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"

**Mathematical Model:**
- Equations ("known")
- Parameters ("known")

**Model solution:**
- Numerical approximation
- Evaluated at "known" parameters
Typical "Inverse Problem"

Mathematical Model:
- Equations ("known")
- Parameters (unknown)

Data:
- Measurements of observables

Statistical model:
- Models the observation error

Optimization:
- Minimize the observation error as a function of the parameters

Parameter estimates

Estimates of parameter uncertainty
Linear vs. Nonlinear Regression

Linear Regression
- Linear model: $y = mx + b$

Nonlinear Regression
- Nonlinear model: $y = f(t, \theta)$
Linear vs. Nonlinear Regression

**Linear Regression**

- Linear model: $y = mx + b$
- Data

**Nonlinear Regression**

- Nonlinear model: $y = f(t, \theta)$
- Data

**Linear Model Residuals**

- Residuals

**Nonlinear Model Residuals**

- Residuals

Kevin Flores
Least squares tutorial
The Mathematical Model We consider inverse or parameter estimation problems in the context of a parameterized (with vector parameter \( q \in \mathbb{R}^{\kappa_q} \)) \( n \)-dimensional vector dynamical system or mathematical model

\[
\frac{dx}{dt}(t) = g(t, x(t), q), \quad (1)
\]
\[
x(t_0) = x_0, \quad (2)
\]

with observation process

\[
f(t; \theta) = C x(t; \theta), \quad (3)
\]

where \( \theta = (q^T, \tilde{x}_0^T)^T \in \mathbb{R}^{\kappa_q + \tilde{n}} = \mathbb{R}^{\kappa_\theta}, \tilde{n} \leq n \), and the observation operator \( C \) maps \( \mathbb{R}^n \) to \( \mathbb{R}^m \). In most of the discussions below we assume without loss of generality that some subset \( \tilde{x}_0 \) of the initial values \( x_0 \) 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 $y_j$ corresponding to

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

In general the corresponding observations or data $\{y_j\}$ will not be exactly $f(t_j; \theta)$. 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

\[ Y_j = f(t_j; \theta_0) + h_j \circ \tilde{E}_j, \quad j = 1, \ldots, N, \]  

(5)

where \( f(t_j; \theta) = C x(t_j; \theta), j = 1, \ldots, N, \) and \( C \) 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 \( \theta \in \mathbb{R}^{\kappa q + \tilde{n}} = \mathbb{R}^{\kappa \theta} \). Here the \( m \)-vector function \( h_j \) is defined by

\[
h_j = \begin{cases} 
(1, \ldots, 1)^T & \text{for the vector OLS case} \\
(w_{1,j}, \ldots, w_{m,j})^T & \text{for the vector WLS case} \\
(f_{1}^{\gamma}(t_j; \theta_0), \ldots, f_{m}^{\gamma}(t_j; \theta_0))^T & \text{for the vector GLS case},
\end{cases}
\]

(6)

for \( j = 1, \ldots, N, \) and \( h_j \circ \tilde{E}_j \) denotes the component-wise multiplication of the vectors \( h_j \) and \( \tilde{E}_j \).
The vector $\theta_0$ represents the “truth” parameter that generates the observations $\{Y_j\}_{j=1}^N$.

The terms $h_j \circ \tilde{E}_j$ 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 $f(t_j; \theta_0)$ from the smooth path $f(t, \theta_0)$.

Since these fluctuations are unknown to the modeler, we will assume that realizations $\tilde{\epsilon}_j$ of $\tilde{E}_j$ are generated from a probability distribution which reflects the assumptions regarding these phenomena.
Thus specific data (realizations) corresponding to (5) will be represented by

\[ y_j = f(t_j; \theta_0) + h_j \circ \tilde{\epsilon}_j, \quad j = 1, \ldots, N. \] (7)

We make standard assumptions about the \( \tilde{\mathcal{E}}_j \) 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 \( \gamma = 1 \).
Discuss methodology related to estimates $\hat{\theta}$ for the true value of the parameter $\theta_0$ from a set $\Omega_\theta$ of admissible parameters, and the dependence of this methodology on what is assumed about the choice of $\gamma$ and the covariance matrices of the errors $\tilde{E}_j$.

