SAS/STAT® 15.1 User's Guide

The LOESS Procedure

This document is an individual chapter from *SAS/STAT® 15.1 User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2018. *SAS/STAT® 15.1 User's Guide*. Cary, NC: SAS Institute Inc.

SAS/STAT® 15.1 User's Guide

Copyright © 2018, SAS Institute Inc., Cary, NC, USA

All Rights Reserved. Produced in the United States of America.

For a hard-copy book: No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

For a web download or e-book: Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

U.S. Government License Rights; Restricted Rights: The Software and its documentation is commercial computer software developed at private expense and is provided with RESTRICTED RIGHTS to the United States Government. Use, duplication, or disclosure of the Software by the United States Government is subject to the license terms of this Agreement pursuant to, as applicable, FAR 12.212, DFAR 227.7202-1(a), DFAR 227.7202-3(a), and DFAR 227.7202-4, and, to the extent required under U.S. federal law, the minimum restricted rights as set out in FAR 52.227-19 (DEC 2007). If FAR 52.227-19 is applicable, this provision serves as notice under clause (c) thereof and no other notice is required to be affixed to the Software or documentation. The Government's rights in Software and documentation shall be only those set forth in this Agreement.

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

November 2018

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

SAS software may be provided with certain third-party software, including but not limited to open-source software, which is licensed under its applicable third-party software license agreement. For license information about third-party software distributed with SAS software, refer to http://support.sas.com/thirdpartylicenses.

Chapter 75
The LOESS Procedure

Contents

Overview: LOESS Procedure . . . **5682**
- Local Regression and the Loess Method . . . 5682

Getting Started: LOESS Procedure . . . **5683**
- Scatter Plot Smoothing . . . 5683

Syntax: LOESS Procedure . . . **5695**
- PROC LOESS Statement . . . 5696
- BY Statement . . . 5700
- ID Statement . . . 5701
- MODEL Statement . . . 5701
- OUTPUT Statement . . . 5706
- SCORE Statement . . . 5708
- WEIGHT Statement . . . 5709

Details: LOESS Procedure . . . **5709**
- Missing Values . . . 5709
- Output Data Sets . . . 5709
- Data Scaling . . . 5711
- Direct versus Interpolated Fitting . . . 5712
- *k*-d Trees and Blending . . . 5712
- Local Weighting . . . 5713
- Iterative Reweighting . . . 5713
- Specifying the Local Polynomials . . . 5713
- Smoothing Matrix . . . 5714
- Model Degrees of Freedom . . . 5714
- Statistical Inference and Lookup Degrees of Freedom . . . 5714
- Automatic Smoothing Parameter Selection . . . 5715
- Sparse and Approximate Degrees of Freedom Computation . . . 5718
- Scoring Data Sets . . . 5719
- ODS Table Names . . . 5719
- ODS Graphics . . . 5720

Examples: LOESS Procedure . . . **5721**
- Example 75.1: Engine Exhaust Emissions . . . 5721
- Example 75.2: Sulfate Deposits in the U.S. for 1990 . . . 5727
- Example 75.3: Catalyst Experiment . . . 5733
- Example 75.4: El Niño Southern Oscillation . . . 5740

References . . . **5748**

Overview: LOESS Procedure

The LOESS procedure implements a nonparametric method for estimating regression surfaces pioneered by Cleveland, Devlin, and Grosse (1988); Cleveland and Grosse (1991); 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 that uses k-d trees
- performs statistical inference
- performs automatic smoothing parameter selection
- performs iterative reweighting to provide robust fitting when there are outliers in the data
- produces graphs with ODS Graphics

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 $\mathbf{x}$ 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 $\mathbf{x}$, the regression function $g(\mathbf{x})$ 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_i.

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 that the error distribution satisfies some basic assumptions. In particular, such analysis is appropriate when the ϵ_i are iid 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. Alternatively, PROC LOESS also provides an OUTPUT statement to create SAS data sets from analysis results.

Getting Started: LOESS Procedure

Scatter Plot Smoothing

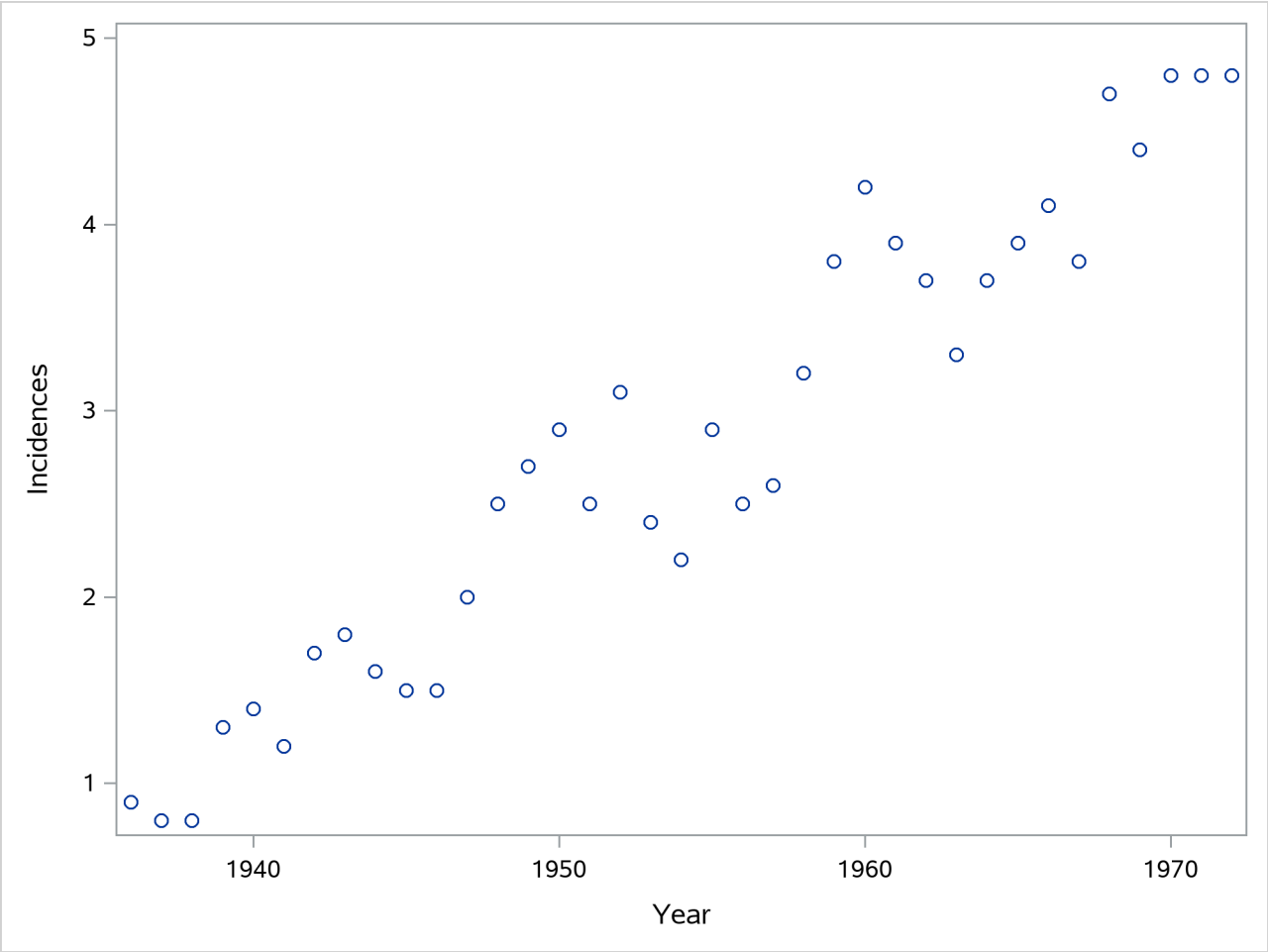
The following data from the Connecticut Tumor Registry presents age-adjusted numbers of melanoma incidences per 100,000 people for the 37 years from 1936 to 1972 (Houghton, Flannery, and Viola 1980).

```
data Melanoma;
   input  Year Incidences @@;
   format Year d4.0;
   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 SGPLOT statements produce the simple scatter plot of these data displayed in Figure 75.1.

```
proc sgplot data=Melanoma;
   scatter y=Incidences x=Year;
run;
```

Figure 75.1 Scatter Plot of the Melanoma Data

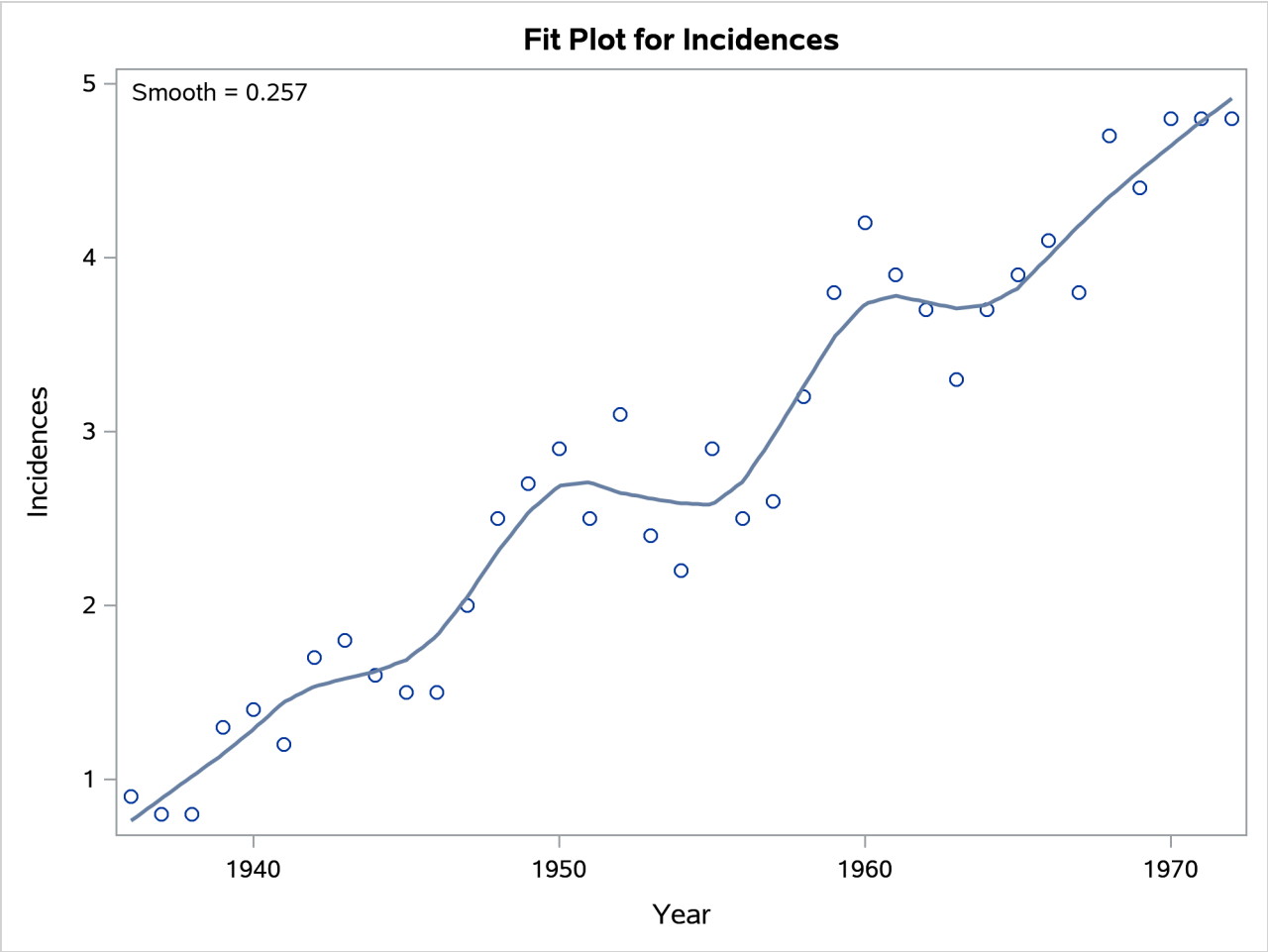


Suppose that you want to smooth the response variable Incidences as a function of the variable Year. The following PROC LOESS statements request this analysis with the default settings:

```
ods graphics on;

proc loess data=Melanoma;
   model Incidences=Year;
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.

Figure 75.2 Default Loess Fit for the Melanoma Data



When ODS Graphics is enabled, PROC LOESS produces several default plots. Figure 75.2 shows the "Fit Plot" that overlays the loess fit on a scatter plot of the data. You can see that the loess fit captures the increasing trend in the data as well as the periodic pattern in the data, which is related to an 11-year sunspot activity cycle.

Figure 75.3 Fit Summary

The LOESS Procedure
Selected Smoothing Parameter: 0.257
Dependent Variable: Incidences

Fit Summary	
Fit Method	kd Tree
Blending	Linear
Number of Observations	37
Number of Fitting Points	37
kd Tree Bucket Size	1
Degree of Local Polynomials	1
Smoothing Parameter	0.25676
Points in Local Neighborhood	9
Residual Sum of Squares	2.03105
Trace[L]	8.62243
GCV	0.00252
AICC	-1.17277

Figure 75.3 shows the "Fit Summary" table. This table details the settings used and provides statistics about the fit that is produced. You can see that smoothing parameter value for this loess fit is 0.257. This smoothing parameter determines the fraction of the data in each local neighborhood. In this example, there are 37 data points and so the smoothing parameter value of 0.257 yields local neighborhoods containing 9 observations.

Figure 75.4 Smoothing Parameter Selection

Optimal Smoothing Criterion	
AICC	**Smoothing Parameter**
-1.17277	0.25676

The "Smoothing Criterion" table provides information about how this smoothing parameter value is selected. The default method implemented in PROC LOESS chooses the smoothing parameter that minimizes the AICC criterion (Hurvich, Simonoff, and Tsai 1998) that strikes a balance between the residual sum of squares and the complexity of the fit.

You use options in the MODEL statement to change the default settings and request optionally displayed tables. For example, the following statements request that the "Model Summary" and "Output Statistics" tables be included in the displayed output. By default, these tables are not displayed.

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

Figure 75.5 Model Summary Table

The LOESS Procedure
Dependent Variable: Incidences

Model Summary				
Smoothing Parameter	Local Points	Residual SS	GCV	AICC
0.41892	15	3.42229	0.00339	-0.96252
0.68919	25	4.05838	0.00359	-0.93459
0.31081	11	2.51054	0.00279	-1.12034
0.20270	7	1.58513	0.00239	-1.12221
0.17568	6	1.56896	0.00241	-1.09706
0.28378	10	2.50487	0.00282	-1.10402
0.20270	7	1.58513	0.00239	-1.12221
0.25676	9	2.03105	0.00252	-1.17277
0.22973	8	2.02965	0.00256	-1.15145
0.25676	9	2.03105	0.00252	-1.17277

The "Model Summary" table shown in Figure 75.5 provides information about all the models that PROC LOESS evaluated in choosing the smoothing parameter value.

Figure 75.6 AICC Criterion by Smoothing Parameter

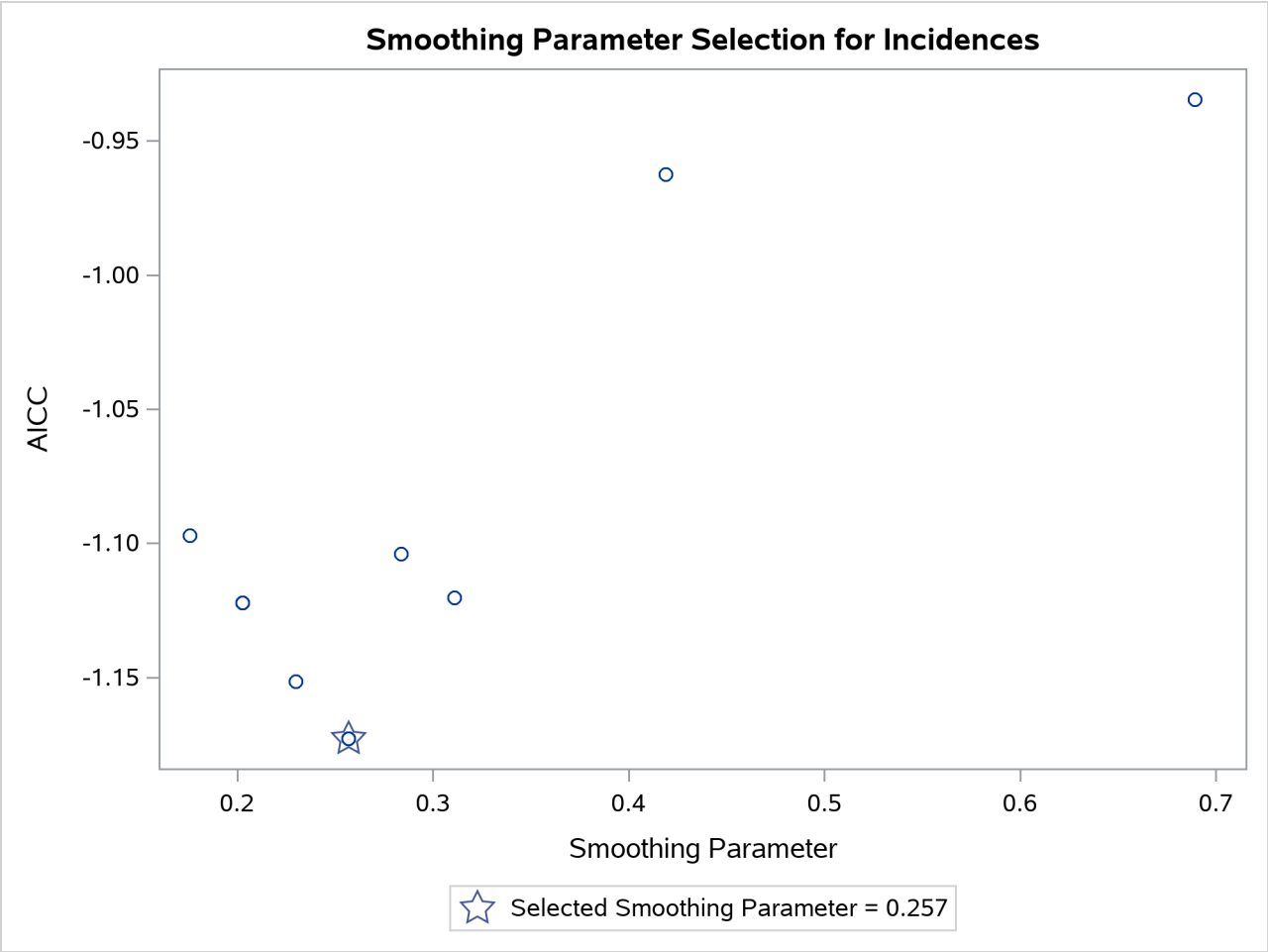


Figure 75.6 shows the "Criterion Plot" that provides a graphical display of the smoothing parameter selection process.

Figure 75.7 Output Statistics

The LOESS Procedure
Selected Smoothing Parameter: 0.257
Dependent Variable: Incidences

Output Statistics				
Obs	Year	Incidences	Predicted Incidences	Residual
1	1936	0.90000	0.76235	0.13765
2	1937	0.80000	0.88992	-0.08992
3	1938	0.80000	1.01764	-0.21764
4	1939	1.30000	1.14303	0.15697
5	1940	1.40000	1.28654	0.11346
6	1941	1.20000	1.44528	-0.24528
7	1942	1.70000	1.53482	0.16518
8	1943	1.80000	1.57895	0.22105
9	1944	1.60000	1.62058	-0.02058
10	1945	1.50000	1.68627	-0.18627
11	1946	1.50000	1.82449	-0.32449
12	1947	2.00000	2.04976	-0.04976
13	1948	2.50000	2.30981	0.19019
14	1949	2.70000	2.53653	0.16347
15	1950	2.90000	2.68921	0.21079
16	1951	2.50000	2.70779	-0.20779
17	1952	3.10000	2.64837	0.45163
18	1953	2.40000	2.61468	-0.21468
19	1954	2.20000	2.58792	-0.38792
20	1955	2.90000	2.57877	0.32123
21	1956	2.50000	2.71078	-0.21078
22	1957	2.60000	2.96981	-0.36981
23	1958	3.20000	3.26005	-0.06005
24	1959	3.80000	3.54143	0.25857
25	1960	4.20000	3.73482	0.46518
26	1961	3.90000	3.78186	0.11814
27	1962	3.70000	3.74362	-0.04362
28	1963	3.30000	3.70904	-0.40904
29	1964	3.70000	3.72917	-0.02917
30	1965	3.90000	3.82382	0.07618
31	1966	4.10000	4.00515	0.09485
32	1967	3.80000	4.18573	-0.38573
33	1968	4.70000	4.35152	0.34848
34	1969	4.40000	4.50284	-0.10284
35	1970	4.80000	4.64413	0.15587
36	1971	4.80000	4.78291	0.01709
37	1972	4.80000	4.91602	-0.11602

Figure 75.7 show the “Output Statistics” table that contains the predicted loess fit value at each observation in the input data set.

Although the default method for selecting the smoothing parameter value is often satisfactory, it is often a good practice to examine how the loess fit varies with the smoothing parameter. In some cases, fits with different smoothing parameters might reveal important features of the data that cannot be discerned by looking at a fit with just a single "best" smoothing parameter. Example 75.4 provides such an example. You can produce the loess fits for a range of smoothing parameters by using the SMOOTH= option in the MODEL statement as follows:

```
proc loess data=Melanoma;
   model Incidences=Year/smooth=0.1 0.25 0.4 0.6 residual;
   ods output OutputStatistics=Results;
run;
```

The RESIDUAL option causes the residuals to be added to the "Output Statistics" table. Note that, even if you do not specify the DETAILS option in the MODEL statement to request the display of the "Output Statistics" table, you can use an ODS OUTPUT statement to output this and other optionally displayed tables as data sets.

