

Recent Enhancements and New Directions in SAS/STAT[®] Software, Part I: Updates

Maura E. Stokes and Robert N. Rodriguez
SAS Institute Inc., Cary, NC

Abstract

Version 7 of the SAS[®] System brings major enhancements to the statistical software. All output is now handled by the Output Delivery System, which gives the user control over the printing of the results, allows all tables and statistics to be output to SAS data sets, and produces web-browsable HTML output. New procedures provide tools for partial least squares analysis and spatial prediction. The GENMOD procedure now provides LSMEANS and ESTIMATE statements, and its GEE facility provides the alternating logistic regression algorithm, produces Wald and score tests for model effects, and handles the ordinal response case. Additional exact tests have been added to several procedures, and even the TTEST procedure has been updated.

In addition to procedures for survey design and analysis, Version 7 also introduces experimental procedures for nonparametric density estimation and nonparametric regression, as discussed in Part II of this paper.

Introduction

Statistical developers have been busy at work on enhancements for Version 7 of the SAS System, targeted for availability during the fourth quarter, 1998. The Output Delivery System is now used by all procedures to handle their results. Instead of directly generating list files and output data sets, procedures generate an output object for every result that can be displayed. There are two components to the output object: the data component, which is the raw results, and the template component, which is a description of how the output should be arranged when formatted on a page.

The default output destinations continue to be the standard list file or SAS output window, and the established OUT= data sets and OUTPUT statements are still supported. However, this system enables you to send output to other destinations such as output directories, modify the output style with the new TEMPLATE procedure, merge pieces of the output into more comprehensive pages, and render the output in HTML, rich-text, Postscript, or PCL format. You can also replay the output after the procedure has already been executed and output to a data set any table or statistic that the procedure computes.

The incorporation of ODS in all of the statistical procedures gives the user long-needed flexibility in the management of analytical results and their inclusion into various document

forms; for more information, refer to Olinger and Tobias (1998) in these proceedings.

Another outcome of the ODS project is a more consistent appearance of the SAS/STAT output. A major effort was undertaken to make the output more readable and similar across the statistical procedures. Table formats, statistical terms, and abbreviations are now much more consistent.

While the ODS work has been a major undertaking, substantial progress has been made with statistical enhancements. These fall into the following categories: new production procedures, enhancements to existing production procedures, and new experimental procedures that take SAS/STAT software in new directions. This paper highlights some of these features.

Partial Least Squares

Partial least squares is a very popular technique in the field of chemometrics. The goal of regular least squares regression is to minimize sample response prediction error, finding linear functions of the predictors that explain as much variation as possible in the response. Predictive partial least squares has the additional goal of accounting for variation in the predictors, since directions in the predictor space that are well sampled should provide better prediction for new observations when the predictors are highly correlated. The PLS procedure, production in Version 7 of the SAS System, extracts successive linear combinations of the predictors, called factors or components, that optimally explain predictor and/or response variation.

Specifically, these techniques are

- principal components regression, which extracts factors to explain as much predictor sample variation as possible
- reduced rank regression, which extracts factors to explain as much response variation as possible
- partial least squares regression, which balances the two objectives of explaining response variation and explaining predictor variation

The data help to determine the number of factors that you extract. You can improve the model fit if you extract more factors, but if you extract too many factors you may overfit the data. With the PLS procedure, you can choose the

number of extracted factors via cross-validation, which is a protocol for fitting the model to part of the data and minimizing prediction error for the unfitted part. One-at-a-time validation, splitting the data into blocks, and test set validation methods are included.

Spectrometric Calibration

As an example, consider the following data reported by Umetrics (1995). Investigators studying pollution in the Baltic Sea wanted to use the spectra of samples of sea water to determine the amounts of three compounds present: lignin sulfonate (LS: pulp industry pollution), humic acids (HA: natural forest products), and optical whitener from detergent (DT). The predictors are the frequencies in sample spectrum, and the responses are the amounts of various chemicals in the sample.

For the purposes of calibrating the model, samples of known compositions are used. The calibration data consist of 16 samples of known concentrations of LS, HA, and DT, with spectra based on 27 frequencies (or, equivalently, wavelengths). The following statements create a SAS data set named SAMPLE.

```
data sample;
  input obsnam $ v1-v27 ls ha dt @@@@;
datalines;
EM1 2766 2610 3306 3630 3600 3438 3213 3051 2907 2844 2796
2787 2760 2754 2670 2520 2310 2100 1917 1755 1602 1467
1353 1260 1167 1101 1017 3.0110 0.0000 0.00
EM2 1492 1419 1369 1158 958 887 905 929 920 887 800
710 617 535 451 368 296 241 190 157 128 106
89 70 65 56 50 0.0000 0.4005 0.00
EM3 2450 2379 2400 2055 1689 1355 1109 908 750 673 644
640 630 618 571 512 440 368 305 247 196 156
120 98 80 61 50 0.0000 0.0000 90.63
EM4 2751 2883 3492 3570 3282 2937 2634 2370 2187 2070 2007
1974 1950 1890 1824 1680 1527 1350 1206 1080 984 888
810 732 669 630 582 1.4820 0.1580 40.00
(12 other samples)
```

To isolate a few underlying spectral factors that provide a good predictive model, you can fit a PLS model to the 16 samples with the following statements.

```
proc pls data=sample;
  model ls ha dt = v1-v27;
run;
```

Percent Variation Accounted for by Partial Least Squares Factors				
Number of Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	97.4607	97.4607	41.9155	41.9155
2	2.1830	99.6436	24.2435	66.1590
3	0.1781	99.8217	24.5339	90.6929
4	0.1197	99.9414	3.7898	94.4827
5	0.0415	99.9829	1.0045	95.4873
6	0.0106	99.9935	2.2808	97.7681
7	0.0017	99.9952	1.1693	98.9374
8	0.0010	99.9961	0.5041	99.4415
9	0.0014	99.9975	0.1229	99.5645
10	0.0010	99.9985	0.1103	99.6747
11	0.0003	99.9988	0.1523	99.8270
12	0.0003	99.9991	0.1291	99.9561
13	0.0002	99.9994	0.0312	99.9873
14	0.0004	99.9998	0.0065	99.9938
15	0.0002	100.0000	0.0062	100.0000

Figure 1. PLS Variation Summary

By default, the PLS procedure extracts as many as 15 factors. The procedure lists the amount of variation accounted for by each of these factors, both individual and cumulative. See Figure 1 for the listing. Almost all of the variation is explained by a relatively small number of factors—one or two for the predictors and three to eight for the responses.

To continue the PLS modeling process, you make a choice about the number of factors. You try to determine the number of factors that sufficiently explain the predictor and response variation without overfitting. One way to do this is with cross-validation, in which you divide the data set into two or more groups. You fit the model to all groups except one, then you check the capability of the model to predict responses for the group omitted. Repeating this for each group, you then can measure the overall capability of a given form of the model. The Predicted RESidual Sum of Squares (PRESS) statistic is based on the residuals generated by this process.

To select the number of extracted factors by cross-validation, you specify the CV= option with an argument that specifies which cross-validation method to use. For example, a common method is split-sample validation, in which the different groups are comprised of every seventh observation beginning with the first, every seventh observation beginning with the second, and so on. You can specify split-sample validation using the CV=SPLIT option, as illustrated in the following statements.

```
proc pls data=sample cv=split;
  model ls ha dt = v1-v27;
run;
```

The resulting output is shown in Figure 2 and Figure 3.

Split-sample Validation for the Number of Extracted Factors		
Number of Extracted Factors	Root Mean PRESS	
0	1.107747	
1	0.957983	
2	0.931314	
3	0.520222	
4	0.530501	
5	0.586786	
6	0.475047	
7	0.477595	
8	0.483138	
9	0.485739	
10	0.48946	
11	0.521445	
12	0.525653	
13	0.531049	
14	0.531049	
15	0.531049	
Minimum root mean PRESS		0.4750
Minimizing number of factors		6

Figure 2. Split-Sample Validated PRESS Statistics for Number of Factors

Percent Variation Accounted for by Partial Least Squares Factors				
Number of Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	97.4607	97.4607	41.9155	41.9155
2	2.1830	99.6436	24.2435	66.1590
3	0.1781	99.8217	24.5339	90.6929
4	0.1197	99.9414	3.7898	94.4827
5	0.0415	99.9829	1.0045	95.4873
6	0.0106	99.9935	2.2808	97.7681

Figure 3. PLS Variation Summary for Split-Sample Validated Model

The absolute minimum PRESS is achieved with six extracted factors. Notice, however, that this is not much smaller than the PRESS for three factors. By using the CVTEST option, you can perform a statistical model comparison suggested by van der Voet (1994) to test whether this difference is significant.

```
proc pls data=sample cv=split cvtest;
  model ls ha dt = v1-v27;
run;
```

The resulting output is shown in Figure 4 and Figure 5.

Split-sample Validation for the Number of Extracted Factors			
Number of Extracted Factors	Root Mean PRESS	T**2	Prob > T**2
0	1.107747	9.272858	0.0010
1	0.957983	10.62305	<.0001
2	0.931314	8.950878	0.0020
3	0.520222	5.133259	0.1340
4	0.530501	5.168427	0.1090
5	0.586786	6.437266	0.0120
6	0.475047	0	1.0000
7	0.477595	2.809763	0.4390
8	0.483138	7.189526	0.0130
9	0.485739	7.931726	0.0060
10	0.48946	6.612597	0.0220
11	0.521445	6.666235	0.0100
12	0.525653	7.092861	0.0060
13	0.531049	7.538298	0.0040
14	0.531049	7.538298	0.0040
15	0.531049	7.538298	0.0040

Minimum root mean PRESS 0.4750
 Minimizing number of factors 6
 Smallest number of factors with p > 0.1 3

Figure 4. Testing Split-Sample Validation for Number of Factors

Percent Variation Accounted for by Partial Least Squares Factors				
Number of Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	97.4607	97.4607	41.9155	41.9155
2	2.1830	99.6436	24.2435	66.1590
3	0.1781	99.8217	24.5339	90.6929

Figure 5. PLS Variation Summary for Tested Split-Sample Validated Model

The p -value of 0.1340 for comparing the cross-validated residuals from models with 6 and 3 factors indicates that the difference between the two models is insignificant; therefore, the model with fewer factors is preferred. You could continue the analysis by applying this model to new samples.

For more information, refer to Tobias (1995).

Tools for Spatial Prediction

Spatial prediction is an analytical technique that is useful in such areas as petroleum exploration, mining, and air and water pollution analysis. In these fields, data are often available at particular spatial locations, such as an experimental station positioned a certain distance in the air or under the ground, and the goal is to predict the quantities at unsampled locations. The unsampled locations are often mapped on a regular grid, and the predictions are used to produce surface plots or contour maps.

