Rapport sur l'étude statistique des Vinhos Verdes

Thibault DUGAUQUIER

Introduction

Le jeu de données sur lequel nous avons basé notre étude statistique provient du site UC Irvine Machine Learning Repository, qui recense plus de 480 jeux de données mis à contribution à des fins d'apprentissage pour les intelligences artificielles. Ces données, récoltées en 2009, concernent différentes variantes du "Vinho Verde", vin produit dans le Minho, région du nord-ouest du Portugal.

Les variables traitées dans ce jeu de données se rapportent à des paramètres physico-chimiques (taux d'alcool, quantité d'acide citrique,...) et à la qualité gustative de chaque vin. À partir de celles-ci, nous allons effectuer quatre tests statistiques et nous allons mettre en place deux modèles de régression.

I/ Présentation des données

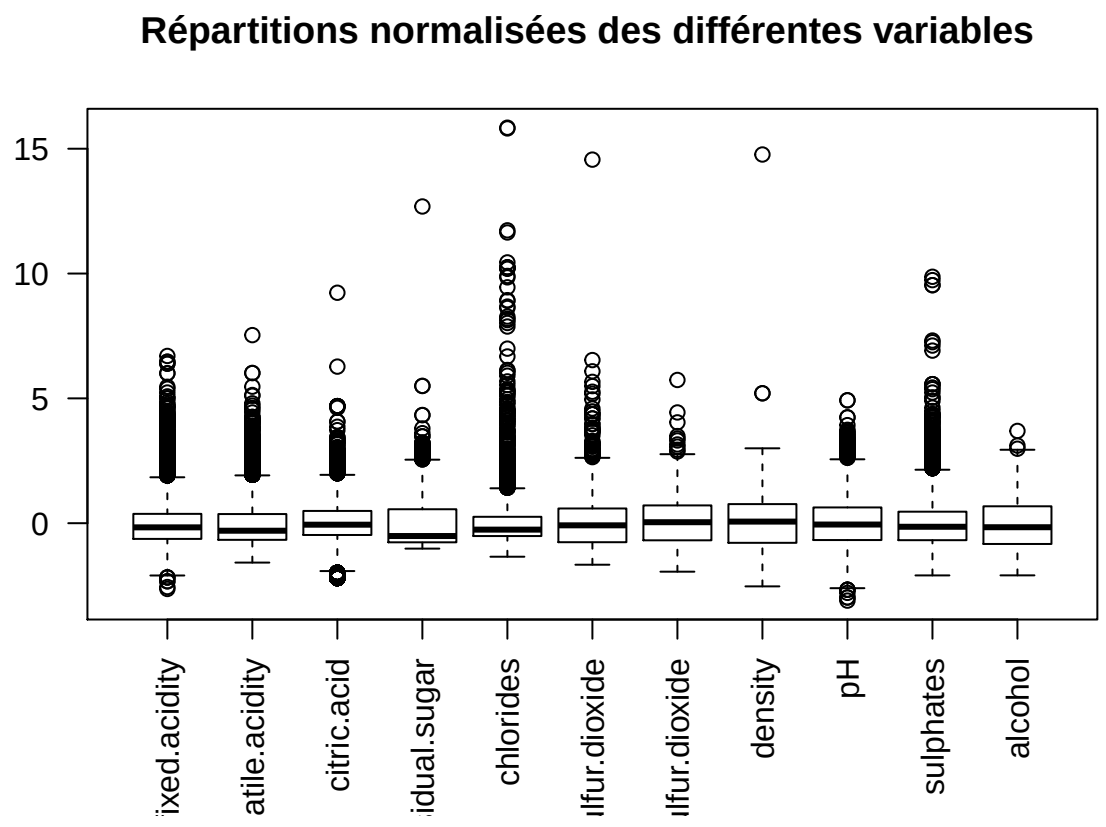
Le jeu de données étudié ici est constitué de deux fichiers, "winequality-white.csv", qui contient des données de 4898 Vinhos Verdes blancs et "winequality-red.csv", qui contient les mêmes données pour 1599 vins rouges. La population analysée est les différentes variantes de Vinhos Verdes et l'unité statistique représente un vin. Les variables prises en compte dans les deux fichiers sont :