PROC PRINT displays the first five observations of the Results data set:

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

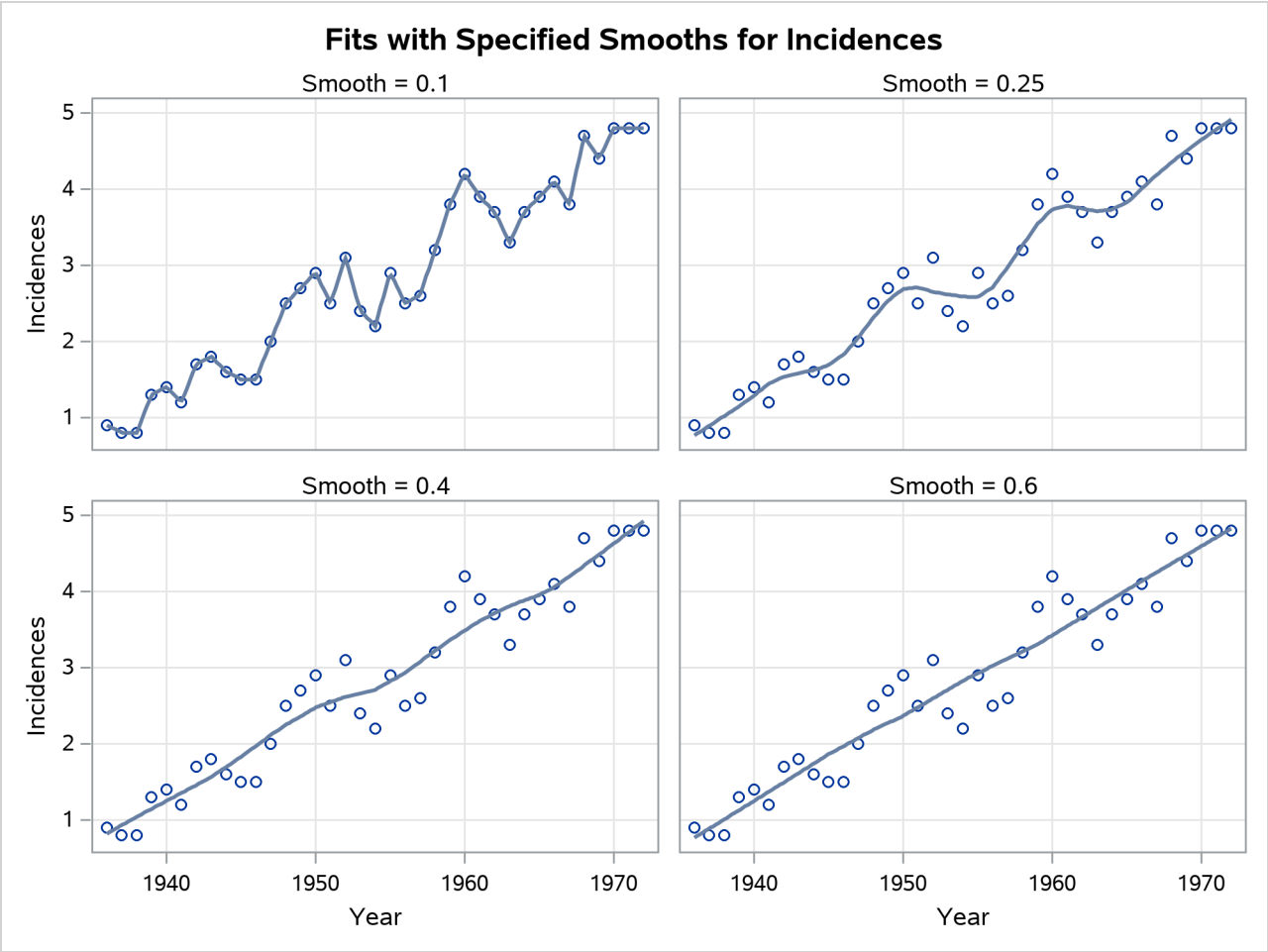
Figure 75.8 PROC PRINT Output of the Results Data Set

Obs	SmoothingParameter	Year	DepVar	Pred	Residual
1	0.1	1936	0.9	0.90000	0
2	0.1	1937	0.8	0.80000	0
3	0.1	1938	0.8	0.80000	0
4	0.1	1939	1.3	1.30000	0
5	0.1	1940	1.4	1.40000	0

Note that the fits for all the smoothing parameters are placed in single data set. A variable named SmoothingParameter that you use to distinguish each fit is included in this data set.

When you specify a list of smoothing parameters for a model and ODS Graphics is enabled, PROC LOESS produces a panel containing up to six plots that show the fit obtained for each value of the smoothing parameter that you specify. If you specify more than six smoothing values, then multiple panels are produced. For each regressor, PROC LOESS also produces panels of the residuals versus each regressor by the smoothing parameters that you specify.

Figure 75.9 Loess Fits for a Range of Smoothing Parameters



If you examine the plots in Figure 75.9, you see that a visually reasonable fit is obtained with smoothing parameter values of 0.25. With smoothing parameter value 0.1, there is gross overfitting in the sense that the original data are exactly interpolated. When the smoothing parameter value is 0.4, you obtain an overly smooth fit where the contribution of the sunspot cycle has been mostly averaged away. At smoothing parameter value 0.6 the fit shows just the increasing trend in the data.

It is also instructive to look at scatter plots of the residuals for each of the fits. These are also produced by default by PROC LOESS when ODS Graphics is enabled.

Figure 75.10 Residuals of Loess Fits for a Range of Smoothing Parameters

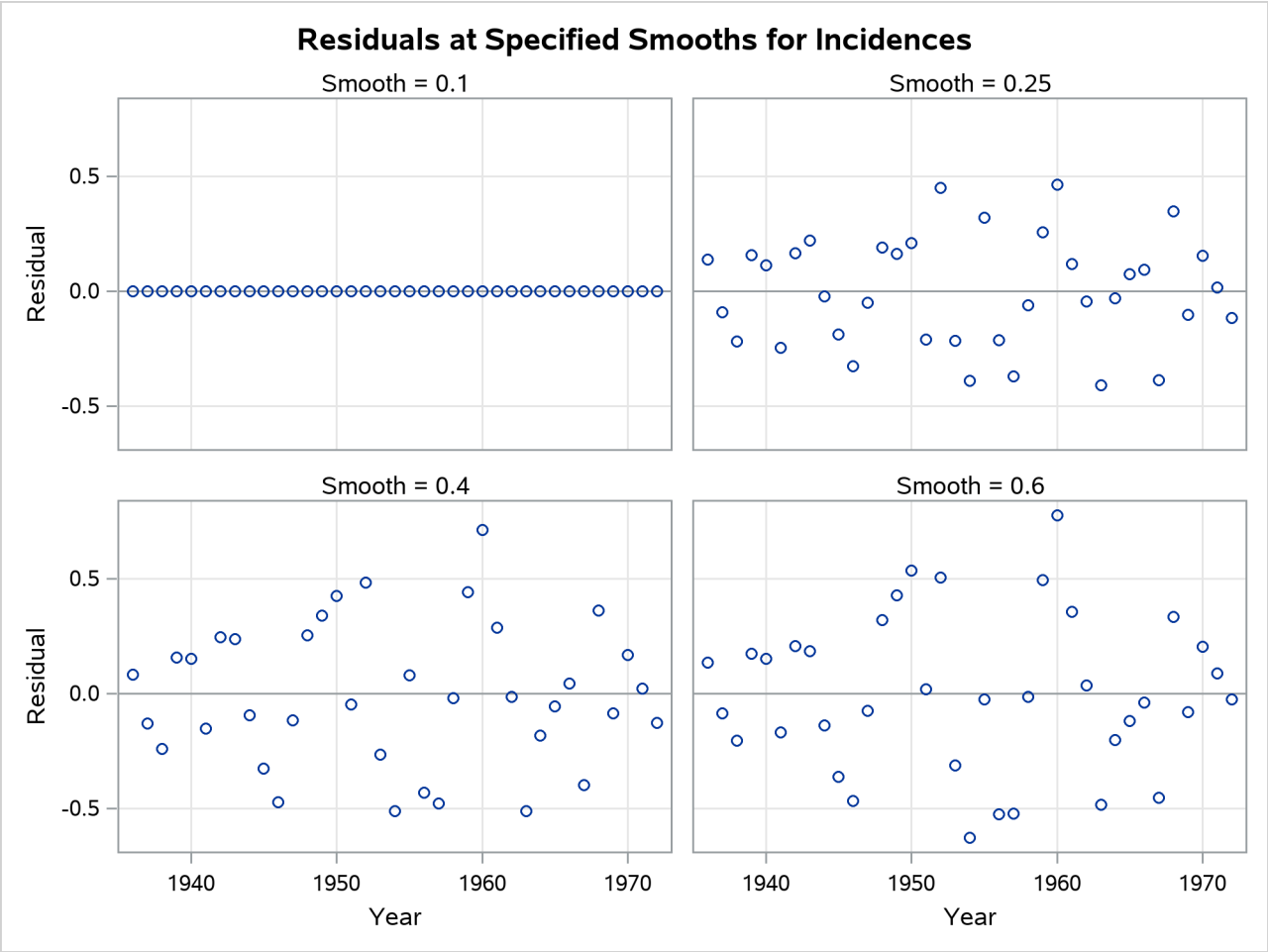
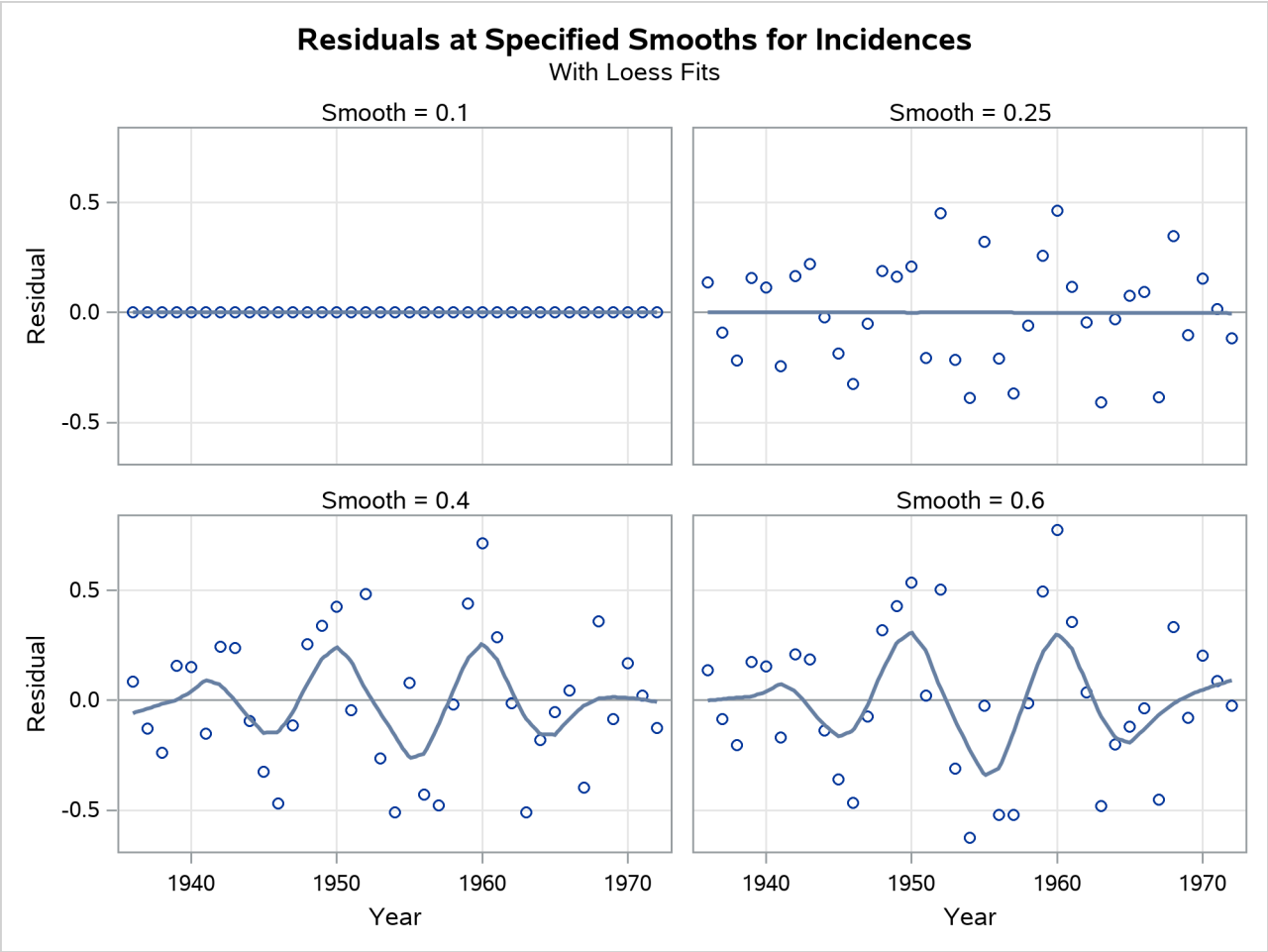


Figure 75.10 shows a scatter plot of the residuals by year for each smoothing parameter value. One way to discern patterns in these residuals is to superimpose a loess fit on each plot in the panel. You request loess fits on the residual plots in this panel by specifying the SMOOTH= suboption of the PLOTS=RESIDUALSBYSMOOTH option in the PROC LOESS statement. Note that the loess fits that are displayed on each of the residual plots are obtained independently of the loess fit that produces these residuals. The following statements show how you do this for the Melanoma data.

```
proc loess data=Melanoma plots=ResidualsBySmooth(smooth);
   model Incidences=Year/smooth=0.1 0.25 0.4 0.6;
run;
```

Figure 75.11 Residuals with Superimposed Loess Fits



The loess fits shown on the plots in Figure 75.11 help confirm the conclusions obtained when you look at Figure 75.9. Note that residuals for smoothing parameter value 0.25 do not exhibit any pattern, confirming that at this value the loess fit of the melanoma data has successfully modeled the variation in this data. By contrast, the residuals for the fit with smoothing parameter 0.6 retain the variation caused by the sunspot cycle.

The examination of the fits and residuals obtained with a range of smoothing parameter values confirms that the value of 0.257 that PROC LOESS selects automatically is appropriate for these data. The next step in this analysis is to examine fit diagnostics and produce confidence limit for the fit. If ODS Graphics is enabled, then a panel of fit diagnostics is produced. Furthermore, you can request prediction confidence limits by adding the CLM option in the MODEL statement. By default 95% limits are produced, but you can use the ALPHA= option in the MODEL statement to change the significance level. The following statements request 90% confidence limits.

```
proc loess data=Melanoma;
   model Incidences=Year/clm alpha=0.1;
run;

ods graphics off;
```

Figure 75.12 Fit Diagnostics

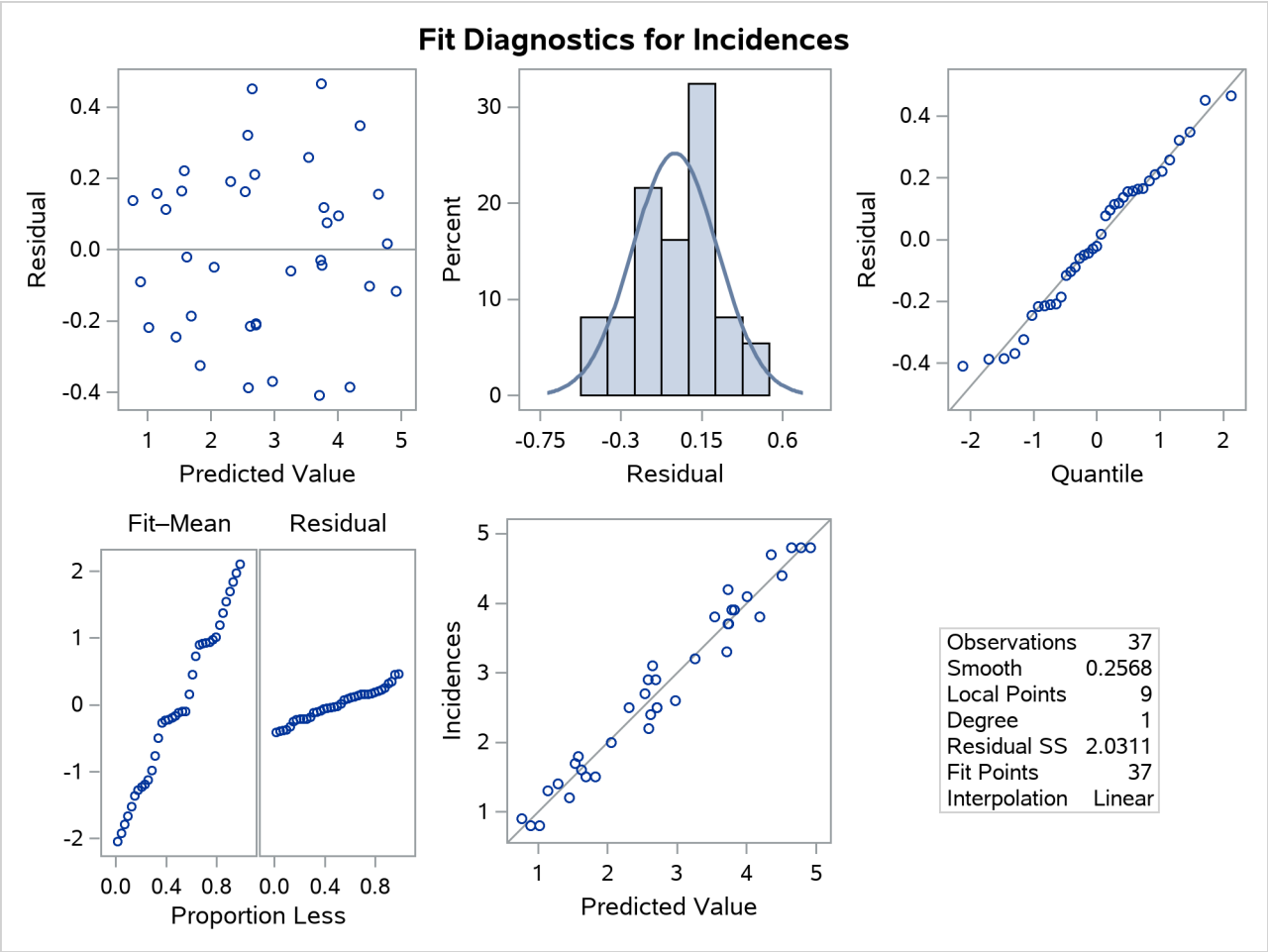
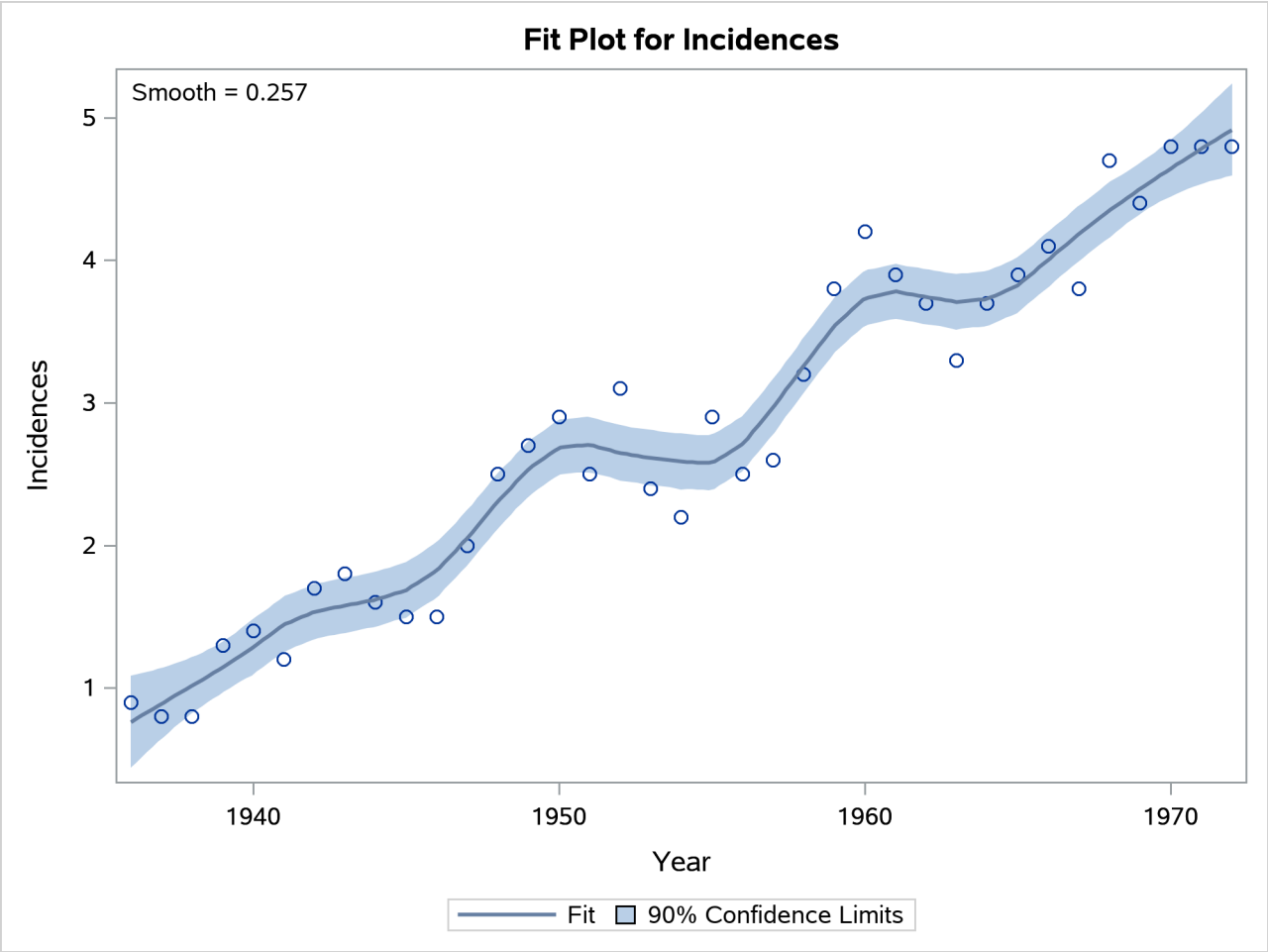


Figure 75.12 shows the fit diagnostics panel. The histogram of the residuals with overlaid normal density estimator and the normal quantile plot show that the residuals do exhibit some small departure from normality. The "Residual-Fit" spread plot shows that the spread in the centered fit is much wider that the spread in the residuals. This indicates that the fit has accounted for most of the variation in the incidences of melanoma in this data. This conclusion is supported by the absence of any clear pattern in the scatter plot of residuals by predicted values and the closeness of the points to the 45-degree reference line in the plot of observed by predicted values.

