Chapter 38
The LOESS Procedure

Chapter Table of Contents

Chapter 38
The LOESS Procedure

Overview

The LOESS procedure implements a nonparametric method for estimating regression surfaces pioneered by Cleveland, Devlin, and Grosse (1988), Cleveland and Grosse (1991), and Cleveland, Grosse, and Shyu (1992). The LOESS procedure allows great flexibility because no assumptions about the parametric form of the regression surface are needed.

The SAS System provides many regression procedures such as the GLM, REG, and NLIN procedures for situations in which you can specify a reasonable parametric model for the regression surface. You can use the LOESS procedure for situations in which you do not know a suitable parametric form of the regression surface. Furthermore, the LOESS procedure is suitable when there are outliers in the data and a robust fitting method is necessary.

The main features of the LOESS procedure are as follows:

- fits nonparametric models
- supports the use of multidimensional data
- supports multiple dependent variables
- supports both direct and interpolated fitting using kd trees
- performs statistical inference
- performs iterative reweighting to provide robust fitting when there are outliers in the data
- supports multiple SCORE statements

Local Regression and the Loess Method

Assume that for $i = 1$ to n, the ith measurement y_i of the response y and the corresponding measurement x_i of the vector x of p predictors are related by

$$y_i = g(x_i) + \epsilon_i$$

where g is the regression function and ϵ_i is a random error. The idea of local regression is that at a predictor x, the regression function $g(x)$ can be locally approximated by the value of a function in some specified parametric class. Such a local approximation is obtained by fitting a regression surface to the data points within a chosen neighborhood of the point x.

In the loess method, weighted least squares is used to fit linear or quadratic functions of the predictors at the centers of neighborhoods. The radius of each neighborhood is chosen so that the neighborhood contains a specified percentage of the data points. The fraction of the data, called the *smoothing parameter*, in each local neighborhood controls the smoothness of the estimated surface. Data points in a given local neighborhood are weighted by a smooth decreasing function of their distance from the center of the neighborhood.

In a direct implementation, such fitting is done at each point at which the regression surface is to be estimated. A much faster computational procedure is to perform such local fitting at a selected sample of points in predictor space and then to blend these local polynomials to obtain a regression surface.

You can use the LOESS procedure to perform statistical inference provided the error distribution satisfies some basic assumptions. In particular, such analysis is appropriate when the ϵ_i are i.i.d. normal random variables with mean 0. By using the iterative reweighting, the LOESS procedure can also provide statistical inference when the error distribution is symmetric but not necessarily normal. Furthermore, by doing iterative reweighting, you can use the LOESS procedure to perform robust fitting in the presence of outliers in the data.

While all output of the LOESS procedure can be optionally displayed, most often the LOESS procedure is used to produce output data sets that will be viewed and manipulated by other SAS procedures. PROC LOESS uses the Output Delivery System (ODS) to place results in output data sets. This is a departure from older SAS procedures that provide OUTPUT statements to create SAS data sets from analysis results.

Getting Started

Scatter Plot Smoothing

The following data from the Connecticut Tumor Registry presents age-adjusted numbers of melanoma incidences per 100,000 people for 37 years from 1936 to 1972 (Houghton, Flannery, and Viola, 1980).

```
data Melanoma;
   input  Year Incidences @@;
   format Year d4.0;
   format DepVar d4.1;
datalines;
1936    0.9   1937   0.8  1938   0.8  1939   1.3
1940    1.4   1941   1.2  1942   1.7  1943   1.8
1944    1.6   1945   1.5  1946   1.5  1947   2.0
1948    2.5   1949   2.7  1950   2.9  1951   2.5
1952    3.1   1953   2.4  1954   2.2  1955   2.9
1956    2.5   1957   2.6  1958   3.2  1959   3.8
1960    4.2   1961   3.9  1962   3.7  1963   3.3
1964    3.7   1965   3.9  1966   4.1  1967   3.8
1968    4.7   1969   4.4  1970   4.8  1971   4.8
1972    4.8
;
```

The following PROC GPLOT statements produce the simple scatter plot of these data displayed in Figure 38.1.

```
symbol1 color=black value=dot ;
proc gplot data=Melanoma;
   title1 'Scatter Plot of Melanoma Data';
   plot Incidences*Year;
run;
```

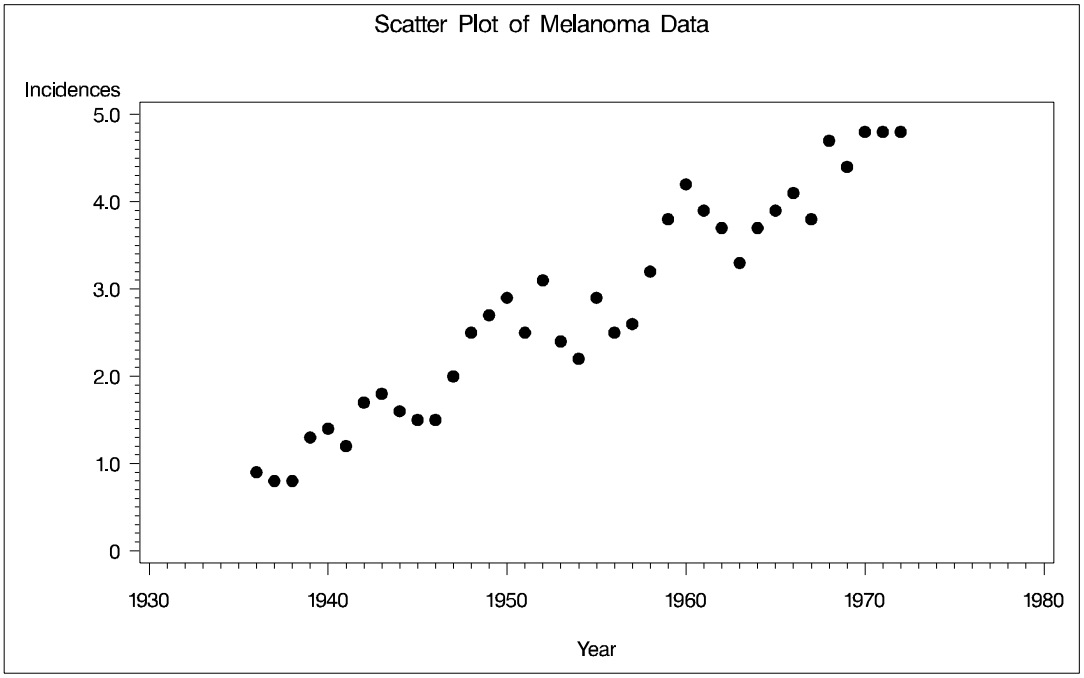


Figure 38.1. Scatter Plot of Incidences versus Year for the Melanoma Data

Suppose that you want to smooth the response variable Incidences as a function of the variable Year. The following PROC LOESS statements request this analysis:

```
proc loess data=Melanoma;
   model Incidences=Year/details(OutputStatistics);
run;
```

You use the PROC LOESS statement to invoke the procedure and specify the data set. The MODEL statement names the dependent and independent variables. You use options in the MODEL statement to specify fitting parameters and control the displayed output. For example, the MODEL statement option DETAILS(OutputStatistics) requests that the "Output Statistics" table be included in the displayed output. By default, this table is not displayed.

The results are displayed in Figure 38.2 and Figure 38.3.

Loess Fit of Melanoma Data

The LOESS Procedure

Independent Variable Scaling

Scaling applied: None

Statistic	Year
Minimum Value	1936
Maximum Value	1972

Loess Fit of Melanoma Data

The LOESS Procedure
Smoothing Parameter: 0.5
Dependent Variable: Incidences

Output Statistics

Obs	Year	Incidences	Predicted Incidences
1	1936	0.9	0.79168
2	1937	0.8	0.90451
3	1938	0.8	1.01734
4	1939	1.3	1.13103
5	1940	1.4	1.24472
6	1941	1.2	1.36308
7	1942	1.7	1.48143
8	1943	1.8	1.59978
9	1944	1.6	1.73162
10	1945	1.5	1.86345
11	1946	1.5	1.97959
12	1947	2.0	2.09573
13	1948	2.5	2.21187
14	1949	2.7	2.30363
15	1950	2.9	2.39539
16	1951	2.5	2.48929
17	1952	3.1	2.58320
18	1953	2.4	2.68985
19	1954	2.2	2.79649
20	1955	2.9	2.89805
21	1956	2.5	2.99960
22	1957	2.6	3.10116
23	1958	3.2	3.20623
24	1959	3.8	3.31130
25	1960	4.2	3.43311
26	1961	3.9	3.55493
27	1962	3.7	3.67934
28	1963	3.3	3.80375
29	1964	3.7	3.91434
30	1965	3.9	4.02493
31	1966	4.1	4.13552
32	1967	3.8	4.24475
33	1968	4.7	4.35398
34	1969	4.4	4.46846
35	1970	4.8	4.58293
36	1971	4.8	4.70316
37	1972	4.8	4.82338

Figure 38.2. Output from PROC LOESS

Loess Fit of Melanoma Data

The LOESS Procedure
Smoothing Parameter: 0.5
Dependent Variable: Incidences

Fit Summary

Fit Method	**Interpolation**
Number of Observations	**37**
Number of Fitting Points	**17**
kd Tree Bucket Size	**3**
Degree of Local Polynomials	**1**
Smoothing Parameter	**0.50000**
Points in Local Neighborhood	**18**
Residual Sum of Squares	**3.97047**

Figure 38.3. Output from PROC LOESS continued

Usually, such displayed results are of limited use. Most frequently the results are needed in an output data set so that they can be displayed graphically and analyzed further. For example, to place the "Output Statistics" table shown in Figure 38.2 in an output data set, you use the ODS OUTPUT statement as follows:

```
proc loess data=Melanoma;
   model Incidences=Year;
   ods output OutputStatistics=Results;
run;
```

The statement

```
ods output OutputStatistics=Results;
```

requests that the "Output Statistics" table that appears in Figure 38.2 be placed in a SAS data set named Results. Note also that the DETAILS(OutputStatistics) option that caused this table to be included in the displayed output need not be specified.

The PRINT procedure displays the first five observations of this data set:

```
title1 'First 5 Observations of the Results Data Set';
proc print data=Results(obs=5);
  id obs;
run;
```

First 5 Observations of the Results Data Set

Obs	Smoothing Parameter	Dependent	Year	Dep Var	Pred
1	0.5	Incidences	1936	0.9	0.79168
2	0.5	Incidences	1937	0.8	0.90451
3	0.5	Incidences	1938	0.8	1.01734
4	0.5	Incidences	1939	1.3	1.13103
5	0.5	Incidences	1940	1.4	1.24472

Figure 38.4. PROC PRINT Output of the Results Data Set

You can now produce a scatter plot including the fitted loess curve as follows:

```
symbol1 color=black value=dot;
symbol2 color=black interpol=join value=none;

/* macro used in subsequent examples */
%let opts=vaxis=axis1 hm=3 vm=3 overlay;

axis1 label=(angle=90 rotate=0);

proc gplot data=Results;
  title1 'Melanoma Data with Default LOESS Fit';
  plot  DepVar*Year Pred*Year/ &opts;
run;
```

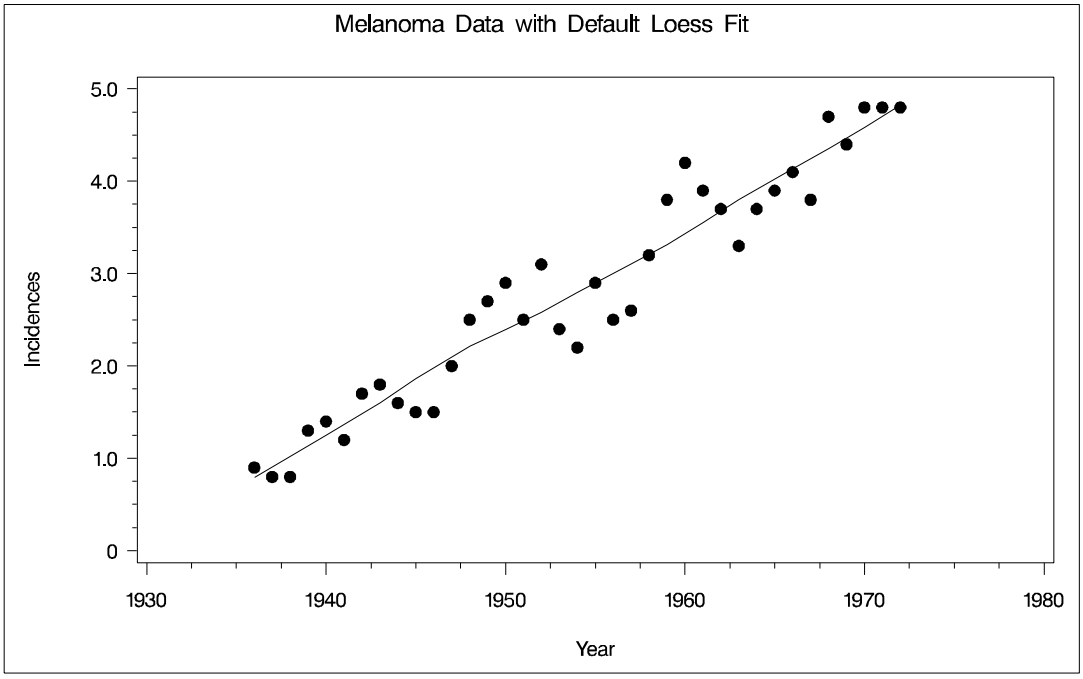


Figure 38.5. Default Loess FIT for Melanoma Data

The loess fit shown in Figure 38.5 was obtained with the default value of the smoothing parameter, which is 0.5. It is evident that this results in a loess fit that is too smooth for the Melanoma data. The loess fit captures the increasing trend in the data but does not reflect the periodic pattern in the data, which is related to an 11-year sunspot activity cycle. By using the SMOOTH= option in the MODEL statement, you can obtain loess fits for a range of smoothing parameters as follows:

```
proc loess data=Melanoma;
   model Incidences=Year/smooth=0.1 0.2 0.3 0.4 residual;
   ods output OutputStatistics=Results;
run;
```

The RESIDUAL option causes the residuals to be added to the "Output Statistics" table. PROC PRINT displays the first five observations of this data set:

```
proc print data=Results(obs=5);
  id obs;
run;
```

First 5 Observations of the Results Data Set

Obs	Smoothing Parameter	Dependent	Year	Dep Var	Pred	Residual
1	0.1	Incidences	1936	0.9	0.90000	0
2	0.1	Incidences	1937	0.8	0.80000	0
3	0.1	Incidences	1938	0.8	0.80000	0
4	0.1	Incidences	1939	1.3	1.30000	0
5	0.1	Incidences	1940	1.4	1.40000	0

Figure 38.6. PROC PRINT Output of the Results Data Set

Note that the fits for all the smoothing parameters are placed in single data set and that ODS has added a SmoothingParameter variable to this data set that you can use to distinguish each fit.