- **"Fixed acidity"** (variable quantitative continue) : correspond au total d'acides présent dans un vin (acide tartrique, acide lactique, acide citrique, acide malique).
- **"Volatile acidity"** (variable quantitative continue) : correspond à la quantité d'acides distillables en vapeur présente dans un vin (acide acétique principalement).
- **"Acide citric"** (variable quantitative continue) : quantité d'acide citrique présente dans un vin.
- **"Residual sugar"** (variable quantitative continue) : quantité de sucre naturellement présente dans le raisin après la fermentation alcoolique.
- **"Chlorides"** (variable quantitative continue) : matière minérale contenue naturellement dans le vin, comme le sel (chlorure de sodium).
- **"Free sulfur dioxide"** (variable quantitative continue) : dioxyde de soufre présent sous forme de molécule libre (c'est à dire non liée à d'autres molécules) dans le vin ; propriétés antiseptiques et antioxydantes.
- **"Total sulfur dioxide"** (variable quantitative continue) : tient compte des molécules de dioxyde de soufre libres mais aussi de celles ayant des liaisons avec d'autres éléments chimiques.
- **"Density"** (variable quantitative continue) : densité d'un vin.
- **"pH"** (variable quantitative continue) : mesure de l'acidité d'un vin, allant de 1 à 14.
- **"Sulphates"** (variable quantitative continue) : similaire au dioxyde de soufre, propriétés antiseptiques et antioxydantes pour le vin.
- **"Alcohol"** (variable quantitative continue) : degré d'alcool présent dans un vin.
- **"Quality"** (variable qualitative ordinale) : degré d'appréciation d'un vin par des experts, allant de 0 (très mauvais) à 10 (excellent).

Afin de réaliser une étude statistique approndie prenant en compte la couleur des vins, nous avons décidé de regrouper les deux fichiers en une grande matrice intitulée **"DataVins"**, qui contient les 6497 vins. Nous avons ajouté une colonne **"Type"** à chaque vin, qui contient la couleur associée au vin en question (variable qualitative nominale).

fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide
Min. : 3.800	Min. :0.0800	Min. :0.0000	Min. : 0.600	Min. :0.00900	Min. : 1.00
1st Qu.: 6.400	1st Qu.:0.2300	1st Qu.:0.2500	1st Qu.: 1.800	1st Qu.:0.03800	1st Qu.: 17.00
Median : 7.000	Median :0.2900	Median :0.3100	Median : 3.000	Median :0.04700	Median : 29.00
Mean : 7.215	Mean :0.3397	Mean :0.3186	Mean : 5.443	Mean :0.05603	Mean : 30.53
3rd Qu.: 7.700	3rd Qu.:0.4000	3rd Qu.:0.3900	3rd Qu.: 8.100	3rd Qu.:0.06500	3rd Qu.: 41.00
Max. :15.900	Max. :1.5800	Max. :1.6600	Max. :65.800	Max. :0.61100	Max. :289.00

total.sulfur.dioxide	density	pH	sulphates	alcohol	quality
Min. : 6.0	Min. :0.9871	Min. :2.720	Min. :0.2200	Min. : 8.00	3: 30
1st Qu.: 77.0	1st Qu.:0.9923	1st Qu.:3.110	1st Qu.:0.4300	1st Qu.: 9.50	4: 216
Median :118.0	Median :0.9949	Median :3.210	Median :0.5100	Median :10.30	5:2138
Mean :115.7	Mean :0.9947	Mean :3.219	Mean :0.5313	Mean :10.49	6:2836
3rd Qu.:156.0	3rd Qu.:0.9970	3rd Qu.:3.320	3rd Qu.:0.6000	3rd Qu.:11.30	7:1079
Max. :440.0	Max. :1.0390	Max. :4.010	Max. :2.0000	Max. :14.90	8: 193
NA	NA	NA	NA	NA	9: 5

