Factorial and Unbalanced Analysis of Variance

Nathaniel E. Helwig

Assistant Professor of Psychology and Statistics
University of Minnesota (Twin Cities)

Updated 04-Jan-2017

Copyright

Outline of Notes

1) Balanced Two-Way ANOVA:
- Model Form & Assumptions
- Least-Squares Estimation
- Basic Inference
- Hypertension Example (pt 1)
- Multiple Comparisons
- Hypertension Example (pt 2)

2) Balanced Three-Way ANOVA:
- Model Form & Estimation
- Hypertension Example (pt 3)

3) Unbalanced ANOVA Models:
- Overview of problem
- Types of sums-of-squares
- Hypertension Example (pt 4)

Balanced Two-Way ANOVA

Two-Way ANOVA Model (cell means form)

The Two-Way Analysis of Variance (ANOVA) model has the form

$$y_{ijk} = \mu_{jk} + e_{ijk}$$

for $i \in \{1, \ldots, n_{jk}\}$, $j \in \{1, \ldots, a\}$, and $k \in \{1, \ldots, b\}$ where

- $y_{ijk} \in \mathbb{R}$ is real-valued response for i-th subject in factor cell (j, k)
- $\mu_{jk} \in \mathbb{R}$ is real-valued population mean for factor cell (j, k)
- $e_{ijk} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ is a Gaussian error term
- n_{jk} is number of subjects in cell (j, k) and $n = \sum_{j=1}^{a} \sum_{k=1}^{b} n_{jk}$ (note: $n_{jk} = n_* \forall j, k$ in balanced two-way ANOVA)
- a and b are number of levels for first and second factors

Implies that $y_{ijk} \overset{\text{ind}}{\sim} \mathrm{N}(\mu_{jk}, \sigma^2)$.

Two-Way ANOVA Model (effect coding)

Using effect coding, the mean for factor cell (j, k) has the form

$$\mu_{jk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk}$$

for $j \in \{1, \ldots, a\}$ and $k \in \{1, \ldots, b\}$ where

- μ is overall population mean
- α_j is main effect of first factor such that $\sum_{j=1}^{a} \alpha_j = 0$
- β_k is main effect of second factor such that $\sum_{k=1}^{b} \beta_k = 0$
- $(\alpha\beta)_{jk}$ is interaction effect such that $\sum_{j=1}^{a} (\alpha\beta)_{jk} = 0 \; \forall k$ and $\sum_{k=1}^{b} (\alpha\beta)_{jk} = 0 \; \forall j$

Set $(\alpha\beta)_{jk} = 0 \; \forall j, k$ to fit an additive model.

Two-Way ANOVA Model (matrix form)

In matrix form, the two-way ANOVA model is $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$ where

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} \cdots x_{1(a-1)} & z_{11} \cdots z_{1(b-1)} & x_{11}z_{11} \cdots x_{1(a-1)}z_{11} & \cdots & x_{11}z_{1(b-1)} \cdots x_{1(a-1)}z_{1(b-1)} \\ 1 & x_{21} \cdots x_{2(a-1)} & z_{21} \cdots z_{2(b-1)} & x_{21}z_{21} \cdots x_{2(a-1)}z_{21} & \cdots & x_{21}z_{2(b-1)} \cdots x_{2(a-1)}z_{2(b-1)} \\ 1 & x_{31} \cdots x_{3(a-1)} & z_{31} \cdots z_{3(b-1)} & x_{31}z_{31} \cdots x_{3(a-1)}z_{31} & \cdots & x_{31}z_{3(b-1)} \cdots x_{3(a-1)}z_{3(b-1)} \\ \vdots & \vdots \ \ddots \ \vdots & \vdots \ \ddots \ \vdots & \vdots \ \ddots \ \vdots & \ddots & \vdots \ \ddots \ \vdots \\ 1 & x_{n1} \cdots x_{n(a-1)} & z_{n1} \cdots z_{n(b-1)} & x_{n1}z_{n1} \cdots x_{n(a-1)}z_{n1} & \cdots & x_{n1}z_{n(b-1)} \cdots x_{n(a-1)}z_{n(b-1)} \end{pmatrix}$$

$$\mathbf{b} = \begin{pmatrix} \mu & \alpha_1 \cdots \alpha_{a-1} & \beta_1 \cdots \beta_{b-1} & (\alpha\beta)_{11} \cdots (\alpha\beta)_{(a-1)1} & \cdots & (\alpha\beta)_{1(b-1)} \cdots (\alpha\beta)_{(a-1)(b-1)} \end{pmatrix}'$$

where $\mathbf{X}$ has $1 + (a-1) + (b-1) + (a-1)(b-1) = ab$ columns

- $x_{ij} = \begin{cases} 1 & \text{if } i\text{-th observation is in } j\text{-th level of first factor} \\ -1 & \text{if } i\text{-th observation is in } a\text{-th level of first factor} \\ 0 & \text{otherwise} \end{cases}$
- $z_{ik} = \begin{cases} 1 & \text{if } i\text{-th observation is in } k\text{-th level of second factor} \\ -1 & \text{if } i\text{-th observation is in } b\text{-th level of second factor} \\ 0 & \text{otherwise} \end{cases}$
- $i \in \{1, \ldots, n\}$ and additional subscripts on y and e are dropped

Implies that $\mathbf{y} \sim \mathrm{N}(\mathbf{X}\mathbf{b}, \sigma^2\mathbf{I}_n)$.

Two-Way ANOVA Model (assumptions)

The fundamental assumptions of the two-way ANOVA model are:

1. x_{ij}, z_{ik} and y_i are observed random variables (known constants)
2. $e_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ is an unobserved random variable
3. μ_{jk} are unknown constants
4. $(y_i | x_{ij}, z_{ik}) \overset{\text{ind}}{\sim} \mathrm{N}(\mu_{jk}, \sigma^2)$
   note: homogeneity of variance

Interpretation of μ_{jk} depends on model form

- Additive: $\mu_{jk} = \mu + \alpha_j + \beta_k$
- Interaction: $\mu_{jk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk}$

Ordinary Least-Squares

We want to find the effect estimates (i.e., $\hat{\mu}$, $\hat{\alpha}_j$, $\hat{\beta}_k$, and $(\hat{\alpha\beta})_{jk}$ terms) that minimize the ordinary least squares criterion

$$SSE = \sum_{k=1}^{b} \sum_{j=1}^{a} \sum_{i=1}^{n_{jk}} (y_{ijk} - \mu - \alpha_j - \beta_k - (\alpha\beta)_{jk})^2$$

If $n_{jk} = n_* \forall j, k$ the least-squares estimates have the form

$$\hat{\mu} = \frac{1}{abn_*} \sum_{k=1}^{b} \sum_{j=1}^{a} \sum_{i=1}^{n_*} y_{ijk} = \bar{y}_{\cdot\cdot}$$

$$\hat{\alpha}_j = \left(\frac{1}{bn_*} \sum_{k=1}^{b} \sum_{i=1}^{n_*} y_{ijk} \right) - \hat{\mu} = \bar{y}_{j\cdot} - \bar{y}_{\cdot\cdot}$$

$$\hat{\beta}_k = \left(\frac{1}{an_*} \sum_{j=1}^{a} \sum_{i=1}^{n_*} y_{ijk} \right) - \hat{\mu} = \bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot}$$

$$(\hat{\alpha\beta})_{jk} = \left(\frac{1}{n_*} \sum_{i=1}^{n_*} y_{ijk} \right) - \hat{\mu} - \hat{\alpha}_j - \hat{\beta}_k = \bar{y}_{jk} - \bar{y}_{j\cdot} - \bar{y}_{\cdot k} + \bar{y}_{\cdot\cdot}$$

which implies that $\hat{y}_{ijk} = \bar{y}_{jk}$ for all (i, j, k). Proof

Fitted Values and Residuals

Form of fitted values depends on fit model:

- Additive: $\hat{\mu}_{jk} = \bar{y}_{j\cdot} + \bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot}$
- Interaction: $\hat{\mu}_{jk} = \bar{y}_{jk}$

Residuals have the form

$$\hat{e}_{ijk} = y_{ijk} - \hat{\mu}_{jk}$$

where form of $\hat{\mu}_{jk}$ depends on fit model (additive versus interaction).

