Least squares Tutorial

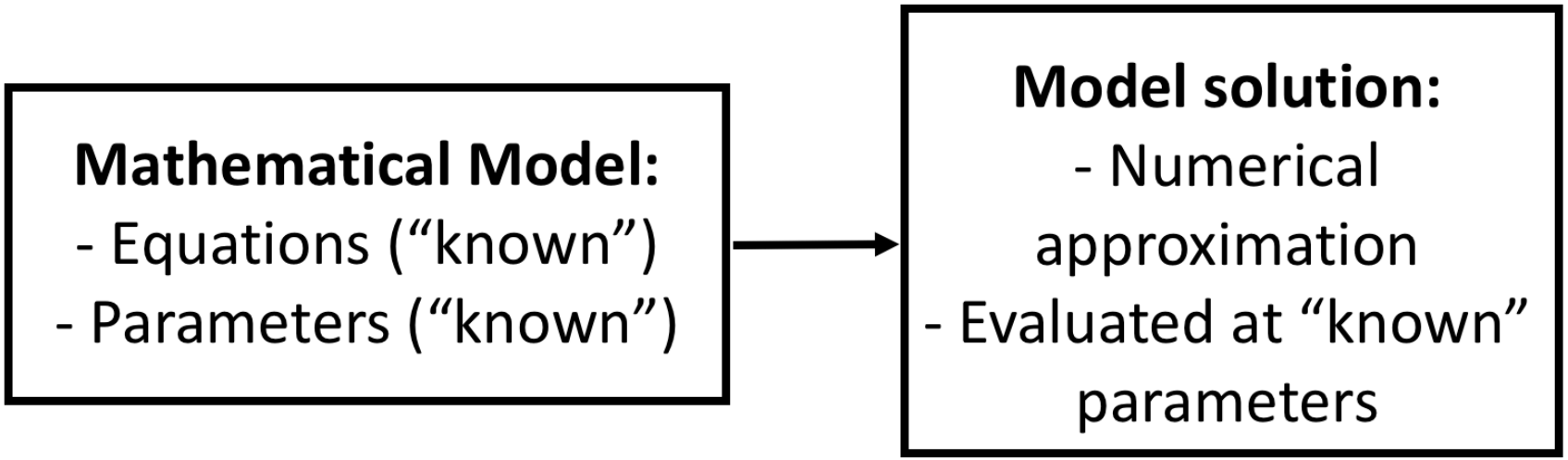
Kevin Flores

Center for Research in Scientific Computation
Center for Quantitative Sciences in Biomedicine
Department of Mathematics
North Carolina State University
Raleigh, NC 27695-8212 USA

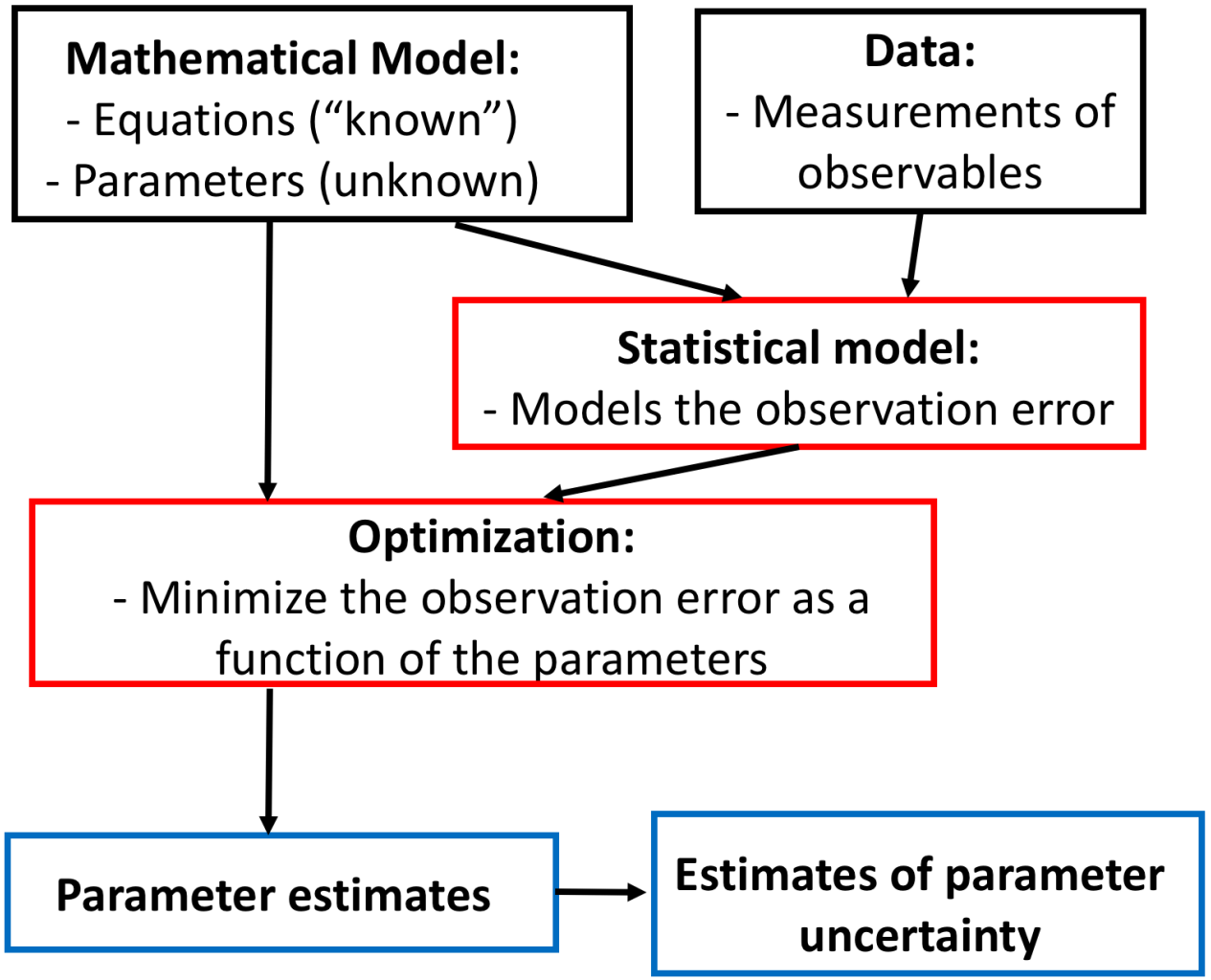
NCSU RTG Tutorial Workshop on Parameter Estimation for Biological Models
July 28, 2016

Acknowledgement: Slides courtesy of H.T. Banks, Zachary Kenz, Keri Rehm

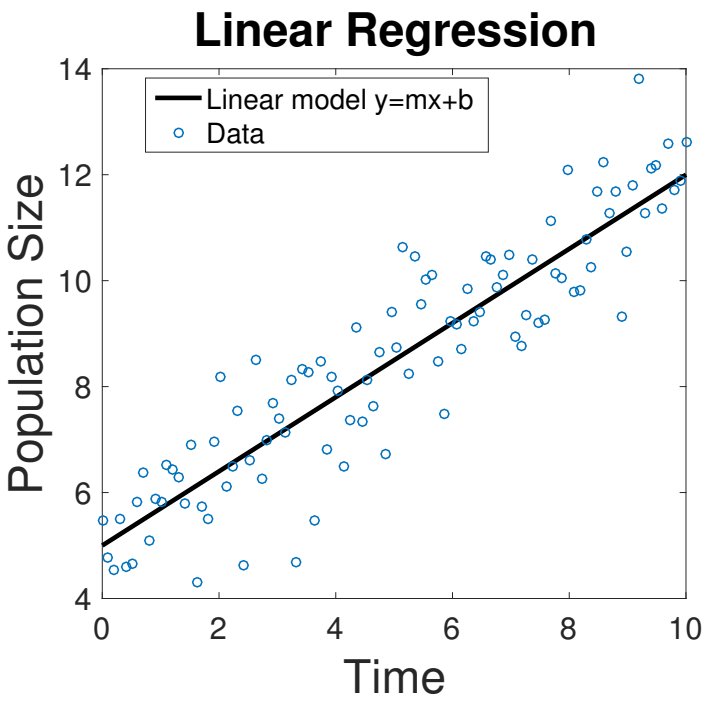
Typical "Forward Problem"

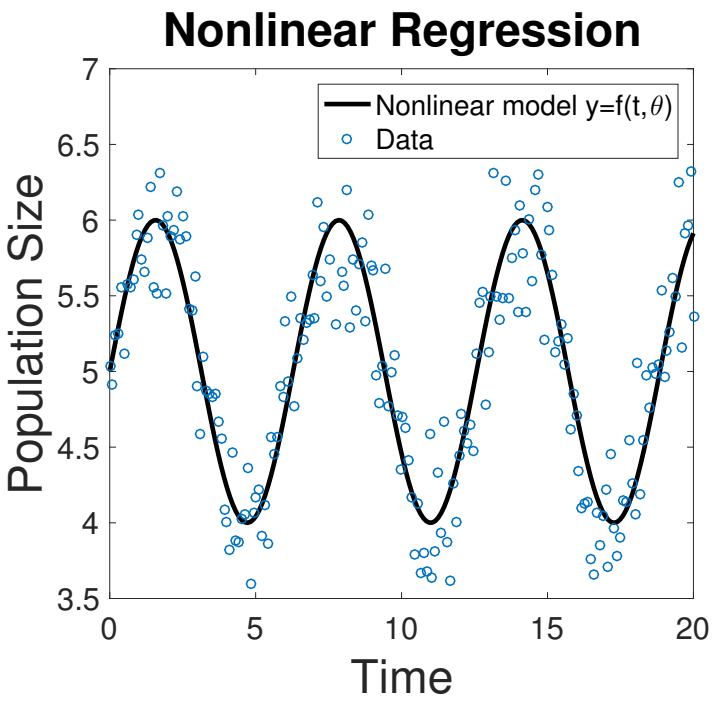


Typical "Inverse Problem"

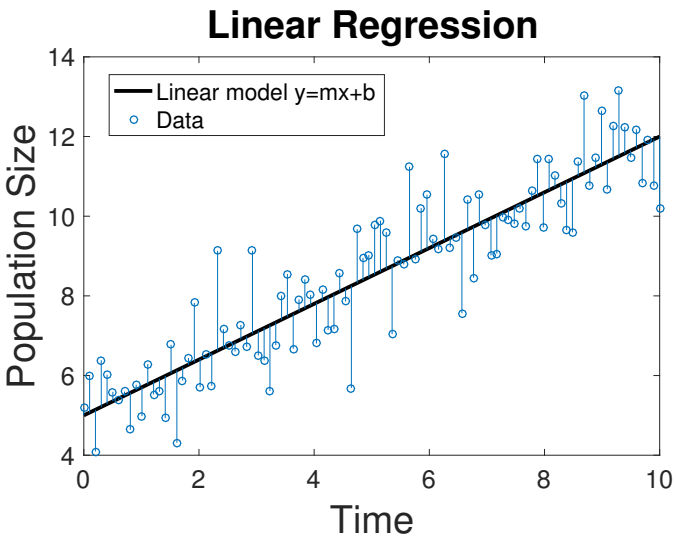


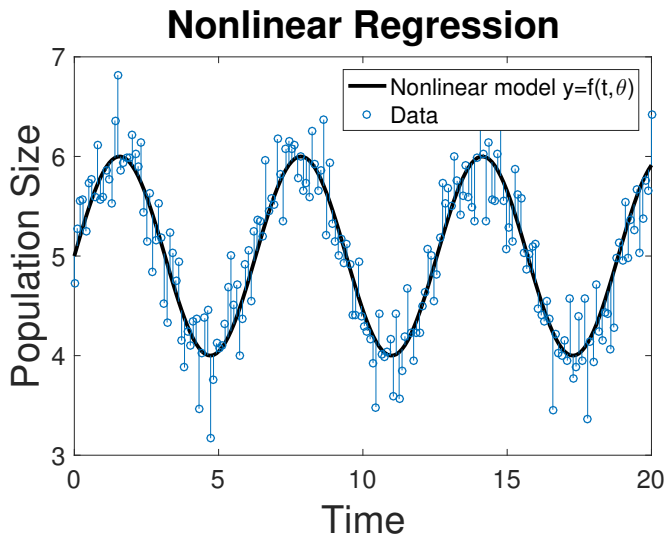
Linear vs. Nonlinear Regression

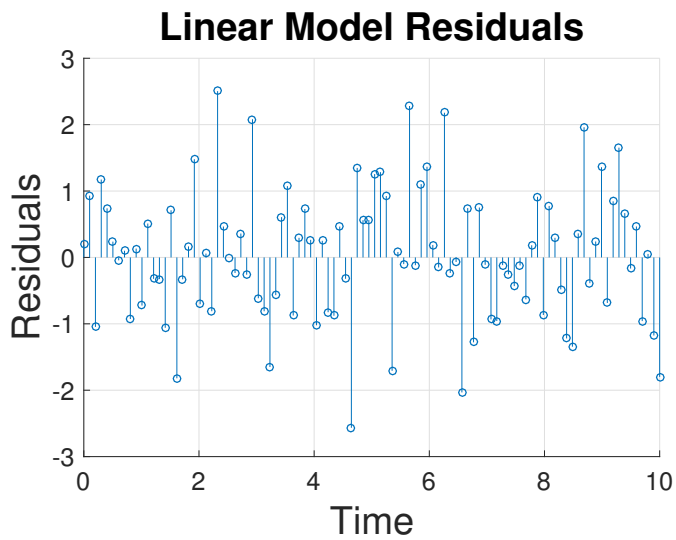


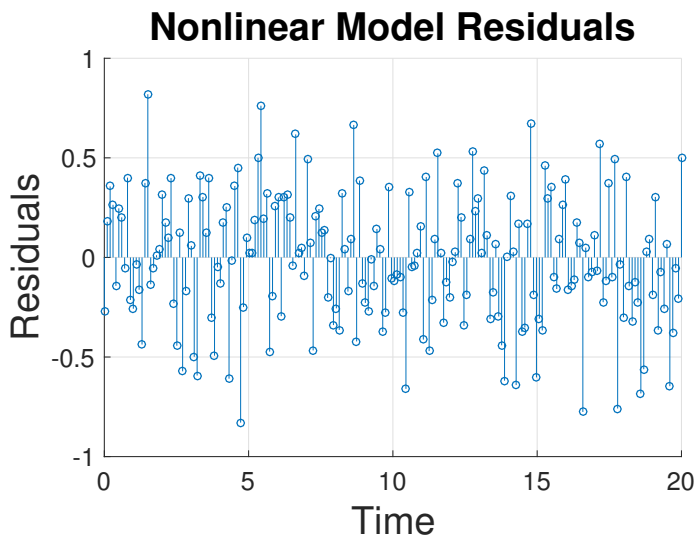


Linear vs. Nonlinear Regression









Least Squares Inverse Problem Formulations

The Mathematical Model We consider inverse or parameter estimation problems in the context of a parameterized (with vector parameter $\boldsymbol{q} \in \mathbb{R}^{\kappa q}$) n-dimensional vector dynamical system or **mathematical model**

