Projet de statistiques

Production vinicole française en 2015 par département

# Introduction

La France est le 2e producteur mondial de vin derrière l’Italie et devant l’Espagne. En 2015, la production française représentait un peu plus de 17% de la production mondiale, soit environ 48 millions d’hectolitres. De plus, comme environ un tiers de cette production est exportée, le secteur des vins et spiritueux est le second secteur d'exportation bénéficiaire en France derrière l’aéronautique. Il s’agit donc d’un secteur économique très important pour le pays.

Cette production est très hétérogène car il y a en France de nombreux terroirs différents, et donc de nombreux types de vins. Pour protéger ces spécificités, il existe des dizaines d’appellations sous forme d’AOP ou d’IGP. Elles permettent de contrôler l’origine du produit et, dans le cas des AOP, le respect d’un cahier de charges.

Les données statistiques relatives à la production vinicole française sont éditées tous les ans par la douane, qui les met à disposition sur son site internet <http://www.douane.gouv.fr/>. Nous allons ici analyser une partie de ces données afin d’essayer de dégager certaines caractéristiques de cette production.

# Analyse des données

## Caractéristiques principales des données

On dispose de 23 variables, dont 2 qualitatives qui correspondent aux noms des départements et des régions :

* *departement* : Nom du département
* *nombre\_declarations* : Nombre de de producteurs déclarés dans ce département
* *superficie* : Superficie totale des vignobles du département (ha)
* *superficie\_aop* : Superficie des vignobles du département contrôlés par une Appellation d’Origine Contrôlée (ha)
* *superficie\_cognac* : Superficie des vignobles du département destinés à la production de cognac (ha)
* *superficie\_igp* : Superficie des vignobles du département contrôlés par une Indication Géographique Protégée (ha)
* *superficie\_vsig* : Superficie des vignobles du département non contrôlés par une quelconque indication géographique (ha)
* *qte\_aop\_blanc*, *qte\_aop\_rouge*, *qte\_aop\_rose* : Production vinicole du département sous une AOP, correspondant respectivement à du vin blanc, du rouge et du rosé (hl)
* *qte\_vci* : Volume Complémentaire Individuel (hl)
* *qte\_igp\_blanc*, *qte\_igp\_rouge*, *qte\_igp\_rose*, *qte\_igp\_vci* : Production vinicole du département sous une IGP, correspondant respectivement à du vin blanc, du rouge, du rosé, et du VCI (hl)
* *qte\_vsig\_blanc*, *qte\_vsig\_rouge*, *qte\_vsig\_rose* : Production vinicole du département sans indication géographique, correspondant respectivement à du vin blanc, du rouge et du rosé (hl)
* *qte\_cognac* : Production de cognac du département (hl)
* *prod\_commerciale* : Production du département commercialisée (hl)
* *prod\_noncommerciale* : Production du département non commercialisée. Cela correespond aux jus de raisins produits par les viticulteurs, les lies et/ou les vins en dépassement de rendement (hl)
* *total* : Production totale du département
* *region* : Région correspondant au département

On remarque que la variable *total* correspond à la somme de toutes les variables exprimant le volume d’un type de production – celles commençant par « qte » - mais aussi la somme des variables *prod\_commerciale* et *prod\_noncommerciale*.

> summary(vins$total)

Min. 1st Qu. Median Mean 3rd Qu. Max.

