SAS/STAT® 9.2 User's Guide
Introduction to Regression Procedures
(Book Excerpt)

SAS® Documentation

This document is an individual chapter from *SAS/STAT® 9.2 User's Guide*.

The correct bibliographic citation for the complete manual is as follows: SAS Institute Inc. 2008. *SAS/STAT® 9.2 User's Guide*. Cary, NC: SAS Institute Inc.

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

1st electronic book, March 2008
2nd electronic book, February 2009

Chapter 4

Introduction to Regression Procedures

Contents

Overview: Regression Procedures **76**
Introduction 76
Introductory Example: Linear Regression 79
Linear Regression: The REG Procedure 85
Response Surface Regression: The RSREG Procedure 88
Partial Least Squares Regression: The PLS Procedure 88
Generalized Linear Regression 88
Logistic Regression 89
Other Generalized Linear Models 90
Regression for Ill-Conditioned Data: The ORTHOREG Procedure 90
Quantile Regression: The QUANTREG Procedure 91
Nonlinear Regression 91
Nonparametric Regression 92
Local Regression: The LOESS Procedure 93
Smooth Function Approximation: The TPSPLINE Procedure 93
Generalized Additive Models: The GAM Procedure 93
Robust Regression: The ROBUSTREG Procedure 93
Regression with Transformations: The TRANSREG Procedure 94
Interactive Features in the CATMOD, GLM, and REG Procedures 94
Statistical Background in Linear Regression **95**
Linear Regression Models 95
Parameter Estimates and Associated Statistics 96
Predicted and Residual Values 100
Testing Linear Hypotheses 102
Multivariate Tests 102
Comments on Interpreting Regression Statistics 107
References **111**

Overview: Regression Procedures

This chapter provides an overview of procedures in SAS/STAT software that perform regression analysis. The REG procedure provides the most general analysis capabilities for the linear regression model. But many other procedures can fit linear regression models, and many procedures are specifically designed for more general regression problems, such as robust regression, generalized linear regression, nonlinear regression, nonparametric regression, regression modeling of survey data, regression modeling of survival data, and regression modeling of transformed variables. The aim of this chapter is to provide a brief road map and delineation of the various SAS/STAT procedures that can fit regression models.

A sampling of the procedures that fall into this category are the CATMOD, GAM, GENMOD, GLIMMIX, GLM, LIFEREG, LOESS, LOGISTIC, MIXED, NLIN, NLMIXED, ORTHOREG, PHREG, PLS, PROBIT, REG, ROBUSTREG, RSREG, SURVEYLOGISTIC, SURVEYREG, and TRANSREG procedures.

This chapter also briefly mentions several procedures in SAS/ETS software.

Introduction

Recall from Chapter 3, "Introduction to Statistical Modeling with SAS/STAT Software," that the general regression problem is to model the mean of a random vector **Y** as a function of a parameters and covariates in a statistical model. The many forms of regression models have their origin in the characteristics of the response variable (discrete or continuous, normal or nonnormal distributed), assumptions about the form of the model (linear, nonlinear, or generalized linear), assumptions about the data generating mechanism (survey, observational, or experimental data), and estimation principles. The following procedures, listed in alphabetical order, perform at least one type of regression analysis.

CATMOD — analyzes data that can be represented by a contingency table. PROC CATMOD fits linear models to functions of response frequencies, and it can be used for linear and logistic regression. See Chapter 8, "Introduction to Categorical Data Analysis Procedures," and Chapter 28, "The CATMOD Procedure," for more information.

GAM — fits generalized additive models. The models fitted with the GAM procedure are nonparametric in that the usual assumption of a linear predictor is relaxed. The name stems from the fact that the models consist of additive, smooth functions in the regression variables. The GAM procedure can fit additive models to nonnormal data. See Chapter 36, "The GAM Procedure," for more information.

GENMOD — fits generalized linear models. PROC GENMOD is especially suited for responses with discrete outcomes, and it performs logistic regression and Poisson regression as well as fitting generalized estimating equations for repeated measures data. Bayesian analysis capabilities for generalized linear models are also available with the GENMOD procedure. See Chapter 8, "Introduction to Categorical Data Analysis Procedures," and Chapter 37, "The GENMOD Procedure," for more information.

GLIMMIX — fits generalized linear mixed models by likelihood-based methods. In addition to many other analyses, PROC GLIMMIX can perform simple, multiple, polynomial, and weighted regression. The GLIMMIX procedure can also fit linear mixed models and models without random effects. See Chapter 38, "The GLIMMIX Procedure," for more information.

GLM — uses the method of least squares to fit general linear models. In addition to many other analyses, PROC GLM can perform simple, multiple, polynomial, and weighted regression. PROC GLM has many of the same input/output capabilities as PROC REG, but it does not provide as many diagnostic tools or allow interactive changes in the model or data. See Chapter 5, "Introduction to Analysis of Variance Procedures," and Chapter 39, "The GLM Procedure," for more information.

LIFEREG — fits parametric models to failure-time data that might be right censored. These types of models are commonly used in survival analysis. See Chapter 14, "Introduction to Survey Procedures," and Chapter 48, "The LIFEREG Procedure," for more information.

LOESS — fits nonparametric models using a local regression method. PROC LOESS is suitable for modeling regression surfaces where the underlying parametric form is unknown and where robustness in the presence of outliers is required. See Chapter 50, "The LOESS Procedure," for more information.

LOGISTIC — fits logistic models for binomial and ordinal outcomes. PROC LOGISTIC provides a wide variety of model-building methods and computes numerous regression diagnostics. See Chapter 8, "Introduction to Categorical Data Analysis Procedures," and Chapter 51, "The LOGISTIC Procedure," for more information.

MIXED — fits linear mixed models by likelihood-based techniques. In addition to many other analyses, PROC MIXED can fit models without random effects; hence, the procedure can perform simple, multiple, polynomial, and weighted regression. See Chapter 56, "The MIXED Procedure," for more information.

NLIN — fits general nonlinear regression models by the method of nonlinear least squares. Several different iterative methods are available. See Chapter 60, "The NLIN Procedure," for more information.

NLMIXED — fits general nonlinear mixed regression models by the method of maximum likelihood. With the NLMIXED procedure you can specify a custom objective function for parameter estimation and fit models with or without random effects. See Chapter 61, "The NLMIXED Procedure," for more information.

ORTHOREG — performs regression by using the Gentleman-Givens computational method. For ill-conditioned data, PROC ORTHOREG can produce more accurate parameter estimates than other procedures such as PROC GLM and PROC REG. See Chapter 63, "The ORTHOREG Procedure," for more information.

PHREG — fits Cox proportional hazards regression models to survival data. See Chapter 64, "The PHREG Procedure," for more information.

PLS — performs partial least squares regression, principal components regression, and reduced rank regression, with cross validation for the number of components. See Chapter 66, "The PLS Procedure," for more information.

PROBIT — performs probit regression as well as logistic regression and ordinal logistic regression. The PROBIT procedure is useful when the dependent variable is either dichotomous or polychotomous and the independent variables are continuous. See Chapter 71, "The PROBIT Procedure," for more information.

QUANTREG — models the effects of covariates on the conditional quantiles of a response variable by means of quantile regression. See Chapter 72, "The QUANTREG Procedure," for more information.

REG — performs linear regression with many diagnostic capabilities, selects models by using one of nine methods, produces scatter plots of raw data and statistics, highlights scatter plots to identify particular observations, and allows interactive changes in both the regression model and the data used to fit the model. See Chapter 73, "The REG Procedure," for more information.

ROBUSTREG — performs robust regression by using Huber M estimation and high breakdown value estimation. PROC ROBUSTREG is suitable for detecting outliers and providing resistant (stable) results in the presence of outliers. See Chapter 74, "The ROBUSTREG Procedure," for more information.

RSREG — builds quadratic response-surface regression models. PROC RSREG analyzes the fitted response surface to determine the factor levels of optimum response and performs a ridge analysis to search for the region of optimum response. See Chapter 75, "The RSREG Procedure," for more information.

SURVEYLOGISTIC — fits logistic models for binary and ordinal outcomes to survey data by maximum likelihood. See Chapter 84, "The SURVEYLOGISTIC Procedure," for more information.

SURVEYREG — fits linear regression models to survey data by generalized least squares by using elementwise regresssion. See Chapter 86, "The SURVEYREG Procedure," for more information.

TRANSREG — fits univariate and multivariate linear models, optionally with spline and other nonlinear transformations. Models include ordinary regression and ANOVA, multiple and multivariate regression, metric and nonmetric conjoint analysis, metric and nonmetric vector and ideal point preference mapping, redundancy analysis, canonical correlation, and response surface regression. See Chapter 90, "The TRANSREG Procedure," for more information.

Several SAS/ETS procedures also perform regression. The following procedures are documented in the *SAS/ETS User's Guide*.

AUTOREG — implements regression models that use time-series data where the errors are autocorrelated. See Chapter 8, "The AUTOREG Procedure" (*SAS/ETS User's Guide*), for more details.

COUNTREG — analyzes regression models in which the dependent variable takes nonnegative integer or count values. See Chapter 10, "The COUNTREG Procedure" (*SAS/ETS User's Guide*), for more details.

MODEL	handles nonlinear simultaneous systems of equations, such as econometric models. See Chapter 18, "The MODEL Procedure" (*SAS/ETS User's Guide*), for more details.
PANEL	analyzes a class of linear econometric models that commonly arise when time series and cross-sectional data are combined. See Chapter 19, "The PANEL Procedure" (*SAS/ETS User's Guide*), for more details.
PDLREG	performs regression analysis with polynomial distributed lags. See Chapter 20, "The PDLREG Procedure" (*SAS/ETS User's Guide*), for more details.
SYSLIN	handles linear simultaneous systems of equations, such as econometric models. See Chapter 26, "The SYSLIN Procedure" (*SAS/ETS User's Guide*), for more details.

Introductory Example: Linear Regression

Regression analysis is the analysis of the relationship between a response or outcome variable and another set of variables. The relationship is expressed through a statistical model equation that predicts a *response variable* (also called a *dependent variable* or *criterion*) from a function of *regressor variables* (also called *independent variables, predictors, explanatory variables, factors,* or *carriers*) and *parameters*. In a linear regression model the predictor function is linear in the parameters (but not necessarily linear in the regressor variables). The parameters are estimated so that a measure of fit is optimized. For example, the equation for the ith observation might be

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where Y_i is the response variable, x_i is a regressor variable, β_0 and β_1 are unknown parameters to be estimated, and ϵ_i is an error term. This model is termed the simple linear regression (SLR) model, because it is linear in β_0 and β_1 and contains only a single regressor variable.

Suppose you are using regression analysis to relate a child's weight to a child's height. One application of a regression model with the response variable Weight is to predict a child's weight for a known height. Suppose you collect data by measuring heights and weights of 19 randomly selected schoolchildren. A simple linear regression model with the response variable weight and the regressor variable height can be written as the following equation:

$$\text{Wght}_i = \beta_0 + \beta_1 \,\text{Height}_i + \epsilon_i$$

where

Wght_i	is the response variable for the ith child.
Height_i	is the regressor variable for the ith child.
β_0, β_1	are the unknown regression parameters.
ϵ_i	is the unobservable random error associated with the ith observation.

The following DATA step creates a SAS data set identifying the children and their observed heights (variable Height) and weights (variable Wght).

```
data class;
   input Name $ Height Wght Age;
   datalines;
Alfred  69.0 112.5 14
Alice   56.5  84.0 13
Barbara 65.3  98.0 13
Carol   62.8 102.5 14
Henry   63.5 102.5 14
James   57.3  83.0 12
Jane    59.8  84.5 12
Janet   62.5 112.5 15
Jeffrey 62.5  84.0 13
John    59.0  99.5 12
Joyce   51.3  50.5 11
Judy    64.3  90.0 14
Louise  56.3  77.0 12
Mary    66.5 112.0 15
Philip  72.0 150.0 16
Robert  64.8 128.0 12
Ronald  67.0 133.0 15
Thomas  57.5  85.0 11
William 66.5 112.0 15
;

ods graphics on;
proc reg data=class;
   model Wght = Height;
run;
```

Figure 4.1 displays the default tabular output of the REG procedure for this model. Nineteen observations are read from the data set and all observations are used in the analysis. The estimates of the two regression parameters are $\widehat{\beta}_0 = -143.02692$ and $\widehat{\beta_1} = 3.89903$. These estimates are obtained by the least squares principle. See the sections "Classical Estimation Principles" and "Linear Model Theory" in Chapter 3, "Introduction to Statistical Modeling with SAS/STAT Software," for details about the principle of least squares estimation and its role in linear model analysis. For a general discussion of the theory of least squares estimation of linear models and its application to regression and analysis of variance, refer to one of the applied regression texts, including Draper and Smith (1981), Daniel and Wood (1980), Johnston (1972), and Weisberg (1985).

