19. Two-Factor ANOVA

- Linear model analysis with two categorical predictors

Example data set: How is size of cash offer (in $100's) for a used car related to the age and gender of the seller? Study involving 36 car dealerships.

Age	Gender: Male	Gender: Female
Young	$21, 23, \ldots, 23$	$21, 22, \ldots, 25$
Middle	$30, 29, \ldots, 27$	$26, 29, \ldots, 29$
Elderly	$25, 22, \ldots, 21$	$23, 19, \ldots, 20$

- Equal sample sizes. Six subjects at each combination of factor levels.

Notation

- Factor A index: $i = 1, \ldots, a$
  Factor B index: $j = 1, \ldots, b$

$$
\begin{aligned}
Y_{ijk} &= k\text{th response in } \textit{cell } (i,j) \\
\bar{Y}_{ij\cdot} &= \text{sample mean in cell } (i,j) \\
s_{ij} &= \text{sample standard deviation in cell } (i,j) \\
\mu_{ij} &= \text{expected (mean) response in cell } (i,j) \\
n_{ij} &= \text{sample size in cell } (i,j) \\
&= n, \text{ if equal sample sizes} \\
n_T &= \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij} = abn
\end{aligned}
$$

Model

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk}\,, \qquad \epsilon_{ijk} \sim \text{i.i.d.} N(0, \sigma^2)$$

Factor level means

$$\mu_{i\cdot} = \frac{1}{b}\sum_{j=1}^{b}\mu_{ij} = i\text{th factor } A \text{ level mean}$$

$$\mu_{\cdot j} = \frac{1}{a}\sum_{i=1}^{a}\mu_{ij} = j\text{th factor } B \text{ level mean}$$

$$\mu_{\cdot\cdot} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij} = \text{overall mean}$$

- These averages give equal weight to all levels!

Main effects

- Factor A: deviation of ith level from overall mean

$$\alpha_i = \mu_{i\cdot} - \mu_{\cdot\cdot}, \quad i = 1, \ldots, a$$

- Factor B: deviation of jth level from overall mean

$$\beta_j = \mu_{\cdot j} - \mu_{\cdot\cdot}, \quad j = 1, \ldots, b$$

- Main effects sum to zero

$$\sum_{i=1}^{a}\alpha_i = \sum_{j=1}^{b}\beta_j = 0$$

$$\alpha_a = -\sum_{i=1}^{a-1}\alpha_i \qquad \beta_b = -\sum_{j=1}^{b-1}\beta_j$$

Main effects model

$$\mu_{ij} = \mu_{\cdot\cdot} + \alpha_i + \beta_j$$

- Implies that the effects of factors A and B are additive.
- The effect of changing the level of one factor is the same at all levels of the other factor – *no interaction*!
- e.g. Used car sales: the *gender effect* is the same at all ages.

Matrix formulation: Main effects model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- e.g. n=2, a=3, b=2

$$\begin{bmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ Y_{211} \\ Y_{212} \\ Y_{221} \\ Y_{222} \\ Y_{311} \\ Y_{312} \\ Y_{321} \\ Y_{322} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & -1 \\ 1 & 1 & 0 & -1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & -1 \\ 1 & 0 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \mu_{\cdot\cdot} \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_{111} \\ \epsilon_{112} \\ \epsilon_{121} \\ \epsilon_{122} \\ \epsilon_{211} \\ \epsilon_{212} \\ \epsilon_{221} \\ \epsilon_{222} \\ \epsilon_{311} \\ \epsilon_{312} \\ \epsilon_{321} \\ \epsilon_{322} \end{bmatrix}$$

Interaction model

- Consider the identity

$$\begin{aligned} \mu_{ij} &= \mu_{\cdot\cdot} + (\mu_{i\cdot} - \mu_{\cdot\cdot}) + (\mu_{\cdot j} - \mu_{\cdot\cdot}) \\ &\quad +(\mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu_{\cdot\cdot}) \\ &= \mu_{\cdot\cdot} + \alpha_i + \beta_j + (\alpha\beta)_{ij} \end{aligned}$$

- The hypothesis of no interaction is $H_0 : (\alpha\beta)_{ij} = 0$ for all (i, j).

