Chapter 7

Frequentist Techniques for Parameter Estimation

The differential equation models of Section 3.2 can be classified as ODE systems

$$\begin{aligned} &\frac{du}{dt} = g(t, u(t), q) \ , \ u(t_0) = u_0 \ , \ u(t,q) \in \mathbb{R}^N \\ &y(t,q) = \mathcal{C}u(t,q) \ , \ \mathcal{C} \in \mathbb{R}^{\nu \times N}, \end{aligned} \tag{7.1}$$

stationary PDE

$$\begin{aligned} \mathcal{N}(u,q) &= F(q) \quad , \quad x \in \mathcal{D} \\ B(u,q) &= G(q) \quad , \quad x \in \partial\mathcal{D} \\ y(x,q) &= \mathcal{C}u(x,q), \end{aligned} \tag{7.2}$$

or evolutionary PDE

$$\begin{aligned} &\frac{\partial u}{\partial t} = \mathcal{N}(u,q) + F(q) && , \ x \in \mathcal{D}, \ t \in [t_0, \infty) \\ &B(u,q) = G(q) && , \ x \in \partial\mathcal{D}, \ t \in [t_0, \infty) \\ &u(t_0, x, q) = I(q) && , \ x \in \mathcal{D}. \end{aligned} \tag{7.3}$$

Here y and q denote observations and parameters and $\mathcal{N}, F, B$ and G denote differential operators, source terms and boundary conditions.

Additionally, we considered algebraic models

$$A(q)u = F(q). \tag{7.4}$$

If $A(q) \in \mathbb{R}^{n \times n}$ is invertible, we can represent the n observations by

$$y(q) = u(q) = A^{-1}(q)F(q). \tag{7.5}$$

Linear regression is a special case in which the parameter dependency is linear so

$$y(q) = Xq.$$

For $q \in \mathbb{R}^p$, $X \in \mathbb{R}^{n \times p}$ is termed the design matrix.

In the statistics and inverse problems literature, the observed model response or quantity of interest is often formulated as

$$y = f(\chi, q) \tag{7.6}$$

where χ are independent variables – e.g., t or x – or other known inputs. In the statistics literature, χ are also referred to as explanatory or regressor variables. The function f generically denotes the map from the independent variables and parameters to the response. We assume that f is fixed and known in the sense that there exists a unique modeled response. For nonlinear, ODE and PDE and algebraic models, however, one can rarely obtain analytic solutions and hence explicit formulations for f. Hence for most problems, we rely on numerical approximations for f.

Throughout our discussion, we assume that we have observations (χ_i, υ_i), $i = 1, \cdots, n$, where the measured quantity of interest υ_i is corrupted by measurement errors ϵ_i so that

$$\upsilon_i = f(\chi_i, q) + \epsilon_i, \quad , \ i = 1, \cdots, n. \tag{7.7}$$

The mathematical inverse problem associated with parameter estimation can then be formulated as follows: given these noisy measurements, determine q in a stable manner. The associated statistical inverse problem – sometimes referred to as *inverse uncertainty quantification* – is to additionally quantify uncertainties associated with q due to the measurement errors. The assumptions required to approximate q and quantify its uncertainty define frequentist and Bayesian techniques for parameter estimation.

For sensitivity analysis and uncertainty propagation, the specific role of the independent variables are of secondary importance and we are instead interested how the model solution varies as a function of the parameters or inputs q. This is facilitated by the representation

$$\upsilon_i = f_i(q) + \epsilon_i \quad , \ i = 1, \cdots, n. \tag{7.8}$$

where $f_i(q) \in \mathbb{R}^\nu$ denotes the observed model response and $\upsilon_i \in \mathbb{R}^\nu$ again denotes measured data. For the models (7.1), (7.2) and (7.4), the model response can be expressed as $n \times \nu$ vector

$$\begin{aligned} f(q) &= [f(t_1, q), \cdots, f(t_n, q)]^T && , \ \text{Evolution Processes} \\ f(q) &= [f(x_1, q), \cdots, f(x_n, q)]^T && , \ \text{Stationary Processes} \\ f(q) &= [f_1(q), \cdots, f_n(q)]^T && , \ \text{Algebraic Models.} \end{aligned} \tag{7.9}$$

Hence the dependence of the observed model response on the independent or regressor variables is suppressed in the notation $f(q)$.

For evolution models, we will have $\nu \geq 1$ experimental measurements and model responses at each time $t_j, j = 1, \cdots, n$. For stationary processes and algebraic models, we consider scalar measurements and model evaluations so $\nu = 1$.

7.1 Parameter Estimation from a Frequentist Perspective

We recall George E.P. Box's quote "Essentially, all models are wrong, but some are useful," page 424 of [38]. Thus the mathematical models will exhibit model errors, which we collectively denote by vector $\delta = [\delta_1, \cdots, \delta_n]^T$, along with measurement errors. To accommodate these errors, we consider statistical models of the form

$$\Upsilon = f(q_0) + \delta + \varepsilon \tag{7.10}$$

where $\Upsilon = [\Upsilon_1, \cdots, \Upsilon_n]^T$ is a random vector whose realization $\upsilon = [\upsilon_1, \cdots, \upsilon_n]^T$ is comprised of measurements from an experiment. Measurement errors are represented by the random vector $\varepsilon = [\varepsilon_1, \cdots, \varepsilon_n]^T$ and errors resulting for a specific experiment are denoted by $\epsilon = [\epsilon_1, \cdots, \epsilon_n]^T$.

As detailed in Section 4.8.1, a basic tenet of frequentist inference is the assumption that parameters are fixed but possibly unknown. Hence q_0 represents the true but unknown value of the parameter set that generated the observations $\upsilon = [\upsilon_1, \cdots, \upsilon_n]^T$. We emphasize that since q_0 is not a random vector, the model response $f(q_0)$ is a deterministic quantity.

If the quantification of modeling errors constitutes one of the goals, then it is necessary to consider the statistical model (7.10) and characterize the modeling errors in an efficient and statistically consistent manner as detailed in Chapter 12. For many applications, however, the modeling and measurement errors can be collectively quantified by the random vector ε, in which case, one would employ the statistical model

$$\Upsilon = f(q_0) + \varepsilon \tag{7.11}$$

in which errors are additive.

To construct likelihoods in the manner detailed in Section 4.3, we typically assume that the random variables ε_i are unbiased and iid which is often not the case if they are comprised of both modeling and measurement errors. For example, we illustrate in Chapter 12 that residuals for a structural model are highly dependent on the magnitude of y even though the model is providing an accurate fit to measured data. Hence for some applications, the statistical model

$$\Upsilon_i = f_i(q_0)(1 + \varepsilon_i) \ , \ j = 1, \cdots, n, \tag{7.12}$$

with multiplicative errors may be more appropriate since $\text{var}(\Upsilon_i)$ will depend on the magnitude of $f_i(q_0)$.

The goal when calibrating models is to determine parameter estimates q so that the model response $f(q)$ fits the data in some optimal sense. We showed in Section 4.3 that this can be achieved by constructing an estimator $\hat{q}$ that estimates q_0 in a statistically reasonable manner.[1] It was demonstrated that ordinary least squares (OLS) estimators

$$\hat{q}_{OLS} = \underset{q \in \mathcal{Q}}{\text{argmin}} \sum_{i=1}^{n} [\Upsilon_i - f_i(q)]^2 \tag{7.13}$$

[1] The notation $\hat{q}$ for the estimator is not universal and many texts denote the estimate by $\hat{q}$. Hence care must be taken to establish the convention employed in a specific text.

and maximum likelihood estimators (MLE) both achieve this goal and are equivalent for certain assumptions regarding the distribution of errors ε_i.