The following statements display the loess fits obtained in a 2 by 2 plot grid:

```
goptions  nodisplay;
proc gplot data=Results;
   by SmoothingParameter;
   plot DepVar*Year=1 Pred*Year/ &opts name='fit';
run; quit;

goptions display;
proc greplay nofs tc=sashelp.templt template=l2r2;
    igout gseg;
    treplay 1:fit 2:fit2 3:fit1 4:fit3;
run; quit;
```

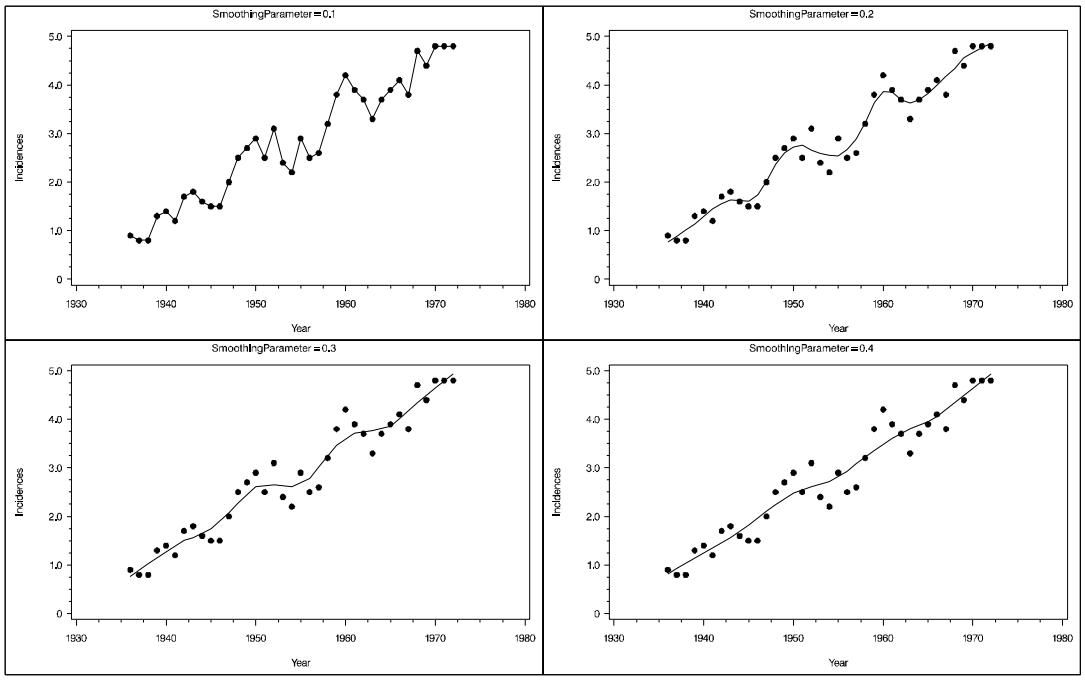


Figure 38.7. Loess Fits with a Range of Smoothing Parameters

If you examine the plots in Figure 38.7, you see that a good fit is obtained with smoothing parameter 0.2. You can gain further insight in how to choose the smoothing parameter by examining scatter plots of the fit residuals versus the year. To aid the interpretation of these scatter plots, you can again use PROC LOESS to smooth the response Residual as a function of Year.

```
proc loess data=Results;
   by SmoothingParameter;
   ods output OutputStatistics=residout;
   model Residual=Year/smooth=0.3;
run;

axis1 label = (angle=90 rotate=0)
      order = (-0.8 to 0.8 by 0.4);
goptions nodisplay;
proc gplot data=residout;
   by SmoothingParameter;
   plot  DepVar*Year Pred*Year / &opts vref=0 lv=2 vm=1
                                 name='resids';
run; quit;

goptions display;
proc greplay nofs tc=sashelp.templt template=l2r2;
    igout gseg;
    treplay 1:resids 2:resids2 3:resids1 4:resids3;
run; quit;
```

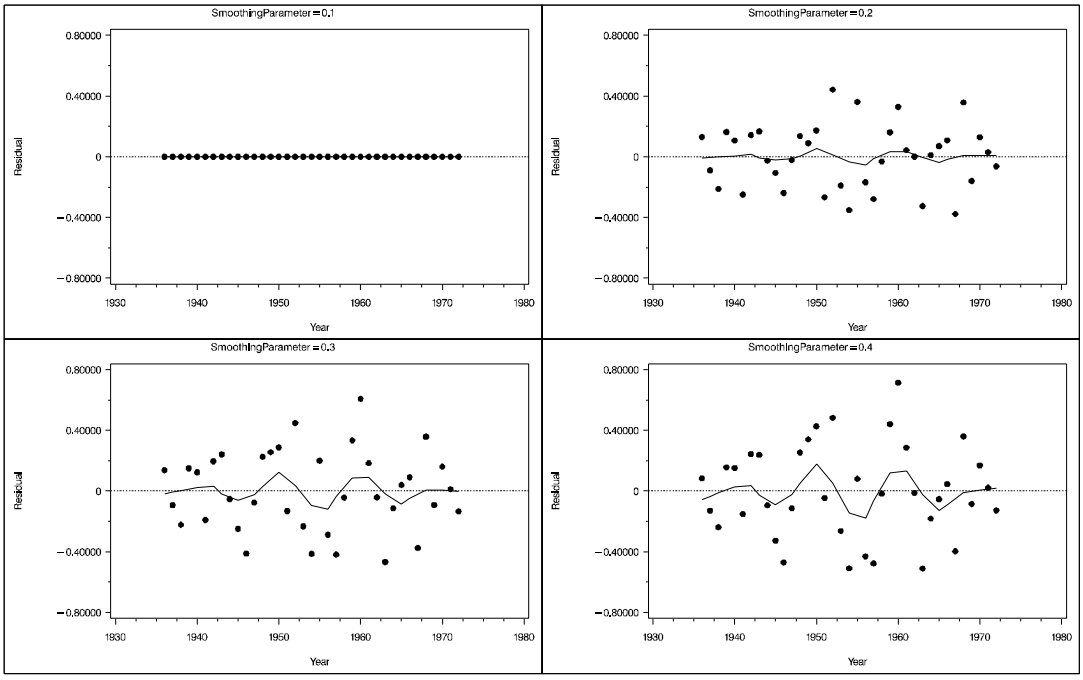


Figure 38.8. Scatter Plots of Residuals versus Year

Looking at the scatter plots in Figure 38.8 confirms that the choice of smoothing parameter 0.2 is reasonable. With smoothing parameter 0.1, there is gross overfitting in

the sense that the original data are exactly interpolated. The loess fits on the Residual versus Year scatter plots for smoothing parameters 0.3 and 0.4 reveal that there is a periodic trend in the residuals that is much weaker when the smoothing parameter is 0.2. This suggests that when the smoothing parameter is above 0.3, an overly smooth fit is obtained that misses essential features in the original data.

Having now decided on a loess fit, you may want to obtain confidence limits for your model predictions. This is done by adding the CLM option in the MODEL statement. By default 95% limits are produced, but this can be changed by using the ALPHA= option in the MODEL statement. The following statements add 90% confidence limits to the Results data set and display the results graphically:

```
proc loess data=Melanoma;
   model Incidences=Year/smooth=0.2 residual clm
                        alpha=0.1;
   ods output OutputStatistics=Results;
run;

symbol3 color=green interpol=join value=none;
symbol4 color=green interpol=join value=none;
axis1 label = (angle=90 rotate=0)
      order = (0 to 6);
title1 'Age-adjusted Melanoma Incidences for 37 Years';



proc gplot data=Results;
  plot  DepVar*Year Pred*Year LowerCl*Year UpperCL*Year
                    /  &opts;
run;
```

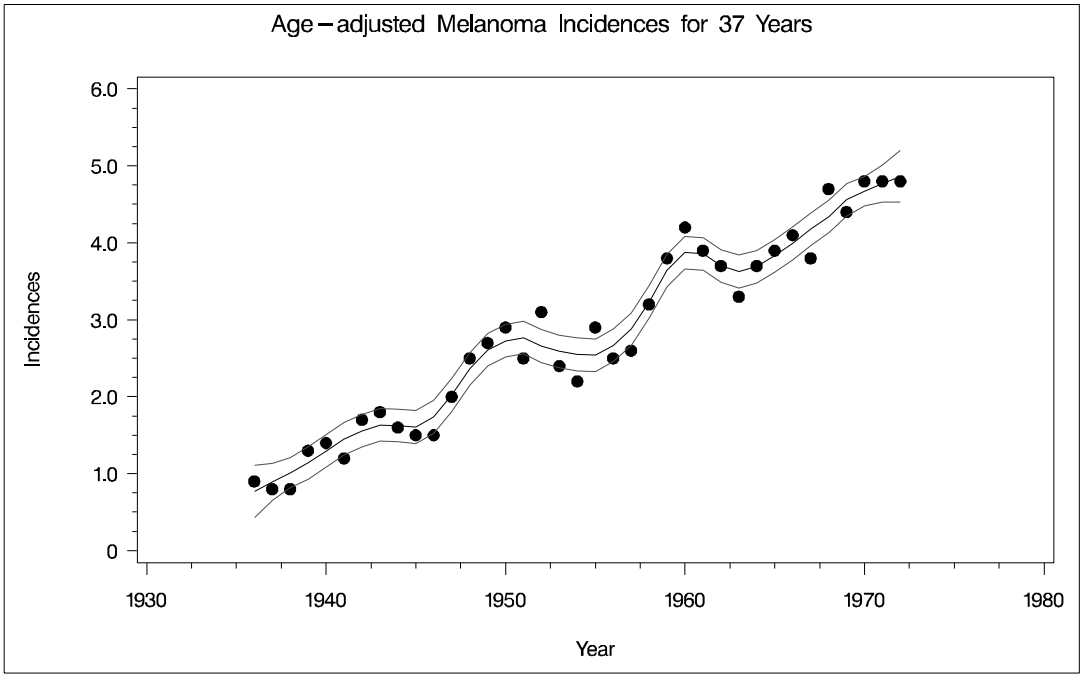


Figure 38.9. Loess fit of Melanoma Data with 90% Confidence Bands

Syntax

The following statements are available in PROC LOESS:

PROC LOESS <*DATA=SAS-data-set*> **;**

MODEL *dependents=regressors* < */ options* > **;**
ID *variables* **;**
BY *variables* **;**
WEIGHT *variable* **;**
SCORE *DATA=SAS-data-set* < *ID=(variable list)* > < */ options* > **;**

The PROC LOESS and MODEL statements are required. The BY, WEIGHT, and ID statements are optional. The SCORE statement is optional, and more than one SCORE statement can be used.

The statements used with the LOESS procedure, in addition to the PROC LOESS statement, are as follows.

BY	specifies variables to define subgroups for the analysis.
ID	names variables to identify observations in the displayed output.
MODEL	specifies the dependent and independent variables in the loess model, details and parameters for the computational algorithm, and the required output.
SCORE	specifies a data set containing observations to be scored.
WEIGHT	declares a variable to weight observations.

PROC LOESS Statement

PROC LOESS <*DATA=SAS-data-set*> **;**

The PROC LOESS statement is required. The only option in this statement is the DATA= option, which names a data set to use for the loess model.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC LOESS to obtain separate analyses on observations in groups defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in the order of the BY variables. The *variables* are one or more variables in the input data set.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data using the SORT procedure with a similar BY statement.
- Specify the BY statement option NOTSORTED. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables using the DATASETS procedure (in Base SAS software).

For more information on the BY statement, refer to the discussion in *SAS Language Reference: Concepts*. For more information on the DATASETS procedure, refer to the discussion in the *SAS Procedures Guide*.

