Modélisation multiniveaux bivariée de la santé mentale perçue et du sentiment d'appartenance au quartier chez les jeunes adultes de Sherbrooke

par

Kossi Ekouagou

mémoire présenté au Département de mathématiques
en vue de l'obtention du grade de maître ès sciences (M.Sc.)

FACULTÉ DES SCIENCES
UNIVERSITÉ DE SHERBROOKE

Sherbrooke, Québec, Canada, février 2023

À MES ENFANTS CHRISTOPHE ET LIGHT
À MON ÉPOUSE ESSE
EN MEMOIRE DE MA MÈRE

Le 1^{er} février 2023

Le jury a accepté le mémoire de Monsieur Kossi Ekouagou dans sa version finale.

Membres du jury

Professeur Félix Camirand Lemyre

Directeur de recherche

Département de mathématiques

Professeure Martine Shareck

Codirectrice de recherche

Département des sciences de la santé communautaire

Professeur Taoufik Bouezmarni

Président-rapporteur

Département de mathématiques

Professeur Klaus Herrmann

Membre interne

Département de mathématiques

SOMMAIRE

Les inégalités sociales de santé recouvrent les différences d'état de santé entre individus ou groupes d'individus, liées à des facteurs individuels et contextuels. La communauté dans laquelle cohabitent et interagissent les individus exerce une influence sur leurs états de santé. Par exemple, l'isolement social et un faible sentiment d'appartenance au quartier de résidence, tous deux associés à divers états de santé, sont plus prévalents chez les personnes plus défavorisées. Plusieurs chercheurs ont tenté de comprendre comment différents aspects de l'environnement physique et social interagissent avec les caractéristiques des individus pour produire une différence de santé entre eux, mais peu d'études ont mené une analyse par approche mixte, notamment l'approche hiérarchique avec un *outcome* bivarié pour mieux comprendre ce problème. Cette approche tient compte de la corrélation entre les individus d'une même grappe et permet d'estimer le lien entre un *outcome* bivarié et les covariables ; ce que ne fait pas le modèle linéaire simple.

L'équipe de Martine Shareck mène actuellement une étude qui vise à mieux comprendre comment l'environnement physique et social des quartiers de Sherbrooke peut influencer le sentiment d'appartenance au quartier, les liens sociaux et le bien-être chez les jeunes et à déterminer si ces éléments divergent entre groupes sociaux. Il est question d'estimer le lien entre un *outcome* bivarié et des covariables avec des données hiérarchiques, les participants à l'étude (niveau 1) imbriqués dans des communautés locales de Sherbrooke prises comme quartiers (niveau 2).

Lorsque les données présentent une structure en grappe, une approche couramment utilisée est le modèle mixte. L'objectif de ce mémoire est de voir comment il est possible d'utiliser l'approche du modèle mixte pour mieux comprendre les différences de santé mentale perçue et du sentiment d'appartenance au quartier chez les jeunes adultes après avoir contrôlé pour les facteurs individuels (âge, genre, éducation, revenu...) et contextuel (niveau de défavorisation matérielle relative au quartier).

Nous avons donc estimé un modèle à deux niveaux avec un *outcome* bivarié qui est un modèle mixte. Les deux *outcomes* sont le score de santé mentale perçue et le score du sentiment d'appartenance au quartier. Les données individuelles proviennent des jeunes participants à l'étude résidents de Sherbrooke (Québec), au moment de remplir le questionnaire en ligne. Les données caractérisant les quartiers de Sherbrooke sont notamment le niveau de défavorisation matérielle et proviennent du tableau de bord des communautés de l'Estrie. Nous avons posé deux principales hypothèses que nous avons testées : les perceptions du quartier exercent différentes influences sur la santé mentale perçue et sur le sentiment d'appartenance au quartier chez les jeunes adultes ; le niveau de défavorisation matérielle exerce plus d'influence sur le score du sentiment d'appartenance au quartier que sur le score de santé mentale perçue des jeunes adultes.

Suivant les principes du modèle mixte et la stratégie d'analyse multiniveaux multivariée, la première série d'analyses a introduit uniquement les deux *outcomes* pour former un modèle de base. Ensuite, nous avons estimé le modèle vide au niveau 1. L'objectif est d'obtenir certaines statistiques et la déviance du modèle qui seront utilisées pour tester l'effet d'une covariable par la suite.

Dans la deuxième série d'analyses, une covariable est incluse dans le modèle. L'objectif est de déterminer s'il y a une différence significative en moyenne d'au moins un des *outcomes* pour une covariable donnée. Si un effet global significatif de covariable est présent, alors l'effet de la covariable pour chaque *outcome* est estimé et testé pour sa signification.

Dans la troisième série d'analyses, nous avons testé si l'impact d'une covariable est le même pour chaque *outcome* en contraignant les effets fixes à être égaux, puis nous avons testé la différence d'ajustement en utilisant les déviances entre ce modèle contraint et le modèle où les effets sont estimés librement.

Le niveau du quartier est inclus finalement dans le modèle d'analyse par la suite avec de multiples prédicteurs. Les paramètres des effets fixes et de variances-covariances ont été estimés. Les paramètres de variances-covariances ont permis d'évaluer la part de variabilité imputable au niveau du quartier pour chacun des deux *outcomes*.

REMERCIEMENTS

Je tiens d'abord à remercier mes codirecteurs de recherche Félix Camirand Lemyre et Martine Shareck pour les nombreux échanges respectivement en statistique, en épidémiologie sociale et géographique qui m'ont été si bénéfiques. Je vous remercie pour la confiance que vous m'avez accordée en acceptant de me confier ce projet de recherche. Merci de vos disponibilités, de vos conseils, de votre encadrement, de vos commentaires éclairés et de votre aide tout au long de ma maîtrise.

Je remercie tous les enseignants ayant intervenu dans mes cours durant la première partie de cette maîtrise et les membres de l'équipe statistique du département de mathématiques de l'Université de Sherbrooke, tant pour les discussions en statistique que pour l'esprit de camaraderie qui a toujours régné au sein de l'équipe.

Je remercie le groupe de recherche interdisciplinaire en informatique de la santé (Griis) pour son soutien financier qui m'a permis en partie de réaliser ce programme de maîtrise de type recherche.

Finalement, j'adresse un grand merci à mon épouse et à mes enfants, qui m'ont accompagné avec force et morale dans la réalisation de cette maîtrise.

Kossi Ekouagou

Sherbrooke, janvier 2023

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

INTRODUCTION

Les inégalités en matière de santé se définissent comme des différences de santé entre les individus liées à des facteurs ou des critères sociaux de différenciation (classe sociale, catégories socioprofessionnelles, catégories de revenu, niveaux d'étude, etc.) [8, 28]. Par exemple, les individus dont le statut socio-économique est moins favorable sont en moins bonne santé que celles dont le statut est plus favorable [8]. La communauté dans laquelle cohabitent et interagissent les individus exerce une influence sur leurs états de santé. Le concept de *défavorisation* souvent utilisé comme critère de différenciation sociale, sert entre autres, à l'évaluation de ces inégalités au niveau contextuel et se réfère à un désavantage face à la communauté locale à laquelle appartient l'individu [24, 60].

Pour expliquer ces inégalités, peu d'études ont eu recours à une analyse par approche hiérarchique avec un *outcome* bivarié comme la santé mentale perçue et le sentiment d'appartenance au quartier alors qu'au niveau macro, il y a une dimension environnementale dans laquelle se trouvent les individus [56]. De fait, les enquêtes répétées sur l'ampleur de ces inégalités adoptent un modèle social de la santé, qui place l'individu au centre entouré de *couches d'influence* liées aux facteurs de style de vie, aux réseaux sociaux et communautaires, aux conditions de vie et de travail et à l'environnement socio-économique et culturel [38, 4].

Plusieurs chercheurs ont tenté de comprendre comment différents aspects de l'environnement physique, social, familial, organisationnel et communautaire interagissent avec les caractéristiques des individus pour produire une différence de santé entre eux [31, 36, 39]. Certains, comme Robert, en 1999, ont documenté le fait que les processus qui produisent des différences en matière de santé peuvent avoir leur origine dans les expositions individuelles ou dans les milieux de vie [47]. D'autres travaux de recherche ont montré que les populations exposées à la défavorisation peuvent percevoir un faible sentiment d'appartenance à leur communauté à celles qui n'y sont pas exposées et que le sentiment d'appartenance à la communauté est fortement corrélé à l'état de santé physique et mentale, même si l'on tient compte de l'effet de l'âge, du statut socio-économique et d'autres facteurs [49, 55]. Même Statistique-Canada, dans l'un de ces rapports officiels, a relevé que les individus ayant un fort sentiment d'appartenance à la communauté étaient nettement plus susceptibles de déclarer leur santé excellente ou très bonne que celles dont le sentiment d'appartenance à la communauté est faible, même si l'on tient compte de l'effet d'autres facteurs qui peuvent éventuellement mêler les cartes.

Il est important de reprendre les études sur les inégalités de santé et du sentiment d'appartenance à la communauté pour évaluer les effets fixes et aléatoires à l'aide d'une méthode quantitative avancée [9] afin de mieux comprendre le phénomène [18, 12]. Ainsi, un ensemble de techniques statistiques, telles que le modèle de régression à effet mixte, dont le modèle multiniveaux, le modèle à équations structurelles, et le modèle bayésien, offrent aux chercheurs d'outils avancés pour mieux aborder ces défis [6, 45].

Si on se penche sur le modèle multiniveaux, il réfère à la présence de plateaux, d'échelons ou de couches. Il s'ensuit que d'un point de vue conceptuel, ce modèle désigne un ensemble de prédictions ou d'explications qui s'étendent sur plusieurs unités d'analyse de systèmes vivants. Du point de vue méthodologique, il renvoie à la mesure ou à la manipulation de variables situées à plusieurs niveaux d'analyse de systèmes vivants.

Du point de vue statistique, il réfère à un ensemble de techniques statistiques qui s'inscrit dans le cadre de la généralisation du *GLM* (*Generalized Linear Model*) et qui permet le traitement de données provenant de plusieurs unités d'analyse [45, 63]. Typiquement, ce modèle permet d'estimer les sources de variance intra-unité et inter-unité à l'aide de la corrélation intra-classe, de déterminer la présence d'effets aléatoires et de quantifier les effets fixes [18]. Bien que les modèles mathématiques à la base de cette analyse aient été élaborés depuis près de trente ans, leur utilisation ne s'est généralisée qu'avec le développement des logiciels performants [18]. Les toutes premières applications portaient sur les déterminants du rendement scolaire [5, 32], celles en santé s'étaient véritablement mises en œuvre à partir des années 90 et prennent de plus en plus d'ampleur [10, 46, 11, 36]. Au cours des vingt dernières années, il y a eu donc une multitude d'applications de cette analyse [9, 29, 39, 61] dans lesquelles les descriptions des techniques ont été développées. Dans un premier temps, on a apparié les individus à leur quartier de résidence, leur communauté ou leur municipalité. Ensuite, on a appliqué une analyse multiniveaux simple où les individus sont imbriqués dans un milieu de vie donné. Une revue critique de 25 études récentes [39] indique que même si les variables individuelles expliquent une proportion importante de la variance de divers indicateurs de santé, on retrouve de façon systématique des effets de contexte, notamment le niveau de favorisation matérielle et sociale qui est associé à de meilleurs états de santé, et ce même après un contrôle des variations compositionnelles. Il y a eu également des applications de ce modèle à plus de deux niveaux où les individus appartiennent à des ménages et ceux-ci à des quartiers [36]. Certaines études comme celles de Raudenbush en 2001 [44] avaient pris une cohorte d'individus suivis longitudinalement où les mesures répétées sont imbriquées dans les individus et ensuite dans les quartiers, dont l'intérêt est de décrire et d'expliquer les profils de changement en fonction des caractéristiques des individus et de leur quartier.

Cependant, peu d'études ont appliqué ce modèle avec deux *outcomes* même si quelques chercheurs anglo-saxons l'ont déjà appliqué en éducation et en sciences sociales [20, 41]. Plus adapté à ces types d'études, l'estimation du modèle multiniveaux avec deux *outcomes* conduit à des erreurs standards plus petites pour les tests des covariables sur un *outcome* donné, par rapport à un modèle à un seul *outcome* ; ce qui augmente la puissance de l'approche lorsque les deux *outcomes* sont surtout corrélés et que les individus ont des données manquantes pour certaines des *outcomes* [57].

Actuellement, Martine Shareck et son équipe mènent une étude nommée *CentrÉS* qui vise à mieux comprendre comment l'environnement physique et social des quartiers de Sherbrooke peut influencer le sentiment d'appartenance au quartier, les liens sociaux et le bien-être chez les jeunes et à déterminer si ces éléments divergent entre groupes sociaux. Il y a en vue de cela un problème d'estimation du lien entre un *outcome* bivarié et des covariables, alors que les données sont structurées hiérarchiquement. Le modèle linéaire simple ne convient pas tout à fait à cette étude puisqu'il ne tient pas compte de la corrélation entre les individus d'une même grappe. Or, dans la littérature en santé, lorsque les données présentent une structure en grappe, une approche couramment employée est de recourir au modèle mixte. Le but de ce mémoire est de voir comment il est possible d'utiliser l'approche par modèle mixte dans ce contexte.

Au premier chapitre du mémoire, le modèle linéaire mixte sera décrit dans sa forme la plus générale avec ses propriétés, à partir de la présentation du modèle linéaire simple. Ensuite, au second chapitre, le modèle hiérarchique sera décrit avec un *outcome* univarié et avec un *outcome* bivarié. Au troisième chapitre, le modèle hiérarchique avec un *outcome* bivarié sera appliqué aux données *CentrÉS*.

Mots-clés : Modèle linéaire mixte, modèle multiniveaux bivarié, santé mentale perçue, sentiment d'appartenance au quartier, appartenance ethnique, défavorisation matérielle

CHAPITRE 1

Le modèle mixte

1.1 Du modèle linéaire simple au modèle linéaire mixte

L'analyse de variance, la régression linéaire et l'analyse de covariance sont des méthodes qui proviennent de la théorie du modèle linéaire. Dans ce modèle, les variables indépendantes apparaissent sous la forme d'effets fixes. Lorsque cette variable est qualitative, on s'intéresse à l'effet particulier de chacun de ses niveaux ou modalités sur la variable dépendante. Ce procédé suppose que l'on introduit dans le modèle tous les niveaux du facteur susceptible d'avoir un intérêt, mais qui n'est pas toujours possible. Si on s'intéresse par exemple à l'effet d'une catégorie de revenu personnel sur la santé mentale perçue, on ne pourra pas le tester sur toute la population. À chaque fois, pour réaliser l'expérience, il faudra prendre un échantillon d'individus et chercher à étendre les résultats obtenus à la population entière. Si on suppose que ces individus ont été tirés au hasard dans la population, on ne s'intéresse plus à l'effet particulier associé à tel individu particulier, mais à la distribution de l'ensemble des effets possibles. L'effet associé à l'individu n'est plus fixe, mais devient aléatoire et il faut en tenir compte dans l'analyse.

Dans ce cas, le modèle contient un mélange d'effets ; on parle alors du modèle mixte. Le modèle mixte constitue donc une extension du modèle linéaire classique. On pourra y faire appel chaque fois que l'on désirera étendre à une population tout entière des résultats obtenus sur un échantillon d'individus pris au hasard dans cette population.

À partir du modèle linéaire simple, nous présentons dans ce chapitre le modèle linéaire mixte dans sa forme la plus générale et nous étudions ses propriétés (estimations et tests) inspirées du livre *Generalized, Linear,and Mixed Models* [53] et des documents scientifiques de recherche de Christèle Robert-Granié et al. [14, 48]. Tout au long du mémoire, l'espérance de la variable de réponse (*outcome*), souvent noté y_{ij}, sera une fonction linéaire de certains paramètres inconnus à estimer. Nous utiliserons alors dans l'ensemble du document une notation spéciale en prenant le même symbole (y_{ij}) pour la variable aléatoire (Y_{ij}) et l'observation, qui est la réalisation de cette variable aléatoire. Les paramètres notés par une lettre grecque seront les paramètres des effets fixes et ceux notés par une lettre romaine seront les paramètres des effets aléatoires.

1.1.1 Définition

Un modèle mixte est un modèle dans lequel toute ou une partie des paramètres associés à certaines unités expérimentales sont traités comme des variables aléatoires. Dans ce modèle, les niveaux des effets fixes étant fixés une fois pour toutes, les effets associés sont des paramètres à estimer qui interviennent dans la moyenne du modèle. Les facteurs à effets aléatoires ont souvent un grand nombre de niveaux, les observations réalisées correspondent à un nombre restreint de ces niveaux, pris aléatoirement. On va modéliser ces niveaux en tant qu'observations d'une variable aléatoire normale, de moyenne nulle et de variance inconnue à estimer. Chaque facteur à effets aléatoires sera caractérisé par un paramètre de variance qu'il faudra estimer en plus de la variance des erreurs du modèle.

1.1.2 Exemples illustratifs

Dans cette sous-section, nous allons illustrer quelques exemples en santé montrant d'abord la nature d'un facteur puis la formulation d'un modèle à effet aléatoire, fixe et mixte.

Exemple 1. Supposons que l'on cherche à comparer 2 groupes d'appartenance ethnique (A = minorité ethnique, B = majorité ethnique) vis-à-vis du score de santé mentale perçue dans une région donnée. Quatre quartiers de résidence ont été sélectionnés pour participer à cette expérience. Dans chaque quartier de résidence, un échantillon d'individus a été tiré au hasard, une moitié font partie du groupe A et l'autre moitié du groupe B. Les analyses ont montré que le groupe B perçoit plus une bonne santé mentale que le groupe A. Que peut-on conclure ? Pour répondre convenablement à cette question, il est nécessaire de préciser la nature du facteur quartier. Si les quartiers ont été choisis, le facteur quartier sera un facteur fixe et les résultats de l'analyse ne peuvent pas être extrapolés à d'autres quartiers. Si les quartiers ont été sélectionnés au hasard parmi tous les quartiers de la ville, le facteur quartier sera alors un facteur aléatoire et les résultats de cette analyse peuvent être extrapolés aux autres quartiers.

Exemple 2. On a obtenu les scores du sentiment d'appartenance au quartier de 25 jeunes adultes résidant dans 5 quartiers d'une ville donnée qui avaient été tirés au sort dans la population. On voudrait savoir dans quelle mesure le score d'appartenance au quartier est influencé par les effets du quartier dans lequel résident ces jeunes.
Il s'agit de répondre à une question concernant toute la population. Pour pouvoir étendre les résultats obtenus sur l'échantillon, il faut que celui-ci soit représentatif de toute la population. Il est obtenu par tirage au sort (indépendants et équiprobables). Il en découle que les quartiers de résidence de l'échantillon sont aléatoires et leurs effets sur les scores du sentiment d'appartenance au quartier sont à fortiori aléatoires. Le modèle s'écrit :

$$y_{ij} = \mu + a_i + e_{ij} \text{ avec } j = 1, \ldots, 25 \text{ et } i = 1, \ldots, 5,$$

où y_{ij} représentent ici le score du sentiment d'appartenance au quartier du jeune adulte j du quartier i, μ : la moyenne générale, a_i : l'effet du quartier i, supposé aléatoire (puisque le quartier est choisi aléatoirement) et indépendant de e_{ij} (résidu indépendant de $e_{ij'}$). On suppose les distributions suivantes : $a_i \sim N(0, \sigma_a^2)\ \forall\ i$; $e_{ij} \sim N(0, \sigma_e^2)$. On appelle σ_a^2 et σ_e^2 les composantes de la variance. La quantité $\frac{\sigma_a^2}{\sigma_a^2+\sigma_e^2}$ représente ici la part de variabilité due à l'hétérogénéité des quartiers par rapport à la variance totale.

Il faut noter que $\mu + a_i$ n'est pas l'espérance de y_{ij}, mais son espérance conditionnelle, soit : $\mathbb{E}[y_{ij}|a_i] = \mu + a_i$. La variance conditionnelle de y_{ij} est : $\mathrm{Var}[y_{ij}|a_i] = \sigma_e^2$.

Il en découle que l'espérance et la variance de y_{ij} sont :

$\mathbb{E}[y_{ij}] = \mathbb{E}[\mathbb{E}[y_{ij}|a_i]] = \mathbb{E}[\mu + a_i] = \mathbb{E}[\mu] + \mathbb{E}[a_i] = \mu$,

$\mathrm{Var}[y_{ij}] = \mathrm{Var}[\mathbb{E}[y_{ij}|a_i]] + \mathbb{E}[\mathrm{Var}[y_{ij}|a_i]] = \mathrm{Var}[\mu + a_i] + \mathbb{E}[\sigma_e^2] = \sigma_a^2 + \sigma_e^2$.

On peut calculer également les covariances : $\mathrm{Cov}(y_{ij}, y_{ij'})$ est la covariance entre deux observations (score du sentiment d'appartenance au quartier) de deux jeunes adultes j et j' du même quartier i, alors que $\mathrm{Cov}(y_{ij}, y_{i'j'})$ est celle entre deux observations (score du sentiment d'appartenance au quartier) de deux jeunes adultes j et j' issues de deux quartiers différents i et i'. En effet, on a :

$$
\begin{aligned}
\mathrm{Cov}(y_{ij}, y_{ij'}) &= \mathrm{Cov}(\mathbb{E}[y_{ij}|a_i], \mathbb{E}[y_{ij'}|a_i]) + \mathbb{E}[\mathrm{Cov}(y_{ij}, y_{ij'}|a_i)] \\
&= \mathrm{Cov}(\mu + a_i, \mu + a_i) + \mathbb{E}[\mathbb{E}[y_{ij}y_{ij'}|a_i] - \mathbb{E}[y_{ij}|a_i]\mathbb{E}[y_{ij'}|a_i]] \\
&= \mathrm{Cov}(a_i, a_i) + \mathbb{E}[\mathbb{E}[y_{ij}y_{ij'}|a_i]] - \mathbb{E}[\mathbb{E}[y_{ij}|a_i]\mathbb{E}[y_{ij'}|a_i]] \\
&= \mathrm{Cov}(a_i, a_i) + \mathbb{E}[y_{ij}y_{ij'}] - \mathbb{E}[\mathbb{E}[y_{ij}|a_i]]\mathbb{E}[\mathbb{E}[y_{ij'}|a_i]] \\
&= \mathrm{Cov}(a_i, a_i) + \mathbb{E}[y_{ij}y_{ij'}] - \mathbb{E}[y_{ij}]\mathbb{E}[y_{ij'}] \\
&= \mathrm{Cov}(a_i, a_i) + \mathbb{E}[y_{ij}]\mathbb{E}[y_{ij'}] - \mathbb{E}[y_{ij}]\mathbb{E}[y_{ij'}] \text{ (car } e_{ij} \text{ est indépendant de } e_{ij'}) \\
&= \mathrm{Cov}(a_i, a_i) + 0 \\
&= \mathrm{Var}[a_i] \\
&= \sigma_a^2,
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Cov}(y_{ij}, y_{i'j'}) &= \mathrm{Cov}(\mathbb{E}[y_{ij}|a_i], \mathbb{E}[y_{i'j'}|a_{i'}]) + \mathbb{E}[\mathrm{Cov}(y_{ij}, y_{i'j'}|a_{i'})] \\
&= \mathrm{Cov}(\mu + a_i, \mu + a_{i'}) + \mathbb{E}[\mathbb{E}[y_{ij}y_{i'j'}|a_{i'}] - \mathbb{E}[y_{ij}|a_{i'}]\mathbb{E}[y_{i'j'}|a_{i'}]] \\
&= \mathrm{Cov}(a_i, a_{i'}) + \mathbb{E}[\mathbb{E}[y_{ij}y_{i'j'}|a_{i'}]] - \mathbb{E}[\mathbb{E}[y_{ij}|a_{i'}]\mathbb{E}[y_{i'j'}|a_{i'}]] \\
&= \mathrm{Cov}(a_i, a_{i'}) + \mathbb{E}[y_{ij}y_{i'j'}] - \mathbb{E}[\mathbb{E}[y_{ij}|a_{i'}]]\mathbb{E}[\mathbb{E}[y_{i'j'}|a_{i'}]] \\
&= \mathrm{Cov}(a_i, a_{i'}) + \mathbb{E}[y_{ij}y_{i'j'}] - \mathbb{E}[y_{ij}]\mathbb{E}[y_{i'j'}] \\
&= 0 + \mathbb{E}[y_{ij}y_{i'j'}] - \mathbb{E}[y_{ij}]\mathbb{E}[y_{i'j'}] \quad (\text{car } a_i \text{ est supposé indépendant de } a_{i'}) \\
&= \mathbb{E}[y_{ij}]\mathbb{E}[y_{i'j'}] - \mathbb{E}[y_{ij}]\mathbb{E}[y_{i'j'}] \quad (\text{car } e_{ij} \text{ est supposé indépendant de } e_{i'j'}) \\
&= 0.
\end{aligned}
$$

Dans cet exemple, on note que deux observations de deux quartiers différents ne sont pas corrélées ($\mathrm{Cov}(y_{ij}, y_{i'j'}) = 0$) alors que deux observations du même quartier sont corrélées ($\mathrm{Cov}(y_{ij}, y_{ij'}) = \sigma_a^2$). Le coefficient de corrélation vaut alors :

$$
\begin{aligned}
\mathrm{Cor}(y_{ij}, y_{ij'}) &= \frac{\mathrm{Cov}(y_{ij}, y_{ij'})}{\sqrt{\mathrm{Var}(y_{ij})\mathrm{Var}(y_{ij'})}} \\
&= \frac{\sigma_a^2}{\sqrt{(\sigma_a^2 + \sigma_e^2)(\sigma_a^2 + \sigma_e^2)}} \\
&= \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}.
\end{aligned}
$$

Exemple 3. Un observatoire du bien-être social a recueilli les perceptions des jeunes adultes sur l'appartenance au quartier de résidence. Pour cela, 10 quartiers de résidence ont été sélectionnés au hasard. Pour chaque quartier de résidence, les scores du sentiment d'appartenance au quartier ont été obtenus chez ces jeunes.

La question posée est : existe-t-il une différence significative du score de sentiment d'appartenance au quartier entre les jeunes hommes et les jeunes femmes ?

Le fichier de données est le suivant :

Quartier	$Hommes$	$Femmes$	Différence ($diff$)
1	10.4	10.1	0.3
2	10.6	10.8	-0.2
3	10.2	10.2	0.0
4	10.1	9.9	0.2
5	10.3	11.0	-0.7
6	10.7	10.2	0.2
7	10.3	10.2	0.1
8	10.9	10.9	0.0
9	10.1	10.4	-0.3
10	9.8	9.9	-0.1

TABLEAU 1.1 – Données des quartiers sélectionnés au hasard

Nous effectuons d'abord les calculs suivants :

$$\overline{y}_{Hommes} = \frac{1}{10}\sum_{i=1}^{10} y_{iHommes} = \frac{1}{10}(10.4 + 10.6 + ... + 9.8) = 10.34,$$

$$\overline{y}_{Femmes} = \frac{1}{10}\sum_{i=1}^{10} y_{iFemmes} = \frac{1}{10}(10.1 + 10.8 + ... + 9.9) = 10.36,$$

$$\overline{diff} = \frac{1}{10}\sum_{i=1}^{10} diff_i = \frac{1}{10}\sum_{i=1}^{10}(y_{iHommes} - y_{iFemmes}) = \frac{1}{10}(0.3 - 0.2 + ... - 0.1) = -0.05,$$

$$\begin{aligned} s_{Hommes} &= \sqrt{\frac{1}{10-1}\sum_{i=1}^{10}(y_{iHommes} - \overline{y}_{Hommes})^2} \\ &= \sqrt{\frac{1}{10-1}[(10.4 - 10.34)^2 + (10.6 - 10.34)^2 + ... + (9.8 - 10.34)^2]} \\ &= 0.3239, \end{aligned}$$

$$\begin{aligned} s_{Femmes} &= \sqrt{\frac{1}{10-1}\sum_{i=1}^{10}(y_{iFemmes} - \overline{y}_{Femmes})^2} \\ &= \sqrt{\frac{1}{10-1}[(10.1 - 10.36)^2 + (10.8 - 10.36)^2 + ... + (9.9 - 10.36)^2]} \\ &= 0.4033, \end{aligned}$$

$$
\begin{aligned}
s_{diff} &= \sqrt{\frac{1}{10-1}\sum_{i=1}^{10}(y_{idiff} - \overline{diff})^2} \\
&= \sqrt{\frac{1}{10-1}[(0.3+0.05)^2 + (-0.2+0.05)^2 + ... + (-0.1+0.05)^2]} \\
&= 0.2953.
\end{aligned}
$$

(a) L'analyse la plus simple consiste à considérer les données comme un échantillon apparié et à utiliser le test de Student correspondant.

La moyenne et l'écart-type des différences $diff$ sont respectivement $\overline{diff} = -0.05$ et $s_{diff} = 0.2953$. La statistique du test de Student à $n-1 = 10-1 = 9$ degrés de liberté vaut $t = \frac{\overline{diff}}{SE_{\overline{diff}}} = \frac{-0.05}{0.2953/\sqrt{10}} = -0.5354$ qui donne une p-valeur $= 0.6054$.

Il n'y a donc pas de différence significative du score du sentiment d'appartenance au quartier entre les jeunes hommes et les jeunes femmes. L'intervalle de confiance à 95% de la différence est :

$\overline{diff} \pm t_{(0.975;\ 9)} \times SE_{\overline{diff}} = -0.05 \pm 2.262 \times 0.2953/\sqrt{10} = -0.05 \pm 0.2112$; soit $[-0.2612; 0.1612]$.