Remark 7.1. Because the estimator $\hat{q}$ is a random variable or random vector, it has a mean, covariance and distribution termed the sampling distribution. We will show that with appropriate assumptions regarding the distribution of ε_i, $\mathbb{E}(\hat{q}) = q_0$ and the covariance will quantify the variability of the errors. Furthermore, confidence limits for the sampling distribution can be used to *quantify the accuracy of the estimation process.*

What the sampling distribution *does not do* is provide a distribution for the model parameters since q_0 is not a random variable in frequentist inference. We will illustrate that, for certain problems, the sampling distribution coincides with the parameter distribution constructed using Bayesian techniques. This makes it tempting to propagate the sampling distribution through the model, using the techniques of Chapters 9 and 10, to quantify the model or response uncertainty. However, this is problematic for two reasons. The first is that there is no convergence theory specifying an asymptotic relation between the sampling distribution and parameter distribution which relies on Bayesian assumptions. Secondly, the sampling distribution is Gaussian which limits its accuracy for quantifying non-Gaussian parameter distributions. Hence this approach should be avoided unless additional analysis indicates an equivalence between the two distributions.

There are two alternatives. From a frequentist perspective, one can assume parametric forms (e.g., Gaussian or Johnson distributions) for the densities associated with model parameters and estimate the augmented parameter set using moment or distribution matching techniques [154,155,240]. For model responses of the form (7.3), this requires that errors ε_i be characterized from independent experiments. We do not provide further details about this approach but rather refer the reader to the cited references. Alternatively, the Bayesian techniques detailed in Chapter 8 can be used to construct parameter densities and moments that can be directly propagated through models.

The estimators $\hat{q}$ can be determined explicitly only for linear parameter dependencies. Whereas applications such as convolution models for acoustics or image processing and X-ray tomography yield linearly parameterized models, general models typically exhibit a nonlinear dependence on q. To illustrate the derivation of relevant theory, we consider the linear regression (linear parameterization) problem first in Section 7.2. We return to the general problem posed here in Section 7.3.

7.2 Linear Regression

We illustrate here fundamental results regarding linear regression to motivate corresponding theory for the nonlinear least squares problem (7.13). Additional details can be found in [95].

We consider the statistical model

$$\Upsilon = Xq_0 + \varepsilon \tag{7.14}$$

where $\Upsilon = [\Upsilon_1, \cdots, \Upsilon_n]^T$ and $\varepsilon = [\varepsilon_1, \cdots, \varepsilon_n]^T$ are random vectors and the $n \times p$ design matrix X is considered deterministic and known. We let q_0 denote the vector of true but unknown parameters and let $\upsilon = [\upsilon_1, \cdots, \upsilon_n]^T$ denote realizations or observations from an experiment in which the realized errors are $\epsilon = [\epsilon_1, \cdots, \epsilon_n]$. Throughout this discussion, we assume that there are more measurements than parameters so that $n > p$.

Assumption 7.2. We make the assumption that errors are unbiased and independent and identically distributed (iid) with variance σ_0^2; hence for $j = 1, \cdots, n$,

$$\begin{aligned}&\text{(i) } \mathbb{E}(\varepsilon_i) = 0\\ &\text{(ii) } \text{var}(\varepsilon_i) = \sigma_0^2\ ,\ \text{cov}(\varepsilon_i, \varepsilon_j) = 0 \text{ for } i \neq j.\end{aligned} \tag{7.15}$$

In accordance with frequentist assumptions, the error variance σ_0^2 is assumed fixed but unknown. At this point, we make no additional assumptions regarding the error distribution.

Our first objective is to construct unbiased estimators $\hat{q}$ and $\hat{\sigma}^2$ for the unknown parameters q_0 and σ_0^2.

7.2.1 Parameter Estimator and Estimate

To construct an estimator $\hat{q}$ for q_0, we seek q that minimizes the ordinary least squares functional

$$\mathcal{J}(q) = (\Upsilon - Xq)^T(\Upsilon - Xq). \tag{7.16}$$

If (7.16) were scalar-valued, we would optimize it by setting the derivative with respect to q equal to 0 and solving for q. For vector-valued problems, this is achieved using the gradient $\nabla_q \mathcal{J}$ of $\mathcal{J}$ with respect to q. Specifically, one sets

$$\nabla_q \mathcal{J} = 2[\nabla_q(\Upsilon - Xq)^T][\Upsilon - Xq] = 0,$$

where

$$\nabla_q(\Upsilon - Xq)^T = -\nabla_q q^T X^T = -X^T,$$

to obtain the least squares estimator

$$\hat{q}_{OLS} = (X^T X)^{-1} X^T \Upsilon. \tag{7.17}$$

The realization

$$q_{OLS} = (X^T X)^{-1} X^T \upsilon \tag{7.18}$$

is the least squares estimate for the unknown true parameter q_0.

Remark 7.3. Throughout this chapter, we will discuss only OLS estimators and estimates. Hence to simplify notation, we will drop the subscript OLS and let $\hat{q} = \hat{q}_{OLS}$ and $q = q_{OLS}$ denote the least squares estimator and estimate.

Whereas the normal equations (7.17) provide an analytic minimum for (7.16), they are typically ill-conditioned for moderate to large numbers of parameters. Hence in practice, it is often numerically advantageous to solve the minimization problem (7.16) to avoid inaccurate results associated with numerically solving ill-conditioned linear systems.

7.2.2 Parameter Estimator Properties

Result 7.4. The parameter estimator $\hat{q}$ has the mean and covariance matrix

$$\begin{aligned} &\text{(i) } \mathbb{E}(\hat{q}) = q_0 \\ &\text{(ii) } V(\hat{q}) = \sigma_0^2 (X^T X)^{-1}. \end{aligned} \tag{7.19}$$

Relation (i) follows directly from (7.17) since

$$\mathbb{E}(\hat{q}) = \mathbb{E}[(X^T X)^{-1} X^T \Upsilon] = (X^T X)^{-1} X^T \mathbb{E}(\Upsilon) = q_0.$$

Hence q provides an unbiased estimate for the true parameter. To establish the covariance relation, we let $A = (X^T X)^{-1} X^T$ and note that

$$\begin{aligned} V(\hat{q}) = &\ \mathbb{E}[(\hat{q} - q_0)(\hat{q} - q_0)^T] \\ = &\ \mathbb{E}[(q_0 + A\varepsilon - q_0)(q_0 + A\varepsilon - q_0)^T] \ , \text{ since } \hat{q} = A\Upsilon = A(Xq_0 + \varepsilon) \\ = &\ A\mathbb{E}(\varepsilon\varepsilon^T)A^T \\ = &\ \sigma_0^2 (X^T X)^{-1}. \end{aligned}$$

As noted previously, the error variance σ_0^2 is assumed to be fixed but unknown. Hence to employ (7.19) to estimate the parameter covariance, we must construct an unbiased estimator $\hat{\sigma}^2$ for σ_0^2.

7.2.3 Error Variance Estimator

Result 7.5. The unbiased error covariance estimator is

$$\hat{\sigma}^2 = \frac{1}{n-p} \widehat{R}^T \widehat{R} \tag{7.20}$$

where

$$\widehat{R} = \Upsilon - X\hat{q} \tag{7.21}$$

denotes the residual estimator.

