2022 ■ Vol. 18 ■ no. 2

Une introduction aux modèles de régressions multiniveaux avec R

Félix Duplessis-Marcotte [a], Raphaël Lapointe [a] & Pier-Olivier Caron [b]

[a]Université du Québec à Montréal
[b]Université TÉLUQ

Abstract ■ La crise de reproductibilité en psychologie est en partie causée par l'utilisation d'analyses statistiques inadaptées aux données récoltées. Les données ont souvent des caractéristiques importantes à considérer, comme lorsque celles-ci sont nichées dans différents groupes (p. ex. recruter plusieurs élèves dans différentes classes). Dans ce cas, cela fait en sorte que le postulat de normalité des modèles linéaires généraux n'est pas respecté. Ignorer ce postulat d'indépendance en utilisant un modèle linéaire général peut mener à des résultats erronés, comme des faux positifs, des biais ou une perte de puissance. Les analyses de régressions multiniveaux répondent à ce problème et assurent la validité des résultats obtenus. Cet article se veut un tutoriel couvrant les principes généraux sous-jacents aux régressions multiniveaux pour analyser des données nichées. Des données pseudoaléatoires sont générées avec R et analysées avec des régressions multiniveaux afin de démontrer la valeur ajoutée de considérer la hiérarchisation des données quant à la validité des résultats. De plus, cet article fournit, étape par étape, la syntaxe R pour faciliter l'utilisation des analyses multiniveaux et l'adaptation de celles-ci aux données des lecteurs.

Keywords ■ Régression, Régression multiniveau, R, modélisation. **Tools** ■ R.

✉ duplessis-marcotte.felix@courrier.uqam.ca

10.20982/tqmp.18.2.p168

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Ce texte a reçu le prix SQRP-TQMP remis lors de la conférence annuelle de la Société québécoise pour la recherche en psychologie *en 2022.*

Introduction

Analyser la relation entre une variable dépendante et une ou plusieurs variables indépendantes par l'entremise d'un modèle linéaire général, comme l'analyse de variance ou la régression, est la modélisation statistique la plus utilisée en psychologie (Blanca, Alarcón & Bono, 2018). L'un des postulats du modèle linéaire stipule que les individus sont échantillonnés indépendamment les uns des autres. Toutefois, plusieurs études y contreviennent en incorporant volontairement ou involontairement des groupes de participants. Par exemple, recruter des participants provenant de différents quartiers selon différents contextes socioéconomiques. Dans ce cas, les citoyens d'un quartier défavorisé risquent de se ressembler plus entre eux qu'avec d'autres personnes issues de milieux très favorisés. Pareillement, des élèves provenant de la même classe partagent le même enseignant et le même environnement de classe faisant en sorte qu'ils partagent une expérience d'enseignement plus similaire en comparaison à des élèves provenant d'autres classes. Et les classes provenant de différentes écoles sont quant à elles représentatives de leur milieu. Il adviendra inévitablement que les individus se retrouvant dans les mêmes groupes (classes, écoles ou quartiers) soient plus similaires en comparaison entre eux qu'aux individus d'autres groupes. Cette structure hiérarchique (c.-à-d. composée de plusieurs niveaux) ne respecte pas le postulat d'indépendance des participants. Quand celui-ci n'est pas respecté, l'analyse multiniveau est une avenue de choix afin de tenir compte de la variabilité à l'intérieur des groupes, mais aussi entre les groupes.

Il est important de noter que l'analyse multiniveau peut être utilisée lorsque les données sont nichées dans différents groupes (p. ex. des patients dans un hôpital), ou qu'elles proviennent d'un même individu (p. ex. plusieurs comportements d'un individu). Dans les deux cas, des niveaux hiérarchiques sont présents. Le présent article se concentre sur la hiérarchisation des données dans différents groupes, mais les principes généraux de l'analyse multiniveau sont les mêmes dans les deux contextes.

Objectifs

Les objectifs de la présente étude sont d'introduire la modélisation sous-jacente à l'analyse multiniveau et de fournir un tutoriel afin de faciliter l'application de cette analyse. Le logiciel privilégié est R, une plateforme d'analyse statistique libre d'accès (R Core Team, 2021). Le mouvement « open-science » monte en popularité et combat la crise de réplicabilité dont souffre le domaine de la psychologie plus particulièrement (van der Zee & Reich, 2018). Ainsi, cet article se veut un tutoriel rendant facilement accessible des analyses qui sont généralement considérées comme complexes.

L'article est divisé en deux grandes parties. D'abord, les principes théoriques et mathématiques sous-jacents à l'analyse multiniveau seront expliqués, en commençant par une révision de la régression linéaire simple. Par la suite, les analyses multiniveaux seront appliquées à un exemple concret. Pour se faire, des données pseudoaléatoires seront générées, pour ensuite les analyser à partir de la modélisation établie. Finalement, les caractéristiques à observer ainsi que l'interprétation de ce type d'analyse seront présentées.

La régression linéaire simple

Les modèles de régressions multiniveaux sont une extension du modèle de la régression linéaire simple (Hox, Moerbeek & Van de Schoot, 2017). Ainsi, il importe d'aborder brièvement la régression linéaire simple (pour une description extensive du modèle de régression linéaire simple, voir Fox & Weisberg, 2018; Montgomery, Peck & Vining, 2021). L'équation d'une régression linéaire simple se caractérise par la présence d'un seul prédicteur et d'une seule variable dépendante qui est représentée de la façon suivante :

$$y_i = b_0 + b_1 x_i + \epsilon_i.$$

Dans cette équation, y représente la variable prédite (variable dépendante), x est le prédicteur (variable indépendante), b_0 correspond à l'ordonnée à l'origine (la valeur de y lorsque $x = 0$), b_1 est le coefficient de régression associé à x (la pente), ϵ représente l'erreur résiduelle et i correspond à la i^{e} observation. La régression doit respecter quatre postulats :

— les résidus doivent être distribués normalement;
— la variance résiduelle est homoscédastique;
— la relation entre les variables doit être linéaire;
— les observations doivent être indépendantes;

Contrevenir à l'un d'eux peut mener à des résultats erronés si la bonne analyse statistique n'est pas utilisée. Il existe des techniques pour remédier à une violation de ces postulats. Le respect des trois premiers postulats peut être facilement identifié par une inspection graphiques. Le dernier postulat « les observations doivent être indépendantes » est davantage méthodologique, puisqu'il concerne l'échantillonnage des données (comment les participants sont recrutés). Y contrevenir peut toutefois entraîner des conséquences substantielles sur les trois autres postulats. L'une de ces conséquences est que les estimations des erreurs standards diminuent, ce qui peut produire des résultats faussement significatifs (Hox et collègues, 2017).

