An Inverse Problem Statistical Methodology Summary

H.T. Banks, M. Davidian, J.R. Samuels, Jr. and Karyn L.Sutton
Center for Research in Scientific Computation
and
Center for Quantitative Sciences in Biomedicine
North Carolina State University
Raleigh, NC 27695-8205
and
Department of Mathematics & Statistics
Arizona State University
Tempe, AZ, 85287-1804

January 12, 2008

Abstract

We discuss statistical and computational aspects of inverse or parameter estimation problems based on Ordinary Least Squares and Generalized Least Squares with appropriate corresponding data noise assumptions of constant variance and nonconstant variance (relative error), respectively. Among the topics included here are mathematical model, statistical model and data assumptions, and some techniques (residual plots, sensitivity analysis, model comparison tests) for verifying these. The ideas are illustrated throughout with the popular logistic growth model of Verhulst and Pearl as well as with a recently developed population level model of pneumococcal disease spread.

Keywords: Inference, least squares inverse problems, parameter estimation, sensitivity and generalized sensitivity functions.

1 Introduction

In this Chapter we discuss mathematical and statistical aspects of **inverse** or **parameter estimation** problems. While we briefly discuss *maximum likelihood estimators* (MLE), our focus here will be on *ordinary least squares* (OLS) and *generalized least squares* (GLS) estimation formulations and issues related to use of these techniques in practice. While we choose a

general nonlinear ordinary differential equation *mathematical model* to discuss concepts and ideas, the discussions are also applicable to partial differential equation models and other deterministic dynamical systems. As we shall explain, the choice of an appropriate *statistical model* is of critical importance, and we discuss at length the difference between constant variance and nonconstant variance noise in the observation process, the consequences for incorrect choices in this regard, and computational techniques for investigating whether a good decision has been made. In particular, we illustrate use of residual plots to suggest whether or not a correct statistical model has been specified in an inverse problem formulation. We illustrate these and other techniques with examples including the well known Verhulst-Pearl logistic population model and a specific epidemiological model (a pneumococcal disease dynamics model). We discuss the use of sensitivity equations coupled with the asymptotic theory for sampling distributions and the computation of associated covariances, standard errors and confidence intervals for the estimators of model parameters. We also discuss *sensitivity functions* (*traditional* and *generalized*) and their emerging use in design of experiments for data specific to models and mechanism investigation. Traditional sensitivity involves sensitivity of outputs to parameters while the recent concept of generalized sensitivity in inverse problems pertains to sensitivity of parameters (to be estimated) to data or observations. That is, generalized sensitivity quantifies the relevance of data measurements for identification of parameters in a typical parameter estimation problem. In a final section we present and illustrate some methods for model comparison.

2 Parameter Estimation: MLE, OLS, and GLS

2.1 The Underlying Mathematical and Statistical Models

We consider inverse or parameter estimation problems in the context of a parameterized (with vector parameter $\vec{\theta}$) dynamical system or **mathematical model**

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

with **observation process**

$$\vec{y}(t) = \mathcal{C}\vec{x}(t; \vec{\theta}). \tag{2}$$

Following usual convention (which agrees with the data usually available from experiments), we assume a discrete form of the observations in which one has n longitudinal observations corresponding to

$$\vec{y}(t_j) = \mathcal{C}\vec{x}(t_j; \vec{\theta}), \quad j = 1, \ldots, n. \tag{3}$$

In general the corresponding observations or data $\{\vec{y}_j\}$ will not be exactly $\vec{y}(t_j)$. Because of 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.

2.2 Description of Statistical Model

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

$$\vec{Y}_j = \vec{f}(t_j, \vec{\theta}_0) + \vec{\epsilon}_j, \quad j = 1, \ldots, n, \tag{4}$$

where $\vec{f}(t_j, \vec{\theta}) = \mathcal{C}\vec{x}(t_j; \vec{\theta})$, $j = 1, \ldots, n$, corresponds to the solution of the mathematical model (1) at the j^{th} covariate for a particular vector of parameters $\vec{\theta} \in R^p, \vec{x} \in R^N, \vec{f} \in R^m$, and $\mathcal{C}$ is an $m \times N$ matrix. The term $\vec{\theta}_0$ represents the "truth" or the parameters that generate the observations $\{\vec{Y}_j\}_{j=1}^n$. (The existence of a truth parameter $\vec{\theta}_0$ is standard in statistical formulations and this along with the assumption that the means $E[\vec{\epsilon}_j]$ are zero yields implicitly that the (1) is a correct description of the process being modeled.) The terms $\vec{\epsilon}_j$ are random variables which can represent measurement error, "system fluctuations" or other phenomena that cause observations to not fall exactly on the points $\vec{f}(t_j, \vec{\theta})$ from the smooth path $\vec{f}(t, \vec{\theta})$. Since these fluctuations are unknown to the modeler, we will assume $\vec{\epsilon}_j$ is generated from a probability distribution (with mean zero throughout our discussions) that reflects the assumptions regarding these phenomena. For instance, in a statistical model for pharmacokinetics of drug in human blood samples, a natural distribution for $\vec{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ might be a multivariate normal distribution. In other applications the distribution for $\vec{\epsilon}$ might be much more complicated [22].

The purpose of our presentation here is to discuss methodology related to the estimation of the true value of the parameters $\vec{\theta}_0$ from a set Θ of admissible parameters, and its dependence on what is assumed about the variance $\text{var}(\vec{\epsilon}_j)$ of the error $\vec{\epsilon}_j$. We discuss two inverse problem methodologies that can be used to calculate estimates $\hat{\theta}$ for $\vec{\theta}_0$: the ordinary least-squares (OLS) and generalized least-squares (GLS) formulations as well as the popular maximum likelihood estimate (MLE) formulation in the case one assumes the distributions of the error process $\{\vec{\epsilon}_j\}$ are known.

2.3 Known error processes: Normally distributed error

In the introduction of the statistical model we initially made no mention of the probability distribution that generates the error $\vec{\epsilon}_j$. In many situations one readily assumes that the errors $\vec{\epsilon}_j = 1, \ldots, n$, are independent and identically distributed (we make the *standing assumptions of independence across* j throughout our discussions in this Chapter). We discuss a case where one is able to make further assumptions on the error, namely that the distribution is known. In this case, maximum likelihood techniques may be used. We discuss first one such case for a scalar observation system, i.e., $m = 1$. If, in addition, there is sufficient evidence to suspect the error is generated by a normal distribution then we may be willing to assume $\epsilon_j \sim \mathcal{N}(0, \sigma_0^2)$, and hence $Y_j \sim \mathcal{N}(f(t_j, \vec{\theta}_0), \sigma_0^2)$. We can then obtain an expression for determining $\vec{\theta}_0$ and σ_0 by seeking the maximum over $(\vec{\theta}, \sigma^2) \in \Theta \times (0, \infty)$ of

the likelihood function for $\epsilon_j = Y_j - f(t_j, \vec{\theta})$ which is defined by

$$L(\vec{\theta}, \sigma^2 | \vec{Y}) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}[Y_j - f(t_j, \vec{\theta})]^2\}. \tag{5}$$

The resulting solutions θ_{MLE} and σ^2_{MLE} are the maximum likelihood **estimators** (MLEs) for $\vec{\theta}_0$ and σ_0^2, respectively. We point out that these solutions $\theta_{\text{MLE}} = \theta^n_{\text{MLE}}(\vec{Y})$ and $\sigma^2_{\text{MLE}} = \sigma^{2\,n}_{\text{MLE}}(\vec{Y})$ are *random variables* by virtue of the fact that $\vec{Y}$ is a random variable. The corresponding maximum likelihood **estimates** are obtained by maximizing (5) with $\vec{Y} = (Y_1, \ldots, Y_n)^T$ replaced by a given realization $\vec{y} = (y_1, \ldots, y_n)^T$ and will be denoted by $\hat{\theta}_{\text{MLE}} = \hat{\theta}^n_{\text{MLE}}$ and $\hat{\sigma}_{\text{MLE}} = \hat{\sigma}^n_{\text{MLE}}$ respectively. In our discussions here and below, almost every quantity of interest is dependent on n, the *size of the set of observations* or the *sampling size*. On occasion we will express this dependence explicitly by use of superscripts or subscripts, especially when we wish to remind the reader of this dependence. However, for notational convenience we will often suppress the notation of explicit dependence on n.

Maximizing (5) is equivalent to maximizing the log likelihood

$$\log L(\vec{\theta}, \sigma^2 | \vec{Y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{j=1}^{n}[Y_j - f(t_j, \vec{\theta})]^2. \tag{6}$$

We determine the maximum of (6) by differentiating with respect to $\vec{\theta}$ (with σ^2 fixed) and with respect to σ^2 (with $\vec{\theta}$ fixed), setting the resulting equations equal to zero and solving for $\vec{\theta}$ and σ^2. With σ^2 fixed we solve $\frac{\partial}{\partial\vec{\theta}} \log L(\vec{\theta}, \sigma^2 | \vec{Y}) = 0$ which is equivalent to

$$\sum_{j=1}^{n}[Y_j - f(t_j, \vec{\theta})]\nabla f(t_j, \vec{\theta}) = 0, \tag{7}$$

where as usual $\nabla f = \frac{\partial}{\partial\vec{\theta}} f = f_{\vec{\theta}}$. We see that solving (7) is the same as the least squares optimization

$$\theta_{\text{MLE}}(\vec{Y}) = \arg\min_{\vec{\theta}\in\Theta} J(\vec{Y}, \vec{\theta}) = \arg\min_{\vec{\theta}\in\Theta} \sum_{j=1}^{n}[Y_j - f(t_j, \vec{\theta})]^2. \tag{8}$$

We next fix $\vec{\theta}$ to be θ_{MLE} and solve $\frac{\partial}{\partial\sigma^2} \log L(\theta_{\text{MLE}}, \sigma^2 | \vec{Y}) = 0$, which yields

$$\sigma^2_{\text{MLE}}(\vec{Y}) = \frac{1}{n} J(\vec{Y}, \theta_{\text{MLE}}). \tag{9}$$

Note that we can solve for θ_{MLE} and σ^2_{MLE} separately – a desirable feature, but one that does not arise in more complicated formulations discussed below. The 2^{nd} derivative test (which is omitted here) verifies that the expressions above for θ_{MLE} and σ^2_{MLE} do indeed maximize (6).

If, however, we have a vector of observations for the j^{th} covariate t_j then the statistical model is reformulated as

$$\vec{Y_j} = \vec{f}(t_j, \vec{\theta_0}) + \vec{\epsilon_j} \tag{10}$$

where $\vec{f} \in R^m$ and

$$V_0 = \text{var}(\vec{\epsilon_j}) = \text{diag}(\sigma_{0,1}^2, \ldots, \sigma_{0,m}^2) \tag{11}$$

for $j = 1, \ldots, n$. In this setting we have allowed for the possibility that the observation coordinates Y_j^i may have different *constant* variances $\sigma_{0,i}^2$, i.e., $\sigma_{0,i}^2$ does not necessarily have to equal $\sigma_{0,k}^2$. If (again) there is sufficient evidence to claim the errors are independent and identically distributed and generated by a normal distribution then $\vec{\epsilon_j} \sim \mathcal{N}_m(0, V_0)$. We thus can obtain the maximum likelihood estimators $\theta_{\text{MLE}}(\{\vec{Y_j}\})$ and $V_{\text{MLE}}(\{\vec{Y_j}\})$ for θ_0 and V_0 by determining the maximum of the log of the likelihood function for $\vec{\epsilon_j} = \vec{Y_j} - \vec{f}(t_j, \vec{\theta})$ defined by

$$\begin{aligned}
\log L(\vec{\theta}, V | \{Y_j^1, \ldots, Y_j^m\}) &= -\frac{n}{2} \sum_{i=1}^{m} \log \sigma_{0,i}^2 - \frac{1}{2} \sum_{i=1}^{m} \frac{1}{\sigma_{0,i}^2} \sum_{j=1}^{n} [Y_j^i - f^i(t_j, \vec{\theta})]^2 \\
&= -\frac{n}{2} \sum_{i=1}^{m} \log \sigma_{0,i}^2 - \sum_{j=1}^{n} [\vec{Y_j} - \vec{f}(t_j, \vec{\theta})]^T V^{-1} [\vec{Y_j} - \vec{f}(t_j, \vec{\theta})].
\end{aligned}$$

Using arguments similar to those given for the scalar case, we determine the maximum likelihood estimators for $\vec{\theta_0}$ and V_0 to be

$$\theta_{\text{MLE}} = \arg\min_{\vec{\theta} \in \Theta} \sum_{j=1}^{n} [\vec{Y_j} - \vec{f}(t_j, \vec{\theta})]^T V_{\text{MLE}}^{-1} [\vec{Y_j} - \vec{f}(t_j, \vec{\theta})] \tag{12}$$

$$V_{\text{MLE}} = \text{diag}\left(\frac{1}{n} \sum_{j=1}^{n} [\vec{Y_j} - \vec{f}(t_j, \theta_{\text{MLE}})][\vec{Y_j} - \vec{f}(t_j, \theta_{\text{MLE}})]^T \right). \tag{13}$$

Unfortunately, this is a coupled system, which requires some care when solving numerically. We will discuss this issue further in Sections 2.4.2 and 2.4.5 below.

2.4 Unspecified Error Distributions and Asymptotic Theory

In Section 2.3 we examined the estimates of $\vec{\theta_0}$ and V_0 under the assumption *that the error is normally distributed, independent and constant longitudinally.* But what if it is suspected that the error is not normally distributed, or the error distribution is unknown to the modeler beyond the assumptions on $E[\vec{Y_j}]$ embodied in the model and the assumptions made on $\text{var}(\vec{\epsilon_j})$ (as in most applications)? How should we proceed in estimating $\vec{\theta_0}$ and σ_0 (or V_0) in these circumstances? In this section we will review two estimation procedures for such situations: ordinary least squares (OLS) and generalized least squares (GLS).

2.4.1 Ordinary Least Squares (OLS)

The statistical model in the scalar case takes the form

$$Y_j = f(t_j, \vec{\theta}_0) + \epsilon_j \tag{14}$$

where the variance $\text{var}(\epsilon_j) = \sigma_0^2$ is assumed constant in longitudinal data (note that the error's distribution is not specified). We also note that the assumption that the observation errors are uncorrelated across j (i.e., time) 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

$$\theta_{\text{OLS}}(\vec{Y}) = \theta^n_{\text{OLS}}(\vec{Y}) = \arg\min_{\vec{\theta}\in\Theta} \sum_{j=1}^{n} [Y_j - f(t_j, \vec{\theta})]^2 \tag{15}$$

then θ_{OLS} can be viewed as minimizing the distance between the data and model where all observations are treated as of equal importance. We note that minimizing in (15) corresponds to solving for $\vec{\theta}$ in

$$\sum_{j=1}^{n} [Y_j - f(t_j, \vec{\theta})] \nabla f(t_j, \vec{\theta}) = 0. \tag{16}$$

We point out that θ_{OLS} is a *random variable* ($\epsilon_j = Y_j - f(t_j, \vec{\theta})$ is a random variable); hence if $\{y_j\}_{j=1}^n$ is a realization of the *random process* $\{Y_j\}_{j=1}^n$ then solving

$$\hat{\theta}_{\text{OLS}} = \hat{\theta}^n_{\text{OLS}} = \arg\min_{\vec{\theta}\in\Theta} \sum_{j=1}^{n} [y_j - f(t_j, \vec{\theta})]^2 \tag{17}$$

provides a realization for θ_{OLS}. (A remark on notation: for a random variable or estimator θ we will always denote a corresponding realization or estimate with an over hat, e.g., $\hat{\theta}$ is an estimate for θ.)

Noting that

$$\sigma_0^2 = \frac{1}{n} E[\sum_{j=1}^{n} [Y_j - f(t_j, \vec{\theta}_0)]^2] \tag{18}$$

suggests that once we have solved for θ_{OLS} in (15), we may obtain an estimate $\hat{\sigma}^2_{\text{OLS}} = \hat{\sigma}^{2\,n}_{\text{MLE}}$ for σ_0^2.

Even though the error's distribution is not specified we can use asymptotic theory to approximate the mean and variance of the random variable θ_{OLS} [31]. As will be explained in more detail below, as $n \to \infty$, we have that

$$\theta_{\text{OLS}} = \theta^n_{\text{OLS}} \sim \mathcal{N}_p(\vec{\theta}_0, \Sigma^n_0) \approx \mathcal{N}_p(\vec{\theta}_0, \sigma_0^2 [\chi^{nT}(\vec{\theta}_0)\chi^n(\vec{\theta}_0)]^{-1}) \tag{19}$$

where the sensitivity matrix $\chi(\vec{\theta}) = \chi^n(\vec{\theta}) = \{\chi^n_{jk}\}$ is defined as

$$\chi^n_{jk}(\vec{\theta}) = \frac{\partial f(t_j, \vec{\theta})}{\partial \vec{\theta}_k}, \quad j = 1, \ldots, n, \quad k = 1, \ldots, p,$$

and

$$\Sigma^n_0 \equiv \sigma^2_0 [n\Omega_0]^{-1} \tag{20}$$

with

$$\Omega_0 \equiv \lim_{n\to\infty} \frac{1}{n} \chi^{nT}(\vec{\theta}_0)\chi^n(\vec{\theta}_0), \tag{21}$$

where the limit is assumed to exist–see [31]. However, $\vec{\theta}_0$ and σ^2_0 are generally unknown, so one usually will instead use the *realization* $\vec{y} = (y_1, \ldots, y_n)^T$ of the random process $\vec{Y}$ to obtain the estimate

$$\hat{\theta}_{\text{OLS}} = \arg\min_{\vec{\theta}\in\Theta} \sum_{j=1}^{n} [y_j - f(t_j, \vec{\theta})]^2 \tag{22}$$

and the *bias adjusted* estimate

$$\hat{\sigma}^2_{\text{OLS}} = \frac{1}{n-p} \sum_{j=1}^{n} [y_j - f(t_j, \hat{\theta})]^2 \tag{23}$$

to use as an approximation in (19).

