Analysis of Variance

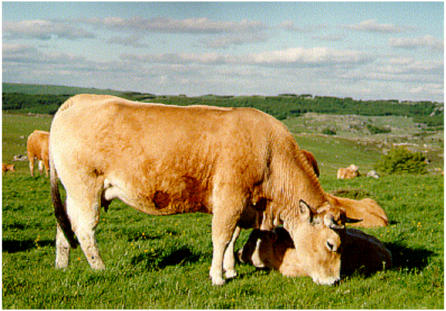
M2 Gen2E

E. Marchadier, C. Dillmann

november 2024

Biological question

Problem : Should a farmer from the french Aubrac region raise calves from across between Aubrac x Charollais breeds rather than pure-breed Aubrac calves ?



Biological question

Problem : Should a farmer from the french Aubrac region raise calves from across between Aubrac x Charollais breeds rather than pure-breed Aubrac calves ?

Data

Problem : Should a farmer from the french Aubrac region raise calves from across between Aubrac x Charollais breeds rather than pure-breed Aubrac calves ?

Data : Average Daily Weight Gain (ADWG) (grams/day) have been measured from newborn calves after summertime in three different pastures in the Aubrac region. The pastures differ for the elevation (1=plains, 2=low mountain, 3=high mountain)

Breed	Aubrac	Aubrac x Charolais
HM (high mountain)	716	671
	679	640
	707	661
	733	693
	700	680
LM (low mountain)	715	770
	757	780
	734	808
	704	791
	725	756
PL (plains)	770	797
	747	834
	741	780
	756	813
	716	806

Random variable and sources of variation

Y_{ijk} : ADWG from animal k from breed i, in pasture j.

- Pasture = 3 different pastures. $j = 1..J$, $J = 3$.
- Breed = 2 different breeds, A and AxC. $i = 1..I$, $I = 2$.
- Individuals = For each breed and each pasture, five calves are randomly chosen and measured, $k = 1..K$, $K = 5$.

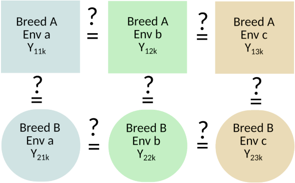
Questions

Y_{ijk} : ADWG from animal k from breed i, in pasture j.

- Are there differences between pastures ?
- Are there differences between breeds ?

ANOVA2 model

We want to separate the effects of the breed and the one of the pasture



Breed A Env a Y_{11k} $\overset{?}{=}$ Breed A Env b Y_{12k} $\overset{?}{=}$ Breed A Env c Y_{13k}

$\overset{?}{=}$ $\overset{?}{=}$ $\overset{?}{=}$

Breed B Env a Y_{21k} $\overset{?}{=}$ Breed B Env b Y_{22k} $\overset{?}{=}$ Breed B Env c Y_{23k}

Regular Model

$$Y_{ijk} \approx \mathcal{N}(\mu_{ij}, \sigma)$$

Singular Model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$

with $\epsilon_{ijk} \approx \mathcal{N}(0, \sigma)$ AND $\sum_i \alpha_i = 0$ AND $\sum_j \beta_j = 0$

Are there significant differences beteewen breeds OR between environments ?,

Summary Statistics

Model : $Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$

Race	A	AxC	
HM	$Y_{11.}$	$Y_{21.}$	$Y_{.1.}$
LM	$Y_{12.}$	$Y_{22.}$	$Y_{.2.}$
PL	$Y_{13.}$	$Y_{23.}$	$Y_{.3.}$
	$Y_{1..}$	$Y_{2..}$	$Y_{...}$

- Mean of breed i in pasture j : $Y_{ij.} = \frac{1}{K}\sum_{k=1}^{K} Y_{ijk}.$
- Mean of breed i : $Y_{i..} = \frac{1}{J}\sum_{j=1}^{J} Y_{ij.}.$
- Mean of pasture j : $Y_{.j.} = \frac{1}{I}\sum_{i=1}^{I} Y_{ij.}.$
- General mean : $Y_{...} = \frac{1}{IJ}\sum_{j=1}^{J}\sum_{i=1}^{I} Y_{ij.}.$

Parameter's estimators

Modèle : $Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$

Race	A	AxC	
HM	$Y_{11.}$	$Y_{21.}$	$Y_{.1.}$
LM	$Y_{12.}$	$Y_{22.}$	$Y_{.2.}$
PL	$Y_{13.}$	$Y_{23.}$	$Y_{.3.}$
	$Y_{1..}$	$Y_{2..}$	$Y_{...}$

- Breed i effect : $\hat{\alpha}_i = Y_{i..} - Y_{...}$
- Pasture j effect : $\hat{\beta}_j = Y_{.j.} - Y_{...}$
- General mean : $\hat{\mu} = Y_{...}$

Model's parameters can be estimated from summary statistics $Y_{ij.}$, $Y_{i..}$, $Y_{.j.}$ and $Y_{...}$.

