# Modèle Linéaire Gaussien

2023-01-21

## 1. Chargement des données

Les données contiennent 9 variables et 288 observations. Il n'y a pas de valeurs manquantes dans les données.

```
##        id Marqueur        Variete Rendement    Huile Proteine   Amidon
## 1 CBD_109        2 Corn_Belt_Dent  294.3652 4.045744 13.26304 66.49133
## 2 CBD_192        2 Corn_Belt_Dent  302.0538 4.096716 12.35128 68.44154
## 3  CBD_85        2 Corn_Belt_Dent  319.3038 3.546716 13.65128 67.31654
## 4 CBD_146        2 Corn_Belt_Dent  343.8788 3.351156 13.63024 68.65412
## 5  CBD_96        2 Corn_Belt_Dent  332.5538 3.246716 14.50128 68.14154
## 6  CBD_20        1 Corn_Belt_Dent  351.6288 3.876156 14.85524 66.47912
##   Floraison Feuilles
## 1    864.26 13.97777
## 2    875.58 14.04315
## 3    984.34 16.63161
## 4   1047.40 17.26623
## 5    963.89 17.32392
## 6   1019.31 18.17008
```

## 2. Analyse Descriptive

Nous avons 3 variables qualitatives et 6 variables quantitatives dans les données.

- Variables qualitatives : *id*, *Marqueur*, *Variete*
- Variables quantitatives : *Rendement*, *Huile*, *Proteine*, *Amidon*, *Floraison*, *Feuilles*

Les variables *Rendement*, *Huile*, *Proteine*, et *Amidon* ont leur moyenne proche à leur médiane et leur écart minimum-médiane proche de leur écart médiane-maximum. Elles semblent avoir une distribution à peu près symétrique.

Les variables *Floraison* et *Feuilles* semblent avoir grande variabilité pour les valeurs au-dessus de la médiane avec un plus grand écart médiane-maximum que l'écart minimum-médiane, d'où une distribution biasée à droite.

```
##       id              Marqueur       Variete            Rendement    
##  Length:288         Min.   :1.000   Length:288         Min.   :252.4  
##  Class :character   1st Qu.:1.000   Class :character   1st Qu.:319.3  
##  Mode  :character   Median :2.000   Mode  :character   Median :337.5  
##                     Mean   :1.632                      Mean   :335.5  
##                     3rd Qu.:2.000                      3rd Qu.:352.8  
##                     Max.   :2.000                      Max.   :394.2
```

```
##      Huile          Proteine          Amidon        Floraison
##  Min.   :1.622   Min.   : 8.579   Min.   :60.03   Min.   : 808.2
##  1st Qu.:3.058   1st Qu.:12.156   1st Qu.:67.43   1st Qu.: 914.8
##  Median :3.509   Median :13.191   Median :69.16   Median : 982.7
##  Mean   :3.516   Mean   :13.144   Mean   :68.98   Mean   :1021.0
##  3rd Qu.:3.922   3rd Qu.:14.110   3rd Qu.:70.58   3rd Qu.:1092.8
##  Max.   :8.201   Max.   :19.425   Max.   :74.81   Max.   :1646.8
##     Feuilles
##  Min.   :12.08
##  1st Qu.:14.76
##  Median :16.57
##  Mean   :17.27
##  3rd Qu.:19.07
##  Max.   :30.58
```

## 2.1. Variables qualitatives

### 2.1.1. Changement des variables qualitatives en facteur

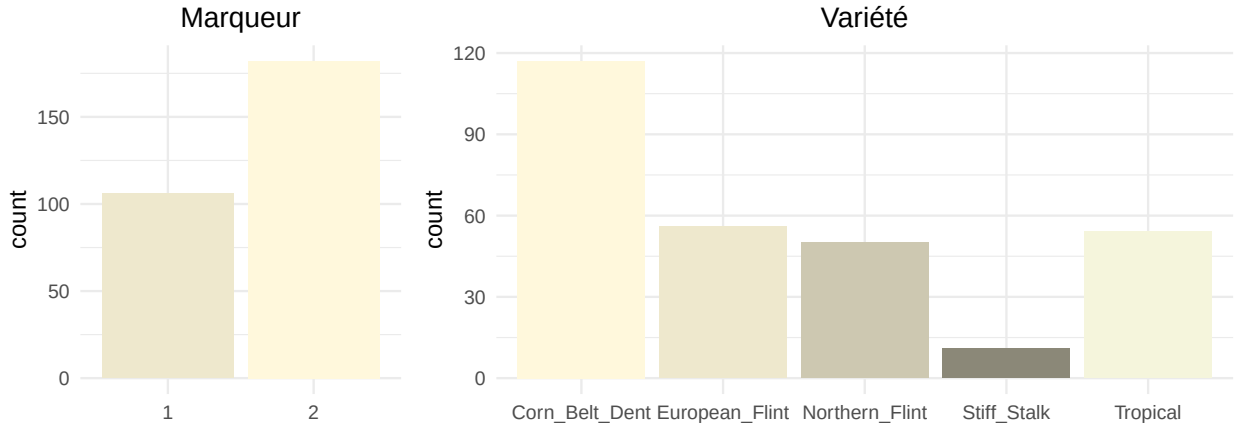
Les variables *Marqueur* et *Variete* sont à convertir en facteur pour pouvoir identifier proprement leurs modalités.

```
Marqueur = as.factor(Marqueur)
Variete = as.factor(Variete)
```

### 2.1.2. Visualisation des variables qualitatives

Nous avons une dispartié du nombre de chaque types de marqueur génétique. Le type 2 est plus présent que le type 1.

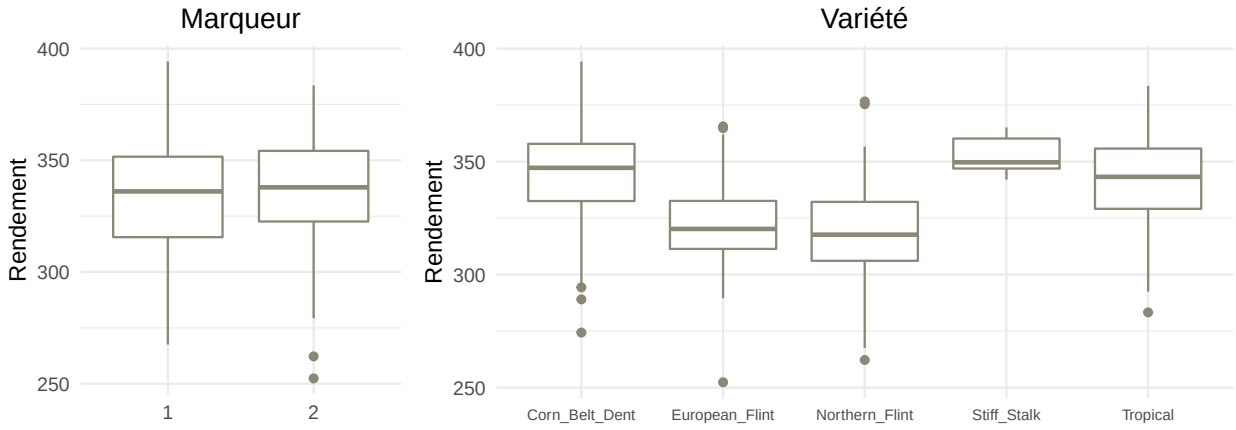
Pour la variété de mais, nous avons 5 modalités différentes et elles sont très déséquilibrées. *Corn_Belt_Dent* est la plus dominante, plus que deux fois de plus, par rapport aux autres modalités et *Stiff_Stalk* est visiblement faible.




### 2.1.3. Lien avec la variable cible

La variable réponse est *Rendement*. Il semble que le niveau de rendement de mais est similaire pour les deux marqueurs génétiques différents.

Pour la variété, la différence du niveau de rendement est apparente entre les modalités. C'est *Stiff_Stalk* qui a un niveau de rendement moyen le plus élevé et *Europe_Flint* et *Nothern_Flint* ont un niveau de rendement moyen le moins élevé. Nous observons également que la variabilité de *Stiff_Stalk* est très petite par rappport aux autres variétés.

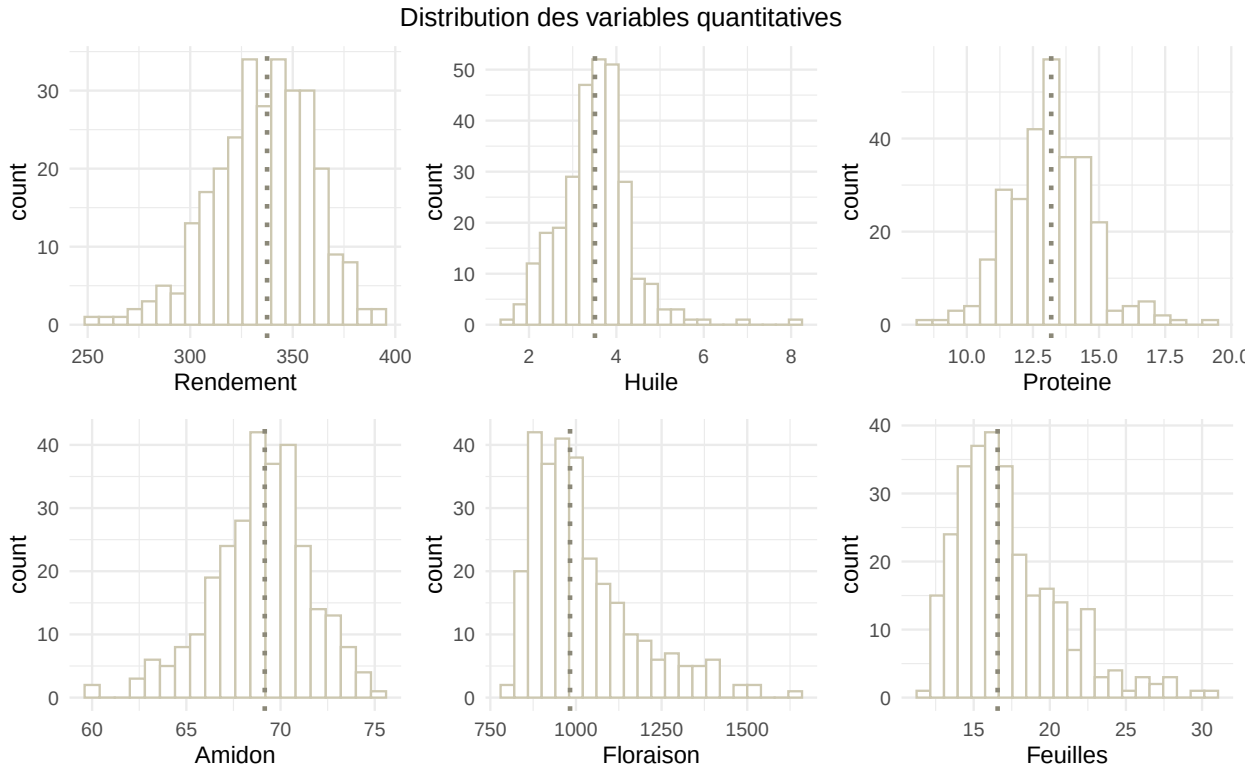



## 2.2. Variables quantitatives

### 2.2.1. Visualisation des variables quantitatives

Comme vu précédemment, les variables *Rendement*, *Huile*, *Proteine* et *Amidon* ont une distribution assez symétrique.

