Tests multidimensionnels pour données répétées avec la procédure GLM de SAS

Alain Baccini
Institut de Mathématiques de Toulouse
Mai 2010

L'objet de cette note est de détailler les tests multidimensionnels réalisés par la procédure GLM de SAS dans le traitement de données répétées. Nous allons illustrer tout ce qui suit au moyen d'un exemple (fictif) de données répétées.

Cette note constitue l'annexe E du cours **Le Modèle Linéaire Gaussien Général***, en libre accès sur ce même site.*

1 Les données

Voici ces données :

```
1 10 15 18 24
1 12 14 15 18
1 14 18 20 24
1 13 15 19 21
1 11 13 16 19
2 21 30 42 50
2 24 36 45 56
2 23 27 30 35
2 26 35 38 45
2 29 38 49 57
2 28 38 45 54
3 50 53 57 59
3 51 54 58 60
3 54 58 62 68
3 50 51 54 57
3 53 54 57 63
3 51 54 55 56
3 52 53 56 58
```

En première colonne figure un unique facteur, noté F, à trois niveaux, notés 1, 2 et 3. Le facteur F est supposé à effets fixes. Les tests que nous allons détailler concernant les effets fixes, nous avons, pour simplifier, considéré un modèle à effets fixes avec un unique facteur. Par contre, certains résultats étant un peu particuliers si le facteur ne comporte que deux niveaux, nous avons considéré un facteur à trois niveaux. De plus, nous avons volontairement considéré un plan déséquilibré, afin d'avoir les résultats les plus généraux possible. Ainsi, les données comportent respectivement 5, 6 et 7 observations dans les trois niveaux de F, soit un échantillon de 18 observations (18 lignes dans le fichier ci-dessus).

Il y a ensuite, dans les quatre colonnes suivantes du fichier, une variable réponse Y observée à quatre instants différents (ces variables seront par la suite notées Y_1, Y_2, Y_3 et Y_4).

2 Traitement avec la commande `repeated` de la procédure GLM

Nous faisons ici un traitement de ces données avec la procédure GLM de SAS, au sein de laquelle nous devons mettre la commande `repeated`. Les données ci-dessus sont dans un fichier appelé `repet.don` et contenu dans le répertoire dans lequel nous avons ouvert SAS.

```
data repet;
infile 'repet.don';
input F $ Y1-Y4;
run;
* ------------------------------------- ;
*          modelisation avec GLM        ;
* ------------------------------------- ;
proc glm data=repet;
class F;
model Y1-Y4 = F / ss3;
repeated temps contrast(1) / printh printe;
run;
```

L'option `ss3` de la commande `model` permet de n'avoir en sortie que les sommes de type 3 (les seules qui nous intéressent ici). L'élément `contrast(1)` de la commande `repeated` permet de calculer les évolutions par rapport au temps 1, au lieu de le faire par rapport au temps 4, comme cela est fait par défaut. Enfin, les options `printh` et `printe` permettent d'obtenir les matrices **H** et **E** qui interviennent dans la définition des statistiques des tests multidimensionnels.

Nous donnons ci-dessous une partie des résultats obtenus. Nous ne faisons pas figurer les premiers d'entre eux qui sont les ANOVA unidimensionnelles réalisées, à chaque instant, selon le facteur F. Notons simplement que ces quatre ANOVA sont très significatives (toutes les *p-values* sont inférieures à 10^{-4}) et sont associées à des coefficients R^2 tous supérieurs à 0.9. On peut donc penser qu'il y aura un effet marginal de F (indépendamment du temps) très significatif.

```
                          The GLM Procedure
                  Repeated Measures Analysis of Variance

                  Repeated Measures Level Information

        Dependent Variable          Y1       Y2       Y3       Y4

            Level of temps           1        2        3        4

---------------------------------------------------------------------------

                           E = Error SSCP Matrix

 temps_N represents the contrast between the nth level of temps and the 1st

                           temps_2       temps_3       temps_4

            temps_2         52.262        80.476       113.190
            temps_3         80.476       196.248       255.019
            temps_4        113.190       255.019       367.848

---------------------------------------------------------------------------

                    H = Type III SSCP Matrix for temps

 temps_N represents the contrast between the nth level of temps and the 1st

                         temps_2           temps_3           temps_4

       temps_2      391.24276814      758.20605251      1166.7347575
       temps_3      758.20605251      1469.3598576      2261.0650645
       temps_4      1166.7347575      2261.0650645      3479.3486426
```

