1

Introductory Concepts

1.1 Illustrative Example – Traditional Linear Least-Squares Regression

Consider this objective: find the best quadratic model, as described by the following equation:

$$y = a + bx + cx^2 \tag{1.1}$$

which matches the data in Figure 1.1.

Here, "x" represents the independent variable and "y" the dependent variable. Often, x and y are respectively termed *cause and effect, input and output, influence and response, property and condition*, and y is termed a function of x. Equation 1.1 is a human's mathematical description of how y responds to x and it is likely that the relation will not exactly match how nature actually works. In regression, in fitting a model to data, the values of the model coefficients (a, b, and c) will be adjusted to create the best model.

Conventionally, the best model is the one that minimizes the sum of squared distances from data point to model curve, where distance is that for the dependent variable (parallel to the vertical y axis). One data-to-model deviation is indicated as "d" on Figure 1.1. The sum of squared deviations (SSD) is defined as

$$SSD = \sum_{i=1}^{N} [y_i - \widetilde{y}(x_i)]^2 \tag{1.2}$$

where N indicates the number of data points on Figure 1.1 and "i" the number of a particular data point within the set of N. The number associated with a data point does not necessarily correspond with either x or y values. More likely the data point number corresponds to the chronological order of experimental trials that implemented the x value and measured the y response, as the sequential trial number is indicated on Figure 1.1. The data set might appear as illustrated in Table 1.1.

Continuing the explanation of Equation 1.2, y_i represents the ith measured y value, the data value, from Table 1.1, and $\widetilde{y}(x_i)$ indicates the model-calculated y value from Equation 1.1 using the ith x value from Table 1.1. The tilde accents on the symbols $\widetilde{y}$ and $\widetilde{y}(x_i)$ are both explicit indications that $\widetilde{y}(\mathrm{x}_i)$ represents the modeled y value. Redundancy in symbols is often not used, and here the $\widetilde{y}(x_i)$ term will be represented by either $\widetilde{y}_i$ or $\widetilde{y}(x_i)$.

Nonlinear Regression Modeling for Engineering Applications: Modeling, Model Validation, and Enabling Design of Experiments, First Edition. R. Russell Rhinehart.

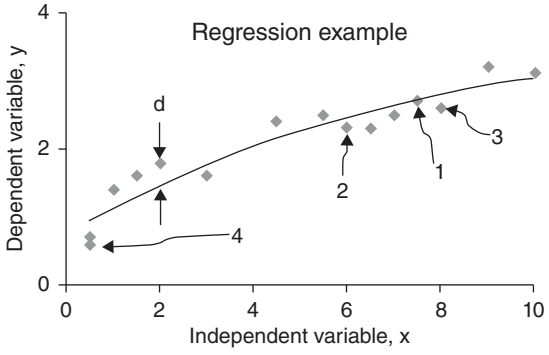

Figure 1.1 Illustration of regression concepts

Table 1.1 Illustration of data for Figure 1.1

Trial number	*X*, Input variable value	*Y*, Response variable value
1	7.5	2.7
2	6	2.3
3	8	2.6
4	0.5	0.7
.	.	.
.	.	.
.	.	.

The objective, find values for coefficients a, b, and c that minimize the SSD, defines an optimization procedure. Conventionally, the optimization application is stated by:

$$\underset{\{a,b,c\}}{Min}\, J = \sum_{i=1}^{N} [y_i - \tilde{y}(x_i)]^2 \tag{1.3}$$

In the jargon of optimization, Equation 1.3 reads, "The objective is to *Min*(imize) the *Objective Function J* (equal to the SSD) by adjustment of values of the decision variables (DVs) a, b, and c." This fully describes the regression "problem." DVs are what you adjust to minimize the objective function (OF) value. The DVs are the model coefficients that are adjusted to make the model best fit the data.

