
MATH 60602 *Analyse multidimensionnelle appliquée*

Examen de pratique

Questionnaire

Examineur: Léo Belzile

Instructions: L'examen est d'une durée de 180 minutes. Aucune documentation écrite n'est permise. L'utilisation d'un ordinateur ou de tout autre matériel électronique est interdit. Une calculatrice non programmable est autorisée.

La répartition des 50 points de l'examen se trouve dans la marge de droite.

Les questions avec une étoile (★) demandent une réponse plus élaborée. Prévoyez du temps en conséquence.

Vous devez rendre ce **questionnaire** et le feuillet avec les annexes à la fin de l'examen.

Aide-mémoire

- Propriétés de l'exponentielle: $\exp(a + b) = \exp(a) \exp(b)$, $\exp(0) = 1$
- Propriétés du logarithme népérien: $\ln(ab) = \ln(a) + \ln(b)$, $\ln(1) = 0$.
- sensibilité: $\Pr(\hat{Y} = 1 \mid Y = 1)$, spécificité: $\Pr(\hat{Y} = 0 \mid Y = 0)$, bonne classification: $\Pr(\hat{Y} = Y)$.
- Critères d'information: $\text{AIC} = -2\ell(\hat{\theta}) + 2p$ et $\text{BIC} = -2\ell(\hat{\theta}) + \ln(n)p$ où $\hat{\theta} = \max_{\theta \in \Theta} \ell(\theta)$, $p = \dim(\theta)$.
- modèle de régression logistique: paramétrisation avec référence $Y = k$

$$\ln \left\{ \frac{\Pr(Y = j \mid \mathbf{X})}{\Pr(Y = k \mid \mathbf{X})} \right\} = \eta_j = \beta_{0j} + \beta_{1j}X_1 + \cdots + \beta_{jp}X_p, \quad j \neq k.$$

- modèle à risques proportionnels de Cox:

$$h(t; \mathbf{X}) = h_0(t) \exp(\beta_1 X_1 + \cdots + \beta_p X_p).$$

Section réservée au correcteur

Question:	1	2	3	4	5	Total
Points:	13	7	10	8	12	50
Score:						

Question 1. Durée des abonnements mensuels de transport en commun

La Société des transports de Montréal (STM) a mandaté la firme Léger pour faire un sondage sur le comportement des usagers et usagères de transport en commun à Montréal. Les informations collectées lors de l'enquête sur $n = 1318$ personnes détentrices d'un abonnement annuel ou mensuel au cours des trois dernières années sont les suivantes:

- *clientele*: variable catégorielle, une parmi
 - jeune de 12 à 17 ou étudiant(e) de moins de 25 ans (jeune),
 - régulier (regulier) et
 - 65 ans et plus (65+).
- *activite*: variable binaire indiquant si la personne est employée ou aux études (oui ou non).
- *nmens*: nombre de passages mensuels moyens
- *nmoisabo*: nombre de mois avec un abonnement
- *abo_valide*: est-ce que l'abonnement était valide au moment de la fin de l'étude, 1 si oui, 0 sinon.

La STM s'intéresse à la durée des abonnements mensuels (*nmoisabo*) en fonction des informations sociodémographiques collectées. Certaines personnes sondées sont toujours titulaires d'un titre mensuel lors de l'enquête; il n'y a aucune perte de suivi.

- 1.1 Selon l'**Annexe 1**, combien de personnes parmi les 1318 détiennent toujours un abonnement à la fin de la période d'étude? [2]

- 1.2 À l'aide des statistiques présentées dans l'**Annexe 1**, rapportez une estimation **fiable** de la moyenne et la médiane de la durée d'abonnement au sein de la population de 65 ans et plus détentrices d'abonnements mensuels. Si ce n'est pas possible, expliquez pourquoi. [2]

1.3 Vous tracez les courbes de survie pour chaque clientèle; le résultat est présenté à la Figure A1 de l'**Annexe 1**. Identifiez la courbe correspondant à chaque clientèle et justifiez votre réponse.

