SAS/STAT® 15.1 User's Guide

Introduction to Statistical Modeling with SAS/STAT Software

This document is an individual chapter from *SAS/STAT® 15.1 User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2018. *SAS/STAT® 15.1 User's Guide*. Cary, NC: SAS Institute Inc.

SAS/STAT® 15.1 User's Guide

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

November 2018

Chapter 3
Introduction to Statistical Modeling with SAS/STAT Software

Contents

Overview: Statistical Modeling

The majority of procedures in SAS/STAT software are dedicated to solving problems in statistical modeling. The goal of this chapter is to provide a roadmap to statistical models and to modeling tasks, enabling you to make informed choices about the appropriate modeling context and tool. This chapter also introduces important terminology, notation, and concepts used throughout this documentation. Subsequent introductory chapters discuss model families and related procedures.

It is difficult to capture the complexity of statistical models in a simple scheme, so the classification used here is necessarily incomplete. It is most practical to classify models in terms of simple criteria, such as the presence of random effects, the presence of nonlinearity, characteristics of the data, and so on. That is the approach used here. After a brief introduction to statistical modeling in general terms, the chapter describes a number of model classifications and relates them to modeling tools in SAS/STAT software.

Statistical Models

Deterministic and Stochastic Models

Purely mathematical models, in which the relationships between inputs and outputs are captured entirely in deterministic fashion, can be important theoretical tools but are impractical for describing observational, experimental, or survey data. For such phenomena, researchers usually allow the model to draw on stochastic as well as deterministic elements. When the uncertainty of realizations leads to the inclusion of random components, the resulting models are called *stochastic* models. A *statistical* model, finally, is a stochastic model that contains *parameters*, which are unknown constants that need to be estimated based on assumptions about the model and the observed data.

There are many reasons why statistical models are preferred over deterministic models. For example:

- Randomness is often introduced into a system in order to achieve a certain balance or representativeness. For example, random assignment of treatments to experimental units allows unbiased inferences about treatment effects. As another example, selecting individuals for a survey sample by random mechanisms ensures a representative sample.

- Even if a deterministic model can be formulated for the phenomenon under study, a stochastic model can provide a more parsimonious and more easily comprehended description. For example, it is possible in principle to capture the result of a coin toss with a deterministic model, taking into account the properties of the coin, the method of tossing, conditions of the medium through which the coin travels and of the surface on which it lands, and so on. A very complex model is required to describe the simple outcome—heads or tails. Alternatively, you can describe the outcome quite simply as the result of a stochastic process, a Bernoulli variable that results in heads with a certain probability.

- It is often sufficient to describe the average behavior of a process, rather than each particular realization. For example, a regression model might be developed to relate plant growth to nutrient availability. The explicit aim of the model might be to describe how the average growth changes with nutrient availability, not to predict the growth of an individual plant. The support for the notion of averaging in a model lies in the nature of expected values, describing typical behavior in the presence of randomness. This, in turn, requires that the model contain stochastic components.

The defining characteristic of statistical models is their dependence on parameters and the incorporation of stochastic terms. The properties of the model and the properties of quantities derived from it must be studied in a long-run, average sense through expectations, variances, and covariances. The fact that the parameters of the model must be estimated from the data introduces a stochastic element in applying a statistical model: because the model is not deterministic but includes randomness, parameters and related quantities derived from the model are likewise random. The properties of parameter estimators can often be described only in an asymptotic sense, imagining that some aspect of the data increases without bound (for example, the number of observations or the number of groups).

The process of estimating the parameters in a statistical model based on your data is called *fitting* the model. For many classes of statistical models there are a number of procedures in SAS/STAT software that can perform the fitting. In many cases, different procedures solve identical estimation problems—that is, their parameter estimates are identical. In some cases, the same model parameters are estimated by different statistical principles, such as least squares versus maximum likelihood estimation. Parameter estimates obtained by different methods typically have different statistical properties—distribution, variance, bias, and so on. The choice between competing estimation principles is often made on the basis of properties of the estimators. Distinguishing properties might include (but are not necessarily limited to) computational ease, interpretive ease, bias, variance, mean squared error, and consistency.

Model-Based and Design-Based Randomness

A statistical model is a description of the data-generating mechanism, not a description of the specific data to which it is applied. The aim of a model is to capture those aspects of a phenomenon that are relevant to inquiry and to explain how the data could have come about as a realization of a random experiment. These relevant aspects might include the genesis of the randomness and the stochastic effects in the phenomenon under study. Different schools of thought can lead to different model formulations, different analytic strategies, and different results. Coarsely, you can distinguish between a viewpoint of *innate* randomness and one of *induced* randomness. This distinction leads to model-based and design-based inference approaches.

In a design-based inference framework, the random variation in the observed data is induced by random *selection* or random *assignment*. Consider the case of a survey sample from a finite population of size N; suppose that $\mathcal{F}_N = \{y_i : i \in U_N\}$ denotes the finite set of possible values and U_N is the index set $U_N = \{1, 2, \ldots, N\}$. Then a sample S, a subset of U_N, is selected by probability rules. The realization of the random experiment is the selection of a particular set S; the associated values selected from $\mathcal{F}_N$ are considered fixed. If properties of a design-based sampling estimator are evaluated, such as bias, variance, and mean squared error, they are evaluated with respect to the distribution induced by the sampling mechanism.

Design-based approaches also play an important role in the analysis of data from controlled experiments by randomization tests. Suppose that k treatments are to be assigned to kr homogeneous experimental units. If you form k sets of r units with equal probability, and you assign the jth treatment to the tth set, a completely randomized experimental design (CRD) results. A design-based view treats the potential response of a particular treatment for a particular experimental unit as a constant. The stochastic nature of the error-control design is induced by randomly selecting one of the potential responses.

Statistical models are often used in the design-based framework. In a survey sample the model is used to motivate the choice of the finite population parameters and their sample-based estimators. In an experimental design, an assumption of additivity of the contributions from treatments, experimental units, observational errors, and experimental errors leads to a linear statistical model. The approach to statistical inference where statistical models are used to construct estimators and their properties are evaluated with respect to

the distribution induced by the sample selection mechanism is known as *model-assisted inference* (Särndal, Swensson, and Wretman 1992).

In a purely model-based framework, the only source of random variation for inference comes from the unknown variation in the responses. Finite population values are thought of as a realization of a superpopulation model that describes random variables $Y_1, Y_2, \ldots$. The observed values $y_1, y_2, \ldots$ are realizations of these random variables. A model-based framework does not imply that there is only one source of random variation in the data. For example, mixed models might contain random terms that represent selection of effects from hierarchical (super-) populations at different granularity. The analysis takes into account the hierarchical structure of the random variation, but it continues to be model based.

A design-based approach is implicit in SAS/STAT procedures whose name commences with SURVEY, such as the SURVEYFREQ, SURVEYMEANS, SURVEYREG, SURVEYLOGISTIC, and SURVEYPHREG procedures. Inferential approaches are model based in other SAS/STAT procedures. For more information about analyzing survey data with SAS/STAT software, see Chapter 14, "Introduction to Survey Procedures."

Model Specification

If the model is accepted as a description of the data-generating mechanism, then its parameters are estimated using the data at hand. Once the parameter estimates are available, you can apply the model to answer questions of interest about the study population. In other words, the model becomes the lens through which you view the problem itself, in order to ask and answer questions of interest. For example, you might use the estimated model to derive new predictions or forecasts, to test hypotheses, to derive confidence intervals, and so on.

Obviously, the model must be "correct" to the extent that it sufficiently describes the data-generating mechanism. Model selection, diagnosis, and discrimination are important steps in the model-building process. This is typically an iterative process, starting with an initial model and refining it. The first important step is thus to formulate your knowledge about the data-generating process and to express the real observed phenomenon in terms of a statistical model. A statistical model describes the distributional properties of one or more variables, the *response* variables. The extent of the required distributional specification depends on the model, estimation technique, and inferential goals. This description often takes the simple form of a model with additive error structure:

$$\text{response} = \text{mean} + \text{error}$$

In mathematical notation this simple model equation becomes

$$Y = f(x_1, \ldots, x_k; \beta_1, \ldots, \beta_p) + \epsilon$$

In this equation Y is the *response* variable, often also called the *dependent* variable or the *outcome* variable. The terms $x_1, \ldots, x_k$ denote the values of k regressor variables, often termed the *covariates* or the "independent" variables. The terms $\beta_1, \ldots, \beta_p$ denote parameters of the model, unknown constants that are to be estimated. The term ϵ denotes the random disturbance of the model; it is also called the residual term or the error term of the model.

In this simple model formulation, stochastic properties are usually associated only with the ϵ term. The covariates $x_1, \ldots, x_k$ are usually known values, not subject to random variation. Even if the covariates are measured with error, so that their values are in principle random, they are considered fixed in most models fit by SAS/STAT software. In other words, stochastic properties under the model are derived conditional on the

xs. If ϵ is the only stochastic term in the model, and if the errors have a mean of zero, then the function $f(\cdot)$ is the *mean function* of the statistical model. More formally,

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

where $\mathrm{E}[\cdot]$ denotes the expectation operator.

In many applications, a simple model formulation is inadequate. It might be necessary to specify not only the stochastic properties of a single error term, but also how model errors associated with different observations relate to each other. A simple additive error model is typically inappropriate to describe the data-generating mechanism if the errors do not have zero mean or if the variance of observations depends on their means. For example, if Y is a Bernoulli random variable that takes on the values 0 and 1 only, a regression model with additive error is not meaningful. Models for such data require more elaborate formulations involving probability distributions.

Classes of Statistical Models

Linear and Nonlinear Models

A statistical estimation problem is nonlinear if the estimating equations—the equations whose solution yields the parameter estimates—depend on the parameters in a nonlinear fashion. Such estimation problems typically have no closed-form solution and must be solved by iterative, numerical techniques.

Nonlinearity in the mean function is often used to distinguish between linear and nonlinear models. A model has a nonlinear mean function if the derivative of the mean function with respect to the parameters depends on at least one other parameter. Consider, for example, the following models that relate a response variable Y to a single regressor variable x:

$$\mathrm{E}[Y|x] = \beta_0 + \beta_1 x$$
$$\mathrm{E}[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2$$
$$\mathrm{E}[Y|x] = \beta + x/\alpha$$

In these expressions, $\mathrm{E}[Y|x]$ denotes the expected value of the response variable Y at the fixed value of x. (The conditioning on x simply indicates that the predictor variables are assumed to be non-random. Conditioning is often omitted for brevity in this and subsequent chapters.)

The first model in the previous list is a simple linear regression (SLR) model. It is linear in the parameters β_0 and β_1 since the model derivatives do not depend on unknowns:

$$\frac{\partial}{\beta_0}(\beta_0 + \beta_1 x) = 1$$
$$\frac{\partial}{\beta_1}(\beta_0 + \beta_1 x) = x$$

The model is also linear in its relationship with x (a straight line). The second model is also linear in the parameters, since

$$\frac{\partial}{\beta_0}\left(\beta_0 + \beta_1 x + \beta_2 x^2\right) = 1$$
$$\frac{\partial}{\beta_1}\left(\beta_0 + \beta_1 x + \beta_2 x^2\right) = x$$
$$\frac{\partial}{\beta_2}\left(\beta_0 + \beta_1 x + \beta_2 x^2\right) = x^2$$

However, this second model is *curvilinear*, since it exhibits a curved relationship when plotted against x. The third model, finally, is a nonlinear model since

$$\frac{\partial}{\beta}\left(\beta + x/\alpha\right) = 1$$
$$\frac{\partial}{\alpha}\left(\beta + x/\alpha\right) = -\frac{x}{\alpha^2}$$

The second of these derivatives depends on a parameter α. A model is nonlinear if it is not linear in at least one parameter. Only the third model is a nonlinear model. A graph of $\mathrm{E}[Y]$ versus the regressor variable thus does not indicate whether a model is nonlinear. A curvilinear relationship in this graph can be achieved by a model that is linear in the parameters.

Nonlinear mean functions lead to nonlinear estimation. It is important to note, however, that nonlinear estimation arises also because of the estimation principle or because the model structure contains nonlinearity in other parts, such as the covariance structure. For example, fitting a simple linear regression model by minimizing the sum of the absolute residuals leads to a nonlinear estimation problem despite the fact that the mean function is linear.

Regression Models and Models with Classification Effects

A linear regression model in the broad sense has the form

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

where $\mathbf{Y}$ is the vector of response values, $\mathbf{X}$ is the matrix of regressor effects, $\boldsymbol{\beta}$ is the vector of regression parameters, and $\boldsymbol{\epsilon}$ is the vector of errors or residuals. A regression model in the narrow sense—as compared to a classification model—is a linear model in which all regressor effects are continuous variables. In other words, each effect in the model contributes a single column to the $\mathbf{X}$ matrix and a single parameter to the overall model. For example, a regression of subjects' weight (Y) on the regressors age (x_1) and body mass index (bmi, x_2) is a regression model in this narrow sense. In symbolic notation you can write this regression model as

weight = age + bmi + error

This symbolic notation expands into the statistical model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

Single parameters are used to model the effects of age (β_1) and bmi (β_2), respectively.

A classification effect, on the other hand, is associated with possibly more than one column of the **X** matrix. Classification with respect to a variable is the process by which each observation is associated with one of k levels; the process of determining these k levels is referred to as *levelization* of the variable. Classification variables are used in models to identify experimental conditions, group membership, treatments, and so on. The actual values of the classification variable are not important, and the variable can be a numeric or a character variable. What is important is the association of discrete values or levels of the classification variable with groups of observations. For example, in the previous illustration, if the regression also takes into account the subjects' gender, this can be incorporated in the model with a two-level classification variable. Suppose that the values of the gender variable are coded as 'F' and 'M', respectively. In symbolic notation the model

$$\text{weight} = \text{age} + \text{bmi} + \text{gender} + \text{error}$$

expands into the statistical model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \tau_1 \mathrm{I}(\text{gender} = \text{'F'}) + \tau_2 \mathrm{I}(\text{gender} = \text{'M'}) + \epsilon_i$$

where I(gender='F') is the indicator function that returns 1 if the value of the gender variable is 'F' and 0 otherwise. Parameters τ_1 and τ_2 are associated with the gender classification effect. This form of parameterizing the gender effect in the model is only one of several different methods of incorporating the levels of a classification variable in the model. This form, the so-called singular parameterization, is the most general approach, and it is used in the GLM, MIXED, and GLIMMIX procedures. Alternatively, classification effects with various forms of nonsingular parameterizations are available in such procedures as GENMOD and LOGISTIC. See the documentation for the individual SAS/STAT procedures on their respective facilities for parameterizing classification variables and the section "Parameterization of Model Effects" on page 393 in Chapter 19, "Shared Concepts and Topics," for general details.

Models that contain only classification effects are often identified with *analysis of variance* (ANOVA) models, because ANOVA methods are frequently used in their analysis. This is particularly true for experimental data where the model effects comprise effects of the treatment and error-control design. However, classification effects appear more widely than in models to which analysis of variance methods are applied. For example, many mixed models, where parameters are estimated by restricted maximum likelihood, consist entirely of classification effects but do not permit the sum of squares decomposition typical for ANOVA techniques.

Many models contain both continuous and classification effects. For example, a continuous-by-class effect consists of at least one continuous variable and at least one classification variable. Such effects are convenient, for example, to vary slopes in a regression model by the levels of a classification variable. Also, recent enhancements to linear modeling syntax in some SAS/STAT procedures (including GLIMMIX and GLMSELECT) enable you to construct sets of columns in **X** matrices from a single continuous variable. An example is modeling with splines where the values of a continuous variable x are expanded into a spline basis that occupies multiple columns in the **X** matrix. For purposes of the analysis you can treat these columns as a single unit or as individual, unrelated columns. For more details, see the section "EFFECT Statement" on page 403 in Chapter 19, "Shared Concepts and Topics."

Univariate and Multivariate Models

A multivariate statistical model is a model in which multiple response variables are modeled jointly. Suppose, for example, that your data consist of heights (h_i) and weights (w_i) of children, collected over several years (t_i). The following separate regressions represent two univariate models:

$$w_i = \beta_{w0} + \beta_{w1} t_i + \epsilon_{wi}$$
$$h_i = \beta_{h0} + \beta_{h1} t_i + \epsilon_{hi}$$

In the univariate setting, no information about the children's heights "flows" to the model about their weights and vice versa. In a multivariate setting, the heights and weights would be modeled jointly. For example:

$$\begin{aligned} \mathbf{Y}_i = \begin{bmatrix} w_i \\ h_i \end{bmatrix} &= \mathbf{X}\boldsymbol{\beta} + \begin{bmatrix} \epsilon_{wi} \\ \epsilon_{hi} \end{bmatrix} \\ &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \\ \boldsymbol{\epsilon}_i &\sim \left(\mathbf{0}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right) \end{aligned}$$

The vectors $\mathbf{Y}_i$ and $\boldsymbol{\epsilon}_i$ collect the responses and errors for the two observation that belong to the same subject. The errors from the same child now have the correlation

$$\text{Corr}[\epsilon_{wi}, \epsilon_{hi}] = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \, \sigma_2^2}}$$

and it is through this correlation that information about heights "flows" to the weights and vice versa. This simple example shows only one approach to modeling multivariate data, through the use of covariance structures. Other techniques involve seemingly unrelated regressions, systems of linear equations, and so on.