As a model of the y response to x, Equation 1.1 is nonlinear. *Nonlinear* means not linear, but does not indicate what the nonlinearity is (quadratic, cubic, reciprocal, exponential, etc.). If the cx^2 term was not in Equation 1.2 then the model would describe a linear y–x relation. However, in regression we adjust coefficient values, not the x or y values, and in Equation 1.1 each coefficient appears linearly (holding all else constant, the value of y is a linear response to the value of either a or b or c). The exponent for each coefficient is +1 and none of the coefficients are imbedded within a functionality that would make it have a nonlinear impact

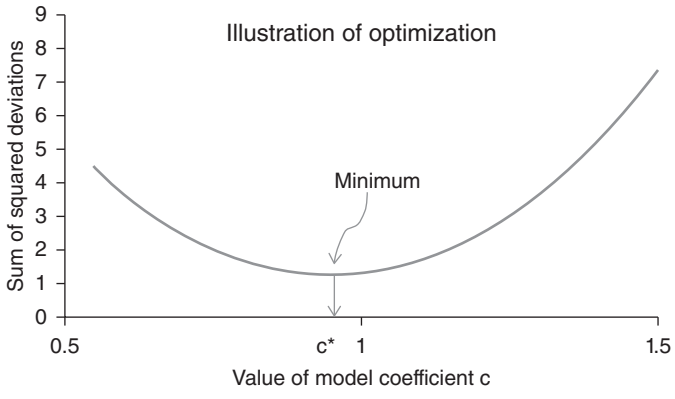

Figure 1.2 Illustration of regression coefficient optimization

on *y*. A formal definition of linearity is given later. Linearity simplifies determination of the optimum values for coefficients.

In linear regression, the model coefficients have a linear impact on the model prediction even if the model has a nonlinear $\tilde{y}(x_i)$ relation.

At the optimum value of each coefficient, SSD is a minimum. This means that any change in the coefficient value (either larger or smaller) makes the SSD larger. This is illustrated in Figure 1.2 for coefficient *c*. The concept is the same for each coefficient. The figure also illustrates that the SSD w.r.t. the coefficient graph has a slope (derivative) with a value of zero at the optimum value, *c**, of the coefficient. This property provides the classic method to determine values of the coefficients.

First, expand Equation 1.2 to explicitly reveal the model coefficients:

$$SSD = \sum_{i=1}^{N} [y_i - a - bx_i - cx_i{}^2]^2 \tag{1.4}$$

where x_i and y_i represent the values from Table 1.1 for the *i*th trial number.

Second, take the derivative of SSD with respect to each of the three coefficients and set the derivative (the slope) of each to zero (the value of the slope at the optimum value for each DV). This yields

$$\begin{aligned} \frac{\partial SSD}{\partial a} &= 0 = -2\sum_{i=1}^{N} [y_i - a - bx_i - cx_i{}^2] \\ \frac{\partial SSD}{\partial b} &= 0 = -2\sum_{i=1}^{N} [y_i - a - bx_i - cx_i{}^2]x_i \\ \frac{\partial SSD}{\partial c} &= 0 = -2\sum_{i=1}^{N} [y_i - a - bx_i - cx_i{}^2]x_i{}^2 \end{aligned} \tag{1.5}$$

Divide each of the three equations by the value of −2 and rearrange to produce the equation set 1.6, often termed the "normal equations." Note that the three model coefficients appear

linearly:

$$\begin{aligned} aN + b\Sigma x_i + c\Sigma {x_i}^2 &= \Sigma y_i \\ a\Sigma x_i + b\Sigma {x_i}^2 + c\Sigma {x_i}^3 &= \Sigma x_i y_i \\ a\Sigma {x_i}^2 + b\Sigma {x_i}^3 + c\Sigma {x_i}^4 &= \Sigma {x_i}^2 y_i \end{aligned} \tag{1.6}$$

Third, solve the three linear equations for the three unknowns, a, b, and c. There are many linear algebra methods for the solution of equation set 1.6, one being Gaussian elimination.

Equation set 1.6 is often presented in matrix-vector, linear algebra notation:

$$\begin{bmatrix} N & \Sigma x_i & \Sigma {x_i}^2 \\ \Sigma x_i & \Sigma {x_i}^2 & \Sigma {x_i}^3 \\ \Sigma {x_i}^2 & \Sigma {x_i}^3 & \Sigma {x_i}^4 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \Sigma y_i \\ \Sigma x_i y_i \\ \Sigma {x_i}^2 y_i \end{bmatrix} \tag{1.7}$$

or as

$$\underline{\underline{M}}\,\underline{c} = \underline{RHS} \tag{1.8}$$

If the reader is seeking to understand classic linear regression, the reader should create a simple problem and perform those three steps.

