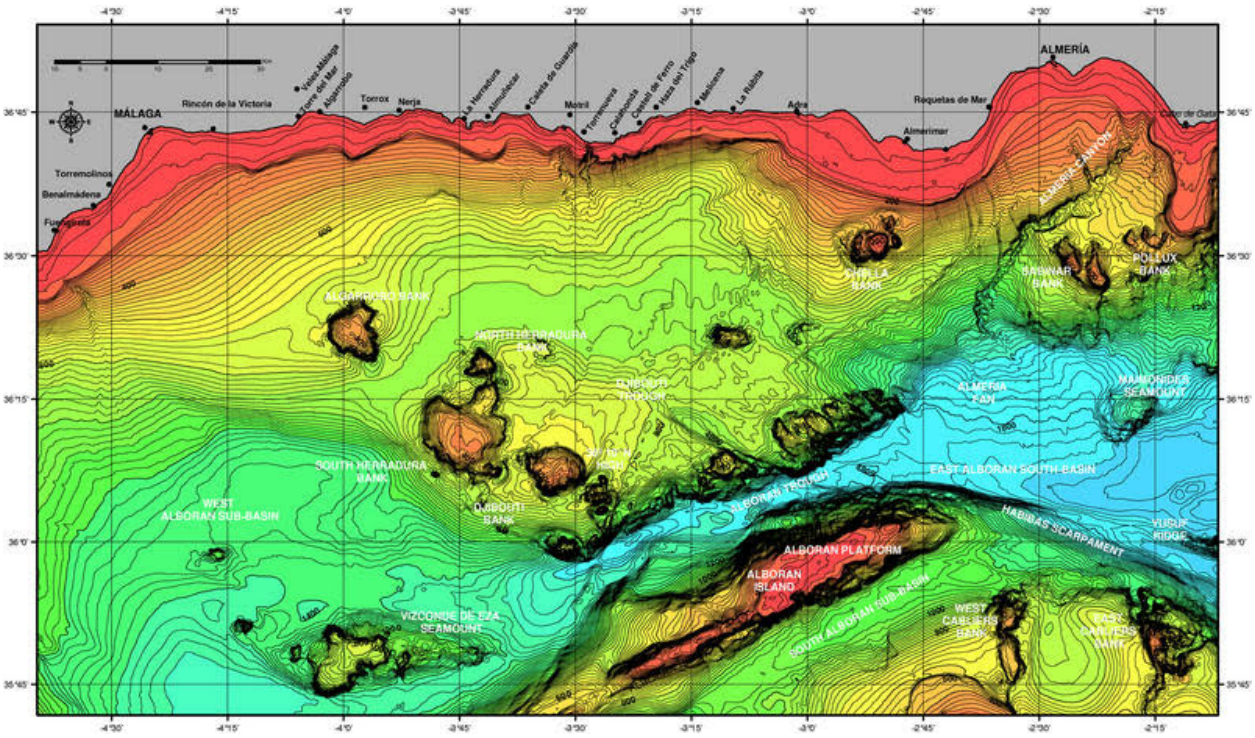
Analyse statistique de données d'aléas marins en mer Méditerranée

Jeremy Uzan et Kevin Faou

Contents

Introduction . . . 2
1. Etat des lieux des données, synthèse du travail effectué l'an passé . . . 2
A. Variables présentes, données manquantes . . . 2
B. Corrélations possibles entre les variables . . . 3
C. Séparation des variables qualitatives et quantitatives . . . 3
D. Synthèse . . . 4
2. Analyse des variables quantitatives et qualitatives . . . 4
A. Régréssion linéaire simple, sélection des variables pertinentes pour prévoir différents Y, "variables ajustée" . . . 4
La surface . . . 4
La surface privé du runout et Hauteur de la cicatrice . . . 4
la log(surface) . . . 5
Le Runout . . . 6
BILAN . . . 7
B. Analyse des variables qualitatives . . . 7
Transformation et sélection du jeu de données . . . 7
Régréssion linéaire simple . . . 7
3. Modèle linéaire généralisé, avec la surface en variable explicative . . . 8
A.Théorie, principe . . . 8
B. Application . . . 9
La log(surface) avec les variables quantitatives . . . 9
GLM de La log(surface) pour les variables qualitatives . . . 9
Conclusion et ouverture de recherche . . . 10
Références . . . 11



Introduction

Nous avons récupéré plusieurs bases de données de MTD (Mass transport deposit) provenant du bassin d'Alboran méridional en méditérannée. En nous appuyant sur les outils statistiques et les travaux de nos prédécesseurs Balint Gersey et Nicolas Prost, nous proposons ici une analyse des données collectées par les géologues de l'ISTEP (Sarah Lafuerza, Elia d'Acremont et Alain Rabaute) de l'Université Pierre et Marie-Curie. Ce TER est mené sous la direction de Maud Thomas.

