Apprentissage de données biologiques
Examen
2021

Tous les documents et la calculatrice sont autorisés.

1. Analyse statistique des défauts fromagers

Objectif

On cherche ici à mettre en évidence les caractéristiques de fromages à pâte cuite qui expliqueraient la présence de défauts les rendant impropres à la commercialisation (défectueux).

Données

On dispose pour cette étude des résultats d'un plan de suivi en usine consistant à déclarer défecteux ou non chaque fromage d'un échantillon (une centaine de fromages prélévés par jour pendant 140 jours, soit au total plus de 15000 fromages). Les valeurs moyennes journalières d'indicateurs de la qualité sanitaire (`San1` et `San2`) d'une part et de la qualité fromagère (`From1`, `From2` et `From3`) d'autre part sont aussi disponibles. Un extrait des données est fourni dans le tableau 1. La variable `Def` donne le statut du fromage (1 si défectueux, 0 sinon). La variable `Jour` donne un numéro identifiant chaque journée du plan de suivi.

R script

```
> str(from)

'data.frame':        15153 obs. of  7 variables:
 $ Def  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ San1 : int  4375 4375 4375 4375 4375 4375 4375 4375 4375 4375 ...
 $ San2 : num  15255 15255 15255 15255 15255 ...
 $ From1: num  30.6 30.6 30.6 30.6 30.6 ...
 $ From2: num  34 34 34 34 34 34 34 34 34 34 ...
 $ From3: num  0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 0.9 ...
 $ Jour : Ord.factor w/ 140 levels "1"<"2"<"3"<"4"<..: 1 1 1 1 1 1 1 1 1 1 ...
```

TABLE 1 – Données de plan de suivi en usine de fabrication de fromages.

1. A partir des variables du tableau 1, comment peut-on définir la variable à expliquer ? Quelles sont les variables explicatives ?

Dans un premier temps, on s'intéresse à l'étude du lien entre la variable à expliquer et la variable `From1` (notée x dans la suite), qui présente un intérêt particulier en pratique.

2. Quel modèle statistique, appelé dans la suite $\mathcal{M}$, est adapté à cette étude (vous donnerez le nom usuel de ce modèle, sa forme mathématique explicite et le nombre de ses paramètres) ?

Les résultats de l'ajustement du modèle $\mathcal{M}$ par la fonction `glm` de `R` sont donnés dans le tableau 2.

Le modèle $\mathcal{M}$ permet de calculer, pour chaque valeur de x, une estimation de la probabilité que le fromage soit défectueux. Par exemple, les 6 premiers jours de l'expérience, le tableau 3 donne les valeurs de x, de la proportion de fromages défectueux et de la probabilité estimée par le modèle $\mathcal{M}$ qu'un fromage soit défectueux.

R script

```
> summary(mod)

Call:
glm(formula = Def ~ From1, family = binomial, data = from)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.8287     1.8481   3.695  0.00022 ***
From1        -0.3265     0.0606  -5.387 7.16e-08 ***

    Null deviance: 5209.1  on 15152  degrees of freedom
Residual deviance: 5181.0  on 15151  degrees of freedom
AIC: 5185

Number of Fisher Scoring iterations: 6
```

TABLE 2 – Résultats de l'ajustement du modèle $\mathcal{M}$.

R script

```
> from1 = tapply(X=from$From1,INDEX=from$Jour,FUN=mean)
> head(from1)

    1     2     3     4     5     6
30.55 30.55 30.63 30.09 30.17 30.13

> propdef = tapply(X=from$Def,INDEX=from$Jour,FUN=mean)
> head(propdef)

         1          2          3          4          5          6
0.00000000 0.11538462 0.02884615 0.04807692 0.28846154 0.20192308

> probadef = predict(mod,type="response",newdata=data.frame(From1=from1))
> head(probadef)

         1          2          3          4          5          6
0.04130694 0.04130694 0.04028501 0.04768083 0.04650886 0.04709138
```

TABLE 3 – Probabilités estimées par le modèle $\mathcal{M}$ qu'un fromage soit défecteux et proportions observées de fromages défectueux lors des 6 premiers jours du plan de suivi.

3. Par quelle opération mathématique est calculée la probabilité, estimée par le modèle $\mathcal{M}$, qu'un fromage soit défectueux lorsque sa valeur de x vaut celle observée le 1er jour du plan de suivi, à savoir 30.55 ?

4. A partir des résultats rapportés dans le tableau 3, diriez-vous que le modèle s'ajuste bien aux données ? Argumentez votre réponse.

