Journal Title: Stevens' Handbook of Experimental Psychology Vol 4

Volume: 
Issue: 
Month/Year: 2004
Pages: 429-460

Article Author: 
Article Title: Mathematical Modeling

Call #: BF181 .H336 2004
Location: evans/ref (1st floor)
Not Wanted Date: 02/19/2006

Status: Faculty
Phone: 845-2503
E-mail: tya@psyc.tamu.edu

Name: Yamauchi, Takashi

Pickup at Evans

Mail Stop 4235
Dept. of Psychology
College Station, TX 77843
CHAPTER 11

Mathematical Modeling

IN JAE MYUNG AND MARK A. PITT

INTRODUCTION

Why Do Mathematical Modeling?

As psychologists, we seek to identify lawful patterns of behavior, infer mental structures and processes underlying these patterns, and develop theories that provide a succinct explanation of the behavior. Within psychology today, most theories exist in a verbal form only. That is, they are a set of statements formulated from observations (i.e., data) about behavior. Verbal modeling is popular in current psychological research for a number of good reasons. First, a verbal model, stated in everyday language, helps readers to grasp the essence of an idea, thereby providing a good conceptual understanding of the phenomenon of interest. Second, verbal modeling is a somewhat conservative approach to knowledge acquisition because specification of the inner workings of the theory does not stretch too far beyond what is, in principle, observable in the data. Such a strategy makes sense when data are scarce or a clear understanding of the phenomenon is lacking, especially in the early stages of research. It is prudent to avoid assuming too much about the

phenomenon for fear of leading the research enterprise astray. Third, considerable mileage can be made with verbal modeling. In conjunction with hypothesis-driven experimentation, a verbal model can be used to identify key variables that affect the phenomenon of interest, such as the influence of word frequency in word recognition and the influence of serial position in memory retrieval.

At its best, a verbal model furnishes testable predictions about the relationship between variables and levels of a variable, while making as few assumptions as possible about the details of the underlying mental process. Verbal modeling can thus lead to good qualitative descriptions of the data, yielding many useful insights into the underlying process.

Although researchers have made many important advances in science and psychology through the use of verbal modeling, a case of diminishing returns can be reached unless the theory is specified in more detail. The lack of precision is a serious shortcoming of verbal modeling (Lewandowsky, 1993; Ratcliff, 1998). Because it is expressed verbally or graphically without making use of explicit mathematical formulations, a verbal model does not provide sufficient information about structural or functional characteristics of the phenomenon being studied. For instance, computational mechanisms may be vaguely specified or sometimes left

The authors thank Barbara Mellers for valuable comments on an earlier version of this chapter. The authors were supported by NIH Grant MH57472.
undefined, making it unclear which predictions follow from the model and which do not (Forster, 1994). Moreover, the lack of precision in verbal modeling opens up the possibility of multiple interpretations of how a model functions, each with distinct predictions, possibly making the model too powerful to test or virtually unfalsifiable (Jacobs & Grainger, 1994). If two competing theories exhibit this property, a form of gridlock can arise, making scientific advancement difficult. Furthermore, although they furnish ordinal information about the effect of interest, most verbal models are silent about its magnitude. That is, verbal models do not specify how exactly the magnitude of an effect may be influenced by other relevant variables. For example, a model of the word frequency effect states that response time in a lexical decision task is a monotonically decreasing function of word frequency: the more frequent a word, the faster the response time. The model, however, is mute on the specific characteristic of the functional relationship between response time and word frequency, such as whether it is linear or nonlinear. From the standpoint of the theory of measurement, variables of verbal models can be specified only on ordinal scales of measurement, not on more precise interval and ratio scales. Consequently, many potentially important questions formulated in terms of magnitude relationships are left untestable.

Mathematical modeling represents an alternative approach that overcomes these limitations (Luce, 1995, 1997; Ratcliff, 1998). Mathematical models, which seek quantitative descriptions of data, attempt to characterize patterns of behavior by directly asking about the form of the underlying mechanism that gives rise to the behavior of interest. These include questions about how stimulus information is represented, what computations are performed on the input, what circuits are involved in information processing, and so on. In mathematical modeling, researchers formulate hypotheses about the underlying mechanisms using closed-form expressions, algorithms, or other simulation procedures, thereby imposing precision and clarity on what is meant. As a result, mathematical modeling enables, even requires, precise predictions be made from the underlying assumptions, which improves the ability to discriminate among models and hypotheses. The virtue of the approach is particularly evident when the predictions and outcomes are not obvious.

As Luce (1995) put it, mathematical modeling is the "opened black box" approach to psychological inquiry, as opposed to the "unopened black box" characteristic of verbal modeling. Because the goal of mathematical modeling is to specify the details in the black box, it may not be the best strategy to use in the early, exploratory stage of research, when a phenomenon is first being investigated. Rather, it is probably most fruitful in the more advanced stage of research, when considerable knowledge has been acquired about the behavior through verbal modeling. Otherwise, one’s quantitative formulation of the process is difficult to justify and most likely will be a poor approximation of the true form of the mental process. The following recent example from the literature demonstrates how research can benefit from mathematical modeling.

A Reinterpretation of Brinley Plots

A Brinley plot (Brinley, 1965) is a plot of mean response time by older adults against mean response time by younger adults on some cognitive task (e.g., mental rotation, memory scanning). A typical observation is that the mean response time for the older adults is slower than that for younger adults. The relationship between these two groups has proven to be very consistent, so much so that the data of older adults can be estimated from the data of younger adults:

$$\text{RT}_{\text{old}} = \alpha \text{RT}_{\text{young}} + \beta$$

where $\beta$, the y-intercept, is positive, and $\alpha$, the slope, is negative. Researchers have fit linear models of this form to data of older subjects and younger subjects, with $\alpha$ reflecting the ratio (i.e., the slope) of RTs of older subjects to RTs of younger subjects. Cicerone and Gray (1983) report a slope of about 1.57, and $\beta = 200$, RTs for young adults would exceed RTs for older adults by about 200 ms; for young adults. Ratcliff and McKoon (2000) showed that the slope of the RT for the task with a slower set size was in the range of about 1.25, suggesting consistency of this finding over a range of tasks. However, the slope was not consistent across-the-board; slopes of the RTs for the older adults.

Although this reinterpretation of the Brinley plot is more convincing, Ratcliff et al. (1991) showed that the old-young association might be more a consequence of the technique itself. The linear regression of analytic methods by Cicerone and Gray (1983) and statistical conditions over time since the plot pattern can also be obtained if the slope was their quantitative measure, which revealed that the slope was not because the distribution of the response time for the older adults is less than that of the younger adults. The intercept is negative, indicating that the response time is more or less identical to RTs of older and younger groups. The linear regression plots derives from the assumption of a linear relationship between the response time for older and younger adults.

Ratcliff et al. (1991) argue that the slopes are more strongly by another technique.
from the data of younger adults by a straightforward linear transformation:

\[ RT_{old} = \alpha RT_{young} - \beta \quad (\alpha > 1, \beta > 0) \quad (1) \]

where \( \beta \), the y intercept, is frequently positive, and \( \alpha \), the slope, often hovers around 1.5. Researchers have sought to explain the cause of this linear relationship, whereby the data of older subjects can be predicted from that of younger subjects through a constant factor (i.e., the slope). For example, for \( \alpha = 1.5 \) and \( \beta = 200 \), \( RT_{old} = 700 \) ms and \( RT_{young} = 600 \) ms would be observed in one cognitive task with a slowing effect of 100 ms. The consistency of this finding has been interpreted as an across-the-board, general slowing of cognitive processes in the elderly.

Although this cognitive-slowing interpretation of the Brinley plot seems clear-cut and convincing, Ratcliff, Spieler, and McKoon (2000) showed that the strong linear association might be an artifact of the analysis technique itself. Using simulations as well as analytic methods, they identified a set of statistical conditions under which the Brinley pattern can be observed. Most enlightening was their quantile-quantile (Q-Q) analysis, which revealed that (a) a Brinley plot is linear because the distributions of responses times for the older adults have about the same shape as that for the younger group, (b) the slope is greater than 1 because the standard deviation of the older group’s response times is greater than that of the younger group’s, and (c) the intercept is negative because the motor response time is more or less the same for both the older and younger groups. The linearity in Brinley plots derives from a constant difference in the variability in response time between the two groups (point b) and is not unambiguous evidence of a static (cognitive-slowing) relationship between the performances of older and younger adults.

Ratcliff et al. (2000) made this point even more strongly by showing that extant models of aging effects, such as Cerella’s (1985) linear model, Meyerson, Hale, Wagstaff, Poon, and Smith’s (1990) information loss model, and even Ratcliff’s (1978) random walk model, can be reinterpreted as being consistent with the Q-Q analysis, making each model capable of accounting for the Brinley pattern of results, though with differing assumptions about the underlying cognitive process. Particularly disturbing is Ratcliff et al.’s (2000) demonstration through simulations that the diffusion model can reproduce the Brinley pattern of results by simply manipulating two parameters (the boundary-position and drift-rate parameters) independently or in combination, even when they were varied in a counterintuitive manner with respect to the effects of aging.

In short, the Brinley pattern can appear on first encounter to provide compelling evidence that performance differences as a function of age are related in a straightforward, linear fashion. Ratcliff et al.’s (2000) investigation shows the pattern can be observed under a variety of modeling assumptions and thus provides only very weak constraints on modeling. Such insights become evident only through quantitative analysis of the data and competing models.

Critical tests of cognitive slowing will require analysis of aspects of the data besides plots of means. Examples of such data are the shape of the response time distributions, the joint relationship between response time and accuracy, and the relative speeds of correct and error responses. Along with these data, researchers need a modeling approach that makes explicit assumptions about the underlying process, generates clearly falsifiable predictions about the shape of the data, and thus provides a means of discriminating among specific hypotheses about the mental process. Mathematical modeling is necessary to address such issues, and the random walk model is one example. For instance, by
varying the appropriate parameters, one can test whether older adults respond slowly because the quality of the information that they use to make decisions is relatively poor or because they set more conservative decision criteria. Comparing the simulated data with the observed data in each case will answer this question. Much more information can be extracted from the data when evaluating quantitative models, allowing the researcher to test many pointed hypotheses, a feat that is not possible in verbal modeling.

Three lessons can be drawn from the above Brinley plot example. First, a reliance on verbal models, while useful initially, can lead to serious misinterpretations of observed patterns of results, because the evidence may be far less constraining than was imagined. Second, a formal mathematical analysis of the problem in question can provide a clearer picture of the theoretical constraints, revealing previously unsuspected insufficiencies in the model and in data interpretation. Third, mathematical modeling avoids the pitfalls of verbal modeling by formally expressing the details of the model and then squeezing as much information from the data as possible to test its accuracy, not just its mean performance. The end product is both a deeper understanding of the mental process of interest and the adequacy of one’s approximation to it (Lewandowsky, 1993).

Overview

The aim of this chapter is to serve as an introduction to the field of mathematical modeling, first covering the different types of models and then discussing how to create (i.e., define) a model, how to test it, and how to compare it with competitors. Examples are provided along the way to illustrate points and serve as brief tutorials. The focus throughout most of the chapter is on statistical models, because these are in widespread use and their popularity is growing. For in-depth, technically rigorous treatments of mathematical modeling in psychology, the reader may consult books on specialized topics and conference proceedings (see, e.g., Ashby, 1992; Brown & Keith-Smith, 1991; Dowling, Roberts, & Theuns, 1998; Estes, 1991; Healy, Kosslyn, & Shiffrin, 1992; Luce, 1986; Marley, 1997; McFall & Townsend, 1998; Townsend & Ashby, 1983; Wickens, 1982).

The chapter was written for graduate students who have completed a year of statistics courses. Some readers may be challenged by the technical details in a few sections, but a thorough understanding of the mathematics is not necessary to follow the discussion.

TYPES OF MODELS

Model Building via Regularity Constraints

In this approach, a model is created by constraining it to be subject to certain regularity conditions at the level of behavior. Functional equation models and axiomatic models are of this type.

