Appendix D

Nonlinear Least Squares Data Fitting

D.1 Introduction

A nonlinear least squares problem is an unconstrained minimization problem of the form

$$\underset{x}{\text{minimize}} \; f(x) = \sum_{i=1}^{m} f_i(x)^2,$$

where the objective function is defined in terms of auxiliary functions $\{ f_i \}$. It is called "least squares" because we are *minimizing* the sum of *squares* of these functions. Looked at in this way, it is just another example of unconstrained minimization, leading one to ask why it should be studied as a separate topic. There are several reasons.

In the context of data fitting, the auxiliary functions $\{ f_i \}$ are not arbitrary nonlinear functions. They correspond to the residuals in a data fitting problem (see Chapter 1). For example, suppose that we had collected data $\{ (t_i, y_i) \}_{i=1}^{m}$ consisting of the size of a population of antelope at various times. Here t_i corresponds to the time at which the population y_i was counted. Suppose we had the data

t_i :	1	2	4	5	8
y_i :	3	4	6	11	20

where the times are measured in years and the populations are measured in hundreds. It is common to model populations using exponential models, and so we might hope that

$$y_i \approx x_1 e^{x_2 t_i}$$

for appropriate choices of the parameters x_1 and x_2. A model of this type is illustrated in Figure D.1.

If least squares were used to select the parameters (see Section 1.5) then we would solve

$$\underset{x_1, x_2}{\text{minimize}} \; f(x_1, x_2) = \sum_{i=1}^{5} (x_1 e^{x_2 t_i} - y_i)^2$$

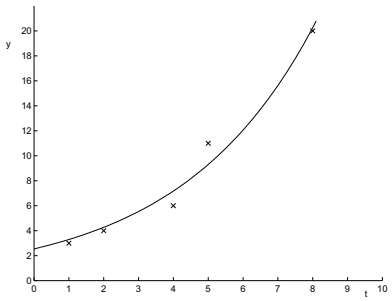

Figure D.1. *Exponential Model of Antelope Population*

so that the i-th function would be

$$f_i(x_1, x_2) = x_1 e^{x_2 t_i} - y_i,$$

that is, it would be the residual for the i-th data point. Most least squares problems are of this form, where the functions $f_i(x)$ are residuals and where the index i indicates the particular data point. This is one way in which least squares problems are distinctive.

Least-squares problems are also distinctive in the way that the solution is interpreted. Least squares problems usually incorporate some assumptions about the errors in the model. For example, we might have

$$y_i = x_1 e^{x_2 t_i} + \epsilon_i,$$

where the errors $\{ \epsilon_i \}$ are assumed to arise from a single probability distribution, often the normal distribution. Associated with our model are the "true" parameters x_1 and x_2, but each time we collect data and solve the least-squares problem we only obtain estimates $\hat{x}_1$ and $\hat{x}_2$ of these true parameters. After computing these estimates, it is common to ask questions about the model such as: What bounds can we place on the values of the true parameters? Does the model adequately fit the data? How sensitive are the parameters to changes in the data? And so on.

Algorithms for least-squares problems are also distinctive. This is a consequence of the special structure of the Hessian matrix for the least-squares objective function. The Hessian in this case is the sum of two terms. The first only involves the gradients of the functions $\{ f_i \}$ and so is easier to compute. The second involves the second derivatives, but is zero if the errors $\{ \epsilon_i \}$ are all zero (that is, if the model fits the data perfectly). It is tempting to approximate the second term in the Hessian, and many algorithms for least squares do this. Additional techniques are used to deal with the first term in a computationally sensible manner.

If least-squares problems were uncommon then even these justifications would not be enough to justify our discussion here. But they are not uncommon. They are one of the most widely encountered unconstrained optimization problems, and amply justify the attention given them.

D.2 Nonlinear Least-Squares Data Fitting

Let us first examine the special form of the derivatives in a least-squares problem. We will write the problem as

$$\underset{x}{\text{minimize}}\ f(x) = \tfrac{1}{2}\sum_{i=1}^{m} f_i(x)^2 \equiv \tfrac{1}{2}F(x)^T F(x)$$

where F is the vector-valued function

$$F(x) = (\ f_1(x) \quad f_2(x) \quad \cdots \quad f_m(x)\ )^T.$$

We have scaled the problem by $\frac{1}{2}$ to make the derivatives less cluttered. The components of $\nabla f(x)$ can be derived using the chain rule:

$$\nabla f(x) = \nabla F(x)\, F(x).$$

$\nabla^2 f(x)$ can be derived by differentiating this formula with respect to x_j:

$$\nabla^2 f(x) = \nabla F(x)\, \nabla F(x)^T + \sum_{i=1}^{m} f_i(x)\nabla^2 f_i(x).$$

These are the formulas for the gradient and Hessian of f.