(b) ANOVA :

Le modèle d'analyse de variance pour cette étude s'écrit :

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \qquad e_{ij} \sim N(0, \sigma^2),$$

où y_{ij} représente le score du sentiment d'appartenance au quartier i du genre j, μ la moyenne générale, α_i l'effet du $i$ème quartier, et β_j l'effet du genre j. La statistique du test F de Fisher est : $F = t^2 = (-0.5354)^2 = 0.2867$.

L'estimation de l'écart-type résiduel est $\hat{\sigma} = 0.2090 = \sqrt{2}s_{diff}$. L'incertitude sur la moyenne des différences est $SE(\overline{y}_{Hommes} - \overline{y}_{Femmes}) = \sqrt{\sigma^2(1/10 + 1/10)} = 0.0934$. Elle est exacte à $SE_{\overline{diff}} = 0.2953/\sqrt{10} = 0.0934$ dans l'approche simple en (a).

Si maintenant, on s'intéresse d'une part à la précision de la valeur moyenne pour les jeunes femmes adultes, l'ANOVA donne $SE(\overline{y}_{Femmes}) = \frac{\hat{\sigma}}{\sqrt{10}} = 0.0661$. D'autre part, si on considère l'échantillon des 10 mesures du score chez les jeunes femmes adultes, on obtient $s_{Femmes} = 0.4012$ et donc $SE(\overline{y}_{Femmes}) = \frac{s_{Femmes}}{\sqrt{10}} = 0.1270$, valeur très différente de 0.0661. L'ANOVA sous-estime donc l'incertitude sur l'estimation de l'effet moyen des jeunes femmes et cela est dû au fait que la variance σ^2 mesure la variabilité résiduelle après que les effets des quartiers ont été corrigés. La différence conceptuelle entre les deux approches est que l'ANOVA considère que les 10 quartiers de résidence n'ont pas été tirés au hasard, alors que la seconde (échantillon des 10 mesures chez les jeunes femmes) considère un tirage aléatoire. L'ANOVA n'est valide que si l'on s'intéresse aux effets spécifiques des 10 quartiers. L'idée du modèle mixte est de combiner les deux approches, c'est-à-dire utiliser un modèle linéaire et y considérer certains facteurs comme aléatoires.

(c) On considère maintenant le modèle linéaire mixte :

$$y_{ij} = \mu + a_i + \beta_j + \varepsilon_{ij} \tag{1.1}$$

où a_i est l'effet aléatoire du $i$ème quartier. Les a_i et ε_{ij} sont supposés iid (indépendants et identiquement distribués) : $a_i \sim N(0, \sigma_a^2)$, $\varepsilon_{ij} \sim N(0, \sigma^2)$. On peut avoir les propriétés suivantes :

	ANOVA	Modèle mixte
$\mathrm{E}[y_{ij}]$	$\mu + a_i + \alpha_i + \beta_j$	$\mu + \beta_j$
$\mathrm{Var}(y_{ij})$	σ^2	$\sigma^2 + \sigma_a^2$
$\mathrm{Cov}(y_{ij}, y_{i'j'}), j \neq j'$	0	$\sigma_a^2 \quad si \quad i = i' \quad et \quad 0 \quad sinon$

TABLEAU 1.2 – Espérance, variance et covariance entre l'ANOVA et le modèle mixte

L'écart-type de la moyenne des valeurs chez les jeunes femmes vaut dans le cadre du modèle linéaire mixte : $SE(\overline{y}_{femmes}) = \frac{\sqrt{\hat{\sigma}_a^2 + \hat{\sigma}^2}}{\sqrt{10}} = 0.115$. Il est plus faible que celui calculé dans les autres modèles.

Le modèle 1.1 peut s'écrire également sous forme matricielle :

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{a} + \boldsymbol{\varepsilon},$$

où $\boldsymbol{y} = (y_{1,Hommes}, ..., y_{10,Hommes}, y_{1,Femmes}, ..., y_{10,Femmes})'$ est le vecteur de dimension (20×1) des données du score du sentiment d'appartenance au quartier,

$\boldsymbol{\beta} = (\beta_{Hommes}, \beta_{Femmes})'$ est le vecteur de dimension (2×1) des effets fixes,

$\boldsymbol{a} = (a_1, ..., a_{10})'$ est le vecteur de dimension (10×1) des effets aléatoires,

$\boldsymbol{\varepsilon} = (\varepsilon_{1,Hommes}, ..., \varepsilon_{10,Hommes}, \varepsilon_{1,Femmes}, ..., \varepsilon_{10,Femmes})'$ est le vecteur de dimension (20×1) des résidus,

$\boldsymbol{X} = (\boldsymbol{x}_{Hommes}, \boldsymbol{x}_{Femmes})$ est la matrice d'incidence associée à $\boldsymbol{\beta}$ et formée des vecteurs colonnes $\boldsymbol{x}_{Hommes}= (1, ..., 1, 0, ...0)'$ et $\boldsymbol{x}_{Femmes} = (0, ..., 0, 1, ...1)'$,

$\boldsymbol{Z} = (\boldsymbol{I}_{10}, \boldsymbol{I}_{10})'$ est la matrice d'incidence associée à $\boldsymbol{a}$ avec $\boldsymbol{I}_{10}$: la matrice identité,

$\boldsymbol{a} \sim N_{10}(0, \sigma_a^2\boldsymbol{I})$ et $\boldsymbol{\varepsilon} \sim N_{20}(0, \sigma_e^2\boldsymbol{I})$ puis $\boldsymbol{a}$ et $\boldsymbol{\varepsilon}$ sont indépendants.

Par ailleurs, le modèle mixte est souvent utilisé dans le cadre de l'analyse de données longitudinales où plusieurs observations sont mesurées sur le même individu au cours du temps en prenant en compte des corrélations entre les observations d'un même individu. Si le nombre d'instants de mesures sur un individu est faible, on utilise le modèle mixte pour prendre en compte la dépendance entre observations issues du même individu. En revanche, si l'évolution dans le temps est le sujet d'intérêt, alors il faudra utiliser une autre catégorie de modèles ou l'analyse des trajectoires latentes.

Au vu des exemples illustrés, le modèle linéaire simple ne convient pas au cadre de l'étude puisqu'il ne tient pas compte de la corrélation entre les individus d'un même quartier. Dans la littérature en santé, lorsque les données présentent une structure en grappe, une approche couramment employée est de recourir au modèle linéaire mixte comme présenté. Dans la section suivante, nous allons étudier les propriétés (estimation et tests) et les critères de sélection du modèle linéaire mixte.

1.2 Spécification d'un modèle linéaire mixte

1.2.1 Écriture générale

Comme nous venons de l'illustrer dans l'exemple 3, un modèle linéaire gaussien mixte, relatif à n observations, s'écrit sous la forme matricielle suivante :

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Za} + \boldsymbol{\varepsilon}.$$

Nous précisons ci-dessous les éléments de cette écriture :

$\boldsymbol{y}$ est une matrice de taille $(n \times 1)$ des observations,

$\boldsymbol{X} \in \mathbb{R}^{n \times p}$, est une matrice d'indicatrices ou de variables continues pour les effets fixes,

$\boldsymbol{\beta}$ est une matrice de taille $(p \times 1)$ des paramètres des effets fixes,

$\boldsymbol{Z}$ est une matrice $(n \times K)$ d'indicatrices ou de variables continues pour les effets aléatoires,

$\boldsymbol{a}$ est une matrice de taille $(K \times 1)$ des paramètres des effets aléatoires,

$\boldsymbol{\varepsilon}$ est une matrice de taille $(n \times 1)$ des erreurs aléatoires résiduelles.

On suppose que :

- les vecteurs $\boldsymbol{a}$ et $\boldsymbol{\varepsilon}$ suivent une loi normale multivariée,
- l'espérance de ces vecteurs est nulle : $\mathbb{E}\left[\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{\varepsilon}\end{bmatrix}\right] = \begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{bmatrix}$,
- la variance-covariance de ces vecteurs est la suivante : $\mathrm{Var}\left[\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{\varepsilon}\end{bmatrix}\right] = \begin{bmatrix}\boldsymbol{G} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{R}\end{bmatrix}$.

Les sous-matrices $\boldsymbol{G}$ et $\boldsymbol{R}$ peuvent prendre différentes formes, auxquelles on référera comme la structure de covariance. Il est courant de poser $\boldsymbol{G} = \mathrm{diag}(\sigma_1^2 \boldsymbol{I}_{q_1} ... \sigma_K^2 \boldsymbol{I}_{q_K})$. Cela permet d'écrire : $\sum_{k=1}^{K} \sigma_k^2 \boldsymbol{Z}_k \boldsymbol{Z}_k' = \boldsymbol{ZGZ'}$ d'où $\boldsymbol{\Sigma} = \boldsymbol{ZGZ'} + \sigma_0^2 \boldsymbol{I}_n$. La partie $\sigma_0^2 \boldsymbol{I}_n$ de la variance est souvent notée $\boldsymbol{R}$.

1.2.2 Moments de $\boldsymbol{y}$

— De façon évidente, il vient : $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{X\beta}$.

— D'autre part : $\boldsymbol{\Sigma} := \text{Var}[\boldsymbol{y}] = \text{Var}[\boldsymbol{Za}] + \text{Var}[\boldsymbol{\varepsilon}] = \sum_{k=1}^{K} \sigma_k^2 \boldsymbol{Z}_k \boldsymbol{Z}_k' + \sigma_0^2 \boldsymbol{I}_n$.

Finalement, nous obtenons $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{\Sigma})$, les composantes de $\boldsymbol{y}$ n'étant pas indépendantes au sein d'un même niveau d'un facteur aléatoire donné.

1.2.3 Estimation par maximum de vraisemblance

L'expression que nous obtenons ici pour l'estimation de $\boldsymbol{\beta}$ fait intervenir l'estimation de la matrice de variances-covariances $\boldsymbol{\Sigma}$ de $\boldsymbol{y}$. Si elle est unique, la valeur correspondante dépend de la méthode d'estimation de $\boldsymbol{\Sigma}$. En fait, l'expression obtenue est aussi celle fournie par la méthode des moindres carrés généralisés notée $GLSE(\boldsymbol{\beta})$ (pour *Generalized Least Squares Estimation*) mais nous allons nous intéresser ici à la méthode du maximum de vraisemblance. La densité de $\boldsymbol{y}$ qui est la vraisemblance est définie par :

$$L(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}) \right).$$

La log-vraisemblance à maximiser est obtenue en prenant le logarithme de L :

$$l(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \frac{-n}{2} \log(2\pi) - \frac{1}{2} \log|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}).$$

En dérivant l par rapport à $\boldsymbol{\beta}$ nous obtenons le système de p équations :
$\frac{\partial l}{\partial \boldsymbol{\beta}} = -\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X\beta} + \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$ dont découlent les équations normales.
On remarque ensuite que $\frac{\partial \boldsymbol{\Sigma}}{\partial \sigma_k^2} = \boldsymbol{Z}_k \boldsymbol{Z}_k'$; on en déduit que
$\frac{\partial l}{\partial \sigma_k^2} = -\frac{1}{2} \text{ tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}_k\boldsymbol{Z}_k') + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta})'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}_k\boldsymbol{Z}_k'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta})$ pour $k = 0, 1, ..., K$.
En égalant à 0 les dérivées de l par rapport à $\boldsymbol{\beta}$ et à σ_k^2, $k = 0, 1, ..., K$, on peut résoudre les systèmes d'équations suivantes et obtenir les estimateurs du maximum de vraisemblance.

$$-\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X\beta} + \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y} = \boldsymbol{0}, \tag{1.2}$$

$$-\frac{1}{2}\text{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}_k\boldsymbol{Z}_k') + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta})'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}_k\boldsymbol{Z}_k'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}) = \boldsymbol{0}. \tag{1.3}$$

La solution du système 1.2 est : $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y} = GLSE(\boldsymbol{\beta})$.
Dans l'expression ci-dessus, $\boldsymbol{\Sigma} = \sum_{k=1}^{K} \sigma_k^2 \boldsymbol{Z}_k \boldsymbol{Z}_k' + \sigma_0^2 \boldsymbol{I}_n$; nous voyons donc que nous devons estimer les composantes de la variance avant de pouvoir estimer le vecteur $\boldsymbol{\beta}$.

Le système 1.3 donne : $\mathrm{tr}(\boldsymbol{\Sigma^{-1}} \boldsymbol{Z}_k \boldsymbol{Z}_k') = (\boldsymbol{y} - \boldsymbol{X\beta})' \boldsymbol{\Sigma^{-1}} \boldsymbol{Z}_k \boldsymbol{Z}_k' \boldsymbol{\Sigma^{-1}} (\boldsymbol{y} - \boldsymbol{X\beta})$,
où $k = 0, 1, ..., K$. Nous obtenons ainsi un système de $K+1+p$ équations non linéaires à $K+1+p$ inconnues que l'on résout par une méthode numérique itérative (de type *Fisher scoring*). Ces procédures numériques fournissent en plus, à la convergence, la matrice de variances-covariances asymptotiques des estimateurs.

1.2.4 Estimation par maximum de vraisemblance restreinte

Nous avons constaté, dans le point précédent, que les différentes équations obtenues lorsqu'on réalise l'estimation des paramètres par maximum de vraisemblance contiennent à la fois le vecteur $\boldsymbol{\beta}$ des paramètres liés à l'espérance de $\boldsymbol{y}$ et la matrice $\boldsymbol{\Sigma}$ des paramètres liés à la variance de $\boldsymbol{y}$. Dans un modèle mixte, c'est ce mélange de paramètres de natures différentes dans les mêmes équations qui engendre un biais systématique dans l'estimation par maximum de vraisemblance des composantes de la variance. L'objet de la méthode du maximum de vraisemblance restreint est précisément de séparer les deux types de paramètres. Le principe est le suivant : aux colonnes de la matrice $\boldsymbol{X}$ sont associés des vecteurs de l'espace vectoriel $F = \mathbb{R}^n$ (n est le nombre d'observations réalisées) ; on munit ce dernier de la métrique identité (associée à la matrice $\boldsymbol{I}_n$ sur la base canonique), ce qui en fait un espace euclidien. En notant $F_{\boldsymbol{X}}$ le sous-espace vectoriel de F engendré par les colonnes de $\boldsymbol{X}$ ($F_{\boldsymbol{X}}$ est supposé de dimension p), on sait que le projecteur orthogonal de F dans $F_{\boldsymbol{X}}$ a pour matrice associée :

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'.$$

Par ailleurs, le projecteur sur $F_{\boldsymbol{X}}^{\perp}$ (supplémentaire orthogonal à $F_{\boldsymbol{X}}$ dans F) est

$$\boldsymbol{H}^{\perp} = \boldsymbol{I}_n - \boldsymbol{H}.$$

Ainsi, en projetant $\boldsymbol{y}$ sur $F_{\boldsymbol{X}}^{\perp}$ et en travaillant avec cette projection (qui est, par définition, orthogonale à toute combinaison linéaire des colonnes de $\boldsymbol{X}$), on se libère de $\boldsymbol{\beta}$ dans l'estimation des composantes de la variance. Toutefois, $F_{\boldsymbol{X}}^{\perp}$ étant de dimension $m = n-p$, le vecteur aléatoire projeté de $\boldsymbol{y}$ sur $F_{\boldsymbol{X}}^{\perp}$ est multinormal d'ordre m. C'est un vecteur aléatoire écrit dans $\mathbb{R}^n$ dont la matrice des variances-covariances est singulière et qu'on doit transformer pour obtenir un vecteur aléatoire directement écrit dans $\mathbb{R}^m$. Soit donc $\boldsymbol{M}_0$ une matrice $m \times n$, de rang m, réalisant cette transformation et soit $\boldsymbol{M} = \boldsymbol{M}_0\boldsymbol{H}^{\perp}$. On considère finalement $\boldsymbol{y}^* = \boldsymbol{M}\boldsymbol{y}$; on a ainsi $\boldsymbol{y}^* \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma}^*)$, avec

$$\boldsymbol{\mu} = \boldsymbol{M}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{M}_0(\boldsymbol{H}^{\perp}\boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0} \text{ (par définition de } \boldsymbol{H}^{\perp}),$$
$$\text{et } \boldsymbol{\Sigma}^* = \boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}' = \boldsymbol{M}_0\boldsymbol{H}^{\perp}\boldsymbol{\Sigma}\boldsymbol{H}^{\perp}\boldsymbol{M}_0'.$$

On peut encore écrire :

$$\boldsymbol{\Sigma}^* = \sum_{k=1}^{K} \sigma_k^2 \boldsymbol{M}\boldsymbol{Z}_k\boldsymbol{Z}_k'\boldsymbol{M}' = \sum_{k=1}^{K} \sigma_k^2 \boldsymbol{Z}_k^*\boldsymbol{Z}_k^{*'} \text{ en posant } \boldsymbol{Z}_k^* = \boldsymbol{M}\boldsymbol{Z}_k.$$

En réécrivant les $K+1$ dernières équations de vraisemblance relatives au vecteur aléatoire $\boldsymbol{y}^*$, il vient maintenant :

$$\mathrm{tr}(\boldsymbol{\Sigma}^{*-1}\boldsymbol{Z}_k^*\boldsymbol{Z}_k^{*'}) = \boldsymbol{y}^{*'}\boldsymbol{\Sigma}^{*-1}\boldsymbol{Z}_k^*\boldsymbol{Z}_k^{*'}\boldsymbol{\Sigma}^{*-1}\boldsymbol{y}^{*'}, \;\; k = 0, 1, ..., K \text{ (avec } \boldsymbol{y}^* = \boldsymbol{M}\boldsymbol{y}).$$

Il s'agit d'un système de $K+1$ équations non linéaires à $K+1$ inconnues (les composantes σ_k^2 de la variance) dans lequel ne figure plus le vecteur $\boldsymbol{\beta}$. Ce système n'admet pas de solution analytique et nécessite une procédure numérique itérative pour sa résolution.

Là encore, la procédure itérative fournit, à la convergence, la matrice des variances-covariances asymptotiques des estimateurs.

Les équations ci-dessus proviennent en fait de l'annulation des dérivées partielles par rapport aux σ_k^2 de la log-vraisemblance de $\boldsymbol{y}$, autrement dit de la log-vraisemblance de la projection de $\boldsymbol{y}$ sur $F_{\boldsymbol{X}}^{\perp}$ ou encore de la restriction de la vraisemblance de $\boldsymbol{y}$ à ce sous-espace. Les estimateurs obtenus par maximisation de cette restriction de la vraisemblance sont, pour cette raison, appelés estimateurs du maximum de vraisemblance restreinte.

1.2.5 Estimation par *MINQUE* et *MIVQUE*

Dans un modèle mixte, il est possible d'estimer les composantes de la variance en utilisant soit la méthode *MINQUE (Minimum Norm Quadratic Unbiased Estimation)* [42], soit la méthode *MIVQUE (Minimum Variance Quadratic Unbiased Estimation)*[62]. Lorsque les variables sont normalement distribuées, *MIVQUE* coïncide avec *MINQUE* sous la norme euclidienne d'une matrice [43]. Le principe général d'estimation reste le même que celui exposé précédemment. Ces méthodes sont peu utilisées dans la pratique, car la méthode du maximum de vraisemblance, notamment restreinte, s'est avérée rapidement comme un passage obligé et une référence dans l'inférence des composantes de la variance en modèle linéaire mixte, au point qu'elle a supplanté en pratique ces estimateurs quadratiques en raison de la faisabilité numérique grâce au développement simultané des ordinateurs et d'algorithmes de calcul performants [14].

1.2.6 Espérance, variance et loi des estimateurs

Supposons que $\boldsymbol{\Sigma}$ est connue (σ_0^2 et σ_k^2 sont connues). Dans ce cas, l'estimateur de $\boldsymbol{\beta}$ est la solution du maximum de vraisemblance $\widehat{\boldsymbol{\beta}}$. On a :

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}.$$

L'espérance de $\widehat{\boldsymbol{\beta}}$ est :

$$\begin{aligned}
\mathbb{E}[\widehat{\boldsymbol{\beta}}] &= \mathrm{E}[(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}] \\
&= (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\mathbb{E}[\boldsymbol{y}] \\
&= (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\boldsymbol{\beta} \\
&= \boldsymbol{\beta}.
\end{aligned}$$

La variance de $\widehat{\boldsymbol{\beta}}$ est :

$$\begin{aligned}
\mathrm{Var}[\widehat{\boldsymbol{\beta}}] &= \mathrm{Var}[(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}] \\
&= (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\mathrm{Var}[\boldsymbol{y}]((\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}))' \\
&= (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}.
\end{aligned}$$

La loi de $\widehat{\boldsymbol{\beta}}$ est : $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \boldsymbol{(X'\Sigma^{-1}X)^{-1}})$.

Supposons que $\boldsymbol{\Sigma}$ n'est pas connue. Pour σ_0^2 et σ_k^2, leur estimateur n'a pas de forme explicite, il est beaucoup plus compliqué de trouver leur espérance. Par contre, il est possible d'utiliser les résultats sur la loi asymptotique des estimateurs du maximum de vraisemblance :

$$\begin{bmatrix}\widehat{\sigma}_0^2 \\ \widehat{\sigma}_k^2\end{bmatrix} \sim N\left(\begin{bmatrix}\sigma_0^2 \\ \sigma_k^2\end{bmatrix}, \left(-\mathbb{E}\begin{bmatrix}l_{\sigma_0^2\sigma_0^2} & l_{\sigma_0^2\sigma_k^2} \\ l_{\sigma_0^2\sigma_k^2} & l_{\sigma_k^2\sigma_k^2}\end{bmatrix}\right)^{-1}\right),$$

où l est la log-vraisemblance et $l_{\sigma_0^2\sigma_0^2} = \frac{\partial l}{\partial \sigma_0^2 \sigma_0^2}$, $l_{\sigma_0^2\sigma_k^2} = \frac{\partial l}{\partial \sigma_0^2 \sigma_k^2}$, $l_{\sigma_k^2\sigma_k^2} = \frac{\partial l}{\partial \sigma_k^2 \sigma_k^2}$.

1.3 Tests d'hypothèses

Dans cette section, nous allons présenter les tests de significativité des effets fixes ($\boldsymbol{\beta}$) et ceux des effets aléatoires (les composantes de la variance $\boldsymbol{\Sigma}$) pour le modèle mixte.

1.3.1 Tests sur les effets fixes

Pour tester la significativité des effets fixes, nous allons rappeler d'abord quelques propriétés de l'estimateur $\widehat{\boldsymbol{\beta}}$, ensuite, nous allons présenter des tests de Wald et du rapport de vraisemblance.

a. Propriétés

(i) L'estimateur $\widehat{\boldsymbol{\beta}}$ du maximum de vraisemblance est sans biais pour $\boldsymbol{\beta}$:

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}.$$

Soit $\boldsymbol{k}'$ une matrice $(r \times p)$ avec $r < p$ dont les r lignes sont linéairement indépendantes.

(ii) Sous l'hypothèse d'indépendance et de variances homogènes ($\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$), $\boldsymbol{k}'\widehat{\boldsymbol{\beta}}$ est le meilleur estimateur linéaire sans biais de $\boldsymbol{k}'\boldsymbol{\beta}$:

$$\mathbb{E}[\boldsymbol{k}'\widehat{\boldsymbol{\beta}}] = \boldsymbol{k}'\boldsymbol{\beta} \text{ et } \mathrm{Var}(\boldsymbol{k}'\widehat{\boldsymbol{\beta}}) = \sigma^2 \boldsymbol{k}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{k}.$$

En plus, sous l'hypothèse de normalité des erreurs, la distribution de l'estimateur $\boldsymbol{k}'\widehat{\boldsymbol{\beta}}$ est aussi normale :

$$\boldsymbol{k}'\widehat{\boldsymbol{\beta}} \sim N\Big(\boldsymbol{k}'\boldsymbol{\beta}, \sigma^2 \boldsymbol{k}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{k}\Big).$$

b. Test de Wald

On considère le test de l'hypothèse nulle : $H_0 : \boldsymbol{k}'\boldsymbol{\beta} = \boldsymbol{m}$ contre son alternative contraire $H_1 : \boldsymbol{k}'\boldsymbol{\beta} \neq \boldsymbol{m}$ où $\boldsymbol{k}'$ est une matrice $(r \times p)$ avec $r < p$ dont les r lignes sont linéairement indépendantes, et $\boldsymbol{m}$ est un vecteur $(r \times 1)$ de constantes, souvent nulles.

- Si $\boldsymbol{\Sigma}$ est connue ; sous H_0, la statistique
  $\boldsymbol{W} = (\boldsymbol{k}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m})'(\mathrm{Var}(\boldsymbol{k}'\widehat{\boldsymbol{\beta}}))^{-1}(\boldsymbol{k}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m}) = (\boldsymbol{k}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m})'[\boldsymbol{k}'(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{k}]^{-1}(\boldsymbol{k}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m})$
  qui est celle de Wald relative au test, suit une loi du khi-deux ($\chi^2_{rang(\boldsymbol{k})}$) à $rang(\boldsymbol{k})$ degrés de liberté, avec $\mathrm{Var}(\boldsymbol{k}'\widehat{\boldsymbol{\beta}}) = \boldsymbol{k}'(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{k}$.

- Si $\boldsymbol{\Sigma}$ est connue à un coefficient près ($\boldsymbol{\Sigma} = \sigma^2\boldsymbol{\Sigma}_0$ avec $\boldsymbol{\Sigma}_0$ une matrice quelconque, mais connue, et σ^2 un paramètre inconnu),

  $$\begin{aligned} \mathrm{Var}(\boldsymbol{k}'\widehat{\boldsymbol{\beta}}) &= \boldsymbol{k}'(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{k} \\ &= \frac{1}{\sigma^2}\boldsymbol{k}'(\boldsymbol{X}'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{X})^{-1}\boldsymbol{k}. \end{aligned}$$

  Donc,

  $$\begin{aligned} \boldsymbol{W} &= (\boldsymbol{k}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m})'(\mathrm{Var}(\boldsymbol{k}'\widehat{\boldsymbol{\beta}}))^{-1}(\boldsymbol{k}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m}) \\ &= \frac{1}{\sigma^2}(\boldsymbol{k}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m})'(\boldsymbol{X}'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{X})^{-1}(\boldsymbol{k}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m}) \\ &= \frac{1}{\sigma^2}\boldsymbol{Q}, \end{aligned}$$

  où $\boldsymbol{Q} = (\boldsymbol{k}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m})'(\boldsymbol{X}'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{X})^{-1}(\boldsymbol{k}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m})$.

  Sous H_0, la statistique $\boldsymbol{Q} = \sigma^2\boldsymbol{W}$ suit une loi $\sigma^2\chi^2_{rang(\boldsymbol{k})}$. Or, on ne connait pas σ^2.

  Par contre, $\frac{(n-rang(\boldsymbol{X}))\hat{\sigma}^2}{\sigma^2} = \frac{\boldsymbol{SSE}}{\sigma^2} \sim \chi^2_{n-rang(\boldsymbol{X})}$ où n est le nombre total d'observations et $\boldsymbol{SSE}$ (*Sum of Squares of Errors*) $= (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$. On peut donc former la statistique de Fisher qui ne dépend pas de la variance σ^2 [53] :

  $$\begin{aligned} F &= \frac{[\boldsymbol{Q}/\sigma^2]/rang(\boldsymbol{k})}{[\boldsymbol{SSE}/\sigma^2]/[n - rang(\boldsymbol{X})]} \\ &= \frac{\boldsymbol{Q}/rang(\boldsymbol{k})}{[\boldsymbol{SSE}]/[n - rang(\boldsymbol{X})]} \\ &= \frac{(\boldsymbol{k}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m})'(\boldsymbol{X}'\boldsymbol{\Sigma}_0^{-1}\boldsymbol{X})^{-1}(\boldsymbol{k}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m})[n - rang(\boldsymbol{X})]}{(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})rang(\boldsymbol{k})} \sim F_{rang(\boldsymbol{k}),\ n-rang(\boldsymbol{X})}. \end{aligned}$$

  L'hypothèse nulle est rejetée au niveau de signification α lorsque $F > F^{(1-\alpha)}_{rang(\boldsymbol{k}),\ n-rang(\boldsymbol{X})}$.

- Si $\boldsymbol{\Sigma}$ est inconnue, on fait usage des résultats asymptotiques. Les tests suivants ne sont alors pas exacts :

  La matrice $\boldsymbol{\Sigma}$ est estimée ici par maximum de vraisemblance $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\widehat{\boldsymbol{\gamma}})$ où $\boldsymbol{\gamma}$ est le vecteur des composantes de la variance. $\widehat{\boldsymbol{\beta}}$ est alors solution de l'équation

  $$(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}.$$

Sous H_0, la statistique du test de Wald

$$\boldsymbol{W}(\widehat{\boldsymbol{\gamma}}) = (\boldsymbol{k}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m})'[\boldsymbol{k}'(\boldsymbol{X}'\widehat{\boldsymbol{\Sigma}}^{-\mathbf{1}}\boldsymbol{X})^{-1}\boldsymbol{k}]^{-1}(\boldsymbol{k}'\widehat{\boldsymbol{\beta}} - \boldsymbol{m}),$$

quand $n \longrightarrow +\infty$, tend vers une loi du khi-deux à $rang(\boldsymbol{k})$ degrés de liberté, $(\chi^2_{rang(\boldsymbol{k})})$. De manière analogue au cas précédent, on peut utiliser le test de type F de Fisher dont, asymptotiquement sous H_0, la statistique est [15] :

$$F = \frac{\boldsymbol{W}(\widehat{\boldsymbol{\gamma}})}{rang(\boldsymbol{k})} \sim F_{rang(\boldsymbol{k}),\ n-rang(\boldsymbol{X})}.$$

Ici, on forme $\frac{\boldsymbol{W}(\widehat{\boldsymbol{\gamma}})}{rang(\boldsymbol{k})}$ qu'on compare à un $F_{rang(\boldsymbol{k}),\ d}$ avec un nombre de degrés de liberté d qui est calculé selon une méthode approchée comme celle de Satterthwaite [50].

