Scénario : Modèle Linéaire Général avec SAS

Résumé

Ce scénario présente des exemples de Modèle Linéaire Général (GLM) : gaussien, binomial et poissonnien traités en SAS et avec une initiation aux stratégies élémentaire de choix de modèle pour une meilleure prévision. Différents types de modèles sont estimés et comparés en prenant en compte le type des variables explicatives quantitatives ou qualitatives. Plusieurs exemples sont traités pour illustrer les fonctionnalités des procédures du module SAS/Stat.

Tous les fichiers de données sont disponibles dans le répertoire `data` du site `wikistat.fr`.

Remarque : Par habitude et pour ne pas remettre en cause de lourds investissements, SAS reste toujours très utilisé, par exemple en biostatistique (recherche pharmaceutique), pour des modèles classiques d'ANOVA et en marketing (GRC) avec l'utilisation de la régression logistique pour estimer des scores d'appétence (exemple de scénario) ou d'attrition. SAS reste aussi très utilisé pour l'estimation de modèles mixtes ou à effet aléatoire, notamment en biostatistique (INRA), associant modélisation de la moyenne et aussi de la variance. Pas étudiés dans ce tutoriel, ils sont plus complexes et l'utilisation de SAS y est largement préférée à celle de R.

1 Modèle gaussien

Des exemples d'estimation avec SAS de modèles gaussiens par régression linéaire et lorsque toutes les variables explicatives sont quantitatives, sont explicités dans d'autres scénarios. Le jeu de données mélange dans l'exemple traité variables explicatives quantitatives et qualitatives.

Attention : La progression est "pédagogique" : anova à un, deux, trois facteurs, ancova mais celle-ci n'est pas justifiée. La bonne pratique consiste à poser et estimer d'emblée le modèle correspondant effectivement au plan expérimental avec ses éventuelles variables blocs.

1.1 Données

Les données (fichier `milk.dat`), extraites de Jobson (1991), sont issues d'une étude marketing visant à étudier l'impact de 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 taille de la famille. Chaque semaine, chacune de ces familles ont 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 chaînes 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 ont été mesurées (en dollars) après deux mois de campagne publicitaire.

Les données initiales se présentent sous la forme d'un tableau à 6 variables : la région géographique, les 4 consommations pour chacune des villes ou campagnes publicitaires diffusées, la taille de la famille.

Elles sont lues et réorganisées pour satisfaire à la structure classique d'un plan factoriel équilibré croisant trois facteurs orthogonaux : région, taille de la famille, type de campagne. L'objectif est d'étudier l'effet du type de publicité sur la consommation des ménages.

```
data sasuser.milk;
infile "milk.dat" delimiter="09"x;
input region camp1 camp2 camp3 camp4 taille;
run;
data sasuser.milkcc;
set sasuser.milk;
array c{4} camp1-camp4;
do pub=1 to 4;
consom=c{pub};
output;
```

```
end;
drop camp1-camp4;
run;
proc print data=sasuser.milkcc;
run;
```

Étudier la distribution de la variable de consommation seule (diagramme boîte) puis en fonction des facteurs des autres variables (diagrammes boîte parallèle).

1.2 Anova à un facteur

Une première étude s'intéresse à l'effet du simple facteur "type de campagne publicitaire". On suppose implicitement que les familles ont été désignées aléatoirement indépendamment de l'appartenance géographique ou de leur taille. La procédure SAS/ANOVA est utilisée dans le programme suivant. Elle est plus particulièrement adaptée aux situations équilibrées comme c'est le cas pour cet exemple. Le cas déséquilibré ne pose pas de problème majeur pour un modèle à un facteur mais pour deux facteurs ou plus, un message signale que les résultats sont fournis sous la "responsabilité de l'utilisateur" car les sommes de carrés ne possèdent plus les bonnes propriétés de décomposition. Dans ce cas, la procédure plus générale SAS/GLM doit être utilisée.

Anova avec comparaison des moyennes par test multiple.

```
proc anova data=sasuser.milkcc;
class pub;
model consom=pub;
means pub/bon scheffe tukey;
```

Cette procédure signale explicitement que des problèmes peuvent apparaître si certains tests, spécifiques au cas équirépété, sont utilisés hors de leur contexte. Différentes options de présentation des résultats sont proposées : tests avec niveau paramétrable (5% par défaut) de significativité, intervalles de confiance des différences ou des moyennes.

Comparer les résultats des différents tests. Quelles sont les moyennes significativement différentes ? Pour quel test ?

