Modèles Linéaires

Correction Séance 8
Licence 3 MIASHS (2022-2023)

Guillaume Metzler, Francesco Amato
Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France
guillaume.metzler@univ-lyon2.fr ; francesco.amato@univ-lyon2.fr

Résumé

Cette quatrième se focalise sur un modèle généralisé des modèles linéaires que l'on appelle **modèle linéaire à effets mixtes**.

Les détails théoriques exploités ici sont présentés dans le polycopie de la fiche de TD, directement disponible sur le site de Stéphane Chrétien. Ce TD est très largement inspiré de le tutoriel d'Alexander Demos.

1 Vers l'intérêt du modèle linéaire à effets mixtes

Pour justifier notre intérêt vers le modèle à effets mixtes on peut commencer par un exemple.

Nous voulons étudier comment les élèves réagissent à un nouveau type de méthode d'apprentissage active (informatisée) en classe de mathématiques. Pour cela, on mesure les résultats des élèves en mathématiques et la proportion de temps qu'ils passent à utiliser l'ordinateur : on s'attende que plus ils passent de temps à faire la méthode d'apprentissage actif, plus leurs résultats aux tests de mathématiques seront élevés.

Imaginez de conduire cet enquête dans une école en sélectionnât 4 classes différentes et 20 élevés pour chaque classe.

Maintenant, comme nous n'avons pas d'étude comme celle-ci, nous allons simuler ces données. Après avoir appris les modèles d'effets mixtes, vous pourrez suivre la procédure de simulation, mais pour l'instant, vous pouvez simplement passer au prochain morceau de code.

```r
#Set seed so your answers are all the same
set.seed(9)
# Sample Per class room people
n1 <- 20; n2 <- 20; n3 <- 20; n4 <- 20
N<-n1+n2+n3+n4 # Total N
# Uniform distrobution of proportion of time per classroom
X1 <- runif(n1,  0, .35)
X2 <- runif(n2, .3, .55)
X3 <- runif(n3, .5, .75)
X4 <- runif(n4, .7,1.0)
# noise per classroom
e1 <- rnorm(n1, 0, sd=2.5)
e2 <- rnorm(n2, 0, sd=2.5)
e3 <- rnorm(n3, 0, sd=2.5)
e4 <- rnorm(n4, 0, sd=2.5)
# Intercepts per classroom
B0.1 <- 80
B0.2 <- 70
B0.3 <- 60
B0.4 <- 50
# Same slope per classroom
B1=10
# Our equation to  create Y for each classroom
Y1 = B1*scale(X1,scale=F)  + B0.1 + e1
Y2 = B1*scale(X2,scale=F)  + B0.2 + e2
Y3 = B1*scale(X3,scale=F)  + B0.3 + e3
Y4 = B1*scale(X4,scale=F)  + B0.4 + e4
# Merge classrooms into 1 data.frame
Math.Data<-data.frame(Math=c(Y1,Y2,Y3,Y4),ActiveTime=c(X1,X2,X3,X4),
                      Classroom=c(rep("C1",n1),
                      rep("C2",n2),
                      rep("C3",n3),
                      rep("C4",n4)),
                      StudentID=as.factor(1:N))
```

1.1 Données

Donc, imaginez avoir recueilli ces données et les avoir stockées dans une tableau :

```r
# View data
dim(Math.Data)
```

```
## [1] 80  4
```

```
## 80 students
```

```
head(Math.Data)
```

```
##       Math  ActiveTime Classroom StudentID
## 1 77.31942 0.077560489        C1         1
## 2 73.60132 0.008481868        C1         2
## 3 77.44306 0.072491657        C1         3
## 4 80.27724 0.075506742        C1         4
## 5 84.44514 0.155303256        C1         5
## 6 78.52651 0.046926653        C1         6
```

Nous avons :