In general, spatial prediction is any prediction method that incorporates spatial dependence. A popular method of spatial prediction is ordinary kriging, which produces both predicted values and associated standard errors. Ordinary kriging requires the complete specification (the form and parameter values) of the spatial dependence of the spatial process in terms of a covariance or semivariogram model. Typically, the semivariogram model is not known in advance and must be estimated, either visually or by some estimation method. Performing spatial prediction requires two steps. First, the theoretical covariance or semivariogram of the spatial process must be determined. This involves choosing both a mathematical form and the values of the associated parameters. Second, the theoretical semivariogram is used in solving the kriging system at a specified set of spatial points, resulting in predicted values and associated standard errors.

Version 7 of SAS/STAT software includes production versions of two procedures that correspond to the two steps described for spatial prediction for two-dimensional data. Both of these procedures were available as experimental procedures in Release 6.12. The VARIOGRAM procedure computes the sample or empirical measures of spatial continuity (the semivariogram or covariance), which is then used in determining the theoretical semivariogram model by graphical or other means. The KRIGE2D procedure performs ordinary kriging at specified points using the theoretical model. Results are usually displayed with the GPLOT and G3D procedures or SAS/INSIGHT® software.

The VARIOGRAM procedure

- produces the sample regular semivariogram, a robust version of the semivariogram and the sample covariance
- saves continuity measures in an output data set, allowing plotting or parameter estimation for theoretical semivariogram or covariance models
- computes isotropic and anisotropic measures
- saves an additional OUTPAIR data set to contain an

observation for each pair of points

- saves an additional OUTDISTANCE data set that contains histogram information on the count of pairs within distance intervals

The KRIGE2D procedure

- handles anisotropic and nested semivariogram models
- supports Gaussian, exponential, spherical, and power models
- provides local kriging through specification of a radius around a grid point or specification of number of nearest neighbors to use
- writes kriging estimates and standard errors to an output data set z

The following is a surface plot of kriged data, obtained by applying the VARIOGRAM and KRIGE2D procedures. For more details, refer to *SAS/STAT® Software Technical Report: Spatial Prediction Using the SAS® System*.

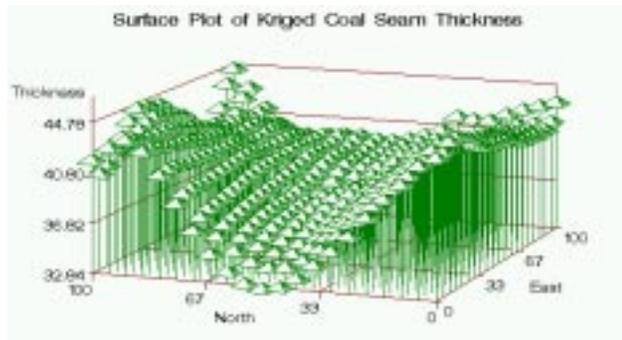


Figure 6. Fitted Surface Using Values in ESTIMATE

New in Release 7 is the SIM2D procedure, which produces a spatial simulation for a Gaussian random field with a specified mean and covariance structure.

Survey Data Analysis

Many researchers use sample surveys to collect their information, relying on probability-based complex sample designs such as stratified selection, clustering, and unequal weighting. This is done to select samples at lowest possible cost that can produce estimates that are precise enough for the purposes of the study. To make statistically valid inferences, the study design must be taken into account in the data analysis. Traditional SAS procedures such as the MEANS and GLM procedures compute statistics under the assumption that the sample is drawn from an infinite population with simple random sampling.

New SAS procedures for survey design and survey data analysis enable the SAS user to work with data based on

complex sampling design. The SURVEYSELECT procedure provides a variety of methods for selecting probability-based random samples as well as samples according to a complex multi-stage design that includes stratification, clustering, and unequal probabilities of selection. You input a SAS data set that includes the sampling frame, the list of units from which the sample is to be selected, and specify the selection method, the desired sample size or sampling rate, and other selection parameters. PROC SURVEYSELECT then selects the sample and produces an output data set that contains the selected units, their selection probabilities, and the sampling weights.

The SURVEYMEANS procedure computes estimates of survey population totals and means, estimates of their variances, confidence limits, and other descriptive statistics. The SURVEYREG procedure performs regression analysis for sample survey data, fitting linear models and computing regression coefficients and the covariance matrix. It provides significance tests for model effects and for specifiable estimable linear functions of the model parameters. Both of these procedures can handle sample designs such as stratification, clustering, and unequal weighting.

The following statements illustrate the SURVEYREG procedure syntax. Besides the familiar MODEL statement, the STRATA statement defines the strata and the WEIGHT statement specifies the variable containing the sampling weights.

```
proc surveyreg data=elder N=StrataTot;
  strata state region / list;
  model DentExpend = income age status;
  weight sweight;
run;
```

For more information, refer to An and Watts (1998) in these proceedings.

Generalized Linear Models

Several users have requested an LSMEANS statement for the GENMOD procedure, and Version 7 provides an extension of least squares means to the generalized linear model. In addition, an ESTIMATE statement has been added, and the negative binomial distribution is now supported through the DIST=NEGBIN option in the MODEL statement. The more recent GEE facilities have also been enhanced.

Least squares means are population marginal estimates, or the class or subclass estimates that you would expect for a balanced design involving the class variable with all covariates at their mean value. In the ANOVA setting, these estimates are means. In the generalized linear model setting, these quantities are the appropriate link function being modeled, such as the logit function in logistic regression. The basic facilities of the LSMEANS statement such as estimating marginal estimates and differences are provided, and the syntax is identical to that in PROC GLM.

The ESTIMATE statement is also now available with the GENMOD procedure and is used to estimate linear functions of parameters of the form Lb where b is the parameter

vector. The syntax is also the same as the ESTIMATE statement in PROC GLM.

As an example, consider the following data (Collett 1991, p. 142). Three different insecticide treatments were applied to flour beetles, in different dosages, and the proportion killed as a result of exposure were recorded. Variable Y is the number killed, N is the total number of beetles in that group, DEPOSIT is the dosage, and TRT is the type of insecticide treatment. Variable LDEP is log of DEPOSIT.

```
data beetle;
  input y n deposit trt$;
  ldep = log(deposit);
  sub=_n_;
  datalines;
3 50 2.00 ddt
5 49 2.64 ddt
19 47 3.48 ddt
19 50 4.59 ddt
24 49 6.06 ddt
35 50 8.00 ddt
2 50 2.00 gbhc
14 49 2.64 gbhc
20 50 3.48 gbhc
27 50 4.59 gbhc
41 50 6.06 gbhc
40 50 8.00 gbhc
28 50 2.00 mix
37 50 2.64 mix
46 50 3.48 mix
48 50 4.59 mix
48 50 6.06 mix
50 50 8.00 mix
;
```

The following statements fit a logistic regression model to these data with LDEP and TRT as the explanatory variables. The LSMEANS statement requests predicted logits for each treatment level as well as differences and the covariance matrix. The first three ESTIMATE statements reproduce the LSMEANS results. The fourth ESTIMATE statement requests the difference between levels *gbhc* and *mix*, as well as the 90% confidence limits. This request is repeated with the EXP option added, which specifies that the estimate is to be exponentiated. The final ESTIMATE statement requests the test of whether the average of the logits corresponding to the first two treatments is equal to the logit for their mixture.

```
proc genmod data=beetle;
  class trt;
  model y/n = ldep trt / dist = binomial;
  lsmeans trt /diff cov;
  estimate 'trt ddt' intercept 1 ldep 1.3859983 trt 1 0 0;
  estimate 'trt gbhc' intercept 1 ldep 1.3859983 trt 0 1 0;
  estimate 'trt mix' intercept 1 ldep 1.3859983 trt 0 0 1 /exp;
  estimate 'trt gbhc-mix' trt 0 1 -1/alpha=.10;
  estimate '1/3(trt1+trt2) - 2/3trt3'
    trt 1 1 -2 / divisor=3;
run;
```

Figure 7 contains the parameter estimates. The model is adequate with a Pearson chi-square goodness-of-fit value of 21.2819 and 14 DF (not shown).

Analysis Of Parameter Estimates					
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	
				Lower	Upper
Intercept	1	-1.4248	0.2851	-1.9835	-0.8661
ldep	1	2.6958	0.2157	2.2730	3.1185
trt	ddt	-3.1305	0.2522	-3.6248	-2.6362
trt	gbhc	-2.4177	0.2381	-2.8844	-1.9510
trt	mix	0.0000	0.0000	0.0000	0.0000
Scale	0	1.0000	0.0000	1.0000	1.0000

Analysis Of Parameter Estimates			
Parameter	Chi-Square	DF	Pr > ChiSq
Intercept	24.98	1	<.0001
ldep	156.21	1	<.0001
trt	154.09	1	<.0001
trt	103.10	1	<.0001
trt	.	.	.
Scale	.	.	.

Figure 7. Parameter Estimates

Figure 8 contains the results, which include the estimates, their differences, and tests of significance.

Least Squares Means							
Effect	trt	Estimate	Standard Error	DF	Chi-Square	Pr > ChiSq	Cov1
trt	ddt	-0.8189	0.1450	1	31.89	<.0001	0.0210
trt	gbhc	-0.1061	0.1361	1	0.61	0.4355	0.0002
trt	mix	2.3116	0.1936	1	142.63	<.0001	-0.0026

Least Squares Means				
Effect	trt	Cov2	Cov3	
trt	ddt	0.0002	-0.0026	
trt	gbhc	0.0185	-0.0004	
trt	mix	-0.0004	0.0375	

Differences of Least Squares Means							
Effect	trt	_trt	Estimate	Standard Error	DF	Chi-Square	Pr > ChiSq
trt	ddt	gbhc	-0.7128	0.1981	1	12.95	0.0003
trt	ddt	mix	-3.1305	0.2522	1	154.09	<.0001
trt	gbhc	mix	-2.4177	0.2381	1	103.10	<.0001

Figure 8. Least Square Mean Results

Figure 9 contains the ESTIMATE statement results. Note that the first three estimates reproduce the LSMEANS results and these estimates are accompanied by a 90% confidence interval. Note that the EXP option produces the exponentiated estimate for the mixture treatment level.

