Examen du 8 février 2011

Durée : 2 heures. Tous les documents ainsi que les calculatrices sont autorisés.

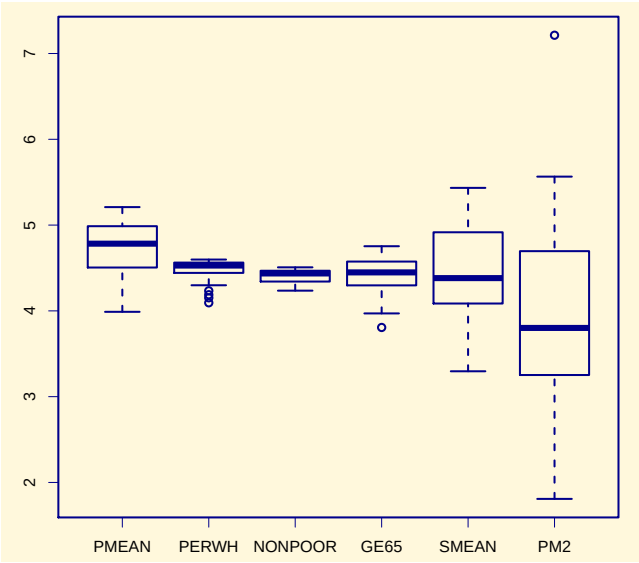
Partie I : Analyse des données multivariées

La pollution de l'eau, de l'air, ... est un des problèmes les plus importants dans le domaine de l'environnement. De nombreuses études relatives à ce type de problème font appel à la Statistique et permettent de répondre à différentes questions sensibles telles que : «Est-ce que la pollution a un impact sur le taux de mortalité ?», «Peut-on construire un indicateur de pollution ?», ou encore «Y a t-il des lieux qui se comportent différemment face à la pollution ?». Pour cela, sur un échantillon de 40 villes des Etats-Unis en 1960, 7 mesures ont été relevées, en plus du taux de mortalité :

TMR : (nombre de décès pour 10000 durant un an)
GE65 : pourcentage ($\times 10$) de la population des 65 ans et plus,
NONPOOR : pourcentage de ménages avec un revenu au dessus du seuil de pauvreté,
PERWH : pourcentage de population blanche,
PMEAN : moyenne arithmétique des relevés réalisés deux fois par semaine de particules suspendues dans l'air ($\mu_g/m^3 \times 10$),
SMEAN : moyenne arithmétique des relevés réalisés deux fois par semaine de sulfate ($\mu_g/m^3 \times 10$),
PM2 : densité de population par mile carré ($\times 0.1$).

1. On commence par effectuer une **Analyse en Composantes Principales** sur l'ensemble des variables dont on exclue le taux de mortalité.

   (a) Au vu des boxplots des variables tracés ci-dessous (sur une échelle logarithmique), qu'est ce qui vous semble plus légitime : une ACP normée ou non normée ? Justifiez votre réponse.

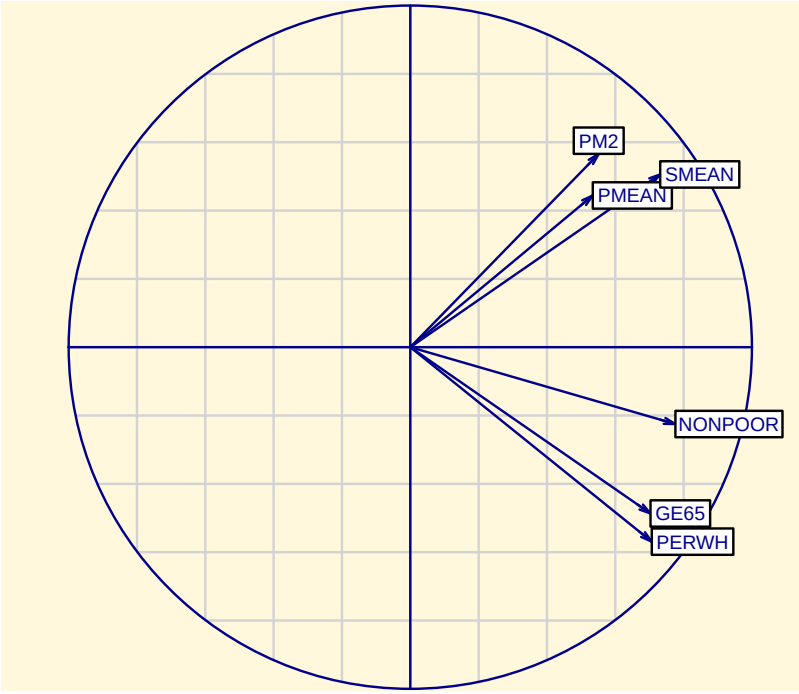


(b) Lorsqu'on effectue une ACP en utilisant la matrice des corrélations, les 6 composantes principales obtenues ont les variances suivantes :

$$\lambda_1 = 2.73, \quad \lambda_2 = 1.38, \quad \lambda_3 = 0.79, \quad \lambda_4 = 0.49, \quad \lambda_5 = 0.35, \quad \lambda_6 = 0.26.$$

