Modèle linéaire

Pierre Gloaguen

16 novembre 2018

Objectifs

- Expliquer les variations d'une variable **quantitative**:
  - Un rendement, une abondance, un taux d'une substance...
- En fonctions d'autres variables:
  - Un fertilisant, une région, un apport chimique...

Avantages

- Formulation mathématique simple permettant de connaître ses propriétés.
- Bonne représentation (en première approximation) de nombreux phénomènes.

Cas d'étude: Rendement de maïs

- On souhaite expliquer le **rendement** de plants de maïs.
- On dispose de 288 parcelles (on a désormais enlevé 11 parcelles).
- Sur chaque parcelle, le maïs a un même *marqueur génétique*:
  - Soit un marqueur de type 1;
  - Soit un marqueur de type 2;
- Sur chaque parcelle, le maïs a une même *variété*:
  - Corn, European, Northern, Tropical.
- Sur chaque parcelle, on mesure différentes caractéristiques:
  - Le **rendement** de la parcelle;
  - La teneur moyenne en *huile* d'un grain de maïs;
  - La teneur moyenne en *proteine* d'un grain de maïs;
  - La teneur moyenne en *amidon* d'un grain de maïs;
  - Le nombre de degrés-jours moyen avant la *floraison* d'un plant de maïs;
  - Le nombre moyen de *feuilles* par plant de maïs;
- Quelles *variables explicatives* donnent des informations sur le **rendement**?

Principes du modèle linéaire

- Modèle mathématique décrivant le lien entre une variable explicative **quantitative** (le rendement) et des variables explicatives (la variété, la teneur en huile,. . .)
- Modèle décrit dans un cadre probabiliste décrivant l'aléa (part non prédite).

Principe d'application

1. Question biologique;
2. Ecriture du modèle;
3. Ajustement (estimation) du modèle grâce aux données;
4. Vérification de la validité des hypothèses faites dans le modèle;
5. Test de la pertinence du modèle linéaire par rapport à un modèle simple;
6. Test de la pertinence des différents éléments du modèle;
7. Critique du modèle;
8. Conclusion sur la question biologique.

Analyse de la variance à 2 facteurs

I) Question biologique

Question biologique: Le **rendement** d'une espèce peut il être expliqué par sa *variété* et son *marqueur génétique*?.

- Le **rendement** est la variable à expliquer;
- La *variété* et le *marqueur* sont des variables explicatives. Ce sont des variables *qualitatives*.
- **Cadre de l'ANOVA 2 facteurs** 1 **variable à expliquer**, quantitative, 2 *variable explicatives*, qualitatives.
- **Première étape:** Une approche descriptive.

