Analyse statistique de données climatiques

Marie-Cécile Selosse

October 10, 2017

Contents

1 La Terre se réchauffe-t-elle ?

On trouve sur le site du Climatic Research Unit les moyennes globales de température de 1856 à 2004 (mois par mois et annuelles). Une copie locale où on n'a conservé que les données de température est disponible localement temperature_globe.txt

Question 1 *Saisir le fichier de données de température.*
Extraire les moyennes annuelles.
Tracer la courbe $t \hookrightarrow T(t)$ donnant l'évolution de ces dernières.

```
//saisie de données de température
Tg=fscanfMat('temperatures_globe.txt')
//on extrait la dernière colonne qui contient les moyennes annuelles
y=Tg(:,$)
// t représente le temps (années)
[m,n]=size(Tg);t=[1:m]';
// graphique des données
xbasc();plot2d(t',y');
```

Question 2 *Calculer les coefficients de la droite de régression $y = \theta_1 + \theta_2 t$ qui est la plus proche de la courbe des températures au sens des moindres carrés.*

```
//on programme une régression linéaire
//formule de régression a la main
x=[ones(t),t];
M=(x'*x)^(-1);
theta=M*x'*y;
// On peut remarquer que Scilab effectue la résolution
// au sens des moindres carrés si x est << tall >>
// i.e. est une matrice qui a plus de lignes que de colonnes
// Le problème est bien posé si son rang est égal à son nombre de colonnes,
// sinon Scilab donne une solution parmi les solutions possibles
theta=x\y
// on peut aussi utiliser reglin
// [a,b,sig]=reglin(t',y')
```

Question 3 *Superposer la courbe des données et la droite de régression.*

```
//superposition graphique des données et de la droite de régression
xbasc();
plot2d(t,y);
plot2d(t,(x*theta),2,"000");
```

Question 4 *La température $T(t)$ est supposée être une réalisation du modèle $T(t) = \theta_1 + \theta_2 t + \varepsilon_t$, où ε_1, ε_2... est une suite de variables aléatoires indépendantes de même loi normale $\mathcal{N}(0, \sigma^2)$.*

Interpréter en terme de paramètres l'hypothèse (H_0) ¡¡ la température est stationnaire ¿¿. Tester cette hypothèse. (On pourra consulter la section 3 a la fin de ce document).

On calcule l'estimateur de σ^2 appellé s^2

```
p=2
Res=y-x*theta;
s2=Res'*Res/(n-p);
```

On calcule l'écart type de $\widehat{\theta}_2$ à partir de la covariance de $\widehat{\theta}$:

```
sigtheta2=sqrt(s2*M(2,2))
```

Alors `theta(2)/sigtheta2` suit une loi de Student à $(n-2)$ degrés de liberté :

```
T=theta(2)/sigtheta2
// Utiliser la fonction cdft pour calculer la
// probabilite qu'une v.a suivant une loi de Student a (n-2)
// degré de liberté dépasse T
```

Question 5 *Examiner graphiquement la courbe des résidus $\widehat{\varepsilon}_t = T(t) - \widehat{\theta}_1 - \widehat{\theta}_2 t$.*

Subsiste-t-il une structure dans les résidus ?

