

Chapter 3

Introduction to Regression Procedures

Chapter Table of Contents

OVERVIEW	29
Introduction	29
Introductory Example	30
General Regression: The REG Procedure	34
Nonlinear Regression: The NLIN Procedure	36
Response Surface Regression: The RSREG Procedure	36
Partial Least Squares Regression: The PLS Procedure	36
Regression for Ill-conditioned Data: The ORTHOREG Procedure	37
Logistic Regression: The LOGISTIC Procedure	37
Regression With Transformations: The TRANSREG Procedure	37
Regression Using the GLM, CATMOD, LOGISTIC, PROBIT, and LIF- EREG Procedures	37
Interactive Features in the CATMOD, GLM, and REG Procedures	38
STATISTICAL BACKGROUND	38
Linear Models	38
Parameter Estimates and Associated Statistics	39
Comments on Interpreting Regression Statistics	42
Predicted and Residual Values	46
Testing Linear Hypotheses	47
Multivariate Tests	48
REFERENCES	50

Chapter 3

Introduction to Regression Procedures

Overview

This chapter reviews SAS/STAT software procedures that are used for regression analysis: CATMOD, GLM, LIFEREG, LOGISTIC, NLIN, ORTHOREG, PLS, PROBIT, REG, RSREG, and TRANSREG. The REG procedure provides the most general analysis capabilities; the other procedures give more specialized analyses. This chapter also briefly mentions several procedures in SAS/ETS software.

Introduction

Many SAS/STAT procedures, each with special features, perform regression analysis. The following procedures perform at least one type of regression analysis:

- | | |
|---------|--|
| CATMOD | analyzes data that can be represented by a contingency table. PROC CATMOD fits linear models to functions of response frequencies, and it can be used for linear and logistic regression. The CATMOD procedure is discussed in detail in Chapter 5, “Introduction to Categorical Data Analysis Procedures.” |
| GENMOD | fits generalized linear models. PROC GENMOD is especially suited for responses with discrete outcomes, and it performs logistic regression and Poisson regression as well as fitting Generalized Estimating Equations for repeated measures data. See Chapter 5, “Introduction to Categorical Data Analysis Procedures,” and Chapter 27, “The GENMOD Procedure,” for more information. |
| GLM | uses the method of least squares to fit general linear models. In addition to many other analyses, PROC GLM can perform simple, multiple, polynomial, and weighted regression. PROC GLM has many of the same input/output capabilities as PROC REG, but it does not provide as many diagnostic tools or allow interactive changes in the model or data. See Chapter 4, “Introduction to Analysis-of-Variance Procedures,” for a more detailed overview of the GLM procedure. |
| LIFEREG | fits parametric models to failure-time data that may be right censored. These types of models are commonly used in survival analysis. See Chapter 10, “Introduction to Survival Analysis |

	Procedures,” for a more detailed overview of the LIFEREG procedure.
LOGISTIC	fits logistic models for binomial and ordinal outcomes. PROC LOGISTIC provides a wide variety of model-building methods and computes numerous regression diagnostics. See Chapter 5, “Introduction to Categorical Data Analysis Procedures,” for a brief comparison of PROC LOGISTIC with other procedures.
NLIN	builds nonlinear regression models. Several different iterative methods are available.
ORTHOREG	performs regression using the Gentleman-Givens computational method. For ill-conditioned data, PROC ORTHOREG can produce more accurate parameter estimates than other procedures such as PROC GLM and PROC REG.
PLS	performs partial least squares regression, principal components regression, and reduced rank regression, with cross validation for the number of components.
PROBIT	performs probit regression as well as logistic regression and ordinal logistic regression. The PROBIT procedure is useful when the dependent variable is either dichotomous or polychotomous and the independent variables are continuous.
REG	performs linear regression with many diagnostic capabilities, selects models using one of nine methods, produces scatter plots of raw data and statistics, highlights scatter plots to identify particular observations, and allows interactive changes in both the regression model and the data used to fit the model.
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.
TRANSREG	fits univariate and multivariate linear models, optionally with spline and other nonlinear transformations. Models include ordinary regression and ANOVA, multiple and multivariate regression, metric and nonmetric conjoint analysis, metric and nonmetric vector and ideal point preference mapping, redundancy analysis, canonical correlation, and response surface regression.

