SAS/STAT® 14.1 User's Guide
Introduction to Regression Procedures

This document is an individual chapter from *SAS/STAT® 14.1 User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2015. *SAS/STAT® 14.1 User's Guide*. Cary, NC: SAS Institute Inc.

SAS/STAT® 14.1 User's Guide

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

July 2015

Chapter 4
Introduction to Regression Procedures

Contents

Overview: Regression Procedures

This chapter provides an overview of SAS/STAT procedures that perform regression analysis. The REG procedure provides extensive capabilities for fitting linear regression models that involve individual numeric independent variables. Many other procedures can also fit regression models, but they focus on more specialized forms of regression, such as robust regression, generalized linear regression, nonlinear regression, nonparametric regression, quantile regression, regression modeling of survey data, regression modeling of survival data, and regression modeling of transformed variables. The SAS/STAT procedures that can fit regression models include the ADAPTIVEREG, CATMOD, GAM, GENMOD, GLIMMIX, GLM, GLMSELECT, LIFEREG, LOESS, LOGISTIC, MIXED, NLIN, NLMIXED, ORTHOREG, PHREG, PLS, PROBIT, QUANTREG, QUANTSELECT, REG, ROBUSTREG, RSREG, SURVEYLOGISTIC, SURVEYPHREG, SURVEYREG, TPSPLINE, and TRANSREG procedures. Several procedures in SAS/ETS software also fit regression models.

Introduction

In a linear regression model, the mean of a response variable **Y** is a function of 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, normally or nonnormally 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. Some models contain classification (or CLASS) variables that enter the model not through their values but through their levels. For an introduction to linear regression models, see Chapter 3, "Introduction to Statistical Modeling with SAS/STAT Software." For information that is common to many of the regression procedures, see Chapter 19, "Shared Concepts and Topics." The following procedures, listed in alphabetical order, perform at least one type of regression analysis.

ADAPTIVEREG — fits multivariate adaptive regression spline models. This is a nonparametric regression technique that combines both regression splines and model selection methods. PROC ADAPTIVEREG produces parsimonious models that do not overfit the data and thus have good predictive power. PROC ADAPTIVEREG supports CLASS variables. For more information, see Chapter 25, "The ADAPTIVEREG Procedure."

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. PROC CATMOD supports CLASS variables. For more information, see Chapter 8, "Introduction to Categorical Data Analysis Procedures," and Chapter 32, "The CATMOD Procedure."

GAM — fits generalized additive models. Generalized additive models are nonparametric in that the usual assumption of linear predictors is relaxed. Generalized additive models consist of additive, smooth functions of the regression variables. PROC GAM can fit additive models to nonnormal data. PROC GAM supports CLASS variables. For more information, see Chapter 41, "The GAM Procedure."

GENMOD — fits generalized linear models. PROC GENMOD is especially suited for responses that have discrete outcomes, and it performs logistic regression and Poisson regres-

	sion in addition to fitting generalized estimating equations for repeated measures data. PROC GENMOD supports CLASS variables and provides Bayesian analysis capabilities. For more information, see Chapter 8, "Introduction to Categorical Data Analysis Procedures," and Chapter 44, "The GENMOD Procedure."
GLIMMIX	uses likelihood-based methods to fit generalized linear mixed models. PROC GLIMMIX can perform simple, multiple, polynomial, and weighted regression, in addition to many other analyses. PROC GLIMMIX can fit linear mixed models, which have random effects, and models that do not have random effects. PROC GLIMMIX supports CLASS variables. For more information, see Chapter 45, "The GLIMMIX Procedure."
GLM	uses the method of least squares to fit general linear models. PROC GLM can perform simple, multiple, polynomial, and weighted regression in addition to many other analyses. 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. PROC GLM supports CLASS variables. For more information, see Chapter 5, "Introduction to Analysis of Variance Procedures," and Chapter 46, "The GLM Procedure."
GLMSELECT	performs variable selection in the framework of general linear models. PROC GLMSELECT supports CLASS variables (like PROC GLM) and model selection (like PROC REG). A variety of model selection methods are available, including forward, backward, stepwise, LASSO, and least angle regression. PROC GLMSELECT provides a variety of selection and stopping criteria. For more information, see Chapter 49, "The GLMSELECT Procedure."
LIFEREG	fits parametric models to failure-time data that might be right-censored. These types of models are commonly used in survival analysis. PROC LIFEREG supports CLASS variables and provides Bayesian analysis capabilities. For more information, see Chapter 13, "Introduction to Survival Analysis Procedures," and Chapter 69, "The LIFEREG Procedure."
LOESS	uses a local regression method to fit nonparametric models. PROC LOESS is suitable for modeling regression surfaces in which the underlying parametric form is unknown and for which robustness in the presence of outliers is required. For more information, see Chapter 71, "The LOESS Procedure."
LOGISTIC	fits logistic models for binomial and ordinal outcomes. PROC LOGISTIC provides a wide variety of model selection methods and computes numerous regression diagnostics. PROC LOGISTIC supports CLASS variables. For more information, see Chapter 8, "Introduction to Categorical Data Analysis Procedures," and Chapter 72, "The LOGISTIC Procedure."
MIXED	uses likelihood-based techniques to fit linear mixed models. PROC MIXED can perform simple, multiple, polynomial, and weighted regression, in addition to many other analyses. PROC MIXED can fit linear mixed models, which have random effects, and models that do not have random effects. PROC MIXED supports CLASS variables. For more information, see Chapter 77, "The MIXED Procedure."
NLIN	uses the method of nonlinear least squares to fit general nonlinear regression models. Several different iterative methods are available. For more information, see Chapter 81, "The NLIN Procedure."