Une première étude approfondie a été faite par nos camarades l'an passé. Nous nous sommes penchés sur la question de trouver un modèle de régression linéaire qui correspond assez bien aux données préparant ainsi à de l'éventuelle prédiction. En nous appuyant sur les résultats de nos prédécesseurs, nous nous sommes concentrés sur les modèles qui sont les plus pertinents à savoir le log(surface) et la surface en variable explicative.

Plan d'étude

Nous allons tout d'abord étudier le jeu de données, et extraire une synthèse du Data Mining et des analyses faites l'année précédente. Cela nous permettra alors de nous concentrer sur la recherche d'un modèle pertinent pour expliquer la variable surface. Dans un premier temps, on utilisera les modèles de régression linéaires simples. Dans un second temps, nous tenterons de proposer un modèle linéaire généralisé.

1. Etat des lieux des données, synthèse du travail effectué l'an passé

A. Variables présentes, données manquantes

Nous avons 20 variables :

- MTD_ID Identifiant du MTD
- LAYER Age chronologique relatif des MTD par rapport aux marqueurs sismiques (en ms)
- AGE_SISM Age relatif par rapport aux marqueurs isotopiques (MIS)
- MTD_CHRON Numéro des MTD dans l'ordre chronologique (1 est le plus ancien)

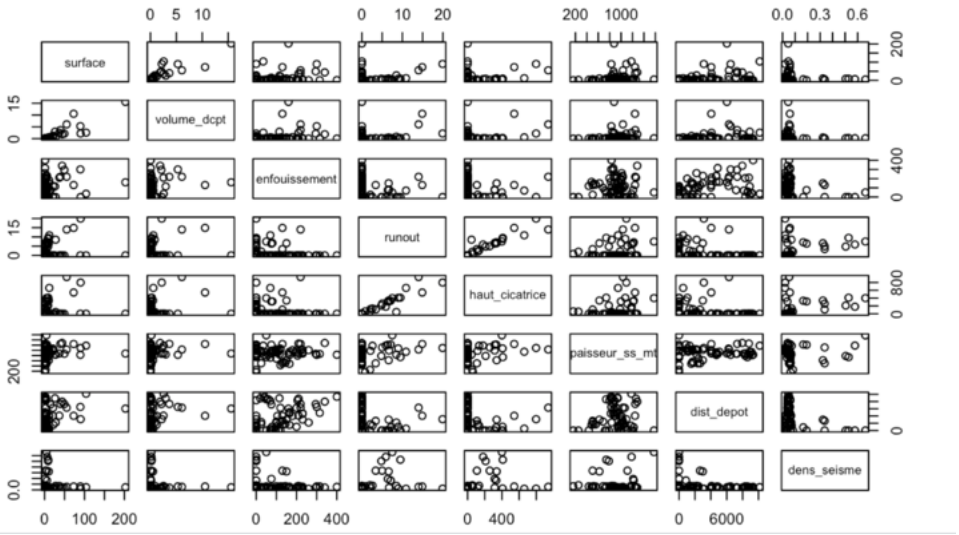
- AGE Age relatif par rapport aux marqueurs sismiques
- SURF_KM3 Surface du MTD (= le dépôt) en km2
- VOL_KM3 Volume calculé du MTD en km3
- PORO Porosité mesurée sur sédiments
- DVOL_KM3 Volume décompacté du MTD en km3
- BURIAL_MS Enfouissement en millisecondes
- RUNOUT_KM Run-out du MTD en KM à partir de la cicatrice d'arrachement observée
- SCHGHT_M Hauteur de la cicatrice d'arrachement en mètres (quand il est possible d'identifier celle-ci)
- SCSLOPE_D Pente de la cicatrice en degrés
- MULTI MTD multiphasé ou non
- MTD_TYPE Type du MTD : DF: debris flow - coulée de débris; S: slide - glissement en masse
- EROD_BASE Présence d'une base érosive ou non
- THCK_PQ_M Epaisseur du Plio-Quaternaire en mètres (sous le MTD)
- RNOUT_KM Run-out maximum du MTD en km à partir de la cicatrice d'arrachement supposée
- DIST_CONT_M Distance du barycentre du MTD par rapport à un dépôt contouritique proche
- EQ_DENS Densité de séismes à l'aplomb du MTD (nombre de séismes par km2, pondération par la magnitude, catalogue 1970-2017, seuls les séismes 0<Mag<=4 et Depth < 40km sont conservés

Nous supprimons la porosité car elle est proportionnelle à l'enfouissement.

