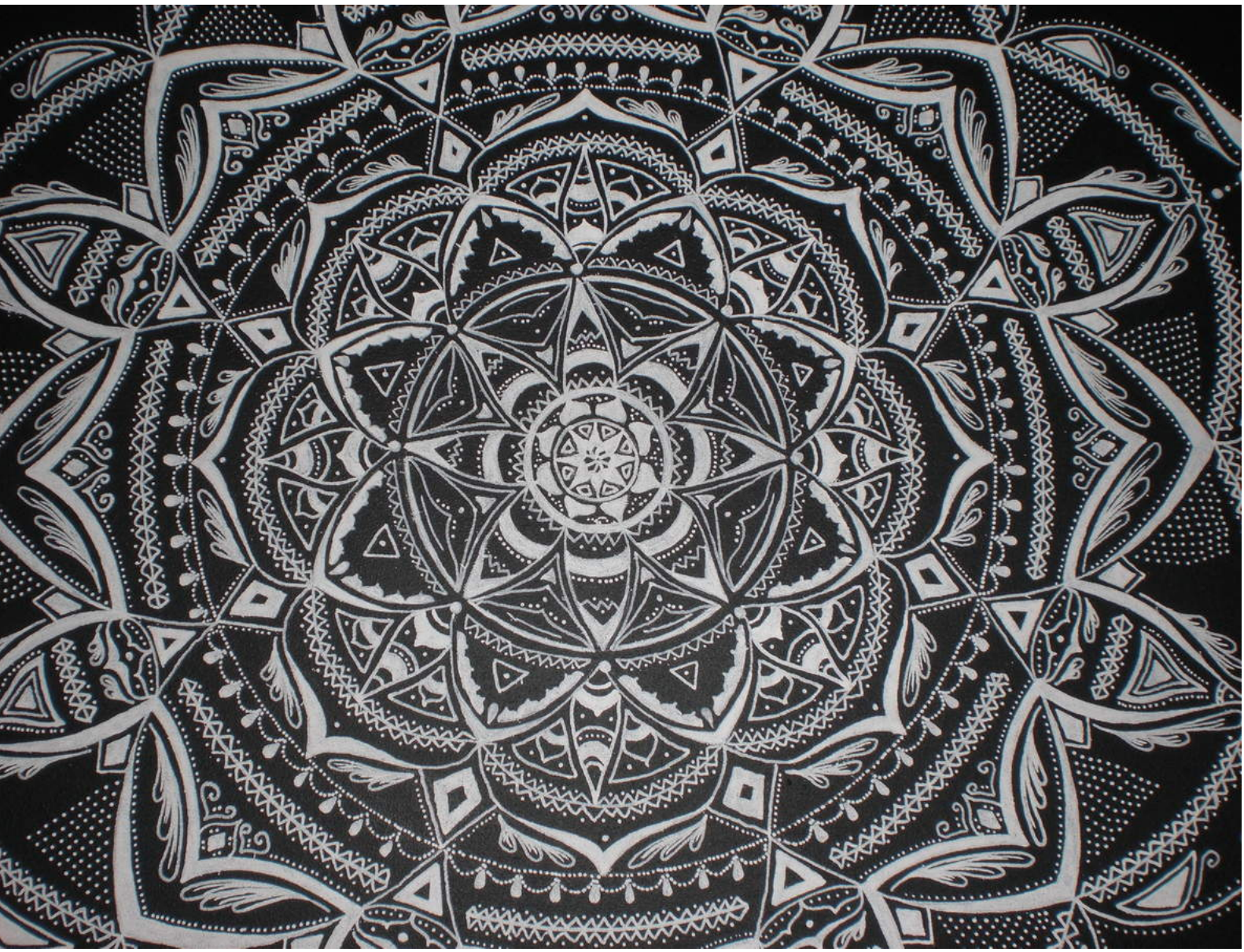
P

RÉGRESSION LINÉAIRE LOGISTIQUE AVEC AJUSTEMENT DES DONNÉES POUR LA DÉTECTION DES MATÉRIAUX

Amin EL GAREH



1. INTRODUCTION

Nous souhaitons expliquer à partir d'un modèle statistique la nature (métallique ou rocheuse) d'un objet visé par sonar. Les mesures effectuées dans les différentes longueurs d'onde ont été recueillies et mises en relation avec le matériau de l'objet ciblé.
Le choix du modèle à retenir se doit d'être en cohérence avec les données, puisque nous nous intéressons à la prédiction d'une variable qui prend deux modalités: 'M' pour métallique et 'R' pour rocheuse, alors la régression logistique, cas particulier du modèle linéaire généralisé, s'avère être adaptée à notre situation. La régression logistique est connue pour avoir été la première méthode utilisée, notamment en marketing pour le scoring et en épidémiologie, pour aborder la modélisation d'une variable binaire binomiale ou de Bernoulli: possession on non d'un produit, décès ou survie d'un patient, absence ou présence d'une pathologie...
Cependant, elle conduit à des interprétations pouvant être complexes mais rentrées dans les usages pour quantifier, par exemple, des facteurs de risque liés à une pathologie, une faillite...
Cette méthode reste donc celle la plus utilisée même si, en terme de qualité prévisionnelle, d'autres approches sont susceptibles, en fonction des données étudiées, d'apporter de bien meilleurs résultats.

2. PRÉSENTATION DU MODÈLE LINÉAIRE GÉNÉRALISÉ & LOGIT

Le modèle linéaire généralisé a été développé à partir de 1972 par Nelder et Wedderburn, dont l'exposé détaillé est présenté dans les ouvrages de Nelder et Mc Cullagh (1983), d'Agresti (1990) ou d'Antoniadis et al. (1992).
L'idée ici est d'introduire le cadre théorique général permettant de regrouper tous les modèles linéaires, en particulier celui dit logit, et qui repose sur le fait d'exprimer l'espérance de la variable à expliquer en fonction d'une combinaison linéaire des variables explicatives.

2.1. Distribution et densité de la variable à expliquer

