Modèles linéaires

Analyses de covariance

Denis Laloë
GABI - PSGen

11 octobre 2016

Résumé des épisodes précédents

Modélisation statistique

- une variable réponse Y est modélisée par des combinaisons linéaires de variables explicatives x (ou leur transformées)
- Une partie fixe : $\mathbb{E}(Y) = \sum \beta_i x_i$
- un résidu aléatoire : $E \sim N(0, \sigma)$

y réalisation d'une variable aléatoire Y

- Estimation des β
- L'estimateur de β est une v.a., dont on a une réalisation, l'estimation.
- Tests d'hypothèse, IC
- Prédiction

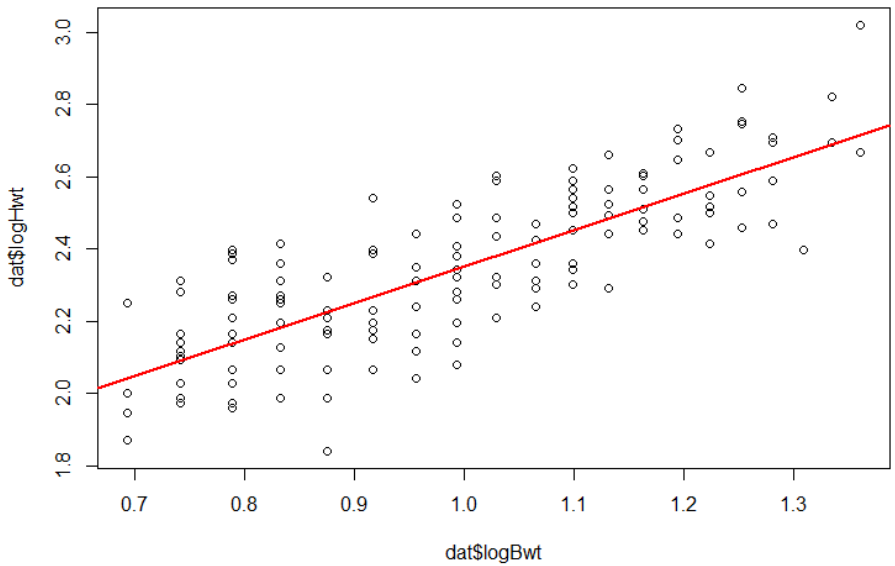
Nature des variables explicatives

- Quantitative : régression
- Qualitatif : analyse de variance
- Quantitatif+Qualitatif : analyse de covariance

Modèle linéaire et type de variables

Variables quantitatives

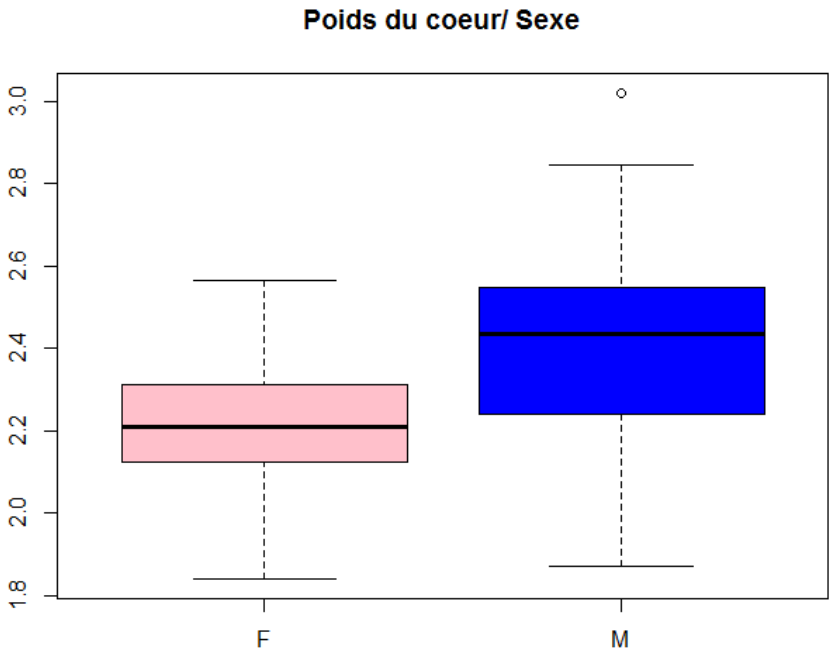
- r=0,80
- Modèle de régression $r^2 = R^2$



Modèle linéaire et type de variables

Variables quantitatives/qualitative

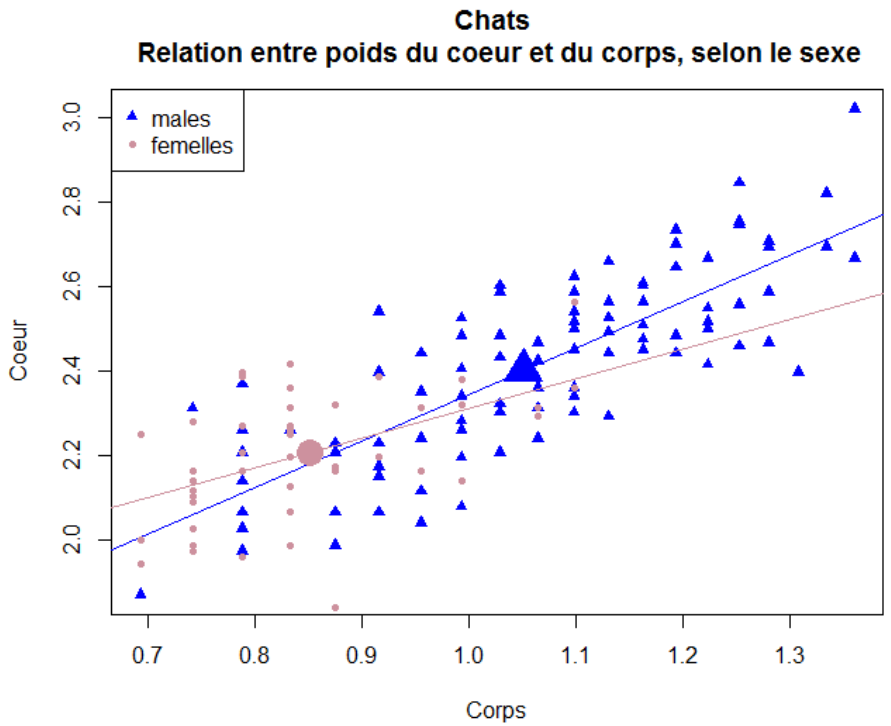
- (Carré du) rapport de corrélation η^2
- Analyse de variance $\eta^2 = R^2$



