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.

Solution

Il y a entre 1 et 198 élèves par école. Sept écoles ont moins de cinq élèves, il est impossible d'estimer de manière fiable un effet fixe pour ces dernières.

(b) Écrivez l'équation du modèle théorique postulé pour la variable `score` incluant les variables `TLL`, `rv`, `sexe`, `type` et `denom` comme effets fixes et un effet aléatoire pour la variable `ecole`. Dans votre modèle, utilisez les catégories de référence `rv=3`, `mixte` pour `type` et `autre` pour `denom`.
N'oubliez pas de spécifier la loi aléatoire des erreurs et des effets aléatoires et leurs relations.

Solution

L'équation du modèle est

$$\begin{aligned}\texttt{score}_{ij} &= \beta_0 + b_i\beta_1\texttt{TLL}_{ij} + \beta_2\texttt{rv}_{ij1} + \beta_3\texttt{rv}_{ij2} + \beta_4\texttt{sexe}_{ij} + \beta_5\texttt{type}_{ij\texttt{garcons}} \\ &\quad + \beta_6\texttt{type}_{ij\texttt{filles}} + \beta_7\texttt{denom}_{ij\texttt{angli}} + \beta_8\texttt{denom}_{ij\texttt{catho}} + \beta_9\texttt{denom}_{ij\texttt{etat}} + \varepsilon_{ij}\end{aligned}$$

où i est l'indice pour l'école et j l'identifiant de l'étudiant au sein de l'école. On postule que $b_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0,\sigma_s^2)$ et $\varepsilon_{ij} \stackrel{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2)$, en plus de supposer l'indépendance entre ordonnées aléatoires (par école) et erreurs.

(c) Interprétez les effets des variables explicatives dont l'effet est statistiquement différent de zéro

Solution

Selon les statistiques de Wald multidimensionnelles (type III), toutes les variables explicatives sauf `type` sont globalement significatives (note : avec d'autres statistique, `denom` pourrait être ou ne pas être significatif). Les estimés des paramètres et leur interprétation sont les suivantes.

— L'ordonnée à l'origine ($\widehat{\beta}_0 = -0,568$) donne le score moyen standardisé des filles d'écoles mixtes de dénomination autre qui ont obtenu la moyenne au test de lecture de Londres.

— L'estimé du coefficient β_1, $\widehat{\beta}_1 = 0,03$, représente l'augmentation du score standardisé pour chaque augmentation du score au test de lecture de Londres de 1 points, *ceteris paribus*.
— Toutes choses étant égales par ailleurs, les scores standardisés des filles sont $0,154$ points plus élevés que ceux des garçons sur l'examen de fin d'année.
— *Ceteris paribus*, la différence de score standardisé pour les étudiants d'écoles anglicanes/catholiques/-d'état par rapport aux dénominations autres sont de $\widehat{\beta}_7 = -0,27/\widehat{\beta}_8 = 0,186/\widehat{\beta}_9 = -0,17$, respectivement.

(d) 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?

Solution
Ici, un effet fixe serait plus logique (mais les deux possibilités sont faisables). Il y a seulement trois catégories et 1918 observations, suffisamment pour estimer l'effet de chaque modalité, qui sont mutuellement exclusives. L'effet est aussi statistiquement significatif. Une différence si on ajoute un effet aléatoire est la paramétrisation (la moyenne des effets aléatoires est zéro, donc on contraint la somme des effets aléatoires à zéro plutôt que d'imposer $\beta_{rv_3} = 0$ — bien que numériquement les résultats ne soient pas identiques, les différences de prédiction moyenne pour `rv=3` versus `rv=1` ou `rv=2` sont presques pareilles. L'ajout d'un effet aléatoire pour `rv` crée une structure d'équacorrélation pour les élèves qui sont dans la même catégorie de score de raisonnement verbal.

(e) 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.