On suggère de modifier le modèle $\mathcal{M}$ en s'affranchissant de l'hypothèse de linéarité de l'effet de x par l'utilisation d'un modèle non paramétrique.

5. Donnez l'expression mathématique du modèle $\mathcal{M}^*$ résultant de cette modification de $\mathcal{M}$.

Une table d'analyse de la déviance non-paramétrique du modèle $\mathcal{M}^*$ est donnée dans le tableau 4.

R script

```
> mod0 = gam(Def~From1,family=binomial,data=from)
> mod = gam(Def~s(From1,df=8),family=binomial,data=from)
> anova(mod0,mod,test="Chisq")

Analysis of Deviance Table

Model 1: Def ~ From1
Model 2: Def ~ s(From1, df = 8)
  Resid. Df Resid. Dev     Df Deviance  Pr(>Chi)
1     15151     5181.0
2     15144     4691.4 7.0003   489.58 < 2.2e-16 ***
```

TABLE 4 – Table d'analyse de la déviance non-paramétrique du modèle $\mathcal{M}^*$.

6. Quelles sont les hypothèses nulle et alternative du test d'analyse de la déviance non-paramétrique donné dans le tableau 4 ?

L'estimation du modèle $\mathcal{M}^*$ introduit un hyperparamètre appelé `df` dans les arguments de la fonction `s` (voir tableau 4).

7. Plus la valeur de `df` est grande, plus les valeurs estimées des probabilités qu'un fromage du plan de suivi soit défectueux sont proches ou au contraire différentes des proportions observées de fromages défectueux ? Argumentez votre réponse.

Les commandes `R` reproduites dans le tableau 5 permettent de choisir la valeur de `df`, parmi trois valeurs candidates.

8. Décrire la procédure implémentée dans le tableau 5. Quel est finalement le meilleur choix pour la valeur de `df` ?

Pour chaque variable explicative, on cherche ainsi le meilleur ajustement possible, linéaire ou non-paramétrique, et le cas échéant, on choisit aussi le nombre de degrés de liberté le plus adapté à l'ajustement. Le tableau 6 donne la trace de la sélection du meilleur modèle non-paramétrique incluant toutes les variables explicatives.

9. D'après le tableau 6, quel est le meilleur modèle parmi ceux n'ayant qu'une variable explicative ? Selon quel critère ce modèle est-il meilleur que tous les autres modèles ?

R script

```
> vecdf = c(1,10,20)
> jours = levels(from$Jour)
> mse = rep(0,length(vecdf))
> for (i in 1:length(vecdf)) {
+   cvproba = rep(0,140)
+   for (k in 1:140) {
+     train = from[from$Jour!=jours[k],]
+     train = droplevels(train)
+     test = from[from$Jour==jours[k],]
+     test = droplevels(test)
+     mod = gam(Def~s(From1,df=vecdf[i]),family=binomial,data=train)
+     cvproba[k] = mean(predict(mod,newdata=test,type="response"))
+   }
+   mse[i] = mean((cvproba-propdef)^2)
+ }
> mse

[1] 0.01107822 0.01047658 0.01802857
```

TABLE 5 – Choix de la valeur du paramètre `df`.

R script

```
> from.gam = gam(Def~1,family=binomial,data=from)
> from.step = step.Gam(from.gam,scope=list("San1"=~1+San1+s(San1,10),
+                                          "San2"=~1+San2+s(San2,10),
+                                          "From1"=~1+From1+s(From1,5),
+                                          "From2"=~1+From2+s(From2,5),
+                                          "From3"=~1+From3+s(From3,2)))

Start:  Def ~ 1; AIC= 5211.112
Step:1 Def ~ San1 ; AIC= 4502.798
Step:2 Def ~ San1 + From2 ; AIC= 4172.978
Step:3 Def ~ s(San1, 10) + From2 ; AIC= 4028.34
Step:4 Def ~ s(San1, 10) + s(From2, 5) ; AIC= 3937.949
Step:5 Def ~ s(San1, 10) + s(From2, 5) + From3 ; AIC= 3895.03
Step:6 Def ~ s(San1, 10) + From1 + s(From2, 5) + From3 ; AIC= 3878.533
Step:7 Def ~ s(San1, 10) + s(From1, 5) + s(From2, 5) + From3 ; AIC= 3809.164
Step:8 Def ~ s(San1, 10) + s(From1, 5) + s(From2, 5) + s(From3, 2) ; AIC= 3796.62
```

TABLE 6 – Sélection du meilleur modèle non-paramétrique.