Dans le cas du modèle linéaire à effets fixes classique, on n'estime que σ_e^2. On a alors :

$$d = n - rang(\boldsymbol{X}).$$

Dans le cas du modèle linéaire mixte, on doit estimer σ_a^2 et σ_e^2.

c. Test du rapport de vraisemblance

Sous l'hypothèse H_0, la statistique

$$\lambda = -2\boldsymbol{l}_R + 2\boldsymbol{l}_C,$$

suit asymptotiquement une loi du khi-deux à r degrés de liberté (χ^2_r) où sous H_0, $\boldsymbol{l}_R$ est le maximum du logarithme de la vraisemblance sous le modèle réduit, $\boldsymbol{l}_C$ est le maximum du logarithme de la vraisemblance sous le modèle complet, sous H_1. Le degré de liberté r est la différence de paramètres à estimer sous H_0 et sous H_1. Les paramètres du modèle sont estimés par maximum de vraisemblance. Il est important de souligner que les deux modèles heurtés vis-à-vis des effets fixes doivent présenter la même structure de variance-covariance. Pour comparer deux modèles emboîtés avec ce test, on compare les log-vraisemblances des deux modèles.

Si la différence entre ces log-vraisemblances est grande, le fait de passer d'un modèle simple (modèle sous H_0) à un modèle plus complexe (modèle sous H_1) apporte un écart significatif, le modèle sous H_1 est donc acceptable. Par contre, si l'écart est faible, cela veut dire que les deux modèles sont voisins, on conserve le modèle sous H_0.

1.3.2 Tests sur les composantes de la variance

Pour effectuer les tests de significativité des composantes de la variance, nous allons présenter d'abord une règle générale dans le cadre des modèles emboîtés, ensuite, nous allons présenter le test de nullité de la variance σ^2 et le test robuste.

a. Règle générale

Lorsque les modèles sont emboîtés, le test couramment utilisé est celui du rapport de vraisemblance. Si on veut tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$, on calcule la statistique :

$$\lambda_n = -2l_R + 2l_C,$$

où l_R est la log-vraisemblance sous le modèle réduit (sous H_0) et l_C est la log-vraisemblance sous le modèle complet (sous H_1). Sous l'hypothèse H_0, la statistique $\lambda_n \sim \chi^2_r$ où $r = dl_C - dl_R$ degrés de libertés, soit la différence entre les nombres de paramètres sous les deux modèles. La distribution asymptotique de λ_n repose sur la normalité asymptotique des estimateurs du maximum de vraisemblance. Si θ_0 est situé sur le bord de l'espace des paramètres, alors la loi asymptotique de $\hat{\theta}$ n'est plus une gaussienne, et donc la loi de λ_n n'est plus une khi-deux.

b. Test de nullité de σ^2

Lorsque l'on teste $H_0 : \sigma^2 = 0$ contre $H_1 : \sigma^2 > 0$, 0 est à la limite de l'espace des paramètres, car une variance σ^2 est définie sur $[0; +\infty)$.

Dans le cas d'un modèle linéaire mixte à un facteur à effets aléatoires autre que la résiduelle, si on veut tester la nullité de la variance de ce facteur, la statistique λ_n du rapport de vraisemblance suit une loi qui est un mélange de khi-deux : $\frac{1}{2}\chi^2_0+\frac{1}{2}\chi^2_1$, où χ^2_0 est en effet la mesure de Dirac en 0 [35]. La procédure revient concrètement à effectuer un test unilatéral. Plus généralement, lorsque le modèle comporte q facteurs à effets aléatoires autres que la résiduelle, et que l'on souhaite tester la nullité d'une des q variances, la distribution sous l'hypothèse nulle de la statistique du rapport de vraisemblance suit un mélange de lois : $\frac{1}{2}\chi^2_{q-1} + \frac{1}{2}\chi^2_q$ [16].

c. Test robuste

Ce test est utile lorsque l'on ne connaît pas la structure de variance-covariance des observations. Le choix d'un modèle global requiert à la fois le choix des effets fixes : modèle de l'espérance $\mathbb{E}$ et celui des effets aléatoires : modèle de variance-covariance $\mathbf{\Sigma}$. Ce choix semble être complexe, car la comparaison de modèles sur l'espérance $\mathbb{E}$ dépend de la structure de variance-covariance $\mathbf{\Sigma}$ et celle de modèles $\mathbf{\Sigma}$ dépend de $\mathbb{E}$. On peut procéder de la manière suivante : (i) On part d'une structure de variance$-$covariance $\mathbf{\Sigma}$ donnée, on choisit un modèle sur l'espérance $\mathbb{E}$, (ii) puis à modèle d'espérance E fixé, on compare différentes structures de variance$-$covariance $\mathbf{\Sigma}$. On itère ensuite (i) et (ii).

On peut aussi inférer sur les effets fixes par des estimateurs "sandwich" [30, 7, 17] vis-à-vis de la structure $\mathbf{\Sigma}$ dont à la base, il faut utiliser les propriétés de l'estimateur du maximum de vraisemblance $(\boldsymbol{X}'\mathbf{\Sigma}^{-1}\boldsymbol{X})\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\mathbf{\Sigma}^{-1}\boldsymbol{y}$ sans avoir fait l'hypothèse que la matrice de variance-covariance $\mathbf{\Sigma}$ est spécifiée.

1.4 Sélection de modèles linéaires mixtes

Pour comparer des modèles mixtes n'ayant pas les mêmes effets fixes, ni la même structure de variance-covariance, nous allons définir des critères pour choisir le meilleur modèle.

Ces critères sont basés sur une estimation de la distance entre chacun des modèles candidats et le vrai modèle (inconnu), puis sur le choix de celui qui minimise cette distance, sachant qu'une mesure de ressemblance entre deux modèles est donnée par l'information de Kullback-Leibler faisant intervenir des logarithmes des densités [27].

Définition 1. Le critère d'Akaike (AIC)

Il se calcule par :

$$AIC = -2l + 2k,$$

où l est la log-vraisemblance maximisée et k est le nombre de paramètres dans le modèle. La déviance du modèle ($-2l$) est pénalisée par 2 fois le nombre de paramètres.

Le premier terme du critère s'interprète comme une mesure de l'inadéquation du modèle aux données, tandis que le second représente une fonction de pénalité mesurant la complexité du modèle à travers son nombre de paramètres. Ce critère permet de comparer des modèles non nécessairement emboîtés dont les paramètres sont estimés par maximum de vraisemblance [1]. Le meilleur modèle est celui qui a le plus faible AIC [1, 2].

Définition 2. Le critère de Schwarz (BIC)

C'est un critère basé sur un raisonnement bayésien [52]. Il s'obtient par :

$$BIC = -2l + k\log(n),$$

où l est la log-vraisemblance maximisée, k est le nombre de paramètres libres du modèle et n est le nombre d'observations dans l'échantillon. Le modèle qui sera sélectionné est celui qui minimise le critère BIC, soit $M_{BIC} = \underset{M}{\operatorname{argmin}}\{BIC(M)\}$ où M désigne le modèle.

1.5 Discussion sur le modèle linéaire mixte

Lorsqu'on cherche un bon modèle pour ajuster les données, il faut trouver à la fois un bon modèle pour l'espérance : $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{X\beta}$ et pour la variance : $\mathrm{Var}[\boldsymbol{y}] = \boldsymbol{\Sigma} = \boldsymbol{ZGZ'} + \boldsymbol{R}$.

La procédure suivante peut être itérée :

(i) On se fixe la structure de variance, et on choisit un bon modèle pour l'espérance : tests des effets fixes ou critères de choix de modèles.

(ii) On se fixe la structure d'espérance, et on choisit un bon modèle pour la variance : tests des effets aléatoires ou critères de choix de modèles.

Cependant, les modèles linéaires mixtes, quand les données sont hiérarchiques, prennent une forme particulière, en raison de l'hypothèse d'échangeabilité des observations au sein d'une même grappe. Cette forme particulière prend le nom du modèle multiniveaux.

Dans le chapitre suivant, nous allons décrire le modèle multiniveaux en présentant dans un premier temps le modèle à deux niveaux avec un *outcome* univarié et ensuite le modèle à deux niveaux avec un *outcome* bivarié.

CHAPITRE 2

Le modèle linéaire multiniveaux

2.1 Présentation

En épidémiologie sociale, les structures de données imbriquées sont très courantes [9]. Les données imbriquées (qui peuvent donner lieu à des observations corrélées) se produisent chaque fois que les participants sont regroupés, comme c'est souvent le cas dans la recherche en sciences sociales. Par exemple, les réponses des individus pour les résidents regroupés dans des communautés d'appartenance dépendront dans une certaine mesure de la dynamique du groupe, ce qui entraîne une dépendance au sein du groupe d'appartenance [26]. Pourtant, une hypothèse inférentielle clé de toutes les techniques statistiques utilisées dans les sciences sociales et traitées au premier chapitre est que les observations sont indépendantes. Kenny et Judd en 1986 [25] avait noté que si la dépendance est généralement traitée comme une nuisance, il y a encore "de nombreuses occasions où la dépendance est le problème de fond que nous essayons de comprendre dans la recherche psychologique". L'interaction sociale implique donc par définition la dépendance.

Si un chercheur souhaite étudier l'interaction sociale, ou le comportement qui se produit dans divers cadres ou contextes, la dépendance inhérente n'est pas tant un problème statistique à résoudre qu'un centre d'intérêt. La figure 2.1 fournit une représentation générale d'un plan imbriqué à deux niveaux, où une telle dépendance est souvent présente. Le niveau inférieur ou niveau 1 comprend les participants, chacun d'entre eux étant membre ou appartenant à une seule grappe (niveau 2). Dans chaque grappe se trouvent n participants (qui peuvent varier d'une grappe à l'autre), et N désigne le nombre de grappes dans le plan. Des exemples de telles conceptions à deux niveaux incluent les individus de l'étude *CentrÉS* au sein des communautés locales de Sherbrooke.

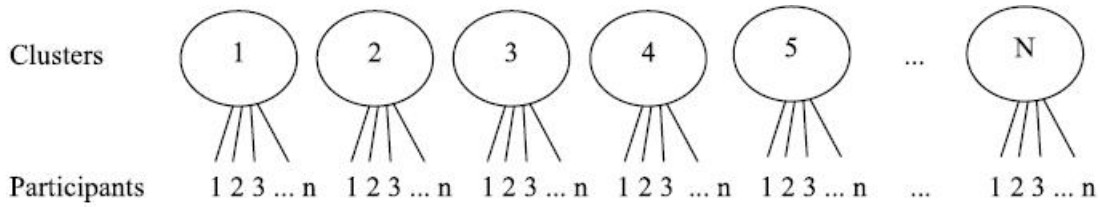

FIGURE 2.1 – Représentation générale d'un plan d'étude à deux niveaux avec des participants imbriqués dans des grappes

Un tel regroupement n'implique pas toujours deux niveaux seulement. Un modèle à quatre niveaux rencontré dans la recherche en santé a impliqué des individus (niveau 1) imbriqués dans des ménages (niveau 2), imbriqués dans les secteurs locaux (niveau 3) regroupés dans les régions sociosanitaires (niveau 4) [36]. Lorsque les données sont regroupées de cette manière, l'usage du modèle multiniveaux est nécessaire pour fournir une inférence statistique précise. Bien que les données imbriquées puissent suggérer l'usage d'un tel modèle, un élément clé à prendre en compte est la nature de la variable de regroupement. Si les participants sont imbriqués dans ce qui est considéré comme un facteur fixe, l'analyse standard peut être appropriée. Un facteur fixe comme le genre peut être considéré comme un facteur pour lequel, si une étude de réplication devait être réalisée, les mêmes niveaux de ce facteur seraient inclus dans l'étude.

Un exemple classique, traité également au chapitre 1, est celui d'un plan d'expérience à deux ou plusieurs groupes, où l'analyse standard de la variance peut être utilisée. Dans ce cas, l'appartenance au groupe est presque toujours considérée comme un facteur fixe, de sorte que si l'étude était reproduite, les mêmes niveaux de ce facteur seraient mis en œuvre. Il est à noter que les groupes d'appartenance ne sont pas considérés comme un échantillon d'une population de conditions qui auraient pu être incluses dans l'étude. Au contraire, toutes les conditions d'intérêt sont incluses dans l'étude. Toutefois, il faut tenir compte de la nature des participants inclus dans le plan d'expérience standard. Si une étude de réplication était menée, un ensemble différent de participants serait probablement inclus dans l'étude. En d'autres termes, les participants sont censés représenter un échantillon d'une population d'intérêt plus large. Les participants sont donc considérés comme un facteur aléatoire. Dans le plan expérimental standard, un facteur aléatoire et un ou plusieurs facteurs fixes d'intérêt (le genre, le niveau d'éducation...) sont inclus.

La clé pour reconnaître la nécessité d'une analyse multiniveaux est qu'une telle analyse est souvent nécessaire lorsque deux ou plusieurs facteurs de la conception sont considérés comme des facteurs aléatoires. Par exemple, avec les jeunes adultes imbriqués dans des quartiers de résidence, les jeunes et les quartiers sont susceptibles d'être considérés comme des facteurs aléatoires, de sorte que si l'étude est reproduite, un ensemble différent de jeunes adultes et de quartiers de résidence serait inclus (ou échantillonné) dans l'étude. En d'autres termes, les jeunes et les quartiers peuvent être considérés comme représentant une population plus large de jeunes et de quartiers, respectivement. Ainsi, c'est l'inclusion de multiples facteurs aléatoires dans un plan de recherche qui signale la nécessité d'une analyse multiniveaux.

2.2 Problèmes liés à l'usage de l'analyse à un seul niveau de données hiérarchiques

Des techniques d'estimation sophistiquées, développées à la fin des années 1970, ont conduit à la création de logiciels ayant facilité l'utilisation du modèle multiniveaux[6, 45]. Avant cette période, les chercheurs utilisaient généralement des modèles de régression à un seul niveau pour examiner les relations entre les variables à différents niveaux, malgré la violation attendue de l'hypothèse d'indépendance. Cette inadéquation entre les caractéristiques de la conception et le modèle d'analyse peut être problématique pour diverses raisons. On suppose qu'un chercheur s'intéresse à la relation entre la santé mentale perçue des participants d'une étude et les caractéristiques des quartiers dans lesquels ils résident. La conception d'une telle étude peut impliquer une sélection aléatoire de quartiers, suivie d'une sélection aléatoire de participants dans chacun des quartiers. Les quartiers et les participants seraient considérés comme des facteurs aléatoires, car chacun d'eux représente un échantillon d'une population plus large, ce qui suggère la nécessité d'une analyse multiniveaux. Lors de l'étude de la question d'intérêt, un chercheur qui choisirait d'ignorer la dépendance dans les données aurait deux choix analytiques en utilisant la modélisation à un seul niveau.

Le chercheur pourrait agréger les données sur les scores de santé mentale perçue au niveau du quartier et utiliser les données résultantes à ce niveau dans une analyse de régression à un seul niveau. Dans ce cas, le résultat serait le score moyen de santé mentale perçue du participant du quartier. Les prédicteurs seraient constitués de descripteurs du quartier. L'un des principaux problèmes d'une telle analyse est la perte d'informations concernant la variabilité des résultats des participants au sein des quartiers, la diminution de la puissance statistique et la compromission de la validité écologique des inférences [26].

Le chercheur pourrait désagréger les données au niveau du participant et du quartier.

Cette forme de modélisation, généralement peu souhaitable, impliquerait d'utiliser les participants de l'étude comme unité d'analyse et d'ignorer la dépendance des scores des participants au sein de chaque quartier. Dans la régression à un seul niveau qui serait utilisée avec des données désagrégées, la variable de réponse serait le score de santé mentale perçue, avec des prédicteurs comprenant les caractéristiques du participant et du quartier. Le problème de cette analyse est que la variation entre les participants et les quartiers est, très probablement, combinée incorrectement dans le terme résiduel du modèle et ne reflète pas correctement la variation au niveau des participants et des quartiers. Une conséquence importante ici est que les erreurs standards associées à l'estimation des effets des prédicteurs peuvent être considérablement sous-estimées. Ces erreurs types mal estimées conduisent alors à des taux d'erreur de type I gonflés et à de mauvaises estimations de l'intervalle de confiance. Le problème s'accentue lorsqu'il y a une plus grande dépendance entre les observations.

On utilise couramment une mesure de degré de dépendance entre les individus appelée corrélation intra-classe (ICC) lorsque des variables explicatives sont incluses dans le modèle. Plus les caractéristiques du contexte sont liées au résultat individuel d'intérêt, plus l'ICC est élevé et plus, on a besoin d'une analyse multiniveaux. Pour les données à deux niveaux, l'ICC, lorsqu'il est positif, ce qui est généralement le cas, peut être interprété comme la proportion de la variance totale du résultat qui se produit entre les regroupements (par opposition à l'intérieur des regroupements). Hedges et Hedberg en 2007 ont noté que les ICC sont généralement compris entre 0.05 et 0.20 [21], même si des valeurs plus petites ou plus grandes peuvent être obtenues. Cependant, même un ICC légèrement supérieur à zéro peut avoir un effet sur les taux d'erreurs de type I [51].

Les chercheurs n'ont pas à choisir entre la perte d'information liée à l'agrégation des données dépendantes et les taux d'erreur de type I élevés liés aux données désagrégées. Au lieu de choisir un seul niveau pour des analyses de données groupées ou hiérarchiques, ils peuvent utiliser la technique appelée modélisation multiniveaux ou hiérarchique.

2.3 Définitions

Des définitions du point de vue conceptuel, méthodologique et statistique ont été proposées par plusieurs chercheurs, notamment Gauvin et Dassa en 2004 [18].
Du point de vue conceptuel, un modèle multiniveaux désigne un ensemble de prédictions ou d'explications qui s'étendent sur plusieurs unités d'analyse de systèmes vivants.
Du point de vue méthodologique, un modèle multiniveaux renvoie à la mesure ou à la manipulation de variables situées à plusieurs niveaux d'analyse de systèmes vivants.
Du point de vue statistique, cette analyse réfère à un ensemble de techniques statistiques qui s'inscrit dans le cadre de la généralisation du *GLM* et qui permet le traitement de données structurées hiérarchiquement, c'est-à-dire qui proviennent de plusieurs unités d'analyse [45, 57] permettant d'estimer les sources de variance intra-unité et inter-unités à l'aide de la corrélation intra-classe, de déterminer la présence d'effets aléatoires et de quantifier les effets fixes. Ils sont des extensions des modèles linéaires et un cas particulier du modèle linéaire mixte.

2.4 Formulation du modèle multiniveaux

Les théories et principes du modèle linéaire mixte présentées tout au long du chapitre 1 de ce mémoire sont les mêmes pour le modèle multiniveaux. C'est en fait la partie aléatoire qui le distingue du modèle classique. Il permet de prendre en compte la corrélation et l'hétéroscédasticité induite par la structure hiérarchique des données.
En effet, il existe deux façons courantes d'afficher les modèles d'analyse pour le modèle multiniveaux. Ces modèles peuvent être exprimés comme un ensemble d'équations à chaque niveau séparément, ou les équations de chaque niveau peuvent être combinées pour fournir une seule expression.

La formulation à niveaux multiples est souvent plus facile à comprendre, surtout lorsqu'on apprend à connaître le modèle, mais l'équation combinée présente également des avantages. Avec l'expression multiniveaux, le modèle de niveau 1 ou de niveau inférieur contient des variables mesurées au niveau micro (par exemple, au niveau du participant d'une étude) tandis que le modèle de niveau 2 ou de niveau supérieur contient des variables au niveau macro (par exemple, au niveau du quartier de résidence du participant). Nous utiliserons à la fois les équations multiniveaux et combinées.

Avant de présenter la formulation à deux niveaux, nous allons expliquer une certaine terminologie. Raudenbush et Bryk en 2002 [45] ont fait la distinction entre les modèles inconditionnels et conditionnels. Un modèle inconditionnel encore appelé modèle vide est un modèle dans lequel aucun prédicteur (à aucun des niveaux) n'est inclus. Un modèle conditionnel inclut au moins un prédicteur à n'importe quel niveau. Un modèle couramment utilisé dans une telle analyse est un modèle inconditionnel au niveau 1 et conditionnel au niveau 2. Nous verrons aussi qu'un tel modèle, lorsqu'il comporte des *intercpet* et pentes variables, est appelé modèle à coefficients aléatoires.
Les résultats obtenus à partir d'un logiciel de modélisation multiniveaux sont souvent séparés en ce que l'on appelle les effets fixes et aléatoires, qui sont liés aux facteurs fixes et aléatoires décrits précédemment. En bref, les effets d'un facteur aléatoire sont résumés par des variances et covariances, tandis que l'effet d'un facteur fixe est résumé (comme dans la régression à un seul niveau) par un coefficient de régression. Nous verrons également qu'il est possible qu'un facteur fixe ait à la fois des effets fixes et aléatoires.

2.4.1 Modèle à deux niveaux avec un *outcome* univarié

Pour formuler le modèle à deux niveaux avec un *outcome* univarié, on adopte une stratégie usuelle de complexification croissante du modèle selon les étapes suivantes :

a. Modèle inconditionnel ou modèle vide

C'est le modèle à deux niveaux le plus simple possible. Il est sans variable explicative et correspond à une analyse de variance à un facteur avec effet aléatoire. Il fournit la décomposition initiale de la variance. Nous avons les formulations suivantes :

$$\text{Niveau 1} : y_{ij} = \beta_{0j} + e_{ij},$$
$$\text{Niveau 2} : \beta_{0j} = \beta_0 + u_{0j},$$
$$\text{soit } y_{ij} = \underbrace{\beta_0}_{Partie\ fixe} + \underbrace{(u_{0j} + e_{ij})}_{Partie\ aléatoire}, \text{ où}$$

- y_{ij} représente la variable dépendante observée chez un individu i dans le groupe j,
- β_0 est la partie fixe des ordonnées pour chaque groupe qui représente l'ordonnée moyenne pour tous les groupes et correspond à la moyenne théorique générale,
- β_{0j} est l'ordonnée à l'origine dans la population des individus pour le groupe j et peut s'interpréter comme la valeur moyenne du groupe j,
- e_{ij} est le terme d'erreur correspondant à l'individu i dans le groupe j,
- u_{0j} est le terme aléatoire traduisant la variabilité de l'ordonnée entre les groupes. Elle représente l'écart de la moyenne de chaque groupe à la moyenne générale. Cette variabilité est mesurée au niveau 2 et est la même pour tous les individus du groupe j.

Chaque donnée individuelle s'éloigne de la moyenne du groupe de la quantité e_{ij} et chaque moyenne du groupe j s'éloigne de la moyenne générale de la quantité u_{0j}.

Suivant le chapitre 1, les hypothèses suivantes sont admises :

- Les e_{ij} sont les réalisations *iid* d'une variable aléatoire de loi $N(0, \sigma_e^2)$ où σ_e^2 traduit la variabilité des individus au sein d'un même groupe. Cette variabilité dite *intra-groupe* est supposée identique au sein de chaque groupe.
- Les u_{0j} sont des réalisations *iid* de loi $N(0, \sigma_{u_0}^2)$ indépendantes des e_{ij} .

D'après les démonstrations faites dans le chapitre 1, on définit ainsi le coefficient de corrélation *intra-groupe* ici par

$$\begin{aligned}\rho(y_{ij}, y_{i'j}) &= \frac{\mathrm{Cov}(y_{ij}, y_{i'j})}{\sqrt{\mathrm{Var}(y_{ij})\mathrm{Var}(y_{i'j})}} \\ &= \frac{\sigma^2_{u_0}}{\sigma^2_{u_0} + \sigma^2_e} \\ &= \frac{Variance\ inter\text{-}groupes}{Variance\ totale}.\end{aligned}$$

Ce coefficient exprime :

- La corrélation entre deux individus d'un même groupe,
- La part de variabilité imputable au niveau 2.

L'intérêt de cette première phase de l'analyse est qu'elle fournit des indications sur la partition de la variance entre les deux niveaux, la part qui revient au niveau 2 (les groupes) et la part qui revient au niveau 1 (les individus). Elle donne donc une estimation du coefficient de corrélation *intra-groupe*. Ceci constitue un début d'indication d'un possible effet contextuel (effet groupe). S'il n'y a pas de variation significative entre macro-unités (on ne rejette pas $H_0 : \sigma^2_{u_0} = 0$) il est inutile de développer une analyse plus complexe.

b. Modèle conditionnel : introduction d'une covariable, *intercept* aléatoire

On développe maintenant un modèle complexe en introduisant une variable explicative X et en supposant que l'*intercept* est aléatoire et la pente est fixe. On rappelle qu'on est dans le cas où la variable explicative X est centrée pour faciliter l'interprétation de l'*intercept*. Nous avons la formulation suivante de nouvelles équations :

$$\begin{aligned}\text{Niveau 1} : y_{ij} &= \beta_{0j} + \beta_{1j}x_{ij} + e_{ij}, \\ \text{Niveau 2} : \beta_{0j} &= \beta_0 + u_{0j}, \\ \text{soit } y_{ij} &= \underbrace{\beta_0 + \beta_{1j}x_{ij}}_{Partie\ \ fixe} + \underbrace{(u_{0j} + e_{ij})}_{Partie\ \ aléatoire}.\end{aligned}$$

La variable X est mesurée au niveau 1 par x_{ij}. Elle est susceptible d'expliquer une partie de la variabilité de l'*outcome* Y. On émet les hypothèses suivantes :

- les e_{ij} sont les réalisations *iid* d'une variable aléatoire de loi $N(0, \sigma_e^2)$,
- les u_{0j} sont des réalisations *iid* de loi $N(0, \sigma_{u_0}^2)$ indépendantes des e_{ij},

Le modèle comporte maintenant deux facteurs fixes β_0 et β_1 et un facteur aléatoire u_0.

c. Modèle conditionnel : *intercept* et pente aléatoires

On développe encore un modèle, mais plus complexe en supposant que l'intercept et la pente sont aléatoires dans le cas où la variable explicative X est centrée. En effet, l'hypothèse de pente identique est levée ici, on suppose alors que l'hétérogénéité entre les groupes ne se résume pas à leur valeur moyenne, mais qu'elle influe également la relation entre les différences de départ et d'arrivée. Nous avons les nouvelles équations suivantes :

$$\text{Niveau 1} : y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + e_{ij},$$
$$\text{Niveau 2} : \beta_{0j} = \beta_0 + u_{0j} \ , \quad \beta_{1j} = \beta_1 + u_{1j},$$

$$\text{soit } y_{ij} = \underbrace{\beta_0 + \beta_1 x_{ij}}_{Partie\ \ fixe} + \underbrace{(u_{0j} + u_{1j} x_{ij} + e_{ij})}_{Partie\ \ aléatoire}, \text{ où}$$

- β_0 est la partie fixe des ordonnées pour chaque groupe qui représente l'ordonnée moyenne pour tous les groupes ;
- β_1 est la partie fixe des pentes pour chaque groupe qui représente la pente moyenne pour tous les groupes ;
- e_{ij} est le terme d'erreur correspondant à l'individu i dans le groupe j,
- u_{0j} est le terme aléatoire traduisant la variabilité de l'ordonnée entre les groupes et représente l'écart de la moyenne de chaque groupe à la moyenne générale mesurée au niveau 2 et identique pour tous les individus du groupe j.
- u_{1j} est le terme aléatoire traduisant la variabilité de la pente entre les groupes. Cette variabilité est mesurée au niveau 2 et est la même pour tous les individus du groupe j.

On émet les hypothèses suivantes :

- les e_{ij} sont les réalisations *iid* d'une variable aléatoire E de loi $N(0, \sigma_e^2)$, où σ_e^2 traduit la variabilité des individus au sein d'un même groupe. Cette variabilité dite intra-groupe est supposée identique au sein de chaque groupe.
- les couples (u_{0j}, u_{1j}) sont des réalisations *iid* d'un vecteur aléatoire gaussien U centré et de matrice de variance-covariance :

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{u_0}^2 & \mathrm{Cov}(u_0, u_1) \\ \mathrm{Cov}(u_0, u_1) & \sigma_{u_1}^2 \end{pmatrix},$$

  où les variances $\sigma_{u_0}^2$ et $\sigma_{u_1}^2$ traduisent respectivement la variabilité de l'ordonnée et de la pente selon les groupes (variabilité dite inter-groupe) et $\mathrm{Cov}(u_0, u_1)$ représente la covariance entre les variables aléatoires u_0 et u_1 qui est égale à la covariance entre les variables aléatoires ordonnée et pente.
- la variable d'erreur E au sein de chaque groupe est indépendante de l'erreur U entre les groupes, c'est-à-dire que les résidus des niveaux 1 et 2 sont supposés indépendants.

Dans ce modèle, la partie aléatoire $(u_{0j} + u_{1j}x_{ij} + e_{ij})$ dépend de la variable X. Nous sommes donc dans une situation d'hétéroscédasticité.
Comme dans le cas précédent, le modèle comporte deux facteurs fixes β_0 et β_1 mais il y a maintenant deux facteurs aléatoires u_0 et u_1 qui ne sont pas supposés indépendants.

d. Estimation

Dans le dernier modèle formulé (modèle conditionnel : *intercept* et pente aléatoires), les paramètres à estimer sont :

- β_0, β_1 liés aux effets fixes,
- σ_e^2, $\sigma_{u_0}^2$, $\sigma_{u_1}^2$, $\mathrm{Cov}(u_0, u_1)$ liés aux effets aléatoires.

