

THE STATA JOURNAL

Editor

H. Joseph Newton
Department of Statistics
Texas A & M University
College Station, Texas 77843
979-845-3142; FAX 979-845-3144
jnewton@stata-journal.com

Associate Editors

Christopher F. Baum
Boston College

Rino Bellocco
Karolinska Institutet, Sweden and
Univ. degli Studi di Milano-Bicocca, Italy

A. Colin Cameron
University of California–Davis

David Clayton
Cambridge Inst. for Medical Research

Mario A. Cleves
Univ. of Arkansas for Medical Sciences

William D. Dupont
Vanderbilt University

Charles Franklin
University of Wisconsin–Madison

Joanne M. Garrett
University of North Carolina

Allan Gregory
Queen's University

James Hardin
University of South Carolina

Ben Jann
ETH Zürich, Switzerland

Stephen Jenkins
University of Essex

Ulrich Kohler
WZB, Berlin

Stata Press Production Manager

Stata Press Copy Editor

Editor

Nicholas J. Cox
Department of Geography
Durham University
South Road
Durham City DH1 3LE UK
n.j.cox@stata-journal.com

Jens Lauritsen
Odense University Hospital

Stanley Lemeshow
Ohio State University

J. Scott Long
Indiana University

Thomas Lumley
University of Washington–Seattle

Roger Newson
Imperial College, London

Marcello Pagano
Harvard School of Public Health

Sophia Rabe-Hesketh
University of California–Berkeley

J. Patrick Royston
MRC Clinical Trials Unit, London

Philip Ryan
University of Adelaide

Mark E. Schaffer
Heriot-Watt University, Edinburgh

Jeroen Weesie
Utrecht University

Nicholas J. G. Winter
University of Virginia

Jeffrey Wooldridge
Michigan State University

Lisa Gilmore
Gabe Waggoner

Copyright Statement: The Stata Journal and the contents of the supporting files (programs, datasets, and help files) are copyright © by StataCorp LP. The contents of the supporting files (programs, datasets, and help files) may be copied or reproduced by any means whatsoever, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the Stata Journal.

The articles appearing in the Stata Journal may be copied or reproduced as printed copies, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the Stata Journal.

Written permission must be obtained from StataCorp if you wish to make electronic copies of the insertions. This precludes placing electronic copies of the Stata Journal, in whole or in part, on publicly accessible web sites, file servers, or other locations where the copy may be accessed by anyone other than the subscriber.

Users of any of the software, ideas, data, or other materials published in the Stata Journal or the supporting files understand that such use is made without warranty of any kind, by either the Stata Journal, the author, or StataCorp. In particular, there is no warranty of fitness of purpose or merchantability, nor for special, incidental, or consequential damages such as loss of profits. The purpose of the Stata Journal is to promote free communication among Stata users.

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press. Stata and Mata are registered trademarks of StataCorp LP.

Speaking Stata: In praise of trigonometric predictors

Nicholas J. Cox
Department of Geography
Durham University
Durham City, UK
n.j.cox@durham.ac.uk

Abstract. Using sine and cosine terms as predictors in modeling periodic time series and other kinds of periodic responses is a long-established technique, but it is often overlooked in many courses or textbooks. Such trigonometric regression is straightforward in Stata through applications of existing commands. I give various examples using classic periodic datasets on the motion of the asteroid Pallas and the daily rhythm of birth numbers. I make a brief connection to polynomial-trigonometric regression.

Keywords: st0116, circular regression, Fourier regression, harmonic regression, periodic regression, polynomial-trigonometric regression, trigonometric regression, sine, cosine

1 Introduction

The last two Speaking Stata columns have illustrated a theme of circular arguments: examining time of day as a circular scale (Cox 2006a) and graphing data to show the structure of seasonality (Cox 2006b). This column completes a trio by focusing on the use of trigonometric predictors, series of sine and cosine terms, in regression-like models. The general topic has also been discussed under the headings of circular, Fourier, harmonic, or periodic regression.

Fourier analysis (Fourier series, transforms, etc.) is one of the largest and most fruitful areas of applied mathematical science as a whole, especially in the physical and engineering sciences. Körner (1988), Bracewell (2000), and Kammler (2000) are just three of many splendid books celebrating the depth and richness of this field. Lanczos (1956, vii) underscored its importance with a moment of melodrama: “If we were asked to abandon all mathematical discoveries save one, we would hardly fail to vote for the Fourier series as the candidate for survival.”

In statistics, on the other hand, the topic to be discussed here sometimes falls between the gaps separating various texts or courses. It often perhaps appears a little too advanced (or too specialized) for elementary treatments and a little too elementary (or too obvious) for advanced ones. The technique may also be too classic (with roots centuries old) to appeal to those permanently in search of what the Australian art critic Robert Hughes called “the shock of the new”.

A more specific categorization problem is also evident. The technique in some ways falls uneasily between time-series and regression treatments, seeming a basic (and limited) application of regression to time-series people and a specialized (and peripheral) application to time series to regression people.

Fitting the first few terms of a Fourier series is a standard warm up on the road to a frequency domain treatment of time series centered on spectral analysis (e.g., Bloomfield 2000). However, spectral analysis appears to be one of those techniques that even time-series people apply either constantly or almost never. Moreover, the mainstream of current time-series analysis flows directly from the idea that time series should be considered realizations of stochastic processes. The use of trigonometric predictors as discussed here has a rather different focus, namely, the exploratory identification of systematic smooth structure in periodic data. It is more an example of nonparametric or semiparametric regression in style, although it is an example of classic parametric regression in substance.

Speculation and even paradox aside, good accounts of trigonometric regression can be found in several statistics texts geared to the needs of climatology and hydrology (e.g., Helsel and Hirsch 1992; Wilks 2006). A further classic reference stuffed with real biological examples is Bliss (1970). For a vignette of Bliss' life and work, see [R] **probit**.

Trigonometric regression raises many points of interest to Stata users. One simple message of this column is that no new Stata commands, official or user written, are needed for what is essentially just another kind of regression-like modeling. You just need to know something of the functionality already provided.

