Projet Regression lineaire 2023

SEFFANE Asmaa

2023-01-24

Introduction

Dans ce projet, je vais expliquer le rendement de plants de maïs. Sur chaque parcelle, le maïs a un même marqueur génétique (1 ou 2) et une même variété. On mesure différentes caractéristiques :

- le rendement de la parcelle,
- la teneur moyenne en huile, en protéine et 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.

Analyse de données

Pour commencer, j'importe les librairies nécessaires pour le traitement des données

1) Chargement des données (Q1)

```
donnee <- read_delim("mais.txt",
                     "\t", escape_double = FALSE,
                     trim_ws = TRUE)

**
```

2) Formatage et analyse descriptive (Q2)

```
attach(donnee)
#definitions des données qualitatives:
donnee$Variete <- as.factor(donnee$Variete)
donnee$Marqueur <- as.factor(donnee$Marqueur)

levels(Variete)

## NULL

levels(Marqueur)

## NULL

#niveau modalité des variables:
table(Variete)

## Variete
## Corn_Belt_Dent European_Flint Northern_Flint    Stiff_Stalk       Tropical
##            117             56             50             11             54

table(Marqueur)

## Marqueur
##   1   2
## 106 182

  #moyenne du rendement par rapport aux variables qualitatives:
RDM_means <- aggregate(list(Rendement = Rendement), list(Variete = Variete,
Marqueur = Marqueur), mean)
  #moyenne du Huile par rapport aux variables qualitatives:
HL_means <- aggregate(list(Huile = Huile), list(Variete = Variete, Marqueur =
Marqueur), mean)
  #moyenne du Proteine par rapport aux variables qualitatives:
PRT_means <- aggregate(list(Proteine = Proteine), list(Variete = Variete,
Marqueur = Marqueur), mean)
```

```r
  #moyenne du Amidon par rapport aux variables qualitatives:
AMD_means <- aggregate(list(Amidon = Amidon), list(Variete = Variete,
Marqueur = Marqueur), mean)
  #moyenne du Floraison par rapport aux variables qualitatives:
FRS_means <- aggregate(list(Floraison = Floraison), list(Variete = Variete,
Marqueur = Marqueur), mean)
  #moyenne du Feuilles par rapport aux variables qualitatives:
FLLE_means <- aggregate(list(Feuilles = Feuilles), list(Variete = Variete,
Marqueur = Marqueur), mean)

#affichage des différentes moyennes
RDM_means
```

```
##           Variete Marqueur Rendement
## 1  Corn_Belt_Dent        1  347.8866
## 2  European_Flint        1  319.9829
## 3  Northern_Flint        1  316.7196
## 4     Stiff_Stalk        1  353.5280
## 5        Tropical        1  342.4842
## 6  Corn_Belt_Dent        2  342.7325
## 7  European_Flint        2  324.4547
## 8  Northern_Flint        2  319.3362
## 9     Stiff_Stalk        2  352.8715
## 10       Tropical        2  341.2252
```

```r
HL_means
```

```
##           Variete Marqueur    Huile
## 1  Corn_Belt_Dent        1 3.516024
## 2  European_Flint        1 3.650809
## 3  Northern_Flint        1 3.848407
## 4     Stiff_Stalk        1 3.687856
## 5        Tropical        1 3.615948
## 6  Corn_Belt_Dent        2 3.317855
## 7  European_Flint        2 3.494534
## 8  Northern_Flint        2 3.959901
## 9     Stiff_Stalk        2 2.965166
## 10       Tropical        2 3.428015
```

```r
PRT_means
```

```
##           Variete Marqueur Proteine
## 1  Corn_Belt_Dent        1 12.72193
## 2  European_Flint        1 12.92312
## 3  Northern_Flint        1 13.71802
## 4     Stiff_Stalk        1 12.14204
## 5        Tropical        1 12.60925
## 6  Corn_Belt_Dent        2 12.97176
## 7  European_Flint        2 13.25200
## 8  Northern_Flint        2 14.02037
```

```
## 9     Stiff_Stalk        2 13.19828
## 10       Tropical        2 13.24608
```

AMD_means

```
##           Variete Marqueur   Amidon
## 1  Corn_Belt_Dent        1 69.05215
## 2  European_Flint        1 69.13650
## 3  Northern_Flint        1 68.60888
## 4     Stiff_Stalk        1 70.34181
## 5        Tropical        1 69.95661
## 6  Corn_Belt_Dent        2 69.39667
## 7  European_Flint        2 69.26752
## 8  Northern_Flint        2 67.00878
## 9     Stiff_Stalk        2 70.05063
## 10       Tropical        2 68.58107
```

FRS_means

```
##           Variete Marqueur Floraison
## 1  Corn_Belt_Dent        1 1063.0124
## 2  European_Flint        1  922.7989
## 3  Northern_Flint        1  913.6430
## 4     Stiff_Stalk        1 1109.5625
## 5        Tropical        1 1159.4975
## 6  Corn_Belt_Dent        2  994.7915
## 7  European_Flint        2  945.9811
## 8  Northern_Flint        2  921.9680
## 9     Stiff_Stalk        2 1025.9500
## 10       Tropical        2 1202.5450
```

FLLE_means

```
##           Variete Marqueur Feuilles
## 1  Corn_Belt_Dent        1 18.41872
## 2  European_Flint        1 14.73859
## 3  Northern_Flint        1 14.57905
## 4     Stiff_Stalk        1 19.68123
## 5        Tropical        1 20.93294
## 6  Corn_Belt_Dent        2 16.78605
## 7  European_Flint        2 15.38878
## 8  Northern_Flint        2 15.14983
## 9     Stiff_Stalk        2 17.59823
## 10       Tropical        2 21.22684
```

```
#Resume des variables quantitatives
summary(Rendement)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   252.4   319.3   337.5   335.5   352.8   394.2
```

```
summary(Huile)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.622   3.058   3.509   3.516   3.922   8.201

summary(Amidon)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   60.03   67.43   69.16   68.98   70.58   74.81

summary(Proteine)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.579  12.156  13.191  13.144  14.110  19.425

summary(Feuilles)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.08   14.76   16.57   17.27   19.07   30.58

summary(Floraison)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   808.2   914.8   982.7  1021.0  1092.8  1646.8
```

Je remarque que les moyennes restent homogènes : pour chaque variable quantitative, les moyennes par variété et marqueur restent du même ordre de grandeur.