MANOVA Test Criteria and Exact F Statistics
for the Hypothesis of no temps Effect
H = Type III SSCP Matrix for temps
E = Error SSCP Matrix

S=1 M=0.5 N=5.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.09087595	43.35	3	13	<.0001
Pillai's Trace	0.90912405	43.35	3	13	<.0001
Hotelling-Lawley Trace	10.00401131	43.35	3	13	<.0001
Roy's Greatest Root	10.00401131	43.35	3	13	<.0001

H = Type III SSCP Matrix for temps*F

temps_N represents the contrast between the nth level of temps and the 1st

	temps_2	temps_3	temps_4
temps_2	157.73809524	271.19047619	388.80952381
temps_3	271.19047619	469.53015873	671.98095238
temps_4	388.80952381	671.98095238	962.15238095

MANOVA Test Criteria and F Approximations
for the Hypothesis of no temps*F Effect
H = Type III SSCP Matrix for temps*F
E = Error SSCP Matrix

S=2 M=0 N=5.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.23139352	4.68	6	26	0.0024
Pillai's Trace	0.80107053	3.12	6	28	0.0182
Hotelling-Lawley Trace	3.18134407	6.69	6	15.676	0.0012
Roy's Greatest Root	3.13661494	14.64	3	14	0.0001

The GLM Procedure
Repeated Measures Analysis of Variance
Tests of Hypotheses for Between Subjects Effects

Source	DF	Type III SS	Mean Square	F Value	Pr > F
F	2	17971.10754	8985.55377	181.81	<.0001
Error	15	741.33690	49.42246		

Signalons que le dernier test réalisé (celui relatif à l'effet marginal du facteur F) est, comme prévu, très significatif, avec encore une *p-value* inférieure à 10^{-4}. Les test multidimensionnels sont, de leur côté, très significatifs pour le temps et assez significatifs pour les interactions entre le temps et le facteur. Par conséquent, tous les effets déclarés dans ce modèle sont significatifs.

Interrogeons nous maintenant sur la façon dont sont construits ces tests multidimensionnels, autrement dit sur la façon dont sont obtenues les matrices $\mathbf{E}$ pour les erreurs du modèle, $\mathbf{H}_T$ pour les tests relatifs au temps et $\mathbf{H}_{T*F}$ pour ceux relatifs aux interactions. Pour cela, nous devons considérer ce que nous allons appeler les **évolutions**, c'est-à-dire les différences entre les observations de la variable réponse Y à chaque instant t variant de 2 à 4 (de façon générale, de 2 à T) et les observations de cette même variable Y à l'instant initial 1 (instant initial souvent appelé

baseline dans le "jargon" de la statistique médicale et parfois noté 0).

3 Traitement multivarié des variables d'évolution

3.1 Introduction

Définissons donc les trois variables d'évolution suivantes :
$Z_2 = Y_2 - Y_1$; $Z_3 = Y_3 - Y_1$; $Z_4 = Y_4 - Y_1$.

Nous allons maintenant réaliser, toujours avec la procédure GLM de SAS, mais cette fois avec la commande `manova`, une analyse multivariée des trois variables Z_2, Z_3 et Z_4, en testant la significativité du facteur F.