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

AUTOREG	implements regression models using time-series data where the errors are autocorrelated.
PDLREG	performs regression analysis with polynomial distributed lags.

SYSLIN	handles linear simultaneous systems of equations, such as econometric models.
MODEL	handles nonlinear simultaneous systems of equations, such as econometric models.

Introductory Example

Regression analysis is the analysis of the relationship between one variable and another set of variables. The relationship is expressed as an 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*. The parameters are adjusted 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.

You might use regression analysis to find out how well you can predict a child's weight if you know that child's height. Suppose you collect your data by measuring heights and weights of 19 school children. You want to estimate the intercept β_0 and the slope β_1 of a line described by the equation

$$\text{Weight} = \beta_0 + \beta_1 \text{Height} + \epsilon$$

where

Weight	is the response variable.
β_0, β_1	are the unknown parameters.
Height	is the regressor variable.
ϵ	is the unknown error.

The data are included in the following program. The results are displayed in Figure 3.1 and Figure 3.2.

```

data class;
  input Name $ Height Weight Age;
  datalines;
Alfred 69.0 112.5 14
Alice 56.5 84.0 13
Barbara 65.3 98.0 13
Carol 62.8 102.5 14
Henry 63.5 102.5 14
James 57.3 83.0 12

```

```

Jane      59.8  84.5 12
Janet    62.5 112.5 15
Jeffrey  62.5  84.0 13
John     59.0  99.5 12
Joyce    51.3  50.5 11
Judy     64.3  90.0 14
Louise   56.3  77.0 12
Mary     66.5 112.0 15
Philip   72.0 150.0 16
Robert   64.8 128.0 12
Ronald   67.0 133.0 15
Thomas   57.5  85.0 11
William  66.5 112.0 15
;
symbol1 v=dot c=blue height=3.5pct;
proc reg;
  model Weight=Height;
  plot Weight*Height/cframe=ligr;
run;

```

The SAS System					
The REG Procedure					
Model: MODEL1					
Dependent Variable: Weight					
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

Figure 3.1. Regression for Weight and Height Data

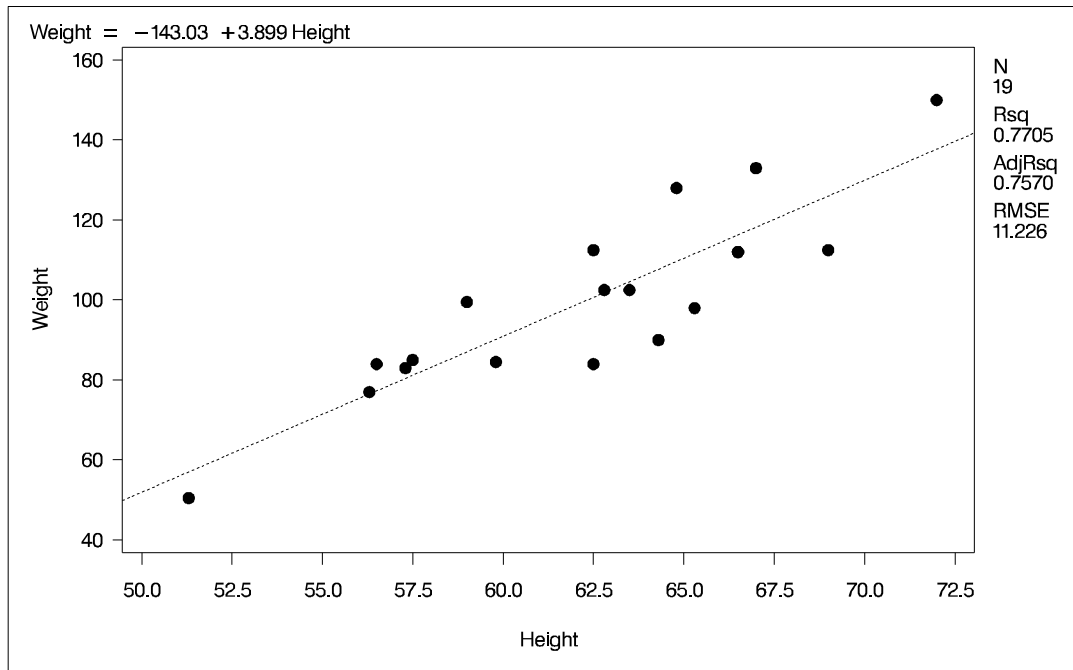


Figure 3.2. Regression for Weight and Height Data

Estimates of β_0 and β_1 for these data are $b_0 = -143.0$ and $b_1 = 3.9$, so the line is described by the equation

$$\text{Weight} = -143.0 + 3.9 * \text{Height}$$

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 scientifically the relationship. See the “Comments on Interpreting Regression Statistics” section on page 42 for more information.

