# 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.

# Modèle de régression simple

# I) Question biologique

**Question biologique:** La teneur moyenne en *amidon* d'un grain de maïs permet elle de prédire le **rendement** d'une parcelle?

- Le **rendement** est la variable à expliquer;
- La teneur en *amidon* est la variable explicative. C'est une variable *quantitative*.
- **Cadre de la régression simple:** 1 **variable à expliquer**, quantitative, 1 *variable explicative*, quantitative.
- **Première étape:** Une approche descriptive.

# Coefficient de corrélation linéaire empirique

On a $n = 288$ observations. On note, pour $1 \leq k \leq 288$

- $x_k$ la mesure de la teneur moyenne en *amidon* sur la parcelle $k$. On note $\mathbf{x}$ l'échantillon complet, et $\bar{x}$ la valeur moyenne de l'échantillon;
- $y_k$ la mesure du **rendement** sur la parcelle $k$. On note $\mathbf{y}$ l'échantillon complet, et $\bar{y}$ la valeur moyenne de l'échantillon.

Le corrélation linéaire empirique est donnée par:

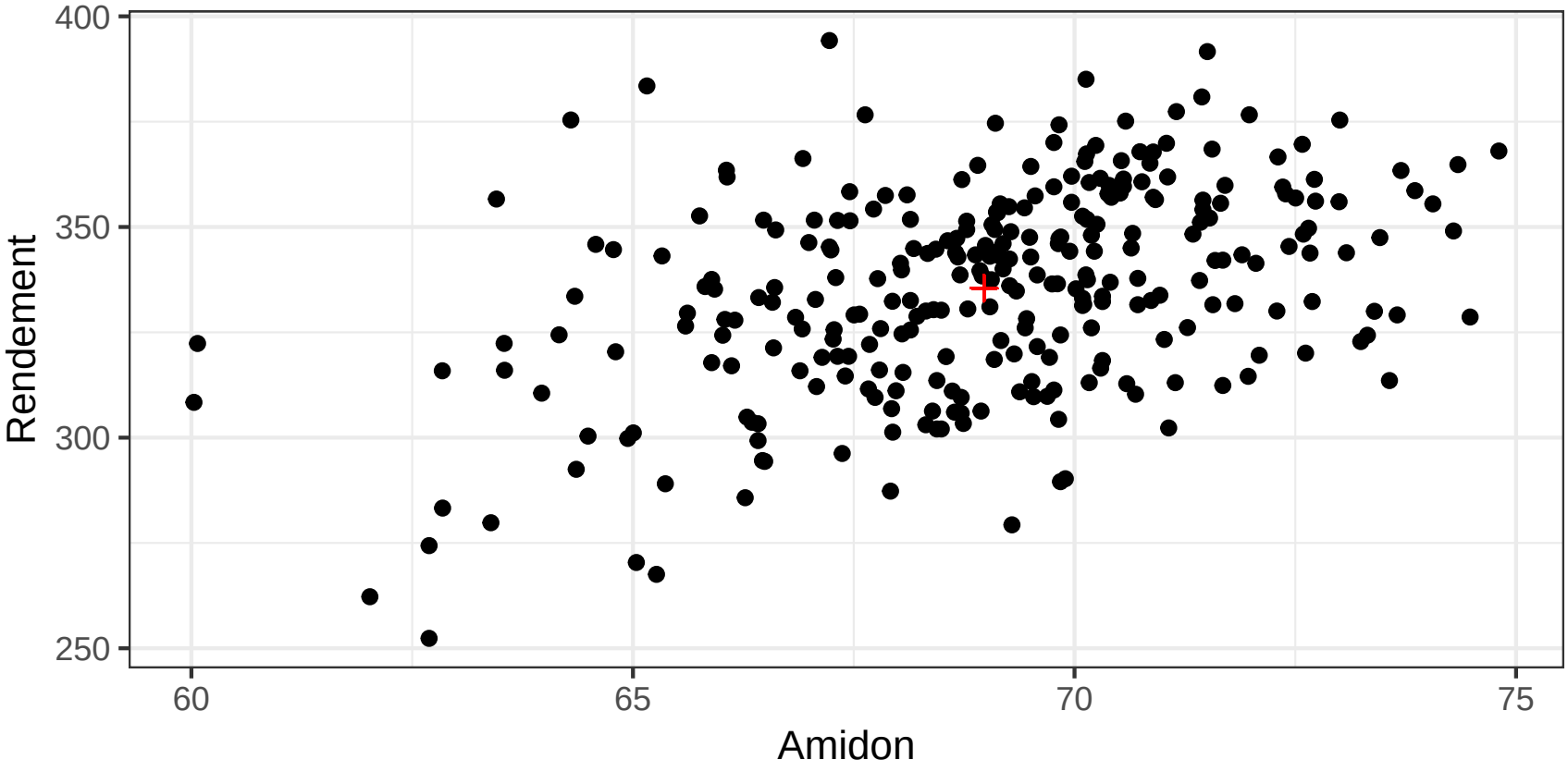
$$\rho(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{n}(x_k - \bar{x})^2 \sum_{k=1}^{n}(y_k - \bar{y})^2}}$$

- $-1 \leq \rho(\mathbf{x}, \mathbf{y}) \leq 1$
- $\rho(\mathbf{x}, \mathbf{y})$ proche de 0: $\mathbf{x}$ et $\mathbf{y}$ ne sont pas corrélés **linéairement**;
- $\rho(\mathbf{x}, \mathbf{y})$ proche de 1: corrélation **linéaire** positive entre $\mathbf{x}$ et $\mathbf{y}$. Quand $\mathbf{x} \nearrow, \mathbf{y} \nearrow$ ;
- $\rho(\mathbf{x}, \mathbf{y})$ proche de $-1$: corrélation **linéaire** négative entre $\mathbf{x}$ et $\mathbf{y}$. Quand $\mathbf{x} \nearrow, \mathbf{y} \searrow$ ;

Ici, on observe: $\bar{x} = 68.98, \bar{y} = 335.5, \rho(\mathbf{x}, \mathbf{y}) = 0.43$

# Visualisation graphique

**Question biologique:** La teneur moyenne en amidon d'un grain de maïs permet elle de prédire le rendement d'une parcelle?



# II) Ecriture du modèle

## Notations

On a $n = 288$ observations. On note, pour $1 \leq k \leq 288$

- $x_k$ la mesure de la teneur moyenne en *amidon* sur la parcelle $k$.
- $y_k$ la mesure du **rendement** sur la parcelle $k$.

## Modèle de régression linéaire simple

On suppose que $y_k$ est la réalisation d'une variable aléatoire $Y_k$ telle que:

$$Y_k = \beta_0 + \beta_1 x_k + E_k, 1 \leq k \leq 288$$

où

- $\beta_0$ est un paramètre inconnu;
- $\beta_1$ est un paramètre inconnu, l'effet de l'amidon sur le rendement;
- $E_k$ une variable aléatoire appelée **résidu**, telle que:
  - Toutes les variables aléatoires $E_1, \ldots, E_n$ sont **indépendantes**;
  - Tous les $E_k$ ont la **même espérance**, égale à **0**;
  - Tous les $E_k$ ont la **même variance**, égale à $\sigma^2$ (paramètre inconnu);
  - Tous les $E_k$ suivent une **loi normale**;
- $\Rightarrow$ les $E_k$ sont indépendants et identiquement distribués de loi $\mathcal{N}(0, \sigma^2)$

# III) Ajustement du modèle

- **Objectif** Trouver $\beta_0$, $\beta_1$ et $\sigma^2$ qui s'ajustent le mieux aux données.

## Estimateurs et estimations de $\beta_0$ et $\beta_1$

- **Estimateurs**: (Variables aléatoires)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_k - \bar{x})(Y_k - \bar{Y})}{\sum_{k=1}^{n}(x_k - \bar{x})^2}, \ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{x}$$

  Ces estimateurs sont **sans biais**. Ces estimateurs suivent des lois normales.

- **Estimations**: (Réalisations sur les données)

$$\hat{\beta}_1^{obs} = \frac{\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})}{\sum_{k=1}^{n}(x_k - \bar{x})^2}, \ \hat{\beta}_0^{obs} = \bar{y} - \hat{\beta}_1^{obs}\bar{x}$$

## Prédicteur et prédiction

- **Prédicteur** (V.A.) Valeur moyenne attendue pour le **rendement** pour une valeur $x_k$ d'*amidon*.

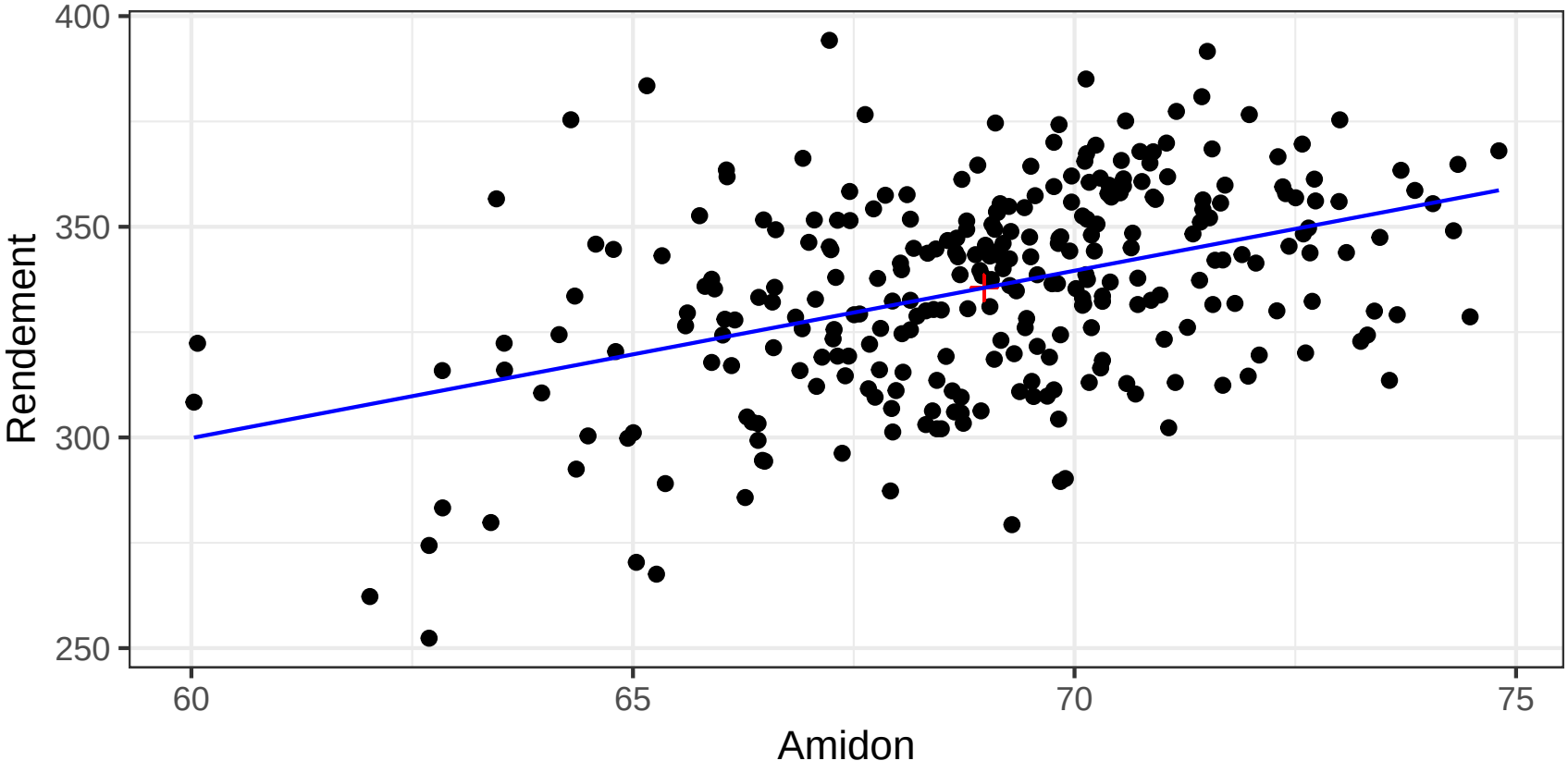
$$\widehat{Y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k$$

- **Prédiction** Réalisation sur les données $\widehat{y}_k = \hat{\beta}_0^{obs} + \hat{\beta}_1^{obs} x_k$