Commençant l'analyse prédictive.

3) Première approche : La teneur moyenne en amidon d'un grain de maïs permet-elle de prédire le rendement d'une parcelle ? (Q3)

En premier temps, on étudie l'effet de la teneur moyenne de l'amidon sur le rendement, donc on a:

- Variable d'intérêt: le rendement
- Variable explicative: moyenne d'amidon (var. quantitative)

on calcule les moyennes des variables puis le coefficient de corrélation linéaires:

```
x_bar <- mean(Amidon)
y_bar <- mean(Rendement)
rho_xy <- cor(Rendement, Amidon)

x_bar

## [1] 68.9755
```

```
y_bar
```

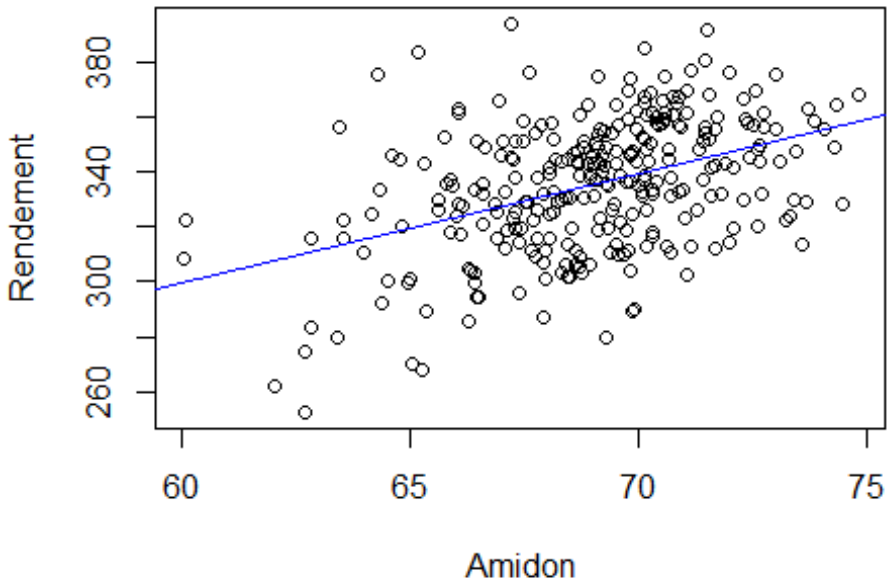
```
## [1] 335.5031
```

```
rho_xy
```

```
## [1] 0.4258431
```

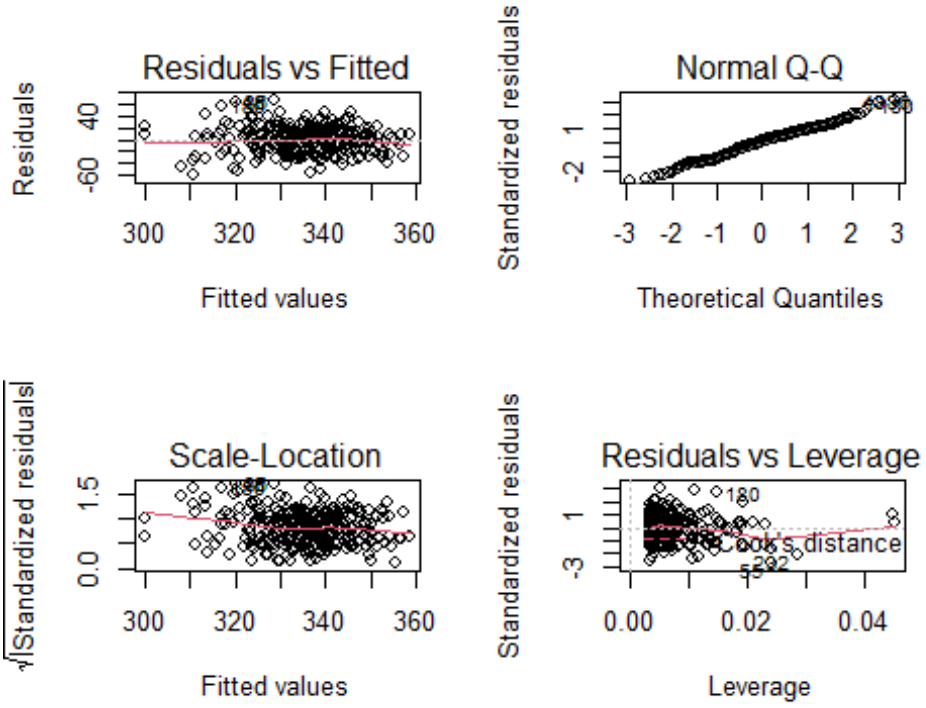
Le coefficient de corrélation ρ(x, y) = 0.425 est positif. ça me permet d'essayer un modèle de régression linéaire simple:

```
reg <- lm(Rendement ~ Amidon)
plot (x = Amidon , y = Rendement)
abline(reg, col = 'blue')
```



Je constate une tendance linéaire confirmée par la droite de régression (en bleu)

Je passe à la validation des quatre hypothèses de notre modèle:



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

- L'hypothèse 1: ne peut être assuré que par le protocole expérimental.
- L'hypothèse 2: vérifiée grâce au graphique en haut à gauche (espérance nulle)
- L'hypothèse 3: vérifiée grâce au graphique en bas à gauche (même variance)
- L'hypothèse 4: vérifiée grâce au graphique en haut à droite (loi normale respectée)

Grace au graphique en bas à droite, je constate qu'il n'y a pas de points aberrants, je garde alors tout mon échantillon

Les hypothèses de notre modèle sont bien vérifiées.

Regardant l'intervalle de confiance

```
confint(reg, level = .95)
```

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

Regardant maintenant avec ce modèle si la moyenne d'amidon n'affecte pas le rendement (Hypothèse H0)

```r
alpha <- 0.05
dim_p <- length(coef(reg))
n <- nrow(donnee)
S2 <- summary(reg)$sigma^2
(n - dim_p) * S2/(qchisq(c(1 - alpha/2, alpha/2), n - dim_p))
```

```
## [1] 409.1220 568.1457
```

```r
summary(reg)
```

```
##
## 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
```

J'ai la P-valeur égale à 4.085e-14 < 0.05, donc la moyenne d'Amidon a un effet sur le rendement (H0 rejetée).

