Théorie et applications des méthodes de régression

Travail 2

Présenté Par

Rahma Jebali
Hajar Taqif
Oumaima Ouffy

Décembre 2020

Table des matières

Table des figures

1 Exercice 1

Dans cet exercice nous traitons les données d'un sous-ensemble des étudiants de 8ème année ayant participé au National Educational Longitudinal Study de 1988 (analysés par Kreft & Leeuw (1998)).
L'objectif est de voir comment les résultats en mathématiques varient en fonction du nombre d'heures de travail à la maison à l'aide d'un ajustement d'un modèle linéaire mixte en fonction des différentes variables disponibles.

1.1 Chargement des données

Les données sont disponibles dans un fichier TXT contenant plusieurs variables comme mentionné dans la figure 1.

schid	stuid	meanses	homework	white	public	ratio	math	sex
6053	1	0.6997727	1	1	0	18	50	2
6053	2	0.6997727	1	1	0	18	43	2
6053	4	0.6997727	3	0	0	18	50	2
6053	11	0.6997727	1	1	0	18	49	2
6053	12	0.6997727	1	1	0	18	62	1
6053	13	0.6997727	1	1	0	18	43	2
6053	18	0.6997727	1	1	0	18	42	1
6053	22	0.6997727	4	1	0	18	68	1
6053	23	0.6997727	1	0	0	18	41	1
6053	24	0.6997727	5	1	0	18	62	1
6053	25	0.6997727	1	1	0	18	69	2
6053	26	0.6997727	0	1	0	18	60	2
6053	27	0.6997727	4	1	0	18	71	1
6053	28	0.6997727	1	1	0	18	56	1
6053	32	0.6997727	1	0	0	18	47	2
6053	33	0.6997727	1	1	0	18	58	2
6053	34	0.6997727	1	1	0	18	66	1
6053	36	0.6997727	2	0	0	18	41	1
6053	39	0.6997727	2	0	0	18	60	2
6053	42	0.6997727	2	1	0	18	69	1
6053	43	0.6997727	1	1	0	18	44	1
6053	44	0.6997727	1	1	0	18	61	2

FIGURE 1 – Échantillon des données d'un sous-ensemble des étudiants de 8ème année ayant participé au National Educational Longitudinal Study de 1988

Dans ce travail, nous allons considérer que les variables :
schid : numéro d'identification de l'école
meanses statut socio-économique moyen des étudiants de l'école
homework : nombre d'heures de travail de l'étudiant à la maison par semaine
white : est 1 pour blanc, 0 pour minorité visible
ratio : nombre d'étudiants par enseignant dans les classes de chaque école i
math : est les résultats de l'étudiant en mathématiques dans l'école (la variable réponse)

1.2 L'effet des heures de travail à la maison sur les résultats en mathématiques sans considérer la variable meanses

1. Modèle linéaire ordinaire

Dans cette étape nous avons ajusté un modèle linéaire ordinaire large (qui contient plusieurs interaction possible) données par l'équation 2 :

$$Y_{ij} = \beta_0 + \beta_1 \, x_{i1} + \beta_2 \, x_{wi} + \beta_3 \, x_{ri} + \beta_4 \, x_{i1} \, x_{ri} + \beta_5 x_{i1} \, x_{wi} + \epsilon_{ij} \qquad (1)$$

Oû :
i : numéro d'identification de l'école (schid)
Y_{ij} : est les résultats de l'étudiant j en mathématiques (math) dans l'école i
x_{ij1} : est le nombre d'heures de travail de l'étudiant i à la maison par semaine (homework)
x_{wi} : est 1 pour blanc, 0 pour minorité visible (white)
x_{ri} : nombre d'étudiants par enseignant dans les classes de chaque école i (ratio).

```
Call:
lm(formula = math ~ homework + white + ratio + +homework:ratio +
    homework:white, data = Q1)

Residuals:
     Min       1Q   Median       3Q      Max
-27.6629  -6.9295   0.2273   6.1853  27.2902

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    45.40758    2.67373  16.983  < 2e-16 ***
homework        3.13681    1.05162   2.983  0.00299 **
white           3.99216    1.62962   2.450  0.01463 *
ratio          -0.11610    0.14056  -0.826  0.40918
homework:ratio -0.09228    0.05296  -1.742  0.08203 .
homework:white  1.31025    0.74666   1.755  0.07989 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.16 on 513 degrees of freedom
Multiple R-squared:  0.2754,	Adjusted R-squared:  0.2683
F-statistic: 38.99 on 5 and 513 DF,  p-value: < 2.2e-16
```