Let x_* be the solution of the least-squares problem, and suppose that at the solution, $f(x_*) = 0$. Then $f_i(x_*) = 0$ for all i, indicating that all the residuals are zero and that the model fits the data with no error. As a result, $F(x_*) = 0$ and hence $\nabla f(x_*) = 0$, confirming that the first-order necessary condition is satisfied. It also follows that

$$\nabla^2 f(x_*) = \nabla F(x_*)\, \nabla F(x_*)^T,$$

so that the Hessian at the solution is positive semi-definite, as expected. If $\nabla F(x_*)$ is a matrix of full rank then $\nabla^2 f(x_*)$ is positive definite.

Example D.1 Gradient and Hessian. For the antelope data and model in Section D.1,

$$F(x) = \begin{pmatrix} x_1 e^{x_2 t_1} - y_1 \\ x_1 e^{x_2 t_2} - y_2 \\ x_1 e^{x_2 t_3} - y_3 \\ x_1 e^{x_2 t_4} - y_4 \\ x_1 e^{x_2 t_5} - y_5 \end{pmatrix} = \begin{pmatrix} x_1 e^{1 x_2} - \ 3 \\ x_1 e^{2 x_2} - \ 4 \\ x_1 e^{4 x_2} - \ 6 \\ x_1 e^{5 x_2} - 11 \\ x_1 e^{8 x_2} - 20 \end{pmatrix}.$$

The formula for the least-squares objective function is

$$f(x_1, x_2) = \tfrac{1}{2}\sum_{i=1}^{5}(x_1 e^{x_2 t_i} - y_i)^2 = \tfrac{1}{2}F(x)^T F(x).$$

The gradient of f is

$$\nabla f(x_1, x_2) = \begin{pmatrix} \displaystyle\sum_{i=1}^{5}(x_1 e^{x_2 t_i} - y_i)e^{x_2 t_i} \\ \displaystyle\sum_{i=1}^{5}(x_1 e^{x_2 t_i} - y_i)x_1 t_i e^{x_2 t_i} \end{pmatrix}.$$

This can be rewritten as $\nabla f(x_1, x_2) =$

$$\begin{pmatrix} e^{x_2 t_1} & e^{x_2 t_2} & e^{x_2 t_3} & e^{x_2 t_4} & e^{x_2 t_5} \\ x_1 t_1 e^{x_2 t_1} & x_1 t_2 e^{x_2 t_2} & x_1 t_3 e^{x_2 t_3} & x_1 t_4 e^{x_2 t_4} & x_1 t_5 e^{x_2 t_5} \end{pmatrix} \begin{pmatrix} x_1 e^{x_2 t_1} - y_1 \\ x_1 e^{x_2 t_2} - y_2 \\ x_1 e^{x_2 t_3} - y_3 \\ x_1 e^{x_2 t_4} - y_4 \\ x_1 e^{x_2 t_5} - y_5 \end{pmatrix}$$

so that $\nabla f(x_1, x_2) = \nabla F(x)\, F(x)$. The Hessian matrix is $\nabla^2 f(x) = \nabla F(x)\, \nabla F(x)^T + \sum_{i=1}^{m} f_i(x) \nabla^2 f_i(x) =$

$$\begin{pmatrix} e^{x_2 t_1} & e^{x_2 t_2} & e^{x_2 t_3} & e^{x_2 t_4} & e^{x_2 t_5} \\ x_1 t_1 e^{x_2 t_1} & x_1 t_2 e^{x_2 t_2} & x_1 t_3 e^{x_2 t_3} & x_1 t_4 e^{x_2 t_4} & x_1 t_5 e^{x_2 t_5} \end{pmatrix} \begin{pmatrix} e^{x_2 t_1} & x_1 t_1 e^{x_2 t_1} \\ e^{x_2 t_2} & x_1 t_2 e^{x_2 t_2} \\ e^{x_2 t_3} & x_1 t_3 e^{x_2 t_3} \\ e^{x_2 t_4} & x_1 t_4 e^{x_2 t_4} \\ e^{x_2 t_5} & x_1 t_5 e^{x_2 t_5} \end{pmatrix}$$
$$+ \sum_{i=1}^{5} (x_1 e^{x_2 t_i} - y_i) \begin{pmatrix} 0 & t_i e^{x_2 t_i} \\ t_i e^{x_2 t_i} & x_1 t_i^2 e^{x_2 t_i} \end{pmatrix}.$$

Note that $\{ t_i \}$ and $\{ y_i \}$ are the data values for the model, while x_1 and x_2 are the variables in the model.