Soit un échantillon constitué de n variables aléatoires $\{Y_i,\ i = 1, ..., n\}$ indépendantes admettant des distributions issues d'une structure exponentielle. Cela signifie que la densité (par rapport à une mesure de comptage ou la mesure de Lebesgue) de la variable Y_i s'écrit sous la forme:

$$f(y_i, \theta_i, \phi) \ = \ exp\left(\ \frac{y\,\theta_i \ - \ b(\theta_i)}{a(\phi)} \ + \ c(y_i, \phi) \ \right)$$

où θ_i est le paramètre de position et ϕ le paramètre de dispersion. On a par ailleurs

$$\mathbb{E}[Y_i] \ = \ \mu_i \ = \ b'(\theta_i) \quad \text{et} \quad Var(Y_i) = b''(\theta_i)\, a(\phi)$$

La densité de la variable Y_i peut aussi se mettre sous forme canonique.
D'abord en remarquant que pour certaines lois la fonction $a(\phi)$ s'écrit:

$$a(\phi) \ = \ \frac{\phi}{\omega_i} \quad \text{avec } \omega_i \text{ sont les poids connus des observations, fixés ici à 1.}$$

Et ensuite en posant $Q(\theta_i) = \frac{\theta_i}{\phi}$, $d(\theta_i) = exp(-\frac{b(\theta_i)}{\phi})$ et $e(y_i) = exp\,(\,c(y_i, \phi)\,)$, on obtient la densité de Y_i sous forme canonique suivante:

$$f(y_i, \theta_i) \; = \; d(\theta_i)\, e(y_i)\, exp(\; y_i\, Q(\theta_i)\;) \tag{1}$$

2.2. Régression linéaire généralisée

Les observations des variables explicatives sont organisées dans la matrice X, et β est un vecteur de $p+1$ paramètres, qui réunit les p coefficients des variables explicatives ainsi que la constante de régression. Le prédicteur linéaire, composante déterministe du modèle, est le vecteur à n composantes :

$$\eta = X\, \beta$$

2.3. Fonction de lien

Les fonctions de lien usuelles sont les fonctions de liens canoniques, supposée monotone et différentiable qui vérifient par définition,

$$g(\mu_i) = \theta_i = \eta_i \quad \text{où} \;\; \mu_i = \mathbb{E}[Y_i]$$

2.4. Régression linéaire généralisée avec la fonction logit

Considérons n variables aléatoires indépendantes notées Y_i, qui sont qualitatives (de modalités 1 ou 0), telles que la probabilité de succès est π_i et d'espérance $\mathbb{E}[Y_i] = \pi_i$.
La fonction de densité de Y_i est élément de la famille:

$$f(y_i, \pi_i) \; = \; \pi_i^{y_i}\, (1-\pi_i)^{1-y_i} \; = \; (1-\pi_i)\, exp\left(\; y_i\, ln\left(\; \frac{\pi_i}{1-\pi_i} \;\right)\right)$$