Voici le programme SAS pour réaliser cette analyse :

```
* ------------------------------------- ;
*        calcul des evolutions :        ;
*      Z2=Y2-Y1  Z3=Y3-Y1  Z4=Y4-Y1     ;
* ------------------------------------- ;
data evol;
set repet;
Z2 = Y2 - Y1;
Z3 = Y3 - Y1;
Z4 = Y4 - Y1;
run;
* ------------------------------------- ;
*         MANOVA des evolutions         ;
* ------------------------------------- ;
proc glm data=evol;
class F;
model Z2-Z4 = F / ss3;
manova H = F / printh printe;
run;
```

Et en voici les résultats.

```
                          The GLM Procedure

                       Class Level Information

                    Class         Levels    Values

                    F                  3    1 2 3

                Number of Observations Read          18


                        E = Error SSCP Matrix

                          Z2                Z3                Z4

          Z2      52.261904762      80.476190476      113.19047619
          Z3      80.476190476      196.24761905      255.01904762
          Z4      113.19047619      255.01904762      367.84761905


                    H = Type III SSCP Matrix for F

                          Z2                Z3                Z4

          Z2      157.73809524      271.19047619      388.80952381
          Z3      271.19047619      469.53015873      671.98095238
          Z4      388.80952381      671.98095238      962.15238095
```

Characteristic Roots and Vectors of: E Inverse * H, where
H = Type III SSCP Matrix for F
E = Error SSCP Matrix

Characteristic Root	Percent	Characteristic Vector V'EV=1 Z2	Z3	Z4
3.13661494	98.59	0.09861045	-0.00613949	0.02147662
0.04472912	1.41	-0.19155986	0.13254247	-0.01485006
0.00000000	0.00	-0.10786212	-0.18554264	0.17317314

MANOVA Test Criteria and F Approximations
for the Hypothesis of No Overall F Effect
H = Type III SSCP Matrix for F
E = Error SSCP Matrix

S=2 M=0 N=5.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.23139352	4.68	6	26	0.0024
Pillai's Trace	0.80107053	3.12	6	28	0.0182
Hotelling-Lawley Trace	3.18134407	6.69	6	15.676	0.0012
Roy's Greatest Root	3.13661494	14.64	3	14	0.0001

Notons tout d'abord que, comme toujours, les premiers résultats fournis par la procédure GLM dans un contexte multidimensionnel sont les ANOVA univariées de chacune des variables Z_2, Z_3 et Z_4 par rapport au facteur F. Elles ne figurent pas ci-dessus, mais nous pouvons indiquer qu'elles sont toutes les trois très significatives (*p-values* inférieures ou égales à 10^{-4}) et possèdent un bon coefficient R^2 (compris entre 0.70 et 0.75).

3.2 Tests des interactions

Regardons maintenant les deux matrices $\mathbf{E}$ et $\mathbf{H}$ obtenues à l'issue de cette analyse. La matrice $\mathbf{E}$ de ce modèle est la même que celle obtenue avec les données initiales (les quatre observations de la variable Y et la commande `repeated`), ce qui signifie que les deux modèles sont équivalents. En effet, en prenant en compte, dans le modèle ci-dessus, les variables d'évolution, on fait bien intervenir le facteur temps et, en déclarant le facteur F, ce dernier intervient également. Les résidus de ce nouveau modèle sont donc logiquement les mêmes que dans le modèle initial, dans lequel intervenaient le temps, le facteur et les interactions. Par ailleurs, on constate également que la matrice $\mathbf{H}$ du modèle relatif aux évolutions (le modèle ci-dessus) est identique à la matrice $\mathbf{H}_{T*F}$ associée aux interactions dans le modèle initial. Par conséquent, dans le modèle pour données répétées traité avec GLM, les tests multidimensionnels relatifs aux interactions `temps * facteur` correspondent aux tests relatifs au facteur dans le modèle prenant en compte les évolutions, ce qui est logique.

