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.
- 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 Belt Dent, European Flint, Northern Flint, Stiff Stalk, 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 un facteur

I) Question biologique

Question biologique: Le **rendement** d'une espèce peut il être expliqué par sa *variété*.

- Le **rendement** est la variable à expliquer;
- La *variété* est la variable explicative. C'est une variable *qualitative*.
- **Cadre de l'ANOVA 1 facteur** 1 **variable à expliquer**, quantitative, 1 *variable explicative*, qualitative.
- **Première étape:** Une approche descriptive.

Coefficient de corrélation linéaire empirique

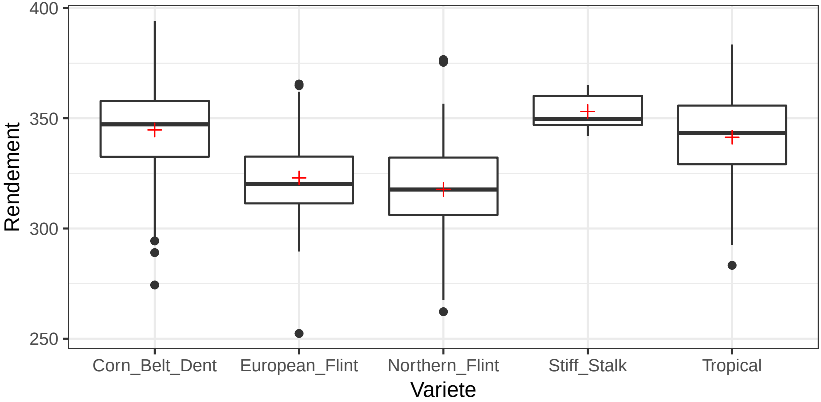
Moyennes et effectif par variété

```
Corn_Belt_Dent European_Flint Northern_Flint    Stiff_Stalk       Tropical
           117             56             50             11             54
```

Niveau	Moyenne
Corn_Belt_Dent	344.7148
European_Flint	322.9375
Northern_Flint	317.7662
Stiff_Stalk	353.1102
Tropical	341.4117

Visualisation graphique

Question biologique: Le **rendement** d'une espèce peut il être expliqué par sa *variété*.



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

Notations

On a $n = 288$ observations. Le facteur explicatif a $I = 5$ niveaux codés ainsi: Corn Belt Dent (i=1), European Flint (i=2), Northern Flint (i=3), Stiff Stalk (i=4), Tropical (i=5). Pour chaque niveau i, on dispose de n_i observations avec

$n_1 = 117, n_2 = 56, n_3 = 50, n_4 = 11, n_5 = 54.$

On note y_{ik} le rendement de la k-ième parcelle pour la variété i $(1 \leq k \leq n_i)$

Modèle

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

$$Y_{ik} = \mu_i + E_{ik}, \; 1 \leq i \leq I, \; 1 \leq k \leq n_i$$

- μ_i est la moyenne attendue de rendement pour la variété i;
- E_{ik} est le résidu (aléa) associé à l'observation Y_{ik}. $E_{ik} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

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

Notations

On a $n = 288$ observations. Le facteur explicatif a $I = 5$ niveaux codés ainsi: Corn Belt Dent (i=1), European Flint (i=2), Northern Flint (i=3), Stiff Stalk (i=4), Tropical (i=5). Pour chaque niveau i, on dispose de n_i observations avec

$n_1 = 117, n_2 = 56, n_3 = 50, n_4 = 11, n_5 = 54.$

On note y_{ik} le rendement de la k-ième parcelle pour la variété i $(1 \leq k \leq n_i)$

Modèle

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

$$Y_{ik} = \mu + \alpha_i + E_{ik},\ 1 \leq i \leq I,\ 1 \leq k \leq n_i$$

- μ est la moyenne de référence de rendement;
- α_i est **l'effet** de la variété i sur le rendement.
- E_{ik} est le résidu (aléa) associé à l'observation Y_{ik}. $E_{ik} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

Toutes les observations sont supposées indépendantes.

Pour un même niveau de facteur les observations sont de **même loi (identiquement distribuées)**.

III) Ajustement du modèle régulier

Estimateurs

- **Estimateurs**: (Variables aléatoires) L'estimateur de μ_i est donné par la moyenne empirique du rendement des parcelles de variété i.

