TP d'Econométrie II avec le logiciel R

Pratique de la régression par moindres carrés et comportement asymptotique des statistiques du modèle linéaire

L'objectif de ce TP est de réaliser des simulations de modèles de régression linéaire multiple, de les analyser par la méthode des moindres carrés et de regarder le comportement des statistiques usuelles (estimateurs et tests) pour un nombre d'individus très grand. Avant cela, une étude sur données réelles est proposée

Traitement d'une base de données

Le fichier de données `bea-2006.csv` (à télécharger depuis le site) contient des informations sur les économies des 366 zones statistiques métropolitaines (villes) des États-Unis en 2006. En particulier, il répertorie, pour chaque ville, la population, la valeur totale de l'ensemble des biens et services produits pour la vente dans la ville cette année-là par personne ("produit métropolitain brut par habitant", pcgmp) et la part de la production économique provenant de quatre industries sélectionnées.

1. Chargez le fichier de données et vérifiez qu'il comporte 366 lignes et 7 colonnes. Pourquoi comporte-t-il 7 colonnes, alors que le paragraphe ci-dessus ne décrit que 6 variables?
2. Calculez les statistiques sommaires pour les six colonnes à valeur numérique. Qu'est-ce que les `NA`? Définir un nouveau fichier sans ces valeurs `NA` en utilisant la commande `na.omit`.
3. Avec la commande `ggpairs`, étudier les liens entre les variables 2 à 2.
4. Effectuer une régression par moindres carrés du GMP par habitant en fonction de la population.
5. Tracer les 4 graphes relatifs à la régression. Conclusions?
6. Effectuer une régression par moindres carrés du GMP par habitant en fonction de toutes les autres variables. Commenter les résultats numériques et tracer les graphes.

Simulations

On commence par des cas simples de régressions polynomiales. Copier les commandes suivantes dans un script que vous pourrez enregister puis utiliser avec des modifications ensuite (on peut compiler une ou plusieurs lignes de commande du script).

```
n=100
i=1:n
Z1=i;Z2=i^2;Z3=i^3
epsilon=5*rnorm(n,0)
Y=5-0.03*Z1+0.002*Z2+epsilon;
plot(i,Y)
```

Qu'a-t-on fait suivant ces lignes de commande? Vérifier en affichant les variables simulées.
On passe maintenant aux résultats de la régression:

```
Y.lm=summary(lm(Y~Z1+Z2))
plot(lm(Y~Z1+Z2))
names(Y.lm)
Y.lm$coeff
Y.lm$fstat
Y.lm$sigma
vcov(lm(Y~Z1+Z2))
```

Comprendre et commenter chacune des commandes tapées. Relancer plusieurs fois ces suites de commandes. Faire une boucle (utiliser la commande `for` pour répéter 100 fois, puis stocker les résultats. Tracer des histogrammes pour voir le comportement en distribution des estimateurs. Changer la loi de `epsilon` par une loi uniforme sur $[-30, 30]$. Recommencer les mêmes étapes dans un tel cas. Vérifier en particulier la normalité asymptotique des estimateurs et la loi asymptotique du χ^2 pour la statistique de Fisher. Essayer ensuite avec une loi de Cauchy (utiliser la méthode d'inversion de la fonction de répartition pour simuler une telle loi). Conclusions?
Refaire la même chose pour $n = 1000$ puis pour $n = 10000$. Vérifier la convergence des estimateurs. Donner un moyen d'approcher la vitesse de convergence en n des estimateurs.

Premier exercice

Recommencer les étapes précédentes (dans le cas gaussien uniquement) avec le même modèle de simulation de `Y` mais en effectuant une régression par rapport à `Z1`, `Z2` et `Z3`. Qu'observe-t-on? De même, recommencer les étapes précédentes (dans le cas gaussien uniquement) avec le même modèle de simulation de `Y` mais en effectuant une régression par rapport à `Z1` seulement. Qu'observe-t-on?

Second exercice

Remplacer `Z2` dans le programme par `Z2`$= 1/i$. Vérifier la non convergence du coefficient lié à `Z2`. Que peut-on dire des autres coefficients estimés? Et de l'estimateur de la variance et de la statistique de test?