Figure 2 – Modéle linéaire mixte

Ce modèle montre que la variable homework a un effet significatif de $\hat{\beta}_1 = 3.13$.

2. Modèle linéaire mixte

Vu que le nombre d'étudiant par école n_i varient d'une école à une autre, il peut y avoir une corrélation entre les données d'une même école. Alors, nous avons ajusté un modèle linéaire mixte.

Pour ce fait, nous avons choisi de commencer avec un modèle mixte d'une ordonnée à l'origine aléatoire et d'une pente aléatoire liée à la variable homework (nous n'avons pas introduit des pentes aléatoires aux interaction et à la variable ratio à cause du message d'erreur de divergence dans l'ajustement du modèle). Ensuite, nous avons procédé à choisir les formes des matrices V et D en comparant les AIC des deux options VC pour V, UN pour D et VC pour V, UN(1) pour D (le AIC le plus faible est le meilleur) (figure 3) :

```
# Choix des formes des matrices D et V
# Option 1: VC pour V, UN pour D
Resultat.VCUN_1 <- lmer(math~homework+ratio+white+homework:ratio+homework:white+
                        (homework|schid),data=Q1,REML=TRUE)
extractAIC(Resultat.VCUN_1)
# Option 1b: VC pour V, UN(1) pour D
Resultat.VCUN0 <-lmer(math~homework+ratio+white+homework:ratio+homework:white+
                        (homework||schid)
                          ,data=Q1,REML=TRUE)
extractAIC(Resultat.VCUN0)
```

10 3647.14287691068

9 3670.59329712432

FIGURE 3 – Choix des formes des matrices V et D

Nous remarquons d'aprés les figure 3 que l'AIC du modèle avec la forme UN pour la matrice D est inférieur à l'AIC du modèle avec la forme UN(1) pour la matrice D ($3647.15 < 3670.59$). Nous avons choisi alors la première option.

Ensuite, nous avons effectué un test d'hypothèse sur la possibilité de réduire le modèle avec seulement une ordonnée à l'origine aléatoire (figure 4) :

```
# Modele avec seulement l'ordonnee a l'origine aleatoire
Resultat.VCUN_2 <- lmer(math~homework+ratio+white+homework:ratio+homework:white+
                        (1|schid),data=Q1,REML=TRUE)
# Test du rapport des vraisemblances pour H_0: Resultat.VCUN_2
# vs H1: Resultat.VCUN_1
xi1 <- 2*(logLik(Resultat.VCUN_1)-logLik(Resultat.VCUN_2))
pval1 <- 0.5*(1-pchisq(xi1,1))
pval1
```

```
'log Lik.' 0 (df=10)
```

FIGURE 4 – test d'hypothèse sur la pente aléatoire

La valeur du pvalue obtenue est près de $0 << 0.05$, donc nous avons rejeté H0 et nous avons gardé la pente aléatoire.
Par la suite, nous avons procédé à sélectionner les effets fixes les plus importantes en utilisant la méthode BACKWARD.

```
# Selection des effets fixes, methode BACKWARD
Anova(Resultat.VCUN_1,type=3)
```

	Chisq	Df	Pr(>Chisq)
(Intercept)	46.64228074	1	8.520194e-12
homework	0.54254721	1	4.613790e-01
ratio	0.03054866	1	8.612511e-01
white	2.99683853	1	8.342717e-02
homework:ratio	0.06440605	1	7.996628e-01
homework:white	0.04997208	1	8.231119e-01