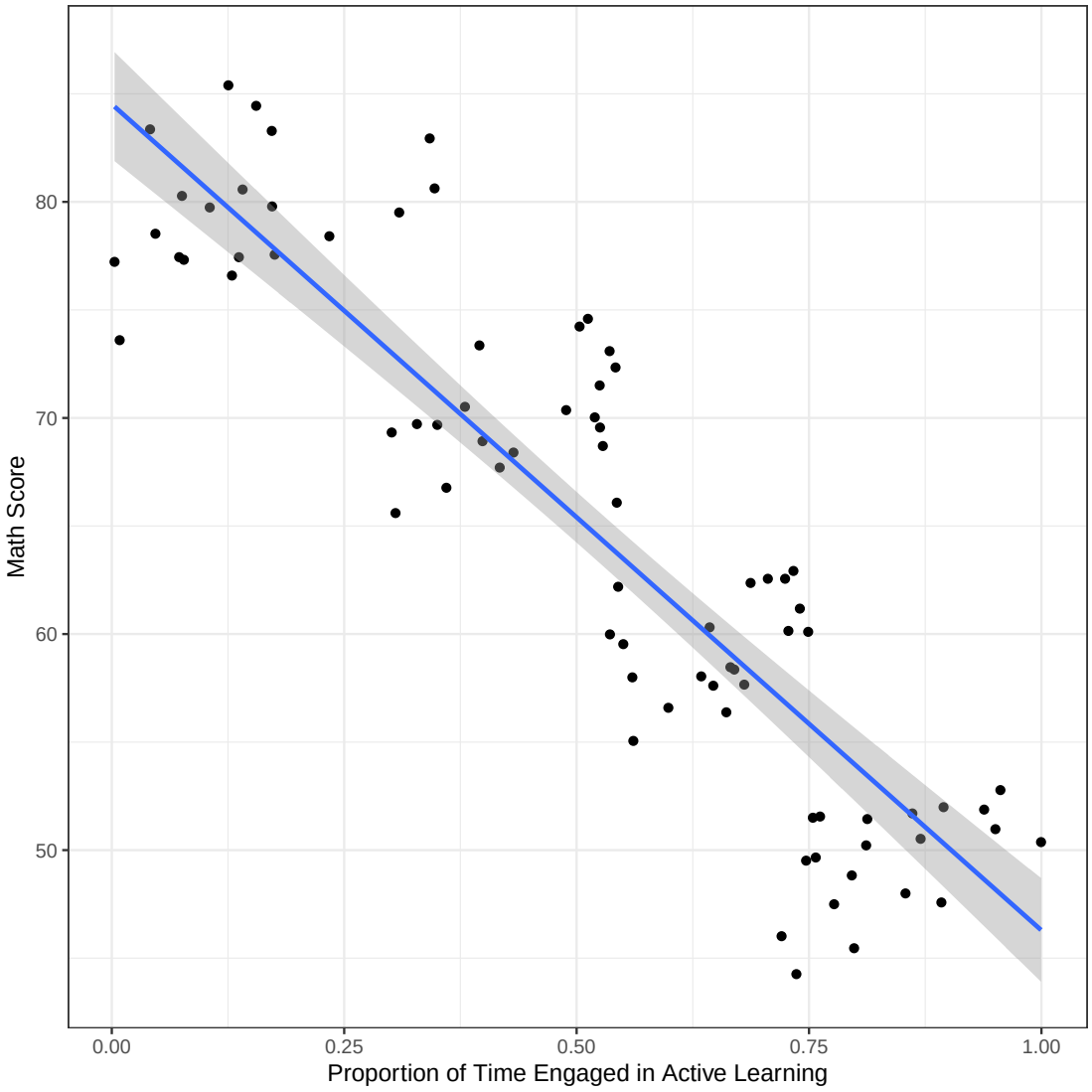
- **Math** : Mesures des résultats des élèves en mathématiques,
- **ActiveTime** : Proportion de temps qu'ils passent à faire la méthode d'apprentissage actif,
- **Classroom** : Classe d'appartenance des élèves, codée avec C1,...,C4,
- **StudentID** : Numéro d'identification de chaque étudiant.

2 Régression simple

Avec nos données, nous somme confiantes de pouvoir conduire notre enquête et nous effectuons une régression linéaire simple :

```
## Plot
library(ggplot2)
# Note: *ggplot2 lets us plot the results with a regression line automatically added
theme_set(theme_bw())
ClassRoom.Plot.1 <-ggplot(data = Math.Data,
                          aes(x = ActiveTime, y=Math))+ #scaffold
  geom_point()+ # add layer of scatterplot
  geom_smooth(method = "lm", se = TRUE)+ # add regression line
  xlab("Proportion of Time Engaged in Active Learning")+
  ylab("Math Score") # add labels
ClassRoom.Plot.1 #call plot
```

```
## `geom_smooth()` using formula 'y ~ x'
```



