MATH60604 *Modélisation statistique*

Examen intratrimestriel

Questionnaire

Examen de pratique

Léo Belzile

Instructions: L'examen est d'une durée de 180 minutes.

Une feuille d'aide mémoire (recto, format lettre) est permise. L'utilisation d'un ordinateur ou de tout autre matériel électronique est interdit. Une calculatrice non programmable est autorisée.

La répartition des 40 points de l'examen se trouve dans la marge de droite.

Nom:

Prénom:

MATRICULE:

Question:	1	2	3	Total
Points:	12	12	16	40
Score:				

Question 1. ..... **12**

En analyse des valeurs extrêmes, la théorie asymptotique dicte que les excès de Y au dessus d'un seuil élevé u est bien approximée par une loi de **Pareto généralisée**, avec $Z = Y - u \dot{\sim} \mathsf{GP}(\sigma, \xi)$ pour des paramètres d'échelle $\sigma > 0$ et de forme $\xi \in \mathbb{R}$. Les fonctions de répartition et de densité de la loi Pareto généralisée sont

$$F(z) = 1 - (1 + \xi z/\sigma)_+^{-1/\xi}, \qquad f(z) = \sigma^{-1}(1 + \xi z/\sigma)_+^{-1/\xi - 1},$$

où $(x)_+ = \max\{x, 0\}$; le cas $\xi = 0$ est défini par continuité (sous-famille exponentielle).

On considère les plus grandes réclamations d'assurance incendies (en millions de couronnes) à la Copenhagen Re, une compagnie de réassurance danoise, soumises entre janvier 1980 et la fin de décembre 1990 ($n_Y = 11$ années de données). On modélise les $n = 109$ excès de de seuils au delà de 10 millions de couronnes, correspond à une proportion de $\zeta = 0.0503$ des données complètes. Notre objectif est de fournir un **niveau de retour** à 100 ans pour une analyse de risque.

Le niveau de retour à T-années, dénoté r_T, est un quantile élevé excédé avec probabilité p, où on prend $p = \zeta n_Y / T$, avec ζ la proportion d'observations au dessus du seuil, n_Y le nombre d'années d'observations et $T = 100$ le nombre d'années de l'horizon considéré. Si on inverse la fonction de répartition, on obtient la fonction quantile et la formule

$$rT = \frac{\sigma}{\xi}\left\{(\zeta n_Y / T)^{-\xi} - 1\right\}$$

1.1 Écrivez la fonction de log vraisemblance pour un échantillon de taille n d'excès de seuils indépendants z_i, $(i = 1, \ldots, n)$ si $\xi \neq 0$. [2]

1.2 Si on reparamétrise le modèle en terme de ξ et $\theta = \xi/\sigma$, montrer que l'on peut dériver une formule explicite pour la log vraisemblance profilée $\ell_{\mathrm{p}}(\theta)$, ce qui permet de réduire le problème d'optimisation de 2D à 1D.. [2]

```
Log-vraisemblance: -374.893
Taille de l'echantillon: 109
Proportion au dessus du seuil: 0.0503

Estimation                Erreurs-type
echelle  forme            echelle   forme
6.975    0.497            1.1135    0.1363
```

Code 1: Estimations du maximum de vraisemblance pour les données de réassurance danoise (loi Pareto généralisée)

1.3 Un algorithme d'optimisation a retourné les estimations de la Code 1. Expliquez comment vous pourriez tester si $\xi = 0$ (modèle exponentiel) en utilisant [2]

(a) un test de Wald avec la sortie de Code 1 et

(b) un test du rapport de vraisemblance, si en plus vous avez l'information suivante:

```
> sum(dexp(y, rate = 1/mean(y), log = TRUE))
-397.2921
```

Code 2: Code pour une log vraisemblance exponentielle

1.4 On peut démontrer que l'information de Fisher d'un échantillon aléatoire simple de taille n de loi Pareto généralisée est [2]

$$\imath(\sigma,\xi) = n\begin{pmatrix} \sigma^{-2}(1+2\xi)^{-1} & \sigma^{-1}(1+\xi)^{-1}(1+2\xi)^{-1} \\ \sigma^{-1}(1+\xi)^{-1}(1+2\xi)^{-1} & 2(1+\xi)^{-1}(1+2\xi)^{-1} \end{pmatrix}$$

Expliquez comment on peut utiliser ce résultat pour obtenir les erreurs-types des paramètres σ et ξ.