FIGURE 5 – Tableau ANOVA du modèle complet

D'après le tableau ANOVA du modèle complet (figure 5), nous remarquons que l'interaction entre les variables homework et white est non significatif au seuil 5%. Alors nous avons procédé à l'enlever.

```
Resultat.VCUN_3 <- lmer(math~homework+ratio+white+homework:ratio+
                        (homework|schid),data=Q1,REML=TRUE)
Anova(Resultat.VCUN_3,type=3)
```

	Chisq	Df	Pr(>Chisq)
(Intercept)	47.28359553	1	6.142413e-12
homework	0.59224050	1	4.415540e-01
ratio	0.03467446	1	8.522795e-01
white	11.61160520	1	6.554156e-04
homework:ratio	0.05822070	1	8.093306e-01

FIGURE 6 – Tableau ANOVA du modèle sans l'interaction entre les variables homework et white

D'après le tableau ANOVA du modèle de la figure 6, nous remarquons que l'interaction entre les variables homework et ratio est non significatifau au seuil 5%. Alors nous avons procédé à l'enlever.

```
Resultat.VCUN_4 <- lmer(math~homework+ratio+white+
                        (homework|schid),data=Q1,REML=TRUE)
Anova(Resultat.VCUN_4,type=3)
```

	Chisq	Df	Pr(>Chisq)
(Intercept)	150.7089152	1	1.213417e-34
homework	4.2965380	1	3.819004e-02
ratio	0.5402388	1	4.623338e-01
white	11.6105547	1	6.557859e-04

FIGURE 7 – Tableau ANOVA du modèle sans l'interaction entre les variables homework et ratio

D'après le tableau ANOVA du modèle de la figure 7, nous remarquons que la variable ratio est non significatif au seuil 5%. Alors nous avons procédé à l'enlever.

```r
Resultat.VCUN_5 <- lmer(math~homework+white+
                      (homework|schid),data=Q1,REML=TRUE)
Anova(Resultat.VCUN_5,type=3)
```

	Chisq	Df	Pr(>Chisq)
(Intercept)	575.50806	1	3.557749e-127
homework	4.30861	1	3.791994e-02
white	11.38244	1	7.414185e-04

FIGURE 8 – Tableau ANOVA du modèle sans la variable ratio

D'après la figure 8, nous remarquons que les deux variables homework et white sont significatifs au seuil 5%. Nous avons gardé alors ce dernier modèle.

```r
summary(Resultat.VCUN_5)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: math ~ homework + white + (homework | schid)
   Data: Q1

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
white         3.3000     0.9781   3.374

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

FIGURE 9 – Résumé du modèle final obtenu

D'après la figure 8 et 9, nous pouvons conclure que les heures de travail à la maison (la variable homework) ont un effet significatif sur les résultats en mathématiques puisque son pvalue $= 3.79\text{e-}02 < < 0.05$ et son $\hat{\beta}_1 = 1.9031$. De plus, l'existence de la pente aléatoire devant cette variable confirme la variation des résultats en mathématiques d'une école à une autre.

1.3 L'effet des heures de travail à la maison sur les résultats en mathématiques en considérant la variable meanses

Dans cette partie, nous avons ajusté un modèle linéaire mixte en considérant la possibilité d'inclure la variable meanses et son interaction avec le nombre d'heures de travail. Nous avons commencé par choisir les formes des matrices V et D. Nous avons comparé les AIC des deux options VC pour V, UN pour D et VC pour V, UN(1) pour D (figure 10) :

```
# Choix des formes des matrices D et V
# Option 1: VC pour V, UN pour D
Resultat.VCUN_b1 <- lmer(math~homework+meanses+white+ratio+homework:meanses+homework:ratio
                +homework:white+(homework|schid)+(homework:meanses|schid),data=Q1,REML=TRUE)
extractAIC(Resultat.VCUN_b1)
# Option 1b: VC pour V, UN(1) pour D
Resultat.VCUN_b0 <- lmer(math~homework+meanses+white+ratio+homework:meanses++homework:ratio
              +homework:white+(homework||schid)+(homework:meanses||schid),data=Q1,REML=TRUE)
extractAIC(Resultat.VCUN_b0)
```