On régard les résultats de la régression :

```
## Run regression
Class.All<-lm(Math~ActiveTime, data = Math.Data)
summary(Class.All)

##
## Call:
## lm(formula = Math ~ ActiveTime, data = Math.Data)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -12.1043  -3.8995  -0.5892   4.2604  11.4905
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   84.522      1.274   66.35   <2e-16 ***
## ActiveTime   -38.234      2.180  -17.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.263 on 78 degrees of freedom
## Multiple R-squared:  0.7978,Adjusted R-squared:  0.7952
## F-statistic: 307.7 on 1 and 78 DF,  p-value: < 2.2e-16
```

Les résultats montrent qu'une intersection élevée $= 84,52$ et une forte pente négative $= -38,23$ et un $R^2 = 0,8$ très élevé $=$ 0,8.
Nous nous attendions à une pente positive, mais le graphique montre une pente clairement négative !

Avec votre R^2 massif et contre-intuitivement forte pente négative, nous nous précipitons pour publier dans *Psych Science* au titre « L'apprentissage actif est un échec ! Ne perdez pas votre temps en classe ! ».
$\Longrightarrow$ Le gouvernement français, désireux d'améliorer les résultats des mathématiques françaises, cesse de promouvoir l'apprentissage actif et chaque professeur de mathématiques en France change son enseignement, et la prochaine génération d'étudiants en mathématiques obtient de **moins** bons résultats que la dernière génération d'étudiants. On a ruiné l'avenir de la France. **Bon travail !**

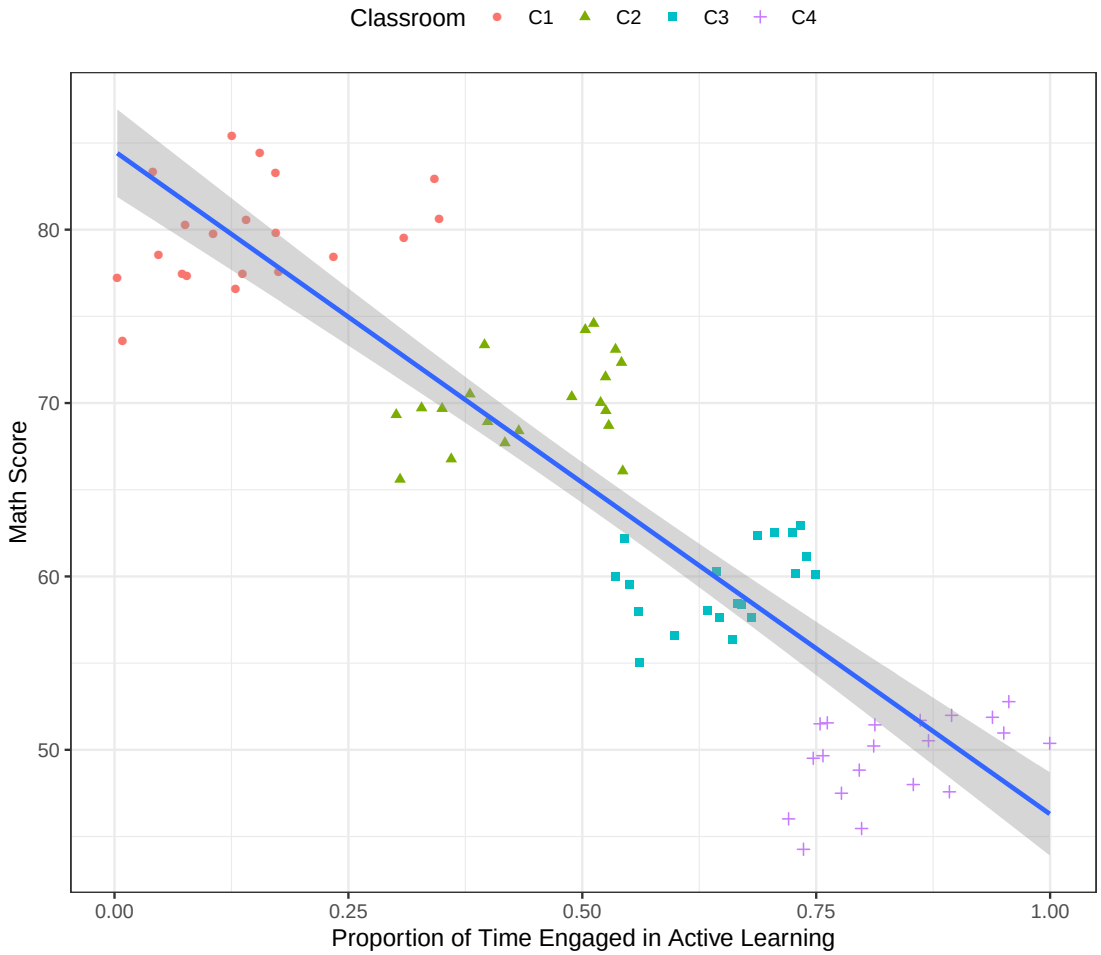
Mais, attendez une seconde, c'est vraiment comme ça ?

