EISTI -Département Mathématiques
Examen de rattrapage Modèle Linéaire : MI
Décembre 2017- durée 2h00

La consultation et l'échange de documents et de calculatrices sont interdits. Deux feuilles manuscrites recto-verso format A4 et les calculatrices sont autorisées.

Exercice 1. On considère le modèle

$$y_i = \beta x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

où les ϵ_i sont des variables aléatoires centrées, de moyenne nulle, non corrélées et de variance constante.

1. En utilisant la méthode des moindres carrés, montrer que l'estimateur de β vaut

$$\widehat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.$$

2. Montrer que la droite de régression passe par l'origine et le centre de gravité $(\overline{x}, \overline{y})$ du nuage de points est $y = \beta^* x$ avec

$$\beta^* = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}$$

3. Vérifier que $\widehat{\beta}$ et β^* peuvent réécrire sous la forme:

$$\widehat{\beta} = \beta + \frac{\sum_{i=1}^n x_i \epsilon_i}{\sum_{i=1}^n x_i^2} \quad \text{et} \quad \beta^* = \beta + \frac{\sum_{i=1}^n \epsilon_i}{\sum_{i=1}^n x_i}. \tag{0.1}$$

En déduire que $\widehat{\beta}$ et β^* sont tous deux des estimateurs sans bias de β.

4. En utilisant l'inégalité de Cauchy-Schwarz

$$\forall\ u, v \in \mathbb{R}^n : \quad \left(\sum_{i=1}^n u_i v_i\right)^2 \leq \left(\sum_{i=1}^n u_i^2\right)\left(\sum_{i=1}^n v_i^2\right),$$

montrer que $V(\beta^*) \geq V(\widehat{\beta})$ sauf dans le cas où tous les x_i sont égaux. Ce résultat était-il prévisible ?

Exercice 2. Considérons le jeu de données sur le ronflement.

AGE	POIDS	TAILLE	ALCOOL	SEXE	TABAC	RONFLE
47	71	158	A	1	1	0
56	58	164	I	0	0	1
46	116	208	M	1	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮

L'objectif est d'expliquer le ronflement d'une personne en fonction des autres caractéristiques.

- RONFLE est la variable cible Y. C'est une variable binaire telle que RONFLE=1 si la personne ronfle et RONFLE=0 sinon
- AGE$\in [20, 80]$, POIDS $\in [42, 120]$, TAILLE $\in [158, 208]$ sont des variables explicatives quantitatives
- SEXE (H=1/F=0), TABAC (oui=1/non=0) sont des variables qualitatives à deux modalités

Si dessous les résultats obtenus pour les différents modèles de régression logistique à une variable explicative.

	variable explicative	$\widehat{a}_0$	$\widehat{\sigma}_0$	$\widehat{a}_1$	$\widehat{\sigma}_1$	Déviance
1	SEXE (H=1)	0.20000	0.09476	0.20000	0.10942	22.00
2	AGE	-0.180638	0.220530	0.010152	0.004123	21.42

1) Modèle 1

a. Explicitez le modèle 1 avec Y=RONFLE et X=SEXE.

b. Calculez l'odds des hommes et l'odds des femmes. Faites une phrase pour expliquer ce que signifient les résultats.

c. Comment interprétez-vous $\widehat{a}_1$?

d. Quelle est l'augmentation du risque des hommes par rapport aux femmes ?

e. Calculez un intervalle de confiance pour l'odds-ratio (le quantile 0.95 de $\mathcal{N}(0, 1)$ vaut 1.96)

2) Modèle 2 : AGE

a. Explicitez le modèle 2 avec Y=RONFLE et X=AGE.

b. Quelle est la probabilité de ronfler pour une personne de 40 ans.

c. Calculez l'odds en fonction de l'âge. En déduire l'odds pour une personne de 80 ans. Qu'est-ce cela signifie (Faites une phrase) ?

d. Comment interprétez-vous $\widehat{a}_1$?

e. Quel est le rapport de chance de ronfler quand on passe de 20 à 30 ans ou de 20 à 60 ans ? Pensez-vous qu'il y a une différence importante ?

f. Afin de confirmer ou infirmer votre conclusion, effectuez un test pour déterminer si a_1 est significativement non nul.