∎

If $F(x_*) = 0$ then it is reasonable to expect that $F(x) \approx 0$ for $x \approx x_*$, implying that

$$\nabla^2 f(x) = \nabla F(x)\, \nabla F(x)^T + \sum_{i=1}^{m} f_i(x) \nabla^2 f_i(x) \approx \nabla F(x)\, \nabla F(x)^T.$$

This final formula only involves the first derivatives of the functions $\{ f_i \}$ and suggests that an approximation to the Hessian matrix can be found using only first derivatives, at least in cases where the model is a good fit to the data. This idea is the basis for a number of specialized methods for nonlinear least squares data fitting.

The simplest of these methods, called the *Gauss-Newton method* uses this approximation directly. It computes a search direction using the formula for Newton's method

$$\nabla^2 f(x) p = -\nabla f(x)$$

but replaces the Hessian with this approximation

$$\nabla F(x)\, \nabla F(x)^T p = -\nabla F(x)\, F(x).$$

In cases where $F(x_*) = 0$ and $\nabla F(x_*)$ is of full rank, the Gauss-Newton method behaves like Newton's method near the solution, but without the costs associated with computing second derivatives.

The Gauss-Newton method can perform poorly when the residuals at the solution are not "small" (that is, when the model does not fit the data well), or if the Jacobian of F is not of full rank at the solution. Loosely speaking, in these cases the Gauss-Newton method is using a poor approximation to the Hessian of f.

Example D.2 Gauss-Newton Method. We apply the Gauss-Newton method to an exponential model of the form

$$y_i \approx x_1 e^{x_2 t_i}$$

with data

$$\begin{aligned} t &= (\,1 \quad 2 \quad 4 \quad 5 \quad 8\,)^T \\ y &= (\,3.2939 \quad 4.2699 \quad 7.1749 \quad 9.3008 \quad 20.259\,)^T. \end{aligned}$$

For this example, the vector y was chosen so that the model would be a good fit to the data, and hence we would expect the Gauss-Newton method to perform much like Newton's method. (In general y will not be chosen, but will be part of the given data for a problem.) We apply the Gauss-Newton method without a line search, using an initial guess that is close to the solution:

$$x = \begin{pmatrix} 2.50 \\ 0.25 \end{pmatrix}.$$

At this point

$$F(x) = \begin{pmatrix} -0.0838 \\ -0.1481 \\ -0.3792 \\ -0.5749 \\ -1.7864 \end{pmatrix} \quad \text{and} \quad \nabla F(x)^T = \begin{pmatrix} 1.2840 & 3.2101 \\ 1.6487 & 8.2436 \\ 2.7183 & 27.1828 \\ 3.4903 & 43.6293 \\ 7.3891 & 147.7811 \end{pmatrix}.$$

Hence

$$\begin{aligned} \nabla f(x) = \nabla F(x)\, F(x) &= \begin{pmatrix} -16.5888 \\ -300.8722 \end{pmatrix} \\ \nabla F(x)\, \nabla F(x)^T &= \begin{pmatrix} 78.5367 & 1335.8479 \\ 1335.8479 & 24559.9419 \end{pmatrix}. \end{aligned}$$

The Gauss-Newton search direction is obtained by solving the linear system

$$\nabla F(x)\, \nabla F(x)^T = -\nabla F(x)\, F(x)$$

and so

$$p = \begin{pmatrix} 0.0381 \\ 0.0102 \end{pmatrix}$$

and the new estimate of the solution is

$$x \leftarrow x + p = \begin{pmatrix} 2.5381 \\ 0.2602 \end{pmatrix}.$$

(For simplicity, we do not use a line search here, although a practical method would require such a globalization strategy.) The complete iteration is given in Table D.1. At the solution,

$$x = \begin{pmatrix} 2.5411 \\ 0.2595 \end{pmatrix}.$$

Table D.1. *Gauss-Newton Iteration (Ideal Data)*

k	$f(x_k)$	$\|\nabla f(x_k)\|$
0	2×10^{0}	3×10^{2}
1	4×10^{-3}	2×10^{1}
2	2×10^{-8}	3×10^{-2}
3	3×10^{-9}	4×10^{-8}
4	3×10^{-9}	3×10^{-13}

Since $f(x) \approx 0$, an approximate global solution has been found to the least-squares problem. (The least-squares objective function cannot be negative.) In general, the Gauss-Newton method is only guaranteed to find a local solution.

