DISCOVERING THE NATURE OF PERIODIC DATA: II. ANALYZING DATA AND JUDGING THE GOODNESS OF FIT USING RESIDUALS

Scott A. Sinex and George S. Perkins
Prince George's Community College

Generating periodic data using a motion detector and the general behavior of the sine function were discussed in Part I of this article, which appeared in the Spring 2001 issue. Now we want to analyze some real temperature data using manual and regression fits and judge the goodness of fit. Keystrokes for the TI-83 Plus graphing calculator are shown in bold with [ ], such as **[ON]**.

Analyzing Periodic Data

Let's examine the 1998-2000 surface water temperature record for Jack Bay on the lower Patuxent River. These data are from the Maryland Department of Natural Resources.

Date	Time in months	Temp. ^{o}C	Date	Time in months	Temp. ^{o}C	Date	Time in months	Temp. ^{o}C
1/98	1	5.90	1/99	13	5.30	1/00	25	1.00
2/98	2	6.10	2/99	14	6.90	2/00	26	1.40
3/98	3	6.95	3/99	15	6.25	3/00	27	8.85
4/98	4	15.00	4/99	16	13.45	4/00	28	14.00
5/98	5	19.90	5/99	17	19.20	5/00	29	19.15
6/98	6	23.70	6/99	18	23.70	6/00	30	24.70
7/98	7	27.45	7/99	19	26.25	7/00	31	27.20
8/98	8	28.05	8/99	20	28.00	8/00	32	25.90
9/98	9	25.95	9/99	21	23.60	9/00	33	22.80
10/98	10	19.10	10/99	22	17.00	10/00	34	17.15
11/98	11	11.70	11/99	23	12.90	11/00	35	11.80
12/98	12	12.90	12/99	24	8.30	12/00	36	6.00

From: http://www.dnr.state.md.us/bay/conditions/jack_wt.html

To enter the time in months in the list L_1 in **[STAT]** EDIT, do the following: move the cursor onto the L_1 header and press **[2nd] [LIST]** OPS and select 5:seq(and press **[ENTER]**. On the header complete the line: seq(X,X,1,36) and then press **[ENTER]**. Use the **[X,T,q,n]** key to get the X. Type the temperature data into L_2.

Plot a graph of temperature against time in months. Select **[2nd] [STAT PLOT]**, then select Plot 1, and set up the plot as a scatter plot (data points only). For the mathematicians, use **[WINDOW]** to manually set the axes, or for the scientists, press **[ZOOM] [9]** to have the calculator do it. What is the independent variable?

What type of relationship is found? Is there anything unusual about January and February of the year 2000?

Sketch and describe the graph.

What is the range in temperature? This is the difference between the maximum and minimum values.

What is the time interval between maxima (or minima)?

Is there a mathematical function that can model the cyclic nature of the data above?

Manually Fitting a Sine Function

Using graphical analysis, we can estimate the four parameters (a, b, c, and d) in the general equation for a sine function. This will allow us to plot a sine curve through the data points.

1. Finding the midline (y = d) or y-axis shift

This is the average of the temperature or y-data in this case. Press **[STAT]** CALC and select 1-Var Stats and press **[ENTER]**. Back on the home screen type **[2nd] [L_2]** and press **[ENTER]**. The first quantity is the average of the temperature, `0`. Place this value into the function editor, **[Y=]** as $Y_1 = d =$ `0`.

d = _____

2. Estimate the period of the data

This is the time between adjacent maxima or adjacent minima. Make as many determinations as you can and average them. Use **[TRACE]** on the data points to do this. Convert the period, P, using the formula to angular frequency, b.

P = _____ b = $2\pi/P$ = _____

3. Estimate the amplitude of the temperature

This is the difference between the maximum and minimum divided by two or the average distance from the midline to maxima or minima. Use **[TRACE]** on the data points to do this.

a = _____

4. Estimate the x-axis shift

This is the time along the midline from the y-axis to the point where the curve would cross the midline. Use **[TRACE]** along the midline, in Y_1, to do this.

c = _____

Now place the values of a, b, c, and d into the general equation in the function editor, **[Y=]**, in Y_2. Then press **[GRAPH]** to see your function. How well did it fit the data? Do you need to make any adjustments to the values of a, b, c, and d? Record your final best-fit equation.

Fitting a Sine Function Using the Sine Regression
The sine regression, under **[STAT]** CALC as "SinReg" is at the bottom of the menu. Select SinReg and press **[ENTER]** and then type **[2nd] [L_1] [,] [2nd] [L_2]** and press **[ENTER]**. The calculator will return a sine regression after a few moments. Record the values listed below.

a = _____ b = _____ c = _____ d = _____

Now we want to plot this regression equation with the data points. To paste the regression equation into the function editor: Press **[Y=]** and select Y_3 , then press **[VARS] [5] [<] [<]** EQ, select RegEq **[ENTER]**. The regression equation should be in the function editor, **[Y=]**. Press **[GRAPH]** and the regression line will be plotted with the data points.

How well does the regression fit the data points?

How well does the regression equation agree with your manual fit equation?