La méthode d'estimation est celle du maximum de vraisemblance et du maximum de vraisemblance restreinte montrée au chapitre 1.

La différence entre ces deux méthodes provient du fait que la méthode d'estimation par maximum de vraisemblance des parties fixes et aléatoires est obtenue à partir de la vraisemblance complète dans le dernier modèle formulé, alors que la méthode d'estimation par maximum de vraisemblance restreinte sépare l'estimation des paramètres de ces deux parties. Il est important de retenir que ces deux méthodes diffèrent peu pour l'estimation des effets fixes, mais peuvent conduire à des différences plus importantes pour les composantes de la variance.

e. Tests d'hypothèses

Nous formulons ici des hypothèses simples ou multiples montrées dans le chapitre 1 sur

- les coefficients β_0, β_1 (effets fixes),
- les paramètres σ_e^2, $\sigma_{u_0}^2$, $\sigma_{u_1}^2$, $\mathrm{Cov}(u_0, u_1)$ (effets aléatoires).

Le test de l'hypothèse H_0 de nullité d'un paramètre revient à comparer deux modèles : le modèle H_0 au modèle H_1. On dit alors que le modèle H_0 est emboîté dans le modèle H_1. Pour tester ici l'hypothèse H_0, nous utilisons les déviances des deux modèles en calculant la statistique : $\lambda_n = -2l_R + 2l_C$, comme nous l'avons montrée au chapitre 1. Sous H_0, λ_n suit une loi de khi-deux χ_r^2 où $r = dl_C - dl_R = 4 - 3 = 1$ degrés de liberté (différence entre les nombres de paramètres sous l'avant-dernier modèle et le dernier modèle).
l_R est la log-vraisemblance sous le modèle réduit (sous H_0) : modèle avec une covariable et *intercept* aléatoire et l_C est la log-vraisemblance sous le modèle complet (sous H_1) : modèle avec *intercept* et pente aléatoires.
Lorsqu'on est amené à comparer des modèles qui ne sont pas emboîtés, on utilise d'autres indices pour choisir le modèle le mieux adapté aux observations. Les plus utilisés sont le critère d'information d'Akaike (AIC) et le critère d'information bayésienne (BIC) définis dans le chapitre 1.

Au vu de la formulation du modèle à deux niveaux avec un *outcome* univarié, nous retenons que ce modèle tient compte en général de la dépendance découlant des données en grappes, ce que ne font pas les modèles linéaires simples. Toutefois, ce modèle à deux niveaux ne peut incorporer qu'un seul *outcome* à partir d'unités. Ainsi, une telle modélisation multiniveaux ne peut pas être utilisé pour un problème d'estimation du lien entre un *outcome* bivarié et des covariables.

2.4.2 Modèle à deux niveaux avec un *outcome* bivarié

L'extension du modèle multiniveaux avec un *outcome* univarié au modèle multiniveaux avec un *outcome* bivarié permet de modéliser simultanément deux *outcomes* en tenant compte de la dépendance des observations résultant de l'imbrication des participants d'une étude dans de différents contextes. La stratégie de modélisation montrée dans cette sous-section est inspirée du livre de Keenan A. Pituch et James P. Stevens [40].

a. Avantages de l'usage du modèle à deux niveaux avec un *outcome* bivarié
Lorsqu'il y a un problème d'estimation du lien entre deux *outcomes* et les covariables avec des données hiérarchiques, on peut mener une analyse multiniveaux bivariée. L'une des raisons d'envisager une telle analyse est de se prémunir contre l'inflation du taux d'erreur de type I global en utilisant un test global initial comme approche de test protégée.

Une deuxième raison est qu'au lieu d'examiner les différences de groupe univariées en utilisant un score total, obtenu en additionnant ou en faisant la moyenne des scores de plusieurs sous-tests, on peut comparer les différences de groupe sur les multiples sous-tests, ce qui peut fournir plus d'informations sur la nature des différences de groupe.

L'analyse multiniveaux avec un *outcome* bivarié n'exige pas qu'un participant de l'étude fournisse des scores pour chaque *outcome*. Au contraire, si un participant fournit un score pour au moins un des *outcomes*, celui-ci peut être inclus dans l'analyse.

Par rapport aux analyses standards, l'analyse multiniveaux bivarié utilise davantage les données disponibles, ce qui peut permettre d'augmenter la puissance du modèle. De plus, les logiciels de statistique comme SAS et SPSS offrent un traitement de vraisemblance maximale des données manquantes, ce qui permet d'obtenir des estimations optimales.

D'après les études de Snijders et Bosker [57], l'usage du modèle multiniveaux avec un *outcome* bivarié peut entraîner des erreurs standards plus petites pour les tests des prédicteurs sur un *outcome* donné par rapport à une analyse multiniveaux avec un *outcome* univarié. La précision supplémentaire et l'augmentation de la puissance du modèle avec un *outcome* bivarié peuvent être pertinentes lorsque les *outcomes* sont corrélées et les participants ont des données manquantes sur certains des *outcomes*.

Lorsque les *outcomes* ont une échelle similaire, le modèle multiniveaux avec un *outcome* bivarié peut être utilisé pour tester si les effets d'une covariable sont les mêmes ou diffèrent entre les *outcomes*. Dans un cadre expérimental, par exemple, un chercheur peut apprendre si les effets du traitement sont plus forts pour un *outcome* que pour l'autre, ce qui peut suggérer de revoir la nature et la mise en œuvre de l'intervention.

Lorsque les participants sont regroupés dans des grappes, le modèle multiniveaux avec un *outcome* bivarié peut être utilisé pour décrire les associations entre les *outcomes* au niveau des participants et des grappes grâce au partitionnement de la variabilité obtenue. Au lieu d'apprendre comment les scores d'un seul *outcome* varient entre les participants et les grappes, comme le cas du modèle multiniveaux avec un *outcome* univarié, le modèle multiniveaux avec un *outcome* bivarié peut informer les chercheurs des associations entre les *outcomes* qui sont à l'intérieur et entre les grappes.

Eu égard à ces avantages, le modèle multiniveaux avec un *outcome* bivarié en particulier et avec un *outcome* multivarié en général reste une procédure d'analyse plus complexe que le modèle multiniveaux avec un *outcome* univarié.

b. Format des données

Le format de l'ensemble de données requis pour la modélisation multiniveaux bivarié en particulier ou multivarié en général est le format long. Ce format des données est la clé de cette modélisation. Nous expliquons de manière générale comment un ensemble de données organisé initialement en format large peut être reformaté dans le format long. Le modèle à deux niveaux est utilisé à cette fin avec deux *outcomes* étiquetés respectivement par Y_1 et Y_2. Le tableau 2.1 montre tout d'abord une esquisse d'un ensemble de données initialement au format large contenant une colonne de l'identifiant de l'individu, une colonne de grappe dont fait partie chacun des individus, une colonne de Y_1, une colonne de Y_2 distincte de celle de Y_1 et une colonne de X.

Identifiant	Grappe	Y_1	Y_2	X
1	1	y_{11}	y_{21}	x_1
2	3	y_{12}	y_{22}	x_2
3	1	y_{13}	y_{23}	x_3
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
n	2	y_{1n}	y_{2n}	x_n

TABLEAU 2.1 – Esquisse de l'ensemble de données initialement en format large

Le format large signifie que chaque individu a un enregistrement et que toutes les variables de cet individu, en particulier Y_1 et Y_2, apparaissent sur ce même enregistrement dans des colonnes différentes. Ainsi, étant donné qu'il y a n individus dans l'ensemble de données, ce dernier contient n enregistrements.

Pour le modèle multiniveaux bivarié, les données doivent être au format long, c'est-à-dire qu'au lieu de faire apparaître les scores des deux *outcomes* Y_1 et Y_2 dans des colonnes séparées, les scores de tous les *outcomes* doivent être placés dans une seule colonne, ce qui crée plusieurs enregistrements pour chaque individu.

Ainsi, dans l'ensemble de données reformaté, un individu aura plusieurs enregistrements, égaux au nombre d'*outcomes* obtenus pour cet individu. Par exemple, si l'échantillon compte 1000 individus et que des données ont été collectées sur deux *outcomes* pour chaque individu, l'ensemble de données nécessaires pour ce modèle comportera 2000 enregistrements. Le tableau 2.2 montre une esquisse des données au format long avec un nombre d'enregistrements égal à $2n$. Les scores de chaque individu apparaissent sur deux lignes distinctes. En outre, les colonnes contenant les valeurs de Y_1 et de Y_2 ont été supprimées. Elles sont remplacées par deux nouvelles variables, $Index$ et Y.

Enregistrement	Identifiant	Grappe	$Index$	Y	X	a_1	a_2
1	1	1	1	y_{11}	x_1	1	0
2	1	1	2	y_{21}	x_1	0	1
3	2	3	1	y_{12}	x_2	1	0
4	2	3	2	y_{22}	x_2	0	1
5	3	1	1	y_{13}	x_3	1	0
6	3	1	2	y_{23}	x_3	0	1
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
$2n-1$	n	2	1	y_{1n}	x_n	1	0
$2n$	n	2	2	y_{2n}	x_n	0	1

TABLEAU 2.2 – Esquisse de l'ensemble de données en format long

La variable Y contient les scores de Y_1 et de Y_2 dans une seule colonne. $Index$ indique la séquence des deux *outcomes* dans la colonne Y (c'est-à-dire Y_1 suivi de Y_2). L'identifiant de l'individu, la grappe dans laquelle se trouve l'individu et la covariable X sont aussi au format long, les valeurs pour un cas donné étant répétées dans tous les enregistrements. L'extrême droite du tableau 2.2 montre deux variables indicatrices a_1 et a_2 codées de façon fictive. Ces variables indicatrices sont requises pour ce modèle multiniveaux.
a_1 est codé 1 lorsqu'un enregistrement donné conserve un score pour Y_1 et 0 sinon
a_2 est codé 1 lorsqu'un enregistrement donné conserve un score pour Y_2 et 0 sinon.

En somme, pour la modélisation multiniveaux multivarié, l'organisation de l'ensemble des données dans le format long et l'utilisation de variables indicatrices codées par des nombres fictifs sont les clés de la conversion d'un modèle multiniveaux univarié standard en un modèle multiniveaux multivarié.

c. Formulation du modèle au niveau 1 avec un *outcome* bivarié

c.1 Intégration des *outcomes* dans le modèle

Pour cette partie, la notation utilisée pour les équations suit celle utilisée dans le livre de Raudenbush et Bryk en 2002 [45]. Pour un modèle à deux niveaux avec un *outcome* bivarié, le modèle de base au niveau 1 est généralement représenté par l'équation suivante :

$$Y_{ij} = \pi_{1j}a_{1j} + \pi_{2j}a_{2j}, \tag{2.1}$$

- où Y_{ij} représente la colonne unique intitulée Y dans le tableau 2.2 contenant les scores de chaque *outcome* i (Y_1 et Y_2), pour un individu j donné,
- a_{1j} et a_{2j} sont des variables indicatrices codées pour un *outcome* i d'un individu j donné, comme indiqué dans le tableau 2.2.
  L'équation 2.1 n'a pas d'intercept ni de terme d'erreur.
- Il est important de comprendre ce que représentent les paramètres π_{1j} et π_{2j} dans 2.1 :
  Si $a_{1j} = 1$, en raison du codage utilisé, $a_{2j} = 0$. Dans ce cas, l'équation 2.1 devient

  $$Y_{ij} = \pi_{1j}(1) \text{ ou simplement } Y_{ij} = \pi_{1j},$$

  ce qui correspond à Y_1, en raison de la structure de la colonne Y et des variables à code fictif. Autrement dit, lorsque $a_{1j} = 1$, l'*outcome* bivarié Y_{ij} dans l'équation 2.1 tire uniquement les observations des enregistrements ayant une valeur pour Y_1 dans la colonne Y, ce qui, dans cet ensemble de données, correspond également à la sélection des seuls enregistrements impairs.

De même, si $a_{2j} = 1$, en raison du codage utilisé, $a_{1j} = 0$. Dans ce cas, l'équation 2.1 devient

$$Y_{ij} = \pi_{2j}(1) \text{ ou simplement } Y_{ij} = \pi_{2j},$$

ce qui correspond à Y_2. Ainsi, lorsque $a_{2j} = 1$, l'*outcome* bivarié Y_{ij} dans l'équation 2.1 utilise seulement les valeurs des enregistrements numérotés pairs dans l'ensemble de données, qui correspondent à Y_2.

π_{1j} et π_{2j} représentent donc respectivement Y_1 et Y_2 en raison des données et du modèle.

L'équation 2.1, est utilisée comme équation de base dans la modélisation multiniveaux avec un *outcome* bivarié.

c.2 Modèle vide au niveau 1

Ce modèle n'inclut aucune variable explicative. Les paramètres π_{1j} et π_{2j} qui représentent les *outcomes* Y_1 et Y_2 peuvent varier d'un individu à l'autre. Nous avons les équations :

$$\pi_{1j} = \beta_{10} + e_{1j}, \tag{2.2}$$

$$\pi_{2j} = \beta_{20} + e_{2j}, \tag{2.3}$$

$$\begin{aligned} \text{soit} \quad Y_{ij} &= \pi_{1j}a_{1j} + \pi_{2j}a_{2j} \\ &= (\beta_{10} + e_{1j})a_{1j} + (\beta_{20} + e_{2j})a_{2j} \\ &= \beta_{10}a_{1j} + e_{1j}a_{1j} + \beta_{20}a_{2j} + e_{2j}a_{2j} \\ &= \beta_{10}a_{1j} + \beta_{20}a_{2j} + e_{1j}a_{1j} + e_{2j}a_{2j}, \end{aligned}$$

où β_{10} et β_{20} représentent respectivement la moyenne de Y_1 et de Y_2.

Les termes résiduels (e_{1j} et e_{2j}) sont supposés suivre une distribution normale bivariée, avec une moyenne attendue de 0 et une certaine variance-covariance.

Nous avons au total 5 paramètres à estimer pour ce modèle vide :

- les deux effets fixes β_{10} et β_{20} qui sont les moyennes de Y_1 et Y_2 respectivement,
- la variance de e_{1j} ($\sigma^2_{e_1}$),
- la variance de e_{2j} ($\sigma^2_{e_2}$),
- la covariance de e_{1j} et e_{2j} ($\text{Cov}(e_1, e_2)$).

La corrélation des résidus indique le degré de corrélation entre Y_1 et Y_2 et elle vaut

$$\rho(e_{1j}, e_{2j}) = \frac{\text{Cov}(e_1, e_2)}{\sqrt{\sigma^2_{e_1}\sigma^2_{e_2}}}.$$

La déviance du modèle pour les 5 paramètres est $-2l$ où l est la log-vraisemblance.
La valeur de la déviance reflète l'ajustement du modèle et sera comparée à la déviance obtenue lorsqu'une covariable X sera ajoutée au modèle vide pour déterminer si l'ajustement s'améliore.
La procédure d'estimation par maximum de vraisemblance complète (telle qu'elle est montrée dans le chapitre 1) est utilisée pour tester les effets fixes.
La signification de chaque élément de la variance-covariance est testée à l'aide de la règle générale de tests sur les composantes de la variance montrée dans le chapitre 1.

c.3 Inclusion d'une covariable dans le modèle au niveau 1

On introduit maintenant une variable X dans le modèle vide. L'objectif est de déterminer s'il existe des effets de X pour tout *outcome*, ce qui sera accompli par un test global de l'hypothèse nulle selon laquelle aucun effet de X n'est présent pour tout *outcome*. Si des effets de X sont présents, alors les effets et les résultats des tests statistiques pour X seront examinés pour chaque *outcome*. Pour ce nouveau modèle, au niveau 1, les équations 2.2 et 2.3 sont modifiées de façon à ce que la variable X soit ajoutée à chacune des équations. Les équations du nouveau modèle sont :

$$\pi_{1j} = \beta_{10} + \beta_{11}X_j + e_{1j}, \tag{2.4}$$

$$\pi_{2j} = \beta_{20} + \beta_{21}X_j + e_{2j}, \tag{2.5}$$

$$\begin{aligned}\text{soit} \quad Y_{ij} &= \pi_{1j}a_{1j} + \pi_{2j}a_{2j} \\ &= (\beta_{10} + \beta_{11}X_j + e_{1j})a_{1j} + (\beta_{20} + \beta_{21}X_j + e_{2j})a_{2j} \\ &= (\beta_{10} + \beta_{11}X_j)a_{1j} + (\beta_{20} + \beta_{21}X_j)a_{2j} + e_{1j}a_{1j} + e_{2j}a_{2j},\end{aligned}$$

où X_j représente la variable explicative pour l'individu j, β_{10} et β_{20} représentent respectivement la moyenne de Y_1 et de Y_2, β_{11} et β_{21} représentent la différence moyenne entre deux groupes caractéristiques de X pour Y_1 et Y_2, respectivement.
Les termes résiduels (e_{1j} et e_{2j}) sont supposés suivre une distribution normale bivariée, avec une matrice de variances-covariances égale entre les groupes des individus.
L'hypothèse nulle bivariée pour le test de X est $H_0 : \beta_{11} = \beta_{21} = 0$. Elle est testée en utilisant les déviances entre le nouveau modèle et le modèle vide.

Pour ce nouveau modèle, les paramètres à estimer sont :

- 4 effets fixes : β_{10}, β_{20} , β_{11} et β_{21} dans les équations 2.4 et 2.5,
- 3 éléments de la variance-covariance : $\sigma^2_{e_1}$,$\sigma^2_{e_2}$, $\text{Cov}(e_1, e_2)$

Pour inclure une variable explicative X dans le modèle de façon à ce que des effets distincts de X soient estimés pour chaque *outcome*, la variable explicative X doit être multipliée par la variable $Index$ dans le tableau 2.2. Ainsi, pour estimer β_{11} et β_{21}, le terme $X \times Index$ doit être ajouté à l'instruction du modèle via le logiciel de statistique.

c.4 Vérifier si l'effet d'une covariable diffère selon les *outcomes*

Cette étape vérifie si l'effet d'une variable explicative X est de la même ampleur pour chaque *outcome*. Si les *outcomes* sont mesurées sur la même échelle, les chercheurs peuvent souhaiter savoir si une nouvelle intervention a des effets plus forts pour certains *outcomes* que pour d'autres. Pour ce faire, il faut d'abord contraindre les effets fixes à être égaux, puis tester la différence d'ajustement en utilisant les déviances entre ce modèle contraint et un modèle où les effets sont estimés librement.

Dans les équations 2.4 et 2.5, les effets fixes de X sont librement estimés (sans contrainte) pour les deux *outcomes* Y_1 (β_{11}) et Y_2 (β_{21}). Nous testons maintenant si ces effets fixes sont identiques ou différents pour Y_1 et Y_2. Le modèle utilisé est essentiellement le même que celui du modèle précédent, mais un effet de covariable commun présumé sera estimé. Pour estimer ce modèle, on remplace le terme $X \times Index$ du modèle précédent par X.

Le nombre de paramètres à estimer est maintenant 6, comprenant :

- 3 effets fixes : β_{10}, β_{20} et $\beta_{11} = \beta_{21}$,
- 3 éléments de la variance-covariance : $\sigma^2_{e_1}$,$\sigma^2_{e_2}$, $\mathrm{Cov}(e_1, e_2)$

d. Formulation du modèle à deux niveaux avec un *outcome* bivarié

La partie précédente a présenté la procédure de cette modélisation au niveau 1 (niveau de l'individu) ; elle n'a pas inclus le niveau 2 (niveau de la grappe) dans le modèle. Nous allons tenir compte maintenant de la dépendance au sein de la grappe et entre les grappes en ajoutant dans le modèle précèdent le niveau 2 (niveau de la grappe).

d.1 Formulation du modèle à deux niveaux pour les effets d'un prédicteur

L'équation 2.1 qui était auparavant l'équation de base, est légèrement modifiée afin de reconnaître l'inclusion du niveau de la grappe. Elle est maintenant :

$$Y_{ijk} = \pi_{1jk} a_{1jk} + \pi_{2jk} a_{2jk}. \tag{2.6}$$

Elle est identique à l'équation 2.1 sauf que l'indice k a été ajouté. Ainsi, π_{1jk} et π_{2jk} représentent les scores de l'*outcome* Y_1 et de l'*outcome* Y_2 respectivement, pour un individu j donné qui appartient à une grappe k donnée.

- Au niveau 1 (niveau de l'individu) le modèle a alors pour équations

$$\pi_{1jk} = \beta_{10k} + e_{1jk}, \tag{2.7}$$

$$\pi_{2jk} = \beta_{20k} + e_{2jk}, \tag{2.8}$$

$$\begin{aligned}\text{soit} \quad Y_{ijk} &= \pi_{1jk}a_{1jk} + \pi_{2jk}a_{2jk} \\ &= (\beta_{10k} + e_{1jk})a_{1jk} + (\beta_{20k} + e_{2jk})a_{2jk} \\ &= \beta_{10k}a_{1jk} + e_{1jk}a_{1jk} + \beta_{20k}a_{2jk} + e_{2jk}a_{2jk} \\ &= \beta_{10k}a_{1jk} + \beta_{20k}a_{2jk} + e_{1jk}a_{1jk} + e_{2jk}a_{2jk},\end{aligned}$$

où β_{10k} et β_{20k} représentent la moyenne pour une grappe k, pour Y_1 et Y_2, respectivement. Les termes résiduels au niveau de l'individu (e_{1jk} et e_{2jk}) sont supposés suivre une distribution normale bivariée, avec une moyenne attendue de 0 et des variances ($\sigma^2_{e_1}$ et $\sigma^2_{e_2}$) et une covariance ($\text{Cov}(e_1, e_2)$).

Supposons que l'affectation du prédicteur X varie d'une grappe à l'autre. La variable indicatrice de X apparaîtra alors dans le modèle au niveau de la grappe.

- Au niveau 2 (niveau de la grappe) nous avons ainsi les équations suivantes du modèle :

$$\beta_{10k} = \gamma_{100} + \gamma_{101}X_k + u_{10k}, \tag{2.9}$$

$$\beta_{20k} = \gamma_{200} + \gamma_{201}X_k + u_{20k}, \tag{2.10}$$

$$\text{soit} \quad Y_{ijk} = \beta_{10k}a_{1jk} + \beta_{20k}a_{2jk} + e_{1jk}a_{1jk} + e_{2jk}a_{2jk}$$

$$Y_{ijk} = (\gamma_{100} + \gamma_{101}X_k + u_{10k})a_{1jk} + (\gamma_{200} + \gamma_{201}X_k + u_{20k})a_{2jk} + e_{1jk}a_{1jk} + e_{2jk}a_{2jk}$$

$$Y_{ijk} = (\gamma_{100} + \gamma_{101}X_k)a_{1jk} + (\gamma_{200} + \gamma_{201}X_k)a_{2jk} + u_{10k}a_{1jk} + u_{20k}a_{2jk} + e_{1jk}a_{1jk} + e_{2jk}a_{2jk},$$

où γ_{100} et γ_{200} représentent la moyenne générale pour Y_1 et Y_2 respectivement.
Les paramètres clés sont γ_{101} et γ_{201} et représentent les différences de moyennes entre deux groupes caractéristiques de X pour Y_1 et Y_2 respectivement.
Les termes résiduels au niveau de la grappe sont u_{10k} et u_{20k} et sont supposés suivre une distribution normale bivariée avec une moyenne attendue de 0 et des variances constantes ($\sigma^2_{u_1}$ et $\sigma^2_{u_2}$) et une covariance ($\text{Cov}(u_1, u_2)$).

Le modèle décrit par les équations 2.6 à 2.10 comporte au total 10 paramètres à estimer :

- 4 effets fixes : γ_{100}, γ_{200}, γ_{101}, γ_{201}
- 6 éléments de variance-covariance : $\sigma^2_{e_1}$, $\sigma^2_{e_2}$, $\text{Cov}(e_1, e_2)$, $\sigma^2_{u_1}$,$\sigma^2_{u_2}$, $\text{Cov}(u_1, u_2)$

Comme définies dans le chapitre 1, la matrice $\boldsymbol{R}$ est ici la matrice de variance-covariance pour les résidus au niveau 1 (niveau de l'individu), et la matrice $\boldsymbol{G}$ est ici la matrice de variance-covariance pour les résidus au niveau 2 (niveau de la grappe). Nous avons alors :

$$\boldsymbol{R} = \begin{pmatrix} \sigma^2_{e_1} & \text{Cov}(e_1, e_2) \\ \text{Cov}(e_2, e_1) & \sigma^2_{e_2} \end{pmatrix},$$

$$\boldsymbol{G} = \begin{pmatrix} \sigma^2_{u_1} & \text{Cov}(u_1, u_2) \\ \text{Cov}(u_2, u_1) & \sigma^2_{u_2} \end{pmatrix}.$$

La corrélation entre les résidus au niveau de l'individu est $\rho(e_{1jk}, e_{2jk}) = \frac{\text{Cov}(e_1,e_2)}{\sqrt{\sigma^2_{e_1}\sigma^2_{e_2}}}$ et celle au niveau de la grappe est $\rho(u_{10k}, u_{20k}) = \frac{\text{Cov}(u_1,u_2)}{\sqrt{\sigma^2_{u_1}\sigma^2_{u_2}}}$.
Il est à noter que si l'on souhaite un modèle vide omettant la variable X, des équations 2.9 et 2.10 pourraient être estimées avant ce modèle. Si cela était fait, les déviances des modèles pourraient être comparées comme cela a été effectué précédemment pour tester l'hypothèse nulle bivariée globale d'absence d'effet de X.

d.2 Formulation du modèle à deux niveaux avec multiples prédicteurs

On inclut maintenant plusieurs variables explicatives. L'équation 2.6 reste le modèle de base. Le modèle au niveau 1 est modifié pour inclure ces variables.

- Le modèle au niveau 1 (niveau de l'individu) est

$$\pi_{1jk} = \beta_{10k} + \sum_{i=1}^{p} \beta_{1ik} X_{ijk} + e_{1jk}, \tag{2.11}$$

$$\pi_{2jk} = \beta_{20k} + \sum_{i=1}^{p} \beta_{2ik} X_{ijk} + e_{2jk}, \tag{2.12}$$

$$\text{avec} \quad Y_{ijk} = \pi_{1jk} a_{1jk} + \pi_{2jk} a_{2jk},$$

où X_{ijk} représente la covariable i pour l'individu j appartenant à une grappe k, pour $i = 1, 2, 3..., p$ covariables,

β_{1ik} représente le coefficient associé à X_{ijk} pour $i = 1, 2, 3..., p$ covariables.

Les termes résiduels au niveau de l'individu ou au sein de la grappe (e_{1jk} et e_{2jk}) sont supposés suivre une distribution normale bivariée avec une moyenne attendue de 0 et une certaine variance-covariance.

- Au niveau 2 (niveau de la grappe),

chacun des coefficients de régression β_{1ik} et β_{2ik} des équations 2.11 et 2.12 peut être considéré comme une variable dépendante à modéliser en fonction de certains prédicteurs. Cependant, cette situation peut produire des effets d'interaction entre deux covariables dans le modèle que nous pouvons tester par l'hypothèse nulle bivariée de l'absence d'interaction entre deux covariables pour les deux *outcomes*.

La matrice de variance-covariance pour les résidus au niveau 1 et celle pour les résidus au niveau 2 sont obtenus respectivement comme précédemment par :

$$\boldsymbol{R} = \begin{pmatrix} \sigma'^2_{e_1} & \mathrm{Cov}'(e_1, e_2) \\ \mathrm{Cov}'(e_2, e_1) & \sigma'^2_{e_2} \end{pmatrix},$$

$$\boldsymbol{G} = \begin{pmatrix} \sigma'^2_{u_1} & \mathrm{Cov}'(u_1, u_2) \\ \mathrm{Cov}'(u_2, u_1) & \sigma'^2_{u_2} \end{pmatrix}$$

La corrélation entre les résidus au niveau 1 est $\rho'(e_{1jk}, e_{2jk}) = \frac{\mathrm{Cov}'(e_1,e_2)}{\sqrt{\sigma'^2_{e_1}\sigma'^2_{e_2}}}$ et celle au niveau 2 est $\rho'(u_{10k}, u_{20k}) = \frac{\mathrm{Cov}'(u_1,u_2)}{\sqrt{\sigma'^2_{u_1}\sigma'^2_{u_2}}}$.

e. Test bivarié pour les variances-covariances multiples

Cette partie de la modélisation teste si la différence intra-grappe entre deux catégories d'individus sur un *outcome* varie entre les grappes. Dans les équations 2.11 et 2.12, β_{1ik} et β_{2ik} représentent chacun la différence attendue entre deux catégories d'individus sur un *outcome* au sein d'une grappe k, en contrôlant les autres covariables du modèle. Cette différence peut être supposée constante d'une grappe à l'autre dans le modèle.

