时间序列分析（TIME SERIES ANALYSIS）

主讲：吴尚
复旦大学管理学院统计与数据科学系

时间序列回归模型

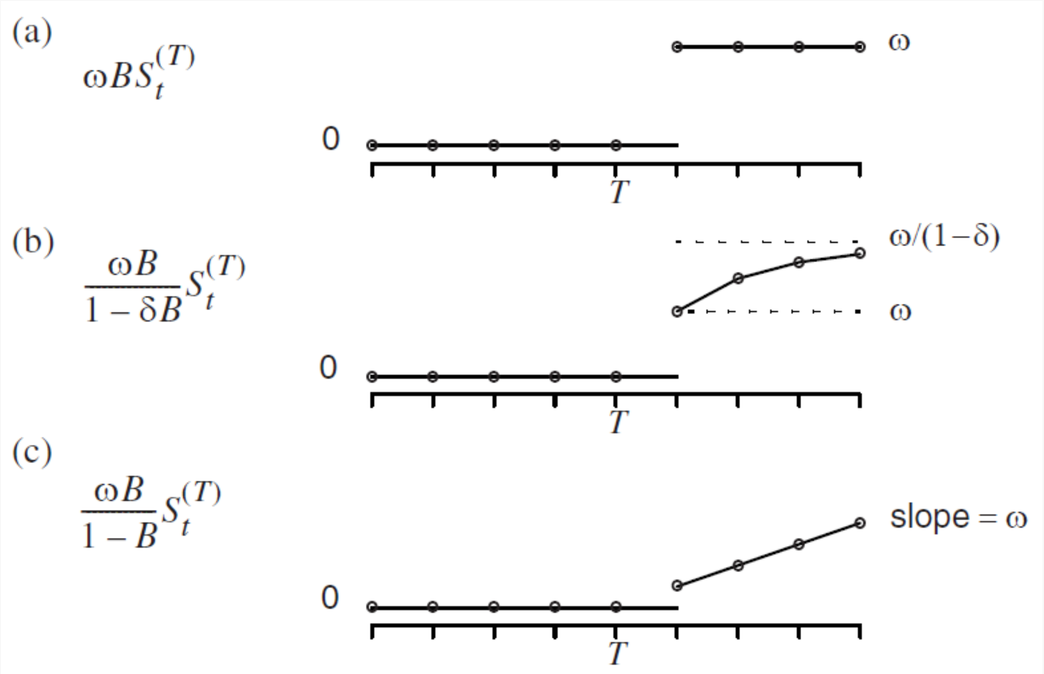
- 干预分析
- 异常值
- 伪相关
- 预白化与随机回归

干预分析

- 假设干预是通过改变时间序列的均值函数或趋势而对过程施加影响，例如收到自然灾害、突发事件的影响序列均值发生变化。
- $Y_t = m_t + N_t$
- 定义阶梯函数$S_t^{(T)} = I(t \geq T)$, 即当$t \geq T$时$S_t^{(T)} = 1$, 否则$S_t^{(T)} = 0$.
- 定义脉冲函数$P_t^{(T)} = I(t = T)$,即当$t = T$时$P_t^{(T)} = 1$, 否则$P_t^{(T)} = 0$. 我们有$P_t^{(T)} = S_t^{(T)} - S_{t-1}^{(T)}$.