$$\frac{d\boldsymbol{x}}{dt}(t) = \boldsymbol{g}(t, \boldsymbol{x}(t), \boldsymbol{q}), \tag{1}$$

$$\boldsymbol{x}(t_0) = \boldsymbol{x}_0, \tag{2}$$

with **observation process**

$$\boldsymbol{f}(t; \theta) = \mathcal{C}\boldsymbol{x}(t; \theta), \tag{3}$$

where $\theta = (\boldsymbol{q}^T, \tilde{\boldsymbol{x}}_0^T)^T \in \mathbb{R}^{\kappa q + \tilde{n}} = \mathbb{R}^{\kappa_\theta}, \tilde{n} \leq n$, and the observation operator $\mathcal{C}$ maps $\mathbb{R}^n$ to $\mathbb{R}^m$. In most of the discussions below we assume without loss of generality that some subset $\tilde{\boldsymbol{x}}_0$ of the initial values $\boldsymbol{x}_0$ are also unknown.

Following usual conventions (which correspond to the form of data usually available from experiments), we assume a discrete form of the observations in which one has N longitudinal observations $\mathbf{y}_j$ corresponding to

$$\boldsymbol{f}(t_j;\boldsymbol{\theta}) = \mathcal{C}\boldsymbol{x}(t_j;\boldsymbol{\theta}), \quad j = 1,\ldots,N. \tag{4}$$

In general the corresponding observations or data $\{\mathbf{y}_j\}$ will not be exactly $\boldsymbol{f}(t_j;\boldsymbol{\theta})$. Due to the nature of the phenomena leading to this discrepancy, we treat this uncertainty pertaining to the observations with a statistical model for the observation process.

The Statistical Model

In our discussions here we consider a **statistical model** of the form

$$\boldsymbol{Y}_j = \boldsymbol{f}(t_j; \boldsymbol{\theta}_0) + \boldsymbol{h}_j \circ \widetilde{\boldsymbol{\mathcal{E}}}_j, \quad j = 1, \ldots, N, \tag{5}$$

where $\boldsymbol{f}(t_j; \boldsymbol{\theta}) = \mathcal{C}\boldsymbol{x}(t_j; \boldsymbol{\theta})$, $j = 1, \ldots, N$, and $\mathcal{C}$ is an $m \times n$ matrix. This corresponds to the observed part of the solution of the mathematical model (1)-(2) at the j^{th} covariate or observation time for a particular vector of parameters $\boldsymbol{\theta} \in \mathbb{R}^{\kappa q + \tilde{n}} = \mathbb{R}^{\kappa_\theta}$. Here the m-vector function $\boldsymbol{h}_j$ is defined by

$$\boldsymbol{h}_j = \begin{cases} (1, \ldots, 1)^T & \text{for the vector OLS case} \\ (w_{1,j}, \ldots, w_{m,j})^T & \text{for the vector WLS case} \\ (f_1^{\gamma}(t_j; \boldsymbol{\theta}_0), \ldots, f_m^{\gamma}(t_j; \boldsymbol{\theta}_0))^T & \text{for the vector GLS case,} \end{cases} \tag{6}$$

for $j = 1, \ldots, N$, and $\boldsymbol{h}_j \circ \widetilde{\boldsymbol{\mathcal{E}}}_j$ denotes the component-wise multiplication of the vectors $\boldsymbol{h}_j$ and $\widetilde{\boldsymbol{\mathcal{E}}}_j$.

- The vector θ_0 represents the "truth" parameter that generates the observations $\{\mathbf{Y}_j\}_{j=1}^N$.
- The terms $\boldsymbol{h}_j \circ \widetilde{\boldsymbol{\mathcal{E}}}_j$ are random variables which can represent observation or measurement error, "system fluctuations" or other phenomena that cause observations to not fall exactly on the points $\boldsymbol{f}(t_j; \theta_0)$ from the smooth path $\boldsymbol{f}(t, \theta_0)$.
- Since these fluctuations are unknown to the modeler, we will assume that realizations $\widetilde{\epsilon}_j$ of $\widetilde{\mathcal{E}}_j$ are generated from a probability distribution which reflects the assumptions regarding these phenomena.

Thus specific data (*realizations*) corresponding to (5) will be represented by

$$\boldsymbol{y}_j = \boldsymbol{f}(t_j; \boldsymbol{\theta}_0) + \boldsymbol{h}_j \circ \widetilde{\boldsymbol{\epsilon}}_j, \quad j = 1, \ldots, N. \tag{7}$$

We make standard assumptions about the $\widetilde{\boldsymbol{\mathcal{E}}}_j$ in that they are independent and identically distributed with mean zero and and constant covariance matrix. This model (7) allows for a fairly wide range of error models including the usual *absolute* (or *constant variance*) error model, when $\gamma = 0$ (the OLS case), as well as the *relative (or constant coefficient of variation) error* model when $\gamma = 1$.

Remaining Slides:

- Discuss methodology related to estimates $\hat{\theta}$ for the true value of the parameter θ_0 from a set Ω_θ of admissible parameters, and the dependence of this methodology on what is assumed about the choice of γ and the covariance matrices of the errors $\widetilde{\mathcal{E}}_j$.
- We discuss a class of inverse problem methodologies that can be used to calculate estimates $\hat{\theta}$ for θ_0: the ordinary, the weighted and the generalized least-squares formulations.
- We are interested in situations (as is the case in most applications) where the error distribution is *unknown* to the modeler beyond the assumptions on $\mathbb{E}(\boldsymbol{Y}_j)$ embodied in the model and the assumptions made on $\text{Var}(\widetilde{\mathcal{E}}_j)$.

Methodology: Ordinary, Weighted and Generalized Least Squares

Scalar Ordinary Least Squares To simplify notation, we first consider the absolute error statistical model ($\gamma = 0$) in the scalar case. This then takes the form

$$Y_j = f(t_j; \theta_0) + \widetilde{\mathcal{E}}_j, \; j = 1, \ldots, N, \tag{8}$$

where the variance $\mathrm{Var}(\widetilde{\mathcal{E}}_j) = \sigma_0^2$ is assumed to be unknown to the modeler. (Note also that the distribution of the error need not be specified.) It is assumed that the observation errors are independent across j (i.e., time), which may be a reasonable one when the observations are taken with sufficient intermittency or when the primary source of error is measurement error.

If we define

$$\boldsymbol{\theta}_{\text{OLS}} = \boldsymbol{\theta}^{N}_{\text{OLS}}(\boldsymbol{Y}) = \arg\min_{\boldsymbol{\theta}\in\Omega_\theta} \sum_{j=1}^{N} [Y_j - f(t_j;\boldsymbol{\theta})]^2, \tag{9}$$

where $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_N)^T$, then $\boldsymbol{\theta}_{\text{OLS}}$ can be viewed as minimizing the distance between the data and model where all observations are treated as being of equal importance.

We note that minimizing the functional in (9) corresponds to solving for θ in

$$\sum_{j=1}^{N}[Y_j - f(t_j;\theta)]\nabla f(t_j;\theta) = 0, \tag{10}$$

the so-called *normal equations* or *estimating equations*. We point out that θ_{OLS} is a *random vector* (because $\widetilde{\mathcal{E}}_j = Y_j - f(t_j;\theta)$ is a random variable); hence if $\{y_j\}_{j=1}^N$ are realizations of the *random variables* $\{Y_j\}_{j=1}^N$ then solving

$$\hat{\theta}_{\text{OLS}} = \hat{\theta}_{\text{OLS}}^N = \arg\min_{\theta\in\Omega_\theta} \sum_{j=1}^{N}[y_j - f(t_j;\theta)]^2 \tag{11}$$

provides a realization for θ_{OLS}.

Notation:

- For a random vector or estimator θ_{OLS}, we will always denote a corresponding realization or estimate with an over hat, e.g., $\hat{\theta}_{\text{OLS}}$ is an estimate for θ_0.
- We sometimes suppress the dependence on N unless it is specifically needed.
- Finally, we drop the subscript OLS for the estimates when it is clearly understood in context.

Returning to (9) and (11) and noting that

$$\sigma_0^2 = \frac{1}{N}\mathbb{E}\left(\sum_{j=1}^{N}[Y_j - f(t_j;\boldsymbol{\theta}_0)]^2\right), \tag{12}$$

we see that once we have solved for $\hat{\boldsymbol{\theta}}_{\mathrm{OLS}}$ in (11), we may readily obtain an estimate $\hat{\sigma}^2_{\mathrm{OLS}}$ for σ_0^2. (Recall that $\mathbb{E}$ denotes the expectation operator.)

Even though the distribution of the error random variables is not specified, we can use asymptotic theory to approximate the mean and covariance of the random vector $\boldsymbol{\theta}_{\text{OLS}}$ [43]. As will be explained in more detail below, as $N \to \infty$, we have that

$$\boldsymbol{\theta}_{\text{OLS}} = \boldsymbol{\theta}_{\text{OLS}}^N \sim \mathcal{N}(\boldsymbol{\theta}_0, \Sigma_0^N) \approx \mathcal{N}(\boldsymbol{\theta}_0, \sigma_0^2[F_{\boldsymbol{\theta}}^N(\boldsymbol{\theta}_0)^T F_{\boldsymbol{\theta}}^N(\boldsymbol{\theta}_0)]^{-1}), \quad (13)$$

where the sensitivity matrix $F_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = F_{\boldsymbol{\theta}}^N(\boldsymbol{\theta}) = \left((F_{\boldsymbol{\theta}}^N)_{jk}(\boldsymbol{\theta})\right)$ is defined by

$$(F_{\boldsymbol{\theta}}^N)_{jk}(\boldsymbol{\theta}) = \frac{\partial f(t_j;\boldsymbol{\theta})}{\partial \theta_k}, \quad j = 1, \ldots, N, \quad k = 1, \ldots, \kappa_\theta, \quad (14)$$

and

$$\Sigma_0^N \equiv \sigma_0^2[N\Omega_0]^{-1}, \quad (15)$$

with

$$\Omega_0 \equiv \lim_{N\to\infty} \frac{1}{N} F_{\boldsymbol{\theta}}^N(\boldsymbol{\theta}_0)^T F_{\boldsymbol{\theta}}^N(\boldsymbol{\theta}_0), \quad (16)$$

where the limit is assumed to exist (see [9, 14, 43]).

- $\boldsymbol{\theta}_{\text{OLS}}$ is approximately distributed as a multivariate normal random variable with mean $\boldsymbol{\theta}_0$ and covariance matrix Σ_0^N.
- The *realization* (data) $\mathbf{y} = (y_1, \ldots, y_N)^T$ of the random vector $\mathbf{Y}$ is used to estimate $\hat{\boldsymbol{\theta}}_{\text{OLS}}$ given by (11) and the *bias adjusted* approximation for σ_0^2:

$$\hat{\sigma}^2_{\text{OLS}} = \frac{1}{N - \kappa_\theta} \sum_{j=1}^{N} [y_j - f(t_j; \hat{\boldsymbol{\theta}}_{\text{OLS}})]^2. \tag{17}$$

- Both $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\text{OLS}}$ and $\hat{\sigma}^2 = \hat{\sigma}^2_{\text{OLS}}$ will then be used to approximate the covariance matrix

$$\Sigma_0^N \approx \hat{\Sigma}^N \equiv \hat{\sigma}^2 [F_{\boldsymbol{\theta}}^N(\hat{\boldsymbol{\theta}})^T F_{\boldsymbol{\theta}}^N(\hat{\boldsymbol{\theta}})]^{-1}. \tag{18}$$

- We can obtain the standard errors $\text{SE}_k(\hat{\boldsymbol{\theta}}_{\text{OLS}})$ (discussed in more detail below) for the k^{th} element of $\hat{\boldsymbol{\theta}}_{\text{OLS}}$ by calculating $\text{SE}_k(\hat{\boldsymbol{\theta}}_{\text{OLS}}) \approx \sqrt{\hat{\Sigma}^N_{kk}}$.

Remarks:

- $\hat{\sigma}^2_{\text{OLS}}$ represents the estimate for σ_0^2 of (12) with the factor $\frac{1}{N}$ replaced by the factor $\frac{1}{N - \kappa_\theta}$.
- In the linear case the estimate with $\frac{1}{N}$ can be shown to be biased downward (i.e., biased too low) and the same behavior can be observed in the general nonlinear case – see Chapter 12 of [43] and p. 28 of [27].
- The subtraction of κ_θ degrees of freedom reflects the fact that $\hat{\boldsymbol{\theta}}$ has been computed to satisfy the κ_θ normal equations (10).

