Modèles linéaires

Analyses de variance

Denis Laloë
GABI - PSGen

27 septembre 2016

Résumé de l'épisode précédent

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

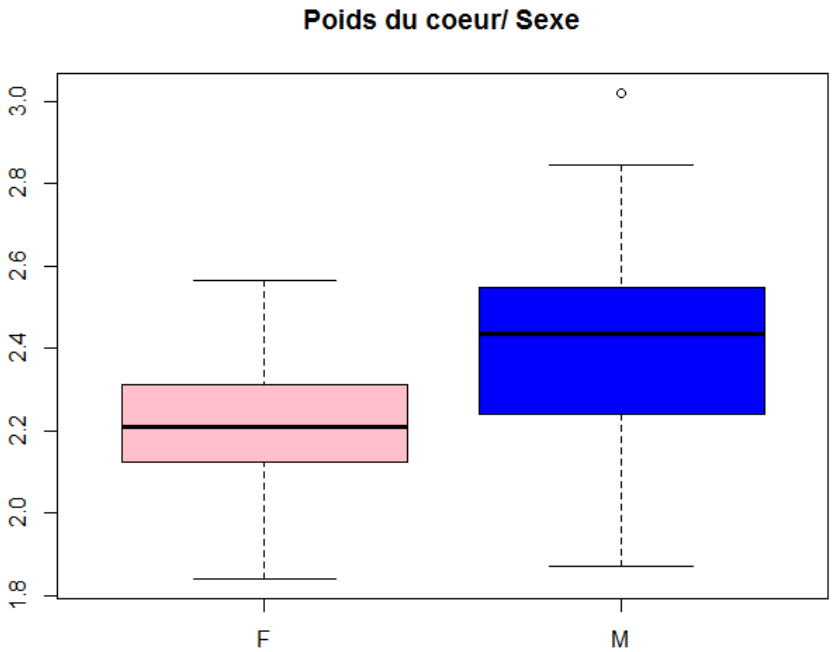
Nature des variables explicatives

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

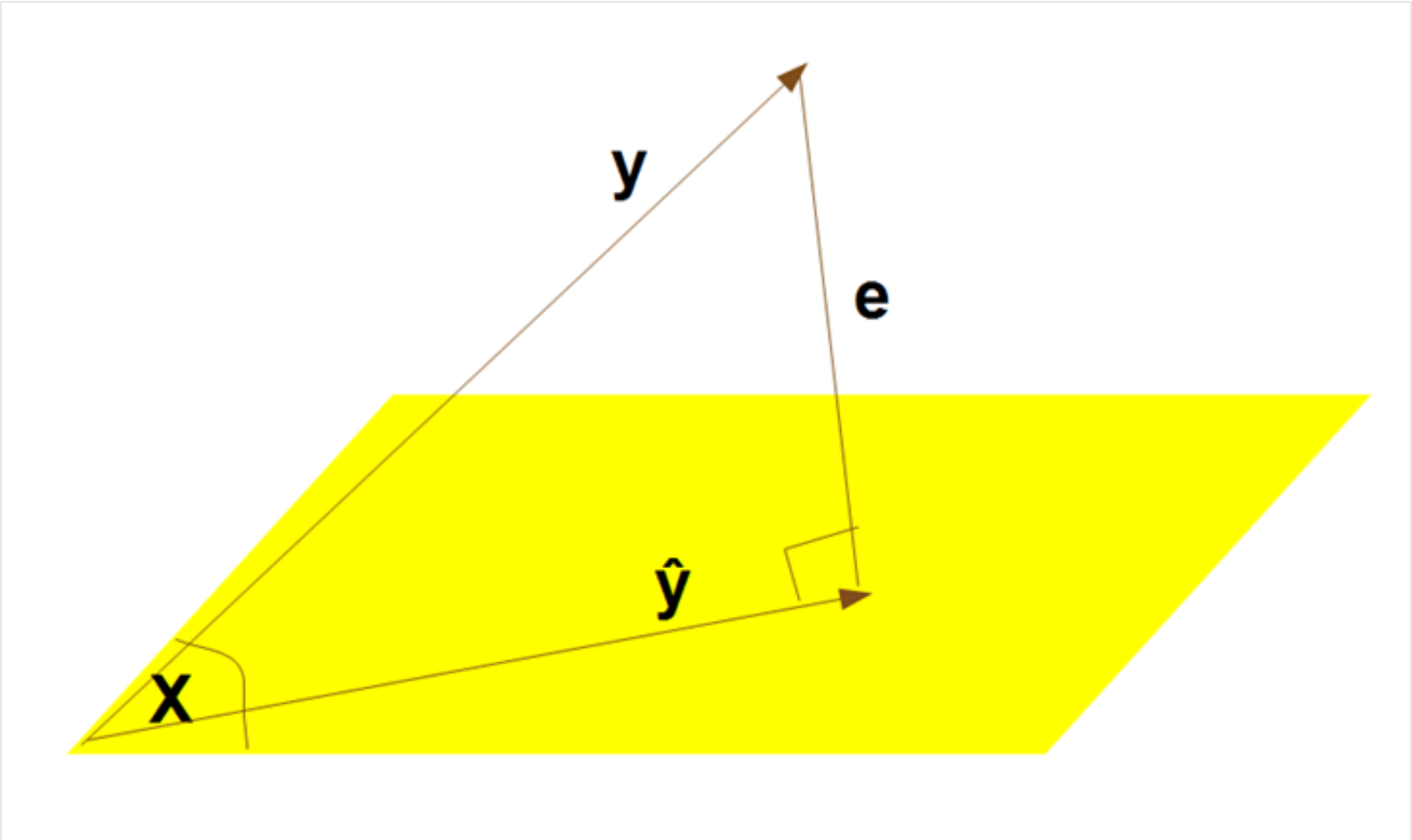
Modèle linéaire et type de variables

Variables quantitatives/qualitative

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



L'approche géométrique



Le modèle linéaire général

Modèle général

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

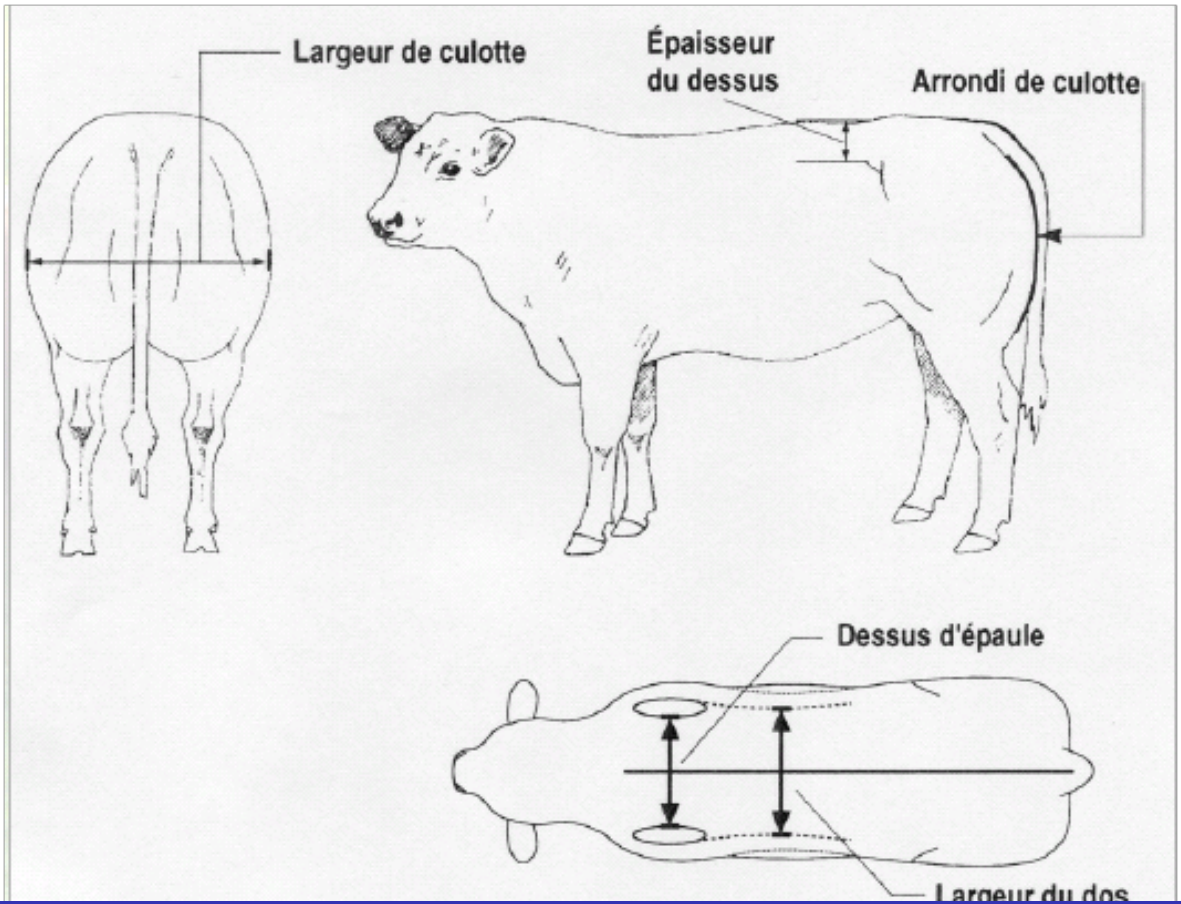
$$\mathbf{Y} = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(p)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(p)} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \beta_p \end{bmatrix} + \mathbf{E}$$

Résolution

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

Un exemple - Morphologie des bovins allaitants

- Variable réponse : dessus d'épaule
- Deux facteurs explicatifs
  - Campagne
  - Pointeur



Le modèle

$$Y_{ij} = \mu + pointeur_i + camp_j + E_{ij}$$

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

$$\mathbf{Y} = \mathbf{X} \begin{bmatrix} \mu \\ pointeur_1 \\ pointeur_2 \\ camp_1 \\ camp_2 \end{bmatrix} + \mathbf{E}$$

Les données

	pointeur	camp	DM
21228	405	2003	3.00
19867	405	2002	7.00
81887	403	2002	4.00