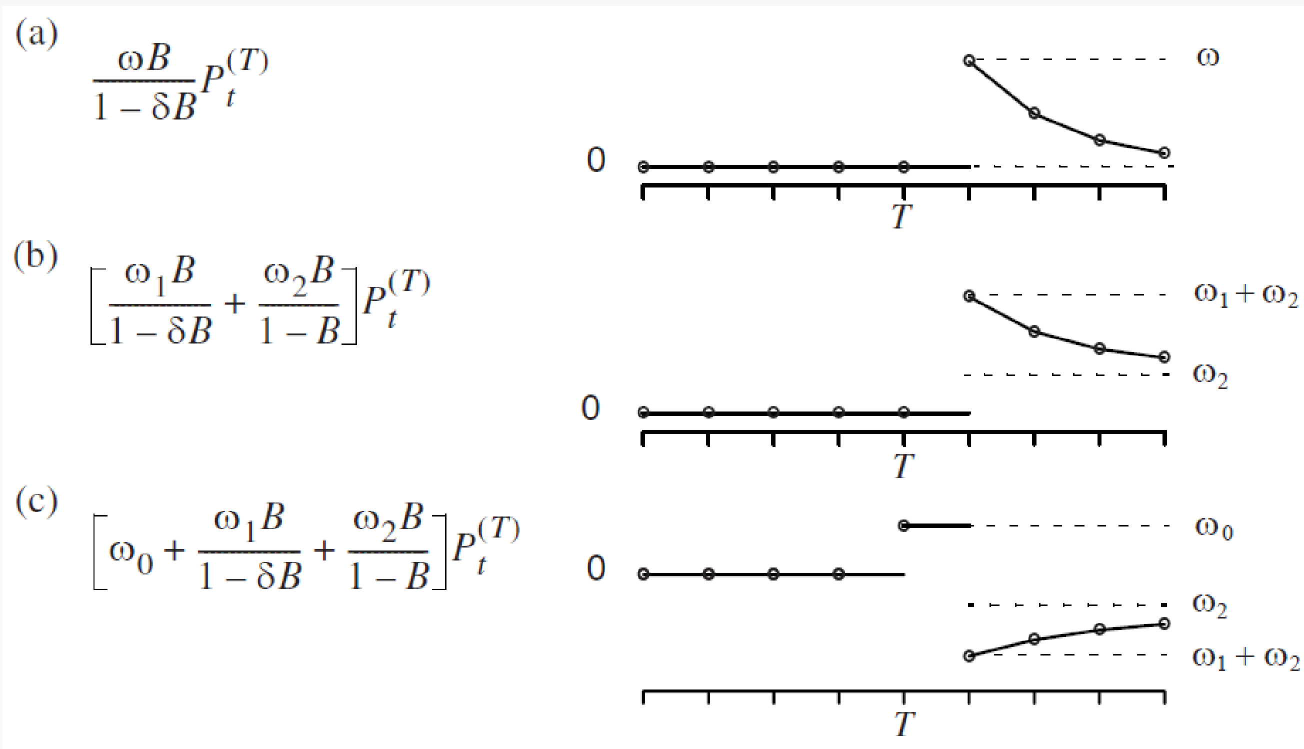
干预类型一

- $m_t = \omega S_t^{(T)}, m_t = \omega S_{t-d}^{(T)} = \omega B^d S_t^{(T)}$.
- $m_t = \delta m_{t-1} + \omega S_{t-1}^{(T)}$, 可写作$(1-\delta B)m_t = \omega B S_t^{(T)}$.
- 初始$m_0 = 0$,则$m_{T+1} = \omega$, 当$0 < \delta < 1$时，m_t单调递增，收敛到$\frac{\omega}{1-\delta}$，当$\delta = 1$时，$m_t = \omega(t-T)$.