ESTIMATE Statement Results				
Label	Estimate	Standard Error	Alpha	Lower
trt ddt	-0.8189	0.1450	0.05	-1.1031
trt gbhc	-0.1061	0.1361	0.05	-0.3729
trt mix	2.3116	0.1936	0.05	1.9322
Exp(trt mix)	10.0903	1.9530	0.05	6.9048
trt gbhc-mix	-2.4177	0.2381	0.1	-2.8094
1/3(trt1+trt2) - 2/3trt3	-1.8494	0.1496	0.05	-2.1426

ESTIMATE Statement Results				
Label	Upper	Chi-Square	Pr >	ChiSq
trt ddt	-0.5347	31.89	<.0001	
trt gbhc	0.1606	0.61	0.4355	
trt mix	2.6909	142.63	<.0001	
Exp(trt mix)	14.7454			
trt gbhc-mix	-2.0261	103.10	<.0001	
1/3(trt1+trt2) - 2/3trt3	-1.5563	152.89	<.0001	

Figure 9. Estimate Results

For Release 6.12 of the SAS System, the GENMOD procedure was enhanced to support Generalized Estimating Equations (GEE), introduced by Liang and Zeger (1986) as a method of dealing with correlated data when, except for the correlation among responses, the data can be modeled with a generalized linear model.

Correlated data can arise from situations such as

- longitudinal studies, in which multiple measurements are taken on the same subject at different points in time
- clustering, where measurements are taken on subjects that share a common category or characteristic that leads to correlation

The correlation must be accounted for by analysis methods appropriate to the data. You model the correlated data by using the same link function and linear predictor as in a generalized linear model for the independent case; you describe the random component by the same variance function. However, in the GEE approach, you also model the covariance structure of the correlated measures.

The GEE facilities have also been extended in Version 7. Type 3 tests are now available for model effects, and the CONTRAST statement now applies to the GEE model estimates. In addition, the LSMEANS and ESTIMATE statements can be used for the GEE parameter estimates.

The method of alternating logistic regression estimation (Carey, Zeger, and Diggle 1993) is now available as well as models for ordinal response data. The proportional odds model is perhaps the most popular of these models for GEE analysis (Lipsitz, Kim, and Zhao 1994) and depends on modeling cumulative logit functions. The GENMOD procedure also models cumulative probits and cumulative complementary log-log functions.

Consider the following SAS data set from Koch et al (1990). A clinical study conducted at several medical centers investigates whether active treatment has an effect on respiratory symptoms, captured as a five point scale from 0 for poor to 4 for excellent. Other variables include base score, age, and gender.

```
data resp;
  input age base gender $ treat $
  center id visit score dichot;
  trt = (treat = 'a');
  gen = (gender = 'female');
datalines;
39 1 female p 1 101 1 2 0
39 1 female p 1 101 2 1 0
39 1 female p 1 101 3 1 0
39 1 female p 1 101 4 2 0
25 2 male a 1 102 1 2 0
25 2 male a 1 102 2 4 1
25 2 male a 1 102 3 4 1
25 2 male a 1 102 4 4 1
58 4 male a 1 103 1 4 1
58 4 male a 1 103 2 4 1
58 4 male a 1 103 3 4 1
58 4 male a 1 103 4 4 1
51 3 female p 1 104 1 4 1
51 3 female p 1 104 2 2 0
51 3 female p 1 104 3 4 1
51 3 female p 1 104 4 4 1
32 3 female p 1 105 1 2 0
32 3 female p 1 105 2 2 0
32 3 female p 1 105 3 3 1
32 3 female p 1 105 4 4 1
45 3 male p 1 106 1 4 1
....
;
```

The REPEATED statement is where the cluster id is specified, as well as the working correlation structure. The LINK=CLOGIT option in the MODEL statement requests cumulative logits, which, with the DIST=MULT specification of the multinomial distribution, specifies the proportional odds model.

```
proc genmod data=resp;
  class id treat gender ;
  model score = visit trt gen center base /
  dist=mult link=clogit itprint;
  repeated subject=id / type=unstr corrw;
run;
```

Analysis Of GEE Parameter Estimates Empirical Standard Error Estimates						
Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > Z
			Lower	Upper		
Intercept1	-2.6672	1.6827	-5.9653	0.6308	-1.59	0.1129
Intercept2	-1.5543	1.6820	-4.8510	1.7425	-0.92	0.3555
Intercept3	0.2224	1.6998	-3.1091	3.5538	0.13	0.8959
Intercept4	1.4065	1.7190	-1.9627	4.7756	0.82	0.4132
visit	-0.0421	0.0554	-0.1506	0.0664	-0.76	0.4469
trt	-1.7737	0.5503	-2.8524	-0.6950	-3.22	0.0013
gen	-0.3600	0.6850	-1.7026	0.9826	-0.53	0.5992
center	1.1326	0.5592	0.0366	2.2286	2.03	0.0428
base	-0.7664	0.1201	-1.0019	-0.5310	-6.38	<.0001

Figure 10. Parameter Estimates

Since there are five outcomes, four cumulative logits are being modeled. The model includes an intercept term corresponding to each cumulative logit and slope terms that apply to all cumulative logits. This analysis indicates that treatment and center are influential effects and that baseline

must be included as a covariate.

Analyzing III-Conditioned Data

Occasionally, you may be faced with badly-scaled or ill-conditioned data. You may use the GLM or REG procedures only to find that you get messages stating that the estimation process can't find solutions. The ORTHOREG procedure was designed to handle these situations for the regression setting, and it uses the QR method to produce numerically precise estimates. This procedure has now been enhanced to accept a CLASS statement and GLM-like model specification so that it can handle a broader range of statistical models. In addition, the results have been upgraded to include additional statistics.

The following example illustrates the use of the ORTHOREG procedure with atomic data. In order to calibrate an instrument for measuring atomic weight, 24 replicate measurements of the atomic weight of silver (chemical symbol Ag) are made with the new instrument and with a reference instrument (Powell, Murphy, and Gramlich 1982).

```
data AgWeight;
  input Instrument AgWeight @@;
  datalines;
  1 107.8681568 1 107.8681465 1 107.8681572 1 107.8681785
  1 107.8681446 1 107.8681903 1 107.8681526 1 107.8681494
  1 107.8681616 1 107.8681587 1 107.8681519 1 107.8681486
  1 107.8681419 1 107.8681569 1 107.8681508 1 107.8681672
  1 107.8681385 1 107.8681518 1 107.8681662 1 107.8681424
  1 107.8681360 1 107.8681333 1 107.8681610 1 107.8681477
  2 107.8681079 2 107.8681344 2 107.8681513 2 107.8681197
  2 107.8681604 2 107.8681385 2 107.8681642 2 107.8681365
  2 107.8681151 2 107.8681082 2 107.8681517 2 107.8681448
  2 107.8681198 2 107.8681482 2 107.8681334 2 107.8681609
  2 107.8681101 2 107.8681512 2 107.8681469 2 107.8681360
  2 107.8681254 2 107.8681261 2 107.8681450 2 107.8681368
  ;
```

Notice that the variation in the atomic weight measurements is several orders of magnitude less than their mean. This is a situation that causes difficulty for standard least squares computations. The following statements invoke the ORTHOREG procedure to perform a simple one-way analysis of variance, testing for differences between the two instruments:

```
proc orthoreg data=AgWeight;
  class Instrument;
  model AgWeight = Instrument;
run;
```

ORTHOREG Regression Procedure					
Dependent Variable: AgWeight					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3.6383419E-9	3.6383419E-9	15.95	0.0002
Error	46	1.0495173E-8	2.281559E-10		
Corrected Total	47	1.4133515E-8			
		Root MSE	0.0000151048		
		R-Square	0.2574265445		

Figure 11. Results for Atomic Weight Example

Parameter	DF	Parameter Estimate	Standard Error	t Value
Intercept	1	107.868136354166	3.0832608E-6	3.499E7
(Instrument='1')	1	0.00001741249999	4.3603893E-6	3.99
(Instrument='2')	0	0	.	.

Parameter	Pr > t
Intercept	<.0001
(Instrument='1')	0.0002
(Instrument='2')	.

Figure 12. Results for Atomic Weight Example

The mean difference between instruments is about 1.74×10^{-5} (the value of the (Instrument='1') parameter in the parameter estimates table), whereas the level of background variation in the measurements is about 1.51×10^{-5} (the value of the root mean squared error). The difference is significant, with a *p*-value of 0.0002.

The following table displays the ANOVA values certified by the National Institute of Standards and Technology (1997) and those produced by the ORTHOREG and GLM procedures.

	Model SS	Error SS
cert	3.6383418750000E-09	1.0495172916667E-08
O	3.6383418747907E-09	1.0495172916797E-08
G	0	1.0331496763990E-08
	Root MSE	R-Square
cert	1.5104831444641E-05	0.25742654453832
O	1.5104831444735E-05	0.25742654452494
G	1.4986585859992E-05	0

The ORTHOREG values are quite close to the certified ones, but the GLM values are not. In fact, since the model sum of squares is so small, the GLM procedure sets it (and consequently R^2) to zero. While the GLM and REG procedures adequately handle most data sets that arise in practice, the ORTHOREG procedure is a useful tool for the exceptional occasion where they do not.

Revisiting T-Tests

While the TTEST procedure has been around since, well, the Statistical Analysis System, that doesn't mean that it couldn't be improved. The TTEST procedure can now perform a *t*-test for one sample, two samples, or paired observations. The one sample *t*-test compares the mean of the sample to a given number. The two sample *t*-test compares the mean of the first sample minus the mean of the second sample to a given number. The paired observations *t*-test compares the mean of the differences in the observations to a given number. FREQ and WEIGHT statements have been added, and confidence intervals for the means, differences of means, and a pooled-variance are available through the OUTPUT data set.

The new PAIRED statement enables you to test the differences of pairs of observations, instead of the difference of means of two groups. The following statements illustrate the specification:

```
paired a*b;
/* Performs t- test on difference A-B. */
paired a*b c*d;
/* Tests differences A-B and C-D. */
```

The CLASS and VAR statement cannot be used with the PAIRED statement.

For an example, consider the following systolic blood pressure data. Researchers recorded blood pressure before and after a stimulus was applied.

```
data pressure;
  input SBPbefore SBPafter @@;
datalines;
121 130 124 131 130 131 118 127
143 134 121 129 144 147 139 140
128 116 127 136 126 130 127 137
;
run;
```

The following statements request a paired *t*-test analysis.

```
proc ttest;
  paired SBPbefore*SBPafter;
run;
```

The PAIRED statement is used to test whether the mean change in systolic blood pressure is significantly different from zero.

Figure 13 contains statistics for the mean difference.