82214	403	2003	9.00
212281	405	2003	3.00
198671	405	2002	7.00
818871	403	2002	4.00
822141	403	2003	9.00

	pointeur	camp	DM
21228	405	2003	3.00
19867	405	2002	7.00
81887	403	2002	4.00
82214	403	2003	9.00
212281	405	2003	3.00
198671	405	2002	7.00
818871	403	2002	4.00
822141	403	2003	9.00

La matrice **X**

	Intercept	pointeur403	pointeur405	camp2002	camp2003
21228	1.00	0.00	1.00	0.00	1.00
19867	1.00	0.00	1.00	1.00	0.00
81887	1.00	1.00	0.00	1.00	0.00
82214	1.00	1.00	0.00	0.00	1.00
212281	1.00	0.00	1.00	0.00	1.00
198671	1.00	0.00	1.00	1.00	0.00
818871	1.00	1.00	0.00	1.00	0.00
822141	1.00	1.00	0.00	0.00	1.00

	pointeur	camp	DM
21228	405	2003	3.00
19867	405	2002	7.00
81887	403	2002	4.00
82214	403	2003	9.00
212281	405	2003	3.00
198671	405	2002	7.00
818871	403	2002	4.00
822141	403	2003	9.00

La matrice $\mathbf{X^tX}$

	Intercept	pointeur403	pointeur405	camp2002	camp2003
Intercept	8.00	4.00	4.00	4.00	4.00
pointeur403	4.00	4.00	0.00	2.00	2.00
pointeur405	4.00	0.00	4.00	2.00	2.00
camp2002	4.00	2.00	2.00	4.00	0.00
camp2003	4.00	2.00	2.00	0.00	4.00

La matrice $\mathbf{X^tY}$

	x
Intercept	46.00
pointeur403	26.00
pointeur405	20.00
camp2002	22.00
camp2003	24.00

Interprétation des matrices

- $\mathbf{X}$: colonnes
- $\mathbf{X^tX}$
- $\mathbf{X^tY}$

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

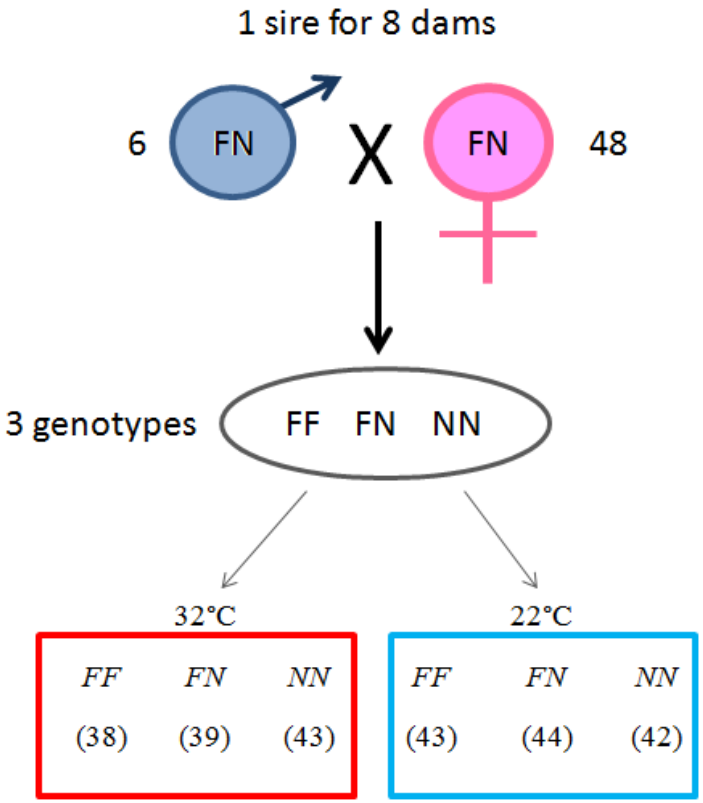
Données fournies par T Zerjal

Mutation Frizzle (F)

