

Chapitre 6

Spécification du modèle

Ghislain Léveillé
École d'actuariat, Université Laval

Copyright © 2017

6.1 Contenu du chapitre

Dans le but de représenter une série de données stationnaire, nous avons développé (au chapitre 4) un premier modèle de série chronologique, soit le modèle ARMA(p,q). Dans le cas où il y aurait absence de stationnarité, nous avons aussi développé au chapitre 5 un modèle de série chronologique plus général, soit le modèle ARIMA(p,d,q), qui nous a permis de ramener la stationnarité suite à certaines opérations préalables sur notre série de données.

Ainsi, dans ce chapitre, notre intention est de spécifier le modèle de série chronologique, i.e. les paramètres p, d et q du processus ARIMA(p,d,q), le « mieux adapté » à notre série de données. Pour ce faire, nous utiliserons dans notre analyse les fonctions d'autocorrélation échantillonnale complète, partielle et étendue. Le test de la racine unitaire de Dickey-Fuller sera aussi utilisé pour nous aider à déterminer si la série est stationnaire ou non.

6.2 Spécification par la fonction d'autocorrélation échantillonnale

6.2.1 Rappel et motivation

Nous avons défini au chapitre 3 la fonction d'autocorrélation échantillonnale (ACF)

$$r_k = \frac{\sum_{t=k+1}^n (Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^n (Y_t - \bar{Y})^2}, \quad k = 1, 2, 3, \dots$$

Notre but est d'utiliser cet estimé de ρ_k , dont les comportements sont connus pour les processus ARMA(p,q) usuels, afin de nous aider à identifier les modèles ARMA qui seraient de « bons » candidats pour représenter notre série de données.

Ainsi, nous savons que $\rho_k = 0$, pour $k > q$ dans le cas d'un modèle MA(q). Par contre, dans le cas d'un modèle AR(p) (faiblement) stationnaire, $\rho_k \downarrow 0$ quand $k \uparrow \infty$. Ainsi, pour un processus ARMA(p,q) (faiblement) stationnaire, $\rho_k \downarrow 0$ quand $k \uparrow \infty$.

Bien sûr, il nous faut aussi être prudent dans l'utilisation de cet estimé de ρ_k et nous aurons besoin de connaître les propriétés de notre estimateur r_k afin d'établir sa « précision » par rapport à ρ_k , ce qui sera l'objet de la prochaine section.

6.2.2 Propriétés de l'estimateur r_k

Considérer le processus linéaire général $\{Y_t; t \in \mathbb{Z}\}$ défini par

$$Y_t = \mu + \sum_{j=0}^{\infty} \psi_j e_{t-j},$$

où

- $\{e_t; t \in \mathbb{Z}\}$ est un bruit blanc (faible)

- $E[e_t] = 0$, $Var[e_t] = \sigma_e^2$

- $\sum_{j=0}^{\infty} |\psi_j| < \infty$, $\sum_{j=0}^{\infty} j\psi_j^2 < \infty$

Alors, pour tout m fixé, la distribution conjointe des variables aléatoires suivantes

$$\sqrt{n}(r_1 - \rho_1), \dots, \sqrt{n}(r_m - \rho_m)$$

converge, quand $n \rightarrow \infty$, vers une loi normale conjointe avec moyennes nulles, variances c_{jj} et covariances c_{ij} ,

où,

$$c_{ij} = \sum_{k=-\infty}^{\infty} (\rho_{k+i}\rho_{k+j} + \rho_{k-i}\rho_{k+j} - 2\rho_i\rho_k\rho_{k+j} - 2\rho_j\rho_k\rho_{k+i} + 2\rho_i\rho_j\rho_k^2) .$$

Preuve : Voir le livre de Shumway and Stoffer (2006, P. 519-520)

Remarques : (1) Pour une variable aléatoire ou un vecteur aléatoire X , nous utiliserons à l'occasion la notation

$$X \sim AN(\mu, \sigma^2)$$

pour indiquer que X est « approximativement » normal avec moyenne (ou vecteur de moyennes) μ et variance (ou matrice de variances-covariances) σ^2 .

2) Pour n grand,

- $r_k \sim AN(\rho_k, c_{kk}/n)$, où

$$c_{kk} = \sum_{i=-\infty}^{\infty} (\rho_{i+k}^2 + \rho_{i-k}\rho_{i+k} - 4\rho_k\rho_i\rho_{i+k} + 2\rho_i^2\rho_k^2).$$

- $\text{Corr}(r_k, r_j) = c_{kj} / \sqrt{c_{kk}c_{jj}}$

(3) De la remarque précédente, nous avons

- Pour n grand, $\text{Var}(r_k)$ est inversement proportionnelle au nombre d'observations (n).
- Pour n grand, $\text{Corr}(r_k, r_j)$ est approximativement constante.

6.2.3 Exemples

(1) $\{Y_t; t \in \mathbb{Z}\}$ est un bruit blanc

- Pour n grand, (... puisque $c_{kk} = \rho_0^2 = 1$)

$$r_k \sim AN\left(0, 1/n\right), \quad k \in \mathbb{N}, \text{ et } \text{Corr}\left(r_k, r_j\right) \approx 0, \quad k \neq j.$$

- Une valeur de r_k tel $|r_k| > 2/\sqrt{n}$ sera donc considérée comme « inhabituelle » sous l'hypothèse d'un bruit blanc.

(2) $\{Y_t; t \in \mathbb{Z}\}$ est un processus stationnaire AR(1)

- Pour n grand,

$$r_k \sim AN\left(\rho_k, \sigma_{r_k}^2\right), \quad k \in \mathbb{N},$$

où,

$$- \rho_k = \varphi^k$$

$$- \sigma_{r_k}^2 \approx \frac{1}{n} \left[\frac{(1+\varphi^2)(1-\varphi^{2k})}{1-\varphi^2} - 2k\varphi^{2k} \right]$$

(3) $\{Y_t; t \in \mathbb{Z}\}$ est un processus inversible MA(1)

- Pour n grand,

$$r_1 \sim AN\left(\rho_1, \sigma_{r_1}^2\right),$$

où,

$$\rho_1 = -\frac{\theta}{1+\theta^2} , \quad \sigma_{r_1}^2 \approx \frac{1-3\rho_1^2+4\rho_1^4}{n} .$$

- Pour n grand et $k > 1$,

$$r_k \sim AN\left(0, \sigma_{r_k}^2\right),$$

où,

$$\sigma_{r_k}^2 \approx \frac{1+2\rho_1^2}{n}.$$

(4) $\{Y_t; t \in \mathbb{Z}\}$ est un processus inversible MA(q)

- Pour n grand et $k > q$,

$$r_k \sim AN\left(0, \sigma_{r_k}^2\right),$$

où,

$$\sigma_{r_k}^2 \approx \frac{1+2\sum_{j=1}^q \rho_j^2}{n}.$$

Remarque : Dans le cas d'un processus MA(q), nous pourrions élaborer un test statistique (basé sur un n grand), comme suit ...

H_0 : MA(q) est un modèle approprié pour nos données

versus

H_1 : MA(q) n'est pas un modèle approprié pour nos données

En effet, si H_0 est vraie, alors

$$r_{q+1} \sim AN\left[0, \frac{1}{n} \left(1 + 2 \sum_{j=1}^q \rho_j^2\right)\right] ,$$

i.e.

$$Z = \frac{r_{q+1}}{\sqrt{\frac{1}{n} \left(1 + 2 \sum_{j=1}^q \rho_j^2\right)}} \sim AN(0,1) .$$

Puisque les ρ_j sont inconnus, nous les remplacerons par les estimés r_j , ce qui ne devrait pas affecter la valeur de Z pour n grand. Ainsi, nous obtenons le test statistique

$$Z^* = \frac{r_{q+1}}{\sqrt{\frac{1}{n} \left(1 + 2 \sum_{j=1}^q r_j^2 \right)}} \sim AN(0,1) .$$

Un test d'hypothèse bilatéral à un niveau α nous amènera à rejeter H_0 en faveur de H_1 si

$$|Z^*| > z_{\alpha/2} ,$$

où $z_{\alpha/2}$ est le $\alpha/2$ quantile supérieur de la distribution $N(0,1)$.

6.3 Les fonctions d'autocorrélation partielle et étendue

6.3.1 La fonction d'autocorrélation partielle (PACF)

Nous avons vu précédemment que la fonction d'autocorrélation d'une MA(q) est nulle pour des écarts dépassant q , ce qui facilite l'identification d'un modèle MA(q) si nous soupçonnons que ce dernier puisse représenter « correctement » les données.

Par contre, l'identification d'un modèle stationnaire AR(p) est plus complexe puisque la fonction d'autocorrélation décroît (en valeur absolue) plutôt vers 0 avec l'augmentation des écarts. Il nous faudrait donc une nouvelle fonction qui se comporterait comme la fonction d'autocorrélation pour le processus MA(q), mais cette fois-ci pour le processus AR(p). Dans ce qui suit, nous allons donc motiver et construire cette nouvelle fonction qui nous aidera à identifier plus facilement le modèle (stationnaire) AR(p).

- Nous considérons d'abord le meilleur prédicteur linéaire de Y_t basé sur les $k-1$ variables intermédiaires entre Y_{t-k} et Y_t , soit $Y_{t-1}, \dots, Y_{t-k+1}$, i.e celui qui minimise l'erreur quadratique moyenne. Nous obtenons ainsi

$$\beta_1 Y_{t-1} + \dots + \beta_{k-1} Y_{t-k+1} .$$

Nous considérons ensuite le meilleur prédicteur linéaire de Y_{t-k} basé sur les mêmes $k-1$ variables intermédiaires entre Y_{t-k} et Y_t . Par hypothèse, notre série chronologique étant stationnaire, nous obtenons

$$\beta_1 Y_{t-k+1} + \dots + \beta_{k-1} Y_{t-1} ,$$

où les coefficients sont les mêmes que précédemment mais dans l'ordre inverse.

En effet, le cas $k = 2$ étant trivial, nous vérifions pour $k = 3$. Nous avons alors à minimiser les 2 espérances quadratiques suivantes :

$$E\left[\left(Y_t - \beta_1 Y_{t-1} - \beta_2 Y_{t-2}\right)^2\right], \quad E\left[\left(Y_{t-3} - \alpha_1 Y_{t-2} - \alpha_2 Y_{t-1}\right)^2\right].$$

Par stationnarité, nous obtenons donc les deux systèmes d'équations

$$\beta_1 \gamma_0 + \beta_2 \gamma_1 = \gamma_1 \quad , \quad \alpha_1 \gamma_0 + \alpha_2 \gamma_1 = \gamma_1$$

$$\beta_1 \gamma_1 + \beta_2 \gamma_0 = \gamma_2 \quad , \quad \alpha_1 \gamma_1 + \alpha_2 \gamma_0 = \gamma_2$$

Nous constatons facilement que

$$\alpha_1 = \beta_1 \quad , \quad \alpha_2 = \beta_2 \quad .$$

- La fonction d'autocorrélation partielle (théorique) d'ordre k (d'écart k) sera alors définie par l'identité suivante :

$$\phi_{kk} = \text{Corr}\left(Y_t - (\beta_1 Y_{t-1} + \dots + \beta_{k-1} Y_{t-k+1}), Y_{t-k} - (\beta_1 Y_{t-k+1} + \dots + \beta_{k-1} Y_{t-1})\right), \quad k > 1,$$

où $\phi_{11} = \text{Corr}(Y_t, Y_{t-1}) = \rho_1$.

Cette fonction représente donc (pour $k > 1$) la corrélation entre les erreurs engendrées par les prédicteurs linéaires de Y_t et de Y_{t-k} , ceux-ci s'appuyant sur les variables intermédiaires $Y_{t-1}, \dots, Y_{t-k+1}$, i.e. la corrélation entre Y_t et Y_{t-k} une fois supprimé la dépendance linéaire avec ces variables intermédiaires.

Ainsi, si la série chronologique peut être modélisée adéquatement par un processus AR(p) stationnaire, nous montrons que

$$\phi_{kk} \neq 0 \text{ pour } k \leq p \text{ et } \phi_{kk} = 0 \text{ pour } k > p.$$

- Pour démontrer la proposition précédente, nous aurons besoin d'un résultat bien connu en régression. En effet, nous savons déjà que si une variable Y doit être prédite par une fonction des n variables X_1, \dots, X_n , soit $f(X_1, \dots, X_n)$, alors le prédicteur minimal au sens des moindres carrés sera donné par la courbe de régression

$$E[Y|X_1, \dots, X_n] .$$

i.e.

$$E\left[\left[Y - f(X_1, \dots, X_n)\right]^2\right] \geq E\left[\left[Y - E[Y|X_1, \dots, X_n]\right]^2\right] .$$

Dans le cas qui nous concerne, si la série chronologique peut être modélisée par un processus (stationnaire) AR(p), i.e. si

$$Y_t = \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} + e_t ,$$

alors le prédicteur minimal de Y_t au sens des moindres carrés sera pour $k > p$

$$E[Y_t | Y_{t-1}, \dots, Y_{t-k+1}] = E[\phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} + e_t | Y_{t-1}, \dots, Y_{t-k+1}] = \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} ,$$

i.e. que dans le cas d'un processus AR(p)

$$\beta_1 Y_{t-1} + \dots + \beta_{k-1} Y_{t-k+1} = \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} , \quad k > p .$$

Ainsi, nous obtenons pour $k > p$ et $h(Y_{t-1}, \dots, Y_{t-k+1})$ le meilleur prédicteur linéaire de Y_{t-k}

$$\begin{aligned} Cov\left(Y_t - (\phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p}), Y_{t-k} - h(Y_{t-1}, \dots, Y_{t-k+1})\right) \\ = Cov(e_t, Y_{t-k} - h(Y_{t-1}, \dots, Y_{t-k+1})) \\ = 0 \end{aligned}$$

Maintenant, nous montrons que $\phi_{kk} = \phi_k \neq 0$, $k \leq p$

- Modèle AR(1) :

. Nous avons $\phi_{11} = \rho_1 = \phi \neq 0$.

- Modèle AR(2) :

. Nous avons $\phi_{11} = \rho_1 = \frac{\phi_1}{1 - \phi_2} \neq 0$.

. Nous avons

$$\begin{aligned}\phi_{22} &= \text{Corr}\left(Y_t - \beta_1 Y_{t-1}, Y_{t-2} - \beta_1 Y_{t-1}\right) \\ &= \frac{\text{Cov}\left(\phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t - \beta_1 Y_{t-1}, Y_{t-2} - \beta_1 Y_{t-1}\right)}{\text{Var}\left(Y_t - \beta_1 Y_{t-1}\right)}\end{aligned}$$

Comme $\beta_1 = \gamma_1 / \gamma_0$ ($= \rho_1$),

$$\begin{aligned}\phi_{22} &= \frac{(\phi_1 - \beta_1)(\gamma_1 - \beta_1 \gamma_0) + \phi_2(\gamma_0 - \beta_1 \gamma_1)}{\gamma_0 + \beta_1^2 \gamma_0 - 2\beta_1 \gamma_1} \\ &= \frac{\phi_2 \gamma_0 (1 - \beta_1^2)}{\gamma_0 (1 - \beta_1^2)} \\ &= \phi_2 \\ &\neq 0\end{aligned}$$

Remarques : (1) Pour un processus MA(q), nous pouvons montrer que la fonction PACF décroît exponentiellement avec l'écart k.

Pour le processus MA(1), nous avons pour $k \geq 1$ (à vérifier)

$$\phi_{kk} = \frac{\theta^k (\theta^2 - 1)}{1 - \theta^{2(k+1)}} \Rightarrow \lim_{k \rightarrow \infty} \phi_{kk} = 0 .$$

Pour $q > 1$, les mathématiques à manipuler étant plus complexes, nous utiliserons la fonction R, ARMAacf, pour illustrer notre propos.

