5.1 **Cartel des prix de l'essences en Gaspésie** : Plusiers maires et préfets gaspésiens ont demandé à la Régie de l'énergie du Québec d'enquêter sur le prix de l'essence, beaucoup plus élevé en 2019 selon eux dans la région qu'ailleurs au Québec. Pour répliquer l'analyse de la Régie, les données suivantes ont été extraites du site de l'organisme gouvernemental pour la période 2014–2019. Les données `renergie` incluent les variables suivantes :

- `region` : région administrative, une parmi Bas-Saint-Laurent (1), Saguenay-Lac-Saint-Jean (2), Capitale-Nationale (3), Mauricie (4), Estrie (5), Montréal (6), Outaouais (7), Abitibi-Témiscamingue (8), Côte-Nord (9), Nord-du-Québec, excluant le Nunavik (10), Gaspésie-Îles-de-la-Madeleine (11), Chaudière-Appalaches (12), Laval (13), Lanaudière (14), Laurentides (15), Montérégie (16) et Centre-du-Québec (17).
- `date` : date hebdomadaire pour les prix minimum et moyens à la pompe en format `aaaa-mm-jj`.
- `pmin` : prix minimum (plancher) calculé par la Régie de l'énergie, incluant les taxes et les frais de transports.
- `pmoy` : prix moyen à la pompe affiché par les détaillants.

Faites une analyse des données longitudinales pour déterminer si la marge de profit des détaillants de la Gaspésie et des Îles-de-la-Madeleine est significativement plus élevée que partout ailleurs à l'aide d'une analyse de variance à un facteur qui prenne en compte la corrélation intra-région. *Indication : dans SAS, utilisez l'option `ddfm=satterth` pour le calcul des degrés de liberté avec la procédure `mixed`.*

(a) Tracez un graphique (a) du prix moyen et (b) de la différence entre le prix moyen et le prix minimum pour chaque région en fonction du temps. Commentez sur les différences observées entre ces deux graphiques.

(b) Sélectionnez un modèle de covariance adéquat pour modéliser la dépendence temporelle intra-région. Vous devez choisir un modèle de covariance parmi (a) indépendance (covariance diagonale), (b) AR(1), (c) équisymmétrie et (d) non-structuré. Justifiez adéquatement votre choix.

(c) Rapportez les erreurs-type de la marge de profit moyenne des détaillants de la région Gaspésie-Îles-de-la-Madeleine (GIM) pour le modèle de régression ordinaire (qui suppose l'indépendance entre observations) et le modèle autorégressif d'ordre 1. Laquelle est la plus grande? expliquez pourquoi.

(d) Calculez la différence entre la marge de profit des détaillants de la Gaspésie-Îles-de-la-Madeleine et des autres régions en prenant en compte la corrélation intra-région. Quelles différences sont statistiquement significatives?

5.2 **Enseignement de la lecture** : les données sont tirées de

> J. Baumann, N. Seifert-Kessell, L. Jones (1992), *Effect of Think-Aloud Instruction on Elementary Students' Comprehension Monitoring Abilities*, Journal of Reading Behavior, **24** (2), pp. 143–172.

Ces chercheurs ont fait une étude pour déterminer l'efficacité relative de méthodes d'apprentissage de la lecture. L'échantillon de 66 élèves de quatrième année primaire comporte 32 filles et 34 garçons qui ont été alloués de façon aléatoire aux trois groupes. On s'intéresse à l'amélioration des capacités de lecture par rapport à une méthode d'enseignement traditionnelle (DR). Deux tests ont été administrés avant et pendant l'expérience pour mesurer l'efficacité des méthodes ; afin de rendre les résultats comparables, ils ont été repondérés de telle sorte à ce que une note parfaite vaille 1.

Les données contiennent des renseignements sur

- `groupe` : unité expérimentale, une parmi lecture-experimental group, une parmi *directed reading-thinking activity* (`DRTA`), méthode de la pensée à voix haute (`TA`) et lecture dirigée (`DR`).
- `mpre` : moyenne du score de prédiction pré-intervention (standardisé) pour les tests de détection d'erreurs et decompréhension.
- `mpost` : même que `mpre`, mais pour les évaluations post-intervention.

Nous sommes intéressés tout d'abord par l'amélioration pour chacune des méthodes à l'aide de deux modèles.

(a) Dans leur article, Baumann *et al.* font une analyse de variance à un facteur pour les scores pré-intervention (`mpre`) avec le facteur `groupe`. Expliquez quel est l'utilité d'un tel test dans le contexte de l'étude.

