TRAVAUX DIRIGES DE STATISTIQUES

INTRODUCTION AU MODELE LINEAIRE

Ecole Nationale Vétérinaire de Toulouse

Département des Sciences biologiques et fonctionnelles

UP de Biométrie D. Concordet

Objectifs du TD

Le modèle linéaire est un terme générique qui recouvre une famille de modèles dans lesquels les paramètres à estimer interviennent linéairement. Cette famille de modèles statistiques est probablement une des plus utilisée quand il s'agit d'analyser pratiquement de vraies données. Il permet de dépasser le cadre souvent très scolaire (et peu utile en pratique) des tests de student, de Fisher et du chi2 traditionnellement enseignés comme base des statistiques (la maîtrise de ces outils de base est un préalable nécessaire à l'utilisation de méthodes plus construites). Son utilisation est facile pour peu qu'un certain nombre de concepts élémentaires aient été bien compris. Les objectifs de ces différentes séances de TD est d'illustrer avec des exemples qui pourraient ressembler à vos projets de statistiques l'utilisation du modèle linéaire. Deux groupes d'exercices vous sont proposés. Les exercices 1,2 et 3 sont des exercices qui devraient vous permettre d'avoir une représentation mentale simple des concepts vus en cours. Ces premiers exercices sont dépouillés de tous les problèmes complexes rencontrés lors de "vraies" analyses. Vous devrez simuler les données puis les analyser. Les exercices suivants ressemblent à des problèmes que vous pouvez rencontrer en pratique.

Exercice N°1.

L'objet de cet exercice est de comprendre comment des données normalement distribuées peuvent être générées par un ordinateur. On peut imaginer que faire une expérience (et donc effectuer des mesures) revient en fait à "faire tourner" le simulateur complexe qu'est la nature. En première approximation ou plus exactement d'un point de vue macroscopique, l'observation de la mesure d'un phénomène peut être considérée comme la réalisation d'une variable aléatoire.

Le logiciel que nous utiliserons pour cet exercice est EXCEL.

Rappelons que pour taper une formule dans excel, il faut tout d'abord taper = puis taper le nom de la fonction (éventuellement ses arguments séparés par des virgules)

Exemple : supposons que le nombre 3 soit dans la cellule B1, vous voulez que son logarithme se trouve dans la cellule C1. Placez le curseur en C1, taper *=ln(B1)*, puis appuyer sur entrée. La cellule C1 contient maintenant le nombre ln(3).

1) Simulation d'une variable aléatoire uniformément distribuée sur]0,1[

Ouvrez XL, nommez *U* la colonne A (ie Ecrivez *U* en A1). Placez le curseur sur A2 et tapez *=alea()*

Copiez la formule que vous venez de taper en A2 jusqu'en A101.

Les nombres de la colonne U sont tirés selon une loi uniforme sur]0,1[.

2) Simulation d'une variable aléatoire N(0,1)

Nous allons effectuer un tirage de 100 réalisations d'une variable aléatoire X distribuée selon une loi N(0,1).

A cet effet, nous allons utiliser la propriété suivante :

Soient une loi L dont la fonction de répartition est F, et U une variable aléatoire uniformément distribuée sur]0,1[alors, $F^{-1}(U)$ est distribuée selon une loi L.

La fonction réciproque de la fonction de répartition d'une loi normale N(0,1) est nommée dans XL :

LOI.NORMALE.STANDARD.INVERSE(). En d'autres termes si vous tapez = *LOI.NORMALE.STANDARD.INVERSE(A2)* dans la cellule B2, le contenu de la cellule B2 sera une réalisation d'une N(0,1).

Nommez eps la colonne B, et faites 100 tirages selon une loi N(0,1).

3) On rappelle que si eps ~ N(0,1) alors Z=a+b eps~ N(a,b^2).

On veut simuler le poids des étudiants de deuxième année du premier cycle. On sait que le poids est normalement distribué. De plus, il est admis qu'un étudiant pèse en moyenne 68 ± 5 kg (moyenne ± SD).

Simulez dans la colonne le poids de 100 étudiants de première année

Calculez la moyenne de cette colonne (=moyenne(C2:C101)).

Calculez l'écart-type de cette colonne (=ecartype(C2:C101))

Tapez sur F9

Que se passe t-il ?

Utilisez vos connaissances en statistiques pour construire un intervalle dans lequel se trouvent 95 % des

- moyennes
- des variances que vous obtenez en frappant sur F9.

Exercice N°2

L'objet de cet exercice est de d'expérimenter l'utilisation de l'analyse de variance.