Pour la AR(1), $Y_t = e_t + 0.6Y_{t-1}$, et la MA(1), $Y_t = e_t - 0.6e_{t-1}$, nous avons

```
> round(ARMAacf(ar=c(0.6), lag.max=10), digits=3)
  0   1   2   3   4   5   6   7   8   9   10
1.000 0.600 0.360 0.216 0.130 0.078 0.047 0.028 0.017 0.010 0.006

> round(ARMAacf(ar=c(0.6), lag.max=10, pacf=TRUE), digits=3)
[1] 0.6 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0

> round(ARMAacf(ma=c(0.6), lag.max=10), digits=3)
  0   1   2   3   4   5   6   7   8   9   10
1.000 0.441 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000

> round(ARMAacf(ma=c(0.6), lag.max=10, pacf=TRUE), digits=3)
[1] 0.441 -0.242 0.141 -0.083 0.050 -0.030 0.018 -0.011 0.006 -0.004
```

Pour la AR(2), $Y_t = e_t + 0.6Y_{t-1} - 0.4Y_{t-2}$, et la MA(2),
 $Y_t = e_t - 0.6e_{t-1} + 0.4e_{t-2}$, nous avons

```
> round(ARMAacf(ar=c(0.6,-0.4), lag.max=10), digits=3)
  0   1   2   3   4   5   6   7   8   9   10
1.000 0.429 -0.143 -0.257 -0.097 0.045 0.066 0.022 -0.013 -0.017 -0.005

> round(ARMAacf(ar=c(0.6,-0.4), lag.max=10, pacf=TRUE), digits=3)
[1] 0.429 -0.400 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000

> round(ARMAacf(ma=c(0.6,-0.4), lag.max=10), digits=3)
  0   1   2   3   4   5   6   7   8   9   10
1.000 0.237 -0.263 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000

> round(ARMAacf(ma=c(0.6,-0.4), lag.max=10, pacf=TRUE), digits=3)
[1] 0.237 -0.338 0.196 -0.189 0.149 -0.134 0.116 -0.105 0.095 -0.086
```

Nous avons donc ces comportements pour les fonctions ACF et PACF, pour les processus AR(p) et MA(q).

	AR(p)	MA(q)
ACF	$\downarrow 0$ quand $k \uparrow \infty$	$= 0$ pour $k > q$
PACF	$= 0$ pour $k > p$	$\downarrow 0$ quand $k \uparrow \infty$

(2) Nous pouvons montrer que la fonction ϕ_{kk} satisfait les équations de Yule-Walker, soit

$$\rho_j = \phi_{k1}\rho_{j-1} + \phi_{k2}\rho_{j-2} + \dots + \phi_{kk}\rho_{j-k} , \quad j = 1, 2, \dots, k$$

où

$$\phi_{kj} = \phi_{k-1,j} - \phi_{kk}\phi_{k-1,k-j} , \quad j = 1, 2, \dots, k-1 .$$

Vérifions cela pour $k = 2$.

$$\begin{aligned}\phi_{22} &= \frac{Cov(Y_t - \rho_1 Y_{t-1}, Y_{t-2} - \rho_1 Y_{t-1})}{\sqrt{Var(Y_t - \rho_1 Y_{t-1})} \sqrt{Var(Y_{t-2} - \rho_1 Y_{t-1})}} \\ &= \frac{\gamma_2 - 2\rho_1\gamma_1 + \rho_1^2\gamma_0}{\gamma_0 + \rho_1^2\gamma_0 - 2\rho_1\gamma_1} \\ &= \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}\end{aligned}$$

De plus,

- $\phi_{21}\rho_0 + \phi_{22}\rho_1 = (1 - \phi_{22})\rho_1 + \phi_{22}\rho_1 = \rho_1 \quad ,$
- $\rho_2 = \phi_{21}\rho_1 + \phi_{22}\rho_0 = (1 - \phi_{22})\rho_1^2 + \phi_{22} \Rightarrow \phi_{22} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}$

(3) Nous pouvons aussi montrer que la fonction ϕ_{kk} satisfait une relation récursive (plus intéressante!), soit

$$\phi_{kk} = \frac{\rho_k - \sum_{j=1}^{k-1} \phi_{k-1,j} \rho_{k-j}}{1 - \sum_{j=1}^{k-1} \phi_{k-1,j} \rho_j} .$$

(4) Si nous remplaçons les ρ_j par les r_j , $j = 1, 2, \dots, k-1$, dans la formule précédente, nous obtiendrons ainsi la fonction d'autocorrélation partielle échantillonnale $\hat{\phi}_{kk}$.

(5) Si le modèle AR(p) est adéquat, alors pour n « grand » nous pouvons montrer que

$$\hat{\phi}_{kk} \sim AN\left(0, 1/n\right) , \quad k > p .$$

Ainsi, nous pourrons utiliser le test bilatéral usuel avec niveau α (et points critiques $\pm z_{\alpha/2} / \sqrt{n}$) pour vérifier l'hypothèse nulle versus l'hypothèse alternative ...

H_0 : AR(p) est un modèle approprié

H_1 : AR(p) n'est pas un modèle approprié

... tout comme nous l'avons fait pour le modèle MA(p) en utilisant les coefficients d'autocorrélation échantillonnale r_k .

6.3.2 La fonction d'autocorrélation étendue (EACF)

Dans la section précédente, nous avons développé les coefficients d'autocorrélation échantillonnale (ACF) et d'autocorrélation échantillonnale partiel (PACF) comme outils efficaces pour nous aider à savoir si nous pouvons modéliser notre série chronologique soit par un processus MA(q) ou soit par un processus AR(p).

Malheureusement ces outils peuvent difficilement nous aider dans le cas où notre série chronologique correspondrait à un processus ARMA(p,q), avec $p > 0, q > 0$. Nous porterons donc notre attention sur une autre méthode, basée sur des fonctions d'autocorrélation « étendue » (EACF), qui consistera d'abord à « estimer » la partie AR du processus ARMA (i.e. ses coefficients) de sorte que à ce que le résidu engendré par cet estimé corresponde approximativement à un modèle MA. Une suite de régressions sera habituellement nécessaire pour obtenir un modèle ARMA adéquat.

Idée de base de la méthode :

Un processus ARMA(p,q) stationnaire $\{Y_t; t \in \mathbb{Z}\}$ peut être exprimé, à l'aide des polynômes caractéristiques ϕ et θ , et de l'opérateur de retard B, comme suit

$$\phi(B)Y_t = \theta(B)e_t ,$$

où $\{e_t; t \in \mathbb{Z}\} \sim WN(0, \sigma_e^2)$, $E[Y_t] = 0$, e_t est indépendant de Y_{t-1}, Y_{t-2}, \dots

C'est donc dire que le processus généré par

$$W_t = \phi(B)Y_t = Y_t - \phi_1 Y_{t-1} - \dots - \phi_p Y_{t-p}$$

suit un modèle MA(q).

Si nous connaissons les paramètres ϕ_1, \dots, ϕ_p , il ne resterait donc qu'à évaluer la fonction d'autocorrélation échantillonnale de W_t pour estimer le nombre (q) de paramètres de la partie MA. Or, en pratique, nous n'avons entre les mains qu'une série chronologique et donc les valeurs de ces paramètres (et même le nombre de ceux-ci) nous sont inconnues.

L'idée serait donc de faire d'abord une hypothèse sur le nombre p de paramètres de la partie AR (nous commençons par de petits nombres, soit 1, 2, 3, ...), puis nous estimons le « mieux possible » les paramètres ϕ_1, \dots, ϕ_p de telle sorte que nos estimateurs $\hat{\phi}_1, \dots, \hat{\phi}_p$ soient des estimateurs consistants de ceux-ci (i.e. que $\hat{\phi}_i \rightarrow \phi_i$ quand $n \rightarrow \infty$ avec probabilité 1, $i = 1, \dots, p$).

Nous nous servirons ensuite de l'erreur sur cette estimation de la partie AR(p) ...

$$\hat{W}_t = Y_t - \hat{\phi}_1 Y_{t-1} - \dots - \hat{\phi}_p Y_{t-p}$$

... pour générer une nouvelle série de n valeurs à partir de laquelle nous estimerons le nombre de paramètres q et ce en calculant les fonctions d'autocorrélation échantillonnelles « étendues » (ESACF).

Une première illustration de la méthode :

Supposons que le vrai modèle qui se cache derrière notre série chronologique est un processus stationnaire ARMA(1,1)

$$Y_t = \phi Y_{t-1} + e_t - \theta e_{t-1} \quad .$$

Alors une première régression linéaire de Y_t sur Y_{t-1} par la méthode des moindres carrés, sans ordonnée à l'origine (intercept), nous donnerait

$$\hat{\beta}_1^{(1)} Y_{t-1} , \text{ où } \hat{\beta}_1^{(1)} = \sum_{t=1}^n Y_t Y_{t-1} \Bigg/ \sum_{t=1}^n Y_{t-1}^2 .$$

De plus, quand $n \rightarrow \infty$, $\hat{\beta}_1^{(1)}$ doit converger vers ϕ pour être consistant. Or, si le vrai processus qui représente notre série chronologique est le processus décrit par $Y_t = \phi Y_{t-1} + e_t - \theta e_{t-1}$, nous avons que $\hat{\beta}_1^{(1)} \rightarrow \rho_1$ quand $n \rightarrow \infty$, où

$$\rho_1 = \frac{(\phi - \theta)(1 - \phi\theta)}{1 - 2\phi\theta + \theta^2} \neq \phi \text{ (sinon } \phi = 1\text{)} ,$$

et donc $\hat{\beta}_1^{(1)}$ n'est pas un estimateur consistant de ϕ .

Si $\varepsilon_t^{(1)}$ est le résidu de cette régression, soit

$$Y_t - \hat{\beta}_1^{(1)} Y_{t-1} = \varepsilon_t^{(1)} ,$$

alors $\varepsilon_t^{(1)}$ ne peut générer le « résidu » MA(1) puisque, avec l'hypothèse du processus ARMA(1,1), nous avons

$$\varepsilon_t^{(1)} = (\phi - \hat{\beta}_1^{(1)}) Y_{t-1} + e_t - \theta e_{t-1} .$$

Cependant, nous observons que $\varepsilon_t^{(1)}$ contient de l'information sur le processus d'erreur généré par e_t (donc sur la valeur de q). Nous allons donc faire une deuxième régression linéaire sur Y_{t-1} et $\varepsilon_{t-1}^{(1)}$, i.e.

$$Y_t = \hat{\beta}_1^{(2)} Y_{t-1} + \hat{\beta}_2^{(2)} \varepsilon_{t-1}^{(1)} .$$

Nous obtenons ainsi,

$$\hat{\beta}_1^{(2)} = \frac{\sum_{t=1}^n Y_t Y_{t-1} \sum_{t=1}^n (\varepsilon_{t-1}^{(1)})^2 - \sum_{t=1}^n Y_t \varepsilon_{t-1}^{(1)} \sum_{t=1}^n Y_{t-1} \varepsilon_{t-1}^{(1)}}{\sum_{t=1}^n Y_{t-1}^2 \sum_{t=1}^n (\varepsilon_{t-1}^{(1)})^2 - \left[\sum_{t=1}^n Y_{t-1} \varepsilon_{t-1}^{(1)} \right]^2},$$

$$\hat{\beta}_2^{(2)} = \frac{\sum_{t=1}^n Y_t Y_{t-1} \sum_{t=1}^n Y_{t-1} \varepsilon_{t-1}^{(1)} - \sum_{t=1}^n Y_t \varepsilon_{t-1}^{(1)} \sum_{t=1}^n Y_{t-1}^2}{\left[\sum_{t=1}^n Y_{t-1} \varepsilon_{t-1}^{(1)} \right]^2 - \sum_{t=1}^n Y_{t-1}^2 \sum_{t=1}^n (\varepsilon_{t-1}^{(1)})^2}.$$

Maintenant, est-ce que $\hat{\beta}_1^{(2)}$ est un estimateur consistant de ϕ ? La réponse est oui, mais nous allons devoir le démontrer avec la formule obtenue pour $\hat{\beta}_1^{(2)}$.

Ainsi, en observant que $\varepsilon_{t-1}^{(1)} = Y_t - \hat{\beta}_1^{(1)}Y_{t-1}$ converge vers $Y_t - \rho_1 Y_{t-1}$ quand $n \rightarrow \infty$, alors pour n « grand »

$$\begin{aligned}\hat{\beta}_1^{(2)} &= \frac{\sum_{t=1}^n Y_t Y_{t-1} \sum_{t=1}^n (\varepsilon_{t-1}^{(1)})^2 - \sum_{t=1}^n Y_t \varepsilon_{t-1}^{(1)} \sum_{t=1}^n Y_{t-1} \varepsilon_{t-1}^{(1)}}{\sum_{t=1}^n Y_{t-1}^2 \sum_{t=1}^n (\varepsilon_{t-1}^{(1)})^2 - \left[\sum_{t=1}^n Y_{t-1} \varepsilon_{t-1}^{(1)} \right]^2} \\ &\approx \frac{\sum_{t=1}^n Y_t Y_{t-1} \sum_{t=1}^n (Y_{t-1} - \rho_1 Y_{t-2})^2 - \sum_{t=1}^n Y_t (Y_{t-1} - \rho_1 Y_{t-2}) \sum_{t=1}^n Y_{t-1} (Y_{t-1} - \rho_1 Y_{t-2})}{\sum_{t=1}^n Y_{t-1}^2 \sum_{t=1}^n (Y_{t-1} - \rho_1 Y_{t-2})^2 - \left[\sum_{t=1}^n Y_{t-1} (Y_{t-1} - \rho_1 Y_{t-2}) \right]^2} \\ &\approx \frac{\gamma_1 (\gamma_0 - 2\rho_1 \gamma_1 + \rho_1^2 \gamma_0) - (\gamma_1 - \rho_1 \gamma_2)(\gamma_0 - \rho_1 \gamma_1)}{\gamma_0 (\gamma_0 - 2\rho_1 \gamma_1 + \rho_1^2 \gamma_0) - (\gamma_0 - \rho_1 \gamma_1)^2}\end{aligned}$$

Et donc,

$$\begin{aligned}\hat{\beta}_1^{(2)} &\approx \frac{-\rho_1\gamma_1^2 + \rho_1^2\gamma_0\gamma_1 + \rho_1\gamma_0\gamma_2 - \rho_1^2\gamma_1\gamma_2}{\rho_1^2(\gamma_0^2 - \gamma_1^2)} \\ &\approx \frac{\rho_2(1 - \rho_1^2)}{\rho_1(1 - \rho_1^2)} \\ &\approx \frac{\phi\rho_1}{\rho_1} \\ &\approx \phi\end{aligned}$$

Ainsi le « résidu » $W_t = Y_t - \hat{\beta}_1^{(2)}Y_{t-1}$ devrait constituer une bonne approximation du processus MA(1) ce qui nous permettra d'estimer le paramètre θ correspondant en se servant de la nouvelle série chronologique générée par W_t .

Qu'arrive-t-il si le vrai modèle qui se cache derrière notre série chronologique est plutôt un processus stationnaire ARMA(1,2)

$$Y_t = \phi Y_{t-1} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} .$$

Nous pourrions montrer (mais ce serait laborieux ...) qu'il faut passer par trois régressions linéaires tel que ...

- $Y_t = \hat{\beta}_1^{(1)} Y_{t-1} + \varepsilon_t^{(1)}$
- $Y_t = \hat{\beta}_1^{(2)} Y_{t-1} + \hat{\beta}_2^{(2)} \varepsilon_{t-1}^{(1)} + \varepsilon_t^{(2)}$
- $Y_t = \hat{\beta}_1^{(3)} Y_{t-1} + \hat{\beta}_2^{(3)} \varepsilon_{t-2}^{(1)} + \hat{\beta}_3^{(3)} \varepsilon_{t-1}^{(2)} + \varepsilon_t^{(3)}$

où, quand $n \rightarrow \infty$,

$$\beta_1^{(1)} \rightarrow \rho_1 \neq \phi , \quad \beta_1^{(2)} \rightarrow \frac{\rho_2}{\rho_1} \neq \phi , \quad \beta_1^{(3)} \rightarrow \phi .$$

Construction générale de la méthode :

(1) Nous utilisons la méthode des moindres carrés pour approximer Y_t par l'estimateur linéaire $\beta_1^{(1)}Y_{t-1} + \dots + \beta_p^{(1)}Y_{t-p}$, et nous obtenons

$$Y_t = \beta_1^{(1)}Y_{t-1} + \dots + \beta_p^{(1)}Y_{t-p} + \varepsilon_t^{(1)} ,$$

où $\varepsilon_t^{(1)}$ est un terme d'erreur.

Si le processus qui se cache derrière notre série chronologique est une ARMA(p,q), alors ces premiers estimés $\beta_1^{(1)}, \dots, \beta_p^{(1)}$ ne peuvent être consistants. Ainsi $\varepsilon_t^{(1)}$ ne peut générer une MA(q) et contient toujours de l'information sur le processus d'erreur généré par e_t (et donc sur la valeur de q).