We note that (23) represents the estimate for σ^2_0 of (18) with the factor $\frac{1}{n}$ replaced by the factor $\frac{1}{n-p}$ (in the linear case the estimate with $\frac{1}{n}$ can be shown to be biased downward and the same behavior can be observed in the general nonlinear case– see Chap. 12 of [31] and p. 28 of [22]). We remark that (18) is true even in the general nonlinear case (it does not rely on any asymptotic theories although it does depend on the assumption of constant variance being correct).

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

$$\Sigma^n_0 \approx \hat{\Sigma}^n \equiv \hat{\sigma}^2 [\chi^{nT}(\hat{\theta})\chi^n(\hat{\theta})]^{-1}. \tag{24}$$

We can obtain the standard errors $SE(\hat{\theta}_{\text{OLS},k})$ (discussed in more detail in the next section) for the k^{th} element of $\hat{\theta}_{\text{OLS}}$ by calculating $SE(\hat{\theta}_{\text{OLS},k}) \approx \sqrt{\hat{\Sigma}^n_{kk}}$. Also note the similarity between the MLE equations (8) and (9), and the scalar OLS equations (22) and (23). That is, under a normality assumption for the error, the MLE and OLS formulations are equivalent.

If, however, we have a vector of observations for the j^{th} covariate t_j and we assume the variance is still constant in longitudinal data, then the statistical model is reformulated as

$$\vec{Y}_j = \vec{f}(t_j, \vec{\theta}_0) + \vec{\epsilon}_j \tag{25}$$

where $\vec{f} \in R^m$ and

$$V_0 = \text{var}(\vec{\epsilon}_j) = \text{diag}(\sigma_{0,1}^2, \ldots, \sigma_{0,m}^2) \tag{26}$$

for $j = 1, \ldots, n$. Just as in the MLE case we have allowed for the possibility that the observation coordinates Y_j^i may have different *constant* variances $\sigma_{0,i}^2$, i.e. $\sigma_{0,i}^2$ does not necessarily have to equal $\sigma_{0,k}^2$. 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)). The problem will consist of finding the minimizer

$$\theta_{\text{OLS}} = \arg\min_{\vec{\theta} \in \Theta} \sum_{j=1}^{n} [\vec{Y}_j - \vec{f}(t_j, \vec{\theta})]^T V_0^{-1} [\vec{Y}_j - \vec{f}(t_j, \vec{\theta})], \tag{27}$$

where the procedure weights elements of the vector $\vec{Y}_j - \vec{f}(t_j, \vec{\theta})$ according to their variability. (Some authors refer to (27) as a generalized least squares (GLS) procedure, but we will make use of this terminology in a different formulation in subsequent discussions). Just as in the scalar OLS case, θ_{OLS} is a *random variable* (again because $\vec{\epsilon}_j = \vec{Y}_j - \vec{f}(t_j, \vec{\theta})$ is); hence if $\{\vec{y}_j\}_{j=1}^n$ is a realization of the *random process* $\{\vec{Y}_j\}_{j=1}^n$ then solving

$$\hat{\theta}_{\text{OLS}} = \arg\min_{\vec{\theta} \in \Theta} \sum_{j=1}^{n} [\vec{y}_j - \vec{f}(t_j, \vec{\theta})]^T V_0^{-1} [\vec{y}_j - \vec{f}(t_j, \vec{\theta})] \tag{28}$$

provides an estimate (realization) $\hat{\theta} = \hat{\theta}_{\text{OLS}}$ for θ_{OLS}. By the definition of variance

$$V_0 = \text{diag } E\left(\frac{1}{n} \sum_{j=1}^{n} [\vec{Y}_j - \vec{f}(t_j, \vec{\theta}_0)][\vec{Y}_j - \vec{f}(t_j, \vec{\theta}_0)]^T \right),$$

so an unbiased estimate of V_0 for the realization $\{\vec{y}_j\}_{j=1}^n$ is

$$\hat{V} = \text{diag}\left(\frac{1}{n-p} \sum_{j=1}^{n} [\vec{y}_j - \vec{f}(t_j, \hat{\theta})][\vec{y}_j - \vec{f}(t_j, \hat{\theta})]^T \right). \tag{29}$$

However, the estimate $\hat{\theta}$ requires the (generally unknown) matrix V_0 and V_0 requires the unknown vector $\vec{\theta}_0$ so we will instead use the following expressions to calculate $\hat{\theta}$ and $\hat{V}$:

$$\vec{\theta}_0 \approx \hat{\theta} = \arg\min_{\vec{\theta} \in \Theta} \sum_{j=1}^{n} [\vec{y}_j - \vec{f}(t_j, \vec{\theta})]^T \hat{V}^{-1} [\vec{y}_j - \vec{f}(t_j, \vec{\theta})] \tag{30}$$

$$V_0 \approx \hat{V} = \text{diag}\left(\frac{1}{n-p} \sum_{j=1}^{n} [\vec{y}_j - \vec{f}(t_j, \hat{\theta})][\vec{y}_j - \vec{f}(t_j, \hat{\theta})]^T \right). \tag{31}$$

Note that the expressions for $\hat{\theta}$ and $\hat{V}$ constitute a coupled system of equations, which will require greater effort in implementing a numerical scheme.

Just as in the scalar case we can determine the asymptotic properties of the OLS estimator (27). As $n \to \infty$, θ_{OLS} has the following asymptotic properties [22, 31]:

$$\theta_{\text{OLS}} \sim \mathcal{N}(\vec{\theta}_0, \Sigma_0^n), \tag{32}$$

where

$$\Sigma_0^n \approx \left(\sum_{j=1}^{n} D_j^T(\vec{\theta}_0) V_0^{-1} D_j(\vec{\theta}_0) \right)^{-1}, \tag{33}$$

and the $m \times p$ matrix $D_j(\vec{\theta}) = D_j^n(\vec{\theta})$ is given by

$$\begin{pmatrix} \frac{\partial f_1(t_j,\vec{\theta})}{\partial \theta_1} & \frac{\partial f_1(t_j,\vec{\theta})}{\partial \theta_2} & \cdots & \frac{\partial f_1(t_j,\vec{\theta})}{\partial \theta_p} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_m(t_j,\vec{\theta})}{\partial \theta_1} & \frac{\partial f_m(t_j,\vec{\theta})}{\partial \theta_2} & \cdots & \frac{\partial f_m(t_j,\vec{\theta})}{\partial \theta_p} \end{pmatrix}.$$

Since the true value of the parameters $\vec{\theta}_0$ and V_0 are unknown their estimates $\hat{\theta}$ and $\hat{V}$ will be used to approximate the asymptotic properties of the least squares estimator θ_{OLS}:

$$\theta_{\text{OLS}} \sim \mathcal{N}_p(\vec{\theta}_0, \Sigma_0^n) \approx \mathcal{N}_p(\hat{\theta}, \hat{\Sigma}^n) \tag{34}$$

where

$$\Sigma_0^n \approx \hat{\Sigma}^n = \left(\sum_{j=1}^{n} D_j^T(\hat{\theta}) \hat{V}^{-1} D_j(\hat{\theta}) \right)^{-1}. \tag{35}$$

The standard errors can then be calculated for the k^{th} element of $\hat{\theta}_{\text{OLS}}$ ($SE(\hat{\theta}_{\text{OLS},k})$) by $SE(\hat{\theta}_{\text{OLS},k}) \approx \sqrt{\hat{\Sigma}_{kk}}$. Again, we point out the similarity between the MLE equations (12) and (13), and the OLS equations (30) and (31) for the vector statistical model (25).

2.4.2 Numerical Implementation of the OLS Procedure

In the scalar statistical model (14), the estimates $\hat{\theta}$ and $\hat{\sigma}$ can be solved for separately (this is also true of the vector OLS in the case $V_0 = \sigma_0^2 I_m$, where I_m is the $m \times m$ identity) and thus the numerical implementation is straightforward - first determine $\hat{\theta}_{\text{OLS}}$ according to (22) and then calculate $\hat{\sigma}^2_{\text{OLS}}$ according to (23). The estimates $\hat{\theta}$ and $\hat{V}$ in the case of the vector statistical model (25), however, require more effort since they are coupled:

$$\hat{\theta} = \arg\min_{\vec{\theta} \in \Theta} \sum_{j=1}^{n} [\vec{y}_j - \vec{f}(t_j, \vec{\theta})]^T \hat{V}^{-1} [\vec{y}_j - \vec{f}(t_j, \vec{\theta})] \tag{36}$$

$$\hat{V} = \text{diag}\left(\frac{1}{n-p} \sum_{j=1}^{n} [\vec{y}_j - \vec{f}(t_j, \hat{\theta})][\vec{y}_j - \vec{f}(t_j, \hat{\theta})]^T \right). \tag{37}$$

To solve this coupled system the following iterative process will be followed:

1. Set $\hat{V}^{(0)} = \mathbf{I}$ and solve for the initial estimate $\hat{\theta}^{(0)}$ using (36). Set $k = 0$.
2. Use $\hat{\theta}^{(k)}$ to calculate $\hat{V}^{(k+1)}$ using (37).
3. Re-estimate $\vec{\theta}$ by solving (36) with $\hat{V} = \hat{V}^{(k+1)}$ to obtain $\hat{\theta}^{(k+1)}$.
4. Set $k = k + 1$ and return to 2. Terminate the process and set $\hat{\theta}_{\text{OLS}} = \hat{\theta}^{(k+1)}$ when two successive estimates for $\hat{\theta}$ are sufficiently close to one another.

2.4.3 Generalized Least Squares (GLS)

Although in Section 2.4.1 the error's distribution remained unspecified, we did however require that the error remain constant in variance in longitudinal data. That assumption may not be appropriate for data sets whose error is not constant in a longitudinal sense. A common relative error model that experimenters use in this instance for the scalar observation case [22] is

$$Y_j = f(t_j, \vec{\theta}_0)\,(1 + \epsilon_j) \tag{38}$$

where $E(Y_j) = f(t_j, \vec{\theta}_0)$ and $\text{var}(Y_j) = \sigma_0^2 f^2(t_j, \vec{\theta}_0)$ which derives from the assumptions that $E[\epsilon_j] = 0$ and $\text{var}(\epsilon_j) = \sigma_0^2$. We will say that the variance generated in this fashion is non-constant variance. The method we will use to estimate $\vec{\theta}_0$ and σ_0^2 can be viewed as a particular form of the Generalized Least Squares (GLS) method.

To define the *random variable* θ_{GLS} the following equation must be solved for the estimator θ_{GLS}:

$$\sum_{j=1}^{n} w_j[Y_j - f(t_j, \theta_{\text{GLS}})]\nabla f(t_j, \theta_{\text{GLS}}) = 0, \tag{39}$$

where Y_j obeys (38) and $w_j = f^{-2}(t_j, \theta_{\text{GLS}})$. The quantity θ_{GLS} is a random variable, hence if $\{y_j\}_{j=1}^n$ is a *realization* of the random process Y_j then solving

$$\sum_{j=1}^{n} f^{-2}(t_j, \hat{\theta})[y_j - f(t_j, \hat{\theta})]\nabla f(t_j, \hat{\theta}) = 0, \tag{40}$$

for $\hat{\theta}$ we obtain an estimate $\hat{\theta}_{\text{GLS}}$ for θ_{GLS}.

The GLS estimator $\theta_{\text{GLS}} = \theta_{\text{GLS}}^n$ has the following asymptotic properties [22]:

$$\theta_{\text{GLS}} \sim \mathcal{N}_p(\vec{\theta}_0, \Sigma_0^n) \tag{41}$$

where

$$\Sigma_0^n \approx \sigma_0^2 \left(F_{\vec{\theta}}^T(\vec{\theta}_0) W(\vec{\theta}_0) F_{\vec{\theta}}(\vec{\theta}_0) \right)^{-1}, \tag{42}$$

$$F_{\vec{\theta}}(\vec{\theta}) = F^n_{\vec{\theta}}(\vec{\theta}) = \begin{pmatrix} \frac{\partial f(t_1,\vec{\theta})}{\partial \theta_1} & \frac{\partial f(t_1,\vec{\theta})}{\partial \theta_2} & \dots & \frac{\partial f(t_1,\vec{\theta})}{\partial \theta_p} \\ \vdots & & & \vdots \\ \frac{\partial f(t_n,\vec{\theta})}{\partial \theta_1} & \frac{\partial f(t_n,\vec{\theta})}{\partial \theta_2} & \dots & \frac{\partial f(t_n,\vec{\theta})}{\partial \theta_p} \end{pmatrix} = \begin{pmatrix} \nabla f(t_1,\vec{\theta})^T \\ \vdots \\ \nabla f(t_n,\vec{\theta})^T \end{pmatrix}$$

and $W^{-1}(\vec{\theta}) = \text{diag}\left(f^2(t_1,\vec{\theta}),\dots,f^2(t_n,\vec{\theta})\right)$. Note that because $\vec{\theta}_0$ and σ_0^2 are unknown, the estimates $\hat{\theta} = \hat{\theta}_{\text{GLS}}$ and $\hat{\sigma}^2 = \hat{\sigma}^2_{\text{GLS}}$ will be used in (42) to calculate

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

where [22] we take the approximation

$$\sigma_0^2 \approx \hat{\sigma}^2_{\text{GLS}} = \frac{1}{n-p}\sum_{j=1}^{n} \frac{1}{f^2(t_j,\hat{\theta})}[y_j - f(t_j,\hat{\theta})]^2.$$

We can then approximate the standard errors of $\hat{\theta}_{\text{GLS}}$ by taking the square roots of the diagonal elements of $\hat{\Sigma}$. We will also mention that the solutions to (30) and (40) depend upon the numerical method used to find the minimum or root, and since Σ_0 depends upon the estimate for $\vec{\theta}_0$, the standard errors are therefore affected by the numerical method chosen.

2.4.4 GLS motivation

We note the similarity between (16) and (40). The GLS equation (40) can be motivated by examining the weighted least squares (WLS) estimator

$$\theta_{\text{WLS}} = \arg\min_{\vec{\theta}\in\Theta} \sum_{j=1}^{n} w_j[Y_j - f(t_j,\vec{\theta})]^2. \tag{43}$$

In many situations where the observation process is well understood, the weights $\{w_j\}$ may be known. The WLS estimate can be thought of minimizing the distance between the data and model while taking into account unequal quality of the observations [22]. If we differentiate the sum of squares in (43) with respect to $\vec{\theta}$, and *then* choose $w_j = f^{-2}(t_j,\vec{\theta})$, an estimate $\hat{\theta}_{\text{GLS}}$ is obtained by solving

$$\sum_{j=1}^{n} w_j[y_j - f(t_j,\vec{\theta})]\nabla f(t_j,\vec{\theta}) = 0$$

for $\vec{\theta}$. However, we note the GLS relationship (40) does *not* follow from minimizing the weighted least squares with weights chosen as $w_j = f^{-2}(t_j,\vec{\theta})$.

Another motivation for the GLS estimating equation (40) can be found in [18]. In the text the authors claim that if the data are distributed according to the gamma distribution, then the maximum-likelihood estimator for $\vec{\theta}$ is the solution to

$$\sum_{j=1}^{n} f^{-2}(t_j,\vec{\theta})[Y_j - f(t_j,\vec{\theta})]\nabla f(t_j,\vec{\theta}) = 0,$$

which is equivalent to (40). The connection between the MLE and our GLS method is reassuring, but it also poses another interesting question: What if the variance of the data is assumed to not depend on the model output $f(t_j, \vec{\theta})$, but rather on some function $g(t_j, \vec{\theta})$ (i.e., $\text{var}(Y_j) = \sigma_0^2 g^2(t_j, \vec{\theta}) = \sigma_0^2/w_j$)? Is there a corresponding maximum likelihood estimator of $\vec{\theta}$ whose form is equivalent to the appropriate GLS estimating equation ($w_j = g^{-2}(t_j, \vec{\theta})$)

$$\sum_{j=1}^{n} g^{-2}(t_j, \vec{\theta})[Y_j - f(t_j, \vec{\theta})]\nabla f(t_j, \vec{\theta}) = 0 \; ? \tag{44}$$

In their text, Carroll and Rupert [18] briefly describe how distributions belonging to the exponential family of distributions generate maximum-likelihood estimating equations equivalent to (44).