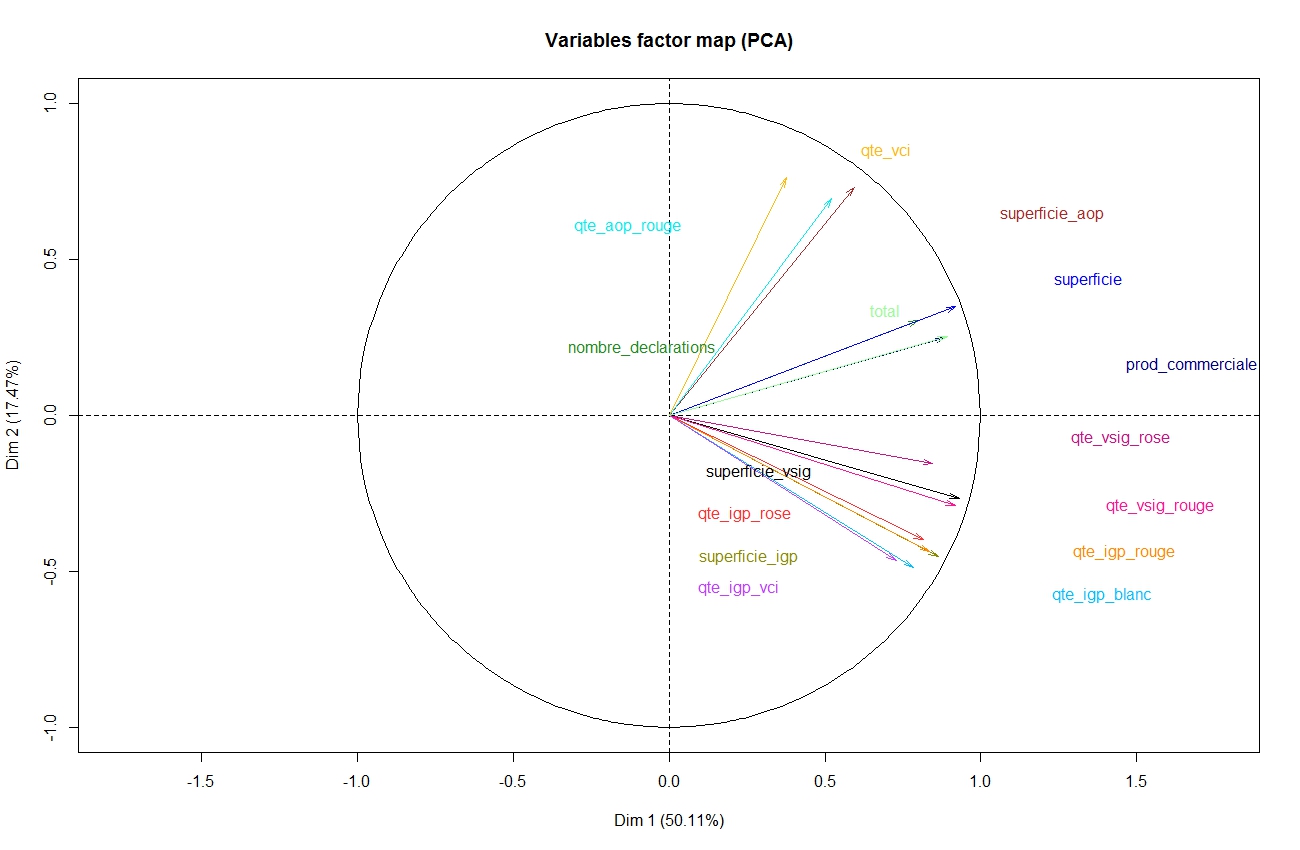
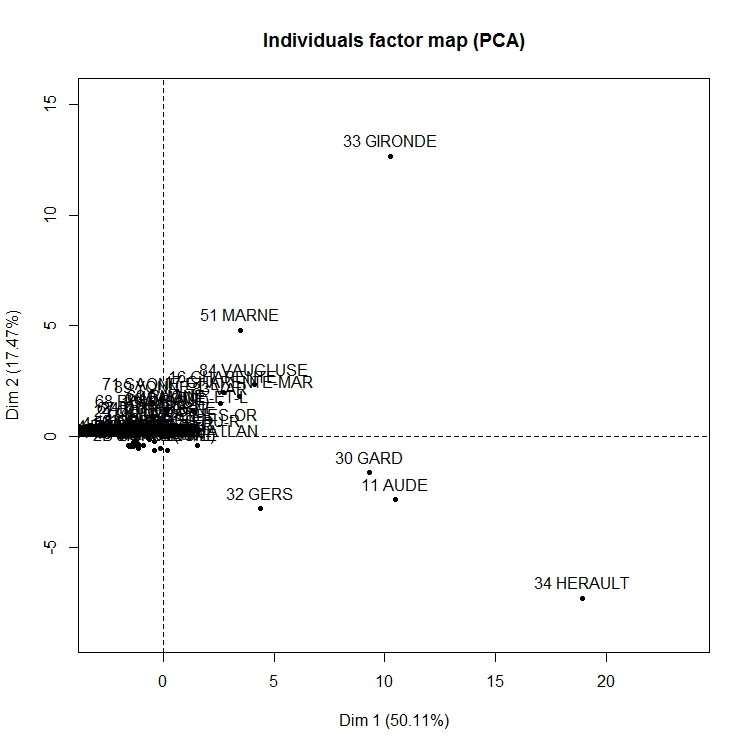
4 3948 66607 605767 516658 5672267