Exercice 3. Un laboratoire d'analyse sensorielle souhaite examiner dans quelle mesure l'appréciation globale d'un cocktail de jus de fruit peut être expliquée par ses saveurs (saveur acide, amère, sucré) et son caractère pulpeux. Un jury d'experts a évalué les saveurs et le caractère pulpeux de 16 cocktails; d'autre part, un jury de consommateurs a noté son degré d'appréciation de ces mêmes cocktails à l'aide d'une échelle de note allant de 0 à 10. Les moyennes des évaluations des experts et des consommateurs ont été calculées pour chacun des cocktails. Le tableau ci-dessous donne un aperçu du début de ces moyennes

	X_1	X_2	X_3	X_4	Y
Produit	Sucre	Acide	Amer	Pulpeux	Appréciation
1	6.21	7.08	2.00	2.54	4.97
2	7.75	3.29	1.54	2.26	6.98
⋮	⋮	⋮	⋮	⋮	⋮

1) Le tableau 1 (en annexe) résume les corrélations entre les différentes variables. Commenter ce tableau.

2) Le tableau 2 résume l'analyse de la variance du modèle de régression linéaire exprimant l'appréciation globale en fonction des 4 variables explicatives que sont les saveurs et la variable pulpeux.

a) Compléter ce tableau

b) Tester l'hypothèse de nullité simultanée des 4 coefficients au seuil $\alpha = 5\%$, sachant que le quantile 0.95 de la loi de Fisher "adapté" vaut 3.357.

3) L'estimation des paramètres du modèle par la méthode des moindres carrés a été effectuée d'une part sur les variables brutes et d'autres sur les variables centrées réduites. Les résultats de ces deux régressions sont données dans le tableau 3.

a) Qu'apportent les résultats obtenus sur les données centrées-réduites ?

b) Quelles sont les variables (ou la variable) les plus influentes sur l'appréciation ?

c) Quel est le test réalisé pour déterminer le caractère significatif d'un coefficient β_j ? Préciser l'hypothèse nulle, l'hypothèse alternative, la statistique de test, la loi de celle-ci sous H_0 et la règle de décision.

d) Quels sont les coefficients significatifs (dans le modèle complet) au niveau de confiance 95% ? Comparer avec la question 2 b) et commenter ? Que suggérez-vous ?

4) On propose dans cette question de procéder au choix d'un meilleur modèle. Plusieurs modèles ont été proposés en annex, tableau 4.

a) Quel modèle vous paraît pertinent pour expliquer l'appréciation ? Justifier votre réponse en expliquant la méthode utilisée.

b) A quoi ça sert le graphique 1 ?

5) On considère maintenant uniquement la variable acide pour expliquer l'appréciation.

a) Écrire le modèle de régression linéaire correspondant en explicitant les différentes hypothèses qui sont faîtes dans ce modèle.

b) Quelle est l'estimation du coefficient de la variable acide dans le modèle? Comment interprétez-vous ce coefficient ?

c) Commenter la qualité d'ajustement des données au modèle et donner une estimation du coefficient de corrélation r entre l'appréciation et l'acide.

d) Tester les hypothèses $H_0 : r = 0$ contre $H_1 : r \neq 0$.

e) Calculer un intervalle de confiance pour la valeur moyenne de Y_0 au niveau 95% connaissant $x_0 = 6.39$, le quantile 0.95 de la loi de Student à 14 degrés de liberté $t_{\alpha/2}(n-2) = 1.761$, $\overline{x} = 5.656875$ et $\sum_{i=1}^{16}(x_i - \overline{x})^2 = 116.7696$.

Annexe

Tableau 1: Matrice de corrélation

	Sucre	Acide	Amer	Pulpeux	Appréciation
Sucre	1.00				
Acide	-0.91	1.00			
Amer	-0.86	0.86	1.00		
Pulpeux	0.74	-0.57	-0.48	1.00	
Appréciation	0.74	-0.83	-0.69	0.43	1.00

Tableau 2: Tableau d'analyse de la variance

Source de Variabilité	Somme des carrés des écarts	Degrés de liberté	Carrés Moyens
Expliqué par le modèle	11.607	...	...
Non expliquée (résiduelle)	...	...	...
Totale	16.877	...	

Tableau 3: Estimation des paramètres de la régression

```
Résultats des régression sur des données brutes
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.5091     4.8970   1.533   0.1534
Sucre         0.1006     0.5589   0.180   0.8605
Acide        -0.4973     0.2624  -1.895   0.0846 .
Amer          0.3869     1.3644   0.284   0.7820
Pulpeux      -0.2816     0.8537  -0.330   0.7477

Résultats des régression sur des données centrées réduites
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.574e-16  1.630e-01   0.000   1.0000
Sucre        1.135e-01  6.312e-01   0.180   0.8605
Acide       -8.699e-01  4.590e-01  -1.895   0.0846 .
Amer         1.072e-01  3.779e-01   0.284   0.7820
Pulpeux     -9.632e-02  2.920e-01  -0.330   0.7477
```

