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 0 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:	Total
Points:	0
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]

Solution: La sensibilité est de 92/141, soit 65%. C'est le pourcentage de personnes atteintes de la Covid que le test rapide permettrait de détecter.

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

Solution: C'est le pourcentage de personnes atteintes de la Covid que le test rapide permettrait de détecter. Oui, dans la mesure où le test rapide est peu fiable et le coefficient reproducteur du virus était très élevé: ne pas détecter une personne contaminée menait à un grand nombre d'infections subséquentes.

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]

> **Solution:** Le code crée une variable catégorielle (`factor`) avec valeur `petit` si le nombre annuel de tests de conduite du centre est inférieur à 1000, `moyen` s'il est entre 1000 et 2499 et `grand` sinon pour 2500 tests annuels et plus; cete variable est ajoutée à la base de données `gbconduite`.

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]

> **Solution:** Il suffit de trouver le coefficient le plus élevé (région [2], Angleterre du Nord-Est) et le moins élevé (région [7], Grand Londres).

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]

> **Solution:** Le rapport de cote petit versus moyen est 1.443/1.137=1.269 et la borne inférieure de l'intervalle de confiance à niveau 95% pour cette différence est de $\exp(0.35 - 0.121) = 1.257$; puisque l'intervalle n'inclut pas un rapport de cote de 1, l'augmentation est significative.

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

> **Solution:** Toute chose étant égale par ailleurs, les hommes ont une probabilité plus élevée de passer leur premis puisque la cote des hommes est 32.4% plus élevée que celle des femmes.

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.

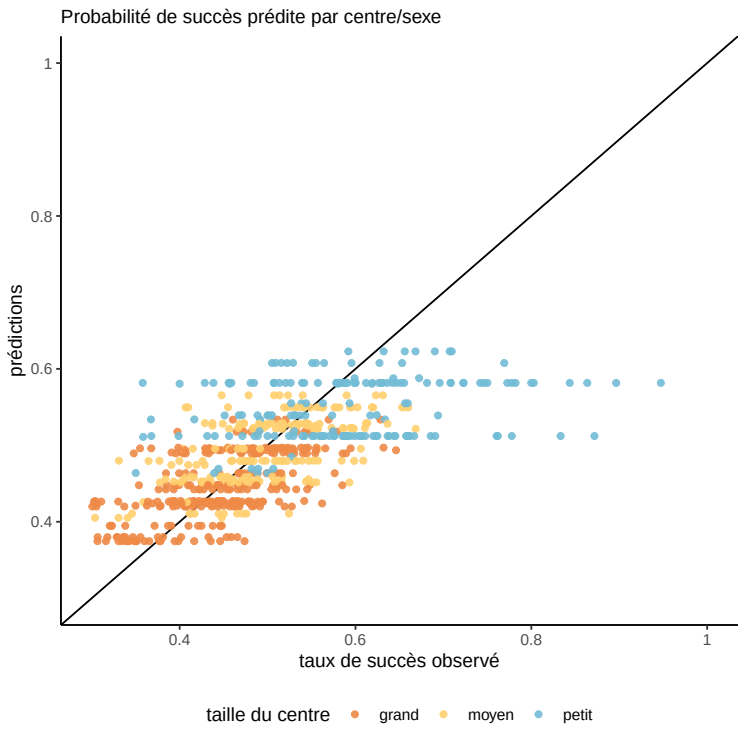

Figure 1 – *Nuage de points des prédictions de la probabilité de réussite versus le taux de 2018, en fonction de la taille du centre.*

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]

> **Solution:** Non. Si la corrélation entre les prédictions des taux de réussite et les taux observés est positive, le modèle est clairement incapable de capturer la plus grande volatilité des petits centres. L'étendu des prédictions est beaucoup plus faible: les plus petits centres ont moins de poids dans l'ajustement parce qu'il y a moins de tests là-bas comparativement à l'ensemble de la Grande-Bretagne.

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.

> **Solution:** On voit qu'il n'y a pas de petits centres dans les régions Est et Sud-Est, donc impossible d'estimer l'effet. Si l'interaction risque d'améliorer le modèle (on a beaucoup d'observations!), cela serait une forme de surajustement . . . il vaudrait plutôt mettre le nombre de tests effectués comme variable.

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.

Solution: On s'attendrait à ce que les données soient ordonnées de manière croissante, mais ce n'est pas le cas. Il y a des données rapportées systématiquement dans les mois subséquents dans la Figure A1, à intervalle régulier. On note également un report avec une tendance haussière pour les douze premiers jours de chaque mois, ce qui sous-tend que les champs pour les mois et les jours sont permutés pour certaines observations.