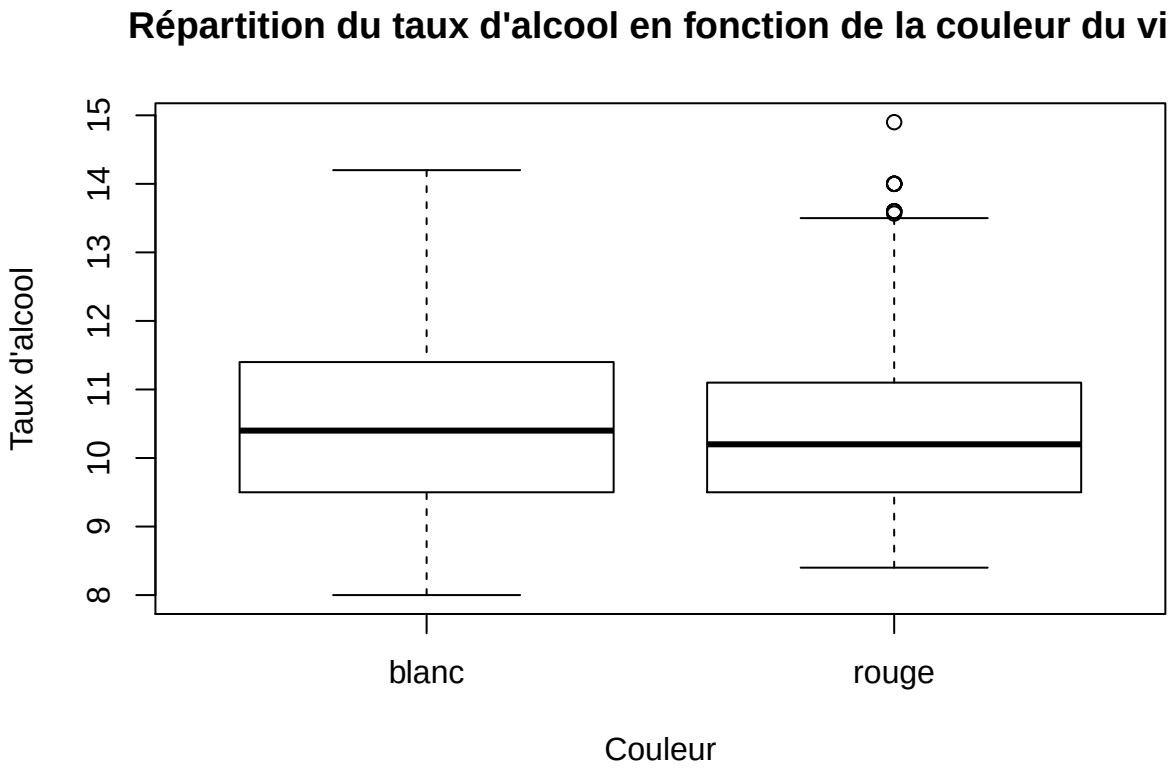


II/ Analyse du jeu de données

A) Test ϵ

Tout d'abord, nous nous sommes posés la question suivante :

Le taux d'alcool présent dans un vin est-il significativement différent en fonction du type (blanc ou rouge) de celui-ci ?



Pour répondre à cette question, nous allons effectuer un test de comparaison de moyennes dans le cas de grands échantillons, c'est à dire un test ϵ. Le but va être de comparer les moyennes d'une même variable quantitative sur deux échantillons distincts.

La variable aléatoire étudiée ici est X : le taux d'alcool présent dans un vin.

Nous formulons les hypothèses H_0 : $\mu_{blanc} = \mu_{rouge}$ et H_1 : μ_{blanc} != μ_{rouge}.

Etant donnée la taille des échantillons (n > 30 individus), nous n'avons pas besoin de vérifier de conditions d'application.

A l'issue du test ϵ réalisé sur ces deux échantillons, on obtient une p-value de 0.008. Au risque $\alpha = 5\%$, le test est donc significatif et on rejette H_0. Il y a bien une différence dans le taux d'alcool des vins blancs et des vins rouges ; les Vinhos Verdes blancs sont plus alcoolisés que les rouges.

B) Test de corrélation de Pearson

Dans un second temps, nous nous sommes interrogés :

Le pH d'un vin dépend-il de son taux d'alcool ?

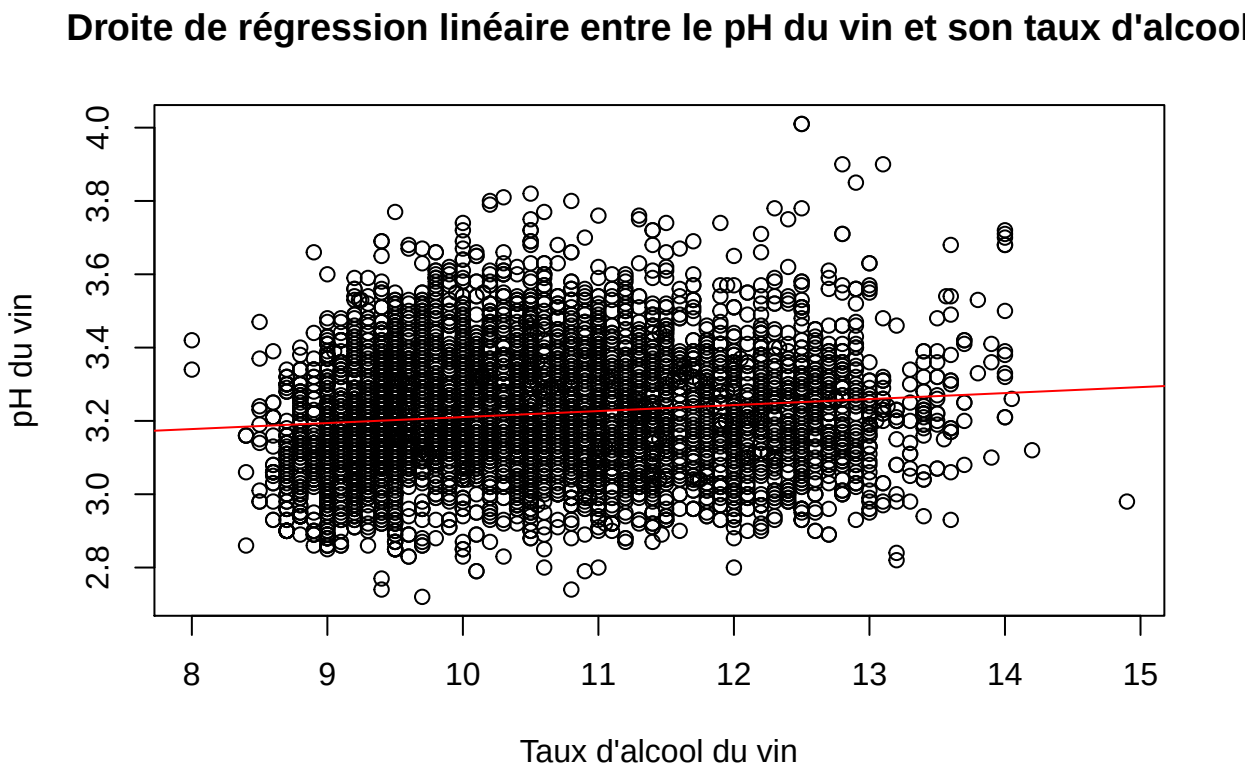
Dans cette optique, nous allons réaliser un test de corrélation de Pearson. Ce test a pour but d'étudier le lien entre deux variables quantitatives.