The method most commonly used to estimate the parameters is to minimize the sum of squares of the differences between the actual response value and the value predicted by the equation. The estimates are called *least-squares estimates*, and the criterion value is called the *error sum of squares*

$$\text{SSE} = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

where b_0 and b_1 are the estimates of β_0 and β_1 that minimize SSE.

For a general discussion of the theory of least-squares estimation of linear models and its application to regression and analysis of variance, refer to one of the applied regression texts, including Draper and Smith (1981), Daniel and Wood (1980), Johnston (1972), and Weisberg (1985).

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

- parameter estimates 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

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

General Regression: The REG Procedure

The REG procedure is a general-purpose procedure for regression that

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

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

Model-selection Methods in PROC REG

The nine methods of model selection implemented in PROC REG are

NONE	no selection. This method is the default and uses the full model given in the MODEL statement to fit the linear regression.
FORWARD	forward selection. This method starts with no variables in the model and adds variables one by one to the model. At each step, the variable added is the one that maximizes the fit of the model. You can also specify groups of variables to treat as a unit during the selection process. An option enables you to specify the criterion for inclusion.
BACKWARD	backward elimination. This method starts with a full model and eliminates variables one by one from the model. At each step, the variable with the smallest contribution to the model is deleted. You can also specify groups of variables to treat as a unit during the selection process. An option enables you to specify the criterion for exclusion.
STEPWISE	stepwise regression, forward and backward. This method is a modification of the forward-selection method in that variables already in the model do not necessarily stay there. You can also specify groups of variables to treat as a unit during the selection process. Again, options enable you to specify criteria for entry into the model and for remaining in the model.
MAXR	maximum R^2 improvement. This method tries to find the best one-variable model, the best two-variable model, and so on. The MAXR method differs from the STEPWISE method in that many more models are evaluated with MAXR, which considers all switches before making any switch. The STEPWISE method may remove the “worst” variable without considering what the “best” remaining variable might accomplish, whereas MAXR would consider what the “best” remaining variable might accomplish. Consequently, MAXR typically takes much longer to run than STEPWISE.
MINR	minimum R^2 improvement. This method closely resembles MAXR, but the switch chosen is the one that produces the smallest increase in R^2 .
RSQUARE	finds a specified number of models having the highest R^2 in each of a range of model sizes.

CP	finds a specified number of models with the lowest C_p within a range of model sizes.
ADJRSQ	finds a specified number of models having the highest adjusted R^2 within a range of model sizes.

Nonlinear Regression: The NLIN Procedure

The NLIN procedure implements iterative methods that attempt to find least-squares estimates for nonlinear models. The default method is Gauss-Newton, although several other methods, such as Gauss or Marquardt, are available. You must specify parameter names, starting values, and expressions for the model. For some iterative methods, you also need to specify expressions for derivatives of the model with respect to the parameters. A grid search is also available to select starting values for the parameters. Since nonlinear models are often difficult to estimate, PROC NLIN may not always find the globally optimal least-squares estimates.

Response Surface Regression: The RSREG Procedure

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

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

Partial Least Squares Regression: The PLS Procedure

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

Regression for Ill-conditioned Data: The ORTHOREG Procedure

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

Logistic Regression: The LOGISTIC Procedure

The LOGISTIC procedure fits logistic models, in which the response can be either dichotomous or polychotomous. Stepwise model selection is available. You can request regression diagnostics, and predicted and residual values.