Functional Equation Models

Rather than making specific assumptions about the form of functions in a mathematical model, this method (Aczel, 1966) makes certain equality restrictions involving an unknown function. The restrictions themselves, however, are so severe that they yield a particular solution as the only possible form for the function. To illustrate this approach, consider models of psychophysics, which aim to capture the relationship between physical scale (e.g., intensity of a tone) and psychological scale (e.g., perceived loudness). Fechner’s logarithmic law and Stevens’s power law are two well-known, well-studied models of psychophysics. Fechner’s law can be derived using a functional equation method. That is, the equation holds for all stimulus strengths.

\[ \psi(x) = k \psi(x) \]

for some unknown function \( \psi \). As \( x \) increases, we observe \( \psi(50) \), \( \psi(25) \), and so on. Functional equation models define the above equation in the form \( \psi(x) = k \psi(x) \), which is just a slightly different way of saying that the model prescribes a monotonic behavior of \( \psi \) over the range of any stimulus over time. Instead of assuming a functional equation as a constraint, \( \psi(x) = k \psi(x) \), this equation gives rise to a model of psychophysics, \( \psi(x) \).

In the above model, all or things under study are assumed to be in a stimulus dimension, which is defined as a discrete (e.g., categorical) set of stimulus strengths, \( \{a, b, c, \ldots\} \), and a set of functions \( \{\psi_a, \psi_b, \psi_c, \ldots\} \) that correspond to each stimulus. As a consequence, we can write the following equation, which is a decision rule of the form that is ambiguous in the context of a stimulus domain.

\[ P(a, b, c) = \begin{cases} P(a, b, c) & \text{if } a \text{ is stimulus } b \text{ is stimulus } c \text{ is stimulus} \\ 0 & \text{otherwise} \end{cases} \]

for all \( a, b, c \). Such equations are satisfied if the product of the probabilities of each stimulus is equal.
be derived using the functional equation method. That is, suppose that the following equation holds for all values \( x \) and \( y \) of physical stimulation strength,

\[
\psi(xy) = \psi(x) + \psi(y)
\]

for some unknown function \( \psi \). For instance, according to the above constraint, we must observe \( \psi(50) = \psi(5) + \psi(10) = \psi(2) + \psi(25) \), and so on, for various pairs of stimulation strengths. It is then shown (F. S. Roberts, 1979, p. 162) that the function that satisfies the above equation must be logarithmic in the form \( \psi(x) = k \ln(x) \) for a constant \( k \), which is just Fechner's law. Note here that the model is derived entirely in terms of overt behavior, with no explicit assumption of any intervening variables or mental processes. Instead of the additive form of the functional equation, if a multiplicative constraint, \( \psi(xy) = \psi(x)\psi(y) \), is imposed, this gives rise to Stevens’s power law of psychophysics, \( \psi(x) = x^k \).

In the above discussion, the set of objects or things under study on which the model equation is defined is continuous. If the set is discrete (e.g., set of people or cities), similar constraints may be used to derive the model equation. As an example, consider a system in which an individual is required to select between a pair of choices, say \( a \) and \( b \). For instance, participants might be asked to decide whether they prefer choice \( a \) to choice \( b \) in a decision making experiment, to classify an ambiguous visual stimulus into category \( a \) or \( b \) in a categorization experiment, or to judge whether stimulus \( a \) is brighter than stimulus \( b \) in a psychophysical task. Let \( P_{ab} \) denote the probability of choosing stimulus \( a \) over \( b \). By definition we should have \( P_{ab} + P_{ba} = 1 \) for all \( a \) and \( b \). Suppes and Zinnes (1963) showed that the product rule

\[
P_{ab}P_{ba}P_{ca} = P_{ba}P_{cb}P_{ac}
\]

for all \( a, b, c \) uniquely derives the following model of choice probability:

\[
P_{ab} = \frac{f(a)}{f(a) + f(b)}
\]

for some real-valued function \( f(a) \). It is important to note that the product rule and the about model are logically equivalent, meaning that one implies the other and vice versa. The \( f(a) \) in the previous equation is interpreted as a measure of response strength and defines a ratio scale (F. S. Roberts, 1979, pp. 281–283). Many models in cognitive psychology are of this form or of its extension, for example, Luce’s (1959) choice model, context models of categorization (Medin & Schaffer, 1978; Nosofsky, 1986), and connectionist models of category learning (Gluck & Bower, 1988).

**Axiomatic Models**

In an axiomatic model, regularity conditions that are imposed upon observed variables are in the form of ordinal relations called axioms, rather than equality relations as in functional equation models. A set of such ordinal relations often sufficiently constrains the possible solution to be uniquely identified. The class of axiomatic models in judgment and decision making (Fishburn, 1982; Luce, 1996; Luce & Fishburn, 1991) is of this type. For example, the expected utility (EU) model of decision making under uncertainty (Von Neumann & Morgenstern, 1944) assumes that individuals select among a set of alternative choices the one that maximizes expected utility, which is defined as

\[
EU(A) = \sum_i p(A_i)u(A_i)
\]

where the sum is over probabilistic events of choice \( A_i \), \( p(A_i) \) is the probability of event \( A_i \), and \( u(A_i) \) is the utility of the event. For instance, the expected utility of a gamble in which one receives $100 with probability 0.2 or $0 with probability 0.8 is calculated as the sum \((0.2)u($100) + (0.8)u($0)\) for some
nondecreasing function \( u(x) \). A set of axioms on choice behavior that guarantees the existence of a utility function \( u(x) \) defined on choice alternatives such that the expected utility model holds includes the following:

1. **Transitivity** (if choice A is preferred to choice B, which in turn is preferred to C, then choice A is preferred to choice C)

2. **Independence** (addition or subtraction of the same event to all choices does not change preference rankings)

3. **Dominance** (if choice A is preferred to choice B, then choice A is preferred to any probabilistic combination of choices A and B)

4. **Solvability** (no choice is infinitely better than any other choices)

For further detail on these and other axioms, see Fishburn (1982).

**Model Building via Processing Assumptions**

In the preceding approach, the model is derived from a set of regularity conditions imposed at the level of behavior. The models described in this section are created by making a set of assumptions about (unobservable) internal processes, which are presumed to be involved in generating an observed response given an input stimulus in an experimental task. These assumptions eventually give rise to a model equation that specifies observed response as a function of some internal variables (i.e., parameters). They are classified into four categories: differential-process, algorithmic, connectionist, and algebraic models.

**Differential-Process Models**

For this class a model is obtained by making assumptions about internal mental processes of interest in terms of changes in behavior. From these assumptions the model is derived by integrating the constraining equations. As an example of this approach, consider again Fechner’s logarithmic law of psychophysics. Fechner derived his law by assuming that every change (i.e., increase or decrease) in the strength of a physical stimulus does not necessarily result in a constant change in perception but, instead, that the change in perception is proportional to the relative change in stimulation, which is known as Weber’s law:

\[
\Delta \psi = k \frac{\Delta x}{x} \quad \text{(Weber’s law)}
\]

where \( \psi \) is the psychological scale, \( x \) is the physical scale, and \( k \) is a positive scaling constant. Note that the internal process assumed to be responsible for this transformation is not directly observable. Rather, the observed response must be derived from such assumptions. This is carried out by integrating both sides of the above equation, resulting in the following logarithmic form:

\[
\psi(x) = k \ln(x) \quad \text{(Fechner’s Law)}
\]

Instead of Weber’s law, if it is assumed that a relative increase in sensation occurs in proportion to the relative change in stimulation, this assumption then leads to the derivation of Stevens’s power law of psychophysics:

\[
\frac{\Delta \psi}{\psi} = k \frac{\Delta x}{x} \Rightarrow \psi(x) = x^k \quad \text{(Stevens’ Law)}
\]

As another example, consider a model of forgetting that assumes that the rate of memory loss, rather than being constant across time, depends on the current memory load. Specifically, the rate of forgetting is proportional to the load: the more items stored in memory, the larger the mean number of items lost during each time interval after storage. This assumption leads to the exponential model of forgetting (Wickelgren, 1970):

\[
\frac{\Delta y}{\Delta t} = -cy \Rightarrow y(t) = y(0)e^{-ct}
\]

(Exponential Model)
where \( y \) denotes memory load at time \( t \) and \( c \) is a positive constant.

**Algorithmic Models**

For this class a model is defined in terms of a simulation procedure (Jacobs & Grainger, 1992). The procedure specifies in detail how specific internal processes interact with one another to yield output behavior as a final result. Often, the processes involved are too complicated to be expressed in closed form; consequently, to derive predictions from the model, the entire process must often be simulated on computer with help of random number generators (see the Appendix to this chapter). A sample of cognitive processes that algorithmic models have been employed to model includes discrimination (Link & Heath, 1975; Smith, 1995), memory retrieval (Ratcliff, 1978), recognition memory (Hintzman, 1988; Shiffrin & Steyvers, 1997), and decision making (Busemeyer & Townsend, 1993; Diederich, 1997). The random walk model of memory retrieval (Ratcliff, 1978) is described in detail to illustrate the idea.

The random walk model makes the assumption that memory retrieval is a search process. Specifically, given a probe item on a recognition memory test, the decision whether the probe is new or old is made by comparing it to each item in the memory search set simultaneously and in parallel. Each individual comparison proceeds by the gradual accumulation of evidence over time via a random walk process. On each trial of the random walk, the probe evokes either a sympathetic or a nonsympathetic “vibration” in a memory-set item, each determined with a fixed probability. The value of this probability is obtained as a random sample drawn from a normal distribution, the mean of which is assumed to be equal to the relatedness value between the probe and the memory-set item. If a sympathetic vibration occurs, then the walk moves upward by one unit, and downward otherwise. Over a series of trials, the random walk of each comparison process moves up and down on the scale until it eventually hits an upper boundary (match) or a lower boundary (non-match).

The decision process is made by combining outcomes of such individual comparison processes; a positive “yes” response is made if any one of the parallel comparisons terminates with a match, and a negative “no” response is made if all comparisons terminate with a nonmatch. The distribution of reaction times for each positive or negative response and also the probability of the response are obtained by simulating the random walk process on computer. By using different parameter values for different experimental conditions, the model can generate simulated data sets for the entire experiment that are compared against observed data to determine the model’s viability.

The algorithmic modeling approach provides an attractive environment in which to design models. Scientists can easily construct many variants of a model and quickly test a hypothesis to observe the model’s behavior. The model can be made as sophisticated as one likes without having to worry about ensuring that there is a closed-form solution. Accordingly, the approach allows the scientist to work with ideas that cannot yet be expressed in precise mathematical form (Estes, 1975). All these features may explain its popularity.

The approach, however, is not without disadvantages. The main disadvantage is the lack of transparency between the parts of the model and their corresponding mental process. A typical algorithmic model makes a host of assumptions about the mental processes involved, which are difficult to verify empirically because they are not directly observable. This is problematic because the adequacy of the model can be evaluated only by relying upon its predictions of output responses, even
models have no predefined meaning and can be difficult to interpret. Second, connectionist models might not be entirely falsifiable, enabling them to fit almost any pattern of data, including idiosyncratic noise. Mathematicians (Hecht-Nielsen, 1989; Hornik, Stinchcomb, & White, 1989, 1990) proved this unlimited flexibility for three-layer feedforward networks with hidden units. Specifically, their results showed that the three-layer network with back-propagation learning and a sufficient number of hidden units can approximate any continuous nonlinear input-output function to any desired degree of accuracy. If used blindly the power of such models can also be their downfall. Care must be taken to ensure that a connectionist model learns only the regularities underlying the data and not the whole data set, in which case it degenerates into a redescription of the data, which provides little useful insight into the phenomenon of interest, if any.

Algebraic Models
In an algebraic model the operation of the underlying cognitive process being modeled in the data is explicitly specified in its parameters and in the model equation. For example, the parameters may specify the relevant psychological or stimulus dimensions to which the underlying process is sensitive. The model equation may describe exactly how these parameters and the input stimulus are combined to produce an output response. This specificity creates a tight link between descriptive (verbal) theory and its computational instantiation because the relationships among the input, output, and parameters are clearly identifiable in the model equation. Accordingly, algebraic models can be easy to understand, and their assumptions can usually be well justified, often axiomatically or through functional equations. Further, quantitative as well as qualitative predictions can oftentimes be derived analytically.

