33

The Method of Least Squares

KEY WORDS *confidence interval, critical sum of squares, dependent variable, empirical model, experimental error, independent variable, joint confidence region, least squares, linear model, linear least squares, mechanistic model, nonlinear model, nonlinear least squares, normal equation, parameter estimation, precision, regression, regressor, residual, residual sum of squares.*

One of the most common problems in statistics is to fit an equation to some data. The problem might be as simple as fitting a straight-line calibration curve where the independent variable is the known concentration of a standard solution and the dependent variable is the observed response of an instrument. Or it might be to fit an unsteady-state nonlinear model, for example, to describe the addition of oxygen to wastewater with a particular kind of aeration device where the independent variables are water depth, air flow rate, mixing intensity, and temperature.

The equation may be an *empirical model* (simply descriptive) or *mechanistic model* (based on fundamental science). A *response variable* or *dependent variable* (y) has been measured at several settings of one or more *independent variables* (x), also called *input variables, regressors*, or *predictor variables*. *Regression* is the process of fitting an equation to the data. Sometimes, regression is called *curve fitting* or *parameter estimation*.

The purpose of this chapter is to explain that certain basic ideas apply to fitting both linear and nonlinear models. Nonlinear regression is neither conceptually different nor more difficult than linear regression. Later chapters will provide specific examples of linear and nonlinear regression. Many books have been written on regression analysis and introductory statistics textbooks explain the method. Because this information is widely known and readily available, some equations are given in this chapter without much explanation or derivation. The reader who wants more details should refer to books listed at the end of the chapter.

Linear and Nonlinear Models

The fitted model may be a simple function with one independent variable, or it may have many independent variables with higher-order and nonlinear terms, as in the examples given below.

Linear models	$\eta = \beta_0 + \beta_1 x + \beta_2 x^2$	$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_2 x_1 x_2$
Nonlinear models	$\eta = \dfrac{\theta_1}{1 - \exp(-\theta_2 x)}$	$\eta = \exp(-\theta x_1)(1 - x_2)^{\theta_2}$

To maintain the distinction between linear and nonlinear we use a different symbol to denote the parameters. In the general linear model, $\eta = f(x, \beta)$, x is a vector of independent variables and β are parameters that will be estimated by regression analysis. The estimated values of the parameters $\beta_1, \beta_2, \ldots$ will be denoted by $b_1, b_2, \ldots$. Likewise, a general nonlinear model is $\eta = f(x, \theta)$ where θ is a vector of parameters, the estimates of which are denoted by $k_1, k_2, \ldots$.

The terms *linear* and *nonlinear* refer to the parameters in the model and not to the independent variables. Once the experiment or survey has been completed, the numerical values of the dependent

and independent variables are known. It is the parameters, the β's and θ's, that are unknown and must be computed. The model $y = \beta x^2$ is nonlinear in x; but once the known value of x^2 is provided, we have an equation that is linear in the parameter β. This is a linear model and it can be fitted by linear regression. In contrast, the model $y = x^\theta$ is nonlinear in θ, and θ must be estimated by nonlinear regression (or we must transform the model to make it linear).

It is usually assumed that a well-conducted experiment produces values of x_i that are essentially without error, while the observations of y_i are affected by random error. Under this assumption, the y_i observed for the ith experimental run is the sum of the true underlying value of the response (η_i) and a residual error (e_i):

$$y_i = \eta_i + e_i \quad i = 1, 2, \ldots, n$$

Suppose that we know, or tentatively propose, the linear model $\eta = \beta_0 + \beta_1 x$. The observed responses to which the model will be fitted are:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

which has residuals:

$$e_i = y_i - \beta_0 - \beta_1 x_i$$

Similarly, if one proposed the nonlinear model $\eta = \theta_1 \exp(-\theta_2 x)$, the observed response is:

$$y_i = \theta_1 \exp(-\theta_2 x_i) + e_i$$

with residuals:

$$e_i = y_i - \theta_1 \exp(-\theta_2 x_i)$$

The relation of the residuals to the data and the fitted model is shown in Figure 33.1. The lines represent the model functions evaluated at particular numerical values of the parameters. The residual ($e_i = y_i - \eta_i$) is the vertical distance from the observation to the value on the line that is calculated from the model. The residuals can be positive or negative.