Pour cet exemple, on a enlevé la variété Stiff Stalk

Coefficient de corrélation linéaire empirique

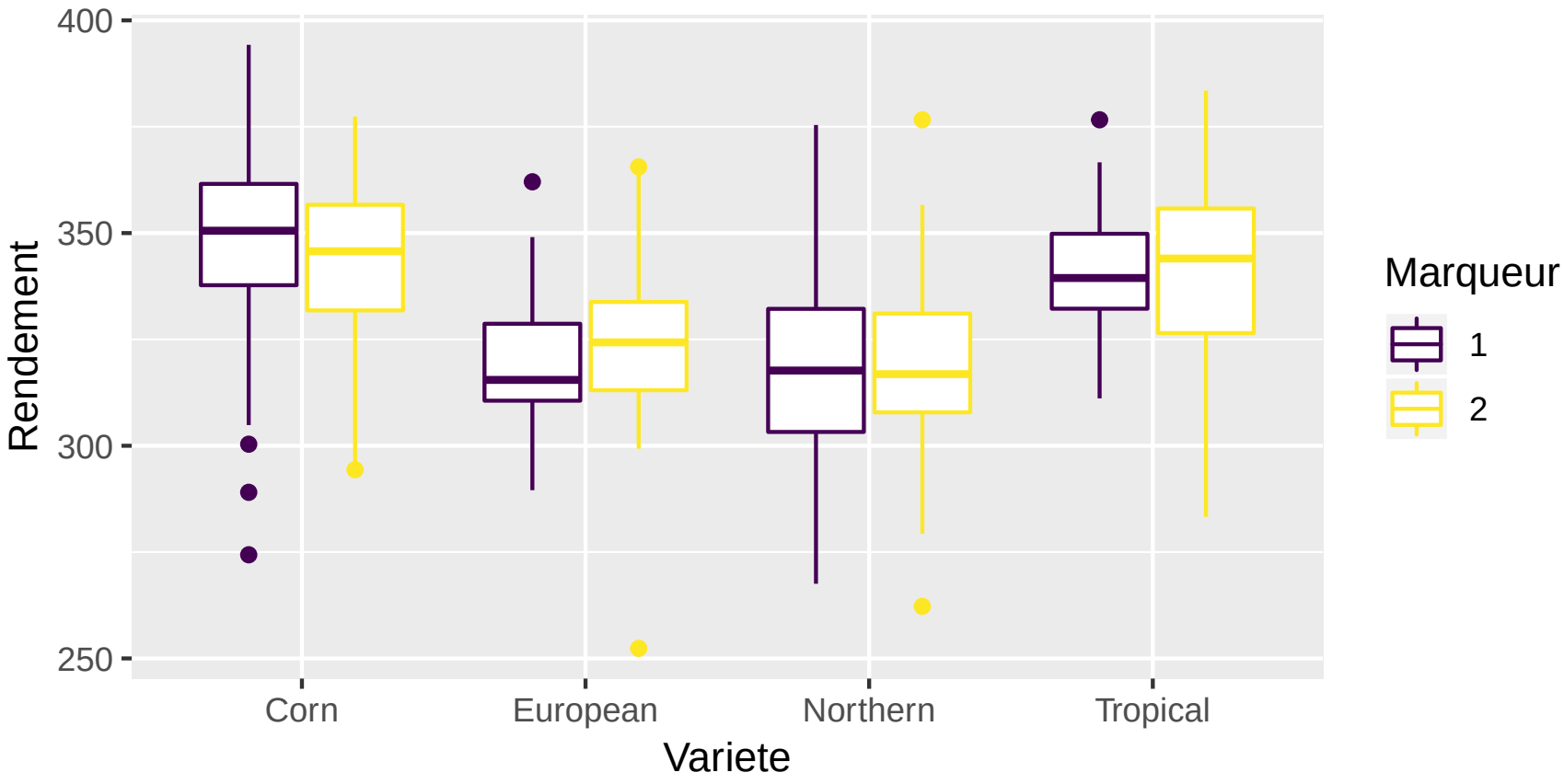
Effectifs et moyennes par variété

Marqueur	Corn	European	Northern	Tropical
	Variete			
1	45	19	30	8
2	72	37	20	46

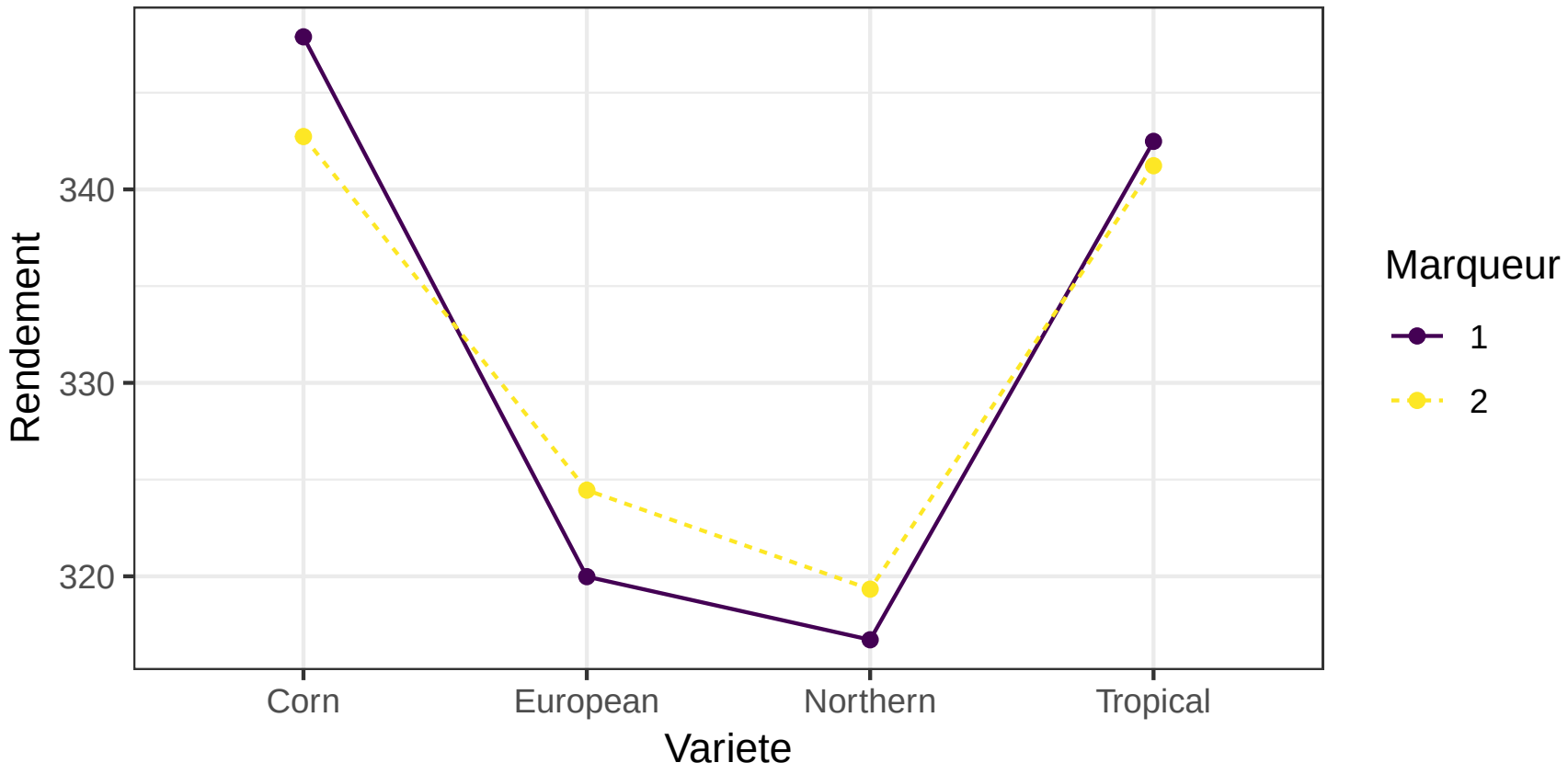
	Variete	Marqueur	Rendement
1	Corn	1	347.8866
2	European	1	319.9829
3	Northern	1	316.7196
4	Tropical	1	342.4842
5	Corn	2	342.7325
6	European	2	324.4547
7	Northern	2	319.3362
8	Tropical	2	341.2252

Visualisation graphique

Question biologique: Le **rendement** d'une espèce peut il être expliqué par sa *variété* et son marqueur génétique?.



Interactions



II) **Ecriture du modèle** (régulier)

Notations

On a $n = 277$ observations. Le facteur Marqueur a $I = 2$ niveaux, codés ainsi: 1 (i=1), 2 (i=2). Le facteur Variete a $J = 4$ niveaux, codés ainsi: Corn (j=1), European (j=2), Northern (j=3), Tropical (j=4).

Pour chaque niveau croisement de niveaux (i, j), on dispose de n_{ij} observations avec

$n_{11} = 45, n_{12} = 19, n_{13} = 30, n_{14} = 8, n_{21} = 72, n_{22} = 37, n_{23} = 20, n_{24} = 46.$

On note y_{ijk} le rendement de la k-ième parcelle pour le Marqueur i et la Variete j.

Modèle

On suppose que y_{ijk} est la réalisation d'une V.A. Y_{ijk} telle que:

$$Y_{ijk} = \mu_{ij} + E_{ijk},\ 1 \leq i \leq I,\ 1 \leq j \leq J,\ 1 \leq k \leq n_{ij}$$

- μ_{ij} est la moyenne attendue de rendement pour le Marqueur i et la Variete j;
- E_{ijk} est le résidu (aléa) associé à l'observation Y_{ijk}. $E_{ijk} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

II) **Ecriture du modèle** (singulier)

Notations

On a $n = 277$ observations. Le facteur Marqueur a $I = 2$ niveaux, codés ainsi: 1 (i=1), 2 (i=2). Le facteur Variete a $J = 4$ niveaux, codés ainsi: Corn (j=1), European (j=2), Northern (j=3), Tropical (j=4).

Pour chaque croisement de niveaux (i, j), on dispose de n_{ij} observations avec

$n_{11} = 45, n_{12} = 19, n_{13} = 30, n_{14} = 8, n_{21} = 72, n_{22} = 37, n_{23} = 20, n_{24} = 46.$

On note y_{ijk} le rendement de la k-ième parcelle pour le Marqueur i et la Variete j.

Modèle

On suppose que y_{ijk} est la réalisation d'une V.A. Y_{ijk} telle que:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + E_{ijk},\ 1 \leq i \leq I,\ 1 \leq k \leq n_{ij}$$

- μ est la rendement moyen de référence;
- α_i est **l'effet principal** du Marqueur de niveau i.
- β_j est **l'effet principal** de la Variete de niveau j;
- γ_{ij} est **l'effet d'interaction** entre le Marqueur i et la Variete j;
- E_{ijk} est le résidu (aléa) associé à l'observation Y_{ijk}. $E_{ijk} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$.

III) Ajustement du modèle régulier

Estimateurs

- **Estimateurs**: Variables aléatoires L'estimateur de μ_{ij} est donné par la moyenne empirique du rendement des parcelles de Marqueur i et de Variete j.

$$\hat{\mu}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} Y_{ijk} = Y_{ij\bullet}$$

- **Estimations**: Réalisation sur les données

$$\hat{\mu}_{ij}^{obs} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk} = y_{ij\bullet}$$

Prédicteur et prédiction

- **Prédicteur** Pour une parcelle e Marqueur i et de Variete j, le rendement prédit est:

$$\widehat{Y}_{ijk} = \hat{\mu}_{ij} = Y_{ij\bullet}$$

- **Prédiction** Réalisation sur les données

$$\widehat{y}_{ik} = \hat{\mu}_{ij}^{obs} = y_{ij\bullet}$$

III) Ajustement du modèle singulier

Il y a trop de paramètres de moyennes! $(1 + I + J + IJ)$ alors qu'on a IJ moyennes!

On doit **poser** $1 + I + J$ **contraintes**, typiquement:

$$\alpha_1 = 0,\ \beta_1 = 0,\ \gamma_{1j} = \gamma_{i1} = 0,\ 1 \leq i \leq I,\ 1 \leq j \leq J.$$

Les 1ers niveaux de chaque facteur sont alors **des niveaux de référence**.

Les estimateurs des $\mu, \alpha_i, \beta_j, \gamma_{ij}$ vont **dépendre** de la contrainte!

Pour la contrainte ci dessus, on a:

$$\hat{\mu} = Y_{11\bullet},\ \hat{\alpha}_i = Y_{i1\bullet} - Y_{11\bullet},\ \hat{\beta}_j = Y_{1j\bullet} - Y_{11\bullet}, \hat{\gamma}_{ij} = Y_{ij\bullet} - Y_{i1\bullet} - Y_{1j\bullet} + Y_{11\bullet}$$

Le prédicteur est alors

$$\widehat{Y}_{ijk} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij} = Y_{ij\bullet}$$

Le prédicteur **ne dépend pas** de la contrainte.