Vector Ordinary Least Squares

We next consider the more general case in which we have a **vector of observations** for the j^{th} covariate t_j. If we still assume the variance is constant in longitudinal data, then the statistical model is reformulated as

$$\boldsymbol{Y}_j = \boldsymbol{f}(t_j; \boldsymbol{\theta}_0) + \widetilde{\boldsymbol{\mathcal{E}}}_j, \tag{19}$$

where $\boldsymbol{f}(t_j; \boldsymbol{\theta}_0) \in \mathbb{R}^m$ and $\widetilde{\boldsymbol{\mathcal{E}}}_j,\ j = 1, \ldots, N$ are independent and identically distributed with zero mean and covariance matrix given by

$$V_0 = \mathrm{Var}(\widetilde{\boldsymbol{\mathcal{E}}}_j) = \mathrm{diag}(\sigma^2_{0,1}, \ldots, \sigma^2_{0,m}), \tag{20}$$

for $j = 1, \ldots, N$. Here we have allowed for the possibility that the observation coordinates $\mathbf{Y}_j$ may have different *constant* variances $\sigma^2_{0,i}$, i.e., $\sigma^2_{0,i}$ does not necessarily have to equal $\sigma^2_{0,k}$.

We note that this formulation also can be used to treat the case where V_0 is used to simply scale the observations, (i.e., $V_0 = \text{diag}(v_1, \ldots, v_m)$ is known). In this case the formulation is simply a *vector OLS* (sometimes also called a weighted least squares (WLS)).

Weighted Least Squares (WLS)

- In the above discussion we required that the measurement error remain constant in variance in longitudinal data.
- This assumption may not be appropriate for data sets whose measurement error is not constant in a longitudinal sense.
- A common *weighted error* model, in which the error is weighted according to some *known* weights, an assumption which might be reasonable when one has data that varies widely in the scale of observations that experimentalists must use for the scalar observation case is

$$Y_j = f(t_j; \theta_0) + w_j \widetilde{\mathcal{E}}_j. \tag{21}$$

Here $\mathbb{E}(Y_j) = f(t_j; \theta_0)$ and $\mathrm{Var}(Y_j) = \sigma_0^2 w_j^2$, which derives from the assumptions that $\mathbb{E}(\widetilde{\mathcal{E}}_j) = 0$ and $\mathrm{Var}(\widetilde{\mathcal{E}}_j) = \sigma_0^2$.

The WLS estimator is defined here by

$$\boldsymbol{\theta}_{\text{WLS}} = \arg \min_{\boldsymbol{\theta} \in \Omega_\theta} \sum_{j=1}^{N} w_j^{-2} [Y_j - f(t_j; \boldsymbol{\theta})]^2, \tag{22}$$

with corresponding estimate

$$\hat{\boldsymbol{\theta}}_{\text{WLS}} = \arg \min_{\boldsymbol{\theta} \in \Omega_\theta} \sum_{j=1}^{N} w_j^{-2} [y_j - f(t_j; \boldsymbol{\theta})]^2. \tag{23}$$

This special form of the WLS estimate can be thought of minimizing the distance between the data and model while taking into account the known but unequal quality of the observations [27].

The WLS estimator $\boldsymbol{\theta}_{\text{WLS}} = \boldsymbol{\theta}^N_{\text{WLS}}$ has the following asymptotic properties [26, 27]:

$$\boldsymbol{\theta}_{\text{WLS}} \sim \mathcal{N}(\boldsymbol{\theta}_0, \Sigma_0^N), \tag{24}$$

where

$$\Sigma_0^N \approx \sigma_0^2 \left(F_{\boldsymbol{\theta}}^T(\boldsymbol{\theta}_0) W F_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0) \right)^{-1}, \tag{25}$$

the sensitivity matrix is given by

$$F_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = F_{\boldsymbol{\theta}}^N(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial f(t_1;\boldsymbol{\theta})}{\partial \theta_1} & \frac{\partial f(t_1;\boldsymbol{\theta})}{\partial \theta_2} & \cdots & \frac{\partial f(t_1;\boldsymbol{\theta})}{\partial \theta_{\kappa_\theta}} \\ \vdots & & & \vdots \\ \frac{\partial f(t_N;\boldsymbol{\theta})}{\partial \theta_1} & \frac{\partial f(t_N;\boldsymbol{\theta})}{\partial \theta_2} & \cdots & \frac{\partial f(t_N;\boldsymbol{\theta})}{\partial \theta_{\kappa_\theta}} \end{pmatrix}, \tag{26}$$

and the matrix W is defined by $W^{-1} = \operatorname{diag}\left(w_1^2, \ldots, w_N^2 \right)$.

Note that because $\boldsymbol{\theta}_0$ and σ_0^2 are unknown, the estimates $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\text{WLS}}$ and $\hat{\sigma}^2 = \hat{\sigma}^2_{\text{WLS}}$ will be used in (25) to calculate

$$\Sigma_0^N \approx \hat{\Sigma}^N = \hat{\sigma}^2 \left(F_{\boldsymbol{\theta}}^T(\hat{\boldsymbol{\theta}}) W F_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) \right)^{-1},$$

where we take the approximation

$$\sigma_0^2 \approx \hat{\sigma}^2_{\text{WLS}} = \frac{1}{N - \kappa_\theta} \sum_{j=1}^{N} \frac{1}{w_j^2} [y_j - f(t_j; \hat{\boldsymbol{\theta}})]^2.$$

We can then approximate the standard errors of $\boldsymbol{\theta}_{\text{WLS}}$ by taking the square roots of the diagonal elements of $\hat{\Sigma}^N$.

Generalized Least Squares: Definition and Motivation

A method motivated by the WLS (as we have presented it above) involves the so-called Generalized Least Squares (GLS) estimator. To define the *random vector* $\boldsymbol{\theta}_{\text{GLS}}$ [26, Chapter 3] and [43, p. 69], the following *normal equations* are solved for the estimator $\boldsymbol{\theta}_{\text{GLS}}$:

$$\sum_{j=1}^{N} f^{-2\gamma}(t_j;\boldsymbol{\theta}_{\text{GLS}})[Y_j - f(t_j;\boldsymbol{\theta}_{\text{GLS}})]\nabla f(t_j;\boldsymbol{\theta}_{\text{GLS}}) = \mathbf{0}_{\kappa_\theta}, \tag{27}$$

where Y_j satisfies

$$Y_j = f(t_j;\boldsymbol{\theta}_0) + f^{\gamma}(t_j;\boldsymbol{\theta}_0)\widetilde{\mathcal{E}}_j,$$

and

$$\nabla f(t_j;\boldsymbol{\theta}) = \left(\frac{\partial f(t_j;\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial f(t_j;\boldsymbol{\theta})}{\partial \theta_{\kappa_\theta}}\right)^T.$$

The quantity θ_{GLS} is a random vector, hence if $\{y_j\}_{j=1}^N$ is a *realization* of $\{Y_j\}_{j=1}^N$, then solving

$$\sum_{j=1}^{N} f^{-2\gamma}(t_j;\theta)[y_j - f(t_j;\theta)]\nabla f(t_j;\theta) = \mathbf{0}_{\kappa_\theta} \tag{28}$$

for θ will provide an estimate for θ_{GLS}.

The GLS equation (28) can be motivated by examining the special weighted least squares estimate

$$\hat{\boldsymbol{\theta}}_{\text{WLS}} = \arg \min_{\boldsymbol{\theta} \in \Omega_\theta} \sum_{j=1}^{N} w_j [y_j - f(t_j; \boldsymbol{\theta})]^2. \tag{29}$$

for a given $\{w_j\}$. If we differentiate the sum of squares in (29) with respect to $\boldsymbol{\theta}$ and *then* choose $w_j = f^{-2\gamma}(t_j; \boldsymbol{\theta})$, an estimate $\hat{\boldsymbol{\theta}}_{\text{GLS}}$ is obtained by solving

$$\sum_{j=1}^{N} w_j [y_j - f(t_j; \boldsymbol{\theta})] \nabla f(t_j; \boldsymbol{\theta}) = \mathbf{0}_{\kappa_\theta}$$

for $\boldsymbol{\theta}$, i.e., solving (28). However, we note the GLS relationship (28) does *not* follow from minimizing the weighted least squares with weights chosen as $w_j = f^{-2\gamma}(t_j; \boldsymbol{\theta})$ (see p. 89 of [43]).

Another motivation for the GLS estimating equations (27) and (28) can be found in [23]. In that text, Carroll and Ruppert claim that if the data are distributed according to the gamma distribution, then the *maximum-likelihood estimate* for θ (a standard approach when one assumes that the distribution for the measurement error is completely known–to be discussed later) is the solution to

$$\sum_{j=1}^{N} f^{-2}(t_j;\theta)[y_j - f(t_j;\theta)]\nabla f(t_j;\theta) = \mathbf{0}_{\kappa_\theta},$$

which is equivalent to the corresponding GLS estimating equations (28) with $\gamma = 1$. (See Chapter 3 of [12])

The GLS estimator $\boldsymbol{\theta}_{\text{GLS}} = \boldsymbol{\theta}^N_{\text{GLS}}$ has the following asymptotic properties [27, 43]:

$$\boldsymbol{\theta}_{\text{GLS}} \sim \mathcal{N}(\boldsymbol{\theta}_0, \Sigma_0^N), \tag{30}$$

where

$$\Sigma_0^N \approx \sigma_0^2 \left(F_{\boldsymbol{\theta}}^T(\boldsymbol{\theta}_0) W(\boldsymbol{\theta}_0) F_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0) \right)^{-1}, \tag{31}$$

the sensitivity matrix is given by (26) and the matrix $W(\boldsymbol{\theta})$ is defined by $W^{-1}(\boldsymbol{\theta}) = \text{diag}\left(f^{2\gamma}(t_1; \boldsymbol{\theta}), \ldots, f^{2\gamma}(t_N; \boldsymbol{\theta}) \right)$.

Note that because $\boldsymbol{\theta}_0$ and σ_0^2 are unknown, the estimates $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\text{GLS}}$ and $\hat{\sigma}^2 = \hat{\sigma}^2_{\text{GLS}}$ will again be used in (31) to calculate

$$\Sigma_0^N \approx \hat{\Sigma}^N = \hat{\sigma}^2 \left(F_{\boldsymbol{\theta}}^T(\hat{\boldsymbol{\theta}}) W(\hat{\boldsymbol{\theta}}) F_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) \right)^{-1},$$

where we take the approximation

$$\sigma_0^2 \approx \hat{\sigma}^2_{\text{GLS}} = \frac{1}{N - \kappa_\theta} \sum_{j=1}^{N} \frac{1}{f^{2\gamma}(t_j; \hat{\boldsymbol{\theta}})} [y_j - f(t_j; \hat{\boldsymbol{\theta}})]^2.$$

We can then approximate the standard errors of $\boldsymbol{\theta}_{\text{GLS}}$ by taking the square roots of the diagonal elements of $\hat{\Sigma}^N$.

Computation of $\hat{\Sigma}^N$, Standard Errors, and Confidence Intervals

- We return to the case of N scalar longitudinal observations and consider the OLS case (the extension of these ideas to vectors is completely straight-forward).
- Recall that in the ordinary least squares approach, we seek to use a realization $\{y_j\}$ of the observation process $\{Y_j\}$ along with the model to determine a vector $\hat{\theta}^N_{\text{OLS}}$ where

$$\hat{\theta}^N_{\text{OLS}} = \arg\min_{\theta \in \Omega_\theta} J^N_{\text{OLS}}(\theta; \mathbf{y}) = \arg\min_{\theta \in \Omega_\theta} \sum_{j=1}^{N} [y_j - f(t_j; \theta)]^2. \quad (32)$$

- Since Y_j is a random variable, the corresponding estimator θ^N_{OLS} (here we wish to emphasize the dependence on the sample size N) is also a random vector with a distribution called the *sampling distribution*.

Remarks on sampling distribution:

- Knowledge of this sampling distribution provides uncertainty information (e.g., standard errors) for the numerical values of $\hat{\theta}^N$ obtained using a specific data set $\{y_j\}$.
- In particular, loosely speaking the sampling distribution characterizes the distribution of possible values the estimator could take on across all possible realizations with data of size N that could be collected.
- The standard errors thus approximate the extent of variability in possible parameter values across all possible realizations, and hence provide a measure of the extent of uncertainty involved in estimating θ using a specific estimator and sample size N in actual data collection.

Computation of sensitivities

- The quantity $F_{\boldsymbol{\theta}}$ is the fundamental entity in computational aspects of this theory.
- There are typically several ways to compute the matrix $F_{\boldsymbol{\theta}}$ (which actually is composed of the well known **sensitivity functions** widely used in applied mathematics and engineering.
- First, the elements of the matrix $F_{\boldsymbol{\theta}} = (F_{\boldsymbol{\theta}jk})$ can always be estimated using the forward difference

$$F_{\boldsymbol{\theta}jk}(\boldsymbol{\theta}) = \frac{\partial f(t_j;\boldsymbol{\theta})}{\partial \theta_k} \approx \frac{f(t_j;\boldsymbol{\theta}+\boldsymbol{h}_k) - f(t_j;\boldsymbol{\theta})}{|\boldsymbol{h}_k|},$$

where $\boldsymbol{h}_k$ is a κ_θ-vector with a nonzero entry in only the k^{th} component which is chosen "small" and $|\cdot|$ is the Euclidean norm in $\mathbb{R}^{\kappa_\theta}$.