The position of the line obviously will depend upon the particular values that are used for β_0 and β_1 in the linear model and for θ_1 and θ_2 in the nonlinear model. The regression problem is to select the values for these parameters that best fit the available observations. "Best" is measured in terms of making the residuals small according to a least squares criterion that will be explained in a moment.

If the model is correct, the residual $e_i = y_i - \eta_i$ will be nothing more than *random measurement error*. If the model is incorrect, e_i will reflect lack-of-fit due to all terms that are needed but missing from the model specification. This means that, after we have fitted a model, the residuals contain diagnostic information.

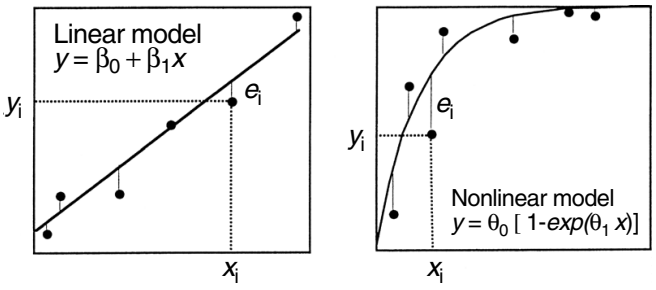

FIGURE 33.1 Definition of residual error for a linear model and a nonlinear model.

Residuals that are normally and independently distributed with constant variance over the range of values studied are persuasive evidence that the proposed model adequately fits the data. If the residuals show some pattern, the pattern will suggest how the model should be modified to improve the fit. One way to check the adequacy of the model is to check the properties of the residuals of the fitted model by plotting them against the predicted values and against the independent variables.

The Method of Least Squares

The best estimates of the model parameters are those that minimize the sum of the squared residuals:

$$S = \sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} (y_i - \eta_i)^2$$

The minimum sum of squares is called the *residual sum of squares* (S_R). This approach to estimating the parameters is known as the *method of least squares*. The method applies equally to linear and nonlinear models. The difference between linear and nonlinear regression lies in how the least squares parameter estimates are calculated. The essential difference is shown by example.

Each term in the summation is the difference between the observed y_i and the η computed from the model at the corresponding values of the independent variables x_i. If the residuals (e_i) are normally and independently distributed with constant variance, the parameter estimates are unbiased and have minimum variance.

For models that are linear in the parameters, there is a simple algebraic solution for the least squares parameter estimates. Suppose that we wish to estimate β in the model $\eta = \beta x$. The sum of squares function is:

$$S(\beta) = \sum (y_i - \beta x_i)^2 = \sum (y_i^2 - 2\beta x_i y_i + \beta^2 x_i^2)$$

The parameter value that minimizes S is the *least squares estimate* of the true value of β. This estimate is denoted by b. We can solve the sum of squares function for this estimate (b) by setting the derivative with respect to β equal to zero and solving for b:

$$\frac{dS(\beta)}{d\beta} = 0 = 2\sum (bx_i^2 - x_i y_i)$$

This equation is called the *normal equation*. Note that this equation is linear with respect to b. The algebraic solution is:

$$b = \frac{\sum x_i y_i}{\sum x_i^2}$$

Because x_i and y_i are known once the experiment is complete, this equation provides a generalized method for direct and exact calculation of the least squares parameter estimate. (Warning: This is not the equation for estimating the slope in a two-parameter model.)

If the linear model has two (or more) parameters to be estimated, there will be two (or more) normal equations. Each normal equation will be linear with respect to the parameters to be estimated and therefore an algebraic solution is possible. As the number of parameters increases, an algebraic solution is still possible, but it is tedious and the linear regression calculations are done using linear algebra (i.e., matrix operations). The matrix formulation was given in Chapter 30.