Connectionist Models
Connectionist models (Grossberg, 1987; Kruschke, 1992; McClelland & Elman, 1986; Rumelhart & McClelland, 1986; Seidenberg & McClelland, 1989) are essentially of the algorithmic type, so all the advantages and disadvantages discussed in the preceding section apply also to this class of models, especially to localist connectionist models. On the other hand, connectionist models—in particular artificial neural networks—possess a few unique features that set them apart from other algorithmic models. First, these connectionist models make few explicit assumptions about the underlying processes in advance, but instead learn the regularities underlying the data through training (e.g., back-propagation rule). Because of this, the parameters (i.e., connection weights and architectural characteristics) of connectionist

where $C$ represents the constant required for certain assumptions, thus reflects sensory-motor processing that explains age-related slowing that for the experiment, reaction times is independent of age; that is, $bM_{young}(a, b) = bM_{old}(a, b)$ after simple algebraic manipulation response time for young subjects of response time as a function of the following form:

$$\text{RT}_{old} = a + bM_{young}$$

Thus the model predicts that for old adults the reaction time for young adults is less than 1 and an increase in age is positively dependent on $M_{young}$.

As discussed previously, the model generally conforms to the empirical studies in a variety of tasks. According to the model, reaction time is to a general slowing in the sensory-motor processes and the conservation of a
meaning and can only be understood, connectionally falsifiable, any pattern of noise. Matheney, 1989; Hornik, 1990) proved the three-layer feedforward units. Specifically, the three-layer learning architecture can approximate the input-output accuracy. If such models are taken to be the learning only, then not the phenomenon derived analytically. Examples of the algebraic modeling approach include the fuzzy model of perception (FLMP; Oden & Massaro, 1978), context models of categorization (Medin & Schaffer, 1978; Nosofsky, 1986), the signal detection theory of perceptual identification (Green & Swets, 1966), the processing tree models of memory (Batchelder & Riefer, 1990), and information processing models of cognitive slowing (Cerella, 1985; Meyerson et al., 1990). Cerella's (1985) linear model of cognitive slowing is described to illustrate the idea.

The model assumes that response time (RT) in a cognitive task is equal to the sum of the durations of two subprocesses:

\[ RT = C + M \]

where \( C \) represents the amount of time required for central cognitive processing and thus reflects task difficulty. \( M \) represents sensory-motor processing time. The model explains age-related slowing by assuming that for the elderly, each of these processing times is increased by a constant proportion; that is, \( C_{old} = aC_{young} \) and \( M_{old} = bM_{young} \) \((a, b > 1)\). With this assumption and after simple algebraic manipulations, one can write response time for older adults in terms of response time for younger adults in the following form:

\[ RT_{old} = aRT_{young} + (b - a)M_{young}. \] (2)

Thus the model predicts that response time for old adults is a linear function of response time for young adults with a slope greater than 1 and an intercept that may be negative or positive depending upon the sign of \((b - a)\). As discussed earlier, these predictions were generally confirmed across many individual studies in a variety of experimental settings. According to the model, this outcome is due to a general slowing of cognitive and sensory-motor processing. In particular, the typical observation of a negative intercept in the linear functional relationship in response time between older and younger adults may suggest that the slowing of central cognitive processes is more severe than that of sensory-motor processes \((i.e., a > b)\).

To summarize, there are a number of approaches to modeling behavior. Each has pros and cons, and the modeler must be aware of these when choosing and using the approach. For example, if one is interested in developing the most accurate redescription of the data without reference to the underlying process \((i.e., mapping the input-output relationship as in psychophysical modeling), then the functional equation approach might be most suitable. On the other hand, if the goal is to model the processes underlying this input-output relationship, then one of the other approaches should be used.

MODEL CONSTRUCTION AND TESTING

The goal of modeling in psychology is to infer the regularity present in given data while at the same time assessing the veridicality of the hypothesized model. From a statistical standpoint, the data \( y = (y_1, \ldots, y_m) \) with \( m \) observations is a random sample generated from a true but unknown probability distribution, which represents the regularity underlying the data. Formally, a model is defined as a family of probability distributions, \( \{f(y | \theta), \theta \in \Gamma\} \) where \( \Gamma \) is the parameter space. The probability distribution function \( f(y | \theta) \) specifies the probability of observing data \( y \) given the parameter \( \theta \) of the model. The same probability curve \( f(y | \theta) \) when expressed as a function of the parameter \( \theta \) given a particular value of data \( y \), is called the likelihood function. The parameter \( \theta = (\theta_1, \ldots, \theta_k) \) may be a vector as an element of a multidimensional parameter space. By varying values of the parameter, different shapes of probability
distribution are generated. In formal terms, a model consists of the collection of all such probability distributions indexed by its parameter vector.

In some circumstances, the data \( y \) can be written as a sum of a deterministic component plus random error:

\[
y_i = g(\theta, x_i) + e_i \quad (i = 1, \ldots, m) \tag{3}
\]

where \( x_i \) is the value of an independent variable, \( g(\theta, x_i) \) is the mean of \( y_i \), and \( e_i \) is a random noise with zero mean.

As an illustrated example, consider again Cerella’s (1985) linear model of cognitive slowing in Equation (2). The model relates response times for younger adults \((x_i)\) to response times for older adult \((y_i)\) across tasks \((i)\) by a linear function,

\[
y_i = \theta_1 x_i + \theta_2 + e_i
\]

(Cerella’s linear model) \tag{4}

where \( i = 1, \ldots, m \). According to this equation, the model assumes two parameters, \( \theta = (\theta_1, \theta_2) \), and the mean function in the form of \( g(\theta, x_i) = \theta_1 x_i + \theta_2 \). Note that the two parameters \( a \) and \( b - a \) of \( M_{\text{young}} \) in Equation (2) now correspond to \( \theta_1 \) and \( \theta_2 \) in the Equation (4), respectively. As such, the slope parameter \( \theta_1 \) is interpreted as a ratio of the cognitive processing time for older adults over younger adults, whereas the intercept parameter \( \theta_2 \) is proportional to the difference in the age-related slowing ratio between sensory-motor and cognitive processing. The model may further assume that the error \( e_i \) is normally distributed with variance \( \sigma^2 \). This implies that for given values of \( \theta_1 \), \( \theta_2 \), and \( x_i \), the data value \( y_i \) is normally distributed with mean \( \theta_1 x_i + \theta_2 \) and variance \( \sigma^2 \). The linear model therefore defines a family of normal probability distributions, \( N(\theta_1 x_i + \theta_2, \sigma^2) \), created by varying the values of the two parameters \( \theta_1 \) and \( \theta_2 \) for a fixed value of \( x_i \). The probability density function of \( y_i \) given all parameter values then takes the following form:

\[
f(y_i | \theta) = \frac{1}{\sqrt{2\pi \sigma}} \exp\left(-\frac{(y_i - \theta_1 x_i - \theta_2)^2}{2\sigma^2}\right) \quad (i = 1, \ldots, m) \tag{5}
\]

where \( \exp(x) \) stands for the exponential function, that is, \( \exp(x) = e^x \). Quite often, the equation \( g(\theta, x_i) \) itself is taken to define a mathematical model and the underlying probability distribution is kept implicit. Assuming independent observations, the probability distribution function of the entire data set, \( y = (y_1, \ldots, y_m) \), can be written as the multiplication of \( m \) individual density functions as follows:

\[
f(y | \theta) = \prod_{i=1}^{m} f(y_i | \theta). \tag{6}
\]

The model-testing approach presented later requires that a model be specified by its parametric family of probability distributions, \( \{f(y | \theta), \theta \in \Gamma\} \), or at least by its mean function \( g(\theta, x_i) \). Most of the models discussed in the preceding section, including algebraic models and connectionist models, satisfy this requirement. Exceptions are some axiomatic models, in which the specification of the probability distributions and the mean function is not possible. In this case, testing of such “non-statistical” models requires an alternative approach, which is described briefly at the end of this section.

Before discussing model testing, a word about what one can reasonably hope to achieve in modeling behavior is in order. Psychological phenomena have the potential to be very complex and may involve many subsystems interacting with one another in a highly nonlinear fashion. It could easily be the case that a mathematical model with at least a dozen parameters is necessary to capture the phenomenon accurately. In the early stages of modeling, all models will most certainly be wrong in many details. Because of this, they are also known as “true” models or “true” models. It may also be worth noting that capturing the correct rate and importance of models and their parameters to the true model is not possible. Thus, it is important to keep this in mind when modeling data in these models.

Model Falsifiability

Once a model has been defined, the question of whether the model is falsifiable is next. A model is falsifiable if it must be true and false at any time. In other words, if a model is falsifiable, it is falsifiable in every possible way. This is a fundamental requirement of any model; otherwise, no point is made in the model. If the model is falsifiable, it is possible to test the model against data. Assume the data set \( \{x_i, y_i\} \). The model is then falsifiable if there is a perfect match, \( \langle i = 1, \ldots, m \rangle \), between the model and the data. In other words, the model is falsifiable if it is correct.
\[ f_i = \frac{\sum_{j=1}^{m} \left( x_i - \theta_j \right)^2}{2\sigma^2} \]

(5)

mental function, the 
to define a 
ly underlying probabilistic. Assuming probabilistic data set, as the multiply functions are also likely to be misspecified, meaning that they do not include as a special case the “true” model that generated the observed data. It may also be impossible to build a model that captures the underlying mental process in every detail. Thus, it is probably most accurate and instructive to think of mathematical models as best approximations to the truth. Practically speaking, the goal of mathematical modeling is to find the best approximation to the truth, with the hope that the winning model is correctly specified (i.e., includes the true model as a special case). It is useful to keep this in mind to avoid placing undue faith in these models.

Model Falsifiability and Identifiability

Once a model is defined along with the family of probability distributions indexed by the model’s parameters, two important issues must be checked before the model is fitted to observed data to assess its validity. They are falsifiability and identifiability.

Falsifiability (Popper, 1959), also called testability, refers to whether there exist potential observations that are inconsistent with the model (i.e., data that it does not predict). This is a necessary precondition for testing a model; unless a model is falsifiable, there is no point in testing the model. An unfalsifiable model is one that can describe all possible patterns of data that can arise in a given experiment. For example, one may wish to test a two-parameter linear model, \( y = \theta_1 + \theta_2 x \), against data that consist of just two observations, \((x_1, y_1)\) and \((x_2, y_2)\). In this case the model is unfalsifiable because it can provide a perfect fit to the data for any \((x_i, y_i)\) pairs \((i = 1, 2)\). On the other hand, if one additional point \((x_3, y_3)\) is added to the data, then the two-parameter model becomes falsifiable. In other words, whether a model is falsifiable or not depends upon the size of the data set.

A rule of thumb for assessing falsifiability is that a model is falsifiable if the number of free parameters is less than the number of observations in the data. Bamber and van Santen (1985) showed, however, that this “counting rule” can be misleading, especially for nonlinear models, and provided counterexamples as evidence. Consider Luce’s (1959) choice model. The model assumes that the probability of choosing choice alternative \(i\) over alternative \(j\) is determined by their respective utility values in the following form:

\[
P_{i \succ j} = \frac{u_i}{u_i + u_j} \quad (u_i > 0; i, j = 1, \ldots, s)
\]

(6)

where \(u_i\) is the utility parameter to be estimated from the data. Note that the number of parameters in the model is equal to the number of choice alternatives \(s\), whereas the number of independent observations is equal to \(s(s - 1)/2\). Hence, for \(s = 3\), both the number of parameters and the number of observations are equal. However, it is easy to show that the model is falsifiable in this case. In another, more dramatic example, Bamber and van Santen (1985, p. 453) showed that the number of parameters \(7\) in a model exceeded the number of data observations \(6\), yet the model was still falsifiable!

Rectifying the apparent limitations of the counting rule of falsifiability, Bamber and van Santen (1985) provided a formal definition of falsifiability and also a criterion for assessing falsifiability, which includes the counting rule as a special case. Specifically, the criterion states that a model is falsifiable if the rank of what is called the Jacobian matrix, defined as

\[
J_{ij}(\theta) = \left[ \frac{\partial E(y_j)}{\partial \theta_i} \right],
\]

\((i = 1, \ldots, k; j = 1, \ldots, m)\),

is less than the number of independent observations \(m\) in the data for all \(\theta\) values.

