Projet de MASTER 2 d'analyse statistique de données réelles

Guillaume SAINT PIERRE
IFSTTAR/COSYS/LIVIC
Institut français des sciences et technologies des transports, de l'aménagement et des réseaux
Département Composants & Systèmes
Laboratoire sur les Interactions Véhicules-Infrastructure-Conducteurs

23 mars 2016

Table des matières

1 Préambule :

Ce projet ne peut pas être réalisé en binôme et un rapport doit être rendu à l'enseignant par chaque étudiant. L'objectif de ce projet est de vous initier à toutes les étapes nécessaires à la production d'un rapport d'analyse de données réelles : du nettoyage et de l'importation des données brutes jusqu'à leur traitement statistique et leur présentation formelle.

1.1 Critères d'évaluation

Le barème de notation est à peu près de 1 point par question. En outre, 1 point sera réservé à l'appréciation générale des commentaires (pertinence de ceux ci par rapport au sujet), ainsi qu'à la présentation et la qualité des tableaux et graphiques contenus dans le rapport. Des points peuvent être enlevés en cas de copie trop évidente d'une page web ou d'une autre copie. La note peut donc être dans un premier temps supérieure à 20, mais est ensuite recalée entre 0 et 20 après réception et notation de toutes les copies.
La totalité des documents sera accessible sur le site http ://perso.lcpc.fr/guillaume.saint-pierre/Master2.html.

1.2 Présentation du rapport

Une attention particulière sera portée à la présentation du rapport. La rédaction du rapport obéira aux règles classiques : page de garde, table des matières, introduction, rédaction, conclusion. Il ne devra pas dépasser 30 pages et pourra comporter 10 pages d'annexes au maximum (soit 40 pages max au total). Le corps du texte comportera quelques sorties numériques et/ou graphiques obtenues avec SAS judicieusement choisies. Quelques précisions sur les sorties pourront être données en annexes. Sur le rapport devront figurer une adresse électronique et un numéro de téléphone où l'on peut joindre les auteurs.

1.3 Remise du projet

La remise de ce sujet est fixée au **vendredi 22 avril 2016**. Elle se fera par un envoi de courrier électronique avec accusé de réception à l'adresse guillaume.saintpierre@ifsttar.fr. Le projet en lui-même consistera en un document au format pdf, ps, ou word, auquel sera joint le code sas (fichier séparé et compressé, afin de pouvoir en vérifier la bonne exécution).

2 Tests d'association pour tables de contingence

Cette première partie s'intéresse aux tests d'association pour les tables de contingences de type 2×2.

Q. 1 *Décrire la formulation mathématique du test du Chi-deux de Pearson utilisé comme une mesure d'association dans les tables de contingence* 2×2. *Donner ses conditions d'application et ses limites.*

Q. 2 *Appliquer le test du Chi-deux de Pearson à la table 1 et décrire les résultats (i.e. quelle conclusion tirer de ces tests).*

Q. 3 *Décrire la formulation mathématique et le cadre d'application du test exact de Fisher.*

Q. 4 *Appliquer le test exact de fisher à la table 2.*

Q. 5 *Dire comment (quelle option de proc freq) l'on peut aussi calculer un test d'association du Chi-deux exact sur la table 2, et le mettre en application (expliquer les résultats).*

Q. 6 *Expliquer quelle erreur a été évitée en utilisant la valeur exacte du test plutôt que son équivalent asymptotique.*

Traitement	Favorable	Défavorable	Total
Placebo	16	48	64
Test	40	20	60

TABLE 1 – Table pour test du chi-deux de Pearson

Traitement	Favorable	Défavorable	Total
Test	10	2	12
Contrôle	2	4	6
Total	12	6	18

TABLE 2 – Table pour test de fisher et test exact du chi-deux

3 Analyse de la covariance

3.1 Consommation de lait

Le fichier `http://perso.lcpc.fr/guillaume.saint-pierre/Enseignement/Projet/milk.dat` contient les résultats d'une étude marketing de l'influence d'une campagne publicitaire sur la consommation de lait chez des familles américaines (Jobson, 1991). Dans 5 régions des Etats-Unis, les consommations de 6 familles ont été étudiées pour 4 types de campagnes publicitaires. Dans chaque région, les familles sont différenciées par leur taille. La première colonne du fichier milk.dat correspond à la `region`, les colonnes 2 à 5 donnent la consommation de lait (en $) pour chaque campagne publicitaire `camp1, camp2, camp3, camp4`, et la 6ème colonne donne la `taille` de chaque famille.