For comparison, we now apply Newton's method to the same problem using the same initial guess

$$x = \begin{pmatrix} 2.50 \\ 0.25 \end{pmatrix}.$$

At this point

$$\nabla f(x) = \begin{pmatrix} -16.5888 \\ -300.8722 \end{pmatrix} \quad \text{and} \quad \nabla^2 f(x) = \begin{pmatrix} 78.5367 & 1215.4991 \\ 1215.4991 & 22278.6570 \end{pmatrix}.$$

(This matrix is similar to the matrix used in the Gauss-Newton method.) The search direction is the solution of

$$\nabla^2 f(x) p = -\nabla f(x)$$

so that

$$p = \begin{pmatrix} 0.0142 \\ 0.0127 \end{pmatrix} \quad \text{and} \quad x \leftarrow x + p = \begin{pmatrix} 2.5142 \\ 0.2627 \end{pmatrix}.$$

The complete iteration is in Table D.2. The solution obtained is almost identical to that obtained by the Gauss-Newton method.

We now consider the same model

$$y_i \approx x_1 e^{x_2 t_i}$$

but with the data

$$\begin{aligned} t &= (\, 1 \quad 2 \quad 4 \quad 5 \quad 8 \quad 4.1 \,)^T, \\ y &= (\, 3 \quad 4 \quad 6 \quad 11 \quad 20 \quad 46 \,)^T. \end{aligned}$$

This corresponds to the antelope data of Section D.1, but with an extraneous data point added. (This point is called an *outlier*, since it is inconsistent with the other data points for this model.) In this case the exponential model will not be a good

Table D.2. *Newton Iteration (Ideal Data)*

k	$f(x_k)$	$\|\nabla f(x_k)\|$
0	2×10^{0}	3×10^{2}
1	1×10^{-1}	5×10^{1}
2	2×10^{-4}	9×10^{-1}
3	5×10^{-9}	6×10^{-3}
4	6×10^{-9}	8×10^{-8}
5	3×10^{-9}	1×10^{-12}

fit to the data, so we would expect the performance of the Gauss-Newton method to deteriorate. The runs corresponding to the initial guess

$$x = \begin{pmatrix} 10 \\ 0.1 \end{pmatrix}$$

are given in Table D.3. As expected, the Gauss-Newton method converges slowly. Both methods find the solution

$$x = \begin{pmatrix} 9.0189 \\ 0.1206 \end{pmatrix}.$$

The initial guess was close to the solution, so that the slow convergence of the Gauss-Newton method was not due to a poor initial guess. Also, the final function value is large, indicating that the model cannot fit the data well in this case. This is to be expected given that an outlier is present.

∎

Many other methods for nonlinear least-squares can be interpreted as using some approximation to the second term in the formula for the Hessian matrix

$$\sum_{i=1}^{m} f_i(x) \nabla^2 f_i(x).$$

The oldest and simplest of these approximations is

$$\sum_{i=1}^{m} f_i(x) \nabla^2 f_i(x) \approx \lambda I,$$

where $\lambda \geq 0$ is some scalar. Then the search direction is obtained by solving the linear system

$$\Big[\nabla F(x)\, \nabla F(x)^T + \lambda I\Big] p = -\nabla F(x)\, F(x).$$

This is referred to as the *Levenberg-Marquardt* method.

Table D.3. *Gauss-Newton (left) and Newton Iterations (right); Data Set with Outlier*

k	$f(x_k)$	$\|\nabla f(x_k)\|$
0	601.90	2×10^{2}
1	599.90	8×10^{0}
2	599.67	3×10^{1}
3	599.65	6×10^{0}
4	599.64	2×10^{0}
5	599.64	6×10^{-1}
⋮	⋮	⋮
16	599.64	1×10^{-6}
17	599.64	4×10^{-7}
18	599.64	1×10^{-7}
19	599.64	4×10^{-8}
20	599.64	1×10^{-8}
21	599.64	3×10^{-9}

k	$f(x_k)$	$\|\nabla f(x_k)\|$
0	601.90	2×10^{2}
1	599.64	1×10^{1}
2	599.64	2×10^{-2}
3	599.64	6×10^{-8}
4	599.64	5×10^{-13}

The Levenberg-Marquardt method is often implemented in the context of a trust-region strategy (see Section 11.6). If this is done then the search direction is obtained by minimizing a quadratic model of the objective function (based on the Gauss-Newton approximation to the Hessian)

$$\underset{p}{\text{minimize}}\ Q(p) = f(x) + p^T \nabla F(x)\, F(x) + \tfrac{1}{2} p^T \nabla F(x)\, \nabla F(x)^T p$$

subject to the constraint

$$\|p\| \leq \Delta$$

for some scalar $\Delta > 0$. This gives a step p that satisfies the Levenberg-Marquardt formula for an appropriate $\lambda \geq 0$. The scalar λ is determined indirectly by picking a value of Δ, as is described in Section 11.6. The scalar Δ can be chosen based on the effectiveness of the Gauss-Newton approximation to the Hessian, and this can be easier than choosing λ directly. An example illustrating a trust-region approach can be found in the same Section.