2.4.5 Numerical Implementation of the GLS Procedure

Recall that an estimate $\hat{\theta}_{\text{GLS}}$ can either be solved for directly according to (40) or iteratively using the equations outlined in Section 2.4.3. The iterative procedure as described in [22] is summarized below:

1. Estimate $\hat{\theta}_{\text{GLS}}$ by $\hat{\theta}^{(0)}$ using the OLS equation (15). Set $k = 0$.
2. Form the weights $\hat{w}_j = f^{-2}(t_j, \hat{\theta}^{(k)})$.
3. Re-estimate $\hat{\theta}$ by solving

$$\hat{\theta}^{(k+1)} = \arg\min_{\theta \in \Theta} \sum_{j=1}^{n} \hat{w}_j \left(y_j - f\big(t_j, \vec{\theta}\big) \right)^2$$

   to obtain the $k+1$ estimate $\hat{\theta}^{(k+1)}$ for $\hat{\theta}_{\text{GLS}}$.
4. Set $k = k + 1$ and return to 2. Terminate the process when two of the successive estimates for $\hat{\theta}_{\text{GLS}}$ are sufficiently close.

We note that the above iterative procedure was formulated by minimizing (over $\vec{\theta} \in \Theta$)

$$\sum_{j=1}^{n} f^{-2}(t_j, \tilde{\theta})[y_j - f(t_j, \vec{\theta})]^2$$

and then updating the weights $w_j = f^{-2}(t_j, \tilde{\theta})$ after each iteration. One would hope that after a sufficient number of iterations $\hat{w}_j$ would converge to $f^{-2}(t_j, \hat{\theta}_{\text{GLS}})$. Fortunately, under reasonable conditions, if the process enumerated above is continued a sufficient number of times [22], then $\hat{w}_j \to f^{-2}(t_j, \hat{\theta}_{\text{GLS}})$.

3 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 of Section 2.4.1 (the extension of these ideas to vectors is completely straight-forward). These n scalar observations are represented by the statistical model

$$Y_j \equiv f(t_j, \vec{\theta}_0) + \epsilon_j, \quad j = 1, 2, \ldots, n, \tag{45}$$

where $f(t_j, \vec{\theta}_0)$ is the model for the observations in terms of the state variables and $\vec{\theta}_0 \in \mathbb{R}^p$ is a set of theoretical "true" parameter values (assumed to exist in a standard statistical approach). We further assume that the errors ϵ_j, $j = 1, 2, \ldots, n$, are independent identically distributed (*i.i.d.*) random variables with mean $E[\epsilon_j] = 0$ and constant variance $\text{var}(\epsilon_j) = \sigma_0^2$, where σ_0^2 is unknown. The observations Y_j are then *i.i.d.* with mean $E[Y_j] = f(t_j, \vec{\theta}_0)$ and variance $\text{var}(Y_j) = \sigma_0^2$.

Recall that in the ordinary least squares (OLS) 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

$$\hat{\theta}^n_{\text{OLS}} = \arg\min J_n(\vec{\theta}) = \sum_{j=1}^{n} [y_j - f(t_j, \vec{\theta})]^2. \tag{46}$$

Since Y_j is a random variable, the corresponding estimator $\theta^n = \theta^n_{\text{OLS}}$ (here we wish to emphasize the dependence on the sample size n) is also a random variable with a distribution called the *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 values across all possible realizations, and hence provide a measure of the extent of uncertainty involved in estimating θ using the specific estimator and sample size n in actual data collection.

Under reasonable assumptions on smoothness and regularity (the smoothness requirements for model solutions are readily verified using continuous dependence results for differential equations in most examples; the regularity requirements include, among others, conditions on *how the observations are taken* as sample size increases, i.e., as $n \to \infty$), the standard nonlinear regression approximation theory ([22, 26, 29], and Chapter 12 of [31]) for **asymptotic** (as $n \to \infty$) **distributions** can be invoked. As stated above, this theory yields that the sampling distribution for the estimator $\theta^n(\vec{Y})$, where $\vec{Y} = (Y_1, \ldots, Y_n)^T$, is approximately a p-multivariate Gaussian with mean $E[\theta^n(\vec{Y})] \approx \vec{\theta}_0$ and covariance matrix $\text{var}(\theta^n(\vec{Y})) \approx \Sigma_0^n = \sigma_0^2[n\Omega_0]^{-1} \approx \sigma_0^2[\chi^{nT}(\vec{\theta}_0)\chi^n(\vec{\theta}_0)]^{-1}$. Here $\chi^n(\vec{\theta}) = F_{\vec{\theta}}(\vec{\theta})$ is the $n \times p$ sensitivity matrix with elements

$$\chi_{jk}(\vec{\theta}) = \frac{\partial f(t_j, \vec{\theta})}{\partial \theta_k} \qquad \text{and} \qquad F_{\vec{\theta}}(\vec{\theta}) \equiv (f_{1\vec{\theta}}(\vec{\theta}), \ldots, f_{n\vec{\theta}}(\vec{\theta}))^T,$$

where $f_{j\vec{\theta}}(\vec{\theta}) = \frac{\partial f}{\partial \vec{\theta}}(t_j, \vec{\theta})$. That is, for n large, the sampling distribution approximately satisfies

$$\theta^n_{\text{OLS}}(\vec{Y}) \sim \mathcal{N}_p(\vec{\theta}_0, \Sigma^n_0) \approx \mathcal{N}_p(\vec{\theta}_0, \sigma_0^2[\chi^{nT}(\vec{\theta}_0)\chi^n(\vec{\theta}_0)]^{-1}). \tag{47}$$

There are typically several ways to compute the matrix $F_{\vec{\theta}}$. First, the elements of the matrix $\chi = (\chi_{jk})$ can always be estimated using the forward difference

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

where h_k is a p-vector with a nonzero entry in only the k^{th} component. But, of course, the choice of h_k can be problematic in practice.

Alternatively, if the $f(t_j, \vec{\theta})$ correspond to longitudinal observations $\vec{y}(t_j) = \mathcal{C}\vec{x}(t_j; \vec{\theta})$ of solutions $\vec{x} \in \mathbb{R}^N$ to a parameterized N-vector differential equation system $\dot{\vec{x}} = \vec{g}(t, \vec{x}(t), \vec{\theta})$ as in (1), then one can use the $N \times p$ matrix **sensitivity equations** (see [4, 9] and the references therein)

$$\frac{d}{dt}\left(\frac{\partial \vec{x}}{\partial \vec{\theta}}\right) = \frac{\partial \vec{g}}{\partial \vec{x}}\frac{\partial \vec{x}}{\partial \vec{\theta}} + \frac{\partial \vec{g}}{\partial \vec{\theta}} \tag{48}$$

to obtain

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

Finally, in some cases the function $f(t_j, \vec{\theta})$ may be sufficiently simple so as to allow one to derive analytical expressions for the components of $F_{\vec{\theta}}$.

Since $\vec{\theta}_0$, σ_0 are unknown, we will use their estimates to make the approximation

$$\Sigma^n_0 \approx \sigma_0^2[\chi^{nT}(\vec{\theta}_0)\chi^n(\vec{\theta}_0)]^{-1} \approx \hat{\Sigma}^n(\hat{\theta}^n_{\text{OLS}}) = \hat{\sigma}^2[\chi^{nT}(\hat{\theta}^n_{\text{OLS}})\chi^n(\hat{\theta}^n_{\text{OLS}})]^{-1}, \tag{49}$$

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

$$\sigma_0^2 \approx \hat{\sigma}^2 = \frac{1}{n-p}\sum_{j=1}^{n}[y_j - f(t_j, \hat{\theta}^n_{\text{OLS}})]^2. \tag{50}$$

Standard errors to be used in the confidence interval calculations are thus given by $SE_k(\hat{\theta}^n) = \sqrt{\Sigma_{kk}(\hat{\theta}^n)}$, $k = 1, 2, \ldots, p$ (see [19]).

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

$$P\{\hat{\theta}^n_k - t_{1-\alpha/2}SE_k(\hat{\theta}^n) < \theta_{0k} < \hat{\theta}^n_k + t_{1-\alpha/2}SE_k(\hat{\theta}^n)\} = 1 - \alpha, \tag{51}$$

where $\alpha \in [0, 1]$ and $t_{1-\alpha/2} \in \mathbb{R}_+$. Given a small α value (e.g., $\alpha = .05$ for 95% confidence intervals), the critical value $t_{1-\alpha/2}$ is computed from the Student's t distribution t^{n-p} with

$n-p$ degrees of freedom. The value of $t_{1-\alpha/2}$ is determined by $P\{T \geq t_{1-\alpha/2}\} = \alpha/2$ where $T \sim t^{n-p}$. 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 θ_0 using the given estimator and sample size n.

When one is taking longitudinal samples corresponding to solutions of a dynamical system, the $n \times p$ sensitivity matrix depends explicitly on where in time the observations are taken when $f(t_j, \vec{\theta}) = \mathcal{C}x(t_j, \vec{\theta})$ as mentioned above. That is, the sensitivity matrix

$$\chi(\vec{\theta}) = F_{\vec{\theta}}(\vec{\theta}) = \left(\frac{\partial f(t_j, \vec{\theta})}{\partial \theta_k} \right)$$

depends on the number n and the nature (for example, how taken) of the sampling times $\{t_j\}$. Moreover, it is the matrix $[\chi^T \chi]^{-1}$ in (49) and the parameter $\hat{\sigma}^2$ in (50) that ultimately determine the standard errors and confidence intervals. At first investigation of (50), it appears that an increased number n of samples might drive $\hat{\sigma}^2$ (and hence the SE) to zero as long as this is done in a way to maintain a bound on the residual sum of squares in (50). However, we observe that the *condition number* of the matrix $\chi^T \chi$ is also very important in these considerations and increasing the sampling could potentially adversely affect the inversion of $\chi^T \chi$. In this regard, we note that among the important hypotheses in the asymptotic statistical theory (see p. 571 of [31]) is the existence of a matrix function $\Omega(\vec{\theta})$ such that

$$\frac{1}{n} \chi^{nT}(\vec{\theta}) \chi^n(\vec{\theta}) \to \Omega(\vec{\theta}) \quad \text{uniformly in } \vec{\theta} \text{ as } n \to \infty,$$

with $\Omega_0 = \Omega(\vec{\theta_0})$ a **nonsingular** matrix. It is this condition that is rather easily violated in practice when one is dealing with data from differential equation systems, especially near an equilibrium or steady state (see the examples of [4]).

All of the above theory readily generalizes to vector systems with partial, non-scalar observations. Suppose now we have the vector system (1) with partial vector observations given by (5.1), that is, we have m coordinate observations where $m \leq N$. In this case, we have

$$\frac{d\vec{x}}{dt}(t) = \vec{g}(t, \vec{x}(t), \vec{\theta}) \tag{52}$$

and

$$\vec{y}_j = \vec{f}(t_j, \vec{\theta}_0) + \vec{\epsilon}_j = \mathcal{C}\vec{x}(t_j, \vec{\theta}_0) + \vec{\epsilon}_j, \tag{53}$$

where $\mathcal{C}$ is an $m \times N$ matrix and $\vec{f} \in R^m, \vec{x} \in R^N$. As already explained in Section 2.4.1, if we assume that different observation coordinates f_i may have different variances σ_i^2 associated with different coordinates of the errors ϵ_j, then we have that $\vec{\epsilon}_j$ is an m-dimensional random vector with

$$E[\vec{\epsilon}_j] = 0, \quad \text{var}(\vec{\epsilon}_j) = V_0,$$

where $V_0 = \text{diag}(\sigma_{0,1}^2, ..., \sigma_{0,m}^2)$, and we may follow a similar asymptotic theory to calculate approximate covariances, standard errors and confidence intervals for parameter estimates.

Since the computations for standard errors and confidence intervals (and also *model comparison tests*) depend on *an asymptotic limit distribution theory*, one should interpret the findings as sometimes crude indicators of uncertainty inherent in the inverse problem findings. Nonetheless, it is useful to consider the formal mathematical requirements underpinning these techniques.

Among the more readily checked hypotheses are those of the statistical model requiring that the errors ϵ_j, $j = 1, 2, \ldots, n$, are independent and identically distributed (*i.i.d.*) random variables with mean $E[\epsilon_j] = 0$ and constant variance $\text{var}(\epsilon_j) = \sigma_0^2$.

- After carrying out the estimation procedures, one can readily plot the *residuals* $r_j = y_j - f(t_j, \hat{\theta}_{OLS}^n)$ *vs. time* t_j and the *residuals vs. the resulting estimated model/ observation* $f(t_j, \hat{\theta}_{OLS}^n)$ *values*. A random pattern for the first is strong support for validity of independence assumption; a non increasing, random pattern for latter suggests assumption of constant variance may be reasonable.

- The underlying assumption that sampling size n must be large (recall the theory is asymptotic in that it holds as $n \to \infty$) is not so readily "verified"–often ignored (albeit at the user's peril in regard to the quality of the uncertainty findings).

Often asymptotic results provide remarkably good approximations to the true sampling distributions for finite n. However, in practice there is no way to ascertain whether theory holds for a specific example.

4 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 $Y_j = f(t_j, \vec{\theta}_0) + \epsilon_j$ whereas the GLS should be used for problems in which we have nonconstant variance observations $Y_j = f(t_j, \vec{\theta}_0)(1 + \epsilon_j)$.

We emphasize that in order 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.

Not doing so can lead to incorrect conclusions. In either case, the standard error calculations are not valid unless the correct formulas (which depends 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, the two residuals tests can be performed *after* the estimation procedure has been completed to assist in concluding whether or not the correct asymptotic statistics were used.

4.1 Residual Plots

One can carry out simulation studies with a proposed mathematical model 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 noise

$$Y_j = f(t_j, \vec{\theta}_0) + \frac{k}{100}\epsilon_j, \qquad \text{Var}(Y_j) = \frac{k^2}{10000}\sigma^2,$$

and nonconstant variance noise

$$Y_j = f(t_j, \vec{\theta}_0)(1 + \frac{k}{100}\epsilon_j), \qquad \text{Var}(Y_j) = \frac{k^2}{10000}\sigma^2\, f^2(t_j, \vec{\theta}_0).$$

We can obtain a data set by considering a *realization* $\{y_j\}_{j=1}^n$ of the random process $\{Y_j\}_{j=1}^n$ through a realization of $\{\epsilon_j\}_{j=1}^n$ and then calculate an estimate $\hat{\theta}$ of $\vec{\theta}_0$ using the OLS or GLS procedure.

We will then 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 error then

$$Y_j = f(t_j, \vec{\theta}_0) + \epsilon_j \quad \text{or} \quad \epsilon_j = Y_j - f(t_j, \vec{\theta}_0).$$

Since it is assumed that the error ϵ_j is *i.i.d.* a plot of the residuals $r_j = y_j - f(t_j, \hat{\theta})$ vs. t_j should be random. Also, 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. 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.

We turn next to questions of what to expect if this residual test is applied to a data set that has nonconstant variance generated error. That is, we wish to investigate what happens if the data are incorrectly assumed to have constant variance error when in fact they have nonconstant variance error. Since in the nonconstant variance example, $R_j = Y_j - f(t_j, \vec{\theta}_0) = f(t_j, \vec{\theta}_0)\,\epsilon_j$ depends upon the deterministic model $f(t_j, \vec{\theta}_0)$, 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 nonconstant variance 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. Thus $r_j = y_j - f(t_j, \hat{\theta})$ vs. $f(t_j, \hat{\theta})$ should have a fan shape in the nonconstant variance case.

In summary, if a data set has nonconstant variance generated data, then

$$Y_j = f(t_j, \vec{\theta}_0) + f(t_j, \vec{\theta}_0)\,\epsilon_j \quad \text{or} \quad \epsilon_j = \frac{Y_j - f(t_j, \vec{\theta}_0)}{f(t_j, \vec{\theta}_0)}.$$

If the distribution ϵ_j is *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 in nonconstant variance generated data. 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.

Another question of interest concerns the case in which the data are incorrectly assumed to have nonconstant variance error when in fact they have constant variance error. Since $Y_j - f(t_j, \vec{\theta}_0) = \epsilon_j$ 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.

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 for the 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). Another common technique when plotting against model predictions is to plot against $\log f(t_j, \hat{\theta})$ 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 $R_j^m = \frac{Y_j - f(t_j, \hat{\theta})}{f(t_j, \hat{\theta})}$ 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 chose here to reduce the data sets (although this sometimes could lead to a deterioration in the estimation results obtained). In our examples below, estimates obtained using a truncated data set will be denoted by $\hat{\theta}_{\text{OLS}}^{\text{tcv}}$ for constant variance data and $\hat{\theta}_{\text{OLS}}^{\text{tncv}}$ for nonconstant variance data.