Figure 4.1 Regression for Weight and Height Data

```
                    The REG Procedure
                      Model: MODEL1
                Dependent Variable: Wght

        Number of Observations Read          19
        Number of Observations Used          19
```

Figure 4.1 *continued*

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7193.24912	7193.24912	57.08	<.0001
Error	17	2142.48772	126.02869		
Corrected Total	18	9335.73684			

Root MSE	11.22625	R-Square	0.7705
Dependent Mean	100.02632	Adj R-Sq	0.7570
Coeff Var	11.22330		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-143.02692	32.27459	-4.43	0.0004
Height	1	3.89903	0.51609	7.55	<.0001

Based on the least squares estimates shown in Figure 4.1, the fitted regression line relating height to weight is described by the equation

$$\widehat{\text{Wght}} = -143.02692 + 3.89903 \times \text{Height}$$

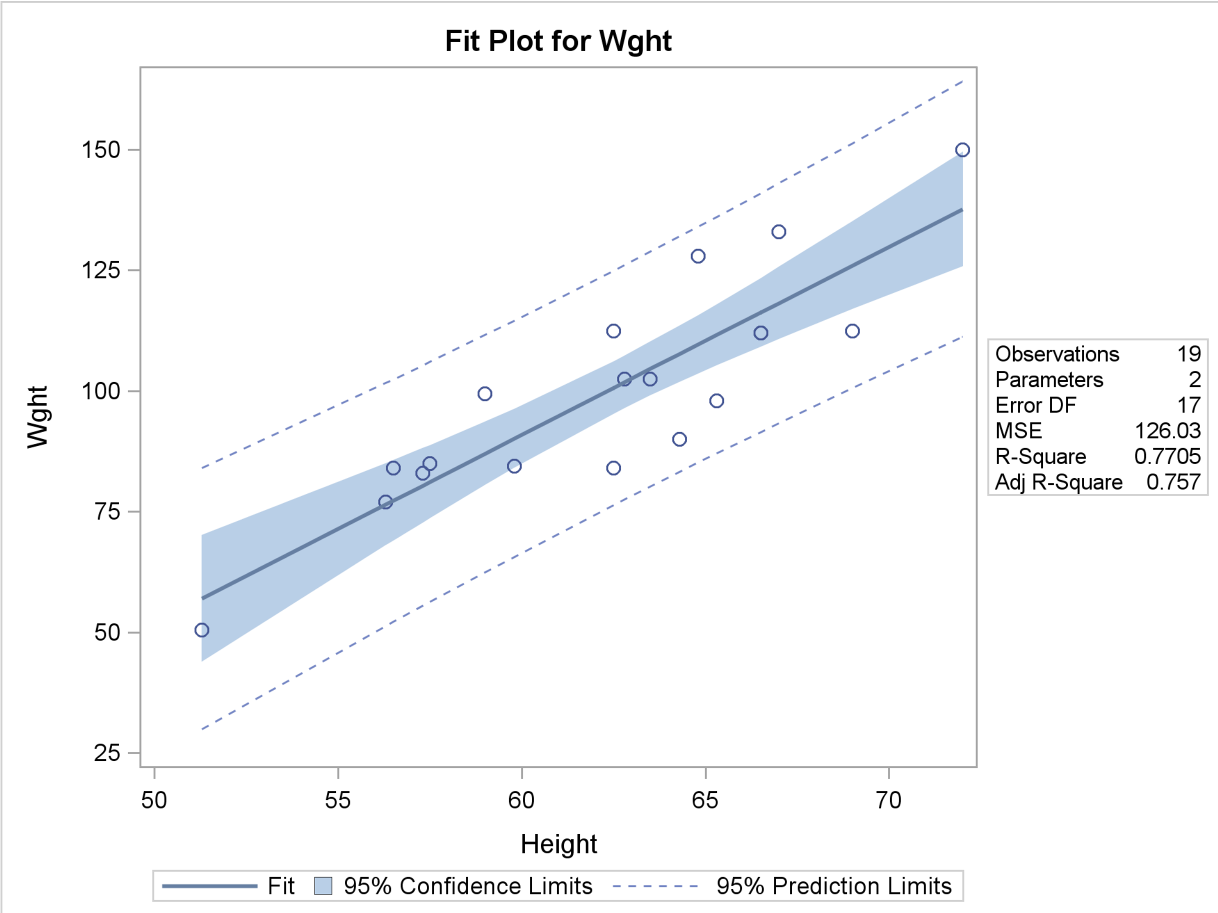
The "hat" notation is used to denote the predicted variable on the left side of the prediction equation to emphasize that $\widehat{\text{Wght}}$ is not one of the original observations but a value predicted under the regression model that has been fit to the data. At the least squares solution the residual sum of squares

$$\text{SSE} = \sum_{i=1}^{19} (\text{Wght}_i - \beta_0 - \beta_1 \text{Height}_i)^2$$

is minimized and the achieved criterion value is displayed in the analysis of variance table as the error sum of squares (2142.48772).

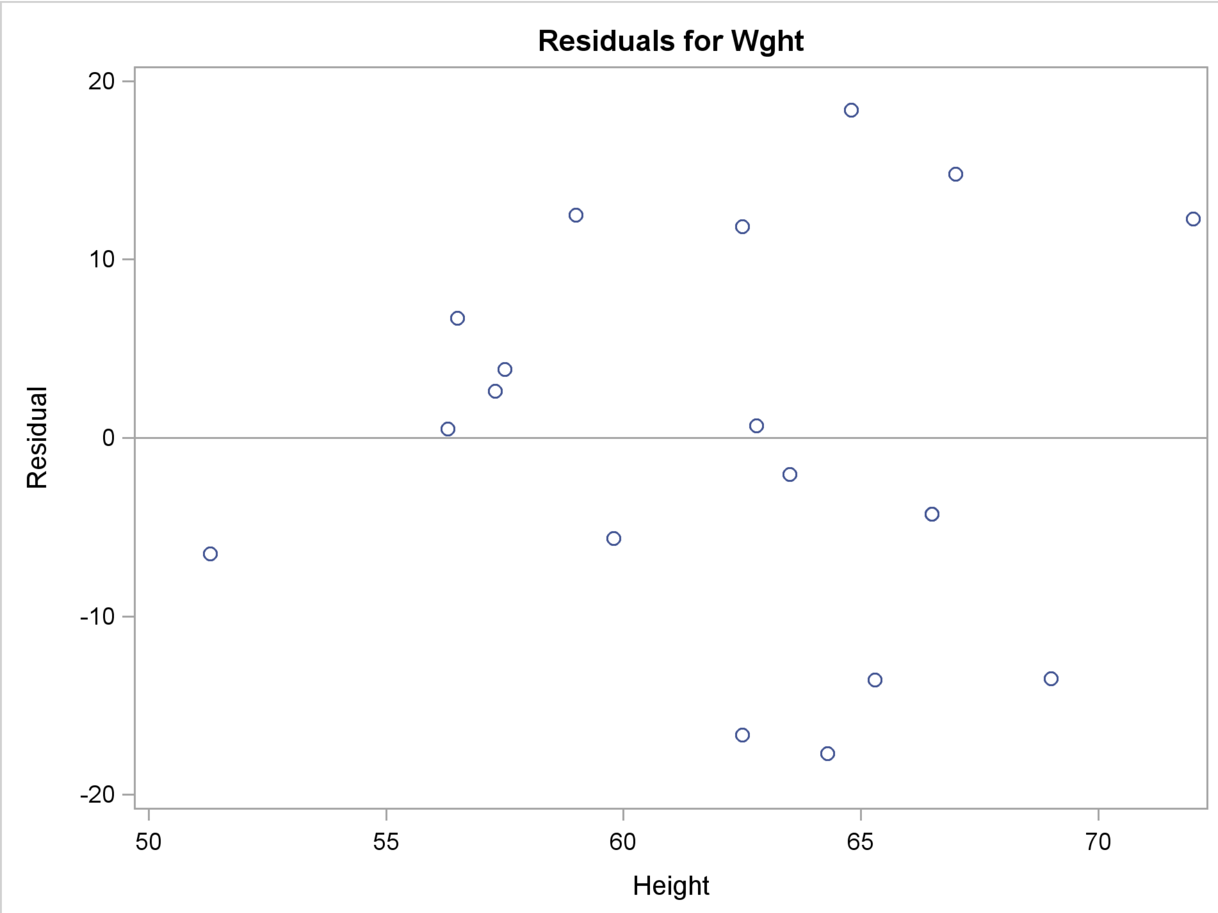
Figure 4.2 displays the fit plot induced by the **ODS GRAPHICS ON;** statement. The fit plot shows the positive slope of the fitted line. The average weight of a child changes by $\widehat{\beta}_1 = 3.89903$ units for each unit change in height. The 95% confidence limits in the fit plot are pointwise limits that cover the mean weight for a particular height with probability 0.95. The prediction limits, which are wider than the confidence limits, show the pointwise limits that cover a new observation for a given height with probability 0.95.

Figure 4.2 Fit Plot for Regression of Weight on Height



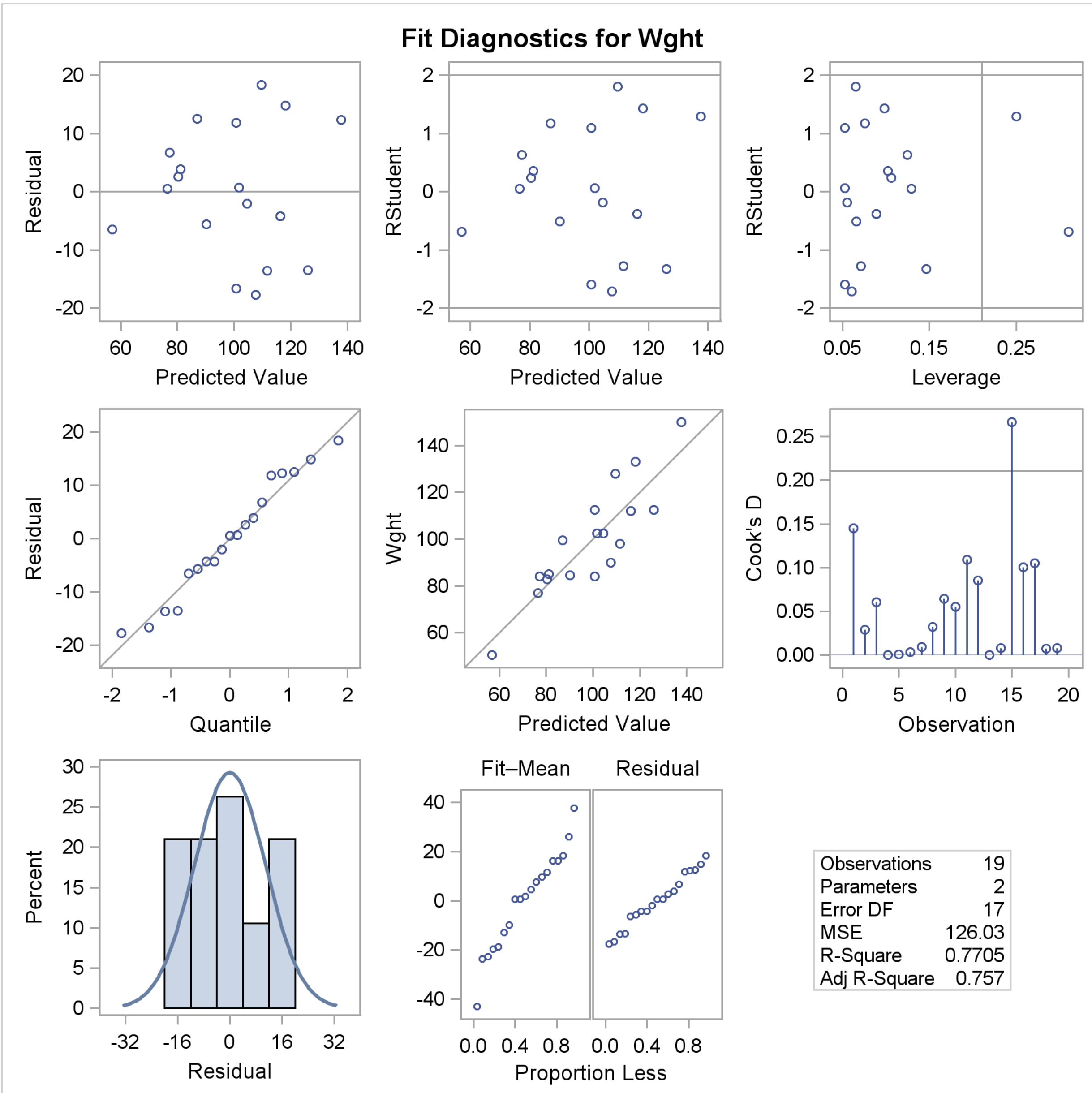
Regression is often used in an exploratory fashion to look for empirical relationships, such as the relationship between Height and Wght. In this example, Height is not the cause of Wght. You would need a controlled experiment to confirm the relationship scientifically. See the section "Comments on Interpreting Regression Statistics" on page 107 for more information. A separate question from a possible cause-and-effect relationship between the two variables involved in this regression is whether the simple linear regression model adequately describes the relationship in these data. If the usual assumptions about the model errors ϵ_i are met in the SLR model, then the errors should have zero mean and equal variance and be uncorrelated. Because the children were randomly selected, the observations from different children are not correlated. If the mean function of the model is correctly specified, the fitted residuals $\text{Wght}_i - \widehat{\text{Wght}}_i$ should scatter about the zero reference line without discernible structure. The residual plot in Figure 4.3 confirms this behavior.