(2) Puisque $\varepsilon_t^{(1)}$ contient de l'information sur la valeur de q, nous ferons une seconde régression linéaire sur Y_{t-1}, \dots, Y_{t-p} et $\varepsilon_{t-1}^{(1)}$, pour ainsi obtenir

$$Y_t = \beta_1^{(2)} Y_{t-1} + \dots + \beta_p^{(2)} Y_{t-p} + \beta_{p+1}^{(2)} \varepsilon_{t-1}^{(1)} + \varepsilon_t^{(2)} ,$$

où $\varepsilon_t^{(2)}$ est un second terme d'erreur.

Si l'ordre réel de la partie MA qui se cache derrière les données est $q = 1$, alors les estimés $\beta_1^{(2)}, \dots, \beta_p^{(2)}$ seront consistants.

Si $q > 1$, alors ces mêmes estimés ne seront pas consistants et le terme d'erreur $\varepsilon_t^{(2)}$ contiendra toujours de l'information sur le processus d'erreur généré par e_t (et donc sur la valeur de q).

(3) Si $q > 1$, nous ferons une troisième régression linéaire sur $Y_{t-1}, \dots, Y_{t-p}, \varepsilon_{t-2}^{(1)}$ et $\varepsilon_{t-1}^{(2)}$, pour ainsi obtenir

$$Y_t = \beta_1^{(3)} Y_{t-1} + \dots + \beta_p^{(3)} Y_{t-p} + \beta_{p+1}^{(3)} \varepsilon_{t-2}^{(1)} + \beta_{p+2}^{(3)} \varepsilon_{t-1}^{(2)} + \varepsilon_t^{(3)},$$

où $\varepsilon_t^{(3)}$ est un troisième terme d'erreur.

Si l'ordre réel de la partie MA qui se cache derrière les données est $q = 2$, alors les estimés $\beta_1^{(3)}, \dots, \beta_p^{(3)}$ seront consistants.

Si $q > 2$, alors ces mêmes estimés ne seront pas consistants et le terme d'erreur $\varepsilon_t^{(3)}$ contiendra toujours de l'information sur le processus d'erreur généré par e_t (et donc sur la valeur de q).

(4) Si $q > 2$, nous continuons le processus itératif avec une quatrième régression linéaire sur $Y_{t-1}, \dots, Y_{t-p}, \varepsilon_{t-3}^{(1)}, \varepsilon_{t-2}^{(2)}$ et $\varepsilon_{t-1}^{(3)}$, pour ainsi obtenir

$$Y_t = \beta_1^{(4)} Y_{t-1} + \dots + \beta_p^{(4)} Y_{t-p} + \beta_{p+1}^{(4)} \varepsilon_{t-3}^{(1)} + \beta_{p+2}^{(4)} \varepsilon_{t-2}^{(2)} + \beta_{p+3}^{(4)} \varepsilon_{t-1}^{(3)} + \varepsilon_t^{(4)} ,$$

où $\varepsilon_t^{(4)}$ est un troisième terme d'erreur.

Et nous poursuivons la même analyse que précédemment, jusqu'à ce que nous parvenions à des estimateurs consistants des coefficients de la partie AR de notre processus ARMA(p,q).

Fonction d'autocorrélation échantillonnale des « résidus »

Puisque les ordres p et q de notre modèle ARMA(p,q) nous sont inconnus, nous aurons donc à les estimer en nous basant sur la stratégie précédente. Ainsi, nous définissons la i -ième fonction d'autocorrélation (échantillonnale) étendue (E(S)ACF) d'écart j

$$\hat{\rho}_{i,j} , \quad i = 0, 1, 2, \dots , \quad j = 1, 2, \dots$$

comme étant la fonction d'autocorrélation (échantillonnale) du processus résiduel

$$\hat{W}_{t,i,j} = Y_t - \left(\beta_1^{(j)} Y_{t-1} + \dots + \beta_i^{(j)} Y_{t-i} \right) ,$$

où i correspond à l'ordre du processus AR considéré et j correspond à la j -ième régression $\beta_1^{(j)} Y_{t-1} + \dots + \beta_i^{(j)} Y_{t-i}$.

Les valeurs ainsi calculées pour une série temporelle apparaissent sous la forme d'un tableau dont les dimensions sont habituellement « raisonnables ».

	MA					
AR	0	1	2	3	...	
0	$\hat{\rho}_{0,1}$	$\hat{\rho}_{0,2}$	$\hat{\rho}_{0,3}$	$\hat{\rho}_{0,4}$...	
1	$\hat{\rho}_{1,1}$	$\hat{\rho}_{1,2}$	$\hat{\rho}_{1,3}$	$\hat{\rho}_{1,4}$...	
2	$\hat{\rho}_{2,1}$	$\hat{\rho}_{2,2}$	$\hat{\rho}_{2,3}$	$\hat{\rho}_{2,4}$...	
3	$\hat{\rho}_{3,1}$	$\hat{\rho}_{3,2}$	$\hat{\rho}_{3,3}$	$\hat{\rho}_{3,4}$...	
:	:	:	:	:	...	

Remarques : (1) Les valeurs de la k-ième colonne correspondent simplement au coefficient d'autocorrélation (échantillonnal) ACF d'ordre k de la partie MA(k-1) de la série chronologique.

(2) Mathématiquement, lorsque $n \rightarrow \infty$, il peut être démontré qu'avec probabilité 1,

- $\hat{\rho}_{i,j} \rightarrow 0$, $0 \leq i-p < j-q$

- $\hat{\rho}_{i,j} \rightarrow d_{ij} \neq 0$, autrement

(3) Le EACF (théorique) d'une ARMA(1,1) se présenterait comme suit

	MA					
AR	0	1	2	3	...	
0	x	x	x	x	...	
1	x	0	0	0	...	
2	x	x	0	0	...	
3	x	x	x	0	...	
:	:	:	:	:	...	

Les x correspondent aux limites non nulles des $\hat{\rho}_{i,j}$ et les 0 correspondent aux limites nulles des $\hat{\rho}_{i,j}$. La « pointe » de la région angulaire formée par les 0, correspondant aux coordonnées $(i,j) = (1,2)$, indique que nous avons affaire à un modèle ARMA(1,1).

Le EACF (théorique) d'une ARMA(2,1) se présenterait quant à lui comme suit

		MA					
AR	0	1	2	3	...		
0	x	x	x	x	...		
1	x	x	x	x	...		
2	x	0	0	0	...		
3	x	x	0	0	...		
4	x	x	x	0			
:	:	:	:	:	...		

(4) Notez que les EACF échantillonaux ne donneront pas une région triangulaire aussi bien définie. Il faudra alors vérifier si certaines valeurs sont significativement différentes de 0. Ainsi, il peut être démontré que, lorsque $n \rightarrow \infty$,

$$\hat{\rho}_{i,j} \sim AN\left(0, \frac{1}{n-i-j}\right) ,$$

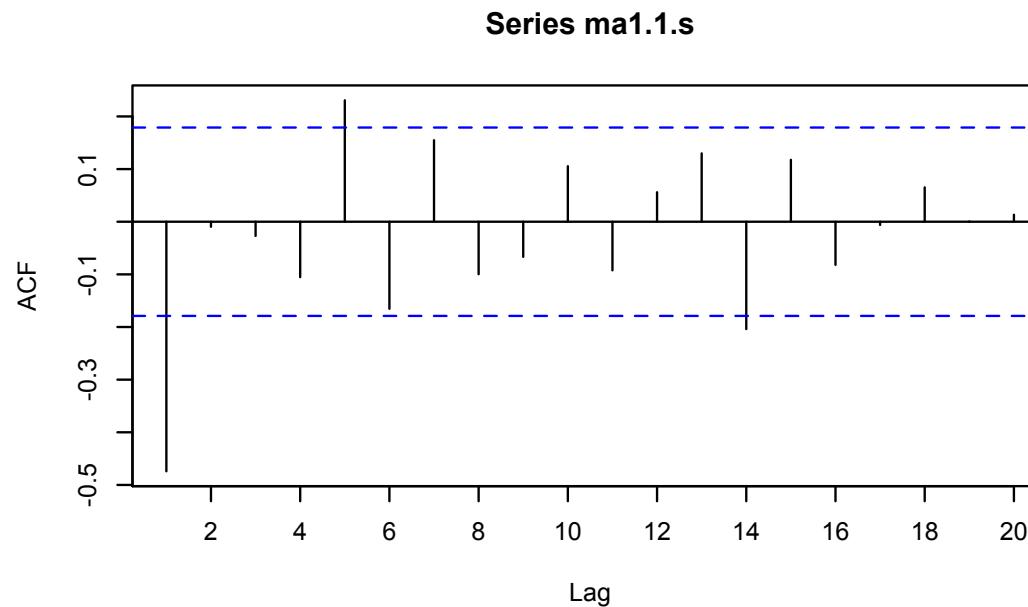
i.e. qu'on s'attend à ce que 95% des valeurs de ce coefficient se retrouvent à $\pm 1.96/\sqrt{n-i-j}$ de 0.

(5) Les calculs du EACF sont heureusement pris en charge par de nombreux logiciels, dont le R avec la fonction `> eacf(....)`, fonction que nous illustrerons dans les sections suivantes par des exemples soit simulés ou provenant de bases de données.

6.4 Spécifications de quelques séries chronologiques simulées

6.4.1 Une MA(1), avec $\theta = 0.9$ et $n = 120$

```
> data(ma1.1.s)
> win.graph(width=4.875, height=3, pointsize=8)
> acf(ma1.1.s, xaxp=c(0,20,10))
```



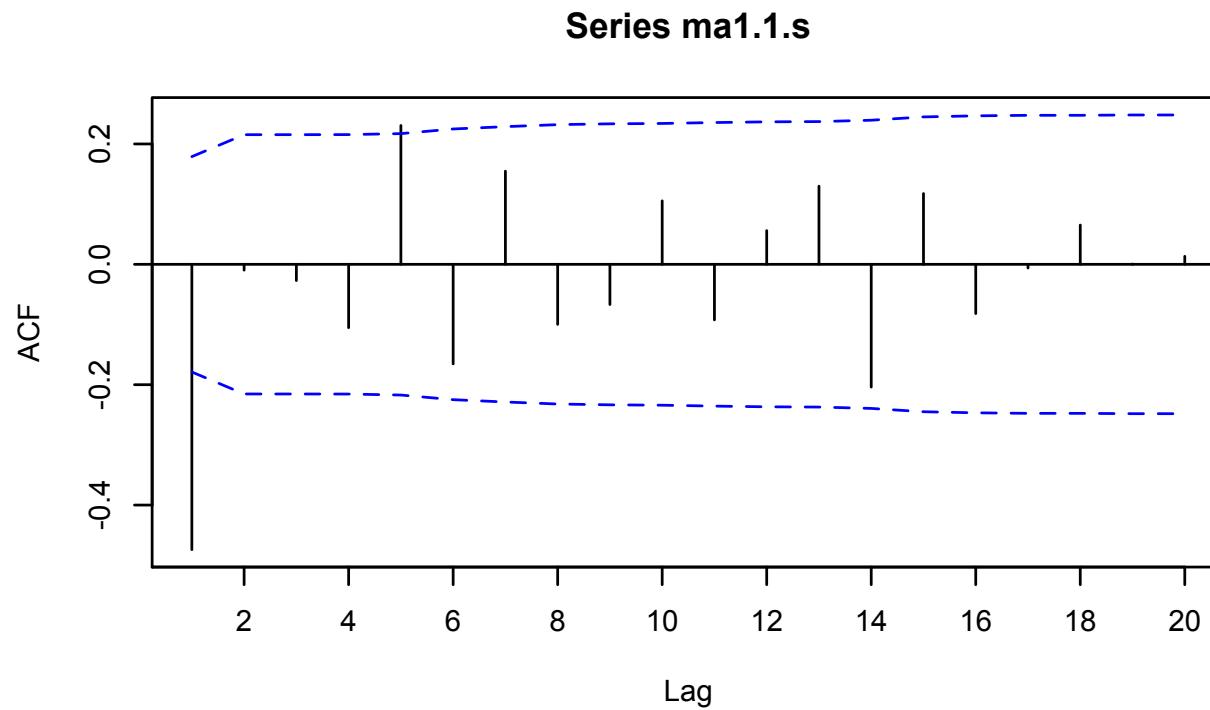
Le graphique précédent exhibe une zone limitée par des pointillés à une distance de $2/\sqrt{n} = 2/\sqrt{120} \approx 0.1826$ du 0, ce qui nous aide quelque peu à valider si les coefficients r_k sont significativement différents de 0.

L'écart-type utilisé (pour n grand, soit $1/\sqrt{n}$) corresponds cependant à celui d'un bruit blanc (voir l'exemple en 6.2.3-(1)), qui est utilisé par défaut par le logiciel R. Il est une approximation grossière des écarts-type réels correspondant aux r_k du processus MA(1) dont les valeurs exhibées dans l'exemple en 6.2.3-(3) sont uniquement les valeurs asymptotiques (n grand), soit

$$\sigma_{r_1} \approx \sqrt{\frac{1 - 3\rho_1^2 + 4\rho_1^4}{n}} , \quad \sigma_{r_k} \approx \sqrt{\frac{1 + 2\rho_1^2}{n}} , \quad k > 1 ,$$

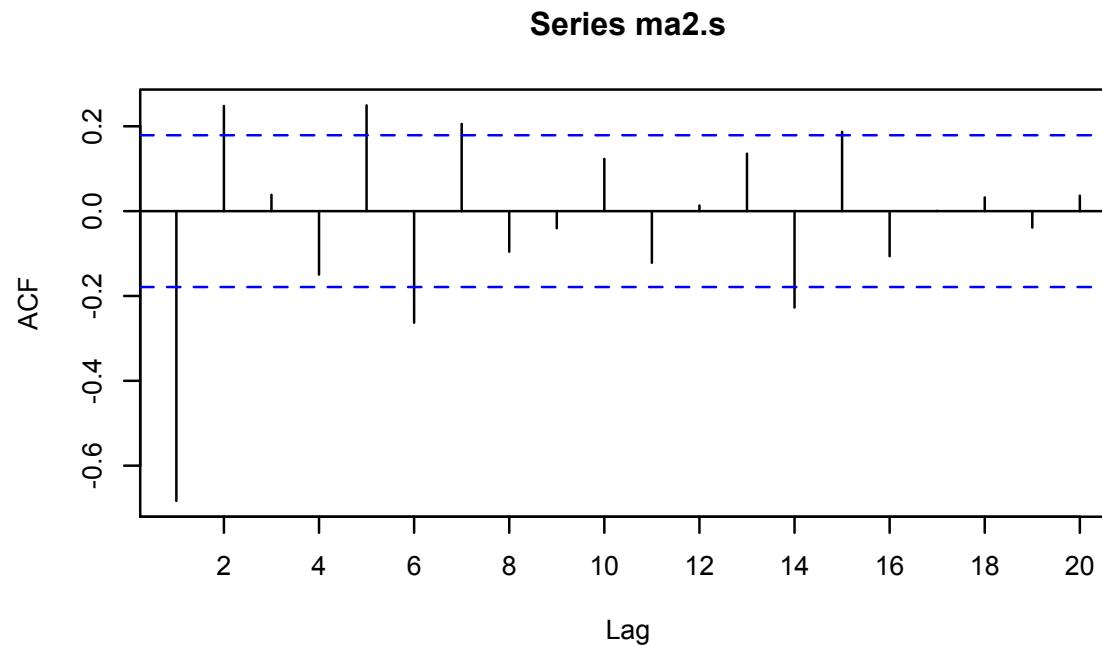
ce qui nous donnerait une zone plus irrégulière limitée par les pointillés décrits ci-après, et plus concluante pour la validation des coefficients r_k ...

```
> acf(ma1.1.s, ci.type='ma', xaxp=c(0,20,10))
```



