6.1 Les données ainsi que la description qui suit ont été adaptées d'un exemple d'OpenBUGS et proviennent de l'étude

> H. Goldstein *et al.* (1993). *A Multilevel Analysis of School Examination Results*, Oxford Review of Education, **19** (4), pp. 425–433.

Les auteurs analysent les résultats d'examens d'écoliers des écoles dans les districts centraux du Grand Londres, et étudient la variabilité des résultats entre les écoles afin d'en faire un classement. Les scores moyens d'examen de 1978 écoliers de 38 écoles différentes ont été standardisés et rendus disponibles pour cette étude. On a pour chaque élève le sexe, son score au test de lecture de Londres (`TLL`) et un rang pour un test de raisonnement verbal (`rv`) passé au début de l'année lorsque l'élève était âgé de 11 ans. Les deux tests sont effectués au début de l'année scolaire. Chaque école a été classifiée selon type (école pour filles, école pour garçons ou école mixte, et selon sa dénomination (école anglicane affiliée à Église d'Angleterre, école catholique, école publique, ou autre). Ces deux critères sont utilisés comme variables catégorielles spécifiques à chaque école. Les données `goldstein` contiennent les variables suivantes :

- `score` : score standardisé à l'examen de fin d'année pour chaque élève,
- `ecole` : identifiant de l'école,
- `TLL` : score au test de lecture de Londres
- `rv` : catégorie du test de raisonnement verbal (`1`, `2` ou `3`, où `1` représente le groupe avec la plus haute aptitude et `3` celui avec la plus faible),
- `sexe` : sexe de l'élève, soit fille (`0`) ou garçon (`1`),
- `type` : variable catégorielle pour le type de l'école, soit `filles`, `garcons` ou `mixte`.
- `denom` : variable catégorielle pour la dénomination de l'école, soit Église d'Angleterre (`angli`), catholique (`catho`), école publique (`etat`), ou `autre`.

(a) Donnez l'étendue du nombre d'écoliers par école et utilisez cette information afin de déterminer s'il est possible d'estimer un effet de groupe fixe pour chaque école.

(b) Écrivez l'équation du modèle postulé pour la variable `score` incluant les variables `TLL`, `rv`, `sexe`, `type` et `denom` comme effets fixes et un effet aléatoire pour la variable `ecode`. Dans votre modèle, utilisez les catégories de référence `rv=3`, `mixte` pour `type` et `autre` pour `denom`.

(c) On pourrait considérer un modèle où la variable `rv` a un effet aléatoire plutôt qu'un effet fixe. Lequel des deux modèles vous semble le plus adéquat et qu'elle est la différence conceptuelle entre ces deux modèles ?

(d) En utilisant le modèle ajusté avec un effet aléatoire pour `ecole`, calculez la matrice de covariance estimée pour l'école 37 et expliquez comment les termes de cette matrice ont été obtenus à partir des estimés des paramètres de covariance. Calculez la proportion de la variance totale qui est due à l'effet de groupe pour l'école.

(e) Produisez un diagramme quantile-quantile normal des effets aléatoires prédits pour l'école. Commentez sur l'hypothèse du modèle quant aux effets aléatoires.

(f) Le but de l'étude de Goldstein *et al.* (1993) était de classer les écoles londoniennes. Quel est le bienfait de combiner toutes les informations provenant des différentes écoles afin d'estimer les scores moyens des écoliers ? Représentez les effets de groupe prédits en fonction de l'identifiant de l'école (du groupe) ainsi que les intervalles de prédiction en vous basant sur la formule $\widehat{b}_i \pm 1.96\text{se}(\widehat{b}_i)$ (il se peut que vous ayez besoin de calculer les bornes des intervalles manuellement). Selon ces prédictions, quel est le classement des cinq meilleurs écoles ? Qu'est-ce qui manque pour que ce classement soit adéquat et quelle sont les limitations de cette approche (ne considérer que l'effet aléatoire conditionnel) ?

6.2 **Résultats de l'examen GSCE** : Au Royaume-Uni, un examen national (GCSE) est administré à chaque année à la fin du secondaire. La réussite de cette évaluation mène à l'obtention du diplôme d'études secondaires. L'objectif de cette étude est de vérifier s'il y a un lien entre le sexe du candidat et sa performance au test. La base de données de cette mise en contexte contient 1905 observations pour les 72 centres d'examen choisis au hasard aux Royaume-Uni. Voici la liste des variables utilisées :

— `centre` : variable catégorielle identifiant les 72 centres d'examen.
— `sexe` : variable binaire indiquant le sexe du candidat, soit garçon (`0`), soit fille (`1`).
— `resultat` : résultat obtenu à l'examen GCSE.
— `score` : résultat cumulé pendant l'année scolaire dans les cours évalués par l'enseignant(e) du candidat ou de la candidate.