There are several efficiencies associated with linear regression. One is the direct extension to more complicated models. For example, if the model represented by Equation 1.1 was a cubic relation between y and x, then there would be an additional x^3 term, with a fourth coefficient, d. In that case, following the same procedure of minimizing SSD by taking the derivative of SSD w.r.t. each coefficient and setting the derivatives to zero, there would be four normal equations with four unknowns. If, for another example, there were two independent variables, x_1 and x_2, in a model with M number of linear coefficients there would be M normal equations. In linear regression there are M normal equations, each a linear relation in the M model coefficients, and each representing the derivative-equal-zero for each of the M coefficients. Regardless of the number of linear coefficients, the representation in Equation 1.8 remains the same and the linear algebra method for solving Equation 1.8 remains the same. This is one convenience of traditional linear regression.

The second convenience is, as long as the determinate of the M matrix is not zero, there is a unique solution that can be obtained from classic linear algebra solution methods.

Linear regression provides one universally applicable method that is guaranteed to return a unique solution.

In regression, the unknowns are the values of the model coefficients. By contrast, when using the model, model coefficients appear as the known values in Equation 1.1; but, in equation set 1.6 their roles are reversed. The a, b, and c values are the unknowns, and since the experimental values for N and each y_i and x_i are known, the values of the multipliers of the model coefficients are the known values. Since the unknowns, a, b, and c, appear linearly within the three equations of equation set 1.6, linear algebra procedures will solve for the coefficient values. This is predicated on the three equations being linearly independent, which requires relatively straightforward planning of the experimental procedure that generated the x–y data. Two heuristics that guide experimental design are to (i) have three or more independent data points for each coefficient to be evaluated and (ii) choose conditions such the data represent the entire x and y range.

The preceding example represents a *batch* data process in which all data are available and the objective is to find the best model to match the batch of data. However, in many applications related to modeling dynamic processes (those that change in time) the long-past data are irrelevant and model coefficients are incrementally updated using the new data at each sampling in an *incremental* or *recursive* manner.

The coefficients appear linearly in Equation 1.1. Although the model output, y, is a nonlinear (quadratic) function of the model input, x, the model output is linearly dependent on each coefficient. If you were to fix the value of x and plot the value of y w.r.t. any of the model coefficients, the graph would be a straight line. Contrasting this in nonlinear regression the sensitivity of model output to model coefficient is not linear.

1.2 How Models Are Used

Mathematical models represent a human conceptual understanding of how Nature works. They relate inputs to outputs, which are alternately termed causes to effects, influences to outcomes, stimulus to responses. Mathematical models are statements of relations such as: at a particular temperature the reaction rate will be such-and-such. When the load exceeds xxx the beam will fail or when the speed is xxx the fuel consumption will be yyy.

We use models in two prediction manners. Once we have the model, we use it to predict, forecast, or anticipate what Nature will do. In developing the model, we start with concepts and then convert the concepts into mathematical statements that are equivalent to our linguistic sentences. If the mathematical model of the human concept is a correct representation of Nature, then the model will match the data. If the concept is wrong the model will not match the data. Therefore, we also use models to test our concepts of how Nature works.

We also use the inverse of the model. Once we have a mathematical "sentence" of how Nature works, we can reverse the sentence to answer the question "If we want a particular outcome, what has to be done?" For instance: "What temperature is required to make the reaction go at a desired rate?" "If I want the beam to support xxx weight, how thick must it be?" In these applications we determine the input that gives the desired output, which is termed the inverse calculation.

1.3 Nonlinear Regression

Often, useful models are not linear in their coefficient influence on the model output and this leads to unique functionalities that do not have a generic method of solution. Consider a simple power law model

$$y = ax^b \tag{1.9}$$

This has two model coefficients, a and b. Following the classic analytical procedure for minimization of SSD, first take the derivatives of SSD w.r.t. each coefficient:

$$\frac{\partial SSD}{\partial a} = -2\sum(y_i - ax_i{}^b)x_i{}^b \tag{1.10}$$

$$\frac{\partial SSD}{\partial b} = -2\sum(y_i - ax_i{}^b)ax_i{}^b \ln(x_i) \tag{1.11}$$

and set each to zero:

$$0 = \sum y_i {x_i}^b - \sum a {x_i}^{2b} \tag{1.12}$$

$$0 = \sum y_i a {x_i}^b \ln(x_i) - \sum a^2 {x_i}^{2b} \ln(x_i) \tag{1.13}$$

This leads to two equations in two unknowns, and in general to M equations in M unknowns, but there does not appear to be a rearrangement that permits a linear algebra solution for the unknowns a and b. The a and b coefficients are neither independent (a and b are in the same term) nor linear (neither are to the first power).

