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

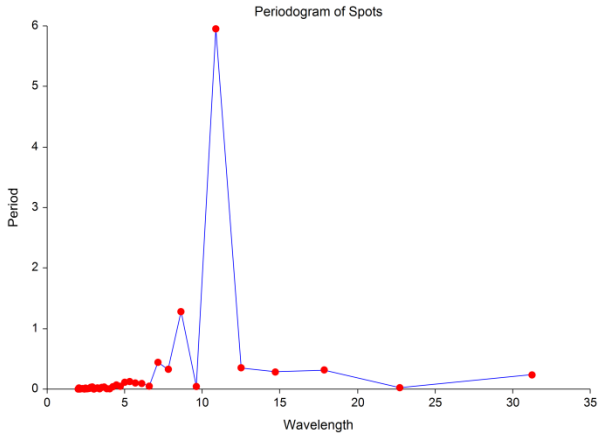
Finally, we can optionally add a trend term to the model to obtain the forecasting equation

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

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 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.

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.

Setup

To run this example, complete the following steps:

1 Open the Sunspot example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Sunspot** and click **OK**.

2 Specify the Harmonic Regression procedure options

- Find and open the **Harmonic Regression** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab	
Y (Dependent)	**Spots**
T (Time)	**Year**
Fit Time Trend	**Unchecked**
Wavelengths	**9.4 9.9 10.6 11.2 57.0 91.0**

Reports Tab	
Run Summary	**Checked**
Regression Coefficients	**Checked**
Harmonic Analysis	**Checked**
Predict Y at Specified Time Values	**Checked**
Time Values	**200 300 400**
Analysis of Variance Table	**Checked**
Parameter Correlations	**Checked**
Predicted Values and Residuals	**Checked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Run Summary Section

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	232		

Estimated Model

(42.4440777001334+(12.6758109351947)*SIN((0.668423968848891*Year))-(1.69709539553923)*COS((0.668423968848891*Year))+(10.2460245379788)*SIN((0.634665182543392*Year))+(8.20186341536487)*COS((0.634665182543392*Year))-(10.22268495783)*SIN((0.592753330865998*Year))+(0.088555879047667)*COS((0.592753330865998*Year))+(0.0702933646411201)*SIN((0.560998688141034*Year))-(0.253894745929229)*COS((0.560998688141034*Year))+(0.174272323629067)*SIN((0.110231321178589*Year))+(0.116888691209716)*COS((0.110231321178589*Year))-(0.283968136405105)*SIN((0.0690459923865888*Year))+(0.0167389449568583)*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

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	42.44408	1.83054	23.19	0.0000	38.83466	46.05350
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)	10.24602	2.33573	4.39	0.0000	5.64048	14.85157
Cos(9.9)	8.20186	2.34181	3.50	0.0003	3.58434	12.81938
Sin(10.6)	-10.22268	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

Harmonic Analysis Section

Wave Length	Frequency	Amplitude	Phase	Sine Term Coefficient	Cosine Term Coefficient
9.400	0.6684	12.78891	-1.70389	12.67581	-1.69710
9.900	0.6347	13.12447	-0.89576	10.24602	8.20186
10.600	0.5928	10.22307	1.56213	-10.22268	0.08856
11.200	0.5610	0.26345	-2.87150	0.07029	-0.25389
57.000	0.1102	0.20984	-0.97999	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

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

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

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

User-Specified Predicted Values Section

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	43.38062	-4.757022	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

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

Correlation Matrix of Regression Coefficients

Section 1

	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 the 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

