SAS/STAT® 9.3 User’s Guide
The LOESS Procedure
(Chapter)
Chapter 52
The LOESS Procedure

Contents
Overview: LOESS Procedure .................................................. 3938
  Local Regression and the Loess Method .................................. 3938
Getting Started: LOESS Procedure ............................................ 3939
  Scatter Plot Smoothing ..................................................... 3939
Syntax: LOESS Procedure ..................................................... 3951
  PROC LOESS Statement ...................................................... 3952
  BY Statement ............................................................... 3957
  ID Statement ............................................................... 3957
  MODEL Statement ........................................................... 3957
  SCORE Statement ........................................................... 3963
  WEIGHT Statement .......................................................... 3964
Details: LOESS Procedure ..................................................... 3964
  Missing Values ............................................................. 3964
  Output Data Sets ........................................................... 3965
  Data Scaling ............................................................... 3966
  Direct versus Interpolated Fitting ....................................... 3967
  kd Trees and Blending ...................................................... 3967
  Local Weighting ............................................................ 3968
  Iterative Reweighting ...................................................... 3968
  Specifying the Local Polynomials ...................................... 3969
  Smoothing Matrix .......................................................... 3969
  Model Degrees of Freedom ................................................ 3969
  Statistical Inference and Lookup Degrees of Freedom ............... 3970
  Automatic Smoothing Parameter Selection ................................ 3971
  Sparse and Approximate Degrees of Freedom Computation ........... 3973
  Scoring Data Sets .......................................................... 3974
  ODS Table Names ........................................................... 3975
  ODS Graphics .............................................................. 3975
Examples: LOESS Procedure .................................................. 3977
  Example 52.1: Engine Exhaust Emissions ................................ 3977
  Example 52.2: Sulfate Deposits in the U.S. for 1990 ................. 3983
  Example 52.3: Catalyst Experiment ..................................... 3987
  Example 52.4: El Niño Southern Oscillation ........................... 3995
References ................................................................. 4003
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), and Cleveland, Grosse, and Shyu (1992). The LOESS procedure allows great flexibility because no assumptions about the parametric form of the regression surface are needed.

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

The main features of the LOESS procedure are as follows:

- fits nonparametric models
- supports the use of multidimensional data
- supports multiple dependent variables
- supports both direct and interpolated fitting that uses kd trees
- performs statistical inference
- performs automatic smoothing parameter selection
- performs iterative reweighting to provide robust fitting when there are outliers in the data
- supports graphical displays produced through ODS Graphics

Local Regression and the Loess Method

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

\[
y_i = g(x_i) + \epsilon_i
\]

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

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

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

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

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

---

**Getting Started: 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).

```sas
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 52.1.
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:

```sas
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.
When ODS Graphics is enabled, PROC LOESS produces several default plots. **Figure 52.2** 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 52.3  Fit Summary

| 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 52.3 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 52.4  Smoothing Parameter Selection

<table>
<thead>
<tr>
<th>Optimal SmoothingCriterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoothing Parameter</td>
</tr>
<tr>
<td>AICC</td>
</tr>
<tr>
<td>-1.17277</td>
</tr>
</tbody>
</table>
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 52.5 Model Summary Table**

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

The “Model Summary” table shown in **Figure 52.5** provides information about all the models that PROC LOESS evaluated in choosing the smoothing parameter value.
Figure 52.6 shows the “Criterion Plot” that provides a graphical display of the smoothing parameter selection process.
Figure 52.7 Output Statistics