[2]

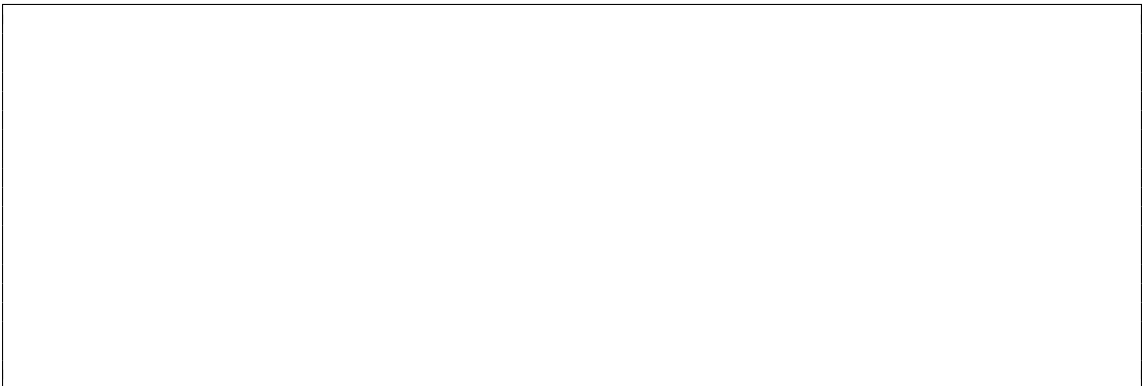
- jeune:
- régulier:
- 65+:

Justifiez votre réponse.



1.4 À partir du Tableau A3 de l'**Annexe 1**, calculez le troisième quartile du temps d'abonnement pour les **jeunes** (c'est-à-dire, le temps à partir duquel 25% des personnes sont désabonnés).

[2]



Vos gestionnaires sont intéressés à savoir comment le profil de la clientèle affecte le temps de détention de laisser-passer mensuels. Vous ajustez un modèle de Cox avec `nmens` et `activite` en stratifiant selon la `clientele`. Les résultats sont fournis dans l'**Annexe 2**.

- 1.5 Interprétez le coefficient pour `nmens` (**en pourcentage d'augmentation ou de diminution du risque**). **Attention à votre interprétation**

[2]

- 1.6 Quels sont les avantages et inconvénients de la stratification par `clientele`? Expliquez en quoi la sortie du modèle avec `clientele` comme variable explicative différerait de celle rapportée dans l'**Annexe 2**.

[3]

Question 2. Analyse factorielle**7**

On vous demande de faire une analyse des résultats du sondage administré aux 1318 participant(e)s contacté(e)s par Léger. Toutes les questions sont mesurées sur des échelles de Likert de 1 à 10: où 1 indique très insatisfait et 10 très satisfait.

2.1 Quels sont les objectifs de l'analyse factorielle exploratoire?

[2]

2.2 L'**Annexe 3** contient les résultats d'un sondage de 13 questions et l'estimation du modèle d'analyse factorielle. Sur la base des sorties, choisissez un nombre adéquat de facteurs. Justifiez votre réponse.

[3]

2.3 À partir du modèle avec le nombre de facteurs choisi, fournissez une typologie des facteurs.

[2]

Question 3. Données manquantes dans une analyse de regroupements**10**

On veut segmenter une base de données pour faire une analyse de regroupements, mais cette dernière contient des valeurs manquantes. Pour pallier à ce problème, on fait de l'imputation multiple et on crée 100 copies de la base de données avec les valeurs manquantes imputées. On procède ensuite à une analyse de regroupements avec la méthode des K -moyennes séparément pour chaque copie imputée de la base de données.

3.1 (★) Quelles sont les principales étapes d'une analyse de regroupement?

[3]

3.2 Suggérez un **critère** pour **choisir** le nombre de regroupements avec la méthode des K -moyennes. Expliquez ce dernier brièvement.

[2]

3.3 Expliquez comment vous adapteriez ce dernier pour choisir un seul nombre de groupes k à partir des résultats pour les 100 imputations.