Further, there is not a unique resulting functionality for the derivative relations. Each new model functionality results in new relations in the derivative equations. Following the procedure for another simple, two-parameter model,

$$y = \frac{a}{x - b} \tag{1.14}$$

leads to

$$0 = \sum \frac{y_i}{x_i - b} - a \sum \frac{1}{(x_i - b)^2} \tag{1.12}$$

$$0 = \sum \frac{y_i}{(x_i - b)^2} - a^2 \sum \frac{1}{(x_i - b)^3} \tag{1.13}$$

Again, this results in two equations in two unknowns, but they are nonlinear and not separable.

There are techniques for solving for the roots of a set of nonlinear equations (roots are the values of the coefficients that make the function equal zero), such as Newton's method, and it would appear that after the derivative equations are determined (in these examples either Equations 1.10 and 1.11 or 1.12 and 1.13), one could apply a standard procedure to solve for the unknowns. However, nonlinear equations can provide relationships that make root finding algorithms unstable and can generate multiple sets of roots. We need a better method.

Further, often in nonlinear regression there are constraints on variables (delays must be greater than zero, for example) and root finding algorithms cannot cope with constraints.

Finally, more often than not, one cannot analytically determine the derivative function. The equivalent of Equations 1.10 and 1.11 or 1.12 and 1.13 may not be available.

However, we desire a universal method for nonlinear regression that does not depend on the user defining analytical derivatives, that is, stable, and that can handle constraints. This book offers numerical optimization procedures as the answer. We will use numerical optimization to determine the numerical values of the coefficients in Equation 1.3.

1.4 Variable Types

Proper choices of model, optimization procedure, and objective function are dependent on the types of variables that are relevant. This section describes variable types.

There are a variety of naming conventions for numbers representing origins in mathematics, philosophy, statistics, and computer languages. However, the concepts are aligned. My explanation of number types comes from an engineering and computer language view. It somewhat

Table 1.2 Nomenclature for variable types

Here	Stephens' names	Other names
Class	Nominal	Classification, Category, Text, String
Rank	Ordinal	Opinion, Scale, Rating
Discrete	Cardinal	Integer
Continuum	Rational	Continuous-Valued, Scientific, Double Precision, Real
Deviation	Interval	Gage, Reference, Relative
Scaled	—	Dimensionless, Normalized

follows the taxonomy proposed by Stephens in 1946, a statistician, which still does not have total agreement within the statistics community. He proposed that variables are of the following types (Table 1.2).

Class represents classification and *Nominal* is derived from Latin for "name," but the name has no quantitative relation to the item. In a rock collection, there is the first rock you found, then the second, third, and so on. However, the numbers are just names and do not represent any quantity related to properties of the rock such as weight, size, value, thermal conductivity, surface roughness, or beauty. The names could have well been "A," "B," "C," or "alpha," "beta," "gamma," or "un," "deux," and "trois," We name trial runs sequentially in chronological order, even though trials should be run in a manner that randomizes the relation to any input variable. The "1" in the name "Trial 1" has no relevance to the flow rate, the impact, the difficulty, the benefit, and so on.

Nominal labels represent class, or category, or type. These could be city names, brand names, people names, social security numbers, colors, process types (Kraft or Solvay), pump types (centrifugal, turbine, positive displacement), separation types (distillation, absorption, crystallization), modeling types (neural network, finite-impulse), and so on. The names refer to items or procedures that have characteristic properties, but there is no correspondence of name spelling to the properties. You cannot mathematically relate the name to the value. In computer programming languages these are called string, text, or class variables.