(b) Soit `dpp` = `mpost` − `mpre` la différence entre résultats standardisés post- et pré-intervention. Ajustez une analyse de variance à un facteur pour `dpp` avec le facteur `groupe` (modèle 1.1). Écrivez l'équation du modèle

ajusté en termes de scores pré- et post-intervention et montrez que le modèle est un cas spécial d'un modèle de régression linéaire pour `mpost` avec un terme de décalage (*offset* en anglais).

(c) Comparez le modèle d'analyse de variance à un facteur pour `dpp` avec `groupe` (modèle 1.1) à un modèle linéaire ayant `mpost` comme variable réponse et `mpre` et `groupe` comme variables explicatives (modèle 1.2). Au vu de l'ajustement de ce dernier, est-ce que le modèle d'analyse de variance est adéquat? Justifiez votre réponse.

(d) Transformez les données de format court à format long; ce dernier est plus convenable pour l'analyse de données longitudinales. En plus de `groupe`, vos données devraient contenir les colonnes suivantes
- `id` : identifiant de l'étudiant.
- `score` : moyenne pour l'évaluation.
- `test` : variable catégorielle, une de `mpost` ou `mpre`, qui renseigne sur le score correspond à la moyenne pré-intervention ou post-intervention.

Le Table 1 présente le format final que vous devriez obtenir.

groupe	test	score	id
DR	mpre	0,23	1
DR	mpost	0,27	1
DR	mpre	0,35	2
DR	mpost	0,42	2
DR	mpre	0,41	3
DR	mpost	0,24	3
DR	mpre	0,57	4
DR	mpost	0,39	4
DR	mpre	0,67	5
DR	mpost	0,56	5

TABLE 1 – Premières 10 lignes des données de Baumann en format long

Comparez deux modèles pour `score` en fonction du `groupe` et de `test`, en incluant un terme d'interaction entre les deux, mais avec des modèles de covariance intra-individus différents :
- modèle 5.2.3 avec une structure d'équicorrélation pour les erreurs;
- modèle 5.2.4 avec une covariance non structurée.

Expliquez quel est la différence fondamentale entre les modèles 5.2.3-.4 et 5.2.2. Écrivez la matrice de covariance des erreurs pour le modèle 5.2.3 et rapportez les corrélations estimées entre pré-interventions et post-interventions pour un(e) étudiant(e).

(e) À l'aide des sorties des modèles 5.2.3 et 5.2.4, testez si la variance des scores moyens pré-intervention et post-intervention sont les mêmes. Écrivez le nom du test que vous utilisez, la valeur numérique de la statistique et calculez la valeur-p avant de conclure dans le contexte du problème.

(f) Puisqu'on a affaire à des données longitudinales, il serait logique de considérer en plus des deux modèles de covariance, un modèle autorégressif d'ordre 1, ou modèle AR(1), pour les erreurs. Est-ce que ça serait utile dans ce cas? Justifiez votre réponse.

(g) Jusqu'à présent, on a supposé que les résultats pré- et post-intervention de tous les élèves avaient la même matrice de covariance. On pourrait cependant supposer que les paramètres de cette matrice de covariance diffèrent d'une méthode d'apprentissage à une autre. Est-ce que les données corroborent cette hypothèse?

(h) Utilisez le modèle 5.2.4 pour déterminer si les résultats pour les méthodes d'enseignement `DRTA` et `TA` sont significativement meilleures que la méthode standard `DR`.

5.3 **Tolérance d'adolescents face à la délinquance** : Les données proviennent du *American National Longitudinal Survey of Youth*, une étude longitudinale qui a démarré en 1997 et qui suit une cohorte de jeunes Américains nés entre

1980 et 1984. Un total de 8984 participants âgés de 12 à 17 sont inclus pour la première fois en 1997 et a été suivie à 15 reprises jusqu'à maintenant.

