Chapter 2
Mathematical Modeling

All mathematics texts include story problems. These are almost always of the sort that I call "applications":

- Applications are narrow in scope because they use fixed numbers rather than parameters and because the questions call for answers that are simply numbers. One example is "If a bacteria colony doubles every hour, how long does it take a single bacterium to become a population of one million?" Sometimes the parameter values must be calculated indirectly, as in "A jar initially contains 1 g of a radioactive substance X. After 1 h, the jar contains only 0.9 g of X. How much more time is required before the jar contains only 0.01 g of X?"
- The mathematical setting in an application is implicitly assumed to be exactly equivalent to the scientific setting. Hence, the mathematical answers are unquestioningly accepted as the answers to the scientific questions.

I use the term "applications" for problems with these characteristics because the emphasis is on the mathematics rather than the setting. Some effort may be required for modeling tasks such as determining parameter values and/or interpreting results in context, but all or most of the effort in application problems is in obtaining mathematical solutions. Even when word problems ask more sophisticated questions, they still often suffer from the common interpretation of mathematical models as "mathematical constructions that describe real phenomena in the physical world."

I view mathematical models as "mathematical constructions based on real world settings and created in the hope that the model behavior will resemble the real world behavior." With this interpretation, mathematical modeling is as much theoretical science as it is mathematics. In contrast to applications, model analysis requires us to determine the extent to which models are able to replicate real-world phenomena before we accept mathematical answers as meaningful. Mathematical modeling can be used for narrow questions that call for numerical answers, but it can also be used for broad questions about general behavior (e.g., for what ranges of parameter values will a population go extinct?).

Mathematical modeling requires interdisciplinary skills that are seldom taught in any courses, and this is one important reason why many students who have always had good grades in mathematics find themselves at a loss when they need to use mathematics to do work in science. The purpose of this chapter is to introduce modeling ideas and skills in a small number of simple settings. These ideas and skills are then utilized in the remainder of the text.

Sections 2.1 and 2.2 introduce the ideas of mathematical modeling. A brief example of scientific data is introduced in Section 2.1 to serve as a concrete focus for the development of the chapter. The section also discusses the interplay between data analysis and modeling

G. Ledder, *Mathematics for the Life Sciences: Calculus, Modeling, Probability, and Dynamical Systems*, Springer Undergraduate Texts in Mathematics and Technology, DOI 10.1007/978-1-4614-7276-6_2,

and the limitations of using deterministic models in biology. Section 2.2 presents a qualitative discussion of mathematical modeling, focusing on the differences and connections between mathematics, mathematical modeling, and theoretical science. These introductory ideas provide a framework with which the reader can construct his/her understanding of mathematical modeling.

In standard application problems, the reader is given a mathematical model. In actual practice, mathematical models are not given by some higher authority; hence, mathematical modeling requires skills in obtaining mathematical models, fitting models to data, and selecting among competing models. The remainder of the chapter is focused on these skills. The different topics are unified by the use of the data set of Section 2.1 and the framework of Section 2.2.

Models can be classified as *empirical* and *mechanistic*. Empirical models define a mathematical relationship between quantities in a data set. These are obtained from the general appearance of a data set without regard for underlying biological ideas. Mechanistic models attempt to show how certain quantities in a data set are causally linked to other quantities, independent of any links suggested by data.[1] The distinction between mechanistic modeling and empirical modeling is a central theme of this chapter.

While empirical models are "identified" rather than "derived," it does not follow that empirical modeling lacks mathematical validity. Empirical modeling requires determination of parameter values from data and selection among several candidate models. Both of these tasks can be done with statistical methods, which are developed in Sections 2.3, 2.4, and 2.7. Unlike in story problems, the data needed to determine parameters in mathematical modeling are not exact, and the experimenter collects a surplus of data to compensate for the uncertainty in each measurement. To assign values to parameters, we must identify a quantitative measure of fitting error and then solve the mathematics problem of minimizing that error. This topic is introduced in Section 2.3 on linear least squares and extended in Section 2.4 to models that are linear in one parameter and nonlinear in another. Selection among competing models is discussed in Section 2.7 in terms of the Akaike information criterion (AIC), a method for determining the statistical support provided for a given model by a specific data set. The method is simple to apply, although the mathematical justification is far beyond the scope of any undergraduate course—perhaps this accounts for the curious absence of AIC from elementary texts in statistics. The value that biologists gain from the use of AIC argues for its inclusion in any statistics text written for general use, as well as those written specifically for biologists.

Section 2.5 focuses on the primary task of mechanistic modeling: developing models from biological assumptions based on biological theory and observational data. In Section 2.6, we examine the use of different algebraic forms for the same mathematical model. Differences in form can be as simple as using different symbols for the same quantity, but they can be more complex as well. The modeler often has a variety of ways to define the parameters in a mathematical model, with the choice of parameters affecting the algebraic form. In particular, the work of modeling is almost always simplified by deriving a dimensionless version of the model. Readers of theoretical science papers will often encounter dimensionless models and need to understand what they are and how they relate to the original model. The point of these two sections is not to make the reader an expert on mechanistic modeling, but to give the reader a feel for how mechanistic modeling is done. In particular, Section 2.5 is long and contains some sophisticated modeling. Some readers will want to focus on the ideas, while readers who want to learn how to construct mechanistic models should devote extra time to understanding the details. Additional discussion and examples of mechanistic modeling appear in Chapters 5–7.

[1] Ideally, a model should be both empirically *and* mechanistically based, but the methods for the two types of modeling are distinct.

There are several sets of related problems:

Section	2.3		2.5		2.6	2.7
Fluorine at South Pole	2.3.4					2.7.7
Grape harvests	2.3.8					2.7.9
Lake ice duration	2.3.9					2.7.10
Chemostat		2.5.3	2.5.5	2.5.8	2.6.9	
SIR disease		2.5.4	2.5.6	2.5.7	2.6.10	

2.1 Mathematics in Biology

After studying this section, you should be able to:

- Identify the role of mathematical modeling in science.
- Discuss the concept of demographic stochasticity and apply this concept to biological experiments.
- Generate questions about an experiment that could possibly be addressed with a mathematical model.

The tremendous success of science stems from the interaction of two enterprises: theory and observation.[2] Theory without observation is nothing more than myth, and observation without theory is nothing more than a collection of disjointed facts. Progress in science is therefore possible only by combining them. Theory is used to explain and unify observations and to predict results of future experiments. Observations are used to motivate and validate theory.

The connection between theory and observation is the realm of mathematical modeling. Theory provides statements of scientific principles, while observation provides numerical data. Mathematics is a language that can bridge the gulf between the two. Metaphorically speaking, mathematical modeling is the tendon that connects the muscle of mathematics to the skeleton of science.

2.1.1 Biological Data

Pick up a calculus or precalculus book and find a story problem with a scientific setting. Most likely, the problem you find has exact data. Real scientific data is not exact, and this difference must be understood before we can do mathematical modeling. We can explain the difference here, but you will understand it much better if you discover it yourself. This section will be unnecessary for practicing biologists, but it will be helpful for those readers who have not collected research data themselves.

The real world is not an easy setting for the collection of biological data. Even ignoring the difficulties in getting a good data set, there is the problem that data collection takes a lot of time and effort. This effort is necessary if we are going to practice real science, but it is a distraction if our purpose is to learn mathematical modeling. An alternative to collecting data from an experiment in the real world is to collect data from an experiment in a virtual world. Virtual worlds can be studied in a comfortable chair in front of a computer, without having to wait for events to occur in natural time. If the virtual world is carefully designed, what we learn from it might even be helpful in understanding the real world.

---

[2] I am using the word "observation" to encompass both observation of the natural world and observation directed by experiments.

A famous virtual-world experiment in ecology was conducted by C. S. Holling in the late 1950s, before the capability of creating virtual worlds with computers. Holling was interested in understanding the relationship between the availability of prey and the amount of prey eaten by predators in a given amount of time. He set up a virtual world consisting of sandpaper discs tacked onto a plywood board. The discs represented insects and a blindfolded student represented a predatory bird. In each experimental run, a student tapped the board with a finger at a steady pace, moving randomly around the board. Each time the student touched a sandpaper disk, he removed it, placed it in a cup, and then returned to tapping. After 1 min, the student recorded the number of disks "eaten" in this manner. The data set consisted of pairs of numbers: disks available and disks "eaten." Holling used the data, along with his observations, to create the predation models that now bear his name [6].

The first time I taught mathematical modeling in biology, I recreated Holling's experiment with my class. It was only partly successful, both because the students focused as much on the activity itself as on careful collection of data and also because the virtual world of Holling's human experiment is not well regulated.[3] For the following year, I wrote a software application that creates a virtual world based on Holling's experiment [7]. The application, called BUGBOX-predator, consists of a Windows executable file with supplementary data files, all of which can be downloaded from my web page [8] or the web page for this book: http://www.springer.com/978-1-4614-7275-9.

The BUGBOX-predator world consists of a grid populated by x virtual aphids (the number x is chosen by the experimenter). In each experiment, a virtual coccinellid (ladybird beetle) moves randomly through the virtual-world environment, stopping to consume any virtual aphids in its path. The experiment outcome y is the number of prey animals eaten in 1 min.[4] The animation is unsophisticated, which has the advantage of helping the experimenter appreciate the extreme simplicity of the BUGBOX-predator world.

The problem set for this section consists largely of experiments using BUGBOX-predator. If you do not have a lot of experience collecting biological data, do these problems before reading the rest of the section.

2.1.2 Overall Patterns in a Random World

Table 2.1.1 BUGBOX-predator consumption (y) for various prey densities (x)

Prey density	x	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140
P. steadius	y	0	2	7	10	9	14	21	20	25	20	30	25	29	35	38
P. speedius	y	0	7	11	19	19	22	25	21	25	26	23	27	29	30	29

Table 2.1.1 contains some data from trials with two of the predator "species" in the BUGBOX world. The same data is plotted in Figure 2.1.1. Like real data, the BUGBOX predation data does not appear to fall on a perfectly smooth curve. The BUGBOX-predator world is extremely simple, but is based on rules for biological behavior rather than a mathematical model of the relationship between the variables. The distribution of prey animals is random, and there is also significant randomness built into the predator's movement algorithm. These elements create uncertainty in the outcome, just as in real experiments. It is important to understand

[3] Few students can resist the impulse to tap faster whenever they are having only minimal success.

[4] Given unit time for the experiment and unit area for the environment, we can interpret y as the consumption rate per predator, in prey animals per unit time, and x as the prey density, in prey animals per unit area.

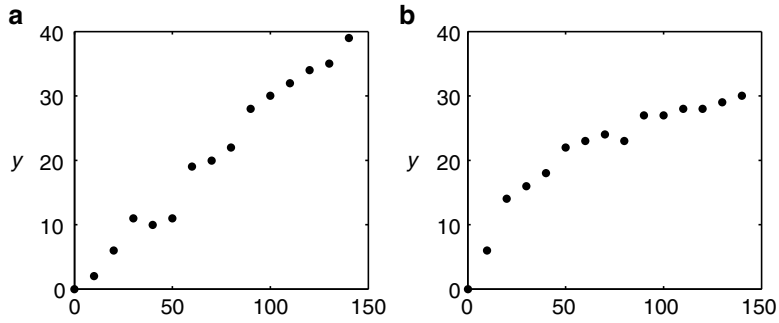

Fig. 2.1.1 Predation data for (**a**) *P. steadius* and (**b**) *P. speedius* from Table 2.1.1

the inevitability of this uncertainty. Real-world experiments can have uncertainty caused by difficulties with experiment design or measurement and by our inability to completely control an experimental environment; our virtual-world experiment is free from these sources of uncertainty. However, there is still variation among individuals. This random effect, called *demographic stochasticity*, is present whenever the number of individuals in an experiment is small and differences in individual behavior are significant. In contrast, experiments in chemistry are free of demographic stochasticity because of the extremely large number of particles in even a small amount of material.[5] This beneficial averaging does not occur in experiments with single individuals. Even if we designed an experiment with a population of 1,000 independently acting predators,[6] the number is insufficient to obtain a fully predictable average.[7]

At first thought, the inevitable randomness in biological data seems to suggest that mathematical modeling is pointless in biology. Mathematics is the most deterministic of disciplines, with many problems defined in such a way that there is a unique solution. If all biological events are affected by random factors, how can mathematics have any value in biology?

Obviously, the highly stochastic nature of biological events limits the possibilities for using mathematical methods in biology. It certainly *is* pointless to attempt to use mathematics to predict how many virtual aphids will be consumed by a virtual coccinellid in a single experimental run. However, mathematics can be used to study the patterns that arise when experiments are repeated many times.

If you run the same BUGBOX-predator experiment (the same predator species and the same initial prey population) hundreds of times, you will see that some outcomes are more likely than others. If your friend also runs the same experiment hundreds of times, the pattern of outcomes obtained by you and your friend will be very similar. In principle, if we repeat a BUGBOX-predator experiment millions of times, the average y for that given x for a particular species might be fully predictable. Averages, and also the expected distribution of results, can be estimated using *descriptive statistics*.[8]

[5] Demographic stochasticity of virus particles is not an issue in a model that tries to predict quantities of these particles in a person suffering from a communicable disease; however, demographic stochasticity in a population of people could be quite significant.

[6] An additional complication occurs if, as would usually be the case, the behavior of each individual is influenced by the behavior of the other individuals in the population.

[7] The connection between number of individuals and predictability is explored in Chapter 4.

[8] See Section 3.1.

2.1.3 Determining Relationships

So far we have only considered the patterns obtained in repetition of a specific experiment. We are often more interested in identifying patterns in relationships between quantities in an experiment. At its most elementary level, this is the goal of ANOVA (analysis of variance), a statistical extension of descriptive statistics. However, ANOVA only seeks to determine the significance of the relationship between quantities. A more ambitious goal is to search for a quantitative description of a relationship. Here again, we have to understand how the highly determined subject of mathematics can possibly be useful in the highly stochastic world of biology. The actual number we would obtain as the average of millions of repeated BUGBOX-predator experiments would be different for different initial numbers of prey. While the data from individual trials show clear evidence of stochasticity (as seen in Figure 2.1.1), it seems reasonable to expect that data from the averages of millions of trials would appear to lie on a smooth curve. That smooth curve is conceivably something that we could identify using mathematics. Think of an actual result of an individual trial as consisting of random variation superimposed on a predictable average y, which depends on the particular x for the experiments. We could then use our limited data set to model that predictable average. **This is the aim of this chapter—to develop methods for obtaining *deterministic* models for the *average* behavior of fundamentally *stochastic* biological processes.**

Examine the data in Figure 2.1.1. Allowing for the overall randomness of the data, there is an obvious difference between the two species. The data for *P. steadius* look approximately linear, but the data for *P. speedius* are definitely not. These kinds of qualitative differences occur in the real world as well as the virtual world. We can attempt to model each species individually; however, a more interesting question is whether we can develop a mathematical model that can explain the data for both predator species. A model that works for a collection of species would have much more value than a model that works for just one.

As noted above, the value of a model depends on the questions it is used to address. The modeler has to develop these questions along with the model. The possibilities are limited primarily by the creativity of the modeler, and the most worthwhile questions are generally not the obvious ones. Here is a brief list, by no means exclusive, of possible questions motivated by the BUGBOX-predator experiments. These are ordered from the most limited to the broadest.

1. If we want to represent the *P. steadius* data by a linear function (a straight line on Figure 2.1.1a), what parameters should we choose?
2. Is there a (clearly nonlinear) function we can use to represent the *P. speedius* data?
3. To what extent can we use the models to predict average predation for x values not given in the data?
4. If we find more than one model for the *P. speedius* data, is there some way to rank the models?
5. Can we identify biological characteristics that could account for the fact that different species have graphs of different shapes?
6. Can we use our predation models as part of a larger model of interacting populations?

The first five of these questions are addressed in the course of this chapter; the last one is addressed in Chapters 5 and 7.

Problems

These problems require BUGBOX-predator, which can be downloaded from http://www.math.unl.edu/~gledder1/BUGBOX/. Save the data sets from these problems, as they will be needed for problems in other sections.

2.1.1. Collect a predation data set for *P. steadius*, using the default choice of no replacement. Use prey values of *approximately* 10, 20, and so on up to 140. It is *not* necessary that the values be exact multiples of 10 (indeed, it is somewhat difficult to accomplish), but they should be roughly evenly spaced. Plot these data on a graph similar to Figure 2.1.1.

(This problem is continued in Problems 2.3.1 and 2.3.7.)

2.1.2. Collect a predation data set for *P. speedius*, using the default choice of no replacement. Use prey values of *approximately* 10, 20, and so on up to 140. It is *not* necessary that the values be exact multiples of 10 (indeed, it is somewhat difficult to accomplish), but they should be roughly evenly spaced. Plot these data on a graph similar to Figure 2.1.1.

(This problem is continued in Problem 2.4.6.)

2.1.3. Repeat Problem 2.1.1, but with replacement.

(This problem is continued in Problems 2.3.1 and 2.3.7.)

2.1.4. From your experience in Problems 2.1.1 and 2.1.3, discuss the significance of the replacement option. Which option would be easier to implement in an experiment with real organisms? Which option allows for unambiguous reporting of data (think about possible differences between what x is supposed to mean and the way we measure it)? This example illustrates the difficulty of designing biological experiments, given the need for practical implementation and the importance of avoiding ambiguity.

2.2 Basic Concepts of Modeling

After studying this section, you should be able to:

- Discuss the relationships between the real world and mathematical models.
- Discuss the distinctions between mechanistic and empirical modeling.
- Discuss the concepts of parameterization, simulation, and characterization with mathematical models.
- Discuss the concepts of the narrow and broad view of mathematical models and the function of parameters in each view.

A *mathematical model* is a self-contained set of formulas and/or equations based on an approximate quantitative description of real phenomena. This definition is useful in the semantic sense, but it fails to distinguish between models that are extremely useful and models that are totally worthless. Instead, I prefer to adopt a working definition that is necessarily vague:

Mathematical model: *a self-contained set of formulas and/or equations based on an approximate quantitative description of real phenomena and created in the hope that the behavior it predicts will be consistent with the real behavior on which it is based.*

Note the tentative language of the added phrase. The emphasis is on the uncertainty in the connection between the mathematical model and the real-world setting to which it is applied. This emphasis means that modeling requires the theoretical science skills of approximation and validation, and it changes the focus of the mathematical skills from proof and solution to characterization (understanding the broad range of possible behaviors) and simulation (visualizing the behavior in specific examples). The thinking you need for mathematical modeling is therefore somewhat different from the thinking associated with mathematics per se and more like the thinking associated with theoretical science, as illustrated in Example 2.2.1.

Example 2.2.1. The Lotka–Volterra model tries to use a linear predation model[9] to explain the quantitative relationship between populations of predators and populations of prey. It was originally developed to explain changes in Mediterranean fish populations that occurred during and after World War I, which it succeeded in doing. Subsequently, it has been used (improperly, as explained in Example 2.2.6 below) in some differential equations textbooks to "prove" that hunting coyotes (to keep them from eating farm animals) increases the population of the coyotes' natural prey *without decreasing the coyote population*. This claim is unsupported by any biological data and is obviously incorrect. □

The coyote–rabbit setting does involve a predator and prey, but it does not follow that just any predator–prey model will be useful. The correct approach is to think of the Lotka–Volterra model as only one *possible* model.[10] Instead of accepting a ridiculous result, such as the impossibility of eliminating predators, we should conclude that the model is wrong.

The lesson of Example 2.2.1 bears frequent repetition:

The value of a model depends on the setting to which it is applied and the questions it is used to address.