Les variables *Floraison* et *Feuilles* ont une distributions plus asymétriques que les quatre précédantes, avec plus de variabilité à droite.



### 2.2.2. Corrélation des variables quantitatives

Le rendement de mais est positivement corrélé avec la teneur moyenne en amidon, le nombre de degré jours moyen avant sa floraison, et son nombre moyen de feuilles. Et il est négativement corrélé avec sa teneur moyenne en huile et en proteine.

Entre les variables explicatives,

- le nombre de degré jours moyen avant la floraison et le nombre moyen de feuille présentent une corrélation positive forte.
- la teneur en amidon et en proteine présente une corrélation négative assez élevée.

```
##            Rendement       Huile    Proteine        Amidon     Floraison
## Rendement  1.0000000 -0.28405866 -0.37125637  0.4258431130  0.4074927563
## Huile     -0.2840587  1.00000000  0.07826465 -0.3855028292 -0.0425452017
## Proteine  -0.3712564  0.07826465  1.00000000 -0.7452303327 -0.1766606738
## Amidon     0.4258431 -0.38550283 -0.74523033  1.0000000000  0.0002449755
## Floraison  0.4074928 -0.04254520 -0.17666067  0.0002449755  1.0000000000
## Feuilles   0.4295111 -0.04421115 -0.16010327 -0.0041345494  0.9328629265
##               Feuilles
## Rendement  0.429511061
## Huile     -0.044211147
## Proteine  -0.160103267
## Amidon    -0.004134549
## Floraison  0.932862926
## Feuilles   1.000000000
```

La forte corrélation entre le nombre de degré jours moyen avant la floraison et le nombre moyen de feuille est significativement différente de 0.

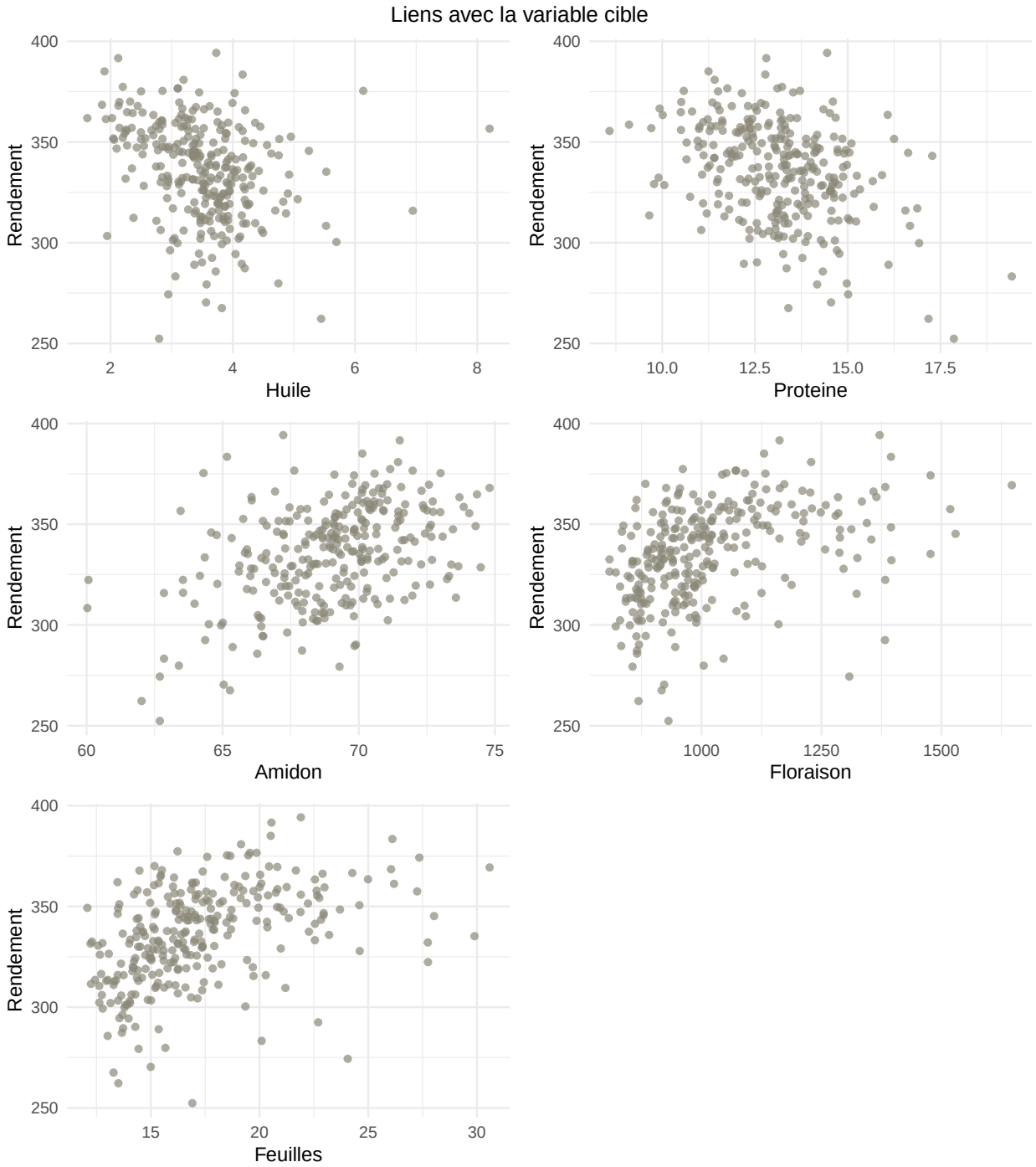
```
##
##  Pearson's product-moment correlation
##
## data:  Floraison and Feuilles
## t = 43.794, df = 286, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9160520 0.9464016
## sample estimates:
##       cor
## 0.9328629
```

La corrélation entre la teneur en amidon et en proteine est aussi significativement différente de 0.

```
##
##  Pearson's product-moment correlation
##
## data:  Amidon and Proteine
## t = -18.901, df = 286, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7925452 -0.6889964
## sample estimates:
##        cor
## -0.7452303
```

### 2.2.3. Lien avec la variable cible

Le rendement semble avoir un lien plus ou moins fort avec tous les variables explicatives quantitatives. Le rendement de mais diminue avec l'augmentation de sa teneur en huile et en proteine. Il augmente avec l'augmentation de sa teneur en amidon, du nombre de degré jours moyen avant sa floraison, et son nombre moyen de feuilles.



# 3. Teneur moyenne en amidon

## 3.1. Résumé numérique

Le rendement de mais et sa teneur moyenne en amidon présente une corrélation positive de 0,42 et cette corrélation est significativement différente de 0.

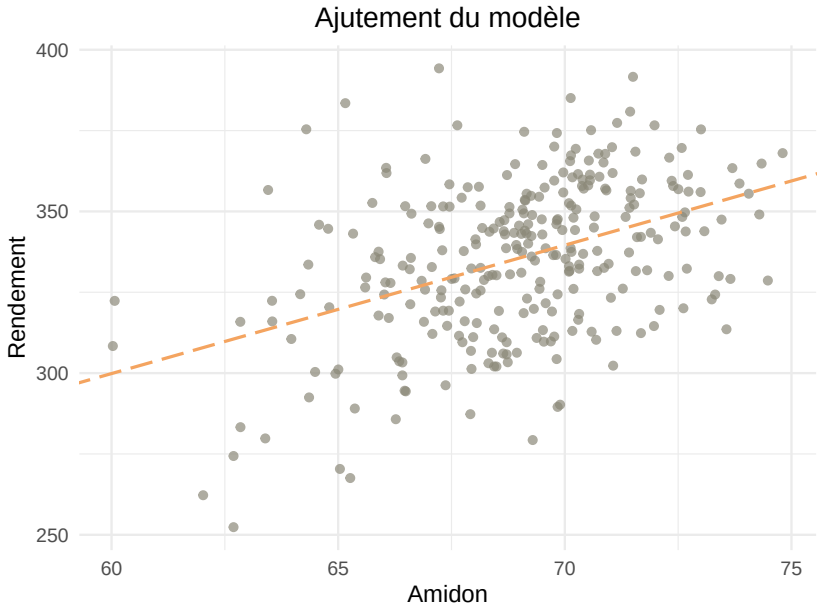
```
##
##  Pearson's product-moment correlation
##
## data:  Rendement and Amidon
## t = 7.9594, df = 286, p-value = 4.085e-14
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3263250 0.5160245
## sample estimates:
##       cor
## 0.4258431
```

## 3.2. Ajustement du modèle

Nous avons vu precedemment qu'il y a une tendance croissante. Plus la teneur en amidon de mais est élevée, plus son rendement est élevé. Nous pouvons donc effectuer l'ajustement du modèle.

