MODELS AND DATA

There are a lot of connections between mathematics and biology, yet most students—and even most mathematicians and biologists—are unaware of these connections. One reason for this is that neither the historical development nor the pedagogical introduction of either subject involves the other. Biology grew out of natural philosophy, which was entirely descriptive. Modern biology curricula generally begin with descriptive biology, either organismal or cellular. The mathematically-rich areas of genetics and ecology make their appearance in advanced courses, after students have come to see biology as a non-mathematical subject. Historically, the development of calculus and calculus-based mathematics was driven by the mathematical needs of physics, and it remains standard practice to use physics to motivate calculus-based mathematics, whenever such motivation is deemed necessary. Other areas of mathematics, such as game theory and difference equations[1], were motivated to some extent by biology, but these topics appear in specialized courses generally taken by mathematics students only. Probability is another mathematical topic with strong connections to biology; however, probability is generally encountered in statistics courses, which usually emphasize social science applications.

In our discussion, we will focus on the nature of the connections, wherever they might be present, rather than the topics in which connections are most easily found. This way, we will be able to identify for ourselves the areas in biology that are amenable to mathematical treatment.

Like all scientists, biologists collect data. Useful results usually come from the analysis of data, rather than from the data itself. The presence of data provides an opportunity to use mathematics in the service of biology. Analysis of data is generally labeled as statistics, but it can also be thought of as mathematics. Statistics involves a combination of careful mathematics with judgment based on experience, and we will examine some useful topics in statistics that are mathematical in nature.

A two-sided connection between mathematics and biology comes from the use of mathematical models. Science is more than just collection and analysis of data. Science is also a search for general principles. These principles have to be supported by the data, but they also benefit from support in the form of a consistent mathematical model that ties together a variety of observations and/or experiments. Mathematical modeling in biology benefits biology by providing theoretical results that can be used to suggest new researches, make predictions, and provide a mathematical framework for the analysis of data. Mathematical modeling in biology also benefits mathematics because biological models often contain interesting mathematics.

These notes describe the way models and data connect mathematics with biology. We begin by discussing the use of mathematics to characterize collections of data. Then we examine some mathematical models to get an idea of the general properties of models and to obtain some experience in understanding specific models. In Section 3, we use calculus to solve the mathematical problem of fitting a simple linear model to data. This work is extended in Section 4 to the problem of fitting simple nonlinear models to data.

[1]Leonardo of Pisa, more commonly known as Fibonacci, developed his famous sequence as the solution of a difference equation modeling population growth.

1 Characterizing Data

Scientific research is concerned with asking questions about the natural and physical world. The questions are asked by making observations or conducting experiments, and the results of these observations and experiments are generally in the form of data. The characterization of data sets is the subject of **descriptive statistics**, whose basic elements we consider here. We begin with the classification of data into different types, continue with a study of how to display data for visualization, and conclude with the quantitative characterization of data.

Types of data

Data comes in many varieties, and we begin with a description of some of the terms used to classify data. Different classes of data need to be treated differently, so a first step in data analysis is data classification. A nonstandard, but highly useful, was to classify data is to begin with two large categories, **arithmetic data** and **non-arithmetic data**.

DEFINITION Data is **arithmetic**[2] if it consists of numerical values that can be averaged in a meaningful way, and **non-arithmetic** otherwise.

Non-arithmetic data can be further subdivided into several classes:

categorical or nominal data data consisting of descriptive terms, such as "male" and "female," or "married," "single," "divorced," and "widowed," or "American," "English," "French,", "German," *etc.* Some authors use "nominal" when the descriptives are assigned an arbitrary number and "categorical" otherwise, but really there is no difference. The order of the categories is arbitrary.

ordinal data numerical data in which the ordering of categories is significant, but the distance between the category descriptions is ambiguous. An example of ordinal data is the data from a survey in which respondents are asked to rate a statement with an integer from 1 through 5, with 1 for "strongly agree" and 5 for "strongly disagree." If one person gives a rating of 3 and another gives a rating of 1, we can *say* that the "average" rating is 2, but this average has no clear meaning in terms of the actual categories. There is no objective reason to think that the difference between "undecided" and "agree" is the same as that between "agree" and "strongly agree."

qualitative data data consisting of text, such as statements made in response to a prompt during an interview.

Statistics is a combination of objective mathematical modeling with subjective rules of thumb. Analysis of non-arithmetic data is necessarily subjective, and we leave such topics to the statisticians. In keeping with the theme of natural science, rather than social science, our interest in statistics will be restricted to the analysis of arithmetic data. One important note of caution, however, is that ratios of arithmetic data may or may not be arithmetic also. Ratios of populations are arithmetic if weighted averages are used. For example, if a population of 100 individuals is 50% male and a population of 50 individuals is 30% male, then we could

[2]The word "arithmetic" used here is the adjective, pronounced "eh-rith-MEH-tic," not the noun of the same spelling that is pronounced "a-RITH-muh-tic."

determine the average percentage of males, although the result is not 40%. Similarly, a wage of \$12.00 per hour is twice as much as a wage of \$6.00 per hour. However, ratios of temperatures are not meaningful; 100°F is not twice as hot as 50°F.

Arithmetic data can be further categorized according to relationships with other variables:

independent data data that does not depend on anything except random variation;

univariate data data in which the values depend on an arithmetic independent variable;

multivariate data data in which the values depend on more than one independent variable.

Any of these kinds of data may also be **stratified**, meaning that the data are split into several different subsets according to a categorical or nominal variable.[3] For example, a data set on the height of adult Americans might be subdivided into male and female subsets. We consider independent data, stratified or not, in the remainder of this section. Univariate data is treated in Sections 3 and 4.

Displaying data

Suppose we want to know how important height is as a predictor of success in women's volleyball. We randomly divide a women's volleyball class into teams of ten women each, six of whom will play at a time with a regular pattern of substitution to rotate the non-starting players. We measure the heights of all the players and consider the data for each team separately. Table 1 shows the heights, in centimeters, for the players on the Spikers team.