NLMIXED	uses the method of maximum likelihood to fit general nonlinear mixed regression models. PROC NLMIXED enables you to specify a custom objective function for parameter estimation and to fit models with or without random effects. For more information, see Chapter 82, "The NLMIXED Procedure."
ORTHOREG	uses the Gentleman-Givens computational method to perform regression. For ill-conditioned data, PROC ORTHOREG can produce more-accurate parameter estimates than procedures such as PROC GLM and PROC REG. PROC ORTHOREG supports CLASS variables. For more information, see Chapter 84, "The ORTHOREG Procedure."
PHREG	fits Cox proportional hazards regression models to survival data. PROC PHREG supports CLASS variables and provides Bayesian analysis capabilities. For more information, see Chapter 13, "Introduction to Survival Analysis Procedures," and Chapter 85, "The PHREG Procedure."
PLS	performs partial least squares regression, principal component regression, and reduced rank regression, along with cross validation for the number of components. PROC PLS supports CLASS variables. For more information, see Chapter 88, "The PLS Procedure."
PROBIT	performs probit regression in addition to logistic regression and ordinal logistic regression. PROC PROBIT is useful when the dependent variable is either dichotomous or polychotomous and the independent variables are continuous. PROC PROBIT supports CLASS variables. For more information, see Chapter 93, "The PROBIT Procedure."
QUANTREG	uses quantile regression to model the effects of covariates on the conditional quantiles of a response variable. PROC QUANTREG supports CLASS variables. For more information, see Chapter 95, "The QUANTREG Procedure."
QUANTSELECT	provides variable selection for quantile regression models. Selection methods include forward, backward, stepwise, and LASSO. The procedure provides a variety of selection and stopping criteria. PROC QUANTSELECT supports CLASS variables. For more information, see Chapter 96, "The QUANTSELECT Procedure."
REG	performs linear regression with many diagnostic capabilities. PROC REG produces fit, residual, and diagnostic plots; heat maps; and many other types of graphs. PROC REG enables you to select models by using any one of nine methods, and you can interactively change both the regression model and the data that are used to fit the model. For more information, see Chapter 97, "The REG Procedure."
ROBUSTREG	uses Huber M estimation and high breakdown value estimation to perform robust regression. PROC ROBUSTREG is suitable for detecting outliers and providing resistant (stable) results in the presence of outliers. PROC ROBUSTREG supports CLASS variables. For more information, see Chapter 98, "The ROBUSTREG Procedure."
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. For more information, see Chapter 99, "The RSREG Procedure."
SURVEYLOGISTIC	uses the method of maximum likelihood to fit logistic models for binary and ordinal outcomes to survey data. PROC SURVEYLOGISTIC supports CLASS variables.

For more information, see Chapter 14, "Introduction to Survey Procedures," and Chapter 111, "The SURVEYLOGISTIC Procedure."

SURVEYPHREG — fits proportional hazards models for survey data by maximizing a partial pseudo-likelihood function that incorporates the sampling weights. The SURVEYPHREG procedure provides design-based variance estimates, confidence intervals, and tests for the estimated proportional hazards regression coefficients. PROC SURVEYPHREG supports CLASS variables. For more information, see Chapter 14, "Introduction to Survey Procedures," Chapter 13, "Introduction to Survival Analysis Procedures," and Chapter 113, "The SURVEYPHREG Procedure."

SURVEYREG — uses elementwise regression to fit linear regression models to survey data by generalized least squares. PROC SURVEYREG supports CLASS variables. For more information, see Chapter 14, "Introduction to Survey Procedures," and Chapter 114, "The SURVEYREG Procedure."

TPSPLINE — uses penalized least squares to fit nonparametric regression models. PROC TPSPLINE makes no assumptions of a parametric form for the model. For more information, see Chapter 116, "The TPSPLINE Procedure."

TRANSREG — fits univariate and multivariate linear models, optionally with spline, Box-Cox, and other nonlinear transformations. Models include regression and ANOVA, conjoint analysis, preference mapping, redundancy analysis, canonical correlation, and penalized B-spline regression. PROC TRANSREG supports CLASS variables. For more information, see Chapter 117, "The TRANSREG Procedure."

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

ARIMA — uses autoregressive moving-average errors to perform multiple regression analysis. For more information, see Chapter 8, "The ARIMA Procedure" (*SAS/ETS User's Guide*).

AUTOREG — implements regression models that use time series data in which the errors are autocorrelated. For more information, see Chapter 9, "The AUTOREG Procedure" (*SAS/ETS User's Guide*).

COUNTREG — analyzes regression models in which the dependent variable takes nonnegative integer or count values. For more information, see Chapter 12, "The COUNTREG Procedure" (*SAS/ETS User's Guide*).

MDC — fits conditional logit, mixed logit, heteroscedastic extreme value, nested logit, and multinomial probit models to discrete choice data. For more information, see Chapter 25, "The MDC Procedure" (*SAS/ETS User's Guide*).

MODEL — handles nonlinear simultaneous systems of equations, such as econometric models. For more information, see Chapter 26, "The MODEL Procedure" (*SAS/ETS User's Guide*).

PANEL — analyzes a class of linear econometric models that commonly arise when time series and cross-sectional data are combined. For more information, see Chapter 27, "The PANEL Procedure" (*SAS/ETS User's Guide*).

PDLREG — fits polynomial distributed lag regression models. For more information, see Chapter 28, "The PDLREG Procedure" (*SAS/ETS User's Guide*).

QLIM	analyzes limited dependent variable models in which dependent variables take discrete values or are observed only in a limited range of values. For more information, see Chapter 29, "The QLIM Procedure" (*SAS/ETS User's Guide*).
SYSLIN	handles linear simultaneous systems of equations, such as econometric models. For more information, see Chapter 36, "The SYSLIN Procedure" (*SAS/ETS User's Guide*).
VARMAX	performs multiple regression analysis for multivariate time series dependent variables by using current and past vectors of dependent and independent variables as predictors, with vector autoregressive moving-average errors, and with modeling of time-varying heteroscedasticity. For more information, see Chapter 42, "The VARMAX Procedure" (*SAS/ETS User's Guide*).

Introductory Example: Linear Regression

Regression analysis models the relationship between a response or outcome variable and another set of variables. This 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 *i*th 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 called 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 the child's height. One application of a regression model that contains 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 that contains the response variable Weight and the regressor variable Height can be written as

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

where

Weight_i	is the response variable for the *i*th child
Height_i	is the regressor variable for the *i*th child
β_0, β_1	are the unknown regression parameters
ϵ_i	is the unobservable random error associated with the *i*th observation

The data set Sashelp.class, which is available in the Sashelp library, identifies the children and their observed heights (the variable Height) and weights (the variable Weight). The following statements perform the regression analysis:

```
ods graphics on;
proc reg data=sashelp.class;
   model Weight = Height;
run;
```

Figure 4.1 displays the default tabular output of PROC REG 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. For more information about the principle of least squares estimation and its role in linear model analysis, see the sections "Classical Estimation Principles" and "Linear Model Theory" in Chapter 3, "Introduction to Statistical Modeling with SAS/STAT Software." Also see an applied regression text such as Draper and Smith (1998); Daniel and Wood (1999); Johnston and DiNardo (1997); Weisberg (2005).