ANOVA Sums-of-Squares

In balanced two-way ANOVA model with interaction:

- $SST = \sum_{k=1}^{b} \sum_{j=1}^{a} \sum_{i=1}^{n_*} (y_{ijk} - \bar{y}_{..})^2 \qquad df = abn_* - 1$
- $SSR = n_* \sum_{k=1}^{b} \sum_{j=1}^{a} (\bar{y}_{jk} - \bar{y}_{..})^2 \qquad df = ab - 1$
- $SSE = \sum_{k=1}^{b} \sum_{j=1}^{a} \sum_{i=1}^{n_*} (y_{ijk} - \bar{y}_{jk})^2 \qquad df = abn_* - ab$

In balanced two-way ANOVA model with no interaction:

- $SST = \sum_{k=1}^{b} \sum_{j=1}^{a} \sum_{i=1}^{n_*} (y_{ijk} - \bar{y}_{..})^2 \qquad df = abn_* - 1$
- $SSR = n_* \sum_{k=1}^{b} \sum_{j=1}^{a} ([\bar{y}_{j\cdot} + \bar{y}_{\cdot k} - \bar{y}_{..}] - \bar{y}_{..})^2 \qquad df = a + b - 2$
- $SSE = \sum_{k=1}^{b} \sum_{j=1}^{a} \sum_{i=1}^{n_*} (y_{ijk} - [\bar{y}_{j\cdot} + \bar{y}_{\cdot k} - \bar{y}_{..}])^2$
  $df = abn_* - (a + b - 1)$

Partitioning the Variance

From MLR notes we know that $SST = SSR + SSE$.

If $n_{jk} = n_* \forall j, k$ can partition $SSR = SSA + SSB + SSAB$ where

- $SSA = bn_* \sum_{j=1}^{a} \hat{\alpha}_j^2$ $\quad df = a - 1$
- $SSB = an_* \sum_{k=1}^{b} \hat{\beta}_k^2$ $\quad df = b - 1$
- $SSAB = n_* \sum_{k=1}^{b} \sum_{j=1}^{a} (\hat{\alpha\beta})_{jk}^2$ $\quad df = (a-1)(b-1)$

Implies that $SSR = SSA + SSB$ for additive model (if $n_{jk} = n_* \forall j, k$).

Proof

Extended ANOVA Table and F Tests

We typically organize the SS information into an ANOVA table:

Source	SS	df	MS	F	p-value
SSR	$n_* \sum_{k=1}^{b} \sum_{j=1}^{a} (\bar{y}_{jk} - \bar{y}_{..})^2$	$ab-1$	MSR	F^*	p^*
SSA	$bn_* \sum_{j=1}^{a} (\bar{y}_{j.} - \bar{y}_{..})^2$	$a-1$	MSA	F_a^*	p_a^*
SSB	$an_* \sum_{k=1}^{b} (\bar{y}_{.k} - \bar{y}_{..})^2$	$b-1$	MSB	F_b^*	p_b^*
SSAB	$n_* \sum_{k=1}^{b} \sum_{j=1}^{a} (\bar{y}_{jk} - \bar{y}_{j.} - \bar{y}_{.k} + \bar{y}_{..})^2$	$(a-1)(b-1)$	$MSAB$	F_{ab}^*	p_{ab}^*
SSE	$\sum_{k=1}^{b} \sum_{j=1}^{a} \sum_{i=1}^{n_*} (y_{ijk} - \bar{y}_{jk})^2$	$ab(n_*-1)$	MSE		
SST	$\sum_{k=1}^{b} \sum_{j=1}^{a} \sum_{i=1}^{n_*} (y_{ijk} - \bar{y}_{..})^2$	abn_*-1			

$MSR = \frac{SSR}{ab-1}$, $MSA = \frac{SSA}{a-1}$, $MSB = \frac{SSB}{b-1}$, $MSAB = \frac{SSAB}{(a-1)(b-1)}$, $MSE = \frac{SSE}{ab(n_*-1)}$,

$F^* = \frac{MSR}{MSE} \sim F_{ab-1,ab(n_*-1)}$ and $p^* = P(F_{ab-1,ab(n_*-1)} > F^*)$,

$F_a^* = \frac{MSA}{MSE} \sim F_{a-1,ab(n_*-1)}$ and $p_a^* = P(F_{a-1,ab(n_*-1)} > F_a^*)$,

$F_b^* = \frac{MSB}{MSE} \sim F_{b-1,ab(n_*-1)}$ and $p_b^* = P(F_{b-1,ab(n_*-1)} > F_b^*)$,

$F_{ab}^* = \frac{MSAB}{MSE} \sim F_{(a-1)(b-1),ab(n_*-1)}$ and $p_{ab}^* = P(F_{(a-1)(b-1),ab(n_*-1)} > F_{ab}^*)$,

ANOVA Table F Tests

F^* and p^*-value are testing $H_0 : \alpha_j = \beta_k = (\alpha\beta)_{jk} = 0 \; \forall j, k$ versus $H_1 : (\exists j, k \in \{1, \ldots, a\} \times \{1, \ldots, b\})(\alpha_j = \beta_k = (\alpha\beta)_{jk} = 0$ is false)

- Equivalent to $H_0 : \mu_{jk} = \mu \; \forall j, k$ versus H_1 : not all μ_{jk} are equal

F_a^* statistic and p_a^*-value are testing $H_0 : \alpha_j = 0 \; \forall j$ versus $H_1 : (\exists j \in \{1, \ldots, a\})(\alpha_j \neq 0)$

- Testing main effect of first factor

F_b^* statistic and p_b^*-value are testing $H_0 : \beta_k = 0 \; \forall k$ versus $H_1 : (\exists k \in \{1, \ldots, b\})(\beta_k \neq 0)$

- Testing main effect of second factor

F_{ab}^* statistic and p_{ab}^*-value are testing $H_0 : (\alpha\beta)_{jk} = 0 \; \forall j, k$ versus $H_1 : (\exists j, k \in \{1, \ldots, a\} \times \{1, \ldots, b\})((\alpha\beta)_{jk} \neq 0)$

- Testing interaction effect

Hypertension Example: Data Description

Hypertension example from Maxwell & Delany (2003).

Total of $n = 72$ subjects participate in hypertension experiment.

- Factor A: `drug` type ($a = 3$ levels: X, Y, Z)
- Factor B: `diet` type ($b = 2$ levels: yes, no)

Randomly assign $n_{jk} = 12$ subjects to each treatment cell:

- Note there are $(ab) = (3)(2) = 6$ treatment cells
- Observations are independent within and between cells

Hypertension Example: Descriptive Statistics

Sum of blood pressure for each treatment cell ($\sum_{i=1}^{12} y_{ijk}$):

	Diet		
Drug	No ($k = 1$)	Yes ($k = 2$)	Total
X ($j = 1$)	2136	2052	4188
Y ($j = 2$)	2424	2154	4578
Z ($j = 3$)	2388	2130	4518
Total	6948	6336	13284

Sum-of-squares of blood pressure for each treatment cell ($\sum_{i=1}^{12} y_{ijk}^2$):

	Diet		
Drug	No ($k = 1$)	Yes ($k = 2$)	Total
X ($j = 1$)	382368	352518	734886
Y ($j = 2$)	491008	388898	879906
Z ($j = 3$)	478238	380462	858700
Total	1351614	1121878	2473492

Hypertension Example: OLS Estimation (by hand)

Least-squares estimates are cell means: $\hat{\mu}_{jk} = \bar{y}_{jk}$ and

$$\hat{\mu} = \frac{1}{abn_*} \sum_{k=1}^{b} \sum_{j=1}^{a} \sum_{i=1}^{n_*} y_{ijk} = \bar{y}_{..} = \tfrac{13284}{72} = 184.5$$

$$\hat{\alpha}_1 = \left(\frac{1}{bn_*} \sum_{k=1}^{b} \sum_{i=1}^{n_*} y_{i1k} \right) - \hat{\mu} = \bar{y}_{1.} - \bar{y}_{..} = \frac{4188}{24} - 184.5 = -10$$

$$\hat{\alpha}_2 = \left(\frac{1}{bn_*} \sum_{k=1}^{b} \sum_{i=1}^{n_*} y_{i2k} \right) - \hat{\mu} = \bar{y}_{2.} - \bar{y}_{..} = \frac{4578}{24} - 184.5 = 6.25$$

$$\hat{\alpha}_3 = \left(\frac{1}{bn_*} \sum_{k=1}^{b} \sum_{i=1}^{n_*} y_{i3k} \right) - \hat{\mu} = \bar{y}_{3.} - \bar{y}_{..} = \frac{4518}{24} - 184.5 = 3.75$$

$$\hat{\beta}_1 = \left(\frac{1}{an_*} \sum_{j=1}^{a} \sum_{i=1}^{n_*} y_{ij1} \right) - \hat{\mu} = \bar{y}_{.1} - \bar{y}_{..} = \frac{6948}{36} - 184.5 = 8.5$$

$$\hat{\beta}_2 = \left(\frac{1}{an_*} \sum_{j=1}^{a} \sum_{i=1}^{n_*} y_{ij2} \right) - \hat{\mu} = \bar{y}_{.2} - \bar{y}_{..} = \frac{6336}{36} - 184.5 = -8.5$$

Hypertension Example: OLS Estimation (by hand)

Continuing from the previous slide...