Regression With Transformations: The TRANSREG Procedure

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

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

Regression Using the GLM, CATMOD, LOGISTIC, PROBIT, and LIFEREG Procedures

The GLM procedure fits general linear models to data, and it can perform regression, analysis of variance, analysis of covariance, and many other analyses. The following features for regression distinguish PROC GLM from other regression procedures:

- direct specification of polynomial effects
- ease of specifying categorical effects (PROC GLM automatically generates dummy variables for class variables)

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

See Chapter 4, “Introduction to Analysis-of-Variance Procedures,” and Chapter 28, “The GLM Procedure,” for more details.

The CATMOD procedure can perform linear regression and logistic regression of response functions for data that can be represented in a contingency table. See Chapter 5, “Introduction to Categorical Data Analysis Procedures,” and Chapter 20, “The CATMOD Procedure,” for more details.

The LOGISTIC and PROBIT procedures can perform logistic and ordinal logistic regression. See Chapter 5, “Introduction to Categorical Data Analysis Procedures,” Chapter 35, “The LOGISTIC Procedure,” and Chapter 49, “The PROBIT Procedure,” for additional details.

The LIFEREG procedure is useful in fitting equations to data that may be right-censored. See Chapter 10, “Introduction to Survival Analysis Procedures,” and Chapter 33, “The LIFEREG Procedure,” for more details.

Interactive Features in the CATMOD, GLM, and REG Procedures

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

Statistical Background

The rest of this chapter outlines the way many SAS/STAT regression procedures calculate various regression quantities. Exceptions and further details are documented with individual procedures.

Linear Models

In matrix algebra 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 unknown errors. The first column of \mathbf{X} is usually a vector of 1s used in estimating the intercept term.

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

Other assumptions are that

- the form of the model is correct (all important explanatory variables have been included)
- regressor variables are measured without error
- the expected value of the errors is zero
- the variance of the errors (and thus the dependent variable) is a constant across observations (called σ^2)
- the errors are uncorrelated across observations

When hypotheses are tested, the additional assumption is made that the errors are normally distributed.

Statistical Model

If the model satisfies all the necessary assumptions, the least-squares estimates are the best linear unbiased estimates (BLUE). In other words, the estimates have minimum variance among the class of estimators that are unbiased and are linear functions of the responses. If the additional assumption that the error term is normally distributed is also satisfied, then

- the statistics that are computed have the proper sampling distributions for hypothesis testing
- parameter estimates are normally distributed
- various sums of squares are distributed proportional to chi-square, at least under proper hypotheses
- ratios of estimates to standard errors are distributed as Student's t under certain hypotheses
- appropriate ratios of sums of squares are distributed as F under certain hypotheses

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

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

Parameter Estimates and Associated Statistics

Parameter estimates are formed using least-squares criteria by solving the normal equations

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}$$

for the parameter estimates \mathbf{b} , yielding

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Assume for the present that $(\mathbf{X}'\mathbf{X})$ is full rank (this assumption is relaxed later). The variance of the error σ^2 is estimated by the mean square error

$$s^2 = \text{MSE} = \frac{\text{SSE}}{n - k} = \frac{1}{n - k} \sum_{i=1}^n (y_i - \mathbf{x}_i \mathbf{b})^2$$

where \mathbf{x}_i is the i th row of regressors. The parameter estimates are unbiased:

$$\begin{aligned} E(\mathbf{b}) &= \boldsymbol{\beta} \\ E(s^2) &= \sigma^2 \end{aligned}$$

The covariance matrix of the estimates is

$$\text{VAR}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2$$

The estimate of the covariance matrix is obtained by replacing σ^2 with its estimate, s^2 , in the formula preceding:

$$\text{COVB} = (\mathbf{X}'\mathbf{X})^{-1}s^2$$

The correlations of the estimates are derived by scaling to 1s on the diagonal.

Let

$$\begin{aligned} \mathbf{S} &= \text{diag}((\mathbf{X}'\mathbf{X})^{-1})^{-\frac{1}{2}} \\ \text{CORRB} &= \mathbf{S}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{S} \end{aligned}$$

Standard errors of the estimates are computed using the equation

$$\text{STDERR}(b_i) = \sqrt{(\mathbf{X}'\mathbf{X})_{ii}^{-1}s^2}$$

where $(\mathbf{X}'\mathbf{X})_{ii}^{-1}$ is the i th diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$. The ratio

$$t = \frac{b_i}{\text{STDERR}(b_i)}$$

is distributed as Student's t under the hypothesis that β_i is zero. Regression procedures display the t ratio and the significance probability, which is the probability under the hypothesis $\beta_i = 0$ of a larger absolute t value than was actually obtained. When the probability is less than some small level, the event is considered so unlikely that the hypothesis is rejected.