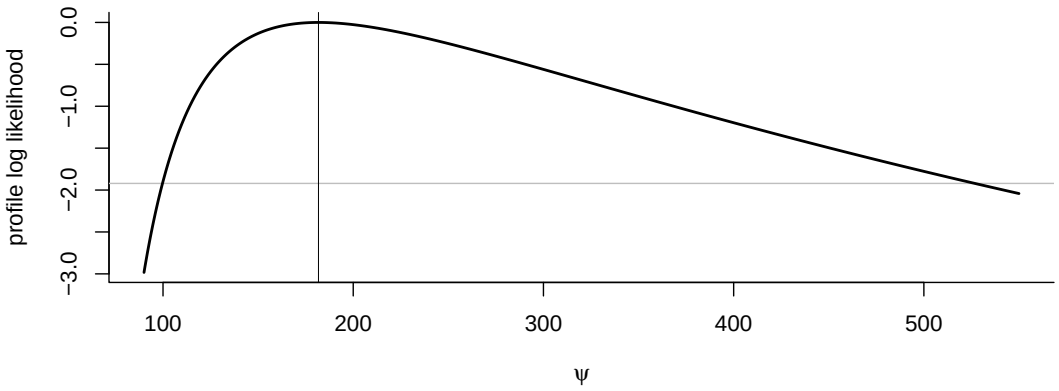

Figure 1: Log vraisemblance profilée pour le niveau de retour à 100 ans. La log vraisemblance profilée a été décalée pour être zéro lorsqu'évaluée aux EMV, et la ligne grise horizontale indique les points de coupure pour un intervalle de confiance basé sur la statistique du rapport de log vraisemblance à niveau 95%.

1.5 Donnez l'estimation du maximum de vraisemblance (EMV) du niveau de retour à 100 ans r_{100} pour les données de réassurance danoises. [2]

1.6 La Figure 1 montre la log vraisemblance profilée pour le niveau de retour à 100 ans (avec $\psi \equiv r_{100}$); la ligne horizontale grise indique les points de coupure pour un intervalle de confiance à 95% basé sur la loi asymptotique χ^2_1 de la statistique du rapport de log vraisemblance profilée. Au vu de la loi d'échantillonnage de ce paramètre, est-ce que l'intervalle de Wald serait semblable? [2]

Question 2. ..... **12**

Grossmann et Kross (2014) étudient la question suivante : « Les gens sont-ils plus sages lorsqu'ils réfléchissent aux problèmes des autres qu'aux leurs ? Ils ont assigné au hasard des participants à

- raisonner sur leur propre problème à partir d'une perspective immersive (condition auto, immersive),
- raisonner sur le problème de leur ami à partir d'une perspective immersive (condition immersion, autre),
- raisonner sur leur propre problème d'un point de vue distancié (condition auto, distanciée), ou
- raisonner sur le problème de leur ami à partir d'une perspective distanciée (condition « autre, distancié).

L'étude ci-dessous a également considéré l'âge comme un facteur distinct.

Les variables comprennent

- `limites` : variable de réponse, le score moyen centré pour la question sur la « reconnaissance des limites de la connaissance ».
- `cible` : facteur, la cible est-elle la personne qui participe (`auto`) ou un.e ami.e. (`autre`).
- `perspective` : facteur, perspective `immersive` ou `distancie`.
- `age` : groupe d'âge, soit `jeune` (20–40 ans) ou `vieux`, (60–80 ans).

