MATH 60602 ***Analyse multidimensionnelle appliquée***

Examen intra de pratique

Questionnaire

Examinateur: 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 30 points de l'examen se trouve dans la marge de droite.

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: $\Pr(\widehat{Y} = 1 \mid Y = 1)$, spécificité: $\Pr(\widehat{Y} = 0 \mid Y = 0)$, bonne classification: $\Pr(\widehat{Y} = Y)$.
- Critères d'information: $\mathsf{AIC} = -2\ell(\widehat{\boldsymbol{\theta}}) + 2p$ et $\mathsf{BIC} = -2\ell(\widehat{\boldsymbol{\theta}}) + \ln(n)p$ où $\widehat{\boldsymbol{\theta}} = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ell(\boldsymbol{\theta})$, $p = \dim(\boldsymbol{\theta})$.
- modèle de régression logistique: paramétrisation avec référence $Y = k$

$$\ln\left\{\frac{\Pr(Y = j \mid \mathbf{X})}{\Pr(Y = k \mid \mathbf{X})}\right\} = \eta_j = \beta_{0j} + \beta_{1j}\mathrm{X}_1 + \cdots + \beta_{jp}\mathrm{X}_p, \qquad j \neq k.$$

Section réservée au correcteur

Question:	1	2	3	4	Total
Points:	3	13	4	10	30
Score:					

Question 1. Politiques publiques et coronavirus **3**

Le gouvernement du Québec a été longtemps réticent à rendre disponible les tests antigéniques rapides lors de la pandémie de Coronavirus. Une étude suisse, qui a comparé les résultats de ces tests dans un milieu clinique avec celui de tests PCR auprès de 1462 personnes, dont 141 ont été diagnostiqués Covid positif (+), est le suivant:

	PCR +	PCR −
rapide +	92	2
rapide −	49	1319
total	141	1321

Tableau 1 – *Matrice de confusion pour la comparaison entre résultats de tests antigéniques rapides (lignes) et tests d'amplification en chaîne par polymérase (PCR).*

1.1 À l'aide de la matrice de confusion, calculez la sensibilité du test rapide en partant du principe que le résultat du test PCR est la valeur de référence. [1 pt]

1.2 Vulgarisez et expliquez ce résultat dans le contexte du problème: est-ce que les inquiétudes du gouvernement étaient fondées? [2 pts]

Question 2. Examens de conduite automobile en Grande-Bretagne **13**

Un article du journal *The Guardian* d'août 2019 mettait en lumière les énormes différences dans les taux de réussite aux examens de conduite: selon les journalistes, il est plus facile d'obtenir un permis dans un petit centre de test rural qu'en ville.

Le jeu de données `gbconduite` contient les résultats de tous les tests de 2018, ventilés par sexe de l'individu et par région administrative anglaise (ou pays pour l'Écosse et le Pays de Galles). L'Annexe 1 contient les résultats de la modélisation.

2.1 Que produit les lignes 1–6 du code de l'Annexe 1.1? [2 pts]

2.2 Selon le modèle logistique ajusté (**Tableau A3**), quelles sont les régions ou pays où le taux de réussite aux examens est le plus fort et le plus faible, *ceteris paribus*? Justifiez votre réponse. [2 pts]

2.3 Quel est le rapport de cote pour les centres de petite versus moyenne taille? Est-ce que la différence entre les deux tailles est significative? [2 pts]

2.4 Interprétez le coefficient `sexe` en pourcentage d'augmentation ou de diminution de la cote. [2 pts]

Suite à l'ajustement, vous obtenez les prédictions du taux de succès dans chaque centre/sexe et vous produisez un nuage de points (**Figure 1**) pour comparer les résultats et les prédictions.