3 L'importance d'être groupé.

Clairement, chaque étudiant ajoute du bruit dans notre analyse. Mais les élèves ajoutent du bruit indépendant ? **Non !** Les élèves ne sont pas indépendants les uns des autres dans la classe : clowns de classe, morale de classe, température de la salle, heure de la classe, les enseignants, sont toutes choses qui peuvent varier d'une classe à l'autre. Chaque classe est donc un cluster où les scores des enfants sont interdépendants en fonction de leur expérience partagée.

3.1 Réexamen par groupe

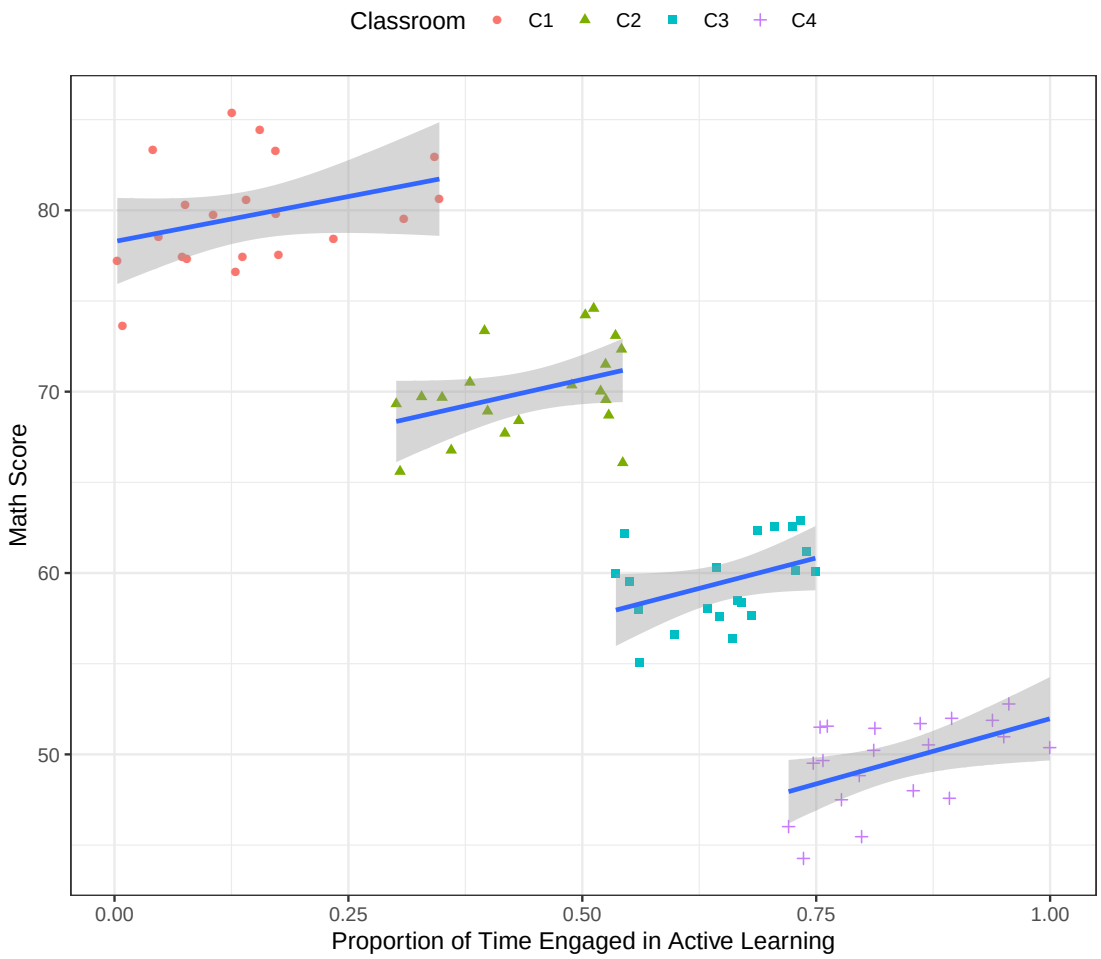
```
ClassRoom.Plot.2 <-ggplot(data = Math.Data,
                          aes(x = ActiveTime, y=Math))+
  geom_point(aes(colour = Classroom, shape=Classroom))+ # we add color by cluster
  geom_smooth(method = "lm", se = TRUE)+
  xlab("Proportion of Time Engaged in Active Learning")+
  ylab("Math Score")+ # add labels
  theme(legend.position = "top")+
  theme(aspect.ratio = 0.8)
ClassRoom.Plot.2

## `geom_smooth()` using formula 'y ~ x'
```