Tableau 1: Coefficients et erreurs-types pour le modèle factoriel complet à trois voies.

	coef.	erreur-type
(cst)	0.207	0.144
cible [auto]	0.024	0.204
perspective [immersive]	−0.121	0.203
age [vieux]	−0.225	0.216
cible [auto]:perspective [immersive]	−0.211	0.284
cible [auto]:age [vieux]	0.059	0.308
perspective [immersive]:age [vieux]	0.020	0.306
cible [auto]:perspective [immersive]:age [vieux]	−0.195	0.433

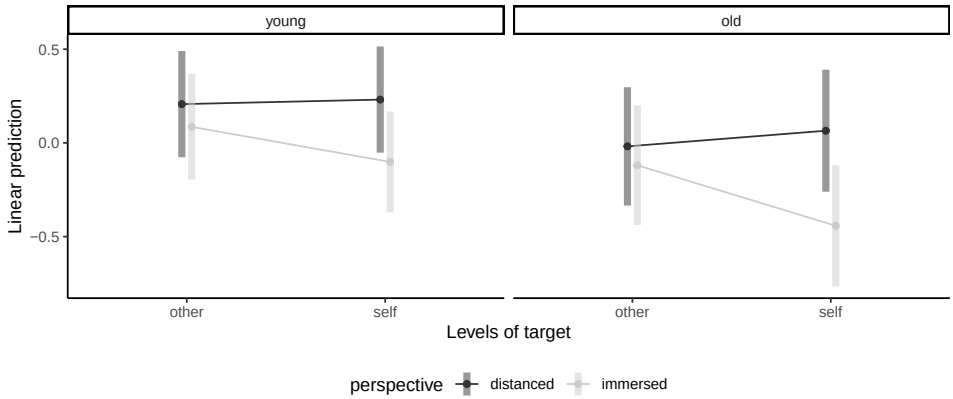

Figure 2: Moyennes marginales avec intervalles de confiance à 95% pour chaque sous-groupe

Tableau 2: Tableau d'analyse de variance (décomposition des carrés de type 2) pour le modèle factoriel complet à trois facteurs.

	somme des carrés	ddl	stat. F	valeur-p
cible	1.12	1	0.87	0.35
perspective	7.59	1	5.88	0.02
age	6.12	1	4.74	0.03
cible:perspective	2.45	1	1.90	0.17
cible:age	0.04	1	0.03	0.86
perspective:age	0.16	1	0.13	0.72
cible:perspective:age	0.26	1	0.20	0.65
residuals	570.17	442		

Tableau 3: Moyennes marginales estimées pour les quatre groupes expérimentaux.

perspective	cible	moyenne marg.	erreur-type	borne inf.	borne sup.
distancée	autre	0.09	0.11	−0.12	0.31
immersive	autre	−0.02	0.11	−0.23	0.20
distancée	auto	0.15	0.11	−0.07	0.36
immersive	auto	−0.27	0.11	−0.48	−0.06

Tableau 4: Contrastes marginaux pour le modèle à deux facteurs.

contraste	estimation	erreur-type	stat. t	valeur-p
C_1	0.311	0.131	2.366	0.018
C_2	−0.109	0.134	−0.818	0.414
C_3	0.420	0.153	2.742	0.006
C_4	0.111	0.153	0.727	0.468