Table 4: Choix de modèle

```
MODÈLE 1:
lm(formula = Appréciation ~ ., data = Cocktail)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.5091     4.8970   1.533   0.1534
Sucre         0.1006     0.5589   0.180   0.8605
Acide        -0.4973     0.2624  -1.895   0.0846 .
Amer          0.3869     1.3644   0.284   0.7820
Pulpeux      -0.2816     0.8537  -0.330   0.7477

MODÈLE 2:
lm(formula = Appréciation ~ Acide + Pulpeux + Amer, data = Cocktail)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.2913     2.1605   3.838  0.00236 **
Acide        -0.5279     0.1918  -2.752  0.01753 *
Pulpeux      -0.1720     0.5735  -0.300  0.76942
Amer          0.2653     1.1365   0.233  0.81935

MODÈLE 3:
lm(formula = Appréciation ~ Acide + Pulpeux, data = Cocktail)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.6007     1.6430   5.235 0.000161 ***
Acide        -0.4915     0.1080  -4.552 0.000543 ***
Pulpeux      -0.1696     0.5522  -0.307 0.763562

MODÈLE 4:
lm(formula = Appréciation ~ Acide, data = Cocktail)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.12265    0.51000  15.927 2.30e-10 ***
Acide       -0.47268    0.08593  -5.501 7.81e-05 ***
Residual standard error: 0.6174 on 14 degrees of freedom
Multiple R-squared:  0.6837,    Adjusted R-squared:  0.6611
F-statistic: 30.26 on 1 and 14 DF,  p-value: 7.81e-05
```

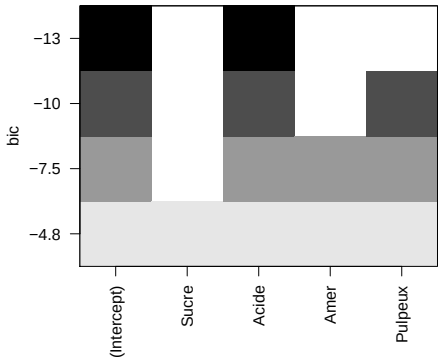

Figure 1: Choix de variables avec le critère de BIC

Solution

Exercice 1.

1) Par définition, l'estimateur des moindres carrés de β vérifié

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta x_i)^2 = \arg\min_{\beta} S(\beta).$$

La fonction S est strictement convexe et admet donc un unique minimum au point où sa dérivée s'annule:

$$S'(\beta) = -2\sum_{i=1}^{n} x_i(y_i - \beta x_i) = 2\left[\beta \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i\right].$$

Ce qui implique

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

2) La droite passant par l'origine et centre de gravité $(\overline{x}, \overline{y})$ du nuage de points admet pour équation $y = \beta^* x$ où

$$\beta^* = \frac{\overline{y}}{\overline{x}} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$$

3) On réécrire les estimateurs obtenus sous la forme:

Par hypothèses sur les erreurs on obtient d'après (0.1) que:

$$\mathbb{E}(\widehat{\beta}) = \mathbb{E}(\beta^*) = \beta.$$

4) On a

$$V(\widehat{\beta}) = \frac{\sum_{i=1}^{n} x_i^2 \sigma^2}{(\sum_{i=1}^{n} x_i^2)^2} = \frac{\sigma 2}{(\sum_{i=1}^{n} x_i^2)}.$$

De même

$$V(\beta^*) = \frac{n\sigma 2}{(\sum_{i=1}^n x_i^2)^2}.$$

En utilisant l'inégalité de Cauchy-Schwarz

$$\forall\ u, v \in \mathbb{R}^n : \quad \left(\sum_{i=1}^n u_i v_i\right)^2 \leq \left(\sum_{i=1}^n u_i^2\right)\left(\sum_{i=1}^n v_i^2\right)$$

appliquée à $u = (x_1, \ldots, x_n)^t$ et $v = (1, \ldots, 1)^t$, on obtient résultat.

L'égalité a lieu si et seulement si u et v sont colinéaires, i.e. si tous les x_i sont égaux. Puisque les deux estimateurs sont linéaires en y et que $\widehat{\beta}$ est celui des moindres carrés qui est optimal, donc ce résultat n'est pas étonnant.

Exercice 3

1) On remarque que les prédicteurs sont corrélés, leurs coefficients dans le modèle ne sont pas aisément interprétables.

2 a) Tableau 2: analyse de la variance

Source de	Somme des carrés	Degrés de	Carrés
Variabilité	des écarts	liberté	Moyens
Expliqué par le modèle	11.607	**4**	**2.902**
Non expliquée (résiduelle)	**5.270**	**11**	**0.479**
Totale	16.877	**15**	

2 b) Pour tester l'hypothèse de nullité simultanée des 4 coefficients au seul $\alpha = 5\%$, on calcule la valeur de la statistique de test $f_{\text{obs}} = \frac{CM_{\text{modele}}}{CM_{\text{résiduelle}}} = 6.058$. Cette valeur étant supérieure au quantile 0.95 de la loi de Fisher à 4 et 11 degrés de liberté (3.357), on rejette l'hypothèse H_0 et on considère qu'au moins un coefficient est non nul.

