Two-way ANOVA with balanced and unbalanced designs

Malachy Campbell

Sept. 9, 2020

One-way ANOVA Recap

$$y_{ij} = \mu + \alpha_i + e_{ij} \quad ; e_{ij} \sim N(0, \sigma^2)$$

- t: No. treatments
- n_t: No. replicates for each trt.
- y_{ij}: Observation for ith treatment and jth replicate
- μ: mean for the first treatment
- α_i: Deviation from of ith treatment from μ
- e_{ij}: Random deviation for the ith treatment and jth replicate from treatment mean

One-way ANOVA Recap

$$y_{ij} = \mu + \alpha_i + e_{ij} \;\; ; e_{ij} \sim N(0, \sigma^2)$$

$$\mathbf{y} = (y_{11}, \ldots, y_{1n_1}, y_{21}, \ldots, y_{2n_2}, \ldots, y_{tn_t})'$$
$$\boldsymbol{\beta} = (\mu, \alpha_2, \alpha_3, \ldots, \alpha_t)'$$

- t: No. treatments
- n_t: No. replicates for each trt.
- y_{ij}: Observation for ith treatment and jth replicate
- μ: mean for the first treatment
- α_i: Deviation from of ith treatment from μ
- e_{ij}: Random deviation for the ith treatment and jth replicate from treatment mean

One-way ANOVA Recap

$$y_{ij} = \mu + \alpha_i + e_{ij} \;\; ; e_{ij} \sim N(0, \sigma^2)$$

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$$
$$\mathbf{e} \sim N(0, \sigma^2 \mathbf{I})$$

$$\mathbf{y} = (y_{11}, \ldots, y_{1n_1}, y_{21}, \ldots, y_{2n_2}, \ldots, y_{tn_t})'$$
$$\beta = (\mu, \alpha_2, \alpha_3, \ldots, \alpha_t)'$$

- t: No. treatments
- n_t: No. replicates for each trt.
- y_{ij}: Observation for ith treatment and jth replicate
- μ: mean for the first treatment
- α_i: Deviation from of ith treatment from μ
- e_{ij}: Random deviation for the ith treatment and jth replicate from treatment mean

One-way ANOVA recap – Incidence matrix

X maps observations to fixed terms in the model

Nit.	Rep.	Yld.
180	1	173.3
180	2	182.9
180	3	169.6
200	1	205.9
200	2	208.5
200	3	203.9
220	1	229.1
220	2	231.3
220	3	208.7

Overall mean:

$$\mathbf{X}_{\mathbf{OM}} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

One-way ANOVA recap – Incidence matrix

X maps observations to fixed terms in the model

Nit.	Rep.	Yld.
180	1	173.3
180	2	182.9
180	3	169.6
200	1	205.9
200	2	208.5
200	3	203.9
220	1	229.1
220	2	231.3
220	3	208.7

Overall mean: **Cell means:**

$$\mathbf{X}_{\mathbf{OM}} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \qquad \mathbf{X}_{\mathbf{CM}} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

One-way ANOVA recap – Incidence matrix

X maps observations to fixed terms in the model

Nit.	Rep.	Yld.
180	1	173.3
180	2	182.9
180	3	169.6
200	1	205.9
200	2	208.5
200	3	203.9
220	1	229.1
220	2	231.3
220	3	208.7

Overall mean: **Cell means:** **Additive:**

$$\mathbf{X}_{\mathbf{OM}} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad \mathbf{X}_{\mathbf{CM}} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{X}_{\mathbf{Add.}} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

$$\bar{y} = \mathbf{X}_{\mathbf{OM}}(\mathbf{X}'_{\mathbf{OM}}\mathbf{X}_{\mathbf{OM}})^{-1}\mathbf{X}'_{\mathbf{OM}}\mathbf{y} = \mathbf{H}_{\mathbf{OM}}\mathbf{y}$$

$$\hat{\mathbf{y}} = \mathbf{X}_{\mathbf{CM}}(\mathbf{X}'_{\mathbf{CM}}\mathbf{X}_{\mathbf{CM}})^{-1}\mathbf{X}'_{\mathbf{CM}}\mathbf{y} = \mathbf{H}_{\mathbf{CM}}\mathbf{y}$$