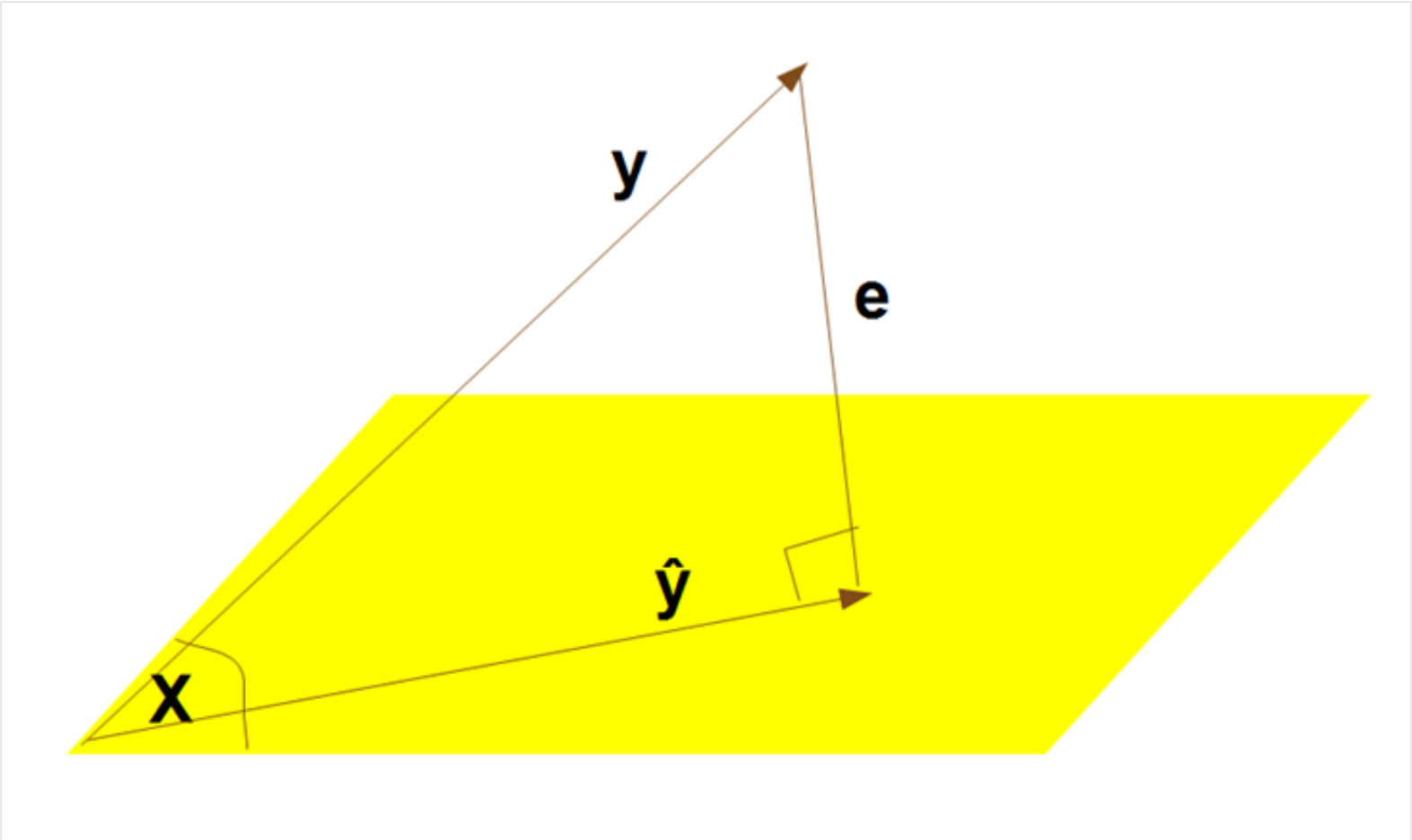
Modèle linéaire et type de variables

Variables quantitatives/quantitative + qualitative

- Analyse de covariance



L'approche géométrique



Le modèle linéaire général

Modèle général

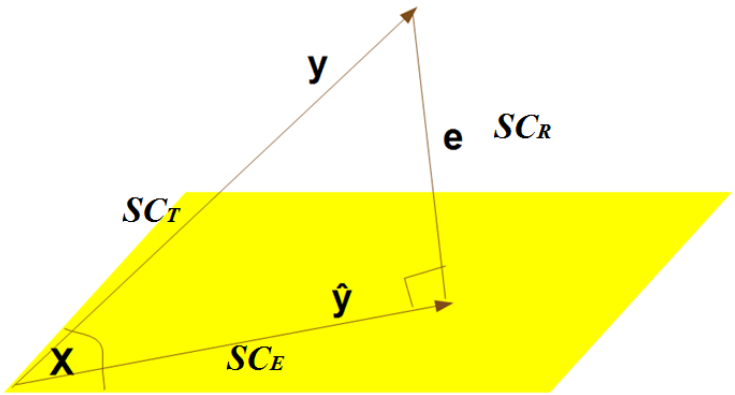
$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{E}$$

$$\mathbf{Y} = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(p)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(p)} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \beta_p \end{bmatrix} + \mathbf{E}$$

Résolution

$$\hat{\beta} = (\mathbf{X^t X})^{-} \mathbf{X^t y}$$

Les tests



Les sommes de carrés

H0 : $\hat{\beta} = 0$

	Somme des carrés	Distribution associée
Modèle	$SC_E = \sum_i (\hat{Y}_i - \bar{Y})^2 = \|\hat{Y} - \bar{Y}\|^2$	$\sigma^2 \chi^2(p-1)$
Résiduelle (R)	$SC_R = \sum_i (Y_i - \hat{Y}_i)^2 = \|Y - \hat{Y}\|^2$	$\sigma^2 \chi^2(n-p)$
Total (T)	$SC_T == \sum_i (Y_i - \bar{Y})^2 = \|Y - \bar{Y}\|^2$	$\sigma^2 \chi^2(n-1)$

Les tests

	Somme des carrés	ddl	Carré moyen	Test
Modèle	$SC_E = \|\|\hat{Y} - \bar{Y}\|\|^2$	$p-1$	$CM_E = \frac{SC_E}{p-1}$	$\frac{CM_E}{CM_R}$
Résiduelle (R)	$SC_R = \|\|Y - \hat{Y}\|\|^2$	$n-p$	$CM_R = \frac{SC_R}{n-p}$	
Total (T)	$SC_T = \|\|Y - \bar{Y}\|\|^2$	$n-1$		

Sous H0, $\frac{CM_E}{CM_R} \sim \mathbb{F}(p-1, n-p)$

Le modèle d'analyse de variance

Modèle général

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{E}$$

$$\mathbf{Y} = \begin{bmatrix} 1 & \delta_1^{(1)} & \delta_1^{(2)} & \dots & \delta_1^{(p)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \delta_n^{(1)} & \delta_n^{(2)} & \dots & \delta_n^{(p)} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \beta_p \end{bmatrix} + \mathbf{E}$$

$\delta_i^{(j)} = 1$ si l'observation i présente la modalité j, 0 sinon

Le modèle d'analyse de variance

Un exemple. Modèle à un facteur et trois modalités

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{E}$$

$$\mathbf{Y} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ : & : & : & \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_1 \\ \hat{\beta}_3 \end{bmatrix} + \mathbf{E}$$

- La somme des colonnes des modalités d'un facteur est égale à 1.
- Il y a des dépendances entre colonnes: seules *p-1* = 2 colonnes sont "informatives".
- $\mathbf{X}$ n'est pas de plein rang, $\mathbf{X^tX}$ n'est pas inversible
- on ne peut estimer que *p-1* =2 paramètres
- Système de contraintes / contrastes

Résolution

$$\hat{\beta} = (\mathbf{X^t X})^{-} \mathbf{X^t y}$$

Signification des matrices

- $\mathbf{X^t X}$ est une matrice d'effectifs
- $\mathbf{X^t Y}$ est une matrice de sommes de performances
- $(\mathbf{X^t X})^{-} \mathbf{X^t y}$ est une matrice de moyennes ajustées

Analyse de covariance - L'exemple des chats. Le fichier

Données concernant des chats (MASS, cats)

cats {MASS}

Anatomical Data from Domestic Cats

Description

The heart and body weights of samples of male and female cats used for *digitalis* experiments. The cats were all adult, over 2 kg body weight.