Sur le graphique ci-dessous, le nuage des points représente les différentes observations et la droite pointillée représente les valeurs ajustées. La droite d'ajustement a une pente de 3,97 et une ordonnée à l'origine de 61,60.

```
## (Intercept)      Amidon
##   61.602614    3.970982
```



## 3.3. Validité des hypothèses

### 3.3.1. Hypothèses d'espérance nulle et d'homoscédasticité

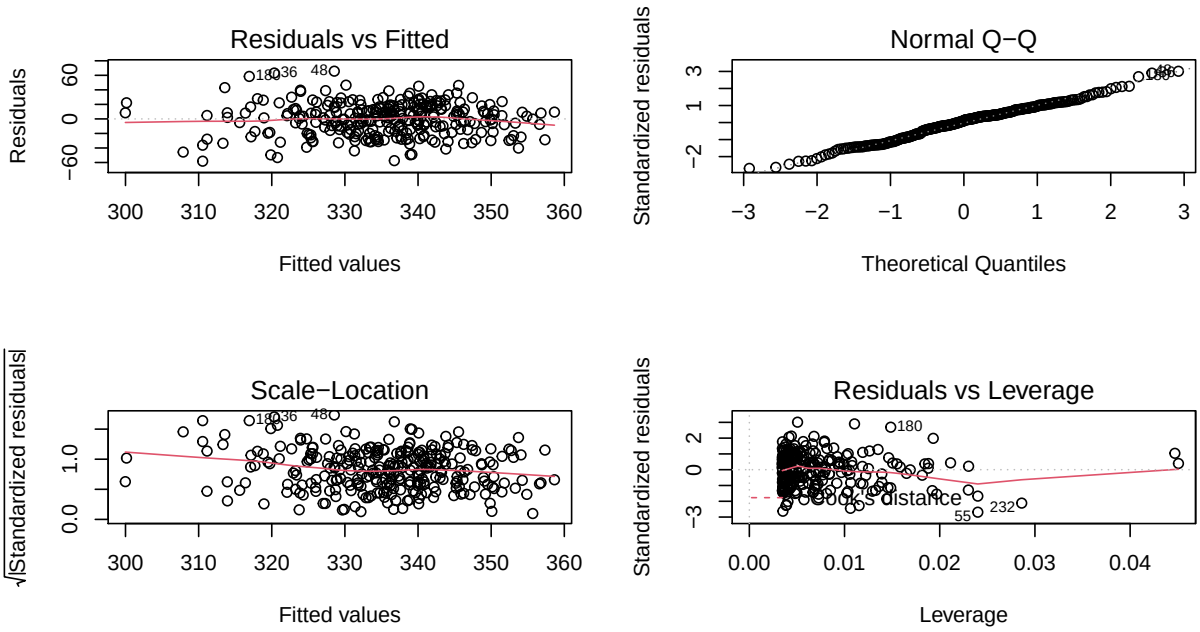
Sur le graphique *Residuals vs Fitted*, nous n'observons pas de structure particulière avec une bande horizontale. Les hypothèses d'espérance nulle et d'homoscédasticité sont vérifiées. On peut également confirmer l'hypothèse d'homoscédasticité par le graphique *Scale-Location* d'où le nuage de points est centré sans structure particulière.

### 3.3.2. Hypothèse de normalité

Sur le graphique *Normale Q-Q*, nous observons que les points sont bien alignés sur la droite qui montre la correspondance entre les quantiles empiriques et les quantiles théoriques. La normalité est vérifiée.

### 3.3.3. Valeurs atypiques

Sur le graphique *Residuals vs Leverage*, nous observons que la distance de Cook des points est inférieure à 0,5. Nous pouvons alors valider qu'il n'y a pas des observations atypiques problèmatiques.




Avec la validation de toutes les hypothèses, nous pouvons alors valider le modèle et effectuer les analyses statistiques.

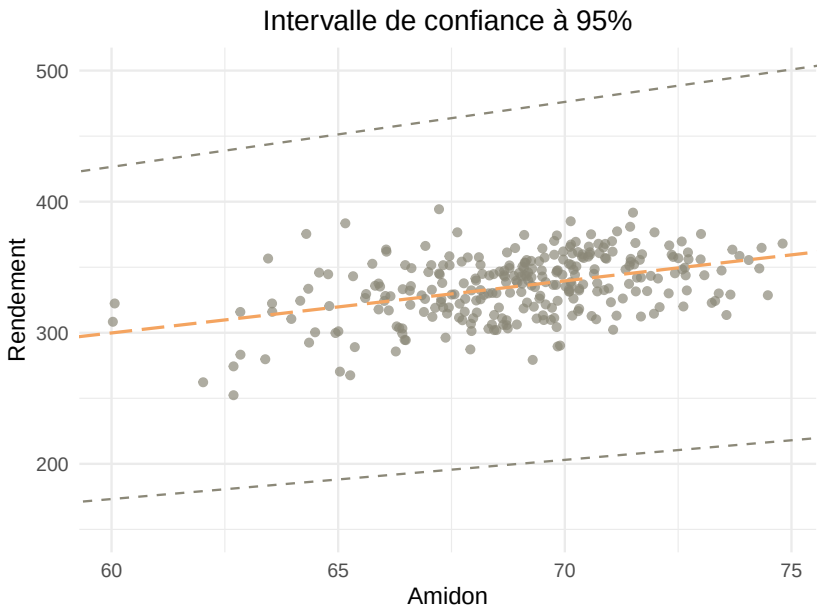
## 3.4. Intervalles de confiance

Au niveau 95%, le coefficient de la pente est différent de 0.

- nous avons l'ordonnée à l'origine entre -6,18 et 129,38
- nous avons la pente entre 2,99 et 4,95

Il semble que la variable explicative *Amidon* a une influence sur le rendement de mais.

```
##                 2.5 %    97.5 %
## (Intercept) -6.178044 129.38327
## Amidon       2.988994   4.95297
```




### 3.5. Tests statistiques

Pour tester si la variable *Amidon* est une variable explicative pertinente, nous allons effectuer un test de nullité du coefficient de *Amidon* obtenu par l'ajustement du modèle.

- L'hypothèse nulle est que le coefficient est égal à 0, c'est à dire, la teneur en amidon n'a pas d'influence sur le rendement de mais.
- L'hypothèse alternative est que le coefficient est différente de 0 qui signifie que la teneur en amidon a une influence sur le rendement de mais.

Nous avons la p-valeur proche de 0, plus petit que 5%. L'hypothèse nulle est rejetée et **nous pouvons considérer que la teneur en amidon a une influence sur le rendement de mais.**

```
##
## Call:
## lm(formula = Rendement ~ Amidon)
##
## 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
```

## 3.6. Prédiction

### 3.6.1. Création de données à prédire

Trois nouvelles observations sont créées pour prédire le rendement.

```
##   Marqueur        Variete Huile Proteine Amidon Floraison Feuilles
## 1        1 Corn_Belt_Dent  3.39     13.0  69.34      1000       17
## 2        1 European_Flint  3.54     13.3  69.41       943       15
## 3        2 Corn_Belt_Dent  2.85     11.8  67.70       934       16
```

### 3.6.2. Prédiction

Le rendement de mais pour les nouvelles observations est prédit à partir de leur teneure en amidon. Étant donné que la teneur en amidon de toutes les trois nouvelles observations est proche de sa moyenne, nous n'observons pas de grande variabilité de prédiction entre les observations.

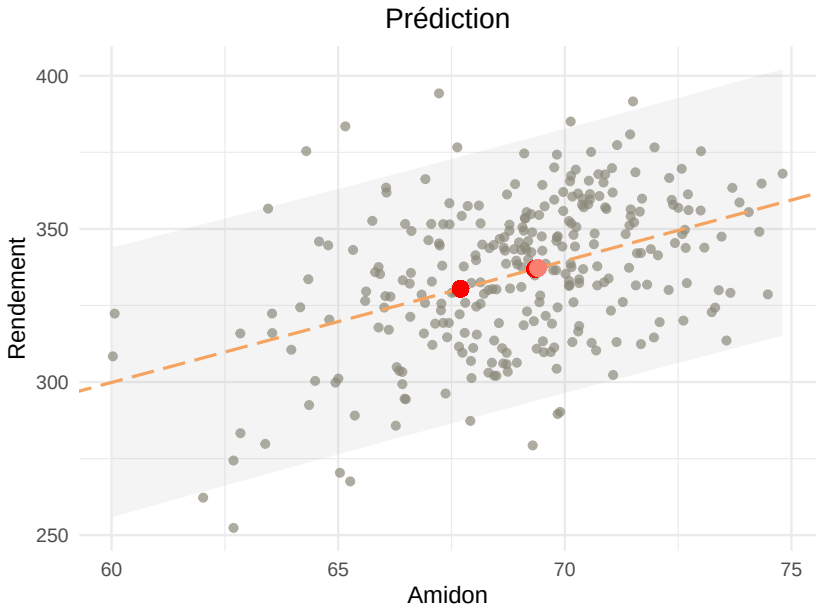
```
predict(reg_amidon, mais_new, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 336.9505 334.3874 339.5136
## 2 337.2285 334.6549 339.8021
## 3 330.4381 327.6079 333.2683
```

```
pred_amidon = predict(reg_amidon, mais_new, interval = "prediction")
pred_amidon
```

```
##        fit      lwr      upr
## 1 336.9505 293.8031 380.0979
## 2 337.2285 294.0804 380.3765
## 3 330.4381 287.2740 373.6022
```

Sur le graphique ci-dessous, la prédiction des 3 nouvelles observations est représenté par les points rouges et orange. Les aléas, c'ést à dire, l'erreur de prévision entre la valeur inconnue de nouvelles observations et leur valeur prédite, sont représentés par la bande grise sur le graphique. Ces aléas quantifient la capacité du modèle à prévoir le rendement de mais en tenant compte des aléas modélisés.




