Bivariate Series: An Introduction

Stat 443: Time Series

Dr. Bruce Dunham

Department of Statistics

Introduction

Often it is natural to examine two time series in tandem. Broadly there are two scenarios:

(a) The series are "on an equal footing", and interest lies mainly in correlations between the two. E.g., monthly male and female deaths due to heart disease.

(b) One series is the *input*, the other series being the corresponding *output*. Such systems are common in engineering.

We focus on situation (a).

Cross–correlation

Notation here mimics the univariate case: we have a *bivariate* stochastic process$\{(X(t), Y(t)), t = 1, 2, \ldots\}$ which has generated data $(x(1), y(1)), \ldots, (x(N), y(N))$.

We start with summary statistics for $(X(t), Y(t))$. We assume this process is (weakly) stationary, meaning here that for all t,

$$\begin{aligned}
E(X(t)) &= \mu_X, \\
E(Y(t)) &= \mu_Y, \\
\mathrm{Cov}(X(t), X(t+k)) &= \gamma_X(k), \\
\mathrm{Cov}(Y(t), Y(t+k)) &= \gamma_Y(k)
\end{aligned}$$

for $k \in \mathbb{Z}$, and moreover the *cross–covariance function,*

$$\begin{aligned}
\gamma_{XY}(k) : &= \mathrm{Cov}(X(t), Y(t+k)) \\
&= E((X(t) - \mu_X)(Y(t+k) - \mu_Y))
\end{aligned}$$

also depends only on the lag k, and not t.

Note that:

1. Many authors use
$$\gamma^*_{XY}(k) := \mathsf{Cov}(X(t), Y(t-k))$$
to define the cross–covariance function. Of course
$$\gamma^*_{XY}(-k) = \gamma_{XY}(k).$$

2. In general
$$\gamma_{XY}(k) \neq \gamma_{XY}(-k).$$
Instead,
$$\begin{aligned}
\gamma_{XY}(-k) &= \mathsf{Cov}(X(t), Y(t-k)) \\
&= \mathsf{Cov}(Y(t-k), X(t)) \\
&= \mathsf{Cov}(Y(s), X(s+k)) \\
&= \gamma_{YX}(k).
\end{aligned}$$

3. In fact both $X(t)$ and $Y(t)$ being stationary implies $(X(t), Y(t))$ is stationary.

The *cross–correlation* at lag k is defined as

$$\begin{aligned}\rho_{XY}(k) :&= \frac{\gamma_{XY}(k)}{(\gamma_X(0)\gamma_Y(0))^{\frac{1}{2}}} \\ &= \frac{\gamma_{XY}(k)}{\sigma_X\sigma_Y}\end{aligned}$$

where $\text{Var}(X(t)) = \sigma_X^2 = \gamma_X(0)$ and $\text{Var}(Y(t)) = \sigma_Y^2$.

Note that

1. From the above,

$$\rho_{XY}(k) = \rho_{YX}(-k).$$

2. It can be shown that

$$|\rho_{XY}(k)| \leq 1.$$

3. The "marginal" correlations say little about $\rho_{XY}(k)$, in particular,

$$\rho_X(0) = \rho_Y(0) = 1 \not\Rightarrow \rho_{XY}(0) = 1.$$

Example 0.1 *Suppose*

$$X(t) = Z(t),$$

$$Y(t) = 0.2Z(t-1) + 0.2Z(t-2),$$

where $Z(t)$ is white noise with variance σ^2. Then

$$\gamma_{XY}(k) = Cov(X(t), Y(t+k))$$

$$= Cov(Z(t), 0.2[Z(t-1+k) + Z(t-2+k)])$$

which is 0.2σ^2 when $k = 1, 2$, zero otherwise. Since

$$\gamma_X(0) = \sigma^2,$$

$$\gamma_Y(0) = 0.2^2\left(\sigma^2 + \sigma^2\right)$$

$$= 0.08\sigma^2$$

then

$$\rho_{XY}(k) = \begin{cases} \frac{0.2}{(0.08)^{\frac{1}{2}}} & k = 1, 2 \\ \\ 0 & \textit{otherwise.} \end{cases}$$

Example 0.2 *Let*

$$X(t) = \alpha X(t-1) + Z_1(t)$$

where $|\alpha| < 1$ *and* $Z_1(t)$ *is white noise with variance* σ_1^2, *and for some integer* j *let*

$$Y(t) = \beta X(t+j) + Z_2(t)$$

where $|\beta| < 1$ *and* $Z_2(t)$ *is white noise with variance* σ_2^2, *independent of* $X(t)$ *and* $Z_1(t)$. *Now we recall that,*

$$\gamma_X(k) = \frac{\sigma_1^2 \alpha^{|k|}}{\left(1 - \alpha^2\right)},$$

and moreover

$$\begin{aligned}
\gamma_Y(k) &= E(Y(t)\,Y(t+k)) \\
&= E((\beta X(t+j) + Z_2(t)) \\
&\qquad \times (\beta X(t+j+k) + Z_2(t+k))) \\
&= \beta^2 E(X(t+j)\,X(t+j+k)) \\
&= \beta^2 \gamma_X(k)
\end{aligned}$$

unless $k = 0$ *when*

$$\begin{aligned}
\gamma_Y(0) &= E\left(Y(t)^2\right) \\
&= \beta^2 E\left(X(t+j)^2\right) + E\left(Z_2(t)^2\right) \\
&= \frac{\beta^2 \sigma_1^2}{\left(1-\alpha^2\right)} + \sigma_2^2 \\
&= \beta^2 \gamma_X(0) + \sigma_2^2.
\end{aligned}$$