Number	1	2	3	4	5	6	7	8	9	10
Height, X	151.4	159.0	148.7	159.0	156.4	155.1	163.6	158.1	156.6	160.3

Table 1.1: Heights, in cm, of the ten players on the Spiker volleyball team

In many areas of mathematics, we must distinguish between the *discrete* and the *continuous*. **Discrete** mathematics has to do with a collection of distinct equally-spaced values, while **continuous** mathematics deals with intervals of real numbers. The distinction between discrete and continuous is a theoretical one. In theory, the height of a person could be any real number within a realistic range, so height should be a continuous random variable. In practice, however, we can only measure a quantity to some standard of precision. It is impossible, for example, to measure a line to be of length $\sqrt{2}$. We might want to use a *model* that treats line length or human height as a continuously varying quantity, but the actual data is always discrete. In Table 1.1, the heights are given to the nearest millimeter.

How should we make sense of the data in Table 1.1? The first important point is that the data is independent, so there is no significance to the order that the heights are listed in. Think of the number in the top row as the arbitrary order in which the players lined up to be measured or the numbers on the players' jerseys. We could just as easily have given the data as a list: 151.4, 159.0, 148.7, ... We could also have sorted the data to present it in ascending or descending order.

[3]The various subsets of stratified independent data are treated in the same way as non-stratified independent data. The issue of comparing stratified data belongs in *inferential*, rather than descriptive, statistics.

The primary difficulty in understanding a mass of data is that the crucial information is overwhelmed by the detail. To see the crucial information, we need a way to filter out the detail. This can be done by breaking up the interval over which the data varies into equal subintervals, called **classes**, and then use the class of each value instead of the value itself. Suppose we use the following classes:

$$[148.0, 150.0), [150.0, 152.0), [152.0, 154.0), \ldots, [162.0, 164.0).$$

We can name these classes "149," "151," ..., "163." In effect, we are rounding the original data off to the nearest two centimeters.[4] With the data grouped by class, we can display it as a table that gives the number of occurrences, or **frequency**, of each class value. The result is a **frequency table** (Table 1.2).

Height	149	151	153	155	157	159	161	163
Frequency	1	1	0	1	2	3	1	1

Table 1.2: Heights, in cm, of the ten players on the Spiker volleyball team

The obvious advantage of Table 1.2 over Table 1.1 is that the data looks simpler. Another important advantage is that the data of Table 1.2 can be meaningfully plotted, using the height classes on the horizontal axis and the frequencies on the vertical axis. While it would be reasonably good to plot these numbers as points, it is best to plot them as a bar graph. A bar graph in which the heights of the bars give the frequencies of the data classes is called a **histogram**. Figure 1.1 is a histogram for the volleyball team data.

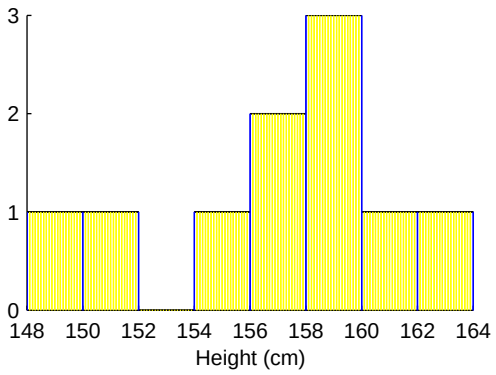

Figure 1.1: Heights of Spiker team players

Measures of central tendency

Suppose you want to summarize the data for the Spikers with just a single number that represents the "average" team height. There are several different ways to do this, but we consider here only two of them: the mean and the median.

[4] If the data set contained 100 values instead of 10, we might prefer rounding off to the nearest one instead of two. If measurements of 100 people were made to the nearest centimeter, we might choose to make each distinct value its own class.

DEFINITION The **mean** μ of a set of data is the arithmetic average of the data values. The median of a set of $2n-1$ data values, ordered by magnitude, is the nth value; the median of a set of $2n$ values is the average of the nth value and the $(n+1)$st value.

Example 1.1
The mean of the Spiker heights is

$$\frac{151.4+159.0+148.7+159.0+156.4+155.1+163.6+158.1+156.6+160.3}{10} = 156.82 \approx 156.8.$$

The fifth and sixth tallest players have heights 158.1 and 156.6, so the median height is

$$\frac{158.1+156.6}{2} = 157.35.$$

$\diamond$

The choice of median or mean depends on the purpose of the data. The mean is the arithmetic average of the values; we can think of the median as the value of the average individual. Consider a set of data that gives the yearly salaries of a major league baseball team. The owner has to pay all of the salaries, so he is most concerned about the mean of the salaries. The players' union represents the views of the players as individuals, so it is most likely concerned with the median salary. This difference of viewpoint helps explain some of the perceptual differences in labor negotiations—salaries in any organization usually include a small number of larger salaries and a large number of smaller salaries, a case for which the mean is larger than the median.

Another issue to consider in measuring central tendency is the importance of **outliers**, which are values that seem rather far from the majority. The mean of a set of data is highly sensitive to the inclusion of outliers, while the median is largely insensitive. In the Spiker height data, one could view the two smallest heights as outliers. Their combined effect is to depress the mean without affecting the median. Replace those two players with players 5 cm taller, and you will leave the median unchanged while raising the mean by 1 cm. This is why instructors should report median exam grades rather than mean exam grades—a single student getting 10 out of 100 on an exam can exert a significant influence on the mean. For scientific work, it is generally important to work with the mean rather than the median, and this may require the scientist to make a decision whether to include or omit outliers. The decision can be based on probability calculations, which need to be deferred for now.

Variability of data

To characterize the variability of a set of data, we need to look at the differences between the data values and the mean. Table 1.3 extends the original data table to include these deviations.

Number	1	2	3	4	5	6	7	8	9	10
Height, X	151.4	159.0	148.7	159.0	156.4	155.1	163.6	158.1	156.6	160.3
Deviations, $X-\mu$	-5.42	2.18	-8.12	2.18	-0.42	-1.72	6.78	1.28	-0.22	3.48

Table 1.3: Heights and deviations from the mean for the Spiker height data

For reasons that only become clear after learning a lot more probability and statistics, the appropriate average of the deviations is the standard deviation, defined as follows:

DEFINITION The **standard deviation** σ of a set of n X-values is given by

$$\sigma = \sqrt{\frac{(X_1 - \mu)^2 + (X_2 - \mu)^2 + \cdots + (X_n - \mu)^2}{n-1}}.$$

Example 1.2
The standard deviation of the Spikers height data is

$$\sqrt{\frac{(-5.42)^2 + (2.18)^2 + (-8.12)^2 + \cdots + (3.48)^2}{9}} \approx 4.3.$$

$\diamond$

Ratios of standard deviations are meaningful; if the heights of the Setters volleyball team has standard deviation twice that of the Spikers, then it makes sense to say that the Setters data is twice as variable as the Spikers data. If the data are normally distributed[5], then about 68% of the data lies within one standard deviation of the mean, but this may not be true for small samples or histograms that are not approximately normal.