En identifiant avec les termes de **(1)**, on remarque que:

$$Q(\theta_i) \; = \; ln\left(\; \frac{\pi_i}{1-\pi_i} \;\right)$$

Or comme $Q(\theta_i) = \frac{\theta_i}{\phi}$, et puisque la mesure de dispersion ϕ est égale à 1 alors

$$Q(\theta_i) = \theta_i = g(\pi_i)$$

En sommes, la fonction de lien dite fonction logit est définie par

$$g(\pi_i) \; = \; ln\left(\; \frac{\pi_i}{1-\pi_i} \;\right)$$

Et le modèle dit de régression logistique s'écrit

$$ln\left(\; \frac{\pi_i}{1-\pi_i} \;\right) \; = \; x_i'\, \beta$$

3. APPLICATION DU MODÈLE LOGIT POUR LA DÉTECTION DES MATÉRIAUX

3.1. Étude exploratoire des données

Nos données sont réunies dans un échantillon de 208 observations prises pour 61 variables. Les 60 premières variables quantifient l'énergie retransmise (après 'normalisation') dans les différentes longueurs d'onde. Tandis que la dernière variable qualifie la nature de l'objet, elle a été binarisée et vaut '1' si l'objet visé est du type rocheux et '0' s'il est métallique. Nous nous sommes intéressés aux variables quantitatives, et nous avons constaté qu'elles étaient fortement corrélées lorsqu'elles étaient voisines. Si on considère de telles variables corrélées comme étant prédictives, alors apparaissent des propriétés hautement indésirables dans notre modèle. Ce qui nous amène à un dilemme, soit on prend cet ensemble de variables pertinentes aussi complet que possible, au risque d'avoir des coefficients ininterprétables (choix exhaustif), soit on prend peu de variables bien qu'étant susceptibles d'être moins significatives (choix par parcimonie). Un autre aspect et non des moindres qui régit nos données est le fait qu'il existe un certain nombre de points influents, augmentant la variabilité des variables. Il est envisageable de réduire leurs variabilités au risque de perdre de l'information.