One-way ANOVA recap

Rank of matrix: Number of linearly independent columns (or rows)

Cell means:

$$\mathbf{X}_{\mathbf{CM}} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

Column rank = 3

Additive:

$$\mathbf{X}_{\mathbf{Add.}} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

Column rank = 3

$$\mathbf{X}_{?} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

Column rank = 3

One-way ANOVA recap

We are partitioning the sums of squares with ANOVA

$$\mathrm{SS}_{\mathrm{Tot.}} = \mathrm{SS}_{\mathrm{Reg.}} + \mathrm{SS}_{\mathrm{Err.}}$$

$$\sum_{i}^{t}\sum_{j}^{n_t}(\bar{y} - y_{ij})^2 = \sum_{i}^{t}(\bar{y} - \hat{y}_i)^2 + \sum_{i}^{t}\sum_{j}^{n_t}(\hat{y}_i - y_{ij})^2$$

One-way ANOVA recap

We are partitioning the sums of squares with ANOVA

$$\mathrm{SS}_{\mathrm{Tot.}} = \mathrm{SS}_{\mathrm{Reg.}} + \mathrm{SS}_{\mathrm{Err.}}$$

$$\sum_{i}^{t}\sum_{j}^{n_t}(\bar{y} - y_{ij})^2 = \sum_{i}^{t}(\bar{y} - \hat{y}_i)^2 + \sum_{i}^{t}\sum_{j}^{n_t}(\hat{y}_i - y_{ij})^2$$

$$\begin{aligned}\bar{y} &= \mathbf{X}_{\mathbf{OM}}(\mathbf{X}'_{\mathbf{OM}}\mathbf{X}_{\mathbf{OM}})^{-1}\mathbf{X}'_{\mathbf{OM}}\mathbf{y} = \mathbf{H}_{\mathbf{OM}}\mathbf{y} \\ &= \mathbf{H}_{\mathbf{OM}}\mathbf{y} \\ \hat{\mathbf{y}} &= \mathbf{X}_{\mathbf{CM}}(\mathbf{X}'_{\mathbf{CM}}\mathbf{X}_{\mathbf{CM}})^{-1}\mathbf{X}'_{\mathbf{CM}}\mathbf{y} \\ &= \mathbf{H}_{\mathbf{CM}}\mathbf{y}\end{aligned}$$

$$\mathbf{y}'(\mathbf{I} - \mathbf{H}_{\mathbf{OM}})\mathbf{y} = \mathbf{y}'(\mathbf{H}_{\mathbf{CM}} - \mathbf{H}_{\mathbf{OM}})\mathbf{y} + \mathbf{y}'(\mathbf{I} - \mathbf{H}_{\mathbf{CM}})\mathbf{y}$$

One-way ANOVA recap

We are partitioning the sums of squares with ANOVA

$$\mathrm{SS}_{\mathrm{Tot.}} = \mathrm{SS}_{\mathrm{Reg.}} + \mathrm{SS}_{\mathrm{Err.}}$$

$$\sum_{i}^{t}\sum_{j}^{n_t}(\bar{y} - y_{ij})^2 = \sum_{i}^{t}(\bar{y} - \hat{y}_i)^2 + \sum_{i}^{t}\sum_{j}^{n_t}(\hat{y}_i - y_{ij})^2$$

$$\begin{aligned}\bar{y} &= \mathbf{X}_{\mathbf{OM}}(\mathbf{X}'_{\mathbf{OM}}\mathbf{X}_{\mathbf{OM}})^{-1}\mathbf{X}'_{\mathbf{OM}}\mathbf{y} = \mathbf{H}_{\mathbf{OM}}\mathbf{y}\\ &= \mathbf{H}_{\mathbf{OM}}\mathbf{y}\\ \hat{\mathbf{y}} &= \mathbf{X}_{\mathbf{CM}}(\mathbf{X}'_{\mathbf{CM}}\mathbf{X}_{\mathbf{CM}})^{-1}\mathbf{X}'_{\mathbf{CM}}\mathbf{y}\\ &= \mathbf{H}_{\mathbf{CM}}\mathbf{y}\end{aligned}$$