<table>
<thead>
<tr>
<th>Obs</th>
<th>Year</th>
<th>Incidences</th>
<th>Predicted Incidences</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1936</td>
<td>0.90000</td>
<td>0.76235</td>
<td>0.13765</td>
</tr>
<tr>
<td>2</td>
<td>1937</td>
<td>0.80000</td>
<td>0.88992</td>
<td>-0.08992</td>
</tr>
<tr>
<td>3</td>
<td>1938</td>
<td>0.80000</td>
<td>1.01764</td>
<td>-0.21764</td>
</tr>
<tr>
<td>4</td>
<td>1939</td>
<td>1.30000</td>
<td>1.14303</td>
<td>0.15697</td>
</tr>
<tr>
<td>5</td>
<td>1940</td>
<td>1.40000</td>
<td>1.28654</td>
<td>0.11346</td>
</tr>
<tr>
<td>6</td>
<td>1941</td>
<td>1.20000</td>
<td>1.44528</td>
<td>-0.24528</td>
</tr>
<tr>
<td>7</td>
<td>1942</td>
<td>1.70000</td>
<td>1.53482</td>
<td>0.16518</td>
</tr>
<tr>
<td>8</td>
<td>1943</td>
<td>1.80000</td>
<td>1.57895</td>
<td>0.22105</td>
</tr>
<tr>
<td>9</td>
<td>1944</td>
<td>1.60000</td>
<td>1.62058</td>
<td>-0.02058</td>
</tr>
<tr>
<td>10</td>
<td>1945</td>
<td>1.50000</td>
<td>1.68627</td>
<td>-0.18627</td>
</tr>
<tr>
<td>11</td>
<td>1946</td>
<td>1.50000</td>
<td>1.82449</td>
<td>-0.32449</td>
</tr>
<tr>
<td>12</td>
<td>1947</td>
<td>2.00000</td>
<td>2.04976</td>
<td>-0.04976</td>
</tr>
<tr>
<td>13</td>
<td>1948</td>
<td>2.50000</td>
<td>2.30981</td>
<td>0.19019</td>
</tr>
<tr>
<td>14</td>
<td>1949</td>
<td>2.70000</td>
<td>2.53653</td>
<td>0.16347</td>
</tr>
<tr>
<td>15</td>
<td>1950</td>
<td>2.90000</td>
<td>2.68921</td>
<td>0.21079</td>
</tr>
<tr>
<td>16</td>
<td>1951</td>
<td>2.50000</td>
<td>2.70779</td>
<td>-0.20779</td>
</tr>
<tr>
<td>17</td>
<td>1952</td>
<td>3.10000</td>
<td>2.64837</td>
<td>0.45163</td>
</tr>
<tr>
<td>18</td>
<td>1953</td>
<td>2.40000</td>
<td>2.61468</td>
<td>-0.21468</td>
</tr>
<tr>
<td>19</td>
<td>1954</td>
<td>2.20000</td>
<td>2.58792</td>
<td>-0.38792</td>
</tr>
<tr>
<td>20</td>
<td>1955</td>
<td>2.90000</td>
<td>2.57877</td>
<td>0.32123</td>
</tr>
<tr>
<td>21</td>
<td>1956</td>
<td>2.50000</td>
<td>2.71078</td>
<td>-0.21078</td>
</tr>
<tr>
<td>22</td>
<td>1957</td>
<td>2.60000</td>
<td>2.96981</td>
<td>-0.36981</td>
</tr>
<tr>
<td>23</td>
<td>1958</td>
<td>3.20000</td>
<td>3.26005</td>
<td>-0.06005</td>
</tr>
<tr>
<td>24</td>
<td>1959</td>
<td>3.80000</td>
<td>3.54143</td>
<td>0.25857</td>
</tr>
<tr>
<td>25</td>
<td>1960</td>
<td>4.20000</td>
<td>3.73482</td>
<td>0.46518</td>
</tr>
<tr>
<td>26</td>
<td>1961</td>
<td>3.90000</td>
<td>3.78186</td>
<td>0.11814</td>
</tr>
<tr>
<td>27</td>
<td>1962</td>
<td>3.70000</td>
<td>3.74362</td>
<td>-0.04362</td>
</tr>
<tr>
<td>28</td>
<td>1963</td>
<td>3.30000</td>
<td>3.70904</td>
<td>-0.40904</td>
</tr>
<tr>
<td>29</td>
<td>1964</td>
<td>3.70000</td>
<td>3.72917</td>
<td>-0.02917</td>
</tr>
<tr>
<td>30</td>
<td>1965</td>
<td>3.90000</td>
<td>3.82382</td>
<td>0.07618</td>
</tr>
<tr>
<td>31</td>
<td>1966</td>
<td>4.10000</td>
<td>4.00515</td>
<td>0.09485</td>
</tr>
<tr>
<td>32</td>
<td>1967</td>
<td>3.80000</td>
<td>4.18573</td>
<td>-0.38573</td>
</tr>
<tr>
<td>33</td>
<td>1968</td>
<td>4.70000</td>
<td>4.35152</td>
<td>0.34848</td>
</tr>
<tr>
<td>34</td>
<td>1969</td>
<td>4.40000</td>
<td>4.50284</td>
<td>-0.10284</td>
</tr>
<tr>
<td>35</td>
<td>1970</td>
<td>4.80000</td>
<td>4.64413</td>
<td>0.15587</td>
</tr>
<tr>
<td>36</td>
<td>1971</td>
<td>4.80000</td>
<td>4.78291</td>
<td>0.01709</td>
</tr>
<tr>
<td>37</td>
<td>1972</td>
<td>4.80000</td>
<td>4.91602</td>
<td>-0.01602</td>
</tr>
</tbody>
</table>

Figure 52.7 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 52.4 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:

```plaintext
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:

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

**Figure 52.8** PROC PRINT Output of the Results Data Set

<table>
<thead>
<tr>
<th>Obs</th>
<th>Smoothing Parameter</th>
<th>Year</th>
<th>Var</th>
<th>Pred</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1</td>
<td>1936</td>
<td>0.9</td>
<td>0.90000</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.1</td>
<td>1937</td>
<td>0.8</td>
<td>0.80000</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0.1</td>
<td>1938</td>
<td>0.8</td>
<td>0.80000</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0.1</td>
<td>1939</td>
<td>1.3</td>
<td>1.30000</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0.1</td>
<td>1940</td>
<td>1.4</td>
<td>1.40000</td>
<td>0</td>
</tr>
</tbody>
</table>

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 param-
eter 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.
If you examine the plots in Figure 52.9, 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 52.10 Residuals of Loess Fits for a Range of Smoothing Parameters

Figure 52.10 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.

```plaintext
proc loess data=Melanoma plots=ResidualsBySmooth(smooth);
   model Incidences=Year/smooth=0.1 0.25 0.4 0.6;
run;
```
The loess fits shown on the plots in Figure 52.11 help confirm the conclusions obtained when you look at Figure 52.9. 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.