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} Y_{ik} = Y_{i\bullet}$$

- **Estimations**: Réalisation sur les données)

$$\hat{\mu}_i^{obs} = \frac{1}{n_i} \sum_{k=1}^{n_i} y_{ik} = y_{i\bullet}$$

Prédicteur et prédiction

- **Prédicteur** Pour une parcelle de variété i, le rendement prédit est:

$$\widehat{Y}_{ik} = \hat{\mu}_i = Y_{i\bullet}$$

- **Prédiction** Réalisation sur les données

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

III) Ajustement du modèle singulier

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

On doit **poser une contrainte**, typiquement $\alpha_1 = 0$.

Le niveau 1 est alors **le niveau de référence**.

Les estimateurs de $\mu, \alpha_1, \ldots, \alpha_I$ vont **dépendre** de la contrainte!

Pour la contrainte $\alpha_1 = 0$, on a:

$$\hat{\mu} = Y_{1\bullet}, \ \hat{\alpha}_i = Y_{i\bullet} - Y_{1\bullet}$$

Le prédicteur est alors

$$\widehat{Y}_{ik} = \hat{\mu} + \hat{\alpha}_i = Y_{i\bullet}$$

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

Estimateur et estimation de la variance σ^2

Résidus observés

$$\hat{e}_{ik} = y_{ik} - \hat{y}_{ik},\ 1 \leq i \leq I,\ 1 \leq k \leq n_i$$

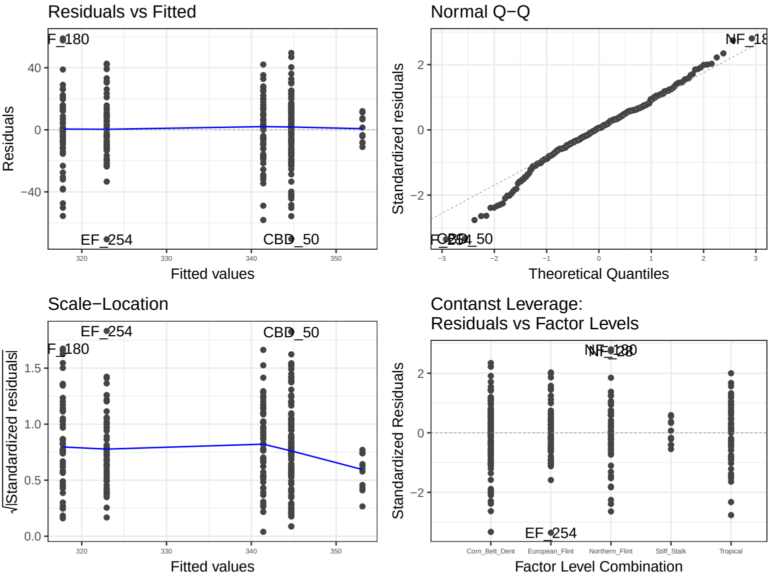
Estimateur

$$S^2 = \frac{\sum_{i=1}^{I} \sum_{k=1}^{n_i} (Y_{ik} - \widehat{Y}_{ik})^2}{n - I}$$

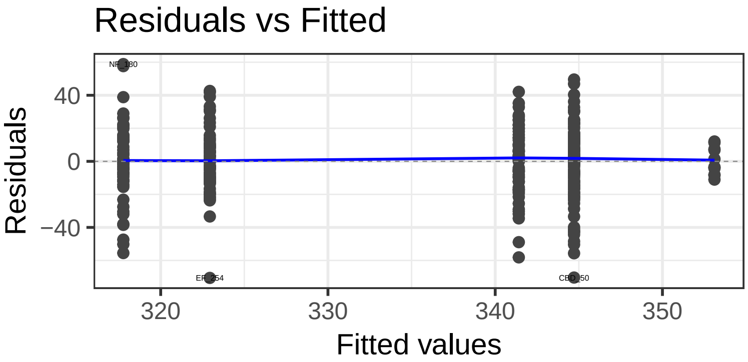
Estimation

$$\hat{\sigma}^2_{obs} = \frac{\sum_{i=1}^{I} \sum_{k=1}^{n_i} (y_{ik} - \widehat{y}_{ik})^2}{n - I} = \frac{\sum_{k=1}^{n_i} \hat{e}^2_{ik}}{n - I} \stackrel{ici}{=} 450.5$$