Figure 4.1 Regression for Weight and Height Data

The REG Procedure
Model: MODEL1
Dependent Variable: Weight

Number of Observations Read	19
Number of Observations Used	19

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 that relates height to weight is described by the equation

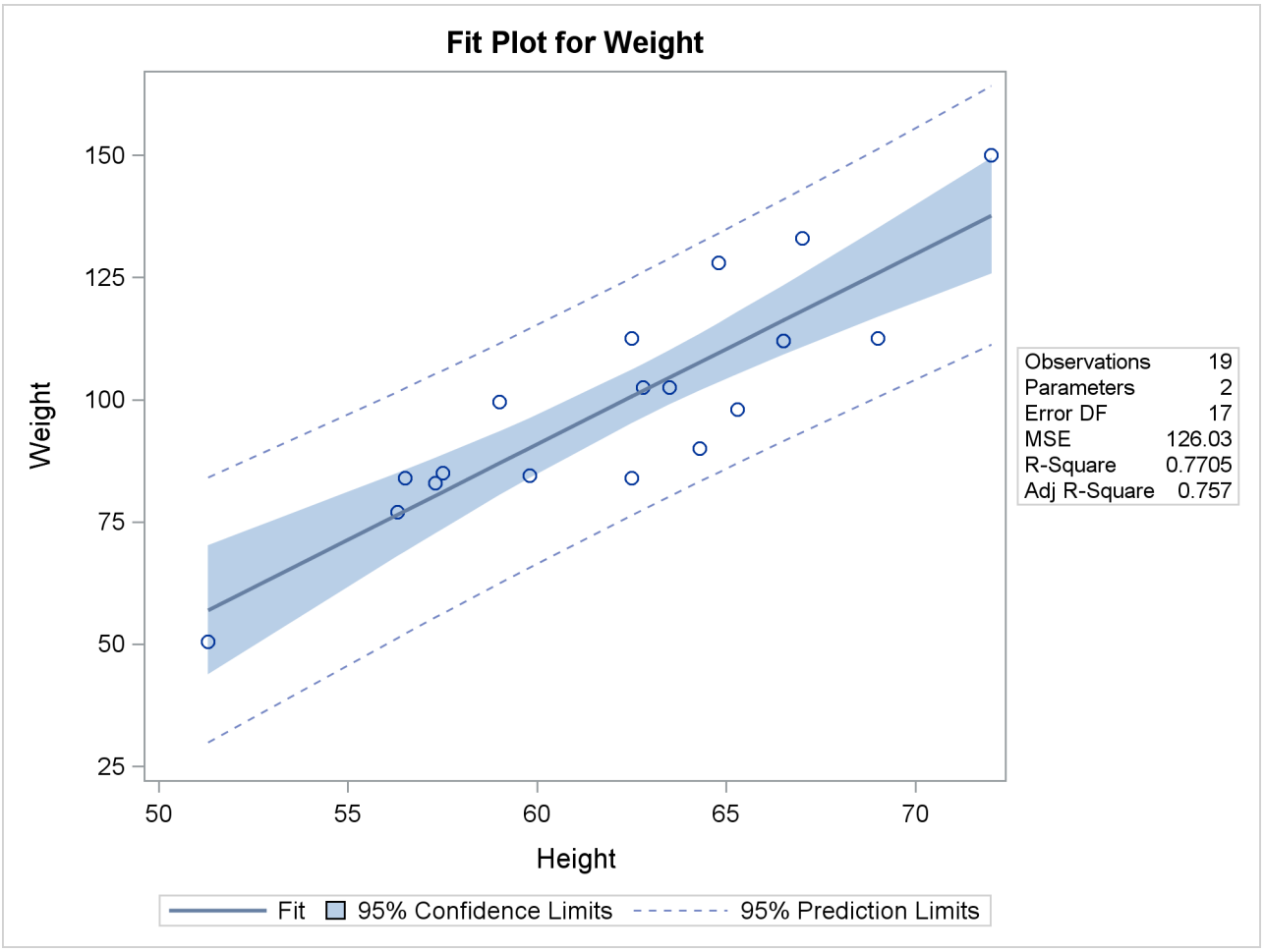
$$\widehat{\text{Weight}} = -143.02692 + 3.89903 \times \text{Height}$$

The "hat" notation is used to emphasize that $\widehat{\text{Weight}}$ is not one of the original observations but a value predicted under the regression model that has been fit to the data. In the least squares solution, the following residual sum of squares is minimized and the achieved criterion value is displayed in the analysis of variance table as the error sum of squares (2142.48772):

