Chapter 460

Harmonic Regression

Introduction

This program calculates the harmonic regression of a time series. That is, it fits designated harmonics (sinusoidal terms of different wavelengths) using our nonlinear regression algorithms. This could be accomplished using NCSS's Multiple Regression procedure by first generating the harmonics using appropriate sine and cosine transformations and then fitting them in a regression analysis. This procedure allows you to avoid generating these trigonometric terms, plus it automatically generates useful reports and plots specific to time series data.

Harmonic regression is discussed in Chatfield (2004) and Bloomfield (1976).

Technical Details

This section provides the technical details of the model that is fit by this procedure.

Time Series Variable

Suppose we believe that a time series, X_t, contains a periodic (cyclic) component. A natural model of the sinusoidal component would be

$$X_t = \mu + R\cos(ft + d) + e_t$$

where

μ is the mean of the series.

R is the amplitude of variation. Normally, the cosine varies between -1 and 1. Hence, if R is 6, then the term would vary between -6 and 6. The impact of the amplitude is in the size (height or magnitude) of the wave. The length of the wave is not influenced by the amplitude.

f is the frequency of periodic variation, measured in number of radians per unit time. This is the 'frequency' scale of the plots. If we divide 2π by f, we get the corresponding *wavelength*. This is the 'wavelength' scale of the plots. The impact of the frequency is to change the length of a cycle. As f increases, the length of the cycle decreases. A model with $f = 2$ would have a cycle length equal to one-half the cycle length of a model with $f = 1$.

d is the phase or horizontal offset. Changing the phase causes a shift in the beginning of the cycle.

e_t is the random error (noise) of the series about the period component.

t is the time period number. Usually, $t=1, 2, 3, ..., N$. **Note that the *sampling interval* is one. If your sampling interval is different from one, you must rescale your time variable so that it is one.**

Since $\cos(ft+d) = \cos(ft)\cos(d) - \sin(ft)\sin(d)$, this model may be written in the alternative form

$$X_t = \mu + a\cos(ft) + b\sin(ft) + e_t$$

where $a = R\cos(d)$ and $b = -R\sin(d)$.

Hence, this nonlinear model can be fit is a linear regression model with two independent variables. In this case, the independent variables are *X1* = cos(*ft*) and *X2* = sin(*ft*). The regression coefficients are *B1* = *a* and *B2* = *b*. In practice, the variation in a time series may be modeled as the sum of several different individual sinusoidal terms occurring at different frequencies.

The generalization of this model to the sum of *k* frequencies may be written symbolically as

$$X_t = \mu + \sum_{j=1}^{k} R_j \cos\left(f_j t + d_j\right) + e_t$$

or, using the alternative form, as

$$X_t = \mu + \sum_{j=1}^{k} a_j \cos\left(f_j t\right) + \sum_{j=1}^{k} b_j \sin\left(f_j t\right) + e_t$$

Note that if the f_j were known constants, and we let $W_{tr} = \cos(f_r t)$ and $Z_{ts} = \sin(f_s t)$, this could be rewritten in the usual multiple regression form

$$X_t = \mu + \sum_{j=1}^{k} a_j W_{tj} + \sum_{j=1}^{k} b_j Z_{tj} + e_t$$

where the *a's* and the *b's* are regression coefficients to be estimated. This is an example of a harmonic regression.

Harmonic Regression Model

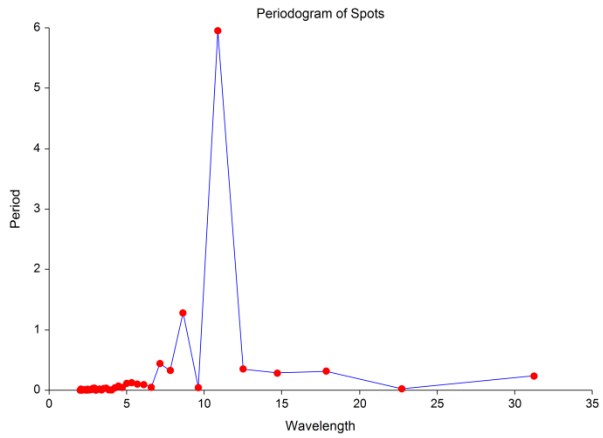
Finally, we can optionally add a trend term to the model to obtain the forecasting equation

$$X_t = \mu + bt + \sum_{j=1}^{k} a_j \cos\left(f_j t\right) + \sum_{j=1}^{k} b_j \sin\left(f_j t\right) + e_t$$