Figure 4.3 Residual Plot for Regression of Weight on Height



An even more detailed look at the model-data agreement is gained with the panel of regression diagnostics in Figure 4.4. The graph in the upper-left panel repeats the raw residual plot in Figure 4.3. The plot of the RSTUDENT residuals shows externally studentized residuals that take into account heterogeneity in the variability of the residuals. RSTUDENT residuals that exceed the threshold values of ± 2 often indicate outlying observations. The residual-by-leverage plot shows that two observations have high leverage—that is, they are unusual in their height values relative to the other children. The normal-probability Q-Q plot in the second row of the panel shows that the normality assumption for the residuals is reasonable. The plot of the Cook's D statistic shows that observation 15 exceeds the threshold value; this indicates that the observation for this child is influential on the regression parameter estimates.

Figure 4.4 Panel of Regression Diagnostics



For detailed information about the interpretation of regression diagnostics and about ODS statistical graphics with the REG procedure, see Chapter 73, “The REG Procedure.”

SAS/STAT regression procedures produce the following information for a typical regression analysis:

- parameter estimates derived by using the least squares criterion
- estimates of the variance of the error term
- estimates of the variance or standard deviation of the sampling distribution of the parameter estimates
- tests of hypotheses about the parameters

SAS/STAT regression procedures can produce many other specialized diagnostic statistics, including the following:

- collinearity diagnostics to measure how strongly regressors are related to other regressors and how this affects the stability and variance of the estimates (REG)
- influence diagnostics to measure how each individual observation contributes to determining the parameter estimates, the SSE, and the fitted values (LOGISTIC, MIXED, REG, RSREG)
- lack-of-fit diagnostics that measure the lack of fit of the regression model by comparing the error variance estimate to another pure error variance that is not dependent on the form of the model (CATMOD, PROBIT, RSREG)
- diagnostic scatter plots that check the fit of the model and highlighted scatter plots that identify particular observations or groups of observations (REG)
- predicted and residual values, and confidence intervals for the mean and for an individual value (GLM, LOGISTIC, REG)
- time-series diagnostics for equally spaced time-series data that measure how much errors might be related across neighboring observations. These diagnostics can also measure functional goodness of fit for data sorted by regressor or response variables (REG, SAS/ETS procedures).

Many SAS/STAT procedures produce general and specialized statistical graphics through ODS Graphics to diagnose the fit of the model and the model-data agreement, and to highlight observations that are influential on the analysis. Figure 4.2 and Figure 4.3, for example, are two of the ODS statistical graphs produced by the REG procedure by default for the simple linear regression model. For general information about ODS Graphics, see Chapter 21, "Statistical Graphics Using ODS." For specific information about the ODS statistical graphs available with a SAS/STAT procedure, see the PLOTS option in the "Syntax" section for the PROC statement and the "ODS Table Names" section in the "Details" section of the individual procedure documentation.

Linear Regression: The REG Procedure

The REG procedure is a general-purpose procedure for linear regression that does the following:

- handles multiple regression models
- provides nine model-selection methods
- allows interactive changes both in the model and in the data used to fit the model
- allows linear equality restrictions on parameters
- tests linear hypotheses and multivariate hypotheses
- produces collinearity diagnostics, influence diagnostics, and partial regression leverage plots

- saves estimates, predicted values, residuals, confidence limits, and other diagnostic statistics in output SAS data sets
- generates plots of data and of various statistics
- "paints" or highlights scatter plots to identify particular observations or groups of observations
- uses, optionally, correlations or crossproducts for input

Model-Selection Methods in Linear Regression Models

An important step in building statistical models is to determine which effects and variables affect the response variable and to form a model that fits the data well without incurring the negative effects of overfitting the model. Models that are overfit—that is, contain too many regressor variables and unimportant regressor variables—have a tendency to be too closely molded to a a particular set of data, have unstable regression coefficients, and possibly have poor predictive precision. In situations where many potential regressor variables are available for inclusion in a regression model, guided, numerical variable-selection methods offer one approach to model building.

The model-selection techniques for linear regression models implemented in the REG procedure are as follows:

NONE — specifies that no selection be made. This method is the default and uses the full model given in the MODEL statement to fit the linear regression.

FORWARD — specifies that variables be selected based on a forward-selection algorithm. This method starts with no variables in the model and adds variables one by one to the model. At each step, the variable added is the one that most improves the fit of the model. You can also specify groups of variables to treat as a unit during the selection process. An option enables you to specify the criterion for inclusion.

BACKWARD — specifies that variables be selected based on a backward-elimination algorithm. This method starts with a full model and eliminates variables one by one from the model. At each step, the variable with the smallest contribution to the model is deleted. You can also specify groups of variables to treat as a unit during the selection process. An option enables you to specify the criterion for exclusion.

STEPWISE — specifies that variables be selected for the model based on a stepwise-regression algorithm, which combines forward-selection and backward-elimination steps. This method is a modification of the forward-selection method in that variables already in the model do not necessarily stay there. You can also specify groups of variables to treat as a unit during the selection process. Again, options enable you to specify criteria for entry into the model and for remaining in the model.

MAXR — specifies that model formation be based on the maximum R^2 improvement. This method tries to find the best one-variable model, the best two-variable model, and so on. The MAXR method differs from the STEPWISE method in that many more models are evaluated. The MAXR method considers all possible variable exchanges before making any exchange. The STEPWISE method might remove the "worst" variable without considering what the "best" remaining variable might accomplish, whereas MAXR would consider what the "best" remaining variable might accomplish. Consequently, model building based on the maximum R^2 improvement typically takes much longer to run than stepwise model building.

MINR — specifies that model formation be based on the minimum R^2 improvement. This method closely resembles MAXR, but the switch chosen is the one that produces the smallest increase in R^2.

RSQUARE — finds a specified number of models having the highest R^2 in each of a range of model sizes.

CP — finds a specified number of models with the lowest C_p within a range of model sizes.

ADJRSQ — finds a specified number of models having the highest adjusted R^2 within a range of model sizes.

The GLMSELECT procedure has been specifically designed for the purpose of model building in linear models. In addition to having a wider array of selection methods and criteria compared to the REG procedure, the GLMSELECT procedure also supports classification variables. See Chapter 42, "The GLMSELECT Procedure," for more information.

Regression with the REG and GLM Procedures

In terms of the assumptions about the basic model and the estimation principles, the REG and GLM procedures are very closely related. Both procedures estimate parameters by ordinary or weighted least squares and assume homoscedastic, uncorrelated model errors with zero mean. An assumption of normality of the model errors is not necessary for parameter estimation, but it is implied in confirmatory inference based on the parameter estimates—that is, the computation of tests, *p*-values, and confidence and prediction intervals.

The GLM procedure supports a CLASS statement for the levelization of classification variables; see the section "Parameterization of Model Effects" on page 368 in Chapter 18, "Shared Concepts and Topics," on the parameterization of classification variables in statistical models. Classification variables are accommodated in the REG procedure by the inclusion of the necessary dummy regressor variables.

Most of the statistics based on predicted and residual values that are available in PROC REG are also available in PROC GLM. However, PROC GLM does not produce collinearity diagnostics, influence diagnostics, or scatter plots. In addition, PROC GLM allows only one model and fits the full model.

Both procedures are interactive, in that they do not stop after processing a RUN statement. The procedures accept statements until a QUIT statement is submitted.

Response Surface Regression: The RSREG Procedure

The RSREG procedure fits a quadratic response-surface model, which is useful in searching for factor values that optimize a response. The following features in PROC RSREG make it preferable to other regression procedures for analyzing response surfaces:

- automatic generation of quadratic effects
- a lack-of-fit test
- solutions for critical values of the surface
- eigenvalues of the associated quadratic form
- a ridge analysis to search for the direction of optimum response

Partial Least Squares Regression: The PLS Procedure

The PLS procedure fits models by using any of a number of linear predictive methods, including *partial least squares* (PLS). Ordinary least squares regression, as implemented in the GLM or REG procedure, has the single goal of minimizing sample response prediction error by seeking linear functions of the predictors that explain as much variation in each response as possible. The techniques implemented in the PLS procedure have the additional goal of accounting for variation in the predictors under the assumption that directions in the predictor space that are well sampled should provide better prediction for *new* observations when the predictors are highly correlated. All of the techniques implemented in the PLS procedure work by extracting successive linear combinations of the predictors, called *factors* (also called *components* or *latent vectors*), which optimally address one or both of these two goals—explaining response variation and explaining predictor variation. In particular, the method of partial least squares balances the two objectives, seeking factors that explain both response and predictor variation.

Generalized Linear Regression

As outlined in the section "Generalized Linear Models" on page 37 of Chapter 3, "Introduction to Statistical Modeling with SAS/STAT Software," the class of generalized linear model generalizes the linear regression model in two ways:

- by allowing the data to come from a distribution that is a member of the exponential family of distributions
- by introducing a link function $g(\cdot)$ that provides a mapping between the linear predictor $\eta = \mathbf{x}'\boldsymbol{\beta}$ and the mean of the data, $g(\mathrm{E}[Y]) = \eta$. The link function is monotonic, so that $\mathrm{E}[Y] = g^{-1}(\eta)$ and $g^{-1}(\cdot)$ is called the inverse link function.

One of the most commonly used generalized linear regression models is the logistic model for binary or binomial data. Suppose that Y denotes a binary outcome variable that takes on the values 1 and 0 with probabilities π and $1 - \pi$, respectively. The probability π is also referred to as the "success probability," supposing that the coding $Y = 1$ corresponds to a success in a Bernoulli experiment. The success probability is also the mean of Y, and one of the aims of logistic regression analysis is to study how regressor variables affect the outcome probabilities or functions thereof, such as odds ratios.

The logistic regression model for π is defined by a linear predictor $\eta = \mathbf{x}'\boldsymbol{\beta}$ and the logit link function:

$$\text{logit}\{\Pr(Y = 0)\} = \log\left\{\frac{\pi}{1-\pi}\right\} = \mathbf{x}'\boldsymbol{\beta}$$

The inversely linked linear predictor function in this model is

$$\Pr(Y = 0) = \frac{1}{1 + \exp\{-\eta\}}$$

An extension of the dichotomous logistic regression model is models for multinomial (polychotomous) data. Two classes of models for multinomial data can be fit with procedures in SAS/STAT software: models for ordinal data that rely on cumulative link functions and models for nominal (unordered) outcomes that rely on generalized logits. The next section briefly discusses SAS/STAT procedures for logistic regression. See Chapter 8, "Introduction to Categorical Data Analysis Procedures," for more information about the comparison of the procedures mentioned there with respect to analysis of categorical responses.

Logistic Regression

The SAS/STAT procedures CATMOD, GENMOD, GLIMMIX, LOGISTIC, and PROBIT can fit generalized linear models for binary, binomial, and multinomial outcomes.

CATMOD — provides maximum likelihood estimation for logistic regression, including the analysis of logits for dichotomous outcomes and the analysis of generalized logits for polychotomous outcomes. The CATMOD procedure can analyze data represented by a contingency table.

GENMOD — is a general modeling procedure for generalized linear models. It estimates parameters by maximum likelihood. Like the LOGISTIC procedure, it uses CLASS and MODEL statements in SAS/STAT procedures to form the statistical model and can fit models to binary and ordinal outcomes. The GENMOD procedure does not fit generalized logit models for nominal outcomes. However, the procedure also provides the capability of solving generalized estimating equations (GEE) to model correlated data and can perform a Bayesian analysis.

GLIMMIX — is a general modeling procedure for generalized linear mixed models. If the model does not contain random effects, the GLIMMIX procedure fits generalized linear models by the method of maximum likelihood. In the class of logistic regression models, the procedure can fit models to binary, binomial, ordinal, and nominal outcomes.

LOGISTIC	is specifically designed for logistic regression and estimates parameters by maximum likelihood. The procedure fits the usual logistic regression model for binary data as well as models with cumulative link function for ordinal data (such as the proportional odds model) and the generalized logit model for nominal data. The LOGISTIC procedure offers a number of variable selection methods and can perform conditional and exact conditional logistic regression analysis.
PROBIT	calculates maximum likelihood estimates of regression parameters and the natural (or threshold) response rate for quantal response data from biological assays or other discrete event data. This includes probit, logit, ordinal logistic, and extreme value (or gompit) regression models.
SURVEYLOGISTIC	is designed for logistic regression and estimates parameters by maximum likelihood. The procedure incorporates complex survey sample designs, including designs with stratification, clustering, and unequal weighting.

Other Generalized Linear Models

When a generalized linear model is formed with distributions other than the binary, binomial, or multinomial, you can use the GENMOD and GLIMMIX procedures for parameter estimation and inference.

Both procedures can accommodate correlated observations, but they use different techniques to accomplish this goal. The GENMOD procedure can fit correlated data models via generalized estimating equations that rely on a first- and second-moment specification for the response data and a working correlation assumption. With the GLIMMIX procedure, you can model correlations between the observations by (1) specifying random effects in the conditional distribution that induce a marginal correlation structure or (2) direct modeling of the marginal dependence. The GLIMMIX procedure employs likelihood-based techniques in parameter estimation.