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

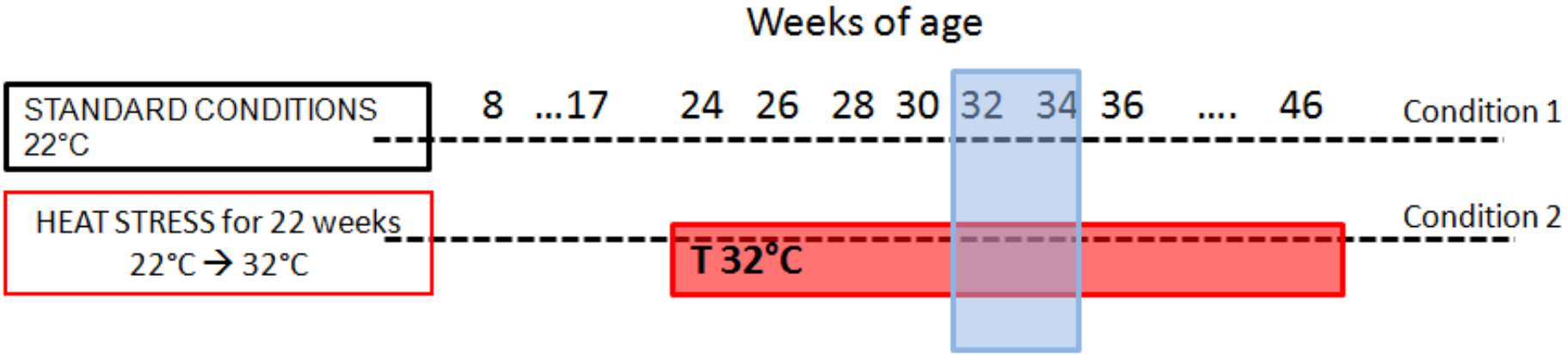
Le protocole d'expérience

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)

Les mesures

Traits measured (23 in total):

Body related traits
Body Weight (BW) at 18, 30 and 46 weeks of age and an average adult BW .
Body weight change (DBW) between 34 and 31 weeks, body change 2(dw2) between of age and body change 3 (dw3) between 46-30 wks

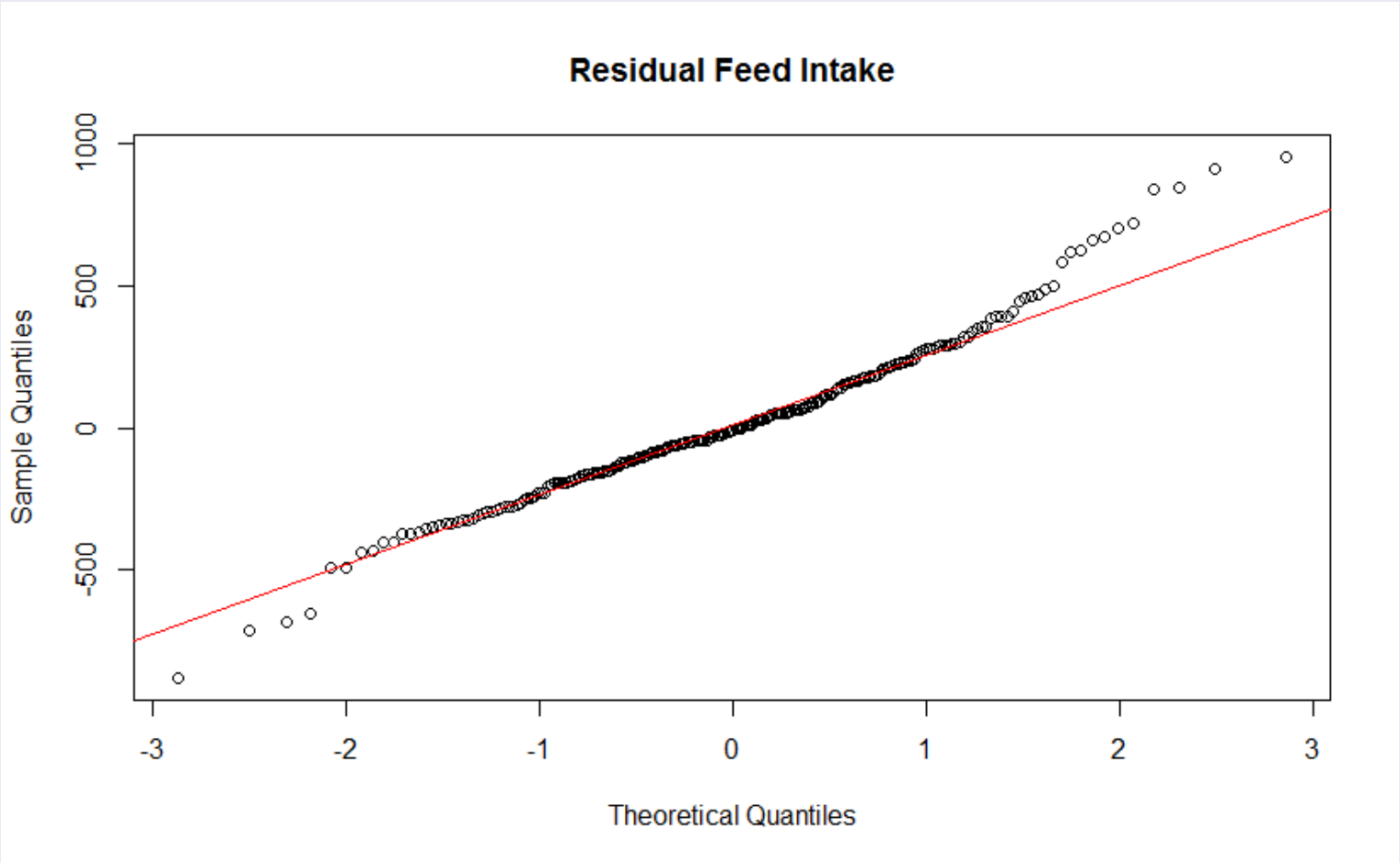
Egg production and quality related traits
Age first egg (PREC); Number of eggs (N_EGGS); Laying rate (IP); % soft eggs (PM); % of cracked eggs (PC); clutch length (CL); percentage of pauses (PP); average egg weight (EW); membrane thickness (MT); shell thickness (ST); egg mass (EM)