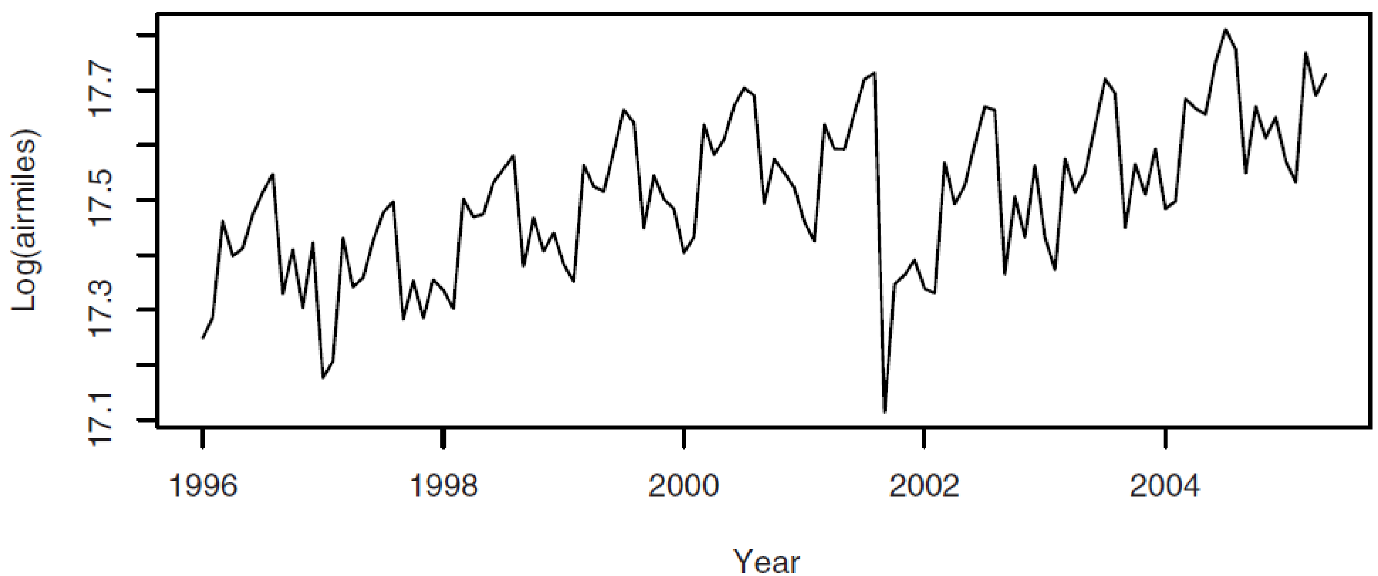
干预类型二

- $m_t = \omega P_t^{(T)}, m_t = \delta m_{t-1} + \omega P_{t-1}^{(T)}$，可写作 $(1 - \delta B)m_t = \omega B P_t^{(T)}$.
- $m_t = \frac{\omega_1 B}{1-\delta B} P_t^{(T)} + \frac{\omega_2 B}{1-B} P_t^{(T)}$，一般地$m_t = \frac{\omega(B)}{\delta(B)} P_t^{(T)}$.



例：美国航空的每月客运里程（对数）

Exhibit 11.1 Monthly U.S. Airline Miles: January 1996 through May 2005

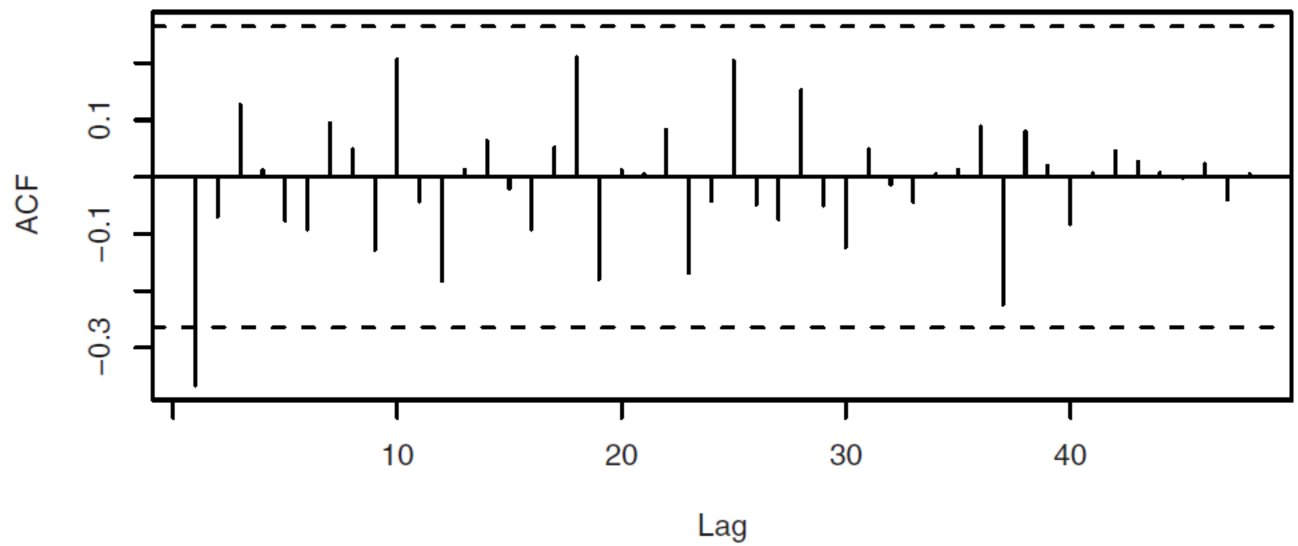



```
> win.graph(width=4.875,height=2.5,pointsize=8)
> data(airmiles)
> plot(log(airmiles),ylab='Log(airmiles)',xlab='Year')
```

例

- 对于预干预数据（未被影响的，T之前的序列），根据一次差分和季节差分后的ACF，为该序列拟合ARIMA(0,1,1)(0,1,0)12模型。

Exhibit 11.5 Sample ACF for $(1-B)(1-B^{12})$ Log(Air Passenger Miles) Over the Preintervention Period




```
> acf(as.vector(diff(diff(window(log(airmiles),end=c(2001,8)),
   12))),lag.max=48)
```

例

- 对拟合模型的诊断表明，模型需要一个季节MA(1)系数，因此为该序列拟合ARIMA(0,1,1)(0,1,1)12模型。
- 考虑干预效应$m_t = \omega_0 P_t^{(T)} + \frac{\omega_1}{1-\omega_2 B} P_t^{(T)}$
- 同时检测到3个可加异常值（即$\omega_A P_t^{(T)}$，后述）。

Exhibit 11.6 Estimation of Intervention Model for Logarithms of Air Miles (Standard errors are shown below the estimates)