# Droite de prédiction

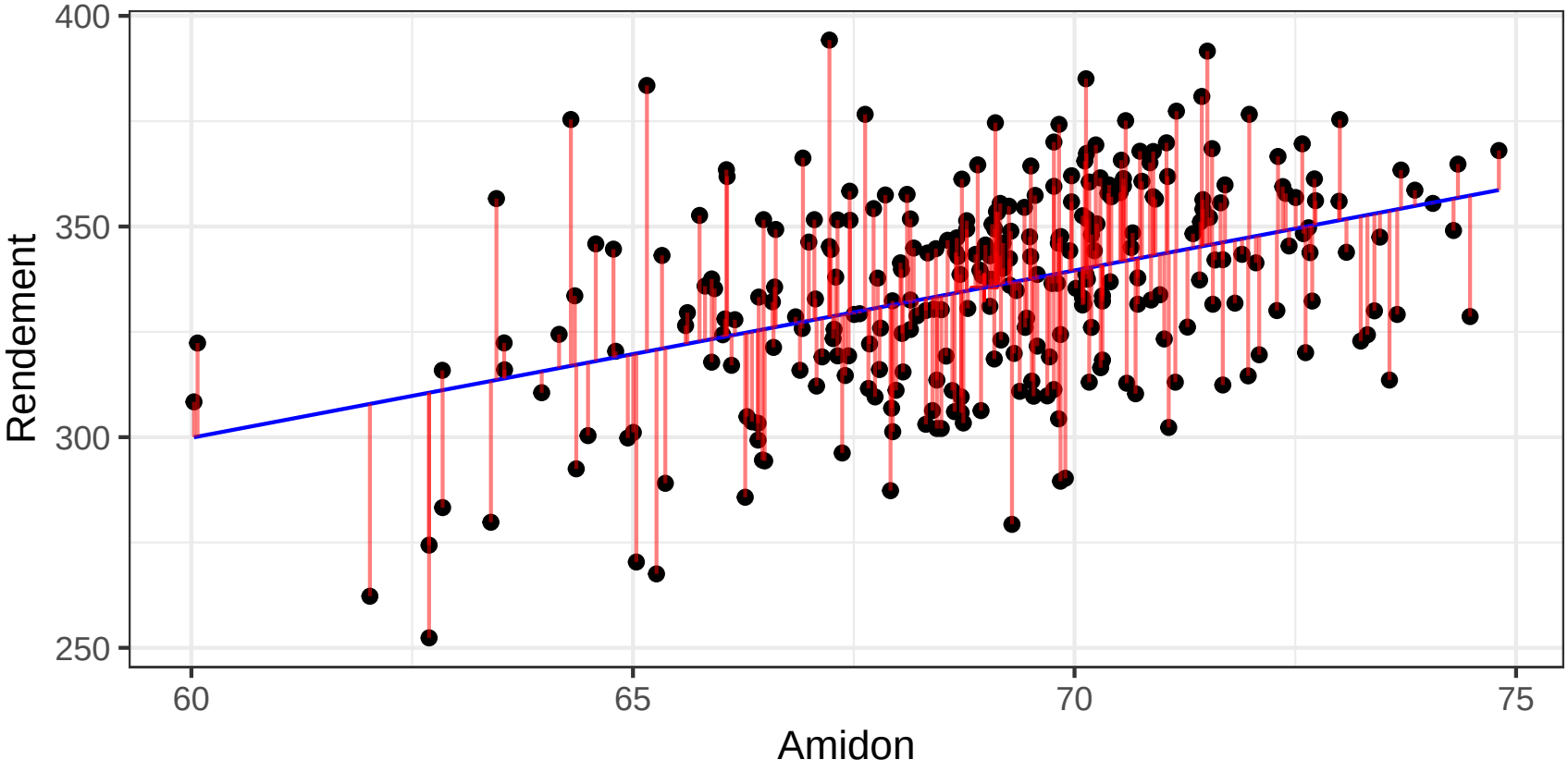
Droite d'équation

$$y = \hat{\beta}_0^{obs} + \hat{\beta}_1^{obs} x \text{ où ici, } \hat{\beta}_0^{obs} = 61.6, \hat{\beta}_1^{obs} = 3.97$$



# Résidus observés

$$\hat{e}_k = y_k - \hat{y}_k, \ 1 \leq k \leq n$$



# Estimateur et estimation de la variance $\sigma^2$

## Estimateur

$$S^2 = \frac{\sum_{k=1}^{n}(Y_k - \widehat{Y}_k)^2}{n-2}$$

## Estimation

$$\hat{\sigma}^2_{obs} = \frac{\sum_{k=1}^{n}(y_k - \widehat{y}_k)^2}{n-2} = \frac{\sum_{k=1}^{n}\hat{e}_k^2}{n-2} \overset{ici}{=} 478.8$$

# IV) **Validité des hypothèses**

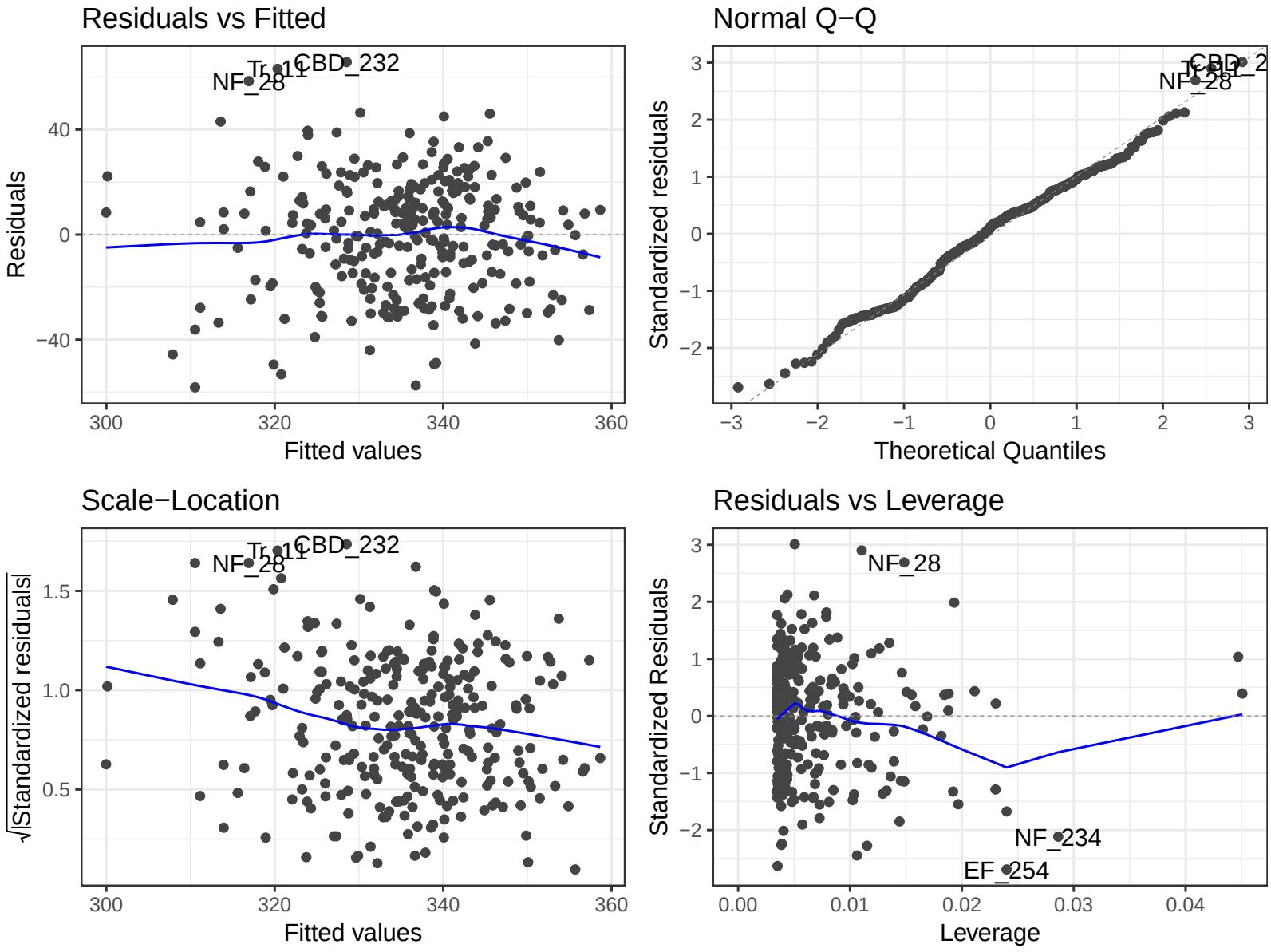
Les résidus observés permettent de valider les hypothèses du modèle linéaire:

- $E_k$ une variable aléatoire appelée **résidu**, telle que:
  - Toutes les variables aléatoires $E_1, \ldots, E_n$ sont **indépendantes**;
  - Tous les $E_k$ ont la **même espérance**, égale à $\mathbf{0}$;
  - Tous les $E_k$ ont la **même variance**, égale à $\sigma^2$ (paramètre inconnu);
  - Tous les $E_k$ suivent une **loi normale**;

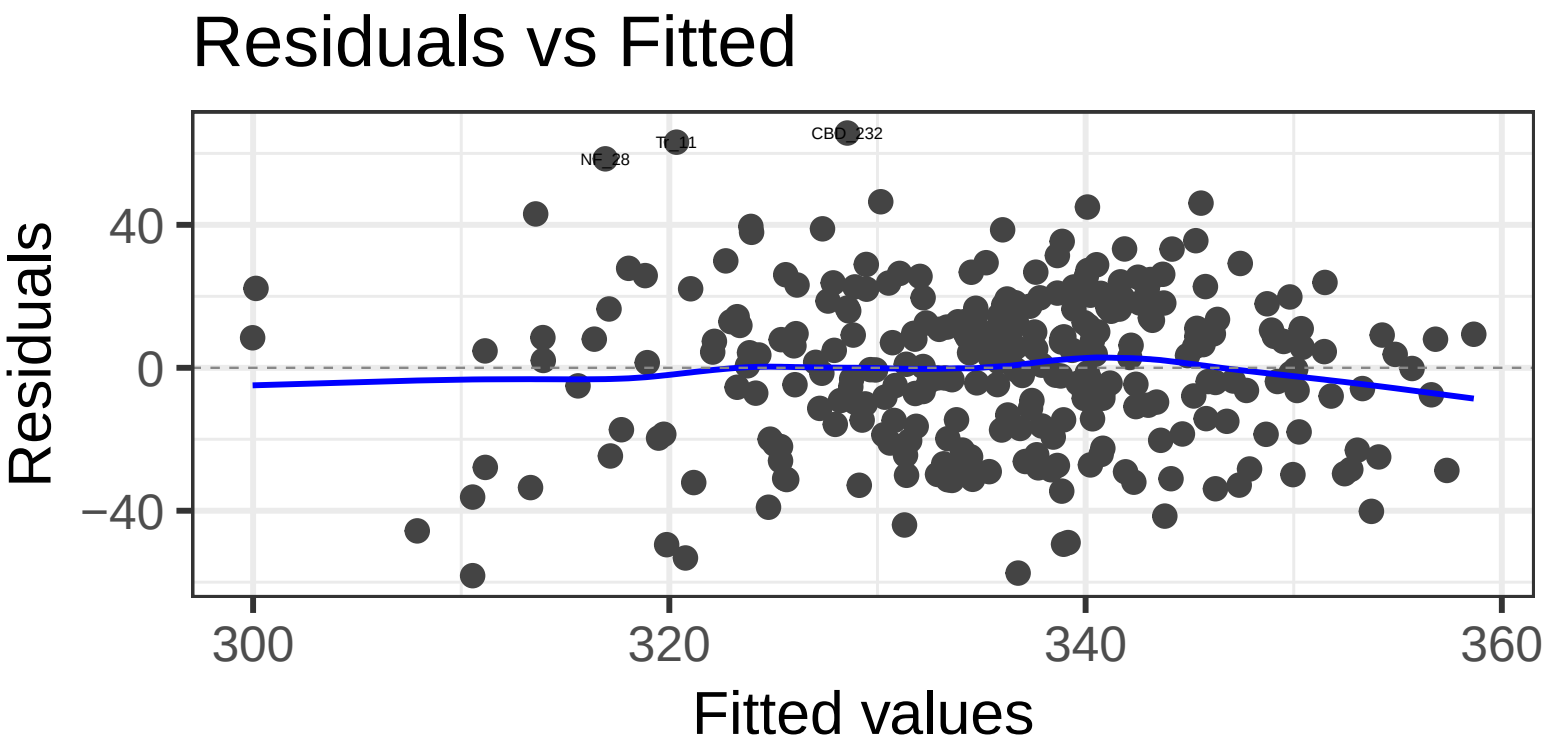
## Validation des hypothèses

- **Hypothèse d'indépendance:** Elle doit être validée par le plan d'expérience!
- **Distribution identique, de loi normale:** Ces hypothèses doivent être vérifiées grâce aux $\hat{e}_k$.
- **En pratique:** diagnostic graphique des résidus

