CC2 d'Elements de modélisation statistique

17/01/2020

Durée : 2h
Les documents, les calculatrices et les téléphones portables ne sont pas autorisés.
Vous prendrez soin à la rédaction de vos réponses.

Exercice 1

Un chercheur s'intéresse à la façon dont des variables, telles que le **gre** (Graduate Record Exam scores), le **gpa** (grade point average) et le prestige de l'établissement de premier cycle (variable **rank**), affectent l'admission aux études supérieures (variable **admit**, 1 si admis / 0 si non admis).

```
##  admit       gre             gpa        rank
##  0:273   Min.   :220.0   Min.   :2.260   1: 61
##  1:127   1st Qu.:520.0   1st Qu.:3.130   2:151
##          Median :580.0   Median :3.395   3:121
##          Mean   :587.7   Mean   :3.390   4: 67
##          3rd Qu.:660.0   3rd Qu.:3.670
##          Max.   :800.0   Max.   :4.000
```

Question 1 : On considère un modèle, appelé *mod* dans la suite, utilisant toutes les variables explicatives sans interaction pour expliquer la variable **admit**. Ecrivez l'équation de ce modèle.

```
mod= glm(admit ~ ., data = mydata, family = "binomial")
summary(mod)
```

```
##
## Call:
## glm(formula = admit ~ ., family = "binomial", data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6268  -0.8662  -0.6388   1.1490   2.0790
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.989979   1.139951  -3.500 0.000465 ***
## gre          0.002264   0.001094   2.070 0.038465 *
## gpa          0.804038   0.331819   2.423 0.015388 *
## rank2       -0.675443   0.316490  -2.134 0.032829 *
## rank3       -1.340204   0.345306  -3.881 0.000104 ***
## rank4       -1.551464   0.417832  -3.713 0.000205 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 499.98  on 399  degrees of freedom
## Residual deviance: 458.52  on 394  degrees of freedom
## AIC: 470.52
##
## Number of Fisher Scoring iterations: 4
```

Question 2 : Donnez la définition du pseudo R^2 et donnez sa valeur sous forme d'une fraction numérique pour le modèle *mod*.

Question 3 : Donnez une interprétation pour les valeurs 1.002 et 0.262 dans la sortie suivante :

```
round(exp(coef(mod)),digits=3)
```

```
## (Intercept)         gre         gpa       rank2       rank3       rank4
##       0.019       1.002       2.235       0.509       0.262       0.212
```

(Vous prendrez soin de justifier votre réponse mathématiquement avant d'interpréter par une phrase.)

Question 4 : Construisez un test statistique pour tester l'influence de la variable **rank** au niveau 5%. On obtient pour ce test une pvaleur de 7.09×10^{-5}, concluez.

Question 5 : On décide de complexifier le modèle en ajoutant un terme d'interaction entre les variables **gre** et **gpa**. Est-ce-que le terme d'interaction entre **gre** et **gpa** est nécessaire ? Vous répondrez en construisant un test adapté et concluez pour une erreur de première espèce de 5%.

```
##
## Call:
## glm(formula = admit ~ . + gre * gpa, family = "binomial", data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4815  -0.8772  -0.6357   1.1243   2.2612
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -13.608810   6.071148  -2.242 0.024990 *
## gre           0.018344   0.009968   1.840 0.065722 .
## gpa           3.652170   1.788448   2.042 0.041143 *
## rank2        -0.721697   0.319150  -2.261 0.023740 *
## rank3        -1.343466   0.346383  -3.879 0.000105 ***
## rank4        -1.606298   0.420991  -3.816 0.000136 ***
## gre:gpa      -0.004719   0.002898  -1.629 0.103418
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 499.98  on 399  degrees of freedom
## Residual deviance: 455.77  on 393  degrees of freedom
## AIC: 469.77
##
## Number of Fisher Scoring iterations: 4
```

Exercice 2

Donnez la définition d'un modèle loglinéaire et expliquez comment vous obtenez les valeurs ajustées pour chaque individu.

Exercice 3

L'absorption de CO2 de six plantes originaires du Québec et de six originaires du Mississippi a été mesurée à plusieurs niveaux de concentration ambiante de CO2. La moitié des plantes de chaque type ont été réfrigérées pendant une nuit avant de mener l'expérience. Les variables sont :

- type : origine de la plante (Québec ou Mississippi)
- treat : indique si la plante a été réfrigérée (chilled) ou pas (nonchilled)
- conc : concentration ambiante de dioxyde de carbone (mL/L).
- taux : taux d'absorption de dioxyde de carbone (umol/m^2 sec).

On pourra noter taux_{ijk} (resp. conc_{ijk}) la valeur de la variable **taux** (resp. **conc**) pour la $k$ème plante prenant la modalité $i \in I = \{$"Québec", "Mississipi"$\}$ pour la variable **type** et la modalité $j \in J = \{$"chilled", "nonchilled"$\}$ pour la variable **treat**.

```
##           type           treat         conc           taux      
##  Quebec     :42   nonchilled:42   Min.   :  95   Min.   : 7.70  
##  Mississippi:42   chilled   :42   1st Qu.: 175   1st Qu.:17.90  
##                                    Median : 350   Median :28.30  
##                                    Mean   : 435   Mean   :27.21  
##                                    3rd Qu.: 675   3rd Qu.:37.12  
##                                    Max.   :1000   Max.   :45.50
```

Partie 1

Question 1 : Proposez un modèle linéaire pour expliquer la variable **taux** en fonction de **treat** et **conc**, et avec interaction des variables explicatives. Ecrivez ce modèle sous forme régulière.
Dans la suite, ce modèle est appelé *mod1* et s'écrit matriciellement $T = X\theta + \varepsilon$.

Question 2 : Construisez un intervalle de prédiction pour le taux d'absorption de dioxyde de carbone d'une plante qui a été réfrigérée la veille et avec une concentration ambiante de dioxyde de carbone de 400 au niveau de confiance de 95%.

Question 3 : Construisez un test pour tester la nullité de la première coordonnée du vecteur θ.

Partie 2

Question 1 : Ecrivez l'équation du modèle linéaire permettant d'expliquer **taux** en fonction de **type** et **treat**, **avec** effet d'interaction entre les variables explicatives. On appellera *mod2* ce modèle par la suite.

Question 2 : Construisez un test pour évaluer l'influence de l'effet d'interaction dans le modèle *mod2*. On a obtenu une pvaleur de 0.064, concluez.

Partie 3

Dans cette partie, on considère le modèle suivant, appelé *mod3* :

$$\begin{cases} \text{taux}_{ijk} = \mu + \alpha_i + \varepsilon_{ijk}, \ \forall i \in I, \forall j \in J, \forall k \in \{1, \ldots, 21\} \\ \varepsilon_{ijk} \text{ i.i.d } \mathcal{N}(0, \sigma^2). \end{cases}$$

Question 1 : Ecrivez la fonction des moindres carrés associée au modèle *mod3*.

Question 2 : Estimez les paramètres du modèle *mod3* sous les contraintes d'orthogonalité.