Samples and populations

In most scientific experiments, we want to determine the characteristics of some population but are unable to study all the individuals in the population. We want to know how tall adult Americans are, but we cannot measure all American adults. We want to determine how many people have HIV, but we can get records only for people who have volunteered to be tested. We want to determine how much oxygen a particular Olympic swimmer uses in 100 m training swims, but our equipment is only available at one pool.[6]

In practice, we must determine the mean $\bar{x}$ and standard deviation s for a *sample* and then approximate the mean μ and standard deviation σ of the *population* by

$$\mu \approx \bar{x}, \qquad \sigma \approx s.$$

Uncertainty in results about sample characteristics is caused only by any uncertainty in the measurement of the data. There are additional sources of uncertainty in the use of sample characteristics to approximate population characteristics.

The educational-television math program called "Square One" had a short cartoon vignette in which two children opened up a CD store to sell recordings of violoncello music. The store failed because it did not have enough customers. The children told the mathematically-adept hero of the cartoon that they paid a consulting firm to do market research and got the report that "2 out of 3 music store customers prefer 'cello music." It turned out that the consultants had gone to a store specializing in classical music and interviewed exactly three customers, two of whom were 'cellists. In biology as well as social science, a set of data is only as good as the

[5] A normal distribution has a "bell-shaped" histogram

[6] The "population" in this case consists of the different 100 m swims for the athlete.

process used to collect it, a process that includes sampling procedures as well as laboratory or field procedures.

When we can't collect data for an entire population, we collect data for a *sample* of the population and then assume that the sample is *representative* of the population. Sampling introduces two types of uncertainty into measurement of population characteristics, and both are illustrated by the music store vignette. First, the sample represents the population of classical music listeners rather than the population of all music listeners, which is the population that was supposedly being studied. Second, the sample is so small that the amount of uncertainty in its correspondence with the population it represents makes the results meaningless. If the study included hundreds of music listeners at a single classical music store, then the sample would adequately represent the customers of that one store. Perhaps it also is representative of classical listeners in general, or perhaps there are differences among classical listeners in different cities or of different economic groups.

In doing an experiment, we collect a sample from a population. The sample is **random** if each member of the population has an equal chance of being selected. A non-random sample can still be representative if there is no systematic difference among portions of the population. In chemistry, for example, we take a sample of pure hydrochloric acid by pouring some out of a bottle. This is not a random sample, because we got all of the molecules out of the same bottle. However, it is a representative sample because there is no systematic difference among bottles of pure hydrochloric acid. In an experiment on living creatures, we take a random sample of some subpopulation of a larger population. For example, if we are studying bison, we might restrict our sample to the subpopulation of bison living in Custer State Park in South Dakota. We need to employ a well-designed procedure to obtain a random sample of the subpopulation, and then we need to be concerned about the possibility of systematic differences between bison who live in Custer State Park and those who live elsewhere.

Even with a random sample, there is still error introduced by sampling. The very nature of randomness means that different samples from a population will have different means. We should expect sample characteristics to be approximately, but not exactly, equal to the characteristics of the overall population.[7] Larger samples should show less difference from the overall population than smaller samples. If the consultants in the Square One vignette had sampled 300 customers, they almost certainly would not have found 200 who prefer 'cello music to other classical music.[8]

In summary, we should anticipate three potential sources of error whenever we use data obtained from a sample to determine characteristics of a larger population: uncertainty in measurement of the data, uncertainty caused by using a random sample, and error caused by poor representation of the population by the sample. Measurement uncertainty must be studied in the context of the specific experiment or observation being used. Representation error is minimized by using appropriate sampling procedures. Both of these issues are important, but beyond the scope of mathematics and mathematical modeling. Sampling uncertainty can be studied mathematically, but only after a thorough background in elementary probability; we will return to this topic later.

[7]The data in Table 1 was obtained by randomly selecting values from a large population having mean 155.4 and standard deviation 7.2.

[8]I am not bad-mouthing 'cello music, which I particularly like. Given the variety of classical music choices, we should not expect any one choice to be preferred by a large percentage of listeners.

2 Mathematical Models

In your previous mathematics experience, you have undoubtedly solved word problems of two types. A small number of word problems are purely mathematical (If Arthur is twice as old as Betty, ...). Most word problems use mathematical models.

Example 2.1
"A train leaves point A at 12:00 and travels toward point B at a speed of 60 km/hr. A second train leaves point B at 1:00 and travels toward point A at a speed of 70 km/hr. The trains meet at 2:00. How far is point B from point A?" This problem is based on a mathematical model because the narrative is impossible as written. A train cannot suddenly start at 60 km/hr from a standing start, nor can it maintain a constant speed when traveling over real terrain and through varied country. The mathematical problem is not about real trains, but rather about theoretical trains, and it is based on a model of trains that can accelerate instantly from a standing start to a final speed and can maintain a constant speed. ◇

Our interest here is in understanding the *idea* of mathematical models, and so we will examine separately the connection between mathematical models and the real situations that inspire them. While not a formal mathematical definition, the following description serves as a working definition of "mathematical model."

DEFINITION A **mathematical model** is 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 resemble the real behavior on which it is based.

This definition fits the train problem nicely. We approximate real train motion as motion at a constant speed and we hope that the result we get will be accurate enough to be useful. Intuitively, it should be very accurate, provided the speeds assumed in the model are close to the *average* speeds of the real trains.

"Spherical turkey" assumptions

Perhaps you have heard a joke that is sometimes told to make fun of mathematics. The joke is about determining how long to cook a turkey, and the punchline is that the mathematician's answer to the question begins with "Assume a spherical turkey." As a mathematical modeler, I am not the least embarrassed by this answer. I will indeed solve the problem by assuming a spherical turkey, and I will interpret the laughter as a sign of failure to understand the value of mathematical modeling. The point is this: a turkey's shape is much closer to spherical than it is to cylindrical or cubic. The equations of heat flow are much simpler when applied to these simple geometric shapes than to arbitrary shapes. By assuming a spherical shape, I can get a solution quickly by hand calculation. If I don't assume a spherical turkey, I will first have to spend a lot of effort coming up with formulas to describe the real shape, and then I will have to write a computer program to approximately solve the heat flow equations on this complicated shape. Because the minor variation in distance of surface points from the center plays only a small role in the heat flow processes, the answers I get by the two methods will be within a few minutes of each other. While not a true mathematical definition, we have a useful working concept.

