Chapter 56
The RSREG Procedure

Chapter Table of Contents

Chapter 56
The RSREG Procedure

Overview

The RSREG procedure uses the method of least squares to fit quadratic response surface regression models. Response surface models are a kind of general linear model in which attention focuses on characteristics of the fit response function and in particular, where optimum estimated response values occur.

In addition to fitting a quadratic function, you can use the RSREG procedure to

- test for lack of fit
- test for the significance of individual factors
- analyze the canonical structure of the estimated response surface
- compute the ridge of optimum response
- predict new values of the response

Comparison to Other SAS Software

Other SAS/STAT procedures can be used to fit the response surface, but the RSREG procedure is more specialized. The following statements model a three-factor response surface in PROC RSREG:

```
proc rsreg;
   model y=x1 x2 x3;
run;
```

These statements are more compact than the statements for other regression procedures in SAS/STAT software. For example, the equivalent statements for the GLM procedure are

```
proc glm;
   model y=x1 x1*x1
           x2 x1*x2 x2*x2
           x3 x1*x3 x2*x3 x3*x3;
run;
```

Additionally, PROC RSREG includes specialized methodology for analyzing the fitted response surface, such as canonical analysis and optimum response ridges.

Note that the ADX Interface in SAS/QC software provides an *interactive* environment for constructing and analyzing many different kinds of experiments, including response surface experiments. The ADX Interface is the preferred interactive SAS System tool for analyzing experiments, since it includes facilities for checking underlying assumptions and graphically optimizing the response surface. The RSREG procedure is appropriate for analyzing experiments in a batch environment.

Terminology

Variables are referred to according to the following conventions:

factor variables	independent variables used in constructing the quadratic response surface. To estimate the necessary parameters, each variable must have at least three distinct values in the data. Independent variables must be numeric.
response variables	the dependent variables to which the quadratic response surface is fit. Dependent variables must be numeric.
covariates	additional independent variables for use in the regression but not in the formation of the quadratic response surface. Covariates must be numeric.
WEIGHT variable	a variable for weighting the observations in the regression. The WEIGHT variable must be numeric.
ID variables	variables not in the above lists that are transferred to an output data set containing statistics for each observation in the input data set. This data set is created using the OUT= option in the PROC RSREG statement. ID variables can be either character or numeric.
BY variables	variables for grouping observations. Separate analyses are obtained for each BY group. BY variables can be either character or numeric.

Getting Started

A Response Surface with a Simple Optimum

This example uses the three-factor quadratic model discussed in John (1971). Schneider and Stockett (1963) performed an experiment aimed at reducing the unpleasant odor of a chemical produced with several factors. The objective is to minimize the unpleasant odor of a chemical. The following statements read the data.

```
title 'Response Surface with a Simple Optimum';
data smell;
   input Odor T R H @@;
   label
      T = "Temperature"
      R = "Gas-Liquid Ratio"
      H = "Packing Height";
   datalines;
 66 40 .3 4     39 120 .3 4     43 40 .7 4     49 120 .7  4
 58 40 .5 2     17 120 .5 2     -5 40 .5 6    -40 120 .5  6
 65 80 .3 2      7  80 .7 2     43 80 .3 6    -22  80 .7  6
-31 80 .5 4    -35  80 .5 4    -26 80 .5 4
;
```

The INPUT statement names the variables contained in the SAS data set smell; the variable Odor is the response, while the variables T, R, and H are the independent factors.

The following statements invoke PROC RSREG on the data set smell. Figure 56.1 through Figure 56.3 display the results of the analysis, including a lack-of-fit test requested with the LACKFIT option.

```
proc rsreg data=smell;
   model Odor = T R H / lackfit;
run;
```

Response Surface with a Simple Optimum

The RSREG Procedure

Coding Coefficients for the Independent Variables

Factor	Subtracted off	Divided by
T	80.000000	40.000000
R	0.500000	0.200000
H	4.000000	2.000000

Response Surface for Variable Odor

Response Mean	15.200000
Root MSE	22.478508
R-Square	0.8820
Coefficient of Variation	147.8849

Regression	DF	Type I Sum of Squares	R-Square	F Value	Pr > F
Linear	3	7143.250000	0.3337	4.71	0.0641
Quadratic	3	11445	0.5346	7.55	0.0264
Crossproduct	3	293.500000	0.0137	0.19	0.8965
Total Model	9	18882	0.8820	4.15	0.0657

Residual	DF	Sum of Squares	Mean Square	F Value	Pr > F
Lack of Fit	3	2485.750000	828.583333	40.75	0.0240
Pure Error	2	40.666667	20.333333		
Total Error	5	2526.416667	505.283333		

Figure 56.1. Summary Statistics and Analysis of Variance

Figure 56.1 displays the coding coefficients for the transformation of the independent variables to lie between -1 and 1, simple statistics for the response variable, hypothesis tests for linear, quadratic, and crossproduct terms, and the lack-of-fit test. The hypothesis tests can be used to gain a rough idea of importance of the effects; here the crossproduct terms are not significant. However, the lack-of-fit for the model is significant, so more complicated modeling or further experimentation with additional variables should be performed before firm statements are made concerning the underlying process.

Response Surface with a Simple Optimum

The RSREG Procedure

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Parameter Estimate from Coded Data
Intercept	1	568.958333	134.609816	4.23	0.0083	-30.666667
T	1	-4.102083	1.489024	-2.75	0.0401	-12.125000
R	1	-1345.833333	335.220685	-4.01	0.0102	-17.000000
H	1	-22.166667	29.780489	-0.74	0.4902	-21.375000
T*T	1	0.020052	0.007311	2.74	0.0407	32.083333
R*T	1	1.031250	1.404907	0.73	0.4959	8.250000
R*R	1	1195.833333	292.454665	4.09	0.0095	47.833333
H*T	1	0.018750	0.140491	0.13	0.8990	1.500000
H*R	1	-4.375000	28.098135	-0.16	0.8824	-1.750000
H*H	1	1.520833	2.924547	0.52	0.6252	6.083333

Factor	DF	Sum of Squares	Mean Square	F Value	Pr > F	Label
T	4	5258.016026	1314.504006	2.60	0.1613	Temperature
R	4	11045	2761.150641	5.46	0.0454	Gas-Liquid Ratio
H	4	3813.016026	953.254006	1.89	0.2510	Packing Height

Figure 56.2. Parameter Estimates and Hypothesis Tests

Parameter estimates and the factor ANOVA are shown in Figure 56.2. Looking at the parameter estimates, you can see that the crossproduct terms are not significantly different from zero, as noted previously. The "Estimate" column contains estimates based on the raw data, and the "Parameter Estimate from Coded Data" column contains those based on the coded data. The factor ANOVA table displays tests for all four parameters corresponding to each factor—the parameters corresponding to the linear effect, the quadratic effect, and the effects of the cross products with each of the other two factors. The only factor with a significant over-all effect is R, indicating that the level of noise left unexplained by the model is still too high to estimate the effects of T and H accurately. This may be due to the lack of fit.

```
          Response Surface with a Simple Optimum

                    The RSREG Procedure
     Canonical Analysis of Response Surface Based on Coded Data

                   Critical Value
     Factor           Coded         Uncoded    Label

     T             0.121913       84.876502    Temperature
     R             0.199575        0.539915    Gas-Liquid Ratio
     H             1.770525        7.541050    Packing Height

          Predicted value at stationary point: -52.024631


                                  Eigenvectors
     Eigenvalues              T               R               H

       48.858807       0.238091        0.971116       -0.015690
       31.103461       0.970696       -0.237384        0.037399
        6.037732      -0.032594        0.024135        0.999177

                  Stationary point is a minimum.
```