θ	Θ	Dec96	Jan97	Dec02	ω_0	ω_1	ω_2
0.383	0.650	0.099	–0.069	0.081	–0.095	–0.27	0.814
(0.093)	(0.119)	(0.023)	(0.022)	(0.020)	(0.046)	(0.044)	(0.098)

σ^2 estimated as 0.000672: log-likelihood = 219.99, AIC= –423.98

```
> air.m1=arimax(log(airmiles),order=c(0,1,1),
   seasonal=list(order=c(0,1,1),period=12),
   xtransf=data.frame(I911=1*(seq(airmiles)==69),
   I911=1*(seq(airmiles)==69)),transfer=list(c(0,0),c(1,0)),
   xreg=data.frame(Dec96=1*(seq(airmiles)==12),
   Jan97=1*(seq(airmiles)==13),Dec02=1*(seq(airmiles)==84)),
   method='ML')
> air.m1
```

例

Exhibit 11.7 Logs of Air Passenger Miles and Fitted Values

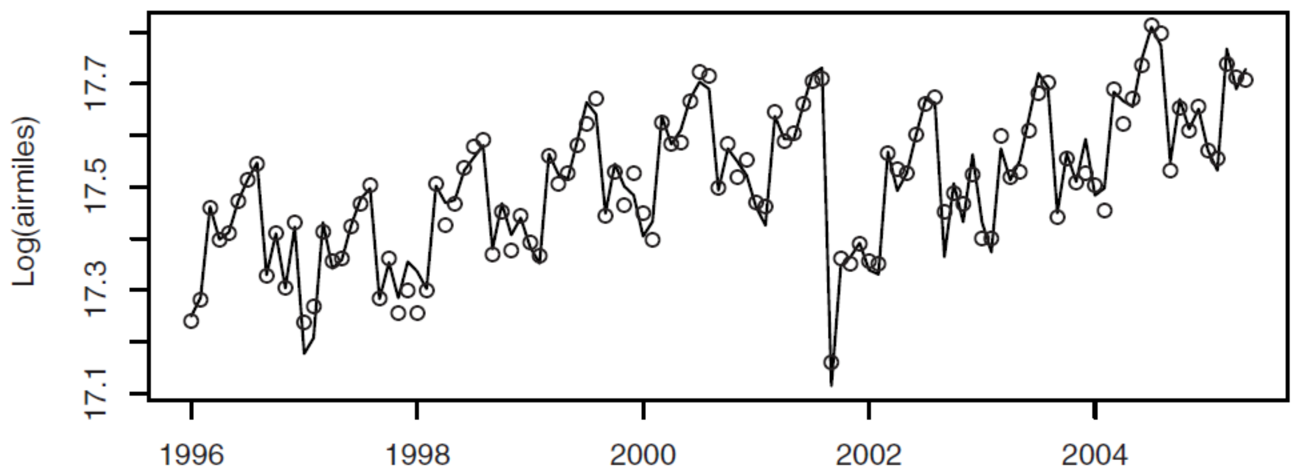
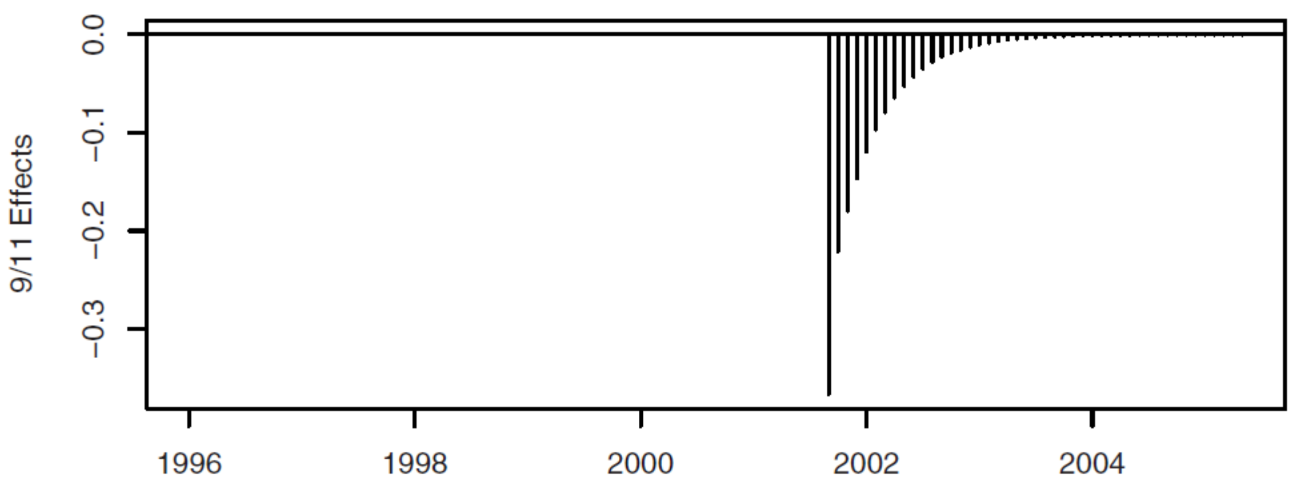