DEFINITION A **spherical turkey** assumption is an assumption that seems foolish, but simplifies a problem without noticeably affecting the solution.

The ignorant may laugh at spherical turkey assumptions, but the enlightened will profit by them.

Conceptual models

One commonly reads in mathematics books a statement to the effect that mathematical models are descriptions of a real physical setting. This is giving mathematical models a loftier status than they deserve. At best, they represent an oversimplified characterization of a real physical setting. They are like expert witnesses testifying in court. As the lawyer for one side, we want to understand and use the evidence they present; as the lawyer for the other side, we want to ask them difficult questions in an attempt to discredit them; as the jury, we want to understand their testimony well enough to render a verdict that is beyond a reasonable doubt.

How do we create mathematical models? According to our working definition, the models should be based on an "approximate quantitative description" of real phenomena. It is helpful to separate out the "approximate" from the "quantitative description." The approximate part does not need to be mathematical, and the quantitative description does not need to be approximate. The separation is made by using the idea of a conceptual model.

DEFINITION A **conceptual model** is a verbal description of an idealized physical setting that is simple enough to be converted into a mathematical model.

Consider again the problem of cooking a turkey. My conceptual model for the problem is that of a solid sphere of turkey-flesh that is heated from the outside by thermal radiation. The conceptual model simplifies the real problem by making the shape spherical, ignoring the distinction between meat, fat, skin, and bones, ignoring the geometry of the heating coils in the oven, and neglecting the influence of the pan or rack that holds the turkey. The resulting conceptual model is simple enough that I can convert it to a mathematical model. The result will only be as accurate as the weakest assumption, which is probably the neglect of the roasting pan. The assumption of a spherical shape is probably the least worrisome of the lot. As a mathematical modeler, I will derive and solve the math problem corresponding to this conceptual modeler. I will then give the result to an experimental scientist (the cook) to use in preparing a turkey dinner. If the experiment result agrees closely enough with the model result that the meal is a success, then there is no need to use a more complicated model. If the turkey turns out to be overdone or underdone, then I need to try again with a more complicated conceptual model. My first thought would be to include the effect of the roasting pan. If that doesn't yield a result that agrees with the experiment, then I'll have to rethink the model. When a mathematical model yields a bad result, it is not always a failure. Replacing a conceptual model that doesn't work with one that does work is the heart of scientific progress. Theoretical science is not about understanding the world; it is about developing useful conceptual models of the real world.

Example 2.2
Careful measurements of the movement of the planet Mercury contradicted results obtained using Newtonian mechanics. Eventually, good agreement with observation was obtained using Einstein's theory of general relativity, which was based on a non-Newtonian conceptual model. This did not

eliminate Newtonian mechanics as a useful subject in science; rather, it refined Newtonian mechanics by placing a limit on the settings in which it could be used. ◇

Standard models

The two most commonly used mathematical models are the linear model,

$$y = b + mx, \tag{2.1}$$

and the exponential model,

$$y = Ae^{kx}. \tag{2.2}$$

Both of these models involve an independent variable x and a dependent variable y. The other quantities in the models are parameters. Note that the choice of symbols for the quantities is irrelevant. The models $y = b + mx$ and $y = b + mt$ and $x = b + mt$ are all the same, mathematically, the difference being only in the choice of symbols for the variables.

DEFINITION An **independent variable** is a quantity that can be systematically varied, and a **dependent variable** is a quantity whose value is determined by the dependent variable. A **parameter** is a quantity that is constant, in the sense of being independent of the independent variable; it does not, however, have a single fixed value. A **fixed constant** is a number, like the circumference-diameter ratio π, that has a unique value which may or may not be known.

The distinction between fixed constants, parameters, and variables is vital to an understanding of mathematical models.

Fixed constants other than π and e are rare in mathematical models. Most constants are either integers, in which case we would use the integer rather than a letter, or else they represent quantities that do not have a single fixed value. Suppose, for example, we want to model motion at a constant speed of 2. If x is the position and t the time, then we have

$$x = 2t.$$

We could just as easily consider motion at a constant speed of 17, which would be

$$x = 17t.$$

It is silly to have a different model for motion at different speeds. It is much better to let the parameter v represent the constant speed and use the model

$$x = vt.$$

We can choose a value for v when a specific problem requires it, but we should leave it unspecified otherwise. Both the independent variable t and the parameter v are independent; the distinction is that the time t varies continuously through a range of values, whereas the parameter v takes one fixed value for all t in one instance of the model and a different fixed value for all t in the next instance. To plot the model $x = vt$, we draw several graphs of x vs t, each with its own value of v. The linear model $y = b+mx$ has two parameters, m and b, which represent the slope and the y intercept of the straight line. Each particular instance of the model is obtained by

choosing values of m and b, and then we can graph the model corresponding to those particular values.

Observe that the linear model is the simplest 2-parameter model for which the derivative dy/dx is constant. The exponential model has a similar property.

DEFINITION The **absolute rate of change** of y with respect to t is $\dfrac{dy}{dt}$. The **relative rate of change** of y with respect to t is $\dfrac{1}{y}\dfrac{dy}{dt}$.

For the exponential model $y = Ae^{kt}$, the relative rate of change is

$$\frac{1}{y}\frac{dy}{dt} = \frac{1}{Ae^{-kt}}\left(kAe^{-kt}\right) = k.$$

The exponential model is the simplest 2-parameter model for which the *relative* rate of change is constant.

Empirical models

There are different ways to obtain formulas for mathematical models, and the respect we should grant to the model depends on how it was obtained. Sometimes models are obtained from simple assumptions that seem plausible. It may be reasonable in some circumstances to assume that a quantity has a constant absolute rate of change, leading to a linear model, or a constant relative rate of change, leading to an exponential model. The Monod growth model, which we consider shortly, is a more complicated biological model that is also derived from basic assumptions. Unfortunately, basic principles are not as common in biology as they are in the physical sciences, and so there are many cases in which we simply don't know enough about the natural process to develop a model from plausible assumptions. In these cases, the best we can do is to choose a model that has the right general shape. With enough parameters, almost any model can be made to fit almost any data; the model will be capable of generating a reasonable graph, although its scientific value is probably minimal.