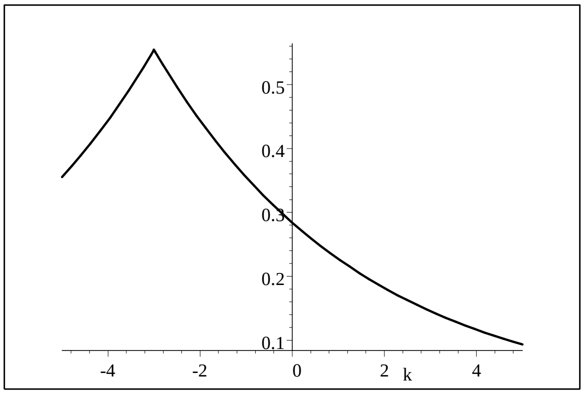
The cross–covariance is

$$\begin{aligned}
\gamma_{XY}(k) &= E\left(X(t)\,Y(t+k)\right) \\
&= E\left(X(t)\left(\beta X(t+j+k)+Z_2(t+k)\right)\right) \\
&= \beta\gamma_X(k+j)
\end{aligned}$$

and the cross–correlation is

$$\begin{aligned}
\rho_{XY}(k) &= \frac{\beta\gamma_X(k+j)}{\left(\gamma_X(0)\left(\beta^2\gamma_X(0)+\sigma_2^2\right)\right)^{\frac{1}{2}}} \\
&= \frac{\gamma_X(k+j)}{\left(\gamma_X(0)\left(\gamma_X(0)+\sigma_2^2/\beta^2\right)\right)^{\frac{1}{2}}} \\
&= \frac{\alpha^{|k+j|}}{\left(1+\sigma_2^2\left(1-\alpha^2\right)/\left(\beta^2\sigma_1^2\right)\right)^{\frac{1}{2}}}
\end{aligned}$$

which is maximised when $k = -j$.



Case with $\alpha = 0.8$, $\beta = 0.4$, $\sigma_1 = \sigma_2 = 1$, $j = 3$

Estimation of the cross–correlation

Given bivariate data $(x(1), y(1)), \ldots, (x(N), y(N))$, how to estimate $\rho_{XY}(k)$? The *sample cross–covariance* function is