Les variables aléatoires que nous utilisons ici sont X : le pH d'un vin et Y : son taux d'alcool.

Les hypothèses du test sont H_0 : il n'y a pas de lien entre les deux variables et H_1 : il existe un lien entre les deux variables.

Comme nous travaillons sur de grands échantillons (n $>$ 30 individus), nous n'avons pas de conditions d'application à vérifier.

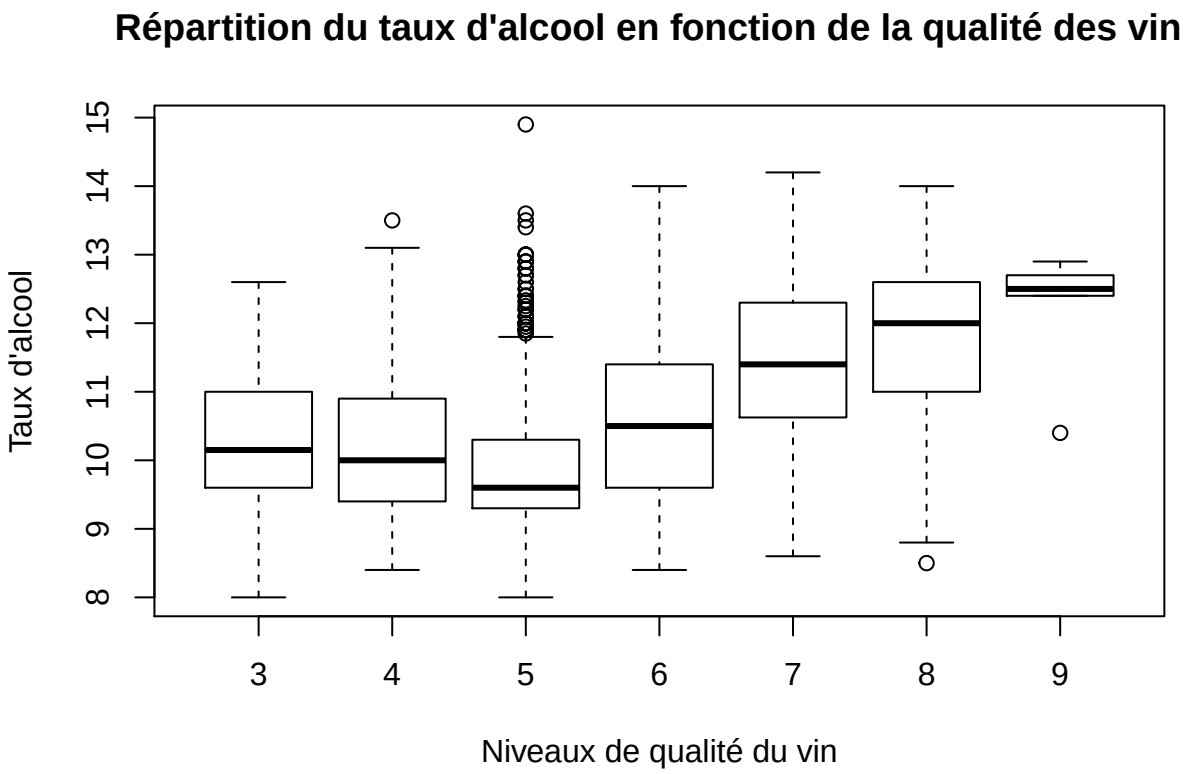
En effectuant ce test de corrélation de Pearson, on a une p-value égale à 1.05×10^{-22}. Par conséquent, au risque $\alpha = 5\%$, le test est significatif et on rejette H_0. Il y a un lien entre le pH et le taux d'alcool. Cependant, ces deux variables ne sont que très faiblement corrélées. En effet, on obtient un coefficient de corrélation de seulement 0.12.



