Université de Montpellier Année 2022-2023
HAX912X - Modèles linéaires généralisés

Examen final

Durée 2h - Sans document
Mercredi 26 octobre 2022

Exercice 1 (4 pts)

Décrire et commenter les résultats obtenus à l'aide du code R ci-dessous. Les données considérées sont décrites en annexe.

```
> library(datasets)
> data(warpbreaks)
> model <- glm(breaks ~ wool + tension, warpbreaks ,
family = poisson(link = "log"))
> summary(model)

Call:
  glm(formula = breaks ~ wool + tension, family =
  poisson(link = "log"), data = warpbreaks)

Deviance Residuals:
  Min       1Q   Median       3Q      Max
-3.6871  -1.6503  -0.4269   1.1902   4.2616

Coefficients:
  Estimate Std. Error z value Pr(>|z|)

  (Intercept)  3.69196    0.04541  81.302  < 2e-16 ***
  woolB       -0.20599    0.05157  -3.994 6.49e-05 ***
  tensionM    -0.32132    0.06027  -5.332 9.73e-08 ***
  tensionH    -0.51849    0.06396  -8.107 5.21e-16 ***

  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 297.37  on 53  degrees of freedom
Residual deviance: 210.39  on 50  degrees of freedom
AIC: 493.06

Number of Fisher Scoring iterations: 4
```

Exercice 2 (6 pts)

Décrire et commenter les résultats obtenus à l'aide du code Python ci-dessous. Les données considérées sont décrites en annexe.

```
import pandas as pd
from sklearn.linear_model import LogisticRegression
data = pd.read_csv('/Users/jmm/TEACHING/2223/M2-GLM-HAX912X/EXAMENS/
    Breast-Cancer-Wisconsin.csv')
data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 33 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       569 non-null    int64
 1   diagnosis                569 non-null    object
 2   radius_mean              569 non-null    float64
 3   texture_mean             569 non-null    float64
 4   perimeter_mean           569 non-null    float64
 5   area_mean                569 non-null    float64
 6   smoothness_mean          569 non-null    float64
 7   compactness_mean         569 non-null    float64
 8   concavity_mean           569 non-null    float64
 9   concave points_mean      569 non-null    float64
 10  symmetry_mean            569 non-null    float64
 11  fractal_dimension_mean   569 non-null    float64
 12  radius_se                569 non-null    float64
 13  texture_se               569 non-null    float64
 14  perimeter_se             569 non-null    float64
 15  area_se                  569 non-null    float64
 16  smoothness_se            569 non-null    float64
 17  compactness_se           569 non-null    float64
 18  concavity_se             569 non-null    float64
 19  concave points_se        569 non-null    float64
 20  symmetry_se              569 non-null    float64
 21  fractal_dimension_se     569 non-null    float64
 22  radius_worst             569 non-null    float64
 23  texture_worst            569 non-null    float64
 24  perimeter_worst          569 non-null    float64
 25  area_worst               569 non-null    float64
 26  smoothness_worst         569 non-null    float64
 27  compactness_worst        569 non-null    float64
 28  concavity_worst          569 non-null    float64
 29  concave points_worst     569 non-null    float64
 30  symmetry_worst           569 non-null    float64
 31  fractal_dimension_worst  569 non-null    float64
 32  Unnamed: 32              0 non-null      float64
dtypes: float64(31), int64(1), object(1)
memory usage: 146.8+ KB
```

```
y = data['diagnosis']
X = data.drop(['id','diagnosis','Unnamed: 32'], axis = 1)
X.head()
   radius_mean  texture_mean  ...  symmetry_worst  fractal_dimension_worst
0        17.99         10.38  ...          0.4601                  0.11890
1        20.57         17.77  ...          0.2750                  0.08902
2        19.69         21.25  ...          0.3613                  0.08758
3        11.42         20.38  ...          0.6638                  0.17300
4        20.29         14.34  ...          0.2364                  0.07678
[5 rows x 30 columns]

x.shape
(569, 30)

model1 = LogisticRegression(penalty='l2', C=1.0, class_weight=None,
dual=False, fit_intercept=True, l1_ratio=None, solver='lbfgs',
max_iter=5000, tol=0.0001)
model1.fit(X,y)
print(model1.score(X,y))
0.9578207381370826

model2 = LogisticRegression(penalty='l2', C=100.0, class_weight=None,
dual=False, fit_intercept=True, l1_ratio=None, solver='lbfgs',
max_iter=5000, tol=0.0001)
model2.fit(X,y)
print(model2.score(X,y))
0.9789103690685413

from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.25)
X_test.shape
(143, 30)

from sklearn.model_selection import RepeatedKFold
from sklearn.model_selection import GridSearchCV
cv = RepeatedKFold(n_splits=5, n_repeats=10)
parameters = {'C':range(1,380,20)}
model3 = GridSearchCV(LogisticRegression(penalty='l2', class_weight=None,
dual=False, fit_intercept=True, l1_ratio=None, solver='lbfgs',
max_iter=5000, tol=0.0001),parameters,cv=cv,n_jobs=-1)
model3.fit(X_train, y_train)

print(model3.best_params_)
{'C': 161}
print(model3.score(X_test,y_test))
0.972027972027972
```