Model identifiability refers to whether the parameters of a model are unique given observed data. As with falsifiability, a model is
identifiable or unidentifiable with respect to data size. When a model is unidentifiable, the problem of equivalent models arises in which there exist multiple sets of parameter values that provide exactly the same fit to the data. When this happens, one cannot meaningfully interpret the parameter values of the model. To illustrate, consider a three-parameter model of $y = \theta_1 + \theta_2 x + \theta_3 x^2$ and suppose that two data points are obtained, say $(x_1, y_1) = (1, 1)$ and $(x_2, y_2) = (2, 5)$. The model is unidentifiable given these data because there exist multiple sets of the model's parameter values $(\theta_1, \theta_2, \theta_3)$ that fit the data perfectly, for example, $(\theta_1, \theta_2, \theta_3) = (-1, 1, 1)$ and $(\theta_1, \theta_2, \theta_3) = (-5, 7, -1)$. There are, in fact, an infinite number of such parameter values of the model that can provide a perfect description of the data. In order for this model to be identifiable, three or more data points are needed.

A rule of thumb for assessing identifiability is that a model is identifiable if the number of free parameters is less than or equal to the number of independent observations. Again, Bamber and van Santen (1985) provide a formal definition of identifiability and show that this rule is also imperfect.

Although it might appear that falsifiability and identifiability should be related to each other, there exists only one consistent relationship between the two: The counting rule of falsifiability is valid if the model is identifiable. If a model is not identifiable, the counting rule for assessing falsifiability may be inapplicable (Bamber & van Santen, 1985).

The lack of a tight relationship between these properties of a model means that a model can be falsifiable but not identifiable. A case in point is FLMP (Oden & Massaro, 1978). To demonstrate this situation, consider a letter recognition experiment in which participants have to classify the stimulus as belonging to one of two categories, A and B. Assume that the probability of classifying a stimulus as a member of category A is a function of the extent to which the two feature dimensions of the stimulus (i and j) support the category response (Massaro & Friedman, 1990). Specifically, FLMP assumes that the response probability $P_{ij}$ is a function of two parameters, $\theta_i$ and $\lambda_j$, each of which represents the degree of support for a category A response given the specific $i$ and $j$ feature dimensions of an input stimulus:

$$P_{ij} = g(\theta_i, \lambda_j) = \frac{\theta_i \lambda_j}{\theta_i \lambda_j + (1 - \theta_i)(1 - \lambda_j)}$$

where $0 < \theta_i, \lambda_j < 1, 1 \leq i \leq s, 1 \leq j \leq v$. $s$ and $v$ represent the number of stimulus levels on the two feature dimensions, $i$ and $j$, respectively, and together constitute the design of the experiment.

FLMP is falsifiable, which can be shown using the falsifiability rule mentioned earlier (Bamber and van Santen, 1985; see also Batchelder, 1997). For example, one can easily come up with a set of $P_{ij}s$ that do not fit into the model equation, such as $P_{ij} = (a_i + b_j)/2$ for $0 < a_i, b_j < 1$ (N. H. Anderson, 1981).

Regarding the identifiability of FLMP, for the $s \times v$ experimental design, the number of independent observations is $sv$, and the number of parameters of FLMP is $(s + v)$. For example, for $s = v = 8$, the number of observations is 64, which far exceeds the number of parameters in the model (16). Surprisingly, however, Crowther, Batchelder, and Hu (1995) have shown that FLMP is not identifiable for all values of $s$ and $v$. According to their analysis, for any given set of parameter values $(\theta_i, \lambda_j)$ that satisfy the above model equation, another set of parameter values $(\theta_i^*, \lambda_j^*)$ that also satisfy the same equation can always be obtained:

$$\theta_i^* = \frac{\theta_i}{1 + c(1 - \theta_i)}; \quad \lambda_j^* = \frac{\lambda_j}{1 + c(1 - \lambda_j)}$$

for any constant $c$. Crowther et al. (1995) identify the equivalent parameter sets $\{\theta_i, \lambda_j\}$ for any $\{\theta_i, \lambda_j\}$ and show that there are an infinite number of such pairs. Specifically, for any $\{\theta_i, \lambda_j\}$, there are an infinite number of equivalent parameter sets, each of which provides an equally good fit to the data. FLMP is not falsifiable.

Can FLMP be falsifiable? Consider the case of one of its parameter sets $\{\theta_i, \lambda_j\}$. For this set, the goodness-of-fit constant $c$ is the only parameter that accommodates the fit to the data. In contrast, for any other parameter set containing the same $\{\theta_i, \lambda_j\}$, the goodness-of-fit constant is not to be accommodated. Instead of this parameter set, the one shown by Crowther et al. (1995) can be used.

**Parameter Estimation**

Once data have been assumed to be generated by a model and are shown to be identifiable, one can then ask how well the model's goodness-of-fit parameter set constant describes the data. Goodness of fit is the extent to which the model contains a parameter set constant that best describes the probability distribution of the data. A distinct set of goodness-of-fit parameter sets is to find a set of parameter constants that best fits the observations (Anderson, 1981). This process is called parameter estimation.

Two general parameter estimation methods are available: (1) ordinary least-squares (OLS), and (2) maximum likelihood estimation (MLE). In OLS, one maximizes the mean squared error (MSE) of the predictions $\hat{y}$ to the observed data:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where $\Sigma(y_i - \hat{y}_i)^2$ represents the sum of squared errors (SSE).

On the other hand, in MLE, one maximizes the log-likelihood function $L$ of the data, $f(y | \theta)$.
for any constant \( c > -1 \). One can easily verify the equivalence by plugging the parameters \((\theta_i, \lambda_i)\) into the model equation. Because there are an infinite number of possible \( c \) values, there will be an equal number of parameter sets, each of which provides exactly the same fit to the observed data, meaning that FLMP is not identifiable.

Can FLMP be made identifiable? The answer to this question is yes. For example, one of its parameters can be fixed to a preset constant (e.g., \( \theta_i = 0.15 \), for some \( i \)). Alternatively, the model equation can be modified to accommodate four response categories instead of two. For further details, consult Crowther et al. (1995).

**Parameter Estimation**

Once data have been collected and the model is shown to be falsifiable as well as identifiable, one is in a position to assess the model's goodness of fit to the experimental data. Goodness of fit refers to how well the model fits the observed data. Given that the model contains many (theoretically infinite) probability distributions, each associated with a distinct set of parameter values, the objective is to find a set of parameter values that best fits the observed data in some defined sense. This process is called parameter estimation.

Two generally accepted methods of parameter estimation are least square estimation (LSE) and maximum likelihood estimation (MLE). In LSE the parameter values that minimize the mean squared error (MSE) between predictions and observations are sought:

\[
\text{MSE} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (y_i - g(\theta, x_i))^2}
\]  

where \( \Sigma (y_i - g(\theta, x_i))^2 \) is the sum of squares error (SSE).

On the other hand, in MLE the likelihood of the data, \( f(y | \theta) \), is maximized with respect to the model's parameter values, as illustrated schematically in Figure 11.1. Assuming independent observations and normal error with a constant variance \( \sigma^2 \), we can write the logarithm of the likelihood, called the log likelihood, as

\[
\ln f(y | \theta) = -\frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - g(\theta, x_i))^2 - m \ln\left(\frac{1}{\sqrt{2\pi\sigma}}\right) = \alpha \text{MSE}^2 + \beta
\]

where \( \alpha = -m/(2\sigma^2) \), \( \beta = -m \ln(\sqrt{2\pi\sigma}) \). Note that \( \alpha \) and \( \beta \) do not depend upon the parameter \( \theta \). Therefore, if \( y, s \) are normally distributed with equal variances, maximizing the likelihood is equivalent to minimizing MSE, and therefore the same parameter values are obtained under either method. Otherwise, the two solutions tend to differ. In general, MLE is a preferred method of estimation, especially when the equal variance assumption is violated (e.g., binomial probability distributions). Throughout this chapter the best-fitting
parameter vector obtained in MLE or LSE is denoted by $\hat{\theta}$.

Finding the parameter values that maximize MLE or minimize MSE usually requires use of nonlinear estimation techniques unless the solution can be found in analytic form. For example, suppose that the model is a linear regression model, $Y = X\theta + E$, where $Y$ and $E$ are $m \times 1$ dependent and error vectors respectively, $X$ is a $m \times k$ design matrix and $\theta$ is a $k \times 1$ parameter vector. In this case, the solution that minimizes MSE can be obtained in analytic form. For nonlinear models, however, it is generally impossible to obtain such analytic form solutions; consequently, a solution must be sought numerically using optimization algorithms (Thisted, 1988).

To illustrate parameter estimation, consider again Cerella's (1985) linear model of cognitive slowing, which has two free parameters, $\theta_1$ (slope) and $\theta_2$ (intercept). We fitted the model to a data set reported in Cerella, Poon, and Fozard (1981). The observed data consist of ten pairs of response times for old and young adults collected under the normal-pairity condition in a letter rotation task, and are shown as open circles in Figure 11.2. The values of the parameters that minimized MSE were $\theta_1 = 2.488$ and $\theta_2 = -0.958$, with MSE minimization equal to 0.109. The model accounted for 95.7% of the total variance in the data. The thicker solid line in Figure 11.2 depicts this best-fitting linear model. What do the parameter values mean? According to the model, the slope parameter ($\theta_1$) is interpreted as the proportion of slowing in central cognitive processing for older adults relative to younger adults, whereas the sign of the intercept parameter ($\theta_2$) indicates whether slowing in sensory-motor processing is more (+) or less (−) severe than slowing in central cognitive processing. The results indicate that in the letter rotation task, there is a significant cognitive processing slowing of sensory-motor type.

**Goodness-of-fit tests**

Although parameter estimation provides a sense of how well the model fits the data, a goodness-of-fit test is needed if one is not particularly satisfied with the model or if an MSE of 0.109 is not considered good enough. Thus, one must consider a test of model fit method of doing so. Generally, there are two tests that can be applied, the likelihood ratio test and the log-likelihood ratio test. The likelihood ratio test is:

$$\chi^2 = G^2$$

In the equation above, $\chi^2$ is the likelihood ratio test statistic or any other goodness-of-fit test statistic such as LSE, and $G^2$ is a general-purpose goodness-of-fit statistic (some use for observations) conclude that when the $\chi^2$ test statistic of a model is (that of the null hypothesis), the goodness-of-fit statistic, $y_i = n f(y_i)$, where $f(y_i)$ is the expected frequencies. When the $\chi^2$ test statistic takes on a significant discrepancy, the hypothesis of log-likelihood becomes non-significant. Under the null hypothesis, the $\chi^2$ test statistic is distributed as a chi-square distribution with degrees of freedom equal to the number of the observed frequencies.
the letter rotation task, the slowing of central cognitive processing is more severe than is slowing of sensory-motor processing.

Goodness-of-Fit Testing

Although parameter estimation yields a measure of how well the model fits the observed data, a goodness-of-fit measure, by itself, is not particularly meaningful. How good is an MSE of 0.109? To answer this question, one must conduct a goodness-of-fit test. One method of doing so is via hypothesis testing, in which the null hypothesis that the model is correct (i.e., viable) is tested against the alternative hypothesis that it is not correct.

Null Hypothesis Testing for Discrete Random Variables

In this test the model’s goodness of fit is assessed by comparing the expected counts under the null hypothesis against the actually observed counts, \(y_i\). The two most popular, general-purpose methods used to test the null hypothesis are the Pearson chi-square \((\chi^2)\) test and the log likelihood ratio \((G^2)\) test:

\[
\chi^2 = \sum_{i=1}^{m} \frac{(y_i - n f(y_i | \hat{\theta}))^2}{n f(y_i | \hat{\theta})};
\]

\[
G^2 = -2 \sum_{i=1}^{m} y_i \ln \frac{n f(y_i | \hat{\theta})}{y_i}.
\]

In the equation, \(\hat{\theta}\) is the MLE parameter estimate or any other reasonable estimate such as LSE, and \(n\) is the sample size, which is the number of repeated random samples (i.e., observations) collected to obtain count \(y_i\). Note that when there is perfect agreement between the null hypothesis and the observed data, that is, \(y_i = n f(y_i | \hat{\theta})\), both statistics become zero, \(\chi^2 = G^2 = 0\), and otherwise they take on positive values. The greater the discrepancies, the larger the value of the statistic. Under the null hypothesis, both statistics are distributed as a chi-square distribution with \((m-k-1)\) degrees of freedom where \(k\) is the number of free parameters. Therefore, the null hypothesis may be retained or rejected by comparing the observed value of \(\chi^2\) or \(G^2\) statistic to the critical \(\chi^2\) value obtained for an appropriate alpha level. If the model is rejected, one may look for some alternative model that is more consistent with the data. Although both statistics, \(\chi^2\) or \(G^2\), are equivalent for large \(n\) under the assumption that the model is correct, we recommend the latter because it is more robust and requires fewer assumptions. Nevertheless, in practice, the difference between the two statistics is rarely large enough to lead to differing conclusions. When the model is not correct, however, the two statistics can yield very different results even with large sample sizes. A more comprehensive treatment of goodness-of-fit tests for discrete random variables, including the \(\chi^2\) and \(G^2\) tests, can be found in Read and Cressie (1988).