We discuss a class of inverse problem methodologies that can be used to calculate estimates $\hat{\theta}$ for $\theta_0$: 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 $\mathbb{E}(Y_j)$ embodied in the model and the assumptions made on $\text{Var}(\tilde{E}_j)$. 
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

\[
Y_j = f(t_j; \theta_0) + \tilde{E}_j, \quad j = 1, \ldots, N,
\]

where the variance \(\text{Var}(\tilde{E}_j) = \sigma_0^2\) 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

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

where $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)^T$, then $\theta_{\text{OLS}}$ 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 $\theta$ in

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

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

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

provides a realization for $\theta_{OLS}$. 
Notation:

- For a random vector or estimator $\theta_{\text{OLS}}$, we will always denote a corresponding realization or estimate with an over hat, e.g., $\hat{\theta}_{\text{OLS}}$ is an estimate for $\theta_0$.
- 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

\[ \sigma_0^2 = \frac{1}{N} \mathbb{E} \left( \sum_{j=1}^{N} [Y_j - f(t_j; \theta_0)]^2 \right) \]  

(12)

we see that once we have solved for \( \hat{\theta}_{\text{OLS}} \) in (11), we may readily obtain an estimate \( \hat{\sigma}_{\text{OLS}}^2 \) for \( \sigma_0^2 \). (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 $\theta_{\text{OLS}}$ [43]. As will be explained in more detail below, as $N \to \infty$, we have that

$$\theta_{\text{OLS}} = \theta_{\text{OLS}}^N \sim \mathcal{N}(\theta_0, \Sigma_0^N) \approx \mathcal{N}(\theta_0, \sigma_0^2 [F_{\theta}^N(\theta_0)^T F_{\theta}^N(\theta_0)]^{-1}), \quad (13)$$

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

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

and

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

with

$$\Omega_0 \equiv \lim_{N \to \infty} \frac{1}{N} F_{\theta}^N(\theta_0)^T F_{\theta}^N(\theta_0), \quad (16)$$

where the limit is assumed to exist (see [9, 14, 43]).
\( \theta_{\text{OLS}} \) is approximately distributed as a multivariate normal random variable with mean \( \theta_0 \) and covariance matrix \( \Sigma_0^N \).

The realization (data) \( y = (y_1, \ldots, y_N)^T \) of the random vector \( Y \) is used to estimate \( \hat{\theta}_{\text{OLS}} \) given by (11) and the bias adjusted approximation for \( \sigma_0^2 \):

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

(17)

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

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

(18)

We can obtain the standard errors \( \text{SE}_k(\hat{\theta}_{\text{OLS}}) \) (discussed in more detail below) for the \( k^{th} \) element of \( \hat{\theta}_{\text{OLS}} \) by calculating

\[
\text{SE}_k(\hat{\theta}_{\text{OLS}}) \approx \sqrt{\hat{\Sigma}_{kk}^N}.
\]
Remarks:

- $\hat{\sigma}^2_{OLS}$ represents the estimate for $\sigma_0^2$ 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 $\kappa\theta$ degrees of freedom reflects the fact that $\hat{\theta}$ has been computed to satisfy the $\kappa\theta$ normal equations (10).
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

$$Y_j = f(t_j; \theta_0) + \tilde{E}_j,$$

where $f(t_j; \theta_0) \in \mathbb{R}^m$ and $\tilde{E}_j$, $j = 1, \ldots, N$ are independent and identically distributed with zero mean and covariance matrix given by

$$V_0 = \text{Var}(\tilde{E}_j) = \text{diag}(\sigma^2_{0,1}, \ldots, \sigma^2_{0,m}),$$

for $j = 1, \ldots, N$. Here we have allowed for the possibility that the observation coordinates $Y_j$ 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)).
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

\[ Y_j = f(t_j; \theta_0) + w_j \tilde{E}_j. \] (21)

Here \( \mathbb{E}(Y_j) = f(t_j; \theta_0) \) and \( \text{Var}(Y_j) = \sigma_0^2 w_j^2 \), which derives from the assumptions that \( \mathbb{E}(\tilde{E}_j) = 0 \) and \( \text{Var}(\tilde{E}_j) = \sigma_0^2 \).
The WLS estimator is defined here by

\[
\theta_{WLS} = \arg \min_{\theta \in \Omega} \sum_{j=1}^{N} w_j^{-2} [Y_j - f(t_j; \theta)]^2,
\]