Feed efficiency related traits
Feed intake (FI); Residual feed intake (RFI)

Physiological traits
rectal temperature (RT)

La capacité d'ingestion résiduelle

Normalité



La capacité d'ingestion résiduelle

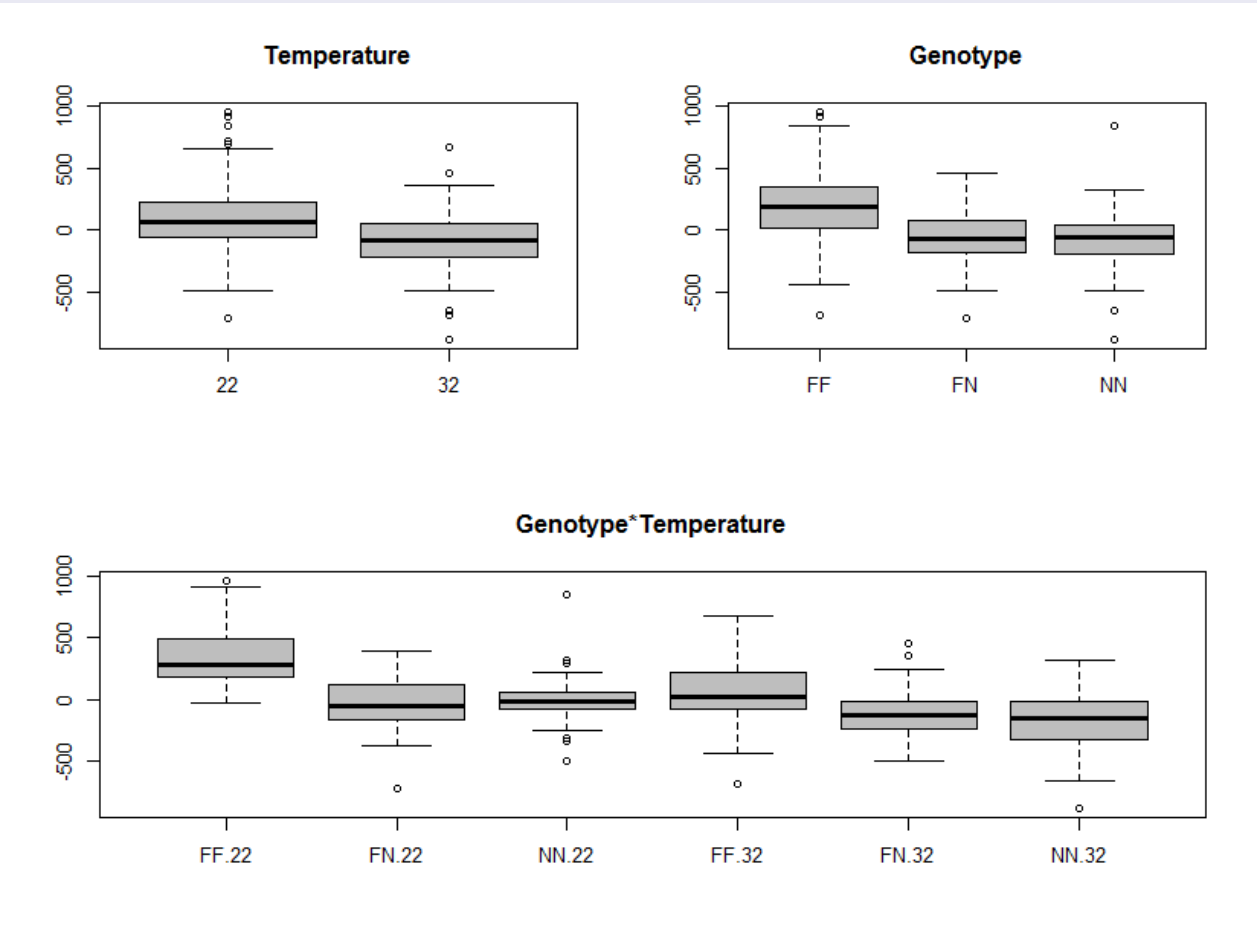
Les données

	sire	genotype	Temperature	R
1	30	FF	32	183.10
2	30	FF	32	672.40
3	30	FN	32	-247.60
4	30	NN	32	-326.30
5	30	FN	32	-356.80
6	30	FF	32	234.20

- *sire*,*genotype*,*Temperature* sont des variables qualitatives / facteurs
- R est une variable quantitative