Comparer avec la version non paramétrique de l'Anova. Justifier les différences observées.

```
proc npar1way data=sasuser.milkcc;
class pub;
var consom;
run;
```

1.3 Anova à deux facteurs

Il est toujours instructif de vérifier graphiquement la présence possible d'interactions : profil moyen et profils de la consommation moyenne de chaque région en fonction du type de campagne de publicité.

```
proc means data=sasuser.milkcc mean stderr;
class pub region;
var consom;
output out=cellmoy mean=moycons;
run;
symbol i=join v=dot cv=black ;
symbol2 i=join v=% cv=black h=2;
symbol3 i=join v='"' cv=black h=2;
symbol4 i=join v=# cv=black h=2;
symbol5 i=join v=$ cv=black h=2;
proc gplot data=cellmoy;
plot moycons*region=pub;
run;
goptions reset=all; quit;
```

Conclure à la présence ou non d'interactions.

Dans le cas équirépété, la procédure SAS/ANOVA reste valide mais SAS/GLM, plus générale, est utilisée et fournit dans ce cas les mêmes résultats. Cette procédure adaptée aux situations complexes fournit également d'autres options (contrastes, estimation des paramètres...).

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

Cette interaction semble-t-elle significative par ce test ?

1.4 AnCoVa

Si la variable `taille` est considérée quantitative, c'est un modèle d'analyse de covariance qui est adapté.

Pourquoi est-il utile et pertinent de privilégier cette solution par rapport à une ANOVA avec la variable *taille* qualitative à 6 classes ?

A l'aide des graphes obtenus ci-dessous, **vérifier visuellement la bonne linéarité de la relation consommation x taille conditionnellement aux variables région et pub.**

```
proc gplot data=sasuser.milkcc;
by region;
symbol i=r v=dot;
plot consom*taille=pub;
run;
```

Peut-on affecter à chaque région la campagne de pub "optimale" ?

La question est alors de savoir si les différences observées entre les droites (ordonnées à l'origine, pentes) sont significatives ou pas.

Modèle (trop) simpliste

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

Interpréter les résultats des tests. Quelles sont les influences remarquables ? Ou plutôt l'ansence d'influence remarquable ?

Néanmoins, pris d'un doute, le même calcul est effectué séparément pour chaque région :

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

- **Que dire de l'influence de la pub ?**
- **Quelle conclusion en tirer sur de possibles interactions concernant les effets région et publicité ?**
- **Choisir la campagne de publicité la mieux adaptée à chaque région.**

Modèle complet

Ceci incite donc à se méfier des *interactions* (l'effet région tend à compenser l'effet publicité). La règle qui en découle est de **toujours** conserver le facteur *bloc* (ici la région) dans une analyse de (co)variance. Une approche complète, considérant *a priori* toutes les variables (3 facteurs) et leurs interactions, est ici nécessaire.

```
proc glm data=sasuser.milkcc;
class pub region;
model consom=taille|region|pub @2;
run;
```

Interpréter finalement ces résultats ; influence des différents facteurs, de leurs interactions ; choix de la campagne de publicité.

AnOVa

La variable "taille" est considérée qualitative, facteur à 6 niveaux. Il s'agit alors d'une anova à trois facteurs. Le nombre insuffisant d'observations (pas de répétition) ne permet pas de considérer l'interaction taille×régions×pub d'ordre trois. Seules les interactions d'ordre 2 sont donc testées.

```
/* Modèle de toutes interactions d'orde 2*/
proc anova data=sasuser.milkcc;
class pub region taille;
model consom=taille|pub|region @2;
means pub/bon scheffe tukey;
run;
```

- **Comment interpréter les résultats des tests ?**
- **Quelles différences avec le modèle d'ANCOVA ? Pourquoi ?**
- **Interpréter les tests multiples.**

- **Quelle différence avec le premier test d'ANOVA avec correction de Scheffe ?**

2 Modèle poissonnien vs. logit

2.1 Les données

On s'intéresse aux résultats (Jobson, 1991) (fichier `ceinture.dat`) d'une étude préalable à la législation sur le port de la ceinture de sécurité dans la province d'Alberta à Edmonton au Canada. Un échantillon de 86 769 rapports d'accidents de voitures ont été compulsés afin d'extraire une table de contingence complète croisant :

1. Gravité des blessures : Gr0 : rien à Gr3 : fatales
2. Risque regroupe Gr3 à Gr1 d'un côté et Gr0 de l'autre.
3. Port de la ceinture : Coui/Cnon
4. Sexe du conducteur : Hom/Fem
5. Etat du conducteur : Ajeu /A_bu

```
data sasuser.ceinture;
infile "ceinture.dat";
input grave $ ceinture $ sexe $ alcool $ effectif;
select (grave);
/* Construction d'une variable binaire*/
when("Gr1","Gr2","Gr3") risque="Rimp";
when("Gr0") risque="Rfai";
otherwise;
end;
run;
```

Vérifier les répartitions des classes des variables (barre plot). L'étude des croisements des variables 2 à 2 (mosaïc plot) sont bienvenus.