Une autre façon, plus rigoureuse, de voir les choses est d'écrire le modèle initial, sur les v.a.r. Y_{ijt}, selon le paramétrage centré :

$$Y_{ijt} = \mu + \alpha_j^1 + \alpha_t^2 + \gamma_{jt} + U_{ijt} \ ,$$

où μ est l'effet (moyen) général, les α_j^1 sont les effets principaux (centrés selon j) du facteur, les α_t^2 les effets principaux (centrés selon t) du temps, les γ_{jt} les effets (doublement centrés) d'interactions et les U_{ijt} des v.a.r. erreurs telles que les vecteurs $U_{ij} = (U_{ij1} \cdots U_{ijT})'$ de $\mathbb{R}^T$ sont indépendants, gaussiens, centrés, avec une structure de covariance $\mathbf{\Sigma}$ constante (indépendante de i et de j). On obtient alors :

$$\begin{aligned} Z_{ijt} = Y_{ijt} - Y_{ij1} &= (\alpha_t^2 - \alpha_1^2) + (\gamma_{jt} - \gamma_{j1}) + (U_{ijt} - U_{ij1}) \\ &= (\alpha_t^2 - \alpha_1^2) + (\gamma_{jt} - \gamma_{j1}) + E_{ijt} \ , \end{aligned}$$

en posant $E_{ijt} = U_{ijt} - U_{ij1}$, les E_{ijt} étant toujours des v.a.r. erreurs telles que les vecteurs $E_{ij} = (E_{ij2} \cdots E_{ijT})'$ de $\mathbb{R}^{T-1}$ sont indépendants, gaussiens, centrés, seule leur structure de covariance ayant changée. Notons maintenant $\overline{Z}_{\bullet jt}$ la moyenne des quantité Z_{ijt} sur l'indice i et $\overline{Z}_{\bullet\bullet t}$ la moyenne des mêmes quantités sur les deux indices i et j. Il vient :

$$\overline{Z}_{\bullet jt} = (\alpha_t^2 - \alpha_1^2) + (\gamma_{jt} - \gamma_{j1}) + \overline{E}_{\bullet jt} \text{ et } \overline{Z}_{\bullet\bullet t} = (\alpha_t^2 - \alpha_1^2) + \overline{E}_{\bullet\bullet t} ,$$

puisque les quantités γ_{jt} sont doublement centrées. Les tests multidimensionnels (dans $\mathbb{R}^{T-1}$) de significativité du facteur F font intervenir les deux matrices $\mathbf{H}$ et $\mathbf{E}$ dont les termes généraux sont définis ci-après. Le terme général de $\mathbf{H}$ est

$$\sum_{j=1}^{J} n_j(\overline{Z}_{\bullet jt} - \overline{Z}_{\bullet\bullet t})(\overline{Z}_{\bullet jt'} - \overline{Z}_{\bullet\bullet t'}) = \sum_{j=1}^{J} n_j[(\gamma_{jt} - \gamma_{j1}) + (\overline{E}_{\bullet jt} - \overline{E}_{\bullet\bullet t})][(\gamma_{jt'} - \gamma_{j1}) + (\overline{E}_{\bullet jt'} - \overline{E}_{\bullet\bullet t'})];$$

celui de $\mathbf{E}$ est

$$\sum_{j=1}^{J}\sum_{i=1}^{n_j}(Z_{ijt} - \overline{Z}_{\bullet jt})(Z_{ijt'} - \overline{Z}_{\bullet jt'}) = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(E_{ijt} - \overline{E}_{\bullet jt})(E_{ijt'} - \overline{E}_{\bullet jt'}) .$$

On voit ainsi pourquoi les tests multidimensionnels relatifs au facteur F sur les données d'évolution sont en fait des tests relatifs aux interactions sur les données initiales.

Intéressons nous maintenant à la matrice $\mathbf{H}_T$ intervenant dans les tests relatifs au temps dans le modèle pour données répétées traité avec GLM, tel qu'il a été introduit en 2. Elle est beaucoup plus délicate à obtenir et nous l'explicitons dans le paragraphe suivant.