Mathematics has the benefit of certainty, as exemplified by proofs of theorems. This is of great value to mathematicians, because it minimizes the time spent arguing about facts. Once a mathematical claim has been proven, everyone is obligated to accept it. However, this certainty only applies to mathematical claims; it does not extend to mathematical modeling. Mathematical results about a model can be confirmed with mathematical proof, and proven results are correct—but only for the model. Conclusions drawn from models are only correct for the real-world setting to the extent that the behavior of the model reflects that of the real world. This question is not mathematical and must be addressed by other means. It is therefore misleading to think of models as "correct" or "true" for a real-world setting. At best, a mathematical model can be *valid*, in the sense of "giving meaningful results under a given set of real-world circumstances." There are almost certainly quantitative differences between model results and real-world empirical results, and there may be important qualitative differences as well. If the differences are small enough in the given setting, we judge the model to be valid and use it with confidence. The model may work for somewhat different settings as well, but we must worry about its validity in the new setting. Where the validation is not satisfactory, we must revise the model and try again.

Example 2.2.2. The exponential decay model

$$y = y_0 e^{-kt}, \qquad k, y_0 > 0 \tag{2.2.1}$$

is valid for a macroscopic amount of a single radioactive substance. The model can also be applied to other settings where a quantity is decreasing or increasing to a fixed value, such as the clearance of medication from the bloodstream of an animal. The ultimate value could be nonzero, in which case we can interpret y as the difference between the current value and the ultimate value. Whatever the context, we have to be careful that the model is appropriate. In lead poisoning, a significant portion of the lead is deposited in the bones, so a more sophisticated model[11] is needed to incorporate this physiological mechanism. Time-release medications are slow to absorb from the digestive system and require a more sophisticated model as well. □

[9] We might use this model for *P. steadius*, but we ought not use it for *P. speedius*.

[10] We return to this scenario later in this section. More appropriate models are presented in Chapter 7.

[11] See Section 7.1.

2.2.1 Mechanistic and Empirical Modeling

Mathematical models can be classified according to the method used to obtain them.

Mechanistic model: *a mathematical model based on assumptions about the scientific principles that underlie the phenomena being modeled.*

Empirical model: *a mathematical model based on examination of numerical data.*

The distinction between the two types of models is sharpened by separating the "approximate quantitative description" in the definition of a mathematical model into two distinct processes: that of approximation and that of quantitative description. To clarify this point, it is helpful to introduce the concept of the *conceptual model*.

Conceptual model: *an approximation of the real world that serves as a verbal description of a mathematical model.*

Conceptual models are seldom explicit in the presentation of mathematical models. Identifying the underlying conceptual model is necessary to understand biological literature that uses mathematics. Identification of conceptual models from examination of mathematical formulas is a recurring theme in the problems sets of this chapter and those of Part III.

Figure 2.2.1 illustrates the processes of mechanistic and empirical modeling. The flow of modeling is not unidirectional. Each of the components feeds into the others, but it is important to note the lack of a direct connection from the mathematical model to the real world. It is this point that distinguishes mathematical modeling from the "applications" of mathematics that appear in most textbooks.

Because of the lack of a clear direction in the flow, we describe these processes in alphabetical order before discussing some key issues in mathematical modeling.

Approximation: An intentional process of choosing features to include in models, analogous to drawing a political cartoon. Caricatures of President Obama always have ears that are much larger than any real person while omitting some of his more subtle features. Yet anyone who has seen the president can easily identify him in a political cartoon. Similarly, a conceptual model in mechanistic modeling focuses on features of the real world that the modeler believes are critical while omitting anything thought to be an unnecessary complication. The hope is that the resemblance of the simple model to the complicated real world will be unmistakable. Approximation is treated in more detail in Section 2.5 and parts of Chapters 5–7.

Characterization: Obtaining general results about a model. Sometimes we can reduce a model to an explicit solution formula. In other cases, we can use graphical or approximation methods to determine how the values chosen for the parameters influence the model behavior. Whatever we learn from characterization applies to the conceptual model, not the real world. Characterization uses techniques of calculus as well as advanced techniques discussed in Parts II and III.

Derivation: Constructing a mathematical model from a verbal description of assumptions and simplifying the model prior to analysis. Sections 2.5 and 2.6 contain examples of model derivation and simplification, as do Chapters 5 and 7. The associated problem sets focus on the reverse skill of identifying the underlying conceptual model from a given mathematical model. Model

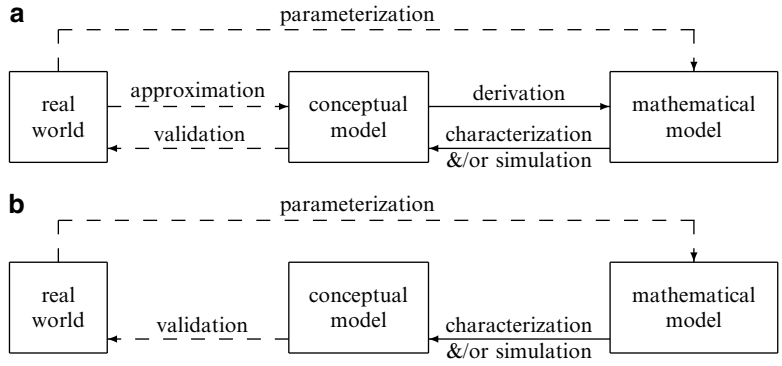

Fig. 2.2.1 Relationships between the real world, a conceptual model, and the corresponding mathematical model. *Solid arrows* indicate processes amenable to mathematical certainty, while *dashed arrows* indicate processes that must be viewed with scientific skepticism. (**a**) Mechanistic modeling. (**b**) Empirical modeling

derivation is a skill needed primarily by mathematical modelers, but anyone who wants to read quantitative biological literature needs to be able to understand the biological assumptions of a model, including assumptions that are only implied.

Model selection: Choosing a mathematical model from multiple options. In mechanistic modeling, construction of a conceptual model constitutes model selection. In empirical modeling, selection of a model should be done with the aid of the Akaike information criterion, a method for quantifying the statistical support a data set gives to a model. This topic is addressed in Section 2.7.

Parameterization: Using data to obtain values for the parameters in a model. Parameterization is necessary if the model analysis is to include simulation. If the analysis is to be general, then it is helpful to have ranges for parameter values; however, precise values fitted to data are unnecessary. Parameterization is addressed in Sections 2.3 and 2.4.

Simulation: Using mathematics and computation to visualize model behavior for a given set of parameters. If an explicit solution formula can be found, then simulation is just a matter of graphing the solution. However, in most cases, it is necessary to implement a numerical method with a computer. Parts II and III contain simulation techniques for specific problem types.

Validation: Determining whether a model reproduces real-world results well enough to be useful. The criteria for validation depend on the purpose of the model. This topic is addressed in more detail later in Section 2.2.2.

Few research projects in biology are simple enough to incorporate all of the processes in Figure 2.2.1a in a single study. One such project is summarized in Example 2.2.3.

Example 2.2.3. The University of Nebraska-Lincoln has a course called *Research Skills for Theoretical Ecology* that focuses on combining experiment and theory to understand the growth of a population of pea aphids.[12]

- Approximation: We assume that the population at day $t+1$ depends only on the population at day t and make specific quantitative assumptions about this dependence.
- Derivation: The assumptions made in the approximation step lead us to a model consisting of formulas for determining day $t+1$ populations from day t populations.

[12] Aphids are particularly convenient organisms for the study of population dynamics because many aphid species exist for long periods as asexual females that reproduce by cloning. Simple models are more likely to yield accurate results when applied to simple systems.

- Parameterization: We use laboratory experiments and statistical analysis to determine values for the various model parameters, such as the average number of offspring an adult produces in 1 day. These parameters are needed so that we can obtain quantitative results.
- Characterization: Independent of parameter values, the model predicts that the population will tend toward stable proportions of nymphs and adults with unrestricted population growth at some fixed rate. This rate is about 32 % per day for the parameter values determined from the laboratory data.[13]
- Simulation: We write a computer program to simulate an experiment in which a population consisting initially of a single adult is allowed to grow for several weeks. Using the parameters determined from laboratory data, the simulation predicts a variable growth rate that approaches 32 % per day over the first 2 weeks.
- Validation: The students measure the quantitative growth rate by experiment and obtain a result of approximately 32 %, as long as the food supply in the experiment is maintained at a high level. However, the unrestricted population growth predicted by the model is qualitatively wrong under conditions where food and space are limited. The model could be used to forecast future aphid populations subject to the condition of virtually unlimited resources and no predators, but not in more general circumstances. □

2.2.2 Aims of Mathematical Modeling

Mathematical models can be used for different purposes, and the aim of the model plays a large role in determining the type of analysis and the criteria for validation. For example, sometimes the goal of modeling is to predict the results of hypothetical experiments.

Example 2.2.4. Mathematicians and biologists at the University of Tennessee have created a sophisticated computer simulation called Across Trophic Level System Simulation (ATLSS) that models populations of animal and plant species in the Florida Everglades [13]. This model has been used by the Army Corps of Engineers and other agencies to predict the effects of environmental policy on the Everglades ecosystem. For example, the model can be used to address the question "What effect would a new housing development in a particular area of the Florida Everglades have on the endangered Florida panther population?" □

A model such as ATLSS needs specific geographic and climate data and specific values for many parameters, such as the average size of a litter of Florida panther cubs and the survival rate for newly hatched whooping cranes. The parameters are estimated for the model because the goal is to predict populations in a hypothetical experiment for a real scenario. Given this goal, the criteria for model validity are quantitative. The model is valid if the results it predicts for experiments are within an acceptable tolerance of the actual experiment results. Demonstrating validity of a model used for quantitative prediction can be difficult. In a laboratory setting, such as that of Example 2.2.3, the model simulation can be designed to match a specific experiment and the results can be directly compared. But we cannot conduct designed experiments for the Everglades. Instead, we look for historical events that can be thought of as experiments, in which case we can match the simulation to the historical event. If we know the effect of a historical housing development on the panther population, we can check to see that our model is quantitatively accurate for that specific case. If so, then we have evidence that our model will correctly predict the effect of a similar hypothetical event.

[13] One could not design a better organism for rapid population growth than the aphid. When reproduction is by cloning, individuals do not need to mature before they begin the reproduction process. Indeed, pea aphids are born pregnant and begin to give birth to live young within hours after becoming adults.

Quantitative prediction for hypothetical experiments is not all that can be done with mathematical models. In Section 1.5, we used a mathematical model to obtain a prediction for foraging behavior in a patchy environment. This required mathematical characterization rather than numerical simulation, and it resulted in the marginal value theorem, which we can think of as a qualitative prediction for a general setting. Broad questions, such as that of optimal foraging behavior, require mathematical characterization rather than numerical simulation. Other questions are less broad but require characterization of a general model because parameter values can vary.

Example 2.2.5. The side effects of chemotherapy and its efficacy against tumors depend on the dosage schedule. We could administer the medication at a constant rate over a fixed time interval, we could make the rate large initially but then decrease it to zero over some interval, or we could choose any other time-dependent dosage schedule. A reasonable goal for modeling is to identify a protocol that minimizes side effects while reducing the tumor at a desired rate. One way to attempt this is to run simulations with a chemotherapy model, but we could only test a few of the infinitely many dosing protocols. Instead, it is possible to use mathematical methods to obtain an approximate solution to the problem of optimizing the dosage schedule. □

The validation of a model intended to address broad theoretical questions is different from that for a model intended for quantitative prediction. For a general model, the task of validation involves trying to confirm that the model behavior is qualitatively consistent with the behavior of the real biological system we are trying to model. Of course, we must specify what we mean by "consistent with the real behavior." For example, a model whose purpose is to study extinction risk for endangered species would need to be checked to ensure that it is actually capable of predicting extinction under some set of circumstances.

Example 2.2.6. Suppose we want to know what effect coyote hunting will have on a coyote population (the question of Example 2.2.1). From our reading of the biology literature, if not from direct experience, it should be obvious that predator extinction is a possibility. However, characterization of the Lotka–Volterra model shows it to be incapable of predicting this possibility. The proper response to this mathematical result is to immediately reject the use of the Lotka–Volterra model for the given setting. Careful examination of the conceptual model can identify the flaw that accounts for its failure to make correct qualitative predictions, which in turn can suggest a better model.[14] □

2.2.3 The Narrow and Broad Views of Mathematical Models

In any particular instance of a mathematical model, we have one or more dependent variables and one or more independent variables, and a set of given values are assigned to the parameters. The focus of a simulation is on determining how the dependent variables depend on the independent variables. This is the *narrow view* of mathematical models. In contrast, there is a

[14] See Problem 2.2.1.

broad view of mathematical models, in which the objective is to understand the effect of the parameter values on the model behavior. The relationship between these views is illustrated in Figure 2.2.2. In the broad view, the role of "independent variable" is played by the parameters and the role of "dependent variable" is played by whatever aspects of the model behavior are of interest.

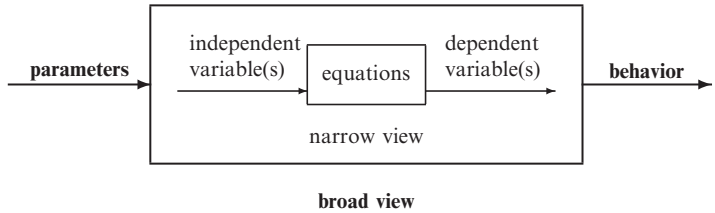

Fig. 2.2.2 Narrow and broad views of mathematical models

Example 2.2.7. Consider the family $y = \sin kt$, which is sometimes used as an empirical model for periodic data. When we choose a specific value for k and plot y as a function of t, we are working in the narrow view. Without choosing k, we can calculate the period to be the smallest time T such that $y(t+T) = y(t)$. Since the sine function repeats as the angle increases by 2π, the period is when $kT = 2\pi$, or $T = 2\pi/k$. If we plot T as a function of k, we are working in the broad view. □

As in Example 2.2.7, parameters function as constants in some aspects of model analysis and as variables in other aspects, corresponding to the narrow and broad views, respectively. This can be very confusing. Generally we are working with the narrow view for simulations and the broad view for characterization. Both are important. We can make use of the full power of computers for simulations, but we can address deeper questions when we retain the broad view.

2.2.4 Accuracy, Precision, and Interpretation of Results

Most people make little distinction in ordinary language between the terms "accuracy" and "precision," but these terms have very distinct meanings in science. *Accuracy* is the extent to which results are correct, while *precision* is the extent to which results are reproducible. Precision is easier than accuracy to measure, but of course it is accuracy that we really need. Nevertheless, one cannot be confident of accuracy in the absence of precision. A repeated theme of this book, starting with the discussion of biological data in Section 2.1 and continuing through the rest of the book, with special emphasis in Chapter 4 on probability distributions of samples, is that precision is limited in most areas of biology. Even where careful measurements are possible, results are not very reproducible. Have your blood pressure taken on five consecutive days, and you will see the point.

The lack of precision in most biological data has strong implications for how we interpret mathematical results. Computers give very precise results—divide 1.0 by 3 and you will not get 0.33, but 0.333333333333. This is alright if the numerator is certain to be very close to 1.0, but in biology it could be that you measured 1.0 when the "correct" value is 0.9. If the data is off by 10 %, then the additional digits beyond the second one are surely meaningless. Mathematics, of course, offers the possibility of infinite precision. In terms of biology, this does

more harm than good. Apply infinitely precise methods to crude results and you get results that are infinitely precise in appearance without being reproducible. It is easy to take this apparent precision more seriously than it deserves. This is what I call the "measure it with your hand, mark it with a pencil, cut it with a laser" fallacy. The risk of this fallacy must be kept firmly in mind whenever we interpret results obtained from mathematical modeling applied to crude data.

Problems

2.2.1. Suppose we want to construct a realistic model for a predator–prey system. This model should allow for a variety of realistic results; in particular, it should predict three possible long-term results:

1. The predator and prey can coexist.
2. The prey can survive while the predator becomes locally extinct.
3. Both species can become locally extinct.

We look in a mathematical biology book and find the Lotka–Volterra model:

$$\frac{dx}{dt} = rx - qxy,$$

$$\frac{dy}{dt} = cqxy - my,$$

where x and y are the biomasses of the prey and predator, respectively,[15] r is the growth rate of the prey, m is the death rate of the predator, q measures the extent of predation, and c is a conversion factor for prey biomass into predator biomass. The model is presented as a predator–prey model, but we recognize the need to check that a model is appropriate for the setting we have in mind.

(a) Suppose the prey and predator populations stabilize to fixed biomasses $X \geq 0$ and $Y \geq 0$. If the biomasses at some time are $x = X$ and $y = Y$, then there should be no further change. This means that the fixed biomasses must make the right-hand sides of the differential equations be 0. Use this idea to find all possible pairs X and Y. Is this model capable of predicting all three possible long-term results? If not, which is missing?
(b) To help see what is wrong with the model,[16] write down the prey equation for the special case where the predator is absent. What does the model predict will happen?

2.2.2. Find an instance of a mathematical model in a biology book or research paper. Describe:

(a) The mathematical model itself,
(b) The conceptual model that corresponds to the mathematical model, and
(c) Features of the real-world setting that do not appear in the conceptual model.

2.2.3.* Some genetic traits are determined by a single gene having two variants (alleles), with one (**A**) dominating the other (**a**). This means that individuals who have two dominant alleles (**AA**) and individuals who have one of each type (**Aa**) both exhibit the physical

[15] Most descriptions of predator–prey models interpret the variables as the numbers of individuals, but the models are more realistic if the variables are viewed as being the total biomass of the individuals.

[16] The point here is that using the Lotka–Volterra model to demonstrate that something can't happen in the real world is a logical fallacy when the model itself contains the assumption that the thing can't happen.

characteristics (phenotype) of the dominant trait, while the recessive phenotype is only found among individuals who have two recessive alleles (**aa**). It is sometimes helpful in genetics to model inheritance as a two-step process: first, all of the parents' genes are assembled into a gene pool; then, pairs of genes are randomly withdrawn from the gene pool for individuals in the next generation.

(a) Suppose q is the fraction of recessive genes in the gene pool. Based on the two-step conceptual model, what will be the fraction of individuals in the next generation who exhibit the recessive trait? What will be the fraction of individuals who have one dominant allele and one recessive allele? What will be the fraction of individuals who do not have the recessive allele? (The combination of these results is called the *Hardy–Weinberg principle*.)
(b) About 13 % of the people of Scotland have red hair. Assuming that red hair is caused by a single recessive gene pair, what does the Hardy–Weinberg principle predict for the fraction of the recessive trait in the gene pool and the fraction of the population who do not have the recessive allele?
(c) Demographers estimate that 60 % of the people of Scotland do not have the recessive allele for red hair. What flaws in the conceptual model might account for the difference between this estimate and the estimate you obtained from the Hardy–Weinberg principle?

2.2.4. The model $y = y_0 e^{-kt}$, with $k > 0$, is often used to model radioactive decay, where y is the amount of radioactive material remaining and y_0 is the initial amount of the material. This model is derived in Section 2.5 using a conceptual model in which the decay rate is k times the quantity of material. Without calculus, we can get some sense of what this means. Consider the specific instance $y = e^{-2t}$. The average rate of decay over the interval $t_1 < t < t_1 + h$ is

$$r_h(t_1) = \frac{y(t_1 + h) - y(t_1)}{h}.$$

For the intervals $0 < t < 0.1$, $0.1 < t < 0.2$, and so on up to $0.9 < t < 1.0$, calculate the average rate of decay and compare it to the average of the quantities of radioactive material at the beginning and end of the time interval. Explain why the results are consistent with the conceptual model as described here.

2.2.5. The model $y = y_0 e^{kt}$, with $k > 0$, is sometimes used to model bacterial growth.

(a) Describe the qualitative predictions made by the model. In particular, show that

$$G(t) = \frac{y(t+1)}{y(t)}$$

does not actually depend on t.
(b) Describe an experiment that tests the prediction of part (a).
(c) Describe a physical setting in which this model for population growth is clearly not appropriate.
(d) Describe a physical setting in which this model for population growth might be appropriate.

2.2.6. Of Problems 2.2.4 and 2.2.5, one works primarily with the narrow view of a model and the other primarily with the broad view. Match these descriptions with the problems, explaining why that view is the focus of the problem.