Il arrive donc souvent qu'on ne dispose pas d'hypothèses à priori solides pour préciser si ces effets sont fixes ou varient d'une grappe à l'autre. Bien qu'il soit souvent prudent de modéliser de tels effets comme étant fixes d'une grappe à l'autre, parce que l'inclusion d'une variation insignifiante peut causer des problèmes d'estimation, la variation réelle, si elle est présente, doit être incluse dans le modèle, car l'exclusion de cette variation peut entraîner une augmentation du taux d'erreur type pour le test des effets fixes. Afin d'utiliser une base empirique pour modéliser la différence entre les deux catégories d'individus comme fixe ou variable, on effectue un test pour déterminer si cette différence sur un *outcome* varie entre les grappes après avoir contrôlé les autres covariables du modèle. Pour mettre en œuvre ce test, un terme aléatoire associé à cette différence doit être ajouté à l'une des équations de β_{1ik} et β_{2ik}. Pour tester si la différence entre deux catégories d'individus données sur Y_1 ou sur Y_2 varie entre les grappes, on aurait l'équation

$$\beta_{1ik} = \gamma_{110} + \gamma_{111}X_{ik} + u_{11k} \text{ ou } \beta_{2ik} = \gamma_{210} + \gamma_{211}X_{ik} + u_{21k}, \text{ respectivement.}$$

En ajoutant un résidu dans l'équation de β_{1ik} (u_{11k}) ou dans celle de β_{2ik} (u_{21k}), la matrice $\boldsymbol{G}$ devient :

$$\boldsymbol{G} = \begin{pmatrix} \sigma^2_{u_1} & \text{Cov}(u_1, u_2) & \text{Cov}(u_1, u_3) \\ \text{Cov}(u_2, u_1) & \sigma^2_{u_2} & \text{Cov}(u_2, u_3) \\ \text{Cov}(u_3, u_1) & \text{Cov}(u_3, u_2) & \sigma^2_{u_3} \end{pmatrix}$$

L'hypothèse nulle suppose que les valeurs des paramètres de tous les nouveaux termes sont nulles dans la population. Ceci peut être testé en comparant la déviance du modèle actuel à la déviance du modèle précédent, comme cela a été effectué dans les parties précédentes. On note que le modèle précédent est emboîté dans le modèle actuel. Sous H_0, si la statistique du khi-deux (différence des deux déviances) ne dépasse pas la valeur critique au seuil $\alpha = 5\%$ et à r degrés de liberté (différence entre les nombres de paramètres sous le modèle précédent et le modèle actuel) alors l'amélioration de l'ajustement obtenue par l'ajout de ce résidu est négligeable. Il n'y aura pas ainsi de support empirique pour inclure ces effets aléatoires supplémentaires.

CHAPITRE 3

Application aux données *CentrÉS*

3.1 Mise en Contexte

Les inégalités sociales de santé sont des différences systématiques dans les états de santé entre les groupes socio-économiques, les régions et communautés, les femmes et les hommes et les groupes ethniques ayant leurs racines dans des arrangements sociaux injustes [37]. En général, les individus dont le statut socio-économique est moins favorable sont en moins bonne santé que celles dont le statut est plus favorable [8]. Les enquêtes répétées sur l'ampleur et les causes de ces inégalités ont adopté un modèle social de la santé, qui place l'individu au centre entouré de *couches d'influence* liées aux facteurs de style de vie, aux réseaux sociaux et communautaires, aux conditions de vie et de travail et à l'environnement physique et socio-économique [4, 38]. Ainsi, la communauté, dans laquelle cohabitent et interagissent les individus, exerce une influence sur leurs états de santé. Le concept de *défavorisation* sert, entre autres, à l'évaluation de ces inégalités au niveau contextuel et réfère au désavantage d'une communauté locale à laquelle appartient un individu relativement aux autres communautés de la société [24, 60].

Pour expliquer les inégalités entre groupes sociaux définis avec les caractéristiques individuelles ou entre quartiers plus ou moins favorisés, peu d'études mènent une analyse multiniveaux multivarié pour estimer par exemple le lien entre un *outcome* bivarié (santé mentale perçue et sentiment d'appartenance au quartier) et les covariables.

3.2 Question de recherche

Les inégalités sociales de santé demeurent incomprises malgré plusieurs études [37, 38]. Par exemple, des variations dans l'état de santé autorapporté persistent notamment en fonction du niveau d'éducation ou de revenu.

Au Québec, une plus grande proportion de personnes moins scolarisées se déclarent en mauvaise santé (18%) comparativement aux personnes plus scolarisées (7%) [**?**], et ce dans tous les groupes d'âge [28]. L'isolement social et un faible sentiment d'appartenance au quartier de résidence, tous deux associés à divers états de santé, sont également plus prévalents chez les personnes plus défavorisées [24, 60, 34].

Pour mieux comprendre comment les facteurs individuels, l'environnement physique et social des quartiers peuvent influencer la santé mentale perçue et le sentiment d'appartenance au quartier des jeunes adultes, il se pose un problème d'estimation du lien entre les deux *outcomes* et les covariables, alors que les données sont structurées hiérarchiquement. Or, dans la littérature en santé, lorsque les données présentent une telle structure, une approche est de recourir aux modèles mixtes. Comment faut-il alors utiliser l'approche par modèle mixte pour mieux comprendre les inégalités sociales dans la santé mentale perçue et dans le sentiment d'appartenance au quartier chez les jeunes adultes ?

3.3 Objectifs et hypothèses

3.3.1 Objectifs

De manière générale, cette recherche a pour but de mieux comprendre les différences de santé mentale perçue et du sentiment d'appartenance au quartier chez les jeunes après avoir contrôlé pour les facteurs individuels et contextuels.

De manière spécifique, cette recherche vise d'une part à déterminer quels facteurs influencent la santé mentale perçue et le sentiment d'appartenance au quartier, d'autre part à déterminer si ces facteurs influencent pareillement ou différemment l'*outcome* bivarié sachant que les deux *outcomes* sont corrélés.

3.3.2 Hypothèses

Grâce à la revue de la littérature sur le sujet, nous allons tester les hypothèses suivantes :

- Les perceptions du quartier ont des influences différentes sur la santé mentale perçue et sur le sentiment d'appartenance au quartier chez les jeunes adultes.
- Le niveau de défavorisation matérielle relative au quartier a plus d'influence sur le sentiment d'appartenance au quartier que sur la santé mentale perçue des jeunes adultes.

3.4 Méthodes

3.4.1 Données et population d'étude

Les données individuelles proviennent des participants à l'étude *CentrÉS* (*CENTRe-ville Équitable et en Santé*) [54].

Le *CentrÉS* est une étude prospective et multidisciplinaire qui vise à mieux comprendre comment l'environnement physique et social des quartiers de Sherbrooke peut influencer le sentiment d'appartenance au quartier, les liens sociaux et le bien-être chez les jeunes adultes et à déterminer si ces éléments divergent entre groupes sociaux.

Les données utilisées dans le cadre du présent mémoire sont des données transversales correspondant à la première vague colligées entre le 7 juillet 2020 et le 31 octobre 2021. La figure 3.1 présente le *Flowchart* illustrant ainsi pour la première vague, le nombre de participants inclus dans la base de données avec les nombres de sorties de l'étude.

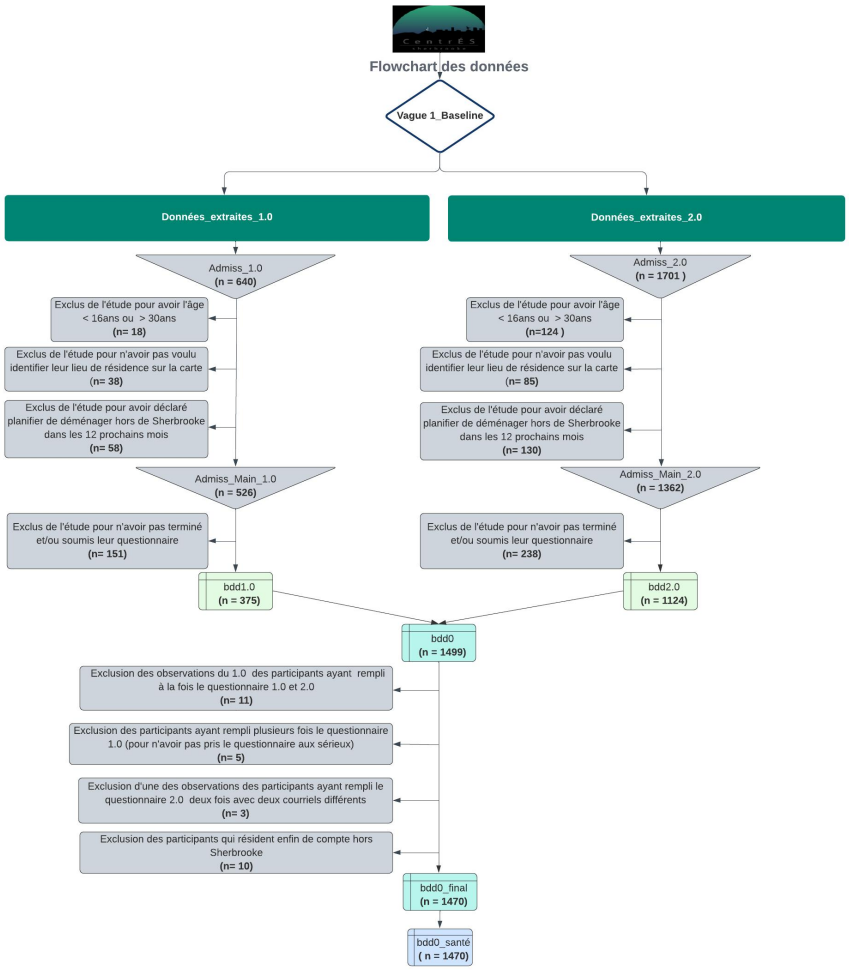

FIGURE 3.1 – *Flowchart*, base de données de l'étude *CentrÉS*

La base de données utilisée contient donc 1470 participants, soient des résidents de Sherbrooke, âgés de 16 à 30 ans au moment de compléter le questionnaire en ligne. Les réponses collectées auprès des participants ont été recueillies à base d'un questionnaire déployé en ligne sur la santé, le quartier et les lieux d'activités.

Les données d'intérêt caractérisant les communautés locales de Sherbrooke, prises comme quartiers dans cette présente étude, sont issues du tableau de bord des communautés de l'Estrie [33]. La figure 3.2 montre la carte de la municipalité de Sherbrooke subdivisée en arrondissements dans lesquels on retrouve ces différentes communautés locales.

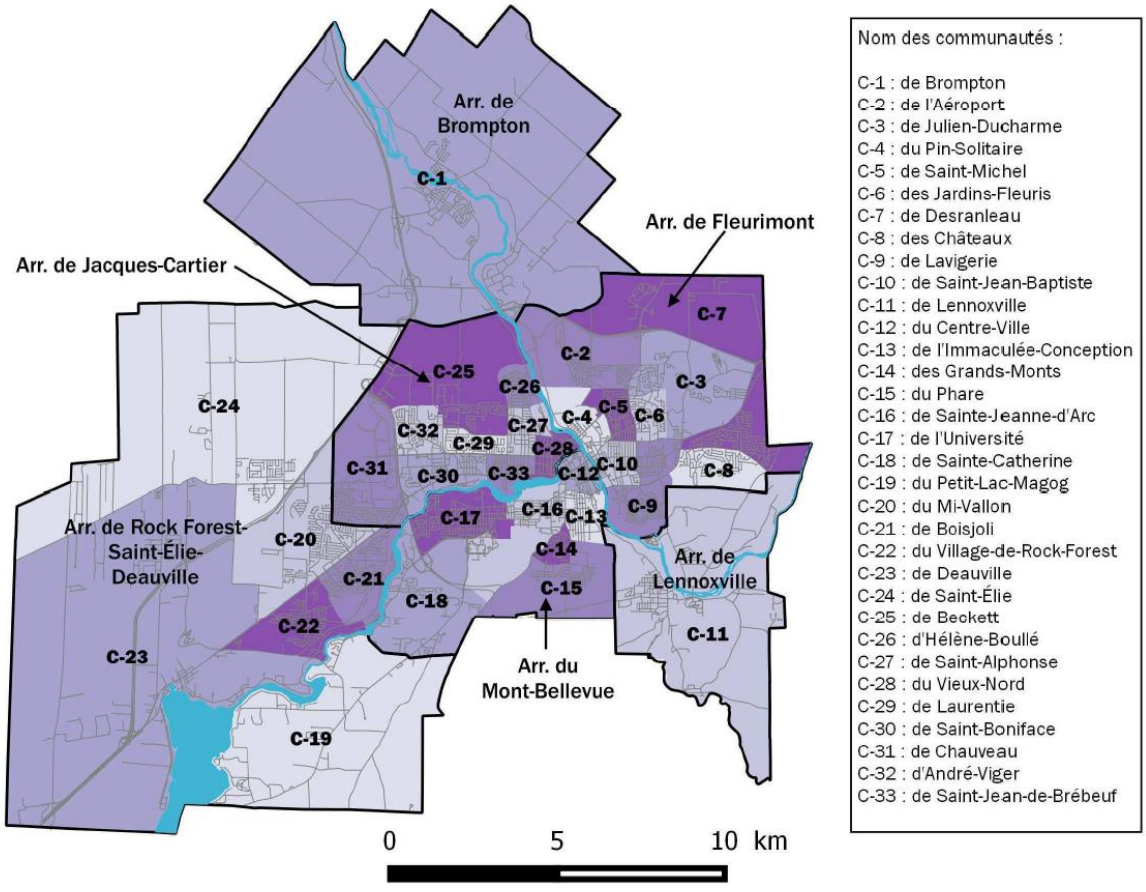

FIGURE 3.2 – Les communautés locales de la ville de Sherbrooke (Québec)

3.4.2 Variables

a. *Outcomes*

Deux *outcomes* sont définis dans cette recherche.

a.1 Score de santé mentale perçue

La santé mentale perçue est une mesure subjective de l'état de santé mentale global. Elle a été recueillie auprès des participants à l'étude *CentrÉS* par une question tirée de l'enquête sur la santé dans les collectivités canadiennes ($ESCC$, GEN_Q005) qui demande d'évaluer la santé mentale actuelle en indiquant si celle-ci est excellente (1), très bonne (2), bonne (3), passable (4), mauvaise (5). Le codage numérique des réponses est considéré ici comme un score exprimant une hiérarchisation des réponses.

a.2 Score du sentiment d'appartenance au quartier

Le sentiment d'appartenance au quartier exprime le sentiment d'affection social des individus et indique l'obligation sociale et la collaboration au sein des quartiers [59]. Il a été recueilli auprès des participants à l'étude *CentrÉS* par des questions portant sur leur quartier de résidence tel qu'ils le perçoivent. Six items du questionnaire de l'étude *CentrÉS* tirés des travaux de Fone DL, et al. [13] et de Greene G et al. [19] et définis en annexe A de ce mémoire, ont permis de calculer le score du sentiment d'appartenance au quartier qui est composite. Ce dernier est obtenu en sommant en ligne les valeurs ordinales observées de ces items, mais l'item 3 est recodé dans l'ordre inverse suivant pour avoir une logique dans le calcul : 5 = tout à fait d'accord ; 4 = d'accord ; 3 = ni d'accord ni en désaccord ; 2 = pas d'accord ; 1 = pas du tout d'accord.

b. Covariables

Nous définissons ici les covariables choisies pour cette présente étude dont leurs présentations se trouvent en annexe B de ce mémoire.

b.1 Au niveau du participant

b.1.1 Variables sociodémographiques et économique

- L'âge : Il est continu et mesuré à partir de la date de naissance du participant, au moment de l'admissibilité à l'étude *CentrÉS*.
- Le genre : Il correspond à l'identité du participant.
- Le niveau d'éducation : C'est le plus haut niveau de scolarité complété.
- Le revenu personnel : C'est le revenu total de l'année dernière du participant avant déductions d'impôts.
- L'appartenance ethnique : Elle identifie l'origine ethnique du participant.

b.1.2 Variables liées à la perception du quartier et au sentiment de sécurité du quartier

Elles sont analysées individuellement. Chacune de leurs réponses est considérée comme un score élémentaire [58]. Nous les avons choisies dans l'analyse, car des études comme celle de TD Hill et al. ont rapporté qu'il y a une association entre le contexte du quartier et la santé mentale [22] ou le sentiment d'appartenance au quartier. Ces variables sont :

- Avoir accès à tout ce dont on a besoin dans mon quartier
- Il y a des choses intéressantes à faire dans mon quartier
- La ville investit dans mon quartier
- Les changements dans mon quartier améliorent ma qualité de vie
- Mon quartier est de plus en plus dynamique
- Les personnes à faible revenu ont du mal à rester dans le quartier
- Je me sens de plus en plus exclus de mon quartier
- Sentiment de sécurité

b.2 Au niveau du quartier

Le niveau de défavorisation matérielle des quartiers de résidence de Sherbrooke est la principale variable contextuelle intégrée dans cette présente étude.

Ce sont des données exogènes se trouvant dans un fichier de données séparé, qu'il a fallu fusionner au fichier de données individuelles. Développé par l'institut national de santé publique du Québec [23], cet indice est calculé sur une base estrienne [33] et est obtenu à partir des trois indicateurs socioéconomiques suivants issus du recensement de 2016 au Canada : le ratio emploi/population chez les 15 ans ou plus, la proportion de personnes de 15 ans ou plus sans certificat ou diplôme d'études secondaires et le revenu moyen des personnes de 15 ans ou plus. Les résultats ont été divisés en quintiles. Le premier quintile représente la population la plus favorisée ($Q1$) et, inversement, le cinquième quintile représente la population la plus défavorisée ($Q5$). L'échelle de valeurs est associée aux : $Q1$ = très favorisé ; $Q2$ = favorisé ; $Q3$ = défavorisation moyenne ; $Q4$ = défavorisation forte ; $Q5$ = défavorisation très forte.

3.4.3 Analyses statistiques

Nous avons réalisé une analyse descriptive dans un premier temps en effectuant le tri-à-plat de toutes les variables d'intérêts. À cet effet, les taux de non-réponses et de valeurs manquantes ont été relevés afin de déterminer la procédure à suivre concernant les données manquantes. Certaines covariables ont été recodées. Nous avons ensuite mené une analyse de corrélation entre les deux *outcomes*.

Dans un second temps, nous avons utilisé l'approche par modèle mixte, notamment le modèle multiniveaux bivarié pour estimer le lien entre l'*outcome* bivarié et les covariables. Nous avons choisi un tel modèle vu la structure des données *CentrÉS*, avec les participants à l'enquête (niveau 1) imbriqués dans leur quartier de résidence (niveau 2), et vu qu'il y a un problème d'estimation du lien entre l'*outcome* bivarié et les covariables. Le modèle linéaire simple ne convient pas à cette analyse puisqu'il ne tient pas compte de la corrélation entre les individus d'un même quartier dans ce contexte.

3.5 Résultats

3.5.1 Analyse descriptive

a. Tris-à-plat et gestion des données manquantes

Les tableaux C.1, C.2, C.3 et C.4 en annexe C de ce mémoire présentent respectivement le profil des 1470 participants à l'étude (C.1), les fréquences de distribution des variables portant sur la perception du quartier (C.2) et sur la sécurité perçue dans le quartier (C.3) puis leur distribution par niveau de défavorisation matérielle des quartiers de résidences à Sherbrooke (C.4). Nous faisons la synthèse suivante :

- L'âge médian [IIQ] des participants au moment de compléter le questionnaire est de 23 $[20-27]$ ans. Parmi ces participants, la majorité ont déclaré être des femmes (70%) et avoir terminé leurs études postsecondaires (75%). Seulement quelques-uns ont répondu n'avoir aucun revenu personnel (4%). La plus grande partie ont déclaré appartenir à l'ethnie caucasienne (82%).
- Plus de la moitié des participants ont déclaré être en accord avec le fait qu'ils ont accès à tout ce dont ils ont besoin dans leur quartier (56%). Moins de la moitié sont généralement en accord avec le fait qu'il y a des choses intéressantes à faire dans leur quartier (44%) ; que la ville investit dans leur quartier (32%) ; que les changements dans leur quartier améliorent leur qualité de vie (27%) ; que leur quartier est de plus en plus dynamique (29%) ; que les personnes à faible revenu ont de la difficulté à rester dans le quartier (35%) et qu'ils se sentent de plus en plus exclus de leur quartier (23%).
- Pour ce qui concerne la sécurité perçue dans le quartier, la plupart de ces participants ont déclaré être en sécurité (70%).
- Un peu moins de la moitié des participants à l'étude semblent résider, en général, dans des quartiers favorisés (47%).

La figure C.1 en annexe C illustre la distribution des fréquences des réponses sur la santé mentale perçue des participants à l'étude. Parmi les 1470 participants, seulement 6% ont répondu qu'ils sont en excellente santé mentale ; 20% en très bonne santé mentale ; 34% en bonne santé mentale ; 28% ont déclaré que leur santé mentale est passable et 12%, une mauvaise santé mentale.

Le codage numérique des réponses sur la santé mentale perçue étant considéré comme des scores exprimant une hiérarchisation des réponses, les statistiques simples sont résumées dans le tableau 3.1. Ainsi, le score moyen de santé mentale perçue des participants à l'étude est de 3.2 (écart-type = 1.1) sur une échelle de mesure de 1 à 5.

Outcome 1 : *score_sante_mentale*							
n	moyenne	écart-type	max	min	1^{er} quartile	médiane	3^{ieme} quartile
1430	3.2	1.1	5	1	2	3	4

TABLEAU 3.1 – Statistiques simples du score de santé mentale perçue

La figure C.2 en annexe C montre la distribution des fréquences pour chacun des 6 items ayant permis de calculer le score composite du sentiment d'appartenance au quartier. Nous avons ainsi les statistiques simples du score du sentiment d'appartenance au quartier dans le tableau 3.2 dont le score moyen est de 17.2 (écart-type = 5.3).

Outcome 2 : *score_appartenance_quartier*							
n	moyenne	écart-type	max	min	1^{er} quartile	médiane	3^{ieme} quartile
1470	17.2	5.3	30	6	13	17	21

TABLEAU 3.2 – Statistiques simples du score d'appartenance au quartier calculé

Le tableau 3.3 montre les variables pour lesquelles il y a des valeurs manquantes et des non-réponses. Parmi les deux *outcomes*, c'est le score de santé mentale perçue qui a des valeurs manquantes, soit 2.7%. Le revenu personnel et l'appartenance ethnique ont des taux de non-réponse respectivement de 5.2% et 2.6%. Les observations ayant des valeurs manquantes ou des non-réponses pour ces variables ne seront pas prises en compte. Après leur suppression, nous avons au total 1323 observations pour la suite de l'analyse.

Variables	Effectif	Valeurs manquantes k (%)	Non réponses k (%)
Score de santé mentale perçue	1470	40 (2.7)	0 (0.0)
Revenu personnel	1470	0 (0.0)	76 (5.2)
Appartenance ethnique	1470	0 (0.0)	38 (2.6)

TABLEAU 3.3 – Valeurs manquantes et non-réponses

b. Recodages de variables

- Recodage de la variable *genre* (*genre_rec*)

Nous avons catégorisé la variable *genre* en hommes (catégories incluses ici), femmes (catégories incluses ici) et autres identifiés de genre (catégories incluses ici). Parmi les 1323 participants obtenus après suppression des valeurs manquantes et des non-réponses, 27% sont des hommes, 70%, des femmes et 3% font partie des autres identifiés de genre.

- Recodage de la variable *niveau d'éducation* (*education_rec*)

Nous avons également catégorisé la variable *niveau d'éducation* en niveau secondaire (catégories incluant ici ceux qui ont le niveau secondaire 4 ou moins et les diplômés d'études secondaires ou équivalent), niveau collégial ou technique (catégories incluant ici ceux qui ont le diplôme ou certificat d'études d'un programme technique au CÉGEP, d'une école de métiers, d'un collège commercial ou privé ou d'un institut technique et les diplômés d'études d'un programme général au CÉGEP), niveau universitaire (catégories incluant ici ceux qui ont obtenu le certificat ou diplôme universitaire de premier cycle, le certificat ou diplôme universitaire de deuxième cycle et le doctorat). Parmi les 1323 participants, 25% ont le niveau secondaire, 34%, le niveau collégial ou technique et 41% le niveau universitaire.

c. Analyse de corrélation entre les deux *outcomes*

Le tableau 3.4 montre le résultat de la corrélation entre les deux *outcomes*. Un test de corrélation de Pearson, sous l'hypothèse nulle ($H_0 : \rho = 0$) montre qu'il existe un lien positif statistiquement significatif, mais faible ($\rho \simeq 17\%$, p-valeur $< 0.0001 << 5\%$) entre le score de santé mentale perçue et le score du sentiment d'appartenance au quartier. L'approche hiérarchique avec un *outcome* bivarié semble être pertinente, ce qui entraînerait des erreurs standards plus petites pour les tests des covariables sur un *outcome* donné par rapport à celle avec un *outcome* univarié. [57]

Coefficient de corrélation de Pearson, $n = 1323$		
	score_sante_mentale	*score_appartenance_quartier*
score_sante_mentale	1.0000	0.1666 (<.0001)
score_appartenance_quartier	0.1666 (<.0001)	1.0000

TABLEAU 3.4 – Corrélation entre les deux *outcomes*

3.5.2 Analyse multiniveaux bivariée

Nous estimons le lien entre l'*outcome* bivarié (le score de santé mentale perçue et le score du sentiment d'appartenance au quartier) et les covariables simultanément par le modèle à deux niveaux en tenant compte de la dépendance des observations résultant de l'imbrication des participants à l'étude dans des quartiers de résidence de Sherbrooke. Chaque série d'analyses se concentrera sur l'estimation et le test des effets d'un facteur pour les deux *outcomes*.

a. Format long des données *CentrÉS*

L'ensemble de données *CentrÉS* initialement dans le format large, est reformaté dans le format long requis. Les scores de santé mentale perçue et ceux du sentiment d'appartenance au quartier sont placés dans une seule colonne, ce qui crée plusieurs enregistrements pour chaque participant à l'étude.

Comme l'échantillon compte 1323 participants et que les données ont été collectées sur deux *outcomes* pour chaque participant, l'ensemble de données nécessaire pour cette modélisation contient $2 \times 1323 = 2646$ enregistrements.
Les figures D.1 et D.2 en annexe D de ce mémoire montrent respectivement les extraits de données *CentrÉS* initialement dans le format large et dans le format long.

Les scores de chaque participant apparaissent sur deux lignes distinctes. Les colonnes contenant les scores de santé mentale perçue et ceux du sentiment d'appartenance au quartier sont remplacés par deux nouvelles variables, $index1$ et $reponse$.
La variable $reponse$ contient les scores de santé mentale perçue et ceux du sentiment d'appartenance au quartier dans une seule colonne. $index1$ indique la séquence de ces deux *outcomes* dans la colonne $reponse$ (c'est-à-dire le score de santé mentale perçue suivi du score du sentiment d'appartenance au quartier). Les $identifiants$ des participants et les covariables sont également au format long, les valeurs pour un cas donné étant répétées dans tous les enregistrements.
Les deux variables indicatrices requises pour cette modélisation sont :

- a_1 codé 1 lorsqu'un enregistrement donné conserve un score pour la santé mentale perçue et 0 sinon,
- a_2 codé 1 lorsqu'un enregistrement donné conserve un score pour le sentiment d'appartenance au quartier et 0 sinon.

b. Intégration des deux *outcomes* dans le modèle

Le modèle de base au niveau du participant est :

$$reponse_j = score\_sante\_mentale_j \; a_{1j} + score\_appartenance\_quartier_j \; a_{2j}, \quad (3.1)$$

- où $reponse_j$ représente la colonne unique intitulée $reponse$ contenant les scores de chaque *outcome* d'un participant j donné,
- les a_{1j} et a_{2j} sont pour un *outcome* d'un participant j donné :

si $a_{1j} = 1$, $a_{2j} = 0$ en raison du codage utilisé. Dans ce cas, l'équation 3.1 devient

$$reponse_j = score\_sante\_mentale_j.$$

si $a_{2j} = 1$, $a_{1j} = 0$ en raison du codage utilisé. L'équation 3.1 devient alors

$$reponse_j = score\_appartenance\_quartier_j.$$

L'équation 3.1 sera utilisée comme équation de base pour toutes les analyses.

c. Estimation du modèle bivarié au niveau du participant

c.1 Estimation du modèle vide

Ce modèle n'inclut aucune covariable. Nous avons les équations suivantes :

$$score\_sante\_mentale_j = \beta_{10} + e_{1j}, \tag{3.2}$$

$$score\_appartenance\_quartier_j = \beta_{20} + e_{2j} \tag{3.3}$$

avec $reponse_j = score\_sante\_mentale_j \; a_{1j} + score\_appartenance\_quartier_j \; a_{2j}$,

où β_{10} et β_{20} représentent respectivement la moyenne du score de santé mentale perçue et la moyenne du score du sentiment d'appartenance au quartier.
Les termes résiduels e_{1j} et e_{2j} sont supposés suivre une distribution normale bivariée avec une moyenne attendue de 0 et une certaine variance-covariance.

Nous avons au total 5 paramètres à estimer pour ce modèle vide :

- les deux effets fixes β_{10} et β_{20},
- la variance de e_{1j} ($\sigma^2_{e_1}$), la variance de e_{2j} ($\sigma^2_{e_2}$) et la covariance de e_{1j} et e_{2j} ($\text{Cov}(e_1, e_2)$).

Le tableau D.3 en annexe D montre l'estimation de ces 5 paramètres :
Les moyennes du score de santé mentale perçue et du sentiment d'appartenance au quartier, telles qu'elles apparaissent dans la table des effets fixes, sont respectivement $\widehat{\beta}_{10} = 3.2101$ et $\widehat{\beta}_{20} = 17.2525$.