Usage

```
cats
```

Format

This data frame contains the following columns:

`Sex`

sex: Factor with evels `"F"` and `"M"`.

`Bwt`

body weight in kg.

`Hwt`

heart weight in g.

Source

R. A. Fisher (1947) The analysis of covariance method for the relation between a part and the whole, *Biometrics* **3**, 65–68.

References

Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth edition. Springer.

[Package *MASS* version 7.3-45 Index]

Analyse de covariance - L'exemple des chats
Poids du coeur en fonction du sexe et du poids du corps

Les données

Sexe	Poids
F	x_1
F	x_2
M	x_3
M	x_4

Analyse de covariance - L'exemple des chats
Poids du coeur en fonction du sexe et du poids du corps

Les données

Sexe	Poids
F	x_1
F	x_2
M	x_3
M	x_4

La matrice **X**

μ	F	Poids
1	1	x_1
1	1	x_2
1	0	x_3
1	0	x_4

La variable x est centrée : $\sum x_i = 0$

Analyse de covariance - L'exemple des chats
Poids du coeur en fonction du sexe et du poids du corps

La matrice $\mathbf{X}$

μ	F	Poids
1	1	x_1
1	1	x_2
1	0	x_3
1	0	x_4

La variable x est centrée : $\sum x_i = 0$

La matrice $\mathbf{X^tX}$

$$\begin{bmatrix} \mu & F & Poids \\ 4 & 2 & 0 \\ 2 & 2 & \sum_F x_i \\ 0 & \sum_F x_i & \sum x_i^2 \end{bmatrix}$$

La matrice $\mathbf{X^tY}$

$$\begin{bmatrix} \sum Y_i \\ \sum_F Y_i \\ \sum Y_i x_i \end{bmatrix}$$

Analyse de covariance - L'exemple des chats
Poids du coeur en fonction du sexe et du poids du corps

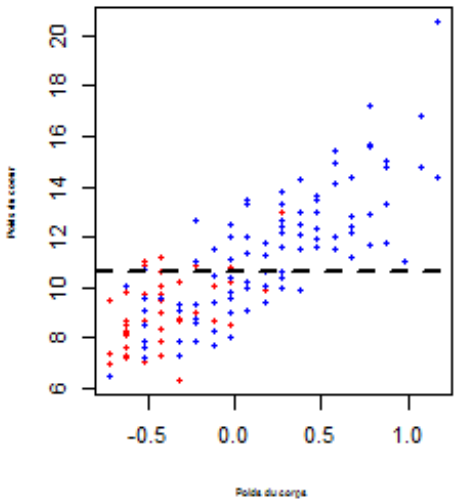
$\mathbf{X^tX}\beta = \mathbf{X^tY}$

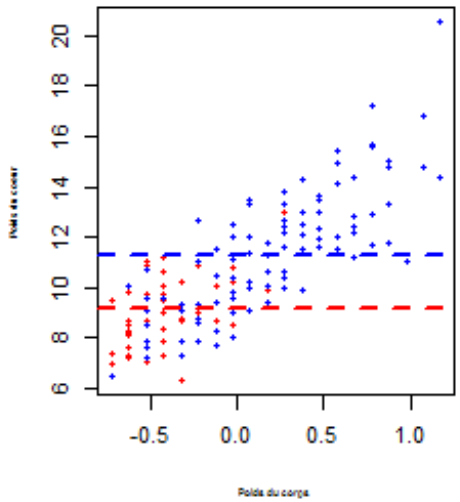
$$\begin{bmatrix} 4 & 2 & 0 \\ 2 & 2 & \sum_F x_i \\ 0 & \sum_F x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\beta}_F \\ \hat{\beta}_x \end{bmatrix} = \begin{bmatrix} 4\hat{\mu} + 2\hat{\beta}_F \\ 2\hat{\mu} + 2\hat{\beta}_F + \hat{\beta}_x \sum_F x_i \\ \sum_F x_i \hat{\beta}_F + \hat{\beta}_x \sum x_i^2 \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum_F Y_i \\ \sum Y_i x_i \end{bmatrix}$$

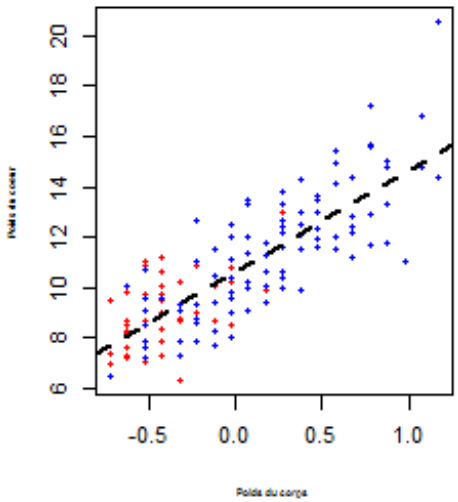
Les solutions des coefficients

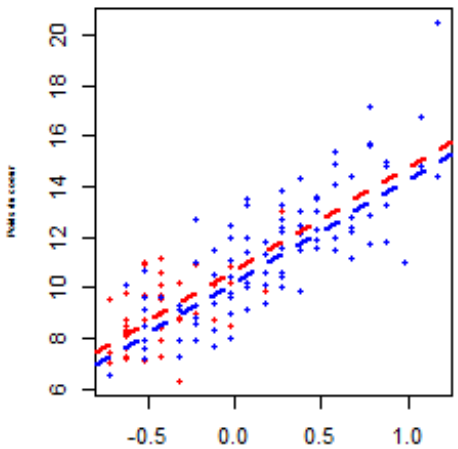
$$\begin{bmatrix} \hat{\mu} = \dfrac{\sum Y_i - 2\hat{\beta}_F}{4} \\ \hat{\beta}_F = \dfrac{\sum_F Y_i - 2\mu - \hat{\beta}_x \sum_F x_i}{2} \\ \hat{\beta}_x = \dfrac{\sum Y_i x_i - \sum_F x_i \hat{\beta}_F}{\sum x_i^2} \end{bmatrix}$$

