

STAT 516 Lec 07

Unbalanced two-way factorial design

Karl Gregory

2024-03-28

Mice example

Change in mRNA expression of *esr1* in wild-type vs gen-modified mice on low- and high-fat diets.

Data from Trey Hope in Dr. Enos' lab in the USC School of Medicine.

WT LFD	ADIPO-ER α LFD	WT HFD	ADIPO-ER α HFD
493	555	595.5	1153
172	302	418.5	890.5
617	434	642	539
534	90	743.5	651.5
500.5	72	1351	585
127	489	1180	783.5
1224	453	938.5	533.5
143	552.5	670	895
852	635.5	1319	579
134	330.9	1007	551
547.5	169.5	589	800
	737.5	481.5	1026
		785	308
		1060	1415
		435	342.5
		535.5	197.5

Does the diet affect the response? The genetic modification? Is the effect of the diet the same in the wild-type and the gen-modified mice?

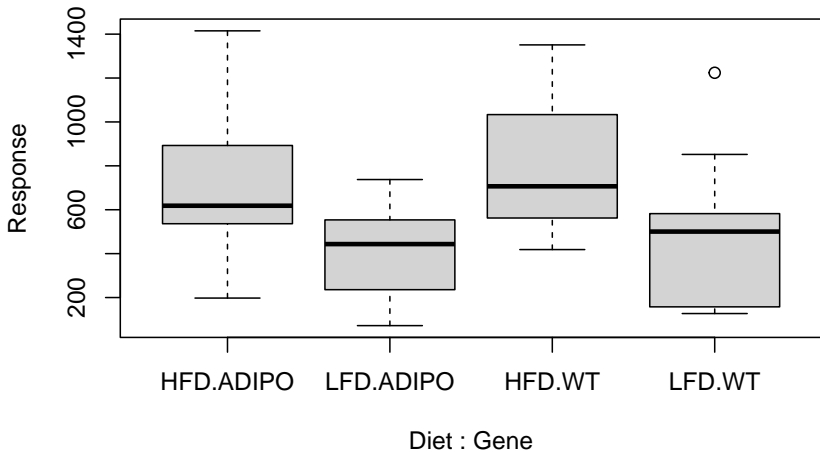
Now in a .csv file with one column for each factor.

```
Response, Gene, Diet  
493, WT, LFD  
172, WT, LFD  
617, WT, LFD  
534, WT, LFD  
500.5, WT, LFD  
127, WT, LFD  
1224, WT, LFD  
143, WT, LFD  
852, WT, LFD  
134, WT, LFD  
547.5, WT, LFD  
555, ADIPO, LFD  
302, ADIPO, LFD  
434, ADIPO, LFD  
90, ADIPO, LFD  
72, ADIPO, LFD  
489, ADIPO, LFD  
453, ADIPO, LFD  
552.5, ADIPO, LFD  
635.5, ADIPO, LFD  
330.9, ADIPO, LFD  
169.5, ADIPO, LFD  
737.5, ADIPO, LFD  
595.5, WT, HFD  
418.5, WT, HFD
```

```
link <- url("https://people.stat.sc.edu/gregorkb/data/ADIPOER.csv")
adipoer <- read.csv(link)
head(adipoer, n = 14)
```

	Response	Gene	Diet
1	493.0	WT	LFD
2	172.0	WT	LFD
3	617.0	WT	LFD
4	534.0	WT	LFD
5	500.5	WT	LFD
6	127.0	WT	LFD
7	1224.0	WT	LFD
8	143.0	WT	LFD
9	852.0	WT	LFD
10	134.0	WT	LFD
11	547.5	WT	LFD
12	555.0	ADIPO	LFD
13	302.0	ADIPO	LFD
14	434.0	ADIPO	LFD

```
boxplot(Response ~ Diet + Gene, data = adipoer)
```



```
grp_means <- aggregate(Response ~ Diet + Gene, data = adipoer, mean)
grp_means
```