Multivariate data can be coarsely classified into three types. The response vectors of *homogeneous multivariate data* consist of observations of the same attribute. Such data are common in repeated measures experiments and longitudinal studies, where the same attribute is measured repeatedly over time. Homogeneous multivariate data also arise in spatial statistics where a set of geostatistical data is the incomplete observation of a single realization of a random experiment that generates a two-dimensional surface. One hundred measurements of soil electrical conductivity collected in a forest stand compose a single observation of a 100-dimensional homogeneous multivariate vector. *Heterogeneous multivariate* observations arise when the responses that are modeled jointly refer to different attributes, such as in the previous example of children's weights and heights. There are two important subtypes of heterogeneous multivariate data. In *homocatanomic multivariate data* the observations come from the same distributional family. For example, the weights and heights might both be assumed to be normally distributed. With *heterocatanomic multivariate data* the observations can come from different distributional families. The following are examples of heterocatanomic multivariate data:

- For each patient you observe blood pressure (a continuous outcome), the number of prior episodes of an illness (a count variable), and whether the patient has a history of diabetes in the family (a binary outcome). A multivariate model that models the three attributes jointly might assume a lognormal distribution for the blood pressure measurements, a Poisson distribution for the count variable and a Bernoulli distribution for the family history.
- In a study of HIV/AIDS survival, you model jointly a patient's CD4 cell count over time—itself a homogeneous multivariate outcome—and the survival of the patient (event-time data).

Fixed, Random, and Mixed Models

Each term in a statistical model represents either a *fixed effect* or a *random effect*. Models in which all effects are fixed are called fixed-effects models. Similarly, models in which all effects are random—apart from possibly an overall intercept term—are called random-effects models. Mixed models, then, are those models that have fixed-effects and random-effects terms. In matrix notation, the linear fixed, linear random, and linear mixed model are represented by the following model equations, respectively:

$$\begin{aligned}
\mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
\mathbf{Y} &= \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \\
\mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}
\end{aligned}$$

In these expressions, $\mathbf{X}$ and $\mathbf{Z}$ are design or regressor matrices associated with the fixed and random effects, respectively. The vector $\boldsymbol{\beta}$ is a vector of fixed-effects parameters, and the vector $\boldsymbol{\gamma}$ represents the random effects. The mixed modeling procedures in SAS/STAT software assume that the random effects $\boldsymbol{\gamma}$ follow a normal distribution with variance-covariance matrix $\mathbf{G}$ and, in most cases, that the random effects have mean zero.

Random effects are often associated with classification effects, but this is not necessary. As an example of random regression effects, you might want to model the slopes in a growth model as consisting of two components: an overall (fixed-effects) slope that represents the slope of the average individual, and individual-specific random deviations from the overall slope. The $\mathbf{X}$ and $\mathbf{Z}$ matrix would then have column entries for the regressor variable associated with the slope. You are modeling fixed and randomly varying regression coefficients.

Having random effects in your model has a number of important consequences:

- Some observations are no longer uncorrelated but instead have a covariance that depends on the variance of the random effects.
- You can and should distinguish between the inference spaces; inferences can be drawn in a broad, intermediate, and narrow inference space. In the narrow inference space, conclusions are drawn about the particular values of the random effects selected in the study. The broad inference space applies if inferences are drawn with respect to all possible levels of the random effects. The intermediate inference space can be applied for effects consisting of more than one random term, when inferences are broad with respect to some factors and narrow with respect to others. In fixed-effects models, there is no corresponding concept to the broad and intermediate inference spaces.
- Depending on the structure of $\mathbf{G}$ and $\mathrm{Var}[\boldsymbol{\epsilon}]$ and also subject to the balance in your data, there might be no closed-form solution for the parameter estimates. Although the model is linear in $\boldsymbol{\beta}$, iterative estimation methods might be required to estimate all parameters of the model.
- Certain concepts, such as least squares means and Type III estimable functions, are meaningful only for fixed effects.
- By using random effects, you are modeling variation through variance. Variation in data simply implies that things are not equal. Variance, on the other hand, describes a feature of a random variable. Random effects in your model are random variables: they model variation through variance.

It is important to properly determine the nature of the model effects as fixed or random. An effect is either fixed or random by its very nature; it is improper to consider it fixed in one analysis and random in another

depending on what type of results you want to produce. If, for example, a treatment effect is random and you are interested in comparing treatment means, and only the levels selected in the study are of interest, then it is not appropriate to model the treatment effect as fixed so that you can draw on least squares mean analysis. The appropriate strategy is to model the treatment effect as random and to compare the solutions for the treatment effects in the narrow inference space.

In determining whether an effect is fixed or random, it is helpful to inquire about the *genesis* of the effect. If the levels of an effect are randomly sampled, then the effect is a random effect. The following are examples:

- In a large clinical trial, drugs A, B, and C are applied to patients in various clinical centers. If the clinical centers are selected at random from a population of possible clinics, their effect on the response is modeled with a random effect.
- In repeated measures experiments with people or animals as subjects, subjects are declared to be random because they are selected from the larger population to which you want to generalize.
- Fertilizers could be applied at a number of levels. Three levels are randomly selected for an experiment to represent the population of possible levels. The fertilizer effects are random effects.

Quite often it is not possible to select effects at random, or it is not known how the values in the data became part of the study. For example, suppose you are presented with a data set consisting of student scores in three school districts, with four to ten schools in each district and two to three classrooms in each school. How do you decide which effects are fixed and which are random? As another example, in an agricultural experiment conducted in successive years at two locations, how do you decide whether location and year effects are fixed or random? In these situations, the fixed or random nature of the effect might be debatable, bearing out the adage that "one modeler's fixed effect is another modeler's random effect." However, this fact does not constitute license to treat as random those effects that are clearly fixed, or vice versa.

When an effect cannot be randomized or it is not known whether its levels have been randomly selected, it can be a random effect if its impact on the outcome variable is of a stochastic nature—that is, if it is the realization of a random process. Again, this line of thinking relates to the genesis of the effect. A random year, location, or school district effect is a placeholder for different environments that cannot be selected at random but whose effects are the cumulative result of many individual random processes. Note that this argument does not imply that effects are random because the experimenter does not know much about them. The key notion is that effects represent something, whether or not that something is known to the modeler. Broadening the inference space beyond the observed levels is thus possible, although you might not be able to articulate what the realizations of the random effects represent.

A consequence of having random effects in your model is that some observations are no longer uncorrelated but instead have a covariance that depends on the variance of the random effect. In fact, in some modeling applications random effects might be used not only to model heterogeneity in the parameters of a model, but also to induce correlations among observations. The typical assumption about random effects in SAS/STAT software is that the effects are normally distributed.

For more information about mixed modeling tools in SAS/STAT software, see Chapter 6, "Introduction to Mixed Modeling Procedures."

Generalized Linear Models

A class of models that has gained increasing importance in the past several decades is the class of generalized linear models. The theory of generalized linear models originated with Nelder and Wedderburn (1972); Wedderburn (1974), and was subsequently made popular in the monograph by McCullagh and Nelder (1989). This class of models extends the theory and methods of linear models to data with nonnormal responses. Before this theory was developed, modeling of nonnormal data typically relied on transformations of the data, and the transformations were chosen to improve symmetry, homogeneity of variance, or normality. Such transformations have to be performed with care because they also have implications for the error structure of the model, Also, back-transforming estimates or predicted values can introduce bias.

Generalized linear models also apply a transformation, known as the *link function*, but it is applied to a deterministic component, the mean of the data. Furthermore, generalized linear models take the distribution of the data into account, rather than assuming that a transformation of the data leads to normally distributed data to which standard linear modeling techniques can be applied.

To put this generalization in place requires a slightly more sophisticated model setup than that required for linear models for normal data:

- The *systematic* component is a linear predictor similar to that in linear models, $\eta = \mathbf{x}'\boldsymbol{\beta}$. The linear predictor is a linear function in the parameters. In contrast to the linear model, η does not represent the mean function of the data.
- The *link function* $g(\cdot)$ relates the linear predictor to the mean, $g(\mu) = \eta$. The link function is a monotonic, invertible function. The mean can thus be expressed as the inversely linked linear predictor, $\mu = g^{-1}(\eta)$. For example, a common link function for binary and binomial data is the logit link, $g(t) = \log\{t/(1-t)\}$. The mean function of a generalized linear model with logit link and a single regressor can thus be written as

$$\log\left\{\frac{\mu}{1-\mu}\right\} = \beta_0 + \beta_1 x$$

$$\mu = \frac{1}{1+\exp\{-\beta_0 - \beta_1 x\}}$$

  This is known as a logistic regression model.
- The *random component* of a generalized linear model is the distribution of the data, assumed to be a member of the exponential family of distributions. Discrete members of this family include the Bernoulli (binary), binomial, Poisson, geometric, and negative binomial (for a given value of the scale parameter) distribution. Continuous members include the normal (Gaussian), beta, gamma, inverse Gaussian, and exponential distribution.

The standard linear model with normally distributed error is a special case of a generalized linear model; the link function is the identity function and the distribution is normal.

Latent Variable Models

Latent variable modeling involves variables that are not observed directly in your research. It has a relatively long history, dating back from the measure of general intelligence by common factor analysis (Spearman 1904) to the emergence of modern-day structural equation modeling (Jöreskog 1973; Keesling 1972; Wiley 1973).

Latent variables are involved in almost all kinds of regression models. In a broad sense, all additive error terms in regression models are latent variables simply because they are not measured in research. Hereafter, however, a narrower sense of latent variables is used when referring to latent variable models. Latent variables are *systematic* unmeasured variables that are also referred to as *factors*. For example, in the following diagram a simple relation between Emotional Intelligence and Career Achievement is shown:

In the diagram, both Emotional Intelligence and Career Achievement are treated as latent factors. They are hypothetical constructs in your model. You hypothesize that Emotional Intelligence is a "causal factor" or predictor of Career Achievement. The symbol β represents the regression coefficient or the effect of Emotional Intelligence on Career Achievement. However, the "causal relationship" or prediction is not perfect. There is an error term δ, which accounts for the unsystematic part of the prediction. You can represent the preceding diagram by using the following linear equation:

$$\mathrm{CA} = \beta \mathrm{EI} + \delta$$

where CA represents Career Achievement and EI represents Emotional Intelligence. The means of the latent factors in the linear model are arbitrary, and so they are assumed to be zero. The error variable δ also has a zero mean with an unknown variance. This equation represents the so-called "structural model," where the "true" relationships among latent factors are theorized.

In order to model this theoretical model with latent factors, some observed variables must somehow relate to these factors. This calls for the measurement models for latent factors. For example, Emotional Intelligence could be measured by some established tests. In these tests, individuals are asked to respond to certain special situations that involve stressful decision making, personal confrontations, and so on. Their responses to these situations are then rated by experts or a standardized scoring system. Suppose there are three such tests and the test scores are labeled as X1, X2 and X3, respectively. The measurement model for the latent factor Emotional Intelligence is specified as follows:

$$\begin{aligned} \mathrm{X1} &= a_1 \mathrm{EI} + e_1 \\ \mathrm{X2} &= a_2 \mathrm{EI} + e_2 \\ \mathrm{X3} &= a_3 \mathrm{EI} + e_3 \end{aligned}$$

where a_1, a_2, and a_3 are regression coefficients and e_1, e_2, and e_3 are measurement errors. Measurement errors are assumed to be independent of the latent factors EI and CA. In the measurement model, X1, X2, and X3 are called the indicators of the latent variable EI. These observed variables are assumed to be centered in the model, and therefore no intercept terms are needed. Each of the indicators is a scaled measurement of the latent factor EI plus a unique error term.

Similarly, you need to have a measurement model for the latent factor CA. Suppose that there are four observed indicators Y1, Y2, Y3, and Y4 (for example, Job Status) for this latent factor. The measurement model for CA is specified as follows:

$$\begin{aligned} \mathrm{Y1} &= a_4 \mathrm{CA} + e_4 \\ \mathrm{Y2} &= a_5 \mathrm{CA} + e_5 \\ \mathrm{Y3} &= a_6 \mathrm{CA} + e_6 \\ \mathrm{Y4} &= a_7 \mathrm{CA} + e_7 \end{aligned}$$

where a_4, a_5, a_6, and a_7 are regression coefficients and e_4, e_5, e_6, and e_7 are error terms. Again, the error terms are assumed to be independent of the latent variables EI and CA, and Y1, Y2, Y3, and Y4 are centered in the equations.

Given the data for the measured variables, you analyze the structural and measurement models simultaneously by the structural equation modeling techniques. In other words, estimation of β, a_1–a_7, and other parameters in the model are carried out simultaneously in the modeling.

Modeling involving the use of latent factors is quite common in social and behavioral sciences, personality assessment, and marketing research. Hypothetical constructs, although not observable, are very important in building theories in these areas.

Another use of latent factors in modeling is to "purify" the predictors in regression analysis. A common assumption in linear regression models is that predictors are measured without errors. That is, in the following linear equation x is assumed to have been measured without errors:

$$y = \alpha + \beta x + \epsilon$$

However, if x has been contaminated with measurement errors that cannot be ignored, the estimate of β might be biased severely so that the true relationship between x and y would be masked.

A measurement model for x provides a solution to such a problem. Let F_x be a "purified" version of x. That is, F_x is the "true" measure of x without measurement errors, as described in the following equation:

$$x = F_x + \delta$$

where δ represents a random measurement error term. Now, the linear relationship of interest is specified in the following new linear regression equation:

$$y = \alpha + \beta F_x + \epsilon$$

In this equation, F_x, which is now free from measurement errors, replaces x in the original equation. With measurement errors taken into account in the simultaneous fitting of the measurement and the new regression equations, estimation of β is unbiased; hence it reflects the true relationship much better.

Certainly, introducing latent factors in models is not a "free lunch." You must pay attention to the identification issues induced by the latent variable methodology. That is, in order to estimate the parameters in structural equation models with latent variables, you must set some identification constraints in these models. There are some established rules or conventions that would lead to proper model identification and estimation. See Chapter 17, "Introduction to Structural Equation Modeling with Latent Variables," for examples and general details.

In addition, because of the nature of latent variables, estimation in structural equation modeling with latent variables does not follow the same form as that of linear regression analysis. Instead of defining the estimators in terms of the data matrices, most estimation methods in structural equation modeling use the fitting of the first- and second- order moments. Hence, estimation principles described in the section "Classical Estimation Principles" on page 37 do not apply to structural equation modeling. However, you can see the section "Estimation Criteria" on page 1695 in Chapter 30, "The CALIS Procedure," for details about estimation in structural equation modeling with latent variables.

Bayesian Models

Statistical models based on the classical (or *frequentist*) paradigm treat the parameters of the model as fixed, unknown constants. They are not random variables, and the notion of probability is derived in an objective sense as a limiting relative frequency. The Bayesian paradigm takes a different approach. Model parameters are random variables, and the probability of an event is defined in a subjective sense as the degree to which you believe that the event is true. This fundamental difference in philosophy leads to profound differences in the statistical content of estimation and inference. In the frequentist framework, you use the data to best estimate the unknown value of a parameter; you are trying to pinpoint a value in the parameter space as well as possible. In the Bayesian framework, you use the data to update your beliefs about the *behavior* of the parameter to assess its distributional properties as well as possible.

Suppose you are interested in estimating θ from data $\mathbf{Y} = [Y_1, \ldots, Y_n]$ by using a statistical model described by a density $p(\mathbf{y}|\theta)$. Bayesian philosophy states that θ cannot be determined exactly, and uncertainty about the parameter is expressed through probability statements and distributions. You can say, for example, that θ follows a normal distribution with mean 0 and variance 1, if you believe that this distribution best describes the uncertainty associated with the parameter.

The following steps describe the essential elements of Bayesian inference:

1. A probability distribution for θ is formulated as $\pi(\theta)$, which is known as the *prior* distribution, or just the prior. The prior distribution expresses your beliefs, for example, on the mean, the spread, the skewness, and so forth, about the parameter prior to examining the data.
2. Given the observed data $\mathbf{Y}$, you choose a statistical model $p(\mathbf{y}|\theta)$ to describe the distribution of $\mathbf{Y}$ given θ.
3. You update your beliefs about θ by combining information from the prior distribution and the data through the calculation of the *posterior* distribution, $p(\theta|\mathbf{y})$.