# 4 graphes et un oeil fin



## Distribution identique, espérance constante et nulle

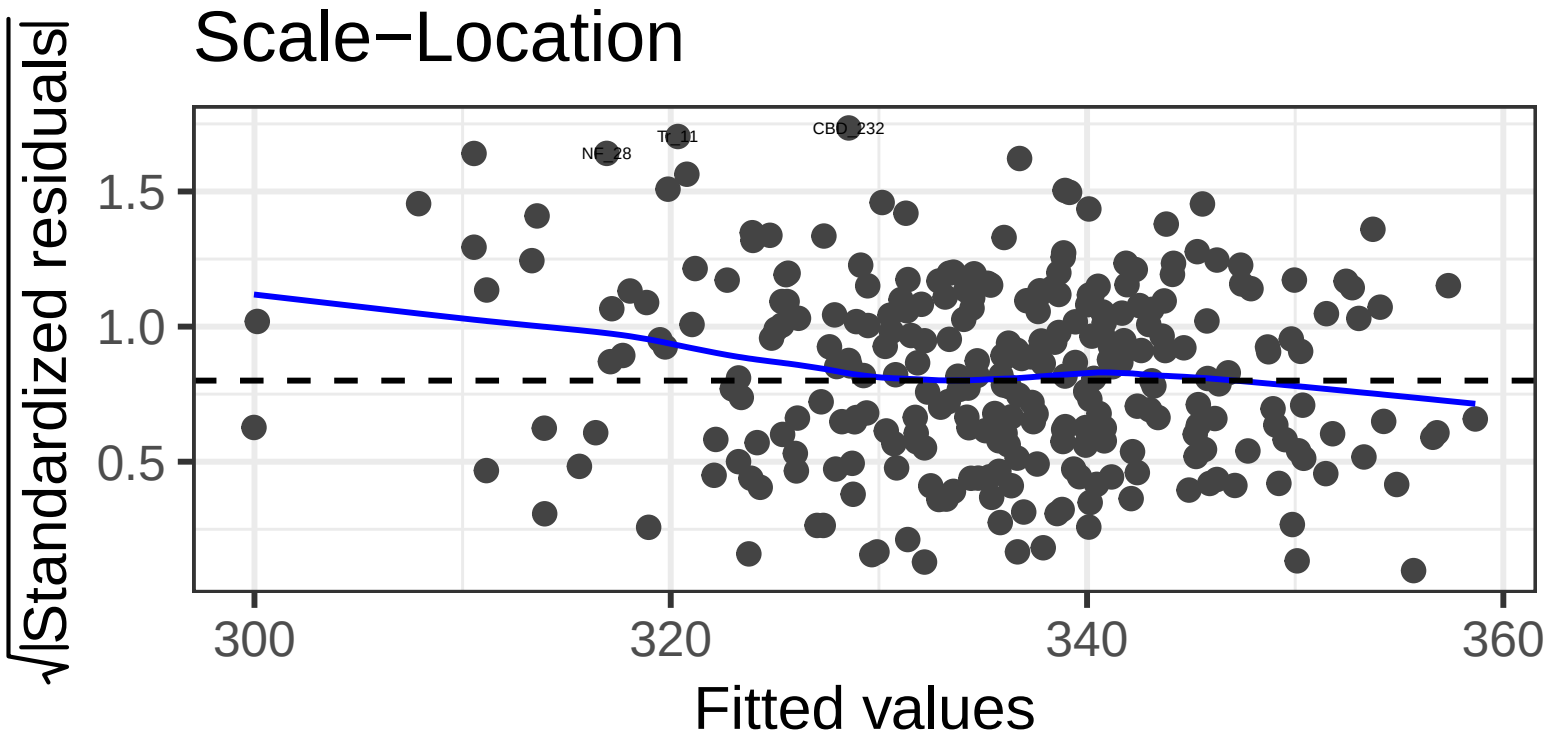



**Ce qu'on regarde:** Les résidus observés $\hat{e}_k$ en fonction des prédictions $\hat{y}_k$.

**Ce qu'on voit:** La valeur des résidus ne semble pas dépendre de la valeur des prédictions (il ne sont donc pas structurés en fonction de la prédiction). Ils sont globalement identiquement distribués autour de 0.

**Ce qu'on conclut:** On valide l'hypothèse d'espérance constante et égale à 0.

## Distribution identique, variance constante

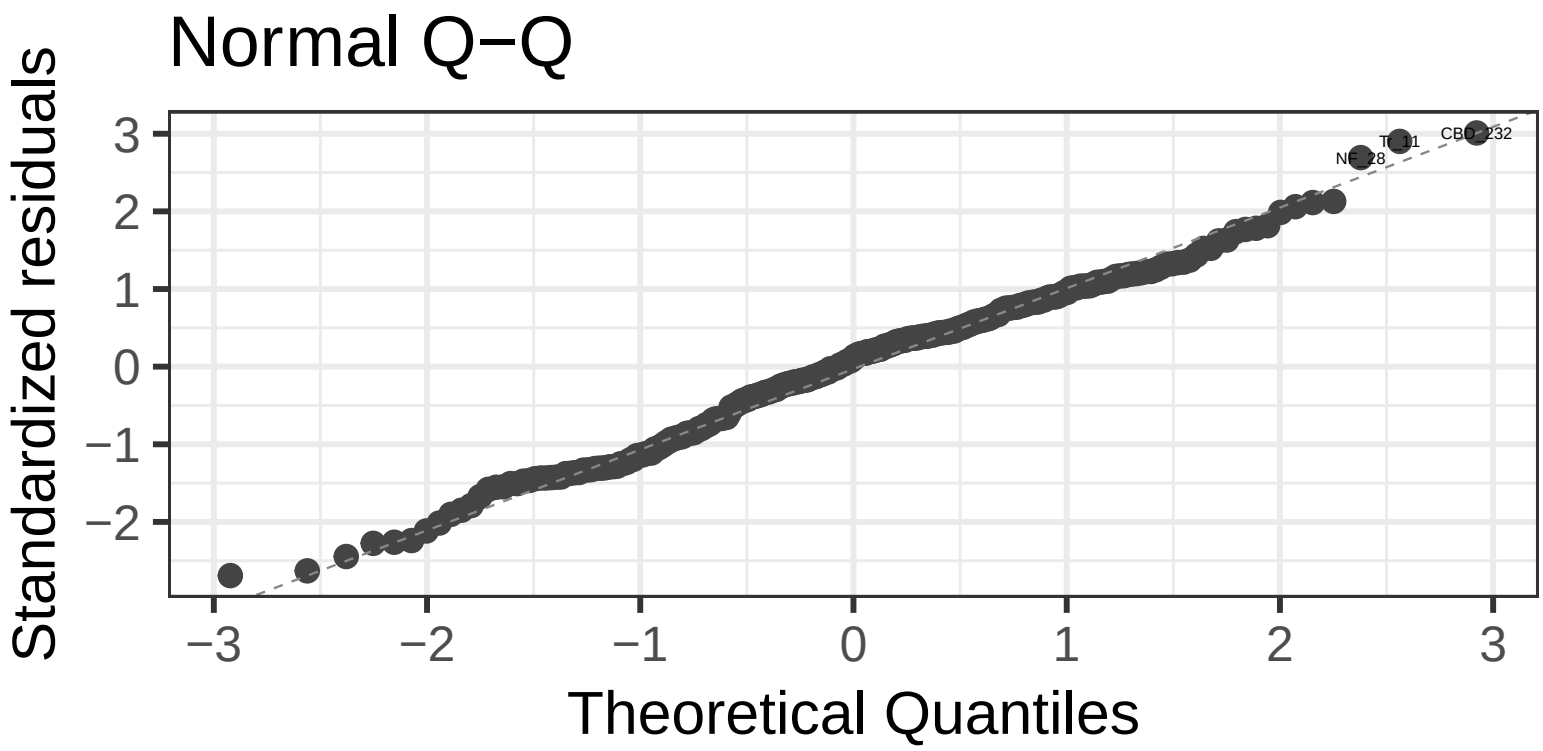



**Ce qu'on regarde:** La valeur absolue des résidus (standardisés) observés en fonction des prédictions $\hat{y}_k$.

**Ce qu'on voit:** La valeur absolue des résidus ne semble pas dépendre de la valeur des prédictions (il ne sont donc pas structurés en fonction de la prédiction). Ils sont globalement identiquement distribués autour de 0.8.

**Ce qu'on conclut:** On valide l'hypothèse de variance constante.

# Distribution normale

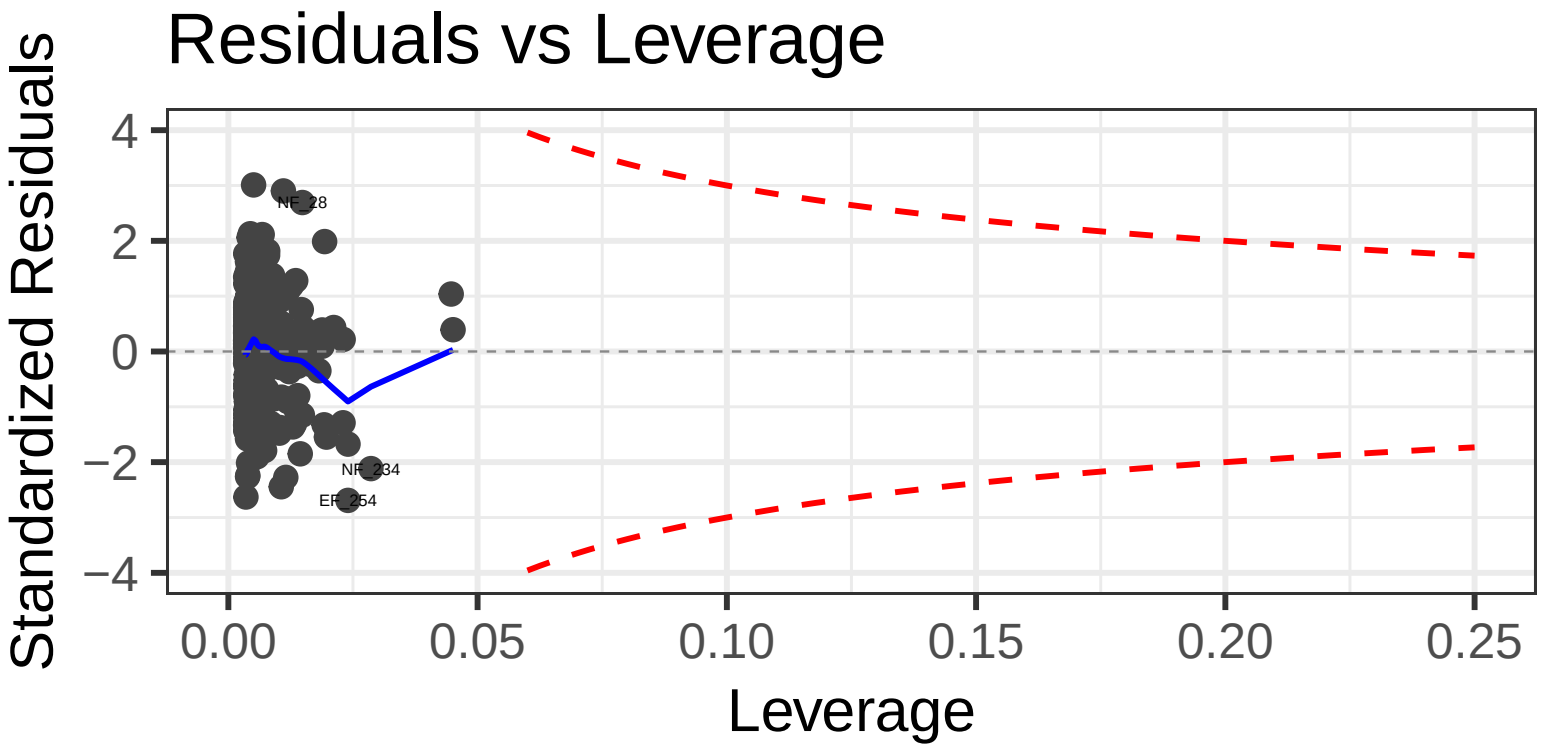



**Ce qu'on regarde:** La valeur des quantiles empiriques des résidus standardisés en fonction de la valeur quantiles théoriques d'une loi normale $\mathcal{N}(0, 1)$.