[2]

3.4 La Figure 1 montre le résultat de l'analyse de regroupements avec $k = 3$ groupes pour deux

[3]

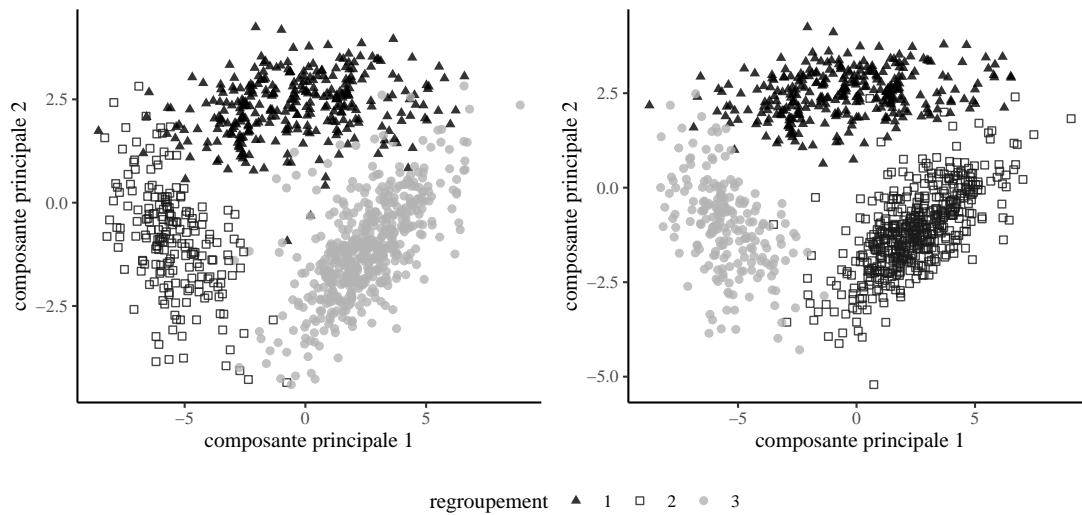


Figure 1: Projection sur deux composantes principales des regroupements obtenus par la méthode hiérarchique avec la méthode des k -moyennes avec $k = 3$ groupes. Chaque graphique représente une base de données imputées (parmi 100). Le numéro de groupe est indiqué dans la légende.

des 100 bases de données imputées. Elle illustre le problème de **permutation des étiquettes**. Suggérez une approche pour combiner les résultats des 100 analyses afin d'avoir une seule segmentation des sujets. *Indication: la sortie du modèle est une étiquette de groupe (G_1, G_2, \dots, G_k) pour chaque copie de la base de données.*

Question 4. Sélection de variables**8**

Pour chacun des énoncés suivants, indiquez s'il est **vrai ou faux**. **Justifiez** brièvement votre réponse.

- 4.1 Le modèle avec le plus petit BIC est toujours **meilleur** ou équivalent au modèle qui a la plus petite valeur d'AIC.

[2]

- 4.2 L'avantage de la validation externe par rapport à la validation croisée est l'absence d'aléatoire: on obtient toujours la même mesure de performance ou d'erreur.

[2]

- 4.3 Dans le cadre de la sélection de variables avec un ensemble de modèles linéaires candidats, on choisit celui qui a la plus petite erreur quadratique moyenne d'entraînement.

[2]

- 4.4 La procédure LASSO mène implicitement à une sélection de variables.

[2]

Question 5. Risque de défaillance et score de crédit**12**

Le terme de score de crédit fait référence à l'utilisation de méthodes statistiques pour classifier des créiteurs en bon ou mauvais risques. Une banque décide de faire une analyse de ses dossiers de crédits (aux particuliers) afin de mieux comprendre les caractéristiques qui font qu'une personne à qui on a accordé un prêt remboursera ce dernier avant l'échéance (defaut=0) ou pas (defaut=1).