Passant aux prédictions:

```r
x1_new <- data.frame("Marqueur" = 1, "Variete" ='Corn_Belt_Dent', "Huile" =
3.39, "Proteine" = 13, "Amidon" = 69.34,                          "Floraison"
= 1000,"Feuilles" = 17)
predict(reg, newdata = x1_new, interval = "confidence", level = .95)
```

```
##        fit      lwr      upr
## 1 336.9505 334.3874 339.5136
```

```r
predict(reg, newdata = x1_new, interval = "prediction", level = .95)
```

```
##        fit      lwr      upr
## 1 336.9505 293.8031 380.0979
```

```r
x2_new <- data.frame("Marqueur" = 1, "Variete" = 'European_Flint' ,  "Huile"
= 3.54,
                     "Proteine" = 13.3, "Amidon" = 69.41, "Floraison" = 943,
                     "Feuilles" = 15)
predict(reg, newdata = x2_new, interval = "confidence", level = .95)
```

```
##        fit      lwr      upr
## 1 337.2285 334.6549 339.8021

predict(reg, newdata = x2_new, interval = "prediction", level = .95)

##        fit      lwr      upr
## 1 337.2285 294.0804 380.3765

x3_new <- data.frame("Marqueur" = 2, "Variete" = 'Corn_Belt' ,  "Huile" =
2.85,
                     "Proteine" = 11.8, "Amidon" = 67.7, "Floraison" = 934,
                     "Feuilles" = 16)
predict(reg, newdata = x3_new, interval = "confidence", level = .95)

##        fit      lwr      upr
## 1 330.4381 327.6079 333.2683

predict(reg, newdata = x3_new, interval = "prediction", level = .95)

##        fit     lwr      upr
## 1 330.4381 287.274 373.6022
```

On remarque qu'il y a une augmentation de rendement quand la moyenne d'amidon augmente, on passe d'un rendement de 330.4 avec Amidon à 67.7 au rendement de 337.2 avec Amidon à 69.4

Maintenant, je regarde l'efficacité de ce premier modèle à prédire. Voyant pour cela le coefficient de détermination ajusté :

```
summary(reg)$r.squared

## [1] 0.1813424
```

Ce coefficient est très faible, mon modèle ne permet pas de bien prédire le rendement en se basant uniquement sur l'Amidon.

4) Seconde approche, changeant de modèle :

Le rendement d'une parcelle peut-il être prédit à l'aide de la teneur en amidon, en huile, en protéine d'un grain de maïs ainsi que du nombre de degrés jours avant floraison, et du nombre de feuilles par plant de maïs ? (Q4)

J'étudie l'effet de la teneur moyenne des covariables citées ci-dessus sur le rendement, donc on a:

- Variable réponse: le rendement
- Variable explicative: moyenne d'amidon, huile, protéine, floraison et feuilles (variable quantitatives)

Je calcule le coefficient de corrélation linéaires:

```
rho_xy <- cor(Rendement, Amidon + Huile + Proteine + Feuilles + Floraison)
rho_xy

## [1] 0.4112425
```

Le coefficient de corrélation ρ(x, y) = 0.411 est positif. ça me permet d'essayer un modèle de régression linéaire multiple:

```
reg_1 <- lm(Rendement ~ Amidon + Huile + Proteine + Feuilles + Floraison)
```

J'ai une régression multiple.

Pour savoir s'il y a un problème de colinéarité entre les différentes variables explicatives:

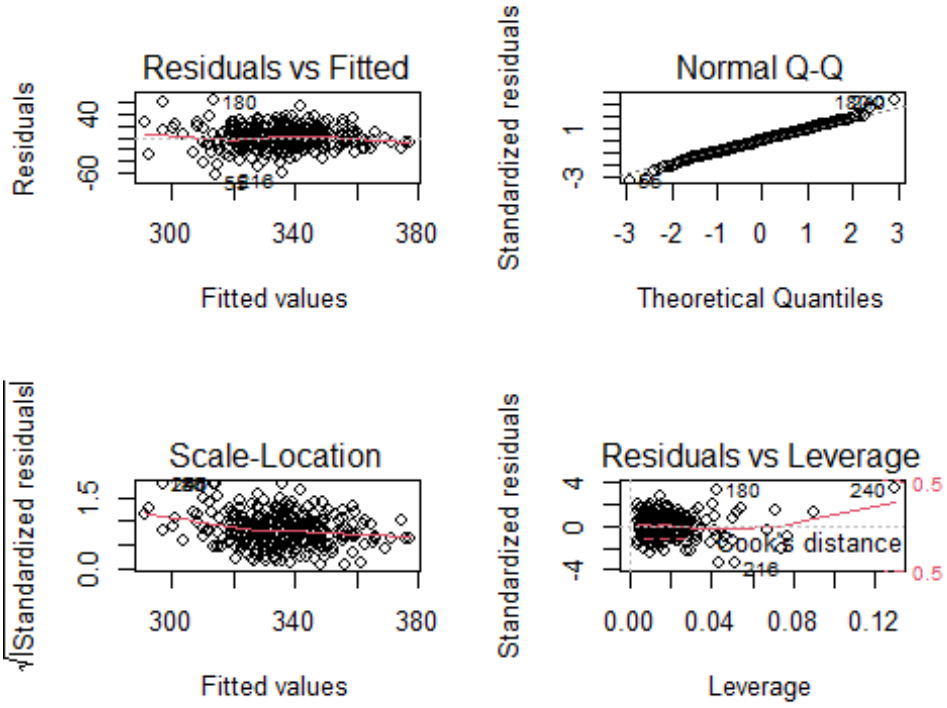
```
vif(reg_1)

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

Je remarque que les deux variables Feuilles et Florisation ont une valeur plus grande que 5, J'élimine du modèle celle qui a le plus grand coefficient. Puis je regénère mon modèle:

```
reg_2 <-  lm(Rendement ~ Amidon + Huile + Proteine + Feuilles)
vif(reg_2)

##   Amidon    Huile Proteine Feuilles
## 3.139589 1.356563 2.756242 1.086766
```

On peut continuer maintenant avec le nouveau modèle.

Je commence par vérifier les quatre hypothèses de mon modèle:



- L'hypothèse 1: ne peut être assuré que par le protocole expérimental.
- L'hypothèse 2: vérifiée grâce au graphique en haut à gauche (espérance nulle)
- L'hypothèse 3: vérifiée grâce au graphique en bas à gauche (même variance)
- L'hypothèse 4: vérifiée grâce au graphique en haut à droite (loi normale respectée)

Grace au graphique en bas à droite, je constate qu'il n'y a pas de points aberrants, je garde alors tout mon échantillon

Les hypothèses de mon modèles sont respectées, je continue.

Regardant l'intervalle de confiance pour les différentes covariables