Figure 56.3. Canonical Analysis and Eigenvectors

Figure 56.3 contains the canonical analysis and eigenvectors. The canonical analysis indicates that the directions of principle orientation for the predicted response surface are along the axes associated with the three factors, confirming the small interaction effect in the Regression ANOVA. The largest eigenvalue (48.8588) corresponds to the eigenvector $\{0.238091,\ 0.971116,\ -0.015690\}$, the largest component of which (0.971116) is associated with R; similarly, the second largest eigenvalue (31.1035) is associated with T. The third eigenvalue (6.0377), associated with H, is quite a bit smaller than the other two, indicating that the response surface is relatively insensitive to changes in this factor. The coded form of the canonical analysis indicates that the estimated response surface is at a minimum when T and R are both near the middle of their respective ranges and H is relatively high; in uncoded, terms, the model predicts that the unpleasant odor will be minimized when $T = 84.876502$, $R = 9.539915$, and $H = 7.541050$.

To plot the response surface with respect to two of the factor variables, first fix H, the least significant factor variable, at its estimated optimum value and generate a grid of points for T and R. To ensure that the grid data do not affect parameter estimates, the response variable (Odor) is set to missing. (See the "Missing Values" section on page 3046.) The following statements produce and graph the necessary data. Initial data steps creates a grid over T and R, with H set to a constant value, and combine this grid with the original data. Then, PROC RSREG is used to create predictions for the combined data. Finally, PROC G3D is used to create a surface plot of the predictions.

```
data grid;
   do;
      Odor =  .  ;
      H    = 7.541;
      do T = 20 to 140 by 5;
```

```
         do R = .1 to .9 by .05;
            output;
         end;
      end;
   end;
data grid;
   set smell grid;
run;

proc rsreg data=grid out=predict noprint;
   model Odor = T R H / predict;
run;

data plot;
   set predict;
   if H = 7.541;
proc g3d data=plot;
   plot T*R=Odor / rotate=38 tilt=75 xticknum=3 yticknum=3
   zmax=300 zmin=-60 ctop=red cbottom=blue caxis=black;
run;
```

The first DATA step creates grid points for T and R at H=7.541 and sets Odor to missing, and the second DATA step concatenates these grid points with the original data. Predicted values are created in the SAS data set predict by invoking the RSREG procedure with the PREDICT option in the MODEL statement. The analysis is not displayed due to the NOPRINT option. The third DATA step subsets the predicted values over just the grid points (excluding the predictions at the original points). PROC G3D is then used to create the three-dimensional plot shown in Figure 56.4.

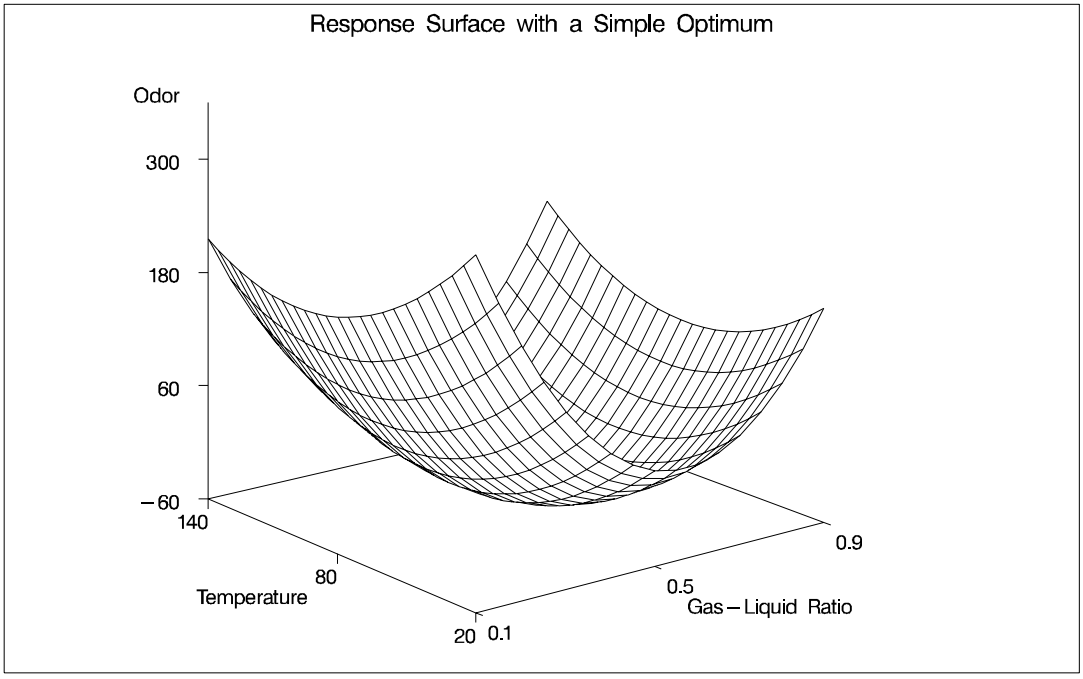


Figure 56.4. The Response Surface Obtained from the PREDICT Option

Syntax

The following statements are available in PROC RSREG.

PROC RSREG < *options* > **;**
MODEL *responses= independents* < */ options* > **;**

RIDGE < *options* > **;**
WEIGHT *variable* **;**
ID *variables* **;**
BY *variables* **;**

The PROC RSREG and MODEL statements are required. The BY, ID, MODEL, RIDGE, and WEIGHT statements are described after the PROC RSREG statement, and they can appear in any order.

PROC RSREG Statement

PROC RSREG < *options* > **;**

The PROC RSREG statement invokes the procedure. You can specify the following options in the PROC RSREG statement.

DATA=*SAS-data-set*
specifies the input SAS data set that contains the data to be analyzed. By default, PROC RSREG uses the most recently created SAS data set.

NOPRINT
suppresses the normal display of results when only the output data set is required. For more information, see the description of the NOPRINT option in the "MODEL Statement" and "RIDGE Statement" sections. Note that this option temporarily disables the Output Delivery System (ODS); see Chapter 15, "Using the Output Delivery System," for more information.

OUT=*SAS-data-set*
creates an output SAS data set that contains statistics for each observation in the input data set. In particular, this data set contains the BY variables, the ID variables, the WEIGHT variable, the variables in the MODEL statement, and the output options requested in the MODEL statement. You must specify output options in the MODEL statement; otherwise, the output data set is created but contains no observations. To create a permanent SAS data set, you must specify a two-level name (refer to the discussion in *SAS Language Reference: Concepts* for more information on permanent SAS data sets). For details on the data set created by PROC RSREG, see the "Output Data Sets" section on page 3050.

BY Statement

BY *variables* **;**

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

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

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

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

ID Statement

ID *variables* **;**

The ID statement names variables that are to be transferred to the data set created by the OUT= option in the PROC RSREG statement.