Solution
Les estimés des paramètres de covariance sont $\widehat{\sigma}^2_s = 0.039$, et $\widehat{\sigma}^2 = 0.5715$ pour respectivement la variance de l'effet aléatoire de l'école et celle des erreurs. L'école 37 ne compte que quatre élèves, donc la matrice de covariance est

$$\mathsf{Var}\left(\boldsymbol{y} \mid \texttt{ecole} = 37\right) = \begin{pmatrix} \widehat{\sigma}^2_s & \widehat{\sigma}^2_s & \widehat{\sigma}^2_s & \widehat{\sigma}^2_s \\ \widehat{\sigma}^2_s & \widehat{\sigma}^2_s & \widehat{\sigma}^2_s & \widehat{\sigma}^2_s \\ \widehat{\sigma}^2_s & \widehat{\sigma}^2_s & \widehat{\sigma}^2_s & \widehat{\sigma}^2_s \\ \widehat{\sigma}^2_s & \widehat{\sigma}^2_s & \widehat{\sigma}^2_s & \widehat{\sigma}^2_s \end{pmatrix} + \begin{pmatrix} \widehat{\sigma}^2 & 0 & 0 & 0 \\ 0 & \widehat{\sigma}^2 & 0 & 0 \\ 0 & 0 & \widehat{\sigma}^2 & 0 \\ 0 & 0 & 0 & \widehat{\sigma}^2 \end{pmatrix} = \begin{pmatrix} 0,611 & 0,0395 & 0,0395 & 0,0395 \\ 0,0395 & 0,611 & 0,0395 & 0,0395 \\ 0,0395 & 0,0395 & 0,611 & 0,0395 \\ 0,0395 & 0,0395 & 0,0395 & 0,611 \end{pmatrix}$$

Le pourcentage estimé de la variance totale due à l'école est de 0,0646; cela correspond à la corrélation du modèle entre deux élèves d'une même école.

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

Solution
Il y a 38 écoles, donc on s'attend à ce que deux points en moyennes sortent des intervalles de confiance ponctuels obtenus par simulation. Le diagramme quantile-quantile du panneau droit de la Figure 1 n'indique aucune violation du postulat de normalité pour l'ordonnée à l'origine aléatoire par école. D'autres diagnostics graphiques des résidus suggèrent que les hypothèses de linéarité, d'hétéroscédasticité et de normalité des erreurs tiennent la route.

(g) 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 à l'aide d'effets aléatoires afin d'estimer les scores moyens des écoliers?

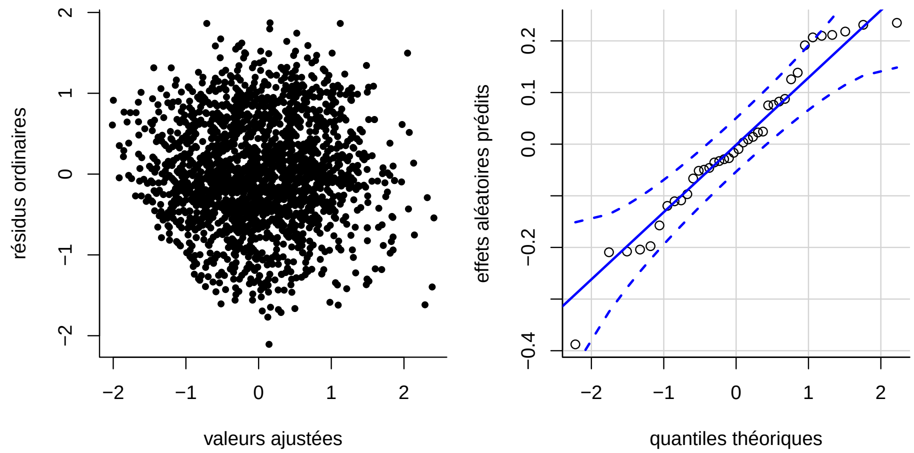

FIGURE 1 – Diagnostics graphiques : valeurs ajustées contre résidus ordinaires (gauche) et diagramme quantile-quantile des effets aléatoires prédits spécifiques à chaque école avec intervalles de confiance à 95% ponctuels (droite).

Solution
Combiner les observations permet de partager de l'information afin de mieux estimer l'effet des variables explicatives partagées et nous renseigne sur des bornes pour la variance dans les cas où on a peu d'élèves par école.