(22)

with corresponding estimate

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

(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 $\theta_{\text{WLS}} = \theta^N_{\text{WLS}}$ has the following asymptotic properties [26, 27]:

$$\theta_{\text{WLS}} \sim \mathcal{N}(\theta_0, \Sigma^N_0),$$

(24)

where

$$\Sigma^N_0 \approx \sigma^2_0 \left( F^T_{\theta}(\theta_0) W F_{\theta}(\theta_0) \right)^{-1},$$

(25)

the sensitivity matrix is given by

$$F_{\theta}(\theta) = F^N_{\theta}(\theta) = \begin{pmatrix}
\frac{\partial f(t_1; \theta)}{\partial \theta_1} & \frac{\partial f(t_1; \theta)}{\partial \theta_2} & \cdots & \frac{\partial f(t_1; \theta)}{\partial \theta_{\kappa \theta}} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial f(t_N; \theta)}{\partial \theta_1} & \frac{\partial f(t_N; \theta)}{\partial \theta_2} & \cdots & \frac{\partial f(t_N; \theta)}{\partial \theta_{\kappa \theta}}
\end{pmatrix},$$

(26)

and the matrix $W$ is defined by $W^{-1} = \text{diag} \left( w^2_1, \ldots, w^2_N \right)$. 
Note that because $\theta_0$ and $\sigma_0^2$ are unknown, the estimates $\hat{\theta} = \hat{\theta}_{\text{WLS}}$ and $\hat{\sigma}^2 = \hat{\sigma}_{\text{WLS}}^2$ will be used in (25) to calculate

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

where we take the approximation

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

We can then approximate the standard errors of $\theta_{\text{WLS}}$ 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 $\theta_{\text{GLS}}$ [26, Chapter 3] and [43, p. 69], the following normal equations are solved for the estimator $\theta_{\text{GLS}}$:

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

(27)

where $Y_j$ satisfies

$$Y_j = f(t_j; \theta_0) + f^\gamma(t_j; \theta_0)\tilde{\epsilon}_j,$$

and

$$\nabla f(t_j; \theta) = \left(\frac{\partial f(t_j; \theta)}{\partial \theta_1}, \ldots, \frac{\partial f(t_j; \theta)}{\partial \theta_{\kappa\theta}}\right)^T.$$
The quantity $\theta_{\text{GLS}}$ is a random vector, hence if $\{y_j\}_{j=1}^N$ is a realization of $\{Y_j\}_{j=1}^N$, then solving

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

for $\theta$ will provide an estimate for $\theta_{\text{GLS}}$. 

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

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

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

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

for \( \theta \), i.e., solving (28). However, we note the GLS relationship (28) does \textit{not} follow from minimizing the weighted least squares with weights chosen as \( w_j = f^{-2\gamma}(t_j; \theta) \) (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 $\theta$ (a standard approach when one assumes that the distribution for the measurement error is completely known—to be discussed later) is the solution to

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

which is equivalent to the corresponding GLS estimating equations (28) with $\gamma = 1$. (See Chapter 3 of [12])
The GLS estimator $\theta_{\text{GLS}} = \theta_{\text{GLS}}^N$ has the following asymptotic properties [27, 43]:

$$\theta_{\text{GLS}} \sim \mathcal{N}(\theta_0, \Sigma_0^N),$$

(30)

where

$$\Sigma_0^N \approx \sigma_0^2 \left( F^T_{\theta}(\theta_0) W(\theta_0) F_{\theta}(\theta_0) \right)^{-1},$$

(31)

the sensitivity matrix is given by (26) and the matrix $W(\theta)$ is defined by $W^{-1}(\theta) = \text{diag} \left( f^{2\gamma}(t_1; \theta), \ldots, f^{2\gamma}(t_N; \theta) \right)$. 
Note that because $\theta_0$ and $\sigma_0^2$ are unknown, the estimates $\hat{\theta} = \hat{\theta}_{\text{GGLS}}$ and $\hat{\sigma}^2 = \hat{\sigma}_{\text{GGLS}}^2$ will again be used in (31) to calculate