2 Trigonometry revisited

Most readers will be familiar with elementary plane trigonometry, but we will review briskly the fundamental ideas of angle measurement and of periodic functions of angles, particularly sine and cosine. Introductory and even popular treatments abound. If you want a refresher or a reference for your students or colleagues, Gullberg (1997) includes many historical details in addition to the standard formulas and graphs. Maor (1998) is also good on history; his survey culminates in an account of Fourier's work.

2.1 Angles and periodic functions

Sine and cosine for our purposes are periodic functions that return results between -1 and 1 as some angle, say, θ , varies. Here angles are defined as measured in an *agreed direction* as the amount of rotation away from a *fixed axis*. Angles can be arbitrarily large, either as positive angles or as negative angles, as an angle corresponds to more and more rotation, either in the agreed direction or in the *opposite direction*, including as many rotations past the fixed axis as we please. In mathematics, the fixed axis is conventionally taken to be horizontal and the agreed direction is conventionally counter-clockwise, the opposite direction being thus clockwise. That these are conventions is

shown by contrast. The usual practice in geography, and the Earth sciences generally, is to take the fixed axis to point north (by hemispherist cartographic custom plotted as a vertical axis) and the agreed direction to be clockwise. The resulting angle is a map bearing.

As a geographer, I find the latter conventions just as appealing as those more standard in mathematics. They have at least one advantage: they are coupled to well-known terminology for compass directions. Forget that our planet is three-dimensional, and think about a map with an axis pointing north and some place on the map that we choose as a fixed origin. Then the location of any other place can be represented either as a distance from that origin, measured directly, and a map bearing; or as the distance north and the distance east from that origin (figure 1). In this scheme, a distance that is south rather than north is a negative distance north, and a distance that is west rather than east is a negative distance east. Sine and cosine are then defined by

$$\text{sine of bearing} = \text{distance east} / \text{distance from origin}$$

and

$$\text{cosine of bearing} = \text{distance north} / \text{distance from origin}.$$

The limits for both sine and cosine of 1 and -1 also follow from such definitions.

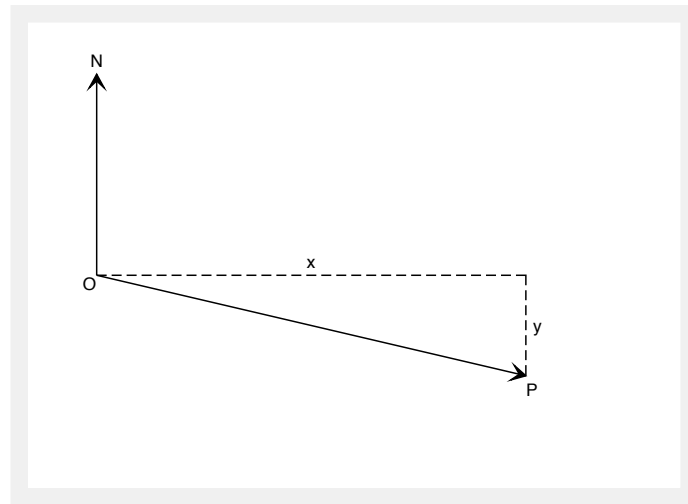


Figure 1: A geographical interpretation of sine and cosine. The position of a place P is given by either its direct distance from the origin OP and the angle or bearing NOP from the axis facing north, or its distance east (x) and distance north (y). The sine of bearing is x/OP and the cosine of bearing is y/OP .

Naturally, these definitions are akin to those often first given in elementary mathematics, as ratios of lengths of sides of right-angled triangles, but they are more general, as the angles concerned can be any fraction of a circle.

Any angle is equivalent to the same angle plus or minus any integer number of rotations of the circle, just as a ballet dancer or skater who spins around, say, once, twice, three times, and so on, is facing the same direction as before after each complete rotation. Periodic functions of angles also have the same results for any angle on the circle and that angle plus or minus so many complete rotations.

Stata follows usual mathematical practice in expecting angles to be supplied in radians. A radian is the angle subtended at the center of a circle by an arc of the same length as the radius of that circle. Because the circumference of a circle is the radius multiplied by 2π , a complete circular rotation is an angle of 2π radians. Scientists are more likely to be familiar with angle measurement in degrees ($^\circ$). (Is there any field that reports *data* in radians?) The relation is that 2π radians equals 360° ; thus, 1 radian is $180^\circ/\pi$ or about 57.3° . In Stata, π is wired in to as much precision as is possible in its calculations. The stored constant has names of both `_pi` and `c(pi)`. Thus you need never (nor should you ever) type in whatever number of digits you can remember from the decimal representation of π (3.14159...).

When we think about periodic time series, we suppress this mental framework of circles and angles measured on circles but focus instead on the idea that values are periodic with respect to some interval, here time. Often this idea is coupled with the ideas that there may be some long-term trend, which is not periodic, and that there may be added irregular or random noise, so that no smooth function will fit exactly. (A key idea from Fourier analysis is that we can get an exact fit if we use enough trigonometric predictors, but statistically this is rarely a good idea in modeling, and it is useless for smoothing.)

Consider two common kinds of examples of periodic variation, over time scales of a day or of a year, respectively. In the first case, responses we are tracking are likely to be recorded at times measured in hours and possibly more finely. I discussed the complications that may arise here and how to handle them in a previous column (Cox 2006a). In the second case, how time is measured may vary considerably: daily, weekly, monthly, and quarterly measurements are all common for different phenomena. And although measurement at or for regularly spaced times is common, the method of trigonometric regression is also applicable when times are irregularly spaced.

The most practical method for treating data like these is simply to express the interval of periodicity (e.g., a day, a year) as a unit and to convert all times to that scale. So, hours of a day, months of the year, or whatever are to be converted to fractions of a day or a year. If data are for intervals, say, the 24 hours of the day or the 12 months of the year, then I tend to use conversions such as $(\text{hour} - 0.5)/24$ or $(\text{month} - 0.5)/12$, so that data are related to the (approximate) center of each interval, although many researchers would be happy to work with $\text{hour}/24$ or $\text{month}/12$. However, these conversions treat months as of equal length, which may or may not appear an adequate approximation.