(h) Représentez graphiquement 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 $\widehat{b}_i \pm 1.96\mathrm{se}(\widehat{b}_i)$.

Solution
Le classement selon les effets aléatoires estimés contient les écoles 2, 13, 18, 9, et 30 ; voir la Figure 2. Plusieurs des intervalles de confiance des effets aléatoires se chevauchent, donc ce classement contient une bonne dose d'incertitude qui le rend incertain. Dans chacune des cinq premières écoles, il y a au moins 35 élèves — cela se reflète dans les intervalles de prédictions plus courts, dus au fait que les erreurs-type des prédictions des effets aléatoires sont plus petites également.

(i) Quel est le classement des cinq meilleurs écoles ?

Solution
Évidemment, un classement devrait inclure le type d'école et sa dénomination, puisque l'effet aléatoire conditionel indique à quel point l'école sous-performe ou sur-performe par rapport aux autres de sa catégorie.
Le classement devrait dont inclure les effets fixes pour les variables spécifiques aux écoles ; on peut forcer les variables propres aux individus à prendre des valeurs identiques, en créant un individu lambda. De cette façon, la prédiction inclura à la fois l'effet aléatoire (via la l'ordonnée à l'origine prédite) et les effets fixes pour `type` et `denom`. Le seul bémol avec cette approche est que les écoles unisexe ne peuvent techniquement accueillir des élèves que du sexe correspondant au type. Le classement final des meilleures écoles est 9, 13, 10, 18 et 19.

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, et est tirée de

Cresswell, M.J. (1990). *Gender Effects in GCSE : Some Initial Analyses*, Associated Examining Board Research Report RAC/517.

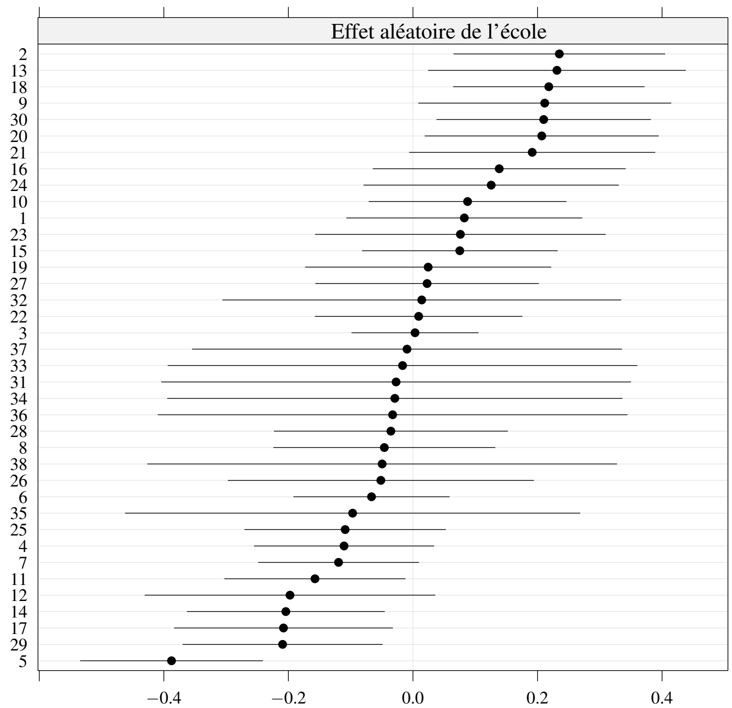

FIGURE 2 – Graphique des ordonnées à l'origine prédite par école avec intervalles de prédiction à 95% (Wald)

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,

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

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

(g) Est-ce qu'un modèle autorégressif d'ordre un, $\mathsf{AR}(1)$, serait adéquat pour modéliser la corrélation intra-groupe dans ce problème. Justifiez brièvement votre réponse.
Solution
Non, parce que le modèle de corrélation $\mathsf{AR}(1)$ 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 garçons (donc les filles ont en moyenne,*ceteris paribus*, un score 8,42 points inférieur à celui des garçons).