	Diet	Gene	Response
1	HFD	ADIPO	703.1250
2	LFD	ADIPO	401.7417
3	HFD	WT	796.9375
4	LFD	WT	485.8182

```
grp_counts <- aggregate(Response ~ Diet + Gene, data = adipoer, length)
grp_counts
```

	Diet	Gene	Response
1	HFD	ADIPO	16
2	LFD	ADIPO	12
3	HFD	WT	16
4	LFD	WT	11

Two-way treatment effects model

The two-way treatment effects model gives

$$Y_{ijk} = \mu + \tau_i + \gamma_j + (\tau\gamma)_{ij} + \varepsilon_{ijk}$$

for $i = 1, \dots, a$, $j = 1, \dots, b$, and $k = 1, \dots, n_{ij}$, where

- ▶ Y_{ijk} is the response for EU k under level i of A and level j of B.
- ▶ μ represents a baseline or overall mean.
- ▶ The τ_i are the main effects for Factor A.
- ▶ The γ_j are the main effects for Factor B.
- ▶ The $(\tau\gamma)_{ij}$ are the interaction effects between A and B.
- ▶ The ε_{ijk} are $\text{Normal}(0, \sigma^2)$ error terms.

Now suppose n_{ij} are not all the same, so the design is unbalanced.

Parameter constraints in the unbalanced case

We have $1 + a + b + ab$ parameters for ab treatment means...

1. To give μ a baseline interpretation, set

$$\tau_a = 0, \quad \gamma_b = 0, \quad \text{and} \quad (\tau\gamma)_{aj} = (\tau\gamma)_{ib} = 0 \text{ for all } i, j.$$

2. To give μ an overall mean interpretation, set

$$\sum_{i=1}^a n_{i.} \tau_i = 0, \quad \sum_{j=1}^b n_{.j} \gamma_j = 0,$$

where $n_{i.} = \sum_{j=1}^b n_{ij}$ and $n_{.j} = \sum_{i=1}^a n_{ij}$, as well as

$$\sum_{j=1}^b n_{ij} (\tau\gamma)_{ij} = 0 \text{ for all } i \text{ and } \sum_{i=1}^a n_{ij} (\tau\gamma)_{ij} = 0 \text{ for all } j.$$

Estimating the error term variance

An unbiased estimator of the error term variance σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{N - ab} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{ij.})^2,$$

where $N = \sum_{i=1}^a \sum_{j=1}^b n_{ij}$.

Mice data (cont)

R uses the μ -as-baseline constraint.

```
lm_out <- lm(Response ~ Diet + Gene + Diet:Gene, data = adipoer)
summary(lm_out)
```

Call:

```
lm(formula = Response ~ Diet + Gene + Diet:Gene, data = adipoer)
```

Residuals:

Min	1Q	Median	3Q	Max
-505.62	-220.09	-11.94	170.32	738.18

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	703.125	75.516	9.311	1.39e-12 ***
DietLFD	-301.383	115.352	-2.613	0.0118 *
GeneWT	93.813	106.795	0.878	0.3838
DietLFD:GeneWT	-9.736	165.238	-0.059	0.9532

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 302.1 on 51 degrees of freedom

Multiple R-squared: 0.2289, Adjusted R-squared: 0.1835

F-statistic: 5.045 on 3 and 51 DF, p-value: 0.003882

Sums of squares in the unbalanced design

Sum of squares	Symbol	Formula
Total	SS_{Tot}	$\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{\dots})^2$
Treatment	SS_{Trt}	$\sum_{i=1}^a \sum_{j=1}^b n_{ij} (\bar{Y}_{ij\cdot} - \bar{Y}_{\dots})^2$
Error	SS_{Error}	$\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{ij\cdot})^2$

We have the decomposition $SS_{\text{Tot}} = SS_{\text{Trt}} + SS_{\text{Error}}$.

The Analysis of Variance (ANOVA) table

Obtain the MS values by dividing the SS values by the Df values.