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

where we take the approximation

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

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

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

Since $Y_j$ is a random variable, the corresponding estimator $\theta_{\text{OLS}}^N$ (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 $\theta$ using a specific estimator and sample size $N$ in actual data collection.
Computation of sensitivities

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

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

where $h_k$ is a $\kappa_\theta$-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 \( h_k \) 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 \( f(t_j; \theta) \) may be sufficiently simple to allow one to derive analytical expressions for \( F_\theta \). Alternatively, if the \( f(t_j; \theta) \) correspond to longitudinal observations \( f(t_j; \theta) = Cx(t_j; \theta) \) of solutions to a parameterized \( n \)-vector differential equation system

\[
\dot{x} = g(t, x(t), q)
\]
as in (1)-(2), then one can use the \( n \times \kappa_\theta \) matrix sensitivity equations (see [5, 7] and the references therein)

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

to obtain

\[
\frac{\partial f(t_j; \theta)}{\partial \theta_k} = C \frac{\partial x(t_j, \theta)}{\partial \theta_k}.
\]
In order to quantify the variation in the state variable $x(t)$ with respect to changes in the parameters $q$ and the initial conditions $x_0$, we are naturally led to consider the individual (traditional) sensitivity functions (TSF) defined by the derivatives

$$s_{q_k}(t) = \frac{\partial x}{\partial q_k}(t) = \frac{\partial x}{\partial q_k}(t, \theta), \quad k = 1, \ldots, \kappa_q, \quad (34)$$

and

$$r_{x_{0_l}}(t) = \frac{\partial x}{\partial x_{0_l}}(t) = \frac{\partial x}{\partial x_{0_l}}(t, \theta), \quad l = 1, \ldots, n, \quad (35)$$

where $x_{0_l}$ is the $l^{th}$ component of the initial condition $x_0$. If the function $g$ is sufficiently regular, the solution $x$ is differentiable with respect to $q_k$ and $x_{0_l}$, and therefore the sensitivity functions $s_{q_k}$ and $r_{x_{0_l}}$ are well–defined.
Remarks on local sensitivities:

- Because they are defined by partial derivatives which have a *local* character, the sensitivity functions are also local in nature.

- Sensitivity and insensitivity ($s_{q_k} = \partial x / \partial q_k$ not close to zero and very close to zero, respectively) depend on the time interval, the state values $x$, and the values of $\theta$ for which they are considered.

- For example, in a certain time subinterval we might find $s_{q_k}$ small so that the state variable $x$ is *insensitive* to the parameter $q_k$ on that particular interval.

- The same function $s_{q_k}$ can take large values on a different subinterval, indicating to us that the state variable $x$ is *very sensitive* to the parameter $q_k$ on the latter interval.
From the sensitivity analysis theory for dynamical systems, one finds that $s = (s_{q_1}, \ldots, s_{q_{\kappa q}})$ is an $n \times \kappa_q$ vector function that satisfies the ODE system

$$\dot{s}(t) = \frac{\partial g}{\partial x}(t, x(t; \theta), q)s(t) + \frac{\partial g}{\partial q}(t, x(t; \theta), q),$$

which is obtained by differentiating (1)-(2) with respect to $q$. Here the dependence of $s$ on $(t, x(t; \theta))$ as well as $q$ is readily apparent.
In a similar manner, the sensitivity functions with respect to the components of the initial condition $x_0$ define an $n \times n$ vector function $r = (r_{x_01}, \ldots, r_{x_0n})$, which satisfies

$$
\dot{r}(t) = \frac{\partial g}{\partial x}(t, x(t; \theta), q)r(t), \quad \text{(37)}
$$

$$
r(t_0) = I_n.
$$

This is obtained by differentiating (1)-(2) with respect to the initial conditions $x_0$. The equations (36) and (37) are used in conjunction with (i.e., usually solved simultaneously with) equations (1)-(2) to numerically compute the sensitivities $s$ and $r$ for general cases when the function $g$ is sufficiently complicated to prohibit a closed form solution by direct integration. These can be succinctly written as a system for

$$
\frac{\partial x}{\partial \theta} = \left( \frac{\partial x}{\partial q}, \frac{\partial x}{\partial x_0} \right)
$$
given by (33).
As we have already noted, since $\theta_0$ and $\sigma_0$ are unknown, we will use their estimates to make the approximation

$$\Sigma_0^N \approx \sigma_0^2 [F^N_\theta(\theta_0)^T F^N_\theta(\theta_0)]^{-1} \approx \hat{\Sigma}_N^N(\hat{\theta}_{\text{OLS}}^N) = \hat{\sigma}^2 [F^N_\theta(\hat{\theta}_{\text{OLS}}^N)^T F^N_\theta(\hat{\theta}_{\text{OLS}}^N)]^{-1},$$  

