Travail pratique 2

Travail présenté à

M. Thierry Duchesne

Dans le cadre du cours

Théorie et applications des méthodes de régression
STT-7125

Réalisé par l'équipe 21 :

Alexandre Lepage
& Amedeo Zito

le 17 décembre 2020

Faculté des sciences et de génie
École d'actuariat
Université Laval

Table des matières

Liste des illustrations

Introduction

Les modèles linéaires (LM) et modèles linéaires généralisés (GLM) sont des outils fort utiles pour modéliser toute sorte de phénomènes et sont largement utilisés dans le milieu statistique. Cependant, ces modèles s'appuient sur l'hypothèse que les observations $Y_1, \ldots, Y_n,\ n > 0$ servant à les entraîner sont indépendantes ; laquelle n'est pas toujours réaliste selon le contexte.

Les modèles linéaires mixtes (LMM) permettent donc d'insérer une structure de dépendance entre les observations d'un LM. Du côté des GLM, il est possible d'effectuer un ajustement au modèle pour que la matrice de variance de celui-ci puisse tenir compte d'éventuelles covariances entre les observations. De tels modèles ajustés sont appelés des GEE (*Generalized Estimating Equation*).

L'objet de ce travail pratique est de mettre en pratique ces deux types de modèles de régression. Ainsi, les questions 1 et 2 abordent le sujet des modèles linéaires mixtes tandis que la question 3 aborde le sujet des GEE.

1 Modèle linéaire mixte pour les résultats en mathématique

Pour la première question de ce travail, on s'intéresse aux données d'un sous-ensemble des étudiants de 8ème année ayant participé au *Nationnal Educationnal Longitudinal Study* de 1988. L'objectif de cette étude est de voir comment les résultats en mathématiques varient en fonction du nombre d'heures de travail à la maison (variable `homework` dans la base de données). Dans ce cas-ci, on a que la variable endogène Y_{ij} correspond au résultat de l'examen de mathématique de l'étudiant j, appartenant à l'école i, $i = 1, \ldots, 23$, $j = 1, \ldots, n_i$, où n_i correspond au nombre d'étudiants appartenant à la i-ème grappe. Dans la base de données, cette variable est désignée comme `math`.

Comme la variable explicative `meanses`, correspondant au statut socio-économique moyen des étudiants de l'école, est fortement corrélée avec l'école d'origine des étudiants (variable de grappe), alors on est intéressé de voir quelle différence il y aurait entre un LMM avec et sans cette variable et de voir si le besoin d'effets aléatoires dans le modèle persiste si on ajoute cette dernière.

1a) Exlusion de la variable `meanses`

Pour débuter l'entraînement d'un modèle, la première étape est de considérer un LM, d'évaluer ses résidus pour voir si les postulats sont respectés et de prendre action autrement. À noter que les LMM permettent de régler les problèmes d'auto-corrélation des résidus et, dans certains cas, de régler l'hétéroscédasticité.

Entraînement d'un modèle linéaire

On considère le modèle linéaire suivant :

$$Y_{ij} = \beta_0 + \beta_1 \text{homework} + \beta_2 \text{white} + \beta_3 \text{ratio} + \epsilon_{ij}. \tag{1}$$

Une analyse de multicolinéarité réalisée conformément à la méthodologie décrite dans le travail pratique 1 du présent cours ne soulève aucun problème. En revanche, lorsque l'on regarde l'illustration 1, on voit que les résidus ont une légère tendance descendante. Le postulat de linéarité n'est donc pas respecté. Pour remédier à ce problème, on regarde les splines générés par un modèle additif généralisé (GAM) à l'aide de la fonction `gam` du *package* du même nom. Ceux-ci sont présentés dans l'illustration 4.

On voit donc dans l'illustration 4a qu'il est possible de passer une droite dans l'intervalle de confiance entourant le spline. Pour cette variable, aucune transformation n'est donc nécessaire. Pour ce qui est de la variable `ratio`, il est impossible de passer une telle droite. Conséquemment, il faudrait ajouter un terme de deuxième degré sur la variable ratio qui aurait été centrée et réduite au préalable. Pour se faire, on pose

$$\text{ratio}^* = \frac{\text{ratio} - 18}{\sigma_{\text{ratio}^*}} \quad \text{et} \quad \text{ratio2} = (\text{ratio}^*)^2.$$

Le modèle linéaire devient alors

$$Y_{ij} = \beta_0 + \beta_1 \text{homework} + \beta_2 \text{white} + \beta_3 \text{ratio}^* + \beta_4 \text{ratio2} + \epsilon_{ij}. \tag{2}$$

Si on refait l'exercice du GAM, on trouve l'illustration 5. Dans celle-ci, on remarque qu'aucune transformation additionnelle n'est nécessaire. Si on se fie à la statistique F produite par la fonction `summary` en `R`, alors on trouve qu'aucune autre transformation n'est significative au seuil de 1%. Par la suite, on peut tenter d'ajouter des interactions. Au seuil de 5%, les interactions ajoutées sont `white:ratio` et `white:ratio2`, ce qui permet d'obtenir le modèle (3)

$$\begin{aligned} Y_{ij} = \beta_0 + \beta_1 \text{homework} + \beta_2 \text{white} + \beta_3 \text{ratio}^* + \beta_4 \text{ratio2} \\ + \beta_5 (\text{white} : \text{ratio}^*) + \beta_6 (\text{white} : \text{ratio2}) + \epsilon_{ij}. \end{aligned} \tag{3}$$

L'illustration 2 permet de voir que la linéarité semble meilleure, mais que le problème est désormais au niveau de l'hétéroscédasticité des résidus. Or, un LMM peut aider à traiter ce genre de problème. Du point de vue de l'auto-corrélation des observations du jeu de données, l'illustration 6 permet de voir que, selon l'école d'appartenance des élèves, les résidus du LM ne sont pas identiquement distribués. En effet, selon la grappe, on voit que les résidus ont une variance et une moyenne qui peut différer. Cette observation vient donc légitimer l'utilisation d'un LMM.

Entraînement d'un modèle linéaire mixte

Maintenant qu'un modèle linéaire a été entraîné et que l'on a observé la nécessité d'y inclure des effets aléatoires pour tenir compte de la corrélation qui existe entre les élèves d'une même école, il est temps d'entraîner un LMM.

Pour se faire, la première étape est de faire les graphiques des résidus en fonction des différentes variables pour voir sur lesquelles d'entre elles il serait intéressant appliquer un effet aléatoire. Bien que cela soit difficile à voir, l'illustration 8 montre que la variable la plus susceptible d'avoir un effet aléatoire est `homework` puisqu'elle est celle ayant le plus de volatilité dans la distribution des résidus selon les valeurs qu'elle peut prendre. Ainsi, avec la fonction `lmer` du `package R lme4`, on entraîne le modèle (4) avec les structures de variance VC [1] pour la variance des résidus et UN [2], de même que UN(1) [3] pour la variance des effets aléatoires.

$$\begin{aligned} Y_{ij} = \beta_0 + \gamma_{i0} + (\beta_1 + \gamma_{i1}) \text{homework} + \beta_2 \text{white} + \beta_3 \text{ratio}^* \\ + \beta_4 \text{ratio2} + \beta_5 (\text{white} : \text{ratio}^*) + \beta_6 (\text{white} : \text{ratio2}) + \epsilon_{ij}. \end{aligned} \tag{4}$$

À noter que l'ajout de trop d'effets aléatoires dans le modèle testé peut entraîner de l'instabilité numérique lors de l'entraînement de celui-ci. C'est pourquoi on se limite à deux effets aléatoires dans le modèle (4). Pour la même raison, si l'effet fixe est trop complexe (trop de termes d'ordre supérieur), la méthode ne convergera pas. C'est pourquoi les interactions trouvées plus tôt ont été sélectionnées selon la méthode algorithmique de type *forward* avec un seuil de test de 5%.

En ce qui attrait à la structure de variance des résidus de type CS [4], comme la fonction `lmer` ne permet pas de l'utiliser, on peut faire appel à la fonction `lme` du *package* `nlme`. Pour ce qui est de la structure AR(1) [5], celle-ci n'a que peu de sens dans ce contexte puisque les observations (les élèves d'une même école) ne peuvent pas être ordonnées selon un ordre chronologique ou spatial. Pour cette raison, on ne considérera pas cette dernière.

Comme les trois modèles testés possèdent tous la même composante fixe ($\boldsymbol{X\beta}$), alors on peut comparer les log-vraisemblances de même que les AIC. Le tableau 1 présente donc l'AIC pour chacun des modèles testés.