Statistics					
Difference	N	Lower CL		Upper CL	Lower CL
		Mean	Mean	Mean	Std Dev
SBPbefore - SBPafter	12	-7.926	-3.333	1.2591	5.1202

Statistics					
Difference	Upper CL		Std Err	Minimum	Maximum
	Std Dev	Std Dev			
SBPbefore - SBPafter	7.2279	12.272	2.0865	-10	12

Figure 13. Statistics

Figure 14 contains the value and *p*-value for the paired *t*-test. The difference is not significantly different from zero.

T-Tests			
Difference	DF	t Value	Pr > t
SBPbefore - SBPafter	11	-1.60	0.1384

Figure 14. Test Statistic

More Exact *p*-Values

In recent years, exact *p*-values have been added for many statistics in the FREQ and NPAR1WAY procedures. Exact *p*-values provide an alternative strategy when data are

sparse, skewed, or unbalanced so that the assumptions required for standard asymptotic tests are violated. Advances in computer performance and developments in network algorithms over the last decade have made exact *p*-values accessible for a number of statistical tests.

This work continues with Version 7.

Monte Carlo simulation is now available for computing exact *p*-values in both procedures; it is useful in some situations where the default exact algorithms are not feasible. This is requested with the MC option in the EXACT statement, and related options include the SEED=, ALPHA=, and N= options. The MAXTIME option in the EXACT statement specifies a time at which to quit if exact computations are not finished.

In addition, PROC NPAR1WAY has been updated to include:

- nonparametric tests for scale differences Siegel-Tukey, Ansari-Bradley, Klotz, Mood
- exact *p*-values for the above
- exact *p*-values for the Kolmogorov-Smirnov test
- FREQ statement

Also, a new SCORES=DATA option enables the user to input raw data as scores, giving the user a lot of flexibility. This option applies to both asymptotic and exact tests.

The FREQ procedure has been enhanced with a TEST statement for tests of the MEASURES and AGREE statistics. In addition, the new BINOMIAL option requests the binomial proportion, standard errors and confidence intervals, and a test of whether it is equal to 0.5 (or another specified value). Both asymptotic and exact tests are available.

SAS users may also be interested to know that SAS/IML® software now includes three new routines for robust regression and outlier detection. The LMS (Least Median of Squares) and LTS (Least Trimmed Squares) routines perform robust regression. They detect outlier and perform least squares regression on the remaining observations. The MVS (Minimum Volume Ellipsoid Estimation) can be used to find the minimum volume ellipsoid estimator, the location and robust covariance matrix that can be used to construct confidence regions and to detect multivariate outliers and leverage points. Refer to Rousseeuw (1984) and Rousseeuw and Leroy (1987) for details on robust estimation theory and methods.

Growing Confidence Intervals

In response to many requests from users, a number of procedures now provide additional support for confidence limits. For example, in the GLM procedure you can specify the CLPARM option in the MODEL statement to request confidence limits for the parameter estimates (if the SOLUTION option is also specified) and for the results of all ESTIMATE statements. Likewise, in the REG procedure

you can specify the CLB option in the MODEL statement to request confidence limits for the parameter estimates.

The UNIVARIATE procedure now computes confidence limits for a variety of distributional parameters. You can request a table of confidence intervals for the mean, variance, and standard deviation by specifying the CIBASIC option in the PROC statement. You can request confidence intervals for percentiles assuming normality with the CIPCTLNORMAL option, and you can request distribution-free confidence intervals for percentiles with the CIPCTLDF option. Refer to Hahn and Meeker (1991) for details of these methods.

The following statements illustrate these options using the batch data of Hahn and Meeker (1991). Note the use of the ODS SELECT statement to display selected tables.

```
data batch;
  input Amount;
datalines;
1.49
1.66
2.05
...
58.11
run;

ods select BasicIntervals Quantiles;
proc univariate data=batch
  cibasic cipctlnormal cipctldf;
  var Amount;
run;
```

The UNIVARIATE Procedure			
Variable: Amount			
Basic Confidence Limits Assuming Normality			
Parameter	Estimate	95% Confidence Limits	
Mean	12.9745	10.87392	15.07508
Std Deviation	10.58646	9.294988	12.29803
Variance	112.0732	86.39681	151.2416

Figure 15. Confidence Intervals

Figure 15 displays the basic confidence intervals. The confidence intervals for quantiles assuming normality are added to the default Quantiles table, as displayed in Figure 16.

Quantiles (Definition 5)			
Quantile	Estimate	95% Confidence Limits Assuming Normality	
100% Max	58.11		
99%	55.77	34.03228972	42.181539031
95%	31.48	27.51018657	33.991855510
90%	27.46	23.98135486	29.676202639
75% Q3	17.60	17.93944898	22.602019948
50% Median	9.23	10.87391576	15.075084236
25% Q1	5.18	3.34698005	8.009551024
10%	3.25	-3.72720264	1.967645138
5%	2.49	-8.04285551	-1.561186566
1%	1.57	-16.23253903	-8.083289719
0% Min	1.49		

Figure 16. Confidence Intervals Assuming Normality

Likewise, the distribution-free confidence intervals for quantiles, together with their corresponding ranks and coverage probabilities, are added to the Quantiles table.

Quantiles (Definition 5)				
Quantile	95% Confidence Limits		LCL Rank	UCL Rank
	Distribution Free			
100% Max				
99%	37.32	58.11	98	100
95%	28.28	58.11	91	100
90%	24.33	33.24	85	97
75% Q3	14.17	23.66	67	84
50% Median	7.81	12.93	41	61
25% Q1	4.09	6.55	17	34
10%	2.24	4.04	4	16
5%	1.49	3.23	1	10
1%	1.49	2.05	1	3
0% Min				

Quantiles (Definition 5)		
Quantile	Coverage	
100% Max		
99%		55.46
95%		96.59
90%		95.23
75% Q3		95.13
50% Median		95.40
25% Q1		95.13
10%		95.23
5%		96.59
1%		55.46
0% Min		

Figure 17. Distribution-Free Confidence Intervals

Robust Methods

The new STDIZE procedure standardizes one or more numeric variables in a SAS data set by subtracting a location measure and dividing by a scale measure. A variety of location and scale measures are provided, including estimates that are resistant to outliers and clustering. In addition, the standardized values can be multiplied by a constant or have a constant added to them, or both. Missing values can be replaced by the location measure or by any specified constant.

SAS users may also be interested to know that SAS/IML® software now includes three new routines for robust regression and outlier detection. The LMS (Least Median of Squares) and LTS (Least Trimmed Squares) routines perform robust regression. They detect outliers and perform least squares regression on the remaining observations. The MVS (Minimum Volume Ellipsoid Estimation) can be used to find the minimum volume ellipsoid estimator, the location and robust covariance matrix that can be used to construct confidence regions and to detect multivariate outliers and leverage points. Refer to Rousseeuw (1984) and Rousseeuw and Leroy (1987) for details on robust estimation theory and methods.

Other Work

Many other statistical procedures have been enhanced in one way or another. The PLAN procedure now generates lists of permutations and combinations. Smoothing splines

have been added to the TRANSREG procedure. And, various minor options have been added to several procedures.

Conclusion

The statistical capabilities of the SAS System continue to grow with Version 7 of the SAS System. The integration of ODS into all procedures makes results management facilities quite powerful. The addition of confidence intervals in several procedures gives users the type of information they have been requesting recently. New production and experimental procedures provide users with new tools for data analysis. The documentation for SAS/STAT software is also undergoing changes and a revised set of manuals will be released with Version 7. Documentation for the experimental procedures will be available in a separate technical report. In addition, documentation will also be available online for ready reference. The URL for the R and D web pages is <http://www.sas.com/rnd/> and these pages contain up-to-date information about the statistical products.

Acknowledgements

We are grateful to Tony An, Bruce Elsheimer, Gordon Johnston, Ann Kuo, Chris Olinger, Donna Sawyer, Randy Tobias, and Donna Watts for their contributions.

References

An, T. and Watts, D. L. (1998), "New SAS Procedures for Analysis of Sample Survey Data," in *Proceedings of the Twenty-Third Annual SAS Users Group International Conference*, Cary, NC: SAS Institute. Inc.

Carey, V., Zeger, S.L., and Diggle, P. (1993), "Modelling Multivariate Binary Data with Alternating Logistic Regressions," *Biometrika*, 517–526.

Collett, D. (1991). *Modelling Binary Data*, London: Chapman and Hall.

Diggle, P.J., Liang, K.-Y. and Zeger, S.L. (1994), *Analysis of Longitudinal Data*, Oxford: Oxford Science.

Hahn, G. J. and Meeker, W. Q. (1991), *Statistical Intervals*, New York: John Wiley & Sons, Inc.

Koch, G. G., Carr, G. J., Amara, I. A., Stokes, M. E., and Uryniak, T. J. (1990), "Categorical data analysis", in *Statistical Methodology in the Pharmaceutical Sciences*, ed. D.A. Berry, New York: Marcel Dekker Inc., 391–475.

Liang, K.-Y. and Zeger, S.L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 13–22

Lipsitz, S. R., Kim, K., and Zhao, L. (1994), "Analysis of Repeated Categorical Data Using Generalized Estimating Equations," *Statistics in Medicine*, 13, 1149–1163.

Miller, M. E., Davis, C. E., and Landis, J. R. (1993), "The Analysis of Longitudinal Polytomous Data: Generalized Estimating Equations and Connections with Weighted Least

Squares," *Biometrics*, 49, 1033–1044.

National Institute of Standards and Technology (1997), "Statistical Reference Datasets," [<http://www.nist.gov/>].

Olinger, C. and Tobias, R. (1997), "It Chops, It Dices, It Makes Julianne Slices! ODS for Data Analysis: Output As-You-Like-It in Version 7," in *Proceedings of the Twenty-Third Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc.

Powell, L.J., Murphy, T.J., and Gramlich, J.W. (1982), "The Absolute Isotopic Abundance and Atomic Weight of a Reference Sample of Silver," *NBS Journal of Research*, 87, 9–19.

Rousseeuw, P. J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, 79, 871–880.

Rousseeuw, P. J. and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, New York: John Wiley & Sons.

SAS Institute, Inc. (1996), *SAS STAT Technical Report: Spatial Prediction Using the SAS System*, Cary, NC: SAS Institute Inc.

Tobias, R. (1995), "An Introduction to Partial Least Squares Regression," in *Proceedings of the Twentieth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc., 1250–1257.

Umetrics, Inc. (1995), *Multivariate Analysis (3-day course)*, Winchester, MA.

van der Voet, H. (1994), "Comparing the Predictive Accuracy of Models Using a Simple Randomization Test," *Chemometrics and Intelligent Laboratory Systems*, 25, 313–323.

Zeger, S.L. and Liang, K.-Y. (1986), "Longitudinal Data Analysis for Discrete and Continuous Outcomes," *Biometrics*, 121–130.