(38)

where the approximation $\hat{\sigma}^2$ of $\sigma_0^2$, as discussed earlier, is given by

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

Standard errors to be used in the confidence interval calculations are given by $SE_k(\hat{\theta}_N^N) = \sqrt{\hat{\Sigma}_{kk}^N(\hat{\theta}_N^N)}$, $k = 1, 2, \ldots, \kappa_{\theta}$ (see [25]).
To compute the confidence intervals (at the 100(1 − α)% level) for the estimated parameters in our example, we define the confidence intervals associated with the estimated parameters so that

\[
\text{Prob}\{\theta_k^N - t_{1-\alpha/2}SE_k(\hat{\theta}^N) < \theta_0k < \theta_k^N + t_{1-\alpha/2}SE_k(\hat{\theta}^N)\} = 1 - \alpha,
\]

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

\[
[\hat{\theta}_k^N - t_{1-\alpha/2}SE_k(\hat{\theta}^N), \hat{\theta}_k^N + t_{1-\alpha/2}SE_k(\hat{\theta}^N)].
\]

Given a small \(\alpha\) 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 \( \theta_0 \).

- Thus, a confidence interval provides further information on the extent of uncertainty involved in estimating \( \theta_0 \) using the given estimator and sample size \( N \).
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 \( Y_j = f(t_j; \theta_0) + \tilde{E}_j \) whereas the GLS should be used for problems in which we have nonconstant variance observations \( Y_j = f(t_j; \theta_0) + \gamma(t_j; \theta_0)\tilde{E}_j \).

- 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 ($\gamma = 0$)

  $$Y_j = f(t_j; \theta_0) + \tilde{E}_j, \quad \text{Var}(Y_j) = \sigma_0^2,$$

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

  $$Y_j = f(t_j; \theta_0)(1 + \tilde{E}_j), \quad \text{Var}(Y_j) = \sigma_0^2 f^2(t_j; \theta_0).$$

- 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 $\{\tilde{E}_j\}_{j=1}^N$, and then calculate an estimate $\hat{\theta}$ of $\theta_0$ 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

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

  and hence the residuals $r_j$ are approximations to realizations of the errors $\tilde{E}_j$ (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 $\tilde{E}_j$ 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.
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, \( R_j = Y_j - f(t_j; \theta_0) = f(t_j; \theta_0) \tilde{E}_j \) 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.
If a data set has nonconstant variance generated data, then

\[ Y_j = f(t_j; \theta_0) + f(t_j; \theta_0) \tilde{\varepsilon}_j \quad \text{or} \quad \tilde{\varepsilon}_j = \frac{Y_j - f(t_j; \theta_0)}{f(t_j; \theta_0)}. \]

Test 1: If the distributions of \( \tilde{\varepsilon}_j \) 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.
What if the data are incorrectly assumed to have non-constant variance error when in fact they have constant variance error?

Since $Y_j - f(t_j; \theta_0) = \tilde{\varepsilon}_j$ 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).
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., $f_{\text{repeat}}$) 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).
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).
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

\[ x(t) = \frac{K x_0 e^{rt}}{K + x_0 (e^{rt} - 1)}. \tag{43} \]
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 $x_0 = 0.1$. 
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 $\tilde{E}_j$) and obtained estimates $\hat{\theta}$ of $\theta_0$ 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 $\theta_0$, 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.

