ING2-MF
Examen De Modèles Linéaires 2019-2020: Durée 2h

Rendu : vous devez déposer sur Arel un document pdf obtenu par scan (lisible) de vos copie ou photo (lisible de bonne qualité) avec numérotation des pages. Le nom de votre fichier pdf doit être de la forme NOM-PRENOM et n'oubliez pas de mettre vos noms et prénoms en entête de votre document.
Tout soupçon de triche m'oblige à faire passer le fichier pdf par un logiciel anti-plagiat qui reconnaît du copier-coller du raisonnement.
J'attacherai aussi beaucoup d'importance à l'analyse que vous allez faire de vos résultats.

Exercice 1. On considère le jeu de données suivant et la matrice des corrélations:

Y	X_1	X_2
17	6	9
12	5	10
14	5	11
13	5	11

corr() :	X_1	X_2
X_2	-0.87	
Y	0.92	-0.64

$$\text{Soit} \qquad Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, i = 1, 2, 3, 4, \quad \text{où} \quad \epsilon_i \hookrightarrow \mathcal{N}(0, \sigma^2), \tag{1}$$

le modèle de régression permettant d'expliquer Y en fonction de X_1 et X_2. Dans ce modèle et dans la suite les variables X_1 et X_2 sont observées et non aléatoires. On donne

$$(X'X)^{-1} = \begin{pmatrix} 578.5 & -54.5 & -28.5 \\ -54.5 & 5.5 & 2.5 \\ -28.5 & 2.5 & 1.5 \end{pmatrix} \quad \text{et} \quad (X'X)^{-1}X' = \begin{pmatrix} -5 & 21 & -7.5 & -7.5 \\ 1 & -2 & 0.5 & 0.5 \\ 0 & -1 & 0.5 & 0.5 \end{pmatrix}$$

où X représente la matrice de schéma du modèle et X' sa transposée.

1) Donnez la forme explicite de X et l'expression matricielle du modèle (1).

2) Donnez l'expression du vecteur de paramètres estimé $\widehat{\beta}$ et calculez-le. Que se passe-t-il si la matrice X n'était pas de plein rang ? Justifier vos réponses.
3) Écrire l'équation du modèle et donnez l'augmentation (en moyenne) prédite de Y, si X_1 augmente d'une unité, la variableX_2 restant constante.

4) Complétez le tableau d'analyse de variance suivant (en le recopiant sur votre copie). Justifiez vos calculs

Source de variation	SC=somme des carrés	d.l.l.	CM=Carrés Moyens	F
Régression	SCE=	.....	.....	F= .....
Résiduelle	SCR=.....	.....	.....	
Total	SCT=.....	.....		

5) Calculez le coefficient de détermination et interpréter le résultat.

6) Le tableau suivant donne des résultats partiels de la régression de Y sur X_1 et X_2:

```
lm(formula = Y ~ X1 + X2)
Coefficients:
              Estimate   Std.Error   t-value   Pr(>|t|)
(Intercept)    ....       .....       .....     0.284
X1             ....       .....       ....      0.159
X2             .....      .....       .....     0.333
Residual standard error: 0.7071 on 1 degrees of freedom
F-statistic:    13.5 on  2 and 1 DF,    p-value: 0.189
```

Interprétez ces résultats. Quelle est votre conclusion ?

Exercice 2. Les données sont celles de l'exercice 1. Une régression de la variable Y sur X_1 a donné le résultat suivant:

```
Coefficients:
              Estimate    Std. Error t-value   Pr(>|t|)
(Intercept)   -7.000      6.083    -1.151      0.3688
X1             4.000      1.155     3.464       0.0742
Residual standard error: 1 on 2 degrees of freedom
```

1) Écrire le modèle obtenu. Comment est mesurée la variabilité des estimateurs ?

2) Donnez l'augmentation (en moyenne) prédite de Y, si X_1 augment d'une unité.
Comment expliquez-vous qu'il y ait une différence entre ce résultat est le résultat obtenu dans la question 3) de l'exercice 1 ?