Nous pouvons refaire la régression dans chaque cluster pour voir les pentes correctes.

```
ClassRoom.Plot <-ggplot(data = Math.Data,
                        aes(x = ActiveTime, y=Math))+
  geom_point(aes(colour = Classroom, shape=Classroom))+
  geom_smooth(method = "lm", se = TRUE, aes(group = Classroom))+# we add group level
  xlab("Proportion of Time Engaged in Active Learning")+ylab("Math Score")+ # add labels
  theme(legend.position = "top")+
  theme(aspect.ratio = 0.8)
ClassRoom.Plot
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Les pentes semblent positives maintenant ! Nous avons besoin d'une méthode de régression qui capture les différences dans les clusters et qui estime correctement nos pentes.

4 Modèle linéaire à effets mixtes

4.1 Généralités

Dans un modèle de régression classique, il s'agit d'étudier la liaison statistique entre une variable à expliquer Y et des variables explicatives X non aléatoire. Soit y_i la réponse de l'individu i et x_i les valeurs prises par les variables explicatives pour cet individu. La relation entre X et Y peut s'écrire sous la forme :

$$y_i = \beta_0 + x_i\beta + \varepsilon_i$$

pour le modèle avec une seule covariable et

$$y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i$$

avec

- ε_i une variable aléatoire distribuée selon une loi normale d'espérance nulle et représentant les résidus du modèle ou erreurs,
- β_0 correspond à ce qu'on appelle l'intercept et
- β représente les coefficients du modèle.

Dans un modèle classique, les erreurs sont supposées être indépendantes et identiquement distribuées (le plus souvent selon une loi normale).

Les modèles mixtes ou modèles à effets aléatoires permettent de généraliser l'approche déjà connue pour l'estimation des paramètres d'un modèle de régression linéaire classique (pour observations indépendantes).

4.2 Modèle à effets aléatoires

Dans la pratique, il peut avoir du sens de supposer que l'effet un modèle à un seul facteur à effets aléatoires (et sans effet fixe autre que l'effet général) s'écrira sous la forme suivante

$$Y_{ij} = \mu + A_j + U_{ij}.$$

Dans l'écriture ci dessus :

- Y_{ij} est la réponse selon les niveaux A_1,...,A_J du facteur A.
- μ est un effet fixe, non aléatoire, à estimer (c'est l'effet général, unique effet fixe entrant dans ce modèle)
- A_j, $j = 1, \ldots, J$ est une variable aléatoire de loi $\mathcal{N}\left(0, \sigma_a^2\right)$

— les différentes v.a.r. A_j sont supposées indépendan-tes et de même loi

— elles sont donc i.i.d., gaussiennes, centrées, de même variance inconnue

— ainsi, les J niveaux du facteur à effets aléatoires considéré sont J observations indépendantes de cette loi $\mathcal{N}\left(0, \sigma_a^2\right)$

— $U_{ij} \sim \mathcal{N}\left(0, \sigma^2\right)$, les U_{ij} étant également i.i.d.

— pour un niveau j fixé du facteur aléatoire, on réalise n_j observations indicées par i, de sorte que
  — l'indice i varie de 1 à n_j et
  — que le nombre total d'observations est $n = \sum_{j=1}^{J} n_j$

— on suppose de plus que chaque variable aléatoire A_j est indépendante de chaque variable aléatoire $U_{ij'}, \forall\ (i, j, j')$.

On déduit ainsi de ce modèle

$$\mathbb{E}\left(Y_{ij}\right) = \mu, \quad \text{Var}\left(Y_{ij}\right) = \sigma_a^2 + \sigma^2$$

les deux termes de variance σ_a^2 et σ^2 étant inconnus, ils devront être estimés, de même que μ.

Remarque : pour un même indice j et deux indices i et i' différents, on a :

$$\text{Cov}\left(Y_{ij}, Y_{i'j}\right) = \text{Cov}\left(A_j + U_{ij}, A_j + U_{i'j}\right) = \sigma_a^2$$

il n'y a donc pas indépendance entre les deux v.a.r. Y_{ij} et $Y_{i'j}$: ceci est un élément nouveau et très important dans les modèles à effets aléatoires, comme dans les modèles mixtes.

4.3 Ecriture matricielle

Comme indiqué plus haut, on suppose que l'on fait n_j observations (non indépendantes) de la v.a.r. réponse Y au niveau j du facteur aléatoire ; on supposera $n_j \geq 1$, de sorte que l'on ne considèrera ici que des plans complets ; le nombre total d'observations réalisées est noté n ($n = \sum_{j=1}^{J} n_j$). Sous forme matricielle, le modèle s'écrit :