With a unit time interval, the conversion to the radian scale is then simply to multiply by 2π . If data arrive measured in degrees, as is common in some fields, the multiplier is $2\pi/360$ or $\pi/180$.

2.2 Sine and cosine plotted in Stata

Conveniently, `twoway function` in Stata by default plots functions with respect to a horizontal unit interval from 0 to 1. Thus for a reminder of the shape of sine and cosine functions, in Stata `sin()` and `cos()`, you can use `twoway function`, regardless of what else may be in memory at the time. (For more examples of the use of `twoway function`, see Cox [2004b].)

Figure 2 shows the basic sine and cosine functions. Among other small details, note how naming the function in the `twoway function` command is echoed automatically in the legend.

```
. twoway function sine = sin(2 * _pi * x) ||
> function cosine = cos(2 * _pi * x),
> lpattern(dash) xlabel(0(0.25)1) ylabel(, angle(h))
> legend(symxsize(*0.6) ring(0) position(7) column(1))
> xtitle(fraction of circle)
```

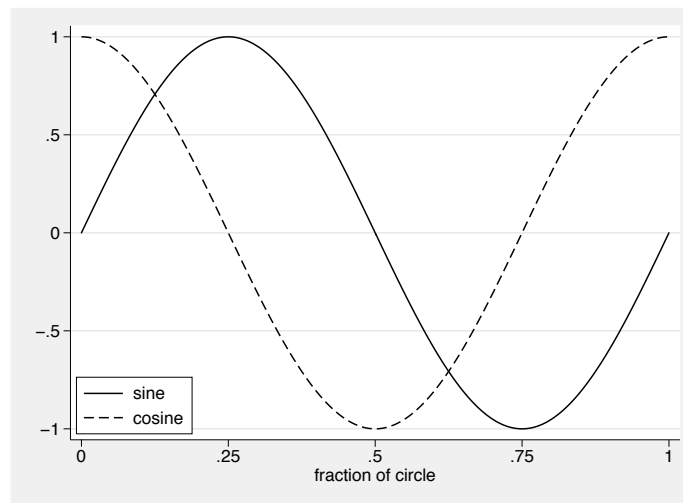


Figure 2: Sine and cosine functions on the circle or unit periodic interval

If we double, triple, or quadruple the argument of sine or cosine, each function repeats itself two, three, or four times over the unit interval, and similarly for any other positive integer. Figure 3 gives an example.

```
. twoway function sin2 = sin(4 * _pi * x) ||
> function cos2 = cos(4 * _pi * x),
> lpattern(dash) xlabel(0(0.25)1) ylabel(, angle(h))
> legend(symxsize(*0.6) ring(0) position(7) column(1))
> xtitle(fraction of circle)
```

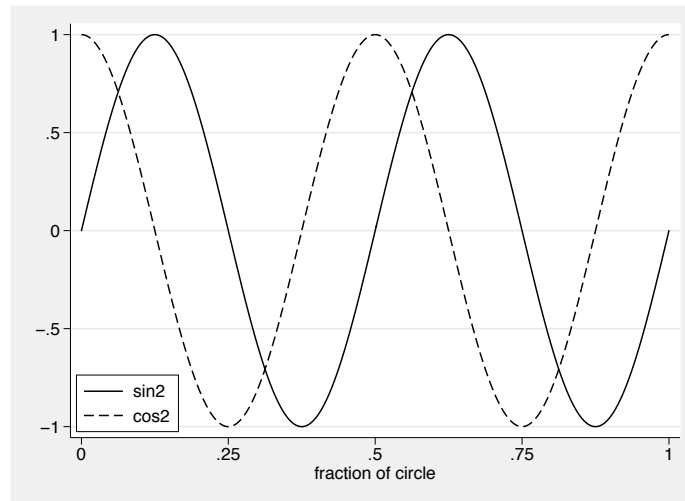


Figure 3: Doubling the angle argument doubles the number of complete cycles of sine and cosine.

If we double the angles, what is plotted are results for angles from 0 to 4π radians, so two complete rotations of the circle are represented.

3 Trigonometric regression

3.1 The basic recipe

The recipe of trigonometric regression, in its simplest application, is based on a combination of ideas. For concreteness, when regression is mentioned, you may like to think about the Stata command `regress`, but the ideas are much more general. The distinctiveness of trigonometric regression is purely a matter of a distinctive set of predictors or covariates and has nothing to do with how the response or dependent variable is treated.

1. The interval over which variations are periodic is scaled to unit length and thus related to radian measure by multiplying it by 2π , producing a horizontal scale. Because the scale often measures time, let us call it t .
2. The regression typically includes as predictors (basis functions, if you like)

1, giving a constant term, as is usual

and some J pairs

$$\sin(2j\pi t), \cos(2j\pi t), \quad j = 1, \dots, J$$

as appropriate for the data and the problem.

3. A linear mixture of a few sine and cosine terms, with coefficients estimated from the data, can often do a good job of fitting some fairly smooth periodic structure in the data. The combination of sine and cosine terms and of different frequencies of repetition within the interval gives flexibility here to match the occurrence of peaks and troughs and other major features. To be fair: minor but systematic irregularities, such as boosted December sales or employment in many economic time series, are more difficult to handle. Such problems require various tricks (e.g., additional dummy variables) or recourse to a different method.
4. There is no problem in adding other predictors if this seems sensible, such as a trend term, as mentioned earlier.

That recipe is really the main idea in this column, and what follows are examples and details. But various general comments are worth making now.

3.2 Sine and cosine terms are taken in pairs

First, sine and cosine terms are taken in pairs. Jeffreys (1961, 343) gives a characteristically concise and direct explanation: “There are many cases where two parameters enter into a law in such a way that it would be practically meaningless to consider one without the other. The typical case is that of a periodicity. If it is present it implies the need for a sine and a cosine. If one is needed the other will be accepted automatically as giving only a determination of phase.” Helsel and Hirsch (1992, 342) make the same point.

