General Factorial Design

Yibi Huang

- General Factorial Design

Three Way Interactions

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

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

E.g., in a 3-way design, based on the means model $y_{ijkl} = \mu_{ijk} + \varepsilon_{ijkl}$, 3-way interaction of 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

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

$$= \underbrace{(\mu_{i_1 j_1 k_1} - \mu_{i_2 j_1 k_1} - \mu_{i_1 j_2 k_1} + \mu_{i_2 j_2 k_1})}_{\text{AB interaction at level } k_1 \text{ of C}} - \underbrace{(\mu_{i_1 j_1 k_2} - \mu_{i_2 j_1 k_2} - \mu_{i_1 j_2 k_2} + \mu_{i_2 j_2 k_2})}_{\text{AB interaction at level } k_2 \text{ of C}}$$

$$= \underbrace{(\mu_{i_1 j_1 k_1} - \mu_{i_1 j_2 k_1} - \mu_{i_1 j_1 k_2} + \mu_{i_1 j_2 k_2})}_{\text{BC interaction at level } i_1 \text{ of A}} - \underbrace{(\mu_{i_2 j_1 k_1} - \mu_{i_2 j_2 k_1} - \mu_{i_2 j_1 k_2} + \mu_{i_2 j_2 k_2})}_{\text{BC interaction at level } i_2 \text{ of A}}$$

$$= \underbrace{(\mu_{i_1 j_1 k_1} - \mu_{i_2 j_1 k_1} - \mu_{i_1 j_1 k_2} + \mu_{i_2 j_1 k_2})}_{\text{AC interaction at level } j_1 \text{ of B}} - \underbrace{(\mu_{i_1 j_2 k_1} - \mu_{i_2 j_2 k_1} - \mu_{i_1 j_2 k_2} + \mu_{i_2 j_2 k_2})}_{\text{AC interaction at level } j_2 \text{ of B}}$$

Higher Order Interactions

- We say 4 factors have **4-way interactions** means the 3-way interaction of any 3 of the 4 factors changes with the levels of a 4th factor.

- 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 kth factor.

Hierarchy

- Since k-way interactions are defined on $(k-1)$-way interactions, we cannot skip orders.

- 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.

General Factorial Designs

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

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

effects model: $y_{ijklm} = \underbrace{\mu}_{\text{grand mean}} + \underbrace{\alpha_i + \beta_j + \gamma_k + \delta_l}_{\text{main effects}}$

$$+ \underbrace{\alpha\beta_{ij} + \alpha\gamma_{ik} + \alpha\delta_{il} + \beta\gamma_{jk} + \beta\delta_{jl} + \gamma\delta_{kl}}_{\text{2-way interactions}}$$

$$+ \underbrace{\alpha\beta\gamma_{ijk} + \alpha\beta\delta_{ijl} + \alpha\gamma\delta_{ikl} + \beta\gamma\delta_{jkl}}_{\text{3-way interactions}}$$

$$+ \quad \underbrace{\alpha\beta\gamma\delta_{ijkl}}_{\text{4-way interaction}} \quad + \underbrace{\varepsilon_{ijklm}}_{\text{error}}$$

General Factorial Designs

$$\begin{aligned}
y_{ijklm} = \mu &+ \alpha_i + \beta_j + \gamma_k + \delta_l \\
&+ \alpha\beta_{ij} + \alpha\gamma_{ik} + \alpha\delta_{il} + \beta\gamma_{jk} + \beta\delta_{jl} + \gamma\delta_{kl} \\
&+ \alpha\beta\gamma_{ijk} + \alpha\beta\delta_{ijl} + \alpha\gamma\delta_{ikl} + \beta\gamma\delta_{jkl} \\
&+ \alpha\beta\gamma\delta_{ijkl} + \varepsilon_{ijklm}
\end{aligned}$$

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

- $\sum_i \alpha_i = \sum_j \beta_j = \sum_k \gamma_k = \sum_l \delta_l = 0$

- $\sum_i \alpha\gamma_{ik} = \sum_k \alpha\gamma_{ik} = 0$, for all i, k,
  so do other 2-way interactions

- $\sum_i \alpha\gamma\delta_{ikl} = \sum_k \alpha\gamma\delta_{ikl} = \sum_l \alpha\gamma\delta_{ikl} = 0$, for all i, k, l,
  so do other 3-way interactions

- $\sum_i \alpha\beta\gamma\delta_{ijkl} = \sum_j \alpha\beta\gamma\delta_{ijkl} = \sum_k \alpha\beta\gamma\delta_{ijkl} = \sum_l \alpha\beta\gamma\delta_{ijkl} = 0$,
  for all i, j, k, l.

