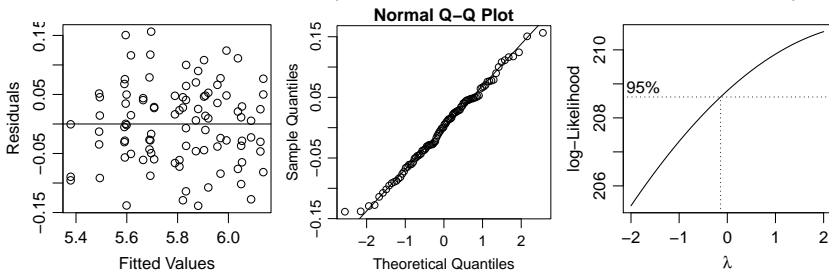
```
lm(log(y)~at*v*gt, data=amyl).
```



- ▶ The non-constant variance problem is alleviated
- ▶ The QQ plot still looks short-tailed. Residuals often appear short-tailed when fitting a “large” model, which tends to overfit the data, making residuals too close to zero.
- ▶ Box-Cox suggests no transformation ($\lambda = 1$ is in the 95% C.I.), i.e., the log-transformed response is fine.

So the ANOVA table based on the log-transformed data seems trustworthy and we can make inference or drop terms based on it.

Let's check the model again after dropping insignificant terms in the log transformed model (i.e. `lm(log(y)~at*gt+gt*v, data=amy1)`).



- ▶ The residual plot looks fine
- ▶ After removing insignificant terms, residuals no longer appear short-tailed
- ▶ Box-Cox 95% C.I. for λ contains 1. Okay.

Remark about Box-Cox: R by default will only plot λ for the range $(-2, 2)$. The range of λ can be changed, like -2 to 10 in steps 0.25 in the command below.

```
> library(MASS)
> boxcox(logfit2, lambda=seq(-2, 10, 0.25))
```

8.9 Single Replicates

Some factorial experiments have only ONE replicate per treatment.

- ▶ No degree of freedom for error, cannot estimate σ^2
- ▶ All sum of squares (SS) can be computed as usual except that $SSE = 0$.
- ▶ ANOVA F -tests for main effects and interactions of all orders cannot be done!

Remedy — Pooling higher order interactions into error

Problem 8.6 (p. 222, Oehlert's)

Response: dry matter yield in hundreds of pounds per acre over a 54-week study period

Factors:

- ▶ height of cut (1, 3, or 6 inches)
- ▶ cutting interval (1, 3, 6, or 9 weeks)
- ▶ amount of nitrogen fertilizer (0, 8, 16, or 32 hundred pounds of ammonium sulfate per acre per year).

		Cutting Interval			
		1 wks.	3 wks.	6 wks.	9 wks.
Ht 1	F 0	74.1	65.4	96.7	147.1
	F 8	87.4	117.7	190.2	188.6
	F 16	96.5	122.2	197.9	232.0
	F 32	107.6	140.5	241.3	192.0
Ht 3	F 0	61.7	83.7	88.8	155.6
	F 8	112.5	129.4	145.0	208.1
	F 16	102.3	137.8	173.6	203.2
	F 32	115.3	154.3	211.2	245.2
Ht 6	F 0	49.9	72.7	113.9	143.4
	F 8	92.9	126.4	175.5	207.5
	F 16	100.8	153.5	184.5	194.2
	F 32	115.8	160.0	224.8	197.5

Let's first load the data and fit a full 3-way model.

```
pr8.6 = read.table(
  "http://www.stat.uchicago.edu/~yibi/s222/pr8_6.txt", h=T)
pr8.6$HT = as.factor(pr8.6$ht)
pr8.6$FERT = as.factor(pr8.6$fert)
pr8.6$INT = as.factor(pr8.6$int)
lm1 = lm(y ~ HT*FERT*INT, data=pr8.6); anova(lm1)
```

The ANOVA table obtained is

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
HT	2	29	14.6		
FERT	3	42072	14023.9		
INT	3	73887	24629.0		
HT:FERT	6	406	67.7		
HT:INT	6	3005	500.9		
FERT:INT	9	5352	594.6		
HT:FERT:INT	18	3155	175.3		
Residuals	0	0			

Warning message:

In anova.lm(lm1) :

ANOVA F-tests on an essentially perfect fit are unreliable

We can pool the 3-way interaction terms as errors to get a *conservative* estimate of σ^2 .

```
> lm2 = lm(y ~ (HT+FERT+INT)^2, data=pr8.6)
```

```
> anova(lm2)
```

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
HT	2	29	14.6	0.0830	0.92068	
FERT	3	42072	14023.9	80.0153	1.334e-10	***
INT	3	73887	24629.0	140.5241	1.120e-12	***
HT:FERT	6	406	67.7	0.3860	0.87835	
HT:INT	6	3005	500.9	2.8578	0.03903	*
FERT:INT	9	5352	594.6	3.3927	0.01313	*
Residuals	18	3155	175.3			

Pooling SS of higher-order interactions into error can be justified since the expected value of the MS for a term (main effect or interaction) can be shown to be $\geq \sigma^2$.

