CENTRE DE RECHERCHE PUBLIC

Modelling periodic data

A note prepared with Mathcad(R)

ORIGIN := 1

Background and an example

A problem that sometimes arises is that data are of a periodic nature. Take the example of the monthly average temperatures in Luxembourg

Input indicators for the 12 months

$i := 1, 2 .. 12$ $\quad$ $Month_i := i$

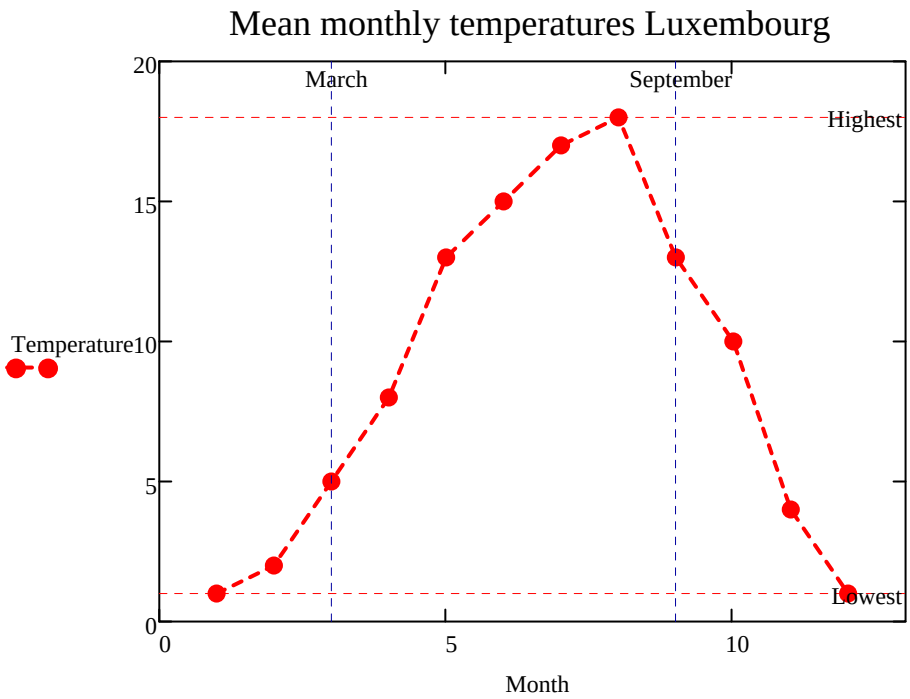
$Month^T = (1\ \ 2\ \ 3\ \ 4\ \ 5\ \ 6\ \ 7\ \ 8\ \ 9\ \ 10\ \ 11\ \ 12)$ $\quad$ $March := 3$ $\quad$ $September := 9$

Input temperature readings

$Temperature := (1\ \ 2\ \ 5\ \ 8\ \ 13\ \ 15\ \ 17\ \ 18\ \ 13\ \ 10\ \ 4\ \ 1)^T$

$Highest := max(Temperature)$ $\quad$ $Lowest := min(Temperature)$

Plot data



CENTRE DE RECHERCHE PUBLIC

Using a periodic function

It might appear to be rather difficult to find a simple function of time of year that enables us to approximate the temperature. However, since the data are periodic (month 1, which is January, is actually closer to month 12, which is December, than it is to month 4, which is April), simple periodic functions will do the job better than standard polynomials.

The simplest such function is a *sine wave* which, when the period is known, we can represent as follows

$$Y(X,\gamma,\beta,\alpha) := \gamma + \beta\cdot\sin(X + \alpha)$$ Define sine wave function as Y( )

Here *X* is a periodic time variable, *Y* is some response, γ is the average value of the *level* of the response, β is the *amplitude* of the sine wave and α is the *phase.*

Interpretation of the parameters of the function

$$\int_0^{2\pi} \gamma + \beta\cdot\sin(X + \alpha)\, dX \rightarrow 2\cdot\gamma\cdot\pi$$ This is the area under the curve (AUC)

Check. Note how the area under the curve (AUC) between 0 and 2π is 2γπ and this means that if we divide by the length of the interval, 2π, we get the mean height of the function as γ. Hence γ is the mean or *level* of the function.