Unlike linear models, no unique algebraic solution of the normal equations exists for nonlinear models. For example, if $\eta = \exp(-\theta x)$, the method of least squares requires that we find the value of θ that minimizes S:

$$S(\theta) = \sum (y_i - \exp(-\theta x_i))^2 = \sum [y_i^2 - 2y_i \exp(-\theta x_i) + (\exp(-\theta x_i))^2]$$

TABLE 33.1

Example Data and the Sum of Squares Calculations for a One-Parameter Linear Model and a One-Parameter Nonlinear Model

Linear Model: $\eta = \beta x$					Nonlinear Model: $\eta_i = \exp(-\theta x_i)$				
x_i	$y_{obs,i}$	$y_{calc,i}$	e_i	$(e_i)^2$	x_i	$y_{obs,i}$	$y_{calc,i}$	e_i	$(e_i)^2$
Trial value: $b = 0.115$					Trial value: $k = 0.32$				
2	0.150	0.230	–0.080	0.0064	2	0.620	0.527	0.093	0.0086
4	0.461	0.460	0.001	0.0000	4	0.510	0.278	0.232	0.0538
6	0.559	0.690	–0.131	0.0172	6	0.260	0.147	0.113	0.0129
10	1.045	1.150	–0.105	0.0110	10	0.180	0.041	0.139	0.0194
14	1.364	1.610	–0.246	0.0605	14	0.025	0.011	0.014	0.0002
19	1.919	2.185	–0.266	0.0708	19	0.041	0.002	0.039	0.0015
		Sum of squares = 0.1659					Sum of squares = 0.0963		
Trial value: $b = 0.1$ (optimal)					Trial value: $k = 0.2$ (optimal)				
2	0.150	0.200	–0.050	0.0025	2	0.620	0.670	–0.050	0.0025
4	0.461	0.400	0.061	0.0037	4	0.510	0.449	0.061	0.0037
6	0.559	0.600	–0.041	0.0017	6	0.260	0.301	–0.041	0.0017
10	1.045	1.000	0.045	0.0020	10	0.180	0.135	0.045	0.0020
14	1.364	1.400	–0.036	0.0013	14	0.025	0.061	–0.036	0.0013
19	1.919	1.900	0.019	0.0004	19	0.041	0.022	0.019	0.0003
		Minimum sum of squares = 0.0116					Minimum sum of squares = 0.0115		

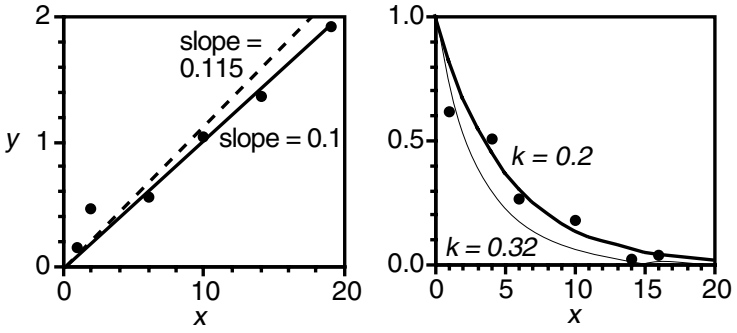

FIGURE 33.2 Plots of data to be fitted to linear (left) and nonlinear (right) models and the curves generated from the initial parameter estimates of $b = 0.115$ and $k = 0.32$ and the minimum least squares values ($b = 0.1$ and $k = 0.2$).

The least squares estimate of θ still satisfies $\partial S/\partial\theta = 0$, but the resulting derivative does not have an algebraic solution. The value of θ that minimizes S is found by iterative numerical search.

Examples

The similarities and differences of linear and nonlinear regression will be shown with side-by-side examples using the data in Table 33.1. Assume there are theoretical reasons why a linear model ($\eta_i = \beta x_i$) fitted to the data in Figure 33.2 should go through the origin, and an exponential decay model ($\eta_i = \exp(-\theta x_i)$) should have $y = 1$ at $t = 0$. The models and their sum of squares functions are:

$$y_i = \beta x_i + e_i \qquad \min S(\beta) = (y_i - \beta x_i)^2$$

$$y_i = \exp(-\theta x_i) + e_i \qquad \min S(\theta) = \sum (y_i - \exp(-\theta x_i))^2$$