C) Anova

Ensuite, nous nous sommes demandés :

Le taux d'alcool est-il significativement différent en fonction des différents niveaux de qualité du vin ?



Pour cela, nous allons faire un test Anova. Ce test a pour but de comparer les moyennes d'une même variable aléatoire quantitative sur plusieurs échantillons.

Nous allons utiliser la variable aléatoire X : le taux d'alcool d'un vin, avec le facteur F : la qualité du vin, à 7 niveaux (de 3 à 9)

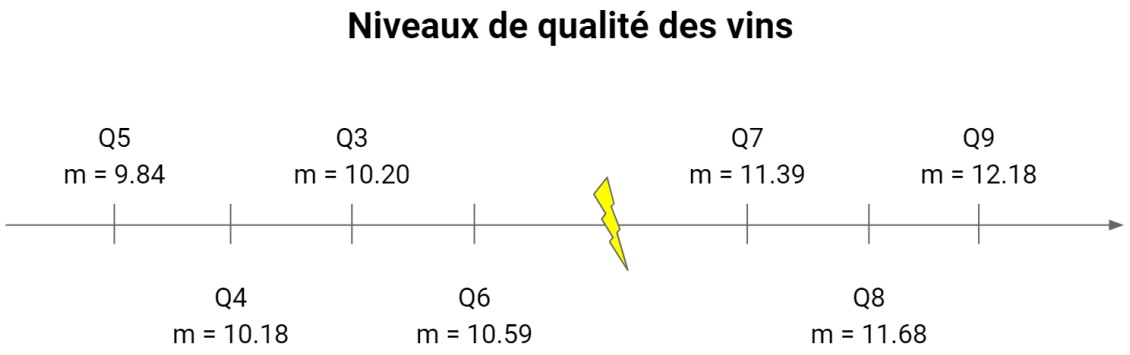
On a les deux hypothèses H_0 : $\mu_3 = \mu_4 = \mu_5 = \ldots = \mu_9$ et H_1 : au moins une des moyennes est différente.

Comme la taille des échantillons est importante, nous n'avons pas de conditions d'application à vérifier.

Ce test Anova nous donne une p-value inférieure à $2e^{-16}$. On peut donc en conclure que le test est significatif pour un risque $\alpha = 5\%$. Par conséquent, on rejette H_0, au moins une des moyennes est différentes. Afin d'identifier les points de différences entre les moyennes, nous allons réaliser des tests post-hoc.