Q. 7 *Créer une table* `milk` *contenant les observations du fichier de données. Réorganiser les données afin de créer une variable* `pub` *décrivant le type de campagne publicitaire et une variable* `consom` *indiquant la consommation de lait.*

Pour cette question, on pourra créer une table `milkcc` contenant les variables `pub` et `milkcc` de la façon suivante

```
data=milkcc;
set milk;
array c{4} camp1-camp4;
do pub=1 to 4;
consom=c{pub}
output;
end;
drop camp1-camp4;
run;
```

Q. 8 *Étudier l'influence de chaque facteur (région, campagne publicitaire et taille des familles) sur la consommation de lait à l'aide d'une analyse de variance à un facteur. Que constatez-vous ?*

Q. 9 *Analyser l'influence et l'interaction des facteurs région et campagne publicitaire sur la consommation de lait. Que constatez-vous ?*

Q. 10 *Représenter la consommation de lait en fonction de la taille de la famille pour chaque type de campagne publicitaire. Que suggèrent ces graphiques ?*

La variable `taille` est une variable quantitative dont on souhaite étudier l'influence sur la consommation de lait à l'aide d'une analyse de la covariance. Sous SAS, la procédure `GLM` permet de mettre en oeuvre un modèle d'analyse de la covariance à l'aide de la syntaxe suivante

```
proc GLM data=tab_sas;
class T;
model Y = X T X*T;
run;
```

où Y est la variable quantitative à expliquer, T est une variable explicative qualitative et X une variable explicative quantitative.

Q. 11 *Ajuster un modèle d'analyse de covariance pour juger de l'influence et de l'interaction des variables* `taille` *et* `pub` *sur la consommation de lait. Comment interprète-on ces résultats ? Que suggèrent-ils ?*

Q. 12 *Représenter la consommation de lait en fonction de la taille de la famille pour chaque type de région. Que suggèrent ces graphiques ?*

Q. 13 *Pour chaque région, ajuster un modèle d'analyse de covariance pour juger de l'influence et de l'interaction des variables* `taille` *et* `pub` *sur la consommation de lait. Que suggèrent ces résultats ? Expliquer les différences avec les résultats obtenus dans la question 11.*

4 Utilisation des méthodes factorielles

L'objectif de cette section est l'utilisation des méthodes factorielles pour établir une classification et une caractérisation d'individus par rapport à la pollution de l'air intérieur d'un échantillon de 567 logements représentatifs du parc des résidences principales en France Métropolitaine. Les données sont disponibles à l'adresse suivante : `http://perso.lcpc.fr/guillaume.saint-pierre/Enseignement/Projet/Donn` `methodes factorielles.zip` et doivent être décompressées puis importées dans une librairie permanente du logiciel SAS.

4.1 Contexte et objectifs

L'Observatoire de la Qualité de l'Air Intérieur (OQAI) a engagé une campagne nationale dans les logements sur un échantillon de 567 logements représentatif du parc des 24 millions de résidences principales de la France continentale métropolitaine. Cette campagne vise à dresser un état de la pollution de l'air dans l'habitat afin de donner les éléments utiles pour l'estimation de l'exposition des populations, la quantification et la hiérarchisation des risques sanitaires associés, l'identification des facteurs prédictifs de la qualité de l'air intérieur.

Au cours de cette campagne, plus de 30 paramètres (chimiques, biologiques, physiques) de pollution ont été mesurés, sur une durée d'une semaine, à plusieurs emplacements à l'intérieur des logements, dans les garages attenants lorsqu'ils existent et à l'extérieur. Dans le même temps des informations détaillées ont été collectées sur les caractéristiques techniques des logements, sur leur environnement ainsi que sur les ménages et leurs activités au travers d'un questionnaire.

L'objectif de ce projet est de définir une typologie des logements enquêtés en tenant compte de l'ensemble des paramètres relevés sur les logements : type, structure, taille, ancienneté, mais également revêtements intérieurs, aménagements..., les occupants et leurs habitudes. Cette typologie a une double visée : elle permet d'une part une description synthétique de l'échantillon des logements enquêtés et peut d'autre part être utilisé comme facteur explicatif dans l'étape de recherche des déterminants des niveaux de pollution intérieurs.

4.2 Structure des bases de données

Les données recueillies au cours de cette campagne logement sont structurées par blocs mixtes (variables à la fois quantitatives et qualitatives) suivant 4 critères (Bloc des logements, bloc des Ménages, Bloc des Habitudes, Bloc des Polluants).

- Le bloc des logements est composé de 71 variables (39 qualitatives et 32 quantitatives). Ces variables décrivent la structure technique des bâtiments à savoir les matériaux entrant dans la construction des logements (les murs en briques, en béton,...) ; le revêtement ; l'aménagement intérieur ...
- Le bloc des ménages regroupe un ensemble de 11 variables (5 qualitatives et 6 quantitatives) caractérisant la structure du ménage (nombre de personnes vivants des les logements, en couple ou vivant seul...).