```sas
proc loess data=Melanoma;
    model Incidences=Year/clm alpha=0.1;
run;
ods graphics off;
```
Figure 52.12 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 than 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.
Finally, Figure 52.13 shows the selected loess fit with 90% confidence limits.

Syntax: LOESS Procedure

The following statements are available in PROC LOESS:

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

The PROC LOESS and MODEL statements are required. The BY, WEIGHT, and ID statements are optional. The SCORE statement is optional, and more than one SCORE statement can be used.
The statements used with the LOESS procedure, in addition to the PROC LOESS statement, are as follows.

**BY** specifies variables to define subgroups for the analysis.

**ID** names variables to identify observations in the displayed output.

**MODEL** specifies the dependent and independent variables in the loess model, details and parameters for the computational algorithm, and the required output.

**SCORE** specifies a data set containing observations to be scored.

**WEIGHT** declares a variable to weight observations.

---

**PROC LOESS Statement**

```
PROC LOESS <options> ;
```

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) > ) >**

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 requesting plots. 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 609 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 52.1  Default Graphs Produced

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

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 SELECTION= 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 with PROC LOESS to obtain separate analyses on 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 for 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

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 52.2 lists the options available in the MODEL statement.
Chapter 52: The LOESS Procedure

Table 52.2  Summary of MODEL Statement Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fit Options</td>
<td></td>
</tr>
<tr>
<td>BUCKET=</td>
<td>specifies the number of points in kd tree buckets</td>
</tr>
<tr>
<td>DEGREE=</td>
<td>specifies the degree of local polynomials (1 or 2)</td>
</tr>
<tr>
<td>DFMETHOD=</td>
<td>specifies the method of computing lookup degrees of freedom</td>
</tr>
<tr>
<td>DIRECT</td>
<td>specifies direct fitting at every data point</td>
</tr>
<tr>
<td>DROPSQUARE=</td>
<td>specifies the variables whose squares are to be dropped from local quadratic polynomials</td>
</tr>
<tr>
<td>INTERP=</td>
<td>specifies the interpolating polynomials (linear or cubic)</td>
</tr>
<tr>
<td>ITERATIONS=</td>
<td>specifies the number of reweighting iterations</td>
</tr>
<tr>
<td>SCALE=</td>
<td>specifies the method used to scale the regressor variables</td>
</tr>
<tr>
<td>SELECT=</td>
<td>specifies that automatic smoothing parameter selection be done</td>
</tr>
<tr>
<td>SMOOTH=</td>
<td>specifies the list of smoothing values</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Output Statistics Table Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
</tr>
<tr>
<td>CLM</td>
</tr>
<tr>
<td>RESIDUAL</td>
</tr>
<tr>
<td>SCALEDINDEP</td>
</tr>
<tr>
<td>STD</td>
</tr>
<tr>
<td>T</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Other options</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALPHA= number</td>
</tr>
<tr>
<td>DETAILS=</td>
</tr>
<tr>
<td>TRACEL</td>
</tr>
</tbody>
</table>

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

**ALL**

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

**ALPHA=number**

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

**BUCKET=number**

specifies the maximum number of points in the leaf nodes of the kd tree. The default value used is \( s = n/5 \), where \( s \) is a smoothing parameter 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
MODEL Statement ◆ 3959

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 kd 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 kd 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 3973 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 kd tree decomposition of the predictor space followed by blending of the local polynomials to obtain a regression surface.

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

For example,

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

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

**INTERP=LINEAR | CUBIC**
specifies the degree of the interpolating polynomials used for blending local polynomial fits at the kd 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).
- **AICCI** specifies the $AIC_{C1}$ 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 $L$ relating the observed and predicted values of the dependent variable. Valid values for the **DFCriterion** are as follows:

- **DF1** specifies $\text{Trace}(L)$.
- **DF2** specifies $\text{Trace}(L^T L)$.
- **DF3** specifies $2\text{Trace}(L) - \text{Trace}(L^T L)$.

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 $|\text{DFCriterion} - \text{target}|$, 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 3971 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 52.4 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>
<thead>
<tr>
<th>Syntax</th>
<th>Search Method</th>
<th>Search Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>default</td>
<td>golden section using AICC</td>
<td>(0, 1)</td>
</tr>
<tr>
<td>SMOOTH=list</td>
<td>no selection</td>
<td>values in list</td>
</tr>
<tr>
<td>SMOOTH=list SELECT=criterion</td>
<td>global</td>
<td>values in list</td>
</tr>
<tr>
<td>SMOOTH=list SELECT=criterion ( RANGE(l, u) )</td>
<td>global</td>
<td>values in list within [l, u]</td>
</tr>
<tr>
<td>SELECT=criterion</td>
<td>golden section</td>
<td>(0, 1)</td>
</tr>
<tr>
<td>SELECT=criterion ( RANGE(l, u) )</td>
<td>golden section</td>
<td>[l, u]</td>
</tr>
<tr>
<td>SELECT=criterion ( GLOBAL )</td>
<td>global</td>
<td>(0, 1)</td>
</tr>
<tr>
<td>SELECT=criterion ( GLOBAL RANGE(l, u) )</td>
<td>global</td>
<td>[l, u]</td>
</tr>
</tbody>
</table>

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 in-
cluded 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.

**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.

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

**CLM**

requests that $100(1 - \alpha)\%$ confidence limits on the mean predicted value be added to the “Score
Results” table. By default the 95% limits are computed; the ALPHA= option in the MODEL statement
can be used to change the 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 depen-
dent 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.

**SCALEINDEP**
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 state-
ment. This enables the procedure to reuse the kd tree for all the dependent variables that appear in the
MODEL statement. If you have multiple dependent variables with different missing value structures for the
same set of independent variables, you 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 3975 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 52.8 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 “kd 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:

```plaintext
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.

```plaintext
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.

```plaintext
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 52.3.2 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.