Class or category can be related to dichotomous (two-valued) values (good/bad, on/off, fail/success, 1/0, tumor/not-tumor), and models are often used to predict whether something will be categorized as 1 or 0.

Rank means placement and *Ordinal* means sequence or order. A runner comes in first place, or second, or third, and so on, in a race. Here the number is a measure of placement, or goodness, or desirability, and relates to some property of the event. Alternately, it could relate to the property of an item (largest, second largest, … , or brightest, second brightest, …). Ordinal numbers reveal order, but they do not indicate a relative or proportional quantification of the property. For example, in a race, first and second may be nearly tied, and far ahead of third. Alternately, first may be far ahead of second and third. Ranking, "It came in second," indicates a relative relationship, but does not represent a ratio of quantity or goodness. Second place may be nearly as desirable as first, or far less desirable than first.

Similarly, opinion ranking scales (rate your opinion of the movie from best 5-star to worst 1-star), often called Likert scales, can place items in order, but the betterness of the 1 (outstanding) over the 2 (very good) may be much greater than the betterness of the 2 over the 3

(average). In school the standard quality points associated with letter grades of "A," "B," "C," "D," and "F" are 4, 3, 2, 1, and 0. These are used to calculate a grade point average (GPA) representing a student's average performance. However, if the normal distribution is used to assign grades, the "A" could represent the top 10 percentile, a "B" the 30 to 10 percentile (a 20% interval) and the "C" a 30% interval. An average of rankings may be a continuous-valued statistic, not just the integers associated with one ranking, but it is likely to be a nonlinear representation of "goodness."

Cardinal numbers, integers, are the counting numbers, the whole numbers, the indications of the number of whole items. They can only have integer values. They can only have values that exist at intervals of unity. The integer value is in direct proportion to the quantity it represents. Twelve is twice as much as six and six is twice three. Further, 10 is 2 more than 8 and 3 is 2 more than 1. The difference in values also represents the same counted quantity. The relation to quantity is preserved with addition, subtraction, multiplication, or division.

Integers are a type of *discrete* number, but the discretization interval is unity. However, the interval could be half, or quarters, or 16ths. The discretization does not have to be unity. The Westminster Clock chimes on the quarter hour. Its interval is $1/4$ of an hour. Or is it an interval of 15 minutes? Either way the time interval is not unity. In bit representation of computer storage, an 8-bit storage location might be filled with the binary sequence 00101101. The discrete interval is 1 bit. However, representing a display range of 0–100% the 2^8 possible numbers in an 8-bit storage have a 1-bit discretization interval that represents a $(100\% - 0\%)/2^8 = 0.390625\%$ interval. Observing such a number display, you would see all reported values as multiples of the 0.390625% interval. This minimum quantity represents discretization error, or discretization uncertainty. Look at a table of viscosity or table of t-statistic critical values and you will find that the table does not report infinite numbers for each entry. It may report one decimal digit, in which case the discretization interval is 0.1, or four decimal digits, in which case the discretization interval is 0.0001.

Whether the discrete numbers have unity or some other discretization interval value, they have several properties. One is that their value is linearly proportional to the quantity they represent. As a consequence differences between any two also relate to the quantity whether the numbers are high or low.

By contrast, a rank (or ordinal) number representing value scale, opinion scale, or ranking does not necessarily have the linearly proportional value. If the highest GPA is a 4.00 and the lowest a 0.00, then the difference between a 4.00 and a 3.50 is not the same as the difference between a 2.50 and a 2.00. The person with a 4.00 did not do twice as much work as one with a 2.00. They both may have completed all assignments. The 4.00 person did not get twice as many answers right as the 2.00 person. The 4.00 person might have got 93% correct on all tests, and the 2.00 person got 77% correct on all tests, representing a 1.2:1 performance advantage.

Continuum, *Real*, or *Rational* numbers refer to a continuum and are proportional to the quantity. We often imagine that properties of length, time, mass, and temperature are continuous and that the property can be divided into infinitesimal intervals. Continuum numbers permit having infinite decimal places. The ratio of 1 to 3 is $0.3333333\ldots$, but on an atomic view, mass is not continuous. If you want to increase the amount of water in a glass, the smallest increment you can add is one molecule of water. If you tried to add half of an H_2O molecule, it would not be a molecule of water. Effectively, on an engineering scale the continuum view seems valid and we consider the measurements of properties to be real numbers or continuum, with the discretization interval effectively zero. In one community these continuous-valued

numbers that are proportional to the quantity they represent are termed rational numbers. In computer programming they are called real numbers or single or double precision numbers.