Nous pouvons voir que les niveaux de production sont extrêmement différents d’un département à l’autre.

## Réalisation d’une analyse en composantes principales

Notre étude est basée sur un nombre conséquent de variables. De plus, mis à part les variables de superficies et de productions, et la variable « total » qui est une somme de variables ; il est difficile de trouver des liens corrélations évidentes entre ces variables. Nous avons donc décidé de réaliser une ACP afin d’avoir une première approche des corrélations potentielles. Voici les résultats que nous avons obtenus :

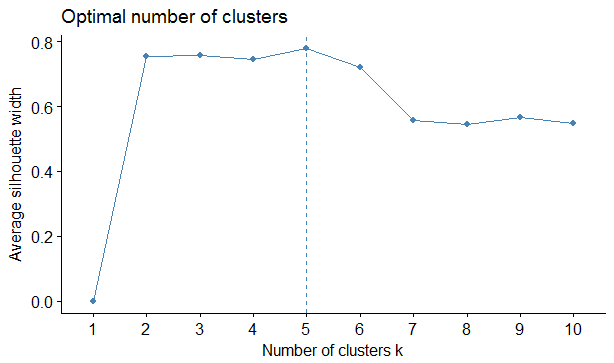
plot.PCA(res.acp,choix="var",1:2,col.hab=c("forestgreen",…),lim.cos2.var = 0.7)



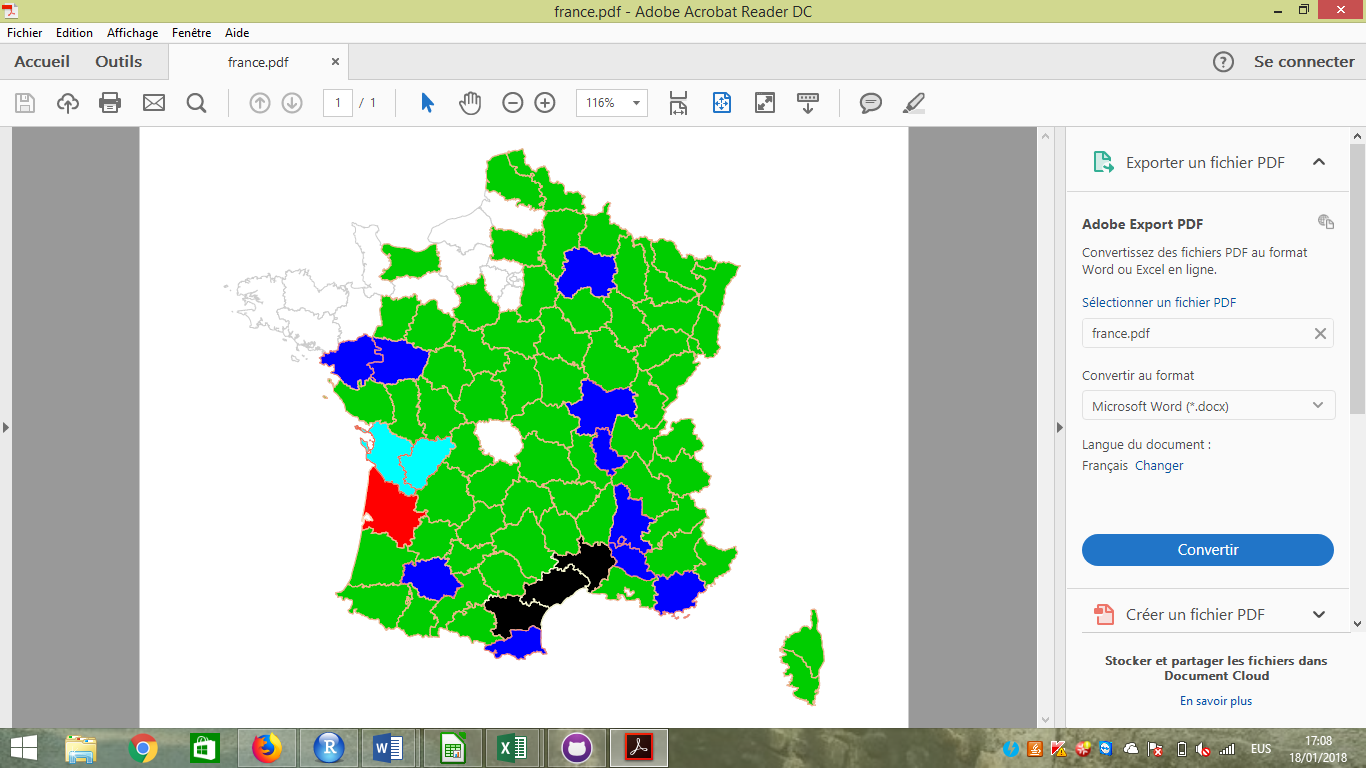
L’axe 1 porte 50,11% de l’information et l’axe 2 17,47%. Ainsi, entre les deux axes, 67,58% de l’information est représentée. D’après le graphique des individus, certains départements ont des productions très différentes des autres. Cela est particulièrement vrai pour la Gironde et l’Hérault. On décide donc de répartir en clusters les différents départements dans le but de prendre en compte au mieux ces différences par la suite.