Estimateur et estimation de la variance σ^2

Résidus observés

$$\hat{e}_{ijk} = y_{ijk} - \hat{y}_{ijk},\ 1 \leq i \leq I,\ 1 \leq j \leq J,\ 1 \leq k \leq n_{ij}$$

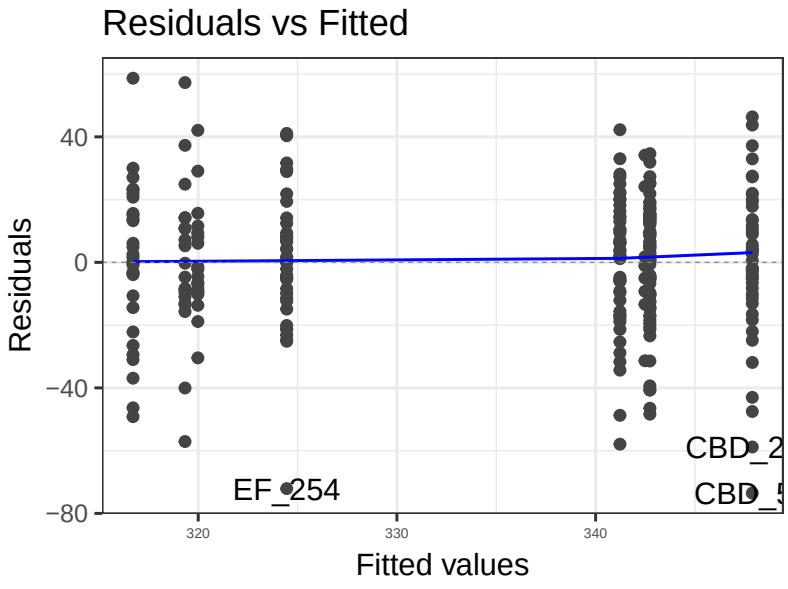
Estimateur

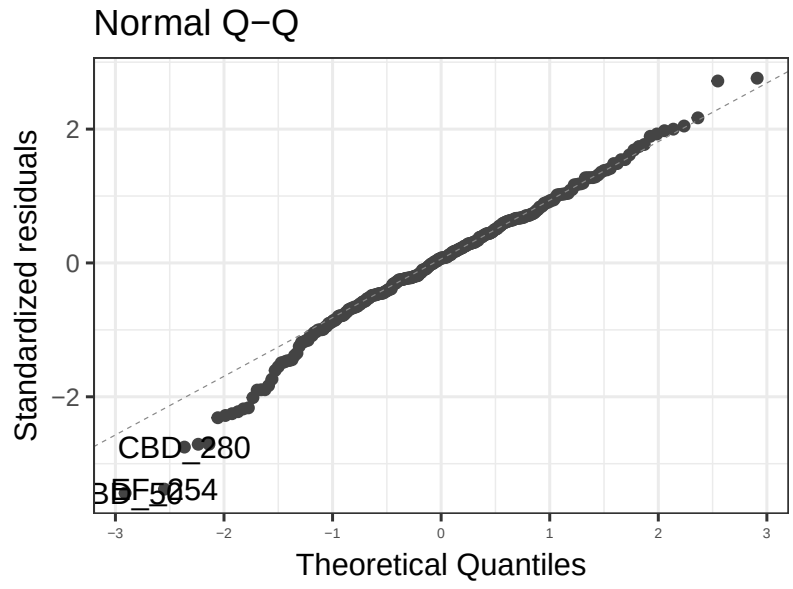
$$S^2 = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}(Y_{ijk} - \widehat{Y}_{ijk})^2}{n - IJ}$$

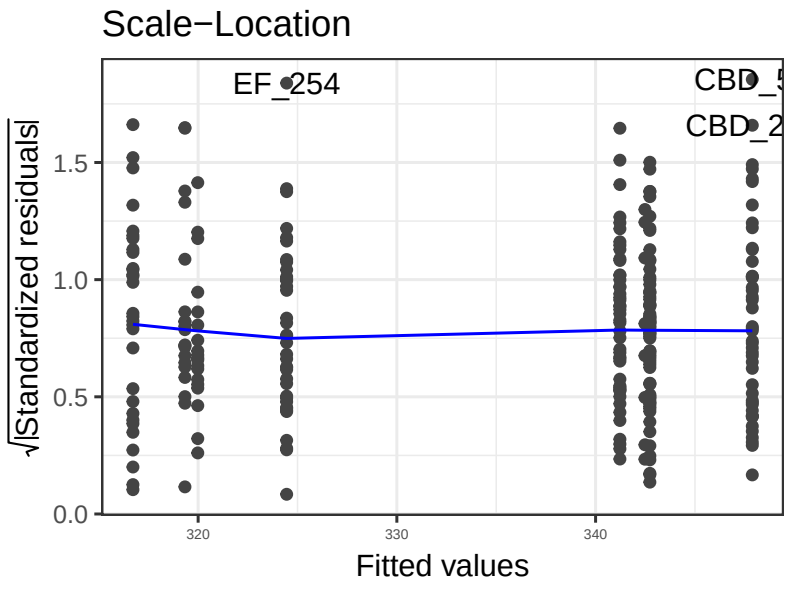
Estimation

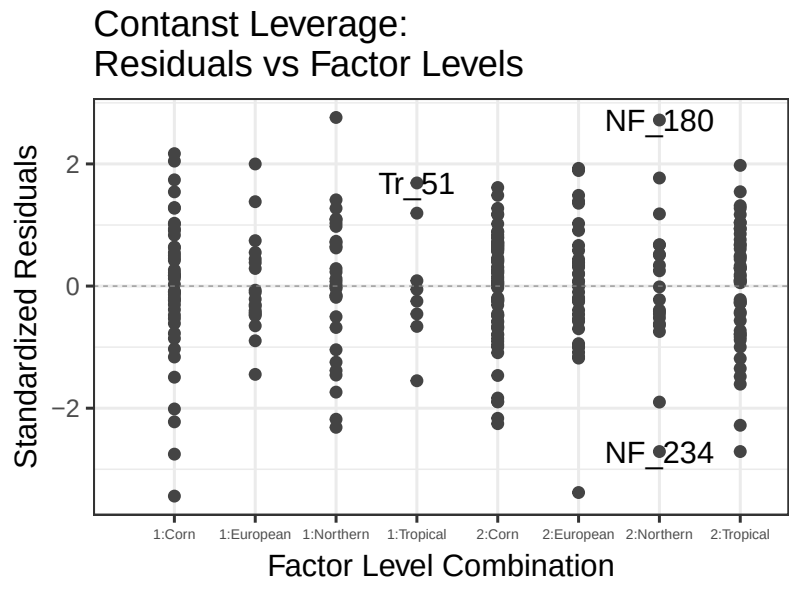
$$\hat{\sigma}^2_{obs} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}(y_{ijk} - \widehat{y}_{ijk})^2}{n - IJ} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}\widehat{e}^2_{ijk}}{n - IJ} \overset{ici}{=} 467.4$$