```
confint(reg_2, level = .95)
```

```
##                   2.5 %      97.5 %
## (Intercept) -69.648453 211.3727189
## Amidon        1.844012   4.8901229
## Huile        -6.927927  -0.5057686
## Proteine     -2.767704   1.9915703
## Feuilles      2.260708   3.5958577
```

Regardant "l'efficacité" de notre modèle

```
alpha <- 0.05
dim_p <- length(coef(reg_2))
```

```r
n <- nrow(donnee)
S2 <- summary(reg_2)$sigma^2
(n - dim_p) * S2/(qchisq(c(1 - alpha/2, alpha/2), n - dim_p))
```

```
## [1] 313.1937 435.6888
```

```r
summary(reg_2)
```

```
##
## Call:
## lm(formula = Rendement ~ Amidon + Huile + Proteine + Feuilles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.773 -11.628   0.151  12.295  61.985
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  70.8621    71.3838   0.993   0.3217
## Amidon        3.3671     0.7738   4.352 1.89e-05 ***
## Huile        -3.7168     1.6313  -2.278   0.0234 *
## Proteine     -0.3881     1.2089  -0.321   0.7484
## Feuilles      2.9283     0.3391   8.634 4.38e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.15 on 283 degrees of freedom
## Multiple R-squared:  0.3794, Adjusted R-squared:  0.3706
## F-statistic: 43.25 on 4 and 283 DF,  p-value: < 2.2e-16
```

J'ai la P-valeur égale à 2.2e-16 < 0.05, donc mon modèle linéaire se basant sur les covariables Amidon, Huile, Proteine et Feuilles permet d'expliquer l'effet sur le rendement.

Passant aux prédictions:

```r
predict(reg_2, newdata = x1_new, interval = "confidence", level = .95)
```

```
##        fit     lwr      upr
## 1 336.4704 334.204 338.7369
```

```r
predict(reg_2, newdata = x2_new, interval = "confidence", level = .95)
```

```
##        fit      lwr     upr
## 1 330.1756 327.4452 332.906
```

```r
predict(reg_2, newdata = x3_new, interval = "confidence", level = .95)
```

```
##        fit     lwr      upr
## 1 330.4929 323.708 337.2778
```

```r
# prediction
predict(reg_2, newdata = x1_new, interval = "prediction", level = .95)
```

```
##        fit      lwr      upr
## 1 336.4704 298.7007 374.2401
```

```
predict(reg_2, newdata = x2_new, interval = "prediction", level = .95)
```

```
##        fit      lwr     upr
## 1 330.1756 292.3752 367.976
```

```
predict(reg_2, newdata = x3_new, interval = "prediction", level = .95)
```

```
##        fit      lwr      upr
## 1 330.4929 292.1856 368.8002
```

On remarque qu'il y a une augmentation de rendement quand les variables Amidon et Feuilles augmentent, ou Huile diminue. Floraison n'a pas d'impact significatif sur le rendement avec ce modèle. on passe d'un rendement de 330.5 au rendement de 336.5

Maintenant, je regarde l'efficacité de ce nouveau modèle à prédire. Voyant pour cela le coefficient de détermination ajusté :

```
summary(reg_2)$adj.r.squared
```

```
## [1] 0.3706023
```

Ce coefficient reste faible, Mon modèle ne permet pas de bien prédire le rendement en se basant sur ces covariables uniquement.

Entre le modèle simple d'une variable et le modèle multiple avec plusieurs variables, j'utilise la fonction anova pour savoir parmi ces deux modèles, lequel est le plus pertinent:

```
anova(reg, reg_2)
```

```
## Analysis of Variance Table
##
## Model 1: Rendement ~ Amidon
## Model 2: Rendement ~ Amidon + Huile + Proteine + Feuilles
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1    286 136950
## 2    283 103822  3     33128 30.1 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La p-valeur est inférieure à 5%, on conserve alors le second modèle qui inclut plusieurs variables explicatives.

L'efficacité de prédiction n'étant toujours pas au rendez-vous,

5) Changeant encore le modèle :

La variété du plan de maïs a-t-elle une influence sur le rendement de l'espèce ?(Q5)

On a donc :

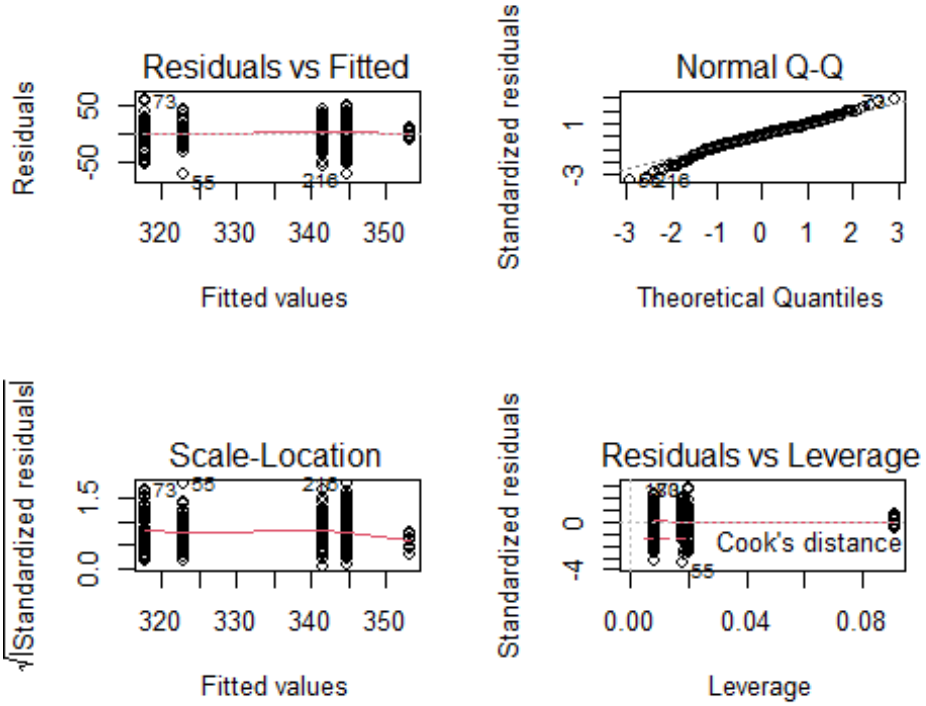
- Variable réponse: le rendement
- Variable explicative: La variété (var. qualitative)

