Modèles statistiques

Un modèle statistique est une fonction mathématique f pour représenter les relations entre les variables dans un système ou processus. La fonction f relie des variables X (variables explicatives ou facteurs) et une ou plusieurs variables de réponse Y (variables à expliquer). Les modèles statistiques sont généralement linéaires dans les paramètres (constantes inconnues). Le modèle général s'écrit :

$$Y = f(X1, X2,\ldots; XA, XB,\ldots; \beta_0, \beta_1, \beta_2,\ldots) + \varepsilon \qquad \varepsilon \sim N(0,\sigma^2) \qquad (1)$$

variables (facteurs) continues : X1, X2, …
variables (facteurs) catégoriques : XA, XB, …
paramètres inconnus : $\beta_0, \beta_1, \beta_2$,…
erreur : ε composante à moyenne nulle et variance constante σ^2
représente l'ensemble de tous les facteurs inconnus
non pris en compte dans le système

La fonction f est inconnue et on utilise des approximations formées par de fonctions polynomiales de degré 1 ou degré 2 (facteurs continus seulement) avec ou sans des termes produits X*X (interactions) entre les facteurs. Le modèle doit réfléter aussi des éléments additionnels ou caractéristiques des facteurs.

Caractéristiques, classification et rôle des facteurs

cacractéristiques de X	définition
nature	continue ou catégorique
modalités (valeurs)	fixes (choisies et contrôlées) ou aléatoires (résultat d'un échantillonnage)
influence sur Y	moyenne (cas fixe) ou variance (cas aléatoire)
relation avec les unités expérimentales (UE)	inter ("betwen") ou intra ("whithin") inter : modalités distribuées sur les UE intra : chaque UE prend toutes les modalités
liens entre les modalités	croisées ou emboitées croisées : toutes les combinaisons des modalités de 2 facteurs sont possibles emboitées : les modalités prises par un facteur sont spécifiquement associées aux modalités d'un autre facteur
bloc	facteur sans interaction avec les autres X peut-être fixe (contrôlé) ou aléatoire (UE)
covariable(rôle)	facteur continu mesuré sans interaction avec tous les autres X continus ou catégoriques
UE	en général ils sont relativement homogènes et on en tient pas compte dans le modèle; lorsqu'ils sont hérérogènes on en tient compte avec des covariables; dans d'autres cas, on peut en tenir compte dans le modèle et considérer UE comme un facteur aléatoire.

Modèles avec facteurs catégoriques et facteurs continus

En présence de facteurs continus et catégoriques, on doit se poser une question préalable pour adopter et ajuster un modèle :

les facteurs continus X1, X2,…on-ils des interactions X1*XA, X1*XB, X2*XA,… avec les facteurs catégoriques XA, XB,…?

Généralement et dans beaucoup de situations pratiques, il n'y a pas pas.
Mais il faut valider cette décision avant d'adopter un des 2 modèles Ma et Ms

Ma : modèle avec des interactions (continu x catégorique) facteurs
Ms : modèle sans interactions, appelé modèle d'analyse de covariance

On peut montrer que l'absence d'interactions X1*XA, X1*XB, X2*XA,…se traduit par un modèle avec des pentes égales (homogènes) sur les facteurs continus X1, X2,..pour les différentes combinaisons des modalités des facteurs catégoriques XA, XB,…