IV) **Validité des hypothèses**

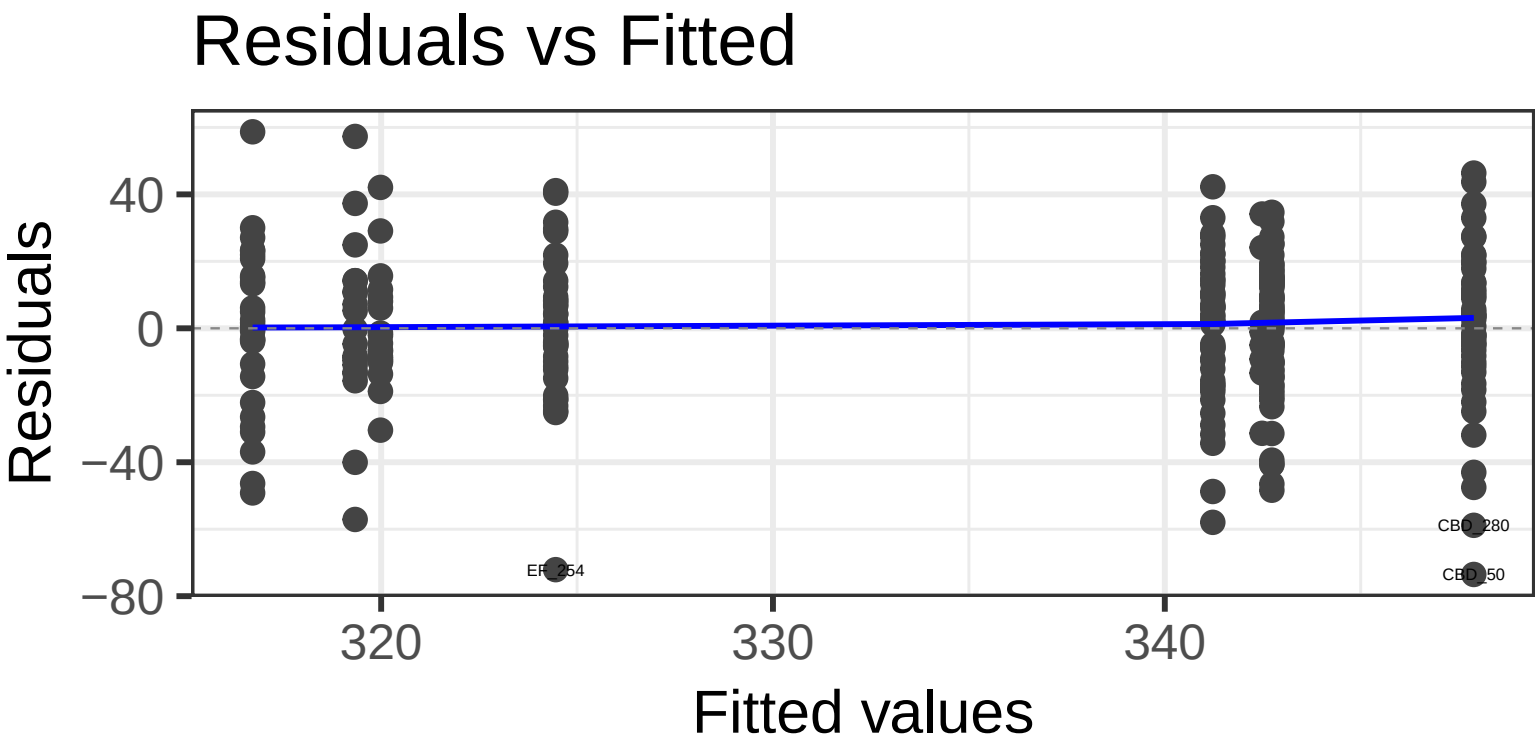








Distribution identique, espérance constante et nulle

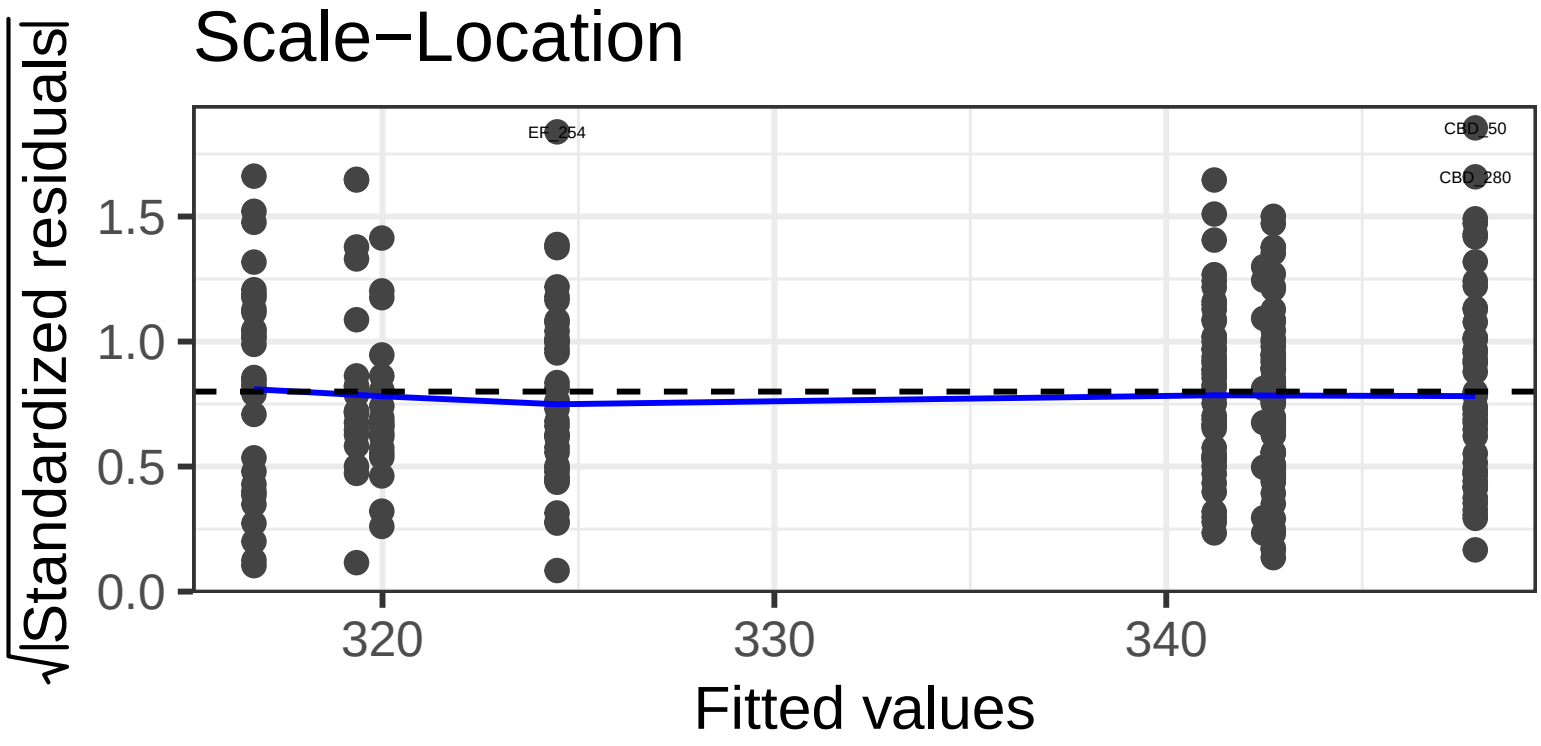


Ce qu'on regarde: Les résidus observés $\hat{e}_{ijk}$ en fonction des prédictions $\hat{y}_{ijk}$ (équivalent à regarder en fonction des croisements de niveaux).

Ce qu'on voit: La distribution des résidus semble comparable dans toutes les variétés. La variance ne semble pas différer selon le croisement de niveaux.

Ce qu'on conclut: L'hypothèse de distribution identique des résidus semble valable.