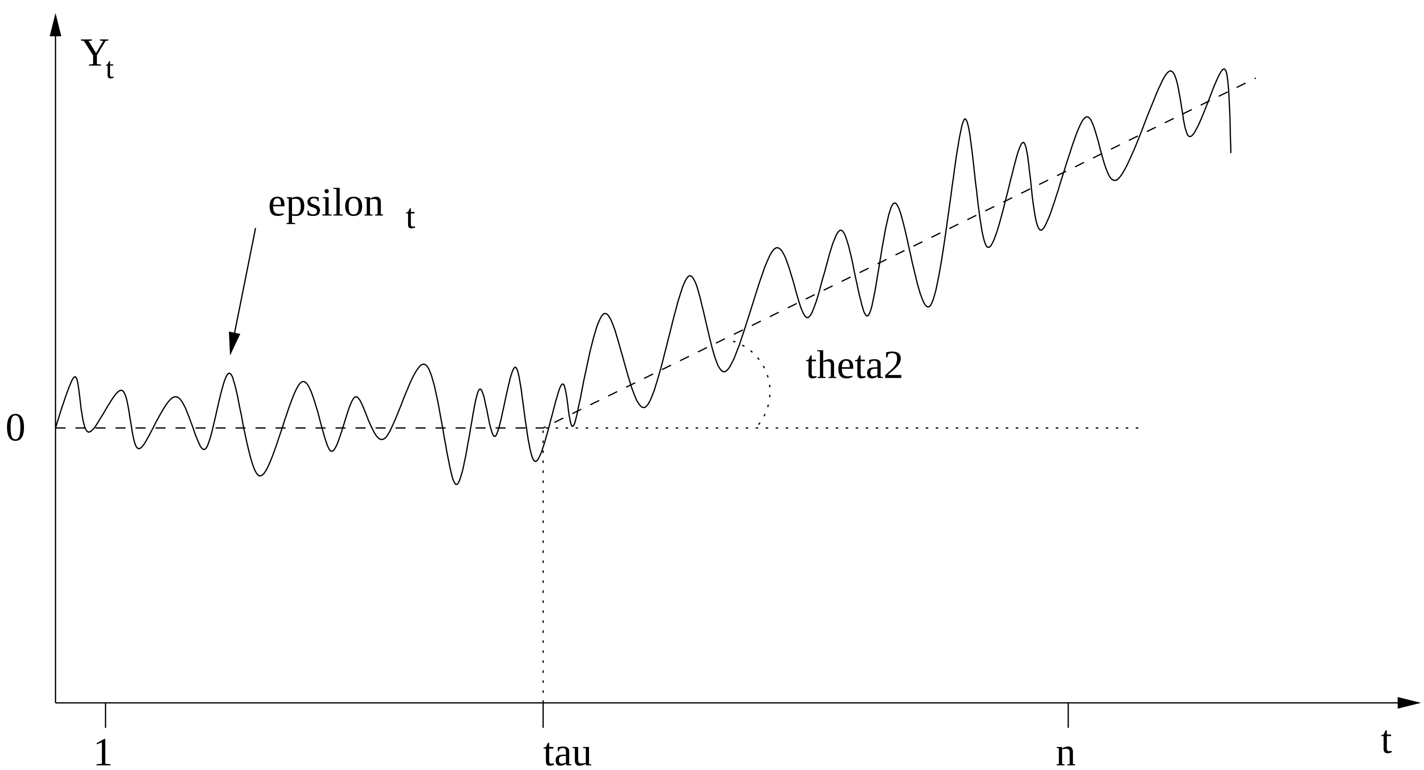
Tracer la courbe $t \mapsto (\widehat{\varepsilon}_t, T(t))$.

À l'examen de l'histogramme des résidus, l'hypothèse de normalité vous semble-t-elle raisonnable ?

```
//examen des résidus
r=y-x*theta
plot(r)
histplot(10,r)
```

2 À quelle date le changement climatique a-t-il commencé ?

2.1 Modèle statistique



$$\begin{cases} Y_t = \theta_1 + \varepsilon_t & 1 \leq t < \tau \\ Y_t = \theta_1 + \theta_2(t-\tau) + \varepsilon_t & \tau \leq t \leq n \end{cases} \quad (1)$$

On suppose les ε_t indépendants et identiquement distribués (i.i.d.) suivant une loi normale $\mathcal{N}(0, \sigma^2)$.

Le vecteur des paramètres est noté

$$\phi = (\theta_1, \theta_2, \sigma^2, \tau)\,. \tag{2}$$

2.2 Position du problème

On estime les quatre paramètres par la méthode du maximum de vraisemblance.

Définition de la fonction de vraisemblance

La fonction de vraisemblance de la loi normale est définie comme suit :

$$L(\phi, Y) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\left\{\sum_{t=1}^{\tau-1}(y_t-\theta_1)^2 + \sum_{t=\tau}^{n}(y_t-\theta_1-\theta_2(t-\tau))^2\right\}\right] \tag{3}$$

Pour calculer le maximum de vraisemblance, c'est-à-dire les valeurs de θ_1, θ_2, σ^2 et τ qui maximisent $\log L(\phi, Y)$, on choisit de :

- fixer d'abord $\tau \in \{1, ..., n\}$;
- calculer les valeurs $\widehat{\theta}_1^*(\tau)$, $\widehat{\theta}_2^*(\tau)$ et $\widehat{\sigma^2}^*(\tau)$ qui maximisent $\log L(\phi, Y)$;
- calculer enfin la valeur τ^* qui maximise $\log L^*(\tau) = \log L(\widehat{\theta}_1^*(\tau), \widehat{\theta}_2^*(\tau), \widehat{\sigma^2}^*(\tau), \tau)$.

On effectue donc les opérations dans l'ordre suivant :

$$\max_{\tau,\theta_1,\theta_2,\sigma^2} \log L(\phi, Y) = \max_{\tau=1,\ldots,n}\left(\max_{\theta_1,\theta_2,\sigma^2} \log L(\phi, Y)\right) = \max_{\tau=1,\ldots,n} \log L(\widehat{\theta}_1^*(\tau), \widehat{\theta}_2^*(\tau), \widehat{\sigma^2}^*(\tau), \tau)\,.$$

Calcul de $\widehat{\theta}_1^*(\tau)$, $\widehat{\theta}_2^*(\tau)$ et $\widehat{\sigma^2}^*(\tau)$