Type I SS and Type II SS measure the contribution of a variable to the reduction in SSE. Type I SS measure the reduction in SSE as that variable is entered into the model in sequence. Type II SS are the increment in SSE that results from removing the variable from the full model. Type II SS are equivalent to the Type III and Type IV SS reported in the GLM procedure. If Type II SS are used in the numerator of an F test, the test is equivalent to the t test for the hypothesis that the parameter is zero. In polynomial models, Type I SS measure the contribution of each polynomial term after it is orthogonalized to the previous terms in the model. The four types of SS are described in Chapter 11, "The Four Types of Estimable Functions."

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.

R^2 is an indicator of how much of the variation in the data is explained by the model. It is defined as

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

where SSE is the sum of squares for error and TSS is the corrected total sum of squares. The Adjusted R^2 statistic is an alternative to R^2 that is adjusted for the number of parameters in the model. This is calculated as

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

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

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 goes to 0 and the variance inflation gets very large. Tolerance (TOL) is 1 minus the R^2 that results from the regression of the other variables in the model on that regressor. Variance inflation (VIF) is the diagonal of $(\mathbf{X}'\mathbf{X})^{-1}$ if $(\mathbf{X}'\mathbf{X})$ is scaled to correlation form. The statistics are related as

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

Models Not of Full Rank

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

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}$$

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

$$E(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-}(\mathbf{X}'\mathbf{X})\boldsymbol{\beta}$$

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

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 may 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 may 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 may nevertheless be highly problematical.

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 1958, it may not give very good predictions in 1988 no matter how well it did in 1958. If it is the future you want to predict, your model must include whatever relevant factors may change over time. If the process you are studying does in fact change over time, you must take observations at several, perhaps many, different times. Analysis of such data is the province of SAS/ETS procedures such as AUTOREG and STATESPACE. Refer to the *SAS/ETS User’s Guide* for more information on these procedures.

The comments in the rest of this section are directed toward linear least-squares regression. Nonlinear regression and non-least-squares regression often introduce further complications.

For more detailed discussions of the interpretation of regression statistics, see Darlington (1968), Mosteller and Tukey (1977), Weisberg (1985), and Younger (1979).

Interpreting Parameter Estimates from a Controlled Experiment

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

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

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

Interpreting Parameter Estimates from an Observational Study

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

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

Comparing Parameter Estimates

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

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

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

Correlated Regressors

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

If the regressors are correlated, it becomes difficult to disentangle the effects of one regressor from another, and the parameter estimates may be highly dependent on which regressors are used in the model. Two correlated regressors may 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 may be possible to recode correlated regressors to make interpretation easier. For example, if X and Y are highly correlated, they could be replaced in a linear regression by $X + Y$ and $X - Y$ without changing the fit of the model or statistics for other regressors.

Errors in the Regressors

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

Probability Values (p -values)

Probability values (p -values) do not necessarily measure the importance of a regressor. An important regressor can have a large (nonsignificant) p -value if the sample

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

The p -values are always approximations. The assumptions required to compute exact p -values are never satisfied in practice.

Interpreting R^2

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

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

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

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

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

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

Incorrect Data Values

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

Predicted and Residual Values

After the model has been fit, predicted and residual values are usually calculated and output. The predicted values are calculated from the estimated regression equation; the residuals are calculated as actual minus predicted. Some procedures can calculate standard errors of residuals, predicted mean values, and individual predicted values.

Consider the i th observation where \mathbf{x}_i is the row of regressors, \mathbf{b} is the vector of parameter estimates, and s^2 is the mean squared error.

