Contrôle des connaissances

Centrale - Statistiques avancées (première partie)

07 Mars 2011

A- Régression sans terme constant

On veut faire la régression d'un vecteur $Y \in \mathbb{R}^n$ sur les p régresseurs $X_1, \ldots, X_p \in \mathbb{R}^n$ sans terme constant : autrement dit, on cherche à minimiser en $\theta \in \mathbb{R}^p$ la quantité

$$\|Y - \theta_1 X_1 - \cdots - \theta_p X_p\|^2_{\mathbb{R}^n} .$$

Montrer comment la régression de $Z = \begin{pmatrix} Y \\ -Y \end{pmatrix} \in \mathbb{R}^{2n}$ sur les régresseurs

$$\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \begin{pmatrix} X_1 \\ -X_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_p \\ -X_p \end{pmatrix} \in \mathbb{R}^{2n}$$

permet d'obtenir la solution.

B- Inégalités de traitement

Le fichier "salaires.csv" contient les salaires de quelques employés choisis au hasard dans une grande entreprise, en fonction de leur sexe. L'étude porte sur 15 hommes et 15 femmes, et elle vise à mettre en évidence une éventuelle différence de traitement entre les deux sexes dans l'entreprise.

```
    Hommes             Femmes
 Min.   :  3649   Min.   : 6031
 1st Qu.: 24350   1st Qu.:15221
 Median : 32925   Median :32359
 Mean   : 41805   Mean   :37618
 3rd Qu.: 56535   3rd Qu.:50074
 Max.   :107432   Max.   :95039
```

Quel type de test peut-on (ou ne peut-on pas) utiliser en pareil cas ? Justifiez votre réponse.

C- Utilisation de R : la loi de Zipf

Le but de cet exercice est de vérifier la bonne compréhension de l'utilisation de R pour la régression linéaire : il ne s'agit en aucun cas de tester la connaissance exhaustive de toutes les commandes utilisées, mais plutôt de vérifier la bonne compréhension de la démarche globale mise en oeuvre et des résultats obtenus avec le logiciel.

Certains linguistes ont cherché à mettre une évidence une loi générale sur la fréquence d'utilisation des mots dans un texte. On associe à chaque mot un unique numéro $i \in \{1, \ldots, n\}$, où n désigne le nombre de mots différents apparaissant dans le texte, et on note F_i le nombre d'occurrences du mot numéro i. On désigne par $R(i)$ désigne le rang du mot i dans le classement des mots par ordre décroissant de fréquence. La loi de Zipf (nommée d'après son auteur George Kingsley Zipf) veut que

$$\forall i \in \{1, \ldots, n\}, \quad F_i = \frac{\beta}{(R_i)^\alpha}$$

où α est caractéristique du style de l'auteur et où β est un paramètre dépendant de la taille du texte.

On a relevé tous les mots de la pièce "Don Juan ou le Festin de Pierre", de Molière, ainsi que le nombre d'occurrences de chacun d'entre eux dans la pièce : voici les quelques premières lignes du fichier ainsi obtenu :

Freq	Mot
14	"TOUTE"
274	"LA"
1	"PHILOSOPHIE"
274	"IL"
186	"N"
295	"EST"
54	"RIEN"

Pour tester la validité de la loi de Zipf, et pour estimer le paramètre α, le code R suivant a été utilisé

```
donJuan <- read.table('donJuan.csv', header=T, sep=",")
nbWords <- dim(donJuan)[1]
n <- max(donJuan$Freq)
h <- hist(donJuan$Freq, breaks=seq(0,n))
# pour 1<=j<=n, h$counts[j] contient le nombre de mots apparaissant j fois dans le texte

rank <- c(1, 1 + cumsum(h$counts[seq(n,2,-1)]))
rank <- rank[seq(n,1,-1)];
donJuan$Rank <- rank[donJuan$Freq]

modeleZipf <- lm(log(Freq) ~ I(log(Rank)), data=donJuan)
summary(modeleZipf)
```

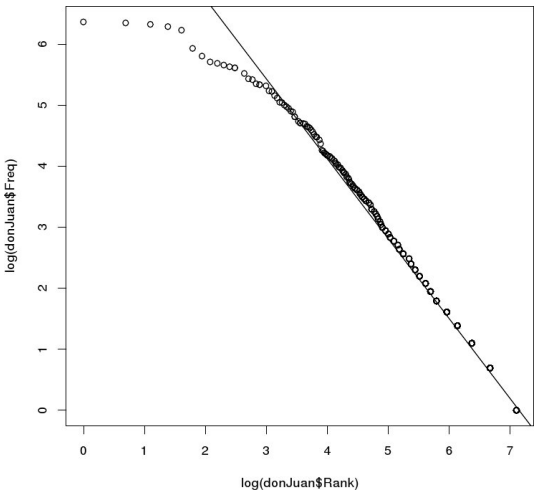
Le résultat renvoyé par R est le suivant :

```
Call:
lm(formula = log(Freq) ~ I(log(Rank)), data = donJuan)

Residuals:
     Min       1Q   Median       3Q      Max
-2.99572 -0.05173 -0.05173  0.07446  0.17798

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.362192   0.017980   520.7   <2e-16 ***
I(log(Rank)) -1.310719   0.002713  -483.2   <2e-16 ***
---
```

```
Signif. codes:  0  *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1164 on 2602 degrees of freedom
Multiple R-squared: 0.989,      Adjusted R-squared: 0.989
F-statistic: 2.335e+05 on 1 and 2602 DF,  p-value: < 2.2e-16
```