- Notice that the $(\alpha\beta)$-terms satisfy the sum constraints

$$\sum_{i=1}^{a} (\alpha\beta)_{ij} = \sum_{j=1}^{b} (\alpha\beta)_{ij} = \sum_{i=1}^{a}\sum_{j=1}^{b} (\alpha\beta)_{ij} = 0$$

- e.g. if $n = 2$, $a = 3$ and $b = 2$ then there are only two *free* interation parameters:

$$(\alpha\beta)_{11} \text{ and } (\alpha\beta)_{21}$$

Matrix formulation: Interaction model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- e.g. n=2, a=3, b=2

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & -1 & -1 & 0 \\ 1 & 1 & 0 & -1 & -1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & -1 & 0 & -1 \\ 1 & 0 & 1 & -1 & 0 & -1 \\ 1 & -1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \mu_{..} \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ (\alpha\beta)_{11} \\ (\alpha\beta)_{21} \end{bmatrix}$$

Regression Model Formulation

e.g. a=3, b=2

$$X_1 = \begin{cases} 1 & \text{level 1, factor A} \\ 0 & \text{level 2, factor A} \\ -1 & \text{level 3, factor A} \end{cases}$$

$$X_2 = \begin{cases} 0 & \text{level 1, factor A} \\ 1 & \text{level 2, factor A} \\ -1 & \text{level 3, factor A} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{level 1, factor B} \\ -1 & \text{level 2, factor B} \end{cases}$$

$$\begin{aligned} E(Y) &= \mu_{.} + \alpha_1 X_1 + \alpha_2 X_2 + \beta_1 X_3 \\ &\quad + (\alpha\beta)_{11} X_1 X_3 + (\alpha\beta)_{21} X_2 X_3 \end{aligned}$$

- The hypothesis of *no interation* is

$$H_0 : (\alpha\beta)_{11} = (\alpha\beta)_{21} = 0$$

Least Squares Estimation

$$Q = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \mu_{ij})^2$$

- Minimizing with respect to the cell means results in

$$\hat{\mu}_{ij} = \bar{Y}_{ij}$$

$$\begin{aligned} \mu_{..} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij} \quad &\Longrightarrow \quad \hat{\mu}_{..} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\bar{Y}_{ij.} = \bar{Y}_{...} \\ \alpha_i = \mu_{i.} - \mu_{..} \quad &\Longrightarrow \quad \hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...} \\ \beta_i = \mu_{.j} - \mu_{..} \quad &\Longrightarrow \quad \hat{\beta}_j = \bar{Y}_{.j.} - \bar{Y}_{...} \end{aligned}$$

and

$$(\alpha\beta)_{ij} = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..}$$

implies

$$(\widehat{\alpha\beta})_{ij} = \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}$$

Two-Way ANOVA

- As for one-way ANOVA with ab “treatment” combinations, the total SS partitions into treatment SS and error SS

$$SSTO = SSTR + SSE$$

$$\begin{aligned} \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \bar{Y}_{...})^2 \quad = \quad & \sum_{i=1}^{a}\sum_{j=1}^{b} n(\bar{Y}_{ij.} - \bar{Y}_{...})^2 \\ & + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \bar{Y}_{ij.})^2 \end{aligned}$$

- The degrees of freedom also partition as in one-way ANOVA

$$n_T - 1 = (ab - 1) + (abn - ab)$$

- Partitioning SSTR

$$\begin{array}{cc}
\text{SSTR} & n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{Y}_{ij\cdot}-\bar{Y}_{\cdots})^2 \\
= & = \\
\text{SSA} & bn\sum_{i=1}^{a}(\bar{Y}_{i\cdot\cdot}-\bar{Y}_{\cdots})^2 \\
+ & + \\
\text{SSB} & an\sum_{j=1}^{b}(\bar{Y}_{\cdot j\cdot}-\bar{Y}_{\cdots})^2 \\
+ & + \\
\text{SSAB} & n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{Y}_{ij\cdot}-\bar{Y}_{i\cdot\cdot}-\bar{Y}_{\cdot j\cdot}+\bar{Y}_{\cdots})^2
\end{array}$$

Two-Way ANOVA Table:

Source of variation	Sum of Squares (SS)	df	Mean Square (MS)
Factor A	SSA	$a-1$	$\frac{SSA}{a-1}$
Factor B	SSB	$b-1$	$\frac{SSB}{b-1}$
Interaction	$SSAB$	$(a-1)(b-1)$	$\frac{SSAB}{(a-1)(b-1)}$
Error	SSE	$ab(n-1)$	$\frac{SSE}{ab(n-1)}$
Total	$SSTO$	$abn-1$	

F-tests

- *No interation*: $H_0 : (\alpha\beta)_{ij} = 0$ for all i, j.

$$F^* = \frac{MSAB}{MSE} \sim F((a-1)(b-1), ab(n-1))$$

- *No main effect of Factor A*: $H_0 : \alpha_i = 0$ for all i.

$$F^* = \frac{MSA}{MSE} \sim F(a-1, ab(n-1))$$

- *No main effect of Factor B*: $H_0 : \beta_j = 0$ for all j.