ID Statement

ID *variables* **;**

The ID statement is optional, and more than one ID statement can be used. The variables listed in any of the ID statements are displayed in the "Output Statistics" table beside each observation. Any variables specified as a regressor or dependent variable in the MODEL statement already appear in the "Output Statistics" table and are not treated as ID variables, even if they appear in the variable list of an ID statement.

MODEL Statement

MODEL *dependents=independent variables* < */ options* > **;**

The MODEL statement names the dependent variables and the independent variables. Variables specified in the MODEL statement must be numeric variables in the data set being analyzed.

Table 38.1 lists the options available in the MODEL statement.

Table 38.1. Model Statement Options

Option	Description
Fitting Parameters	
DIRECT	specifies direct fitting at every data point
SMOOTH=	specifies the list of smoothing values
DEGREE=	specifies the degree of local polynomials (1 or 2)
DROPSQUARE=	specifies the variables whose squares are to be dropped from local quadratic polynomials
BUCKET=	specifies the number of points in kd tree buckets
ITERATIONS=	specifies the number of reweighting iterations
DFMETHOD=	specifies the method of computing lookup degrees of freedom
Residuals and Confidence limits	
ALL	requests the following options: CLM, RESIDUAL, STD, SCALEDINDEP
CLM	displays $100(1-\alpha)$% confidence interval for the mean predicted value
RESIDUAL	displays residual statistics
STD	displays estimated prediction standard deviation
T	displays t statistics
Display Options	
DETAILS=	specifies which tables are to be displayed

Table 38.1. (continued)

Option	Description
Other options	
ALPHA=	sets significance value for confidence intervals
SCALE=	specifies the method used to scale the regressor variables
SCALEDINDEP	displays scaled independent variable coordinates

The following options are available in the MODEL statement after a slash (/).

ALL
requests all these options: CLM, RESIDUAL, SCALEDINDEP, STD, and T.

ALPHA=*number*
sets the significance level used for the construction of confidence intervals for the current MODEL statement. The value must be between 0 and 1; the default value of 0.05 results in 95% intervals.

BUCKET=*number*
specifies the maximum number of points in the leaf nodes of the kd tree. The default value used is $s * n/5$, where s is a smoothing parameter specified using the SMOOTH= option and n is the number of observations being used in the current BY group. The BUCKET= option is ignored if the DIRECT option is specified.

CLM
requests that $100(1-\alpha)$ confidence limits on the mean predicted value be added to the "Output Statistics" table. By default, 95% limits are computed; the ALPHA= option in the MODEL statement can be used to change the α-level. The use of this option implicitly selects the model option DFMETHOD=EXACT if the DFMETHOD= option has not been explicitly used.

DEGREE= 1 | 2
sets the degree of the local polynomials to use for each local regression. The valid values are 1 for local linear fitting or 2 for local quadratic fitting, with 1 being the default.

DETAILS < **(** *tables* **)** >
selects which tables to display, where *tables* is one or more of kdTree (or TREE), PredAtVertices (or FITPOINTS), and OutputStatistics (or STATOUT). A specification of kdTree outputs the kd tree structure, PredAtVertices outputs fitted values and coordinates of the kd tree vertices where the local least squares fitting is done, and OutputStatistics outputs the predicted values and other requested statistics at the points in the input data set. The kdTree and PredAtVertices specifications are ignored if the DIRECT option is specified in the MODEL statement. Specifying the option DETAILS with no qualifying list outputs all tables.

DFMETHOD= NONE | EXACT
specifies the method used to calculate the "lookup" degrees of freedom used in performing statistical inference. The default is DFMETHOD=NONE. Approximate methods for computing the "lookup" degrees of freedom are not currently supported. The use of any of the MODEL statement options ALL, CLM or T or any SCORE statement CLM option implicitly selects the DFMETHOD=EXACT option.

DIRECT
specifies that local least squares fits are to be done at every point in the input data set. When the direct option is not specified, a computationally faster method is used. This faster method performs local fitting at vertices of a kd tree decomposition of the predictor space followed by blending of the local polynomials to obtain a regression surface.

DROPSQUARE=(*variables***)**
specifies the quadratic monomials to exclude from the local quadratic fits. This option is ignored unless the DEGREE=2 option has been specified. For example,

```
model z=x y / degree=2 dropsquare=(y)
```

uses the monomials 1, x, y, x^2, and xy in performing the local fitting.

ITERATIONS=*number*
specifies the number of iterative reweighting steps to be done. Such iterations are appropriate when there are outliers in the data or when the error distribution is a symmetric long-tailed distribution. The default number of iterations is 1.

RESIDUAL | R
specifies that residuals are to be included in the "Output Statistics" table.

SCALE= NONE | SD $<$ **(***number***)** $>$
specifies the scaling method to be applied to scale the regressors. The default is NONE, in which case no scaling is applied. A specification of SD(*number*) indicates that a trimmed standard deviation is to be used as a measure of scale, where *number* is the trimming fraction. A specification of SD with no qualification defaults to 10% trimmed standard deviation.

SCALEDINDEP
specifies that scaled regressor coordinates be included in the output tables. This option is ignored if the SCALE= model option is not used or if SCALE=NONE is specified.

SMOOTH=*value-list*
specifies a list of positive smoothing parameter values. A separate fit is obtained for each smoothing value specified.

STD
specifies that standardized errors are to be included in the "Output Statistics" table.

T
specifies that t statistics are to be included in the "Output Statistics" table.

SCORE Statement

SCORE *<DATA=SAS-data-set> < ID=(variable list) >* $<$ */ options* $>$ **;**

The fitted loess model is used to score the data in the specified SAS data set. This data set must contain all the regressor variables specified in the MODEL statement. Furthermore, when a BY statement is used, the score data set must also contain all the BY variables sorted in the order of the BY variables. A SCORE statement is optional, and more than one SCORE statement can be used. SCORE statements cannot be used if the DIRECT option is specified in the MODEL statement. The optional ID=(*variable list*) specifies ID variables to be included in the "Score Results" table.

You find the results of the SCORE statement in the "Score Results" table. This table contains all the data in the data set named in the SCORE statement, including observations with missing values. However, only those observations with nonmissing regressor variables are scored. If no data set is named in the SCORE statement, the data set named in the PROC LOESS statement is scored. You use the PRINT option in the SCORE statement to request that the "Score Results" table be displayed. You can place the "Score Results" table in an output data set using an ODS OUTPUT statement even if this table is not displayed.

The following options are available in the SCORE statement after a slash (/).

CLM
requests that $100(1-\alpha)$ confidence limits on the mean predicted value be added to the "Score Results" table. By default the 95% limits are computed; the ALPHA= option in the MODEL statement can be used to change the α-level. The use of this option implicitly selects the model option DFMETHOD=EXACT if the DFMETHOD= option has not been explicitly used.

PRINT < (*variables* **)>**
specifies that the "Score Results" table is to be displayed. By default only the variables named in the MODEL statement, the variables listed in the ID list in the SCORE statement, and the scored dependent variables are displayed. The optional list in the PRINT option specifies additional variables in the score data set that are to be included in the displayed output. Note however that all columns in the SCORE data set are placed in the SCORE results table, even if you do not request that they be included in the displayed output.

WEIGHT Statement

WEIGHT *variable* **;**

The WEIGHT statement specifies a variable in the input data set that contains values to be used as a priori weights for a loess fit.

The values of the weight variable must be nonnegative. If an observation's weight is zero, negative, or missing, the observation is deleted from the analysis.

Details

Missing Values

PROC LOESS deletes any observation with missing values for any variable specified in the MODEL statement. This enables the procedure to reuse the kd tree for all the dependent variables that appear in the MODEL statement. If you have multiple dependent variables with different missing value structures for the same set of independent variables, you may want to use separate PROC LOESS steps for each dependent variable.

Output Data Sets

PROC LOESS assigns a name to each table it creates. You can use the ODS OUTPUT statement to place one or more of these tables in output data sets. See the section "ODS Table Names" on page 1877 for a list of the table names created by PROC LOESS. For detailed information on ODS, see Chapter 15, "Using the Output Delivery System."

For example, the following statements create an output data set named MyOutStats containing the OutputStatistics table and an output data set named MySummary containing the FitSummary table.

```
proc loess data=Melanoma;
   model Incidences=Year;
   ods output OutputStatistics = MyOutStats
              FitSummary       = MySummary;
run;
```

Often, a single MODEL statement describes more than one model. For example, the following statements fit eight different models (4 smoothing parameters for each dependent variable).

```
proc loess data=notReal;
   model y1 y2 = x1 x2 x3/smooth =0.1 to 0.7 by 0.2;
   ods output OutputStatistics = MyOutStats;
run;
```

The eight OutputStatistics tables for these models are stacked in a single data set called MyOutStats. The data set contains a column named DepVarName and a column named SmoothingParameter that distinguish each model (see Figure 38.4 on page 1860 for an example). If you want the OutputStatistics table for each model to be in its own data set, you can do so by using the MATCH_ALL option in the ODS OUTPUT statement. The following statements create eight data sets named MyOutStats, MyOutStats1, ... , MyOutStats7.

```
proc loess data=notReal;
   model y1 y2 = x1 x2 x3/smooth =0.1 to 0.7 by 0.2;
   ods output OutputStatistics(match_all) = MyOutStats;
run;
```

For further options available in the ODS OUTPUT statement, see Chapter 15, "Using the Output Delivery System."

Only the ScaleDetails and FitSummary tables are displayed by default. The other tables are optionally displayed by using the DETAILS option in the MODEL statement and the PRINT option in the SCORE statement. Note that it is not necessary to display a table in order for that table to be used in an ODS OUTPUT statement. For example, the following statements display the OutputStatistics and kdTree tables but place the OutputStatistics and PredAtVertices tables in output data sets.

```
proc loess data=Melanoma;
   model Incidences=Year/details(OutputStatistics kdTree);
   ods output OutputStatistics = MyOutStats
              PredAtVertices   = MyVerticesOut;
run;
```

Using the DETAILS option alone causes all tables to be displayed.

The MODEL statement options CLM, RESIDUAL, STD, SCALEDINDEP, and T control which optional columns are added to the OutputStatistics table. For example, to obtain an OutputStatistics output data set containing residuals and confidence limits in addition to the model variables and predicted value, you need to specify the RESIDUAL and CLM options in the MODEL statement as in the following example:

```
proc loess data=Melanoma;
   model Incidences=Year/residual clm;
   ods output OutputStatistics = MyOutStats;
run;
```

Finally, note that the ALL option in the MODEL statement includes all optional columns in the output. Also, ID columns can be added to the OutputStatistics table by using the ID statement.

Data Scaling

The loess algorithm to obtain a predicted value at a given point in the predictor space proceeds by doing a least squares fit using all data points that are close to the given point. Thus the algorithm depends critically on the metric used to define closeness. This has the consequence that if you have more than one predictor variable and these predictor variables have significantly different scales, then closeness depends almost entirely on the variable with the largest scaling. It also means that merely changing the units of one of your predictors can significantly change the loess model fit.

To circumvent this problem, it is necessary to standardize the scale of the independent variables in the loess model. The SCALE= option in the MODEL statement is provided for this purpose. PROC LOESS uses a symmetrically trimmed standard deviation as the scale estimate for each independent variable of the loess model. This is a robust scale estimator in that extreme values of a variable are discarded before estimating the data scaling. For example, to compute a 10% trimmed standard deviation of a sample, you discard the smallest and largest 5% of the data and compute the standard deviation of the remaining 90% of the data points. In this case, the trimming fraction is 0.1.

For example, the following statements specify that the variables Temperature and Catalyst are scaled before performing the loess fitting. In this case, because the trimming fraction is 0.1, the scale estimate used for each of these variables is a 10% trimmed standard deviation.

```
model Yield=Temperature Catalyst / scale = SD(0.1);
```

The default trimming fraction used by PROC LOESS is 0.1 and need not be specified by the SCALE= option. Thus the following MODEL statement is equivalent to the previous MODEL statement.

```
model Yield=Temperature Catalyst / scale = SD;
```

If the SCALE= option is not specified, no scaling of the independent variables is done. This is appropriate when there is only a single independent variable or when all the independent variables are a priori scaled similarly.

When the SCALE= option is specified, the scaling details for each independent variable are added to the ScaleDetails table (see Output 38.3.2 on page 1892 for an example). By default, this table contains only the minimum and maximum values of each independent variable in the model. Finally, note that when the SCALE= option is used, specifying the SCALEDINDEP option in the MODEL statement adds the scaled values of the independent variables to the OutputStatistics, PredAtVertices, and ScoreResults tables. By default, only the unscaled values are placed in these tables.

Direct versus Interpolated Fitting

Local regression to obtain a predicted value at a given point in the predictor space is done by doing a least squares fit using all data points in a local neighborhood of the given point. This method is computationally expensive because a local neighborhood must be determined and a least squares problem solved for each point at which a fitted value is required. A faster method is to obtain such fits at a representative sample of points in the predictor space and to obtain fitted values at all other points by interpolation.

PROC LOESS can fit models using either of these two paradigms. By default, PROC LOESS uses fitting at a sample of points and interpolation. The method fitting a local model at every data point is selected by specifying the DIRECT option in the MODEL statement.

kd Trees and Blending

PROC LOESS uses a kd tree to divide the box (also called the *initial cell* or *bucket*) enclosing all the predictor data points into rectangular cells. The vertices of these cells are the points at which local least squares fitting is done.