En premier temps, je calcule la moyenne de rendement par rapport aux variétés puis je définie mon modèle :

```
table(Variete)

## Variete
## Corn_Belt_Dent European_Flint Northern_Flint    Stiff_Stalk       Tropical
##            117             56             50             11             54

aggregate(
list(Rendement = Rendement),
list(Variete = Variete),
mean
)

##          Variete Rendement
## 1 Corn_Belt_Dent  344.7148
## 2 European_Flint  322.9375
## 3 Northern_Flint  317.7662
## 4    Stiff_Stalk  353.1102
## 5       Tropical  341.4117
```

On remarque qu'il y a une différence de rendement par variété. Regardant si cela est significatif. Pour cela j'utilise un nouveau modèle linéaire (reg_3).

On commence par les hypothèses de notre modèle si elle sont respectées:



- L'hypothèse 1: ne peut être assuré que par le protocole expérimental.
- L'hypothèse 2: vérifiée grâce au graphique en haut à gauche (espérance nulle)
- L'hypothèse 3: vérifiée grâce au graphique en bas à gauche (même variance)
- L'hypothèse 4: vérifiée grâce au graphique en haut à droite (loi normale respectée)

Grace au graphique en bas à droite, je constate qu'il n'y a pas de points aberrants, je garde alors tout mon échantillon

Les hypothèses sont bien respectées.

Regardant l'intervalle de confiance pour les covariables de notre modèle:

```
confint(reg_3, level = .95)
```

```
##                              2.5 %     97.5 %
## (Intercept)            340.852395 348.577270
## VarieteEuropean_Flint  -28.566116 -14.988595
## VarieteNorthern_Flint  -34.007447 -19.889723
## VarieteStiff_Stalk      -4.780184  21.570997
## VarieteTropical        -10.176390   3.570126
```

je remarque que pour l'intervalle de confiance de "Steff_Stalk" et "Tropical" incluent le zéro

```r
alpha <- 0.05
dim_p <- length(coef(reg_3))
n <- nrow(donnee)
S2 <- summary(reg_3)$sigma^2
(n - dim_p) * S2/(qchisq(c(1 - alpha/2, alpha/2), n - dim_p))
```

```
## [1] 384.5931 535.0136
```

```r
summary(reg_3)
```

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

J'ai la P-valeur égale à 7.005e-16 < 0.05, donc la variété a un effet sur le rendement.

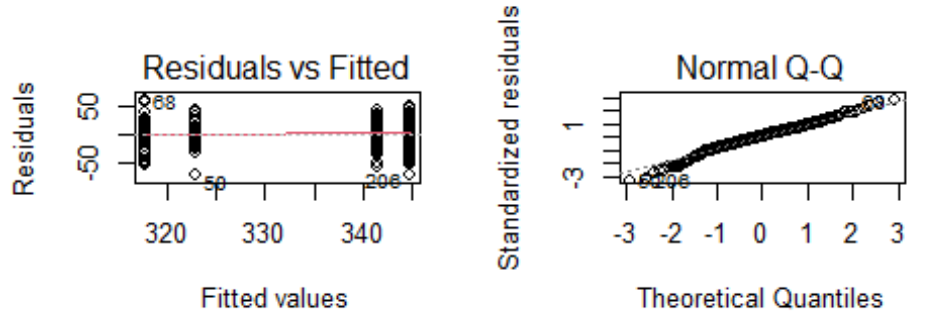
Je remarque aussi que l'erreur pour "Stiff_Stalk" est plus forte par rapport aux autres variétés. voyons les coefficients de chacune des variétés:

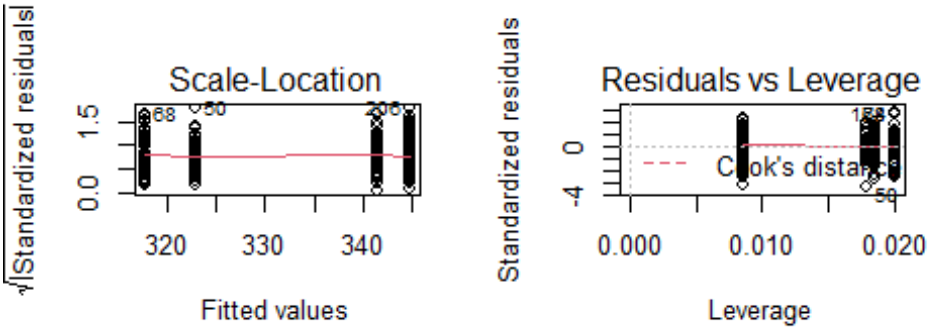
```r
reg_3$coefficients
```

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

J'enlève cette variété "Stiff_Stalk" et je regarde si ça "améliore" notre modèle:

```r
donnee_1 <- subset(donnee, donnee$Variete != 'Stiff_Stalk')

reg_4 <- lm(Rendement ~ Variete, data = donnee_1)

par(mfrow = c(2, 2))
plot(reg_4)
```





- L'hypothèse 1: ne peut être assuré que par le protocole expérimental.
- L'hypothèse 2: vérifiée grâce au graphique en haut à gauche (espérance nulle)
- L'hypothèse 3: vérifiée grâce au graphique en bas à gauche (même variance)
- L'hypothèse 4: vérifiée grâce au graphique en haut à droite (loi normale respectée)

Grace au graphique en bas à droite, je constate qu'il n'y a pas de points aberrants, je garde alors tout mon échantillon

Les hypothèses sont bien respectées.

Regardant l'intervalle de confiance

```
confint(reg_4)
```

```
##                              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
```

```
alpha <- 0.05
dim_p <- length(coef(reg_4))
n <- nrow(donnee_1)
S2 <- summary(reg_4)$sigma^2
(n - dim_p) * S2/(qchisq(c(1 - alpha/2, alpha/2), n - dim_p))
```

```
## [1] 395.4771 553.4726
```

```
summary(reg_4)
```

```
##
## Call:
## lm(formula = Rendement ~ Variete, data = donnee_1)
##
## 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
```

J'ai la P-valeur égale à 7.864e-15 < 0.05, donc la variété a un effet sur le rendement.