DEFINITION An **empirical model** is a model chosen because it has the right qualitative features to fit the data, even though it cannot be derived from reasonable assumptions nor supported by theoretical arguments.

Empirical models of the form

$$y = ax^p, \tag{2.3}$$

with p a real number, are common in biology, for reasons that will become clear in Section 4.

The Monod growth function

In 1949, Jacques Monod introduced a conceptually-based mathematical model for bacterial growth in cultures.[9] Let S be the concentration of substrate (food) in the culture and let r be the rate of food intake of a microorganism. Our aim is to determine a function $r(S)$ that models the dependence of the food intake rate of a microorganism on the amount of food available. It

[9] J. Monod, The growth of bacterial cultures, *Annual Review of Microbiology*, 3, 371–394, 1949.

is almost always helpful to keep track of the dimensions of the quantities in a model. Here, S is measured in units of food per area and r in units of food per time.

Instead of thinking about microorganisms, it might help to do a thought experiment in which you put yourself in the role of the microorganism. Imagine yourself at a buffet restaurant where individual servings of food are set out on small tables scattered through the room. Each time you eat a serving of food, a restaurant worker sets out another serving in some randomly chosen location within the room. This last assumption is necessary to keep the food concentration S at a constant level. How fast can you eat? Assume that you never get full, because you can always split into two if you get too big (remember that you are playing the role of a microorganism).

A first try Most likely, there are no servings immediately within your reach; you walk to a serving, eat the food, and then walk to another serving. Think of yourself as being very nearsighted.[10] Instead of walking directly to a serving of food, you have to wander around until you get close enough to see the food. If you walk really slowly, you will obviously eat less than if you walk fast, so let's define a parameter a that measures the area you can search per unit time. Now, clearly your ability to consume food will be better if there is more food available and if you search faster. A good first guess is that your consumption rate is given by

$$r = aS.$$

This means that doubling the concentration of food or your search speed doubles your consumption rate. This seems very reasonable, and it makes mathematical sense too. Multiplying servings per unit area by area per unit time will give a result of servings per unit time.

Does the simple formula $r = aS$ make sense? It seems pretty good, considering the description of the procedure. But we want a formula that works for all values of S. Suppose the tables are set out so close together that you can reach one from any point in the room. Now the search process is almost unnecessary, so your search speed shouldn't matter. Moreover, even with an infinite stomach capacity, there is a limit to how much food you can consume in an hour.[11]

A better model The problem with our first try is that it is based on the assumption that all of our time can be spent in searching for food. This is only true if it doesn't take any time to eat the food. A better model would begin with the formula

$$r = f \cdot aS,$$

where f is the fraction of the time that is available for searching. Consider that we have two activities, searching and eating. The amount of time spent eating depends on the amount of food that we have collected, which depends on the rate r. Suppose you are capable of eating c units of food in one unit of time. With r as the food intake per *total* time and c the food intake per *eating* time, the ratio r/c is the ratio of eating time per total time. Thus, the fraction of time you spend searching, rather than eating, is $f = 1 - r/c$. Thus, we have

$$r = \left(1 - \frac{r}{c}\right) aS.$$

[10] This makes our thought experiment more like nature, in which finding food is a nontrivial matter.

[11] It is the combination of stomach size and eating speed that determine your performance in the annual world championship hot dog-eating contest. A contestant with an infinite stomach and a small mouth is not going to win.

This equation relates the unknown quantity r to the independent variable S and the two parameters a and c. We can solve it algebraically for r, and we obtain the solution

$$r = \frac{aS}{1 + \frac{a}{c}S}. \tag{2.4}$$

This formula has some pleasing properties. Note what happens in the limits as $S \to 0$ and $S \to \infty$. As $S \to 0$, the formula simplifies to $r = aS$, which is what we originally got before we started worrying about the effect of an overabundance of food. If food is in short supply, $r = aS$ is a reasonable simplification. As $S \to \infty$, the second term dominates the denominator, and we have $r \approx aS/(aS/c) = c$. If food is overly abundant, $r = c$ is a reasonable approximation. In this case, we spend very little time searching and are able to eat at our maximum capacity. Equation 2.4 is one version of the Monod growth function.

Note that we are not saying that Equation 2.4 is the correct quantitative description of the consumption of food by microorganisms. We are saying that it is the correct quantitative description of a *conceptual model* in which organisms can eat food at a maximum rate of c and spend all of their time either eating food or searching for it. Equation 2.4 is only as good as the conceptual model that it corresponds to. This is an issue to be determined by experiment.

The Monod growth function is generally written in a different form. If we multiply the numerator and denominator of the right hand side by c/a, we obtain

$$r = \frac{cS}{\frac{c}{a} + S}.$$

Now define $K = c/a$, and we obtain the alternative form

$$r = \frac{cS}{K + S}. \tag{2.5}$$

Equations 2.4 and 2.5 are equivalent versions of the same model. They differ in the specification of the parameters. The parameter K is called the *semi-saturation constant*. The name stems from the observation that when we take $S = K$, we get

$$r(K) = \frac{cK}{K + K} = \frac{c}{2}.$$

Thus, the food intake when $S = K$ is half of that when $S \to \infty$. One advantage of Equation 2.5 for the Monod growth function is that both parameters c and K indicate directly-observable features of the graph of the function (see Figure 2.1).

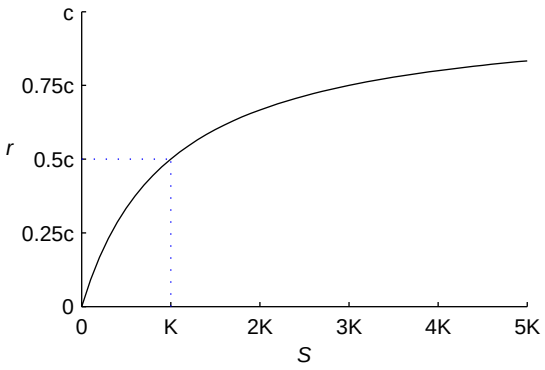

Figure 2.1: The Monod growth function

3 Linear Regression

In a 1977[12] paper, Joe Kiceniuk and David R. Jones reported on their study of the oxygen transport system in trout.[13] Their data included measurements of the volume of oxygen consumed (V) and the rate of blood flow (Q), as seen in Table 3.1 and plotted in Figure 3.1.