$$(\hat{\alpha\beta})_{11} = \bar{y}_{11} - \bar{y}_{1\cdot} - \bar{y}_{\cdot 1} + \bar{y}_{\cdot\cdot} = \frac{2136}{12} - \frac{4188}{24} - \frac{6948}{36} + 184.5 = -5$$

$$(\hat{\alpha\beta})_{12} = \bar{y}_{12} - \bar{y}_{1\cdot} - \bar{y}_{\cdot 2} + \bar{y}_{\cdot\cdot} = \frac{2052}{12} - \frac{4188}{24} - \frac{6336}{36} + 184.5 = 5$$

$$(\hat{\alpha\beta})_{21} = \bar{y}_{21} - \bar{y}_{2\cdot} - \bar{y}_{\cdot 1} + \bar{y}_{\cdot\cdot} = \frac{2424}{12} - \frac{4578}{24} - \frac{6948}{36} + 184.5 = 2.75$$

$$(\hat{\alpha\beta})_{22} = \bar{y}_{22} - \bar{y}_{2\cdot} - \bar{y}_{\cdot 2} + \bar{y}_{\cdot\cdot} = \frac{2154}{12} - \frac{4578}{24} - \frac{6336}{36} + 184.5 = -2.75$$

$$(\hat{\alpha\beta})_{31} = \bar{y}_{31} - \bar{y}_{3\cdot} - \bar{y}_{\cdot 1} + \bar{y}_{\cdot\cdot} = \frac{2388}{12} - \frac{4518}{24} - \frac{6948}{36} + 184.5 = 2.25$$

$$(\hat{\alpha\beta})_{32} = \bar{y}_{32} - \bar{y}_{3\cdot} - \bar{y}_{\cdot 2} + \bar{y}_{\cdot\cdot} = \frac{2130}{12} - \frac{4518}{24} - \frac{6336}{36} + 184.5 = -2.25$$

Hypertension Example: Enter Data (in R)

```
> bp = scan("/Users/Nate/Desktop/hypertension.dat")
Read 72 items
> diet = factor(rep(rep(c("no","yes"),each=6),6))
> drug = factor(rep(rep(c("X","Y","Z"),each=12),2))
> biof = factor(rep(c("present","absent"),each=36))
> hyper = data.frame(bp=bp, diet=diet, drug=drug, biof=biof)
> hyper[1:20,]
   bp diet drug    biof
1  170   no    X present
2  175   no    X present
3  165   no    X present
4  180   no    X present
5  160   no    X present
6  158   no    X present
7  161  yes    X present
8  173  yes    X present
9  157  yes    X present
10 152  yes    X present
11 181  yes    X present
12 190  yes    X present
13 186   no    Y present
14 194   no    Y present
15 201   no    Y present
16 215   no    Y present
17 219   no    Y present
18 209   no    Y present
19 164  yes    Y present
20 166  yes    Y present
```

Hypertension Example: OLS Estimation (in R)

Effect coding for `drug` and `diet`:

```
> contrasts(hyper$drug) <- contr.sum(3)
> contrasts(hyper$drug)
  [,1] [,2]
X    1    0
Y    0    1
Z   -1   -1
> contrasts(hyper$diet) <- contr.sum(2)
> contrasts(hyper$diet)
    [,1]
no     1
yes   -1
> mymod = lm(bp ~ drug * diet, data=hyper)
> summary(mymod)     # I deleted some output

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  184.500      1.642 112.355  < 2e-16 ***
drug1        -10.000      2.322  -4.306 5.64e-05 ***
drug2          6.250      2.322   2.691  0.00901 **
diet1          8.500      1.642   5.176 2.30e-06 ***
drug1:diet1   -5.000      2.322  -2.153  0.03498 *
drug2:diet1    2.750      2.322   1.184  0.24059
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.93 on 66 degrees of freedom
Multiple R-squared:  0.4329, Adjusted R-squared:  0.3899
F-statistic: 10.07 on 5 and 66 DF,  p-value: 3.385e-07
```

Hypertension Example: Sums-of-Squares (by hand 1)

Defining $n = \sum_{k=1}^{b} \sum_{j=1}^{a} n_{jk}$, the relevant sums-of-squares are

$$SST = \sum_{k=1}^{b} \sum_{j=1}^{a} \sum_{i=1}^{n_{jk}} (y_{ijk} - \bar{y}_{\cdot\cdot})^2 = \sum_{k=1}^{b} \sum_{j=1}^{a} \sum_{i=1}^{n_{jk}} y_{ijk}^2 - \frac{1}{n} \left(\sum_{k=1}^{b} \sum_{j=1}^{a} \sum_{i=1}^{n_{jk}} y_{ijk} \right)^2$$
$$= 2473492 - \frac{1}{72}(13284)^2 = 22594$$

$$SSE = \sum_{k=1}^{b} \sum_{j=1}^{a} \sum_{i=1}^{n_{jk}} (y_{ijk} - \bar{y}_{jk})^2 = \sum_{k=1}^{b} \sum_{j=1}^{a} \sum_{i=1}^{n_{jk}} y_{ijk}^2 - \sum_{k=1}^{b} \sum_{j=1}^{a} \frac{\left(\sum_{i=1}^{n_{jk}} y_{ijk} \right)^2}{n_{jk}}$$
$$= 2473492 - \left[2136^2 + 2052^2 + 2424^2 + 2154^2 + 2388^2 + 2130^2 \right] / 12 = 12814$$

$$SSR = SST - SSE = 22594 - 12814 = 9780$$

Hypertension Example: Sums-of-Squares (by hand 2)

The sums-of-squares for the main and interaction effects are given by

$$\begin{aligned} SSA &= bn_* \sum_{j=1}^{a} (\bar{y}_{j\cdot} - \bar{y}_{\cdot\cdot})^2 = bn_* \sum_{j=1}^{a} \hat{\alpha}_j^2 \\ &= (2)(12)\left[(-10)^2 + 6.25^2 + 3.75^2\right] = 3675 \end{aligned}$$

$$\begin{aligned} SSB &= an_* \sum_{k=1}^{b} (\bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot})^2 = an_* \sum_{k=1}^{b} \hat{\beta}_k^2 \\ &= (3)(12)\left[(-8.5)^2 + 8.5^2\right] = 5202 \end{aligned}$$

$$\begin{aligned} SSAB &= n_* \sum_{k=1}^{b} \sum_{j=1}^{a} (\bar{y}_{jk} - \bar{y}_{j\cdot} - \bar{y}_{\cdot k} + \bar{y}_{\cdot\cdot})^2 = n_* \sum_{k=1}^{b} \sum_{j=1}^{a} (\hat{\alpha\beta})_{jk}^2 \\ &= 12\left[(-5)^2 + 5^2 + 2.75^2 + (-2.75)^2 + 2.25^2 + (-2.25)^2\right] = 903 \end{aligned}$$

and since $n_{jk} = n_* = 12 \; \forall j, k$, we have

$$SSR = SSA + SSB + SSAB$$

$$9780 = 3675 + 5202 + 903$$

Hypertension Example: ANOVA Table (by hand)

Putting things together in ANOVA table:

Source	SS	df	MS	F	p-value
SSR	9780	5	1956.0	10.07	$< .0001$
SSA	3675	2	1837.5	9.46	0.0002
SSB	5202	1	5202.0	26.79	$< .0001$
SSAB	903	2	451.5	2.33	0.1057
SSE	12814	66	194.2		
SST	22594	71			

