Démarche statistique

Session 6 - Sélection de modèles

David Causeur
Institut Agro Rennes Angers
IRMAR UMR 6625 CNRS

05 janvier, 2025

Modèle linéaire

Une grande diversité de problématiques

Illustration :

- La valeur commerciale d'une carcasse de porc est définie à partir de son taux de viande maigre (LMP)
- Mesurer le LMP d'une carcasse de porc est très coûteux (plusieurs heures de dissection)
- On cherche donc à l'approcher au mieux à partir d'informations plus accessibles (épaisseurs de tissus gras et maigres)
- Pour construire une règle de prédiction du LMP, on dispose de mesures sur 354 carcasses

```
dta <- read.table("./data/pig10.txt",stringsAsFactors=TRUE)
str(dta)
```

```
'data.frame':   354 obs. of  10 variables:
 $ GENOTYPE   : Factor w/ 3 levels "P0","P25","P50": 3 3 1 3 3 1 1 1 1 1 ...
 $ SEX        : Factor w/ 2 levels "F","M": 2 1 1 1 2 2 1 2 1 2 ...
 $ SplitFat   : num  11.12 8.77 16.51 10.37 16.98 ...
 $ SplitMuscle: num  69.3 83.6 74.1 67.4 73.7 ...
 $ LV23Fat    : num  14.1 13.4 20.4 15.1 24.8 ...
 $ LR23Fat    : num  13.96 7.38 15.01 8.85 16.21 ...
 $ LR23Muscle : num  58.2 71.1 60.8 62.8 61 ...
 $ LR34Fat    : num  14.84 8.38 17.3 11.97 20.33 ...
 $ LR34Muscle : num  56.2 68.2 56.8 57.9 54.2 ...
 $ LMP        : num  81.8 87.3 79.1 84.1 76.1 ...
```

Effet groupe

Illustration : les teneurs en viande maigre moyennes par génotype sont-elles les mêmes ?

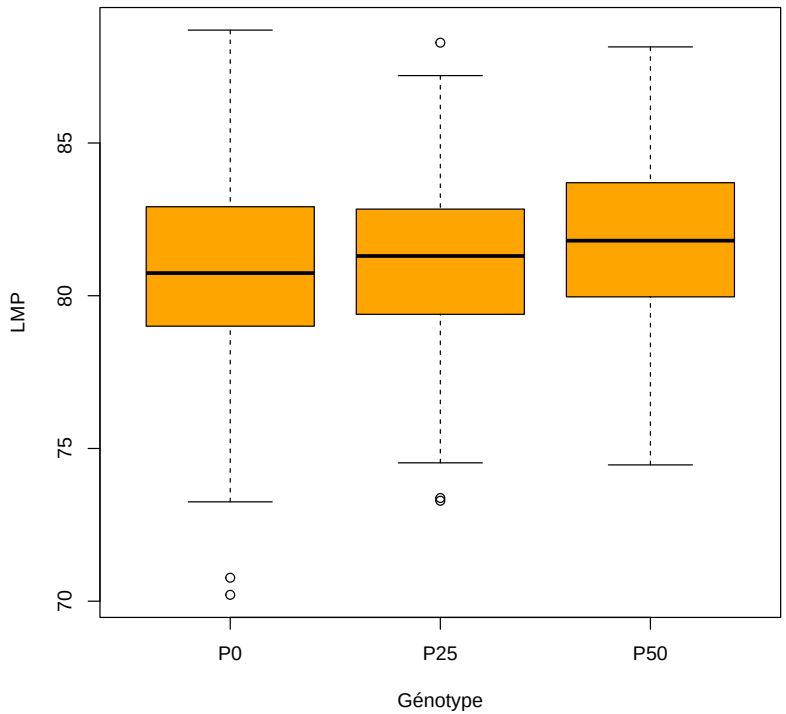
Analyse de la variance à un facteur

Pour $i = 1, 2, 3$, $j = 1, \ldots, n_i$,

$$Y_{ij} \quad = \quad \mu + \alpha_i + \varepsilon_{ij},$$

où

- Y_{ij} : teneur en viande maigre de la $j$ème carcasse de génotype i
- $\varepsilon_{ij} \sim \mathcal{N}(0; \sigma)$
- $\alpha_1 = 0$: μ est la teneur en viande maigre moyenne des carcasses de génotype P_0



Analyse de la variance à un facteur dans R

Ajustement du modèle

```
mod <- lm(LMP~GENOTYPE,data=dta)
```

Test de Fisher : 2 options (équivalentes)

```
anova(mod)
```

```
Analysis of Variance Table

Response: LMP
           Df Sum Sq Mean Sq F value  Pr(>F)
GENOTYPE    2   70.0  35.011  3.5553 0.02961
Residuals 351 3456.6   9.848
```

```
mod0 <- lm(LMP~1,data=dta)
anova(mod0,mod)
```

