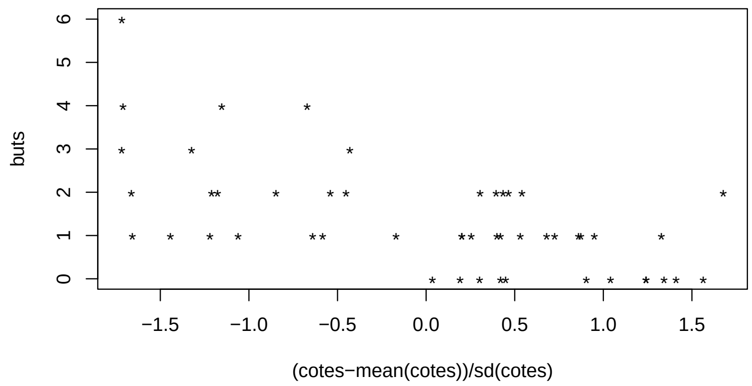
Université de Montpellier　　　　　　　　　　　　　　　　　　　　Année 2021-2022
HAX912X - Modèles linéaires généralisés

Examen final

Durée 2h - Sans document - 27 octobre 2021

Exercice 1 (4 pts)

Soit Y une variable de comptage correspondant au nombre de buts marqués à domicile par des équipes de football et x la cote de la victoire de l'équipe jouant à domicile. Les observations sont représentées dans le graphe ci-dessous.



Proposer un modèle linéaire généralisé permettant de prédire Y en fonction de x.

Exercice 2 (4 pts)

1) (1 pt) Expliquer l'objectif d'une modélisation log-linéaire graphique ?

2) (1 pt) Représenter le modèle log-linéaire graphique associé à l'équation suivante

$$x_1 * x_2 * x_3 + x_1 * x_4 + x_4 * x_5 .$$

Donner deux relations d'indépendance conditionnelle entre les variables qui le composent.

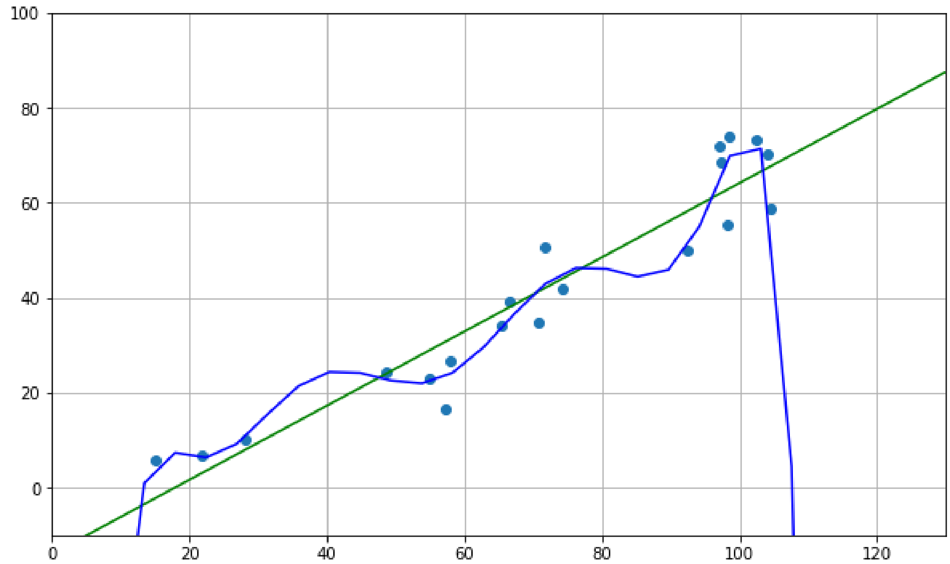
3) (2 pts) Nous considérons 100 réalisations suivant deux variables qualitatives ayant chacune deux modalités. Donner la vraisemblance associée au modèle log-linéaire graphique poissonien d'indépendance. Montrer qu'il s'agit bien d'un modèle linéaire généralisé.