Let

$$h_i = \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i' \quad (\text{the leverage})$$

Then

$$\begin{aligned} \hat{y}_i &= \mathbf{x}_i\mathbf{b} \quad (\text{the predicted mean value}) \\ \text{STDERR}(\hat{y}_i) &= \sqrt{h_i s^2} \quad (\text{the standard error of the predicted mean}) \end{aligned}$$

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

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

The residual is defined as

$$\begin{aligned} \text{RESID}_i &= y_i - \mathbf{x}_i\mathbf{b} \quad (\text{the residual}) \\ \text{STDERR}(\text{RESID}_i) &= \sqrt{(1 - h_i)s^2} \quad (\text{the standard error of the residual}) \end{aligned}$$

The ratio of the residual to its standard error, called the *studentized residual*, is sometimes shown as

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

There are two kinds of confidence intervals for predicted values. One type of confidence interval is an interval for the mean value of the response. The other type, sometimes called a *prediction* or *forecasting interval*, is an interval for the actual value of a response, which is the mean value plus error.

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

$$\begin{aligned} \text{LowerM} &= \mathbf{x}_i\mathbf{b} - t_{\alpha/2}\sqrt{h_i s^2} \\ \text{UpperM} &= \mathbf{x}_i\mathbf{b} + t_{\alpha/2}\sqrt{h_i s^2} \end{aligned}$$

where $t_{\alpha/2}$ is the tabulated t statistic with degrees of freedom equal to the degrees of freedom for the mean squared error.

The limits for the confidence interval for an actual individual response are

$$\begin{aligned}\text{LowerI} &= \mathbf{x}_i\mathbf{b} - t_{\alpha/2}\sqrt{(1+h_i)s^2} \\ \text{UpperI} &= \mathbf{x}_i\mathbf{b} + t_{\alpha/2}\sqrt{(1+h_i)s^2}\end{aligned}$$

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

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

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

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

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

Testing Linear Hypotheses

The general form of a linear hypothesis for the parameters is

$$\mathbf{H}_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{c}$$

where \mathbf{L} is $q \times k$, $\boldsymbol{\beta}$ is $k \times 1$, and \mathbf{c} is $q \times 1$. To test this hypothesis, the linear function is taken with respect to the parameter estimates:

$$\mathbf{L}\mathbf{b} - \mathbf{c}$$

This has variance

$$\text{Var}(\mathbf{L}\mathbf{b} - \mathbf{c}) = \mathbf{L}\text{Var}(\mathbf{b})\mathbf{L}' = \mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'\sigma^2$$

where \mathbf{b} is the estimate of $\boldsymbol{\beta}$.

A quadratic form called the *sum of squares due to the hypothesis* is calculated:

$$SS(\mathbf{Lb} - \mathbf{c}) = (\mathbf{Lb} - \mathbf{c})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{Lb} - \mathbf{c})$$

If you assume that this is testable, the SS can be used as a numerator of the F test:

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

This is compared with an F distribution with q and dfe degrees of freedom, where dfe is the degrees of freedom for residual error.

Multivariate Tests

Multivariate hypotheses involve several dependent variables in the form

$$\mathbf{H}_0 : \mathbf{L}\boldsymbol{\beta}\mathbf{M} = \mathbf{d}$$

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 written

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

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

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

These multivariate tests are covered in detail in Morrison (1976); Timm (1975); Mardia, Kent, and Bibby (1979); Bock (1975); and other works cited in Chapter 6, "Introduction to Multivariate Procedures."

To test this hypothesis, construct two matrices, \mathbf{H} and \mathbf{E} , that correspond to the numerator and denominator of a univariate F test:

$$\begin{aligned}\mathbf{H} &= \mathbf{M}'(\mathbf{LB} - \mathbf{c}\mathbf{j})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{LB} - \mathbf{c}\mathbf{j})\mathbf{M} \\ \mathbf{E} &= \mathbf{M}'(\mathbf{Y}'\mathbf{Y} - \mathbf{B}'(\mathbf{X}'\mathbf{X})\mathbf{B})\mathbf{M}\end{aligned}$$

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 i th 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 $\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'$. Let v be the error degrees of freedom and $s = \min(p, q)$. Let $m = (|p - q| - 1)/2$, and let $n = (v - p - 1)/2$. Then the following statistics have the approximate F statistics as shown.