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]

> **Solution:** Faux; la procédure séquentielle utilise un algorithm glouton et ne considère typiquement qu'un seul modèle avec un nombre donné de variables. L'énoncé décrit la procédure exhaustive.

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]

> **Solution:** Vrai; la procédure ajoute la variable qui améliore le plus l'ajustement parmi celles qui ne sont pas déjà dans le modèle à chaque étape, pourvu que la valeur du critère d'Akaike diminue. Or, ce critère pénalise moins et l'historique BIC sera terminé avant ou à la même itération que celui de l'AIC.

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]

> **Solution:** Faux, maximiser la sensibilité revient à déclarer toutes les observations comme des cas positifs, cette stratégie ne nécessite pas de modèle.

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]

> **Solution:** La courbe qui remonte plus rapidement correspond au BIC, la courbe traitillée au AIC qui pénalise moins.

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]

Laquelle des sorties correspond au résultat attendu?

> **Solution:** L'aire sous la courbe de la fonction ROC est nécessairement inférieure à 1, puisque ce graphique représente le couple (sensibilité, 1-spécificité). C'est donc la courbe A.

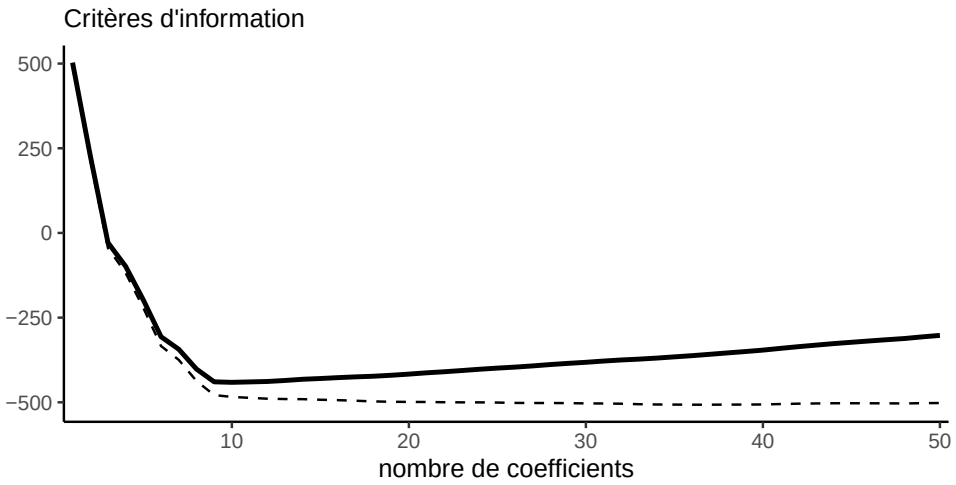

Figure 2 – *Valeurs des critères d'information AIC et BIC en fonction du nombre de coefficients du modèle.*

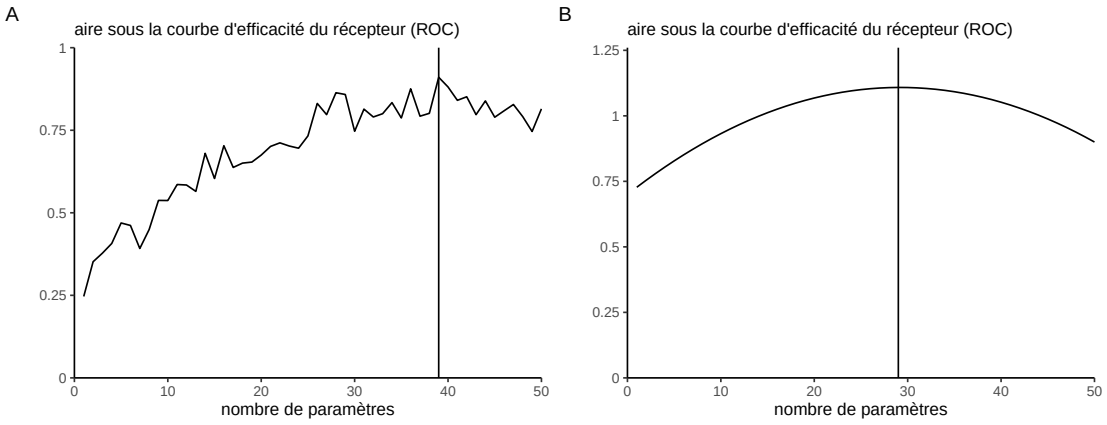

Figure 3 – *Courbes candidates en sortie de la procédure décrite à la question 4.5*