On dispose des variables explicatives suivantes:

- **score**: score de crédit, une valeur entre 300 et 850; un score plus élevé dénote un meilleur dossier de crédit.
- **proprio**: est-ce que la personne qui a un emprunt est propriétaire, oui (1) ou non (0).
- **stabilite**: variable catégorielle pour la stabilité de l'emploi, soit faible (**faible**), modérée (**modere**) ou élevée (**eleve**).
- **tauxendet**: taux d'endettement (en pourcentage)

On divise la base de données en deux (échantillons d'entraînement et de validation) pour évaluer de manière fiable la performance.

Les coefficients du modèle logistique ajusté, les tests de significativité globale des coefficients et le tableau du lift sont présentés dans l'**Annexe 4**.

5.1 Interprétez l'effet du score de crédit sur le risque de défaut.

[2]

5.2 Rapportez le rapport de cote pour locataire versus propriétaire.

[2]

- 5.3 Est-ce que l'effet de la stabilité de l'emploi impacte globalement le risque de défaillance? Justifiez votre réponse.

[2]

- 5.4 Le lift pour le modèle de prédiction est retourné dans l'**Annexe 4**. Interprétez le lift pour 30%.

[2]

- 5.5 Votre collègue a décidé de faire de la sélection de variable et obtient trois modèles. Il calcule un point de coupure optimal pour chaque modèle. Les tableaux qui suivent résument la performance des trois modèles sur l'échantillon de validation contenant $n = 500$ observations (pour rappel, une valeur de $Y = 1$ indique un **défaut** de paiement).

Tableau 1: Tableaux de classification pour les différents modèles proposés.

	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	2	8
$\hat{Y} = 0$	38	452

(a) Modèle 1

	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	16	93
$\hat{Y} = 0$	24	367

(b) Modèle 2

	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	30	40
$\hat{Y} = 0$	10	420

(c) Modèle 3

- i. Si vous cherchez à minimiser le taux de mauvaise classification, quel modèle est préférable? Justifiez votre réponse. [1]

- ii. La banque ne prêtera pas aux personnes qui sont identifiées comme mauvais créditeurs. Selon leur estimation, un prêt moyen permet un retour sur investissement de 2%, mais on perd en moyenne 10% si une personne ne le rembourse pas à échéance (autrement dit, il est cinq fois plus grave de classer une observation comme bon créditeur alors que la personne fera défaut). Selon ce scénario, quel modèle est alors préférable? Justifiez votre réponse. [3]

Annexe 1 - Estimation de la fonction de survie

```
library(survival)
# Statistiques descriptives usuelles, Tableau A1
summary(stm$nmoisabo)
km <- survfit(
  formula = Surv(time = nmoisabo, event = !abo_valide) ~ clientele,
  data = stm)
print(km) # Tableau A2
summary(km) # Tableau A3
```

Tableau A1: Statistiques descriptives usuelles de l'ensemble des observations pour la variable nmoisabo (durée d'abonnement) en fonction de la clientèle

clientèle	médiane	moyenne	écart-type	min	max
jeune	20	27.08	22.80	1	96
regulier	19	34.74	50.30	1	385
65+	12	16.48	18.25	1	183

Tableau A2: Statistiques descriptives de la survie des abonnements en fonction de la clientèle (nombre d'observation, nombre d'événements, moyenne restreinte et son erreur-type, médiane).

	nombre	événement	médiane	moyenne r.	écart-type
jeune	117	78	27	71.00	19.78
regulier	309	184	36	73.37	8.59
65+	892	522	19	30.48	1.86