Passant aux predictions:

```
predict(reg_4, newdata = x1_new, interval = "confidence", level = .95)
```

```
##        fit      lwr      upr
## 1 344.7148 340.7921 348.6376
```

```
predict(reg_4, newdata = x2_new, interval = "confidence", level = .95)
```

```
##        fit      lwr      upr
## 1 322.9375 317.2674 328.6075
```

```
predict(reg_4, newdata = x1_new, interval = "prediction", level = .95)
```

```
##        fit      lwr      upr
## 1 344.7148 302.1032 387.3265
```

```
predict(reg_4, newdata = x2_new, interval = "prediction", level = .95)
```

```
##        fit      lwr      upr
## 1 322.9375 280.1296 365.7454
```

```
# cette prédiction n'est pas possible car elle contient une variété non
utilisé dans notre modèle
# predict(reg_5, newdata = x3_new, interval = "prediction", level = .95)
```

du fait du coefficient négatif de la variété 'European_Flint', le rendement passe de 344.7 à 322.9

NB : La prédiction pour l'échantillon 3 (x3_new) n'est pas possible car elle contient une variété non utilisé dans notre modèle.

Maintenant, je regarde l'efficacité de ce modèle à prédire.

Voyant pour cela le coefficient de détermination ajusté :

```
summary(reg_4)$adj.r.squared
```

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

Ce coefficient reste encore faible,

Ce nouveau modèle ne permet pas de bien prédire le rendement en se basant uniquement sur la variété.

Regardant la comparaison 2 à 2 des moyennes par modalité:

```
pairwise.t.test(donnee_1$Rendement, donnee_1$Variete, "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  donnee_1$Rendement and donnee_1$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
```

Je remarque qu'il n'y a pas de différence d'effet entre "Corn_Belt_Dent" et "Tropical", et entre "European_Flint" et "Northen_Flint". puisque leur P_valeur est supérieur à 0.05

6) Une autre approche pour la construction du modèle :

Le rendement d'une espèce peut-il être expliqué par sa variété et son marqueur génétique?(Q6)

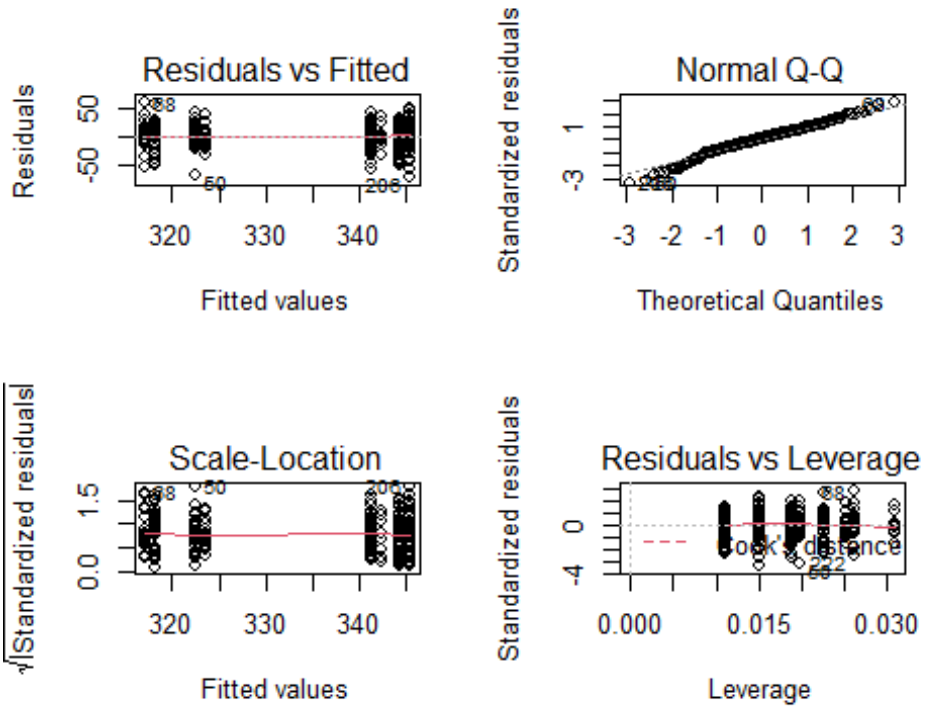
donc on a:

- Variable réponse: le rendement
- Variable explicative: variété et marqueur génétique(var. qualitative)

Je construits le modèle puis je valide ses quatre hypothèses:

```
reg_5 <- lm(Rendement ~Variete + Marqueur, data = donnee_1)
```

```
par(mfrow = c(2, 2))
plot(reg_5)
```



Grace aux graphiques ci-dessus, je constate qu'il n'y a pas de points aberrants, et que les hypothèses de mon modèle sont bien respectées.

Regardant l'intervalle de confiance

```
confint(reg_5)
```

```
##                              2.5 %     97.5 %
## (Intercept)            340.181517 350.577086
## VarieteEuropean_Flint  -28.638579 -14.818241
## VarieteNorthern_Flint  -34.459635 -19.902664
## VarieteTropical        -10.160610   4.065003
## Marqueur2               -6.609662   4.450139
```

```
alpha <- 0.05
dim_p <- length(coef(reg_5))
n <- nrow(donnee_1)
S2 <- summary(reg_5)$sigma^2
(n - dim_p) * S2/(qchisq(c(1 - alpha/2, alpha/2), n - dim_p))
```

```
## [1] 396.6031 555.3924
```

```
summary(reg_5)
```

```
##
## Call:
## lm(formula = Rendement ~ Variete + Marqueur, data = donnee_1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -71.01 -12.00   1.15  13.25  59.51
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            345.379      2.640 130.817  < 2e-16 ***
## VarieteEuropean_Flint  -21.728      3.510  -6.190  2.2e-09 ***
## VarieteNorthern_Flint  -27.181      3.697  -7.352  2.3e-12 ***
## VarieteTropical         -3.048      3.613  -0.844    0.400
## Marqueur2               -1.080      2.809  -0.384    0.701
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.59 on 272 degrees of freedom
## Multiple R-squared:  0.2227, Adjusted R-squared:  0.2113
## F-statistic: 19.49 on 4 and 272 DF,  p-value: 4.106e-14
```

J'ai la P-valeur égale à 4.106e-14 < 0.05, donc la variété et le marqueur génétique ont un effet sur le rendement.

Maintenant, je regarde l'efficacité de ce modèle à prédire.