Exercice 3 (6 pts)

On considère n variables aléatoires indépendantes $y_1, \ldots, y_n$ telles que y_i est distribuée suivant une loi inverse gaussienne de paramètres (μ_i, σ^2) où $\mu_i > 0$ et $\sigma^2 > 0$.
Le terme inverse ne doit pas être mal interprété, la loi est inverse dans le sens suivant : la valeur du mouvement brownien à un temps fixé est de loi normale, à l'inverse, le temps en lequel le mouvement brownien avec une dérive positive atteint une valeur fixée est de loi inverse gaussienne. Nous avons

$$f(y_i; \mu_i, \sigma^2) = \frac{1}{\sqrt{2\pi y_i^3}\sigma} \exp\left(-\frac{(y_i - \mu_i)^2}{2(\mu_i\sigma)^2 y_i}\right) \mathbf{1}_{y_i>0}\,.$$

1 (2 pts) Montrer que la loi inverse gausienne appartient à la famille exponentielle avec un paramètre de nuisance.

2 (2 pts) Calculer $\mathbb{E}(y_i)$ et $\mathbb{V}(y_i)$.

Pour tout $i = 1, \ldots, n$, on suppose que $\log(\mu_i) = a + bx_i$ où $x_i \in \mathbb{R}$ est supposé connu, a et b sont des paramètres inconnus.

3 (2 pts) Montrer qu'il s'agit d'un modèle linéaire généralisé. La fonction de lien canonique a-t-elle été utilisée ?

Exercice 4 (4 pts - QCM - Bonne réponse +1 pt - Mauvaise réponse -0.5 pt)

1 (1 pt) Lors d'une regression linéaire, si le $R^2 = 1$, les points sont-ils alignés ?

A) non
B) oui
C) pas obligatoirement

2 (1 pt) La droite des moindres carrées d'une regression simple passe par le point moyen

A) toujours
B) jamais
C) parfois

3 (1 pt) L'ensemble des modèles log-linaires grahiques par rapport à celui des modèles log-linéaires hiérarchiques est

A) plus riche
B) moins riche
C) le même

4 (1 pt) Dans une régression linéaire multiple, on utilise le critère BIC pour

A) prédire
B) expliquer

Annexe

Breast cancer data (Données sur le cancer du sein)

About Dataset

Features are computed from a digitized image of a fine needle aspirate (FNA)
of a breast mass. They describe characteristics of the cell nuclei present
in the image. n the 3-dimensional space is that described in:
[K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination
of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-
34].

Attribute Information:

1) ID number
2) Diagnosis (M = malignant, B = benign)

3-32)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three
largest values) of these features were computed for each image,
resulting in 30 features. For instance, field 3 is Mean Radius, field
13 is Radius SE, field 23 is Worst Radius.

All feature values are recoded with four significant digits.

Missing attribute values: none

Class distribution: 357 benign, 212 malignant

Warp breaks per loom data (ruptures de chaîne par métier à tisser)

warpbreaks {datasets} R Documentation
The Number of Breaks in Yarn during Weaving

Description

This data set gives the number of warp breaks per loom,
where a loom corresponds to a fixed length of yarn.

Usage
warpbreaks
Format
A data frame with 54 observations on 3 variables.

[,1] breaks numeric The number of breaks
[,2] wool factor The type of wool (A or B)
[,3] tension factor The level of tension (L, M, H)
There are measurements on 9 looms for each of the six types
of warp (AL, AM, AH, BL, BM, BH).

Correction Examen Final HAX912X 26/10/2022

Exercice 1

Nous considérons un jeu de données contenant 54 individus et 3 variables. Nous cherchons à expliquer la nature des types de fil de laine utilisés dans la confection d'un métier à tisser. Les 2 variables explicatives sont qualitatives et correspondent au type de laine (2 modalités) et au niveau d'étirement du fil, le plus élevé étant également le plus coûteux (3 modalités).