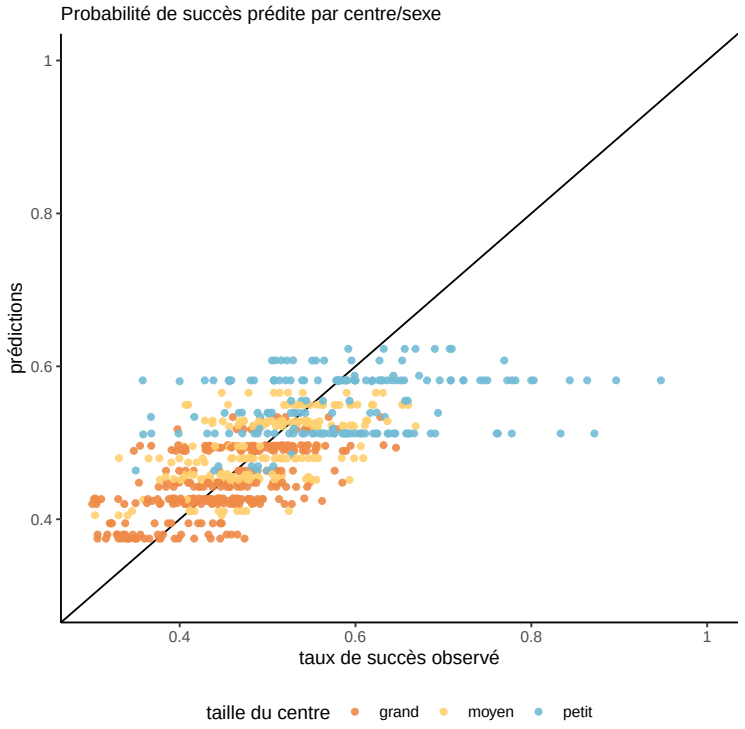

Figure 1 – *Nuage de points des prédictions de la probabilité de réussite versus le taux de 2018, en fonction de la taille du centre.*

2.5 Commentez sur la qualité du modèle et le constat du Guardian en vous basant sur les sorties et notamment la **Figure 1**: est-ce que le modèle parvient à expliquer les haut taux de succès? [3 pts]

2.6 Le **Tableau A1** montre des différences notables dans la taille de centres selon les régions. Pour améliorer le modèle, vous aimeriez considérer une interaction entre les variables catégorielles taille et région. Vous ajustez le modèle, mais obtenez le message suivant dans la sortie: [2 pts]

```
Coefficients: (2 not defined because of singularities)
```

et deux coefficients ajustés sont manquants. Expliquez la cause de ce problème.

Question 3. Inoccupation des refuges à Toronto **4**

Les données ouvertes de la ville de Toronto contiennent des statistiques sur l'occupation des refuges pour l'année 2017. Une analyse exploratoire préliminaire visant à s'assurer de la conformité des données est effectuée.

Chaque ligne de la base de données contient le type de refuge, le décompte et la capacité pour chaque jour. Les données sont ordonnées lors de l'importation en ordre chronologique.

Décrivez un problème potentiel soulevé par les Figures A1 et A2 et fournissez une explication plausible au vu des tendances observées.

Question 4. Sélection de variables **10**

Les énoncés suivants portent sur la sélection de variables à partir d'un ensemble de variables explicatives données. Pour chacun d'entre eux, indiquez s'il est vrai ou faux et **justifiez votre réponse**. Si l'énoncé est faux, corrigez-le.

4.1 La procédure séquentielle avec le critère d'information bayésien (BIC) retourne forcément le modèle avec le plus petit BIC pour les données d'entraînement parmi tous les modèles envisageables. [2 pts]

4.2 Si on fait une recherche ascendante avec le critère d'information d'Akaike (AIC), l'ensemble des modèles considérés lors de la recherche incluera nécessairement le résultat de la recherche ascendante avec le BIC. [2 pts]

4.3 Si on veut prédire une variable réponse binaire et qu'on s'intéresse davantage aux succès (valeur de 1), utiliser la sensibilité comme critère de sélection est un choix sensé pour comparer plusieurs modèles entre eux. [2 pts]