- The choice of $\boldsymbol{h}_k$ can be problematic in practice, i.e., what does "small" mean, especially when the parameters may vary by orders of magnitude?
- In some cases the function $f(t_j;\boldsymbol{\theta})$ may be sufficiently simple to allow one to derive analytical expressions for $F_{\boldsymbol{\theta}}$. Alternatively, if the $f(t_j;\boldsymbol{\theta})$ correspond to longitudinal observations $f(t_j;\boldsymbol{\theta}) = \mathcal{C}\boldsymbol{x}(t_j;\boldsymbol{\theta})$ of solutions to a parameterized n-vector differential equation system $\dot{\boldsymbol{x}} = \boldsymbol{g}(t,\boldsymbol{x}(t),\boldsymbol{q})$ as in (1)-(2), then one can use the $n \times \kappa_\theta$ matrix **sensitivity equations** (see [5, 7] and the references therein)

$$\frac{d}{dt}\left(\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}}\right) = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}} + \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{\theta}} \tag{33}$$

to obtain

$$\frac{\partial f(t_j;\boldsymbol{\theta})}{\partial \theta_k} = \mathcal{C}\frac{\partial \boldsymbol{x}(t_j,\boldsymbol{\theta})}{\partial \theta_k}.$$

In order to quantify the variation in the state variable $\boldsymbol{x}(t)$ with respect to changes in the parameters $\boldsymbol{q}$ and the initial conditions $\boldsymbol{x}_0$, we are naturally led to consider the individual *(traditional) sensitivity functions (TSF)* defined by the derivatives

$$\boldsymbol{s}_{q_k}(t) = \frac{\partial \boldsymbol{x}}{\partial q_k}(t) = \frac{\partial \boldsymbol{x}}{\partial q_k}(t, \boldsymbol{\theta}), \qquad k = 1, \ldots, \kappa_q, \tag{34}$$

and

$$\boldsymbol{r}_{x_{0l}}(t) = \frac{\partial \boldsymbol{x}}{\partial x_{0l}}(t) = \frac{\partial \boldsymbol{x}}{\partial x_{0l}}(t, \boldsymbol{\theta}), \qquad l = 1, \ldots, n, \tag{35}$$

where x_{0l} is the l^{th} component of the initial condition $\boldsymbol{x}_0$. If the function $\boldsymbol{g}$ is sufficiently regular, the solution $\boldsymbol{x}$ is differentiable with respect to q_k and x_{0l}, and therefore the sensitivity functions $\boldsymbol{s}_{q_k}$ and $\boldsymbol{r}_{x_{0l}}$ are well–defined.

Remarks on local sensitivites:

- Because they are defined by partial derivatives which have a *local* character, the sensitivity functions are also local in nature.
- Sensitivity and insensitivity ($\boldsymbol{s}_{q_k} = \partial \boldsymbol{x}/\partial q_k$ not close to zero and very close to zero, respectively) depend on the time interval, the state values $\boldsymbol{x}$, and the values of $\boldsymbol{\theta}$ for which they are considered.
- For example, in a certain time subinterval we might find $\boldsymbol{s}_{q_k}$ small so that the state variable $\boldsymbol{x}$ is *insensitive* to the parameter q_k on that particular interval.
- The same function $\boldsymbol{s}_{q_k}$ can take large values on a different subinterval, indicating to us that the state variable $\boldsymbol{x}$ is *very sensitive* to the parameter q_k on the latter interval.

From the sensitivity analysis theory for dynamical systems, one finds that $\boldsymbol{s} = (\boldsymbol{s}_{q_1}, \ldots, \boldsymbol{s}_{q_{\kappa q}})$ is an $n \times \kappa_q$ vector function that satisfies the ODE system

$$\begin{aligned} \dot{\boldsymbol{s}}(t) &= \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}(t, \boldsymbol{x}(t;\theta), \boldsymbol{q})\boldsymbol{s}(t) + \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{q}}(t, x(t;\theta), \boldsymbol{q}), \qquad (36) \\ \boldsymbol{s}(t_0) &= \boldsymbol{0}_{n \times \kappa_q}, \end{aligned}$$

which is obtained by differentiating (1)-(2) with respect to $\boldsymbol{q}$. Here the dependence of $\boldsymbol{s}$ on $(t, \boldsymbol{x}(t;\theta))$ as well as $\boldsymbol{q}$ is readily apparent.

In a similar manner, the sensitivity functions with respect to the components of the initial condition $\boldsymbol{x}_0$ define an $n \times n$ vector function $\boldsymbol{r} = (\boldsymbol{r}_{x_{01}}, \ldots, \boldsymbol{r}_{x_{0n}})$, which satisfies

$$\begin{aligned} \dot{\boldsymbol{r}}(t) &= \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}(t, \boldsymbol{x}(t;\theta), \boldsymbol{q})\boldsymbol{r}(t), \\ \boldsymbol{r}(t_0) &= \mathbf{I}_n. \end{aligned} \tag{37}$$

This is obtained by differentiating (1)-(2) with respect to the initial conditions $\boldsymbol{x}_0$. The equations (36) and (37) are used in conjunction with (i.e., usually solved simultaneously with) equations (1)-(2) to numerically compute the sensitivities $\boldsymbol{s}$ and $\boldsymbol{r}$ for general cases when the function $\boldsymbol{g}$ is sufficiently complicated to prohibit a closed form solution by direct integration. These can be succinctly written as a system for $\dfrac{\partial \boldsymbol{x}}{\partial \boldsymbol{\theta}} = \left(\dfrac{\partial \boldsymbol{x}}{\partial \boldsymbol{q}}, \dfrac{\partial \boldsymbol{x}}{\partial \boldsymbol{x}_0} \right)$ given by (33).

As we have already noted, since $\boldsymbol{\theta}_0$ and σ_0 are unknown, we will use their estimates to make the approximation

$$\Sigma_0^N \approx \sigma_0^2[F_{\boldsymbol{\theta}}^N(\boldsymbol{\theta}_0)^T F_{\boldsymbol{\theta}}^N(\boldsymbol{\theta}_0)]^{-1} \approx \hat{\Sigma}^N(\hat{\boldsymbol{\theta}}_{\text{OLS}}^N) = \hat{\sigma}^2[F_{\boldsymbol{\theta}}^N(\hat{\boldsymbol{\theta}}_{\text{OLS}}^N)^T F_{\boldsymbol{\theta}}^N(\hat{\boldsymbol{\theta}}_{\text{OLS}}^N)]^{-1}, \tag{38}$$

where the approximation $\hat{\sigma}^2$ of σ_0^2, as discussed earlier, is given by

$$\sigma_0^2 \approx \hat{\sigma}^2 = \frac{1}{N - \kappa_\theta} \sum_{j=1}^{N} [y_j - f(t_j; \hat{\boldsymbol{\theta}}_{\text{OLS}}^N)]^2. \tag{39}$$

Standard errors to be used in the confidence interval calculations are given by $\text{SE}_k(\hat{\boldsymbol{\theta}}^N) = \sqrt{\hat{\Sigma}_{kk}^N(\hat{\boldsymbol{\theta}}^N)}$, $k = 1, 2, \ldots, \kappa_\theta$ (see [25]).

To compute the confidence intervals (at the $100(1-\alpha)\%$ level) for the estimated parameters in our example, we define the confidence intervals associated with the estimated parameters so that

$$\text{Prob}\{\theta_k^N - t_{1-\alpha/2}\text{SE}_k(\hat{\boldsymbol{\theta}}^N) < \theta_{0k} < \theta_k^N + t_{1-\alpha/2}\text{SE}_k(\hat{\boldsymbol{\theta}}^N)\} = 1-\alpha, \quad (40$$

where $\alpha \in [0,1]$ and $t_{1-\alpha/2} \in \mathbb{R}^+$. For a realization $\boldsymbol{y}$ and estimates $\hat{\boldsymbol{\theta}}^N$, the corresponding confidence intervals are given by

$$[\hat{\theta}_k^N - t_{1-\alpha/2}\text{SE}_k(\hat{\boldsymbol{\theta}}^N), \hat{\theta}_k^N + t_{1-\alpha/2}\text{SE}_k(\hat{\boldsymbol{\theta}}^N)]. \quad (41)$$

Given a small α value (e.g., $\alpha = 0.05$ for 95% confidence intervals), the critical value $t_{1-\alpha/2}$ is computed from the *student's t* distribution $t^{N-\kappa_\theta}$ with $N-\kappa_\theta$ degrees of freedom. The value of $t_{1-\alpha/2}$ is determined by $\text{Prob}\{T \geq t_{1-\alpha/2}\} = \alpha/2$ where $T \sim t^{N-\kappa_\theta}$.

Remarks on asymptotic confidence intervals:

- In general, a confidence interval is constructed so that, if the confidence interval could be constructed for each possible realization of data of size N that could have been collected, $100(1-\alpha)\%$ of the intervals so constructed would contain the true value θ_{0k}.
- Thus, a confidence interval provides further information on the extent of uncertainty involved in estimating $\boldsymbol{\theta}_0$ using the given estimator and sample size N.

Begin Sensitivities Interlude...

Example of how to calculate sensitivities and construct the Fisher Information Matrix for the logistic model.

End Sensitivities Interlude...

Investigation of Statistical Assumptions

- The form of error in the data (which of course is rarely known) dictates which method from those discussed above one should choose.
- The OLS method is most appropriate for constant variance observations of the form $Y_j = f(t_j; \theta_0) + \widetilde{\mathcal{E}}_j$ whereas the GLS should be used for problems in which we have nonconstant variance observations $Y_j = f(t_j; \theta_0) + f^{\gamma}(t_j; \theta_0)\widetilde{\mathcal{E}}_j$.
- We emphasize that to obtain *the correct standard errors* in an inverse problem calculation, the OLS method (and *corresponding asymptotic formulas*) must be used with constant variance generated data, while the GLS method (and *corresponding asymptotic formulas*) should be applied to nonconstant variance generated data.

- An incorrect error model can lead to *incorrect conclusions*.
- In either case, the standard error calculations are not valid unless the correct formulas (which depend on the error structure) are employed.
- Unfortunately, it is very difficult to ascertain the structure of the error, and hence the correct method to use, without *a priori* information.
- Although the error structure cannot definitively be determined, two residual tests can be performed *after* the estimation procedure has been completed to assist in concluding whether or not the correct asymptotic statistics were used.

Residual Plots

- We will show results from simulation studies in these slides to assist in understanding the behavior of the model in inverse problems with different types of data with respect to mis-specification of the statistical model.
- For example, we consider a statistical model with constant variance (CV) noise ($\gamma = 0$)

$$Y_j = f(t_j; \boldsymbol{\theta}_0) + \widetilde{\mathcal{E}}_j, \qquad \text{Var}(Y_j) = \sigma_0^2,$$

and another with nonconstant variance (NCV) noise ($\gamma = 1$)

$$Y_j = f(t_j; \boldsymbol{\theta}_0)(1 + \widetilde{\mathcal{E}}_j), \qquad \text{Var}(Y_j) = \sigma_0^2 f^2(t_j; \boldsymbol{\theta}_0).$$

- We obtain a data set by considering a *realization* $\{y_j\}_{j=1}^N$ of the random variables $\{Y_j\}_{j=1}^N$ through a realization of $\{\widetilde{\mathcal{E}}_j\}_{j=1}^N$, and then calculate an estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$ using the OLS or GLS procedure.

Testing for a constant variance error model

- We will use the *residuals* $r_j = y_j - f(t_j; \hat{\theta})$ to test whether the data set is *i.i.d.* and possesses the assumed variance structure.
- If a data set has constant variance then

$$Y_j = f(t_j; \theta_0) + \widetilde{\mathcal{E}}_j \quad \text{or} \quad \widetilde{\mathcal{E}}_j = Y_j - f(t_j; \theta_0),$$

and hence the residuals r_j are approximations to realizations of the errors $\widetilde{\mathcal{E}}_j$ (when it is tacitly assumed that $\hat{\theta} \approx \theta_0$).

Testing for a constant variance error model

- Test 1: Since it is assumed that the errors $\widetilde{\varepsilon}_j$ are *i.i.d.*, a plot of the residuals $r_j = y_j - f(t_j; \hat{\theta})$ vs. t_j should be random (and neither increasing nor decreasing with time).
- Test 2: The error in the constant variance case does not depend on $f(t_j; \theta_0)$, and so a plot of the residuals $r_j = y_j - f(t_j; \hat{\theta})$ vs. $f(t_j; \hat{\theta})$ should also be random (and neither increasing nor decreasing).
- Therefore, *if* the error has constant variance, then a plot of the residuals $r_j = y_j - f(t_j; \hat{\theta})$ against t_j and against $f(t_j; \hat{\theta})$ should both be random.
- If not, then the constant variance assumption is suspect.

Testing for a constant variance error model

- What to expect if this residual test is applied to a data set that has nonconstant variance (NCV) generated error?
- What happens if the data are incorrectly assumed to have CV error when in fact they have NCV error?
- Since in the NCV example, $R_j = Y_j - f(t_j;\theta_0) = f(t_j;\theta_0)\,\widetilde{\mathcal{E}}_j$ depends upon the deterministic model $f(t_j;\theta_0)$, we should expect that a plot of the residuals $r_j = y_j - f(t_j;\hat{\theta})$ vs. t_j should exhibit some type of pattern.
- Also, the residuals actually depend on $f(t_j;\hat{\theta})$ in the NCV case, and so as $f(t_j;\hat{\theta})$ increases the variation of the residuals $r_j = y_j - f(t_j;\hat{\theta})$ should increase as well.

Testing for a non-constant variance error model

- If a data set has nonconstant variance generated data, then

$$Y_j = f(t_j;\theta_0) + f(t_j;\theta_0)\,\widetilde{\varepsilon}_j \quad \text{or} \quad \widetilde{\varepsilon}_j = \frac{Y_j - f(t_j;\theta_0)}{f(t_j;\theta_0)}.$$

- Test 1: If the distributions of $\widetilde{\varepsilon}_j$ are *i.i.d.*, then a plot of the *modified residuals* $r_j^m = (y_j - f(t_j;\hat{\theta}))/f(t_j;\hat{\theta})$ vs. t_j should be random for nonconstant variance generated data.
- Test 2: A plot of $r_j^m = (y_j - f(t_j;\hat{\theta}))/f(t_j;\hat{\theta})$ vs. $f(t_j;\hat{\theta})$ should also be random.