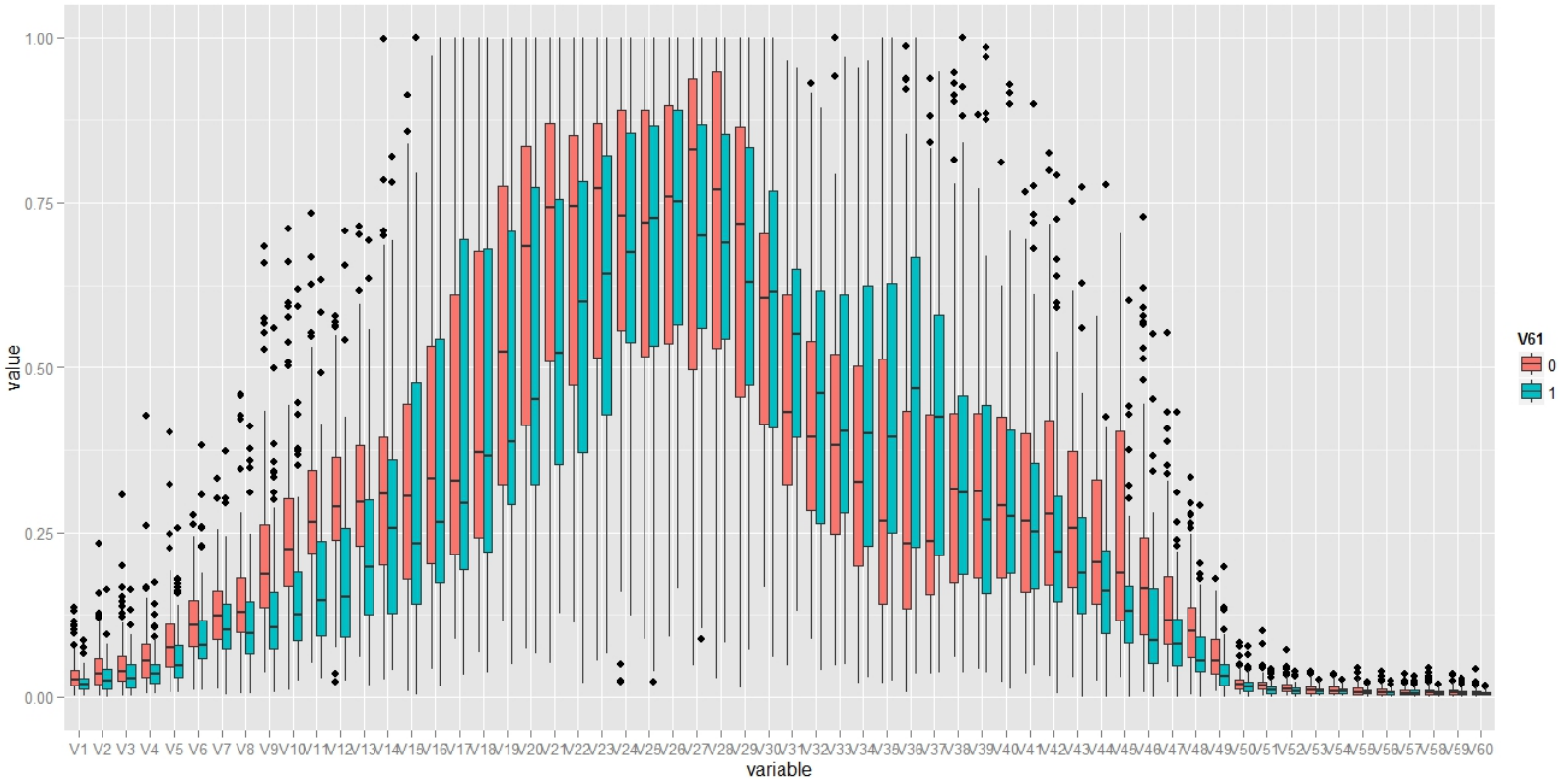


Fig.1 - Répartition des mesures en fonction du matériau, représentée pour les 60 variables

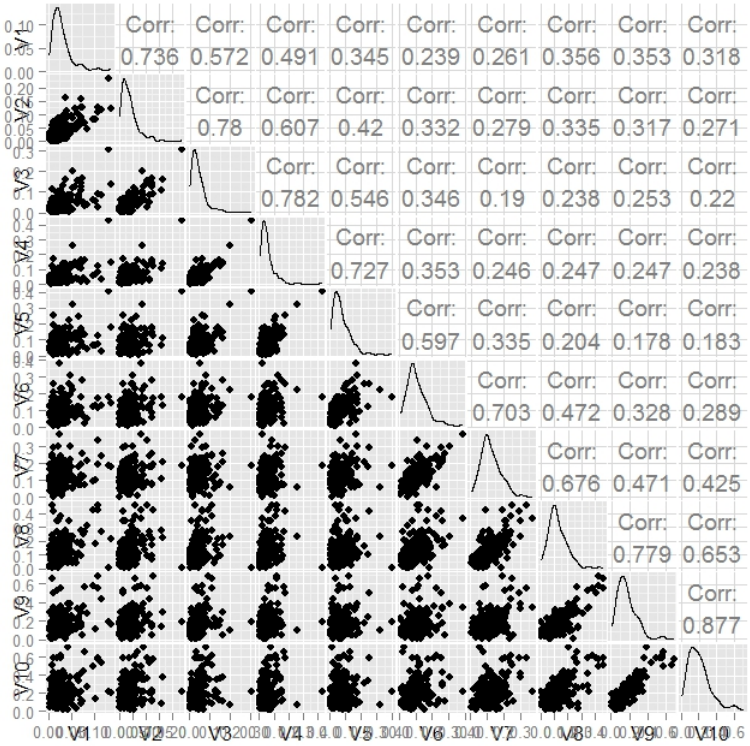


Fig.2 - Graphique de corrélations de variables 2 à 2, représenté pour les 10 premières variables

3.2. Construction du modèle logit

► Étape 1 : Sélection des variables à partir du modèle complet

- **Modèle complet**

```
> res.glm.init=glm(V61~.,family=binomial,data=dat.sonar)
> summary(res.glm.init)
```

```
Warning messages:
1: glm.fit: l'algorithme n'a pas convergé
2: glm.fit: des probabilités ont été ajustées numériquement à 0 ou 1
```

• **Procédure de sélection des variables**

Il existe plusieurs méthodes de sélection automatique de variables sur le logiciel **R**, leur l'objectif est de choisir le meilleur ensemble de variables explicatives.