## Distribution des départements en clusters

Notre objectif est de rassembler les départements en clusters dans le but de diminuer par la suite le nombre de variable utilisé dans les futures modélisations. Nous avons utilisé une détermination automatique du nombre de cluster afin d’avoir le meilleur nombre possible. Nous avons obtenu le résultat suivant :



Nous avons donc formé 5 clusters qui regroupent les différents départements en fonction des caractéristiques de leurs productions vinicoles. Ces clusters sont représentés dans la carte suivante :



Numéro de cluster :

a

b

c

d

e

**Représentation cartographique des clusters**

# Les déterminants de la production non commerciale7

La production non commerciale représente un peu plus de 2% de la production totale. C’est donc une quantité non négligeable qui ne peut pas être commercialisé comme le reste des produits, et qu’il faudrait probablement essayer de réduire. Nous voudrions essayer de savoir pourquoi il y a autant de production non commerciale. Pour cela, nous chercherons à déterminer à quel type de vin correspond majoritairement cette production.

## Tests de corrélation avec les indications géographiques

Le but est ici de savoir si une des indications géographiques contribue plus que les autres à cette production non commerciale. En effet, cela nous renseignerait sur l’origine et peut-être même les raisons de l’existence d’une telle production non commerciale. On réalise donc des tests de corrélation entre la proportion de production non commerciale et celles des 3 différentes indications géographiques, soit les AOP, les IGP et les VSIG. Nous préférons utiliser des proportions car les niveaux de production sont très différents d’un département à l’autre, et qu’il est évident que la production totale varie dans le même sens que toutes ces productions. De plus, nous n’avons pas effectué de test concernant le cognac car les 2 départements qui en produisent l’immense majorité ont une très faible proportion de production non commerciale. On applique les lignes de codes suivantes pour les 3 :

> p.aop=(vins$qte\_aop\_blanc+vins$qte\_aop\_rouge+vins$qte\_aop\_rose)/vins$total

> p.noncom=vins$prod\_noncommerciale/vins$total

> test.aop=cor.test(p.aop,p.noncom)

> test.aop

Voici les résultats que nous avons obtenus :

|  |  |  |
| --- | --- | --- |
| AOP | IGP | VSIG |
| cor  0.2377902 | cor  0.07756447 | cor  -0.37675 |