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

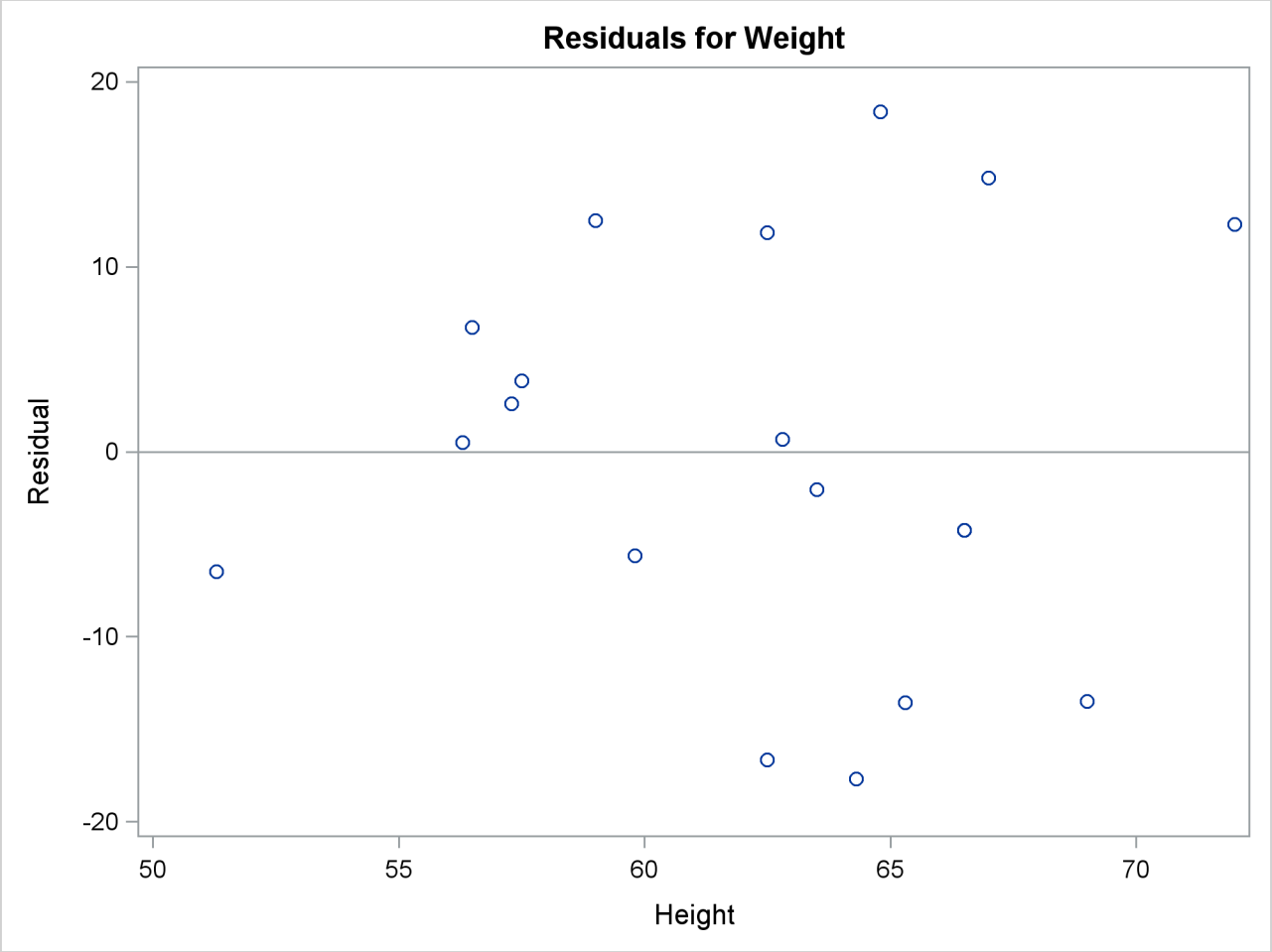
Figure 4.2 displays the fit plot that is produced by ODS Graphics. The fit plot shows the positive slope of the fitted line. The average weight of a child changes by $\widehat{\beta}_1 = 3.89903$ units for each unit change in height. The 95% confidence limits in the fit plot are pointwise limits that 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



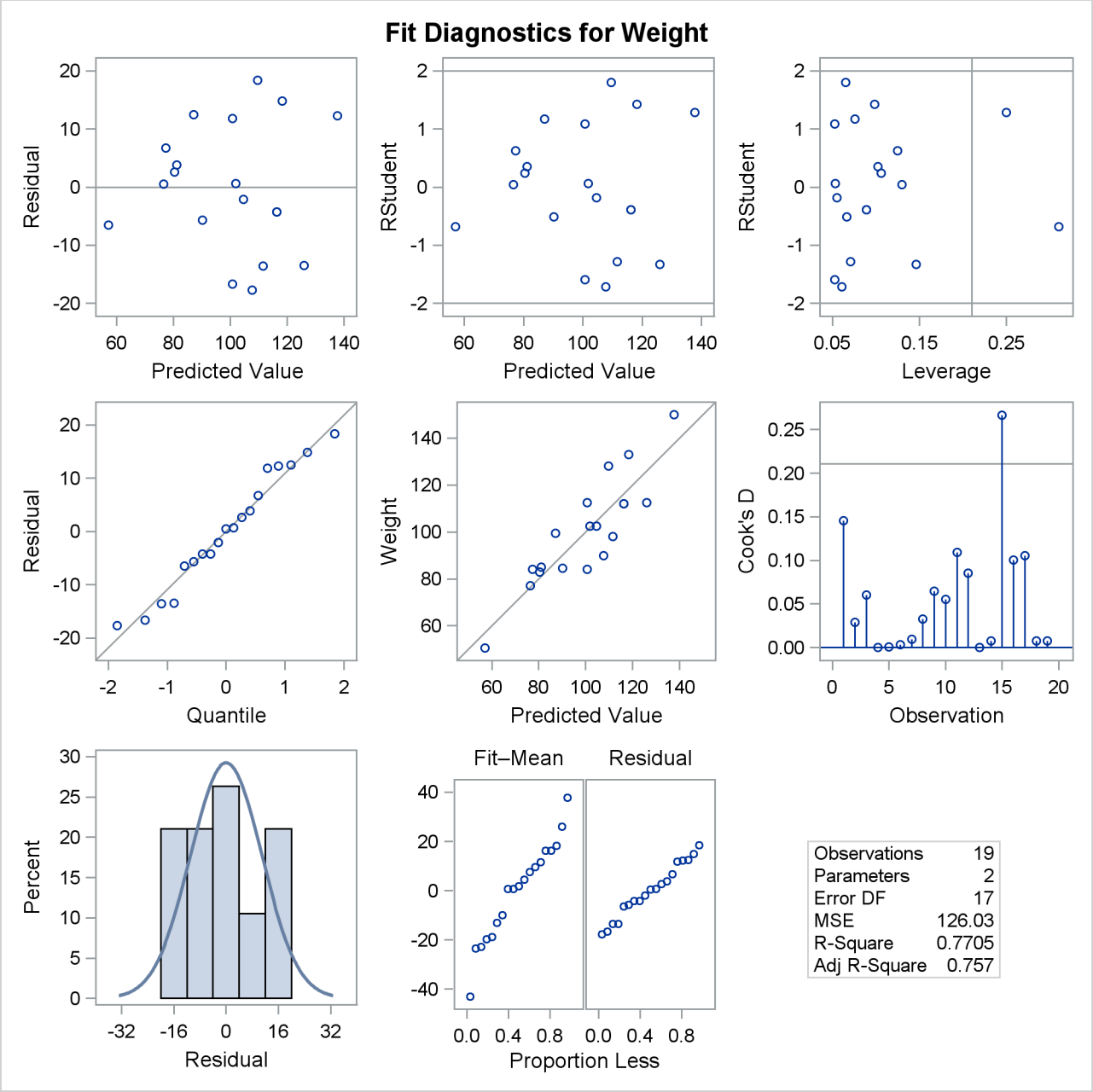
Regression is often used in an exploratory fashion to look for empirical relationships, such as the relationship between Height and Weight. In this example, Height is not the cause of Weight. You would need a controlled experiment to confirm the relationship scientifically. For more information, see the section "Comments on Interpreting Regression Statistics" on page 99. A separate question from whether there is a cause-and-effect relationship between the two variables that are involved in this regression is whether the simple linear regression model adequately describes the relationship among these data. If the SLR model makes the usual assumptions about the model errors ϵ_i, 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 $\text{Weight}_i - \widehat{\text{Weight}}_i$ should scatter around the zero reference line without discernible structure. The residual plot in Figure 4.3 confirms this.