To obtain this result, we first note that the residual can be expressed as

$$\widehat{R} = (I_n - H)\Upsilon$$

where I_n denotes the $n \times n$ identity matrix and

$$H \equiv X(X^T X)^{-1} X^T.$$

It is straightforward to show that H satisfies the properties

$$\begin{aligned} &H^T = H \quad \text{(Symmetric)},\\ &H^2 = H \quad \text{(Idempotent)},\\ &(I_n - H)^2 = I_n - H,\\ &(I_n - H)X = 0. \end{aligned} \tag{7.22}$$

From (7.14) and (7.22), it follows that

$$\widehat{R} = (I_n - H)\varepsilon$$

so that

$$\widehat{R}^T\widehat{R} = \varepsilon^T(I_n - H)\varepsilon. \tag{7.23}$$

If we generically denote the ij entry of $I_n - H$ by h_{ij}, the quadratic form (7.23) can be expressed as

$$\widehat{R}^T\widehat{R} = \sum_{i=1}^n\sum_{j=1}^n h_{ij}\varepsilon_i\varepsilon_j.$$

It then follows that

$$\begin{aligned} \mathbb{E}(\widehat{R}^T\widehat{R}) &= \sum_{i=1}^n\sum_{j=1}^n h_{ij}\mathbb{E}(\varepsilon_i\varepsilon_j)\\ &= \sum_{i=1}^n\sum_{j=1}^n h_{ij}\text{cov}(\varepsilon_i,\varepsilon_j) \quad \text{, follows from (4.14) with } \mathbb{E}(\varepsilon_j) = \mathbb{E}(\varepsilon_i) = 0\\ &= \sum_{i=1}^n h_{ii}\text{var}(\varepsilon_i) \quad \text{, } \varepsilon_i \text{ independent}\\ &= \sigma_0^2\text{tr}(I_n - H) \quad \text{, } \varepsilon \text{ identically distributed with variance } \sigma_0^2. \end{aligned}$$

Since the trace operator satisfies the properties $\text{tr}(A + B) = \text{tr}(A) + \text{tr}(B)$ and $\text{tr}(AB) = \text{tr}(BA)$, it follows that

$$\begin{aligned} \text{tr}(I_n - H) &= n - \text{tr}[X(X^TX)^{-1}X^T]\\ &= n - \text{tr}[(X^TX)^{-1}X^TX]\\ &= n - p. \end{aligned} \tag{7.24}$$

Thus $\hat{\sigma}^2 = \frac{1}{n-p}\widehat{R}^T\widehat{R}$ is an unbiased estimator for σ_0^2. Furthermore, we can conclude from (7.24) that the eigenvalues of H are 0 or 1.

Example 7.6. Consider the height-weight data from the *1975 World Almanac and Book Facts* that is compiled in Table 7.1. To model this data, we employ the quadratic relation

$$\Upsilon_i = q_1 + q_2(x_i/12) + q_3(x_i/12)^2 + \varepsilon_i \tag{7.25}$$

Height (in)	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72
Weight (lbs)	115	117	120	123	126	129	132	135	139	142	146	150	154	159	164

Table 7.1. *Height-weight data from the 1975 World Almanac and Book Facts.*

where x_i is the height in inches and Υ_i is the corresponding weight. Solution of the normal equations (7.18) yields the parameter values $q = [261.88, -88.18, 11.96]^T$. We note that the conditioning of the 3×3 matrix $X^T X$ is 6.7×10^7 thus illustrating the ill-conditioning of the normal equations. The variance estimate provided by (7.20) is $\sigma^2 = 0.15$ which yields the covariance matrix estimate

$$V = \begin{bmatrix} 634.88 & -235.04 & 21.66 \\ -235.04 & 87.09 & -8.03 \\ 21.66 & -8.03 & 0.74 \end{bmatrix}.$$

The estimated parameter values, plus and minus two standard deviations, are thus

$$\begin{aligned} q_1 &= 261.88 \pm 50.39 & & q_1 \in [211.48, 312.27] \\ q_2 &= -88.18 \pm 18.66 & \Rightarrow \quad & q_2 \in [-106.84, -69.51] \\ q_3 &= 11.96 \pm 1.72 & & q_3 \in [10.24, 13.68]. \end{aligned} \tag{7.26}$$

7.2.4 Sampling Distribution for $\hat{q}$

As detailed in Section 4.2, the estimator $\hat{q}$ has a distribution, termed the sampling distribution, which we will use to construct confidence intervals for the estimation process. The assumptions required to specify a sampling distribution are more stringent than those in Assumption 7.2 and require either that errors are normally distributed or that samples are sufficiently large that the central limit theorem can be invoked for averaged error relations.

Assumption 7.7. The sampling distribution for $\hat{q}$ can be directly specified for problems in which errors are iid and $\varepsilon_i \sim N(0, \sigma_0^2)$ where σ_0 is fixed, but likely unknown.

Property 7.8 (Sampling Distribution for $\hat{q}$). With Assumption 7.7, $\hat{q}$ has the sampling distribution $\hat{q} \sim N(q_0, \sigma_0^2 (X^T X)^{-1})$. Furthermore, if we let δ_k denote the k^{th} diagonal element of $(X^T X)^{-1}$ and q_{0_k} denote the k^{th} element of the true parameter vector q_0, then $\hat{q}_k \sim N(q_{0_k}, \sigma_0^2 \delta_k)$.

To verify this property, we note from [95] that because each component $\hat{q}_k$ is the linear combination of independent random variables Υ_k, it follows that $\hat{q}$ has a joint multivariate normal distribution. When combined with the fact that $\mathbb{E}(\hat{q}) = q_0$ and $\text{cov}(\hat{q}) = \sigma_0^2 (X^T X)^{-1}$, it follows that $\hat{q} \sim N(q_0, \sigma_0^2 (X^T X)^{-1})$.

For numerous applications, errors may be iid with variance σ_0^2 but not normally distributed. For sufficiently large sample sizes, asymptotic theory yields a result similar to Property 7.8.

Property 7.9 (Asymptotic Sampling Distribution for $\hat{q}$). Consider the model (7.14) with errors which are iid with variance σ_0^2. For sufficiently large n, the sampling distribution for $\hat{q}$ is asymptotically normal which we denote by $\hat{q} \overset{a}{\sim} N(q_0, \sigma_0^2(X^TX)^{-1})$.

Rather than provide a complete proof of Property 7.9, we instead summarize the approach and refer the reader to [216] for additional details. We first note that substitution of (7.14) into (7.17) yields $\hat{q} - q_0 = (X^TX)^{-1}X^T\varepsilon$ so that

$$\sqrt{n}(\hat{q} - q_0) = \left(\frac{1}{n}X^TX\right)^{-1} \frac{1}{\sqrt{n}}X^T\varepsilon.$$

Because the first right-hand side term can be interpreted as an average, the law of large numbers is used to establish that

$$\frac{1}{n}X^TX \xrightarrow{P} \mathcal{Y}$$

where $\mathcal{Y}$ is positive definite. Since $\mathbb{E}(\frac{1}{\sqrt{n}}X^T\varepsilon) = 0$, it follows that

$$\text{var}\left(\frac{1}{\sqrt{n}}X^T\varepsilon\right) = \mathbb{E}\left(\frac{1}{n}X^T\varepsilon\varepsilon^TX\right) \xrightarrow{P} \sigma_0^2\mathcal{Y}.$$

The central limit theorem, discussed in Section 4.4, is then invoked to establish that

$$\frac{1}{\sqrt{n}}X^T\varepsilon \xrightarrow{D} Z$$

where $Z \sim N(0, \sigma_0^2\mathcal{Y})$, so that $\sqrt{n}(\hat{q} - q_0) \overset{a}{\sim} N(0, \sigma_0^2\mathcal{Y}^{-1})$. Finally, one shows that $\frac{1}{n}X^TX$ is a strongly consistent estimator of $\mathcal{Y}$ to obtain the asymptotic result.

An obvious practical question concerns the size n required to justify using these asymptotic results. This is problem dependent and alternative methods, such as Bayesian analysis, may be required to establish the normality of distributions when sample sizes are small.

