POLYTECHNIQUE MONTRÉAL
LE GÉNIE EN PREMIÈRE CLASSE

DÉPARTEMENT DE MATHÉMATIQUES ET DE GÉNIE INDUSTRIEL

MTH2302D - PROBABILITÉS ET STATISTIQUE

Devoir - Automne 2024

Date de remise : 4 décembre avant 23h59 (dans Moodle)

DIRECTIVES :

- ✓ Vous devez remettre un rapport par équipe de 2 (au maximum) au plus tard mercredi le **4 décembre avant 23h59**, dans Moodle, sous la forme d'un fichier électronique de format PDF nommé **`matricule1_matricule2.pdf`**. Le rapport doit contenir votre nom, prénom, matricule et toutes les informations requises sur la page de présentation dont un modèle est disponible sur le site du cours. Aucune remise papier ou par courriel ne sera acceptée. La note **zéro** sera attribuée à toute remise qui ne respecte pas les directives, ainsi qu'à toute équipe constituée de **plus de 2** membres.
- ✓ Pour vos analyses, vous devez d'abord obtenir avec votre matricule (un des deux matricules) un ensemble personnalisé de données. Toutes vos réponses doivent correspondre à votre ensemble de données. **Veuillez consulter les instructions** de la procédure sur la page de présentation qui est disponible sur le site Moodle.
- ✓ Chacune de vos réponses doit être **complète**, **expliquée** et **justifiée**. Lors de la correction, il sera tenu compte de la qualité de la présentation, la pertinence des analyses et l'initiative dont vous ferez preuve dans votre rapport.
  Sur les **40** points, **38** sont alloués aux analyses, commentaires pertinents, etc. et **2** à la présentation.
- ✓ Les analyses, les tableaux et les graphiques du rapport doivent être produits avec le logiciel R (ou Python).
- ✓ Lorsque nécessaire et selon le contexte, utiliser un seuil critique de **5 %**, ou un niveau de confiance de **95 %**.
- ✓ **Rappel** : L'usage des SIAG (systèmes d'intelligence artificielle générative, ex : ChatGPT, OpenAI Codex, GitHub Copilot, DALL-E, Midjourney, etc.) est totalement proscrit dans ce cours. Tout cas de soupçon de fraude sera automatiquement rapporté au comité de discipline étudiante et pourrait mener à l'attribution d'une note **F** au cours.

CONTEXTE.

Le devoir est une étude de cas qui consiste en une analyse de données (recueillies durant une période de temps) relatives aux ventes de sièges d'automobile pour enfants d'un fabricant.

Les données. Les données à analyser sont constituées d'un échantillon de **200** observations (points de vente) formées de quatre variables mesurant un certain nombre de caractéristiques socio-économiques à différents points de vente répartis dans plusieurs villes. Le Tableau 1 ci-dessous présente les variables de l'étude (numéro de colonne dans le fichier, symbole, nom, et description).

Col. n°	Symbole	Nom	Description
1	$--$	Identification	Un code assigné à l'observation (ne pas en tenir compte)
2	Y	Ventes (`Sales`)	Nombre de sièges vendus (en milliers) au point de vente
3	X_1	Prix (`Price`)	Prix du siège du fabricant au point de vente (en $)
4	X_2	Âge (`Age`)	L'âge moyen de la population au point de vente.
5	X_3	Lieu (`Region`)	Lieu du point de vente : urbain (**1**) ou rural (**0**).

Tableau 1 : Les variables de l'analyse.

Vous êtes chargé d'effectuer une analyse de ces données afin de d'établir les liens possibles entre différentes variables et de déterminer un modèle statistique permettant de prévoir les ventes (*Sales*) des sièges d'automobile en fonction de certaines variables.

Phase 1 **: Analyse statistique descriptive et inférence**.