```
boundary (singular) fit: see ?isSingular
```

15   3645.42660532887

```
boundary (singular) fit: see ?isSingular
```

13   3676.0639754391

FIGURE 10 – Choix des matrice V et D

Nous remarquons d'aprés les figure 10 que l'AIC du modèle avec la forme UN pour la matrice D est inférieur à l'AIC du modèle avec la forme UN(1) pour la matrice D ($3645.43 < 3676.06$). Nous avons choisi alors la première option. Ensuite, nous avons effectué des tests d'hypothèses sur la possibilité de la réduction des effets aléatoires du modèle. Nous avons commencé par tester la possibil d'enlever la pente aléatoire de l'interaction entre les variables homework et meanses :

```
# Modele avec seulement l'ordonnee la pente aleatoire de l'interaction
Resultat.VCUN_b2<- lmer(math~homework+meanses+white+ratio+homework:meanses+homework:ratio
                +homework:white+(homework|schid),data=Q1,REML=TRUE)
# Test du rapport des vraisemblances pour H_0: Resultat.VCUN_b2
# vs H1: Resultat.VCUN_b1
xib1 <- 2*(logLik(Resultat.VCUN_b1)-logLik(Resultat.VCUN_b2))
pvalb1 <- 0.5*(1-pchisq(xib1,1))
pvalb1
```

```
'log Lik.' 0.137837 (df=15)
```

FIGURE 11 – Test d'hypothèse sur la pente aléatoire de l'interaction entre les variables homework et meanses

La valeur du p_{val} obtenue est près de 0.138$>>$0.05, nous ne rejetons pas H0 et nous n'avons pas besoin d'avoir la pente aléatoire de cette interaction.

Puis, nous avons testé si nous pouvons enlever la pente aléatoire de la variable homework.

```
# Modele avec seulement l'ordonnee la pente aleatoire de l'interaction
Resultat.VCUN_b3<- lmer(math~homework+meanses+white+ratio+homework:meanses+homework:ratio
                +homework:white+(1|schid),data=Q1,REML=TRUE)
# Test du rapport des vraisemblances pour H_0: Resultat.VCUN_b3
# vs H1: Resultat.VCUN_b2
xib2 <- 2*(logLik(Resultat.VCUN_b2)-logLik(Resultat.VCUN_b3))
pvalb2 <- 0.5*(1-pchisq(xib2,1))
pvalb2
```

```
'log Lik.' 0 (df=12)
```

FIGURE 12 – Test d'hypothèse sur la pente aléatoire de la variable homework

La valeur du pvalue obtenue est près de 0< <0.05, nous rejetons H0 et nous avons gardé la pente aléatoire de la variable homework.
Par la suite, nous avons procédé à sélectionner les effets fixes les plus importantes en utilisant la méthode BACKWARD :

```
Anova(Resultat.VCUN_b2,type=3)
```

	Chisq	Df	Pr(>Chisq)
(Intercept)	44.752293624	1	2.236049e-11
homework	0.310356134	1	5.774617e-01
meanses	1.642924767	1	1.999247e-01
white	2.345103934	1	1.256772e-01
ratio	0.010230928	1	9.194329e-01
homework:meanses	0.070611984	1	7.904478e-01
homework:ratio	0.004893832	1	9.442287e-01
homework:white	0.079828176	1	7.775304e-01

FIGURE 13 – Tableau ANOVA du modèle complet

```r
Resultat.VCUN_b4<- lmer(math~homework+meanses+white+ratio+homework:meanses+homework:ratio
                   +(homework|schid),data=Q1,REML=TRUE)
Anova(Resultat.VCUN_b4,type=3)
```

	Chisq	Df	Pr(>Chisq)
(Intercept)	44.962820772	1	2.008111e-11
homework	0.344741006	1	5.571053e-01
meanses	1.585150761	1	2.080203e-01
white	10.179569383	1	1.420053e-03
ratio	0.006434945	1	9.360638e-01
homework:meanses	0.090423336	1	7.636397e-01
homework:ratio	0.002124293	1	9.632385e-01

FIGURE 14 – Tableau ANOVA du modèle sans l'interaction entre les variables homework et white