The mention of phase reminds us that the parameterization of predictors (point 2 above) is at least in part a convenience that keeps the model linear in the parameters. But each pair of terms, one sine and one cosine, can be represented as a function of two angles, say, θ and ϕ , for example as

$$\sin(\theta + \phi) = \sin \theta \cos \phi + \cos \theta \sin \phi$$

Thus in various ways, mathematically and substantively, pairs of sine and cosine terms can be considered yoked together.

There is a complementary view. As mentioned, in geography and the Earth sciences, the fixed axis points north and map bearings are measured clockwise from north. Following our earlier geographical definitions, cosine coefficients based on such bearings measure effects operating north–south (meridional, in one jargon) and sine coefficients measure effects operating east–west (zonal, in the same jargon). Evans and Cox (2005) give one Stata-based exploration of the effects of direction on glacier altitudes worldwide. The story there is related to contrasts in radiation, shade, and so forth, so that cosine and sine have different physical interpretations.

3.3 Predictions satisfy boundary conditions

The predictions from a regression on sine and cosine terms are automatically identical at the beginning and the end of the interval. This is exactly as it should be, because we are dealing with circular scales in which midnight or each New Year, say, is the end of one period and the beginning of the next, so discontinuities in predicted response would be absurd whenever we are dealing with a periodicity.

The basis for this statement will turn out to be useful as well. Consider a prediction, using notation that fits the purpose, that is

$$b_0 + \sum_{j=1}^J s_j \sin(2j\pi t) + \sum_{j=1}^J c_j \cos(2j\pi t)$$

Here b_0 is the constant or intercept and the s_j and the c_j are the other coefficients to be estimated from the data. For any integer j , $\sin(2j\pi t)$ at $t = 0$ or 1 is always 0 , and so at those boundaries $\sum_{j=1}^J s_j \sin(2j\pi t)$ is 0 , whatever the s_j may be. Similarly, $\cos(2j\pi t)$ at $t = 0$ or 1 is always 1 , and so at those boundaries $\sum_{j=1}^J c_j \cos(2j\pi t)$ is identically $\sum_{j=1}^J c_j$. Thus the predicted value at the boundaries is the sum of the fitted constant b_0 and the estimated coefficients of the cosine terms $\sum_{j=1}^J c_j$ and therefore a constant.

The stipulation of integers j is vital here. If j were not an integer, sine and cosine would be partway through a cycle at $t = 1$ and their values there would differ from those at $t = 0$. Similarly, this property does not hold if there are other predictors in the model. In that circumstance it applies only to the sum of the trigonometric terms.

More positively, the use of a link function (in generalized linear model terminology) in the model does not affect the issue. For example, if exponentiation maps linear predictions back onto the scale of the response variable, then a constant is mapped to another constant, and the boundary conditions remain satisfied.

3.4 Orthogonality of predictors

An attractive property of the sine and cosine terms is their orthogonality or lack of correlation. This property can be shown in a slightly unusual way by a scatterplot matrix (figure 4). Along the way, we can see a simple method of producing a set of trigonometric predictors with a `forvalues` loop. The temptation to write a command to encapsulate this loop should be declined, as being able to do it easily from first principles scores over the disadvantage of having to remember or rediscover the precise syntax of an extra command.

```

. set obs 101
. range t 0 1
. forval j = 1/3 {
.     gen sin'j' = sin('j' * 2 * _pi * t)
.     gen cos'j' = cos('j' * 2 * _pi * t)
. }
. graph matrix sin1-cos3, ysize(4) xsize(4)

```

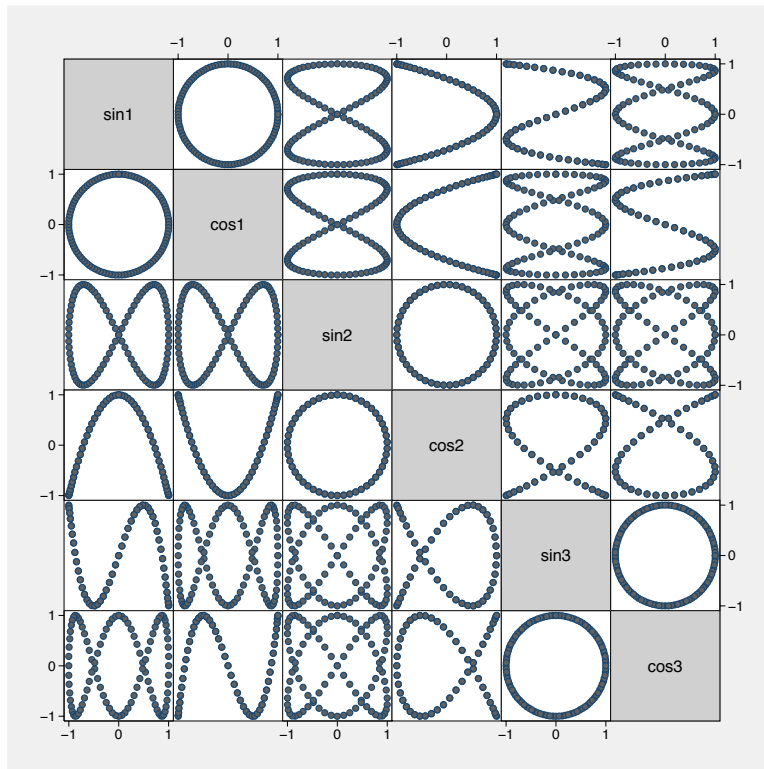


Figure 4: Scatterplot matrix of the first three pairs of sine and cosine terms

If you would like more explanation of the `forvalues` loop, here it is. See also [P] `forvalues` and Cox (2002) as desired. The loop is over the integers 1/3, i.e., 1, 2, and 3. The counter (strictly, a local macro) `j` is set to 1 the first time around the loop. Its value is substituted in the first `generate` statement, which becomes

```
gen sin1 = sin(1 * 2 * _pi * t)
```

and then similarly in the second statement. Notice how the value of 1 will be used by Stata both as text (as part of the new name `sin1`) and as a number within the argument of `sin()` or `cos()`. On the second and third times around the loop, `j` is set in turn to 2 and then 3. Thus the loop is a way of repeating two commands but substituting in turn 1, 2, and 3 as desired.