Distribution identique, variance constante

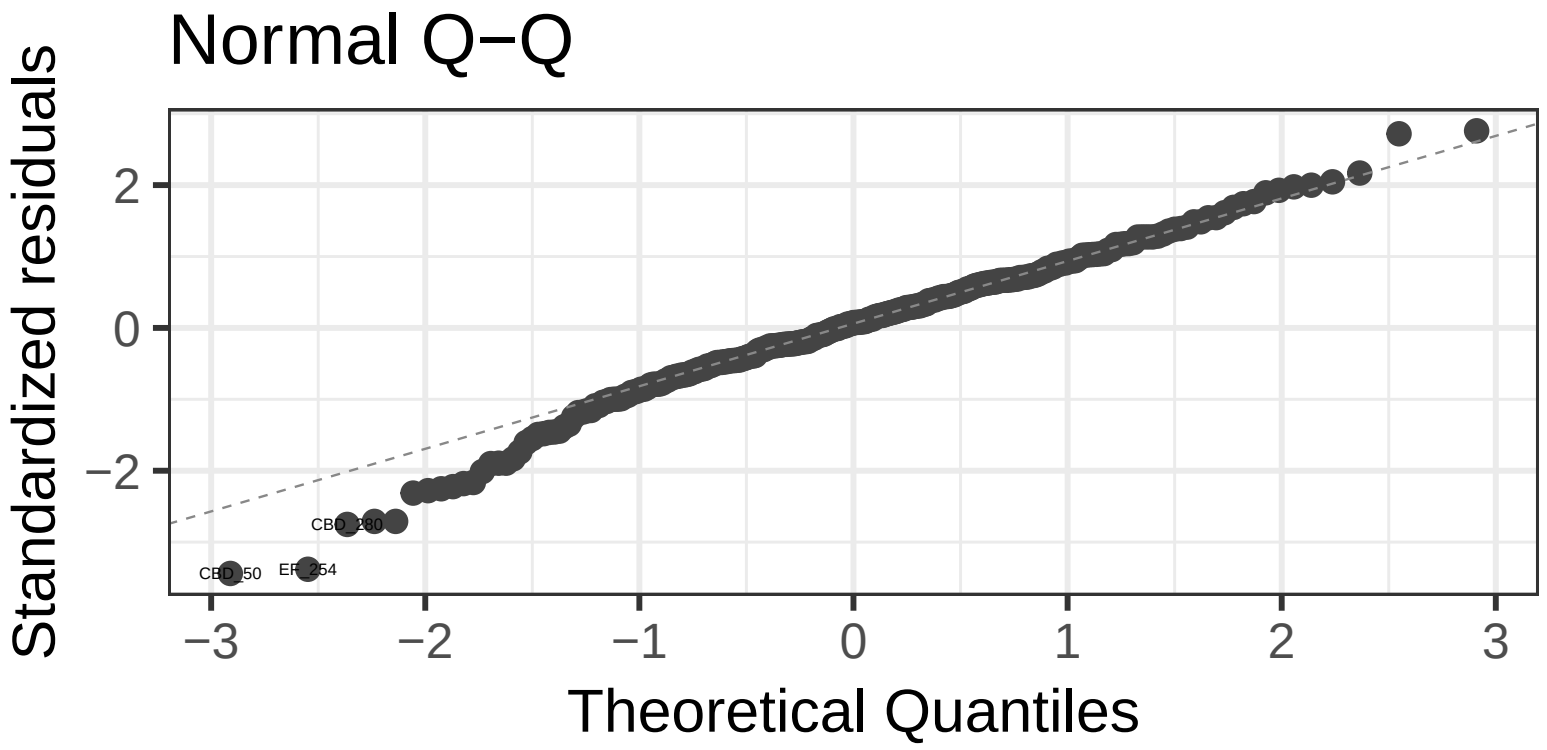


Ce qu'on regarde: La valeur absolue des résidus (standardisés) observés en fonction des prédictions $\hat{y}_k$ (équivalent à regarder en fonction des croisements de niveaux).

Ce qu'on voit: La valeur absolue des résidus semble comparable dans toutes les variétés (autour de 0.8). La variance ne semble pas différer selon le croisement de niveaux.

Ce qu'on conclut: L'hypothèse de variance identique des résidus semble valable.

Distribution normale

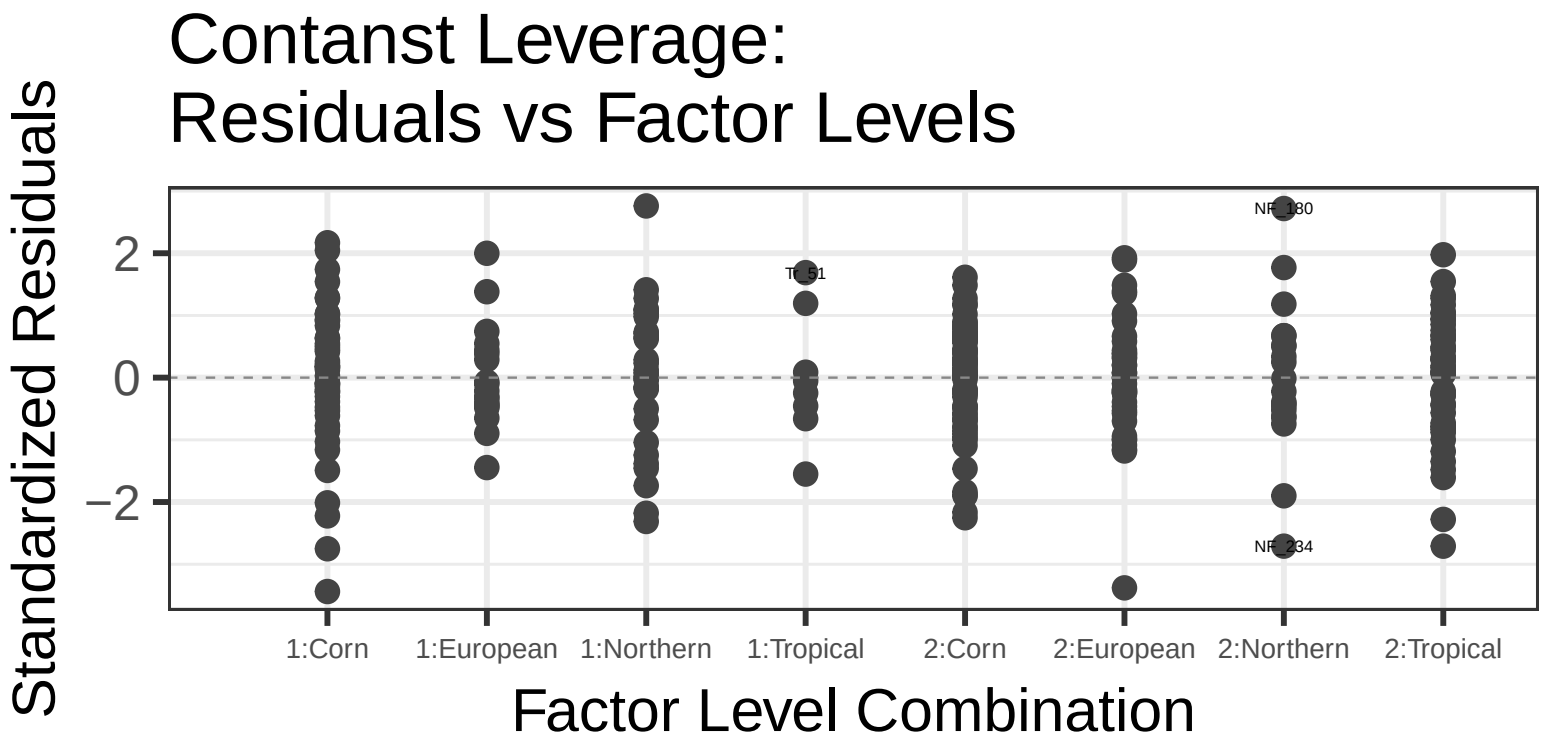


Ce qu'on regarde: La valeur des quantiles empiriques des résidus standardisés en fonction de la valeur quantiles théoriques d'une loi normale $\mathcal{N}(0, 1)$.