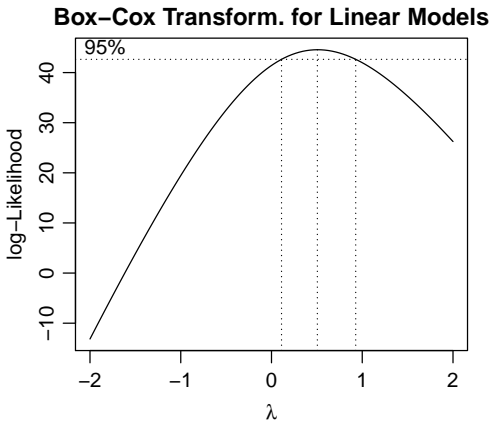
If we divide the MS for other terms by an “MSE” that tend to overestimates σ^2 , the F -value tend to be too small, less significant.

Before examining the treatment effects, first check if the model assumptions are met.

```
> library(MASS)
```

```
> boxcox(lm2)
```

The Box-Cox method below suggest a square-root transformation of the response since 0.5 is in the 95% confidence interval for λ .



Should check model assumption again after fitting the new model below.

```
> lm2s= lm(sqrt(y) ~ (HT+FERT+INT)^2, data=pr8.6)
> anova(lm2s)
```

Analysis of Variance Table

Response: sqrt(y)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
HT	2	0.103	0.052	0.1763	0.83979
FERT	3	82.222	27.407	93.8199	3.510e-11 ***
INT	3	132.738	44.246	151.4617	5.865e-13 ***
HT:FERT	6	0.537	0.089	0.3062	0.92553
HT:INT	6	4.873	0.812	2.7800	0.04303 *
FERT:INT	9	6.868	0.763	2.6123	0.03962 *
Residuals	18	5.258	0.292		

One possible model is keeping the two significant interactions **HT:INT** and **FERT:INT**. In this case, one needs to keep the insignificant **HT** main effects to maintain the hierarchy.

```
lm3s = lm(sqrt(y) ~ HT + INT + FERT + HT:INT + FERT:INT, data=pr8.6)
```

Another possible model is dropping all two-way interactions because the evidence for `HT:INT` and `FERT:INT` interactions is just at the borderline. In this case, we can drop the insignificant `HT` main effects and only keep the `FERT` and `INT` main effects

```
> lm4s = lm(sqrt(y) ~ FERT + INT, data=pr8.6)
```

```
> anova(lm2s, lm4s)
```

Analysis of Variance Table

```
Model 1: sqrt(y) ~ FERT + INT
```

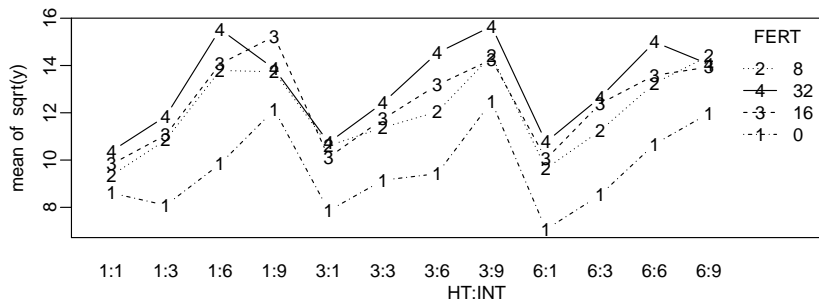
```
Model 2: sqrt(y) ~ (HT + FERT + INT)^2
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	41	17.6385				
2	18	5.2583	23	12.38	1.8426	0.09473 .

This simple model fits the data not much worse than the model with all 2-way interactions with a P -value of 0.095, which is preferred for its simplicity.

3-Way Interaction Plot

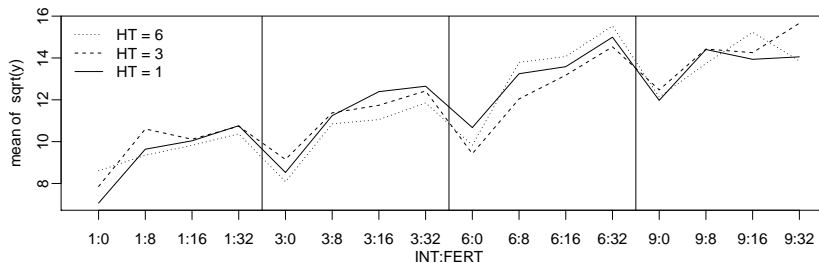
```
with(pr8.6, interaction.plot(HT:INT,FERT,sqrt(y),type="b"))
```



- ▶ Little **INT:FERT** interactions for the 3 HT level
⇒ little evidence of **INT:FERT:HT** interactions
- ▶ Large gap between lines ⇒ significant **FERT** main effect
- ▶ Lines are not horizontal ⇒ significant **INT** main effect

3-Way Interaction Plot

```
with(pr8.6, interaction.plot(INT:FERT,HT,sqrt(y),legend=F))  
legend("topleft",lty=c(3,2,1),paste("HT =", c(6,3,1)),bty="n")  
for(i in 1:4){abline(v=4*i+0.5)}
```



- ▶ The 3 lines entangled \Rightarrow little HT main effect
- ▶ Lines are closed to parallel \Rightarrow Little HT:FERT interactions