<table>
<thead>
<tr>
<th>$\theta_{\text{init}}$</th>
<th>$\theta_0$</th>
<th>$\hat{\theta}^{\text{CV}}_{\text{OLS}}$</th>
<th>SE($\hat{\theta}^{\text{CV}}_{\text{OLS}}$)</th>
<th>$\hat{\theta}^{\text{TCV}}_{\text{OLS}}$</th>
<th>SE($\hat{\theta}^{\text{TCV}}_{\text{OLS}}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>17.5</td>
<td>1.75e+001</td>
<td>1.58e-003</td>
<td>1.74e+001</td>
<td>6.42e-003</td>
</tr>
<tr>
<td>.8</td>
<td>.7</td>
<td>7.00e-001</td>
<td>4.28e-004</td>
<td>7.00e-001</td>
<td>6.58e-004</td>
</tr>
</tbody>
</table>
Estimation using the GLS procedure with CV data.

<table>
<thead>
<tr>
<th>$\theta_{\text{init}}$</th>
<th>$\theta_0$</th>
<th>$\hat{\theta}_{\text{GLS}}^{\text{CV}}$</th>
<th>SE($\hat{\theta}_{\text{GLS}}^{\text{CV}}$)</th>
<th>$\hat{\theta}_{\text{GLS}}^{\text{TCV}}$</th>
<th>SE($\hat{\theta}_{\text{GLS}}^{\text{TCV}}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>17.5</td>
<td>1.75e+001</td>
<td>1.38e-004</td>
<td>1.75e+001</td>
<td>9.12e-005</td>
</tr>
<tr>
<td>.8</td>
<td>.7</td>
<td>7.00e-001</td>
<td>7.81e-005</td>
<td>7.01e-001</td>
<td>1.60e-005</td>
</tr>
<tr>
<td>1.2</td>
<td>.1</td>
<td>9.99e-002</td>
<td>6.61e-005</td>
<td>9.97e-002</td>
<td>1.21e-005</td>
</tr>
</tbody>
</table>
An Example Using Residual Plots: Logistic Growth

Estimation using the OLS procedure with NCV data.

<table>
<thead>
<tr>
<th>$\theta_{init}$</th>
<th>$\theta_0$</th>
<th>$\hat{\theta}_{NCV}^{OLS}$</th>
<th>SE($\hat{\theta}_{NCV}^{OLS}$)</th>
<th>$\hat{\theta}_{OLSNCV}$</th>
<th>SE($\hat{\theta}_{OLSNCV}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>17.5</td>
<td>1.75e+001</td>
<td>2.27e-002</td>
<td>1.74e+001</td>
<td>7.16e-002</td>
</tr>
<tr>
<td>.8</td>
<td>.7</td>
<td>7.02e-001</td>
<td>6.18e-003</td>
<td>7.09e-001</td>
<td>7.60e-003</td>
</tr>
<tr>
<td>1.2</td>
<td>.1</td>
<td>9.95e-002</td>
<td>4.51e-003</td>
<td>9.49e-002</td>
<td>4.83e-003</td>
</tr>
</tbody>
</table>
An Example Using Residual Plots: Logistic Growth

Estimation using the GLS procedure with NCV data.

<table>
<thead>
<tr>
<th>( \theta_{\text{init}} )</th>
<th>( \theta_0 )</th>
<th>( \hat{\theta}_{\text{NCV}}^{\text{GLS}} )</th>
<th>SE(( \hat{\theta}_{\text{NCV}}^{\text{GLS}} ))</th>
<th>( \hat{\theta}_{\text{GLS}}^{\text{TNCV}} )</th>
<th>SE(( \hat{\theta}_{\text{GLS}}^{\text{TNCV}} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>17.5</td>
<td>1.75e+001</td>
<td>9.44e-005</td>
<td>1.74e+001</td>
<td>3.13e-004</td>
</tr>
<tr>
<td>.8</td>
<td>.7</td>
<td>7.02e-001</td>
<td>5.36e-005</td>
<td>7.09e-001</td>
<td>5.72e-005</td>
</tr>
<tr>
<td>1.2</td>
<td>.1</td>
<td>9.93e-002</td>
<td>4.49e-005</td>
<td>9.49e-002</td>
<td>4.12e-005</td>
</tr>
</tbody>
</table>
Figure: Residual vs. time plots in tests for independence: Original and truncated logistic curve for $\hat{\theta}^{\text{CV}}_{\text{OLS}}$. 
Figure: Residual vs. model plots in tests of form of variance: Original and truncated logistic curve for $\hat{\theta}_{CV}^{\text{OLS}}$. 