Le jeu de données comprend deux tableaux de respectivement 28 et 38 MTD avec les mêmes variables.

B. Corrélations possibles entre les variables

Nous commençons par manipuler les variables et observer des premières corrélations entre variables quantitatives.



VOLUME corrélé à la surface
RUNOUT corrélé à la hauteur de la cicatrice
AGE corrélé à l'enfouissement (négativement !) et la porosité
L'épaisseur corrélée à 1) dist depot (0,7)

Surf en kmˆ3: la variable la plus corrélée à cette variable est le Vol_kmˆ3 et le DVOL_km$^{3\ \mathrm{RUNOUT_km}}$3: les deux variables les plus corrélées sont SCSLOPE_D (Pente de la cicatrice en degrés) et SCHGT_M (Hauteur de la cicatrice d'arrachement en mètres (quand il est possible d'identifier celle-ci))

C. Séparation des variables qualitatives et quantitatives

Nous avons 13 variables quantitatives et 7 variables qualitatives.

D. Synthèse

- Le volume décompacté et le volume sont proportionnels, et il en est de même pour la porosité. Ce sont donc des variables qu'on retire de l'analyse.
- Il apparait souvent la valeur 0 sur les variables Runout, Hauteur, Pente. On suppose que ces valeurs sont une absence de résultats.
- Dans plusieurs résultats du RUNOUTmax apparaît "<", c'est une variable qu'on choisit d'écarter de l'analyse, malgré son intérêt potentiel. On garde néanmoins la variable Runout
- la variable qualitative MULTI possède une grande majorité de N (95%) dans les deux jeux de données. On choisit de la garder.
- Nos prédécesseurs ont conjecturé que la loi log normale concorde relativement bien aux jeux de données, notamment pour la surface.

2. Analyse des variables quantitatives et qualitatives

A. Régréssion linéaire simple, sélection des variables pertinentes pour prévoir différents Y, "variables ajustée"

Rappel théorique
On tente de modéliser une variable y ("variable à expliquer"), comme une fonction d'autres variables x=(x1,x2,x3,...,xi)', "variables explicatives" sous forme de vecteur colonne de taille le nombre de variables explicatives (ici i)
Donc y=g(x) et le but est de retrouver la fonction g.
Dans le modèle linéaire simple, g est affine, donc il existe un vecteur colonne t=(t1,t2,t3,..,ti)' de taille le nombre de variables explicatives tel que
y=$\sum_{i=1}^{n} t_i x_i$+ (où epsilon est l'erreur de mesure) On dispose d'un échantillon de taille n et de (i+1)-uplets (x,y) et l'objectif est d'estimer le paramètre t

Etant donné que nous disposons de deux jeux de données, nous effectuons une première analyse avec le premier jeu, une autre avec le deuxieme, et une autre en groupant les deux jeux de données