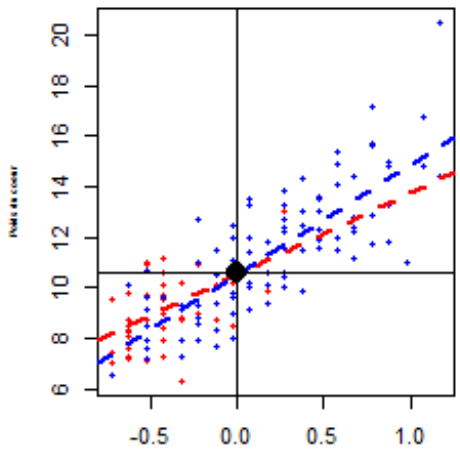
Les différents modèles

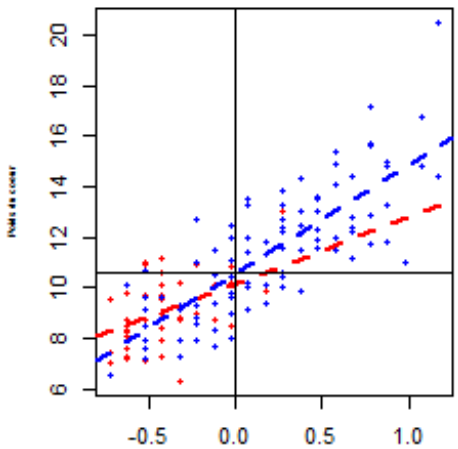




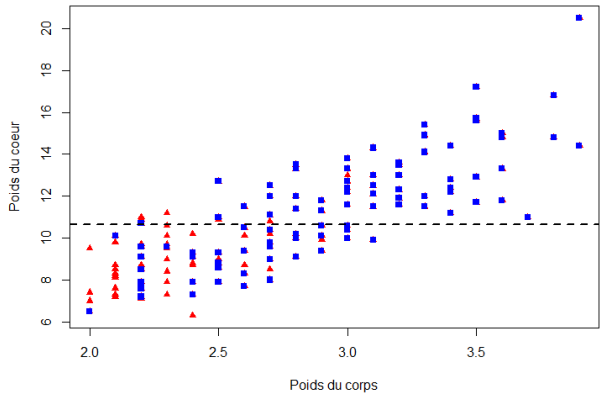








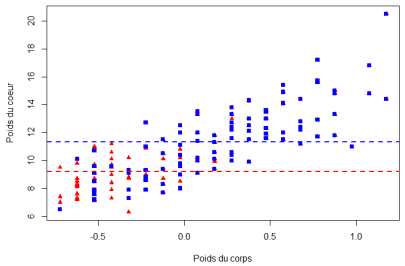
Les différents modèles : Pas d'effet



mu.lm=lm(Hwt 1,data=dat)
anova(mu.lm)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Residuals	143	847.63	5.93		

Les différents modèles : Effet du sexe



sexe.lm=lm(Hwt∼Sex,data=dat) ; anova(sexe.lm)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Sex	1	142.37	142.37	28.66	0.0000
Residuals	142	705.26	4.97		

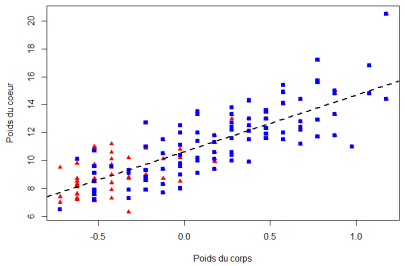
anova(mu.lm,sexe.lm)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	143	847.63				
2	142	705.26	1	142.37	28.66	0.0000

summary(sex.lm)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.2021	0.3251	28.31	0.0000
SexM	2.1206	0.3961	5.35	0.0000

Les différents modèles : Régression simple



reg.lm=lm(Hwt $\sim$ Bwtm,data=dat) ; anova(reg.lm)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Bwtm	1	548.09	548.09	259.83	0.0000
Residuals	142	299.53	2.11		

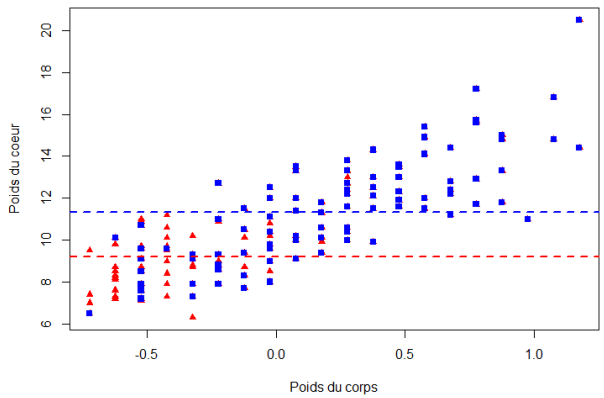
anova(mu.lm,sexe.lm)

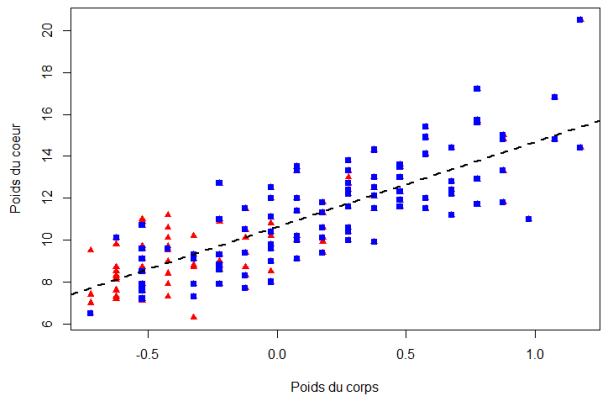
	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	143	847.63				
2	142	705.26	1	142.37	28.66	0.0000

summary(simple.lm)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.6306	0.1210	87.83	0.0000
Bwtm	4.0341	0.2503	16.12	0.0000

Comparaison Sexe / Régression





Test de Fisher partiel. Modèles emboîtés Table d'analyse de la variance

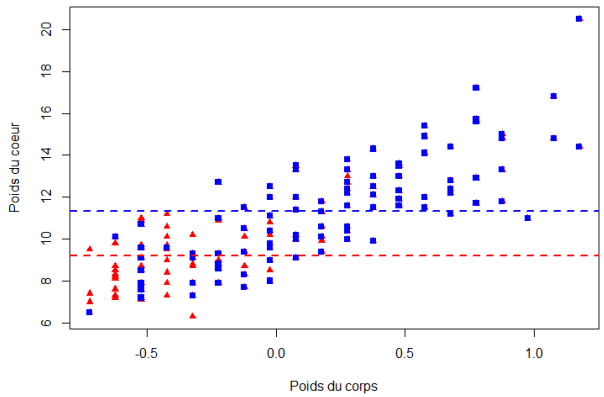
Critère d'Akaike AIC $AIC = -2\log(L) + 2k = n[\log(\frac{2\pi SCR}{n}) + 1] + 2k$

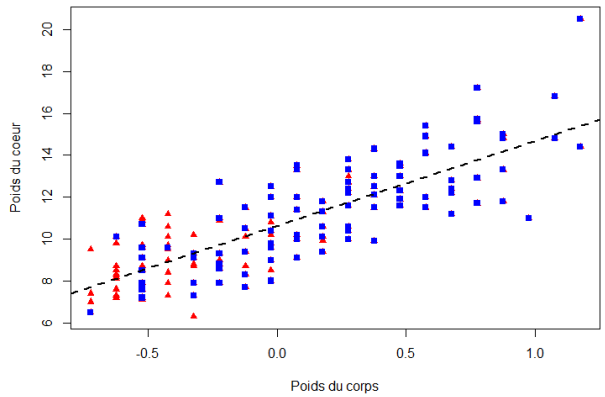
(L vraisemblance maximisée, k nombre de paramètres du modèle)

Critère d'Information Bayésien BIC $BIC = -2\log(L) + k\log(n)$

(L vraisemblance maximisée, k nombre de paramètres du modèle)

Comparaison Sexe / Régression





	AIC	BIC
Régression	520	529
Sexe	643	652

Les différents modèles : Sexe + Poids du corps

Mêmes pentes, moyennes différentes

memepente.lm=lm(Hwt $\sim$ Sexe+Bwtm,data=dat)

memepente1.lm=lm(Hwt $\sim$ Bwtm+Sexe,data=dat)

anova(memepente.lm)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Sex	1	142.37	142.37	67.05	0.0000
Bwtm	1	405.88	405.88	191.16	0.0000
Residuals	141	299.38	2.12		

anova(memepente1.lm)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Bwtm	1	548.09	548.09	258.14	0.0000
Sex	1	0.15	0.15	0.07	0.7875
Residuals	141	299.38	2.12		

summary(memepente.lm)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.6859	0.2381	44.88	0.0000
SexM	-0.0821	0.3040	-0.27	0.7875
Bwtm	4.0758	0.2948	13.83	0.0000