2.2.7. In this problem, we develop and study a model to predict the future effects of changes in the average number of children per adult and the average age for childbirth on human population

growth rates in a highly developed country. The model is based on the Euler–Lotka equation, which was developed in Section 1.7:

$$\int_0^\infty e^{-rt}\ell(t)m(t)\,dt = 1\ ,$$

where r is the unknown population growth rate resulting from a fecundity of $m(t)$ births per year per parent of age t, given a probability of $\ell(t)$ of survival to age t.

(a) For the sake of a simple thought experiment, we assume particularly simple forms for ℓ and m. We take $\ell = 1$ since survival to adulthood is high in highly developed countries. For m we assume a piecewise linear function with peak at age a, tapering to 0 at ages $a-5$ and $a+5$, and having a total (integrated over time without the factor e^{-rt}) of n. This will allow us to modify just the parameters a and n rather than the whole fecundity function m. Show that the function

$$m(t) = 0.04n \begin{cases} 0, & t < a-5 \\ 5+t-a, & a-5 < t < a \\ 5-t+a, & a < t < a+5 \\ 0, & t > a+5 \end{cases}$$

has all of the desired properties.

(b) Substitute $\ell = 1$ and the function m from part (a) into the Euler–Lotka equation to obtain the integral equation

$$0.2n\left[\int_{a-5}^{a}\left(1+\frac{t-a}{5}\right)e^{-rt}\,dt + \int_a^{a+5}\left(1-\frac{t-a}{5}\right)e^{-rt}\,dt\right] = 1.$$

(c) Make the substitution $x = (t-a)/5$ to simplify the integrals. You should now have one integral with $-1 < x < 0$ and one with $0 < x < 1$.

(d)* Make an additional substitution $y = -x$ in the integral on $-1 < x < 0$ and combine the two integrals into a single integral on the interval $(0,1)$. See Problem 1.9.9.

(e) You should now have an equation of the form

$$ne^{-ra}F(r) = 1\ ,$$

where $F(r)$ is a complicated definite integral. Show that $F(0) = 1$ and $F(0.02) < 1.001$. You can do the latter by calculating the integral (see Problem 1.8.12) or by numerical approximation (see Problem 1.7.9). The function F is strictly increasing, so the approximation $F(r) = 1$ has error less than 0.1 % if $r \le 0.02$.

(f) Indicate at least one biological assumption in this model that we can expect to introduce more than a 0.1 % error; conclude that the approximation $F = 1$ is fully justified.

(g) Plot a graph of r against n, with three curves using different values of a, and use this curve to discuss the effects of average number of children and average age of reproduction on population growth. To do this intelligently, you must choose reasonable low, medium, and high values for a and a reasonable range of n values for the horizontal axis.

(h) Repeat part (g), reversing the roles of n and a.

2.2.8. Suppose individuals of group X and individuals of group Y interact randomly. If x and y are the numbers of individuals in the respective groups, it is reasonable to expect each member of group Y to have kx interactions with members of group X, where $k > 0$ is a parameter. (This says that doubling the membership of group X should double the contact rate with group X for a member of group Y.)

(a) Use the information about contact rates for individuals in group Y to find a model for the overall rate at which members of the two groups interact.
(b) Suppose p is the (fixed) fraction of encounters between individuals of the two groups that results in some particular event occurring between the individuals. Use this assumption to create a model for the rate R at which the events occur in the population.
(c) If the model of part (b) is used for the rate of infection of human populations with some communicable disease, what do the groups X and Y represent?
(d) The model of part (b) was used successfully to model an influenza outbreak in a small boarding school in rural England. Why do you think the model worked well in this case?
(e) The Center for Disease Control in Atlanta did not use the model of part (b) to make predictions about the spread of the H1N1 virus in the United States in 2009. Explain why the model would not have been appropriate in this case.
(f) Describe some real-world settings other than epidemiology that could conceivably use this interaction model [Hint: This model finds common usage in chemistry and ecology as well as epidemiology.] How accurate do you expect the model to be in these different settings?

2.3 Empirical Modeling I: Fitting Linear Models to Data

After studying this section, you should be able to:

- Use the linear least squares method to obtain the best-fit parameter values for the linear models $y = mx$ and $y = b + mx$.
- Discuss the assumptions made in claiming that the results of the linear least squares method are the best parameter values for the data.

Simulations require values for the model parameters, which raises the question of how parameter values should be determined. Occasionally they can be measured directly, but more often they can only be inferred from their effects. This is done by collecting experimental data for the independent and dependent variables and then using a mathematical procedure to determine the parameter values that give the best fit for the data. In this section, we develop a parameterization method for two linear models:

$$y = mx, \qquad y = b + mx\,, \tag{2.3.1}$$

where m and b are parameters. These models, together with the exponential model,

$$y = Ae^{kx}\,, \tag{2.3.2}$$

and the power function model,

$$y = Ax^{p}\,, \tag{2.3.3}$$

comprise the principal empirical models commonly encountered.[17] Parameterization of exponential and power function models is considered in Section 2.4.

[17] The symbols in these models are generic; that is, they represent whatever actual variables are in a given model. For example, a model $H = CL^q$ is a power function model with independent variable L, dependent variable H, exponent parameter q, and coefficient parameter C. Most symbols in mathematics are not standard, so the reader must be able to identify models as equivalent when the only difference is the choice of symbols. This theme is extended much further in Section 2.6.

2.3.1 The Basic Linear Least Squares Method ($y = mx$)

Table 2.3.1 Predation rate y for prey density x

x	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140
y	0	2	7	10	9	14	21	20	25	20	30	25	29	35	38

Table 2.3.1 reproduces the *P. steadius* data from Table 2.1.1. Because the data points appear to lie roughly on a straight line through the origin (Figure 2.1.1a), it makes sense to try a linear model without a parameter to represent the y intercept; that is, $y = mx$ rather than $y = b + mx$.[18] The parameterization process consists of finding and solving a mathematics problem to determine a value of m for this data set.

Obviously there is no single value of m for which the model $y = mx$ fits the data exactly. For any given value of m, some or all of the data points lie off the graph of the model. Figure 2.3.1 shows the data with several possible straight lines. Clearly, the slope m for the top line is too large and that for the bottom is too small.

2.3.1.1 Overview of the Method

Any optimization problem has a modeling step and an analysis step:

1. Determine a function that expresses the quantity to be maximized or minimized in terms of one or more variable quantities.
2. Determine the set of values for the variable quantities that yields the maximum or minimum function value among permissible values of the variables.

Optimization problems for mathematical models are conceptually more difficult than optimization problems in calculus because of the different roles played by variables and

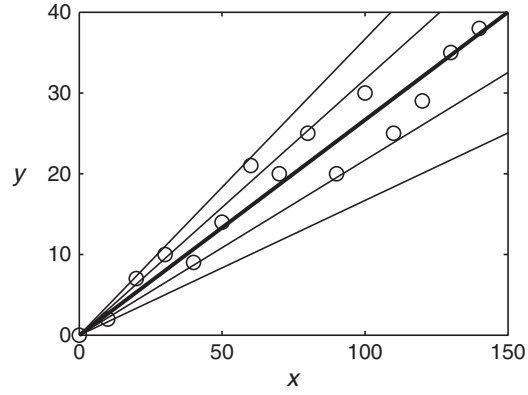

Fig. 2.3.1 Consumption rate y for prey density x from Table 2.3.1, showing several instances of the model $y = mx$; the *heavy line* is the instance that will emerge as the best fit

[18] There are two advantages to omitting the parameter b. Mathematically, it is much easier to find one parameter from data than two. More importantly, the model $y = mx$ may be more appropriate on biological and/or statistical grounds, as will be seen in Sections 2.5 and 2.7, respectively.

parameters. In our current example, the variables x and y represent specific data points, while the parameter m is unknown.

When parameterizing a mathematical model from data, the *parameters* in the model are the *variables* in the optimization problem, while model variables appear in the optimization problem only as labels of values in the data set.

The imagery of the narrow and broad views of mathematical modeling (Figure 2.2.2) is helpful in thinking about the problem of determining a best value of m. In step 1, we assume a fixed value of the parameter m, generate a set of "theoretical" (x,y) data points using the model $y = mx$ with the x values from the actual data, and calculate some quantitative measure of the total discrepancy between the actual data and the data obtained using the model. This step occurs within the narrow view because m is fixed. Once we have a formula for calculating the total discrepancy, we change our perspective. Now we think of the *data* as fixed and the total discrepancy for that fixed data set as a function $F(m)$. We then obtain the optimal value of m using methods of ordinary calculus. Treating m as a variable locates step 2 within the broad view.

2.3.1.2 Quantifying Total Discrepancy

There are several reasonable ways to measure quantitative discrepancy. The standard choice is

$$F(m) = (\Delta y_1)^2 + (\Delta y_2)^2 + \cdots + (\Delta y_n)^2 \,, \tag{2.3.4}$$

where the *residuals* Δy_i are the vertical distances between the data points and the corresponding points from the model. For the model $y = mx$, we have

$$\Delta y_i = |mx_i - y_i| \,. \tag{2.3.5}$$

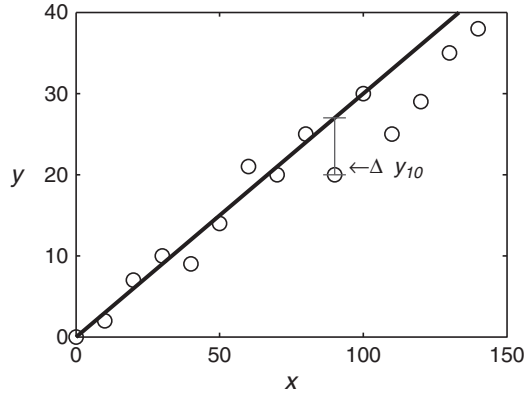

Fig. 2.3.2 Consumption rate y for prey density x, showing the model $y = 0.3x$ and the residual for $x = 90$

Example 2.3.1. Let $m = 0.3$. The data, model, and one of the residuals are shown in Figure 2.3.2. The total discrepancy is $F(0.3) = 218$, calculated as the sum of the bottom row of Table 2.3.2. □

Table 2.3.2 Total discrepancy calculations for $y = 0.3x$ with the *P. steadius* data

x	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140
$0.3x$	0	3	6	9	12	15	18	21	24	27	30	33	36	39	42
y	0	2	7	10	9	14	21	20	25	20	30	25	29	35	38
Δy	0	1	-1	-1	3	1	-3	1	-1	7	0	8	7	4	4
$(\Delta y)^2$	0	1	1	1	9	1	9	1	1	49	0	64	49	16	16

Check Your Understanding 2.3.1:
Repeat Example 2.3.1 for the model $y = 0.25x$. Is this model better or worse than $y = 0.3x$?

Total discrepancy calculations are easily automated with a spreadsheet; however, we cannot do the calculation for all possible values of m. We shall see that the optimal m can be determined mathematically, without actually computing any values of $F(m)$.

2.3.1.3 Minimizing Total Discrepancy

The problem of choosing m to minimize F is surprisingly easy to solve. Substituting (2.3.5) into (2.3.4) and expanding the squares, we have

$$F(m) = (m^2x_1^2 - 2mx_1y_1 + y_1^2) + \cdots + (m^2x_n^2 - 2mx_ny_n + y_n^2),$$

which we can rearrange as

$$F(m) = (x_1^2 + x_2^2 + \cdots + x_n^2)m^2 - 2(x_1y_1 + x_2y_2 + \cdots + x_ny_n)m + (y_1^2 + y_2^2 + \cdots + y_n^2).$$

We can simplify this formula using summation notation[19]:

$$F(m) = \left(\sum x^2\right) m^2 - 2\left(\sum xy\right) m + \left(\sum y^2\right) , \tag{2.3.6}$$

where the sums are understood to be over all of the data points. In the context for (2.3.6), the data points are known; hence, the total discrepancy formula is a function of a single variable m.[20] The function is a simple parabola pointing upward, so we need only find the vertex of that parabola to obtain the important mathematical result[21]:

[19] For ease of reading, I use a simplified form of summation notation. What I have as $\sum xy$, for example, is more properly given as $\sum_{i=1}^{n} x_iy_i$. In the given context, the extra notation decreases readability unnecessarily.

[20] Context is crucial. As noted earlier, the parameter m functions as a *constant* in the model $y = mx$ (narrow view) but as a *variable* in the total discrepancy function F (broad view). Meanwhile, x and y are variables in the model, but the data points (x_i, y_i) function as parameters in the total discrepancy calculation because we have a fixed set of data.

[21] The proof of Theorem 2.3.1 is given as Problem 2.3.10.

Theorem 2.3.1 (Linear Least Squares Fit for the Model $y = mx$). *Given a set of points (x_i, y_i) for $i = 1, 2, \cdots, n$, the value of m that minimizes the total discrepancy function for the model $y = mx$ is*

$$m = \frac{\sum xy}{\sum x^2} \; ; \tag{2.3.7}$$

the corresponding residual sum of squares is

$$RSS = \sum y^2 - \frac{(\sum xy)^2}{\sum x^2} = \sum y^2 - m \sum xy \; . \tag{2.3.8}$$

The **residual sum of squares** is the total discrepancy for the model when the best value of m is used; that is, it is the minimum value of the function F. It will be needed for the semilinear data fitting scheme of Section 2.4 and the model selection scheme of Section 2.7.

We now have the mathematical tools needed to find the optimal value of m for the *P. steadius* data set.

Example 2.3.2. For the data of Table 2.3.1, we obtain the results

$$\sum x^2 = 101,500, \qquad \sum xy = 27,080, \qquad \sum y^2 = 7,331 \; ;$$

therefore, (2.3.7) and (2.3.8) yield the results $m \approx 0.267$ and RSS ≈ 106.1. The best-fit line is the heavy one in Figure 2.3.1. □

Check Your Understanding 2.3.2:
Verify the values given in Example 2.3.2.

2.3.2 Adapting the Method to the General Linear Model

Most straight lines in a plane do not pass through the origin. While there are theoretical reasons for insisting that the predation model pass through the origin, this is obviously not valid for *all* linear models; hence, Theorem 2.3.1 would seem to be of limited use. However, the problem of fitting the model $y = b + mx$ to data can be reduced to the problem of fitting the model $y = mx$ to data. The derivation of the mathematical result for this two-parameter model is given as Problem 2.3.11.

Theorem 2.3.2 (Linear Least Squares Fit for the General Linear Model $y = b + mx$). *Let $\bar{x}$ be the mean of the values $x_1, x_2, \ldots, x_n$, let $\bar{y}$ be the mean of the values $y_1, y_2, \ldots, y_n$, and define shifted data points by*

$$X_i = x_i - \bar{x} \; , \quad Y_i = y_i - \bar{y} \; , \quad \text{for } i = 1, 2, \ldots, n \; . \tag{2.3.9}$$

Then

1. *The best-fit slope and residual sum of squares for the model $y = b + mx$ can be found by fitting the XY data to the model $Y = mX$;*
2. *The best-fit intercept b is given by*

$$b = \bar{y} - m\bar{x} \,. \tag{2.3.10}$$

Example 2.3.3. To fit the model $y = b + mx$ to the data of Table 2.3.1, we first compute the means $\bar{x} = 70$ and $\bar{y} = 19$. Then we subtract the means from the original data set to obtain a shifted data set, as shown in Table 2.3.3. Applying Theorem 2.3.1 to the shifted data yields the results

$$m = 0.255 \,, \qquad \text{RSS} \approx 100.4 \,.$$

By Theorem 2.3.2, these results hold for the model $y = b + mx$ with the original data, and we calculate b from (2.3.10):

$$b = 1.175 \,.$$

□

Table 2.3.3 Consumption rate y for prey density x, along with the shifted data $X = x - \bar{x}$, $Y = y - \bar{y}$

x	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140
y	0	2	7	10	9	14	21	20	25	20	30	25	29	35	38
X	−70	−60	−50	−40	−30	−20	−10	0	10	20	30	40	50	60	70
Y	−19	−17	−12	−9	−10	−5	2	1	6	1	11	6	10	16	19

Check Your Understanding 2.3.3:
Verify the values given in Example 2.3.3.

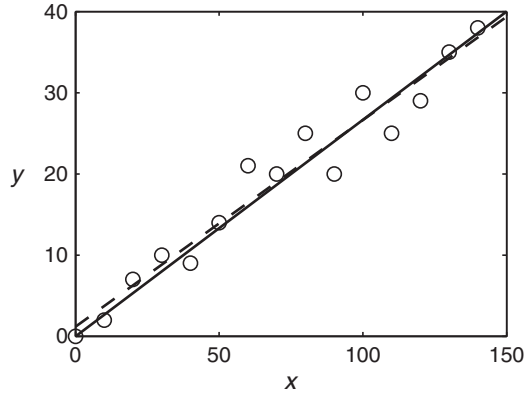

Fig. 2.3.3 Consumption rate y for prey density x, showing the linear least squares fits for the models $y = mx$ (*solid*) and $y = b + mx$ (*dashed*)

We now have two best-fit results for the Table 2.3.1 data: the line $y = 0.267x$ from Example 2.3.2, with residual sum of squares 106.1, and the line $y = 1.175 + 0.255x$ from Example 2.3.3, with residual sum of squares 100.4. Figure 2.3.3 shows both of these lines along with the data. Does the slightly lower residual sum of squares mean that the two-parameter model is better than the one-parameter model? Not necessarily. The calculation of the residual sum of squares treats all data equally. However, the data point $(0,0)$ is free of experimental uncertainty, so perhaps we should be less tolerant of the discrepancy $\Delta y_1 = 1.175$ than the discrepancies at the other points. Perhaps we should insist that $y(0) = 0$ is a *requirement* for our model, even though doing so slightly increases the residual sum of squares. We take up this issue in Section 2.7 after we have laid more groundwork.

2.3.3 Implied Assumptions of Least Squares

The sum of squares of vertical residuals is not the only function that could be used to quantify the total discrepancy between an instance of a model and a set of data. We need some way to measure the distance between each point and the model line, but why not measure it horizontally or along a normal to the line? In choosing Δy as the measure of discrepancy, we are implicitly assuming that we have very accurate measurements of x and uncertainty primarily in y. This is frequently true, but not always. However, it is common to use $(\Delta y)^2$ even in cases where there is significant uncertainty in x values because the least squares formula is so easy to apply.[22]

Check Your Understanding Answers

Table 2.3.4 Total discrepancy calculations for $y = 0.25x$ with the *P. steadius* data

x	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140
$0.25x$	0	2.5	5	7.5	10	12.5	15	17.5	20	22.5	25	27.5	30	32.5	35
y	0	2	7	10	9	14	21	20	25	20	30	25	29	35	38
Δy	0	0.5	2	2.5	1	1.5	6	2.5	5	2.5	5	2.5	1	2.5	3
$(\Delta y)^2$	0	0.25	4	6.25	1	2.25	36	6.25	25	6.25	25	6.25	1	6.25	9

1. The total discrepancy is $F(0.25) = 134.75$, which is less than $F(0.3) = 218$.

Problems

2.3.1. (Continued from Problems 2.1.1 and 2.1.3.)

(a) Fit the model $y = mx$ to your *P. steadius* data from Problem 2.1.1.
(b) Fit the model $y = mx$ to your *P. steadius* data from Problem 2.1.3.
(c) How much different (in percentage) are your results from the result in the text?
(d) Discuss whether or not replacement appears to be a significant source of differences in the results.

(This problem is continued in Problem 2.7.4.)

[22] See [10] for a much more complete discussion of this topic.

2.3.2. One of the data sets in Table 2.3.5 has the origin as its mean point. Find the equation of the straight line that best fits that data. Plot the data and the best-fit line together on a graph.