Confidence Intervals

It was shown in Section 3.3 that chi-squared and t-distributions are required to construct confidence intervals. This is established for our estimators in the next two properties.

Property 7.10. For $\hat{\sigma}^2$ given by (7.20), the random variable $\nu = \frac{(n-p)\hat{\sigma}^2}{\sigma_0^2}$ has a chi-squared distribution with $n - p$ degrees of freedom.

To establish this, we note that

$$\begin{aligned}\frac{(n-p)\hat{\sigma}^2}{\sigma_0^2} &= \frac{1}{\sigma_0^2}\widehat{R}^T\widehat{R}\\ &= \frac{1}{\sigma_0^2}\varepsilon^T(I_n - H)\varepsilon\\ &= \frac{1}{\sigma_0^2}\left\langle \varepsilon, U\Lambda U^T\varepsilon\right\rangle \qquad , \ I_n - H = U\Lambda U^T \text{ since symmetric}\\ &= \frac{1}{\sigma_0^2}\left\langle U^T\varepsilon, \Lambda U^T\varepsilon\right\rangle .\end{aligned}$$

Since $\mathrm{tr}(I_n - H) = \mathrm{rank}(I_n - H) = n-p$, we can express Λ as

$$\Lambda = \left[\begin{array}{cc} I_{n-p} & 0\\ 0 & 0\end{array}\right]$$

where I_{n-p} is the $n-p$ identity matrix. Moreover, it is proven in [95] that since U^T is an orthogonal matrix and $\varepsilon \sim N(0,\sigma_0^2)$, then $u = U^T\varepsilon$ is a vector of $N(0,\sigma_0^2)$ random variables. Because

$$\nu = \frac{(n-p)\hat{\sigma}^2}{\sigma_0^2} = \frac{\langle u, \Lambda u\rangle}{\sigma_0^2} = \sum_{i=1}^{n-p}\frac{u_i^2}{\sigma_0^2}$$

is the sum of squares of $n-p$ independent $N(0,1)$ random variables, it thus has a chi-squared distribution with $n-p$ degrees of freedom.

Property 7.11. The random variable

$$T_k = \frac{\hat{q}_k - q_{0_k}}{\hat{\sigma}\sqrt{\delta_k}}$$

has a t-distribution with $n-p$ degrees of freedom.

To verify Property 7.11, we note from Property 7.8 that $Z = \frac{\hat{q}_k - q_{0_k}}{\sigma_0\sqrt{\delta_k}} \sim N(0,1)$. It then follows from Definition 4.12 that

$$\begin{aligned}T_k &= \frac{\hat{q}_k - q_{0_k}}{\hat{\sigma}\sqrt{\delta_k}}\\ &= \frac{\hat{q}_k - q_{0_k}}{\sigma_0\sqrt{\delta_k}}\frac{\sigma_0}{\hat{\sigma}\sqrt{n-p}}\cdot\sqrt{n-p}\\ &= \frac{Z}{\sqrt{\nu/(n-p)}} \quad , \ Z \sim N(0,1) \quad , \ \nu \sim \chi^2(n-p),\end{aligned}$$

has a t-distribution with $n-p$ degrees of freedom.

To construct a $(1-\alpha)\times 100\%$ confidence interval, we employ the techniques of Example 4.33, with $T_k = \frac{\hat{q}_k - q_{0_k}}{\hat{\sigma}\sqrt{\delta_k}}$, to obtain

$$P\left(\hat{q}_k - t_{n-p,1-\alpha/2}\cdot\hat{\sigma}\sqrt{\delta_k} < q_{0_k} < \hat{q}_k + t_{n-p,1-\alpha/2}\cdot\hat{\sigma}\sqrt{\delta_k}\right) = 1-\alpha.$$

We then employ the parameter estimate $q = (X^TX)^{-1}X^T\upsilon$ and variance estimate $\sigma^2 = \frac{1}{n-p}R^TR$, where $R = \upsilon - Xq$, to obtain

$$\left[q_k - t_{n-p,1-\alpha/2} \cdot \sigma\sqrt{\delta_k}, q_k + t_{n-p,1-\alpha/2} \cdot \sigma\sqrt{\delta_k}\right]. \tag{7.27}$$

We note that this is often expressed as

$$\left[q_k - t_{n-p,1-\alpha/2} \cdot SE_k, q_k + t_{n-p,1-\alpha/2} \cdot SE_k\right] \tag{7.28}$$

where $SE_k \equiv \sigma\sqrt{\delta_k}$ is termed the *standard error.* To construct (7.27) or (7.28), one uses a table of t-distributions or t-value calculator to look up or compute values of $t_{n-p,1-\alpha/2}$ for specified values of n, p and α where $n-p$ is the *degrees of freedom.* We caution the reader that whereas most tables are compiled in terms of one tail $(1-\alpha/2)$, some provide values for both tails $(1-\alpha)$. Hence care must be taken to employ α consistent with the table.

Example 7.12. We revisit Example 7.6 and use the t-distribution to construct 90% confidence intervals for the parameters q_1, q_2 and q_3 in the quadratic model (7.25). Here we have $n = 15$ observations and $p = 3$ parameters. For $\alpha = 0.05$, we obtain the value $t_{n-p,1-\alpha/2} = 2.2$ from a table of t-values. This yields the 95% confidence intervals

$$\begin{aligned} q_1 &\in [206.45, 317.31] \\ q_2 &\in [-108.71, -67.65] \\ q_3 &\in [10.07, 13.86]. \end{aligned}$$

These intervals are slightly larger than those in (7.26) for two reasons: the intervals in (7.26) reflect $2\sigma \approx 94.45\%$ confidence intervals and the t-distribution has heavier tails than the normal distribution as illustrated in Figure 4.3(b).

The statistical model, estimators, and statistical properties of the linear regression model are summarized in Table 7.2. This provides motivation and a basis for comparison for the nonlinear theory summarized in the next section.

7.3 Nonlinear Parameter Estimation Problem

We return to the evolutionary process model (7.1), stationary process model (7.2) and algebraic model (7.4), which exhibit nonlinear parameter dependencies, along with the associated statistical model

$$\Upsilon = f(q_0) + \varepsilon. \tag{7.29}$$

The model responses $f(q_0)$ for the three regimes are summarized in (7.3). As before, we take $q \in \mathbb{R}^p$ and let q_0 designate the true but unknown parameter that generates the response $\upsilon \in \mathbb{R}^n$. As in Section 7.2, we assume that there are more measurements than parameters so that $n > p$. We let $\mathbb{Q}$ denote the admissible parameter space and $\mathcal{Q}$ denote the space associated with the estimator $\hat{q}$. Since both specify admissible parameter values, $\mathbb{Q}$ and $\mathcal{Q}$ will coincide for reasonable estimators.