### 3.7. Critique du modèle

Le coefficient de détermination du modèle est à l'ordre de 0,18. La teneur en amidon explique 18% de la variabilité du rendement de mais. Donc la capacité prédictive du modèle est faible. Seule la teneur en amidon prédit mal les données. Nous pouvons supposer qu'il faut faire intervenir d'autres variables explicatives dans le modèle.

```
summary(reg_amidon)$adj.r.squared
```

```
## [1] 0.1784799
```

## 4. Ensemble de variables quantitatives

### 4.1. Colinéarité

Nous avons vu précedemment que tous les variables quantitatives semblent avoir un lien avec la variable réponse. Nous avons testé qu'il y avait une colinéarité importante signficative entre les variables :

- *Floraison* et *Feuilles*
- *Amidon* et *Proteine*

Nous allons également tester les colinéarités moins fortes des variables explicatives :

- *Proteine* et *Floraison* avec un coefficient de corrélation de -0,17
- *Proteine* et *Feuille* avec un coefficient de corrélation de -0,16
- *Amidon* et *Huile* avec un coefficient de corrélation de -0,38

```
##            Rendement       Huile    Proteine        Amidon     Floraison
## Rendement  1.0000000 -0.28405866 -0.37125637  0.4258431130  0.4074927563
## Huile     -0.2840587  1.00000000  0.07826465 -0.3855028292 -0.0425452017
## Proteine  -0.3712564  0.07826465  1.00000000 -0.7452303327 -0.1766606738
## Amidon     0.4258431 -0.38550283 -0.74523033  1.0000000000  0.0002449755
## Floraison  0.4074928 -0.04254520 -0.17666067  0.0002449755  1.0000000000
## Feuilles   0.4295111 -0.04421115 -0.16010327 -0.0041345494  0.9328629265
##               Feuilles
## Rendement  0.429511061
## Huile     -0.044211147
## Proteine  -0.160103267
## Amidon    -0.004134549
## Floraison  0.932862926
## Feuilles   1.000000000
```

La corrélation etnre *Proteine* et *Floraison* est significativement différente de 0.

```
##
##  Pearson's product-moment correlation
##
## data:  Proteine and Floraison
## t = -3.0353, df = 286, p-value = 0.002624
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.28639244 -0.06235443
## sample estimates:
##        cor
## -0.1766607
```

La corrélation etnre *Proteine* et *Feuilles* est significativement différente de 0.

```
##
##  Pearson's product-moment correlation
##
## data:  Proteine and Feuilles
## t = -2.743, df = 286, p-value = 0.006473
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.27067396 -0.04536328
## sample estimates:
##        cor
## -0.1601033
```

La corrélation etnre *Amidon* et *Huile* est significativement différente de 0.

```
##
##  Pearson's product-moment correlation
##
## data:  Amidon and Huile
## t = -7.0656, df = 286, p-value = 1.223e-11
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

```
##  -0.4797083 -0.2825110
## sample estimates:
##        cor
## -0.3855028
```

#### 4.1.1. Suppression des variables corrélées

La variable *Proteine* est signficativement corrélée avec plusieurs variables, donc nous allons éliminer *Proteine*.

Entre *Feuilles* et *Floraison* qui sont significativement corrélées, nous allons garder la variable *Feuilles* car elle a un coefficient de corrélation avec la variable réponse plus élevée.

Pour la même raison, nous allons garder la variable *Amidon* et éliminer la variable *Huile*.

An final, il nous reste *Amidon* et *Feuilles* comme les variables explicatives. Nous avons déjà vu que *Amidon* était significativement corrélé avec la variable réponse. Donc nous allons tester la corrélation de *Feuilles* avec la variable réponse.

La corrélation entre *Rendement* et *Feuilles* est significativement différente de 0.

```
##
##  Pearson's product-moment correlation
##
## data:  Rendement and Feuilles
## t = 8.0434, df = 286, p-value = 2.34e-14
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3303301 0.5193105
## sample estimates:
##       cor
## 0.4295111
```

#### 4.1.2. Suppression des variables corrélées avec le test de multicolinéarité

Nous pouvons aussi éliminer les variables corrélées en calculant le vif. Les variables ayant une valeur supérieure à 5 sont considérées corrélées avec d'autres variables explicative. Avec cette méthode, nous pouvons retirer *Floraison* et *Feuilles*.

```
##     Huile    Amidon  Proteine Floraison  Feuilles
##  1.358728  3.160711  2.790891  7.804027  7.709960
```

### 4.2. Ajustement du modèle

Nous pouvons créér deux modèles linéaires multiples différents avec les variables obtenues par les deux raisonnements précédants.

#### 4.2.1. Ajustement du modèle multiple 1

Nous avons vu précédemment qu'il y a une tendance croissante entre la variable réponse et la teneur en amidon et le nombre de feuilles. Plus la teneur en amidon de mais et le nombre de feuilles sont élevés, plus le rendement de mais est élevé. Nous pouvons donc effectuer l'ajustement du modèle avec les deux variables explicatives.

```
## (Intercept)      Amidon    Feuilles
##    8.720243    3.987610    2.996121
```

### 4.2.2. Ajustement du modèle multiple 2

Avec la variable *Huile* et *Proteine*, nous avons vu qu'il y a une tendance décroissante. Nous allons créer un deuxième modèle avec les variables *Huile*, *Proteine*, et *Amidon* pour expliquer la variable réponse.

```
## (Intercept)       Huile      Amidon    Proteine
##  274.570777   -5.767767    1.805243   -3.294864
```

## 4.3. Validité des hypothèses

### 4.3.1. Hypothèses d'espérance nulle et d'homoscédasticité

Sur le graphique *Residuals vs Fitted* et *Scale-Location*, le nuage de points est centré et aligné en bande sans structure particulière. Les hypothèses d'espérance nulle et d'homoscédasticité sont vérifiées

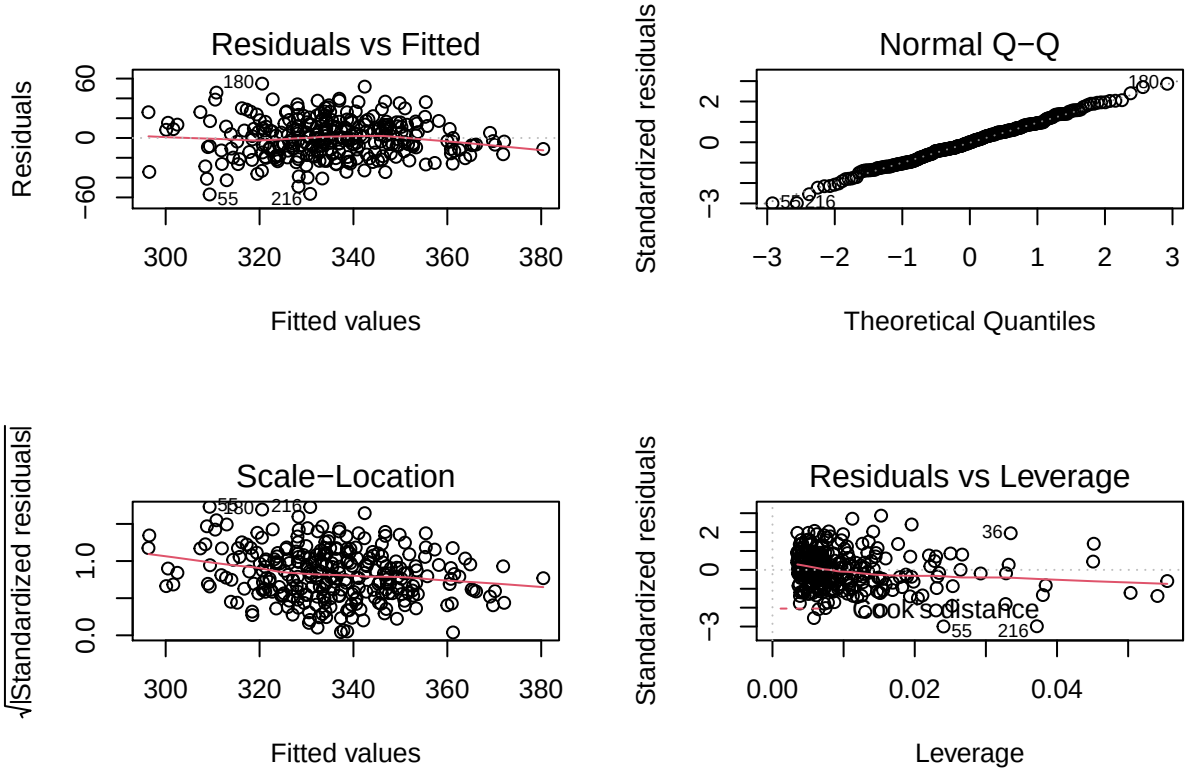
### 4.3.2. Hypothèse de normalité

Sur le graphique *Normale Q-Q*, nous observons que les points sont bien alignés sur la droite d'où la normalité.

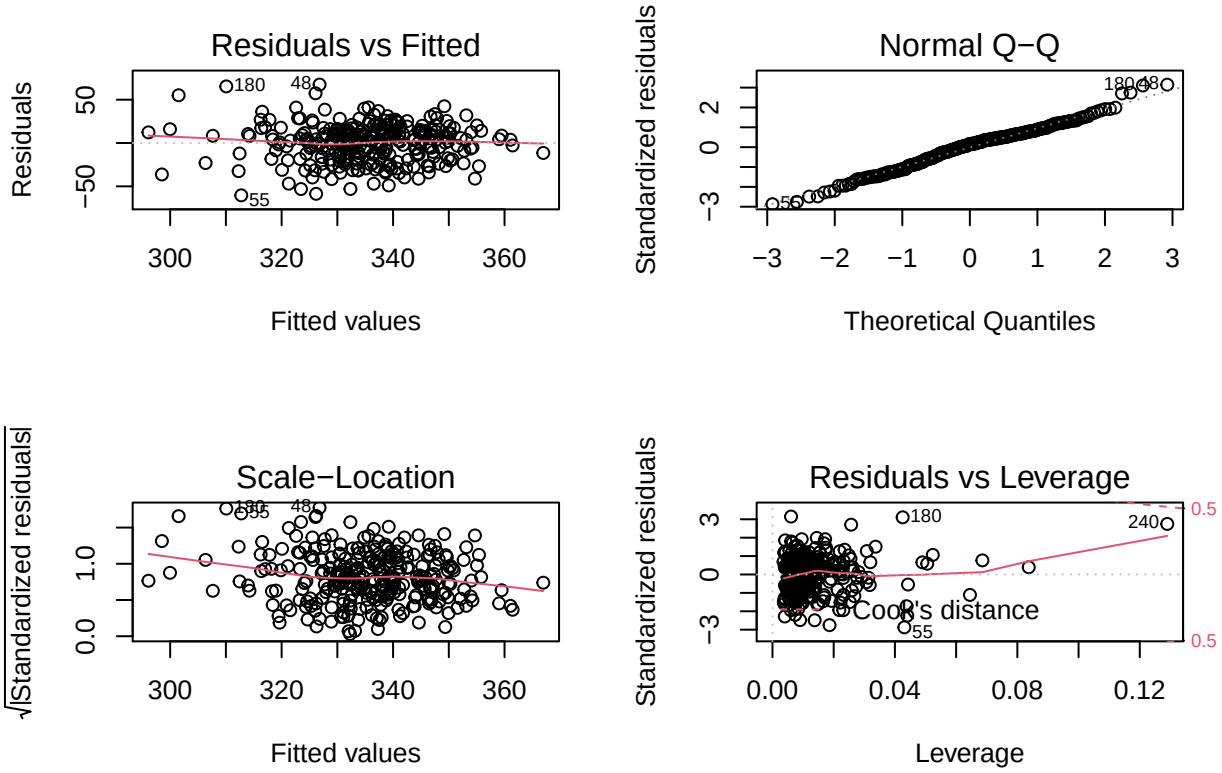
### 4.3.3. Valeurs atypiques

Sur le graphique *Residuals vs Leverage*, nous observons qu'il n'y a pas des observations atypiques problèmatiques.

- Modèle multiple 1



- Modèle multiple 2




Avec la validation de toutes les hypothèses, nous pouvons effectuer les analyses statistiques du modèle.

## 4.4. Intervalles de confiance

Nous avons une intervalle de confiance différente pour les deux modèles. L'intervalle est beaucoup plus grande pour le deuxième modèle.

- Modèle multiple 1