1. *Variance Components* : indépendance entre les résidus.
2. *Unstructured* : chaque combinaison d'effets aléatoires a un coefficient de corrélation différent.
3. *Diagonales principales* : les effets aléatoires sont indépendants l'un de l'autre.
4. *Compound symmetry* : la corrélation entre les résidus est la même partout.
5. *Auto-régression d'ordre 1* : la corrélation diminue selon un aspect d'éloignement (généralement pour les observations qui sont étudiées à travers le temps ou l'espace).

$\text{Var}(\boldsymbol{\epsilon})$	$\text{Var}(\boldsymbol{\gamma})$	*dl*	AIC
VC	UN	11.00	3630.72
VC	UN(1)	10.00	3658.96
CS	UN	12.00	3621.86

Tableau 1 – AIC des trois modèles testés en fonction de (4) avec le nombre de degrés de liberté *dl* associé à chacun d'eux.

Avec le tableau 1, on voit que la structure de variance qui minimise l'AIC est CS/UN. Cependant, avec la fonction `summary` de `R`, on voit que le coefficient de corrélation des résidus d'une même classe est de 5.126496×10^{-18}, ce qui est très près de zéro. On peut donc simplifier le modèle et prendre la structure VC/UN. Puis on voit que la corrélation entre les effets aléatoires γ_{i0} et γ_{i1} est de -0.89, ce qui confirme qu'il existe un lien de dépendance significatif entre ces variables aléatoires et que la structure de variance UN est approprié. En somme, la corrélation entre les étudiants d'une même école est négligeable et il existe un lien de dépendance significatif entre les effets aléatoires du modèle.

Après avoir sélectionné les structures de variance du LMM, il faut tester si les effets aléatoires du modèle (4) sont nécessaires. Pour se faire, il s'agit de procéder à un test du ratio des vraisemblances. Soit les hypothèses de test suivantes :

$$H_0 : \text{Le modèle simple est suffisant ;}$$
$$H_1 : \text{Le modèle complet représente mieux les données.}$$

Soit l_0 et l_1, la log-vraisemblance sous H_0 et celle sous H_1. On définit Δ_{dl} comme la différence du nombre de paramètres entre les deux modèles. Le calcul de la *p-value* du test est effectué avec (5).

$$p\text{−value} = 0.5\left[2 - \mathbb{P}(\chi^2_{\Delta_{dl}-1} > \xi) - \mathbb{P}(\chi^2_{\Delta_{dl}} > \xi)\right], \tag{5}$$

On applique ainsi (5) pour évaluer si l'effet aléatoire γ_{i1} est significatif et on trouve une statistique de test de 92.92 avec $\Delta_{dl} = 2$, ce qui permet de calculer un seuil observé de 0. Conséquemment, on rejette fortement H_0 et on conserve l'effet aléatoire γ_{i1}. De plus, puisque γ_{i1} est conservé, on ne peut retirer l'ordonnée à l'origine aléatoire. Le modèle obtenu suite à cette étape de construction du LMM correspond ainsi à (6).

$$\begin{aligned} Y_{ij} = {} & \beta_0 + \gamma_{i0} + (\beta_1 + \gamma_{i1})\text{homework} + \beta_2\text{white} + \beta_3\text{ratio}^* \\ & + \beta_4\text{ratio2} + \beta_5(\text{white} : \text{ratio}^*) + \beta_6(\text{white} : \text{ratio2}) + \epsilon_{ij}. \end{aligned} \tag{6}$$

Il ne reste plus qu'à sélectionner les effets fixes. Pour se faire, on utilise le test de Wald de type III utilisé par la fonction `Anova` du *package* `car`. On remarque alors que la variable `ratio`* possède un seuil de test de 15.69%. Cependant, comme on ne peut la retirer sans avoir retiré les variables dépendantes d'elle au préalable, c.-à-d. `white:ratio2`, `white:ratio`* et `ratio2`, on ne peut pas l'enlever. Conséquemment, on va commencer par retirer l'interaction `white:ratio2` avant de réeffectuer le test. Puis, on retire aussi `white:ratio`* puisque la variable `ratio`* n'est toujours pas significative au seuil de 5%. On fait de même avec `ratio2` pour finalement retirer `ratio`*. On trouve ainsi le modèle final (7).

$$Y_{ij} = \beta_0 + \gamma_{i0} + (\beta_1 + \gamma_{i1})\text{homework} + \beta_2\text{white} + \epsilon_{ij}. \tag{7}$$

Avec la fonction `summary` de `R`, on obtient les résultats présentés dans l'illustration 10. D'une part, on a les effets fixes pour lesquels un intervalle de confiance à 95% est calculé dans le tableau 2.

	Estimateur	Écart-type	IC 95%	
β_0	44.02	1.83	40.42	47.62
β_1	1.90	0.92	0.11	3.70
β_2	3.30	0.98	1.38	5.22

Tableau 2 – Estimateurs des poids pour les effets fixes du LMM ainsi que leurs intervalles de confiance à 95%.

D'autre part, on a

$$\boldsymbol{D}_i = \text{Var}(\boldsymbol{\gamma}_i) = \begin{pmatrix} 58.20797 & -27.01225 \\ -27.01225 & 17.25707 \end{pmatrix}, i = 1, \ldots, 23 \text{ et } \boldsymbol{D} = \begin{pmatrix} D_1 & 0 & \ldots & 0 \\ 0 & D_2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \ldots & 0 & D_{23} \end{pmatrix} \tag{8}$$

De plus,

$$\boldsymbol{V} = \text{Var}(\boldsymbol{\epsilon}) = 52.66\, \boldsymbol{I}_{n\times n},\ n = \sum_{i=1}^{23} n_i = 519. \tag{9}$$

Discussion

Avec le tableau 2, on voit que, toute autre chose étant égale, chaque heure de travail supplémentaire à la maison contribue à augmenter la note moyenne d'un étudiant pour son examen de mathématique de 1.90%. Cet effet est significatif puisque l'intervalle de confiance à 95% de l'estimateur n'inclut pas la valeur 0. Par ailleurs, comme on a pu l'observer lors de l'étape du test des effets aléatoires, l'effet du nombre d'heures de travail à la maison peut varier d'une école à l'autre.

1b) Inclusion de la variable `meanses`

Comme pour la partie 1a), on commence par entraîner un modèle linéaire mixte et on procède de façon similaire pour trouver le modèle (10).

$$\begin{aligned} Y_{ij} = {} & \beta_0 + \beta_1 \text{meanses} + \beta_2 \text{homework} + \beta_3 \text{white} + \beta_4 \text{ratio}^* + \beta_5 \text{ratio2} \\ & + \beta_6(\text{white} : \text{ratio}) + \beta_7(\text{white} : \text{ratio2}) + \beta_8(\text{meanses} : \text{white}) + \beta_9(\text{meanses} : \text{ratio2}) + \epsilon_{ij}. \end{aligned} \tag{10}$$

À noter que l'interaction des variables `homework` et `meanses` n'est pas significatif au seuil de 5%.

En comparant les illustrations 6 et 7, on voit que l'ajout de la variable `meanses` au modèle semble, a priori, régler le problème de corrélation entre les observations. En effet, on voit dans l'illustration 7 que les résidus de chacune des écoles semblent centrées à zéro. Cependant les variance varient encore quelque peu. Voyons maintenant si l'ajout d'effets aléatoires serait significatif.

Entraînement d'un modèle linéaire mixte

Pour débuter, l'illustration 9 montre que, encore une fois, seule la variable `homework` est susceptible de recevoir un effet aléatoire. Afin de confirmer cette observation, on peut entraîner un LMM ne comportant que deux effets aléatoires, soit une ordonnée à l'origine et un effet sur l'une des variables explicatives parmi `homework`, `meanses` et `ratio`*. On teste ainsi chacune des variables avec les trois structures mentionnées ci-haut, soit VC/UN, VC/UN(1) et CS/UN. Il en découle que le modèle qui minimise l'AIC, est celui incluant un effet aléatoire à la variable `homework`. Par la suite, si on tente d'ajouter un troisième effet aléatoire, on obtient que les fonctions `lmer` et `lme` deviennent instables numériquement. On s'en tiendra donc à 2 effets. En ce qui attrait aux structures de variances, celle qui minimise l'AIC est la structure CS/UN. Cependant, comme dans la section 1a), on a un coefficient de corrélation pour la variance de $\boldsymbol{\epsilon}_i$ qui est de 5.126496×10^{-18}. La dépendance entre les résidus d'une même grappe est donc négligeable et on peut simplifier le modèle en adoptant la structure VC/UN. Plus encore, la corrélation entre l'ordonnée à l'origine aléatoire et l'effet appliqué à la variable `homework` est de -0.91 confirmant ainsi que la structure UN est adéquate pour la variance de $\boldsymbol{\gamma}$.