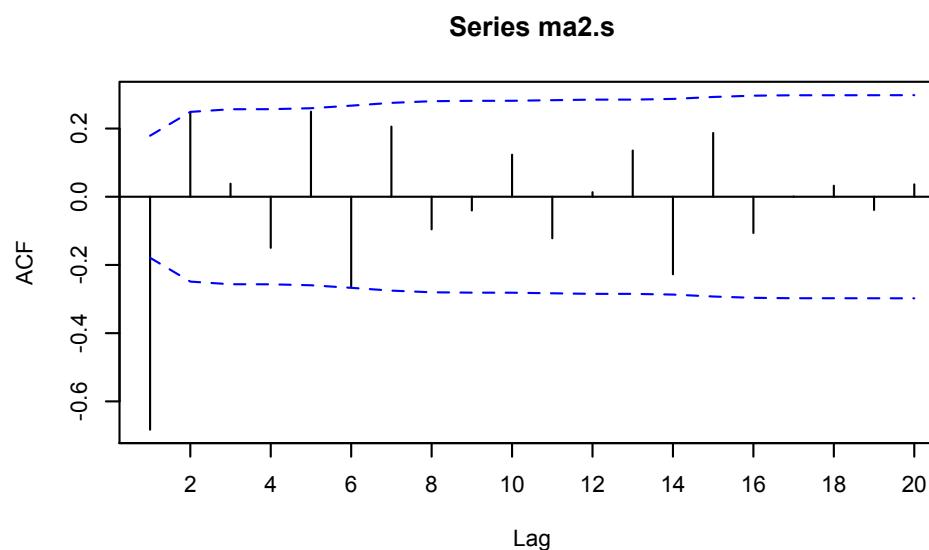
6.4.2 Une MA(2), avec $\theta_1 = 1$, $\theta_2 = -0.6$ et $n = 120$

```
> data(ma2.s)
> win.graph(width=4.875, height=3, pointsize=8)
> acf(ma2.s, xaxp=c(0,20,10))
```



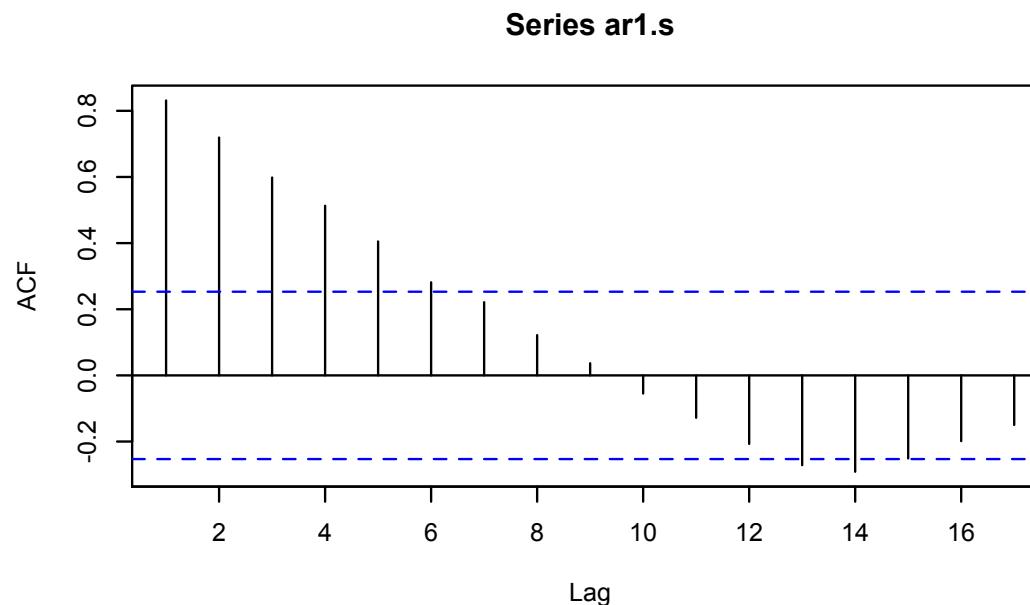
À nouveau, le graphique précédent exhibe une zone similaire à celle de l'exemple précédent pour la validation des r_k . En se basant plutôt sur les écarts-type réels correspondant aux r_k du processus MA(2), dont les valeurs asymptotiques sont données dans l'exemple en 6.2.3-(4), nous obtiendrons une zone plus concluante pour la validation des coefficients r_k ...

```
> acf(ma2.s, ci.type='ma', xaxp=c(0,20,10))
```



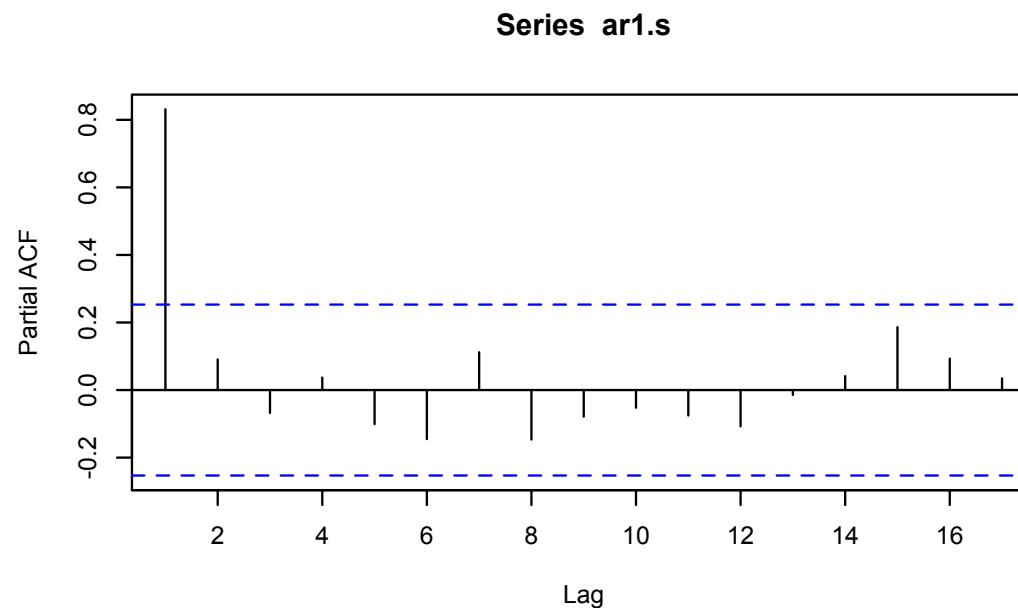
6.4.3 Une AR(1), avec $\phi = 0.9$ et $n = 60$

```
> data(ar1.s)
> win.graph(width=4.875, height=3, pointsize=8)
> acf(ar1.s, xaxp=c(0,20,10))
```



Comme attendu, peu d'informations peuvent être extraites du graphique précédent, en rapport avec la fonction ACF. Nous allons donc examiner plutôt la fonction PACF.

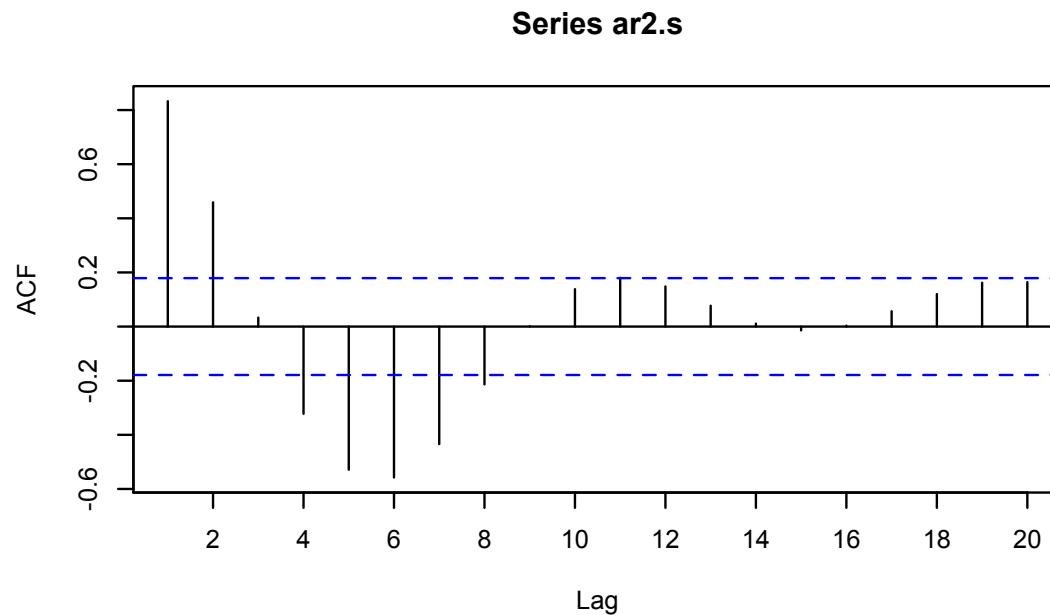
```
> pacf(ar1.s, xaxp=c(0,20,10))
```



Le graphique suggère fortement le comportement d'une AR(1).

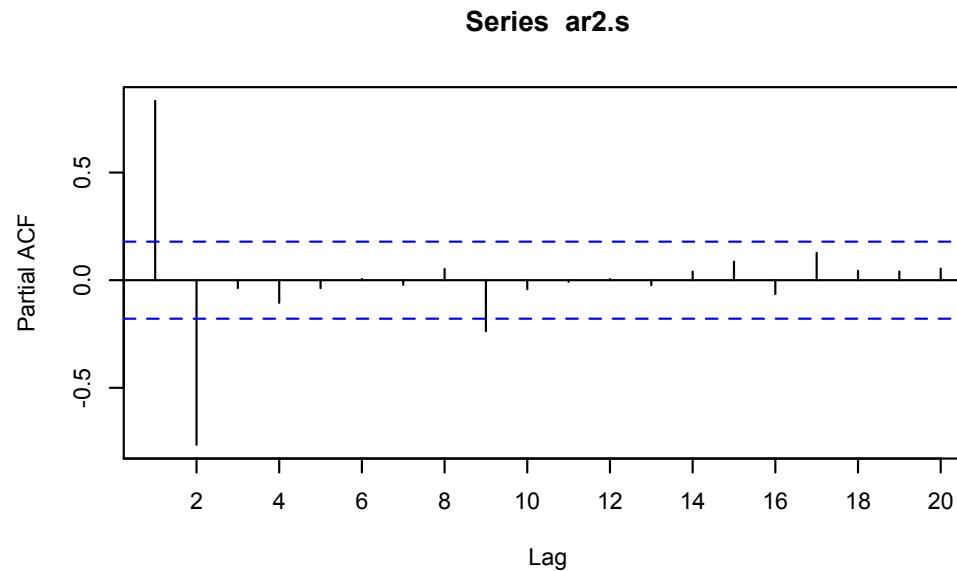
6.4.4 Une AR(2), avec $\phi_1 = 1.5$, $\phi_2 = -0.75$ et $n = 120$

```
> data(ar2.s)
> win.graph(width=4.875, height=3, pointsize=8)
> acf(ar2.s, xaxp=c(0,20,10))
```



À nouveau, peu d'informations peuvent être obtenues de ce graphique de la fonction ACF. Nous allons donc examiner plutôt la fonction PACF.

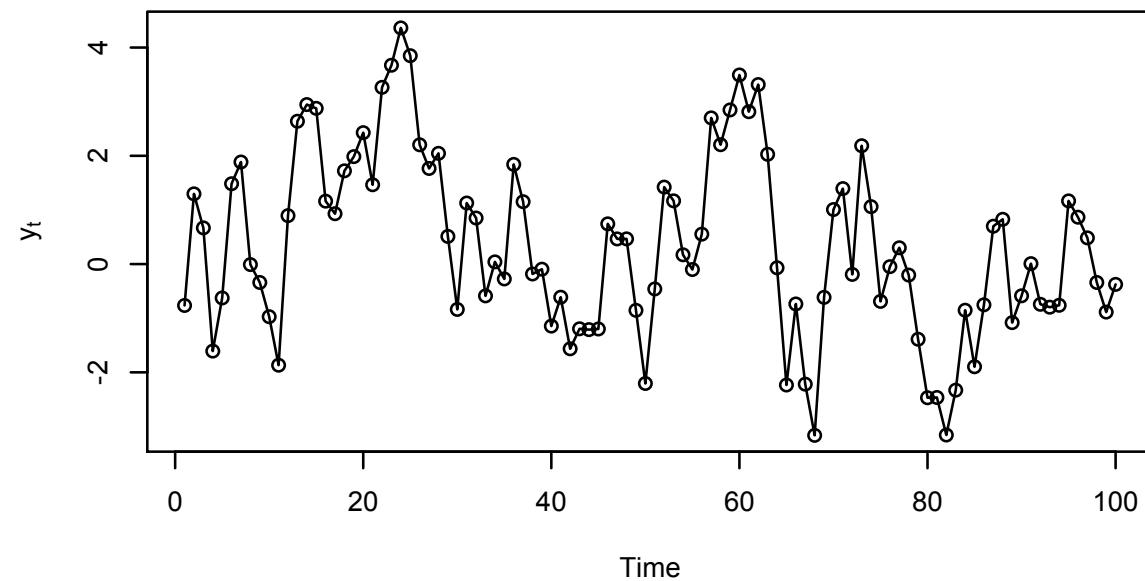
```
> pacf(ar2.s, xaxp=c(0,20,10))
```



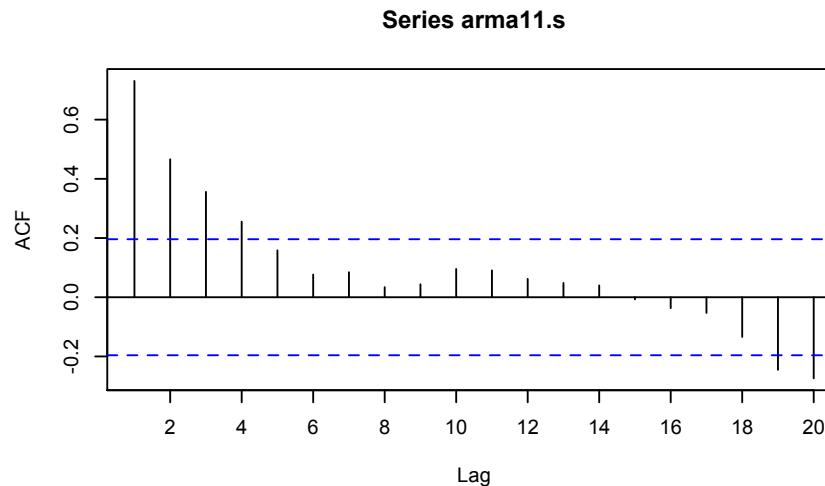
Le graphique suggère fortement le comportement d'une AR(2).

6.4.4 Une ARMA(1,1), avec $\phi = 0.6$, $\theta = -0.3$ et $n = 100$

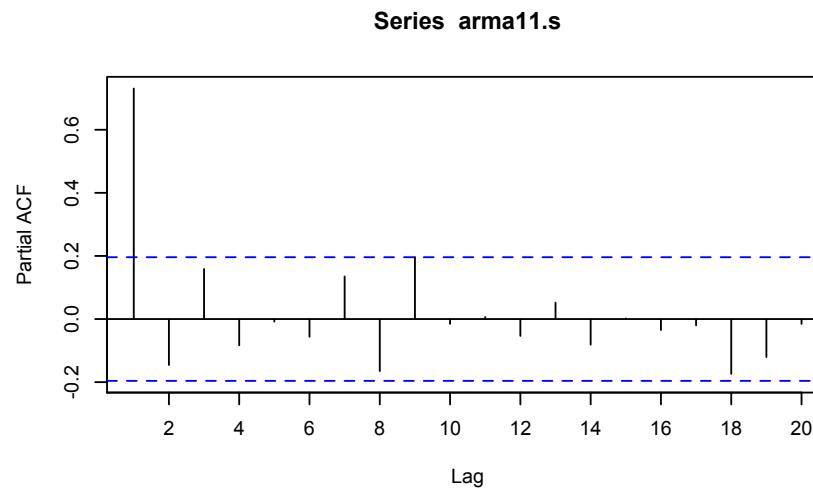
```
> data(arma11.s)
> win.graph(width=4.875, height=3, pointsize=8)
> plot(arma11.s, type='b', ylab=expression(y[t]))
```



```
> acf(arma11.s, xaxp=c(0,20,10))
```



```
> pacf(arma11.s, xaxp=c(0,20,10))
```



À nouveau, peu d'informations peuvent être obtenues du graphique de la fonction ACF. Le graphique de la fonction PACF semble suggérer un modèle AR(1). Nous allons vérifier cela avec le tableau de la fonction EACF.

```
> eacf(arma11.s)
```

AR/MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	x	x	x	o	o	o	o	o	o	o	o	o	o
1	x	o	o	o	o	o	o	o	o	o	o	o	o	o
2	x	o	o	o	o	o	o	o	o	o	o	o	o	o
3	x	x	o	o	o	o	o	o	o	o	o	o	o	o
4	x	o	x	o	o	o	o	o	o	o	o	o	o	o
5	x	o	o	o	o	o	o	o	o	o	o	o	o	o
6	x	o	o	o	x	o	o	o	o	o	o	o	o	o
7	x	o	o	o	x	o	o	o	o	o	o	o	o	o

Ce tableau suggère qu'une ARMA(1,1) ou une ARMA(2,1) seraient aussi de bons candidats pour modéliser nos données.