**Ce qu'on voit:** Les points sont globalement alignés sur la droite $y = x$. Les quantiles empiriques sont donc à peu près égaux aux quantiles théoriques (si les hypothèses du modèle sont vraies).

**Ce qu'on conclut:** On valide l'hypothèse de distribution normale des résidus.

## Points influents ou aberrants

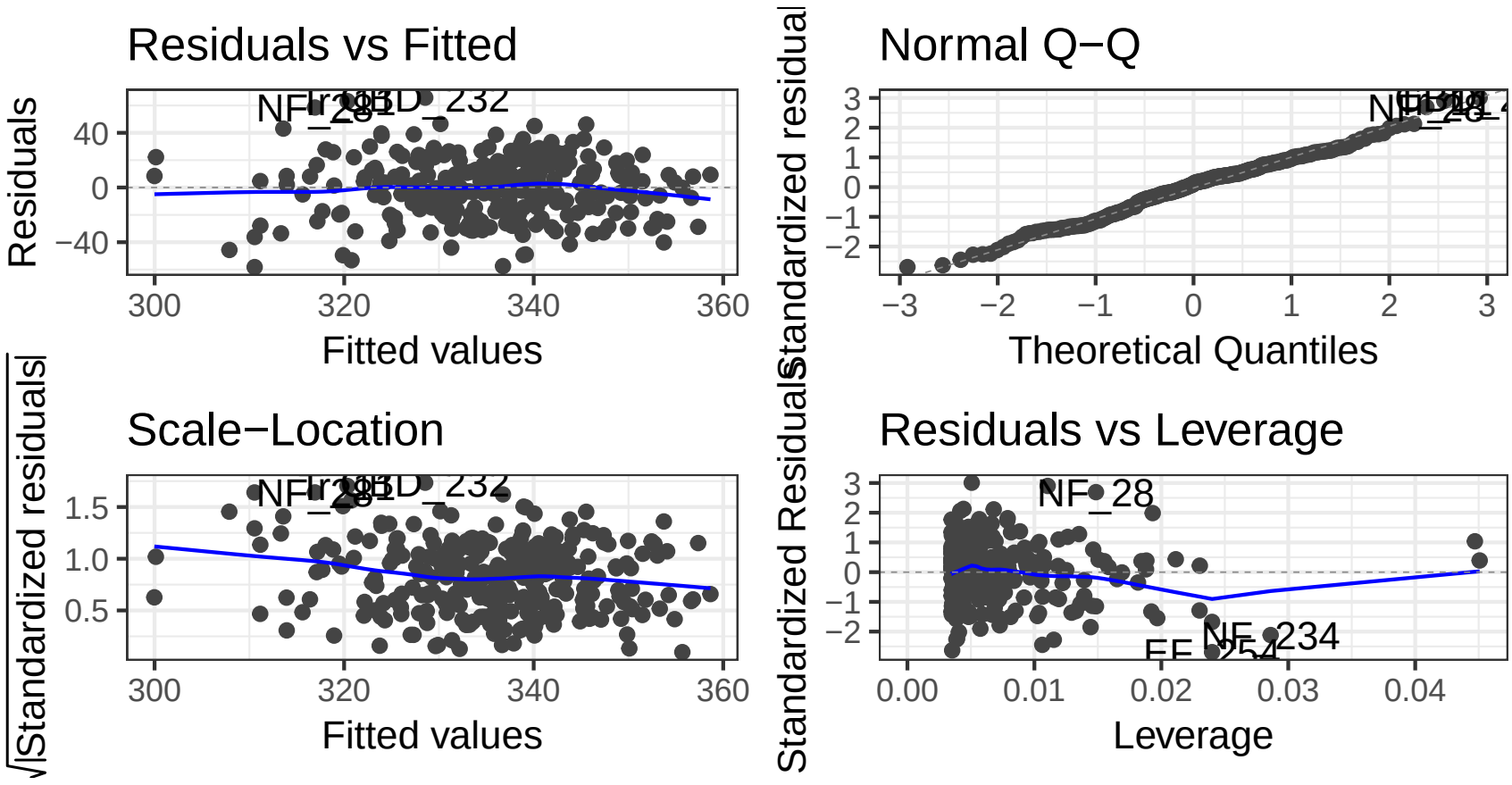



**Ce qu'on regarde:** La valeur des résidus (standardisés) en fonction du levier de l'observation (poids d'une observation dans l'estimation de sa prédiction).

**Ce qu'on voit:** Les points ont tous un petit levier, donc aucun point n'influe trop sur la droite. Aucun point n'est en dehors de l'enveloppe délimitée par les hyperboles rouges, représentant les lignes de niveau 0.5 de la distance de Cook.

**Ce qu'on conclut:** Aucun point n'est aberrant ou trop influent.

# 4 graphes et un oeil fin




Donc on valide les hypothèses du modèle pour notre exemple.

On peut maintenant tester la pertinence du modèle.

# V) Test du modèle

On veut tester si notre modèle impliquant l'*amidon* explique mieux le **rendement** qu'un modèle simple, où le **rendement** ne dépend pas de l'*amidon*.

## Hypothèses du test

On teste:

$$\begin{array}{lll} & H_0 : & Y_k = \beta_0 + E_k \qquad \text{Modèle } M_0 \\ \text{contre} & H_1 : & Y_k = \beta_0 + \beta_1 x_k + E_k \qquad \text{Modèle } M_1 \end{array}$$

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

$$\overset{SCT}{\sum_{k=1}^{n}(Y_k - \bar{Y})^2} = \overset{SCM}{\sum_{k=1}^{n}(\widehat{Y}_k - \bar{Y})^2} + \overset{SCR}{\sum_{k=1}^{n}(Y_k - \widehat{Y}_k)^2}$$

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

# Test du modèle

## Hypothèses du test

$$\begin{array}{rlll} & H_0 : & Y_k = \beta_0 + E_k & \text{Modèle } M_0 \\ \text{contre} & H_1 : & Y_k = \beta_0 + \beta_1 x_k + E_k & \text{Modèle } M_1 \end{array}$$

## 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 $\alpha$ 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 ~ Amidon
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    287 167286
2    286 136950  1     30336 63.352 4.085e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Conclusion du test:** On rejette $H_0$ et on conclut à une relation entre le **rendement** et l'*amidon*.

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

# VI) Test sur les paramètres de moyenne.

## Hypothèses de test

Pour le paramètre de moyenne $\beta_1$, on teste:

$$\begin{array}{rlll} & H_0 : & Y_k = \beta_0 + E_k & \text{Modèle } M_0 \\ \text{contre} & H_1 : & Y_k = \beta_0 + \beta_1 x_k + E_k & \text{Modèle } M_1 \end{array}$$

## Statistique de test

$$T = \frac{\hat{\beta}_1}{\sqrt{\widehat{\mathbb{V}[\hat{\beta}_1]}}}$$

où $\widehat{\mathbb{V}[\hat{\beta}_1]}$ est l'estimateur de la variance de $\hat{\beta}_1$.

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 $\alpha$ 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_reg_simple, data = donnees)

Residuals:
    Min      1Q  Median      3Q     Max
-58.184 -15.953   2.466  14.664  65.699