For the linear model, the sum of squares function expanded in terms of the observed data and the parameter β is:

$$S(\beta) = (0.15 - 2\beta)^2 + (0.461 - 4\beta)^2 + (0.559 - 6\beta)^2 + (1.045 - 10\beta)^2 + (1.361 - 14\beta)^2 + (1.919 - 19\beta)^2$$

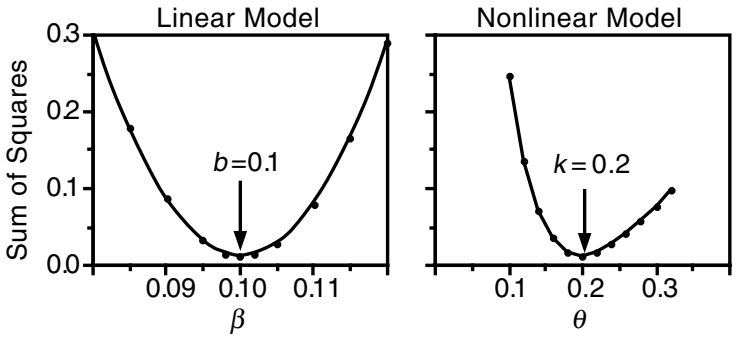

FIGURE 33.3 The values of the sum of squares plotted as a function of the trial parameter values. The least squares estimates are $b = 0.1$ and $k = 0.2$. The sum of squares function is symmetric (parabolic) for the linear model (left) and asymmetric for the nonlinear model (right).

For the nonlinear model it is:

$$S(\theta) = (0.62 - e^{-2\theta})^2 + (0.51 - e^{-4\theta})^2 + (0.26 - e^{6\theta})^2 + (0.18 - e^{-10\theta})^2 + (0.025 - e^{-14\theta})^2 + (0.041 - e^{-19\theta})^2$$

An algebraic solution exists for the linear model, but to show the essential similarity between linear and nonlinear parameter estimation, the least squares parameter estimates of both models will be determined by a straightforward numerical search of the sum of squares functions. We simply plot $S(\beta)$ over a range of values of β, and do the same for $S(\theta)$ over a range of θ.

Two iterations of this calculation are shown in Table 33.1. The top part of the table shows the trial calculations for initial parameter estimates of $b = 0.115$ and $k = 0.32$. One clue that these are poor estimates is that the residuals are not random; too many of the linear model regression residuals are negative and all the nonlinear model residuals are positive. The bottom part of the table is for $b = 0.1$ and $k = 0.2$, the parameter values that give the minimum sum of squares.

Figure 33.3 shows the smooth sum of squares curves obtained by following this approach. The minimum sum of squares — the minimum point on the curve — is called the *residual sum of squares* and the corresponding parameter values are called the *least squares estimates*. The least squares estimate of β is $b = 0.1$. The least squares estimate of θ is $k = 0.2$. The fitted models are $\hat{y} = 0.1x$ and $\hat{y} = \exp(-0.2x)$. $\hat{y}$ is the predicted value of the model using the least squares parameter estimate.

The sum of squares function of a linear model is always symmetric. For a univariate model it will be a parabola. The curve in Figure 33.3a is a parabola. The sum of squares function for nonlinear models is not symmetric, as can be seen in Figure 33.3b.

When a model has two parameters, the sum of squares function can be drawn as a surface in three dimensions, or as a contour map in two dimensions. For a two-parameter linear model, the surface will be a parabaloid and the contour map of S will be concentric ellipses. For nonlinear models, the sum of squares surface is not defined by any regular geometric function and it may have very interesting contours.

The Precision of Estimates of a Linear Model

Calculating the "best" values of the parameters is only part of the job. The precision of the parameter estimates needs to be understood. Figure 33.3 is the basis for showing the confidence interval of the example one-parameter models.

For the *one-parameter* linear model *through the origin*, the variance of b is:

$$\text{Var}(b) = \frac{\sigma^2}{\sum x_i^2}$$

The summation is over all squares of the settings of the independent variable x. σ^2 is the *experimental error variance*. (Warning: This equation does not give the variance for the slope of a two-parameter linear model.)