$$Y = \mu 1_n + \mathbf{Z}A + U$$

avec

— Y et U sont des vecteurs aléatoires de $\mathbb{R}^n$ (que l'on supposera muni de la base canonique) dont les composantes sont, respectivement, les v.a.r. Y_{ij} et U_{ij}
— le vecteur 1_n est le vecteur de $\mathbb{R}^n$ donné par

$$1_n = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.$$

— la matrice $\mathbf{Z}$, de dimension $n \times J$, comporte dans ses colonnes les indicatrices des niveaux du facteur considéré : elle ne contient donc que des 0 et des 1 ;
— enfin, le vecteur

$$A = \begin{pmatrix} A_1 \\ \vdots \\ A_J \end{pmatrix}$$

est un vecteur aléatoire gaussien de $\mathbb{R}^J$: $A \sim \mathcal{N}_J\left(0, \sigma_a^2 \mathbf{I}_J\right)$, où $\mathbf{I}_J$ est la matrice identité d'ordre J.

Considérons le cas très simple dans lequel $J = 2, n_1 = 2\ n_2 = 3$ (on a donc $n = 5$). On a par exemple :

$$\mathbf{Z} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \qquad A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$$

et la matrice de covariance de Y est donnée par

$$\mathbf{V}_Y = \left[\begin{array}{cc|ccc} \sigma_a^2 + \sigma^2 & \sigma_a^2 & 0 & 0 & 0 \\ \sigma_a^2 & \sigma_a^2 + \sigma^2 & 0 & 0 & 0 \\ \hline 0 & 0 & \sigma_a^2 + \sigma^2 & \sigma_a^2 & \sigma_a^2 \\ 0 & 0 & \sigma_a^2 & \sigma_a^2 + \sigma^2 & \sigma_a^2 \\ 0 & 0 & \sigma_a^2 & \sigma_a^2 & \sigma_a^2 + \sigma^2 \end{array}\right]$$

4.4 Modèles à effets mixtes

On appelle modèle à effets mixtes un modèle statistique dans lequel on considère à la fois des facteurs à effets fixes (qui vont intervenir au niveau de l'espérance du modèle) et des facteurs à effets aléatoires (qui vont intervenir au niveau de la variance du modèle).

Un modèle linéaire gaussien mixte, relatif à n observations, est donné par

$$Y_{i,k,l} = \mu + \sum_{j=1^{p_{i,k,l}}} x_{i,k,l,j}\beta_j + A_{k,l} + U_{i,k,l}$$

où, $j = 1, \ldots, J$, $k = 1, \ldots, K$, $\ell = 1, \ldots, L_k$ et $i = 1, \ldots, I_{j,k,l}$. Il s'écrit sous la forme matricielle suivante

$$Y = \mathbf{X}\beta + \sum_{k=1}^{K} \mathbf{Z}_k A_k + U.$$

avec

- X est la matrice $n \times p$ relative aux effets fixes du modèle (figurant en colonnes) ; p est donc le nombre total d'effets fixes pris en compte dans le modèle ; ainsi, dans le cas présent, la matrice $\mathbf{X}$ est analogue à la matrice d'incidence dans une ANOVA.

- β est le vecteur des p effets fixes β_j, $j = 1, \ldots, p$.

- $\mathbf{Z}_k$ est la matrice des indicatrices (disposées en colonnes) des niveaux du k -ième facteur à effets aléatoires $k = 1, \ldots, K$. On notera L_k le nombre de niveaux de ce facteur $\mathbf{Z}_k$ est donc de dimension $n \times L_k$

- $A_{k\ell}$ sont des variables aléatoires associées au ℓ-ième niveau du k-ième facteur à effets aléatoires $\ell = 1, \ldots, L_k$.

  - Pour tout ℓ et tout k, on a

    $$A_{k\ell} \sim \mathcal{N}\left(0, \sigma_k^2\right)$$

    et,

  - pour un indice k donné (autrement dit, pour un facteur déterminé), les v.a.r. $A_{k\ell}$ sont supposées i.i.d.

  - Ainsi, la réponse Y faites au même niveau ℓ du k -ième facteur sont corrélées, leur covariance comportant le terme σ_k^2 et, éventuellement, d'autres composantes de la variance ; par ailleurs, pour deux indices k et k' distincts, $A_{k\ell}$ et $A_{k'\ell'}$ sont indépendantes, pour tous les niveaux ℓ et ℓ'.

  - On a posé

    $$A_k = \begin{bmatrix} A_{k1} \\ \vdots \\ A_{kL_k} \end{bmatrix}$$

de sorte que l'on a $A_k \sim \mathcal{N}_{q_k}\left(0, \sigma_k^2 \mathbf{I}_{L_k}\right)$, les vecteurs aléatoires A_k étant mutuellement indépendants.

— U est le vecteur aléatoire des erreurs du modèle :

$$U = \begin{bmatrix} U_{1,1,1} \\ \vdots \\ U_{I_{J,K,L_K},J,K,L_K} \end{bmatrix}.$$

Il vérifie $U \sim \mathcal{N}_n\left(0, \sigma^2 \mathbf{I}_n\right)$ de plus, il est supposé indépendant des A_k.

5 Rétour à notre example

D'abord, on peut chercher d'utiliser un modèle comme lequel dans la section 4.2. Ce modèle ne considère pas la variable du temps d'exposition à l'apprentissage actif, mais il est un modèle à un seul facteur à effets aléatoires et sans effet fixe autre que l'effet général.