Kevin Flores
Least squares tutorial
Figure: Residual vs. time plots in tests for independence: Original and truncated logistic curve for $\hat{\theta}_{\text{NCV}}^\text{OLS}$.
Figure: Residual vs. model plots in tests of form of variance: Original and truncated logistic curve for $\hat{\theta}_{\text{NCV}}^\text{OLS}$.
Figure: Residual vs. time plots in tests for independence: Original and truncated logistic curve for $\hat{\theta}_{\text{CV}}$ and $\hat{\theta}_{\text{GLS}}$. 
Figure: Modified residual vs. model plots in tests of form of variance: Original and truncated logistic curve for $\hat{\theta}_{CV}^{\text{GLS}}$. 
Figure: Modified residual vs. time plots in tests for independence: Original and truncated logistic curve for $\hat{\theta}^{NCV}_{GLS}$. 
Figure: Modified residual vs. model plots in tests of form of variance: Original and truncated logistic curve for $\hat{\theta}_{GLS}^{NCV}$.
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 $\theta_0$. 
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

\[
Y_j = f(t_j; \theta_0) + \tilde{\epsilon}_j, \quad j = 1, \ldots, N,
\]  

(44)

where the \(\tilde{\epsilon}_j\) are independent and identically distributed (i.i.d.) with mean zero (\(\mathbb{E}(\epsilon_j) = 0\)) and constant variance \(\sigma_0^2\), and \(\theta_0\) 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

\[
y_j = f(t_j; \theta_0) + \epsilon_j.
\]

The following algorithm [23, 24, 26, p. 285–287] can be used to compute the bootstrapping estimate \(\hat{\theta}_{\text{BOOT}}\) of \(\theta_0\) and its empirical distribution.
First estimate $\hat{\theta}^0$ from the entire sample $\{y_j\}_{j=1}^N$ using OLS.

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 $m = 0$.

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 $\{\bar{r}^m_1, \ldots, \bar{r}^m_N\}$.

Create bootstrap sample points

$$y^m_j = f(t_j; \hat{\theta}^0) + r^m_j,$$

where $j = 1, \ldots, N$.

Obtain a new estimate $\hat{\theta}^{m+1}$ from the bootstrapping sample $\{y^m_j\}$ using OLS.

Set $m = m + 1$ and repeat steps 3–5 until $m \geq M$ (e.g., typically $M = 1000$ as in our calculations below).
We then calculate the mean, standard error, and confidence intervals using the formulae

\[ \hat{\theta}_{BOOT} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}^m, \]

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

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

where \( \theta_{BOOT} \) denotes the bootstrapping estimator.
In the above procedures, the \{\bar{r}_1, \ldots, \bar{r}_N\} are realizations of \textit{i.i.d.} random variables \bar{R}_j with the empirical distribution function \( F_N \). It can be shown that

\[
\mathbb{E}(\bar{R}_j | F_N) = N^{-1} \sum_{j=1}^{N} \bar{r}_j = 0, \quad \text{Var}(\bar{R}_j | F_N) = N^{-1} \sum_{j=1}^{N} \bar{r}_j^2 = \hat{\sigma}^2.
\]
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 \textit{ode45} where
  \[
  \theta = (K, r, x_0)^T.
  \]

- A noise vector of length \( N \) with noise level \( \sigma_0 \), was taken from a random number generator for \( \mathcal{N}(0, \sigma_0^2) \).

- The constant variance data sets were obtained from the equation
  \[
  y_j = f(t_j; \theta_0) + \tilde{\epsilon}_j.
  \]

- 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 $SE_k(\hat{\theta})$ and corresponding confidence intervals $[\hat{\theta}_k - 1.96SE_k(\hat{\theta}), \hat{\theta}_k + 1.96SE_k(\hat{\theta})]$ 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, \( \sigma_0 = 0.01 \).