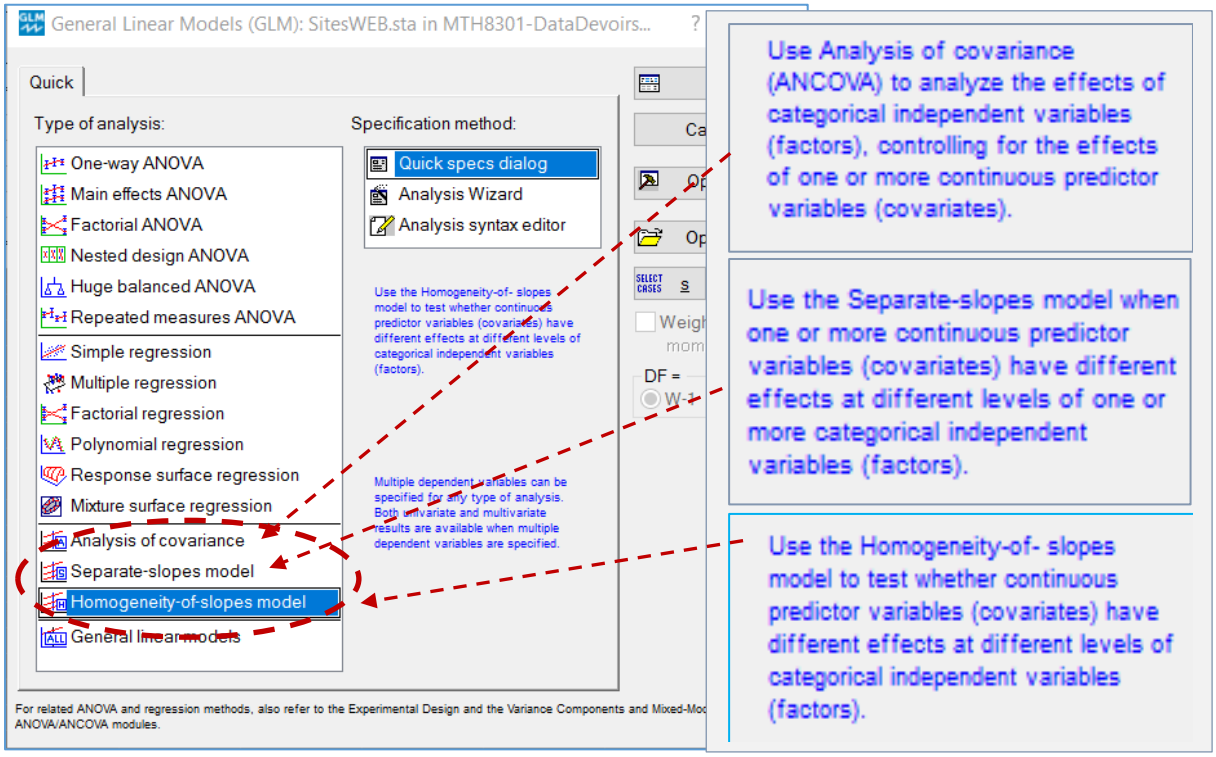
On peut prendre cette décision en comparant les 2 modèles Ma Ms.
En pratique, on commence par ajuster le modèle Ma et l'on teste la significativité des effets d'interactions. Si au moins un des test est significatif , on conserve le modèle Ma. Sinon, on peut adopter le modèle Ms qui est plus simple que le modèle Ma.

***STATISTICA* offre 3 modules dans GLM pour traiter la situation :**

1. ***Homogenity* of slopes model (test d'égalité des pentes) : Ma ou Ms?**
2. ***Separate* slopes model : Ma**
3. ***Analysis* of covariance (equal slopes) : Ms**

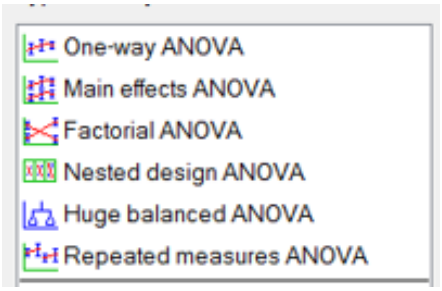
On adoptera le modèle Ma ou Ms selon le résultat de l'analyse *homogeneity.*
Si on ne rejette pas le test de pentes égales, alors on adoptera le modèle Ms et on développera une analyse avec un modèle classique aussi appelé modèle d'analyse de covariance. Dans le cas contraire, on adoptera Ms et on fera son ajustement.

STATISTICA



Syntaxe pour la spécification des effets inter ("*between effects*")

Lors de la mise en œuvre du module GLM (General Linear Model) on doit spécifier les variables X et Y. Si on est dans le cas avec seulement des facteurs fixes, il y a plusieurs sous-modules que l'on peut choisir (voir page 2). Le choix est dicté en autres par la présence de facteurs inter et de facteurs intra dans nos données ou si on a des facteurs emboités et croisés. L'écran suivant donne la liste des sous-modules ANOVA de GLM.



Pour illustrer la syntaxe pour spécifier les effets inter, nous allons utiliser les données suivantes dont on voit une partie. Les données représentent la mesure du stress (v9) en tenant compte de 7 facteurs dont 4 sont catégoriques (v2, v3, v4, v5) et 3 sont continues (v6, v7, v8). Il n'y a pas de facteurs intra dans cet exemple, donc pas de mesures répétées sur les UE.