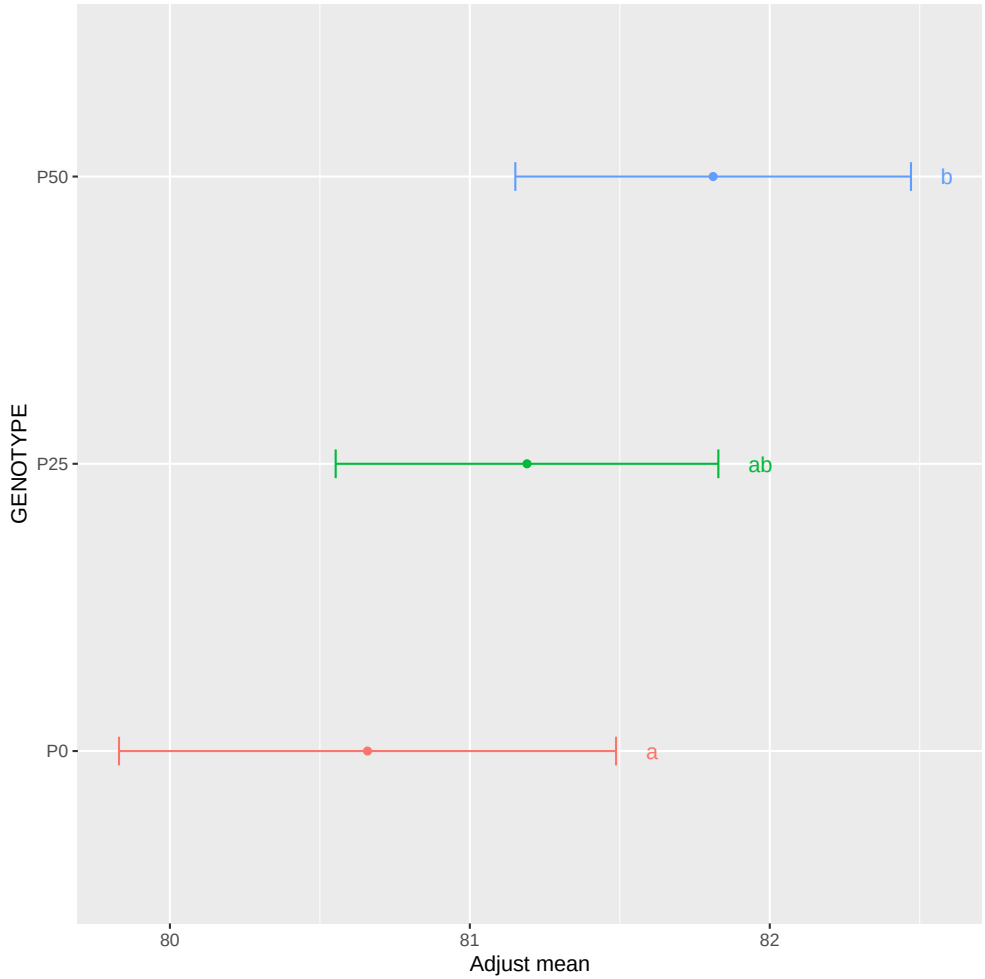
```
Analysis of Variance Table

Model 1: LMP ~ 1
Model 2: LMP ~ GENOTYPE
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    353 3526.6
2    351 3456.6  2    70.023 3.5553 0.02961
```

Description d'un effet groupe

Tests post-hoc : tests de Student de comparaison des génotypes par paires

```
posthoc <- meansComp(mod, ~ GENOTYPE,adjust="bonferroni",graph=TRUE)
```



Effet d'interaction entre deux variables catégorielles

Illustration : les différences entre les teneurs en viande maigre moyennes par génotype sont-elles les mêmes pour les mâles et les femelles ?