There are other common definitions for the terms rational and real. From a mathematical view, real numbers are either rational (meaning ratio) if they can be represented by a ratio of integers or irrational if they cannot be represented by a ratio of integers. When the decimal part has a repeated pattern the real number is rational. If there is no repeated pattern (such as the square root of 2, the golden ratio, the base of the natural logarithm, or pi), the real number with infinite decimal digits is irrational. Further, in mathematics, real numbers are distinguished from imaginary or complex numbers that contain $\sqrt{-1}$ as an element. However, this text will use the term real to mean a continuous-valued number that is linearly proportional to the quantity it represents.

Next in the table comes a *deviation* variable, Stephens' category of *interval.* The zero of the interval or deviation scale does not represent that the property or characteristic of the item is zero. Zero degrees Fahrenheit does not mean zero thermal energy. You can still remove thermal energy and cool it to minus 17 °F. Zero gage pressure does not mean zero pressure. "Sea level" does not mean zero height. You cannot use deviation variables or interval values in many scientific calculations. Reaction kinetics are based on absolute temperature and pressure. Gravitational pull is based on distance from the center of the Earth. However, deviation variables preserve relative order and relative changes. The number of molecules needed to go from 5 to 6 psig is the same as that needed to go from 32 to 33 psig (ideally). The amount of thermal energy needed to raise the temperature from 38 to 39 °C is the same that is required to raise the temperature from 18 to 19 °C (ideally). In subtraction, deviation variables preserve relative difference, but they cannot be used in multiplication, addition, or division. Four degrees Celsius is not twice as hot as 2 °C. They do not preserve proportion or relative quantity.

Continuum numbers (rational or irrational) and discrete numbers (which includes integers) are directly related, linearly proportional, to the property of the event or item (time, mass, weight, intensity, value, etc.). As a result, real and discrete numbers have a ratio or proportional property. If the numerical representation of one event is twice in magnitude as the numerical representation of the other, then the one event has twice the magnitude, impact, value, of the other. Real and discrete numbers preserve this ratio or proportional relation, making them useful for modeling.

Finally, there is the category of *scaled* numbers. Usually these are either continuum or discrete numbers that are expressed as a ratio of full scale. A continuum example is "the glass is half full." An integer example is "we filled 90% of the seats." In those examples scaled variables are a fraction of the maximum possible value. Alternately, classic dimensionless groups, such as the Reynolds number or Fourier number, scale the extent of one mechanism with another. Here the scaled variable value is not bounded between 0 and 1. One can also scale deviation variables by their low to high value, and alternately in neural network modeling they are scaled from low to high on a −0.8 to +0.8 basis. Scaled variables have the same properties of ratio and relative proportion as their basis continuum, discrete, or deviation variables.

This textbook is mainly about the use of discrete (cardinal, integer) and continuum numbers (continuous valued, real, rational), both of which are proportional to the characteristic, within engineering models about the physical and chemical world. This is about how to generate the data, how to generate the models, and how to evaluate the models.

By contrast, there are methods related to the use of class or category or nominal variables as either the input or output (cause or effect, independent or dependent) variables in a model.

Although this text contains some discussion of such operations, the content of this book is not about that.

1.5 Simulation

The aspects of regression modeling can be simulated, as Chapter 5 more thoroughly directs, and doing so permits the reader to rapidly explore methods of this text.

With experimental data you do not know whether your regression modeling results are correct or not, because you do not know the truth about Nature. However, with simulation, you assume the truth about Nature (as represented with equations and coefficient values) and generate simulated experimental data. Then you can compare regression models to the simulated truth to see if the results of the procedures are correct. If the regression procedure consistently provides correct results over a range of critically devised tests, then you have a high confidence that the procedure is valid and will provide correct results on an unknown application.