Quels nombres se cachent derrière le tableau précédent?

```
> res=eacf(arma11.s)
```

```
> res
```

```
$eacf
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]      [,8]      [,9]
[1,] 0.7306776 0.4660120 0.35608109 0.255438796 0.15855639 0.07671343 0.0845558970 0.03425112 0.04385366
[2,] 0.2095843 -0.2024517 0.03089843 0.076548162 0.08227227 -0.16809867 0.0888265424 -0.09530559 -0.01675742
[3,] 0.5040561 -0.1108708 0.06332267 0.012965322 0.02753716 -0.05821879 -0.0008396188 -0.06149982 -0.03426046
[4,] 0.3674307 0.2258370 0.13673940 -0.010634281 0.03502736 -0.04934294 -0.0020820351 -0.18003648 0.02060658
[5,] -0.2431391 -0.1967037 0.23577748 0.053755999 0.01355323 -0.02844293 -0.0086894767 -0.16966037 0.06603945
[6,] -0.2738476 -0.1045300 0.18663583 0.007894465 0.01097416 -0.03923744 -0.0021858164 -0.19103072 0.06846740
[7,] 0.3584490 0.1947224 0.16888088 0.003309865 0.30169475 -0.03965688 0.0138995618 -0.21468358 0.05780874
[8,] 0.5069406 0.0313976 -0.18568982 -0.019630435 0.37137623 0.11942316 -0.0015135671 -0.16198402 -0.02325274

      [,10]     [,11]     [,12]     [,13]     [,14]
[1,] 0.09543669 0.090861366 0.062083089 4.819929e-02 0.0399006628
[2,] 0.09515195 0.071992733 0.010955894 -3.116564e-05 0.0481609830
[3,] 0.07408780 0.034227976 0.003369403 4.803725e-04 0.0614502739
[4,] 0.05557071 0.047902402 -0.001928911 6.143485e-03 0.0628228467
[5,] 0.03508668 -0.003550315 0.015328572 -3.231288e-02 0.0006093876
[6,] 0.07683705 0.010722453 0.046820091 -3.232496e-02 -0.0042187234
[7,] 0.05679359 -0.004548155 0.027150408 -4.363450e-02 -0.0167052767
[8,] 0.10554957 0.084177867 0.051593290 -3.621192e-02 -0.0007006141
```

La i-ième ligne correspond à l'ordre i - 1 de la partie AR et la j-ième colonne correspond à l'ordre j - 1 de la partie MA.

Ces nombres représentent les calculs des coefficients d'autocorrélation étendus. Pour la correspondance avec le tableau des 0 et des x, rappelons que nous conclurons à un 0 si le résultat du calcul de $\hat{\rho}_{k,l}$ est dans l'intervalle de confiance bilatéral à 95% de la loi normale centrée à 0 et de variance $1/(n-k-l)$.

Prenons la donnée -0.2024517 (ligne 2, colonne 2). Pour déterminer s'il s'agit d'un 0, ce nombre doit se retrouver à une distance inférieure à $1.96/\sqrt{n-k-l} = 1.96/\sqrt{100-1-2} \approx 0.199$. C'est presqu'un 0, et il a été admis comme tel.

Prenons la donnée -0.2258370 (ligne 4, colonne 2). Pour déterminer s'il s'agit d'un 0, ce nombre doit se retrouver à une distance inférieure à $1.96/\sqrt{n-k-l} = 1.96/\sqrt{100-3-2} \approx 0.1949$. Ce nombre est plus hors de l'intervalle, c'est donc un x.

6.5 Non stationnarité

6.5.1 Surdifférenciation avec les processus ARIMA

Nous avons vu aux sections précédentes comment identifier les valeurs possibles des paramètres p et q d'une ARMA(p,q) stationnaire, soit

- Le coefficient ACF échantillonnal pour déterminer l'ordre q d'un processus MA(q).
- Le coefficient PACF échantillonnal pour déterminer l'ordre p d'un processus AR(p).
- Le coefficient EACF échantillonnal pour déterminer les ordres p et q d'un processus ARMA(p,q).

Nous avons aussi vu que les séries chronologiques non stationnaires présentaient souvent des tendances et/ou des variances non constantes. De plus nous avons observé que leurs coefficients ACF échantillonnaux montraient une décroissance très lente avec l'augmentation des écarts. Ainsi ...

- S'il y a une tendance claire dans la série chronologique (linéaire , quadratique, etc ...), si la variance est très irrégulière (augmente avec le temps, par exemple) et que la fonction ACF échantillonnale décroît très lentement, il faudrait alors considérer une première différence.
- Si la fonction ACF échantillonnale de cette première différence ressemble à un processus ARMA, i.e qu'elle décroît très rapidement (en valeur absolue), alors vous considérez un processus ARIMA($p,1,q$) et vous évaluez aussi les coefficients ACF, PACF et EACF sur cette première différence afin d'identifier les paramètres p et q .

- Si la fonction ACF échantillonnale de cette première différence décroît plutôt lentement, alors vous considérez une deuxième différence, et donc un processus ARIMA(p,2,q). Vous évaluez ensuite les coefficients ACF, PACF et EACF sur cette deuxième différence afin d'identifier les paramètres p et q.
- Et ainsi de suite ... malgré qu'il n'est pas si fréquent en pratique qu'on ait à utiliser une troisième différenciation.

Remarques : (1) Il peut être souvent approprié de transformer votre série chronologique avant d'utiliser la différenciation. Vous utiliserez alors la transformée Box-Cox.

(2) Il pourrait arriver, dans votre évaluation du modèle ARIMA, que vous alliez trop loin dans la différenciation, i.e. que vous choisissiez un « d » trop grand.

Par exemple, supposons qu'un modèle IMA(1,1) serait approprié pour une série chronologique, soit

$$Y_t = Y_{t-1} + e_t - \theta e_{t-1} ,$$

où $|\theta| < 1$, $e_t \sim WN(0, \sigma_e^2)$ et e_t est indépendant de Y_{t-1}, Y_{t-2}, \dots

La première différence donne (en terme de l'opérateur B)

$$\nabla Y_t = (1 - \theta B) e_t ,$$

laquelle est un processus MA(1) inversible.

Une deuxième différence donne

$$\nabla^2 Y_t = [1 - (1 + \theta)B + \theta B^2] e_t = (1 - B)(1 - \theta B) e_t .$$

Cette deuxième différence donne une MA(2) non inversible puisque 1 est une racine. Nous avons donc converti une MA(1) inversible en une MA(2) non inversible, ce qui aura comme conséquence que nous aurons maintenant 2 paramètres à estimer et qu'en plus ils ne seront pas uniques.

Il est donc possible que vous soyez confronté à des situations « limites », où le graphique de votre série chronologique et la fonction ACF échantillonnale soient difficiles à interpréter en terme de la stationnarité. Une alternative est à notre disposition pour tester si la série chronologique présente de la stationnarité ou non: le test (augmenté) de Dickey-Fuller.

6.5.2 Test (augmenté) de Dickey-Fuller (ADF)

Afin de motiver quelque peu la construction du test ADF, nous utiliserons le modèle particulier suivant

$$Y_t = \alpha Y_{t-1} + X_t ,$$

où X_t génère un processus stationnaire AR(k), tel que

$$X_t = \phi_1 X_{t-1} + \dots + \phi_k X_{t-k} + e_t .$$

Nous pouvons donc réécrire Y_t comme suit

$$\begin{aligned} Y_t &= \alpha Y_{t-1} + \phi_1 X_{t-1} + \dots + \phi_k X_{t-k} + e_t \\ &= \alpha Y_{t-1} + \phi_1 (Y_{t-1} - \alpha Y_{t-2}) + \dots + \phi_k (Y_{t-k} - \alpha Y_{t-k-1}) + e_t \end{aligned}$$

En terme de l'opérateur B , nous avons

$$\phi(B)(1-\alpha B)Y_t = e_t .$$

Ainsi, si $\alpha=1$, la série générée par Y_t est non stationnaire (et $\nabla Y_t = X_t$ est donc stationnaire). Nous pouvons alors écrire

$$\nabla Y_t = \phi_1 \nabla Y_{t-1} + \dots + \phi_k \nabla Y_{t-k} + e_t .$$

Si $|\alpha| < 1$, la série générée par Y_t est stationnaire (et Y_t génère alors une AR($k+1$)).

Ainsi, en posant $a = \alpha - 1$, nous pouvons construire un test de non-stationnarité basé sur la réécriture du modèle AR comme suit

$$\nabla Y_t = a Y_{t-1} + c_1 \nabla Y_{t-1} + \dots + c_{k-1} \nabla Y_{t-k} + e_t ,$$

i.e. en faisant une régression (des moindres carrés) sur ∇Y_t à partir de $Y_{t-1}, \nabla Y_{t-1}, \dots, \nabla Y_{t-k}$.

La statistique de Dickey-Fuller sera construite à partir de l'estimé \hat{a} du coefficient de Y_{t-1} , soit

$$t_{DF} = \frac{\hat{a}}{\hat{\sigma}_{\hat{a}}} ,$$

et elle servira à vérifier les hypothèses suivantes :

$H_0 : a = 0$ (non stationnarité \Rightarrow la série doit être différenciée)

$H_1 : a < 0$ (stationnarité \Rightarrow la série n'a pas à être différenciée)

Donc, si t_{DF} est significativement plus petit que 0 (à un niveau de confiance de 95%, par exemple), H_0 est rejeté et nous concluons que la série est stationnaire. Si t_{DF} n'est pas significativement plus petit que 0, nous concluons que la série est non stationnaire.

Remarques : (1) La statistique t_{DF} est similaire au test statistique de Student construit à partir de la méthode des moindres carrés mais sa distribution asymptotique (n grand) est plus complexe qu'une distribution de Student. Le logiciel R fournit heureusement une fonction (`> adf(...)`) qui opérera ce test pour nous.

(2) Le test ADF est un test unilatéral à gauche, car le graphique de la densité de cette statistique est asymétrique et a une queue plus allongée vers la gauche. Notons que ce test indique surtout la nécessité d'une (première) différenciation pour rendre stationnaire la série chronologique initiale.

(3) Le logiciel R a une fonction (`> ar(...)`) qui cherche l'ordre du modèle AR qui s'ajustera le mieux aux données d'une série chronologique.

(4) Le modèle utilisé pour illustrer la construction du test ADF est en fait un des cas particuliers du test ADF. Ce modèle est utilisé quand la série chronologique est relativement « aplatie », i.e. n'a pas de tendance linéaire (ou généralement polynomiale) et « gravite » autour de 0.

Un deuxième modèle à considérer est celui où la série chronologique est relativement aplatie, mais gravite autour d'une valeur non nulle. Nous obtenons alors comme modèle de construction

$$\nabla Y_t = a + bY_{t-1} + c_1\nabla Y_{t-1} + \dots + c_{k-1}\nabla Y_{t-k+1} + e_t .$$

Un troisième modèle à considérer est celui où la série chronologique a une tendance linéaire et gravite autour de sa droite (trend). Nous obtenons alors comme modèle de construction

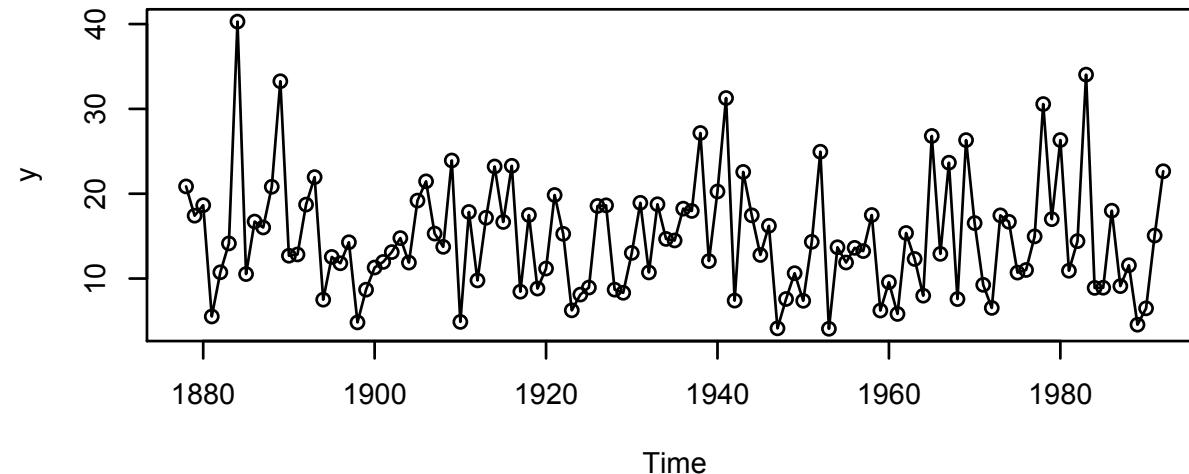
$$\nabla Y_t = a + bt + cY_{t-1} + c_1\nabla Y_{t-1} + \dots + c_{k-1}\nabla Y_{t-k+1} + e_t .$$

(5) Test ADF: valeurs critiques, avec probabilité à gauche

	n	1%	2.5%	5%	10%	90%	95%	97.5%	99%
Modèle 1	25	-2.66	-2.26	-1.95	-1.60	0.92	1.33	1.70	2.16
	50	-2.62	-2.25	-1.95	-1.61	0.91	1.31	1.66	2.08
	100	-2.60	-2.24	-1.95	-1.61	0.90	1.29	1.64	2.03
	250	-2.58	-2.23	-1.95	-1.61	0.89	1.28	1.63	2.01
	500	-2.58	-2.23	-1.95	-1.61	0.89	1.28	1.62	2.00
	> 500	-2.58	-2.23	-1.95	-1.61	0.89	1.28	1.62	2.00
Modèle 2	25	-3.75	-3.33	-3.00	-2.62	-0.37	0.00	0.34	0.72
	50	-3.58	-3.22	-2.93	-2.60	-0.40	-0.03	0.29	0.66
	100	-3.51	-3.17	-2.89	-2.58	-0.42	-0.05	0.26	0.63
	250	-3.46	-3.14	-2.88	-2.57	-0.42	-0.06	0.24	0.62
	500	-3.44	-3.13	-2.87	-2.57	-0.43	-0.07	0.24	0.61
	> 500	-3.43	-3.12	-2.86	-2.57	-0.44	-0.07	0.23	0.60
Modèle 3	25	-4.38	-3.95	-3.60	-3.24	-1.14	-0.80	-0.50	-0.15
	50	-4.15	-3.80	-3.50	-3.18	-1.19	-0.87	-0.58	-0.24
	100	-4.04	-3.73	-3.45	-3.15	-1.22	-0.90	-0.62	-0.28
	250	-3.99	-3.69	-3.43	-3.13	-1.23	-0.92	-0.64	-0.31
	500	-3.98	-3.68	-3.42	-3.13	-1.24	-0.93	-0.65	-0.32
	> 500	-3.96	-3.66	-3.41	-3.12	-1.25	-0.94	-0.66	-0.33

Exemple 1 : Considérer le fichier (larain) des quantités de pluie tombées annuellement à Los Angeles entre les années 1878 et 1992 ($\Rightarrow n = 115$).

```
> data(larain)
> win.graph(width=4.875, height=2.5, pointsize=8)
> plot(larain, type='o', ylab='y')
```



Avant d'utiliser le test ADF, nous allons déterminer l'ordre k du processus AR qui s'ajustera le mieux à `diff(larain)`. Pour ce faire, nous choisissons la fonction « > ar(...) » de la librairie « tseries ».

```
> library(tseries)
> ar(diff(larain))
```

Call:

```
ar(x = diff(larain))
```

Coefficients:

1	2	3	4
-0.8601	-0.6386	-0.4473	-0.2684

Order selected 4 sigma² estimated as 56.06

* R recommande d'utiliser k = 4.