Hypertension Example: ANOVA Table (in R)

```
> anova(mymod)
Analysis of Variance Table

Response: bp
          Df Sum Sq Mean Sq F value    Pr(>F)
drug       2   3675  1837.5  9.4643 0.0002433 ***
diet       1   5202  5202.0 26.7935 2.305e-06 ***
drug:diet  2    903   451.5  2.3255 0.1056925
Residuals 66  12814   194.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple Comparisons Overview

Still have multiple comparison problem:

- Overall test is not very informative
- Can examine effect estimates for group differences
- Need follow-up tests to examine linear combinations of means

Still can use the same tools as before:

- Bonferroni
- Tukey (Tukey-Kramer)
- Scheffé

Two-Way ANOVA Linear Combinations

Assuming interaction model, we now have

$$\hat{L} = \sum_{k=1}^{b} \sum_{j=1}^{a} c_{jk} \bar{y}_{jk} \qquad \text{and} \qquad \hat{V}(\hat{L}) = \hat{\sigma}^2 \sum_{k=1}^{b} \sum_{j=1}^{a} c_{jk}^2 / n_{jk}$$

where c_{jk} are the coefficients and $\hat{\sigma}^2$ is the MSE.

Assuming the additive model, we have

$$\hat{L}_a = \sum_{j=1}^{a} c_j \bar{y}_{j\cdot} \qquad \text{and} \qquad \hat{V}(\hat{L}_a) = \hat{\sigma}^2 \sum_{j=1}^{a} c_j^2 / n_{j\cdot}$$
$$\hat{L}_b = \sum_{k=1}^{b} c_k \bar{y}_{\cdot k} \qquad \text{and} \qquad \hat{V}(\hat{L}_b) = \hat{\sigma}^2 \sum_{k=1}^{b} c_k^2 / n_{\cdot k}$$

where c_j and c_k are main effect coefficients, $\hat{\sigma}^2$ is the MSE, and $n_{j\cdot} = \sum_{k=1}^{b} n_{jk}$ and $n_{\cdot k} = \sum_{j=1}^{a} n_{jk}$ are the marginal sample sizes.

Two-Way Multiple Comparisons in Practice

For interaction model, you follow-up on $\hat{\mu}_{jk} = \bar{y}_{jk}$

- Bonferroni for any f tests (independent or not)
- Tukey (Tukey-Kramer) for all pairwise comparisons
- Scheffé for all possible contrasts

For additive model, you follow-up on $\hat{\mu}_j = \bar{y}_{j\cdot}$ and $\hat{\mu}_k = \bar{y}_{\cdot k}$

- Bonferroni for any f tests (independent or not)
- Tukey (Tukey-Kramer) for all pairwise comparisons
- Scheffé for all possible contrasts

For additive model, Tukey and Scheffé control FWER for each main effect family separately.

- Use Bonferroni in combination with Tukey/Scheffé to control FWER for both families simultaneously

Hypertension Example: Interaction (by hand)

All $ab(ab-1)/2 = 15$ possible pairwise comparisons of $\hat{\mu}_{jk}$:

$$\hat{L} = \bar{y}_{jk} - \bar{y}_{j'k'} \qquad \text{and} \qquad \hat{V}(\hat{L}) = 194.2(2/12) = 32.36667$$

and we know that $\frac{\sqrt{2}(\hat{L})}{\sqrt{\hat{V}(\hat{L})}} \sim q_{ab,abn_*-ab}$, so $100(1-\alpha)\%$ CI is given by

$$\hat{L} \pm \frac{1}{\sqrt{2}} q^{(\alpha)}_{ab,abn_*-ab} \sqrt{\hat{V}(\hat{L})}$$

where $q^{(\alpha)}_{ab,abn_*-ab}$ is critical value from studentized range.

For example, 95% CI for $\mu_{21} - \mu_{11}$ is given by:

$$(\hat{\mu}_{21} - \hat{\mu}_{11}) \pm \frac{1}{\sqrt{2}} q^{(.05)}_{6,66} \sqrt{\hat{V}(\hat{L})}$$

$$\left(\frac{2424}{12} - \frac{2136}{12}\right) \pm \frac{1}{\sqrt{2}}(4.150851)\sqrt{32.36667} = [7.303829;\ \ 40.69617]$$

Hypertension Example: Interaction (in R)

All $ab(ab-1)/2 = 15$ possible pairwise comparisons of $\hat{\mu}_{jk}$:

```
> mymod = aov(bp ~ drug * diet, data=hyper)
> TukeyHSD(mymod, "drug:diet")
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = bp ~ drug * diet, data = hyper)

$`drug:diet`
             diff        lwr        upr     p adj
Y:no-X:no    24.0   7.303829  40.696171 0.0010415
Z:no-X:no    21.0   4.303829  37.696171 0.0058124
X:yes-X:no   -7.0 -23.696171   9.696171 0.8203137
Y:yes-X:no    1.5 -15.196171  18.196171 0.9998189
Z:yes-X:no   -0.5 -17.196171  16.196171 0.9999992
Z:no-Y:no    -3.0 -19.696171  13.696171 0.9948741
X:yes-Y:no  -31.0 -47.696171 -14.303829 0.0000117
Y:yes-Y:no  -22.5 -39.196171  -5.803829 0.0025081
Z:yes-Y:no  -24.5 -41.196171  -7.803829 0.0007710
X:yes-Z:no  -28.0 -44.696171 -11.303829 0.0000856
Y:yes-Z:no  -19.5 -36.196171  -2.803829 0.0128988
Z:yes-Z:no  -21.5 -38.196171  -4.803829 0.0044123
Y:yes-X:yes   8.5  -8.196171  25.196171 0.6690751
Z:yes-X:yes   6.5 -10.196171  23.196171 0.8616371
Z:yes-Y:yes  -2.0 -18.696171  14.696171 0.9992610
```

Hypertension Example: Additive (by hand A part 1)

All $a(a-1)/2 = 3$ possible pairwise comparisons of $\hat{\mu}_j$:

$$\begin{aligned}
\mathtt{Y} - \mathtt{X}: &\quad \hat{L}_{a_1} = \frac{4578}{24} - \frac{4188}{24} = 16.25 \\
\mathtt{Z} - \mathtt{X}: &\quad \hat{L}_{a_2} = \frac{4518}{24} - \frac{4188}{24} = 13.75 \\
\mathtt{Z} - \mathtt{Y}: &\quad \hat{L}_{a_3} = \frac{4518}{24} - \frac{4578}{24} = -2.5
\end{aligned}$$

and the variance is given by

$$\hat{V}(\hat{L}_{a_j}) = \hat{\sigma}^2 \textstyle\sum_{j=1}^{a} c_j^2 / n_{j\cdot} = (201.7206)(2/24) = 16.81005$$

where $\hat{\sigma}^2 = \frac{SSE+SSAB}{abn_* - (a+b-1)} = \frac{12814+903}{68} = 201.7206$

Hypertension Example: Additive (by hand A part 2)

Note $\frac{\sqrt{2}(\hat{L}_{a_j})}{\sqrt{\hat{V}(\hat{L}_{a_j})}} \sim q_{a,abn_*-(a+b-1)}$, so $100(1-\alpha)\%$ CI is given by

$$\hat{L}_{a_j} \pm \frac{1}{\sqrt{2}} q^{(\alpha)}_{a,abn_*-(a+b-1)} \sqrt{\hat{V}(\hat{L}_{a_j})}$$

where $q^{(\alpha)}_{a,abn_*-(a+b-1)}$ is critical value from studentized range.

The 95% CI for all three pairwise comparisons is given by

$$\hat{L}_{a_1} \pm \frac{1}{\sqrt{2}} q^{(.05)}_{3,68} \sqrt{\hat{V}(\hat{L}_{a_1})} = 16.25 \pm \frac{1}{\sqrt{2}}(3.388576)\sqrt{16.81005} = [6.426037;\ \ 26.07396]$$

$$\hat{L}_{a_2} \pm \frac{1}{\sqrt{2}} q^{(.05)}_{3,68} \sqrt{\hat{V}(\hat{L}_{a_2})} = 13.75 \pm \frac{1}{\sqrt{2}}(3.388576)\sqrt{16.81005} = [3.926037;\ \ 23.57396]$$

$$\hat{L}_{a_3} \pm \frac{1}{\sqrt{2}} q^{(.05)}_{3,68} \sqrt{\hat{V}(\hat{L}_{a_3})} = -2.5 \pm \frac{1}{\sqrt{2}}(3.388576)\sqrt{16.81005} = [-12.32396;\ \ 7.323963]$$

Hypertension Example: Additive (by hand B part 1)

All $b(b-1)/2 = 1$ possible pairwise comparison of $\hat{\mu}_k$:

$$\texttt{yes} - \texttt{no}: \qquad \hat{L}_b = \frac{6336}{36} - \frac{6948}{36} = -17$$

and the variance is given by

$$\hat{V}(\hat{L}_b) = \hat{\sigma}^2 \sum_{k=1}^{b} c_k^2 / n_{\cdot k} = (201.7206)(2/36) = 11.2067$$

where $\hat{\sigma}^2 = \frac{SSE+SSAB}{abn_* - (a+b-1)} = \frac{12814+903}{68} = 201.7206$

Hypertension Example: Additive (by hand B part 2)

Note $\frac{\sqrt{2}(\hat{L}_b)}{\sqrt{\hat{V}(\hat{L}_b)}} \sim q_{b,abn_*-(a+b-1)}$, so $100(1-\alpha)\%$ CI is given by

$$\hat{L}_b \pm \frac{1}{\sqrt{2}} q^{(\alpha)}_{b,abn_*-(a+b-1)} \sqrt{\hat{V}(\hat{L}_b)}$$

where $q^{(\alpha)}_{b,abn_*-(a+b-1)}$ is critical value from studentized range.

The 95% CI for pairwise comparison is given by

$$\begin{aligned}\hat{L}_b \pm \frac{1}{\sqrt{2}} q^{(.05)}_{2,68} \sqrt{\hat{V}(\hat{L}_b)} &= -17 \pm \frac{1}{\sqrt{2}}(2.822019)\sqrt{11.2067} \\ &= [-23.68011;\ \ -10.31989]\end{aligned}$$

Hypertension Example: Additive (in R)

All $a(a-1)/2 = 3$ possible pairwise comparisons of $\hat{\mu}_j$:

```
> mymod = aov(bp ~ drug + diet, data=hyper)
> TukeyHSD(mymod, "drug")
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = bp ~ drug + diet, data = hyper)

$drug
     diff        lwr       upr     p adj
Y-X 16.25   6.426037 26.073963 0.0005220
Z-X 13.75   3.926037 23.573963 0.0036941
Z-Y -2.50 -12.323963  7.323963 0.8152941
```

All $b(b-1)/2 = 1$ possible pairwise comparison of $\hat{\mu}_k$:

```
> mymod = aov(bp ~ drug + diet, data=hyper)
> TukeyHSD(mymod,"diet")
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = bp ~ drug + diet, data = hyper)