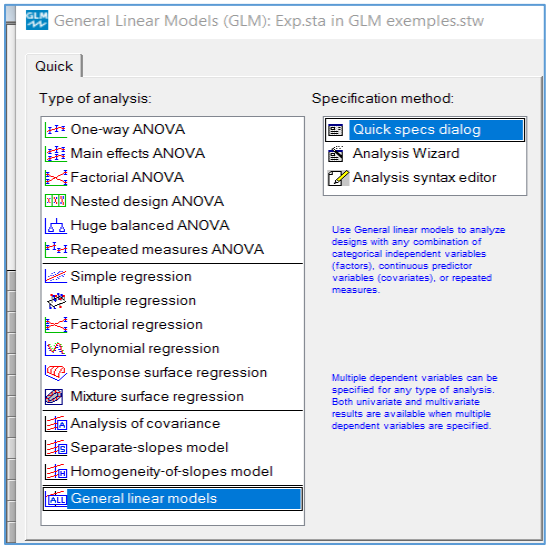
Données

Performance on a memory test under stressful conditions
v1: subject_ID (s01,...,s48)
v2: XA_group (experimental, control)
v3: XB_genre (male, female)
v4: XC_time reward (before, after)
v5: XD_whether the subject was offered money (not_paid, paid)
v6: X1_nb_correct_P1 (number of corrected answers on problem 1)
v7: X2_nb_correct_P2 (number of corrected answers on problem 2)
v8: X3_nb_correct_P3 (number of corrected answers on problem 3)
v8: Y_stress (measur of stress)

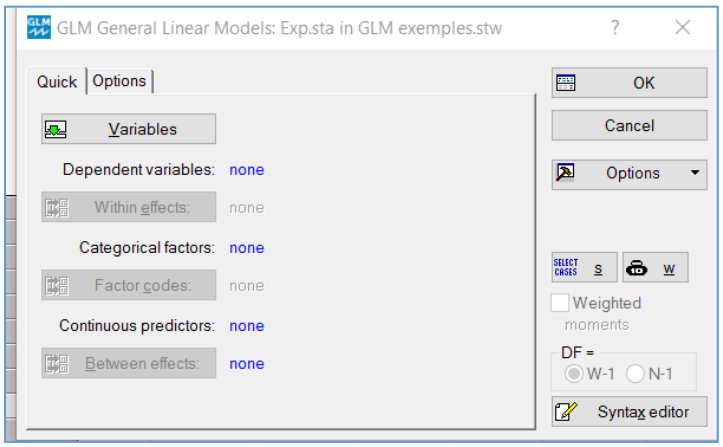
	1 subject_ID	2 XA_group	3 XB_genre	4 XC_time reward	5 XD_paid	6 X1_nb_correct_P1	7 X2_nb_correct_P2	8 X3_nb_correct_P3	9 Y_stress
1	s01	EXPERMTL	MALE	BEFORE	NOT_PAID	12	4	6	1,41
2	s02	EXPERMTL	MALE	BEFORE	NOT_PAID	3	3	7	1,73
3	s03	EXPERMTL	MALE	BEFORE	PAID	7	6	0	0,00
4	s04	EXPERMTL	MALE	BEFORE	PAID	11	7	3	1,41
5	s05	EXPERMTL	MALE	AFTER_1	NOT_PAID	8	2	7	12,83

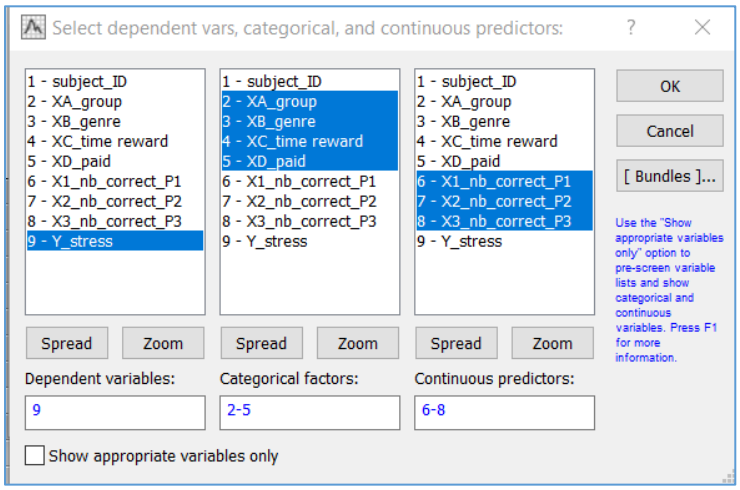
Nous avons choisi le sous module *General Linear Model* pour faire l'analyse.
On aurait pu aussi employer l'un des 3 sous modules d'analyse de covariance.