Du fait qu'on ne peut évaluer le modèle complet puisque l'algorithme de Fisher scoring n'a pas convergé, alors il semble être plus judiscieux d'effectuer une sélection ascendante avec réévaluation du modèle courant plutôt qu'une sélection descendante.
La procédure sur **R** le permettant est: **stepwise** (voir package 'Rcmdr') avec l'option **direction="forward/backward"**, et qui s'éxecute de la manière suivante:

Le modèle de départ est le modèle comprenant une constante, et auquel on a ajouté une variable. A chaque étape de la procédure, on examine à la fois si une nouvelle variable doit être ajoutée selon un seuil d'entrée fixé, et si une des variables déjà incluses doit être éliminée selon un seuil de sortie fixé. Cette méthode permet de retirer du modèle d'éventuelles variables qui seraient devenues moins indispensables du fait de la présence de celles nouvellement introduites. La procédure s'arrête lorsque aucune variable ne peut être rajoutée ou retirée du modèle selon les critères choisis.

• **Critère d'information**

Le critère d'information que nous allons utiliser est le critère d'Akaike,

$$AIC = 2k - 2ln(L)$$

où k est le nombre de paramètres à estimer et L est le maximum de la fonction de vraisemblance du modèle.

• **Modèle sélectionné**

```
> res.glm=stepwise(res.glm.init,direction="forward/backward",criterion="AIC" )
> summary(res.glm)
```

```
Call:
glm(formula = V61 ~ V36 + V45 + V4 + V21 + V51 + V8 + V49 + V50 +
    V1 + V3 + V54 + V23 + V31 + V12 + V30 + V32 + V53 + V7 +
    V16 + V26 + V9 + V58 + V13, family = binomial, data = dat.sonar)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.28942  -0.16492  -0.00004   0.08388   2.20561

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   13.038      3.455   3.774 0.000161 ***
V36           13.254      3.141   4.220 2.44e-05 ***
V45          -22.027      5.716  -3.854 0.000116 ***
V4           -37.113     12.561  -2.955 0.003131 **
V21           -3.290      2.141  -1.537 0.124314
V51         -156.180     53.531  -2.918 0.003528 **
V8            25.915      8.287   3.127 0.001765 **
```

```
V49         -111.746     27.842  -4.014 5.98e-05 ***
V50          241.400     61.808   3.906 9.40e-05 ***
V1           -85.424     23.100  -3.698 0.000217 ***
V3            64.056     19.978   3.206 0.001345 **
V54         -119.596     60.469  -1.978 0.047950 *
V23           -9.562      2.470  -3.871 0.000108 ***
V31           24.834      6.006   4.135 3.56e-05 ***
V12          -25.477      6.553  -3.888 0.000101 ***
V30          -16.202      3.878  -4.178 2.94e-05 ***
V32          -10.963      3.692  -2.970 0.002981 **
V53         -213.721     72.013  -2.968 0.002999 **
V7            18.865      7.608   2.480 0.013153 *
V16            8.506      2.491   3.415 0.000638 ***
V26            3.800      1.819   2.089 0.036734 *
V9           -13.572      5.016  -2.706 0.006810 **
V58         -120.318     69.615  -1.728 0.083931 .
V13            6.378      4.499   1.418 0.156259
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 287.406  on 207  degrees of freedom
Residual deviance:  85.022  on 184  degrees of freedom
AIC: 133.02

Number of Fisher Scoring iterations: 8
```

- **Discussion:**

On parle de succès de prédiction lorsque la probabilité de prédire un objet de type rocheux est supérieure à 0.5, et lorsque que la probabilité de prédire un objet de type métallique est inférieure ou égale à 0.5.

```
      predicted
actual  FALSE  TRUE
     0     101    10
     1      12    85
```

Dans $87,6\%$ des cas, on a réussi à prédire la nature rocheuse de l'objet visé, et dans 91% des cas sa nature métallique.

```
> sum(res.glm$fitted.values[1:97]>.8)/97*100
[1] 76.28866
> sum(res.glm$fitted.values[98:208]<.2)/111*100
[1] 79.27928
```

On peut tout de même discuter de la qualité du modèle, puisque seulement $\sim 80\%$ des valeurs prédictes ont une forte probabilité de correspondre à la vraie valeur, voir **Fig.3**. Ici on parle de forte probabilité lorsque la probabilité de prédire un objet de type rocheux est comprise entre $0,8$ et 1, et lorsque la probabilité de prédire un objet métallique est comprise entre 0 et $0,2$.