$diet
       diff       lwr       upr   p adj
yes-no  -17 -23.68011 -10.31989 3.2e-06
```

Balanced Three-Way ANOVA

Three-Way ANOVA Model (cell means form)

The Three-Way Analysis of Variance (ANOVA) model has the form

$$y_{ijkl} = \mu_{jkl} + e_{ijkl}$$

for $i \in \{1, \ldots, n_{jkl}\}$, $j \in \{1, \ldots, a\}$, $k \in \{1, \ldots, b\}$, $l \in \{1, \ldots, c\}$, where

- $y_{ijkl} \in \mathbb{R}$ is response for i-th subject in factor cell (j, k, l)
- $\mu_{jkl} \in \mathbb{R}$ is population mean for factor cell (j, k, l)
- $e_{ijkl} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$ is a Gaussian error term
- n_{jkl} is number of subjects in cell (j, k, l)
  (note: $n_{jkl} = n_* \forall j, k, l$ in balanced three-way ANOVA)
- (a, b, c) is number of factor levels for Factors (A, B, C)

Implies that $y_{ijkl} \overset{\text{ind}}{\sim} \mathrm{N}(\mu_{jkl}, \sigma^2)$.

OLS Estimation (cell means form)

Similar to balanced two-way ANOVA, we want to minimize

$$\sum_{l=1}^{c}\sum_{k=1}^{b}\sum_{j=1}^{a}\sum_{i=1}^{n_{jkl}}(y_{ijkl}-\mu_{jkl})^2$$

which is equivalent to minimizing $\sum_{i=1}^{n_{jkl}}(y_{ijkl}-\mu_{jkl})^2$ for all j, k, l

Taking the derivative of $SSE_{jkl}=\sum_{i=1}^{n_{jkl}}(y_{ijkl}-\mu_{jkl})^2$, we see that

$$\frac{\mathrm{d}SSE_{jkl}}{\mathrm{d}\mu_{jkl}}=-2\sum_{i=1}^{n_{jkl}}y_{ijkl}+2n_{jkl}\mu_{jkl}$$

and setting to zero and solving gives $\hat{\mu}_{jkl}=\frac{1}{n_{jkl}}\sum_{i=1}^{n_{jkl}}y_{ijkl}=\bar{y}_{jkl}$

Three-Way ANOVA Model (effect coding)

The three-way ANOVA with all interactions assumes that

$$\mu_{jkl} = \mu + \alpha_j + \beta_k + \gamma_l + (\alpha\beta)_{jk} + (\alpha\gamma)_{jl} + (\beta\gamma)_{kl} + (\alpha\beta\gamma)_{jkl}$$

for $j \in \{1, \ldots, a\}$, $k \in \{1, \ldots, b\}$, and $l \in \{1, \ldots, c\}$ where

- μ is overall population mean
- α_j is main effect of first factor such that $\sum_{j=1}^{a} \alpha_j = 0$
- β_k is main effect of second factor such that $\sum_{k=1}^{b} \beta_k = 0$
- γ_l is main effect of third factor such that $\sum_{l=1}^{c} \gamma_l = 0$
- $(\alpha\beta)_{jk}$ is $A * B$ interaction effect such that $\sum_{j=1}^{a} (\alpha\beta)_{jk} = 0 \; \forall k$ and $\sum_{k=1}^{b} (\alpha\beta)_{jk} = 0 \; \forall j$
- $(\alpha\gamma)_{jl}$ is $A * C$ interaction effect such that $\sum_{j=1}^{a} (\alpha\gamma)_{jl} = 0 \; \forall l$ and $\sum_{l=1}^{c} (\alpha\gamma)_{jl} = 0 \; \forall j$
- $(\beta\gamma)_{kl}$ is $B * C$ interaction effect such that $\sum_{k=1}^{b} (\beta\gamma)_{kl} = 0 \; \forall l$ and $\sum_{l=1}^{c} (\beta\gamma)_{kl} = 0 \; \forall k$
- $(\alpha\beta\gamma)_{jkl}$ is $A * B * C$ interaction effect such that $\sum_{j=1}^{a} (\alpha\beta\gamma)_{jkl} = 0 \; \forall k, l$ and $\sum_{k=1}^{b} (\alpha\beta\gamma)_{jkl} = 0 \; \forall j, l$ and $\sum_{l=1}^{c} (\alpha\beta\gamma)_{jkl} = 0 \; \forall j, k$

OLS Estimation (effect coding)

The OLS estimates of the various effects are given by

$$
\begin{aligned}
\hat{\mu} &= \bar{y}_{...} \\
\hat{\alpha}_j &= \bar{y}_{j..} - \bar{y}_{...} \\
\hat{\beta}_k &= \bar{y}_{.k.} - \bar{y}_{...} \\
\hat{\gamma}_l &= \bar{y}_{..l} - \bar{y}_{...} \\
(\hat{\alpha\beta})_{jk} &= \bar{y}_{jk.} - \bar{y}_{j..} - \bar{y}_{.k.} + \bar{y}_{...} \\
(\hat{\alpha\gamma})_{jl} &= \bar{y}_{j.l} - \bar{y}_{j..} - \bar{y}_{..l} + \bar{y}_{...} \\
(\hat{\beta\gamma})_{kl} &= \bar{y}_{.kl} - \bar{y}_{.k.} - \bar{y}_{..l} + \bar{y}_{...} \\
(\hat{\alpha\beta\gamma})_{jkl} &= \bar{y}_{jkl} - [\hat{\mu} + \hat{\alpha}_j + \hat{\beta}_k + \hat{\gamma}_l + (\hat{\alpha\beta})_{jk} + (\hat{\alpha\gamma})_{jl} + (\hat{\beta\gamma})_{kl}]
\end{aligned}
$$

Fitted Values and Residuals

Form of fitted values depends on fit model:

- Additive: $\hat{\mu}_{jkl} = \hat{\mu} + \hat{\alpha}_j + \hat{\beta}_k + \hat{\gamma}_l$
- All 2-way Int: $\hat{\mu}_{jkl} = \hat{\mu} + \hat{\alpha}_j + \hat{\beta}_k + \hat{\gamma}_l + (\hat{\alpha\beta})_{jk} + (\hat{\alpha\gamma})_{jl} + (\hat{\beta\gamma})_{kl}$
- 3-way Int: $\hat{\mu}_{jkl} = \bar{y}_{jkl}$

Residuals have the form

$$\hat{e}_{ijkl} = y_{ijkl} - \hat{\mu}_{jkl}$$

where form of $\hat{\mu}_{jkl}$ depends on fit model (additive versus interaction).

ANOVA Sums-of-Squares

Defining $N = abcn_*$, the three-way ANOVA sums-of-squares are:

- $SST = \sum_{l=1}^{c} \sum_{k=1}^{b} \sum_{j=1}^{a} \sum_{i=1}^{n_*} (y_{ijkl} - \bar{y}_{...})^2 \qquad df = N - 1$
- $SSR = n_* \sum_{l=1}^{c} \sum_{k=1}^{b} \sum_{j=1}^{a} (\bar{y}_{jkl} - \bar{y}_{...})^2 \qquad df = abc - 1$
- $SSE = \sum_{l=1}^{c} \sum_{k=1}^{b} \sum_{j=1}^{a} \sum_{i=1}^{n_*} (y_{ijkl} - \bar{y}_{jkl})^2 \qquad df = N - abc$

$$SSR = SS_A + SS_B + SS_C + SS_{AB} + SS_{AC} + SS_{BC} + SS_{ABC}$$

- $SS_A = bcn_* \sum_{j=1}^{a} \hat{\alpha}_j^2 \qquad df = a - 1$
- $SS_B = acn_* \sum_{k=1}^{b} \hat{\beta}_k^2 \qquad df = b - 1$
- $SS_C = abn_* \sum_{l=1}^{c} \hat{\gamma}_l^2 \qquad df = c - 1$
- $SS_{AB} = cn_* \sum_{k=1}^{b} \sum_{j=1}^{a} (\widehat{\alpha\beta})_{jk}^2 \qquad df = (a-1)(b-1)$
- $SS_{AC} = bn_* \sum_{l=1}^{c} \sum_{j=1}^{a} (\widehat{\alpha\gamma})_{jl}^2 \qquad df = (a-1)(c-1)$
- $SS_{BC} = an_* \sum_{l=1}^{c} \sum_{k=1}^{b} (\widehat{\beta\gamma})_{kl}^2 \qquad df = (b-1)(c-1)$
- $SS_{ABC} = n_* \sum_{l=1}^{c} \sum_{k=1}^{b} \sum_{j=1}^{a} (\widehat{\alpha\beta\gamma})_{jkl}^2$
  $df = (a-1)(b-1)(c-1)$

Memory Example: Data Description (revisited)

Hypertension example from Maxwell & Delany (2003).

Total of $N = 72$ subjects participate in hypertension experiment.

- Factor A: `drug` type ($a = 3$ levels: X, Y, Z)
- Factor B: `diet` type ($b = 2$ levels: yes, no)
- Factor C: `biof` type ($c = 2$ levels: present, absent)

Randomly assign $n_{jkl} = 6$ subjects to each treatment cell:

- Note there are $(abc) = (3)(2)(2) = 12$ treatment cells
- Observations are independent within and between cells

Hypertension Example: Look at Data