Figure 4.3 Residual Plot for Regression of Weight on Height



The panel of regression diagnostics in Figure 4.4 provides an even more detailed look at the model-data agreement. 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, indicating that the observation for this child has a strong influence on the regression parameter estimates.

Figure 4.4 Panel of Regression Diagnostics



For more information about the interpretation of regression diagnostics and about ODS statistical graphics with PROC REG, see Chapter 97, "The REG Procedure."

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

- parameter estimates that are 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 relationship affects the stability and variance of the estimates (REG procedure)
- influence diagnostics to measure how each individual observation contributes to determining the parameter estimates, the SSE, and the fitted values (GENMOD, GLM, LOGISTIC, MIXED, NLIN, PHREG, REG, and RSREG procedures)
- 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 does not depend on the form of the model (CATMOD, LOGISTIC, PROBIT, and RSREG procedures)
- diagnostic plots that check the fit of the model (GLM, LOESS, PLS, REG, RSREG, and TPSPLINE procedures)
- predicted and residual values, and confidence intervals for the mean and for an individual value (GLIMMIX, GLM, LOESS, LOGISTIC, NLIN, PLS, REG, RSREG, TPSPLINE, and TRANSREG procedures)
- time series diagnostics for equally spaced time series data that measure how closely errors might be related across neighboring observations. These diagnostics can also measure functional goodness of fit for data that are sorted by regressor or response variables (REG and 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 strongly influence the analysis. Figure 4.2, Figure 4.3, and Figure 4.4, for example, show three of the ODS statistical graphs that are produced by PROC REG 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 Graphics" section in the "Details" section of the individual procedure documentation.

Model Selection Methods

Statistical model selection (or model building) involves forming a model from a set of regressor variables that fits the data well but without overfitting. Models are overfit when they contain too many unimportant regressor variables. Overfit models are too closely molded to a particular data set. As a result, overfit models have unstable regression coefficients and are quite likely to have poor predictive power. Guided, numerical variable selection methods offer one approach to building models in situations where many potential regressor variables are available for inclusion in a regression model.

Both the REG and GLMSELECT procedures provide extensive options for model selection in ordinary linear regression models.[1] PROC GLMSELECT provides the most modern and flexible options for model selection. PROC GLMSELECT provides more selection options and criteria than PROC REG, and PROC

[1] The QUANTSELECT, PHREG, and LOGISTIC procedures provide model selection for quantile, proportional hazards, and logistic regression, respectively.

GLMSELECT also supports CLASS variables. For more information about PROC GLMSELECT, see Chapter 49, "The GLMSELECT Procedure." For more information about PROC REG, see Chapter 97, "The REG Procedure."

SAS/STAT procedures provide the following model selection options for regression models:

NONE performs no model selection. This method uses the full model given in the MODEL statement to fit the model. This selection method is available in the GLMSELECT, LOGISTIC, PHREG, QUANTSELECT, and REG procedures.

FORWARD uses a forward-selection algorithm to select variables. This method starts with no variables in the model and adds variables one by one to the model. At each step, the variable that is 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. This selection method is available in the GLMSELECT, LOGISTIC, PHREG, QUANTSELECT, and REG procedures.

BACKWARD uses a backward-elimination algorithm to select variables. This method starts with a full model and eliminates variables one by one from the model. At each step, the variable that makes 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. This selection method is available in the GLMSELECT, LOGISTIC, PHREG, QUANTSELECT, and REG procedures.

STEPWISE uses a stepwise-regression algorithm to select variables; it combines the 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 entering the model and for remaining in the model. This selection method is available in the GLMSELECT, LOGISTIC, PHREG, QUANTSELECT, and REG procedures.

LASSO adds and deletes parameters by using a version of ordinary least squares in which the sum of the absolute regression coefficients is constrained. If the model contains CLASS variables, then these CLASS variables are split. This selection method is available in the GLMSELECT and QUANTSELECT procedures.

LAR uses least angle regression to select variables. Like forward selection, this method starts with no effects in the model and adds effects. The parameter estimates at any step are "shrunk" when compared to the corresponding least squares estimates. If the model contains CLASS variables, then these CLASS variables are split. This selection method is available in PROC GLMSELECT.

MAXR uses maximum R-square improvement to select models. 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 it evaluates many more models. 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 considers what the "best" remaining variable might accomplish. Consequently, model building based on the maximum R-square improvement usually takes much longer than stepwise model building. This selection method is available in PROC REG.

MINR	uses minimum R-square improvement to select models. This method closely resembles MAXR, but the switch that is chosen is the one that produces the smallest increase in R square. This selection method is available in PROC REG.
RSQUARE	selects a specified number of models that have the highest R square in each of a range of model sizes. This selection method is available in PROC REG.
CP	selects a specified number of models that have the lowest C_p within a range of model sizes. This selection method is available in PROC REG.
ADJRSQ	selects a specified number of models that have the highest adjusted R square within a range of model sizes. This selection method is available in PROC REG.
SCORE	selects models based on a branch-and-bound algorithm that finds a specified number of models that have the highest likelihood score for all possible model sizes, from one up to a model that contains all the explanatory effects. This selection method is available in PROC LOGISTIC and PROC PHREG.