Q	17.6	28.4	34.8	42.9	52.6
V	25.0	67.8	84.8	139.3	193.7

Table 3.1: Data for blood flow rate (Q) and oxygen consumption (V) in trout

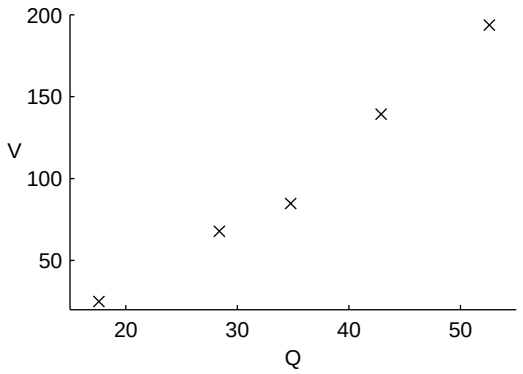

Figure 3.1: Oxygen consumption V as a function of blood flow rate Q in trout

The quantities Q and V seem to lie close to a straight line. Indeed, theoretical arguments suggest that the graph should be linear. Since measurement is never exact, it makes sense to determine the equation for the straight line that best fits the data, and then use that function, rather than the data itself, for further calculations. In mathematical terms, we have an example of the following general problem.

Problem 1 *Given the data points (x_1, y_1), (x_2, y_2), ..., (x_n, y_n), find the numbers b and m that give the "best fit" of the linear model $y = b + mx$ to the data.*[14]

In the context of this general problem, b and m are variables that select properties of the graph of y vs x, and we are looking at the set of all functions of the form $y = b + mx$. It makes sense to think of b and m as independent variables and think of properties of the graph as dependent variables.

[12]Lest you think this is old stuff, I note that an article appearing in 2005 in the Journal of Experimental Biology cited Kiceniuk and Jones. The 2005 article established a clear connection between swimming performance and heart health in trout. The authors discovered that the common practice (among fish farmers) of breeding trout to encourage more fat also encourages a specific heart defect, leading to larger numbers of fish not healthy enough to make it to harvest time.

[13]Kiceniuk, J. and D.R. Jones, The oxygen transport system in trout (*Salmo gairdneri*) during sustained exercise, *Journal of Experimental Biology*, 69: 247-260, 1977.

[14]The symbols Q and V to represent the data are arbitrary and have been chosen to agree with notation in physiology textbooks. Any planar graph can always be thought of as having axes called x and y, regardless of the symbols that are used for a specific problem. In our example, thought of as an instance of Problem 1, we have $n = 5$, $x_1 = 17.6$, $y_1 = 25.0$, and so on.

Example 3.1
For example, let a be the x intercept of the graph of $y = b + mx$; thus, a is a property of the graph. Substitute the point $(a, 0)$, which is the definition of a as the x intercept, into the model $y = b + mx$. We get

$$0 = b + ma.$$

This equation determines the graph property a for known b and m, so we solve it for a and obtain

$$a = -\frac{b}{m}.$$

Hence, a is a function of b and m. ◇

Simplifying the problem

Problem 1 is not a complete math problem because we have not given a precise definition to the phrase "best fit." Before we do that, it is convenient to change the problem to one that is a little bit simpler. Let $\bar{x}$ and $\bar{y}$ be the average values of x and y. Thus,

$$\bar{x} = \frac{\sum_{k=1}^{n} x_k}{n}, \qquad \bar{y} = \frac{\sum_{k=1}^{n} y_k}{n}. \tag{3.1}$$

Example 3.2
For the data of Table 3.1, using x for Q and y for V, we have

$$\bar{x} = \frac{17.6 + 28.4 + 34.8 + 42.9 + 52.6}{5} = 35.26, \qquad \bar{y} = \cdots = 102.12.$$

◇

Now define new variables X and Y by

$$X = x - \bar{x}, \qquad Y = y - \bar{y}. \tag{3.2}$$

We can compute the values of X and Y for the data of Table 3.1. Table 3.2 contains the same data in terms of the new variables. The data is illustrated in Figure 3.2.

X	-17.66	-6.86	-0.46	7.64	17.34
Y	-77.12	-34.32	-17.32	37.18	91.58

Table 3.2: The example data, in terms of X and Y

Let's now focus on finding the straight line that best fits the XY data, postponing consideration of the original problem. Note that the XY data is centered about the origin, in the sense that the average X value and the average Y value are both 0. It seems reasonable to include as part of the definition of "best fit" the requirement that the line should go through the point $(X, Y) = (0, 0)$. This requirement means that we will restrict consideration of possible best fit lines to those that are of the form $Y = mX$. We now have a problem with only one unknown parameter.

Problem 2 *Given the data points (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n), find the number m that gives the "best fit" of the linear model $Y = mX$ to the data.*

The plot of Figure 3.2 shows the lines with $m = 4.5$ and $m = 5.3$. The best value of m appears to lie somewhere between these two values.

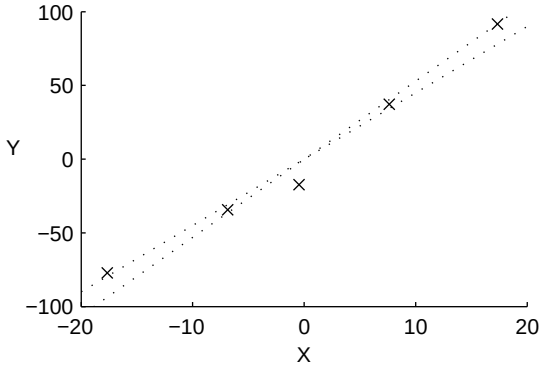

Figure 3.2: The example data, along with the lines $Y = 4.5X$ and $Y = 5.3X$

Defining "best fit"

Think about the problem from a graphical point of view. On a graph, we plot the points (X_k, Y_k). These points are scattered about, but we can still get a sense of where the best fit line should be. We have assumed that it should go through the origin, but the slope m is not known. To convert Problem 2 into a well-posed mathematical problem, we need to define a function $f(m)$ that measures the deviation of the data from the line $Y = mX$ to the data. Note that f is a function of m alone, since m is the only quantity that we are free to choose. The formula for $f(m)$ will make use of the fixed constants X_k and Y_k.