<table>
<thead>
<tr>
<th>( \hat{\theta} )</th>
<th>( \hat{\theta} )</th>
<th>SE(( \hat{\theta} ))</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{K}_{asy} )</td>
<td>17.498576</td>
<td>0.002021</td>
<td>(17.494615, 17.502537)</td>
</tr>
<tr>
<td>( \hat{r}_{asy} )</td>
<td>0.700186</td>
<td>0.000553</td>
<td>(0.699103, 0.701270)</td>
</tr>
<tr>
<td>( \hat{x}<em>0 )</em>{asy}</td>
<td>0.100044</td>
<td>0.000407</td>
<td>(0.099247, 0.100841)</td>
</tr>
<tr>
<td>( \hat{K}_{boot} )</td>
<td>17.498464</td>
<td>0.001973</td>
<td>(17.494597, 17.502331)</td>
</tr>
<tr>
<td>( \hat{r}_{boot} )</td>
<td>0.700193</td>
<td>0.000548</td>
<td>(0.699118, 0.701268)</td>
</tr>
<tr>
<td>( \hat{x}<em>0 )</em>{boot}</td>
<td>0.100034</td>
<td>0.000399</td>
<td>(0.099252, 0.100815)</td>
</tr>
</tbody>
</table>
## Results of Numerical Simulations

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

<table>
<thead>
<tr>
<th>( \theta )</th>
<th>( \hat{\theta} )</th>
<th>( \text{SE}(\hat{\theta}) )</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{K}_{asy} )</td>
<td>17.486571</td>
<td>0.010269</td>
<td>(17.466444, 17.506699)</td>
</tr>
<tr>
<td>( \hat{r}_{asy} )</td>
<td>0.702352</td>
<td>0.002825</td>
<td>(0.696815, 0.707889)</td>
</tr>
<tr>
<td>(( \hat{x}<em>0 ))</em>{asy}</td>
<td>0.098757</td>
<td>0.002050</td>
<td>(0.0947386, 0.102775)</td>
</tr>
<tr>
<td>( \hat{K}_{boot} )</td>
<td>17.489658</td>
<td>0.010247</td>
<td>(17.469574, 17.509742)</td>
</tr>
<tr>
<td>( \hat{r}_{boot} )</td>
<td>0.702098</td>
<td>0.002938</td>
<td>(0.696339, 0.707857)</td>
</tr>
<tr>
<td>(( \hat{x}<em>0 ))</em>{boot}</td>
<td>0.0990520</td>
<td>0.002152</td>
<td>(0.094834, 0.103270)</td>
</tr>
</tbody>
</table>
Asymptotic and bootstrap OLS estimates for NCV data, $\sigma_0 = 0.1$.

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>$\hat{\theta}$</th>
<th>SE($\hat{\theta}$)</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{K}_{asy}$</td>
<td>17.081926</td>
<td>0.262907</td>
<td>(16.566629, 17.597223)</td>
</tr>
<tr>
<td>$\hat{r}_{asy}$</td>
<td>0.727602</td>
<td>0.078513</td>
<td>(0.573717, 0.881487)</td>
</tr>
<tr>
<td>$(\hat{x}<em>0)</em>{asy}$</td>
<td>0.082935</td>
<td>0.047591</td>
<td>(-0.010343, 0.176213)</td>
</tr>
<tr>
<td>$\hat{K}_{boot}$</td>
<td>17.095648</td>
<td>0.250940</td>
<td>(16.603807, 17.587490)</td>
</tr>
<tr>
<td>$\hat{r}_{boot}$</td>
<td>0.733657</td>
<td>0.081852</td>
<td>(0.573228, 0.894087)</td>
</tr>
<tr>
<td>$(\hat{x}<em>0)</em>{boot}$</td>
<td>0.094020</td>
<td>0.054849</td>
<td>(-0.013484, 0.201524)</td>
</tr>
</tbody>
</table>
## Results of Numerical Simulations

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

<table>
<thead>
<tr>
<th>Noise Level</th>
<th>Asymptotic Theory</th>
<th>Bootstrapping</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>0.017320</td>
<td>4.285640</td>
</tr>
<tr>
<td>5%</td>
<td>0.009386</td>
<td>4.625428</td>
</tr>
<tr>
<td>10%</td>
<td>0.008806</td>
<td>4.914146</td>
</tr>
</tbody>
</table>
Figure: Bootstrap parameter distributions corresponding to 5% noise with CV.
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.


[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.


