

BE-TP R de Statistiques

Emmanuel Rachelson and Matthieu Vignes

8 novembre 2013, SupAero - ISAE

1 Analyses descriptives, tests et modélisation linéaire pour identifier le meilleur protocole d'une culture de bactéries

1.1 Motivation, présentation des données

Le but de ce TP va être d'analyser un jeu de données portant sur une culture de bactérie. Un article (Binnie, J. Stat. Educ., 2004) qui parle du sujet peut être consulté là :

<http://www.amstat.org/publications/jse/v12n2/datasets.binnie.html>.

L'histoire résumée est la suivante : le staphylocoque dor est une bactérie pathogène responsable d'infections nosocomiales. En cas d'infection post-opératoire, une culture en labo est réalisée pour préciser le diagnostic de l'infection et prescrire le bon antibiotique (qui peut être très cher). Le jeu de données comprend les comptages de 5 souches mises en culture dans différentes conditions. On suspecte en effet que les conditions utilisées traditionnellement ne sont pas optimales. Pour chaque souche, on a utilisé chacune des combinaisons des niveaux des facteurs suivants :

- le temps (*Times*) à 24 et 48h,
- la température (*Temp*) à 27, 35 et 43°C et
- la concentration du gel en nutriment (*Conc*): 0.6, 0.8, 1.0, 1.2 et 1.4%.

On va essayer de répondre aux questions suivantes :

- la souche influence-t-elle la concentration ?
- Les facteurs (*Times*, *Temp*, *Conc*) ont-ils un effet ?
- En fixant le temps à 24h, construire le "meilleur" modèle de prévision et en déduire les valeurs optimales des autres paramètres.

Télécharger les données depuis :
<http://www.amstat.org/publications/jse/datasets/Tryptone.dat.txt>.
 Réorganiser les données pour arriver à un `data.frame` de 150 lignes dont les colonnes seront : le type de souche (**Souche**), le temps, la température, la concentration de nutriment et le comptage des *S. aureus*.
 Aide R : on pourra utiliser la fonction `rbind()`.

1.2 Exploration

Exploration unidimensionnelle

it-onconfianceÃ lascience

Avec les commandes `hist()` et `boxplot()`, regarder la variable de dénombrement des bactéries et tester l'intérêt d'une éventuelle transformation des données.

Que fournit la commande suivante : `t.test(Count, conf.level=0.95)` ?

Vérifier les propriétés de la distribution de la variable et de sa transformation ; sont-elles gaussiennes ? (`qqnorm()`, `qqline()`, `shapiro.test()`)

Que dire de la transformation ?

Exploration bidimensionnelle

Croiser les différents facteurs/variables explicatifs avec le dénombrement ; que dire de l'influence possible de ces facteurs ? De la linéarité de l'effet ?

Exemple de commande utile : `boxplot(Count~Time)`.

Commenter.

1.3 Influence du type de souche

On va ici mettre en œuvre une ANOVA et en vérifier les conditions d'applications (que sont-elles ?).

Commenter le résultat de `barlett.test(Count~Souche)`.

Puis estimer le modèle, par exemple avec `res.anova.bact <- aov(Count~Souche)`.

Quelles informations dans le `summary(res.anova.bact)` ?

Étudier le comportement des résidus (`qqnorm()`, `qqline()`, `shapiro.test()`)

Le test est-il pertinent ? Quelle est sa significativité ? Peut-on conclure ?

`kruskal.test(Sount~ Souche)` est-il une bonne alternative ?

1.4 Influence des autres facteurs

On va ici regrouper les échantillons ayant des types de souches différents. On aurait pu tout tester en une seule fois par une analyse de covariance mais nous commençons par des choses simples. Nous allons réaliser plusieurs tests (notion de tests multiples) sur le même échantillon. Que pensez vous

du risque de première espèce global ? Pouvez vous trouver une correction à cet effet de tests multiples (*e.g.* essayer de comprendre une correction de Bonferroni ou du FDR).

Les facteurs autres que la souche sont quantitatifs, un modèle de régression (linéaire encore une fois pour faire simple au début puis quadratique) est plus approprié en terme de puissance qu'une ANOVA qui estime davantage de paramètres.

Modèle linéaire

La commande `res1.reg.bact <- lm(Count~Time+Temp+Conc)` fait presque tout le boulot.

A nouveau, regarder la normalité des résidus. Puis étudier le nuage des résidus :

```
res.student.bact <- rstudent(res1.reg.bact)
yhat <- res1.reg.bact$fitted.values
plot(res.student.bact~yhat, ylab="residus")
abline(h=c(-2,0,2),lty=c(2,1,2))
```

Pour étudier les points influents, expliquer ce que font les commandes :

```
cook <- cooks.distance(res1.reg.bact)
plot(cook~yhat,ylab="distance de Cook")
abline(h=c(0,1),lty=c(1,2))
```

Que conclure sur la validité du modèle ? Avec le `summary(res1.reg.bact)`, que dire sur l'influence des facteurs ? De la qualité de l'ajustement ? Des qualités d'éventuelles prévisions ?