Determining the Appropriate Frequencies using Spectral Analysis

The most difficult task for you the analyst is to determine the appropriate set of frequencies to fit in the harmonic regression model. This is most easily accomplished using the Spectral Analysis program. By inspecting the periodogram, you can determine those frequencies (or wavelengths) that should be represented in the regression model.

For example, spikes appear to occur in the following periodogram at wavelengths of about 11, 9, and perhaps 7. Usually, the scale of the horizontal axis would be changed focus on the wavelengths of interest. In this example, we would create a second periodogram showing wavelengths between 6 and 15. This would allow us to better determine the exact wavelengths we would want to use in a harmonic regression analysis.



Data Structure

The data are entered in a two variables: one containing time values and the other containing the value of the dependent variable.

Missing Values

Missing values are ignored. If only the response value is missing, the value predicted by the model will be generated in the Predicted Values report.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variables on which to run the analysis.

Variables

Y (Dependent) Variable

Specify the column containing the dependent (Y) variable. This variable is to be predicted by the harmonic regression.

The actual values fed into the prediction equation depend on which transformation (if any) is selected for this variable in the Transformation box to the right.

Transformation

Specifies a power transformation for the indicated variable.

Available transformations are

Y' = 1/(Y^2) = 1/(Y*Y)

Y' = 1/Y

Y' = 1/SQRT(Y)

Y' = LN(Y)

Y' = SQRT(Y)

Y' = Y (None)

Y' = Y^2 = Y*Y

Avoid Creating Missing Values

Care must be taken so that you don’t apply a transformation that omits much of your data. For example, you cannot take the square root of a negative number, so if you apply this transformation to a dependent variable containing negative values, those observations will be treated as missing values and ignored. Similarly, you cannot have a zero in the denominator of a fraction such as 1/Y and you cannot take the logarithm of a number less than or equal to zero.

Model

Fit Time Trend

Check this option to add a straight-line time-trend term to the model.

Wavelengths

The wavelength of a sinusoidal function is the horizontal distance between successive peaks. Each wavelength entered generates both a sine term and a cosine term.

This box contains a list of wavelengths that are to be included in the regression model. For example, if you want to fit a 5-period term and an 11-period term, you would enter '5 11.'

Appropriate wavelengths can be determined by trial and error. However, they are most easily determined by inspecting a periodogram constructed by the Spectral Analysis procedure.

You can enter as many wavelengths as you want, but you should avoid overfitting your data by entering too many wavelengths.

Range of Values

The minimum wavelength is 2. The maximum wavelength is N/2, where N is the length of your time series.

List Syntax

You can enter a consecutive list using a colon with the increment in parentheses. For example, 2:8(2) results in 2 4 6 8. If the parentheses are omitted, the increment is set to one. For example, 2:8 results in 2 3 4 5 6 7 8.

Options Tab

The following options control the nonlinear regression algorithm.

Options

Lambda

This is the starting value of the lambda parameter as defined in Marquardt's procedure. We recommend that you do not change this value unless you are very familiar with both your model and the Marquardt nonlinear regression procedure. Changing this value will influence the speed at which the algorithm converges.

Nash Phi

Nash supplies a factor he calls *phi* for modifying lambda. When the residual sum of squares is large, increasing this value may speed convergence.

Lambda Inc

This is a factor used for increasing lambda when necessary. It influences the rate at which the algorithm converges.

Lambda Dec

This is a factor used for decreasing lambda when necessary. It also influences the rate at which the algorithm converges.

Max Iterations

This sets the maximum number of iterations before the program aborts. If the starting values you have supplied are not appropriate or the model does not fit the data, the algorithm may diverge. Setting this value to an appropriate number (say 500) causes the algorithm to abort after this many iterations.