Figure 75.13 Loess Fit of Melanoma Data with 90% Confidence Limits



Finally, Figure 75.13 shows the selected loess fit with 90% confidence limits.

Syntax: LOESS Procedure

The following statements are available in the LOESS procedure:

PROC LOESS < **DATA=***SAS-data-set* > **;**
  MODEL *dependents* **=** *regressors* < / *options* > **;**
  OUTPUT < **OUT=***SAS-data-set* > < *keyword* < **=***name* > > < . . . *keyword* < **=***name* > > < / *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 OUTPUT, 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.
OUTPUT	creates an output data set containing predicted values, residuals, and results of statistical inference.
SCORE	specifies a data set containing observations to be scored.
WEIGHT	declares a variable to weight observations.

PROC LOESS Statement

PROC LOESS < *options* > **;**

The PROC LOESS statement invokes the LOESS procedure. The PROC LOESS statement is required. You can specify the following *options* in the PROC LOESS statement:

DATA=*SAS-data-set*
names the SAS data set to be used by PROC LOESS. If the DATA= option is not specified, PROC LOESS uses the most recently created SAS data set.

PLOTS < (*global-plot-options*) > < = *plot-request* < (*options*) > >
PLOTS < (*global-plot-options*) > < = (*plot-request* < (*options*) > < ... *plot-request* < (*options*) > >) >
controls the plots produced through ODS Graphics. When you specify only one plot request, you can omit the parentheses around the plot request. Here are some examples:

```
plots=none
plots=residuals(smooth)
plots(unpack)=diagnostics
plots(only)=(fit residualHistogram)
```

ODS Graphics must be enabled before plots can be requested. For example:

```
ods graphics on;

proc loess;
   model y = x;
run;

ods graphics off;
```

For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" on page 623 in Chapter 21, "Statistical Graphics Using ODS."

If ODS Graphics is enabled but you but do not specify the PLOTS= option, then PROC LOESS produces a default set of plots. The following table lists the default set of plots produced.

Table 75.1 Default Graphs Produced

Plot	Conditional On
ContourFitPanel	SMOOTH= option specified in the MODEL statement
ContourFit	Model with two regressors
CriterionPlot	Smoothing parameter selection performed
DiagnosticsPanel	Unconditional
ResidualsBySmooth	SMOOTH= option specified in the MODEL statement
ResidualPanel	Unconditional
FitPanel	SMOOTH= option specified in the MODEL statement
FitPlot	Model with one regressor
ScorePlot	One or more SCORE statements and a model with one regressor

For models with multiple dependent variables, separate plots are produced for each dependent variable. For models where multiple smoothing parameters are requested with the SMOOTH= option in the MODEL statement and smoothing parameter value selection is not requested, separate plots are produced for each smoothing parameter. If smoothing parameter value selection is requested with the SELECT= option in the MODEL statement, then the plots are produced for the selected model only. However, if you specify the STEPS suboption of the SELECT= option, then plots are produced for all smoothing parameters examined in the selection process.

The *global-plot-options* apply to all relevant plots generated by the LOESS procedure, unless they are overridden with a *specific-plot-option*. The *global-plot-options* supported by the LOESS procedure follow.

Global Plot Options

MAXPOINTS=NONE | *number*

specifies that plots with elements that require processing more than *number* points are suppressed. The default is MAXPOINTS=5000. This cutoff is ignored if you specify MAXPOINTS=NONE.

ONLY

suppresses the default plots. Only the plots specifically requested are produced.

UNPACK

suppresses paneling. By default, multiple plots can appear in some output panels. Specify UNPACK to get each plot individually. You can specify PLOTS(UNPACK) to unpack the default plots. You can also specify UNPACK as a suboption with CONTOURFITPANEL, DIAGNOSTICS, FITPANEL, RESIDUALS and RESIDUALSBYSMOOTH.

Specific Plot Options

The following listing describes the specific plots and their *options*.

ALL

requests that all plots appropriate for the particular analysis be produced. You can specify other options with ALL; for example, to request all plots and unpack only the residuals, specify PLOTS=(ALL RESIDUALS(UNPACK)).

CONTOURFIT < (*contour-options*) >

produces a contour plot of the fitted surface overlaid with a scatter plot of the data for models with two regressors. Contour plots are not produced if you specify the DIRECT option in the MODEL statement. You can use the following *contour-options* to control how the observations are displayed:

OBS=GRADIENT

specifies that observations be displayed as circles colored by the observed response. The same color gradient is used to display the fitted surface and the observations. Observations where the predicted response is close to the observed response have similar colors—the greater the contrast between the color of an observation and the surface, the larger the residual is at that point. OBS=GRADIENT is the default if you do not specify any *contour-options*.

OBS=NONE

suppresses the observations.

OBS=OUTLINE

specifies that observations be displayed as circles with a border but with a completely transparent fill.

OBS=OUTLINEGRADIENT

is the same as OBS=GRADIENT except that a border is shown around each observation. This option is useful to identify the location of observations where the residuals are small, because at these points the color of the observations and the color of the surface are indistinguishable.

CONTOURFITPANEL < (< UNPACK > < *contour-options* >) >

produces panels of contour plots overlaid with a scatter plot of the data for each smoothing parameter specified in the SMOOTH= option in the MODEL statement, for models with two regressors. This plot is not produced if you specify the DIRECT option in the MODEL statement. If you do not specify the SMOOTH= option or if the model does not have two regressors, then this plot is not produced. If you specify the SELECT= option in addition to the SMOOTH= option in the MODEL statement, then you need to additionally specify the STEPS suboption of the SELECT= option to obtain this plot. Note that each panel contains at most six plots, and multiple panels are used in the case that there are more than six smoothing parameters in the SMOOTH= option in the MODEL statement. See the CONTOURFIT option for a description of the individual plots in this panel. The UNPACK option suppresses paneling, and the *contour-options* are the same as for the CONTOURFIT option.

CRITERIONPLOT | CRITERION

displays a scatter plot of the value of the SELECT= criterion versus the smoothing parameter value for all smoothing parameter values examined in the selection process. This plot is not produced if smoothing parameter selection is not done.

DIAGNOSTICSPANEL | DIAGNOSTICS < (UNPACK) >

produces a summary panel of fit diagnostics consisting of the following:

- residuals versus the predicted values
- histogram of the residuals
- normal quantile plot of the residuals
- a “Residual-Fit” (or RF) plot consisting of side-by-side quantile plots of the centered fit and the residuals.

- dependent variable values versus the predicted values

You can request the five plots in this panel as individual plots by specifying the UNPACK option. You can also request individual plots in the panel by name without having to unpack the panel. Note that the fit diagnostics panel is produced by default whenever ODS Graphics is enabled.

FITPANEL < (UNPACK) >
produces panels of plots showing the fitted LOESS curve overlaid on a scatter plot of the input data for each smoothing parameter specified in the SMOOTH= option in the MODEL statement. If you do not specify the SMOOTH= option or the model has more than one regressor, then this plot is not produced. If you specify the SELECT= option in addition to the SMOOTH= option in the MODEL statement, then you need to additionally specify the STEPS suboption of the SELECT= option to obtain this plot. Note that each panel contains at most six plots, and multiple panels are used in the case that there are more than six smoothing parameters in the SMOOTH= option in the MODEL statement. If the CLM option is specified in the MODEL statement, then a confidence band at the significance level specified in the ALPHA= option is included in each plot in the panels. If you specify the UNPACK option, then all fit panels are unpacked.

FITPLOT | FIT
produces a scatter plot of the input data with the fitted LOESS curve overlaid for models with a single regressor. If the CLM option is specified in the MODEL statement, then a confidence band at the significance level specified in the ALPHA= option is included in the plot.

NONE
suppresses all plots.

OBSERVEDBYPREDICTED
produces a scatter plot of the dependent variable values by the predicted values.

QQPLOT | QQ
produces a normal quantile plot of the residuals.

RESIDUALSBYSMOOTH < (< UNPACK > < SMOOTH >) >
produces for each regressor panels of plots showing the residuals of the LOESS fit versus the regressor for each smoothing parameter specified in the SMOOTH= option in the MODEL statement. If you do not specify the SMOOTH= option, then this plot is not produced. If you specify the SELECT= option in addition to the SMOOTH= option in the MODEL statement, then you need to additionally specify the STEPS suboption of the SELECT= option to obtain this plot. Note that each panel contains at most six plots, and multiple panels are used in the case that there are more than six smoothing parameters in the SMOOTH= option in the MODEL statement. If you specify the UNPACK option, then all RESIDUALSBYSMOOTH panels are unpacked.

The SMOOTH option requests that a nonparametric fit line be shown in each plot in the panel. The type of nonparametric fit and the options used are controlled by the template that underlies this plot. In the standard template that is provided, the nonparametric smooth is specified to be a loess fit corresponding to the default options of PROC LOESS, except that the PRESEARCH suboption is always used. It is important to note that the loess fit that is shown in each of the residual plots is computed independently of the loess fit that is used to obtain the residuals.

RESIDUALBYPREDICTED
produces a scatter plot of the residuals by the predicted values.

RESIDUALHISTOGRAM
produces a histogram of the residuals.

RESIDUALPANEL | RESIDUALS **<(***residual-options***)>**
produces panels of the residuals versus the regressors in the model. Note that each panel contains at most six plots, and multiple panels are used when there are more than six regressors in the model.

The following *residual-options* are available:

SMOOTH
requests that a nonparametric fit line be shown in each plot in the panel. The type of nonparametric fit and the options used are controlled by the template that underlies this plot. In the standard template that is provided, the nonparametric smooth is specified to be a loess fit corresponding to the default options of PROC LOESS, except that the PRESEARCH suboption is always used. It is important to note that the loess fit that is shown in each of the residual plots is computed independently of the loess fit that is used to obtain the residuals.

UNPACK
suppresses paneling.

RFPLOT | RF
produces a "Residual-Fit" (or RF) plot consisting of side-by-side quantile plots of the centered fit and the residuals. This plot "shows how much variation in the data is explained by the fit and how much remains in the residuals" (Cleveland 1993).

SCOREPLOT | SCORE
produces a scatter plot of the scored values at the score points for each SCORE statement. SCORE plots are not produced for models with more than one regressor. If the CLM option is specified in the MODEL statement, then confidence bars at the significance level specified in the ALPHA= option are shown at score data points.

BY Statement

BY *variables* **;**

You can specify a BY statement in PROC LOESS to obtain separate analyses of observations in groups that are defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. If you specify more than one BY statement, only the last one specified is used.

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

- Sort the data by using the SORT procedure with a similar BY statement.
- Specify the NOTSORTED or DESCENDING option in the BY statement in the LOESS procedure. 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 by using the DATASETS procedure (in Base SAS software).

For more information about BY-group processing, see the discussion in *SAS Language Reference: Concepts*. For more information about the DATASETS procedure, see the discussion in the *Base 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* **=** *regressors* < / *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 75.2 summarizes the *options* available in the MODEL statement.

Table 75.2 Summary of MODEL Statement Options

Option	Description
Fit Options	
BUCKET=	specifies the number of points in *k*-d tree buckets
DEGREE=	specifies the degree of local polynomials (1 or 2)
DFMETHOD=	specifies the method of computing lookup degrees of freedom
DIRECT	specifies direct fitting at every data point
DROPSQUARE=	specifies the variables whose squares are to be dropped from local quadratic polynomials
INTERP=	specifies the interpolating polynomials (linear or cubic)
ITERATIONS=	specifies the number of reweighting iterations
SCALE=	specifies the method used to scale the regressor variables
SELECT=	specifies that automatic smoothing parameter selection be done
SMOOTH=	specifies the list of smoothing values
Output Statistics Table Options	
ALL	requests CLM, RESIDUAL, SCALEDINDEP, STD, and T options
CLM	displays confidence limits for mean predictions
RESIDUAL	displays residuals
SCALEDINDEP	displays scaled independent variable coordinates
STD	displays standard errors of the mean predicted values
T	displays *t* statistics

Table 75.2 *continued*

Option	Description
Other Options	
ALPHA=	sets significance level for confidence intervals
DETAILS=	specifies which tables are to be displayed
TRACEL	displays the trace of the smoothing matrix

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 k-d tree. The default value used is $s * n/5$, where s is a smoothing parameter value 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 significance 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 and 2 for local quadratic fitting, with 1 being the default.

DETAILS <(*tables* **)>**
selects which tables to display, where *tables* is one or more of the specifications KDTREE, MODELSUMMARY, OUTPUTSTATISTICS, and PREDATVERTICES:

- KDTREE displays the k-d tree structure.
- MODELSUMMARY displays the fit criteria for all smoothing parameter values that are specified in the SMOOTH= option in the MODEL statement, or that are fit with automatic smoothing parameter selection.
- OUTPUTSTATISTICS displays the predicted values and other requested statistics at the points in the input data set.
- PREDATVERTICES displays fitted values and coordinates of the k-d tree vertices where the local least squares fitting is done.

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 | APPROX <(*approx-options*)>

specifies the method used to calculate the lookup degrees of freedom used in performing statistical inference. The default is DFMETHOD=NONE, unless you specify any of the MODEL statement options ALL, CLM, STD, and T, or any SCORE statement CLM option, in which case the default is DFMETHOD=EXACT.

You can specify the following *approx-options* in parentheses after the DFMETHOD=APPROX option:

QUANTILE=*number*

specifies that the smallest 100(*number*)% of the nonzero coefficients in the smoothing matrix be set to zero in computing the approximate lookup degrees of freedom. The default value is QUANTILE=0.9.

CUTOFF=*number*

specifies that coefficients in the smoothing matrix whose magnitude is less than the specified value be set to zero in computing the approximate lookup degrees of freedom. Using the CUTOFF= option overrides the QUANTILE= option.

See the section "Sparse and Approximate Degrees of Freedom Computation" on page 5718 for a description of the method used when the DFMETHOD=APPROX option is specified.

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 k-d 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.

INTERP=LINEAR | CUBIC

specifies the degree of the interpolating polynomials used for blending local polynomial fits at the k-d tree vertices. This option is ignored if the DIRECT option is specified in the model statement. INTERP=CUBIC is not supported for models with more than two regressors. The default is INTERP=LINEAR.

ITERATIONS=*number*

specifies the total number of iterations to be done. The first iteration performs an initial LOESS fit. Subsequent iterations perform iterative reweighting. 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 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.

SELECT=*criterion* < (< GLOBAL > < PRESEARCH > < STEPS > < RANGE(*lower*,*upper*) >) >
SELECT=*DFCriterion* < (*target* < GLOBAL > < PRESEARCH > < STEPS > < RANGE(*lower*,*upper*) >) >

specifies that automatic smoothing parameter selection be done using the named *criterion* or *DFCriterion*. Valid values for the *criterion* are as follows:

AICC specifies the AIC_C criterion (Hurvich, Simonoff, and Tsai 1998).

AICC1 specifies the AIC_{C_1} criterion (Hurvich, Simonoff, and Tsai 1998).

GCV specifies the generalized cross validation criterion (Craven and Wahba 1979).

The *DFCriterion* specifies the measure used to estimate the model degrees of freedom. The measures implemented in PROC LOESS all depend on prediction matrix $\mathbf{L}$ relating the observed and predicted values of the dependent variable. Valid values for the *DFCriterion* are as follows:

DF1 specifies $\mathrm{Trace}(\mathbf{L})$.

DF2 specifies $\mathrm{Trace}(\mathbf{L}'\mathbf{L})$.

DF3 specifies $2\mathrm{Trace}(\mathbf{L}) - \mathrm{Trace}(\mathbf{L}'\mathbf{L})$.

For both types of selection, the smoothing parameter value is selected to yield a minimum of an optimization criterion. If you specify *criterion* as one of AICC, AICC1, or GCV, the optimization criterion is the specified *criterion*. If you specify *DFCriterion* as one of DF1, DF2, or DF3, the optimization criterion is $|\mathit{DFCriterion} - \mathit{target}|$, where *target* is a specified target degree of freedom value. Note that if you specify a *DFCriterion*, then you must also specify a target value. See the section "Automatic Smoothing Parameter Selection" on page 5715 for definitions and properties of the selection criteria.

The selection is done as follows:

- If you specify the SMOOTH=*value-list* option, then PROC LOESS selects the largest value in this list that yields the global minimum of the specified optimization criterion.
- If you do not specify the SMOOTH= option, then PROC LOESS finds a local minimum of the specified optimization criterion by using a golden section search of values less than or equal to one.

You can specify the following suboptions in parentheses after the specified criterion to alter the behavior of the SELECT= option:

GLOBAL
specifies that a global minimum be found within the range of smoothing parameter values examined. This suboption has no effect if you also specify the SMOOTH= option in the MODEL statement.

PRESEARCH
requests an initial grid search to find a smoothing parameter range within which the subsequent golden section search is done. The initial point in this grid is the smoothing parameter value corresponding to the smallest number of points, n, in the local neighborhoods that yields a fit that does not interpolate all the data points. Subsequent fits with number of local points $n + 1$, $n + 2$, $n + 4$, $n + 8$, ... are evaluated until either the number of local points exceeds the number of fitting points or the SELECT=criterion starts increasing. This suboption is ignored if you additionally specify the GLOBAL suboption of the SELECT= option or if you specify the SMOOTH= option in the MODEL statement. If you additionally specify the RANGE= suboption, then the golden section search is done on the intersection of the range found by this grid search and the range that you specify in the RANGE= suboption. This option is useful for data exhibiting features at multiple scales, because in such cases the SELECT= criterion often has multiple local minima. Using the PRESEARCH option increases the likelihood that the golden section search will find the global minimum of the SELECT= criterion. See Example 75.4 for such an example.

RANGE(*lower*,*upper*)
specifies that only smoothing parameter values greater than or equal to *lower* and less than or equal to *upper* be examined.

STEPS
specifies that all models evaluated in the selection process be displayed.

For models with one dependent variable, if you specify neither the SELECT= nor the SMOOTH= options in the MODEL statement, then PROC LOESS uses SELECT=AICC.

The following table summarizes how the smoothing parameter values are chosen for various combinations of the SMOOTH= option, the SELECT= option, and the SELECT= option modifiers.

Table 75.3 Smoothing Parameter Value(s) Used for Combinations of SMOOTH= and SELECT= OPTIONS for Models with One Dependent Variable

Syntax	Search Method	Search Domain
default	golden section using AICC	$(0, 1]$
SMOOTH=*list*	no selection	values in *list*
SMOOTH=*list* **SELECT=***criterion*	global	values in *list*
SMOOTH=*list* **SELECT=***criterion* (**RANGE**(l, u))	global	values in *list* within $[l, u]$
SELECT=*criterion*	golden section	$(0, 1]$
SELECT=*criterion* (**RANGE**(l,u))	golden section	$[l, u]$
SELECT=*criterion* (**GLOBAL**)	global	$(0, 1]$
SELECT=*criterion* (**GLOBAL RANGE**(l, u))	global	$[l, u]$

Some examples of using the SELECT= option follow:

SELECT=GCV	specifies selection that uses the GCV *criterion*.
SELECT=DF1(6.3)	specifies selection that uses the DF1 *DFCriterion* with target value 6.3.
SELECT=AICC(STEPS)	specifies selection that uses the AICC *criterion*, showing all step details.
SELECT=DF2(7 GLOBAL)	specifies selection that uses a global search algorithm to find the smoothing parameter that yields the DF2 *DFCriterion* closest to the target value 7.

NOTE: The SELECT= option cannot be used for models with more than one dependent variable.

SMOOTH=*value-list*

specifies a list of positive smoothing parameter values. If you do not specify the SELECT= option in the MODEL statement, then a separate fit is obtained for each SMOOTH= value specified. If you do specify the SELECT= option, then models with all values specified in the SMOOTH= list are examined, and PROC LOESS selects the value that minimizes the criterion specified in the SELECT= option.

For models with two or more dependent variables, if the SMOOTH= option is not specified in the MODEL statement, then SMOOTH=0.5 is used as a default.