Ce qu'on voit: Les points sont globalement alignés sur la droite $y = x$, avec quelques problèmes pour les premiers quantiles.

Ce qu'on conclut: On peut valider l'hypothèse de distribution normale des résidus.

Résidus par variété

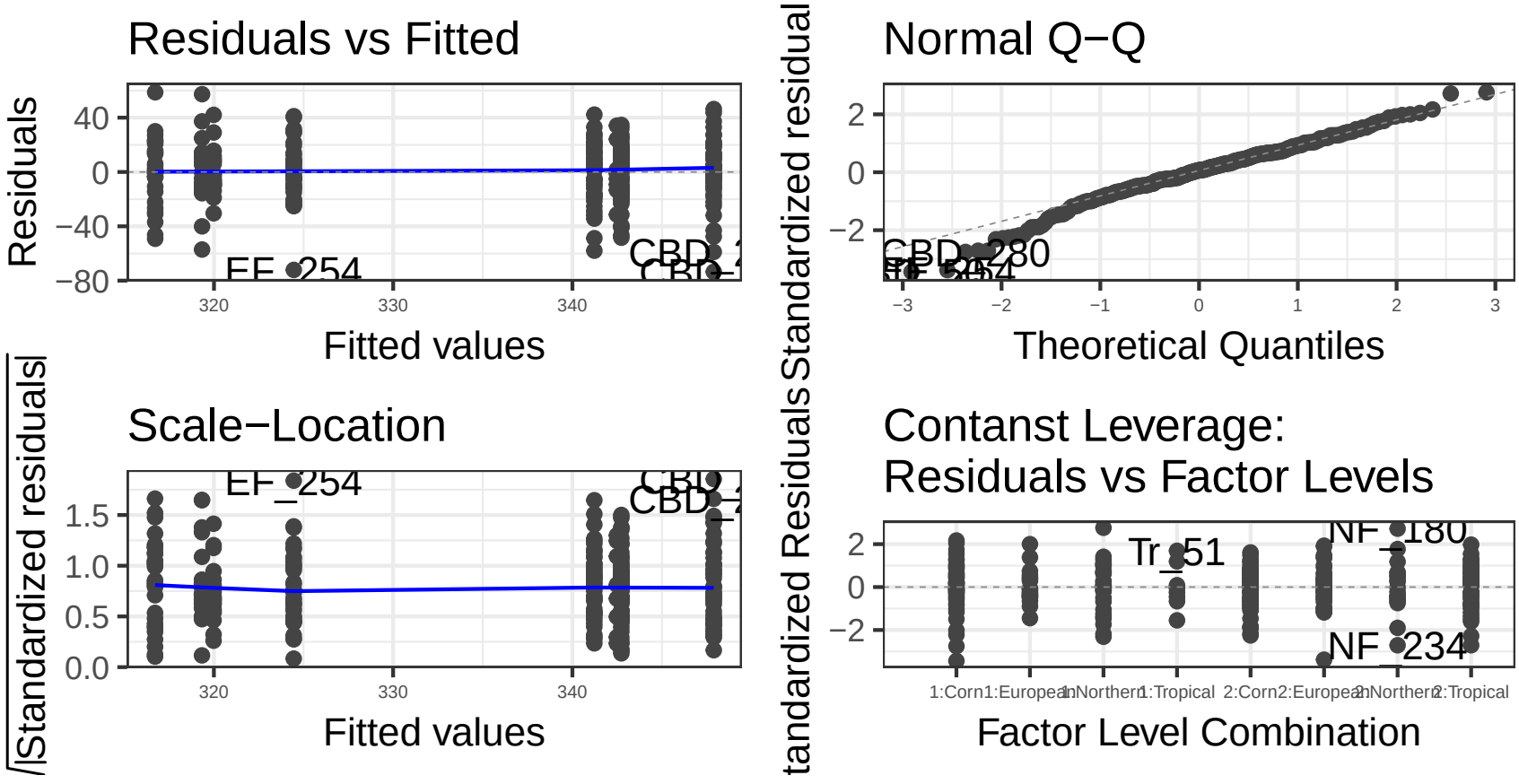


Ce qu'on regarde: Même graphique (réarrangé) que le 1er.

Ce qu'on voit: Même chose qu'au premier.

Ce qu'on conclut: Même chose qu'au premier.

4 graphes et un oeil fin



Rien dans ces quatre graphiques ne permet de contredire les hypothèses faites sur les résidus.

V) Test du modèle

On veut tester si notre modèle impliquant le niveau de Marqueur et de Variete du maïs explique mieux le **rendement** qu'un modèle simple, où le **rendement** est constant.

Hypothèses du test

On teste:

$$\begin{array}{rlll} & H_0 : & Y_{ijk} = \mu + E_{ijk} & \text{Modèle } M_0 \\ \text{contre} & H_1 : & Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + E_{ijk} & \text{Modèle } M_1 \end{array}$$

Pour tester cette hypothèse, on va décomposer la variabilité des données:

$$\overbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}(Y_{ijk}-\bar{Y})^2}^{SCT} = \overbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}(\widehat{Y}_{ijk}-\bar{Y})^2}^{SCM} + \overbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}}(Y_{ijk}-\widehat{Y}_{ijk})^2}^{SCR}$$

Somme	*ddl*	Réalisation
SCT:	n - 1	$SCT_{obs} = \sum_{i,j,k}(y_{ijk} - \bar{y})^2$
SCM:	IJ - 1	$SCM_{obs} = \sum_{i,j,k}(\hat{y}_{ijk} - \bar{y})^2$
SCR:	n - IJ	$SCR_{obs} = \sum_{i,j,k}(y_{ijk} - \hat{y}_{ijk})^2$

Test du modèle

Hypothèses du test

$$\begin{array}{rlll} & H_0 : & Y_{ijk} = \mu + E_{ijk} & \text{Modèle } M_0 \\ \text{contre} & H_1 : & Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + E_{ijk} & \text{Modèle } M_1 \end{array}$$

Statistique de test

On considère la statistique de test

$$F = \frac{SCM/ddl(SCM)}{SCR/ddl(SCR)}$$

Si H_0 est vraie, alors $F \overset{H_0}{\sim} Fisher(ddl(SCM), ddl(SCR))$.