$$\mathbf{y}'(\mathbf{I} - \mathbf{H}_{\mathbf{OM}})\mathbf{y} = \mathbf{y}'(\mathbf{H}_{\mathbf{CM}} - \mathbf{H}_{\mathbf{OM}})\mathbf{y} + \mathbf{y}'(\mathbf{I} - \mathbf{H}_{\mathbf{CM}})\mathbf{y}$$

Since $\mathrm{SS}_{\mathrm{Reg.}}$ is orthogonal ($\perp$) to $\mathrm{SS}_{\mathrm{Err.}}$, then $(\mathbf{H}_{\mathbf{CM}} - \mathbf{H}_{\mathbf{OM}}) \perp (\mathbf{I} - \mathbf{H}_{\mathbf{CM}})$

- $(\mathbf{H}_{\mathbf{CM}} - \mathbf{H}_{\mathbf{OM}})(\mathbf{I} - \mathbf{H}_{\mathbf{CM}}) = 0$

An example in R

Using dummy variables:

```
> dummyDF
  Int N200 N220      Yld
1   1    0    0 173.7374
2   1    0    0 182.9028
3   1    0    0 169.5862
4   1    1    0 205.9120
5   1    1    0 208.5298
6   1    1    0 203.9329
7   1    0    1 229.1282
8   1    0    1 231.3177
9   1    0    1 208.7234
>
```

Using `aov()`:

An example in R

Using dummy variables:

```
> summary(lm(Yld ~ 1 + N200 + N220, data = dummyDF))

Call:
lm(formula = Yld ~ 1 + N200 + N220, data = dummyDF)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.3330  -2.1920  -0.2129   6.0717   8.2612 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  175.409      4.796  36.574 2.79e-08 ***
N200          30.716      6.783   4.529 0.003981 ** 
N220          47.648      6.783   7.025 0.000415 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.307 on 6 degrees of freedom
Multiple R-squared:  0.8942,	Adjusted R-squared:  0.859 
F-statistic: 25.36 on 2 and 6 DF,  p-value: 0.001183
```

