Scénario: Optimisation d'une culture bactérienne

Résumé

*Initiation à la pratique des techniques élémentaires de statistique par l'étude d'un jeu de données à l'aide du logiciel R. L'objectif est de décrire, modéliser, pour optimiser les conditions de culture de souches de bactéries (*Staphylococcus aureus*) : description élémentaire, estimation, tests, régression linéaire, analyse de variance (ANOVA), analyse en composantes principales, régression multiple. Si nécessaire, un tutoriel de démarrage avec R est disponible.*

1 Introduction

Dans un article pédagogique, Neil Binnie (2004) [1] détaille le problème de la façon suivante :

> *A person may have a boil or an infected wound from an operation. If it is not responding to antibiotics, they will often be required to supply a swab from the wound for a medical laboratory to investigate. A sterile cotton bud swab is wiped across the wound and wetted with the pus that contains myriads of bacteria cells. The cotton bud is then swirled in a salt enrichment broth and incubated for a period of time. This broth may then be diluted and a drop wiped across the surface of a solid nutritive medium so that bacteria are transferred to the medium surface. The laboratory cultures the sample on a variety of media one of which is suitable for the growth of Methicillin resistant Staphylococcus aureus (MRSA). The aim is to culture the offending bacterium quickly so that it can be identified and susceptibility testing performed so that a suitable antibiotic can be used.*
>
> *MRSA is a bacterium, whose strains are resistant to penicillin and sensitive only to the expensive antibiotic vancomycin. This bacterium is cultured in a 1.5% (by weight) sodium chloride enrichment broth. The salt in the broth inhibits the growth of most normal flora that may be present at the site from which the specimen was collected. This ensures that it will be easier to identify the Staphylococcus aureus if it is present.*
>
> *Tryptone is a source of amino acids derived from casein, which is included in the broth as a nutrient. It contains nitrogen, in a form readily available to the bacteria, which encourages growth. The current protocol requires tryptone to be included in the broth at a concentration of 1.0% (by weight) and cultured at 35 degrees Celsius for 24 hours. The goal is to decide if these are the best conditions for culturing the MRSA.*
>
> *Five strains of MRSA were used so that the conditions for optimum growth could be investigated for each strain separately. These strains were WSPP1 MRSA, WSPP2 MRSA, Akh2 MRSA, Phage pattern 52/52A/79/.., and Phage pattern 29/52/77/+. Because of these curious names they are simply referred to in the data as strain 1, 2, 3, 4 and 5 respectively.*

En résumé : Le staphylocoque doré est une bactérie pathogène responsable d'infections nosocomiales. En cas d'infection postopératoire, une culture en laboratoire est réalisée afin de préciser le diagnostic et prescrire le bon antibiotique. Le jeu de données contient les comptages des 5 souches mises en culture. Pour chaque souche on a essayé une fois chacune des différentes combinaisons possibles des 3 conditions expérimentales :

- le temps : (Times) 24h et 48h
- a température : (Temp) 27°, 35° et 43°
- la concentration du gel en nutriment (Conc) : 0,6% 0,8% 1,0% 1,2% et 1,4%

L'objectif est de trouver les conditions optimales de culture de ces souches. Il s'agit de répondre aux questions biologiques suivantes :

- La souche influence-t-elle la concentration ?
- Les facteurs (température, temps, concentration) ont-ils un effet ?
- Construire le meilleur modèle de prévision et en déduire les valeurs optimales des paramètres sensibles en fixant le temps à 24h.

2 Gestion des données

Les données sont téléchargées à partir de l'URL :
http ://www.amstat.org/publications/jse/datasets/Tryptone.dat.txt

Télécharger le fichier dans le répertoire courant de R.

Les données se présentent dans un fichier texte avec le format suivant. Cela nécessite de les réorganiser pour faciliter toute analyse statistique à l'aide d'un logiciel.

Row	Count1	Count2	Count3	Count4	Count5	Time	Temp	Conc
1	9	3	10	14	33	24	27	0.6
2	16	12	26	20	31	24	27	0.8
...								
29	232	216	234	177	201	48	43	1.2
30	163	141	172	212	184	48	43	1.4

1. Statistical analysis of microbiology data : optimal conditions for culturing different strains of *Staphylococcus aureus*, *Journal of Statistics Education*, Volume 12, Number 2.