Linear Regression: The REG Procedure

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

- handles simple and multiple regression models
- provides nine model selection methods
- allows interactive changes both in the model and in the data that are 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 fit, of data, and of various statistics
- uses data, correlations, or crossproducts for input

Regression with the REG and GLM Procedures

The REG and GLM procedures are closely related; they make the same assumptions about the basic model and use the same estimation principles. 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.

PROC GLM provides a CLASS statement for the levelization of classification variables; see the section "Parameterization of Model Effects" on page 391 in Chapter 19, "Shared Concepts and Topics," on the parameterization of classification variables in statistical models. In most cases, you should directly use PROC GLM, PROC GLMSELECT, or some other procedure when you fit models that have classification variables. However, you can fit models that have classification variables in PROC REG by first coding the classification variables by using PROC GLMSELECT, PROC TRANSREG, PROC GLMMOD, or some other method, and then including the coded variables in the MODEL statement in PROC REG.

Most of the statistics based on predicted and residual values that are available in PROC REG are also available in PROC GLM. However, PROC REG provides more diagnostic information. In addition, PROC GLM allows only one model and does not provide model selection.

Both PROC REG and PROC GLM are interactive, in that they do not stop after processing a RUN statement. Both procedures accept statements until a QUIT statement is submitted. For more information about interactive procedures, see the section "Interactive Features in the CATMOD, GLM, and REG Procedures" on page 87.

Model Selection: The GLMSELECT Procedure

PROC GLMSELECT performs model selection in the framework of general linear models. A variety of model selection methods are available, including forward, backward, stepwise, the LASSO method of Tibshirani (1996), and the related least angle regression method of Efron et al. (2004). The GLMSELECT procedure offers extensive capabilities for customizing the selection by providing a wide variety of selection and stopping criteria, including significance level–based and validation-based criteria. The procedure also provides graphical summaries of the selection process.

PROC GLMSELECT compares most closely with PROC REG and PROC GLM. PROC REG supports a variety of model selection methods but does not provide a CLASS statement. PROC GLM provides a CLASS statement but does not provide model selection methods. PROC GLMSELECT fills this gap. PROC GLMSELECT focuses on the standard general linear model for univariate responses with independently and identically distributed errors. PROC GLMSELECT provides results (tables, output data sets, and macro variables) that make it easy to explore the selected model in more detail in a subsequent procedure such as PROC REG or PROC GLM.

Response Surface Regression: The RSREG Procedure

PROC RSREG 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

PROC PLS 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 that are implemented in PROC PLS 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 the techniques that are implemented in PROC PLS 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 variation and predictor variation.

Generalized Linear Regression

As outlined in the section "Generalized Linear Models" on page 33 in Chapter 3, "Introduction to Statistical Modeling with SAS/STAT Software," the class of generalized linear models generalizes the linear regression models 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 $\mathrm{E}[Y] = g^{-1}(\eta)$; $g^{-1}(\cdot)$ 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 the 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 the linear predictor $\eta = \mathbf{x}'\boldsymbol{\beta}$ and the logit link function:

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

The inversely linked linear predictor function in this model is

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

The dichotomous logistic regression model can be extended to multinomial (polychotomous) data. Two classes of models for multinomial data can be fit by using 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 sections briefly discuss SAS/STAT procedures for logistic regression. For more information about the comparison of the procedures mentioned there with respect to analysis of categorical responses, see Chapter 8, "Introduction to Categorical Data Analysis Procedures."

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

Contingency Table Data: The CATMOD Procedure

PROC CATMOD fits models to data that are represented by a contingency table by using weighted least squares and a variety of link functions. Although PROC CATMOD also provides maximum likelihood estimation for logistic regression, PROC LOGISTIC is more efficient for most analyses.

Generalized Linear Models: The GENMOD Procedure

PROC GENMOD is a generalized linear modeling procedure that estimates parameters by maximum likelihood. It uses CLASS and MODEL statements 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, PROC GENMOD can solve generalized estimating equations (GEE) to model correlated data and can perform a Bayesian analysis.

Generalized Linear Mixed Models: The GLIMMIX Procedure

PROC GLIMMIX fits generalized linear mixed models. If the model does not contain random effects, PROC GLIMMIX fits generalized linear models by using the method of maximum likelihood. In the class of logistic regression models, PROC GLIMMIX can fit models to binary, binomial, ordinal, and nominal outcomes.

Logistic Regression: The LOGISTIC Procedure

PROC LOGISTIC fits logistic regression models and estimates parameters by maximum likelihood. The procedure fits the usual logistic regression model for binary data in addition to models with the cumulative link function for ordinal data (such as the proportional odds model) and the generalized logit model for nominal data. PROC LOGISTIC offers a number of variable selection methods and can perform conditional and exact conditional logistic regression analysis.

Discrete Event Data: The PROBIT Procedure

PROC 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. PROC PROBIT fits probit, logit, ordinal logistic, and extreme value (or gompit) regression models.

Correlated Data: The GENMOD and GLIMMIX Procedures

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

Both PROC GENMOD and PROC GLIMMIX can accommodate correlated observations, but they use different techniques. PROC GENMOD fits correlated data models by using generalized estimating equations that rely on a first- and second-moment specification for the response data and a working correlation assumption. With PROC GLIMMIX, you can model correlations between the observations either by specifying random effects in the conditional distribution that induce a marginal correlation structure or by direct modeling of the marginal dependence. PROC GLIMMIX uses likelihood-based techniques to estimate parameters.

PROC GENMOD supports a Bayesian analysis through its BAYES statement.

With PROC GLIMMIX, you can vary the distribution or link function from one observation to the next.

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

Ill-Conditioned Data: The ORTHOREG Procedure

PROC ORTHOREG performs linear least squares regression by using the Gentleman-Givens computational method (Gentleman 1972, 1973), 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 PROC ORTHOREG. The collinearity diagnostics in PROC REG can help you determine whether PROC ORTHOREG would be useful.

Quantile Regression: The QUANTREG and QUANTSELECT Procedures

PROC QUANTREG models the effects of covariates on the conditional quantiles of a response variable by using quantile regression. PROC QUANTSELECT performs quantile regression with model selection.

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 in modeling data that have heterogeneous conditional distributions. Data of this type occur in many fields, including biomedicine, econometrics, and ecology.