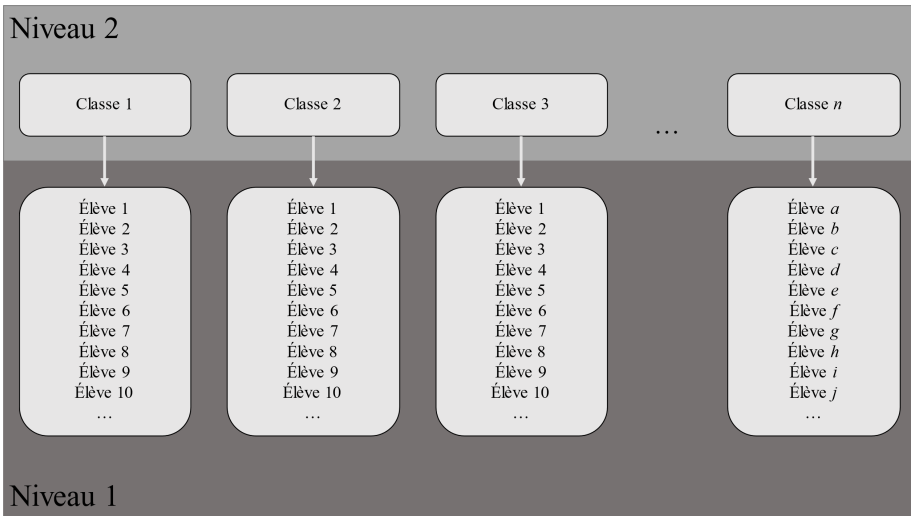
Lorsqu'il y a une hiérarchie dans les données, les variables contextuelles introduisent une dépendance entre les observations. Cela fait en sorte que le postulat d'indépendance selon lequel les unités d'un modèle linéaire doivent être indépendantes les unes des autres n'est pas respecté. Par exemple, les élèves qui fréquentent la même classe ne sont pas indépendants, car ils partagent plusieurs éléments en communs (ils sont dans le même local, avec les mêmes pairs et le même enseignant) ce qui peut influencer leurs comportements. Dans cet exemple, les classes sont une variable contextuelle, car les élèves y sont regroupés (nichés). Les observations auprès d'enfants d'une même classe seront plus similaires entre elles que les observations effectuées auprès d'enfants provenant d'autres classes. Cela entraîne une hiérarchie dans la structure des données où elles sont organisées en deux niveaux. La Figure 1 montre que les élèves (niveau 1) sont nichés dans des classes (niveau 2).

Une structure hiérarchique peut contenir plus de deux niveaux. Par exemple, les élèves (niveau 1) se retrouvent dans différentes classes (niveau 2) qui elles se trouvent dans différentes écoles (niveau 3). Dans ce cas, la hiérarchie ressemble à celle présentée dans la Figure 2. Heureusement, les modèles multiniveaux permettent de prendre en compte le non-respect du postulat d'indépendance afin d'obtenir des résultats valides lorsqu'il y a présence de hiérarchisation dans les données.

Régressions multiniveaux

Dans le modèle de régression linéaire simple, les paramètres b_0 et b_1 sont implicitement fixes : il est assumé que le modèle s'applique adéquatement à l'ensemble de l'échantillon. Cependant, lorsque des individus sont échantillonnés à partir du même contexte, il est probable que le postulat de l'indépendance ne soit pas respecté. Dans ces circonstances, il est inadéquat de considérer que les paramètres b_0 et b_1 sont fixes, car les ordonnées et les pentes pourraient varier selon les contextes (dû à la dépendance des observations dans un même contexte). C'est ici que les modèles multiniveaux interviennent : ils surmontent le problème lié à la dépendance des observations en permettant aux paramètres b_0 et/ou b_1 de varier selon les variables contextuelles. Lorsque les pa-

Figure 1 ■ Structure hiérarchique à deux niveaux. Les élèves (niveau 1) sont nichés dans les classes (niveau 2), qui représentent la variable contextuelle (inspiré de Field, Miles & Field, 2012).



ramètres varient, ils ne sont plus fixes, ils sont maintenant « aléatoires ». Autrement dit, plutôt que de fixer un ou des paramètres du modèle, ceux-ci peuvent être estimés différentiellement selon la variable contextuelle. Par exemple, chaque classe aura son ou ses estimateurs qui lui seront propres.

Afin de déterminer l'influence des variables contextuelles dans un échantillon, il est possible d'ajouter des coefficients aléatoires dans le modèle de régression. La Figure 3 montre les trois cas possibles : 1) inclure des ordonnées aléatoires tout en gardant des pentes fixes, 2) garder des ordonnées fixes en incluant des pentes aléatoires, et 3) inclure des ordonnées et des pentes aléatoires.

L'équation suivante représente le premier cas de figure, où les ordonnées sont aléatoires et les pentes sont fixes :

$$y_{ij} = b_{0j} + b_1 x_{ij} + \epsilon_{ij}$$
$$b_{0j} = b_0 + \mu_{0j}$$

Force est de constater que cette équation est similaire à celle du modèle de régression linéaire simple présenté plus tôt. Dans cette nouvelle équation, j représente la variable contextuelle (les classes), alors que μ_{0j} correspond à la variabilité dans l'ordonnée selon j. Sur la deuxième ligne de l'équation, b_0 (l'ordonnée) est associé à μ_{0j}, formant ainsi b_{0j}, indiquant que l'ordonnée varie selon la variable contextuelle. En d'autres mots, ce modèle tient compte de la variabilité dans l'ordonnée causée par le regroupement des données (p. ex. dans différentes classes).

Dans un second cas de figure, il est possible d'inclure des pentes aléatoires au lieu d'ordonnées aléatoires. Voici à quoi ressemblerait un modèle avec pentes aléatoires :

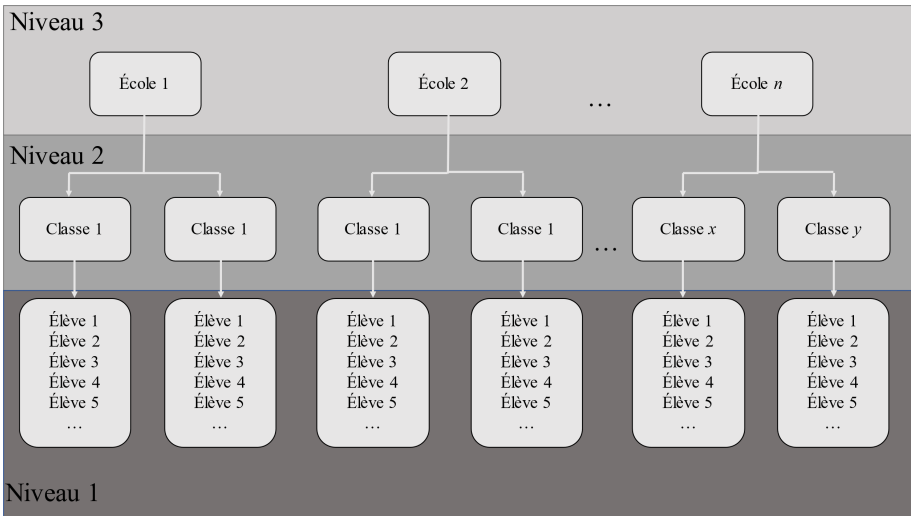
$$y_{ij} = b_0 + b_{1j} x_{ij} + \epsilon_{ij}$$
$$b_{1j} = b_1 + \mu_{1j}$$

Dans cette équation, j représente encore la variable contextuelle, soit les classes, alors que μ_{1j} correspond à la variabilité dans la pente selon j. Sur la deuxième ligne de l'équation, b_1 (la pente) est associé à μ_{1j}, formant ainsi b_{1j}, indiquant que la pente varie selon la variable contextuelle. Autrement dit, ce modèle tient compte de la variabilité dans la pente qui est causée par la hiérarchisation des données.

Lorsque des pentes aléatoires sont incluses dans un modèle, il est assumé que les ordonnées le seront aussi, car la variabilité dans les pentes devrait normalement créer de la variabilité dans les ordonnées (Gelman & Hill, 2007). Au lieu de seulement inclure des pentes aléatoires dans un modèle, il est généralement recommandé d'inclure à la fois des pentes et des ordonnées aléatoires dans la même étape. Cela mène à la troisième équation où les ordonnées et les pentes du modèle sont aléatoires. Dans ce dernier modèle, il s'agit de combiner les éléments des deux modèles précédents, c'est-à-dire, l'ordonnée (b_{0j}) et la pente (b_{1j}) aléatoires :

$$y_{i,j} = b_{0j} + b_{1j} x_{ij} + \epsilon_{ij}$$
$$b_{0j} = b_0 + \mu_{0j}$$
$$b_{1j} = b_1 + \mu_{1j}$$

Figure 2 ■ Structure hiérarchique à trois niveaux. Les élèves (niveau 1) sont nichés dans les classes (niveau 2), qui sont elles aussi nichées dans les écoles (niveau 3), une seconde variable contextuelle (inspiré de Field, Miles & Field, 2012).