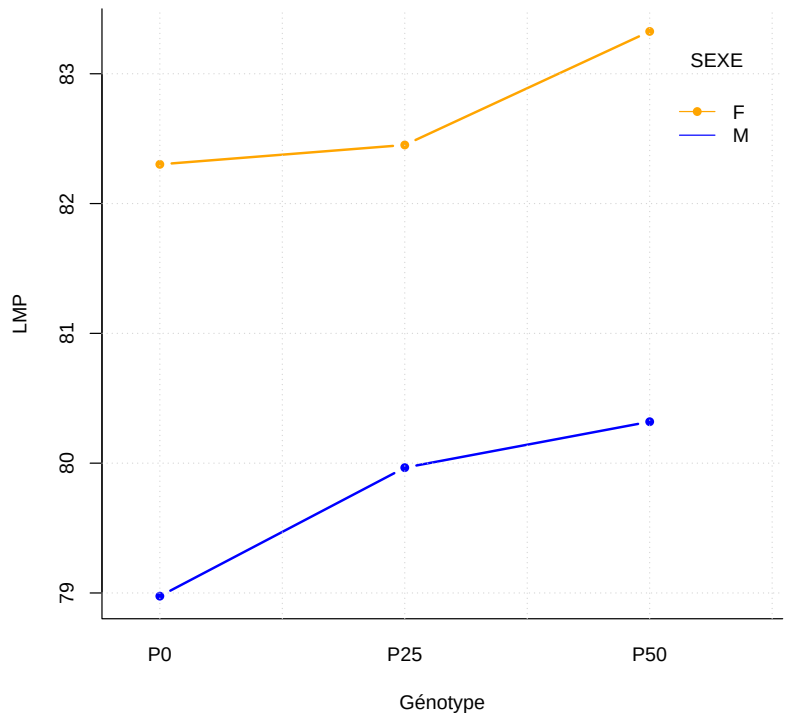
Analyse de la variance à deux facteurs

Pour $i = 1, 2, 3$, $j = 1, 2$, $k = 1, \ldots, n_{ij}$,

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk},$$

où

- Y_{ijk} : teneur en viande maigre de la $k$ème carcasse ayant le génotype i et de sexe j
- $\varepsilon_{ijk} \sim \mathcal{N}(0; \sigma)$
- $\alpha_1 = 0$, $\beta_1 = 0$
- $(\alpha\beta)_{1j} = 0$ pour $j = 1, 2$, $(\alpha\beta)_{i1} = 0$ pour $i = 1, 2, 3$



Analyse de la variance à deux facteurs dans R

Ajustement du modèle avec interaction

```
mod <- lm(LMP~SEX+GENOTYPE+GENOTYPE:SEX,data=dta)
anova(mod)
```

```
Analysis of Variance Table

Response: LMP
               Df  Sum Sq Mean Sq F value Pr(>F)
SEX             1  729.01  729.01 93.4530 <2e-16
GENOTYPE        2   72.81   36.40  4.6666  0.010
SEX:GENOTYPE    2   10.10    5.05  0.6474  0.524
Residuals     348 2714.68    7.80
```

Ajustement du modèle sans interaction

```
mod <- lm(LMP~SEX+GENOTYPE,data=dta)
anova(mod)
```

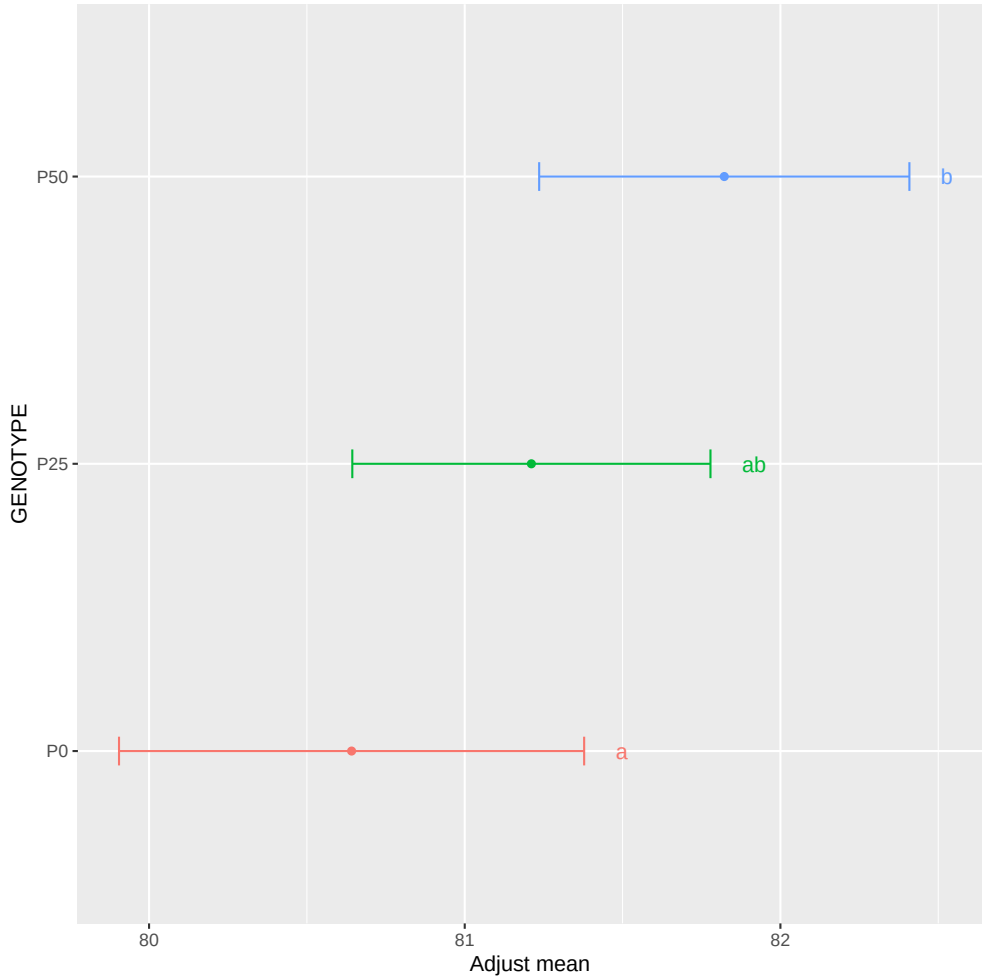
```
Analysis of Variance Table

Response: LMP
            Df  Sum Sq Mean Sq F value    Pr(>F)
SEX          1  729.01  729.01  93.642 < 2.2e-16
GENOTYPE     2   72.81   36.40   4.676  0.009906
Residuals  350 2724.78    7.79
```