Parameter Estimates

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

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

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

$$
SS_C = \sum_{ijklm}(\widehat{\gamma}_k)^2 = abdn \sum_k (\widehat{\gamma}_k)^2
$$

$$
SS_{BC} = \sum_{ijklm}(\widehat{\beta\gamma}_{jk})^2 = abn \sum_{jk} (\widehat{\beta\gamma}_{jk})^2
$$

$$
SS_{ACD} = \sum_{ijklm}(\widehat{\alpha\gamma\delta}_{ikl})^2 = bn \sum_{ikl} (\widehat{\alpha\gamma\delta}_{ikl})^2
$$

$$
SS_{ABCD} = \sum_{ijklm}(\widehat{\alpha\beta\gamma\delta}_{ijkl})^2 = n \sum_{ijkl} (\widehat{\alpha\beta\gamma\delta}_{ijkl})^2
$$

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)
- Data file: `amylaze.txt` is posted on Canvas

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).

|		Analysis Temperature							
GT	Var.	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

Example 8.10 Amylase data — Interaction Plots
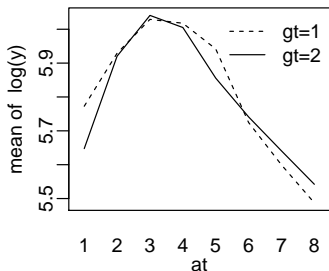
- ▶ Does the main effect `at` appear significant?
- ▶ How about the main effect `gt`?
- ▶ How about `at:gt` interaction?
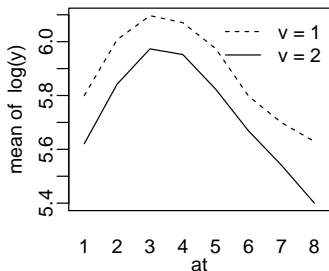
- ▶ Does the main effect `at` appear significant?
- ▶ and the main effect `v`?
- ▶ `at:v` interaction?

Example 8.10 Amylase data — Interaction Plots (2)
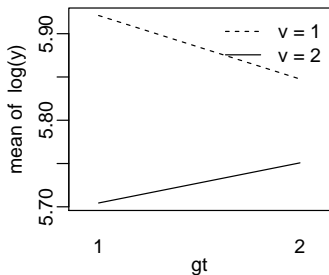
- Does the main effect `gt` appear significant?
- and main effect `v`?
- `gt:v` interaction?

General Factorial In R

```r
amyl = read.table("amylaze.txt", h=T)
amyl$at = as.factor(amyl$atemp)
amyl$gt = as.factor(amyl$gtemp)
amyl$v  = as.factor(amyl$variety)
```

To make the variance constant, the response is log-transformed (see p.215-216 in the textbook).

We fit a full model with all 2-way and 3-way interactions.

```r
logfit1 = lm(log(y) ~ at+gt+v+at:gt+at:v+gt:v + at:gt:v, data=amyl)
```

A simpler syntax is

```r
logfit1 = lm(log(y) ~ at*gt*v, data=amyl)
```

The syntax `at*gt*v` will automatically include all relevant main effects and lower order interactions in the model.

Example 8.10 Amylase data — ANOVA Table

```
> logfit1 = lm(log(y) ~ at*gt*v, 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 ***
gt         1 0.00438 0.00438   0.8016 0.3739757
v          1 0.58957 0.58957 107.9085 2.305e-15 ***
at:gt      7 0.08106 0.01158   2.1195 0.0539203 .
at:v       7 0.02758 0.00394   0.7212 0.6543993
gt:v       1 0.08599 0.08599  15.7392 0.0001863 ***
at:gt:v    7 0.04764 0.00681   1.2457 0.2916176
Residuals 64 0.34967 0.00546
```

Only analysis temperature (at), variety (v), and the growth
temperature by variety interaction (gt:v) are highly significant.
Can I fit a model like $y_{ijkl} = \mu + \alpha_i + \gamma_k + \beta\gamma_{jk} + \varepsilon_{ijkl}$?

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

- $y_{ijkl} = \mu + \alpha_i + \gamma_k + \beta\gamma_{jk} + \varepsilon_{ijkl}$ is not hierarchical because including the term $\beta\gamma_{jk}$ must includes *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 included main effects: $\alpha_i + \beta_j + \gamma_k$
    - and the included two-way effects: $\alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk}$.

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

- This is because a *k*-way interaction in defined upon its composite lower-order terms. It is strange to consider a ABC interaction while claiming A and B have no 2-way interaction.

Back to the Amylase Data

Here is a model that is still hierarchical, but leaves off non-significant model terms.