STD

specifies that standard errors of the mean predicted values be included in the "Output Statistics" table. The use of this option implicitly selects the model option DFMETHOD=EXACT if the DFMETHOD= option has not been explicitly used.

T

specifies that *t* statistics are to be included in the "Output Statistics" table. The use of this option implicitly selects the model option DFMETHOD=EXACT if the DFMETHOD= option has not been explicitly used.

TRACEL

specifies that the trace of the prediction matrix as well as the GCV and AICC statistics be included in the "Fit Summary" table. The use of any of the MODEL statement options ALL, CLM, DFMETHOD=EXACT, DIRECT, SELECT=, STD, and T implicitly selects the TRACEL option.

OUTPUT Statement

OUTPUT <**OUT=** *SAS-data-set*> <*keyword* <**=** *name*>> <. . . *keyword* <**=***name*>> <**/** *options*> **;**

The OUTPUT statement creates a new SAS data set that saves the predicted values and other requested statistics that are calculated after models for all smoothing parameter values that are specified in the SMOOTH= option in the MODEL statement have been fit. If you do not specify a *keyword*, then only the predicted response is included.

All the variables in the original data set are included in the new data set, along with variables created by the OUTPUT statement. These new variables contain the predicted values and a variety of other statistics that are calculated for each observation in the data set.

If you want to create a SAS data set in a permanent library, you must specify a two-level name. For more information about permanent libraries and SAS data sets, see *SAS Language Reference: Concepts*.

You can specify the following *options* in the OUTPUT statement:

OUT=*SAS data set*
specifies the name of the new data set. By default, the procedure uses the DATA*n* convention to name the new data set.

keyword < **=***name* >
specifies the statistics to include in the output data set as new variables and optionally names the new variables. Specify a *keyword* for each desired statistic (see the following list of *keywords*), followed optionally by an equal sign and a variable to contain the statistic.

The new variables are named as follows: If you specify *keyword***=***name*, the new variable has the specified *name*. If you omit the optional **=***name* after a *keyword*, then the new variable name is formed by using a default character string that identifies the statistic. In either case, if you also specify the ROWWISE option after a slash and you specify more than one dependent variable or smoothing value in the MODEL statement, the variable name is appended with an order number. For details, see the ROWWISE option.

The *keywords* allowed and the statistics they represent are as follows:

PREDICTED | P creates a new variable that contains predicted values. The default *name* is Predicted.

RESIDUAL | R creates a new variable that contains residual values, which are calculated as ACTUAL – PREDICTED. The default *name* is Residual.

STD creates a new variable that contains standard errors of the mean predicted values. The use of this option implicitly selects the model option DFMETHOD=EXACT even if the DFMETHOD= option has not been explicitly used. The default *name* is StdErr.

T creates a new variable that contains t statistics. The use of this option implicitly selects the model option DFMETHOD=EXACT even if the DFMETHOD= option has not been explicitly used. The default *name* is tValue.

LCLM creates a new variable that contains the lower part of $100(1-\alpha)\%$ confidence limits on the mean predicted value. By default, the 95% limits are computed; the ALPHA= option in the MODEL statement can be used to change the significance level. The use of this option implicitly selects the model option DFMETHOD=EXACT even if the DFMETHOD= option has not been explicitly used. The default *name* is LowerCL.

UCLM creates a new variable that contains the upper part of $100(1-\alpha)\%$ confidence limits on the mean predicted value. By default, the 95% limits are computed; the ALPHA= option in the MODEL statement can be used to change the significance level. The use of this option implicitly selects the model option DFMETHOD=EXACT even if the DFMETHOD= option has not been explicitly used. The default *name* is UpperCL.

You can specify the following *options* in the OUTPUT statement after a slash (/).

ALL

requests all these *keywords*: PREDICTED, RESIDUAL, STD, T, LCLM, and UCLM.

ROWWISE | ROW

arranges the created OUTPUT data set in rowwise format. For each dependent variable and each smoothing value specified in the SMOOTH= option in the MODEL statement, one variable is generated for each specified *keyword* and the variable name is appended with an order number if there are multiple occurrences of the requested statistic. Those variables appear in an order that corresponds to the specified order of the dependent variables and the smoothing values in the MODEL statement. For each variable generated, a label is also created automatically; the label contains the default name of the represented statistic, the name of the dependent variable selected to be modeled, and the smoothing value used for calculating the represented statistic.

By default, the OUTPUT data set is created in columnwise format, where the input data is repeated for each dependent variable and for each smoothing value. Three extra columns, named SmoothingParameter for smoothing parameter values, DepVar for dependent variable names, and Obs for observation numbers, are also added to the OUTPUT data set to distinguish each model.

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 by using an ODS OUTPUT statement even if this table is not displayed.

You can specify the following *options* 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 significance level. The use of this option implicitly selects the model option DFMETHOD=EXACT if the DFMETHOD= option has not been explicitly used.

PRINT < (VAR=*variables* **) >**

specifies that the "Score Results" table 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. You can use the VAR= option to specify 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.

RESIDUAL | R
requests that residuals be added to the "Score Results" table. If the data set you specify in DATA= option in the SCORE statement does not contain one or more of the model dependent variables, then the corresponding residual values in the "Score Results" table are set to missing.

SCALEDINDEP
specifies that scaled regressor coordinates be included in the "Score Results" table. This option is ignored if the SCALE= option is not specified in the MODEL statement.

STEPS
requests that all models evaluated during smoothing parameter value selection be scored, provided that the SELECT= option together with the STEPS modifier is specified in the MODEL statement. By default only the selected model is scored.

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: LOESS Procedure

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 k-d 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 might 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 5719 for a list of the table names created by PROC LOESS. For detailed information about ODS, see Chapter 20, "Using the Output Delivery System."

For example, the following statements create an output data set named MyOutStats containing the "Output Statistics" table and an output data set named MySummary containing the "Fit Summary" 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 (four smoothing parameter values for each dependent variable).

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

The eight "Output Statistics" 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 75.8 for an example). If you want the "Output Statistics" table for each model to be in its own data set, you can use the MATCH_ALL option in the ODS OUTPUT statement. The following statements create eight data sets named MyOutStats, MyOutStats1, . . . , MyOutStats7.

```
proc loess;
   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 20, "Using the Output Delivery System."

Only the "Scale Details" and "Fit Summary" 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 "Output Statistics" and "*k*-d Tree" tables but place the "Output Statistics" and "Prediction at Vertices" 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 using the ALL option in the MODEL statement causes all optional columns to be included 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 that uses all data points 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 statement specifies 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 75.3.2 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 and PredAtVertices tables. If the SCALEDINDEP option is specified in the SCORE statement, then scaled values of the independent variables are included in the ScoreResults table. 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 that uses 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 must be 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 by using either of these two methods. 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.

k-d Trees and Blending

PROC LOESS uses a *k*-d 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. This procedure is repeated for each child cell that has more than a prespecified number of points, called the *bucket size* of the *k*-d tree.

You can specify the bucket size with the BUCKET= option in the MODEL statement. If you do not specify the BUCKET= option, the default value used is the largest integer less than or equal to $ns/5$, where n is the number of observations and s is the value of the smoothing parameter. Note that if fitting is being done for a range of smoothing parameter values, the bucket size can change for each value.

The set of vertices of all the cells of the *k*-d 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 *k*-d tree cell that contains that data point.

The univariate blending methods available in PROC LOESS are linear and cubic polynomial interpolation, with linear interpolation being the default. You can request cubic interpolation by specifying the INTERP=CUBIC option in the MODEL statement. In this case, PROC LOESS uses the unique cubic polynomial whose values and first derivatives match those of the fitted local polynomials evaluated at the two endpoints of the *k*-d tree cell edge.

In the multivariate case, such univariate interpolating polynomials are computed on each edge of the *k*-d tree cells and are combined using blending functions (Gordon 1971). In the case of two regressors, if you specify INTERP=CUBIC in the MODEL statement, PROC LOESS uses Hermite cubic polynomials as blending functions. If you do not specify INTERP=CUBIC, or if you specify a model with more than two regressors, then PROC LOESS uses linear polynomials as blending functions. In these cases, the blending method reduces to tensor product interpolation from the 2^p vertices of each *k*-d tree cell, where p is the number of regressors.

While the details of the *k*-d tree and the fitted values at the vertices of the *k*-d tree are implementation details that seldom need to be examined, PROC LOESS does provide options for their display. Each *k*-d tree subdivision of the data used by PROC LOESS is placed in the kdTree table. The predicted values at the

vertices of each k-d tree are placed in the PredAtVertices table. You can request these tables by using the DETAILS option in the MODEL statement.

Local Weighting

The size of the local neighborhoods that PROC LOESS uses in performing local fitting is determined by the smoothing parameter value s. When $s < 1$, the local neighborhood used at a point x_i contains the s fraction of the data points closest to the point x_i. 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_i. The point at distance d_i from x_i is given a weight w_i in the local regression that decreases as the distance from x_i increases. PROC LOESS uses a tricube weight function to define

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

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. See 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;
   model y= x1 x2/ degree=2 dropsquare=(x1);
run;
```

Smoothing Matrix

When no iterative reweighting is done, the "Smoothing Matrix" denoted by $\mathbf{L}$ defines the linear relationship between the fitted and observed dependent variable values of a loess model. You can obtain the predicted values of a loess fit from the observed values via

$$\hat{\mathbf{y}} = \mathbf{L}\mathbf{y}$$

where $\mathbf{y}$ is the vector of observed values and $\hat{\mathbf{y}}$ is the corresponding vector of predicted values of the dependent variable. Note that $\mathbf{L}$ is an n by n matrix, where n is the number of observations in the analysis. PROC LOESS does not explicitly form $\mathbf{L}$ if the DFMETHOD=EXACT option is not explicitly or implicitly selected.

Model Degrees of Freedom

The approximate model degrees of freedom in a nonparametric fit is a number that is analogous to the number of free parameters in a parametric model. There are three commonly used measures of model degrees of freedom in nonparametric models. These criteria are as follows:

$$\begin{aligned}
\mathrm{DF1} &\equiv \mathrm{Trace}(\mathbf{L}) \\
\mathrm{DF2} &\equiv \mathrm{Trace}(\mathbf{L}'\mathbf{L}) \\
\mathrm{DF3} &\equiv 2\mathrm{Trace}\mathbf{L} - \mathrm{Trace}(\mathbf{L}'\mathbf{L})
\end{aligned}$$

A discussion of their properties can be found in Hastie and Tibshirani (1990). DF2 is also referred to as the "Equivalent Number of Parameters," and this is the name that PROC LOESS uses for DF2 when it appears in the "Fit Summary" table.

Statistical Inference and Lookup Degrees of Freedom

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. Formulas for doing statistical inference under the preceding conditions can be found in Cleveland and Grosse (1991) and Cleveland, Grosse, and Shyu (1992). Cleveland and Grosse (1991) show that standardized residuals for a loess model follow a t distribution with ρ degrees of freedom where

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

The residual standard error that you find in the "Fit Summary" table is defined by

$$\text{Residual Standard Error} \equiv \sqrt{\text{Residual SS}/\delta_1}$$

The determination of ρ is computationally expensive and is not done by default. It is computed if you specify the DFMETHOD=EXACT or DFMETHOD=APPROX option in the MODEL statement. It is also computed if you specify any of the options CLM, STD, and T in the MODEL statement. Note that the values of δ_1, δ_2, and ρ are reported in the "Fit Summary" table.

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.

Automatic Smoothing Parameter Selection

There are several methodologies for automatic smoothing parameter selection. One class of methods chooses the smoothing parameter value to minimize a criterion that incorporates both the tightness of the fit and model complexity. Such a criterion can usually be written as a function of the error mean square, $\hat{\sigma}^2$, and a penalty function designed to decrease with increasing smoothness of the fit. This penalty function is usually defined in terms of the smoothing matrix $\mathbf{L}$ (see the section "Smoothing Matrix" on page 5714).

Examples of specific criteria are generalized cross validation (Craven and Wahba 1979) and the Akaike information criterion (Akaike 1973). These classical selectors have two undesirable properties when used with local polynomial and kernel estimators: they tend to undersmooth small data sets and tend to be nonrobust in the sense that small variations of the input data can change the choice of smoothing parameter value significantly. Hurvich, Simonoff, and Tsai (1998) obtained several corrected AIC criteria that address the small-sample bias and perform comparably with the *plug-in selectors* (Ruppert, Sheather, and Wand 1995). PROC LOESS provides automatic smoothing parameter selection that uses two of these corrected AIC criteria, named AIC_{C_1} and AIC_C in Hurvich, Simonoff, and Tsai (1998), and generalized cross validation, denoted by GCV.

The relevant formulas are

$$
\begin{aligned}
\mathrm{AIC}_{C_1} &= n\log(\hat{\sigma}^2) + n\frac{\delta_1/\delta_2(n+\nu_1)}{\delta_1^2/\delta_2 - 2} \\
\mathrm{AIC}_{C} &= \log(\hat{\sigma}^2) + 1 + \frac{2\,(\mathrm{Trace}(\mathbf{L})+1)}{n - \mathrm{Trace}(\mathbf{L}) - 2} \\
\mathrm{GCV} &= \frac{n\hat{\sigma}^2}{(n - \mathrm{Trace}(\mathbf{L}))^2}
\end{aligned}
$$

where n is the number of observations and

$$
\begin{aligned}
\delta_1 &\equiv \mathrm{Trace}(\mathbf{I}-\mathbf{L})'(\mathbf{I}-\mathbf{L}) \\
\delta_2 &\equiv \mathrm{Trace}\left((\mathbf{I}-\mathbf{L})'(\mathbf{I}-\mathbf{L})\right)^2 \\
\nu_1 &\equiv \text{Equivalent Number of Parameters} \\
&\equiv \mathrm{Trace}(\mathbf{L}'\mathbf{L})
\end{aligned}
$$

You invoke these methods for automatic smoothing parameter selection by specifying the SELECT=*criterion* option in the MODEL statement, where *criterion* is AICC1, AICC, or GCV. The LOESS procedure evaluates the specified criterion for a sequence of smoothing parameter values and selects the value in this sequence that minimizes the specified criterion. If multiple values yield the optimum, then the largest of these values is selected.

A second class of methods seeks to set an approximate measure of model degrees of freedom to a specified target value. These methods are useful for making meaningful comparisons between loess fits and other nonparametric and parametric fits. Three approximate model degrees of freedom for a loess model are defined in the section "Model Degrees of Freedom" on page 5714. You invoke these methods by specifying the SELECT=*DFCriterion*(*target*) option in the MODEL statement, where *DFCriterion* is DF1, DF2, or DF3. The criterion that is minimized is given in the following table.

Table 75.4 Minimization Criteria

Syntax	Minimization Criterion
SELECT=DF1(*target*)	$\lvert\,\mathrm{Trace}(\mathbf{L}) - \mathit{target}\rvert$
SELECT=DF2(*target*)	$\lvert\,\mathrm{Trace}(\mathbf{L}'\mathbf{L}) - \mathit{target}\rvert$
SELECT=DF3(*target*)	$\lvert 2\mathrm{Trace}(\mathbf{L}) - \mathrm{Trace}(\mathbf{L}'\mathbf{L}) - \mathit{target}\rvert$

The results are summarized in the "Smoothing Criterion" table. This table is displayed whenever automatic smoothing parameter selection is performed. You can obtain details of the sequence of models examined by specifying the DETAILS(MODELSUMMARY) option in the MODEL statement to display the "Model Summary" table.

There are several ways in which you can control the sequence of models examined by PROC LOESS. If you specify the SMOOTH=*value-list* option in the MODEL statement, then only the values in this list are

examined in performing the selection. For example, the following statements select the model that minimizes the AICC1 criterion among the three models with smoothing parameter values 0.1, 0.3, and 0.4:

```
proc loess;
   model y= x1/ smooth=0.1 0.3 0.4 select=AICC1;
run;
```

If you do not specify the SMOOTH= option in the MODEL statement, then by default PROC LOESS uses a golden section search method to find a local minimum of the specified criterion in the range $(0, 1]$. You can use the RANGE(*lower*,*upper*) modifier in the SELECT= option to change the interval in which the golden section search is performed. For example, the following statements request a golden section search to find a local minimizer of the GCV criterion for smoothing parameter values in the interval [0.1,0.5]:

```
proc loess;
   model y= x1/select=GCV( range(0.1,0.5) );
run;
```

If you want to be sure of obtaining a global minimum in the range of smoothing parameter values examined, you can specify the GLOBAL modifier in the SELECT= option. For example, the following statements request that a global minimizer of the AICC criterion be obtained for smoothing parameter values in the interval $[0.2, 0.8]$:

```
proc loess;
   model y= x1/select=AICC( global range(0.2,0.8) );
run;
```

Note that even though the smoothing parameter is a continuous variable, a given range of smoothing parameter values corresponds to a finite set of local models. For example, for a data set with 100 observations, the range $[0.2, 0.4]$ corresponds to models with $20, 21, 22, \ldots, 40$ points in the local neighborhoods. If the GLOBAL modifier is specified, all possible models in the range are evaluated sequentially.

Note that by default PROC LOESS displays a "Fit Summary" and other optionally requested tables only for the selected model. You can request that these tables be displayed for all models in the selection process by adding the STEPS modifier in the SELECT= option. Also note that by default scoring requested with SCORE statements is done only for the selected model. However, if you specify the STEPS in both the MODEL and SCORE statements, then all models evaluated in the selection process are scored.

In terms of computation, AIC_C, GCV, and DF1 depend on the smoothing matrix $\mathbf{L}$ only through its trace. In the direct method, this trace can be computed efficiently. In the interpolated method that uses k-d trees, there is some additional computational cost but the overall work is not significant compared to the rest of the computation. In contrast, the quantities δ_1, δ_2, and ν_1 that appear in the AIC_{C_1} criterion, and the DF2 and DF3 criteria, depend on the entire $\mathbf{L}$ matrix and for this reason, the time needed to compute these quantities dominates the time required for the model fitting. Hence SELECT=AICC1, SELECT=DF2, and SELECT=DF3 are much more computationally expensive than SELECT=AICC, SELECT=GCV, and SELECT=DF1, especially when combined with the GLOBAL modifier. Hurvich, Simonoff, and Tsai (1998) note that AIC_C can be regarded as an approximation of AIC_{C_1} and that "the AIC_C selector generally performs well in all circumstances."

For models with one dependent variable, PROC LOESS uses SELECT=AICC as its default, if you specify neither the SMOOTH= nor the SELECT= option in the MODEL statement. With two or more dependent

variables, automatic smoothing parameter selection needs to be done separately for each dependent variable. For this reason automatic smoothing parameter selection is not available for models with multiple dependent variables. In such cases you should use a separate PROC LOESS step for each dependent variable, if you want to use automatic smoothing parameter selection.

Sparse and Approximate Degrees of Freedom Computation

As noted in the section "Statistical Inference and Lookup Degrees of Freedom" on page 5714, obtaining confidence limits in loess models requires the computation of the lookup degrees of freedom. This in turn requires the computation of

$$\delta_2 \quad \equiv \quad \text{Trace}\left((\mathbf{I}-\mathbf{L})'(\mathbf{I}-\mathbf{L})\right)^2$$

where $\mathbf{L}$ is the loess smoothing matrix (see the section "Smoothing Matrix" on page 5714).

The work in a direct implementation of this formula grows as n^3, where n is the number of observations in analysis. For large n, this work dominates the time needed to fit the loess model itself. To alleviate this computational bottleneck, Cleveland and Grosse (1991) and Cleveland, Grosse, and Shyu (1992) developed approximate methods for estimating this quantity in terms of more readily computable statistics. A different approach to obtaining a computationally cheap estimate of δ_2 has been implemented in PROC LOESS.

For large data sets with significant local structure, the loess model is often used with small values of the smoothing parameter. Recalling that the smoothing parameter defines the fraction of the data used in each local regression, this means that the loess fit at any point in regressor space depends on only a small fraction of the data. This is reflected in the smoothing matrix $\mathbf{L}$ whose (i, j) entry is nonzero only if the ith and jth observations lie in at least one common local neighborhood. Hence the smoothing matrix is a sparse matrix (has mostly zero entries) in such cases. By exploiting this sparsity, PROC LOESS now computes δ_2 orders of magnitude faster than in previous implementations.

When each local neighborhood contains a large subset of the data—i.e., when the smoothing parameter is large—then it is no longer true that the smoothing matrix is sparse. However, since a point in a local neighborhood is given a local weight that decreases with its distance from the center of the neighborhood, many of the coefficients in the smoothing matrix turn out to be nonzero but with orders of magnitude smaller than that of the larger coefficients in the matrix. The approximate method for computing δ_2 that has been implemented in PROC LOESS exploits these disparities in magnitudes of the elements in the smoothing matrix by setting the small elements to zero. This creates a sparse approximation of the smoothing matrix to which the fast sparse methods can be applied.

In order to decide the threshold at which elements in the smoothing matrix are set to zero, PROC LOESS samples the elements in the smoothing matrix to obtain the value of the element in a specified lower quantile in this sample. The magnitude of the element at this quantile is used as a cutoff value, and all elements in the smoothing matrix whose magnitude is less than this cutoff are set to zero for the approximate computation. By default all elements in the lower ninetieth percentile are set to zero. You can use the DFMETHOD=APPROX(QUANTILE=) option in the MODEL statement to change this value. As you increase the value for the quantile to be zeroed, you speed up the degrees of freedom computation at the expense of increasing approximation errors. You can also use the DFMETHOD=APPROX(CUTOFF=) option in the MODEL statement to specify the cutoff value directly.