3. Le fait de centrer et réduire les données ne modifié en rien la pertinence statistique du modèle: toutes les probabilités critiques sont inchangées, exceptée celle relative à la constante du modèle puisque cette dernière est nulle par construction.

a) L'intérêt du "centrage-réduction" est de rendre les coefficients $\widehat{\beta}_j$ comparables entre eux.

b) les variables les plus influentes sont celles qui ont un coefficient élevé. On considérera donc que c'est la variable acide qui influe le plus sur l'appréciation.

c) Le test d'un paramètre se construit de même façon que le test de β en régression simple. Les hypothèses testées sont $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$.

La statistique de test est: $\frac{\widehat{\beta}_j}{\widehat{\sigma}_{\widehat{\beta}_j}}$.

Sous l'hypothèse nulle H_0, cette statistique suit une loi de Student à $n-p-1$ degrés de liberté (i.e. les degrés de liberté de la résiduelle), soit 11 dans cet exemple.

Règle de décision: si la valeur absolue de la statistique de test est supérieure au quantile 0.975 de la loi de Student à 11 degrés de liberté, alors on rejette l'hypothèse H_0, sinon, on l'accepte. En pratique, il est plus facile de comparer la probabilité critique au seuil α choisi.

d) Au seuil $\alpha = 5\%$, aucun coefficient n'est significatif. Ceci est une contradiction avec la réponse de la question 2 b).

Commentaire:

Pris séparément, les tests de nullité d'un coefficient montrent qu'aucun coefficient n'est significativement différent de 0 au seuil 5%. Ceci est en apparente contradiction avec le test de nullité de tous les coefficients (test F) puisqu'on avait considéré qu'au moins un coefficient était non nul. Mais, dans le test d'un coefficient β_j, on teste si ce coefficient nul, toutes les variables étant présentes dans le modèle. L'hypothèse testée peut se lire "la variable j n'apporte pas d'information complémentaire pour expliquer Y, i.e., une information non déjà expliquée par les autres variables". Ainsi, dans l'exemple, chaque prédicteur n'a pas d'apport significatif compte tenu de la présence de tous les autres.

On peut alors se demander s'il ne serait pas judicieux de se limiter à un sous ensemble de prédicteurs.

4).

a) D'après la méthode pas à pas descendante le meilleur modèle est celui avec une seule variable "acide"

b) Le graphique est le résultat du choix automatique de la fonction regsubsets de R, laquelle est basée sur le critère de BIC qui consiste à choisir le modèle qui minimise la quantité BIC. On en déduit d'après le graphique que le meilleur modèle est celui qui contient la variable "acid".

5)

a) Le modèle s'écrit

$$\forall\ i \quad y_i = \beta_0 + \beta x_i + \epsilon$$

où y_i est la note d'appréciation du cocktail i, x_i son saveur et son caractère pulpeux. Les hypothèses sont: les erreurs gaussiennes, centrées, indépendantes et de même variance:

$$\epsilon_i \hookrightarrow \mathcal{N}(0,\ \sigma^2 I_{16}).$$

b) L'estimation de la pente de la droite de régression est : $\widehat{\beta} = -0.47268$. Ce coefficient est significativement différent de 0 (p-valeur ~ 0) et s'interprète comme suit: la note de degré

d'appréciation d'un cocktail diminuera en moyenne de 0.47268 pour une augmentation de sa saveur d'acidité et de son caractère pulpuex de 1

c) Le coefficient de de détermination R^2, qui s'interprète comme le pourcentage de variabilité expliquée, vaut 0.6837: l'influence est donc importante. Le coefficient de corrélation $r = \sqrt{R^2} = \sqrt{0.6837} = 0.8268615$.

d) Le test proposé est équivalente à tester $\beta = 0$ qui s'en déduit directement de la p-valeur donnée dans la sortie du R

e) On a:

$$\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta} x_0 = 8.12265 - 0.47268 \times 6.39 = 5,102225.$$

$$IC(\mathbb{E}(Y_0)) = [\widehat{Y}_0 - z, \widehat{Y}_0 + z] = [5.102225 - z; \quad 5.102225 - z]$$

$$z = t_{\alpha/2} s_\sigma \sqrt{\frac{1}{16} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{16}(x_i - \overline{x})^2}}$$

avec $s_\sigma = 0.6174 \quad t_{\alpha/2}(n-2) = 1.761$, on obtient

$$z = 1.761 \times 0.6174 \times 0.04602 = 0.059366 \quad \text{et} \quad IC(\mathbb{E}(Y_0)) = [5.042859, \quad 5.161591]$$