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  dataVins$alcohol and dataVins$quality
##
##   3       4       5       6       7       8
## 4 1.00000 -       -       -       -       -
## 5 1.00000 0.00010 -       -       -       -
## 6 1.00000 7.9e-07 < 2e-16 -       -       -
## 7 3.5e-08 < 2e-16 < 2e-16 < 2e-16 -       -
## 8 2.6e-11 < 2e-16 < 2e-16 < 2e-16 0.00743 -
## 9 0.00220 0.00052 1.3e-05 0.01452 1.00000 1.00000
##
## P value adjustment method: bonferroni
```

Grâce à ces informations, il nous est possible d'identifier les différences significatives entre les moyennes, représentées sur la figure ci-dessous.



D) Test du khi2

La quatrième question que nous nous sommes posés est la suivante :

La qualité d'un vin est-elle indépendante de sa couleur ?

Pour répondre à cette interrogation, nous allons devoir effectuer un test du khi2. L'objectif de ce type de tests est d'étudier l'indépendance entre deux variables quantitatives.

Nous avons ici deux variables aléatoires, X : la couleur du vin et Y : la qualité du vin.

On formule les deux hypothèses H_0 : les variables sont indépendantes et H_1 : les variables sont liées.

Lorsque l'on effectue le test du khi2 sur nos données, on obtient le tableau des effectifs théoriques suivant :

	blanc	rouge
3	22.616592	7.383408
4	162.839464	53.160536
5	1611.809143	526.190857
6	2138.021856	697.978144
7	813.443435	265.556565
8	145.500077	47.499923
9	3.769432	1.230568

Dans les effectifs théoriques, on remarque que la case de qualité 9 pour les vins blancs et les vins rouges compte moins de 5 individus, nous allons donc refaire un test du khi2 en regroupant les classes de qualité 8 et 9.

Après avoir réalisé le test du khi2 avec ce nouveau tableau d'effectifs regroupés, on obtient une p-value de 1.9×10^{-23}. Au risque $\alpha = 5\%$, on peut donc affirmer que le test est significatif. La qualité d'un vin est influencée par sa couleur; les Vinhos Verdes blancs ont tendance à être meilleurs que les rouges.

E) Modèles de régression

Nous allons à présent créer des modèles de prédiction de différentes variables. Dans cette étude, nous nous intéresserons à deux types de modèles de régression. D'un côté, nous avons les modèles de régression linéaire, qui consistent à prédire une variable quantitative en fonction d'une ou plusieurs variables quantitatives. De l'autre, on a les modèles de régression logistique binaire, qui ont pour finalité la prédiction d'une variable qualitative à deux classes en fonction d'une ou plusieurs variables qualitatives ou quantitatives.

La première étape, pour la création d'un modèle de régression, consiste à effectuer un nettoyage des données. Pour cela, on commence par supprimer les NA. On observe qu'il n'existe aucune donnée de ce type dans la base de données. Ensuite, on supprime les variables à variance nulle. De même, nous n'en avons pas dans notre jeu de données. La dernière étape consiste à se débarrasser des valeurs aberrantes. Après avoir affiché la matrice de données en boite à moustache (boxplot), on remarque que certaines valeurs, pour residual.sugar, free.sulfur.dioxide et density, sont très élevées par rapport au reste de la population. On choisit donc de supprimer les vins pour lesquels ces variables sont trop importantes. Cela correspond à un residual.sugar supérieur à 40, à un free.sulfur.dioxide supérieur à 200, et à une density supérieure à 1.005. On vérifie ensuite la taille de la matrice, 4 vins ont été supprimés.