Bien que les ordonnées et les pentes peuvent prendre des valeurs en fonction de la variable contextuelle, celles-ci doivent respecter une distribution normale multivariée ayant comme paramètres leurs moyennes μ_0 et μ_1 et la matrice de covariance $\mathbf{\Sigma}$ qui exprime leur variance et leur covariance, dont on peut donner la forme suivante :

$$\begin{pmatrix} b_{0j} \\ b_{1j} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix}, \mathbf{\Sigma} \right).$$

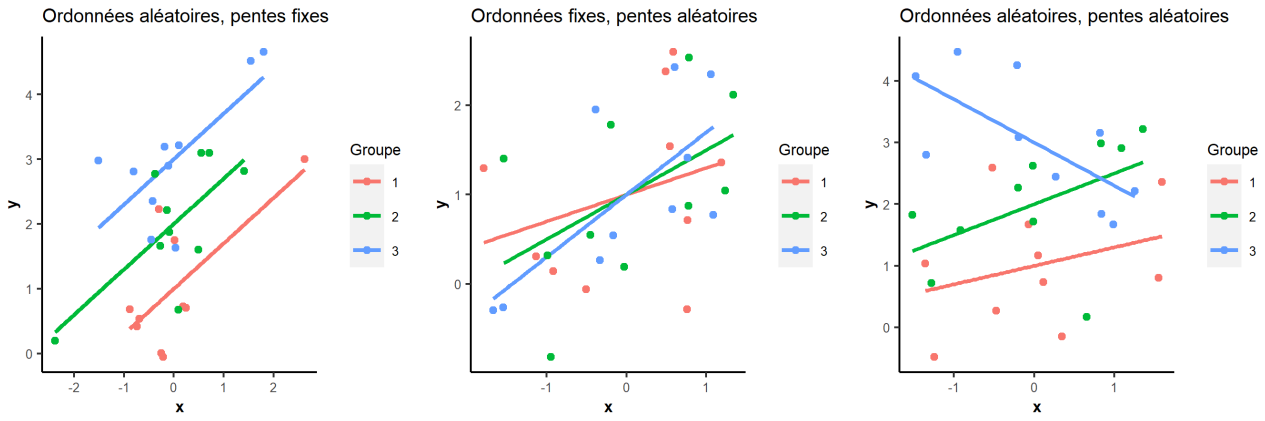
Dans une régression linéaire simple, μ_0 et μ_1 sont équivalents aux paramètres d'ordonnée et de pente de la population, b_0 et b_1, et sont fixes. En spécifiant une matrice de covariance, l'analyse multiniveau permet à ces paramètres de varier en fonction de la matrice de covariance et ainsi d'attribuer des valeurs différentes des paramètres b_0 et b_1 selon leur niche j.

Quand recourir aux analyses multiniveaux ?

Avant de procéder aux analyses multiniveaux et à l'interprétation de celles-ci, il est recommandé de tester les modèles ci-dessus afin de les comparer entre eux pour déterminer lequel s'applique le mieux aux données (Raudenbush & Bryk, 2002 ; Twisk, 2006). En effet, selon le principe de parcimonie (Bates, Kliegl, Vasishth & Baayen, 2018), il importe d'utiliser le modèle qui représente le mieux les données avec le moins de paramètres libres possible. Un modèle plus simple (contenant moins de paramètres libres) facilite l'interprétation des résultats. Il y a quatre indices qui sont souvent utilisés pour comparer les modèles. La fonction de log-vraisemblance (*log-likelihood* ; LL), qui sert de base pour déterminer les trois autres indices, indique à quel point il reste de l'information qui n'est pas expliquée par le modèle après l'avoir appliqué sur les données. À partir de la fonction de vraisemblance est obtenu un indice de déviance (-2LL), qui correspond à une valeur de moins deux fois la fonction de log-vraisemblance (-2 × LL). Notons aussi que l'indice de déviance suit une distribution du χ^2, ce qui permet d'obtenir une valeur p, pour ensuite faciliter la comparaison entre les modèles. Il est également possible d'obtenir le critère d'information d'Akaike (AIC), qui permet d'estimer la qualité de l'ajustement du modèle en considérant le nombre de paramètres estimés. Finalement, il est possible d'obtenir le critère bayésien de Schwarz (BIC), qui est similaire à l'AIC, mais en étant plus conservateur, car il corrige plus sévèrement pour le nombre de paramètres estimés dans le modèle. Une fois que ces indices seront obtenus, les indices d'ajustement des différents modèles seront comparés entre eux. L'interprétation des comparaisons entre les modèles sera abordée plus en détail dans les dernières sections.

Il est également recommandé de vérifier la dépendance entre les sujets avant d'entreprendre des régressions multiniveaux. La dépendance entre les résultats des participants dans un échantillon peut être estimée grâce à certains indices statistiques, comme la corrélation intraclasse (ICC). Dans l'exemple des élèves nichés dans les classes, la corrélation intraclasse représente la proportion de la variabilité totale dans la variable dépendante (p. ex. le rendement des élèves) qui est attribuable à la variable contex-

Figure 3 ■ Comparaison entre trois modèles avec différents effets aléatoires. Un modèle de régressions multiniveaux avec des ordonnées aléatoires et la pente fixe (panneau de gauche). Un modèle de régressions multiniveaux avec l'ordonnée à l'origine fixe et les pentes aléatoires (panneau du centre). Un modèle de régressions multiniveaux avec des ordonnées et des pentes aléatoires (panneau de droite).



tuelle, c'est-à-dire, aux classes (Hox et collègues, 2017). La corrélation intraclasse est représentée par l'équation suivante :

$$ICC = \frac{\sigma^2_{u0}}{\sigma^2_{u0} + \sigma^2_{\epsilon}}$$

Dans cette équation, σ^2_{u0} représente la variance des résiduels de niveau 2 (les classes) et σ^2_{ϵ} représente la variance des résiduels de niveau 1 (les élèves). La corrélation intraclasse obtenue peut ensuite être utilisée afin de calculer le *design effect* ($def\,f$) d'un paramètre (Muthén & Satorra, 1995). Le *design effect* permet d'obtenir un indicateur qui illustre à quel point la nature d'un échantillonnage qui comprend des regroupements (*clusters*) peut biaiser l'estimation d'un paramètre (Kish, 1965). Dans le cadre d'un modèle multiniveau, le *design effect*, d, se calcule ainsi :

$$d = 1 + (c - 1)\,\mathrm{ICC}$$

Dans cette équation, c représente le nombre moyen de personnes se retrouvant dans les groupes (classes), alors que l'ICC correspond à la corrélation intraclasse d'une variable (Muthén & Satorra, 1995). La valeur obtenue peut ensuite être comparée à une norme fréquemment utilisée dans le domaine de la recherche en éducation (Lai & Kwok, 2015), mais aussi dans le domaine de la psychologie (p. ex. Weisz, Bearman, Santucci & Jensen-Doss, 2017) : lorsque le *design effect* est supérieur à 2, il est nécessaire d'utiliser des analyses multiniveaux, car les risques d'obtenir des résultats biaisés sont trop élevés. Cette norme n'est toutefois pas recommandée si 1) le nombre moyen de personnes se trouvant dans les groupes est inférieur à 10, et 2) si l'intérêt porte sur l'effet de prédicteurs se retrouvant à un niveau supérieur dans la hiérarchie des données (p. ex. s'intéresser à l'effet des pratiques des enseignants [niveau 2] sur le rendement des élèves [niveau 1]; Lai & Kwok, 2015).