```
##                   2.5 %    97.5 %
## (Intercept) -52.043953 69.484439
## Amidon        3.122820  4.852400
## Feuilles      2.351858  3.640384
```

- Modèle multiple 2

```
##                    2.5 %      97.5 %
## (Intercept) 125.7744606 423.3670941
## Huile        -9.3322777  -2.2032554
## Amidon        0.1436911   3.4667945
## Proteine     -5.8592564  -0.7304707
```

## 4.5. Tests statistiques

### 4.5.1. Test de Fisher

Pour vérifier la pertinence de l'ensemble de coefficients, nous pouvons tester les hypothèses suivantes :

- Hypothèse nulle : les coefficients de variables explicatives sont simultanément égaux à 0.
- Hypothèse alternative : les coefficients de variables explicatives ne sont pas simultanément nuls.

Le test de Fisher entre le modèle linéaire simple et les deux modèles linéaires multiples permet effectuer le test.

- Modèle simple vs Modèle multiple 1

Nous avons une p-valeur inférieure à 5% donc nous pouvons conserver le modèle multiple avec les variables explicatives *Amidon* et *Feuilles*.

```
## Analysis of Variance Table
##
## Model 1: Rendement ~ Amidon
## Model 2: Rendement ~ Amidon + Feuilles
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    286 136950
## 2    285 105835  1     31115 83.788 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Modèle simple vs Modèle multiple 2

Nous avons une p-valeur inférieure à 5% donc nous pouvons conserver le modèle multiple avec les variables explicatives *Amidon*, *Huile* et *Proteine*.

```
## Analysis of Variance Table
##
## Model 1: Rendement ~ Amidon
## Model 2: Rendement ~ Huile + Amidon + Proteine
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    286 136950
## 2    284 131171  2    5778.6 6.2556 0.002195 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### 4.5.2. Pertinence de coefficients et test du modèle

Nous pouvons effectuer le test de pertinence de coefficient de chaque variable explicative et de l'ensemble des variables explicatives.

- Modèle multiple 1

Pour les variables explicatives *Amidon* et *Feuilles*, nous avons une p-valeur est inférieure à 5% donc l'hypothèse nulle est rejetté au niveau 5%. Nous pouvons conclure que la teneur en amidon et le nombre de feuilles de mais a une influence sur son rendement.

La p-valeur du modèle est inférieure à 5% alors le modèle linéaire multiple explique mieux les données qu'un modèle constant qui n'a que des effets de bruits.

```
##
## Call:
## lm(formula = Rendement ~ Amidon + Feuilles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.004 -13.087  -0.044  12.269  54.849
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.7202    30.8711   0.282    0.778
## Amidon        3.9876     0.4394   9.076   <2e-16 ***
## Feuilles      2.9961     0.3273   9.154   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.27 on 285 degrees of freedom
## Multiple R-squared:  0.3673, Adjusted R-squared:  0.3629
## F-statistic: 82.74 on 2 and 285 DF,  p-value: < 2.2e-16
```

- Modèle multiple 2

Pour les trois variables explicatives, nous avons une p-valeur est inférieure à 5% donc l'hypothèse nulle est rejetté au niveau 5%. Nous pouvons conclure que l'ensemble de la teneur en huile, en amidon et en proteine de mais a une influence sur son rendement.

La p-valeur du modèle est inférieure à 5% alors le modèle linéaire multiple explique mieux les données qu'un modèle constant qui n'aurait que de variabilité dû aux bruits.

```
##
## Call:
## lm(formula = Rendement ~ Huile + Amidon + Proteine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.397 -14.261   3.375  13.619  67.437
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 274.5708    75.5944   3.632 0.000333 ***
## Huile        -5.7678     1.8109  -3.185 0.001609 **
## Amidon        1.8052     0.8441   2.139 0.033323 *
## Proteine     -3.2949     1.3028  -2.529 0.011979 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.49 on 284 degrees of freedom
## Multiple R-squared:  0.2159, Adjusted R-squared:  0.2076
## F-statistic: 26.06 on 3 and 284 DF,  p-value: 6.37e-15
```

## 4.6. Prédiction

### 4.6.1. Les données à prédire

Nous pouvons effectuer la prédiction avec les trois nouvelles observations créés précédemment.

```
##   Marqueur        Variete Huile Proteine Amidon Floraison Feuilles
## 1        1 Corn_Belt_Dent  3.39     13.0  69.34      1000       17
## 2        1 European_Flint  3.54     13.3  69.41       943       15
## 3        2 Corn_Belt_Dent  2.85     11.8  67.70       934       16
```

### 4.6.2. Prédiction de nouvelles observations

Nous remarquons que le premier modèle multiple prédit un rendement plus bas avec moins de variabilité de prédiction par rapport au deuxième modèle.

- Modèle multiple 1

```
##        fit      lwr      upr
## 1 336.1552 298.1572 374.1531
## 2 330.4421 292.4159 368.4682
## 3 326.6194 288.5982 364.6405
```

- Modèle multiple 2

```
##        fit      lwr      upr
## 1 337.3604 294.9824 379.7383
## 2 335.6331 293.2431 378.0231
## 3 341.4682 298.5783 384.3581
```

## 4.7. Critique du modèle

Le coefficient de détermination du premier modèle est à l'ordre de 0,36 et celui du deuxième modèle est à l'ordre de 0,21. Le premier modèle prédit mieux que le deuxième modèle mais sa capacité prédictive n'est pas satisfaisante. L'ensemble de la teneur en amidon et du nombre de feuilles de mais explique 36% de la variabilité de son rendement.

- Modèle multiple 1

```
summary(reg_amidon_feuille)$adj.r.squared
```

```
## [1] 0.3629011
```

- Modèle multiple 2

```
summary(reg_amidon_huile_proteine)$adj.r.squared
```

```
## [1] 0.2076027
```

# 5. Variété

## 5.1. Détection de problème

Lors de l'analyse descriptive, nous avons remarqué pour la variable *Variete*, une faible présence et une petite variabilité d'une modalité par rapport aux autres modalités. Nous pouvons supposer que ce deséquilibre peut probablement poser un problème.

Nous allons effectuer l'ajustement du modèle avec la variable *Variete* et vérifier les hypothèses du modèle linéaire gaussien.

### 5.1.1. Ajustement du modèle

```
##           (Intercept) VarieteEuropean_Flint VarieteNorthern_Flint
##             344.714833            -21.777356            -26.948585
##    VarieteStiff_Stalk       VarieteTropical
##               8.395406             -3.303132
```

### 5.1.2. Vérification d'hypothèses

- Hypothèses d'espérance nulle et d'homoscédasticité

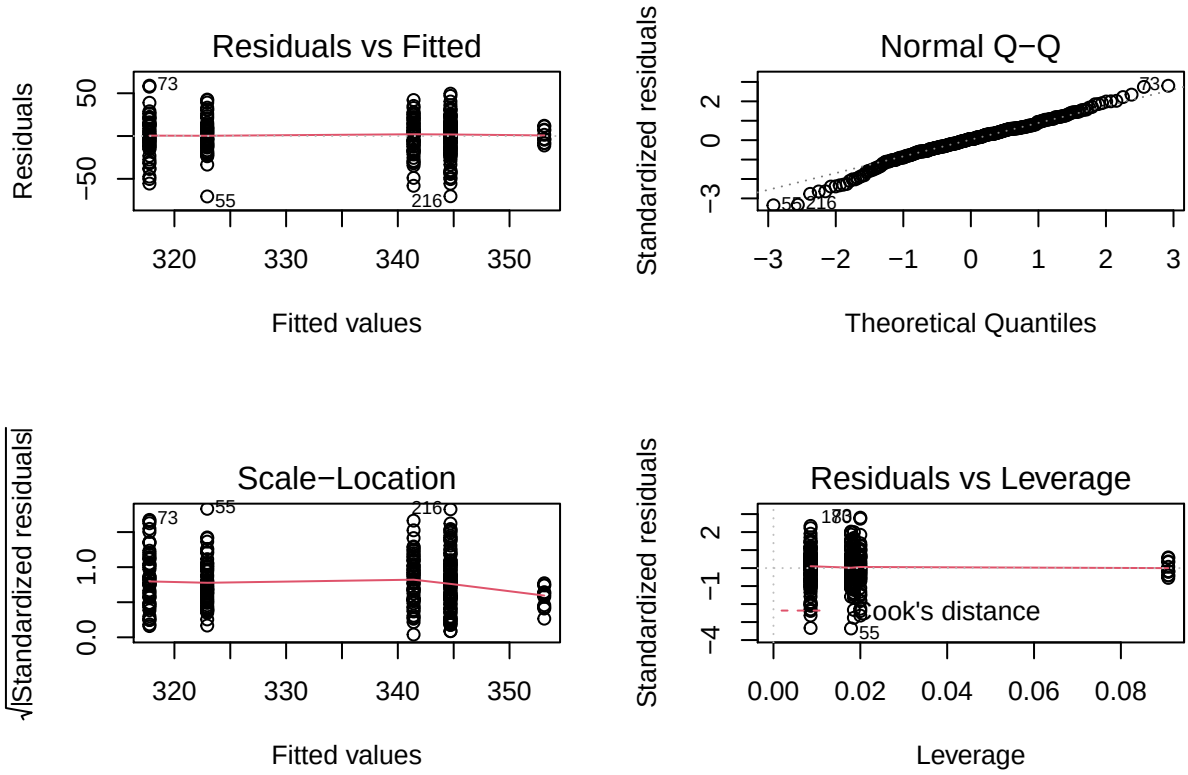
Sur le graphique *Residuals vs Fitted*, les points sont centrés mais ne forme pas d'une bande horizontale à cause d'un groupe de points plus court que les autres. L'hypothèse d'espérance nulle est validée mais pas celle d'homoscédasticité. Nous observons également une structure qui penche à la droite sur la modalité minoritaire sur le graphique *Scale-Location*. Il y a un problème d'hétérosédasticité.

- Hypothèse de normalité

Sur le graphique *Normale Q-Q*, nous observons que les points sont bien alignés sur la droite qui montre la correspondance entre les quantiles empiriques et les quantiles théoriques. La normalité est vérifiée.

- Valeurs atypiques

Sur le graphique *Residuals vs Leverage*, nous observons que la distance de Cook des points est inférieure à 1. Nous pouvons alors valider qu'il n'y a pas des observations atypiques problèmatiques.




La présence de variété minoritaire, *Stiff_Stalk* pose le problème d'hétérosédasticité.

#### 5.1.3. Création de nouvelles données

Nous allons alors créer nouvelles données sans la variété *Stiff_Stalk*.