The property of orthogonality stands in stark contrast to the typically high correlations between various powers of a standard power series polynomial, namely, t , t^2 , t^3 , and so forth. Turn and turn about, that is, as is well known, the motivation for using alternative families of polynomial in modeling or smoothing.

In practice, data that are irregularly distributed over the unit interval may show some nonzero correlations. Such correlations may be the indirect signal of data ill-suited to determine the coefficients of a trigonometric regression, especially if they are concentrated on only part of the interval.

3.5 Not only smooth but also differentiable

Apart from signs and constants, the derivative of any sine is always a cosine, and vice versa. It follows that a fitted function that is a sum of sine and cosine terms can be differentiated as many times as you please. This differential can be useful to those interested partly or even primarily in rates of change. Not only will they get a smooth fit from the first few Fourier terms, but the fit is defined by a closed-form expression and so are all its derivatives. (As a matter of convenience, you might prefer numeric differentiation, but that is your choice.) Naturally, if a relatively smooth function also includes kinks, jumps, or small bumps or ruts, then a trigonometric fit might not be best for getting at rates of change, but again that is your choice.

4 The orbit of Pallas

For a first example with data, let us revisit a dataset used by Gauss on the right ascension and declination, i.e., the longitude and latitude of the corresponding point on the celestial sphere, of the asteroid Pallas. This was cutting-edge astronomy 200 years ago, especially given the relative faintness of the object and increasing awareness of new planets and other objects at that time. You can find more on Gauss, including his astronomical work, in standard biographies such as that by Dunnington (1955). The data are given by Kammler (2000, 70). The declinations (in minutes) at right ascensions $0(30)330^\circ$ are 408, 89, -66 , 10, 338, 807, 1238, 1511, 1583, 1462, 1183, and 804. Our variables are thus right ascension `asc` and declination `dec`. We create a unit interval t by division by 360 and then look at the data (figure 5).

```
. generate t = asc / 360
. scatter dec asc, xlabel(0(30)360) ylabel(, angle(h)) msymbol(oh)
```

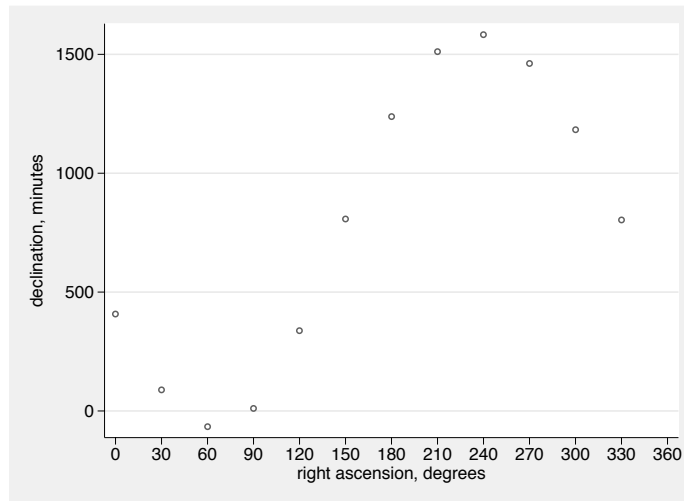


Figure 5: Declination of Pallas as a function of right ascension

The smoothness of the orbit is guaranteed by classical dynamics, modulo collisions. The problem is as much one of interpolation as one of smoothing, but it is still worth a decent job. We can use a `forvalues` loop like that given earlier to generate a bundle of sine and cosine terms. We will adopt the naming convention `sin1`, `cos1`, `sin2`, `cos2`, etc. The first fit is just

```
. regress dec sin1 cos1
```

Source	SS	df	MS	Number of obs =	12
Model	4125966.15	2	2062983.08	F(2, 9) =	1594.16
Residual	11646.7643	9	1294.08492	Prob > F =	0.0000
Total	4137612.92	11	376146.629	R-squared =	0.9972
				Adj R-squared =	0.9966
				Root MSE =	35.973

declination	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
sin1	-720.2279	14.68608	-49.04	0.000	-753.4501 -687.0057
cos1	-411.0144	14.68608	-27.99	0.000	-444.2366 -377.7922
_cons	780.5833	10.38462	75.17	0.000	757.0917 804.075

R^2 at 99.72% is the object of fantasy for many, and there might seem little scope for improvement, but seeing whether there is any systematic structure lurking in the residuals is always a good idea. One useful command for plotting results discussed in an earlier column (Cox 2004a) is `regplot`. `regplot` is a postestimation command that you can issue after `regress` (and many similar commands). By default, it plots the response and predicted values for the response on the vertical axis against the *first* predictor named (which in the simplest case is the only one) on the horizontal axis. It can also be used with a named alternative variable on the horizontal axis, which is useful here, given that right ascension is the forcing variable, but not included in the model so far as Stata knows. The plot is given in figure 6.

```
. regplot asc, xlabel(0(30)360) ylabel(, angle(h))
```

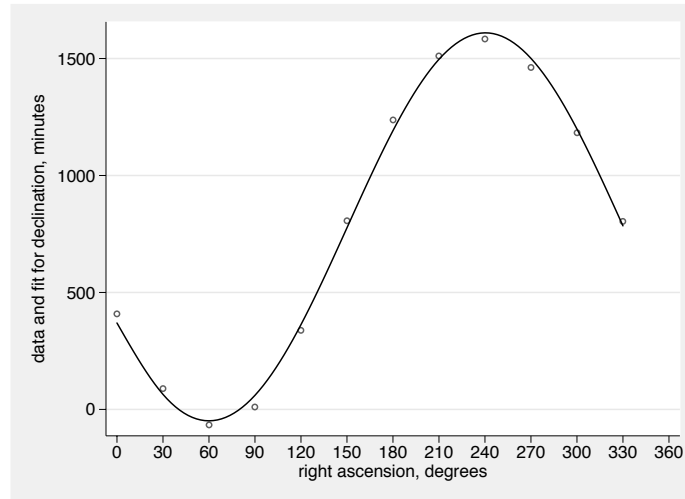


Figure 6: Declination of Pallas as a function of one sine and one cosine term

A plot of residuals versus right ascension is given in figure 7. Official Stata's `rvpplot` (see [R] **regress postestimation**) will not play here, as `asc` was not in the last model fitted, but the alternative `rvpplot2` (Cox 2004a) will oblige, so long as you show awareness that `asc` was not in the model by specifying the `force` option.

```
. rvpplot2 asc, force xlabel(0(30)360) ylabel(, angle(h))
```