4.2 Example using Residual Plots

We illustrate residual plot techniques by exploring a widely studied model - the logistic population growth model of Verhulst/Pearl

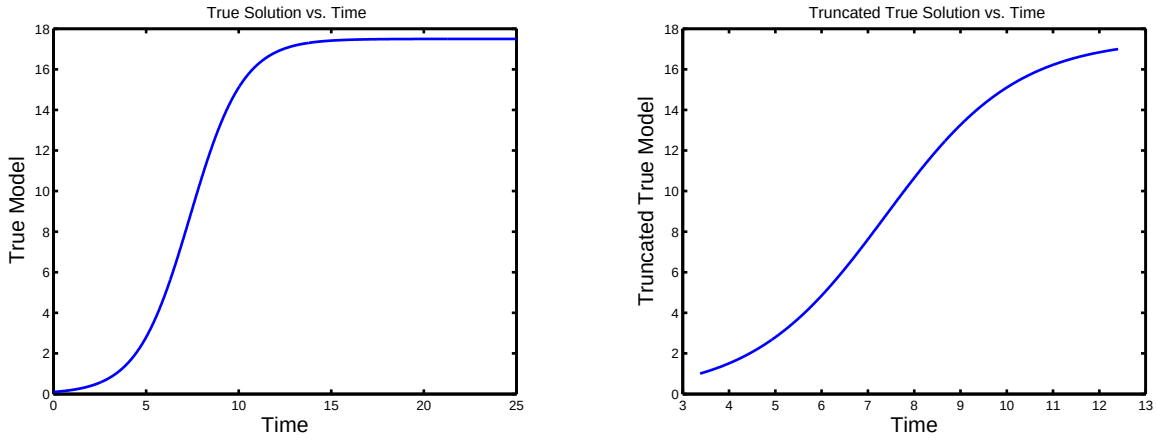
$$\dot{x} = rx(1 - \frac{x}{K}), \quad x(0) = x_0. \tag{54}$$

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

$$x(t) = \frac{K\, x_0 e^{rt}}{K + x_0\,(e^{rt} - 1)}. \tag{55}$$

The left plot in Figure 1 depicts the solution of the logistic model for $K = 17.5$, $r = .7$ and $x_0 = 1$ for $0 \le t \le 25$. If high frequency repeated or nearly repeated values (i.e., near the initial value x_0 or near the the asymptote $x = K$) are removed from the original plot, the resulting truncated plot is given in the right panel of Figure 1 (there are no near zero values for this function).

Figure 1: Original and truncated logistic curve with $K = 17.5$, $r = .7$ and $x_0 = .1$.



For this example we generated both constant variance and nonconstant variance noisy data (we sampled from $\mathcal{N}(0,1)$ random variables to obtain realizations of ϵ_j) and obtained estimates $\hat{\theta}$ of $\vec{\theta}_0 = (K, r, x_0)$ by applying either the OLS or GLS method to a realization $\{y_j\}_{j=1}^n$ of the random process $\{Y_j\}_{j=1}^n$. The initial guesses $\vec{\theta}_{init} = \hat{\theta}^{(0)}$ along with estimates for each method and error structure are given in Tables 1-4 (the superscript tcv and tncv denote the estimate obtained using the truncated data set). As expected, both methods do a good job of estimating $\vec{\theta}_0$, however the error structure was not always correctly specified since incorrect asymptotic formulas were used in some cases.

Table 1: Estimation using the OLS procedure with constant variance data for $k = 5$.

k	$\vec{\theta}_{\text{init}}$	$\vec{\theta}_0$	$\hat{\theta}^{\text{cv}}_{\text{OLS}}$	SE($\hat{\theta}^{\text{cv}}_{\text{OLS}}$)	$\hat{\theta}^{\text{tcv}}_{\text{OLS}}$	SE($\hat{\theta}^{\text{tcv}}_{\text{OLS}}$)
5	17	17.5	1.7500e+001	1.5800e-003	1.7494e+001	6.4215e-003
5	.8	.7	7.0018e-001	4.2841e-004	7.0062e-001	6.5796e-004
5	1.2	.1	9.9958e-002	3.1483e-004	9.9702e-002	4.3898e-004

Table 2: Estimation using the GLS procedure with constant variance data for $k = 5$.

k	$\vec{\theta}_{\text{init}}$	$\vec{\theta}_0$	$\hat{\theta}^{\text{cv}}_{\text{GLS}}$	SE($\hat{\theta}^{\text{cv}}_{\text{GLS}}$)	$\hat{\theta}^{\text{tcv}}_{\text{GLS}}$	SE($\hat{\theta}^{\text{tcv}}_{\text{GLS}}$)
5	17	17.5	1.7500e+001	1.3824e-004	1.7494e+001	9.1213e-005
5	.8	.7	7.0021e-001	7.8139e-005	7.0060e-001	1.6009e-005
5	1.2	.1	9.9938e-002	6.6068e-005	9.9718e-002	1.2130e-005

Table 3: Estimation using the OLS procedure with nonconstant variance data for $k = 5$.

k	$\vec{\theta}_{\text{init}}$	$\vec{\theta}_0$	$\hat{\theta}^{\text{ncv}}_{\text{OLS}}$	SE($\hat{\theta}^{\text{ncv}}_{\text{OLS}}$)	$\hat{\theta}^{\text{tncv}}_{\text{OLS}}$	SE($\hat{\theta}^{\text{tncv}}_{\text{OLS}}$)
5	17	17.5	1.7499e+001	2.2678e-002	1.7411e+001	7.1584e-002
5	.8	.7	7.0192e-001	6.1770e-003	7.0955e-001	7.6039e-003
5	1.2	.1	9.9496e-002	4.5115e-003	9.4967e-002	4.8295e-003

Table 4: Estimation using the GLS procedure with nonconstant variance data for $k = 5$.

k	$\vec{\theta}_{\text{init}}$	$\vec{\theta}_0$	$\hat{\theta}^{\text{ncv}}_{\text{GLS}}$	SE($\hat{\theta}^{\text{ncv}}_{\text{GLS}}$)	$\hat{\theta}^{\text{tncv}}_{\text{GLS}}$	SE($\hat{\theta}^{\text{tncv}}_{\text{GLS}}$)
5	17	17.5	1.7498e+001	9.4366e-005	1.7411e+001	3.1271e-004
5	.8	.7	7.0217e-001	5.3616e-005	7.0959e-001	5.7181e-005
5	1.2	.1	9.9314e-002	4.4976e-005	9.4944e-002	4.1205e-005

When the OLS method was applied to nonconstant variance data and the GLS method was applied to constant variance data, the residual plots given below do reveal that the error structure was misspecified. For instance, the plot of the residuals for $\hat{\theta}^{\text{ncv}}_{\text{OLS}}$ given in Figures 4 and 5 reveal a fan shaped pattern, which indicates the constant variance assumption is suspect. In addition, the plot of the residuals for $\hat{\theta}^{\text{cv}}_{\text{GLS}}$ given in Figures 6 and 7 reveal an inverted fan shaped pattern, which indicates the nonconstant variance assumption is suspect. As expected, when the correct error structure is specified, the *i.i.d.* test and the model dependence test each display a random pattern (Figures 2, 3 and Figures 8, 9).

Also, included in the right panel of Figures 2 - 9 are the residual plots with the truncated data sets. In those plots only model values between one and seventeen were considered (i.e. $1 \leq y_j \leq 17$). Doing so removed the dense vertical lines in the plots with $f(t_j, \hat{\theta})$ along the x-axis. Nonetheless, the conclusions regarding the error structure remain the same.

Figure 2: Residual plots: Original and truncated logistic curve for $\hat{\theta}^{cv}_{\text{OLS}}$ with $k = 5$.

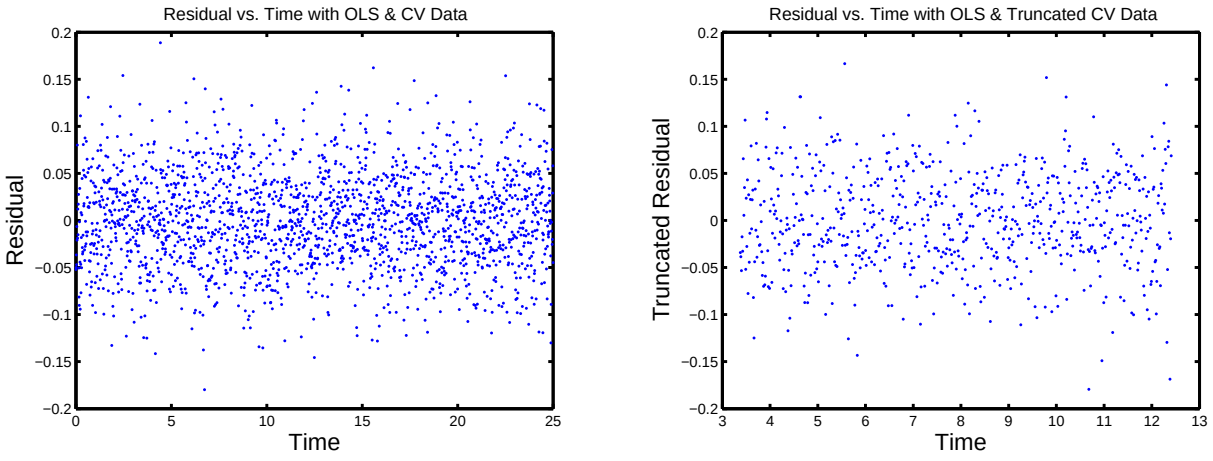


Figure 3: Original and truncated logistic curve for $\hat{\theta}^{cv}_{\text{OLS}}$ with $k = 5$.

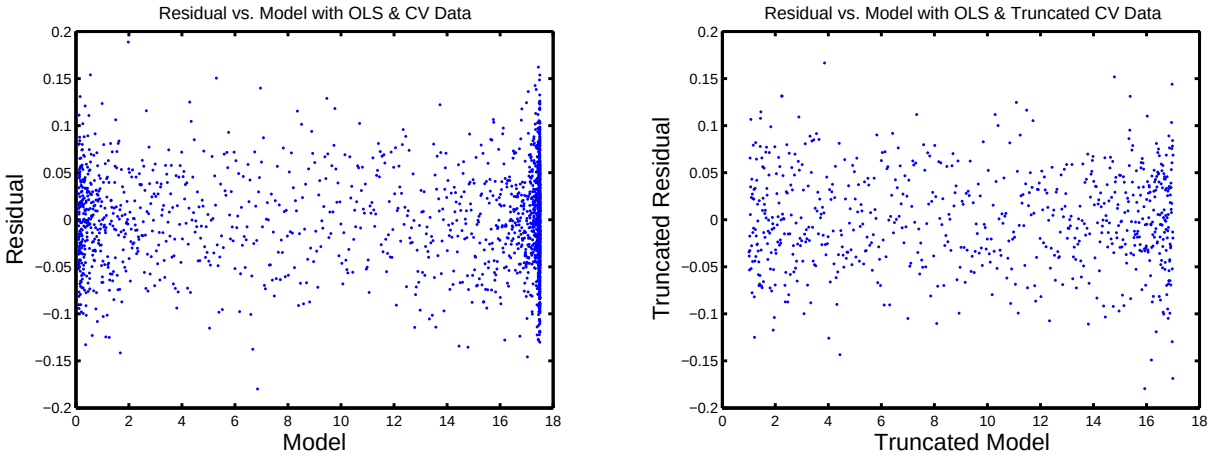


Figure 4: Original and truncated logistic curve for $\hat{\theta}^{ncv}_{\text{OLS}}$ with $k = 5$.

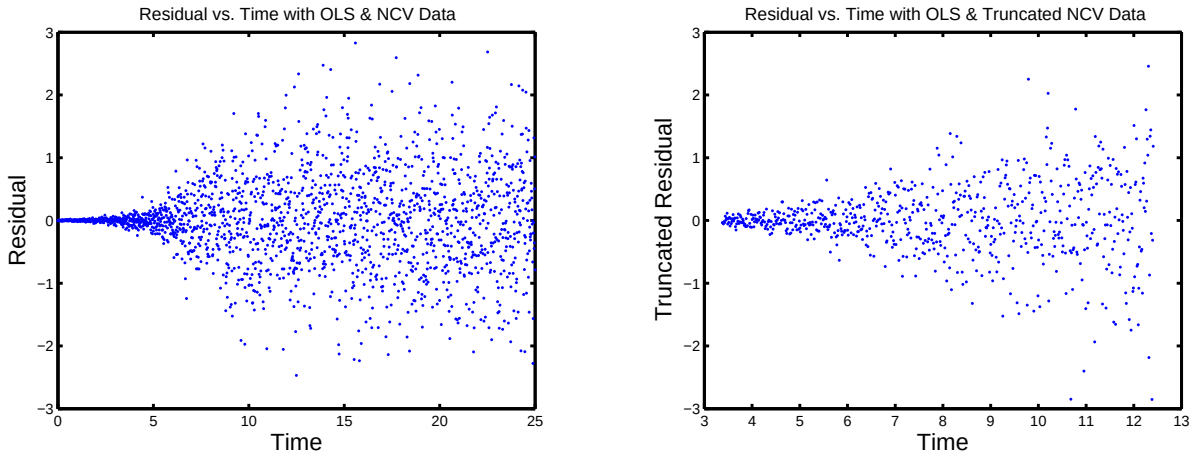


Figure 5: Original and truncated logistic curve for $\hat{\theta}^{ncv}_{\text{OLS}}$ with $k = 5$.

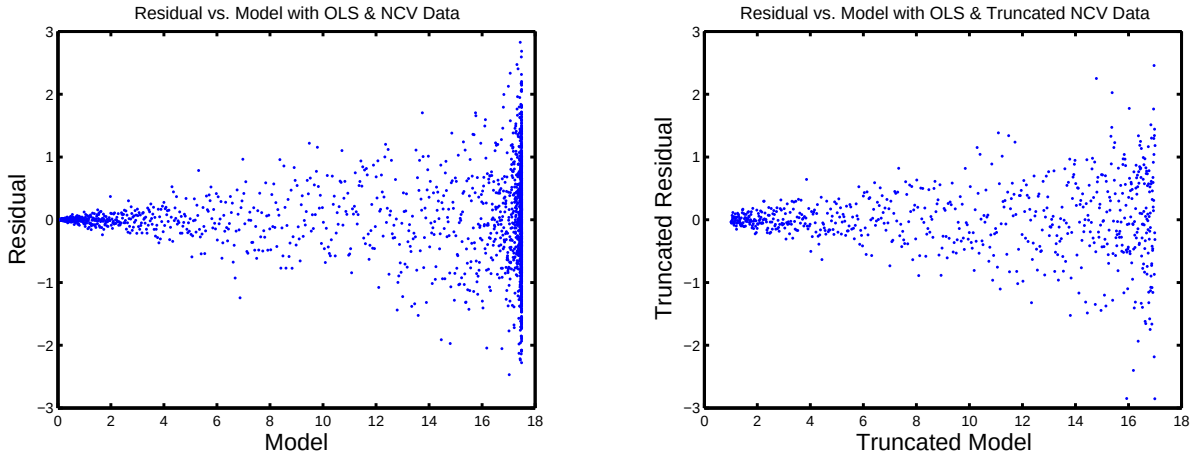
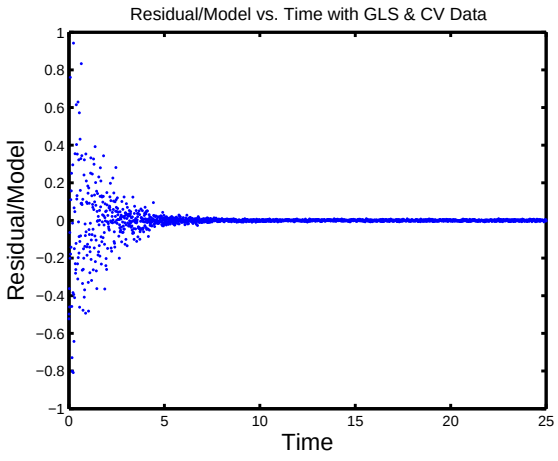


Figure 6: Original and truncated logistic curve for $\hat{\theta}^{cv}_{\text{GLS}}$ with $k = 5$.



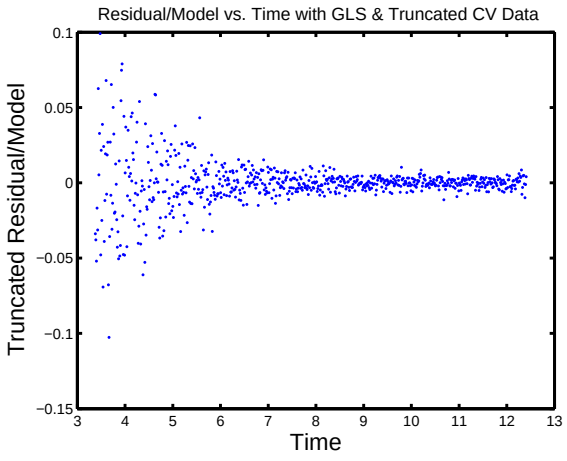
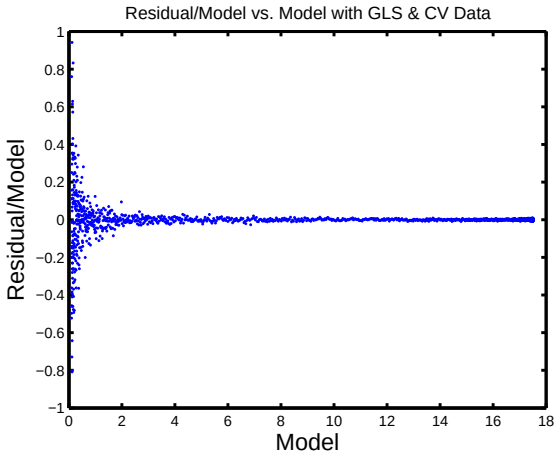


Figure 7: Original and truncated logistic curve for $\hat{\theta}^{cv}_{\text{GLS}}$ with $k = 5$.