Par la suite, on effectue le test du ratio des vraisemblances dont le calcul du seuil observé est présenté en (5). On trouve ainsi une statistique de test de 90.95 avec $\Delta_{dl} = 2$, pour un seuil observé de 0. L'évidence est donc forte contre l'hypothèse nulle et on peut en conclure que l'effet aléatoire ajouté à la variable `homework` est significatif.

Pour ce qui est de la sélection des effets fixes, on a que la variable ayant le plus grand seuil observé avec le test de Wald de type III est `ratio`*. Cependant, comme pour la section 1a), on doit gérer les variables d'ordre supérieur qui dépendent de celle-ci avant de pouvoir la retirer. On commence donc par retirer l'interaction `white:ratio2`. Puis, on refait le test pour retirer `ratio2` ; ainsi de suite jusqu'à trouver le modèle (11) où tous les effets fixes sont significatifs au seuil de 5%.

$$Y_{ij} = \beta_0 + \gamma_{i0} + \beta_1 \text{meanses} + (\beta_2 + \gamma_{i2})\text{homework} + \beta_3 \text{white} + \epsilon_{ij}. \tag{11}$$

Si on essaie d'intégrer l'interaction `meanses:homework`, on trouve un seuil de test de 0.7368 ; on ne l'inclue donc pas dans le modèle. La sortie `R` de la fonction `summary` appliquée sur le modèle ainsi obtenu est présentée dans l'illustration 11. Les effets fixes sont décrits dans le tableau 3 et les matrices de variances sont présentées dans (12) et (13), lesquelles sont très similaires que dans la section 1a).

	Estimateurs	Écarts-types	IC 95%	
β_0	44.70	1.79	41.20	48.21
β_1	4.89	1.34	2.26	7.52
β_2	1.93	0.90	0.17	3.68
β_3	3.11	0.96	1.24	4.99

Tableau 3 – Estimateurs des poids pour les effets fixes du LMM (11) ainsi que leurs intervalles de confiance à 95%.

$$\boldsymbol{D}_i = \text{Var}(\boldsymbol{\gamma}_i) = \begin{pmatrix} 53.57617 & -27.00012 \\ -27.00012 & 16.39948 \end{pmatrix}, i = 1, \ldots, 23 \quad \text{et} \quad \boldsymbol{D} = \begin{pmatrix} D_1 & 0 & \ldots & 0 \\ 0 & D_2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \ldots & 0 & D_{23} \end{pmatrix} \tag{12}$$

$$\boldsymbol{V} = \text{Var}(\boldsymbol{\epsilon}) = 52.79033\, \boldsymbol{I}_{n\times n}, \; n = \sum_{i=1}^{23} n_i = 519. \tag{13}$$

Discussion

Comme on a pu l'observer avec l'illustration 7, l'ajout de la variable `meanses` qui est très fortement corrélée avec les identifiants des écoles (les grappes) a réduit considérablement le besoin d'ajouter des effets aléatoires au modèle puisque les résidus sont maintenant centrés autour de zéro. Néanmoins, avec le test du ratio des vraisemblances, on a pu voir que l'effet aléatoire appliqué à la variable `homework`, de même que l'ordonnée à l'origine aléatoire, sont utiles.

Au final, on trouve qu'une personne qui n'a pas étudié peut espérer, en moyenne, une note de 44.70%. Cette moyenne augmente à 47.81% pour une personne de couleur blanche et elle augmente encore d'avantage si l'élève est issu d'un milieu aisé. Finalement, l'ajout d'une heure supplémentaire d'étude augmente, en moyenne, l'espérance de la note en mathématique de 1.925% et cet effet varie d'une école à l'autre (effet aléatoire).

2 Modèle linéaire mixte pour la grandeur de jeunes filles

Pour cette deuxième question, le jeu de données à l'étude présente 20 courbes de la croissance de jeunes filles mesurées annuellement entre les âges 6 à 10 ans. Celui-ci a été publié par Gildstein (1979). Dans ce cas-ci, la variable endogène Y_{ij} correspond à la taille de la i-ème fille lors de sa j-ème mesure à l'âge $5 + j$, $i = 1, \ldots, 20,\ j = 1, \ldots, 5$. On cherche ici à déterminer si la grandeur de la mère de l'enfant (représentée par la variable `group`) influence la grandeur de la fille.

Entraînement d'un modèle linéaire

Afin de voir si la relation qui existe entre l'âge des jeunes filles et leur grandeur est linéaire, on regarde l'illustration 12. Comme celle-ci l'est effectivement, on entraîne le modèle (14).

$$Y_{ij} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{group}_2 + \beta_3 \text{group}_3 + \beta_4 (\text{group}_2 : \text{age}) + \beta_5 (\text{group}_3 : \text{age}) + \epsilon_{ij}. \tag{14}$$

Comme (14) possède un facteur d'inflation de la variance généralisé ($\text{GVIF}_j^{1/(2p_j)}$) supérieur à $\sqrt{10} = 3.16$, on est en présence de multicolinéarité. Pour remédier à ce problème, on peut simplement tronquer la variable `age` de la manière suivante :

$$\text{temps} = \text{age} - 6.$$

Le modèle (14) devient alors (15).

$$Y_{ij} = \beta_0 + \beta_1 \text{temps} + \beta_2 \text{group}_2 + \beta_3 \text{group}_3 + \beta_4 (\text{group}_2 : \text{temps}) + \beta_5 (\text{group}_3 : \text{temps}) + \epsilon_{ij}. \tag{15}$$

Avec ce dernier, on calcule les résidus studentisés de manière à générer l'illustrations 13. Dans un premier temps, on remarque avec l'illustration 13a que les résidus ne sont pas tous centrés autour de zéro. Dépendamment de la fillette, ceux-ci ont une moyenne qui diffère grandement, ce qui laisse présager une corrélation entres les observations d'une même fillette. Cela suggère qu'un LMM pourrait régler le problème d'auto-corrélation des résidus.

Entraînement d'un modèle linéaire mixte

Dans un deuxième temps, on remarque avec les illustrations 13a et 13b que les deux graphiques sont pratiquement identiques, laissant présager qu'un effet aléatoire sur la variable `temps` n'aurait aucune incidence sur les résidus. Plus encore, avec l'illustration 13c, on voit que les résidus varient énormément selon la valeur de la variable `group`. Cela pourrait expliquer en partie les ordonnées à l'origine des résidus qui diffèrent dans 13a.

Voyons maintenant si ces observations s'avèrent réalistes en entraînant un LMM avec les structures de variances VC/UN, CS/UN et AR(1)/UN. À noter que la structure AR(1) pour la variance des résidus est particulièrement intéressante dans ce contexte puisque les observations d'une même fillettes peuvent être ordonnées chronologiquement. À cet effet, avec ce dernier, l'ajout de la variable `temps` dans les effets aléatoires engendre des problèmes de convergence avec la fonction `lme`. Conséquemment, le modèle entraîné à cette étape consiste en (16).

$$\begin{aligned} Y_{ij} = {} & \beta_0 + \gamma_{i0} + \beta_1 \text{temps} + (\beta_2 + \gamma_{i2}) \text{group}_2 \\ & + (\beta_3 + \gamma_{i3}) \text{group}_3 + \beta_4 (\text{group}_2 : \text{temps}) + \beta_5 (\text{group}_3 : \text{temps}) + \epsilon_{ij}. \end{aligned} \tag{16}$$

L'AIC calculé pour chacun des modèles entraîné est présenté dans le tableau 4

$\text{Var}(\boldsymbol{\epsilon})$	$\text{Var}(\boldsymbol{\gamma})$	dl	AIC
VC	UN	13.00	356.72
CS	UN	14.00	360.87
AR(1)	UN	14.00	336.09

Tableau 4 – AIC des trois modèles testés en fonction de (16) avec le nombre de degrés de liberté dl associé à chacun d'eux.