MODEL Statement

MODEL *responses=independents* < */ options* > **;**

The MODEL statement lists response (dependent) variables followed by an equal sign and then lists independent variables, some of which may be covariates. The output options to the MODEL statement specify which statistics are output to the data set created using the OUT= option in the PROC RSREG statement. If none of the options are selected, the data set is created but contains no observations. The option keywords become values of the special variable _TYPE_ in the output data set. Any of the following options can be specified.

Task	Options
Analyze Original Data	NOCODE
Fit Model to First BY Group Only	BYOUT
Declare Covariates	COVAR=
Request Additional Statistics	PRESS
Request Additional Tests	LACKFIT
Suppress Displayed Output	NOANOVA NOOPTIMAL NOPRINT
Output Statistics	ACTUAL PREDICT RESIDUAL L95 U95 L95M U95M D

ACTUAL
specifies that the observed response values from the input data set be written to the output data set.

BYOUT
uses only the first BY group to estimate the model. Subsequent BY groups have scoring statistics computed in the output data set only. The BYOUT option is used only when a BY statement is specified.

COVAR=*n*
declares that the first n variables on the right-hand side of the model are simple linear regressors (covariates) and not factors in the quadratic response surface. By default, PROC RSREG forms quadratic and crossproduct effects for all regressor variables in the MODEL statement. See the "Handling Covariates" section on page 3048 for more details and Example 56.2 on page 3058 for an example using covariates.

D
specifies that Cook's D influence statistic be written to the output data set. See Chapter 3, "Introduction to Regression Procedures," for details and formulas.

LACKFIT
performs a lack-of-fit test. Refer to Draper and Smith (1981) for a discussion of lack-of-fit tests.

L95
specifies that the lower bound of a 95% confidence interval for an individual predicted value be written to the output data set. The variance used in calculating this bound is a function of both the mean square error and the variance of the parameter estimates. See Chapter 3 for details and formulas.

L95M
specifies that the lower bound of a 95% confidence interval for the expected value of the dependent variable be written to the output data set. The variance used in calculating this bound is a function of the variance of the parameter estimates. See Chapter 3 for details and formulas.

NOANOVA
NOAOV
suppresses the display of the analysis of variance and parameter estimates from the model fit.

NOCODE
performs the canonical and ridge analyses with the parameter estimates derived from fitting the response to the original values of the factors variables, rather than their coded values (see the "Coding the Factor Variables" section on page 3046 for more details.) Use this option if the data are already stored in a coded form.

NOOPTIMAL
NOOPT
suppresses the display of the canonical analysis for the quadratic response surface.

NOPRINT
suppresses the display of both the analysis of variance and the canonical analysis.

PREDICT
specifies that the values predicted by the model be written to the output data set.

PRESS
computes and displays the predicted residual sum of squares (PRESS) statistic for each dependent variable in the model. The PRESS statistic is added to the summary information at the beginning of the analysis of variance, so if the NOANOVA or NOPRINT option is specified, PRESS has no effect. See Chapter 3 for details and formulas.

RESIDUAL
specifies that the residuals, calculated as ACTUAL − PREDICTED, be written to the output data set.

U95
specifies that the upper bound of a 95% confidence interval for an individual predicted value be written to the output data set. The variance used in calculating this bound is a function of both the mean square error and the variance of the parameter estimates. See Chapter 3 for details and formulas.

U95M
specifies that the upper bound of a 95% confidence interval for the expected value of the dependent variable be written to the output data set. The variance used in calculating this bound is a function of the variance of the parameter estimates. See Chapter 3 for details and formulas.

RIDGE Statement

RIDGE < *options* > **;**

A RIDGE statement computes the ridge of optimum response. The ridge starts at a given point $\mathbf{x}_0$, and the point on the ridge at radius r from $\mathbf{x}_0$ is the collection of factor settings that optimizes the predicted response at this radius. You can think of the ridge as climbing or falling as fast as possible on the surface of predicted response. Thus, the ridge analysis can be used as a tool to help interpret an existing response surface or to indicate the direction in which further experimentation should be performed.

The default starting point, $\mathbf{x}_0$, has each coordinate equal to the point midway between the highest and lowest values of the factor in the design. The default radii at which the ridge is computed are 0, 0.1, . . . , 0.9, 1. If, as usual, the ridge analysis is based on the response surface fit to coded values for the factor variables (see the "Coding the Factor Variables" section on page 3046 for details), then this results in a ridge that starts at the point with a coded zero value for each coordinate and extends toward, but not beyond, the edge of the range of experimentation. Alternatively, both the center point for the ridge and the radii at which it is to be computed can be specified.

You can specify the following options in the RIDGE statement:

CENTER=*uncoded-factor-values*
gives the coordinates of the point $\mathbf{x}_0$ from which to begin the ridge. The coordinates should be given in the original (uncoded) factor variable values and should be separated by commas. There must be as many coordinates specified as there are factors in the model, and the order of the coordinates must be the same as that used in the MODEL statement. This starting point should be well inside the range of experimentation. The default sets each coordinate equal to the value midway between the highest and lowest values for the associated factor.

MAXIMUM
MAX
computes the ridge of maximum response. Both the MIN and MAX options can be specified; at least one must be specified.

MINIMUM
MIN
computes the ridge of minimum response. Both the MIN and MAX options can be specified; at least one must be specified.

NOPRINT
suppresses the display of the ridge analysis when only an output data set is required.

OUTR=*SAS-data-set*
creates an output SAS data set containing the computed optimum ridge. For details, see the "Output Data Sets" section on page 3050.

RADIUS=*coded-radii*
gives the distances from the ridge starting point at which to compute the optimum.

The values in the list represent distances between coded points. The list can take any of the following forms or can be composed of mixtures of them:

$m_1, m_2, \ldots, m_n$	several values
m TO n	a sequence where m equals the starting value, n equals the ending value, and the increment equals 1
m TO n BY i	a sequence where m equals the starting value, n equals the ending value, and i equals the increment

Mixtures of the preceding forms should be separated by commas. The default list runs from 0 to 1 by increments of 0.1. The following are examples of valid lists.

```
radius=0 to 5 by .5;
radius=0, .2, .25, .3, .5 to 1.0 by .1;
```

WEIGHT Statement

WEIGHT *variable* **;**

When a WEIGHT statement is used, a weighted residual sum of squares

$$\sum_i w_i(y_i - \hat{y}_i)^2$$

is minimized, where w_i is the value of the variable specified in the WEIGHT statement, y_i is the observed value of the response variable, and $\hat{y}_i$ is the predicted value of the response variable.

The observation is used in the analysis only if the value of the WEIGHT statement variable is greater than zero. The WEIGHT statement has no effect on degrees of freedom or number of observations. If the weights for the observations are proportional to the reciprocals of the error variances, then the weighted least-squares estimates are best linear unbiased estimators (BLUE).

Details

Introduction to Response Surface Experiments

Many industrial experiments are conducted to discover which values of given factor variables optimize a response. If each factor is measured at three or more values, a quadratic response surface can be estimated by least-squares regression. The predicted optimal value can be found from the estimated surface if the surface is shaped like a simple hill or a valley. If the estimated surface is more complicated, or if the predicted optimum is far from the region of experimentation, then the shape of the surface can be analyzed to indicate the directions in which new experiments should be performed.