Il s'agit de concaténer les sous-matrices en introduisant une variable identifiant la souche plutôt qu'une colonne par souche. Il est très classique d'avoir à faire ce traitement à partir de données gérées avec Excel. Voici la solution avec R :

```
#Lire les données brutes
trip=read.table("Tryptone.dat.txt",header=T)
#transformer les données en empilant par type de souche
tryp=rbind(data.frame(Souche=rep("strain1",30),trip[,7:9],Count=trip[,2]),
           data.frame(Souche=rep("strain2",30),trip[,7:9],Count=trip[,3]),
           data.frame(Souche=rep("strain3",30),trip[,7:9],Count=trip[,4]),
           data.frame(Souche=rep("strain4",30),trip[,7:9],Count=trip[,5]),
           data.frame(Souche=rep("strain5",30),trip[,7:9],Count=trip[,6]))
```

Produisant des données sous la forme :

```
     Souche Time Temp Conc Count
1   strain1   24   27  0.6     9
2   strain1   24   27  0.8    16
3   strain1   24   27  1.0    22
...
148 strain5   48   43  1.0   127
149 strain5   48   43  1.2   201
150 strain5   48   43  1.4   184
```

Dont il faut encore vérifier la bonne cohérence :

```
summary(tryp)
# commande pour pouvoir identifier chaque variable
# de la base tryp
attach(tryp)
```

3 Exploration

3.1 Unidimensionnelle

Le commandes suivantes permettent de vérifier la distribution de la variable de dénombrement des bactéries et de tester l'éventuel intérêt d'une transformation.

```
hist(Count)
hist(sqrt(Count))
boxplot(Count)
boxplot(sqrt(Count))
```

La commande suivante de test sur un échantillon fournit un intervalle de confiance de la moyenne :

```
t.test(Count, conf.level=.95)
```

Vérifier les propriétés (droite de Henri et test) de la distribution de la variable et de sa transformation. Sont-elles gaussiennes ?

```
qqnorm(Count)
qqline(Count)
qqnorm(sqrt(Count))
qqline(sqrt(Count)Count)
shapiro.test(Count)
shapiro.test(sqrt(Count))
```

Que dire de la transformation ?

3.2 Bidimensionnelle

Croiser (diagramme boîte) les différentes variables ou facteurs avec le dénombrement. Que dire de l'influence possible de ces facteurs, de la linéarité de l'effet ?

```
boxplot(Count~Time)
boxplot(Count~Temp)
boxplot(Count~Conc)
boxplot(Count~Souche)
```

Commentaires.

4 Influence du type de la souche

Il s'agit de metre en œuvre une ANOVA et donc d'en vérifier les bonnes conditions d'application.

```
bartlett.test(Count~Souche)
```

Commentaire sur ce résultat avant d'estimer le modèle :

```
res.anova=aov(Count~Souche)
```

et étudier le comportement des résidus :

```
qqnorm(res.anova$residuals)
qqline(res.anova$residuals)
shapiro.test(res.anova$residuals)
```

Que dire de la validité de l'ANOVA ?

```
#Résultas
summary(res.anova)
```

Conclure sur la pertinence et la significativité du test. Que conclure sur la question biologique ?

Dans quelle situation, le test ci-dessous serait pertinent ?

```
kruskal.test(Count~Souche)
```

5 Influence des autres facteurs

Compte tenu du résultat précédent, la variable "type de souche" n'est pas pris en compte et les échantillons regroupés. Il serait plus judicieux de tout tester en une seule fois dans un même modèle par une "analyse de covariance" mais comme ce modèle est plus complexe, le travail a été réalisé en deux phases. Attention, ceci n'est pas complètement rigoureux pour s'assurer du bon niveau du risque de première espèce car plusieurs tests (tests multiples) sont réalisés sur le même échantillon.

Les autres facteurs sont quantitatifs, un modèle linéaire de régression est donc plus appropié, en terme de puissance de test, qu'une ANOVA car celle-ci estime plus de paramètre et donc conduit à restreindre le nombre de degrés de liberté.