Table 2.3.5 Two *xy* data sets for Problem 2.3.2

x	−4	−1	0	2	3
y_1	−5	−2	0	2	4
y_2	−5	−2	1	2	4

2.3.3.* The data sets of Table 2.3.6 contain a parameter c that perturbs some of the data points away from the straight line $y = x$, while still maintaining an average y of 0. By examining the change in slope m as a function of c, we can measure the effect of measurement error on the result of the least squares procedure. To do this, plot the linear regression slope m as a function of the parameter c, where $0 \leq c \leq 1$, for each data set. How do measurement errors affect the least squares line? In particular, which errors are the least squares line more sensitive to?

Table 2.3.6 Two data sets for Problem 2.3.3

x_1	−2	−1	0	1	2
y_1	−2+c	−1	0	1	2−c
x_2	−2	−1	0	1	2
y_2	−2	−1+c	0	1−c	2

2.3.4. (a) Fit a linear model to the data of Table 2.3.7 (using a calculator to do the computations). The data gives the concentration C, in parts per trillion, of the trace gas F-12 at the South Pole from 1976 to 1980.
(b) Discuss the quality of the fit of the model to the data.
(This problem is continued in Problem 2.7.7.)

Table 2.3.7 Concentration of F-12 at the South Pole by year, with 1976 as year 0 [11]

t	0	1	2	3	4
C	195	216	244	260	284

2.3.5.*

(a) Fit a linear model to the data of Table 2.3.8 (using a calculator to do the computations).
(b) Discuss the quality of the fit of the model to the data.

(This problem is continued in Problem 2.7.8.)

Table 2.3.8 A data set for Problem 2.3.5

t	8.7	9	11	18	19	22	28
C	25	25	26	48	65	90	100

2.3.6. In a convenient programming environment, write a program that inputs a file with two columns of data and uses the least squares procedure to fit the model $y = b + mx$ using the first column for the x values and the second column for the y values. Test the program with the data from Problems 2.3.4 or 2.3.5.

2.3.7. (Continued from Problems 2.1.1 and 2.1.3.)
Repeat Problem 2.3.1 with the model $y = b + mx$ using the program of Problem 2.3.6.

2.3.8. The National Oceanographic and Atmospheric Administration (NOAA) has a data set on its web site that gives the dates of the beginning of the grape harvests in Burgundy from 1370 to 2003 [5]. This data offers a crude, but long-term, look at global climate change.

(a) Fit a linear model to the data using the program of Problem 2.3.6. Do three different calculations: (1) the years 1800–1950, (2) the years 1951–1977, and (3) the years 1979–2003. Of particular interest is the slope.
(b) On the average, grape harvest dates have been getting earlier since 1800. By how many days did the expected grape harvest date change in each of the three periods?
(c) Plot all of the data from 1800 to 2003 as points on a graph. Explain why it is a mistake to connect the points to make a dot-to-dot graph.
(d) Add the three linear regression lines to the plot, being careful to use only the appropriate time interval for each.
(e) What do the results appear to say about global climate change?
(f) Offer at least one possible explanation for the results that does not involve global climate change. [Hint: Think about possible biological explanations.]

(This problem is continued in Problem 2.7.9.)

2.3.9. The National Snow and Ice Data Center has a data set on its web site that gives the duration of ice cover for a number of Northern Hemisphere lakes dating back to the 1800s in some cases. These data sets offer a look at global climate change that is shorter term than the grape harvest data in Problem 2.3.8 but which has fewer confounding factors.

(a) Go to the search engine for the Global Lake and River Ice Phenology Database: http://nsidc.org/data/lake_river_ice/freezethaw.html
Type a name code in the appropriate box (in capitals). Suitable lakes for this study include JGL03 (Iowa); DMR1, JJM22, JJM27, and JJM33 (Wisconsin); KMS10, KMS11, and KMS14 (New York); GW240, GW341, GW369, GW512, and GW592 (Sweden); JK02, JK03, JK09, JK17, JK31, JK48 (Finland); and NG1 (Russia). In the output parameter options list, choose Ice Off Date, Ice Duration, Latitude, Longitude, Lake Name, and Country Name. In the output sort options list, choose Ice Season. Click the Submit Request button.
(b) Copy the data and paste it into a text file. Save the data with the file extension csv (comma-delimited).
Of course you will want to look at the lake name, latitude, longitude, and country name for context, but these columns are not part of the data to be analyzed. Open the file in a spreadsheet program and delete all but the first and fourth columns, which give the year and number of days of ice cover. Also delete any rows in which the ice duration is given as -999, which means that the data is unavailable. Save the file again.
(c) Fit a linear model to the data using the program of Problem 2.3.6. Do three different calculations: (1) the full data set, (2) the years 1930–1970, and (3) the years 1970 to the present. Of particular interest is the slope.
(d) Plot all of the data as points on a graph. Explain why it is a mistake to connect the points to make a dot-to-dot graph.

(e) Add the three linear regression lines to the plot, being careful to use only the appropriate time interval for each.
(f) What do the results appear to say about global climate change?

(This problem is continued in Problem 2.7.10.)

2.3.10. Derive the results of the linear least squares method for the model $y = mx$,

$$m = \frac{\sum xy}{\sum x^2}, \qquad \text{RSS} = \sum y^2 - \frac{(\sum xy)^2}{\sum x^2} = \sum y^2 - m\sum xy,$$

by applying optimization methods from calculus to the total discrepancy function

$$F(m) = \left(\sum x^2\right) m^2 - 2\left(\sum xy\right) m + \left(\sum y^2\right) .$$

2.3.11. Derive the general linear least squares results (Theorem 2.3.2) by using the results for the $y = mx$ case.

2.4 Empirical Modeling II: Fitting Semilinear Models to Data

After studying this section, you should be able to:

- Use the least squares method to fit the linearized versions of the models $z = Ae^{\pm kt}$ and $y = Ax^p$ to data, where A, k, and p are parameters.
- Use a variant of the linear least squares method to fit models of the form $y = qf(x, p)$ to data, where q and p are parameters.

In Section 2.3, we learned the basic least squares method for linear models. The method can be adapted for some nonlinear models.

2.4.1 Fitting the Exponential Model by Linear Least Squares

Taking natural logarithms of both sides of the model

$$z = Ae^{-kt} \tag{2.4.1}$$

changes the model equation to

$$\ln z = \ln A - kt \, . \tag{2.4.2}$$

Now suppose we define a new set of variables and parameters by the equations

$$y = \ln z, \qquad x = t \, , \qquad m = -k \, , \qquad b = \ln A \, . \tag{2.4.3}$$

The result is the standard linear model

$$y = b + mx \, . \tag{2.4.4}$$

The algebraic equivalence of the original model (Equation (2.4.1)) and the linearized model (Equation (2.4.4)) allows us to fit the exponential model to data using linear least squares.[23]

[23] Equivalent models are the subject of Section 2.6.

Algorithm 2.4.1 *Linear least squares fit for the exponential model* $z = Ae^{-kt}$

1. *Convert the tz data to xy data using* $y = \ln z$ *and* $x = t$.
2. *Convert the xy data to XY data using* $X = x - \bar{x}$ *and* $Y = y - \bar{y}$, *where* $\bar{x}$ *and* $\bar{y}$ *are the means of the x and y values, respectively.*
3. *Find the parameters m and b for the xy model using the linear least squares formulas:*

$$m = \frac{\sum XY}{\sum X^2}\,, \qquad b = \bar{y} - m\bar{x}\,. \tag{2.4.5}$$

4. *Calculate the parameters for the exponential model:* $A = e^b$ *and* $k = -m$.

Example 2.4.1. Table 2.4.1 shows data from a radioactive decay simulation of 1,000 particles, each of which had an 8 % chance of decaying in any given time step. The original data consists of the time t and number of remaining particles z for each time. The xy and XY data sets were calculated in steps 1 and 2 of Algorithm 2.4.1, with $\bar{x} = 4.5$ and $\bar{y} = 6.542$. The linear least squares formulas (2.3.7) and (2.3.8) yield the results

$$m = -0.0835\,, \qquad b = 6.917\,,$$

from which we obtain

$$k = 0.0835\,, \qquad A = 1010\,.$$

Figure 2.4.1 shows the data and best-fit model in both the xy and tz planes. □

Table 2.4.1 Data sets for the exponential model in Example 2.4.1

t	0	1	2	3	4	5	6	7	8	9
z	1,000	929	855	785	731	664	616	568	515	471
x	0	1	2	3	4	5	6	7	8	9
y	6.908	6.834	6.751	6.666	6.594	6.498	6.423	6.342	6.244	6.155
X	−4.5	−3.5	−2.5	−1.5	−0.5	0.5	1.5	2.5	3.5	4.5
Y	0.366	0.292	0.209	0.124	0.053	−0.043	−0.118	−0.200	−0.297	−0.387

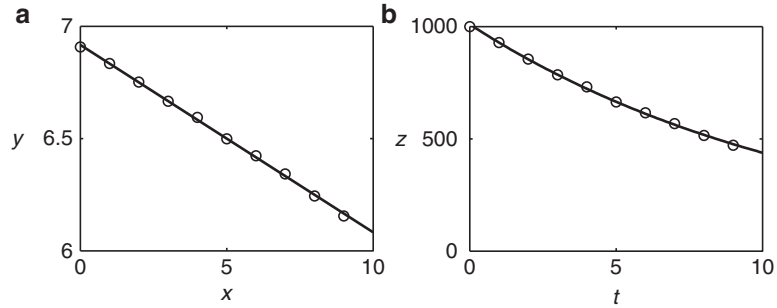

Fig. 2.4.1 The exponential model fit to the data of Table 2.4.1 using linear least squares (**a**) $y = b + mx$, (**b**) $z = Ae^{-kt}$

2.4.2 Linear Least Squares Fit for the Power Function Model $y = Ax^p$

The predation data for *P. speedius* (Table 2.1.1 and Figure 2.1.1) appears superficially to resemble a square root graph. This suggests a model of the form

$$y = Ax^p\,, \qquad A, p > 0\,. \tag{2.4.6}$$

This model can be fit using the same linearization technique used for the exponential model. However, we have to be careful about notation. Often a particular symbol has different meanings in two or more formulas needed to solve a particular problem. Here, the symbol x represents the number of prey animals in the biological setting (and hence in the model (2.4.6)), but it also represents the generic independent variable in the generic models $y = b + mx$ and $y = mx$. This kind of duplication is unavoidable because many formulas have their own standard notation. One way to avoid error in these cases is to rewrite the generic formulas using different symbols. In this case, let's use U, V, u, and v in place of X, Y, x, and y in the generic linear least squares formulation.

Taking a natural logarithm of (2.4.6) yields

$$\ln y = \ln A + p \ln x\,,$$

which is equivalent to the linear model $v = b + mu$ using the definitions

$$u = \ln x\,, \quad v = \ln y\,, \quad m = p\,, \quad b = \ln A\,.$$

We can then formulate an algorithm for fitting a power function model using linearized least squares.

Algorithm 2.4.2 *Linear least squares fit for the power function model $y = Ax^p$*

1. *Convert the xy data to uv data using $u = \ln x$ and $v = \ln y$.*
2. *Convert the uv data to UV data using $U = u - \bar{u}$ and $V = v - \bar{v}$, where $\bar{u}$ and $\bar{v}$ are the means of the u and v values, respectively.*
3. *Find the parameters m and b for the uv model using the linear least squares formulas:*

$$m = \frac{\sum UV}{\sum U^2}\,, \qquad b = \bar{v} - m\bar{u}\,. \tag{2.4.7}$$

4. *Calculate the parameters for the power function model: $A = e^b$ and $p = m$.*

Notice that all the symbols used in Algorithm 2.4.2 are defined within the algorithm statement. The meaning of a biological symbol is seldom clear from the context alone, so it is good modeling practice to define all symbols in the statement of a model or algorithm.

Example 2.4.2. In fitting the model $y = Ax^p$ to the *P. speedius* data from Table 2.1.1, we run into a problem. The change-of-variables formulas $u = \ln x$ and $v = \ln y$ do not work for the point $(0,0)$. This is not a serious problem, because $(0,0)$ satisfies $y = Ax^p$ exactly for any values of the parameters. Omitting $(0,0)$, we obtain the linearized least squares result

$$y = 2.71x^{0.499}\,. \tag{2.4.8}$$

This model is plotted together with the data in Figure 2.4.2. The first plot shows the data and model as a plot of y versus x, while the second plot shows the linearized data and model as a plot of $\ln y$ versus $\ln x$. □

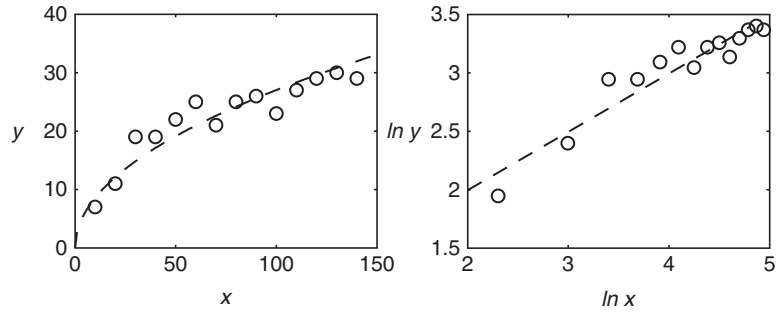

Fig. 2.4.2 The power function model fit to the *P. speedius* data [without $(0,0)$] using linear least squares

Check Your Understanding 2.4.1:
Construct a data table, similar to Table 2.4.1, to verify the results of Example 2.4.2. Note that the data point $(0,0)$ must be omitted.

2.4.3 Semilinear Least Squares

In Example 2.4.2, the parameter values $A = 2.71$ and $p = 0.499$ minimize the fitting error on the graph of $\ln y$ versus $\ln x$. This is not the same thing as minimizing the fitting error on the graph of y versus x. It is tempting to accept the results as optimal, but this is not necessarily appropriate. If we want to minimize the fitting error in the original data, we must find a way to do so without linearizing the model. For this purpose, we need to adapt the linear least squares method to apply to what we can call *semilinear* models: models of the form $y = qf(x,p)$, where f is a nonlinear function of the independent variable x and one of the parameters, while the other parameter appears as a multiplicative factor.[24]

2.4.3.1 Finding the Best *A* for Known *p*

The semilinear regression method for the model $y = Ax^p$ involves two distinct mathematics problems: first we find the best A in terms of an arbitrary choice of p, and then we find the best p. The first of these problems can be solved for any particular p by using linear least squares on a data set that has been modified to account for the chosen value of p.

Example 2.4.3. Suppose we *assume* $p = 0.5$. Then, for each data point, we can calculate the quantity x^p exactly. Defining $z = x^{0.5}$, we can rewrite the model $y = Ax^{0.5}$ as $y = Az$. This allows us to convert the original xy data into the zy data of Table 2.4.2. The mathematical result for models of the form $y = mx$ (with z playing the role of x and A playing the role of m) then yields the slope and residual sum of squares:

$$A = \frac{\sum zy}{\sum z^2} \approx 2.67, \quad \text{RSS} = \sum y^2 - A\sum zy \approx 80.1.$$

[24] This could be done with the exponential model $z = Ae^{-kt}$ as well, if the goal is to minimize the fitting error in the original data. However, for reasons beyond the scope of this discussion, it is usually better to fit exponential models in the linearized form $\ln z = \ln A - kt$ rather than the original form.

In comparison, the residual sum of squares for $y = 2.71x^{0.499}$ is approximately 80.7. Thus, the model $y = 2.67x^{0.5}$ is a little more accurate on a graph in the xy plane than the best fit obtained by linearization. □

Table 2.4.2 Data points for the model $y = Ax^{0.5}$ of Example 2.4.3

$z = x^{0.5}$	0	3.16	4.47	5.48	6.32	7.07	7.75	8.37
y	0	7	11	19	19	22	25	21
$z = x^{0.5}$	8.94	9.49	10.0	10.49	10.95	11.40	11.83	
y	25	26	23	27	29	30	29	

The very slight improvement we found in Example 2.4.3 is not enough to justify the more complicated procedure for finding the parameter values. However, we only *guessed* the value $p = 0.5$; what we really need is a way to find the *best* p.

2.4.3.2 Finding the Best *p*

When fitting the model $y = Ax^p$ to data, we define a residual sum of squares in terms of the parameters A and p. The goal is to choose the pair (p,A) that minimizes the residual sum of squares. Optimization problems for two-parameter nonlinear models are usually very difficult, but in this case we already know how to find the best value of A for any given choice of p. If we assume that we will always use the best A, then we can think of the residual sum of squares as a function of p only.

Formally, define a residual sum of squares function F by

$$F(p) = \min_A (\mathrm{RSS}(p,A)) = \mathrm{RSS}(p,\hat{A}(p)), \tag{2.4.9}$$

where $\hat{A}(p)$ is found as in Example 2.4.3. From the calculation in the example, we have $\hat{A}(0.5) = 2.67$ and $\mathrm{RSS}(0.5, 2.67) = 80.1$; therefore, $F(0.5) = 80.1$. For any given value of p, we have to create a modified data set and use the linear least squares formulas to get the corresponding F for that p. That is a lot of work, but it is a type of work for which computers are ideally suited. The simplest way to use a computer to identify the best parameters is to calculate a lot of points on the graph of F and identify (approximately) the smallest F from that set of points.

Check Your Understanding 2.4.2:
Repeat the calculation of Example 2.4.3 to obtain the result $F(0.4) = 63.1$. This result shows that $p = 0.4$ is closer to the optimal value than our original guess of $p = 0.5$.

Example 2.4.4. Table 2.4.3 shows some values of $F(p)$ for the model $y = Ax^p$ using the *P. speedius* data from Table 2.1.1. From this table, we can see that the optimal p value is somewhere in the interval $0.35 \le p \le 0.45$. We then compute the values of $F(p)$ for $p = 0.350, 0.351, 0.352, \ldots 0.450$. These are plotted in Figure 2.4.3. By examining the list of F values, we see that the optimal p value, to three significant figures, is 0.409, with a corresponding residual sum of squares of 62.9. We then obtain A by linear least squares, as in Example 2.4.3, leading to the result

$$y = 4.01x^{0.409}, \qquad \mathrm{RSS} = 62.9\,. \tag{2.4.10}$$

This new model is plotted along with that of Example 2.4.2 in Figure 2.4.4. □

A visual examination of Figure 2.4.4 shows the clear superiority of the semilinear least squares method compared to the linear least squares method for power function models. The

Table 2.4.3 Some values of the residual sum of squares function $F(p)$ for the model $y = Ax^p$ using the *P. speedius* data from Table 2.1.1

p	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70
$F(p)$	92.3	71.2	63.1	66.6	80.1	102.4	132.3	168.7	210.7

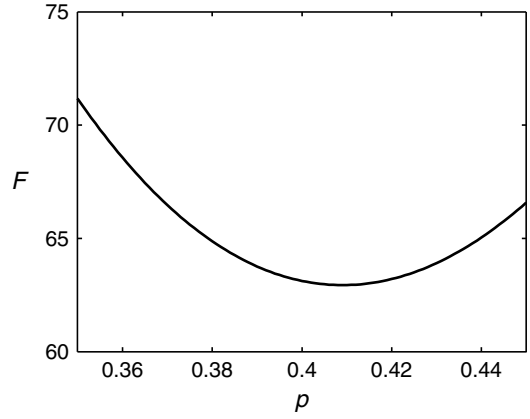

Fig. 2.4.3 The minimum residual sum of squares (F in (2.4.9)) for the *P. speedius* predation data with the power function model (Example 2.4.4)

semilinear method is also superior to the Lineweaver–Burk linearization that is commonly used to determine parameters for Michaelis–Menten reactions in biochemistry.[25]

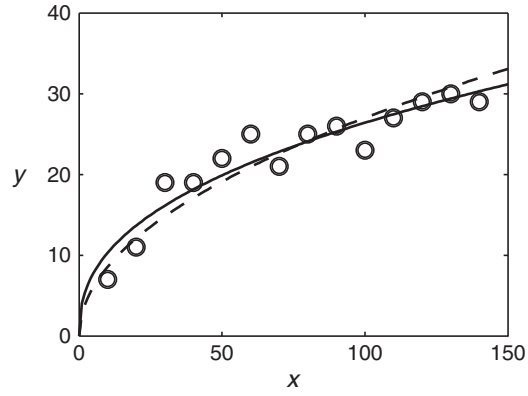

Fig. 2.4.4 The power function model fit to the *P. speedius* data (without $(0,0)$) using linear least squares (*dashed*) and semilinear least squares (*solid*)

[25] See Problem 2.4.8 for an illustration of how important this is. Other authors (see [10], for example) also state that one should use nonlinear regression rather than using linear regression on a linearized model, but they don't always explain the reason carefully or present an illustrative example.