The GENMOD procedure supports a Bayesian analysis through its BAYES statement.

With the GLIMMIX procedure you can vary the distribution or link function on an observation-by-observation basis.

To fit a generalized linear model with a distribution that is not available in the GENMOD or GLIMMIX procedure, you can use the NLMIXED procedure and code the log-likelihood function of an observation with SAS programming statements.

Regression for Ill-Conditioned Data: The ORTHOREG Procedure

The ORTHOREG procedure performs linear least squares regression by using the Gentleman-Givens computational method, and it can produce more accurate parameter estimates for ill-conditioned data. PROC GLM and PROC REG produce very accurate estimates for most problems. However, if you have very ill-conditioned data, consider using the ORTHOREG procedure. The collinearity diagnostics in PROC REG can help you to determine whether PROC ORTHOREG would be useful.

Quantile Regression: The QUANTREG Procedure

The QUANTREG procedure models the effects of covariates on the conditional quantiles of a response variable by means of quantile regression.

Ordinary least squares regression models the relationship between one or more covariates X and the conditional mean of the response variable $\mathrm{E}[Y|X = x]$. Quantile regression extends the regression model to conditional quantiles of the response variable, such as the 90th percentile. Quantile regression is particularly useful when the rate of change in the conditional quantile, expressed by the regression coefficients, depends on the quantile. An advantage of quantile regression over least squares regression is its flexibility for modeling data with heterogeneous conditional distributions. Data of this type occur in many fields, including biomedicine, econometrics, and ecology.

Features that you will find in the QUANTREG procedure include the following:

- simplex, interior point, and smoothing algorithms for estimation
- sparsity, rank, and resampling methods for confidence intervals
- asymptotic and bootstrap methods to estimate covariance and correlation matrices of the parameter estimates
- Wald and likelihood ratio tests for the regression parameter estimates
- regression quantile spline fits

Nonlinear Regression

Recall from Chapter 3, "Introduction to Statistical Modeling with SAS/STAT Software," that a nonlinear regression model is a statistical model in which the mean function depends on the model parameters in a nonlinear function. The SAS/STAT procedures that can fit general, nonlinear models are the NLIN and NLMIXED procedures. The procedures have the following in common:

- Nonlinear models are fit by iterative methods.
- You must provide an expression for the model through programming statements.
- Analytic derivatives of the objective function with respect to the parameters are calculated automatically.
- A grid search is available to select the best starting values for the parameters from a set of starting points that you provide.

The following items reflect some important differences between the NLIN and NLMIXED procedures:

- Parameters are estimated by nonlinear least squares with the NLIN procedure and by maximum likelihood with the NLMIXED procedure.
- The NLMIXED procedure enables you to construct nonlinear models that contain normally distributed random effects.
- The NLIN procedure requires that you declare all model parameters in the PARAMETERS statement and assign starting values. The NLMIXED procedure determines the parameters in your model based on the PARAMETER statement and the other modeling statements. It is not necessary to supply starting values for all parameters in the NLMIXED procedure, but it is highly recommended.
- The residual variance is not a parameter in models fit with the NLIN procedure, but it is in models fit with the NLMIXED procedure.
- The default iterative optimization method in the NLIN procedure is the Gauss-Newton method; the default method in the NLMIXED procedure is the quasi-Newton method. Other optimization techniques are available in both procedures.

Nonlinear models are fit with iterative techniques that begin from starting values and attempt to iteratively improve on the estimates by updating the estimates. There is no guarantee that the solution achieved when the iterative algorithm converges will correspond to a global optimum.

Nonparametric Regression

Regression models that suppose a parametric form express the mean of an observation as a function of regressor variables $x_1, \cdots, x_k$ and parameters $\beta_1, \cdots, \beta_p$:

$$\mathrm{E}[Y] = f(x_1, \cdots, x_k; \beta_1, \cdots, \beta_p)$$

Nonparametric regression techniques not only relax the assumption of linearity in the regression parameters, but they also do not require that you specify a precise functional form for the relationship between response and regressor variables. Consider a regression problem where the relationship between response Y and regressor X is to be modeled. It is assumed that $\mathrm{E}[Y_i] = g(x_i) + \epsilon_i$, where $g(\cdot)$ is an unspecified regression function. Two primary approaches in nonparametric regression modeling are as follows:

- approximate $g(x_i)$ locally by a parametric function constructed from information in a local neighborhood of x_i
- approximate the unknown function $g(x_i)$ by a smooth, flexible function and determine the necessary smoothness and continuity properties from the data

The SAS/STAT procedures LOESS, GAM, and TPSPLINE fit nonparametric regression models by one of these methods.

Local Regression: The LOESS Procedure

The LOESS procedure implements a local regression approach for estimating regression surfaces pioneered by Cleveland, Devlin, and Grosse (1988). No assumptions about the parametric form of the entire regression surface are made with the LOESS procedure. Only a parametric form of the local approximation is specified by the user. Furthermore, the LOESS procedure is suitable when there are outliers in the data and a robust fitting method is necessary.

Smooth Function Approximation: The TPSPLINE Procedure

The TPSPLINE procedure decomposes the regressor contributions to the mean function into parametric components and into smooth functional components. Suppose that the regressor variables are collected into the vector $\mathbf{x}$ and that this vector is partitioned as $\mathbf{x} = [\mathbf{x}_1' \ \mathbf{x}_2']'$. The relationship between Y and $\mathbf{x}_2$ is linear (parametric) and the relationship between Y and $\mathbf{x}_1$ is nonparametric. The TPSPLINE procedure fits models of the form

$$\mathrm{E}[Y] = g(\mathbf{x}_1) + \mathbf{x}_2'\boldsymbol{\beta}$$

The function $g(\cdot)$ can be represented as a sequence of spline basis functions.

The parameters are estimated by a penalized least squares method. The penalty is applied to the usual least squares criterion to obtain a regression estimate that fits the data well and to prevent the fit from attempting to interpolate the data (fit the data too closely).

Generalized Additive Models: The GAM Procedure

Generalized additive models are nonparametric models in which one or more regressor variables are present and can make different smooth contributions to the mean function. For example, if $\mathbf{x}_i = [x_{i1}, x_{i2}, \cdots, x_{ik}]$ is a vector of k regressor for the ith observation, then an additive model represents the mean function as

$$\mathrm{E}[Y] = f_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_3(x_{i3})$$

The individual functions f_j can have a parametric or nonparametric form. If all f_j are parametric, the GAM procedure fits a fully parametric model. If some f_j are nonparametric, the GAM procedure fits a semiparametric model. Otherwise, the models are fully nonparametric.

The generalization of additive models is akin to the generalization for linear models: nonnormal data are accommodated by explicitly modeling the distribution of the data as a member of the exponential family and by applying a monotonic link function that provides a mapping between the predictor and the mean of the data.

Robust Regression: The ROBUSTREG Procedure

The ROBUSTREG procedure implements algorithms to detect outliers and provide resistant (stable) results in the presence of outliers. The ROBUSTREG procedure provides four such methods: M estimation, LTS estimation, S estimation, and MM estimation.

- M estimation was introduced by Huber (1973), and it is the simplest approach both computationally and theoretically. Although it is not robust with respect to leverage points, it is still used extensively in analyzing data for which it can be assumed that the contamination is mainly in the response direction.
- Least trimmed squares (LTS) estimation is a high breakdown value method introduced by Rousseeuw (1984). The breakdown value is a measure of the proportion of contamination that an estimation method can withstand and still maintain its robustness.
- S estimation is a high breakdown value method introduced by Rousseeuw and Yohai (1984). With the same breakdown value, it has a higher statistical efficiency than LTS estimation.
- MM estimation, introduced by Yohai (1987), combines high breakdown value estimation and M estimation. It has both the high breakdown property and a higher statistical efficiency than S estimation.

Regression with Transformations: The TRANSREG Procedure

The TRANSREG procedure can fit many standard linear models. In addition, PROC TRANSREG can find nonlinear transformations of the data and fit a linear model to the transformed variables. This is in contrast to PROC REG and PROC GLM, which fit linear models to data, or PROC NLIN, which fits nonlinear models to data. The TRANSREG procedure fits many types of linear models, including the following:

- ordinary regression and ANOVA
- metric and nonmetric conjoint analysis
- metric and nonmetric vector and ideal point preference mapping
- simple, multiple, and multivariate regression with variable transformations
- redundancy analysis with variable transformations
- canonical correlation analysis with variable transformations
- response surface regression with variable transformations

Interactive Features in the CATMOD, GLM, and REG Procedures

The CATMOD, GLM, and REG procedures do not stop after processing a RUN statement. More statements can be submitted as a continuation of the previous statements. Many new features in these procedures are useful to request after you have reviewed the results from previous statements. The procedures stop if a DATA step or another procedure is requested or if a QUIT statement is submitted.

Statistical Background in Linear Regression

The remainder of this chapter outlines the way in which many SAS/STAT regression procedures calculate various regression quantities. The discussion focuses on the linear regression models. General statistical background about linear statistical models can be found in the section "Linear Model Theory" of Chapter 3, "Introduction to Statistical Modeling with SAS/STAT Software."

Exceptions and further details are documented with individual procedures.

Linear Regression Models

In matrix notation, a linear model is written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{X}$ is the $(n \times k)$ design matrix (rows are observations and columns are the regressors), $\boldsymbol{\beta}$ is the $(k \times 1)$ vector of unknown parameters, and $\boldsymbol{\epsilon}$ is the $(n \times 1)$ vector of unobservable model errors. The first column of $\mathbf{X}$ is usually a vector of 1s and is used to estimate the intercept term.

The statistical theory of linear models is based on strict classical assumptions. Ideally, the response is measured with all the factors controlled in an experimentally determined environment. If you cannot control the factors experimentally, some tests must be interpreted as being conditional on the observed values of the regressors.

Other assumptions are as follows:

- The form of the model is correct (all important explanatory variables have been included). This assumption is reflected mathematically in the assumption of a zero mean of the model errors, $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$.
- Regressor variables are measured without error.
- The expected value of the errors is zero.
- The variance of the error (and thus the dependent variable) for the ith observation is σ^2/w_i, where w_i is a known weight factor. Usually, $w_i = 1$ for all i and thus σ^2 is the common, constant variance.
- The errors are uncorrelated across observations.

When hypotheses are tested, or when confidence and prediction intervals are computed, an additional assumption is made that the errors are normally distributed.

Parameter Estimates and Associated Statistics

The Least Squares Estimators

Least squares estimators of the regression parameters are found by solving the normal equations

$$(\mathbf{X}'\mathbf{W}\mathbf{X})\,\boldsymbol{\beta} = \mathbf{X}'\mathbf{W}\mathbf{Y}$$

for the vector $\boldsymbol{\beta}$, where $\mathbf{W}$ is a diagonal matrix with the observed weights on the diagonal. The resulting estimator of the parameter vector is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{W}\mathbf{Y}$$

This is an unbiased estimator, since

$$\begin{aligned} \mathrm{E}\left[\widehat{\boldsymbol{\beta}}\right] &= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{W}\mathrm{E}[\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{W}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} \end{aligned}$$

Notice that the only assumption necessary in order for the least squares estimators to be unbiased is that of a zero mean of the model errors. If the estimator is evaluated at the observed data, it is referred to Subidxleast squaresestimate (Introduction to Regression)as the least squares estimate,

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{W}\mathbf{y}$$

If the standard classical assumptions are met, the least squares estimators of the regression parameters are the best linear unbiased estimators (BLUE). In other words, the estimators have minimum variance in the class of estimators that are unbiased and that are linear functions of the responses. If the additional assumption of normally distributed errors is satisfied, then the following are true:

- The statistics that are computed have the proper sampling distributions for hypothesis testing.
- Parameter estimators are normally distributed.
- Various sums of squares are distributed proportional to chi-square, at least under proper hypotheses.
- Ratios of estimators to standard errors follow the Student's t distribution under certain hypotheses.
- Appropriate ratios of sums of squares follow an F distribution for certain hypotheses.

When regression analysis is used to model data that do not meet the assumptions, the results should be interpreted in a cautious, exploratory fashion. The significance probabilities under these circumstances are unreliable.

Box (1966) and Mosteller and Tukey (1977, Chapters 12 and 13) discuss the problems that are encountered with regression data, especially when the data are not under experimental control.

Estimating the Precision

Assume for the present that $\mathbf{X}'\mathbf{W}\mathbf{X}$ has full column rank k (this assumption is relaxed later). The variance of the error terms, $\mathrm{Var}[\epsilon_i] = \sigma^2$, is then estimated by the mean square error

$$s^2 = \mathrm{MSE} = \frac{\mathrm{SSE}}{n-k} = \frac{1}{n-k} \sum_{i=1}^{n} w_i \left(y_i - \mathbf{x}_i' \boldsymbol{\beta} \right)^2$$

where $\mathbf{x}_i'$ is the ith row of the design matrix $\mathbf{X}$. The residual variance estimate is also unbiased: $\mathrm{E}[s^2] = \sigma^2$.