```
##
## Corn_Belt_Dent European_Flint Northern_Flint       Tropical
##            117             56             50             54
```

## 5.2. Ajustement du modèle

Nous avons observé précédemment qu'il est probable d'avoir un lien entre le rendement et les différentes variétés. Nous allons donc ajuster le modèle avec les nouvelles données.

```
##           (Intercept) VarieteEuropean_Flint VarieteNorthern_Flint
##             344.714833           -21.777356           -26.948585
##       VarieteTropical
##            -3.303132
```

## 5.3. Validité des hypothèses

Avec nouvelles données, nous pouvons valider tous les hypothèses du modèle linéaire guassien.

### 5.3.1. Hypothèses d'espérance nulle et d'homoscédasticité

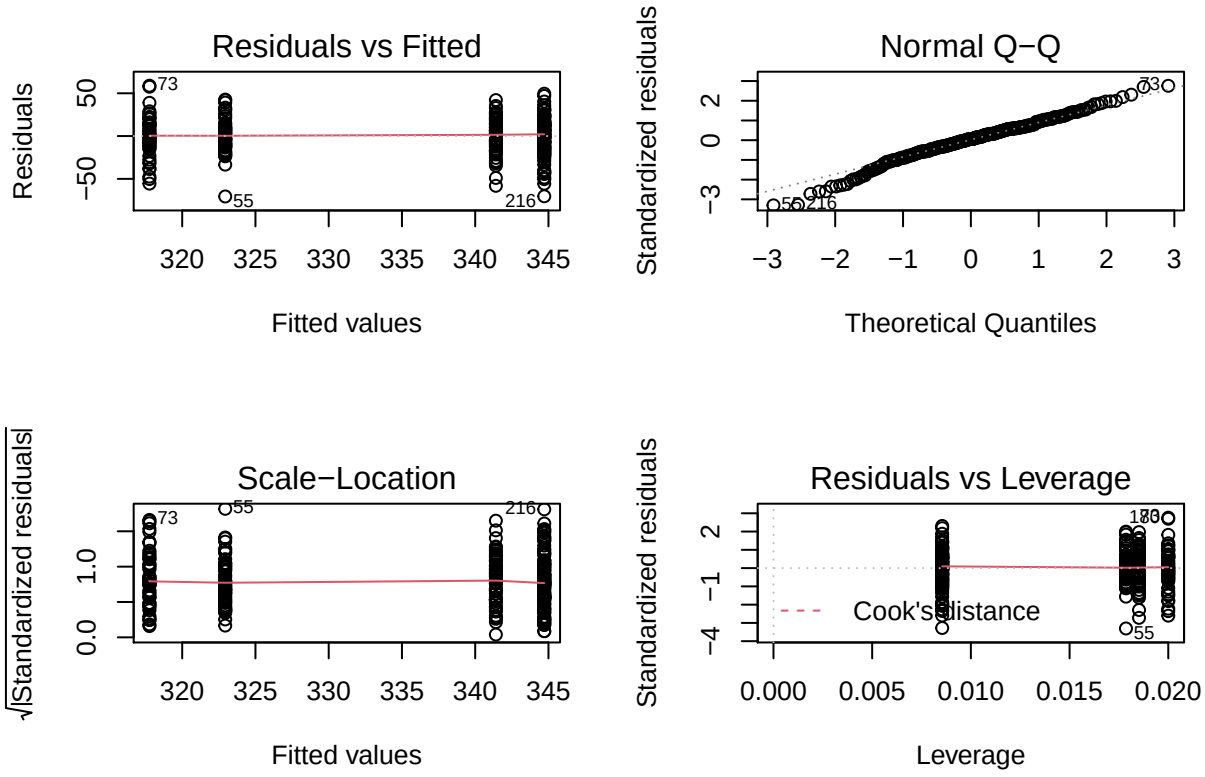
Sur le graphique *Residuals vs Fitted*, nous n'observons pas de structure particulière avec une bande horizontale. Les hypothèses d'espérance nulle et d'homoscédasticité sont vérifiées. On peut également confirmer l'hypothèse d'homoscédasticité par le graphique *Scale-Location* qui ne montre plus une structure particulière.

### 5.3.2. Hypothèse de normalité

Sur le graphique *Normale Q-Q*, nous observons que les points sont bien alignés sur la droite qui montre la correspondance entre les quantiles empiriques et les quantiles théoriques. La normalité est vérifiée.

### 5.3.3. Valeurs atypiques

Sur le graphique *Residuals vs Leverage*, nous observons que la distance de Cook des points est inférieure à 1. Nous pouvons alors valider qu'il n'y a pas des observations atypiques problèmatiques.




## 5.4. Intervalles de confiance

Dans les 95% de cas, les observations appartiennent dans l'intervalle de confiance ci-dessus.

```
##                            2.5 %     97.5 %
## (Intercept)           340.79211 348.637556
## VarieteEuropean_Flint -28.67208 -14.882635
## VarieteNorthern_Flint -34.11762 -19.779547
## VarieteTropical       -10.28367   3.677405
```

## 5.5. Tests statistiques

### 5.5.1. Test du modèle

Pour vérifier si le type de variété a une influence sur le rendement de mais, nous pouvons effectuer le test de Fisher qui permettra de voir si la variation du modèle est dû à la variation interne de chaque variété ou si elle est dû à la variation entre les variétés.

- L'hypothèse nulle est que les coefficients de tous les groupes sont égaux à 0, c'est à dire que la fluctuation du modèle est dû aux bruits.
- L'hypothèse alternative est que les coefficients d'au moins d'un couple de variétés sont différents de 0 qui signifie que la différence de variété a une influence sur le rendement de mais.

Nous avons la p-valeur plus petit que 5%. L'hypothèse nulle est rejetée et **nous pouvons considérer que la variété a une influence sur le rendement de mais.**

```
##
## Call:
## lm(formula = Rendement ~ Variete, data = mais_bis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -70.572 -12.059   1.339  13.150  58.863
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             344.715      1.993 173.002  < 2e-16 ***
## VarieteEuropean_Flint  -21.777      3.502  -6.218 1.88e-09 ***
## VarieteNorthern_Flint  -26.949      3.642  -7.400 1.68e-12 ***
## VarieteTropical         -3.303      3.546  -0.932    0.352
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.55 on 273 degrees of freedom
## Multiple R-squared:  0.2223, Adjusted R-squared:  0.2138
## F-statistic: 26.01 on 3 and 273 DF,  p-value: 7.864e-15
```

### 5.5.2. Test d'égalité des moyennes.

Maintenant que nous connaissons qu'il existe au moins deux modalités avec une moyenne différente, nous pouvons identifier les couples ayant une moyenne différente par le test de Student.

Pour les paires *European_Flint & Northern_Flint* et *Corn_Belt_Dent & Tropical*, les p-valeurs sont visiblement supérieures à 5%. Nous ne pouvons pas confirmer qu'ils ont une moyenne différente.

Pours tous les autres paires, c'est à dire, *Corn_Belt_Dent & European_Flint*, *Corn_Belt_Dent & Northern_Flint*, *European_Flint & Tropical*, *Northern_Flint & Tropical*, les p-valeurs sont inférieures à 5%. Ils ont une comportement différents au niveau 5%.

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  mais_bis$Rendement and mais_bis$Variete
```

```
##
##                Corn_Belt_Dent European_Flint Northern_Flint
## European_Flint 1.1e-08        -              -
## Northern_Flint 1.0e-11        1              -
## Tropical       1              6.2e-05        3.3e-07
##
## P value adjustment method: bonferroni
```

### 5.6. Prédiction

Toutes les nouvelles observations seraient prédites à la moyenne du groupe.

```
##        fit      lwr      upr
## 1 344.7148 302.1032 387.3265
## 2 322.9375 280.1296 365.7454
## 3 344.7148 302.1032 387.3265
```

### 5.7. Critique du modèle

Le coefficient de détermination du modèle est à l'ordre de 0,21. La variété explique 21% de la variabilité du rendement de mais. Le modèle avec la variété seule a une capacité prédictive faible.

```
summary(reg_variete)$adj.r.squared
```

```
## [1] 0.2137581
```

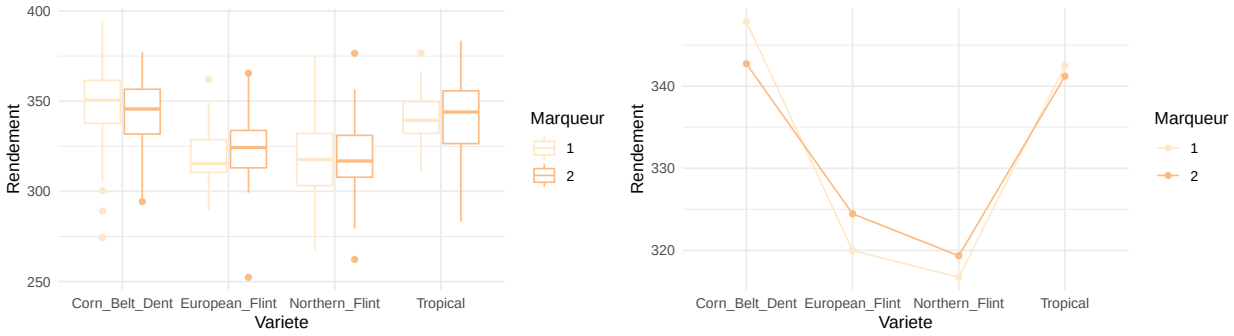
## 6. Variété & Marqueur génétique

### 6.1. Vérification préliminaire

```
##          Variete Marqueur Rendement
## 1 Corn_Belt_Dent        1  347.8866
## 2 European_Flint        1  319.9829
## 3 Northern_Flint        1  316.7196
## 4       Tropical        1  342.4842
## 5 Corn_Belt_Dent        2  342.7325
## 6 European_Flint        2  324.4547
## 7 Northern_Flint        2  319.3362
## 8       Tropical        2  341.2252
```

Sur le graphique de boîte à moustache nous observons que le niveau de rendement est différent selon la variété de mais. La variété semble avoir un effet sur le rendement.