On considère 16 individus qui ont répondu aux cinq premières vagues d'entrevue entre l'âge de 11 et 15 ans, avec un suivi annuel. , et en particulier à certaines variables mesurant la tolérance des jeunes face aux comportements délinquants. Les données disponibles nous permettent de suivre l'évolution de 16 jeunes chaque année entre l'âge de 11 ans et 15 ans (donc 5 mesures pour chaque individu). Chaque année au début de l'étude puis tous les 2 ans ensuite, les participants ont rempli un questionnaire permettant d'évaluer leur tolérance face aux comportements délinquants. À l'aide d'une échelle à 4 points où les choix sont « très mal » (1), « mal » (2), « un peu mal » (3) et « tout à fait acceptable » (4), les jeunes indiquaient s'ils qualifiaient de « mal » pour quelqu'un de leur âge de (a) tricher à un examen, (b) détruire le bien d'autrui à dessein, (c) fumer de la marijuana, (d) voler quelque chose d'une valeur de moins de 5 dollars, (e) frapper ou menacer quelqu'un sans raison, (f) consommer de l'alcool, (g) entrer par effraction dans un bâtiment ou véhicule afin de voler, (h) vendre des drogues dures et (i) voler quelque chose d'une valeur de plus de 50$. Chaque score a été mesuré sur une échelle de Likert allant de très mal (1) à complètement acceptable (4). Les données `tolerance` incluent les variables suivantes :

- `id` : identifiant du participant.
- `age` : âge du participant lors du suivi
- `tolerance` : moyenne du score pour les neuf questions sur la tolérance à la délinquence.
- `sexe` : indicateur binaire, un pour les hommes, zéro pour les femmes.
- `exposition` : score moyen de l'exposition du participant à 11 ans aux comportements délinquants dans son entourage. Cette variable est une estimation du participant de la proportion de ses ami(e)s qui ont été impliqué(e)s dans chacune des activités (a) à (i) décrites plus haut.

(a) Présentez et interprétez les statistiques descriptives des variables `tolerance`, `sexe` and `exposition`.

(b) Évaluez graphiquement la relation entre `tolerance` et les variables `sexe`, `exposition` et `age`. Résumez brièvement vos observations.

(c) Faites un graphique des trajectoires de la tolérance aux comportements délinquants en fonction de l'âge du participant et commentez.

(d) En prenant en compte la corrélation intra-sujet (au cours des cinq années) et en assumant cette dernière fixe peu importe l'année, ajustez le modèle

$$\begin{aligned} \texttt{tolerance}_{ij} &= \beta_0 + \beta_1 \texttt{sex}_{ij} + \beta_2 \texttt{exposure}_{ij} + \beta_3 \texttt{age}_{ij} + \varepsilon_{ij}, \\ \boldsymbol{\varepsilon}_i &\overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}_5, \Sigma_i), \boldsymbol{\Sigma}_i \sim \mathsf{CS}, i = 1, \ldots, 16; j = 1, \ldots, 5. \end{aligned} \tag{M_1}$$

   i. Interprétez l'effet des variables du modèle et commentez sur les résultats.

   ii. Calculez la corrélation entre deux valeurs de tolérances pour les mesures à l'âge 11 et 12 ans, de même qu'entre 11 et 15 ans.

(e) En supposant une structure autorégressive d'ordre 1 pour les erreurs, ajustez le modèle

$$\begin{aligned} \texttt{tolerance}_{ij} &= \beta_0 + \beta_1 \texttt{sex}_{ij} + \beta_2 \texttt{exposure}_{ij} + \beta_3 \texttt{age}_{ij} + \varepsilon_{ij}, \\ \boldsymbol{\varepsilon}_i &\overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}_5, \Sigma_i), \boldsymbol{\Sigma}_i \sim \mathsf{AR}_1, i = 1, \ldots, 16; j = 1, \ldots, 5. \end{aligned} \tag{M_2}$$

   i. Identifiez tous les paramètres de la matrice de covariance.

   ii. Calculez la corrélation entre deux valeurs de tolérances pour les mesures à l'âge 11 et 12 ans, de même qu'entre 11 et 15 ans et comparez ces corrélations avec celles obtenues pour le modèle d'équicorrélation.

(f) En supposant l'indépendance des observations, ajustez le modèle

$$\begin{aligned} \texttt{tolerance}_{ij} &= \beta_0 + \beta_1 \texttt{sex}_{ij} + \beta_2 \texttt{exposure}_{ij} + \beta_3 \texttt{age}_{ij} + \varepsilon_{ij}, \\ \boldsymbol{\varepsilon}_i &\overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}_5, \sigma^2 \mathbf{I}_5), i = 1, \ldots, 16; j = 1, \ldots, 5, \end{aligned} \tag{M$_3$}$$

Lequel des modèles M_1, M_2 et M_3 choisiriez- vous? Justifiez votre choix.

Compilé le 02/10/2020 à 12:58