Null Hypothesis Testing for Continuous Random Variables

Testing the goodness of fit of a model with the dependent variable \(y\) measured on a continuous scale is a bit complicated. No general-purpose method of testing the validity of a single model exists unless probability distributions are restricted to a few known families, such as exponential and normal families (D’Agostino & Stephens, 1986). On the other hand, the relative ability of two nested models to account for observed data can be tested via the generalized likelihood ratio test (Wilks, 1938). Two models are nested if one model can be reduced to a special case of the other. For example, a two-parameter linear model of the form \(y = \theta_1 + \theta_2 x + \text{error}\) and a three-parameter quadratic model of the form \(y = \theta_1 + \theta_2 x + \theta_3 x^2 + \text{error}\) with the same probability distribution of the error are nested.
This is because the former is obtained from the latter as its special case by setting \( \theta_3 = 0 \). On the other hand, an exponential model of the form \( y = \theta_1 + \theta_2 \exp(-\theta_3 x) + \epsilon \) and \( y = \theta_1 + \theta_2 x + \theta_3 x^2 + \epsilon \) are nonnested.

The basic idea of the generalized likelihood ratio test is to create two models, \( M_1 \) (restricted model) and \( M_2 \) (full model), in such a way that \( M_1 \) is nested within \( M_2 \). For example, \( M_1 \) might be created by holding constant values of one or several parameters of \( M_2 \), rather than treating them as free parameters. Each of the two models is then fitted to the data, and its best-fitting parameter values are obtained via MLE. Let \( L_1 \) and \( L_2 \) denote the maximized likelihood value of \( M_1 \) and \( M_2 \), respectively. The generalized likelihood ratio test is based on the \( G^2 \) statistic defined as

\[
G^2 = 2(\ln L_2 - \ln L_1) \tag{10}
\]

which takes on nonnegative values. Under the null hypothesis that \( M_1 \) is correct, this statistic is distributed as a chi-square distribution with degrees of freedom equal to the difference in the number of free parameters between the two models. The standard hypothesis testing procedure is then applied to decide whether to retain or reject the null hypothesis. If the null hypothesis is retained, one concludes that the reduced model \( M_1 \) offers a sufficiently good description of the data, and thus the extra free parameters of the full model \( M_2 \) appear to provide no real improvement in fit and therefore may be unjustified. On the other hand, if the null hypothesis is rejected, one concludes that the extra parameters may be necessary to account for the observed pattern of data. This generalized likelihood test based on \( G^2 \) can also be applied to discrete random variables. It is worth noting that no comparable test based on \( \chi^2 \) for the same purpose exists.

As an illustrated example of the above generalized likelihood test, suppose that one is interested in testing the adequacy of Cerella’s (1985) linear model of age-related deficits. In order to apply the method, we created the following nonlinear model that yields the linear model as a special case (i.e., \( \theta_2 = \theta_4 = 1 \) and \( \theta_3 = 0 \)):

\[
y_i = \theta_1 (\theta_2 x_i + \theta_3)^{\theta_4} + \theta_5 + \epsilon_i \quad (i = 1, \ldots, m).
\]

This model is motivated from and is a generalized version of Meyerson et al.’s (1990) information loss model of cognitive aging. The latter model assumes that a constant proportion of information is lost in each successive step of cognitive processing and, further, that the proportion that is greater for older adults than for younger adults. Specifically, the parameter \( \theta_4 \) represents the ratio of information loss between older and younger adults; similarly, other parameters of the model can be related to information loss (see Meyerson et al., 1990, for details). Best-fit parameter values of the model for the same data set from Cerella et al. (1981) were obtained as \( \theta_1 = 0.467, \theta_2 = 1.512, \theta_3 = -0.162, \theta_4 = 2.143, \theta_5 = 0.588 \), with the minimized MSE equal to 0.089. The thinner solid line in Figure 11.2 represents the best-fitting nonlinear model. This five-parameter model fit the data almost perfectly, according for 97.1\% of the variance, which is an increase of 1.4\% over Cerella et al.’s (1981) two-parameter linear model. The generalized likelihood test can then be performed to determine whether the mere 1.4\% increase represents meaningful improvement in fit. To obtain the required \( G^2 \) statistic, the MSE values must be converted to the corresponding maximized log likelihood values. This was done using Equation (10), and the resulting log likelihood values were equal to 7.99 and 10.0 for the restricted and full models, respectively. \( G^2 \) was then calculated and found to be equal to 4.02, which is smaller than 7.81, the critical value of \( \chi^2(df = 3, \alpha = 0.05) \). Therefore, the null hypothesis

Although this method is useful and provides a form of hypothesis testing from its own standpoint, it is somewhat unsatisfactory for applied research. In Chap. 10, we discussed the limitations of the likelihood ratio test. In this chapter, we will focus on hypothesis testing of linear models. In particular, we will examine the null hypothesis that the effects of the independent variables are not significantly different from zero. Jumping to the results without investigating the data may lead to incorrect conclusions. Therefore, it is important to check the assumptions before applying such methods.

As mentioned earlier, the likelihood ratio test can be nested in the same way as the traditional chi-square test, but it does not have the same number of degrees of freedom. Therefore, the model equation predicted in the previous section is not comparable to \( \chi^2 \) if more than one variable is added to the model. In practice, the generalized likelihood ratio test is used to compare nested models. In such cases, the likelihood ratio test is usually more powerful than the chi-square test. In summary, we will discuss the principles of hypothesis testing and model selection.

### Testing Null Hypothesis

So far we have only discussed the likelihood ratio test. A similar test can be conducted to determine if the null hypothesis is true. In this section, we will discuss how to conduct such a test and what assumptions are necessary to make the test valid.

1. **Assumptions of the Test**: For the test to be valid, certain assumptions about the data must be satisfied. These assumptions include:
   - The data are normally distributed.
   - The residuals are independent and identically distributed.
   - There is no multicollinearity among the independent variables.

2. **Calculating the Test Statistic**: The test statistic is calculated using the following formula:

\[
G^2 = 2(\ln L_2 - \ln L_1)
\]

where \( L_1 \) and \( L_2 \) are the likelihood values for the null and alternative hypotheses, respectively.

3. **Determining the Significance Level**: The significance level is determined by comparing the calculated test statistic to the critical value from the chi-square distribution. The critical value is determined using the significance level and the degrees of freedom.

4. **Interpreting the Results**: If the calculated test statistic is greater than the critical value, the null hypothesis is rejected. If the calculated test statistic is less than the critical value, the null hypothesis is not rejected.

In summary, the likelihood ratio test is a powerful tool for testing hypotheses about the parameters of a model. However, it is important to ensure that the assumptions of the test are met before applying it. Additionally, the test statistic is calculated using the likelihood values for the null and alternative hypotheses. The significance level is determined by comparing the calculated test statistic to the critical value from the chi-square distribution. Finally, the results are interpreted by comparing the calculated test statistic to the critical value.
hypothesis that the linear model is a correct description of the data should be retained. The three extra parameters \( \theta_2, \theta_3, \theta_4 \) do not seem necessary to account for the observed data.

Although null hypothesis testing is easy to use and provides a reasonable and informative assessment of the validity of a model in its own way, the usefulness of the method is somewhat over-sold, especially in the behavioral sciences. The reader should be aware of the limitations and criticisms of the method (Chap. 10; Berger & Berry, 1988; Cohen, 1994; but see Hagan, 1997) and is cautioned against possible misinterpretations of hypothesis testing results. For example, retaining the null hypothesis (i.e., failing to reject the null hypothesis) does not necessarily imply that the hypothesis is more likely to hold than is the alternative hypothesis, let alone that it is confirmed to be the correct (i.e., true) model. Jumping to such conclusions is still commonplace in the psychological literature, so it is particularly important to guard against making such errors in reasoning.

As mentioned earlier, the generalized likelihood ratio test requires that the two models be nested. If they are not, if both have the same number of parameters but differ in their model equation (e.g., \( y = \theta x \) vs. \( y = x^0 \)), or if more than two models are being compared, the generalized likelihood test is not appropriate. In such cases, another method of statistical inference must be used. This and related issues are discussed in the section titled "Model Selection."

### Testing Nonstatistical Models

So far we have dealt with statistical models only. A statistical model specifies the probability of observing data—that is, \( f(y \mid \theta) \)—given the model's parameter values. This allows a probabilistic formulation for testing the validity of the model using null hypothesis testing. On the other hand, there are classes of nonstatistical models for which the probability distribution is not specified. Some axiomatic models that are formulated in the form of ordinal predictions fall into this category. For such qualitative models, it is not entirely clear how to construct a probabilistic formulation for model testing. To illustrate, suppose that an axiomatic model assumes transitivity and that when the axiom was tested against observed data sets, it was found that it held up pretty well, with violations observed in only 3% of the data. Should this be considered sufficient evidence for retaining the model? If this question were answered by performing a statistical test, one would need to calculate the probability of observing violations of the axiom in 3% or more of the data under the assumption that the model, or an appropriately chosen null model, holds. Because the model specifies no error theory for its axioms, it is not possible to calculate this probability.

Monte Carlo methodology may present a possible remedy for testing nonstatistical models. For example, Nygren (1983) proposed a Monte Carlo approach by which the likelihood of violating an axiom by chance is estimated through numerical simulations. Specifically, to assess the fit of an axiomatic model to observed data, we first generate artificial data sets under a random response model with no particularly meaningful structure. An axiom of interest is then tested individually in these random data sets, and the proportion of violations that would be expected under the random model is obtained. This baseline violation rate is then used as a benchmark against the empirically observed proportion. An appropriate statistical test, such as the \( r \) test using the binomial distribution, may be performed to examine whether the observed data represent a significantly better fit than would be expected under the null hypothesis of the random response model. The procedure may be repeated for each of the
Interpreting Good Fits

What does it mean when a model fits the data well? It is important not to jump to the conclusion that one has identified the form of the underlying process and therefore discovered the true model. As S. Roberts and Pashler (2000) noted, a good fit is only one of many conditions that must be satisfied before such a conclusion should even be contemplated. At first blush, a good fit would seem to be sufficient to claim that the model is accurate. After all, by fitting the data well, the model has demonstrated its ability to mimic the phenomenon of interest. However, a good fit merely qualifies the model as a true model, placing it in the set of models that could be true. As the example on interpreting the linearity of Brinley plots shows, this information is not terribly informative by itself, for there are without a doubt many models in this set.

Simply put, verification of a model's predictions (i.e., data fitting) can never amount to a sufficiency test of the model. What constitutes a sufficiency test? Given that a model is defined as a set of assumptions about the underlying mental processes of interest, the model is sufficient only when all of its assumptions are tested and validated independently, which could be a challenging task. The model itself must be well understood before such a claim can be made. For starters, we need to be able to answer the question, "Why does the model fit the data well?" The answer should be "because it is a good approximation of the mental process." As discussed in the next section, good fits can be achieved for reasons that have nothing to do with the model's exactness. For this reason we recommend that the more appropriate measure of a model's adequacy is a test of its generalizability, not its goodness of fit.

MODEL SELECTION

The preceding discussion should make it clear that the objective in model testing is to test the viability of a model, not to take the additional step and conclude that it is the correct model. To do so is unwarranted because one's model has not been demonstrated to be superior to others. Model selection, on the other hand, involves a set of competing models, all of which have passed goodness-of-fit tests and have been found to provide a "good" description of the data. The objective in model selection is to decide which one is the best model in the sense that it most closely approximates the underlying mental process. A more detailed and technically rigorous discussion of some of the issues presented here can be found in the book *Model Selection* (Linhart & Zucchini, 1986) and in a special issue on model selection of the *Journal of Mathematical Psychology* (Myung, Forster, & Browne, 2000).