Both the Gauss-Newton and Levenberg-Marquardt methods use an approximation to the Hessian $\nabla^2 f(x)$. If this approximation is not accurate then the methods will converge more slowly than Newton's method; in fact, they will converge at a linear rate.

Other approximations to the Hessian of $f(x)$ are also possible. For example, a quasi-Newton approximation to

$$\sum_{i=1}^{m} f_i(x) \nabla^2 f_i(x)$$

could be used.

There is one other computational detail associated with the Gauss-Newton method that should be mentioned. The formula for the search direction in a Gauss-Newton method

$$\nabla F(x)\,\nabla F(x)^T p = -\nabla F(x)\,F(x)$$

is equivalent to the solution of a linear least-squares problem[19]

$$\underset{p}{\text{minimize}}\ \left\|\nabla F(x)^T p + F(x)\right\|_2^2 = \Big[\nabla F(x)^T p + F(x)\Big]^T \Big[\nabla F(x)^T p + F(x)\Big].$$

If we set the gradient of this function with respect to p equal to zero, we obtain the Gauss-Newton formula. The Gauss-Newton formula corresponds to a system of linear equations called the *normal equations* for the linear least-squares problem. If the normal equations are solved on a computer then the computed search direction will have an error bound proportional to

$$\text{cond}(\nabla F(x)\,\nabla F(x)^T) = \text{cond}(\nabla F(x))^2$$

if the 2-norm is used to define the condition number (see Section A.8). However if the search direction is computed directly as the solution to the linear least-squares problem without explicitly forming the normal equations, then in many cases a better error bound can be derived.

Working directly with the linear least-squares problem is especially important in cases where $\nabla F(x)$ is not of full rank, or is close (in the sense of rounding error) to a matrix that is not of full rank. In this case the matrix

$$\nabla F(x)\,\nabla F(x)^T$$

will be singular or nearly singular, causing difficulties in solving

$$\nabla F(x)\,\nabla F(x)^T p = -\nabla F(x)\,F(x).$$

The corresponding linear least-squares problem is well defined, however, even when $\nabla F(x)$ is not of full rank (although the solution may not be unique). A similar approach can be used in the computation of the search direction in the Levenberg-Marquardt method (see the Exercises).

Exercises

2.1. Let $x = (2, 1)^T$. Calculate

$$\begin{array}{rrr} F(x), & f(x), & \nabla F(x), \\ \nabla f(x), & \nabla F(x)\,\nabla F(x)^T, & \nabla^2 f(x) \end{array}$$

for the antelope model.

[19] A least-squares model is "linear" if the variables x appear linearly. Thus the model $y \approx x_1 + x_2 t^2$ is linear even though it includes a nonlinear function of the independent variable t.

2.2. Consider the least-squares model

$$y \approx x_1 e^{x_2 t} + x_3 + x_4 t.$$

Determine the formulas for $F(x)$, $f(x)$, $\nabla F(x)$, $\nabla f(x)$, $\nabla F(x)\,\nabla F(x)^T$, and $\nabla^2 f(x)$ for a general data set $\{\, t_i, y_i \,\}_{i=1}^m$.

2.3. Verify the results in Example D.2 for the ideal data set.

2.4. Verify the results in Example D.2 for the data set containing the outlier.

2.5. Modify the least-squares model used in Example D.2 for the data set containing the outlier. Use the model

$$y \approx x_1 e^{x_2 t} + x_3 (t - 4.1)^{10}.$$

Apply the Gauss-Newton method with no line search to this problem. Use the initial guess $(2.5, .25, .1)^T$. Does the method converge rapidly? Does this model fit the data well?

2.6. Apply Newton's method and the Gauss-Newton method to the antelope model, using the initial guess $x = (2.5, .25)^T$. Do not use a line search. Terminate the algorithm when the norm of the gradient is less than 10^{-6}. Compute the difference between the Hessian and the Gauss-Newton approximation at the initial and final points.

2.7. Repeat the previous exercise, but use the back-tracking line search described in Section 11.5 with $\mu = 0.1$.

2.8. Prove that the Gauss-Newton method is the same as Newton's method for a linear least-squares problem.

2.9. Consider the formula for the Levenberg-Marquardt direction:

$$\Big[\nabla F(x)\,\nabla F(x)^T + \lambda I\Big] p = -\nabla F(x)\, F(x).$$

Show that p can be computed as the solution to a linear least-squares problem with coefficient matrix

$$\begin{pmatrix} \nabla F(x)^T \\ \sqrt{\lambda} I \end{pmatrix}.$$

Why might it be preferable to compute p this way?