L'estimation des variances $\sigma^2_{e_1}$ et $\sigma^2_{e_2}$ sont respectivement $\widehat{\sigma}^2_{e_1} = 1.1516$ et $\widehat{\sigma}^2_{e_2} = 27.6339$ comme le montre la matrice $\boldsymbol{R}$ estimée pour le participant d'identifiant = 100.
L'estimation de la covariance $\mathrm{Cov}(e_1, e_2)$ est, $\widehat{\mathrm{Cov}}(e_1, e_2) = 0.9401$ comme le montrent les tables de la valeur estimée du paramètre de covariance et la matrice de covariance $\boldsymbol{R}$ ou résiduelle.
Les écarts types du score de santé mentale perçue et du score du sentiment d'appartenance au quartier sont, respectivement, $\sqrt{\widehat{\sigma}^2_{e_1}} = 1.0731$ et $\sqrt{\widehat{\sigma}^2_{e_2}} = 5.2568$. La corrélation estimée des résidus est

$$\widehat{\rho}(e_{1j}, e_{2j}) = \frac{\widehat{\mathrm{Cov}}(e_1, e_2)}{\sqrt{\widehat{\sigma}^2_{e_1}\widehat{\sigma}^2_{e_2}}} = \frac{0.9401}{1.0731 \times 5.2568} = 0.1667.$$

Cette valeur de la corrélation des résidus indique que le score de santé mentale perçue et le score du sentiment d'appartenance au quartier sont positivement corrélées. La déviance du modèle pour les 5 paramètres est $-2l = 12049.6$ (comme le montre le tableau D.3) et reflète l'ajustement du modèle. Elle sera comparée à la déviance obtenue lorsqu'une covariable sera ajoutée au modèle vide pour déterminer si l'ajustement s'améliore.

c.2 Inclusion d'une covariable dans le modèle au niveau du participant

Nous introduisons la variable *ethnie* dans le modèle vide. L'appartenance ethnique a souvent un effet significatif dans l'analyse d'une part de l'espace social et d'autre part du champ de la santé mentale d'un individu [3]. L'objectif ici est de déterminer s'il existe des effets de l'appartenance ethnique pour tout *outcome*. Si ces effets sont présents, alors ils seront examinés pour chaque *outcome* ainsi que les résultats des tests. Pour ce nouveau modèle, au niveau du participant, les équations 3.2 et 3.3 sont modifiées. Nous avons :

$$score\_sante\_mentale_j = \beta_{10} + \beta_{11}\ ethnie_j + e_{1j}, \tag{3.4}$$

$$score\_appartenance\_quartier_j = \beta_{20} + \beta_{21}\ ethnie_j + e_{2j} \tag{3.5}$$

avec $reponse_j = score\_sante\_mentale_j\ a_{1j} + score\_appartenance\_quartier_j\ a_{2j}$,

où $ethnie_j$ représente l'appartenance ethnique pour le participant j.
β_{10} et β_{20} représentent respectivement la moyenne pour le score de santé mentale perçue et la moyenne pour le score du sentiment d'appartenance au quartier.
β_{11} et β_{21} représentent la différence moyenne entre les participants du groupe d'ethnie non caucasienne et du groupe d'ethnie caucasienne pour le score de santé mentale perçue et pour le score du sentiment d'appartenance au quartier, respectivement.
Les termes résiduels e_{1j} et e_{2j} sont supposés suivre une distribution normale bivariée, avec une matrice de variance-covariance égale entre les groupes des participants.
L'hypothèse nulle pour le test de la variable *ethnie* est $H_0 : \beta_{11} = \beta_{21} = 0$. Elle est testée en utilisant les déviances par rapport à ce modèle et au modèle vide.

Pour ce modèle, nous allons estimer 7 paramètres :

- 4 effets fixes : β_{10}, β_{20} , β_{11} et β_{21} dans les équations 3.4 et 3.5,
- 3 éléments de la variance-covariance : la variance de e_{1j} ($\sigma^2_{e_1}$), la variance de e_{2j} ($\sigma^2_{e_2}$) et la covariance de e_{1j} et e_{2j} ($\mathrm{Cov}(e_1, e_2)$).

Pour inclure la variable *ethnie* dans le modèle de façon à ce que des effets distincts de cette variable soient estimés pour chaque *outcome*, celle-ci a été multipliée par $index1$. Pour estimer β_{11} et β_{21} le terme $ethnie_j \times index1$ a été ajouté à l'instruction du modèle. Le tableau D.4 en annexe D montre les résultats des effets de la variable *ethnie* introduite dans le modèle vide. La déviance pour ce modèle est $-2l = 12042.5$, avec 7 paramètres estimés. La déviance du modèle vide était $-2l = 12049.6$ avec 5 paramètres estimés.
Le test global pour l'hypothèse nulle qu'aucun effet de la variable *ethnie* n'est présent pour aucun des deux *outcomes* ($H_0 : \beta_{11} = \beta_{21} = 0$) est testée en calculant la différence de ces déviances qui est distribuée sous la forme d'une valeur de khi-deux ayant un degré de liberté d égale à la différence du nombre de paramètres estimés pour ces modèles, c'est-à-dire $d = 7 - 5 = 2$. Ce test de déviance est utilisé ici, car le modèle vide peut être obtenu à partir du modèle actuel en contraignant les effets de la variable *ethnie* à 0.

Le calcul de la différence entre les déviances du modèle donne une valeur de khi-deux de $\Delta(-2l) = 12049.6 - 12042.5 = 7.1$, ce qui est statistiquement significatif (p-valeur $< 5\%$) car cette valeur dépasse la valeur critique du khi-deux $= 5.99$ ($\alpha = 5\%$, $d = 2$).

Puisque le rejet de l'hypothèse nulle globale bivariée suggère que les effets de l'appartenance ethnique sont présents pour au moins l'un des deux *outcomes*, nous considérons maintenant les estimations et les résultats des tests statistiques de l'effet de la variable *ethnie* pour chaque *outcome*. Les effets de l'appartenance ethnique sont donc $\widehat{\beta}_{11} = -0.2074$ (erreur-type $= 0.08133$) pour le score de santé mentale perçue et $\widehat{\beta}_{21} = 0.1429$ (erreur-type $= 0.3994$) pour le score du sentiment d'appartenance au quartier, pour lesquels le groupe d'ethnie non caucasienne est la référence.

Les statistiques t d'environ -2.55 (p-valeur $< 5\%$) et 0.36 (p-valeur $> 5\%$), respectivement, pour le score de santé mentale perçue et pour le score du sentiment d'appartenance au quartier, suggèrent la présence d'effet significatif de l'appartenance ethnique seulement sur le score de santé mentale perçue des participants à l'étude. Ainsi, pour le score de santé mentale perçue, la moyenne du groupe des caucasiens pour laquelle le groupe des non-caucasiens est la référence donne :

$$(\overline{score\_sante\_mentale}_j)_{caucasiens} = 3.3854 - 0.2074(1) = 3.1780.$$

Si le groupe des caucasiens est pris comme référence, alors la moyenne du groupe des non-caucasiens sera :

$$(\overline{score\_sante\_mentale}_j)_{non\ caucasiens} = 3.1780 + 0.2074(1) = 3.3854.$$

Le groupe des caucasiens a tendance à avoir une meilleure santé mentale (3.1780) que le groupe des non-caucasiens (3.3854). L'estimation des variances $\sigma^2_{e_1}$ et $\sigma^2_{e_2}$ sont respectivement $\widehat{\sigma}^2_{e_1} = 1.1460$ et $\widehat{\sigma}^2_{e_2} = 27.6312$. Ces estimations ont légèrement diminué par rapport à celles du modèle vide, ceci dû à l'introduction de la variable *ethnie*.

L'estimation de la covariance $\mathrm{Cov}(e_1, e_2)$ est $\widehat{\mathrm{Cov}}(e_1, e_2) = 0.9440$.

Les écarts types à nouveau du score de santé mentale perçue et du score du sentiment d'appartenance au quartier sont, respectivement, $\sqrt{\widehat{\sigma}^2_{e_1}} = 1.0705$ et $\sqrt{\widehat{\sigma}^2_{e_2}} = 5.2565$.
La nouvelle corrélation estimée des résidus est donc

$$\widehat{\rho}(e_{1j}, e_{2j}) = \frac{\widehat{\text{Cov}}(e_1, e_2)}{\sqrt{\widehat{\sigma}^2_{e_1}\widehat{\sigma}^2_{e_2}}} = \frac{0.9440}{1.0705 \times 5.2565} = 0.1678.$$

c.3 Vérifier si l'effet de l'appartenance ethnique diffère selon les *outcomes*

Cette étape de l'analyse vérifie si l'effet de l'appartenance ethnique est de la même ampleur pour chaque *outcome*. Nous avons d'abord contraint les effets fixes à être égaux, puis tester la différence d'ajustement en utilisant les déviances entre ce modèle contraint et le modèle où les effets sont estimés librement.
Dans les équations 3.4 et 3.5, les effets fixes de la variable *ethnie* sont librement estimés (sans contrainte) pour les deux *outcomes* : le score de santé mentale perçue (β_{11}) et le score du sentiment d'appartenance au quartier (β_{21}). Nous allons tester maintenant si ces effets fixes sont identiques ou différents pour les deux *outcomes*.
Le modèle utilisé est essentiellement le même que celui de l'analyse précédente, mais un effet commun présumé de la variable *ethnie* sera estimé. Pour estimer ce modèle, nous avons remplacé le terme $ethnie_j \times index1$ du modèle précédent par $ethnie_j$.

Le nombre de paramètres à estimer est maintenant de 6, comprenant :

- 3 effets fixes : β_{10}, β_{20} et $\beta_{11} = \beta_{21}$ (une seule estimation de l'effet de la variable *ethnie*),
- 3 éléments de la matrice de variance-covariance : la variance de e_{1j} ($\sigma^2_{e_1}$), la variance de e_{2j} ($\sigma^2_{e_2}$) et la covariance de e_{1j} et e_{2j} ($\text{Cov}(e_1, e_2)$).

Le tableau D.5 en annexe D montre ces estimations. La déviance associée à ce modèle contraint d'effets de l'appartenance ethnique est $-2l = 12043.3$, ce qui n'est que légèrement plus grand (reflétant un moins bon ajustement) que le modèle précédent qui fournissait des estimations séparées de ces effets.

Plus précisément, la différence de déviances des modèles, qui est distribuée sous la forme d'une valeur de khi-deux, est $\Delta(-2l) = 12043.3 - 12042.5 = 0.8$, ce qui ne dépasse pas la valeur critique du khi-deux $= 3.84$ ($\alpha = 5\%, d = 7 - 6 = 1$). Ainsi, ce test ne suggère pas que ces deux modèles ont un ajustement différent. Il existe des preuves soutenant l'hypothèse selon laquelle l'effet de l'appartenance ethnique est similaire pour les scores de santé mentale perçue et du sentiment d'appartenance au quartier. L'effet de l'appartenance ethnique commun est estimé à $\widehat{\beta}_{11} = \widehat{\beta}_{21} = -0.2047$ (erreur-standard $= 0.08128$) et est statistiquement significatif (p-valeur $< 5\%$). Les estimations de β_{10} et de β_{20} sont maintenant $\widehat{\beta}_{10} = 3.3831$ et $\widehat{\beta}_{20} = 17.4255$.

L'estimation des variances de e_{1j} et de e_{2j} n'ont quasiment pas changé. Elles sont respectivement $\widehat{\sigma}^2_{e_1} = 1.1460$ et $\widehat{\sigma}^2_{e_2} = 27.6470$. L'estimation de la covariance $\mathrm{Cov}(e_1, e_2)$ a presque la même valeur, $\widehat{\mathrm{Cov}}(e_1, e_2) = 0.9439$.

d. Estimation du modèle bivarié à deux niveaux

Dans cette partie de l'analyse, nous allons tenir compte de la dépendance au sein du quartier et entre les quartiers en ajoutant maintenant dans le modèle précèdent le niveau 2 (niveau du quartier). Les participants à l'étude sont imbriqués dans l'une des 36 communautés locales de Sherbrooke (Québec) prises comme des quartiers de résidence.

d.1 Estimation du modèle à deux niveaux avec l'appartenance ethnique

L'équation 3.1 est légèrement modifiée ici afin de reconnaître l'inclusion du niveau du quartier. Le modèle de base devient :

$$reponse_{jk} = score\_sante\_mentale_{jk}\ a_{1jk} + score\_appartenance\_quartier_{jk}\ a_{2jk}, \qquad (3.6)$$

où $reponse_{jk}$ représente la colonne unique intitulée *reponse* contenant les scores de chaque *outcome* pour un participant j dans un quartier k donné.

Les a_{1jk} et a_{2jk} sont pour un *outcome* d'un participant j dans un quartier k donné.

- Au niveau 1 (niveau du participant), le modèle sans covariables a pour équations :

$$score\_sante\_mentale_{jk} = \beta_{10k} + e_{1jk}, \tag{3.7}$$

$$score\_appartenance\_quartier_{jk} = \beta_{20k} + e_{2jk}, \tag{3.8}$$

avec $reponse_{jk} = score\_sante\_mentale_{jk}\ a_{1jk} + score\_appartenance\_quartier_{jk}\ a_{2jk}$,

où β_{10k} et β_{20k} représentent la moyenne pour un quartier k donné pour le score de santé mentale perçue et pour le score du sentiment d'appartenance au quartier, respectivement.
Les termes résiduels au niveau du participant résidant dans un quartier, e_{1jk} et e_{2jk} sont supposés suivre une distribution normale bivariée, avec une moyenne attendue de 0 et des variances $\sigma^2_{e_1}$, $\sigma^2_{e_2}$ et une covariance $\text{Cov}(e_1, e_2)$.
L'affectation de l'appartenance ethnique peut varier d'un quartier de résidence à l'autre. La variable *ethnie* peut alors apparaître dans le modèle au niveau du quartier.

- Au niveau 2 (niveau du quartier) nous avons les équations suivantes du modèle :

$$\beta_{10k} = \gamma_{100} + \gamma_{101}\ ethnie_k + u_{10k}, \tag{3.9}$$

$$\beta_{20k} = \gamma_{200} + \gamma_{201}\ ethnie_k + u_{20k}, \tag{3.10}$$

où γ_{100} et γ_{200} représentent la moyenne générale pour le score de santé mentale perçue et pour le score du sentiment d'appartenance au quartier, respectivement.
Les paramètres clés γ_{101} et γ_{201} représentent les différences de moyenne entre le groupe d'ethnie caucasienne et le groupe d'ethnie non caucasienne pour le score de santé mentale perçue et pour le score du sentiment d'appartenance au quartier, respectivement.
Les termes résiduels au niveau du quartier sont u_{10k} et u_{20k} et sont supposés suivre une distribution normale bivariée avec une moyenne attendue de 0 et des variances constantes ($\sigma^2_{u_1}$ et $\sigma^2_{u_2}$), et une covariance ($\text{Cov}(u_1, u_2)$).

Nous avons 10 paramètres à estimer pour le modèle décrit par les équations 3.6 à 3.10 :

- 4 effets fixes : γ_{100}, γ_{200}, γ_{101}, γ_{201},
- 6 éléments de variance-covariance : $\sigma^2_{e_1}$, $\sigma^2_{e_2}$, $\text{Cov}(e_1, e_2)$, $\sigma^2_{u_1}$ et $\sigma^2_{u_2}$ et $\text{Cov}(u_1, u_2)$.

Le tableau D.6 en annexe D montre les estimations de ces 10 paramètres :
Le résultat des effets fixes montre que chez les participants appartenant à l'ethnie caucasienne, les scores diminuent en moyenne de 0.2145 pour la santé mentale perçue ($\hat{\gamma}_{101} = -0.2145$, p-valeur $< 5\%$) que chez les participants non caucasiens. Ces scores augmentent en moyenne de 0.0719 chez les participants caucasiens que chez les participants non caucasiens pour le sentiment d'appartenance au quartier, mais la différence n'est pas statistiquement significative ($\hat{\gamma}_{201} = 0.0719$, p-valeur $> 5\%$).
Les estimations des variances et covariances apparaissent dans les tables des paramètres de covariance et dans les matrices $\boldsymbol{R}$ et $\boldsymbol{G}$, soient

$$\boldsymbol{R} = \begin{pmatrix} \widehat{\sigma}^2_{e_1} & \widehat{\mathrm{Cov}}(e_1, e_2) \\ \widehat{\mathrm{Cov}}(e_2, e_1) & \widehat{\sigma}^2_{e_2} \end{pmatrix} = \begin{pmatrix} 1.1431 & 0.8579 \\ 0.8579 & 26.5572 \end{pmatrix},$$

$$\boldsymbol{G} = \begin{pmatrix} \widehat{\sigma}^2_{u_1} & \widehat{\mathrm{Cov}}(u_1, u_2) \\ \widehat{\mathrm{Cov}}(u_2, u_1) & \widehat{\sigma}^2_{u_2} \end{pmatrix} = \begin{pmatrix} 0.002401 & 0.08512 \\ 0.08512 & 1.1270 \end{pmatrix}$$

Après avoir pris en compte l'appartenance ethnique au niveau du quartier, les résultats montrent que la plus grande partie de la variabilité se situe au sein des quartiers pour le score du sentiment d'appartenance au quartier uniquement. Environ 4.07% de la variabilité est due à l'hétérogénéité des quartiers. En effet, on a :

$$\frac{\widehat{\sigma}^2_{u_2}}{\widehat{\sigma}^2_{u_2} + \widehat{\sigma}^2_{e_2}} = \frac{1.1270}{1.1270 + 26.5572} \simeq 4.07\%.$$

Pour le score de santé mentale perçue, la part de variabilité imputable au niveau du quartier est quasiment nulle, soit

$$\frac{\widehat{\sigma}^2_{u_1}}{\widehat{\sigma}^2_{u_1} + \widehat{\sigma}^2_{e_1}} = \frac{0.002401}{0.002401 + 1.1431} \simeq 0.21\%.$$

d.2 Estimation du modèle à deux niveaux avec multiples prédicteurs

Nous incluons maintenant dans le modèle outes les autres covariables et l'hypothèse nulle bivariée de l'absence d'interaction entre l'âge et le genre des participants à l'étude.

L'équation 3.6 est modifiée pour inclure les covariables décrites dans cette recherche. Du fait d'un grand nombre de covariables, nous écrivons les équations de la manière suivante :

$$score\_sante\_mentale_{jk} = \beta_{1(0)k} + \sum_{(h)=1}^{12} \beta_{1(h)k}\ x_{(h)jk} + e_{1jk}, \tag{3.11}$$

$$score\_appartenance\_quartier_{jk} = \beta_{2(0)k} + \sum_{(h)=1}^{12} \beta_{2(h)k}\ x_{(h)jk} + e_{1jk}, \tag{3.12}$$

avec $reponse_{jk} = score\_sante\_mentale_{jk}\ a_{1jk} + score\_appartenance\_quartier_{jk}\ a_{2jk}$,

où les vecteurs de paramètres $(\beta_{1(1)k}, ..., \beta_{1(12)k})'$ et $(\beta_{2(1)k}, ..., \beta_{2(12)k})'$ sont des effets fixes respectivement pour le score de santé mentale perçue et pour le score du sentiment d'appartenance au quartier du participant j dans le quartier k,

le vecteur de variables $\begin{pmatrix} x_{(1)jk} \\ x_{(2)jk} \\ x_{(3)jk} \\ x_{(4)jk} \\ x_{(5)jk} \\ x_{(6)jk} \\ x_{(7)jk} \\ x_{(8)jk} \\ x_{(9)jk} \\ x_{(10)jk} \\ x_{(11)jk} \\ x_{(12)jk} \end{pmatrix} = \begin{pmatrix} genre\_rec \\ education\_rec \\ ethnie \\ revenu \\ acces\_tout\_besoin\_quartier \\ choses\_interessant\_quartier \\ ville\_investit\_quartier \\ changement\_qualite\_vie\_quartier \\ plus\_dynamique\_quartier \\ revenu\_difficulte\_quartier \\ sentir\_exclus\_quartier \\ securite\_percue \end{pmatrix}$ est associé

à $(\beta_{1(1)k}, ..., \beta_{1(12)k})'$ et à $(\beta_{2(1)k}, ..., \beta_{2(12)k})'$, respectivement pour le score de santé mentale perçue et pour le score du sentiment d'appartenance au quartier.
Les termes résiduels au niveau 1, e_{1jk} et e_{2jk} sont supposés suivre une distribution normale bivariée avec une moyenne attendue de 0 et une certaine variance-covariance.

Au niveau 2, chacun des coefficients de régression des équations 3.11 et 3.12, peut être considéré comme une variable dépendante à modéliser.

Cependant, on suppose que l'association entre chacune des covariables et chacun des deux *outcomes* est la même dans tous les quartiers, donc $\beta_{1(1)k}$, $\beta_{1(2)k}$, ..., $\beta_{1(12)k}$ et $\beta_{2(1)k}$, $\beta_{2(2)k}$, ..., $\beta_{2(12)k}$ sont modélisés comme des effets fixes dans le modèle au niveau 1.
La modélisation de l'interaction entre l'âge et le genre nous amène à ajouter la variable *age* dans le modèle pour β_{11k} et β_{21k}. De plus, le niveau de défavorisation matérielle ($defavorisation$) des quartiers de résidence de Sherbrooke affecté aux participants est inclus dans le modèle pour β_{10k} et β_{20k} afin qu'il puisse servir de prédicteur pour ces derniers. Les différentes équations correspondantes à ce modèle au niveau 2 sont :

$$\beta_{1(0)k} = \gamma_{1(0)0} + \gamma_{1(0)1}\ age_k + \gamma_{1(0)2}\ defavorisation_k + u_{1(0)k}, \tag{3.13}$$

$$\beta_{1(1)k} = \gamma_{1(1)0} + \gamma_{1(1)1}\ age_k, \tag{3.14}$$

$$\beta_{1(2)k} = \gamma_{1(2)0},\ \beta_{1(3)k} = \gamma_{1(3)0}, ..., \beta_{1(12)k} = \gamma_{1(12)0}, \tag{3.15}$$

$$\beta_{2(0)k} = \gamma_{2(0)0} + \gamma_{2(0)1}\ age_k + \gamma_{2(0)2}\ defavorisation_k + u_{2(0)k}, \tag{3.16}$$

$$\beta_{2(1)k} = \gamma_{2(1)0} + \gamma_{2(1)1}\ age_k, \tag{3.17}$$

$$\beta_{2(2)k} = \gamma_{2(2)0},\ \beta_{2(3)k} = \gamma_{2(3)0}, ..., \beta_{2(12)k} = \gamma_{2(12)0}. \tag{3.18}$$

Il n'y a pas de termes résiduels inclus dans les équations 3.14 et 3.17, ce qui suggère que toute variabilité systématique entre les quartiers (si elle est significative) pour la santé mentale perçue et pour le sentiment d'appartenance au quartier des hommes et des femmes, est due à l'âge. Ainsi, les équations 3.13 à 3.18 ont deux termes résiduels, $u_{1(0)k}$ et $u_{2(0)k}$ qui sont supposés suivre une distribution normale bivariée avec une moyenne attendue de 0 et certaines variances-covariance constantes.

Nous pouvons reconnaître les coefficients qui représentent l'interaction entre l'âge et le genre en formant des équations pour le score de santé mentale perçue et pour le score du sentiment d'appartenance au quartier séparément avec les expressions 3.13 à 3.18.

Les paramètres $\gamma_{1(1)1}$ et $\gamma_{2(1)1}$ représentent ainsi les effets d'interaction entre l'âge et le genre (interaction entre niveaux croisés) pour le score de santé mentale perçue et pour le score du sentiment d'appartenance au quartier, respectivement. Les paramètres $\beta_{1(1)k}$ et $\beta_{2(1)k}$ ne reflètent que les différences du genre à l'intérieur du quartier sur les deux *outcomes*, respectivement.

Dans ce modèle, les principaux paramètres à estimer sont :

- 16 effets fixes pour l'*outcome* 1 ($\gamma_{1(0)0}$, $\gamma_{1(0)1}$, $\gamma_{1(0)2}$, $\gamma_{1(1)0}$, $\gamma_{1(1)1}$, $\gamma_{1(2)0}$, $\gamma_{1(3)0}$,...,$\gamma_{1(12)0}$),
- 16 effets fixes pour l'*outcome* 2 ($\gamma_{2(0)0}$, $\gamma_{2(0)1}$, $\gamma_{2(0)2}$, $\gamma_{2(1)0}$, $\gamma_{2(1)1}$, $\gamma_{2(2)0}$, $\gamma_{2(3)0}$,...,$\gamma_{2(12)0}$),
- 6 éléments de variance-covariance, avec trois de ces termes à chacun des niveaux du participant et du quartier ($\sigma^2_{e_1}$, $\sigma^2_{e_2}$, $\text{Cov}(e_1, e_2)$), ($\sigma^2_{u_1}$, $\sigma^2_{u_2}$, $\text{Cov}(u_1, u_2)$).

Le tableau D.7 en annexe D montre les estimations de ces paramètres. L'amélioration significative de l'ajustement obtenue en tenant compte de l'interaction entre l'âge et le genre pour les deux *outcomes* suggère la présence des effets d'interaction entre l'âge et le genre pour au moins l'un des deux *outcomes*. L'examen de sorties des solutions pour les effets fixes montre que l'estimation ponctuelle de ces effets d'interaction pour le score de santé mentale perçue est de -0.0525 (erreur-standard $= 0.0161$) et pour le score du sentiment d'appartenance au quartier est de 0.0584 (erreur-standard $= 0.0616$) chez les femmes et pour lesquelles les hommes sont le groupe de référence. Les statistiques t correspondantes, -3.2600 et 0.9500 et les p-valeurs dont pour le score de santé mentale perçue est inférieure au seuil 5% et pour le score du sentiment d'appartenance au quartier est supérieure au seuil 5%, suggèrent que l'interaction entre l'âge et le genre est statistiquement significative pour le score de santé mentale perçue mais ne l'est pas pour le score du sentiment d'appartenance au quartier. Le coefficient du terme d'interaction (-0.0525) pour le score de santé mentale perçue représente la différence de l'effet de l'âge sur le score de santé mentale perçue entre les hommes et les femmes. L'effet de l'âge sur le score de santé mentale perçue est plus faible parmi les femmes que parmi les hommes.

- Pour le score de santé mentale perçue, les 16 effets fixes estimés sont résumés dans le tableau 3.5. Les effets statistiquement significatifs au seuil p $< 5\%$ sont indiqués par un astérisque (*). Si on s'intéresse aux participants à l'étude âgés de 30 ans, l'écart estimé du score de santé mentale perçue des femmes par rapport aux hommes serait : $\hat{\gamma}_{1(1)0} + \hat{\gamma}_{1(1)1}(age)_k = 1.5087 - 0.0525(30) = 0.0663$. A l'âge de 30 ans, les femmes ont en moyenne un score de santé mentale perçue inférieur de 0.0663 par rapport aux hommes du même âge et une baisse du score de santé mentale perçue indique une tendance vers une meilleure santé mentale.

Effets fixes	Estimations
$Intercept$	$\hat{\gamma}_{1(0)0} = 2.3295$*
age	$\hat{\gamma}_{1(0)1} = 0.0482$
$genre\_rec$ (femmes, hommes = ref)	$\hat{\gamma}_{1(1)0} = 1.5087$*
$age \times genre\_rec$ (femmes, hommes = ref)	$\hat{\gamma}_{1(1)1} = -0.0525$*
$education\_rec$ (secondaire, universitaire = ref)	$\hat{\gamma}_{1(2)0} = 0.0981$
$ethnie$ (caucasiens, non caucasiens = ref)	$\hat{\gamma}_{1(3)0} = -0.1769$*
$revenu$ (100000$ et plus, sans revenu personnel = ref)	$\hat{\gamma}_{1(4)0} = -0.6770$*
$acces\_tout\_besoin\_quartier$	$\hat{\gamma}_{1(5)0} = -0.0209$
$choses\_interessant\_quartier$	$\hat{\gamma}_{1(6)0} = -0.0067$
$ville\_investit\_quartier$	$\hat{\gamma}_{1(7)0} = 0.1103$*
$changement\_qualite\_vie\_quartier$	$\hat{\gamma}_{1(8)0} = 0.0033$
$plus\_dynamique\_quartier$	$\hat{\gamma}_{1(9)0} = 0.0287$
$revenu\_difficulte\_quartier$	$\hat{\gamma}_{1(10)0} = -0.0171$
$sentir\_exclus\_quartier$	$\hat{\gamma}_{1(11)0} = -0.1566$*
$securite\_percue$	$\hat{\gamma}_{1(12)0} = 0.0159$
$defavorisation$	$\hat{\gamma}_{1(0)2} = 0.0216$

ref : référence, (*) : effet statistiquement significatif au seuil p $< 5\%$

TABLEAU 3.5 – Estimation des effets fixes pour le score de santé mentale perçue

- Pour le score du sentiment d'appartenance au quartier, les 16 effets fixes estimés sont résumés dans le tableau 3.6. Les effets statistiquement significatifs au seuil p $< 5\%$ sont indiqués par un astérisque (*).

Effets fixes	Estimations
Intercept	$\hat{\gamma}_{2(0)0} = 11.8442$*
age	$\hat{\gamma}_{2(0)1} = -0.1102$
genre_rec (femmes, hommes = *ref*)	$\hat{\gamma}_{2(1)0} = -1.1868$
age × *genre_rec* (femmes, hommes = *ref*)	$\hat{\gamma}_{2(1)1} = 0.0584$
education_rec (secondaire, universitaire = *ref*)	$\hat{\gamma}_{2(2)0} = -0.7789$*
ethnie (caucasiens, non caucasiens = *ref*)	$\hat{\gamma}_{2(3)0} = 0.1307$
revenu (50000$ à 99999$, sans revenu personnel = *ref*)	$\hat{\gamma}_{2(4)0} = 1.5553$*
acces_tout_besoin_quartier	$\hat{\gamma}_{2(5)0} = 0.2268$*
choses_interessant_quartier	$\hat{\gamma}_{2(6)0} = 0.7739$*
ville_investit_quartier	$\hat{\gamma}_{2(7)0} = 0.1446$
changement_qualite_vie_quartier	$\hat{\gamma}_{2(8)0} = 0.6557$
plus_dynamique_quartier	$\hat{\gamma}_{2(9)0} = 1.0404$
revenu_difficulte_quartier	$\hat{\gamma}_{2(10)0} = 0.7972$
sentir_exclus_quartier	$\hat{\gamma}_{2(11)0} = -1.5494$*
securite_percue	$\hat{\gamma}_{2(12)0} = 0.3600$
defavorisation	$\hat{\gamma}_{2(0)2} = 0.3067$*

ref : référence, (*) : effet statistiquement significatif au seuil p $< 5\%$

TABLEAU 3.6 – Estimation des effets fixes pour le score du sentiment d'appartenance au quartier

L'intersection des deux parties de résultats montre qu'il y a des covariables significativement liées à l'*outcome* bivarié. Ce sont les covariables *revenu* et *sentir_exclus_quartier* qui ont simultanément des effets significatifs sur le score de santé mentale perçue et sur le score du sentiment d'appartenance au quartier.