There are a number of criteria for choosing among mathematical models (Jacobs & Grainger, 1994): (a) explanatory adequacy (is the theoretical explanation of the model reasonable and consistent with established findings?); (b) plausibility (are the assumptions of the model biologically and psychologically plausible?); (c) interpretability (do the parameters of the model make sense and have meaningful interpretations?); (d) goodness of fit or descriptive adequacy (does the model provide a good description of the observed data?); (e) generalizability (does the model predict well the statistics of new, as yet unseen, data?); and (f) complexity or simplicity (does the model capture the phenomenon in the simplest possible manner?). Although each of these criteria is important to consider in model selection, the last three (goodness of fit, generalizability, and complexity) are particularly pertinent to choosing among mathematical models, and quantitative methods have been developed with this in mind, beginning with the criteria of fit and then developing relationships in an illustrative manner.

Goodness of fit

Goodness of fit (GOF) is often based on the likelihood (ML) approach to model selection. The likelihood method links the concept of estimation to the concept of goodness of fit. A model whose output fits the data well can be considered "good." Of course, validating the goodness of fit is not enough for a model selection criterion; the selection based solely on goodness of fit may lead to the wrong choice of model. Why? Because the goodness of fit is a property of the sample itself and comes from the process of inference.

Statistical models are a sample, and therefore with finite variation: variation due to sampling variation, due to the variability of the population, variation due to measurement error, and variation due to lack of independence of the estimates. An important consequence of this is that the model should not be blind to this variation, yet goodness of fit criteria reflect a distinction between models that are blind to variation as MSE tied goodness of fit criteria. What we need, then, is a way to use these uncertainties in the model selection process to assess the generalizability aspect of the model.

Generalizability

Generalizability is a property of a model that is not captured in its fit not only to the
developed with this purpose in mind. We begin by defining these criteria in more detail and then demonstrating their interrelationships in an illustrated example.

**Model Selection Criteria**

**Goodness of Fit**

Goodness of fit, such as MSE and maximized likelihood (ML), is a necessary component of model selection. Because data are our only link to the cognitive process under investigation, a model must be able to describe well the output from this process if a model is to be considered seriously. Failure to do so invalidates the model. As stated earlier, goodness of fit is not a sufficient condition for model selection. This is because model selection based solely on goodness of fit will result in the choice of a model that overfits the data. Why? Because the model will capture variability present in the particular data set that comes from sources other than the underlying process of interest.

Statistically speaking, the observed data are a sample generated from a population and therefore will contain at least three types of variation: variation due to sampling error because the sample is only an estimate of the population, variation due to individual differences, and variation due to the cognitive process of interest. Most of the time it is only the third of these that one is interested in modeling, yet goodness-of-fit measures do not distinguish between any of them. Measures such as MSE treat all variation identically. They are blind to its source and try to absorb as much of it as possible (this is demonstrated later). What is needed is a means of filtering out these unwanted sources of noise. Generalizability achieves this.

**Generalizability**

Generalizability refers to a model’s ability to fit not only the observed data in hand but also future, unseen data sets generated from the same underlying process. To illustrate, suppose that the model is fitted to the initial set of data and that its best-fitting parameter values are obtained. With these parameter values held constant, if the model also provides a good fit to additional data samples collected from replications of that same experiment (i.e., the same underlying probability distribution or regularity), then the model generalizes well. Only under such circumstances can one be sure that a model is accurately capturing the underlying process, and not the idiosyncrasies of a particular sample.

The superiority of this criterion becomes readily apparent in the following illustration. In Figure 11.3 the solid circles represent observed data points, and the curves represent

![Model A](image1)

![Model B](image2)

**Figure 11.3** Illustration of the trade-off between goodness of fit and generalizability.

**NOTE:** Two models (curves) are fitted to the same data set (solid circles). New observations are shown by the open circle.
best-fits by two hypothetical models. Model A not only captures the general trend in the current data but also does a good job of capturing new observations (open circles). On the other hand, Model B provides a much better fit than model A, but it does so by fitting the random fluctuations of each data point as well as the general trend, and consequently suffers in fit when new observations are introduced into the sample. As the example shows, generalizability is a very reliable way to overcome the problem of noise and extract the regularity present in the data. Further examples later will demonstrate why generalizability should be adopted as the primary quantitative criterion on which the adequacy of a model is evaluated.

**Complexity**

Intuitively, model complexity refers to the flexibility inherent in a model that enables it to fit diverse patterns of data. For the moment, think of it as a continuum, with simple models at one end and complex models at the other. A simple model assumes that a relatively narrow range of more of less similar patterns will be present in the data. When the data exhibit one of these few patterns, the model fits the data very well; otherwise, its fit will be rather poor. All other things being equal, simple models are attractive because they are sufficiently constrained to make them easily falsifiable, requiring a small number of data points to disprove the model. In contrast, a complex model, usually one with many parameters that are combined in a highly nonlinear fashion, do not assume a single structure in the data. Rather, like a chameleon, the model is capable of assuming multiple structures by finely adjusting its parameter values. This enables the model to fit a wide range of data patterns.

There are at least two independent dimensions of model complexity: the number of free parameters in a model and its functional form, which refers to the way in which the parameters are combined in the model equation. For example, \( y = \theta x \) and \( y = x^2 \) have the same number of parameters (1) but differ in functional form. The two dimensions of model complexity, and their interplay, can improve a model's fit to the data but—strangely though it may seem—not improve generalizability. This is illustrated next.

As shown in Table 11.1, four models were compared on their ability to fit two data samples generated by the two-parameter linear model \((P_2)\), which by definition is the true model. Goodness of fit was assessed by finding parameter values for each model that gave the best fit to the first sample. With these parameters fixed, generalizability was assessed by fitting the models to the second sample. In the first row of Table 11.1 are each model's mean fit to data drawn from \(P_2\). As can be seen, \(P_2\) fitted better than \(P_1\), which is an incorrect model having one fewer parameter than the true model. The results for \(P_3\) and \(P_4\) are more interesting. These two models have two more parameters than \(P_2\) and contain the true model as a special case. Note that they both provided a better fit to the data than \(P_2\) itself. Given that the data were generated by \(P_2\), one would have expected \(P_2\) to fit its own data best at least some of the time. But this never happened. Instead, \(P_3\) and \(P_4\) always fitted better. The improvement in fit of \(P_3\) and \(P_4\) over \(P_2\) represents the degree to which the data were overfitted. The two extra parameters in the two models enabled them to absorb nonsystematic variation (i.e., random error) in the data, thus improving fit beyond what is needed to capture the underlying regularity. Note also that \(P_4\) provided a better fit than \(P_3\) (0.79 vs. 0.91), and did so much more often (99% vs. 1%). This difference in fit must be due to functional form because these two models differ only in how the parameters and data are combined in the model equation.

The results in the second row of Table 11.1 demonstrate that overfitting a specific sample of data results in significantly worse generalizability. MSEs are not a good guide for \(P_2\); also, the data yielded the best fit to \(P_2\) less often than expected.

This example demonstrates why the complexity of a fitting model should be considered as well as the best and the most general model that significantly affect generalization. Because overfitting can result in a model that cannot fit new data well, it is essential to find a model that can fit new data. Figure 11.4 illustrates the relationship among model complexity, model goodness of fit, and percent variance accounted for the underlying regularity.
Table 11.1 Goodness of Fit and Generalizability of Models Differing in Complexity

<table>
<thead>
<tr>
<th>Model</th>
<th>$P_1$ (true)</th>
<th>$P_3$</th>
<th>$P_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goodness of fit</td>
<td>1.32 (0%)</td>
<td>0.89 (0%)</td>
<td>0.91 (1%)</td>
</tr>
<tr>
<td>Generalizability</td>
<td>1.39 (6%)</td>
<td>1.06 (52%)</td>
<td>1.14 (21%)</td>
</tr>
</tbody>
</table>

NOTE: Mean squared error of the fit of each model to the data and the percentage of samples in which the particular model fitted the data best (in parenthesis). The four models are as follows: $P_1: y = \theta_1 x + \epsilon$; $P_2: y = \theta_1 x + \theta_2 + \epsilon$; $P_3: y = \theta_1 (\theta_2 x + 1)^{0.5} + \theta_4 + \epsilon$; $P_4: y = \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 + \epsilon$. The error $\epsilon$ was normally distributed with a mean of zero and a standard deviation of 1. A thousand pairs of samples were generated from $P_1$ (true model) using $\theta_1 = 4$ and $\theta_2 = -2.5$ on the same 12 points for $x$, which ranged from 3 to 14 in increments of 1.

of data results in a loss of generalizability. MSEs are now greater for $P_3$ and $P_4$ than for $P_2$; also, the two overly complex models yielded the best fit to the second sample much less often than the true model, $P_2$.

This example demonstrates that the best-fitting model does not necessarily generalize the best and that model complexity can significantly affect generalizability and goodness of fit. Because of its extra flexibility, a complex model can fit a single data set better than a simple model. The cost of the superior fit shows up in a loss of generalizability when fitted to new data sets, precisely because it overfitted the first data set by absorbing random error. Figure 11.4 illustrates the intricate relationship among goodness of fit, generalizability, and model complexity. Fit index such as percent variance accounted for is represented along the vertical axis, and model complexity is represented along the horizontal axis. Goodness of fit increases as complexity increases. Generalizability also increases positively with complexity but only up to the point where the model is sufficiently complex to capture the regularities underlying in the data. Additional complexity beyond this point will cause a drop in generalizability as the model begins to capture random noise, thereby overfitting the data.

In conclusion, a model must not be chosen solely on the basis of its goodness of fit. To do so risks selecting an overly complex model that generalizes poorly to other data generated from the same underlying process. If the goal is to develop a model that resembles the underlying process, then the model must be able to fit all current and future data reasonably well. Only generalizability can measure this property of the model, and thus it should be used in model selection. The next section introduces techniques for measuring generalizability and demonstrates their application.

**Measures of Generalizability**

The trade-off between goodness of fit and complexity illustrated in the preceding example is what makes model selection so difficult. The model must be complex enough to describe the variation in any data sample that is due to the underlying process, yet not overfit the data by absorbing noise and thus lose generalizability. Conversely, the model
must not be too simple to capture the underlying process and thereby underfit the data, which will also lower generalizability. The goal of model selection methods is to estimate a model’s generalizability by weighting fit against complexity. This goal is realized by defining a selection criterion that makes an appropriate adjustment to its goodness of fit by taking into account the contribution of complexity. Five representative methods that are currently in use are the Akaike information criterion (AIC; Akaike, 1973; Bozdogan, 2000), the Bayesian information criterion (BIC; Schwarz, 1978), cross-validation (CV; Browne, 2000; Stone, 1974), Bayesian model selection (BMS; Kass & Raftery, 1995; Myung & Pitt, 1997; Wasserman, 2000), and minimum description length (MDL; Grunwald, 2000; Rissanen, 1983, 1996). The first two of these (AIC, BIC) are limited in their application because they take into account only the number of parameters in their complexity measure. The other three (CV, BMS, MDL) consider the functional form dimension of model complexity as well, either implicitly (CV, BMS) or explicitly (MDL).

The first two methods are defined as follows:

\[
\text{AIC} = -2 \ln f(y | \hat{\theta}) + 2k
\]
\[
\text{BIC} = -2 \ln f(y | \hat{\theta}) + k \ln n
\]

where \( \ln f(y | \hat{\theta}) \) is the natural logarithm of a model’s maximized likelihood, \( k \) is the number of free parameters in the model, and \( n \) is the sample size. When errors are normally distributed, the first term of AIC and BIC, \(-2 \ln f(y | \hat{\theta})\), can be replaced by \((n \cdot \ln \text{SSE} + \text{constant})\). These selection methods prescribe that the model minimizing a given criterion should be chosen.

Note that each of these two criteria consists of two terms: The first represents lack of fit, and the second term is naturally interpreted as model complexity (i.e., \( 2k \) for AIC and \( k \cdot \ln(n) \) for BIC). Model selection is carried out by trading lack of fit for complexity. A complex model with many parameters, having a large value in the complexity term, will not be chosen unless its fit justifies the extra complexity.

AIC and BIC are simple and easy to compute and are by far the most commonly used criteria. The only difference between the two is that BIC includes an extra complexity penalty term for sample size. The BIC is derived as a large sample approximation of BMS, described later.