A first step is to generate data that would simulate experimentally obtained data. The concept is that the experimenter would set the independent variables (experimental conditions), execute the experiment, and measure the response (dependent) variable. The simulation of this is illustrated in Figure 1.3, in which x is the independent variable and y is the response variable. Also indicated in the figure is that the measured response is corrupted by both noise (random perturbations) and a deviation, d_2 (systematic bias), and that the input to the process, P, is also corrupted by a deviation, d_1. In reality, the values of the deviations and noise are not knowable; however, for a simulation the experimenter can choose those values. The deviations would represent longer term calibration drifts, or the effects of input material variability, ambient conditions, device warm-up, operator training, operator fatigue, or equipment aging. These values would change slowly over time. By contrast, "noise" would represent sample-to-sample independent perturbations due to random events such as the vagaries of incomplete mixing, turbulence, sampling, mechanical vibrations, stray electronic field disturbances, and so on.

Figure 1.3 represents a traditional view with no uncertainty on x, and drift and noise linearly added to the otherwise truth.

To create a simulator using this model, first define the equation or procedure to obtain y_{true}, the truth about nature, from x:

$$y_{true} = f(x_{true}) \tag{1.14}$$

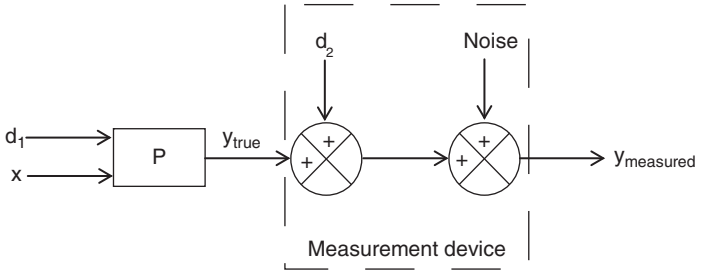

Figure 1.3 Simulation concept

Then perturb the values by the disturbances and noise:

$$y_{meas,i} = f(x_{nominal,i} + d_{1,i}) + d_{2,\,i} + n_{2,i} \tag{1.15}$$

You could add disturbance and noise to the input x in any of several appropriate ways. See Chapter 5 for a variety of options, and for models for noise and disturbance.

1.6 Issues

There are many types of models, but the best for engineering utility are often based on fundamental (alternately labeled scientific, first-principles, physical, theoretical, mechanistic, or phenomenological) derivations. In these, the coefficients that are adjusted to best fit the data often appear nonlinearly. Nonlinear regression is desired, but linear regression is easier.

Linear regression leads to the normal equations, equation set 1.6, that are deterministically solved, an aspect that makes it preferred over nonlinear regression. Often, to employ linear regression, we linearize functions by log-transforming power-law or exponential functions, or use other approaches that are mathematically correct when there is no uncertainty in either the x or y data. However, mathematically transformed data alters the relative locations and variance, which leads to biased coefficient values when there is experimental error in the data. By contrast, nonlinear regression can manage the equation in its nonlinear form. However, nonlinear optimization is an iterative procedure, which requires user-selected stopping criteria and which may become trapped in local minima. Normally, stopping criteria on either (or both) OF and DVs is scale-dependent and requires *a priori* user knowledge of the problem. Subsequent chapters show how to accommodate these aspects of nonlinear regression.

Should data be fitted with a linear, quadratic, cubic, or higher order model? Should it be a power law model? Does the model include the right input variables? There needs to be a method to determine the proper, justified, rational model structure. These are usually qualitative human choices and subsequent chapters provide a basis for making them.

You can fit a linear model to any data set. You can find a best model of any functional relation for any data set. Just because optimization finds the best coefficient values does not mean that the model is either *right* or *useful. Right* would imply that the model properly captures the natural phenomena that it seeks to represent. *Useful* would imply that the model balances perfection with sufficiency, that is provides a good-enough representation and that it is functional (convenient, reliable, sufficiently accurate) in use. Measures of *useful* are not normally included in lessons about the linear regression technique, and measures of *right* in linear regression lessons are usually based on r-square or other measures of removing variance. However, these statistical measures are independent of mechanistic reasoning. Subsequent sections show a technique for deciding whether the model is *right* for selecting the better from several equations and provide a set of criteria for evaluating *utility.*

Is there really no uncertainty on the x value? Perform any trial to generate data representing a y response to an x influence. Perhaps you wish to explore the influence of your kitchen oven temperature on the time it takes for the turkey thermometer to read 265 °F. You turn a knob or digitally specify the value of x, oven temperature. Is the oven really at that temperature when the power blinks on, then off? Or does the temperature control device lead you to believe it is at the specified temperature? In any experiment there is uncertainty on both the input and output. By contrast, conventional regression ascribes all of the uncertainty to the y response.