On sait qu'un grand nombre de facteurs ont une influence sur le poids des étudiants de deuxième année de premier cycle. En particulier, le sexe et l'équilibre de l'alimentation en sucre et en graisse ont entre autres une influence importante.

On peut imaginer que la distribution des poids est normale d'écart-type 5 kg et que le poids moyen varie de la façon suivante :

	Alimentation		
Sexe	Pauvre	Equilibrée	Riche
Garçon	65	70	80
Fille	60	70	75

Pour simplifier, on suppose que dans une promotion de 120, le nombre d'étudiants qui appartiennent à chacune des cellules du tableau est fixe (non aléatoire) et égale à 20 (cette hypothèse est bien entendue complètement fausse).

1) Simuler les poids observables dans une promo de deuxième année du premier cycle. A cet effet créez 3 colonnes intitulées Sexe, Alimentation, poids qui contiendront respectivement

le sexe de l'étudiant :

1 pour un garçon

2 pour une fille

le type d'alimentation

1 si pauvre

2 si équilibré

3 si riche

le poids simulé correspondant.

2) Etablissez le tableau des moyennes de poids par sexe et par type d'alimentation. A partir de ce tableau construisez le graphique des poids moyens par type d'alimentation et par sexe.

Appuyer sur la touche F9. Expliquez ce qui se passe.

3) Copier les colonnes sexe, alimentation, poids et collez en **les valeurs** dans une autre feuille.

Ecrivez le modèle qui vous semble le mieux approprié pour analyser ces données. Faites les hypothèses adéquates.

4) A l'aide du modèle que vous venez d'écrire, analysez les données (table d'analyse de la variance, R^2).

5) Testez respectivement l'effet de l'interaction, du sexe et du type d'alimentation. Concluez.

6) Copiez les données dans un fichier SYSTAT. Vous allez maintenant pouvoir vérifier si l'analyse que vous venez d'effectuer (à la main) est sans erreur. A cet effet, vous allez analyser les données à l'aide de systat. Systat a besoin de connaître le modèle à utiliser pour analyser les données. Le modèle que vous avez proposé à la question 3 devrait ressembler à

$$Y_{i,j,k} = \mu + S_i + A_j + (S * A)_{i,j} + \varepsilon_{i,j,k}$$

où $Y_{i,j,k}$ est le poids du k ième étudiant de sexe i qui a une alimentation de type j,

μ est l'effet moyen général,

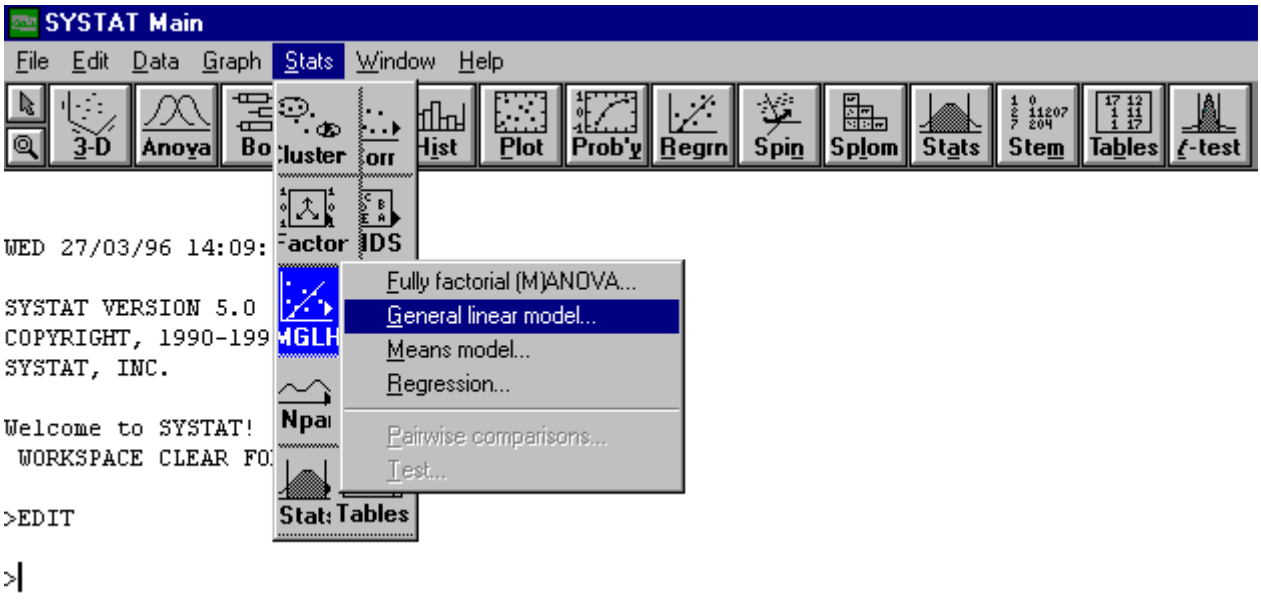
S_i est l'effet différentiel du niveau i du facteur sexe,

