Functions from Data

One of the most important procedures in applied science is using experimental data to discover relationships between variables. In this module we will discuss some mathematical techniques for doing this, and we will use these ideas to investigate principles of planetary motion and the cooling of liquids.

Fitting Curves to Data

Suppose that a scientist is looking for a relationship between two variables x and y and that measurements of corresponding values of these variables have produced a set of n data points

$$(x_1, y_1),\quad (x_2, y_2),\quad (x_3, y_3), \ldots,\quad (x_n, y_n)$$

If the scientist uses the data in some way to obtain a relationship $y = f(x)$ between x and y, then this equation is called a ***mathematical model*** for the data.

One way to obtain a mathematical model for a set of data is to look for a function f whose graph passes through all of the data points; this is called an ***interpolating function***. Although interpolating functions are appropriate in certain situations, they do not adequately account for measurement errors in the data. For example, suppose that the relationship between x and y is known to be linear but that accuracy limitations in the measuring devices and random variations in experimental conditions produce a scatter plot such as that shown in Figure 1a. With the help of a computer, one can find a polynomial of degree 10 whose graph passes through all of the data points (Figure 1b). However, this polynomial model does not successfully convey the underlying linear relationship; a better approach is to look for a linear equation $y = mx + b$ whose graph more accurately describes the linear relationship, even if it does not pass through all (or any) of the data points (Figure 1c).

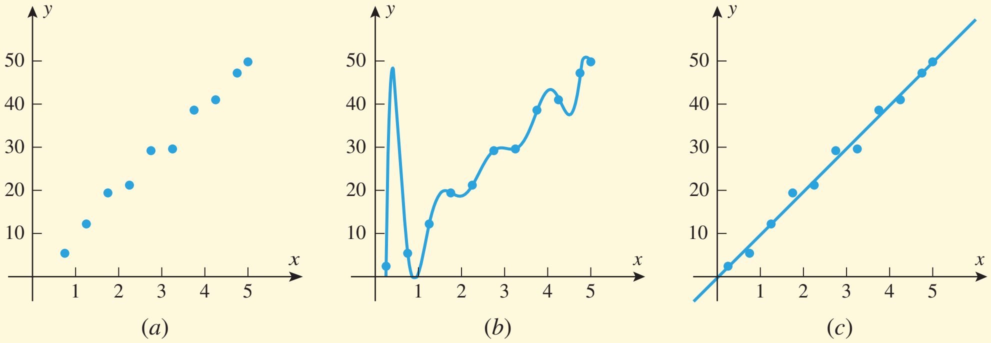

Figure 1

Finding Mathematical Models

The most challenging part of finding a model $y = f(x)$ for experimental data is coming up with an appropriate form for the function f. Sometimes the form of the function will be suggested by the visual appearance of the scatter plot, and sometimes it will be dictated by a known physical law that relates x and y. For example, Figure 1a strongly suggests that the relationship between x and y is linear, so in absence of additional information it would be natural to look for a linear model $y = mx + b$. In contrast, the scatter plot of U.S. population growth in Figure 2 strongly suggests some nonlinear relationship, so we must look for a nonlinear function for the model. The possibilities for nonlinear models are endless; however, there are theories in the study of population growth which suggest that in absence of environmental constraints, populations of people can be modeled over time by equations of the form $P = P_0e^{kt}$, so in this case we might look for an equation of this form to model the data.

Linear Models

The most important methods for finding linear models are based on the following idea: For any proposed linear model $y = mx + b$, draw a vertical connector from each data point (x_i, y_i) to the line, and consider the differences $y_i - y$ (Figure 3). These differences, which are called ***residuals***, may be viewed as "errors" that result when the line is used to model the data. Points above the line have positive errors, points below the line have negative errors, and points on the line have no error.

One way to choose a linear model is to look for a line $y = mx + b$ in which the sum of the residuals is zero, the logic being that this makes the positive and negative errors balance out. However, one can find examples where this procedure produces unacceptably poor models, so for reasons that we cannot discuss here the most common method for finding a linear model is to look for a line $y = mx + b$ in which the *sum of the squares* of the residuals is as small as possible. This is called the ***least-squares line of best fit*** or the ***regression line***.

Exercise 1 One of the lines in Figure 4 is the regression line. Which one is it?