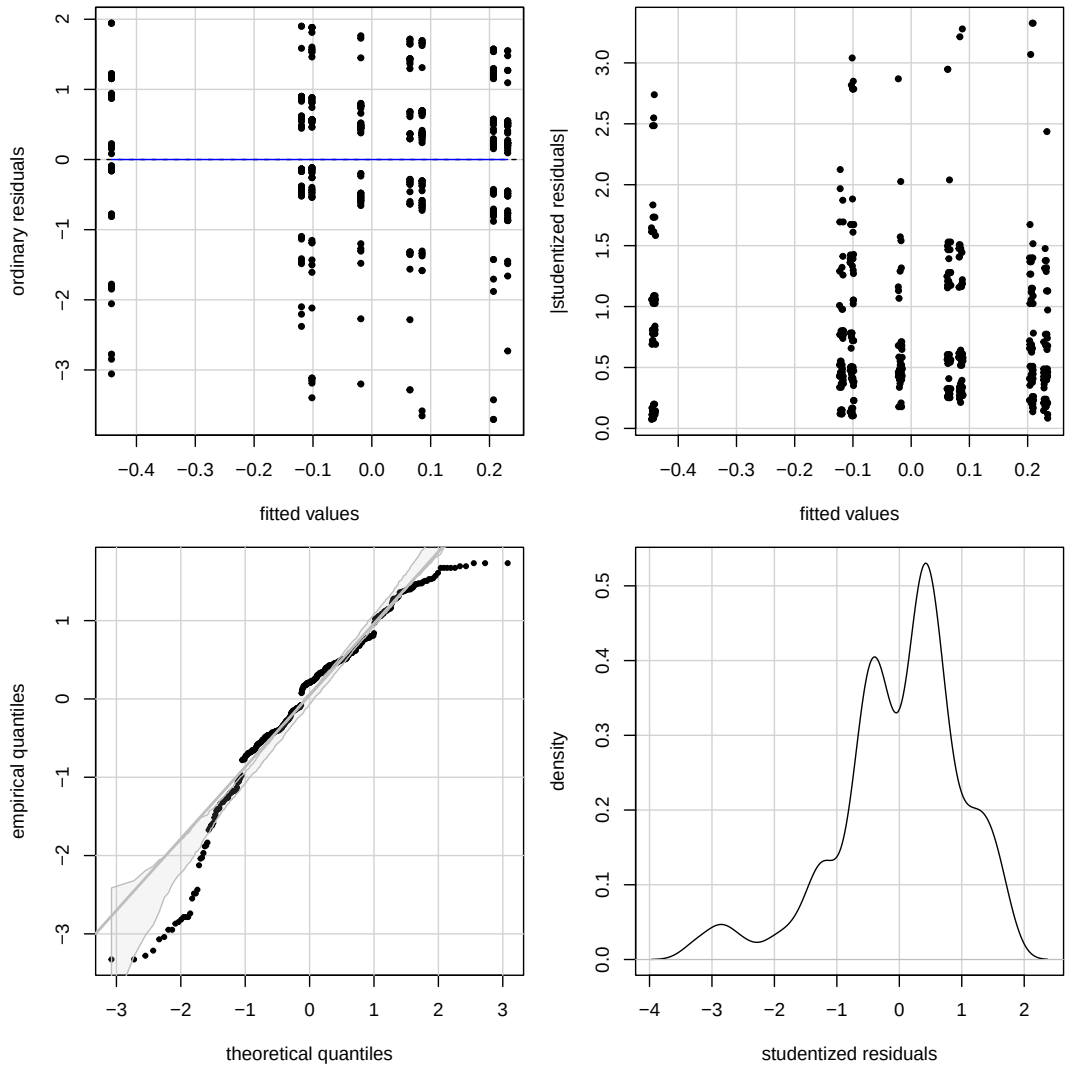

Figure 3: Diagnostics graphiques pour le modèle ANOVA à trois facteurs: nuage de point des résidus ordinaires en fonction des valeurs ajustées (en haut à gauche), valeur absolue des résidus studentisés externes vs valeurs ajustés(en haut à droite), diagramme quantile-quantile Student des résidus studentisés externe (en bas à gauche) et densité empirique des résidus studentisés externes (en bas à droite).