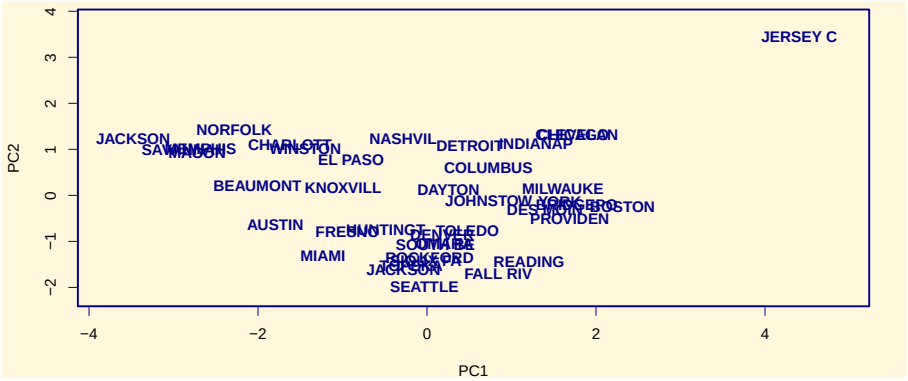
Quelle est la part de l'inertie expliquée par les deux premières composantes principales ?

(c) Lorsqu'on trace le disque des corrélations correspondantes aux 2 premières Composantes Principales (CP), on obtient le résultat suivant :



- Quelles sont les deux variables les mieux représentées par les 2 premières CP ?
- Y a t-il des variables fortement positivement corrélées ?
- Quelles sont les deux variables les plus faiblement corrélées ?

(d) Selon le graph suivant représentant la projection des individus sur le premier plan factoriel, les données contiennent-elles des observations aberrantes ?



2. On effectue maintenant une régression linéaire multiple en considérant le taux de mortalité comme variable à expliquer. La sortie `R` de la commande `lm` est donnée ci-dessous.

```
Call: lm(formula = TMR ~ PMEAN + PERWH + NONPOOR + GE65 + SMEAN + PM2)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 638.4144   134.6606   4.741 3.95e-05 ***
PMEAN         0.6110     0.3474   1.759  0.08790 .
PERWH        -5.1584     1.5333  -3.364  0.00196 **
NONPOOR      -0.2792     2.2423  -0.125  0.90166                          (MODELE 1)
GE65          7.4588     0.7655   9.744 3.09e-11 ***
SMEAN         0.6507     0.2553   2.549  0.01565 *
PM2          -0.1909     0.2434  -0.784  0.43850
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 57.67 on 33 degrees of freedom
Multiple R-squared: 0.8181,     Adjusted R-squared: 0.785
```

(a) Quelles sont les valeurs de l'estimateur des moindres carrés et de l'estimateur de la variance des erreurs dans le modèle :

$$\texttt{TMR} = \beta + \alpha_1 \times \texttt{PMEAN} + \alpha_2 \times \texttt{PERWH} + \alpha_3 \times \texttt{NONPOOR} + \alpha_4 \times \texttt{GE65} + \alpha_5 \times \texttt{SMEAN} + \alpha_6 \times \texttt{PM2} + \varepsilon ?$$

(b) Etant donné la faible valeur de l'estimateur de α_6, on se demande si la variable PM2 est utile pour exoliquer le taux de mortalité. D'après la p-valeur du test de Student calculée par la commande `lm` c-dessus, peut-on accepter, au seuil de 5%, l'hypothèse que la variable PM2 est inutile ?

(c) Si l'on devait supprimer une variable explicative afin de simplifier le modèle, laquelle des 6 variables explicatives supprimeriez-vous ? Argumentez la réponse.

3. Dans le but d'avoir un modèle plus simple, on supprime les variables PMEAN, NONPOOR et PM2. On effectue une nouvelle régression linéaire avec les trois variables explicatives restantes. La sortie de la commande `lm` est présentée ci-dessous.

```
Call: lm(formula = TMR ~ SMEAN + PERWH + GE65)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 637.2325    78.9098   8.075 1.35e-09 ***
SMEAN         0.6974     0.1934   3.605 0.000937 ***
PERWH        -4.4106     1.1817  -3.733 0.000653 ***                   (MODELE 2)
GE65          7.0718     0.7396   9.562 2.03e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 58.04 on 36 degrees of freedom
Multiple R-squared: 0.7989,     Adjusted R-squared: 0.7822
```

(a) Indiquer la statistique, ainsi que ces valeurs pour les deux modèles explicatifs utilisés (Modèles 1 et 2), qui suggère que le Modèle 2 malgré sa simplicité a une qualité explicative comparable à celle du Modèle 1.