The procedure described here works for any model of the general form $y = qf(x,p)$, where x is the independent variable, y is the dependent variable, and q and p are parameters.[26] We summarize the result here.

Theorem 2.4.1 (Semilinear Least Squares). *Given a model $y = qf(x,p)$ with data points (x_1,y_1), (x_2,y_2), …, (x_n,y_n), define a function F by*

$$F(p) = \min_q (RSS(p,q)) = \sum y^2 - \frac{[\sum yf(x,p)]^2}{\sum f^2(x,p)}.$$

Let $\hat{p}$ be the value of p that yields the minimum value of F. Then the minimum residual sum of squares on the graph in the xy plane is achieved with parameter values

$$p = \hat{p}\,, \qquad q = \frac{\sum yf(x,\hat{p})}{\sum f^2(x,\hat{p})}\,.$$

Check Your Understanding Answers

Table 2.4.4 Total discrepancy calculations for $y = 0.25x$ with the *P. speedius* data

x	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140
$0.25x$	0	2.5	5	7.5	10	12.5	15	17.5	20	22.5	25	27.5	30	32.5	35
y	0	2	7	10	9	14	21	20	25	20	30	25	29	35	38
Δy	0	0.5	2	2.5	1	1.5	6	2.5	5	2.5	5	2.5	1	2.5	3
$(\Delta y)^2$	0	0.25	4	6.25	1	2.25	36	6.25	25	6.25	25	6.25	1	6.25	9

1. The total discrepancy is $F(0.25) = 134.75$, which is less than $F(0.3)$ (Table 2.4.4).

Problems

2.4.1.* The data of Table 2.4.5 gives the population of a bacteria colony as a function of time. Find the exponential function that best fits the data set using the linearization method. Plot the linearized data with the best-fit line. Plot the original data with the exponential curve corresponding to the best-fit line.

Table 2.4.5 Population of bacteria after t hours

t	0	1	2	3	4
N	6.0	9.0	13.0	21.0	29.0

2.4.2.* Use semilinear least squares with the data in Problem 2.4.1 to fit the exponential model without linearization. Compare the parameter results from the two calculations. Plot the two resulting models as y versus t and again as $\ln y$ versus t. Discuss the results.

[26] In Example 2.7.3, we will use this method with a model derived in Section 2.5.

2.4.3. (Continued from Problem 2.3.4.)

(a) Use the semilinear method to fit the model $y = Ax^p$ to the data from Problem 2.3.4 and determine the residual sum of squares.
(b) What is clearly wrong with using the model $y = Ax^p$ for this data set?

2.4.4. (Continued from Problem 2.3.5.)

(a) Use the linearization method to fit the model $y = Ae^{-kt}$ to the data from Problem 2.3.5.
(b) Use the semilinear method to fit the model and data from part (a).
(c) Find the residual sums of squares for the results of parts (a) and (b) on graphs of y versus t and on graphs of $\ln y$ versus t.
(d) Which is better, the linearization result or the semilinear result?

2.4.5:* Use semilinear least squares with the data in Table 2.4.1 to fit the exponential model without linearization. Compare the parameter results with Example 2.4.1, which used the same data. Why are the results different? Plot the two resulting models as y versus t and again as $\ln y$ versus t. Draw a reasonable conclusion from your observations.

2.4.6. (Continued from Problem 2.1.2.)

(a) Fit the model $y = Ax^p$ to your *P. speedius* data from Problem 2.1.2 using linearized least squares.
(b) Fit the model $y = Ax^p$ to the same data using the semilinear method.
(c) Plot the model from part (a) along with the data on a graph of $\ln y$ versus $\ln x$. Repeat for the model from part (b). Compare the visual appearances of the two plots.
(d) Plot the model from part (a) along with the data on a graph of y versus x. Repeat for the model from part (b). Compare the visual appearances of the two plots.
(e) Describe and explain any conclusions you can draw from this set of graphs.

(This problem is continued in Problem 2.7.5.)

2.4.7. Table 2.4.6 shows data for average lengths in centimeters of Atlantic croakers (a species of fish) caught off the coasts of three states. Use this data to fit the von Bertalanffy growth equation,

$$x(t) = x_\infty(1 - e^{-rt}),$$

where $x(t)$ is the length of the fish, x_∞ is the asymptotic maximum length, and r is a positive parameter.[27] How well does the model fit the data?

Table 2.4.6 Average length in centimeters of Atlantic croakers from New Jersey, Virginia, and North Carolina [4]

Age	1	2	3	4	5	6	7	8	9	10
NJ	30.3	31.1	32.4	34.2	35.0	34.8	37.4	36.6	36.1	37.4
VA	25.8	28.9	31.8	34.0	35.2	36.1	37.4	40.2	40.2	40.3
NC	24.6	27.3	29.7	33.1	35.2	37.2	37.8	38.4	37.7	38.1

2.4.8. Michaelis–Menten reactions are enzyme-catalyzed reactions in biochemistry. The initial reaction rate v depends on the concentration S of the principal reactant (called the *substrate*), according to the Briggs–Haldane model

[27] It would be better to fit the data for individual lengths rather than averages; however, the raw data sets are quite large and not generally available.

$$v = \frac{v_m S}{K_M + S}, \qquad S, v_m, K_M > 0,$$

where v_m is the maximum rate (corresponding to a very large substrate concentration) and K_M is the semisaturation parameter (see Section 2.6). Lineweaver and Burk rewrite the reaction rate equation as

$$\frac{1}{v} = \frac{1}{v_m} + \frac{K_M}{v_m}\frac{1}{S}$$

and then fit the data to this form [9].[28]

(a) Derive the Lineweaver–Burk formula from the original model.
(b) Use the Lineweaver–Burk linearization to determine the parameter values using the data in Table 2.4.7; that is, use the data to create a data set for variables $y = 1/v$ and $x = 1/S$. After fitting the model $y = b + mx$ to this data using linear least squares, determine the corresponding parameters v_m and K_M from the original model.
(c) Find the parameter values for the model using the data in Table 2.4.7 with the semilinear least squares method.
(d) Prepare a plot that includes:

   a. The data points (S, v);
   b. The Briggs–Haldane curve, using the parameter values obtained in part (b); and
   c. The Briggs–Haldane curve, using the parameter values obtained in part (c).

(e) Calculate the residual sum of squares for each parameter set. Discuss the results. In particular, which of the curves seems to fit the data better, and why is this the case?

Table 2.4.7 Substrate concentration S and reaction velocity v for nicotinamide mononucleotide adenylyltransferase in pig livers [3]

S	0.138	0.220	0.291	0.560	0.766	1.46
v	0.148	0.171	0.234	0.324	0.390	0.493

2.4.9. (Continued from Problem 1.1.13.)
In Problem 1.1.13, we considered the Sinclair coefficient used to compare weightlifters of different sizes:

$$C(m) = 10^{A[\log_{10}(m/b)]^2},$$

[28] A 1975 paper presents compelling evidence that other methods in use in the mid-1970s were preferable to the Lineweaver–Burk method [2]. Unfortunately, Lineweaver–Burk was entrenched by then, and scientific progress has failed in this case to overcome the inertia of standard practice. The Lineweaver–Burk method remains in common usage today. The 1975 tests included an implementation due to Wilkinson of the nonlinear method [14], which of course produced the best results with simulated data under reasonable assumptions about the types of error in the data. There can be no question that the semilinear least squares method produces the best fit on a plot of the original data, nor is there any reason in the world of fast computing to settle for a method that is not as good simply because it is faster for hand computation. In general, a good understanding of the theories of various methods for solving problems helps us to identify cases, such as this one, where older methods should be replaced by newer computer-intensive methods.

where m is the mass, b is the mass of the heavyweight world record holder, and A is a positive parameter of unclear meaning. The formula is based on the assumption that the two-lift total should fit the model

$$T(m) = C\,10^{-A[\log_{10}(m/b)]^2}.$$

(a) Although this model could be fit as a three-parameter model, which would be outside the scope of our semilinear method, the parameter b in practice is fixed at $b = 173.961$ as of this writing. The value of the parameter A should be chosen to make the model fit the data optimally. To test this point, use the semilinear method to fit the model for $T(m)$ to the data in Table 2.4.8, which gives the official post-1998 world records for each weight class as of June 2012. Report A to the nearest four decimal digits and compare with the official value $A = 0.7848$.
(b) Redo Problem 1.1.13 using the new value of A. Does this change the top three spots in the overall ranking?

Table 2.4.8 Post-1998 world records for men's weightlifting as of June 2012, with T the total amount lifted in two events by lifters of mass m kg

m	56	62	69	77	85	94	105	174
T	305	326	357	378	394	412	436	472

2.5 Mechanistic Modeling I: Creating Models from Biological Principles

After studying this section, you should be able to:

- Discuss the relationship between biological observations, a conceptual model, and a mechanistic mathematical model.
- Discuss the conceptual model for the Holling type II predation function.
- Explain the mathematical derivation of the Holling type II predation function.

In Section 2.4, we modeled radioactive decay with an exponential model. We were able to fit the data quite well, but the empirical justification for the model limits its explanatory value. Would an exponential model be a good fit with a different data set for the same substance? What about a data set that extends the total time of the experiment or a data set for a different radioactive substance? Empirical modeling cannot answer these questions, because empirical reasoning must *begin* with the data.

The alternative approach to modeling is *mechanistic modeling*, in which we obtain a model from assumptions based on theoretical principles. Sometimes a mechanistic justification can be found for a model we have already identified empirically, as we will see with our exponential model for radioactive decay. In this case, the model gains explanatory value. In other cases, we may be able to discover a model not previously identified empirically.

2.5.1 Constructing Mechanistic Models

Textbooks in algebra and calculus include story problems, where you have to derive mathematical models from a verbal statement. These verbal statements are conceptual models, as discussed in Section 2.2. They can be translated into mathematical statements simply by following some basic rules. This process is routine, in the sense that mastery of the rules is all that is needed for it. The hard part of model construction is getting the conceptual model in the first place—this work is done by the author of the story problem rather than the student.

How do we write our own conceptual models? In practice, we typically use whatever we have learned from prior experience in modeling. More fundamentally, conceptual models come from qualitative observation and measurement of data. The observations and data have to be generalized, and the conceptual model is obtained by restating these generalizations as fact. As an example, we consider the development of a model for radioactive decay. Imagine that radioactivity has only just recently been discovered. We have observations and data from experimental scientists, and we must use this information to construct a conceptual model.

Example 2.5.1. From examining data on radium decay rates, we obtain a specific observation.

Specific Observation:

All samples of radium-226 lose approximately 0.043 % of their number in 1 year.

By itself, this observation is interesting enough to suggest more studies with other time intervals and other radioactive isotopes. All such studies yield similar observations, except that the percentage loss and time unit are different for each. Taken together, these experiments allow us to write a much stronger statement.

General Observation[29]:

The percentage of particles lost by a given radioactive substance in one time unit is approximately constant.

Idealizing the general observation yields a mechanistic assumption. In this simple example, only one mechanistic assumption is needed to complete the conceptual model.

Conceptual Model 1:

The fraction of particles lost by a given radioactive substance in one time unit is fixed.

This conceptual model differs only slightly from the corresponding general observation: the latter uses the qualifier "approximately," while the former reads as a statement of fact. Models are not necessarily correct, but they are always specific.

Now that we have a conceptual model, we need only rewrite it in mathematical notation.

Mathematical Model 1:

If $y(t)$ is the amount of radioactive material at time t, then

$$\frac{y(t)-y(t+1)}{y(t)} = K\,, \tag{2.5.1}$$

for some $K > 0$.

Equation (2.5.1) is a direct translation of the conceptual model. The quantity $y(t)-y(t+1)$ is the number of particles lost between time t and time $t+1$. Dividing this quantity by $y(t)$ gives the fraction of particles lost between time t and time $t+1$. It is easier to work with fractions than percentages, so we simply assign the parameter K to be the constant fraction lost in one time unit. In a specific example, we need to know the value of K; otherwise we keep it unspecified so that we can apply the mathematical model to any radioactive substance.

Equation (2.5.1) is a discrete mathematical model for radioactive decay. We can rearrange the model so that it makes a prediction about the number of particles at time $t+1$:

$$y(t+1) = (1-K)y(t)\,.$$

Given y_0 particles at time 0, the model predicts

[29] The simulated experiment in Example 2.4.1 was based on this general observation, with 8 % used as the percentage for the hypothetical substance.

$$y(0) = y_0\,, \quad y(1) = y_0(1-K), \quad y(2) = y_1(1-K) = y_0(1-K)^2\,, \quad \cdots\,.$$

There is a clear pattern—the number of particles at time t is[30]

$$y(t) = y_0(1-K)^t\,. \tag{2.5.2}$$

□

Note that the mathematical model in Example 2.5.1 did not come *directly* from the data of one or more experiments, but *indirectly* from a set of assumptions suggested by the results of many experiments. These assumptions are the conceptual model.

The result (2.5.2) of the discrete model in Example 2.5.1 is not the usual exponential function $y = y_0e^{-kt}$; however, we can obtain the exponential function model by means of algebraic manipulation.

Example 2.5.2. Beginning with the model $y = y_0(1-K)^t$, we can replace the factor $1-K$ with the factor e^{-k}. There is no harm in doing this; K is an arbitrary parameter, and it is alright to use a different arbitrary parameter if we prefer the new one. The advantage of k over K is that we can then write the model as $y = y_0(e^{-k})^t = y_0e^{-kt}$. If we already know the value of K for a specific substance, we can use the equation $1-K = e^{-k}$ to calculate the corresponding value of k. Solving for k gives us the formula

$$k = -\ln(1-K) \tag{2.5.3}$$

which gives the exponential rate parameter in terms of the fraction of particles lost in one time unit. For the simulation used for Example 2.4.3, we have $K = 0.08$, so $k = 0.0834$. The empirical result $k = 0.0835$ obtained in the example agrees very nicely with this theoretical result. □

The model $y = y_0e^{-kt}$ of Example 2.5.2 can also be obtained using calculus.

Example 2.5.3. To derive the model $y = y_0e^{-kt}$ directly, we need a more sophisticated conceptual model. Suppose we could measure the instantaneous rate of radioactive decay (particles per time) rather than the average rate of decay (particles in one unit of time). Because equal fractions of particles decay in equal units of time, we expect a sample of 100 atoms to have twice as many decays in a given amount of time as a sample of 50 atoms. Thus, the rate of decay for a sample should be proportional to the sample size.

Conceptual Model 2:

In any radioactive decay process, the rate of decay is proportional to the amount of the substance.

Translating this into mathematical notation gives us the corresponding mathematical model.

Mathematical Model 2:

If $y(t)$ is the amount of radioactive material at time t, then

$$\frac{dy}{dt} = -ky \tag{2.5.4}$$

for some $k > 0$.

Note that we can identify the family of functions $y = y_0e^{-kt}$ as the functions that satisfy (2.5.3).

[30] Formally, this result can be proved by *mathematical induction*.

To understand the meaning of the parameter k, we can rewrite the model as

$$-\frac{1}{y}\frac{dy}{dt} = k\,. \tag{2.5.5}$$

Thus, k is the decay rate $-dy/dt$ divided by the amount y, which is called the *relative rate of decay*. An alternative statement of Conceptual Model 2:

The relative decay rate of a radioactive substance is constant. □

The derivation of Example 2.5.3 is preferable to that of Examples 2.5.1 and 2.5.2 on general principles:

Conceptual models for continuous processes should be based on rates of change rather than discrete change.

2.5.2 Dimensional Analysis

Dimensional analysis is a powerful tool that simplifies the construction of mathematical models. The idea is that all models must be dimensionally consistent; colloquially, "You can't add apples to oranges." There are three requirements for dimensional consistency.[31]

Rules for Dimensional Consistency

1. Quantities can be added together or set equal to each other only if they have the same dimension.
2. The dimension of a product is the product of the dimensions.
3. The argument of a transcendental function, such as a trigonometric, exponential, or logarithmic function, must be dimensionless.

Example 2.5.4. The circumference of a circle of radius r has length $2\pi r$. The ratio of circumference to radius is 2π, which is, by definition, the radian measure of a full circle. The radian measure of any other angle is similarly defined as the ratio of the length of the corresponding circular arc to the radius of the circle. Because both circumference and radius are lengths, the radian measure of an angle is dimensionless. The rule that arguments of trigonometric functions must be dimensionless suggests that these arguments should be given in radians. □

Example 2.5.5. In functions such as $\sin kt$, the argument does not necessarily correspond to a geometric angle, so the term "radians" does not really apply. Nevertheless, dimensional consistency requires the quantity k to have dimension 1/time. □

Example 2.5.6. The mathematical model $\frac{dy}{dt} = -ky$ must be dimensionally consistent. The independent variable t is a time. The dependent variable y can be considered to be either a mass or a count of atoms; to be specific, let's take it to be a mass. Hence, the derivative dy/dt has dimension mass per time. The right side of the model must also be mass per time, and so the dimension of k must be 1/time.

[31] Note the distinction between dimensions, such as length, and the associated units of measurement, such as meters, feet, and light-years.

The related mathematical model, $y(t) = y_0\,(1-K)^t$, also has to be dimensionally consistent, so the parameter K and the variable t must both be dimensionless. In general, time is dimensionless in any discrete model. □

2.5.3 A Mechanistic Model for Resource Consumption

We are now ready to develop mathematical models for resource consumption, which we hope will explain the differences between the predation data for the two species of predator introduced in Section 2.1.

Consider a situation in which an organism is placed in an environment with a fixed concentration of resource x. Since the organism consumes the resource, it is therefore necessary for the resource to be replenished. We measure the rate y with which we have to replenish the resource; this is equivalent to measuring the rate at which the resource is being consumed. Our goal is to construct a mathematical model that relates the intake rate y to the resource concentration x. Note that it doesn't make any difference how the organism obtains the resource. The organism could be a predator that feeds by hunting, an herbivore that feeds by grazing, or a single cell that feeds by absorbing nutrients through its surface.

To develop a conceptual model for this experiment, it helps to create a narrative version that appeals to human experience. Imagine yourself as the organism in the experiment. You live alone in a building with numerous hiding places that could contain servings of food. A total of x servings of food are distributed randomly throughout the building. Each time you eat a serving, a restaurant worker hides another serving in some randomly chosen location within the building, thereby keeping the food concentration at a constant level.[32] In this setting, our question is "How much do you eat per unit time in an environment with a constant food supply x per unit area?"[33] We denote the amount eaten per unit time by y. Thus, we seek a model to predict y as a function of x.

We begin with the simplest conceptual model.

Example 2.5.7. You have to get to the food in order to eat it. For the simplest conceptual model, we imagine that you continually search the building, eating the food as you find it. It seems reasonable that you will search some fixed amount of space per unit of time. The amount of space you can search in one unit of time could vary a lot, depending on the size of the rooms and the possible number of hiding places. However, it does *not* depend on the amount of food available. Every experiment in a given building with a given person can use the same search rate, regardless of x. Thus, the search rate is a parameter, which we designate as s. The food density x could be measured in servings per square meter, and the intake rate y could be measured in servings consumed per hour. The rate at which you locate food should be the product of the food density and the search rate; thus, we have the model

$$y = sx\,. \tag{2.5.6}$$

Note that the model also makes sense if we replace the quantities by their dimensions:

$$\frac{\text{food}}{\text{time}} = \frac{\text{area}}{\text{time}} \times \frac{\text{food}}{\text{area}}.$$

□

[32] Of course, this does not happen in a real feeding scenario; however, this is an assumption in the conceptual model. In practice, this discrepancy between the real experiment and the conceptual model causes difficulties in the measurement of the parameters.