Il y a donc une faible corrélation positive entre les proportions d’AOP des départements et leur production non commerciale. Cela peut-être dû aux limites de rendements imposés aux producteurs par ces organismes. Ainsi, l’éventuel excès de production passerait en production non commerciale. De plus, il y a aussi une corrélation négative entre la proportion de VSIG et celle de production non commerciale. Cela est probablement dû à la même raison car les producteurs de vins VSIG n’ont pas de limite de rendement à suivre. Enfin, la proportion de vin IGP ne semble pas être corrélée à celle de la production non commerciale. Les producteurs de vins IGP doivent aussi respecter des limites de production, entre 90 et 120 hL/ha en général. Mais ces limitations sont souvent plus élevées que celles des AOP. Il est possible qu’elles soient trop élevées pour induire une importante production de vin non commerciale, sous réserve que notre hypothèse sur l’influence de ces limitations soit vraie.

## Tests de corrélation avec la production par couleur

Nous essayons maintenant de voir si au sein des AOP et des VSIG une couleur en particulier participe plus aux variations de production non commerciale que les autres. Pour cela, nous appliquons le même type de test, mais cette fois aux quantités produites par couleur et par appellation. Nous obtenons les résultats suivants :

|  |  |  |  |
| --- | --- | --- | --- |
| Appellation | Rouge | Rosé | Blanc |
| AOP | cor  -0.01616693 | cor  -0.02926049 | cor  0.3128075 |
| VSIG | cor  -0.3226081 | cor  0.0137295 | cor  -0.01510555 |

Nous remarquons que la seule corrélation positive relativement importante (même si elle reste faible) et celle concernant la proportion de vins blancs AOP. La proportion de production non commerciale serait donc légèrement corrélée à celle de vin blanc AOP. Il se pourrait que cela soit lié aux faibles rendements limites imposés par certaines AOP renommées, comme le Jurançon par exemple, dont le rendement ne doit pas dépasser 40 hL/ha.

Au contraire, ce sont les vins rouges VSIG qui seraient négativement corrélées à la proportion de production non commerciale. On peut penser que ce sont les vins ayant rendement le plus élevé à l’hectare. En effet, ces vins le plus souvent vendus en supermarché sont plus reconnus pour leur prix que pour leur qualité.

En somme, la proportion non commerciale est positivement corrélée à la proportion de vins AOP, et plus particulièrement à celle de vins blancs AOP, alors que sa corrélation est négative avec la proportion de vins VSIG, et plus particulièrement avec celle de vins rouges VSIG. Cela pourrait être dû aux réglementations des AOP qui imposent une limite de rendement à leurs producteurs afin de garantir la qualité du produit. Cependant, ce ne sont que des suppositions et ces résultats sont à prendre avec des pincettes car les corrélations sont tout de même relativement faibles.

# Modélisation de la production totale

La variable la plus importante de ce jeu de données est probablement la production totale. Dans cette partie, nous chercherons à créer un modèle juste et efficace pour modéliser cette variable.

## Modélisation en utilisant les clusters

Pour qu’un modèle soit efficace, il faut que le nombre de variables utilisés soit réduit. Nous pensons donc qu’en utilisant des clusters, qui intègrent les données des autres variables, nous obtiendrons un bon modèle.

Tout d’abord, nous avons commencé par comparer le modèle intégrant les clusters avec le modèle nul.

Yi,k : Production totale pour le département k du cluster i (variable *total*)

 : Effet additif moyen

 : Effet additif lié au cluster i

ei,k : Erreur standard résiduelle. Elles sont indépendantes et identiquement distribuées :

Voici les résultats que nous avons obtenus :

> anC<-lm(formula=total~Cluster,data=Data\_Vins) #modèle avec clusters

> an0<-lm(formula=total~1,data=Data\_Vins) #modèle nul

> anova(anC,an0)

Analysis of Variance Table

Model 1: total ~ Cluster

Model 2: total ~ 1

Res.Df RSS Df Sum of Sq F Pr(>F)

1 74 7.3343e+12

2 78 1.2709e+14 -4 -1.1976e+14 302.08 < 2.2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Comme la p-value est très faible, nous conservons ce modèle par rapport au modèle nul. Comme on pouvait le penser, on peut donc modéliser la production totale avec les clusters. Cependant, si en effectuant un sommaire du modèle testé, nous observons que la p-value correspondant au cluster e est supérieure à 5%.