Les différents modèles : Régressions séparées

Pentes et moyennes différentes

separ1.lm=lm(Hwt ∼ Sexe/Bwtm,data=dat)

anova(separ1.lm)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Sex	1	142.37	142.37	68.48	0.0000
Sex:Bwtm	2	414.21	207.11	99.62	0.0000
Residuals	140	291.05	2.08		

summary(separ1.lm)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.1619	0.3522	28.86	0.0000
SexM	0.4001	0.3854	1.04	0.3010
SexF:Bwtm	2.6364	0.7759	3.40	0.0009
SexM:Bwtm	4.3127	0.3148	13.70	0.0000

Les différents modèles : Régressions séparées

Fichiers séparés

fem.lm = lm(Hwt ∼ Bwtm,data=fem) ; mal.lm = lm(Hwt ∼ Bwtm,data=mal)
summary(fem.lm)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.1619	0.2839	35.80	0.0000
Bwtm	2.6364	0.6254	4.22	0.0001

summary(mal.lm)

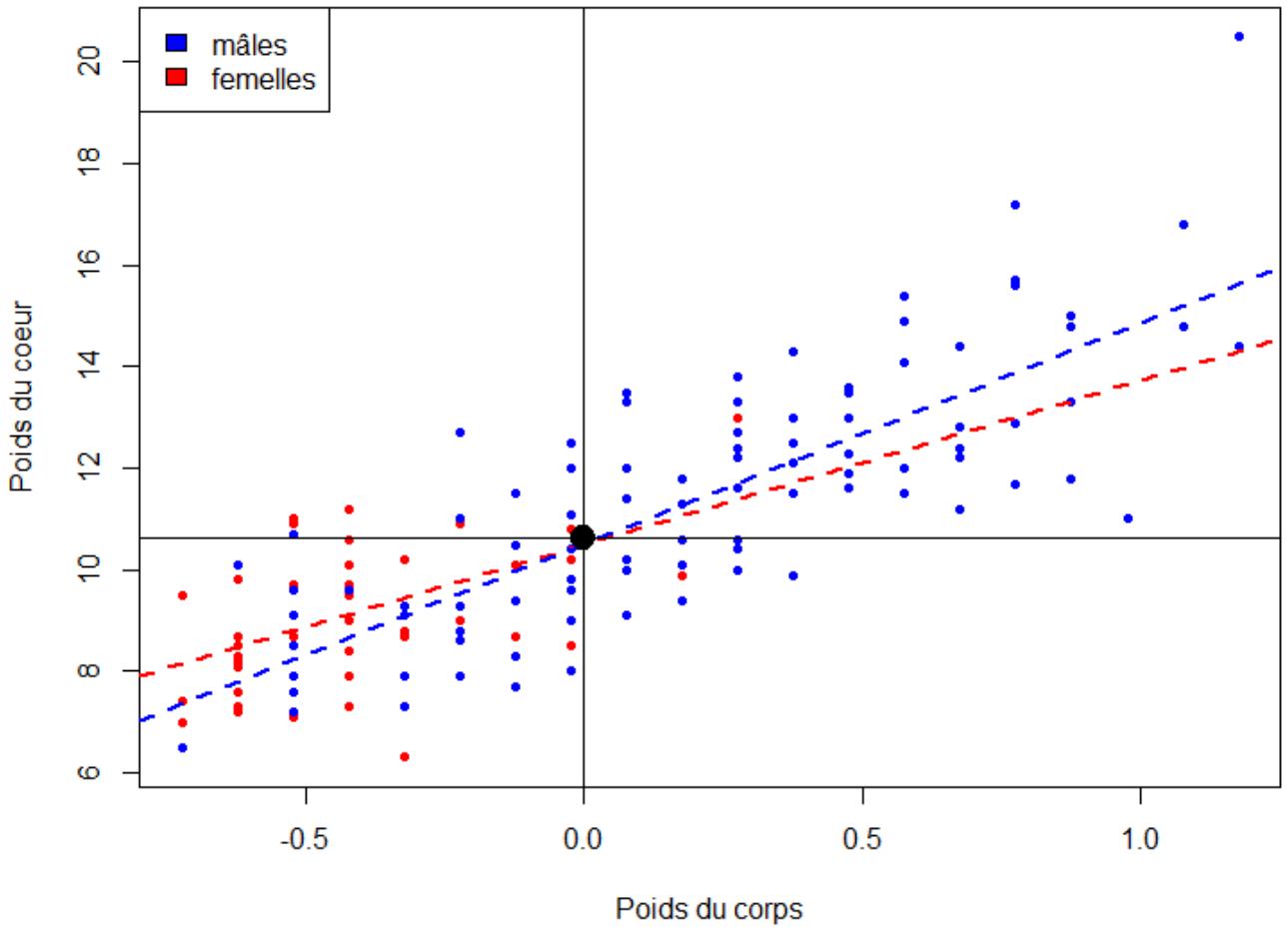
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.5620	0.1691	62.47	0.0000
Bwtm	4.3127	0.3399	12.69	0.0000

Fichier commun

summary(separ1.lm)

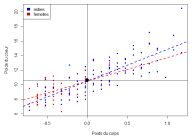
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.1619	0.3522	28.86	0.0000
SexM	0.4001	0.3854	1.04	0.3010
SexF:Bwtm	2.6364	0.7759	3.40	0.0009
SexM:Bwtm	4.3127	0.3148	13.70	0.0000

Même moyenne, pentes différentes



separ0.lm=lm(Hwt ~ Bwtm /Sexe,data=dat)

Même moyenne, pentes différentes



separ0.lm=lm(Hwt ∼ Bwtm/Sexe,data=dat)
anova(separ0.lm)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Bwtm	1	548.09	548.09	263.50	0.0000
Bwtm:Sex	1	6.25	6.25	3.00	0.0853
Residuals	141	293.29	2.08		

summary(separ0.lm)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.4959	0.1431	73.34	0.0000
Bwtm	3.2268	0.5280	6.11	0.0000
Bwtm:SexM	1.1330	0.6538	1.73	0.0853

Choix du modèle

Régression *vs* Pentes séparées (même intercept)

anova(reg.lm,separ0.lm)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	142	299.53				
2	141	293.29	1	6.25	3.00	0.0853

Pentes séparées (même intercept *vs* intercepts séparés

anova(separ0.lm,separ1.lm)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	141	293.29				
2	140	291.05	1	2.24	1.08	0.3010

Régression simple *vs* intercepts et pentes séparés

anova(reg.lm,separ1.lm)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	142	299.53				
2	140	291.05	2	8.49	2.04	0.1337