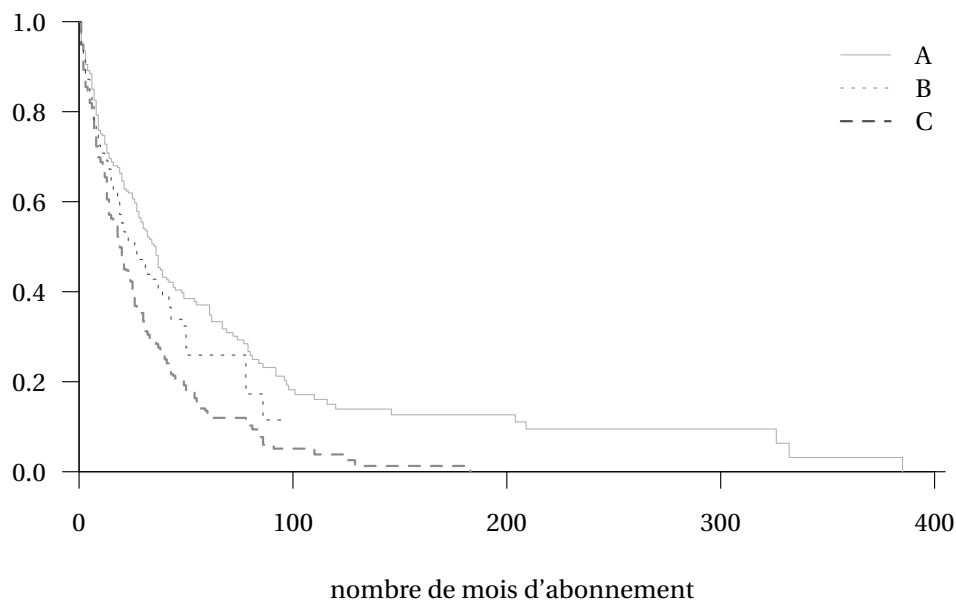


Figure A1: Courbes de survie selon la clientèle.

Tableau A3: Estimation du maximum de vraisemblance nonparamétrique de la fonction de survie (Kaplan–Meier) pour la clientèle jeune (sortie tronquée)

temps	nb à risque	nb d'événements	survie	erreur-type
1	117	6	0.949	0.020
2	111	4	0.915	0.026
3	107	4	0.880	0.030
4	103	1	0.872	0.031
5	102	3	0.846	0.033
6	98	3	0.820	0.036
7	95	6	0.768	0.039
8	89	2	0.751	0.040
9	87	3	0.725	0.041
20	60	2	0.552	0.047
21	56	2	0.533	0.047
23	54	2	0.513	0.047
26	51	1	0.503	0.048
27	48	3	0.472	0.048
31	43	2	0.450	0.048
50	20	3	0.275	0.049
51	17	1	0.259	0.049
78	6	2	0.173	0.059
86	3	1	0.115	0.062

Annexe 2 - Modèle à risques proportionnels de Cox

```
# Modèle avec stratification
rpcox_strat <- coxph(
  formula = Surv(time = nmoisabo,
                 event = !abo_valide) ~
    strata(clientele) + nmens + activite,
  data = stm)
# Coefficients, etc. - Tableau A4
summary(rpcox_strat)
```

Tableau A4: Estimations du modèle à risques proportionnels de Cox: coefficients, erreur-types et intervalles de confiance à 95% de Wald.

	coefficient	exp(coef.)	erreur-type	borne inf.	borne sup.
nmens	-0.039	0.962	0.007	-0.053	-0.025
activite [non]	0.473	1.604	0.153	0.172	0.773

Annexe 3 - Modèle d'analyse factorielle pour sondage

```
# Critères de sélection
hecmulti::ajustement_factanal(factors = 1:5, x = sondage_stm) #Tableau A5
factanal(sondage_stm, factors = 3) # Tableau A6
factanal(sondage_stm, factors = 4) # Tableau A6
factanal(sondage_stm, factors = 5) # Tableau A7
```

Tableau A5: Diagnostics pour la sélection du nombre de facteurs: critères d'information, valeur- p du test de rapport de vraisemblance pour le modèle saturé, nombre de paramètres de covariance et indicateur pour cas de (quasi)-Heywood.

k	AIC	BIC	valeur- p	nb par.	Heywood
1	5183.16	5317.94	$< 10^{-12}$	26	0
2	701.95	898.94	$< 10^{-12}$	38	0
3	-1039.54	-785.53	$< 10^{-12}$	49	0
4	-2062.68	-1756.83	0.019	59	0
5	-2070.94	-1718.44	0.374	68	0

Tableau A6: Estimés des chargements (multipliés par 100) pour les modèles à trois facteurs avec rotation varimax estimé à l'aide de la méthode du maximum de vraisemblance. Les chargements inférieurs à 0.2 sont omis.