Let ΔY_k be the vertical distance to the point (X_k, Y_k) from the point (X_k, mX_k), which is either directly below it or directly above it. Then

$$\Delta Y_k = Y_k - mX_k.$$

Example 3.3
Let m=2.5. For the example data, with $k = 1$, we have $X_1 = -17.66$ and $Y_1 = -77.12$. The corresponding point on the line $Y = mX$ has $Y = mX_1 = -17.66m = -44.15$. Thus, $\Delta Y_1 = Y_1 - mX_1 = -32.97$. Similarly, $\Delta Y_2 = Y_2 - mX_2 = -34.32 - 2.5(-6.86) = -17.17$ and $\Delta Y_3 = -16.17$, $\Delta Y_4 = 18.08$, and $\Delta Y_5 = 48.23$. The line segments corresponding to these distances are shown in Figure 3.3. ◇

Each ΔY_k indicates the deviation of the data from the line for one point. The deviation function that best accounts for the total deviation is obtained by adding the squares of the deviations. We define the function

$$f(m) = (Y_1 - mX_1)^2 + (Y_2 - mX_2)^2 + \cdots + (Y_n - mX_n)^2. \tag{3.3}$$

Example 3.4
In our example,

$$f(2.5) = (-32.97)^2 + \cdots + (48.23)^2 = 4296.3.$$

In comparison to this value, we can determine $f(4.5) = 441.4$ and $f(5.3) = 508.2$. ◇

Now that we have defined a measure of how well a line fits a set of data, the best fit problem becomes a standard optimization problem.

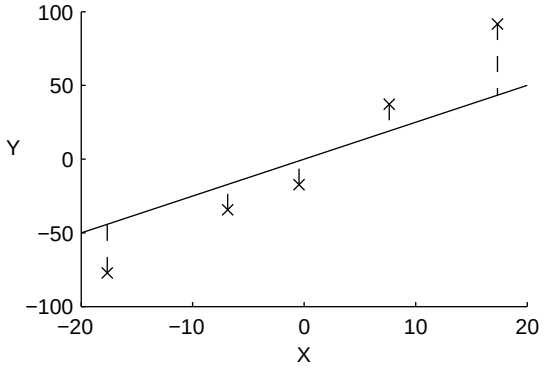

Figure 3.3: The example data, along with the line $Y = 2.5X$ and line segments showing the lengths of each ΔY_k

Problem 3 *Given the data points* (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n), *find the number* m *that gives the global minimum of the function*

$$f(m) = (Y_1 - mX_1)^2 + (Y_2 - mX_2)^2 + \cdots + (Y_n - mX_n)^2, \qquad -\infty < m < \infty.$$

The solution of the problem

Problem 3 asks for the global minimum of a function on an unbounded interval, so we need to be concerned about whether there actually is a solution. In this case, we have $\lim_{m\to\pm\infty} f = \infty$ and f is continuous. Taken together, these facts guarantee that there is a global minimum, and that it must occur at a critical point. Multiplying out the squares and collecting powers of m, we have

$$f(m) = (Y_1^2 + Y_2^2 + \cdots + Y_n^2) - 2m(X_1Y_1 + X_2Y_2 + \cdots + X_nY_n) + m^2(X_1^2 + X_2^2 + \cdots + X_n^2).$$

Then

$$f'(m) = 2m(X_1^2 + X_2^2 + \cdots + X_n^2) - 2(X_1Y_1 + X_2Y_2 + \cdots + X_nY_n).$$

Setting this derivative equal to 0, we find that there is a single critical point,

$$m = \frac{X_1Y_1 + X_2Y_2 + \cdots + X_nY_n}{X_1^2 + X_2^2 + \cdots + X_n^2} = \frac{\sum_{k=1}^n X_kY_k}{\sum_{k=1}^n X_k^2}, \tag{3.4}$$

which must be the global minimum.

Now that we have solved Problem 3, it is a simple matter to solve the original problem. We got X and Y from x and y using Equations 3.2. Then we found m, as given by Equation 3.4. The relationship between x and y is

$$y = Y + \bar{y} = \bar{y} + mX = \bar{y} + m(x - \bar{x}) = (\bar{y} - m\bar{x}) + mx.$$

The best fit for a line $y = b + mx$ to the original xy data is obtained by using the values

$$m = \frac{\sum_{k=1}^n X_kY_k}{\sum_{k=1}^n X_k^2}, \qquad b = \bar{y} - m\bar{x}. \tag{3.5}$$

Example 3.5
For our example data set, we have $\sum_{k=1}^{n} X_k Y_k = 3477.4$ and $\sum_{k=1}^{n} X_k^2 = 718.2$, leading to the best fit value $m = 4.84$. Using this value, we obtain $b = -68.6$. Figure 3.4 shows the linear regression line with the original data set. ◇

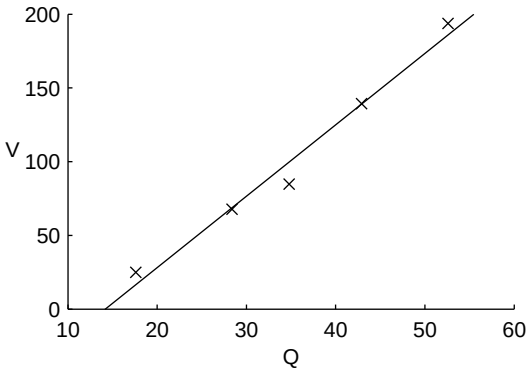

Figure 3.4: Oxygen consumption V as a function of blood flow rate Q in trout

Summary

We have now derived a procedure for finding the best fit line to a set of data points (x_k, y_k).

Algorithm 3.1 *To compute the numbers b and m for the best fit of a straight line $y = b + mx$ for data (x_k, y_k):*

1. *Compute the average values $\bar{x}$ and $\bar{y}$;*
2. *Create a modified table of data by replacing x and y by $X = x - \bar{x}$ and $Y = y - \bar{y}$;*
3. *Compute $\sum_{k=1}^{n} X_k Y_k$ and $\sum_{k=1}^{n} X_k^2$ and then*

$$m = \frac{\sum_{k=1}^{n} X_k Y_k}{\sum_{k=1}^{n} X_k^2}, \qquad b = \bar{y} - m\bar{x}.$$

The line obtained from this procedure is called the **linear regression** line. Any time you see a graph that includes a straight line derived from data, it is almost certainly the linear regression line.