Starting from the initial cell, the direction of the longest cell edge is selected as the split direction. The median of this coordinate of the data in the cell is the split value. The data in the starting cell are partitioned into two child cells. The left child consists of all data from the parent cell whose coordinate in the split direction is less than the split value. The above procedure is repeated for each child cell that has more than a prespecified number of points, called the *bucket size* of the kd tree.

The value of the bucket size used by PROC LOESS can be specified using the BUCKET= option in the MODEL statement. If the BUCKET= option is not specified, the default value used is

$$\text{floor}\left(\frac{ns}{5}\right)$$

where n is the number of observations and s is the smoothing parameter. Note that if fitting is being done for a range of smoothing parameters, the bucket size may change for each smoothing parameter.

The set of vertices of all the cells of the kd tree are the points at which PROC LOESS performs its local fitting. The fitted value at an original data point (or at any other point within the original data cell) is obtained by blending the fitted values at the vertices of the kd tree cell that contains that data point. Currently, PROC LOESS uses linear interpolation from the enclosing kd tree cell vertex values. Future releases of PROC LOESS will incorporate higher-order blending methods.

While the details of the kd tree and the fitted values at the vertices of the kd tree are implementation details that seldom need to be examined, PROC LOESS does provide options for their display. Each kd tree subdivision of the data used by PROC LOESS is placed in a kdTree table. The predicted values at the vertices of each kd tree are placed in a PredAtVertices table. These tables can be optionally displayed or placed

in output data sets (as described in the section "Output Data Sets" on page 1871), or both.

Local Weighting

The size of the local neighborhoods that PROC LOESS uses in performing local fitting is determined by the smoothing parameter s. When $s < 1$, the local neighborhood used at a point x contains the s fraction of the data points closest to the point x. When $s \geq 1$, all data points are used.

Suppose q denotes the number of points in the local neighborhoods and $d_1, d_2, \ldots, d_q$ denote the distances in increasing order of the q points closest to x. The point at distance d_i from x is given a weight w_i in the local regression that decreases as the distance from x increases. PROC LOESS uses a tricube weight function to define

$$w_i = \frac{32}{5}\left(1 - (\frac{d_i}{d_q})^3\right)^3$$

If $s > 1$, then d_q is replaced by $d_q s^{1/p}$ in the previous formula, where p is the number of predictors in the model.

Finally, note that if a weight variable has been specified using a WEIGHT statement, then w_i is multiplied by the corresponding value of the specified weight variable.

Iterative Reweighting

PROC LOESS can do iterative reweighting to improve the robustness of the fit in the presence of outliers in the data. Iterative reweighting is also appropriate when statistical inference is requested and the error distribution is symmetric but not Gaussian.

The number of iterations is specified by the ITERATIONS= option in the MODEL statement. The default is ITERATIONS=1, which corresponds to no reweighting.

At iterations beyond the first iteration, the local weights w_i of the previous section are replaced by $r_i w_i$ where r_i is a weight that decreases as the residual of the fitted value at the previous iteration at the point corresponding to d_i increases. Refer to Cleveland and Grosse (1991) and Cleveland, Grosse, and Shyu (1992) for details.

Specifying the Local Polynomials

PROC LOESS uses linear or quadratic polynomials in doing the local least squares fitting. The option DEGREE = in the MODEL statement is used to specify the degree of the local polynomials used by PROC LOESS, with DEGREE = 1 being the default. In addition, when DEGREE = 2 is specified, the MODEL statement DROPSQUARE= option can be used to exclude specific monomials during the least squares fitting.

For example, the following statements use the monomials 1, x1, x2, x1*x2, and x2*x2 for the local least squares fitting.

```
proc loess data=notReal;
   model y= x1 x2/ degree=2 dropsquare=(x1);
run;
```

Statistical Inference

If you denote the ith measurement of the response by y_i and the corresponding measurement of predictors by x_i, then

$$y_i = g(x_i) + \epsilon_i$$

where g is the regression function and ϵ_i are independent random errors with mean zero. If the errors are normally distributed with constant variance, then you can obtain confidence intervals for the predictions from PROC LOESS. You can also obtain confidence limits in the case where ϵ_i is heteroscedastic but $a_i\epsilon_i$ has constant variance and a_i are a priori weights that are specified using the WEIGHT statement of PROC LOESS. You can do inference in the case in which the error distribution is symmetric by using iterative reweighting.

Formulae for doing statistical inference under the preceding conditions can be found in Cleveland and Grosse (1991) and Cleveland, Grosse, and Shyu (1992). The main result of their analysis is that a standardized residual for a loess model follows a t distribution with ρ degrees of freedom, where ρ is called the "lookup degrees of freedom." ρ is a function of the smoothing matrix L, which defines the linear relationship between the fitted and observed dependent variable values of a loess model.

The determination of ρ is computationally expensive and is not done by default. It is computed if you specify the DFMETHOD=EXACT option in the MODEL statement. It is also computed if you specify any of the options CLM, STD, or T in the MODEL statement.

If you specify the CLM option in the MODEL statement, confidence limits are added to the OutputStatistics table. By default, 95% limits are computed, but you can change this by using the ALPHA= option in the MODEL statement.

Scoring Data Sets

One or more SCORE statements can be used with PROC LOESS. A data set that includes all the variables specified in the MODEL and BY statements must be specified in each SCORE statement. Score results are placed in the ScoreResults table. This table is not displayed by default, but specifying the PRINT option in the SCORE statement produces this table. If you specify the CLM option in the SCORE statement, confidence intervals are included in the ScoreResults table.

Note that scoring is not supported when the DIRECT option is specified in the MODEL statement. Scoring at a point specified in a score data set is done by first finding the cell in the kd tree containing this point and then interpolating the scored

value from the predicted values at the vertices of this cell. This methodology precludes scoring any points that are not contained in the box that surrounds the data used in fitting the loess model.

ODS Table Names

PROC LOESS assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in the following table. For more information on ODS, see Chapter 15, "Using the Output Delivery System."

Table 38.2. ODS Tables Produced by PROC LOESS

ODS Table Name	Description	Statement	Option
FitSummary	Specified fit parameters and fit summary		default
kdTree	Structure of kd tree used	MODEL	DETAILS(kdTree)
OutputStatistics	Coordinates and fit results at input data points	MODEL	DETAILS(OutputStatistics)
PredAtVertices	Coordinates and fitted values at kd tree vertices	MODEL	DETAILS(PredAtVertices)
ScaleDetails	Extent and scaling of the independent variables		default
ScoreResults	Coordinates and fit results at scoring points	SCORE	PRINT

Examples

Example 38.1. Engine Exhaust Emissions

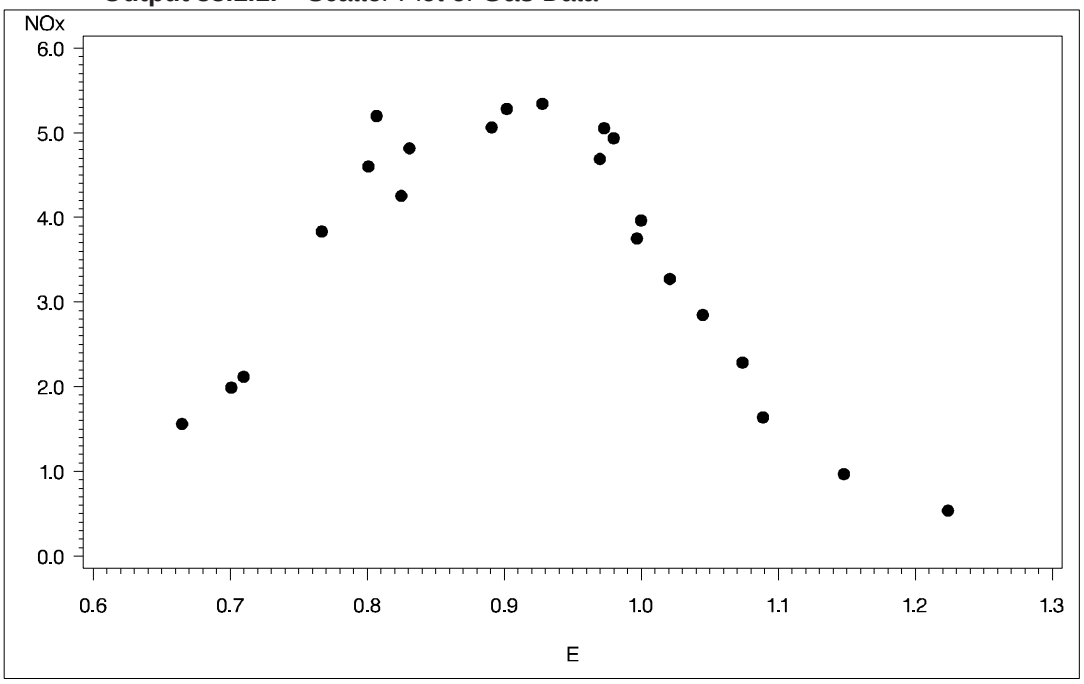
Investigators studied the exhaust emissions of a one cylinder engine (Brinkman 1981). The SAS data set Gas contains the results data. The dependent variable, NOx, measures the concentration, in micrograms per joule, of nitric oxide and nitrogen dioxide normalized by the amount of work of the engine. The independent variable, E, is a measure of the richness of the air and fuel mixture.

```
data Gas;
   input NOx E;
   format NOx f3.1;
   format E f3.1;
datalines;
4.818  0.831
2.849  1.045
3.275  1.021
4.691  0.97
4.255  0.825
5.064  0.891
2.118  0.71
4.602  0.801
2.286  1.074
0.97   1.148
3.965  1
5.344  0.928
3.834  0.767
1.99   0.701
5.199  0.807
5.283  0.902
3.752  0.997
0.537  1.224
1.64   1.089
5.055  0.973
4.937  0.98
1.561  0.665
;
```

The following PROC GPLOT statements produce the simple scatter plot of these data, displayed in Output 38.1.1.

```
symbol1 color=black value=dot ;
proc gplot data=Gas;
   plot NOx*E;
run;
```

Output 38.1.1. Scatter Plot of Gas Data



The following statements fit two loess models for these data. Because this is a small data set, it is reasonable to do direct fitting at every data point. As there is substantial curvature in the data, quadratic local polynomials are used. An ODS OUTPUT statement creates two output data sets containing the "Output Statistics" and "Fit Summary" tables.

```
proc loess data=Gas;
   ods output OutputStatistics = GasFit
              FitSummary=Summary;
   model NOx = E / degree=2 direct smooth = 0.6 1.0
                   alpha=.01 all details;
run;
```

The "Fit Summary" table for smoothing parameter 0.6, shown in Output 38.1.2, records the fitting parameters specified and some overall fit statistics.

Output 38.1.2. Fit Summary Table

The LOESS Procedure
Smoothing Parameter: 0.6
Dependent Variable: NOx

Fit Summary

Fit Method	Direct
Number of Observations	22
Degree of Local Polynomials	2
Smoothing Parameter	0.60000
Points in Local Neighborhood	13
Residual Sum of Squares	1.71852
Trace[L]	6.42184
Delta1	15.12582
Delta2	14.73089
Equivalent Number of Parameters	5.96950
Lookup Degrees of Freedom	15.53133
Residual Standard Error	0.33707

The matrix L referenced in the "Fit Summary" table is the smoothing matrix. This matrix satisfies

$$\hat{y} = Ly$$

where y is the vector of observed values and $\hat{y}$ is the corresponding vector of predicted values of the dependent variable. The quantities

$$\begin{aligned} \delta_1 &\equiv \text{Trace}(I-L)^T(I-L) \\ \delta_2 &\equiv \text{Trace}\left((I-L)^T(I-L)\right)^2 \\ \rho &\equiv \text{Lookup Degrees of Freedom} \\ &\equiv \delta_1^2/\delta_2 \end{aligned}$$

in the "Fit Summary" table are used in doing statistical inference.

The equivalent number of parameters and residual standard error in the "Fit Summary" table are defined by

$$\begin{aligned} \text{Equivalent Number of Parameters} &\equiv \text{Trace}L^TL \\ \text{Residual Standard Error} &\equiv \sqrt{\text{Residual SS}/\delta_1} \end{aligned}$$

The "Output Statistics" table for smoothing parameter 0.6 is shown in Output 38.1.3. Note that, as the ALL option in the MODEL statement is specified, this table includes all the relevant optional columns. Furthermore, because the ALPHA=0.01 option is specified in the MODEL statement, the confidence limits in this table are 99% limits.

