Devoir 2

Partie théorique

T-1

#1

On veut montrer que $\bar{e} = \sum_{i=1}^{n} \frac{e_i}{n} = 0$. On peut donc développer la formule de $\bar{e}$:

$$\begin{aligned}\bar{e} &= \sum_{i=1}^{n} \frac{e_i}{n} = \sum_{i=1}^{n} \frac{Y_i - \hat{Y}_i}{n} \\ &= \bar{Y} - \frac{\sum_{i=1}^{n} x_i^{'}\hat{\beta}}{n}\end{aligned}$$

Nous pouvons écrire la somme de droite de cette façon:

$$\begin{aligned}\sum_{i=1}^{n} x_i^{'}\hat{\beta} &= \hat{\beta}_0 \sum_{i=1}^{n} 1 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i,1} + ... + \hat{\beta}_{p'} \sum_{i=1}^{n} x_{i,p'} \\ \frac{\sum_{i=1}^{n} x_i^{'}\hat{\beta}}{n} &= \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + ... + \hat{\beta}_{p'} \bar{x}_{p'}\end{aligned}$$

Tel que donnée dans l'énoncé, on peut prendre pour acquis que:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}_1 - ... - \hat{\beta_{p'}} \bar{x}_{p'}$$

Ainsi, en remplaçant les valeurs obtenues avec l'équation plus haut, on obtient:

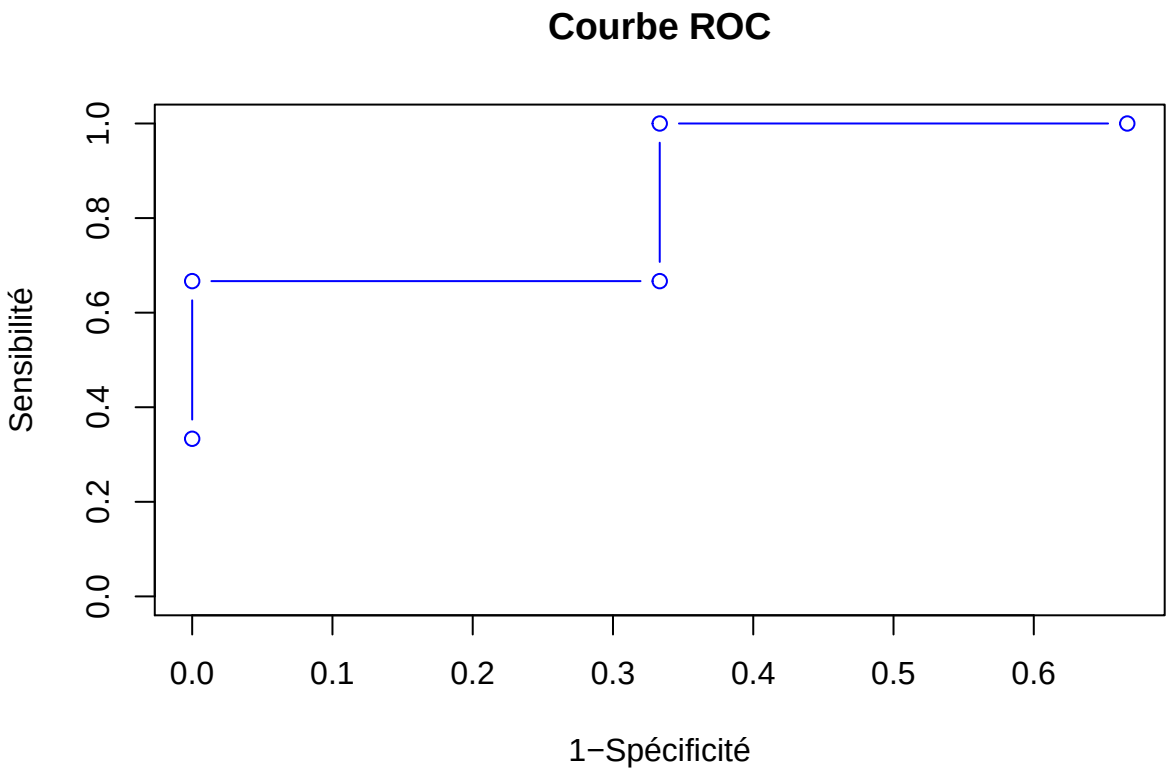
$$\begin{aligned}\bar{e} &= \bar{Y} - \frac{\sum_{i=1}^{n} x_i^{'}\hat{\beta}}{n} \\ &= \bar{Y} - \left(\hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + ... \hat{\beta}_{p'} \bar{x}_{p'}\right) \\ &= \bar{Y} - \left(\left[\bar{Y} - \hat{\beta}_1 \bar{x}_1 - ... - \hat{\beta_{p'}} \bar{x}_{p'}\right] + \hat{\beta}_1 \bar{x}_1 + ... + \hat{\beta}_{p'} \bar{x}_{p'}\right) \\ &= \bar{Y} - \bar{Y} \\ \bar{e} &= 0\end{aligned}$$