2.10. Consider a nonlinear least-squares problem

$$\min f(x) = F(x)^T F(x)$$

with n variables and $m > n$ nonlinear functions $f_i(x)$, and assume that $\nabla F(x)$ is a full-rank matrix for all values of x. Let p_{GN} be the Gauss-Newton search direction, let $p_{LM}(\lambda)$ be the Levenberg-Marquardt search direction for a particular value of λ, and let p_{SD} be the steepest-descent direction. Prove that

$$\lim_{\lambda \to 0} p_{LM}(\lambda) = p_{GN}$$

and

$$\lim_{\lambda \to \infty} \frac{p_{LM}(\lambda)}{\|p_{LM}(\lambda)\|} = \frac{p_{SD}}{\|p_{SD}\|}.$$

D.3 Statistical Tests

[20]Let us return to the exponential model discussed in Section D.1, and assume that the functions $\{ f_i \}$ are residuals of the form

$$f_i(x_1, x_2) = x_1 e^{x_2 t_i} - y_i.$$

We are assuming that

$$y_i = x_1 e^{x_2 t_i} + \epsilon_i,$$

or in more general terms that

$$[\text{Observation}] = [\text{Model}] + [\text{Error}].$$

If assumptions are made about the behavior of the errors, then it is possible to draw conclusions about the fit. If the model is *linear*, that is, if the variables x appear linearly in the functions $f_i(x)$, then the statistical conclusions are precise. If the model is nonlinear, however, standard techniques only produce linear approximations to exact results.

The assumptions about the errors are not based on the particular set of errors that correspond to the given data set, but rather on the errors that would be obtained if a very large, or even infinite, data set had been collected. For example, it is common to assume that the expected value of the errors is zero, that is, that the data are unbiased. Also, it is common to assume that the errors all have the same variance (perhaps unknown), and that the errors are independent of one another. These assumptions can sometimes be guaranteed by careful data collection, or by transforming the data set in a straightforward manner (see the Exercises).

More is required, however. If we interpret the errors to be random, then we would like to know their underlying probability distribution. We will assume that the errors follow a normal distribution with mean 0 and known variance σ^2. The normal distribution is appropriate in many cases where the data come from measurements (such as measurements made with a ruler or some sort of gauge). In addition, the central limit theorem indicates that, at least as the sample size increases, many other probability distributions can be approximated by a normal distribution.

If the errors are assumed to be normally distributed, then least squares minimization is an appropriate technique for choosing the parameters x. Use of least squares corresponds to maximizing the "likelihood" or probability that the parameters have been chosen correctly, based on the given data. If a different probability distribution were given, then maximizing the likelihood would give rise to a different optimization problem, one involving some other function of the residuals.

If least squares is applied to a linear model, then many properties of the parameters and the resulting fit can be analyzed. For example, one could ask for a "confidence interval" that contained the true value of the parameter x_1, with probability 95%. (Since the errors are random and they are only known in a statistical sense, all of the conclusions that can be drawn will be probabilistic.) Or one could

[20]This Section assumes knowledge of statistics.

ask if the true value of the parameter x_2 were nonzero, with 99% probability. This is referred to as a "hypothesis test." (If the true value of this parameter were zero, then the parameter could be removed from the model.) Or, for some given value of the independent variable t, one could ask for a confidence interval that contained the true value of the model y, again with some probability.

The answers to all of these questions depend on the "variance-covariance matrix." In the case of a linear model it is equal to

$$\sigma^2[\nabla^2 f(x)]^{-1} = \sigma^2[\nabla F(x)\,\nabla F(x)^T]^{-1},$$

where σ^2 is the variance of the errors $\{\,\epsilon_i\,\}$. In this case, it is a constant matrix. For a nonlinear model, $\nabla^2 f(x)$ is not constant, and as a result the calculations required to determine confidence intervals or to do hypothesis testing are more difficult and more computationally expensive. It is possible to apply the same formulas that are used for linear models, using either

$$[\nabla^2 f(x_*)]^{-1} \quad \text{or} \quad [\nabla F(x_*)\,\nabla F(x_*)^T]^{-1}$$

in place of $[\nabla^2 f(x)]^{-1}$, but the resulting intervals and tests are then only approximations, and in some cases the approximations can be poor. Using additional computations, it is possible to analyze a model to determine how nonlinear it is, and hence detect if these linear approximations are effective. If they are not, then alternative techniques can be applied, but again at a cost of additional computations.

Exercises

3.1. Determine the matrices

$$[\nabla^2 f(x_*)]^{-1} \quad \text{and} \quad [\nabla F(x_*)\,\nabla F(x_*)^T]^{-1}$$

for the two data sets in Example D.2. Are the two matrices "close" to each other?

3.2. Prove that the matrices

$$[\nabla^2 f(x_*)]^{-1} \quad \text{and} \quad [\nabla F(x_*)\,\nabla F(x_*)^T]^{-1}$$

are the same for a linear least-squares problem.