For small data sets, the approximate computation is not needed and would be rougher than for larger data sets. Hence PROC LOESS performs the exact computation for analyses with fewer than 500 points, even if DFMETHOD=APPROX is specified in the model statement. Also, for small values of the smoothing parameter, elements in the lower specified quantile might already all be zero. In such cases the approximate method is the same as the exact method. PROC LOESS labels as approximate any statistics that depend on the approximate computation of δ_2 only in the cases where the approximate computation was used and is different from the exact computation.

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 the 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 k-d 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 about ODS, see Chapter 20, "Using the Output Delivery System."

Table 75.5 ODS Tables Produced by PROC LOESS

ODS Table Name	Description	Statement	Option
FitSummary	Specified fit parameters and fit summary		default
kdTree	Structure of k-d tree used	MODEL	DETAILS(kdTree)
ModelSummary	Summary of all models evaluated	MODEL	DETAILS(ModelSummary)
OutputStatistics	Coordinates and fit results at input data points	MODEL	DETAILS(OutputStatistics)
PredAtVertices	Coordinates and fitted values at k-d tree vertices	MODEL	DETAILS(PredAtVertices)
ScaleDetails	Extent and scaling of the independent variables		default
ScoreResults	Coordinates and fit results at scoring points	SCORE	PRINT
SmoothingCriterion	Criterion value and selected smoothing parameter	MODEL	SELECT

ODS Graphics

Statistical procedures use ODS Graphics to create graphs as part of their output. ODS Graphics is described in detail in Chapter 21, "Statistical Graphics Using ODS."

Before you create graphs, ODS Graphics must be enabled (for example, by specifying the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section "Enabling and Disabling ODS Graphics" on page 623 in Chapter 21, "Statistical Graphics Using ODS."

The overall appearance of graphs is controlled by ODS styles. Styles and other aspects of using ODS Graphics are discussed in the section "A Primer on ODS Statistical Graphics" on page 622 in Chapter 21, "Statistical Graphics Using ODS."

You can reference every graph produced through ODS Graphics with a name. The names of the graphs that PROC LOESS generates are listed in Table 75.6, along with the relevant PLOTS= options.

Table 75.6 Graphs Produced by PROC LOESS

ODS Graph Name	Plot Description	PLOTS Option
ContourFitPanel	Panel of loess contour surfaces overlaid on scatter plots of data	CONTOURFITPANEL
ContourFit	Loess contour surface overlaid on scatter plot of data	CONTOURFITPANEL
DiagnosticsPanel	Panel of fit diagnostics	DIAGNOSTICS
FitPanel	Panel of loess curves overlaid on scatter plots of data	FITPANEL
FitPlot	Loess curve overlaid on scatter plot of data	FIT
ObservedByPredicted	Dependent variable versus loess fit	OBSERVEDBYPREDICTED
QQPlot	Normal quantile plot of residuals	QQPLOT
ResidualsBySmooth	Panel of residuals versus regressor by smoothing parameter values	RESIDUALSBYSMOOTH
ResidualByPredicted	Residuals versus loess fit	RESIDUALBYPREDICTED
ResidualHistogram	Histogram of fit residuals	RESIDUALHISTOGRAM
ResidualPanel	Panel of residuals versus regressors for fixed smoothing parameter value	RESIDUALS
ResidualPlot	Plot of residuals versus regressor	RESIDUALS
RFPlot	Side-by-side plots of quantiles of centered fit and residuals	RFPLOT
ScorePlot	Loess fit evaluated at scoring points	SCOREPLOT
CriterionPlot	Selection criterion versus smoothing parameter	CRITERION

Examples: LOESS Procedure

Example 75.1: Engine Exhaust Emissions

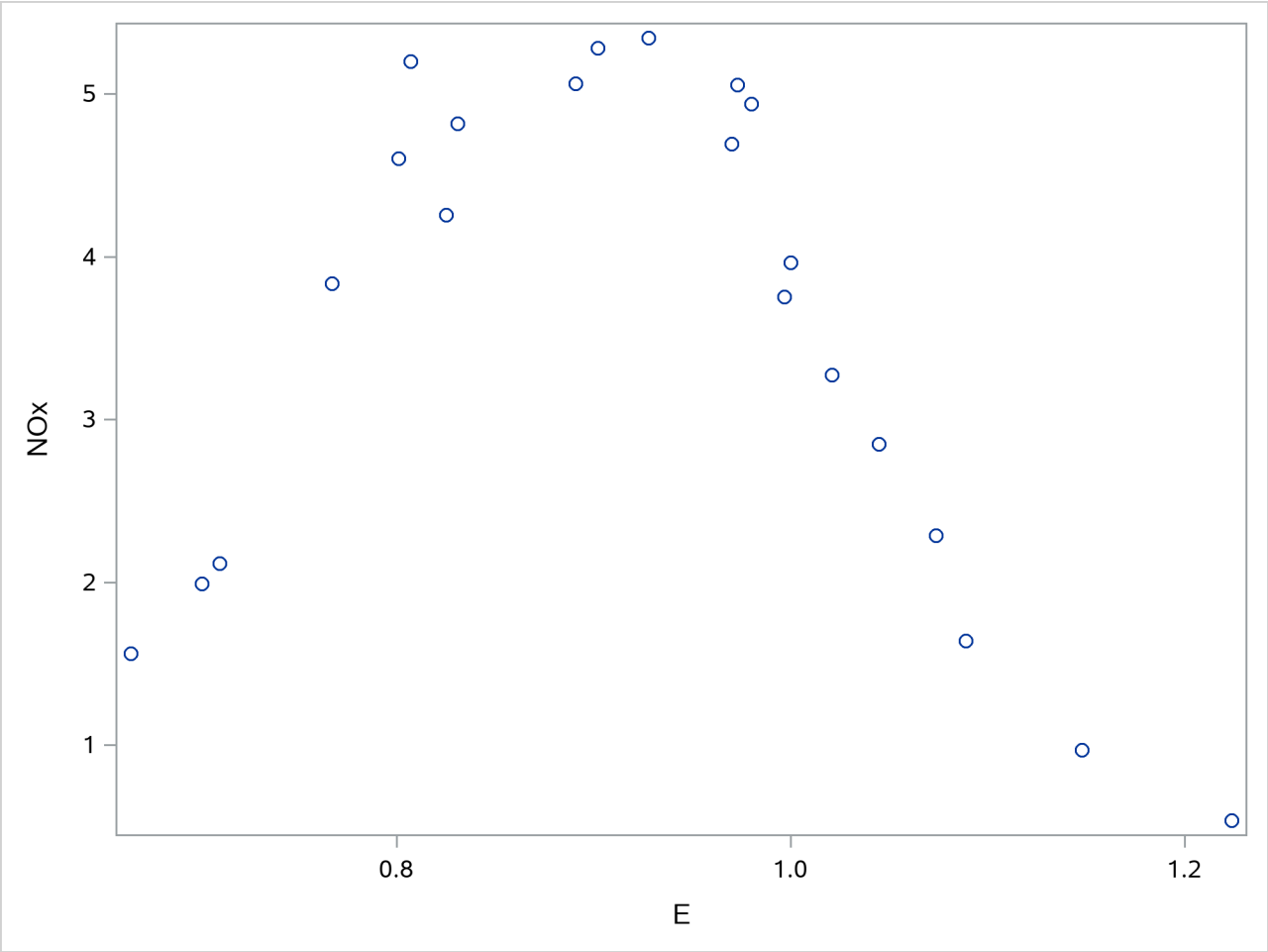
Investigators studied the exhaust emissions of a one-cylinder engine (Brinkman 1981). The SAS data set Gas contains the results data. The dependent variable, NOx, measures the concentration, in micrograms per joule, of nitric oxide and nitrogen dioxide normalized by the amount of work of the engine. 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 SGPLOT statements produce the simple scatter plot of these data displayed in Output 75.1.1.

```
proc sgplot data=Gas;
   scatter x=E y=NOx;
run;
```

Output 75.1.1 Scatter Plot of the 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.

```
ods graphics on;

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

ods graphics off;
```

Output 75.1.2 Fit Summary Table

The LOESS Procedure
Selected 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
GCV	0.00708
AICC	-0.45637
AICC1	-9.39715
Delta1	15.12582
Delta2	14.73089
Equivalent Number of Parameters	5.96950
Lookup Degrees of Freedom	15.53133
Residual Standard Error	0.33707

The "Fit Summary" table for smoothing parameter value 0.6, shown in Output 75.1.2, records the fitting parameters specified and some overall fit statistics. See the section "Smoothing Matrix" on page 5714 for a definition of the smoothing matrix **L**, and the sections "Model Degrees of Freedom" on page 5714 and "Statistical Inference and Lookup Degrees of Freedom" on page 5714 for definitions of the statistics that appear this table.

The "Output Statistics" table for smoothing parameter value 0.6 is shown in Output 75.1.3. Note that, because the ALL option is specified in the MODEL statement, 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 75.1.3 Output Statistics Table

The LOESS Procedure
Selected Smoothing Parameter: 0.6
Dependent Variable: NOx

			Output Statistics					
Obs	E	NOx	Predicted NOx	Estimated Prediction Std Deviation	Residual	t Value	99% Confidence Limits	
1	0.8	4.8	4.87377	0.15528	-0.05577	-0.36	4.41841	5.32912
2	1.0	2.8	2.81984	0.15380	0.02916	0.19	2.36883	3.27085
3	1.0	3.3	3.48153	0.15187	-0.20653	-1.36	3.03617	3.92689
4	1.0	4.7	4.73249	0.13923	-0.04149	-0.30	4.32419	5.14079
5	0.8	4.3	4.82305	0.15278	-0.56805	-3.72	4.37503	5.27107
6	0.9	5.1	5.18561	0.19337	-0.12161	-0.63	4.61855	5.75266
7	0.7	2.1	2.51120	0.15528	-0.39320	-2.53	2.05585	2.96655
8	0.8	4.6	4.48267	0.15285	0.11933	0.78	4.03444	4.93089
9	1.1	2.3	2.12619	0.16683	0.15981	0.96	1.63697	2.61541
10	1.1	1.0	0.97120	0.18134	-0.00120	-0.01	0.43942	1.50298
11	1.0	4.0	4.09987	0.13477	-0.13487	-1.00	3.70467	4.49507
12	0.9	5.3	5.31258	0.17283	0.03142	0.18	4.80576	5.81940
13	0.8	3.8	3.84572	0.14929	-0.01172	-0.08	3.40794	4.28350
14	0.7	2.0	2.26578	0.16712	-0.27578	-1.65	1.77571	2.75584
15	0.8	5.2	4.58394	0.15363	0.61506	4.00	4.13342	5.03445
16	0.9	5.3	5.24741	0.19319	0.03559	0.18	4.68089	5.81393
17	1.0	3.8	4.16979	0.13478	-0.41779	-3.10	3.77457	4.56502
18	1.2	0.5	0.53059	0.32170	0.00641	0.02	-0.41278	1.47397
19	1.1	1.6	1.83157	0.17127	-0.19157	-1.12	1.32933	2.33380
20	1.0	5.1	4.66733	0.13735	0.38767	2.82	4.26456	5.07010
21	1.0	4.9	4.52385	0.13556	0.41315	3.05	4.12632	4.92139
22	0.7	1.6	1.19888	0.26774	0.36212	1.35	0.41375	1.98401

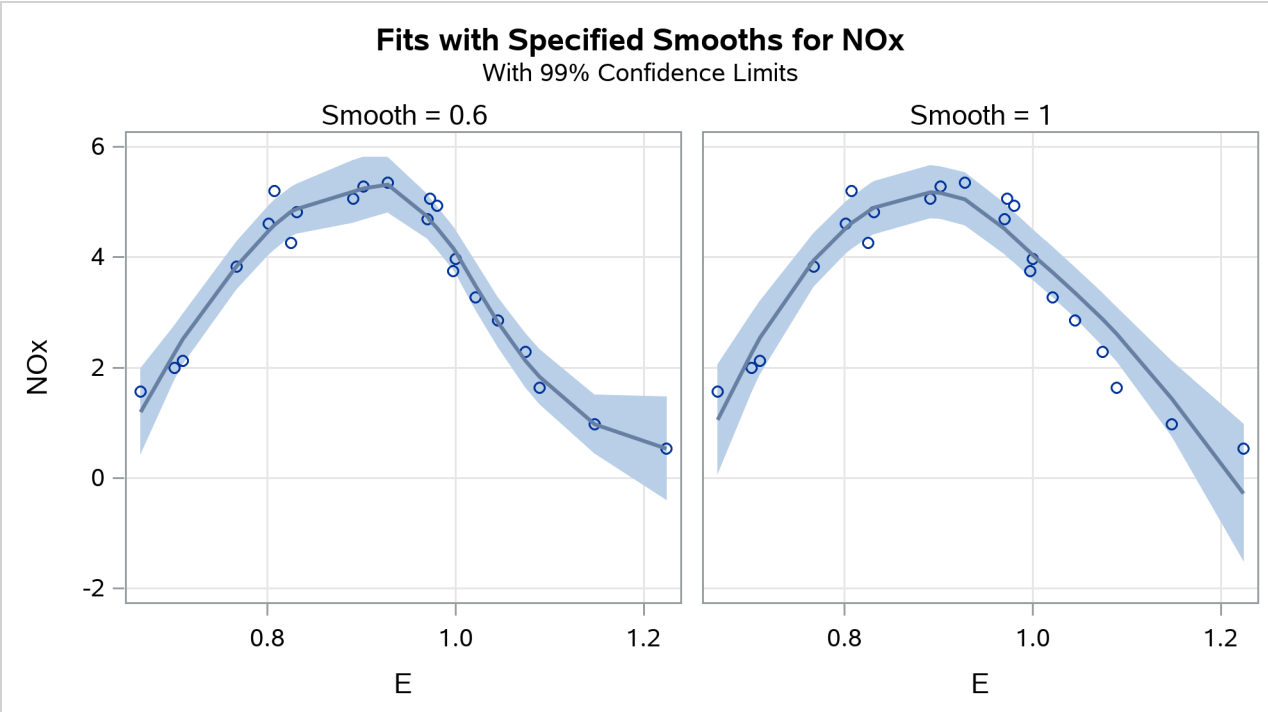
Output 75.1.4 Output Statistics Table

Optimal Smoothing Criterion	
AICC	Smoothing Parameter
-0.45637	0.60000

The combination of the options SELECT=AICC and SMOOTH=0.6 1 in the MODEL statement specifies that PROC LOESS fit models with smoothing parameters of 0.6 and 1 and select the model that yields the smaller value of the AICC statistic. The "Smoothing Criterion" shown in Output 75.1.4 shows that PROC LOESS selects the model with smoothing parameter value 0.6 as it yields the smaller value of the AICC statistic.

With ODS Graphics enabled, PROC LOESS produces a panel of fit plots whenever you specify the SMOOTH= option in the MODEL statement. These fit plots include confidence limits if you additionally specify the CLM option in the MODEL statement.

Output 75.1.5 Loess Fits with 99% Confidence Limits for the Gas Data



Output 75.1.5 shows the "Fit Panel" that displays the fitted models with 99% confidence limits overlaid on scatter plots of the data.

Based on the AICC criterion, the model with smoothing parameter 0.6 is preferred. You can address the question of whether the differences between these models are significant using analysis of variance. You do this by using the model with smoothing parameter value 1 as the null model.

The statistic

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

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 defined in the section "Statistical Inference and Lookup Degrees of Freedom" on page 5714.

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 by 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;
run;

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

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

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

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';
run;

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 75.1.6.

Output 75.1.6 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 small p-value confirms that the fit with smoothing parameter value 0.6 is significantly different from the loess model with smoothing parameter value 1.

Alternatively, you can use the OUTPUT statement to generate the statistics you want to include in the output data set. The following statements produce essentially the same results as the ODS OUTPUT statement does, except all the statistics for each of the two smoothing parameter values are included because the SELECT= option is not specified in the MODEL statement. In addition, with the ROW option specified, the output data set is arranged in rowwise format which enables you to compare statistics side-by-side for a sequence of smoothing values. The ALL option after the slash produces all the statistics (predicted values, residual values, standard errors of the mean predicted values, t statistics, and the lower and upper parts of $100(1-\alpha)\%$ confidence limits on the mean predicted value). All these requested statistics are given their respective default names in the output data set except the predicted value. The P=PREDVAL option causes the name for the predicted value to start with predval.

```
proc loess data=Gas;
   model NOx = E / degree=2 smooth = 0.6 1.0
                   direct alpha=.01;
   output out=GasFit p=predval /all row;
run;
```

Example 75.2: Sulfate Deposits in the U.S. 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 contiguous states.

```
data SO4;
   input Latitude Longitude SO4 @@;
   format Latitude f4.0;
   format Longitude f4.0;
   format SO4 f4.1;
   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
34.17944  93.09861 2.168 36.08389  92.58694 1.578

   ... more lines ...

43.87333 104.19222 0.306 44.91722 110.42028 0.210
45.07611  72.67556 2.646
;
```

As longitudes decrease from west to east in the western hemisphere, the roles of east and west get interchanged if you use these longitudes on the horizontal axis of a plot. You can address this by using negative values to represent longitudes in the western hemisphere. The following statements change the sign of longitude in the SO4 data set and define a format to display these negative values with a suffix of “W”.

```
proc format;
   picture latitude  -90 - 0  = '000S'
                      0  - 90 = '000N';
   picture longitude -180  - 0    = '000W'
                      0    - 180  = '000E';
run;
data SO4;
   set SO4;
   format longitude longitude. latitude  latitude.;
   longitude = -longitude;
run;
```

The following statements use ODS Graphics to plot the locations of the sulfate measurements. The circles indicating the locations are colored using a gradient that denotes the value of SO4.

```
proc template;
   define statgraph gradientScatter;
      beginGraph;
         layout overlay;
            scatterPlot x=longitude y=latitude /
               markercolorgradient = SO4
               markerattrs         = (symbol=circleFilled)
               colormodel          =  ThreeColorRamp
               name                =  "Scatter";
            scatterPlot x=longitude y=latitude /
               markerattrs         = (symbol=circle);
            continuousLegend "Scatter"/title= "SO4";
         endlayout;
      endgraph;
   end;
run;

proc sgrender data=SO4 template=gradientScatter;
run;
```

Output 75.2.1 Sulfate Measurements

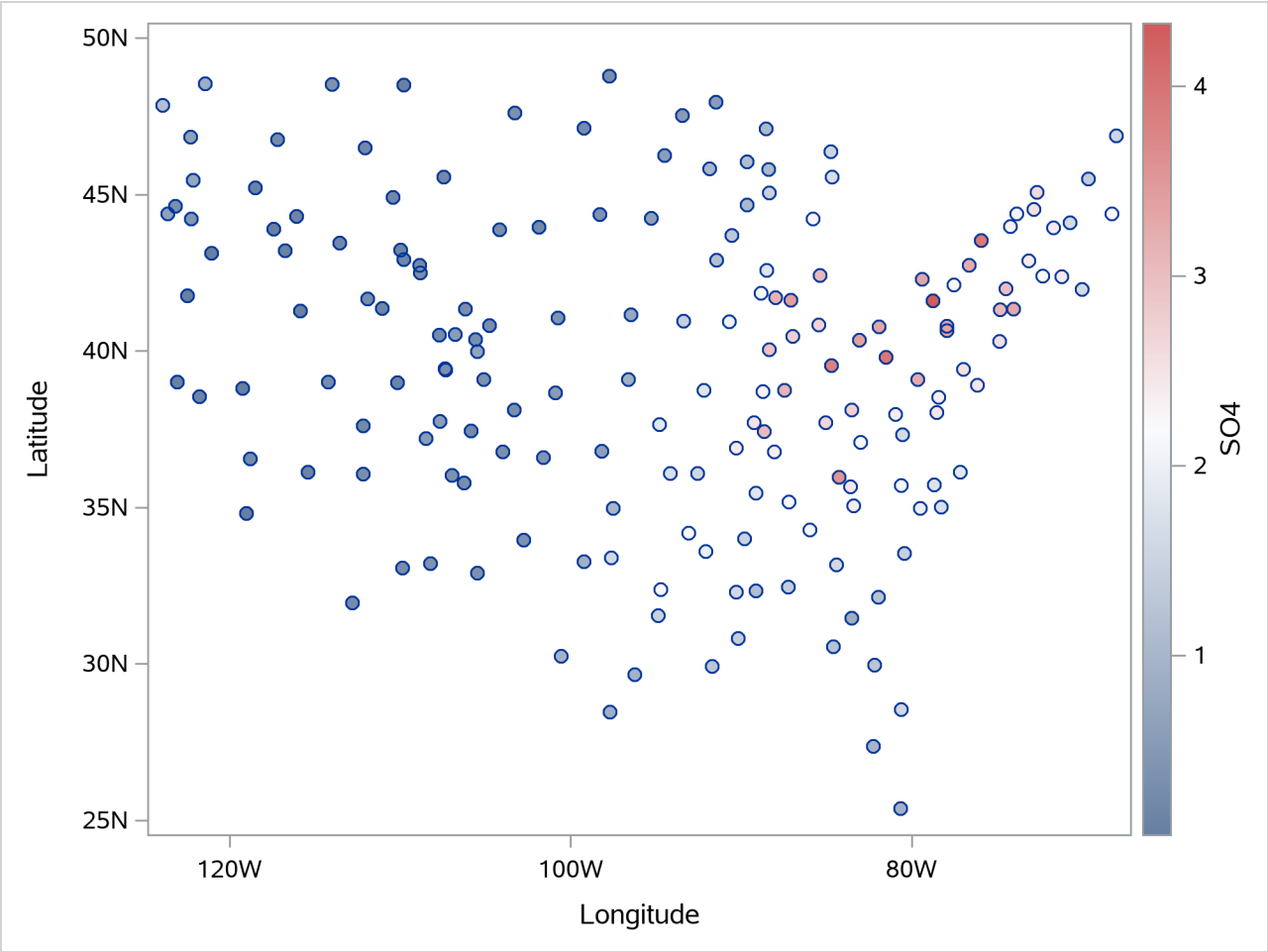


Figure 75.2.1 shows that the largest concentrations of sulfate deposits occur in the northeastern United States. The following statements fit a loess model.

```
ods graphics on;

proc loess data=SO4;
   model SO4=Longitude Latitude / degree=2 interp=cubic;
run;

ods graphics off;
```