The third step is carried out by using Bayes' theorem, from which this branch of statistical philosophy derives its name. The theorem enables you to combine the prior distribution and the model in the following way:

$$p(\theta|\mathbf{y}) = \frac{p(\theta, \mathbf{y})}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\theta)\pi(\theta)}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\theta)\pi(\theta)}{\int p(\mathbf{y}|\theta)\pi(\theta)\, d\theta}$$

The quantity $p(\mathbf{y}) = \int p(\mathbf{y}|\theta)\pi(\theta)\, d\theta$ is the normalizing constant of the posterior distribution. It is also the marginal distribution of $\mathbf{Y}$, and it is sometimes called the marginal distribution of the data.

The likelihood function of θ is any function proportional to $p(\mathbf{y}|\theta)$—that is, $L(\theta) \propto p(\mathbf{y}|\theta)$. Another way of writing Bayes' theorem is

$$p(\theta|\mathbf{y}) = \frac{L(\theta)\pi(\theta)}{\int L(\theta)\pi(\theta)\, d\theta}$$

The marginal distribution $p(\mathbf{y})$ is an integral; therefore, provided that it is finite, the particular value of the integral does not yield any additional information about the posterior distribution. Hence, $p(\theta|\mathbf{y})$ can be written up to an arbitrary constant, presented here in proportional form, as

$$p(\theta|\mathbf{y}) \propto L(\theta)\pi(\theta)$$

Bayes' theorem instructs you how to update existing knowledge with new information. You start from a prior belief $\pi(\theta)$, and, after learning information from data $\mathbf{y}$, you change or update the belief on θ and obtain $p(\theta|\mathbf{y})$. These are the essential elements of the Bayesian approach to data analysis.

In theory, Bayesian methods offer a very simple alternative to statistical inference—all inferences follow from the posterior distribution $p(\theta|\mathbf{y})$. However, in practice, only the most elementary problems enable you to obtain the posterior distribution analytically. Most Bayesian analyses require sophisticated computations, including the use of simulation methods. You generate samples from the posterior distribution and use these samples to estimate the quantities of interest.

Both Bayesian and classical analysis methods have their advantages and disadvantages. Your choice of method might depend on the goals of your data analysis. If prior information is available, such as in the form of expert opinion or historical knowledge, and you want to incorporate this information into the analysis, then you might consider Bayesian methods. In addition, if you want to communicate your findings in terms of probability notions that can be more easily understood by nonstatisticians, Bayesian methods might be appropriate. The Bayesian paradigm can provide a framework for answering specific scientific questions that a single point estimate cannot sufficiently address. On the other hand, if you are interested in estimating parameters and in formulating inferences based on the properties of the parameter estimators, then there is no need to use Bayesian analysis. When the sample size is large, Bayesian inference often provides results for parametric models that are very similar to the results produced by classical, frequentist methods.

For more information, see Chapter 7, "Introduction to Bayesian Analysis Procedures."

Classical Estimation Principles

An estimation principle captures the set of rules and procedures by which parameter estimates are derived. When an estimation principle "meets" a statistical model, the result is an *estimation problem*, the solution of which are the parameter estimates. For example, if you apply the estimation principle of least squares to the SLR model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, the estimation problem is to find those values $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that minimize

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

The solutions are the least squares estimators.

The two most important classes of estimation principles in statistical modeling are the least squares principle and the likelihood principle. All principles have in common that they provide a metric by which you measure the distance between the data and the model. They differ in the nature of the metric; least squares relies on a geometric measure of distance, while likelihood inference is based on a distance that measures plausibility.

Least Squares

The idea of the ordinary least squares (OLS) principle is to choose parameter estimates that minimize the squared distance between the data and the model. In terms of the general, additive model,

$$Y_i = f(x_{i1}, \ldots, x_{ik}; \beta_1, \ldots, \beta_p) + \epsilon_i$$

the OLS principle minimizes

$$\text{SSE} = \sum_{i=1}^{n} \left(y_i - f(x_{i1}, \ldots, x_{ik}; \beta_1, \ldots, \beta_p)\right)^2$$

The least squares principle is sometimes called "nonparametric" in the sense that it does not require the distributional specification of the response or the error term, but it might be better termed "distributionally agnostic." In an additive-error model it is only required that the model errors have zero mean. For example, the specification

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \\ \mathrm{E}[\epsilon_i] &= 0 \end{aligned}$$

is sufficient to derive ordinary least squares (OLS) estimators for β_0 and β_1 and to study a number of their properties. It is easy to show that the OLS estimators in this SLR model are

$$\begin{aligned} \widehat{\beta}_1 &= \left(\sum_{i=1}^{n} (Y_i - \overline{Y}) \sum_{i=1}^{n} (x_i - \overline{x}) \right) \Big/ \sum_{i=1}^{n} (x_i - \overline{x})^2 \\ \widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{x} \end{aligned}$$

Based on the assumption of a zero mean of the model errors, you can show that these estimators are unbiased, $\mathrm{E}[\widehat{\beta}_1] = \beta_1$, $\mathrm{E}[\widehat{\beta}_0] = \beta_0$. However, without further assumptions about the distribution of the ϵ_i, you cannot derive the variability of the least squares estimators or perform statistical inferences such as hypothesis tests or confidence intervals. In addition, depending on the distribution of the ϵ_i, other forms of least squares estimation can be more efficient than OLS estimation.

The conditions for which ordinary least squares estimation is efficient are zero mean, homoscedastic, uncorrelated model errors. Mathematically,

$$\begin{aligned} \mathrm{E}[\epsilon_i] &= 0 \\ \mathrm{Var}[\epsilon_i] &= \sigma^2 \\ \mathrm{Cov}[\epsilon_i, \epsilon_j] &= 0 \text{ if } i \neq j \end{aligned}$$

The second and third assumption are met if the errors have an iid distribution—that is, if they are independent and identically distributed. Note, however, that the notion of stochastic independence is stronger than that of absence of correlation. Only if the data are normally distributed does the latter imply the former.

The various other forms of the least squares principle are motivated by different extensions of these assumptions in order to find more efficient estimators.

Weighted Least Squares

The objective function in weighted least squares (WLS) estimation is

$$\mathrm{SSE}_w = \sum_{i=1}^{n} w_i \left(Y_i - f(x_{i1}, \ldots, x_{ik}; \beta_1, \ldots, \beta_p) \right)^2$$

where w_i is a weight associated with the ith observation. A situation where WLS estimation is appropriate is when the errors are uncorrelated but not homoscedastic. If the weights for the observations are proportional to the reciprocals of the error variances, $\mathrm{Var}[\epsilon_i] = \sigma^2 / w_i$, then the weighted least squares estimates are best linear unbiased estimators (BLUE). Suppose that the weights w_i are collected in the diagonal matrix $\mathbf{W}$ and that the mean function has the form of a linear model. The weighted sum of squares criterion then can be written as

$$\mathrm{SSE}_w = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{W} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

which gives rise to the weighted normal equations

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

The resulting WLS estimator of $\boldsymbol{\beta}$ is

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

Iteratively Reweighted Least Squares

If the weights in a least squares problem depend on the parameters, then a change in the parameters also changes the weight structure of the model. Iteratively reweighted least squares (IRLS) estimation is an iterative technique that solves a series of weighted least squares problems, where the weights are recomputed between iterations. IRLS estimation can be used, for example, to derive maximum likelihood estimates in generalized linear models.

Generalized Least Squares

The previously discussed least squares methods have in common that the observations are assumed to be uncorrelated—that is, $\mathrm{Cov}[\epsilon_i, \epsilon_j] = 0$, whenever $i \neq j$. The weighted least squares estimation problem is a special case of a more general least squares problem, where the model errors have a general covariance matrix, $\mathrm{Var}[\boldsymbol{\epsilon}] = \boldsymbol{\Sigma}$. Suppose again that the mean function is linear, so that the model becomes

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \sim (\mathbf{0}, \boldsymbol{\Sigma})$$

The generalized least squares (GLS) principle is to minimize the generalized error sum of squares

$$\mathrm{SSE}_g = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

This leads to the generalized normal equations

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

and the GLS estimator

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

Obviously, WLS estimation is a special case of GLS estimation, where $\boldsymbol{\Sigma} = \sigma^2\mathbf{W}^{-1}$—that is, the model is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \sim \left(\mathbf{0}, \sigma^2\mathbf{W}^{-1}\right)$$

Likelihood

There are several forms of likelihood estimation and a large number of offshoot principles derived from it, such as pseudo-likelihood, quasi-likelihood, composite likelihood, etc. The basic likelihood principle is *maximum likelihood*, which asks to estimate the model parameters by those quantities that maximize the likelihood function of the data. The likelihood function is the joint distribution of the data, but in contrast to a probability mass or density function, it is thought of as a function of the parameters, given the data. The heuristic appeal of the maximum likelihood estimates (MLE) is that these are the values that make the observed data "most likely." Especially for discrete response data, the value of the likelihood function is the ordinate of a probability mass function, even if the likelihood is not a probability function. Since a statistical

model is thought of as a representation of the data-generating mechanism, what could be more preferable as parameter estimates than those values that make it most likely that the data at hand will be observed?

Maximum likelihood estimates, if they exist, have appealing statistical properties. Under fairly mild conditions, they are best-asymptotic-normal (BAN) estimates—that is, their asymptotic distribution is normal, and no other estimator has a smaller asymptotic variance. However, their statistical behavior in finite samples is often difficult to establish, and you have to appeal to the asymptotic results that hold as the sample size tends to infinity. For example, maximum likelihood estimates are often biased estimates and the bias disappears as the sample size grows. A famous example is random sampling from a normal distribution. The corresponding statistical model is

$$
\begin{aligned}
Y_i &= \mu + \epsilon_i \\
\epsilon_i &\sim \text{iid} N(0, \sigma^2)
\end{aligned}
$$

where the symbol $\sim$ is read as "is distributed as" and iid is read as "independent and identically distributed." Under the normality assumption, the density function of y_i is

$$
f(y_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2} \left(\frac{y_i - \mu}{\sigma} \right)^2 \right\}
$$

and the likelihood for a random sample of size n is

$$
L(\mu, \sigma^2; \mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2} \left(\frac{y_i - \mu}{\sigma} \right)^2 \right\}
$$

Maximizing the likelihood function $L(\mu, \sigma^2; \mathbf{y})$ is equivalent to maximizing the log-likelihood function $\log L = l(\mu, \sigma^2; \mathbf{y})$,

$$
\begin{aligned}
l(\mu, \sigma^2; \mathbf{y}) &= \sum_{i=1}^{n} -\frac{1}{2} \left(\log\{2\pi\} + \frac{(y_i - \mu)^2}{\sigma^2} + \log\{\sigma^2\} \right) \\
&= -\frac{1}{2} \left(n \log\{2\pi\} + n \log\{\sigma^2\} + \sum_{i=1}^{n} (y_i - \mu)^2 / \sigma^2 \right)
\end{aligned}
$$

The maximum likelihood estimators of μ and σ^2 are thus

$$
\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} Y_i = \overline{Y} \qquad \widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{\mu})^2
$$

The MLE of the mean μ is the sample mean, and it is an unbiased estimator of μ. However, the MLE of the variance σ^2 is not an unbiased estimator. It has bias

$$
\mathrm{E}\left[\widehat{\sigma}^2 - \sigma^2 \right] = -\frac{1}{n} \sigma^2
$$

As the sample size n increases, the bias vanishes.

For certain classes of models, special forms of likelihood estimation have been developed to maintain the appeal of likelihood-based statistical inference and to address specific properties that are believed to be shortcomings:

- The bias in maximum likelihood parameter estimators of variances and covariances has led to the development of restricted (or residual) maximum likelihood (REML) estimators that play an important role in mixed models.
- Quasi-likelihood methods do not require that the joint distribution of the data be specified. These methods derive estimators based on only the first two moments (mean and variance) of the joint distributions and play an important role in the analysis of correlated data.
- The idea of composite likelihood is applied in situations where the likelihood of the vector of responses is intractable but the likelihood of components or functions of the full-data likelihood are tractable. For example, instead of the likelihood of $\mathbf{Y}$, you might consider the likelihood of pairwise differences $Y_i - Y_j$.
- The pseudo-likelihood concept is also applied when the likelihood function is intractable, but the likelihood of a related, simpler model is available. An important difference between quasi-likelihood and pseudo-likelihood techniques is that the latter make distributional assumptions to obtain a likelihood function in the pseudo-model. Quasi-likelihood methods do not specify the distributional family.
- The penalized likelihood principle is applied when additional constraints and conditions need to be imposed on the parameter estimates or the resulting model fit. For example, you might augment the likelihood with conditions that govern the smoothness of the predictions or that prevent overfitting of the model.

Least Squares or Likelihood

For many statistical modeling problems, you have a choice between a least squares principle and the maximum likelihood principle. Table 3.1 compares these two basic principles.

Table 3.1 Least Squares and Maximum Likelihood

Criterion	Least Squares	Maximum Likelihood
Requires specification of joint distribution of data	No, but in order to perform confirmatory inference (tests, confidence intervals), a distributional assumption is needed, or an appeal to asymptotics.	Yes, no progress can be made with the genuine likelihood principle without knowing the distribution of the data.
All parameters of the model are estimated	No. In the additive-error type models, least squares provides estimates of only the parameters in the mean function. The residual variance, for example, must be estimated by some other method—typically by using the mean squared error of the model.	Yes
Estimates always exist	Yes, but they might not be unique, such as when the $\mathbf{X}$ matrix is singular.	No, maximum likelihood estimates do not exist for all estimation problems.
Estimators are biased	Unbiased, provided that the model is correct—that is, the errors have zero mean.	Often biased, but asymptotically unbiased

Table 3.1 *continued*

Criterion	Least Squares	Maximum Likelihood
Estimators are consistent	Not necessarily, but often true. Sometimes estimators are consistent even in a misspecified model, such as when misspecification is in the covariance structure.	Almost always
Estimators are best linear unbiased estimates (BLUE)	Typically, if the least squares assumptions are met.	Not necessarily: estimators are often nonlinear in the data and are often biased.
Asymptotically most efficient	Not necessarily	Typically
Easy to compute	Yes	No

Inference Principles for Survey Data

Design-based and model-assisted statistical inference for survey data requires that the randomness due to the selection mechanism be taken into account. This can require special estimation principles and techniques.

The SURVEYMEANS, SURVEYFREQ, SURVEYREG, SURVEYLOGISTIC, and SURVEYPHREG procedures support design-based and/or model-assisted inference for sample surveys. Suppose π_i is the selection probability for unit i in sample S. The inverse of the inclusion probability is known as sampling weight and is denoted by w_i. Briefly, the idea is to apply a relationship that exists in the population to the sample and to take into account the sampling weights. For example, to estimate the finite population total $T_N = \sum_{i \in U_N} y_i$ based on the sample S, you can accumulate the sampled values while properly weighting: $\widehat{T}_\pi = \sum_{i \in S} w_i y_i$. It is easy to verify that $\widehat{T}_\pi$ is design-unbiased in the sense that $\mathrm{E}[\widehat{T}_\pi | \mathcal{F}_N] = T_N$ (see Cochran 1977).

When a statistical model is present, similar ideas apply. For example, if β_{N0} and β_{N1} are finite population quantities for a simple linear regression working model that minimize the sum of squares

$$\sum_{i \in U_N} (y_i - \beta_{0N} - \beta_{1N} x_i)^2$$

in the population, then the sample-based estimators $\widehat{\beta}_{0S}$ and $\widehat{\beta}_{1S}$ are obtained by minimizing the weighted sum of squares

$$\sum_{i \in S} w_i (y_i - \hat{\beta}_{0S} - \hat{\beta}_{1S} x_i)^2$$

in the sample, taking into account the inclusion probabilities.

In model-assisted inference, weighted least squares or pseudo-maximum likelihood estimators are commonly used to solve such estimation problems. Maximum pseudo-likelihood or weighted maximum likelihood

estimators for survey data maximize a sample-based estimator of the population likelihood. Assume a working model with uncorrelated responses such that the finite population log-likelihood is

$$\sum_{i \in U_N} l(\theta_{1N}, \ldots, \theta_{pN}; y_i),$$

where $\theta_{1N}, \ldots, \theta_{pN}$ are finite population quantities. For independent sampling, one possible sample-based estimator of the population log likelihood is

$$\sum_{i \in S} w_i l(\theta_{1N}, \ldots, \theta_{pN}; y_i)$$

Sample-based estimators $\widehat{\theta}_{1S}, \ldots, \widehat{\theta}_{pS}$ are obtained by maximizing this expression.