Preuve: On a $Z = \big(\mathbb{I}, Z1, Z2)\big)$ d'où après calcul

$$ {}^tZ\,Z = \begin{pmatrix} n & \frac{1}{2}n(n+1) & \sum_{i=1}^n \frac{1}{i} \\ \frac{1}{2}n(n+1) & \frac{1}{6}n(n+1)(2n+1) & n \\ \sum_{i=1}^n \frac{1}{i} & n & \sum_{i=1}^n \frac{1}{i^2} \end{pmatrix} \simeq \begin{pmatrix} n & \frac{1}{2}n(n+1) & \log(n) \\ \frac{1}{2}n(n+1) & \frac{1}{6}n(n+1)(2n+1) & n \\ \log(n) & n & \frac{\pi^2}{6} \end{pmatrix} $$

On montre que $\det({}^tZ\,Z) \simeq \frac{\pi^2}{72}\, n^4$ et le terme $(3,3)$ de $({}^tZ\,Z)^{-1}$ est donc $\simeq \frac{1}{12}\, n^4 / \det({}^tZ\,Z) \simeq \frac{6}{\pi^2}$ qui ne tend pas vers 0.

Troisième exercice

Remplacer `Z1` par des réalisations indépendantes de v.a.i.i.d. (par exemple de loi exponentielle de paramètre 1) indépendantes de `epsilon` et `Z2` par des réalisations indépendantes de v.a.i.i.d. (par exemple de loi binomiale $\mathcal{B}(10, 1/3)$), indépendantes de `Z1` et de `epsilon`. Utiliser cette fois-ci le modèle `Y`$= 5 + 2$ `Z1`-3 `Z2` $+$ `epsilon` avec `epsilon` qui suit une loi normale $\mathcal{N}(0, 4)$. Réécrire le modèle en recentrant les variables `Z1` et `Z2`. Vérifier numériquement la convergence des coefficients. Expliquer théoriquement pourquoi.

Preuve: On a $\mathbb{E}$[`Z1`] $= 1$ et $\mathbb{E}$[`Z2`] $= \frac{10}{3}$. D'où `Y`=-3+2 `Z1`$' - 3$ `Z2`$' +$ `epsilon` avec `Z1`$' =$ `Z1` $- 1$ et `Z2`$' =$ `Z2`$' - \frac{10}{3}$. Alors:

$$ {}^tZ\,Z = \begin{pmatrix} n & \sum_{i=1}^n Z1_i & \sum_{i=1}^n Z2_i \\ \sum_{i=1}^n Z1_i & \sum_{i=1}^n (Z1_i)^2 & \sum_{i=1}^n Z1_i Z2_i \\ \sum_{i=1}^n Z2_i & \sum_{i=1}^n Z1_i Z2_i & \sum_{i=1}^n (Z2_i)^2 \end{pmatrix} \simeq n \begin{pmatrix} 1 & \overline{Z1} & \overline{Z2} \\ \overline{Z1} & \overline{Z1^2} & \overline{Z1Z2} \\ \overline{Z2} & \overline{Z1Z2} & \overline{Z2^2} \end{pmatrix} $$

Par la loi forte des grands nombres, les moyennes empiriques convergent vers les espérances et on obtient:

$$\begin{pmatrix} 1 & \overline{Z1} & \overline{Z2} \\ \overline{Z1} & \overline{Z1^2} & \overline{Z1Z2} \\ \overline{Z2} & \overline{Z1Z2} & \overline{Z2^2} \end{pmatrix} \underset{n\to+\infty}{\overset{\mathcal{P}}{\longrightarrow}} \begin{pmatrix} 1 & 1 & \frac{10}{3} \\ 1 & 1 & \frac{10}{3} \\ \frac{10}{3} & \frac{10}{3} & \frac{40}{3} \end{pmatrix} = A$$

en utilisant le fait que $\mathbb{E}[Z1\,Z2] = \mathbb{E}[Z1]\,\mathbb{E}[Z2]$ par l'indépendance. Cette dernière matrice A est inversible, d'où $n\,({}^tZ\,Z)^{-1} \underset{n\to+\infty}{\overset{\mathcal{P}}{\longrightarrow}} A^{-1}$: il y a convergence à vitesse $1/\sqrt{n}$ de $\widehat{\theta}$ vers θ^*.