2.1 Interprétez l'ordonnée à l'origine du modèle factoriel ajusté dont les coefficients sont rapportés dans le Tableau 1. [2]

2.2 Selon le Tableau 2, combien de participants ont été recrutés pour l'étude? [2]

2.3 Sur la base de l'analyse de variance rapportée au Tableau 2, est-il correct de marginaliser et combiner les deux groupes d'âge afin d'étudier l'effet de cible et perspective (effets marginaux)? Indiquez l'avantage de cette façon de faire. [2]

2.4 Écrivez les vecteurs de poids pour les quatre contrastes suivants, en respectant l'ordre des sous-groupes répertoriées dans le Tableau 3: [2]

C_1 (_____, _____, _____, _____) autre (immersive et distance) vs auto-immersive
C_2 (_____, _____, _____, _____) autre (immersive et distancée) vs auto-distancée
C_3 (_____, _____, _____, _____) auto-distancée vs auto-immersive
C_4 (_____, _____, _____, _____) autre-distancée vs autre-immersive

2.5 À l'aide de la sortie du Tableau 4, commentez sur l'analyse des contrastes. [2]

2.6 Sur la base de la Figure 3, quel postulat de validité du modèle linéaire n'est pas respecté? Justifiez votre réponse et discutez des impacts de ce postulat sur l'inférence. [2]

- spécification incorrecte de la moyenne
- additivité
- homoscédasticité
- normalité
- absence de valeurs aberrantes

Question 3. ..... **16**

La stratégie "réfléchir à voix haute" et la compréhension de lecture. Les données de Baumann *et coll.* (1992) étudient l'effet de différentes méthodes d'instruction sur la compréhension de lecture. La base de données avec $n = 66$ observations de l'échantillon équilibré contient les variables suivantes:

- `pretest2`: score (sur 15) sur le questionnaire de compréhension de lecture, administré avant l'expérience
- `posttest2`: response, score (sur 18) sur une version améliorée du questionnaire de compréhension de lecture, administré après l'expérience
- `groupe`: groupe expérimental, soit « Think-Aloud » (TA), dans lequel les élèves ont appris diverses stratégies de contrôle de la compréhension pour la lecture d'histoires (par exemple : auto-questionnement, prédiction, relecture) par le biais de la réflexion à haute voix; (b) un groupe lecture dirigée-activité de réflexion (DRTA), dans lequel les élèves ont appris une stratégie de prédiction-vérification pour lire et répondre aux histoires; ou (c) un groupe activité de lecture dirigée (DRA), un groupe contrôle dans lequel les élèves se sont engagés dans une lecture guidée non interactive d'histoires.

On ajuste des modèles pour le score `posttest2` avec la contrainte de **somme à zéro** pour la variable catégorielle `groupe`. Les matrices de modèles incluent les valeurs binaires suivantes comme au Tableau 5 pour les grois groupes.

	(cst)	groupe1	groupe2
DR	1	1	0
DRTA	1	0	1
TA	1	−1	−1

Tableau 5: Matrice du modèle pour les trois variables binaires avec contrainte de somme à zéro.

```
> model1 <- lm(posttest2 ~ groupe, data = BSJ92)
> model2 <- lm(posttest2 ~ offset(pretest2) + groupe, data = BSJ92)
> model3 <- lm(posttest2 ~ pretest2 + groupe, data = BSJ92)
> model4 <- lm(posttest2 ~ pretest2 * groupe, data = BSJ92)
```

Code 3: Syntaxe **R** pour les quatre modèles ajustés

```
> anova(model3, model4)
Analysis of Variance Table

Model 3: posttest2 ~ pretest2 + groupe
Model 4: posttest2 ~ pretest2 * groupe
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     62 332.19
2     60 331.07  2    1.1264 0.1021 0.9031
```

Code 4: Comparaison de modèles 3 et 4

Tableau 6: Coefficients et erreurs-type pour différents modèles linéaires pour la variable réponse `posttest2`.