Output 75.2.2 Fit Plot for the SO4 Data

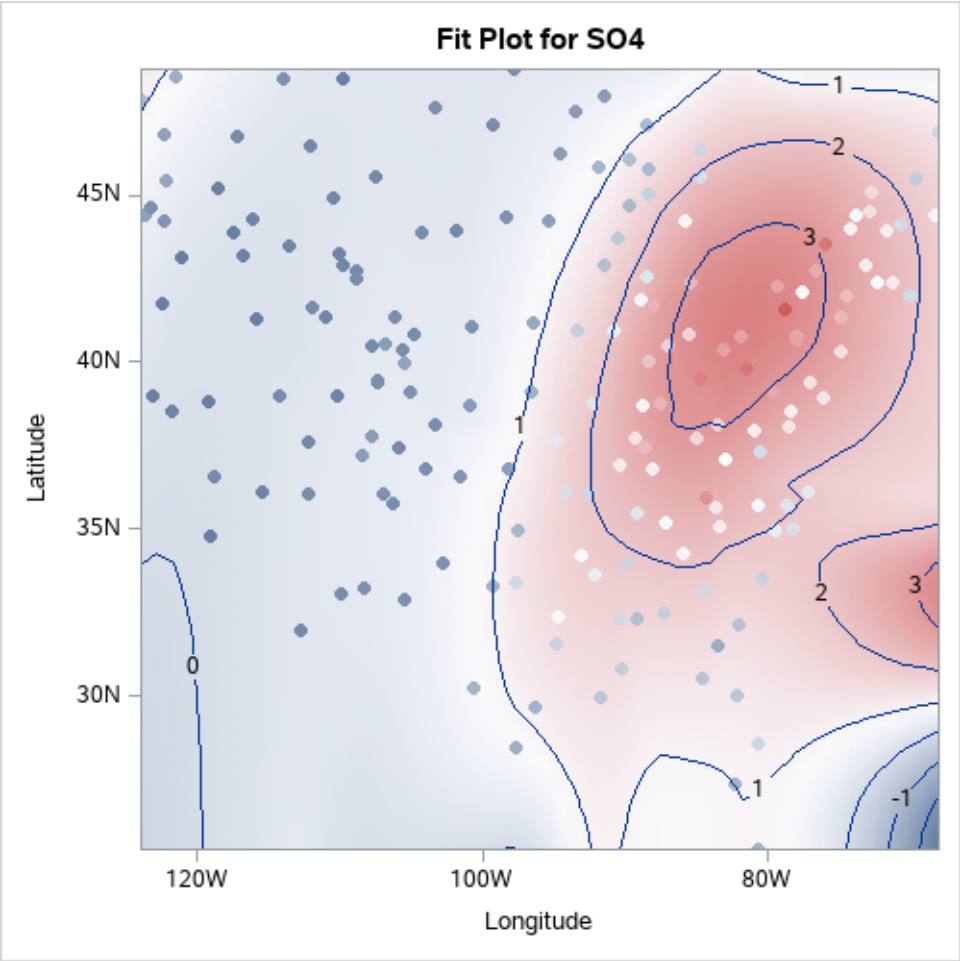


Figure 75.2.2 shows a contour plot of the fitted loess surface overlaid with a scatter plot of the data. The data are colored by the observed sulfate concentrations, using the same color gradient as the gradient-filled contour plot of the fitted surface. Note that for observations where the residual is small, the observations blend in with the contour plot. The greater the size of the residual, the greater the contrast between the observation color and the surface color.

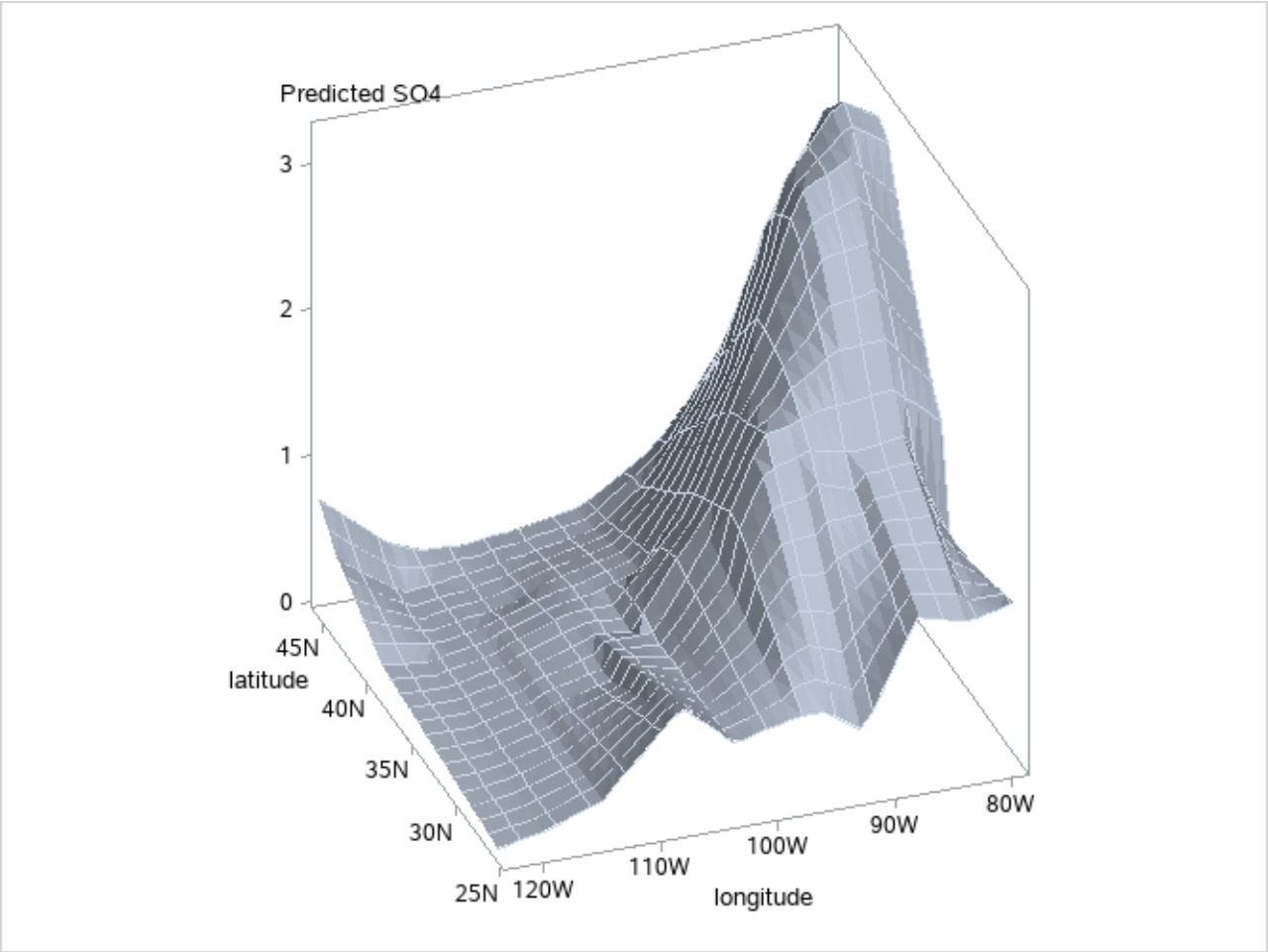
The sulfate measurements are irregularly spaced. To facilitate producing a plot of the fitted loess surface, you can create a data set containing a regular grid of longitudes and latitudes and then use the SCORE statement to evaluate the loess surface at these points. The following statements show how you do this:

```
data PredPoints;
   format longitude longitude.
          latitude  latitude.;
   do Latitude = 26 to 46 by 1;
      do Longitude = -79 to -123 by -1;
         output;
      end;
   end;
run;

proc loess data=SO4;
   model SO4=Longitude Latitude;
   score data=PredPoints / print;
   ods Output ScoreResults=ScoreOut;
run;
```

The PRINT option in the SCORE statement requests that the "Score Results" table be displayed as part of the PROC LOESS output. The ODS OUTPUT statement outputs this table to a data set named ScoreOut. If you do not want to display the score results but you do want the score results in an output data set, then you can omit the PRINT option from the SCORE statement. To plot the surface shown in Figure 75.2.3 by using ODS Graphics, use the following statements:

```
proc template;
   define statgraph surface;
      begingraph;
         layout overlay3d / rotate=340 tilt=30 cube=false;
            surfaceplotparm x=Longitude y=Latitude z=p_SO4;
         endlayout;
      endgraph;
   end;
run;

proc sgrender data=ScoreOut template=surface;
run;
```

Output 75.2.3 Loess Fit of SO4 Surface



Example 75.3: 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.

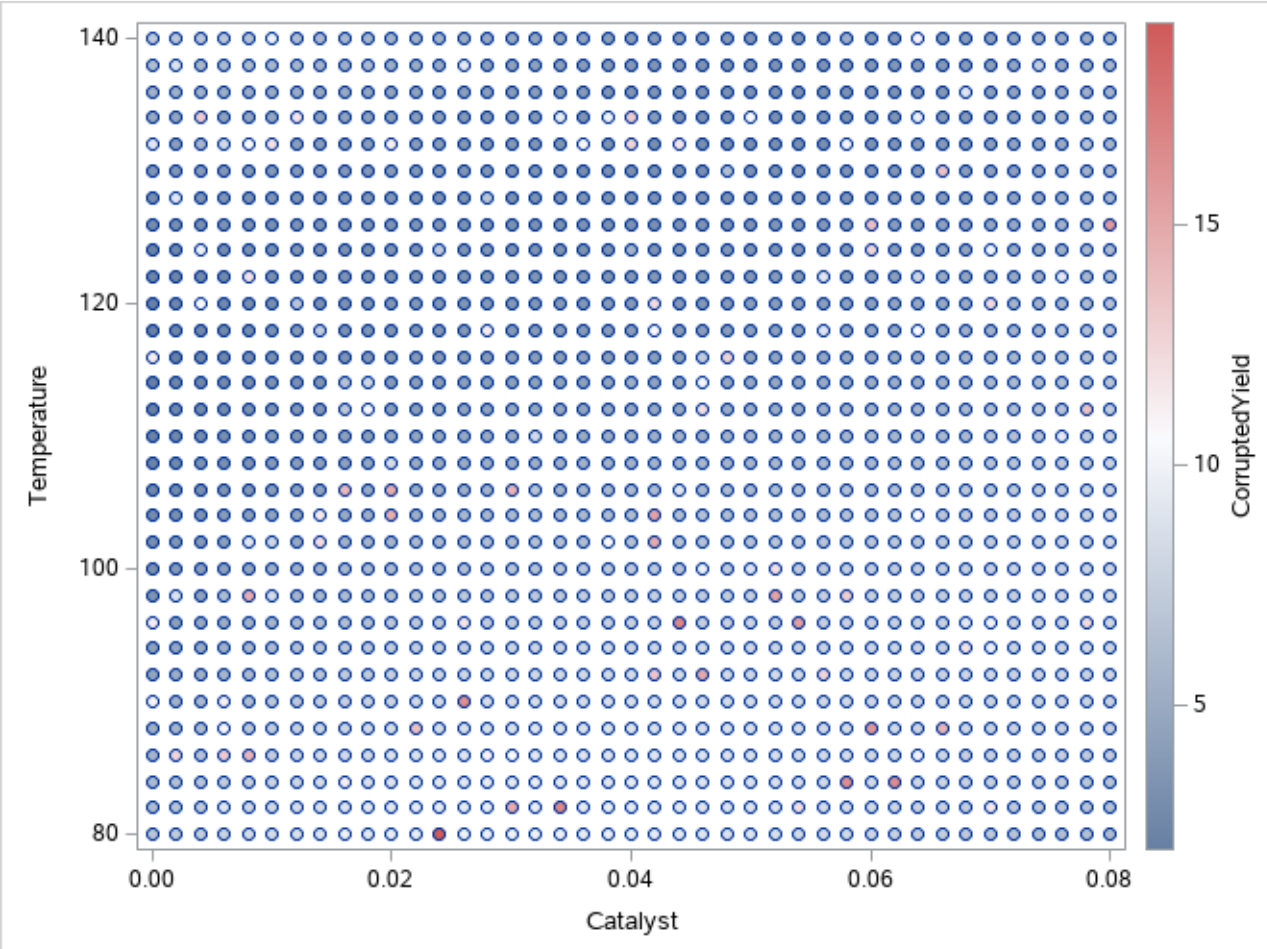
```
data Experiment;
   input Temperature Catalyst MeasuredYield;
   if ranuni(1) < 0.1
      then CorruptedYield = MeasuredYield + 10 * ranuni(1);
      else CorruptedYield = MeasuredYield;
   datalines;
 80    0.000     6.85601
 80    0.002     7.26355
 80    0.004     7.41448
 80    0.006     7.82640

   ... more lines ...

140    0.078     5.20562
140    0.080     5.49371
;
```

The aim of this example is to show how you can use PROC LOESS for robust fitting in the presence of outliers. To simulate an intermittent equipment malfunction, the variable CorruptedYield is the same as the variable MeasuredYield except for about 10% of the observations where an offset has been added. This example shows how you can use PROC LOESS obtain a fit for CorruptedYield that is close to the fit you obtain for MeasuredYield.

The following statements produce a scatter plot of Temperature by Catalyst where the observations are colored by CorruptedYield:

```
proc template;
   define statgraph gradientScatter;
      beginGraph;
         layout overlay;
            scatterPlot x=Catalyst y=Temperature /
                markercolorgradient = CorruptedYield
                markerattrs         = (symbol=circleFilled)
                colormodel          =  ThreeColorRamp
                name                =  "Yield";

            scatterPlot x=Catalyst y=Temperature /
                markerattrs         = (symbol=circle);

            continuousLegend "Yield" / title= "CorruptedYield";
         endlayout;
      endgraph;
   end;
run;

proc sgrender data=Experiment template=gradientScatter;
run;
```

Output 75.3.1 Scatter Plot of Experiment Data Colored by CorruptedYield



Output 75.3.1 shows a scatter plot of the data where the observations are shaded by the value of CorruptedYield. The darkly shaded points that are surrounded by lightly shaded points are points where the simulated incorrect measurements occur.

The following code fits a loess model to the measured data:

```
ods graphics on;

proc loess data=Experiment;
   model MeasuredYield = Temperature Catalyst / scale=sd(0.1);
run;
```

Output 75.3.2 Scale Details for the Experiment Data

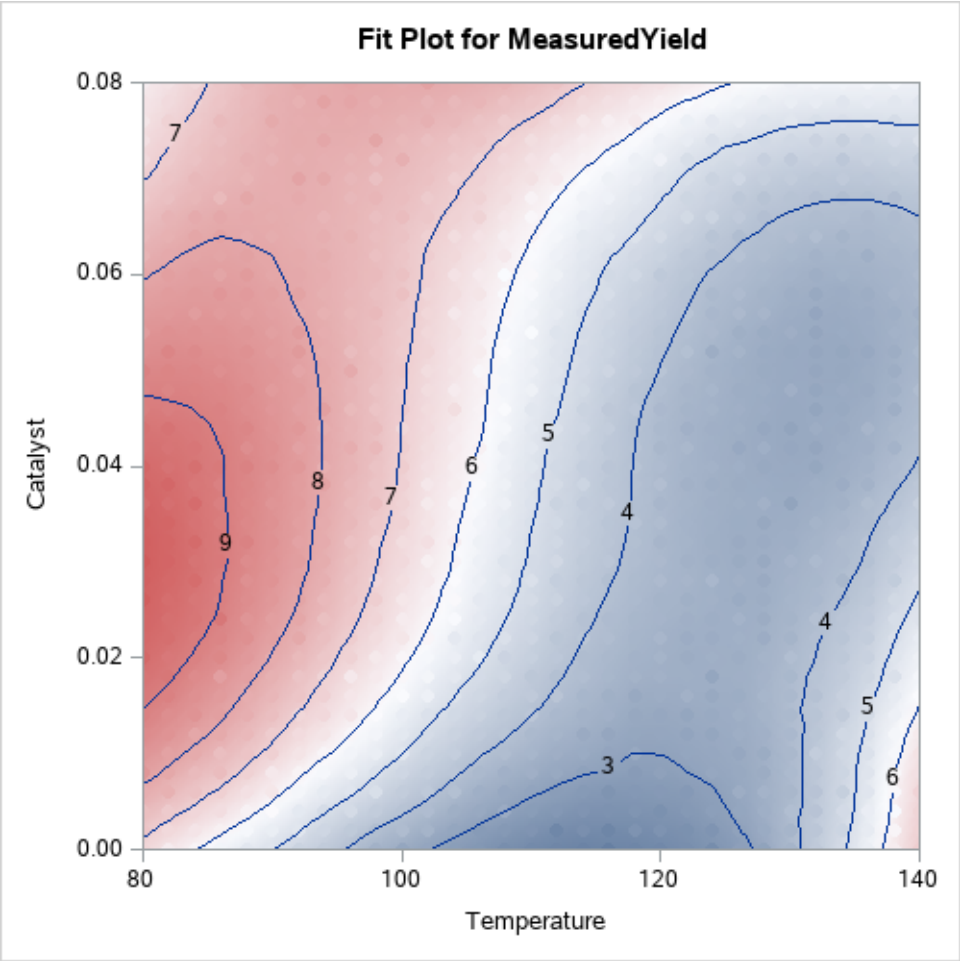
The LOESS Procedure

Independent Variable Scaling		
Scaling applied: 10% trimmed standard deviation		
Statistic	**Temperature**	**Catalyst**
Minimum Value	80.00000	0
Maximum Value	140.00000	0.08000
Trimmed Mean	110.00000	0.04000
Trimmed Standard Deviation	14.32149	0.01894

The SCALE=SD(0.1) option in the MODEL statement 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 because the independent variables Temperature and Catalyst are not similarly scaled. The "Scale Details" table in Output 75.3.2 displays the details of ranges of the regressors and the scale factors applied to each regressor.

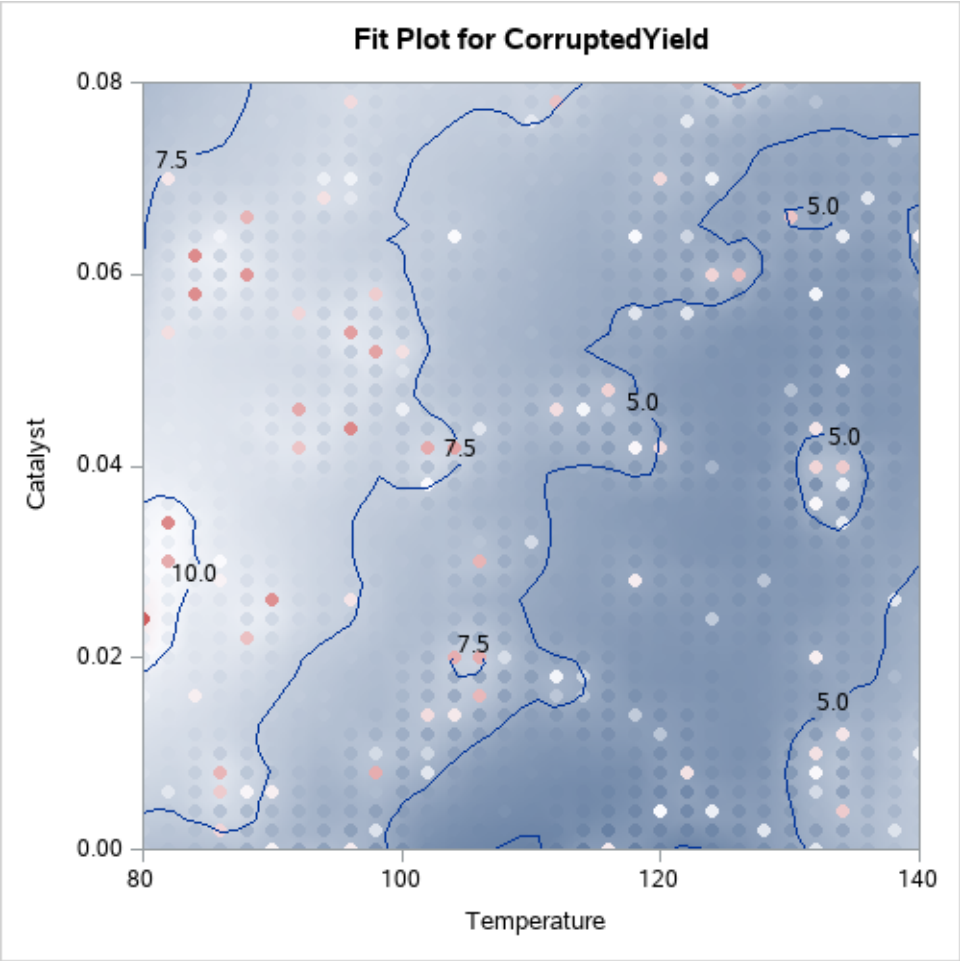
Output 75.3.3 displays the loess fit. Because the fitted surface is a good fit of the observed data, the observations on this plot are not clearly distinguishable from the fitted surface. The results are dramatically different when the outliers are included. The following statements fit a loess model to the corrupted response, using the same smoothing parameter that was selected for the measured response.