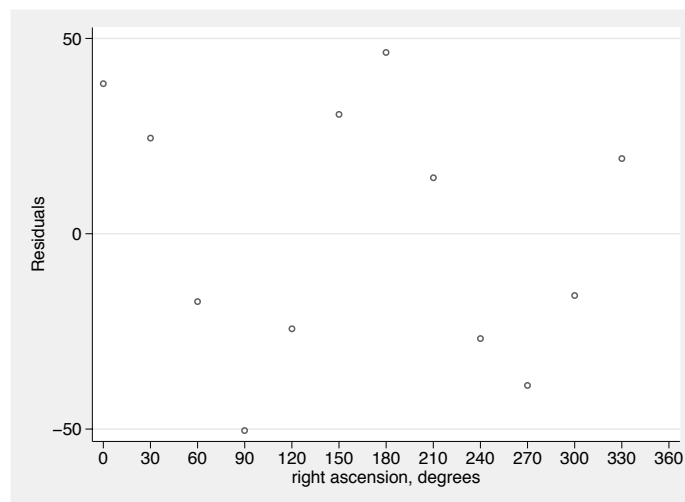


Figure 7: Residuals from first regression plotted against right ascension

The residual plot shows a periodicity that is missed by the model. Adding `sin2` and `cos2` boosts R^2 to 99.99% and more importantly slices root mean squared error from 35.97' to 6.64' (the prime, ', indicates the units of minutes). The `test` (`[R] test`) command could be used to assess the significance of each pair of terms. By the third and fourth pair of terms, R^2 is 1 to four decimal places. It is better to keep an eye on root mean squared error, which is reduced further to 1.70' and then 0.65'. Although the t statistics and p -values remain delightful up to the four-pair fit, the risk of overfitting is also evident. Somewhat arbitrarily, I close with the fit for just two pairs of sine and cosine terms, that is, using `sin1`, `cos1`, `sin2`, and `cos2` as predictors (figure 8).

```
. regress dec sin1 cos2 sin2 cos2
(output omitted)
. regplot asc, xla(0(30)360) yla(, ang(h))
```

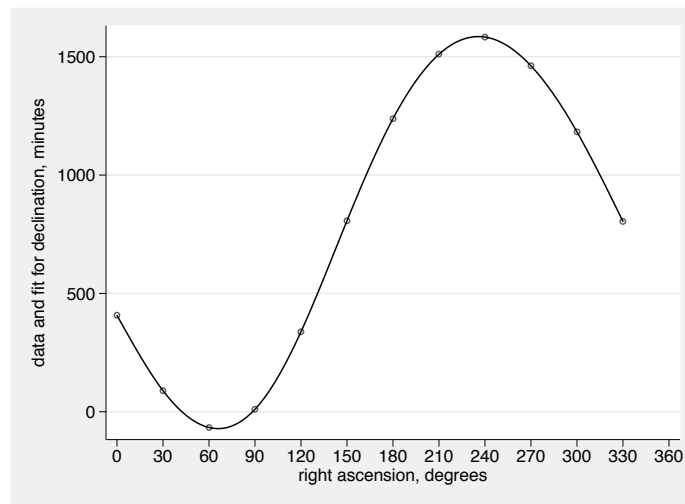


Figure 8: Declination of Pallas as a function of two sine and two cosine terms

Picking up a point made earlier: the prediction at the boundaries can be calculated from

```
. display _b[_cons] + _b[cos1] + _b[cos2]
```

5 A time to be born

Bliss (1970) gives many intriguing examples and we will focus on various data series relating number of births to time of day (p. 279). His source was Kaiser and Halberg (1962). Data are reported from four studies, each as square roots of the original numbers. I reversed this by `round(varname^2)`. There is occasional ambiguity as to which integer is correct, which we trust is trivial. For comparison, the four data series are shown relative to their own means (figure 9). Note the logarithmic scale.

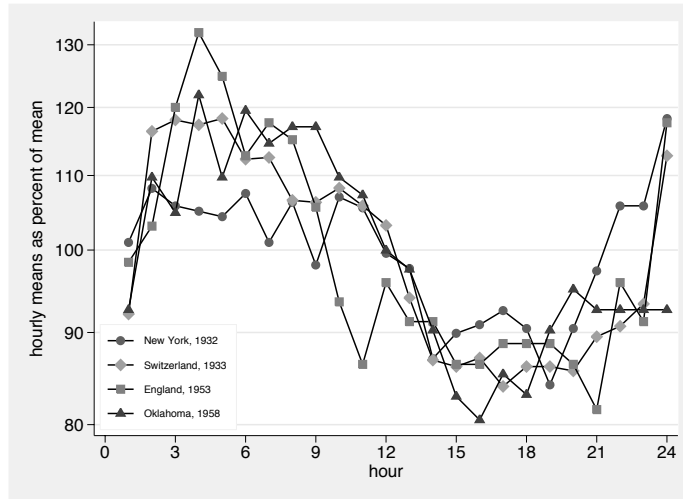


Figure 9: Daily rhythms in number of births from four studies cited by Bliss

Despite some irregularities, the family resemblance is clear. Indeed, these rhythms are well known and serve here largely as a second example of technique. Given a bundle of sine and cosine variables calculated from $(\text{hour} - 0.5)/24$, a natural technique for such count data appears to be Poisson regression. I prefer to use `glm` ([R] `glm`) for this purpose, rather than `poisson`, because it leaves more in its wake—in particular, various kinds of residuals. We select the series from a study from Switzerland in 1933 for more detailed analysis.

```
. glm Switz sin1 cos1, family(poisson) link(log)
Iteration 0:  log likelihood = -539.00041
Iteration 1:  log likelihood = -538.37668
Iteration 2:  log likelihood = -538.37668

Generalized linear models
Optimization      : ML
Deviance          = 802.714701
Pearson           = 799.5965547
Variance function: V(u) = u
Link function     : g(u) = ln(u)

No. of obs       = 24
Residual df      = 21
Scale parameter  = 1
(1/df) Deviance = 38.22451
(1/df) Pearson  = 38.07603
[Poisson]
[Log]
AIC              = 45.11472
BIC              = 735.9756
Log likelihood   = -538.3766791
```