Sans expliquer la manière dont
nous avons fait le choix (modèle 4), on
utilise une régression de Poisson
avec le log comme fonction
de lien. Nous supposons ainsi
que $\log(E(y)) = \alpha + \beta_B \mathbb{1}_{\{x_1 = "B"\}}$

$+ \beta_{T,\Pi} \mathbb{1}_{\{x_2 = "\Pi"\}}$

$+ \beta_{T,H} \mathbb{1}_{\{x_2 = "H"\}}$

où x_1 = type de bois
x_2 = niveau de Tension du fil
À l'aide des 54 observations, on obtient
$\hat{\alpha} \# 3,69$; $\hat{\beta}_B \# -0,21$, $\hat{\beta}_{T,\Pi} \# -0,32$
et $\hat{\beta}_{T,H} \# -0,52$
Sur l'effet d'interaction, modalités de
référence sont A et L.

Exercice 2

Nous considérons ici un jeu de données contenant 569 individus et 30 variables. Il s'agit de construire un prédicteur de la présence de cellules cancéreuses en fonction de ces caractéristiques décrites à l'image. Dans le jeu de données 212 images sur 569 contiennent des cellules cancéreuses (cancer du sein).

Le premier modèle correspond à une régression logistique régularisée ridge avec un paramètre de régularisation C fixé à 1 (modèle 1).

Le deuxième modèle est aussi une régression logistique régularisée ridge avec $C = 100$ (modèle 2)

Deux pourcentages de tous
éléments sont donnés.
Ils sous-évaluent l'erreur car ils
sont calculés sur les mêmes données qui
elles ayant été utilisées pour
ajuster les modèles 1 et 2.

On isole 25% des données
à savoir 143 individus.
Sur les 426 données restantes,
on effectue une régression logistique
régularisée ridge pour laquelle la
puissance de régularisation est
calculé en utilisant une technique
de validation croisée à 5 ensembles
répété 10 fois. On obtient $\hat{C} = \underline{161}$
Enfin, on calcule le taux de fois dans
lequel un modèle correspondant maximisant

l'échantillon du test précédemment isolé. On obtient un classifieur qui a un taux estimé de bons classements de 97% environ.

Exercice 3

1) $f(y|\mu,\sigma^2) = \exp\left\{-\frac{(y-\mu)^2}{2\mu^2\sigma^2 y} - \frac{1}{2}\log(\sigma^2)\right\} \underbrace{\frac{1}{\sqrt{2\pi y^3}}\mathbb{1}_{\{y>0\}}}_{\text{mesure de référence } \nu(dy)}$

$$f(y|\mu,\sigma^2) = \exp\left\{-\frac{y^2 - 2\mu y + \mu^2}{2\mu^2\sigma^2 y} - \frac{1}{2}\log(\sigma^2)\right\} \nu(dy)$$

$$f(y|\mu,\sigma^2) = \exp\left\{-\frac{1}{\sigma^2}\left[\frac{y}{2\mu^2} - \frac{1}{\mu}\right] - \frac{1}{2y\sigma^2} - \frac{1}{2}\log(\sigma^2)\right\} \nu(dy)$$

$\phi = \sigma^2$, $\theta = -\frac{1}{2\mu^2}$, $h(\theta) = -\frac{1}{\mu}$

et $c(y,\sigma^2) = -\frac{1}{2y\sigma^2} - \frac{1}{2}\log(\sigma^2)$

$\mu^2 = -\frac{1}{2\theta} \Rightarrow \mu = \frac{1}{\sqrt{-2\theta}}$ et donc $h(\theta) = -\sqrt{2\theta}$

Ainsi la loi inverse gaussienne appartient à la famille exponentielle avec un paramètre de nuisance.

2] D'après les propriétés des familles exponentielles, nous savons que

$$E(y) = b'(\theta) = -\frac{-2}{2\sqrt{-2\theta}} = \frac{1}{\sqrt{-2\theta}} = \mu$$

$$V(y) = \phi\, b''(\theta) = \sigma^2 \left[\frac{-(-2)\frac{1}{2\sqrt{-2\theta}}}{(-2\theta)} \right]$$

$$V(y) = \sigma^2 \left[\frac{1}{(-2\theta)^{3/2}} \right] = \sigma^2 \mu^3$$

3] $\log(\mu) = \alpha + \beta x$

$\Rightarrow \mu = \exp(\alpha + \beta x)$

$\Rightarrow E(\mu) = \exp(\alpha + \beta x)$

C'est bien un modèle linéaire généralisé.

$$\theta = -\frac{1}{2\mu^2} = -\frac{1}{2} e^{-2a-2bx}$$

$$\neq a+bx$$

(7)

Ce n'est pas la bien commique qui a été utilisé.

Exercice 4

1| C

2| C

3| B

4| B