L'effet significatif de la covariable *revenu* sur le score de santé mentale perçue est celui du *revenu* = 100000$ et plus ($\hat{\gamma}_{1(4)0} = -0.6770$) pour lequel le groupe des jeunes sans revenu personnel est la référence.

Pour le score du sentiment d'appartenance au quartier, l'effet significatif de la covariable *revenu* est celui du *revenu* = 50000$ et 99999$ ($\hat{\gamma}_{1(4)0} = 1.5553$) pour lequel le groupe des jeunes sans revenu personnel est la référence.

Pour la covariable *sentir_exclus_quartier*, son effet sur le score de santé mentale perçue est en moyenne moins faible ($\hat{\gamma}_{1(11)0} = -0.1566$). Il est plus faible en moyenne ($\hat{\gamma}_{2(11)0} = -1.5494$) sur le score du sentiment d'appartenance au quartier. Les jeunes adultes qui se sentent de plus en plus exclus de leurs quartiers ont tendance à percevoir en moyenne un plus faible sentiment d'appartenance au quartier.

- Les 6 éléments de variance-covariance estimés sont dans les matrices $\boldsymbol{R}$ et $\boldsymbol{G}$ suivantes :

$$\boldsymbol{R} = \begin{pmatrix} \widehat{\sigma}^2_{e_1} & \widehat{\mathrm{Cov}}(e_1, e_2) \\ \widehat{\mathrm{Cov}}(e_2, e_1) & \widehat{\sigma}^2_{e_2} \end{pmatrix} = \begin{pmatrix} 1.0788 & 0.2899 \\ 0.2899 & 15.6460 \end{pmatrix}$$

et

$$\boldsymbol{G} = \begin{pmatrix} \widehat{\sigma}^2_{u_1} & \widehat{\mathrm{Cov}}(u_1, u_2) \\ \widehat{\mathrm{Cov}}(u_2, u_1) & \widehat{\sigma}^2_{u_2} \end{pmatrix} = \begin{pmatrix} 0.008801 & -0.00446 \\ -0.00446 & 0.2213 \end{pmatrix}$$

Les variances et les covariances sont généralement plus petites dans ce modèle que dans le précédent, ce qui indique que l'inclusion de covariables supplémentaires explique davantage de variation et de covariation du score de santé mentale perçue et du score du sentiment d'appartenance au quartier.

e. Test pour les variances-covariances multiples

Nous testons maintenant si la différence intra-quartier sur le score du sentiment d'appartenance au quartier entre les jeunes qui se sentent exclus de leur quartier et les jeunes qui ne se sentent pas exclus, varie entre les quartiers. Dans l'équation 3.12, $\beta_{2(11)k}$ représente la différence attendue entre ces deux groupes sur le score du sentiment d'appartenance au quartier au sein d'un quartier k en contrôlant les autres covariables.

Cette différence est supposée être constante d'un quartier à l'autre. De fait, nous ne disposons pas d'hypothèses à priori pour préciser si ces effets sont fixes ou varient d'un quartier à l'autre. Afin d'utiliser une base empirique pour modéliser la différence entre ces deux groupes comme fixe ou variable, nous réalisons un test pour déterminer si cette différence sur le score du sentiment d'appartenance au quartier varie entre les quartiers, après avoir contrôlé les autres variables. Pour mettre en œuvre ce test, un terme aléatoire ($u_{2(11)k}$) associé à cette différence entre ces deux groupes est ajouté à l'équation $\beta_{2(11)k} = \gamma_{2(11)0}$. Nous avons alors :

$$\beta_{2(11)k} = \gamma_{2(11)0} + u_{2(11)k}.$$

Le tableau D.8 en annexe D présente les matrices de variance-covariance et les statistiques d'ajustement estimées pour ce modèle. En ajoutant un résidu ($u_{2(11)k}$), la matrice $\boldsymbol{G}$ est maintenant

$$\boldsymbol{G} = \begin{pmatrix} \widehat{\sigma}^2_{u_1} & \widehat{\mathrm{Cov}}(u_1, u_2) & \widehat{\mathrm{Cov}}(u_1, u_3) \\ \widehat{\mathrm{Cov}}(u_2, u_1) & \widehat{\sigma}^2_{u_2} & \widehat{\mathrm{Cov}}(u_2, u_3) \\ \widehat{\mathrm{Cov}}(u_3, u_1) & \widehat{\mathrm{Cov}}(u_3, u_2) & \widehat{\sigma}^2_{u_3} \end{pmatrix} = \begin{pmatrix} 0.008801 & -0.00981 & 0.002384 \\ -0.00981 & 0.6628 & -0.1168 \\ 0.002384 & -0.1168 & 0.03072 \end{pmatrix}.$$

L'estimation de la variance résiduelle au niveau du quartier ($u_{2(11)k}$) pour des effets de la covariable *sentir_exclus_quartier* est $\widehat{\sigma}^2_{u_3} = 0.03072$.

Les deux nouvelles covariances pour des effets de la covariable *sentir_exclus_quartier* au niveau du quartier sont estimées par $\widehat{\mathrm{Cov}}(u_3, u_1) = 0.002384$, $\widehat{\mathrm{Cov}}(u_3, u_2) = -0.1168$.

L'hypothèse nulle ici suppose que les valeurs des paramètres de tous les nouveaux termes sont nulles. Celle-ci est testée en comparant la déviance du modèle actuel à la déviance du modèle précédent. La déviance du modèle précédent était $-2l = 11340.8$ et celle de ce modèle est $-2l = 11340.3$.

L'amélioration de l'ajustement obtenue par l'ajout de ces trois termes est négligeable, car la statistique du test du khi-deux est $\Delta(-2l) = 11340.8 - 11340.3 = 0.5$, ce qui ne dépasse pas la valeur critique de 7.815 ($\alpha = 5\%$ et dl= 3). Il n'y a alors pas de support empirique pour inclure ces effets aléatoires supplémentaires.

3.5.3 Épilogue de l'application du modèle

Les premiers obstacles à surmonter pour l'application d'un modèle multiniveaux multivarié en général sont de comprendre comment les données doivent être formatées, avec des *outcomes* multivariés apparaissant dans une seule colonne et séquencées au sein des individus, et comment le modèle de niveau 1 peut être spécifié pour inclure des *outcomes* multivariés dans un modèle multiniveaux standard. Ce modèle de niveau 1 diffère des modèles standards, car il n'inclut pas d'ordonnée à l'origine, n'a pas de terme résiduel et utilise des variables indicatrices codées de façon à ce que les paramètres du modèle deviennent des variables dépendantes spécifiques. Cependant, une fois ces obstacles surmontés, l'inclusion de covariables est similaire à la modélisation multiniveaux en général. L'application de ce modèle aux données *CentrÉS* est réaliste avec une certaine complexité croissante. En outre, nous avons souligné que le modèle multiniveaux multivarié est un outil d'analyse important pour les principales raisons suivantes. Premièrement, si des données manquantes sur certains *outcomes* sont présentes, il utilise plus d'observations que ne le ferait une application standard. Cela permet d'obtenir plus de puissance pour tester les différences entre les groupes. Nous pouvons également tester l'égalité des effets d'une covariable sur plusieurs *outcomes* comme montré au niveau des résultats.
Deuxièmement, si les individus sont imbriqués dans des contextes, tels que des quartiers de résidence, l'utilisation de ce modèle est généralement préférée au modèle linéaire simple ou à la modélisation multiniveaux standard, car celui-ci peut modéliser correctement la dépendance des observations que de tels contextes partagés produisent tout, en incorporant des *outcomes* multivariés.
Troisièmement, à l'instar du modèle multiniveaux standard, nous pouvons utiliser le modèle multiniveaux multivarié pour fournir des tests globaux de plusieurs paramètres afin d'aider à contrôler l'inflation du taux d'erreur de type I associé à des tests plus nombreux de paramètres individuels comme le cas dans cette présente étude.

CONCLUSION

L'étude présentée dans ce mémoire avait pour but de mieux comprendre les différences de santé mentale perçue et du sentiment d'appartenance au quartier chez les jeunes adultes imbriqués dans différents quartiers, après avoir contrôlé pour les facteurs individuels (âge, genre, niveau d'éducation, appartenance ethnique, revenu personnel, perceptions du quartier de résidence) et le niveau de défavorisation matérielle relative du quartier.

Avec les données de la première vague de l'enquête *CentrÉS* réalisée auprès des jeunes adultes de Sherbrooke âgés de 16 à 30 ans, deux grandes phases d'analyse avaient été menées, à savoir l'analyse descriptive et l'analyse explicative.

Dans l'analyse descriptive, nous avons d'une part présenté le profil sociodémographique et le profil des perceptions du quartier de résidence des participants. La population d'étude était constituée finalement de 1323 participants dont 27% étaient des hommes et 70% étaient des femmes. D'autre part, nous avons testé la liaison entre la santé mentale perçue et le sentiment d'appartenance au quartier. Le test de corrélation de Pearson, sous l'hypothèse nulle, nous a montré qu'il y a un lien positif statistiquement significatif entre le score de santé mentale perçue et le score du sentiment d'appartenance au quartier qui sont les deux *outcomes* de l'étude. Cela était un apport potentiel pour faire recours à une analyse par approche hiérarchique avec un *outcome* bivarié.

Dans l'analyse explicative, nous avons mené une analyse à deux niveaux avec un *outcome* bivarié pour tenir compte de la dépendance des observations résultant de l'imbrication des participants dans des quartiers en modélisant simultanément le score de santé mentale perçue et le score du sentiment d'appartenance au quartier selon les principes du modèle linéaire gaussien à effets mixtes et la stratégie d'analyse multiniveaux.

L'examen des résultats de cette analyse menée en séries de modélisation a permis de relever tout d'abord la présence d'effet significatif de l'appartenance ethnique seulement sur le score de santé mentale perçue. Il y a donc une différence moyenne statistiquement significative entre les jeunes adultes du groupe d'ethnie caucasienne et ceux du groupe d'ethnie non caucasienne pour le score de santé mentale perçue uniquement. Les résultats ont montré que pour un quartier donné, le groupe des Caucasiens avait tendance à avoir en moyenne une meilleure santé mentale que le groupe des non-Caucasiens.

Ensuite, nous avons testé l'hypothèse de l'effet d'interaction entre l'âge et le genre des participants. Il y a eu donc une différence significative de l'effet de l'âge uniquement sur le score de santé mentale perçue entre le groupe des hommes et le groupe des femmes. L'effet de l'âge sur le score de santé mentale perçue est plus faible parmi le groupe des femmes que parmi le groupe des hommes.

Plus particulièrement, les résultats ont rapporté que ce sont le revenu personnel du participant à l'étude et le fait de se sentir de plus en plus exclu de son quartier qui ont des effets significatifs sur le score de santé mentale perçue et sur le score du sentiment d'appartenance au quartier simultanément. Le score de santé mentale perçue diminuait en moyenne chez le groupe des jeunes ayant déclaré avoir un revenu élevé et pour laquelle le groupe des jeunes sans revenu personnel est la référence. Le groupe des jeunes à revenu élevé avait tendance à avoir en moyenne une meilleure santé mentale. Quant au score du sentiment d'appartenance au quartier, il augmentait significativement chez le groupe des jeunes à revenu personnel élevé. Le groupe des jeunes à revenu élevé avait donc tendance à avoir en moyenne un fort sentiment d'appartenance au quartier.

L'effet de l'indice de se sentir de plus en plus exclus de son quartier était en moyenne moins faible sur le score de santé mentale perçue, mais plus faible sur le score du sentiment d'appartenance au quartier. Les jeunes qui se sentent de plus en plus exclus de leur quartier avaient donc en moyenne un plus faible sentiment d'appartenance au quartier.

Le facteur contextuel tel que le niveau de défavorisation matérielle relative au quartier dans lequel résident les participants n'avait d'effet significatif que sur le score du sentiment d'appartenance au quartier. Cet effet est plus élevé parmi les jeunes qui résident dans des quartiers favorisés que parmi les jeunes qui résident dans des quartiers défavorisés.

Nous retenons finalement pour cette présente étude que la plus grande partie de la variabilité se situait au sein des quartiers et non entre les quartiers, car la part de variabilité imputable au niveau du quartier (corrélation résiduelle au niveau du participant) était significative. En somme, la corrélation résiduelle au niveau du quartier n'était pas statistiquement significative avec cette première vague de données *CentrÉS* (en transversale) mais a influencé positivement l'ajustement du modèle multiniveaux bivarié estimé après avoir effectué des tests statistiques successifs sur les déviances des modèles.

En perspective, nous suggérons une continuité de cette analyse multiniveaux lorsque les données *CentrÉS* seront longitudinales. Plusieurs observations seront mesurées sur le même individu au cours du temps et apporteront des changements. Les corrélations entre les observations et celles entre les quartiers dans lesquels résident les participants de l'étude devront être réévaluées par de nouveaux tests de significativité sur les nouvelles estimations des éléments de la matrice de variances-covariances.

Bibliographie

[1] Akaike, H. Information theory and an extension of the maximum likelihood principle,[w :] proceedings of the 2nd international symposium on information, bn petrow, f. *Czaki, Akademiai Kiado, Budapest* (1973).

[2] Akaike, H. Information measures and model selection. *Int Stat Inst 44* (1983), 277–291.

[3] Bibeau, G., Sabatier, C., Corin, E., Tousignant, M., and Saucier, J.-F. La recherche sociale anglo-saxonne en santé mentale : tendances, limites et impasses. *Santé mentale au Québec 14*, 1 (1989), 103–120.

[4] Brabant, L. Ginette paquet, partir du bas de l'échelle. des pistes pour atteindre l'égalité sociale en matière de santé, montréal, presses de l'université de montréal, 2005, 152 p. *Recherches sociographiques 48*, 1 (2007), 178–180.

[5] Bressoux, P., Coustère, P., and Leroy-Audouin, C. Les modèles multiniveau dans l'analyse écologique : le cas de la recherche en éducation. *Revue française de sociologie* (1997), 67–96.

[6] Bryk, A. S., and Raudenbush, S. W. *Hierarchical linear models : Applications and data analysis methods.* Sage Publications, Inc, 1992.

[7] Carroll, R., Wang, S., Simpson, D., Stromberg, A., and Ruppert, D. The sandwich (robust covariance matrix) estimator. *Unpublished manuscript* (1998).

[8] Daveluy, C., and de la statistique du Québec, I. *Enquête sociale et de santé 1998*. Institut de la statistique du Québec, 2001.

[9] Diez Roux, A. V. Multilevel analysis in public health research.

[10] Duncan, C., Jones, K., and Moon, G. Health-related behaviour in context : a multilevel modelling approach. *Social science & medicine 42*, 6 (1996), 817–830.

[11] Duncan, C., Jones, K., and Moon, G. Context, composition and heterogeneity : using multilevel models in health research. *Social science & medicine 46*, 1 (1998), 97–117.

[12] Evans, R., et al. Interpreting and addressing inequalities in health : from black to acheson to blair to... ? *Monographs* (2002).

[13] Fone, D. L., Farewell, D. M., and Dunstan, F. D. An ecometric analysis of neighbourhood cohesion. *Population health metrics 4*, 1 (2006), 1–17.

[14] Foulley, J.-L., Delmas, C., and Robert-Granié, C. Méthodes du maximum de vraisemblance en modèle linéaire mixte. *Journal de la société française de statistique 143*, 1-2 (2002), 5–52.

[15] Foulley, J.-L., Delmas, C., and Robert-Granié, C. Méthodes du maximum de vraisemblance en modèle linéaire mixte. *Journal de la société française de statistique 143*, 1-2 (2002), 20.

[16] Foulley, J.-L., Delmas, C., and Robert-Granié, C. Méthodes du maximum de vraisemblance en modèle linéaire mixte. *Journal de la société française de statistique 143*, 1-2 (2002), 39–40.

[17] Freedman, D. A. On the so-called "huber sandwich estimator" and "robust standard errors". *The American Statistician 60*, 4 (2006), 299–302.

[18] Gauvin, L., and Dassa, C. L'analyse multiniveaux : avancées récentes et retombées anticipées pour l'étude des inégalités sociales et de santé. *Santé, société et solidarité 3*, 2 (2004), 187–195.

[19] Greene, G., Fone, D., Farewell, D., Rodgers, S., Paranjothy, S., Carter, B., and White, J. Improving mental health through neighbourhood regeneration : the role of cohesion, belonging, quality and disorder. *European journal of public health 30*, 5 (2020), 964–966.

[20] Grilli, L., Pennoni, F., Rampichini, C., and Romeo, I. Exploiting timss and pirls combined data : multivariate multilevel modelling of student achievement. *The Annals of Applied Statistics 10*, 4 (2016), 2405–2426.

[21] Hedges, L. V., and Hedberg, E. C. Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis 29*, 1 (2007), 60–87.

[22] Hill, T. D., and Maimon, D. Neighborhood context and mental health. In *Handbook of the sociology of mental health*. Springer, 2013, pp. 479–501.

[23] INSPQ. "défavorisation". www.inspq.qc.ca/defavorisationref.

[24] jeunesse de Montréal, C. *MONTRÉAL, MA VILLE, MON CHOIX ? : Le sentiment d'appartenance des jeunes Montréalais*. Conseil jeunesse de Montréal, 2006.

[25] Kenny, D. A., and Judd, C. M. Consequences of violating the independence assumption in analysis of variance. *Psychological bulletin 99*, 3 (1986), 422.

[26] Kreft, I. G., and De Leeuw, J. *Introducing multilevel modeling*. Sage, 1998.

[27] Kullback, S., and Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics 22*, 1 (1951), 79–86.

[28] Levasseur, M. Perception de l'état de santé. *la santé et le bien-être* (2001), 259.

[29] Leyland, A. H., and Goldstein, H. *Multilevel modelling of health statistics*. Wiley, 2001.

[30] Liang, K.-Y., and Zeger, S. L. Longitudinal data analysis using generalized linear models. *Biometrika 73*, 1 (1986), 13–22.

[31] Macintyre, S., and Ellaway, A. Ecological approaches : rediscovering the role of the physical and social environment. *Social epidemiology 9*, 5 (2000), 332–348.

[32] Mourji, F., and Abbaia, A. Les déterminants du rendement scolaire en mathématiques chez les élèves de l'enseignement secondaire collégial au maroc : une analyse multiniveaux. *Revue d'économie du développement 21*, 1 (2013), 127–158.

[33] oedc. "portrait des communautés". www.santeestrie.qc.ca.

[34] Painter, C. V. Le sentiment d'appartenance : revue de la littérature.

[35] Pajot, H., and Russ, E. 1. éléments de théorie de la mesure. In *Analyse dans les espaces métriques*. EDP Sciences, 2021, pp. 7–106.

[36] Pampalon, R., Duncan, C., Subramanian, S., and Jones, K. Geographies of health perception in quebec : a multilevel perspective. *Social science & medicine 48*, 10 (1999), 1483–1490.

[37] Pampalon, R., Hamel, D., and Gamache, P. Les inégalités sociales de santé augmentent-elles au québec.

[38] Paquet, G. *Partir du bas de l'échelle : des pistes pour atteindre l'égalité sociale en matière de santé*. PUM, 2005.

[39] Pickett, K. E., and Pearl, M. Multilevel analyses of neighbourhood socioeconomic context and health outcomes : a critical review. *Journal of Epidemiology & Community Health 55*, 2 (2001), 111–122.

[40] Pituch, K. A., and Stevens, J. P. *Applied multivariate statistics for the social sciences : Analyses with SAS and IBM's SPSS*. Routledge, 2015.

[41] Ranathunga, K. N., and Sooriyarachchi, R. Multivariate multilevel modeling of age related diseases. *Journal of Modern Applied Statistical Methods 16*, 1 (2017), 28.

[42] Rao, C. R. Estimation of variance and covariance components—minque theory. *Journal of multivariate analysis 1*, 3 (1971), 257–275.

[43] Rao, C. R. Minimum variance quadratic unbiased estimation of variance components. *Journal of Multivariate Analysis 1*, 4 (1971), 445–456.

[44] Raudenbush, S. W. Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual review of psychology 52* (2001), 501.

[45] Raudenbush, S. W., and Bryk, A. S. *Hierarchical linear models : Applications and data analysis methods*, vol. 1. sage, 2002.

[46] Rice, N., and Leyland, A. Multilevel models : applications to health data. *Journal of health services research & policy 1*, 3 (1996), 154–164.

[47] Robert, S. A. Socioeconomic position and health : the independent contribution of community socioeconomic context. *Annual review of sociology* (1999), 489–516.

[48] Robert Granié, C., and Servin, B. Modèle linéaire mixte gaussien.

[49] Ross, N. Appartenance à la collectivité et santé. *Rapports sur la santé 13*, 3 (2002), 35.

[50] Satterthwaite, F. E. Synthesis of variance. *Psychometrika 6*, 5 (1941), 309–316.

[51] Scariano, S. M., and Davenport, J. M. The effects of violations of independence assumptions in the one-way anova. *The American Statistician 41*, 2 (1987), 123–129.

[52] Schwarz, G. Estimating the dimension of a model. *The annals of statistics* (1978), 461–464.

[53] Searle, S. R., McCulloch, C. E., and Neuhaus, J. M. *Generalized, linear, and mixed models*. John Wiley & Sons, 2011.

[54] Shareck, M. "etude centrÉs". www.shareck-lab.ca/etude-centres.

[55] Shields, M. Appartenance à la communauté et autoévaluation de l'état de santé. *Rapports sur la santé 19*, 2 (2008), 1–10.

[56] Smedley, B. D., and Syme, S. Committee on capitalizing on social science and behavioral research to improve the public's health. *Promoting health : intervention strategies from social and behavioral research. Am J Health Promot 15*, 3 (2001), 149–166.

[57] Snijders, T. A., and Bosker, R. J. *Multilevel analysis : An introduction to basic and advanced multilevel modeling.* sage, 2011.

[58] Spielberger, C. D. *Inventaire d'anxiété État-Trait : forme Y.* ECPA, les Éditions du centre de Psychologie Appliquée, 1993.

[59] statcan. "appartenance à la communauté". www150.statcan.gc.ca/n1/pub/82-229-x/2009001/envir/cob-fra.htm.

[60] Stewart, M. J., Makwarimba, E., Reutter, L. I., Veenstra, G., Raphael, D., and Love, R. Poverty, sense of belonging and experiences of social isolation. *Journal of Poverty 13*, 2 (2009), 173–195.

[61] Subramanian, S., Jones, K., Duncan, C., et al. *Multilevel methods for public health research.* Neighborhoods and health. New York : Oxford University Press, 2003.

[62] Swallow, W. H., and Searle, S. Minimum variance quadratic unbiased estimation (mivque) of variance components. *Technometrics 20*, 3 (1978), 265–272.

[63] Tom, A., Bosker, T. A. S. R. J., and Bosker, R. J. *Multilevel analysis : an introduction to basic and advanced multilevel modeling.* sage, 1999.

Annexe A

Score du sentiment d'appartenance au quartier : les items

1. De manière générale, je suis attiré par le fait de vivre dans ce quartier ($aq8\_a$)
2. J'ai le sentiment d'appartenir à ce quartier ($aq8\_b$)
3. Si l'occasion se présente, j'aimerais déménager ailleurs ($aq8\_c$)
4. Je prévois rester résident de ce quartier pendant plusieurs années ($aq8\_d$)
5. Je me considère semblable aux personnes qui vivent dans ce quartier ($aq8\_e$)
6. Vivre dans ce quartier me donne un sentiment d'appartenance au quartier ($aq8\_f$)

Pour chaque item, on a les réponses ordinales suivantes : 1 = tout à fait d'accord ; 2 = d'accord ; 3 = ni d'accord ni en désaccord ; 4 = pas d'accord ; 5 = pas du tout d'accord.

Annexe B

Présentation des covariables de l'étude

Variables
Age (*age*)
Genre (*genre*) 1 = Homme 2 = Femme 3 = Homme trans 4 = Femme trans 5 = De genre queer ou non conforme au genre 6 = Identité différente
Niveau d'éducation (*education*) 1 = Secondaire 4 ou moins 2 = Diplôme d'études secondaires ou équivalent 3 = Diplôme ou certificat d'études d'un programme technique au CÉGEP, d'une école de métiers, d'un collège commercial ou privé ou d'un institut technique 4 = Diplôme d'études d'un programme général au CÉGEP 5 = Certificat ou diplôme universitaire de premier cycle 6 = Certificat ou diplôme universitaire de deuxième cycle 7 = Doctorat
Revenu personnel (*revenu*) 1 = Aucun revenu personnel 2 = 1 $ à 4 999 $ 3 = 5 000 $ à 9 999 $ 4 = 10 000 $ à 14 999 $ 5 = 15 000 $ à 19 999 $ 6 = 20 000 $ à 29 999 $ 7 = 30 000 $ à 39 999 $ 8 = 40 000 $ à 49 999 $ 9 = 50 000 $ à 99 999 $ 10 = 100 000 $ et plus 11 = Je préfère ne pas répondre
Appartenance ethnique (*ethnie*) 1 = Caucasiens (blancs) 2 = Non caucasiens 3 = Je ne sais pas 4 = Je préfère ne pas répondre

Tableau B.1 – Variables socio-démographiques et économique

Variables
Avoir accès à tout ce dont on a besoin dans mon quartier (*acces_tout_besoin_quartier*) 1=Tout à fait d'accord 2=D'accord 3=Ni d'accord ni en désaccord 4=Pas d'accord 5=Pas du tout d'accord
Il y a des choses intéressantes à faire dans mon quartier (*choses_interessant_quartier*) 1=Tout à fait d'accord 2=D'accord 3=Ni d'accord ni en désaccord 4=Pas d'accord 5=Pas du tout d'accord
La ville investit dans mon quartier (*ville_investit_quartier*) 1=Tout à fait d'accord 2=D'accord 3=Ni d'accord ni en désaccord 4=Pas d'accord 5=Pas du tout d'accord
Les changements dans mon quartier améliorent ma qualité de vie (*changement_qualite_vie_quartier*) 1=Tout à fait d'accord 2=D'accord 3=Ni d'accord ni en désaccord 4=Pas d'accord 5=Pas du tout d'accord
Mon quartier est de plus en plus dynamique (*plus_dynamique_quartier*) 1=Tout à fait d'accord 2=D'accord 3=Ni d'accord ni en désaccord 4=Pas d'accord 5=Pas du tout d'accord
Les personnes à faible revenu ont de la difficulté à rester dans le quartier (*faible_revenu_difficulte_quartier*) 1=Tout à fait d'accord 2=D'accord 3=Ni d'accord ni en désaccord 4=Pas d'accord 5=Pas du tout d'accord
Je me sens de plus en plus exclus de mon quartier (*sentir_exclus_quartier*) 1=Tout à fait d'accord 2=D'accord 3=Ni d'accord ni en désaccord 4=Pas d'accord 5=Pas du tout d'accord

Tableau B.2 – Variables liées à la perception du quartier

Variable
Sentiment de sécurité (*securite_percue*)
1=Tout à fait en sécurité
2=Plutôt en sécurité
3=Pas très en sécurité
4=Pas du tout en sécurité
5=Je ne marche jamais seul(e)

TABLEAU B.3 – Variable liée au sentiment de sécurité dans le quartier

Variable
Niveau de défavorisation matérielle (*defavorisation*)
1=Très favorisé ($Q1$)
2=Favorisé ($Q2$)
3=Défavorisation moyenne ($Q3$)
4=Défavorisation forte ($Q4$)
5=Défavorisation très forte ($Q5$)

TABLEAU B.4 – Variable relative au contexte du quartier

Annexe C

Résultats de l'analyse descriptive

Variables	$n_0(\%)$
Age (*age*)	
Moyenne (±Ecart-type)	23.3(4.1)
Médiane [IIQ]	23[20 – 27]
Genre (*genre*)	
Homme	387(26.3)
Femme	1032(70.2)
Homme trans	5(0.3)
Femme trans	1(0.1)
De genre queer ou non conforme au genre	34(2.3)
Identité différente	11(0.8)
Niveau d'éducation (*education*)	
Secondaire 4 ou moins	136(9.2)
Diplôme d'études secondaires ou équivalent	232(15.8)
Diplôme ou certificat d'études d'un programme technique au CÉGEP, d'une école de métiers, d'un collège commercial ou privé ou d'un institut technique	199(13.5)
Diplôme d'études d'un programme général au CÉGEP	298(20.3)
Certificat ou diplôme universitaire de premier cycle	401(27.3)
Certificat ou diplôme universitaire de deuxième cycle	191(13.0)
Doctorat	13(0.9)
Revenu personnel (*revenu*)	
Aucun revenu personnel	58(3.9)
1 $ à 4 999 $	122(8.3)
5 000 $ à 9 999 $	198(13.5)
10 000 $ à 14 999 $	263(17.9)
15 000 $ à 19 999 $	191(13.0)
20 000 $ à 29 999 $	185(12.6)
30 000 $ à 39 999 $	133(9.0)
40 000 $ à 49 999 $	108(7.3)
50 000 $ à 99 999 $	122(8.3)
100 000 $ et plus	14(1.0)
Je préfère ne pas répondre	76(5.2)
Appartenance ethnique (*ethnie*)	
Caucasiens (blancs)	1200(81.6)
Non caucasiens	232(15.8)
Je ne sais pas	26(1.8)
Je préfère ne pas répondre	12(0.8)