Deux autres caractéristiques importantes à considérer sont le nombre d'observations par groupe (*clusters*) et le nombre de groupes (Cousineau & Laurencelle, 2015). Un trop faible nombre de l'un ou de l'autre peut nuire à la validité des analyses multiniveaux, plus précisément, à l'estimation de l'effet du niveau concerné (Gelman & Hill, 2007). Une analyse multiniveau devrait faire aussi bien qu'une analyse de régression multiple au minimum. Il existe plusieurs études statistiques sur cette question (notamment, Cousineau & Laurencelle, 2015; Moeyaert, Rindskopf, Onghena & Van den Noortgate, 2017; McNeish & Stapleton, 2016). Le contexte de recrutement des participants devrait être parmi les premiers éléments à considérer.

Exemple hypothétique : Professeur X

Afin de concrétiser les principes et l'utilité des analyses de régressions multiniveaux, prenons l'exemple de Professeur X, un jeune chercheur qui s'intéresse à la réussite scolaire des adolescents québécois. Dernièrement, il s'est impliqué dans l'élaboration d'un devis où il recrutera 250 élèves dans 50 classes de 4e secondaire en mathématiques dans une école de la région. Il vise à recruter un total de 12500 élèves. Quoique cette taille d'échantillon soit peu réaliste, elle permet d'obtenir des résultats qui seront plus proches des paramètres définis, ce qui permettra de confir-

mer le succès des analyses effectuées. Dans le devis fictif de Professeur X, les élèves rempliront un questionnaire mesurant leur anxiété de performance, et les résultats scolaires en mathématiques de ces mêmes élèves seront obtenus par la suite. Il poursuit l'élaboration de son devis en réfléchissant à un plan d'analyse : il désire prédire le rendement en mathématiques (variable dépendante) en fonction de l'anxiété de performance (variable indépendante). Considérant le devis et le plan d'analyse de Professeur X, il est possible de le diriger vers l'utilisation des régressions multiniveaux.

Génération de données pseudoaléatoires avec R

Dans le but de présenter la pertinence des analyses multiniveaux à Professeur X, des données pseudoaléatoires seront générées. Contrairement à l'utilisation d'une banque de données publique et naturellement récoltée, générer des données pseudoaléatoires permet de vérifier la performance du modèle créé. En effet, lors de la génération de données, les coefficients des paramètres seront spécifiés, ce qui permettra de vérifier si le modèle réussira à retrouver ces mêmes coefficients. La création de l'échantillon est la première étape de la génération de données pseudoaléatoires. La fonction `set.seed()` permet de pouvoir répliquer les présents résultats. Le nombre de groupes (`ngroupes`), le nombre d'élèves par groupe (`eleves`), le nombre total d'élèves (`n`) ainsi que l'appartenance des élèves à un groupe (`groupes`) sont spécifiés dans la syntaxe suivante.

```
set.seed(365)
ngroupes <- 50
eleves <- 250
n <- ngroupes * eleves
groupes <- rep(1:ngroupes, each=eleves)
```

La deuxième étape est de créer les variables. Dans le cas suivant, celles-ci seront standardisées avec une moyenne de 0 et un écart-type de 1. Pour une description approfondie permettant de créer des modèles standardisés, voir Caron et Lemardelet (2021). Pour ce faire, il faut créer les paramètres fixes de la régression, soit l'ordonnée à l'origine (μ_0; `mu0`) et la pente (μ_1; `mu1`). Des valeurs de 0,2 et -0,5 leur sont attribuées respectivement. Pour que les données reflètent bien la réalité du scénario, la valeur de la pente (μ_1) est négative, car la littérature scientifique (p. ex. von der Embse, Jester, Roy & Post, 2018) souligne que l'anxiété de performance est négativement associée avec le rendement scolaire. La pente et l'ordonnée à l'origine sont ensuite mises dans un vecteur (`mu`) afin de faciliter les prochaines étapes.

```
mu0 = .2
mu1 = -.5
mu = c(mu0, mu1)
```

Il faut ensuite déterminer la matrice de covariance des paramètres aléatoires (`Sigma`). La matrice de covariance indique la variance des variables (les éléments dans la diagonale) et la covariance entre les variables (les éléments hors diagonale). La matrice de covariance détermine à quel point les données varient autour de la moyenne de la population. Puisque les données sont générées pseudoaléatoirement, la matrice de covariance a été arbitrairement choisie, mais elle pourrait être différente de ce présent exemple.

```
Sigma = matrix(c(.2, .1,
                 .1, .4),
               ncol = 2, nrow = 2)
```

Il est maintenant possible de créer les paramètres aléatoires pour les groupes (`parametres`). Comme deux paramètres sont créés pour 50 groupes, ceux-ci seront générés à partir d'une distribution normale multivariée aléatoire (`mvrnorm()`) avec `R` en utilisant le package `MASS` (Venables & Ripley, 2002). Notons que `mu` (μ) correspond au vecteur `mu` créé plus tôt, et que `Sigma` ($\mathbf{\Sigma}$) correspond à la matrice de covariance établie lors de la dernière étape (`Sigma`). La fonction `colnames()` permet de renommer adéquatement les colonnes de la nouvelle matrice de données créée. Ces lignes permettent de générer les ordonnées et les pentes des groupes :

```
parametres <- MASS::mvrnorm(
                n = ngroupes,
                mu = mu,
                Sigma = Sigma)
colnames(parametres) <- c("beta0",
                          "beta1")
```

Par la suite, la variable indépendante (l'anxiété de performance; `AP`) est créée à partir d'une distribution normale centrée réduite en utilisant la fonction `rnorm()` de `R`.

```
AP <- rnorm(n)
```

Il faut ensuite créer la variable dépendante (le rendement en mathématiques). Pour maintenir un scénario standardisé, la variance résiduelle de la variable dépendante (`var_emath`) doit d'abord être obtenue. Pour ce faire, il faut soustraire la somme des variances des paramètres (`sum(diag(Sigma))`) et l'effet au carré de la variable indépendante (`beta1^2`) à la variance standardisée du rendement en mathématiques (voir Caron & Lemardelet, 2021, pour une explication approfondie). La variance résiduelle obtenue est utilisée par la suite afin de générer l'erreur résiduelle de y (`e`). L'erreur résiduelle de y doit suivre une distribution normale, d'où l'utilisa-

tion de `rnorm()`, et son écart-type (`sd`) équivaut à la racine carrée de la variance de l'erreur du rendement en mathématiques (`sqrt(var_emath)`). Finalement, le rendement en mathématiques (`math`) peut être généré en assignant les différentes ordonnées à l'origine et les pentes d'anxiété de performance aux groupes, en multipliant les scores d'anxiété de performance à sa pente, et en ajoutant l'erreur résiduelle (`e`).

```
var_emath <- 1-(sum(diag(Sigma)) +
                beta1^2)
e  <- rnorm(n = n, sd = sqrt(var_emath))
math <- parametres[groupes, "beta0"] +
     parametres[groupes, "beta1"] * AP +
     e
```

La troisième étape de la génération de données est d'utiliser la fonction `data.frame()` afin de grouper l'ensemble des variables créées dans un jeu de données en y incluant la variable dépendante (`math`), la variable indépendante (`AP`) et une variable catégorielle spécifiant le groupe (`g`) en tant que facteur (`as.factor()`).

```
jd <- data.frame(y = math,
                 x = AP,
                 g = as.factor(groupes))
```