Testing for a non-constant variance error model

- What if the data are incorrectly assumed to have non-constant variance error when in fact they have constant variance error?
- Since $Y_j - f(t_j;\theta_0) = \widetilde{\mathcal{E}}_j$ in the constant variance case, we should expect that a plot of $r_j^m = (y_j - f(t_j;\hat{\theta}))/f(t_j;\hat{\theta})$ vs. t_j as well as that for $r_j^m = (y_j - f(t_j;\hat{\theta}))/f(t_j;\hat{\theta})$ vs. $f(t_j;\hat{\theta})$ will possess some distinct pattern (such as a fan shape).

Troubleshooting residual plot analysis

- There are two further issues regarding residual plots. As we shall see by examples, some data sets might have values that are repeated or nearly repeated a large number of times (for example when sampling near an equilibrium of a mathematical model or when sampling a periodic system over many periods).
- If a certain value is repeated numerous times (e.g., f_{repeat}) then any plot with $f(t_j; \hat{\theta})$ along the horizontal axis should have a cluster of values along the vertical line $x = f_{\text{repeat}}$.
- This feature can easily be removed by excluding the data points corresponding to these high frequency values (or simply excluding the corresponding points in the residual plots).

Troubleshooting residual plot analysis

- Another common technique when plotting against model predictions is to plot against $\ln(f(t_j;\hat{\theta}))$ instead of $f(t_j;\hat{\theta})$ itself which has the effect of "stretching out" plots at the ends.
- Also, note that the model value $f(t_j;\hat{\theta})$ could possibly be zero or very near zero, in which case the modified residuals $r_j^m = (y_j - f(t_j;\hat{\theta}))/f(t_j;\hat{\theta})$ would be undefined or extremely large.
- To remedy this situation one might exclude values very close to zero (in either the plots or in the data themselves).

An Example Using Residual Plots: Logistic Growth

We illustrate residual plot techniques by exploring a widely used model – the logistic population growth model of Verhulst/Pearl [36]

$$\dot{x} = rx\left(1 - \frac{x}{K}\right), \quad x(0) = x_0. \tag{42}$$

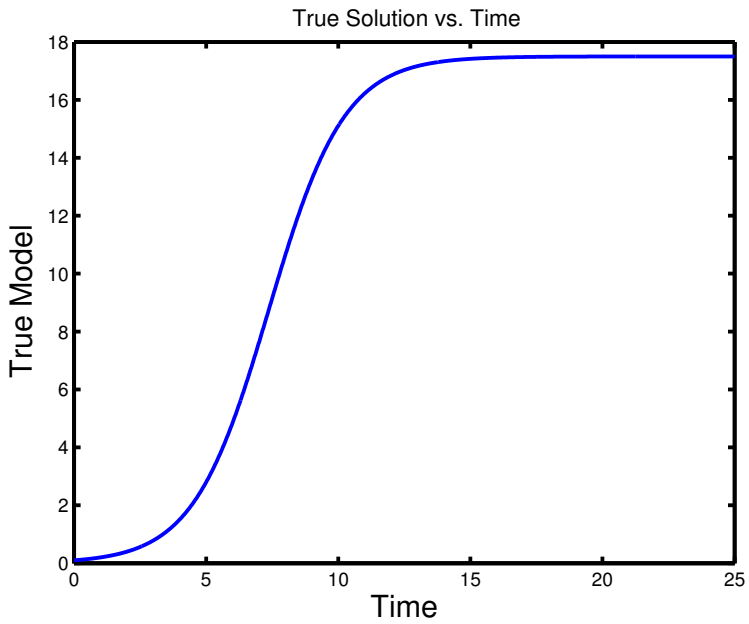
Here K is the population's carrying capacity, r is the intrinsic growth rate and x_0 is the initial population size. This well-known logistic model describes how populations grow when constrained by resources or competition. The closed form solution of this simple model is given by

$$x(t) = \frac{K\, x_0 e^{rt}}{K + x_0\left(e^{rt} - 1\right)}. \tag{43}$$

An Example Using Residual Plots: Logistic Growth

- The left plot in Figure 1 depicts the solution of the logistic model with $K = 17.5$, $r = 0.7$ and $x_0 = 0.1$ for $0 \leq t \leq 25$.
- If high frequency repeated or nearly repeated values (i.e., near the initial value x_0 or near the asymptote $x = K$) are removed from the original plot, the resulting truncated plot is given in the right panel of the figure on the next slide (there are no near zero values for this function).

An Example Using Residual Plots: Logistic Growth



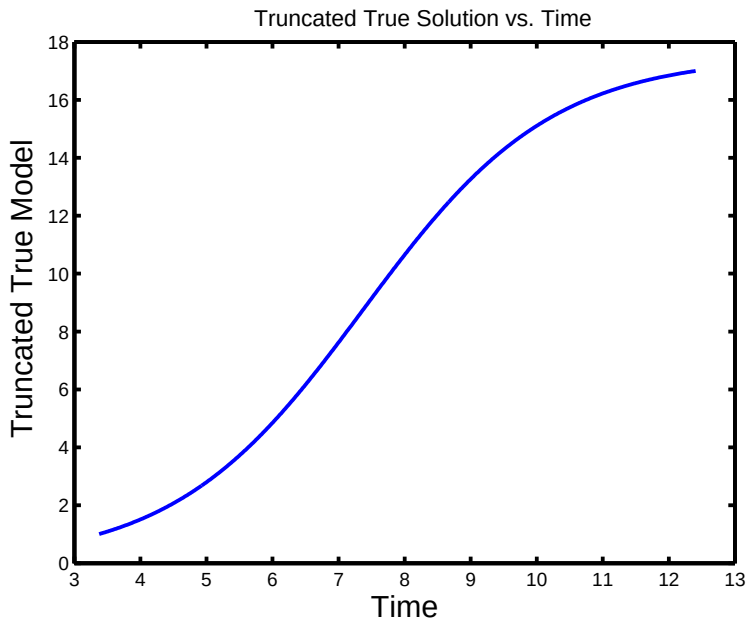


Figure: Original and truncated logistic curve with $K = 17.5$, $r = 0.7$ and $x_0 = 0.1$.

An Example Using Residual Plots: Logistic Growth

- For this example we generated both CV and NCV noisy data (we sampled from $\mathcal{N}(0, 25 \times 10^{-4})$ distributed random variables to obtain realizations of $\widetilde{\mathcal{E}}_j$) and obtained estimates $\hat{\theta}$ of θ_0 by applying either the OLS or GLS method to a realization $\{y_j\}_{j=1}^N$ of the random variables $\{Y_j\}_{j=1}^N$.
- The initial guesses $\theta_{init} = \hat{\theta}^{(0)}$ along with estimates for each method and error structure are given in the Tables below.
- Result: As expected, both methods do a good job of estimating θ_0, however the error structure was not always correctly specified since incorrect asymptotic formulas were used in some cases.

An Example Using Residual Plots: Logistic Growth

Estimation using the OLS procedure with CV data.

θ_{init}	θ_0	$\hat{\theta}_{\text{OLS}}^{\text{CV}}$	$\text{SE}(\hat{\theta}_{\text{OLS}}^{\text{CV}})$	$\hat{\theta}_{\text{OLS}}^{\text{TCV}}$	$\text{SE}(\hat{\theta}_{\text{OLS}}^{\text{TCV}})$
17	17.5	1.75e+001	1.58e-003	1.74e+001	6.42e-003
.8	.7	7.00e-001	4.28e-004	7.00e-001	6.58e-004
1.2	.1	9.99e-002	3.15e-004	9.97e-002	4.39e-004

An Example Using Residual Plots: Logistic Growth

Estimation using the GLS procedure with CV data.

θ_{init}	θ_0	$\hat{\theta}^{\text{CV}}_{\text{GLS}}$	$\text{SE}(\hat{\theta}^{\text{CV}}_{\text{GLS}})$	$\hat{\theta}^{\text{TCV}}_{\text{GLS}}$	$\text{SE}(\hat{\theta}^{\text{TCV}}_{\text{GLS}})$
17	17.5	1.75e+001	1.38e-004	1.75e+001	9.12e-005
.8	.7	7.00e-001	7.81e-005	7.01e-001	1.60e-005
1.2	.1	9.99e-002	6.61e-005	9.97e-002	1.21e-005

An Example Using Residual Plots: Logistic Growth

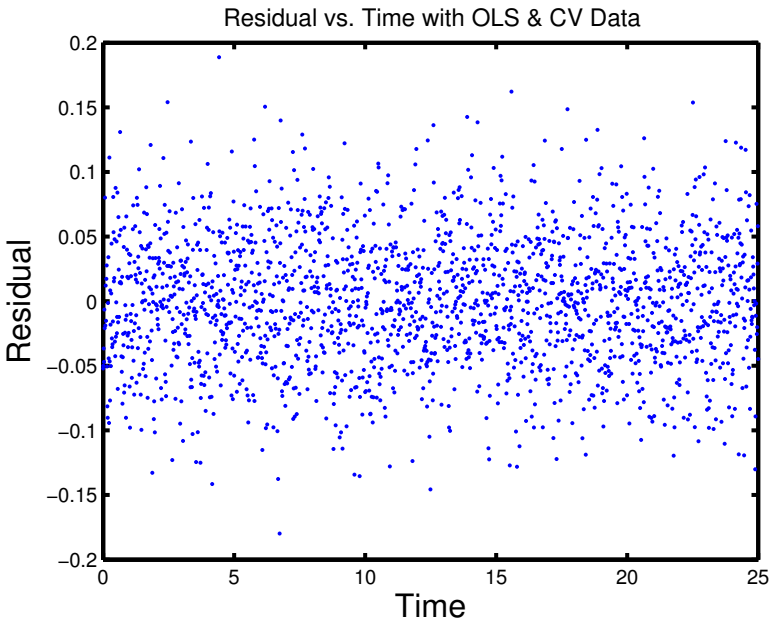
Estimation using the OLS procedure with NCV data.

θ_{init}	θ_0	$\hat{\theta}_{\text{OLS}}^{\text{NCV}}$	$\text{SE}(\hat{\theta}_{\text{OLS}}^{\text{NCV}})$	$\hat{\theta}_{\text{OLS}}^{\text{TNCV}}$	$\text{SE}(\hat{\theta}_{\text{OLS}}^{\text{TNCV}})$
17	17.5	1.75e+001	2.27e-002	1.74e+001	7.16e-002
.8	.7	7.02e-001	6.18e-003	7.09e-001	7.60e-003
1.2	.1	9.95e-002	4.51e-003	9.49e-002	4.83e-003

An Example Using Residual Plots: Logistic Growth

Estimation using the GLS procedure with NCV data.

θ_{init}	θ_0	$\hat{\theta}_{\text{GLS}}^{\text{NCV}}$	$\text{SE}(\hat{\theta}_{\text{GLS}}^{\text{NCV}})$	$\hat{\theta}_{\text{GLS}}^{\text{TNCV}}$	$\text{SE}(\hat{\theta}_{\text{GLS}}^{\text{TNCV}})$
17	17.5	1.75e+001	9.44e-005	1.74e+001	3.13e-004
.8	.7	7.02e-001	5.36e-005	7.09e-001	5.72e-005
1.2	.1	9.93e-002	4.49e-005	9.49e-002	4.12e-005



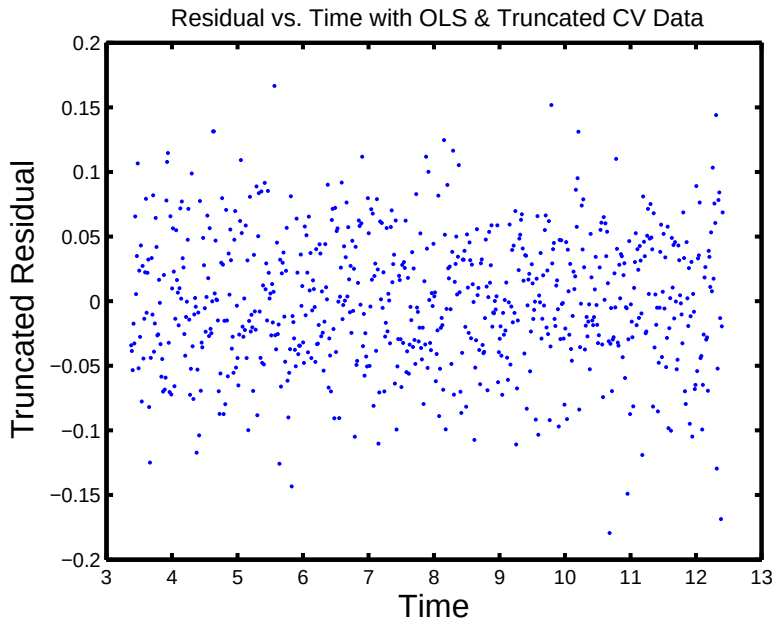
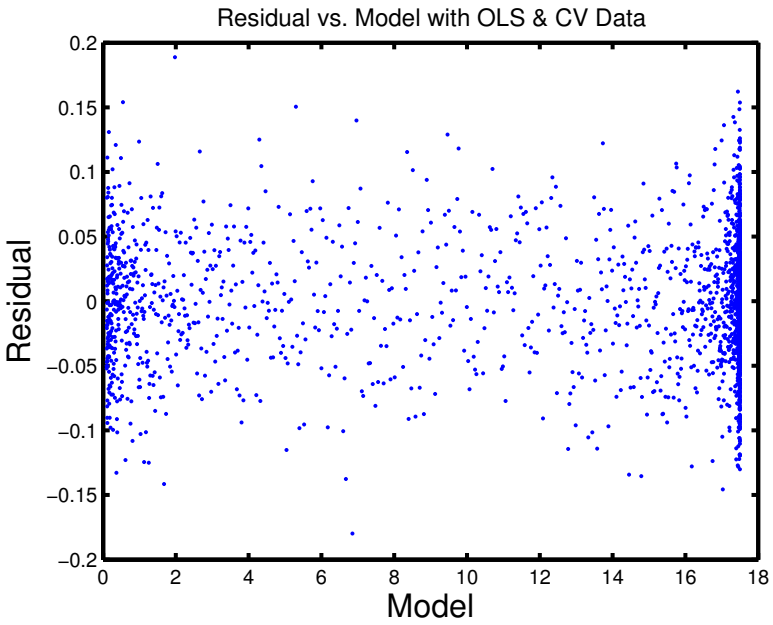


Figure: Residual vs. time plots in tests for independence: Original and truncated logistic curve for $\hat{\theta}^{\mathrm{CV}}_{\mathrm{OLS}}$.