Statistical Model:

$$\Upsilon = Xq_0 + \varepsilon \ , \ q \in \mathbb{R}^p$$

$$\upsilon = Xq_0 + \epsilon \ , \ \text{(realization)}$$

Assumptions: $\mathbb{E}(\varepsilon_i) = 0$, ε_i iid with $\text{var}(\varepsilon_i) = \sigma_0^2$

Least Squares Estimator and Estimate:

$$\hat{q} = (X^T X)^{-1} X^T \Upsilon \quad , \ \mathbb{E}(\hat{q}) = q_0 \ , \ V(\hat{q}) = \sigma_0^2 (X^T X)^{-1}$$

$$q = (X^T X)^{-1} X^T \upsilon$$

Error Variance Estimator and Estimate: $\widehat{R} = \Upsilon - X\hat{q}$, $R = \upsilon - Xq$

$$\hat{\sigma}^2 = \frac{1}{n-p} \widehat{R}^T \widehat{R} \quad , \quad \sigma^2 = \frac{1}{n-p} R^T R$$

Covariance Matrix Estimator and Estimate:

$$V(\hat{q}) = \hat{\sigma}^2 (X^T X)^{-1} \quad , \quad V = \sigma^2 (X^T X)^{-1}$$

Sampling Distribution: Requires $\varepsilon_i \sim N(0, \sigma_0^2)$ or sufficiently large n

- $\hat{q} \sim N(q_0, \sigma_0^2 (X^T X)^{-1})$
- $(1-\alpha) \times 100\%$ Confidence Intervals: $\delta_k = [(X^T X)^{-1}]_{kk}$

$$\left[q_k - t_{n-p,1-\alpha/2} \sigma \sqrt{\delta_k} \, , \ q_k + t_{n-p,1-\alpha/2} \sigma \sqrt{\delta_k} \right]$$

Table 7.2. *Statistical model, estimators, and statistical properties of the linear regression model. As noted in Remark 7.3, $\hat{q} = \hat{q}_{OLS}$ and $q = q_{OLS}$ are the OLS estimator and estimate.*

As noted in Section 7.1, the OLS estimate for the scalar case is obtained by minimizing the functional

$$\mathcal{J}(q) = \sum_{i=1}^{n} [\upsilon_i - f_i(q)]^2 \tag{7.30}$$

subject to $q \in \mathbb{Q}$.

The difficulty is that analytic expressions for these minimizers generally cannot be obtained for nonlinearly parameterized problems. Instead, estimates must be obtained by minimizing the least squares functional. Rather than provide a detailed analysis of the nonlinear problem, we summarize results that are analogous to the linear theory and refer readers to [24, 26, 216] for details regarding the nonlinear problem.

7.3.1 Parameter and Error Variance Estimators – Scalar Observations

Assumption 7.13. To construct parameter and error variance estimators, we require ε_i to be iid with zero mean and fixed but unknown variance σ_0^2. With this assumption, it follows that $\mathbb{E}(\Upsilon_i) = f_i(q_0)$ and $\text{var}(\Upsilon_i) = \sigma_0^2$.

Parameter Estimator and Estimate

Unlike the linear case, which can be solved explicitly using the normal equations, the determination of an OLS estimator and estimate,

$$\hat{q}_{OLS} = \underset{q \in \mathcal{Q}}{\text{argmin}} \sum_{i=1}^{n} [\Upsilon_i - f_i(q)]^2 \quad , \quad q_{OLS} = \underset{q \in \mathbb{Q}}{\text{argmin}} \sum_{i=1}^{n} [\upsilon_i - f_i(q)]^2, \tag{7.31}$$

requires numerical optimization techniques. The restriction $q \in \mathbb{Q}$ can produce constraints that must be enforced during optimization.

It was noted in Example 3.3 that parameter values for physical or biological problems can easily vary over 10 orders of magnitude. The direct optimization of (7.30) using standard software will be highly inefficient or fail for such problems. To address this, we employ scaled parameters $q_s = q./s$ where ./ denotes componentwise division and s is a vector whose components are the scale or magnitude of each parameter. Point estimates for the scaled parameters are then given by

$$q_{OLS} = \underset{q_s \in \mathbb{Q}_s}{\text{argmin}} \sum_{i=1}^{n} [\upsilon_i - f_i(q_s. \times s)]^2 \tag{7.32}$$

where $.\times$ denotes componentwise multiplication and $\mathbb{Q}_s$ is the scaled admissible parameter space. We employ (7.32) for physical problems where the magnitude of parameters vary significantly.

Remark 7.14. As noted in Remark 7.3, we will consider OLS estimators and estimates in this chapter. To simplify notation, we thus take $\hat{q} = \hat{q}_{OLS}$ and $q = q_{OLS}$ for the remainder of the discussion.

One approach for obtaining least squares estimates is to employ stochastic optimization techniques such as genetic algorithms, simulated annealing, and differential evolution [229]. These techniques reduce the reliance on accurate initial parameter estimates and, in theory, provide global convergence. However, their convergence rates are slower — they may require infinite time for convergence — and, because they are nondeterministic, multiple optimizations can yield varying final parameter values.

Alternatively, one can employ gradient-based methods such as the interior-reflective Newton, Levenberg–Marquardt, or sequential quadratic programming algorithms employed in the MATLAB routines `lsqnonlin` and `fmincon`. The efficiency and success of gradient-based optimization methods is predicated on determining good initial parameter estimates and being able to accurately determine gradients. The advantage of gradient-based methods is that once they are near the minimum, they can exhibit quadratic convergence rates which is vastly more efficient than stochastic optimization techniques. We note that one alternative is to employ the hybrid approaches in which the stochastic techniques are used to provide reasonable initial estimates for the gradient-based algorithms which then provide fast convergence to final parameter estimates.

Parameter Estimator Mean and Variance

For the linear model with design matrix X, we showed in (7.19) that $\mathbb{E}(\hat{q}) = q_0$ and $V(\hat{q}) = \sigma_0^2(X^TX)^{-1}$. In the nonlinear theory, linearization about q_0 yields the approximate covariance relations

$$V(\hat{q}) \approx \sigma_0^2 \left[\mathcal{X}^T(q_0)\mathcal{X}(q_0)\right]^{-1} \approx \hat{\sigma}^2 \left[\mathcal{X}^T(q)\mathcal{X}(q)\right]^{-1}. \tag{7.33}$$

Here $\mathcal{X}(q)$ denotes the $n \times p$ sensitivity matrix whose elements are

$$\mathcal{X}_{ik}(q) = \frac{\partial f_i(q)}{\partial q_k}. \tag{7.34}$$

Sensitivity Matrix Construction

The sensitivity matrix can be constructed using three techniques: (i) finite difference approximations, (ii) solution of sensitivity equations, or (iii) automatic differentiation. Ideally, one would compare matrices resulting from at least two of the methods to verify results.

The simplest conceptually is to approximate the derivatives using finite difference relations

$$\mathcal{X}_{ik}(q) = \frac{\partial f_i(q)}{\partial q_k} \approx \frac{f_i(q+h_k) - f_i(q)}{|h_k|} \tag{7.35}$$

where h_k is a p-vector having a nonzero k^{th} element. The difficulty is that the accuracy of (7.35) is highly dependent on the choice of h_k which also must be correctly scaled according to the magnitude of q. Hence the accuracy of results should be verified through comparison with the other techniques.

Sensitivity equations can be constructed using various techniques. In Chapter 14, we illustrate their formulation using Gâteaux differentials. More formally, they can be constructed by differentiating the evolution equation $\frac{du}{dt} = g(t, u(t), q)$ with respect to the components q_k of q, and switching the order of integration, to obtain

$$\frac{\partial u_{q_k}}{\partial t} = \frac{\partial g}{\partial u} u_{q_k} + \frac{\partial g}{\partial q_k} \tag{7.36}$$

where $u_{q_k} \equiv \frac{\partial u}{\partial q_k}$. The matrix component $\mathcal{X}_{ik}(q) = \mathcal{C}\frac{\partial u(t_i,q)}{\partial q_k}$ is easily constructed once one has numerically integrated (7.36) to obtain $u_{q_k}(t_i, q)$. This approach has the advantage that it eliminates the uncertainty associated with choosing stepsizes h_k to provide accurate finite difference approximations. However, if the original system has N differential equations, the solution of (7.36) will involve $N \cdot p$ additional differential equations. Moreover, the analytic differentiation of the original system to construct the sensitivity equations is often difficult for complex systems.