Output 38.1.3. Output Statistics Table

The LOESS Procedure
Smoothing Parameter: 0.6
Dependent Variable: NOx

Output Statistics

Obs	E	NOx	Predicted NOx	Estimated Prediction Std Deviation	Residual	t Value
1	0.8	4.8	4.87377	0.15528	-0.05577	-0.36
2	1.0	2.8	2.81984	0.15380	0.02916	0.19
3	1.0	3.3	3.48153	0.15187	-0.20653	-1.36
4	1.0	4.7	4.73249	0.13923	-0.04149	-0.30
5	0.8	4.3	4.82305	0.15278	-0.56805	-3.72
6	0.9	5.1	5.18561	0.19337	-0.12161	-0.63
7	0.7	2.1	2.51120	0.15528	-0.39320	-2.53
8	0.8	4.6	4.48267	0.15285	0.11933	0.78
9	1.1	2.3	2.12619	0.16683	0.15981	0.96
10	1.1	1.0	0.97120	0.18134	-0.00120	-0.01
11	1.0	4.0	4.09987	0.13477	-0.13487	-1.00
12	0.9	5.3	5.31258	0.17283	0.03142	0.18
13	0.8	3.8	3.84572	0.14929	-0.01172	-0.08
14	0.7	2.0	2.26578	0.16712	-0.27578	-1.65
15	0.8	5.2	4.58394	0.15363	0.61506	4.00
16	0.9	5.3	5.24741	0.19319	0.03559	0.18
17	1.0	3.8	4.16979	0.13478	-0.41779	-3.10
18	1.2	0.5	0.53059	0.32170	0.00641	0.02
19	1.1	1.6	1.83157	0.17127	-0.19157	-1.12
20	1.0	5.1	4.66733	0.13735	0.38767	2.82
21	1.0	4.9	4.52385	0.13556	0.41315	3.05
22	0.7	1.6	1.19888	0.26774	0.36212	1.35

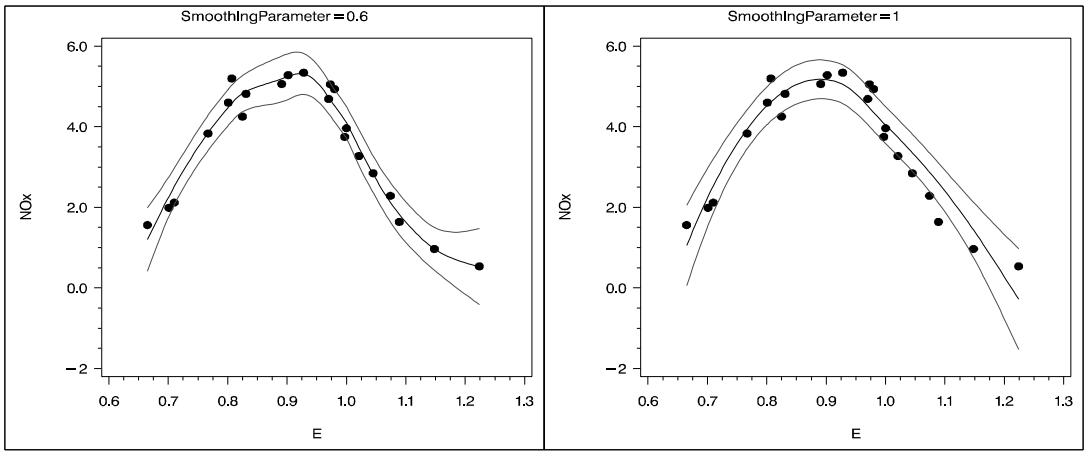
Output Statistics

Obs	99% Confidence Limits	
1	4.41841	5.32912
2	2.36883	3.27085
3	3.03617	3.92689
4	4.32419	5.14079
5	4.37503	5.27107
6	4.61855	5.75266
7	2.05585	2.96655
8	4.03444	4.93089
9	1.63697	2.61541
10	0.43942	1.50298
11	3.70467	4.49507
12	4.80576	5.81940
13	3.40794	4.28350
14	1.77571	2.75584
15	4.13342	5.03445
16	4.68089	5.81393
17	3.77457	4.56502
18	-0.41278	1.47397
19	1.32933	2.33380
20	4.26456	5.07010
21	4.12632	4.92139
22	0.41375	1.98401

Plots of the data points and fitted models with 99% confidence bands are shown in Output 38.1.4.

```
proc sort data=GasFit;
   by SmoothingParameter E;
run;

symbol1 color=black value=dot ;
symbol2 color=black interpol=spline value=none;
symbol3 color=green interpol=spline value=none;
symbol4 color=green interpol=spline value=none;

%let opts=vaxis=axis1 hm=3 vm=3 overlay;

goptions nodisplay hsize=3.75;
axis1 label=(angle=90 rotate=0);

proc gplot data=GasFit;
   by SmoothingParameter;
   plot (DepVar Pred LowerCL UpperCL)*E/ &opts name='fitGas';
run; quit;

goptions display hsize=0 hpos=0;
proc greplay nofs tc=sashelp.templt template=h2;
    igout gseg;
    treplay 1:fitGas 2:fitGas1;
run; quit;
```

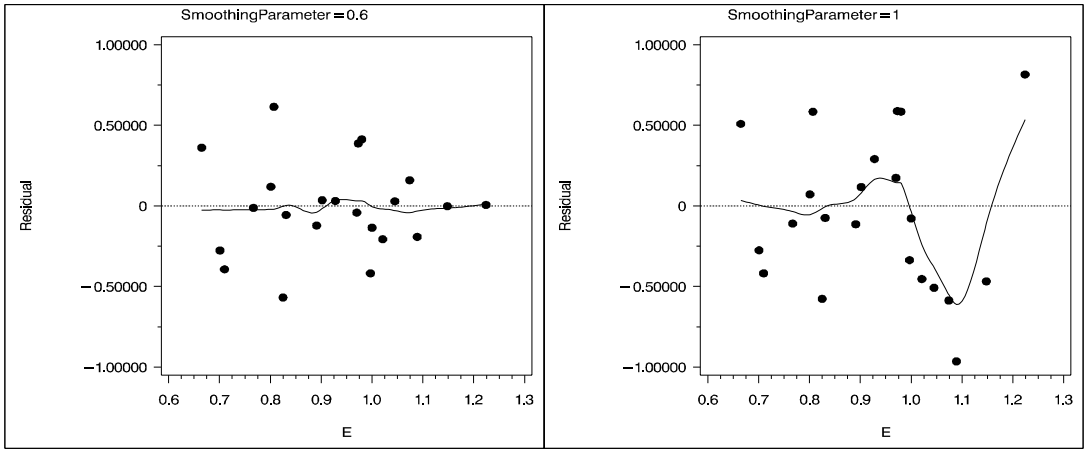
Output 38.1.4. Loess Fits with 99% Confidence Bands for Gas Data



It is evident from the preceding figure that the better fit is obtained with smoothing parameter 0.6. Scatter plots of the fit residuals confirm this observation. Note also that PROC LOESS is again used to produce the Residual variable on these plots.

```
proc loess data=GasFit;
   by SmoothingParameter;
   ods output OutputStatistics=residout;
   model Residual=E;
run;

axis1 label = (angle=90 rotate=0)
      order = (-1 to 1 by 0.5);
goptions nodisplay hsize=3.75;
proc gplot data=residout;
   by SmoothingParameter;
   plot  DepVar*E Pred*E/ &opts vref=0 lv=2 vm=1
                          name='resGas';
run; quit;

goptions display hsize=0 hpos=0;
proc greplay nofs tc=sashelp.templt template=h2;
    igout gseg;
    treplay 1:resGas 2:resGas1;
run; quit;
```

Output 38.1.5. Scatter Plots of Loess Fit Residuals



The residual plots show that with smoothing parameter 1, the loess model exhibits a lack of fit. Analysis of variance can be used to compare the model with smoothing parameter 1, which serves as the null model, to the model with smoothing parameter 0.6.

The statistic

$$F = \frac{(\text{rss}^{(n)} - \text{rss})/(\delta_1^{(n)} - \delta_1)}{\text{rss}/\delta_1}$$

has a distribution that is well approximated by an F distribution with

$$\nu = \frac{(\delta_1^{(n)} - \delta_1)^2}{\delta_2^{(n)} - \delta_2}$$

numerator degrees of freedom and ρ denominator degrees of freedom (Cleveland and Grosse 1991). Here quantities with superscript n refer to the null model, rss is the residual sum of squares, and δ_1, δ_2, and ρ are as previously defined.

The "Fit Summary" tables contain the information needed to carry out such an analysis. These tables have been captured in the output data set named Summary using an ODS OUTPUT statement. The following statements extract the relevant information from this data set and carry out the analysis of variance:

```
data h0 h1;
  set Summary(keep=SmoothingParameter Label1 nValue1
              where=(Label1 in ('Residual Sum of Squares',
                     'Delta1',
                     'Delta2',
                     'Lookup Degrees of Freedom')));
  if SmoothingParameter = 1 then output h0;
  else output h1;
run;

proc transpose data=h0(drop=SmoothingParameter Label1)
               out=h0;

data h0(drop=_NAME_); set h0;
  rename Col1 = RSSNull
         Col2 = delta1Null
         Col3 = delta2Null;
```

```
proc transpose data=h1(drop=SmoothingParameter Label1)
               out=h1;

data h1(drop=_NAME_); set h1;
  rename Col1 = RSS
         Col2 = delta1
         Col3 = delta2
         Col4 = rho;

data ftest; merge h0 h1;
  nu = (delta1Null - delta1)**2 / (delta2Null - delta2);
  Numerator = (RSSNull - RSS)/(delta1Null - delta1);
  Denominator = RSS/delta1;
  FValue = Numerator / Denominator;
  PValue = 1 - ProbF(FValue, nu, rho);
  label nu = 'Num DF'
        rho = 'Den DF'
        FValue = 'F Value'
        PValue = 'Pr > F';

proc print data=ftest label;
  var nu rho Numerator Denominator FValue PValue;
  format nu rho FValue 7.2 PValue 6.4;
run;
```

The results are shown in Output 38.1.6.

Output 38.1.6. Test ANOVA for LOESS MODELS of Gas Data

Obs	Num DF	Den DF	Numerator	Denominator	F Value	Pr > F
1	2.67	15.53	1.05946	0.11362	9.32	0.0012

The highly significant p-value confirms that the loess model with smoothing parameter 0.6 provides a better fit than the model with smoothing parameter 1.

Example 38.2. Sulfate Deposits in the USA for 1990

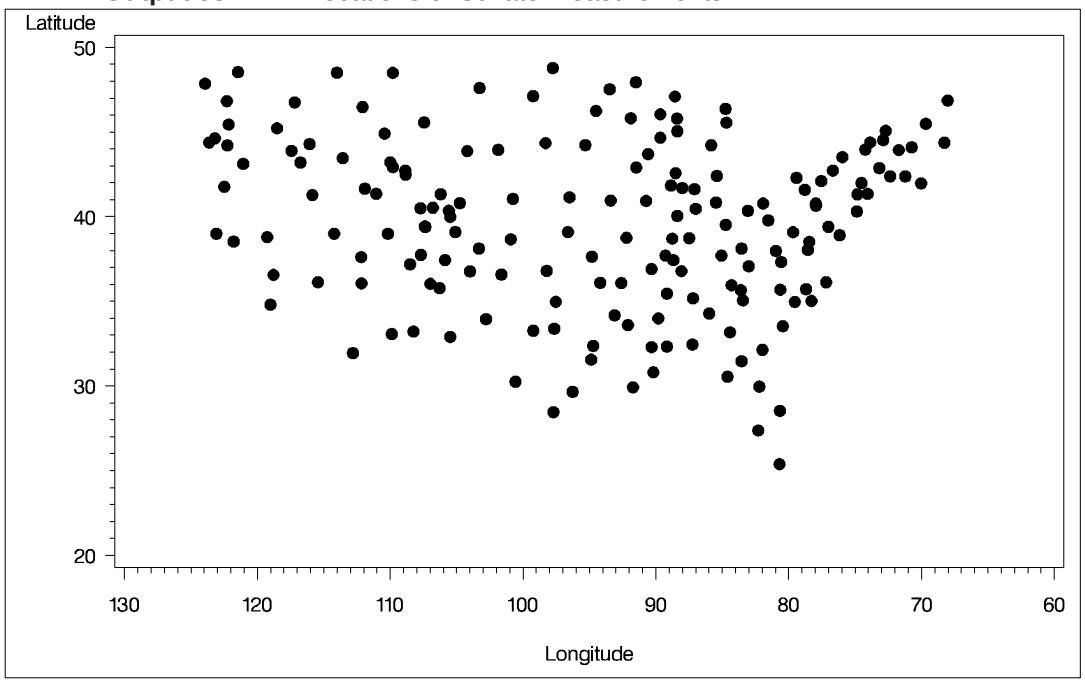
The following data set contains measurements in grams per square meter of sulfate (SO_4) deposits during 1990 at 179 sites throughout the 48 states.

```
data SO4;
   input Latitude Longitude SO4 @@;
datalines;
32.45833  87.24222 1.403 34.28778  85.96889 2.103
33.07139 109.86472 0.299 36.07167 112.15500 0.304
31.95056 112.80000 0.263 33.60500  92.09722 1.950
.
.      more data lines
.
42.92889 109.78667 0.182 43.22278 109.99111 0.161
43.87333 104.19222 0.306 44.91722 110.42028 0.210
45.07611  72.67556 2.646
;
```