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  61.6026    34.4363   1.789   0.0747 .
Amidon        3.9710     0.4989   7.959 4.09e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.88 on 286 degrees of freedom
Multiple R-squared:  0.1813,    Adjusted R-squared:  0.1785
F-statistic: 63.35 on 1 and 286 DF,  p-value: 4.085e-14
```

# VII) **Critique du modèle**

## Capacité prédictive

Indicateur de la capacité de prédiction du modèle:

$$R^2_{aj} = 1 - \frac{SCR/ddl(SCR)}{SCT/ddl(SCT)}$$

- $R^2_{aj}$ proche de 1, le modèle prédit bien les données (on prédit bien le rendement avec l'amidon).
- $R^2_{aj}$ proche de 0, le modèle prédit mal les données (on prédit mal le rendement avec l'amidon).

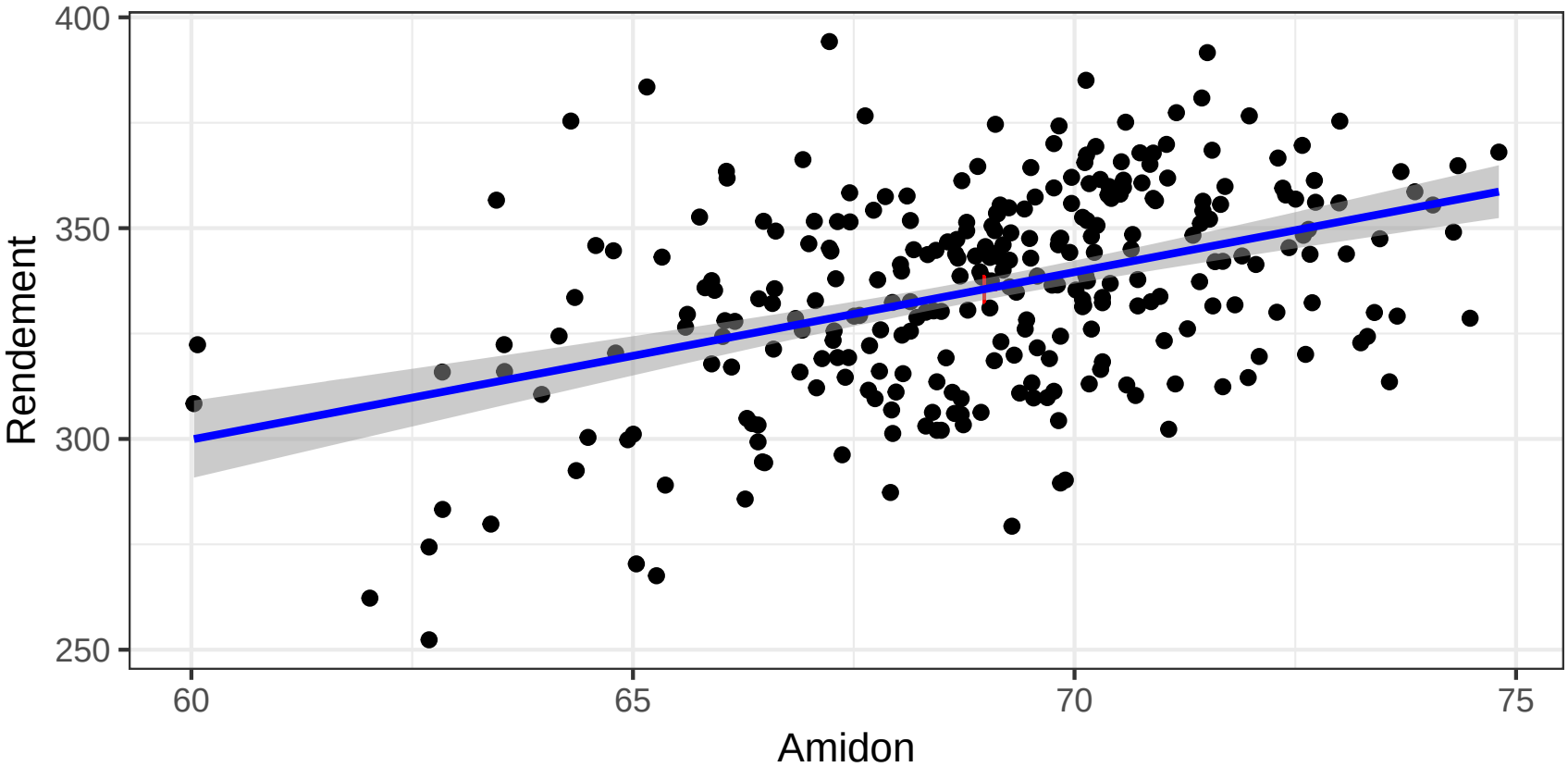
Ici $R^2_{aj} = round(0.1784799, 4)$, le modèle prédit donc mal les données.

# Intervalle de confiance pour $\hat{Y}_k$

## Loi de $\hat{Y}_k$

$$\hat{Y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k \Rightarrow \hat{Y}_k \sim \mathcal{N}(\beta_0 + \beta_1 x_k, \mathbb{V}[\hat{\beta}_0] + x_k^2 \mathbb{V}[\hat{\beta}_1] + 2x_k \mathbb{C}ov(\hat{\beta}_0, \hat{\beta}_1))$$

## Intervalle de confiance à 95% pour $\hat{Y}_k$



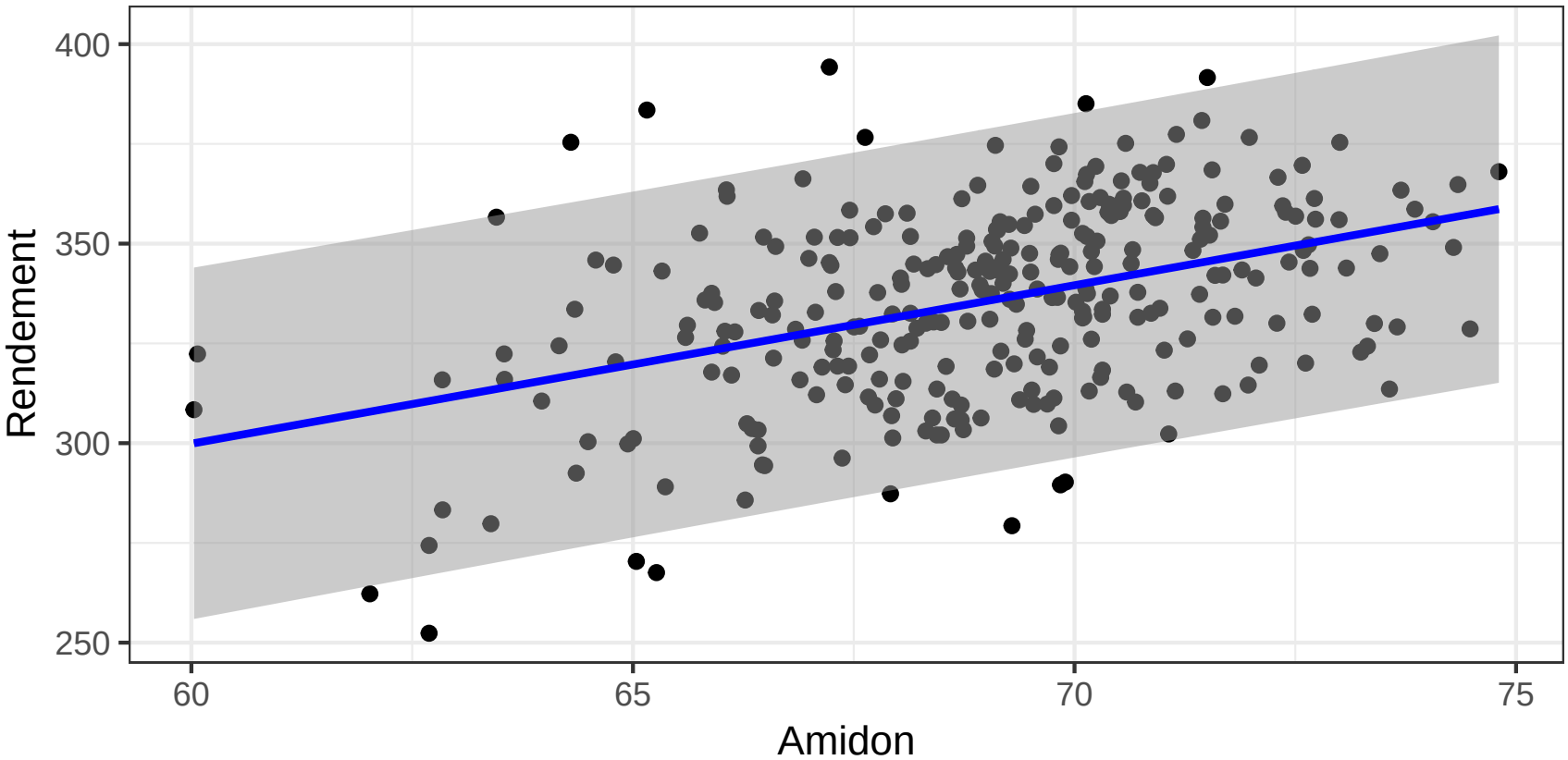
# Intervalle de prédiction

## Ajout de l'incertitude du modèle

Ajout de l'aléa estimé, on trace la loi

$$\mathcal{N}(\beta_0 + \beta_1 x_k, \mathbb{V}[\hat{Y}_k] + S^2)$$

## Intervalle de prédiction à 95 %



# VIII) **Conclusions biologiques**

On conclut que la connaissance de la teneur moyenne en amidon donne une information sur le rendement de la parcelle associée.

Ainsi, une forte teneur en amidon semble s'accompagner d'une augmentation du rendement.

Ceci étant, il existe encore beaucoup de variance non expliquée, et la prédiction associée à un modèle ne comprenant que l'amidon comme variable explicative est très incertaine. Il pourrait être intéressant d'ajouter des variables explicatives.

# Régression multiple

# 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**?

# I) Question biologique

**Question biologique:** Le rendement d'une parcelle peut il est prédit lorsque l'on connaît la teneur en *amidon*, en *huile*, en *proteine* d'un grain de maïs ainsi que le nombre de degrés jours avant *floraison*, et le *nombre de feuilles* par plant de maïs?

- Le **rendement** est la variable à expliquer;
- Les teneurs en *amidon*, *huile*, et *proteine*, la *floraison* et le *nombre de feuilles* sont des variables explicatives. Elles sont *quantitatives*.
- **Cadre de la régression multiple:** 1 **variable à expliquer**, quantitative, $p$ (ici $p = 5$) *variables explicative*, quantitatives.
- **Première étape:** Une approche descriptive.

# Coefficient de corrélation linéaire empirique

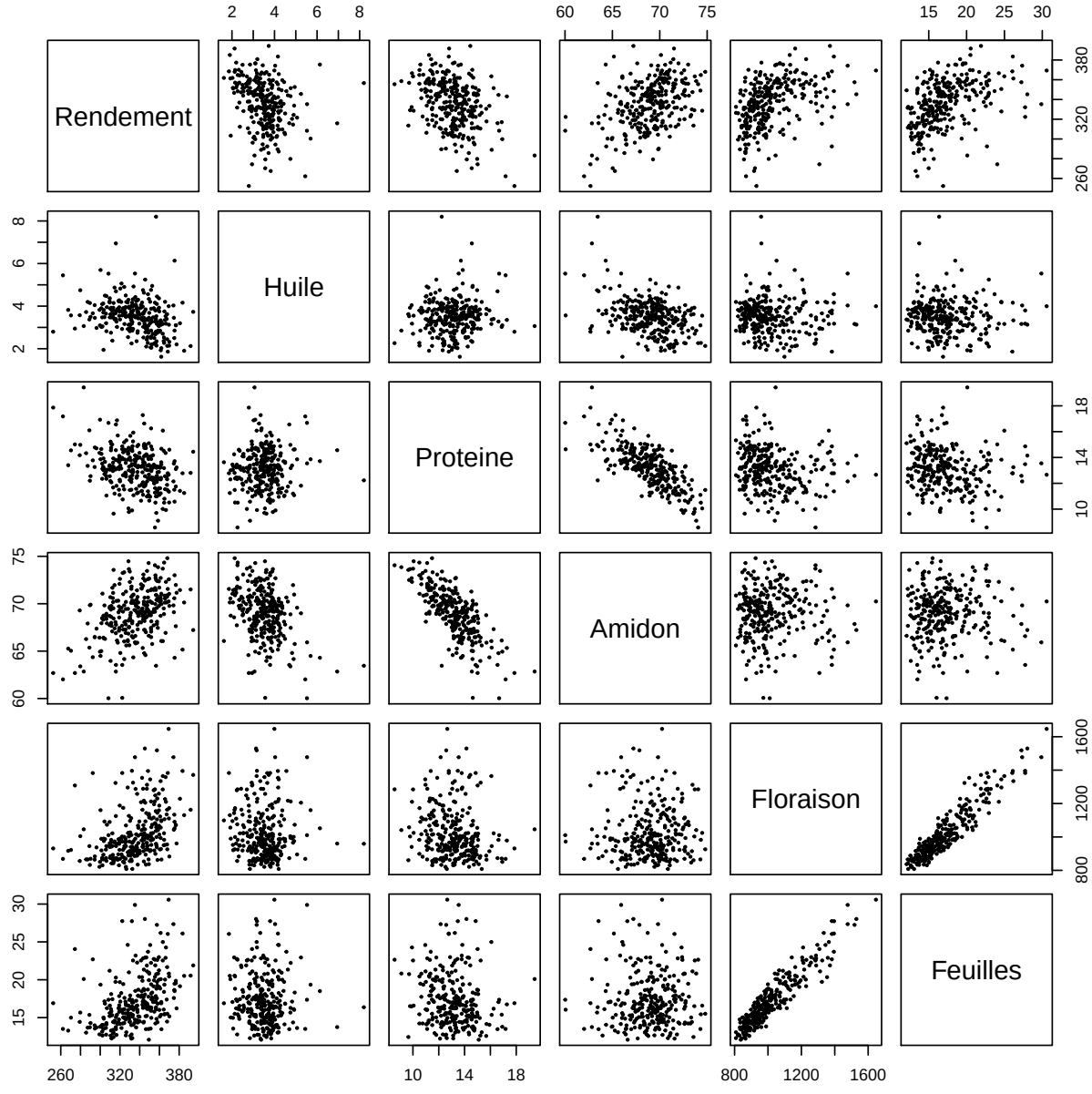
| | Rendement | Huile | Proteine | Amidon | Floraison | Feuilles |
|---|---|---|---|---|---|---|
| Rendement | 1.00 | -0.28 | -0.37 | 0.43 | 0.41 | 0.43 |
| Huile | -0.28 | 1.00 | 0.08 | -0.39 | -0.04 | -0.04 |
| Proteine | -0.37 | 0.08 | 1.00 | -0.75 | -0.18 | -0.16 |
| Amidon | 0.43 | -0.39 | -0.75 | 1.00 | 0.00 | 0.00 |
| Floraison | 0.41 | -0.04 | -0.18 | 0.00 | 1.00 | 0.93 |
| Feuilles | 0.43 | -0.04 | -0.16 | 0.00 | 0.93 | 1.00 |

# Coefficient de corrélation linéaire empirique partiel

Corrélation partielle (corrélation empirique sachant les autres variables)

| | Rendement | Huile | Proteine | Amidon | Floraison | Feuilles |
|---|---|---|---|---|---|---|
| Rendement | 1.00 | -0.13 | -0.02 | 0.25 | 0.02 | 0.18 |
| Huile | -0.13 | 1.00 | -0.36 | -0.45 | -0.04 | 0.01 |
| Proteine | -0.02 | -0.36 | 1.00 | -0.76 | -0.11 | 0.00 |
| Amidon | 0.25 | -0.45 | -0.76 | 1.00 | -0.08 | -0.06 |
| Floraison | 0.02 | -0.04 | -0.11 | -0.08 | 1.00 | 0.91 |
| Feuilles | 0.18 | 0.01 | 0.00 | -0.06 | 0.91 | 1.00 |

# Graphique 2 à 2



# II) Ecriture du modèle

## Notations

On a $n = 288$ observations. On dispose de $p = 5$ variables explicatives. On note, pour $1 \leq k \leq 288$ et $1 \leq i \leq 5$

- $x_{ik}$ la $k - i$ème mesure de la $i$-ème variable quantitative, avec les codes: Huile (i=1), Proteine (i=2), Amidon (i=3), Floraison (i=4), Feuilles (i=5).