For certain problems, automatic differentiation (AD) codes can be used to construct the sensitivity equations in a form that can be directly incorporated in ODE software. In such cases, the use of AD software to construct the sensitivity matrix $\mathcal{X}(q)$ can avoid the inaccuracy associated with finite difference approximations and the potential for errors when formulating and solving the sensitivity equations.

Error Variance Estimator

Since the error variance σ_0^2 in (7.33) is unknown, we construct a variance estimator analogous to that in the linear case. Specifically, we consider the unbiased variance estimator and estimate

$$\hat{\sigma}^2 = \frac{1}{n-p}\widehat{R}^T\widehat{R} \quad , \quad \sigma^2 = \frac{1}{n-p}R^T R \tag{7.37}$$

where $\widehat{R} = \Upsilon_i - f_i(\hat{q})$ and $R = \upsilon_i - f_i(q)$ are the residual estimator and estimate. This yields the estimate

$$V = \sigma^2 \left[\mathcal{X}^T(q)\mathcal{X}(q)\right]^{-1} \tag{7.38}$$

for the covariance matrix.

Sampling Distribution

To specify a sampling distribution for $\hat{q}$, we again require either Assumption 7.7, which stipulates that errors are iid and $\varepsilon \sim N(0, \sigma_0^2)$, or that n is sufficiently large that we can invoke the central limit theorem in the sense of Property 7.9. This directly or asymptotically establishes that

$$\hat{q} \sim N\left(q_0, \sigma_0^2\left[\mathcal{X}^T(q_0)\mathcal{X}(q_0)\right]^{-1}\right) \tag{7.39}$$

where the covariance matrix is approximated by (7.38).

Confidence Intervals

The construction of $(1-\alpha) \times 100\%$ confidence intervals is analogous to the formulation (7.27) or (7.28) for the linearly parameterized model. If we let δ_k denote the k^{th} diagonal element of $[\mathcal{X}^T(q)\mathcal{X}(q)]^{-1}$, then the $(1-\alpha) \times 100\%$ confidence interval is

$$\left[q_k - t_{n-p,1-\alpha/2}\sigma\sqrt{\delta_k}\,,\ q_k + t_{n-p,1-\alpha/2}\sigma\sqrt{\delta_k}\right]. \tag{7.40}$$

where σ is given by (7.37). As noted in Section 7.2.4, t-calculators or tables can be used to calculate or look up $t_{n-p,1-\alpha/2}$ given values of n, p and α.

The properties of the least squares estimator $\hat{q}$ for the nonlinear statistical model (7.13) are compiled in Table 7.3. These can be compared with analogous properties for the linear regression problem summarized in Table 7.2.

Statistical Model:

$$\Upsilon = f(q_0) + \varepsilon \ , \ q \in \mathbb{R}^p, \ \upsilon \in \mathbb{R}^1$$

$$\upsilon = f(q_0) + \epsilon \ , \ \text{(realization)}$$

Assumptions: $\mathbb{E}(\varepsilon_i) = 0$, ε_i iid with $\text{var}(\varepsilon_i) = \sigma_0^2$

Least Squares Estimator and Estimate:

$$\hat{q} = \underset{q \in \mathcal{Q}}{\text{argmin}} \sum_{i=1}^{n} [\Upsilon_i - f_i(q)]^2 \quad , \quad q = \underset{q \in \mathbb{Q}}{\text{argmin}} \sum_{i=1}^{n} [\upsilon_i - f_i(q)]^2$$

Error Variance Estimator and Estimate: $\widehat{R} = \Upsilon - f(\hat{q})$, $R = \upsilon - f(q)$

$$\hat{\sigma}^2 = \frac{1}{n-p} \widehat{R}^T \widehat{R} \quad , \quad \sigma^2 = \frac{1}{n-p} R^T R$$

Covariance Matrix Estimator and Estimate: $\mathcal{X}_{ik}(q) = \frac{\partial f_i(q)}{\partial q_k}$

$$V(\hat{q}) = \hat{\sigma}^2 [\mathcal{X}^T(\hat{q}) \mathcal{X}(\hat{q})]^{-1} \quad , \quad V = \sigma^2 [\mathcal{X}^T(q) \mathcal{X}(q)]^{-1}$$

Statistical Properties: Requires $\varepsilon_i \sim N(0, \sigma_0^2)$ or sufficiently large n

- $\hat{q} \sim N\left(q_0, \sigma_0^2 \left[\mathcal{X}^T(q_0) \mathcal{X}(q_0)\right]^{-1}\right)$
- $(1-\alpha) \times 100\%$ Confidence Intervals: $\delta_k = [(\mathcal{X}^T(q)\mathcal{X}(q))^{-1}]_{kk}$

$$\left[q_k - t_{n-p,1-\alpha/2} \sigma \sqrt{\delta_k} \ , \ q_k + t_{n-p,1-\alpha/2} \sigma \sqrt{\delta_k}\right]$$

Table 7.3. *Statistical model, estimators, and statistical properties of the nonlinearly parameterized model (7.13) with scalar observations. As noted in Remark 7.14, $\hat{q} = \hat{q}_{OLS}$ and $q = q_{OLS}$ are the OLS estimator and estimate.*

Example 7.15. Consider the spring model

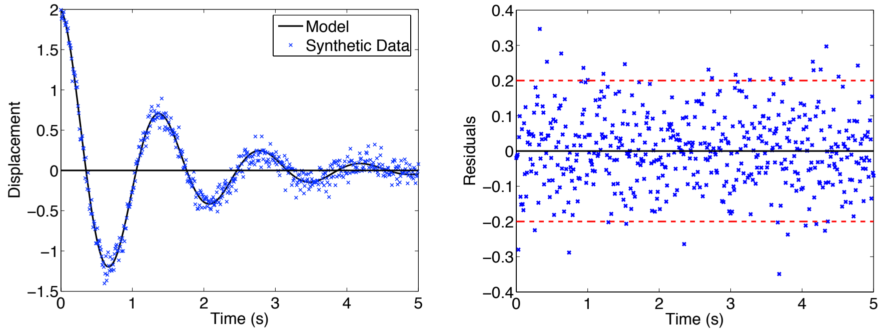
$$\begin{aligned} &\ddot{z} + C\dot{z} + Kz = 0 \\ &z(0) = 2 \ , \ \dot{z}(0) = -C \end{aligned} \tag{7.41}$$

with displacement observations so that

$$y = [1 \ \ 0] \begin{bmatrix} z \\ \dot{z} \end{bmatrix} = z.$$

We showed in Example 3.2 that (7.41) has the solution

$$z(t) = 2e^{-Ct/2} \cos\left(\sqrt{K - C^2/4} \cdot t\right) \tag{7.42}$$



(a) (b)

Figure 7.1. *(a) Synthetic data and modeled displacement and (b) residuals at* $n = 501$ *points.*

when $C^2 - 4K < 0$. We take $K = 20.5$ to be known and let $q = C$ be the parameter considered in the statistical analysis. We note that although the model exhibits a linear dependence on the states z and $\dot{z}$, the dependence of $z(t, q)$ on q is nonlinear.

To numerically generate synthetic data, we employ $C_0 = 1.5$ and add noise $\varepsilon \sim N(0, \sigma_0^2)$ where $\sigma_0 = 0.1$. The model and one realization of the data at $n = 501$ points are plotted in Figure 7.1(a) and the residuals are plotted in Figure 7.1(b). By construction, the residuals are iid with 94.4% of the values lying with the 2σ interval indicated by the horizontal lines.