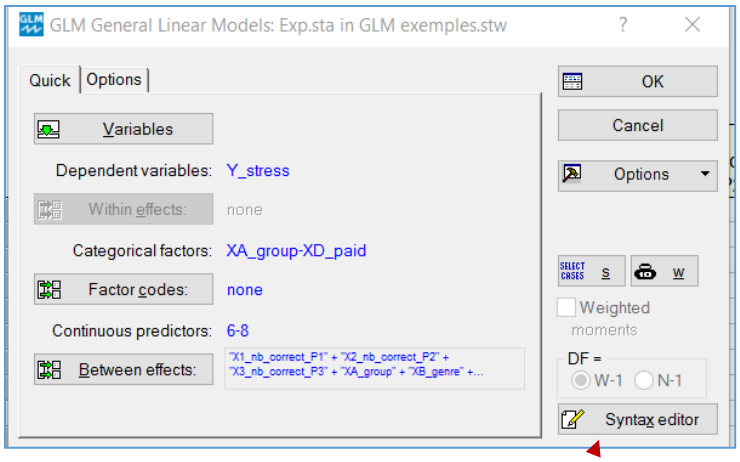
GLM



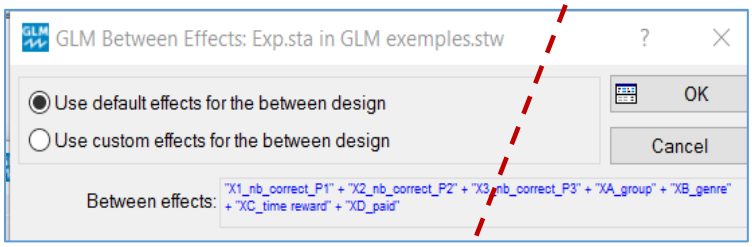
La mise en œuvre est amorcée par la spécification des variables X et Y.





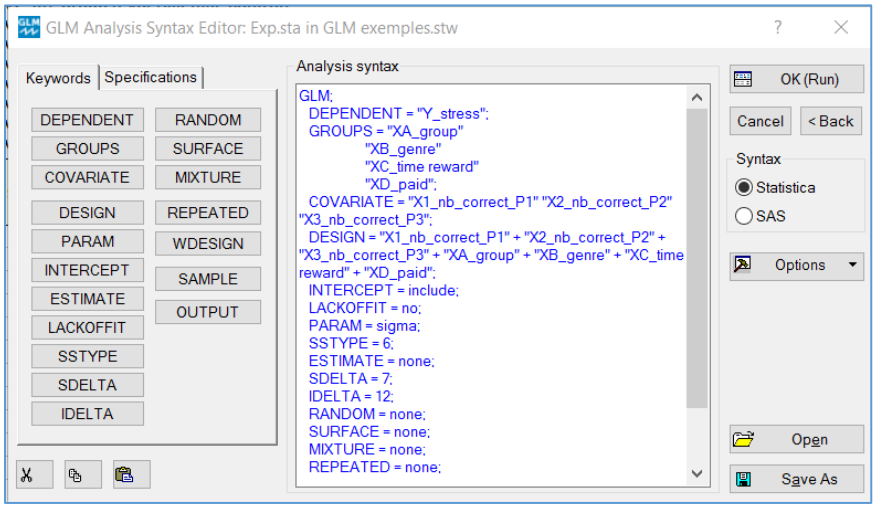


Le modèle choisit par *STATISTICA* pour les effets inter ("*Between*") est additif (ordre 1).

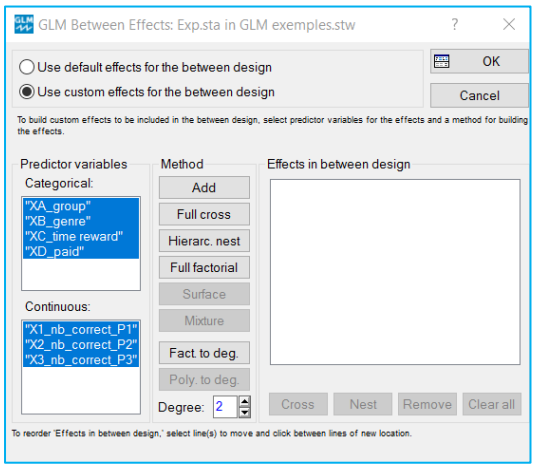


Si on clique sur le bouton "*Syntax editor* "on peut voir toutes les commandes employées pour le traitement de l'analyse.