---

**kd Trees and Blending**

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

Starting from the initial cell, the direction of the longest cell edge is selected as the split direction. The median of this coordinate of the data in the cell is the split value. The data in the starting cell are partitioned into two child cells. The left child consists of all data from the parent cell whose coordinate in the split direction is less than the split value. This procedure is repeated for each child cell that has more than a prespecified number of points, called the bucket size of the kd 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 \(n s/5\), 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 kd tree are the points at which PROC LOESS performs its local fitting. The fitted value at an original data point (or at any other point within the original data cell) is obtained by blending the fitted values at the vertices of the kd tree cell that contains that data point.

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 kd tree cell edge.

In the multivariate case, such univariate interpolating polynomials are computed on each edge of the kd 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 kd tree cell, where \(p\) is the number of regressors.

While the details of the kd tree and the fitted values at the vertices of the kd tree are implementation details that seldom need to be examined, PROC LOESS does provide options for their display. Each kd tree subdivision of the data used by PROC LOESS is placed in the “kdTree” table. The predicted values at the vertices of each kd 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 \(x\) contains the \(s\) fraction of the data points closest to the point \(x\). When \(s \geq 1\), all data points are used.

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

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

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. Refer to Cleveland and Grosse (1991) and Cleveland, Grosse, and Shyu (1992) for details.
Specifying the Local Polynomials

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

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

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

Smoothing Matrix

When no iterative reweighting is done, the “Smoothing Matrix” denoted by $L$ 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

$$\hat{y} = Ly$$

where $y$ is the vector of observed values and $\hat{y}$ is the corresponding vector of predicted values of the dependent variable. Note that $L$ is an $n$ by $n$ matrix, where $n$ is the number of observations in the analysis. PROC LOESS does not explicitly form $L$ 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:

$$
\begin{align*}
\text{DF1} & \equiv \text{Trace}(L) \\
\text{DF2} & \equiv \text{Trace}(L^T L) \\
\text{DF3} & \equiv 2\text{Trace}L - \text{Trace}(L^T L)
\end{align*}
$$
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 $i$th 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 $\epsilon_i$ 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 $\epsilon_i$ 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 $\rho$ degrees of freedom where

$$\delta_1 = \text{Trace}(I - L)^T(I - L)$$

$$\delta_2 = \text{Trace}\left((I - L)^T(I - L)\right)^2$$

$$\rho = \text{Lookup Degrees of Freedom}$$

$$= \delta_1^2/\delta_2$$

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

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

The determination of $\rho$ 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 $\delta_1$, $\delta_2$, and $\rho$ 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 $L$ (see the section “Smoothing Matrix” on page 3969).

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 $AICC_1$ and $AICC$ in Hurvich, Simonoff, and Tsai (1998), and generalized cross validation, denoted by GCV.

The relevant formulas are

$$AICC_1 = n \log(\hat{\sigma}^2) + n \frac{\delta_1/\delta_2(n + \nu_1)}{\delta_1^2/\delta_2 - 2}$$

$$AICC = \log(\hat{\sigma}^2) + 1 + \frac{2 (\text{Trace}(L) + 1)}{n - \text{Trace}(L) - 2}$$

$$GCV = \frac{n \hat{\sigma}^2}{(n - \text{Trace}(L))^2}$$

where $n$ is the number of observations and

$$\delta_1 \equiv \text{Trace}(I - L)^T (I - L)$$

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

$$\nu_1 \equiv \text{Equivalent Number of Parameters}$$

$$= \text{Trace}(L^T L)$$

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 3969. 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 52.4 Minimization Criteria

<table>
<thead>
<tr>
<th>Syntax</th>
<th>Minimization Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>SELECT=DF1(target)</td>
<td></td>
</tr>
<tr>
<td>SELECT=DF2(target)</td>
<td></td>
</tr>
<tr>
<td>SELECT=DF3(target)</td>
<td></td>
</tr>
</tbody>
</table>

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:

```plaintext
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]:

```plaintext
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]:

```plaintext
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, ..., 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 $L$ only through its trace. In the direct method, this trace can be computed efficiently. In the interpolated method that uses kd 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 $\delta_1$, $\delta_2$, and $v_1$ that appear in the $AIC_{C_1}$ criterion, and the DF2 and DF3 criteria, depend on the entire $L$ 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 3970, obtaining confidence limits in loess models requires the computation of the lookup degrees of freedom. This in turn requires the computation of

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

where $L$ is the loess smoothing matrix (see the section “Smoothing Matrix” on page 3969).

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 $\delta_2$ 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 $L$ whose $(i, j)$th entry is nonzero only if the $i$th and $j$th 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 $\delta_2$ 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 $\delta_2$ 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 90th 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 $\delta_2$ 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 kd tree containing this point and then interpolating the scored value from the predicted values at the vertices of this cell. This methodology precludes scoring any points that are not contained in the box that surrounds the data used in fitting the loess model.
ODS Table Names

PROC LOESS assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in the following table. For more information about ODS, see Chapter 20, “Using the Output Delivery System.”

Table 52.5  ODS Tables Produced by PROC LOESS

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

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, with the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section “Enabling and Disabling ODS Graphics” on page 609 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 608 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 52.6, along with the relevant PLOTS= options.
### Table 52.6  Graphs Produced by PROC LOESS

<table>
<thead>
<tr>
<th>ODS Graph Name</th>
<th>Plot Description</th>
<th>PLOTS Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>ContourFitPanel</td>
<td>Panel of loess contour surfaces overlaid on scatter plots of data</td>
<td>CONTOURFITPANEL</td>
</tr>
<tr>
<td>ContourFit</td>
<td>Loess contour surface overlaid on scatter plot of data</td>
<td>CONTOURFITPANEL</td>
</tr>
<tr>
<td>DiagnosticsPanel</td>
<td>Panel of fit diagnostics</td>
<td>DIAGNOSTICS</td>
</tr>
<tr>
<td>FitPanel</td>
<td>Panel of loess curves overlaid on scatter plots of data</td>
<td>FITPANEL</td>
</tr>
<tr>
<td>FitPlot</td>
<td>Loess curve overlaid on scatter plot of data</td>
<td>FIT</td>
</tr>
<tr>
<td>ObservedByPredicted</td>
<td>Dependent variable versus loess fit</td>
<td>OBSERVEDBYPREDICTED</td>
</tr>
<tr>
<td>QQPlot</td>
<td>Normal quantile plot of residuals</td>
<td>QQPLOT</td>
</tr>
<tr>
<td>ResidualsBySmooth</td>
<td>Panel of residuals versus regressor by smoothing parameter values</td>
<td>RESIDUALSBYSMOOTH</td>
</tr>
<tr>
<td>ResidualByPredicted</td>
<td>Residuals versus loess fit</td>
<td>RESIDUALBYPREDICTED</td>
</tr>
<tr>
<td>ResidualHistogram</td>
<td>Histogram of fit residuals</td>
<td>RESIDUALHISTOGRAM</td>
</tr>
<tr>
<td>ResidualPanel</td>
<td>Panel of residuals versus regressors for fixed smoothing parameter value</td>
<td>RESIDUALS</td>
</tr>
<tr>
<td>ResidualPlot</td>
<td>Plot of residuals versus regressor</td>
<td>RESIDUALS</td>
</tr>
<tr>
<td>RFPlot</td>
<td>Side-by-side plots of quantiles of centered fit and residuals</td>
<td>RFPLOT</td>
</tr>
<tr>
<td>ScorePlot</td>
<td>Loess fit evaluated at scoring points</td>
<td>SCOREPLOT</td>
</tr>
<tr>
<td>CriterionPlot</td>
<td>Selection criterion versus smoothing parameter</td>
<td>CRITERION</td>
</tr>
</tbody>
</table>
Example 52.1: 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.

```sas
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 52.1.1.

```sas
proc sgplot data=Gas;
  scatter x=E y=NOx;
run;
```
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 52.1.2 Fit Summary Table

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

Fit Summary

<table>
<thead>
<tr>
<th>Fit Method</th>
<th>Direct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Observations</td>
<td>22</td>
</tr>
<tr>
<td>Degree of Local Polynomials</td>
<td>2</td>
</tr>
<tr>
<td>Smoothing Parameter</td>
<td>0.60000</td>
</tr>
<tr>
<td>Points in Local Neighborhood</td>
<td>13</td>
</tr>
<tr>
<td>Residual Sum of Squares</td>
<td>1.71852</td>
</tr>
<tr>
<td>Trace[L]</td>
<td>6.42184</td>
</tr>
<tr>
<td>GCV</td>
<td>0.00708</td>
</tr>
<tr>
<td>AICC</td>
<td>-0.45637</td>
</tr>
<tr>
<td>AICCI</td>
<td>-9.39715</td>
</tr>
<tr>
<td>Delta1</td>
<td>15.12582</td>
</tr>
<tr>
<td>Delta2</td>
<td>14.73089</td>
</tr>
<tr>
<td>Equivalent Number of Parameters</td>
<td>5.96950</td>
</tr>
<tr>
<td>Lookup Degrees of Freedom</td>
<td>15.53133</td>
</tr>
<tr>
<td>Residual Standard Error</td>
<td>0.33707</td>
</tr>
</tbody>
</table>

The “Fit Summary” table for smoothing parameter value 0.6, shown in Output 52.1.2, records the fitting parameters specified and some overall fit statistics. See the section “Smoothing Matrix” on page 3969 for a definition of the smoothing matrix $L$, and the sections “Model Degrees of Freedom” on page 3969 and “Statistical Inference and Lookup Degrees of Freedom” on page 3970 for definitions of the statistics that appear this table.

The “Output Statistics” table for smoothing parameter value 0.6 is shown in Output 52.1.3. 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.
Chapter 52: The LOESS Procedure

Output Statistics Table

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

Output Statistics

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

Output Statistics

<table>
<thead>
<tr>
<th>Obs</th>
<th>99% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.41841 - 5.32912</td>
</tr>
<tr>
<td>2</td>
<td>2.36883 - 3.27085</td>
</tr>
<tr>
<td>3</td>
<td>3.03617 - 3.92689</td>
</tr>
<tr>
<td>4</td>
<td>4.32419 - 5.14079</td>
</tr>
<tr>
<td>5</td>
<td>4.37503 - 5.27107</td>
</tr>
<tr>
<td>6</td>
<td>4.61855 - 5.75266</td>
</tr>
<tr>
<td>7</td>
<td>2.05585 - 2.96655</td>
</tr>
<tr>
<td>8</td>
<td>4.03444 - 4.93089</td>
</tr>
<tr>
<td>9</td>
<td>1.63697 - 2.61541</td>
</tr>
<tr>
<td>10</td>
<td>0.43942 - 1.50298</td>
</tr>
<tr>
<td>11</td>
<td>3.70467 - 4.49507</td>
</tr>
<tr>
<td>12</td>
<td>4.80576 - 5.81940</td>
</tr>
<tr>
<td>13</td>
<td>3.40794 - 4.28350</td>
</tr>
<tr>
<td>14</td>
<td>1.77571 - 2.75584</td>
</tr>
<tr>
<td>15</td>
<td>4.13342 - 5.03445</td>
</tr>
<tr>
<td>16</td>
<td>4.68089 - 5.81393</td>
</tr>
<tr>
<td>17</td>
<td>3.77457 - 4.56502</td>
</tr>
<tr>
<td>18</td>
<td>-0.41278 - 1.47397</td>
</tr>
<tr>
<td>19</td>
<td>1.32933 - 2.33380</td>
</tr>
<tr>
<td>20</td>
<td>4.26456 - 5.07010</td>
</tr>
<tr>
<td>21</td>
<td>4.12632 - 4.92139</td>
</tr>
<tr>
<td>22</td>
<td>0.41375 - 1.98401</td>
</tr>
</tbody>
</table>
Output 52.1.4  Output Statistics Table