Suppose that a response variable y is measured at combinations of values of two factor variables, x_1 and x_2. The quadratic response-surface model for this variable is written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon$$

The steps in the analysis for such data are

1. model fitting and analysis of variance to estimate parameters
2. canonical analysis to investigate the shape of the predicted response surface
3. ridge analysis to search for the region of optimum response

Model Fitting and Analysis of Variance

The first task in analyzing the response surface is to estimate the parameters of the model by least-squares regression and to obtain information about the fit in the form of an analysis of variance. The estimated surface is typically curved: a "hill" whose peak occurs at the unique estimated point of maximum response, a "valley," or a "saddle-surface" with no unique minimum or maximum. Use the results of this phase of the analysis to answer the following questions:

- What is the contribution of each type of effect—linear, quadratic, and crossproduct—to the statistical fit? The ANOVA table with sources labeled "Regression" addresses this question.
- What part of the residual error is due to lack of fit? Does the quadratic response model adequately represent the true response surface? If you specify the LACKFIT option in the MODEL statement, then the ANOVA table with sources labeled "Residual" addresses this question.
- What is the contribution of each factor variable to the statistical fit? Can the response be predicted as well if the variable is removed? The ANOVA table with sources labeled "Factor" addresses this question.
- What are the predicted responses for a grid of factor values? (See the section "Plotting the Surface" on page 3046 and the "Searching for Multiple Response Conditions" section on page 3046.)

Lack-of-Fit Test

The test for lack-of-fit compares the variation around the model with "pure" variation within replicated observations. This measures the adequacy of the quadratic response surface model. In particular, if there are n_i replicated observations $Y_{i1}, \ldots, Y_{in_i}$ of the response all at the same values $\mathbf{x}_i$ of the factors, then we can predict the true response at $\mathbf{x}_i$ either by using the predicted value $\hat{Y}_i$ based on the model or by using the mean $\bar{Y}_i$ of the replicated values. The test for lack-of-fit decomposes the residual error into a component due to the variation of the replications around their mean value (the "pure" error), and a component due to the variation of the mean values around the model prediction (the "bias" error):

$$\sum_i \sum_{j=1}^{n_i} \left(Y_{ij} - \hat{Y}_i\right)^2 = \sum_i \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_i\right)^2 + \sum_i n_i \left(\bar{Y}_i - \hat{Y}_i\right)^2$$

If the model is adequate, then both components estimate the nominal level of error; however, if the bias component of error is much larger than the pure error, then this constitutes evidence that there is significant lack of fit.

If some observations in your design are replicated, you can test for lack of fit by specifying the LACKFIT option in the MODEL statement. Note that, since all other tests use total error rather than pure error, you may want to hand-calculate the tests with respect to pure error if the lack-of-fit is significant. On the other hand, significant lack-of-fit indicates the quadratic model is inadequate, so if this is a problem you can also try to refine the model, possibly using PROC GLM for general polynomial modeling; refer to Chapter 30, "The GLM Procedure," for more information. Example 56.1 on page 3053 illustrates the use of the LACKFIT option.

Canonical Analysis

The second task in analyzing the response surface is to examine the overall shape of the curve and determine whether the estimated stationary point is a maximum, a minimum, or a saddle point. The canonical analysis can be used to answer the following questions:

- Is the surface shaped like a hill, a valley, a saddle surface, or a flat surface?
- If there is a unique optimum combination of factor values, where is it?
- To which factor or factors are the predicted responses most sensitive?

The eigenvalues and eigenvectors in the matrix of second-order parameters characterize the shape of the response surface. The eigenvectors point in the directions of principle orientation for the surface, and the signs and magnitudes of the associated eigenvalues give the shape of the surface in these directions. Positive eigenvalues indicate directions of upward curvature, and negative eigenvalues indicate directions of downward curvature. The larger an eigenvalue is in absolute value, the more pronounced is the curvature of the response surface in the associated direction. Often, all of the coefficients of an eigenvector except for one are relatively small, indicating that the vector points roughly along the axis associated with the factor corresponding to the single large coefficient. In this case, the canonical analysis can be used to determine the relative sensitivity of the predicted response surface to variations in that factor. (See the "Getting Started" section on page 3033 for an example.)

Ridge Analysis

If the estimated surface is found to have a simple optimum well within the range of experimentation, the analysis performed by the preceding two steps may be sufficient. In more complicated situations, further search for the region of optimum response is required. The method of ridge analysis computes the estimated ridge of optimum response for increasing radii from the center of the original design. The ridge analysis answers the following question:

- If there is not a unique optimum of the response surface within the range of experimentation, in which direction should further searching be done in order to locate the optimum?

You can use the RIDGE statement to compute the ridge of maximum or minimum response.

Coding the Factor Variables

For the results of the canonical and ridge analyses to be interpretable, the values of different factor variables should be comparable. This is because the canonical and ridge analyses of the response surface are not invariant with respect to differences in scale and location of the factor variables. The analysis of variance is not affected by these changes. Although the actual predicted surface does not change, its parameterization does. The usual solution to this problem is to code each factor variable so that its minimum in the experiment is -1 and its maximum is 1 and to carry through the analysis with the coded values instead of the original ones. This practice has the added benefit of making 1 a reasonable boundary radius for the ridge analysis since 1 represents approximately the edge of the experimental region. By default, PROC RSREG computes the linear transformation to perform this coding as the data are initially read in, and the canonical and ridge analyses are performed on the model fit to the coded data. The actual form of the coding operation for each value of a variable is

$$\textit{coded value} \quad = \quad (\textit{original value} - M)/S$$

where M is the average of the highest and lowest values for the variable in the design and S is half their difference.

Missing Values

If an observation has missing data for any of the variables used by the procedure, then that observation is not used in the estimation process. If one or more response variables are missing, but no factor or covariate variables are missing, then predicted values and confidence limits are computed for the output data set, but the residual and Cook's D statistic are missing.

Plotting the Surface

You can generate predicted values for a grid of points with the PREDICT option (see the "Getting Started" section on page 3033 for an example) and then use these values to create a contour plot or a three-dimensional plot of the response surface over a two-dimensional grid. Any two factor variables can be chosen to form the grid for the plot. Several plots can be generated by using different pairs of factor variables.