```r
Resultat.VCUN_b5<- lmer(math~homework+meanses+white+ratio+homework:meanses
                   +(homework|schid),data=Q1,REML=TRUE)
Anova(Resultat.VCUN_b5,type=3)
```

	Chisq	Df	Pr(>Chisq)
(Intercept)	1.800075e+02	1	4.828132e-41
homework	4.506308e+00	1	3.377006e-02
meanses	1.712308e+00	1	1.906862e-01
white	1.020998e+01	1	1.396829e-03
ratio	8.098602e-03	1	9.282934e-01
homework:meanses	1.140864e-01	1	7.355389e-01

FIGURE 15 – Tableau ANOVA du modèle sans l'interaction entre les variables homework et ratio

```r
Resultat.VCUN_b6<- lmer(math~homework+meanses+white+ratio
                        +(homework|schid),data=Q1,REML=TRUE)
Anova(Resultat.VCUN_b6,type=3)
```

	Chisq	Df	Pr(>Chisq)
(Intercept)	1.854784e+02	1	3.085743e-42
homework	4.611443e+00	1	3.175929e-02
meanses	1.169027e+01	1	6.282762e-04
white	1.021992e+01	1	1.389320e-03
ratio	5.945941e-03	1	9.385361e-01

FIGURE 16 – Tableau ANOVA du modèle sans l'interaction entre les variables homework et meanses

```r
Resultat.VCUN_b7<- lmer(math~homework+meanses+white
                        +(homework|schid),data=Q1,REML=TRUE)
Anova(Resultat.VCUN_b7,type=3)
```

	Chisq	Df	Pr(>Chisq)
(Intercept)	625.583541	1	4.564214e-138
homework	4.624741	1	3.151403e-02
meanses	13.318675	1	2.627758e-04
white	10.594001	1	1.134552e-03

FIGURE 17 – Tableau ANOVA du modèle sans la variable ratio

Comme mentionner dans les figures 14,15,16,17, selon la significativité des variables au seuil 5%, nous avons commencé par enlever l'interaction entre les variables homework et white ensuite l'interaction entre les variables homework et ratio ensuite l'interaction entre les variables homework et meanses et finalement nous avons enlevé la variable ratio.
Nous concluons que l'ajout de la caractéristique statut socio-économique moyen des étudiants de l'école au modèle n'a pas diminué le besoin d'inclure des effets aléatoires. En effet, nous avons toujours l'ordonnée à l'origine aléatoire et la pente aléatoire de la variable homework.

```r
summary(Resultat.VCUN_b7)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: math ~ homework + meanses + white + (homework | schid)
   Data: Q1

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
homework      1.9251     0.8952   2.151
meanses       4.8925     1.3406   3.649
white         3.1149     0.9570   3.255

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

FIGURE 18 – Résumé du modèle final obtenu

D'après la figure 17 et 18, nous pouvons conclure que les heures de travail

à la maison ont un effet significatif sur les résultats en mathématiques avec un pvalue = 3.15e-02 $<<$ 0.05 et un $\hat{\beta_1} = 1.9251$ et l'existence de la pente aléatoire devant cette variable confirme la variation des résultats en mathématiques d'une école à une autre.
Nous concluons aussi que la variance des effets aléatoires de l'ordonnée à l'origine et de la pente de la variable homework a diminué en rajoutant la variable meanses, respectivement de 58.21 et 17.26 (pour le modèle sans la variable meanses) en 53.58 et 16.40(pour le modèle avec la variable meanses) .

2 Exercice 2

L'objectif de cet exercice est de construire un modèle linéaire mixte qui estime la croissance des filles et de savoir si la croissance des filles est liée à la taille de leur mère.