#4

Pour faire la courbe ROC, il faut calculer la valeur de $\hat{\pi}_i = P(Y_i = 1, x_i) = \frac{exp(-4.2+1.2x_i)}{1+exp(-4.2+1.2x_i)}$. Par la suite on calcule une prédiction de $\hat{Y}_i$ selon un certain seuil u_k : si $\hat{\pi}_i > u_k$, alors $\hat{Y}_i = 1$, sinon $\hat{Y}_i = 0$. On calule alors la sensibilié et la spécificité pour les différents seuils. Voici un tableau des résulats ($\hat{Y}_i$ est représenté comme étant $\hat{Y}_u$ selon la valeur de u_k) pour les valeurs de u_k =0.1, 0.2, 0.5, 0.8 et 0.9 ainsi que le graphique de la courbe ROC:

Observation	x_i	$\hat{\pi}_i$	Y_i	$\hat{Y}_{u=0.1}$	$\hat{Y}_{u=0.2}$	$\hat{Y}_{u=0.5}$	$\hat{Y}_{u=0.8}$	$\hat{Y}_{u=0.9}$
obs 1	1	0.0474259	0	0	0	0	0	0
obs 2	2	0.1418511	0	1	0	0	0	0
obs 3	3	0.3543437	1	1	1	0	0	0
obs 4	4	0.6456563	0	1	1	1	0	0
obs 5	5	0.8581489	1	1	1	1	1	0
obs 6	6	0.9525741	1	1	1	1	1	1

Métriques	$\hat{Y}_{u=0.1}$	$\hat{Y}_{u=0.2}$	$\hat{Y}_{u=0.5}$	$\hat{Y}_{u=0.8}$	$\hat{Y}_{u=0.9}$
VP	3.0000000	3.0000000	2.0000000	2.0000000	1.0000000
FN	0.0000000	0.0000000	1.0000000	1.0000000	2.0000000
VN	1.0000000	2.0000000	2.0000000	3.0000000	3.0000000
FP	2.0000000	1.0000000	1.0000000	0.0000000	0.0000000
Sensibilité	1.0000000	1.0000000	0.6666667	0.6666667	0.3333333
Spécificité	0.3333333	0.6666667	0.6666667	1.0000000	1.0000000



T-2

#4

Le modèle est défini de la façon suivante:

$$Y_{ij} = \beta_0 + \gamma_{i0} + (\beta_1 + \gamma_{i1})x_{ij} + \epsilon_{ij}$$

On peut définir la matrice de betas et de gammas pour nos tests d'hypothèse qui prendront tous la même forme:

$$\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \gamma_{10} \\ \gamma_{11} \\ \gamma_{20} \\ \gamma_{21} \end{bmatrix}$$

$$H_0 : \boldsymbol{L} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix} = \boldsymbol{d}$$

$$H_1 : \boldsymbol{L} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix} \neq \boldsymbol{d}$$

a)

Pour tester si l'effet moyen du médicament sur la pression sanguine moyenne dans cette population de souris est nul, on doit regarder l'espérance de Y lorsque les effets aléatoires sont nuls. Alors,

$$E[Y; x_0 = x + 1] - E[Y; x_0 = x] = 0$$

Ce qui revient à tester si $\beta_1 = 0$.

$$\boldsymbol{L} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$
$$\boldsymbol{d} = \begin{bmatrix} 0 \end{bmatrix}$$

b)

Pour tester si l'effet du médicament sur la pression sanguine moyenne dans la famille 1 est nul, ceci revient à tester si la valeur devant x_{1j} est nulle pour la famille 1. Ceci revient alors à tester si

$$\beta_1 + \gamma_{11} = 0$$
$$\boldsymbol{L} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix}$$
$$\boldsymbol{d} = \begin{bmatrix} 0 \end{bmatrix}$$

c)