Design-based and model-based statistical analysis might employ the same statistical model (for example, a linear regression) and the same estimation principle (for example, weighted least squares), and arrive at the same estimates. The design-based estimation of the precision of the estimators differs from the model-based estimation, however. For complex surveys, design-based variance estimates are in general different from their model-based counterpart. The SAS/STAT procedures for survey data (SURVEYMEANS, SURVEYFREQ, SURVEYREG, SURVEYLOGISTIC, and SURVEYPHREG procedures) compute design-based variance estimates for complex survey data. See the section "Variance Estimation" on page 254, in Chapter 14, "Introduction to Survey Procedures," for details about design-based variance estimation.

Statistical Background

Hypothesis Testing and Power

In statistical hypothesis testing, you typically express the belief that some effect exists in a population by specifying an alternative hypothesis H_1. You state a null hypothesis H_0 as the assertion that the effect does *not* exist and attempt to gather evidence to reject H_0 in favor of H_1. Evidence is gathered in the form of sample data, and a statistical test is used to assess H_0. If H_0 is rejected but there really is *no* effect, this is called a *Type I error*. The probability of a Type I error is usually designated "alpha" or α, and statistical tests are designed to ensure that α is suitably small (for example, less than 0.05).

If there is an effect in the population but H_0 is *not* rejected in the statistical test, then a *Type II error* has been committed. The probability of a Type II error is usually designated "beta" or β. The probability $1 - \beta$ of avoiding a Type II error—that is, correctly rejecting H_0 and achieving statistical significance, is called the *power* of the test.

An important goal in study planning is to ensure an acceptably high level of power. Sample size plays a prominent role in power computations because the focus is often on determining a sufficient sample size to achieve a certain power, or assessing the power for a range of different sample sizes.

There are several tools available in SAS/STAT software for power and sample size analysis. PROC POWER covers a variety of analyses such as *t* tests, equivalence tests, confidence intervals, binomial proportions, multiple regression, one-way ANOVA, survival analysis, logistic regression, and the Wilcoxon rank-sum test. PROC GLMPOWER supports more complex linear models. The Power and Sample Size application provides a user interface and implements many of the analyses supported in the procedures.

Important Linear Algebra Concepts

A *matrix* $\mathbf{A}$ is a rectangular array of numbers. The *order* of a matrix with n rows and k columns is $(n \times k)$. The element in row i, column j of $\mathbf{A}$ is denoted as a_{ij}, and the notation $\left[a_{ij}\right]$ is sometimes used to refer to the two-dimensional row-column array

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1k} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2k} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nk} \end{bmatrix} = \left[a_{ij}\right]$$

A *vector* is a one-dimensional array of numbers. A *column vector* has a single column ($k = 1$). A *row vector* has a single row ($n = 1$). A *scalar* is a matrix of order (1×1)—that is, a single number. A *square* matrix has the same row and column order, $n = k$. A *diagonal* matrix is a square matrix where all off-diagonal elements are zero, $a_{ij} = 0$ if $i \neq j$. The *identity* matrix $\mathbf{I}$ is a diagonal matrix with $a_{ii} = 1$ for all i. The *unit vector* $\mathbf{1}$ is a vector where all elements are 1. The *unit matrix* $\mathbf{J}$ is a matrix of all 1s. Similarly, the elements of the null vector and the null matrix are all 0.

Basic matrix operations are as follows:

Addition — If $\mathbf{A}$ and $\mathbf{B}$ are of the same order, then $\mathbf{A} + \mathbf{B}$ is the matrix of elementwise sums,

$$\mathbf{A} + \mathbf{B} = \left[a_{ij} + b_{ij}\right]$$

Subtraction — If $\mathbf{A}$ and $\mathbf{B}$ are of the same order, then $\mathbf{A} - \mathbf{B}$ is the matrix of elementwise differences,

$$\mathbf{A} - \mathbf{B} = \left[a_{ij} - b_{ij}\right]$$

Dot product — The dot product of two n-vectors $\mathbf{a}$ and $\mathbf{b}$ is the sum of their elementwise products,

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i$$

The dot product is also known as the *inner product* of $\mathbf{a}$ and $\mathbf{b}$. Two vectors are said to be *orthogonal* if their dot product is zero.

Multiplication — Matrices $\mathbf{A}$ and $\mathbf{B}$ are said to be conformable for $\mathbf{AB}$ multiplication if the number of columns in $\mathbf{A}$ equals the number of rows in $\mathbf{B}$. Suppose that $\mathbf{A}$ is of order $(n \times k)$ and that $\mathbf{B}$ is of order $(k \times p)$. The product $\mathbf{AB}$ is then defined as the $(n \times p)$ matrix of the dot products of the ith row of $\mathbf{A}$ and the jth column of $\mathbf{B}$,

$$\mathbf{AB} = \left[\mathbf{a_i} \cdot \mathbf{b_j}\right]_{n \times p}$$

Transposition — The transpose of the $(n \times k)$ matrix $\mathbf{A}$ is denoted as $\mathbf{A}'$ and is obtained by interchanging the rows and columns,

$$\mathbf{A}' = \begin{bmatrix} a_{11} & a_{21} & a_{31} & \cdots & a_{n1} \\ a_{12} & a_{22} & a_{23} & \cdots & a_{n2} \\ a_{13} & a_{23} & a_{33} & \cdots & a_{n3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{1k} & a_{2k} & a_{3k} & \cdots & a_{nk} \end{bmatrix} = \left[a_{ji}\right]$$

A *symmetric* matrix is equal to its transpose, $\mathbf{A} = \mathbf{A}'$. The inner product of two $(n \times 1)$ column vectors $\mathbf{a}$ and $\mathbf{b}$ is $\mathbf{a} \cdot \mathbf{b} = \mathbf{a}'\mathbf{b}$.

Matrix Inversion

Regular Inverses

The right inverse of a matrix $\mathbf{A}$ is the matrix that yields the identity when $\mathbf{A}$ is postmultiplied by it. Similarly, the left inverse of $\mathbf{A}$ yields the identity if $\mathbf{A}$ is premultiplied by it. $\mathbf{A}$ is said to be invertible and $\mathbf{B}$ is said to be the inverse of $\mathbf{A}$, if $\mathbf{B}$ is its right and left inverse, $\mathbf{BA} = \mathbf{AB} = \mathbf{I}$. This requires $\mathbf{A}$ to be square and nonsingular. The inverse of a matrix $\mathbf{A}$ is commonly denoted as $\mathbf{A}^{-1}$. The following results are useful in manipulating inverse matrices (assuming both $\mathbf{A}$ and $\mathbf{C}$ are invertible):

$$\begin{aligned}
\mathbf{A}\mathbf{A}^{-1} &= \mathbf{A}^{-1}\mathbf{A} = \mathbf{I} \\
\left(\mathbf{A}'\right)^{-1} &= \left(\mathbf{A}^{-1}\right)' \\
\left(\mathbf{A}^{-1}\right)^{-1} &= \mathbf{A} \\
(\mathbf{AC})^{-1} &= \mathbf{C}^{-1}\mathbf{A}^{-1} \\
\operatorname{rank}(\mathbf{A}) &= \operatorname{rank}\left(\mathbf{A}^{-1}\right)
\end{aligned}$$

If $\mathbf{D}$ is a diagonal matrix with nonzero entries on the diagonal—that is, $\mathbf{D} = \operatorname{diag}(d_1, \ldots, d_n)$—then $\mathbf{D}^{-1} = \operatorname{diag}(1/d_1, \ldots, 1/d_n)$. If $\mathbf{D}$ is a block-diagonal matrix whose blocks are invertible, then

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{D}_3 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{D}_n \end{bmatrix} \qquad \mathbf{D}^{-1} = \begin{bmatrix} \mathbf{D}_1^{-1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2^{-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{D}_3^{-1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{D}_n^{-1} \end{bmatrix}$$

In statistical applications the following two results are particularly important, because they can significantly reduce the computational burden in working with inverse matrices.

Partitioned Matrix Suppose $\mathbf{A}$ is a nonsingular matrix that is partitioned as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

Then, provided that all the inverses exist, the inverse of $\mathbf{A}$ is given by

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix}$$

where $\mathbf{B}_{11} = \left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)^{-1}$, $\mathbf{B}_{12} = -\mathbf{B}_{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}$, $\mathbf{B}_{21} = -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{B}_{11}$, and $\mathbf{B}_{22} = \left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)^{-1}$.

Patterned Sum Suppose $\mathbf{R}$ is $(n \times n)$ nonsingular, $\mathbf{G}$ is $(k \times k)$ nonsingular, and $\mathbf{B}$ and $\mathbf{C}$ are $(n \times k)$ and $(k \times n)$ matrices, respectively. Then the inverse of $\mathbf{R} + \mathbf{BGC}$ is given by

$$(\mathbf{R} + \mathbf{BGC})^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{B}\left(\mathbf{G}^{-1} + \mathbf{CR}^{-1}\mathbf{B}\right)^{-1}\mathbf{CR}^{-1}$$

This formula is particularly useful if $k << n$ and $\mathbf{R}$ has a simple form that is easy to invert. This case arises, for example, in mixed models where $\mathbf{R}$ might be a diagonal or block-diagonal matrix, and $\mathbf{B} = \mathbf{C}'$.

Another situation where this formula plays a critical role is in the computation of regression diagnostics, such as in determining the effect of removing an observation from the analysis. Suppose that $\mathbf{A} = \mathbf{X}'\mathbf{X}$ represents the crossproduct matrix in the linear model $\mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$. If $\mathbf{x}_i'$ is the ith row of the $\mathbf{X}$ matrix, then $(\mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}_i')$ is the crossproduct matrix in the same model with the ith observation removed. Identifying $\mathbf{B} = -\mathbf{x}_i$, $\mathbf{C} = \mathbf{x}_i'$, and $\mathbf{G} = \mathbf{I}$ in the preceding inversion formula, you can obtain the expression for the inverse of the crossproduct matrix:

$$\left(\mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}_i'\right)^{-1} = \mathbf{X}'\mathbf{X} + \frac{\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_i\mathbf{x}_i'\left(\mathbf{X}'\mathbf{X}\right)^{-1}}{1 - \mathbf{x}_i'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_i}$$

This expression for the inverse of the reduced data crossproduct matrix enables you to compute "leave-one-out" deletion diagnostics in linear models without refitting the model.

Generalized Inverse Matrices

If $\mathbf{A}$ is rectangular (not square) or singular, then it is not invertible and the matrix $\mathbf{A}^{-1}$ does not exist. Suppose you want to find a solution to simultaneous linear equations of the form

$$\mathbf{A}\mathbf{b} = \mathbf{c}$$

If $\mathbf{A}$ is square and nonsingular, then the unique solution is $\mathbf{b} = \mathbf{A}^{-1}\mathbf{c}$. In statistical applications, the case where $\mathbf{A}$ is $(n \times k)$ rectangular is less important than the case where $\mathbf{A}$ is a $(k \times k)$ square matrix of rank less than k. For example, the normal equations in ordinary least squares (OLS) estimation in the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ are

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

A *generalized inverse* matrix is a matrix $\mathbf{A}^-$ such that $\mathbf{A}^-\mathbf{c}$ is a solution to the linear system. In the OLS example, a solution can be found as $\left(\mathbf{X}'\mathbf{X}\right)^-\mathbf{X}'\mathbf{Y}$, where $\left(\mathbf{X}'\mathbf{X}\right)^-$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$.

The following four conditions are often associated with generalized inverses. For the square or rectangular matrix $\mathbf{A}$ there exist matrices $\mathbf{G}$ that satisfy

$$\begin{array}{lrcl} \text{(i)} & \mathbf{AGA} & = & \mathbf{A} \\ \text{(ii)} & \mathbf{GAG} & = & \mathbf{G} \\ \text{(iii)} & (\mathbf{AG})' & = & \mathbf{AG} \\ \text{(iv)} & (\mathbf{GA})' & = & \mathbf{GA} \end{array}$$

The matrix $\mathbf{G}$ that satisfies all four conditions is unique and is called the Moore-Penrose inverse, after the first published work on generalized inverses by Moore (1920) and the subsequent definition by Penrose (1955). Only the first condition is required, however, to provide a solution to the linear system above.

Pringle and Rayner (1971) introduced a numbering system to distinguish between different types of generalized inverses. A matrix that satisfies only condition (i) is a g_1-inverse. The g_2-inverse satisfies conditions (i) and (ii). It is also called a *reflexive* generalized inverse. Matrices satisfying conditions (i)–(iii) or conditions (i), (ii), and (iv) are g_3-inverses. Note that a matrix that satisfies the first three conditions is a right generalized inverse, and a matrix that satisfies conditions (i), (ii), and (iv) is a left generalized inverse. For example, if $\mathbf{B}$ is $(n \times k)$ of rank k, then $\left(\mathbf{B}'\mathbf{B}\right)^{-1}\mathbf{B}'$ is a left generalized inverse of $\mathbf{B}$. The notation g_4-inverse for the Moore-Penrose inverse, satisfying conditions (i)–(iv), is often used by extension, but note that Pringle and Rayner (1971) do not use it; rather, they call such a matrix "the" generalized inverse.

If the $(n \times k)$ matrix $\mathbf{X}$ is rank-deficient—that is, $\text{rank}(\mathbf{X}) < \min\{n, k\}$—then the system of equations

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

does not have a unique solution. A particular solution depends on the choice of the generalized inverse. However, some aspects of the statistical inference are invariant to the choice of the generalized inverse. If $\mathbf{G}$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$, then $\mathbf{X}\mathbf{G}\mathbf{X}'$ is invariant to the choice of $\mathbf{G}$. This result comes into play, for example, when you are computing predictions in an OLS model with a rank-deficient $\mathbf{X}$ matrix, since it implies that the predicted values

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

are invariant to the choice of $\left(\mathbf{X}'\mathbf{X}\right)^{-}$.

Matrix Differentiation

Taking the derivative of expressions involving matrices is a frequent task in statistical estimation. Objective functions that are to be minimized or maximized are usually written in terms of model matrices and/or vectors whose elements depend on the unknowns of the estimation problem. Suppose that $\mathbf{A}$ and $\mathbf{B}$ are real matrices whose elements depend on the scalar quantities β and θ—that is, $\mathbf{A} = \left[a_{ij}(\beta, \theta)\right]$, and similarly for $\mathbf{B}$.

The following are useful results in finding the derivative of elements of a matrix and of functions involving a matrix. For more in-depth discussion of matrix differentiation and matrix calculus, see, for example, Magnus and Neudecker (1999) and Harville (1997).