```
> bp = scan("/Users/Nate/Desktop/hypertension.dat")
Read 72 items
> diet = factor(rep(rep(c("no","yes"),each=6),6))
> drug = factor(rep(rep(c("X","Y","Z"),each=12),2))
> biof = factor(rep(c("present","absent"),each=36))
> hyper = data.frame(bp=bp, diet=diet, drug=drug, biof=biof)
> contrasts(hyper$drug) <- contr.sum(3)
> contrasts(hyper$diet) <- contr.sum(2)
> contrasts(hyper$biof) <- contr.sum(2)
> hyper[1:15,]
   bp diet drug    biof
1  170   no    X present
2  175   no    X present
3  165   no    X present
4  180   no    X present
5  160   no    X present
6  158   no    X present
7  161  yes    X present
8  173  yes    X present
9  157  yes    X present
10 152  yes    X present
11 181  yes    X present
12 190  yes    X present
13 186   no    Y present
14 194   no    Y present
15 201   no    Y present
```

Hypertension Example: All Interactions

```
> mymod = lm(bp ~ drug * diet * biof, data=hyper)
> anova(mymod)
Analysis of Variance Table

Response: bp
               Df Sum Sq Mean Sq F value    Pr(>F)
drug            2   3675  1837.5 11.7287 5.019e-05 ***
diet            1   5202  5202.0 33.2043 3.053e-07 ***
biof            1   2048  2048.0 13.0723 0.0006151 ***
drug:diet       2    903   451.5  2.8819 0.0638153 .
drug:biof       2    259   129.5  0.8266 0.4424565
diet:biof       1     32    32.0  0.2043 0.6529374
drug:diet:biof  2   1075   537.5  3.4309 0.0388342 *
Residuals      60   9400   156.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypertension Example: All 2-Way Interactions

```
> mymod = lm(bp ~ drug * diet + drug * biof + diet * biof, data=hyper)
> anova(mymod)
Analysis of Variance Table

Response: bp
          Df Sum Sq Mean Sq F value    Pr(>F)
drug       2   3675  1837.5 10.8759 8.940e-05 ***
diet       1   5202  5202.0 30.7899 6.345e-07 ***
biof       1   2048  2048.0 12.1218  0.000919 ***
drug:diet  2    903   451.5  2.6724  0.077043 .
drug:biof  2    259   129.5  0.7665  0.468992
diet:biof  1     32    32.0  0.1894  0.664925
Residuals 62  10475   169.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

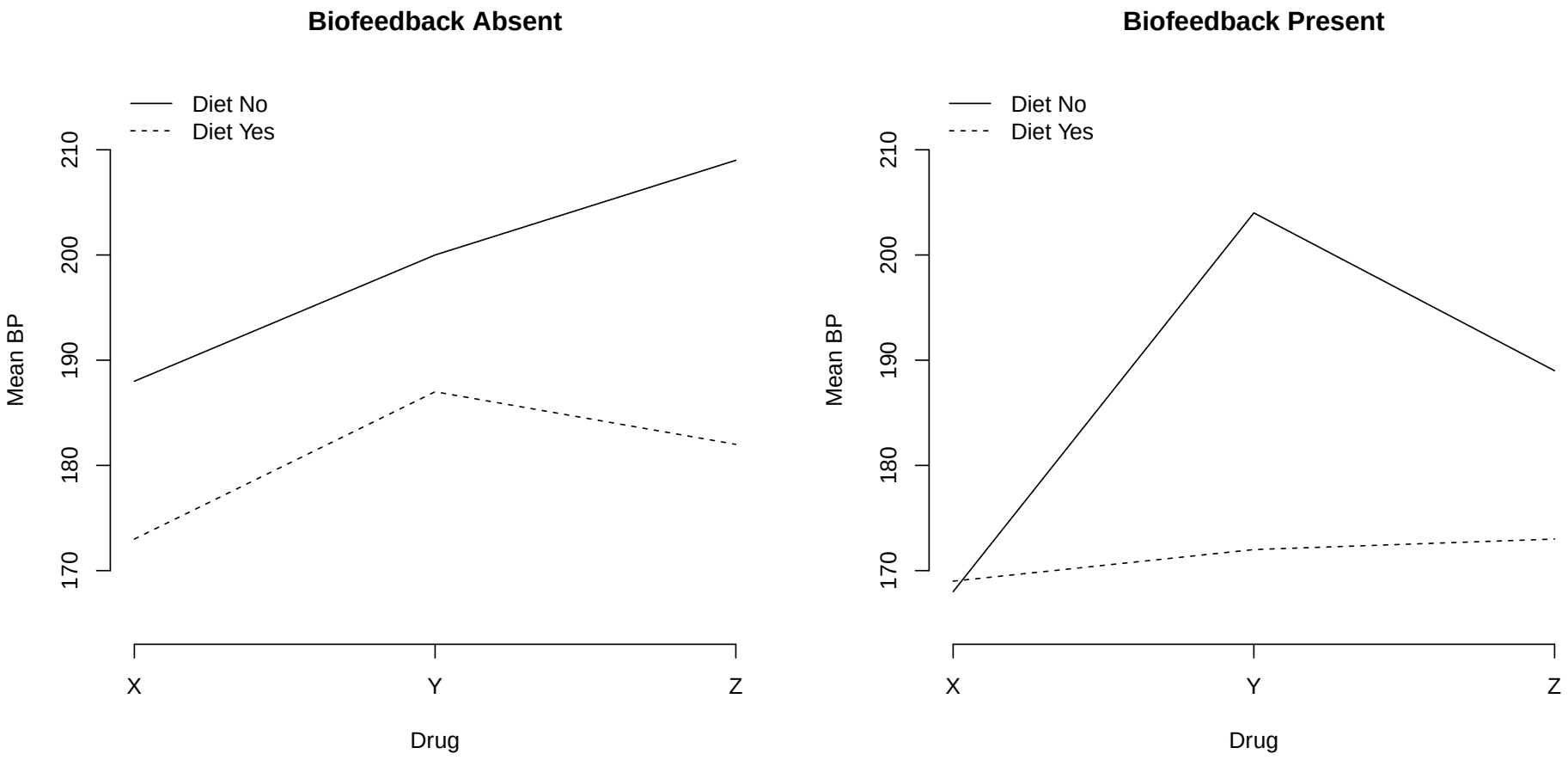
Hypertension Example: Additive Model

```
> mymod = lm(bp ~ drug + diet + biof, data=hyper)
> anova(mymod)
Analysis of Variance Table

Response: bp
          Df Sum Sq Mean Sq F value    Pr(>F)
drug       2   3675  1837.5  10.550 0.0001039 ***
diet       1   5202  5202.0  29.868 7.346e-07 ***
biof       1   2048  2048.0  11.759 0.0010403 **
Residuals 67  11669   174.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypertension Example: Interaction Plot

If you choose the three-way interaction model, you could visualize the interaction using an interaction plot.



Hypertension Example: Interaction Plot (R code)

```
yhat=tapply(hyper$bp,list(hyper$drug,hyper$diet,hyper$biof),mean)
par(mfrow=c(1,2))
mytitles=c("Biofeedback Absent","Biofeedback Present")
for(k in 1:2){
  plot(1:3,yhat[,1,k],ylim=c(165,215),xlab="Drug",
       ylab="Mean BP",main=mytitles[k],axes=FALSE,type="l")
  lines(1:3,yhat[,2,k],lty=2)
  legend("topleft",c("Diet No","Diet Yes"),lty=1:2,bty="n")
  axis(1,at=1:3,labels=c("X","Y","Z"))
  axis(2)
}
```

Hypertension Example: Multiple Comparisons

Assuming we chose the additive model, we would perform follow-up tests on the marginal means.