TABLEAU C.1 – Profil de la population d'étude

Avoir accès à tout ce dont on a besoin dans mon quartier (*acces_tout_besoin_quartier*)	
Tout à fait d'accord	270(18.4)
D'accord	549(37.3)
Ni d'accord ni en désaccord	226(15.4)
Pas d'accord	312(21.2)
Pas du tout d'accord	113(7.7)
Il y a des choses intéressantes à faire dans mon quartier (*choses_interessant_quartier*)	
Tout à fait d'accord	189(12.9)
D'accord	454(30.9)
Ni d'accord ni en désaccord	405(27.6)
Pas d'accord	317(21.6)
Pas du tout d'accord	105(7.0)
La ville investit dans mon quartier (*ville_investit_quartier*)	
Tout à fait d'accord	76(5.2)
D'accord	400(27.2)
Ni d'accord ni en désaccord	557(37.9)
Pas d'accord	352(23.9)
Pas du tout d'accord	85(5.8)
Les changements dans mon quartier améliorent ma qualité de vie (*changement_qualite_vie_quartier*)	
Tout à fait d'accord	67(4.6)
D'accord	324(22.0)
Ni d'accord ni en désaccord	751(51.1)
Pas d'accord	259(17.6)
Pas du tout d'accord	69(4.7)
Mon quartier est de plus en plus dynamique (*plus_dynamique_quartier*)	
Tout à fait d'accord	68(4.6)
D'accord	352(24.0)
Ni d'accord ni en désaccord	551(37.5)
Pas d'accord	412(28.0)
Pas du tout d'accord	87(5.9)
Les personnes à faible revenu ont de la difficulté à rester dans le quartier (*revenu_difficulte_quartier*)	
Tout à fait d'accord	163(11.1)
D'accord	351(23.9)
Ni d'accord ni en désaccord	463(31.5)
Pas d'accord	345(23.5)
Pas du tout d'accord	148(10.0)
Je me sens de plus en plus exclus de mon quartier (*sentir_exclus_quartier*)	
Tout à fait d'accord	104(7.1)
D'accord	234(15.9)
Ni d'accord ni en désaccord	449(30.5)
Pas d'accord	464(31.6)
Pas du tout d'accord	219(14.9)

TABLEAU C.2 – Perceptions du quartier de résidence

Sentiment de sécurité de marcher dans le quartier lorsqu'il fait noir (*securite_percue*)	
Tout à fait en sécurité	424(28.8)
Plutôt en sécurité	607(41.3)
Pas très en sécurité	253(17.2)
Pas du tout en sécurité	101(6.9)
Je ne marche jamais seul(e)	85(5.8)

TABLEAU C.3 – Perception de sécurité dans le quartier

Niveau de défavorisation matérielle (*defavorisation*)	
Très favorisé ($Q1$)	363(24.7)
Favorisé ($Q2$)	319(21.7)
Défavorisation moyenne ($Q3$)	249(16.9)
Défavorisation forte ($Q4$)	179(12.2)
Défavorisation très forte ($Q5$)	360(24.5)

TABLEAU C.4 – Distribution de la population d'étude par niveau de défavorisation matérielle des communautés de résidence à Sherbrooke

Communauté de sherbrooke	**n**	**(%)**
André-Viger	48	(3.27)
Aéroport	20	(1.36)
Beaulieu	19	(1.29)
Beckett	17	(1.16)
Boisé-Fabi	21	(1.43)
Brompton	71	(4.83)
Centre-ville de Sherbrooke	89	(6.05)
Chauveau	26	(1.77)
Châteaux d'eau	12	(0.82)
Deauville	14	(0.95)
Desranleau	28	(1.90)
Grands-Monts	116	(7.89)

Hélène-Boullé	9	(0.61)
Immaculée-Conception	73	(4.97)
Jardins-Fleuris	29	(1.97)
Julien-Ducharme	57	(3.88)
Laurentie	45	(3.06)
Lennoxville	30	(2.04)
Marie-Reine	27	(1.84)
Mi-Vallon	29	(1.97)
Nouveau village de Saint-Élie	19	(1.29)
Parc central	17	(1.16)
Petit-Lac-Magog	8	(0.54)
Pin-Solitaire	44	(2.99)
Saint-Alphonse	43	(2.93)
Saint-Boniface	20	(1.36)
Saint-Jean-Baptiste	58	(3.95)
Saint-Jean-de-Brébeuf	27	(1.84)
Saint-Michel	53	(3.61)
Saint-Élie	10	(0.68)
Sainte-Catherine	33	(2.24)
Sainte-Jeanne-d'Arc	121	(8.23)
Université	149	(10.14)
Vieux-Nord	47	(3.20)
des Châteaux	19	(1.29)
du Phare	22	(1.50)

TABLEAU C.5 – Répartition de la population d'étude par communauté à Sherbrooke

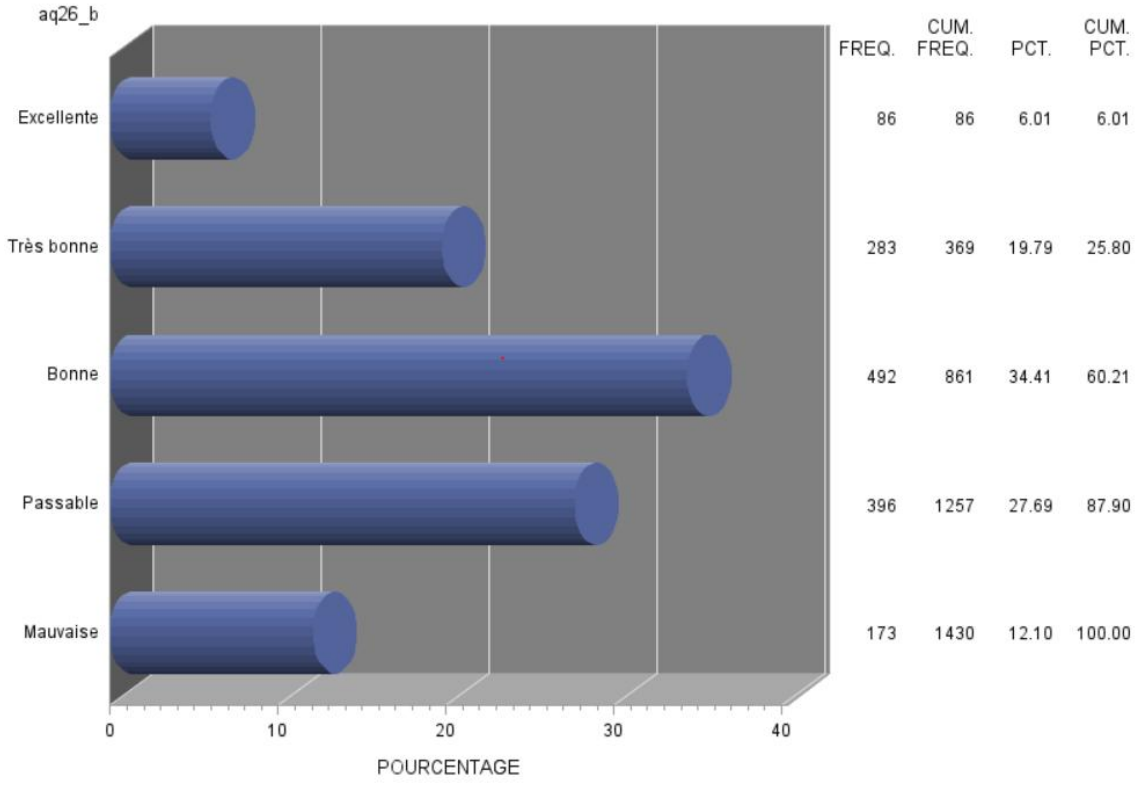

FIGURE C.1 – Pourcentage des participants selon la santé mentale perçue

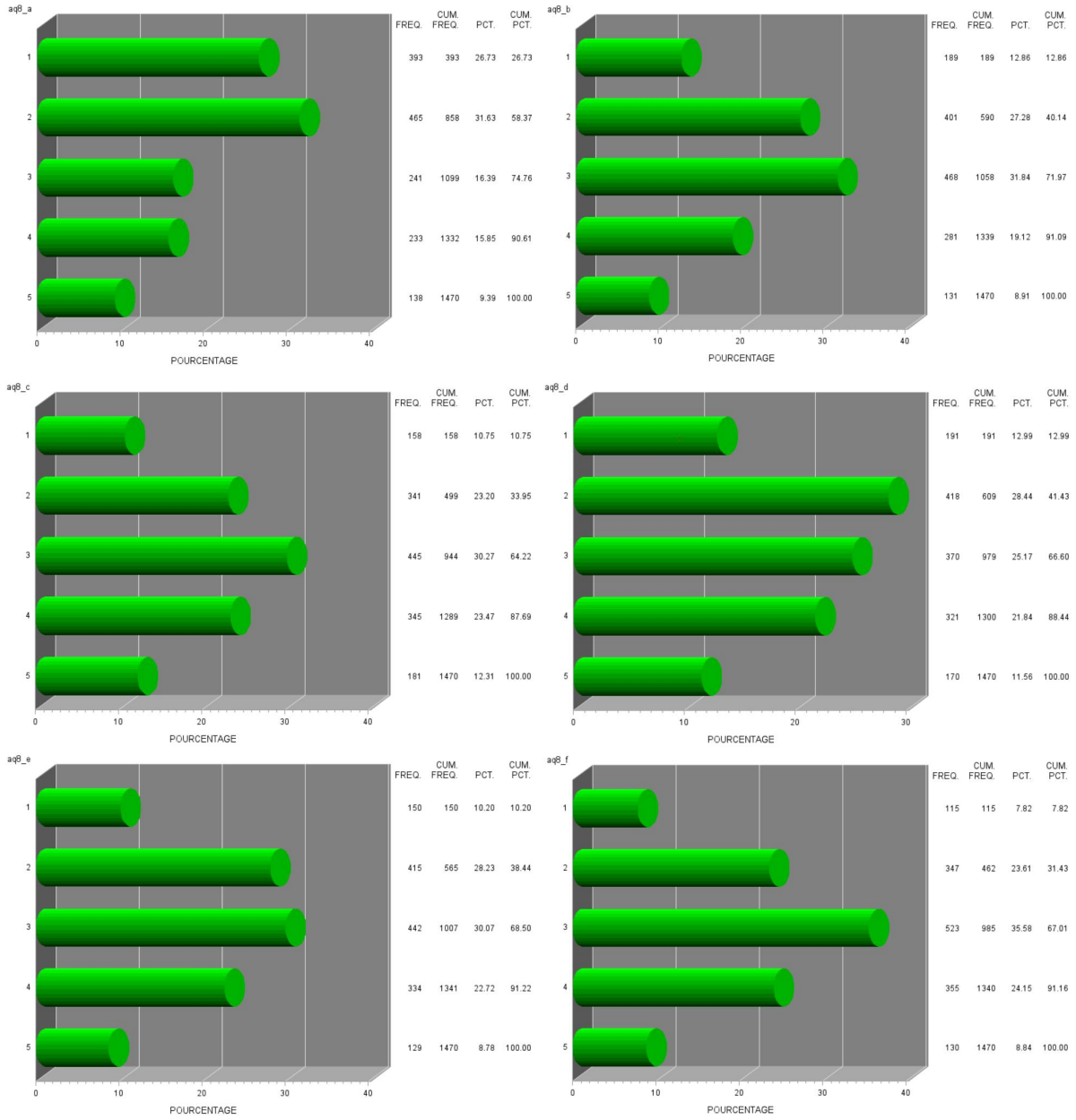

FIGURE C.2 – Pourcentage des participants selon chacun des items du sentiment d'appartenance au quartier

Annexe D

Résultats de l'analyse multiniveaux bivariée

	Identifiant	Age revolu: date de début d'éligibilité - date de naissance	Niveau d'éducation terminé	Genre	score de santé mentale perçue	Score du sentiment d'appartenance au quartier	...
1	72	28	Universitaire	Femme	3	27	
2	74	29	Universitaire	Homme	3	24	
3	76	16	Secondaire	Femme	4	28	
4	77	30	Universitaire	Femme	4	21	
5	81	22	Collegial ou technique	Femme	2	22	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
1320	5058	20	Collegial ou technique	Femme	3	21	...
1321	5093	23	Secondaire	Homme	2	9	...
1322	5311	30	Collegial ou technique	Femme	3	16	
1323	5315	25	Collegial ou technique	Homme	2	24	

TABLEAU D.1 – Ensemble de données *CentrÉS* en format large

	Identifiant	Age revolu: date de début d'éligibilité - date de naissance	Niveau d'éducation terminé	Genre	index1	response	a1	a2
1	72	28	Universitaire	Femme	1	3	1	0
2	72	28	Universitaire	Femme	2	27	0	1
3	74	29	Universitaire	Homme	1	3	1	0
4	74	29	Universitaire	Homme	2	24	0	1
5	76	16	Secondaire	Femme	1	4	1	0
6	76	16	Secondaire	Femme	2	28	0	1
7	77	30	Universitaire	Femme	1	4	1	0
8	77	30	Universitaire	Femme	2	21	0	1
9	81	22	Collegial ou technique	Femme	1	2	1	0
10	81	22	Collegial ou technique	Femme	2	22	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2639	5058	20	Collegial ou technique	Femme	1	3	1	0
2640	5058	20	Collegial ou technique	Femme	2	21	0	1
2641	5093	23	Secondaire	Homme	1	2	1	0
2642	5093	23	Secondaire	Homme	2	9	0	1
2643	5311	30	Collegial ou technique	Femme	1	3	1	0
2644	5311	30	Collegial ou technique	Femme	2	16	0	1
2645	5315	25	Collegial ou technique	Homme	1	2	1	0
2646	5315	25	Collegial ou technique	Homme	2	24	0	1

TABLEAU D.2 – Ensemble de données *CentrÉS* en format long

Matrice R estimée pour user_id 100		
Ligne	Col1	Col2
1	1.1516	0.9401
2	0.9401	27.6339

Tests d'ajustement	
-2 log-vraisemblance	12049.6
AIC (préférer les petites valeurs)	12059.6
AICC (préférer les petites valeurs)	12059.6
BIC (préférer les petites valeurs)	12085.6

Solution pour effets fixes						
Effet	index1	Estimation	Erreur type	DDL	Valeur du test t	Pr > |t|
index1	1	3.2101	0.02950	1323	108.81	<.0001
index1	2	17.2525	0.1445	1323	119.37	<.0001

Valeur estimée du paramètre de covariance					
Param. de cov.	Sujet	Estimation	Erreur type	Valeur Z	Pr Z
UN(1,1)	user_id	1.1516	0.04478	25.72	<.0001
UN(2,1)	user_id	0.9401	0.1572	5.98	<.0001
UN(2,2)	user_id	27.6339	1.0744	25.72	<.0001

Tableau D.3 – Estimation du modèle bivarié vide au niveau du participant

Matrice R estimée pour user_id 100		
Ligne	Col1	Col2
1	1.1460	0.9440
2	0.9440	27.6312

Valeur estimée du paramètre de covariance					
Param. de cov.	Sujet	Estimation	Erreur type	Valeur Z	Pr Z
UN(1,1)	user_id	1.1460	0.04456	25.72	<.0001
UN(2,1)	user_id	0.9440	0.1569	6.02	<.0001
UN(2,2)	user_id	27.6312	1.0743	25.72	<.0001

Tests d'ajustement	
-2 log-vraisemblance	12042.5
AIC (préférer les petites valeurs)	12056.5
AICC (préférer les petites valeurs)	12056.5
BIC (préférer les petites valeurs)	12092.8

Solution pour effets fixes							
Effet	index1	Appartenance ethnique	Estimation	Erreur type	DDL	Valeur du test t	Pr > |t|
index1	1		3.3854	0.07477	1323	45.28	<.0001
index1	2		17.1317	0.3671	1323	46.66	<.0001
index1*ethnie	1	Caucasiens	-0.2074	0.08133	1323	-2.55	0.0109
index1*ethnie	1	Non caucasiens	0	.	.	.	.
index1*ethnie	2	Caucasiens	0.1429	0.3994	1323	0.36	0.7206
index1*ethnie	2	Non caucasiens	0	.	.	.	.

Tests des effets fixes de type 3				
Effet	DDL num.	DDL den.	Valeur F	Pr > F
index1	2	1323	5968.58	<.0001
index1*ethnie	2	1323	3.57	0.0285

TABLEAU D.4 – Estimation du modèle bivarié au niveau du participant avec inclusion de la variable *ethnie*

Matrice R estimée pour user_id 100		
Ligne	Col1	Col2
1	1.1460	0.9439
2	0.9439	27.6470

Valeur estimée du paramètre de covariance					
Param. de cov.	Sujet	Estimation	Erreur type	Valeur Z	Pr Z
UN(1,1)	user_id	1.1460	0.04455	25.72	<.0001
UN(2,1)	user_id	0.9439	0.1570	6.01	<.0001
UN(2,2)	user_id	27.6470	1.0750	25.72	<.0001

Tests d'ajustement	
-2 log-vraisemblance	12043.3
AIC (préférer les petites valeurs)	12053.3
AICC (préférer les petites valeurs)	12053.3
BIC (préférer les petites valeurs)	12079.2

Solution pour effets fixes							
Effet	index1	Appartenance ethnique	Estimation	Erreur type	DDL	Valeur du test t	Pr > |t|
index1	1		3.3831	0.07472	1322	45.27	<.0001
index1	2		17.4255	0.1600	1322	108.88	<.0001
ethnie		Caucasiens	-0.2047	0.08128	1322	-2.52	0.0119
ethnie		Non caucasiens	0	.	.	.	.

Tests des effets fixes de type 3				
Effet	DDL num.	DDL den.	Valeur F	Pr > F
index1	1	1322	9696.23	<.0001
ethnie	1	1322	6.35	0.0119

TABLEAU D.5 – Estimation du modèle bivarié au niveau du participant avec effets de la variable *ethnie* contraints d'être égaux

Solution pour effets fixes							
Effet	index1	Appartenance ethnique	Estimation	Erreur type	DDL	Valeur du test t	Pr > |t|
index1	1		3.3842	0.07460	70	45.37	<.0001
index1	2		17.0705	0.4126	70	41.37	<.0001
index1*ethnie	1	Caucasiens	-0.2145	0.08116	2572	-2.64	0.0083
index1*ethnie	1	Non caucasiens	0	.	.	.	.
index1*ethnie	2	Caucasiens	0.07190	0.3943	2572	0.18	0.8553
index1*ethnie	2	Non caucasiens	0	.	.	.	.

Matrice R estimée pour user_id(NOM_COMM) 105 André-Viger		
Ligne	Col1	Col2
1	1.1431	0.8579
2	0.8579	26.5572

Matrice G estimée					
Ligne	Effet	Communauté d'appartenance a sherbrooke	index1	Col1	Col2
1	index1	André-Viger	1	0.002401	0.08512
2	index1	André-Viger	2	0.08512	1.1270

Valeur estimée du paramètre de covariance					
Param. de cov.	Sujet	Estimation	Erreur type	Valeur Z	Pr Z
UN(1,1)	NOM_COMM	0.002401	0.007456	0.32	0.3737
UN(2,1)	NOM_COMM	0.08512	0.04677	1.82	0.0688
UN(2,2)	NOM_COMM	1.1270	0.4630	2.43	0.0075
UN(1,1)	user_id(NOM_COMM)	1.1431	0.04483	25.50	<.0001
UN(2,1)	user_id(NOM_COMM)	0.8579	0.1550	5.54	<.0001
UN(2,2)	user_id(NOM_COMM)	26.5572	1.0446	25.42	<.0001

Tests d'ajustement	
-2 log-vraisemblance	12018.0
AIC (préférer les petites valeurs)	12038.0
AICC (préférer les petites valeurs)	12038.1
BIC (préférer les petites valeurs)	12053.8

TABLEAU D.6 – Estimation du modèle bivarié à deux niveaux avec la variable *ethnie*

Solution pour effets fixes										
Effet	Niveau d'éducation terminé	Genre	index1	Appartenance ethnique	Approximativement, quel était votre revenu personnel total l'année dernière, avant déductions d'impôts ?	Estimation	Erreur type	DDL	Valeur du test t	Pr > |t|
index1			1			2.3295	0.4617	68	5.05	<.0001
index1			2			11.8442	1.7678	68	6.70	<.0001
age*index1			1			0.04815	0.01557	2524	3.09	0.0020
age*index1			2			-0.1102	0.05958	2524	-1.85	0.0645
index1*education_rec	Collegial ou technique		1			-0.02286	0.07254	2524	-0.32	0.7527
index1*education_rec	Secondaire		1			0.09813	0.09883	2524	0.99	0.3209
index1*education_rec	Universitaire		1			0	.	.	.	.
index1*education_rec	Collegial ou technique		2			-0.2296	0.2768	2524	-0.83	0.4070
index1*education_rec	Secondaire		2			-0.7789	0.3788	2524	-2.06	0.0398
index1*education_rec	Universitaire		2			0	.	.	.	.
index1*genre_rec		Autres genres	1			-0.7931	1.0635	2524	-0.75	0.4559
index1*genre_rec		Femme	1			1.5087	0.3851	2524	3.92	<.0001
index1*genre_rec		Homme	1			0	.	.	.	.
index1*genre_rec		Autres genres	2			1.0558	4.0547	2524	0.26	0.7946
index1*genre_rec		Femme	2			-1.1868	1.4705	2524	-0.81	0.4197
index1*genre_rec		Homme	2			0	.	.	.	.
age*index1*genre_rec		Autres genres	1			0.05731	0.04610	2524	1.24	0.2139
age*index1*genre_rec		Femme	1			-0.05251	0.01612	2524	-3.26	0.0011
age*index1*genre_rec		Homme	1			0	.	.	.	.
age*index1*genre_rec		Autres genres	2			-0.06291	0.1758	2524	-0.36	0.7204
age*index1*genre_rec		Femme	2			0.05842	0.06155	2524	0.95	0.3427
age*index1*genre_rec		Homme	2			0	.	.	.	.
index1*ethnie			1	Caucasiens		-0.1769	0.08057	2524	-2.20	0.0282
index1*ethnie			1	Non caucasiens		0	.	.	.	.
index1*ethnie			2	Caucasiens		0.1307	0.3072	2524	0.43	0.6707
index1*ethnie			2	Non caucasiens		0	.	.	.	.
index1*revenu			1		1 $ à 4 999 $	-0.2215	0.1773	2524	-1.25	0.2116
index1*revenu			1		10 000 $ à 14 999 $	-0.1877	0.1674	2524	-1.12	0.2622
index1*revenu			1		100 000 $ et plus	-0.6770	0.3326	2524	-2.04	0.0419
index1*revenu			1		15 000 $ à 19 999 $	-0.1452	0.1747	2524	-0.83	0.4058
index1*revenu			1		20 000 $ à 29 999 $	-0.1677	0.1767	2524	-0.95	0.3427
index1*revenu			1		30 000 $ à 39 999 $	-0.1500	0.1877	2524	-0.80	0.4244
index1*revenu			1		40 000 $ à 49 999 $	-0.2640	0.1973	2524	-1.34	0.1810
index1*revenu			1		5 000 $ à 9 999 $	-0.2855	0.1679	2524	-1.70	0.0892
index1*revenu			1		50 000 $ à 99 999 $	-0.3165	0.1977	2524	-1.60	0.1095
index1*revenu			1		Aucun revenu personnel	0	.	.	.	.
index1*revenu			2		1 $ à 4 999 $	1.0376	0.6759	2524	1.54	0.1249
index1*revenu			2		10 000 $ à 14 999 $	0.7595	0.6380	2524	1.19	0.2340
index1*revenu			2		100 000 $ et plus	1.5327	1.2679	2524	1.21	0.2268
index1*revenu			2		15 000 $ à 19 999 $	-0.05404	0.6661	2524	-0.08	0.9353
index1*revenu			2		20 000 $ à 29 999 $	0.8461	0.6737	2524	1.26	0.2093
index1*revenu			2		30 000 $ à 39 999 $	0.9149	0.7161	2524	1.28	0.2015
index1*revenu			2		40 000 $ à 49 999 $	1.1211	0.7530	2524	1.49	0.1366
index1*revenu			2		5 000 $ à 9 999 $	0.7536	0.6402	2524	1.18	0.2392
index1*revenu			2		50 000 $ à 99 999 $	1.5553	0.7547	2524	2.06	0.0394
index1*revenu			2		Aucun revenu personnel	0	.	.	.	.

acces_tout_be*index1			1			-0.02085	0.02959	2524	-0.70	0.4812
acces_tout_be*index1			2			0.2268	0.1129	2524	2.01	0.0446
choses_intere*index1			1			-0.00668	0.03095	2524	-0.22	0.8292
choses_intere*index1			2			0.7739	0.1181	2524	6.55	<.0001
ville_investi*index1			1			0.1103	0.03638	2524	3.03	0.0024
ville_investi*index1			2			0.1446	0.1388	2524	1.04	0.2976
changement_qu*index1			1			0.003313	0.03833	2524	0.09	0.9311
changement_qu*index1			2			0.6557	0.1461	2524	4.49	<.0001
plus_dynamiqu*index1			1			0.02866	0.03510	2524	0.82	0.4143
plus_dynamiqu*index1			2			1.0404	0.1339	2524	7.77	<.0001
revenu_diffic*index1			1			-0.01709	0.02597	2524	-0.66	0.5107
revenu_diffic*index1			2			0.7972	0.09904	2524	8.05	<.0001
sentir_exclus*index1			1			-0.1566	0.03023	2524	-5.18	<.0001
sentir_exclus*index1			2			-1.5494	0.1153	2524	-13.43	<.0001
securite_perc*index1			1			0.01592	0.02933	2524	0.54	0.5874
securite_perc*index1			2			0.3600	0.1120	2524	3.21	0.0013
defavorisatio*index1			1			0.02161	0.02408	2524	0.90	0.3695
defavorisatio*index1			2			0.3067	0.09973	2524	3.08	0.0021

Matrice R estimée pour user_id(NOM_COMM) 105 André-Viger

Ligne	Col1	Col2
1	1.0788	0.2899
2	0.2899	15.6460

Matrice G estimée

Ligne	Effet	Communauté d'appartenance a sherbrooke	index1	Col1	Col2
1	index1	André-Viger	1	0.008801	-0.00446
2	index1	André-Viger	2	-0.00446	0.2213

Valeur estimée du paramètre de covariance

Param. de cov.	Sujet	Estimation	Erreur type	Valeur Z	Pr Z
UN(1,1)	NOM_COMM	0.008801	0	.	.
UN(2,1)	NOM_COMM	-0.00446	0.04195	-0.11	0.9154
UN(2,2)	NOM_COMM	0.2213	0.1688	1.31	0.0950
UN(1,1)	user_id(NOM_COMM)	1.0788	0.04278	25.22	<.0001
UN(2,1)	user_id(NOM_COMM)	0.2899	0.1158	2.50	0.0123
UN(2,2)	user_id(NOM_COMM)	15.6460	0.6230	25.11	<.0001

Tests d'ajustement

-2 log-vraisemblance restreinte	11340.8
AIC (préférer les petites valeurs)	11352.8
AICC (préférer les petites valeurs)	11352.8
BIC (préférer les petites valeurs)	11362.3

TABLEAU D.7 – Estimation du modèle bivarié à deux niveaux avec prédicteurs multiples et avec interaction

Nombre d'observations	
Nb d'observations lues	2646
Nb d'obs. utilisées	2646
Nb d'obs. non utilisées	0

Matrice R estimée pour user_id(NOM_COMM) 105 André-Viger		
Ligne	Col1	Col2
1	1.0788	0.2899
2	0.2899	15.4896

Matrice G estimée						
Ligne	Effet	Communauté d'appartenance a sherbrooke	index1	Col1	Col2	Col3
1	index1	André-Viger	1	0.008801	0.01533	-0.00602
2	index1	André-Viger	2	0.01533	1.0525	-0.3390
3	sentir_exclus_qua*a2	André-Viger		-0.00602	-0.3390	0.1438

Valeur estimée du paramètre de covariance					
Param. de cov.	Sujet	Estimation	Erreur type	Valeur Z	Pr Z
UN(1,1)	NOM_COMM	0.008801	0	.	.
UN(2,1)	NOM_COMM	0.01533	0.2813	0.05	0.9565
UN(2,2)	NOM_COMM	1.0525	1.4293	0.74	0.2308
UN(3,1)	NOM_COMM	-0.00602	0.09124	-0.07	0.9474
UN(3,2)	NOM_COMM	-0.3390	0.4093	-0.83	0.4076
UN(3,3)	NOM_COMM	0.1438	0.1367	1.05	0.1464
UN(1,1)	user_id(NOM_COMM)	1.0788	0.04284	25.18	<.0001
UN(2,1)	user_id(NOM_COMM)	0.2899	0.1167	2.48	0.0130
UN(2,2)	user_id(NOM_COMM)	15.4896	0.6228	24.87	<.0001

Tests d'ajustement	
-2 log-vraisemblance restreinte	11339.6
AIC (préférer les petites valeurs)	11357.6
AICC (préférer les petites valeurs)	11357.6
BIC (préférer les petites valeurs)	11371.8

TABLEAU D.8 – Test pour les variances-covariances multiples : indice de se sentir exclure de son quartier en aléatoire pour le score du sentiment d'appartenance au quartier