Modèle quadratique

On va créer un `data.frame` qui contient les données des variables, leur carré et les variables d'interactions (les produits).

Puis on va estimer le modèle plus complet que celui de la sous-section du modèle linéaires avec une instruction du type :

```
res2.reg.bact <- lm(Count~., data=data.quadrat), où data.quadrat
```

est justement la `data.frame` qui vient d'être créé.

Etudier à nouveau la normalité des résidus, le nuage de points et regarder les points influents comme pour le modèle linéaire.

Que dire du coefficient de détermination par rapport au modèle précédent ? Et de la forme du nuage des résidus ? Quelle conclusion sur la validité du modèle ?

Modèle après transformation des données

On va essayer une transformation des données pour pallier le problème que nous venons de rencontrer. On va inclure les mêmes données que

précédemment mais le comptage (variable à expliquer) est remplacé par sa racine carrée. Le `data.frame` correspondant est nommé `data.3`.

Comme à la sous-section précédente, on pourra estimer le modèle, regarder la normalité des résidus (`shapiro.test(res3.reg.bact$residuals)`) et étudier le nuage de ces résidus.

Commenter ces résultats (normalité, forme) par rapport au modèle précédent. Que dire de la qualité de l'ajustement du modèle ? De l'influence des interactions ?

1.5 Conditions optimales

Que calcule la fonction suivante :

```
press=function(model) {  
  h=influence(model)$hat  
  e=influence(model)$wt.res  
  n=length(e)  
  sum((e/(1-h))^2)/n  
} ?
```

On estime successivement les 3 modèles suivants pour les comparer :

```
res3.reg.bact <- lm(sqrt.Count.~Time+Temp+Conc+Conc.2+Temp.2+  
+ Temp...Conc+Temp...Time+Time...Conc,data=data.3)  
summary(res3.reg.bact) (déjà vu...)  
res4.reg.bact <- lm(sqrt.Count.~Time+Temp+Conc+Conc.2+Temp.2+  
+ Temp...Conc+Temp...Time,data=data.3)  
summary(res4.reg.bact)  
press(res4.reg.bact)  
res5.reg.bact <- lm(sqrt.Count.~Time+Temp+Conc+Conc.2+Temp.2+  
+ Temp...Time,data=data.3)  
summary(res5.reg.bact)  
press(res5.reg.bact)
```

Les commenter en essayant de motiver leurs choix.

Retenir le “meilleur” modèle de prévision ; c’est la *surface de réponse* du problème d’optimisation. Expliquer ce concept.

En posant `Time=24` dans ce modèle, estimer le comptage du nombre de bactéries en fonction de la température et de la concentration en tryptone.

Pour quel couple de valeurs (température et concentration) maximise-t-on ce comptage ?

2 Fait-on confiance à la science ?

1. Importer le fichier `sciences.csv`. Ce fichier contient des données issues d’une enquête réalisée en 1993 auprès de $n = 871$ sondés sur leur avis sur le rôle de la science. Ils ont dû répondre s’ils sont d’accord ou non avec les déclarations suivantes :

- A : nous croyons trop souvent en la science et pas assez à l'intuition (aux sentiments et aux croyances).
- B : globalement les sciences font plus de mal que de bien.
- C : tous les changements que les hommes font à la nature, aussi scientifiques soient-ils, sont susceptibles d'empirer les choses.
- D : la science va résoudre les problèmes d'environnement sans affecter notre façon de vivre.

Il y a cinq réponses possibles : 1 tout à fait d'accord; 2 d'accord; 3 pas d'avis; 4 pas d'accord; 5 absolument pas d'accord. Puis des questions informationnelles leur ont été posées :

- Sexe : Homme (sex1), Femme (sex2)
- Age (6 classes) : 16-24 (age1), 25-34, 35-44, 45-54 ,55-64 ,65 et plus (age6)
- Niveau d'éducation (6 niveaux) : enseignement primaire (edu1), niveau collège, niveau lycée, niveau Bac + 2, niveau Bac + 5, niveau Bac + 8 (edu6)

Le but de l'étude est de caractériser, de décrire la population de cette enquête, de faire une typologie des sondés en fonction de leur attitude vis-à-vis des sciences.

2. Réaliser et commenter l'analyse du jeu de données `sciences.csv`. Tous les coups sont permis!
3. Interpréter finement la proximité des modalités "C5" et "B5".
4. Y-a-t'il un lien entre le niveau d'éducation et le comportement face à la science ?
5. Comment sont positionnées les modalités Homme et Femme ?
6. Comment peut-on faire pour étudier le comportement des femmes qui ont un haut niveau d'éducation et des hommes qui ont un haut niveau d'éducation? Que concluez-vous?
7. Comment interpréter la distance entre deux individus ? Quelle distance est utilisée en AFCM ?
8. Comment peut-on expliquer le faible rapport de corrélation entre la quatrième variable D et les deux premières dimensions? Ou comment peut-on interpréter la position des modalités de la variables D?
9. Peut-on avoir un test de significativité des contributions des variables ?