Using aov():

```
>
> summary(aov(Yld ~ Nit, data = Ndata))
            Df Sum Sq Mean Sq F value  Pr(>F)   
Nit          2   3500    1750   25.36 0.00118 **
Residuals    6    414      69                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Two-way ANOVA

Two-way ANOVA: ANOVA with two factors (A and B)

$$\begin{aligned}\mathrm{SS_{Tot.}} &= \mathrm{SS_{Reg.}} + \mathrm{SS_{Err.}}\\ &= \mathrm{SS_{A}} + \mathrm{SS_{B}} + \mathrm{SS_{Err.}}\end{aligned}$$

$$\mathrm{SS_A} \perp \mathrm{SS_B}$$

$\mathrm{SS_A} \perp \mathrm{SS_B}$ in specific designs

Balanced Two-way ANOVA w/ **no interaction**

Researchers are interested in studying the effects of two N regimes (20, 30) on yield for two oat varieties (Corral, Belinda). The field was split into 24 plots (experimental units) and treatment combinations were randomly assigned to each plot. Six observations were recorded for each combination of N level and line ($N = 4 \times 6 = 24$).

Balanced Two-way ANOVA w/ **no interaction**

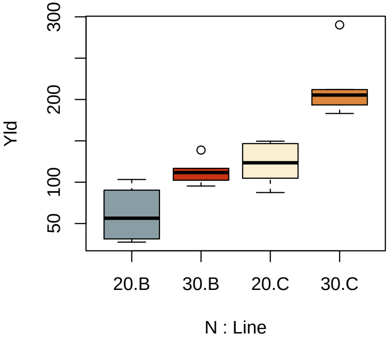
Researchers are interested in studying the effects of two N regimes (20, 30) on yield for two oat varieties (Corral, Belinda). The field was split into 24 plots (experimental units) and treatment combinations were randomly assigned to each plot. Six observations were recorded for each combination of N level and line ($N = 4 \times 6 = 24$).

- **Treatment design:** Each Nitro.-Line treatment combo is assigned to at least one experimental unit $\rightarrow$ *full factorial treatment design*
  - Two factors: Nitrogen and Line
  - Two levels for Nitro.: 20 and 30
  - Two levels for Line: Belinda and Corral

- **Experimental Design:** Treatments are randomly assigned and all are observed an equal number of times $\rightarrow$ *balanced complete randomized design*

Balanced Two-way ANOVA w/ **no interaction**

Researchers are interested in studying the effects of two N regimes (20, 30) on yield for two oat varieties (Corral, Belinda). The field was split into 24 plots (experimental units) and treatment combinations were randomly assigned to each plot. Six observations were recorded for each combination of N level and line ($N = 4 \times 6 = 24$).



Balanced Two-way ANOVA w/ **no interaction**

- $2 \times 2 = 4$ possible treatment combinations

Line	**N**	$\boldsymbol{\mu_i}$
Belinda	20	μ_1
Corral	20	μ_2
Belinda	30	μ_3
Corral	30	μ_4

$$y_{ij} = \mu_i + e_{ij}$$

Cell means model

$$y_{ij} = \mu_i + e_{ij}$$

	Belinda	**Corral**
N: 20	μ_1	μ_2
N: 30	μ_3	μ_4

Marginal Means:

Nitrogen:

- $\mu_{N=20} = \frac{\mu_1+\mu_2}{2}$
- $\mu_{N=30} = \frac{\mu_2+\mu_4}{2}$

Lines:

- $\mu_{Belinda} = \frac{\mu_1+\mu_3}{2}$
- $\mu_{Corral} = \frac{\mu_2+\mu_4}{2}$

Cell means model

$$y_{ij} = \mu_i + e_{ij}$$

	Belinda	**Corral**
N: 20	μ_1	μ_2
N: 30	μ_3	μ_4

Simple effects: Difference between means for levels at specific level of second factor

Compare ***cell means*** *within row or columns*

- Nitro. effects for Belinda $\mu_1 - \mu_3$
- Line effects at low Nitro. $(\mathrm{Nitro.} = 20)$ $\mu_1 - \mu_2$

Cell means model

$$y_{ij} = \mu_i + e_{ij}$$

	Belinda	**Corral**
N: 20	μ_1	μ_2
N: 30	μ_3	μ_4

Main effects: Difference between two levels of a factor across levels of second factor

*Compare **marginal means** for factor*

- Nitro. effects $\frac{\mu_1+\mu_2}{2} - \frac{\mu_2+\mu_4}{2}$
- Line effects $\frac{\mu_1+\mu_3}{2} - \frac{\mu_2+\mu_4}{2}$

Cell means model

$$y_{ij} = \mu_i + e_{ij}$$

	Belinda	**Corral**
N: 20	μ_1	μ_2
N: 30	μ_3	μ_4

Interaction effects: Response for level of one factor is dependant on level of second (e.g. Belinda is insensitive to Nitro. fertilizer, but Corral is sensitive.)

*Comparison of **simple effects***

- $(\mu_1 - \mu_3) - (\mu_2 - \mu_4)$
- If difference is effectively zero, the the simple effects are the same for each line and we can interpret main effects

Balanced Two-way ANOVA w/ **no interaction**

Line	N	Yld
B	20	90.4
B	30	95.4
B	20	27.5
B	30	138.8
⋮	⋮	⋮
C	20	104.8
C	30	183.1
C	20	118.7
C	30	193.5

Overall Mean

$$\mathbf{X}_{\mathbf{OM}} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Balanced Two-way ANOVA w/ **no interaction**

Line	**N**	**Yld**
B	20	90.4
B	30	95.4
B	20	27.5
B	30	138.8
⋮	⋮	⋮
C	20	104.8
C	30	183.1
C	20	118.7
C	30	193.5

Treatments

$$\mathbf{X}_{\mathbf{Nit.}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} ; \mathbf{X}_{\mathbf{Line}} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

Balanced Two-way ANOVA w/ **no interaction**

Line	N	Yld
B	20	90.4
B	30	95.4
B	20	27.5
B	30	138.8
⋮	⋮	⋮
C	20	104.8
C	30	183.1
C	20	118.7
C	30	193.5

Full model

$$\mathbf{X}_{\textbf{Model}} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ \vdots & \vdots & & \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

Balanced Two-way ANOVA w/ **no interaction**