- $y_k$ la mesure du **rendement** sur la parcelle $k$.

## Modèle de régression linéaire multiple

On suppose que $y_k$ est la réalisation d'une variable aléatoire $Y_k$ telle que:

$$Y_k = \beta_0 + \beta_1 x_{1k} + \cdots + \beta_p x_{pk} + E_k, 1 \leq k \leq n$$

où

- $\beta_0$ est un paramètre inconnu;
- $\beta_1, \ldots, \beta_p$ sont des paramètres inconnu, les effets des variables explicatives sur le rendement.
- $E_k$ une variable aléatoire appelée **résidu**, telle que tous les $E_k$ sont indépendants et identiquement distribués de loi $\mathcal{N}(0, \sigma^2)$

# Ecriture matricielle du modèle

Le modèle peut s'écrire simplement

$$Y = X\theta + E$$

avec

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \; X = \begin{pmatrix} 1 & x_{11} & \dots & x_{p1} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{1n} & \dots & x_{pn} \end{pmatrix}, \; \theta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \; E = \begin{pmatrix} E_1 \\ \vdots \\ E_n \end{pmatrix}$$

Cette notation simplifie grandement le problème d'estimation.

C'est une écriture générique du modèle linéaire.

# III) Ajustement du modèle

## Estimateurs

- **Estimateurs**: (Variables aléatoires) L'estimateur de $\theta$ est donné par la résolution des équations normales:

$$X^t X\theta = X^t Y$$

Si $X^t X$ est inversible, alors on a

$$\hat{\theta} = (X^t X)^{-1} X^t Y$$

Cet estimateur est sans biais. De plus, cet estimateur est de la normale (multivariée) dont on connaît la matrice de variance covariance:

$$\hat{\theta} \sim \mathcal{N}(\theta, (X^t X)^{-1}\sigma^2)$$

## Prédicteur et prédiction

- **Prédicteur** Pour des mesures de variables explicatives $x_{ik}$:

$$\widehat{Y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_{1k} + \cdots + \hat{\beta}_p x_{pk}$$

- **Prédiction** Réalisation sur les données $\widehat{y}_k = \hat{\beta}_0^{obs} + \hat{\beta}_1^{obs} x_{ik} + \cdots + \hat{\beta}_p^{obs} x_{ip}$

# Estimateur et estimation de la variance $\sigma^2$

## Résidus observés

Comme en régression simple:

$$\hat{e}_k = y_k - \hat{y}_k, \ 1 \leq k \leq n$$

## Estimateur

$$S^2 = \frac{\sum_{k=1}^{n} (Y_k - \widehat{Y}_k)^2}{n - (p+1)}$$

## Estimation

$$\hat{\sigma}^2_{obs} = \frac{\sum_{k=1}^{n} (y_k - \widehat{y}_k)^2}{n - (p+1)} = \frac{\sum_{k=1}^{n} \hat{e}_k^2}{n - (p+1)} \overset{ici}{=} 368.1$$

# IV) **Validité des hypothèses**

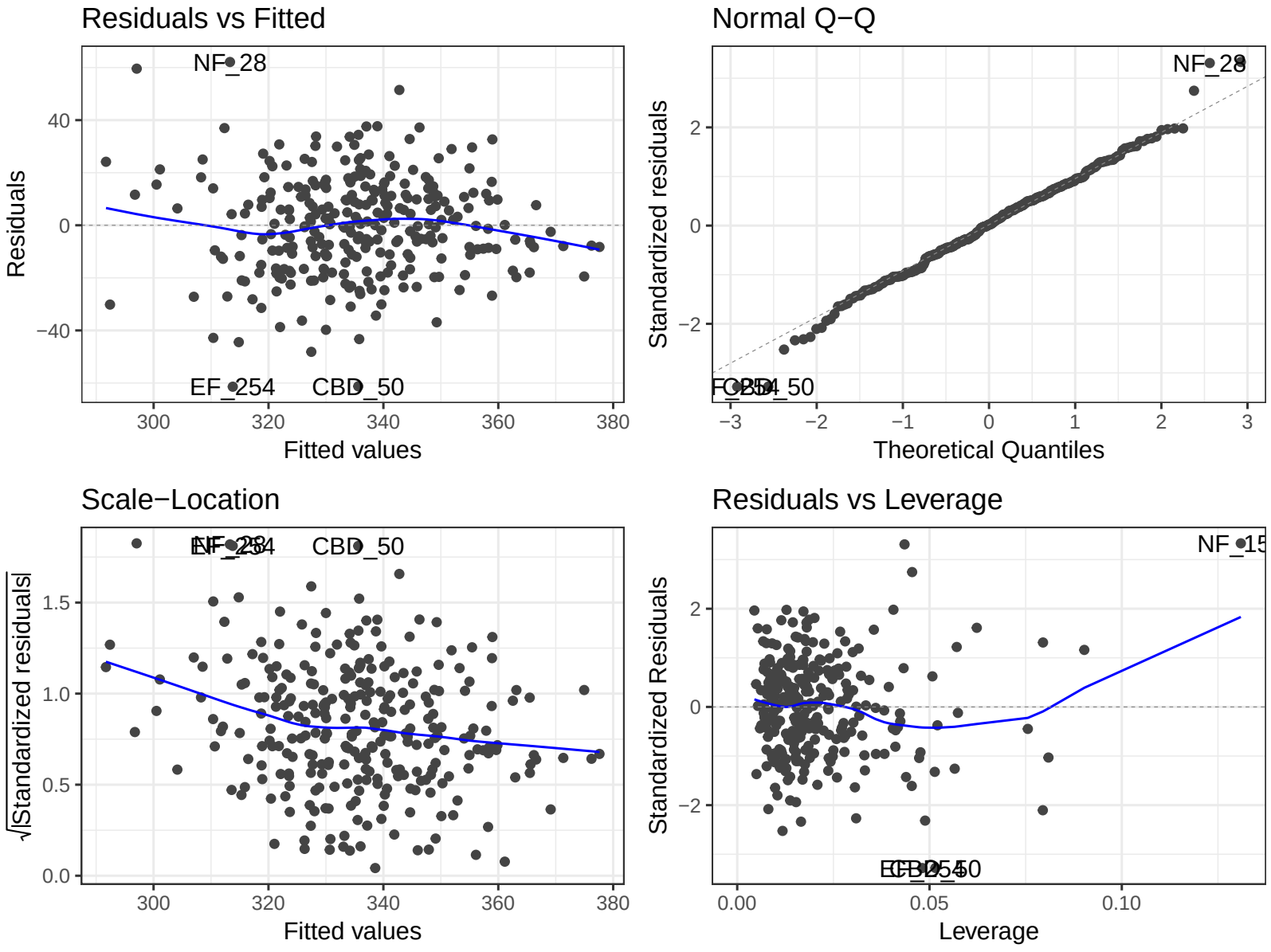
Les résidus observés permettent de valider les hypothèses du modèle linéaire:

- $E_k$ une variable aléatoire appelée **résidu**, telle que:
  - Toutes les variables aléatoires $E_1, \ldots, E_n$ sont **indépendantes**;
  - Tous les $E_k$ ont la **même espérance**, égale à $\mathbf{0}$;
  - Tous les $E_k$ ont la **même variance**, égale à $\sigma^2$ (paramètre inconnu);
  - Tous les $E_k$ suivent une **loi normale**;

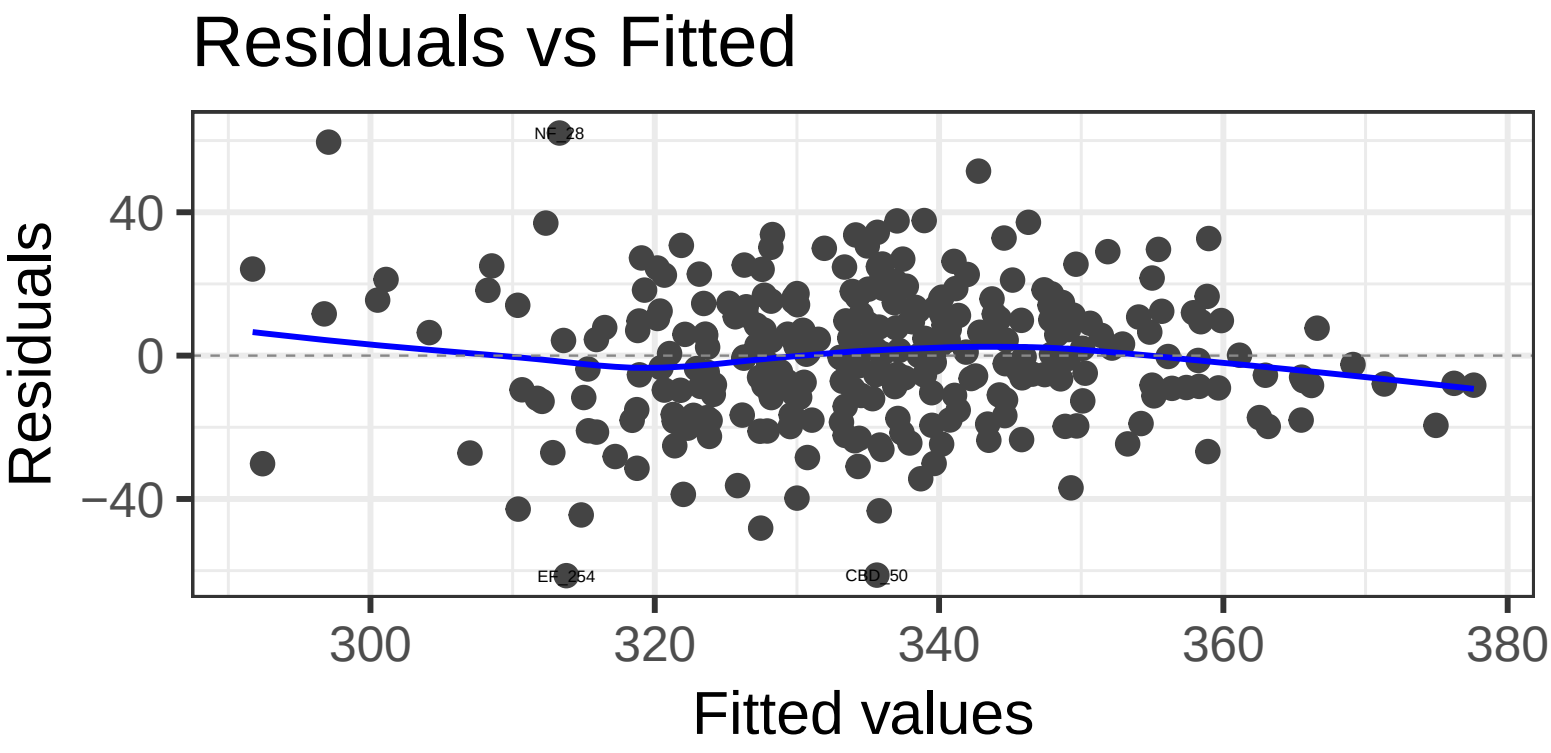
## Validation des hypothèses

- **Hypothèse d'indépendance:** Elle doit être validée par le plan d'expérience!
- **Distribution identique, de loi normale:** Ces hypothèses doivent être vérifiées grâce aux $\hat{e}_k$.
- **En pratique:** diagnostic graphique des résidus

# 4 graphes et un oeil fin



## Distribution identique, espérance constante et nulle

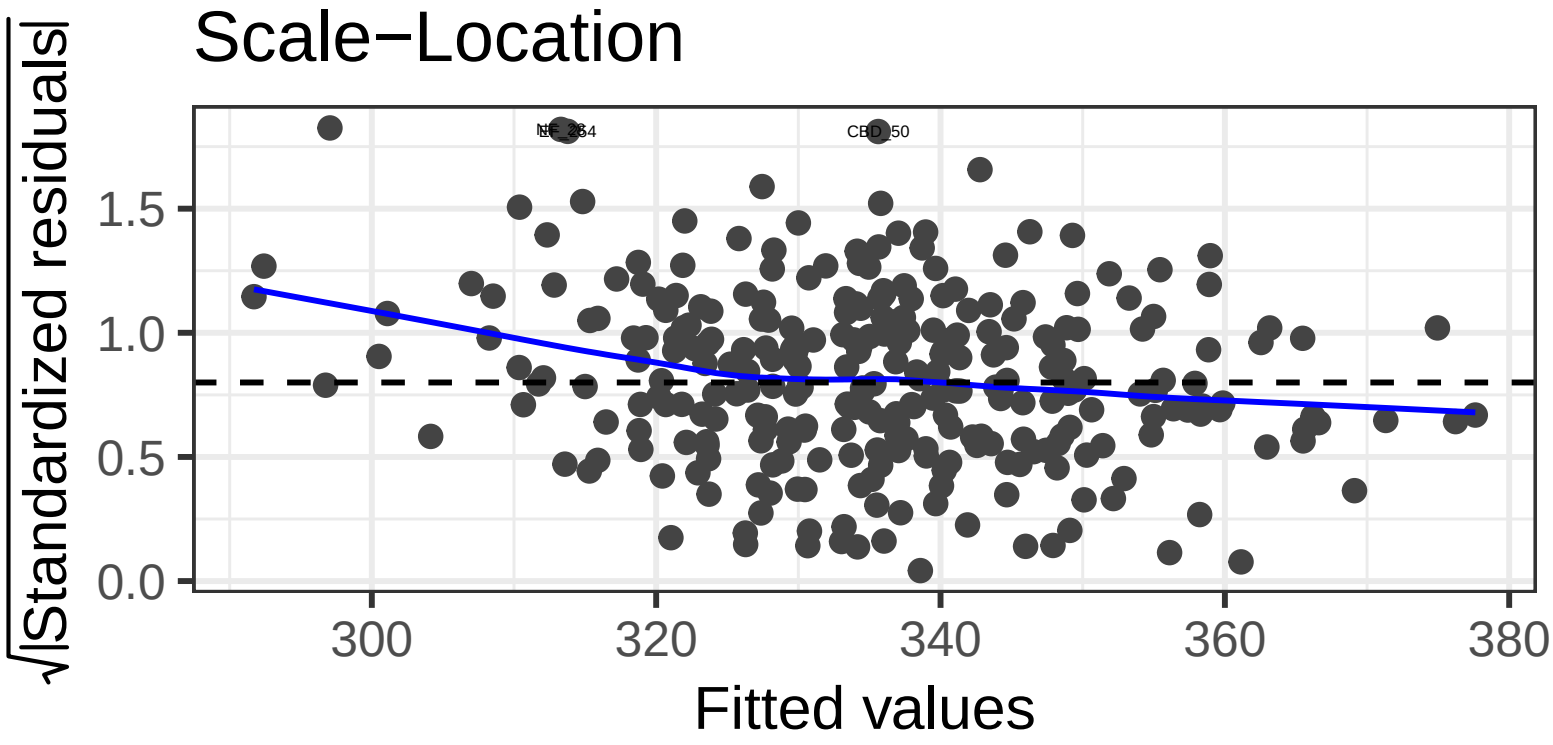



**Ce qu'on regarde:** Les résidus observés $\hat{e}_k$ en fonction des prédictions $\hat{y}_k$.

**Ce qu'on voit:** La valeur des résidus ne semble pas dépendre de la valeur des prédictions (il ne sont donc pas structurés en fonction de la prédiction). Ils sont globalement identiquement distribués autour de 0.

**Ce qu'on conclut:** On valide l'hypothèse d'espérance constante et égale à 0.

## Distribution identique, variance constante

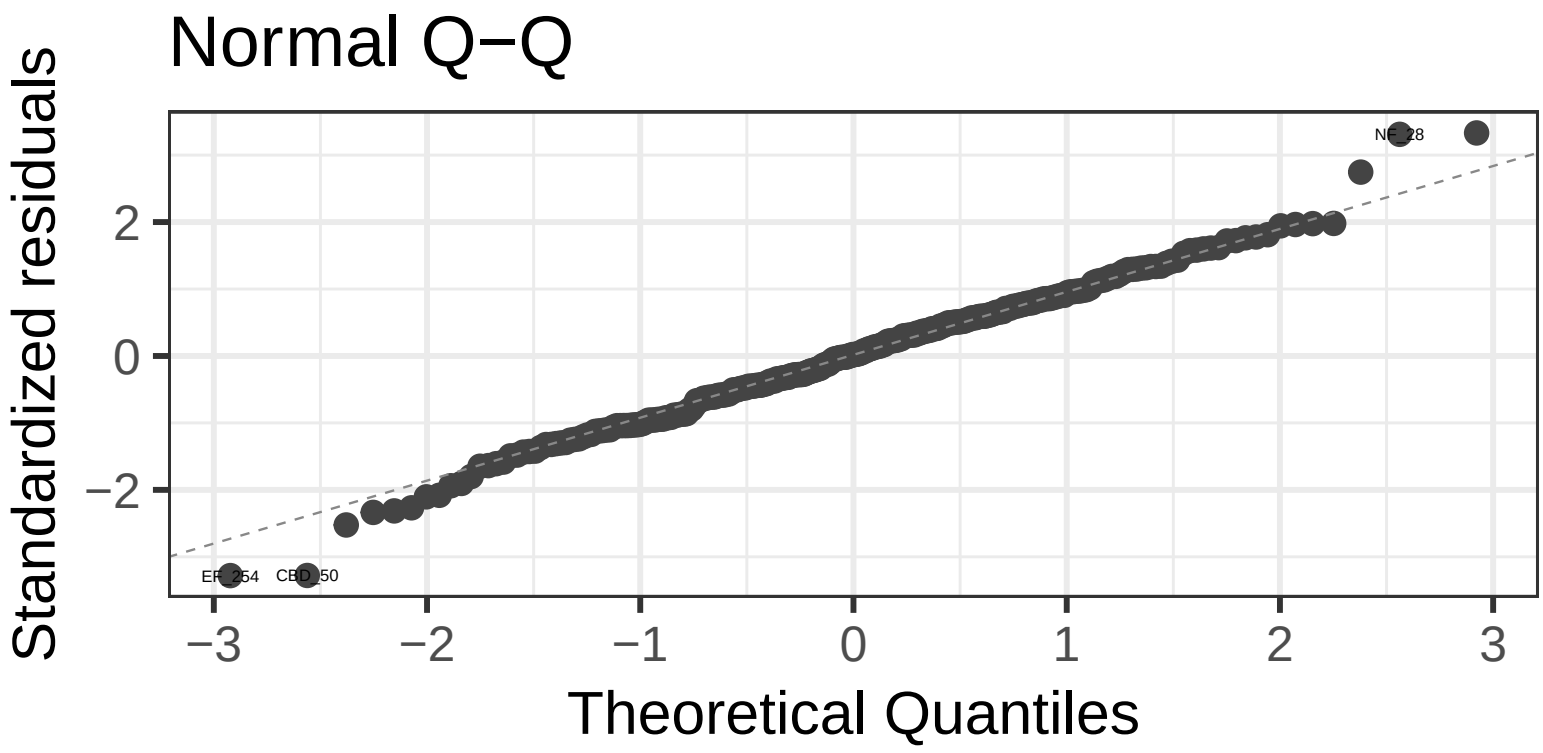



**Ce qu'on regarde:** La valeur absolue des résidus (standardisés) observés en fonction des prédictions $\hat{y}_k$.

**Ce qu'on voit:** La valeur absolue des résidus ne semble pas dépendre de la valeur des prédictions (il ne sont donc pas structurés en fonction de la prédiction). Ils sont globalement identiquement distribués autour de 0.8.

**Ce qu'on conclut:** On valide l'hypothèse de variance constante.

## Distribution normale

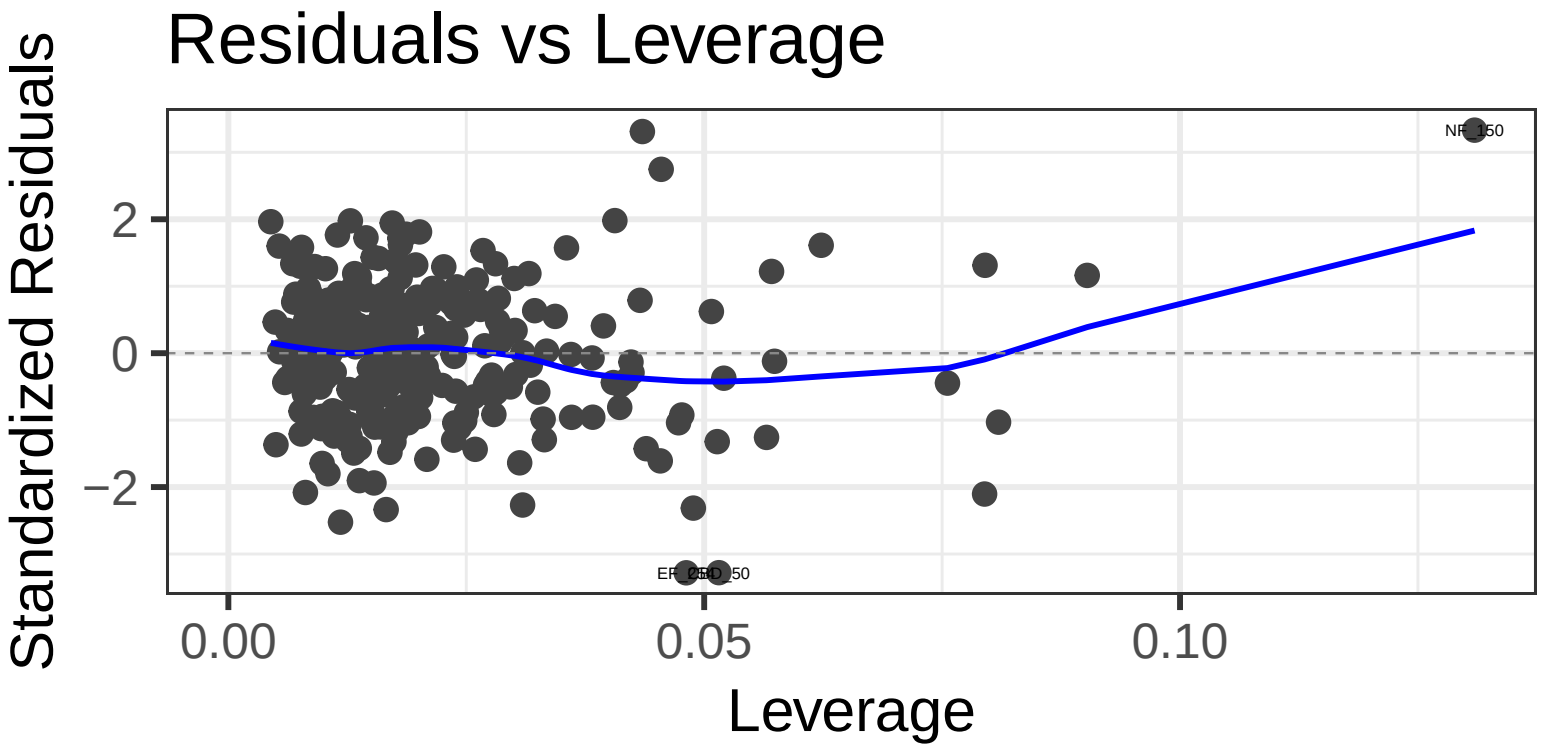