Afin de tester si l'effet du médicament sur la pression sanguine moyenne est nul chez chacune des deux familles de souris, alors on souhaite tester si les valeurs devant x_{ij} pour les 2 familles sont nulles. Alors, on peut tester simulatément ceci:

$$\beta_1 + \gamma_{11} = 0 \text{ et } \beta_1 + \gamma_{21} = 0$$
$$\boldsymbol{L} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$
$$\boldsymbol{d} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

d)

Afin de tester si une hausse d'une unité de la dose de médicament dans la famille 1 est équivalente à une hausse d'une unité de la dose de médicament dans la famille 2 en termes d'effet sur la pression sanguine moyenne, on peut réécrire le tout de la manière suivante:

$$E[Y_1; x_{10}^* = x_{10} + 1] - E[Y_1; x_{10}^* = x_{10}] = E[Y_2; x_{20}^* = x_{20} + 1] - E[Y_2; x_{20}^* = x_{20}]$$

Alors ceci revient à tester si

$$\beta_1 + \gamma_{11} = \beta_1 + \gamma_{21}$$
$$\gamma_{11} - \gamma_{21} = 0$$

$$\boldsymbol{L} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix}$$
$$\boldsymbol{d} = \begin{bmatrix} 0 \end{bmatrix}$$

#5

a)

On peut définir le modèle de cette façon: $Y_{ij} = \beta_0 + \gamma_{i0} + \beta_1 x_{ij} + \epsilon_{ij}$. Voici la notation matricielle:

$$\mathbf{Y}' = \begin{bmatrix} Y_{11} & Y_{12} & Y_{21} & Y_{22} & Y_{31} & Y_{32} \end{bmatrix}$$
$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{12} \\ 1 & x_{21} \\ 1 & x_{22} \\ 1 & x_{31} \\ 1 & x_{32} \end{bmatrix}$$
$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$
$$\boldsymbol{\gamma} = \begin{bmatrix} \gamma_{10} \\ \gamma_{20} \\ \gamma_{30} \end{bmatrix}$$
$$\mathbf{Z} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$
$$\epsilon' = \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{21} & \epsilon_{22} & \epsilon_{31} & \epsilon_{32} \end{bmatrix}$$

On peut également remplacer les valeurs symboliques de $\mathbf{Y}$ et $\mathbf{X}$ par leurs valeurs numériques:

$$\mathbf{Y}' = \begin{bmatrix} 70 & 80 & 50 & 60 & 100 & 70 \end{bmatrix}$$
$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$$

b)

On trouve les matrices de variance:

$$\mathbf{D} = Var(\boldsymbol{\gamma}) = \begin{bmatrix} \sigma_0^2 & 0 & 0 \\ 0 & \sigma_0^2 & 0 \\ 0 & 0 & \sigma_0^2 \end{bmatrix}$$

$$\mathbf{V} = Var(\boldsymbol{\epsilon}) = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = Var(\mathbf{Y}) = \mathbf{ZDZ}^{'} + \mathbf{V}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 + \sigma_0^2 & \sigma_0^2 & 0 & 0 & 0 & 0 \\ \sigma_0^2 & \sigma^2 + \sigma_0^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 + \sigma_0^2 & \sigma_0^2 & 0 & 0 \\ 0 & 0 & \sigma_0^2 & \sigma^2 + \sigma_0^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 + \sigma_0^2 & \sigma_0^2 \\ 0 & 0 & 0 & 0 & \sigma_0^2 & \sigma^2 + \sigma_0^2 \end{bmatrix}$$

c)

Ces deux valeurs peuvent s'estimer avec les formules suivantes:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{'}\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{X})^{-1}\mathbf{X}^{'}{\boldsymbol{\Sigma}}^{-1}\mathbf{Y}$$

$$\hat{\boldsymbol{\gamma}} = \mathbf{DZ}^{'}{\boldsymbol{\Sigma}}^{-1}(\mathbf{Y} - \boldsymbol{X\beta})$$

On peut utiliser R pour calculer les valeurs numériques de ces estimés:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 80 \\ -16.67 \end{bmatrix}$$

$$\hat{\boldsymbol{\gamma}} = \begin{bmatrix} 2.67 \\ -13.33 \\ 10.67 \end{bmatrix}$$

d)

On peut calculer l'estimé à partir de cette équation:

$$\hat{\boldsymbol{Y_i}} = \boldsymbol{V_i}{\boldsymbol{\Sigma_i}}^{-1}\boldsymbol{X_i^{'}}\hat{\boldsymbol{\beta}} + (\boldsymbol{I}_{n_i * n_i} - \boldsymbol{V_i}{\boldsymbol{\Sigma_i}}^{-1})\boldsymbol{Y_i}$$

Où:

$$\boldsymbol{V_i}{\boldsymbol{\Sigma_i}}^{-1} = \begin{bmatrix} 0.6 & -0.4 \\ -0.4 & 0.6 \end{bmatrix}$$

De cette façon, on obtient la valeur moyenne de $\hat{\boldsymbol{Y_i}} = 56.67$ pour notre estimé de la note moyenne obtenue par l'individu 2 dans les cours où il utilise un manuel de langue anglaise.