Source	Df	SS	MS	F value
Treatments	$ab - 1$	SS_{Trt}	MS_{Trt}	$F_{\text{stat}} = MS_{\text{Trt}} / MS_{\text{Error}}$
Error	$N - ab$	SS_{Error}	MS_{Error}	
Total	$N - 1$	SS_{Tot}		

Reject $H_0: \mu_{ij}$ all the same if $F_{\text{stat}} > F_{ab-1, N-ab, \alpha}$.

In the unbalanced case we have $N = \sum_{i=1}^a \sum_{j=1}^b n_{ij}$ in place of abn .

Main effect sums of squares under unbalancedness

- ▶ In the balanced case, we decomposed $SS_{T_{rt}}$ as $SS_A + SS_B + SS_{AB}$.
- ▶ No such “clean” decomposition of $SS_{T_{rt}}$ under unbalancedness.
- ▶ Specifically, it is not obvious how to define SS_A and SS_B .
- ▶ Three versions of SS_A and SS_B measure main effect importance:
 1. Sequentially
 2. After other main effect in the *absence* of interaction
 3. After other main effect in the *presence* of interaction

These are called Type I, Type II, and Type III sums of squares.

Type I SS (meaningless for unbalanced designs)

```
anova(lm(Response ~ Gene + Diet + Gene:Diet, data = adipoeer))
```

Analysis of Variance Table

Response: Response

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Gene	1	127271	127271	1.3949	0.2430621
Diet	1	1253492	1253492	13.7381	0.0005184 ***
Gene:Diet	1	317	317	0.0035	0.9532452
Residuals	51	4653342	91242		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
anova(lm(Response ~ Diet + Gene + Diet:Gene, data = adipoeer))
```

Analysis of Variance Table

Response: Response

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Diet	1	1270105	1270105	13.9202	0.0004804 ***
Gene	1	110659	110659	1.2128	0.2759481
Diet:Gene	1	317	317	0.0035	0.9532452
Residuals	51	4653342	91242		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type II SS (assumes no interaction)

```
# fit model with both main effects
lm_Diet_Gene <- lm(Response ~ Diet + Gene, data = adipoer)
SSE_Diet_Gene <- sum(lm_Diet_Gene$residuals^2)

# fit model with just Gene
lm_Gene <- lm(Response ~ Gene, data = adipoer)
SSE_Gene <- sum(lm_Gene$residuals^2)
SS_Diet <- SSE_Gene - SSE_Diet_Gene # take difference in SSE
SS_Diet
```

```
[1] 1253492
```

```
# fit model with just Diet
lm_Diet <- lm(Response ~ Diet, data = adipoer)
SSE_Diet <- sum(lm_Diet$residuals^2)
SS_Gene <- SSE_Diet - SSE_Diet_Gene # take difference in SSE
SS_Gene
```

```
[1] 110658.6
```

Obtain Type II SS with `Anova()` function from R package `car`.