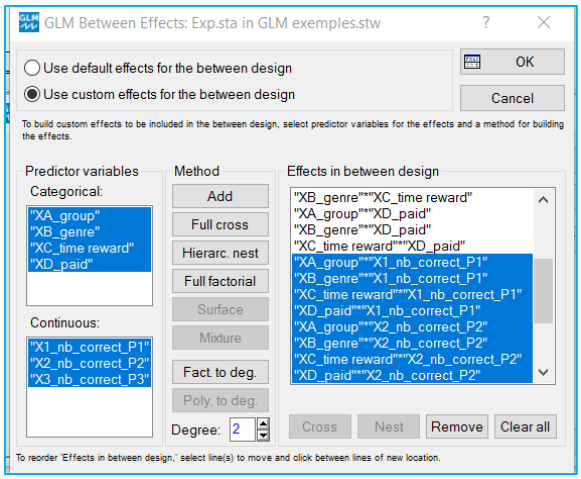
Voici la copie de cet écran. Il est complètement éditable en particulier la commande *DESIGN.*



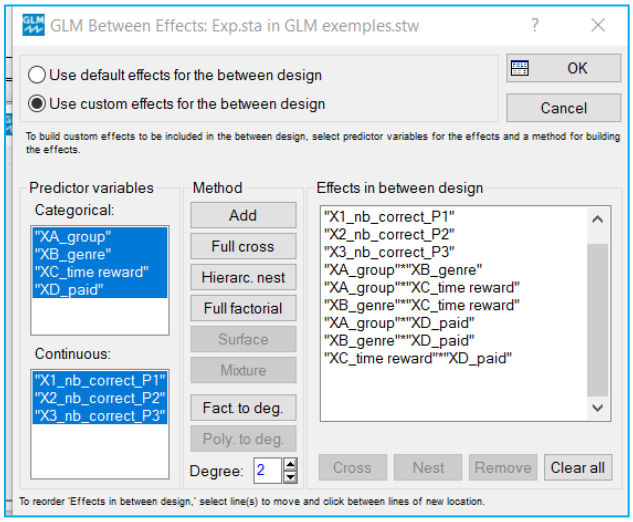
On peut modifier le modèle en cliquant sur l'onglet "*Between*" de l'écran précédent.
On obtient alors une boite de dialogue qui permet de changer le modèle.
Par exemple, on pourrait ajouter des effets d'interaction entre les facteurs catégoriques ce qui donnerait le modèle classique d'analyse de covariance.



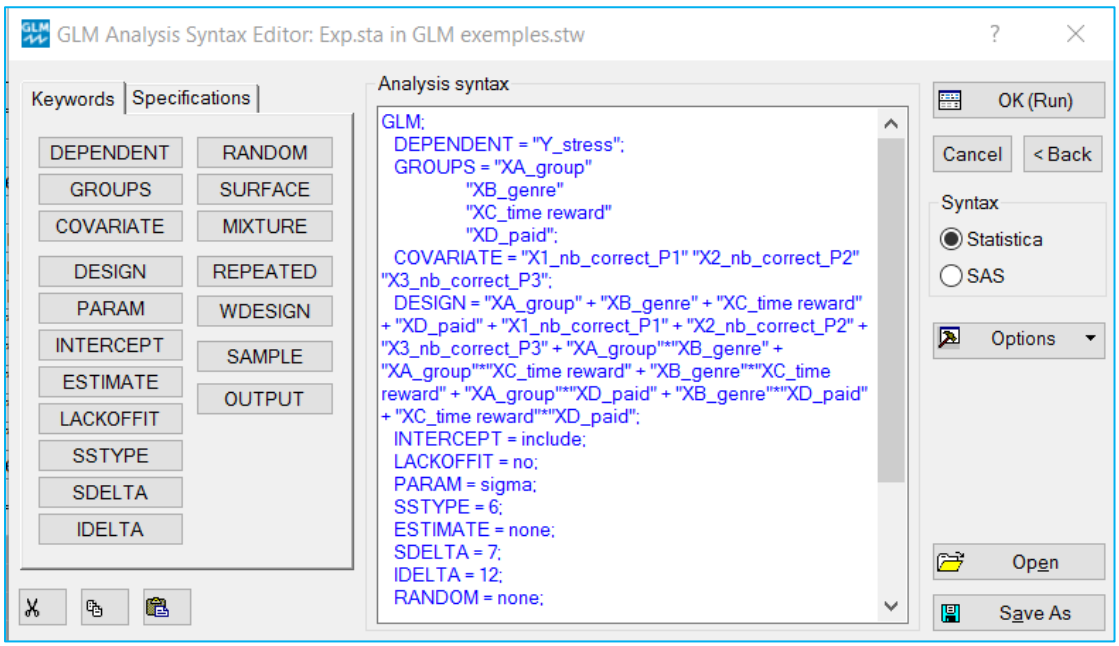
Nous avons demandé un modèle factoriel de degré 2 : tous les effets principaux des 7 facteurs + tous les effets d'interaction (il y en a 42) entre les 7 facteurs. Mais on doit enlever les 36 effets d'interactions entre les facteurs catégoriques et les facteurs continus. Il restera seulement 6 effets d'interaction entre les facteurs catégoriques. On y arrive en édidant la liste des effets avec la commande *remove.*