2.1 Chargement des données

Le jeu de données GirlsGrowth.data est un fichier contenant les données de 3 variables explicatives (age ,group,child) et la grandeur des filles comme variable réponse (height).

	height	child	age	group
1	111.0	1	6	1
2	110.0	2	6	1
3	113.7	3	6	1
4	114.0	4	6	1
5	114.5	5	6	1
6	112.0	6	6	1
7	116.0	7	6	2
8	117.6	8	6	2
9	121.0	9	6	2
10	114.5	10	6	2
11	117.4	11	6	2
12	113.7	12	6	2
13	113.6	13	6	2
14	120.4	14	6	3
15	120.2	15	6	3
16	118.9	16	6	3
17	120.7	17	6	3
18	121.0	18	6	3
19	115.9	19	6	3
20	125.1	20	6	3
21	116.4	1	7	1
22	115.8	2	7	1

FIGURE 19 – Échantillon des données GirlsGrowth

2.2 Modèle linéaire ordinaire et visualisation des résidus

Dans cette étape nous avons ajusté un modèle linéaire ordinaire :

$$Y_{ij} = \beta_0 + \beta_1\, x_{ij1} + \beta_2\, x_{Gi} + \epsilon_{ij} \tag{2}$$

Oû :
i est l'enfant (child)

Y_{ij} est la grandeur de l'enfant i à l'age j (height)
x_{ij1} est l'âge de l'enfant au moment de la mesure
x_{Gi} est une variable qui vaut 1 si la mère est de petite taille, 2 si la mère est de taille moyenne et 3 si la mère est grande (group).
Afin de déterminer le besion des effets aléatoires, nous avons visualisé les résidus en fonction de la variable x_{ij1}

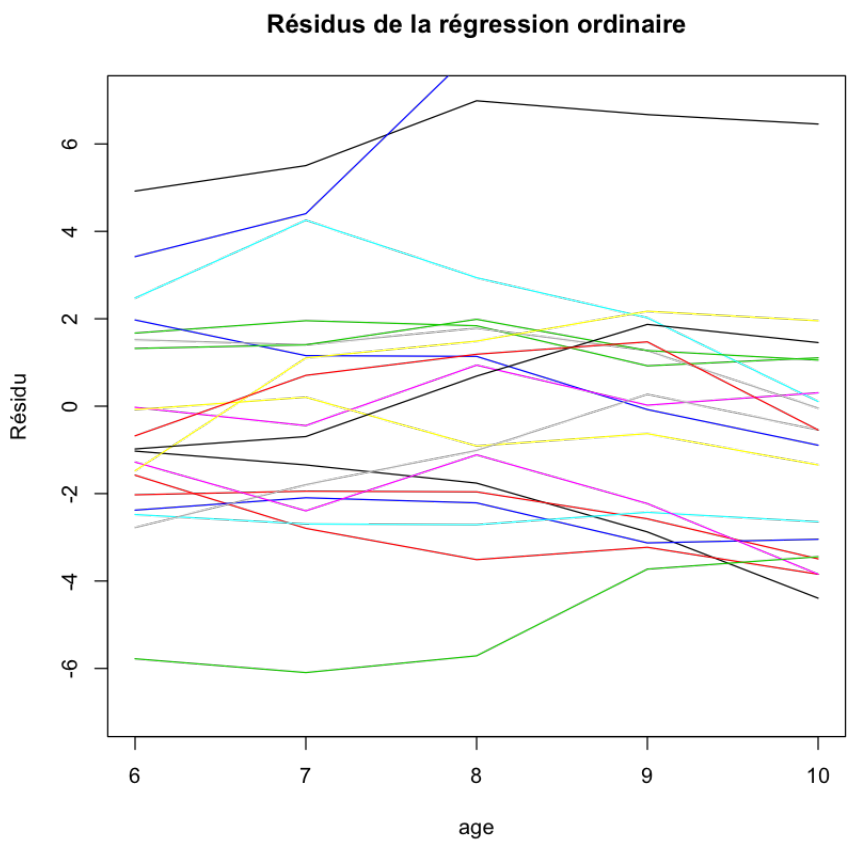

FIGURE 20 – Graphe des résidus en de la variable age

Nous remarquons d'après la figure 20 que l'ordonnée à l'origine et la pente varient d'un individu à l'autre, ce qui confirme la possibilité d'une de corrélation à l'intérieur de résidu d'une même fille. Alors nous avons procédé à ajuster un modèle linéaire mixte.