(b) Y a t-il des variables inutiles dans le Modèle 2 ?

(c) Préciser la valeur de l'estimateur du coefficient de SMEAN. Peut-on déduire de ce résultat que le taux de mortalité élevé dans certaines villes est dû, en partie, à la pollution.

(d) Comment interprétez-vous le signe de l'estimateur du coefficient de PERWH ?

Partie II : Statistique Paramétrique

À la fin du XIXe siècle, l'économiste italien Vilfredo Pareto analyse les données fiscales de l'Angleterre, la Russie, la France, la Suisse et l'Italie. Il remarque que la répartition statistique de la richesse suit une même loi mathématique : le logarithme du pourcentage des personnes ayant un revenu supérieur à une valeur x est une fonction affine (avec un coefficient de pente négative) du logarithme de cette valeur x. Il en déduit que le pourcentage de la population dont les revenus sont supérieurs à une valeur x est toujours proportionnel à $1/x^\beta$ (la valeur de β, entre 2 et 3, varie selon le pays).

On suppose donc que si l'on choisit une personne au hasard en France, son revenu annuel représenté par une variable aléatoire X admet la densité

$$p(x) = \frac{C}{x^{1+\beta}} \mathbb{1}_{[a,+\infty[}(x),$$

où $\beta > 0$ et $a > 0$ (a représente le revenu minimal). On dit alors que X suit une loi de Pareto de paramètres (β, a) et on écrit $X \sim \text{Pareto}(\beta, a)$.

1. Déterminer la valeur de C en fonction de β et de a. Pour $\beta > 1$, calculer l'espérance de X.
2. Trouver un sous-ensemble A de $]0, \infty[$ tel que la variable aléatoire X est de carré intégrable si et seulement si $\beta \in A$. Pour tout $\beta \in A$, déterminer la variance de X.
3. Soient $X_1, \ldots, X_n$ des variables i.i.d. de loi Pareto$(\beta, 1)$. On cherche à estimer le paramètre β.
   (a) Calculer l'estimateur du maximum de vraisemblance, noté $\hat{\beta}_n$, de β.
   (b) Prouver que $\hat{\beta}_n$ est fortement convergent.
   (c) Etablir la normalité asymptotique de la suite $\sqrt{n}(\hat{\beta}_n^{-1} - \beta^{-1})$ et calculer sa variance asymptotique. En déduire, en utilisant l'identité $\sqrt{n}(\hat{\beta}_n - \beta) = (-\hat{\beta}_n \beta)\sqrt{n}(\hat{\beta}_n^{-1} - \beta^{-1})$, que l'estimateur $\hat{\beta}_n$ est asymptotiquement normal.
   (d) Trouver un intervalle de confiance de niveau asymptotique $1 - \alpha$ pour β.
   (e) Soit $X \sim \text{Pareto}(\beta, 1)$. Déterminer la loi de $Y = \beta \ln(X)$.
   (f) En déduire [1] la loi de $\beta/\hat{\beta}_n$. En considérant les quantiles de la loi Gamma connus, proposer un intervalle de confiance de niveau (non-asymptotique) $1 - \alpha$ pour le paramètre β.
4. Soient $X_1, \ldots, X_n$ des variables i.i.d. de loi Pareto(β, a).
   (a) Quel est le comportement asymptotique de $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ pour $n \to \infty$?
   (b) Quel est l'estimateur du maximum de vraisemblance du couple (a, β) ?
   Indication : commencera par maximiser la vraisemblance par rapport à a, pour β fixé.
   (c) Soit $\hat{a}_n$ l'estimateur du maximum de vraisemblance de a. En admettant que $\hat{a}_n$ est fortement convergent, montrer que la suite $\widetilde{\beta}_n = \frac{\overline{X}_n}{\overline{X}_n - \hat{a}_n}$ converge presque sûrement vers une limite à préciser.
   (d) A.N. [2] : selon un sondage réalisé sur un échantillon de grande taille, le salaire (net) moyen annuel en 2008 des personnes interrogées était de 24400€ alors que le salaire minimal était de 15700€. En supposant que les salaires sont répartis selon une loi de Pareto, déterminer une valeur approchée du paramètre β.
   (e) Prouver que $\hat{a}_n$ est un estimateur faiblement convergent de a.

1. On admettra que si $Z_1, \ldots, Z_n$ sont iid de loi exponentielle de moyenne λ alors $Z_1 + \ldots + Z_n$ suit la loi gamma $\Gamma(n, \lambda^{-1})$.
2. Données disponibles sur le site `http://www.insee.fr/`