Switz	OIM			P> z	[95% Conf. Interval]	
	Coef.	Std. Err.	z			
sin1	.1547047	.0024011	64.43	0.000	.1499986	.1594108
cos1	.0401815	.0023944	16.78	0.000	.0354885	.0448745
_cons	9.581521	.0017009	5633.27	0.000	9.578187	9.584854

A first stab is encouraging, both ways, in structure captured and structure not captured by the model. Figures 10, 11, and 12 are produced by `regplot` and `rvpplot2` as before. Figures 10 and 11 make it clear that structure has been missed by the first pair of terms. A problem is now evident in the form of aberrant first and last data points. Bliss (1970, 278) identifies “a desire to assign births occurring shortly after midnight to a day earlier”. Why anyone would want other people to be even a day older than they really are is a mystery to me in middle age, but the effect appears genuine.

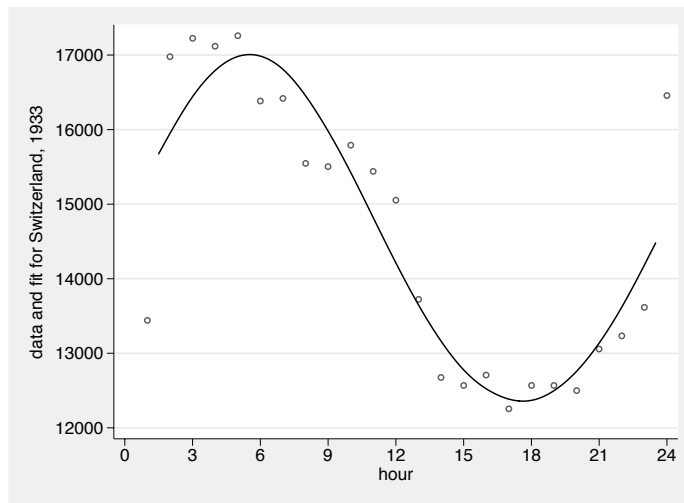


Figure 10: Births in a Swiss study as a function of one sine and one cosine term

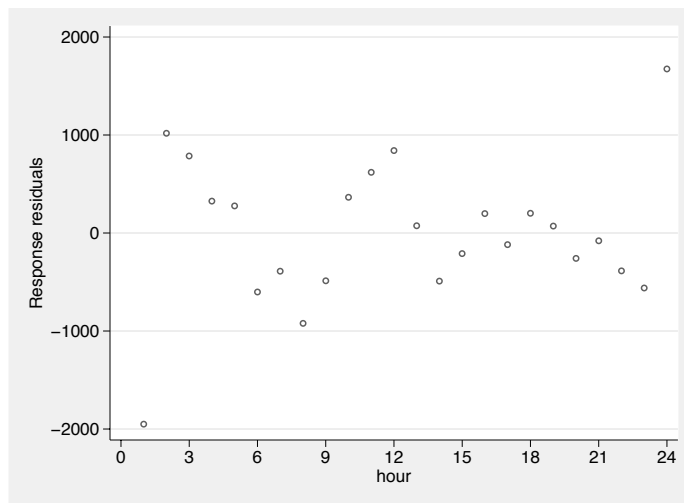


Figure 11: Structure is apparent in the residuals. Note in particular the aberrant first and last data points.

Figure 12 shows that two pairs of terms do a fair amount better but that there is still room for improvement.

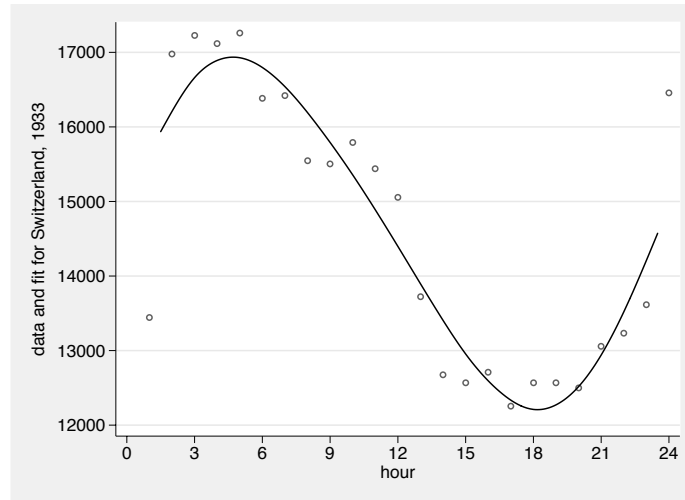


Figure 12: Births in a Swiss study as a function of two sine and two cosine terms

The last model, or as will be seen the last pair of models, to be shown here is based on three pairs of sine and cosine terms. We do this twice: once with all the data and once with first and last data points omitted.