Sur les données on observe

$$f_{obs} = \frac{SCM_{obs}/ddl(SCM)}{SCR_{obs}/ddl(SCR)}.$$

On rejette H_0 au risque de première espèce α si

$$\overbrace{\mathbb{P}(F > f_{obs})}^{\text{p-valeur}} < \alpha$$

Table d'analyse de la variance

ddl	Somme	*ddl*	Somme	Stat. test.	p-valeur
$ddl(SCT)$	SCT_{obs}				
$ddl(SCR)$	SCR_{obs}	$ddl(SCM)$	SCM_{obs}	f_{obs}	$\mathbb{P}(F > f_{obs})$

```
Analysis of Variance Table

Model 1: Rendement ~ 1
Model 2: Rendement ~ Marqueur * Variete
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    276 163064
2    269 125735  7     37329 11.409 1.135e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion du test: On rejette H_0 et on conclut que le modèle d'ANOVA explique mieux les données qu'un modèle où le rendement est constant.

REMARQUE: $\hat{\sigma}^2 = S^2 = SCR/ddl(SCR)$

VI) Tests des effets

Chacun des effets dans le modèle est il indispensable?

- Effet d'interaction?
- Effet principal du Marqueur?
- Effet principal de la Variete?

Test de l'effet d'interaction

On testera si l'ajout d'un effet d'interaction à un modèle avec les effets principaux apporte de l'information sur le rendement.

Test des effets principaux

2 manières de voir le problème?

- L'ajout de l'effet principal du facteur Marqueur est il intéressant par rapport à un modèle constant? (Test de type I)
- L'ajout de l'effet principal du facteur Marqueur est il intéressant par rapport à un modèle ayant l'effet Variete? (Test de type II)

Notion de réduction

Notations

Modèle	Equation	SCM
M_μ	$Y_{ijk} = \mu + E_{ijk}$	SCM_μ
$M_{\mu,\alpha}$	$Y_{ijk} = \mu + \alpha_i + E_{ijk}$	$SCM_{\mu,\alpha}$
$M_{\mu,\beta}$	$Y_{ijk} = \mu + \beta_j + E_{ijk}$	$SCM_{\mu,\beta}$
$M_{\mu,\alpha,\beta}$	$Y_{ijk} = \mu + \alpha_i + \beta_j + E_{ijk}$	$SCM_{\mu,\alpha,\beta}$
$M_{comp} = M_{\mu,\alpha,\beta,\gamma}$	$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + E_{ijk}$	$SCM_{\mu,\alpha,\beta,\gamma}$

Réduction

On appelle réduction associée α dans un modèle contenant μ la quantité:

$$R(\alpha|\mu) = SCM_{\mu,\alpha} - SCM_\mu$$

Son **degré de liberté** est le nombre de paramètres supplémentaires à estimé (après contraintes) dans $M_{\mu,\alpha}$ par rapport à M_μ, donc, $I - 1$ pour cet exemple.

On généralise cette définition pour tout modèle M_1 **emboîté** dans un modèle M_2 (noté $M_1 \subset M_2$), i.e. tel que M_2 consiste en l'**ajout** de paramètres à M_1, en notant:

$$R(M_2|M_1) = SCM_2 - SCM_1$$

où M_2 est **au plus** le modèle complet M_{comp}.

Test associé à une réduction

Pour tester l'intérêt d'un modèle M_2 par rapport à un modèle M_1 ($M_1 \subset M_2$), on testera:

$$\begin{array}{rll} & H_0 : & \text{Le vrai modèle est } M_1 \\ \text{contre} & H_1 : & \text{Le vrai modèle est } M_2 \end{array}$$

Statistique de test

On considère la statistique de test

$$F = \frac{R(M_2|M_1)/ddl(R(M_2|M_1))}{SCR/ddl(SCR)}$$

où SCR est la somme des carrés résiduelles du modèle **complet** M_{comp}.

Si H_0 est vraie, alors $F \overset{H_0}{\sim} Fisher(ddl(R(M_2|M_1), ddl(SCR))$.

Sur les données on observe f_{obs} comme une réalisation de F.

On rejette H_0 au risque de première espèce α si $\overbrace{\mathbb{P}(F > f_{obs})}^{\text{p-valeur}} < \alpha$

Dans ce cas, on conclue que les effets de M_2 non présents dans M_1 apporte de l'information sur la variable explicative.

Tests de type I et II

Type I

Effet testé	H_0	H_1	Réduction
α	$Y_{ijk} = \mu + E_{ijk}$	$Y_{ijk} = \mu + \alpha_i + E_{ijk}$	$R(\alpha\|\mu)$
β	$Y_{ijk} = \mu + \alpha_i + E_{ijk}$	$Y_{ijk} = \mu + \alpha_i + \beta_j + E_{ijk}$	$R(\beta\|\mu,\alpha)$
γ	$Y_{ijk} = \mu + \alpha_i + \beta_j + E_{ijk}$	$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} E_{ijk}$	$R(\gamma\|\mu,\alpha,\beta)$

Le test des effets principaux dépend de l'ordre d'entrée des facteurs!

Type II

Effet testé	H_0	H_1	Réduction
α	$Y_{ijk} = \mu + \beta_j + E_{ijk}$	$Y_{ijk} = \mu + \alpha_i + \beta_j + E_{ijk}$	$R(\alpha\|\mu,\beta)$
β	$Y_{ijk} = \mu + \alpha_i + E_{ijk}$	$Y_{ijk} = \mu + \alpha_i + \beta_j + E_{ijk}$	$R(\beta\|\mu,\alpha)$
γ	$Y_{ijk} = \mu + \alpha_i + \beta_j + E_{ijk}$	$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} E_{ijk}$	$R(\gamma\|\mu,\alpha,\beta)$

Test des effets Marqueur et Variete (Type I)

Effet	*ddl*	Réduction	Ratio	Stat. test.	p-valeur
α	$ddl(R(\cdot))$	$R(\alpha\|\mu)$	F	f_{obs}	$\mathbb{P}(F > f_{obs})$
β	$ddl(R(\cdot))$	$R(\beta\|\mu,\alpha)$	F	f_{obs}	$\mathbb{P}(F > f_{obs})$
γ	$ddl(R(\cdot))$	$R(\gamma\|\mu,\alpha,\beta)$	F	f_{obs}	$\mathbb{P}(F > f_{obs})$
	$ddl(SCR)$	SCR	S^2		

```
Analysis of Variance Table

Response: Rendement
                  Df Sum Sq Mean Sq F value    Pr(>F)
Marqueur           1    470   469.6  1.0048    0.3171
Variete            3  35849 11949.7 25.5655 1.393e-14 ***
Marqueur:Variete   3   1011   336.9  0.7208    0.5403
Residuals        269 125735   467.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Au risque $\alpha = 0.01$

- Pas d'effet d'interaction;
- Pas d'effet principal de type I pour le facteur Marqueur;
- Effet du facteur Variete;

Test des effets Marqueur et Variete (Type II)

Effet	Réduction	*ddl*	Stat. test.	p-valeur
α	$R(\alpha\|\mu)$	$ddl(R(\cdot))$	f_{obs}	$\mathbb{P}(F > f_{obs})$
β	$R(\beta\|\mu,\alpha)$	$ddl(R(\cdot))$	f_{obs}	$\mathbb{P}(F > f_{obs})$
γ	$R(\gamma\|\mu,\alpha,\beta)$	$ddl(R(\cdot))$	f_{obs}	$\mathbb{P}(F > f_{obs})$
	SCR	$ddl(SCR)$		

```
Anova Table (Type II tests)

Response: Rendement
                 Sum Sq  Df F value    Pr(>F)
Marqueur             69   1  0.1473    0.7014
Variete           35849   3 25.5655 1.393e-14 ***
Marqueur:Variete   1011   3  0.7208    0.5403
Residuals        125735 269
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Au risque $\alpha = 0.01$

- Pas d'effet d'interaction;
- Pas d'effet principal de type II pour le facteur Marqueur;
- Effet du facteur Variete;

Notion d'orthogonalité

Facteur 1 / Facteur 2	1	...	J	Somme
1	n_{11}	...	n_{1J}	n_{1+}
$\vdots$	$\vdots$	...	...	$\vdots$
I	n_{I1}	...	n_{IJ}	n_{I+}
Somme	n_{+1}	...	n_{+J}	n

Le plan est **orthogonal** si pour tout i et j

$$n_{ij} = \frac{n_{i+} \times n_{+j}}{n}$$

En particulier, si le plan d'expérience est **équilibré**, il est orthogonal.

Si le plan est orthogonal, les tests de type I et II sont les mêmes.

VI) Test sur les paramètres de moyenne.

Hypothèses de test

Pour chaque paramètre de moyenne (**une fois les contraintes posées!**) on teste:

$$\begin{aligned} & H_0 : \quad \text{paramètre} = 0 \\ \text{contre} \quad & H_1 : \quad \text{paramètre} \neq 0 \end{aligned}$$

Ce test a en pratique peu d'intérêt car il dépend de la contrainte

Statistique de test

Exemple pour α_i, on utilise $T = \frac{\hat{\alpha}_i}{\sqrt{\widehat{\mathbb{V}[\hat{\alpha}_i]}}}$

où $\widehat{\mathbb{V}[\hat{\alpha}_i]}$ est l'estimateur de la variance de $\hat{\alpha}_i$ (calculé grâce à la réalisation de S^2).

Si H_0 est vraie, $T \overset{H_0}{\sim} Student(ddl(SCR))$.

On observe la réalisation t_{obs} de T. On rejette H_0 au risque α si:

$$\mathbb{P}(T < t_{obs}) < \alpha/2 \text{ ou } \mathbb{P}(T > t_{obs}) < \alpha/2$$

Ou, de manière équivalente: $\overbrace{2\mathbb{P}(T > |t_{obs}|)}^{\text{p-valeur dans R}} < \alpha$

Estimations du modèle