Authors

Maura E. Stokes, SAS Institute Inc., SAS Campus Drive, Cary, NC 27513. Phone (919)-677-8000 ext 7172. FAX (919)-677-4444. Email sasmzs@wnt.sas.com

Robert N. Rodriguez, SAS Institute Inc., SAS Campus Drive, Cary, NC 27513. Phone (919)-677-8000 ext 7650. FAX (919)-677-4444. Email sasnr@unx.sas.com

SAS, SAS/STAT, SAS/IML, and SAS/INSIGHT are registered trademarks of SAS Institute Inc. in the USA and in other countries. ® indicates USA registration.

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

Recent Enhancements and New Directions in SAS/STAT® Software, Part II: Nonparametric Modeling Procedures

Robert N. Rodriguez and Maura E. Stokes
SAS Institute Inc.
Cary, North Carolina, USA

Introduction

Part II of this paper describes new SAS® procedures for nonparametric density estimation and nonparametric regression, two of the new directions in which statistical software is being developed for Version 7. These procedures are preliminary steps toward comprehensive support for modern nonparametric data analysis methods within the SAS System. It is anticipated that the coverage described here will expand to include a variety of other important methods. Some of the techniques provided by the new procedures are also being implemented as functions in SAS/IML® software and with interactive graphics in SAS/INSIGHT® software. The SUGI23 Proceedings paper by Cohen *et al.* (1998) describes parallel development in SAS/INSIGHT software.

The sections that follow discuss the scope of the new procedures and illustrate their use with basic examples. The procedures will be available as experimental software with the initial release of Version 7, and updated information will be provided on the Institute's Research and Development Web site at <http://www.sas.com/rnd/>. Complete documentation of syntax and computational details will be provided in a technical report.

Nonparametric Density Estimation: The KDE Procedure

The KDE procedure computes nonparametric estimates of univariate and bivariate probability density functions using the method of kernel density estimation. The procedure saves the density estimate in a SAS data set for subsequent plotting or analysis. In the bivariate case, the procedure also computes contours of the estimated density function.

For a univariate sample, $X_i, i = 1, 2, \dots, n$ with probability density function $f(x)$, the general form of the kernel density estimate of $f(x)$ is

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)$$

where h is the so-called bandwidth, and $K(x)$ is referred to as the kernel function. The kernel function continuously "smears out" the mass $1/n$ at each of the observations, and the estimate is formed by summing these masses. The

default kernel function used by the KDE procedure is

$$K(x) = \varphi(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$$

which is the standard normal density function. For an introduction to kernel density estimation, refer to Silverman (1986).

For a bivariate sample $(X_i, Y_i), i = 1, 2, \dots, n$ with joint probability density function $f(x, y)$, the kernel density estimate used for $f(x, y)$ is

$$\hat{f}(x, y) = \frac{1}{nh_x h_y} \sum_{i=1}^n \varphi\left(\frac{x - X_i}{h_x}, \frac{y - Y_i}{h_y}\right)$$

where $h_x > 0$ and $h_y > 0$ are a pair of bandwidths, and where $\varphi(x, y)$ is the bivariate normal density function

$$\varphi(x, y) = \frac{1}{2\pi} \exp\left(-\frac{x^2 + y^2}{2}\right)$$

The approach used by the KDE procedure in the bivariate case follows the development of Wand and Jones (1993) but is expected to evolve with ongoing research.

The following example illustrates the basic features of PROC KDE. An automotive industry study was carried out to assess the octane requirements of a group of customer-owned cars as determined by trained raters and the customers themselves; refer to Rodriguez and Taniguchi (1980). Based on previous studies, it was surmised that a significant fraction of customers should be experiencing knock on gasoline with an average octane number of 92.6. However, the low level of customer complaints about knock suggested that this level satisfied most customers. Consequently, a preliminary stage of the analysis was to explore the joint distribution of customer and rater octane requirements.

The following SAS statements create a data set named OCTANE which contains the requirements.

```
data octane;
  input Rater Customer;
datalines;
94.5 92.0
94.0 88.0
94.0 90.0
... ..
run;
```

The following statements compute a bivariate kernel density estimate from these data.

```
proc kde data=octane out=octden;
  var Customer Rater;
run;
```

The output from this analysis is as follows. The Inputs table lists basic information concerning the fit.

The KDE Procedure	
Inputs	
Data Set	WORK.OCTANE
Number of Observations Used	229
Variable 1	Customer
Variable 2	Rater
Estimation Method	Bivariate Kernel

The Controls table lists the parameters controlling the fit, which is computed for a 60×60 grid over the entire range of the data with a default bandwidth. You can use the `NGRID = numlist` option to specify the number of grid points associated with the variable(s) in the VAR statement. The default values are 401 when there is a single variable and 60 when there are two variables. You can use the `GRIDL = numlist` option to specify the lower bound for the grid, expressed as a percentage of the range of the corresponding VAR variable (the default is 0, indicating the minimum value of the variable). Likewise, you can use the `GRIDU = numlist` option to specify the upper bound for the grid, expressed as a percentage of the range of the corresponding VAR variable (the default is 100, indicating the maximum value of the variable).

Controls		
	Customer	Rater
Grid Points	60	60
Lower Grid Percentage	0	0
Upper Grid Percentage	100	100
Bandwidth Multiplier	1	1

The Statistics table provides standard univariate statistics for each variable.

Statistics		
	Customer	Rater
Mean	86.35	92.20
Variance	15.29	11.16
Standard Deviation	3.91	3.34
Range	21.60	17.50
Interquartile Range	5.00	5.00
Lower Grid Value	76.60	82.00
Upper Grid Value	98.20	99.50
Bandwidth	1.58	1.35

The Bivariate Statistics table provides the covariance and correlation. Note that the correlation is mild (0.56).

Bivariate Statistics	
Covariance	7.29
Correlation	0.56

The Percentiles table lists percentiles for each variable. You can specify the percentiles with the `PERCENTILES=` option in the PROC statement.

Percentiles		
	Customer	Rater
0.5	76.60	83.00
1.0	77.00	84.00
2.5	77.00	85.00
...
99.5	95.00	99.00

The Levels table lists density values corresponding to contours that enclose given percents of the data. For example, 90 percent of the observations have a density value less than 0.01091. Note that the contours need to be interpreted with caution because quantiles based on smoothed density estimates are biased estimates of population quantiles. You can specify the percents for the table with the `LEVELS=` option in the PROC statement.

Levels					
Percent	Density	Lower1	Lower2	Upper1	Upper2
1	0.000703	76.23	82.00	96.74	99.80
5	0.001315	76.23	84.67	95.64	99.50
10	0.001756	77.33	84.97	95.27	99.20
50	0.007350	82.82	87.93	91.98	96.83
90	0.01091	84.65	90.01	89.78	95.05
95	0.01111	85.02	90.31	87.22	92.97
99	0.01157	85.39	90.90	86.48	92.08
100	0.01166	86.12	91.79	86.12	91.79

The output data set OCTDEN contains the 3600 points at which the kernel density estimate was evaluated. You can display surface and contour plots of the estimate as follows:

```
title 'Distribution of Octane Requirements';
proc g3d data=octden;
  plot Rater*Customer=density;

proc gcontour data=octden;
  plot y*x=density;
run;
```

These plots are displayed in Figure 1 and Figure 2. Figure 1 reveals that the data were slightly censored for low octane requirements; in fact, there were 17 cars for which the customer requirement was less than 76.6 RON (the lowest octane gasoline used in the study). Both plots suggest that the density is slightly bimodal. They also reveal that the conditional distributions of customer requirements given rater requirements are heteroscedastic.

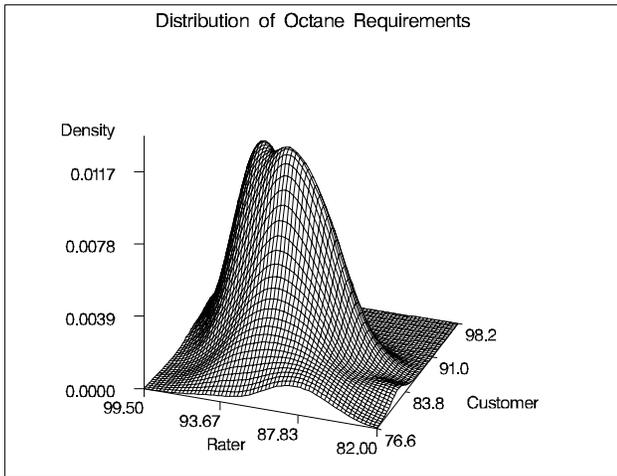


Figure 1. Surface of Density Estimate

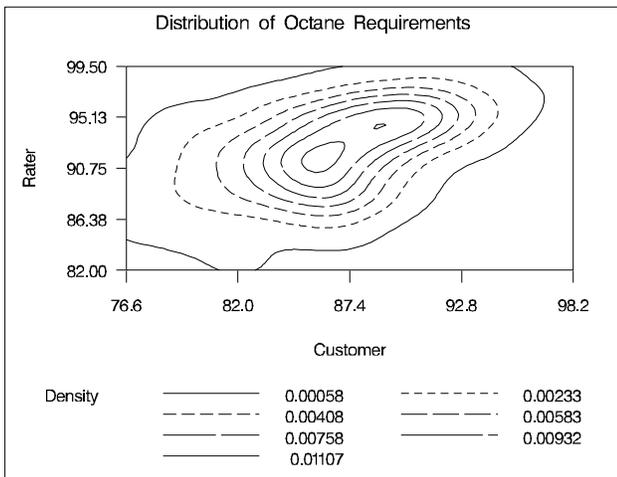


Figure 2. Contours of Density Estimate

An important issue in the application of kernel density estimates is the choice of the bandwidth. In the univariate case, this has been the topic of considerable research; refer to Marron (1989) for a survey. For this case the KDE procedure provides several methods for automatic bandwidth selection, including the method provided by Silverman (1986) and the more recent SJPI method recommended by Jones *et al.* (1996). You can use the BWM= option in the PROC statement to specify a multiplier for the default bandwidth.

Wand and Jones (1993) note that automatic bandwidth selection in the bivariate case is both difficult and computationally expensive. However, their study also shows that using two bandwidths, one in each coordinate direction, is often adequate. The KDE procedure allows you to adjust the two bandwidths by using the BWM= option to specify multipliers for the default bandwidths recommended by Bowman and Foster (1992):

$$h_X = \sigma_X n^{-1/6}$$

$$h_Y = \sigma_Y n^{-1/6}$$

Here, $\hat{\sigma}_X$ and $\hat{\sigma}_Y$ are the sample standard deviations, respectively. These are the optimal bandwidths for two independent normal variables that have the same variances as X and Y , respectively. They are conservative in the sense that they tend to over-smooth the surface. It is good practice to work with a range of bandwidths since, as recommended by Marron (1998), important information is available at a number of different smoothing levels.