4.4 Vous avez produit le graphique suivant pour illustrer l'estimation des critères d'information d'Akaike (AIC) et de Schwarz (BIC), mais sans étiqueter les courbes. Associez chaque critère avec la courbe (ligne traitillée ou pleine). [2 pts]

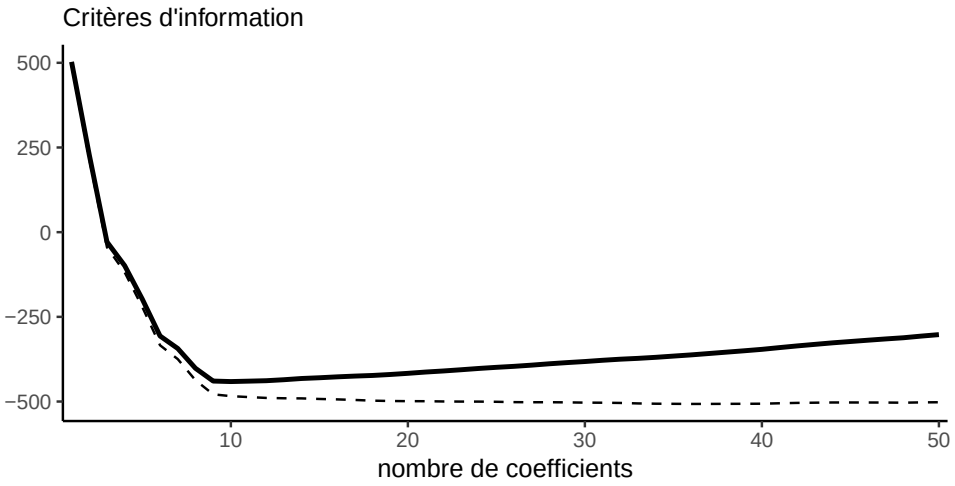

Figure 2 – *Valeurs des critères d'information AIC et BIC en fonction du nombre de coefficients du modèle.*

4.5 Supposons qu'on fait une recherche ascendante avec le critère d'information d'Akaike et que l'on estime ensuite, pour chaque modèle avec de 1 à 40 variables explicatives, l'aire sous la courbe d'efficacité du récepteur par validation croisée pour chaque modèle. [2 pts]

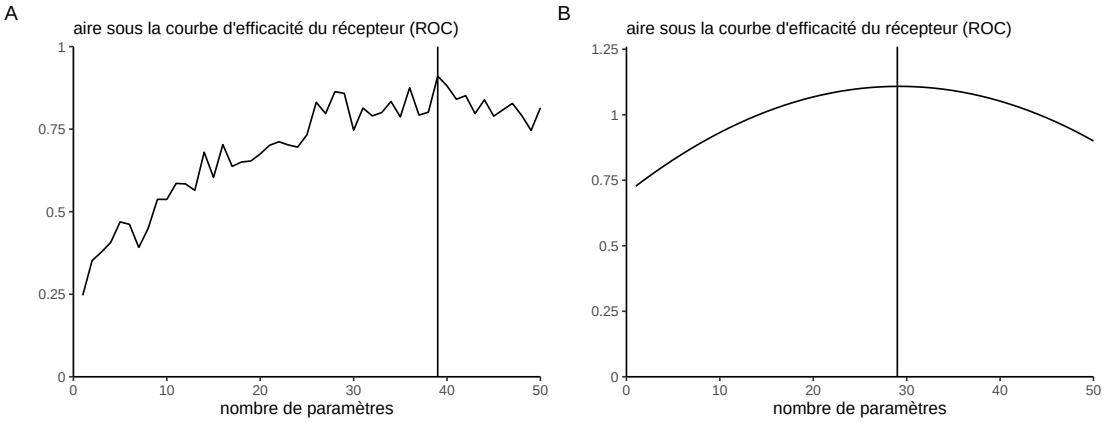

Figure 3 – *Courbes candidates en sortie de la procédure décrite à la question 4.5*

Laquelle des sorties correspond au résultat attendu?