```
library(lme4)

## Loading required package: Matrix

Model.Null<-lmer(Math ~1+(1|Classroom),
                 data=Math.Data, REML=FALSE)
summary(Model.Null)

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: Math ~ 1 + (1 | Classroom)
##    Data: Math.Data
##
##      AIC      BIC   logLik deviance df.resid
##    408.6    415.7   -201.3    402.6       77
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.34755 -0.72263  0.01011  0.73079  2.21975
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  Classroom (Intercept) 126.884  11.264
##  Residual                6.668   2.582
## Number of obs: 80, groups:  Classroom, 4
##
```

```
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)    64.70       5.64   11.47
```

```
# nous examinons les interceptions pour chaque classe (effet aléatoire)
ranef(Model.Null)
```

```
## $Classroom
##    (Intercept)
## C1   14.959676
## C2    5.307215
## C3   -5.190059
## C4  -15.076832
##
## with conditional variances for "Classroom"
```

Par contre, en ajoutant la variable de du temps d'exposition à l'apprentissage actif, et donc un autre effet fixe, en plus de l'intercepte :

```
Model.1<-lmer(Math ~ActiveTime+(1|Classroom),
              data=Math.Data, REML=FALSE)
summary(Model.1)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: Math ~ ActiveTime + (1 | Classroom)
##    Data: Math.Data
##
##      AIC      BIC   logLik deviance df.resid
##    399.3    408.8   -195.6    391.3       76
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.13111 -0.77650 -0.04481  0.61218  2.50097
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  Classroom (Intercept) 200.00   14.142
##  Residual                5.61    2.368
## Number of obs: 80, groups:  Classroom, 4
##
## Fixed effects:
##             Estimate Std. Error t value
```