Ideally, σ^2 would be estimated from independent replicate experiments at some settings of the x variable. There are no replicate measurements in our example, so another approach is used. The residual sum of squares can be used to estimate σ^2 if one is willing to assume that the model is correct. In this case, the residuals are random errors and the average of these residuals squared is an estimate of the error variance σ^2. Thus, σ^2 may be estimated by dividing the residual sum of squares (S_R) by its degrees of freedom $(\nu = n - p)$, where n is the number of observations and p is the number of estimated parameters.

In this example, $S_R = 0.0116$, $p = 1$ parameter, $n = 6$, $\nu = 6 - 1 = 5$ degrees of freedom, and the estimate of the experimental error variance is:

$$s^2 = \frac{S_R}{n-p} = \frac{0.0116}{5} = 0.00232$$

The estimated variance of b is:

$$\text{Var}(b) = \frac{s^2}{\sum x_i^2} = \frac{0.00232}{713} = 0.0000033$$

and the standard error of b is:

$$\text{SE}(b) = \sqrt{\text{Var}(b)} = \sqrt{0.0000032} = 0.0018$$

The $(1-\alpha)100\%$ confidence limits for the true value β are:

$$b \pm t_{\nu,\alpha/2}\text{SE}(b)$$

For $\alpha = 0.05$, $\nu = 5$, we find $t_{5,0.025} = 2.571$, and the 95% confidence limits are $0.1 \pm 2.571(0.0018) = 0.1 \pm 0.0046$.

Figure 33.4a expands the scale of Figure 33.3a to show more clearly the confidence interval computed from the t statistic. The sum of squares function and the confidence interval computed using the t statistic are both symmetric about the minimum of the curve. The upper and lower bounds of the confidence interval define two intersections with the sum of squares curve. The sum of squares at these two points is identical because of the symmetry that always exists for a linear model. This level of the sum of squares function is the *critical sum of squares, S_c*. All values of β that give $S < S_c$ fall within the 95% confidence interval.

Here we used the easily calculated confidence interval to define the critical sum of squares. Usually the procedure is reversed, with the critical sum of squares being used to determine the boundary of the confidence region for two or more parameters. Chapters 34 and 35 explain how this is done. The *F statistic* is used instead of the t statistic.

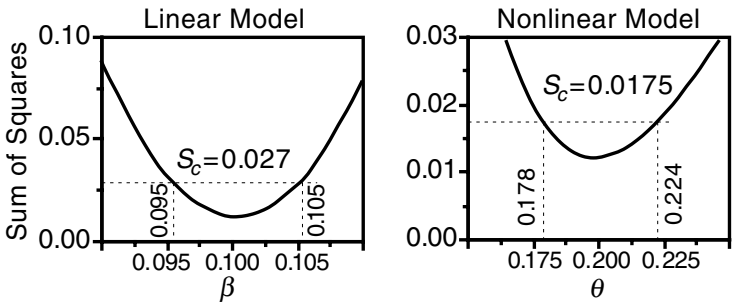

FIGURE 33.4 Sum of squares functions from Figure 33.3 replotted on a larger scale to show the confidence intervals of β for the linear model (left) and θ for the nonlinear model (right).

The Precision of Estimates of a Nonlinear Model

The sum of squares function for the nonlinear model (Figure 33.3) is not symmetrical about the least squares parameter estimate. As a result, the confidence interval for the parameter θ is not symmetric. This is shown in Figure 33.4, where the confidence interval is 0.20 – 0.022 to 0.20 + 0.024, or [0.178, 0.224].

The asymmetry near the minimum is very modest in this example, and a symmetric *linear approximation* of the confidence interval would not be misleading. This usually is not the case when two or more parameters are estimated. Nevertheless, many computer programs do report confidence intervals for nonlinear models that are based on symmetric linear approximations. These intervals are useful as long as one understands what they are.

This asymmetry is one difference between the linear and nonlinear parameter estimation problems. The essential similarity, however, is that we can still define a critical sum of squares and it will still be true that all parameter values giving $S \leq S_c$ fall within the confidence interval. Chapter 35 explains how the critical sum of squares is determined from the minimum sum of squares and an estimate of the experimental error variance.