```
> # Get incd. for each
> X_I <- model.matrix(~ 1, dataSet)
> X_N <- model.matrix(~ 0 + N, dataSet)
> X_Line <- model.matrix(~ 0 + Line, dataSet)
> X_full <- model.matrix(~ 0 + Line + N, dataSet)
>
> # Get projection matricies
> Pi <-  X_I %*% solve(t(X_I) %*% X_I) %*% t(X_I)
> Pn <- X_N %*% solve(t(X_N) %*% X_N) %*% t(X_N)
> Pl <- X_Line %*% solve(t(X_Line) %*% X_Line) %*% t(X_Line)
> Pfull <- X_full %*% solve(t(X_full) %*% X_full) %*% t(X_full)
>
> ## SS
> Iden <- diag(1, 24, 24)
> SSl <- t(dataSet$Yld) %*% (Pl - Pi) %*% dataSet$Yld
> SSn <- t(dataSet$Yld) %*% (Pn - Pi) %*% dataSet$Yld
> SSr <- t(dataSet$Yld) %*% (Iden - Pfull) %*% dataSet$Yld
> SSt <- t(dataSet$Yld) %*% (Iden - Pi) %*% dataSet$Yld
>
```

$$\mathrm{SS}_{\mathrm{Nit.}} = \mathbf{y}'(\mathbf{H}_{\mathbf{L}} - \mathbf{H}_{\mathbf{OM}})\mathbf{y}$$
$$\mathrm{SS}_{\mathrm{Line}} = \mathbf{y}'(\mathbf{H}_{\mathbf{L}} - \mathbf{H}_{\mathbf{OM}})\mathbf{y}$$
$$\mathrm{SS}_{\mathrm{Err.}} = \mathbf{y}'(\mathbf{I} - \mathbf{H}_{\mathbf{Model}})\mathbf{y}$$

Balanced Two-way ANOVA w/ **no interaction**

```
> data.frame(Term = c("Line", "Nit.", "Err.", "Tot."),
+            SS = c(SSl, SSn, SSr, SSt),
+            df = c(qr(Pl - Pi)$rank,
+                   qr(Pn - Pi)$rank,
+                   qr(Iden - Pfull)$rank,
+                   qr(Iden - Pi)$rank))
  Term       SS df
1 Line 40268.54  1
2 Nit. 31195.08  1
3 Err. 19759.33 21
4 Tot. 91222.94 23
>
>
> aovSum <- summary(aov(Yld ~ Line + N, dataSet))
> aovSum
            Df Sum Sq Mean Sq F value   Pr(>F)    
Line         1  40269   40269   42.80 1.76e-06 ***
N            1  31195   31195   33.15 1.03e-05 ***
Residuals   21  19759     941                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> sum(aovSum[[1]]$`Sum Sq`)
[1] 91222.94
>
```

Unbalanced Two-way ANOVA w/ **no interaction**

Researchers are interested in studying the effects of two N regimes (20, 30) on yield for two oat varieties (Corral, Belinda). The field was split into 24 plots (experimental units) and treatment combinations were randomly assigned to each plot. Each combination of N level and line, was replicated six times. A hail storm destroyed three plots.

	N=20	**N=30**
Bel.	5	6
Cor.	6	4

Unbalanced Two-way ANOVA w/ **no interaction**