Suppose after viewing the preceding figures, you decide that you would like a slightly smoother estimate. You could rerun the analysis with a larger bandwidth pair:

```
ods output Levels=OutLevels;
proc kde data=octane out=octden2
  bwm=2,2
  levels=25 50 75 95;
  var Customer Rater;
run;
```

The BWM=2,2 option requests bandwidth multipliers of 2 for both Customer and Rater. The results of this fit are displayed in Figure 3. This estimate is unimodal, although heteroscedasticity is still evident.

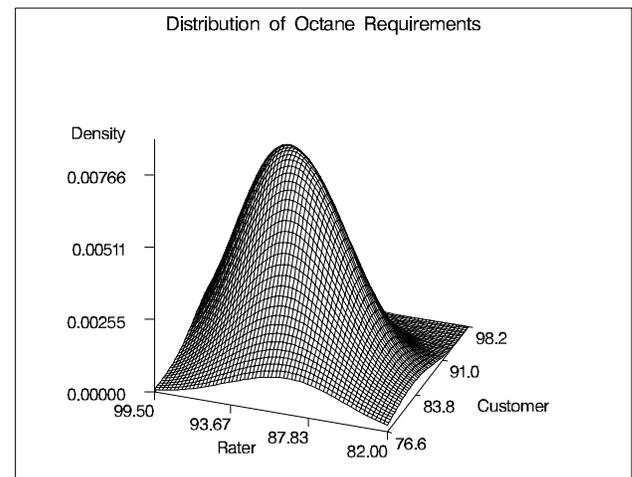


Figure 3. Surface of Density Estimate for BWM=2,2

You can also use the results from the Levels table to plot specific contours corresponding to percentiles of the data. The Levels table from the output using BWM=2,2 is as follows:

Levels					
Percent	Density	Lower1	Lower2	Upper1	Upper2
25	0.003680	80.26	86.75	93.07	98.02
50	0.005471	82.46	88.53	91.24	96.53
75	0.006612	83.92	89.71	89.78	95.35
95	0.007451	85.39	91.19	88.32	93.86

You can plot the contour levels shown in Figure 4 as follows.

```

data OutLevels;
  set OutLevels;
  if Percent = 25 then
    call symput('den25', left(density) );
  else if Percent = 50 then
    call symput('den50', left(density) );
  else if Percent = 75 then
    call symput('den75', left(density) );
  else if Percent = 95 then
    call symput('den95', left(density) );
run;

proc gcontour data=octden;
  plot Rater*Customer=density /
    levels = &den25 &den50 &den75 &den95
    vminor = 0
    hminor = 0
    vaxis = axis1
    legend = legend1;
  axis1 label = ( r=0 a=90 );
  legend1 label = ( 'Levels' )
    value = ( '95' '75' '50' '25' );
run;

```

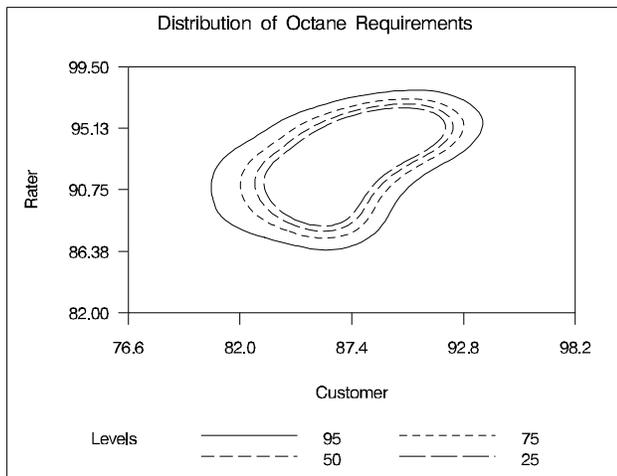


Figure 4. Level Contours for BWM=2,2

For large data sets, the number of kernel evaluations can be prohibitive in the bivariate case. To avoid this problem, the KDE procedure uses a binning method based on the Fast Fourier Transform which is practically as accurate as direct evaluation; for details, refer to Fan and Marron (1993)

Note that facilities for kernel density estimation in the univariate case are currently available in SAS/INSIGHT software and in the CAPABILITY procedure in SAS/QC software; refer to SAS Institute Inc. (1995a, 1995b). Support for the bivariate case, along with interactive 3D graphics, is being added in SAS/INSIGHT software; see Cohen *et al.* (1998).

Nonparametric Regression: The LOESS Procedure

The LOESS procedure implements a nonparametric method for estimating local regression surfaces pioneered by Cleveland (1979); also refer to Cleveland *et al.* (1988) and Cleveland and Grosse (1991). This method is commonly referred to as *loess*, which is short for *local regression*.

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 ϵ_i is a random error. The idea of local regression is that at a predictor x , the regression function $g(x)$ can be locally approximated by the value of a function in some specified parametric class.

More specifically, the method of weighted least squares is used to fit linear or quadratic functions of the predictors at the centers of neighborhoods whose radii are chosen so that each neighborhood contains a specified percentage of the data points. The fraction of the data in each local neighborhood, called the smoothing parameter, controls the smoothness of the estimated surface. Data points in a given local neighborhood are weighted by a function of their distance from the center of the neighborhood that decreases smoothly from one at the center to zero on the boundary 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 local fitting at a selected sample of points in the predictor space and then blend these local polynomials to obtain a regression surface. The points at which the least squares fitting is done are chosen as the vertices of cells of a k -d tree decomposition of the regressor data.

The first step in the decomposition is to select the regressor with widest range and to divide the data into two cells about the median of this regressor. This step is then applied recursively to each of two resulting cells. The process terminates when all the cells contain fewer than a specified number of points.

Statistical inference can be done when the ϵ_i are iid normal random variables with zero mean. Furthermore, robustness to outliers in the data can be achieved and inference can be done when the ϵ_i have a symmetric, long-tailed distribution by performing iterative reweighting. In all but the first iteration the i th data point, x_i , is weighted by an appropriate function of the residual at that point at the previous iteration.

The following example illustrates the use of the LOESS procedure for a single regressor. During an earthquake, both its magnitude and duration are recorded. The following SAS statements create a data set named QUAKES which contains the magnitudes (measured on the Richter scale) and the logs (base 10) of the durations in seconds for 225 earthquakes which occurred on the Island of Hawaii in 1975 and 1976; refer to Bevins and Wright (1992).

```

data Quake;
  input Magnitude logDuration;
datalines;
3.35 3.5
3.35 3.4
3.35 3.3
... ..

```

A plot of the data shown in Figure 5 shows that there is a strong relationship between Magnitude and logDuration.

The following statements compute a loess fit for the data.

```
ods output OutputStatistics=OutQuake;
proc loess data=Quake;
  model logDuration = Magnitude /
    cli
    smooth = 0.1;
run;
```

The MODEL statement specifies the dependent variable and the regressor variables, which are separated by an equal sign. A linear function (the default) is to be fit locally, and the SMOOTH= option specifies the smoothing parameter. The CLI option requests pointwise 95% confidence limits. The following statements create the plot displayed in Figure 5.

```
symbol1 v=none i=join w=2;
symbol2 v=none i=join w=2 l=2;
symbol3 v=none i=join w=2 l=2;
symbol4 v=plus h=2.5 pct;

proc gplot data=OutQuake;
  plot ( Pred LowerCL UpperCL DepVar ) * Magnitude /
    overlay
    hminor = 0
    vminor = 0
    vaxis = axis1
    frame;
  axis1 label = ( r = 0 a = 90 );
  format Pred Magnitude 3.1 ;
run;
```

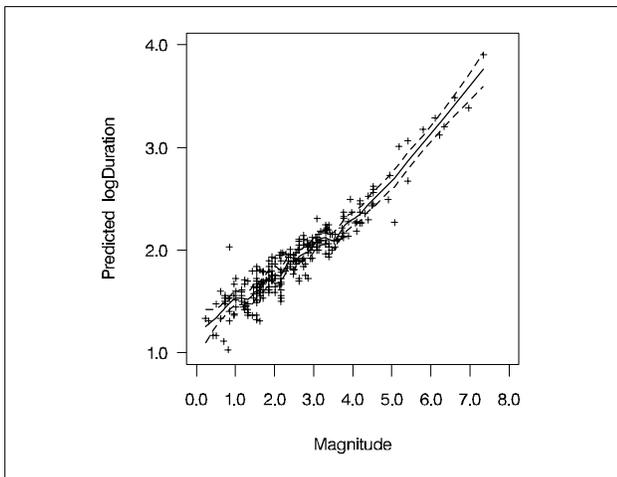


Figure 5. Loess Fit for Earthquake Data

For clarity, Figure 6 shows the fit without the data. Figure 6 reveals a slight bend in the relationship between logDuration and Magnitude, which is otherwise nearly linear.

The next example illustrates the use of the LOESS procedure in fitting a highly nonlinear surface in the presence of significant noise and outliers. The statements below create and display a data set named HATIRREGULAR in which the variables X, Y, and Z are constructed by irregularly sampling a "cowboy hat" surface with an off-centered elliptical spike, white noise, and random spikes.