Searching for Multiple Response Conditions

Suppose you want to find the factor setting that produces responses in a certain region. For example, you have the following data with two factors and three responses:

```
data a;
   input x1 x2 y1 y2 y3;
   datalines;
-1      -1         1.8 1.940  3.6398
-1       1         2.6 1.843  4.9123
 1      -1         5.4 1.063  6.0128
```

```
 1       1         0.7 1.639  2.3629
 0       0         8.5 0.134  9.0910
 0       0         3.0 0.545  3.7349
 0       0         9.8 0.453 10.4412
 0       0         4.1 1.117  5.0042
 0       0         4.8 1.690  6.6245
 0       0         5.9 1.165  6.9420
 0       0         7.3 1.013  8.7442
 0       0         9.3 1.179 10.2762
 1.4142  0         3.9 0.945  5.0245
-1.4142  0         1.7 0.333  2.4041
 0       1.4142    3.0 1.869  5.2695
 0      -1.4142    5.7 0.099  5.4346
;
```

You want to find the values of x1 and x2 that maximize y1 subject to y2<2 and y3<y2+y1. The exact answer is not easy to obtain analytically, but you can obtain a practically feasible solution by checking conditions across a grid of values in the range of interest. First, append a grid of factor values to the observed data, with missing values for the responses.

```
data b;
   set a end=eof;
   output;
   if eof then do;
      y1=.;
      y2=.;
      y3=.;
      do x1=-2 to 2 by .1;
         do x2=-2 to 2 by .1;
            output;
         end;
      end;
   end;
run;
```

Next, use PROC RSREG to fit a response surface model to the data and to compute predicted values for both the observed data and the grid, putting the predicted values in a data set c.

```
proc rsreg data=b out=c;
   model y1 y2 y3=x1 x2 / predict;
run;
```

Finally, find the subset of predicted values that satisfy the constraints, sort by the unconstrained variable, and display the top five predictions.

```
data d;
   set c;
   if y2<2;
```

```
   if y3<y2+y1;

proc sort data=d;
   by descending y1;
run;

data d; set d;
   i = _n_;
proc print;
   where (i <= 5);
run;
```

The final results are displayed in Figure 56.5. They indicate that optimal values of the factors are around 0.3 for x1 and around -0.5 for x2.

Obs	x1	x2	_TYPE_	y1	y2	y3	i
1	0.3	-0.5	PREDICT	6.92570	0.75784	7.60471	1
2	0.3	-0.6	PREDICT	6.91424	0.74174	7.54194	2
3	0.3	-0.4	PREDICT	6.91003	0.77870	7.64341	3
4	0.4	-0.6	PREDICT	6.90769	0.73357	7.51836	4
5	0.4	-0.5	PREDICT	6.90540	0.75135	7.56883	5

Figure 56.5. Top Five Predictions

Handling Covariates

Covariate regressors are added to a response surface model because they are believed to account for a sizable yet relatively uninteresting portion of the variation in the data. What the experimenter is really interested in is the response corrected for the effect of the covariates. A common example is the block effect in a block design. In the canonical and ridge analyses of a response surface, which estimate responses at hypothetical levels of the factor variables, the actual value of the predicted response is computed using the average values of the covariates. The estimated response values do optimize the estimated surface of the response corrected for covariates, but true prediction of the response requires actual values for the covariates. You can use the COVAR= option in the MODEL statement to include covariates in the response surface model. Example 56.2 on page 3058 illustrates the use of this option.

Computational Method

Canonical Analysis

For each response variable, the model can be written in the form

$$y_i = \mathbf{x}_i' \mathbf{A} \mathbf{x}_i + \mathbf{b}' \mathbf{x}_i + \mathbf{c}' \mathbf{z}_i + \epsilon_i$$

where

y_i is the ith observation of the response variable.

$\mathbf{x}_i$ $= (x_{i1}, x_{i2}, \ldots, x_{ik})'$ are the k factor variables for the ith observation.

$\mathbf{z}_i$ $= (z_{i1}, z_{i2}, \ldots, z_{iL})'$ are the L covariates, including the intercept term.

$\mathbf{A}$ is the $k \times k$ symmetrized matrix of quadratic parameters, with diagonal elements equal to the coefficients of the pure quadratic terms in the model and off-diagonal elements equal to half the coefficient of the corresponding cross product.

$\mathbf{b}$ is the $k \times 1$ vector of linear parameters.

$\mathbf{c}$ is the $L \times 1$ vector of covariate parameters, one of which is the intercept.

ϵ_i is the error associated with the ith observation. Tests performed by PROC RSREG assume that errors are independently and normally distributed with mean zero and variance σ^2.

The parameters in $\mathbf{A}$, $\mathbf{b}$, and $\mathbf{c}$ are estimated by least squares. To optimize $\mathbf{y}$ with respect to $\mathbf{x}$, take partial derivatives, set them to zero, and solve:

$$\frac{\partial y}{\partial \mathbf{x}} = 2\mathbf{x}'\mathbf{A} + \mathbf{b}' = \mathbf{0} \quad \Longrightarrow \quad \mathbf{x} = -\frac{1}{2}\mathbf{A}^{-1}\mathbf{b}$$

You can determine if the solution is a maximum or minimum by looking at the eigenvalues of $\mathbf{A}$:

If the eigenvalues. . .	**then the solution is**. . .
are all negative	a maximum
are all positive	a minimum
have mixed signs	a saddle point
contain zeros	in a flat area

Ridge Analysis

The eigenvector for the largest eigenvalue gives the direction of steepest ascent from the stationary point, if positive, or steepest descent, if negative. The eigenvectors corresponding to small or zero eigenvalues point in directions of relative flatness.

The point on the optimum response ridge at a given radius R from the ridge origin is found by optimizing

$$(\mathbf{x}_0 + \mathbf{d})'\mathbf{A}(\mathbf{x}_0 + \mathbf{d}) + \mathbf{b}'(\mathbf{x}_0 + \mathbf{d})$$

over $\mathbf{d}$ satisfying $\mathbf{d}'\mathbf{d} = R^2$, where $\mathbf{x}_0$ is the $k \times 1$ vector containing the ridge origin and $\mathbf{A}$ and $\mathbf{b}$ are as previously discussed. By the method of Lagrange multipliers, the optimal $\mathbf{d}$ has the form

$$\mathbf{d} = -(\mathbf{A} - \mu\mathbf{I})^{-1}(\mathbf{A}\mathbf{x}_0 + 0.5\mathbf{b})$$

where $\mathbf{I}$ is the $k \times k$ identity matrix and μ is chosen so that $\mathbf{d}'\mathbf{d} = R^2$. There may be several values of μ that satisfy this constraint; the right one depends on which sort of response ridge is of interest. If you are searching for the ridge of maximum response, then the appropriate μ is the unique one that satisfies the constraint and is greater than all the eigenvalues of $\mathbf{A}$. Similarly, the appropriate μ for the ridge of minimum response satisfies the constraint and is less than all the eigenvalues of $\mathbf{A}$. (Refer to Myers and Montgomery (1995) for details.)

Output Data Sets

OUT=SAS-data-set

An output data set containing statistics requested with options in the MODEL statement for each observation in the input data set is created whenever the OUT= option is specified in the PROC RSREG statement. The data set contains the following variables.

- the BY variables
- the ID variables
- the WEIGHT variable
- the independent variables in the MODEL statement
- the variable _TYPE_, which identifies the observation type in the output data set. _TYPE_ is a character variable with a length of eight, and it takes on the values 'ACTUAL', 'PREDICT', 'RESIDUAL', 'U95M', 'L95M', 'U95', 'L95', and 'D', corresponding to the options specified.
- the response variables containing special output values identified by the _TYPE_ variable

All confidence limits use the two-tailed Student's t value.

OUTR=SAS-data-set

An output data set containing the optimum response ridge is created when the OUTR= option is specified in the RIDGE statement. The data set contains the following variables:

- the current values of the BY variables
- a character variable _DEPVAR_ containing the name of the dependent variable
- a character variable _TYPE_ identifying the type of ridge being computed, MINIMUM or MAXIMUM. If both MAXIMUM and MINIMUM are specified, the data set contains observations for the minimum ridge followed by observations for the maximum ridge.
- a numeric variable _RADIUS_ giving the distance from the ridge starting point
- the values of the model factors at the estimated optimum point at distance _RADIUS_ from the ridge starting point
- a numeric variable _PRED_, which is the estimated expected value of the dependent variable at the optimum
- a numeric variable _STDERR_, which is the standard error of the estimated expected value

Displayed Output

All estimates and hypothesis tests assume that the model is correctly specified and the errors are distributed according to classical statistical assumptions.