[33] Note that x need not be very large. In our human version, we could have a restaurant the size of a shopping mall with an average of one serving of food in each of the stores.

2.5.4 A More Sophisticated Model for Food Consumption

We now have a mechanistic formulation for the linear model that we used in Section 2.3 for *P. steadius*. The model is obviously not useful for *P. speedius* because the plot of the data (Figure 2.1.1) is not approximately linear. We must conclude that the conceptual model of Example 2.5.7 is missing some biologically significant features. The crucial step in finding a better model for *P. speedius* is the search for a more realistic conceptual model.

Example 2.5.8. The model (2.5.6) is based on the assumption that you spend all of your time searching for food. This might be the case if food is extremely scarce. However, if you live in a buffet restaurant, where food is plentiful and easy to find, you spend only a tiny fraction of your time searching for it. You don't keep eating just because you can find more food! Instead, you spend nearly all of your time digesting and hardly any of it searching for more. The time required for digestion is a feature of the actual experiment that our first conceptual model lacks. Run another BUGBOX-predator trial with *P. speedius* and a large prey density. Notice that the predator spends only a small portion of the experiment time searching, because it pauses whenever it locates prey.

We need a conceptual model in which time is partitioned into two types: search time and "handling" time. With this distinction between two types of time, our first model is no longer dimensionally consistent; the "time" in y is total time, while the "time" in s is search time. Our dimensional equation needs an additional factor to account for the distinction:

$$\frac{\text{resource}}{\text{total time}} = \frac{\text{search time}}{\text{total time}} \times \frac{\text{area}}{\text{search time}} \times \frac{\text{resource}}{\text{area}}.$$

Using f to denote the fraction of time spent searching, we have

$$y = fsx\,. \tag{2.5.7}$$

If f is a parameter, like s, then we are done. However, this is not the case. We've already observed that f is approximately 1 when the resource is scarce, but approximately 0 when it is plentiful. Thus, f is a second dependent variable. We therefore need a second equation. From our conceptual model,

$$\text{search time} + \text{handling time} = \text{total time}.$$

Dividing by total time, we have

$$\frac{\text{search time}}{\text{total time}} + \frac{\text{handling time}}{\text{total time}} = 1\,.$$

It seems reasonable that the amount of handling time should be proportional to the amount of resource consumed, so we define a new parameter h to be the time required to handle one unit of the resource. Now we can think of the dimensional equation as

$$\frac{\text{search time}}{\text{total time}} + \left(\frac{\text{handling time}}{\text{resource}} \times \frac{\text{resource}}{\text{total time}}\right) = 1.$$

In symbols, this is

$$f + hy = 1\,. \tag{2.5.8}$$

Equations (2.5.7) and (2.5.8) are a pair of equations for a pair of dependent variables. Thus, we have the right number of equations for a complete model. Substituting from Equation (2.5.8) into (2.5.7), we have

$$y = (1 - hy)sx = sx - shxy\,,$$

or

$$y + shxy = sx\,.$$

Thus, we arrive at the model

$$y = \frac{sx}{1 + shx}\,. \tag{2.5.9}$$

One instance of this model is shown in Figure 2.5.1. The model shows the right general shape to fit the *P. speedius* data. □

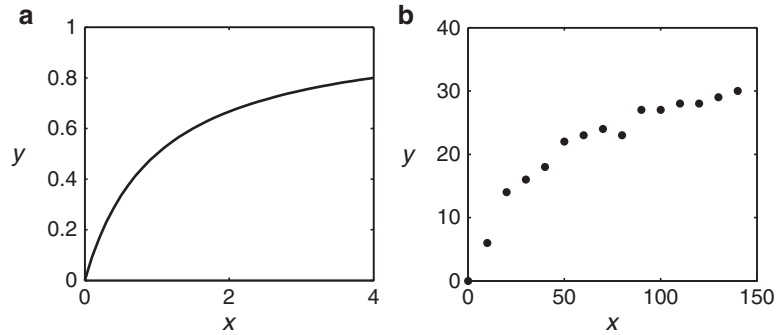

Fig. 2.5.1 Side-by-side comparison of (**a**) the Holling type II model $y = x/(1+x)$ and (**b**) the predation data for *P. speedius*, showing the same general shape

Equation (2.5.9) is one form of the model known in ecology as the *Holling type II* predation function and in microbiology as the *Monod growth function*. In its current form, the model is not semilinear, so we cannot use the method of Section 2.4 to fit it to data. In Section 2.6, we obtain the more common semilinear form of the model and fit this form to the *P. speedius* data. The model is also mathematically equivalent to the Briggs–Haldane model used to determine reaction rates for Michaelis–Menten reactions.

2.5.5 A Compartment Model for Pollution in a Lake

Suppose we want to model the concentration of pollutants in a lake. There is a general modeling framework, called *compartment analysis*, that is used when the primary structure of the model consists of accounting for changes in one or more quantities. For simplicity, we consider a single body of water and a single pollutant. We can think of the amount of pollutant in two ways, as the total mass Q and as the concentration C, which is the mass per unit volume. It is easier to think of mass while doing the accounting, but it is more meaningful to think of concentration if we want to understand the biological significance of the pollution.

The idea of compartment analysis is that the net rate of increase (or decrease) of a quantity is the sum of the various processes that work to increase or decrease that quantity. In other words, we have

$$\begin{matrix}\text{net rate of}\\ \text{increase}\end{matrix} = \begin{matrix}\text{input}\\ \text{rate}\end{matrix} - \begin{matrix}\text{output}\\ \text{rate}\end{matrix} + \begin{matrix}\text{production}\\ \text{rate}\end{matrix} - \begin{matrix}\text{decay}\\ \text{rate}\end{matrix}\,, \tag{2.5.10}$$

where each of the rates is expressed as mass of pollutant per unit time. Lakes are continually fed by some streams and drained by others, so we can expect to have inflow and outflow of water. If the inflow is polluted, then the rate equation will include an input rate of pollutant, which we will have to express in terms of basic quantities. The outflow will, of course, carry some of the pollutant along with the water. There is probably no mechanism for production of pollutants within the lake, but there may be decay because of chemical instability or a chemical reaction with something else in the water. The term "decay" should be broadly interpreted as any process that removes pollutant from the lake other than flow with the water, which could include processes whereby the pollutant is absorbed onto rocks or sinks into the sediment. The exact mechanism is unimportant as long as we can make a simple assumption about the rates of the processes collected together as pollutant decay.

Real lakes are complicated, with multiple input streams that vary in flow rates as the seasons change and the weather varies between dry and rainy. The concentration of pollutant is higher in areas near polluted inflow streams and lower in areas near clean inflow streams, so C depends on location as well as time. There may be multiple decay processes, some involving complicated chemical reactions. In some circumstances, we may want to account for all of these details, but our model will be of more general value if we keep it simple. We therefore adopt a simple conceptual model. Suppose we have a tank in a laboratory playing the role of the lake, and we imagine the simplest experiment we can do that might be similar to what happens in a real lake. Our conceptual model includes the following assumptions:

1. The water in the tank is stirred so that there are no spatial variations in concentration.
2. Any decay processes are proportional to the amount present, as in radioactive decay. This assumption requires a proportionality constant, which we will call R.
3. The input and output flow rates are constant in time and equal so that the volume in the tank is constant. We take V to be the volume and F to be the flow rate (volume/time).
4. The input stream has a fixed pollutant concentration C_{in} (mass/volume).

The net rate of increase is simply dQ/dT, where we are using T for time to facilitate some additional analysis that will appear in Section 2.6. Since Q is a mass, we know by the requirement of dimensional consistency that each of the other terms in the equation must have the dimension mass/time. This observation helps us quantify the rates of the three nonzero processes on the right side of the equation. The decay rate, by assumption 2, is RQ, with R having the dimension of 1/time. The input rate of pollutant should be larger for larger water flow rates as well as for larger concentrations. The product FC_{in} has a dimension of (volume/time)*(mass/volume) = (mass/time), so this quantity has the right dimension to be the input rate. Similarly, the output rate is $FC(T)$, where the assumption of adequate mixing justifies equating the concentration of the pollutant in the output stream with that in the lake. Combining all of these formulas, we have the equation

$$\frac{dQ}{dT} = FC_{\text{in}} - FC(T) - RQ(T) .$$

This equation is not quite what we need, because it contains both of the related quantities Q and C. Given that C is mass of pollutant per volume, these quantities are related by the equation $C(T) = Q(T)/V$. We can use this to eliminate either of the variables; given that the concentration is of greater biological importance, we replace Q by VC and divide by V to get

$$\frac{dC}{dT} = \frac{F}{V}C_{\text{in}} - \frac{F}{V}C(T) - RC(T) . \tag{2.5.11}$$

We will do some preliminary analysis with this model in Problem 2.6.8.

2.5.6 *"Let the Buyer Beware"*

As noted in Section 2.2, all mathematical models come with a disclaimer. Results obtained from their analysis are only guaranteed to be true for the corresponding conceptual models. Whether they are also true for the real-world setting depends on the quality of the approximation process, which is non-mathematical. Each of the models developed in this section requires validation to justify application to a real problem. The radioactive decay model has been validated innumerable times and is accurate for any macroscopic quantity to within any degree of accuracy that has been achieved by any experiment. Likewise, Figure 2.5.1 serves as a qualitative validation of the Holling type II predation model for the BUGBOX data, though similar data is needed to validate the model for a real biological setting. The pollutant model (2.5.11) also requires experimental validation. It would be easy to validate it for a laboratory experiment, but that would not guarantee its validity for use as a model of pollutants in a real lake. Most likely, historical data collected to monitor improvement in polluted lakes would serve for validation, but this would depend on the extent to which the assumptions in the conceptual model are actually true for that lake. A lake that is poorly mixed and has highly variable pollutant inflow and water flow rates will behave less like the model than one that has steady currents and nearly constant pollutant inflow and water flow rates.

Problems

2.5.1. Two possible models for the dynamics of a renewable resource (biotic or abiotic) are

$$\frac{dx}{dt} = 0.1 - \frac{xy}{1+x} \quad \text{and} \quad \frac{dx}{dt} = 0.1x - \frac{xy}{1+x}\,,$$

where $x(t)$ is the amount of resource present at time t and y is the number of consumers.

(a) For each of these models, describe a mechanism that accounts for the growth of the resource in a way that is consistent with the model.
(b) Explain the assumption the models make about the consumers.

2.5.2. The populations $x(t)$ and $y(t)$ of two interacting species are modeled using the equations

$$\frac{dx}{dt} = ax + bxy\,, \qquad \frac{dy}{dt} = cx + dxy\,,$$

where a, b, c, and d are parameters, not necessarily positive.

(a) Suppose the species are herbivores that compete for a common plant resource. Which of the four parameters should be positive and which negative? Explain.
(b) Repeat (a) for the case where x is an herbivore and y is a predator that eats the herbivore.

Problems 2.5.3–2.5.8 are based on models for a chemostat and an SIR disease.

- A *chemostat* is a device that grows bacteria in a steady environment by continuously adding fresh nutrient solution at rate Q and removing the whole mixture at the same rate. One possible model for a chemostat is given by the equation

$$\frac{dN}{dT} = RN\left(1 - \frac{N}{K}\right) - QN\,, \qquad R,K,Q > 0\,,$$

where N is the population of bacteria at time T, R is the maximum bacterial growth rate, K is the maximum population that the chemostat can support, and Q is the flow rate of the mixture.

- The *SIR disease model* tries to predict the sizes of three subgroups—Susceptible, Infective, and Recovered—of a population of constant size N subjected to a disease pathogen. The model is

$$\frac{dS}{dT} = -pBSI\,, \qquad \frac{dI}{dT} = pBSI - KI\,, \frac{dR}{dT} = KI\,, \qquad B, K > 0\,, \quad 0 < p \leq 1,$$

where S, I, and R are the sizes of the three subgroup populations at time T, B is a parameter that quantifies the rate at which encounters among population members occurs, p is the probability of transmission in an encounter between an infected and a susceptible individual, and K is the recovery rate parameter.

2.5.3. Explain the chemostat model by comparing it with the exponential growth model

$$\frac{dN}{dT} = RN\,.$$

Specifically,

(a) What is the effect of the extra factor $1 - N/K$ in the growth term?
(b) What physical process in the chemostat is represented by the term $-QN$?
(c) What assumption does the algebraic form $-QN$ make about the process it describes?
(d) Fluid flows out of the chemostat at the rate Q, but the model does not include the volume as a dependent variable. Why not?

2.5.4. Explain the SIR model. Specifically,

(a) Why is the change in S proportional to both S and I?
(b) Why do the terms $pBSI$ and KI appear in two different equations, and why are they positive in one instance and negative in the other?
(c) The model

$$\frac{dR}{dT} = KR$$

represents exponential growth of R. Why does

$$\frac{dR}{dT} = KI$$

not represent exponential growth?
(d) Does the model clearly limit how large R can be? What do you expect will happen to I after a long time?

2.5.5. Consider a conceptual model similar to the chemostat model, but with two differences:

1. There is no fluid flow.
2. There is a predator that feeds on the bacteria. Assume that the population of predators is fixed at P and that the rate of predation per predator is given by the Holling type II model:

$$y(N) = \frac{sN}{1 + shN}\,, \qquad s, h > 0\,.$$

Write down the mathematical model that corresponds to this conceptual model.

2.5.6.* Suppose scientists discover that immunity to a disease is lost at a rate proportional to the recovered population with rate constant Q. Rewrite the SIR model equations to include this additional mechanistic process.

2.5.7. Suppose the original SIR model is changed by including vaccination of susceptibles at a rate VS, where V is the vaccination rate constant. Rewrite the model equations for this revised conceptual model.

2.5.8. For the chemostat model:

(a) Explain how we know that the dimension of K is the same as the dimension of N.
(b) Explain how we know that the dimension of R is 1/time.
(c) Determine the dimension of Q.

2.6 Mechanistic Modeling II: Equivalent Forms

After studying this section, you should be able to:

- Identify equivalent forms of a mathematical model.
- Convert a mathematical model to dimensionless form, given appropriate choices for reference quantities.
- Explain why equivalent forms do not always produce equivalent parameter values when fit to a data set.

In physics, almost everyone chooses the same way to write Newton's Second Law of Motion: $F = ma$. The model could be written in other ways, but this form has become standard. This uniformity makes it easy for a reader to compare material written by different authors. In contrast, uniformity of model appearance is rare in biology. Different authors inevitably choose different ways to write the same model. Hence, the reader who wants to understand mathematical models in biology needs to develop the skill of discriminating between different models and different forms of the same model. This is a skill that mathematicians take for granted, but it is problematic for most students. In this section, we consider different ways of writing the same model, working our way up from forms that differ only in notation to forms with more substantive differences.

2.6.1 Notation

Any given model can be written with a variety of notations. Sometimes these differences are due to the different settings in which the model can arise; other times, they are due simply to lack of standardization.

Example 2.6.1. Many important biochemical reactions are of the Michaelis–Menten type. The initial rate of reaction depends on the concentration of the principal reactant, called the substrate. In Chapter 7, we derive the well-known Briggs–Haldane model that relates the rate of reaction and substrate concentration. This model is generally written as

$$v_0 = \frac{v\max[S]}{K_M + [S]}\,, \qquad [S], v_{max}, K_M > 0\,,$$

where v_0 is the initial rate of the chemical reaction, $[S]$ is the concentration of one of the chemical reactants, v_{max} is the initial rate of chemical reaction for infinite $[S]$, and K_M is called the semisaturation parameter.

The Monod growth function in microbiology is used to model the rate of nutrient uptake by a microorganism as a function of the concentration of nutrients in the environment. The Monod model, in one of several common notations, is

$$r = \frac{qS}{A+S}\,, \qquad S,q,A > 0\,,$$

where r is the rate of nutrient uptake, S is the concentration of the nutrient, q is the maximum uptake rate, and A is the semisaturation parameter.

The Briggs–Haldane and Monod models were designed for different contexts. Although the interpretation of each model is based on its own context, the models are mathematically identical, differing only in notation. □

The notation for the Briggs–Haldane equation is now standard, while the Monod growth function has almost as many systems of notation as the number of authors who have written about it. Although the two formulas use different symbols, they are identical in mathematical form. *Both formulas say that the dependent variable is a rational function of the independent variable, with the numerator of the form "parameter times independent variable" and the denominator of the form "parameter plus independent variable."* If we focus on the *symbols*, we are misled by the apparent differences. If we focus on the *roles* of the symbols, we see that the formulas are identical. Few symbols in mathematics have a fixed meaning. Even π, which is the universal symbol for the ratio of circle circumference to circle diameter, is occasionally used for other purposes. This lack of standardization makes it necessary to provide the meaning of each symbol for any given context.

2.6.2 Algebraic Equivalence

Occasionally, it is possible to write the same model in two different ways by performing algebraic operations. This is most common with formulas that include a logarithmic or exponential function.

Example 2.6.2. In fitting an exponential model by linearization in Section 2.4, we converted the model

$$z = Ae^{-kt}$$

to the form

$$\ln z = \ln A - kt\,.$$

These forms are equivalent in the sense that they both produce the same z values for any given t value. □

Example 2.6.3. A certain problem in the life history of plants eventually reduces to the algebraic equation

$$Pe^{-P} = e^{-R}\,,$$

where R is the independent variable and P is the dependent variable. Hence, the formula defines P implicitly in terms of R. Various algebraic operations can be performed to change this equation. For example, we can multiply by e^P and e^R to obtain the equation

$$Pe^R = e^P \, .$$

We can then take the natural logarithm on both sides, obtaining

$$\ln P + R = P \, .$$

Further rearrangement yields

$$P - \ln P = R \, .$$

All of these forms are algebraically equivalent, although they appear different at first glance. □

None of the formulas in Example 2.6.3 allows us to solve algebraically for P. The choice among them is a matter of taste. I prefer the last one, because that form allows us to graph the relationship by calculating values of R for given values of P.

2.6.3 Different Parameters

More subtle than differences in algebraic presentation are differences resulting from parameter choices. Sometimes it requires some algebraic manipulation to see that two similar models are actually equivalent.

Example 2.6.4. In Section 2.5, we developed the Holling type II model

$$y(x) = \frac{sx}{1+hsx} \, , \qquad x,s,h > 0 \, , \tag{2.6.1}$$

for the relationship between the number of prey animals eaten by a predator in a unit amount of time (y) and the number of prey animals available in a specific region (x), where s is the amount of habitat that a predator can search per unit time and h is the time required for the predator to process one prey animal. The same model can also be written as

$$y(x) = \frac{qx}{a+x} \, , \qquad x,q,a > 0 \, , \tag{2.6.2}$$

where the parameters q and a are different from the original s and h. The *functions* in (2.6.1) and (2.6.2) are mathematically different, so the models are not *identical*. However, the *models* represented by the functions are *equivalent* if we define the parameters in (2.6.2) by $q = 1/h$ and $a = 1/(hs)$. Note that the Holling type II model is also equivalent to the Briggs–Haldane and Monod models, with yet another context. □

Check Your Understanding 2.6.1:
Substitute $q = 1/h$ and $a = 1/(hs)$ into (2.6.2) and simplify the result to obtain (2.6.1).

Each of the equivalent forms (2.6.1) and (2.6.2) is preferable from a particular point of view. Form (2.6.2) is preferable from a graphical point of view, because the parameters q and a indicate properties directly visible on the graph (see Problem 1.1.10). It has the additional advantage of being semilinear (as defined in Section 2.4), so we can fit it to empirical data by the semilinear least squares method.[34] Form (2.6.1) is better from a biological point of view, because it allows us to study the effects of search speed and maximum consumption rate separately.