2.3 Modèle linéaire mixte

Nous avons commencé par un modèle qui contient une ordonnée à l'origine alétoire et une pente alétoire devant la variable age.
Vue que chaque fille appartient au même groupe de mère et que la variable groupe ne varie pas dans la même grappe i, nous n'avons pas considéré un effet aléatoire devant la variable group.

Choix des formes des matrices D et V
Nous avons choisi pour la matrice V la forme VC et pour la matrice D nous avons séléctinné entre les deux options UN et UN(1) en comparant le AIC trouvé entre les deux modèles (le meilleur AIC est le plus faible).
Option 1 : VC pour V, UN pour D

```
> GirlsGrowth.VCUN0 <- lmer(height~age+factor(group)+(age|child),data=GirlsGrowth,REML=TRUE)
> extractAIC(GirlsGrowth.VCUN0)
[1]   8.0000 346.3052
```

FIGURE 21 – Option 1 : VC pour V, UN pour D

Option 2 : VC pour V, UN(1) pour D

```
> GirlsGrowth.VCUN1 <- lmer(height~age+factor(group)+(age||child),data=GirlsGrowth,REML=TRUE)
> extractAIC(GirlsGrowth.VCUN1)
[1]   7.0000 349.0742
```

FIGURE 22 – Option 2 : VC pour V, UN(1) pour D

Nous remarquons d'aprés les figure 21 et 22 que l'AIC du modèle avec la forme UN pour la matrice D est inférieur à l'AIC du modèle avec la forme UN(1) pour la matrice D ($346.3052 < 349.0742$). Alors, nous avons choisi la matrice VC pour V et la matrice UN pour D.

Sélection des effets aléatoires
Ensuite, nous avons effectué un test d'hypothèse sur la possibilité de réduire le modèle avec seulement une ordonnée à l'origine aléatoire. Nous avons utilisé le test du rapport des vraisemblances pour H0 : l'ordonnée à l'origine aléatoire VS H1 : effet aléatoire devant âge et l'ordonnée à l'origine aléatoire.

```
> GirlsGrowth.VCUN2 <- lmer(height~age+factor(group)+(1|child),data=GirlsGrowth,REML=TRUE)
> xi1 <- 2*(logLik(GirlsGrowth.VCUN0)-logLik(GirlsGrowth.VCUN2))
> pval1 <- 0.5*(1-pchisq(xi1,1))
> pval1
'log Lik.' 2.663653e-10 (df=8)
```

FIGURE 23 – Test du rapport des vraisemblances

Nous avons trouvé que la pvalue égale à 2.663653e-10 qui est inférieur à 0.05 donc nous avons rejeté H0 et nous avons accepté H1. Par la suite, nous avons conservé le modèle avec l'ordonnée à l'origine aléatoire et l'effet aléatoire devant la variable âge.

Réduction de la partie fixe
Nous avons procédé à sélectionner les variables exogènes significatives qui contribuent à l'explication du modèle en utilisant la méthode BACKWARD :

```
Anova(GirlsGrowth.VCUN0,type=3)
```

	Chisq	Df	Pr(>Chisq)
(Intercept)	4148.25792	1	0.0000000000
age	1938.85167	1	0.0000000000
factor(group)	18.38379	2	0.0001018618

FIGURE 24 – ANOVA du modèle complet

Nous remarquons que les deux variables âge et group sont significatifs au seuil 5%, donc nous avons conservé le modèle avec les deux variables explicatives.