Figure 13 shows the data and fit for all hours. Figure 14 shows the data and fit with hours 2–23 inclusive but with the data for hours 1 and 24 superimposed for comparison. To get the fit at the boundaries in the latter case, add the appropriate coefficients and then exponentiate to reverse the effects of working on the link scale.

```
. display exp(_b[_cons] + _b[cos1] + _b[cos2] + _b[cos3])
```

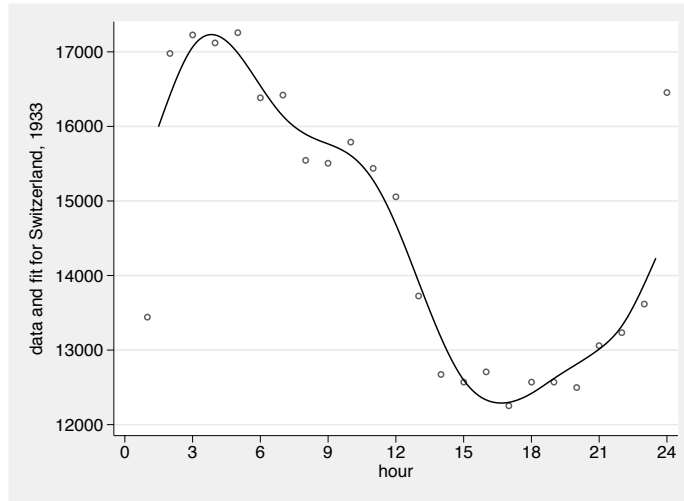


Figure 13: Births in a Swiss study as a function of three sine and three cosine terms: all hours

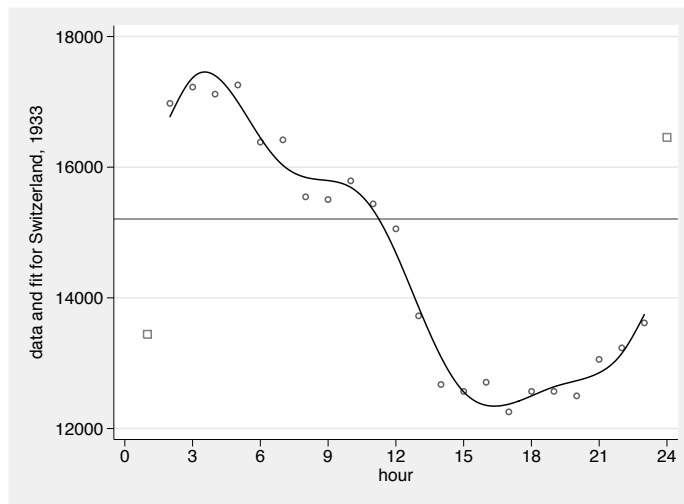


Figure 14: Births in a Swiss study as a function of three sine and three cosine terms: hours 2–23. The added horizontal line shows the fitted hourly rate at midnight.

It is too late to ask the parents or the medical staff about the extent of misassignment to the previous day, but a data analyst can thus make various stabs at estimating how birth numbers behave around midnight. It seems especially appropriate that a method with good model behavior at its boundaries should be used to cast light on data with bad behavior at the same boundaries.

6 Polynomial-trigonometric regression

I make one brief, final connection to polynomial-trigonometric regression, particularly as discussed by Eubank and Speckman (1990). They explain and exemplify the use of a hybrid method combining power series polynomial (usually only linear and quadratic) and trigonometric terms (as many as needed). As in much of this column, their general emphasis is on using parametric regression in a nonparametric or semiparametric style. It should be evident that adding linear and quadratic terms to a mix of sine and cosine terms would be straightforward in Stata. What might be called the “hybrid vigor” of their technique may remind many readers of the fractional polynomials implemented in Stata (see [R] `fracpoly`).

7 Conclusion

Periodicity is common in many time-series and circular datasets. Such structure should be matched by models allowing similar structure. However, trigonometric (circular, Fourier, harmonic, or periodic) regression is not at all exotic but really just another kind of regression modeling with distinctive predictors. Hence, for Stata application, you need only work out what is needed in terms of previously supplied commands.

8 Acknowledgment

Ian S. Evans and I have had many enjoyable interactions applying these ideas, particularly to the effects of aspect (compass direction). He was one of the last students of Chester I. Bliss.

9 References

- Bliss, C. I. 1970. *Statistics in Biology: Volume II*. New York: McGraw–Hill.
- Bloomfield, P. 2000. *Fourier Analysis of Time Series: An Introduction*. New York: Wiley.
- Bracewell, R. N. 2000. *The Fourier Transform and Its Applications*. New York: McGraw–Hill.
- Cox, N. J. 2002. Speaking Stata: How to face lists with fortitude. *Stata Journal* 2: 202–222.
- . 2004a. Speaking Stata: Graphing model diagnostics. *Stata Journal* 4: 449–475.
- . 2004b. Stata tip 15: Function graphs on the fly. *Stata Journal* 4: 488–489.
- . 2006a. Speaking Stata: Time of day. *Stata Journal* 6: 124–137.
- . 2006b. Speaking Stata: Graphs for all seasons. *Stata Journal* 6: 397–419.

- Dunnington, G. W. 1955. *Gauss: Titan of Science*. New York: Hafner.
- Eubank, R. L., and P. Speckman. 1990. Curve fitting by polynomial-trigonometric regression. *Biometrika* 77: 1–9.
- Evans, I. S., and N. J. Cox. 2005. Global variations of local asymmetry in glacier altitude: Separation of north-south and east-west components. *Journal of Glaciology* 51: 469–482.
- Gullberg, J. 1997. *Mathematics: From the Birth of Numbers*. New York: W. W. Norton.
- Helsel, D. R., and R. M. Hirsch. 1992. *Statistical Methods in Water Resources*. Amsterdam: Elsevier. <http://pubs.usgs.gov/twri/twri4a3/>.
- Jeffreys, H. 1961. *Theory of Probability*. Oxford: Oxford University Press.
- Kaiser, I. H., and F. Halberg. 1962. Circadian periodic aspects of birth. *Annals of the New York Academy of Sciences* 98: 1056–1068.
- Kammler, D. W. 2000. *A First Course in Fourier Analysis*. Upper Saddle River, NJ: Prentice Hall.
- Körner, T. W. 1988. *Fourier Analysis*. Cambridge: Cambridge University Press.
- Lanczos, C. 1956. *Applied Analysis*. Englewood Cliffs, NJ: Prentice Hall.
- Maor, E. 1998. *Trigonometric Delights*. Princeton, NJ: Princeton University Press.
- Wilks, D. S. 2006. *Statistical Methods in the Atmospheric Sciences*. Burlington, MA: Academic Press.

About the author

Nicholas Cox is a statistically minded geographer at Durham University. He contributes talks, postings, FAQs, and programs to the Stata user community. He has also coauthored 15 commands in official Stata. He was an author of several inserts in the *Stata Technical Bulletin* and is an editor of the *Stata Journal*.