Scénario : analyses de variance et covariance élémentaires

Résumé

Analyse de variance et de covariance sont des cas de modèles gaussiens. Trois exemples d'AnOVa à deux facteurs et d'une AnCoVa à trois facteurs sont traités avec SAS.

1 ANOVA à deux facteurs

Une analyse de variance à deux facteurs soulèvent et réponds à une succession de questions classiques ; Sur chacun des jeux de données suivants,

1. caractériser les propriétés du plan (bloc, répétition, équilibre, complétude),
2. lire les données à partir du répertoire `http://wikistat/data`
3. tracer les profils illustrant les interactions,
4. validité du modèle,
5. comparaisons multiples des moyennes.

1.1 Orge

Le premier jeu de données décrit les résultats d'une expérimentation de six fertilisants azotés :
1=(NH4)2 SO2, 2=NH4NO3, 3=CO(NH2)2, 4=CA(NO3)2, 5=NaNO3, 6=Rien,
sur 4 types de sol. Les données sont contenues dans le fichier : orge.dat, avec dans l'ordre : traitement, bloc, rendement.

```
/* Lire les données */
data sasuser.orge;
infile "orge.dat";
input Engrais Sol Rend;
run;
```

Le modèle peut être estimé avec le module SAS/INSIGHT :

```
Solutions > Analyse > Analyse interactive des don.
   Sasuser > ORGE > Open
/* Déclarer Engrais et Sol "nominales" */
Engrais : Int > Nom ; Sol : Int > Nom
/* Estimation du modèle */
Analyse > Fit >  Sélect. Rend en Y
    Sélect. Engrais SOl en X > OK
Graphs > Residual Normal QQ
```

Les mêmes résultats sont obtenus avec le programme :

```
proc anova data=sasuser.orge;
class Engrais Sol;
model Rend=Engrais Sol ;
means Engrais/bon scheffe tukey;
run;
```

Interpréter.

1.2 Carottes

On étudie le temps de germinations de différentes variétés de carottes en fonction du type de sol. Les données sont contenues dans le fichier : carotte.dat, avec dans l'ordre : type de sol, variété de carotte, jour de germination.

```
/* Lire les données */
data sasuser.carotte;
infile "carotte.dat";
input Sol Variete Jour;
run;
```

Estimer le modèle avec le programme ci-dessous ou le module SAS/Insight. Attention, le plan n'est pas équilibré la procédures SAS/ANOVA n'est plus légitime et remplacée par la procédure SAS/GLM et l'interaction doit être prise en compte.

```
proc glm data=sasuser.carotte;
```

```
class Variete Sol;
model Jour=Variete Sol variete*Sol;
run;
```

Interpréter.

1.3 Corrosion

On étudie la résistance a la corrosion de différents tuyaux en fonction de deux facteurs : la nature du sol dans lesquels ils se trouvent et le type de protection (peinture) qu'ils ont reçu. Les données sont dans le fichier`corrosion.dat`, avec dans l'ordre : corrosion, nature du sol, protection.

```
/* Lire les données */
data sasuser.corrosion;
infile "corrosion.dat";
input Corrosion Sol Protection;
run;
```

Estimer le modèle avec le programme ci-dessous ou le module SAS/Insight.

```
proc anova data=sasuser.corrosion;
class Sol Protection;
model Corrosion = Sol Protection Sol*Protection;
run;
```

Interpréter.

2 Analyse de covariance

2.1 Les données

Les données (Jobson, 1991) sont issues d'une étude maketing visant à étudier l'impact de différentes campagnes publicitaires sur les ventes de différents aliments. Un échantillon ou "panel" de familles a été constitué en tenant compte du lieu d'habitation ainsi que de la constitution de la famille. Chaque semaine, chacune de ces familles a rempli un questionnaire décrivant les achats réalisés.

Nous nous limitons ici à l'étude de l'impact sur la consommation de lait de quatre campagnes diffusées sur des chaines locales de télévision. Quatre villes, une par campagne publicitaire, ont été choisies dans cinq différentes régions géographiques. Les consommations en lait par chacune des six familles par ville alors été mesurées (en dollars) après deux mois de campagne.

Les données se présentent sous la forme d'un plan factoriel équilibré croisant trois facteurs : région, taille de la famille, type de campagne. Le nombre insuffisant d'observations (pas de répétition) ne permet pas de considérer l'interaction d'ordre trois taille$\times$régions$\times$pub ; seules les interactions d'ordre 2 seront donc testées.