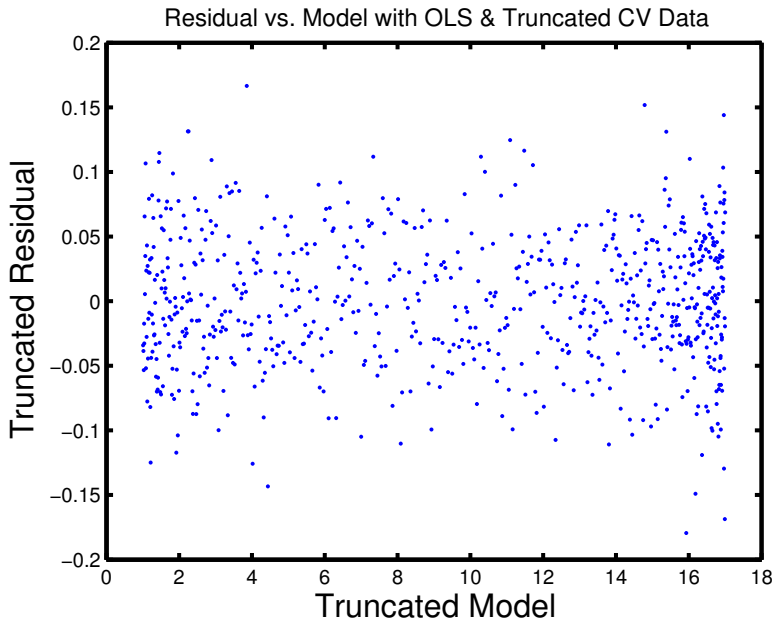
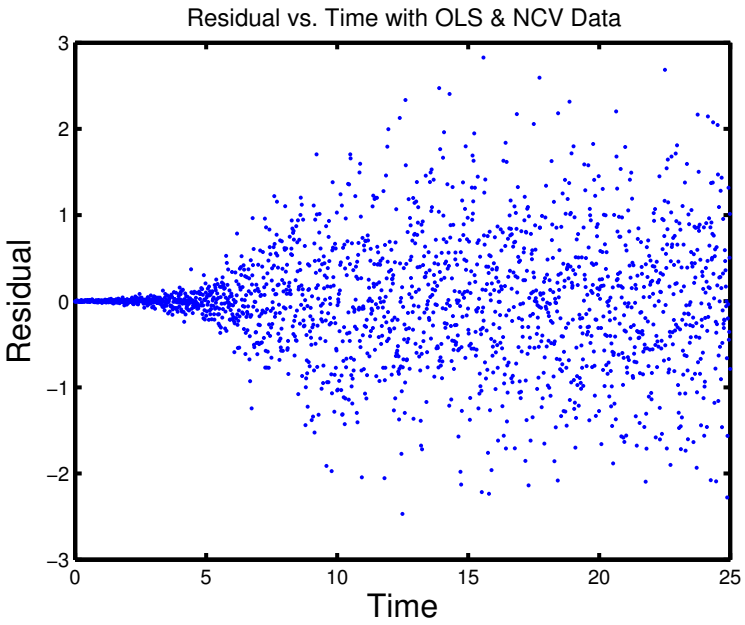


Figure: Residual vs. model plots in tests of form of variance: Original and truncated logistic curve for $\hat{\theta}^{\text{CV}}_{\text{OLS}}$.



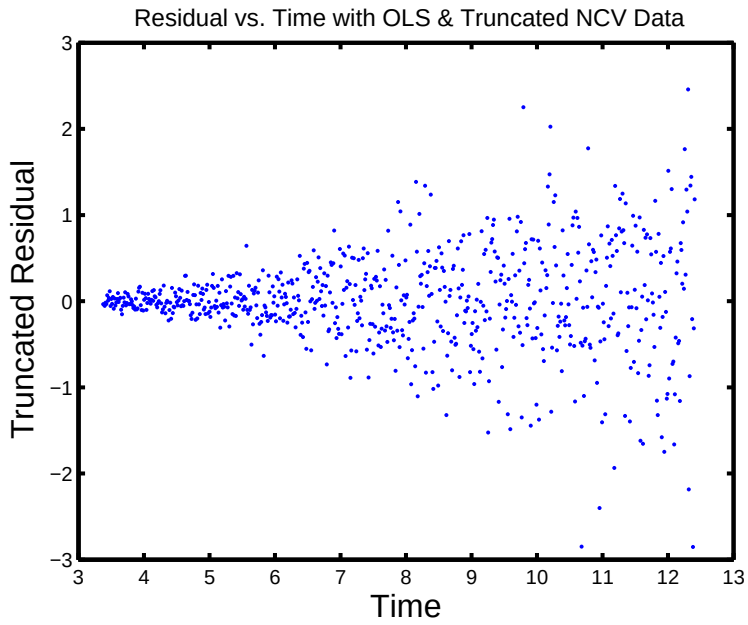
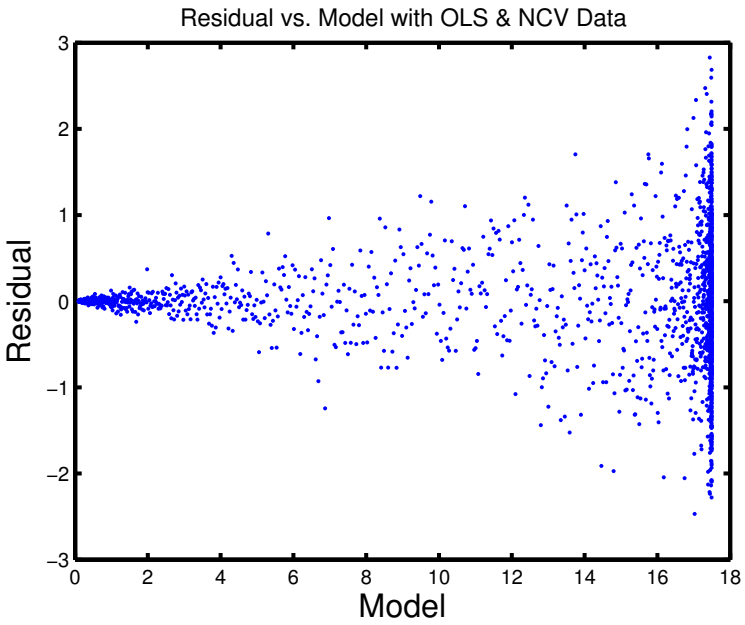


Figure: Residual vs. time plots in tests for independence: Original and truncated logistic curve for $\hat{\theta}_{\text{OLS}}^{\text{NCV}}$.



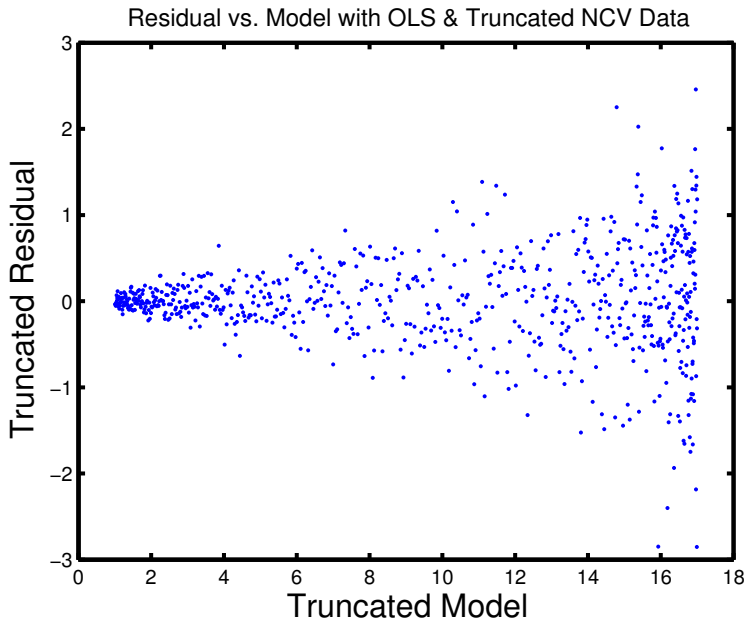
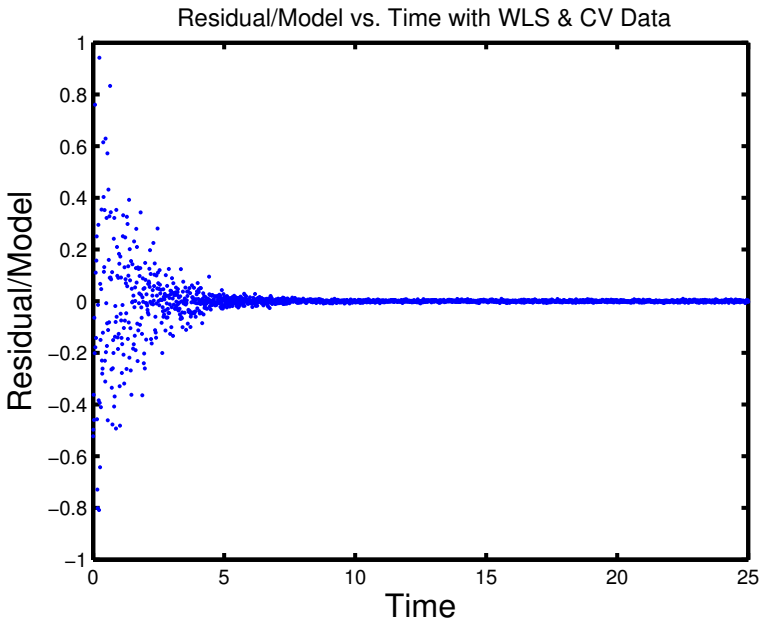


Figure: Residual vs. model plots in tests of form of variance: Original and truncated logistic curve for $\hat{\theta}_{\text{OLS}}^{\text{NCV}}$.



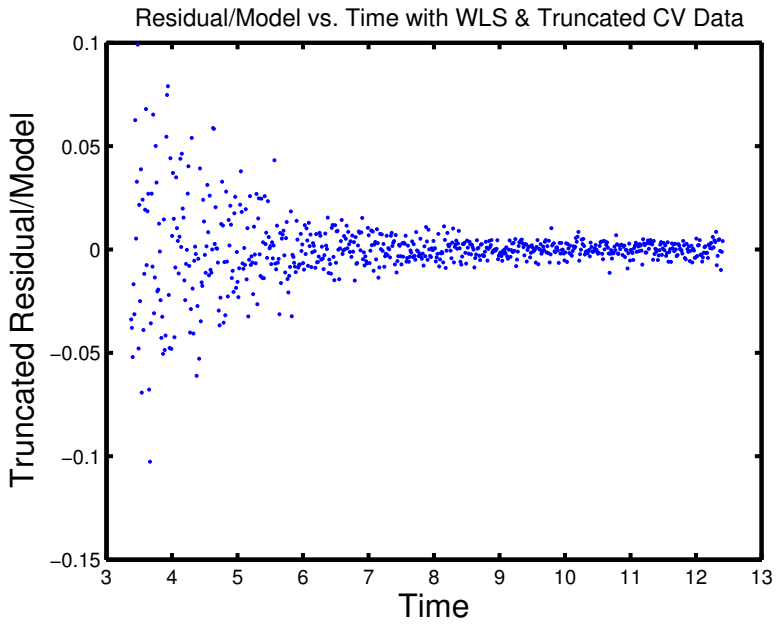
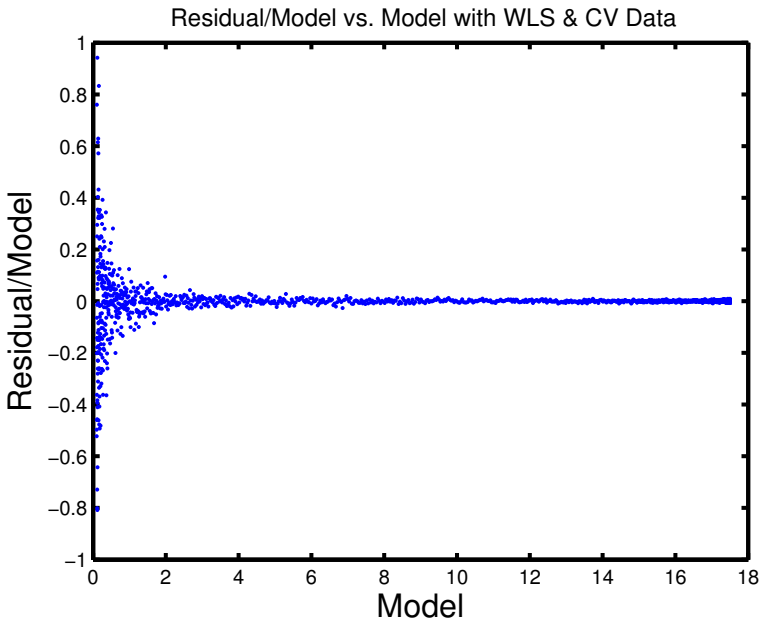


Figure: Residual vs. time plots in tests for independence: Original and truncated logistic curve for $\hat{\theta}^{\mathrm{CV}}_{\mathrm{GLS}}$.