IV) Validité des hypothèses



Distribution identique, espérance constante et nulle

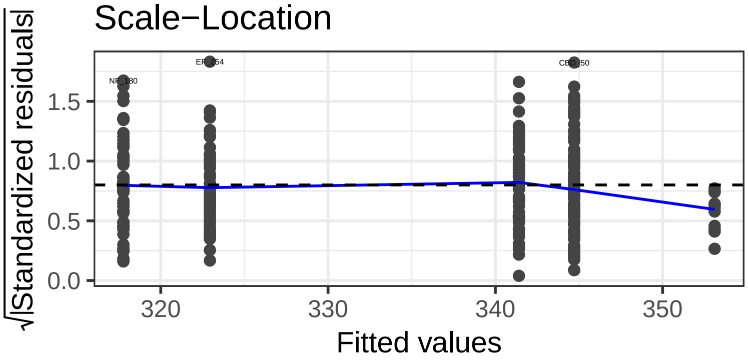


Ce qu'on regarde: Les résidus observés $\hat{e}_{ik}$ en fonction des prédictions $\hat{y}_{ik}$ (équivalent à regarder en fonction de la variété).

Ce qu'on voit: La distribution des résidus semble comparable dans toutes les variétés sauf celle de moyenne maximale (Stiff stalk).

Ce qu'on conclut: L'hypothèse de distribution identique ne semble valable que pour 4 variétés sur 5.

Distribution identique, variance constante

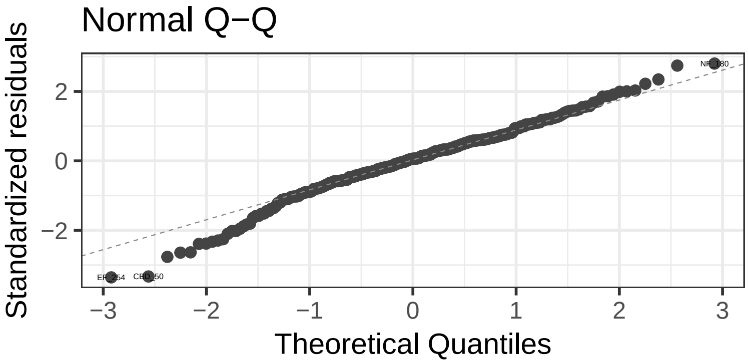


Ce qu'on regarde: La valeur absolue des résidus (standardisés) observés en fonction des prédictions $\hat{y}_{ik}$ (équivalent à regarder en fonction de la variété).

Ce qu'on voit: La valeur absolue des résidus semble comparable dans toutes les variétés (autour de 0.8) sauf celle de moyenne maximale (Stiff stalk).

Ce qu'on conclut: L'hypothèse de variance identique ne semble valable que pour 4 variétés sur 5.

Distribution normale

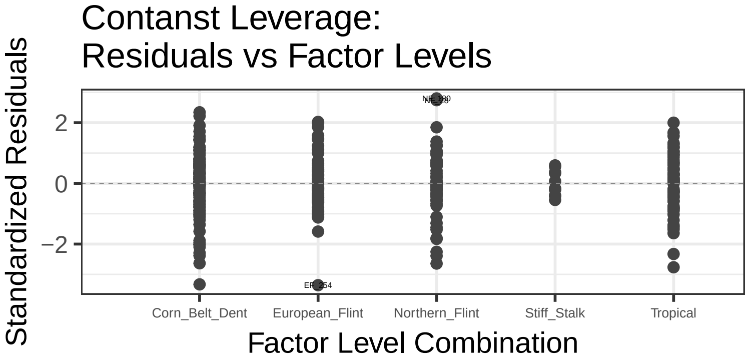


Ce qu'on regarde: La valeur des quantiles empiriques des résidus standardisés en fonction de la valeur quantiles théoriques d'une loi normale $\mathcal{N}(0, 1)$.

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 (à voir si le problème vient aussi de la variété Stiff Stalk).