Comments

The method of least squares is used in the analysis of data from planned experiments and in the analysis of data from unplanned happenings. For the least squares parameter estimates to be unbiased, the residual errors ($e = y - \eta$) must be random and independent with constant variance. It is the tacit assumption that these requirements are satisfied for unplanned data that produce a great deal of trouble (Box, 1966). Whether the data are planned or unplanned, the residual (e) includes the effect of latent variables (lurking variables) which we know nothing about.

There are many conceptual similarities between linear least squares regression and nonlinear regression. In both, the parameters are estimated by minimizing the sum of squares function, which was illustrated in this chapter using one-parameter models. The basic concepts extend to models with more parameters.

For linear models, just as there is an exact solution for the parameter estimates, there is an exact solution for the $100(1 - \alpha)\%$ confidence interval. In the case of linear models, the linear algebra used to compute the parameter estimates is so efficient that the work effort is not noticeably different to estimate one or ten parameters.

For nonlinear models, the sum of squares surface can have some interesting shapes, but the precision of the estimated parameters is still evaluated by attempting to visualize the sum of squares surface, preferably by making contour maps and tracing approximate joint confidence regions on this surface.

Evaluating the precision of parameter estimates in multiparameter models is discussed in Chapters 34 and 35. If there are two or more parameters, the sum of squares function defines a surface. A joint confidence region for the parameters can be constructed by tracing along this surface at the critical sum of squares level. If the model is linear, the joint confidence regions are still based on parabolic geometry. For two parameters, a contour map of the joint confidence region will be described by ellipses. In higher dimensions, it is described by ellipsoids.

References

Box, G. E. P. (1966). "The Use and Abuse of Regression," *Technometrics,* 8, 625–629.
Chatterjee, S. and B. Price (1977). *Regression Analysis by Example,* New York, John Wiley.
Draper, N. R. and H. Smith, (1998). *Applied Regression Analysis,* 3rd ed., New York, John Wiley.
Meyers, R. H. (1986). *Classical and Modern Regression with Applications,* Boston, MA, Duxbury Press.

Mosteller, F. and J. W. Tukey (1977). *Data Analysis and Regression: A Second Course in Statistics,* Reading, MA, Addison-Wesley Publishing Co.

Neter, J., W. Wasserman, and M. H. Kutner (1983). *Applied Regression Models,* Homewood, IL, Richard D. Irwin Co.

Rawlings, J. O. (1988). *Applied Regression Analysis: A Research Tool,* Pacific Grove, CA, Wadsworth and Brooks/Cole.

Exercises

33.1 Model Structure. Are the following models linear or nonlinear in the parameters?

(a) $\eta = \beta_0 + \beta_1 x^2$

(b) $\eta = \beta_0 + \beta_1 2^x$

(c) $\eta = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \dfrac{\beta_4}{x - 60}$

(d) $\eta = \dfrac{\beta_0}{x + \beta_1 x}$

(e) $\eta = \beta_0(1 + \beta_1 x_1)(1 + \beta_2 x_2)$

(f) $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{123} x_1 x_2 x_3$

(g) $\eta = \beta_0[1 - \exp(-\beta_1 x)]$

(h) $\eta = \beta_0[1 - \beta_1 \exp(-x)]$

(i) $\ln(\eta) = \beta_0 + \beta_1 x$

(j) $\dfrac{1}{\eta} = \beta_0 + \dfrac{\beta_1}{x}$

33.2 Fitting Models. Using the data below, determine the least squares estimates of β and θ by plotting the sum of squares for these models: $\eta_1 = \beta x^2$ and $\eta_2 = 1 - \exp(-\theta x)$.

x	y_1	y_2
2	2.8	0.44
4	6.2	0.71
6	10.4	0.81
8	17.7	0.93

33.3 Normal Equations. Derive the two normal equations to obtain the least squares estimates of the parameters in $y = \beta_0 + \beta_1 x$. Solve the simultaneous equations to get expressions for b_0 and b_1, which estimate the parameters β_0 and β_1.