

Cours de Scoring - Séance 4 (Présentation des procédures SAS pour le Scoring)

Ibrahim TOURE,
Ingénieur Statisticien

Université d'Evry Val d'Essone

12 Janvier 2017

Sommaire

1 Constitution de la base d'analyse

2 Etude exploratoire

3 Transformation des variables

4 Estimation du modèle

5 Analyse des performances

Prise en main de la base d'analyse

```
Proc Contents DATA = TABLE; /* Table en entrée */  
RUN;
```

Détection des doublons

```
PROC SORT DATA = TABLE /* Table en entrée */  
NODUPKEY /* Option pour supprimer les doublons */  
OUT = TABLE2 /* Table en sortie */  
DUPOUT = TABLE3 /* Table en sortie contenant les doublons */;  
BY VAR1 VAR2 ... /* Variables analysées */;  
RUN;
```

Détection des valeurs manquantes (1/2)

```
PROC FREQ DATA = TABLE; /* Table en entrée */  
TABLES VarQualitative1 VarQualitative2 ... / MISSING; /* Variables  
qualitatives */  
RUN;
```

- Le nombre d'observations manquantes s'affichera dans l'output sous la forme d'une modalité spécifique
- Remarque : le critère à modéliser est une variable qualitative

Détection des valeurs manquantes (2/2)

```
PROC MEANS DATA = TABLE /* Table en entrée */ NMISS;  
TABLES VarQuantitative1 VarQuantitative2 ... / MISSING; /* Variables  
qualitatives */  
RUN;  
***
```

- NMISS restitue le nombre d'observations manquantes par variables
- Remarque : il est indispensable d'étudier l'influence des valeurs manquantes sur la distribution du critère à modéliser

Détection des valeurs aberrantes

```
PROC FREQ DATA = TABLE /* Table en entrée */;  
TABLES VarQualitative1 VarQualitative2 ...; /* Variables qualitatives */  
RUN;  
****  
PROC UNIVARIATE DATA = TABLE /* Table en entrée */;  
VAR VarQuantitative1 VarQuantitative2; /* Variables qualitatives */  
HISTOGRAM VarQuantitative1 VarQuantitative2 ... / KERNEL; /*  
Variables quantitatives et édition d'une courbe de densité */ RUN;
```

Data Splitting

```
PROC SURVEYSELECT DATA = TABLE /* Table en entrée */;  
METHOD = SRS /* Simple Random Sampling */  
SEED = NombreQuelconque /* Nombre initial ; par exemple : 63 */  
SAMPRATE = RATE /* Taux d'échantillonnage ; compris entre 65 et 80 */  
*/  
OUT= TABLE1 /* Table en sortie */  
OUTALL /* Conserve toutes les observations en créant une indicatrice */  
; STRATA Critere; /* Critère à modéliser (votre variable y) */  
RUN
```

- Remarque 1 : L'option STRATA permet de réaliser un échantillonnage stratifié proportionnel (sur le critère à modéliser dans notre cas). Sans elle, il s'agira d'un échantillonnage simple sans remise.
- Remarque 2 : il convient, au préalable, de trier la table en entrée selon la ou les variables de stratification
- Remarque 3 : il est indispensable d'étudier l'influence de l'échantillon sur la distribution des variables (dont le critère)

Sommaire

1 Constitution de la base d'analyse

2 Etude exploratoire

3 Transformation des variables

4 Estimation du modèle

5 Analyse des performances

Liaison Var. quantitative * Var. quantitative

```
PROC CORR DATA = TABLE /* Table en entrée */
PEARSON /* Coefficient de corrélation linéaire */
SPEARMAN; /* Coefficient de corrélation des rangs de Spearman */
VAR VarQuantitative1 VarQuantitative2 ...; /* Variables quantitatives */
*/
RUN;
```

Il est bon de rappeler que la valeur de ces coefficients est comprise entre -1 et 1.

Liaison Var. qualitative * Var. qualitative

```
PROC FREQ DATA = TABLE /* Table en entrée */  
TABLE VarQuali1 * VarQuali2 / CHISQ; /* Variables qualitatives 1 et 2  
*/  
OUTPUT OUT = TABLE2 CHISQ; /* Table en sortie */  
RUN;
```

- Le Coefficient V de Cramer est directement disponible
- Récupérer la statistique du Khi-2, le nombre d'observations, le nombre de modalités de la variable qualitative 1 et 2, pour calculer le coefficient T de Tschuprow compris entre 0 et 1

Liaison Var. qualitative * Var. quantitative

Il s'agit d'un test **non paramétrique** : test de WILCOXON

```
PROC NPAR1WAY WILCOXON DATA = TABLE /* Table en entrée
*/ CORRECT = NO ;
CLASS VarQualitative; /* Variable qualitative */
VAR VarQuantitative; /* Variable quantitative */
RUN ;
***
```

Règle de décision :

Si la P-Value est inférieure à 0.05, on rejette l'hypothèse nulle.