Description d'un effet groupe par les moyennes ajustées

Tests post-hoc : tests de Student de comparaison des génotypes par paires

```
posthoc <- meansComp(mod, ~ GENOTYPE,adjust="bonferroni",graph=TRUE)
```



Effet linéaire

Illustration : la teneur en viande maigre d'une carcasse dépend-elle de son épaisseur de gras ?

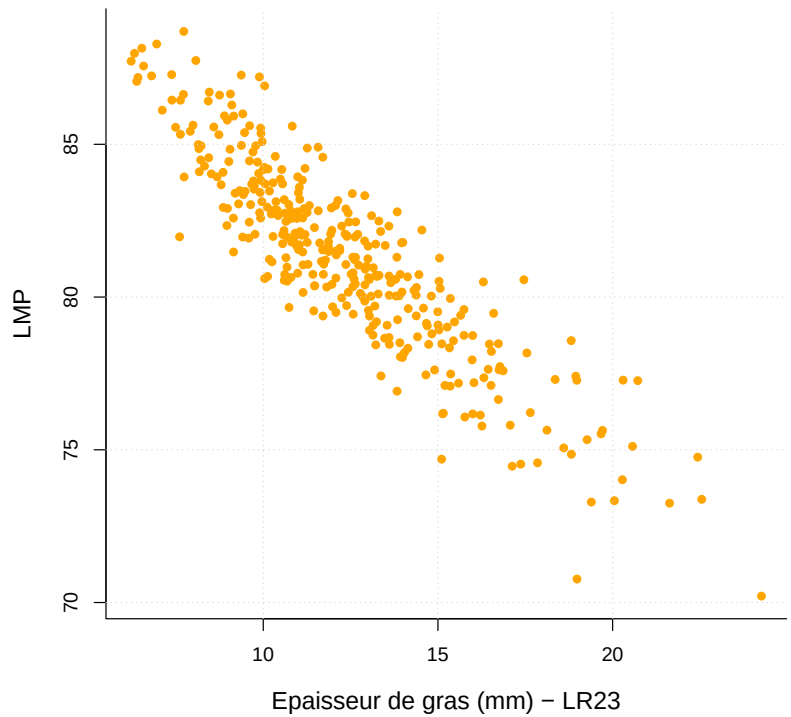
Régression linéaire simple

Pour $i = 1, \ldots, n$,

$$Y_i \quad = \quad \beta_0 + \beta_1 x_i + \varepsilon_i,$$

où

- Y_i : teneur en viande maigre de la $i$ème carcasse
- x_i : épaisseur de gras de la $i$ème carcasse
- $\varepsilon_i \sim \mathcal{N}(0; \sigma)$



Régression linéaire dans R

Ajustement du modèle

```
mod <- lm(LMP~LR23Fat,data=dta)
```

Test de Fisher : 2 options (équivalentes)

```
anova(mod)
```

```
Analysis of Variance Table

Response: LMP
           Df  Sum Sq Mean Sq F value    Pr(>F)
LR23Fat     1 2816.22 2816.22  1395.5 < 2.2e-16
Residuals 352  710.37    2.02
```

```
mod0 <- lm(LMP~1,data=dta)
anova(mod0,mod)
```

```
Analysis of Variance Table

Model 1: LMP ~ 1
Model 2: LMP ~ LR23Fat
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    353 3526.6
2    352  710.4  1    2816.2 1395.5 < 2.2e-16
```

Régression linéaire par groupes

Illustration : la relation entre la teneur en viande maigre d'une carcasse et son épaisseur de gras est-elle la même pour les mâles et les femelles ?