Création des modèles

Une fois les données générées, des modèles peuvent être créés et testés. Reprenons tout d'abord le modèle de régression linéaire simple. Dans le cas de Professeur X, l'équation suivante sera attribuée au premier modèle (modèle 1) qui sera testé dans les étapes ultérieures du papier :

$$\begin{aligned}\text{Rendement en mathématiques}_i &= b_0 \\ &+ b_1 \text{Anxiété de performance}_i \\ &+ \epsilon_i\end{aligned}$$

Quoique la régression linéaire simple réponde à la question de recherche de Professeur X, il est possible d'améliorer le modèle. Comme mentionné précédemment, le regroupement des élèves dans les différentes classes enfreint le postulat d'indépendance. En effet, le coefficient fixe b_0 signifie que le modèle prédit les données en assumant une seule moyenne pour toutes les classes. Cela dit, il est fort probable que les classes varient dans leur rendement en mathématiques. Il est possible d'ajouter un effet aléatoire (μ_{0j}) pour permettre à l'intercepte des classes (la moyenne du rendement en mathématique de chaque classe lorsque l'anxiété de performance est tenue à la moyenne) de varier. L'équation suivante sera attribuée au deuxième modèle (modèle 2) qui sera testé dans les étapes ultérieures du present article :

$$\begin{aligned}\text{Rendement en mathématiques}_{ij} &= b_{0j} \\ &+ b_1 \text{Anxiété de performance}_{ij} \\ &+ \epsilon_{ij} \\ b_{0j} &= b_0 + \mu_{0j}\end{aligned}$$

Finalement, il est possible que l'anxiété de performance n'influence pas le rendement en mathématiques de la même façon dans toutes les classes (p. ex. à cause d'une meilleure préparation de la part d'un professeur). Par conséquent, le dernier modèle permet à l'ordonnée et à la pente de varier librement selon la variable contextuelle (les classes). Ce troisième et dernier modèle (modèle 3) sera testé dans les étapes ultérieures et s'exprime avec l'équation suivante :

$$\begin{aligned}\text{Rendement en mathématiques}_{ij} &= b_{0j} \\ &+ b_{1j} \text{Anxiété de performance}_{ij} \\ &+ \epsilon_{ij} \\ b_{0j} &= b_0 + \mu_{0j} \\ b_{1j} &= b_1 + \mu_{1j}\end{aligned}$$

Analyse du modèle

Afin d'analyser ces données avec R, les packages nécessaires pour l'analyse des trois modèles doivent être importés. Le package `lme4` (Bates, Mächler, Bolker & Walker, 2015) permet d'utiliser et d'analyser des modèles linéaires mixtes (dont les régressions multiniveaux) et `lmerTest` (Kuznetsova, Brockhoff & Christensen, 2017) permet d'obtenir les valeurs *p* pour les effets fixes des régressions multiniveaux.

```
library(lme4)
library(lmerTest)
```

Modèle 0, la corrélation intraclasse et le design effect

Avant d'entreprendre l'aventure des analyses multiniveaux, il est d'abord pertinent pour Professeur X de commencer par déterminer s'il y a de la variance à expliquer dans son échantillon et si cette variance se situe dans la variable de regroupement (les classes). Pour se faire, il faut créer un modèle inconditionnel (modèle 0) afin d'obtenir les paramètres nécessaires pour calculer la corrélation intraclasse et le *design effect*. Ce modèle ne contient aucun prédicteur et permet à l'ordonnée de varier (effet aléatoire).

$$\begin{aligned}\text{Rendement en mathématiques}_{ij} &= b_{0j} + \epsilon_{ij} \\ b_{0j} &= b_0 + u_{0j}\end{aligned}$$

Pour créer le modèle 0 dans R, l'utilisation d'une régression multiniveau est appelée avec la fonction

`lmer()`. Minimalement, la fonction nécessite une équation (« formula ») et un jeu de données. Dans l'équation, le symbole ∼, signifiant « prédit par », départage à gauche, la variable dépendante et à droite, toutes les autres variables (effets fixes et aléatoires). Le `1` signifie l'ordonnée, alors que `(1|g)` signifie que le modèle laisse l'ordonnée à l'origine varier en fonction des groupes. Le symbole | indique les effets aléatoires. La fonction `summary()` permet d'obtenir le sommaire des résultats de l'analyse.

```
modele0 <- lmer(y ~ 1 + (1 | g),
                data = jd)
summary(modele0)
```

Deux éléments de la sortie du modèle 0 sont importants pour obtenir l'ICC, soit la variance des résiduels de niveau 2 (les classes; $\sigma^2_{u0} = 0,8551$) et la variance des résiduels de niveau 1 (les élèves; $\sigma^2_\epsilon = 0.1348$). Maintenant que ces indices sont connus, il est possible d'obtenir le coefficient de corrélation intraclasse (ICC) en plaçant les valeurs dans l'équation suivante (notons que le package `performance` Lüdecke, Ben-Shachar, Patil, Waggoner & Makowski, 2021, permet également de calculer $l'ICC$) :

$$ICC = \frac{\sigma^2_{u0}}{\sigma^2_{u0} + \sigma^2_\epsilon} = \frac{0,1348}{0,1348 + 0,8551} = 0,136$$

```
icc <- performance::icc(modele0)
[1] Adjusted ICC: 0.136
icc <- icc$ICC_adjusted
```

La sortie précédente indique que la valeur de $l'ICC$ est bien de 0,136. Cela signifie qu'environ 14% de la variabilité dans la variable dépendante (le rendement en mathématiques) est attribuable à la variable contextuelle, autrement dit, aux classes. Maintenant que $l'ICC$ est connu, il est possible de calculer le *design effect*. Rappelons que dans l'équation suivante, c correspond au nombre moyen de personnes se retrouvant dans les groupes. Dans le présent exemple, c = 250.

$$d = 1 + (c - 1) * ICC$$

```
deff <- 1 + (eleves-1)*icc
[1] 34.864
```

La valeur du *design effect* est supérieure à 2, ce qui justifie l'utilisation des régressions multiniveaux (Lai & Kwok, 2015). Même si cette valeur indique que l'utilisation d'une régression linéaire simple est inadéquate, le modèle 1 sera tout de même testé afin de le comparer aux autres modèles et de voir la valeur ajoutée de l'inclusion des effets aléatoires.

Modèle 1 : Régression simple

Comme mentionné précédemment, le modèle le plus simple est la régression linéaire simple, représentée par la fonction `lm()`, où les estimateurs de l'ordonnée et de la pente sont fixes. Rappelons que y correspond au rendement en mathématiques alors que x représente l'anxiété de performance.

```
modele1 <- lm(y ~ x, data = jd)
```

Modèle 2 : Ordonnée aléatoire et pente fixe

Le deuxième modèle maintient le paramètre de la pente fixe, mais permet à l'ordonnée de varier (effet aléatoire) selon la variable contextuelle (les classes). C'est ce qui est présenté dans la syntaxe ci-dessous, où le terme `(1|g)` signifie que l'ordonnée (`1`) varie selon les classes (`g`). Ces termes sont inclus dans la fonction de régressions multiniveaux `lmer()`.

```
modele2 <- lmer(y ~ x + (1 | g),
                data = jd)
```

Modèle 3 : Ordonnée et pente aléatoires

Dans le troisième modèle, les paramètres de la pente et de l'ordonnée sont libres de varier (effets aléatoires) selon l'appartenance aux classes. C'est ce qui est représenté dans la syntaxe ci-dessous, où la pente est maintenant incluse dans le terme (`1 + x|g`), signifiant que l'ordonnée (`1`) et la pente de l'anxiété de performance (`x`) peuvent varier selon les groupes (`g`).

```
modele3 <- lmer(y ~ x + (1 + x | g),
                data = jd)
```