The covariance matrix of the least squares estimators is

$$\mathrm{Var}\left[\widehat{\boldsymbol{\beta}}\right] = \sigma^2 \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}$$

An estimate of the covariance matrix is obtained by replacing σ^2 with its estimate, s^2 in the preceding formula. This estimate is often referred to as COVB in SAS/STAT modeling procedures:

$$\mathrm{COVB} = \widehat{\mathrm{Var}}\left[\widehat{\boldsymbol{\beta}}\right] = s^2 \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}$$

The correlation matrix of the estimates, often referred to as CORRB, is derived by scaling the covariance matrix: Let $\mathbf{S} = \mathrm{diag}\left(\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\right)^{-\frac{1}{2}}$. Then the correlation matrix of the estimates is

$$\mathrm{CORRB} = \mathbf{S} \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1} \mathbf{S}$$

The estimated standard error of the ith parameter estimator is obtained as the square root of the ith diagonal element of the COVB matrix. Formally,

$$\mathrm{STDERR}\left(\widehat{\boldsymbol{\beta}}_i\right) = \sqrt{\left[s^2 (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\right]_{ii}}$$

The ratio

$$t = \frac{\widehat{\boldsymbol{\beta}}_i}{\mathrm{STDERR}\left(\widehat{\boldsymbol{\beta}}_i\right)}$$

follows a Student's t distribution with $(n-k)$ degrees of freedom under the hypothesis that β_i is zero and provided that the model errors are normally distributed.

Regression procedures display the t ratio and the significance probability, which is the probability under the hypothesis $H\colon \beta_i = 0$ of a larger absolute t value than was actually obtained. When the probability is less than some small level, the event is considered so unlikely that the hypothesis is rejected.

Type I SS and Type II SS measure the contribution of a variable to the reduction in SSE. Type I SS measure the reduction in SSE as that variable is entered into the model in sequence. Type II SS are the increment in SSE that results from removing the variable from the full model. Type II SS are equivalent to the Type III and Type IV SS reported in the GLM procedure. If Type II SS are used in the numerator of an F test, the test is equivalent to the t test for the hypothesis that the parameter is zero. In polynomial models, Type I SS measure the contribution of each polynomial term after it is orthogonalized to the previous terms in the model. The four types of SS are described in Chapter 15, "The Four Types of Estimable Functions."

Coefficient of Determination

The coefficient of determination in a regression model, also known as the R-square statistic (R^2), measures the proportion of variability in the response that is explained by the regressor variables. In a linear regression model with intercept, R^2 is defined as

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

where SSE is the residual (error) sum of squares and SST is the total sum of squares corrected for the mean. The adjusted R^2 statistic is an alternative to R^2 that takes into account the number of parameters in the model. This statistic is calculated as

$$\text{ADJRSQ} = 1 - \frac{n - i}{n - p}\left(1 - R^2\right)$$

where n is the number of observations used to fit the model, p is the number of parameters in the model (including the intercept), and i is 1 if the model includes an intercept term, and 0 otherwise.

R^2 statistics also play an important indirect role in regression calculations. For example, the proportion of variability explained by regressing all other variables in a model on a particular regressor can provide insights into the interrelationship among the regressors.

Tolerances and variance inflation factors measure the strength of interrelationships among the regressor variables in the model. If all variables are orthogonal to each other, both tolerance and variance inflation are 1. If a variable is very closely related to other variables, the tolerance approaches 0 and the variance inflation gets very large. Tolerance (TOL) is 1 minus the R^2 that results from the regression of the other variables in the model on that regressor. Variance inflation (VIF) is the diagonal of $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$, if $(\mathbf{X}'\mathbf{W}\mathbf{X})$ is scaled to correlation form. The statistics are related as

$$\text{VIF} = \frac{1}{\text{TOL}}$$

Explicit and Implicit Intercepts

A linear model contains an *explicit* intercept if the $\mathbf{X}$ matrix contains a column whose nonzero values do not vary, typically a column of ones. Many SAS/STAT procedures automatically add this column of ones as the first column in the $\mathbf{X}$ matrix. Procedures that support a NOINT option in the MODEL statement provide the capability to suppress the automatic addition of the intercept column.

In general, models without intercept should be the exception, especially if your model does not contain classification variables. An overall intercept is provided in many models to adjust for the grand total or overall mean in your data. A simple linear regression without intercept, such as

$$\text{E}[Y_i] = \beta_1 x_i + \epsilon_i$$

assumes that Y has mean zero if X takes on the value zero. This might not be a reasonable assumption.

If you explicitly suppress the intercept in a statistical model, the calculation and interpretation of your results can change. For example, the exclusion of the intercept in the following PROC REG statements leads to a different calculation of the R-square statistic. It also affects the calculation of the sum of squares in the analysis of variance for the model. For example, the model and error sum of squares add up to the uncorrected total sum of squares in the absence of an intercept.

```
proc reg;
   model y = x / noint;
quit;
```

Many statistical models contain an *implicit* intercept. This occurs when a linear function of one or more columns in the **X** matrix produces a column of constant, nonzero values. For example, the presence of a CLASS variable in the GLM parameterization always implies an intercept in the model. If a model contains an implicit intercept, adding an intercept to the model does not alter the quality of the model fit, but it changes the interpretation (and number) of the parameter estimates.

The way in which the implicit intercept is detected and accounted for in the analysis depends on the procedure. For example, the following statements in the GLM procedure lead to an implied intercept:

```
proc glm;
   class a;
   model y = a / solution noint;
run;
```

Whereas the analysis of variance table uses the uncorrected total sum of squares (due to the NOINT option), the implied intercept does not lead to a redefinition or recalculation of the R-square statistic (compared to the model without the NOINT option). Also, because the intercept is implied by the presence of the CLASS variable a in the model, the same error sum of squares results whether the NOINT option is specified or not.

A different approach is taken, for example, by the TRANSREG procedure. The ZERO=NONE option in the CLASS parameterization of the following statements leads to an implicit intercept model:

```
proc transreg;
   model ide(y) = class(a / zero=none) / ss2;
run;
```

The analysis of variance table or the regression fit statistics are not affected in the TRANSREG procedure. Only the interpretation of the parameter estimates changes because of the way in which the intercept is accounted for in the model.

Implied intercepts not only occur when classification effects are present in the model. They also occur with B-splines and other sets of constructed columns.

Models Not of Full Rank

If the **X** matrix is not of full rank, then a generalized inverse can be used to solve the normal equations to minimize the SSE:

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{W}\mathbf{y}$$

However, these estimates are not unique since there are an infinite number of solutions corresponding to different generalized inverses. PROC REG and other regression procedures choose a nonzero solution for all variables that are linearly independent of previous variables and a zero solution for other variables. This corresponds to using a generalized inverse in the normal equations, and the expected values of the estimates are the Hermite normal form of $\mathbf{X}'\mathbf{W}\mathbf{X}$ multiplied by the true parameters:

$$\mathrm{E}\left[\widehat{\boldsymbol{\beta}}\right] = \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-}(\mathbf{X}'\mathbf{W}\mathbf{X})\boldsymbol{\beta}$$

Degrees of freedom for the estimates that correspond to singularities are not counted (reported as zero). The hypotheses that are not testable have t tests displayed as missing. The message that the model is not of full rank includes a display of the relations that exist in the matrix.

See the sections "Generalized Inverse Matrices" and "Linear Model Theory" in Chapter 3, "Introduction to Statistical Modeling with SAS/STAT Software," on the nature and construction of generalized inverses and their importance for statistical inference in linear models.

Predicted and Residual Values

After the model has been fit, predicted and residual values are usually calculated, graphed, and output. The predicted values are calculated from the estimated regression equation; the raw residuals are calculated as the observed minus the predicted value. Often other forms of residuals are used for model diagnostics, such as studentized or cumulative residuals. Some procedures can calculate standard errors of residuals, predicted mean values, and individual predicted values.

Consider the ith observation where $\mathbf{x}_i'$ is the row of regressors, $\widehat{\boldsymbol{\beta}}$ is the vector of parameter estimates, and s^2 is the estimate of the residual variance (the mean squared error). The *leverage* value of the ith observation is defined as

$$h_i = w_i \mathbf{x}_i'(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{x}_i$$

where $\mathbf{X}$ is the design matrix for the observed data, $\mathbf{x}_i'$ is an arbitrary regressor vector (possibly but not necessarily a row of $\mathbf{X}$), $\mathbf{W}$ is a diagonal matrix with the observed weights on the diagonal, and w_i is the weight corresponding to $\mathbf{x}_i'$.

Then the predicted mean and the standard error of the predicted mean are

$$\widehat{y}_i = \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$$

$$\mathrm{STDERR}(\widehat{y}_i) = \sqrt{s^2 h_i / w_i}$$

The standard error of the individual (future) predicted value y_i is

$$\mathrm{STDERR}(y_i) = \sqrt{s^2(1 + h_i)/w_i}$$

If the predictor vector $\mathbf{x}_i$ corresponds to an observation in the analysis data, then the raw residual for that observation and the standard error of the raw residual are defined as

$$\mathrm{RESID}_i = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$$

$$\mathrm{STDERR}(\mathrm{RESID}_i) = \sqrt{s^2(1 - h_i)/w_i}$$

The *studentized residual* is the ratio of the raw residual and its estimated standard error. Symbolically,

$$\text{STUDENT}_i = \frac{\text{RESID}_i}{\text{STDERR}(\text{RESID}_i)}$$

There are two kinds of intervals involving predicted values that are associated with a measure of confidence: the *confidence* interval for the mean value of the response and the *prediction* (or *forecasting*) interval for an individual observation. As discussed in the section "Mean Squared Error" in Chapter 3, "Introduction to Statistical Modeling with SAS/STAT Software," both intervals are based on the mean squared error of predicting a target based on the result of the model fit. The difference in the expressions for the confidence interval and the prediction interval comes about because the target of estimation is a constant in the case of the confidence interval (the mean of an observation) and the target is a random variable in the case of the prediction interval (a new observation).

For example, you can construct a confidence interval for the ith observation that contains the true mean value of the response with probability $1 - \alpha$. The upper and lower limits of the confidence interval for the mean value are

$$\begin{aligned} \text{LowerM} &= \mathbf{x}_i' \widehat{\boldsymbol{\beta}} - t_{\alpha/2,\nu} \sqrt{s^2 h_i / w_i} \\ \text{UpperM} &= \mathbf{x}_i' \widehat{\boldsymbol{\beta}} + t_{\alpha/2,\nu} \sqrt{s^2 h_i / w_i} \end{aligned}$$

where $t_{\alpha/2,\nu}$ is the tabulated t quantile with degrees of freedom equal to the degrees of freedom for the mean squared error, $\nu = n - \text{rank}(\mathbf{X})$.

The limits for the prediction interval for an individual response are

$$\begin{aligned} \text{LowerI} &= \mathbf{x}_i' \widehat{\boldsymbol{\beta}} - t_{\alpha/2,\nu} \sqrt{s^2 (1 + h_i) / w_i} \\ \text{UpperI} &= \mathbf{x}_i' \widehat{\boldsymbol{\beta}} + t_{\alpha/2,\nu} \sqrt{s^2 (1 + h_i) / w_i} \end{aligned}$$

Influential observations are those that, according to various criteria, appear to have a large influence on the analysis. One measure of influence, Cook's D, measures the change to the estimates that results from deleting an observation:

$$\text{COOKD}_i = \frac{1}{k} \text{STUDENT}_i^2 \left(\frac{\text{STDERR}(\widehat{y}_i)}{\text{STDERR}(\text{RESID}_i)} \right)^2$$

where k is the number of parameters in the model (including the intercept). For more information, see Cook (1977, 1979).

The *predicted residual* for observation i is defined as the residual for the ith observation that results from dropping the ith observation from the parameter estimates. The sum of squares of predicted residual errors is called the *PRESS statistic*:

$$\begin{aligned} \text{PRESID}_i &= \frac{\text{RESID}_i}{1 - h_i} \\ \text{PRESS} &= \sum_{i=1}^{n} w_i \text{PRESID}_i^2 \end{aligned}$$

Testing Linear Hypotheses

Testing of linear hypothesis based on estimable functions is discussed in the section "Test of Hypotheses" on page 66 in Chapter 3, "Introduction to Statistical Modeling with SAS/STAT Software," and the construction of special sets of estimable functions corresponding to Type I–Type IV hypotheses is discussed in Chapter 15, "The Four Types of Estimable Functions." In linear regression models, testing of general linear hypotheses follows along the same lines. Test statistics are usually formed based on sums of squares associated with the hypothesis in question. Furthermore, when $\mathbf{X}$ is of full rank—as is the case in many regression models—the consistency of the linear hypothesis is guaranteed.