$$\frac{d}{d\alpha}(\gamma + \beta\cdot\sin(X + \alpha)) \rightarrow \beta\cdot\cos(X + \alpha)$$ This is the derivative of the function. Since cos(π/2+*n*π) = 0, (where *n* is an integer) the function reaches a maximum and a minimum when X + α = π/2+*n*π. This is illustrated below

$$n := -3, -2 .. 3$$

n =	$\cos\left(\frac{\pi}{2} + n\cdot\pi\right)$ =
-3	0
-2	0
-1	0
0	0
1	0
2	0
3	0

Hence α governs the point at which the function reaches a maximum. It is referred to as the *phase* of the function.

$$\gamma + \beta\cdot\sin\left(\frac{\pi}{2}\right) - \left(\gamma + \beta\cdot\sin\left(\frac{3\pi}{2}\right)\right) \rightarrow 2\cdot\beta$$

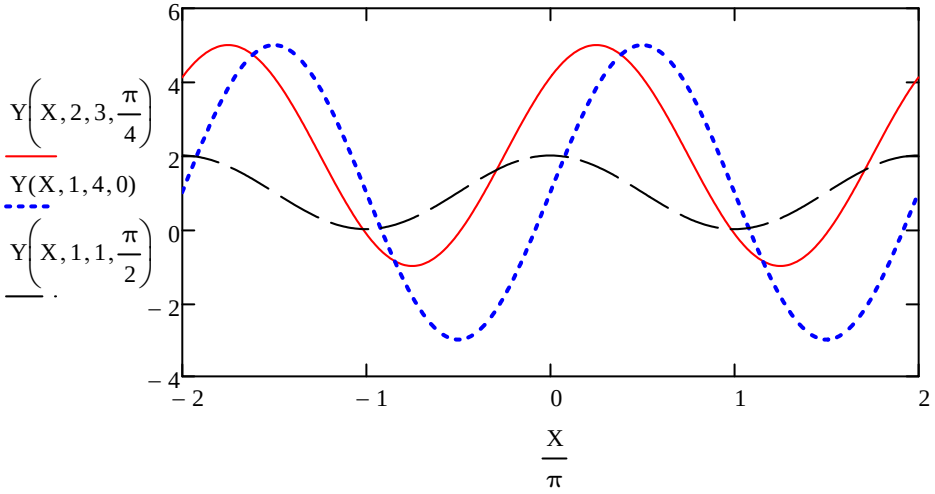
This is the difference between the peak and the trough of the function and is thus twice the *amplitude*. Hence β is the amplitude of the function

Produce various plots to illustrate behaviour of the function

Define range of plot in radians.

$X := -2\cdot\pi, -1.99\pi .. 2\cdot\pi$ This should cover two full cycles

Sine wave illustrated: various parameter settings



Time in radians of the cycle

Fitting the function

Obviously we need to find values of γ,α and β to fit the function. We can do this by linearising the function and using least squares

NB It is *essential* that the predictor time variable be measured in radians over the period. Thus, in our example we need to transform the time variable first by using 2πmonth/12.

CENTRE DE RECHERCHE PUBLIC

Linearisation

We use the following simple general trignometric relationship. (One of the famous 'double angle' formulae

$$\sin(x + \alpha) \text{ expand } \rightarrow \cos(\alpha)\cdot\sin(x) + \sin(\alpha)\cdot\cos(x)$$

Thus we create two new variables *sin(x)* and *cos(x)* and fit these in a multiple regression. This multiple regression will produce estimates of parameters as follows

1. Intercept γ
2. Coefficient of *sin(x)*, $\beta cos(\alpha)$ (which we can call θ_1)
3. Coefficient of *cos(x)*, $\beta sin(\alpha)$ (which we can call θ_2)

In other words we simply create predictor variables *sin*(x) and *cos*(x) and regress *Y* on these to obtain estimates of γ, θ_1 and θ_2.