Comparaison des modèles 1, 2 et 3

Maintenant que les trois modèles ont été obtenus, ceux-ci seront comparés grâce à la fonction `anova()`. Celui s'ajustant le mieux aux données sera conservé, et les résultats de ce modèle seront interprétés par la suite. Le modèle 1 sera comparé au modèle 2, et le modèle 2 sera comparé au modèle 3.

```
anova(modele3, modele2, modele1)
```

Le Tableau 1 présente les différents indices de comparaison utilisés par `R` (AIC, BIC, LL et -2LL) lors de la commande `anova()`. On y retrouve également le nombre de paramètres par modèle (npar), la valeur du khi carré (χ^2), le nombre de degrés de liberté (dl) ainsi que la valeur *p*. Lorsque deux modèles sont comparés, celui ayant des indices de comparaison (AIC et BIC) plus faibles doit être considéré comme étant le modèle qui s'applique le mieux aux données. Pour savoir si la différence d'indice de

Tableau 1 ■ Indices de comparaison des modèles

Modèle	npar	AIC	BIC	LL	-2LL	χ^2	dl
1	3	30796	30819	-15395,1	30790		
2	4	28258	28288	-14125,2	28250	2539,9***	1
3	6	12158	12203	-6073,2	12136	16104***	2

Note. Comparaison des trois modèles en utilisant la fonction `anova()`. Les indices de comparaison et le nombre de paramètres de chaque modèle (`npar`) sont comparés par un test de khi carré. *** $p < 0{,}001$

comparaison est significative, il est nécessaire de se tourner vers la valeur *p*. Dans le présent exemple, le modèle 2 s'applique mieux aux données que le modèle 1, car 1) ses indices sont plus faibles que ceux du modèle 1, et 2) la différence entre les indices de ces deux modèles est significative, $p < 0,001$. En ce qui concerne la comparaison entre le modèle 2 et 3, le modèle 3 est significativement meilleur que le modèle 2. En somme, parmi l'ensemble des modèles, le modèle 3 est celui qui s'applique le mieux aux données. Les postulats de l'analyse multiniveau seront vérifiés à partir de ce modèle.

Vérification des postulats : Modèle 3

Un principal postulat concerne les analyses multiniveaux : les ordonnées et les pentes aléatoires doivent être distribuées normalement autour des paramètres du modèle global. Dans le présent exemple, ce postulat est respecté considérant que les paramètres aléatoires ont été générés par la fonction `mvrnorm()` en y incluant les coefficients des effets fixes et la matrice de covariance des effets aléatoires. Les autres postulats des analyses multiniveaux sont les mêmes que la régression linéaire. Parmi ceux-ci se retrouve notamment la normalité des résiduels, qui est un postulat de la méthode d'estimation des moindres carrés : pour obtenir une estimation optimale des paramètres, les résiduels doivent suivre une distribution normale (Gelman & Hill, 2007). Dans le cadre de l'article, il est intéressant de comparer la distribution des résiduels entre le modèle 1 (régression linéaire simple qui ne prend pas en compte la hiérarchisation des données) et le modèle 3 (régression multiniveau). Pour vérifier le respect de ce postulat, il suffit d'extraire les résiduels d'un modèle désiré à l'aide de la fonction `resid()` et les attribuer à une nouvelle variable (`residuels`) dans la base de données.

```
jd$residuelsLM = resid(modele1)
jd$residuels = resid(modele3)
```

Pour réaliser les figures, le package `ggplot2` (Wickham, 2016) permet de créer rapidement et simplement plusieurs types de graphiques. Dans le cas présent, il est utile de générer un histogramme afin d'observer la distribution des résiduels. La syntaxe détaillée pour produire l'histogramme selon les normes APA est présentée dans le Listing 1.

L'analyse visuelle de la Figure 4 (générée par la syntaxe du Listing 1) confirme que le postulat de la normalité des résiduels est bien respecté pour le modèle 1 et le modèle 3, ce qui n'est pas surprenant considérant que les résiduels ont été générés grâce à la fonction `rnorm()`. Il est à noter que les erreurs résiduelles du modèle 1 (Figure 4a) sont plus variables que celles du modèle 3 (Figure 4b). Cela signifie que le modèle 3 est supérieur au modèle 1 puisque les effets aléatoires expliquent une plus grande partie de la variance.

Un autre postulat qu'il importe de vérifier est celui portant sur l'homoscédasticité des résiduels, qui affirme que les résiduels doivent avoir une variance similaire à chaque niveau du prédicteur (Gelman & Hill, 2007). Si ce postulat n'est pas respecté, l'estimation des paramètres effectuée par la méthode des moindres carrées n'est pas optimale. Dans le cadre de l'article, il est intéressant de comparer l'homoscédasticité des résiduels du modèle 1 et du modèle 3. Afin de vérifier si ce postulat est respecté, il suffit d'extraire les valeurs prédites d'un modèle désiré à l'aide de la fonction `predict()` et les attribuer à une nouvelle variable (`predits`) dans la base de données.

```
jd$preditsLM = predict(modele1)
jd$predits = predict(modele3)
```

Encore avec l'aide de `ggplot2` il est possible de générer un diagramme de dispersion pour observer la relation entre les résiduels et les valeurs prédites. La syntaxe pour produire les histogrammes de la Figure 5 selon les normes APA est présentée dans le Listing 1. L'observation de la Figure 5a illustre que le postulat d'homoscédasticité n'est pas respecté pour le modèle 1, car ils ont la forme d'un nœud papillon. En moyenne, les résidus ont tendance à augmenter au fur et à mesure que les valeurs prédites augmentent (en termes absolus). Cela est un signe d'hétéroscédasticité. La figure 5b démontre que le postulat est bien respecté pour le modèle 3.

Interprétation

Une fois les postulats des analyses multiniveaux vérifiés, il est maintenant temps d'aborder les ca-

Figure 4 ■ Histogramme des résiduels du modèle 1 et du modèle 3. Les histogrammes des résiduels montrent que le postulat de normalité des résiduels est respecté pour le modèle 1 (a) et pour le modèle 3 (b). L'histogramme du modèle 3 montre que l'inclusion des effets aléatoires explique plus de variance dans les résiduels que le modèle 1.

