EMERGING THEMES
IN EPIDEMIOLOGY

METHODOLOGY　　　　　　　　　　　　　　　　　　　　　**Open Access**

The coefficient of cyclic variation: a novel statistic to measure the magnitude of cyclic variation

Anthony JC Fulford[1,2]

Abstract

Background: Periodic or cyclic data of known periodicity are frequently encountered in epidemiological and biomedical research: for instance, seasonality provides a useful experiment of nature while diurnal rhythms play an important role in endocrine secretion. There is, however, little consensus on how to analysis these data and less still on how to measure association or effect size for the often complex patterns seen.

Results: A simple statistic, readily derived from Fourier regression models, provides a readily-understood measure cyclic variation in a wide variety of situations.

Conclusion: The coefficient of cyclic variation or similar statistics derived from the variance of a Fourier series could provide a universal means of summarising the magnitude of periodic variation.

Keywords: Periodic regression, Circular data, Truncated fourier series, Coefficient of circular variation, Cosinor analysis

Introduction

Fourier or trigonometric regression is one of the most powerful methods for the analysis of periodic data when the cycle length is known. It is a natural generalisation of the familiar cosinor regression [1]. The method fits a linear regression model in which the cyclic component, $c(\theta_i)$, is represented by a truncated Fourier series, e.g.:

$$
\begin{aligned}
y_i &= \beta_0 + \beta^T x_i + c(\theta_i) + \varepsilon_i \\
&= \beta_0 + \beta^T x_i + \sum_{j}^{P} \left[\alpha_j sin(j\theta_i) + \gamma_j cos(j\theta_i) \right] \\
&\quad + \varepsilon_i
\end{aligned}
\tag{1}
$$

where θ_i is the angle in radians corresponding to the point in the cycle at which the ith point was measured, α_i and γ_j are the coefficients to be estimated and P is the number of pairs of terms in the truncate Fourier series. The $\beta^T x_i$ term is the linear combination of the other covariates (if any) fitted by the regression. Such models are simple to use, may be implemented with almost any statistical software and have found many and varied application (e.g. a thorough presentation by Fernandez *et al* [2];

possibly the first implementation was by Bliss [3]), although are possibly still not as widely used as they should be [4]. Their form is naturally cyclic and smooth, avoiding the unrealistic steps introduced when the period is discretised. Also, by varying the point at which the Fourier series is truncated, it is possible to determine the degree of detail fitted and avoid over-parameterising the model. Indeed, higher terms represent higher frequencies, which are often noise: by filtering them out the method automatically provides its own smoothing. By contrast, models based on discretising the cycle almost always become less realistic the more parsimonious the model. If, as is often the case, observations are reasonably uniformly distributed across the cycle, all the Fourier terms will be almost orthogonal to one another and to the intercept thus greatly simplifying model building. Furthermore, the Fourier representation is mathematically versatile allowing us under certain circumstances to deconvolve the underlying cyclic pattern when all we can observe is the cumulative effect of exposure to its influence [5].

All models of periodicity by their very nature require more than one parameter to describe them: at least one parameter each is needed in order to fit phase and

Correspondence: tony.fulford.lshtm@gmail.com
[1]MRC Keneba, MRC Unit, P.O. Box 273, Banjul, The Gambia
[2]MRC International Nutrition Group, Department Public Health, London School of Hygiene & Tropical Medicine, London, UK

amplitude. Fourier regression is no exception. Consequently the extent of periodicity is not generally represented by a single model parameter; a summary statistic needs to be derived from the fitted model. This article is concerned with the search for a suitable statistic to summarise and compare the amplitude of cyclic patterns.

Method

A common choice of statistic to measure the magnitude of cyclic patterns is the crude difference between the maximum and minimum values of the mean. While this has its uses it also has its drawbacks: it is not always as straightforward as it may seem to locate and measure the extrema accurately or to estimate the standard error of the difference. It is also focuses on one narrow aspect of the periodic function and ignores information over much of the cycle.