$$c_{XY}(k) := \frac{1}{N} \sum_{t=1}^{N-k} (x(t) - \bar{x})(y(t+k) - \bar{y})$$

for $k = 0, 1, \ldots, N$ and

$$c_{XY}(k) := \frac{1}{N} \sum_{t=1-k}^{N} (x(t) - \bar{x})(y(t+k) - \bar{y})$$

for $k = -1, \ldots, -(N-1)$, the *sample cross–correlation* being

$$r_{XY}(k) := \frac{c_{XY}(k)}{s_X s_Y}.$$

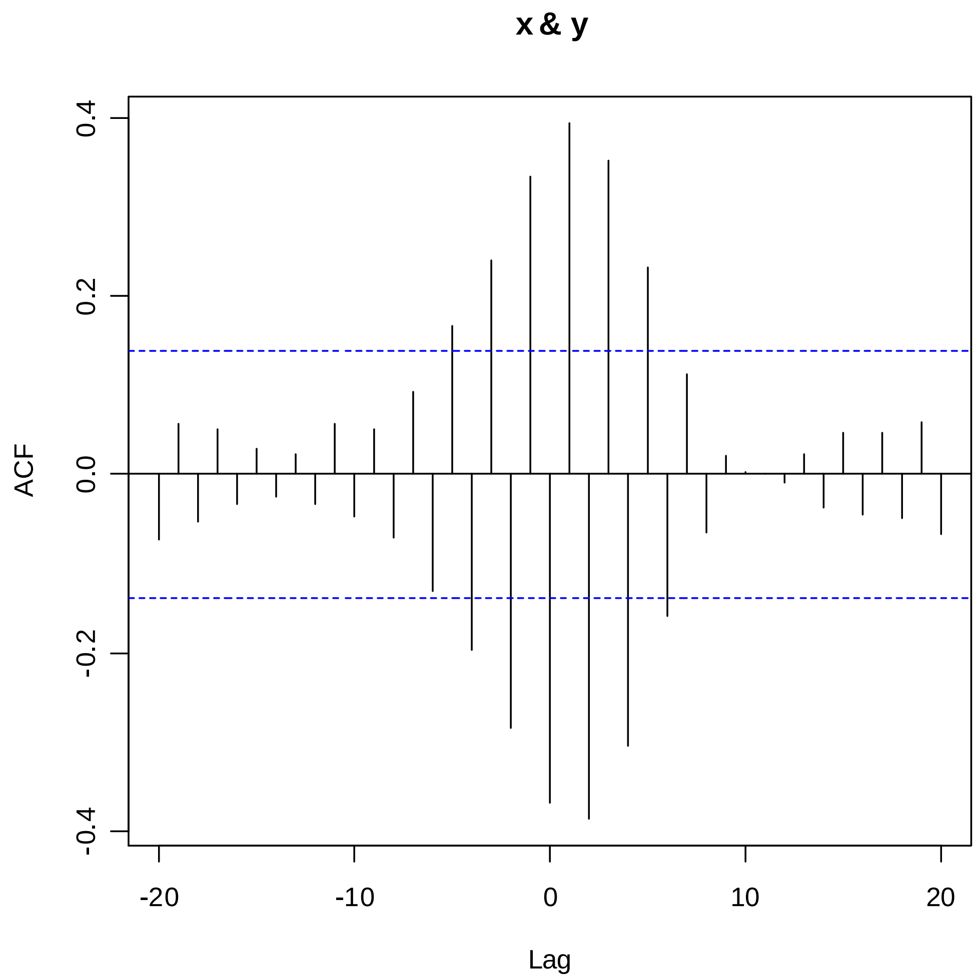
Now

- $r_{XY}(k)$ is asymptotically unbiased and consistent for $\rho_{XY}(k)$ but

- values at neighbouring lags are correlated.

- The acf for the individual series can inflate the values of r_{XY}, even when the two series are mutually independent.

Example 0.3 *Suppose* $\{x\,(t)\}$ *and* $\{y\,(t)\}$ *are both from an AR*(1) *with* $\alpha = -0.9$, *but are mutually independent. Then both series will tend to alternate in (and sometimes out of) phase, making absolute values of* r_{XY} *large.*



r_{XY} *for independent samples of 200 from AR*(1), $\alpha = -0.9$.

The suggestion is to *pre-whiten* at least one of the series, to make it (or both) resemble a white noise realisation. We can either

1. Apply a linear filter (like a moving average), and subtract the smoothed series from the original.

2. Fit an ARIMA model to the data, then subtract the fitted values leaving the residuals.

For two mutually uncorrelated series (at least) one of which is white noise, then

$$E\left(r_{XY}(k)\right) \approx 0,$$

$$\mathrm{Var}\left(r_{XY}(k)\right) \approx \frac{1}{N},$$

and by the CLT values outside $\pm 2/N^{\frac{1}{2}}$ could be deemed significant.

Remark 0.1 *Pre-whitening does affect the cross–correlation, however. In general, suppose*

$$U(t) = \sum_{i=0}^{\infty} a_i X(t+i)$$

and

$$V(t) = \sum_{j=0}^{\infty} b_i Y(t+j)$$

are stationary processes, then

$$\gamma_{UV}(k) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} a_i b_j \gamma_{XY}(k+i-j).$$

Bivariate series in R

Time–domain analysis of bivariate series in R is now described, and interpreting the cross–correlogram discussed. Bivariate (and multivariate) time series can be created from two univariate series in two ways: Firstly

> `ts.union(series 1, series 2, ..., dFRAME = F)`
combines two series with over overlapping time spans, padding with `NA`'s if necessary – it acts like `cbind`. Output may be stored as a dataframe.

> `ts.intersect(series 1, series 2, ..., dFRAME = F)`
is similar, but retains values only at times where both series are observed.

Care should be taken when plotting. The `plot` command has several options:

> `plot(x, y, plot.type = c(''multiple'', ''single''), xy.labels, xy.lines, panel = lines, axes = TRUE, ...)`
where `x` and `y` are time series, or `x` is bivariate and `y=NULL`.

To find the cross–correlation function, use either

> `ccf(x, y, lag.max = NULL, type = c(''correlation'', ''covariance''),`

```
plot = TRUE,...)
```