4 Tests relatifs au temps

Il s'agit des tests multidimensionnels obtenus dans le modèle initial. Ils font intervenir la matrice $\mathbf{E}$, définie dans le point précédent, et la matrice $\mathbf{H}_T$ que nous précisons ci-dessous.

4.1 Expression de la matrice $\mathbf{H}_T$

L'hypothèse nulle correspond à la non influence du temps sur les mesures Y_{ijt}, autrement dit à la constance de ces mesures au cours du temps, autrement dit encore à la nullité des variables d'évolution Z_{ijt}. C'est sur ces dernières que l'on va définir l'hypothèse nulle, puisque les matrices $\mathbf{H}_{T*F}$ et $\mathbf{E}$ ont déjà été définies à partir des variables d'évolution.

En utilisant le paramétrage centré $Z_{ijt} = (\alpha_t^2 - \alpha_1^2) + (\gamma_{jt} - \gamma_{j1}) + E_{ijt}$, la non influence du temps correspond, en fait, à la nullité des $T-1$ paramètres $\alpha_t^2 - \alpha_1^2$ (on notera, ici encore, qu'un tel test n'a de sens que dans la mesure où les interactions entre le temps et le facteur sont elles-même supposées nulles). L'hypothèse nulle peut donc s'énoncer sous la forme suivante :

$$\{H_0 : \alpha_2^2 - \alpha_1^2 = \cdots = \alpha_T^2 - \alpha_1^2 = 0\}.$$

Compte tenu de l'expression donnée plus haut pour les $\overline{Z}_{\bullet\bullet t}$, on peut vérifier sans difficulté que, pour tout t, l'estimateur maximum de vraisemblance de $\alpha_t^2 - \alpha_1^2$ est $\overline{Z}_{\bullet\bullet t} = \dfrac{1}{J}\displaystyle\sum_{j=1}^{J}\overline{Z}_{\bullet jt}$, de sorte que, en pratique, H_0 s'écrit $\{H_0 : \overline{Z}_{\bullet\bullet 2} = \cdots = \overline{Z}_{\bullet\bullet T} = 0\}$, soit encore $\{H_0 : \sum_{j=1}^{J}\overline{Z}_{\bullet j2} = \cdots = \sum_{j=1}^{J}\overline{Z}_{\bullet jT} = 0\}$, que l'on peut réécrire sous la forme $\{H_0 : C'\mathbf{Z} = 0\}$, où $\mathbf{Z}$ est la matrice $J \times (T-1)$ de terme général $\overline{Z}_{\bullet jt}$ et où $C = \mathbb{1}_J$, vecteur de $\mathbb{R}^J$ dont toutes les composantes sont égales à 1.

Pour définir la matrice $\mathbf{H}_T$ correspondant à l'hypothèse nulle considérée, il est préférable ici de revenir à l'expression de la statistique du test de Fisher dans le modèle linéaire gaussien classique. Rappelons que la matrice $\mathbf{H}$ intervenant dans les tests multidimensionnels généralise le numérateur N de la statistique de Fisher et que ce dernier peut s'écrire de différentes façons, l'une des plus commodes étant la suivante :

$$N = \hat{B}'\mathbf{C}[\mathbf{C}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}]^{-1}\mathbf{C}'\hat{B}.$$

Dans cette expression, $\mathbf{X}$ est la matrice d'incidence du modèle : elle est $n \times p$, si n est la taille de l'échantillon considéré et si p désigne le nombre total d'effets fixes (indépendants) pris en compte dans le modèle (ici, $p = J$) ; $\hat{B}$ est l'estimateur maximum de vraisemblance du vecteur β de $\mathbb{R}^p$ des paramètres du modèle ; $\mathbf{C}$ est une matrice $p \times q$ de rang q ($1 \leq q < p$) définissant l'hypothèse nulle sous la forme $\{H_0 : \mathbf{C}'\beta = 0\}$.