Zero

This is the value used as zero by the nonlinear algorithm. Because of rounding error, values lower than this value are reset to zero. If unexpected results are obtained, you might try using a smaller value, such as 1E-16. Note that 1E-5 is an abbreviation for the number 0.00001.

Reports Tab

The following options control which reports are displayed.

Select Reports

Run Summary – Predicted Values and Residuals

These options specify which reports are displayed.

Time Values

Enter an optional list of time values at which to report the predicted value of Y and corresponding confidence interval.

You can enter a single number or a list of numbers. The list can be separated with commas or spaces. The list can also be of the form "XX:YY(ZZ)" which means XX to YY by ZZ. If ZZ is omitted, it is assumed to be one.

Examples

10 20 30

1:10

10:90(10)

Report Options

Confidence Level

Specify the confidence level (as a percentage) of all confidence intervals that are reported. Typical confidence lev els are 90, 95, or 99, with 95 being the most common.

Note that you do not need to enter the percent sign.

Variable Names

Specify whether to use variable names or (the longer) variable labels in report headings.

Plots Tab

This section controls the inclusion and the settings of the plots.

Select Plots

Function Plot with Actual Y – Probability Plot with Transformed Y

Each of these options specifies whether the indicated plot is displayed. Click the plot format button to change the plot settings.

Storage Tab

The predicted values, prediction limits, and residuals may be stored on the current database for further analysis. This group of options lets you designate which statistics (if any) should be stored and which variables should receive these statistics. The selected statistics are automatically stored to the current database while the program is executing.

Note that existing data is replaced. Be careful that you do not specify variables that contain important data.

Storage Variables

Store Predicted Values, Residuals, Lower Prediction Limit, and Upper Prediction Limit

The predicted (Yhat) values, residuals (Y-Yhat), lower 100(1-alpha) prediction limits, and upper 100(1-alpha) prediction limits may be stored in the columns specified here.

Example 1 – Harmonic Regression Analysis

This section presents an example of how to run a harmonic regression of a time series. The Spots variable in the Sunspot dataset will be used as the dependent variable. An inspection of the periodogram created by the Spectral Analysis procedure led to the following wavelengths: 9.4, 9.9, 10.6, 11.2, 57.0, and 91.0.

You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu of the Harmonic Regression window.

1 Open the Sunspot dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **Sunspot.NCSS**.
- Click **Open**.

2 Open the Harmonic Regression window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Harmonic Regression** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- Select the **Variables tab**.
- Double-click in the **Y (Dependent)** box. This will bring up the variable selection window.
- Select **Spots** from the list of variables and then click **Ok**.
- Double-click in the **T (Time)** box. This will bring up the variable selection window.
- Select **Year** from the list of variables and then click **Ok**.
- Uncheck the **Fit Time Trend** option
- Set the **Wavelengths** to **9.4 9.9 10.6 11.2 57 91**.

4 Select the Reports

- Select the **Reports tab**.
- Check all reports.
- Set the **Time Values** to **200 300 400**.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Run Summary Section

Item	Value	Item	Value
Dependent Variable	Spots	Total Rows	215
Time Variable	Year	Rows with Missing Values	0
R^2	0.6608	Rows Used	215
Maximum Iterations	1000		
Iterations Used	245		

Estimated Model

(42.4440741649885+(12.6758109351939)*SIN((0.668423968848891*Year))-(1.69709539553941)*COS((0.668423968848891*Year))
+(10.2460302454465)*SIN((0.634665182543392*Year))+(8.20186341536895)*COS((0.634665182543392*Year))
-(10.2226854877311)*SIN((0.592753330865998*Year))+(0.088555879047715)*COS((0.592753330865998*Year))
+(0.0702933646410777)*SIN((0.560998688141034*Year))-(0.253894745929267)*COS((0.560998688141034*Year))
+(0.174272323629069)*SIN((0.110231321178589*Year))+(0.116888691209774)*COS((0.110231321178589*Year))
-(0.283968136405124)*SIN((0.0690459923865888*Year))+(0.0167389449569679)*COS((0.0690459923865888*Year)))

This section shows the variables used, the R^2 value achieved, the number of iterations used, and the number of rows processed. Pay particular attention to whether the R2 value is high (that is, if the model is useful) and whether the algorithm converged before the maximum number of iterations was reached (if it did not, rerun with a higher Maximum Iterations value).