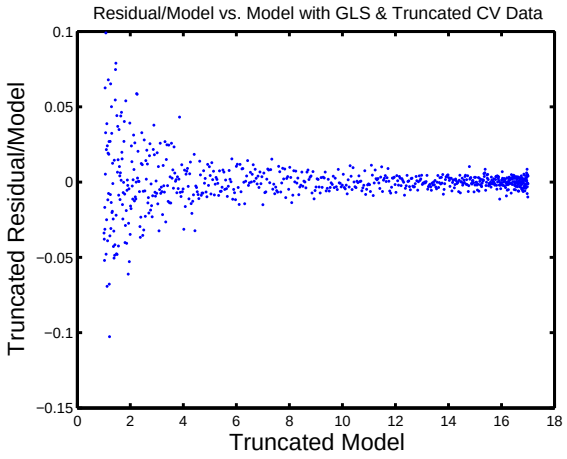


Figure 8: Original and truncated logistic curve for $\hat{\theta}^{ncv}_{\text{GLS}}$ with $k = 5$.

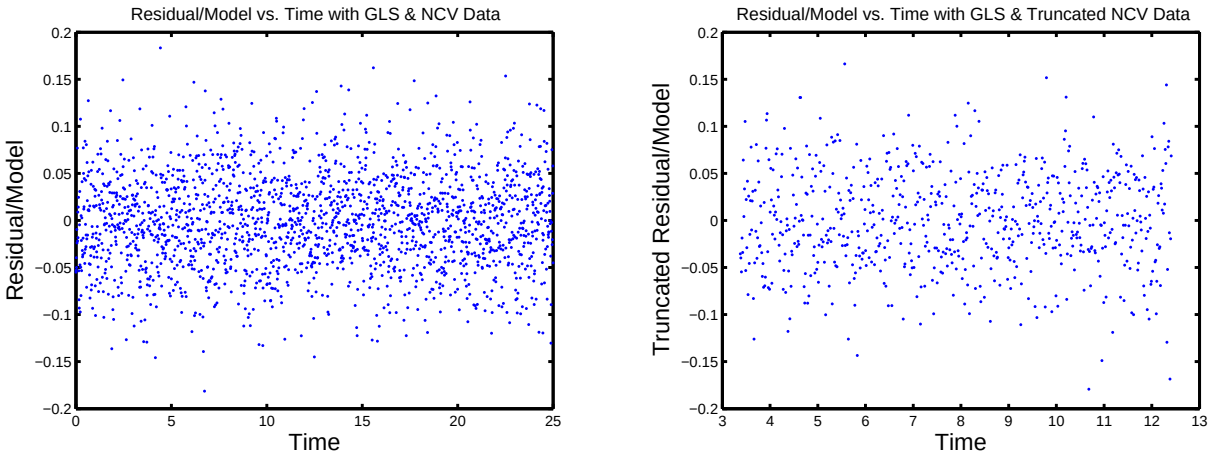
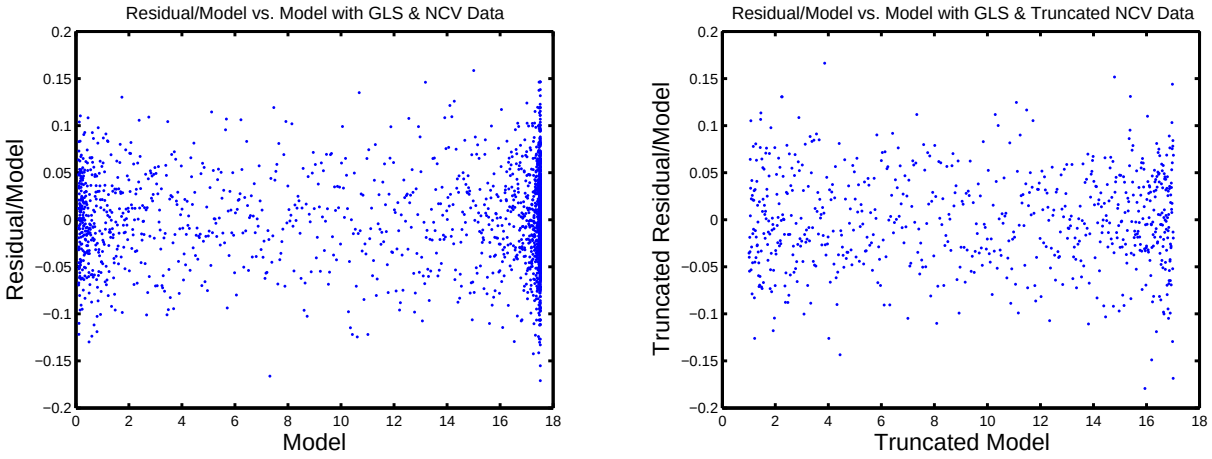


Figure 9: Original and truncated logistic curve for $\hat{\theta}^{ncv}_{\text{GLS}}$ with $k = 5$.

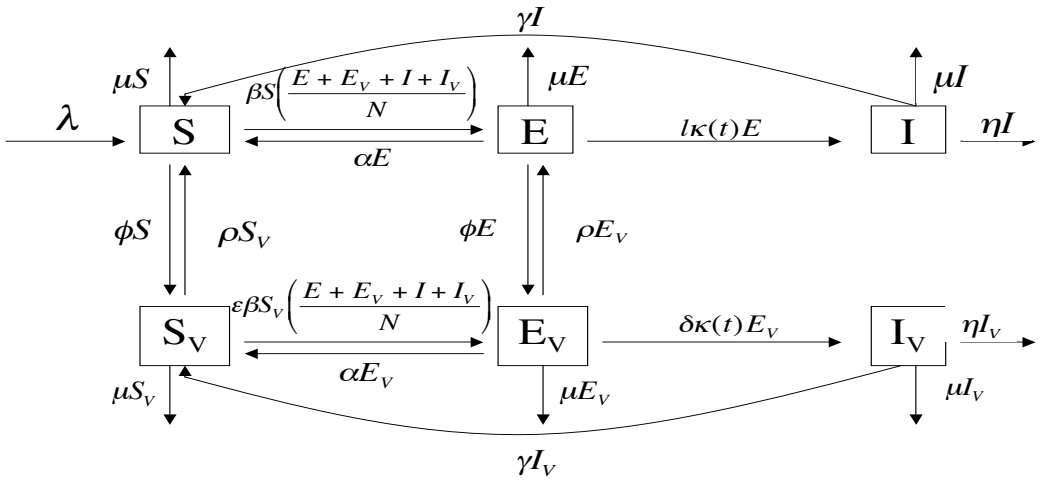


In addition to the residual plots, we can also compare the standard errors obtained for each simulation. At a quick glance of Tables 1 - 4, the standard error of the parameter K in the truncated data set is larger than the standard error of K in the original data set. This behavior is expected. If we remove the "flat" region in the logistic curve, we actually discard measurements with high information content about the carrying capacity K [4]. Doing so reduces the quality of the estimator K. Another interesting observation is that the standard errors of the GLS estimate are more optimistic than that of the OLS estimate, even when the non-constant variance assumption is wrong. This example further solidifies the conclusion we will make with the epidemiological model described below - before one reports an estimate and corresponding standard errors, there needs to be some assurance that the proper error structure has been specified.

5 Pneumococcal Disease Dynamics Model

To explore these ideas in the context of epidemiology, we discuss a population level model of pneumococcal disease dynamics as an example. This model has previously been applied to surveillance data available via the Australian National Notifiable Diseases Surveillance System in [32]. Monthly case notifications of invasive pneumococcal disease (IPD) and annual vaccination information were used to estimate unknown model parameters and to assess the impact of a newly implemented vaccination policy. Here we illustrate, with this example, the effects of incorrect versus correct statistical models assumed to represent observed data in reporting parameter values and their corresponding standard errors. Most importantly, we discuss relevant residual plots and how to use these to determine if reasonable assumptions on observed error have been made.

Figure 10: Pneumococcal infection dynamics with vaccination.



In this model, shown in Figure 10, individuals are classified according to their epidemiological status with respect to invasive pneumococcal diseases, which include pneumonia, bacteremia, meningitis and are defined as the presence of *Streptococcus pneumoniae* in any normal sterile fluid in the body. Individuals are considered susceptible, or in the S class, in the absence of this bacteria. The E class represents individuals whose nasopharyngeal regions are asymptomatically colonized by *S. pneumoniae*, a stage that is typically transient, but always precedes infection. Should a colony of *S. pneumoniae* be successful in establishing an infection, the individual then exhibits a clinical condition described above, and is then considered infected or in the I class. We consider vaccines which prevent progression to infection, or possibly, asymptomatic colonization. This protection is not complete, and the efficacy with which this is accomplished is $1-\delta$ and $1-\epsilon$, respectively. Once vaccinated, individuals may enter any of the epidemiological states, S_V, E_V, and I_V, although they do so with altered rates. The model equations (for detailed derivations, see [32]) are given by

$$\frac{dS}{dt} = \lambda - \beta S\frac{E + E_V + I + I_V}{N} + \alpha E + \gamma I - \phi S + \rho S_V - \mu S \tag{56}$$

$$\frac{dE}{dt} = \beta S\frac{E + E_V + I + I_V}{N} - \alpha E - l\kappa(t)E - \phi E + \rho E_V - \mu E \tag{57}$$

$$\frac{dS_V}{dt} = \phi S - \epsilon\beta S_V\frac{E + E_V + I + I_V}{N} + \alpha E_V + \gamma I_V - \rho S_V - \mu S_V \tag{58}$$

$$\frac{dE_V}{dt} = \epsilon\beta S_V\frac{E + E_V + I + I_V}{N} - \alpha E_V + \phi E - \rho E_V - \delta\kappa(t)E_V - \mu E_V \tag{59}$$

$$\frac{dI}{dt} = l\kappa(t)E - (\gamma + \eta + \mu)I \tag{60}$$

$$\frac{dI_V}{dt} = \delta\kappa(t)E_V - (\gamma + \eta + \mu)I_V. \tag{61}$$

Seasonality of invasive pneumococcal diseases has been observed and studies support a seasonal infection rate, κ, rather than a seasonal effective contact rate, β. Thus, we assume the form

$$\kappa(t) = \kappa_0\left(1 + \kappa_1\cos[\omega(t - \tau)]\right),$$

for $\kappa(t)$ to reflect seasonal changes in host susceptibility to pneumococcal infection.

5.1 Statistical Models of Case Notification Data

Monthly case notifications $f(t_j, \vec{\theta})$ are best represented as integrals of the new infection rates,

$$f(t_j, \vec{\theta}) = \int_{t_j}^{t_{j+1}} \left[l\kappa(s)E(s) + \delta\kappa(s)E_V(s)\right] ds,$$

(including those in the vaccinated class) over each month, since they represent the number of cases reported during the month and do not provide any information on how long individuals remain in an infected state. We use these data to estimate $\vec{\theta} = (\beta, \kappa_0, \kappa_1)^T$. Before using the model with surveillance data, we test the model and methodology capabilities with simulated "data". Following the procedures in the logistic example discussions in Section 4, we generate data according to two statistical models:

$$Y_j = f(t_j, \theta_0) + \epsilon_j, \tag{62}$$

$$Y_j = f(t_j, \theta_0)(1 + \epsilon_j), \tag{63}$$

for $j = 1, ..., n$, where $\vec{\theta}_0$ are the 'true' values of the parameters used to generate the data. In both (62) and (63), the ϵ_j are independent and identically distributed (*i.i.d.*) random variables with $E[\epsilon_j] = 0$ and $\text{var}(\epsilon_j) = \sigma_0^2$. In model (62), however, the residual is then $R_j = Y_j - f(t_j, \vec{\theta}_0) = \epsilon_j$ and thus R_j satisfies $E[R_j] = 0$ and $\text{var}(R_j) = \sigma_0^2$. As before, we

will refer to this error with *constant variance*, or CV. The second case, (63), has residuals of the form $R_j = Y_j - f(t_j, \vec{\theta}_0) = \epsilon_j f(t_j, \vec{\theta}_0)$, so the residual is actually proportional to the model, $f(t_j, \vec{\theta}_0)$, at each time point t_j, and thus this is an example of error with *nonconstant variance*, or NCV. We note that in this case $E[R_j] = 0$ and $\text{var}(R_j) = \sigma_0^2 f^2(t_j, \vec{\theta}_0)$ or $\frac{R_j}{f(t_j, \vec{\theta}_0)}$ has mean zero and variance σ_0^2.

For illustration, we consider the same four cases as with the logistic example in Section 4:

1. OLS estimation of $\hat{\theta}$ using data generated by model (62) with constant variance observational error: $\theta_{OLS}(Y_{CV})$,
2. OLS estimation of $\hat{\theta}$ using data generated by model (63) with nonconstant variance observational error: $\theta_{OLS}(Y_{NCV})$,
3. GLS estimation of $\hat{\theta}$ using data generated by model (62) with constant variance observational error: $\theta_{GLS}(Y_{CV})$,
4. GLS estimation of $\hat{\theta}$ using data generated by model (63) with nonconstant variance observational error: $\theta_{GLS}(Y_{NCV})$.

We compare the parameter estimates $\hat{\theta}$ and standard errors $SE(\hat{\theta})$ obtained in each case. Further we discuss how to interpret plots of $r_j = y_j - f(t_j, \hat{\theta})$ versus t_j and $f(t_j, \hat{\theta})$ to assess whether reasonable assumptions have been made in assuming the statistical model for the data.

5.2 Inverse Problem Results: Simulated Data

Data were generated with $n = 60$ time points (equivalent to five years of data), with the set of parameters

$$\vec{\theta}_0 = \begin{pmatrix} \beta \\ \kappa_0 \\ \kappa_1 \end{pmatrix} = \begin{pmatrix} 1.5 \\ 1.4e^{-3} \\ 0.55 \end{pmatrix}.$$

Error was added to the forward solution according to two statistical models, as described in Section 5.1. In the case of constant variance observational error, the error is scaled to the magnitude of the model but not in a time-dependent manner. In this case we generated noisy data by sampling from a $\mathcal{N}(0, 1)$ distribution (we could of course have sampled from any other random variable). Therefore, for constant variance error of about $k\%$ of the average magnitude of the $f(t_j, \vec{\theta}_0)$,

$$\epsilon_j \sim \frac{k}{100} \text{avg}_j f(t_j, \vec{\theta}_0) \mathcal{N}(0, 1).$$

So in this case $\epsilon_j \sim \mathcal{N}(0, [\frac{k}{100} \text{avg}_j f(t_j, \vec{\theta}_0)]^2)$ with ϵ_j (and also R_j) *i.i.d.* In the second statistical model, the error depends on time and is scaled by the model at each time point, i.e., the error is *relative*. In this case the error is added to the observations by

Table 5: Parameter estimates from data with constant variance CV error.

$\vec{\theta}$	$\vec{\theta}_0$	$\vec{\theta}_{\text{init}}$	$\hat{\theta}_{OLS}$	$SE(\hat{\theta}_{OLS})$	$\hat{\theta}_{GLS}$	$SE(\hat{\theta}_{GLS})$
β	1.5	1.55	1.4845	0.038	1.51186	0.017
κ_0	$1.4e^{-3}$	$1.3e^{-3}$	$1.4188e^{-3}$	$2.1e^{-4}$	$1.3203e^{-3}$	$1.2e^{-4}$
κ_1	0.55	0.65	0.56203	0.050	0.56047	0.019
RSS			$1.6831e^4$		$1.722e^4$	

$$R_j = f(t_j, \vec{\theta}_0)\epsilon_j \sim f(t_j, \vec{\theta}_0)\frac{k}{100}\mathcal{N}(0,1),$$

with $\epsilon_j \sim \mathcal{N}(0, [\frac{k}{100}f(t_j, \vec{\theta}_0)]^2)$, and again the ϵ_j are $i.i.d.$, but now the R_j are not $i.i.d.$ This enables us to compare different types of error on the same scale: one independent of time and observation magnitude, and one dependent on observation magnitude, and thus time. With the present examples, we have taken $k = 10$.

The results from using an OLS and GLS estimator with data generated with constant variance error are displayed in Table 5, and the fitted model solutions displayed in Figure 11. Both estimators do an arguably similar job at producing the true values, that is $\hat{\theta}_{OLS}$ and $\hat{\theta}_{GLS}$ are comparably close to θ_0. The standard errors $SE(\hat{\theta}_{GLS})$ for the GLS estimator however, are all smaller, and seem to indicate that the corresponding estimates are more "reliable". This, however, is not true because they are based on incorrect formulae, as we shall see in our examination of the error plots for both of these cases. Note that from Figure 11 and the residual sum of squares, RSS, in both cases, there is no clear argument from these results as to which estimator is better suited for use with the data.