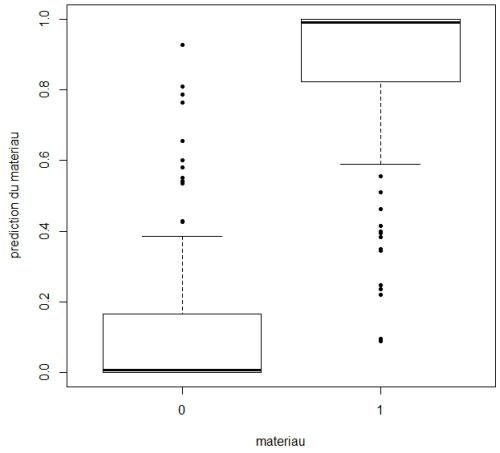


Fig.3 - Répartition des prédictions en fonction du matériau

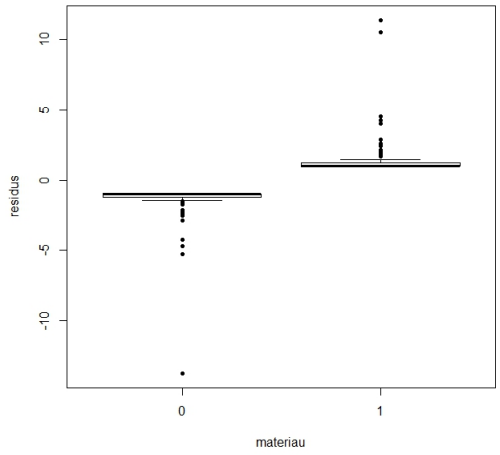


Fig.4 - Répartition des résidus en fonction du matériau

► Étape 2 : Sélection des variables à partir du modèle ajusté

De nombreux indicateurs existent afin d'évaluer la qualité et la robustesse des modèles estimés, leurs rôles sont de détecter les valeurs influentes.

• Effet de levier

On construit la matrice de projection (hat matrix),

$$H = W^{\frac{1}{2}} X \left(X'WX\right)^{-1} X' W^{\frac{1}{2}}$$

relative au produit scalaire de la matrice de 'pondération' de diagonale $W_{ii} = \frac{1}{Var(Y_i)} \left(\frac{\mu_i}{\eta_i}\right)^2$, sur le sous-espace engendré par les variables explicatives.

L'effet de levier consiste à étudier les termes diagonaux H_{ii}, ceux qui sont supérieurs à $\frac{3(p+1)}{n}$ sont considérés comme influents, ajuster les données revient à retirer les observations faites pour ces valeurs.