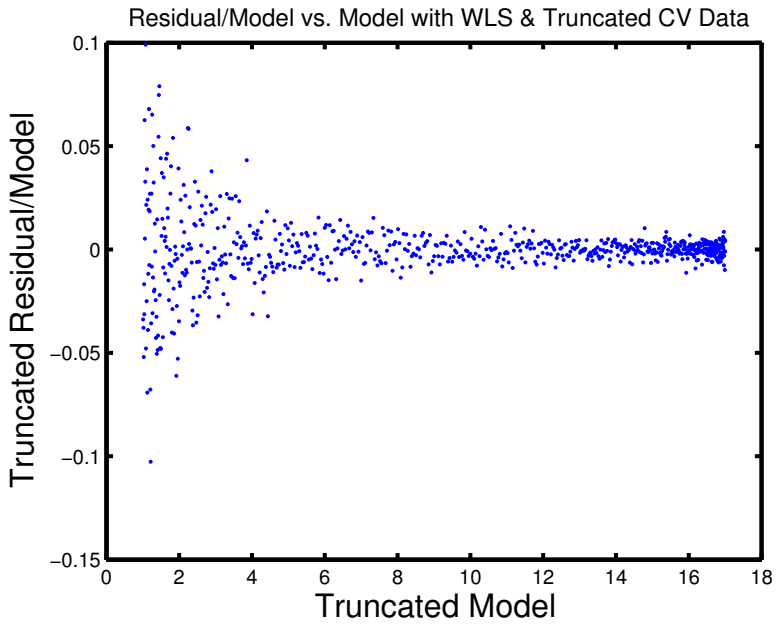
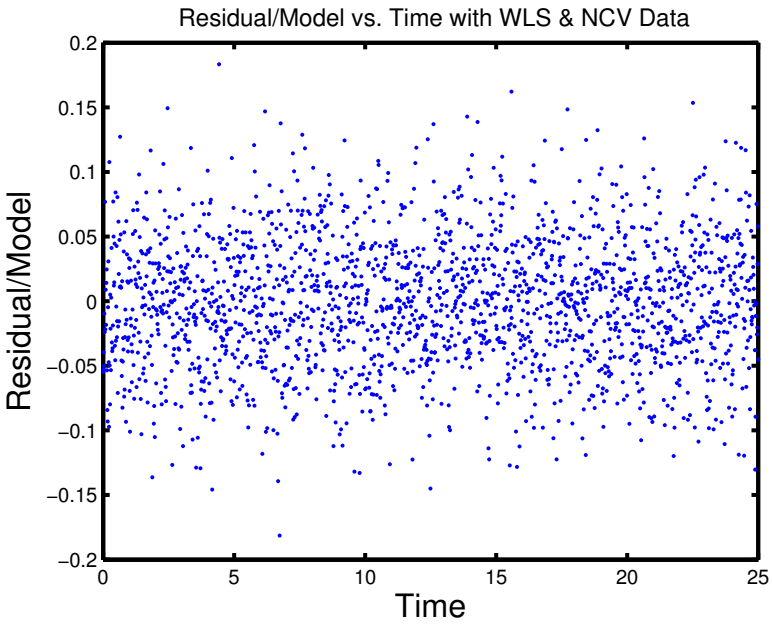


Figure: Modified residual vs. model plots in tests of form of variance: Original and truncated logistic curve for $\hat{\theta}^{\mathrm{CV}}_{\mathrm{GLS}}$.



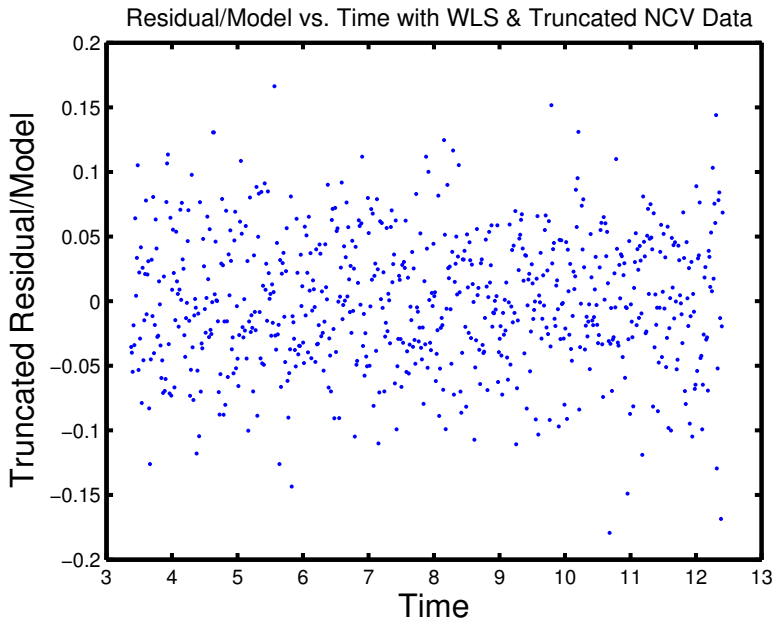
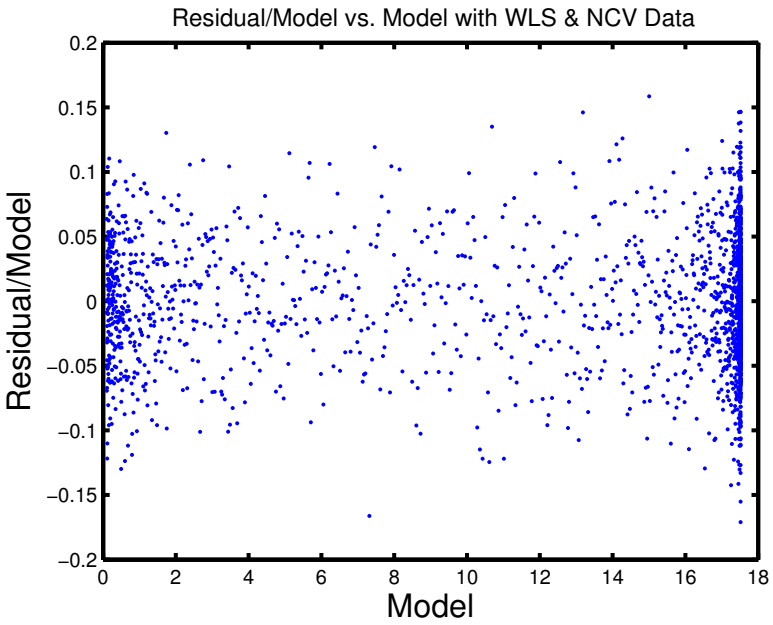


Figure: Modified residual vs. time plots in tests for independence: Original and truncated logistic curve for $\hat{\theta}^{\text{NCV}}_{\text{GLS}}$.



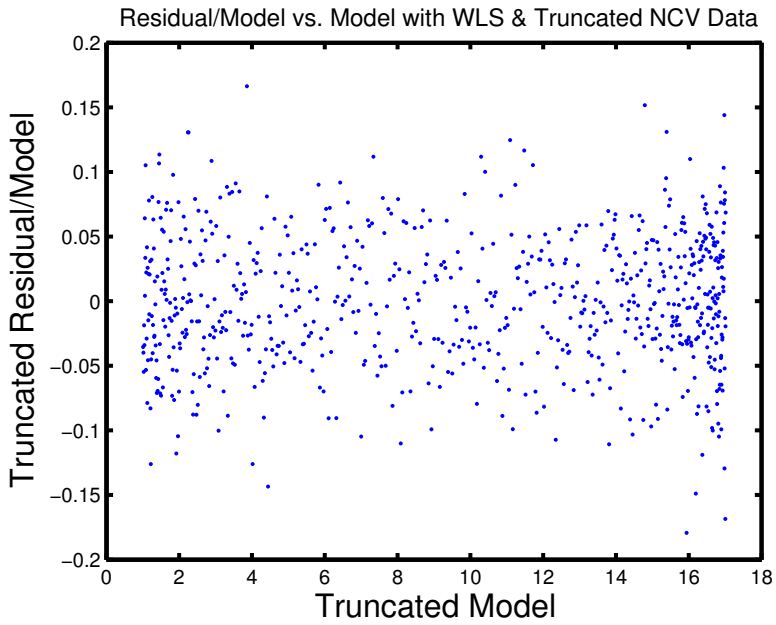


Figure: Modified residual vs. model plots in tests of form of variance: Original and truncated logistic curve for $\hat{\theta}_{\text{GLS}}^{\text{NCV}}$.

Bootstrapping vs. Asymptotic Error Analysis

- In the above discussions we used asymptotic theory to compute uncertainty features for parameter estimates.
- One popular alternative to the asymptotic theory is *bootstrapping* wherein one uses the residuals from an initial estimation to construct a family of samples or simulated data sets.
- One then uses these samples to construct an empirical distribution for the parameters from which the means, standard errors and hence the associated confidence intervals can be readily obtained for the underlying true parameters θ_0.

Bootstrapping Algorithm: Constant Variance Data

Assume we are given experimental data $(t_1, y_1), \ldots, (t_N, y_N)$ for a dynamical system (e.g., the logistic growth model) from an underlying observation process

$$Y_j = f(t_j; \boldsymbol{\theta}_0) + \widetilde{\mathcal{E}}_j, \quad j = 1, \ldots, N, \tag{44}$$

where the $\widetilde{\mathcal{E}}_j$ are independent and identically distributed *(i.i.d.)* with mean zero ($\mathbb{E}(\mathcal{E}_j) = 0$) and constant variance σ_0^2, and $\boldsymbol{\theta}_0$ is the "true value" hypothesized to exist in statistical treatments of data. Associated corresponding realizations $\{y_j\}$ of the random variables $\{Y_j\}$ are given by

$$y_j = f(t_j; \boldsymbol{\theta}_0) + \widetilde{\epsilon}_j.$$

The following algorithm [23, 24, 26, p. 285–287] can be used to compute the *bootstrapping estimate* $\hat{\boldsymbol{\theta}}_{\text{BOOT}}$ of $\boldsymbol{\theta}_0$ and its empirical distribution.

1. First estimate $\hat{\theta}^0$ from the entire sample $\{y_j\}_{j=1}^N$ using OLS.
2. Using this estimate define the standardized residuals

$$\bar{r}_j = \sqrt{\frac{N}{N-\kappa_\theta}}\left(y_j - f(t_j;\hat{\theta}^0)\right)$$

for $j = 1,\ldots,N$. Set $m = 0$.
3. Create a bootstrapping sample of size N using random sampling with replacement from the data (realizations) $\{\bar{r}_1,\ldots,\bar{r}_N\}$ to form a bootstrapping sample $\{r_1^m,\ldots,r_N^m\}$.
4. Create bootstrap sample points

$$y_j^m = f(t_j;\hat{\theta}^0) + r_j^m,$$

where $j = 1,\ldots,N$.
5. Obtain a new estimate $\hat{\theta}^{m+1}$ from the bootstrapping sample $\{y_j^m\}$ using OLS.
6. Set $m = m+1$ and repeat steps 3–5 until $m \geq M$ (e.g., typically $M = 1000$ as in our calculations below).

Bootstrapping Algorithm: Constant Variance Data

We then calculate the mean, standard error, and confidence intervals using the formulae

$$\hat{\theta}_{\text{BOOT}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}^m,$$

$$\text{Var}(\theta_{\text{BOOT}}) = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}^m - \hat{\theta}_{\text{BOOT}})^T(\hat{\theta}^m - \hat{\theta}_{\text{BOOT}}), \quad (45)$$

$$\text{SE}_k(\hat{\theta}_{\text{BOOT}}) = \sqrt{\text{Var}(\theta_{\text{BOOT}})_{kk}}.$$

where θ_{BOOT} denotes the bootstrapping estimator.

Bootstrapping Algorithm: Constant Variance Data

In the above procedures, the $\{\bar{r}_1, \ldots, \bar{r}_N\}$ are realizations of *i.i.d.* random variables $\bar{R}_j$ with the empirical distribution function F_N. It can be shown that

$$\mathbb{E}(\bar{R}_j|F_N) = N^{-1}\sum_{j=1}^{N}\bar{r}_j = 0, \qquad \text{Var}(\bar{R}_j|F_N) = N^{-1}\sum_{j=1}^{N}\bar{r}_j^2 = \hat{\sigma}^2.$$

Results of Numerical Simulations

- We created noisy data sets for the logistic mode using = simulations and a time vector of length $N = 50$ [10].
- The underlying logistic model with the true parameter values $\theta_0 = (17.5, 0.7, 0.1)^T$ was solved for $f(t_j; \theta_0) = x(t_j; \theta_0)$ using the Matlab function *ode45* where $\theta = (K, r, x_0)^T$.
- A noise vector of length N with noise level σ_0, was taken from a random number generator for $\mathcal{N}(0, \sigma_0^2)$. T
- The constant variance data sets were obtained from the equation

$$y_j = f(t_j; \theta_0) + \widetilde{\epsilon}_j.$$

- Constant variance data sets were created for 1%, 5%, and 10% noise, i.e., $\sigma_0 = 0.01, 0.05,$ and 0.1.

Results of Numerical Simulations

- We used the constant variance (CV) data with OLS to carry out the parameter estimation calculations.
- The bootstrapping estimates were computed with $M = 1000$. We use $M = 1000$ because we are computing confidence intervals and not only estimates and standard errors, and Efron and Tibirshani [28] recommend that $M = 1000$ when confidence intervals are to be computed.
- The standard errors $\mathrm{SE}_k(\hat{\theta})$ and corresponding confidence intervals $[\hat{\theta}_k - 1.96\mathrm{SE}_k(\hat{\theta}), \hat{\theta}_k + 1.96\mathrm{SE}_k(\hat{\theta})]$ are listed in tables below.
- We plot the empirical distributions for the case $\sigma_0 = 0.05$; plots in the other two cases are quite similar.

Results of Numerical Simulations

Asymptotic and bootstrap OLS estimates for CV data, $\sigma_0 = 0.01$.

θ	$\hat{\theta}$	$\mathrm{SE}(\hat{\theta})$	95% CI
$\hat{K}_{asy}$	17.498576	0.002021	(17.494615, 17.502537)
$\hat{r}_{asy}$	0.700186	0.000553	(0.699103, 0.701270)
$(\hat{x}_0)_{asy}$	0.100044	0.000407	(0.099247, 0.100841)
$\hat{K}_{boot}$	17.498464	0.001973	(17.494597, 17.502331)
$\hat{r}_{boot}$	0.700193	0.000548	(0.699118, 0.701268)
$(\hat{x}_0)_{boot}$	0.100034	0.000399	(0.099252, 0.100815)

Results of Numerical Simulations

Asymptotic and bootstrap OLS estimates for CV data, $\sigma_0 = 0.05$.