Figure 11: Best fit model solutions to monthly case notifications with constant variance CV error.

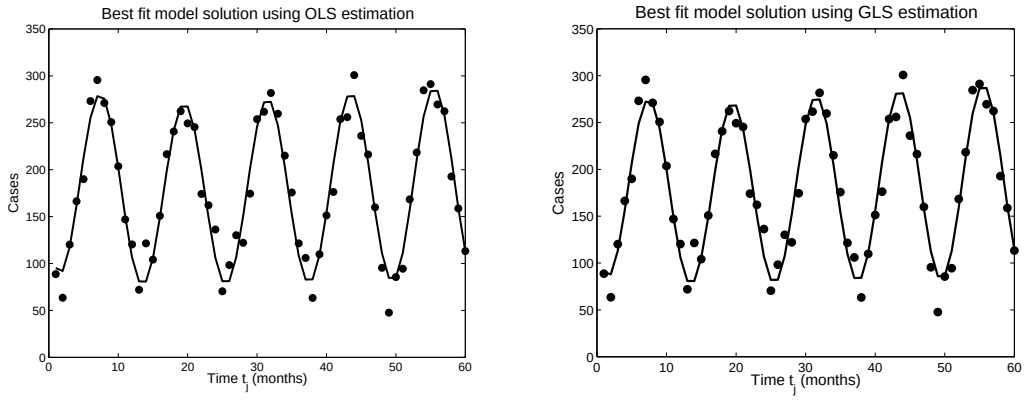
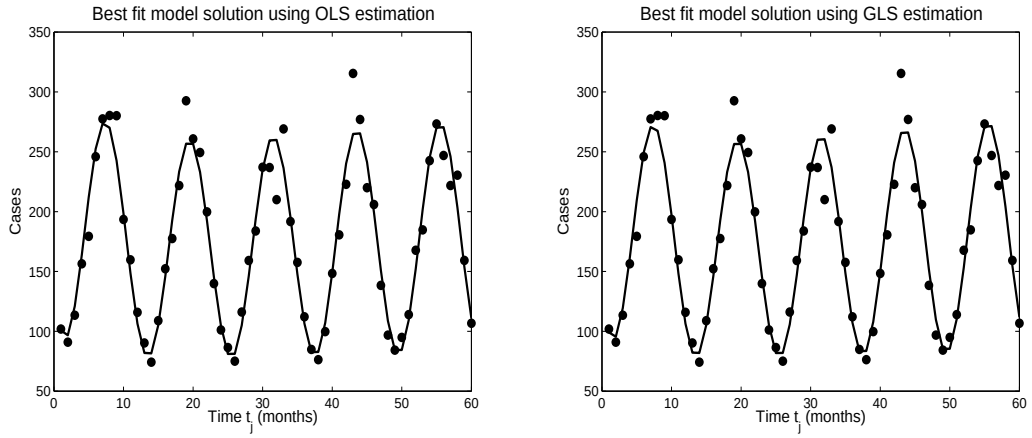


Table 6: Parameter estimates from data with nonconstant variance NCV error.

$\vec{\theta}$	$\vec{\theta}_0$	$\vec{\theta}_{\text{init}}$	$\hat{\theta}_{OLS}$	$SE(\hat{\theta}_{OLS})$	$\hat{\theta}_{GLS}$	$SE(\hat{\theta}_{GLS})$
β	1.5	1.55	1.4876	0.037	1.4923	0.0079
κ_0	$1.4e^{-3}$	$1.3e^{-3}$	$1.4703e^{-3}$	$2.0e^{-4}$	$1.4301e^{-3}$	$7e^{-5}$
κ_1	0.55	0.65	0.54531	0.047	0.54232	0.012
RSS			$1.6692e^4$		$1.676e^4$	

Figure 12: Best fit model solutions to monthly case notifications with nonconstant variance NCV error.



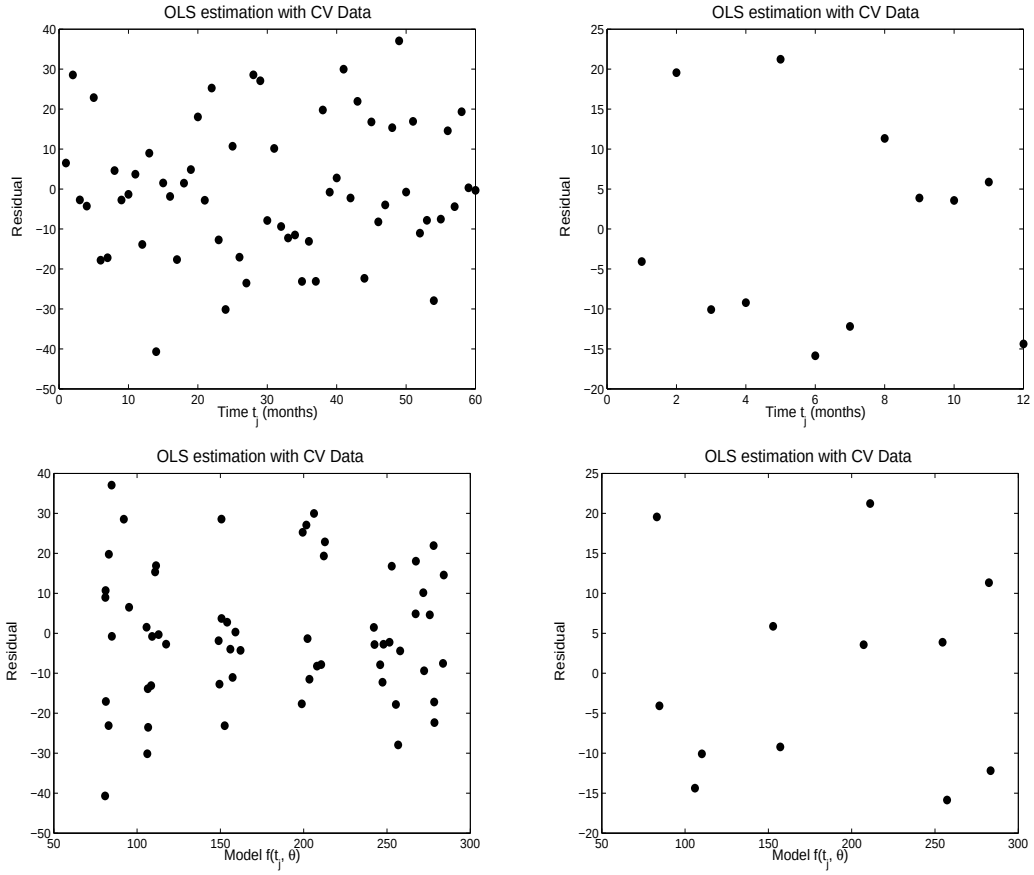
When OLS and GLS estimation are each used with data with nonconstant variance error, the parameters and standard errors in Table 6 are obtained, and the plot of these model solutions over the generated data is given in Figure 12. Again, one estimator does not do a clearly better job over the other in terms of predicting parameter values closer to those used to generate the data. However, again, the standard errors from the GLS estimation are smaller as compared to those of the OLS estimation. From this, it would seem that the GLS estimation would always give 'better' parameter values, or do a better job at producing reliable results. However, we know that in the case of constant variance error, the GLS estimation makes some incorrect assumptions on the data generation and therefore, the standard errors reported there would give a false sense of confidence in the values (indeed they are based on incorrect asymptotic formulae).

5.2.1 Residual Plots

Here we illustrate use of residual plots to investigate whether our assumptions on the errors incurred in observation of data are correct - that is, whether the ϵ_j are $i.i.d.$ for all $j = 1, ..., n$, and also are independent of the observation magnitude. As we have already discussed in

Section 4, if the errors are *i.i.d.* then a plot of the residuals $r_j = y_j - f(t_j, \hat{\theta})$ versus time t_j should show no discernible pattern. Similarly, a plot of the residual r_j as a function of the model values $f(t_j, \hat{\theta})$ should be random if there is no relationship between these two quantities. While use of the OLS estimation tacitly assumes the statistical model (62), and therefore the residual is a realization of the error random variable, this is not true of the GLS estimation. In that case, the assumed statistical model is shown in (62) with ϵ_j *i.i.d.* but the residual r_j are *not i.i.d.* for all $j = 1, ..., n$. Therefore, in the case of GLS we should investigate plots of the the residual/model values, $R_j = \frac{Y_j - f(t_j, \theta_0)}{f(t_j, \theta_0)}$ instead of the residuals.

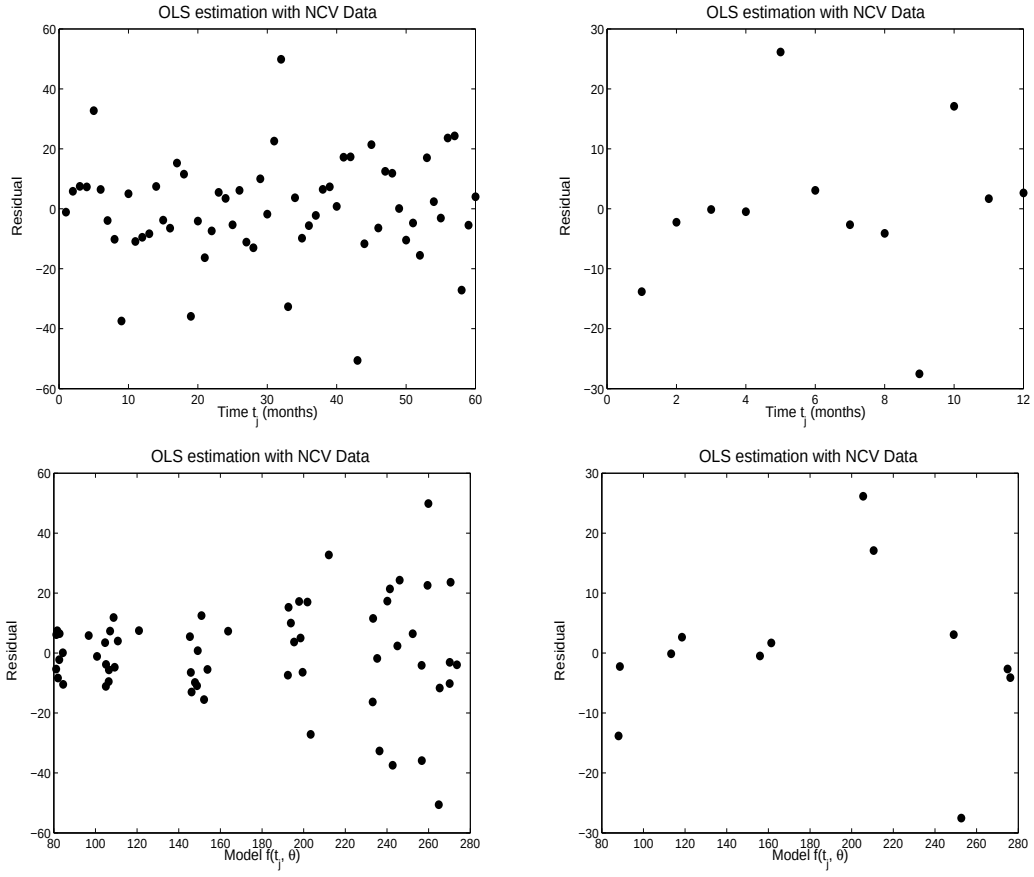
Figure 13: Residual ($r_j = y_j - f(t_j, \hat{\theta})$) plots of the OLS estimation with CV data ($\epsilon_j = Y_j - f(t_j, \vec{\theta_0})$); Left: nontruncated, Right: truncated.



In Figure 13, we see the relationship between the residuals and time, and that between residuals and the model values when the OLS estimation procedure is applied to data which has been generated with constant variance error. In both the top and bottom panels on the left, the full set of $n = 60$ points are used, while on the right hand side, only one year, or $n = 12$ points have been used for the estimation. Both top panels show a random

pattern, so the errors are clearly *i.i.d.* But in the bottom left plot, we observe clustering of residuals around certain model values, although there is no clear pattern in the dependent variable, just in the independent variable, $f(t_j, \hat{\theta})$. However, we recognize that this is due to the seasonality of the data and model, so that at regular repeated time points over many periods, there are going to be repeated values of the model. As evidence of this, we see that when only one period is plotted (the bottom right panel), a random pattern is seen, and we confirm that the errors are not dependent on the model values. Thus, if there are vertical bands on a plot such as this, it can be attributed to certain model values repeating and does not indicate any dependence of the error on the model value. To check, one can simply reduce the number of data points used in the estimation so that there are few or no repeated values.

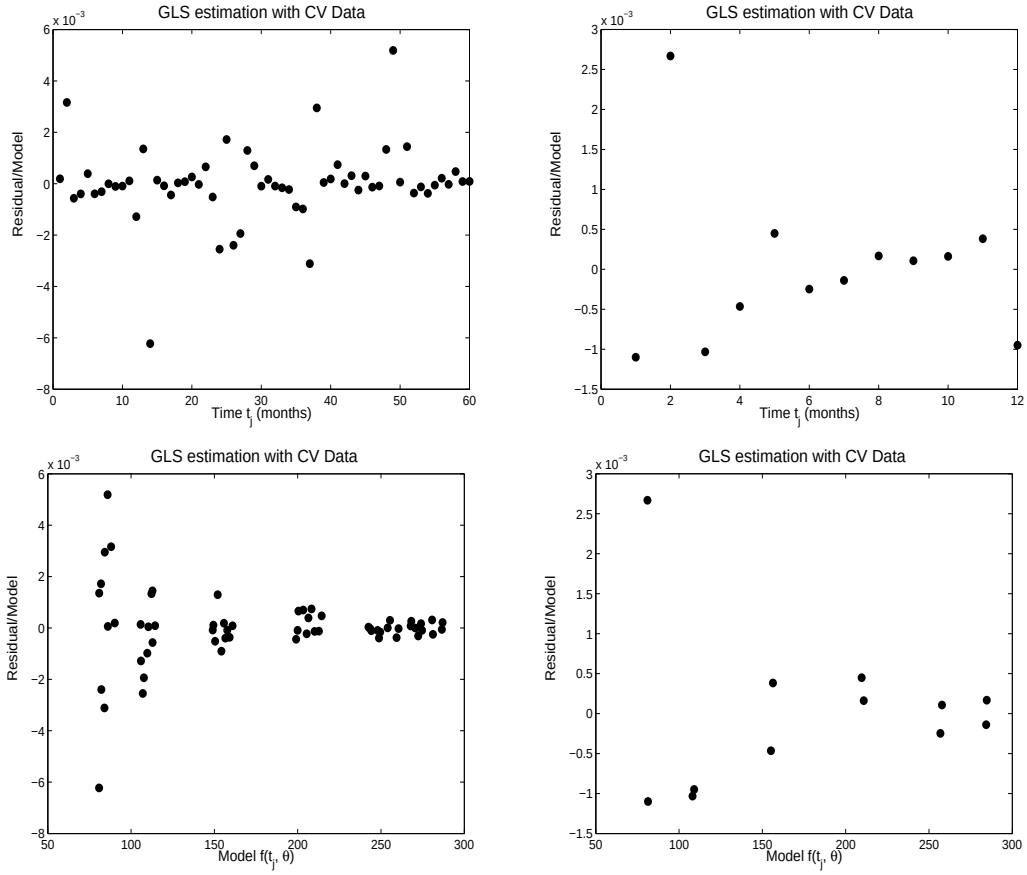
Figure 14: Residual ($r_j = y_j - f(t_j, \hat{\theta})$) plots of the OLS estimation to NCV data ($\epsilon_j = \frac{Y_j - f(t_j, \vec{\theta_0})}{f(t_j, \vec{\theta_0})}$); Left: nontruncated, Right: truncated.



When OLS estimation is carried out with data that has been generated according to the statistical model (63), however, the independence of the error from time is not so clear, as

these graphs (Figure 14) do not show a random pattern. While there is no clear relationship, there is some randomness in the residuals, and the band of residuals are tighter, not homogeneously distributed across the plot as in Figure 13. The dependence of the residuals on model value magnitude (seen in the bottom panels) is apparent as the r_j clearly increase with increasing model values, producing a fan shape. In this case the OLS estimation is used incorrectly, and the residual plots exhibit a clear dependence on model values and do not confirm independence from time.