Autrement dit, une différence significative existe entre les groupes de la variable qualitative.

Sommaire

1 Constitution de la base d'analyse

2 Etude exploratoire

3 Transformation des variables

4 Estimation du modèle

5 Analyse des performances

Discrétisation des variables quantitatives (approche graphique)

La PROC KDE est à effectuer par variables quantitatives.

```
PROC KDE WILCOXON DATA = TABLE /* Table en entrée */  
CORRECT = NO;  
UNIVAR VarQuantitative; /* Variable quantitative */  
/ OUT = TableOutKDE /* Table en sortie */  
BY Critere; /* Critère à modéliser */  
RUN;  
***
```

```
PROC GPLOT DATA = TableGraph /* Table en entrée */;  
PLOT (DENSITY-0 DENSITY-1) * VALUE / OVERLAY; RUN;
```

NB : La table « TableGraph »est une table issue de la concaténation de deux tables créées à partir de « TableOutKDE »(une table pour chaque modalité du critère).

```
RUN;
```

Sommaire

1 Constitution de la base d'analyse

2 Etude exploratoire

3 Transformation des variables

4 Estimation du modèle

5 Analyse des performances

Méthodologie de modélisation et analyse des résultats

```
PROC LOGISTIC DATA = TABLE /* Table en entrée */;  
CLASS VarQualitativeExplicative /* Variables explicatives qualitatives */  
/ PARAM = REF; /* Contrainte d'identifiabilité */  
MODEL CRITERE (EVENT ="1") = VarExplicative1 VarExplicative 2  
... /* Critère à modéliser (0/1) ; Var explicative qualitative et  
quantitative */;  
/ LINK = LOGIT /* Fonction de lien */ SELECTION = BACKWARD  
SLS = 0.05 /* Selection automatique */  
RSQUARE /*R2 */;  
RUN;
```

- Remarque 1 : EVENT = "1" : 1 correspond à un client sain
- Remarque 2 : privilégier une approche "manuelle" de sélection des variables

Méthodologie de modélisation et analyse des résultats

Contribution d'échelle (en %) : participation de la variable explicative à l'étendue des valeurs du score

$$C^j = \frac{\text{Max}\{x_1^j, \dots, x_p^j, 0\} - \text{Min}\{x_1^j, \dots, x_p^j, 0\}}{\sum_{j=1}^k \text{Max}\{x_1^j, \dots, x_p^j, 0\} - \text{Min}\{x_1^j, \dots, x_p^j\}} \quad (1)$$

Sommaire

- 1 Constitution de la base d'analyse
- 2 Etude exploratoire
- 3 Transformation des variables
- 4 Estimation du modèle
- 5 Analyse des performances

Table pour l'analyse des performances

```
PROC LOGISTIC DATA = TABLE /* Table en entrée */;  
CLASS VarQualitativeExplicative /* Variables explicatives qualitatives *//  
/ PARAM = REF; /* Contrainte d'identifiabilité */  
MODEL CRITERE (EVENT = "1") = VarExplicative1 VarExplicative 2  
... /* Critère à modéliser (0/1) ; Var explicative qualitative et  
quantitative */;  
OUTPUT OUT = TABLE2 /* Rable en sortie */  
PREDICTED = YChapeau XBETA = YEtoileChapea;  
WEIGHT SELECTED; /* Astuce pour évaluer la performance de  
l'échantillon test */  
RUN;
```

- Remarque 1 : EVENT = "1" : 1 correspond à un client sain
- Remarque 2 : privilégier une approche "manuelle" de sélection des variables

Courbe ROC et indice C

```
PROC LOGISTIC DATA = TABLE /* Table en entrée */;  
CLASS VarQualitativeExplicative /* Variables explicatives qualitatives *//  
/ PARAM = REF; /* Contrainte d'identifiabilité */  
MODEL CRITERE (EVENT ="1") = VarExplicative1 VarExplicative 2  
... /* Critère à modéliser (0/1); Var explicative qualitative et  
quantitative */;  
OUTROC = TableOUTRoc /* Table sortie données courbe ROC *//  
RUN;  
PROC GPLOT DATA = TableOUTRoc; /* Table en entrée */ PLOT  
(un)SENSIT(un)* ((un)1MSPEC(un) (un)SENSIT(un)) / OVERLAY;  
RUN; QUIT;
```

- Remarque 1 : l'indice C est l'indice correspondant à la courbe ROC
- Remarque 2 : D de Somer = $(2*C) - 1$