```
> # Unbalanced
> aovSum <- summary(aov(Yld ~ N + Line, dataSet_unbal))
> aovSum
            Df Sum Sq Mean Sq F value   Pr(>F)
N            1  19559   19559   18.71 0.000408 ***
Line         1  35319   35319   33.78 1.66e-05 ***
Residuals   18  18818    1045
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> aovSum <- summary(aov(Yld ~ Line + N, dataSet_unbal))
> aovSum
            Df Sum Sq Mean Sq F value   Pr(>F)
Line         1  27421   27421   26.23 7.14e-05 ***
N            1  27457   27457   26.26 7.09e-05 ***
Residuals   18  18818    1045
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

```
>
> # Balanced
> aovSum <- summary(aov(Yld ~ N + Line, dataSet))
> aovSum
            Df Sum Sq Mean Sq F value   Pr(>F)
N            1  31195   31195   33.15 1.03e-05 ***
Line         1  40269   40269   42.80 1.76e-06 ***
Residuals   21  19759     941
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> aovSum <- summary(aov(Yld ~ Line + N, dataSet))
> aovSum
            Df Sum Sq Mean Sq F value   Pr(>F)
Line         1  40269   40269   42.80 1.76e-06 ***
N            1  31195   31195   33.15 1.03e-05 ***
Residuals   21  19759     941
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

The order of the terms matters!

Unbalanced Two-way ANOVA w/ **no interaction**

```
> # Unbalanced
> aovSum <- summary(aov(Yld ~ N + Line, dataSet_unbal))
> aovSum
            Df Sum Sq Mean Sq F value   Pr(>F)    
N            1  19559   19559   18.71 0.000408 ***
Line         1  35319   35319   33.78 1.66e-05 ***
Residuals   18  18818    1045                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
> aovSum <- summary(aov(Yld ~ Line + N, dataSet_unbal))
> aovSum
            Df Sum Sq Mean Sq F value   Pr(>F)    
Line         1  27421   27421   26.23 7.14e-05 ***
N            1  27457   27457   26.26 7.09e-05 ***
Residuals   18  18818    1045                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

```
> 
> # Balanced
> aovSum <- summary(aov(Yld ~ N + Line, dataSet))
> aovSum
            Df Sum Sq Mean Sq F value   Pr(>F)    
N            1  31195   31195   33.15 1.03e-05 ***
Line         1  40269   40269   42.80 1.76e-06 ***
Residuals   21  19759     941                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> aovSum <- summary(aov(Yld ~ Line + N, dataSet))
> aovSum
            Df Sum Sq Mean Sq F value   Pr(>F)    
Line         1  40269   40269   42.80 1.76e-06 ***
N            1  31195   31195   33.15 1.03e-05 ***
Residuals   21  19759     941                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

The order of the terms matters!

Why?

Unbalanced Two-way ANOVA w/ **no interaction**

```
>
> ## Balanced
> # Get incd. for each
> X_I <- model.matrix(~ 1, dataSet)
> X_N <- model.matrix(~ 0 + N, dataSet)
> X_Line <- model.matrix(~ 0 + Line, dataSet)
> X_full <- model.matrix(~ 0 + Line + N, dataSet)
>
> # Get projection matricies
> Pi <-  X_I %*% solve(t(X_I) %*% X_I) %*% t(X_I)
> Pn <- X_N %*% solve(t(X_N) %*% X_N) %*% t(X_N)
> Pl <- X_Line %*% solve(t(X_Line) %*% X_Line) %*% t(X_Line)
> Pfull <- X_full %*% solve(t(X_full) %*% X_full) %*% t(X_full)
>
> # Check for orthogonality
> Pni <- Pn - Pi
> Pli <- Pl - Pi
> sum(round(Pni %*% Pli, 3))
[1] 0
>
```