Recall from Chapter 3, "Introduction to Statistical Modeling with SAS/STAT Software," that the general form of a linear hypothesis for the parameters is $H\colon \mathbf{L}\boldsymbol{\beta} = \mathbf{d}$, where $\mathbf{L}$ is $(q \times k)$, $\boldsymbol{\beta}$ is $(k \times 1)$, and $\mathbf{d}$ is $(q \times 1)$. To test this hypothesis, the linear function is taken with respect to the parameter estimates: $\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{d}$. This linear function in $\widehat{\boldsymbol{\beta}}$ has variance

$$\mathrm{Var}\left[\mathbf{L}\widehat{\boldsymbol{\beta}}\right] = \mathbf{L}\mathrm{Var}[\widehat{\boldsymbol{\beta}}]\mathbf{L}' = \sigma^2\mathbf{L}\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-}\mathbf{L}'$$

The *sum of squares due to the hypothesis* is a simple quadratic form:

$$\mathrm{SS}(H) = \mathrm{SS}(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{d}) = \left(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{d}\right)' \left(\mathbf{L}\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-}\mathbf{L}'\right)^{-1}\left(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{d}\right)$$

If this hypothesis is testable, then $\mathrm{SS}(H)$ can be used in the numerator of an F statistic:

$$F = \frac{\mathrm{SS}(H)/q}{s^2} = \frac{\mathrm{SS}(\mathbf{Lb} - \mathbf{d})/q}{s^2}$$

If $\widehat{\boldsymbol{\beta}}$ is normally distributed, which follows as a consequence of normally distributed model errors, then this statistic follows an F distribution with q numerator degrees of freedom and $n - \mathrm{rank}(\mathbf{X})$ denominator degrees of freedom. Note that it was assumed in this derivation that $\mathbf{L}$ is of full row rank q.

Multivariate Tests

Multivariate hypotheses involve several dependent variables in the form

$$H\colon \mathbf{L}\boldsymbol{\beta}\mathbf{M} = \mathbf{d}$$

where $\mathbf{L}$ is a linear function on the regressor side, $\boldsymbol{\beta}$ is a matrix of parameters, $\mathbf{M}$ is a linear function on the dependent side, and $\mathbf{d}$ is a matrix of constants. The special case (handled by PROC REG) in which the constants are the same for each dependent variable is expressed as

$$(\mathbf{L}\boldsymbol{\beta} - \mathbf{cj})\mathbf{M} = \mathbf{0}$$

where $\mathbf{c}$ is a column vector of constants and $\mathbf{j}$ is a row vector of 1s. The special case in which the constants are 0 is then

$$\mathbf{L}\boldsymbol{\beta}\mathbf{M} = \mathbf{0}$$

These multivariate tests are covered in detail in Morrison (1976), Timm (1975), Mardia, Kent, and Bibby (1979), Bock (1975), and other works cited in Chapter 9, "Introduction to Multivariate Procedures."

Notice that in contrast to the tests discussed in the preceding section, $\boldsymbol{\beta}$ here is a matrix of parameter estimates. Suppose that the matrix of estimates is denoted as $\mathbf{B}$. To test the multivariate hypothesis, construct two matrices, $\mathbf{H}$ and $\mathbf{E}$, that correspond to the numerator and denominator of a univariate F test:

$$\mathbf{H} = \mathbf{M}'(\mathbf{LB} - \mathbf{cj})'(\mathbf{L}(\mathbf{X}'\mathbf{WX})^{-}\mathbf{L}')^{-1}(\mathbf{LB} - \mathbf{cj})\mathbf{M}$$
$$\mathbf{E} = \mathbf{M}'\left(\mathbf{Y}'\mathbf{WY} - \mathbf{B}'(\mathbf{X}'\mathbf{WX})\mathbf{B}\right)\mathbf{M}$$

Four test statistics, based on the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$ or $(\mathbf{E}+\mathbf{H})^{-1}\mathbf{H}$, are formed. Let λ_i be the ordered eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$ (if the inverse exists), and let ξ_i be the ordered eigenvalues of $(\mathbf{E}+\mathbf{H})^{-1}\mathbf{H}$. It happens that $\xi_i = \lambda_i/(1+\lambda_i)$ and $\lambda_i = \xi_i/(1-\xi_i)$, and it turns out that $\rho_i = \sqrt{\xi_i}$ is the ith canonical correlation.

Let p be the rank of $(\mathbf{H}+\mathbf{E})$, which is less than or equal to the number of columns of $\mathbf{M}$. Let q be the rank of $\mathbf{L}(\mathbf{X}'\mathbf{WX})^{-}\mathbf{L}'$. Let v be the error degrees of freedom and $s = \min(p, q)$. Let $m = (|p - q| - 1)/2$, and let $n = (v - p - 1)/2$. Then the following statistics test the multivariate hypothesis in various ways, and their p-values can be approximated by F distributions. Note that in the special case that the rank of $\mathbf{H}$ is 1, all of these F statistics will be the same and the corresponding p-values will in fact be exact, since in this case the hypothesis is really univariate.

Wilks' Lambda

If

$$\Lambda = \frac{\det(\mathbf{E})}{\det(\mathbf{H}+\mathbf{E})} = \prod_{i=1}^{n} \frac{1}{1+\lambda_i} = \prod_{i=1}^{n}(1-\xi_i)$$

then

$$F = \frac{1-\Lambda^{1/t}}{\Lambda^{1/t}} \cdot \frac{rt-2u}{pq}$$

is approximately F distributed, where

$$r = v - \frac{p-q+1}{2}$$
$$u = \frac{pq-2}{4}$$
$$t = \begin{cases} \sqrt{\frac{p^2q^2-4}{p^2+q^2-5}} & \text{if} p^2+q^2-5 > 0 \\ 1 & \text{otherwise} \end{cases}$$

The degrees of freedom are pq and $rt - 2u$. The distribution is exact if $\min(p, q) \le 2$. (See Rao 1973, p. 556.)

Pillai's Trace

If

$$\mathbf{V} = \text{trace}\left(\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}\right) = \sum_{i=1}^{n} \frac{\lambda_i}{1+\lambda_i} = \sum_{i=1}^{n} \xi_i$$

then

$$F = \frac{2n+s+1}{2m+s+1} \cdot \frac{\mathbf{V}}{s-\mathbf{V}}$$

is approximately F distributed with $s(2m+s+1)$ and $s(2n+s+1)$ degrees of freedom.

Hotelling-Lawley Trace

If

$$U = \text{trace}\left(\mathbf{E}^{-1}\mathbf{H}\right) = \sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} \frac{\xi_i}{1-\xi_i}$$

then for $n > 0$

$$F = (U/c)((4+(pq+2)/(b-1))/(pq))$$

is approximately F distributed with pq and $4+(pq+2)/(b-1)$ degrees of freedom, where $b = (p+2n)(q+2n)/(2(2n+1)(n-1))$ and $c = (2+(pq+2)/(b-1))/(2n)$; while for $n \leq 0$

$$F = \frac{2(sn+1)\mathbf{U}}{s^2(2m+s+1)}$$

is approximately F with $s(2m+s+1)$ and $2(sn+1)$ degrees of freedom.

Roy's Maximum Root

If $\Theta = \lambda_1$, then

$$F = \Theta \frac{v-r+q}{r}$$

where $r = \max(p,q)$ is an upper bound on F that yields a lower bound on the significance level. Degrees of freedom are r for the numerator and $v-r+q$ for the denominator.

Tables of critical values for these statistics are found in Pillai (1960).

Exact Multivariate Tests

Beginning with SAS 9, if you specify the MSTAT=EXACT option in the appropriate statement, p-values for three of the four tests are computed exactly (Wilks' lambda, the Hotelling-Lawley trace, and Roy's greatest root), and the p-values for the fourth (Pillai's trace) are based on an F approximation that is more accurate (but occasionally slightly more liberal) than the default. The exact p-values for Roy's greatest root benefit the most, since in this case the F approximation provides only a lower bound for the p-value. If you use the F-based p-value for this test in the usual way, declaring a test significant if $p < 0.05$, then your decisions might be very liberal. For example, instead of the nominal 5% Type I error rate, such a procedure can easily have an actual Type I error rate in excess of 30%. By contrast, basing such a procedure on the exact p-values will result in the appropriate 5% Type I error rate, under the usual regression assumptions.

The MSTAT=EXACT option is supported in the ANOVA, CANCORR, CANDISC, GLM, and REG procedures.

The exact p-values are based on the following sources:

- **Wilks' lambda:** Lee (1972), Davis (1979)
- **Pillai's trace:** Muller (1998)
- **Hotelling-Lawley trace:** Davis (1970), Davis (1980)
- **Roy's greatest root:** Davis (1972), Pillai and Flury (1984)

Note that, although the MSTAT=EXACT p-value for Pillai's trace is still approximate, it has "substantially greater accuracy" than the default approximation (Muller 1998).

Since most of the MSTAT=EXACT p-values are not based on the F distribution, the columns in the multivariate tests table corresponding to this approximation—in particular, the F value and the numerator and denominator degrees of freedom—are no longer displayed, and the column containing the p-values is labeled "P Value" instead of "Pr > F." Suppose, for example, you use the following PROC ANOVA statements to perform a multivariate analysis of an archaeological data set:

```
data Skulls;
   input Loc $20. Basal Occ Max;
   datalines;
Minas Graes, Brazil  2.068 2.070 1.580
Minas Graes, Brazil  2.068 2.074 1.602
Minas Graes, Brazil  2.090 2.090 1.613
Minas Graes, Brazil  2.097 2.093 1.613
Minas Graes, Brazil  2.117 2.125 1.663
Minas Graes, Brazil  2.140 2.146 1.681
Matto Grosso, Brazil 2.045 2.054 1.580
Matto Grosso, Brazil 2.076 2.088 1.602
Matto Grosso, Brazil 2.090 2.093 1.643
Matto Grosso, Brazil 2.111 2.114 1.643
Santa Cruz, Bolivia  2.093 2.098 1.653
Santa Cruz, Bolivia  2.100 2.106 1.623
Santa Cruz, Bolivia  2.104 2.101 1.653
;
```

```
proc anova data=Skulls;
   class Loc;
   model Basal Occ Max = Loc / nouni;
   manova h=Loc;
   ods select MultStat;
run;
```

The default multivariate tests, based on the F approximations, are shown in Figure 4.5.

Figure 4.5 Default Multivariate Tests

```
                          The ANOVA Procedure
                   Multivariate Analysis of Variance

              MANOVA Test Criteria and F Approximations for
                 the Hypothesis of No Overall Loc Effect
                      H = Anova SSCP Matrix for Loc
                          E = Error SSCP Matrix

                            S=2    M=0    N=3

Statistic                        Value    F Value    Num DF    Den DF    Pr > F

Wilks' Lambda               0.60143661       0.77         6        16    0.6032
Pillai's Trace              0.44702843       0.86         6        18    0.5397
Hotelling-Lawley Trace      0.58210348       0.75         6    9.0909    0.6272
Roy's Greatest Root         0.35530890       1.07         3         9    0.4109

         NOTE: F Statistic for Roy's Greatest Root is an upper bound.
                 NOTE: F Statistic for Wilks' Lambda is exact.
```

If you specify MSTAT=EXACT in the MANOVA statement, as in the following statements, then the displayed output is the much simpler table shown in Figure 4.6.

```
proc anova data=Skulls;
   class Loc;
   model Basal Occ Max = Loc / nouni;
   manova h=Loc / mstat=exact;
   ods select MultStat;
run;
```

Figure 4.6 Multivariate Tests with MSTAT=EXACT

```
                        The ANOVA Procedure
                 Multivariate Analysis of Variance

      MANOVA Tests for the Hypothesis of No Overall Loc Effect
                   H = Anova SSCP Matrix for Loc
                       E = Error SSCP Matrix

                         S=2    M=0    N=3

          Statistic                       Value    P-Value

          Wilks' Lambda              0.60143661     0.6032
          Pillai's Trace             0.44702843     0.5521
          Hotelling-Lawley Trace     0.58210348     0.6337
          Roy's Greatest Root        0.35530890     0.7641
```

Notice that the p-value for Roy's greatest root is substantially larger in the new table, and correspondingly more in line with the p-values for the other tests.

If you reference the underlying ODS output object for the table of multivariate statistics, it is important to note that its structure does not depend on the value of the MSTAT= specification. In particular, it always contains columns corresponding to both the default MSTAT=FAPPROX and the MSTAT=EXACT tests. Moreover, since the MSTAT=FAPPROX tests are relatively cheap to compute, the columns corresponding to them are always filled in, even though they are not displayed when you specify MSTAT=EXACT. On the other hand, for MSTAT=FAPPROX (which is the default), the column of exact p-values contains missing values, and is not displayed.

Comments on Interpreting Regression Statistics

In most applications, regression models are merely useful approximations. Reality is often so complicated that you cannot know what the true model is. You might have to choose a model more on the basis of what variables can be measured and what kinds of models can be estimated than on a rigorous theory that explains how the universe really works. However, even in cases where theory is lacking, a regression model can be an excellent predictor of the response if the model is carefully formulated from a large sample. The interpretation of statistics such as parameter estimates might nevertheless be highly problematic.