Orthogonality

In fact these new constructed variables have an interesting property in that they are *orthogonal*. This means that their sums and the sum of their products over one period of the function is equal to zero

Illustration of orthogonality for the continuous mathematical function

$$\int_0^{2\cdot\pi} \sin(x)\,dx \rightarrow 0 \qquad \int_0^{2\cdot\pi} \cos(x)\,dx \rightarrow 0 \qquad \int_0^{2\cdot\pi} \sin(x)\cdot\cos(x)\,dx \rightarrow 0$$

Illustration of orthogonality for the statistical function applied to the example of 12 months

$$\sum_{j=1}^{12} \sin\left(\frac{j\cdot 2\pi}{12}\right) \rightarrow 0 \qquad \sum_{j=1}^{12} \cos\left(\frac{j\cdot 2\pi}{12}\right) \rightarrow 0 \qquad \sum_{j=1}^{12} \left(\sin\left(\frac{j\cdot 2\pi}{12}\right)\cdot\cos\left(\frac{j\cdot 2\pi}{12}\right)\right) \rightarrow 0$$

Solution to our example

This means that in our example a particularly simple estimation process is possible. All the parameters can be obtained as simple weighted means

$$\gamma := \frac{\sum_{i=1}^{12} \text{Temperature}_i}{12} \qquad \gamma = 8.917$$

$$\theta_1 := \frac{\sum_{i=1}^{12}\left[\left(\text{Temperature}_i\right)\cdot\sin\left(\frac{\text{Month}_i\cdot 2\cdot\pi}{12}\right)\right]}{\sum_{i=1}^{12}\sin\left(\frac{\text{Month}_i\cdot 2\cdot\pi}{12}\right)^2}$$

This is based on the standard formula for a regression

$\theta_1 = -4.515$

$$\theta_2 := \frac{\sum_{i=1}^{12}\left[\left(\text{Temperature}_i\right)\cdot\cos\left(\frac{\text{Month}_i\cdot 2.\pi}{12}\right)\right]}{\sum_{i=1}^{12}\cos\left(\frac{\text{Month}_i\cdot 2\cdot\pi}{12}\right)^2}$$

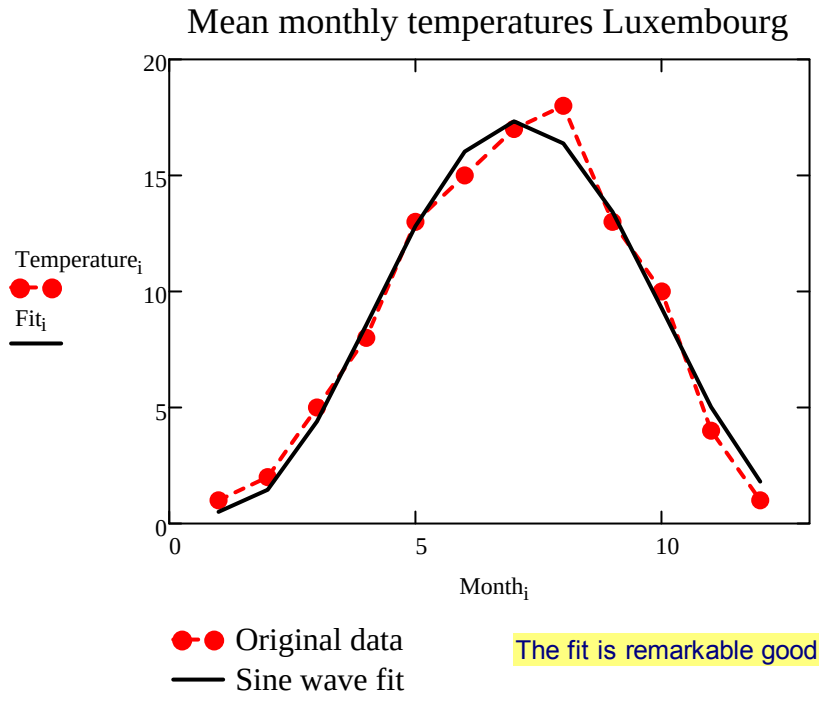
$\theta_2 = -7.108$

However, in practice one would not do this since it would be simpler to use a standard regression procedure available in any package.

Displaying the fit

First we calculate the fitted values

$$\text{Fit}_i := \gamma + \theta_1\cdot\sin\left(\frac{\text{Month}_i\cdot 2\cdot\pi}{12}\right) + \theta_2\cdot\cos\left(\frac{\text{Month}_i\cdot 2\cdot\pi}{12}\right)$$

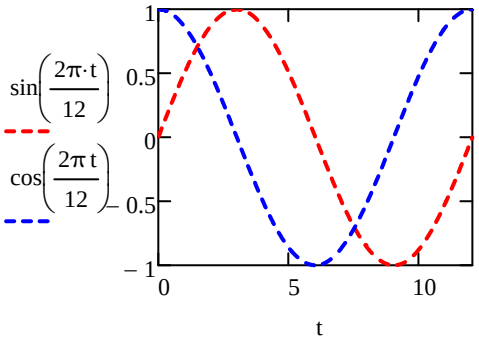


The fit is remarkable good!