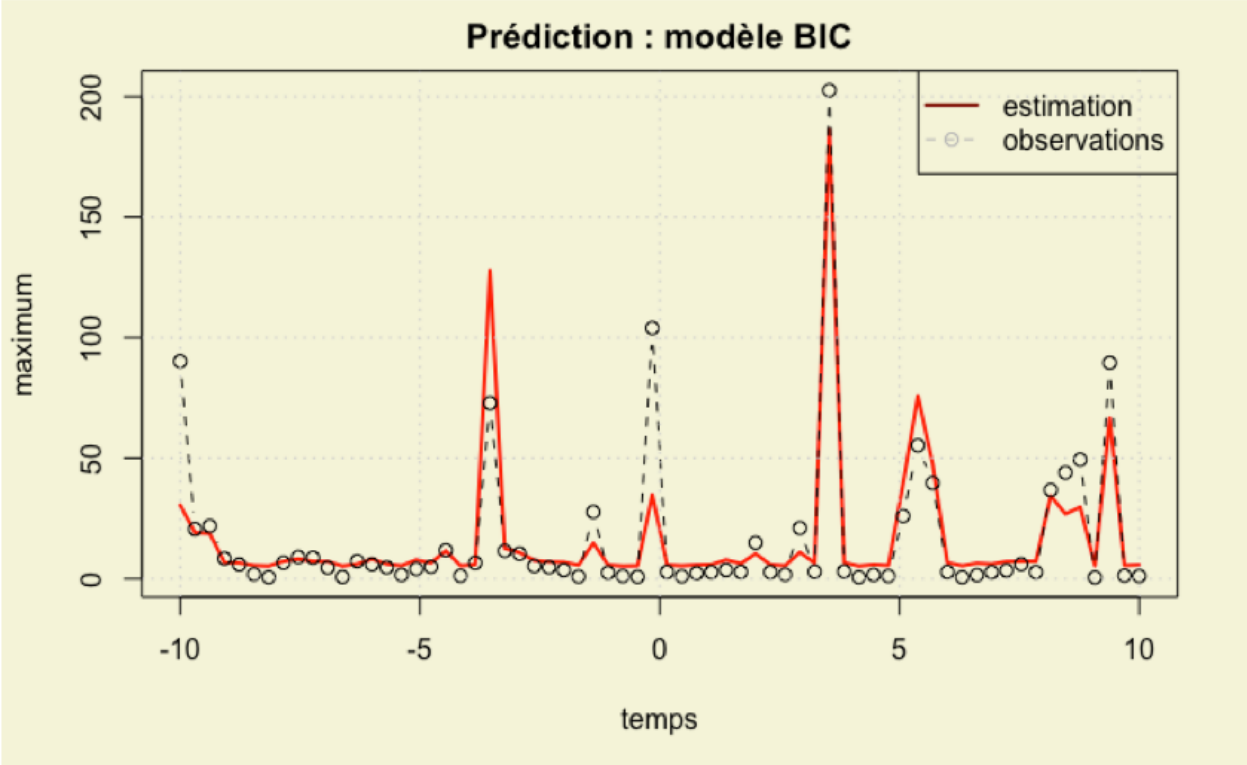
Le test de Fisher global permet de tester l'apport global et conjoint de l'ensemble des variables explicatives présentes dans le modèle pour "expliquer" les variations de Y . L'hypothèse nulle ici est H0:1==p=0 (l'ensemble des p variables explicatives n'apporte pas une information utile pour la prédiction de Y (qui est dans la suite la surface, puis le renout, puis le volume) sous le modèle linéaire). L'assertion d'intérêt est
H1: au moins l'un des coefficients j est significativement différent de zéro (au moins une des variables explicatives et associée à Y après ajustement sur les autres variables explicatives).

La surface

Que se soit avec le critère AIC, BIC ou CP, on obtient toujours la même conclusion: un modèle à une seule variable : le volume décompacté.

La surface privé du runout et Hauteur de la cicatrice

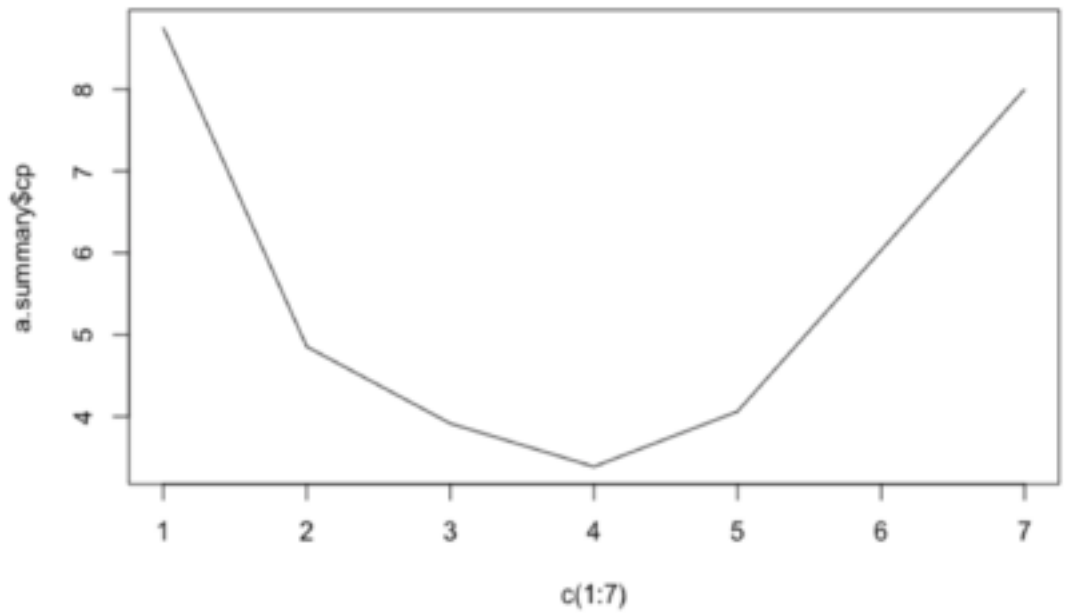
On obtient le même résultat: volume_dcpt