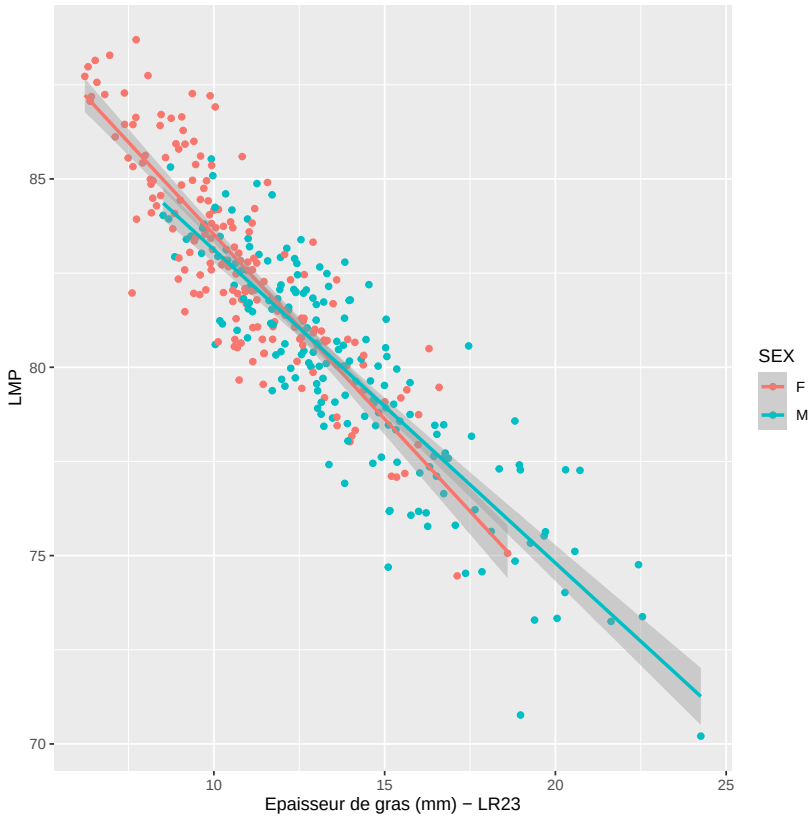
Modèle linéaire avec interaction entre l'épaisseur de gras et le sexe

Pour $i = 1, 2$, $j = 1, \ldots, n_i$,

$$Y_{ij} = \beta_0 + \alpha_i + (\beta_1 + \gamma_i) x_{ij} + \varepsilon_{ij},$$

où

- Y_{ij} : teneur en viande maigre de la $j$ème carcasse ayant le sexe i
- x_{ij} : épaisseur de gras de la $j$ème carcasse ayant le sexe i
- $\varepsilon_{ij} \sim \mathcal{N}(0; \sigma)$
- $\alpha_1 = 0$, $\gamma_1 = 0$



Régression linéaire par groupes dans R

Ajustement du modèle

```
mod <- lm(LMP~LR23Fat+SEX+LR23Fat:SEX,data=dta)
```

Test de Fisher :

```
options(width = 300)
anova(mod)
```

```
Analysis of Variance Table

Response: LMP
              Df  Sum Sq Mean Sq   F value    Pr(>F)
LR23Fat        1 2816.22 2816.22 1416.2404 < 2.2e-16
SEX            1    0.83    0.83    0.4174  0.518663
LR23Fat:SEX    1   13.56   13.56    6.8207  0.009399
Residuals    350  695.98    1.99
```

Choix de modèles

Choix entre deux variables explicatives

Illustration : quelle épaisseur de gras pour expliquer les variations de la teneur en viande maigre, LR23 ou LR34 ?

Modèles avec une variable explicative

```
fat_1 <- lm(LMP~LR34Fat,data=dta)
anova(fat_1)
```

```
Analysis of Variance Table

Response: LMP
           Df  Sum Sq Mean Sq F value    Pr(>F)
LR34Fat     1 2601.63 2601.63  990.07 < 2.2e-16
Residuals 352  924.96    2.63
```

```
fat_2 <- lm(LMP~LR23Fat,data=dta)
anova(fat_2)
```

```
Analysis of Variance Table

Response: LMP
           Df  Sum Sq Mean Sq F value    Pr(>F)
LR23Fat     1 2816.22 2816.22  1395.5 < 2.2e-16
Residuals 352  710.37    2.02
```

Conclusion : l'épaisseur de gras mesurée en LR23 explique mieux les variations de LMP que celle mesurée en LR34

Choisir entre deux variables explicatives ou garder les deux ?