Parameter versus Estimator

Model : $Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$

What is the difference between α_i et $\hat{\alpha}_i$?

$$\hat{\alpha}_i = \alpha_i + \epsilon_{i..} - \epsilon_{...}$$

$$\hat{\beta}_j = \beta_j + \epsilon_{.j.} - \epsilon_{...}$$

Estimators are random variable. Innately, they contain part of the residual variation. Were all $\alpha_i = 0$, $\hat{\alpha}_i$ have a null expectation, but a POSITIVE variance, that depend on the residual variance σ^2.

Parameter's estimation

Model : $Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$

Estimator

- Breed i effect : $\hat{\alpha}_i = Y_{i..} - Y_{...}$.
- Pasture j effect : $\hat{\beta}_j = Y_{.j.} - Y_{...}$.
- General mean : $\hat{\mu} = Y_{...}$.

Estimation

- $\hat{\alpha}_{1_{obs}} = -12.7$, $\hat{\alpha}_{2_{obs}} = +12.7$.
- $\hat{\beta}_{1_{obs}} = +36.7$, $\hat{\beta}_{2\,obs} = +14.7$, $\hat{\beta}_{3_{obs}} = -51.4$.
- $\hat{\mu}_{obs} = 739.3$.

Are those differences significant ?

Building-up the test statistics

Model : $Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$

$$SCT = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - Y_{...})^2$$

$$SCA = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{i..} - Y_{...})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\alpha_i + \epsilon_{i..} - \epsilon_{...})^2$$

$$SCB = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{.j.} - Y_{...})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\beta_j + \epsilon_{.j.} - \epsilon_{...})^2$$

$$SCR = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - Y_{i..} - Y_{.j.} + Y_{...})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\epsilon_{ijk} - \epsilon_{.j.} - \epsilon_{.j.} + \epsilon_{...})^2$$

$$SCT = SCA + SCB + SCR$$

ANOVA2. Test stastitics and probability law under H0

Model : $Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$

Source of variation	H0	probability law under H0	
Breed	$\alpha_1 = \alpha_2 = 0$	$\frac{\frac{SCA}{I-1}}{\sigma^2} \approx \chi^2_{I-1}$	$\frac{\frac{SCA}{I-1}}{\frac{SCR}{IJK-I-J+1}} \approx \mathcal{F}^{I-1}_{IJK-I-J+1}$
Pasture	$\beta_1 = \beta_2 = \beta_3 = 0$	$\frac{\frac{SCB}{J-1}}{\sigma^2} \approx \chi^2_{J-1}$	$\frac{\frac{SCB}{J-1}}{\frac{SCR}{IJK-I-J+1}} \approx \mathcal{F}^{I-1}_{IJK-I-J+1}$
Residual		$\frac{\frac{SCR}{IJK-I-J+1}}{\sigma^2} \approx \chi^2_{IJK-I-J+1}$	

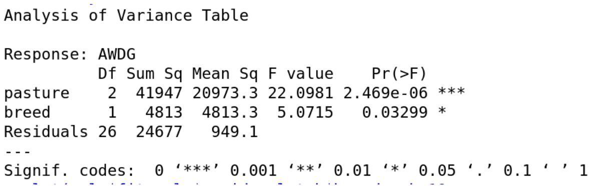
ANOVA2 : additive model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$
$$H0 : \alpha_1 = ... = \alpha_I = 0$$
$$H0' : \beta_1 = ... = \beta_J = 0$$

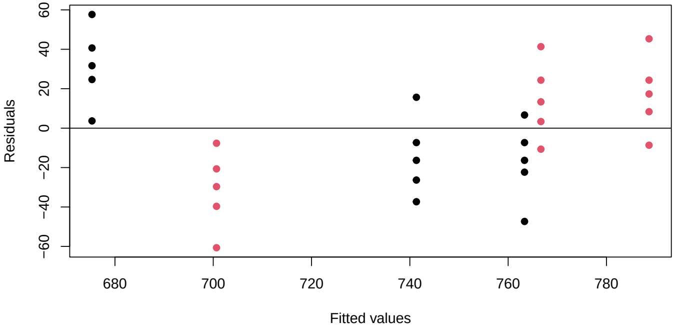
rscripts/rscriptcalf.R

```
## Read the datafile
tab <- read.table("cattle_data.csv",sep=";",header=TRUE)
## Define the categorical variables
tab$breed <- as.factor(tab$breed)
tab$pasture <- as.factor(tab$pasture)
summary(tab)

## anova : additive model
mylm <- lm(AWDG~pasture+breed,data=tab)
anova(mylm)
plot(mylm$fit,mylm$resid,col=tab$breed,pch=19,
     xlab="Fitted values",ylab="Residuals")
abline(h=0)


boxplot(AWDG~pasture+breed,data=tab,las=3,xlab="")
```