The derivative of $\mathbf{A}$ with respect to β is denoted $\dot{\mathbf{A}}_\beta$ and is the matrix of the first derivatives of the elements of $\mathbf{A}$:

$$\dot{\mathbf{A}}_\beta = \frac{\partial}{\partial\beta}\mathbf{A} = \left[\frac{\partial a_{ij}(\beta, \theta)}{\partial\beta}\right]$$

Similarly, the second derivative of $\mathbf{A}$ with respect to β and θ is the matrix of the second derivatives

$$\ddot{\mathbf{A}}_{\beta\theta} = \frac{\partial^2}{\partial\beta\partial\theta}\mathbf{A} = \left[\frac{\partial^2 a_{ij}(\beta, \theta)}{\partial\beta\partial\theta}\right]$$

The following are some basic results involving sums, products, and traces of matrices:

$$\frac{\partial}{\partial\beta}c_1\mathbf{A} = c_1\dot{\mathbf{A}}_\beta$$

$$\frac{\partial}{\partial\beta}(\mathbf{A}+\mathbf{B}) = \dot{\mathbf{A}}_\beta + \dot{\mathbf{B}}_\beta$$

$$\frac{\partial}{\partial\beta}(c_1\mathbf{A}+c_2\mathbf{B}) = c_1\dot{\mathbf{A}}_\beta + c_2\dot{\mathbf{B}}_\beta$$

$$\frac{\partial}{\partial\beta}\mathbf{AB} = \mathbf{A}\dot{\mathbf{B}}_\beta + \dot{\mathbf{A}}_\beta\mathbf{B}$$

$$\frac{\partial}{\partial\beta}\mathrm{trace}(\mathbf{A}) = \mathrm{trace}\left(\dot{\mathbf{A}}_\beta\right)$$

$$\frac{\partial}{\partial\beta}\mathrm{trace}(\mathbf{AB}) = \mathrm{trace}\left(\mathbf{A}\dot{\mathbf{B}}_\beta\right) + \mathrm{trace}\left(\dot{\mathbf{A}}_\beta\mathbf{B}\right)$$

The next set of results is useful in finding the derivative of elements of $\mathbf{A}$ and of functions of $\mathbf{A}$, if $\mathbf{A}$ is a nonsingular matrix:

$$\frac{\partial}{\partial\beta}\mathbf{x}'\mathbf{A}^{-1}\mathbf{x} = -\mathbf{x}'\mathbf{A}^{-1}\dot{\mathbf{A}}_\beta\mathbf{A}^{-1}\mathbf{x}$$

$$\frac{\partial}{\partial\beta}\mathbf{A}^{-1} = -\mathbf{A}^{-1}\dot{\mathbf{A}}_\beta\mathbf{A}^{-1}$$

$$\frac{\partial}{\partial\beta}|\mathbf{A}| = |\mathbf{A}|\,\mathrm{trace}\left(\mathbf{A}^{-1}\dot{\mathbf{A}}_\beta\right)$$

$$\frac{\partial}{\partial\beta}\log\{|\mathbf{A}|\} = \frac{1}{|\mathbf{A}|}\frac{\partial}{\partial\beta}\mathbf{A} = \mathrm{trace}\left(\mathbf{A}^{-1}\dot{\mathbf{A}}_\beta\right)$$

$$\frac{\partial^2}{\partial\beta\partial\theta}\mathbf{A}^{-1} = -\mathbf{A}^{-1}\ddot{\mathbf{A}}_{\beta\theta}\mathbf{A}^{-1} + \mathbf{A}^{-1}\dot{\mathbf{A}}_\beta\mathbf{A}^{-1}\dot{\mathbf{A}}_\theta\mathbf{A}^{-1} + \mathbf{A}^{-1}\dot{\mathbf{A}}_\theta\mathbf{A}^{-1}\dot{\mathbf{A}}_\beta\mathbf{A}^{-1}$$

$$\frac{\partial^2}{\partial\beta\partial\theta}\log\{|\mathbf{A}|\} = \mathrm{trace}\left(\mathbf{A}^{-1}\ddot{\mathbf{A}}_{\beta\theta}\right) - \mathrm{trace}\left(\mathbf{A}^{-1}\dot{\mathbf{A}}_\beta\mathbf{A}^{-1}\dot{\mathbf{A}}_\theta\right)$$

Now suppose that $\mathbf{a}$ and $\mathbf{b}$ are column vectors that depend on β and/or θ and that $\mathbf{x}$ is a vector of constants. The following results are useful for manipulating derivatives of linear and quadratic forms:

$$\frac{\partial}{\partial\mathbf{x}}\mathbf{a}'\mathbf{x} = \mathbf{a}$$

$$\frac{\partial}{\partial\mathbf{x}'}\mathbf{B}\mathbf{x} = \mathbf{B}$$

$$\frac{\partial}{\partial\mathbf{x}}\mathbf{x}'\mathbf{B}\mathbf{x} = \left(\mathbf{B}+\mathbf{B}'\right)\mathbf{x}$$

$$\frac{\partial^2}{\partial\mathbf{x}\partial\mathbf{x}'}\mathbf{x}'\mathbf{B}\mathbf{x} = \mathbf{B}+\mathbf{B}'$$

Matrix Decompositions

To decompose a matrix is to express it as a function—typically a product—of other matrices that have particular properties such as orthogonality, diagonality, triangularity. For example, the Cholesky decomposition of a symmetric positive definite matrix $\mathbf{A}$ is $\mathbf{CC}' = \mathbf{A}$, where $\mathbf{C}$ is a lower-triangular matrix. The spectral decomposition of a symmetric matrix is $\mathbf{A} = \mathbf{PDP}'$, where $\mathbf{D}$ is a diagonal matrix and $\mathbf{P}$ is an orthogonal matrix.

Matrix decomposition play an important role in statistical theory as well as in statistical computations. Calculations in terms of decompositions can have greater numerical stability. Decompositions are often necessary to extract information about matrices, such as matrix rank, eigenvalues, or eigenvectors. Decompositions are also used to form special transformations of matrices, such as to form a "square-root" matrix. This section briefly mentions several decompositions that are particularly prevalent and important.

LDU, LU, and Cholesky Decomposition

Every square matrix $\mathbf{A}$, whether it is positive definite or not, can be expressed in the form $\mathbf{A} = \mathbf{LDU}$, where $\mathbf{L}$ is a unit lower-triangular matrix, $\mathbf{D}$ is a diagonal matrix, and $\mathbf{U}$ is a unit upper-triangular matrix. (The diagonal elements of a unit triangular matrix are 1.) Because of the arrangement of the matrices, the decomposition is called the LDU decomposition. Since you can absorb the diagonal matrix into the triangular matrices, the decomposition

$$\mathbf{A} = \mathbf{L}\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{U} = \mathbf{L}^*\mathbf{U}^*$$

is also referred to as the LU decomposition of $\mathbf{A}$.

If the matrix $\mathbf{A}$ is positive definite, then the diagonal elements of $\mathbf{D}$ are positive and the LDU decomposition is unique. If $\mathbf{A}$ is also symmetric, then the unique decomposition takes the form $\mathbf{A} = \mathbf{U}'\mathbf{DU}$, where $\mathbf{U}$ is unit upper-triangular and $\mathbf{D}$ is diagonal with positive elements. Absorbing the square root of $\mathbf{D}$ into $\mathbf{U}$, $\mathbf{C} = \mathbf{D}^{1/2}\mathbf{U}$, the decomposition is known as the *Cholesky* decomposition of a positive-definite matrix:

$$\mathbf{A} = \mathbf{U}'\mathbf{D}^{1/2}\,\mathbf{D}^{1/2}\mathbf{U} = \mathbf{C}'\mathbf{C}$$

where $\mathbf{C}$ is upper triangular.

If $\mathbf{A}$ is symmetric but only nonnegative definite of rank k, rather than being positive definite of full rank, then it has an extended Cholesky decomposition as follows. Let $\mathbf{C}^*$ denote the lower-triangular matrix such that

$$\mathbf{C}^* = \begin{bmatrix} \mathbf{C}_{k\times k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

Then $\mathbf{A} = \mathbf{CC}'$.

Spectral Decomposition

Suppose that $\mathbf{A}$ is an $(n \times n)$ symmetric matrix. Then there exists an orthogonal matrix $\mathbf{Q}$ and a diagonal matrix $\mathbf{D}$ such that $\mathbf{A} = \mathbf{QDQ}'$. Of particular importance is the case where the orthogonal matrix is also orthonormal—that is, its column vectors have unit norm. Denote this orthonormal matrix as $\mathbf{P}$. Then the corresponding diagonal matrix—$\mathbf{\Lambda} = \mathrm{diag}(\lambda_i, \ldots, \lambda_n)$, say—contains the eigenvalues of $\mathbf{A}$. The spectral decomposition of $\mathbf{A}$ can be written as

$$\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}' = \sum_{i=1}^{n} \lambda_i \mathbf{p}_i \mathbf{p}_i'$$

where $\mathbf{p}_i$ denotes the ith column vector of $\mathbf{P}$. The right-side expression decomposes $\mathbf{A}$ into a sum of rank-1 matrices, and the weight of each contribution is equal to the eigenvalue associated with the ith eigenvector. The sum furthermore emphasizes that the rank of $\mathbf{A}$ is equal to the number of nonzero eigenvalues.

Harville (1997, p. 538) refers to the spectral decomposition of $\mathbf{A}$ as the decomposition that takes the previous sum one step further and accumulates contributions associated with the distinct eigenvalues. If $\lambda_i^*, \ldots, \lambda_k^*$ are the distinct eigenvalues and $\mathbf{E}_j = \sum \mathbf{p}_i \mathbf{p}_i'$, where the sum is taken over the set of columns for which $\lambda_i = \lambda_j^*$, then

$$\mathbf{A} = \sum_{i=1}^{k} \lambda_j^* \mathbf{E}_j$$

You can employ the spectral decomposition of a nonnegative definite symmetric matrix to form a "square-root" matrix of $\mathbf{A}$. Suppose that $\mathbf{\Lambda}^{1/2}$ is the diagonal matrix containing the square roots of the λ_i. Then $\mathbf{B} = \mathbf{P}\mathbf{\Lambda}^{1/2}\mathbf{P}'$ is a square-root matrix of $\mathbf{A}$ in the sense that $\mathbf{BB} = \mathbf{A}$, because

$$\mathbf{BB} = \mathbf{P}\mathbf{\Lambda}^{1/2}\mathbf{P}'\mathbf{P}\mathbf{\Lambda}^{1/2}\mathbf{P}' = \mathbf{P}\mathbf{\Lambda}^{1/2}\mathbf{\Lambda}^{1/2}\mathbf{P}' = \mathbf{P}\mathbf{\Lambda}\mathbf{P}'$$

Generating the Moore-Penrose inverse of a matrix based on the spectral decomposition is also simple. Denote as $\mathbf{\Delta}$ the diagonal matrix with typical element

$$\delta_i = \begin{cases} 1/\lambda_i & \lambda_i \neq 0 \\ 0 & \lambda_i = 0 \end{cases}$$

Then the matrix $\mathbf{P}\mathbf{\Delta}\mathbf{P}' = \sum \delta_i \mathbf{p}_i \mathbf{p}_i'$ is the Moore-Penrose (g_4-generalized) inverse of $\mathbf{A}$.

Singular-Value Decomposition

The singular-value decomposition is related to the spectral decomposition of a matrix, but it is more general. The singular-value decomposition can be applied to any matrix. Let $\mathbf{B}$ be an $(n \times p)$ matrix of rank k. Then there exist orthogonal matrices P and Q of order $(n \times n)$ and $(p \times p)$, respectively, and a diagonal matrix $\mathbf{D}$ such that

$$\mathbf{P}'\mathbf{B}\mathbf{Q} = \mathbf{D} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

where $\mathbf{D}_1$ is a diagonal matrix of order k. The diagonal elements of $\mathbf{D}_1$ are strictly positive. As with the spectral decomposition, this result can be written as a decomposition of $\mathbf{B}$ into a weighted sum of rank-1 matrices

$$\mathbf{B} = -\mathbf{P}\mathbf{D}\mathbf{Q}' = \sum_{i=1}^{n} d_i \mathbf{p}_i \mathbf{q}_i'$$

The scalars $d_1, \ldots, d_k$ are called the *singular values* of the matrix $\mathbf{B}$. They are the positive square roots of the nonzero eigenvalues of the matrix $\mathbf{B}'\mathbf{B}$. If the singular-value decomposition is applied to a symmetric, nonnegative definite matrix $\mathbf{A}$, then the singular values $d_1, \ldots, d_n$ are the nonzero eigenvalues of $\mathbf{A}$ and the singular-value decomposition is the same as the spectral decomposition.

As with the spectral decomposition, you can use the results of the singular-value decomposition to generate the Moore-Penrose inverse of a matrix. If $\mathbf{B}$ is $(n \times p)$ with singular-value decomposition $\mathbf{P}\mathbf{D}\mathbf{Q}'$, and if $\mathbf{\Delta}$ is a diagonal matrix with typical element

$$\delta_i = \begin{cases} 1/d_i & |d_i| \neq 0 \\ 0 & d_i = 0 \end{cases}$$

then $\mathbf{Q}\mathbf{\Delta}\mathbf{P}'$ is the g_4-generalized inverse of $\mathbf{B}$.

Expectations of Random Variables and Vectors

If Y is a discrete random variable with mass function $p(y)$ and support (possible values) $y_1, y_2, \ldots$, then the expectation (expected value) of Y is defined as

$$\mathrm{E}[Y] = \sum_{j=1}^{\infty} y_j \; p(y_j)$$

provided that $\sum |y_j| p(y_j) < \infty$, otherwise the sum in the definition is not well-defined. The expected value of a function $h(y)$ is similarly defined: provided that $\sum |h(y_j)| p(y_j) < \infty$,

$$\mathrm{E}[h(Y)] = \sum_{j=1}^{\infty} h(y_j) \; p(y_j)$$

For continuous random variables, similar definitions apply, but summation is replaced by integration over the support of the random variable. If X is a continuous random variable with density function $f(x)$, and $\int |x| f(x)\, dx < \infty$, then the expectation of X is defined as

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x)\, dx$$

The expected value of a random variable is also called its *mean* or its first moment. A particularly important function of a random variable is $h(Y) = (Y - \mathrm{E}[Y])^2$. The expectation of $h(Y)$ is called the *variance* of Y or the second central moment of Y. When you study the properties of multiple random variables, then you might be interested in aspects of their joint distribution. The covariance between random variables Y and X is defined as the expected value of the function $(Y - \mathrm{E}[Y])(X - \mathrm{E}[X])$, where the expectation is taken under the bivariate joint distribution of Y and X:

$$\mathrm{Cov}[Y, X] = \mathrm{E}[(Y - \mathrm{E}[Y])(X - \mathrm{E}[X])] = \mathrm{E}[YX] - \mathrm{E}[Y]\mathrm{E}[X] = \int \int x\, y f(x, y)\; dx\, dy - \mathrm{E}[Y]\mathrm{E}[X]$$

The *covariance* between a random variable and itself is the variance, $\mathrm{Cov}[Y, Y] = \mathrm{Var}[Y]$.

In statistical applications and formulas, random variables are often collected into vectors. For example, a random sample of size n from the distribution of Y generates a random vector of order $(n \times 1)$,

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

The expected value of the $(n \times 1)$ random vector $\mathbf{Y}$ is the vector of the means of the elements of $\mathbf{Y}$:

$$\mathbf{E}[\mathbf{Y}] = [\mathrm{E}[Y_i]] = \begin{bmatrix} \mathrm{E}[Y_1] \\ \mathrm{E}[Y_2] \\ \vdots \\ \mathrm{E}[Y_n] \end{bmatrix}$$

It is often useful to directly apply rules about working with means, variances, and covariances of random vectors. To develop these rules, suppose that $\mathbf{Y}$ and $\mathbf{U}$ denote two random vectors with typical elements

$Y_1, \ldots, Y_n$ and $U_1, \ldots, U_k$. Further suppose that $\mathbf{A}$ and $\mathbf{B}$ are constant (nonstochastic) matrices, that $\mathbf{a}$ is a constant vector, and that the c_i are scalar constants.

The following rules enable you to derive the mean of a linear function of a random vector:

$$\begin{aligned} \mathrm{E}[\mathbf{A}] &= \mathbf{A} \\ \mathrm{E}[\mathbf{AY} + \mathbf{a}] &= \mathbf{A}\mathrm{E}[\mathbf{Y}] + \mathbf{a} \\ \mathrm{E}[\mathbf{Y} + \mathbf{U}] &= \mathrm{E}[\mathbf{Y}] + \mathrm{E}[\mathbf{U}] \end{aligned}$$

The *covariance matrix* of $\mathbf{Y}$ and $\mathbf{U}$ is the $(n \times k)$ matrix whose typical element in row i, column j is the covariance between Y_i and U_j. The covariance matrix between two random vectors is frequently denoted with the Cov "operator."

$$\begin{aligned} \mathrm{Cov}[\mathbf{Y}, \mathbf{U}] &= \left[\mathrm{Cov}[Y_i, U_j]\right] \\ &= \mathrm{E}\left[(\mathbf{Y} - \mathrm{E}[\mathbf{Y}])\,(\mathbf{U} - \mathrm{E}[\mathbf{U}])'\right] = \mathrm{E}\left[\mathbf{YU}'\right] - \mathrm{E}[\mathbf{Y}]\mathrm{E}[\mathbf{U}]' \\ &= \begin{bmatrix} \mathrm{Cov}[Y_1, U_1] & \mathrm{Cov}[Y_1, U_2] & \mathrm{Cov}[Y_1, U_3] & \cdots & \mathrm{Cov}[Y_1, U_k] \\ \mathrm{Cov}[Y_2, U_1] & \mathrm{Cov}[Y_2, U_2] & \mathrm{Cov}[Y_2, U_3] & \cdots & \mathrm{Cov}[Y_2, U_k] \\ \mathrm{Cov}[Y_3, U_1] & \mathrm{Cov}[Y_3, U_2] & \mathrm{Cov}[Y_3, U_3] & \cdots & \mathrm{Cov}[Y_3, U_k] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[Y_n, U_1] & \mathrm{Cov}[Y_n, U_2] & \mathrm{Cov}[Y_n, U_3] & \cdots & \mathrm{Cov}[Y_n, U_k] \end{bmatrix} \end{aligned}$$

The *variance matrix* of a random vector $\mathbf{Y}$ is the covariance matrix between $\mathbf{Y}$ and itself. The variance matrix is frequently denoted with the Var "operator."

$$\begin{aligned} \mathrm{Var}[\mathbf{Y}] &= \mathrm{Cov}[\mathbf{Y}, \mathbf{Y}] = \left[\mathrm{Cov}[Y_i, Y_j]\right] \\ &= \mathrm{E}\left[(\mathbf{Y} - \mathrm{E}[\mathbf{Y}])\,(\mathbf{Y} - \mathrm{E}[\mathbf{Y}])'\right] = \mathrm{E}\left[\mathbf{YY}'\right] - \mathrm{E}[\mathbf{Y}]\mathrm{E}[\mathbf{Y}]' \\ &= \begin{bmatrix} \mathrm{Cov}[Y_1, Y_1] & \mathrm{Cov}[Y_1, Y_2] & \mathrm{Cov}[Y_1, Y_3] & \cdots & \mathrm{Cov}[Y_1, Y_n] \\ \mathrm{Cov}[Y_2, Y_1] & \mathrm{Cov}[Y_2, Y_2] & \mathrm{Cov}[Y_2, Y_3] & \cdots & \mathrm{Cov}[Y_2, Y_n] \\ \mathrm{Cov}[Y_3, Y_1] & \mathrm{Cov}[Y_3, Y_2] & \mathrm{Cov}[Y_3, Y_3] & \cdots & \mathrm{Cov}[Y_3, Y_n] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[Y_n, Y_1] & \mathrm{Cov}[Y_n, Y_2] & \mathrm{Cov}[Y_n, Y_3] & \cdots & \mathrm{Cov}[Y_n, Y_n] \end{bmatrix} \\ &= \begin{bmatrix} \mathrm{Var}[Y_1] & \mathrm{Cov}[Y_1, Y_2] & \mathrm{Cov}[Y_1, Y_3] & \cdots & \mathrm{Cov}[Y_1, Y_n] \\ \mathrm{Cov}[Y_2, Y_1] & \mathrm{Var}[Y_2] & \mathrm{Cov}[Y_2, Y_3] & \cdots & \mathrm{Cov}[Y_2, Y_n] \\ \mathrm{Cov}[Y_3, Y_1] & \mathrm{Cov}[Y_3, Y_2] & \mathrm{Var}[Y_3] & \cdots & \mathrm{Cov}[Y_3, Y_n] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[Y_n, Y_1] & \mathrm{Cov}[Y_n, Y_2] & \mathrm{Cov}[Y_n, Y_3] & \cdots & \mathrm{Var}[Y_n] \end{bmatrix} \end{aligned}$$

Because the variance matrix contains variances on the diagonal and covariances in the off-diagonal positions, it is also referred to as the *variance-covariance matrix* of the random vector $\mathbf{Y}$.