A_j est l'effet différentiel du niveau j du facteur type d'alimentation,

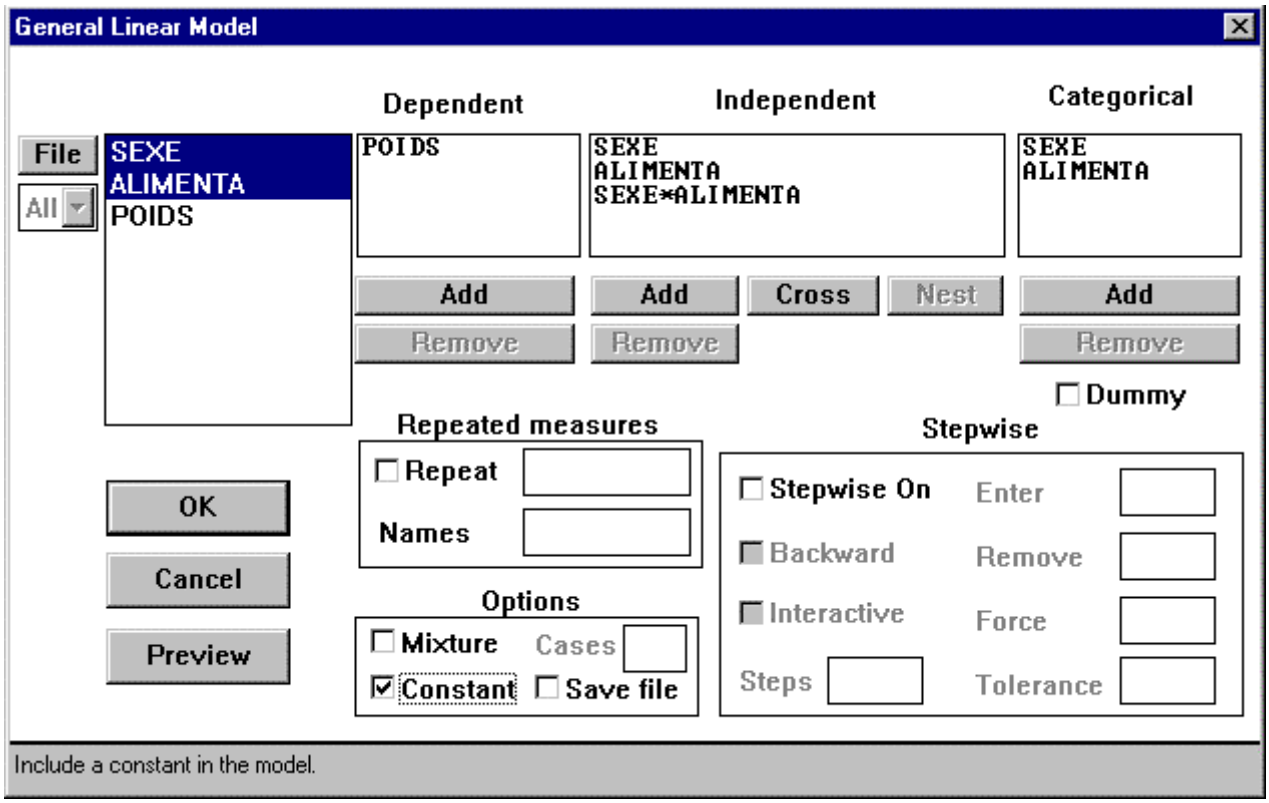
$(S*A)_{ij}$ est l'effet de l'interaction entre sexe et type d'alimentation

et ε_{ijk} est un terme résiduel

Pour le déclarer à systat cliquez sur les options montrées sur le graphique suivant :



Vous obtenez alors une boite de dialogue que vous devez remplir comme indiqué ci-dessous



La variable à expliquer (Dependent) est le poids, les variables explicatives (indépendent) sont le poids, l'alimentation, et l'interaction entre les deux. Ces variables explicatives ne sont pas continues, ce sont des facteurs (Categorical). Le modèle comporte une constante μ, ce qu'il faut déclarer au logiciel en cliquant sur la case à cocher intitulée Constant.

