SY09 Printemps 2015

TP 1
Statistique descriptive, Analyse en composantes principales

1 Statistique descriptive

1.1 Données babies

Données

Le jeu de données considéré babies23.data est constitué de 1236 bébés décrits par 23 variables. Charger le jeu de données et sélectionner 8 variables en utilisant le code R qui suit :

```
babies<-read.table("babies23.txt",header=T)
babies<-babies[c(7,5,8,10,12,13,21,11)]
names(babies)<-c("bwt","gestation","parity","age","height","weight","smoke","education")
```

Remplacer les codes des données non disponibles par NA (*not available*) :

```
babies[babies$bwt == 999, 1] <- NA
babies[babies$gestation == 999, 2] <- NA
babies[babies$age == 99, 4] <- NA
babies[babies$height == 99, 5] <- NA
babies[babies$weight == 999, 6] <- NA
babies[babies$smoke == 9, 7] <- NA
babies[babies$education == 9, 8] <- NA
```

Enfin, déclarer les variables qualitatives comme facteurs :

```
babies$smoke<-factor(c("NonSmoking","Smoking","NonSmoking","NonSmoking")[babies$smoke+1])
babies$education<-factor(babies$education,ordered=T)
```

Les variables disponibles deviennent alors :

1. bwt : le poids de naissance (birth weight) en onces,
2. gestation, la durée de la gestation en jours,
3. parity : le nombre de grossesses précédentes,
4. age : l'âge de la mère à la fin de la grossesse,
5. height : la taille de la mère en pouces,
6. weight : le poids de la mère en livres,
7. smoke : est-ce-que la mère fume ? (0 : never ; 1 : smokes now ; 2 : until current pregnancy ; 3 : once did, not now),
8. ed : le niveau d'éducation de la mère (0 : less than 8th grade ; 1 : 8th to 12th grade - did not graduate ; 2 : High School graduate, no other schooling ; 3 : High School + trade ; 4 : High School + some college ; 5 : College graduate ; 6 and 7 : Trade school, HS unclear).

Questions

Effectuer une analyse exploratoire des données. On cherchera en particulier à analyser les liens entre tabagisme et niveau d'étude, poids du nouveau-né, ou encore temps de gestation.
Pour répondre à ces questions, vous pouvez vous inspirer des suggestions suivantes :
- faire un résumé numérique des variables en distinguant les bébés nés de femmes qui fumaient ou non durant leur grossesse,
- utiliser des méthodes graphiques pour comparer dans ces deux cas les distributions des poids et des temps de gestation (il est recommandé de choisir une échelle commune),
- évaluer la significativité des différences observées.

Résumer les investigations concernant les données. Inclure les sorties graphiques les plus pertinentes. Rapprocher les trouvailles des observations réalisées lors d'études préalables.

Annexe : extrait de l'édition du New York Times datée du 1er mars 1995

Infant deaths tied to premature births
Low weights not solely to blame
A new study of more than 7.5 million births has challenged the assumption that low birth weights per se are the cause of the high infant mortality rate in the United States. Rather, the new findings indicate, prematurity is the principal culprit.
Being born too soon, rather than too small, is the main underlying cause of stillbirth and infant deaths within four weeks of birth.
Each year in the United States about 31,000 fetuses die before delivery and 22,000 newborns die during the first 27 days of life.
The United States has a higher infant mortality rate than those in 19 other countries, and this poor standing has long been attributed mainly to the large number of babies born too small, including a large proportion who are born "small for date", or weighing less than they should for the length of time they were in the womb. The researchers found that American-born babies, on the average, weigh less than babies born in Norway, even when the length of pregnancy is the same. But for a given length of pregnancy, the lighter American babies are no more likely to die than are the slightly heavier Norwegian babies.
The researchers, directed Dr. Allen Wilcox of the National Institute of Environmental Health Sciences in Research Triangle Park, N.C., concluded that improving the nation's infant mortality rate would depend on preventing preterm births, not on increasing the average weight of newborns. Furthermore, he cited an earlier study in which he compared survival rates among low-birthweight babies of women who smoked during the pregnancy.
Ounce for ounce, he said, "the babies of smoking mother had a higher survival rate". As he explained this paradoxical finding although smoking interferes with weight gain, it does not shorten pregnancy.