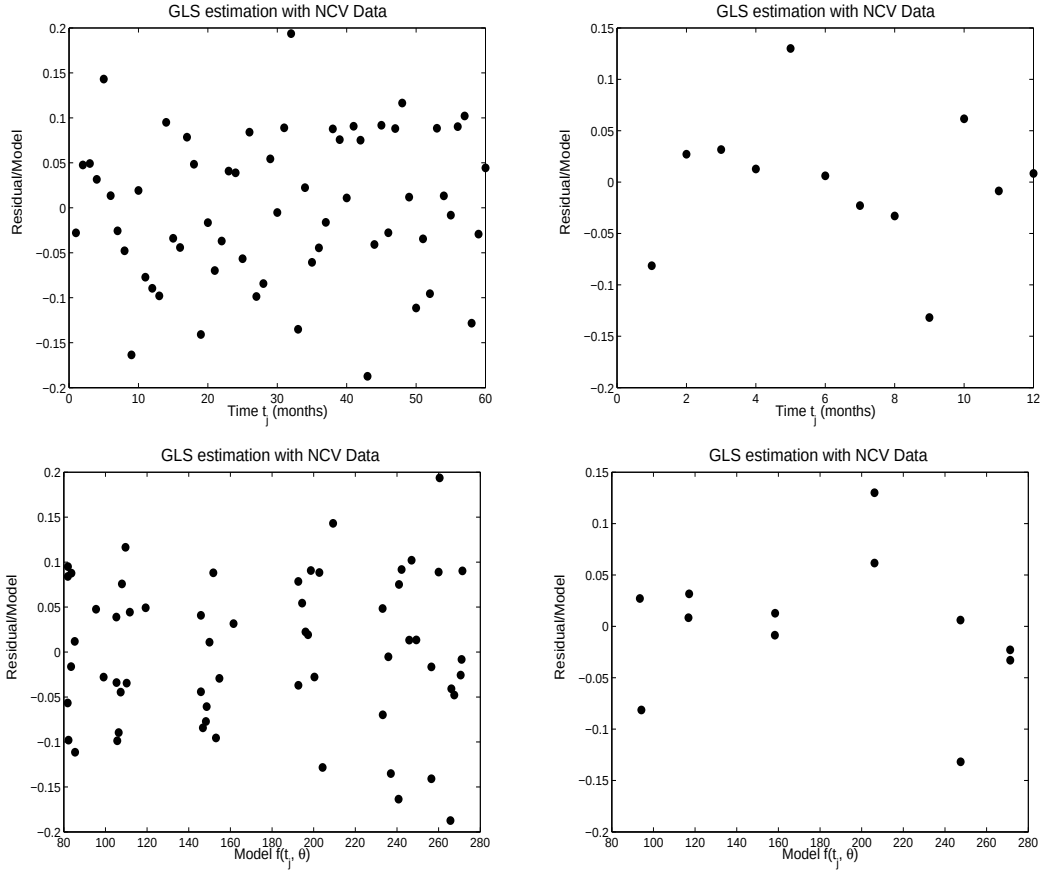
Figure 15: Residual/Model ($\frac{r_j}{f(t_j,\hat{\theta})}$) plots of the GLS estimation to CV data ($\epsilon_j = Y_j - f(t_j, \vec{\theta}_0)$); Left: nontruncated, Right: truncated.



The GLS estimation procedure, however, gave smaller standard errors regardless of the data set used, and therefore, more confidence in the parameter estimates. However, in Figure 15, we see evidence again of the dependence of the residuals on time and model quantities, thus indicating that our assumptions have been incorrect for GLS estimation. In this case, we would have assumed that the errors are proportional to the observations, thus motivating a GLS estimator. If the variance is constant across time and model values, and the GLS estimator is used, we should expect a systematic behavior in the residual plots. Indeed, the

plots in Figure 15 reveal a tight band of points in the $\frac{r_j}{f(t_j,\hat{\theta})}$ versus t_j plots and the reverse fan shape of the plot of the residual/model $\frac{r_j}{f(t_j,\hat{\theta})}$ versus the model values $f(t_j,\hat{\theta})$. This indicates that the relations which give us the parameter estimates and their standard errors no longer hold and we are essentially reporting incorrect values. As we saw in Section 5.2, while the parameter estimates may not necessarily be poor, the reliability provided by the standard errors is incorrect.

Figure 16: Residual/Model ($\frac{r_j}{f(t_j,\hat{\theta})}$) plots of the GLS estimation to NCV data ($\epsilon_j = \frac{Y_j - f(t_j,\vec{\theta}_0)}{f(t_j,\vec{\theta}_0)}$); Left: nontruncated, Right: truncated.



When the GLS estimator is used appropriately, however, the randomness of the error plots suggest reasonability of assumptions, as seen in Figure 16. Here, the error in the data has been generated proportional to the model values, and therefore, not longitudinally constant. So when we plot the ratios $\frac{y_j - f(t_j,\hat{\theta})}{f(t_j,\hat{\theta})}$, we allow for this dependence and see the random patterns we would expect when plotting realizations of a random variable. Again, the vertical bands seen in the bottom left panel indicate repeated model values, as can be

seen by the bottom right panel, where the repetitions have been excluded from the data set.

5.3 Inverse Problem Results: Australian Surveillance Data

Using the iterative weighted least squares procedure described in Section 2.4.2, we carried out inverse problem calculations with the model and observations as outlined in the previous section using Australian IPD data in place of the simulated data. In this case we assumed constant variance noise in the data and hence used WLS , e.g., see (27), for our estimation procedure. Details are given in [32]. We discuss here the case where we used data for the period 2002-2004 (36 months of monthly data $n_1 = 36$, and $n_2 = 6$ of annual vaccinated or unvaccinated cases) and estimated $\vec{\theta} = (\beta, \kappa_0, \kappa_1, \delta)^T$ along with σ_1, σ_2 in a weighted least squares (WLS) functional

$$J_{42}(\vec{\theta}, \sigma_1^2, \sigma_2^2) = \frac{1}{\sigma_1^2} \sum_{j=1}^{36} \left| Y_j^{(1)} - f_j^{(1)} \right|^2 + \frac{9}{\sigma_2^2} \sum_{k=1}^{3} \left\{ \left| Y_k^{(2)} - f_k^{(2)} \right|^2 + \left| Y_k^{(3)} - f_k^{(3)} \right|^2 \right\}. \tag{64}$$

As usual, we assume there exists a 'true' parameter $\vec{\theta}_0$ which generated the data, and our statistical model is then given by

$$Y_j^{(1)} \equiv f^{(1)}(t_j, \vec{\theta}_0) + \epsilon_j^{(1)} \qquad j = 1, ..., 36, \tag{65}$$

$$Y_k^{(2)} \equiv f^{(2)}(t_k, \vec{\theta}_0) + \epsilon_k^{(2)} \qquad k = 1, 2, 3, \tag{66}$$

$$Y_k^{(3)} \equiv f^{(3)}(t_k, \vec{\theta}_0) + \epsilon_k^{(3)} \qquad k = 1, 2, 3. \tag{67}$$

The errors ($\epsilon_j^{(i)}$ in (65) - (67) for $i = 1, 2, 3$) in the above model are assumed to be random variables with means $E[\epsilon_j^{(i)}] = 0$ and constant variances $var(\epsilon_j^{(i)}) = \sigma_{0,i}^2$, where $\sigma_{0,1} = \sigma_1,\ \sigma_{0,2} = \sigma_{0,3} = \sigma_2$ are unknown. Thus we have assumed that the size of the errors committed at each time for a given kind of "measurement" is constant and also does not depend on the magnitude of the measurement itself. We also assume that $\epsilon_j^{(i)}$ are independent and identically distributed (*i.i.d.*) random variables for each fixed i. The observations and the model quantities are related by

- $Y_j^{(1)} \sim f^{(1)}(t_j, \vec{\theta}) = \int_{t_j}^{t_{j+1}} [\kappa(s)E(s) + \delta\kappa(s)E_V(s)]\, ds$ for $j = 1, 2, .., 36$ (monthly cases),
- $Y_k^{(2)} \sim f^{(2)}(t_k, \vec{\theta}) = \int_{t_k}^{t_{k+1}} \kappa(s)E(s)ds$ for $k = 1, 2, 3$ (yearly unvaccinated cases),
- $Y_k^{(3)} \sim f^{(3)}(t_k, \vec{\theta}) = \int_{t_k}^{t_{k+1}} \delta\kappa(s)E_V(s)ds$ for $k = 1, 2, 3$ (yearly vaccinated cases).

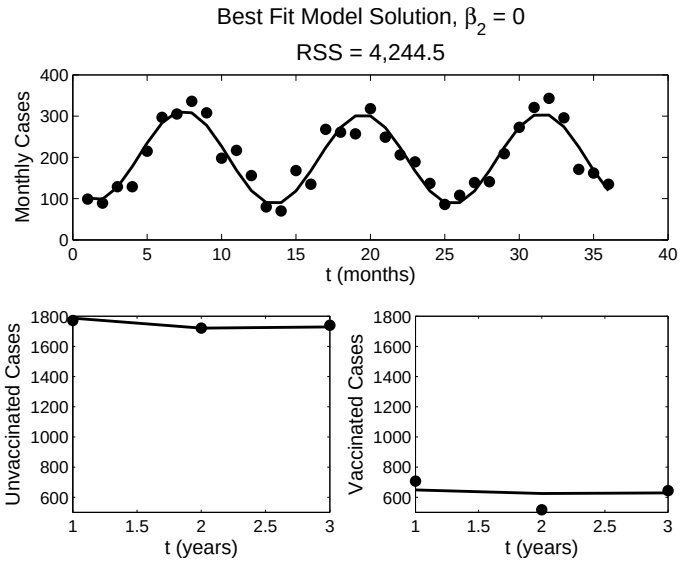
The data fits in Figure 17 reveal that the model solution with the parameters shown in Table 7 fits the Australian surveillance data from 2002-2004, with the top panel showing the fit to the monthly case notification data, the bottom left panel the unvaccinated cases reported

Table 7: Model calibration to Australian IPD data from 2002-2004; estimation of $\hat{\psi} = (\hat{\theta}, \hat{\sigma}_1, \hat{\sigma}_2)^T = (\hat{\beta}, \hat{\kappa}_0, \hat{\kappa}_1, \hat{\delta}, \hat{\sigma}_1, \hat{\sigma}_2)^T$.

ψ	$\hat{\psi}$	$SE(\hat{\theta})$
β	1.52175	0.02
κ_0	$1.3656e^{-3}$	$1.3e^{-4}$
κ_1	0.56444	0.04
δ	0.7197	0.06
σ_1	28.924	
σ_2	86.386	

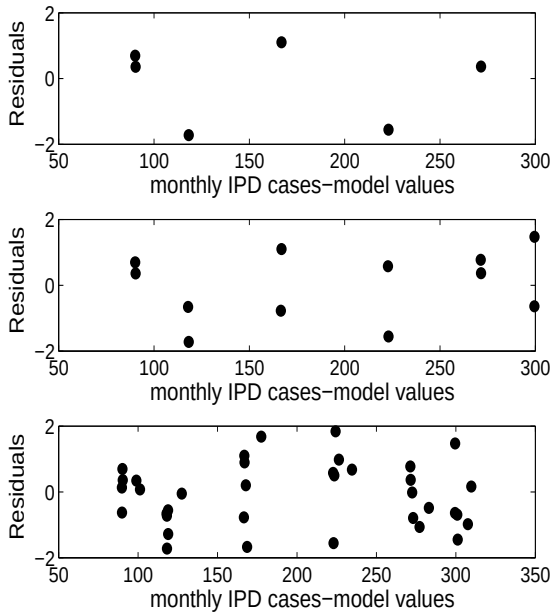
annually, and the bottom right the annual vaccinated cases. The model solution and data agree well, and parameter values are on the scale of our initial guesses, although their values differ slightly to minimize the cost function in the functional (64). Further, the standard errors are relatively small and suggests that the estimates obtained here are reliable.

Figure 17: Best fit solution to Australian IPD data with parameters shown in Table 7. Top panel: monthly cases; bottom left panel: annual unvaccinated cases; bottom right panel: annual vaccinated cases.



To test the assumptions of the statistical model that we have chosen to represent our data, we plotted the residuals between the model and observations as a function of the model, that is, $r_j = y_j^{(1)} - f^{(1)}(t_j, \hat{\theta})$ vs. the model values $f^{(1)}(t_j, \hat{\theta})$ (Figure 18). The lack of a clear relationship between these two quantities suggests that our assumptions are reasonable and the residuals of each observation do not depend on the model values. However, we see six groups of points, which can be explained by the oscillatory pattern of the infections. In the top panel we have plotted just one half of the period of the infection rate and see a completely random pattern, indicating no relationship among these quantities. When we extend this time period for another half of a period, thus plotting an entire period in the middle panel, we see that there are two points in each group of points. Thus, the pattern observed is driven by the seasonality of the infections and not by any incorrect assumptions. On the contrary, only a pattern in the dependent variable (the residuals) would suggest that incorrect assumptions have been made. This analysis suggests that it is reasonable to assume constant variance among observations of the same type, providing support for the statistical model underlying the parameter estimation procedure.

Figure 18: Residuals as a function of model values. Top panel is over the period January 2003 through June 2003, middle panel is for January 2003 through December 2003, and bottom panel is for all three years.



6 Sensitivity Functions

The sensitivity matrices $\chi = F_{\vec{\theta}}$ introduced in Section 3 to define covariances for sampling distributions and associated standard errors are actually well known in the applied mathematics and engineering literature, where they arise in routine sensitivity analysis.

In actuality, *sensitivity analysis* is an ensemble of techniques [30] that can provide information on parameter dependent model behavior, yielding a much better understanding of the underlying mathematical model with a resulting marked improvement in the estimation results obtained using the models in simulations and inverse problems. Traditionally, sensitivity analysis referred to a procedure used in simulation studies (direct problems) where one evaluated the effects of parameter variations on the time course of model outputs and identified the parameters or the initial conditions to which the model is most/least sensitive. In recent years however, investigators' attention has also recently turned to the sensitivity of the solutions to inverse problems with respect to data, in a quest for optimal selection of data measurements in *experimental design*. As part of model validation and verification, one typically needs to estimate model parameters from data measurements, and a related question of paramount interest is related to sampling; specifically, at which time points the measurements are most informative in the estimation of a given parameter. Due to the fact that in practice the components of the parameter estimates are often correlated, *traditional sensitivity functions* (TSF) used alone are not very efficient in answering this question because TSF do not take into account how model output variations affect parameter estimates in inverse problems. Investigators [11, 34] recently proposed a new class of sensitivity functions, called *generalized sensitivity functions* (GSF), which provide information on the relevance of measurements of output variables of a system for the identification of specific parameters. For a given set of time observations, Thomaseth and Cobelli use theoretical information criteria (the Fisher information matrix) to establish a relationship between the monotonicity of the GSF curves with respect to the model parameters and the information content of these observations. Here our interest is in how to use this information content tool along with TSF to improve data collection for estimation of parameters in inverse problems. It is, of course, intuitive that sampling more data points from the region indicated by the GSF to be the "most informative" with respect to a given parameter would result in more information about that parameter, and therefore provide more accurate estimates for it.

To define and discuss these sensitivity functions we consider the general mathematical model (1) with N-vector solutions $\vec{x}$ depending on p-vector parameters $\vec{\theta}$.

6.1 Traditional Sensitivity Functions

Traditional sensitivity functions (TSF) are classical sensitivity functions used in mathematical modeling to investigate variations in the output of a model resulting from variations in the parameters and the initial conditions.

In order to quantify the variation in the state variable $\vec{x}(t)$ with respect to changes in the parameter $\vec{\theta}$ and the initial condition $\vec{x}_0$, we are naturally led to consider *traditional*

sensitivity functions (TSF) as defined by the derivatives

$$\vec{s}_{\theta_k}(t) = \frac{\partial \vec{x}}{\partial \theta_k}(t), \qquad k = 1, \ldots, p, \tag{68}$$

and

$$\vec{r}_{x_{0l}}(t) = \frac{\partial \vec{x}}{\partial x_{0l}}(t), \qquad l = 1, \ldots, N, \tag{69}$$

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

In practice, the model under investigation often is sufficiently simple to allow one to compute analytically the sensitivity functions (68) and (69). This is precisely the case (see (55)) with the logistic growth population example of (54) to be discussed below. However, when one deals with a more complex model, as with the epidemiological example of Section 5, it is often preferable to consider these sensitivity functions separately for clarity purposes.

The sensitivity functions are local in nature because they are defined by partial derivatives which have a *local* character. Thus sensitivity and insensitivity (i.e., $\vec{s}_{\theta_k} = \partial \vec{x}/\partial \theta_k$ very close to zero) depend on the time interval, the state values $\vec{x}$, and the values of $\vec{\theta}$ for which they are considered. For example in a certain time subinterval we might find that $\vec{s}_{\theta_k}$ is small so that the state variable $\vec{x}$ is *insensitive* to the parameter θ_k on that particular interval. The same function $\vec{s}_{\theta_k}$ can take large values on a different subinterval, indicating that the state variable $\vec{x}$ is *quite sensitive* to the parameter θ_k on the latter interval. From the sensitivity analysis theory for dynamical systems, one finds (e.g., see (48)) that $s = (\vec{s}_{\theta_1}, \ldots, \vec{s}_{\theta_p})$ is an $N \times p$ vector function that satisfies the matrix ODE system

$$\begin{aligned} \dot{s}(t) &= \vec{g}_{\vec{x}}(t, \vec{x}(t), \vec{\theta}) s(t) + \vec{g}_{\vec{\theta}}(t, \vec{x}(t), \vec{\theta}), \\ s(t_0) &= 0_{N \times p}, \end{aligned} \tag{70}$$

so that the dependence of s on $(t, \vec{x}(t))$ as well as $\vec{\theta}$ is readily apparent. Here we have used $\vec{g}_{\vec{x}} = \partial \vec{g}/\partial \vec{x}$ and $\vec{g}_{\vec{\theta}} = \partial \vec{g}/\partial \vec{\theta}$ to denote the derivatives of $\vec{g}$ with respect to $\vec{x}$ and $\vec{\theta}$, respectively.