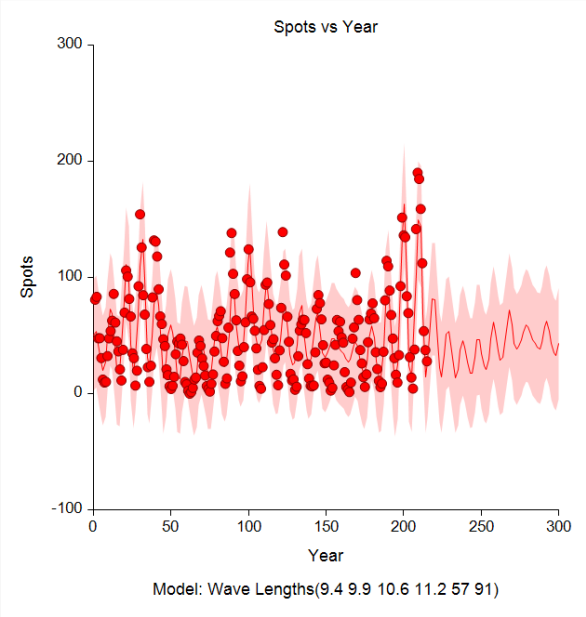
Predicted Values and Residuals Section

Row No.	Year	Spots	Predicted Value	Lower 95% Pred. Limit	Upper 95% Pred. Limit	Residual
1	1	80.9	54.39597	6.579711	102.2122	26.50403
2	2	83.4	51.18402	3.60088	98.76717	32.21598
3	3	47.7	42.85989	-4.537166	90.25696	4.840106
4	4	47.8	32.8028	-14.4662	80.0718	14.9972
5	5	30.7	24.29735	-22.88925	71.48395	6.40265
6	6	12.2	20.14161	-27.04813	67.33135	-7.941611
7	7	9.6	22.56337	-24.7057	69.83244	-12.96337
8	8	10.2	32.6922	-14.67924	80.06364	-22.4922
9	9	32.4	49.11188	1.560225	96.66353	-16.71188
10	10	47.6	66.08909	18.21012	113.968	-18.48909
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

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)

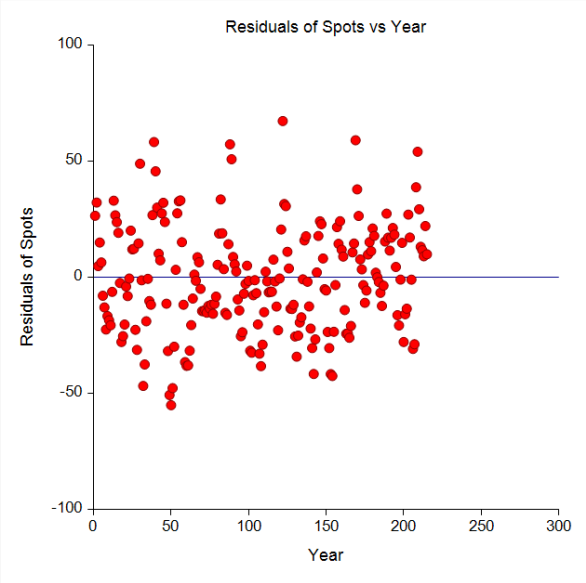
Function Plot(s)



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)

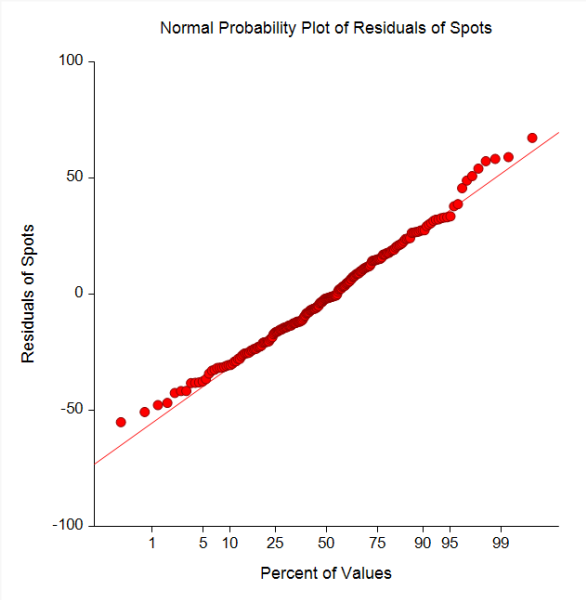
Residual Plot(s)



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

Probability Plot(s)

Probability Plot(s)



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