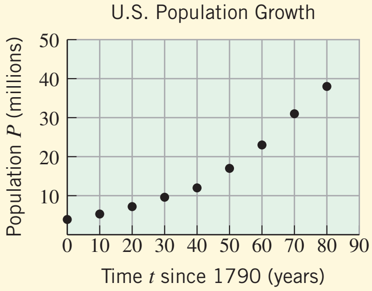

Figure 2

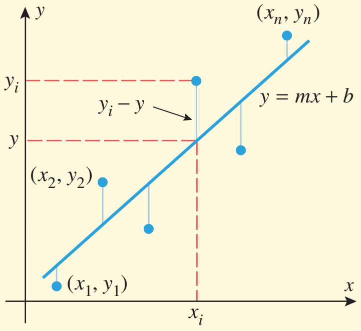

Figure 3

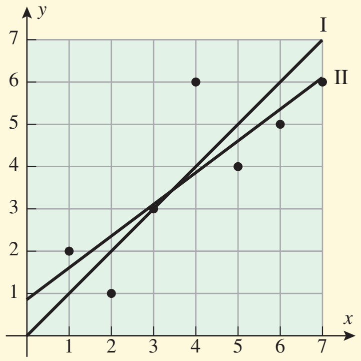

Figure 4

Exercise 2

(a) Most scientific calculators and CAS programs provide a method for finding regression lines. Read the documentation for your calculator or CAS to determine how to do this, and then find the regression line for the following (x, y) data:

x	1.0	1.5	2.0	2.5	3.0	3.5	4.0
y	1.0	2.5	6.0	9.0	10.5	14.5	15.0

(b) Make a scatter plot of the data together with the regression line.

How Good Is the Linear Model?

It is possible to compute a regression line, even in cases where the data have no apparent linear pattern. Thus, it is important to have some quantitative method of determining whether a linear model is appropriate for the data. The most common measure of linearity in data is called the ***correlation coefficient***, which is usually denoted by the letter r. A detailed explanation of correlation coefficients and the formula used to compute them is outside the scope of this text. However, here are some of the basic ideas:

- The values of r are in the interval $-1 \le r \le 1$, where r has the same sign as the slope of the regression line.
- If $r = \pm 1$, then the data points all lie on a line, so a linear model is a perfect fit for the data.

- If $r = 0$, then the data points exhibit no linear tendency, so a linear model is inappropriate for the data.

The closer r is to ±1, the more tightly the data points hug the regression line and the more appropriate the regression line is as a model; the closer r is to 0, the more scattered the points and the less appropriate the regression line is as a model (Figure 5). Roughly stated, the value of r^2 is a measure of the percentage of data points that fall in a "tight linear band." Thus, $r = 0.5$ means that 25% of the points fall in a tight linear band, and $r = 0.9$ means that 81% of the points fall in a tight linear band. A precise explanation of what is meant by a "tight linear band" requires ideas from statistics.

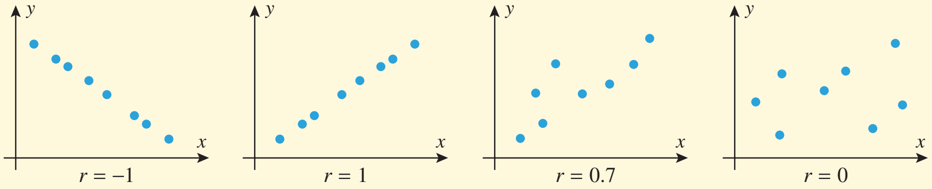

Figure 5

Exercise 3 If you have a scientific calculator, read the documentation to determine whether it produces the correlation coefficient when it computes a regression line. If you have a CAS, then the chances are that it will not automatically produce the correlation coefficient. However, on our website we have provided some CAS "miniprograms" that can be used to find regression lines with their associated correlation coefficients.

Exercise 4 Find the correlation coefficient for the data in Exercise 2.

Exercise 5

(a) Table 0.1.1 of Chapter 0 gives the Indianapolis 500 qualifying speeds S from 1989 to 2006. Take 1989 to be $t = 0$, and find the regression line and correlation coefficient for S versus t.

(b) Do you think that a linear model is reasonable for the data? Explain your reasoning.

(c) Predict the qualifying speed for the year 2012.

(d) What assumptions did you make in part (c)?

Nonlinear Models

Three of the most important nonlinear models are

- ***Exponential models*** $(y = ae^{bx})$
- ***Logarithmic models*** $(y = a + b \ln x)$
- ***Power function models*** $(y = ax^b)$

Many scientific calculators and computer programs can fit models of these types to data by the method of least squares. However, a useful alternative approach is to use logarithms to transform the original data into a form where linear models can be applied. This procedure, called ***linearizing*** the data, is based on the following idea:

- A set of (x_i, y_i) data will have an exponential model if the transformed data $(x_i, \log y_i)$ have a linear model.
- A set of (x_i, y_i) data will have a logarithmic model if the transformed data $(\log x_i, y_i)$ have a linear model.