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 , where

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

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

Pillai's Trace

If

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

then

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

is approximately F with $s(2m + s + 1)$ and $s(2n + s + 1)$ degrees of freedom.**Hotelling-Lawley Trace**

If

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

then

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

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

Roy's Maximum Root

If

$$\Theta = \lambda_1$$

then

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

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

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

References

- Allen, D.M. (1971), "Mean Square Error of Prediction as a Criterion for Selecting Variables," *Technometrics*, 13, 469–475.
- Allen, D.M. and Cady, F.B. (1982), *Analyzing Experimental Data by Regression*, Belmont, CA: Lifetime Learning Publications.
- Belsley, D.A., Kuh, E., and Welsch, R.E. (1980), *Regression Diagnostics*, New York: John Wiley & Sons, Inc.
- Bock, R.D. (1975), *Multivariate Statistical Methods in Behavioral Research*, New York: McGraw-Hill Book Co.
- Box, G.E.P. (1966), "The Use and Abuse of Regression," *Technometrics*, 8, 625–629.
- Cook, R.D. (1977), "Detection of Influential Observations in Linear Regression," *Technometrics*, 19, 15–18.
- Cook, R.D. (1979), "Influential Observations in Linear Regression," *Journal of the American Statistical Association*, 74, 169–174.
- Daniel, C. and Wood, F. (1980), *Fitting Equations to Data*, Revised Edition, New York: John Wiley & Sons, Inc.
- Darlington, R.B. (1968), "Multiple Regression in Psychological Research and Practice," *Psychological Bulletin*, 69, 161–182.
- Draper, N. and Smith, H. (1981), *Applied Regression Analysis*, Second Edition, New York: John Wiley & Sons, Inc.
- Durbin, J. and Watson, G.S. (1951), "Testing for Serial Correlation in Least Squares Regression," *Biometrika*, 37, 409–428.
- Freund, R.J., Littell, R.C., and Spector P.C. (1991), *SAS System for Linear Models*, Cary, NC: SAS Institute Inc.

- Freund, R.J. and Littell, R.C. (1986), *SAS System for Regression, 1986 Edition*, Cary, NC: SAS Institute Inc.
- Goodnight, J.H. (1979), "A Tutorial on the SWEEP Operator," *The American Statistician*, 33, 149–158. (Also available as SAS Technical Report R-106, *The Sweep Operator: Its Importance in Statistical Computing*, Cary, NC: SAS Institute Inc.)
- Hawkins, D.M. (1980), "A Note on Fitting a Regression With No Intercept Term," *The American Statistician*, 34, 233.
- Hosmer, D.W, Jr and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: John Wiley & Sons, Inc.
- Johnston, J. (1972), *Econometric Methods*, New York: McGraw-Hill Book Co.
- Kennedy, W.J. and Gentle, J.E. (1980), *Statistical Computing*, New York: Marcel Dekker, Inc.
- Kvalseth, T.O. (1985), "Cautionary Note About R^2 ," *The American Statistician*, 39, 279.
- Mallows, C.L. (1973), "Some Comments on C_p ," *Technometrics*, 15, 661–75.
- Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979), *Multivariate Analysis*, London: Academic Press.
- Morrison, D.F. (1976), *Multivariate Statistical Methods*, Second Edition, New York: McGraw-Hill Book Co.
- Mosteller, F. and Tukey, J.W. (1977), *Data Analysis and Regression*, Reading, MA: Addison-Wesley Publishing Co., Inc.
- Neter, J. and Wasserman, W. (1974), *Applied Linear Statistical Models*, Homewood, IL: Irwin.
- Pillai, K.C.S. (1960), *Statistical Table for Tests of Multivariate Hypotheses*, Manila: The Statistical Center, University of Philippines.
- Pindyck, R.S. and Rubinfeld, D.L. (1981), *Econometric Models and Econometric Forecasts*, Second Edition, New York: McGraw-Hill Book Co.
- Rao, C.R. (1973), *Linear Statistical Inference and Its Applications*, Second Edition, New York: John Wiley & Sons, Inc.
- Rawlings, J.O. (1988), *Applied Regression Analysis: A Research Tool*, Pacific Grove, California: Wadsworth & Brooks/Cole Advanced Books & Software.
- Timm, N.H. (1975), *Multivariate Analysis with Applications in Education and Psychology*, Monterey, CA: Brooks-Cole Publishing Co.
- Weisberg, S. (1985), *Applied Linear Regression*, Second Edition. New York: John Wiley & Sons, Inc.
- Younger, M.S. (1979), *Handbook for Linear Regression*, North Scituate, MA: Duxbury Press.