	F1	F2	F3
fréquence de passage	82		29
correspondance entre lignes	82		44
temps de trajet	83		39
offre de service	90	21	
achalandage	88	21	28
accessibilité des installations	27	88	
propreté des installations		85	
confort des installations		90	
disponibilité des places assises	39	80	
service au guichet	34	22	83
sécurité lors des déplacements	28		78
tarification	43	30	62
offres de titres de transports	24	51	42

Tableau A7: Estimés des chargements (multipliés par 100) pour les modèles à quatre facteurs avec rotation varimax estimé à l'aide de la méthode du maximum de vraisemblance. Les chargements inférieurs à 0.2 sont omis.

	F1	F2	F3	F4
fréquence de passage	84		28	
correspondance entre lignes	84		33	27
temps de trajet	85		31	21
offre de service	90			
achalandage	89			
accessibilité des installations	28	89		
propreté des installations		82		33
confort des installations		90		
disponibilité des places assises	39	77		
service au guichet	39	20	71	31
sécurité lors des déplacements	31		89	
tarification	44	23	45	57
offres de titres de transports	21	42		79

Tableau A8: Estimés des chargements (multipliés par 100) pour les modèles à cinq facteurs avec rotation varimax estimé à l'aide de la méthode du maximum de vraisemblance. Les chargements inférieurs à 0.2 sont omis.

	F1	F2	F3	F4	F5
fréquence de passage	84		28		
correspondance entre lignes	84		33	27	
temps de trajet	85		31	21	
offre de service	90				
achalandage	89				
accessibilité des installations	28	91			
propreté des installations		81		34	
confort des installations		91			
disponibilité des places assises	39	76		21	
service au guichet	39		71	31	
sécurité lors des déplacements	31		89		
tarification	44	23	45	57	
offres de titres de transports	21	41		80	

Annexe 4 - risque de défaut de paiement

```
logist <- glm(
  formula = défaut ~ proprio * tauxendet + stabilite + score,
  data = scorecredit_entrainement,
  family = binomial(link = "logit"))
# Tableau A9
summary(logist)
confint(logist)
# Tableau A10
car::Anova(logist, type = 3)

pred_prob <- predict(
  object = logist,
  newdata = scorecredit_validation,
  type = "response")
# Tableau A11
hecmulti::courbe_lift(
  prob = pred_prob,
  resp = scorecredit_validation$defaut)
```

Tableau A9: Estimations du modèle logistique: coefficients et intervalles de confiance profilés à 95% pour les coefficients.

variable	coef.	exp(coef.)	borne inf.	borne sup.
const	0.1690	1.1841	-0.6480	0.9745
proprio	-0.7686	0.4636	-1.5561	0.0135
tauxendet	0.0077	1.0077	0.0008	0.0146
stabilite [élevée]	-0.3754	0.6870	-0.7120	-0.0506
stabilite [modérée]	-0.3624	0.6960	-0.6122	-0.1118
score	-0.0054	0.9946	-0.0066	-0.0042
proprio:tauxendet	0.0003	1.0003	-0.0073	0.0078

Tableau A10: Tests du rapport de vraisemblance (effets de type III).

	statistique	ddl	valeur- <i>p</i>
proprio	3.71	1	0.05
tauxendet	4.84	1	0.03
stabilite	9.23	2	0.01
score	78.46	1	<0.001
proprio:tauxendet	0.01	1	0.94

Tableau A11: Tableau du lift.

pourcentage	hasard	modèle	lift
10%	4	9	2.25
20%	8	15	1.88
30%	12	20	1.67
40%	16	24	1.50
50%	20	25	1.25
60%	24	29	1.21
70%	28	35	1.25
80%	32	36	1.12
90%	36	38	1.06