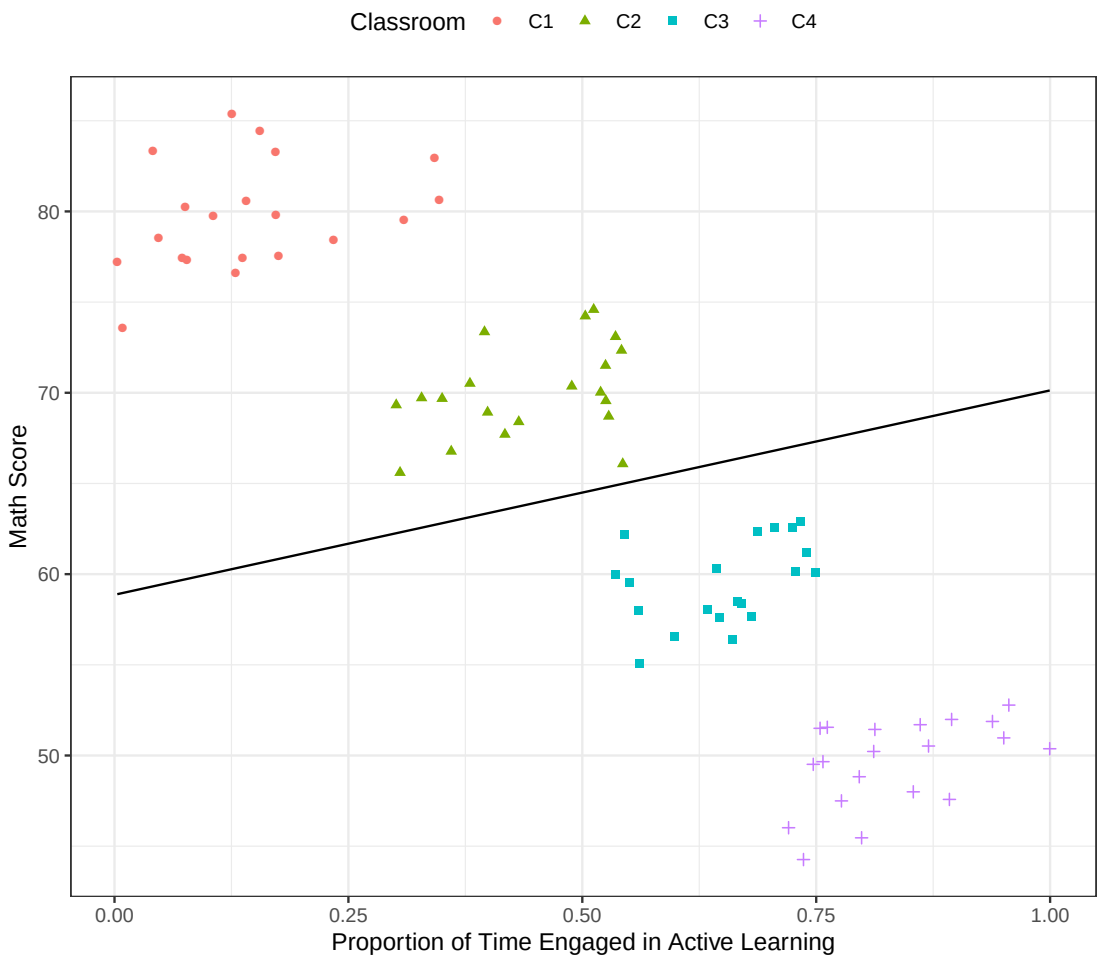
```
## (Intercept)   58.862      7.261   8.107
## ActiveTime    11.269      3.140   3.589
##
## Correlation of Fixed Effects:
##             (Intr)
## ActiveTime -0.224

# Pour voir quel modèle est préferible, on peut
# régarder le BIC du modèle mixte et lequel de
# la régression simple

BIC(Class.All) # Plus haut

## [1] 503.8633

# nous examinons les interceptions pour chaque classe (effet aléatoire)
ranef(Model.1)

## $Classroom
##    (Intercept)
## C1   19.196192
## C2    6.144442
## C3   -6.688270
## C4  -18.652363
##
## with conditional variances for "Classroom"
```

Notez que maintenant une covariance entre les deux effets fixes a été estimé également. Justement, en incluant un effet fixe, qui intervient au niveau de l'espérance du modèle, l'estimation pour les interceptes changent, ainsi que les variances, car l'effet fixe "explique" une partie de la variance. Même les interceptes au niveau de groupe ont changé.

Enfin, On trace l'effet fixe sur les données brutes :

```
#Fixed only, removed random effects
Math.Data$Model.1.Fitted<-predict(Model.1, re.form=NA)

ClassRoom.Plot <-ggplot(data = Math.Data)+
  geom_point( aes(x = ActiveTime, y=Math,
                                    colour = Classroom, shape=Classroom))+
  geom_line(aes(x = ActiveTime, y=Model.1.Fitted))+
```

```
  xlab("Proportion of Time Engaged in Active Learning")+
  ylab("Math Score")+
  theme(legend.position = "top")+
  theme(aspect.ratio = 0.8)
ClassRoom.Plot
```



Une ligne avec une pente comme ça semble contre-intuitif avec nos données, mais c'est la droite qu'on obtienne en utilisant les estimations pour nos effets fixes, une fois que on maîtrise la variance pour les groupes.