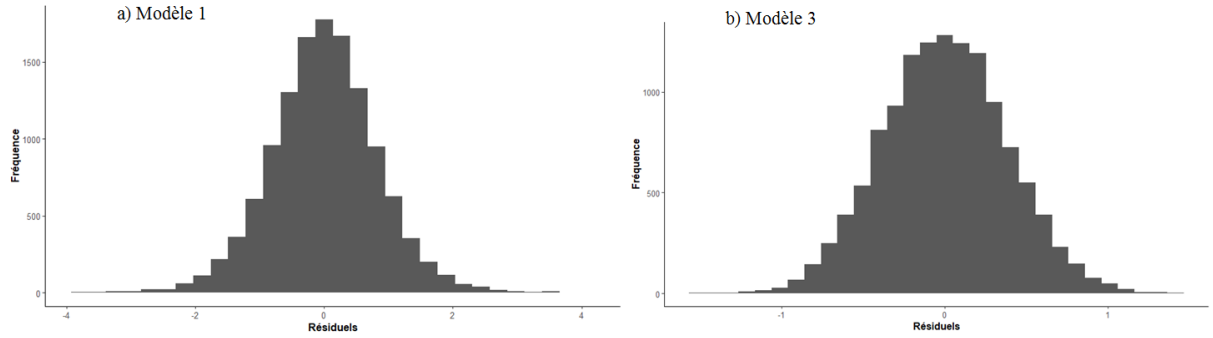
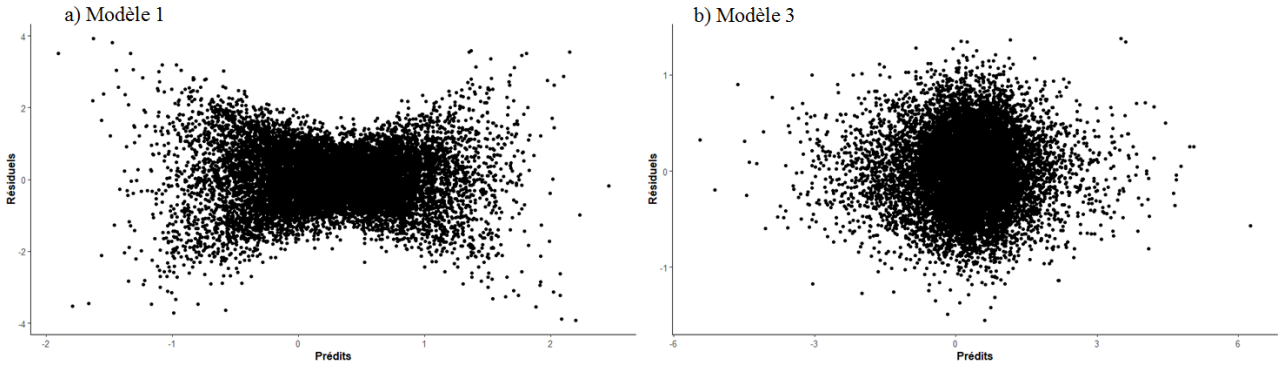


Figure 5 ■ Diagramme de dispersion des valeurs prédites par résiduels pour le modèle 1 et le modèle 3. Les diagrammes montrent que le postulat de l'homoscédasticité des résiduels n'est pas respecté pour le modèle 1 (a), alors qu'il est respecté pour le modèle 3 (b).



ractéristiques à observer dans le modèle 3 ainsi que leurs interprétations. Les effets aléatoires seront d'abord rapportés, suivis des effets fixes.

En premier lieu, il est important d'interpréter les effets aléatoires du modèle. Le Tableau 2 illustre que les ordonnées à l'origine des classes varient significativement, $ET = 0,38,\ \chi^2(1) = 2539,9,\ p < 0,001$ (le χ^2 et la valeur *p* proviennent du Tableau 1). Ce résultat suggère qu'ajouter un effet aléatoire pour l'ordonnée à l'origine est justifié puisque les classes ont un effet significatif sur le rendement en mathématiques des élèves avant même de considérer l'anxiété de performance. Les pentes de la variable d'anxiété de performance varient elles aussi significativement entre les classes, $ET = 0,63,\ \chi^2(2) = 16104,0,\ p < 0,001$. Cela suggère que l'anxiété de performance affecte les élèves différemment dépendamment des classes. Ainsi, le modèle avec les ordonnées et les pentes aléatoires permet de contrôler pour ces variabilités qui ne respectent pas le postulat d'indépendance entre les sujets.

L'effet fixe de l'anxiété de performance doit également être rapporté tel qu'illustré dans le Tableau 3. Dans le modèle, l'anxiété de performance prédit négativement et significativement le rendement en mathématiques, $b = -0,56,\quad t(49) = -6,24,\quad p < 0,001$. De manière générale, cela suggère que plus l'anxiété de performance est élevée, moins le rendement en mathématiques est bon. Il est également possible de confirmer que l'analyse a bien été construite puisque le coefficient de l'ordonnée à l'origine du modèle 3 ($\overline{b_0} = 0,26$) ainsi que celui de la pente ($\overline{b_{11}} = -0,56$), sont très proches des coefficients établis pour la génération des données pseudoaléatoires ($\mu_0 = 0,2$; $\mu_1 = -0,5$).

Tableau 2 ■ Effets aléatoires du modèle 3

	Paramètres	Variance	Écart-type
Groupes	Ordonnée	0,1430	0,3781
	Pente	0,3984	0,6312
Résiduels		0,1475	0,3841

Note. La variance pour les groupes (classes) est montrée. La variance dans l'ordonnée montre que les classes varient dans leurs niveaux de rendements en mathématiques. La variance dans la pente montre que les classes varient dans la mesure que l'anxiété de performance les affectes. Il est aussi intéressant de noter que la variance de l'ordonnée et de la pente sont proches des valeurs qui ont été établies pour la matrice de covariance.

Tableau 3 ■ Effets fixes du modèle 3

	Coefficients	Err. std.	t
Ordonnée	0,26195	0,05359	4,888***
Anxiété de performance	-0,55753	0,08934	-6,241***

Note. Les coefficients, les erreurs standards (Err. Std.) et le test t des effets fixes du modèle 3 sont présentés. L'anxiété de performance est significativement et négativement associée avec le rendement en mathématiques. Les coefficients de régression sont très similaires à ceux préalablement définis lors de la génération de données. *** $p < 0{,}001$

Le Tableau 3 présente également les erreurs standards du modèle 3. Il est intéressant de les comparer à celles du modèle 1 (régression linéaire simple) qui sont présentées dans le Tableau 4. Force est de constater que les erreurs standards du modèle 1 sont inférieures à celle du modèle 3. Comme attendu, le non-respect du postulat d'indépendance cause une sous-estimation des erreurs standards de l'ordonnée à l'origine et de l'anxiété de performance pour le modèle 1. Cela cause une surestimation de la valeur associée au test statistique, ce qui a pour effet d'augmenter les chances d'obtenir une erreur de type I (faux positif).

Finalement, il est possible d'obtenir certains coefficients de détermination (R^2) du modèle 3 grâce au package `MuMin` (Barton, 2020). Ce package est nécessaire, car les coefficients de déterminations des modèles multiniveaux sont différents de ceux obtenus en régression linéaire multiple (Rights & Sterba, 2018). Les R^2 souffrent de plusieurs lacunes : (a) les relations analytiques entre les mesures existantes n'ont pas été établies; (b) un partitionnement complet de la variance n'est pas utilisé pour créer des mesures; (c) une approche unifiée pour interpréter et choisir parmi les mesures ne fait pas consensus et (d) les logiciels statistiques n'intègrent pas toutes les mesures disponibles de manière cohérente.

```
library(MuMIn)
r.squaredGLMM(modele3)
```

En utilisant la fonction `r.squaredGLMM()`, deux coefficients sont obtenus : le coefficient de détermination marginal ($R^2_m = 0,31$) et conditionnel ($R^2_c = 0,85$). R^2_m représente la variance expliquée par les effets fixes du modèle, alors que R^2_c représente la variance expliquée par la totalité du modèle, incluant les effets fixes et aléatoires. On peut conclure que les effets fixes du modèle 3 expliquent à eux seuls environ 31% de la variance. Lorsqu'on combine l'apport des effets fixes et aléatoires du modèle 3, la variance expliquée atteint maintenant 85%.

À la lumière de cette modélisation pour la proposition de recherche de Professeur X, il est sans équivoque que les régressions multiniveaux soient un choix judicieux pour analyser les données que Professeur X récoltera. En effet, les résultats révèlent qu'il est important de prendre en considération le regroupement des données dans les différentes classes afin de modéliser plus adéquatement la relation entre l'anxiété de performance et le rendement en mathématiques.