Annexe 1: examens de conduite automobile en Grande-Bretagne

La base de données tirée de GovUK donne le taux de succès aux examens pratique de conduite en Grande-Bretagne pour 2018 dans 346 centres de conduite. Un total de 761 750 personnes ont réussi l'examen pratique, pour 1 663 897 essais.

Les données sont regroupées par centre et par sexe:

- `reussite`: nombre de personnes ayant réussi l'examen.
- `nbtests`: nombre total de tests dans le centre par sexe.
- `total`: nombre total de tests dans le centre.
- `region`: facteur indiquant une des neuf régions d'Angleterre, le Pays de Galles ou l'Écosse.
- `sexe`: sexe, soit homme ou femme.

Annexe 1.1: manipulation de la base de données et statistiques descriptives

```
library(dplyr)
gbconduite <- gbconduite |>
   mutate(taille =
     factor(case_when(total < 1000 ~ "petit",
                      total < 2500 ~ "moyen",
                      TRUE ~ "grand")))
# Tableau A1
with(gbconduite, table(region, taille))
# Tableau A2
gbconduite |> group_by(region) |>
  summarize(taux = mean(reussite/nbtests),
            ecarttype = sd(reussite/nbtests))
```

Tableau A1 – *Décompte du nombre de centre d'examens de conduite par région anglaise (incluant le Pays de Galles et l'Écosse) et par grosseur de centre.*

no	région	taille grand	moyen	petit
1	Angleterre de l'Est	39	15	0
2	Angleterre du Nord-Est	13	16	13
3	Angleterre du Nord-Ouest	38	25	5
4	Angleterre du Sud-Est	55	23	0
5	Angleterre du Sud-Ouest	24	20	6
6	Écosse	16	25	111
7	Grand Londres	40	8	10
8	Pays de Galles	6	24	18
9	Midlands de l'Est	22	18	6
10	Midlands de l'Ouest	32	22	8
11	Yorkshire et Humber	26	6	2

Tableau A2 – *Taux de réussite et écart-type associé aux examens de conduite par région.*

no de région	1	2	3	4	5	6	7	8	9	10	11
taux	0.46	0.54	0.47	0.46	0.50	0.57	0.41	0.49	0.44	0.53	0.43
écart-type	0.05	0.08	0.07	0.06	0.07	0.11	0.06	0.07	0.07	0.07	0.06

Annexe 1.2: modèle de régression logistique

```
# Modèle de régression logistique
logist_mod <- glm(
  formula = cbind(reussite, nbtests - reussite) ~ region + sexe + taille,
  family = binomial,
  data = gbconduite)
# Tableau A1: coefficients et intervalles de confiance
cbind(coef(logist_mod), exp(coef(logist_mod)), confint(logist_mod))
# Tableau A3- tests de rapport de vraisemblance
car::Anova(logist_mod, type = 3)
```

Tableau A3 – *Estimations des coefficients et intervalles de confiance profilés à 95% pour le modèle de régression logistique. Pour faire la correspondance entre numéros et noms des régions, voir le Tableau A1.*

variable	coef.	exp(coef.)	borne inf.	borne sup.
const	−0.306	0.736	−0.317	−0.296
`region` [2]	0.160	1.174	0.143	0.178
`region` [3]	−0.017	0.984	−0.030	−0.003
`region` [4]	0.010	1.010	−0.003	0.022
`region` [5]	0.097	1.102	0.082	0.112
`region` [6]	−0.012	0.988	−0.027	0.004
`region` [7]	−0.206	0.814	−0.219	−0.193
`region` [8]	0.013	1.013	−0.002	0.028
`region` [9]	−0.183	0.833	−0.197	−0.170
`region` [10]	0.096	1.101	0.078	0.114
`region` [11]	−0.121	0.886	−0.136	−0.106
`sexe` [homme]	0.281	1.324	0.275	0.287
`taille` [moyen]	0.128	1.137	0.121	0.136
`taille` [petit]	0.367	1.443	0.350	0.384