Comme vous pouvez le constater, systat n'utilise pas de notation indicielle.

Cliquez sur OK et décrivez les résultats obtenus.

7) Nous allons maintenant nous assurez que les postulats nécessaires à la bonne interprétation des résultats sont raisonnables. Pour ce faire, recommencez la même "manip" qu'à la question précédente et cliquez sur la case à cocher nommée Save file. Ceci a pour effet de sauvegarder les résidus de votre modèle. Donnez un nom au fichier qui doit contenir ces résidus, et cliquez sur l'option Residuals and diagnostics.

Ouvrez le fichier que vous venez de créer et vérifier (sur les résidus) les 3 postulats fondamentaux.

8) Concluez

Exercice N°3

La régression linéaire simple est la plus simple des méthodes statistiques qui permet l'analyse de la liaison entre deux variables quantitatives. Elle est souvent utilisée en première

approximation pour décrire ou prédire. L'objet de cet exercice est de vous familiariser avec son utilisation.

Lors de la demande d'AMM de certains médicaments vétérinaires, l'industrie pharmaceutique doit fournir (proposer) un délai d'attente. Le délai d'attente est le temps qu'il faut attendre après la dernière administration du médicament pour que la concentration de ce médicament dans les tissus consommables soit inférieure à une limite fixée : les LMR (limites maximales de résidus). Des considérations pharmacocinétiques permettent de modéliser la décroissance de ces concentrations dans les tissus consommables avec un modèle compartimental. Considérons un médicament dont la courbe moyenne de déplétion dans le foie est de la forme :

$$Y_t = \exp(a - bt + \varepsilon_t)$$

a et b sont des paramètres inconnus et ε_t est une variable aléatoire distribuée selon une loi $N(0, \sigma^2)$. Afin de pouvoir doser la concentration du médicament dans le foie, il est nécessaire d'abattre les animaux qui ont reçu le médicament en question. Supposons que dans la population,

$a=ln\ 100$, $b= 0.1$, $var(\varepsilon_t)=0.01= \sigma^2$

1) Le modèle (mod) est il un modèle linéaire ? Est il homoscédastique ? La distribution de Y_t est elle normale ?

2) Six temps d'abattage ont été retenus : t1=0, t2= 5, t3=10, t4=20, t5=30, t6= 60 et à chacun de ces temps, quatre animaux doivent être abattus.

Simuler dans XL la concentration mesurée sur chaque animal à chacun de ces temps d'abattages.

3) Copier les données que vous venez de simuler dans systat. Estimer à l'aide d'un modèle convenablement choisi les paramètres a,b, σ^2. Vérifiez les 3 hypothèses fondamentales.

4) Donnez un intervalle de confiance de sécurité 95% de $\hat{a}$ et $\hat{b}$. Les valeurs a=ln 100 et b= 0.1 sont elles raisonnables ?

5) La LMR de ce produit est fixée à 10.

Jusqu'à ces dernières années, le délai d'attente a été défini comme le temps à partir duquel on pouvait affirmer avec un risque inférieur à 5% que la concentration moyenne de résidu dans le tissu choisi (ici le foie) est inférieur à la LMR. Calculer le délai d'attente.

L'objet des exercices qui suivent est de vous entraîner à la construction de modèles statistiques. Pour ce faire, nous allons considérer un certain nombre de situations

différentes pour lesquelles vous devrez tout d'abord comprendre le plan d'expérience, écrire le modèle, analyser les données, vérifier les 3 postulats et conclure.

Exercice N°4

Dans le but d'étudier l'effet tératogène de 4 traitements (T1, T2, T3, T4) chez la souris l'étude suivante à été menée :

Chacun des traitements a été administré à 5 souris différentes (au total 20 souris ont donc été utilisée). Le diamètre d'un os particulier a ensuite été mesuré sur trois embryons de chaque souris. En effet, l'effet tératogène des ces traitements s'il existe devrait modifier le diamètre des os des embryons.

1) Décrivez le plan d'expérience utilisé. Y a t-il des facteurs aléatoires, fixes ?

2) Pour un traitement donné, écrivez le modèle qui décrit les variations du diamètre des os en fonction du facteur mère.

3) Ecrivez le modèle général qui permet d'analyser le diamètre des os en fonction du facteur traitement.

Le fichier EX4 contient les données relatives à cette expérience. Analysez ces données en utilisant le modèle que vous venez d'écrire.