Judging the Goodness of Fit for a Sine Function
The **residual** is the difference between the actual y-value and the y-value obtained from the regression equation when the appropriate x-value is placed in the equation. It is the error between the measured y-value and that predicted by the model or regression equation.

$$\text{residual} = Y_{\text{actual data value}} - Y_{\text{regression equation}}$$

Residuals, or errors as they are called by some authors, can be used with a mathematical function to judge fit. However, you would have to calculate the residuals in the list editor by taking the difference between the actual data value and the value computed from the manually fit function.

A simple graphical way to judge best fit is to plot the residuals. The TI-83/83 Plus automatically calculates and stores the residuals after a regression analysis is performed. The residuals, RESID, are stored under **[2nd] [LIST]** NAMES. A plot of the residual vs. the x-variable is done via **[2nd] [STAT PLOT]**, where x-variable is the x list and RESID is the y list. Set this plot up in Plot 2 and press **[ZOOM] [9]**. Turn Plot 1 off and unhighlight **[Y=]** equations (turn off).

Best fit is judged when the residuals are at a minimum, and the plot is random as a patternless horizontal band (random noise). A pattern in the residuals plot indicates that the regression is not the best fit of the data or there is a better fitting regression. Outliers will stand out because of their large distance away from the horizontal axis on a residuals plot. Residuals can be calculated for any function, and are a good way to distinguish curvilinear from linear relationships.

Since the sum of the residuals is always zero, another way to judge the goodness of fit is to minimize the sum of the squared residuals for best fit (SUM is found under **[2nd] [LIST]** MATH). This is sometimes referred to as the sum of the squared errors or SSE. For the regression complete the line: SUM ($LRESID^2$) and then press **[ENTER]**.

$$\text{SUM}(LRESID^2) = \_\_\_\_\_$$

Residuals provide an easy and understandable way to judge goodness of fit and help provide a simple explanation to describe regression analysis.

For the manual fit curve, we can calculate the residuals in the following manner. On the home screen type- **[VARS]** cursor over to Y-VARS then select 1:Function and then 2:Y_2.

Now complete this line: Y_2 (L_1) **þ** L_3 (use the **[STO ▶]** to get the **þ**). This will use your manual fit sine function, which is in Y_2 of the function editor and place the time (stored in L_1) into the function, evaluate it and store it into the list L_3. In L_4, the residuals would be calculated as L_2 - L_3. In Plot 3 set up a scatter plot with the x list as L_1 and the y list as L_4 (manual fit residuals).

For the sum of the squared residuals of the manual fit complete the line: SUM (L_4^2) and then press **[ENTER]**.

SUM (L_4^2) = _____

If the residuals are randomly distributed about zero error (a perfect fit), then the positive error should balance the negative or the sum of the residuals should equal zero or at least a very small number.

SUM (LRESID) = _____ SUM (L_4) = _____

Bias, positive or negative, in the sum of the residuals indicates a miss fit of the model to the data.

Another way of assessing the quality of fit of a regression is by looking into the analysis of variance (ANOVA). The purpose of ANOVA is to partition the total errors created when developing a regression model into error due to the regression and random errors within the data.

Summary

The level of treatment of periodic data can vary from just simple graphical analysis to more advanced treatment of the data. The simple graphical analysis can be performed on computer spreadsheets such as Excel. The fitting of a sine function and the statistics of goodness of fit are more suitable for advanced classes. Many natural cycles such as temperature, tides, salinity, and dissolved oxygen in the Chesapeake Bay, can be discovered using periodic data. In physics, the oscillating mass on a spring and the pendulum generate periodic data.

Data-driven analysis and modeling show the elegant connection between mathematics and science and are in stride with national standards in both mathematics and science.

Internet Sources of Periodic Data

Maryland Department of Natural Resources (DNR) for 20 sites on the Bay and its tributaries (temperature, dissolved oxygen, salinity) http://www.dnr.state.md.us/bay/conditions

Historical weather data such as average air temperature, rainfall, and other data in table form by location at http://www.weatherpost.com

Carbon dioxide data collected at Mauna Loa Observatory in Hawaii from C.D. Keeling and T.P. Whorf of Scripps Institute of Oceanography at http://cdiac.esd.ornl.gov/new/keel_page.html (Data can be downloaded and placed into

Excel for handling) The monthly carbon dioxide levels show a changing midline over many years; hence the midline takes the form of: $y = mx + b$

National Solar Observatory at Kitt Peak has a monthly record of sunspot numbers from 1953 to present at http://argo.tuc.noao.edu/nsokp/dataarch.html (Click on sunspot numbers at bottom of page) This is noisy real data!

Chesapeake Bay Observing System (CBOS) has real-time data (temperature, salinity, current, precipitation, winds) available from a number of buoys in the Bay at http://www.cbos.org (Data can be downloaded as an Excel file) This is a great site with easy download to Excel! Do air-sea temperature differences exist?

NOAA Chesapeake Bay ports tidal data at http://www.co-ops.nos.noaa.gov/cbports/upper.html (can be downloaded) Shows graphs of predicted tides generated by a complex mathematical model of tides and then the actual tidal data – watch for differences!