Exhibit 11.8 The Estimated 9/11 Effects for the Air Passenger Series



异常值

- 可加异常值（AO）：$Y_t' = Y_t + \omega_A P_t^{(T)}$
- 新息异常值（IO）： $e_t' = e_t + \omega_I P_t^{(T)}$
  - *此时：* $Y_t' = Y_t + \psi_{t-T}\omega_I$.
- 识别异常值，哪个更容易?
- 定义残差$a_t = Y_t' - \pi_1 Y_{t-1}' - \pi_2 Y_{t-2}' - \cdots$

识别IO

- 若序列在T时刻有IO，则残差$a_t = e_t + \omega_I P_t^{(T)}$，零假设（$\omega_I = 0$）下$a_t$应满足均值为0，方差为$\sigma^2 = \sigma_e^2$的正态分布，定义时刻$T$上IO的检验统计量为

$$\lambda_{1,T} = \frac{a_T}{\sigma}$$

- 需要找到所有时刻当中真正特别极端的异常值，所以临界值用Bonferroni准则，取标准正态分布的上百分位数$\frac{0.025}{n} \times 100\%$.
- 由于异常值会导致噪声标准差估计偏大，所以可用残差绝对均值乘以$\sqrt{\pi/2}$得到σ更为稳健的估计。

识别AO

- $Y_t' = Y_t + \omega_A P_t^{(T)}$，$a_t = Y_t' - \pi_1 Y_{t-1}' - \pi_2 Y_{t-2}' - \cdots$， $e_t = Y_t - \pi_1 Y_{t-1} - \pi_2 Y_{t-2} - \cdots$
- 结合可得，$a_t - e_t = -\pi_{t-T}(Y_T' - Y_T) = -\pi_{t-T}\omega_A$，这里$\pi_0 = -1$, 当$j < 0$时$\pi_j = 0$.
- 使用最小二乘法，最小化$\sum_{t=1}^n e_t^2 = \sum_{t=1}^n (a_t + \pi_{t-T}\omega_A)^2$，可得$\widetilde{\omega}_{T,A} = -\rho^2 \sum_{t=1}^n \pi_{t-T} a_t$，其中$\rho^2 = (1 + \pi_1^2 + \cdots + \pi_{n-T}^2)^{-1}$.
- 零假设下$Var(\widetilde{\omega}_{T,A}) = \rho^4 \sum_{t=1}^n \pi_{t-T}^2 \sigma^2 = \rho^2\sigma^2$，定义
$$\lambda_{2,T} = \frac{\widetilde{\omega}_{T,A}}{\rho\sigma}$$
- 同样用Bonferroni准则和稳健估计。
- $|\lambda_{1,T}| > |\lambda_{2,T}|$时，识别为IO，否则为AO.

例

Exhibit 11.10 ARIMA(0,1,1)×(0,1,1)$_{12}$ Model with IO at $t = 57$ for CO_2 Series

Coefficient	θ	Θ	IO-57
Estimate	0.5925	0.8274	2.6770
Standard Error	0.0775	0.1016	0.7246