[34] See Example 2.7.2.

2.6.4 Visualizing Models with Graphs

The appearance of a graph depends on the ranges chosen for the variables. In making a choice, it is important to have a purpose in mind. Consider three instances of the predation model (2.6.2):

$$y_1 = \frac{2x}{0.5+x}\,, \qquad y_2 = \frac{2x}{2+x}\,, \qquad y_3 = \frac{0.5x}{2+x}\,.$$

If our purpose is to see what effect the different choices of q and a have on the graph, we should plot the three models together, as in Figure 2.6.1a. But suppose instead that we want to plot each curve separately in a way that best shows the behavior of the function. The ranges chosen for Figure 2.6.1a look best for y_2, so we plot this curve with the same ranges (Figure 2.6.1d).

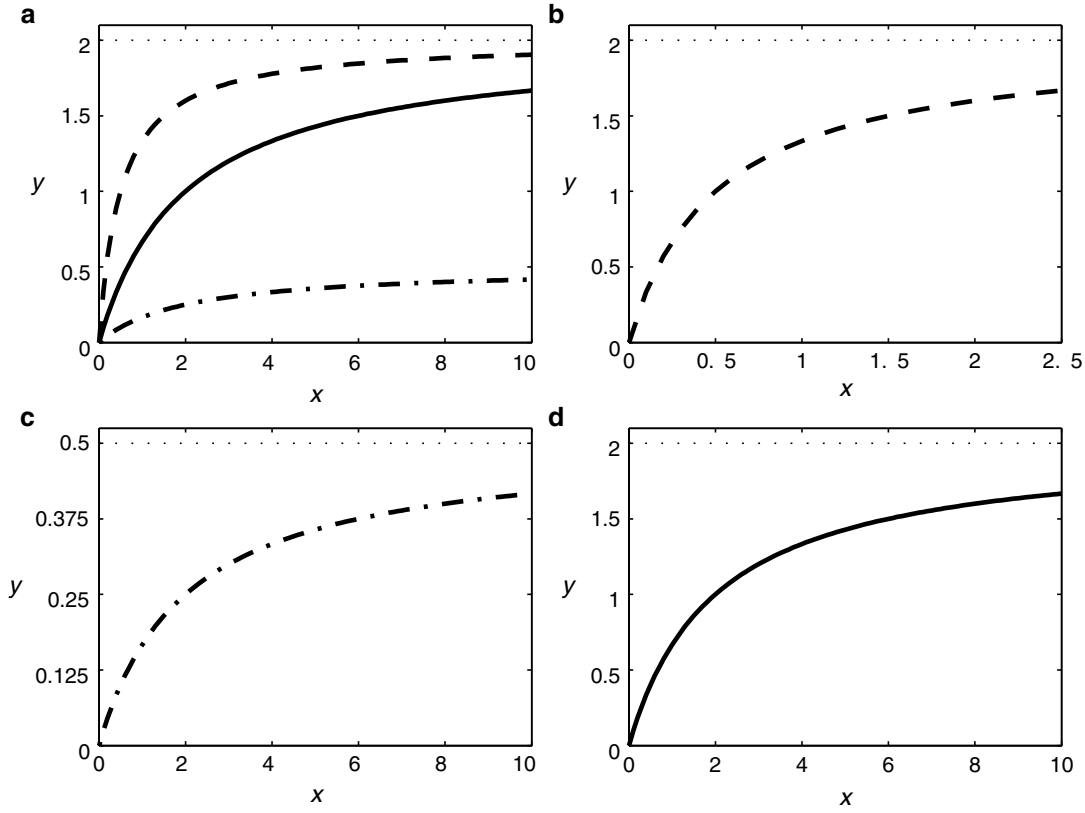

Fig. 2.6.1 The model $y(x) = qx/(a+x)$, with (q,a) values of $(2,0.5)$ (*dashed*), $(2,2)$ (*solid*), $(0.5,2)$ (*dash-dotted*)

The function y_3 has a much lower maximum value than y_2, so we might choose a narrower range for the vertical axis (Figure 2.6.1c). On the other hand, most of the variation in y_1 occurs on the left side of Figure 2.6.1a, so we might choose a narrower range for the horizontal axis (Figure 2.6.1b). Notice that the three individual graphs are now identical, except for the numbers on the axes.

2.6.5 Dimensionless Variables

Look again at Figure 2.6.1. Notice that the vertical axis range is $0 \le y \le q$ for each of the single-curve plots. Similarly, the horizontal axis range for each of the curves is $0 \le x \le 5a$. We can make use of this observation to produce one plot with axis values that are correct for these three cases, and all others as well. One way is to incorporate the parameters q and a into the axis values (Figure 2.6.2a). Alternatively, we could incorporate the parameter values into the variables themselves, as in Figure 2.6.2b.

The quantities y/q and x/a are dimensionless versions of the original quantities y and x. This means that they represent the same things (food consumption rate and food density, respectively), but are measured differently. Where y is the food consumption rate measured in

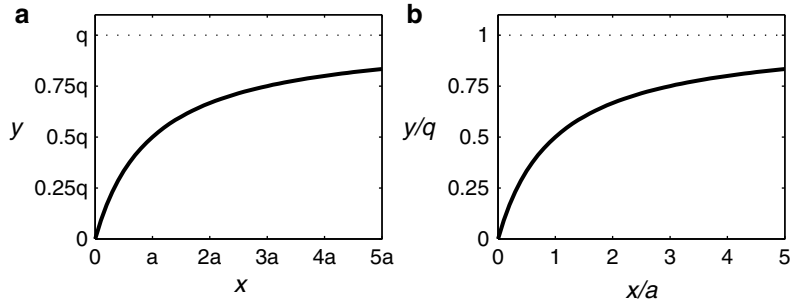

Fig. 2.6.2 The model $y(x) = qx/(a+x)$, using two different labeling schemes: (**a**) the factors a and q are in the axis values; (**b**) the factors a and q are in the axis labels

terms of some convenient but arbitrary unit of measurement (prey animals per week or grams per day, for example), y/q is the food consumption rate measured as a fraction of the maximum food consumption rate q. Where x is the nutrient concentration in a convenient but arbitrary unit, x/a is the nutrient concentration relative to the semisaturation concentration a.

2.6.6 Dimensionless Forms

The quantities y/q and x/a have an algebraic benefit as well as a graphical benefit. To see this benefit, we define dimensionless variables Y and X by[35]

$$Y = \frac{y}{q}\,, \qquad X = \frac{x}{a}\,. \tag{2.6.3}$$

Rearranging these definitions yields substitution formulas:

$$y = qY\,, \qquad x = aX\,. \tag{2.6.4}$$

Now we replace y and x in (2.6.2) using the formulas of (2.6.4). This yields the equation

[35] In this text, we will usually adopt the practice of using one case for all of the original variables in a model and the opposite case for the corresponding dimensionless variables. Other systems, such as those that add accent marks for either the dimensional or the dimensionless quantities, have greater flexibility but other disadvantages. Distinguishing by case has the advantage of easy identification of corresponding quantities without the clumsiness of accent marks.

$$qY = \frac{qaX}{a+aX} . \tag{2.6.5}$$

Removing common factors yields the dimensionless form of the predation model:

$$Y = \frac{X}{1+X} . \tag{2.6.6}$$

The graph of the dimensionless model is the same as Figure 2.6.2b, except that the axis labels are X and Y instead of x/a and y/q. The dimensionless variables measure quantities in terms of units that are intrinsic to the physical setting. The statement $y = 0.5$ cannot be understood without a unit of measurement or a value of q for comparison. The statement $Y = 0.5$ can immediately be understood as a predation rate that is half of the maximum predation rate.

Example 2.6.5. Consider the energy budget model from Example 1.5.1:

$$q = ax^2 - bx^3 , \tag{2.6.7}$$

where x is the length of an organism, q is the surplus energy available for growth or reproduction, and a and b are parameters. Suppose we want to obtain a dimensionless form for this model. One way to do this is to identify quantities made from the parameters that have the same dimensions as x and q. We would first have to find the dimensions of a and b using the rules for dimensional consistency introduced in Section 2.5.

An alternative method is to assume a substitution formula of the form

$$\text{dimensional variable} \;=\; \text{dimensional reference quantity} \;\times\; \text{dimensionless variable}$$

using an unknown dimensional reference quantity, which can be chosen later. With this in mind, we write the substitution formulas

$$q = q_r Q, \qquad x = x_r X ,$$

where Q and X are to be the dimensionless forms of q and x, and q_r and x_r are the yet-to-be-determined reference quantities. With these substitutions, the model becomes

$$q_r Q = a x_r^2 X^2 - b x_r^3 X^3 .$$

The three terms in this equation must all have the same dimension. Since Q and X are dimensionless, this means that q_r, ax_r^2, and bx_r^3 all have the same dimension as q. At this point, the choices of q_r and x_r are free, so we can either choose them directly or we can specify two equations that they must satisfy. The simplest option is to make all three coefficients equal. Thus, we have

$$q_r = a x_r^2 = b x_r^3 .$$

From these equations, we get the results

$$x_r = \frac{a}{b} , \qquad q_r = \frac{a^3}{b^2} . \tag{2.6.8}$$

Given these choices, the final dimensionless model is

$$Q = X^2 - X^3 . \tag{2.6.9}$$

□

Differential equation models can also be nondimensionalized. This requires substitution formulas for the independent variable of derivatives. It is easiest to write these substitution formulas in terms of the differentials.

Example 2.6.6. Consider the differential equation

$$\frac{dy}{dt} = -ky\,.$$

We can define new variables Y and T by

$$Y = \frac{y}{y_0}\,, \qquad T = kt\,.$$

These formulas corresponds to the differential substitution formulas $dy = y_0\,dY$ and $dT = k\,dt$, so we have the substitution formulas

$$\frac{d}{dt} = k\frac{d}{dT}\,, \qquad \frac{dy}{dt} = ky_0\frac{dY}{dT}\,.$$

The final result is

$$\frac{dY}{dT} = -Y\,. \tag{2.6.10}$$

□

Dimensionless models are not always parameter-free like (2.6.6), (2.6.9), and (2.6.10), but they always have fewer parameters than the original dimensional model. That alone makes them preferable for model characterization. Sometimes making a model dimensionless provides valuable insight into the behavior of the model, even before any analysis is performed.[36]

Problems

2.6.1. Derive (2.6.2) from (2.6.1). In so doing, find the algebraic formulas that define h and s in terms of the parameters q and a.

2.6.2. Explain the graphical significance of the parameter a in the model

$$y = \frac{qx}{a+x}\,.$$

[Hint: What is y if you set $x = a$?]

2.6.3.* Rewrite the model $y = y_0e^{-kt}$ in dimensionless form by choosing appropriate dimensionless quantities to replace y and t.

2.6.4. Create a set of four plots similar to Figure 2.6.1 using the functions

$$y_1 = \frac{4x}{2+x}\,, \qquad y_2 = \frac{2x}{2+x}\,, \qquad y_3 = \frac{2x}{4+x}\,.$$

Your plots of the individual functions should be identical to those of Figure 2.6.1, except for the numerical values on the axes.

[36] See Chapter 7.

2.6.5. (a) Redo the nondimensionalization of the energy budget model $q = ax^2 - x^3$ using as reference quantities the value $x^\star$ that maximizes q and the corresponding maximum value $q^\star = q(x^\star)$.
(b) Create a plot of the model $q = ax^2 - x^3$ similar to Figure 2.6.2, using the same reference quantities as part (a).

2.6.6. (Continued from Problem 1.5.3.)
Redo Problem 1.5.3, but nondimensionalize the function in part (c) before solving the optimization problem in part (d).

2.6.7. (Continued from Problem 1.6.7.)
Repeat Problem 1.6.7, but without assuming $r = 1$. Instead, replace the quantities $x(t)$, $z(t)$, and s using $x = YX$, $z = YZ$, $t = T/r$, and $s = rS$ and determine the optimal jumping point $X^\star$ in terms of the dimensionless swim speed S. Note that the key equations for the model are $z^2 = x^2 + Y^2$ and $dx/dt = r$, which are used to determine the rate of change of z with respect to t while running, and $dz/dt = s$, which defines the rate of change of z with respect to t while swimming.

2.6.8.* In Section 2.5, we derived the model

$$\frac{dC}{dT} = \frac{F}{V}C_{\text{in}} - \frac{F}{V}C(T) - RC(T)$$

for the concentration $C(T)$ of pollutant in a lake of volume V, where F is the flow rate of water in and out of the lake and R is the decay rate of the pollutant. Nondimensionalize the differential equation using C_{in} for the reference concentration and V/F for the reference time. Your model will contain one dimensionless parameter. Explain the biological meaning of that parameter.

Problems 2.6.9–2.6.11 use the chemostat and SIR models that were introduced in the Section 2.5 problem set.

2.6.9. (a) Dimensionless variables for the chemostat model can be defined by

$$n = \frac{N}{K}, \qquad t = RT\ .$$

Use these definitions to write appropriate substitution formulas to replace N and T in the model.
(b) Obtain a dimensionless model by making the substitutions for N and T. Note that you will have to divide through by RK so that the left side of the model will be dn/dt. The last term of the equation should include a dimensionless parameter q, which you will need to define.

2.6.10. (a) Determine the dimension of B and the dimension of K in the SIR model. Show that the quantities

$$s = \frac{S}{N}, \qquad i = \frac{I}{N}, \qquad t = KT$$

are dimensionless. (Note that p is dimensionless.)
(b) Use the definitions of s, i, and t to obtain substitution formulas for S, I, and d/dT.
(c) Use the substitution formulas of part (b) to derive a dimensionless model for the variables s and i. (You need not consider the R equation, as the first two form a self-contained model and $R = N - S - I$.) Your model should include a single dimensionless parameter, b, which represents the rate of spread of the infection.

2.6.11. An SI epidemic model is similar to an SIR epidemic model, except that recovered individuals are assumed to be susceptible again. The model can be written as

$$\frac{dS}{dT} = -pBSI + KI, \qquad \frac{dI}{dT} = pBSI - KI, B, K > 0, \quad 0 < p \leq 1\ ,$$

where S and I are the sizes of the subgroups at time T, B is the encounter rate parameter, p is the transmission probability, and K is the recovery rate parameter.

(a) Let $N = S + I$. Show that $dN/dT = 0$.
(b) Use the fact that N is constant to eliminate the variable S from the I equation. The result is a model consisting of a single equation for I with four parameters (p, B, K, and N).
(c) Determine the dimension of B and the dimension of K. Show that the quantities

$$i = \frac{I}{N}\ , \qquad t = KT$$

are dimensionless. (Note that p is dimensionless.)
(d) Use the definitions of i and t to obtain substitution formulas for I and d/dT.
(e) Use the substitution formulas of part (c) to derive a dimensionless model for the variable i. Your model should include a single dimensionless parameter, b, which represents the rate of spread of the infection.

2.7 Empirical Modeling III: Choosing Among Models

After studying this section, you should be able to:

- Use the Akaike information criterion to compare the statistical validity of models as applied to a given set of data.
- Choose among models based on a variety of criteria.

In this section, we consider the issue of how to choose from among several different models for a given setting. There are three key criteria: quantitative accuracy, complexity, and the availability of a mechanistic model.

2.7.1 Quantitative Accuracy

Quantitative accuracy is clearly an important criterion for choosing among possible models. It can easily be measured by the residual sum of squares, but comparisons can only be made when the residual sums of squares are measured on the same graph.

Example 2.7.1. In Examples 2.4.2 and 2.4.4, we fit a power function model using linear least squares on the linearized data and semilinear least squares on the original data. Each set of parameter values wins on its "home field"; the semilinear result is clearly better when viewed on a plot of y versus x, while the linearized result is clearly better when viewed on a plot of $\ln y$ versus $\ln x$. The question of better fit in this case is not settled mathematically, but by the choice

of which graph to use to measure errors. Unless we have a scientific reason for considering a graph of logarithms to be more meaningful than a graph of original values, we should use the semilinear method.[37] □

Example 2.7.2. Using the semilinear least squares method to fit the Holling type II model

$$y = \frac{qx}{a+x}$$

to the *P. speedius* data from Table 2.1.1, we obtain

$$a = 36.4\,, \qquad q = 36.3\,, \qquad \text{RSS} = 43.2\,. \tag{2.7.1}$$

Previously we fit the power function model $y = Ax^p$ to the same data, with the result RSS = 62.9.[38] The Holling type II model has a significantly smaller residual sum of squares than the power function model. Quantitative accuracy favors the Holling model; other criteria will be considered below. □

2.7.2 Complexity

Before we conclude that model selection is usually a simple matter, we turn to the question "Shouldn't we *always* use the model that has the smallest residual sum of squares?" The answer to this question is an emphatic *NO*, as the following example makes clear.

Table 2.7.1 A small partial data set for consumption rate y as a function of prey density x

x	0	20	40	60	80
y	0	7	9	21	25

Table 2.7.2 A larger partial data set for consumption rate y as a function of prey density x

x	0	10	20	30	40	50	60	70	80
y	0	2	7	10	9	14	21	20	25

Example 2.7.3. Suppose we collected the *P. steadius* data from Section 2.1 in stages. The initial data set, in Table 2.7.1, consists of every other point from the first part of the full data set. We could fit the models $y = mx$ and $y = b + mx$ to this data set, and we would find that the second model yields a smaller residual sum of squares than the first model. In fact, it is possible to show that adding a parameter to a model *always* decreases the residual sum of squares until it falls to 0. So if our goal is minimum residual sum of squares, we can just add more parameters. A fourth degree polynomial has five parameters, which means that we can find one whose graph

[37] For models that have more than one nonlinear parameter, one can use a fully nonlinear method. These can be found in any mathematical software package, such as R or Maple. Other packages, such as spreadsheets, that fit exponential or power function models to data use the linearization method.

[38] See Example 2.4.4.

passes through all five data points, giving us a residual sum of squares of 0. The result is the model

$$y \approx 1.1375x - 0.0628x^2 + 0.00134x^3 - 0.00000859x^4 .$$

On the basis of quantitative accuracy for the data points used to obtain the parameters, this model is perfect. However, the graph of this model shows a problem. Figure 2.7.1a shows the five points used for the fit, the fourth degree polynomial that passes through the points, and the least squares line $y = 0.313x$. The sharp curves in the polynomial are suspicious.

Table 2.7.2 includes the full set of data. A useful model should remain reasonably good when more data is added; however, Figure 2.7.1b shows that the fourth degree polynomial has very little predictive value for the extra data in this case. If we fit a fourth degree polynomial to the larger set of data, the graph of the best fit will be significantly changed. Meanwhile, the model $y = 0.313x$ looks reasonably good with the larger set of data. □

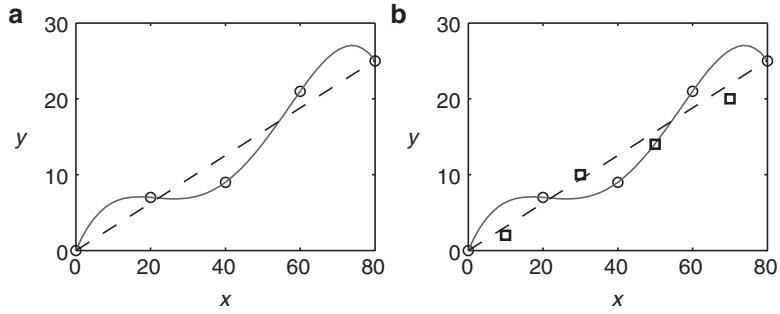

Fig. 2.7.1 The linear and fourth degree polynomial models fit to the data of Table 2.7.1, with the additional data of Table 2.7.2 included in (**b**)

Example 2.7.3 shows that there is such a thing as too much accuracy in fitting a data set. Adding parameters to a model increases the accuracy by allowing the model graph to conform more closely to the individual data points. However, real data has measurement uncertainty as well as the demographic stochasticity discussed in Section 2.1. This means that bending the graph to make it conform to the data can be overdone, resulting in a loss of predictive power. We can use this discovery to formulate a general principle of model selection[39]:

Additional parameters should not be added to models unless the increased accuracy justifies the increase in complexity.