#6

a)

Dans cette situation, le paramètre d'intérêt est β_3 puisque celui-ci estimera l'effet sur la valeur de Y_{ij} au fil du temps seulement lorsque $x_i = 1$.

b)

Il serait raisonnable de choisir les strutures AR(1) et UN(1) puisque cette paire a la plus petite valeur d'AIC pour la méthode ML.

c)

On procède à un test d'hypothèse formel en testant un modèle complet (ligne 1) et un modèle réduit sans la pente aléatoire (ligne 2):

$$H_0 : Y_{ij} = \beta_0 + \gamma_{i0} + \beta_1 x_i + \beta_2 t_j + \beta_3 x_i t_j + \epsilon_{ij}$$
$$H_1 : Y_{ij} = \beta_0 + \gamma_{i0} + \beta_1 x_i + (\beta_2 + \gamma_{i2}) t_j + \beta_3 x_i t_j + \epsilon_{ij}$$

On pose : $\epsilon = 2(\ell_1 - \ell_0) = 2(-88 + 89.5) = 3$. (On pourrait également utiliser les mesures REML pour des résultats semblables).

Nous rejeterons l'hypothèse si la p-value du test est trop élevée:

$$pvalue = 0.5P[\chi^2_{m_1 - m_0 - 1} > \epsilon] + 0.5P[\chi^2_{m_1 - m_0} > \epsilon]$$

$m_0 = 1$(1 variance) et $m_1 = 2$(2 variances). Par conséquent:

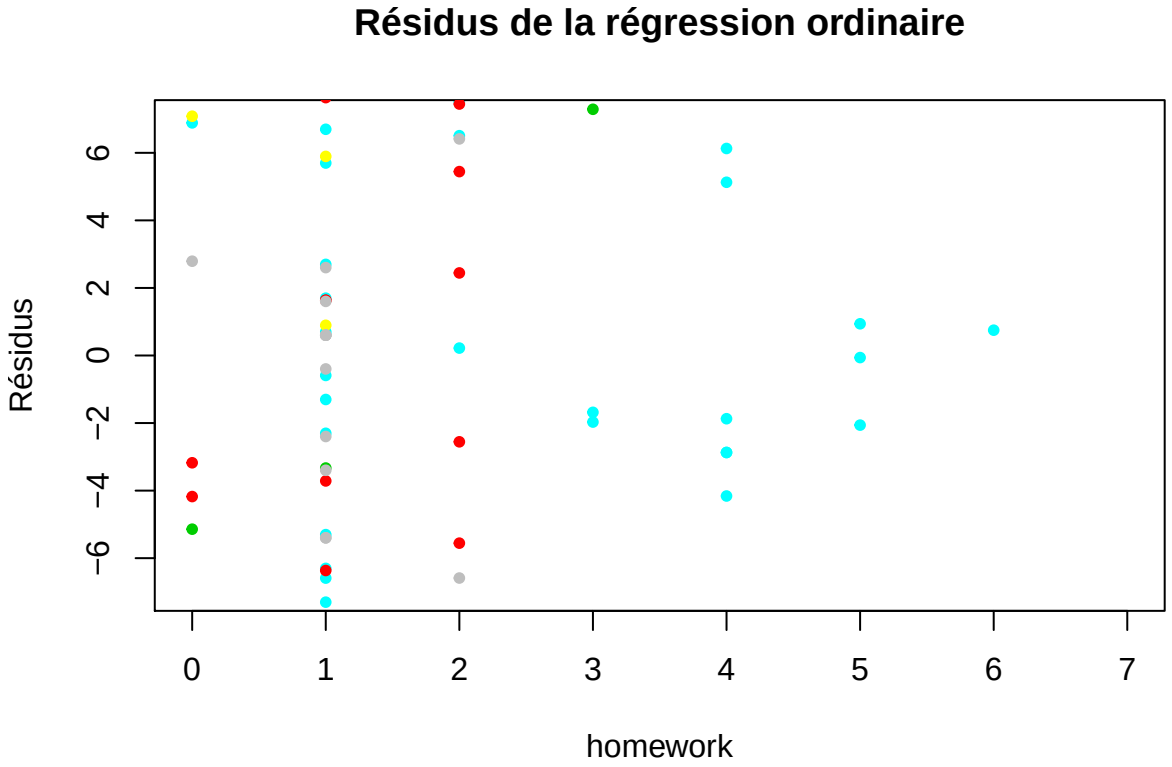
$$p = 0.5P[\chi^2_0 > \epsilon] + 0.5P[\chi^2_1 > \epsilon] = 0 + 0.5 * 0.08326 = 0.04163226$$