**Ce qu'on regarde:** La valeur des quantiles empiriques des résidus standardisés en fonction de la valeur quantiles théoriques d'une loi normale $\mathcal{N}(0, 1)$.

**Ce qu'on voit:** Les points sont globalement alignés sur la droite $y = x$. Les quantiles empiriques sont donc à peu près égaux aux quantiles théoriques (si les hypothèses du modèle sont vraies).

**Ce qu'on conclut:** On valide l'hypothèse de distribution normale des résidus.

## Points influents ou aberrants

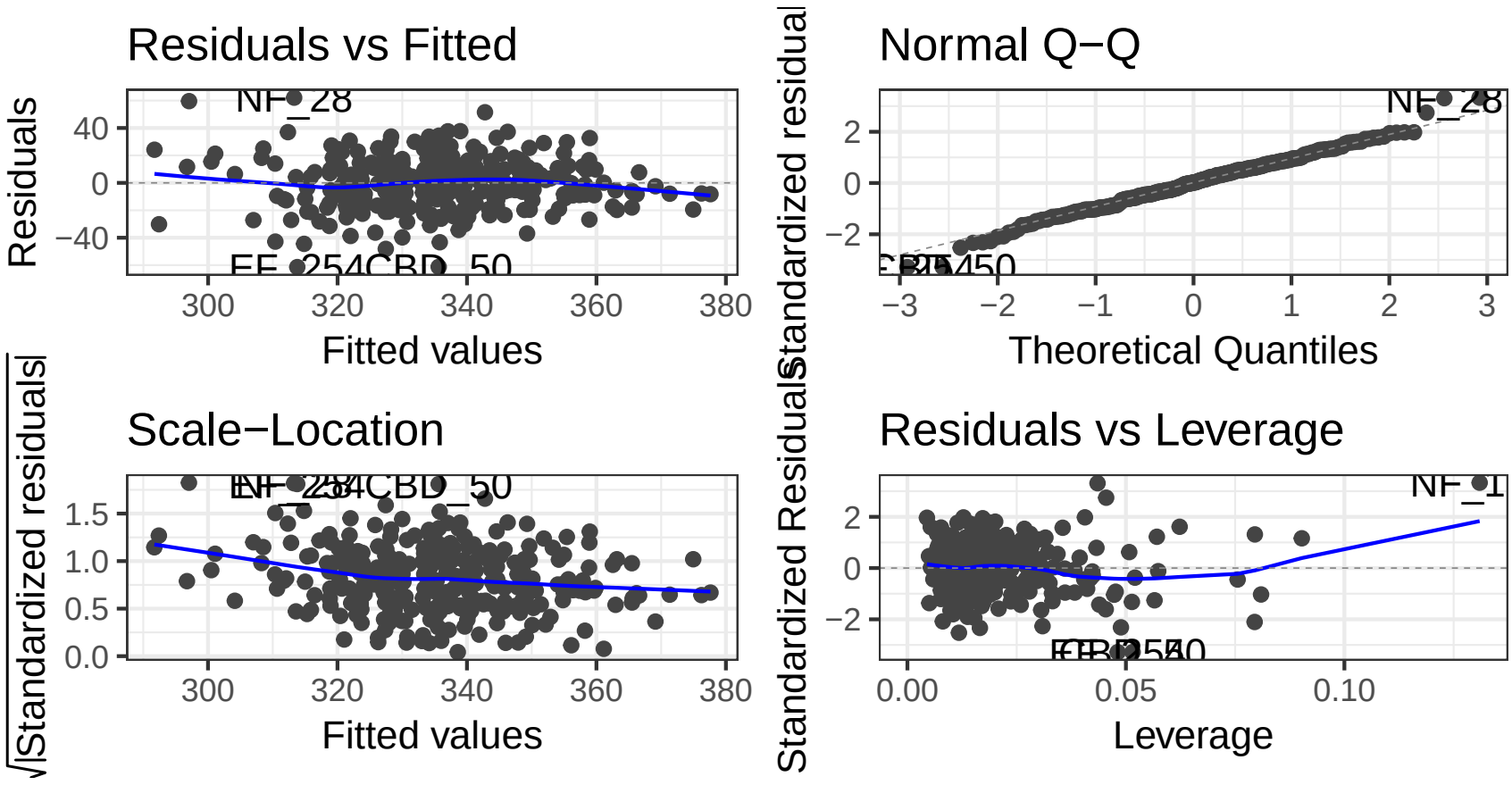



**Ce qu'on regarde:** La valeur des résidus (standardisés) en fonction du levier de l'observation (poids d'une observation dans l'estimation de sa prédiction).

**Ce qu'on voit:** Les points ont tous un petit levier, donc aucun point n'influe trop sur la droite. Aucun point n'est en dehors de l'enveloppe délimitée par les hyperboles rouges, représentant les lignes de niveau 0.5 de la distance de Cook.

**Ce qu'on conclut:** Aucun point n'est aberrant ou trop influent.

# 4 graphes et un oeil fin




Donc on valide les hypothèses du modèle pour notre exemple.

On peut maintenant tester la pertinence du modèle.

# V) Test du modèle

On veut tester si notre modèle impliquant les 5 varaibles explicatives explique mieux le **rendement** qu'un modèle simple, où le **rendement** ne dépend d'aucune de ces variables.

## Hypothèses du test

On teste:

$$\begin{array}{lll} & H_0 : & Y_k = \beta_0 + E_k & \text{Modèle } M_0 \\ \text{contre} & H_1 : & Y_k = \beta_0 + \beta_1 x_{1k} + \ldots \beta_p x_{pk} + E_k & \text{Modèle } M_1 \end{array}$$

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

$$\overset{SCT}{\sum_{k=1}^{n}(Y_k - \bar{Y})^2} = \overset{SCM}{\sum_{k=1}^{n}(\widehat{Y}_k - \bar{Y})^2} + \overset{SCR}{\sum_{k=1}^{n}(Y_k - \widehat{Y}_k)^2}$$

| Somme des carrés | Degrés de liberté (*ddl*) | Réalisation |
|---|---|---|
| *SCT*: Totale | n - 1 | $SCT_{obs} = \sum_{k=1}^{n}(y_k - \bar{y})^2$ |
| *SCM*: Modèle | p | $SCM_{obs} = \sum_{k=1}^{n}(\hat{y}_k - \bar{y})^2$ |
| *SCR*: Résiduelle | n - (p + 1) | $SCR_{obs} = \sum_{k=1}^{n}(y_k - \hat{y}_k)^2$ |

# Test du modèle

## Hypothèses du test

$$\begin{array}{rlll} & H_0 : & Y_k = \beta_0 + E_k & \text{Modèle } M_0 \\ \text{contre} & H_1 : & Y_k = \beta_0 + \beta_1 x_{1k} + \ldots \beta_p x_{pk} + E_k & \text{Modèle } M_1 \end{array}$$

## 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 $\alpha$ 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 ~ Huile + Proteine + Amidon + Floraison + Feuilles
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    287 167286
2    282 103795  5     63490 34.499 < 2.2e-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 de régression multiple 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 le paramètre de moyenne $\beta_i$ $(1 \leq i \leq 5)$, on teste:

$$\begin{array}{rll} & H_0 : & Y_k = \beta_0 + \beta_1 x_{1k} + \cdots + \beta_{i-1} x_{(i-1)k} + \beta_{i+1} x_{(i+1)k} + \cdots + \beta_p x_{pk} + E_k \\ \text{contre} & H_1 : & Y_k = \beta_0 + \beta_1 x_{1k} + \ldots \beta_p x_{pk} + E_k \end{array}$$

## Statistique de test

$$T = \frac{\hat{\beta}_i}{\sqrt{\widehat{\mathbb{V}[\hat{\beta}_i]}}}$$

où $\widehat{\mathbb{V}[\hat{\beta}_i]}$ est l'estimateur de la variance de $\hat{\beta}_i$.

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 $\alpha$ 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_reg_mult, data = donnees)

Residuals:
   Min     1Q Median     3Q    Max
-61.41 -11.72   0.39  12.38  62.11

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 67.416855  72.651009   0.928  0.35422
Huile       -3.699378   1.635315  -2.262  0.02445 *
Proteine    -0.351734   1.218506  -0.289  0.77305
Amidon       3.384079   0.777635   4.352 1.89e-05 ***
Floraison    0.005493   0.020525   0.268  0.78920
Feuilles     2.703860   0.904820   2.988  0.00305 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.19 on 282 degrees of freedom
Multiple R-squared:  0.3795,    Adjusted R-squared:  0.3685
F-statistic:  34.5 on 5 and 282 DF,  p-value: < 2.2e-16
```