On obtient la nouvelle liste finale des 13effets (7 pincipaux + 6 interactions) du modèle d'analyse de covariance.



On peut aussi voir le modèle final dans l'écran *syntax editor* :



Syntaxe pour la spécification d'un modèle statistique en *STATISTICA*

Afin de spécifier ou éditer le modèle dans la boite *syntax éditor* on peut utiliser des commandes dont voici les principales. Les facteurs catégoriques ou continus sont notés par des lettres A, B, …

Termes simples et opérateurs

À la droite de l'énoncé *Design*, on peut faire la liste des effets séparés par le signe + (plus) operateur. Les effets simples peuvent être spécifiés comme suit :

A	**effet principal du facteur A**
A*B	**A par B effet d'interaction entre A et B**
A(B)	**A emboité dans B; les modalités du facteur catégorique A sont emboités dans les modalités du facteurs catégorique B**

Opérateurs macros commandes (raccourci)

Opérateur Bar |

A|B modèle factoriel complet de A et B : to A+B+A*B

A|B|C modèle factoriel complet des facteurs A, B, C :
A + B + C + A*B + A*C + B*C + A*B*C

Opérateur factoriel de degré @k k = 1, 2, 3…

A|B|C @2 modèle factoriel de A, B, et C, jusqu'au degré 2;
donnera le modèle A + B + C + A*B + A*C + B*C
si k = 1 on aura un modèle de degré 1 contenant seulement des effets principaux

Groupement de termes et opérateurs

A | (B+C) signifie le modèle factoriel A + B + C + A*B + A*C

Opérateur de soustraction (-)

A|B|C-C modèle factoriel complet pour A, B, et C sans l'effet principal de C
donne le modèle : A + B + A*B + A*C + B*C + A*B*C

A|B|C-|C modèle factoriel complet pour A, B, et C sans toutes les interactions impliquant C: le modèle est A + B + C + A*B

A|B|C-(A*C) modèle factoriel pour A, B, et C sans l'interaction A*C :
donne le modèle : A + B + C + A*B + B*C + A*B*C

A|B|C-|(A*C) modèle factoriel complet pour A, B, et C sans l'interaction A*C
donne le modèle : A + B + C + A*B + A*C + B*C
l'interaction A*B*C est absente

Modèles linéaires mixtes : facteurs fixes et facteurs aléatoires

On distingue 2 formulation du modèle linéaire selon la présence ou non de facteurs aléatoires. Soit la notation suivante :

X : facteurs fixes
Z : facteurs aléatoires

CAS 1 : modèle linéaire avec uniquement des facteurs fixes X

$$Y = X\beta + \varepsilon \qquad (2)$$

où

Y est le vecteur des valeurs de la réponse
X est la matrice de design qui représente tous les effets
β est le vecteur inconnu de paramètres
ε est le vecteur d'erreur

CAS 2 : modèle linéaire mixte avec des facteurs fixes X et des facteurs aléatoires Z

$$Y = X\beta + Z\gamma + \varepsilon \qquad (3)$$

où

Y est le vecteur des valeurs de la réponse
X est la matrice de design qui représente tous les effets
β est le vecteur inconnu de paramètres pour les effets fixes
Z est matrice de design qui représente les effets aléatoires
γ est le vecteur inconnu de paramètres pour les effets aléatoires
ε est le vecteur d'erreur