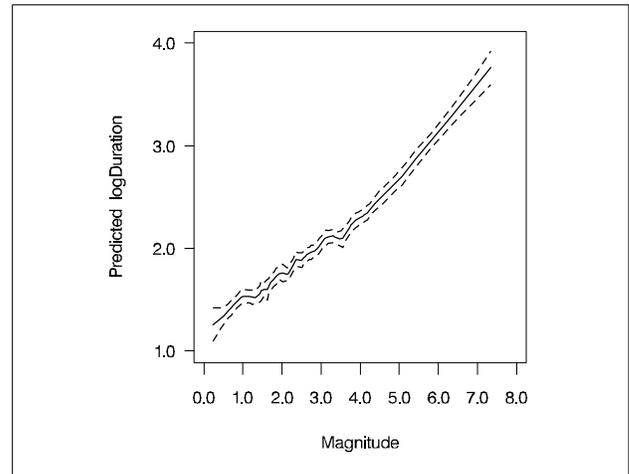


Figure 6. Loess Fit for Earthquake Data

```
data HatIrregular(drop=i);
do i=1 to 1000;
  x = -5+10*ranuni(12345);
  y = -5+10*ranuni(12345);
  z = sin(sqrt(x*x+y*y)) +
    5*exp(-4*(x-2)*(x-2)-y*y) +
    rannor(123);
  if ( ranuni(123) < 0.2 )
    then z = z+10*(ranuni(123)-0.5);
output;
  x = 1 + 3*ranuni(12345);
  y = -1.5 + 3*ranuni(1234567);
  z = sin(sqrt(x*x+y*y)) +
    5*exp(-4*(x-2)*(x-2)-y*y) +
    rannor(123);
output;
end;

title "Scatter Plot of Hat Surface Data";
proc g3d data=HatIrregular;
  scatter y*x = z /
    zticknum = 5
    zmin = -8
    zmax = 8;
run;
```

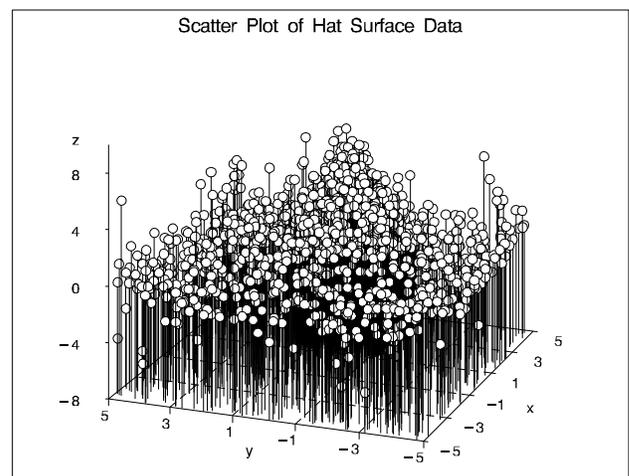


Figure 7. Scatter Plot of Cowboy Hat Surface Data

An additional data set named SCOREHAT provides a regular grid of values at which the fitted model will be scored.

```

data ScoreHat;
  do x = -4 to 4 by 0.2;
    do y = -4 to 4 by 0.2;
      zTrue = sin(sqrt(x*x+y*y)) +
        5*exp(-4*(x-2)*(x-2)-y*y);
    output;
  end;
end;

title "Plot of True Surface";
proc g3d data=ScoreHat;
  plot y*x=zTrue / tilt      = 75
      rotate    = 45
      zticknum = 5
      zmin     = -2
      zmax     = 6;
run;

```

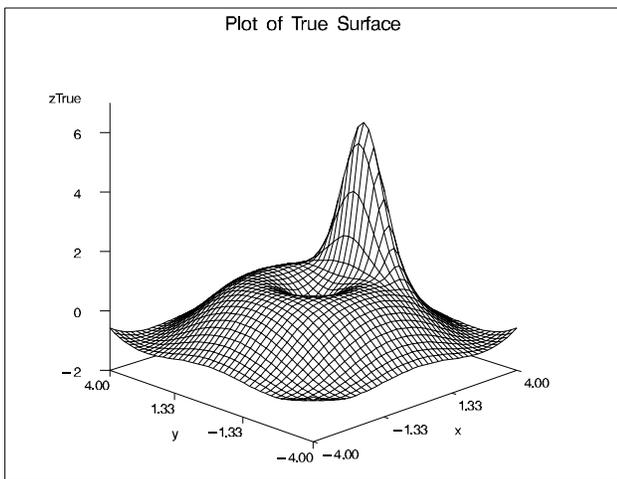


Figure 8. Plot of True Surface

The following statements fit a loess model to the data in HATIRREGULAR and score the model at the points in SCOREHAT. The ODS OUTPUT statement creates an output data set containing the scored data. Here, the option DEGREE=2 in the MODEL statement requests a quadratic fit, the BUCKET= option specifies the number of points in k-d tree buckets, and the ITERATIONS= option specifies the number of reweighting iterations.

```

ods output ScoreResults=OutScore;
proc loess data=HatIrregular;
  model z=x y / degree=2
        smooth=0.2
        bucket=15
        iterations=3;
  score data=ScoreHat;
run;
proc g3d data=OutScore;
  title "Plot of Scored LOESS Surface";
  plot y*x=p_z / tilt=75
      rotate=45
      zticknum=5
      zmin=-2
      zmax=6;
run;

```

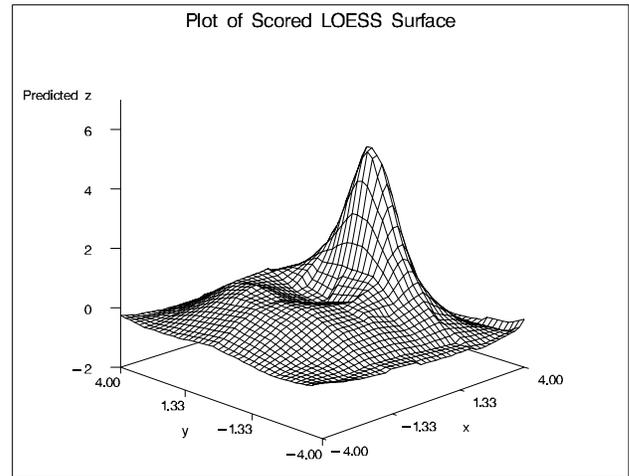


Figure 9. Plot of Scored LOESS Surface

Note that an interactive facility for loess fitting with a single regressor is available in SAS/INSIGHT software; refer to SAS Institute Inc. (1995a).

Nonparametric Regression: The TPSPLINE Procedure

The TPSPLINE procedure uses a penalized least squares method to estimate multivariate regression surfaces with thin-plate smoothing splines. The TPSPLINE procedure allows great flexibility in the form of the regression surface and requires no assumptions of a parametric form for the model. The generalized cross validation (GCV) function is used to select the smoothing parameter.

The TPSPLINE procedure complements the methods provided by standard SAS regression procedures such as the GLM, REG, and NLIN procedures. These procedures can handle most situations in which the user can specify the regression model and the model is known up to a finite number of parameters. However, when the user has no prior knowledge about the model or knows that the data cannot be represented by a model with a finite number of parameters, the TPSPLINE procedure can be used to explore the data.

Smoothing splines are local in nature, as is the case with other non-parametric regression methods. In kernel smoothing, the smoother uses an explicitly defined set of local weights, defined by the kernel, to produce the estimate at each target value. Usually a kernel smoother uses weights that decrease in a smooth fashion as one moves away from the target points. The regression spline represents the fit as a piecewise polynomial. The regions that define the pieces are separated by a sequence of knots, and it is customary to force the piecewise polynomials to join smoothly at these knots. By allowing more knots, the family of curves becomes more flexible.

Mathematically, smoothing splines emerge as the solution to an optimization problem. They were generally regarded as numerical analysis tools until extensive research, pi-

oneered by Grace Wahba, demonstrated that they have useful statistical properties and deserve consideration as a method for performing non-parametric regression analysis. It is now well-recognized that smoothing splines and their variants provide extremely flexible data analysis tools. For more details, refer to Wahba (1990), Duchon (1976), Bates *et. al* (1987), Hastie and Tibshirani (1990), Eubank (1989), Wand and Jones (1995), Hardle and Mammen (1993) and papers referenced there.

You can use the TPSPLINE procedure to fit either a non-parametric model or a semi-parametric model. For the i th observation, define y_i as the response value associated with $(\mathbf{x}_i, \mathbf{z}_i)$, where \mathbf{x}_i is a d -dimensional covariate vector and \mathbf{z}_i is a p -dimensional covariate vector. Assuming that the relation between \mathbf{z}_i and y_i is linear but the relation between \mathbf{x}_i and y_i is not known, you can fit the data using the semi-parametric model

$$y_i = f(\mathbf{X}_i) + \mathbf{z}_i \beta + \epsilon_i,$$

where f is an unknown function which is assumed to be reasonably smooth and $\epsilon_i, i = 1, \dots, n$ are independent, zero-mean random errors and β is a p -dimensional vector of unknown parameters. Here, $\mathbf{z}_i \beta$ is the parametric portion of the model, and \mathbf{z}_i represents the regression variables. The function $f(\mathbf{x}_i)$ is the non-parametric part of the model, and \mathbf{x}_i represents the smoothing variables.

In order to obtain an estimate which fits the data well and, at the same time, has some degree of smoothness, the penalized least squares method is used. This method minimizes the quantity

$$S_\lambda(f) = \frac{1}{n} \sum_{i=1}^n (y_i - f(\mathbf{x}_i) - \mathbf{z}_i \beta)^2 + \lambda J_2(f),$$

where $J_2(f)$ is the penalty on the roughness of f , which is typically defined as the integral of the square of the second derivative of f . The first term measures the goodness-of-fit to the data, and the second term measures the smoothness of f . The multiplier λ is called the smoothing parameter because it governs the tradeoff between smoothness and goodness of fit. A large value of λ penalizes estimates with large second derivatives, and conversely, a small value of λ rewards goodness of fit.

The estimate f_λ is selected from a reproducing kernel Hilbert space, and it can be represented as a linear combination of a sequence of basis functions. Hence, the final estimate of f can be written as

$$f_\lambda(\mathbf{x}_i) = \theta_0 + \sum_{j=1}^d \theta_j x_{ij} + \sum_{j=1}^n \delta_j B_j(\mathbf{x}_i),$$

where B_j is a basis function which depends on \mathbf{x}_j , and θ_j and δ_j are coefficients to be estimated.

The smoothing parameter can be chosen by minimizing the generalized cross validation (GCV) function.

If one expresses the fit as a linear operation

$$\hat{\mathbf{y}} = \mathbf{A}(\lambda)\mathbf{y},$$

then $\mathbf{A}(\lambda)$ is referred to as the “hat” matrix, and the GCV function $V(\lambda)$ is defined as

$$V(\lambda) = \frac{(1/n) \|(\mathbf{I} - \mathbf{A}(\lambda))\mathbf{y}\|^2}{[(1/n) \text{tr}(\mathbf{I} - \mathbf{A}(\lambda))]^2}.$$

For a fixed λ , the coefficients (θ, δ, β) are estimated by solving an $n \times n$ system.

The syntax for the TPSPLINE procedure is similar to that of other regression procedures in the SAS System. For simple applications, only the PROC TPSPLINE and MODEL statements are required, as illustrated in the following example which uses data provided by Bates *et. al* (1987).

The following example illustrates the use of the TPSPLINE procedure with a data set named MEASURE which contains the variables X1, X2, and Y.

```
data measure;
input x1 x2 y @@;
datalines;
-1.0 -1.0 15.54483570 -1.0 -1.0 15.76312613
-.5 -1.0 18.67397826 -.5 -1.0 18.49722167
0 -1.0 19.66086310 0 -1.0 19.80231311
... ..
run;
```

The goal is to fit a surface by using X1 and X2 to model Y. The values of X1 and X2 are distributed regularly on a $[-1 \times 1] \times [-1 \times 1]$ square, and the values of Y were generated by adding a random error to a function $f(x_1, x_2)$. The data are plotted in Figure 10.