2.2 Lire et réorganiser les données

Les données ne sont pas ordonnées correctement pour être traitées par une procédure statistique ; un traitement préliminaire est nécessaire.

```
data sasuser.milk;
infile "milk.dat" delimiter="09"x;
/* Lire les données */
input region camp1 camp2 camp3 camp4 taille;
array c{4} camp1-camp4;
/* Ecrire : 1 variable par colonne */
do pub=1 to 4;
consom=c{pub};
output;
end;
drop camp1-camp4;
run;
```

2.3 Analyse de covariance

Deux approches sont possibles selon que la variable : `taille` de la famille, est considére qualitative ou quantitative. Le choix est fait ici de la considérer quantitative car cette option est réaliste compte tenu de la croissance linéaire observée en fonction de la taille (le vérifier) et que d'autre part, cela "économise" des paramètres. Seul un paramètre est associé à la variable quantitative `taille` au lieu de $5 = 6 - 1$ si elle est considérée comme un facteur. Moins de paramètres sont estimés et donc les tests plus puissants.

Approche exploratoire

Suivre les instructions ci-dessous pour explorer les données avec le module SAS/INSIGHT ; représenter le nuage de points croisant consommation et taille, associer une couleur à chaque type de campagne, un symbole à chaque région. Vérifier visuellement la bonne linéarité de la relation conditionnellement aux variables région et pub : à cet effet construire conjointement les colonnes (mosaic) croisant régions et pub, la sélection d'une cellule permet de vérifier le bon alignement des points ainsi sélectionnés sur le graphe.

```
Solutions > Analyse > Analyse interactive des don.
   Sasuser > MILKCC > Open
/* Déclarer region et pub "nominales" */
pub : Int > Nom ; region : Int > Nom
/* Nuage de points Consom * taille */
Analyse > Scatter plot > Sélect. Consom en Y,
   Taille en X
/* Identifier les points  */
/* Couleur par campagne de pub */
Edit > Windows > Tools >
   Arc-en-ciel > pub > OK
/* Caractère par région */
   Caract. > region > OK
/* Ajoute un mosaïc plot */
Analyse > Mosaïc plot > Sélect. region en X,
   pub en Y
```

Il suffit alors de cliquer sur l'une des cases croisant `region X pb` pour apprécier l'alignement raisonnable de toutes les observations partageant ces critères.

Tests

Il reste à vérifier la significativité des différences (pentes, constantes à l'origine) observées. Estimer (avec GLM ou INSIGHT) les paramètres du modèle expliquant la consommation en fonction de la taille par type de campagne et par région. Introduire toutes les interactions sauf celle d'ordre trois.

```
Analyse > Fit > Sélect. consom en Y
   Sélect. region pub taille > Expand
   Sélect. taille*taille dans Y > Remove > OK
/* Validité du modèle */
Graphs > Residual Normal QQ /*Que conclure?*/
```

Analyser les tests de type III pour conclure sur l'effet de la pub sur la consommation et de des interactions.

Autres approches

Remarquer que le programme ci-dessous conduit aux mêmes résultats.

```
proc glm data=sasuser.milkcc;
class pub region;
model consom= taille region pub pub*region
   taille*pub taille*region;
run;
```

Résultats légèrement différents de ceux de ce programme (noter, expliquer les différences) :

```
proc glm data=sasuser.milkcc;
class pub region taille;
model consom= taille region pub pub*region
   taille*pub taille*region;
run;
```

L'erreur commune serait de ne pas prendre en compte la variable `région` qui joue le rôle de variable "Bloc" dans le modèle :

```
proc glm data=sasuser.milkcc;
class pub ;
model consom= taille pub  taille*pub ;
run;
```

Conclure dans ce cas erroné sur l'influence du facteur `pub` !

2.4 Choix de la campagne de pub par région

Le programme ci-dessous construit le facteur obtenu par croisement de region et pub. Il représente conjointement les droites de régression et es-

time, pour chaque cellule `région X pub`, le modèle linéaire associé et dont l'équation apparaît dans la fenêtre `log`.

```
data milk2;
set sasuser.milk;
reg_pub=pub+4*(region-1);
run;
proc gplot;
by region;
symbol i=r v=dot;
plot consom*taille=reg_pub;
run;
```

Choisir la campagne de publicité la mieux adaptée à chaque région.