```
library(car)
Anova(lm(Response ~ Diet + Gene + Diet:Gene, data = adipoer), type = "II")
```

Anova Table (Type II tests)

Response: Response

	Sum Sq	Df	F value	Pr(>F)
Diet	1253492	1	13.7381	0.0005184 ***
Gene	110659	1	1.2128	0.2759481
Diet:Gene	317	1	0.0035	0.9532452
Residuals	4653342	51		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type III SS (Use this one for unbalanced designs)

- ▶ In the unbalanced case we should use the so-called Type III SS.
- ▶ We will denote them by SS_A^{III} and SS_B^{III} .
- ▶ Type III SS are tedious to obtain in the unbalanced case; they are like full-reduced model differences in the error sum of squares, where the reduced model is the model without the main effect in question.
- ▶ The `anova()` or `aov()` functions in R give sequential SS, which are not meaningful in the unbalanced case; however, the Type III SS are equal to the sequential (i.e. Type I) SS when $n_{ij} = n \forall ij$.
- ▶ Under unbalancedness $SS_{Trt} \neq SS_A^{III} + SS_B^{III} + SS_{AB}$.

Type III sums of squares for unbalanced designs

ANOVA table with Type III SS and MS for A and B:

Source	Df	SS	MS	F value
A	$a - 1$	SS_A^{III}	MS_A^{III}	$F_A = MS_A^{III} / MS_{Error}$
B	$b - 1$	SS_B^{III}	MS_B^{III}	$F_B = MS_B^{III} / MS_{Error}$
AB	$(a - 1)(b - 1)$	SS_{AB}	MS_{AB}	$F_{AB} = MS_{AB} / MS_{Error}$
Error	$N - ab$	SS_{Error}	MS_{Error}	
Total	$N - 1$	SS_{Tot}		

1. Reject $H_0: \bar{\mu}_{.1} = \dots = \bar{\mu}_{.a}$ if $F_A > F_{a-1, N-ab, \alpha}$.
2. Reject $H_0: \bar{\mu}_{.1} = \dots = \bar{\mu}_{.b}$ if $F_B > F_{b-1, N-ab, \alpha}$.
3. Reject $H_0: \mu_{ij} = \bar{\mu}_{.i} + \bar{\mu}_{.j} - \bar{\mu}_{..} \quad \forall ij$ if $F_{AB} > F_{(a-1)(b-1), N-ab, \alpha}$.

Tensile strength data (cont)

Obtain Type III SS with `Anova()` from the R package `car`.

Important to specify all the options exactly!

```
library(car)
Anova(lm(Response ~ Gene + Diet + Gene:Diet,
         contrasts = list(Gene = contr.sum, Diet = contr.sum), # specify constraint
         data = adipoer),
      type = "III")
```

Anova Table (Type III tests)

Response: Response

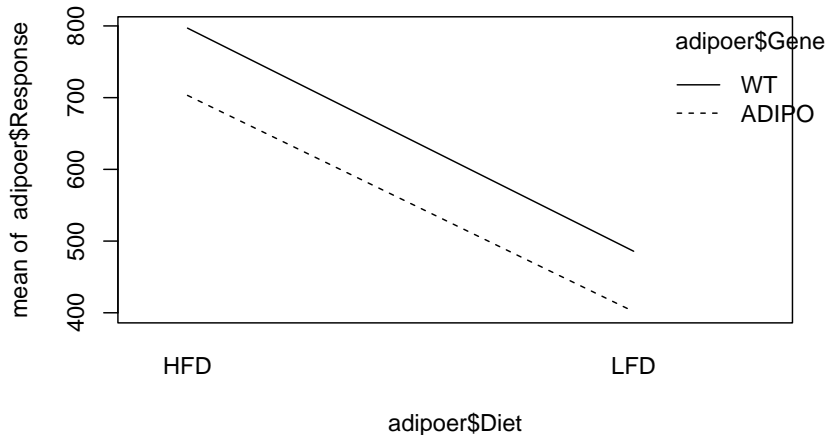
	Sum Sq	Df	F value	Pr(>F)
(Intercept)	19050576	1	208.7917	< 2.2e-16 ***
Gene	105749	1	1.1590	0.2867411
Diet	1253698	1	13.7404	0.0005179 ***
Gene:Diet	317	1	0.0035	0.9532452
Residuals	4653342	51		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Since interaction is not significant, we can focus on main effects...

Mice data (cont)