3.3. Consider the nonlinear least-squares problem

$$\text{minimize} \sum_{i=1}^{m} f_i(x)^2$$

where each f_i represents a residual with error ϵ_i. Assume that all of the errors are independent, with mean zero. Also assume that the i-th error is normally distributed with known variance σ_i^2, but do not assume that all of these variances are equal. Show how to transform this least-squares problem to one where all the errors have the same variance.

D.4 Orthogonal Distance Regression

So far, we have assumed that the data-fitting errors were of the form

$$y = x_1 e^{x_2 t} + \epsilon.$$

That is, either the model is incomplete (there are additional terms or parameters that are being ignored) or the observations contain errors, perhaps due to inaccurate measurements. But it is also possible that there are errors in the independent variable t. For example, the measurements might be recorded as having been taken once an hour at exactly thirty minutes past the hour, but in fact were taken sometime between twenty-five and thirty-five minutes past the hour.

If this is true—if we believe that the independent variable is subject to error—we should use a model of the form

$$y = x_1 e^{x_2 (t+\delta)} + \epsilon.$$

If we assume that the errors δ are normally distributed with mean 0 and constant variance, then it is appropriate to use a least-squares technique to estimate the parameters x and δ:

$$\underset{x,\delta}{\text{minimize}} \; f(x_1, x_2) = \sum_{i=1}^{5} [(x_1 e^{x_2 (t_i + \delta_i)} - y_i)^2 + \delta_i^2] = \sum_{i=1}^{5} [\epsilon_i^2 + \delta_i^2].$$

More generally, if our original least-squares problem were of the form

$$\underset{x}{\text{minimize}} \; f(x) = \sum_{i=1}^{m} f_i(x; t_i)^2,$$

where t_i is a vector of independent variables that is subject to error, then the revised least-squares problem could be written as

$$\underset{x,\delta}{\text{minimize}} \; f(x) = \sum_{i=1}^{m} f_i(x; t_i + \delta_i)^2 + \|\delta\|_2^2 .$$

One might also consider scaling $\|\delta\|$ by some constant to reflect any difference in the variances of the two types of errors.

Problems of this type are sometimes called *errors in variables* models, for obvious reasons, but we will use the term *orthogonal distance regression*. To explain this, consider again the graph of our antelope data (see Figure D.2). If we assume that all the error is in the model or the observation, then the residual measures the *vertical* distance between the data point and the curve. However if there is also error in the independent variable then geometrically we are minimizing the *orthogonal* distance between the data point and the curve.

If the model is changing rapidly—as in an exponential model when the exponent is large, or near a singularity in a model—the vertical distance can be large even though the orthogonal distance is small. A data point in such a region can easily have a large vertical residual and thus can exert extraordinary influence in the ordinary least-squares model, perhaps to the extent that the parameter estimate is strongly influenced by that single data point. Using orthogonal distance regression can alleviate this form of difficulty.

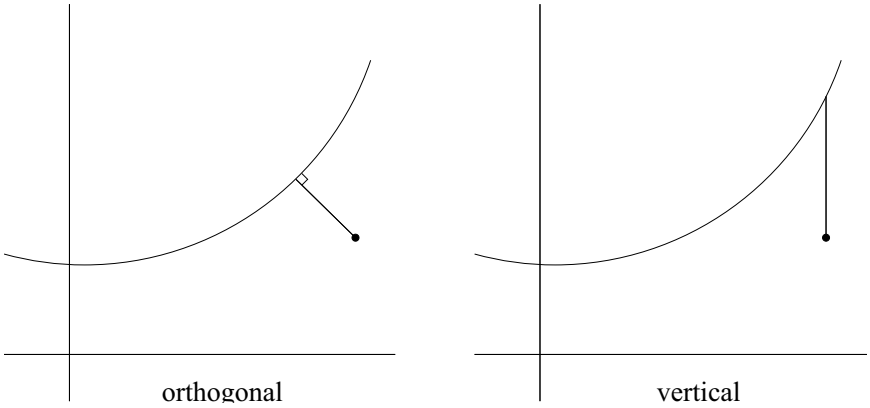

Figure D.2. *Orthogonal and Vertical Distances*

Exercises

4.1. Consider the function $f(x) = 1/x$. Let $\epsilon = 10^{-2}$ and $x = \frac{3}{2}\epsilon$. Determine the vertical and orthogonal distance from the point $(x - \epsilon, 1/x)^T$ to the graph of f.

4.2. Consider the ideal data set in Example D.2. Apply Newton's method (or any other minimization method) to solve

$$\underset{x,\delta}{\text{minimize}}\ f(x) = \sum_{i=1}^{m} f_i(x; t_i + \delta_i)^2 + \|\delta\|_2^2 .$$

Compare your solution to that obtained in Example D.2.