Further, the uncertainty may vary from one level to another. For example, differential pressure variability (noise) on orifice flow rate measurement increases with turbulence, which increases with flow rate. As another example of variability that changes with the *y* value, pH data may be highly sensitive to acid/base concentration variation near neutrality, but have minimal sensitivity in adjacent regions of nearly the same acid-to-base ratio. If one operating region produces data that is highly variable, the regression should reduce the weight of that data in setting model coefficients. Maximum likelihood approaches to the OF definition can accommodate these aspects. However, as a subsequent section reveals, they add complexity that is often not justified by utility aspects, by the improved functionality of the model due to the improved accuracy of the coefficient values.

Finally, my two favorite issues can be summed in the phrases "The model is wrong" and "Nature lies." *The model is wrong.* The mathematical model represents our human concept, an ideal version of the complexity of Nature. Regardless of how rigorous we try to make models, we always seem to find that our landmark breakthroughs still have an incompleteness. Nature is mysteriously complex, yet simple. In addition, we often simplify the models to obtain a mathematically tractable version. For example, we truncate the value of pi. We linearize. We truncate infinite series. We use numerical methods. Often, we intentionally make the model wrong, in a judgment that balances perfection with sufficiency.

Then, *Nature lies.* (Not to be disrespectful, Mother, but you seem to mask the reality about yourself.) Measurements are corrupted with uncertainty and calibration drifts, and experimental outcomes are corrupted with uncontrolled influences and variability of input materials and conditions. We call these noise or disturbances, random or systematic error.

Our objective in regression modeling is to use confounded data to determine coefficient values of an incomplete model, so that the model becomes functional.

Questions related to regression are:

- What should be used to determine goodness of the model fit to data?
- What model choices (model architecture and input variables) are best?
- Of competing model types, which is best?
- How many data points are best?
- Where should the data points be best located?
- What optimization approach is best?

To answer the questions, to use confounded data to create a functional model, we must understand:

- Models, model types, and measures of functionality.
- Systematic and random perturbations and how to quantify them.
- Nonlinear optimization, how to find the global optima, how to define convergence.
- How to propagate uncertainty in experimental conditions to data variability and from model coefficient variability to uncertainty of model prediction.
- How to test data for legitimacy and outliers.
- How to test model fit to the data.
- How to design experiments to generate data that provide defensible and useful models.

This book attempts to provide that knowledge, understanding, and tools.

1.7 Takeaway

Nature is usually nonlinear. Understand the variable types relevant to your application. Since the truth about Nature is unknown, simulations are a strong method to develop confidence and establish credibility.

Your model is wrong (not absolutely true) and the data are corrupted by fluctuations, noise, error, and so on. What you are attempting in regression is to fit a wrong model to data that misrepresents the truth. Quantifying the uncertainty in models is important to those who will want to use the models.

Exercises

1.1 Consider a linear $y(x)$ model and remove the term cx^2 from Equation 1.1. Follow the procedure to minimize SSD, obtain the two "normal equations" and solve for the a and b coefficient values.

1.2 Consider a cubic $y(x)$ model and add the term dx^3 to Equation 1.1. Show that minimizing SSD provides a linear set of equations in the form of Equation 1.8.

1.3 List some situations in which the modeled value is Class, Rank, Discrete, Real, or Deviation.

1.4 Provide several reasons for random error, noise on a measurement.

1.5 Provide several reasons for bias or systematic error on a measurement.

1.6 Provide several reasons for disturbance error, uncontrolled inputs, that affect experimental outcomes.

1.7 Consider that the objective is to find a linear regression model, $\widetilde{y} = a + bx$, to make the average residual zero, $0 = \overline{r} = 1/N \sum_{i=1}^{N} r_i = 1/N \sum_{i=1}^{N} (y_i - \widetilde{y}_i)$. Show that there are an infinite number of solutions.