Régression linéaire multiple

Le premier modèle de prédiction créé est un modèle de régression linéaire multiple. Le but de celui-ci est de prédire le taux d'alcool d'un vin en fonction de toutes les autres variables quantitatives du jeu de données.

On a donc comme variables aléatoires X : le taux d'alcool d'un vin et Y_1 : le pH du vin, Y_2 : sa densité, Y_3 : le total de dioxyde de soufre, etc...

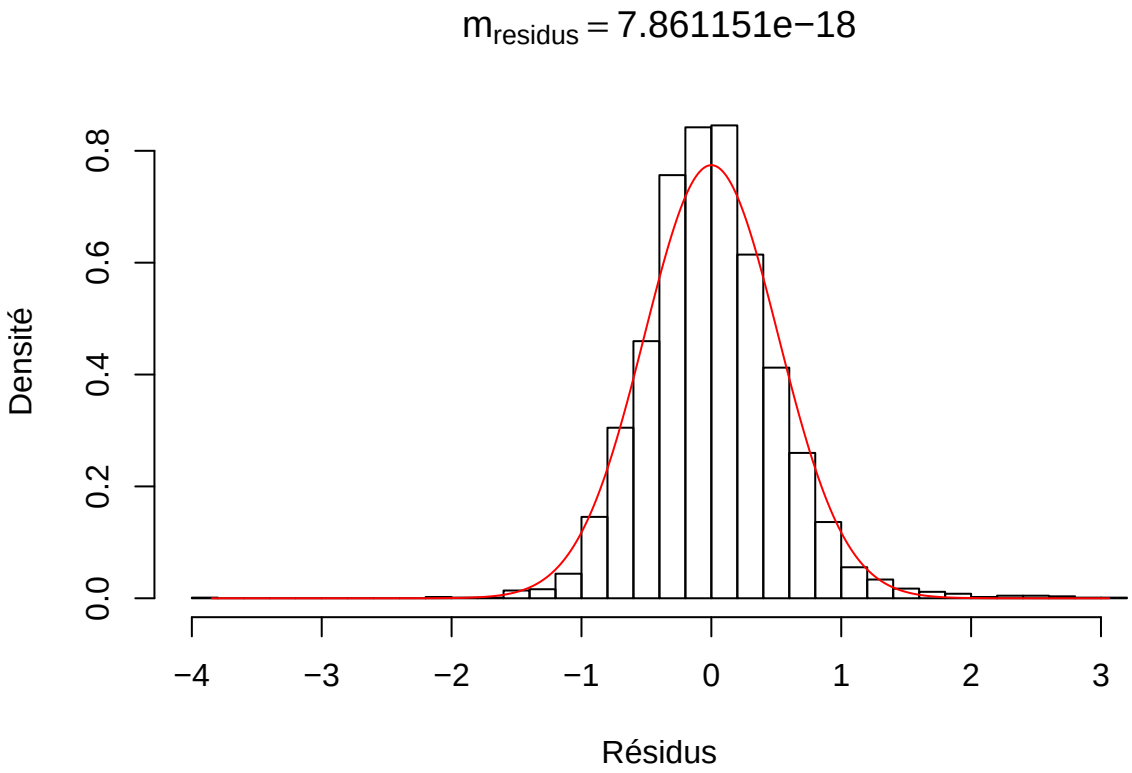
On crée tout d'abord une matrice contenant uniquement les variables que nous étudirons pour les modèles. Ensuite, on sépare la population en deux échantillons : l'échantillon d'apprentissage et l'échantillon test. L'échantillon d'apprentissage contient deux tiers de la population, tirés au hasard. On crée, par la suite, l'échantillon test contenant le reste des individus.

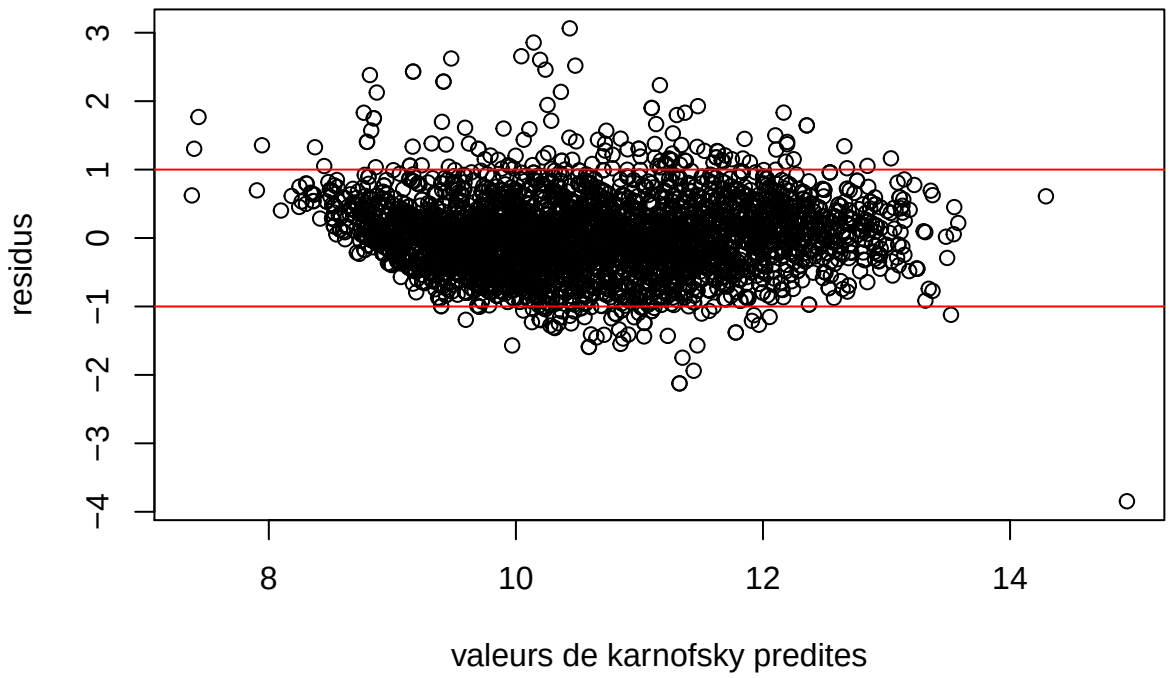
On crée ensuite le modèle à partir de l'échantillon d'apprentissage et on étudie l'implication des descripteurs. On obtient les résultats suivants :

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	616.5233004	5.4498846	113.1259361	0.0000000
fixed.acidity	0.6029091	0.0106236	56.7517746	0.0000000
volatile.acidity	1.4161669	0.0629398	22.5003317	0.0000000
citric.acid	0.3238433	0.0674818	4.7989746	0.0000016
residual.sugar	0.2071916	0.0032529	63.6940818	0.0000000
chlorides	0.0950959	0.2860722	0.3324192	0.7395889
free.sulfur.dioxide	0.0015433	0.0006429	2.4003082	0.0164233
total.sulfur.dioxide	-0.0044165	0.0002261	-19.5299715	0.0000000
density	-625.3943106	5.6504231	-110.6809698	0.0000000
pH	2.9976995	0.0642765	46.6375324	0.0000000
sulphates	1.4908959	0.0625562	23.8329207	0.0000000

On voit que les descripteurs chlorides et free.sulfur.dioxide ont une p-value supérieure à 5%. On peut donc en conclure que ces descripteurs ne permettent pas de prédire le taux d'alcool d'un vin. En revanche, on observe que toutes les autres variables ont une p-value très proche de 0 (les résultats obtenus sur le tableau ci-dessus sont arrondis). Par conséquent, nous allons créer un nouveau modèle en gardant uniquement les descripteurs significatifs.

Pour vérifier la validité de ce modèle, il faut vérifier que les résidus observés ϵ_i (avec $\epsilon_i = y_i - \hat{y}_i$) reflètent les propriétés des vraies erreurs inconnues ϵ_i. Pour cela, il faut que : La répartition des résidus constitue une loi normale, L'espérance des résidus soit nulle, L'homoscédasticité des résidus soit vérifiée, Les résidus soient indépendants les uns des autres. Si ces conditions sont respectées, cela signifie que le modèle appris est vraisemblablement correct.





Graphiquement, on voit que les résidus s'alignent sur une loi normale. De plus, l'espérance est très proche de 0 (quasiment nulle). De même, on observe que l'homoscédasticité est respectée. Les résidus montrent donc que le modèle est exploitable. En ce qui concerne les performances du modèle, on trouve $R^2 = 0.81$ et $R^2 ajusté = 0.81$. Comme ces valeurs sont proches de 1, on peut en conclure que le modèle créé est fiable et utilisable.

Régression logistique multiple

Le second modèle de prédiction réalisé est, quant à lui, un modèle de régression logistique multiple. Le but de celui-ci est de prédire la qualité d'un vin (bon ou mauvais), en fonction de toutes les variables quantitatives du jeu de données.

On a donc comme variables aléatoires X : la qualité d'un vin et Y_1 : le pH du vin, Y_2 : sa densité, Y_3 : le total de dioxyde de soufre, etc...