Why does this work?

It works because any sine wave function can be expressed as the weighted sum of a sine and cosine.

$j := 1, 2 .. 1000$ $\quad t_j := \frac{12 \cdot j}{1000}$



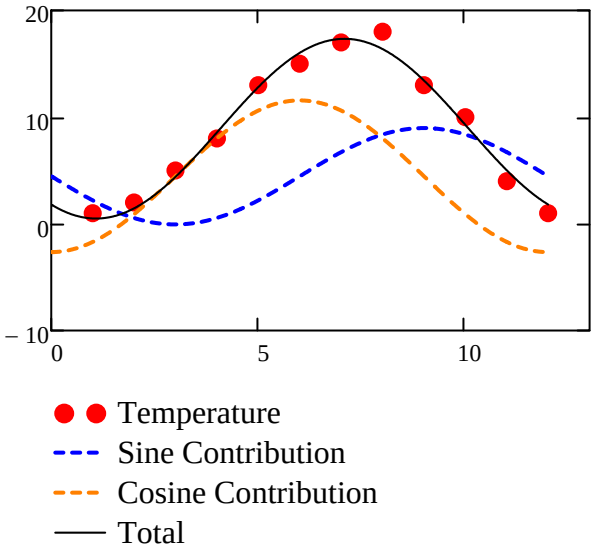
The fundamental sine and cosine functions

We can illustrate this using the plot below

$f_1(x) := \frac{\gamma}{2} + \theta_1 \cdot \sin\left(\frac{2 \cdot \pi \cdot x}{12}\right)$ Sine contribution NB $\theta_1 = -4.515$

$f_2(x) := \frac{\gamma}{2} + \theta_2 \cdot \cos\left(\frac{2 \cdot \pi \cdot x}{12}\right)$ Cosine contribution $\theta_2 = -7.108$

$f_3(x) := f_1(x) + f_2(x)$ Total of the two



Note how the basic sine and cosine functions remain. They have been inverted (maximum becomes minimum and *vice versa*) because θ_1 and θ_2 are both negative. Apart from this they have not shifted position in the horizontal axis. Their average has been lifted by an amount $\gamma/2$ and their amplitude has been modified by the absolute values of θ_1 and θ_2 .

Back transformation

It will not usually be the case that we actually need the values of α and β, however, they can be obtained as follows. We note the following

$$(\beta\cdot\cos(\alpha))^2 + (\beta\cdot\sin(\alpha))^2 \text{ simplify } \rightarrow \beta^2$$

$$\frac{\beta\cdot\sin(\alpha)}{\beta\cdot\cos(\alpha)} \text{ simplify } \rightarrow \tan(\alpha)$$

Hence we can calculate β as the square root of $\theta_1^2 + \theta_2^2$.

We can calculate α as $\tan^{-1}(\theta_1/\theta_2)$.

Application to our example

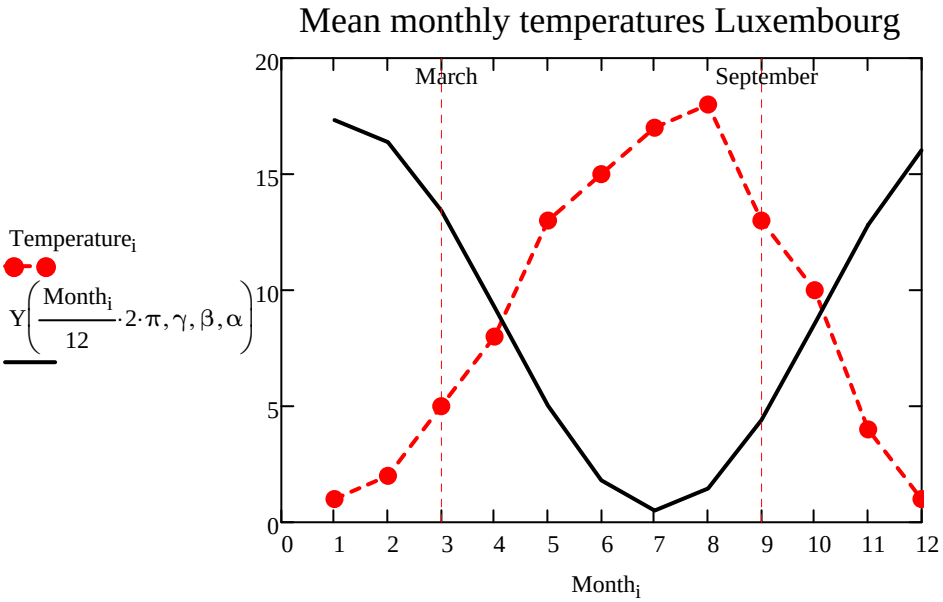
$$\sqrt{\theta_1^2 + \theta_2^2} = 8.421 \qquad \text{atan}\left(\frac{\theta_2}{\theta_1}\right) = 1.005$$