La p-value est inférieur à 0.05, par conséquent on rejète l'hypothèse du modèle réduit avec ce seuil de 5% et on conserve l'effet aléatoire γ_{i2}. Il est toutefois à noter que la p-value est assez proche de 0.05 avec une valeur de 0.0416.

Partie pratique

P-2

Tout d'abord, il est important de commencer l'analyse en considérant les effets aléatoires possibles dans le modèle. Puisque les variables `ratio` ainsi que `meanses` ne varient pas en fonction de l'école, ces variables ne sont pas utilisable comme pente aléatoire dans le modèle. Par ailleurs, il serait probable que le nombre d'heures de travail de l'étudiant à la maison par semaine varient selon l'école. Afin de constaté ceci, il est possible de faire un modèle complet de régression standard et visualiser sur un graphique les résidus en fonction de la variable `homework` pour ces différentes écoles. Afin d'avoir une vue allégée du graphique, seulement les 5 premières écoles sont dans le graphique des résidus plus bas.



Bien qu'il soit difficile à analyser, on semble voir une certaine dépendance avec l'école.

(a)

Commençons en créant un modèle linéaire mixte sans considérer la variable `meanses`.

Alors, avec le modèle suivant:

$$Y_{ij} = \beta_0 + \gamma_{i0} + \beta_1 x_{ij1} + \beta_2 x_{i2} + (\beta_3 + \gamma_{i3}) x_{ij3} + \epsilon_{ij}$$

où
Y_{ij} = résultat de l'étudiant j dans l'école i en mathématique.
x_{ij1} = l'indicateur de 1 pour blanc pour l'étudiant j dans l'école i.
x_{i2} = est le ratio du nomde d'étudiants par enseignant dans les classes de l'école i.
x_{ij3} = le nombre d'heures de travail à la maison de l'étudiant j dans l'école i.

on peut tester si on obtient un AIC minimal avec les blocs de la matrice D non-structuré (UN) ou diagonale principale (UN(1)). Par simplicité et limitation du package `lme4`, la matrice V sera de composante de variance (VC).

```
##     VC et UN      VC et UN(1)
## AIC 3643.262         3666.629
```

Alors, on choisit le modèle avec les blocs de la matrice D non-structuré (UN).

On peut poursuivre en effectuant un test de la nécessité des effets aléatoires avec le test suivant:

$$H_0 : Y_{ij} = \beta_0 + \gamma_{i0} + \beta_1 x_{ij1} + \beta_2 x_{i2} + \beta_3 x_{ij3} + \epsilon_{ij}$$

$$H_1 : Y_{ij} = \beta_0 + \gamma_{i0} + \beta_1 x_{ij1} + \beta_2 x_{i2} + (\beta_3 + \gamma_{i3}) x_{ij3} + \epsilon_{ij}$$

Et nous pourrons obtenir la p-value avec cette équation:

$$p = 0.5P[\chi_1^2 > \epsilon] + 0.5P[\chi_2^2 > \epsilon]$$

puisque nous avons que $m_0 = 1$(1 variance) et $m_1 = 3$(2 variances et 1 covariance). On obtient une valeur de $\epsilon = 2(\ell_1 - \ell_0) = 89.74556$ et on obtient une p-value très près de 0. Alors, on rejette le l'hypothèse nulle et nous ne pouvons pas simplifier le modèle.