Sur le graphique d'interaction, nous observons croisement des droites qui décrivent le niveau de rendement en fonction de variété pour chaque marqueur. Il semble donc qu'il existe une interaction entre *Variete* et *Marqueur*.




## 6.2. Ajustement du modèle

Nous allons donc ajuster le modèle avec les variables *Variete* et *Marqueur*.

```
##                       (Intercept)           VarieteEuropean_Flint
##                        347.886554                     -27.903652
##           VarieteNorthern_Flint                  VarieteTropical
##                       -31.166936                      -5.402369
##                         Marqueur2 VarieteEuropean_Flint:Marqueur2
##                        -5.154046                       9.625836
## VarieteNorthern_Flint:Marqueur2       VarieteTropical:Marqueur2
##                         7.770622                       3.895044
```

## 6.3. Validité des hypothèses

### 6.3.1. Hypothèses d'espérance nulle et d'homoscédasticité

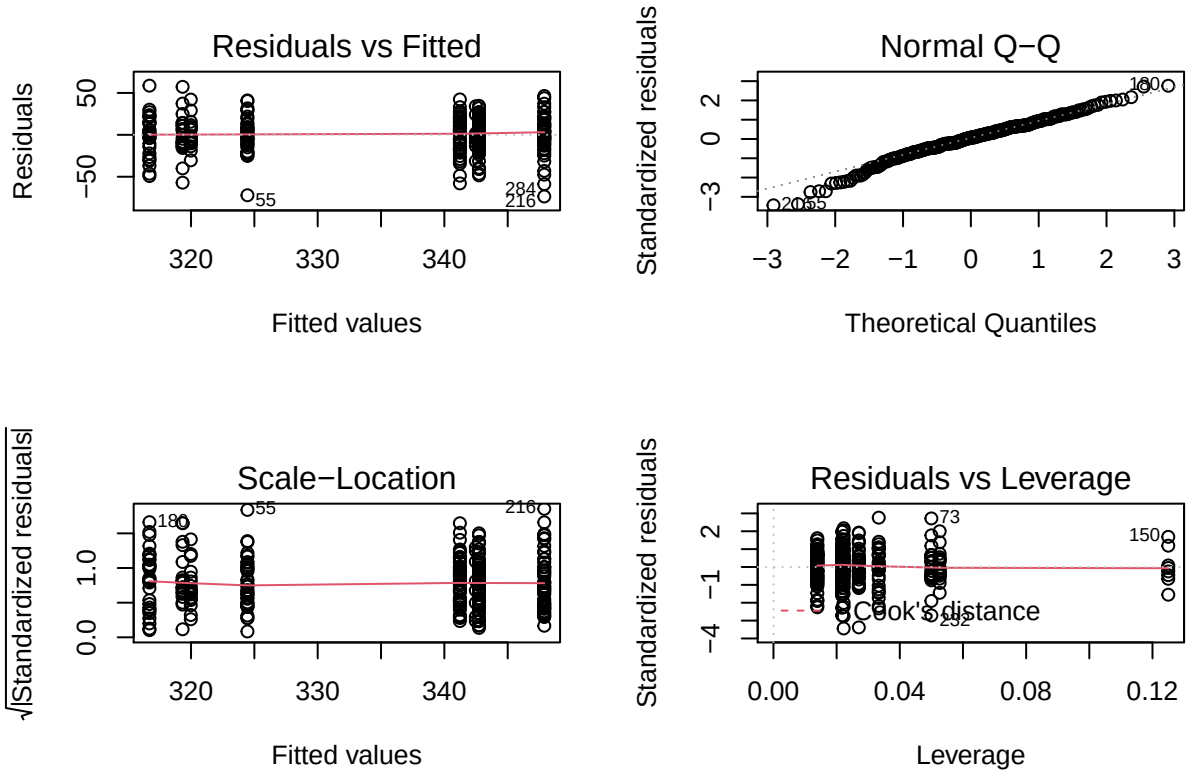
Sur le graphique *Residuals vs Fitted*, nous n'observons pas de structure particulière avec une bande horizontale. Les hypothèses d'espérance nulle et d'homoscédasticité sont vérifiées. On peut également confirmer l'hypothèse d'homoscédasticité par le graphique *Scale-Location* où aucune structure particulière est présente.

### 6.3.2. Hypothèse de normalité

Sur le graphique *Normale Q-Q*, nous observons que les points sont bien alignés sur la droite. La normalité est vérifiée.

### 6.3.3. Valeurs atypiques

Sur le graphique *Residuals vs Leverage*, nous observons que la distance de Cook des points est inférieure à 1. Nous pouvons alors valider qu'il n'y a pas des observations atypiques problèmatiques.




## 6.4. Intervalles de confiance

Nous trouvons l'intervalle de confiance suivante pour le modèle :

```
##                                      2.5 %     97.5 %
## (Intercept)                      341.541260 354.231847
## VarieteEuropean_Flint            -39.549333 -16.257971
## VarieteNorthern_Flint            -41.199726 -21.134146
## VarieteTropical                  -21.734567  10.929829
## Marqueur2                        -13.242740   2.934647
## VarieteEuropean_Flint:Marqueur2   -4.857061  24.108733
## VarieteNorthern_Flint:Marqueur2   -6.940336  22.481579
## VarieteTropical:Marqueur2        -14.306395  22.096483
```

## 6.5. Tests statistiques

### 6.5.1. Test du modèle

Pour vérifier si l'ensemble de variété et de marqueur a une influence sur le rendement de mais, nous pouvons effectuer le test de Fisher.

Nous avons la p-valeur plus petite que 5%. L'hypothèse nulle est rejetée et **nous pouvons considérer que l'ensemble variété et marqueur génétique de mais a une influence sur son rendement**.

```
##
## Call:
## lm(formula = Rendement ~ Variete * Marqueur, data = mais_bis)
##
## 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 ***
## VarieteEuropean_Flint            -27.904      5.915  -4.717 3.84e-06 ***
## VarieteNorthern_Flint            -31.167      5.096  -6.116 3.36e-09 ***
## VarieteTropical                   -5.402      8.295  -0.651    0.515
## Marqueur2                         -5.154      4.108  -1.255    0.211
## VarieteEuropean_Flint:Marqueur2    9.626      7.356   1.309    0.192
## VarieteNorthern_Flint:Marqueur2    7.771      7.472   1.040    0.299
## VarieteTropical:Marqueur2          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
```

### 6.5.2. Test d'effet des variables explicatives

Avec une p-valeur inférieure à 5%, l'ajout de variable *Variete* dans le modèle est pertinent. L'ajout de variable *Marqueur* et l'interaction des variables *Variete* et *Marqueur* ont respectivement une p-valeur de 70% et de 54% donc ces derniers n'ont pas de pouvoir explicatif.

Donc le modèle avec une seule variable explicative *Variete* serait plus pertinent que le modèle avec les deux variables *Variete* et *Marqueur*.

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

## 6.6. Prédiction

Toutes les nouvelles observations seraient prédites à la moyenne du groupe.

```
##        fit      lwr      upr
## 1 347.8866 304.8507 390.9224
## 2 319.9829 276.3116 363.6542
## 3 342.7325 299.8724 385.5926
```

### 6.7. Critique du modèle

Le coefficient de détermination du modèle est à l'ordre de 0,21. L'ensemble de variété et marqueur des mais explique 21% de la variabilité de son rendement. Le modèle a une capacité prédictive faible, similaire au modèle précédant avec seule variable explicative *Variete*. Cela montre bien que l'ajout de la variable *Marqueur* au modèle ayant la variable *Variete* n'est pertinent.

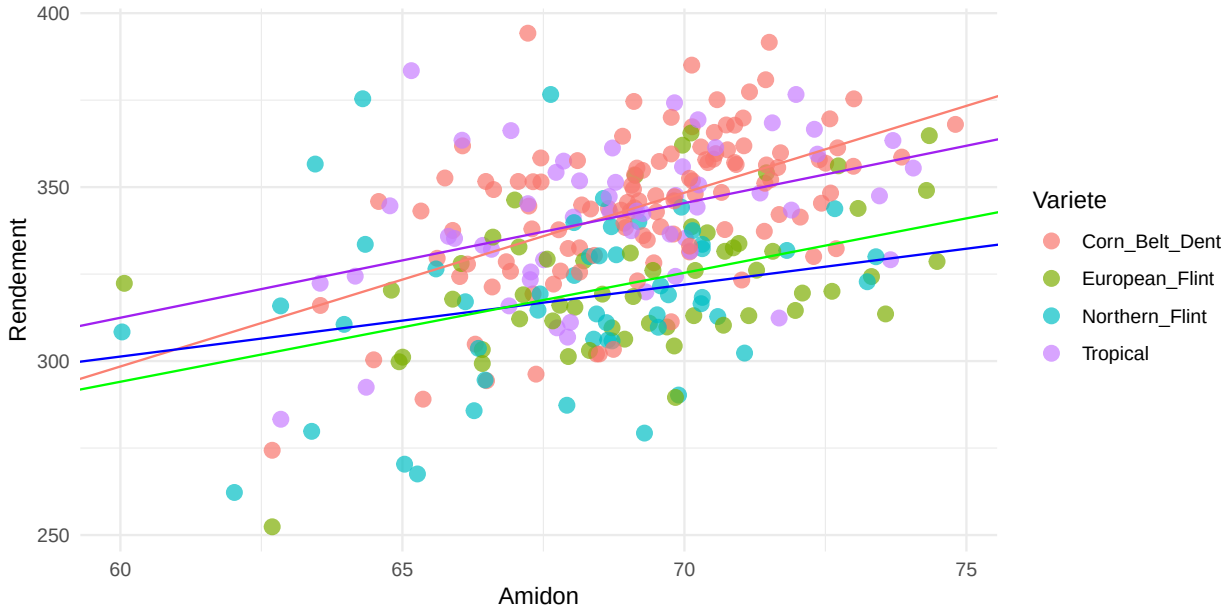
```
summary(reg_variete_marqueur)$adj.r.squared
```

```
## [1] 0.2088599
```

## 7. Variété & Amidon

### 7.1. Vérification préliminaire

Sur le graphique, il semble avoir un niveau différent de rendement par rapport à la teneur en amidon selon la variété de mais et chaque variété semble avoir une pente différente d'où la différence en évolution de rendement par rapport à la teneur en amidon.