Voyant pour cela le coefficient de détermination ajusté :

```
summary(reg_5)$adj.r.squared
```

```
## [1] 0.211296
```

Ce coefficient est encore faible, Mon modèle ne permet pas de bien prédire le rendement en se basant uniquement sur la variété et le marqueur génétique.

J'applique la fonction anova pour savoir quel modèle parmi le modèle simple (reg_4) ou le modèle multiple(reg_5) est plus pertinent:

```
anova(reg_4, reg_5)
```

```
## Analysis of Variance Table
##
## Model 1: Rendement ~ Variete
## Model 2: Rendement ~ Variete + Marqueur
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    273 126814
## 2    272 126746  1    68.858 0.1478  0.701
```

La p-valeur est 0.701 > 0.05, donc on conserve le modèle réduit c'est à dire le modèle simple "reg_4"

7) Essayons encore une autre approche:

Le rendement d'une espèce peut-il être expliqué par sa variété et sa teneur en amidon ?(Q7)

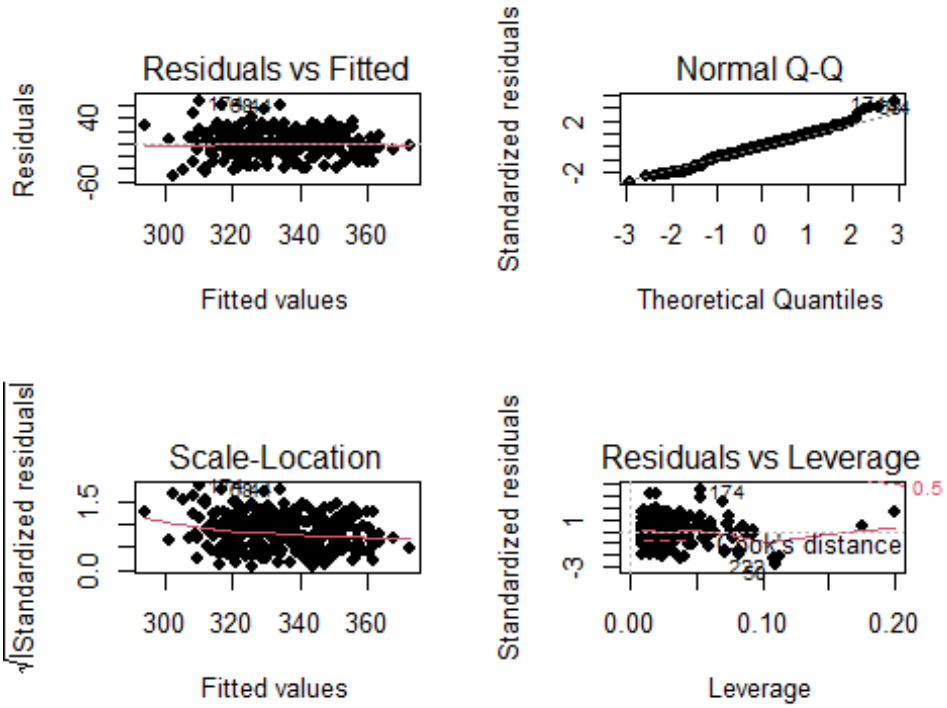
donc on a:

- Variable réponse: le rendement
- Variable explicative: moyenne d'amidon et variété (var. quantitative et qualitative)

Construisant ce nouveau modèle

```
reg_7 <- lm(Rendement ~ Variete * Amidon, data = donnee_1)
```

Validation des hypothèses du modèle :



Grace aux graphiques ci-dessus, je constate qu'il n'y a pas de points aberrants, et que les hypothèses de mon modèle sont bien respectées.

Regardant l'intervalle de confiance

```
confint(reg_7)
```

```
##                                  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
```

```
alpha <- 0.05
dim_p <- length(coef(reg_7))
n <- nrow(donnee_1)
S2 <- summary(reg_7)$sigma^2
(n - dim_p) * S2/(qchisq(c(1 - alpha/2, alpha/2), n - dim_p))
```

```
## [1] 322.0066 451.7777
```

```
summary(reg_7)
```

```
##
## Call:
## lm(formula = Rendement ~ Variete * Amidon, data = donnee_1)
##
## 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
```

J'ai la P-valeur égale à 2.2e-16 < 0.05, donc la moyenne d'Amidon et la variété ont un effet sur le rendement.

Passant aux predictions:

```
predict(reg_7, newdata = x1_new, interval = "prediction", level = .95)
```

```
##        fit      lwr      upr
## 1 345.0935 306.6185 383.5686
```

```
predict(reg_7, newdata = x2_new, interval = "prediction", level = .95)
```

```
##        fit      lwr      upr
## 1 323.5228 284.8693 362.1763
```

```
predict(reg_7, newdata = x1_new, interval = "confidence", level = .95)
```

```
##        fit      lwr      upr
## 1 345.0935 341.5496 348.6375
```

```
predict(reg_7, newdata = x2_new, interval = "confidence", level = .95)
```

```
##        fit      lwr      upr
## 1 323.5228 318.3924 328.6532
```

je remarque que les variables ont un effet sur le rendement en changeant de variété et Amidon.

Maintenant, je regarde l'efficacité de ce modèle à prédire. Voyant pour cela le coefficient de détermination ajusté :

```
summary(reg_7)$adj.r.squared
```

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

Ce coefficient reste un peu faible, Ce nouveau modèle ne permet pas de bien prédire le rendement en se basant uniquement sur l'Amidon et les variétés.

on fait un anova entre le modèle (reg_5) et le modèle (reg_7) pour savoir qui est le plus pertinent:

```
anova(reg_5, reg_7)
```

```
## Analysis of Variance Table
##
## Model 1: Rendement ~ Variete + Marqueur
## Model 2: Rendement ~ Variete * Amidon
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    272 126746
## 2    269 101859  3     24887 21.908 1.005e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-valeur 1.257e-10 < 0.05, on conserve le modèle reg_7 (Rendement ~ Variete * Amidon)

Maintenant j'applique l'ancova sur le modèle dernier:

```
anova(reg_7)
```

```
## Analysis of Variance Table
##
## Response: Rendement
##                 Df Sum Sq Mean Sq F value    Pr(>F)
## Variete          3  36250 12083.3 31.9109 < 2.2e-16 ***
## Amidon           1  22737 22736.5 60.0453 1.905e-13 ***
```