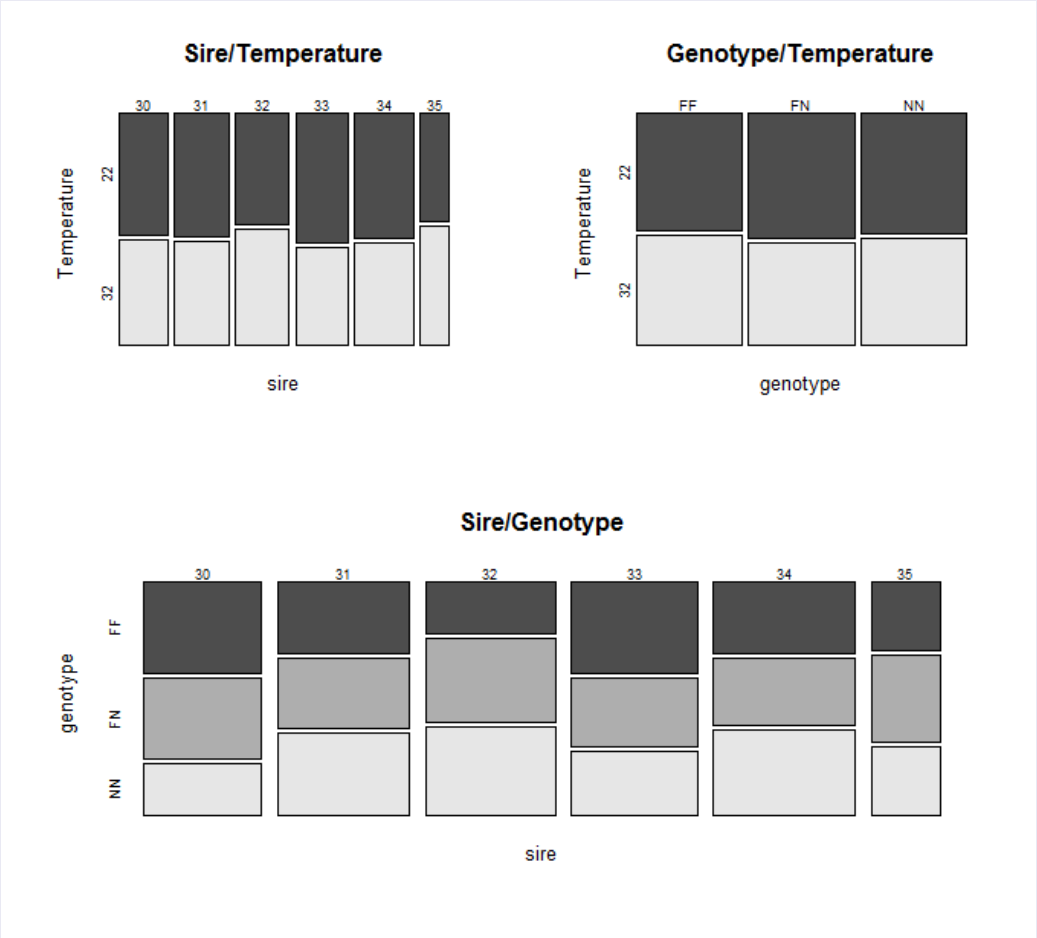
La capacité d'ingestion résiduelle

Boxplots



La capacité d'ingestion résiduelle

Mosaicplots

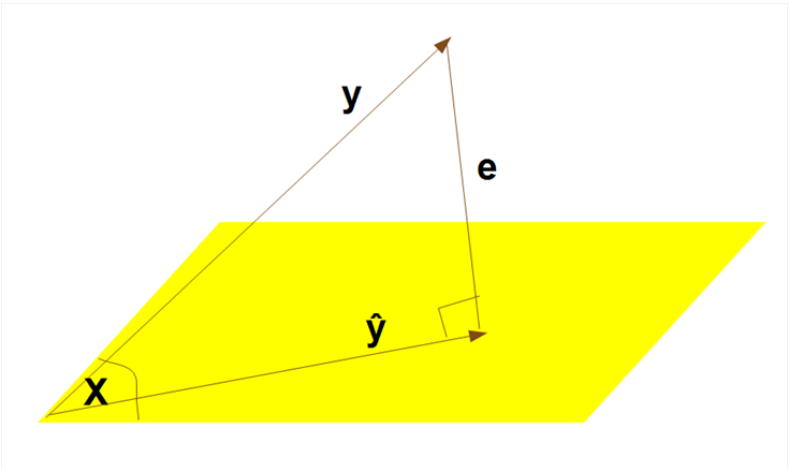


Analyse de variance

Un premier modèle

$Y_{ij} = Temperature_i + genotype_j + E_{ij}$

lm(formula = R ∼ Temperature + genotype , data = fic)



Analyse de variance

Analyse de variance 1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Modèle	3	5778053	1841033	35.67	0.0000
Residuals	234	12636168	504001		

- $R^2 = 0,314$
- $R^2_{adj} = 0,305$

Anova : Tests séquentiels

$Y_{ij} = Temperature_i + genotype_j + E_{ij}$

	Sum Sq
Température (T)	$SC_T = \|\hat{Y}_T - \bar{Y}\|^2$
Génotype (G) | Température	$SC_G = \|\hat{Y}_{T+G} - \hat{Y}_T\|^2$
Résiduelle (R)	$SC_R = \|Y - \hat{Y}_{T+G}\|^2$

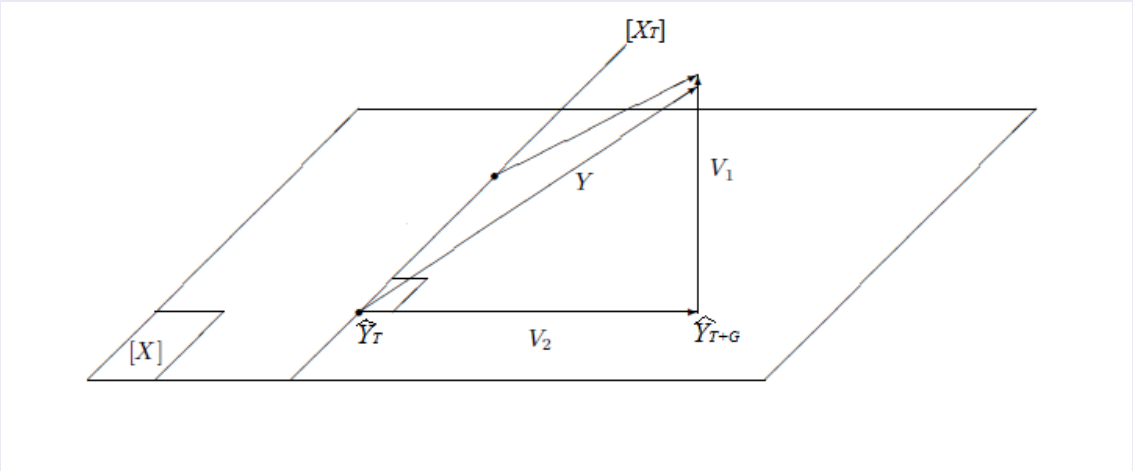
$Y_{ji} = genotype_j + Temperature_i + E_{ji}$

	Sum Sq
Genotype (G)	$SC_G = \|\hat{Y}_G - \bar{Y}\|^2$
Température | Génotype (G)	$SC_T = \|\hat{Y}_{T+G} - \hat{Y}_G\|^2$
Résiduelle (R)	$SC_R = \|Y - \hat{Y}_{T+G}\|^2$

- $\hat{Y}_T - \bar{Y}$, $\hat{Y}_G - \bar{Y}$,$\hat{Y}_{T+G} - \hat{Y}_G$, $\hat{Y}_{T+G} - \hat{Y}_T$ dans l'espace des prédicteurs (colonnes de $\mathbf{X}$)
- $Y - \hat{Y}_{T+G}$ est la résiduelle, orthogonale aux précédentes
- Sommes de carrés indépendantes, Test de Fisher

Anova : Tests séquentiels

$Y_{ij} = Temperature_i + genotype_j + E_{ij}$



d'après *Azais et Bardet*

- $\hat{Y}_T - \bar{Y}$, $\hat{Y}_G - \bar{Y}$, $\hat{Y}_{T+G} - \hat{Y}_G$, $\hat{Y}_{T+G} - \hat{Y}_T$ dans l'espace des prédicteurs (colonnes de $\mathbf{X}$)
- $Y - \hat{Y}_{T+G}$ est la résiduelle, orthogonale aux précédentes
- Sommes de carrés indépendantes, Test de Fisher