θ	$\hat{\theta}$	$\mathrm{SE}(\hat{\theta})$	95% CI
$\hat{K}_{asy}$	17.486571	0.010269	(17.466444, 17.506699)
$\hat{r}_{asy}$	0.702352	0.002825	(0.696815, 0.707889)
$(\hat{x}_0)_{asy}$	0.098757	0.002050	(0.0947386, 0.102775)
$\hat{K}_{boot}$	17.489658	0.010247	(17.469574, 17.509742)
$\hat{r}_{boot}$	0.702098	0.002938	(0.696339, 0.707857)
$(\hat{x}_0)_{boot}$	0.0990520	0.002152	(0.094834, 0.103270)

Results of Numerical Simulations

Asymptotic and bootstrap OLS estimates for NCV data, $\sigma_0 = 0.1$.

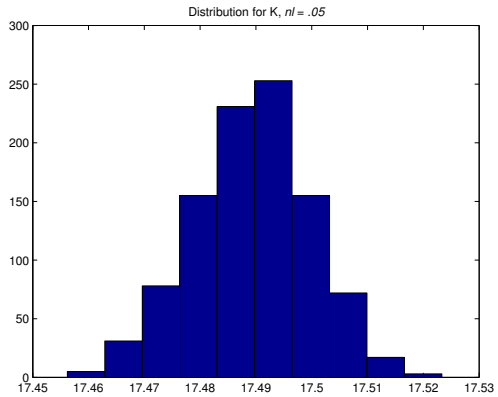
θ	$\hat{\theta}$	$\mathrm{SE}(\hat{\theta})$	95% CI
$\hat{K}_{asy}$	17.081926	0.262907	(16.566629, 17.597223)
$\hat{r}_{asy}$	0.727602	0.078513	(0.573717, 0.881487)
$(\hat{x}_0)_{asy}$	0.082935	0.047591	(-0.010343, 0.176213)
$\hat{K}_{boot}$	17.095648	0.250940	(16.603807, 17.587490)
$\hat{r}_{boot}$	0.733657	0.081852	(0.573228, 0.894087)
$(\hat{x}_0)_{boot}$	0.094020	0.054849	(-0.013484, 0.201524)

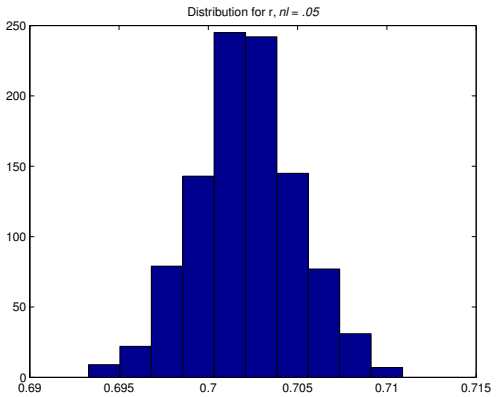
Results of Numerical Simulations

Computational times (sec) for asymptotic theory vs. bootstrapping.

Noise Level	Asymptotic Theory	Bootstrapping
1%	0.017320	4.285640
5%	0.009386	4.625428
10%	0.008806	4.914146

Results of Numerical Simulations





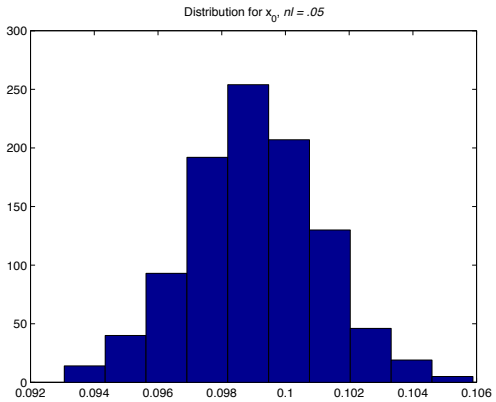


Figure: Bootstrap parameter distributions corresponding to 5% noise with CV.

Results of Numerical Simulations

Remarks:

- The parameter estimates and standard errors are comparable between the asymptotic theory and the bootstrapping theory for this case of constant variance.
- The computational times are two to three orders of magnitude greater for the bootstrapping method as compared to those for the asymptotic theory.
- The asymptotic approach would appear to be the more advantageous method for this simple example.

[1] Takeshi Amemiya, Nonlinear regression models, Chapter 6 in *Handbook of Econometrics, Volume I*, Z. Griliches and M. D. Intriligator, Eds. North Holland, Amsterdam, (1983), 333–389.

[2] H.T. Banks, J.E. Banks, L.K. Dick and J.D. Stark, Estimation of dynamic rate parameters in insect populations undergoing sublethal exposure to pesticides, CRSC-TR05-22, May, 2005; *Bulletin of Mathematical Biology*, **69** (2007), 2139–2180.

[3] H.T. Banks, Amanda Choi, Tori Huffman, John Nardini, Laura Poag and W. Clayton Thompson, Modeling CFSE label decay in flow cytometry data, CRSC-TR12-20, November, 2012; *Applied Mathematical Letters*, **26** (2013), 571–577.

[4] H. T. Banks, M. Davidian, J.R. Samuels Jr., and K.L. Sutton, An inverse problem statistical methodology summary,

CRSC-TR08-01, January, 2008; Chapter 11 in *Statistical Estimation Approaches in Epidemiology*, (edited by Gerardo Chowell, Mac Hyman, Nick Hengartner, Luis M.A Bettencourt and Carlos Castillo-Chavez), Springer, Berlin Heidelberg New York, 2009, 249–302.

[5] H.T. Banks, S. Dediu and S.E. Ernstberger, Sensitivity functions and their uses in inverse problems, *J. Inverse and Ill-posed Problems*, **15** (2007), 683–708.

[6] H. T. Banks, S. Dediu, S.L. Ernstberger, F. Kappel, A new approach to optimal design problems, CRSC-TR08-12, September, 2008.

[7] H. T. Banks, S.L. Ernstberger and S.L. Grove, Standard errors and confidence intervals in inverse problems: sensitivity and associated pitfalls, *J. Inverse and Ill-posed Problems*, **15** (2007), 1–18.

[8] H.T. Banks and B.G. Fitzpatrick, Inverse problems for distributed systems: statistical tests and ANOVA, LCDS/CCS Rep. 88-16, July, 1988, Brown University; *Proc. International Symposium on Math. Approaches to Envir. and Ecol. Problems*, Springer Lecture Note in Biomath., **81** 1989, 262–273.

[9] H.T. Banks and B.G. Fitzpatrick, Statistical methods for model comparison in parameter estimation problems for distributed systems, CAMS Tech. Rep. 89-4, September, 1989, University of Southern California; *J. Math. Biol.*, **28** (1990), 501–527.

[10] H.T. Banks, K. Holm and D. Robbins, Standard error computations for uncertainty quantification in inverse problems: Asymptotic theory vs. bootstrapping, CRSC-TR09-13, June, 2009; Revised August, 2009;

Revised, May, 2010; *Mathematical and Computer Modeling*, **52** (2010), 1610–1625.

[11] H.T. Banks, D.F. Kapraun, W. Clayton Thompson, Cristina Peligero, Jordi Argilaguet and Andreas Meyerhans, A novel statistical analysis and interpretation of flow cytometry data, CRSC-TR12-23, December, 2012; *J. Biological Dynamics*, **7** (2013), 96–132.

[12] H.T. Banks, S. Hu, W. Clayton Thompson, Modeling and inverse problems in the presence of uncertainty. CRC Press, 2014.

[13] H.T. Banks and P. Kareiva, Parameter estimation techniques for transport equations with application to population dispersal and tissue bulk flow models, *J. Math. Biol.*, **17** (1983), 253–272.

[14] H.T. Banks, Z.R. Kenz and W.C. Thompson, An extension of RSS-based model comparison tests for weighted least

squares, CRSC-TR12-18, August, 2012; *Intl. J. Pure and Appl. Math.,* **79** (2012), 155–183.

[15] H.T. Banks, Z.R. Kenz, and W.C. Thompson, A review of selected techniques in inverse problem nonparametric probability distribution estimation, CRSC-TR12-13, May 2012; *J. Inverse and Ill-Posed Problems*, **20** (2012), 429–460.

[16] H.T. Banks and K. Kunisch, *Estimation Techniques for Distributed Parameter Systems*, Birkhäuser, Boston, 1989.

[17] H.T. Banks, L. Potter, and K.L. Rehm, Modeling plant growth using a system of enzyme kinetic equations, to appear.

[18] H.T. Banks, Karyn L. Sutton, W. Clayton Thompson, G. Bocharov, Marie Doumic, Tim Schenkel, Jordi Argilaguet, Sandra Giest, Cristina Peligero, and Andreas Meyerhans, A new model for the estimation of cell proliferation dynamics

using CFSE data, CRSC-TR11-05, Revised July 2011; *J. Immunological Methods,* **373** (2011), 143–160.

[19] H.T. Banks and W. Clayton Thompson, A division-dependent compartmental model with cyton and intracellular label dynamics, CRSC-TR12-12, Revised August 2012; *Intl. J. Pure and Appl. Math*, **77** (2012), 119–147.

[20] H.T. Banks and W. Clayton Thompson, Mathematical models of dividing cell populations: Application to CFSE data, CRSC-TR12-10, April 2012; *J. Math. Modeling of Natural Phenomena*, **7** (2012), 24–52.

[21] H.T. Banks and H.T. Tran, *Mathematical and Experimental Modeling of Physical and Biological Processes*, CRC Press, Boca Raton, FL, 2009.

[22] D. M. Bates, *Nonlinear Regression Analysis and its Applications*, J. Wiley & Sons, Somerset, NJ, 1988.

[23] R.J. Carroll and D. Ruppert, *Transformation and Weighting in Regression,* Chapman & Hall, New York, 1988.

[24] R.J. Carroll, C.F.J. Wu and D. Ruppert, The effect of estimating weights in Weighted Least Squares, *J. Amer. Statistical Assoc.*, **83** (1988), 1045–1054.

[25] G. Casella and R.L. Berger, *Statistical Inference,* Duxbury, California, 2002.

[26] M. Davidian, *Nonlinear Models for Univariate and Multivariate Response*, ST 762 Lecture Notes, Chapters 2, 3, 9 and 11, 2007; http://www4.stat.ncsu.edu/ davidian/courses.html

[27] M. Davidian and D. Giltinan, *Nonlinear Models for Repeated Measurement Data*, Chapman & Hall, London, 1998.

[28] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall/CRC Press, Boca Raton, 1998.

[29] M. Fink. myAD, Retrieved August 2011, from http://www.mathworks.com/matlabcentral/fileexchange/15235-automatic-differentiation-for-matlab.

[30] B. Fitzpatrick, *Statistical Methods in Parameter Identification and Model Selection*, Ph.D. Thesis, Division of Applied Mathematics, Brown University, Providence, RI, 1988.

[31] A.R. Gallant, *Nonlinear Statistical Models*, Wiley, New York, 1987.

[32] F. Graybill, *Theory and Application of the Linear Model*, Duxbury, North Scituate, MA, 1976.

[33] J. Hasenauer, D. Schittler, and F. Allgöwer, A computational model for proliferation dynamics of division-

and label-structured populations, arXive.org, arXiv:1202.4923v1, 22 February 2012.

[34] E.D. Hawkins, M.L. Turner, M.R. Dowling, C. van Gend, and P.D. Hodgkin, A model of immune regulation as a consequence of randomized lymphocyte division and death times, *Proc. Natl. Acad. Sci.*, **104** (2007), 5032–5037.

[35] R.I. Jennrich, Asymptotic properties of non-linear least squares estimators, *Ann. Math. Statist.*, **40** (1969), 633–643.

[36] M. Kot, *Elements of Mathematical Ecology*, Cambridge University Press, Cambridge, 2001.

[37] T. Luzyanina, D. Roose, T. Schenkel, M. Sester, S. Ehl, A. Meyerhans, and G. Bocharov, Numerical modelling of label-structured cell population growth using CFSE distribution data, *Theoretical Biology and Medical Modelling*, **4** (2007), Published Online.

[38] N. Matloff, R. Rose, R. Tai, A comparison of two methods for estimating optimal weights in regression analysis, *J. Statist. Comput. Simul.*, **19** (1984), 265–274.

[39] T. J. Rothenberg, Approximate normality of generalized least squares estimates, *Econometrica*, **52**(4) (1984), 811–825.

[40] W. Rudin, *Principles of Mathematical Analysis*, 2nd edition, McGraw-Hill, New York, 1964.

[41] D. Schittler, J. Hasenauer, and F. Allgöwer, A generalized model for cell proliferation: Integrating division numbers and label dynamics, *Proc. Eighth International Workshop on Computational Systems Biology (WCSB 2011)*, June 2011, Zurich, Switzerland, p. 165–168.

[42] G.A.F. Seber and A.J. Lee, *Linear Regression Analysis*, Wiley, Hoboken, 2003.

[43] G.A.F. Seber and C.J. Wild, *Nonlinear Regression*, J. Wiley & Sons, Hoboken, NJ, 2003.

[44] J. Shao and D. Tu, *The Jackknife and Bootstrap*, Springer-Verlag, New York, 1995.

[45] K. Thomaseth and C. Cobelli, Generalized sensitivity functions in physiological system identification, *Annals of Biomedical Engineering*, **27** (1999), 607–616.

[46] W. Clayton Thompson, *Partial Differential Equation Modeling of Flow Cytometry Data from CFSE-based Proliferation Assays,* Ph.D. Dissertation, Dept. of Mathematics, North Carolina State University, Raleigh, December, 2011.