5.1 Modèle linéaire

Excuter les commandes pour ajuster un modèle linéaire et étudier les propriétés du modèle obtenu.

```
res1.reg=lm(Count ~ Time + Temp + Conc)
```

```
# Normalité des résidus
qqnorm(res1.reg$residuals)
qqline(res1.reg$residuals)
shapiro.test(res1.reg$residuals)
# Nuage des résidus
res.student=rstudent(res1.reg)
ychap=res1.reg$fitted.values
plot(res.student~ychap,ylab="Résidus")
abline(h=c(-2,0,2),lty=c(2,1,2))
# Points influents
cook=cooks.distance(res1.reg)
plot(cook~ychap,ylab="Distance de Cook")
abline(h=c(0,1),lty=c(1,2))
```

Que conclure sur la validité du modèle ?

```
summary(res1.reg)
```

Que conclure sur l'influence des facteurs, la qualité d'ajustement, les qualités potentielles de prévision ?

5.2 Modèle quadratique

La qualité d'ajustement étant relativement faible dans le modèle précédent, un modèle polynomial de degré 2 est construit puis testé.

```
# Construction des variables
tryp2=data.frame(tryp[,-1],Conc**2,Temp**2,
   Temp*Conc,Temp*Time,Time*Conc)
# estimation
res2.reg=lm(Count ~ ., data=tryp2)
# Normalité des résidus
qqnorm(res2.reg$residuals)
qqline(res2.reg$residuals)
shapiro.test(res2.reg$residuals)
# Nuage des résidus
res.student=rstudent(res2.reg)
ychap=res2.reg$fitted.values
plot(res.student~ychap,ylab="Résidus")
```

```
abline(h=c(-2,0,2),lty=c(2,1,2))
# Points influents
cook=cooks.distance(res2.reg)
plot(cook~ychap,ylab="Distance de Cook")
abline(h=c(0,1),lty=c(1,2))
```

```
summary(res2.reg)
```

Que dire du R^2 par rapport au modèle précédent mais surtout de la forme du nuage des résidus et donc de la validité du modèle ?

5.3 Modèle après transformation

Une transformation est appliquée afin de tenter de répondre au problème soulevé par le modèle précédent.

```
# Gestion des données
tryp3=data.frame(tryp[,-c(1,5)],sqrt(Count),
  Conc**2,Temp**2,Temp*Conc,Temp*Time,Time*Conc)
summary(tryp3)
# Estimation
res3.reg=lm(sqrt.Count. ~ ., data=tryp3)
# Normalité des résidus
shapiro.test(res3.reg$residuals)
# Nuage des résidus
res.student=rstudent(res3.reg)
ychap=res3.reg$fitted.values
plot(res.student~ychap,ylab="Résidus")
abline(h=c(-2,0,2),lty=c(2,1,2))
```

Que dire des résidus (normalité, forme) par rapport au modèle précédent, que dire de la qualité d'ajustement de ce modèle et de l'influence des intreractions ?

6 Conditions optimales

La fonction suivante est utile. Que calcule-t-elle ?

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

6.1 Modèle de prévision

Les modèles suivants sont successivements estimés et comparés. Commenter.

```
res3.reg=lm(sqrt.Count. ~ Time+Temp+Conc+
   Conc.2+Temp.2+Temp...Conc+Temp...Time+
   Time...Conc, data=tryp3)
summary(res3.reg)
```

Pourquoi estimer le modèle ci-dessous ?

```
res4.reg=lm(sqrt.Count. ~ Time+Temp+Conc+
   Conc.2+Temp.2+Temp...Conc+Temp...Time,
   data=tryp3)
summary(res4.reg)
press(res4.reg)
```

Même chose avec ce dernier modèle :

```
res5.reg=lm(sqrt.Count. ~ Time+Temp+Conc+
   Conc.2+Temp.2+Temp...Time, data=tryp3)
summary(res5.reg)
press(res5.reg)
```

Retenir le "meilleur" modèle de prévision, il constitue la *surface de réponse* du problème d'optimisation.

Poser "Time=24" dans ce modèle pour en déduire l'expression approchée au mieux de la racine du nombre de bactérie en fonction de la température et de la concentration de tryptone.

Quel est le couple de valeurs (température, concentration) qui maximise cette fonction ?