```
interaction.plot(adipoer$Diet, adipoer$Gene, adipoer$Response)
```



Estimates of cell and marginal means in unbalanced case

The estimators of the cell and marginal means are given by

$$\blacktriangleright \hat{\mu}_{ij} = \bar{Y}_{ij}, \quad i = 1, \dots, a, \quad j = 1, \dots, b.$$

$$\blacktriangleright \hat{\mu}_{i.} = \frac{1}{b} \sum_{j=1}^b \hat{\mu}_{ij}, \quad i = 1, \dots, a.$$

$$\blacktriangleright \hat{\mu}_{.j} = \frac{1}{a} \sum_{i=1}^a \hat{\mu}_{ij}, \quad j = 1, \dots, b.$$

We estimate $\hat{\mu}_{i.}$ with $\bar{Y}_{i.}$ (and $\hat{\mu}_{.j}$ with $\bar{Y}_{.j}$.) only when $n_{ij} = n \quad \forall ij$.

Exercise: Write $\hat{\mu}_{i.}$ in terms of Y_{ijk} and find $\text{Var} \hat{\mu}_{i.}$.

Some CI formulas (without familywise adjustment)

Under unbalancedness our CI formulas become more complicated.

Target	$(1 - \alpha)100\%$ confidence interval
μ_{ij}	$\hat{\mu}_{ij} \pm t_{N-ab, \alpha/2} \hat{\sigma} \sqrt{\frac{1}{n_{ij}}}$
$\mu_{ij} - \mu_{i'j'}$	$\hat{\mu}_{ij} - \hat{\mu}_{i'j'} \pm t_{N-ab, \alpha/2} \hat{\sigma} \sqrt{\frac{1}{n_{ij}} + \frac{1}{n_{i'j'}}}$
$\bar{\mu}_{i.}$	$\hat{\bar{\mu}}_{i.} \pm t_{N-ab, \alpha/2} \hat{\sigma} \sqrt{\frac{1}{b^2} \sum_{j=1}^b \frac{1}{n_{ij}}}$
$\bar{\mu}_{.j}$	$\hat{\bar{\mu}}_{.j} \pm t_{N-ab, \alpha/2} \hat{\sigma} \sqrt{\frac{1}{a^2} \sum_{i=1}^a \frac{1}{n_{ij}}}$
$\bar{\mu}_{i.} - \bar{\mu}_{i'.}$	$\hat{\bar{\mu}}_{i.} - \hat{\bar{\mu}}_{i'.} \pm t_{N-ab, \alpha/2} \hat{\sigma} \sqrt{\frac{1}{b^2} \sum_{j=1}^b \left(\frac{1}{n_{ij}} + \frac{1}{n_{i'j}} \right)}$
$\bar{\mu}_{.j} - \bar{\mu}_{.j'}$	$\hat{\bar{\mu}}_{.j} - \hat{\bar{\mu}}_{.j'} \pm t_{N-ab, \alpha/2} \hat{\sigma} \sqrt{\frac{1}{a^2} \sum_{i=1}^a \left(\frac{1}{n_{ij}} + \frac{1}{n_{i'j}} \right)}$

In the above $\hat{\sigma} = \sqrt{\text{MS}_{\text{Error}}}$.

Mice data (cont)

The interaction appears to be negligible, so we can focus on main effects.

Build 95% CIs for the differences in marginal means for

- ▶ Diet: $\hat{\mu}_{1.} - \hat{\mu}_{2.} \pm t_{55-4, 0.05/2} \hat{\sigma} \sqrt{\frac{1}{2^2} \sum_{j=1}^2 \left(\frac{1}{n_{1j}} + \frac{1}{n_{2j}} \right)}$
- ▶ Gene: $\hat{\mu}_{.1} - \hat{\mu}_{.2} \pm t_{55-4, 0.05/2} \hat{\sigma} \sqrt{\frac{1}{2^2} \sum_{i=1}^2 \left(\frac{1}{n_{i1}} + \frac{1}{n_{i2}} \right)}$

where $\hat{\sigma} = \sqrt{\text{MS}_{\text{Error}}}$.