# Sélection de variables

Le modèle précédent a peut être trop de variables. Il faut un critère pour sélectionner le meilleur modèle, c'est à dire:

- Qui explique bien les données;
- Qui contient le moins de variables possibles

## Critère AIC

Pour un modèle $M$, on peut définir le critère $AIC$

$$AIC(M) = -\ell_M(\mathbf{y}, \hat{\theta}^{obs}, \hat{\sigma}^2_{obs}) + 2D_M$$

où

- $\ell_M(\mathbf{y}, \hat{\theta}^{obs}$ est la log-vraisemblance du modèle , évaluée en l'estimation des paramètres (idéalement, grande!);
- $D_M$ est le nombre de paramètres du modèle $M$ (idéalement, petit!).

Le meilleur modèle parmi les $2^p$ modèles possibles est celui ayant l'$AIC$ le plus faible.

# Sélection pas à pas

```
Start:  AIC=1707.52
Rendement ~ Huile + Proteine + Amidon + Floraison + Feuilles

            Df Sum of Sq    RSS    AIC
- Floraison  1      26.4 103822 1705.6
- Proteine   1      30.7 103826 1705.6
<none>                  103795 1707.5
- Huile      1    1883.6 105679 1710.7
- Feuilles   1    3286.8 107082 1714.5
- Amidon     1    6970.4 110766 1724.2

Step:  AIC=1705.59
Rendement ~ Huile + Proteine + Amidon + Feuilles

           Df Sum of Sq    RSS    AIC
- Proteine  1      37.8 103859 1703.7
<none>                  103822 1705.6
- Huile     1    1904.4 105726 1708.8
- Amidon    1    6946.9 110769 1722.2
- Feuilles  1   27349.3 131171 1770.9

Step:  AIC=1703.7
Rendement ~ Huile + Amidon + Feuilles

           Df Sum of Sq    RSS    AIC
<none>                  103859 1703.7
- Huile     1    1975.2 105835 1707.1
- Amidon    1   20788.4 124648 1754.2
- Feuilles  1   30265.6 134125 1775.3
```

# Estimations dans le modèle final

```
Call:
lm(formula = Rendement ~ Huile + Amidon + Feuilles, data = donnees)

Residuals:
    Min      1Q  Median      3Q     Max
-62.222 -11.865   0.159  12.422  62.147

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  51.0175    35.6337   1.432   0.1533
Huile        -3.5266     1.5174  -2.324   0.0208 *
Amidon        3.5635     0.4726   7.540 6.38e-13 ***
Feuilles      2.9586     0.3252   9.097  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.12 on 284 degrees of freedom
Multiple R-squared:  0.3791,    Adjusted R-squared:  0.3726
F-statistic: 57.81 on 3 and 284 DF,  p-value: < 2.2e-16
```

# VII) Interprétation et critiques

Le modèle final nous dit que l'*huile*, l'*amidon* et le *nombre de feuilles* apportent des informations sur le rendement d'une parcelle:

- À teneur en *amidon* et *nombre de feuilles* fixés, une augmentation de la quantité d'*huile* semble s'accompagner d'une baisse de rendement (même si l'effet est faiblement significatif);
- À teneur en *huile* et *nombre de feuilles* fixés, une augmentation de la quantité d'*amidon* semble s'accompagner d'une hausse des rendements (l'effet semble cette fois fortement significatif);
- À teneur en *huile* et *amidon* fixées, une augmentation du *nombre de feuilles* semble s'accompagner d'une hausse des rendements (l'effet semble cette fois fortement significatif);

De plus, ces 3 variables explicatives sont peu corrélées, on peut donc espérer bien en distinguer les effets.

Corrélation entre les 3 variables:

| | Huile | Amidon | Feuilles |
|---|---|---|---|
| Huile | 1.00 | -0.39 | -0.04 |
| Amidon | -0.39 | 1.00 | 0.00 |
| Feuilles | -0.04 | 0.00 | 1.00 |

Le pouvoir prédictif du modèle reste cependant assez faible

# Ajustement sur un modèle standardisé

Pour chaque variable explicative $\mathbf{x_i}$, $1 \leq i \leq p$, on centre et on standardise les données, on travaillera en utilisant les mesures $\tilde{x}_{ik}$ telles que:

$$\tilde{x}_{ik} = \frac{x_{ik} - x_{i\bullet}}{\sqrt{\sum_{i=1}^{n}(x_{ik} - x_{i\bullet})^2}}$$

où $x_{i\bullet}$ est la moyenne empirique de la variable $\mathbf{x_i}$.

Cette transformation permettra de rendre les estimations $\hat{\beta}_i^{obs}$ comparables.

```
Call:
lm(formula = formule_modele_final, data = donnes_stand)

Residuals:
    Min      1Q  Median      3Q     Max
-62.222 -11.865   0.159  12.422  62.147

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  335.503      1.127 297.734  < 2e-16 ***
Huile         -2.847      1.225  -2.324   0.0208 *
Amidon         9.226      1.224   7.540 6.38e-13 ***
Feuilles      10.282      1.130   9.097  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.12 on 284 degrees of freedom
Multiple R-squared:  0.3791,    Adjusted R-squared:  0.3726
F-statistic: 57.81 on 3 and 284 DF,  p-value: < 2.2e-16
```

# VIII) **Conclusions biologiques**

On conclut que 3 variables parmi les 5 proposées semblent apportées de l'information sur le rendement.

En particulier, la teneur moyenne en amidon et le nombre de feuilles moyen par plant, qui sortent comme fortement significatives (avec une corrélation positive) dans la structuration des rendements.

De plus, il semblerait que la teneur en huile puisse avoir un léger effet négatif sur le rendement.

Il peut être intéressant maintenant de s'intéresser à l'origine du maïs et à sa potentielle influence sur le rendement.