- A set of (x_i, y_i) data will have a power function model if the transformed data $(\log x_i, \log y_i)$ have a linear model.

The following exercise explains the reason for this.

Exercise 6

(a) Suppose that $y = ae^{bx}$, and let $Y = \ln y$. Show that the graph of Y versus x is a line of slope b and Y-intercept $\ln a$.

(b) Suppose that $y = a + b \ln x$, and let $X = \ln x$. Show that the graph of y versus X is a line of slope b and y-intercept a.

(c) Suppose that $y = ax^b$, and let $Y = \ln y$ and $X = \ln x$. Show that the graph of Y versus X is a line of slope b and Y-intercept $\ln a$.

(d) Show that in parts (a), (b), and (c) the statements remain true if the natural logarithm "ln" is replaced by the common logarithm "log".

Exercise 7

(a) Find an exponential model $y = ae^{bx}$ for the following data by linearizing the data, finding the regression line for the linearized data, and then applying part (a) of Exercise 6 to find a and b.

x	0	1	2	3	4	5	6	7
y	3.9	5.3	7.2	9.6	12	17	23	31

(b) Make a scatter plot of the data together with the exponential model.

Exercise 8 The table in Figure 6 shows the relationship between the time T that it takes for each planet in our solar system to make one revolution around the Sun and the mean distance d between the planet and the Sun during one revolution. The graph in Figure 6 is a plot of $\log T$ versus $\log d$.

PLANET	MEAN DISTANCE d FROM THE SUN	TIME T FOR ONE REVOLUTION
Mercury	0.387	0.241
Venus	0.723	0.615
Earth	1.000	1.000
Mars	1.523	1.881
Jupiter	5.203	11.861
Saturn	9.541	29.457
Uranus	19.190	84.008
Neptune	30.086	164.784
Pluto	39.507	248.350

Note: Distances are measured in astronomical units (AU); 1 AU ≈ 93,000,000 mi. Time is measured in Earth years.

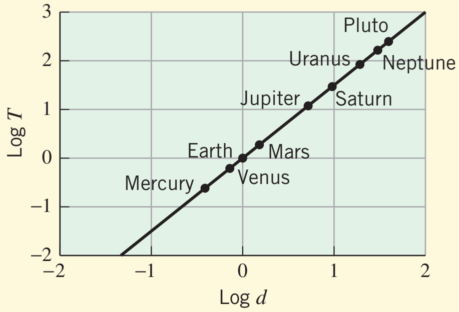

Figure 6

(a) What type of model for T as a function of d is suggested by the graph?

(b) Find the regression line for the $(\log d, \log T)$ data.

(c) Use the appropriate part of Exercise 6 to express T as a function of d.

(d) In part (c) you discovered Kepler's Third Law of Planetary Motion. Find some information about this law, and state the law in words.

Modeling Cooling

If a cup of hot coffee is left on a table to cool, then the graph of its temperature T versus the elapsed time t will have the general shape shown in Figure 7. The graph suggests that the coffee will cool quickly at first and then more and more slowly as its temperature approaches that of the room. To be more precise, it is shown in Physics that if the temperature of a liquid at time $t = 0$ is T_0 and if the room has a constant temperature of T_a, where $T_a < T_0$ (the room is cooler than the liquid), then the temperature T of the liquid at time t is given by

$$T = T_a + (T_0 - T_a)e^{-kt}$$

where k is a *positive* constant whose value depends on the physical characteristics of the liquid. This equation, called ***Newton's Law of Cooling***, can also be written in the form

$$T - T_a = (T_0 - T_a)e^{-kt}$$

which states that the difference between the temperature of the liquid and the temperature of the room has an exponential model.

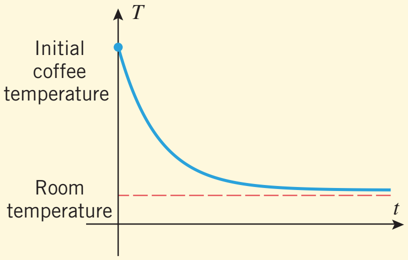

Figure 7

Exercise 9 Table 1 shows temperature measurements of a cup of coffee at 1-minute intervals after it was placed in a room with a constant temperature of 27°C.

(a) Find a model for the temperature T as a function of the elapsed time t.

(b) Estimate the temperature of the coffee at the time it was placed in the room.

(c) Approximately how long will it take until the coffee temperature is within 5°C of the room temperature?

Table 1

t (min)	1	2	3	4	5	6	7	8	9	10
T (°C)	82.2	79.6	77.3	75.0	73.1	70.7	69.2	66.9	65.3	63.3

Module by Mary Ann Connors, USMA, West Point, and Howard Anton, Drexel University