Features of PROC QUANTREG 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

The QUANTSELECT procedure shares most of its syntax and output format with PROC GLMSELECT and PROC QUANTREG. Features of PROC QUANTSELECT include the following:

- a variety of selection methods, including forward, backward, stepwise, and LASSO
- a variety of criteria, including AIC, AICC, ADJR1, SBC, significance level, testing ACL, and validation ACL
- effect selection both for individual quantile levels and for the entire quantile process
- a SAS data set that contains the design matrix
- macro variables that enable you to easily specify the selected model by using a subsequent PROC QUANTREG step

Nonlinear Regression: The NLIN and NLMIXED Procedures

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. These procedures have the following similarities:

- Nonlinear models are fit by iterative methods.
- You must provide an expression for the model by using SAS 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 procedures have the following differences:

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

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

Nonparametric Regression

Parametric regression models express the mean of an observation as a function of the regressor variables $x_1, \ldots, x_k$ and the parameters $\beta_1, \ldots, \beta_p$:

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

Not only do nonparametric regression techniques 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 in which 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 that is 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 ADAPTIVEREG, LOESS, TPSPLINE, and GAM fit nonparametric regression models by one of these methods.

Adaptive Regression: The ADAPTIVEREG Procedure

PROC ADAPTIVEREG fits multivariate adaptive regression splines as defined by Friedman (1991). The method is a nonparametric regression technique that combines regression splines and model selection methods. It does not assume parametric model forms and does not require specification of knot values for constructing regression spline terms. Instead, it constructs spline basis functions in an adaptive way by automatically selecting appropriate knot values for different variables and obtains reduced models by applying model selection techniques. PROC ADAPTIVEREG supports models that contain classification variables.

Local Regression: The LOESS Procedure

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

Thin Plate Smoothing Splines: The TPSPLINE Procedure

PROC TPSPLINE 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. PROC TPSPLINE 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 $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{ik}]$

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, PROC GAM fits a fully parametric model. If some f_j are nonparametric, PROC GAM 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

PROC ROBUSTREG 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, which was introduced by Huber (1973), 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 that was introduced by Rousseeuw (1984). The breakdown value is a measure of the proportion of observations that are contaminated by outliers that an estimation method can withstand and still maintain its robustness.
- S estimation is a high breakdown value method that was introduced by Rousseeuw and Yohai (1984). When the breakdown value is the same, it has a higher statistical efficiency than LTS estimation.
- MM estimation, which was introduced by Yohai (1987), combines high breakdown value estimation and M estimation. It has the high breakdown property of S estimation but a higher statistical efficiency.

For diagnostic purposes, PROC ROBUSTREG also implements robust leverage-point detection based on the robust Mahalanobis distance. The robust distance is computed by using a generalized minimum covariance determinant (MCD) algorithm.

Regression with Transformations: The TRANSREG Procedure

PROC TRANSREG fits linear models to data. In addition, PROC TRANSREG can find nonlinear transformations of the data and fit a linear model to the transformed data. This is in contrast to PROC REG and PROC GLM, which fit linear models to data, and PROC NLIN, which fits nonlinear models to data. PROC TRANSREG fits a variety of 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 optional variable transformations
- redundancy analysis with optional variable transformations
- canonical correlation analysis with optional variable transformations
- simple and multiple regression models with a Box-Cox (1964) transformation of the dependent variable
- regression models with penalized B-splines (Eilers and Marx 1996)

Interactive Features in the CATMOD, GLM, and REG Procedures

The CATMOD, GLM, and REG procedures are interactive: they do not stop after processing a RUN statement. You can submit more statements to continue the analysis that was started by the previous statements. Both interactive and noninteractive procedures stop if a DATA step, a new procedure step, or a QUIT or ENDSAS statement is submitted. Noninteractive SAS procedures also stop if a RUN statement is submitted.

Placement of statements is critical in interactive procedures, particularly for ODS statements. For example, if you submit the following steps, no ODS OUTPUT data set is created:

```
proc reg data=sashelp.class;
   model weight = height;
run;

ods output parameterestimates=pm;
proc glm data=sashelp.class;
   class sex;
   model weight = sex | height / solution;
run;
```

You can cause the PROC GLM step to create an output data set by adding a QUIT statement to the PROC REG step or by moving the ODS OUTPUT statement so that it follows the PROC GLM statement. For more information about the placement of RUN, QUIT, and ODS statements in interactive procedures, see Example 20.6 in Chapter 20, "Using the Output Delivery System."

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. For general statistical background about linear statistical models, see the section "Linear Model Theory" in Chapter 3, "Introduction to Statistical Modeling with SAS/STAT Software."

Exceptions and further details are documented for 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, you measure the response after controlling all factors 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 0.
- 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

Least Squares Estimators

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

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

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

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

This is an unbiased estimator, because

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

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 as the least squares estimate,

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

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, you should interpret the results 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 in 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 that are reported in PROC GLM. 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, measures the proportion of variability in the response that is explained by the regressor variables. In a linear regression model with intercept, R square is defined as

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

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

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

where n is the number of observations that are 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-square statistics also play an important indirect role in regression calculations. For example, the proportion of variability that is 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 square 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

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

Explicit and Implicit Intercepts

A linear model contains an *explicit* intercept if the **X** matrix contains a column whose nonzero values do not vary, usually a column of ones. Many SAS/STAT procedures automatically add this column of ones as the first column in the **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 an intercept should be the exception, especially if your model does not contain classification (CLASS) 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

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

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, which 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, then 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 PROC GLM 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 (because of 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 PROC TRANSREG. 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 PROC TRANSREG. 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 $\mathbf{X}$ matrix is not of full rank, then a generalized inverse can be used to solve the normal equations to minimize the SSE:

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

However, these estimates are not unique, because there are an infinite number of solutions that correspond 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 approach 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:

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

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.

For information about generalized inverses and their importance for statistical inference in linear models, see the sections "Generalized Inverse Matrices" and "Linear Model Theory" in Chapter 3, "Introduction to Statistical Modeling with SAS/STAT Software."

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 value minus the predicted value. Often other forms of residuals, such as studentized or cumulative residuals, are used for model diagnostics. 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 of observed weights, 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

$$\begin{aligned} \mathrm{RESID}_i &= y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}} \\ \mathrm{STDERR}(\mathrm{RESID}_i) &= \sqrt{s^2(1 - h_i)/w_i} \end{aligned}$$

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

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

Two types of intervals provide a measure of confidence for prediction: 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 occurs 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 *i*th 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

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

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, $\nu = n - \mathrm{rank}(\mathbf{X})$.