If the elements of the covariance matrix $\mathrm{Cov}[\mathbf{Y}, \mathbf{U}]$ are zero, the random vectors are uncorrelated. If $\mathbf{Y}$ and $\mathbf{U}$ are normally distributed, then a zero covariance matrix implies that the vectors are stochastically independent.

If the off-diagonal elements of the variance matrix $\mathrm{Var}[\mathbf{Y}]$ are zero, the elements of the random vector $\mathbf{Y}$ are uncorrelated. If $\mathbf{Y}$ is normally distributed, then a diagonal variance matrix implies that its elements are stochastically independent.

Suppose that $\mathbf{A}$ and $\mathbf{B}$ are constant (nonstochastic) matrices and that c_i denotes a scalar constant. The following results are useful in manipulating covariance matrices:

$$\begin{aligned}
\mathrm{Cov}[\mathbf{AY}, \mathbf{U}] &= \mathbf{A}\mathrm{Cov}[\mathbf{Y}, \mathbf{U}] \\
\mathrm{Cov}[\mathbf{Y}, \mathbf{BU}] &= \mathrm{Cov}[\mathbf{Y}, \mathbf{U}]\mathbf{B}' \\
\mathrm{Cov}[\mathbf{AY}, \mathbf{BU}] &= \mathbf{A}\mathrm{Cov}[\mathbf{Y}, \mathbf{U}]\mathbf{B}' \\
\mathrm{Cov}[c_1\mathbf{Y}_1 + c_2\mathbf{U}_1, c_3\mathbf{Y}_2 + c_4\mathbf{U}_2] &= c_1c_3\mathrm{Cov}[\mathbf{Y}_1, \mathbf{Y}_2] + c_1c_4\mathrm{Cov}[\mathbf{Y}_1, \mathbf{U}_2] \\
&\quad + c_2c_3\mathrm{Cov}[\mathbf{U}_1, \mathbf{Y}_2] + c_2c_4\mathrm{Cov}[\mathbf{U}_1, \mathbf{U}_2]
\end{aligned}$$

Since $\mathrm{Cov}[\mathbf{Y}, \mathbf{Y}] = \mathrm{Var}[\mathbf{Y}]$, these results can be applied to produce the following results, useful in manipulating variances of random vectors:

$$\begin{aligned}
\mathrm{Var}[\mathbf{A}] &= \mathbf{0} \\
\mathrm{Var}[\mathbf{AY}] &= \mathbf{A}\mathrm{Var}[\mathbf{Y}]\mathbf{A}' \\
\mathrm{Var}[\mathbf{Y} + \mathbf{x}] &= \mathrm{Var}[\mathbf{Y}] \\
\mathrm{Var}[\mathbf{x}'\mathbf{Y}] &= \mathbf{x}'\mathrm{Var}[\mathbf{Y}]\mathbf{x} \\
\mathrm{Var}[c_1\mathbf{Y}] &= c_1^2\mathrm{Var}[\mathbf{Y}] \\
\mathrm{Var}[c_1\mathbf{Y} + c_2\mathbf{U}] &= c_1^2\mathrm{Var}[\mathbf{Y}] + c_2^2\mathrm{Var}[\mathbf{U}] + 2c_1c_2\mathrm{Cov}[\mathbf{Y}, \mathbf{U}]
\end{aligned}$$

Another area where expectation rules are helpful is quadratic forms in random variables. These forms arise particularly in the study of linear statistical models and in linear statistical inference. Linear inference is statistical inference about linear function of random variables, even if those random variables are defined through nonlinear models. For example, the parameter estimator $\widehat{\boldsymbol{\theta}}$ might be derived in a nonlinear model, but this does not prevent statistical questions from being raised that can be expressed through linear functions of $\boldsymbol{\theta}$; for example,

$$H_0: \begin{cases} \theta_1 - 2\theta_2 = 0 \\ \theta_2 - \theta_3 = 0 \end{cases}$$

if $\mathbf{A}$ is a matrix of constants and $\mathbf{Y}$ is a random vector, then

$$\mathrm{E}[\mathbf{Y}'\mathbf{AY}] = \mathrm{trace}(\mathbf{A}\mathrm{Var}[\mathbf{Y}]) + \mathrm{E}[\mathbf{Y}]'\mathbf{A}\mathrm{E}[\mathbf{Y}]$$

Mean Squared Error

The mean squared error is arguably the most important criterion used to evaluate the performance of a predictor or an estimator. (The subtle distinction between predictors and estimators is that random variables are predicted and constants are estimated.) The mean squared error is also useful to relay the concepts of bias, precision, and accuracy in statistical estimation. In order to examine a mean squared error, you need a target of estimation or prediction, and a predictor or estimator that is a function of the data. Suppose that the

target, whether a constant or a random variable, is denoted as U. The mean squared error of the estimator or predictor $T(\mathbf{Y})$ for U is

$$\mathrm{MSE}\left[T(\mathbf{Y}); U\right] = \mathrm{E}\left[(T(\mathbf{Y}) - U)^2\right]$$

The reason for using a squared difference to measure the "loss" between $T(\mathbf{Y})$ and U is mostly convenience; properties of squared differences involving random variables are more easily examined than, say, absolute differences. The reason for taking an expectation is to remove the randomness of the squared difference by averaging over the distribution of the data.

Consider first the case where the target U is a constant—say, the parameter β—and denote the mean of the estimator $T(\mathbf{Y})$ as μ_T. The mean squared error can then be decomposed as

$$\begin{aligned}\mathrm{MSE}[T(\mathbf{Y}); \beta] &= \mathrm{E}\left[(T(\mathbf{Y}) - \beta)^2\right] \\ &= \mathrm{E}\left[(T(\mathbf{Y}) - \mu_T)^2\right] - \mathrm{E}\left[(\beta - \mu_T)^2\right] \\ &= \mathrm{Var}[T(\mathbf{Y})] + (\beta - \mu_T)^2\end{aligned}$$

The mean squared error thus comprises the variance of the estimator and the squared bias. The two components can be associated with an estimator's precision (small variance) and its accuracy (small bias).

If $T(\mathbf{Y})$ is an unbiased estimator of β—that is, if $\mathrm{E}[T(\mathbf{Y})] = \beta$—then the mean squared error is simply the variance of the estimator. By choosing an estimator that has minimum variance, you also choose an estimator that has minimum mean squared error among all unbiased estimators. However, as you can see from the previous expression, bias is also an "average" property; it is defined as an expectation. It is quite possible to find estimators in some statistical modeling problems that have smaller mean squared error than a minimum variance unbiased estimator; these are estimators that permit a certain amount of bias but improve on the variance. For example, in models where regressors are highly collinear, the ordinary least squares estimator continues to be unbiased. However, the presence of collinearity can induce poor precision and lead to an erratic estimator. Ridge regression stabilizes the regression estimates in this situation, and the coefficient estimates are somewhat biased, but the bias is more than offset by the gains in precision.

When the target U is a random variable, you need to carefully define what an unbiased prediction means. If the statistic and the target have the same expectation, $\mathrm{E}[U] = \mathrm{E}[T(\mathbf{Y})]$, then

$$\mathrm{MSE}\left[T(\mathbf{Y}); U\right] = \mathrm{Var}[T(\mathbf{Y})] + \mathrm{Var}[U] - 2\mathrm{Cov}[T(\mathbf{Y}), U]$$

In many instances the target U is a new observation that was not part of the analysis. If the data are uncorrelated, then it is reasonable to assume in that instance that the new observation is also not correlated with the data. The mean squared error then reduces to the sum of the two variances. For example, in a linear regression model where U is a new observation Y_0 and $T(\mathbf{Y})$ is the regression estimator

$$\widehat{Y}_0 = \mathbf{x}_0' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{Y}$$

with variance $\mathrm{Var}[Y_0] = \sigma^2 \mathbf{x}_0' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{x}_0$, the mean squared prediction error for Y_0 is

$$\mathrm{MSE}\left[\widehat{Y}; Y_0\right] = \sigma^2 \left(\mathbf{x}_0' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{x}_0 + 1\right)$$

and the mean squared prediction error for predicting the mean $\mathrm{E}[Y_0]$ is

$$\mathrm{MSE}\left[\widehat{Y}; \mathrm{E}[Y_0]\right] = \sigma^2 \mathbf{x}_0' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{x}_0$$

Linear Model Theory

This section presents some basic statistical concepts and results for the linear model with homoscedastic, uncorrelated errors in which the parameters are estimated by ordinary least squares. The model can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \sim (0, \sigma^2\mathbf{I})$$

where $\mathbf{Y}$ is an $(n \times 1)$ vector and $\mathbf{X}$ is an $(n \times k)$ matrix of known constants. The model equation implies the following expected values:

$$\mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$$
$$\mathrm{Var}[\mathbf{Y}] = \sigma^2\mathbf{I} \Leftrightarrow \mathrm{Cov}[Y_i, Y_j] = \begin{cases} \sigma^2 & i = j \\ 0 & \text{otherwise} \end{cases}$$

Finding the Least Squares Estimators

Finding the least squares estimator of $\boldsymbol{\beta}$ can be motivated as a calculus problem or by considering the geometry of least squares. The former approach simply states that the OLS estimator is the vector $\widehat{\boldsymbol{\beta}}$ that minimizes the objective function

$$\mathrm{SSE} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

Applying the differentiation rules from the section "Matrix Differentiation" on page 47 leads to

$$\begin{aligned} \frac{\partial}{\partial\boldsymbol{\beta}}\mathrm{SSE} &= \frac{\partial}{\partial\boldsymbol{\beta}}(\mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{0} - 2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}}\mathrm{SSE} &= \mathbf{X}'\mathbf{X} \end{aligned}$$

Consequently, the solution to the *normal equations*, $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$, solves $\frac{\partial}{\partial\boldsymbol{\beta}}\mathrm{SSE} = 0$, and the fact that the second derivative is nonnegative definite guarantees that this solution minimizes SSE. The geometric argument to motivate ordinary least squares estimation is as follows. Assume that $\mathbf{X}$ is of rank k. For any value of $\boldsymbol{\beta}$, such as $\tilde{\boldsymbol{\beta}}$, the following identity holds:

$$\mathbf{Y} = \mathbf{X}\tilde{\boldsymbol{\beta}} + \left(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\right)$$

The vector $\mathbf{X}\tilde{\boldsymbol{\beta}}$ is a point in a k-dimensional subspace of R^n, and the residual $(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$ is a point in an $(n-k)$-dimensional subspace. The OLS estimator is the value $\widehat{\boldsymbol{\beta}}$ that minimizes the distance of $\mathbf{X}\tilde{\boldsymbol{\beta}}$ from $\mathbf{Y}$, implying that $\mathbf{X}\widehat{\boldsymbol{\beta}}$ and $(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$ are orthogonal to each other; that is,

$(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{0}$. This in turn implies that $\widehat{\boldsymbol{\beta}}$ satisfies the normal equations, since

$$\widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} \Leftrightarrow \mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$$

Full-Rank Case

If $\mathbf{X}$ is of full column rank, the OLS estimator is unique and given by

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

The OLS estimator is an unbiased estimator of $\boldsymbol{\beta}$—that is,

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

Note that this result holds if $\mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$; in other words, the condition that the model errors have mean zero is sufficient for the OLS estimator to be unbiased. If the errors are homoscedastic and uncorrelated, the OLS estimator is indeed the *best* linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$—that is, no other estimator that is a linear function of $\mathbf{Y}$ has a smaller mean squared error. The fact that the estimator is unbiased implies that no other linear estimator has a smaller variance. If, furthermore, the model errors are normally distributed, then the OLS estimator has minimum variance among all unbiased estimators of $\boldsymbol{\beta}$, whether they are linear or not. Such an estimator is called a *uniformly minimum variance unbiased estimator*, or UMVUE.

Rank-Deficient Case

In the case of a rank-deficient $\mathbf{X}$ matrix, a generalized inverse is used to solve the normal equations:

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

Although a g_1-inverse is sufficient to solve a linear system, computational expedience and interpretation of the results often dictate the use of a generalized inverse with reflexive properties (that is, a g_2-inverse; see the section "Generalized Inverse Matrices" on page 46 for details). Suppose, for example, that the $\mathbf{X}$ matrix is partitioned as $\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2]$, where $\mathbf{X}_1$ is of full column rank and each column in $\mathbf{X}_2$ is a linear combination of the columns of $\mathbf{X}_1$. The matrix

$$\mathbf{G}_1 = \begin{bmatrix} \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1} & \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{X}_2 \\ -\mathbf{X}_2'\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1} & \mathbf{0} \end{bmatrix}$$

is a g_1-inverse of $\mathbf{X}'\mathbf{X}$ and

$$\mathbf{G}_2 = \begin{bmatrix} \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

is a g_2-inverse. If the least squares solution is computed with the g_1-inverse, then computing the variance of the estimator requires additional matrix operations and storage. On the other hand, the variance of the solution that uses a g_2-inverse is proportional to G_2.

$$\begin{aligned}\mathrm{Var}\left[\mathbf{G}_1\mathbf{X}'\mathbf{Y}\right] &= \sigma^2\mathbf{G}_1\mathbf{X}'\mathbf{X}\mathbf{G}_1 \\ \mathrm{Var}\left[\mathbf{G}_2\mathbf{X}'\mathbf{Y}\right] &= \sigma^2\mathbf{G}_2\mathbf{X}'\mathbf{X}\mathbf{G}_2 = \sigma^2\mathbf{G}_2\end{aligned}$$

If a generalized inverse $\mathbf{G}$ of $\mathbf{X}'\mathbf{X}$ is used to solve the normal equations, then the resulting solution is a biased estimator of $\boldsymbol{\beta}$ (unless $\mathbf{X}'\mathbf{X}$ is of full rank, in which case the generalized inverse is "the" inverse), since $\mathrm{E}\left[\widehat{\boldsymbol{\beta}}\right] = \mathbf{G}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$, which is not in general equal to $\boldsymbol{\beta}$.

If you think of estimation as "estimation without bias," then $\widehat{\boldsymbol{\beta}}$ is the estimator of something, namely $\mathbf{G}\mathbf{X}\boldsymbol{\beta}$. Since this is not a quantity of interest and since it is not unique—it depends on your choice of $\mathbf{G}$—Searle (1971, p. 169) cautions that in the less-than-full-rank case, $\widehat{\boldsymbol{\beta}}$ is a solution to the normal equations and "nothing more."