The output displayed by PROC RSREG includes the following.

Estimation and Analysis of Variance

- The actual form of the coding operation for each value of a variable is

$$\textit{coded value} = \frac{1}{S}(\textit{original value} - M)$$

  where M is the average of the highest and lowest values for the variable in the design and S is half their difference. The Subtracted off column contains the M values for this formula for each factor variable, and S is found in the Divided by column.

- The summary table for the response variable contains the following information.
  - Response Mean is the mean of the response variable in the sample. When a WEIGHT statement is used, the mean $\bar{y}$ is calculated by

$$\bar{y} = \frac{\sum_i w_i y_i}{\sum_i w_i}$$

  - Root MSE estimates the standard deviation of the response variable and is calculated as the square root of the Total Error mean square.
  - The R-Square value is R^2, or the coefficient of determination. R^2 measures the proportion of the variation in the response that is attributed to the model rather than to random error.
  - The Coefficient of Variation is 100 times the ratio of the Root MSE to the Response Mean.

- A table analyzing the significance of the terms of the regression is displayed. Terms are brought into the regression in four steps: (1) the Intercept and any covariates in the model, (2) Linear terms like X1 and X2, (3) pure Quadratic terms like X1*X1 or X2*X2, and (4) Crossproduct terms like X1*X2.
  - The Degrees of Freedom should be the same as the number of corresponding parameters unless one or more of the parameters are not estimable.
  - Type I Sum of Squares, also called the sequential sums of squares, measure the reduction in the error sum of squares as sets of terms (Linear, Quadratic, and so forth) are added to the model.
  - R-Square measures the portion of total R^2 contributed as each set of terms (Linear, Quadratic, and so forth) is added to the model.
  - Each F Value tests the null hypothesis that all parameters in the term are zero using the Total Error mean square as the denominator. This item is a test of a Type I hypothesis, containing the usual F test numerator, conditional on the effects of subsequent variables not being in the model.
  - Pr > F is the significance value or probability of obtaining at least as great an F ratio given that the null hypothesis is true.

- The Total Error Sum of Squares can be partitioned into Lack of Fit and Pure Error. When Lack of Fit is significant, there is variation around the model other than random error (such as cubic effects of the factor variables).
  - The Total Error Mean Square estimates σ^2, the variance.
  - F Value tests the null hypothesis that the variation is adequately described by random error.
- A table containing the parameter estimates from the model is displayed.
  - The Parameter Estimate column contains the parameter estimates based on the *uncoded* values of the factor variables. If an effect is a linear combination of previous effects, the parameter for the effect is not estimable. When this happens, the degrees of freedom are zero, the parameter estimate is set to zero, and the estimates and tests on other parameters are conditional on this parameter being zero.
  - The Standard Error column contains the estimated standard deviations of the parameter estimates based on *uncoded* data.
  - The t Value column contains t values of a test of the null hypothesis that the true parameter is zero when the *uncoded* values of the factor variables are used.
  - Pr > |T| gives the significance value or probability of a greater absolute t ratio given that the true parameter is zero.
  - The Parameter Estimate from Coded Data column contains the parameter estimates based on the *coded* values of the factor variables. These are the estimates used in the subsequent canonical and ridge analyses.
- The sum of squares are partitioned by the Factors in the model, and an analysis table is displayed. The test on a factor, say X1, is a joint test on all the parameters involving that factor. For example, the test for X1 tests the null hypothesis that the true parameters for X1, X1*X1, and X1*X2 are all zero.

Canonical Analysis

- The Critical Value columns contains the values of the factor variables that correspond to the stationary point of the fitted response surface. The critical values can be at a minimum, maximum, or saddle point.
- The Eigenvalues and Eigenvectors are from the matrix of quadratic parameter estimates based on the coded data. They characterize the shape of the response surface.

Ridge Analysis

- Coded Radius is the distance from the coded version of the associated point to the coded version of the origin of the ridge. The origin is given by the point at radius zero.
- Estimated Response is the estimated value of the response variable at the associated point. The Standard Error of this estimate is also given. This quantity is useful for assessing the relative credibility of the prediction at a given radius. Typically, this standard error increases rapidly as the ridge moves up to

and beyond the design perimeter, reflecting the inherent difficulty of making predictions beyond the range of experimentation.

- Uncoded Factor Values are the values of the uncoded factor variables that give the optimum response at this radius from the ridge origin.

ODS Table Names

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

Table 56.1. ODS Tables Produced in PROC RSREG

ODS Table Name	Description	Statement
Coding	Coding coefficients for the independent variables	default
ErrorANOVA	Error analysis of variance	default
FactorANOVA	Factor analysis of variance	default
FitStatistics	Overall statistics for fit	default
ModelANOVA	Model analysis of variance	default
ParameterEstimates	Estimated linear parameters	default
Ridge	Ridge analysis for optimum response	RIDGE
Spectral	Spectral analysis	default
StationaryPoint	Stationary point of response surface	default

Examples

Example 56.1. A Saddle-Surface Response Using Ridge Analysis

Frankel (1961) reports an experiment aimed at maximizing the yield of *mercaptobenzothiazole* (MBT) by varying processing time and temperature. Myers (1976) uses a two-factor model in which the estimated surface does not have a unique optimum. A ridge analysis is used to determine the region in which the optimum lies. The objective is to find the settings of time and temperature in the processing of a chemical that maximize the yield. The following statements read the data and invoke PROC RSREG. These statements produce Output 56.1.1 through Output 56.1.5:

```
data d;
   input Time Temp MBT;
   label Time = "Reaction Time (Hours)"
         Temp = "Temperature (Degrees Centigrade)"
         MBT  = "Percent Yield Mercaptobenzothiazole";
   datalines;
 4.0   250   83.8
20.0   250   81.7
12.0   250   82.4
```

```
12.0   250   82.9
12.0   220   84.7
12.0   280   57.9
12.0   250   81.2
 6.3   229   81.3
 6.3   271   83.1
17.7   229   85.3
17.7   271   72.7
 4.0   250   82.0
;
proc sort;
   by Time Temp;
run;

proc rsreg;
   model MBT=Time Temp / lackfit;
   ridge max;
run;
```

Output 56.1.1. Coding and Response Variable Information

```
                          The RSREG Procedure

            Coding Coefficients for the Independent Variables

                Factor    Subtracted off      Divided by

                Time           12.000000        8.000000
                Temp          250.000000       30.000000


Response Surface for Variable MBT: Percent Yield Mercaptobenzothiazole

                Response Mean                 79.916667
                Root MSE                       4.615964
                R-Square                       0.8003
                Coefficient of Variation       5.7760
```

Output 56.1.2. Analyses of Variance

The RSREG Procedure

Regression	DF	Type I Sum of Squares	R-Square	F Value	Pr > F
Linear	2	313.585803	0.4899	7.36	0.0243
Quadratic	2	146.768144	0.2293	3.44	0.1009
Crossproduct	1	51.840000	0.0810	2.43	0.1698
Total Model	5	512.193947	0.8003	4.81	0.0410