> summary(an)

Call:

lm(formula = total ~ Cluster, data = Data\_Vins)

Residuals:

Min 1Q Median 3Q Max

-899612 -122215 -99932 84462 1283864

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4269014 181762 23.487 < 2e-16 \*\*\*

Clusterb 1403253 363525 3.860 0.00024 \*\*\*

Clusterc -4146230 186040 -22.287 < 2e-16 \*\*\*

Clusterd -3070088 207241 -14.814 < 2e-16 \*\*\*

Clustere 556832 287391 1.938 0.05650 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 314800 on 74 degrees of freedom

Multiple R-squared: 0.9423, Adjusted R-squared: 0.9392

F-statistic: 302.1 on 4 and 74 DF, p-value: < 2.2e-16

Ce cluster ne permet donc pas de modéliser la production totale pour les deux départements concernés avec le niveau de confiance que nous recherchons. Il s’agit de la Charente et la Charente-Maritime. Cette anomalie est très probablement due à leur production de cognac très importante.

A l’inverse, pour les autres clusters, nous pouvons penser que la production totale a joué un rôle important dans leur détermination, notamment via toutes les autres variables de production. C’est un fait que nous pouvons visualiser grâce au cluster a qui regroupe des départements qu’on connaît comme étant à forte productions vinicoles.

Cependant, ce modèle est critiquable car pour réaliser ces clusters nous avons utilisé toutes les variables disponibles. Il n’est donc pas tout à fait vrai que nous avons obtenu un modèle simple car les clusters intègrent toutes les autres variables, dont les variables de production. Notre modèle peut être utile si nous disposons de toutes les autres variables car il est en apparence plus simple. Mais si ce n’est pas le cas, il est inutilisable. Nous allons donc par la suite essayer de créer un modèle qui n’utilise pas de variable de production.

## Modélisation sans utiliser les variables de production

Nous avons dans notre jeu de données 6 variables qui ne sont pas des variables de production. Notre but est toujours de créer le modèle le plus efficace possible, soit le modèle qui a le moins de variables possible mais qui s’ajuste le mieux possible aux données. Nous pensons que certaines variables, comme celles concernant les surfaces, pourraient être redondantes. On pourrait peut-être supprimer certaine(s) d’entre elles. Pour cela, le meilleur moyen selon nous est d’utiliser une méthode « pas à pas » avec le critère d’Akaike. Nous avons choisi de procéder à une procédure descendante. Nous avons obtenu les résultats suivants :

> total.lm=lm(total~nombre\_declarations+superficie+superficie\_aop+superficie\_cognac+superficie\_igp+superficie\_vsig,data=Data\_Vins)

> par(mfrow=c(2,2))

> step(total.lm)

Start: AIC=1830.44

total ~ nombre\_declarations + superficie + superficie\_aop + superficie\_cognac +

superficie\_igp + superficie\_vsig

Df Sum of Sq RSS AIC

- superficie 1 7.8411e+09 7.7227e+11 1829.2

- superficie\_aop 1 7.8741e+09 7.7230e+11 1829.2

- superficie\_igp 1 7.8804e+09 7.7231e+11 1829.2

- superficie\_cognac 1 7.9343e+09 7.7236e+11 1829.3

- superficie\_vsig 1 7.9982e+09 7.7242e+11 1829.3

<none> 7.6443e+11 1830.4

- nombre\_declarations 1 4.8281e+11 1.2472e+12 1867.1

Step: AIC=1829.25

total ~ nombre\_declarations + superficie\_aop + superficie\_cognac +

superficie\_igp + superficie\_vsig

Df Sum of Sq RSS AIC

<none> 7.7227e+11 1829.2

- superficie\_vsig 1 3.4815e+11 1.1204e+12 1856.7

- nombre\_declarations 1 4.7498e+11 1.2473e+12 1865.1

- superficie\_igp 1 1.9862e+12 2.7585e+12 1927.8

- superficie\_aop 1 1.1624e+13 1.2396e+13 2046.5

- superficie\_cognac 1 2.8341e+13 2.9113e+13 2114.0

Call:

lm(formula = total ~ nombre\_declarations + superficie\_aop + superficie\_cognac +

superficie\_igp + superficie\_vsig, data = Data\_Vins)