Question 6 *Montrer, par dérivation de (3), que $\widehat{\theta}_1^*(\tau)$ et $\widehat{\theta}_2^*(\tau)$ sont solution de*

$$\left\{\begin{array}{rcrcl} \widehat{\theta}_1^*(\tau)n & + & \widehat{\theta}_2^*(\tau)\displaystyle\sum_{t=\tau}^{n}(t-\tau) & = & \displaystyle\sum_{t=1}^{n} y_t \\ \widehat{\theta}_1^*(\tau)\displaystyle\sum_{t=\tau}^{n}(t-\tau) & + & \widehat{\theta}_2^*(\tau)\displaystyle\sum_{t=\tau}^{n}(t-\tau)^2 & = & \displaystyle\sum_{t=\tau}^{n}(t-\tau)y_t \end{array}\right. \tag{4}$$

et que

$$\widehat{\sigma^2}^*(\tau) = \frac{1}{n}\left\{\sum_{t=1}^{\tau-1}(y_t-\widehat{\theta}_1^*(\tau))^2 + \sum_{t=\tau}^{n}(y_t-\widehat{\theta}_1^*(\tau)-(t-\tau)\widehat{\theta}_2^*(\tau))^2\right\}\,. \tag{5}$$

Calcul de τ^*

Question 7 *Programmer en Scilab le calcul de τ^* solution de*

$$\max_{\tau=1,\dots,n} \log L(\widehat{\theta}_1^*(\tau), \widehat{\theta}_2^*(\tau), \widehat{\sigma^2}^*(\tau), \tau)$$

```
function [ttheta1,ttheta2,ssigma2]=rupture(tau,y)
  n=size(y,"*");// taille de y
  A=[n,sum(1:n-tau);sum(1:n-tau),sum((1:n-tau)^2)]
  B=[sum(y);(1:n-tau)*y(tau+1:$)'];
  x=A\B;
  ttheta1=x(1,:);
  ttheta2=x(2,:);
  ssigma2=1/n* ...
          (sum((y(1:tau-1)-ttheta1)^2)+sum((y(tau+1:$)-ttheta1-ttheta2*(1:n-tau))^2));
endfunction
```

3 Test de nullité d'un coefficient dans un modèle de régression linéaire simple

Soient n couples de réels (x_1, Y_1),..., (x_n, Y_n) où seule la seconde composante Y_t est aléatoire, supposés suivre le modèle

$$Y_t = \theta_1 + \theta_2 x_t + \varepsilon_t \tag{6}$$

où ε_1, ε_2... est une suite de variables aléatoires indépendantes de même loi normale $\mathcal{N}(0, \sigma^2)$.

En notations vectorielles, on pose

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{pmatrix} \qquad \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$$

soit

$$Y = X\theta + \varepsilon\,.$$

On appelle *estimateur de Gauss-Markov* (ou *estimateur des moindres carrés*)

$$\widehat{\theta} = (X'X)^{-1}X'Y \tag{7}$$

où on a supposé X de rang plein égal à 2. On note

$$U = (X'X)^{-1}X' \tag{8}$$

qui satisfait aux relations

$$UX = I\,, \quad UU' = (X'X)^{-1}\,, \quad XU = U'X' = XUU'X' = X(X'X)^{-1}X'\,. \tag{9}$$

On a
$$\widehat{\theta} = \theta + U\varepsilon \tag{10}$$
si bien que
$$\widehat{\theta} \sim \mathcal{N}(\theta, \sigma^2(X'X)^{-1})\,. \tag{11}$$
On note le vecteur des résidus
$$\widehat{\varepsilon} = Y - X\widehat{\theta} \tag{12}$$
On a
$$\widehat{\varepsilon} = (I - XU)\varepsilon \tag{13}$$
si bien que
$$\widehat{\varepsilon} \sim \mathcal{N}(0, \sigma^2(I - X(X'X)^{-1}X'))\,. \tag{14}$$
Comme $X'\,\mathsf{D}(\widehat{\varepsilon})X = 0$, $\mathsf{D}(\widehat{\varepsilon})$ est de rang $n-2$ car on a supposé X de rang plein égal à 2. Donc
$$\frac{\|\widehat{\varepsilon}\|^2}{\sigma^2} \sim \chi^2(n-2)\,. \tag{15}$$
On pose
$$\widehat{\sigma^2} = \frac{\|\widehat{\varepsilon}\|^2}{n-2} \tag{16}$$
dont on peut vérifier $\mathbb{E}(\widehat{\sigma^2}) = \sigma^2$.

Les vecteurs $\widehat{\varepsilon} = Y - X\widehat{\theta}$ et $\theta - \widehat{\theta}$ sont indépendants, car le couple
$$(Y - X\widehat{\theta}, \widehat{\theta} - \theta) = ((I - XU)\varepsilon, U\varepsilon) \tag{17}$$
est gaussien décorrélé (propriété de la projection).

Donc le rapport
$$T = \frac{\widehat{\theta}_2 - \theta_2}{\sqrt{\widehat{\sigma^2}(X'X)^{-1}_{2,2}}} \tag{18}$$
suit une loi de Student à $n-2$ degrés de liberté.