Un exemple chez la poule. Gene *Frizzle* et thermotolérance

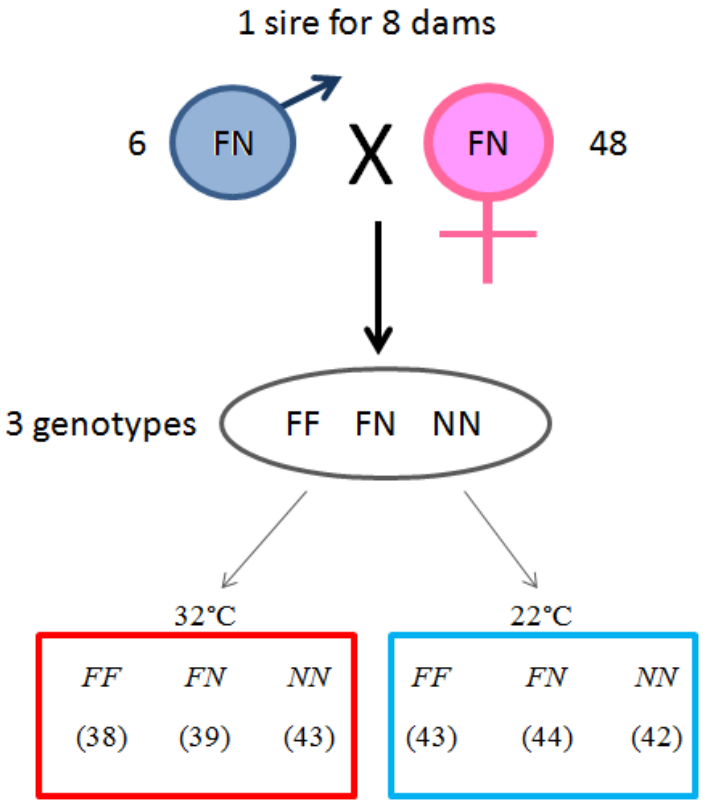
Données fournies par T Zerjal

Mutation Frizzle (F)

- Dominance incomplète
- Effet sur la structure des plumes
- Réduction de la masse du plumage
- Augmentation de la thermotolérance ?