2.4 Conclusion

```
> summary(GirlsGrowth.VCUNO)
Linear mixed model fit by REML ['lmerMod']
Formula: height ~ age + factor(group) + (age | child)
   Data: GirlsGrowth

REML criterion at convergence: 327.2

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.75984 -0.53164 -0.04614  0.50942  2.74595

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 child    (Intercept) 10.8605  3.2955
          age          0.2895  0.5381   -0.65
 Residual              0.4758  0.6898
Number of obs: 100, groups:  child, 20

Fixed effects:
               Estimate Std. Error t value
(Intercept)     79.2624     1.2306  64.407
age              5.7165     0.1298  44.032
factor(group)2   3.0303     1.4700   2.061
factor(group)3   6.2885     1.4700   4.278

Correlation of Fixed Effects:
            (Intr) age    fct()2
age         -0.481
factr(grp)2 -0.643  0.000
factr(grp)3 -0.643  0.000  0.538
```

FIGURE 25 – Résumé du modèle

Nous remarquons d'aprés la figure 23 et 24 que la p value de la variable groupe est égale à 0.000101 alors la taille des mères contribue dans l'explication de la grandeur des filles. De plus, les estimates de la variable group sont positifs. Donc nous pouvons conclure que la croissance des filles est positivement liée à la taille de leur mère.

3 Exercice 3

Le but de cet exercice est d'ajuster un modèle de régression de Poisson décrivant le mieux possible la moyenne du nombre de doses auto-administrées, par des patients d'analgésie auto-administrée, en fonction de la dose et de la période à l'aide de la méthode des équations d'estimation généralisées
a- La variable réponse Y est une variable de dénombrement, donc le modèle que nous allons adopter est un modèle de poisson avec un lien log.

$$\log(Y_{ij}) = \beta_0 + \beta_1 \, x_i + \beta_2 \, t_{ij} + \beta_3 \, x_i t_{ij} \tag{3}$$

Y_{ij} est le nombre de doses auto-administrées par le patient i dans la période j.
x_i est égale à 0 si le patient i est dans le groupe 2 mg et est égale à 1 si le patient i est dans le groupe 1 mg. La modalité de réference est 2mg.
t_{ij} est la période j du patient i.
Puisque, ce sont les mêmes personnes dans des périodes différentes, alors il existe une corrélation entre Y_{ij} et Y_{ik} pour $i \neq k$.
Alors, pour estimer les paramètres du modèle, nous allons utiliser la méthode d'équations d'éstimation généralisées (GEE).
Le modèle est une série temporelle, alors la structure que nous allons utiliser pour la corrélation est la structure auto-regressive AR(1).

Coefficients:

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	2.29202737	0.10280798	22.2942560	0.11286832	20.3070928
groupe1mg	-0.12972789	0.16098791	-0.8058238	0.16315700	-0.7951108
time	-0.04868423	0.01467471	-3.3175601	0.01415532	-3.4392888
groupe1mg:time	-0.03215506	0.02389187	-1.3458582	0.02123399	-1.5143207

Figure 26 – GEE avec toutes les variables

Nous allons effectuer la sélection des variables explicatives par la méthode d'exclusion.
L'effet de l'interaction entre le temps et la dose n'est pas significatif au seuil 5% car la valeur absolue de la statistique mesurée $|z_3| = 1.5143207$ est inférieure à $z_{\alpha/2} = 1.96$.

Coefficients:

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	2.36120547	0.08728500	27.051676	0.10055061	23.482756
groupe1mg	-0.30742284	0.09460575	-3.249515	0.12603931	-2.439103
time	-0.06091784	0.01158702	-5.257421	0.01091007	-5.583636

Figure 27 – GEE en éliminant l'effet de l'interaction

Toutes les variables explicatives sont significatives au seuil 5%.
Le modèle finale est alors :

$$\log(Y_{ij}) = \beta_0 + \beta_1 x_i + \beta_2 t_{ij} \tag{4}$$

b-

```
> exp(t(L)%*%gee_ar1$coefficients)
          [,1]
[1,] 0.7353396
```

FIGURE 28 – Estimation ponctuelle

L'estimation ponctuelle de l'effet moyen de la dose est 0.7353396.

```
> exp(t(L)%*%gee_ar1$coefficients+c(-1,1)*1.96*sqrt(t(L)%*%gee_ar1$robust.variance%*%L))
[1] 0.5743824 0.9414013
```

FIGURE 29 – Intervalle de confiance

L'intervalle de confiance de l'effet moyen de la dose est [0.5743824 , 0.9414013].