- Le bloc des habitudes regroupe un ensemble de 44 variables (23 qualitatives et 21 quantitatives) de caractérisant les activités des personnes vivant à l'intérieur du logement.
- Le bloc des polluants regroupe les concentrations des 14 polluants mesurés dans les 567 logements en France métropolitaine.

4.3 Traitement préalable des données

Afin de vous familiariser avec les différentes tables, il est nécessaire de connaître la structure (distribution) des variables.

Q. 14 *Faire une analyse univariée des différentes variables pour chacune des tables (Voir les PROC FREQ et UNIVARIATE sous SAS). Si la variabilité des variables vous semble trop grande pensez à standardiser les données.*

Q. 15 *Créer une table TQUANT qui concatène l'ensemble des variables quantitatives contenues dans chacune des tables Logement, Menage, Habitude.*

Q. 16 *Créer une table TQUAL qui concatène l'ensemble des variables qualitatives contenues dans chacune des tables Logement, Menage, Habitude.*

4.4 Etude des variables Quantitatives

L'ACP (Analyse en Composantes Principale) est une analyse multivariée qui :
- Cherche à identifier les axes principaux qui expliquent le mieux des corrélations entre variables descriptives.
- Recherche des vecteurs propres (combinaison linéaire des variables descriptives) d'une matrice de dispersion (covariance) ou de corrélation.
- Préserve les distances euclidiennes entre les objets.

Q. 17 *Sur la table TQUANT appliquer une ACP en utilisant la procédure PROC PRINCOMP. Représenter graphiquement dans le cercle des corrélations une projection des variables et Interpréter les résultats de l'ACP en faisant ressortir les liens (variables les plus fortement corrélées) entre les variables des blocs logement, ménage et habitude. Dans le rapport faire apparaitre le cercle des corrélations.*

Q. 18 *Créer une table TQUAL qui concatène l'ensemble des variables qualitatives contenues dans chacune des tables Logement, Menage, Habitude.*

4.5 Etude des variables Qualitatives

L'Analyse des Correspondances Multiples (ACM) est une méthode qui permet d'étudier l'association entre au moins deux variables qualitatives. Elle est aux variables qualitatives ce que l'ACP est aux variables quantitatives. Elle permet en effet d'aboutir à des cartes de représentation sur lesquelles on peut visuellement observer les proximités entre les catégories des variables qualitatives et les observations.

La construction du tableau disjonctif complet est l'une des étapes préalables au calcul de l'Analyse des Correspondances Multiples. Les p variables qualitatives sont éclatées en p tableaux disjonctifs Z1, Z2, ..., Zp, composés d'autant de colonnes qu'il y a de modalités pour chacune des variables. A chaque fois qu'une modalité m de la j ième variable correspond à un individu i, on affecte 1 à Zj(i,m). Les autres valeurs de Zj sont nulles. Les p tableaux disjonctifs sont alors concaténés en un tableau disjonctif complet.
A partir du tableau disjonctif complet sont calculées les coordonnées des modalités des variables qualitatives, ainsi que les coordonnées des observations dans un espace de représentation optimal pour le critère d'inertie.
Cette méthode sera utilisée pour déterminer les relations existantes entre les modalités des variables qualitatives.

Q. 19 *Sur la table TQUAL appliquer une ACM en utilisant la procédure PROC CORRESP. Représenter graphiquement sur le premier plan factoriel la projection des modalités variables. Interpréter les résultats de l'ACM en faisant ressortir les liens (variables les plus fortement corrélées) entre les variables qualitatives des blocs logement, ménage et habitude.*

Q. 20 *Stocker dans une table TQUALACM les axes factoriels issus de l'ACM.*

4.6 Classification des individus

Q. 21 *Concaténer les tables TQUALACM et TQUANTACP dans une nouvelle table TCLUST.*

Q. 22 *Sur la table TCLUST, utiliser la procédure PROC FASTCLUS pour réaliser une Classification Ascendante Hiérarchique (CAH) sur les données. Définir le nombre de classes retenues en spécifiant le critère utilisé dans la CAH. Caractérisées les classes obtenues dans la CAH par rapport aux variables des bases Logements, Ménage et Habitudes.*

Q. 23 *Croiser les classes de la CAH avec la liste des polluants prioritaires de la table polluant (POL) en spécifiant les classes plus où moins polluées par rapport à certains polluants de l'air intérieur. (Spécifier dans le rapport les tests utilisés pour la comparaison des proportions dans vos différentes classes)*

La compréhension de la théorie sous jacente aux méthodes ACP, ACM et CAH est nécessaire à la compréhension et à l'interprétation des résultats.

Bon Projet et Bon Courage à tous !!