Another obvious choice, at least for continuous dependent variables, would be the partial R-squared. Pewsey *et al.* review a number of exotic correlation coefficients devised for circular data [6]. These, however, essentially restrict their attention to the first pair of Fourier terms and do not take account of covariates. R-squared can be thought of as the variance of the fitted values expressed on the scale of the overall variance of the dependent variable in the sample. While this can be a logical scale to work on, it can also present a number of problems of interpretation. Often a large component of the variance is due to measurement imprecision; a scale based on the arbitrary degree of noise can have little meaning. Furthermore both precision of measurement and the underlying variance within the population will often vary between studies thus invalidating direct comparison of R-squared values. The statistical power of studies of cyclic patterns is often greatly improved by observing each individual at several different points in the cycle. In such multi-level designs there is more than one variance to consider and a difficult choice to be made as to which provides the most relevant scale on which to measure the magnitude of the periodicity.

A better choice of scale is therefore needed. We desire something familiar, universal and stable. Simply expressing the explained variance as its square root (i.e. as the standard deviation) places it on the scale of the original variable. That would be familiar and stable but not universally useful: while it might be useful when comparing the same variable measured in different studies, it is usually useless when comparing different variables even within the same study. Another approach to standardising the scale is to divide the standard deviation by mean. The coefficient of variation, frequently used to assess assay precision, is of this form, although obviously in this case it is error rather than explained variation that

is being standardised. This approach is particularly appropriate when, as is often the case, proportional changes in the variable in question are important. Such variables are usually analysed in the logarithm.

Turning our attention to the variation explained by the cyclic pattern, rather than the scale on which it is measured, it is not always the case that the variance of the fitted values of the data will be appropriate. Indeed, if the data are not uniformly distributed across the cycle (had, for instance, we chosen to sample more densely where the periodic function is thought to change the most rapidly) the variance of the fitted values would not yield an unbiased estimate of the variance of an individual's experience over the cycle. (Seasonal data associated with births may be an important exception: birth frequency is itself often seasonal and estimating seasonality of statistics associated with births, such as birth weight, might need to take that into account. An estimate of magnitude based on the variance of predicted values of the data might then be a simple solution provided the sampling frequency follows the same seasonal pattern as the birth frequency.) Instead it may often be preferable to work with the variance of the fitted function across the cycle:

$$\mathrm{var}_\theta(c(\theta)) = \int_0^{2\pi} f(\theta)c(\theta)^2 d\theta \qquad (2)$$

where $f(\theta)$ is the density function. Such an approach seems reasonable and simple but is not widely used in regression analysis probably because the underlying distribution density of the predictor variables is not generally known. However, with cyclic data the relevant underlying distribution (e.g. when considering an individual's experience over the full cycle) is usually uniform, $f(\theta) = 1/2\pi$, and easy to work with. Under this assumption the variance of a Fourier series over a full cycle turns out to be very simply half the sum of the squares of the coefficients. Thus, for the parameters in model (1):

$$\mathrm{var}_\theta(c(\theta)) = \frac{1}{2}\sum_j^P \left[\alpha_j^2 + \gamma_j^2\right] \qquad (3)$$

When the outcome variable is measured on a common, recognisable scale (e.g. an anthropometric z-score) this statistic serves as an adequate measure of the amplitude of the periodicity. In other cases it would be useful to rescale the variance. This suggests a statistic analogous to the coefficient of variation, which I will call the Coefficient of Cyclic Variation (*ccv*): sd($c(\theta)$)/mean. Provided the mean of x_i in the Fourier regression model, (1), is zero (i.e. the covariates are centred) and the data

are a reasonable representation of the population then the mean will be given by β_0:

$$ccv = \sqrt{\frac{1}{2}\sum_{j}^{P}\left[\alpha_j^2 + \gamma_j^2\right]} \Bigg/ \beta_0 \qquad (4)$$

The *ccv* is thus a simple function of the parameters in the Fourier regression model. Its standard error and confidence intervals may be readily calculated either using the delta method or the bootstrap (Stata's *nlcom* command, for instance, can be used to estimate the statistic and its confidence interval based on the delta method – StataCorp, College Station, TX).