Analysis of Variance

The identity

$$\mathbf{Y} = \mathbf{X}\tilde{\boldsymbol{\beta}} + \left(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\right)$$

holds for all vectors $\tilde{\boldsymbol{\beta}}$, but only for the least squares solution is the residual $(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$ orthogonal to the predicted value $\mathbf{X}\widehat{\boldsymbol{\beta}}$. Because of this orthogonality, the additive identity holds not only for the vectors themselves, but also for their lengths (Pythagorean theorem):

$$||\mathbf{Y}||^2 = ||\mathbf{X}\widehat{\boldsymbol{\beta}}||^2 + ||(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})||^2$$

Note that $\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$ and note that $\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{H})\mathbf{Y} = \mathbf{M}\mathbf{Y}$. The matrices $\mathbf{H}$ and $\mathbf{M} = \mathbf{I} - \mathbf{H}$ play an important role in the theory of linear models and in statistical computations. Both are *projection* matrices—that is, they are symmetric and idempotent. (An idempotent matrix $\mathbf{A}$ is a square matrix that satisfies $\mathbf{A}\mathbf{A} = \mathbf{A}$. The eigenvalues of an idempotent matrix take on the values 1 and 0 only.) The matrix $\mathbf{H}$ projects onto the subspace of R^n that is spanned by the columns of $\mathbf{X}$. The matrix $\mathbf{M}$ projects onto the orthogonal complement of that space. Because of these properties you have $\mathbf{H}' = \mathbf{H}$, $\mathbf{H}\mathbf{H} = \mathbf{H}$, $\mathbf{M}' = \mathbf{M}$, $\mathbf{M}\mathbf{M} = \mathbf{M}$, $\mathbf{H}\mathbf{M} = \mathbf{0}$.

The Pythagorean relationship now can be written in terms of $\mathbf{H}$ and $\mathbf{M}$ as follows:

$$||\mathbf{Y}||^2 = \mathbf{Y}'\mathbf{Y} = ||\mathbf{H}\mathbf{Y}||^2 + ||\mathbf{M}\mathbf{Y}||^2 = \mathbf{Y}'\mathbf{H}'\mathbf{H}\mathbf{Y} + \mathbf{Y}'\mathbf{M}'\mathbf{M}\mathbf{Y} = \mathbf{Y}'\mathbf{H}\mathbf{Y} + \mathbf{Y}'\mathbf{M}\mathbf{Y}$$

If $\mathbf{X}'\mathbf{X}$ is deficient in rank and a generalized inverse is used to solve the normal equations, then you work instead with the projection matrices $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'$. Note that if $\mathbf{G}$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$, then $\mathbf{X}\mathbf{G}\mathbf{X}'$, and hence also $\mathbf{H}$ and $\mathbf{M}$, are invariant to the choice of $\mathbf{G}$.

The matrix $\mathbf{H}$ is sometimes referred to as the "hat" matrix because when you premultiply the vector of observations with $\mathbf{H}$, you produce the fitted values, which are commonly denoted by placing a "hat" over the $\mathbf{Y}$ vector, $\widehat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$.

The term $\mathbf{Y}'\mathbf{Y}$ is the uncorrected total sum of squares (SST) of the linear model, $\mathbf{Y}'\mathbf{M}\mathbf{Y}$ is the error (residual) sum of squares (SSR), and $\mathbf{Y}'\mathbf{H}\mathbf{Y}$ is the uncorrected model sum of squares. This leads to the analysis of variance table shown in Table 3.2.

Table 3.2 Analysis of Variance with Uncorrected Sums of Squares

Source	df	Sum of Squares
Model	$\mathrm{rank}(\mathbf{X})$	$\mathrm{SSM} = \mathbf{Y}'\mathbf{H}\mathbf{Y} = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$
Residual	$n - \mathrm{rank}(\mathbf{X})$	$\mathrm{SSR} = \mathbf{Y}'\mathbf{M}\mathbf{Y} = \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}\mathbf{X}'\mathbf{Y} = \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$
Uncorr. Total	n	$\mathrm{SST} = \mathbf{Y}'\mathbf{Y} = \sum_{i=1}^{n} Y_i^2$

When the model contains an intercept term, then the analysis of variance is usually corrected for the mean, as shown in Table 3.3.

Table 3.3 Analysis of Variance with Corrected Sums of Squares

Source	df	Sum of Squares
Model	$\mathrm{rank}(\mathbf{X}) - 1$	$\mathrm{SSM}_c = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - n\overline{Y}^2 = \sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}_i\right)^2$
Residual	$n - \mathrm{rank}(\mathbf{X})$	$\mathrm{SSR} = \mathbf{Y}'\mathbf{M}\mathbf{Y} = \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}\mathbf{X}'\mathbf{Y} = \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$
Corrected Total	$n - 1$	$\mathrm{SST}_c = \mathbf{Y}'\mathbf{Y} - n\overline{Y}^2 = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$

The *coefficient of determination*, also called the R-square statistic, measures the proportion of the total variation explained by the linear model. In models with intercept, it is defined as the ratio

$$R^2 = 1 - \frac{\mathrm{SSR}}{\mathrm{SST}_c} = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}$$

In models without intercept, the R-square statistic is a ratio of the uncorrected sums of squares

$$R^2 = 1 - \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}{\sum_{i=1}^{n} Y_i^2}$$

Estimating the Error Variance

The least squares principle does not provide for a parameter estimator for σ^2. The usual approach is to use a method-of-moments estimator that is based on the sum of squared residuals. If the model is correct, then the mean square for error, defined to be SSR divided by its degrees of freedom,

$$\begin{aligned}\widehat{\sigma}^2 &= \frac{1}{n - \mathrm{rank}(\mathbf{X})}\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)\\ &= \mathrm{SSR}/(n - \mathrm{rank}(\mathbf{X}))\end{aligned}$$

is an unbiased estimator of σ^2.

Maximum Likelihood Estimation

To estimate the parameters in a linear model with mean function $\mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$ by maximum likelihood, you need to specify the distribution of the response vector $\mathbf{Y}$. In the linear model with a continuous response variable, it is commonly assumed that the response is normally distributed. In that case, the estimation problem is completely defined by specifying the mean and variance of $\mathbf{Y}$ in addition to the normality assumption. The model can be written as $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, where the notation $N(\mathbf{a}, \mathbf{V})$ indicates a multivariate normal distribution with mean vector $\mathbf{a}$ and variance matrix $\mathbf{V}$. The log likelihood for $\mathbf{Y}$ then can be written as

$$l(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = -\frac{n}{2}\log\{2\pi\} - \frac{n}{2}\log\{\sigma^2\} - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

This function is maximized in $\boldsymbol{\beta}$ when the sum of squares $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ is minimized. The maximum likelihood estimator of $\boldsymbol{\beta}$ is thus identical to the ordinary least squares estimator. To maximize $l(\boldsymbol{\beta}, \sigma^2; \mathbf{y})$

with respect to σ^2, note that

$$\frac{\partial l(\boldsymbol{\beta}, \sigma^2; \mathbf{y})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Hence the MLE of σ^2 is the estimator

$$\begin{aligned} \widehat{\sigma}^2_M &= \frac{1}{n} \left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)' \left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \\ &= \mathrm{SSR}/n \end{aligned}$$

This is a biased estimator of σ^2, with a bias that decreases with n.

Estimable Functions

A function $\mathbf{L}\boldsymbol{\beta}$ is said to be estimable if there exists a linear combination of the expected value of $\mathbf{Y}$, such as $\mathbf{K}\mathrm{E}[\mathbf{Y}]$, that equals $\mathbf{L}\boldsymbol{\beta}$. Since $\mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$, the definition of estimability implies that $\mathbf{L}\boldsymbol{\beta}$ is estimable if there is a matrix $\mathbf{K}$ such that $\mathbf{L} = \mathbf{K}\mathbf{X}$. Another way of looking at this result is that the rows of $\mathbf{X}$ form a generating set from which all estimable functions can be constructed.

The concept of estimability of functions is important in the theory and application of linear models because hypotheses of interest are often expressed as linear combinations of the parameter estimates (for example, hypotheses of equality between parameters, $\beta_1 = \beta_2 \Leftrightarrow \beta_1 - \beta_2 = 0$). Since estimability is not related to the particular value of the parameter estimate, but to the row space of $\mathbf{X}$, you can test only hypotheses that consist of estimable functions. Further, because estimability is not related to the value of $\boldsymbol{\beta}$ (Searle 1971, p. 181), the choice of the generalized inverse in a situation with rank-deficient $\mathbf{X}'\mathbf{X}$ matrix is immaterial, since

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

where $\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}$ is invariant to the choice of generalized inverse.

$\mathbf{L}\boldsymbol{\beta}$ is estimable if and only if $\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}(\mathbf{X}'\mathbf{X}) = \mathbf{L}$ (see, for example, Searle 1971, p. 185). If $\mathbf{X}$ is of full rank, then the *Hermite* matrix $(\mathbf{X}'\mathbf{X})^{-}(\mathbf{X}'\mathbf{X})$ is the identity, which implies that all linear functions are estimable in the full-rank case.

See Chapter 15, "The Four Types of Estimable Functions," for many details about the various forms of estimable functions in SAS/STAT.

Test of Hypotheses

Consider a general linear hypothesis of the form $H: \mathbf{L}\boldsymbol{\beta} = \mathbf{d}$, where $\mathbf{L}$ is a $(k \times p)$ matrix. It is assumed that $\mathbf{d}$ is such that this hypothesis is linearly consistent—that is, that there exists *some* $\boldsymbol{\beta}$ for which $\mathbf{L}\boldsymbol{\beta} = \mathbf{d}$. This is always the case if $\mathbf{d}$ is in the column space of $\mathbf{L}$, if $\mathbf{L}$ has full row rank, or if $\mathbf{d} = \mathbf{0}$; the latter is the most common case. Since many linear models have a rank-deficient $\mathbf{X}$ matrix, the question arises whether the hypothesis is testable. The idea of testability of a hypothesis is—not surprisingly—connected to the concept of estimability as introduced previously. The hypothesis $H: \mathbf{L}\boldsymbol{\beta} = \mathbf{d}$ is testable if it consists of estimable functions.

There are two important approaches to testing hypotheses in statistical applications—the reduction principle and the linear inference approach. The reduction principle states that the validity of the hypothesis can be inferred by comparing a suitably chosen summary statistic between the model at hand and a reduced model in which the constraint $\mathbf{L}\boldsymbol{\beta} = \mathbf{d}$ is imposed. The linear inference approach relies on the fact that $\widehat{\boldsymbol{\beta}}$ is an

estimator of $\boldsymbol{\beta}$ and its stochastic properties are known, at least approximately. A test statistic can then be formed using $\widehat{\boldsymbol{\beta}}$, and its behavior under the restriction $\mathbf{L}\boldsymbol{\beta} = \mathbf{d}$ can be ascertained.

The two principles lead to identical results in certain—for example, least squares estimation in the classical linear model. In more complex situations the two approaches lead to similar but not identical results. This is the case, for example, when weights or unequal variances are involved, or when $\widehat{\boldsymbol{\beta}}$ is a nonlinear estimator.

Reduction Tests

The two main reduction principles are the sum of squares reduction test and the likelihood ratio test. The test statistic in the former is proportional to the difference of the residual sum of squares between the reduced model and the full model. The test statistic in the likelihood ratio test is proportional to the difference of the log likelihoods between the full and reduced models. To fix these ideas, suppose that you are fitting the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Suppose that SSR denotes the residual sum of squares in this model and that SSR_H is the residual sum of squares in the model for which $\mathbf{L}\boldsymbol{\beta} = \mathbf{d}$ holds. Then under the hypothesis the ratio

$$(\mathrm{SSR}_H - \mathrm{SSR})/\sigma^2$$

follows a chi-square distribution with degrees of freedom equal to the rank of $\mathbf{L}$. Maybe surprisingly, the residual sum of squares in the full model is distributed independently of this quantity, so that under the hypothesis,

$$F = \frac{(\mathrm{SSR}_H - \mathrm{SSR})/\mathrm{rank}(\mathbf{L})}{\mathrm{SSR}/(n - \mathrm{rank}(\mathbf{X}))}$$

follows an F distribution with $\mathrm{rank}(\mathbf{L})$ numerator and $n - \mathrm{rank}(\mathbf{X})$ denominator degrees of freedom. Note that the quantity in the denominator of the F statistic is a particular estimator of σ^2—namely, the unbiased moment-based estimator that is customarily associated with least squares estimation. It is also the restricted maximum likelihood estimator of σ^2 if $\mathbf{Y}$ is normally distributed.

In the case of the likelihood ratio test, suppose that $l(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2; \mathbf{y})$ denotes the log likelihood evaluated at the ML estimators. Also suppose that $l(\widehat{\boldsymbol{\beta}}_H, \widehat{\sigma}^2_H; \mathbf{y})$ denotes the log likelihood in the model for which $\mathbf{L}\boldsymbol{\beta} = \mathbf{d}$ holds. Then under the hypothesis the statistic

$$\lambda = 2\left(l(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2; \mathbf{y}) - l(\widehat{\boldsymbol{\beta}}_H, \widehat{\sigma}^2_H; \mathbf{y})\right)$$

follows approximately a chi-square distribution with degrees of freedom equal to the rank of $\mathbf{L}$. In the case of a normally distributed response, the log-likelihood function can be profiled with respect to $\boldsymbol{\beta}$. The resulting profile log likelihood is

$$l(\widehat{\sigma}^2; \mathbf{y}) = -\frac{n}{2}\log\{2\pi\} - \frac{n}{2}\left(\log\{\widehat{\sigma}^2\}\right)$$

and the likelihood ratio test statistic becomes

$$\lambda = n\left(\log\left\{\widehat{\sigma}^2_H\right\} - \log\left\{\widehat{\sigma}^2\right\}\right) = n\left(\log\left\{\mathrm{SSR}_H\right\} - \log\left\{\mathrm{SSR}\right\}\right) = n\left(\log\left\{\mathrm{SSR}_H/\mathrm{SSR}\right\}\right)$$

The preceding expressions show that, in the case of normally distributed data, both reduction principles lead to simple functions of the residual sums of squares in two models. As Pawitan (2001, p. 151) puts it, there is, however, an important difference not in the computations but in the statistical content. The least squares principle, where sum of squares reduction tests are widely used, does not require a distributional specification.

Assumptions about the distribution of the data are added to provide a framework for confirmatory inferences, such as the testing of hypotheses. This framework stems directly from the assumption about the data's distribution, or from the sampling distribution of the least squares estimators. The likelihood principle, on the other hand, requires a distributional specification at the outset. Inference about the parameters is implicit in the model; it is the result of further *computations* following the estimation of the parameters. In the least squares framework, inference about the parameters is the result of further *assumptions*.

Linear Inference

The principle of linear inference is to formulate a test statistic for $H\colon \mathbf{L}\boldsymbol{\beta} = \mathbf{d}$ that builds on the linearity of the hypothesis about $\boldsymbol{\beta}$. For many models that have linear components, the estimator $\mathbf{L}\widehat{\boldsymbol{\beta}}$ is also linear in $\mathbf{Y}$. It is then simple to establish the distributional properties of $\mathbf{L}\widehat{\boldsymbol{\beta}}$ based on the distributional assumptions about $\mathbf{Y}$ or based on large-sample arguments. For example, $\widehat{\boldsymbol{\beta}}$ might be a nonlinear estimator, but it is known to asymptotically follow a normal distribution; this is the case in many nonlinear and generalized linear models.

If the sampling distribution or the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ is normal, then one can easily derive quadratic forms with known distributional properties. For example, if the random vector $\mathbf{U}$ is distributed as $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{U}'\mathbf{A}\mathbf{U}$ follows a chi-square distribution with $\mathrm{rank}(\mathbf{A})$ degrees of freedom and noncentrality parameter $1/2\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$, provided that $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\Sigma} = \mathbf{A}\boldsymbol{\Sigma}$.

In the classical linear model, suppose that $\mathbf{X}$ is deficient in rank and that $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{Y}$ is a solution to the normal equations. Then, if the errors are normally distributed,

$$\widehat{\boldsymbol{\beta}} \sim N\left(\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}, \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-}\right)$$

Because $H\colon \mathbf{L}\boldsymbol{\beta} = \mathbf{d}$ is testable, $\mathbf{L}\boldsymbol{\beta}$ is estimable, and thus $\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X} = \mathbf{L}$, as established in the previous section. Hence,

$$\mathbf{L}\widehat{\boldsymbol{\beta}} \sim N\left(\mathbf{L}\boldsymbol{\beta}, \sigma^2\mathbf{L}\left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{L}'\right)$$

The conditions for a chi-square distribution of the quadratic form

$$(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{d})'\left(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}'\right)^{-}(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{d})$$

are thus met, provided that

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

This condition is obviously met if $\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}'$ is of full rank. The condition is also met if $\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}'^{-}$ is a reflexive inverse (a g_2-inverse) of $\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}$.