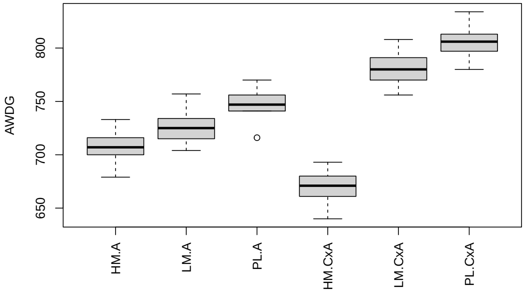
6. Compute the observed values



```
Analysis of Variance Table

Response: AWDG
          Df Sum Sq Mean Sq F value    Pr(>F)    
pasture    2  41947 20973.3 22.0981 2.469e-06 ***
breed      1   4813  4813.3  5.0715   0.03299 *  
Residuals 26  24677   949.1                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



7. Interpret the results



Pasture effects are not constant over Breeds ⟶ genotype by environment interactions ! ! !

ANOVA2 : interactions

Y_{ijk} : ADWG from animal k, breed i, pasture j.

Model :

$$y_{ijk} = \mu + \alpha_i + \beta_j + \theta_{ij} + \epsilon_{ijk}$$

Hypotheses :

ϵ_{ijk} are random variables. They are idependent and identically distributed $iid\mathcal{N}(0, \sigma^2)$.

Beware the experimental design !

Constraints :

$$\sum_{i=1}^{I} \alpha_i = 0,\ \sum_{j=1}^{J} \beta_j = 0,\ \sum_i \theta_{ij} = \sum_j \theta_{ij} = 0.$$

Parameer's estimators

Modèle : $Y_{ijk} = \mu + \alpha_i + \beta_j + \theta_{ij} + \epsilon_{ijk}$

Race	A	AxC	
HM	$Y_{11.}$	$Y_{21.}$	$Y_{.1.}$
LM	$Y_{12.}$	$Y_{22.}$	$Y_{.2.}$
PL	$Y_{13.}$	$Y_{23.}$	$Y_{.3.}$
	$Y_{1..}$	$Y_{2..}$	$Y_{...}$

- Interaction effect : $\hat{\theta}_{ij} = Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...}$
- Breed i effect : $\hat{\alpha}_i = Y_{i..} - Y_{...}$
- Pasture j effect : $\hat{\beta}_j = Y_{.j.} - Y_{...}$
- General mean : $\hat{\mu} = Y_{...}$

Model's parameters can be estimated from summary statistics $Y_{ij.}$, $Y_{i..}$, $Y_{.j.}$ and $Y_{...}$.

ANOVA2 with interactions : buildng-up the test statistics

Modèle : $Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$

$$SCT = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - Y_{...})^2$$

$$SCA = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{i..} - Y_{...})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\alpha_i + \epsilon_{i..} - \epsilon_{...})^2$$

$$SCB = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{.j.} - Y_{...})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\beta_j + \epsilon_{.j.} - \epsilon_{...})^2$$

$$SCI = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ij.} - Y_{i..} - Y_{.j.} + Y_{...})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\theta_{ij} + \epsilon_{ij.} - \epsilon_{.j.} - \epsilon_{.j.} + \epsilon_{...})^2$$

$$SCR = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - Y_{ij.})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\epsilon_{ijk} - \epsilon_{ij.})^2$$

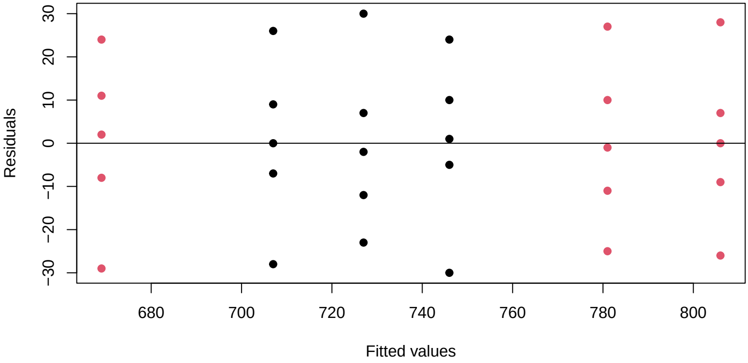
$$SCT = SCA + SCB + SCI + SCR$$

Two-factors anova with interactions

$$H0 : \alpha_1 = ... = \alpha_I = 0$$
$$H0' : \beta_1 = ... = \beta_J = 0$$
$$H0'' : \theta_{11} = ... = \theta_{IJ} = 0$$

Conclusions : Statistics