On voit avec le tableau 4 que la structure de variance la plus appropriée selon le critère de l'AIC pour le modèle (16) est AR(1)/UN. De plus, le coefficient de corrélation liant les résidus d'une même fillette est de 0.9041, ce qui est hautement significatif. Plus encore, la matrice des coefficients de corrélation des effets aléatoire s'exprime comme

$$\boldsymbol{\rho}(\boldsymbol{\gamma}_i) = \begin{pmatrix} 1 & 0.380 & -0.147 \\ 0.380 & 1 & -0.813 \\ -0.147 & -0.813 & 1 \end{pmatrix}, \ i = 1, \ldots, 20.$$

Ainsi, la structure de variance non structurée (UN) est justifiée pour les effets aléatoires puisque les coefficients de corrélations sont significativement différents de zéro.

Si on fait un test du ratio des vraisemblances pour l'effet aléatoire appliqué sur la variable `group`, on trouve une statistique de test de 1.39 pour $\Delta_{dl} = 5$, ce qui donne un seuil observé de 0.885. Ainsi, on ne peut rejeter l'hypothèse nulle que l'effet aléatoire associé à la variable `group` n'est pas significatif et on peut le retirer. Si on refait le test sur l'ordonnée à l'origine aléatoire, on trouve une statistique de 185.54 pour $\Delta_{dl} = 2$, ce qui donne un seuil observé de 0. On ne peut donc pas retirer l'ordonnée à l'origine aléatoire. Mentionnons que pour ce dernier test, comme on compare un modèle linéaire (H_0) à un modèle mixte (H_1), on ne peut utiliser la méthode REML pour calculer la log-vraisemblance de ce dernier lors du test puisque le modèle linéaire associé à l'hypothèse nulle utilise le maximum de vraisemblance.

Au niveau des effets fixes, le test de Wald de type III indique que tous les effets fixes sont significatifs au seuil de 5%. On obtient donc le modèle final (17)

$$\begin{aligned} Y_{ij} = \beta_0 + \gamma_{i0} + \beta_1 \text{temps} + \beta_2 \text{group}_2 + \beta_3 \text{group}_3 \\ + \beta_4(\text{group}_2 : \text{temps}) + \beta_5(\text{group}_3 : \text{temps}) + \epsilon_{ij}. \end{aligned} \tag{17}$$

La sortie `R` de la fonction `summary` appliquée sur le modèle ainsi obtenu est présentée dans l'illustration 14. Les effets fixes sont décrits dans le tableau 5 et les matrices de variances sont présentées dans (18) et (19).

	Estimateurs	Écarts-types	IC 95%	
β_0	112.57	1.22	110.18	114.96
β_1	3.70	1.66	0.44	6.96
β_2	7.79	1.66	4.54	11.05
β_3	5.29	0.19	4.92	5.65
β_4	0.26	0.25	-0.23	0.76
β_5	0.87	0.25	0.37	1.37

Tableau 5 – Estimateurs des poids pour les effets fixes du LMM (17) ainsi que leurs intervalles de confiance à 95%.

$$\boldsymbol{D} = 5.088543 \times 10^{-6}\, \boldsymbol{I}_{20\times 20} \tag{18}$$

et

$$\boldsymbol{V}_i = \begin{pmatrix} 2.988 & 2.838 & 2.696 & 2.560 & 2.432 \\ 2.838 & 2.988 & 2.838 & 2.696 & 2.560 \\ 2.696 & 2.838 & 2.988 & 2.838 & 2.696 \\ 2.560 & 2.696 & 2.838 & 2.988 & 2.838 \\ 2.432 & 2.560 & 2.696 & 2.838 & 2.988 \end{pmatrix}, \; i = 1, \ldots, 20, \; \boldsymbol{V} = \begin{pmatrix} V_1 & 0 & \ldots & 0 \\ 0 & V_2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \ldots & 0 & V_{20} \end{pmatrix} \tag{19}$$

Discussion

L'interprétation des résultats obtenus dans le tableau 5 est synthétisée dans le tableau 6.

Taille de la mère	Taille de la fille à 6 ans	Taux de croissance annuel	Grandeur de la fille à 10 ans
petite	112.57	3.70	127.37
moyenne	120.36	3.96	136.20
grande	117.86	4.57	136.14

Tableau 6 – Mesures moyennes(en cm) pour une petite fille qui est née en 1973 selon la grandeur de la mère.

À la lumière de ces résultats, on peut répondre à la question de Goldstein (1979) en affirmant que oui, la croissance des filles est liée à la taille de la mère. L'interaction `group:temps` qui est significative au seuil de 0.001731 en atteste et les résultats du tableau 6 l'illustre bien.

3 GEE pour le nombre d'auto-administrations de doses analgésiques

Cette question a comme objectif d'utiliser un GEE afin de modéliser l'effet addictif des analgésiques auto-administrés par des patients suite à une chirurgie à l'abdomen. Pour ce contexte, la variable endogène Y_{ij} représente le nombre de doses auto-administrées par le i-ème patient lors de la j-ème période de 4 heures de l'étude, $i = 1, \ldots, 65$, $j = 1, \ldots, 12$. Les patients sont divisés en deux groupes : l'un qui reçoit une dose de 1 mg par injection ($x_i = 0$) et l'autre qui reçoit 2 mg de morphine ($x_i = 1$).

3a) Entraînement du modèle GEE

Comme il s'agit d'un modèle de dénombrement, la famille de GLM qui est la plus appropriée est la loi de Poisson et son lien canonique est le lien logarithmique. On définit ainsi le modèle complet comme

$$\mu_{ij} = \exp\left\{\beta_0 + \beta_1 x_i + \beta_2 j + \beta_3 x_i j\right\}, \quad Y_{ij}|x_i \sim \text{Poisson}(\mu_{ij}) \tag{20}$$

Cependant, comme il a été mentionné en introduction, les modèles de type GLM supposent que toutes les observations d'un jeu de données sont indépendantes l'une de l'autre. Or, comme les 780 observations du jeu de données sont réparties en 65 grappes d'observations corrélées (puisqu'elles proviennent d'un même patient), alors cette hypothèse n'est plus valide. Il faut donc ajuster le modèle de régression à cette réalité avec un GEE. Pour se faire, la fonction `gee` du *package* `R` du même nom permet de réaliser l'entraînement du modèle et de calculer les différentes quantités dont nous aurons besoin.

L'entraînement d'un tel modèle débute par la sélection de la matrice de corrélation de travaille. Comme la taille des grappes $\{n_i\}_{i=1}^K$ est relativement petit et est égal pour chacune des grappes ($n_i = 12\ \forall i$), alors une façon efficace de sélectionner la structure adéquate est de commencer par entraîner un modèle avec une matrice non structurée (UN) pour avoir un aperçu des corrélations empiriques. Puis, si on ne voit aucun schéma dans cette matrice, on peut simplement la conserver. Si on voit que les corrélations sont très près de zéro, on utilisera la structure d'indépendance. Si la corrélation semble uniforme pour tous les couples d'observation $\{(j, j');\ j, j' = 1, \ldots, 12,\ j \neq j'\}$, on sélectionnera la structure échangeable. Finalement, si les coefficients de corrélation semblent s'atténuer au fil des observations d'une même grappe, alors on pourra considérer la structure auto-régressive d'ordre 1 (AR(1)). Pour le cas présent, on présente la matrice des corrélations pour les 6 premières observations parmi les 12 de chacune des grappes dans (21).

$$R_i(\alpha)^{\text{UN}}_{[1:6,1:6]} = \begin{pmatrix} 1.000 & 0.478 & 0.447 & 0.349 & 0.635 & 0.313 \\ 0.478 & 1.000 & 0.431 & 0.381 & 0.264 & 0.135 \\ 0.447 & 0.431 & 1.000 & 0.733 & 0.503 & 0.432 \\ 0.349 & 0.381 & 0.733 & 1.000 & 0.651 & 0.518 \\ 0.635 & 0.264 & 0.503 & 0.651 & 1.000 & 0.651 \\ 0.313 & 0.135 & 0.432 & 0.518 & 0.651 & 1.000 \end{pmatrix} \tag{21}$$

Ainsi, on voit que la corrélation tend à diminuer lorsque les observations sont distancées. Conséquemment, la structure de corrélation de type auto-régressive d'ordre 1 serait appropriée. Par ailleurs, ce résultat est peu surprenant puisque les observations d'une grappe correspondent à des périodes de temps pouvant être ordonnancées chronologiquement et l'interprétation de ce résultat est cohérent dans l'optique où l'état du patient a évolué dans le temps, de même que son niveau de dépendance. C'est donc ce type de matrice de corrélation de travail qui est choisit pour la suite. En entraînant le GEE avec cette structure, on obtient un coefficient de corrélation, ici dénoté α, qui vaut 0.5164258. De ce fait, la matrice de corrélation de travail s'exprime comme

$$R_i(\alpha)^{\text{AR(1)}} = \begin{pmatrix} 1 & \alpha & \ldots & \alpha^{11} \\ \alpha & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \alpha \\ \alpha^{11} & \ldots & \alpha & 1 \end{pmatrix}, \ i = 1, \ldots, 65, \ \alpha = 0.5164258295. \tag{22}$$