1.2 Données crabs

Données

Le jeu de données considéré, disponible dans la bibliothèque de fonctions MASS, est constitué de 200 crabes décrits par huit variables (trois variables qualitatives, et cinq quantitatives). Charger le jeu de données et sélectionner les variables quantitatives en utilisant le code R suivant :

```
> library(MASS)
> data(crabs)
> crabsquant <- crabs[,4:8]
```

Questions

1. Effectuer dans un premier temps une analyse descriptive des données. Existe-t-il des différences de caractéristiques morphologiques selon l'espèce ou le sexe ? Semble-t-il possible d'identifier l'espèce ou le sexe d'un crabe à partir d'une ou plusieurs mesures de ces caractéristiques ?
2. Dans un second temps, on étudiera la corrélation entre les différentes variables. Quelle en est vraisemblablement la cause ? Quel traitement est-il possible d'appliquer aux données pour s'affranchir de ce phénomène ?

2 Analyse en composantes principales

2.1 Exercice théorique

Trois variables mesurées sur quatre individus fournissent le tableau suivant

$$\begin{pmatrix} 3 & 4 & 3 \\ 1 & 4 & 3 \\ 2 & 3 & 6 \\ 2 & 1 & 2 \end{pmatrix}.$$

On associe les mêmes pondérations à tous les individus et on munit $\mathbb{R}^p$ de la métrique identité.

1. Calculer les axes factoriels de l'ACP du nuage ainsi défini. Quels sont les pourcentages d'inertie expliquée par chacun de ces axes.
2. Calculer les composantes principales ; en déduire la représentation des quatre individus dans le premier plan factoriel.
3. Tracer la représentation des trois variables dans le premier plan factoriel.
4. Calculer l'expression $\sum_{\alpha=1}^{k} \mathbf{c}_\alpha \mathbf{u}'_\alpha$ pour les valeurs $k = 1, 2$ et 3.

2.2 Utilisation des outils R

L'objectif de cet exercice est de se familiariser avec les fonctions R permettant d'effectuer une ACP, en particulier les fonctions `princomp`, `summary`, `loadings`, `plot` et `biplot`. Remarquons qu'il existe une autre fonction `prcomp` qui effectue les calculs de manière différente ; on ne l'utilisera pas ici.

– En utilisant ces fonctions, effectuer l'ACP du jeu de données notes étudiées en cours. Montrer comment on peut retrouver tous les résultats alors obtenus (valeurs propres, axes principaux, composantes principales, représentations graphiques, ...).
– On s'intéresse à l'affichage des résultats de la fonction `princomp`. Qu'affichent les fonctions `plot` et `biplot` ? Détailler plus particulièrement le fonctionnement de la fonction `biplot` redéfinie pour la classe `princomp` (accessible par `biplot.princomp`) et de ses différentes options.

2.3 Traitement des données Crabs

Données

Comme dans l'exercice 1, on s'intéressera aux données crabs, et plus particulièrement aux descripteurs quantitatifs. On commencera donc par charger les données et sélectionner les variables quantitatives en utilisant le code R suivant :

```
> library(MASS)
> data(crabs)
> crabsquant<-crabs[,4:8]
```

Questions

Cette étude vise à utiliser l'ACP pour trouver une représentation des crabes qui permettent de distinguer visuellement différents groupes, liés à l'espèce et au sexe.

1. Tester tout d'abord l'ACP sur `crabsquant` sans traitement préalable. Que constatez vous ?
2. Trouver une solution pour améliorer la qualité de votre représentation en termes de visualisation des différents groupes.