```
Analysis of Variance Table

Response: AWDG
             Df Sum Sq Mean Sq F value    Pr(>F)    
pasture       2  41947 20973.3  52.488 1.724e-09 ***
breed         1   4813  4813.3  12.046  0.001981 ** 
pasture:breed  2  15087  7543.3  18.878 1.187e-05 ***
Residuals    24   9590   399.6                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

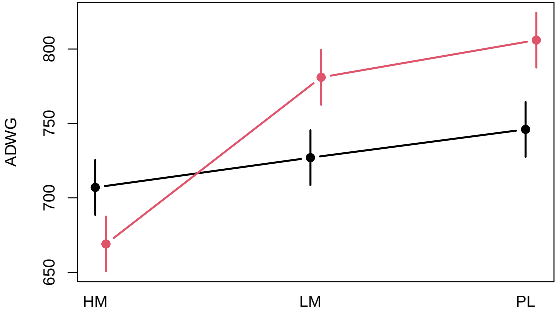


Conclusions : Biology

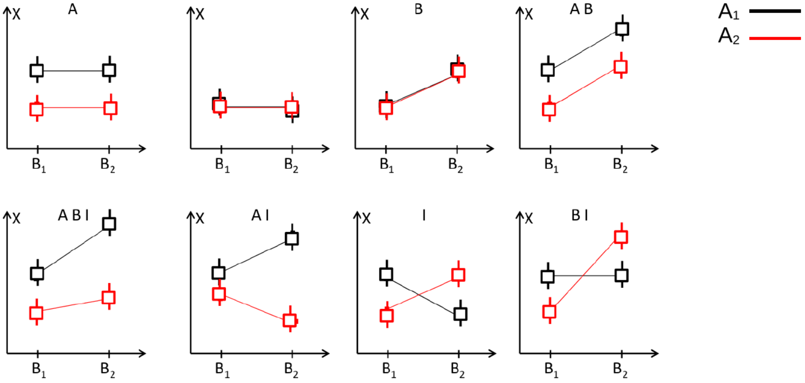
Because of significant interactions, breed do not behave the same according to pastures

Conclusions : Biology

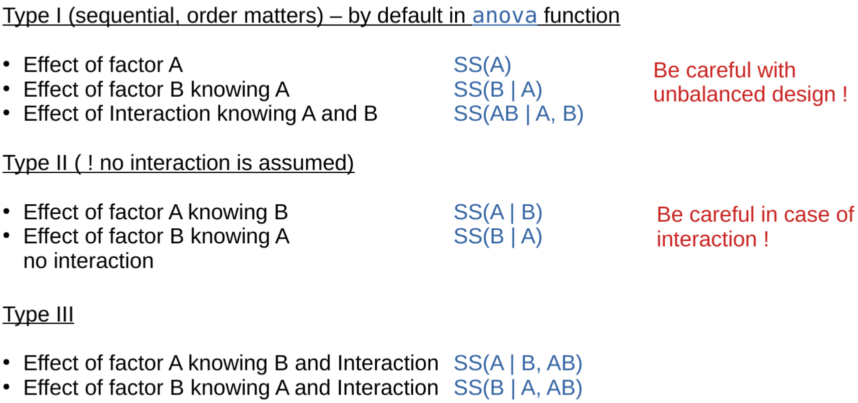
Because of significant interactions, breed do not behave the same according to pastures



Interactions : effect plots



Unbalanced designs



Type I (sequential, order matters) – by default in anova function

Effect of factor A	SS(A)
Effect of factor B knowing A	SS(B | A)
Effect of Interaction knowing A and B	SS(AB | A, B)

Be careful with unbalanced design !

Type II (! no interaction is assumed)

Effect of factor A knowing B	SS(A | B)
Effect of factor B knowing A no interaction	SS(B | A)

Be careful in case of interaction !

Type III

Effect of factor A knowing B and Interaction	SS(A | B, AB)
Effect of factor B knowing A and Interaction	SS(B | A, AB)

Contrasts : Means comparisons