Exercice 3 (6 pts)

Décrire et commenter les résultats obtenus à l'aide du code Python ci-dessous.

```
import numpy as np
import sklearn.linear_model as lm
from sklearn.preprocessing import PolynomialFeatures
import sklearn.metrics as metrics
from math import sqrt
```

```
import matplotlib.pyplot as plt

data = np.loadtxt("data.txt")
data.shape
# Out[]: (20, 2)

X_train = data[:,0].reshape(len(data),1)
Y_train = data[:,1].reshape(len(data),1)

model1 = lm.LinearRegression(normalize=True)
model1.fit(X_train, Y_train)
X = np.linspace(0,130,num=30).reshape(30,1)
Y_pred1 = model1.predict(X)

poly = PolynomialFeatures(degree=8,include_bias=False)
X_train_new = poly.fit_transform(X_train)
model2 = lm.LinearRegression(normalize=True)
model2.fit(X_train_new,Y_train)
X_new = poly.fit_transform(X)
Y_pred2 = model2.predict(X_new)

plt.figure(figsize=(10,6))
plt.plot(X_train, Y_train,'o')
plt.plot(X, Y_pred1, '-g')
plt.plot(X, Y_pred2, '-b')
plt.xlim(0, 130)
plt.ylim(-10, 100)
plt.grid()
```




```
sqrt(metrics.mean_squared_error(Y_train,model1.predict(X_train)))
# Out[]: 6.997902295162095

sqrt(metrics.mean_squared_error(Y_train,model2.predict(X_train_new)))
# Out[]: 4.780839743089144
```

```python
from sklearn.model_selection import cross_val_score

-cross_val_score(model1, X_train, Y_train, cv = 20,
scoring ='neg_root_mean_squared_error').mean()
# Out[]: 6.774129056767689

-cross_val_score(model2, X_train_new, Y_train, cv = 20,
scoring ='neg_root_mean_squared_error').mean()
# Out[]: 19.446094299067326

from sklearn.model_selection import GridSearchCV

parameters = {'alpha':np.linspace(0,1,num=101)}
model3 = GridSearchCV(lm.Ridge(normalize=True), parameters,
scoring='neg_root_mean_squared_error',cv=20)
model3.fit(X_train_new,Y_train)

model3.best_params_
# Out[]: {'alpha': 0.03}

model3 = model3.best_estimator_
sqrt(metrics.mean_squared_error(Y_train,model3.predict(X_train_new)))
# Out[]: 6.224898455388499

-cross_val_score(model3, X_train_new, Y_train, cv = 20,
scoring ='neg_root_mean_squared_error').mean()
# Out[]: 5.659471737238336
```

Exercice 4 (4 pts)

On considère un n-échantillons $(Y_1, x_1), \dots, (Y_n, x_n)$ du couple (Y, x) où Y est une variable aléatoire à valeurs dans $\{0, 1\}$ et x est une variable fixée. On suppose qu'il existe une variable latente Y_i^* telle que $Y_i^*|x_i \sim \mathcal{L}(a + bx_i)$ et

$$Y_i = \begin{cases} 1 & \text{si} \quad Y_i^* > 0 \\ 0 & \text{si} \quad Y_i^* \leq 0 \end{cases} .$$

$\mathcal{L}(\mu)$ est la loi de Laplace de moyenne μ ayant pour densité

$$f(y; \mu) = \frac{1}{2} \exp(-|y - \mu|) .$$

1 (2 pts) Montrer qu'il s'agit d'un modèle linéaire généralisée.

2 (2 pts) Donner les équations de vraisemblance.

Exercice 5 (2 pts)

On considère le modèle de régression suivant, pour tout $i = 1, \dots, n$

$$Y_i = \theta x_i + U_i$$

avec $\mathbb{E}(U_i) = 0$, $\mathbb{V}(U_i) = \sigma_i^2$, $\mathbb{C}(U_i, U_j) = 0$ si $i \neq j$, x_i et θ des réels. Donner l'expression de l'estimateur des moindres carrés généralisés de θ.