The Estimated Model provides a text version of the estimated model that can be used directly by a transformation.

Regression Coefficients Section

Independent Variable	Regression Coefficient b(i)	Standard Error sb(i)	T-Statistic to Test H0: β(i)=0	Prob Level	Lower 95% Conf. Limit of β(i)	Upper 95% Conf. Limit of β(i)
Intercept	42.44407	1.83054	23.19	0.0000	38.83465	46.05349
Sin(9.4)	12.67581	2.24096	5.66	0.0000	8.25714	17.09448
Cos(9.4)	-1.69710	2.36053	-0.72	0.7635	-6.35154	2.95735
Sin(9.9)	10.24603	2.33573	4.39	0.0000	5.64048	14.85158
Cos(9.9)	8.20186	2.34181	3.50	0.0003	3.58434	12.81938
Sin(10.6)	-10.22269	2.08326	-4.91	1.0000	-14.33040	-6.11497
Cos(10.6)	0.08856	0.02835	3.12	0.0010	0.03265	0.14446
Sin(11.2)	0.07029	0.03707	1.90	0.0297	-0.00279	0.14338
Cos(11.2)	-0.25389	0.03458	-7.34	1.0000	-0.32209	-0.18570
Sin(57)	0.17427	0.03004	5.80	0.0000	0.11503	0.23351
Cos(57)	0.11689	0.03141	3.72	0.0001	0.05495	0.17883
Sin(91)	-0.28397	0.03902	-7.28	1.0000	-0.36091	-0.20702
Cos(91)	0.01674	0.03542	0.47	0.3185	-0.05311	0.08658

This section gives the values of the regression coefficients along with their standard errors, t-values, probability levels, and confidence intervals. Remember that terms must be removed in sine and cosine pairs, so you would consider removing wavelengths that were not significant for either the sine term or the cosine term.

Harmonic Analysis Section

Wave Length	Frequency	Amplitude	Phase	Sine Term Coefficient	Cosine Term Coefficient
9.400	0.6684	12.78891	4.57930	12.67581	-1.69710
9.900	0.6347	13.12447	5.38743	10.24603	8.20186
10.600	0.5928	10.22307	1.56213	-10.22269	0.08856
11.200	0.5610	0.26345	3.41169	0.07029	-0.25389
57.000	0.1102	0.20984	5.30320	0.17427	0.11689
91.000	0.0690	0.28446	1.51192	-0.28397	0.01674

This section gives the frequency, amplitude, and phase for each wavelength computed from the regression coefficients. If we let *w* be the wavelength, *a* be the regression coefficient of the sine term, and *b* be the regression coefficient of the cosine term, the formulas for the other quantities are

Harmonic Regression

$$\text{Frequency} = 2\pi/\text{w}$$

$$\text{Amplitude} = \sqrt{a^2 + b^2}$$

$$\text{Phase} = tan^{-1}(-b/a) \text{ in radians}$$

User-Specified Predicted Values Section

Row No.	Year	Predicted Value	Lower 95% Conf. Limit	Upper 95% Conf. Limit
1	200	164.1713	112.7035	215.6392
2	300	43.38063	-4.75702	91.51827
3	400	52.88081	4.537998	101.2236

This report gives the predicted value (the forecast) for the user-specified time values.

Analysis of Variance Table

	DF	Sum of Squares	Mean Squares
Intercept	1	520496.6417	520496.6417
Model	13	740528.1544	844764.9596
Model (Adjusted)	12	220031.5127	18335.9594
Error	202	112923.2056	559.0258
Total (Adjusted)	214	332954.7183	
Total	215	853451.3600	

This report gives the ANOVA table.