Residual	DF	Sum of Squares	Mean Square	F Value	Pr > F
Lack of Fit	3	124.696053	41.565351	39.63	0.0065
Pure Error	3	3.146667	1.048889		
Total Error	6	127.842720	21.307120		

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Parameter Estimate from Coded Data
Intercept	1	-545.867976	277.145373	-1.97	0.0964	82.173110
Time	1	6.872863	5.004928	1.37	0.2188	-1.014287
Temp	1	4.989743	2.165839	2.30	0.0608	-8.676768
Time*Time	1	0.021631	0.056784	0.38	0.7164	1.384394
Temp*Time	1	-0.030075	0.019281	-1.56	0.1698	-7.218045
Temp*Temp	1	-0.009836	0.004304	-2.29	0.0623	-8.852519

Factor	DF	Sum of Squares	Mean Square	F Value	Pr > F	Label
Time	3	61.290957	20.430319	0.96	0.4704	Reaction Time (Hours)
Temp	3	461.250925	153.750308	7.22	0.0205	Temperature (Degrees Centigrade)

Output 56.1.2 shows that the lack of fit for the model is highly significant. Since the quadratic model does not fit the data very well, firm statements about the underlying process should not be based only on the current analysis. Note from the analysis of variance for the model that the test for the time factor is not significant. If further experimentation is undertaken, it might be best to fix Time at a moderate to high value and to concentrate on the effect of temperature. In the actual experiment discussed here, extra runs were made that confirmed the results of the following analysis.

Output 56.1.3. Canonical Analysis

The RSREG Procedure
Canonical Analysis of Response Surface Based on Coded Data

Factor	Critical Value Coded	Uncoded	Label
Time	-0.441758	8.465935	Reaction Time (Hours)
Temp	-0.309976	240.700718	Temperature (Degrees Centigrade)

Predicted value at stationary point: 83.741940

Eigenvalues	Eigenvectors Time	Temp
2.528816	0.953223	-0.302267
-9.996940	0.302267	0.953223

Stationary point is a saddle point.

The canonical analysis (Output 56.1.3) indicates that the predicted response surface is shaped like a saddle. The eigenvalue of 2.5 shows that the valley orientation of the saddle is less curved than the hill orientation, with eigenvalue of -9.99. The coefficients of the associated eigenvectors show that the valley is more aligned with Time and the hill with Temp. Because the canonical analysis resulted in a saddle point, the estimated surface does not have a unique optimum.

Output 56.1.4. Ridge Analysis

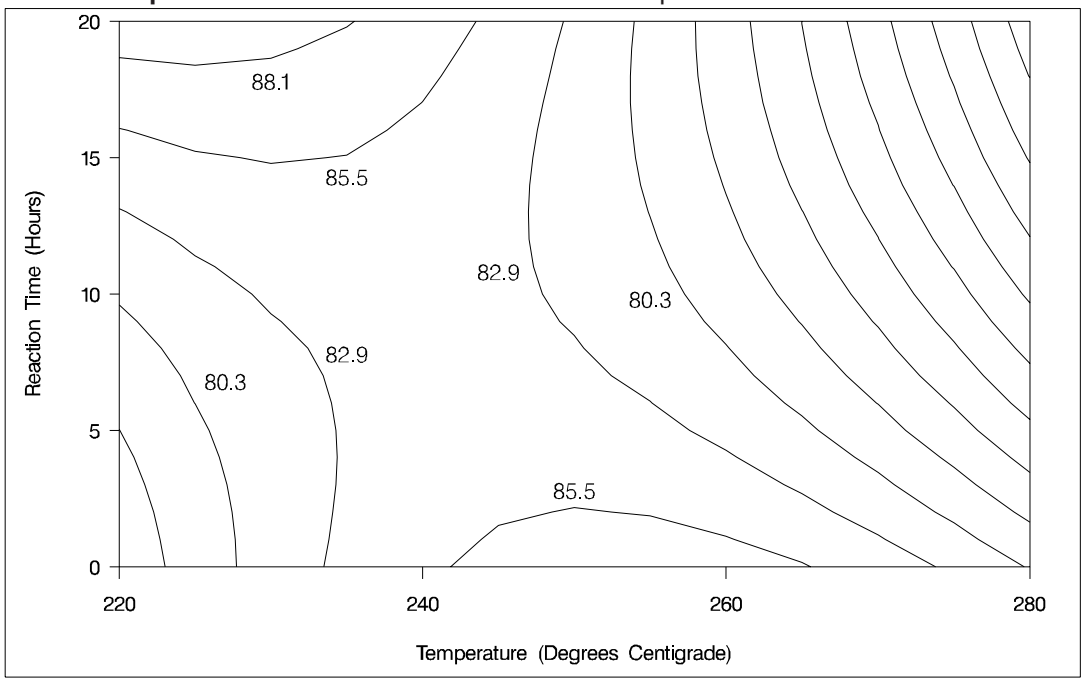
The RSREG Procedure

Estimated Ridge of Maximum Response for Variable MBT: Percent Yield Mercaptobenzothiazole

Coded Radius	Estimated Response	Standard Error	Uncoded Factor Values Time	Temp
0.0	82.173110	2.665023	12.000000	250.000000
0.1	82.952909	2.648671	11.964493	247.002956
0.2	83.558260	2.602270	12.142790	244.023941
0.3	84.037098	2.533296	12.704153	241.396084
0.4	84.470454	2.457836	13.517555	239.435227
0.5	84.914099	2.404616	14.370977	237.919138
0.6	85.390012	2.410981	15.212247	236.624811
0.7	85.906767	2.516619	16.037822	235.449230
0.8	86.468277	2.752355	16.850813	234.344204
0.9	87.076587	3.130961	17.654321	233.284652
1.0	87.732874	3.648568	18.450682	232.256238

However, the ridge analysis in Output 56.1.4 indicates that maximum yields will result from relatively high reaction times and low temperatures. A contour plot of the predicted response surface, shown in Output 56.1.5, confirms this conclusion.

Output 56.1.5. Contour Plot of Predicted Response Surface



The statements that produce this plot follow. Note that contour and three-dimensional plots can be created interactively using SAS/INSIGHT software or the ADX Interface

in SAS/QC software. Initial DATA steps create a grid over Time and Temp and combine this grid with the original data, using a variable flag to indicate the grid. Then, PROC RSREG is used to create predictions for the combined data. Finally, a series of statements subsets the predictions over just the grid and uses PROC GCONTOUR to display a contour plot. An ANNOTATE data set adds level values to the contours.

```
data b;
   set d;
   flag=1;
   MBT=.;
   do Time=0 to 20 by 1;
      do Temp=220 to 280 by 5;
         output;
      end;
   end;
data c;
   set d b;
run;

proc rsreg data=c out=e noprint;
  model MBT=Time Temp / predict;
  id flag;
run;

data f;
   set e;
   if flag=1;
data annote;
   length function color style $8 text $8;
   retain hsys ysys xsys '2' size 1 function 'label'
          color 'black' style 'swissl' position '5';
   x=255; y=10  ; text='80.3'; output;
   x=245; y=11  ; text='82.9'; output;
   x=227; y= 7  ; text='80.3'; output;
   x=235; y= 8  ; text='82.9'; output;
   x=235; y=14.5; text='85.5'; output;
   x=230; y=18  ; text='88.1'; output;
   x=250; y= 3  ; text='85.5'; output;
run;
axis1 label=(angle=90) minor=none;
axis2 order=(220 to 280 by 20) minor=none;

proc gcontour data=f annotate=annote;
   plot Time*Temp=MBT
        / nlevels=12 vaxis=axis1 haxis=axis2 nolegend
          llevels=2 2 2   1 1 1   1 1 1   1 1 1   ;
run;
```

Example 56.2. Response Surface Analysis with Covariates

One way of viewing covariates is as extra sources of variation in the dependent variable that may mask the variation due to primary factors. This example demonstrates the use of the COVAR= option in PROC RSREG to fit a response surface model to the dependent variable values corrected for the covariates.