atan is Mathcad's built in inverse tangent or *arc tan* function

$$\beta := \sqrt{\theta_1^2 + \theta_2^2} \qquad \alpha := \text{atan}\left(\frac{\theta_2}{\theta_1}\right)$$

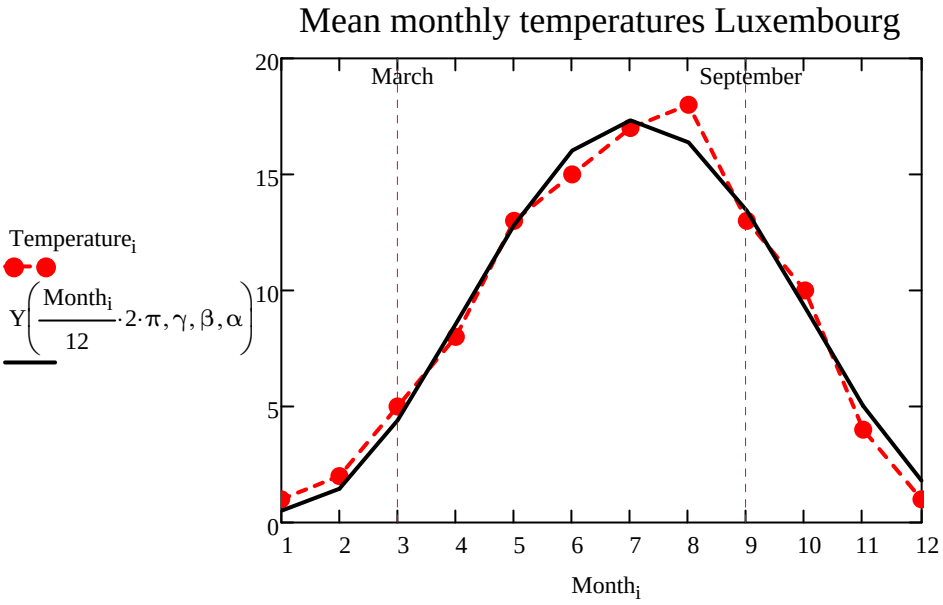
Note that if we wish to measure α in months rather than radians, we need to multiply by 12 and divide by 2π

$$\alpha_{\text{months}} := \frac{\alpha\cdot 12}{2\cdot\pi} \qquad \alpha_{\text{months}} = 1.919$$