Correction examen final HAX912X
Modèles linéaires généralisés
27 octobre 2021

Exercice 1

$Y =$ "nombre de buts marqués à domicile"

$x =$ "cote de la victoire de l'équipe jouant à domicile"

On cherche un modèle caractérisant la relation existant entre Y et x

Y variable aléatoire
x variable fixée

Nous pouvons utiliser un modèle de régression Poissonienne

$$Y \mid x \sim \mathcal{P}(\lambda(x))$$

où $\lambda(x) = \exp\{\alpha \tilde{x} + \beta\}$

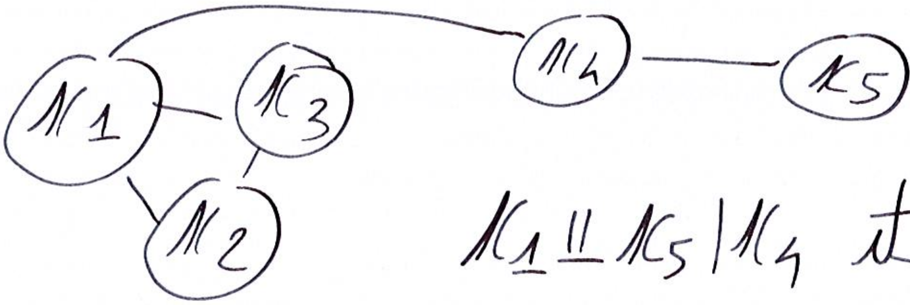
où $\tilde{x} = \left[\frac{x - \bar{x}}{\sqrt{\mathbb{V}e(x)}}\right]$

$\tilde{x}$ est la cote centrée réduite empiriquement.

Exercice 2

1) Un modèle log-linéaire graphique n'est pas un modèle de régression mais un modèle d'association. Il permet d'inférer les relations entre plusieurs variables qualitatives.

2]



$X_1 \perp\!\!\!\perp X_5 \mid X_4$ et $X_3 \perp\!\!\!\perp X_4 \mid X_1$

3] Table de contingence avec $J=4$ cases

$X_1 \in \{1,2\}$ et $X_2 \in \{1,2\}$

Soit $\mu_{i,j}$ l'effectif moyen associé à la case (i,j) et $n_{i,j}$ la valeur observée.

Modèle d'indépendance

$$\log(\mu_{i,j}) = \mu + \alpha_{1,i} + \alpha_{2,j}$$

et $\alpha_{1,1} = -\alpha_{1,2}$ et $\alpha_{2,1} = -\alpha_{2,2}$

Modèle Poissonien

$$n_{i,j} \sim \mathcal{P}(\mu_{i,j})$$

loi de Poisson appartient à la famille exponentielle scalaire et

$$\mathbb{E}(n_{i,j}) = \exp(\mu + \alpha_{1,i} + \alpha_{2,j})$$

C'est bien un modèle linéaire généralisé.

Vraisemblance

$$f(n_{1,1}, n_{1,2}, n_{2,1}, n_{2,2} \mid \mu, \alpha_{1,1}, \alpha_{2,1}) = \prod_{i=1}^{2} \prod_{j=1}^{2} \left[\frac{\mu_{i,j}^{n_{i,j}} e^{-\mu_{i,j}}}{n_{i,j}!} \right]$$

où $\mu_{i,j} = e^{\mu + \alpha_{1,i} + \alpha_{2,j}}$

Exercice 3

$n = 20$, $y = (y_1, \dots, y_{20})$, $x = (x_1, \dots, x_{20})$

→ Mise en œuvre d'un modèle de régression linéaire simple [model 1]

variable à expliquer y

variable explicative $\tilde{x} = \dfrac{x - \bar{x}}{\sqrt{Var(x)}}$

Prédictions pour 30 valeurs régulièrement espacées entre 0 et 130

→ Mise en œuvre d'un modèle de régression polynomiale jusqu'au degré 8 [model 2]

variable à expliquer y

variables explicatives

$\tilde{x}, \tilde{x^2}, \tilde{x^3}, \tilde{x^4}, \tilde{x^5}, \tilde{x^6}, \tilde{x^7}, \tilde{x^8}$

Prédictions pour 30 valeurs de x régulièrement espacées entre 0 et 130

→ Sur l'échantillon d'apprentissage model 2 est meilleur que model 1 sur-apprentissage (voir graphe)

→ Comparaison de model 1 et model 2 par validation à 20 ensembles (leave-one-out cross validation) model 1 est nettement meilleur

[model 3]

→ Mise en oeuvre d'un modèle de régression ridge où le paramètre de régularisation est sélectionné par leave-one-out cross validation, régression ridge sur les variables explicatives polynomiales

(6)

→ Par validation croisée on voit que le modèle ridge est le meilleur.