### 7.2. Ajustement du modèle

Nous allons alors ajuster le modèle avec les variables *Variete* et *Amidon*.

```
##                   (Intercept)        VarieteEuropean_Flint
##                     -1.178092                   107.368050
##        VarieteNorthern_Flint              VarieteTropical
##                    178.495260                   115.902977
##                       Amidon VarieteEuropean_Flint:Amidon
##                     4.993822                    -1.862676
## VarieteNorthern_Flint:Amidon       VarieteTropical:Amidon
##                    -2.927448                    -1.698230
```

## 7.3. Validité des hypothèses

### 7.3.1. Hypothèses d'espérance nulle et d'homoscédasticité

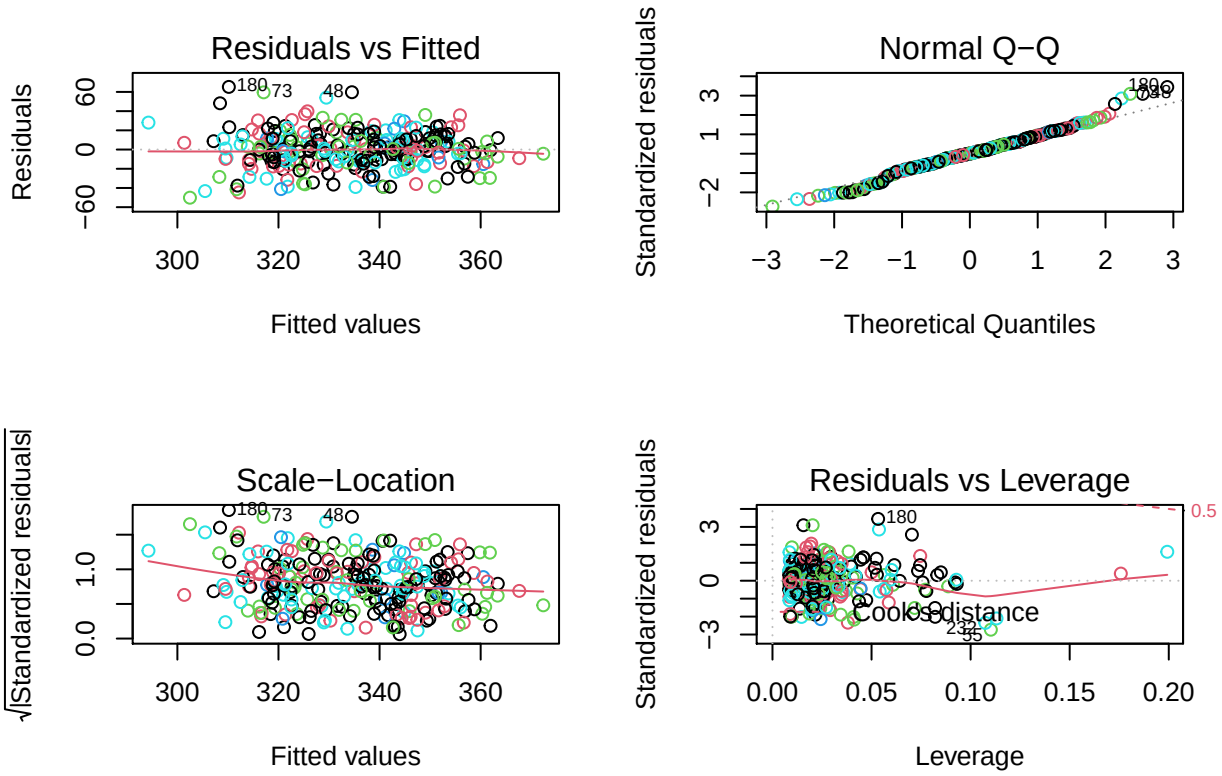
Sur les graphiques *Residuals vs Fitted* et *Scale-Location*, nous avons une homogénéité entre différentes couleurs. Nous n'observons pas de structure particulière avec une bande horizontale. Nous pouvons donc valider les hypothèses d'espérance nulle et d'homoscédasticité.

### 7.3.2. Hypothèse de normalité

Sur le graphique *Normale Q-Q*, nous observons que les points sont bien alignés sur la droite d'où la normalité.

### 7.3.3. Valeurs atypiques

Sur le graphique *Residuals vs Leverage*, nous observons que la distance de Cook des points est inférieure à 1. Nous pouvons alors valider qu'il n'y a pas des observations atypiques problèmatiques.




## 7.4. Intervalles de confiance

Les intervalles de confiance de notre modèles sont suivantes :

```
##                                2.5 %      97.5 %
## (Intercept)                -110.733367 108.3771832
## VarieteEuropean_Flint       -57.724113 272.4602128
```

```
## VarieteNorthern_Flint           8.817155 348.1733640
## VarieteTropical               -59.134166 290.9401188
## Amidon                          3.412947   6.5746976
## VarieteEuropean_Flint:Amidon   -4.245287   0.5199346
## VarieteNorthern_Flint:Amidon   -5.402691  -0.4522044
## VarieteTropical:Amidon         -4.234404   0.8379430
```

## 7.5. Tests statistiques

### 7.5.1. Test du modèle

Nous pouvons effectuer le test de Fisher pour vérifier si l'ensemble de variété et de teneur en amidon a une influence sur le rendement de mais.

Nous avons la p-valeur plus petite que 5%. L'hypothèse nulle est rejetée. Il était pertinent de considérer le modèle ANCOVA avec les variables *Variete* et *Amidon* qu'un modèle constant qui aurait une tendance commune à tous les groupes de modalités.

**Nous pouvons considérer que l'ensemble de variété et de teneur en amidon de mais a une influence sur son rendement**.

```
##
## Call:
## lm(formula = Rendement ~ Variete * Amidon, data = mais_bis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.120 -11.451   0.084  11.296  65.223
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -1.178     55.645  -0.021   0.9831
## VarieteEuropean_Flint          107.368     83.853   1.280   0.2015
## VarieteNorthern_Flint          178.495     86.183   2.071   0.0393 *
## VarieteTropical                115.903     88.904   1.304   0.1935
## Amidon                           4.994      0.803   6.219  1.9e-09 ***
## VarieteEuropean_Flint:Amidon    -1.863      1.210  -1.539   0.1249
## VarieteNorthern_Flint:Amidon    -2.927      1.257  -2.329   0.0206 *
## VarieteTropical:Amidon          -1.698      1.288  -1.318   0.1885
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.46 on 269 degrees of freedom
## Multiple R-squared:  0.3753, Adjusted R-squared:  0.3591
## F-statistic: 23.09 on 7 and 269 DF,  p-value: < 2.2e-16
```

### 7.5.2. Test d'effet des variables explicatives

Nous allons tester l'effet principal de *Variete*, l'effet principal de *Amidon* et l'effet d'interaction de *Variete* et de *Amidon*.

Avec une p-valeur inférieure à 5%, l'ajout de la variable *Variete* dans le modèle est pertinent. L'ajout de variable *Amidon* a également une p-valeur inférieure à 5% donc sa présence est pertinente dans le modèle.

Néanmoins, l'interaction des variables *Variete* et *Amidon* a une p-valeur de 12% donc ce dernier n'a pas de pouvoir explicatif.

Le modèle additif avec seuls les effets principal de *Variete* et *Amidon* serait plus pertinent que le modèle complet qui inclut l'effet d'interactions de ces deux variables.

```
## Anova Table (Type II tests)
##
## Response: Rendement
##                 Sum Sq  Df F value    Pr(>F)
## Variete          30008   3 26.4166 5.239e-15 ***
## Amidon           22737   1 60.0453 1.905e-13 ***
## Variete:Amidon    2219   3  1.9536    0.1213
## Residuals       101859 269
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testons la pertinence du modèle additif par rapport au modèle complet. Avec une p-valeur supérieure 12%, nous observons bien que le modèle additif est plus pertinent que le modèle complet.

```
reg_variete_amidon_add = lm(Rendement ~ Variete + Amidon, mais_bis)
anova(reg_variete_amidon_add, reg_variete_amidon)
```

```
## Analysis of Variance Table
##
## Model 1: Rendement ~ Variete + Amidon
## Model 2: Rendement ~ Variete * Amidon
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    272 104078
## 2    269 101859  3    2219.2 1.9536 0.1213
```

Nous obtenons le même résultat dans le modèle additif où la variété est ajoutée au modèle qui contient déjà la teneur en amidon.

```
reg_amidon_variete_add = lm(Rendement ~ Amidon + Variete, mais_bis)
anova(reg_amidon_variete_add, reg_variete_amidon)
```

```
## Analysis of Variance Table
##
## Model 1: Rendement ~ Amidon + Variete
## Model 2: Rendement ~ Variete * Amidon
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    272 104078
## 2    269 101859  3    2219.2 1.9536 0.1213
```

## 7.6. Prédiction

Pour les nouvelles observations, la prédiction nous donnerait un résultat de régression linéaire entre la teneur en amidon et son rendement pour chaque modalité de variété.

```
##        fit      lwr      upr
## 1 345.0935 306.6185 383.5686
## 2 323.5228 284.8693 362.1763
## 3 336.9037 298.3494 375.4579
```

## 7.7. Critique du modèle

Que ce soit additif ou complet, le coefficient de détermination du modèle est à l'ordre de 0,35. L'ensemble de variété et de teneur en amidon des mais explique 35% de la variabilité de son rendement. Bien que ce soit le modèle qui a un coefficient de détermination le plus élevé parmi tous les modèles testés, sa capacité prédictive n'est toujours pas satisfaisante. Nous pouvons supposer que ce ne soit pas le modèle linéaire qui serait le plus adapté pour expliquer le rendement de mais avec les variables explicatives que nous avons.

```
summary(reg_variete_amidon)$adj.r.squared
```

```
## [1] 0.3590916
```

```
summary(reg_variete_amidon_add)$adj.r.squared
```

```
## [1] 0.352351
```