Modèle avec deux variables explicatives

```
fat_12 <- lm(LMP~LR34Fat+LR23Fat,data=dta)
anova(fat_12)
```

```
Analysis of Variance Table

Response: LMP
           Df  Sum Sq Mean Sq F value    Pr(>F)
LR34Fat     1 2601.63 2601.63 1292.94 < 2.2e-16
LR23Fat     1  218.69  218.69  108.68 < 2.2e-16
Residuals 351  706.27    2.01
```

```
Anova(fat_12)
```

```
Anova Table (Type II tests)

Response: LMP
          Sum Sq  Df F value Pr(>F)
LR34Fat     4.10   1   2.038 0.1543
LR23Fat   218.69   1 108.682 <2e-16
Residuals 706.27 351
```

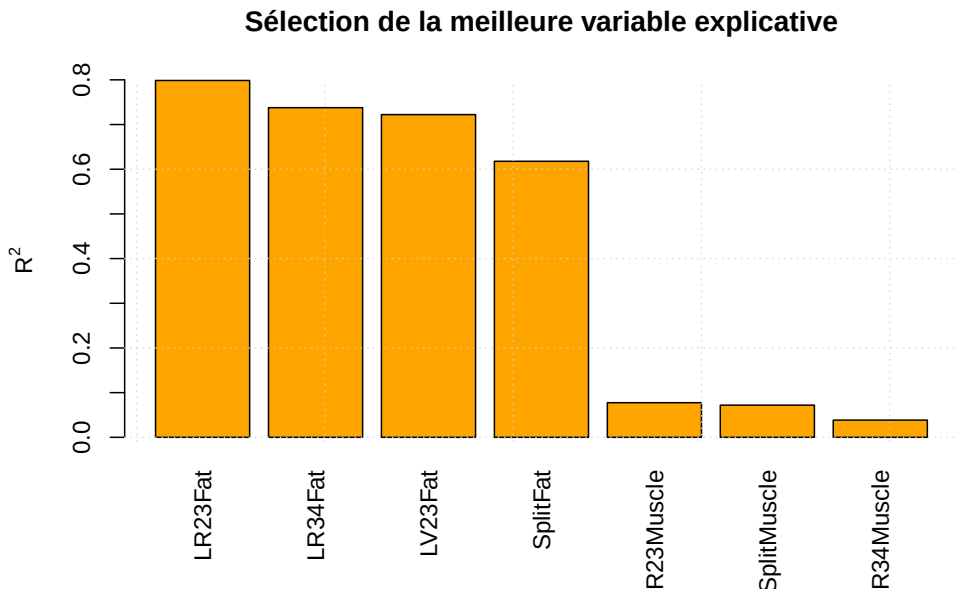
Deux conclusions antagonistes

- Le modèle est mieux ajusté (sommes des carrés des résidus plus faible)
- Une seule des deux épaisseurs de gras est suffisante (LR23) pour expliquer les variations de la teneur en viande maigre

Choix de la meilleure variable explicative

Illustration : quelle épaisseur de gras ou de muscle explique le mieux les variations de la teneur en viande maigre ?

```
R2 <- cor(dta$LMP,dta[,-c(1:2,10)])^2
ord <- order(R2,decreasing=TRUE)
barplot(R2[ord],col="orange",names.arg = colnames(R2)[ord],las=3,
        ylab=expression(R^2),yaxp=c(0,0.8,8),
        main="Sélection de la meilleure variable explicative")
grid()
```

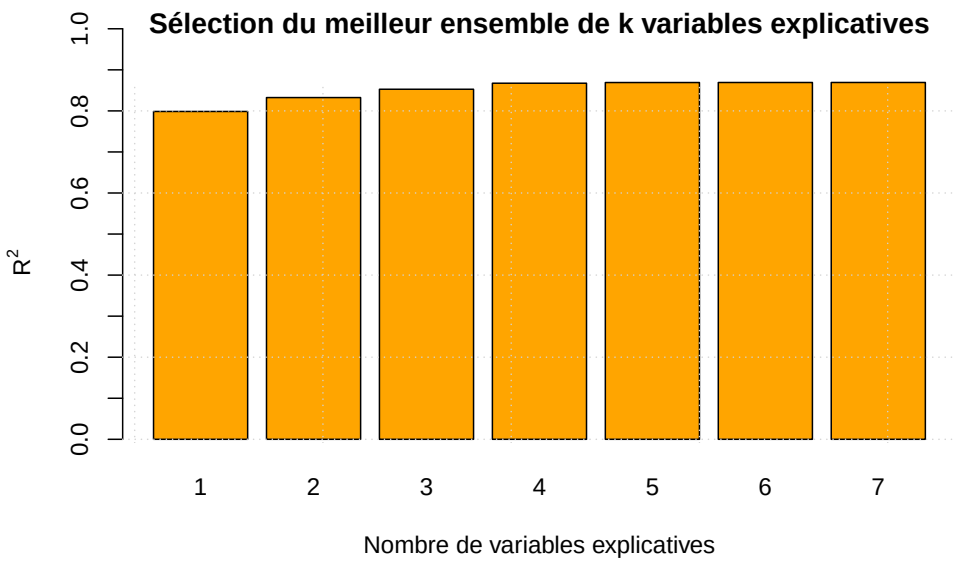


Choix du meilleur ensemble de k variables explicatives

Quel modèle $\mathcal{M}_k$ avec $k \leq K$ variables explicatives est suffisant pour expliquer les variations de teneurs en viande maigre ?