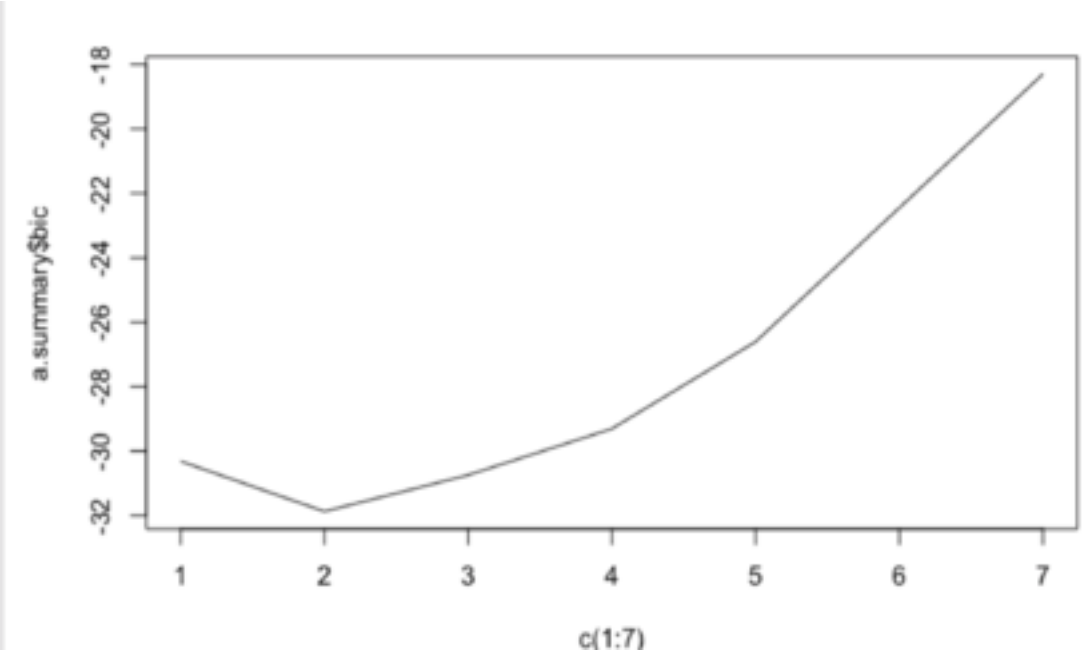
Le modèle avec uniquement le volume_dcpt semble bien adapté aux données de la surface, sauf sur quelques points extrêmes.

la log(surface)

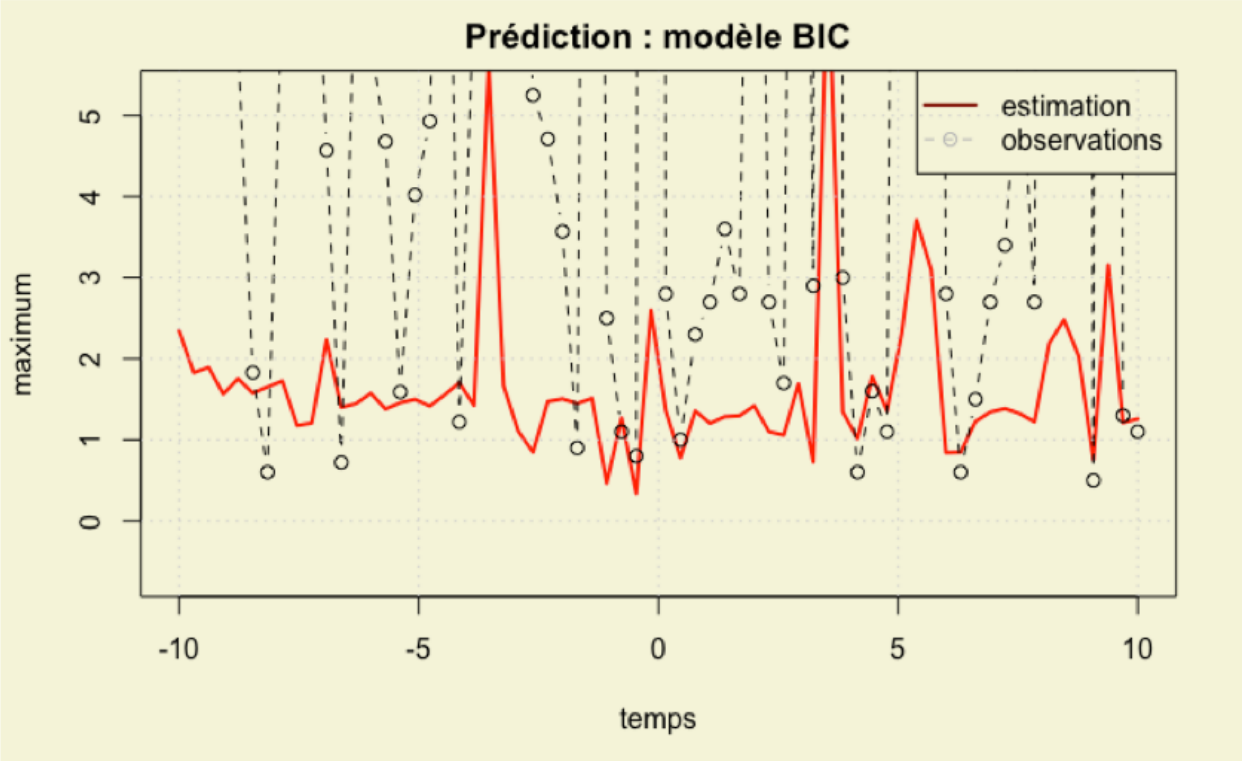
Utilisation de la fonction STEP BILAN 1/ le modèle avec les variables volume décompacté, runout, dist_depot et épaisseur est pertinent pour modéliser les valeurs prises par variable surface qui représente le dépot en km^2