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

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

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

Comparing the reduced model with the full 3-way model, the large
P-value indicates the adequacy of the reduced model.

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

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

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

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

Here is another way to "leave out" the 3-way interaction

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

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

```
> fit1 = lm(y ~ at*gt*v, 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 ***
gt         1   1155    1155  1.7988 0.1845968
v          1  63809   63809 99.3801 1.192e-14 ***
at:gt      7   7158    1023  1.5925 0.1537663
at:v       7   1174     168  0.2611 0.9665902
gt:v       1  10648   10648 16.5839 0.0001305 ***
at:gt:v    7   6257     894  1.3922 0.2240596
Residuals 64  41092     642
```
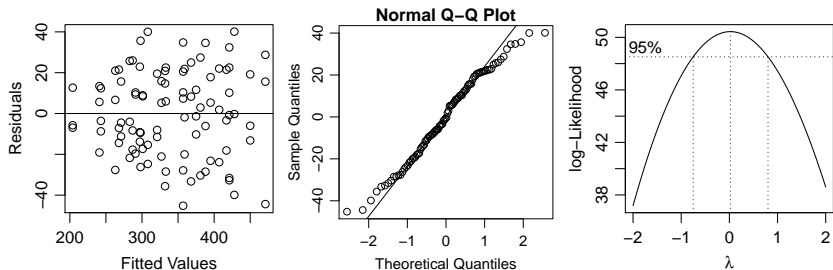
Don't drop non-significant terms before checking model assumptions. If any assumption is violated, the ANOVA table is not reliable.
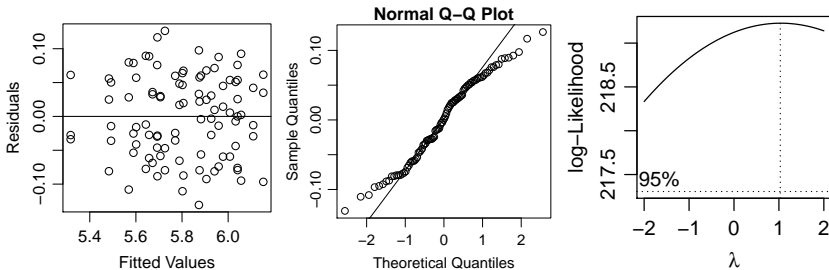
Example 8.10 Amylase data — Model Checking (2)

- The residual plot indicates non-constant variance — 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 <u>before</u> 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=amyl)`).
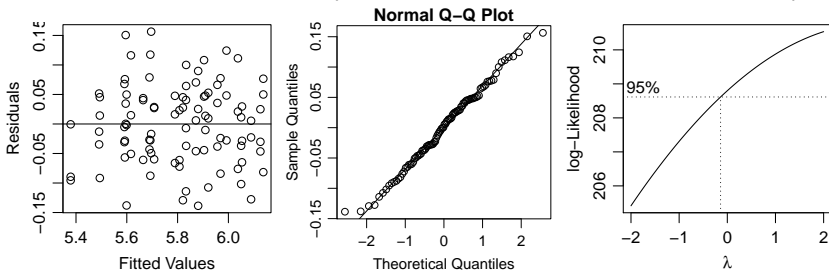
- ▶ The residual plot looks fine
- ▶ After removing significant terms, residuals no longer appear short-tailed
- ▶ Box-Cox shows the 95% C.I. contains $\lambda = 1$. Okay.

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))
```

Single Replicate

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 — Pool higher order interactions into error

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

The data file `pr8_6.txt` is posted on Canvas Let's first load the data and fit a full 3-way model.

```
pr8.6 = read.table("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 pool the 3-way interaction terms as errors to get a *conservative* estimate of the MSE.

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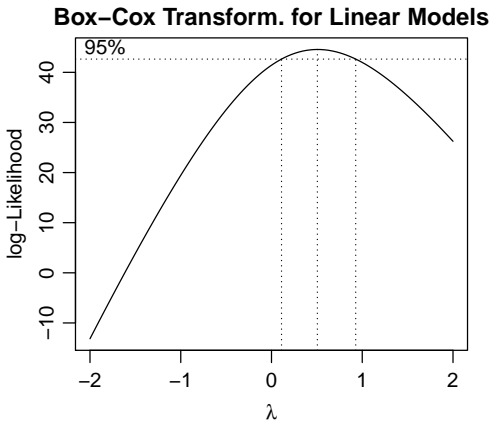
Before examining the treatment effects, first check if the model assumptions are met.

```
> 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).

Box–Cox Transform. for Linear Models

```
> 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
---
```