Table d'analyse de variance

$$Y_{ij} = Temperature_i + genotype_j + E_{ij}$$

		Df	Sum Sq	Mean Sq	F value
Température (T)	$\hat{Y}_T - \bar{Y}$	$df_T = n_T - 1$	$SC_T = \|\hat{Y}_T - \bar{Y}\|^2$	$CM_T = \frac{SC_T}{Df_T}$	$\frac{CM_T}{CM_R}$
Génotype (G)	$\hat{Y}_{T+G} - \hat{Y}_T$	$df_G = n_G - 1$	$SC_G = \|\hat{Y}_{T+G} - \hat{Y}_T\|^2$	$CM_G = \frac{SC_G}{Df_G}$	$\frac{CM_G}{CM_R}$
Résiduelle (R)	$Y - \hat{Y}_{T+G}$	$df_R = n_r - n_T - n_G + 1$	$SC_R = \|Y - \hat{Y}_{T+G}\|^2$	$CM_R = \frac{SC_R}{Df_R}$	

Tables d'analyse de variance

Analyse de variance 1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Modèle	3	5778053	1841033	35.67	0.0000
Residuals	234	12636168	504001		

$Y_{ij} = Temperature_i + genotype_j + E_{ij}$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Temperature	1	1841032.53	1841032.53	34.09	0.0000
genotype	2	3937020.11	1968510.06	36.45	0.0000
Residuals	234	12636167.74	54000.72		

$Y_{ji} = genotype_j + Temperature_i + E_{ij}$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
genotype	2	3829236.71	1914618.36	35.46	0.0000
Temperature	1	1948815.93	1948815.93	36.09	0.0000
Residuals	234	12636167.74	54000.72		

Estimation des effets

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

$Y_{ij} = Temperature_i + genotype_j + E_{ij}$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	284.7333	29.9104	9.52	0.0000
Temperature32	-181.4481	30.2042	-6.01	0.0000
genotypeFN	-265.7591	36.8706	-7.21	0.0000
genotypeNN	-280.1196	36.9764	-7.58	0.0000

$Y_{ji} = genotype_j + Temperature_i + E_{ij}$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	284.7333	29.9104	9.52	0.0000
genotypeFN	-265.7591	36.8706	-7.21	0.0000
genotypeNN	-280.1196	36.9764	-7.58	0.0000
Temperature32	-181.4481	30.2042	-6.01	0.0000

Le codage

$$Y_{ijk} = Temperature_i + genotype_j + sire_k + E_{ijk}$$

La matrice **X**

	(Intercept)	Temperature22	Temperature32	genotypeNN	genotypeFN	genotypeNN
1	1.00	0.00	1.00	1.00	0.00	0.00
2	1.00	0.00	1.00	1.00	0.00	0.00
3	1.00	0.00	1.00	0.00	1.00	0.00
...	...	...	...	...	...	...
249	1.00	1.00	0.00	0.00	0.00	1.00

Le codage (par défaut) dans R

	(Intercept)	Temperature32	genotypeFN	genotypeNN
1	1.00	1.00	0.00	0.00
2	1.00	1.00	0.00	0.00
3	1.00	1.00	1.00	0.00
...	...	...	...	...
249	1.00	0.00	0.00	1.00

- Intra-facteur, la somme des colonnes est égal au vecteur $\mathbf{1}$
- Notion de contraintes / contrastes

Le codage et les contrastes

Le codage par défaut dans R

- Le premier effet de chaque facteur est mis à 0 : effet "témoin"
- contraste *contr.treatment*

	(Intercept)	Temperature32	genotypeFN	genotypeNN
1	1.00	1.00	0.00	0.00
2	1.00	1.00	0.00	0.00
3	1.00	1.00	1.00	0.00
4	1.00	1.00	0.00	1.00
5	1.00	1.00	1.00	0.00
6	1.00	1.00	0.00	0.00
7	1.00	1.00	1.00	0.00
8	1.00	1.00	1.00	0.00
...	...	...	...	...
249	1.00	0.00	0.00	1.00

Les estimées

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	284.7333	29.9104	9.52	0.0000
Temperature32	-181.4481	30.2042	-6.01	0.0000
genotypeFN	-265.7591	36.8706	-7.21	0.0000
genotypeNN	-280.1196	36.9764	-7.58	0.0000

Le codage et les contrastes

Changement de la référence

facteur $\leftarrow$ relevel(*facteur*,ref=*"nivref"*)

	(Intercept)	Temperature32	genotypeFF	genotypeFN
1	1.00	1.00	1.00	0.00
2	1.00	1.00	1.00	0.00
...	...	...	...	...
249	1.00	0.00	0.00	0.00

La matrice **X**

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.6137	29.7266	0.16	0.8768
Temperature32	-181.4481	30.2042	-6.01	0.0000
genotypeFF	280.1196	36.9764	7.58	0.0000
genotypeFN	14.3605	36.8628	0.39	0.6972

Le codage et les contrastes

Référence : FF

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	284.7333	29.9104	9.52	0.0000
Temperature32	-181.4481	30.2042	-6.01	0.0000
genotypeFN	-265.7591	36.8706	-7.21	0.0000
genotypeNN	-280.1196	36.9764	-7.58	0.0000

Référence : NN

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.6137	29.7266	0.16	0.8768
Temperature32	-181.4481	30.2042	-6.01	0.0000
genotypeFF	280.1196	36.9764	7.58	0.0000
genotypeFN	14.3605	36.8628	0.39	0.6972

La sélection de modèles : le R^2 ?

La carte et le territoire

Les collèges de cartographes levèrent une carte de l'Empire, qui avait le format de l'empire et qui coincidait avec lui point par point.

in J L Borges, Histoire universelle de l'infamie/Histoire de l'éternité

La version "modèle linéaire"

iden $\leftarrow$ as.factor(as.numeric(1:238))
borges.lm(formula = R $\sim$ iden,data=fic)
summary(borges.lm)$r.squared
[1] 1
R^2 s'accroît automatiquement au fur et à mesure de l'introduction de nouveaux facteurs dans le modèle.