$\hat{\sigma}_e^2 = 0.4869$: log-likelihood $= -133.08$, AIC $= 272.16$

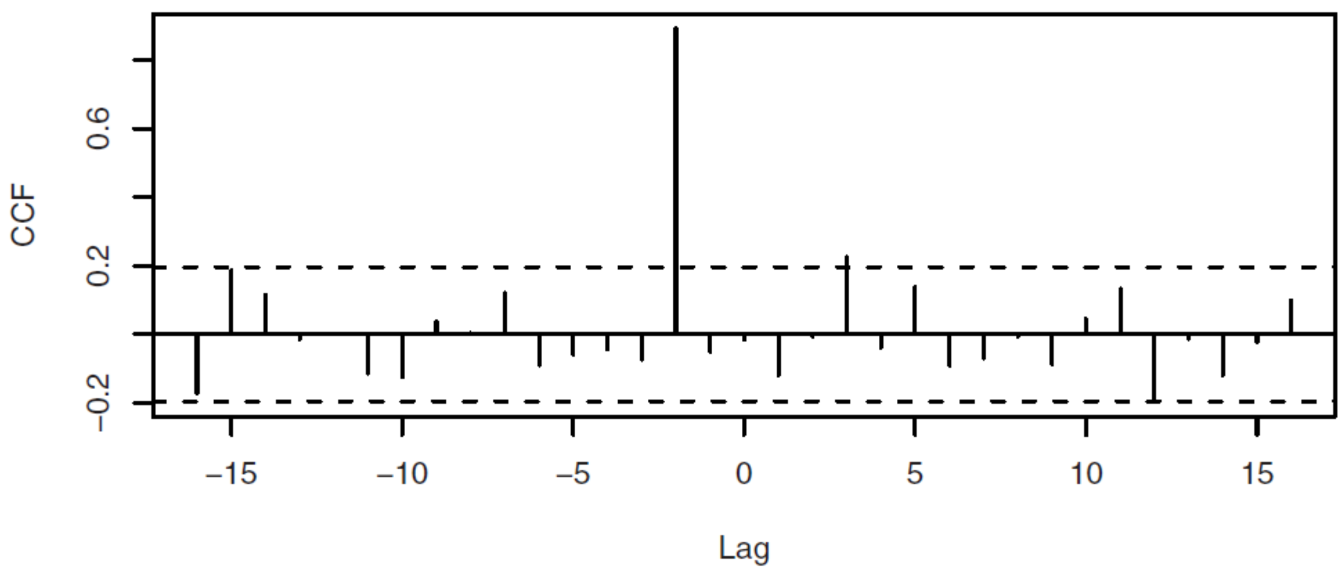
```
> m1.co2=arima(co2,order=c(0,1,1),seasonal=list(order=c(0,1,1),
   period=12)); m1.co2
> detectAO(m1.co2); detectIO(m1.co2)
> m4.co2=arimax(co2,order=c(0,1,1),seasonal=list(order=c(0,1,1),
   period=12),io=c(57)); m4.co2
```

伪相关

- 假设两个平稳序列的协方差只和时差有关，则定义互相关函数为$\rho_k(X,Y)=Corr(X_t,Y_{t-k})$，互相关函数不必是偶函数。
- 考虑如下回归模型，假设X_t,e_t是独立的白噪声
$$Y_t=\beta_0+\beta_1 X_{t-d}+e_t$$
- 则$\rho_{-d}(X,Y)=Corr(X_{t-d},Y_t)=\beta_1\sigma_X/\sigma_Y$.
- 样本互相关函数定义为
$$r_k(X,Y)=\frac{\sum_{t=k+1}^{n}(X_t-\bar{Y})(Y_{t-k}-\bar{Y})}{\sqrt{\sum_{t=1}^{n}(X_t-\bar{X})^2}\sqrt{\sum_{t=1}^{n}(Y_t-\bar{Y})^2}}$$
- 当$\beta_1=0$时，$r_k(X,Y)$渐近服从均值为0，方差为1/n的正态分布。

模拟数据

Exhibit 11.11 Sample Cross-Correlation from Equation (11.3.1) with *d* = 2




```
> win.graph(width=4.875,height=2.5,pointsize=8)
> set.seed(12345); X=rnorm(105); Y=zlag(X,2)+.5*rnorm(105)
> X=ts(X[-(1:5)],start=1,freq=1); Y=ts(Y[-(1:5)],start=1,freq=1)
> ccf(X,Y,ylab='CCF')
```