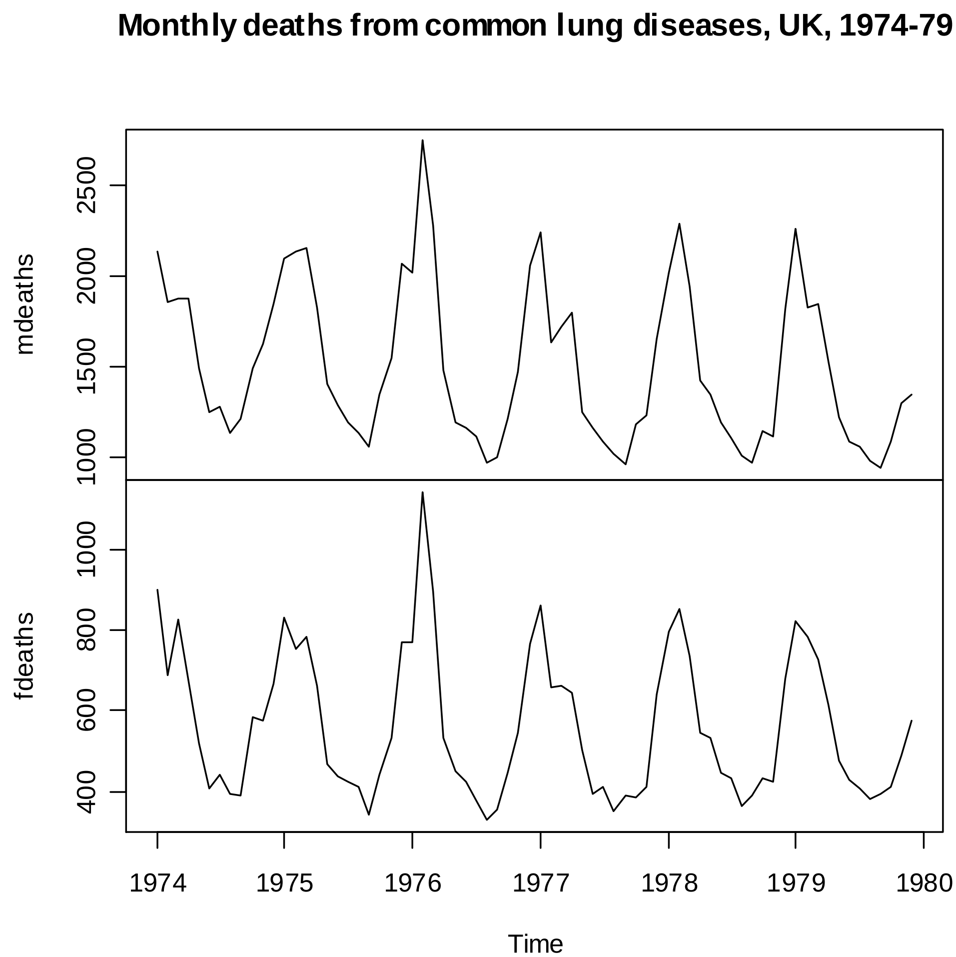
or

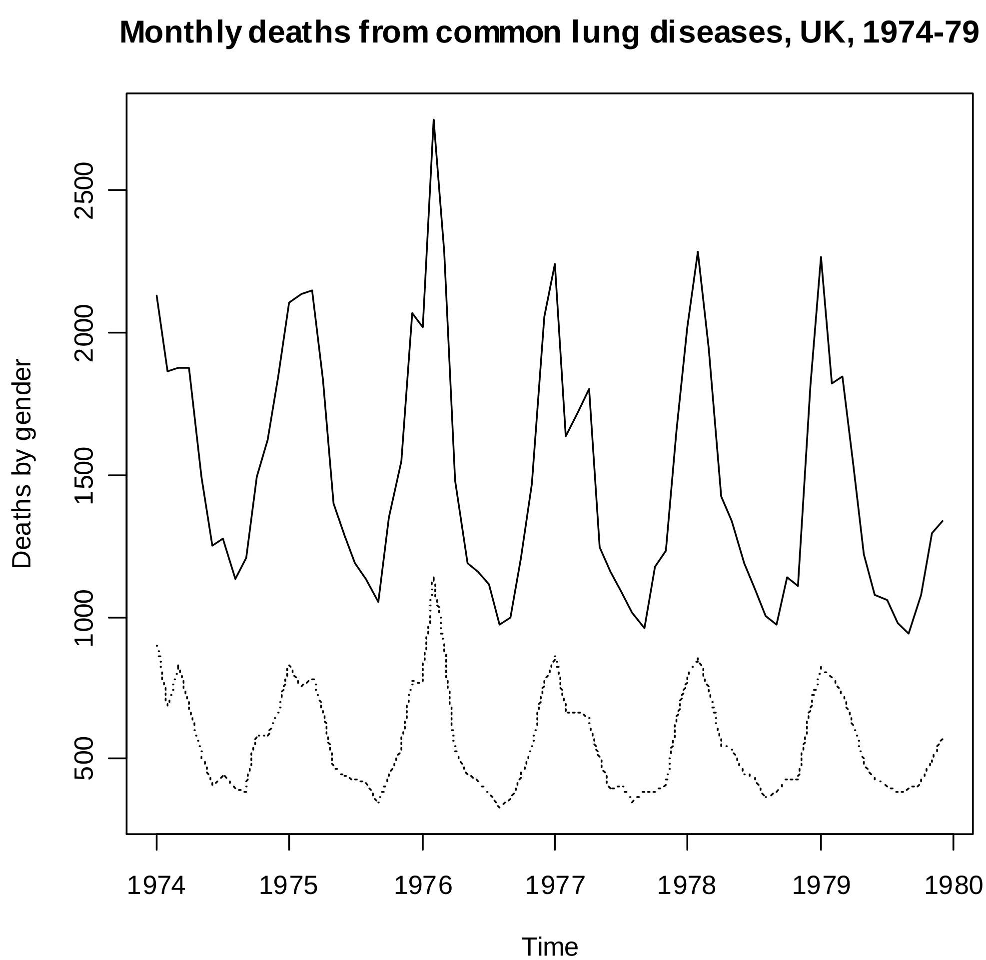
```
> acf(x, ...)
```

on a bivariate series. The latter produces the acf for both univariate series as well as the cross–cf. – perhaps confusingly with c_{XY} for negative lags and c_{YX} for positive.