Ajustez les trois modèles suivants afin expliquer la variable réponse `resultat` en fonction de `sexe` et `score` :

— modèle 2.1 : modèle de régression linéaire avec des erreurs homoscédastiques et équicorrélées.
— modèle 2.2 : modèle de régression linéaire avec la variable explicative additionnelle `centre` pour tenir compte de l'effet du centre, en supposant que les erreurs sont indépendantes.
— modèle 2.3 : modèle de régression linéaire mixte avec un effet aléatoire sur l'ordonnée à l'origine du modèle et en supposant que les erreurs sont indépendantes.

(a) Expliquez brièvement pourquoi il serait légitime de modéliser la corrélation dans ces données. Dites ce que le « groupe » représente dans ce contexte.

Solution
Les différents centres peuvent capturer des disparités géographiques pour le revenu ou les ressources, lesquelles peuvent affecter la performance des élèves. Le « groupe » est `centre`.

(b) Quel est l'avantage principal de l'utilisation du modèle 6.2.3 par rapport au modèle 6.2.1 ?

Solution
On peut obtenir une prédiction pour les effets de groupe pour chaque `centre`.

(c) Quels sont les deux avantages principaux du modèle 6.2.3 par rapport au modèle 6.2.2 ?

Solution
L'ordonnée à l'origine aléatoire induit naturellement de la corrélation entre observations au sein d'un même centre. Comme les ordonnées à l'origine aléatoire ne sont pas des paramètres estimés, mais qu'on peut les prédire, cela permet d'estimer des coefficients même si la taille de l'échantillon est petite ou si d'autres coefficients ne sont pas estimables.

(d) Pour chacun des modèles 6.2.1 et 6.2.3, écrivez les matrices postulées de covariance **intra-centre** pour un centre de trois candidats pour la variable réponse Y **et** les erreurs ε.

Solution
Pour le modèle 6.2.1,

$$\mathsf{Cov}\left(\boldsymbol{y}\right) = \begin{pmatrix} \sigma^2+\tau & \tau & \tau \\ \tau & \sigma^2+\tau & \tau \\ \tau & \tau & \sigma^2+\tau \end{pmatrix}, \qquad \mathsf{Cov}\left(\boldsymbol{\varepsilon}\right) = \mathsf{Cov}\left(\boldsymbol{y}\right),$$

tandis que pour le modèle 6.2.3,

$$\mathsf{Cov}\left(\boldsymbol{y}\right) = \begin{pmatrix} \sigma^2+\sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma^2+\sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma^2+\sigma_b^2 \end{pmatrix}, \qquad \mathsf{Cov}\left(\boldsymbol{\varepsilon}\right) = \sigma^2\mathbf{I}_3.$$

(e) À combien estime-t-on la corrélation entre deux résultats de deux candidats ayant passé leur examens dans deux centres différents selon le modèle 6.2.3 ?

Solution
Zéro, puisqu'on assume que les individus de centres différents sont indépendants.

(f) Quelle est la corrélation estimée entre deux individus d'un même centre selon le modèle 6.2.3 ?

Compilé le 02/10/2020 à 13:40

Solution
La corrélation estimée est $\widehat{\sigma}_b^2/(\widehat{\sigma}_b^2+\widehat{\sigma}^2) = 107,20/(240,72+107,2) = 0,308$.

(g) Est-ce qu'un modèle autorégressif d'ordre un, AR(1), serait adéquat pour modéliser la corrélation intra-groupe dans ce problème. Justifiez brièvement votre réponse.

Solution
Non, parce que le modèle de corrélation AR(1) suppose que la corrélation décroît selon la distance, or les individus sont échangeables.

(h) À l'aide du modèle 6.2.3, prédisez le résultat au GSCE pour une fille qui passe son examen au centre numéro 2 et qui a obtenu un résultat cumulé de 91 points dans ses cours durant l'année.

Solution
Le score prédit est $34,2301 + 91 \times 0,5993 - 8,4188 - 7,9898 = 72,36$ points.

(i) À l'aide du modèle 6.2.3, prédisez le résultat au GSCE pour un garçon qui a obtenu un résultat cumulé de 100 points et qui passe son examen dans un nouveau centre.

Solution
Le score moyen prédit pour le GSCE est de $34,23 + 59,93 = 94,16$ points.

(j) Est-ce que les résultats des élèves au GSCE diffèrent significativement selon leur `sexe` selon le modèle 6.2.3? Justifiez votre réponse et interprétez le coefficient estimé.

Solution
Selon la sortie, la statistique de Wald pour le coefficient `sexe`, qui correspond à $\mathscr{H}_0 : \beta_{\texttt{sexe}} = 0$ vaut $-10,96$ (c'est également la racine carrée de la statistique-F, qui vaut $120,17$). Il y a donc une différence significative entre garçons et filles (valeur-p négligeable) avec une différence estimée de $8,42$ points en faveur des filles (donc les filles ont en moyenne,*ceteris paribus*, un score 8,42 points supérieur à celui des garçons).

Compilé le 02/10/2020 à 13:40