This is a very bad fit! So what has gone wrong? Let us try the other root

$$\beta := -\sqrt{\theta_1^{\,2} + \theta_2^{\,2}}$$

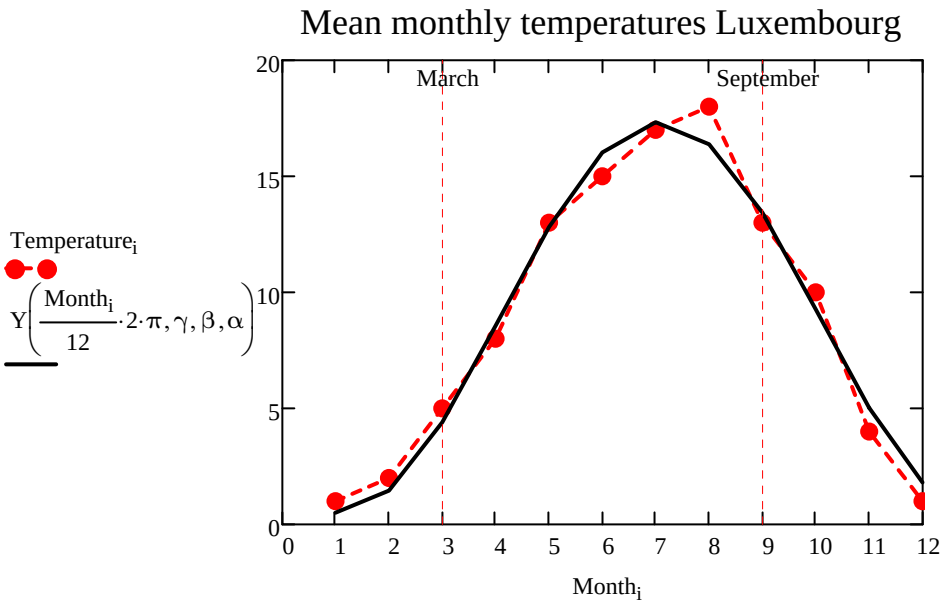


That's better !

Alternatively we can change the value of α. The fact that θ_1,θ_2 are both negative implies that we are in the third quadrant of the cycle so we have

$$\alpha := \operatorname{atan}\left(\frac{\theta_2}{\theta_1}\right) + \pi \qquad \tan(\alpha) = 1.574 \qquad \frac{\theta_2}{\theta_1} = 1.574$$ Check

$$\alpha_{months} := \frac{\alpha \cdot 12}{2 \cdot \pi} \qquad \alpha_{months} = 7.919 \qquad \beta := \sqrt{\theta_1^{\,2} + \theta_2^{\,2}}$$



Testing for significance of a periodic trend

In order to test for the significance of a periodic trend we simply compare the model with a trend to the model without. Analysis of deviance is the general approach one may use for this and we just need to note that we need a 2 DF test to compare the models. This is illustrated in the extract of an analysis with Genstat(R) below.

CENTRE DE RECHERCHE PUBLIC

Adding period terms

Regression analysis

Response variate:	Temperature
Fitted terms:	Constant + CosDate + SinDate

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	2	425.477	212.7384	257.35	<.001
Residual	9	7.440	0.8266		
Total	11	432.917	39.3561		
Change	-2	-425.477	212.7384	257.35	<.001

Percentage variance accounted for 97.9
Standard error of observations is estimated to be 0.909.

Here *CosDate* and *SinDate* are the trigonometric transformations of month and the change in the sums of squares by dropping them from the model 425.5 on 2 degrees of freedom giving a mean square error of 212.7 and a variance ratio of 212.7/0.827 = 257.4, which not surprisingly is highly significant.

Warning

It is ***not*** generally appropriate to use Wald tests ***separately*** on θ_1 and θ_2 and it is not appropriate to drop either the sine or the cosine term from the model because not significant. The one exception would be where we knew for sure what the phase of the model should be. In our example we know it approximately. We know that the function should reach a maximum towards the middle of the year and should reach a minimum towards the end but we also know that temperature tends to lag behind insolation, so that although we know that 21 June and 21 December are the solstices we do not know that these the maximum and minimum points. Hence it is wise to model as if we do not know when the maximum will be and let the data decide.

Comments, applications and extensions

This approach may seem complicated but all that this is is a simple transformation of a predictor variable. It may seem strange that two new variables are created from one but of course, when we add a variable X^2 to a model in which we already have X we are in fact using two predictor variables. Here we could try use a quadratic as an alternative to the sine wave but it would not fit as well and it would require us to make a judgement as to what point to use as the start of the period.

The sine wave is a simple example of Fourier analysis and by adding further periodic terms it would be possible to improve the fit.

Many periodic phenomena can be modelled using this general approach. Examples could include:

weekly cases of hospitalisation with an annual cycle
hourly blood pressure readings if a circadian (daily) rhythm is suspected
daily hormone levels with a monthly cycle

However, note that if a phenomenon is believed to be periodic but its period is not known then a modification as follows is needed

$$\gamma + \beta \cdot \sin\left(\frac{x + \alpha}{\delta}\right)$$

A function of this form is *not* linear in the parameters and cannot be analysed using ordinary least squares. A genuine non-linear optimisation program is needed.

If there is a time series running over many cycles then a general trend can be included. For example one could have daily hospitalisation rates over many years. The cycle could be set at one year, there could be a general linear trend term and there could be a dummy variable to allow for the difference between weekends and weekdays. In other words the use of sine and cosine terms can be incorporated into a more general framework.