Output 75.3.3 Fit for MeasuredYield




```
proc loess data=Experiment;
   model CorruptedYield = Temperature Catalyst /
                  scale=sd(0.1) smooth=0.018;
run;
```

Output 75.3.4 Fit for CorruptedYield

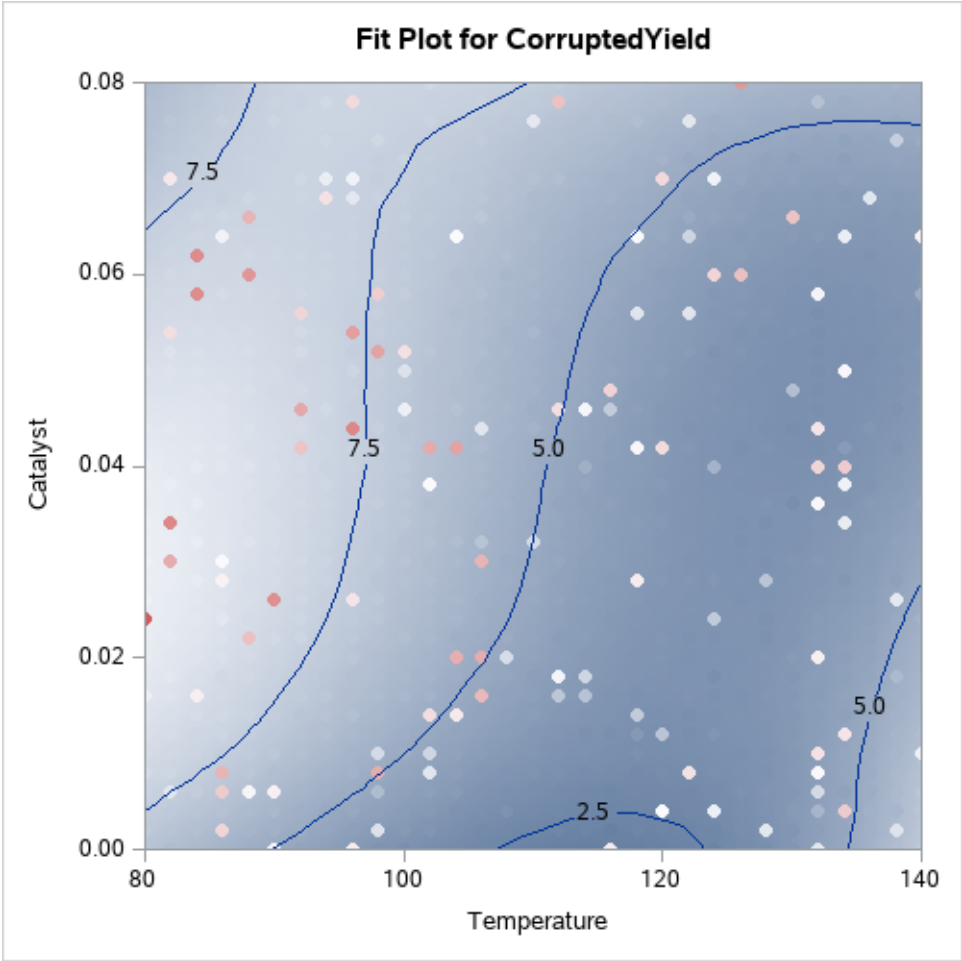


Output 75.3.4 displays the loess fit. The fit is pulled upward in the neighborhoods of these outliers. If you use a larger smoothing parameter value, then these local perturbations in the fit get smoothed out, but at the expense of smoothing away the information in the underlying measured response. In such cases a robust fitting method is indicated. The following statements show how you do this:

```
proc loess data=Experiment;
   model CorruptedYield = Temperature Catalyst /
                          scale  = sd(0.1)
                          smooth = 0.018
                          iterations=4;
run;
```

The ITERATIONS=4 option in the MODEL statement requests the initial loess fit followed by three iteratively reweighted iterations.

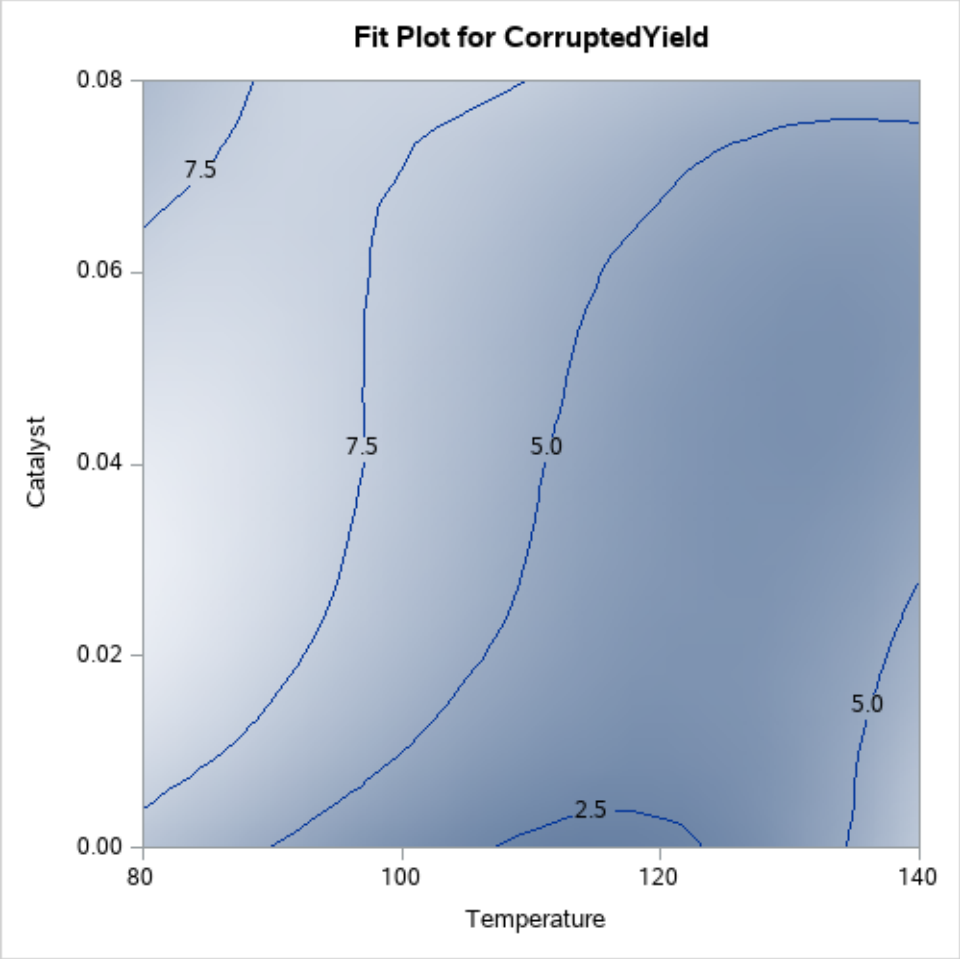
Output 75.3.5 Robust Fit for CorruptedYield



You can see the impact of the robust fitting by comparing the robust fit shown in Output 75.3.5 with the nonrobust fit in Output 75.3.4. In the robust fit you see that the local perturbations caused by the outliers have been eliminated as these the outlying observations get down-weighted during the robustness iterations. By comparing the labeled contours on the fit plot for the uncorrupted response shown in Output 75.3.3 with the labeled contours for the corrupted response shown in Output 75.3.4, you can see that the robust fit has produced a reasonable fit for the underlying measured data. The color gradient in Output 75.3.5 is chosen to accommodate the outliers that are present in the observed data, and so you cannot easily compare the color gradient in this plot with that in Output 75.3.3. The following statements repeat the robust analysis with an option added to suppress the display of the observations on the fit plot:

```
proc loess data=Experiment plots=contourFit(obs=none);
   model CorruptedYield = Temperature Catalyst /
                             scale  = sd(0.1)
                             smooth = 0.018
                             iterations=4;
run;

ods graphics off;
```

Output 75.3.6 Robust Fit for CorruptedYield with Observations Suppressed



Output 75.3.6 shows the robust fit with the observations suppressed. The range of the fitted surface values in this plot is similar to the range in Output 75.3.3. By comparing this contour plot with the contour plot in Output 75.3.3, you clearly see that the robust loess fit has successfully modeled the underlying surface despite the presence of the outliers.

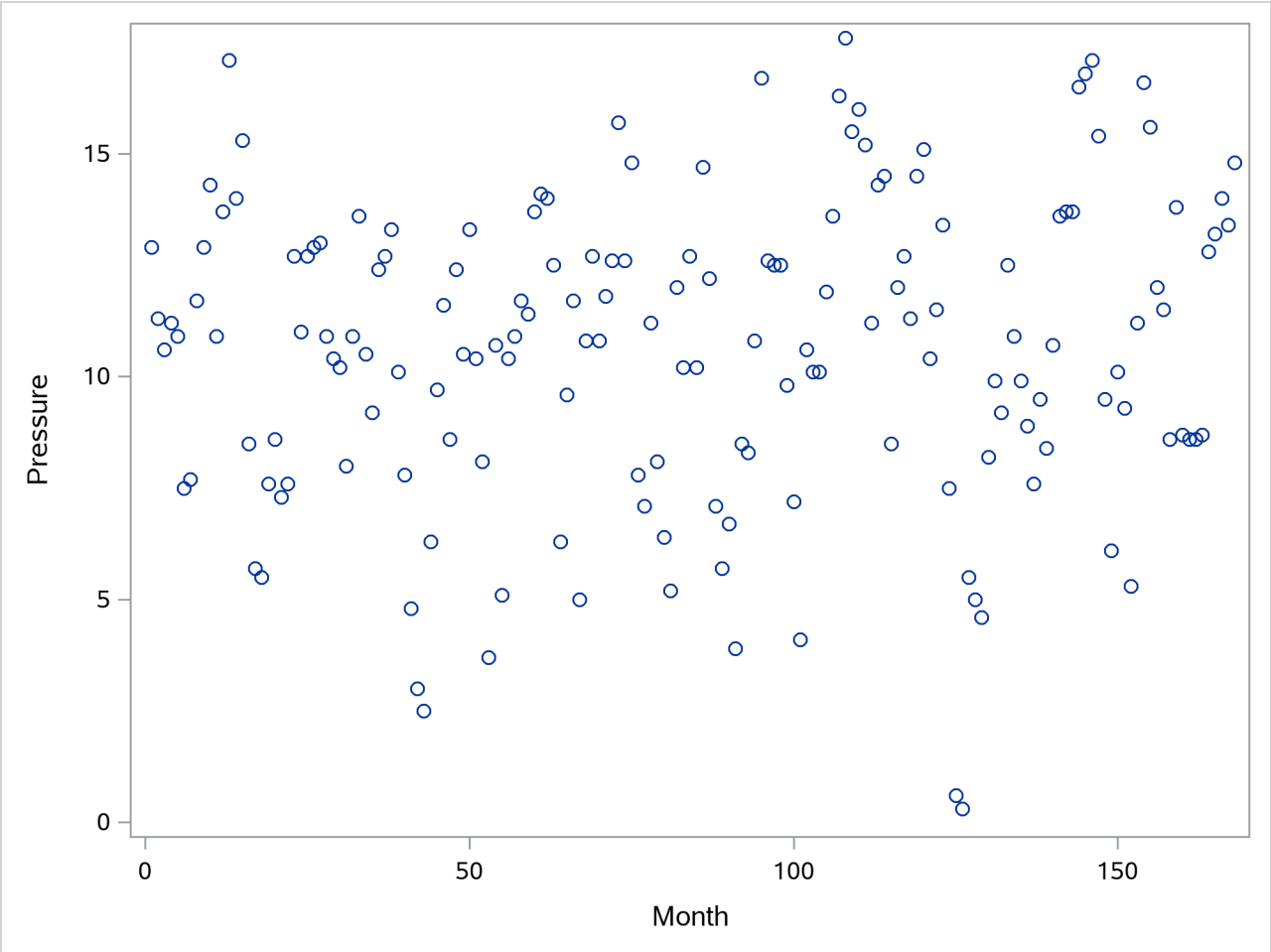
Example 75.4: El Niño Southern Oscillation

The data set sashelp.ENSO, which is available in the Sashelp library, contains measurements of monthly averaged atmospheric pressure differences between Easter Island and Darwin, Australia, for a period of 168 months (National Institute of Standards and Technology 1998).

The following PROC SGPLOT statements produce the simple scatter plot of the ENSO data, displayed in Output 75.4.1.

```
proc sgplot data=sashelp.ENSO;
   scatter y=Pressure x=Month;
run;
```

Output 75.4.1 Scatter Plot of ENSO Data



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

```
ods graphics on;

proc loess data=sashelp.ENSO plots=residuals(smooth);
   model Pressure=Month;
run;
```

The "Smoothing Criterion" and "Fit Summary" tables are shown in Output 75.4.2, and the fit plot is shown in Output 75.4.3.

Output 75.4.2 Output from PROC LOESS

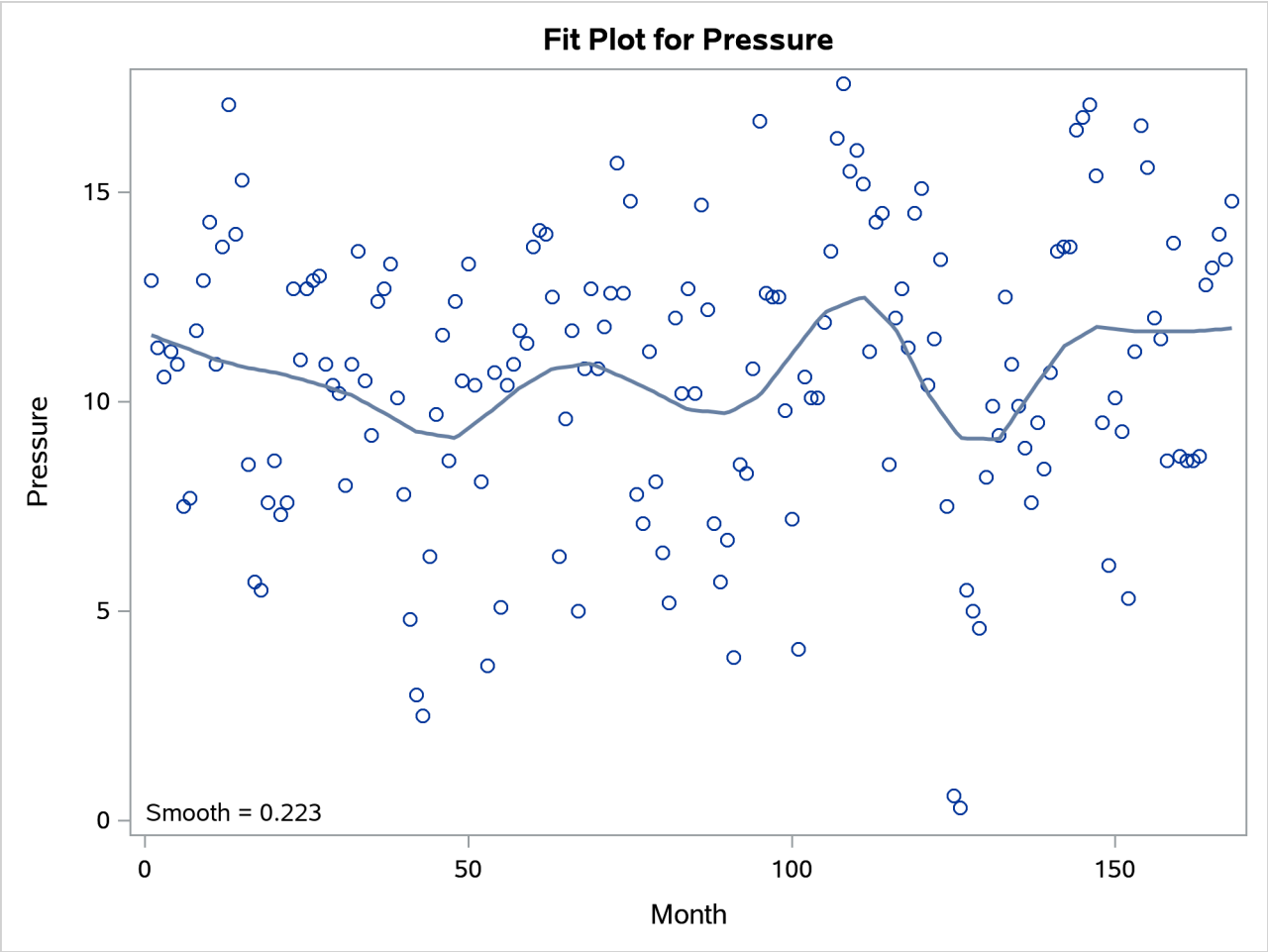
The LOESS Procedure
Dependent Variable: Pressure

Optimal Smoothing Criterion	
AICC	Smoothing Parameter
3.41105	0.22321

The LOESS Procedure
Selected Smoothing Parameter: 0.223
Dependent Variable: Pressure

Fit Summary	
Fit Method	kd Tree
Blending	Linear
Number of Observations	168
Number of Fitting Points	33
kd Tree Bucket Size	7
Degree of Local Polynomials	1
Smoothing Parameter	0.22321
Points in Local Neighborhood	37
Residual Sum of Squares	1654.27725
Trace[L]	8.74180
GCV	0.06522
AICC	3.41105

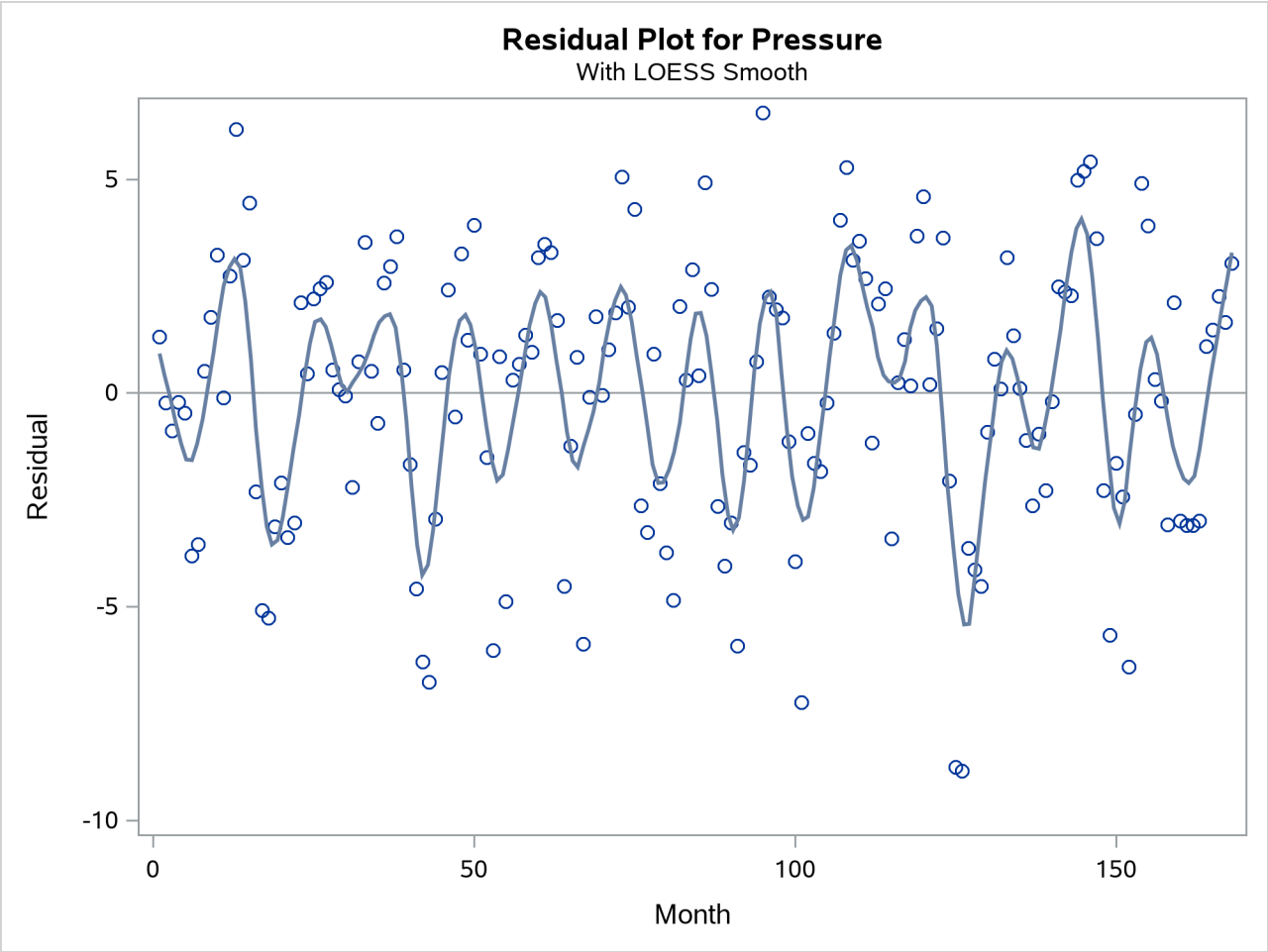
Output 75.4.3 Oversmoothed Loess Fit for the ENSO Data



This weather-related data should exhibit an annual cycle. However, the loess fit in Output 75.4.3 indicates a longer cycle but no annual cycle. This suggests that the loess fit is oversmoothed. One way to detect oversmoothing is to look for patterns in the fit residuals. With ODS Graphics enabled, PROC LOESS produces a scatter plot of the residuals versus each regressor in the model. To aid in visually detecting patterns in these scatter plots, it is useful to superimpose a nonparametric fit on these scatter plots. You can request this by specifying the SMOOTH suboption of the PLOTS=RESIDUALS option in the PROC LOESS statement. The nonparametric fit that is produced is again a loess fit that is produced independently of the loess fit used to obtain these residuals.