```
>
> # Unbalanced
> # Get incd. for each
> X_I <- model.matrix(~ 1, dataSet_unbal)
> X_N <- model.matrix(~ 0 + N, dataSet_unbal)
> X_Line <- model.matrix(~ 0 + Line, dataSet_unbal)
> X_full <- model.matrix(~ 0 + Line + N, dataSet_unbal)
>
> # Get projection matricies
> Pi <-  X_I %*% solve(t(X_I) %*% X_I) %*% t(X_I)
> Pn <- X_N %*% solve(t(X_N) %*% X_N) %*% t(X_N)
> Pl <- X_Line %*% solve(t(X_Line) %*% X_Line) %*% t(X_Line)
> Pfull <- X_full %*% solve(t(X_full) %*% X_full) %*% t(X_full)
>
> Pni <- Pn - Pi
> Pli <- Pl - Pi
> sum(round(Pni %*% Pli, 3))
[1] 0.014
>
>
```

Only the balanced design is orthogonal.

Unbalanced Two-way ANOVA w/ **no interaction**

```
>
> ## Balanced
> # Get incd. for each
> X_I <- model.matrix(~ 1, dataSet)
> X_N <- model.matrix(~ 0 + N, dataSet)
> X_Line <- model.matrix(~ 0 + Line, dataSet)
> X_full <- model.matrix(~ 0 + Line + N, dataSet)
>
> # Get projection matricies
> Pi <-  X_I %*% solve(t(X_I) %*% X_I) %*% t(X_I)
> Pn <- X_N %*% solve(t(X_N) %*% X_N) %*% t(X_N)
> Pl <- X_Line %*% solve(t(X_Line) %*% X_Line) %*% t(X_Line)
> Pfull <- X_full %*% solve(t(X_full) %*% X_full) %*% t(X_full)
>
> # Check for orthogonality
> Pni <- Pn - Pi
> Pli <- Pl - Pi
> sum(round(Pni %*% Pli, 3))
[1] 0
>
```

```
>
> # Unbalanced
> # Get incd. for each
> X_I <- model.matrix(~ 1, dataSet_unbal)
> X_N <- model.matrix(~ 0 + N, dataSet_unbal)
> X_Line <- model.matrix(~ 0 + Line, dataSet_unbal)
> X_full <- model.matrix(~ 0 + Line + N, dataSet_unbal)
>
> # Get projection matricies
> Pi <-  X_I %*% solve(t(X_I) %*% X_I) %*% t(X_I)
> Pn <- X_N %*% solve(t(X_N) %*% X_N) %*% t(X_N)
> Pl <- X_Line %*% solve(t(X_Line) %*% X_Line) %*% t(X_Line)
> Pfull <- X_full %*% solve(t(X_full) %*% X_full) %*% t(X_full)
>
> Pni <- Pn - Pi
> Pli <- Pl - Pi
> sum(round(Pni %*% Pli, 3))
[1] 0.014
>
>
```

Only the balanced design is orthogonal.

The balanced design allows you to vary one factor and keep other constant.

Unbalanced Two-way ANOVA w/ **no interaction**

```
>
> ## Balanced
> # Get incd. for each
> X_I <- model.matrix(~ 1, dataSet)
> X_N <- model.matrix(~ 0 + N, dataSet)
> X_Line <- model.matrix(~ 0 + Line, dataSet)
> X_full <- model.matrix(~ 0 + Line + N, dataSet)
>
> # Get projection matricies
> Pi <-  X_I %*% solve(t(X_I) %*% X_I) %*% t(X_I)
> Pn <- X_N %*% solve(t(X_N) %*% X_N) %*% t(X_N)
> Pl <- X_Line %*% solve(t(X_Line) %*% X_Line) %*% t(X_Line)
> Pfull <- X_full %*% solve(t(X_full) %*% X_full) %*% t(X_full)
>
> # Check for orthogonality
> Pni <- Pn - Pi
> Pli <- Pl - Pi
> sum(round(Pni %*% Pli, 3))
[1] 0
>
```