Résidus par variété

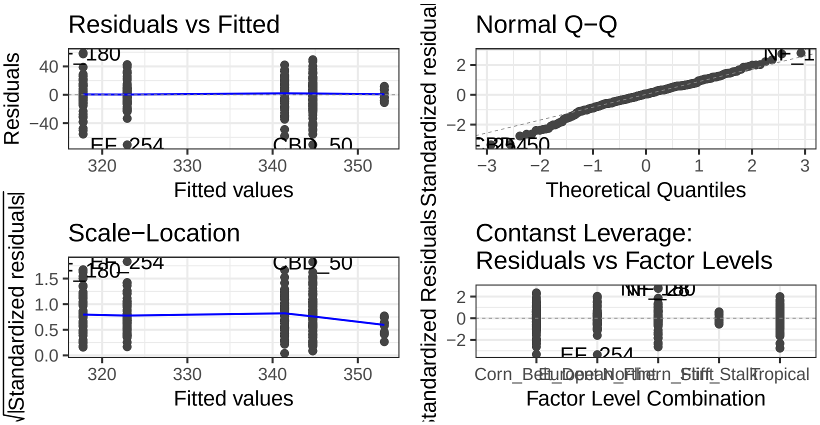


Ce qu'on regarde: Même graphique (réarrangé) que le 1er.

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

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

4 graphes et un oeil fin



La variété Stiff Stalk semble être un problème, il est peut être préférable de les retirer de l'étude (faible effectif, ne réprésente que 11 observations).

V) Test du modèle

On veut tester si notre modèle impliquant la variété de maïs explique mieux le **rendement** qu'un modèle simple, où le **rendement** est constant.

Hypothèses du test

On teste:

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

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

$$\overbrace{\sum_{i=1}^{I}\sum_{k=1}^{n_i}(Y_{ik} - \bar{Y})^2}^{SCT} = \overbrace{\sum_{i=1}^{I}\sum_{k=1}^{n_i}(\widehat{Y}_{ik} - \bar{Y})^2}^{SCM} + \overbrace{\sum_{i=1}^{I}\sum_{k=1}^{n_i}(Y_{ik} - \widehat{Y}_{ik})^2}^{SCR}$$

Somme des carrés	Degrés de liberté (*ddl*)	Réalisation
SCT: Totale	n - 1	$SCT_{obs} = \sum_{k,i}(y_{ik} - \bar{y})^2$
SCM: Modèle	I - 1	$SCM_{obs} = \sum_{k,i}(\hat{y}_{ik} - \bar{y})^2$
SCR: Résiduelle	n - I	$SCR_{obs} = \sum_{k,i}(y_{ik} - \hat{y}_{ik})^2$

Test du modèle

Hypothèses du test

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

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 ~ Variete
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    287 167286
2    283 127490  4     39795 22.084 7.005e-16 ***
---
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) Test sur les paramètres de moyenne.

Hypothèses de test

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

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

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_anov1, data = donnees)

Residuals:
    Min      1Q  Median      3Q     Max
-70.572 -11.674   1.313  12.800  58.863

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            344.715      1.962 175.674  < 2e-16 ***
VarieteEuropean_Flint  -21.777      3.449  -6.314 1.05e-09 ***
VarieteNorthern_Flint  -26.949      3.586  -7.515 7.54e-13 ***
VarieteStiff_Stalk       8.395      6.694   1.254    0.211
VarieteTropical         -3.303      3.492  -0.946    0.345
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.22 on 283 degrees of freedom
Multiple R-squared:  0.2379,    Adjusted R-squared:  0.2271
F-statistic: 22.08 on 4 and 283 DF,  p-value: 7.005e-16
```

Comparaison des moyennes par variété

Pour chaque variété i et i' on veut tester

$$\begin{array}{rl} H_0 : & \mu_i = \mu_{i'} \\ \text{contre} \quad H_1 : & \mu_i \neq \mu_{i'} \end{array}$$

Statistique de test

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

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

- $\frac{I\times(I-1)}{2}$ 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

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

Cette correction est appelée correction de Bonferroni.

Test 2 à 2

```
    Pairwise comparisons using t tests with pooled SD

data:  donnees[, reponse] and donnees[, fact1]

               Corn_Belt_Dent European_Flint Northern_Flint Stiff_Stalk
European_Flint 1.0e-08        -              -              -
Northern_Flint 7.5e-12        1.00000        -              -
Stiff_Stalk    1.00000        0.00023        1.0e-05        -
Tropical       1.00000        7.5e-05        3.4e-07        0.96783

P value adjustment method: bonferroni
```

VII) Critique du modèle

Le modèle prédit mal mais est largement significatif.

Cependant, l'ajustement en prenant en compte la variété Stiff Stalk ne satisfait pas l'hypothèse de variance homogène

VIII) Conclusion biologique

La connaissance de la variété semble donner une information sur le rendement, mais la prédiction reste mauvaise.

Il pourrait être utile de prendre en compte le marquer génétique