An important warning

Not all relationships between variables are linear, but you can always obtain a linear regression line. This means that you have to use your judgement to decide if the linear regression line is really a good description of the data, or if instead the data indicates a nonlinear relationship. Table 3.3 shows a simple data set, and Figure 3.5 illustrates the data along with the linear regression line. Here x is the biomass of grass in a pasture and y is the rate of consumption of grass by a fixed population of cows. The correct graph is the dotted curve in the figure. The important point here is that normally the correct curve is not known, but can only be inferred

x	0	0.5	1.0	1.5	2.0	3.0
y	0	0.2	0.5	0.7	0.8	0.9

Table 3.3: An idealized data set for consumption of grass by cows

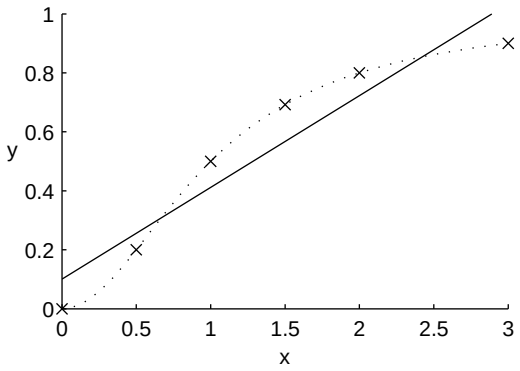

Figure 3.5: Consumption of grass by cows, showing the correct curve and the linear regression line

from the data. If we assume a linear relationship, we'll get a linear regression line. By itself, this doesn't count for anything. The result we get is only as good as the assumption of linearity. If the data doesn't fit the linear regression line really well, we might want to consider other assumptions about the relationship. We'll re-examine this data set in Section 4 and determine the correct relationship.

4 Fitting Nonlinear Models to Data

Very few real-world scenarios are simple enough to be well-approximated by a linear model; therefore, fitting models to data generally requires something beyond linear regression. There are two approaches that could be taken to nonlinear regression: one could try to find a different technique for each nonlinear curve, or one could try to adapt linear regression to nonlinear models. Either approach can be used, but the approach of adapting linear regression is more easily generalized and more appealing to mathematicians. The results obtained by the two approaches could be different, and variants of the two approaches can lead to results that are still different. It all comes down to the definition of the function measuring the deviation of a point from the desired curve. The parameter values obtained from a data set and a nonlinear model are determined by the subjective choice of deviation function as well as by the data set and model that comprise the mathematical problem.

Fitting exponential and power models

In Section 2, we explained the importance of the exponential model by noting that they follow from the elementary assumption of a constant relative rate of change. Accordingly, the exponential model is the most commonly used nonlinear model. As an example, suppose we are studying the biology of zebra mussels, an invasive species that is rapidly spreading in some North American lakes. We start a colony of zebra mussels in a laboratory. If there is enough food and space, it is reasonable to expect the population to satisfy an exponential growth model

$$N(t) = N_0 e^{kt}. \tag{4.1}$$

Our goal is to determine the growth parameter k. Presumably we could take N_0 to be the number of mussels at time 0, but we will instead take it as a second unknown parameter. A large difference between the computed value and the experiment history will indicate that the exponential model is not a good fit to the data. Suppose our measurements yield the following data:

t	0	3	5	8	10
N	20	28	36	50	62

Table 4.1: A data set for a zebra mussel population

We can't use linear regression for the nonlinear model, but we can change the model to a linear one. Taking natural logarithms on both sides of Equation 1 yields the equation

$$\ln N = \ln(N_0 e^{kt}) = \ln(N_0) + \ln(e^{kt}) = \ln(N_0) + kt.$$

Now define

$$y = \ln N, \qquad x = t, \qquad b = \ln(N_0), \qquad m = k, \tag{4.2}$$

which changes the model to the standard linear model $y = b + mx$. The procedure for the regression is essentially that of linear regression, with some minor changes. We first create a new table of data for the linear variables. Then we apply Algorithm 3.1 to find the parameters m and b. We finish by computing the parameters in the nonlinear model by solving the equations that define m and b.

Example 4.1
The data in Table 4.1 yields the data of Table 4.2. The algorithm yields $m = 0.1137$ and $b = 2.999$.

x	0	3	5	8	10
y	2.996	3.332	3.584	3.912	4.127

Table 4.2: The linearized data set for the zebra mussel population

Thus, $k = 0.1137$ and $N_0 = e^b = 20.1$. The fit of the parameters to the data is best seen on a plot of y vs x rather than N vs t. ◇

The power function model $v = au^p$ can also be fit using logarithms. Historically, it is more common to use the base ten logarithm rather than the natural logarithm; either works. We have

$$\log(v) = \log(a) + p\log(u), \qquad y = \log(v), \quad x = \log(u), \quad b = \log(a), \quad m = p.$$

Linearization in general

The general idea of linearization is that the variables and parameters in a model can be changed, as long as they are not mixed up. The variables have a known table of values, from which any function of the variables can be computed to use in place of the original variables. The parameters have specific, but unknown, values, from which any function of the parameters can be used as new parameters. Linearization works if the nonlinear model can be manipulated into the form

$$\text{function of variable(s)} = \text{function of parameter(s)} + \text{function of parameter(s)} * \text{function of variable(s)}. \tag{4.3}$$

Michaelis-Menton reaction velocity

Another common nonlinear model is that used for the so-called "reaction velocity" (rate at which product forms) in an enzyme-catalyzed biochemical reaction. There are many different enzymes, each of which catalyzes one or more reactions in a living organism. The reaction velocity v is an increasing function of the substrate concentration s. Since the reaction uses up some of the substrate, the velocity tends to decrease until the substrate is replenished. The model for enzyme kinetics is

$$v = \frac{v_{\max} s}{K_m + s}, \tag{4.4}$$

where $v_{\max}$ is the maximum reaction velocity and K_m is the substrate concentration at which the reaction speed is half of the maximum. Note that the model is identical to the Monod growth function! One of the exciting features of mathematics is that the same mathematical models sometimes arise from entirely different situations.

There are several ways to linearize the Michaelis-Menten velocity equation so that it can be fit to data. The most common is the Lineweaver-Burk linearization. Taking the reciprocal on both sides of Equation 4.4, we have

$$\frac{1}{v} = \frac{1}{v_{\max}} + \frac{K_m}{v_{\max}} \frac{1}{s}.$$

This is just the linear model, with $y = 1/v$, $x = 1/s$, $b = 1/v_{\max}$, and $m = K_m/v_{\max}$.