Tableau A4 – *Tests du rapport de vraisemblance pour le modèle de régression logistique*

	statistique	ddl	valeur-p
`region`	4413.69	10	<0.001
`sexe`	7994.88	1	<0.001
`taille`	2345.19	2	<0.001

Annexe 2: Inoccupation des refuges à Toronto

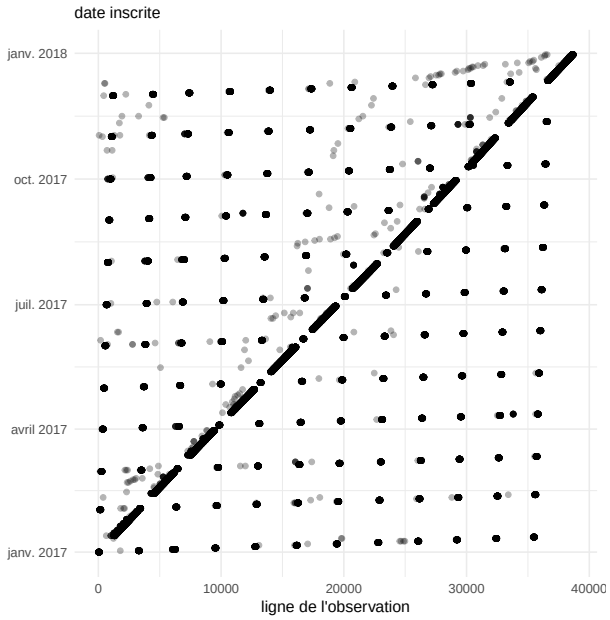

Figure A1 – *Nuage de points de la date inscrite pour le décompte par refuge en ordre chronologique (ordonnée) en fonction du numéro de ligne de l'observation (abscisse).*

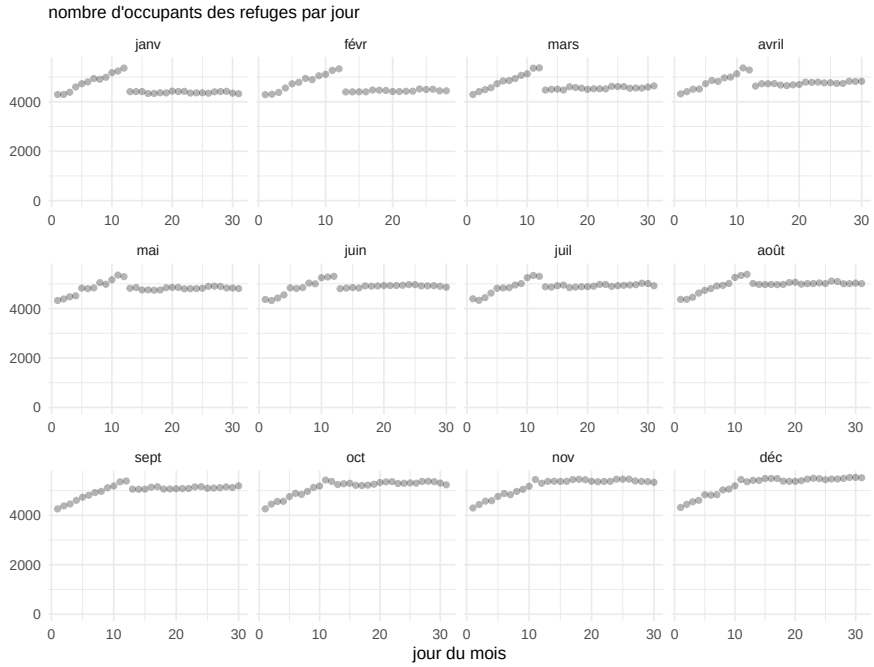

Figure A2 – *Décompte de la fréquentation des refuges de Toronto en 2017 par mois et jour du mois.*