4) Vérifiez sur les résidus si les 3 postulats fondamentaux sont raisonnables.

5) En fait les traitements T1, T2, T3, T4 correspondent à l'administration (ou à la non administration) de deux traitements A et B. Le codage qui a été utilisé est le suivant :

	Pas de Traitement A	Traitement A
Pas de traitement B	T1	T2
Traitement B	T3	T4

Ecrivez le modèle adéquat pour analyser ces données (ie pour savoir quel traitement (A ou B ou les 2) est tératogène.

Le fichier EX4bis contient les mêmes données que le fichier EX4 mais la colonne T a été remplacée par deux colonnes A et B (1 signifie absence de traitement, 2 signifie présence de traitement).

6) Analysez les données avec ce modèle.

Exercice N°5

Une expérience est organisée pour mesurer l'influence de 4 régimes sur la croissance d'animaux de rente. Pour chaque régime huit animaux ont été utilisés puis pesés à des âges différents.

1) Décrivez le plan d'expérience utilisé. Y a t-il des facteurs aléatoires, fixes ?

2) Pour un régime donné, écrivez le modèle qui décrit les variations du poids en fonction de l'âge de l'animal.

3) Ecrivez le modèle général qui permet d'analyser le poids en fonction de son âge et du régime qu'il a suivi.

Le fichier EX5 contient les données relatives à cette expérience. Analysez ces données en utilisant le modèle que vous venez d'écrire.

4) Vérifiez sur les résidus si les 3 postulats fondamentaux sont raisonnables.

5) En fait les régimes R1, R2, R3, R4 correspondent à la complémentation (ou à la non complémentation) en vitamine C et en fer. Le codage qui a été utilisé est le suivant :

	Pas de complément en vit. C	Complément en vit. C
Pas de complément en fer	T1	T2
Complément en fer	T3	T4

Ecrivez le modèle adéquat pour analyser ces données (ie pour savoir quel complémentation) a une influence sur la croissance.

Le fichier EX5bis contient les mêmes données que le fichier EX5 mais la colonne R a été remplacée par deux colonnes C et F (1 signifie absence de complémentation, 2 signifie présence de complémentation).

6) Analysez les données avec ce modèle.

Exercice N°6

Des notes cliniques ont été attribuées à des animaux suivis dans le temps (plan en mesures répétées). On veut étudier l'évolution de ces notes en fonction du traitement que ces animaux ont reçu.

Le plan d'expérience utilisé est le suivant :

		J0	J1	J2	J3
Trt1	an1	.	.	.	.
	an2	.	.	.	.
	an3	.	.	.	.
	an4	.	.	.	.
	an5	.	.	.	.
	an6	.	.	.	.
	an7	.	.	.	.

	an8	.	.	.	.
Trt2	an9	.	.	.	.
	an10	.	.	.	.
	an11	.	.	.	.
	an12	.	.	.	.

Chaque point représente une donnée.

1) Décrivez le plan d'expérience utilisé. Y a t-il des facteurs aléatoires, fixes ?

2) Pour un traitement donné, écrivez le modèle qui décrit les variations de la note en fonction du temps et de l'animal.

3) Ecrivez le modèle général qui permet d'analyser les variations de la note en fonction de l'animal, du traitement et du temps.

Le fichier EX6 contient les données relatives à cette expérience. Analysez ces données en utilisant le modèle que vous venez d'écrire.

4) Vérifiez sur les résidus si les 3 postulats fondamentaux sont raisonnables.

5) Les traitements ont ils le même effet ?

Exercice N°7

L'objet de cet exercice est d'apprendre à modéliser les variations d'une réponse à partir de variables quantitatives seulement. L'outil idéal pour cela est la régression multiple.

On veut prévoir la consommation en oxygène d'athlètes qui courent le 2000 mètres.

Le fichier (EX7) contient 7 variables : âge, poids, oxy (consommation en oxygène), temps (temps de la course), fcrep (fréquence cardiaque au repos), fccours (fréquence cardiaque moyenne en course), fcmax (fréquence cardiaque maximale en course).

L'objet de cet exercice est de trouver le meilleur modèle de prévision de la consommation en oxygène. On utilise comme indice le *ADJUST SQUARED R* qui est proportionnel à la variance d'une observation future ainsi que le coefficient TOLERANCE.

Faites une régression ascendante, descendante, en pas à pas.

Dans une régression en pas à pas, on utilise le modèle additif avec toutes les variables et pour toutes les variables, les tests sont réalisés en tenant compte des autres.