La première ligne de cette matrice correspond donc à

$$R_i(\alpha)^{\text{AR(1)}}_{[1,1:12]} = \begin{pmatrix} 1.000 & 0.516 & 0.267 & 0.138 & 0.071 & 0.037 & 0.019 & 0.010 & 0.005 & 0.003 & 0.001 & 0.001 \end{pmatrix}.$$

Il ne reste plus qu'à effectuer la sélection des variables explicatives. D'abord, on a

$$\hat{\boldsymbol{\beta}} \approx N(\boldsymbol{\beta}, \boldsymbol{V_S}), \tag{23}$$

où $\boldsymbol{V_S}$ est la matrice de variance robuste. Cette dernière est calculée par la fonction `gee` et la racine carrée des composantes de sa diagonale son présentées dans l'attribut `coefficients`, dans la colonne `Robust S.E.` Sous l'hypothèse nulle que les coefficients de régression valent zéro, alors la statistique de test pour la l-ème variable explicative du modèle (20) correspond à (24).

$$z_l = \frac{\hat{\beta}_l}{\sqrt{V_{ll}}},\ l = 0, 1, 2, 3, \tag{24}$$

où V_{ll} correspond à la composante de la diagonale de la matrice $\boldsymbol{V_S}$ assignée à la l-ème variable explicative. Cette statistique est calculée par la fonction `gee` et est présentée également avec l'attribut `coefficients` dans la colonne `Robust z`. Ainsi, avec ces outils en main, il est aisé de calculer le seuil observé du test pour vérifier l'hypothèse nulle que la l-ème variable explicative est inutile à la régression. Pour se faire, il suffit d'utiliser (25) pour effectuer un test bilatéral.

$$p\text{−value} = 2\mathbb{P}\left[N(0,1) > |z_l|\right] \tag{25}$$

Ainsi, avec (25), on trouve que le seuil observé du test est de 0.4265 pour la variable x_i, assignée au groupe de dosage et de 0.1299 pour l'interaction entre les deux variables explicative. Comme pour les questions précédentes du présent travail, on ne peut retirer une variable qui est utilisée pour une interaction avant d'avoir retiré la dite interaction. Conséquemment, c'est cette dernière qui est retirée en premier. En repassant le test sur le modèle ajusté (sans l'interaction), on trouve désormais que la seuil observé de la variable x_i est descendu à 0.0147. On la conserve donc au seuil de 5%. La variable j (assignée à la période du traitement), quant à elle, possède un seuil observé de zéro. Le modèle final est donc présenté en (26).

$$\mu_{ij} = \exp\left\{\beta_0 + \beta_1 x_i + \beta_2 j\right\} \tag{26}$$

et les estimateurs des paramètres de régression sont

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 2.05378263 \\ 0.30742284 \\ -0.06091784 \end{pmatrix}.$$

À la lumière de ces résultats, on observe qu'un patient qui reçoit la dose minimale (1 mg) commencera par s'administrer, en moyenne, environ 7 doses. Pour une personne qui reçoit 2 mg par dose, elle débutera, en moyenne, avec près de 10 doses. Puis, dans les deux cas, la demande moyenne en dose diminue de 94.09% à toutes les périodes de 4 heures.

3b) Prédiction pour la population

On désire désormais donner un estimé ponctuel et un intervalle de confiance à 95% pour l'effet moyen de la dose dans cette population de patients. Pour se faire, on définit le vecteur $c' = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}$. Puis on calcule la prédiction pour le prédicteur linéaire avec (27).

$$\hat{\eta} = c'\hat{\boldsymbol{\beta}}. \tag{27}$$

Par la suite, on trouve l'intervalle de confiance pour η avec (28).

$$\eta \in \hat{\eta} \pm z_{0.975}\sqrt{c'\boldsymbol{V_S}c}. \tag{28}$$

Finalement, l'estimé ponctuel et l'intervalle de confiance pour l'effet moyen de la dose est trouvé en portant $\hat{\eta}$ ainsi que son intervalle de confiance à l'exposant. De cette manière, on obtient un effet moyen de 1.359916 avec un intervalle de confiance à 95% de $\begin{pmatrix} 1.062251 & 1.740993 \end{pmatrix}$. En conclusion, le fait d'avoir une dose de 2 mg augmente la demande moyenne en doses de 36% par rapport à quelqu'un qui reçoit des doses de 1 mg. Cet effet peut varier entre 6% et 74%.

A Graphiques

A.1 Question 1

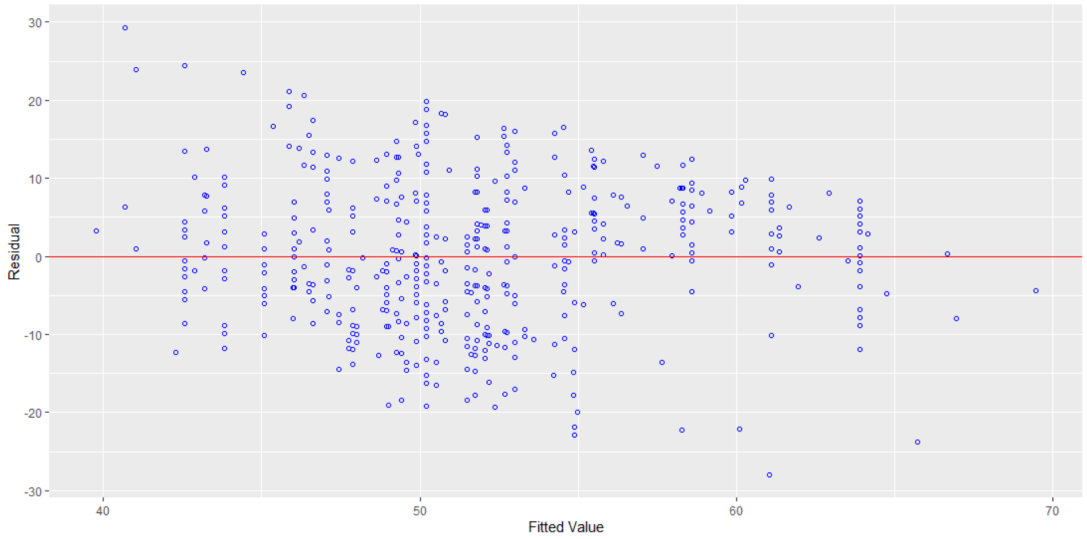


Illustration 1 – Résidus en fonction des valeurs prédites pour le modèle (1).

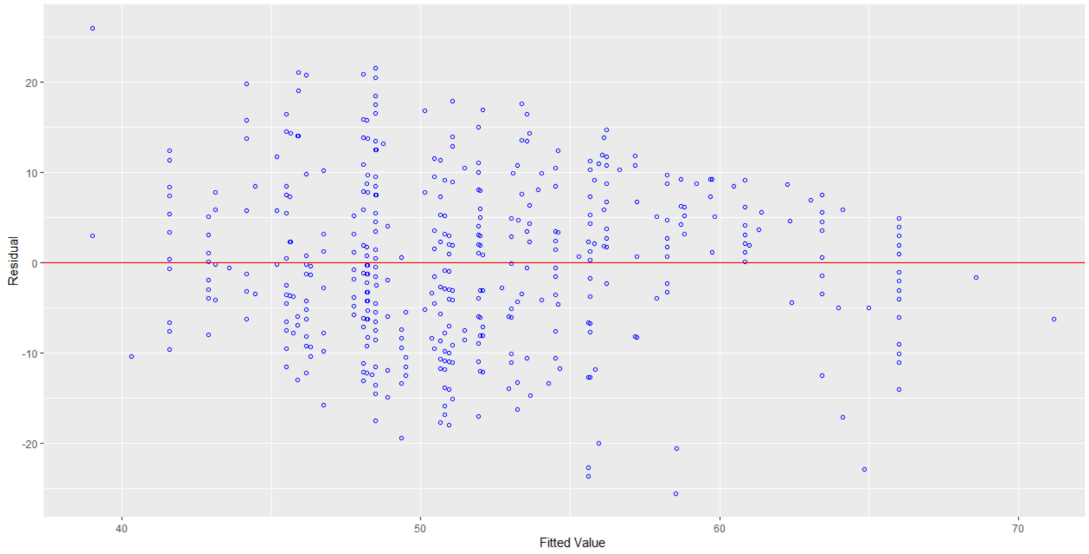


Illustration 2 – Résidus en fonction des valeurs prédites pour le modèle (3).