Bilan2/ avec le critère BIC, on a seulement un modèle à 2 variables: volume_dcpt et epaisseur sous mtd



on regarde la prédiction :



Le résultat est moins probant que précédemment.

Le Runout

On effectue le même travail en posant le Runout comme variable à expliquer. Cette fois-ci, le modèle le

plus pertinent est un modèle à 5 variables avec notamment l'age (MTD chron), la hauteur de la cicatrice, le volume et la pente de la cicatrice.

BILAN

On se concentre désormais sur le modèle avec la surface en variable réponse Y et tentons de reproduire un modèle qui correspond le mieux. En effet, le modèle linéaire simple suppose que les vairbales explicatives (comme le volume, la porosité) sont non aléatoires et mesurables sans erreur. De plus, l'erreur de mesure ne suit pas nécessairement une distribution normale et la variance n'est pas tout le temps constante. Enfin, si la variable Y (ici la surface) ne suit pas la distribution d'une loi normale, mais une autre famille de la classe exponentielle, on utilise les modèles linéaires généralisés.

B. Analyse des variables qualitatives

Transformation et sélection du jeu de données

Pour les variables qualitatives, nous choisissons de séléctionner les données du tableau qui contiennent le plus d'information. En effet, on remarque que l'age chronologique est plus facile à lire dans le tableau Albojer53.xlsx. On a dans ce tableau différents Age chronologique (relatif aux MTD par rapport aux marqueurs sismiques en miliseconde) très variables avec des données facilement comparables. Dans l'autre tableau, on observe les résultats "MTD superficiel", ou "surf" ou même des tranche 100-150, ce qui est plus difficile à étudier.

Régréssion linéaire simple

Nous allons mettre en variable réponse la surface pour rester cohérent avec l'analyse quantitative et ainsi permettre de regrouper les résultats de l'analyse qualitative et quantitative. Nous cherchons le meilleur modèle avec des facteurs qualitatifs qui permette d'expliquer la donnée surface.

L'anova nous permet de constater que la surface est linéairement liée au caractère multiphasé ou non des MTD