```
## Variete:Amidon   3   2219   739.7  1.9536    0.1213
## Residuals      269 101859   378.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Le modèle contenant l'influence conjointe de la variété et de la moyenne d'Amidon est pertinent, Il n'y a pas d'effet d'interaction car la P_valeur 0.1213 > 0.05

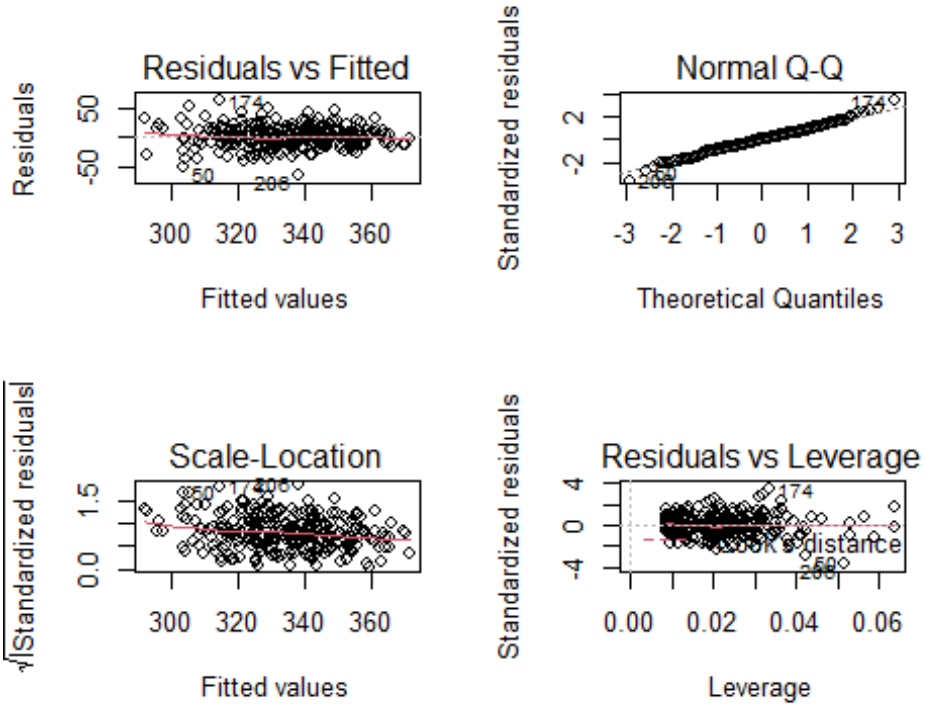
#Constat:

Avec tous ces modèles, le R square reste faible et ne dépassant pas 0.4. Tous ces modèle ne permettent pas une bonne prédiction.

8) Extra :

Intuitivement, je pense à essayer un modèle contenant uniquement les variables 'favorables' déduites des précédents modèles Je considère le modèle linéaire suivant: Rendement ~ Variete + Amidon + Feuilles avec le jeu de données donnee_1

```
reg_T <- lm( Rendement ~ Variete + Amidon  + Feuilles, data = donnee_1)

par(mfrow = c(2, 2))
plot(reg_T)
```



Les hypothèses sont toutes vérifiées, je passe aux intervalles de confiances:

```
confint(reg_T)
```

```
##                             2.5 %     97.5 %
## (Intercept)           -15.230965 105.630932
## VarieteEuropean_Flint -21.751020  -9.575586
## VarieteNorthern_Flint -21.793433  -8.786908
## VarieteTropical       -18.172172  -4.947501
## Amidon                  2.817710   4.495511
## Feuilles                1.862015   3.448977
```

```
alpha <- 0.05
dim_p <- length(coef(reg_T))
n <- nrow(donnee_1)
S2 <- summary(reg_T)$sigma^2
(n - dim_p) * S2/(qchisq(c(1 - alpha/2, alpha/2), n - dim_p))
```

```
## [1] 281.6624 394.6783
```

```
summary(reg_T)
```

```
##
## Call:
## lm(formula = Rendement ~ Variete + Amidon + Feuilles, data = donnee_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.957 -11.690   0.525  11.177  61.244
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            45.2000    30.6950   1.473 0.142032
## VarieteEuropean_Flint -15.6633     3.0922  -5.065 7.54e-07 ***
## VarieteNorthern_Flint -15.2902     3.3032  -4.629 5.71e-06 ***
## VarieteTropical       -11.5598     3.3586  -3.442 0.000669 ***
## Amidon                  3.6566     0.4261   8.581 7.39e-16 ***
## Feuilles                2.6555     0.4030   6.589 2.30e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.19 on 271 degrees of freedom
## Multiple R-squared:  0.4499, Adjusted R-squared:  0.4397
## F-statistic: 44.32 on 5 and 271 DF,  p-value: < 2.2e-16
```

J'ai la P-valeur égale à 2.2e-16 < 0.05,

Je regarde de R square

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

```
## [1] 0.439713
```

Ce coefficient est le plus grands de tous les coefficients précédents:

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

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

```
summary(reg_2)$adj.r.squared
```

```
## [1] 0.3706023
```

```
summary(reg_3)$adj.r.squared
```

```
## [1] 0.2271173
```

```
summary(reg_4)$adj.r.squared
```

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

```
summary(reg_5)$adj.r.squared
```

```
## [1] 0.211296
```

```
summary(reg_7)$adj.r.squared
```

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

La dernière étape est de faire la prédiction:

```
predict(reg_T, newdata = x1_new, interval = "confidence", level = .95)
```

```
##        fit      lwr      upr
## 1 343.8928 340.5647 347.2209
```

```
predict(reg_T, newdata = x2_new, interval = "confidence", level = .95)
```

```
##        fit      lwr      upr
## 1 323.1744 318.3836 327.9653
```

```
predict(reg_T, newdata = x1_new, interval = "prediction", level = .95)
```

```
##        fit      lwr      upr
## 1 343.8928 307.9188 379.8667
```

```
predict(reg_T, newdata = x2_new, interval = "prediction", level = .95)
```

```
##        fit      lwr      upr
## 1 323.1744 287.0358 359.3131
```

Conclusion

En regardant l'efficacité de ce dernier modèle, je constate qu'il est plus efficace que les précédents, en matière de prédiction, je préférerai utiliser celui-là plus que les autres.

Mais surement, il y a des méthodes plus que intuitives qui permettent de trouver un meilleur modèle : Lasso, Ridge, ...