The $n \times 1$ sensitivity matrix (vector) is

$$\mathcal{X}(q) = \left[\frac{\partial y}{\partial C}(t_1, q), \cdots, \frac{\partial y}{\partial C}(t_n, q)\right]^T \tag{7.43}$$

where

$$\frac{\partial y}{\partial C} = e^{-Ct/2}\left[\frac{Ct}{\sqrt{4K - C^2}} \sin\left(\sqrt{K - C^2/4} \cdot t\right) - t \cos\left(\sqrt{K - C^2/4} \cdot t\right)\right] \tag{7.44}$$

results from differentiating (7.42). The construction of $\mathcal{X}(q)$ by constructing and solving the corresponding sensitivity equations is addressed in Exercise 7.1.

Because we know σ_0^2, we obtain the covariance value

$$V = \sigma_c^2 = \sigma_0^2 \left[\mathcal{X}^T(q)\mathcal{X}(q)\right]^{-1} = 3.35 \times 10^{-4}$$

so that $\sigma_c = 0.0183$. Since $\varepsilon_i \sim N(0, \sigma_0^2)$, the random variable $\widehat{C}$ has the sampling distribution

$$\widehat{C} \sim N\left(C_0, \sigma_c^2\right) \tag{7.45}$$

which is plotted in Figure 7.2. The parameter estimated by minimizing (7.31) for the data plotted in Figure 7.1 is $C = 1.4792$ and the 95% confidence interval given by (7.40) is [1.4433, 1.5150].

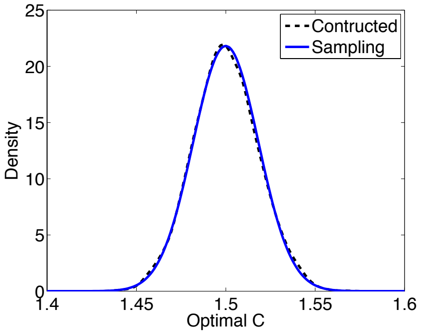


Figure 7.2. *(a) Sampling density $N\left(C_0, \sigma_c^2\right)$ for $\widehat{C}$ and density constructed from 10,000 simulations.*

It was noted in Sections 4.8.1 and 7.1 that in frequentist inference, the 95% confidence interval has the following interpretation in the context of parameter estimation; if the procedure is repeated ℓ times, 0.95ℓ of the computed intervals will contain the true parameter q_0. This is illustrated in Figure 4.11(a). To demonstrate for this example, we generated $10,000$ sets of numerical data using the true parameter values C_0 and $\varepsilon_i \sim N(0, \sigma_0^2)$ with $\sigma_0 = 0.1$. For each data set, we optimized (7.31) to obtain a point estimate C and corresponding 95% confidence interval. In this set of numerical experiments, 9455 of the intervals contained C_0. Using the 10,000 estimated values of C, we used the kernel estimation techniques discussed in Section 4.1.1 to construct the density which is plotted in Figure 7.2. As expected, the kernel density estimate matches the representation (7.45) for the sampling distribution.

Example 7.16. We showed in Example 3.5 that the boundary value problem

$$\frac{d^2T_s}{dx^2} = \frac{2(a+b)}{ab}\frac{h}{k}\left[T_s(x) - T_{amb}\right]$$
$$\frac{dT_s}{dx}(0) = \frac{\Phi}{k} \quad , \quad \frac{dT_s}{dx}(L) = \frac{h}{k}[T_{amb} - T_s(L)]$$

models the steady state temperature of an uninsulated rod with source heat flux Φ at $x = 0$ and ambient air temperature T_{amb}. The model parameters to be estimated and statistically analyzed are $q = [\Phi, h]$ where h is the convective heat transfer coefficient.

The rod used in these experiments was aluminum with cross-sectional dimensions $a = b = 0.95$ cm and length $L = 70$ cm. The temperature measurements y_i, compiled in Table 3.2, were made at 15 equally spaced spatial locations $x_i = x_0 + (i-1)\Delta x$ where $x_0 = 10$ cm and $\Delta x = 4$ cm. The observed solution is

$$y_i(q) = T_s(x_i, q) = c_1(q)e^{-\gamma x_i} + c_2(q)e^{\gamma x_i} + T_{amb}$$

where $\gamma = \sqrt{\frac{2(a+b)h}{abk}}$ and

$$c_1(q) = -\frac{\Phi}{k\gamma}\left[\frac{e^{\gamma L}(h+k\gamma)}{e^{-\gamma L}(h-k\gamma)+e^{\gamma L}(h+k\gamma)}\right] \quad , \quad c_2(q) = \frac{\Phi}{k\gamma} + c_1(q).$$

We suppress the parameter dependence of γ to clarify the notation. We employ the thermal conductivity value $k = 2.37\frac{W}{cm \cdot C}$ reported for aluminum and the measured ambient room temperature $T_{amb} = 21.29^o$C.

A least squares fit to the data yielded the parameter estimates $\Phi = -18.41$ and $h = 0.00191$ and the model fit shown in Figure 7.3(a). We note that this value of h falls within the range $2.8 \times 10^{-4} - 0.0023\frac{W}{cm^2 \cdot C}$ reported for still air. The residuals plotted in Figure 7.3(b) exhibit no discernible pattern thus motivating the assumption that the errors ε_i are iid. We assume that errors are normally distributed when constructing a sampling distribution.

The error variance estimate is $\sigma^2 = 0.0627$ and the covariance matrix, computed using analytic sensitivity relations, as derived in Exercise 7.4 and illustrated in Figure 7.4, is

$$V = \begin{bmatrix} 2.1034 \times 10^{-2} & -2.0286 \times 10^{-6} \\ -2.0286 \times 10^{-6} & 2.0972 \times 10^{-10} \end{bmatrix}. \tag{7.46}$$

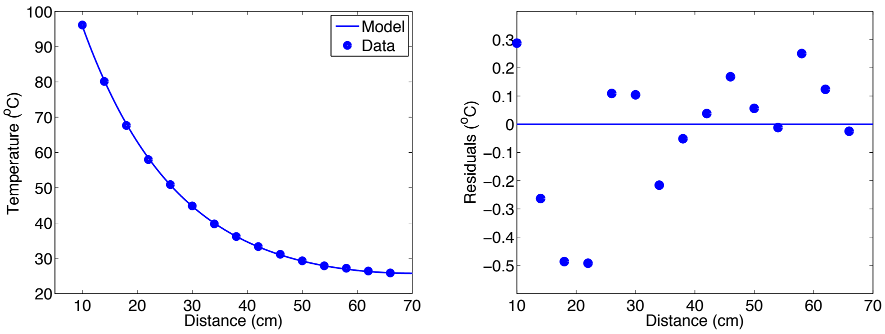
The standard deviations for the errors and sampling distribution are

$$\sigma = 0.2504 \; , \sigma_\Phi = 0.1450 \; , \; \sigma_h = 1.4482 \times 10^{-5}. \tag{7.47}$$

Since $n = 15$ and $p = 2$, the 95% confidence intervals are

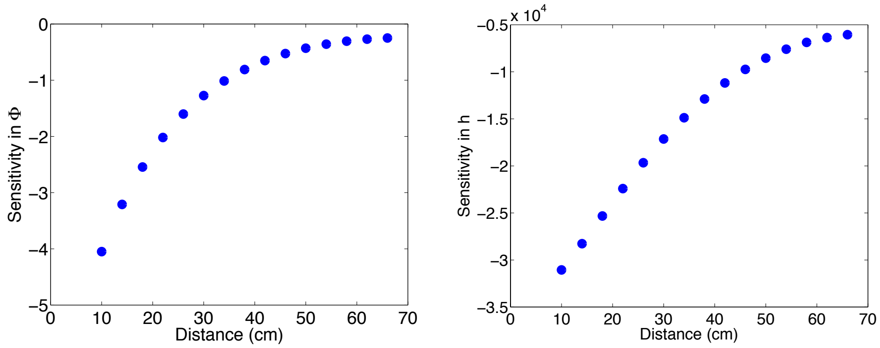
$$[-18.7233, -18.0967] \; , \; [1.8787 \times 10^{-3}, 1.9413 \times 10^{-3}].$$

In Example 8.12, we revisit this example in the context of Bayesian analysis.