Nous pouvons maintenant appliquer le test ADF ...

```
> adf.test(larain, k=4)
```

Augmented Dickey & Fuller test

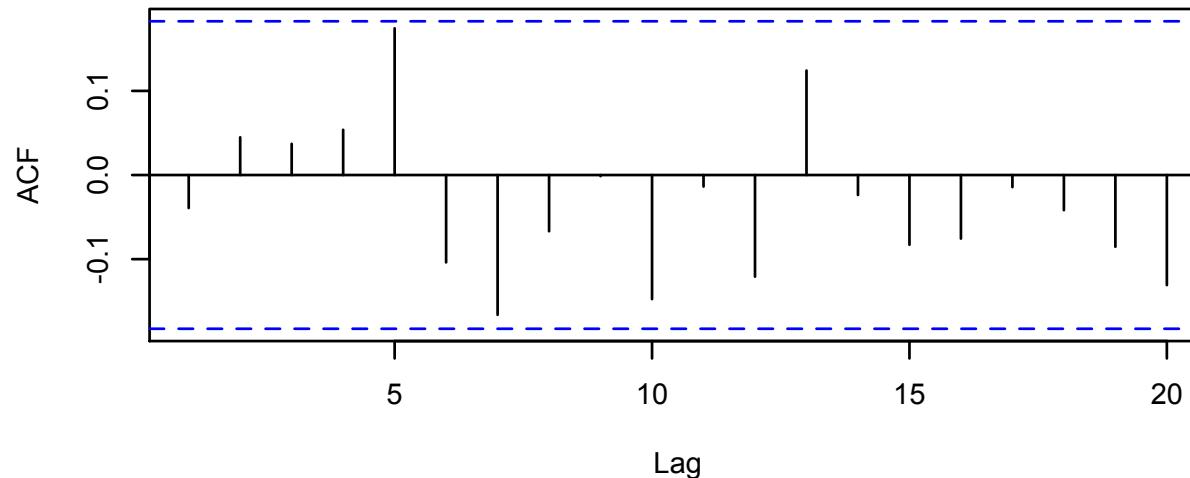
data : larain

Dickey-Fuller = -3.4351 , Lag order = 4 , p-value = 0.05218
alternative hypothesis : stationary

De ce test ADF à 95%, nous obtenons comme estimé (centré réduit) de a/σ_a , $t_{DF} = -3.4351$, avec une p-value de 0.05218. Le test semble corroborer que la série chronologique est non-stationnaire, i.e. qu'une première différenciation devrait s'appliquer, mais il n'est pas très convaincant.

Qu'en est-il de la fonction ACF de cette série?

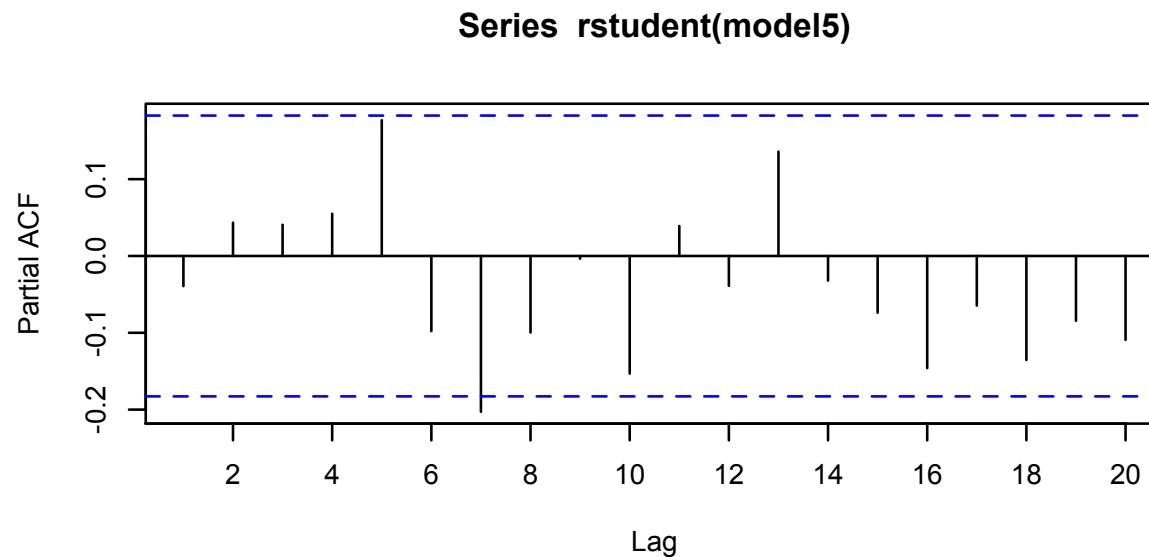
```
> model5=lm(larain~time(larain))  
> acf(rstudent(model5), main="")
```



Les coefficients d'autocorrélation r_k sont à $\pm 2/\sqrt{115} \approx 0.187$ de 0, ce qui correspond à 2 fois l'écart-type d'un bruit blanc soit une distance acceptable pour un bruit blanc.

Qu'en est-il de la fonction PACF de cette série?

```
> pacf(rstudent(model5), xaxp=c(0,20,10))
```



Les coefficients d'autocorrélation partielle ne viennent en rien contredire que nous puissions avoir affaire à un bruit blanc.

Qu'en est-il de la fonction EACF de cette série?

```
> eacf(rstudent(model5))
```

AR/MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	o	o	o	o	o	o	o	o	o	o	o	o	o	o
1	x	o	o	o	o	o	o	o	o	o	o	o	o	o
2	x	o	o	o	o	o	o	o	o	o	o	o	o	o
3	x	o	o	o	o	o	o	o	o	o	o	o	o	o
4	x	o	o	x	o	o	o	o	o	o	o	o	o	o
5	x	x	x	x	o	o	o	o	o	o	o	o	o	o
6	x	x	o	o	x	x	o	o	o	o	o	o	o	o
7	x	x	o	o	x	o	x	o	o	o	o	o	o	o

La fonction EACF montre clairement qu'un bruit blanc semble tout à fait approprié. Donc aucune nécessité de différencier la série chronologique « larain ».

Remarque : Reprenons le test ADF, mais à $k = 0$...

```
> adf.test(larain, k=0)
```

Augmented Dickey & Fuller test

data : larain

Dickey-Fuller = -10.899 , Lag order = 0 , p-value = 0.01

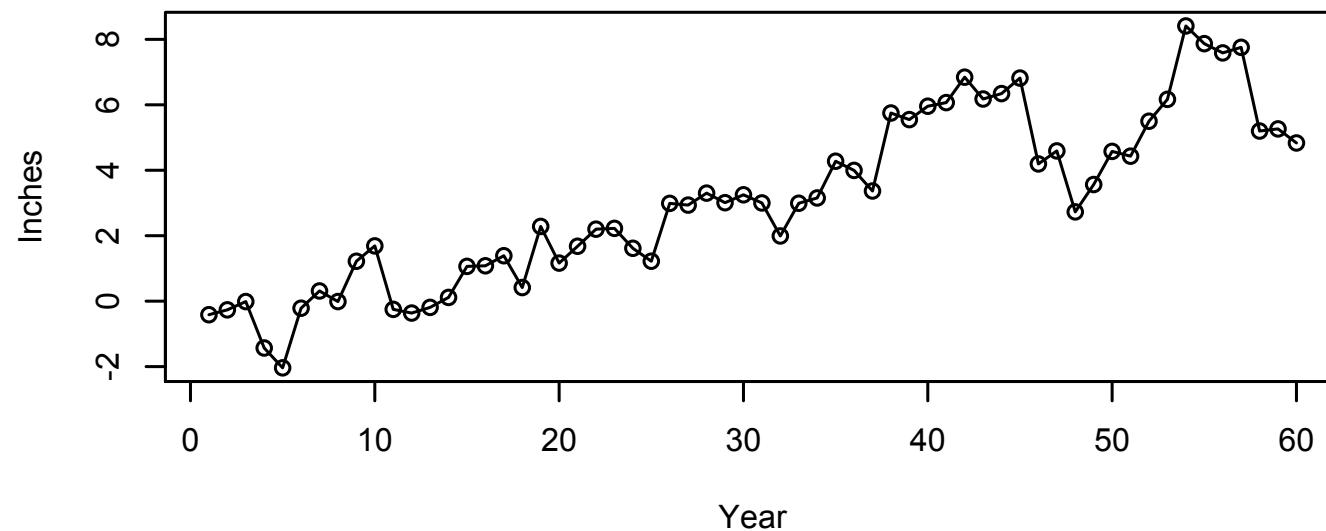
alternative hypothesis : stationary

De ce test ADF à 95%, nous obtenons comme estimé (centré réduit) de a/σ_a , $t_{DF} = -10.899$, avec une p-value de 0.01. Le test corrobore fortement que la série chronologique est stationnaire.

* Donc, il peut arriver que le test ADF ne départage pas bien entre la stationnarité ou la non stationnarité d'une série chronologique.

Exemple 2 : Considérer le fichier (rwalk) de la première marche aléatoire vue au chapitre 2, exemple 1, page 6 ($\Rightarrow n = 60$).

```
> data(rwalk)
> win.graph(width=4.875, height=2.5, pointsize=8)
> plot(rwalk, ylab='Inches', xlab='Year', type='o')
```



À nouveau, afin d'utiliser le test ADF, nous allons déterminer le l'ordre du processus AR qui s'ajustera le mieux à `diff(rwalk)`.

```
> library(tseries)  
> ar(diff(rwalk))
```

Call:

```
ar(x = diff(rwalk))
```

Coefficients:

1	2	3	4	5	6	7	8
-0.1622	-0.1234	0.0036	-0.1961	-0.1597	-0.3418	0.0652	-0.3228

Order selected 8 sigma² estimated as 0.8394

* R recommande d'utiliser k = 8.

Nous appliquons le test ADF (avec une constante, modèle 2) ...

```
> adf.test(rwalk, k=8)
```

Augmented Dickey & Fuller test

data : rwalk

Dickey-Fuller = -2.2892 , Lag order = 8 , p-value = 0.4579
alternative hypothesis : stationary

De ce test ADF à 95%, nous obtenons comme estimé (centré réduit) de b/σ_b , $t_{DF} = -2.2892$, avec une p-value de 0.4579. Le test corrobore fortement que la série chronologique est non-stationnaire, et il y a donc nécessité de différencier la série chronologique « rwalk ».

6.6 Autres méthodes de spécification

6.6.1 Le critère d'information d'Akaike (AIC)

Le critère AIC indique de choisir le modèle ARMA(p,q) qui minimise la fonction

$$AIC = -2 \ln L + 2k ,$$

où $\ln L$ est le logarithme naturel de la fonction du maximum de vraisemblance, fonction calculée sous l'hypothèse d'une distribution conjointe pour Y_1, Y_2, \dots, Y_n , et k est le nombre de paramètres du modèle (excluant la variance du bruit blanc, σ_e^2). Ainsi, dans un modèle ARMA(p,q) sans constante, il y a $k = p + q$ paramètres, et dans un modèle avec constante, il y a $k = p + q + 1$ paramètres.

Remarques : (1) La fonction de vraisemblance L est construite d'une manière passablement complexe à partir de l'hypothèse que Y_t génère un processus ARMA(p,q) Gaussien et en utilisant des estimés particuliers des paramètres ϕ_1, \dots, ϕ_p et $\theta_1, \dots, \theta_q$ (voir le livre de Brockwell & Davis, sections 5.1 et 5.2).

La théorie justifiant ce critère couvrant beaucoup trop d'aspects techniques, nous ne nous attarderons pas aux fondements de cette fonction AIC et nous laisserons le R faire les calculs pour nous! Ce critère est incorporé à la fonction R « ar(x, ...) ».

(2) Le terme $2k$ sert de « pénalité », i.e. que nous ne voulons pas de modèles ARMA avec trop de paramètres.

6.6.2 Le critère d'information Bayesienne (BIC)

Le critère BIC indique de choisir le modèle ARMA(p,q) qui minimise la fonction

$$BIC = -2 \ln L + k \ln n ,$$

où L et k sont définis (et L est construit) comme précédemment.

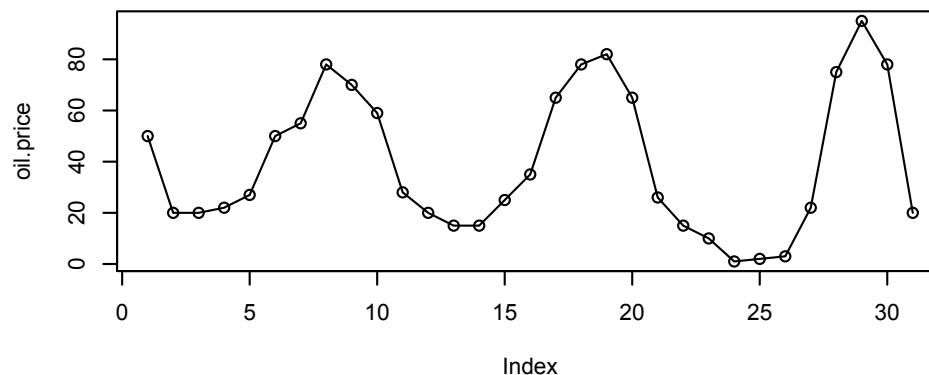
Remarques : Si nous comparons le critère AIC avec le critère BIC, le deuxième critère pénalise plus fortement le nombre de paramètres du modèle ARMA puisque $\ln n$ dépasse le nombre 2 dans la très grande majorité des cas.

6.7 Spécifications de séries chrono. de notre base de données

6.7.1 Exemple 1

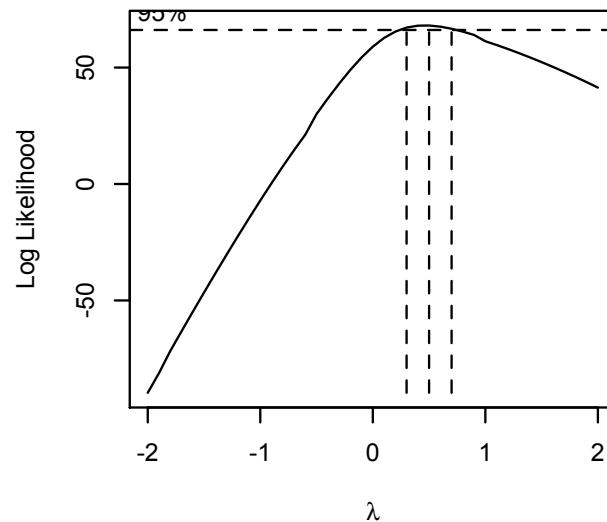
Considérer le fichier (hare), recensant la population annuelle de lièvres au Canada de 1905 à 1935 ($\Rightarrow n = 31$)

```
> data(hare)
> win.graph(width=4.875,height=2.5,pointsize=8)
> plot(as.vector(hare), type='o', ylab='oil.price')
```



Nous allons d'abord vérifier si une transformation des données serait appropriée.

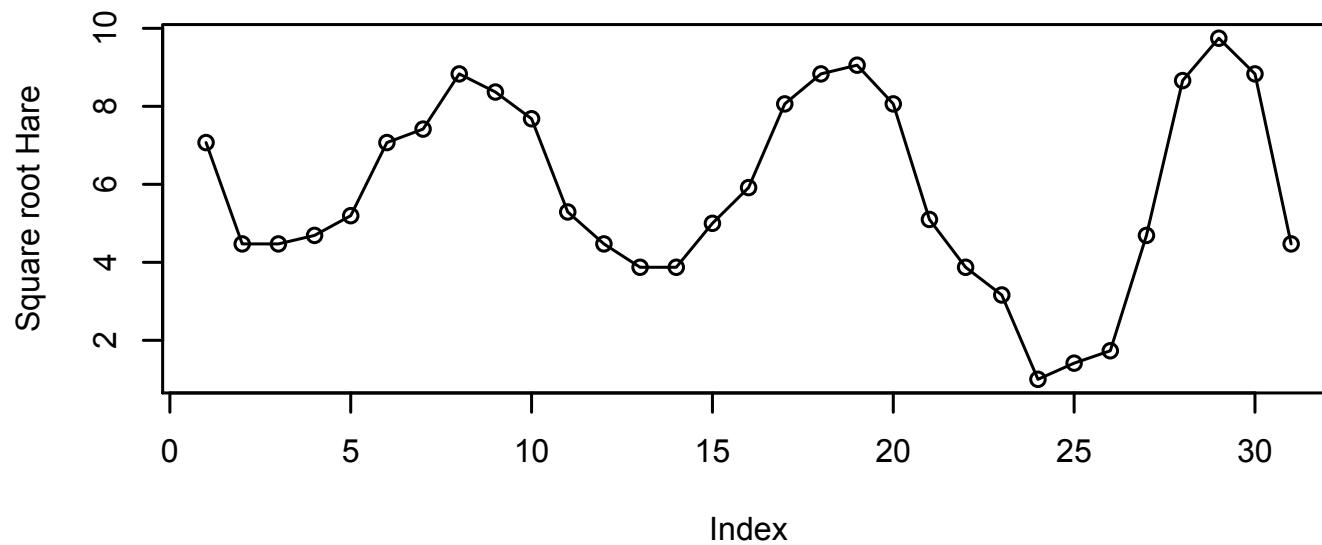
```
> win.graph(width=3, height=3, pointsize=8)
> BoxCox.ar(hare)
```



La fonction de log vraisemblance donne $\lambda = 0.4$, à un niveau de confiance de 95%. Nous prendrons $\lambda = 0.5$, puisqu'il se trouve dans l'intervalle de confiance de λ .