You have a chemical process with a yield that you hypothesize to be dependent on three factors: reaction time, reaction temperature, and reaction pressure. You perform an experiment to measure this dependence. You are willing to include up to 20 runs in your experiment, but you can perform no more than 8 runs on the same day, so the design for the experiment is composed of three blocks. Additionally, you know that the grade of raw material for the reaction has a significant impact on the yield. You have no control over this, but you keep track of it. The following statements create a SAS data set containing the results of the experiment:

```
data Experiment;
   input Day Grade Time Temp Pressure Yield;
   datalines;
1 67      -1     -1     -1         32.98
1 68      -1      1      1         47.04
1 70       1     -1      1         67.11
1 66       1      1     -1         26.94
1 74       0      0      0        103.22
1 68       0      0      0         42.94
2 75      -1     -1      1        122.93
2 69      -1      1     -1         62.97
2 70       1     -1     -1         72.96
2 71       1      1      1         94.93
2 72       0      0      0         93.11
2 74       0      0      0        112.97
3 69       1.633  0      0         78.88
3 67      -1.633  0      0         52.53
3 68       0      1.633  0         68.96
3 71       0     -1.633  0         92.56
3 70       0      0      1.633     88.99
3 72       0      0     -1.633    102.50
3 70       0      0      0         82.84
3 72       0      0      0        103.12
;
```

Your first analysis neglects to take the covariates into account. The following statements use PROC RSREG to fit a response surface to the observed yield, but note that Day and Grade are omitted.

```
proc rsreg data=Experiment;
   model Yield = Time Temp Pressure;
run;
```

The ANOVA results (shown in Output 56.2.1) indicate that *no* process variable effects are significantly larger than the background noise.

Output 56.2.1. Analysis of Variance Ignoring Covariates

The RSREG Procedure

Regression	DF	Type I Sum of Squares	R-Square	F Value	Pr > F
Linear	3	1880.842426	0.1353	0.67	0.5915
Quadratic	3	2370.438681	0.1706	0.84	0.5023
Crossproduct	3	241.873250	0.0174	0.09	0.9663
Total Model	9	4493.154356	0.3233	0.53	0.8226

Residual	DF	Sum of Squares	Mean Square
Total Error	10	9405.129724	940.512972

However, when the yields are adjusted for covariate effects of day and grade of raw material, very strong process variable effects are revealed. The following statements produce the ANOVA results in Output 56.2.2. Note that in order to include the effects of the classification factor Day as covariates, you need to create dummy variables indicating each day separately.

```
data Experiment; set Experiment;
   d1 = (Day = 1);
   d2 = (Day = 2);
   d3 = (Day = 3);
proc rsreg data=Experiment;
   model Yield = d1-d3 Grade Time Temp Pressure / covar=4;
run;
```

Output 56.2.2. Analysis of Variance Including Covariates

The RSREG Procedure

Regression	DF	Type I Sum of Squares	R-Square	F Value	Pr > F
Covariates	3	13695	0.9854	316957	<.0001
Linear	3	156.524497	0.0113	3622.53	<.0001
Quadratic	3	22.989775	0.0017	532.06	<.0001
Crossproduct	3	23.403614	0.0017	541.64	<.0001
Total Model	12	13898	1.0000	80413.2	<.0001

Residual	DF	Sum of Squares	Mean Square
Total Error	7	0.100820	0.014403

The results show very strong effects due to both the covariates and the process variables.

References

Box, G.E.P. (1954), "The Exploration and Exploitation of Response Surfaces: some General Considerations," *Biometrics*, 10, 16.

Box, G.E.P. (1987), *Empirical Model Building and Response Surfaces*, New York: John Wiley & Sons, Inc.

Box, G.E.P. and Draper, N.R. (1982), "Measures of Lack of Fit for Response Surface Designs and Predictor Variable Transformations," *Technometrics*, 24, 1–8.

Box, G.E.P. and Hunter, J.S. (1957), "Multifactor Experimental Designs for Exploring Response Surfaces," *Annals of Mathematical Statistics*, 28, 195–242.

Box, G.E.P., Hunter, W.G., and Hunter, J.S. (1978), *Statistics for Experimenters*, New York: John Wiley & Sons, Inc.

Box, G.E.P. and Wilson, K.J. (1951), "On the Experimental Attainment of Optimum Conditions," *Journal of the Royal Statistical Society*, Ser. B, 13, 1–45.

Cochran, W.G. and Cox, G.M. (1957), *Experimental Designs*, Second Edition, New York: John Wiley & Sons, Inc.

Draper, N.R. (1963), "Ridge Analysis of Response Surfaces," *Technometrics* 5, 469–479.

Draper, N.R. and John, J.A. (1988), "Response Surface Designs for Quantitative and Qualitative Variables," *Technometrics*, 30 (4), 423–428.

Draper, N.R. and Smith, H. (1981), *Applied Regression Analysis, Second Edition*, New York: John Wiley & Sons, Inc.

John, P.W.M. (1971), *Statistical Design and Analysis of Experiments*, New York: Macmillan Publishing Co., Inc.

Mead, R. and Pike, D.J. (1975), "A review of Response Surface Methodology from a Biometric Point of View," *Biometrics*, 31, 803.

Meyer, D.C. (1963), "Response Surface Methodology in Education and Psychology," *Journal of Experimental Education* 31, 329.

Myers, R.H. (1976), *Response Surface Methodology*, Blacksburg, VA: Virginia Polytechnic Institute and State University.

Myers, R.H. and Montgomery, D.C. (1995), *Response Surface Methodology*, New York: John Wiley & Sons, Inc.

Schneider, A.M. and Stockett, A.L. (1963), "An Experiment to Select Optimum Operating Conditions on the Basis of Arbitrary Preference Ratings," *Chemical Engineering Progress Symposium Series* 59.

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

SAS/STAT® User's Guide, Version 8

Copyright © 1999 by SAS Institute Inc., Cary, NC, USA.

ISBN 1–58025–494–2

All rights reserved. Produced in the United States of America. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

U.S. Government Restricted Rights Notice. Use, duplication, or disclosure of the software and related documentation by the U.S. government is subject to the Agreement with SAS Institute and the restrictions set forth in FAR 52.227–19 Commercial Computer Software-Restricted Rights (June 1987).

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

1st printing, October 1999

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

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

The Institute is a private company devoted to the support and further development of its software and related services.