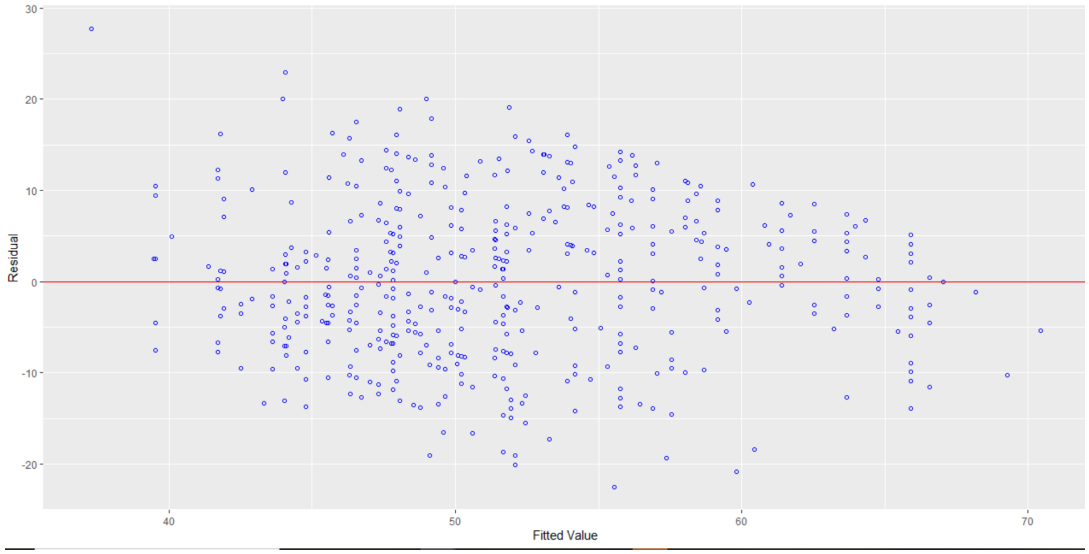
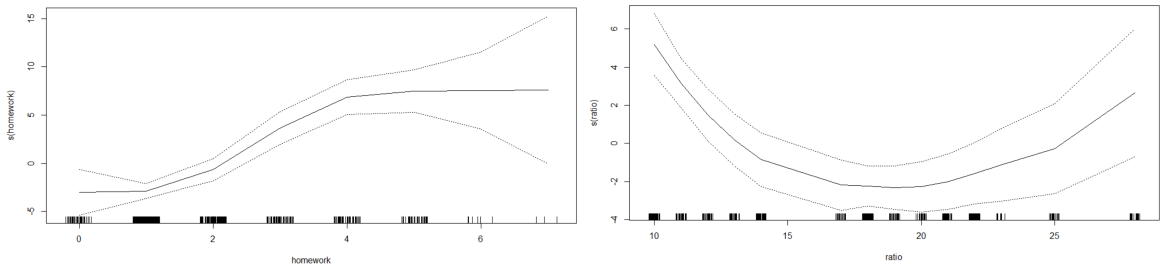


Illustration 3 – Résidus en fonction des valeurs prédites pour le modèle (10).



(a) `homework` (b) `ratio`

Illustration 4 – Splines réalisés sur les variables `homework` et `ratio` lors de l'entraînement d'un GAM.

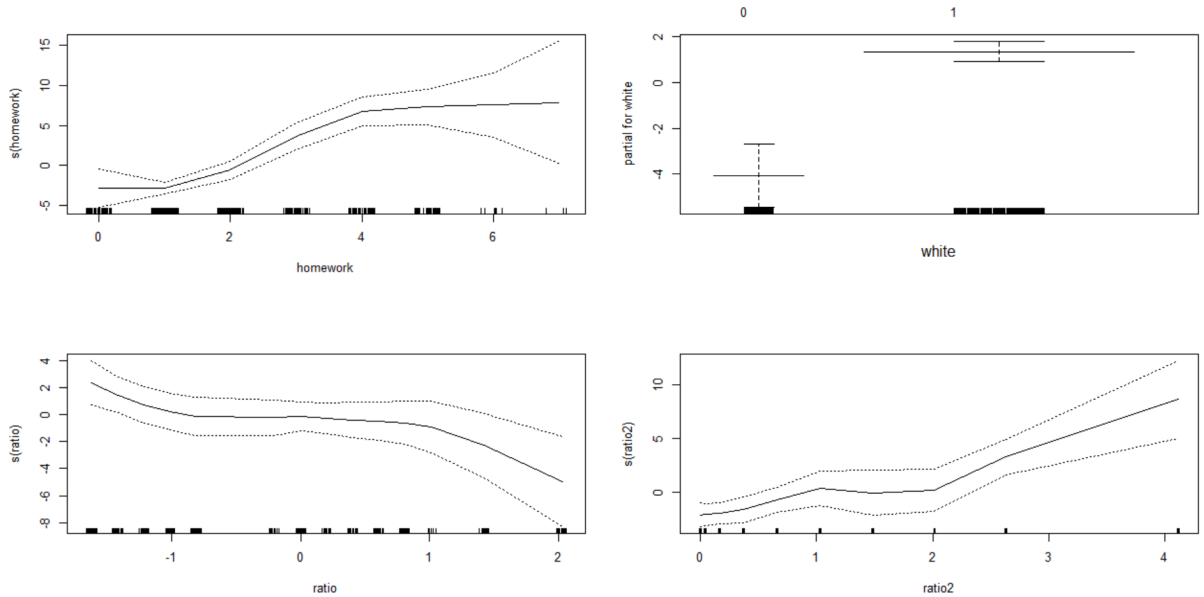


Illustration 5 – Visualisation des Splines suite à l'entraînement d'un GAM utilisant (2).

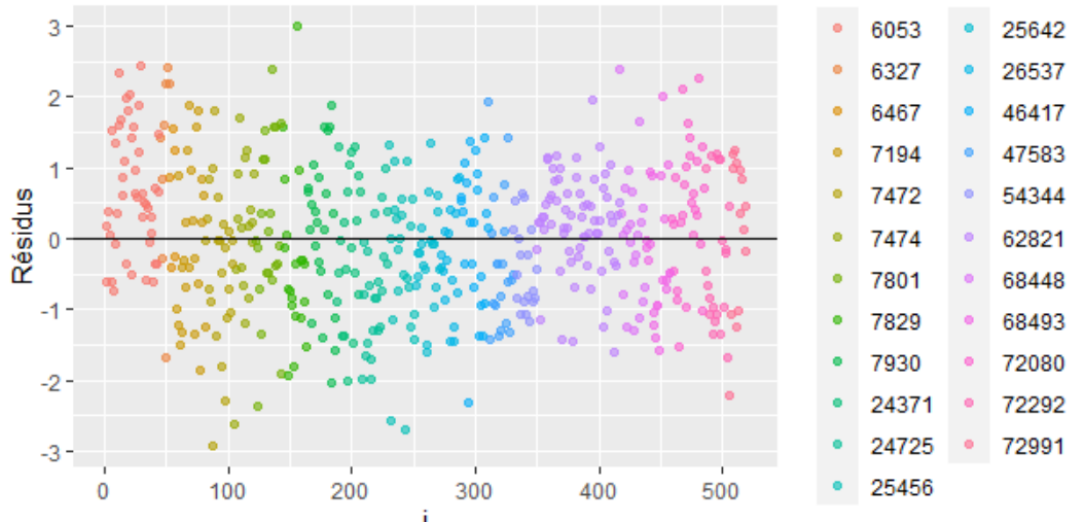


Illustration 6 – Résidus studentisés en fonction de l'index des observations pour le modèle 3.