$$F^* = \frac{MSB}{MSE} \sim F(b-1, ab(n-1))$$

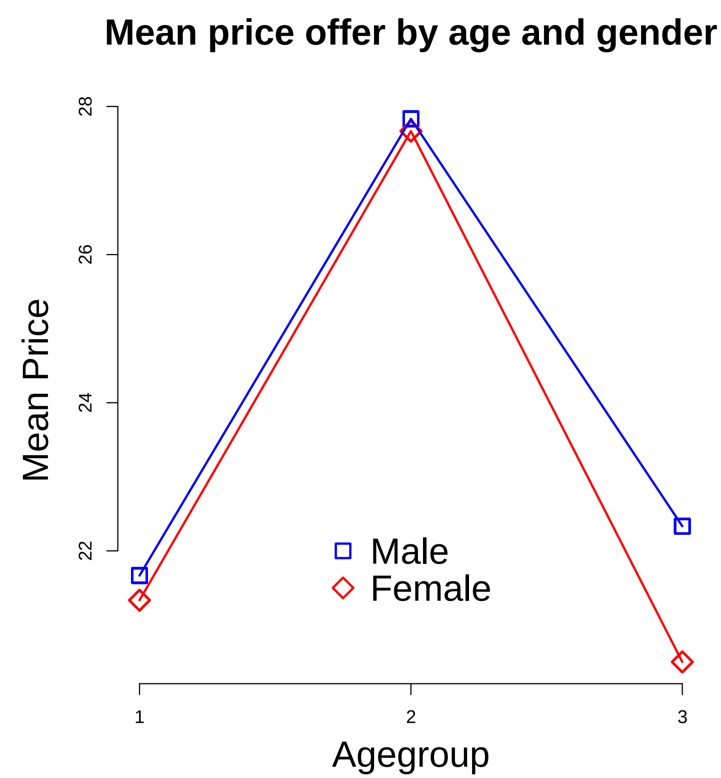
- If there is significant interaction there it doesn't make sense to test the main effects!

- If the interaction is not significant, consider pooling $SSAB$ with SSE!

Used Car Data

```
> car.df <- read.table("data/usedcar.dat")
   price agegroup gender time
1     21        1      1    1
2     23        1      1    2
3     19        1      1    3

35    20        3      2    5
36    20        3      2    6
> colnames(car.df) <- c("price","agegroup","gender","time")
> attach(car.df)
> age <- as.factor(agegroup)
> levels(age) <- c("Young","Middle","Elderly")
> sex <- as.factor(gender)
> levels(sex) <- c("Male","Female")
> fit.i <- lm(price~age*sex)
> anova(fit.i)
Analysis of Variance Table
Response: price
          Df Sum Sq Mean Sq F value   Pr(>F)
age        2 316.72  158.36 66.2907 9.79e-12 ***
sex        1   5.44    5.44  2.2791   0.1416
age:sex    2   5.06    2.53  1.0581   0.3597
Residuals 30  71.67    2.39
```

Linear contrasts

- Let $\bar{Y}_i$ be the mean of a sample of size n_i from a $N(\mu_i, \sigma^2)$ population, $i = 1, \ldots, r$, and suppose the samples are independent.

- A linear contrast among the group means has the form

$$C = \sum_{i=1}^{r} c_i \mu_i \quad \text{where} \quad \sum_{i=1}^{r} c_i = 0$$

- An estimate of C is

$$\hat{C} = \sum_{i=1}^{r} c_i \bar{Y}_i$$

- The expected value and variance of $\hat{C}$ is

$$E(\hat{C}) = \sum_{i=1}^{r} c_i \mu_i \quad \text{where} \quad \text{var}(\hat{C}) = \sigma^2 \sum_{i=1}^{r} \frac{c_i^2}{n_i}$$

Used Car Data

- Pairwise comparisons

```
> agens <- sapply(split(price,age),length)
> agems <- sapply(split(price,age),mean)
> agesds <- sapply(split(price,age),sd)
> summary.df <- signif(data.frame(agens,agems,agesds),4)
> colnames(summary.df) <- c("n","mean","sd")
> summary.df
         n  mean    sd
Young   12 21.50 1.732
Middle  12 27.75 1.288
Elderly 12 21.42 1.676
> mse <- sum(fit.i$residuals^2)/fit.i$df.residual
> mse
[1] 2.388889
# Least significant difference for pairwise comparisons
> lsd <- qt(.975,fit.i$df.residual)*sqrt(mse*(1/12+1/12))
> lsd
[1] 1.288653
```