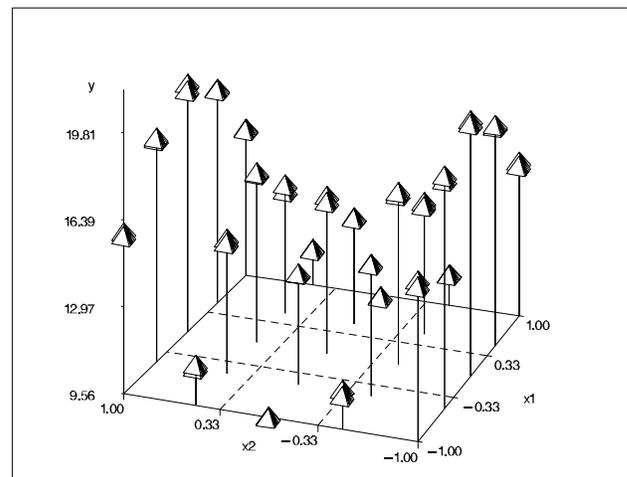


Figure 10. Plot of Data Set MEASURE

The following statements fit a thin plate spline to the data:

```
proc tpspline data=measure;
model y=(x1 x2) / lambda = -4 to -2 by 0.2;
output out=estimate pred 195 u95;
run;
```

In the MODEL statement, the variables X1 and X2 are enclosed by parentheses to indicate that they are smoothing variables as opposed to regression variables. The

LAMBDA= option requests a list of GCV values with $\log_{10}(n\lambda)$ ranging from -4 to -2. The OUTPUT statement specifies that the predicted values and 95% confidence limits are to be saved in an output data set named ESTIMATE. Output from the procedure is displayed in Figure 11, and a partial listing of ESTIMATE is shown in Figure 12.

Obs	x	x2	y	y_p	y_u95
1	-1.0	-1.0	15.5448	15.6474	15.5115
2	-1.0	-1.0	15.7631	15.6474	15.5115
3	-0.5	-1.0	18.6740	18.5783	18.4430
...					
50	1.0	1.0	15.9014	15.8761	15.7402

Figure 12. Data Set ESTIMATE

```

The TPSPLINE Procedure

Summary of Input Values

Number of observations           50
Number of unique observations    25
Number of independent variables  1
Number of regression variables  in the model  0
Number of smoothing variables  in the model  2
Dimension of polynomial space    3

GCV Function
log10 of (nLambda)      y
-4      0.019215
-3.8    0.019148
-3.6    0.019082
-3.4    0.019074
-3.2    0.019286
-3      0.020117
-2.8    0.022462
-2.6    0.028132
-2.4    0.040411
-2.2    0.064699
-2      0.109387

Summary Statistics
of Final Estimation

Parameters      y
Lambda          0.000006681
Smoothing Penalty 2558.143225
RSS             0.246110
Tr(I-A)        25.406797
DF             24.593203
Standard Deviation 0.098421

```

Figure 11. Output from the TPSPLINE Procedure

The data set MEASURE contains 50 observations with 25 unique design points. The value of λ that minimizes the GCV function is around $10^{-3.5}/50$. The final fit is based on $\lambda = 0.000006681$. The residual sum of squares (RSS) for this fit is 0.246110, and the degrees of freedom is 24.593203. The standard deviation, defined as $\text{RSS}/(\text{Tr}(I-A))$, is 0.098421. These values differ slightly from those obtained by Bates *et al.* (1987) who used somewhat different stopping criteria in the search for λ ; however, the final fits themselves agree closely. A plot of the fitted surface in ESTIMATE is shown in Figure 13; the surface is coarse because the data points are sparse. The following statements produce a smoother surface. First, the DATA step is used to generate a finer grid. Then the SCORE statement is used to evaluate the fitted surface at these design points. The fitted values saved in PRED_Y are displayed in Figure 14, which suggests that a quadratic parametric model would also provide a good fit.

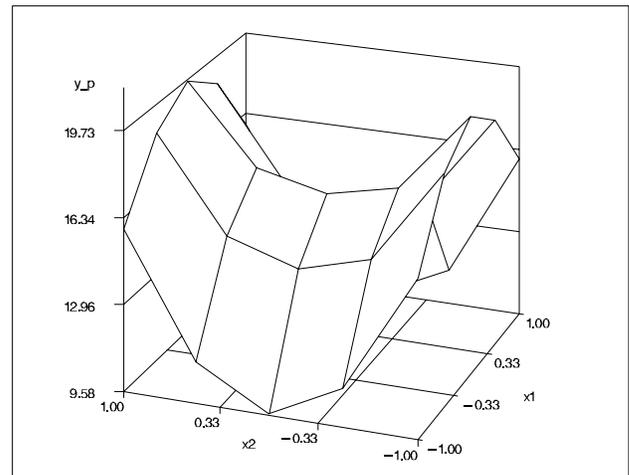


Figure 13. Fitted Surface Using Values in ESTIMATE

```

data pred;
  do x1=-1 to 1 by 0.1;
    do x2=-1 to 1 by 0.1; output;
    end;
  end;

proc tpspline data=measure;
  model y = (x1 x2) / lambda = -4 to -2 by 0.1;
  score data=pred out=pred_y;
run;

```

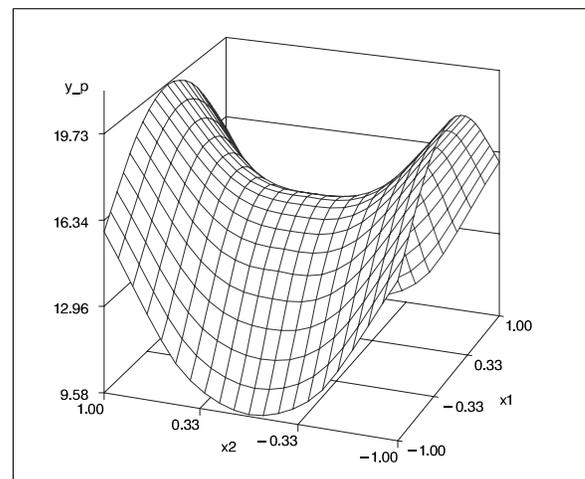


Figure 14. Fitted Surface Using Values in PRED_Y

Note that the computational facilities of PROC TPSPLINE are being made available in SAS/INSIGHT software and as functions in SAS/IML software.

ACKNOWLEDGEMENTS

We are grateful to Robert Cohen, Donna Sawyer, Russell Wolfinger, and Dong Xiang of SAS Institute Inc. for valuable assistance in the preparation of this paper.

REFERENCES

Bates, D., Lindstrom, M., Wahba, G., and Yandell, B. (1987), "GCVPACK-routines for generalized cross validation," *Communications in Statistics B--Simulation and Computing*, 16, 263-297.

Bevens, D. and Wright, T. L. (1992), *The Thomas A. Jaggar Museum Guidebook, Hawaii Volcanoes National Park*, Hawaii Natural History Association.

Bowman, A. and Foster, P. (1992), *Density Based Exploration of Bivariate Data*, Department of Statistics, University of Glasgow, Technical Report No. 92-1.

Cohen, M., Chen, H., Yuan, Y., and Wicklin, F. (1998), "New Features in SAS/INSIGHT Software in Version 7," *Proceedings of the 23rd SAS Users Group International Conference*, to appear.

Cleveland, W. S. (1979), "Robust locally-weighted regression and smoothing scatterplots," *Journal of the American Statistical Association*, 74, 829-836.

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

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

Duchon, J. (1976), "Splines Minimizing Rotation-Invariant Semi-Norms in Sobolev Spaces," in *Constructive Theory of Functions of Several Variables*, eds. W. Schempp and K. Zeller, 85-100.

Eubank, R. (1989), *Spline Smoothing and Nonparametric Regression*, New York: Marcel Dekker.

Fan, J. and Marron, J.S. (1993), "Fast Implementations of Nonparametric Curve Estimators," *Journal of Computational and Graphical Statistics*, 3, 35-56.

Green, P. and Silverman, B. (1994), *Nonparametric Regression and Generalized Linear Models*, New York: Chapman and Hall.

Hardle, W. and Mammen, E. (1993), "Comparing nonparametric versus parametric regression fits," *The Annals of Statistics*, 21, 1926-1947.

Hastie, T. and Tibshirani, R. (1990), *Generalized Additive*

Models, New York: Chapman and Hall.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1996), "A Brief Survey of Bandwidth Selection for Density Estimation," *Journal of the American Statistical Association*, 91, 401-407.

Marron, J. S. (1989), "Automatic smoothing parameter selection: a survey," *Empirical Economics*, 13, 187-208.

Marron, J. S. (1998), personal communication with the authors.

Rodriguez, R. N. and Taniguchi, B. Y. (1980), "A New Statistical Model for Predicting Customer Octane Satisfaction Using Trained-Rater Observations," SAE Technical Paper 801356, *Transactions of the Society of Automotive Engineers*, 4213-4240.

SAS Institute Inc. (1995a), *SAS/INSIGHT User's Guide, Version 6, Third Edition*, Cary, NC: SAS Institute Inc.

SAS Institute Inc. (1995b), *SAS/QC Software: Usage and Reference, Version 6, First Edition*, Cary, NC: SAS Institute Inc.

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

Wand, M.P. (1993), *Fast Computation of Multivariate Kernel Estimators*, University of New South Wales, Australian Graduate School of Management, Working Paper Series 93-007.

Wand, M.P. and Jones, M.C. (1993), "Comparison of Smoothing Parameterizations in Bivariate Kernel Density Estimation," *Journal of the American Statistical Association*, 88, 520-528.

Wand, M. P. and Jones, M. C. (1995), *Kernel Smoothing*, New York: Chapman and Hall.

Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia: Society for Industrial and Applied Mathematics.

AUTHORS

Robert N. Rodriguez, SAS Institute Inc., SAS Campus Drive, Cary, NC 27513. Phone (919) 677-8000 x7650. FAX (919) 677-4444. Email sasnr@wnx.sas.com

Maura E. Stokes, SAS Institute Inc., SAS Campus Drive, Cary, NC 27513. Phone (919) 677-8000 x7172. FAX (919) 677-4444. Email sasmzs@wnx.sas.com

SAS, SAS/STAT, SAS/IML, SAS/INSIGHT, and SAS/QC are registered trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.