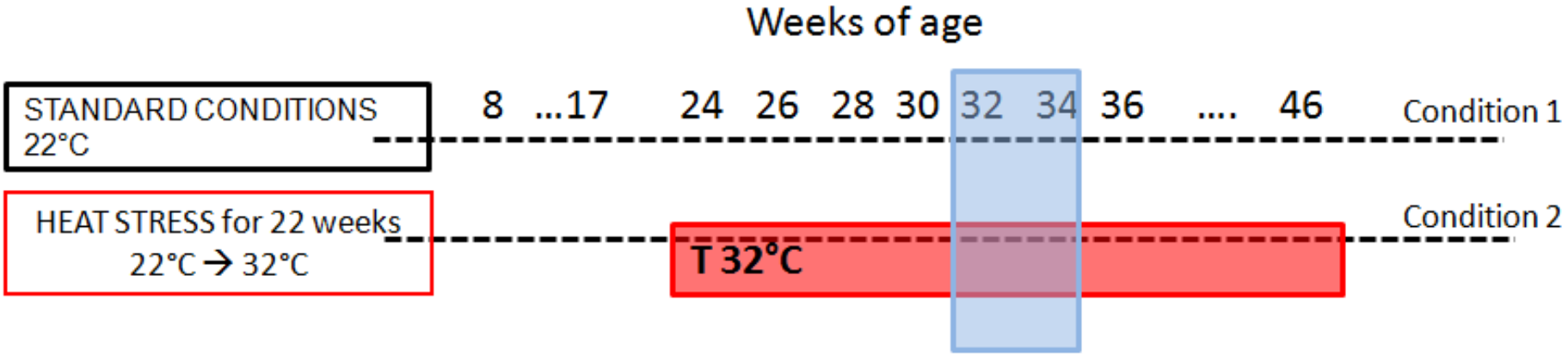
Le protocole d'expérience

Experimental design: 3 genotypes exposed to 2 environmental conditions



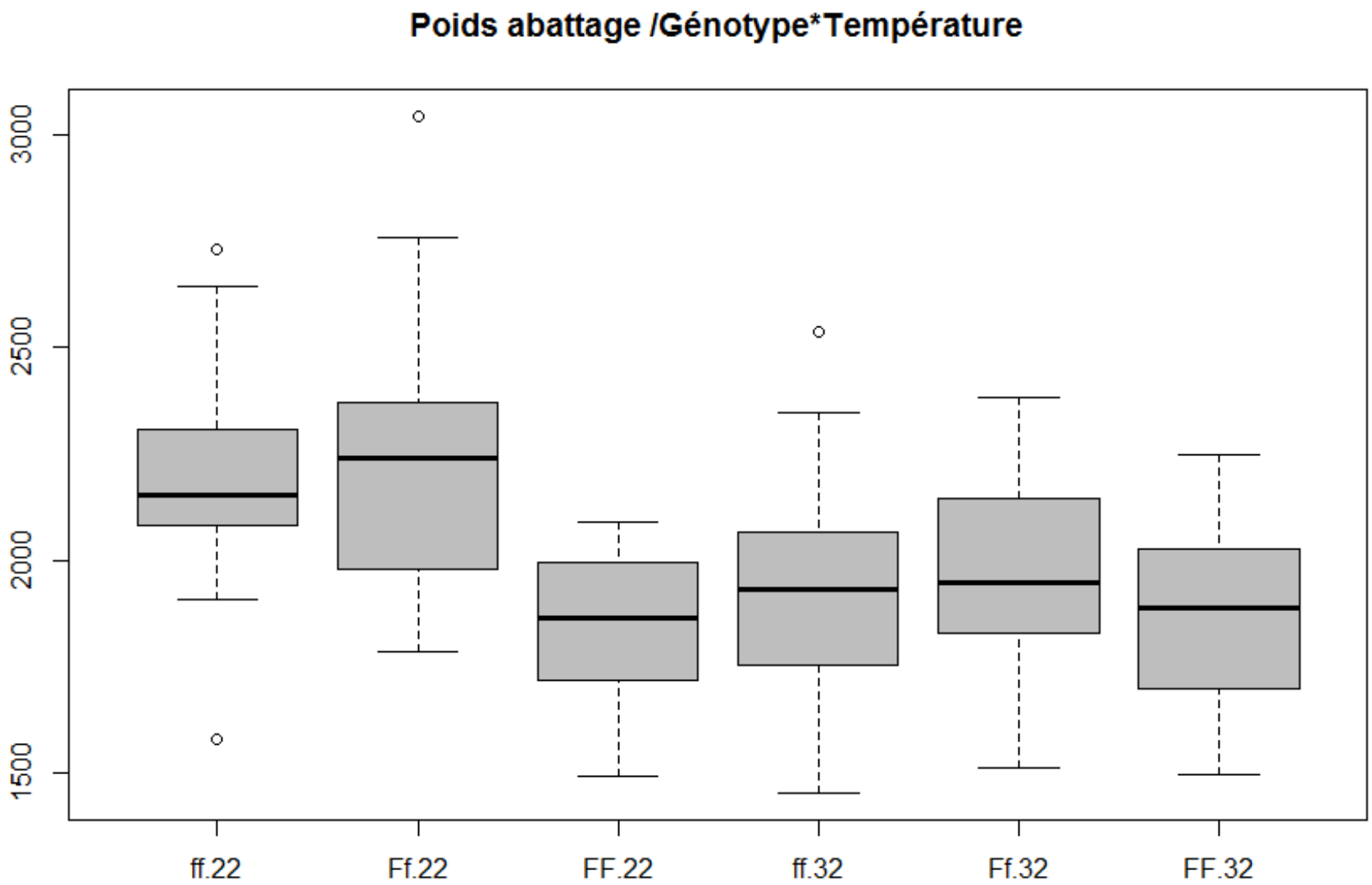
Le protocole d'expérience

The experimental planning

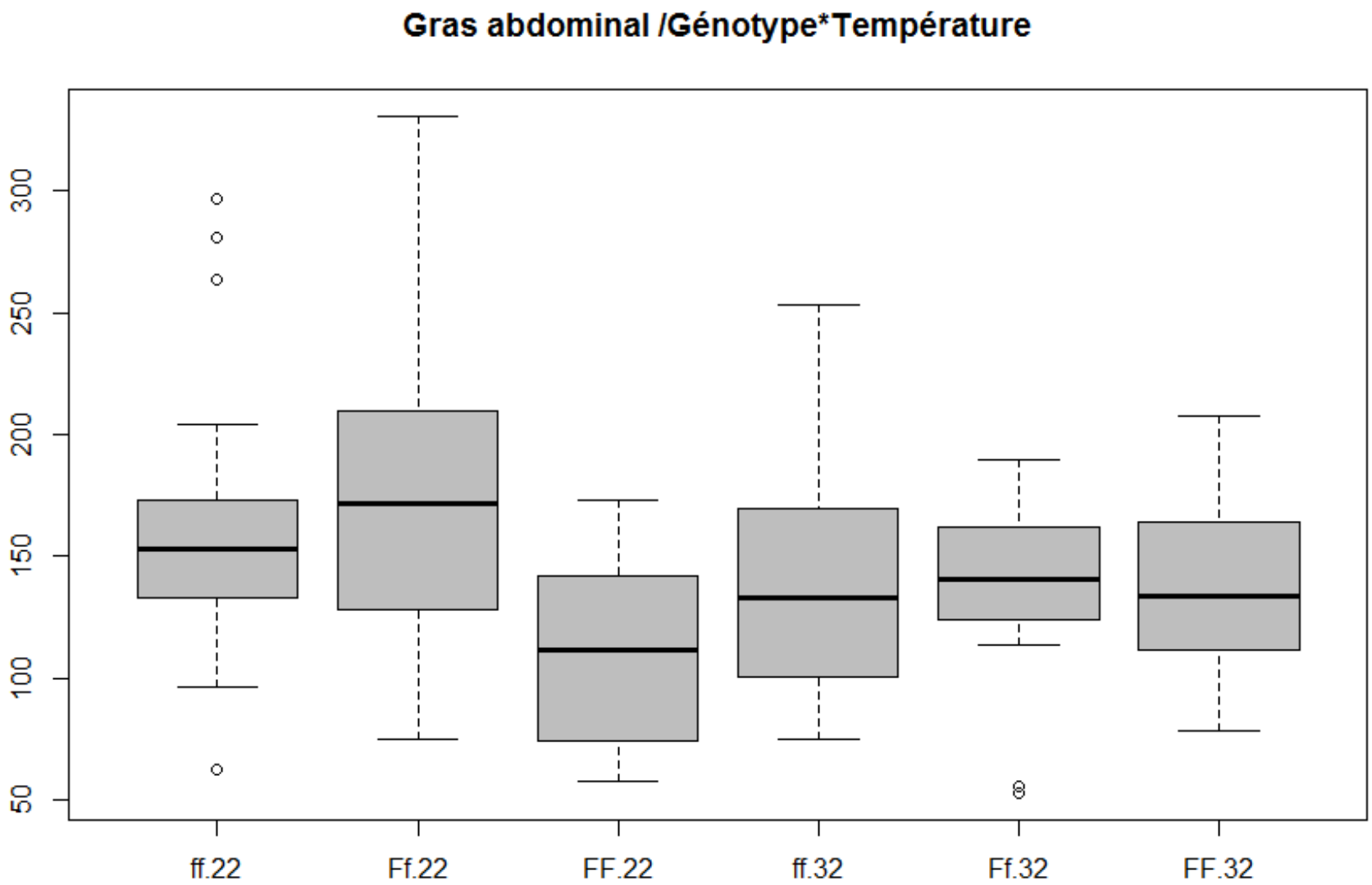


Feed efficiency measures:
BW at 31, BW at 34, 4 weeks FI and Egg Mass→ Residual Feed Consumption
(Fraction of total feed intake (FI) which is not explained by maintenance requirements and production)