La sélection de modèles. Autres critères

Test de Fisher partiel Table d'analyse de la variance

R^2 ajusté $R^2_{adj} = 1 - \frac{SCR}{SCT}\frac{n-1}{n-p}$

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)

Sélection du modèle

Les fonctions

- *add1* : ajout d'un facteur
- *drop1* : retrait d'un facteur
- *step* : recherche automatique

Les critères

- test F
- AIC
- BIC

Les données

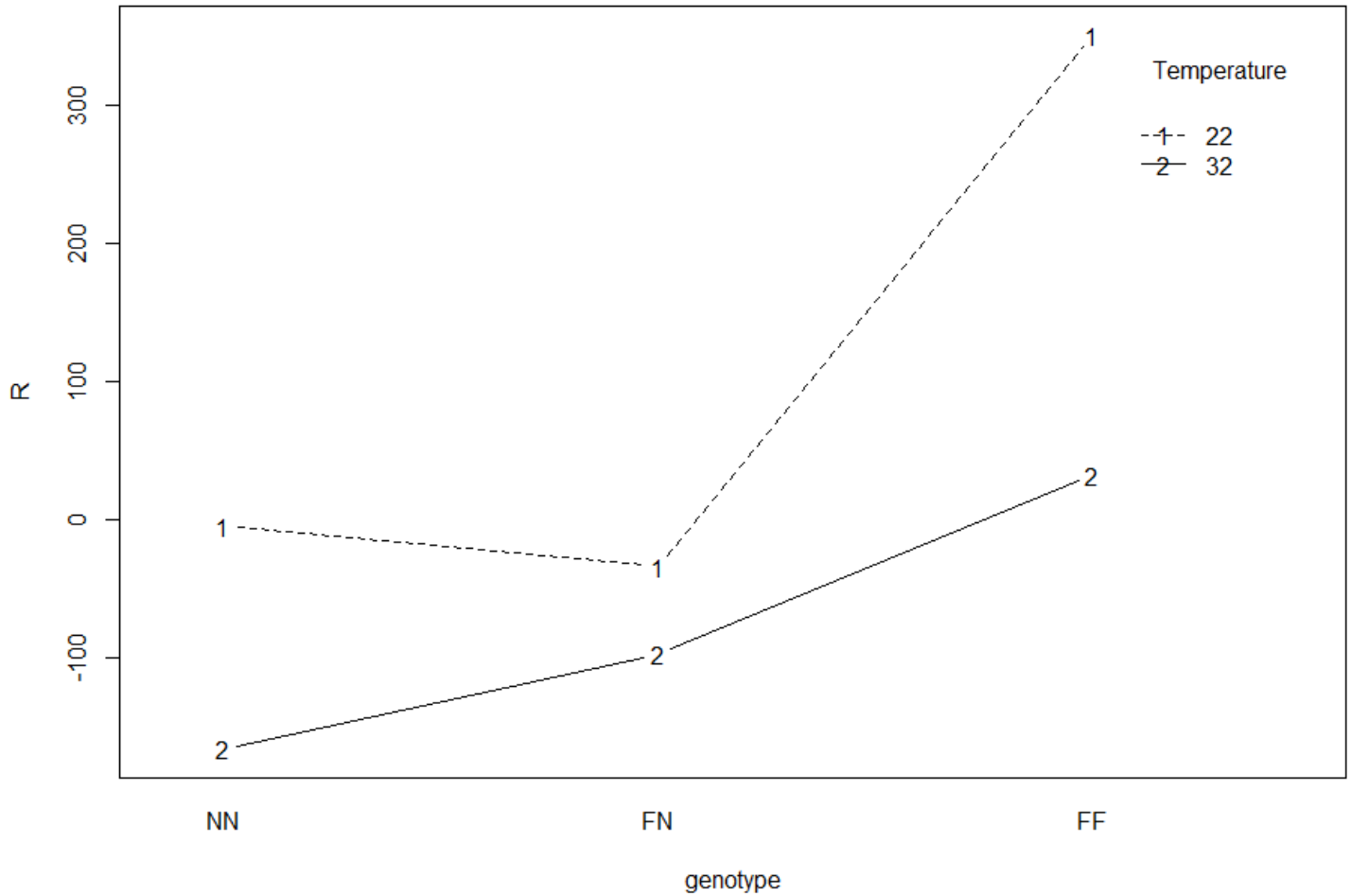
Les données

	sire	genotype	Temperature	R
1	30	FF	32	183.10
2	30	FF	32	672.40
3	30	FN	32	-247.60
4	30	NN	32	-326.30
5	30	FN	32	-356.80
6	30	FF	32	234.20

- Facteurs : *sire*,*genotype*,*Temperature* sont des variables qualitatives
- $+$ interactions

Interaction

interaction.plot



Sélection du modèle

$$Y_{ij} = Temperature_i + genotype_j + E_{ij}$$

Ajout du facteur "sire"

$$Y_{ijk} = Temperature_i + genotype_j + sire_k + E_{ijk}$$

add1(tg.lm,"sire",test="F")

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			12636167.74	2597.39		
sire	5	525882.94	12110284.79	2597.28	1.99	0.0812

Ajout de l'interaction "Température * génotype"

$$Y_{ij} = Temperature_i + genotype_j + Temperature * genotype_{ij} + E_{ij}$$

add1(tg.lm,"Temperature:genotype",test="F")

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			12636167.74	2597.39		
Temperature:genotype	2	658097.10	11978070.64	2588.66	6.37	0.0020

Sélection automatique du modèle

fonction "step"

$$Y_{ijk} = Temperature_i + genotype_j + E_{ijk}$$

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
+ Temperature:genotype	2	658097	11978071	2588.7	6.3733	0.002021
+ sire	5	525883	12110285	2597.3	1.9888	0.081174
<none>			12636167.74	2597.39		
- Temperature	1	1948816	14584984	2629.5	36.0887	7.157e-09
- genotype	2	3937020	16573188	2657.9	36.4534	1.654e-14

Step: AIC=2588.66

$$Y_{ij} = Temperature_i + genotype_j + Temperature * genotype_{ij} + E_{ijk}$$

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>			11978071	2588.7		
+ sire	5	479775	11498296	2588.9	1.8943	0.096233
- Temperature:genotype	2	658097	12636168	2597.4	6.3733	0.002021