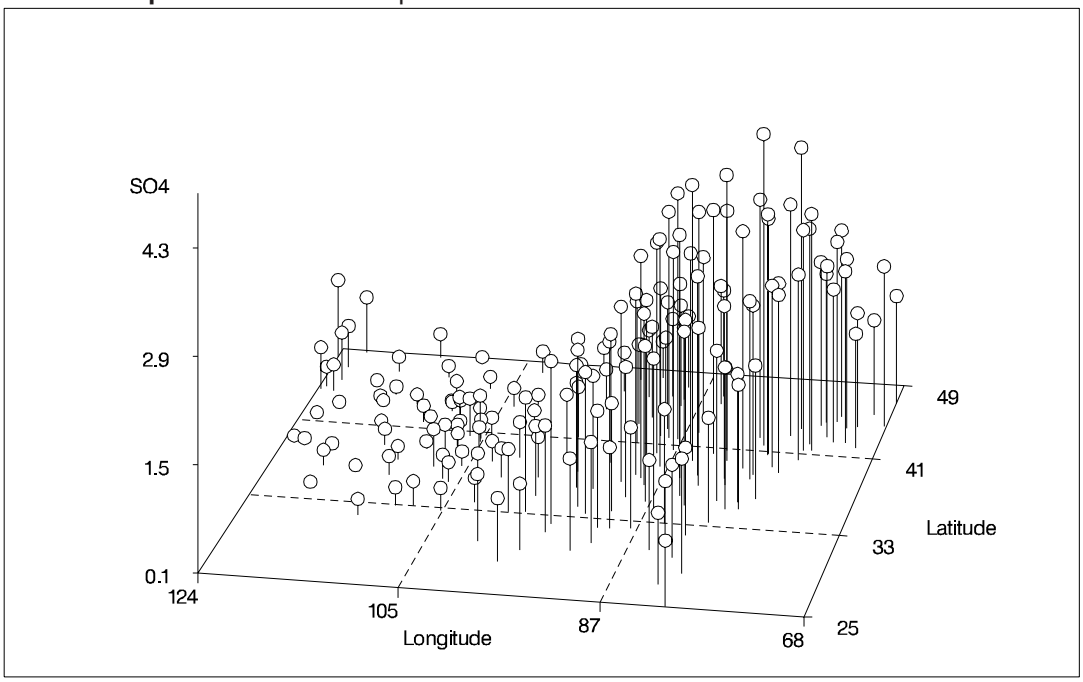
The following statements produce the two scatter plots of the SO4 data shown in Output 38.2.1 and Output 38.2.2:

```
symbol1 color=black value=dot ;
proc gplot data=SO4;
   plot Latitude*Longitude/hreverse;
run;

proc g3d data=SO4;
   format SO4 f4.1;
   scatter Longitude*Latitude=SO4 /
         shape='balloon'
         size=0.35
         rotate=80
         tilt=60;
run;
```

Output 38.2.1. Locations of Sulfate Measurements



Output 38.2.2. Scatter plot of SO4 Data



From these scatter plots, it is clear that the largest concentrations are in the northeastern United States. These plots also indicate that a nonparametric surface, such as a loess fit, is appropriate for these data.

The sulfate measurements are irregularly spaced. The following statements create a SAS data set containing a regular grid of points that will be used in the SCORE statement:

```
data PredPoints;
   do Latitude = 26 to 46 by 1;
      do Longitude = 79 to 123 by 1;
         output;
      end;
   end;
```

The following statements fit loess models for two values of the smoothing parameter and save the results in output data sets:

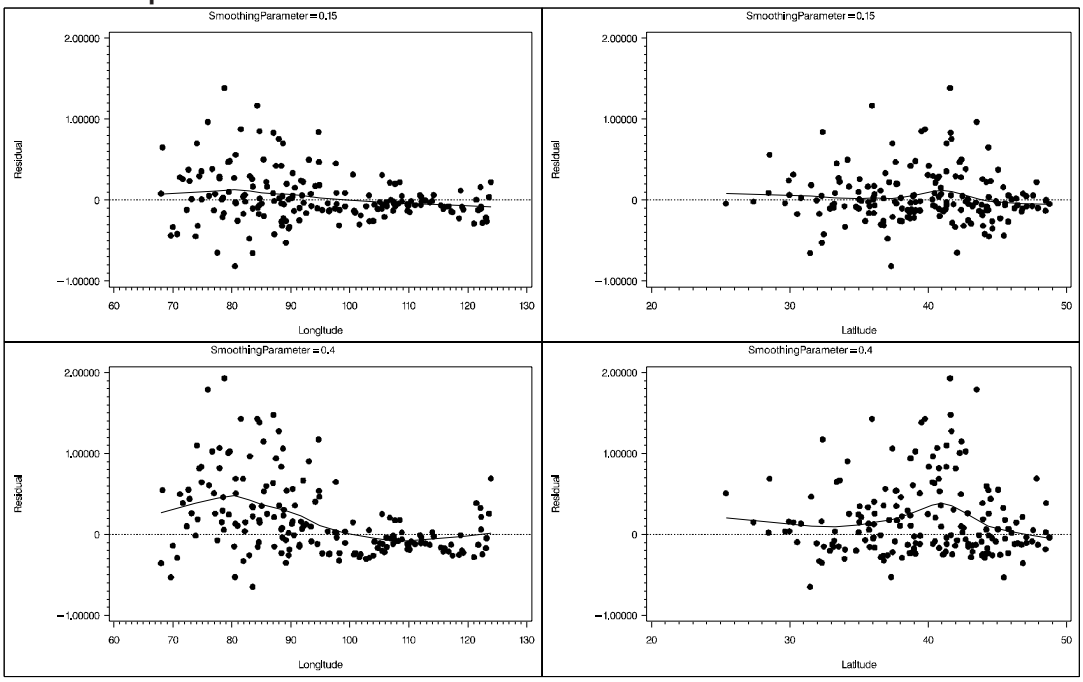
```
proc loess data=SO4;
   ods Output ScoreResults=ScoreOut
              OutputStatistics=StatOut;
   model SO4=Latitude Longitude/smooth=0.15 0.4 residual;
   score data=PredPoints;
run;
```

Notice that even though there are two predictors in the model, the SCALE= option is not appropriate because the predictors (Latitude and Longitude) are identically scaled.

Output 38.2.3 shows scatter plots of the fit residuals versus each of the predictors for the two smoothing parameters specified. A loess fit of the residuals is also shown on these scatter plots and is obtained using PROC LOESS with the StatOut data set generated by the previous PROC LOESS step.

```
proc loess data=StatOut;
   by SmoothingParameter;
   ods output OutputStatistics=ResidLatOut;
   model residual=Latitude;
run;
proc loess data=StatOut;
   by SmoothingParameter;
   ods output OutputStatistics=ResidLongOut;
   model residual=Longitude;
run;
proc sort data=ResidLatOut;
   by SmoothingParameter Latitude;
run;
proc sort data=ResidLongOut;
   by SmoothingParameter Longitude;
run;

goptions nodisplay;
symbol1 color=black value=dot ;
symbol2 color=black interpol=join value=none;
%let opts = vaxis=axis1 overlay vref=0 lv=2;
axis1 label = (angle=90 rotate=0);

proc gplot data=ResidLatOut;
   by smoothingParameter;
   plot (DepVar Pred) * Latitude / &opts name='lat';
run;

proc gplot data=ResidLongOut;
   by smoothingParameter;
   plot (DepVar Pred) * Longitude / &opts name='long';
run;

goptions display;
proc greplay nofs tc=sashelp.templt template=l2r2;
    igout gseg;
    treplay 1:long 2:long1 3:lat 4:lat1;
run; quit ;
```

Output 38.2.3. Scatter Plots of Loess Fit Residuals

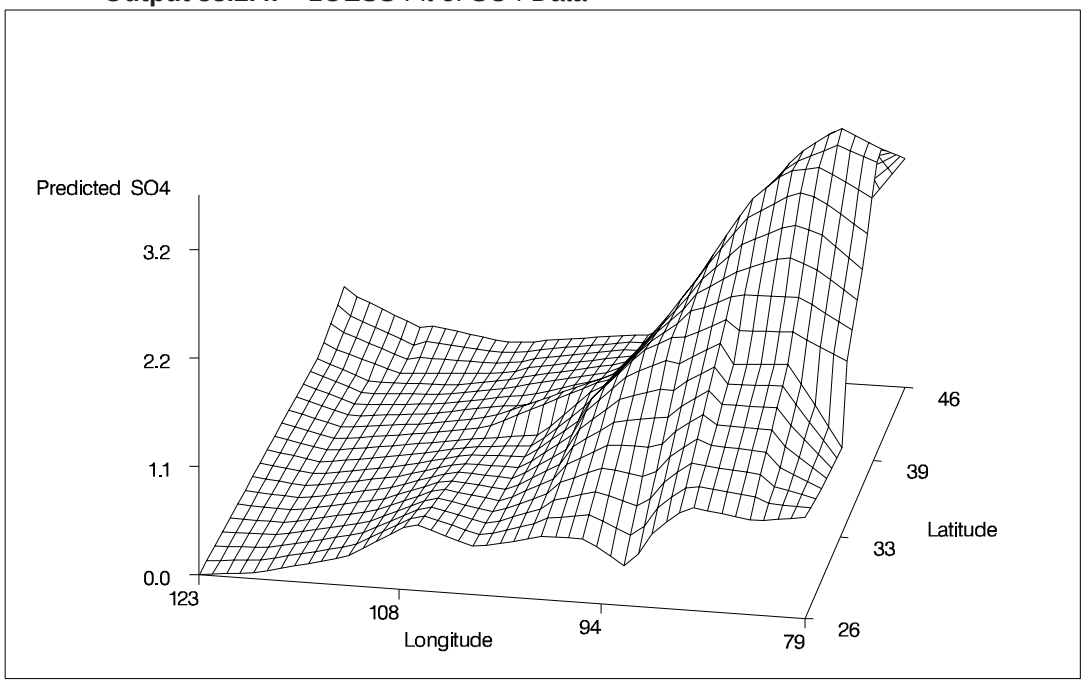


The scatter plots in Output 38.2.3 reveal that, with smoothing parameter 0.4, there is significant information in the data that is not being captured by the loess model. By contrast, the residuals for the more localized smoothing parameter 0.15 show a better fit.

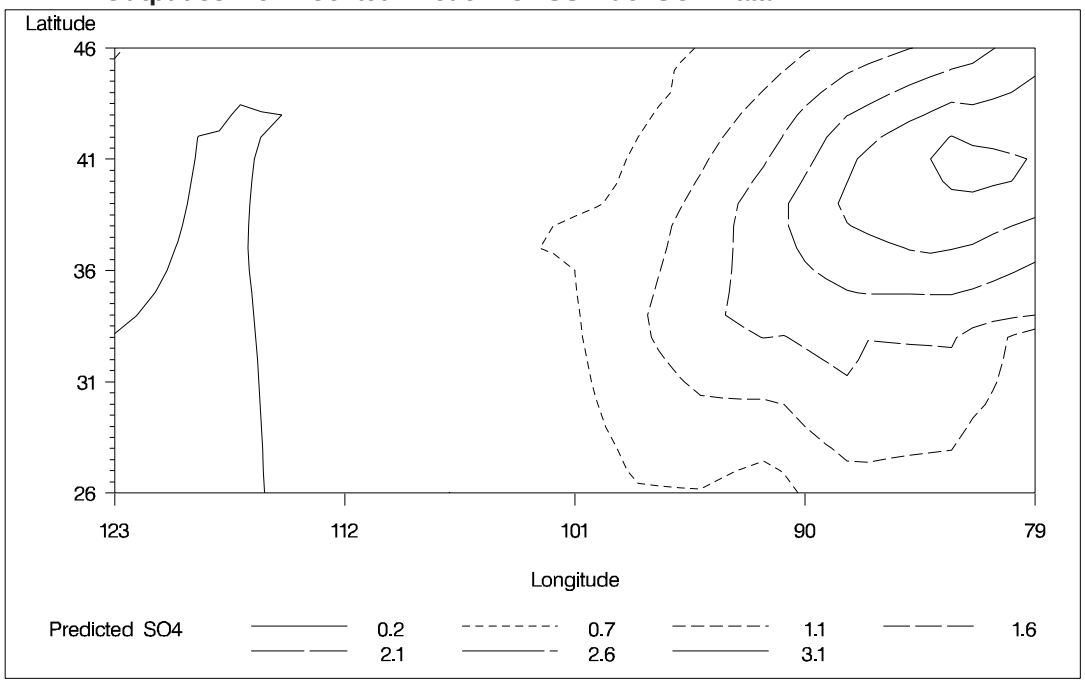
The ScoreOut data set contains the model predictions at the grid defined in the PredPoints data set. The following statements request a fitted surface and a contour plot of this surface with a smoothing parameter of 0.15:

```
proc g3d data=ScoreOut(where= (smoothingParameter=0.15));
   format Latitude f4.0;
   format Longitude f4.0;
   format p_SO4 f4.1;
   plot Longitude*Latitude=p_SO4/tilt=60 rotate=80;
run;

proc gcontour data=ScoreOut(where= (smoothingParameter=0.15));
   format latitude f4.0;
   format longitude f4.0;
   format p_SO4 f4.1;
   plot Latitude*Longitude = p_SO4/hreverse;
run;
```