Statisticians usually use the word "prediction" in a technical sense. *Prediction* in this sense does not refer to "predicting the future" (statisticians call that *forecasting*) but rather to guessing the response from the values of the regressors in an observation taken under the same circumstances as the sample from which the regression equation was estimated. If you developed a regression model for predicting consumer preferences in 1977, it might not give very good predictions in 2007 no matter how well it did in 1977. If it is the future you want to predict, your model must include whatever relevant factors might change over time. If the process you are studying does in fact change over time, you must take observations at several, perhaps many, different times. Analysis of such data is the province of SAS/STAT procedures such as MIXED and GLIMMIX and SAS/ETS procedures such as AUTOREG and STATESPACE. See Chapter 38, “The GLIMMIX Procedure,” and Chapter 56, “The MIXED Procedure,” for more information about modeling serial correlation in longitudinal, repeated measures, or time series data with SAS/STAT mixed modeling procedures. See the *SAS/ETS User’s Guide* for more information about the AUTOREG and STATESPACE procedures.

The comments in the rest of this section are directed toward linear least squares regression. For more detailed discussions of the interpretation of regression statistics, see Darlington (1968), Mosteller and Tukey (1977), Weisberg (1985), and Younger (1979).

Interpreting Parameter Estimates from a Controlled Experiment

Parameter estimates are easiest to interpret in a controlled experiment in which the regressors are manipulated independently of each other. In a well-designed experiment, such as a randomized factorial design with replications in each cell, you can use lack-of-fit tests and estimates of the standard error of prediction to determine whether the model describes the experimental process with adequate precision. If so, a regression coefficient estimates the amount by which the mean response changes when the regressor is changed by one unit while all the other regressors are unchanged. However, if the model involves interactions or polynomial terms, it might not be possible to interpret individual regression coefficients. For example, if the equation includes both linear and quadratic terms for a given variable, you cannot physically change the value of the linear term without also changing the value of the quadratic term. Sometimes it might be possible to recode the regressors, such as by using orthogonal polynomials, to simplify the interpretation.

If the nonstatistical aspects of the experiment are also treated with sufficient care (such as the use of placebos and double blinds), then you can state conclusions in causal terms; that is, this change in a regressor causes that change in the response. Causality can never be inferred from statistical results alone or from an observational study.

If the model you fit is not the true model, then the parameter estimates can depend strongly on the particular values of the regressors used in the experiment. For example, if the response is actually a quadratic function of a regressor but you fit a linear function, the estimated slope can be a large negative value if you use only small values of the regressor, a large positive value if you use only large values of the regressor, or near zero if you use both large and small regressor values. When you report the results of an experiment, it is important to include the values of the regressors. It is also important to avoid extrapolating the regression equation outside the range of regressors in the sample.

Interpreting Parameter Estimates from an Observational Study

In an observational study, parameter estimates can be interpreted as the expected difference in response of two observations that differ by one unit on the regressor in question and that have the same values for all other regressors. You cannot make inferences about “changes” in an observational study since you have not actually changed anything. It might not be possible even in principle to change one regressor independently of all the others. Neither can you draw conclusions about causality without experimental manipulation.

If you conduct an observational study and you do not know the true form of the model, interpretation of parameter estimates becomes even more convoluted. A coefficient must then be interpreted as an average over the sampled population of expected differences in response of observations that differ by one unit on only one regressor. The considerations that are discussed under controlled experiments for which the true model is not known also apply.

Comparing Parameter Estimates

Two coefficients in the same model can be directly compared only if the regressors are measured in the same units. You can make any coefficient large or small just by changing the units. If you convert a regressor from feet to miles, the parameter estimate is multiplied by 5280.

Sometimes standardized regression coefficients are used to compare the effects of regressors measured in different units. Standardized estimates are defined as the estimates that result when all variables are standardized to a mean of 0 and a variance of 1. Standardized estimates are computed by multiplying the original estimates by the sample standard deviation of the regressor variable and dividing by the sample standard deviation of the dependent variable.

Standardizing the variables effectively makes the standard deviation the unit of measurement. This makes sense only if the standard deviation is a meaningful quantity, which usually is the case only if the observations are sampled from a well-defined population. In a controlled experiment, the standard deviation of a regressor depends on the values of the regressor selected by the experimenter. Thus, you can make a standardized regression coefficient large by using a large range of values for the regressor.

In some applications you might be able to compare regression coefficients in terms of the practical range of variation of a regressor. Suppose that each independent variable in an industrial process can be set to values only within a certain range. You can rescale the variables so that the smallest possible value is zero and the largest possible value is one. Then the unit of measurement for each regressor is the maximum possible range of the regressor, and the parameter estimates are comparable in that sense. Another possibility is to scale the regressors in terms of the cost of setting a regressor to a particular value, so comparisons can be made in monetary terms.

Correlated Regressors

In an experiment, you can often select values for the regressors such that the regressors are orthogonal (not correlated with each other). Orthogonal designs have enormous advantages in interpretation. With orthogonal regressors, the parameter estimate for a given regressor does not depend on which other regressors are included in the model, although other statistics such as standard errors and p-values might change.

If the regressors are correlated, it becomes difficult to disentangle the effects of one regressor from another, and the parameter estimates can be highly dependent on which regressors are used in the model. Two correlated regressors might be nonsignificant when tested separately but highly significant when considered together. If two regressors have a correlation of 1.0, it is impossible to separate their effects.

It might be possible to recode correlated regressors to make interpretation easier. For example, if X and Y are highly correlated, they could be replaced in a linear regression by $X + Y$ and $X - Y$ without changing the fit of the model or statistics for other regressors.

Errors in the Regressors

If there is error in the measurements of the regressors, the parameter estimates must be interpreted with respect to the measured values of the regressors, not the true values. A regressor might be statistically nonsignificant when measured with error even though it would have been highly significant if measured accurately.

Probability Values (*p*-Values)

Probability values (p-values) do not necessarily measure the importance of a regressor. An important regressor can have a large (nonsignificant) p-value if the sample is small, if the regressor is measured over a narrow range, if there are large measurement errors, or if another closely related regressor is included in the equation. An unimportant regressor can have a very small p-value in a large sample. Computing a confidence interval for a parameter estimate gives you more useful information than just looking at the p-value, but confidence intervals do not solve problems of measurement errors in the regressors or highly correlated regressors.

Interpreting R^2

R^2 is usually defined as the proportion of variance of the response that is predictable from (can be explained by) the regressor variables. It might be easier to interpret $\sqrt{1 - R^2}$, which is approximately the factor by which the standard error of prediction is reduced by the introduction of the regressor variables.

R^2 is easiest to interpret when the observations, including the values of both the regressors and response, are randomly sampled from a well-defined population. Nonrandom sampling can greatly distort R^2. For example, excessively large values of R^2 can be obtained by omitting from the sample observations with regressor values near the mean.

In a controlled experiment, R^2 depends on the values chosen for the regressors. A wide range of regressor values generally yields a larger R^2 than a narrow range. In comparing the results of two experiments on the same variables but with different ranges for the regressors, you should look at the standard error of prediction (root mean square error) rather than R^2.

Whether a given R^2 value is considered to be large or small depends on the context of the particular study. A social scientist might consider an R^2 of 0.30 to be large, while a physicist might consider 0.98 to be small.

You can always get an R^2 arbitrarily close to 1.0 by including a large number of completely unrelated regressors in the equation. If the number of regressors is close to the sample size, R^2 is very biased. In such cases, the adjusted R^2 and related statistics discussed by Darlington (1968) are less misleading.

If you fit many different models and choose the model with the largest R^2, all the statistics are biased and the p-values for the parameter estimates are not valid. Caution must be taken with the interpretation of R^2 for models with no intercept term. As a general rule, no-intercept models should be fit only when theoretical justification exists and the data appear to fit a no-intercept framework. The R^2 in those cases is measuring something different (see Kvalseth 1985).

Incorrect Data Values

All regression statistics can be seriously distorted by a single incorrect data value. A decimal point in the wrong place can completely change the parameter estimates, R^2, and other statistics. It is important to check your data for outliers and influential observations. Residual and influence diagnostics are particularly useful in this regard.

When a data point is declared as influential or as outlying as measured by a particular model diagnostic, this does not imply that the case should be excluded from the analysis. The label "outlier" does not have a negative connotation. It means that a data point is unusual with respect to the model at hand. If your data follow a strong curved trend and you fit a linear regression, then many data points might be labeled as outliers not because they are "bad" or incorrect data values, but because your model is not appropriate.

References

Allen, D. M. (1971), "Mean Square Error of Prediction as a Criterion for Selecting Variables," *Technometrics*, 13, 469–475.

Allen, D. M. and Cady, F. B. (1982), *Analyzing Experimental Data by Regression*, Belmont, CA: Lifetime Learning Publications.

Belsley, D. A., Kuh, E., and Welsch, R .E. (1980), *Regression Diagnostics*, New York: John Wiley & Sons.

Bock, R. D. (1975), *Multivariate Statistical Methods in Behavioral Research*, New York: McGraw-Hill.

Box, G. E. P. (1966), "The Use and Abuse of Regression," *Technometrics*, 8, 625–629.

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

Cook, R. D. (1977), "Detection of Influential Observations in Linear Regression," *Technometrics*, 19, 15–18.

Cook, R. D. (1979), "Influential Observations in Linear Regression," *Journal of the American Statistical Association*, 74, 169–174.

Daniel, C. and Wood, F. (1980), *Fitting Equations to Data*, Revised Edition, New York: John Wiley & Sons.

Darlington, R. B. (1968), "Multiple Regression in Psychological Research and Practice," *Psychological Bulletin*, 69, 161–182.

Davis, A. W. (1970), "Differential Equation of Hotelling's Generalized T^2," *Annals of Statistics*, 39, 815–832.

Davis, A. W. (1972), "On the Marginal Distributions of the Latent Roots of the Multivariate Beta Matrix," *Biometrika*, 43, 1664–1670.

Davis, A. W. (1979), "On the Differential Equation for Meijer $G^{0,0}_{p,p}$ Function, and Further Wilks's Likelihood Ratio Criterion," *Biometrika*, 66, 519–531.

Davis, A. W. (1980), "Further Tabulation of Hotelling's Generalized T^2," *Communications in Statistics, Part B*, 9, 321–336.

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

Durbin, J. and Watson, G. S. (1951), "Testing for Serial Correlation in Least Squares Regression," *Biometrika*, 37, 409–428.

Freund, R. J., Littell, R. C., and Spector P. C. (1991), *SAS System for Linear Models*, Cary, NC: SAS Institute Inc.

Freund, R. J. and Littell, R. C. (1986), *SAS System for Regression, 1986 Edition*, Cary, NC: SAS Institute Inc.

Goodnight, J. H. (1979), "A Tutorial on the SWEEP Operator," *The American Statistician*, 33, 149−158. (Also available as SAS Technical Report R-106, *The Sweep Operator: Its Importance in Statistical Computing*, Cary, NC: SAS Institute Inc.)

Hawkins, D. M. (1980), "A Note on Fitting a Regression with No Intercept Term," *The American Statistician*, 34, 233.

Hosmer, D. W., Jr. and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: John Wiley & Sons.

Huber, P. J. (1973), "Robust Regression: Asymptotics, Conjectures and Monte Carlo," *Annals of Statistics*, 1, 799-821.

Johnston, J. (1972), *Econometric Methods*, New York: McGraw-Hill.

Kennedy, W. J. and Gentle, J. E. (1980), *Statistical Computing*, New York: Marcel Dekker.

Kvalseth, T. O. (1985), "Cautionary Note about R^2," *The American Statistician*, 39, 279–285.

Lee, Y. (1972), "Some Results on the Distribution of Wilk's Likelihood Ratio Criterion," *Biometrika*, 95, 649.

Mallows, C. L. (1973), "Some Comments on C_p," *Technometrics*, 15, 661–75.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), *Multivariate Analysis*, London: Academic Press.

Morrison, D. F. (1976), *Multivariate Statistical Methods*, Second Edition, New York: McGraw-Hill.

Mosteller, F. and Tukey, J. W. (1977), *Data Analysis and Regression*, Reading, MA: Addison-Wesley.

Muller, K. (1998), "A New F Approximation for the Pillai-Bartlett Trace Under H0," *Journal of Computational and Graphical Statistics*, 7, 131–137.

Neter, J. and Wasserman, W. (1974), *Applied Linear Statistical Models*, Homewood, IL: Irwin.

Pillai, K. C. S. (1960), *Statistical Table for Tests of Multivariate Hypotheses*, Manila: The Statistical Center, University of Philippines.

Pillai, K. C. S. and Flury, B. N. (1984), "Percentage Points of the Largest Characteristic Root of the Multivariate Beta Matrix," *Communications in Statistics, Part A*, 13, 2199–2237.

Pindyck, R. S. and Rubinfeld, D. L. (1981), *Econometric Models and Econometric Forecasts*, Second Edition, New York: McGraw-Hill.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications*, Second Edition, New York: John Wiley & Sons.

Rawlings, J. O. (1988), *Applied Regression Analysis: A Research Tool*, Pacific Grove, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

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

Rousseeuw, P. J. and Yohai, V. (1984), "Robust Regression by Means of S Estimators," in *Robust and Nonlinear Time Series Analysis*, ed. J. Franke, W. Härdle, and R. D. Martin, Lecture Notes in Statistics, 26, New York: Springer-Verlag, 256–274.