Correlation Matrix of Regression Coefficients

	Intercept	Sin(9.4)	Cos(9.4)	Sin(9.9)	Cos(9.9)	Sin(10.6)
Intercept	1.000000	0.066011	-0.059928	0.251837	0.030540	-0.143220
Sin(9.4)	0.066011	1.000000	0.036881	-0.100383	-0.071487	0.247582
Cos(9.4)	-0.059928	0.036881	1.000000	0.021000	-0.011833	0.289158
Sin(9.9)	0.251837	-0.100383	0.021000	1.000000	-0.033422	-0.016148
Cos(9.9)	0.030540	-0.071487	-0.011833	-0.033422	1.000000	0.162717
Sin(10.6)	-0.143220	0.247582	0.289158	-0.016148	0.162717	1.000000
Cos(10.6)	-0.197905	0.148133	-0.244638	0.003784	-0.112719	0.036038
Sin(11.2)	0.003491	-0.167979	0.052957	-0.027518	-0.424011	-0.049825
Cos(11.2)	0.251568	-0.112511	-0.318650	0.332287	-0.236088	-0.066055
Sin(57)	-0.057524	0.285503	0.367727	-0.047382	0.006949	0.146731
Cos(57)	-0.187754	-0.177177	0.216559	0.093898	-0.026476	0.057712
Sin(91)	0.409967	0.020678	0.274172	0.308611	-0.086039	-0.087717
Cos(91)	-0.100740	-0.028987	0.040076	-0.013052	0.469344	0.121039

(Report Continues for other coefficients)

This report displays the asymptotic correlations of the parameter estimates. When these correlations are high (absolute value greater than 0.95), the precision of the parameter estimates is suspect.