<table>
<thead>
<tr>
<th>Optimal Smoothing Criterion</th>
<th>Smoothing Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>AICC</td>
<td>0.60000</td>
</tr>
<tr>
<td>-0.45637</td>
<td>0.60000</td>
</tr>
</tbody>
</table>

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 52.1.4 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 52.1.5  Loess Fits with 99% Confidence Limits for the Gas Data

Output 52.1.5 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
\[ F = \frac{(\text{rss}^{(n)} - \text{rss})/(\delta_1^{(n)} - \delta_1)}{\frac{\text{rss}}{\delta_1}} \]
has a distribution that is well approximated by an \( F \) distribution with \( v = \frac{(\delta_1^{(n)} - \delta_1)^2}{\delta_2^{(n)} - \delta_2} \) numerator degrees of freedom and \( \rho \) 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 \( \delta_1, \delta_2, \) and \( \rho \) are defined in the section “Statistical Inference and Lookup Degrees of Freedom” on page 3970.

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:

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

proc transpose data=h0(drop=SmoothingParameter Label1) out=h0;
   data h0(drop=_NAME_); set h0;
      rename Col1 = RSSNull
            Col2 = delta1Null
            Col3 = delta2Null;
   proc transpose data=h1(drop=SmoothingParameter Label1) out=h1;
      data h1(drop=_NAME_); set h1;
         rename Col1 = RSS
            Col2 = delta1
            Col3 = delta2
            Col4 = rho;
   data ftest; merge h0 h1;
      nu = (delta1Null - delta1)**2 / (delta2Null - delta2);
      Numerator = (RSSNull - RSS)/(delta1Null - delta1);
      Denominator = RSS/delta1;
      FValue = Numerator / Denominator;
      PValue = 1 - ProbF(FValue, nu, rho);
      label nu = 'Num DF'
            rho = 'Den DF'
            FValue = 'F Value'
            PValue = 'Pr > F';
   proc print data=ftest label;
      var nu rho Numerator Denominator FValue PValue;
      format nu rho FValue 7.2 PValue 6.4;
   run;
\end{verbatim}

The results are shown in Output 52.1.6.
```
Output 52.1.6  Test ANOVA for Loess Models of Gas Data

<table>
<thead>
<tr>
<th></th>
<th>Obs</th>
<th>Num DF</th>
<th>Den DF</th>
<th>Numerator</th>
<th>Denominator</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2.67</td>
<td>15.53</td>
<td>1.05946</td>
<td>0.11362</td>
<td>9.32</td>
<td>0.0012</td>
</tr>
</tbody>
</table>

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.

Example 52.2: 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.

```sas
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
  ... 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”.

```sas
proc format;
  picture latitude  -90  = '000S'
                   0   = '000N';
  picture longitude -180  = '000W'
                    0   = '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.
Chapter 52: The LOESS Procedure

proc template;
  define statgraph gradientScatter;
  beginGraph;
    beginLayout overlay;
      scatterPlot x=longitude y=latitude /
        markergradient = 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 52.2.1 Sulfate Measurements
Figure 52.2.1 shows that the largest concentrations of sulfate deposits occur in the northeastern United States.

The following statements fit a loess model.

```sas
ods graphics on;

proc loess data=SO4;
    model SO4=Longitude Latitude / degree=2 interp=cubic;
run;
ods graphics off;
```

**Output 52.2.2** Fit Plot for the SO4 Data

![Fit Plot for SO4](image)

Figure 52.2.2 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:

```sas
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 52.2.3 by using ODS Graphics, use the following statements:

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

proc sgrender data=ScoreOut template=surface;
run;
```
Example 52.3: 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
  ... more lines ...
  140 0.078 5.20562
  140 0.080 5.49371
```

Output 52.2.3 Loess Fit of SO4 Surface
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 \texttt{CorruptedYield} is the same as the variable \texttt{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 \texttt{CorruptedYield} that is close to the fit you obtain for \texttt{MeasuredYield}.