An example

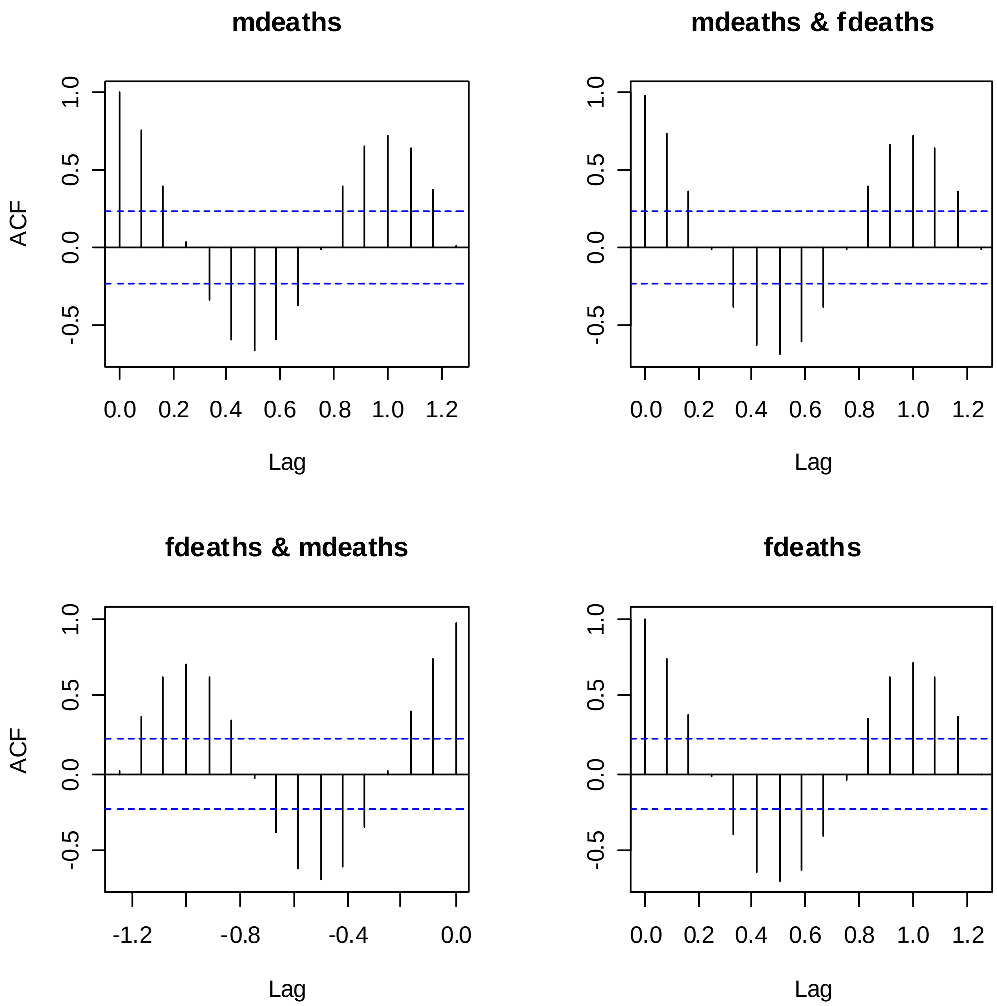
Data–set `deaths` contains monthly deaths from bronchitis, emphysema and asthma in the UK, 1974–1979, for both sexes.