Dans le cas considéré ici, nous avons écrit l'hypothèse nulle sous la forme $\{H_0 : C'\mathbf{Z} = 0\}$ et la transposition du terme N au cas multidimensionnel d'ordre $T-1$ nous donne l'expression suivante pour la matrice $\mathbf{H}_T$:

$$\mathbf{H}_T = \mathbf{Z}'C[C'(\mathbf{X}'\mathbf{X})^{-1}C]^{-1}C'\mathbf{Z}.$$

Cette expression de $\mathbf{H}_T$ se simplifie en remarquant que la matrice d'incidence $\mathbf{X}$, de dimension $n \times J$, comporte en colonnes les indicatrices des niveaux du facteur F (les indicatrices des cellules dans le cas général), de sorte que $\mathbf{X}'\mathbf{X} = \text{diag}\,(n_1 \cdots n_J)$, $(\mathbf{X}'\mathbf{X})^{-1} = \text{diag}\,(\frac{1}{n_1} \cdots \frac{1}{n_J})$, $C'(\mathbf{X}'\mathbf{X})^{-1}C = \frac{1}{n_1} + \cdots + \frac{1}{n_J}$ et $[C'(\mathbf{X}'\mathbf{X})^{-1}C]^{-1} = \frac{1}{\frac{1}{n_1} + \cdots + \frac{1}{n_J}} = n^*$, où n^* est la moyenne harmonique des effectifs n_j, divisée par J. Finalement, il vient :

$$\mathbf{H}_T = n^*\mathbf{Z}'\mathbb{1}_{J\times J}\mathbf{Z},$$

où $\mathbb{1}_{J\times J}$ est la matrice carrée d'ordre J ne comportant que des 1.

Remarque 1 *On notera que, dans un plan équilibré avec n_0 observations par cellule ($n = Jn_0$), il vient : $n^* = \frac{n_0}{J}$. De plus, dans ce cas, le terme général de la matrice $\mathbf{H}_T$ s'écrit $n\overline{Z}_{\bullet\bullet t}\overline{Z}_{\bullet\bullet t'}$.*

Remarque 2 *Il est important de remarquer que l'hypothèse nulle $\{H_0 : \mathbf{C}'\mathbf{Z} = 0\}$ est définie par une seule contrainte sur $\mathbf{Z}$, la matrice $\mathbf{C}$ ne comportant qu'une seule colonne ($q = 1$). En fait, H_0 exprime le centrage, dans $\mathbb{R}^J$, des $T-1$ vecteurs $\overline{Z}_{\bullet t}$ ($t = 2, \ldots, T$), de coordonnées $\overline{Z}_{\bullet jt}$. L'hypothèse nulle signifie donc que les vecteurs $\overline{Z}_{\bullet t}$ sont dans un hyperplan de $\mathbb{R}^J$, ce qui correspond bien à une contrainte unique. Par conséquent, le d.d.l. associé à H_0 (noté ν_H dans les tests multidimensionnels) vaut toujours 1 dans ce cas : $\nu_H = 1$. Par suite, la matrice $\mathbf{H}_T\mathbf{E}^{-1}$ n'admet qu'une seule valeur propre.*

4.2 Application

Revenons maintenant aux données traitées depuis le début. Pour chaque variable d'évolution, nous calculons ses moyennes partielles relativement aux trois niveaux du facteur F, afin de déterminer la matrice $\mathbf{Z}$ définie plus haut.