The limits for the prediction interval for an individual response are

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

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,

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

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*:

$$\mathrm{PRESID}_i = \frac{\mathrm{RESID}_i}{1 - h_i}$$

$$\mathrm{PRESS} = \sum_{i=1}^{n} w_i \mathrm{PRESID}_i^2$$

Testing Linear Hypotheses

Testing of linear hypotheses based on estimable functions is discussed in the section "Test of Hypotheses" on page 59 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 that are 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, you take the linear function 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:

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

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 (2004); Timm (2002); 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 let $s = \min(p,q)$. 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 these F statistics are the same and the corresponding p-values are exact, because 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

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

then

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

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 $n \leq 0$

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

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

If you specify the MSTAT=EXACT option in the appropriate statement, p-values for three of the four tests (Wilks' lambda, the Hotelling-Lawley trace, and Roy's greatest root) are computed exactly, and the p-values for the fourth test (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, because 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 results 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, 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).

Because most of the MSTAT=EXACT *p*-values are not based on the *F* distribution, the columns in the multivariate tests table that correspond to this approximation—in particular, the *F* value and the numerator and denominator degrees of freedom—are no longer displayed, and the column that contains the *p*-values is labeled "P Value" instead of "Pr > F." Suppose, for example, that 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= option. In particular, it always contains columns that correspond to both the default MSTAT=FAPPROX and the MSTAT=EXACT tests. Moreover, because the MSTAT=FAPPROX tests are relatively cheap to compute, the columns that correspond 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 1997, it might not give very good predictions in 2017 no matter how well it did in 1997. If you want to predict the future, your model must include whatever relevant factors might change over time. If the process that 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. For more information about modeling serial correlation in longitudinal, repeated measures, or time series data with SAS/STAT mixed modeling procedures, see Chapter 45, "The GLIMMIX Procedure," and Chapter 77, "The MIXED Procedure." For more information about the AUTOREG and STATESPACE procedures, see the *SAS/ETS User's Guide*.

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); 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 one another. 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 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 by using placebos and double-blind procedures), 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 that you fit is not the true model, then the parameter estimates can depend strongly on the particular values of the regressors that are used in the experiment. For example, if the response is actually a quadratic function of a regressor but you fit a linear function, then 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 because you have not actually changed anything. It might not be possible even in principle to change one regressor independently of all the others. Nor 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. For example, if you convert a regressor from feet to miles, the parameter estimate is multiplied by 5,280.

Sometimes standardized regression coefficients are used to compare the effects of regressors that are 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 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 that are 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 0 and the largest possible value is 1. 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 that 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 one another). 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 the measurements of the regressors contain errors, 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 errors 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 Square

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

In a controlled experiment, the R square depends on the values that are chosen for the regressors. A wide range of regressor values generally yields a larger R square 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 the R square.

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

You can always get an R square 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, the R square is very biased. In such cases, the adjusted R square and related statistics discussed by Darlington (1968) are less misleading.

If you fit many different models and choose the model that has the largest R square, all the statistics are biased and the p-values for the parameter estimates are not valid. You must use caution in interpreting the R square for models that have 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. R square 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, the R square, 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 an outlier 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 the 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: Identifying Influential Data and Sources of Collinearity*. 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.

Box, G. E. P., and Cox, D. R. (1964). "An Analysis of Transformations." *Journal of the Royal Statistical Society, Series B* 26:211–234.

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. S. (1999). *Fitting Equations to Data: Computer Analysis of Multifactor Data*. 2nd ed. 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's $G_{p,p}^{p,0}$ Function, and Further Tables of Wilks's Likelihood Ratio Criterion." *Biometrika* 66:519–531.

Davis, A. W. (1980). "Further Tabulation of Hotelling's Generalized T^2." *Communications in Statistics—Simulation and Computation* 9:321–336.

Draper, N. R., and Smith, H. (1998). *Applied Regression Analysis*. 3rd ed. New York: John Wiley & Sons.

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

Efron, B., Hastie, T. J., Johnstone, I. M., and Tibshirani, R. (2004). "Least Angle Regression (with Discussion)." *Annals of Statistics* 32:407–499.

Eilers, P. H. C., and Marx, B. D. (1996). "Flexible Smoothing with *B*-Splines and Penalties." *Statistical Science* 11:89–121. With discussion.

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

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

Friedman, J. H. (1991). "Multivariate Adaptive Regression Splines." *Annals of Statistics* 19:1–67.

Gentleman, W. M. (1972). *Basic Procedures for Large, Sparse, or Weighted Least Squares Problems*. Technical Report CSRR-2068, University of Waterloo, Ontario.

Gentleman, W. M. (1973). "Least Squares Computations by Givens Transformations without Square Roots." *Journal of the Institute of Mathematics and Its Applications* 12:329–336.

Goodnight, J. H. (1979). "A Tutorial on the Sweep Operator." *American Statistician* 33:149–158.

Hawkins, D. M. (1980). "A Note on Fitting a Regression with No Intercept Term." *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., and DiNardo, J. (1997). *Econometric Methods*. 4th ed. New York: McGraw-Hill.

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

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

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

Mallows, C. L. (1973). "Some Comments on C_p." *Technometrics* 15:661–675.

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

Morrison, D. F. (2004). *Multivariate Statistical Methods*. 4th ed. New York: Duxbury Press.

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

Muller, K. E. (1998). "A New F Approximation for the Pillai-Bartlett Trace under H_0." *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: University of Philippines Statistical Center.

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

Pindyck, R. S., and Rubinfeld, D. L. (1981). *Econometric Models and Econometric Forecasts*. 2nd ed. New York: McGraw-Hill.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. 2nd ed. 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*, edited by J. Franke, W. Härdle, and R. D. Martin, 256–274. Vol. 26 of Lecture Notes in Statistics. Berlin: Springer-Verlag.

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society, Series B* 58:267–288.

Timm, N. H. (2002). *Applied Multivariate Analysis*. New York: Springer.

Weisberg, S. (1985). *Applied Linear Regression*. 2nd ed. New York: John Wiley & Sons.

Weisberg, S. (2005). *Applied Linear Regression*. 3rd ed. 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