3) Au vu des résultats donnés au tableau précédent, on propose d'étudier le modèle suivant reliant la variable Y avec X_1:

$$Y_i = bX_{1i} + \epsilon_i, \qquad 1 \leq i \leq n \tag{2}$$

où les ϵ_i sont telles que $E[\epsilon_i] = 0$, $Var(\epsilon_i) = \sigma^2$ et $Cov(\epsilon_i, \epsilon_j) = 0$ pour $i \neq j$.
Justifier l'introduction de ce modèle.

Les résultats de ce nouveau modèle sont fournis dans le tableau suivant:

```
Coefficients:
           Estimate   Std. Error  t-value    Pr(>|t|)
X1         2.67568    0.09991     26.78      0.000114 ***

Residual standard error: 1.053 on 3 degrees of freedom
Multiple R-squared:  0.9958,    Adjusted R-squared:  0.9944
F-statistic: 717.1 on 1 and 3 DF,  p-value: 0.0001143
```

4) Écrire ce modèle sous forme matricielle et en déduire de la question 2) de l'exercice 1 l'expression de l'estimateur des moindres carrés $\widehat{b}$ en fonction des données $(X_{i1},\ Y_i)_{1\leq i\leq n}$.

5) Quelle est la valeur prédite de Y_0 pour une valeur donnée X_0. Quelle est la variance de l'erreur de prédiction ?
6) Donnez l'expression de la pente b^* de la droite d'équation $Y = b^*X_1$ passant par le centre de gravité du nuage de points $(X_{i1}, Y_i)_{1\leq i\leq n}$.

7) Montrer que $\widehat{b}$ et b^* sont tous deux des estimateurs sans biais de b.
8) Montrer que $V(b^*) > V(\widehat{b})$ sauf dans le cas où tous les X_{i1} sont égaux. Comment interprétez-vous ce résultat ? Ce résultat était-il prévisible ?

Exercice 3. Le jeu de données contient des données médicales de 120 patients atteints ou non d'une maladie du système urinaire. Les variables explicatives sont :

- Temp : Temperature of patient $\{35C - 42C\}$

- Nausea : Occurrence of nausea $\{yes, no\}$

- Lumbar : Lumbar pain $\{yes, no\}$

- Urine : Urine pushing (continuous need for urination) $\{yes, no\}$

- Mic : Micturition pains $\{yes, no\}$

- Uret : Burning of urethra, itch, swelling of urethra outlet $\{yes, no\}$

La variable cible est :

- Diag : Diagnosis of Nephritis of renal pelvis origin $\{yes, no\}$

Temp	Nausea	Lumbar	Urine	Mic	Uret	Diag
35.5	no	yes	no	no	no	no
35.9	no	no	yes	yes	yes	no
35.9	no	yes	no	no	no	no
36	no	no	yes	yes	yes	no
36	no	yes	no	no	no	no
...	...	...	...	...	...	

Extrait du jeu de données

1) Dans un premier temps, nous allons construire un modèle de prévision à partir de la température. Les résultats pour la régression logistique effectuée avec R sont donnés dans la figure (1).

1.1) Écrire le modèle.

1.2) La prévision des individus de la base d'apprentissage avec ce modèle donne le graphique de la figure 2.

Quelle est la règle obtenue ? Expliquer comment vous l'obtenez

1.3) Calculer l'odds en fonction de la température. En déduire l'odds d'une personne ayant 38.2. Qu'est-ce que cela signifie ?

2. On ajuste maintenant un modèle avec uniquement la variable explicative Uret.

2.1) Écrire le modèle.

2.2) Donner le diagnostic obtenu avec ce modèle pour le 1er patient de l'extrait du jeu de données. Expliciter vos calculs.

2.3) Calculer l'odds en fonction de Uret. Interpréter les résultats.

2.4) Calculez un intervalle de confiance pour l'odds-ration

Figure 1

```
Call:
glm(formula = Diag ~ Temp, family = "binomial", data = Mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3539  -0.4725  -0.2792   0.5694   1.8583

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -47.5295     7.2809  -6.528 6.67e-11 ***
Temp          1.2105     0.1856   6.522 6.95e-11 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 163.006  on 119  degrees of freedom
Residual deviance:  85.863  on 118  degrees of freedom
AIC: 89.863

Number of Fisher Scoring iterations: 5
```

Figure 2