Coefficients:

(Intercept) nombre\_declarations superficie\_aop superficie\_cognac

-38053.84 62.67 42.42 119.48

superficie\_igp superficie\_vsig

50.51 201.00

Le meilleur modèle selon ce critère serait donc celui qui inclue toutes les variables quantitatives ne concernant pas directement la production, sauf la superficie totale qui est donc un surparamètre. Nous pourrions la formaliser comme ceci :

Où :

* Yi: La variable aléatoire modélisant la production vinicole totale du département i,
* : La valeur de la k-ème variable pour le département i. Ici,
* Ei : Les erreurs résiduelles standards. Elles sont toutes iid avec

Nous avons quand même quelques doutes par rapport à ce modèle. Est-il vraiment utile de garder toutes ces variables modélisant les superficies alors que la variable *superficie* les intègre toutes ?

## Ergonomie du modèle

Nous voulons estimer la différence de fiabilité entre le modèle obtenu dans le modèle précédent et un modèle qui remplacerait toutes les variables modélisant des superficies par la somme de toutes, soit *superficie*. Pour cela, nous effectuons une ANOVA entre ces deux modèles :

|  |
| --- |
| > an2=lm(formula=total~nombre\_declarations+superficie,data=Data\_Vins)  > an3=lm(formula=total~nombre\_declarations+superficie\_aop+superficie\_igp+superficie\_cognac+superficie\_vsig,data=Data\_Vins)  > anova(an2,an3)  Analysis of Variance Table  Model 1: total ~ nombre\_declarations + superficie  Model 2: total ~ nombre\_declarations + superficie\_aop + superficie\_igp +  superficie\_cognac + superficie\_vsig  Res.Df RSS Df Sum of Sq F Pr(>F)  1 76 1.5035e+13  2 73 7.7227e+11 3 1.4263e+13 449.4 < 2.2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
|  |
|  |

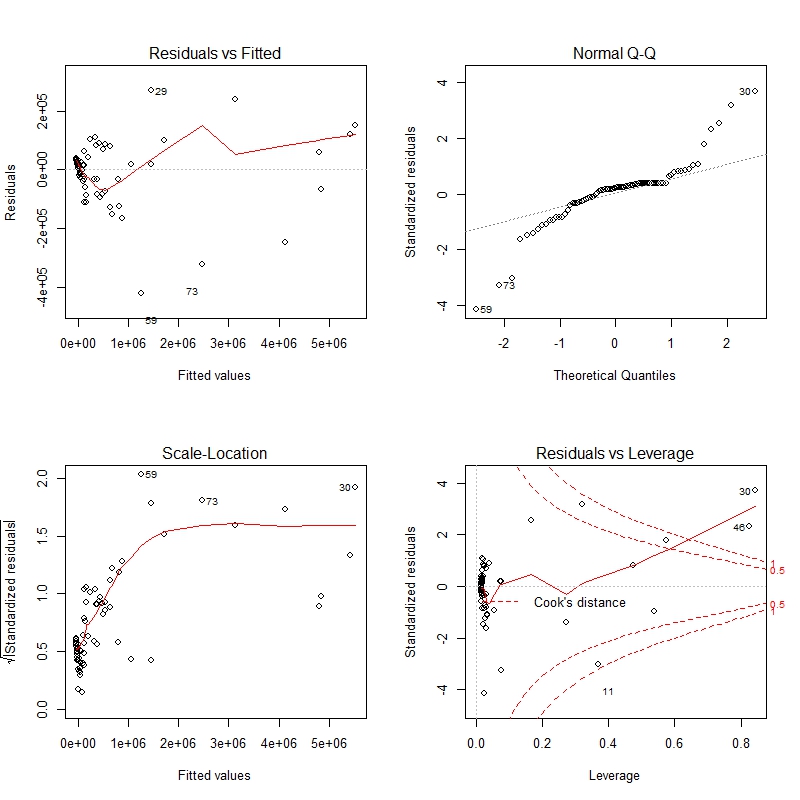
La p-value obtenue est très faible. Notre intuition était donc mauvaise et la variable superficie à elle seule n’offre pas une précision suffisante. Le modèle retenu est donc celui obtenu dans la partie précédente.