- Factor A: $\hat{\mu}_{a_j} = \hat{\mu} + \hat{\alpha}_j = \bar{y}_{j\cdot\cdot}$
- Factor B: $\hat{\mu}_{b_k} = \hat{\mu} + \hat{\beta}_k = \bar{y}_{\cdot k\cdot}$
- Factor C: $\hat{\mu}_{c_l} = \hat{\mu} + \hat{\gamma}_l = \bar{y}_{\cdot\cdot l}$

If we chose three-way interaction model, we would perform follow-up tests on the individual cell means.

$$\begin{aligned}\hat{\mu}_{jkl} &= \hat{\mu} + \hat{\alpha}_j + \hat{\beta}_k + \hat{\gamma}_l + (\widehat{\alpha\beta})_{jk} + (\widehat{\alpha\gamma})_{jl} + (\widehat{\beta\gamma})_{kl} + (\widehat{\alpha\beta\gamma})_{jkl} \\ &= \bar{y}_{jkl}\end{aligned}$$

Hypertension Example: Multiple Comparisons

```
> mymod = aov(bp ~ drug + diet + biof, data=hyper)

> TukeyHSD(mymod, "drug")      # I deleted some output
  Tukey multiple comparisons of means
    95% family-wise confidence level

$drug
     diff        lwr       upr     p adj
Y-X 16.25   7.118642 25.381358 0.0001874
Z-X 13.75   4.618642 22.881358 0.0016810
Z-Y -2.50 -11.631358  6.631358 0.7894946

> TukeyHSD(mymod, "diet")      # I deleted some output
  Tukey multiple comparisons of means
    95% family-wise confidence level

$diet
       diff       lwr       upr p adj
yes-no  -17 -23.20877 -10.79123 7e-07

> TukeyHSD(mymod, "biof")      # I deleted some output
  Tukey multiple comparisons of means
    95% family-wise confidence level

$biof
                   diff       lwr       upr     p adj
present-absent -10.66667 -16.87544 -4.457897 0.0010403
```

Unbalanced ANOVA Models

Unbalanced ANOVA: Model Form

Unbalanced ANOVA has same model form as balanced, but unequal sample sizes in each cell.

- 1-way: $n_j \neq n_{j'}$ for some j, j'
- 2-way: $n_{jk} \neq n_{j'k'}$ for some $(jk), (j'k')$
- 3-way: $n_{jkl} \neq n_{j'k'l'}$ for some $(jkl), (j'k'l')$

In the previous slides, we assumed $n_{jk} = n_* \forall j, k$ (two-way ANOVA) or $n_{jkl} = n_* \forall j, k, l$ (three-way ANOVA), which made life easy.

- Effects were orthogonal in balanced design
- Parameter estimates had simple relation to cell/marginal means

Unbalanced ANOVA: Implications

Main consequence for two-way (and higher-way) unbalanced design:

- Non-orthogonal SS (e.g., $SSR \neq SSA + SSB + SSAB$)
- Design is less efficient (larger variances of parameter estimates)

Unbalanced design also affects our estimation and follow-up tests

- Parameter estimates require matrix inversion: $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
- Need to do follow-up tests on least-squares means

Unbalanced ANOVA: Testing Effects

Because of non-orthogonality, cannot test effects using $F = \frac{MS?}{MSE}$.

Instead we use the General Linear Model (GLM) F test statistic:

$$F = \frac{SSE_R - SSE_F}{df_R - df_F} \div \frac{SSE_F}{df_F}$$
$$\sim F_{(df_R - df_F, df_F)}$$

where

- SSE_R is sum-of-squares error for reduced model
- SSE_F is sum-of-squares error for full model
- df_R is error degrees of freedom for reduced model
- df_F is error degrees of freedom for full model

Unbalanced ANOVA: Testing Example

Consider two-way ANOVA and all 7 possible models

$$y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + e_{ijk} \tag{1}$$
$$y_{ijk} = \mu + \alpha_j + \beta_k + e_{ijk} \tag{2}$$
$$y_{ijk} = \mu + \alpha_j + (\alpha\beta)_{jk} + e_{ijk} \tag{3}$$
$$y_{ijk} = \mu + \beta_k + (\alpha\beta)_{jk} + e_{ijk} \tag{4}$$
$$y_{ijk} = \mu + \alpha_j + e_{ijk} \tag{5}$$
$$y_{ijk} = \mu + \beta_k + e_{ijk} \tag{6}$$
$$y_{ijk} = \mu + e_{ijk} \tag{7}$$

Unbalanced ANOVA: Testing Example (continued)

To test effect, use F test comparing full and reduced models.

To test each effect there are multiple choices we could use for full and reduced models:

- A: F=1 and R=4 or F=2 and R=6 or F=5 and R=7
- B: F=1 and R=3 or F=2 and R=5 or F=6 and R=7
- AB: F=1 and R=2 or F=3 and R=5 or F=4 and R=6

Types of Sum-of-Squares

Type I SS

- Amount of additional variation explained by the model when a term is added to the model (aka sequential sum-of-squares).
- In two-way ANOVA, type I SS would compare:
  (a) Main Effect A: F=5 and R=7
  (b) Main Effect B: F=2 and R=5
  (c) Interaction Effect: F=1 and R=2

Type II SS

- Amount of variation a term adds to the model when all other terms are included except terms that "contain" the effect being tested (e.g., $(\alpha\beta)_{jk}$ contains α_j and β_k).
- In two-way ANOVA, type II SS would compare:
  (a) Main Effect A: F=2 and R=6
  (b) Main Effect B: F=2 and R=5
  (c) Interaction Effect: F=1 and R=2

Type III SS

- Amount of variation a term adds to the model when all other terms are included, which is sometimes called partial sum-of-squares.
- In two-way ANOVA, type III SS would compare:
  (a) Main Effect A: F=1 and R=4
  (b) Main Effect B: F=1 and R=3
  (c) Interaction Effect: F=1 and R=2

Types of Sum-of-Squares (in R)

When fitting multi-way ANOVAs, `anova` function gives Type I SS.

- Order matters in unbalanced design!
- `bp = drug + diet` produces different Type I SS tests than `bp = diet + drug` if design is unbalanced

Use `Anova` function in `car` package for Type II and Type III SS.

- Function performs Type II SS tests by default
- Use `type=3` option for Type III SS tests

Hypertension Example: Type I

```
> mymod = lm(bp ~ drug * diet * biof, data=hyper[1:71,])
> anova(mymod)
Analysis of Variance Table

Response: bp
               Df Sum Sq Mean Sq F value    Pr(>F)
drug            2 3733.6  1866.8 11.7306 5.138e-05 ***
diet            1 5113.3  5113.3 32.1311 4.558e-07 ***
biof            1 2087.2  2087.2 13.1154 0.0006101 ***
drug:diet       2  879.5   439.8  2.7633 0.0712569 .
drug:biof       2  280.5   140.3  0.8813 0.4196123
diet:biof       1   24.2    24.2  0.1522 0.6978384
drug:diet:biof  2 1055.8   527.9  3.3172 0.0431275 *
Residuals      59 9389.2   159.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypertension Example: Type II

```
> library(car)
> Anova(mymod, type=2)
Anova Table (Type II tests)