The number of parameters is the only dimension of complexity that is considered by these two methods. As discussed earlier, functional form can also significantly affect model fit and therefore needs to be taken into account in model selection. The selection methods described next are sensitive to functional form as well as the number of parameters.

The CV, BMS, and MDL methods are defined as follows:

\[
\text{CV} = - \ln f(y_{\text{cal}} | \hat{\theta}_{\text{cal}})
\]
\[
\text{BMS} = - \ln \int f(y | \theta) \pi(\theta) d\theta
\]
\[
\text{MDL} = - \ln f(y | \hat{\theta}) + \frac{k}{2} \ln \left( \frac{n}{2\pi} \right) + \ln \int \sqrt{\det I(\theta)} d\theta.
\]

In the equation, \( \pi(\theta) \) is the prior density of the model parameter, and \( \det I(\theta) \) is the determinant of the Fisher information matrix defined as

\[
I_{ij}(\theta) = - \frac{1}{n} E \left( \frac{\partial^2 \ln f(y | \theta)}{\partial \theta_i \partial \theta_j} \right)
\]

for \( i, j = 1, \ldots, k \) where the expectation \( E \) is taken over \( y \) (e.g., see Robert, 1994, p. 114).

These methods prescribe that the model minimizing a given criterion should be selected.

CV estimates a model’s generalizability without explicitly considering model complexity. In this method, one first randomly

\text{Table 11.2 Cross validation and least squares} \begin{array}{c}
\hline
X \\
\hline
0.806 \\
0.820 \\
0.910 \\
0.911 \\
1.030 \\
1.030 \\
1.155 \\
1.205 \\
1.295 \\
1.475 \\
\hline
\end{array}
\]

\text{NOTE: In this example, the cross validation and least squares estimates define the calibration sample variance accounted for, showing on the last column, } 89.9\%

divides the observation samples, calibration and validation, the former to estimate parameter values of the model, \( \hat{\theta}_{\text{cal}} \), are the parameters to fit the validation data, \( y_{\text{val}} \), yielding the model that represents an estimate of generalizability. This includes the appropriate fit through log likelihood, \( \ln f(y_{\text{val}} | \hat{\theta}_{\text{cal}}) \), and variance accounted for, \( \text{VAF} \). The CV index is defined as a function of validation and calibration data, the unwelcome degree of freedom, and at least minimizing cross validation error for the calibration and validation data, therefore, cannot be minimized by CV. Nevertheless, the term generalizability (being applied...
selection is car-
parameters, hav-
extra com-
extra com-
proximation of
is the only di-
early, func-
account en-
metrics de-
divides the ob-
these values, den-
are then fixed and
fit the validation
equation of the
defined by
model mini-
selected, model
randomly

\[ \hat{\theta}_C \]

\[ \frac{n}{2\pi} \]

\[ \theta \]

\[ E \]

\[ \text{Model Selection} \]

Table 11.2 Cross Validation Example

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y_{cal}$</th>
<th>$Y_{pred}$</th>
<th>$Y_{cal}$</th>
<th>$(Y_{cal} - Y_{pred})^2$</th>
<th>$(Y_{cal} - Y_{pred})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.806</td>
<td>1.132</td>
<td>1.048</td>
<td>1.026</td>
<td>0.007</td>
<td>0.000</td>
</tr>
<tr>
<td>0.820</td>
<td>1.182</td>
<td>1.082</td>
<td>1.217</td>
<td>0.010</td>
<td>0.018</td>
</tr>
<tr>
<td>0.910</td>
<td>1.268</td>
<td>1.306</td>
<td>1.433</td>
<td>0.001</td>
<td>0.016</td>
</tr>
<tr>
<td>0.911</td>
<td>1.293</td>
<td>1.309</td>
<td>1.150</td>
<td>0.000</td>
<td>0.025</td>
</tr>
<tr>
<td>1.030</td>
<td>1.553</td>
<td>1.605</td>
<td>1.745</td>
<td>0.003</td>
<td>0.020</td>
</tr>
<tr>
<td>1.030</td>
<td>1.564</td>
<td>1.605</td>
<td>1.369</td>
<td>0.002</td>
<td>0.056</td>
</tr>
<tr>
<td>1.155</td>
<td>1.819</td>
<td>1.916</td>
<td>1.736</td>
<td>0.009</td>
<td>0.032</td>
</tr>
<tr>
<td>1.205</td>
<td>1.836</td>
<td>2.040</td>
<td>1.940</td>
<td>0.042</td>
<td>0.010</td>
</tr>
<tr>
<td>1.295</td>
<td>2.465</td>
<td>2.264</td>
<td>2.583</td>
<td>0.040</td>
<td>0.102</td>
</tr>
<tr>
<td>1.475</td>
<td>2.775</td>
<td>2.712</td>
<td>2.619</td>
<td>0.004</td>
<td>0.009</td>
</tr>
</tbody>
</table>

NOTE: In this example, the observed data sample was divided into two sub-samples of equal size, calibration ($Y_{cal}$) and validation ($Y_{val}$). Then a two-parameter linear model, $y = \theta_1 x + \theta_2 + \text{error}$, was fitted to the calibration sample, and least square estimates of the parameter values were obtained as $\hat{\theta}_1 = 2.448$ and $\hat{\theta}_2 = -0.958$. These parameter estimates define the model’s prediction ($Y_{pred}$) as $Y_{pred} = 2.448x - 0.958$. From this, mean squared error (MSE) for the calibration sample is obtained as $\text{MSE} = \sqrt{\text{SSE}/m} = \sqrt{0.119}/10 = 0.109$, where SSE stands for the sum of squares error, $\text{SSE} = \sum(Y_{cal} - Y_{pred})^2$, as shown on the fifth column of the table. This MSE is translated into 95.7% of the total variance accounted for (i.e., $1 - \text{SSE}/\text{SST} = 1 - 0.119/2.756 = 0.897$). Similarly, from the SSE, $\Sigma(Y_{cal} - Y_{pred})^2$, shown on the last column of the table, MSE for the validation sample is obtained as $\text{MSE} = \sqrt{0.288}/10 = 0.170$, or equivalently, 89.9% variance accounted for (i.e., $1 - \text{SSE}/\text{SST} = 1 - 0.288/2.853 = 0.899$).

The CV method somehow takes into account the effects of functional form, but how it does this is not clear. Complexity, therefore, cannot be independently measured in CV. Nevertheless, its ease of use and versatility (being applicable for comparing algebraic, algorithmic, and differential-process models) make it an extremely attractive method. Therefore, we recommend its use in model testing, especially when comparing nonnested models. A detailed example of how to use the procedure is provided here and in Table 11.2. Let us again consider Cerella’s (1985) linear model of age-related deficits. Suppose that one wishes to calculate the CV index of this model given the data of Cerella et al. (1981), and further imagine that one divided the original data into two subsamples (calibration and validation), each consisting of ten $(x, y)$ pairs. The open circles and plus signs in Figure 11.2 represent the calibration and validation samples, respectively. Application of CV requires that we first fit the model to the calibration sample and obtain its best-fitting parameter values. This part has already been completed, as described in the section titled "Parameter Estimation." The best-fitting parameter values were $\hat{\theta}_1 = 2.448$ and $\hat{\theta}_2 = -0.958$ with MSE = 0.109, or, equivalently, 95.7% variance accounted for. Now, using these parameter values, predictions from
the model were generated, that is, \( y_{i \in i} = 2.448x_i - 0.958, i = 1, \ldots, 10 \). The predictions, as represented by the thicker solid line in Figure 11.2, were then fitted to the validation sample (plus signs), with no further parameter tuning. When this was done, the resulting MSE was 0.170, with 89.9% variance accounted for, yielding a poorer fit. The latter value represents an estimated generalizability measure of the model in the sense that it is expected, on average, that the linear model will account for about 89.9% of the variance when fitted to all potential data samples generated from the same underlying process. Consequently, the 5.8\% (= 95.7 – 89.9) difference between the calibration and validation samples is the amount by which the calibration sample was overfitted—the amount of random error in the data that was absorbed by Cerella’s 1981 model (Cerella, 1985). Finally, the value of the CV index defined earlier is obtained by combining the MSE value of the validation sample into the minus log likelihood value using Equation (8), in which the parameter standard deviation \( \sigma \) is replaced by the MSE value (0.170), its sample estimate. The resulting value is -3.55.

In BMS the goal is to select the one model among the set of models that maximizes the posterior probability of the model given the data in hand. Under the assumption of equal model priors, the inference leads to maximization of what is called the marginal likelihood, which is the average probability of the data given the model, weighted by the parameter prior density function, \( \pi(\theta) \). BMS itself is defined as the minus logarithm of the marginal likelihood. Under the assumptions of normality and large sample, BMS can be written as

\[
\text{BMS} \approx -\ln f(y | \hat{\theta}) + (1/2) \ln \det(H(\theta)).
\]

In the above equation, \( H(\theta) \) denotes the observed Hessian matrix, whose elements consist of the second derivatives of the minus log likelihood, \( -\ln f(y | \theta) \), differentiated with respect to the parameter vector \( \theta \). The second term of the above expression can be interpreted as a Bayesian complexity measure. Note that the value of the Hessian matrix depends on the functional form of the model’s likelihood function as well as on the number of parameters in the model, as does the Bayesian complexity measure. When the sample size is sufficiently large, BMS is simply reduced to one half of the BIC.

Finally, MDL was developed within the domain of algorithmic coding theory in computer science, where the goal of model selection is to choose the model that permits the greatest compression of data in its description. The assumption underlying this approach is that regularities or patterns in data imply redundancy. The more the data can be compressed by extracting this redundancy, the more we learn about the underlying regularities governing the cognitive process of interest. As with the other selection methods, the first term of MDL is the lack of fit measure. The second and third together constitute the intrinsic complexity of the model. The model that minimizes MDL uncovers the greatest amount of regularity in the data and thus should be preferred.

BMS and MDL are theoretically related to each other (Vitanyi & Li, 2000) and often perform similarly in practice. One drawback in the application of these methods is that they can be computationally intensive, as both require evaluation of numerical integration (see, e.g., Gilks, Richardson, & Spiegelhalter, 1996; Thisted, 1988).

Given the variety of selection methods, it is reasonable to wonder when each is appropriate to use. We end this section by offering a few guidelines. In the ideal situation in which (a) the models being compared are all nested within one another, (b) one of them is correctly specified, and (c) the sample size is sufficiently large (e.g., 200), all five criteria should perform well. However, in most cases compared are not nested and performance will depend on the specific methods (AIC vs. MDL). It is wise to consult different methods (AIC vs. MDL). It is wise to consult the

OTHER IS... Mathematical Modeling

Because of space limitations, we do not discuss a number of model selection methods. The modeller should consult the

 consulted the
should perform equivalently and pick the true model most often. When models being compared are nonnested, functional form has the potential to play a significant role in model performance. In this situation, the first two methods (AIC, BIC) will in general perform worse than the other three (CV, BMS, MDL), which are sensitive to this dimension of model complexity. As a rule of thumb, the latter three may be the safest to use, though there is no guarantee that they will always perform the best. Relative performance of these selection methods can vary considerably depending on the specific set of models being compared, such as nested versus nonnested and correctly specified versus misspecified, and on the sample size, level of random noise, and other characteristics of the data.

Computational considerations will also influence the choice of method. The most general-purpose methods are likely to be MDL and BMS, which perform most accurately across a range of conditions. Unfortunately, they are difficult to implement and require substantial mathematical sophistication to use. The other methods are easier to implement and are likely to perform satisfactorily under restricted conditions. For example, when models have the same number of parameters but differ in functional form, CV is recommended because it, unlike AIC or BIC, is sensitive to this dimension of complexity. If models differ only in number of parameters, then AIC and BIC should do a good job.

**Individual Differences**

Individual differences are an important, though often neglected, topic in mathematical modeling (Luce, 1997). They arise when participants’ data can be fit by the same model but with different values of the model’s parameters. For example, suppose that in a forgetting study one participant’s performance measured by proportion recall decreases according to a power curve with a forgetting rate of 0.25, \(y_{t, \text{sub.1}} = t^{-0.25}\), as a function of the retention interval \(t\). Another participant’s performance may follow the same power curve but with a different value of the forgetting parameter, for example, \(y_{t, \text{sub.2}} = t^{-0.40}\). Whenever individual differences are suspected, averaging across individuals’ data should be done with extreme care; otherwise it can yield a distorted view of the underlying psychological structure, especially for nonlinear models (see, e.g., R. B. Anderson & Teweny, 1997; Ashby, Maddox, & Lee, 1994; Estes, 1956; Hintzman, 1980; Melton, 1936; Myung, Kim, & Pitt, 2000; Siegler, 1987; Singh, 1996). Perhaps a better solution is to analyze the data using a method that takes into account individual differences, such as the hierarchical modeling approach (Bryk & Raudenbush, 1992).