1. Expliquer comment le modèle linéaire a été utilisé ici.
2. Interpréter le résultat renvoyé par R.
3. La figure ci-dessous a été obtenue grâce aux commandes suivantes :

```
plot(log(donJuan$Rank), log(donJuan$Freq))
abline(modeleZipf)
```



   Expliquer et commenter la figure.
4. Dans une version forte, la loi de Zipf stipule que $\alpha = 1$. Peut-on l'exclure dans le cas étudié ici ?

D- Un théorème de Doob élémentaire

On se propose dans cet exercice de montrer le théorème de Doob dans un cadre élémentaire. On rappelle l'inégalité de Jensen *: pour toute variable aléatoire réelle Z bornée et pour toute fonction h convexe sur $\mathbb{R}$, $\mathbb{E}[h(Z)] \geq h(\mathbb{E}[Z])$.*

Soit Θ un ensemble fini, et soit π une loi de probabilité sur Θ. Soit $\mathcal{X}$ un autre ensemble fini, et pour chaque paramètre $\theta \in \Theta$ soit p_θ une loi de probabilité sur $\mathcal{X}$. On cherche à faire l'estimation bayésienne du paramètre θ, et on suppose que le modèle est identifiable : si $\theta \neq \theta'$, $p_\theta \neq p_{\theta'}$. Pour tout $\theta \in \Theta$, on note

$$H(\theta) = -\sum_{x \in \mathcal{X}} p_\theta(x) \log p_\theta(x),$$

et

$$K(\theta, \theta') = \sum_{x \in \mathcal{X}} p_\theta(x) \log \frac{p_\theta(x)}{p_{\theta'}(x)} \, .$$

Dans cette estimation bayésienne, l'expérience est la suivant suivante :
- d'abord, une variable aléatoire T est tirée suivant la loi π ;
- ensuite, sont tirées des variables aléatoires $X_1, X_2, \ldots$ qui, conditionnellement à l'événement $\{T = \theta\}$, sont indépendantes et de même loi p_θ.

Pour chaque valeur entière n, on apelle P^n la loi du $n+1$-uplet $(T, X_1, \ldots, X_n)$ sur l'ensemble $\Theta \times \mathcal{X}^n$, et pour tout $(x_1, \ldots, x_n) \in \mathcal{X}^n$ on note $\Pi^n(\cdot|x_1, \ldots, x_n)$ la loi conditionnelle de T sachant $\{X_1 = x_1, \ldots X_n = x_n\}$: on rappelle qu'elle est définie par

$$\forall \theta \in \Theta, \quad \Pi^n(\theta|x_1, \ldots, x_n) = \frac{P^n(T = \theta, X_1 = x_1, \ldots, X_n = x_n)}{\sum_{\theta' \in \Theta} P^n(T = \theta', X_1 = x_1, \ldots, X_n = x_n)} \ .$$

1. Montrer : $\forall \theta \in \Theta, H(\theta) \geq 0$.
2. Montrer : $\forall \theta, \theta' \in \Theta, \quad \theta \neq \theta' \implies K(\theta, \theta') > 0$.
3. Montrer que, pour tout $(\theta, x_1, \ldots, x_n) \in \Theta \times \mathcal{X}^n$,

$$P^n(T = \theta, X_1 = x_1, \ldots, X_n = x_n) = \pi(\theta) \prod_{k=1}^{n} p_\theta(x_i) \ .$$

4. On note $L_n(\theta, x_1, \ldots, x_n) = \log P^n(T = \theta, X_1 = x_1, \ldots, X_n = x_n)$. Montrer que, conditionnellement à $\{T = \theta_0\}$, $\frac{1}{n} L_n(\theta, X_1, \ldots, X_n)$ converge presque sûrement vers $-H(\theta_0) - K(\theta_0, \theta)$ pour tout $\theta \in \Theta$.
5. Montrer que, quand n tend vers l'infini, $\Pi^n(\theta|X_1, \ldots, X_n)$ converge presque-sûrement vers une variable aléatoire que l'on identifiera.
6. Peut-on généraliser cette preuve à des hypothèses (sur $\mathcal{X}$ et Θ) moins restrictives ?