Choix parmi 2^K modèles (ici, $K = 7$, soit 128 modèles)

```
best <- regsubsets(LMP~.,data=dta[,-(1:2)],nvmax=7)
R2 <- summary(best)$rsq
barplot(R2,col="orange",xlab="Nombre de variables explicatives",
        names.arg = 1:7,ylab=expression(R^2),yaxp=c(0,1,10),
        main="Sélection du meilleur ensemble de k variables explicatives")
grid()
```



Meilleurs modèles à k variables explicatives

Liste des meilleurs modèles à k variables explicatives, $k = 1, \ldots, 7$

```
options(width=300)
summary(best)$which
```

```
  (Intercept) SplitFat SplitMuscle LV23Fat LR23Fat LR23Muscle LR34Fat LR34Muscle
1        TRUE    FALSE       FALSE   FALSE    TRUE       FALSE   FALSE       FALSE
2        TRUE     TRUE       FALSE   FALSE    TRUE       FALSE   FALSE       FALSE
3        TRUE     TRUE        TRUE   FALSE    TRUE       FALSE   FALSE       FALSE
4        TRUE     TRUE        TRUE    TRUE    TRUE       FALSE   FALSE       FALSE
5        TRUE     TRUE        TRUE    TRUE    TRUE        TRUE   FALSE       FALSE
6        TRUE     TRUE        TRUE    TRUE    TRUE        TRUE   FALSE        TRUE
7        TRUE     TRUE        TRUE    TRUE    TRUE        TRUE    TRUE        TRUE
```

Comment choisir parmi ces 7 modèles ? Combien de variables explicatives doit-on garder ?

Compromis entre qualité d'ajustement et complexité du modèle

Modèle linéaire à k variables explicatives et p_k paramètres

$$\text{Pour } i = 1, \ldots, n, \;\; Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i, \text{ où } \varepsilon_i \sim \mathcal{N}(0; \sigma)$$

Critères d'information

- Akaike Information Criterion (AIC), pour prédire

$$\text{AIC} \quad = \quad n\log\left(\frac{RSS}{n}\right) + 2p_k$$

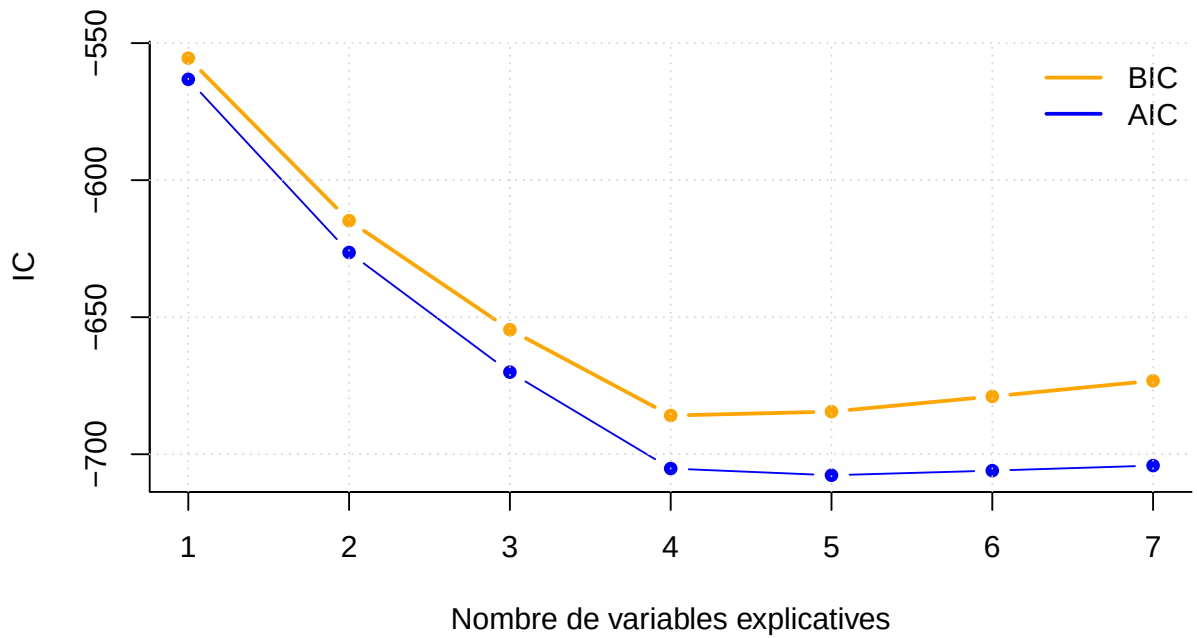
- Bayesian Information Criterion (BIC), pour expliquer

$$\text{BIC} \quad = \quad n\log\left(\frac{RSS}{n}\right) + \log(n)p_k$$

Remarque : comme $\log(n) > 2$ en général, le modèle ayant le BIC le plus faible contient moins de variables que celui ayant l'AIC le plus faible