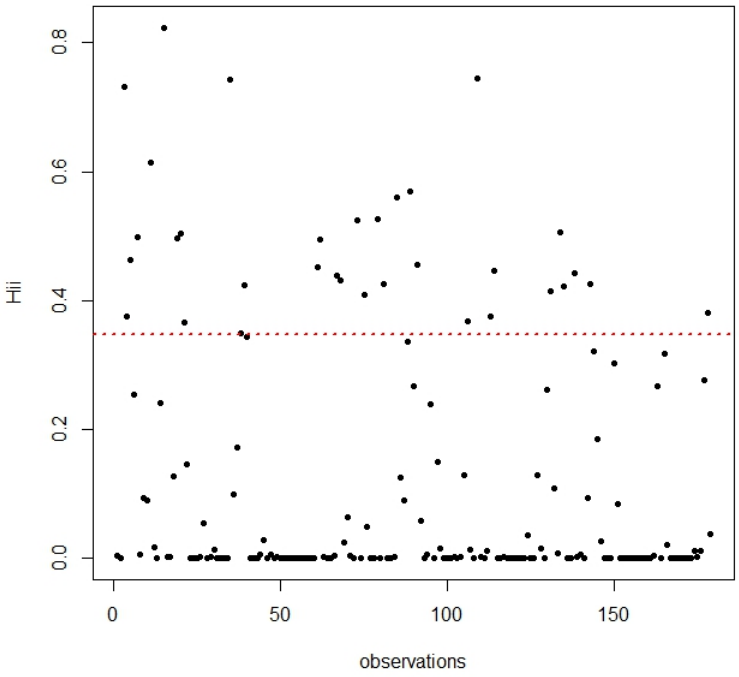


Fig.5 - H_{ii} en fonction des observations

• **Modèle sélectionné après ajustement**

```
> res.glm.init=glm(formula = V61 ~ V36 + V45 + V4 + V21 + V51 + V8 + V49 + V50 +
                   V1 + V3 + V54 + V23 + V31 + V12 + V30 + V32 + V53 + V7 +
                   V16 + V26 + V9 + V58 + V13, family = binomial,
                         data = new_dat.sonar)
> res.glm=stepwise(res.glm.init,direction="forward/backward",criterion="AIC" )
> summary(res.glm)
```

```
Call:
glm(formula = V61 ~ V12 + V45 + V36 + V58 + V23 + V9 + V16 +
    V8 + V51 + V31 + V26 + V50 + V49 + V30 + V53 + V4 + V7 +
    V32 + V13 + V21 + V1, family = binomial, data = new_dat.sonar)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.92799  -0.00009   0.00000   0.00001   2.16389

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    35.359     13.993   2.527  0.01151 *
V12           -97.565     36.285  -2.689  0.00717 **
V45           -82.247     35.017  -2.349  0.01883 *
V36            47.206     19.116   2.469  0.01353 *
V58          -461.952    279.277  -1.654  0.09811 .
V23           -35.230     14.579  -2.416  0.01568 *
V9            -59.465     28.018  -2.122  0.03380 *
V16            39.223     15.707   2.497  0.01252 *
V8             85.806     34.449   2.491  0.01275 *
V51         -1070.165    431.466  -2.480  0.01313 *
V31            77.079     31.046   2.483  0.01304 *
V26            26.056     10.689   2.438  0.01478 *
V50           616.218    251.137   2.454  0.01414 *
V49          -307.491    126.934  -2.422  0.01542 *
V30           -42.780     15.954  -2.681  0.00733 **
V53          -481.317    250.047  -1.925  0.05424 .
V4            -80.096     38.475  -2.082  0.03736 *
V7             55.480     31.753   1.747  0.08059 .
V32           -32.253     15.258  -2.114  0.03453 *
V13            20.465     11.529   1.775  0.07587 .
V21           -11.905      6.583  -1.808  0.07055 .
V1            -85.959     64.356  -1.336  0.18166
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 258.043  on 187  degrees of freedom
Residual deviance:  33.176  on 166  degrees of freedom
AIC: 77.176
```

```
Number of Fisher Scoring iterations: 12
```

- **Discussion:**

```
       predicted
actual FALSE TRUE
     0   101    4
     1     4   79
```

Dans $95,2\%$ des cas, on a réussi à prédire la nature rocheuse de l'objet visé, et dans $96,2\%$ des cas sa nature métallique.

```
> sum(res.glm$fitted.values[1:83]>.8)/83*100
[1] 86.74699
> sum(res.glm$fitted.values[84:188]<.2)/105*100
[1] 90.47619
```

Cette fois-ci, la qualité du modèle est indiscutable puisque suite à l'ajustement des données on a $\sim 90\%$ des valeurs prédictes qui ont une forte probabilité de correspondre à la vraie valeur, voir **Fig.6**. Cependant, pour arriver à ce type de résultat on a dû retirer (par effet de levier) 20 observations parmi les 208 observations de l'échantillon.