伪相关

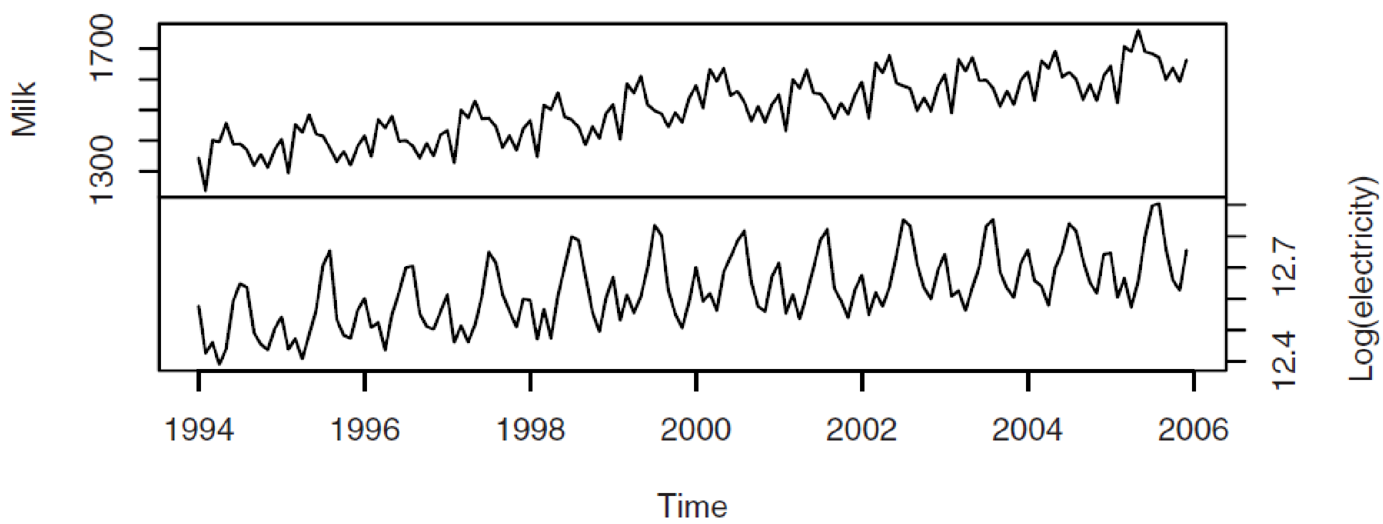
- 问题在于，两个独立的平稳序列（或非平稳序列），其互相关函数一定不显著吗？（基于方差1/n的正态分布）
- 答案是否定的，X和Y是两个独立的平稳序列，则$\sqrt{n}r_k(X,Y)$的方差渐近等于

$$1+2\sum_{k=1}^{\infty}\rho_k(X)\rho_k(Y)$$

- 例如，X和Y是独立的AR(1)模型时，上式等于$\frac{1+\phi_X\phi_Y}{1-\phi_X\phi_Y}$，可以任意大！

伪相关

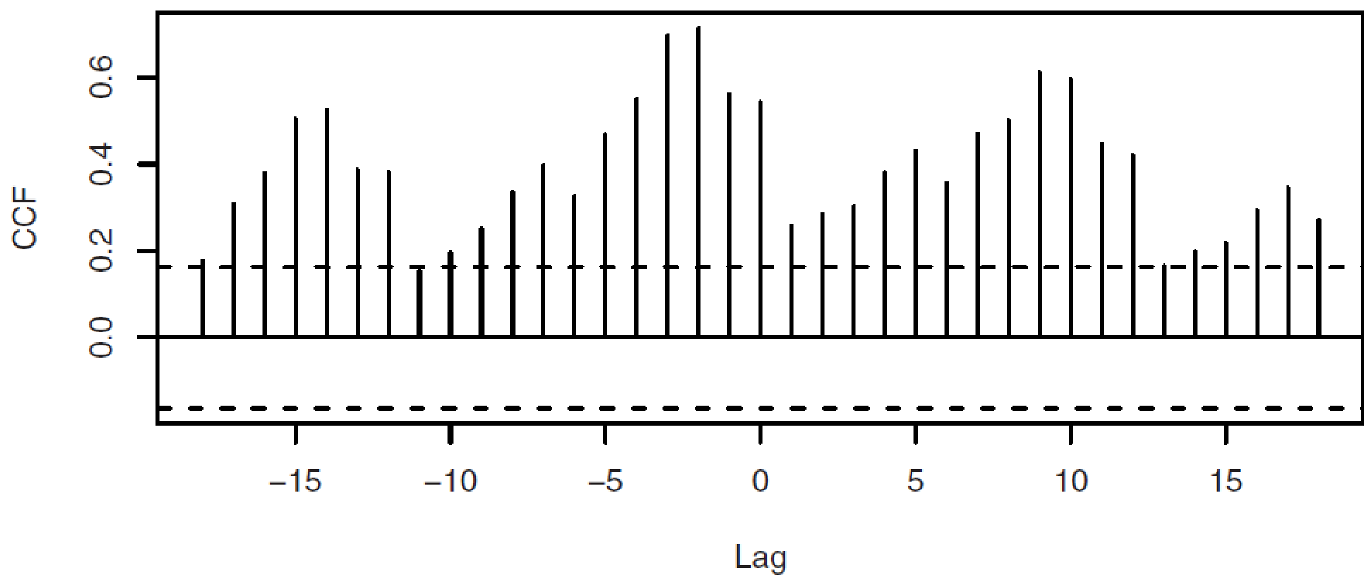
Exhibit 11.14 Monthly Milk Production and Logarithms of Monthly Electricity Production in the U.S.




```
> data(milk); data(electricity)
> milk.electricity=ts.intersect(milk,log(electricity))
> plot(milk.electricity,yax.flip=T)
```

伪相关

Exhibit 11.15 Sample Cross-Correlation Between Monthly Milk Production and Logarithm of Monthly Electricity Production in the U.S.