As mentioned above, scaling by the mean is most appropriate when proportional changes in the variable are important. In these cases the data are often analysed in the logarithm, i.e. y_i is replaced by $\log(y_i)$ in equation (1).
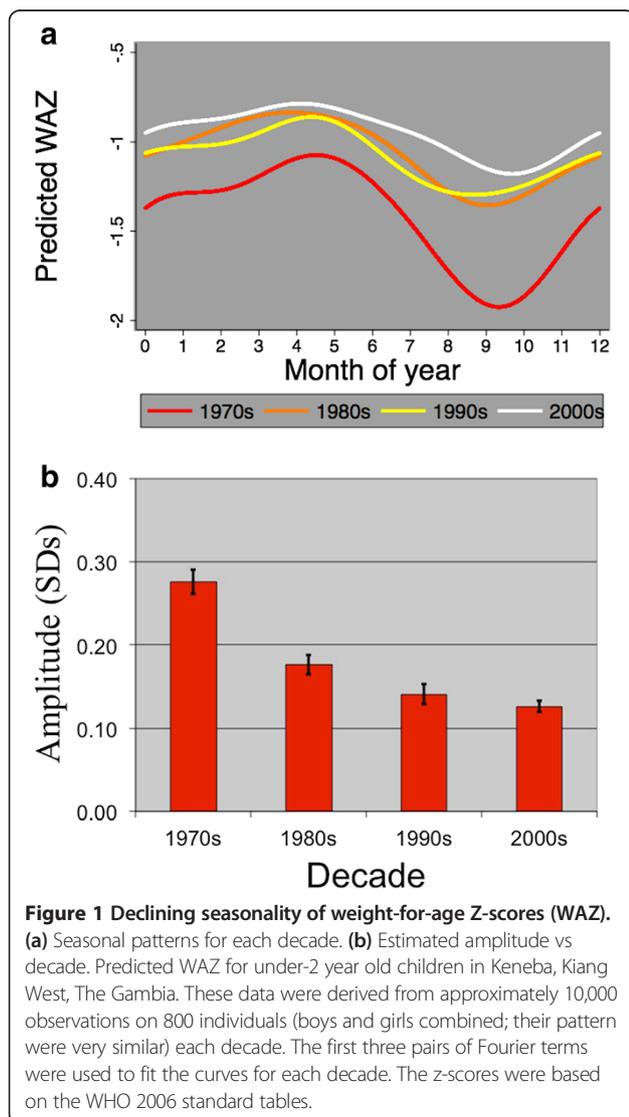

Figure 1 Declining seasonality of weight-for-age Z-scores (WAZ).
(a) Seasonal patterns for each decade. **(b)** Estimated amplitude vs decade. Predicted WAZ for under-2 year old children in Keneba, Kiang West, The Gambia. These data were derived from approximately 10,000 observations on 800 individuals (boys and girls combined; their pattern were very similar) each decade. The first three pairs of Fourier terms were used to fit the curves for each decade. The z-scores were based on the WHO 2006 standard tables.
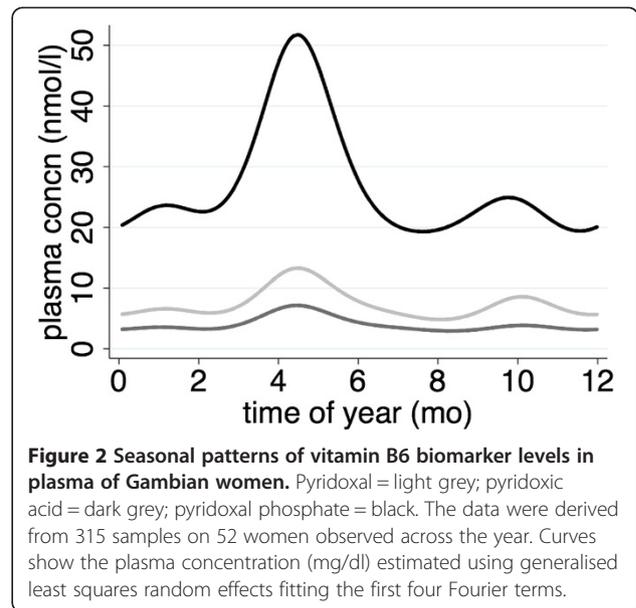