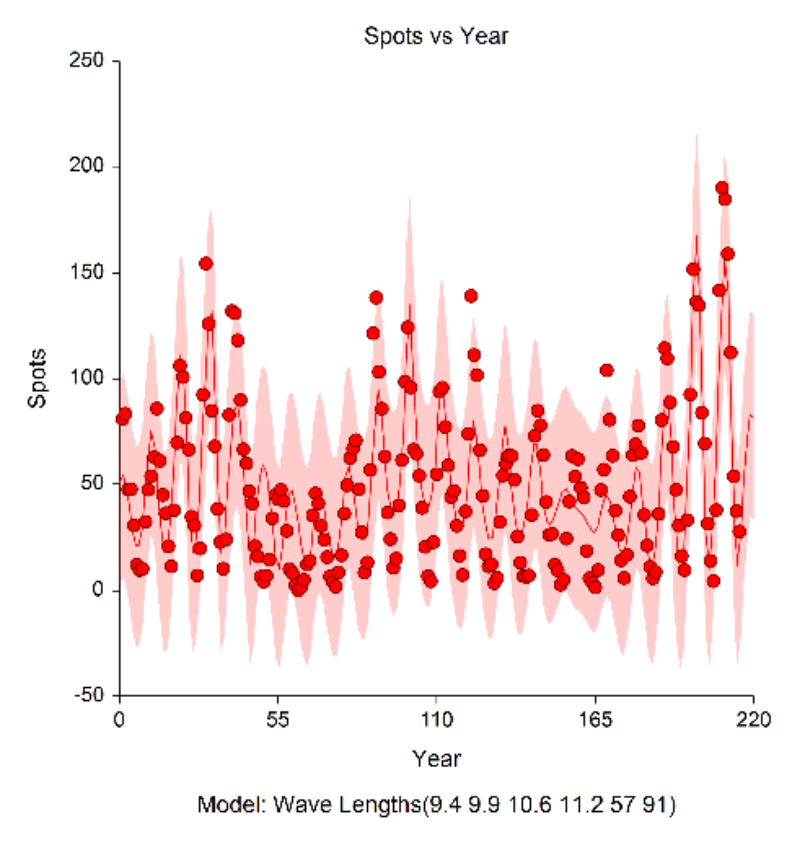
Predicted Values and Residuals Section

Row No.	Year	Spots	Predicted Value	Lower 95% Pred. Limit	Upper 95% Pred. Limit	Residual
1	1	80.9	54.39597	6.57971	102.2122	26.50403
2	2	83.4	51.18402	3.600881	98.76717	32.21598
3	3	47.7	42.8599	-4.537165	90.25696	4.840104
4	4	47.8	32.8028	-14.4662	80.0718	14.9972
5	5	30.7	24.29735	-22.88925	71.48395	6.402653
6	6	12.2	20.14161	-27.04814	67.33134	-7.941605
7	7	9.6	22.56336	-24.7057	69.83243	-12.96336
8	8	10.2	32.6922	-14.67924	80.06363	-22.49219
9	9	32.4	49.11187	1.560219	96.66352	-16.71187
10	10	47.6	66.08908	18.21012	113.968	-18.48908

(Report Continues)

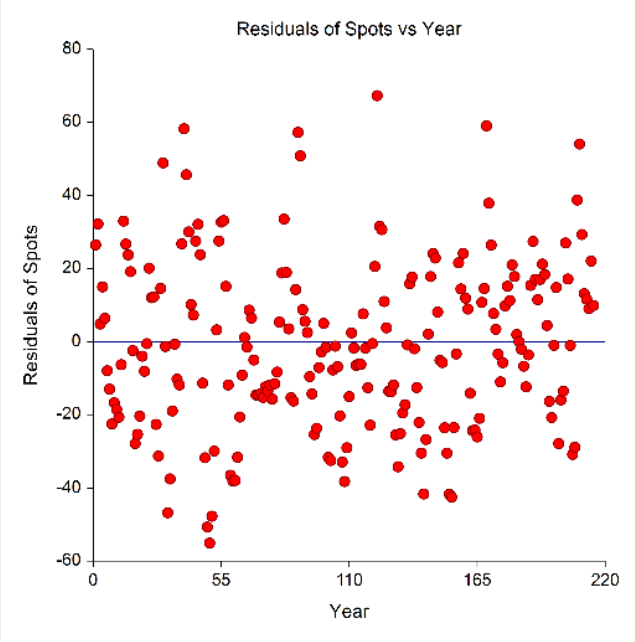
The section shows the predicted value, prediction interval, and residual for each row. If you have observations in which the independent variable is given, but the dependent (Y) variable was left blank, a predicted value and prediction limits will be generated and displayed in this report.

Function Plot(s)



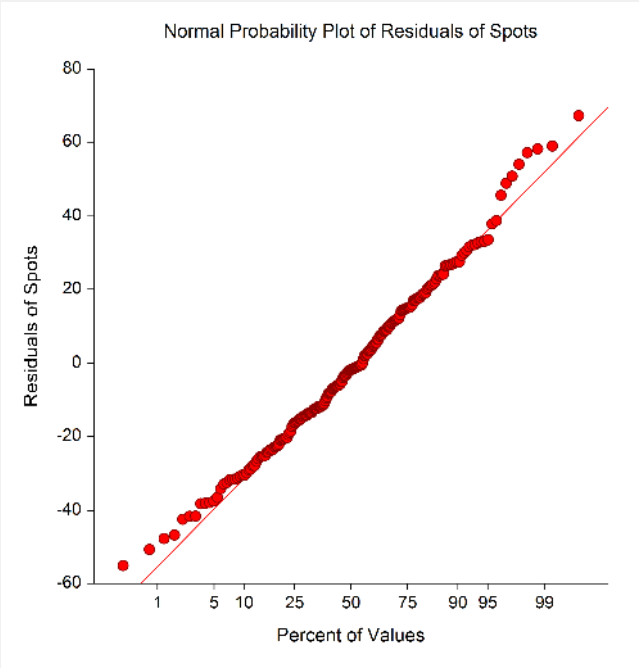
This plot lets you visually assess the fit. It shows the time series as dots, the model as a line, and the prediction limits as a shaded region.

Residual Plot(s)



This plot lets you visually assess the fit. It shows the residuals across time.

Probability Plot(s)



This plot allows you to assess the normality of the residuals.