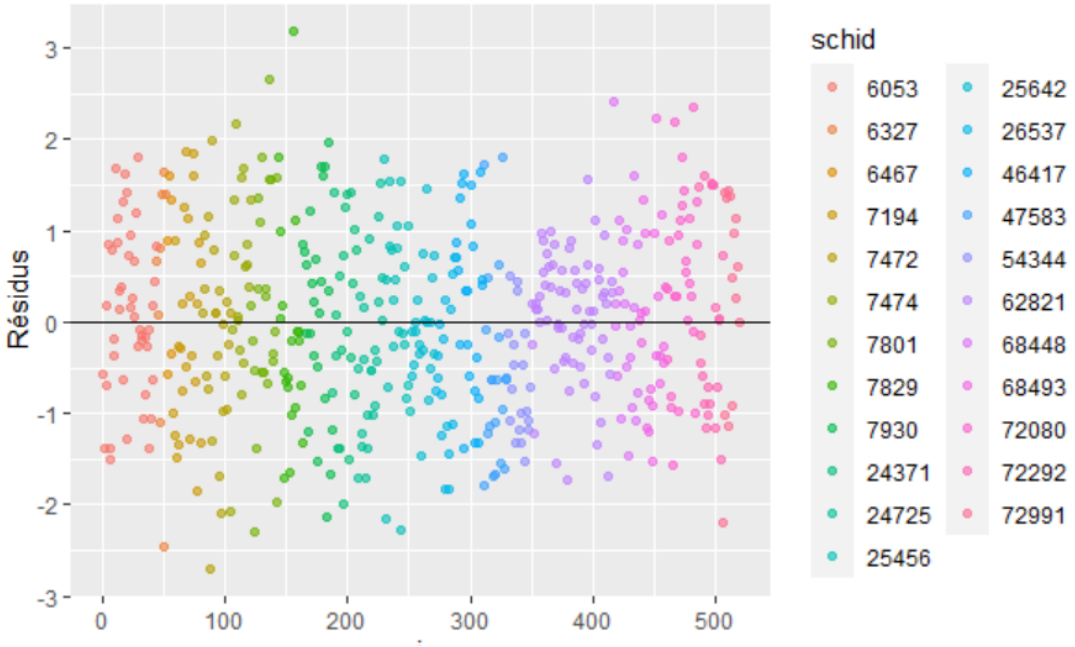
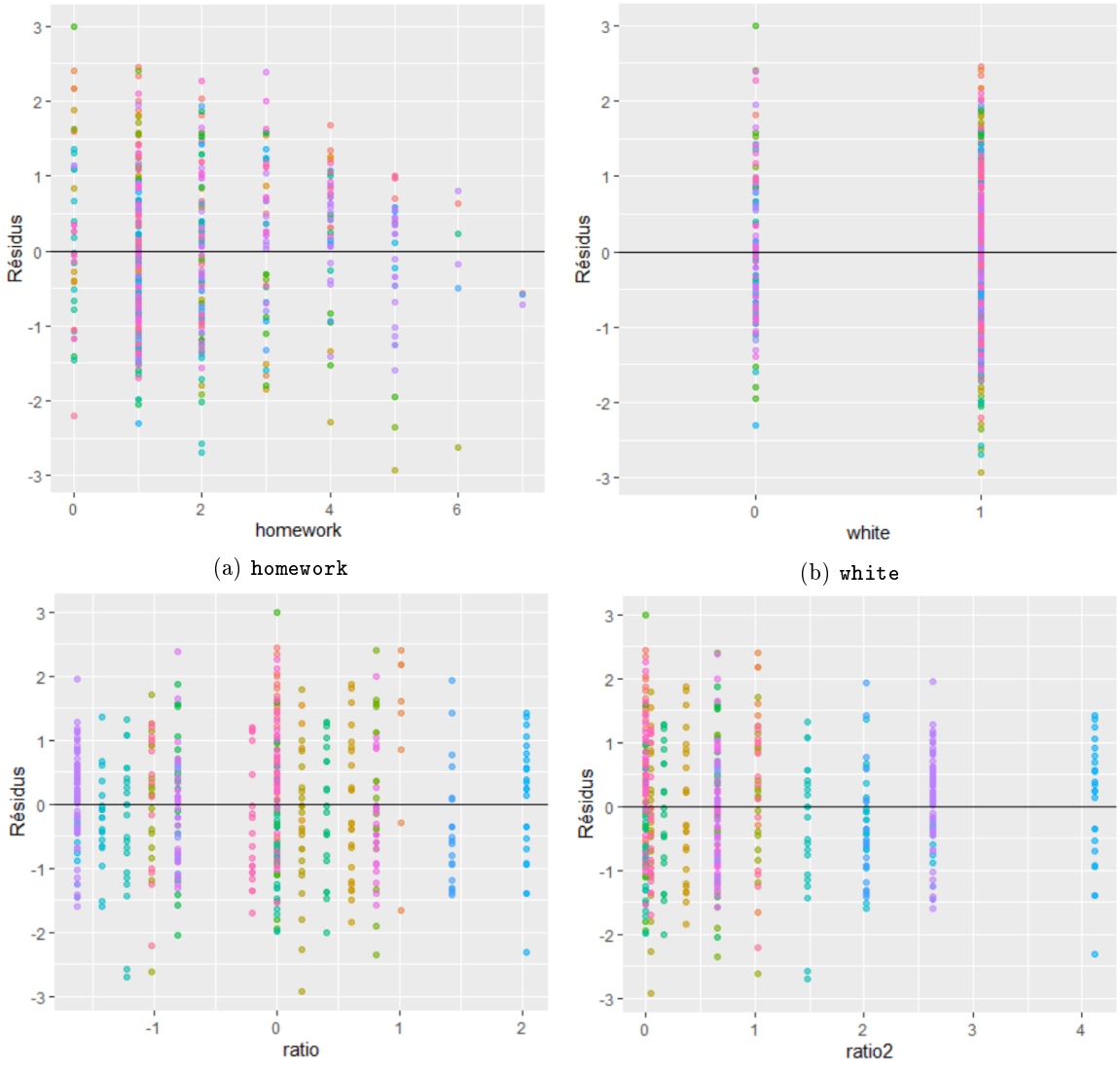


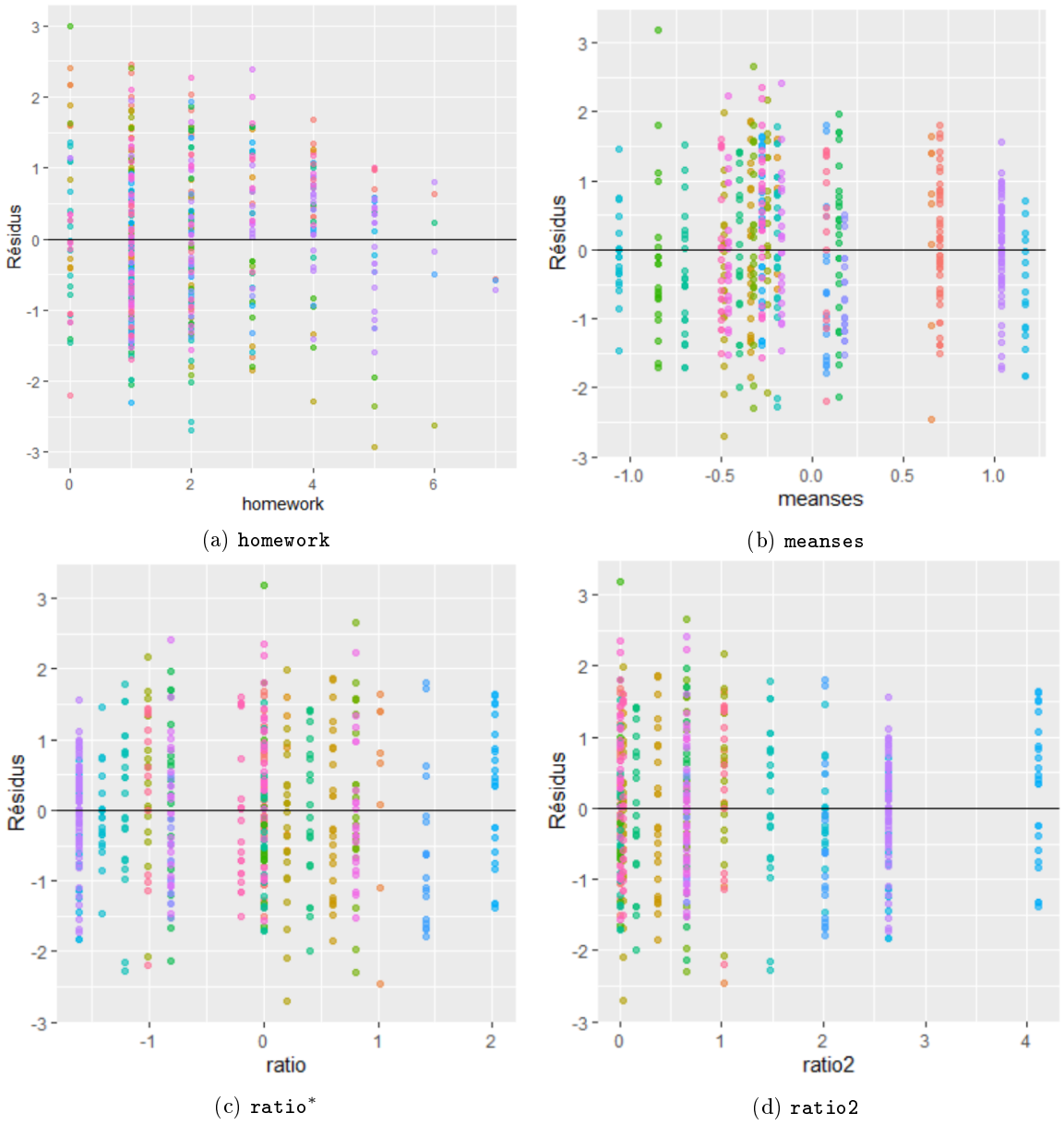
Illustration 7 – Résidus studentisés en fonction de l'index des observations pour le modèle 10.