Output 38.2.4. LOESS Fit of SO4 Data



Output 38.2.5. Contour Plot of LOESS Fit of SO4 Data



Example 38.3. Catalyst Experiment

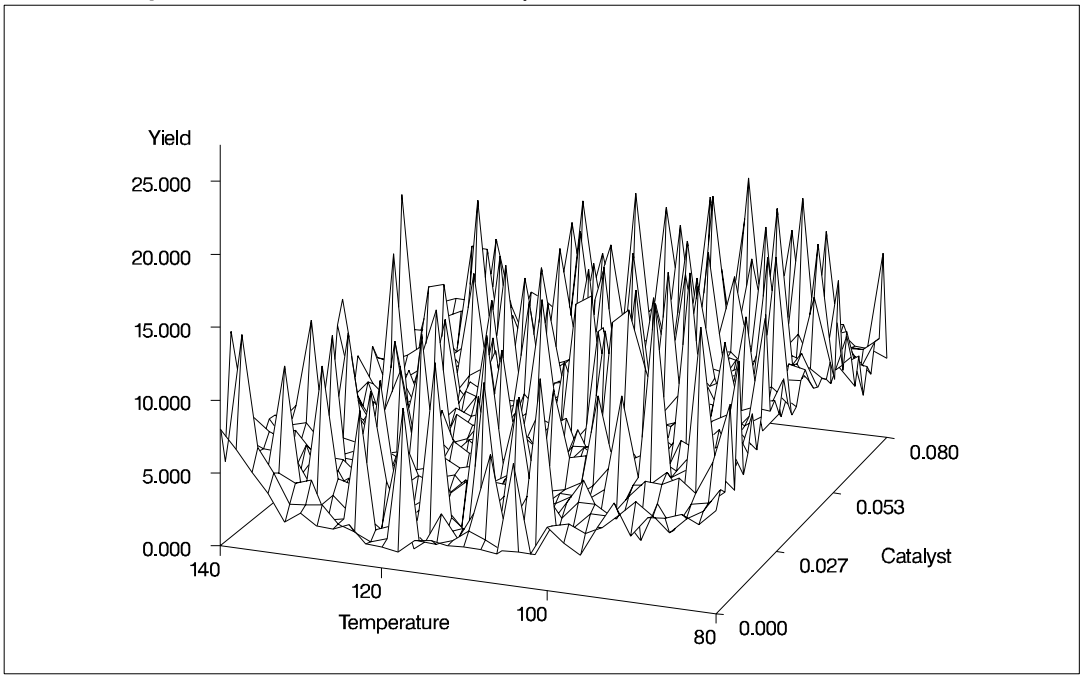
The following data set records the results of an experiment to determine how the yield of a chemical reaction varies with temperature and amount of a catalyst used.

```
data Experiment;
   input Temperature Catalyst Yield;
datalines;
  80      0.000      6.842
  80      0.002      7.944
  .
  .     more data lines
  .
 140      0.078      4.012
 140      0.080      5.212
;
```

Researchers know that about 10% of the yield measurements are corrupted due to an intermittent equipment problem. This can be seen in the surface plot of raw data shown in Output 38.3.1.

```
proc g3d data=Experiment;
   plot Temperature*Catalyst=Yield/zmin=0 zmax=25 zticknum=6;
run;
```

Output 38.3.1. Surface Plot of Experiment Data



A robust fitting method is needed to estimate the underlying surface in the presence of data outliers. The following statements invoke PROC LOESS with iterative reweighting to fit a surface to these data:

```
proc loess data=Experiment;
   ods output OutputStatistics=Results;
   model Yield = Temperature Catalyst /
                         scale=sd(0.1)
                         smooth=0.1
                         iterations=3;
run;
```

The ITERATIONS=3 option in the MODEL statement requests two iteratively reweighted iterations. Note the use of the SCALE=SD(0.1) option in the MODEL statement. This specifies that the independent variables in the model are to be divided by their respective 10% trimmed standard deviations before the fitted model is computed. This is appropriate as the independent variables Temperature and Catalyst are not similarly scaled. The "Scale Details" table produced by PROC LOESS is shown in Output 38.3.2.

Output 38.3.2. Scale Details Table

The LOESS Procedure

Independent Variable Scaling

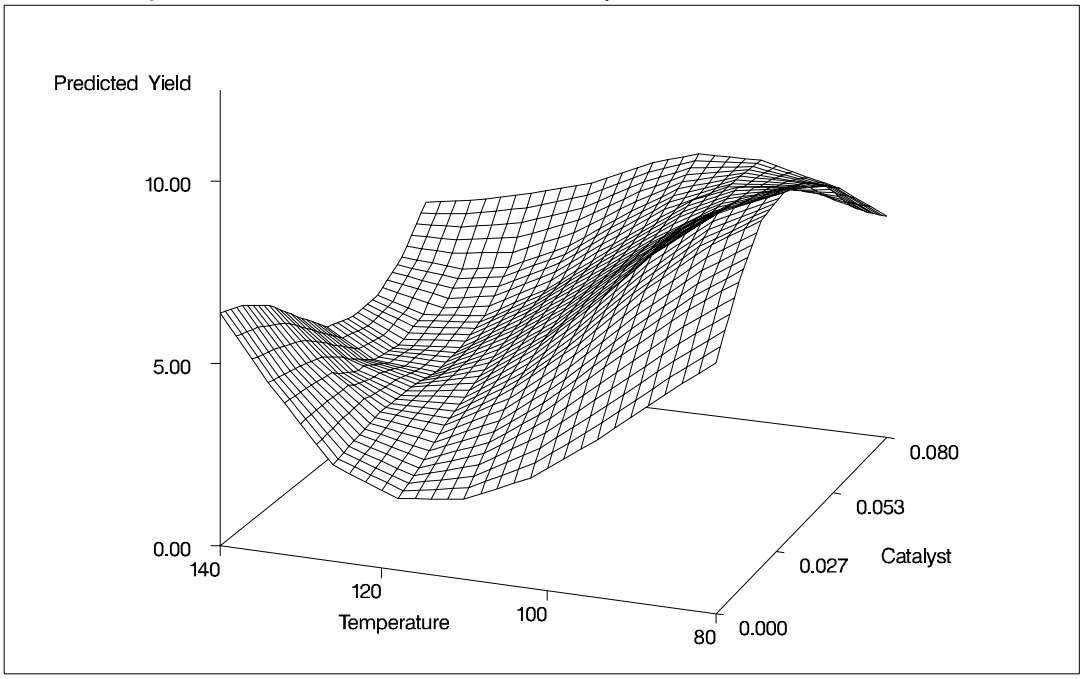
Scaling applied: 10% trimmed standard deviation

Statistic	Temperature	Catalyst
Minimum Value	80	0.000
Maximum Value	140	0.080
Trimmed Mean	110	0.040
Trimmed Standard Deviation	14	0.019

The following statements use the G3D procedure to plot the fitted surface shown in Output 38.3.3.

```
proc g3d data=Results;
   format Temperature f4.0;
   format Catalyst f6.3;
   format pred f5.2;
   plot Temperature*Catalyst=pred/zmin=0 zmax=10 zticknum=3;
run;
```

Output 38.3.3. Fitted Surface Plot for Experiment Data



Example 38.4. Automatic Smoothing Parameter Selection

The following data set contains measurements of monthly averaged atmospheric pressure differences between Easter Island and Darwin, Australia for a period of 168 months (NIST 1998).

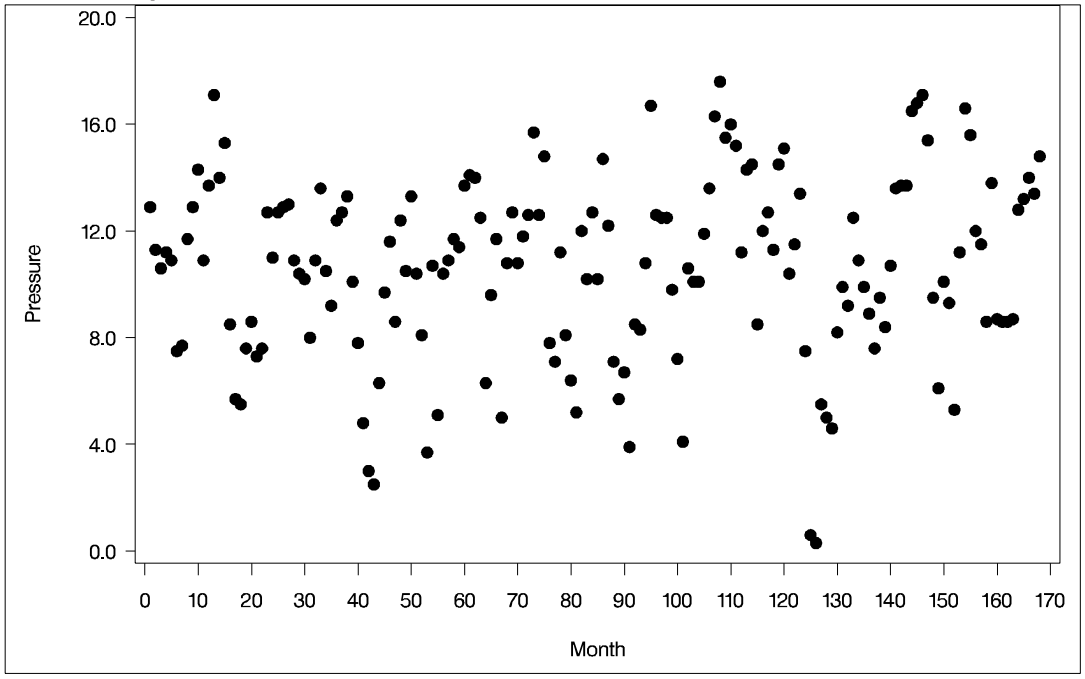
```
data ENSO;
  input Pressure @@;
  Month=_N_;
  format Pressure 4.1;
  format Month 3.0;
datalines;
12.9  11.3  10.6  11.2  10.9   7.5   7.7  11.7
12.9  14.3  10.9  13.7  17.1  14.0  15.3   8.5
 5.7   5.5   7.6   8.6   7.3   7.6  12.7  11.0
12.7  12.9  13.0  10.9  10.4  10.2   8.0  10.9
13.6  10.5   9.2  12.4  12.7  13.3  10.1   7.8
 4.8   3.0   2.5   6.3   9.7  11.6   8.6  12.4
10.5  13.3  10.4   8.1   3.7  10.7   5.1  10.4
10.9  11.7  11.4  13.7  14.1  14.0  12.5   6.3
 9.6  11.7   5.0  10.8  12.7  10.8  11.8  12.6
15.7  12.6  14.8   7.8   7.1  11.2   8.1   6.4
 5.2  12.0  10.2  12.7  10.2  14.7  12.2   7.1
 5.7   6.7   3.9   8.5   8.3  10.8  16.7  12.6
12.5  12.5   9.8   7.2   4.1  10.6  10.1  10.1
11.9  13.6  16.3  17.6  15.5  16.0  15.2  11.2
14.3  14.5   8.5  12.0  12.7  11.3  14.5  15.1
10.4  11.5  13.4   7.5   0.6   0.3   5.5   5.0
 4.6   8.2   9.9   9.2  12.5  10.9   9.9   8.9
```

```
 7.6   9.5   8.4  10.7  13.6  13.7  13.7  16.5
16.8  17.1  15.4   9.5   6.1  10.1   9.3   5.3
11.2  16.6  15.6  12.0  11.5   8.6  13.8   8.7
 8.6   8.6   8.7  12.8  13.2  14.0  13.4  14.8
;
```

The following PROC GPLOT statements produce the simple scatter plot of these data, displayed in Output 38.4.1.

```
symbol1 color=black value=dot ;
proc gplot data=ENSO;
   plot Pressure*Month /
        hminor = 0
        vminor = 0
        vaxis  = axis1
        frame;
        axis1 label = ( r=0 a=90 ) order=(0 to 20  by 4);;
run;
```

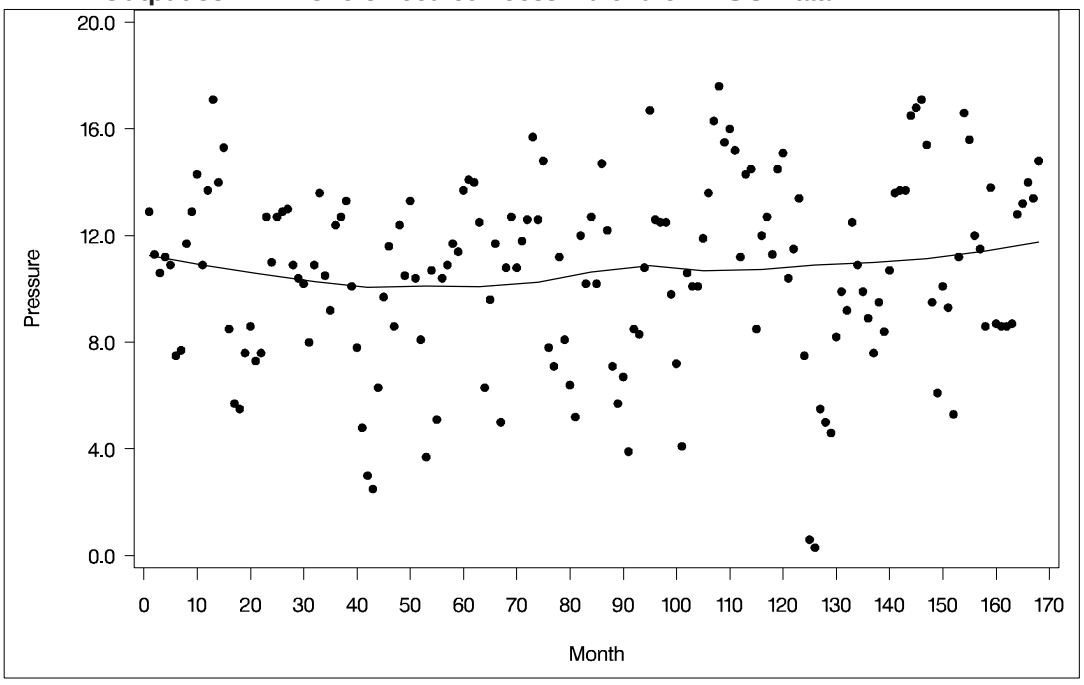
Output 38.4.1. Scatter Plot of ENSO Data



You can compute a loess fit and plot the results for these data using the following statements:

```
ods output OutputStatistics=ENSOstats;

proc loess data=ENSO;
   model Pressure=Month ;
run;

symbol1 color=black value=dot h=2.5 pct;
symbol2 color=black interpol=join value=none width=2;
proc gplot data=ENSOstats;
   plot (depvar pred)*Month / overlay
        hminor = 0
        vminor = 0
        vaxis  = axis1
        frame;
        axis1 label = ( r=0 a=90 ) order=(0 to 20  by 4);
run; quit;
```

Output 38.4.2. Oversmoothed Loess Fit for the ENSO Data



You see that the default smoothing parameter 0.5 is too large. There are several strategies that you can use to select the smoothing parameter. The strategy used in the previous examples is to examine plots of the fit residuals versus the predictor variable and to choose the largest smoothing parameter that yields no clearly discernible trends in the fit residuals.