## Vérification de la qualité du modèle

Nous venons donc de trouver le modèle le plus efficace parmi ceux que nous avons obtenus. Cependant, il nous faut encore vérifier les hypothèses que nous avons faites. Pour cela, nous utilisons les 4 graphes de diagnostic fournis par le logiciel R.

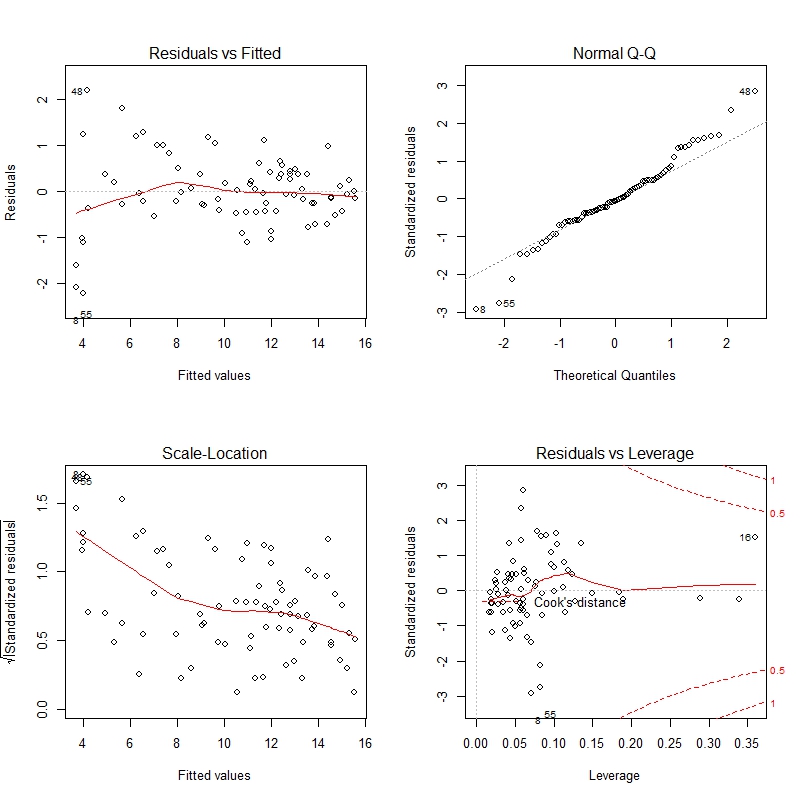
> par(mfrow=c(2,2))

> plot(an3)

Ces graphes correspondent à un modèle où l’hypothèse d’homoscédasticité ne serait pas respectée. Notre modèle est donc à revoir. Une des transformations les plus courantes dans ces cas-là est le passage au logarithme. Nous testons donc le modèle suivant :

où les notations sont les mêmes que précédemment.

Nous avons rajouté 1 à chaque logarithme car certaines valeurs prises par les variables modélisant des surfaces sont nulles et leur logarithme est donc. Or, R ne peut pas gérer une telle modélisation avec des valeurs infinies. Comme les valeurs que nous avons sont très grands par rapport à 1, nous négligeons la différence induite. Voici les graphes de diagnostic que nous avons obtenus :

 Ces résultats sont meilleurs que ceux obtenus avec le modèle précédent, même s’ils ne sont toujours pas parfaits. Nous retenons donc ce modèle par rapport à l’autre.

Remarque : Nous avons aussi essayé de transformer le modèle en passant à la racine, mais nous avons obtenus de meilleurs résultats avec le logarithme.