(a) (b)

Figure 7.3. *(a) Model fit to the steady-state temperature data, and (b) residuals at the 15 spatial locations.*



(a) (b)

Figure 7.4. *Analytic sensitivity values: (a)* $\frac{\partial y}{\partial \Phi}(x_i, q)$ *and (b)* $\frac{\partial y}{\partial h}(x_i, q)$.

7.3.2 Parameter and Error Variance Estimators for Evolution Models – Multiple Responses

In this section, we consider the evolution equation (7.1) with $\nu > 1$ data measurements and model responses specified by an $\nu \times N$ matrix $\mathcal{C}$. The statistical model in this case is

$$\Upsilon_i = f(t_i, q_0) + \varepsilon_i\ ,\ \ j = 1, \cdots, n$$

where Υ_i and ε_i are random ν-vectors.

Assumption 7.17. To accommodate the possibility that error distributions associated with individual components of the observations could differ, we let $\sigma_{0_j}^2$ denote the fixed but unknown variance of the error associated with the j^{th} observation. These values are compiled in the $\nu \times \nu$ diagonal measurement error covariance matrix $V_0 = \text{diag}[\sigma_{0_1}^2, \cdots, \sigma_{0_\nu}^2]$. As before, errors are assumed to be unbiased. We remind the reader that V_0 is fixed but typically unknown.

The construction of parameter and covariance estimators is similar in theory to the scalar case $\nu = 1$ but is complicated by the coupling induced by the potentially differing variances of the error components. We provide an overview of the estimators, estimates and sampling distribution for $\nu > 1$ and refer the reader to [24, 26] for details.

Example 7.18. It was noted in Example 3.2 that for vibrating systems modeled as a simple harmonic oscillator (3.11), displacements and velocities can be respectively measured using a proximity sensor and laser vibrometer. If both sets of measurements are available, the modeled observations will be

$$\begin{bmatrix} y_1(t_i, q) \\ y_2(t_i, q) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} z_1(t_i, q) \\ z_2(t_i, q) \end{bmatrix}$$

which is just the parameter-dependent states. Given the differing nature of the measurement devices, one would expect different error distributions to be associated with the experimental measurements y_1 and y_2. Hence we would employ

$$V_0 = \begin{bmatrix} \sigma_{0_1}^2 & 0 \\ 0 & \sigma_{0_2}^2 \end{bmatrix}. \tag{7.48}$$

Example 7.19. For the HIV model (3.15) of Example 3.3, one can typically only measure the total number $T_1 + T_1^*$ of T-lymphocytes and the viral load V. Hence

$$\mathcal{C} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

and $y(t,q) \in \mathbb{R}^2$. The error covariance matrix would again have the structure (7.48).

Parameter and Error Covariance Estimators

The OLS estimator and estimate are taken to be

$$\begin{aligned} \hat{q}_{OLS} &= \underset{q\in\mathcal{Q}}{\operatorname{argmin}} \sum_{i=1}^{n} [\Upsilon_i - f(t_i,q)]^T \, V_0^{-1} \, [\Upsilon_i - f(t_i,q)] \\ q_{OLS} &= \underset{q\in\mathbb{Q}}{\operatorname{argmin}} \sum_{i=1}^{n} [\upsilon_i - f(t_i,q)]^T \, V_0^{-1} \, [\upsilon_i - f(t_i,q)] \end{aligned} \tag{7.49}$$

where V_0^{-1} weights the response components by the reciprocals of the corresponding error variance associated with each component. Since V_0 is typically unknown, it too must be estimated. Motivated by (7.37), the estimate $V \approx V_0$ is provided by the relation

$$V = \operatorname{diag}\left(\frac{1}{n-p} \sum_{i=1}^{n} [\upsilon_i - f(t_i, q_{OLS})][\upsilon_i - f(t_i, q_{OLS})]^T \right). \tag{7.50}$$

Unlike the scalar response relations (7.31) and (7.37), the multiple response relations (7.49) and (7.50) are coupled due to the fact that $V_0 \neq \sigma_0^2 I$, and hence they must be solved as a coupled system.

Sampling Distribution

To specify a sampling distribution, we need an assumption analogous to Assumption 7.7.

Assumption 7.20. Let ε_{ij} denote the error in the i^{th} component of Υ_i at time t_i. We make the assumption that $\varepsilon_{ij} \sim N(0, \sigma_{0_j}^2)$ so that $\varepsilon_i \sim N(0, V_0)$. For n sufficiently large, the central limit theorem can be invoked in the manner detailed in Property 7.9 to obtain similar asymptotic results.

With this assumption, it is shown in [24, 26] that

$$\hat{q}_{OLS} \sim N(q_0, \mathcal{V}_0) \approx N(q_{OLS}, \mathcal{V})$$

where

$$\mathcal{V}_0 \approx \left(\sum_{j=1}^{n} \mathcal{X}_j^T(q_0) V_0^{-1} \mathcal{X}_j(q_0) \right)^{-1}$$

is the $p \times p$ covariance matrix and

$$\mathcal{X}_j(q) = \begin{bmatrix} \frac{\partial f_1(t_i,q)}{dq_1} & \cdots & \frac{\partial f_1(t_i,q)}{dq_p} \\ \vdots & & \vdots \\ \frac{\partial f_\nu(t_i,q)}{dq_1} & \cdots & \frac{\partial f_\nu(t_i,q)}{dq_p} \end{bmatrix} \tag{7.51}$$

is the $\nu \times p$ sensitivity matrix at time t_i. For implementation, $\mathcal{V}_0$ is approximated by

$$\mathcal{V} = \left(\sum_{j=1}^{n} \mathcal{X}_j^T(q_{OLS}) V^{-1} \mathcal{X}_j(q_{OLS}) \right)^{-1}$$

where (7.51) must be evaluated at each time step. The $(1-\alpha) \times 100\%$ confidence intervals are

$$[q_{OLS,k} - t_{n-p,1-\alpha/2} SE, q_{OLS,k} + t_{n-p,1-\alpha/2} SE]$$

where $q_{OLS,k}$ is the k^{th} element of q_{OLS} and the standard error is

$$SE \approx \sqrt{\mathcal{V}_k}.$$

Here $\mathcal{V}_k$ is the k^{th} diagonal element of $\mathcal{V}$.

7.4 Notes and References

The parameter estimation techniques discussed in this chapter are based on linear and nonlinear regression for which there are numerous excellent texts. The text [95] provides a very nice introduction to linear regression and has the advantage that the authors use different notation to delineate between random variables and their realizations. This is also a good resource for obtaining additional background regarding the confidence and prediction intervals discussed in Chapter 9. Asymptotic theory for nonlinear regression problems is detailed in the classic book [216]. We refer readers to [24, 26] for details regarding the construction of estimators and specification of sampling distributions for parameters in nonlinear evolution models.

For brevity, we do not discuss the following topics: infinite-dimensional inverse problems associated with parameter estimation, regularization, or optimization methods for inverse problems. The reader is referred to [25] for theory and estimation techniques for distributed parameter systems and [15, 126, 173, 241, 253] for details regarding regularization, computational algorithms and case studies pertaining to parameter estimation and inverse problems. The texts [61, 129, 130, 229] cover a variety of optimization techniques that are appropriate for this class of problems.