4.3. Repeat the previous exercise for the data set in Example D.2 containing an outlier.

D.5 Notes

Gauss-Newton Method—We believe that the Gauss-Newton method was invented by Gauss; it is described in his 1809 book on planetary motion. A more modern discussion can be found, for example, in the paper by Dennis (1977). The Levenberg-Marquardt method was first described in the papers of Levenberg (1944) and Marquardt (1963). Its implementation in terms of a trust-region method is described in the paper of Moré (1977). Other approximations to the Hessian of the least-squares problem are described in the papers by Ruhe (1979), Gill and Murray (1978), and Dennis, Gay, and Welsch (1981), and in the book by Bates and Watts (1988). The computational difficulties associated with the normal equations are described in the book by Golub and Van Loan (1996).

In many applications, only a subset of the parameters will occur nonlinearly. In such cases, efficient algorithms are available that treat the linear parameters in

a special way. For further information, see the papers by Golub and Pereyra (1973) and Kaufman (1975). Software implementing these techniques is available from NETLIB.

Statistical Tests—A brief introduction to statistical techniques in nonlinear least-squares data fitting can be found in the article by Watts and Bates (1985). For a more extensive discussion see the book by Bard (1974). A comparison of various techniques for computing confidence intervals can be found in the paper by Donaldson and Schnabel (1987).

Orthogonal Distance Regression—An extensive discussion of orthogonal distance regression can be found in the book by Fuller (1987). Algorithms and software for orthogonal distance regression are closely related to those used for ordinary least-squares regression. For a discussion of these techniques, see the paper by Boggs and Rogers (1990). In the linear case, orthogonal distance regression is often referred to as "total least squares." For information about this case, see the book by Van Huffel and Vandewalle (1991).

D.5.1 References

- Y. BARD, *Nonlinear Parameter Estimation*, Academic Press, New York, 1974.
- D.M. BATES AND D.G. WATTS, *Nonlinear Regression Analysis and its Applications*, Wiley, New York, 1988.
- PAUL T. BOGGS AND JANET E. ROGERS, *Orthogonal Distance Regression*, Contemporary Mathematics, 112 (1990), pp. 183–194
- J.E. DENNIS, JR., *Nonlinear least squares*, in *The State of the Art in Numerical Analysis*, D. Jacobs, editor, Academic Press (New York), pp. 269–312, 1977.
- J.E. DENNIS, JR., D.M. GAY, AND R.E. WELSCH, *An adaptive nonlinear least-squares algorithm*, ACM Transactions on Mathematical Software, 7 (1981), pp. 348–368
- JANET R. DONALDSON AND ROBERT B. SCHNABEL, *Computational experience with confidence regions and confidence intervals for nonlinear least squares*, Technometrics, 29 (1987), pp. 67–82
- W.A. FULLER, *Measurement Error Models*, John Wiley and Sons, New York, 1987.
- C.F. GAUSS, *Theoria Motus Corporum Cœlestium in Sectionibus Conicus Solem Ambientum*, Dover, New York, 1809.
- P.E. GILL AND W. MURRAY, *Algorithms for the solution of the nonlinear least-squares problem*, SIAM Journal on Numerical Analysis, 15 (1978), pp. 977–992
- GENE H. GOLUB AND VICTOR PEREYRA, *The differentiation of pseudo-inverses and nonlinear least-squares problems whose variables separate*, SIAM Journal on Numerical Analysis, 10 (1973), pp. 413–432
- GENE H. GOLUB AND C. VAN LOAN, *Matrix Computations (third edition)*, The Johns Hopkins University Press, Baltimore, 1996.

- Linda Kaufman, *A variable projection method for solving separable nonlinear least squares problems*, BIT, 15 (1975), pp. 49–57
- K. Levenberg, *A method for the solution of certain problems in least squares*, Quarterly of Applied Mathematics, 2 (1944), pp. 164–168
- D. Marquardt, *An algorithm for least-squares estimation of nonlinear parameters*, SIAM Journal of Applied Mathematics, 11 (1963), pp. 431–441
- J.J. Moré, *The Levenberg-Marquardt algorithm: implementation and theory*, in *Numerical Analysis*, G. A. Watson, editor, Lecture Notes in Mathematics 630, Springer-Verlag (Berlin), pp. 105–116, 1977.
- A. Ruhe, *Accelerated Gauss-Newton algorithms for nonlinear least-squares problems*, SIAM Review, 22 (1980), pp. 318–337
- S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, SIAM, Philadelphia, 1991.
- D.G. Watts and D.M. Bates, *Nonlinear regression*, in the *Encyclopedia of Statistical Sciences*, Vol. 6, Samuel Kotz and Normal L. Johnson, editors, John Wiley and Sons (New York), p. 306–312, 1985.