	estimation	erreur-type
(cst)	6.712	0.293
groupe1	−1.167	0.414
groupe2	−0.485	0.414

(a) Coefficients pour le Modèle 1 pour le score post-test 2 en fonction du groupe expérimental (paramétrisation somme nulle).

	estimation	erreur-type
(cst)	5.301	0.722
pretest2	0.276	0.130
groupe1	−1.213	0.404
groupe2	−0.481	0.403

(b) Coefficients pour le Modèle 3 pour le score post-test 2 en fonction du groupe expérimental (paramétrisation somme nulle) et `pretest2`.

	estimation	erreur-type
(cst)	5.398	0.765
pretest2	0.256	0.140
groupe1	−1.619	0.994
groupe2	−0.236	1.113
pretest2:groupe1	0.079	0.176
pretest2:groupe2	−0.047	0.204

(c) Coefficients pour le Modèle 4 pour le score post-test 2 en fonction du groupe expérimental (paramétrisation somme nulle), `pretest2` et leur interaction.

Tableau 7: Moyennes marginales estimées pour groupe pour le Modèle 3.

groupe	moyenne marginale	erreur-type	ddl	borne inf.	borne sup.
DR	5.499	0.494	62	4.512	6.487
DRTA	6.231	0.494	62	5.245	7.218
TA	8.406	0.494	62	7.418	9.393

Tableau 8: Contrastes par paires (différences deux à deux) basés sur les moyennes marginales du Modèle 3.

contraste	estimation	erreur-type	ddl	stat. t	valeur-p‘
DR − DRTA	−0.732	0.698	62	−1.048	0.550
DR − TA	−2.906	0.699	62	−4.157	$< 10^{-3}$
DRTA − TA	−2.174	0.698	62	−3.114	0.008

3.1 Calculez la moyenne empirique du score post-test 2 pour chaque groupe expérimental à l'aide du Tableau 6. [2]

3.2 Il est fréquent (mais souvent invalide) d'ajuster un modèle de variance pour la différence post/pré, soit posttest2−pretest2 (Modèle 2) plutôt que d'ajuster un modèle linéaire comme le Modèle 3. Le Modèle 2 pour la différence est équivalent à un modèle avec un terme de décalage (une variable explicative avec une coefficient fixe de 1). Est-ce que les données supportent le choix du Modèle 2. [2]

3.3 Supposons que l'on veut comparer les quatre modèles répertoriés dans le Code 3: lesquels sont emboîtés? [2]

3.4 Écrivez l'équation de la moyenne théorique de chaque groupe pour le Modèle 4 et montrez que le Modèle 3 est une simplification de ce dernier. Écrivez les hypothèses nulles et alternatives en fonction des paramètres du modèle et concluez sur la base de la sortie de Code 4. [2]

3.5 Les auteurs ont calculé les moyennes marginales par groupe sur la base du Modèle 3, et les différences entre paires. Sur la base de ses dernières, est-ce qu'on peut deviser un classement de la méthode d'enseignement la plus effective (sachant que les scores les plus élevés sont préférables)? [2]

3.6 Si les pentes de `pretest2` pour chaque `groupe` ne sont parallèles, expliquez pourquoi la comparaison de moyennes marginales est trompeuse. [2]

3.7 On rapporte le résultat du test de Levene pour l'homogénéité de variance, [2]

```
> car::leveneTest(rstudent(model3) ~ groupe,
+                 data = BSJ92,
+                 center = "mean")
```

Code 5: Syntaxe **R** pour le test de Levene

et le tableau résultant indique la statistique $F(2,63) = 1.51$, et une valeur-p de $p = 0.23$. À quoi sert ce test? Concluez quant à l'hypothèse et expliquez les impacts sur vos conclusions, si aucun.

3.8 Étant donné que `pretest2` et `posttest2` sont fortement corrélées, est-ce que cela n'est pas une violation du postulat d'indépendance entre observations? Discutez. [2]