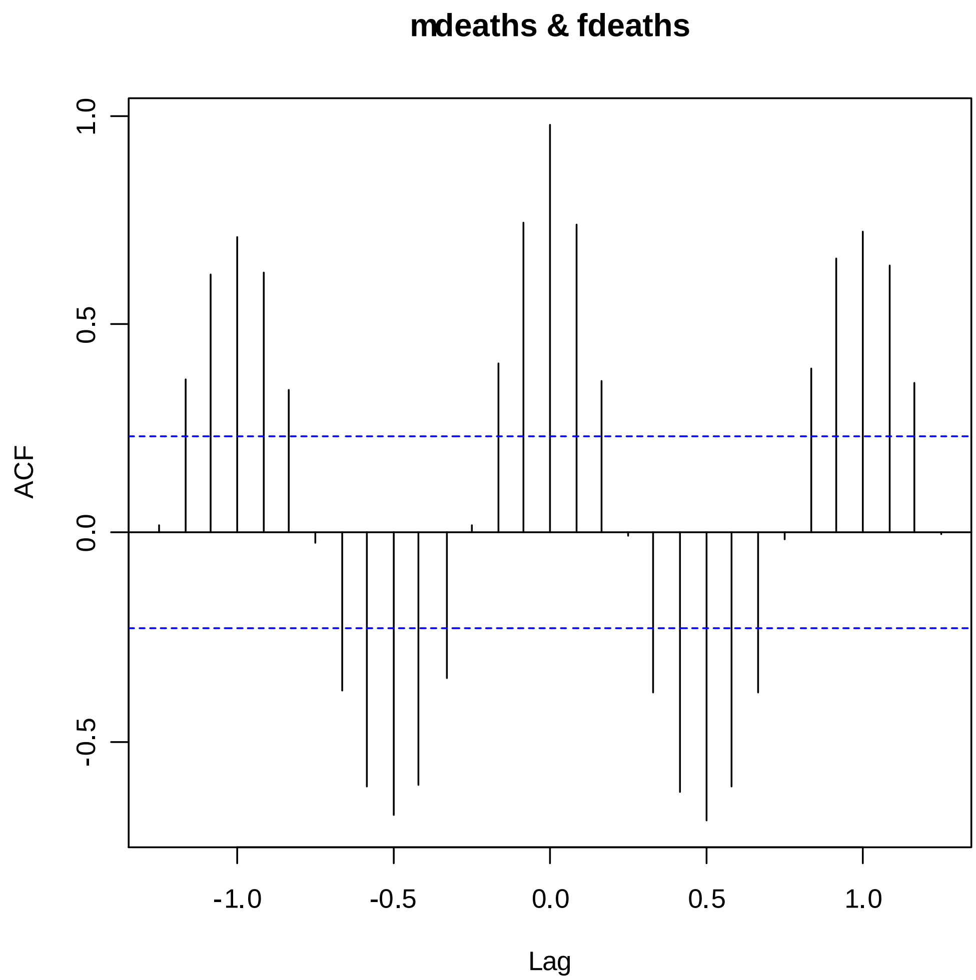




Solid line shows male deaths, dotted female deaths.

The acfs and cross–correlations are given below:

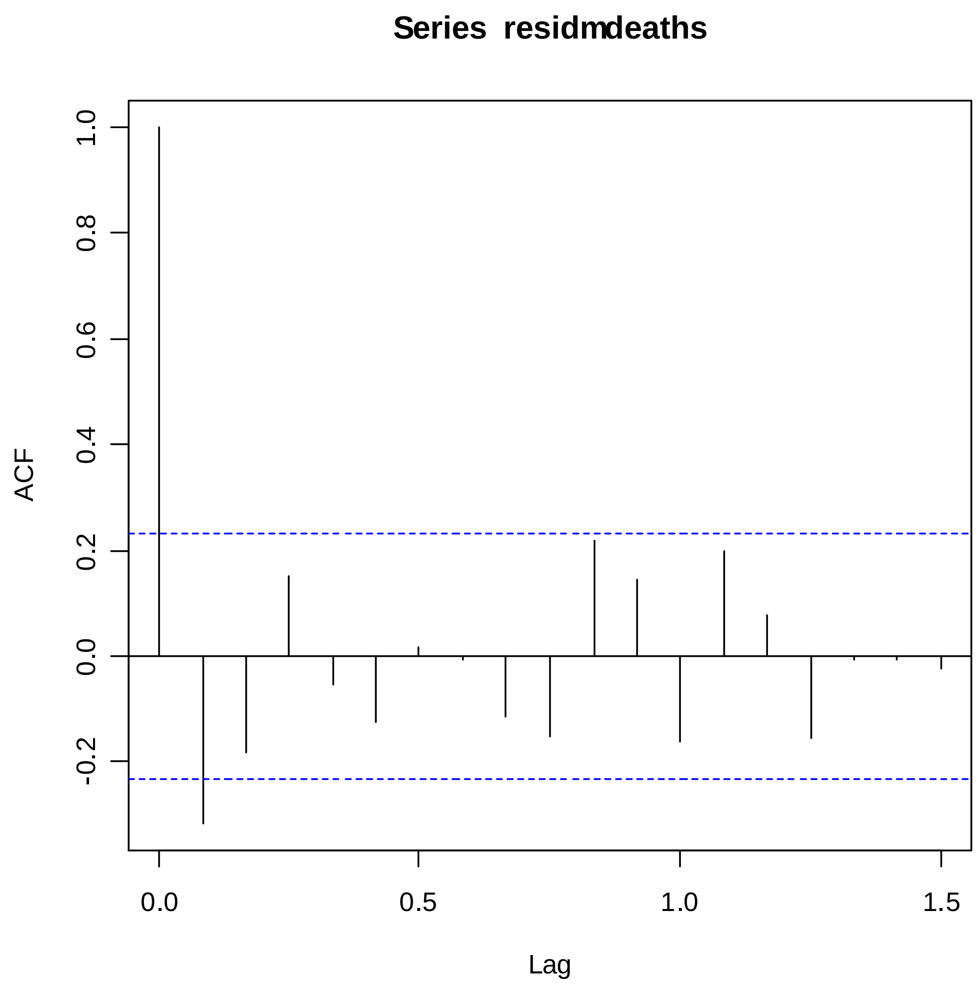


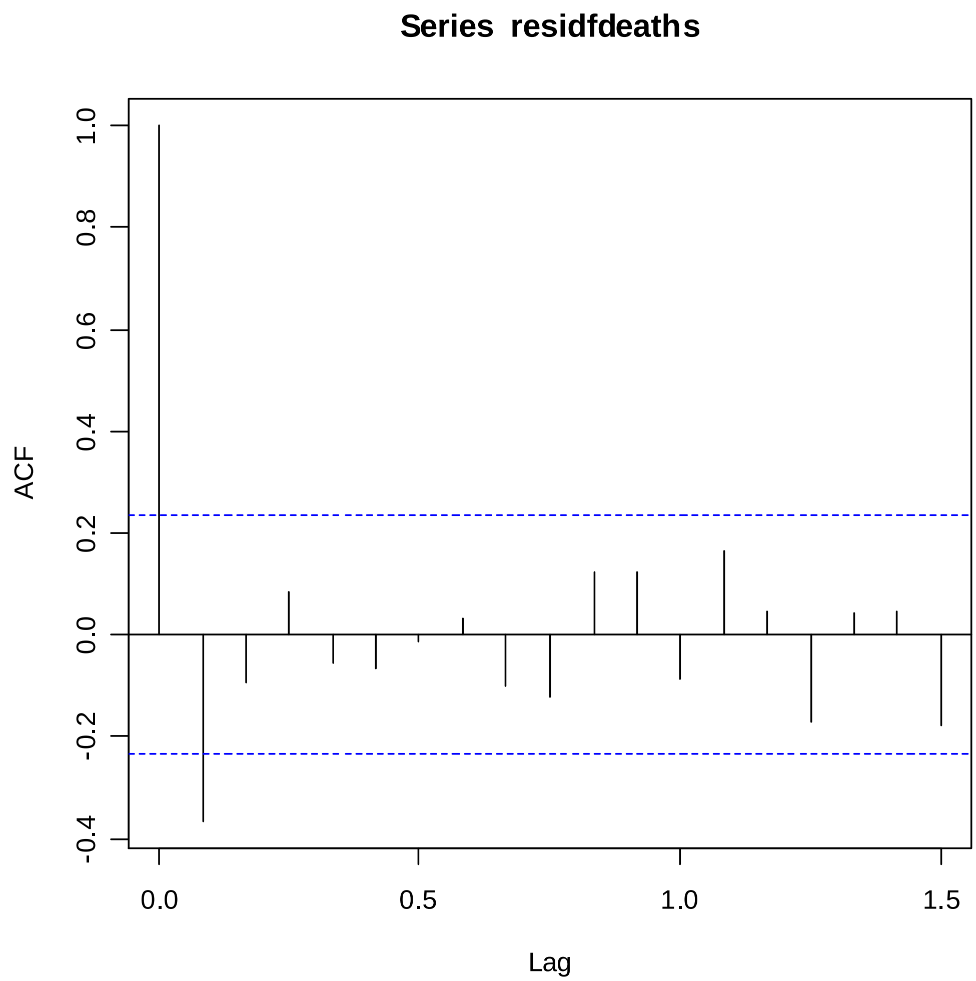


Typical for seasonal series, the acf/ccf values do not dampen down for large lags.

More interesting is to pre-whiten the two series by removing an estimate of the seasonal variation from each – we can do this by applying a 3–point moving average to each, then subtracting from the original series to leave the residuals. (I used the `rollmean` command for this, but `stl` is more sophisticated.)

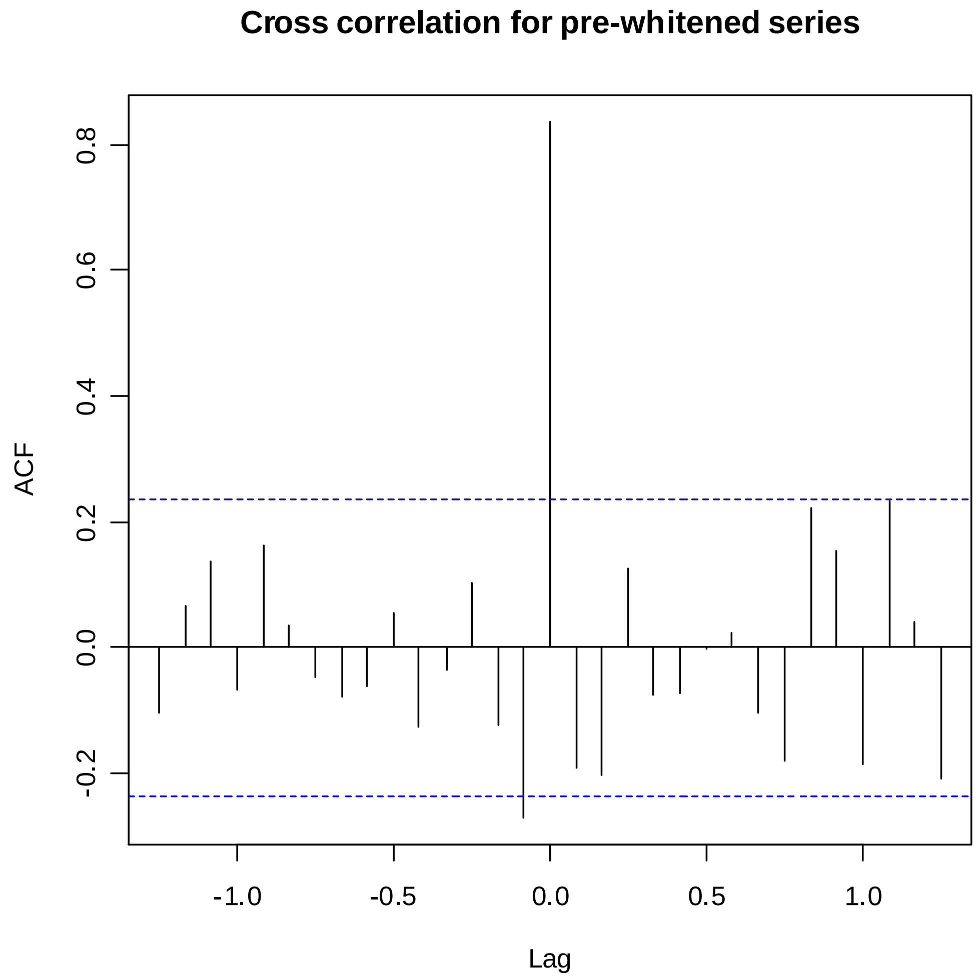
The acfs of the residuals are below:



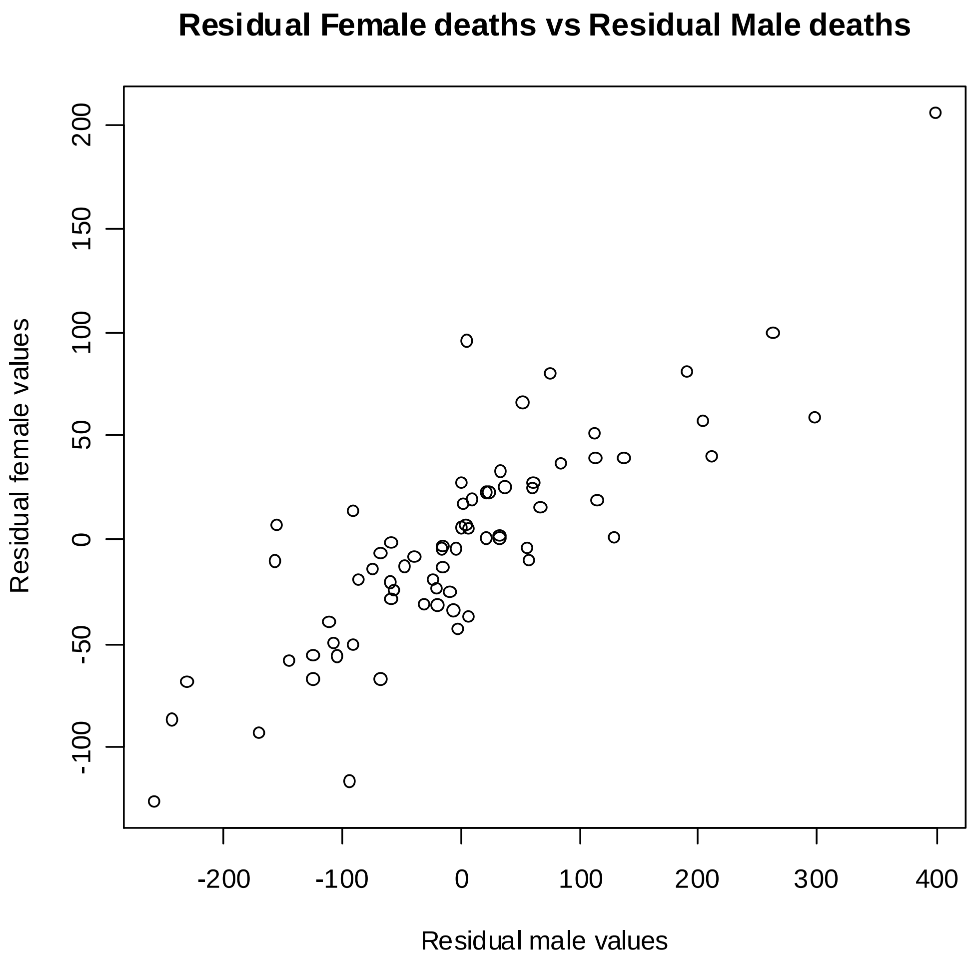


Note pre-whitening not totally successful – both residual series have significant autocorrelation at lag –1.

The cross–correlation is now:



There is a significant positive correlation at lag 0, with a possibly significant negative correlation at lag –1. The former is illustrated below, indicating an approximately linear relationship:



The far–flung point is January 1976.

The above suggests that time – and presumably the associated weather – greatly determines the number of deaths, rather than, say, the number of susceptibles in each gender.

Difficulties with the cross–correlation

The sample cross–correlation function can be difficult to interpret, even with pre-whitened data.

1. A significant value at lag k indicates the two series are related when one is delayed by k time units.

2. However, neighbouring lags may also appear significant.

3. The sampling distribution of r_{XY} depends on the acfs of the individual series. In particular, Var($r_{XY}(k)$) is

$$\frac{1}{N}\sum_{j=-\infty}^{\infty}\left[\rho_X(j)\,\rho_Y(j)+\rho_{XY}(j+k)\,\rho_{XY}(j-k)\right].$$

4. When (at least one) series not pre-whitened, analysis is tricky, especially if series are non–stationary.

5. If data are “aggregated” (say, from monthly to quarterly) key information may be lost.