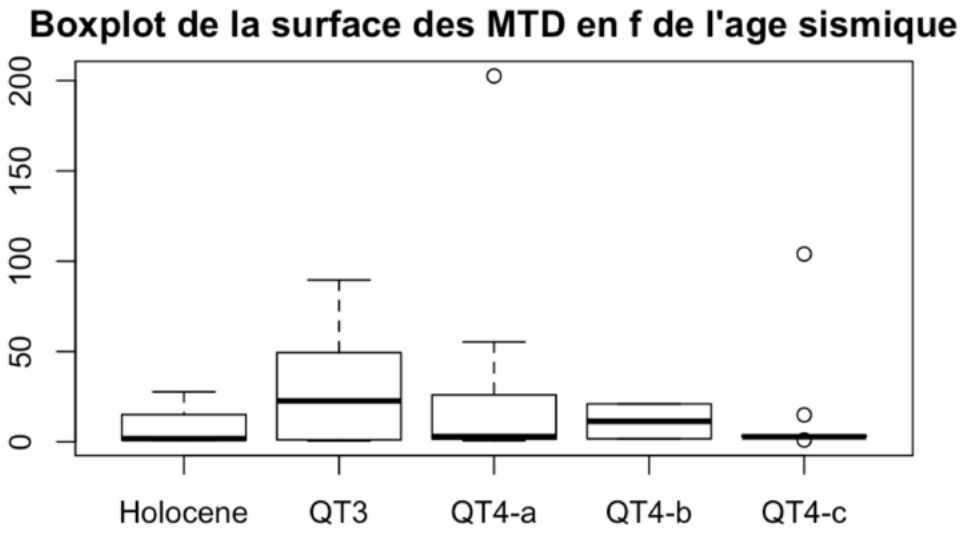
Une deuxieme analyse de la variance nous permet de constater que la log(surface) est encore plus linéairement liée au caractère multiphasé ou non des MTD

```
> reg_multi<-lm(log(surface)~multi,data=geo)
> anova(reg_multi)
Analysis of Variance Table

Response: log(surface)
          Df Sum Sq Mean Sq F value    Pr(>F)
multi      1 37.342  37.342  20.921 6.086e-05 ***
Residuals 34 60.687   1.785
```

Etude qualitative "BIS":L'age sismique et la surface

On peut remarquer que les MTD faisant parties de la classe d'âge QT3 ont une surface plus étalée, et relativement plus élevée en moyenne comparée à la surface des MTD dont les classes d'âge sont QT4-C ou Halocene



3. Modèle linéaire généralisé, avec la surface en variable explicative

Ces modèles ont été introduit par Nelder et Wedderburn en 1972.

A.Théorie, principe

Les modèles linéaires généralisés sont une extension des modèles linéaires classiques. Le tableau proposé dans le cours de l'ISFA de Lyon permet d'observer rapidement les différence entre ces deux modèles:

On suppose toujours que les observations sont indépendantes (ou au moins non corréles). Mais la variance n'est plus constante et la normalité n'est plus nécessaire.

Les observations $Y_1, \ldots, Y_n$ sont indépendantes et suivent une loi qui appartient à la famille exponentielle (Normale, Poisson, Binomiale, Gamma, etc) Ceci est la première des trois composantes des modèles linéaires généralisés La deuxième est la composante systématique, elle attribue à chaque obeservation un prédicteur linéaire

$$\eta_i = \beta_0 + \sum_{j=1}^{p-1} x_{ij}\beta_j$$

Enfin, l'espérance i de Yi (attribut de la première composante) est liée au prédicteur linéaire i par une "fonction de lien" g. Le troisième composant fait donc le lien entre les deux premier.

$$\eta_i = g(\mu_i) = \beta_0 + \sum_{j=1}^{p-1} x_{ij}\,\beta_j$$

On peut donc obtenir en transformant par la fonction de lien. l'espérance des Yi est donc combinaison linéaire des paramètres par une fonction g. On obtient les prédictions en faisant l'inverse de la fonction de lien, g1, qu'on appelle la fonction "de réponse". Les fonctions de lien les plus fréquemment utilisées sont :

log	$g(\mu_i) = \log \mu_i$
logit	$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$
probit	$g(\mu_i) = \Phi^{-1}(\mu_i)$, où $\Phi(.)$ est la fdc $\mathcal{N}(0,1)$
complementary log-log	$g(\mu_i) = \log\left(-\log(1-\mu_i)\right)$
log-log	$g(\mu_i) = \log(-\log(\mu_i))$

B. Application

La log(surface) avec les variables quantitatives

Nous essayons différentes familles de lois exponentielles pour le modèle GLM, mais aucune ne permet d'extraire un ou plusieurs modèles viables, à cause des la nature des données (beaucoup de 0) Les lois discrètes (binomiale, poisson) ne correspondent pas non plus, puisque les résultats des variables ne sont pas des valeurs entières. Seule la loi gaussienne peut fonctionner, mais cela revient à effectuer une régression linéaire simple.