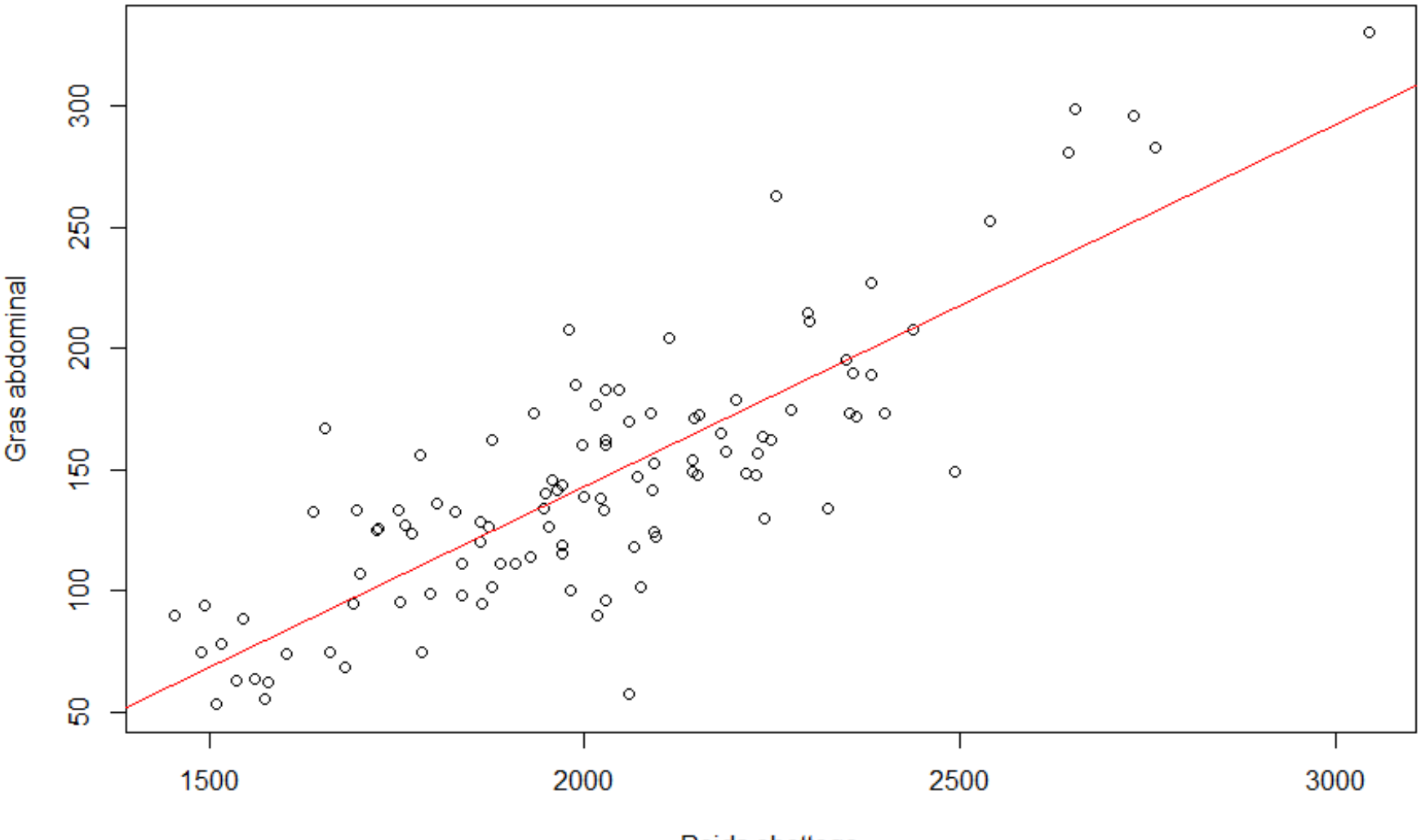
Poids à l'abattage



Gras abdominal



Gras abdominal et poids d'abattage



Gras abdominal / Génotype et Température

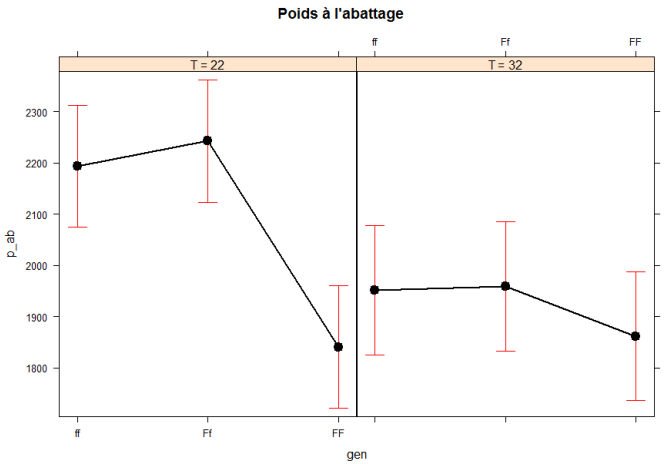
sans poids abattage

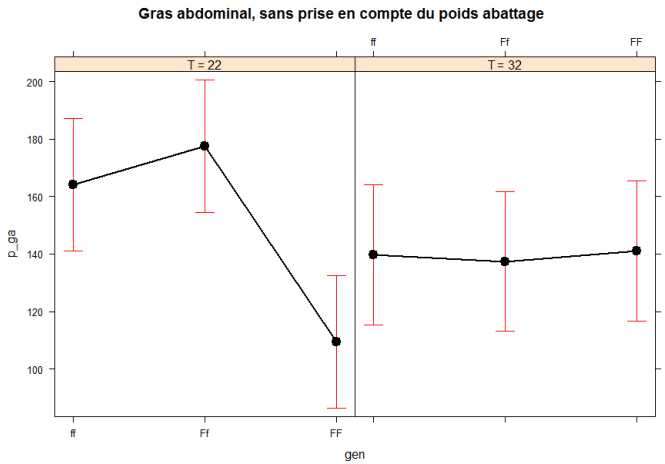
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
gen	2	24171.57	12085.78	4.72	0.0110
T	1	3257.28	3257.28	1.27	0.2622
gen:T	2	25517.32	12758.66	4.98	0.0086
Residuals	102	261346.19	2562.22		

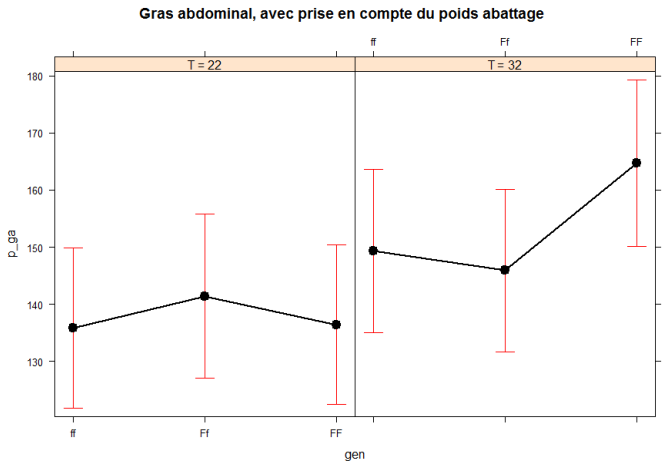
avec poids abattage

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
gen	2	24171.57	12085.78	13.87	0.0000
T	1	3257.28	3257.28	3.74	0.0560
pm_ab	1	196404.61	196404.61	225.33	0.0000
gen:T	2	2422.62	1211.31	1.39	0.2539
Residuals	101	88036.28	871.65		

Gras abdominal / Génotype et Température







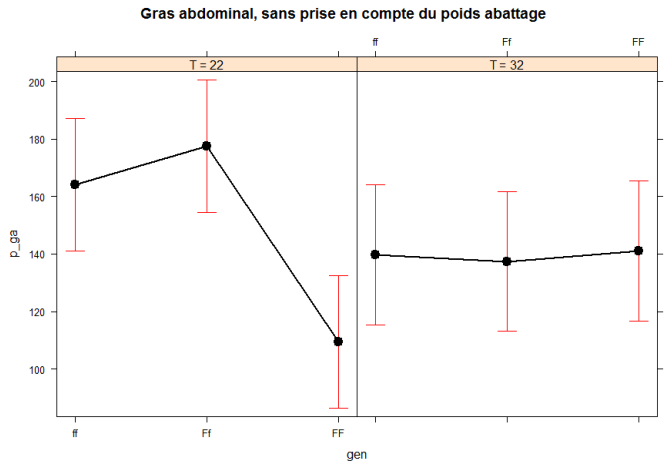
Gras abdominal / Génotype et Température

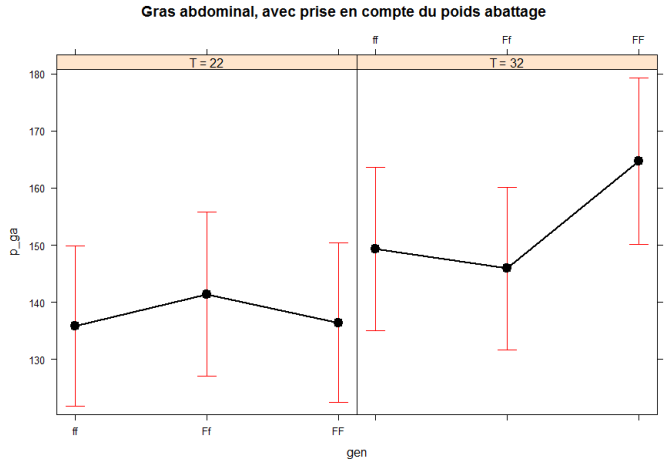
sans poids abattage

	Pr(>F)
gen	0.0110
T	0.2622
gen:T	0.0086
Residuals	

avec poids abattage

	Pr(>F)
gen	0.0000
T	0.0560
pm_ab	0.0000
gen:T	0.2539
Residuals	





Modèles et formules

Table 6: Model Formulae

Expression	Description
`y ~ x`	Simple regression
`y ~ 1+x`	Explicit intercept
`y ~ -1 + x`	Through the origin
`y ~ x +` x^2	Quadratic regression
`y ~ x1 + x2 + x3`	Multiple regression
`y ~ G + x1 + x2`	Parallel regressions
`y ~ G/(x1+x2)`	Separate regressions
`sqrt(y) ~ x +` x^2	Transformed
`y ~ G`	Single Classification
`y ~ A+B`	Randomized block
`y ~ B+N*P`	Factorial in blocks
`y ~ x+B+N*P`	with covariate
`y ~ .-X1`	All variables except `X1`
`. ~ .+A:B`	Add interaction (update)
`Nitrogen ~ Times*(River/Site)`	More complex design

Kuhnert et Venables, 2005

References

- P Kuhnert et B Venables, 2005. An introduction to R, software for statistical modeling and computing, CSIRO, Australia. téléchargeable sur les sites CRAN
- Documentation disponible sur le site ade4