Nous dressons le graphique de la racine carré de nos données.

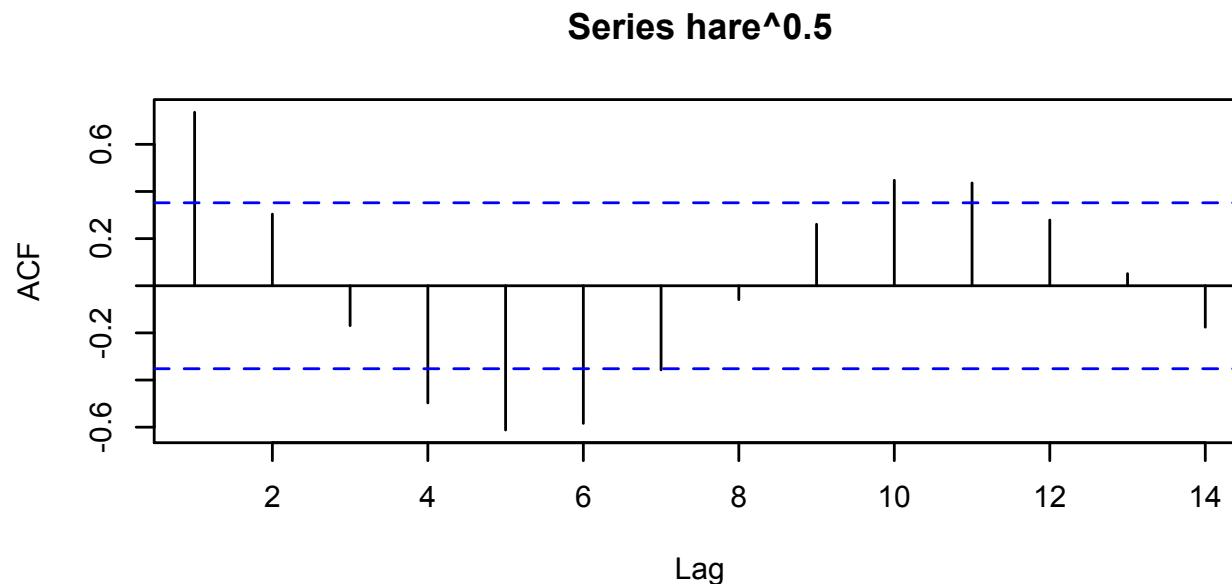
```
> win.graph(width=4.875, height=2.5, pointsize=8)
> plot(as.vector(hare^0.5), type='o', ylab='Square root Hare')
```



Pas encore d'informations précises à cette étape.

Nous examinons la fonction d'autocorrélation de « `hare^0.5` ».

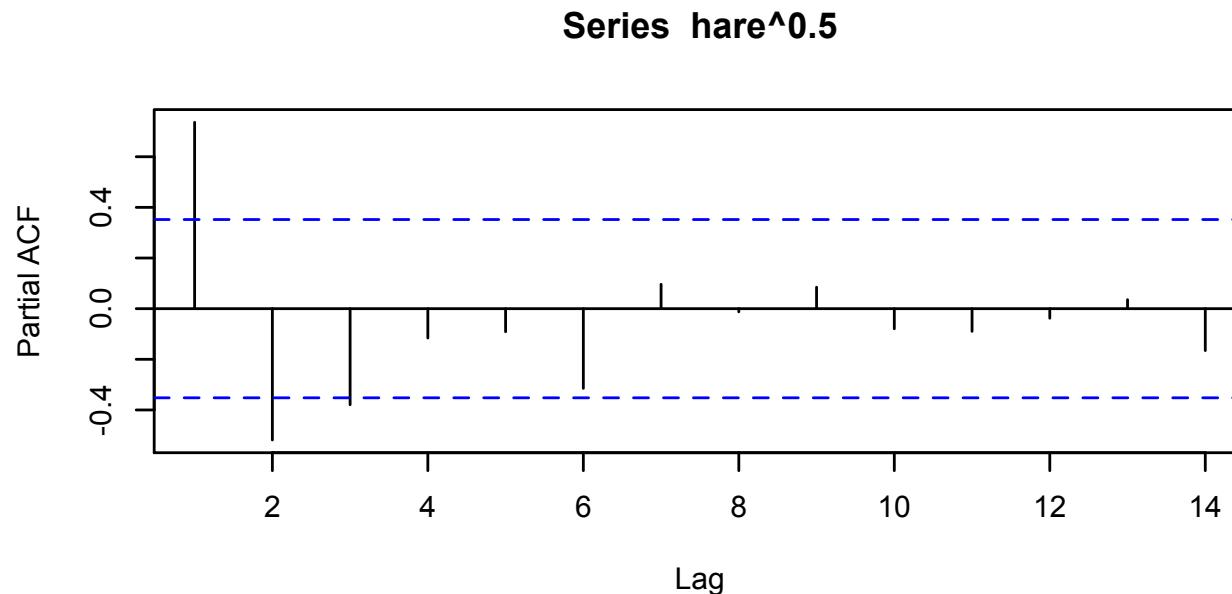
```
> acf(hare^0.5)
```



La fonction ACF oscille passablement. Il y a sans doute une partie autorégressive présente.

Nous dressons un graphique PACF pour le vérifier.

```
> pacf(hare^0.5)
```



Il y a au moins un processus autorégressif d'ordre 2, probablement d'ordre 3 ou même d'ordre 6. Il n'y a probablement pas besoin d'ajouter une partie MA(q), si on se réfère au graphique précédent.

Par curiosité, que nous dit le test EACF?

```
> eacf(hare^0.5,ar.max=4,ma.max=4)
```

AR/MA

	0	1	2	3	4
0	x	o	o	x	x
1	x	o	o	x	x
2	o	o	o	o	o
3	o	o	o	o	o
4	o	o	o	o	o

Le processus autorégressif d'ordre 2 semble bien être un des modèles à considérer.

Puisqu'il ne semble pas nécessaire de différencier notre série chronologique $\text{hare}^{0.5}$, nous déterminons maintenant l'ordre du processus AR qui s'ajustera le mieux à $\text{hare}^{0.5}$.

```
> library(tseries)
> ar(hare^0.5)
```

Call:

```
ar(x = hare^0.5)
```

Coefficients:

1	2	3
0.9208	-0.0945	-0.3795

Order selected 3 sigma² estimated as 1.873

* R recommande d'utiliser k = 3.

Le test ADF est appliqué avec k = 3.

```
> adf.test(hare^0.5, k=3)
```

Augmented Dickey & Fuller test

data : hare^0.5

Dickey-Fuller = -4.479 , Lag order = 3 , p-value = 0.01

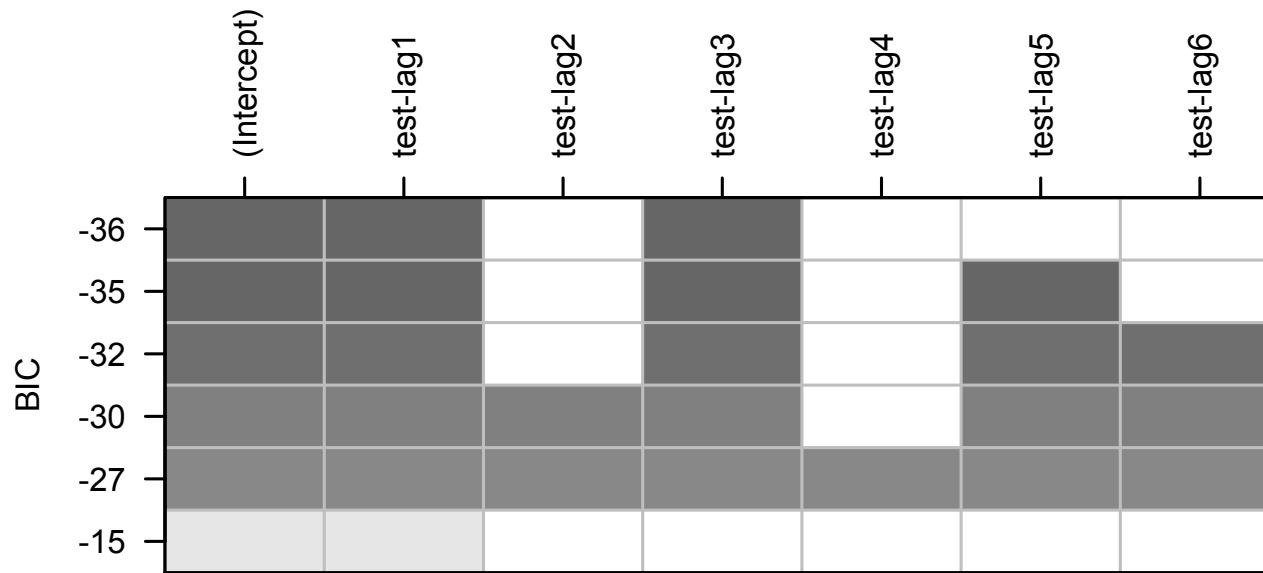
alternative hypothesis : stationary

Ce test ADF confirme fortement que la série « hare^0.5 » est stationnaire. Donc un processus AR(3) semblerait tout indiqué pour représenter notre série transformée, i.e. si Y_t représente les données du fichier « hare », alors nous pouvons considérer le modèle suivant

$$Y_t^{1/2} = c + \phi_1 Y_{t-1}^{1/2} + \phi_2 Y_{t-2}^{1/2} + \phi_3 Y_{t-3}^{1/2} + e_t .$$

Est-ce que le critère BIC pourrait nous confirmer cela et (ou) nous donner d'autres possibilités?

```
> res=armasubsets(y=hare^0.5, nar=6, nma=0, y.name='test',  
ar.method='ols')  
> plot(res)
```



Le meilleur modèle AR à considérer (première ligne!) est un modèle avec constante et avec éléments d'écart 1 et 3, i.e le modèle suivant

$$Y_t^{1/2} = c + \phi_1 Y_{t-1}^{1/2} + \phi_3 Y_{t-3}^{1/2} + e_t .$$

Le deuxième meilleur modèle AR à considérer (deuxième ligne!) est un modèle avec constante et avec éléments d'écart 1, 3 et 5, i.e. le modèle suivant

$$Y_t^{1/2} = c + \phi_1 Y_{t-1}^{1/2} + \phi_3 Y_{t-3}^{1/2} + \phi_5 Y_{t-5}^{1/2} + e_t .$$

Le troisième meilleur modèle AR à considérer (troisième ligne!) est un modèle avec constante et avec éléments d'écart 1, 3, 5 et 6 i.e. le modèle suivant

$$Y_t^{1/2} = c + \phi_1 Y_{t-1}^{1/2} + \phi_3 Y_{t-3}^{1/2} + \phi_5 Y_{t-5}^{1/2} + \phi_6 Y_{t-6}^{1/2} + e_t .$$

Et ainsi de suite ...

* Parmi tous ces modèles, les deux premiers semblent être les meilleurs candidats potentiels, soit

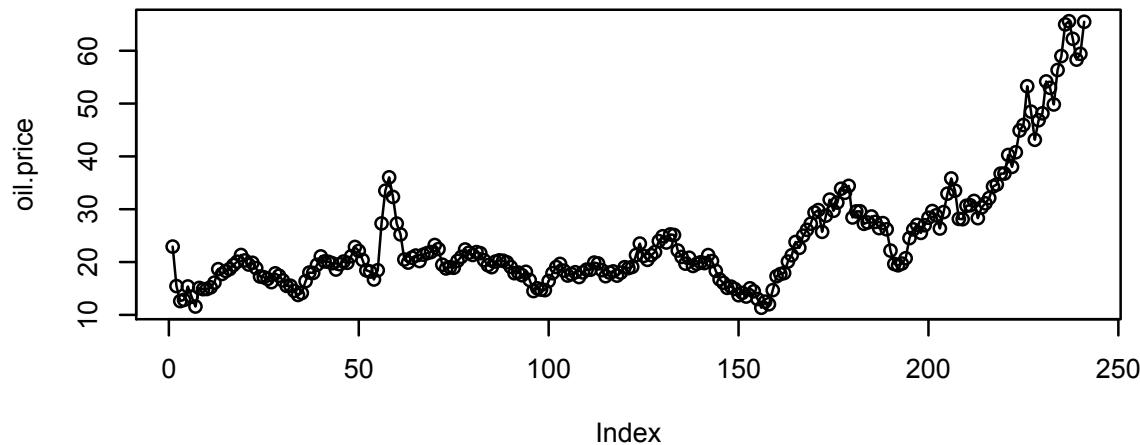
- $Y_t^{1/2} = c + \phi_1 Y_{t-1}^{1/2} + \phi_2 Y_{t-2}^{1/2} + \phi_3 Y_{t-3}^{1/2} + e_t$ (suivant le graphique PACF)
- $Y_t^{1/2} = c + \phi_1 Y_{t-1}^{1/2} + \phi_3 Y_{t-3}^{1/2} + e_t$ (suivant le critère BIC)

Si ϕ_2 est très petit, le deuxième modèle serait probablement celui qui serait retenu pour l'estimation, la validation et la prédiction.

6.7.2 Exemple 2

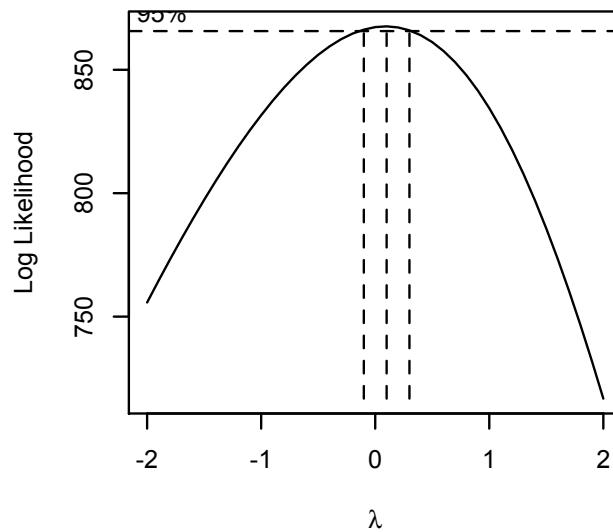
Considérer le fichier (oil.price), donnant le prix mensuel du baril de pétrole brut de janvier 1986 à janvier 2006.

```
> data(oil.price)
> win.graph(width=4.875, height=2.5, pointsize=8)
> plot(as.vector(oil.price), type='o', ylab='oil.price')
```



Nous allons d'abord vérifier si une transformation des données serait appropriée.

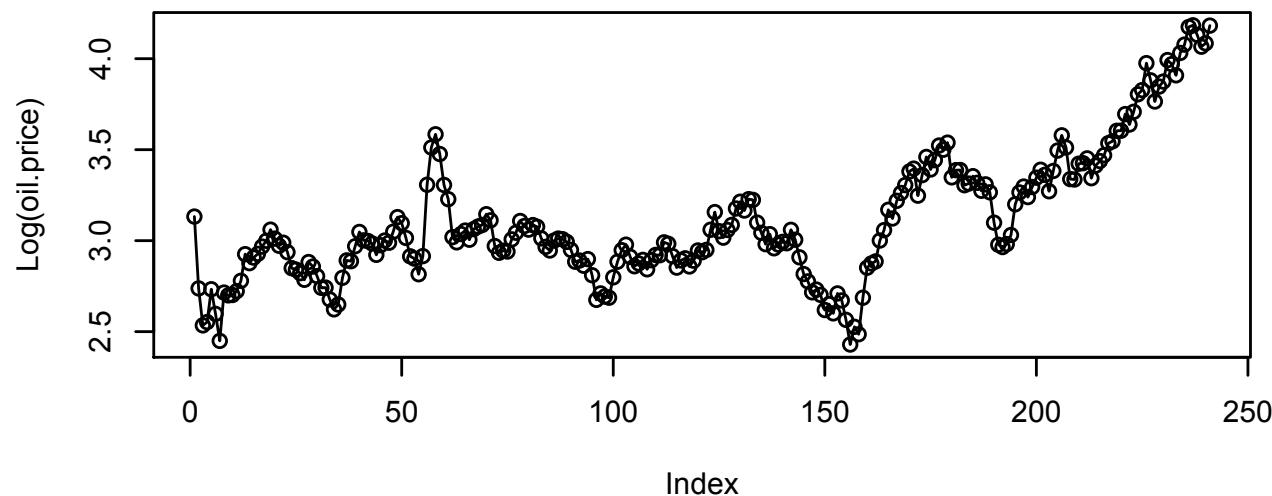
```
> win.graph(width=3, height=3, pointsize=8)
> BoxCox.ar(oil.price)
```



La fonction de log vraisemblance donne $\lambda = 0.1$, à un niveau de confiance de 95%. Nous prendrons $\lambda = 0$, puisqu'il se trouve dans l'intervalle de confiance de λ .