Exercice 4

1] $y \sim \mathcal{B}(p)$

$$f(y|p) = p^{y}(1-p)^{1-y}\,\mathbb{1}_{\{0,1\}}(y)$$

$$f(y|p) = \exp\left\{y\log\left(\frac{p}{1-p}\right) + \log(1-p)\right\}\mathbb{1}_{\{0,1\}}(y)$$

$$\theta = \log\left[\frac{p}{1-p}\right] \Rightarrow p = \frac{e^{\theta}}{1+e^{\theta}}$$

$$b(\theta) = \log(1+e^{\theta})$$

La loi de Bernoulli appartient à la famille exponentielle scalaire.

Par ailleurs, $\mathbb{P}(y=1|x) = \mathbb{P}(y^{*}>0|x)$

7

$$\Leftrightarrow \mathbb{P}(y=1|x) = \int_0^{+\infty} \frac{1}{2} \exp\left[-|y-a-bx|\right] dy$$

$$\Leftrightarrow \mathbb{P}(y=1|x) = \int_{-a-bx}^{+\infty} \frac{1}{2} \exp\left[-|y|\right] dy$$

$$\Leftrightarrow \mathbb{P}(y=1|x) = 1 - F_{\mathcal{L}(0)}(-a-bx)$$

$$\Leftrightarrow \mathbb{P}(y=1|x) = F_{\mathcal{L}(0)}(a+bx)$$

où $F_{\mathcal{L}(0)}(\cdot)$ est la fonction de répartition de la loi de Laplace-centrée

Ainsi $\mathbb{E}(y|x) = \mathbb{P}(y=1|x)$

$$= F_{\mathcal{L}(0)}(a+bx) = f(a+bx)$$

Il s'agit bien d'un modèle linéaire généralisé.

2)

$$\mathcal{L}V(a,b) = \sum_{i=1}^{n} \left[y_i \log\left(F_{\mathcal{L}(0)}(a+bx_i)\right) + (1-y_i) \log\left(1-F_{\mathcal{L}(0)}(a+bx_i)\right)\right]$$

$$\frac{dLV}{da}(a,b) = \sum_{i=1}^{n} y_i \left[\frac{1}{2} \exp\left(-\left|a+bx_i\right|\right)\right] \Big/ F_{L(0)}(a+bx_i)$$

$$- \sum_{i=1}^{n} (1-y_i)\left[\frac{1}{2} \exp\left(-\left|a+bx_i\right|\right)\right] \Big/ \left(1-F_{L(0)}(a+bx_i)\right)$$

$$\frac{dLV}{da}(a,b) = \sum_{i=1}^{n} y_i \frac{\frac{1}{2}\exp(-|a+bx_i|)}{F_{L(0)}(a+bx_i)\left(1-F_{L(0)}(a+bx_i)\right)}$$

$$- \sum_{i=1}^{n} \frac{\frac{1}{2}\exp(-|a+bx_i|)}{\left(1-F_{L(0)}(a+bx_i)\right)}$$

$$\frac{dLV}{db}(a,b) = \sum_{i=1}^{n} y_i x_i \frac{\frac{1}{2}\exp(-|a+bx_i|)}{F_{L(0)}(a+bx_i)\left(1-F_{L(0)}(a+bx_i)\right)}$$

$$- \sum_{i=1}^{n} x_i \frac{\frac{1}{2}\exp(-|a+bx_i|)}{\left(1-F_{L(0)}(a+bx_i)\right)}$$

Ερώτηση 5

$$y_i = \theta x_i + u_i$$

$$V(y_i) = V(u_i) = \sigma_i^2$$

$$C(y_i, y_j) = C(u_i, u_j) = 0 \quad \forall i \neq j$$

$$\hat{\theta}^{\Pi C \omega} = \sum_{i=1}^{n}\left[y_i \frac{x_i}{\sigma_i^2}\right] \Big/ \sum_{i=1}^{n}\left[\frac{x_i^2}{\sigma_i^2}\right]$$