```
>
> # Unbalanced
> # Get incd. for each
> X_I <- model.matrix(~ 1, dataSet_unbal)
> X_N <- model.matrix(~ 0 + N, dataSet_unbal)
> X_Line <- model.matrix(~ 0 + Line, dataSet_unbal)
> X_full <- model.matrix(~ 0 + Line + N, dataSet_unbal)
>
> # Get projection matricies
> Pi <-  X_I %*% solve(t(X_I) %*% X_I) %*% t(X_I)
> Pn <- X_N %*% solve(t(X_N) %*% X_N) %*% t(X_N)
> Pl <- X_Line %*% solve(t(X_Line) %*% X_Line) %*% t(X_Line)
> Pfull <- X_full %*% solve(t(X_full) %*% X_full) %*% t(X_full)
>
> Pni <- Pn - Pi
> Pli <- Pl - Pi
> sum(round(Pni %*% Pli, 3))
[1] 0.014
>
>
```

Only the balanced design is orthogonal.

Is the F-test valid in the unbalanced design?

Type I ANOVA

The `lm()` function uses a type I approach. Type I ANOVA sequentially adds terms to the model.

For a two-way ANOVA with interaction:

$$
\begin{aligned}
E(y_{ijk}) &= \mu \\
E(y_{ijk}) &= \mu + \alpha_i \\
E(y_{ijk}) &= \mu + \alpha_i + \beta_j \\
E(y_{ijk}) &= \mu + \alpha_i + \beta_j = \mu_{ij}
\end{aligned}
$$

This is why order matters in `lm()`!

Type III ANOVA - unbalanced data

We need to use Type III ANOVA when our data is unbalanced (and we have interaction in model)

- Type III ANOVA considers how much information a term contributes while considering all other terms in the model
- This is done by viewing the full model as a series of nested models
- F-test compares the reduced model with the full model
- In R either use the `car::Anova()` function or call `lm()` followed by `drop1()`

Type III ANOVA – F-test

$$F = \left(\frac{SSE_{red} - SSE_{full}}{df_{red} - df_{full}} \right) \left(\frac{SSE_{full}}{df_{full}} \right)^{-1}$$

- df_{red} and df_{full} are the error df for reduced and full models
- The sampling distribution for F is an F-distribution parameterized by numerator df ($df_{red} - df_{full}$) and denominator df (df_{full})

Type II ANOVA

The issue with a Type III ANOVA is that it doesn't make much sense to interpret main effects for a factor when that factor is included in higher terms.

Type II ANOVA seeks to determine the amount of variation a term explains when all other terms *except* those that depend on it are considered.

e.g. We have an experiment with three factors (A, B, C). We can construct a model that includes all lower and higher order terms (A + B + C + AB + BC + AC + ABC). To determine how much is gained by including A in our model, we compare

- Reduced: B + C + BC
- Full: A + B + C + BC

In summary...

For a three-way ANOVA (A + B + C + AB + AC + BC + ABC)

Type I

$SS(A|1)$
$SS(B|1, A)$
$SS(C|1, A, B)$
$SS(AB|1, A, B, C)$
$SS(AC|1, A, B, C, AB)$
$SS(BC|1, A, B, C, AB, AC)$
$SS(ABC|1, A, B, C, AB, AC, BC)$

Type II

$SS(A|1, B, C, BC)$
$SS(B|1, A, C, AC)$
$SS(C|1, A, B, AB)$
$SS(AB|1, A, B, C, AC, BC)$
$SS(AC|1, A, B, C, AB, BC)$
$SS(BC|1, A, B, C, AB, AC)$
$SS(ABC|1, A, B, C, AB, AC, BC)$

Type III

$SS(A|1, B, C, AB, AC, BC)$
$SS(B|1, A, C, AB, AC, BC)$
$SS(C|1, A, B, AB, AC, BC)$
$SS(AB|1, A, B, C, AC, BC)$
$SS(AC|1, A, B, C, AB, BC)$
$SS(BC|1, A, B, C, AB, AC)$
$SS(ABC|1, A, B, C, AB, AC, BC)$