**Random Error in Nonlinear Models**

Random error in a nonlinear model can create the illusion of a statistically reliable and reproducible effect, which in actuality is an artifact. An implicit assumption behind the standard notation of a model, \(y = g(\theta, x) + e\), is that the random error \(e\) is additive to the observed dependent variable \(y\). It may be instructive to examine the integration of other types of random error. For instance, random error could arise inside the mean function \(g(\theta, x)\), instead of outside:

\[y = g(\theta, x_i, e_i) \quad (i = 1, \ldots, m)\]
For example, a model may assume that the data are generated according to the following model equation: \( y_i = \theta_1(x_i + e_i)^2 + \theta_2 \) (\( i = 1, \ldots, m \)). In this equation, \( x_i \) might represent the strength/magnitude of a variable representing an internal mental process (e.g., perceived loudness of a stimulus tone, subjective likelihood of an uncertain event), and because of the way it is coded internally in the brain (e.g., frequency of neuronal spikes), \( x_i \) itself is a random variable with an associated probability distribution. Or, simply, it may not be possible to obtain the exact value of the independent variable \( x_i \), and instead, it is measured with error included. In either case, misspecification of the error for a nonlinear form of the mean function \( g(\theta, x) \) can create an artifactual effect that has no relation to the underlying process that is being modeled. See Erev, Wallsten, and Budescu (1994) for an example. In short, to avoid such pitfalls, one must develop a proper error theory when modeling cognition and take it into account when analyzing data (Busemeyer, 1980; Busemeyer & Jones, 1983; Luce, 1995).

**Equivalent Models**

We often distinguish one model from another by the model's equation. For instance, the power model, \( y = x^\theta + e \), is clearly different from the exponential model, \( y = e^{\theta x} + e \). The look of the model equation can be deceiving, however, because models with distinct equations can often be the same model. For example, the exponential model is equivalent to the model \( y = \alpha^x + e \), as the former can be obtained from the latter through a parameter transformation \( \alpha = e^\theta \). When one model is transformed into another through such a reparameterization, both become equivalent models in the sense that they will fit any given data set with identical precision. Consequently, they are indistinguishable from one another. A similar problem of equivalent models, though in a different sense, arises in covariance structure modeling (see MacCallum, Wegener, Uchina, & Fabrigar, 1993).

As another example of reparameterization-equivalent models, consider the FLMP model (Oden & Massaro, 1978) and the Rasch (1960) model of aptitude defined as follows:

**FLMP:**
\[
g(\theta_i, \lambda_j) = \frac{\theta_i \lambda_j}{\theta_i \lambda_j + (1 - \theta_i)(1 - \lambda_j)}
\]
\[
0 < \theta_i, \lambda_j < 1
\]

**Rasch:**
\[
g(\alpha_i, \beta_j) = \frac{1}{1 + \exp(\alpha_i + \beta_j)}
\]
\[
0 < \alpha_i, \beta_j < \infty.
\]

FLMP the two parameters \( \theta \) and \( \lambda \) represent dimensions of input stimuli (e.g., auditory and visual) and are combined multiplicatively, which is represented by the mean function \( g(\theta_i, \lambda_j) \). On the other hand, in the Rasch model the parameters \( \alpha \) and \( \beta \) are combined in an additive fashion. Despite this difference, however, they are equivalent models under the following parameter transformation (Batchelder, 1997):

\[
\alpha_i = \ln(1 - \theta_i)/\theta_i; \beta_j = \ln(1 - \lambda_j)/\lambda_j.
\]

Thus one can perfectly mimic the other. Therefore, the question of whether the sensory dimensions are combined multiplicatively or additively cannot be answered using statistical tests alone. Answering the question will require use of some nonstatistical means such as an experimental manipulation. In short, for a model with at least one reparameterization-equivalent model, the specific form of the model's equation may not be identifiable.

**CONCLUSION**

The purpose of mathematical modeling is to add precision and clarity to the study of behavior. It forces the scientist to be explicit about the architectural characteristics of the processing system. The data are the detail that alone is known about the mental data. Once the data have been scrutinized in this way (e.g., identifying patterns by comparing human data and the model's predictions, such testing may be used to understand what is valid and—just as importantly—invalid. Another thing that has been to alert the psychologist to what must be ignored. The wider field of software and hardware discipline. Consider the entire piece intended to aid and assist there.

**APPENDIX: GENERATORS FOR NORMAL DISTRIBUTIONS**

In this appendix we present some pseudorandom number generators for five distributions that are commonly used in experimental psychology. Appendix B presents the pseudorandom number generators for other distributions.

In the present approach, we follow Bratley, Fox, and Schrage (1983), useful reference for any interested in an access to a more complete set of numbers, \( U \)'s uniform distribution on \([0, 1]\) is given by

\[
f(y | \mu, \sigma^2) = \frac{1}{\sqrt{2 \pi \sigma^2}} \exp\left(-\frac{(y - \mu)^2}{2 \sigma^2}\right)
\]

where \( \exp(x) \) is the exponential function (that is, e \(^x\)).
processing system, specifying its structure in detail that almost always goes beyond what is known about the process from experimental data. Once specified, the model can be scrutinized in tests of internal consistency (e.g., identifiability) and rigorously evaluated by comparing model performance with human data and then with other competing models. Such tests are only meaningful when we understand what the tests actually measure and—just as importantly—what they do not measure. Another purpose of this chapter has been to alert the reader to the tough problems that must be tackled in order to do modeling. The widespread availability of modeling software bodes well for the future of the discipline. Consider this chapter a companion piece intended to guide would-be modelers and assist them in making informed decisions.

APPENDIX: RANDOM NUMBER GENERATORS

In this appendix we list random number generators for five selected probability distributions that are often assumed in modeling mental processes. For random number generators for other distributions not listed here, see Appendix B of Robert (1994), upon which the present appendix is based. The book by Bratley, Fox, and Schrage (1983) is also a useful reference. We assume that the user has an access to a routine that generates random numbers, $U_i$’s, on the uniform probability distribution on $[0, 1]$.

Normal Distribution. The normal probability distribution of mean $\mu$ and variance $\sigma^2$ is given by

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

where $\exp(x)$ stands for the exponential function (that is, $\exp(x) = e^x$ and $-\infty < y < \infty$.

The following algorithm generates random numbers, $y$, that are normally distributed with mean $\mu$ and variance $\sigma^2$:

Step 1. Generate $U_1, U_2$.
Step 2. Take $x = \sqrt{-2 \ln(U_1)} \cos(2\pi U_2)$.
Step 3. Take $y = \mu + \sigma x$.

Note in the above algorithm that the function $\ln(x)$ denotes the natural logarithm of base $e$, not the logarithm of base 10.

Exponential Distribution. The exponential probability distribution of mean $\alpha$ and variance $\alpha^2$ is given by

$$f(y | \alpha) = \frac{1}{\alpha} \exp(-y/\alpha)$$

where $0 < y < \infty$ and $\alpha > 0$. The algorithm for exponential random numbers is as follows:

Step 1. Generate $U$.
Step 2. Take $y = -\ln(U)/\alpha$.

Beta Distribution. The beta probability distribution of mean $\alpha/(\alpha + \beta)$ and variance $\alpha\beta/((\alpha + \beta)^2(\alpha + \beta + 1))$ is defined as

$$f(y | \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1} (1 - y)^{\beta-1}$$

where $0 \leq y \leq 1$ and $\alpha, \beta > 0$. In the equation, $\Gamma(x)$ denotes the gamma function whose value is equal to $(x-1)(x-2) \cdots 2 \cdot 1$ for a positive integer $x$ but otherwise must be evaluated numerically using a recursion formula. Note that the uniform distribution is obtained as a special case of the beta distribution for $\alpha = \beta = 1$.

The following algorithm generates beta random numbers:

Step 1. Generate $G_\alpha, G_\beta$ (see following)
Step 2. Take $y = G_\alpha/(G_\alpha + G_\beta)$.

The required random number, $G_\alpha$ ($x > 0$), in the above algorithm is generated from one of the following routines depending upon the
value of \( x \):

**Case 1: \( x > 1 \)**

Step 1. Define \( a = x - 1, b = (x - (1/6x))/a, c = 2/a, d = 1 + c, \) and \( e = 1/\sqrt{x} \).

Step 2. Generate \( U_1, U_2 \).

If \( x \leq 2.5 \), then proceed to Step 4. Otherwise, take \( U_1 = U_2 + e(1 - 1.86 U_1) \).

Step 3. If \( U_1 \leq 0 \) or \( U_1 \geq 1 \), then go to Step 2.

Step 4. If \((cU_1 + bU_2/U_1 + U_1/(bU_2)) \leq d \) or \((c \ln(U_1) - \ln(bU_2/U_1) + bU_2/U_1) \leq 1 \), then take \( G_x = aU_2/U_1 \). Otherwise, go to Step 2.

**Case 2: \( x = 1 \)**

In this case, generate the desired random number, \( G_{x=1} \), from the exponential distribution with \( \alpha = 1 \).

**Case 3: \( x < 1 \)**

Step 1. Generate \( U \).

Step 2. Generate \( z = G_{x+1} \) using the above routine for \((x + 1) > 1 \).

Step 3. Take \( G_x = zU^{1/x} \).

**Binomial Distribution.** The binomial probability distribution of mean \( np \) and variance \( np(1-p) \) is given by

\[
 f(y \mid p) = \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y}
\]

where \( y = 0, 1, 2, \ldots, n \), \( 0 \leq p \leq 1 \). The following algorithm generates binomial random numbers:

Step 1. Generate \( U_1, U_2, \ldots, U_n \).

Step 2. Define \( x_i = 1 \) if \( U_i < p \), and \( x_i = 0 \) otherwise, for \( i = 1, \ldots, n \).

Step 3. Take \( y = x_1 + x_2 + \cdots + x_n \).

In essence, this algorithm counts the number of \( n \) uniform random numbers whose values are less than \( p \).

**Poisson Distribution.** The Poisson probability distribution of mean \( \alpha \) and variance \( \alpha \) is given by

\[
 f(y \mid \alpha) = \frac{e^{-\alpha} \alpha^y}{y!}
\]

where \( y = 0, 1, 2, \ldots, \infty \), \( \alpha > 0 \). The following algorithm generates Poisson random numbers:

Step 1. Initialize \( a = 1, b = 0 \).

Step 2. Generate \( U \)

Let \( a = aU, b = b + 1 \)

If \( a \geq e^{-\alpha} \), then go to Step 2. Otherwise, go to Step 3.

Step 3. \( y = b \)

**REFERENCES**


Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive...
Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.


Singh, R. (1996). Subtractive versus ratio model of "fair" allocation: Can the group level analyses...
be misleading? *Organizational Behavior and Human Decision Processes*, 68, 123–144.


CHAPTER 17

**Analysis**

TRISHA VAN ZALK

Response time (RT) data are important measures in experimental psychology and provide a useful basis for testing hypotheses about cognitive processing. The RT data presented in the 24 studies analyzed for this chapter were compiled from the January 2000 issue of *Journal of Experimental Psychology: Human Performance* used as a base of data. All of the data are collected from a large number of experiments in behavioral, developmental, and clinical psychology. The RT data were collected from both human and non-human subjects. Over 27,000 trials were included in the database spanning experiments conducted from April 2000 make up the database.

One reason that RT data are so important in experimental psychology is that they provide a rich source of information like other physical and biological data such as height, weight, or metabolic rate. RT data also provide a scale (Townsend & Masson, 1982) that one can bring to bear on the problem of understanding the psychological machinery that underlies human performance.

One can devise profound research questions

While writing this chapter, I gratefully acknowledge the guidance of the following colleagues who were kind enough to have advice and insights about the research questions, topics, and the development of the tools and theories that underlie the research. I am very grateful for their input. I am also grateful to the many members of my research group who provided feedback and comments on the draft of this chapter. This chapter was also made possible by the generous support of the National Science Foundation.