Analyse statistique de données qualitatives et quantitatives en sciences sociales : TP

MODÈLE LOG-LINÉAIRE (CHAPITRE 2)

Modèles log-linéaires à réaliser

- Table 2x2 (modèle saturé 4 paramètres)
  - Evaldemo2 (Y) et blgetmg (X)
- Table 2x3 (modèle saturé 6 paramètres)
  - Evaldemo2 (Y) et reg3 (X)

- Modèle saturé : $\log \mu_{ij} = \gamma + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$

avec $\lambda_1^X = \lambda_1^Y = \lambda_{1j}^{XY} = \lambda_{i1}^{XY} = 0$

- Modèle d'indépendance : $\log \mu_{ij} = \gamma + \lambda_i^X + \lambda_j^Y$

- Table 2x2x3 (modèle saturé 12 paramètres)
  - Evaldemo2 (X) , blgetmg (Y), reg3(Z)
  - Evaldemo2 (X) , polintr2 (Y), reg3(Z)
- Table 4x2x3 (modèle saturé 24 paramètres)
  - Evaldemo4 (X) , polintr2 (Y), reg3(Z)

- Modèle saturé:

$$\log \mu_{ijk} = \gamma + \lambda_i^X + (\lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}) + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{ijk}^{XYZ}$$

- Modèle sans interaction du 3ème ordre (retrait de λ_{ijk}^{XYZ})
- Indépendance conditionnelle :
  - $X \perp\!\!\!\perp Z|Y$: $\log \mu_{ijk} = \gamma + \lambda_i^X + (\lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}) + \lambda_{ij}^{XY}$
  - $X \perp\!\!\!\perp Y|Z$: $\log \mu_{ijk} = \gamma + \lambda_i^X + (\lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}) + \lambda_{ik}^{XZ}$
- Modèle d'indépendance entre X et les 2 autres variables :

$$\log \mu_{ijk} = \gamma + \lambda_i^X + (\lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ})$$

Table 2xJ

- Point de départ : table des fréquences

proc freq data = *SAS-dataset1*;

tables X**Y/*nopercent nocol norow

out = *SAS-dataset2*;

run;

- Objectif : déterminer si l'hypothèse d'indépendance est plausible

→ comparer les qualités prédictives du modèle d'indépendance avec le modèle saturé via un test de rapport de vraisemblance

  - Modèle d'indépendance : le nombre de personnes attendues en moyenne dans la cellule (*i,j*) ne dépend que des marges de la table de contingence : $\hat{\mu}_{ij} = n_{++} \times \hat{\pi}_{i+} \times \hat{\pi}_{+j}$
  - Modèle saturé : les fréquences attendues $\hat{\mu}_{ij}$ sont exactement égales aux fréquences observées dans chaque cellules n_{ij}

Table 2xJ

```
title "Modèle d'indépendance";
proc genmod data = SAS-dataset2;      /*NB : SAS-dataset2 = proc freq output : table de contingence*/
        class X Y  /     param = ref   ref = first  ;
        model count = X Y /    dist=poisson    link = log ;
run;
```

Régression log-linéaire

```
title "Modèle saturé";
proc genmod data = SAS-dataset2  ;
        class X Y   /   param = ref    ref = first;
        model count = X   Y   X*Y   /   dist=poisson   link = log   type1   type3;
run;
```

NB : alternative pour le choix des catégories de référence

proc genmod data = *SAS-dataset2* ;

class X (ref = “cat_de_ref”) Y / param = ref <ref = first> ;

model count = X Y X*Y / dist=poisson link = log type1 type3;

run;

Table IxJxK

- Point de départ : table des fréquences

proc freq data = *SAS-dataset1*;

tables Y*Z*X*/*nopercent nocol norow

out = *SAS-dataset2*;

run;

- Sélection du modèle : on part du plus complexe (modèle saturé) et on évalue progressivement le retrait des termes d'interaction vers le modèle le plus simple (indépendance entre X et les 2 autres variables).

!!!! Distinction entre sorties de type 1 et de type 3 pour la sélection du modèle !!!!

'CONTRÔLE' SEQUENTIEL   'CONTRÔLE' DE TOUS LES AUTRES TERMES

Sélection du modèle

1. **Modèle saturé**

proc genmod data = *SAS-dataset2*;
class X Y Z (ref=*'cat_de_ref'*)/param = ref ref = first;
model count = X Y Z **Y*Z** X*Y X*Z X*Y*Z / dist=poisson link = log type1 type3;
run;

*NB : Alternative pour l'écriture de l'équation : count = X **Y |Z** X*Y X*Z X*Y*Z ; (cf. cours théorique)

Si la p-valeur associée à XYZ > 0,05 :

2. **Modèle sans l'interaction du 3ème ordre**

proc genmod data = *SAS-dataset2*;
class X Y Z (ref=*'cat_de_ref'*)/param = ref ref = first;
model count = X Y Z Y*Z X*Y X*Z / dist=poisson link = log type1 type3;
run;

Si la p-valeur associée à XY|Z > 0,05 :

3. Indépendance de X et Y conditionnelle à Z

```
proc genmod data = SAS-dataset2;
        class X Y Z (ref='cat_de_ref')/param = ref ref = first;
        model count = X Y Z  Y*Z X*Z/dist=poisson link = log type1 type3;
run;
```

OU si la p-valeur associée à XZ|Y > 0,05 :

3. Indépendance de X et Z conditionnelle à Y

```
proc genmod data = SAS-dataset2;
        class X Y Z (ref=' cat_de_ref ')/param = ref ref = first;
        model count = X Y Z Y*Z X*Y/dist=poisson link = log type1 type3;
run;
```

Si les p-valeurs associées à XY|Z ET à XZ|Y > 0,05 :

4. Indépendance entre X et les 2 autres variables

```
proc genmod data = SAS-dataset2;
    class X Y Z (ref='cat_de_ref')/param = ref ref = first;
    model count = X Y Z  Y*Z /dist=poisson link = log type1 type3;
run;
```

Après avoir sélectionné le modèle, on examine les coefficients pour le modèle retenu