Il est possible d'effectuer une sélection des effets fixes avec la méthodes backward, où nous rejetons la variable `ratio`.

```
## Loading required package: carData
```

```
## [1] "Résumé des coefficients des effets fixes"
```

```
## (Intercept)       white    homework
##   44.019758    3.300014    1.903103
```

```
## [1] "Résumé des effets aléatoires"
```

```
## $schid
##       (Intercept)   homework
## 6053    4.4055629  0.2448821
## 6327   13.8592245 -7.8453317
## 6467   -5.0779757  4.1820446
## 7194    5.1343844 -4.2449442
## 7472    3.1258114 -4.8279724
## 7474   10.6523351 -4.4787070
## 7801    7.6905919 -4.6275450
## 7829    4.8144131 -4.5816031
## 7930   -7.2576202  5.4227929
## 24371  -6.2371857  2.4271454
## 24725 -11.3814602  3.3947014
## 25456   5.5535448 -4.9940643
## 25642   4.1265536 -3.3944956
## 26537  -0.8034290  1.8455423
## 46417   0.5194043  2.2395233
## 47583  -4.2400423  1.1974074
## 54344  -7.4448939  1.3461689
## 62821  10.7107189 -0.4949450
## 68448  -7.4166165  3.7705810
## 68493  -6.2215976  3.2229537
## 72080  -1.9702841  2.3861839
## 72292  -8.7018273  4.1025471
## 72991  -3.8396126  3.7071342
```

Alors, pour répondre aux questions de l'énoncé (a), l'effet du nombre d'heure est positif, alors plus un étudiant passe d'heure d'étude par semaine, meilleur sont ses résultats en mathématiques. De plus, l'augmentation moyenne sur le résultat en mathématique d'une heure d'étude supplémentaire par semaine sur l'ensemble des étudiant est de 1.903103($\hat{\beta}_3$). Par ailleurs, puisque nous avons rejeté le modèle sans la pente aléatoire devant la variable x_{ij3}, on peut conclure qu'il est raisonnable de croire que l'effet du nombre d'heures varient d'une école à une autre.

(b)

En procédant de la même façon que les étapes en (a) et en y rajoutant la variable `meanses`, on conserve la structure des variances VC et UN puisqu'on obtient un AIC minimal. De plus, en effectuant le test de la nécessité des effets aléatoires, on obtient un résultat identique, c'est-à-dire qu'on conserve la pente aléatoire devant la variable x_{ij3} (nombre d'heures de travail par semaine tel que défini plus haut).

En utilisant la méthode backward pour la sélection des effets fixes, on rejette, en 2 étapes, la variable `ratio` ainsi que l'interaction `meanses*homework`

```
## [1] "Résumé des coefficients des effets fixes"
```

```
## (Intercept)       white    homework     meanses
##   44.702203    3.114922    1.925085    4.892483
```

```
## [1] "Résumé des effets aléatoires"
```

```
## $schid
##       (Intercept)   homework
## 6053    0.7955302  0.1671122
## 6327   10.6841274 -6.9633524
## 6467   -5.5008579  4.2295462
## 7194    6.1971941 -4.2164797
## 7472    4.9610280 -4.7656103
## 7474   10.8284267 -4.4721868
## 7801    8.5316493 -4.6276834
## 7829    7.3630183 -4.3946052
## 7930   -8.3637902  5.3596980
## 24371  -5.0440909  2.4523099
## 24725  -8.6148618  3.3969895
## 25456   5.9455044 -4.7524984
## 25642   8.3953235 -3.8543474
## 26537  -4.6855665  1.4033056
## 46417   0.7681519  2.2498679
## 47583  -4.3887822  1.0435040
## 54344  -8.2255629  1.3886616
## 62821   5.4262853 -0.5889649
## 68448  -7.1480214  3.7467238
## 68493  -4.7023652  3.1041034
## 72080  -1.4669293  2.3654021
## 72292  -7.0099571  4.0740467
## 72991  -4.7454535  3.6544576
```

Alors, pour répondre aux questions du numéro, l'ajout de la variable `meanses` au modèle n'a pas diminué le besoin d'inclure des effets aléatoires. De plus, l'effet moyen sur le résultat en mathématique d'une heure d'étude supplémentaire par semaine sur l'ensemble des étudiants est environ le même qu'obtenu précédemment, soit $\hat{\beta}_3 = 1.925085$. Par ailleurs, puisque nous avons rejeté le modèle sans la pente aléatoire devant la variable x_{ij3}, on peut conclure qu'il est raisonnable de croire que l'effet du nombre d'heures varient d'une école à une

autre.
Alors, le fait d'inclure ou non une variable unique à chaque école dans le modèle n'a pas réduit l'importance d'avoir des effets aléatoires dans le modèle et n'a non plus pas changé les conclusions.