Timm, N. H. (1975), *Multivariate Analysis with Applications in Education and Psychology*, Monterey, CA: Brooks-Cole.

Weisberg, S. (1985), *Applied Linear Regression*, Second Edition. New York: John Wiley & Sons.

Yohai V. J. (1987), "High Breakdown Point and High Efficiency Robust Estimates for Regression," *Annals of Statistics*, 15, 642–656.

Younger, M. S. (1979), *Handbook for Linear Regression*, North Scituate, MA: Duxbury Press.

Index

binary data
  Introduction to Regression, 89

coefficient of determination
  definition (Introduction to Regression), 98, 110

estimability
  definition (Introduction to Regression), 102
estimable function
  definition (Introduction to Regression), 102
exponential family
  Introduction to Regression, 88

function
  estimable, definition (Introduction to Regression), 102

general linear model
  Introduction to Regression, 77
generalized estimating equations
  Introduction to Regression, 76, 89, 90
generalized linear mixed model
  Introduction to Regression, 77, 89
generalized linear model
  Introduction to Regression, 76, 88, 89, 93

independent variable
  Introduction to Regression, 79
Introduction to Regression
  adj. R-square selection, 87
  adjusted R-square, 98
  assumptions, 87, 95
  backward elimination, 86
  Bayesian analysis, 89, 90
  binary data, 89
  breakdown value, 94
  canonical correlation, 78, 94
  coefficient of determination, 98, 110
  collinearity, 109
  collinearity diagnostics, 85
  conditional logistic, 90
  confidence interval, 101
  conjoint analysis, 78, 94
  contingency table, 76
  controlled experiment, 108
  correlation matrix, 97
  covariance matrix, 97
  Cox model, 77
  Cp selection, 87
  cross validation, 78
  diagnostics, 77, 83
  dichotomous response, 78, 89
  errors-in-variable, 110
  estimable, 102
  estimate of precision, 97
  exact conditional logistic, 90
  exponential family, 88
  extreme value regression, 90
  failure-time data, 77
  forecasting, 101
  forward selection, 86
  function approximation, 92
  GEE, 76, 89, 90
  general linear model, 77
  generalized additive model, 76, 93
  generalized estimating equations, 76, 89, 90
  generalized least squares, 78
  generalized linear mixed model, 77, 89
  generalized linear model, 76, 88, 89, 93
  generalized logit, 89, 90
  Gentleman-Givens algorithm, 77, 90
  gompit regression, 90
  heterogeneous conditional distribution, 91
  homoscedasticity, 87
  Hotelling-Lawley trace, 104
  Huber M estimation, 78, 94
  hypothesis testing, 102
  ideal point preference mapping, 78, 94
  ill-conditioned data, 90
  independent variable, 79
  influence diagnostics, 85
  interactive procedures, 87, 94
  intercept, 98
  inverse link function, 88
  lack of fit, 85, 88
  least trimmed squares, 94
  levelization, 87
  leverage, 83, 100
  linear mixed model, 77
  linear regression, 78
  link function, 88
  local regression, 77, 92, 93
  LOESS, 77, 93
  logistic regression, 76, 78, 89
  LTS estimation, 94
  M estimation, 78, 94

max R-square selection, 86
min R-square selection, 87
MM estimation, 94
model selection, 86
model selection, adj. R-square, 87
model selection, backward, 86
model selection, Cp, 87
model selection, forward, 86
model selection, max R-square, 86
model selection, min R-square, 87
model selection, R-square, 87
model selection, stepwise, 86
multivariate tests, 102, 104
nonlinear least squares, 77, 92
nonlinear mixed model, 77
nonlinear model, 77, 91
nonparametric, 76, 77, 92
normal equations, 96
observational study, 108
odds ratio, 89
orthogonal regressors, 109
outcome variable, 79
outlier detection, 78
partial least squares, 78, 88
penalized least squares, 93
Pillai's trace, 104
Poisson regression, 76
polychotomous response, 78, 89
polynomial model, 77
predicted value, 100
prediction interval, 101
predictor variable, 79
principal component regression, 78
probit regression, 78, 90
proportional hazard, 77
proportional odds model, 90
quantal response, 90
quantile regression, 78, 91
R-square, 98, 110
R-square selection, 87
R-square, adjusted, 98
raw residual, 83, 100
reduced rank regression, 78
redundancy analysis, 78, 94
regressor variable, 79
residual, 100
residual plot, 82
residual variance, 97
residual, raw, 100
residual, studentized, 100, 101
response surface regression, 78, 88
response variable, 79
ridge regression, 88
robust regression, 78, 93
Roy's maximum root, 104
S estimation, 94
semiparametric model, 93
spline basis function, 93
spline transformation, 78, 94
standard error of prediction, 100
standard error, estimated, 97
statistical graphics, 85
stepwise selection, 86
stratification, 90
studentized residual, 83, 100, 101
success probability, 89
survey data, 78, 90
survival analysis, 77
survival data, 77
time-series diagnostics, 85
transformation, 78, 94
Type I sum of squares, 97, 102
Type II sum of squares, 97, 102
variable selection, 86
variance inflation, 98
Wilk's Lambda, 103

least squares
correlation matrix (Introduction to Regression), 97
covariance matrix (Introduction to Regression), 97
estimator (Introduction to Regression), 96
nonlinear (Introduction to Regression), 77, 92
normal equations (Introduction to Regression), 96
ordinary (Introduction to Regression), 88
partial (Introduction to Regression), 78, 88
penalized (Introduction to Regression), 93
levelization
Introduction to Regression, 87
linear mixed model
Introduction to Regression, 77
link function
cumulative (Introduction to Regression), 89
Introduction to Regression, 88
inverse (Introduction to Regression), 88
LOESS
Introduction to Regression, 77, 93
logistic
diagnostics (Introduction to Regression), 77
regression (Introduction to Regression), 76, 78, 89
regression, diagnostics (Introduction to Regression), 77
regression, ordinal (Introduction to Regression), 78

regression, survey data (Introduction to Regression), 78, 90
logistic regression
Introduction to Regression, 76, 89

matrix
correlation (Introduction to Regression), 97
covariance (Introduction to Regression), 97
diagonal (Introduction to Regression), 96

nonlinear model
Introduction to Regression, 77, 91

odds ratio
Introduction to Regression, 89

Poisson regression
Introduction to Regression, 76
polynomial model
Introduction to Regression, 77
proportional hazard
Introduction to Regression, 77

R-square
definition (Introduction to Regression), 98, 110
regression
adj. R-square selection (Introduction to Regression), 87
adjusted R-square (Introduction to Regression), 98
assumptions (Introduction to Regression), 87, 95
backward elimination (Introduction to Regression), 86
Bayesian analysis (Introduction to Regression), 89, 90
breakdown value (Introduction to Regression), 94
canonical correlation (Introduction to Regression), 78, 94
collinearity (Introduction to Regression), 109
collinearity diagnostics (Introduction to Regression), 85
conditional logistic (Introduction to Regression), 90
confidence interval (Introduction to Regression), 101
conjoint analysis (Introduction to Regression), 78, 94
contingency table (Introduction to Regression), 76
controlled experiment (Introduction to Regression), 108
Cook's D (Introduction to Regression), 83
correlation matrix (Introduction to Regression), 97
covariance matrix (Introduction to Regression), 97
Cox model (Introduction to Regression), 77
Cp selection (Introduction to Regression), 87
diagnostics (Introduction to Regression), 77, 83
diagnostics, collinearity (Introduction to Regression), 85
diagnostics, influence (Introduction to Regression), 85
diagnostics, logistic (Introduction to Regression), 77
errors-in-variable (Introduction to Regression), 110
estimate of precision (Introduction to Regression), 97
exact conditional logistic (Introduction to Regression), 90
failure-time data (Introduction to Regression), 77
forecasting (Introduction to Regression), 101
forward selection (Introduction to Regression), 86
function approximation (Introduction to Regression), 92
GEE (Introduction to Regression), 76, 89, 90
general linear model (Introduction to Regression), 77
generalized additive model (Introduction to Regression), 76, 93
generalized estimating equations (Introduction to Regression), 76, 89, 90
generalized least squares (Introduction to Regression), 78
generalized linear mixed model (Introduction to Regression), 77, 89
generalized linear model (Introduction to Regression), 76, 88, 89, 93
generalized logit (Introduction to Regression), 89, 90
Gentleman-Givens algorithm (Introduction to Regression), 77, 90
gompit (Introduction to Regression), 90
heterogeneous conditional distribution (Introduction to Regression), 91
homoscedasticity (Introduction to Regression), 87
Hotelling Lawley trace (Introduction to Regression), 104

ideal point preference mapping (Introduction to Regression), 78, 94
ill-conditioned data (Introduction to Regression), 90
influence diagnostics (Introduction to Regression), 85
intercept (Introduction to Regression), 98
lack-of-fit (Introduction to Regression), 85, 88
least trimmed squares (Introduction to Regression), 94
leverage (Introduction to Regression), 83, 100
linear (Introduction to Regression), 78
linear mixed model (Introduction to Regression), 77
linear, survey data (Introduction to Regression), 78
local (Introduction to Regression), 77, 92, 93
logistic (Introduction to Regression), 76, 78
logistic, conditional (Introduction to Regression), 90
logistic, exact conditional (Introduction to Regression), 90
LTS estimation (Introduction to Regression), 94
M estimation (Introduction to Regression), 78, 94
max R-square selection (Introduction to Regression), 86
min R-square selection (Introduction to Regression), 87
MN estimation (Introduction to Regression), 94
model selection, adj. R-square (Introduction to Regression), 87
model selection, backward (Introduction to Regression), 86
model selection, Cp (Introduction to Regression), 87
model selection, forward (Introduction to Regression), 86
model selection, max R-square (Introduction to Regression), 86
model selection, min R-square (Introduction to Regression), 87
model selection, R-square (Introduction to Regression), 87
model selection, stepwise (Introduction to Regression), 86
multivariate tests (Introduction to Regression), 102, 104
nonlinear (Introduction to Regression), 77, 91
nonlinear least squares (Introduction to Regression), 77, 92
nonlinear mixed model (Introduction to Regression), 77
nonparametric (Introduction to Regression), 76, 92
normal equations (Introduction to Regression), 96
observational study (Introduction to Regression), 108
orthogonal regressors (Introduction to Regression), 109
partial least squares (Introduction to Regression), 78
Pillai's trace (Introduction to Regression), 104
Poisson (Introduction to Regression), 76
polynomial (Introduction to Regression), 77
precision, estimate (Introduction to Regression), 97
predicted value (Introduction to Regression), 100
prediction interval (Introduction to Regression), 101
principal components (Introduction to Regression), 78
probit (Introduction to Regression), 78, 90
proportional hazard (Introduction to Regression), 77
proportional odds model (Introduction to Regression), 90
quantal (Introduction to Regression), 90
quantile (Introduction to Regression), 78, 91
R-square (Introduction to Regression), 98, 110
R-square selection (Introduction to Regression), 87
R-square, adjusted (Introduction to Regression), 98
raw residual (Introduction to Regression), 100
redundancy analysis (Introduction to Regression), 78, 94
regressor variable (Introduction to Regression), 79
residual (Introduction to Regression), 100
residual plot (Introduction to Regression), 82
residual variance (Introduction to Regression), 97
residual, raw (Introduction to Regression), 100
residual, studentized (Introduction to Regression), 100

response surface (Introduction to Regression), 78, 88
ridge (Introduction to Regression), 88
robust (Introduction to Regression), 78, 93
Roy's maximum root (Introduction to Regression), 104
S estimation (Introduction to Regression), 94
semiparametric model (Introduction to Regression), 93
spline (Introduction to Regression), 93
spline transformation (Introduction to Regression), 78, 94
spline, basis function(Introduction to Regression), 93
standard error of prediction (Introduction to Regression), 100
standard error, estimated (Introduction to Regression), 97
stepwise selection (Introduction to Regression), 86
stratification (Introduction to Regression), 90
studentized residual (Introduction to Regression), 100
sum of squares, Type I (Introduction to Regression), 97, 102
sum of squares, Type II (Introduction to Regression), 97, 102
surface (Introduction to Regression), 77
survey data (Introduction to Regression), 78, 90
survival data (Introduction to Regression), 77
testing hypotheses (Introduction to Regression), 102
transformation (Introduction to Regression), 78, 94
Type I sum of squares (Introduction to Regression), 97, 102
Type II sum of squares (Introduction to Regression), 97, 102
variance inflation (Introduction to Regression), 98
Wilk's Lambda (Introduction to Regression), 103
regressor variable
Introduction to Regression, 79
residual
Cook's D (Introduction to Regression), 83
raw (Introduction to Regression), 83
studentized (Introduction to Regression), 83
residuals
raw (Introduction to Regression), 100
studentized (Introduction to Regression), 100, 101

survival analysis
Introduction to Regression, 77
survival data
Introduction to Regression, 77

Your Turn

We welcome your feedback.

- If you have comments about this book, please send them to **yourturn@sas.com**. Include the full title and page numbers (if applicable).
- If you have comments about the software, please send them to **suggest@sas.com**.