Conclusion

Les analyses de régressions multiniveaux continuent de gagner en popularité depuis les années 1970 (Goldstein, 1995; Raudenbush & Bryk, 2002). Ayant d'abord émergées dans le domaine de l'éducation, elles sont maintenant présentes dans la plupart des domaines de recherche, plus particulièrement lorsque le devis de recherche implique une ou plusieurs variables contextuelles. Il est fréquent que les données soient hiérarchisées (des élèves dans différentes classes, des participants provenant de différentes villes, des observations répétées de chaque participant, etc.). Si ce regroupement des données n'est pas pris en compte par l'entremise des analyses multiniveaux, les analyses statistiques peuvent s'avérer invalides et l'in-

Tableau 4 ■ Effets fixes du modèle 1

	Coefficients	Err. std.	t
Ordonnée	0,266364	0,007417	35,91***
Anxiété de performance	-0,54753	0,007422	-73,80***

Note. La sous-estimation des erreurs standards (Err. Std.) est illustré dans la sortie des effets fixes du modèle 1. *** p < 0,001

terprétation des résultats incorrecte.

Les analyses de régressions multiniveaux sont parfois chronophages, semblent difficiles à utiliser pour le nouvel utilisateur et l'interprétation des résultats peut se complexifier en comparaison aux analyses plus répandues comme la régression linéaire simple. Afin de démocratiser son utilisation, cet article tutoriel a couvert les principes généraux sous-jacents aux analyses de régressions multiniveaux pour analyser des données nichées. En utilisant la génération de données pseudoaléatoires et un exemple hypothétique, l'article démontre l'avantage de considérer la hiérarchisation des données quant à la validité des résultats. De plus, cet article incorpore la syntaxe R nécessaire pour faciliter l'utilisation des analyses multiniveaux et l'adaptation de celles-ci aux données des lecteurs. Cet article s'inscrit dans le courant désirant rendre les analyses dans le domaine de la recherche en psychologie plus rigoureuses, espérant ainsi combattre la crise de la réplicabilité.

Note des auteurs

Les deux premiers auteurs ont contribué également au présent article.

Références

Barton, K. (2020). MuMIn : Multi-Model Inference [R package] (Version 1.43.17). Récupérée à partir de https://CRAN.R-project.org/package=MuMIn

Bates, D., Kliegl, R., Vasishth, S. & Baayen, H. (2018). *Parsimonious mixed models*. online : arxiv. arXiv : 1506.04967. Récupérée à partir de http://arxiv.org/abs/1506.04967

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. doi :10.18637/jss.v067.i01

Blanca, M. J., Alarcón, R. & Bono, R. (2018). Current practices in data analysis procedures in psychology : What Has Changed ? *Frontiers in Psychology*, *9*, 9-99. doi :10.3389/fpsyg.2018.02558

Caron, P.-O. & Lemardelet, L. (2021). The variance sum law and its implication for modelling. *The Quantitative Methods for Psychology*, *17*(2), 80-85. doi :10.20982/tqmp.17.2.p080

Cousineau, D. & Laurencelle, L. (2015). A correction factor for the impact of cluster randomized sampling and its applications. *Psychological Methods*, *21*(1), 121-135. doi :10.1037/met0000055

Field, A., Miles, J. & Field, Z. (2012). *Discovering statistics using R*. Thousand Oaks : SAGE.

Fox, J. & Weisberg, S. (2018). *An R companion to applied regression*. Thousand Oaks : SAGE.

Gelman, A. & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge : Cambridge University Press.

Goldstein, H. (1995). *Multilevel statistical models*. London : Edward Arnold.

Hox, J. J., Moerbeek, M. & Van de Schoot, R. (2017). *Multilevel analysis : Techniques and applications*. New York : Routledge.

Kish, L. (1965). *Survey sampling*. New York : Wiley.

Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. (2017). lmerTest package : Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1-26. doi :10.18637/jss.v082.i13

Lai, M. H. & Kwok, O. M. (2015). Examining the rule of thumb of not using multilevel modeling : The "design effect smaller than two" rule. *The Journal of Experimental Education*, *83*(3), 423-438. doi :10.1080/00220973.2014.907229

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P. & Makowski, D. (2021). performance : An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, *6*(60), 1-8. doi :10.21105/joss.03139

McNeish, D. M. & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates : A review and illustration. *Educational Psychology Review*, *28*(2), 295-314. doi :10.1007/s10648-014-9287-x

Moeyaert, M., Rindskopf, D., Onghena, P. & Van den Noortgate, W. (2017). Multilevel modeling of single-case data : A comparison of maximum likelihood and Bayesian estimation. *Psychological Methods*, *22*, 760-799. doi :10.1037/met0000136

Montgomery, D. C., Peck, E. A. & Vining, G. G. (2021). *Introduction to linear regression analysis*. New York : Wiley.

Muthén, B. O. & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, *25*, 267-316. doi :10.2307/271070

R Core Team. (2021). *R : A language and environment for statistical computing*. Vienna, Austria : R Foundation for Statistical Computing. Récupérée à partir de https://www.R-project.org/

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models*. Thousand Oaks : Sage.

Rights, J. D. & Sterba, S. K. (2018). Quantifying explained variance in multilevel models : an integrative framework for defining r-squared measures. *Psychological Methods*, *23*(3), 434-457. doi :10.1037/met0000184

Twisk, J. W. R. (2006). *Applied multilevel analysis : A practical guide*. Cambridge : Cambridge University Press.

van der Zee, T. & Reich, J. (2018). Open education science. *AERA Open*, *4*(3), 1-15. doi :10.1177/2332858418787466

Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S*. New York : Springer.

von der Embse, N., Jester, D., Roy, D. & Post, J. (2018). Test anxiety effects, predictors, and correlates : A 30-year meta-analytic review. *Journal of affective disorders*, *227*, 483-493. doi :10.1016/j.jad.2017.11.048

Weisz, J., Bearman, S. K., Santucci, L. C. & Jensen-Doss, A. (2017). Initial test of a principle-guided approach to transdiagnostic psychotherapy with children and adolescents. *Journal of Clinical Child & Adolescent Psychology*, *46*(1), 44-58. doi :10.1080/15374416.2016.1163708

Wickham, H. (2016). *ggplot2 : Elegant graphics for data analysis*. New York : Springer-Verlag.

Appendix : Listings

```
ggplot(data = jd) +
  geom_histogram(mapping = aes(x = RésiduelsLM)) + xlab("Résiduels") + ylab("Fréquence") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
      panel.background = element_blank(), axis.line = element_line(colour = "black"),
      text = element_text(size = 12), axis.title = element_text(face = "bold"))

ggplot(data = jd) +
  geom_histogram(mapping = aes(x = residuels)) + xlab("Résiduels") + ylab("Fréquence") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.background = element_blank(), axis.line = element_line(colour = "black"),
        text = element_text(size = 12), axis.title = element_text(face = "bold"))
```

Note. Syntaxe R pour produire les histogrammes des erreurs résiduelles des modèles 1 et 3.

```
ggplot (jd) + (mapping = aes(x = PréditsLM, y = RésiduelsLM)) + xlab("Prédits") + ylab("Résiduels") +
  geom_point() +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.background = element_blank(), axis.line = element_line(colour = "black"),
        text = element_text(size = 12), axis.title = element_text(face = "bold"))

ggplot (jd) + (mapping = aes(x = Prédits, y = Résiduels)) + xlab("Prédits") + ylab("Résiduels") +
  geom_point() +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.background = element_blank(), axis.line = element_line(colour = "black"),
        text = element_text(size = 12), axis.title = element_text(face = "bold"))
```

Note. Syntaxe R pour produire les diagrammes de dispersion des valeurs prédites par résiduels des modèles 1 et 3.

Citation

Duplessis-Marcotte, F., Lapointe, R. & Caron, P.-O. (2022). Une introduction aux modèles de régressions multiniveaux avec R. *The Quantitative Methods for Psychology*, *18*(2), 168-180. doi :10.20982/tqmp.18.2.p168

Received: 13/04/2022 ~ Accepted: 22/04/2022