```

mu_hat <- grp_means$Response
nn <- grp_counts$Response
mu11_hat <- mu_hat[1] ; n11 <- nn[1] # HFD x ADIPO
mu21_hat <- mu_hat[2] ; n21 <- nn[2] # LFD x ADIPO
mu12_hat <- mu_hat[3] ; n12 <- nn[3] # HFD x WT
mu22_hat <- mu_hat[4] ; n22 <- nn[4] # LFD x WT

mu1.hat <- (mu11_hat + mu12_hat)/2 # HFD mean
mu2.hat <- (mu21_hat + mu22_hat)/2 # LFD mean
mu.1hat <- (mu11_hat + mu21_hat)/2 # ADIPO mean
mu.2hat <- (mu12_hat + mu22_hat)/2 # WT mean

N <- sum(nn)
a <- 2
b <- 2
MSE <- sum(lm_out$residuals^2)/(N - a*b)
alpha <- 0.05

me <- qt(1-alpha/2,N-a*b)*sqrt(MSE)*sqrt(1/2^2*(1/n11 + 1/n12 + 1/n11 + 1/n22))

ci_diet <- c(mu1.hat - mu2.hat - me, mu1.hat - mu2.hat + me)
ci_gene <- c(mu.1hat - mu.2hat - me, mu.1hat - mu.2hat + me)

```

We obtain the intervals:

- ▶ Diet: (146.27, 466.24)
- ▶ Gene: (-248.93, 71.04)

Unbalanced tensile strength data from Kuehl (2000)

Table 6.19 Tensile strength (psi) of asphaltic concrete specimens for two aggregate types with each of three kneading compaction methods

Aggregate Type	Compaction Method			Aggregate Means ($\bar{y}_{i..}$)
	Kneading			
	Regular	Low	Very Low	
Basalt	106	93	56	
	108	101		
		98		
Means ($\bar{y}_{1j.}$)	107.0	97.3	56	93.7
Silicious	107	63	40	
	110	60	41	
	116		44	
Means ($\bar{y}_{2j.}$)	111.0	61.5	41.7	72.6
Compaction means ($\bar{y}_{.j.}$)	109.4	83.0	45.3	

```
y <- c(106,108,93,101,98,56,107,110,116,63,60,40,41,44)
agg <- as.factor(c("b","b","b","b","b","b","s","s","s","s","s","s","s","s"))
comp <- as.factor(c("r","r","l","l","l","v1","r","r","r","l","l","v1","v1","v1"))
tensile <- data.frame(y = y, agg = agg, comp = comp)
```

Compute group means:

```
aggregate(y ~ agg + comp, data = tensile, mean)
```

	agg	comp	y
1	b	l	97.33333
2	s	l	61.50000
3	b	r	107.00000
4	s	r	111.00000
5	b	v1	56.00000
6	s	v1	41.66667

Unbalanced tensile strength data (cont)

Again the meaninglessness of sequential SS under unbalancedness:

```
anova(lm(y ~ agg + comp + agg:comp, data = tensile))
```

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
agg	1	1518.0	1518.0	135.184	2.726e-06 ***
comp	2	8401.9	4201.0	374.112	1.252e-08 ***
agg:comp	2	953.4	476.7	42.454	5.497e-05 ***
Residuals	8	89.8	11.2		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
anova(lm(y ~ comp + agg + agg:comp, data = tensile))
```

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
comp	2	9159.3	4579.6	407.834	8.899e-09 ***
agg	1	760.7	760.7	67.740	3.557e-05 ***
comp:agg	2	953.4	476.7	42.454	5.497e-05 ***
Residuals	8	89.8	11.2		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tensile strength data (cont)

Obtain Type III SS with `Anova()` from the R package `car`.

```
library(car)
Anova(lm(y ~ agg + comp + agg:comp, data = tensile,
         contrasts = list(agg = contr.sum, comp = contr.sum)),
      type = "III")
```

Anova Table (Type III tests)

Response: y

	Sum Sq	Df	F value	Pr(>F)	
(Intercept)	75050	1	6683.495	5.589e-13	***
agg	710	1	63.269	4.551e-05	***
comp	6806	2	303.070	2.879e-08	***
agg:comp	953	2	42.454	5.497e-05	***
Residuals	90	8			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Now we could make comparisons of means, noting the interaction.

References

Kuehl, R. O. 2000. *Design of Experiments: Statistical Principles of Research Design and Analysis*. Duxbury/Thomson Learning.