The test statistic to test the linear hypothesis $H\colon \mathbf{L}\boldsymbol{\beta} = \mathbf{d}$ is thus

$$F = \frac{(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{d})'\left(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}'\right)^{-}(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{d})/\mathrm{rank}(\mathbf{L})}{\mathrm{SSR}/(n - \mathrm{rank}(\mathbf{X}))}$$

and it follows an F distribution with $\mathrm{rank}(\mathbf{L})$ numerator and $n - \mathrm{rank}(\mathbf{X})$ denominator degrees of freedom under the hypothesis.

This test statistic looks very similar to the F statistic for the sum of squares reduction test. This is no accident. If the model is linear and parameters are estimated by ordinary least squares, then you can show that the quadratic form $(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{d})'\left(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}'\right)^{-}(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{d})$ equals the differences in the residual sum of squares, $\mathrm{SSR}_H - \mathrm{SSR}$, where SSR_H is obtained as the residual sum of squares from OLS estimation in a model that

satisfies $\mathbf{L}\boldsymbol{\beta} = \mathbf{d}$. However, this correspondence between the two test formulations does not apply when a different estimation principle is used. For example, assume that $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{V})$ and that $\boldsymbol{\beta}$ is estimated by generalized least squares:

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

The construction of $\mathbf{L}$ matrices associated with hypotheses in SAS/STAT software is frequently based on the properties of the $\mathbf{X}$ matrix, not of $\mathbf{X}'\mathbf{V}^{-}\mathbf{X}$. In other words, the construction of the $\mathbf{L}$ matrix is governed only by the design. A sum of squares reduction test for $H\colon \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$ that uses the generalized residual sum of squares $(\mathbf{Y} - \widehat{\boldsymbol{\beta}}_g)'\mathbf{V}^{-1}(\mathbf{Y} - \widehat{\boldsymbol{\beta}}_g)$ is not identical to a linear hypothesis test with the statistic

$$F^* = \frac{\widehat{\boldsymbol{\beta}}_g'\mathbf{L}'\left(\mathbf{L}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-}\mathbf{L}'\right)^{-}\mathbf{L}\widehat{\boldsymbol{\beta}}_g}{\mathrm{rank}(\mathbf{L})}$$

Furthermore, $\mathbf{V}$ is usually unknown and must be estimated as well. The estimate for $\mathbf{V}$ depends on the model, and imposing a constraint on the model would change the estimate. The asymptotic distribution of the statistic F^* is a chi-square distribution. However, in practical applications the F distribution with $\mathrm{rank}(\mathbf{L})$ numerator and ν denominator degrees of freedom is often used because it provides a better approximation to the sampling distribution of F^* in finite samples. The computation of the denominator degrees of freedom ν, however, is a matter of considerable discussion. A number of methods have been proposed and are implemented in various forms in SAS/STAT (see, for example, the degrees-of-freedom methods in the MIXED and GLIMMIX procedures).

Residual Analysis

The model errors $\boldsymbol{\epsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$ are unobservable. Yet important features of the statistical model are connected to them, such as the distribution of the data, the correlation among observations, and the constancy of variance. It is customary to diagnose and investigate features of the model errors through the fitted residuals $\widehat{\boldsymbol{\epsilon}} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = \mathbf{M}\mathbf{Y}$. These residuals are projections of the data onto the null space of $\mathbf{X}$ and are also referred to as the "raw" residuals to contrast them with other forms of residuals that are transformations of $\widehat{\boldsymbol{\epsilon}}$. For the classical linear model, the statistical properties of $\widehat{\boldsymbol{\epsilon}}$ are affected by the features of that projection and can be summarized as follows:

$$\begin{aligned} \mathrm{E}[\widehat{\boldsymbol{\epsilon}}] &= \mathbf{0} \\ \mathrm{Var}[\widehat{\boldsymbol{\epsilon}}] &= \sigma^2\mathbf{M} \\ \mathrm{rank}(\mathbf{M}) &= n - \mathrm{rank}(\mathbf{X}) \end{aligned}$$

Furthermore, if $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, then $\widehat{\boldsymbol{\epsilon}} \sim N(\mathbf{0}, \sigma^2\mathbf{M})$.

Because $\mathbf{M} = \mathbf{I} - \mathbf{H}$, and the "hat" matrix $\mathbf{H}$ satisfies $\partial\widehat{\mathbf{Y}}/\partial\mathbf{Y}$, the hat matrix is also the leverage matrix of the model. If h_{ii} denotes the ith diagonal element of $\mathbf{H}$ (the leverage of observation i), then the leverages are bounded in a model with intercept, $1/n \leq h_{ii} \leq 1$. Consequently, the variance of a raw residual is less than that of an observation: $\mathrm{Var}[\widehat{\epsilon}_i] = \sigma^2(1 - h_{ii}) < \sigma^2$. In applications where the variability of the data is estimated from fitted residuals, the estimate is invariably biased low. An example is the computation of an empirical semivariogram based on fitted (detrended) residuals.

More important, the diagonal entries of $\mathbf{H}$ are not necessarily identical; the residuals are heteroscedastic. The "hat" matrix is also not a diagonal matrix; the residuals are correlated. In summary, the only property that the fitted residuals $\widehat{\boldsymbol{\epsilon}}$ share with the model errors is a zero mean. It is thus commonplace to use transformations of the fitted residuals for diagnostic purposes.

Raw and Studentized Residuals

A *standardized* residual is a raw residual that is divided by its standard deviation:

$$\widehat{\epsilon}_i^* = \frac{Y_i - \widehat{Y}_i}{\sqrt{\mathrm{Var}[Y_i - \widehat{Y}_i]}} = \frac{\widehat{\epsilon}_i}{\sqrt{\sigma^2(1-h_{ii})}}$$

Because σ^2 is unknown, residual standardization is usually not practical. A *studentized* residual is a raw residual that is divided by its estimated standard deviation. If the estimate of the standard deviation is based on the same data that were used in fitting the model, the residual is also called an *internally studentized* residual:

$$\widehat{\epsilon}_{is} = \frac{Y_i - \widehat{Y}_i}{\sqrt{\widehat{\mathrm{Var}}[Y_i - \widehat{Y}_i]}} = \frac{\widehat{\epsilon}_i}{\sqrt{\widehat{\sigma}^2(1-h_{ii})}}$$

If the estimate of the residual's variance does not involve the ith observation, it is called an *externally studentized* residual. Suppose that $\widehat{\sigma}^2_{-i}$ denotes the estimate of the residual variance obtained without the ith observation; then the externally studentized residual is

$$\widehat{\epsilon}_{ir} = \frac{\widehat{\epsilon}_i}{\sqrt{\widehat{\sigma}^2_{-i}(1-h_{ii})}}$$

Scaled Residuals

A scaled residual is simply a raw residual divided by a scalar quantity that is not an estimate of the variance of the residual. For example, residuals divided by the standard deviation of the response variable are scaled and referred to as Pearson or Pearson-type residuals:

$$\widehat{\epsilon}_{ic} = \frac{Y_i - \widehat{Y}_i}{\sqrt{\widehat{\mathrm{Var}}[Y_i]}}$$

In generalized linear models, where the variance of an observation is a function of the mean μ and possibly of an extra scale parameter, $\mathrm{Var}[Y] = a(\mu)\phi$, the Pearson residual is

$$\widehat{\epsilon}_{iP} = \frac{Y_i - \widehat{\mu}_i}{\sqrt{a(\widehat{\mu})}}$$

because the sum of the squared Pearson residuals equals the Pearson X^2 statistic:

$$X^2 = \sum_{i=1}^{n} \widehat{\epsilon}_{iP}^2$$

When the scale parameter ϕ participates in the scaling, the residual is also referred to as a Pearson-type residual:

$$\widehat{\epsilon}_{iP} = \frac{Y_i - \widehat{\mu}_i}{\sqrt{a(\widehat{\mu})\phi}}$$

Other Residuals

You might encounter other residuals in SAS/STAT software. A "leave-one-out" residual is the difference between the observed value and the residual obtained from fitting a model in which the observation in question did not participate. If $\widehat{Y}_i$ is the predicted value of the ith observation and $\widehat{Y}_{i,-i}$ is the predicted value if Y_i is removed from the analysis, then the "leave-one-out" residual is

$$\widehat{\epsilon}_{i,-i} = Y_i - \widehat{Y}_{i,-i}$$

Since the sum of the squared "leave-one-out" residuals is the PRESS statistic (prediction sum of squares; Allen 1974), $\widehat{\epsilon}_{i,-i}$ is also called the PRESS residual. The concept of the PRESS residual can be generalized if the deletion residual can be based on the removal of sets of observations. In the classical linear model, the PRESS residual for case deletion has a particularly simple form:

$$\widehat{\epsilon}_{i,-i} = Y_i - \widehat{Y}_{i,-i} = \frac{\widehat{\epsilon}_i}{1 - h_{ii}}$$

That is, the PRESS residual is simply a scaled form of the raw residual, where the scaling factor is a function of the leverage of the observation.

When data are correlated, $\mathrm{Var}[\mathbf{Y}] = \mathbf{V}$, you can scale the vector of residuals rather than scale each residual separately. This takes the covariances among the observations into account. This form of scaling is accomplished by forming the Cholesky root $\mathbf{C}'\mathbf{C} = \mathbf{V}$, where $\mathbf{C}'$ is a lower-triangular matrix. Then $\mathbf{C}'^{-1}\mathbf{Y}$ is a vector of uncorrelated variables with unit variance. The Cholesky residuals in the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ are

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

In generalized linear models, the fit of a model can be measured by the scaled deviance statistic D^*. It measures the difference between the log likelihood under the model and the maximum log likelihood that is achievable. In models with a scale parameter ϕ, the deviance is $D = \phi \times D^* = \sum_{i=1}^{n} d_i$. The deviance residuals are the signed square roots of the contributions to the deviance statistic:

$$\widehat{\epsilon}_{id} = \mathrm{sign}\{y_i - \widehat{\mu}_i\}\sqrt{d_i}$$

Sweep Operator

The sweep operator (Goodnight 1979) is closely related to Gauss-Jordan elimination and the Forward Doolittle procedure. The fact that a sweep operation can produce a generalized inverse by in-place mapping with minimal storage and that its application invariably leads to some form of matrix inversion is important, but this observation does not do justice to the pervasive relevance of sweeping to statistical computing. In this section the sweep operator is discussed as a conceptual tool for further insight into linear model operations. Consider the nonnegative definite, symmetric, partitioned matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}'_{12} & \mathbf{A}_{22} \end{bmatrix}$$

Sweeping a matrix consists of performing a series of row operations akin to Gauss-Jordan elimination. Basic row operations are the multiplication of a row by a constant and the addition of a multiple of one row to another. The sweep operator restricts row operations to pivots on the diagonal elements of a matrix; further

details about the elementary operations can be found in Goodnight (1979). The process of sweeping the matrix $\mathbf{A}$ on its leading partition is denoted as $\text{Sweep}(\mathbf{A}, \mathbf{A}_{11})$ and leads to

$$\text{Sweep}(\mathbf{A}, \mathbf{A}_{11}) = \begin{bmatrix} \mathbf{A}_{11}^{-} & \mathbf{A}_{11}^{-}\mathbf{A}_{12} \\ -\mathbf{A}_{12}'\mathbf{A}_{11}^{-} & \mathbf{A}_{22} - \mathbf{A}_{12}'\mathbf{A}_{11}^{-}\mathbf{A}_{12} \end{bmatrix}$$

If the kth row and column are set to zero when the pivot is zero (or in practice, less than some singularity tolerance), the generalized inverse in the leading position of the swept matrix is a reflexive, g_2-inverse. Suppose that the crossproduct matrix of the linear model is augmented with a "Y-border" as follows:

$$\mathbf{C} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Y} \\ \mathbf{Y}'\mathbf{X} & \mathbf{Y}'\mathbf{Y} \end{bmatrix}$$

Then the result of sweeping on the rows of $\mathbf{X}$ is

$$\begin{aligned} \text{Sweep}(\mathbf{C}, \mathbf{X}) &= \begin{bmatrix} (\mathbf{X}'\mathbf{X})^{-} & (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y} \\ -\mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-} & \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y} \end{bmatrix} \\ &= \begin{bmatrix} (\mathbf{X}'\mathbf{X})^{-} & \widehat{\boldsymbol{\beta}} \\ -\widehat{\boldsymbol{\beta}} & \mathbf{Y}'\mathbf{M}\mathbf{Y} \end{bmatrix} = \begin{bmatrix} (\mathbf{X}'\mathbf{X})^{-} & \widehat{\boldsymbol{\beta}} \\ -\widehat{\boldsymbol{\beta}} & \text{SSR} \end{bmatrix} \end{aligned}$$

The "Y-border" has been transformed into the least squares solution and the residual sum of squares.

Partial sweeps are common in model selection. Suppose that the $\mathbf{X}$ matrix is partitioned as $[\mathbf{X}_1\ \mathbf{X}_2]$, and consider the augmented crossproduct matrix

$$\mathbf{C} = \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 & \mathbf{X}_1'\mathbf{Y} \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 & \mathbf{X}_2'\mathbf{Y} \\ \mathbf{Y}'\mathbf{X}_1 & \mathbf{Y}'\mathbf{X}_2 & \mathbf{Y}'\mathbf{Y} \end{bmatrix}$$

Sweeping on the $\mathbf{X}_1$ partition yields

$$\text{Sweep}(\mathbf{C}, \mathbf{X}_1) = \begin{bmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-} & (\mathbf{X}_1'\mathbf{X}_1)^{-}\mathbf{X}_1'\mathbf{X}_2 & (\mathbf{X}_1'\mathbf{X}_1)^{-}\mathbf{X}_1'\mathbf{Y} \\ -\mathbf{X}_2'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-} & \mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2 & \mathbf{X}_2'\mathbf{M}_1\mathbf{Y} \\ -\mathbf{Y}'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-} & -\mathbf{Y}'\mathbf{M}_1\mathbf{X}_2 & \mathbf{Y}'\mathbf{M}_1\mathbf{Y} \end{bmatrix}$$

where $\mathbf{M}_1 = \mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-}\mathbf{X}_1'$. The entries in the first row of this partition are the generalized inverse of $\mathbf{X}'\mathbf{X}$, the coefficients for regressing $\mathbf{X}_2$ on $\mathbf{X}_1$, and the coefficients for regressing $\mathbf{Y}$ on $\mathbf{X}_1$. The diagonal entries $\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2$ and $\mathbf{Y}'\mathbf{M}_1\mathbf{Y}$ are the sum of squares and crossproduct matrices for regressing $\mathbf{X}_2$ on $\mathbf{X}_1$ and for regressing $\mathbf{Y}$ on $\mathbf{X}_1$, respectively. As you continue to sweep the matrix, the last cell in the partition contains the residual sum of square of a model in which $\mathbf{Y}$ is regressed on all columns swept up to that point.

The sweep operator is not only useful to conceptualize the computation of least squares solutions, Type I and Type II sums of squares, and generalized inverses. It can also be used to obtain other statistical information. For example, adding the logarithms of the pivots of the rows that are swept yields the log determinant of the matrix.

References

Allen, D. M. (1974). "The Relationship between Variable Selection and Data Augmentation and a Method of Prediction." *Technometrics* 16:125–127.

Cochran, W. G. (1977). *Sampling Techniques*. 3rd ed. New York: John Wiley & Sons.

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

Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. New York: Springer-Verlag.

Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag.

Jöreskog, K. G. (1973). "A General Method for Estimating a Linear Structural Equation System." In *Structural Equation Models in the Social Sciences*, edited by A. S. Goldberger and O. D. Duncan, 85–112. New York: Academic Press.

Keesling, J. W. (1972). "Maximum Likelihood Approaches to Causal Analysis." Ph.D. diss., University of Chicago.

Magnus, J. R., and Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. New York: John Wiley & Sons.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd ed. London: Chapman & Hall.

Moore, E. H. (1920). "On the Reciprocal of the General Algebraic Matrix." *Bulletin of the American Mathematical Society* 26:394–395.

Nelder, J. A., and Wedderburn, R. W. M. (1972). "Generalized Linear Models." *Journal of the Royal Statistical Society, Series A* 135:370–384.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford: Clarendon Press.

Penrose, R. A. (1955). "A Generalized Inverse for Matrices." *Proceedings of the Cambridge Philosophical Society* 51:406–413.

Pringle, R. M., and Rayner, A. A. (1971). *Generalized Inverse Matrices with Applications to Statistics*. New York: Hafner Publishing.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Searle, S. R. (1971). *Linear Models*. New York: John Wiley & Sons.

Spearman, C. (1904). "General Intelligence Objectively Determined and Measured." *American Journal of Psychology* 15:201–293.

Wedderburn, R. W. M. (1974). "Quasi-likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method." *Biometrika* 61:439–447.

Wiley, D. E. (1973). "The Identification Problem for Structural Equation Models with Unmeasured Variables." In *Structural Equation Models in the Social Sciences*, edited by A. S. Goldberger and O. D. Duncan, 69–83. New York: Academic Press.

Index