Alternatively, a variety of automatic methods for choosing the smoothing parameter exist. Many of these methods choose a smoothing parameter that minimizes a criterion that incorporates both the tightness of the fit and model complexity of the form

$$\log(\hat{\sigma}^2) + \psi(L)$$

where $\hat{\sigma}^2$ is the average residual sum of squares and $\psi(\cdot)$ is a penalty function designed to decrease with increasing smoothness of the fit. Here L is the smoothing matrix of the method. This matrix satisfies

$$\hat{y} = Ly$$

where y is the vector of observed values and $\hat{y}$ is the corresponding vector of predicted values of the dependent variable. Examples of specific criteria obtained with this methodology are generalized cross validation (Craven and Wahba 1979), the Akaike information criterion (Akaike 1973), and the bias corrected Akaike information criteria (Hurvich and Simonoff 1998).

Hurvich and Simonoff (1998) show that the bias corrected Akaike information criteria avoid the tendency to undersmooth that often occurs when using the classical Akaike information criterion or generalized cross validation. One such criterion is given by

$$AIC_{C_1} = n\log(\hat{\sigma}^2) + n\frac{\delta_1/\delta_2(n+\nu_1)}{\delta_1^2/\delta_2 - 2}$$

where

$$\begin{aligned}
n &\equiv \text{Number of observations} \\
\delta_1 &\equiv \text{Trace}(I-L)^T(I-L) \\
\delta_2 &\equiv \text{Trace}\left((I-L)^T(I-L)\right)^2 \\
\nu_1 &\equiv \text{Equivalent number of parameters} \\
&\equiv \text{Trace}(L^TL)
\end{aligned}$$

The statistics that you use to compute AIC_{C_1} are also used to perform statistical inference in loess models. You can find these numbers in the "Fit Summary" table of PROC LOESS, provided that you specify at least one of the options ALL, CLM, DFMETHOD=EXACT, STD, and T in the MODEL statement. As with all tables of output, you can capture the "Fit Summary" table numbers in a SAS data set by using an ODS output. The following SAS macro takes such an output data set as an argument. It finds the smoothing parameter that yields the smallest AIC_{C_1} statistic for all the smoothing parameters that you specify in the MODEL statement.

```
%macro SmoothSelect(data);
options nonotes;
ods listing close;

data temp;
   set &data(keep = Label1 nValue1 smoothingParameter
      where=(Label1 in
            ('Number of Observations',
             'Residual Sum of Squares',
             'Trace[L]'
             'Equivalent Number of Parameters',
             'Delta1',
             'Delta2',
             'Lookup Degrees of Freedom')));
run;

proc transpose data=temp(drop=Label1) out=temp;
   by smoothingParameter;
run;

data temp(drop=_NAME_); set temp;
   rename Col1 = n
          Col2 = rss
          Col3 = traceL
          Col4 = delta1
          Col5 = delta2
          Col6 = nu1
          Col7 = lkdf;

data SmoothCriteria(keep = SmoothingParameter aicc1);
   set temp;
   sigmaHat=rss/n;
   aicc1=n*(log(sigmaHat) + (delta1/delta2)*(n+nu1)/(lkdf-2) );

proc sort data=SmoothCriteria(where=(aicc1^=.))
               out=Aicc1Results;
   by aicc1;
run;

ods listing;
proc print data=Aicc1Results(obs=1);
   title2 'Smoothing Parameter Minimizing the AICC1 Statistic';
   id SmoothingParameter;
run;

proc sort data=Aicc1Results;
  by SmoothingParameter;
run;
```

```
title1 'AICC1 Criterion';
symbol1 c=black i=join value=none width=2;
proc gplot data=SmoothCriteria;
   plot aicc1*SmoothingParameter/
      hminor = 0
      vminor = 0
      vaxis  = axis1
      frame;
      axis1 label = ( r=0 a=90 );
run; quit;

title1;
%mend;
```

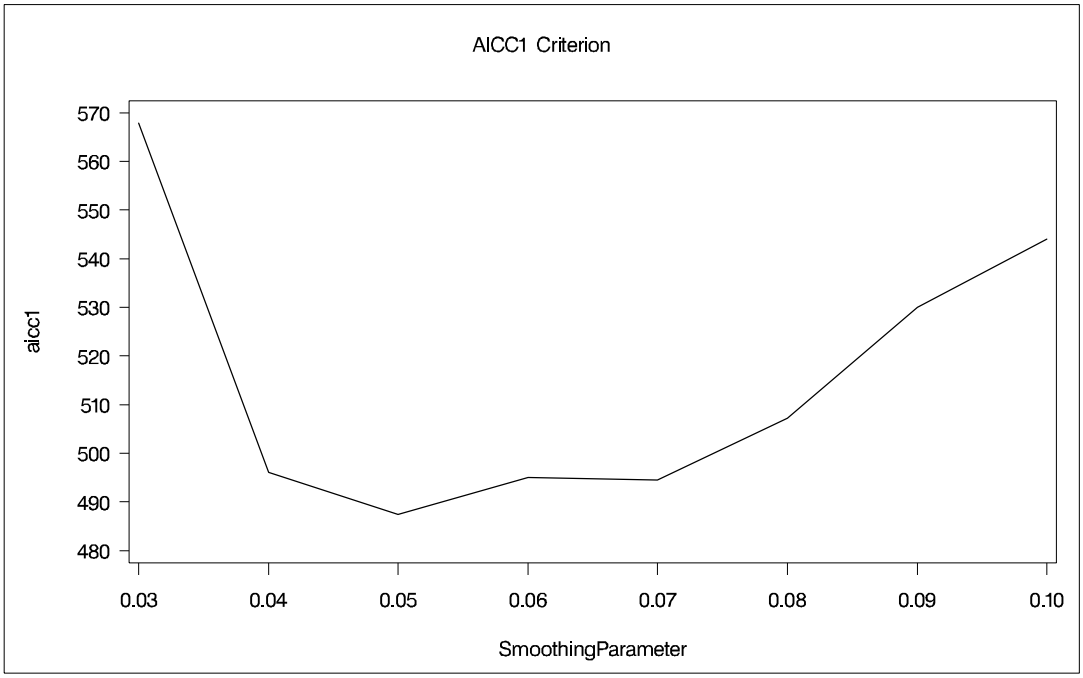
You can use this macro as follows:

```
proc loess data=ENSO;
   ods output fitsummary=ENSOsummary;
   model Pressure=Month /
       smooth = 0.03 to 0.1 by 0.01
       dfmethod=exact;
run;

%SmoothSelect(EnsoSummary);
```

You use the DFMETHOD=EXACT option in the MODEL statement to produce the statistics needed to compute the $AICC_1$ criterion. You use the SMOOTH= option to specify the list of smoothing parameters you want to examine. You use the ODS OUTPUT statement to capture the "Fit Summary" tables for these smoothing parameters in the data set that you use as the argument for the SMOOTHSELECT macro.

You obtain the plot shown in Output 38.4.3 and the output shown in Output 38.4.4.

Output 38.4.3. AIC_{C_1} Statistic for the ENSO data



Output 38.4.4. Smoothing Parameter Minimizing the AIC_{C_1} Statistic

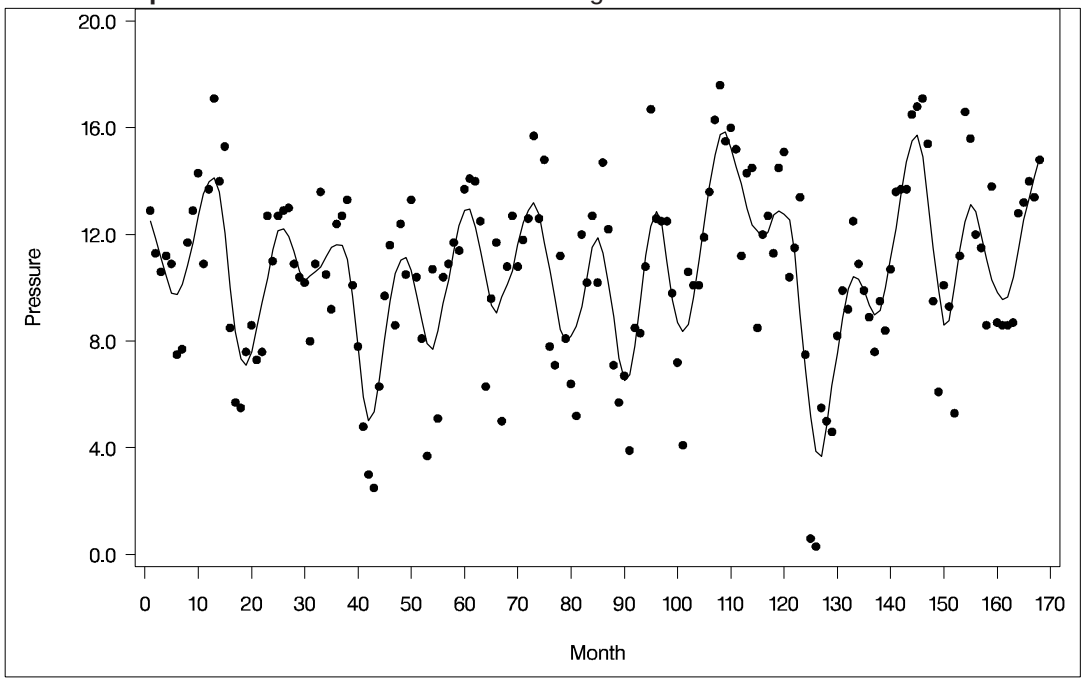
Smoothing Parameter Minimizing the AICC1 Statistic

Smoothing Parameter	aicc1
0.05	487.424

From Output 38.4.4 you see that 0.05 is the best smoothing parameter among the loess fits you requested. You can plot the loess fit for this smoothing parameter with the following statements:

```
ods output OutputStatistics=ENSOstats;

proc loess data=ENSO;
   model Pressure=Month/ smooth =0.05;
run;

symbol1 color=black value=dot h=2.5 pct;
symbol2 color=black interpol=join value=none width=2;
proc gplot data=ENSOstats;
   plot (depvar pred)*Month / overlay
        hminor = 0
        vminor = 0
        vaxis  = axis1
        frame;
        axis1 label = ( r=0 a=90 ) order=(0 to 20  by 4);
run; quit;
run;
```

Output 38.4.5. Loess Fit with Smoothing Parameter 0.05



References

Brinkman, N.D. (1981), "Ethanol Fuel - a Single Engine Study of Efficiency and Exhaust Emissions," *SAE Transactions* 90, No. 810345, 1410–1424.

Cleveland, W.S., Devlin, S.J., and Grosse, E. (1988), "Regression By Local Fitting," *Journal of Econometrics*, 37, 87–114.

Cleveland, W.S. and Grosse, E. (1991), "Computational Methods for Local Regression," *Statistics and Computing*, 1, 47–62.

Cleveland, W.S., Grosse, E., and Ming-Jen Shyu (1992), "A Package of C and Fortran Routines for Fitting Local Regression Models," unpublished paper.

Craven, P. and Wahba, G. (1979), "Smoothing Noisy Data with Spline Functions," *Numer. Math.*, 31, 377–403.

Houghton, A.N., Flannery, J., and Viola, M.V. (1980), "Malignant Melanoma in Connecticut and Denmark," *International Journal of Cancer*, 25, 95–104.

Hurvich, C. M., and Simonoff, J. S. (1998), "Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion" *Journal of the Royal Statistical Society B*, 60, 271–293.

NIST (1998), "Statistical Reference Data Sets," [http://www.nist.gov/itl/div898/strd], accessed 20 January 1999.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/STAT® User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999.

SAS/STAT® User's Guide, Version 8

ISBN 1–58025–494–2

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999