```
Call:
lm(formula = formule_anov2, data = donnees)

Residuals:
    Min      1Q  Median      3Q     Max
-73.521 -11.365   1.584  13.834  58.682

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 347.887      3.223 107.942  < 2e-16 ***
Marqueur2                    -5.154      4.108  -1.255    0.211
VarieteEuropean             -27.904      5.915  -4.717 3.84e-06 ***
VarieteNorthern             -31.167      5.096  -6.116 3.36e-09 ***
VarieteTropical              -5.402      8.295  -0.651    0.515
Marqueur2:VarieteEuropean     9.626      7.356   1.309    0.192
Marqueur2:VarieteNorthern     7.771      7.472   1.040    0.299
Marqueur2:VarieteTropical     3.895      9.245   0.421    0.674
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.62 on 269 degrees of freedom
Multiple R-squared:  0.2289,    Adjusted R-squared:  0.2089
F-statistic: 11.41 on 7 and 269 DF,  p-value: 1.135e-12
```

Comparaison des moyennes par variété

Pour chaque couple de croisement de niveau (i, j) et (i', j') on veut tester

$$\begin{aligned} & H_0 : \quad \mu_{ij} = \mu_{i'j'} \\ \text{contre} \quad & H_1 : \quad \mu_{ij} \neq \mu_{i'j'} \end{aligned}$$

Statistique de test

$T = \frac{\hat{\mu}_{ij} - \hat{\mu}_{i'j'}}{\sqrt{S^2}}$

où S^2 est l'estimateur de la variance du modèle d'ANOVA.

Si H_0 est vraie, $T \overset{H_0}{\sim} Student(ddl(SCR))$.

On observe la réalisation t_{obs} de T. On rejette H_0 au risque α si:

$$\mathbb{P}(T < t_{obs}) < \alpha/2 \text{ ou } \mathbb{P}(T > t_{obs}) < \alpha/2$$

Ou, de manière équivalente:

$$\overbrace{2\mathbb{P}(T > |t_{obs}|)}^{\text{p-valeur dans R}} < \alpha$$

Correction pour les tests multiples

- $\frac{IJ \times (IJ-1)}{2}$ tests sont effectués;
- Mécaniquement, la multitude de test amènera à rejeter H_0;
- Il faut **corriger** la p-valeur!

Correction de Bonferroni

On ajuste la p - valeur par le nombre de tests. Pour un risque de 1ère espèce α, on rejettera H_0 si

$$\overbrace{\frac{IJ(IJ-1)}{2} \times 2\mathbb{P}(T > |t_{obs}|)}^{\text{p-valeur corrigée}} < \alpha$$

Cette correction est appelée correction de Bonferroni.

Test 2 à 2

contrast	estimate	SE	df	t.ratio	p.value
1,Corn - 2,Corn	5.154046	4.108393	269	1.255	1.0000
1,Corn - 1,European	27.903652	5.915050	269	4.717	0.0001
1,Corn - 2,European	23.431862	4.797901	269	4.884	0.0001
1,Corn - 1,Northern	31.166936	5.095834	269	6.116	<.0001
1,Corn - 2,Northern	28.550361	5.810144	269	4.914	<.0001
1,Corn - 1,Tropical	5.402369	8.295416	269	0.651	1.0000
1,Corn - 2,Tropical	6.661371	4.533014	269	1.470	1.0000
2,Corn - 1,European	22.749605	5.576082	269	4.080	0.0017
2,Corn - 2,European	18.277816	4.373183	269	4.180	0.0011
2,Corn - 1,Northern	26.012890	4.698127	269	5.537	<.0001
2,Corn - 2,Northern	23.396314	5.464673	269	4.281	0.0007
2,Corn - 1,Tropical	0.248323	8.057221	269	0.031	1.0000
2,Corn - 2,Tropical	1.507325	4.080819	269	0.369	1.0000
1,European - 2,European	-4.471790	6.101939	269	-0.733	1.0000
1,European - 1,Northern	3.263284	6.338874	269	0.515	1.0000
1,European - 2,Northern	0.646709	6.926152	269	0.093	1.0000
1,European - 1,Tropical	-22.501282	9.111956	269	-2.469	0.3963
1,European - 2,Tropical	-21.242280	5.895932	269	-3.603	0.0105
2,European - 1,Northern	7.735074	5.311625	269	1.456	1.0000
2,European - 2,Northern	5.118499	6.000302	269	0.853	1.0000
2,European - 1,Tropical	-18.029493	8.429696	269	-2.139	0.9339
2,European - 2,Tropical	-16.770491	4.774312	269	-3.513	0.0146
1,Northern - 2,Northern	-2.616575	6.241096	269	-0.419	1.0000
1,Northern - 1,Tropical	-25.764567	8.602758	269	-2.995	0.0840
1,Northern - 2,Tropical	-24.505565	5.073630	269	-4.830	0.0001
2,Northern - 1,Tropical	-23.147991	9.044208	269	-2.559	0.3089
2,Northern - 2,Tropical	-21.888989	5.790680	269	-3.780	0.0054
1,Tropical - 2,Tropical	1.259002	8.281795	269	0.152	1.0000

P value adjustment: bonferroni method for 28 tests

Comparaison des moyennes par Marqueur

On a les moyennes observées par type de marqueur

```
  Marqueur  Moyenne
1        1 333.0984
2        2 335.7980
```

Si on souhaite comparer les moyennes pour l'effet d'un facteur, il faut *ajuster* par l'effet de l'autre facteur et l'effet d'interaction.

Pour les moyennes de rendement par Marqueur, on comparera, pour tout i:

$$\tilde{\mu}_{i\bullet} = \mu + \alpha_i + \overbrace{\frac{1}{J}\sum_{j=1}^{J}\beta_j + \frac{1}{J}\sum_{j=1}^{J}\gamma_{ij}}^{Ajustement}$$

```
$lsmeans
 Marqueur   lsmean       SE  df lower.CL upper.CL
 1        331.7683 2.610020 269 326.6297 336.9070
 2        331.9371 1.814123 269 328.3655 335.5088

Results are averaged over the levels of: Variete
Confidence level used: 0.95

$contrasts
 contrast   estimate       SE  df t.ratio p.value
 1 - 2    -0.1688292 3.178561 269  -0.053  0.9577

Results are averaged over the levels of: Variete
```

VII) Critique du modèle

Le modèle est ici inutilement compliqué, il ne semble pas utile d'inclure le marqueur génétique pour expliquer le rendement.

Parmi les deux facteurs explicatifs, seul la variable variété semble importante.

VIII) Conclusion biologique

La connaissance de la variété semble donner une information sur le rendement, cependant la connaissance du marqueur génétique ne semble pas être un facteur important.

On pourra désormais voir si un modèle comprenant les variables quantitatives **et** l'effet variété permet une bonne prédiction des données.