With the superimposed loess fit shown in Output 75.4.4, you can clearly identify an annual cycle in the residuals, which confirms that the loess fit for the ENSO is oversmoothed. What accounts for this poor fit?

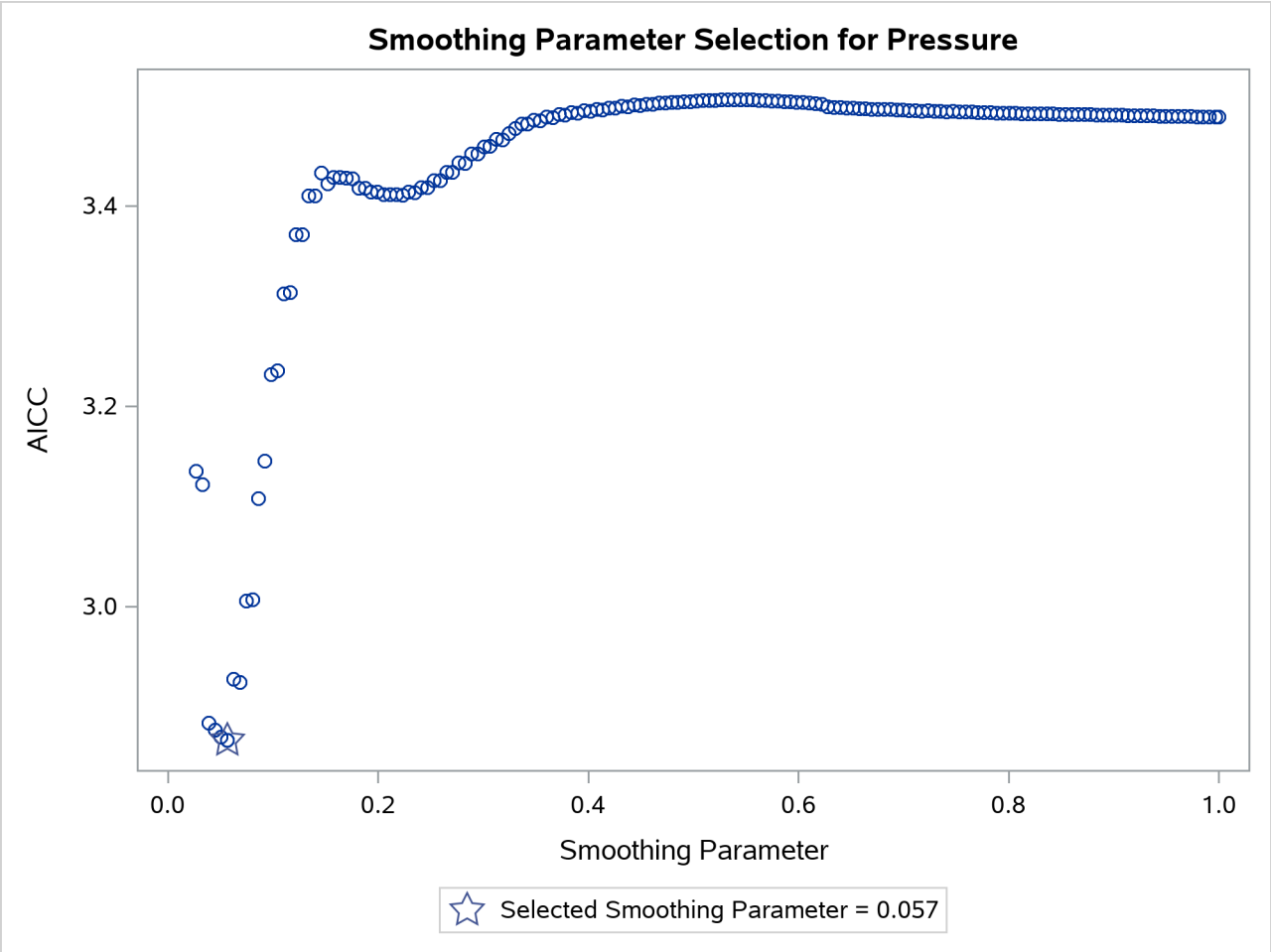
Output 75.4.4 Residuals for the Loess Fit for the ENSO Data



The smoothing parameter value used for the loess fit shown in Output 75.4.3 was chosen using the default method of PROC LOESS, namely a golden section minimization of the AICC criterion over the interval $(0, 1]$. One possibility is that the golden section search has found a local rather than a global minimum of the AICC criterion. You can test this by redoing the fit requesting a global minimum. You do this with the following statements:

```
proc loess data=sashelp.ENSO;
   model Pressure=Month/select=AICC(global);
run;
```

Output 75.4.5 AICC versus Smoothing Parameter Showing Local Minima



The explanation for the oversmoothed fit in Output 75.4.3 is now apparent. Output 75.4.5 shows that the golden section search algorithm found the local minimum that occurs near the value 0.22 of the smoothing parameter rather than the global minimum that occurs near 0.06. Note that if you restrict the range of smoothing parameter values examined to lie below 0.2, then the golden section search finds the global minimum, as the following statements demonstrate:

```
proc loess data=sashelp.ENSO;
   model Pressure=Month/select=AICC(range(0.03,0.2));
run;
```

Output 75.4.6 Selected Smoothing Parameter Value

The LOESS Procedure
Dependent Variable: Pressure

Optimal Smoothing Criterion	
AICC	Smoothing Parameter
2.86660	0.05655

Output 75.4.6 shows that with the restricted range of smoothing parameter values examined, PROC LOESS finds the global minimum of the AICC criterion. Often you might not know an appropriate range of smoothing parameter values to examine. In such cases, you can use the PRESEARCH suboption of the SELECT= option in the MODEL statement. When you specify this option, PROC LOESS does a preliminary search to try to locate a smoothing parameter value range that contains just the first local minimum of the criterion being used for the selection. The following statements provide an example.

```
proc loess data=sashelp.ENSO plots=residuals(smooth);
   model Pressure=Month/select=AICC(presearch);
run;

ods graphics off;
```

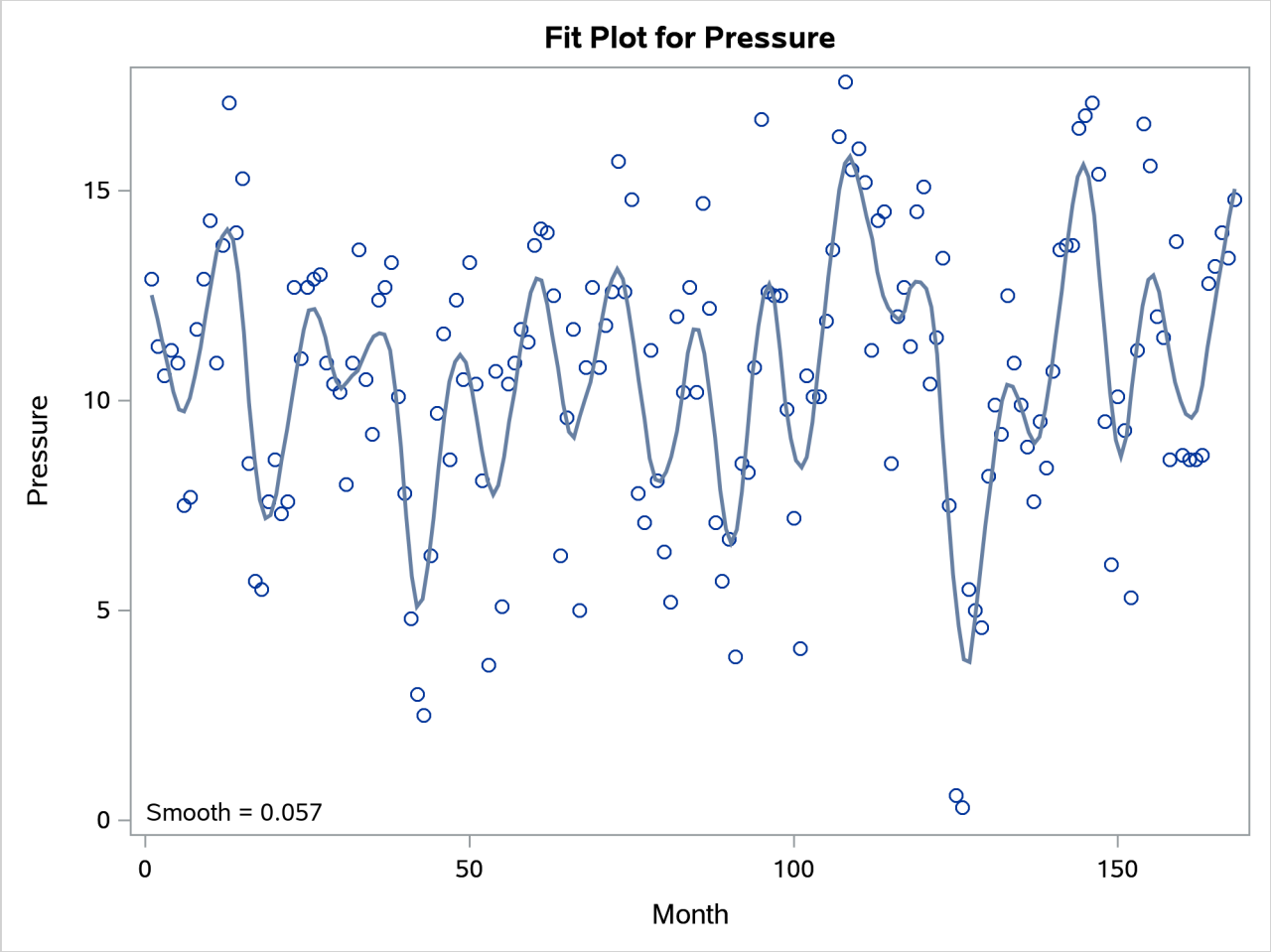
Output 75.4.7 Selected Smoothing Parameter Value When Presearch Is Specified

The LOESS Procedure
Dependent Variable: Pressure

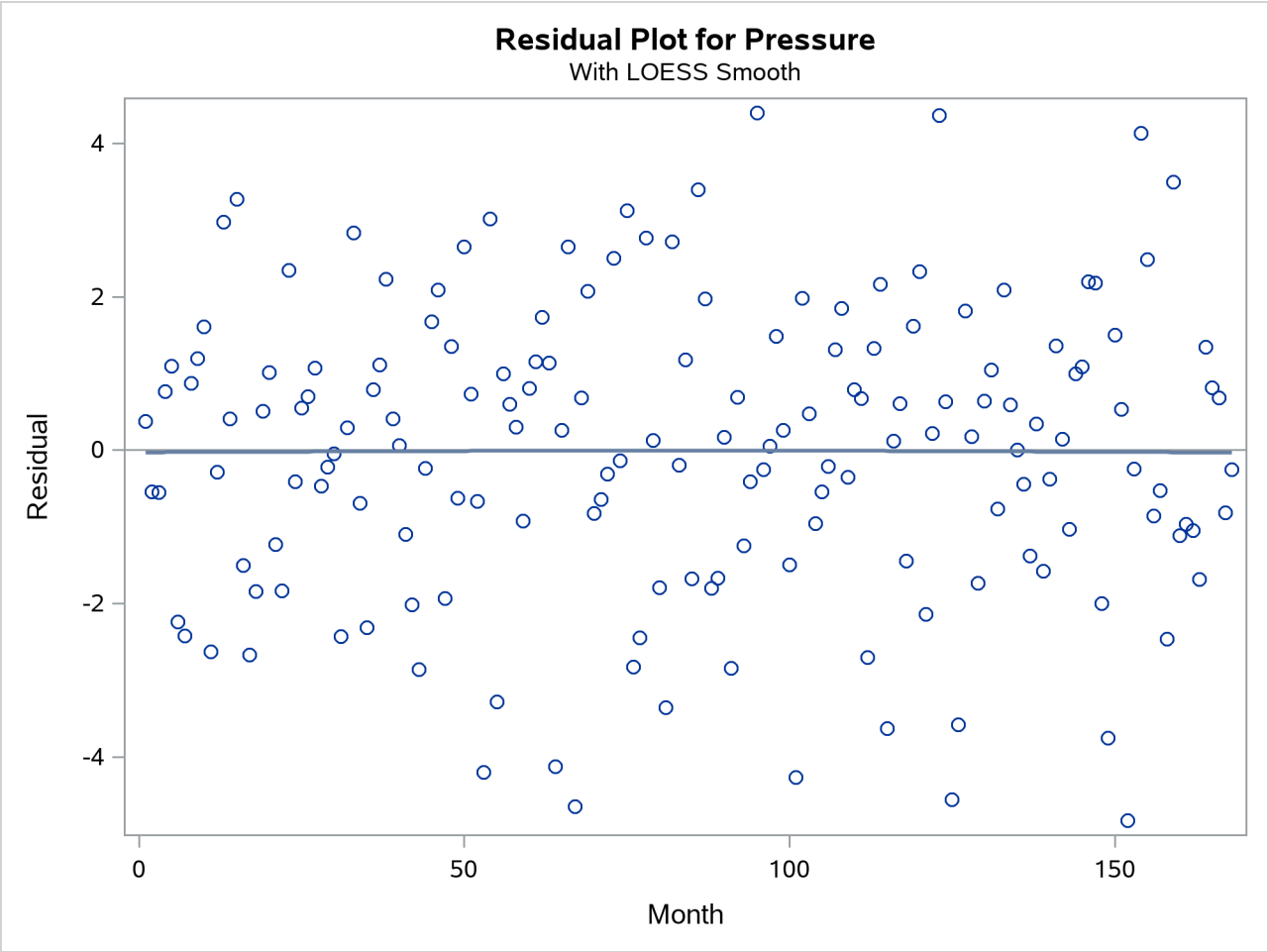
Optimal Smoothing Criterion	
AICC	Smoothing Parameter
2.86660	0.05655

Output 75.4.7 shows that with the PRESEARCH suboption specified, PROC LOESS selects the smoothing parameter value that yields the global minimum of the AICC criterion. The fit obtained is shown in Output 75.4.8, and a plot of the residuals with a superimposed loess fit is shown in Output 75.4.9.

Output 75.4.8 Loess Fit Showing an Annual Cycle



Output 75.4.9 Residuals of the Selected Model



In contrast to the residual plot show in Output 75.4.4, the residuals plotted in Output 75.4.9 do not exhibit any pattern, indicating that the corresponding loess fit has captured all the systematic variation in the data.

An interesting question is whether there is some phenomenon captured in the data that would explain the presence of the local minimum near 0.22 in the AICC curve. Note that there is some evidence of a cycle of about 42 months in the oversmoothed fit in Output 75.4.3. You can see this cycle because the strong annual cycle in Output 75.4.8 has been smoothed out. The physical phenomenon that accounts for the existence of this cycle has been identified as the periodic warming of the Pacific Ocean known as “El Niño.”

References

Akaike, H. (1973). "Information Theory and an Extension of the Maximum Likelihood Principle." In *Proceedings of the Second International Symposium on Information Theory*, edited by B. N. Petrov and F. Csáki, 267–281. Budapest: Akademiai Kiado.

Brinkman, N. D. (1981). "Ethanol Fuel: A Single-Cylinder Engine Study of Efficiency and Exhaust Emissions." *Society of Automotive Engineers Transactions* 90:1410–1424.

Cleveland, W. S. (1993). *Visualizing Data*. Summit, NJ: Hobart Press.

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 Shyu, M.-J. (1992). "A Package of C and Fortran Routines for Fitting Local Regression Models." Unpublished manuscript.

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

Gordon, W. J. (1971). "Blending-Function Methods of Bivariate and Multivariate Interpolation and Approximation." *SIAM Journal on Numerical Analysis* 8:158–177.

Hastie, T. J., and Tibshirani, R. J. (1990). *Generalized Additive Models*. New York: Chapman & Hall.

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

National Institute of Standards and Technology (1998). "Statistical Reference Data Sets." Accessed June 6, 2011. http://www.itl.nist.gov/div898/strd/general/dataarchive.html.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995). "An Effective Bandwidth Selector for Local Least Squares Regression." *Journal of the American Statistical Association* 90:1257–1270.

Subject Index

LOESS procedure
- approximate degrees of freedom, 5718
- automatic smoothing parameter selection, 5715
- data scaling, 5711
- degrees of freedom, 5714
- direct fitting method, 5712
- introductory example, 5683
- iterative reweighting, 5713
- *k*-d trees and blending, 5712
- local polynomials, 5713
- local weighting, 5713
- lookup degrees of freedom, 5714
- missing values, 5709
- ODS graph names, 5720
- ODS Graphics, 5696, 5720
- output data sets, 5709
- output table names, 5719
- scoring data sets, 5719
- smoothing matrix, 5714
- statistical graphics, 5720
- statistical inference, 5714

ODS graph names
- LOESS procedure, 5720

ODS Graphics
- LOESS procedure, 5696, 5720

options summary
- MODEL statement (LOESS), 5701

statistical graphics
- LOESS procedure, 5720

Syntax Index

ALL option
    MODEL statement (LOESS), 5702
    OUTPUT statement (LOESS), 5708
ALPHA= option
    MODEL statement (LOESS), 5702

BUCKET= option
    MODEL statement (LOESS), 5702
BY statement
    LOESS procedure, 5700

CLM option
    MODEL statement (LOESS), 5702
    SCORE statement (LOESS), 5708

DATA= option
    PROC LOESS statement, 5696
DEGREE= option
    MODEL statement (LOESS), 5702
DETAILS option
    MODEL statement (LOESS), 5702
DFMETHOD= option
    MODEL statement (LOESS), 5703
DFMETHOD=APPROX(Cutoff=) option
    MODEL statement (LOESS), 5703
DFMETHOD=APPROX(Quantile=) option
    MODEL statement (LOESS), 5703
DIRECT option
    MODEL statement (LOESS), 5703
DROPSQUARE= option
    MODEL statement (LOESS), 5703

ID statement
    LOESS procedure, 5701
INTERP= option
    MODEL statement (LOESS), 5703
ITERATIONS= option
    MODEL statement (LOESS), 5703

keyword option
    OUTPUT statement (LOESS), 5707

LCLM keyword
    OUTPUT statement (LOESS), 5707
LOESS procedure, BY statement, 5700
LOESS procedure, ID statement, 5701
LOESS procedure, MODEL statement, 5701
    ALL option, 5702
    ALPHA= option, 5702
    BUCKET= option, 5702
    CLM= option, 5702
    DEGREE= option, 5702
    DETAILS option, 5702
    DFMETHOD= option, 5703
    DFMETHOD=APPROX(Cutoff=) option, 5703
    DFMETHOD=APPROX(Quantile=) option, 5703
    DIRECT option, 5703
    DROPSQUARE= option, 5703
    INTERP= option, 5703
    ITERATIONS= option, 5703
    RESIDUAL option, 5704
    SCALE= option, 5704
    SCALEDINDEP option, 5704
    SELECT= option, 5704
    SMOOTH= option, 5706
    STD option, 5706
    T option, 5706
    TRACEL option, 5706
LOESS procedure, OUTPUT statement, 5706
    ALL option, 5708
    keyword option, 5707
    LCLM keyword, 5707
    OUT= option, 5707
    PREDICTED keyword, 5707
    RESIDUAL keyword, 5707
    ROWWISE option, 5708
    STD keyword, 5707
    T keyword, 5707
    UCLM keyword, 5707
LOESS procedure, PROC LOESS statement, 5696
    DATA= option, 5696
    PLOT option, 5696
    PLOTS option, 5696
LOESS procedure, SCORE statement, 5708
    CLM option, 5708
    PRINT option, 5708
    RESIDUAL option, 5709
    SCALEDINDEP option, 5709
    STEPS option, 5709
LOESS procedure, WEIGHT statement, 5709

MODEL statement
    LOESS procedure, 5701

OUT= option
    OUTPUT statement (LOESS), 5707
OUTPUT statement

LOESS procedure, 5706

PLOT option
PROC LOESS statement, 5696
PLOTS option
PROC LOESS statement, 5696
PREDICTED keyword
OUTPUT statement (LOESS), 5707
PRINT option
SCORE statement (LOESS), 5708
PROC LOESS statement, *see* LOESS procedure

RESIDUAL keyword
OUTPUT statement (LOESS), 5707
RESIDUAL option
MODEL statement (LOESS), 5704
SCORE statement (LOESS), 5709
ROWWISE option
OUTPUT statement (LOESS), 5708

SCALE= option
MODEL statement (LOESS), 5704
SCALEDINDEP option
MODEL statement (LOESS), 5704
SCORE statement (LOESS), 5709
SCORE statement
LOESS procedure, 5708
SELECT= option
MODEL statement (LOESS), 5704
SMOOTH= option
MODEL statement (LOESS), 5706
STD keyword
OUTPUT statement (LOESS), 5707
STD option
MODEL statement (LOESS), 5706
STEPS option
SCORE statement (LOESS), 5709

T keyword
OUTPUT statement (LOESS), 5707
T option
MODEL statement (LOESS), 5706
TRACEL option
MODEL statement (LOESS), 5706

UCLM keyword
OUTPUT statement (LOESS), 5707

WEIGHT statement
LOESS procedure, 5709