On vous demande de répondre aux questions suivantes en utilisant des techniques appropriées de statistique (statistique descriptive et inférence), illustrées avec des diagrammes pertinents.

a) (7 points) Pour la variable *Ventes* (*Sales*), en utilisant l'ensemble des données, produisez les graphiques et les tableaux demandés et interprétez brièvement les résultats dans chacun des cas suivants :

- un histogramme et un diagramme de Tukey (ou «Box Plot») ;
- une droite de Henry (ou «Normal Probability Plot») et un test de normalité (Shapiro-Wilk) ;
- un tableau de statistiques descriptives comprenant : *moyenne, quartiles, écart type, intervalle de confiance pour la moyenne* ;

b) (10 points) Afin de vérifier si les ventes (*Sales*) sont affectées par le lieu du point de vente, on peut considérer deux groupes de points de vente selon la variable *Region* et effectuer une comparaison des ventes des deux groupes en termes de moyenne, symétrie et variabilité. Pour ce faire, effectuez les analyses suivantes et donnez une brève conclusion :

- deux histogrammes juxtaposés et deux diagrammes de Tukey (ou «Box Plot») juxtaposés ;
- un tableau des statistiques descriptives par groupe : *moyenne, quartiles, variance, écart type, intervalle de confiance pour la moyenne* ;
- un test d'hypothèses sur l'égalité des variances pour les deux groupes et concluez ;
- un test d'hypothèses sur l'égalité des moyennes pour les deux groupes et concluez.

***Phase 2* : Recherche du meilleur modèle**.

On s'intéresse dans cette phase à la détermination d'un modèle permettant d'expliquer le niveau des ventes en fonction des différentes variables considérées. Pour ce faire, on envisage des modèles de régression simple en considérant les ventes (`Sales`) comme variable dépendante Y.

c) (15 points) On considère les huit modèles suivants :

	Modèle 1 : $Y = \beta_0 + \beta_1 X_1 + \varepsilon;$	**Modèle 2 :** $Y = \beta_0 + \beta_1 X_2 + \varepsilon;$
	Modèle 3 : $Y = \beta_0 + \beta_1 X_1^2 + \varepsilon;$	**Modèle 4 :** $Y = \beta_0 + \beta_1 X_2^2 + \varepsilon;$
modèle 3	**Modèle 5 :** $Y = \beta_0 e^{\beta_1 X_1 + \varepsilon};$	**Modèle 6 :** $Y = \beta_0 e^{\beta_1 X_2 + \varepsilon};$
modèle 2	**Modèle 7 :** $Y = \beta_0 X_1^{\beta_1} e^{\varepsilon};$	**Modèle 8 :** $Y = \beta_0 X_2^{\beta_1} e^{\varepsilon},$

où, dans chaque modèle, β_0 et β_1 sont des paramètres et ε une erreur aléatoire que l'on suppose de loi $N(0,\ \sigma^2)$.

Remarque : Les coefficients β_0 et β_1 ainsi que l'erreur ε ne sont pas les mêmes d'un modèle à l'autre.

Pour chacun des huit modèles ci-dessus :

- *(5 points)* Effectuez l'ajustement (i.e. obtenir le tableau des coefficients de régression, le tableau d'analyse de la variance).
- *(5 points)* Tester la signification du modèle et effectuez une analyse des résidus (normalité, homoscédasticité, points atypiques, etc.)
- *(2 points)* Donner un intervalle de confiance pour chacun des paramètres β_0 et β_1.
- *(3 points)* En conclusion : effectuez une comparaison et dire lequel des six modèles est préférable aux autres. Justifiez votre choix en précisant les critères utilisés.

d) (6 points) Sur la base du meilleur modèle que vous avez obtenu en **c)**, calculez un intervalle de prévision des ventes pour un point de vente ayant les caractéristiques suivantes : $X_1 = 115; X_2 = 35; X_3 = 1$. Commentez brièvement votre résultat.

Remarque. Notez que le modèle que vous avez obtenu en **c)** n'utilise pas nécessairement toutes les valeurs ci-dessus.