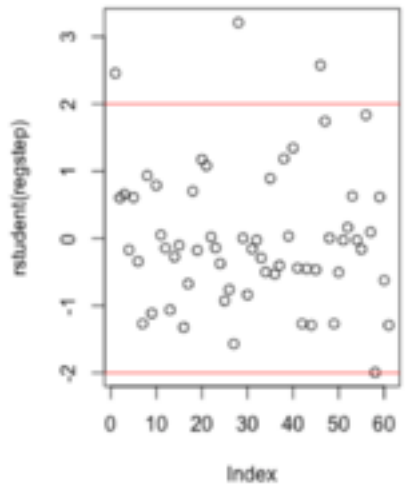
On opte pour le modèle linéaire gaussien classique et on essaye le GLM avec les variables qualitatives.

GLM de La log(surface) pour les variables qualitatives

On ne garde que les donnees qualitatives et complete. On met les variables suivantes en factor: age_sism multi type ero_base pock On effectue une regression lineaire

Puis on effectue un test pour le GLM avec la fonction Gamma.

DEFINITION: On appelle deviance d'un modele lineaire generalisé la valeur D=-2(lm - ls), où ls est la log vraisemblance du modèle parfaitement ajusté et lm la log vraisemblance du modèle considéré On peut donc voir la déviance comme une extension de la somme des carrés des résidus de modèle linéaire gaussien d'ailleurs si l'on applique glm(Y A+B+C,family="gaussian") on obtient les memes resultats qu'un modele lineaire gaussien



Les résidus sont pour la majorité en dessous du seuil de rejet.

GLM:

- On ne peut pas mettre age sismique car l'algo produit des NANs.

- les tests donnent que multi et type sont significatifs tandis que pock ne l'est pas

- On compare les modèle GLM : surface avec les variable multi et type contre surface avec les variables multi, type ET pock

```
> anova(regstepglm,reglm_gamma,test="Chisq" )
Analysis of Deviance Table

Model 1: surface ~ multi + type
Model 2: surface ~ multi + pock + type
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        58     102.88
2        57     102.86  1 0.019599   0.9359
> print(qchisq(.95,1))
[1] 3.841459
```

Le quantile d'une khi2 à 1 degré de liberté est plus grand que la déviance; ainsi on préfère le modèle sans le facteur pock qui n'est pas significatif

```
Analysis of Deviance Table

Model 1: surface ~ multi
Model 2: surface ~ multi + type
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        59     112.73
2        58     102.88  1   9.8497  0.06854 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> print(qchisq(.95,1))# <deviance donc reglsteplglm meilleur
[1] 3.841459
```

Conclusions

Si l'on veut proposer un modèle linéaire généralisé ou la famille de loi est une gamma, le meilleur en terme de deviance et perte de degrés de liberté serait le modèle composé des facteurs type et multi

Conclusion et ouverture de recherche

Les travaux précédents et l'analyse effectué dans ce projet ont permis d'extraire plusieurs conclusions. La variable surface est manifestement une variable REPONSE et la variable du volume_dcpt permet d'estimer (avec une certaine erreur) ladite variable surface.

D'autres variables réponses comme le Runout peuvent être expliquées par les variables volume, la hauteur et la pente de la cicatrice.

Il est par ailleurs important de signaler l'importance des variables comme l'age, la surface, le volume_dcpt comparées à d'autres variables comme pockmark ou encore type, qui apparaissent plus isolée, évoluant "indépendamment des autres"

Nos recherches nous pourrions essayer de tester le modèle avec une variable qualitative (multi) et une variable quantitative:

log(surface)=0+1volume_dcpt+2multi+epsilon.

Ouverture: Utiliser la fonction CUT En utilisant la fonction Cut, on pourrait transformer les variables quantitatives en qualitative (en créer des groupes) et tester le modèle avec la variable volume_dcpt et la variable multi.

Références

- *Rapport : Analyse de données de glissements de terrain sous-marins* Balint Gersey, Nicolas Prost
- *Rapport : Étude des glissements gravitaires sous-marins dans le bassin d'Alboran méridional* Martin Jollivet-Castelot
- *Étude des glissements gravitaires sous-marins dans le bassin d'Alboran méridional*
- *Generalized Linear Models* P. McCullagh and J.A. Nelder
- *Régression avec R* (Pierre-André Cornillon et Eric Matzner-Lober) (*langage Rstudio*)
- Cours :*Statistique mathématique* Arnaud Guyader (*cours théorique de statistique*)
- *Submarine landslides of the Mediterranean Sea: Trigger mechanisms, dynamics, and frequency-magnitude distribution* Roger Urgeles and Angelo Camerlenghi
- *An Introduction to generalized linear model* Annette J.Dobson
- Cours: *Modèles linéaires & GLMs analyse Logit & régression de Poisson Analyse d'un portefeuille d'assurance Algorithm IRWLS avec R* Julien Tomas