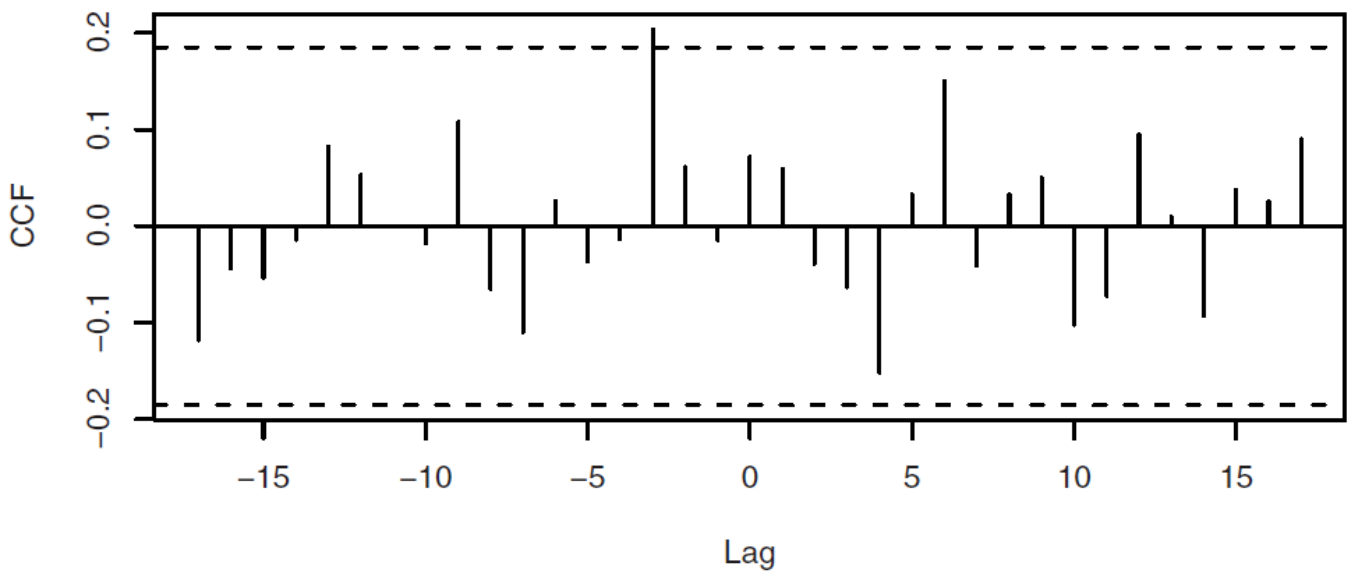

```
> ccf(as.vector(milk.electricity[,1]),
   as.vector(milk.electricity[,2]),ylab='CCF')
```

预白化与随机回归

- 由于$\sqrt{n}r_k(X,Y)$的方差渐近等于$1+2\sum_{k=1}^{\infty}\rho_k(X)\rho_k(Y)$，只要$X$和$Y$有一个是白噪声，问题可解！
- 预白化，就是为X序列拟合一个AR模型（可先做差分），其残差为白噪声，对Y序列做相同的操作，然后根据预白化后样本的互相关函数的显著性来选择回归变量。
- $\tilde{X}_t = X_t - \pi_1 X_{t-1} - \pi_2 X_{t-2} - \cdots$，$\tilde{X}_t$近似为白噪声。
- $\tilde{Y}_t = Y_t - \pi_1 Y_{t-1} - \pi_2 Y_{t-2} - \cdots$
- 由于预白化为线性运算，故原序列的任何线性关系都可保留。
- 预白化只是为了确定回归变量，回归时还是用原序列。

预白化与随机回归

Exhibit 11.16 Sample CCF of Prewhitened Milk and Electricity Production




```
> me.dif=ts.intersect(diff(diff(milk,12)),
   diff(diff(log(electricity),12)))
> prewhiten(as.vector(me.dif[,1]),as.vector(me.dif[,2]),
   ylab='CCF')
```

预白化与随机回归

方法总结如下：

1. 首先对X进行差分（如果需要）和建立AR模型，然后进行预白化，根据$\tilde{X}_t$和$\tilde{Y}_t$的样本CCF的显著性确定滞后项作为回归变量。
2. 进行简单线性回归，对于回归残差识别ARMA模型（包括季节模型）。
3. 统一拟合ARMA和回归模型，识别异常值，进行模型诊断，检验参数显著性等，直到没有异常值、通过模型诊断、所有参数都显著为止。

（实例见教材和演示）