On commence tout d'abord par séparer les valeurs de qualité des vins en deux classes distinctes, bon (pour une qualité supérieure à 5) et mauvais (pour une qualité inférieure ou égale à 5).

Ensuite, comme pour le précédent modèle, on crée tout d'abord une matrice contenant uniquement les variables que nous étudions pour les modèles. Ensuite, on sépare la population en deux échantillons : l'échantillon d'apprentissage et l'échantillon test. L'échantillon d'apprentissage contient deux tiers de la population, tirés au hasard. On crée, par la suite, l'échantillon test contenant le reste des individus.

Lorsque l'on crée le modèle en fonction de l'échantillon d'apprentissage et que l'on étudie l'implication des descripteurs on obtient les résultats suivants :

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-71.4266606	52.3044859	-1.365593	0.1720666

	Estimate	Std. Error	z value	Pr(>|z|)
fixed.acidity	-0.1180881	0.0656180	-1.799630	0.0719191
volatile.acidity	4.3876401	0.3427723	12.800453	0.0000000
citric.acid	0.6894650	0.3110952	2.216251	0.0266743
residual.sugar	-0.0955033	0.0213385	-4.475639	0.0000076
chlorides	1.6215629	1.2079724	1.342384	0.1794715
free.sulfur.dioxide	-0.0184385	0.0030994	-5.949003	0.0000000
total.sulfur.dioxide	0.0082815	0.0010920	7.583680	0.0000000
density	82.8769854	53.3098083	1.554629	0.1200344
pH	-0.7929609	0.3734481	-2.123350	0.0337246
sulphates	-2.3998745	0.3315967	-7.237329	0.0000000
alcohol	-0.8336568	0.0771156	-10.810476	0.0000000

On observe que les descripteurs fixed.acidity, citric.acid, chlorides, density et pH ont une p-value supérieure à 5%. On peut donc en conclure que ces descripteurs ne permettent pas de prédire la qualité d'un vin. En revanche, on observe que toutes les autres variables ont une p-value très proche de 0 (les résultats obtenus sur le tableau ci-dessus sont arrondis). Par conséquent, nous allons créer un nouveau modèle en gardant uniquement les descripteurs significatifs.

Afin de connaître la pertinence du modèle, nous allons effectuer des calculs de performance en resubstitution (c'est-à-dire sur l'échantillon d'apprentissage) et en généralisation (sur l'échantillon de test). Pour réaliser ces calculs, nous avons besoin de connaître les biens prédits et les mals prédits, répartis selon le tableau suivant :

	Positifs observés	Négatifs observés
Positifs prédits	Vrais positifs (VP)	Faux positifs (FP)
Négatifs prédits	Faux négatifs (FN)	Vrais négatifs (VN)

A partir de ces effectifs, nous sommes capables de calculer trois valeurs : le Taux de Bien Prédits (TBP), la Sensibilité (Se) et la Spécificité (Sp).

$$TBP = \frac{VN + VP}{VP + VN + FP + FN}$$

$$Se = \frac{VP}{VP + FN}$$

$$Sp = \frac{VN}{VN + FP}$$

En resubstitution, on trouve TBP = 0.75, Se = 0.77 et Sp = 0.69. En généralisation, on a TBP = 0.73, Se = 0.76 et Sp = 0.66. Ces résultats sont assez moyens (0.5 étant le hasard), ils traduisent une prédiction relativement peu fiable. L'utilisation de ce modèle est donc peu pertinente.

Conclusion

Pour conclure, l'étude de ce jeu de données était très intéressante. En effet, elle a donné lieu à de nombreux résultats significatifs sur les données, ce qui nous a permit d'effectuer de bonnes analyses. Toutefois, on pourra s'interroger sur la pertinence de certains résultats, comme le test de corrélation de Pearson ou le test du khi2, en particulier à cause de la surreprésentation de vins blancs par rapport aux vins rouges.