(a) `homework` (b) `white`

(c) `ratio`* (d) `ratio2`

Illustration 8 – Résidus studentisés en fonction des différentes variables explicatives du modèle 3.



(a) `homework` (b) `meanses`

(c) `ratio`* (d) `ratio2`

Illustration 9 – Résidus studentisés en fonction des différentes variables explicatives du modèle 10.

```
Linear mixed model fit by REML ['lmerMod']
Formula: math ~ homework + white + (homework | schid)
   Data: data

REML criterion at convergence: 3622.8

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.30724 -0.66634 -0.03254  0.68200  3.03431

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 schid    (Intercept) 58.21    7.629
          homework    17.26    4.154    -0.85
 Residual             52.66    7.257
Number of obs: 519, groups:  schid, 23

Fixed effects:
            Estimate Std. Error t value
(Intercept)  44.0198     1.8349  23.990
homework      1.9031     0.9168   2.076
white1        3.3000     0.9781   3.374

Correlation of Fixed Effects:
          (Intr) homwrk
homework -0.773
white1   -0.371 -0.027
```

Illustration 10 – Sortie `R` de la fonction `summary` pour le modèle (7).

```
Linear mixed model fit by REML ['lmerMod']
Formula: math ~ meanses + homework + white + (homework | schid)

REML criterion at convergence: 3610

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.29715 -0.68843 -0.01309  0.68012  2.98973

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 schid    (Intercept) 53.58    7.320
          homework    16.40    4.050    -0.91
 Residual             52.79    7.266
Number of obs: 519, groups:  schid, 23

Fixed effects:
            Estimate Std. Error t value
(Intercept)  44.7022     1.7873  25.012
meanses       4.8925     1.3406   3.649
homework      1.9251     0.8952   2.151
white1        3.1149     0.9570   3.255

Correlation of Fixed Effects:
         (Intr) meanss homwrk
meanses   0.139
homework -0.813 -0.006
white1   -0.384 -0.126 -0.026
```

Illustration 11 – Sortie `R` de la fonction `summary` pour le modèle (11).

A.2 Question 2

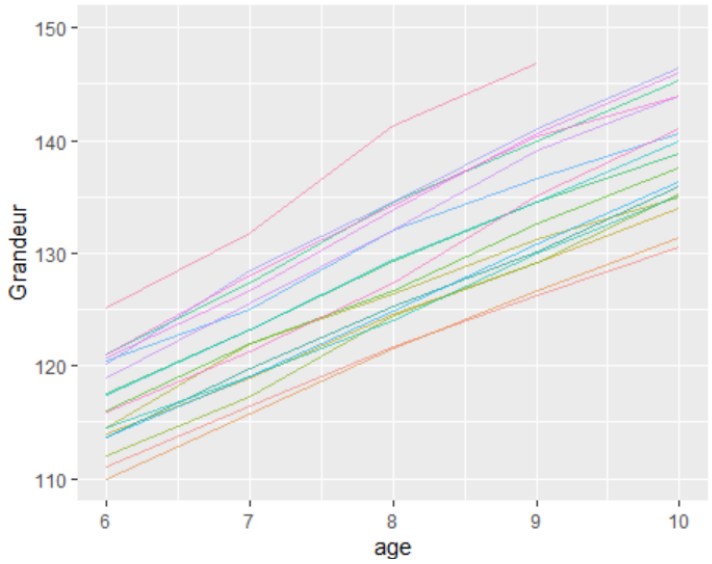
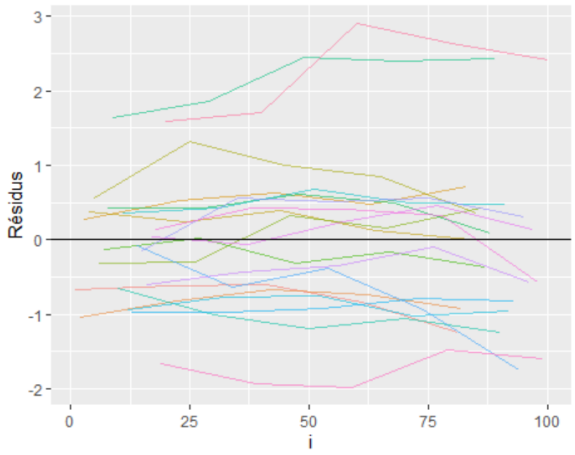
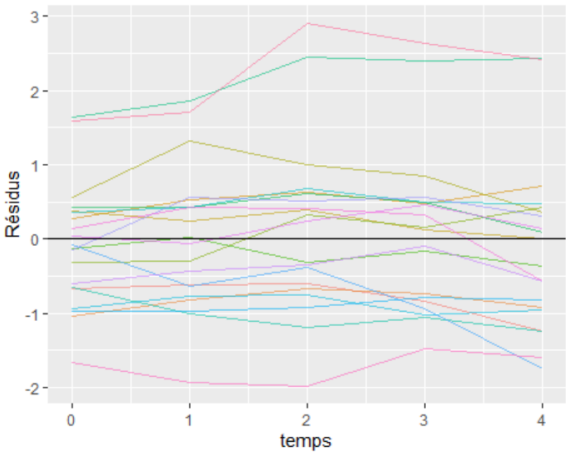


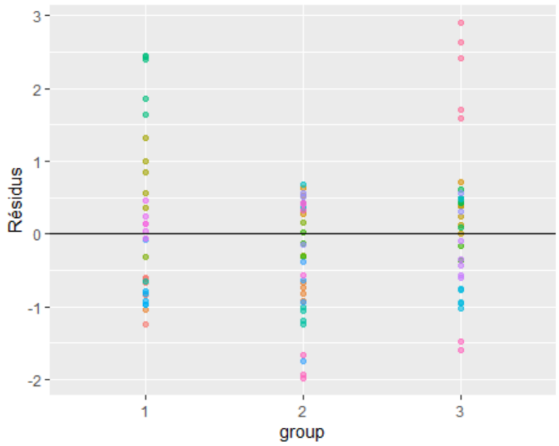
Illustration 12 – Relation de la grandeur en fonction de l'âge pour chacune des jeunes filles.



(a) `i`



(b) `temps`



(c) `group`

Illustration 13 – Graphiques de résidus générés à partir du modèle (15).

```
Linear mixed-effects model fit by REML
 Data: data
       AIC      BIC    logLik
  327.4851 350.3747 -154.7425

Random effects:
 Formula: ~1 | child
        (Intercept) Residual
StdDev:  0.00225578 2.987839

Correlation Structure: AR(1)
 Formula: ~1 | child
 Parameter estimate(s):
      Phi
0.9498482
Fixed effects: height ~ group + temps + group:temps
                Value Std.Error DF  t-value p-value
(Intercept)  112.57455 1.2196105 77 92.30369  0.0000
group2         3.70038 1.6620490 17  2.22640  0.0398
group3         7.79495 1.6620490 17  4.68996  0.0002
temps          5.28740 0.1859984 77 28.42710  0.0000
group2:temps   0.26271 0.2534731 77  1.03643  0.3032
group3:temps   0.87029 0.2534731 77  3.43348  0.0010
 Correlation:
             (Intr) group2 group3 temps  grp2:t
group2       -0.734
group3       -0.734  0.538
temps        -0.305  0.224  0.224
group2:temps  0.224 -0.305 -0.164 -0.734
group3:temps  0.224 -0.164 -0.305 -0.734  0.538
```

Illustration 14 – Sortie `R` de la fonction `summary` pour le modèle (17).