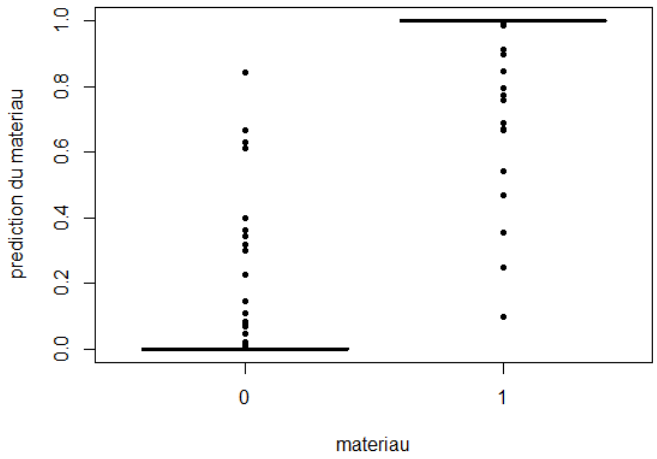


Fig.6 - Répartition des prédictions en fonction du matériau

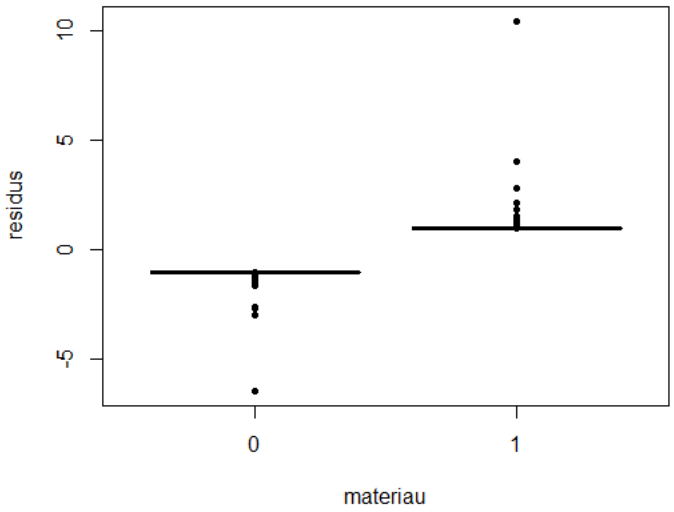


Fig.7 - Répartition des résidus en fonction du matériau

4. ANNEXE

```
###
#
library(Rcmdr)
library(GGally)
library(ggplot2)
library(reshape2)
graphics.off()

###
dat.sonar=read.table("sonar.txt",sep = ",",header=F)
levels(dat.sonar$V61)[1]=0
levels(dat.sonar$V61)[2]=1

### graphique de correlations de variables 2 a 2 pour les 10 premieres variables
x11()
ggpairs(dat.sonar[,1:10])

### affichage de la repartition des mesures en fonction du materiau
df.m = melt(dat.sonar, V61 = "Label")
x11()
ggplot(data = df.m, aes(x=variable, y=value)) + geom_boxplot(aes(fill=V61))

### selection des variables a partir du modele complet
res.glm.init=glm(V61~.,family=binomial,data=dat.sonar)
res.glm=stepwise(res.glm.init, direction="forward/backward", criterion="AIC" )
summary(res.glm)

### table de verite du modele selectionne
table(actual=dat.sonar$V61, predicted=res.glm$fitted.values>.5)

### qualite des valeurs predites en pourcentage
sum(res.glm$fitted.values[1:97]>.8)/97*100
sum(res.glm$fitted.values[98:208]<.2)/111*100

###
x11()
plot(dat.sonar$V61,res.glm$fitted.values, pch=20,
     xlab="materiau",ylab="prediction du materiau")
x11()
plot(dat.sonar$V61,res.glm$residuals, pch=20,
     xlab="materiau",ylab="residus")

### etude des points influents
inf.temp=influence(res.glm)

### affichage des points influents
x11()
plot(inf.temp$hat,pch=20, xlab="observations", ylab="Hii")
```

```
seuil=3*24/208
abline(seuil,0,col="red",lty="dotted",lwd = 2)

### ajustement des donnees
i=which(inf.temp$hat > seuil )
new_dat.sonar=dat.sonar[-i,]

### selection des variables a partir du modele ajuste
res.glm.init=glm(formula = V61 ~ V36 + V45 + V4 + V21 + V51 + V8 + V49 + V50 +
                   V1 + V3 + V54 + V23 + V31 + V12 + V30 + V32 + V53 + V7 +
                   V16 + V26 + V9 + V58 + V13, family = binomial,
                          data = new_dat.sonar)
res.glm=stepwise(res.glm.init, direction="forward/backward", criterion="AIC" )
summary(res.glm)

### table de verite du modele selectionne apres ajustement
table(actual=new_dat.sonar$V61, predicted=res.glm$fitted.values>.5)

### qualite des valeurs predites en pourcentage
names(res.glm$fitted.values)=1:188
sum(res.glm$fitted.values[1:83]>.8)/83*100
sum(res.glm$fitted.values[84:188]<.2)/105*100

###
x11()
plot(new_dat.sonar$V61,res.glm$fitted.values,pch=20,
     xlab="materiau",ylab="prediction du materiau")
x11()
plot(new_dat.sonar$V61,res.glm$residuals,pch=20,
     xlab="materiau",ylab="residus")
```

5. CONCLUSION

En conclusion, cette présentation sommaire de la régression logistique rappelle qu'elle constitue une excellente technique lorsqu'il s'agit de déterminer des prédicteurs d'un phénomène. Bien qu'elle compte certains postulats et qu'elle exige une interprétation rigoureuse, elle s'applique dans une multitude de recherches. La création de modèles à partir de celle-ci exige une réflexion sur la problématique de même qu'une analyse minutieuse des résultats afin de divulguer une explication juste et détaillée du phénomène à l'étude.

6. RÉFÉRENCES

Introduction au modèle linéaire général. Université de Toulouse.
http://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-modlin-mlg.pdf.
L'analyse de régression logistique, Julie Desjardins. Université de Montréal.
http://www.tqmp.org/Content/vol01-1/p035/p035.pdf.