2.2 Modèle binomial

Plusieurs modélisations sont testées avec les procédures `genmod` et `logistic`.

Compte tenu des effectifs très déséquilibrés des modalités de la variable `risque`, les données ont été simplifiées pour ne considérer que deux états de gravité : aucune blessure ou blessure plus ou moins grave à fatale. Les deux procédures sont exécutées ci-dessous.

- **Quelle est la distribution utilisée dans genmod ?**
- **Quelle est la fonction lien canonique ?**
- **Que dire de la légétimité des tests du chi2 ?**
- **Vérifier que même si les paramètres estimés sont différents (pourquoi ?), les tests de significativité conduisent aux mêmes conclusions.**

```
proc logistic data=sasuser.ceinture;
class sexe alcool ceinture;
model risque=sexe|alcool|ceinture@2 ;
freq effectif;
run;

proc genmod data=sasuser.ceinture;
class sexe alcool ceinture ;
model risque=sexe|alcool|ceinture@2 /type3
dist=bin;
freq effectif;
run;
```

- **Que pensez vous de la présence des interactions ?**
- **Utiliser la procédure logistic pour réduire (procédure *backward*) le modèle : retirer une à une l'interaction la moins significative et s'arrêter quand tous les termes sont significatifs.**

Le modèle ci-dessous excluant les interactions permet d'estimer les rapports de cote ou *odds ratio* ainsi que la courbe ROC.

```
proc logistic data=sasuser.ceinture descending;
class sexe alcool ceinture;
model risque=sexe alcool ceinture/ outroc=roc1;
freq effectif;
run;
```

- **Interpréter l'influence des facteurs sur la gravité de l'accident en utilisant les rapports de cote (*odss ratio*).**

- **Qu'est-ce qui est le plus dangereux ?**
- **Que dire de la qualité d'ajustement du modèle ?**
- **De ses qualités prédictives ?**

2.3 Modèle poissonien

Exécuter le code ci-dessous qui estime un modèle binomial et un modèle poissonnien sur les mêmes données.

```
proc genmod data=sasuser.ceinture;
class sexe alcool ceinture;
model risque = sexe alcool ceinture
          /type3  dist=bin;
freq effectif;
run;
proc genmod data=sasuser.ceinture;
class risque sexe alcool ceinture;
model effectif=sexe alcool ceinture
   risque|sexe risque|alcool risque|ceinture
   /type3  dist=poisson;
freq effectif;
run;
```

- **Quelle est la fonction lien par défaut du modèle de poisson ?**
- **Interpréter les résultats des tests.**
- **Que dire de l'influence des facteurs, des interactions ?**
- **Les conclusions des deux modélisations (poisson, binomial) sont-elles cohérentes ?**
- **Quelle stratégie ou modèle semble le plus pertinent ?**

2.4 Régression polytomique

Compte tenu de la nature de la variable à expliquer qui est qualitative *ordinale*, la première chose à faire serait d'utiliser la procédure suivante qui estime, par défaut, une régression logistique ordinale ou polytomique. Cette approche modélise globalement la probabilité de passer d'un niveau de gravité donné à un niveau plus élevé. Cela revient à estimer autant de modèles que le nombre de modalités moins un.

La validité de ce modèle repose sur une hypothèse dite d'*homogénéité des rapports de cote*. Cela signifie que chacun des modèles ne diffère que par le terme constant, les paramètres ou coefficients des variables sont considérés égaux. Il n'est pas possible sinon de résumer cette situation à travers "un seul" modèle ou plutôt une famille de modèles ne différent que par le terme constant (coefficient β_0).

```
proc logistic data=sasuser.ceinture ;
class sexe alcool ceinture grave;
model grave=sexe alcool ceinture ;
freq effectif;
run;
```

L'hypothèse est-elle vérifiée ?

Ce modèle, lorsque les hypothèses en sont vérifiées, est largement utilisé dans les enquêtes de satisfaction.