Nous dressons le graphique du logarithme de nos données.

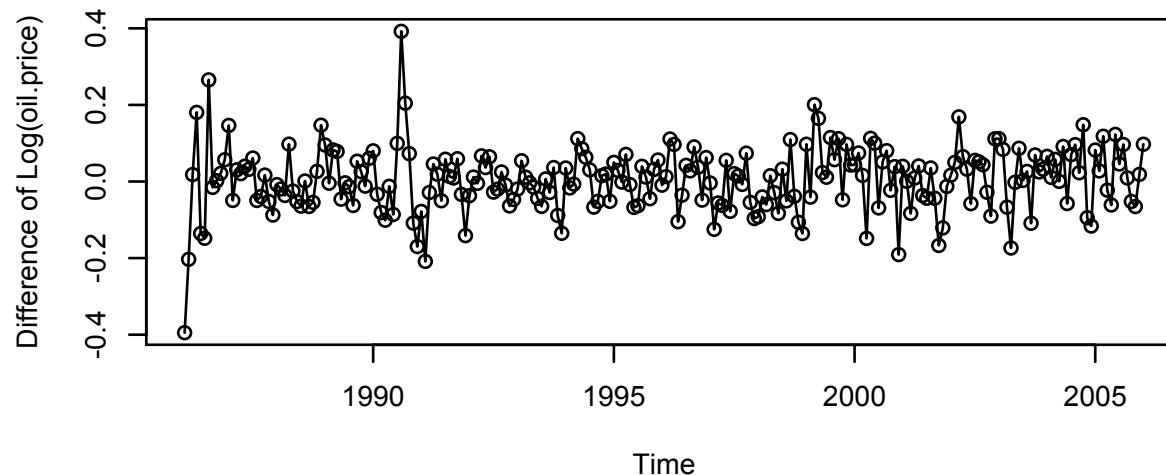
```
> win.graph(width=4.875, height=2.5, pointsize=8)
> plot(as.vector(log(oil.price)), type='o', ylab='Log(oil.price)')
```



Pas encore d'informations précises à cette étape.

Un graphique de la différence du « log(oil.price) » serait probablement approprié.

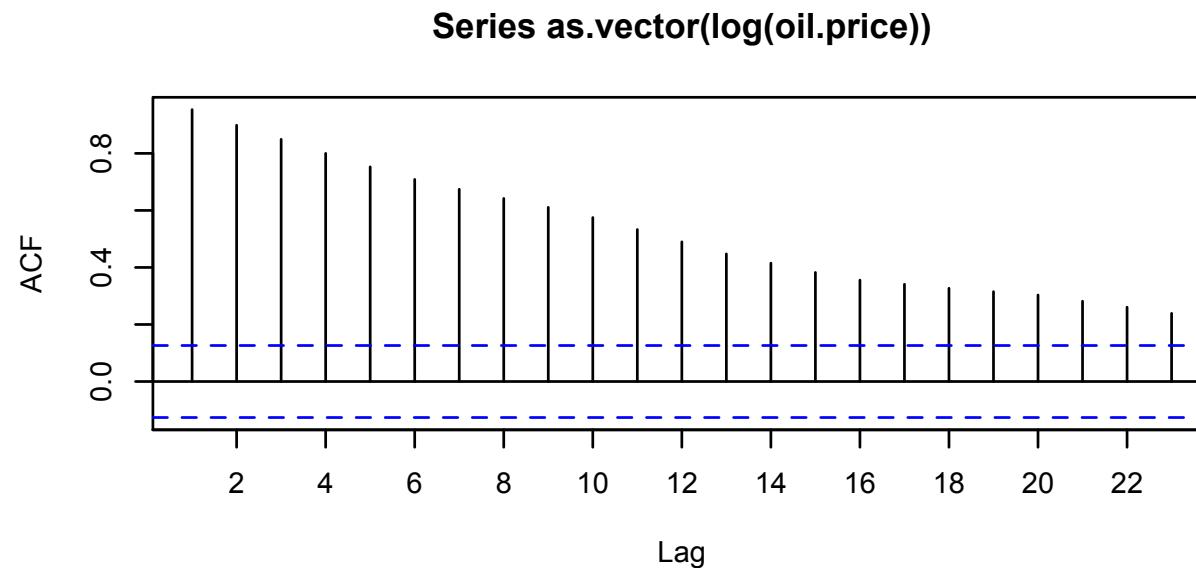
```
> plot(as.vector(diff(log(oil.price))), ylab='Difference of Log(oil.price)', type='o')
```



Nous sommes sur la bonne voie.

Y a-t-il stationnarité? Nous examinons la fonction d'autocorrélation de « log(oil.price) ».

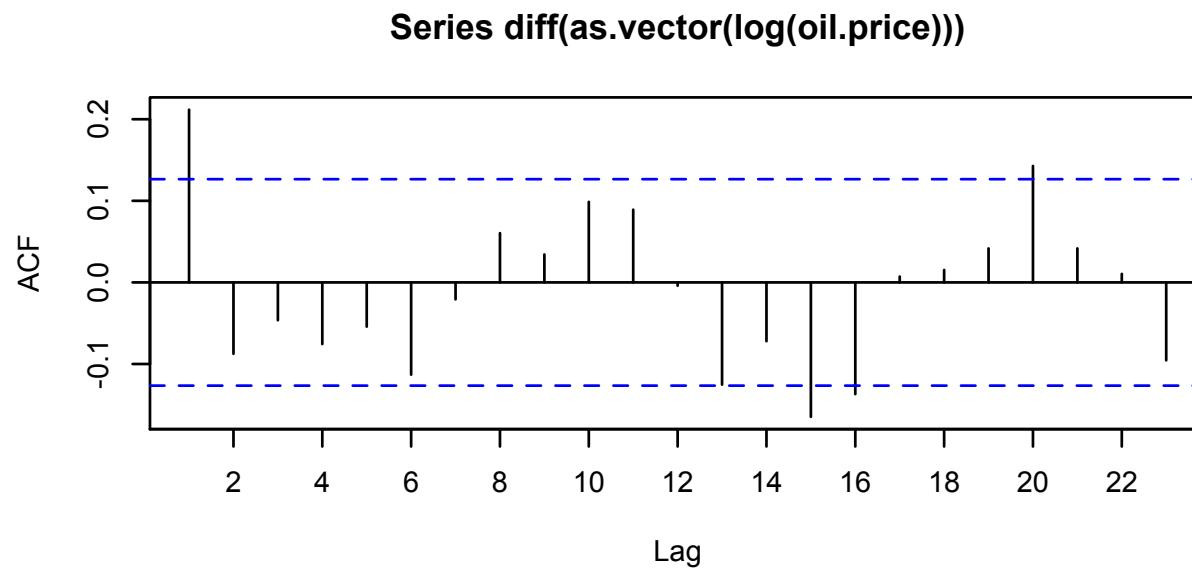
```
> acf(as.vector(log(oil.price)), xaxp=c(0,24,12))
```



La fonction ACF décroît très lentement, signe de non stationnarité.

Nous examinons la fonction d'autocorrélation de la différence de « log(oil.price) ».

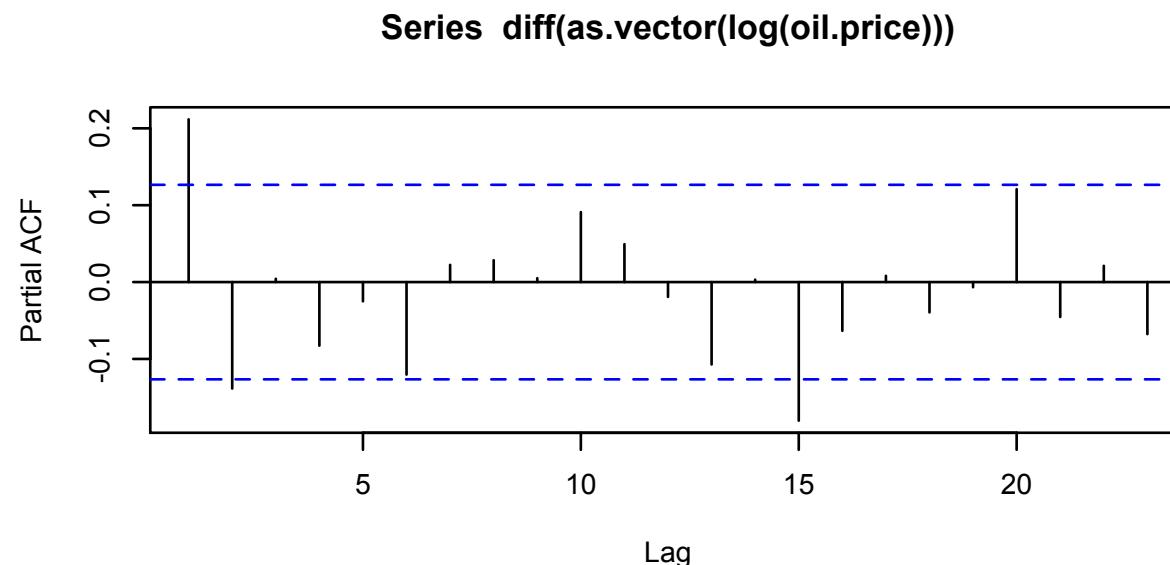
```
> acf(diff(as.vector(log(oil.price))), xaxp=c(0,24,12))
```



Le graphique ACF suggère une MA(1) pour la différence du « log(oil.price) », malgré une certaine oscillation de la fonction qui pourrait introduire une partie autorégressive.

Nous dressons un graphique PACF de la différence de « log(oil.price) ».

```
> pacf(diff(as.vector(log(oil.price))))
```



Le graphique PACF suggère un processus AR(1), et peut-être AR(2) ou même AR(6), pour la différence du « log(oil.price) », malgré l'écart 15 hors de l'intervalle de confiance.

Nous examinons la fonction EACF de la différence de « log(oil.price) ».

```
> eacf(diff(log(oil.price)))
```

AR/MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	o	o	o	o	o	o	o	o	o	o	o	o	o
1	x	x	o	o	o	o	o	o	o	x	o	o	o	o
2	o	x	o	o	o	o	o	o	o	o	o	o	o	o
3	o	x	o	o	o	o	o	o	o	o	o	o	o	o
4	o	x	x	o	o	o	o	o	o	o	o	o	o	o
5	o	x	o	x	o	o	o	o	o	o	o	o	o	o
6	o	x	o	x	o	o	o	o	o	o	o	o	o	o
7	x	x	o	x	o	o	o	o	o	o	o	o	o	o

Le graphique EACF suggère fortement une MA(1) pour la différence du « log(oil.price) ».

Nous déterminons maintenant l'ordre du processus AR qui s'ajustera le mieux à $\text{diff}(\log(\text{oil.price}))$.

```
> library(tseries)
> ar(diff(log(oil.price)))
```

Call:

```
ar(x = diff(log(oil.price)))
```

Coefficients:

1	2
0.2410	-0.1385

Order selected 2 sigma² estimated as 0.006767

* Le meilleur k est k = 2.

Le test ADF est maintenant appliqué, avec $k = 2$, sur les données « log(oil.price) »

```
> adf.test(log(oil.price), k=2)
```

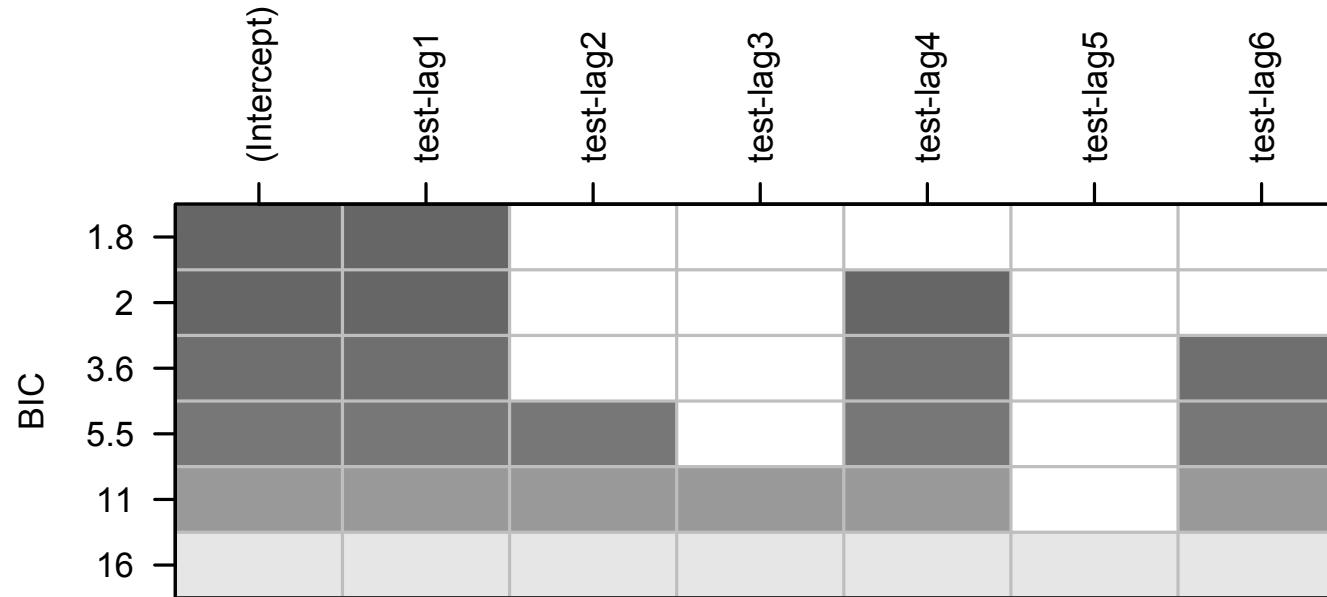
Augmented Dickey & Fuller test

```
data: log(oil.price)
Dickey-Fuller = -1.9401, Lag order = 2, p-value = 0.6011
alternative hypothesis: stationary
```

Ce test ADF confirme fortement que la série « log(oil.price) » est non stationnaire.

Voyons ce que le critère BIC nous donnera ...

```
> res=armasubsets(y=diff(log(oil.price)), nar=6, nma=1,  
+ y.name='test', ar.method='ols')  
> plot(res)
```



Le meilleur modèle AR à considérer (première ligne!) pour les données de la différence du log(oil.price) est un modèle avec constante et élément d'écart 1, i.e le modèle suivant

$$\nabla \log Y_t = c + \phi \nabla \log Y_{t-1} + e_t .$$

Le deuxième meilleur modèle à considérer (deuxième ligne!) pour les données de la différence du log(oil.price) est un modèle sans constante ni tendance, avec éléments d'écart 1 et 4, i.e le modèle suivant

$$\nabla \log Y_t = c + \phi_1 \nabla \log Y_{t-1} + \phi_4 \nabla \log Y_{t-4} + e_t .$$

Nous avons donc comme modèles potentiels pour la différence du « log(oil.price) »

- $\nabla \log Y_t = e_t - \theta e_{t-1}$ (suivant le graphique ACF de la diff. ... et le tableau EACF)
- $\nabla \log Y_t = \phi_1 \nabla \log Y_{t-1} + \phi_2 \nabla \log Y_{t-2} + e_t$ (suivant le graphique PACF de la diff. ... et le critère adf)
- $\nabla \log Y_t = c + \phi \nabla \log Y_{t-1} + e_t$ (suivant le critère BIC)
- $\nabla \log Y_t = c + \phi_1 \nabla \log Y_{t-1} + \phi_4 \nabla \log Y_{t-4} + e_t$ (suivant le critère BIC)

* Nous obtenons plusieurs modèles qu'il sera nécessaire d'estimer, de valider, pour ne retenir qu'un seul modèle (si possible) qui nous servira pour la prédiction.