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

```sas
proc template;
    define statgraph gradientScatter;
    beginGraph;
        layout overlay;
        scatterPlot x=Catalyst y=Temperature /
            marker = 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 52.3.1 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:

```r
ods graphics on;
proc loess data=Experiment;
   model MeasuredYield = Temperature Catalyst / scale=sd(0.1);
run;
```
Output 52.3.2 Scale Details for the Experiment Data

The LOESS Procedure

Independent Variable Scaling

Scaling applied: 10% trimmed standard deviation

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Temperature</th>
<th>Catalyst</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum Value</td>
<td>80.00000</td>
<td>0</td>
</tr>
<tr>
<td>Maximum Value</td>
<td>140.00000</td>
<td>0.08000</td>
</tr>
<tr>
<td>Trimmed Mean</td>
<td>110.00000</td>
<td>0.04000</td>
</tr>
<tr>
<td>Trimmed Standard Deviation</td>
<td>14.32149</td>
<td>0.01894</td>
</tr>
</tbody>
</table>

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 52.3.2 displays the details of ranges of the regressors and the scale factors applied to each regressor.

Output 52.3.3 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 52.3.3  Fit for MeasuredYield

proc loess data=Experiment;
  model CorruptedYield = Temperature Catalyst /
    scale=sd(0.1) smooth=0.018;
run;
Output 52.3.4 Fit for CorruptedYield

Output 52.3.4 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:

```plaintext
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.
You can see the impact of the robust fitting by comparing the robust fit shown in Output 52.3.5 with the nonrobust fit in Output 52.3.4. 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 52.3.3 with the labeled contours for the corrupted response shown in Output 52.3.4, you can see that the robust fit has produced a reasonable fit for the underlying measured data. The color gradient in Output 52.3.5 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 52.3.3. The following statements repeat the robust analysis with an option added to suppress the display of the observations on the fit plot:

```plaintext
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 52.3.6 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 52.3.3. By comparing this contour plot with the contour plot in Output 52.3.3, you clearly see that the robust loess fit has successfully modeled the underlying surface despite the presence of the outliers.
Example 52.4: 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 52.4.1.

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

Output 52.4.1 Scatter Plot of ENSO Data
You can compute a loess fit and obtain graphical results for these data by using the following statements:

```latex
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 52.4.2, and the fit plot is shown in Output 52.4.3.

**Output 52.4.2** Output from PROC LOESS

<table>
<thead>
<tr>
<th>The LOESS Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent Variable: Pressure</td>
</tr>
</tbody>
</table>

Optimal Smoothing Criterion

<table>
<thead>
<tr>
<th>Smoothing Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>AICC Parameter</td>
</tr>
<tr>
<td>3.41105</td>
</tr>
<tr>
<td>0.22321</td>
</tr>
</tbody>
</table>

The LOESS Procedure

Selected Smoothing Parameter: 0.223

<table>
<thead>
<tr>
<th>Dependent Variable: Pressure</th>
</tr>
</thead>
</table>

Fit Summary

<table>
<thead>
<tr>
<th>Fit Method</th>
<th>kd Tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blending</td>
<td>Linear</td>
</tr>
<tr>
<td>Number of Observations</td>
<td>168</td>
</tr>
<tr>
<td>Number of Fitting Points</td>
<td>33</td>
</tr>
<tr>
<td>kd Tree Bucket Size</td>
<td>7</td>
</tr>
<tr>
<td>Degree of Local Polynomials</td>
<td>1</td>
</tr>
<tr>
<td>Smoothing Parameter</td>
<td>0.22321</td>
</tr>
<tr>
<td>Points in Local Neighborhood</td>
<td>37</td>
</tr>
<tr>
<td>Residual Sum of Squares</td>
<td>1654.27725</td>
</tr>
<tr>
<td>Trace[L]</td>
<td>8.74180</td>
</tr>
<tr>
<td>GCV</td>
<td>0.06522</td>
</tr>
<tr>
<td>AICC</td>
<td>3.41105</td>
</tr>
</tbody>
</table>
This weather-related data should exhibit an annual cycle. However, the loess fit in Output 52.4.3 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 52.4.4, 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?
The smoothing parameter value used for the loess fit shown in Output 52.4.3 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;
```
The explanation for the oversmoothed fit in Output 52.4.3 is now apparent. Output 52.4.5 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:

```sas
proc loess data=sashelp.ENSO;
    model Pressure=Month/select=AICC(range(0.03,0.2));
run;
```
Output 52.4.6  Selected Smoothing Parameter Value

<table>
<thead>
<tr>
<th>The LOESS Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent Variable: Pressure</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Optimal Smoothing</td>
</tr>
<tr>
<td>Criterion</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>AICC Parameter</td>
</tr>
<tr>
<td>Smoothing</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>2.86660 0.05655</td>
</tr>
</tbody>
</table>

Output 52.4.6 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.

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

Output 52.4.7  Selected Smoothing Parameter Value When Presearch Is Specified

<table>
<thead>
<tr>
<th>The LOESS Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent Variable: Pressure</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Optimal Smoothing</td>
</tr>
<tr>
<td>Criterion</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>AICC Parameter</td>
</tr>
<tr>
<td>Smoothing</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>2.86660 0.05655</td>
</tr>
</tbody>
</table>

Output 52.4.7 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 52.4.8, and a plot of the residuals with a superimposed loess fit is shown in Output 52.4.9.
Output 52.4.8  Loess Fit Showing an Annual Cycle
In contrast to the residual plot shown in Output 52.4.4, the residuals plotted in Output 52.4.9 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 52.4.3. You can see this cycle because the strong annual cycle in Output 52.4.8 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


Subject Index

L

LOESS procedure
  approximate degrees of freedom, 3973
  automatic smoothing parameter selection, 3971
  data scaling, 3966
  degrees of freedom, 3969
  direct fitting method, 3967
  introductory example, 3939
  iterative reweighting, 3968
  kd trees and blending, 3967
  local polynomials, 3969
  local weighting, 3968
  lookup degrees of freedom, 3970
  missing values, 3964
  ODS graph names, 3975
  ODS Graphics, 3952, 3975
  output data sets, 3965
  output table names, 3975
  scoring data sets, 3974
  smoothing matrix, 3969
  statistical graphics, 3975
  statistical inference, 3970

O

ODS graph names
  LOESS procedure, 3975

ODS Graphics
  LOESS procedure, 3952, 3975

options summary
  MODEL statement (LOESS), 3957

S

statistical graphics
  LOESS procedure, 3975
Syntax Index

A

ALL option
  MODEL statement (LOESS), 3958
ALPHA= option
  MODEL statement (LOESS), 3958

B

BUCKET= option
  MODEL statement (LOESS), 3958
BY statement
  LOESS procedure, 3957

C

CLM option
  MODEL statement (LOESS), 3958
  SCORE statement (LOESS), 3963

D

DATA= option
  PROC LOESS statement, 3952
DEGREE= option
  MODEL statement (LOESS), 3959
DETAILS option
  MODEL statement (LOESS), 3959
DFMETHOD= option
  MODEL statement (LOESS), 3959
DFMETHOD=APPROX(Cutoff= ) option
  MODEL statement (LOESS), 3959
DFMETHOD=APPROX(Quantile= ) option
  MODEL statement (LOESS), 3959
DIRECT option
  MODEL statement (LOESS), 3959
DROPSQUARE= option
  MODEL statement (LOESS), 3960

I

ID statement
  LOESS procedure, 3957
INTERP= option
  MODEL statement (LOESS), 3960
ITERATIONS= option
  MODEL statement (LOESS), 3960

L

LOESS procedure, BY statement, 3957
LOESS procedure, ID statement, 3957
LOESS procedure, MODEL statement, 3957
  ALL option, 3958
  ALPHA= option, 3958
  BUCKET= option, 3958
  CLM= option, 3958
  DEGREE= option, 3959
  DETAILS option, 3959
  DFMETHOD= option, 3959
  DFMETHOD=APPROX(Cutoff= ) option, 3959
  DFMETHOD=APPROX(Quantile= ) option, 3959
  DIRECT option, 3959
  DROPSQUARE= option, 3960
  INTERP= option, 3960
  ITERATIONS= option, 3960
  RESIDUAL option, 3960
  SCALE= option, 3960
  SCALEDINDEP option, 3960
  SELECT= option, 3960
  SMOOTH= option, 3962
  STD option, 3963
  T option, 3963
  TRACEL option, 3963

LOESS procedure, PROC LOESS statement, 3952
  DATA= option, 3952
  PLOT option, 3952
  PLOTS option, 3952
LOESS procedure, SCORE statement, 3963
  CLM option, 3963
  PRINT option, 3963
  RESIDUAL option, 3964
  SCALEDINDEP option, 3964
  STEPS option, 3964
LOESS procedure, WEIGHT statement, 3964

M

MODEL statement
  LOESS procedure, 3957

P

PLOT option
  PROC LOESS statement, 3952
PLOTS option
  PROC LOESS statement, 3952
PRINT option
   SCORE statement (LOESS), 3963
PROC LOESS statement, see LOESS procedure

R

RESIDUAL option
   MODEL statement (LOESS), 3960
   SCORE statement (LOESS), 3964

S

SCALE= option
   MODEL statement (LOESS), 3960
SCALEDINDEP option
   MODEL statement (LOESS), 3960
   SCORE statement (LOESS), 3964
SCORE statement
   LOESS procedure, 3963
SELECT= option
   MODEL statement (LOESS), 3960
SMOOTH= option
   MODEL statement (LOESS), 3962
STD option
   MODEL statement (LOESS), 3963
STEPS option
   SCORE statement (LOESS), 3964

T

T option
   MODEL statement (LOESS), 3963
TRACEL option
   MODEL statement (LOESS), 3963

W

WEIGHT statement
   LOESS procedure, 3964
Your Turn

We welcome your feedback.

- If you have comments about this book, please send them to yourturn@sas.com. Include the full title and page numbers (if applicable).
- If you have comments about the software, please send them to suggest@sas.com.
Whether you are new to the work force or an experienced professional, you need to distinguish yourself in this rapidly changing and competitive job market. SAS® Publishing provides you with a wide range of resources to help you set yourself apart. Visit us online at support.sas.com/bookstore.

**SAS® Press**

Need to learn the basics? Struggling with a programming problem? You’ll find the expert answers that you need in example-rich books from SAS Press. Written by experienced SAS professionals from around the world, SAS Press books deliver real-world insights on a broad range of topics for all skill levels.

[www.support.sas.com/saspress](http://www.support.sas.com/saspress)

**SAS® Documentation**

To successfully implement applications using SAS software, companies in every industry and on every continent all turn to the one source for accurate, timely, and reliable information: SAS documentation. We currently produce the following types of reference documentation to improve your work experience:

- Online help that is built into the software.
- Tutorials that are integrated into the product.
- Reference documentation delivered in HTML and PDF – free on the Web.
- Hard-copy books.

[www.support.sas.com/publishing](http://www.support.sas.com/publishing)

**SAS® Publishing News**

Subscribe to SAS Publishing News to receive up-to-date information about all new SAS titles, author podcasts, and new Web site features via e-mail. Complete instructions on how to subscribe, as well as access to past issues, are available at our Web site.

[www.support.sas.com/spn](http://www.support.sas.com/spn)