Optimisation des critères d'information dans R

```
bic <- summary(best)$bic
aic <- bic+(2-log(nrow(dta)))*(2:8)
plot(1:7,bic,pch=16,bty="l",lwd=2,type="b",col="orange",ylab="IC",
     ylim=range(c(aic,bic)),xlab="Nombre de variables explicatives")
points(1:7,aic,type="b",pch=16,col="blue")
legend("topright",bty="n",col=c("orange","blue"),lwd=2,
       legend=c("BIC","AIC"))
grid()
```



Remarque : la 2ème meilleure variable explicative (LR34Fat) n'est pas choisie

Modèles optimaux

Modèle ayant le BIC le plus faible

```
best_bic <- lm(LMP~SplitFat+SplitMuscle+LV23Fat+LR23Fat,data=dta)
Anova(best_bic)
```

```
Anova Table (Type II tests)

Response: LMP
            Sum Sq  Df F value    Pr(>F)
SplitFat     74.04   1  55.258 8.240e-13
SplitMuscle  86.40   1  64.484 1.507e-14
LV23Fat      51.77   1  38.638 1.455e-09
LR23Fat     170.30   1 127.097 < 2.2e-16
Residuals   467.64 349
```

Modèle ayant le BIC le plus faible

```
best_aic <- lm(LMP~SplitFat+SplitMuscle+LV23Fat+LR23Fat+LR23Muscle,data=dta)
Anova(best_aic)
```

```
Anova Table (Type II tests)

Response: LMP
            Sum Sq  Df  F value    Pr(>F)
SplitFat     79.07   1  59.5881 1.252e-13
SplitMuscle  46.33   1  34.9122 8.219e-09
LV23Fat      49.50   1  37.3004 2.708e-09
LR23Fat     160.59   1 121.0198 < 2.2e-16
LR23Muscle    5.84   1   4.3973   0.03672
Residuals   461.80 348
```

Méthodes alternatives de recherche du meilleur modèle

Méthodes dites pas-à-pas (stepwise) si $K > 50$

Recherche ascendante

- **Etape 1** : $\mathcal{M}_1^\star$, meilleur modèle à une variable explicative
- **Etape** k : $\mathcal{M}_k^\star$, meilleur modèle parmi ceux complétant $\mathcal{M}_{k-1}^\star$ en ajoutant une variable explicative.
- **Stop** si le BIC de $\mathcal{M}_k^\star$ est plus grand que celui de $\mathcal{M}_{k-1}^\star$.

Recherche descendante

- **Etape 1** : $\mathcal{M}_K^\star$, modèle contenant toutes les variables explicatives
- **Etape** k : $\mathcal{M}_{K-k+1}^\star$, meilleur modèle parmi ceux obtenus à partir de $\mathcal{M}_{K-k+2}^\star$ en enlevant une variable explicative.
- **Stop** si le BIC de $\mathcal{M}_{K-k+1}^\star$ est plus grand que celui de $\mathcal{M}_{K-k+2}^\star$.

Méthodes pas-à-pas dans R

Recherches exhaustive, ascendante et descendante

```
best_exh <- regsubsets(LMP~.,data=dta[,-(1:2)],nvmax=7,
                       method="exhaustive")
best_fwd <- regsubsets(LMP~.,data=dta[,-(1:2)],nvmax=7,
                       method="forward")
best_bwd <- regsubsets(LMP~.,data=dta[,-(1:2)],nvmax=7,
                       method="backward")
```

Remarques :

- Dans le cas présent, les trois options de recherche conduisent au même résultat
- Toutefois, en général, la recherche pas-à-pas ne garantit pas de trouver le meilleur modèle

Ce qu'il faut retenir

Messages principaux

Lorsque la problématique cible précisément l'effet d'une variable explicative

- Recenser les confusions potentielles avec l'effet de cette variable explicative et les effets d'interaction avec d'autres variables explicatives
- Procéder par des tests de Fisher de comparaison de modèles

Lorsque la problématique ne cible pas en particulier une variable explicative

- Comparer les modèles construits à partir de tous les sous-ensembles possibles de variables explicatives
- Choisir le modèle optimisant un critère d'information (AIC ou BIC) pour réaliser le meilleur compromis entre qualité d'ajustement et complexité du modèle

Pour aller plus loin : voir le principe philosophique du rasoir d'Ockham