Note that we are not arguing that simpler models are better than complex models for philosophical or aesthetic reasons. We are saying that greater complexity can sometimes lead to an undesirable loss of predictive power.

[39] This statement is a mathematical version of *Occam's razor*, a well-known scientific principle attributed to the fourteenth-century philosopher William of Ockham, although not found in his extant writings and actually appearing in some form before Ockham. The most common form was written by John Punch in 1639 and translates literally from the Latin original as "Entities must not be multiplied beyond necessity." My interpretation is much more in keeping with the actual statement than its more common renderings in English.

2.7.3 The Akaike Information Criterion

We have now established the idea that quantitative accuracy needs to be balanced against complexity. We can quantify accuracy using the residual sum of squares and complexity using the number of parameters. The challenge is to combine these measures in a meaningful way. As an example, consider the *P. steadius* data set, which we fit in Section 2.1 using the linear models $y = mx$ and $y = b + mx$, with residual sums of squares 106.1 and 100.4, respectively. Is the improvement in accuracy for the model $y = b + mx$ worth the added complexity?

The issue of balancing accuracy and complexity was addressed by Hirotugu Akaike using information theory and statistics in a 1974 paper, which defined what has come to be known as the Akaike information criterion, or AIC [1]. An alternative measure, the corrected Akaike information criterion (AICc), is also in common use. Based on recent work of Shane Richards [12], I recommend using the original version, which we can write as

$$\text{AIC} = n \ln\left(\frac{\text{RSS}}{n}\right) + 2(k+1)\,, \tag{2.7.2}$$

where RSS is the residual sum of squares for the specific model, k is the number of parameters that are in the specific model, and n is the fixed number of data points used to determine the model parameters.[40] The actual values of AIC are unimportant. What matters is that smaller values represent a higher level of statistical support for the corresponding model. A comparison can only be made when both models are fit to the same data set using a residual sum of squares defined in the same manner.

Armed with the AIC, we are now ready to address the issue of model selection for the *P. steadius* predation data.

Example 2.7.4. In Section 2.3, we obtained two results from linear least squares for the full *P. steadius* data set[41]:

$$y = 0.267x\,,$$

$$y = 1.175 + 0.255x\,.$$

As an additional empirical option, we can also determine (using methods beyond the scope of this section) the best-fit parabola to be

$$y = 0.197 + 0.300x - 0.00032x^2\,.$$

Finally, the best fit for the Holling Type II model is

$$y = \frac{196x}{628 + x}\,.$$

Table 2.7.3 summarizes the results and Figure 2.7.2 displays the data and the various models. The model $y = mx$ has the smallest AIC value. The very slight improvements in accuracy for the other models do not quite compensate for the additional complexity. □

[40] The usual formula for AIC ends with $2K$ rather than $2(k+1)$, where K is the number of parameters that have to be fit using statistics. This includes the statistical variance along with the k model parameters. For those readers not well versed in statistics, it is easier to count the number of model parameters than the number of statistical parameters, so our version builds the extra parameter into the formula.

[41] See Table 2.1.1.

Table 2.7.3 Comparison of three models for the *P. steadius* data of Table 2.1.1 using AIC

Model	RSS	n	k	$n \ln\left(\frac{\text{RSS}}{n}\right)$	$2(k+1)$	AIC
$y = mx$	106.1	15	1	29.3	4	33.3
$y = b + mx$	100.4	15	2	28.5	6	34.5
$y = b + mx + ax^2$	96.1	15	3	27.9	8	35.9
$y = qx/(a+x)$	95.5	15	2	27.8	6	33.8

2.7.4 Choosing Among Models

Choosing a model is a matter of informed judgment. The best choice using empirical criteria is determined by the AIC; however, AIC results are only as good as the data set being fit. When two models have very close AIC scores with a given data set, there is a possibility that the rankings would be reversed with a different data set. It is reasonable to ask how much of a difference in AIC should be considered definitive. We must consider two separate cases.

A pair of models is said to be *nested* if one can be obtained from the other merely by setting one of its parameters to zero. For example, the pair $y = mx$ and $y = b + mx$ is nested, as is the pair $y = mx$ and $y = qx/(a+x)$; however, the pair $y = b + mx$ and $y = qx/(a+x)$ is not nested. For a nested pair, the best the simpler model can do is to achieve the same residual sum of squares. In this case, the complexity penalty gives the simpler model an AIC advantage of 2. Thus, a difference of 2 must be considered significant for a nested pair. As a rule of thumb, an AIC difference of 1 in a nested pair should probably be taken as definitive.

In contrast, it is harder to be confident in the significance of AIC differences for models that are not a nested pair. A difference of 6 is generally taken to be definitive, but this seems far too conservative, as a difference of 6 is achieved when two models have the same residual

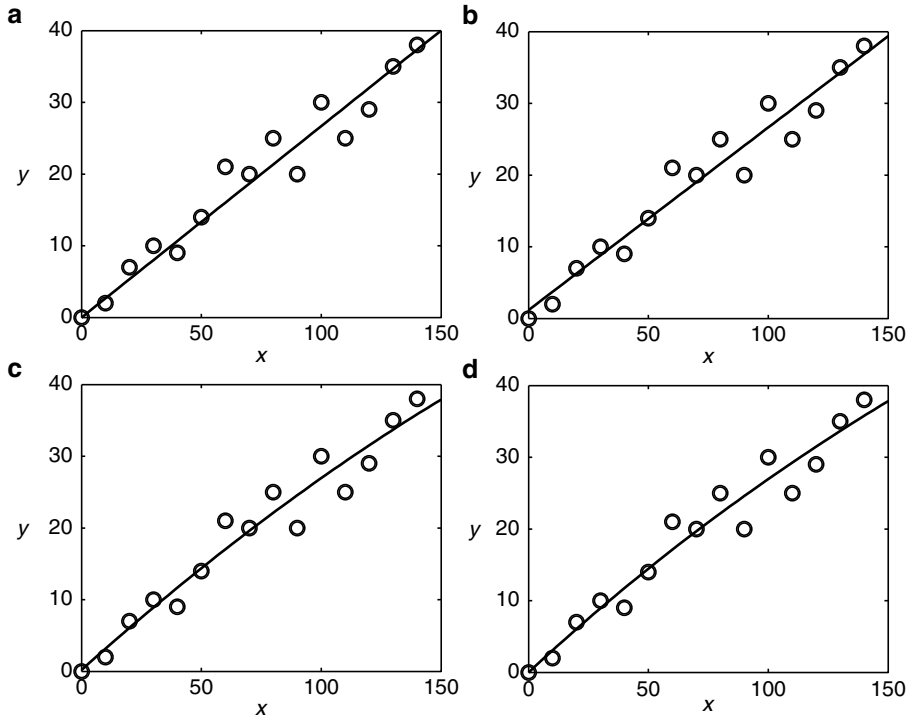

Fig. 2.7.2 The data and models of Example 2.7.4: (**a**) $y = mx$, (**b**) linear, (**c**) quadratic, (**d**) Holling type II

sum of squares but one has three more parameters than the other. It seems intuitively clear that the model with three fewer parameters is the better choice, pending new information such as additional data. In practice, a difference of 2 is probably enough in most cases. It is hard to argue on empirical grounds that a model with an extra parameter that fails to achieve a lower residual sum of squares than a simpler model can be the best choice, even when the models are not nested.

AIC has the advantage of being quantitative, but we should not overestimate its value. Non-quantitative criteria must often be considered. We should usually prefer models that have a mechanistic justification, even if such models have AIC values that are not significantly better than the best empirical model.

Example 2.7.5. Q) Which model should we use for *P. steadius*?

A) In Example 2.7.4, we found the lowest AIC score for the model $y = mx$, but with only a slightly higher score for the Holling type II model. A different data set for the same experiment could yield a lower AIC score for the Holling model, or the difference could be greater than in the example. Both models have a mechanistic justification, which the other two models lack. Either is arguably the best choice. In practice, it is generally wise not to use a more complicated model than necessary, so the simple $y = mx$ is probably the better choice. □

Example 2.7.6. Q) Which model should we use for *P. speedius*?

A) We have already obtained[42] best fits for the power function and Holling type II models. We can also try polynomials of various degrees. Table 2.7.4 summarizes the results and Figure 2.7.3 illustrates the four models with the original data. According to the AIC values, the Holling type II model is the best. The cubic curve scores close to the Holling type II; however, the shape of the cubic curve is clearly not quite right, nor is it supported by a mechanistic justification. Note how the curve turns sharply upward at the right end of the interval. □

Table 2.7.4 Comparison of four models for the *P. speedius* data of Table 2.1.1 using AIC

Model	RSS	n	k	$n \ln\left(\frac{\text{RSS}}{n}\right)$	$2(k+1)$	AIC
$y = ax^p$	62.9	15	2	21.5	6	27.5
$y = qx/(a+x)$	43.2	15	2	15.9	6	21.9
$y = b + mx + ax^2$	88.1	15	3	26.6	8	34.6
$y = b + mx + ax^2 + cx^3$	38.5	15	4	14.1	10	24.1

2.7.5 Some Recommendations

The Akaike information criterion should be thought of as a theoretical tool for empirical modeling. When we want to determine calculated values to represent theoretical data for an experiment, we should usually choose the model with the smallest AIC. Nevertheless, there are pitfalls that we must avoid:

- Empirical models that fit data well over a given range may not work over a larger range. Empirical models should not be extrapolated.

[42] See Examples 2.4.4 and 2.7.3.

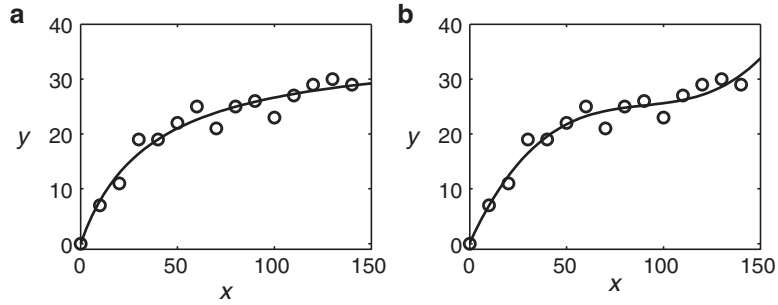

Fig. 2.7.3 The data and the two best models of Example 2.7.6: (**a**) Holling type II, (**b**) cubic

- The ranking of models for a set of data depends on the location of the points on a graph, which can be expected to vary because of random factors. If we collect a new set of data from the same experiment, the model selected by AIC for the first data set may not be the best fit for the second set.
- Part of the amazing success of models in the physical sciences owes to their mechanistic derivation from physical principles. Empirical models, by definition, do not attempt to do this. We can have more confidence in a model that can be obtained mechanistically than in one that is strictly empirical.

For theoretical work, we should always prefer a mechanistic model, even if its AIC value for a given data set is a little larger than that of a strictly empirical model. Of course, we must also be careful with mechanistic models. If the data from experiments fits an empirical model significantly better than it fits a mechanistic model, then either the experimental data is severely flawed or the mechanistic model is not valid for that experimental setting.

Problems

2.7.1.(a) Compute the residual sum of squares for the model

$$y \approx 1.1375x - 0.0628x^2 + 0.00134x^3 - 0.00000859x^4$$

with the full data set of Table 2.7.2.

(b) Compute the residual sum of squares for the model $y = 0.313x$ with the full data set of Table 2.7.2.

(c) Both of the models used in parts (a) and (b) were obtained by fitting to the partial data set of Table 2.7.1. Which has the better quantitative accuracy for the full data set?

2.7.2.* The points corresponding to x values of 0, 10, 20, and 30 in Table 2.7.2 seem to lie close to the least squares line $y = 0.313x$. We can find a polynomial of the form $y = ax + bx^2 + cx^3$ that passes through the points $(0,0)$, (h,y_1), $(2h,y_2)$, and $(3h,y_3)$ by the *method of successive differences*. The method yields simple formulas for the three coefficients:

$$c = \frac{y_3 - 3y_2 + 3y_1}{6h^3}\,, \qquad b = \frac{y_2 - 2y_1}{2h^2} - 3hc\,, \qquad a = \frac{y_1}{h} - hb - h^2c\,.$$

(a) Use these coefficient formulas to fit the data from Table 2.7.2 for the x values 10, 20, and 30.
(b) Plot the polynomial from part (a), along with the three data points, on a common graph.
(c) Does the cubic polynomial model work well for these data? Why or why not?

2.7.3. Using the *method of successive differences*, relatively simple formulas can be found to determine the coefficients of a polynomial of degree n that passes through $n+1$ points having equally spaced x values. Suppose the points are (x_1,y_1), (x_2,y_2), (x_3,y_3), and (x_4,y_4), and $x_4 - x_3 = x_3 - x_2 = x_2 - x_1 = h$. Then the polynomial $y = a + bx + cx^2 + dx^3$ has coefficients given by

$$d = \frac{y_4 - 3y_3 + 3y_2 - y_1}{6h^3}\,, \qquad c = \frac{y_3 - 2y_2 + y_1}{2h^2} - 3x_2 d\,,$$
$$b = \frac{y_2 - y_1}{h} - (x_1 + x_2)c - (3x_1x_2 + h^2)d\,, \qquad a = y_1 - x_1 b - x_1^2 c - x_1^3 d\,.$$

(a) Use these coefficient formulas to fit the data from Table 2.7.2 for the x values 10, 30, 50, and 70.
(b) Repeat part (a) using the data for the x values 20, 40, 60, and 80.
(c) Plot the polynomials from parts (a) and (b), along with all of the data, on a common graph.
(d) How good are these two models for the full data set?

2.7.4. (Continued from Problems 2.3.1 and 2.3.7.)

(a) Use your *P. steadius* data set from Problem 2.1.1 to fit the Holling type II model.
(b) Compute the AIC for the Holling type II model, the model $y = mx$ (Problem 2.3.1a), and the model $y = b + mx$ (Problem 2.3.7a).
(c) Which model gives the lowest AIC with your data? Combining these results with Example 2.7.4, which model do you recommend should be used for *P. steadius*? [Note: Comparison of AIC values is only meaningful when those values were obtained from one data set.]

2.7.5. (Continued from Problem 2.4.6.)

(a) Use your *P. speedius* data set from Problem 2.1.2 to fit the Holling type II model.
(b) Compute the AIC for the Holling type II model and the model $y = Ax^p$ (Problem 2.4.6b).
(c) Which model gives the lowest AIC with your data? Combining these results with Example 2.7.5, which model do you recommend should be used for *P. speedius*? [Note: Comparison of AIC values is only meaningful when those values were obtained from one data set.]

2.7.6.* Suppose we have a set of data points (x_i,y_i), which we would like to fit to a model $y = b + mx + ax^2$.

(a) Suppose further that we construct a new data set (X_i,Y_i) with $X_i = x_i - \bar{x}$ and $Y_i = y_i - \bar{y}$. By substituting $x = X + \bar{x}$ and $y = Y + \bar{y}$ into the model $y = b + mx + ax^2$, find the parameters B and M, in terms of b, m, c, $\bar{x}$, and $\bar{y}$, such that the model is equivalent to $Y = B + MX + aX^2$.
(b) Use the result from part (a) to obtain formulas for b and m in terms of B, M, a, $\bar{x}$, and $\bar{y}$.
(c) Set up a spreadsheet to fit the model $y = b + mx + ax^2$ to a set of data by converting the data to X and Y as in part (a), calculating B, M, and a, and then finding b and m as in part (b). The formulas for B, M, and a can be shown (using calculus) to be

$$a = \frac{\sum X^2 Y \sum X^2 - \sum XY \sum X^3}{\sum X^2 \sum X^4 - (\sum X^3)^2 - (\sum X^2)^3/n}\,, \quad M = \frac{\sum XY - a\sum X^3}{\sum X^2}\,, \quad B = -\frac{a\sum X^2}{n}\,,$$

where n is the number of data points. Test your spreadsheet by confirming the results of Example 2.7.4.

2.7.7. (Continued from Problems 2.3.4 and 2.4.3.)

(a) Use the spreadsheet from Problem 2.7.6 to fit the model $y = b + mx + ax^2$ to the data from Problem 2.3.4.
(b) Use the AIC to compare the results of part (a) with the linear model $y = b + mx$ and the results from Problem 2.4.3 for the model $y = Ax^p$. Are the AIC differences significant? Is the best model very good?

2.7.8. (Continued from Problems 2.3.5 and 2.4.4.)

(a) Use the spreadsheet from Problem 2.7.6 to fit the model $y = b + mx + ax^2$ to the data from Problem 2.3.5.
(b) Use the AIC to compare the results of parts (a) with the linear model $y = b + mx$ and the results from Problem 2.4.4 for the models $y = Ax^p$ and $y = Ae^{kt}$. Are the AIC differences significant? Is the best model very good?

2.7.9. (Continued from Problem 2.3.8.)

(a) Use the spreadsheet from Problem 2.7.6 to fit the model $y = b + mx + ax^2$ to the data from Problem 2.3.8, from 1979 to 2003.
(b) Use the AIC to compare the results with the linear model.

2.7.10. (Continued from Problem 2.3.9.)

(a) Use the spreadsheet from Problem 2.7.6 to fit the model $y = b + mx + ax^2$ to the data from Problem 2.3.9, from 1970 to the present.
(b) Use the AIC to compare the results with the linear model.

References

1. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**: 716–723 (1974)
2. Atkins GL and IA Nimmo. A comparison of seven methods for fitting the Michaelis–Menten equation. *Biochem J.*, **149**, 775–777 (1975)
3. Atkinson, MR, JF Jackson, and RK Morton. Nicotinamide mononucleotide adenylyltransferase of pig-liver nuclei: The effects of nicotinamide mononucleotide concentration and pH on dinucleotide synthesis. *Biochem J.*, **80**, 318–323 (1980)
4. Atlantic States Marine Fisheries Commission. Atlantic Croaker 2010 Stock Assessment Report. *Southeast Fisheries Science Center, National Oceanic and Atmospheric Administration* (2010). http://www.sefsc.noaa.gov/sedar/Sedar_Workshops.jsp?WorkshopNum=20 Cited in Nov 2012
5. Chuine I, P Yiou, N Viovy, B Seguin, V Daux, and EL Ladurie. Grape ripening as a past climate indicator. *Nature*, **432**, 18 (2004)
6. Holling CS. Some characteristics of simple types of predation and parasitism. *Canadian Entomologist*, **91**: 385–398 (1959)
7. Ledder G. An experimental approach to mathematical modeling in biology. *PRIMUS*, **18**, 119–138 (2007)
8. Ledder G. BUGBOX-predator (2007). http://www.math.unl.edu/~gledder1/BUGBOX/ Cited Sep 2012
9. Lineweaver H and D Burk. The determination of enzyme dissociation constants. *Journal of the American Chemical Society*, **56**, 658–666 (1934)
10. Motulsky H and A Christopoulos. *Fitting Models to Biological Data Using Linear and Nonlinear Regression*. Oxford University Press, Oxford, UK (2004)
11. Rasmussen RA. Atmospheric trace gases in Antarctica. *Science*, **211**, 285–287 (1981)
12. Richards S. Testing ecological theory using the information-theoretic approach: Examples and cautionary results. *Ecology*, **86**, 2805–2814 (2005)
13. University of Tennessee. Across Trophic Level System Simulation (1996). http://atlss.org Cited in Nov 2012
14. Wilkinson GN. Statistical estimations in enzyme kinetics. *Biochem J.*, **80**, 324–332 (1961)

http://www.springer.com/978-1-4614-7275-9

Mathematics for the Life Sciences
Calculus, Modeling, Probability, and Dynamical Systems
Ledder, G.
2013, XX, 431 p. 109 illus., Hardcover
ISBN: 978-1-4614-7275-9