Response: bp
               Sum Sq Df F value    Pr(>F)
drug           3704.1  2 11.6378 5.491e-05 ***
diet           4975.9  1 31.2676 6.085e-07 ***
biof           2061.8  1 12.9561 0.0006541 ***
drug:diet       872.5  2  2.7413 0.0727049 .
drug:biof       277.7  2  0.8726 0.4231893
diet:biof        24.2  1  0.1522 0.6978384
drug:diet:biof 1055.8  2  3.3172 0.0431275 *
Residuals      9389.2 59
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hypertension Example: Type III

```
> library(car)
> Anova(mymod, type=3)
Anova Table (Type III tests)

Response: bp
                Sum Sq Df    F value    Pr(>F)
(Intercept)    2412026  1 15156.7271 < 2.2e-16 ***
drug              3685  2    11.5784 5.730e-05 ***
diet              5057  1    31.7754 5.132e-07 ***
biof              2052  1    12.8967 0.0006713 ***
drug:diet          882  2     2.7705 0.0707910 .
drug:biof          268  2     0.8434 0.4353639
diet:biof           27  1     0.1692 0.6822873
drug:diet:biof    1056  2     3.3172 0.0431275 *
Residuals         9389 59
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix

Ordinary Least-Squares (proof for μ)

Expanding the first summation produces

$$SSE = \sum_{k=1}^{b}\sum_{j=1}^{a}\left[\sum_{i=1}^{n_*} y_{ijk}^2 - 2(\mu + \alpha_j + \beta_k + (\alpha\beta)_{jk})\sum_{i=1}^{n_*} y_{ijk} + n_*(\mu + \alpha_j + \beta_k + (\alpha\beta)_{jk})^2\right]$$

Taking the derivative with respect to μ we have

$$\begin{aligned}\frac{\mathrm{d}SSE}{\mathrm{d}\mu} &= \textstyle\sum_{k=1}^{b}\sum_{j=1}^{a}\left[-2\sum_{i=1}^{n_*} y_{ijk} + 2n_*\mu + 2n_*(\alpha_j + \beta_k + (\alpha\beta)_{jk})\right] \\ &= -2\left(\textstyle\sum_{k=1}^{b}\sum_{j=1}^{a}\sum_{i=1}^{n_*} y_{ijk}\right) + 2abn_*\mu\end{aligned}$$

and setting to zero and solving for μ gives
$\hat{\mu} = \frac{1}{abn_*}\sum_{k=1}^{b}\sum_{j=1}^{a}\sum_{i=1}^{n_*} y_{ijk} = \bar{y}_{..}$

Ordinary Least-Squares (proof for α_j)

Taking the derivative with respect to α_j we have

$$\frac{\mathrm{d}SSE}{\mathrm{d}\alpha_j} = \sum_{k=1}^{b} \left[-2\sum_{i=1}^{n_*} y_{ijk} + 2n_*\alpha_j + 2n_*(\mu + \beta_k + (\alpha\beta)_{jk})\right]$$
$$= -2\left(\sum_{k=1}^{b}\sum_{i=1}^{n_*} y_{ijk}\right) + 2bn_*\alpha_j + 2bn_*\mu$$

and setting to zero, using $\hat{\mu}$ for μ, and solving for α_j gives
$\hat{\alpha}_j = \frac{1}{bn_*}(\sum_{k=1}^{b}\sum_{i=1}^{n_*} y_{ijk}) - \hat{\mu} = \bar{y}_{j\cdot} - \bar{y}_{\cdot\cdot}$

Ordinary Least-Squares (proof for β_k)

Taking the derivative with respect to β_k we have

$$\begin{aligned}\frac{\mathrm{d}SSE}{\mathrm{d}\beta_k} &= \sum_{j=1}^{a}\left[-2\sum_{i=1}^{n_*} y_{ijk} + 2n_*\beta_k + 2n_*(\mu + \alpha_j + (\alpha\beta)_{jk})\right] \\ &= -2\left(\sum_{j=1}^{a}\sum_{i=1}^{n_*} y_{ijk}\right) + 2an_*\beta_k + 2an_*\mu\end{aligned}$$

and setting to zero, using $\hat{\mu}$ for μ, and solving for β_k gives $\hat{\beta}_k = \frac{1}{an_*}(\sum_{j=1}^{a}\sum_{i=1}^{n_*} y_{ijk}) - \hat{\mu} = \bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot}$

Ordinary Least-Squares (proof for $(\alpha\beta)_{jk}$)

Taking the derivative with respect to $(\alpha\beta)_{jk}$ we have

$$\frac{\mathrm{d}SSE}{\mathrm{d}(\alpha\beta)_{jk}} = -2\sum_{i=1}^{n_*} y_{ijk} + 2n_*(\alpha\beta)_{jk} + 2n_*(\mu + \alpha_j + \beta_k)$$

and setting to zero, using $(\hat{\mu}, \hat{\alpha}_j, \hat{\beta}_k)$ for (μ, α_j, β_k), and solving for $(\alpha\beta)_{jk}$ gives $(\hat{\alpha\beta})_{jk} = \frac{1}{n_*}(\sum_{i=1}^{n_*} y_{ijk}) - \hat{\mu} - \hat{\alpha}_j - \hat{\beta}_k = \bar{y}_{jk} - \bar{y}_{j\cdot} - \bar{y}_{\cdot k} + \bar{y}_{\cdot\cdot}$

Return

Partitioning the Variance (proof part 1)

To prove $SSR = SSA + SSB + SSAB$ when $n_{jk} = n_* \forall j, k$, note that

$$y_{ijk} - \bar{y}_{\cdot\cdot} = (y_{ijk} - \bar{y}_{jk}) + (\bar{y}_{jk} - [\bar{y}_{j\cdot} + \bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot}]) + (\bar{y}_{j\cdot} - \bar{y}_{\cdot\cdot}) + (\bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot})$$

Now if we square both sides we have

$$\begin{aligned}(y_{ijk} - \bar{y}_{\cdot\cdot})^2 = {} & (y_{ijk} - \bar{y}_{jk})^2 + (\bar{y}_{jk} - [\bar{y}_{j\cdot} + \bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot}])^2 + (\bar{y}_{j\cdot} - \bar{y}_{\cdot\cdot})^2 + (\bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot})^2 \\ & + 2(y_{ijk} - \bar{y}_{jk}) \{(\bar{y}_{jk} - [\bar{y}_{j\cdot} + \bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot}]) + (\bar{y}_{j\cdot} - \bar{y}_{\cdot\cdot}) + (\bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot})\} \\ & + 2(\bar{y}_{jk} - [\bar{y}_{j\cdot} + \bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot}]) [(\bar{y}_{j\cdot} - \bar{y}_{\cdot\cdot}) + (\bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot})] \\ & + 2(\bar{y}_{j\cdot} - \bar{y}_{\cdot\cdot})(\bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot})\end{aligned}$$

Now if we apply the triple summation we have SST

$$SST = \sum_{k=1}^{b} \sum_{j=1}^{a} \sum_{i=1}^{n_{jk}} (y_{ijk} - \bar{y}_{\cdot\cdot})^2$$

Partitioning the Variance (proof part 2)

First, note that we have

$$SSE = \sum_{k=1}^{b} \sum_{j=1}^{a} \sum_{i=1}^{n_{jk}} (y_{ijk} - \bar{y}_{jk})^2$$
$$SSAB = \sum_{k=1}^{b} \sum_{j=1}^{a} \sum_{i=1}^{n_{jk}} (\bar{y}_{jk} - [\bar{y}_{j\cdot} + \bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot}])^2$$
$$SSA = \sum_{k=1}^{b} \sum_{j=1}^{a} \sum_{i=1}^{n_{jk}} (\bar{y}_{j\cdot} - \bar{y}_{\cdot\cdot})^2$$
$$SSB = \sum_{k=1}^{b} \sum_{j=1}^{a} \sum_{i=1}^{n_{jk}} (\bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot})^2$$

so we need to prove that the crossproduct terms are orthogonal.

To prove that the first crossproduct term sums to zero, define $(\alpha\beta)_{jk} = (\bar{y}_{jk} - [\bar{y}_{j\cdot} + \bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot}]) + (\bar{y}_{j\cdot} - \bar{y}_{\cdot\cdot}) + (\bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot})$ and note that

$$\begin{aligned}\sum_{k=1}^{b} \sum_{j=1}^{a} \sum_{i=1}^{n_{jk}} 2(y_{ijk} - \bar{y}_{jk})(\alpha\beta)_{jk} &= 2 \sum_{k=1}^{b} \sum_{j=1}^{a} (\alpha\beta)_{jk} \sum_{i=1}^{n_{jk}} (y_{ijk} - \bar{y}_{jk}) \\ &= 2 \sum_{k=1}^{b} \sum_{j=1}^{a} (\alpha\beta)_{jk}(0) = 0\end{aligned}$$

because we are summing mean-centered variable.

Partitioning the Variance (proof part 3)

To prove that the second crossproduct term sums to zero, note that $(\hat{\alpha\beta})_{jk} = (\bar{y}_{jk} - [\bar{y}_{j\cdot} + \bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot}])$, $\hat{\alpha}_j = (\bar{y}_{j\cdot} - \bar{y}_{\cdot\cdot})$, and $\hat{\beta}_k = (\bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot})$, so

$$\sum_{k=1}^{b} \sum_{j=1}^{a} \sum_{i=1}^{n_{jk}} 2(\hat{\alpha\beta})_{jk}(\hat{\alpha}_j + \hat{\beta}_k) = 2 \sum_{k=1}^{b} \sum_{j=1}^{a} n_{jk}(\hat{\alpha\beta})_{jk}(\hat{\alpha}_j + \hat{\beta}_k)$$

Now assuming that $n_{jk} = n_* \forall j, k$

$$\sum_{k=1}^{b} \sum_{j=1}^{a} n_{jk}(\hat{\alpha\beta})_{jk}\hat{\alpha}_j = n_* \sum_{j=1}^{a} \hat{\alpha}_j \left(\sum_{k=1}^{b} (\hat{\alpha\beta})_{jk} \right) = n_* \sum_{j=1}^{a} \hat{\alpha}_j(0) = 0$$

$$\sum_{k=1}^{b} \sum_{j=1}^{a} n_{jk}(\hat{\alpha\beta})_{jk}\hat{\beta}_k = n_* \sum_{k=1}^{b} \hat{\beta}_k \left(\sum_{j=1}^{a} (\hat{\alpha\beta})_{jk} \right) = n_* \sum_{k=1}^{b} \hat{\beta}_k(0) = 0$$

Partitioning the Variance (proof part 4)

To prove that the third crossproduct term sums to zero, note that

$$\sum_{k=1}^{b} \sum_{j=1}^{a} \sum_{i=1}^{n_{jk}} 2(\bar{y}_{j\cdot} - \bar{y}_{\cdot\cdot})(\bar{y}_{\cdot k} - \bar{y}_{\cdot\cdot}) = 2 \sum_{k=1}^{b} \sum_{j=1}^{a} n_{jk} \hat{\alpha}_j \hat{\beta}_k$$

and if $n_{jk} = n_* \forall j, k$ we have that

$$\begin{aligned} 2 \sum_{k=1}^{b} \sum_{j=1}^{a} n_{jk} \hat{\alpha}_j \hat{\beta}_k &= 2n_* \sum_{k=1}^{b} \sum_{j=1}^{a} \hat{\alpha}_j \hat{\beta}_k \\ &= 2n_* \sum_{k=1}^{b} \hat{\beta}_k \left(\sum_{j=1}^{a} \hat{\alpha}_j \right) \\ &= 2n_* \sum_{k=1}^{b} \hat{\beta}_k (0) = 0 \end{aligned}$$

which completes the proof.

Return