Modèle retenu

$$Y_{ijk} = Temperature_i + genotype_j + Temperature * genotype_{ij} + E_{ij}$$

Anova

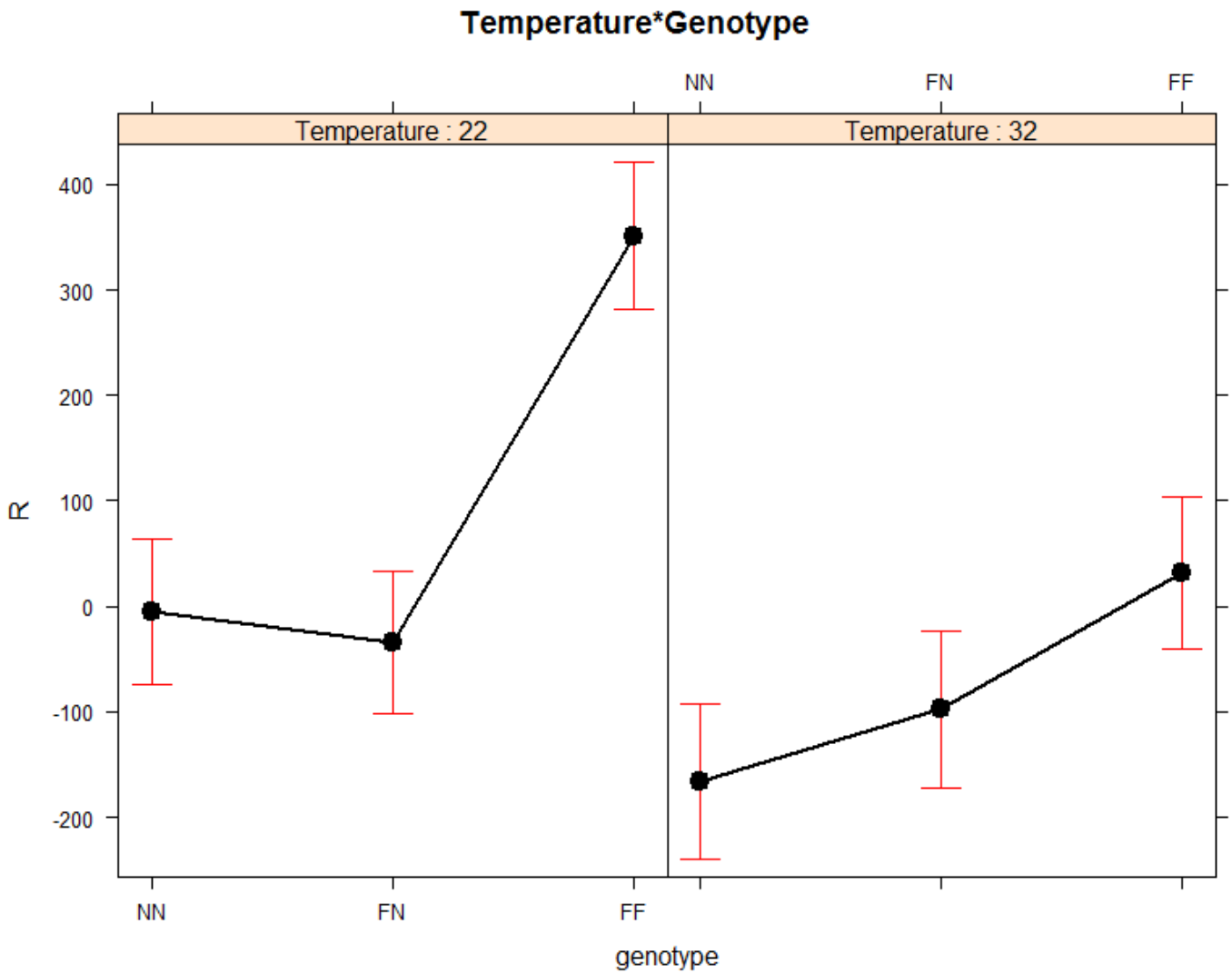
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Temperature	1	1841032.53	1841032.53	35.66	0.0000
genotype	2	3937020.11	1968510.06	38.13	0.0000
Temperature:genotype	2	658097.10	329048.55	6.37	0.0020
Residuals	232	11978070.64	51629.61		

Summary

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.6667	35.0610	-0.13	0.8942
Temperature32	-161.6333	51.2315	-3.15	0.0018
genotypeFF	355.7886	49.8852	7.13	0.0000
genotypeFN	-29.3629	49.0171	-0.60	0.5497
Temperature32:genotypeFF	-157.8334	72.4058	-2.18	0.0303
Temperature32:genotypeFN	97.9712	72.3341	1.35	0.1769

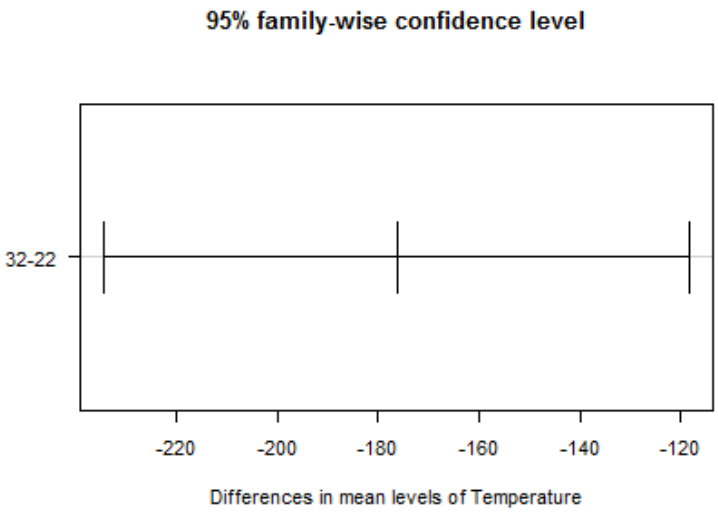
Visualisation de l'interaction

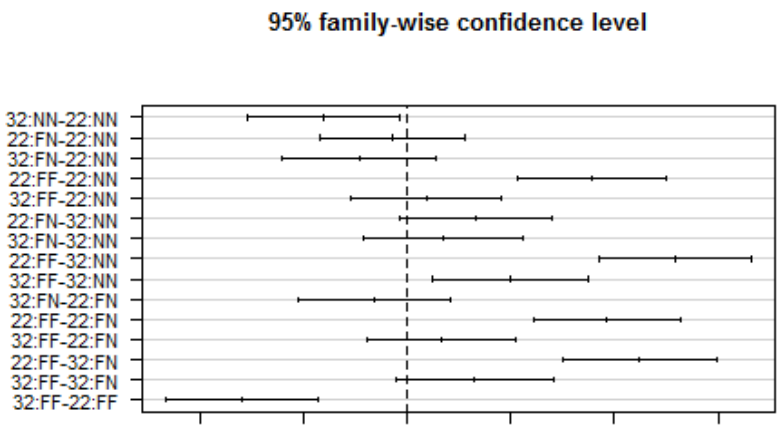
Package effects



Visualisation des effets et de leur significativité

Fonction TukeyHSD (Abdi et Williams, 2010)





References

- H Abdi et L Williams. Tukey's Honestly Significant Difference. *in Neil Salkind (Ed), Encyclopedia of Research Design, Thousand Oaks, CA : Sage, 2010*.
- J M Azais et J M Bardet. Le modèle linéaire par l'exemple. *Publications du laboratoire de Statistique et Probabilité, univ. Paul Sabatier, Toulouse III*.
- Documentation générale : http://pbil.univ-lyon1.fr/ade4/