Voici le programme SAS réalisant ce calcul :

```
proc means data=evol;
var Z2-Z4;
by F;
run;
```

Et voici les résultats obtenus :

```
---------------------------------- F=1 ----------------------------------

                          The MEANS Procedure

 Variable    N           Mean        Std Dev        Minimum        Maximum
 ------------------------------------------------------------------------
 Z2          5      3.0000000      1.4142136      2.0000000      5.0000000
 Z3          5      5.6000000      1.8165902      3.0000000      8.0000000
 Z4          5      9.2000000      3.0331502      6.0000000     14.0000000
 ------------------------------------------------------------------------
```

```
---------------------------------- F=2 ----------------------------------

 Variable     N           Mean        Std Dev        Minimum        Maximum
 ------------------------------------------------------------------------
 Z2           6      8.8333333      2.6394444      4.0000000     12.0000000
 Z3           6     16.3333333      5.7154761      7.0000000     21.0000000
 Z4           6     24.3333333      7.4475947     12.0000000     32.0000000
 ------------------------------------------------------------------------

---------------------------------- F=3 ----------------------------------

 Variable     N           Mean        Std Dev        Minimum        Maximum
 ------------------------------------------------------------------------
 Z2           7      2.2857143      1.2535663      1.0000000      4.0000000
 Z3           7      5.4285714      1.8126539      4.0000000      8.0000000
 Z4           7      8.5714286      2.9920530      5.0000000     14.0000000
 ------------------------------------------------------------------------
```

La matrice $\mathbf{Z}$, de dimension $J \times (T-1)$, soit ici 3×3, s'obtient à partir des moyennes partielles (`Mean`) obtenues ci-dessus. Il vient :

$$\mathbf{Z}' = \begin{pmatrix} 3.00 & 8.83 & 2.29 \\ 5.60 & 16.33 & 5.43 \\ 9.20 & 24.33 & 8.57 \end{pmatrix}.$$

On en déduit :

$$\mathbf{Z}'\mathbb{1}_{3\times 3}\mathbf{Z} = \begin{pmatrix} 199.35 & 386.32 & 594.48 \\ 386.32 & 748.67 & 1152.07 \\ 594.48 & 1152.07 & 1772.81 \end{pmatrix}.$$

On peut par ailleurs vérifier que $n^* = \dfrac{210}{107}$, ce qui permet de calculer (aux erreurs d'arrondi près) :

$$\mathbf{H}_T = n^*\mathbf{Z}'\mathbb{1}_{3\times 3}\mathbf{Z} = \begin{pmatrix} 391.24 & 758.21 & 1166.73 \\ 758.21 & 1469.36 & 2261.07 \\ 1166.73 & 2261.07 & 3479.35 \end{pmatrix}.$$

On retrouve bien ainsi la matrice $\mathbf{H}_T$ fournie en 2 par la procédure GLM de SAS avec la commande `repeated`.

Remarque 3 *Pour le calcul des 4 statistiques de Fisher associées aux tests multidimensionnels et de leurs d.d.l., on utilisera ici $\nu_H = 1$ comme indiqué plus haut, $\nu_E = n - J$ (15 ici, puisque la matrice $\mathbf{E}$ est la même, quelle que soit l'hypothèse testée dans le modèle) et $D = T - 1$ (3 ici).*

5 Bilan

Dans les points 3 et 4 ci-dessus, on a explicité le calcul des matrices $\mathbf{H}_T$, $\mathbf{H}_{T*F}$ et $\mathbf{E}$ permettant de déterminer les statistiques des tests multidimensionnels (et, principalement, le test de Wilks), ces matrices intervenant dans la procédure GLM du logiciel statistique SAS lorsqu'on déclare la commande `repeated` pour traiter des données répétées. La logique de ces tests apparaît ainsi clairement, et il semble tout indiqué de les utiliser pour tester la significativité des effets fixes dans des modèles linéaires mixtes pour données répétées.

On notera que ces tests multidimensionnels peuvent être des tests approchés mais, en général, les approximations obtenues sont bonnes. On notera également que ces tests ne dépendent pas de la structure de covariance choisie pour les données répétées (matrice $\mathbf{R}$) et qu'on peut donc les mettre en œuvre avant de choisir cette dernière. Pour le choix de la matrice $\mathbf{R}$, c'est la procédure MIXED de SAS qui est la plus appropriée, de même que pour l'estimation des composantes de la variance correspondant aux effets aléatoires.