Figure 2 Seasonal patterns of vitamin B6 biomarker levels in plasma of Gambian women. Pyridoxal = light grey; pyridoxic acid = dark grey; pyridoxal phosphate = black. The data were derived from 315 samples on 52 women observed across the year. Curves show the plasma concentration (mg/dl) estimated using generalised least squares random effects fitting the first four Fourier terms.

In that case, provided sd is small relative to the mean, the *ccv* for the original untransformed variable is given approximately by:

$$ccv \approx \sqrt{\frac{1}{2}\sum_{j}^{P}\left[\alpha_j^2 + \gamma_j^2\right]} \qquad (5)$$

Two examples

1. The use of the unscaled variance is illustrated by the example of the change in seasonality of weight-for-age z-scores among Gambian children in recent decades. Figure 1a shows how not only has wasting in these children reduced since the 1970s but the children also appear to be less susceptible to seasonal changes (probably buffered by the "remittance economy"). This is confirmed by the plot of the amplitude of seasonality shown in Figure 1b.

2. This example employing the *ccv* is shown in Figure 2. Here the seasonal patterns of plasma pyridoxal, pyridoxal phosphate and pyridoxic acid derived from 315 observations of 52 Gambian women (data courtesy of Paula Dominguez-Salas). The logarithms of the assay

Table 1 The seasonal variation in vitamin B6 biomarkers among Gambia women

Biomarker	*ccv*	CI 95% (delta method)	CI 95% (bootstrap*)
Pyridoxal	25.3%	(21.9%, 28.6%)	(21.5%, 28.9%)
Pyridoxic acid	27.3%	(23.3%, 31.4%)	(22.6%, 32.0%)
Pyridoxal phosphate	29.0%	(25.0%, 33.0%)	(25.0%, 33.1%)

*Employing 50 re-samplings of the data.

values were fitted by random effects GLS regression using the first four Fourier terms and controlling for age; the fitted values were anti-logged to yield the plotted curves. All three biomarkers appear to follow the same seasonal influences peaking in May but are they affected equally? Although a thorough analysis would make allowance for the fact that these biomarkers were measured in the same women, from the *ccv*s and their 95% confidence intervals tabulated in Table 1 it is clear that the seasonal patterns are of very similar magnitude. Note also the consistency between the two methods used to estimate the confidence intervals.

Conclusion

There are numerous ways to exploit the simple (but apparently little known) formula for the variance of a Fourier series given in equation (3). I suggest that in many, probably the great majority, of cases when seasonal or diurnal patterns are investigated in epidemiological studies, the *ccv* or related statistic will provide a simple and useful measure of the size of periodic variation. Its simplicity has the potential for this approach to provide the universally recognised standard statistics to be reported in such studies.

References
1. Nelson W, Tong YL, Lee JK, Halberg F: **Methods for cosinor-rhythmometry.** *Chronobiologia* 1979, **6**:305–323.
2. Fernández JR, Hermida RC, Mojón A: **Chronobiological analysis techniques. Application to blood pressure.** *Phil Trans R Soc A* 2009, **367**:431–445.
3. Bliss CI: **Periodic regression in biology and climatology.** *Bull Conn Agric Exp Station New Haven* 1958, **615**:1–55.
4. Cox NJ: **Speaking stata: in praise of trigonometric predictors.** *Stata J* 2006, **6**(4):561–579.
5. Fulford AJC, Rayco-Solon P, Prentice AM: **Statistical modelling of the seasonality of preterm delivery and intrauterine growth restriction in rural Gambia.** *Paediatr Perinat Epidemiol* 2006, **20**(3):251–259.
6. Pewsey A, Neuhäuser M, Ruxton GD: **Correlation and Regression.** In *Circular Statistics in R.* Oxford: Oxford University Press; 2013:149–170.