The sensitivity functions with respect to the components of the initial condition $\vec{x}_0$ define an $N \times N$ vector function $r = (\vec{r}_{x_{01}}, \ldots, \vec{r}_{x_{0N}})$, which satisfies the matrix system

$$\begin{aligned} \dot{r}(t) &= \vec{g}_{\vec{x}}(t, \vec{x}(t), \vec{\theta}) r(t), \\ r(t_0) &= I_{N \times N}. \end{aligned} \tag{71}$$

Equations (70) and (71) can be used in conjunction with equation (1) to numerically compute the sensitivities s and r for general cases when the function $\vec{g}$ is sufficiently complicated to prohibit an analytical solution.

In many cases the parameters have different units and the state variables may have varying orders of magnitude, and thus in practice it is sometimes more convenient to work with the scaled versions of the TSF, referred to as *relative sensitivity functions* (RSF). However, here we will focus solely on the non-scaled sensitivities, i.e., TSF.

6.2 Generalized Sensitivity Functions

Recently generalized sensitivity functions were proposed by Thomaseth and Cobelli [34] as a new tool in identification studies to analyze the distribution of the information content (with respect to the model parameters) of the output variables of a system for a given set of observations. These are formulated in the context of an OLS inverse problem framework in [11, 34].

We consider here a scalar observation model with discrete time measurements. When $m = 1$ and $\mathcal{C}$ is a $1 \times N$ array in (4)), the *generalized sensitivity functions* (GSF) are defined as

$$\mathbf{gs}(t_l) = \sum_{i=1}^{l} \frac{1}{\sigma^2(t_i)} [F^{-1} \times \nabla_{\vec{\theta}} f(t_i, \vec{\theta}_0)] \bullet \nabla_{\vec{\theta}} f(t_i, \vec{\theta}_0), \tag{72}$$

where $\{t_l\}$, $l = 1, \ldots, n$ are the times when the measurements are taken,

$$F = \sum_{j=1}^{n} \frac{1}{\sigma^2(t_j)} \nabla_{\vec{\theta}} f(t_j, \vec{\theta}_0) \nabla_{\vec{\theta}} f(t_j, \vec{\theta}_0)^T \tag{73}$$

is the corresponding $p \times p$ *Fisher information matrix* and $\sigma^2(t_j)$ is the observation time dependent variance. The symbol "$\bullet$" represents element-by-element vector multiplication (for motivation and details which lead to the definition above, the interested reader may consult [11, 34]). The Fisher information matrix measures the information content of the data corresponding to the model parameters. In (72) we see that this information is contained in the GSF, making them appropriate tools to indicate the relevance of the measurements to estimation of a parameter in inverse problems.

We observe that the generalized sensitivity functions (72) are vector-valued functions with the same dimension as $\vec{\theta}$. The k-th component gs_k of the vector function **gs** represents the generalized sensitivity function with respect to θ_k. The GSF in (72) are defined only at the discrete time points $\{t_j, j = 1, \ldots, n\}$ and they are cumulative functions involving at time t_l only the contributions of those measurements up to and including t_l; thus gs_k calculates the influence of measurements up to t_l on the parameter estimate for θ_k.

It is readily seen from the definition that all the components of **gs** are one at the final time point t_n, i.e., $\mathbf{gs}(t_n) = \mathbf{1}$. If one defines $\mathbf{gs}(t) = \mathbf{0}$ for $t < t_1$ (naturally, **gs** is zero when no measurements are collected), then each component of **gs** transitions (not necessarily monotonically) from zero to one. As developed in [11, 34], the time subinterval during which the change in gs_k has the *sharpest increase* corresponds to the *observations which provide the most information in the estimation of* θ_k. That is, regions of sharp increases in gs_k indicate a high concentration of information in the data about θ_k. Thus, the utility of these functions in design of experiments is rather obvious.

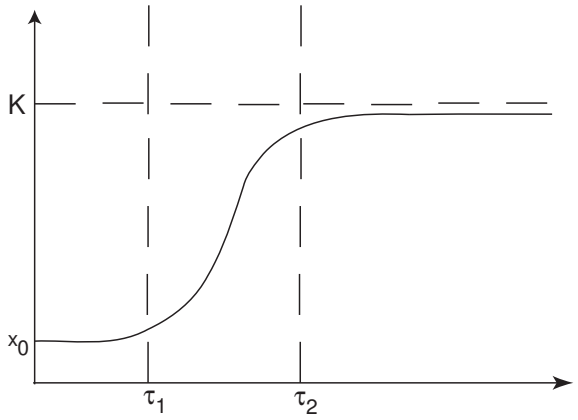
The numerical implementation of the generalized sensitivity functions (72) is straightforward, since the gradient of f with respect to $\vec{\theta}$ (or $\vec{x}_0$) is simply the Jacobian of $\vec{x}$ with respect to $\vec{\theta}$ (or $\vec{x}_0$) multiplied by the observation operator $\mathcal{C}$. These Jacobian matrices can be obtained by numerically solving the sensitivity ODE system (70) or (71) coupled with the

system (1). One would need to use this approach to compute the GSF for the epidemiological model of Section 5. For the the logistic model used below to illustrate ideas, the solution of equation (54) given by (55) is sufficiently simple to permit an analytical representation of the Jacobians.

6.3 TSF and GSF for the Logistic Equation

The Verhulst-Pearl logistic equation (54) is a relatively simple example with easily determined dynamics that is useful in demonstrating the utility of the traditional sensitivity functions as well as the generalized sensitivity functions in inverse problems (see [2, 4] for more discussions on TSF and GSF for this example). Unless data are sampled from regions with changing dynamics, it is possible that some of the parameters will be difficult to estimate. Moreover, the parameters that are obtainable may have high standard errors as a result of introducing redundancy in the sampling region (this is illustrated in [2]). In order to investigate sensitivity for the logistic growth example, we will examine varying behavior in the model depending on the region from which t_j is sampled. We consider points τ_1 and τ_2, as depicted in Figure 19, partitioning the logistic solution curve into three distinct regions: $0 < t_j < \tau_1$, $\tau_1 < t_j < \tau_2$, and $\tau_2 < t_j < T$, with T sufficiently large for our solution to be near its asymptote $x = K$. Based on the changing dynamics of the curve in Figure 19, we expect differences in the ability to estimate parameters depending on the region in which the solution is observed.

Figure 19: Regions with different growth in the Verhulst-Pearl solution curve.



We consider the logistic model with true parameters $\vec{\theta}_0 = (17.5, 0.7, 0.1)$. We analyze the TSF corresponding to each parameter in the initial region of the curve, where the solution approaches x_0 as $t \to 0$. When we consider the initial region of the curve, where $0 < t_j < \tau_1$ for $j = 1, \ldots, n$, we have

$$\frac{\partial x(t_j)}{\partial K} \approx 0, \qquad \frac{\partial x(t_j)}{\partial r} \approx 0, \qquad \frac{\partial x(t_j)}{\partial x_0} \approx 1;$$

this follows from considering the limits of the readily computed analytical sensitivity functions as $t \to 0$. Based on these analytical findings, which indicate low sensitivities with respect to K and r, we expect to have little ability to determine these parameters when we sample data from $[0, \tau_1]$; however we should be able to estimate x_0. This is confirmed by the computational examples in [2, 4].

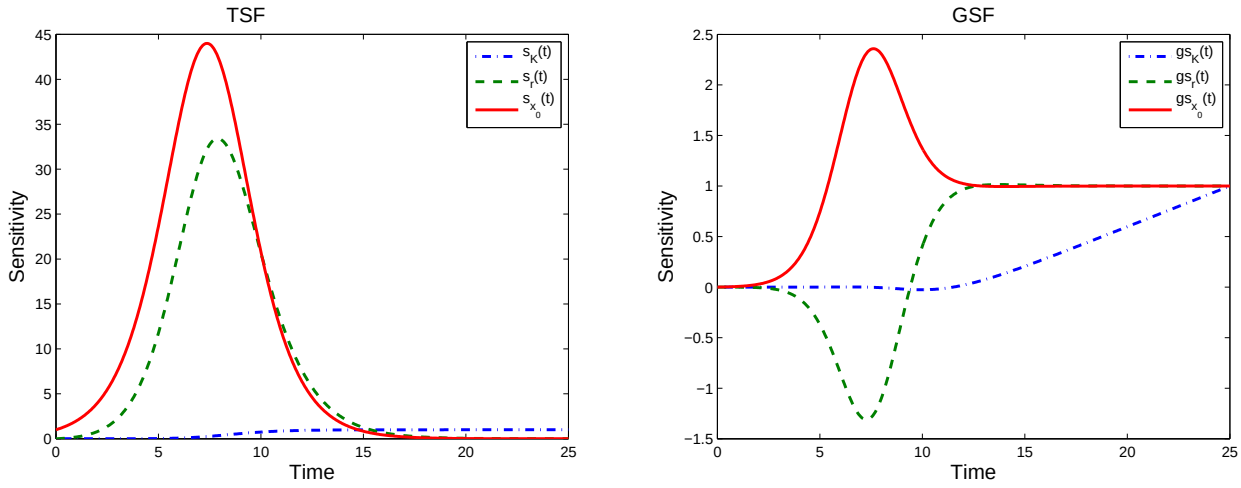
We next consider the region of the curve which is near the asymptote at $x = K$, in this case for $\tau_2 < t_j < T$, $j = 1, \ldots, n$. Here we find that by considering the limits as $t \to \infty$, we have the approximations

$$\frac{\partial x(t_j)}{\partial K} \approx 1, \qquad \frac{\partial x(t_j)}{\partial r} \approx 0, \qquad \frac{\partial x(t_j)}{\partial x_0} \approx 0.$$

Based on these approximations, we expect to be able to estimate K well when we sample data from $[\tau_2, T]$. However, using data only from this region, we do not expect to be able to estimate very well either x_0 or r. Again these expectations are readily confirmed by the inverse problem calculations presented in [2, 4].

Finally, we consider the part of the solution curve where $\tau_1 < t_j < \tau_2$ for $j = 1, \ldots, n$ and where it has nontrivially changing dynamics. We find that the partial derivative values differ greatly from the values in regions $[0, \tau_1]$ and $[\tau_2, T]$. When $[\tau_1, \tau_2]$ is included in the sampling region we expect to recover good estimates for all three parameters (expectations that are met in [2, 4]).

Figure 20: (a) TSF and (b) GSF corresponding to each parameter for the logistic curve with $\vec{\theta_0} = (17.5, 0.7, 0.1)$.



Our analytical observations are fully consistent with information contained in the graphs of the TSF illustrated in Figure 20(a) for $T = 25$. We note that the curve s_K slowly increases

with time and it appears that the solution is insensitive to K until around the flex point of the logistic curve, which occurs shortly after $t = 7$ in this case. The sensitivities s_K and s_r both are close to zero when t is near the origin, and hence we deduce that both K and r will be difficult or impossible to obtain using data in that region. Also, we observe that s_{x_0} and s_r are nearly zero in [15,25], which suggests that we will be unable to estimate x_0 or r using observations in that region.

We numerically computed the GSF using equation (72) with $\sigma = 1$ and the true value parameters $\vec{\theta}_0 = (17.5, 0.7, 0.1)$. The plots of these functions are shown in Figure 20(b) where one can observe obvious regions of steep increase in each curve. For the curves $gs_{x_0}(t)$, $gs_r(t)$ and $gs_K(t)$, we find by visual inspection that these regions are approximately $[4.5, 7.5]$, $[7, 11]$ and $[12, 25]$, respectively. By the generalized sensitivity theory, if we increase the number of data points sampled in one of these regions, the estimation of the corresponding parameter is expected to improve. This is precisely what happens in the computational examples found in [2, 4].

While the general algorithms are still under development, the following scenario involving TSF and GSF in design of experiments for data to be used in OLS and GLS formulations are envisioned:

1. One proposes a mechanism, interaction, etc., as represented by a term or terms (such as a nonlinearity, probability distribution, etc.) in a model (ODE, PDE, etc.). One then uses methodology based on the TSF, GSF and the Fisher Information Matrix (FIM) calculations to suggest design of experiments to collect data (duration of experiment, sampling sizes, frequency in time, space, age/size class, etc.) to be used in inverse problem/parameter estimation techniques to investigate the mechanistic based terms.

2. One then designs and carries out the experiments resulting from 1. with guidance in data collection (variables required to be observed, sampling frequency, measurement accuracy needed, etc) being provided for each class of models to be used with the data; questions and models usually will be driven by mechanism based formulation.

3. Finally, one can carry out post experimental modeling analysis (parameter estimation and inverse problems with both OLS and GLS , statistical analysis of variance in data and model fits with residual plots, hypothesis testing and model comparison as described in the next several sections, Kullback-Leibler distance based and information content based model selection techniques such as AIC and recent generalizations [16, 17] and improvements, etc.) to provide a modeling framework and methodology for future investigations of the type proposed here. In the post analysis one can also carry out verification and validation type studies as well as testing predictive capabilities. This can be done in part by comparing the models with data that was not used in the inverse problems for estimation of parameters.

7 Statistically Based Model Comparison Techniques

In previous sections we have discussed techniques (e.g., residual plots) for investigating correctness of the assumed *statistical model* underlying the estimation (OLS or GLS) procedures used in inverse problems. To this point we have not discussed correctness issues related to choice of the *mathematical model*. However there are a number of ways in which questions related to the mathematical model may arise. In general, modeling studies [7, 8] can raise questions as to whether a mathematical model can be improved by *more detail* and/or *further refinement?* For example, one might ask whether one can improve the mathematical model by assuming more *detail* in a given mechanism (constant rate vs. time or spatially dependent rate – e.g., see [1] for questions related to time dependent mortality rates during sub-lethal damage in insect populations exposed to various levels of pesticides). Or one might question whether an *additional mechanism* in the model might produce a better fit to data–see [5, 6, 7] for *diffusion alone* or *diffusion plus convection* in cat brain transport in grey vs. white matter considerations.

Before continuing an important point must be made: In model comparison results outlined below, there are really <u>two models</u> being compared: the *mathematical model* and the *statistical model*. If one embeds the mathematical model in the *wrong statistical model* (for example, assumes constant variance when this really isn't true), then the mathematical model comparison results using the techniques presented here will be *invalid* (e.g., *worthless*). An important remark in all this is that you must have the mathematical model you want to simplify or improve (e.g., test whether $\mathcal{V} = 0$ or not in the example below) embedded in the *correct statistical model* (determined in large part by the observation process), so that the comparison really is *only with regard to the mathematical model*.

To provide specific motivation, we illustrate the formulation of hypothesis testing by considering a mathematical model for a diffusion-convection process. This model was proposed for use with experiments designed to study substance (labeled sucrose) transport in cat brains, which are heterogeneous, containing grey and white matter [7]. In general, the transport of substance in cat's brains can be described by a PDE describing *change in time and space*. This convection/diffusion model, which is widely discussed in the applied mathematics and engineering literature, has the form

$$\frac{\partial u}{\partial t} + \mathcal{V}\frac{\partial u}{\partial x} = \mathcal{D}\frac{\partial^2 u}{\partial x^2}. \tag{74}$$

Here, the parameter $\vec{\theta} = (\mathcal{D}, \mathcal{V})$, which belongs to some admissible parameter set Θ, denotes the diffusion coefficient $\mathcal{D}$ and the bulk velocity $\mathcal{V}$ of the fluid, respectively. Our problem: test whether the parameter $\mathcal{V}$ plays a significant role in the mathematical model. That is, if the model (74) represents a diffusion-convection process, we seek to determine whether diffusion alone or diffusion plus convection best describes transport phenomena represented in cat brain data sets $\{y_{ij}\}$ for $\{u(t_i, x_j; \vec{\theta})\}$, the concentration of labeled sucrose at times $\{t_i\}$ and location $\{x_j\}$. We then may take $H_0 : \mathcal{V} = 0$ and the alternative $H_A : \mathcal{V} \neq 0$.

Consequently, the restricted parameter set $\Theta_H \subset \Theta$ defined by

$$\Theta_H = \{\vec{\theta} \in \Theta : \mathcal{V} = 0\}$$

will be important. To carry out these determinations, we will need some model comparison tests of analysis of variance (ANOVA) type from statistics involving residual sum of squares (RSS).

7.1 RSS Based Statistical Tests

In general, we assume an inverse problem with mathematical model $f(t, \vec{\theta})$ and n observations $\vec{Y} = \{Y_j\}_{j=1}^n$. We define an OLS performance criterion

$$J_n(\vec{\theta}) = J_n(\vec{Y}, \vec{\theta}) = \frac{1}{n} \sum_{j=1}^{n} [Y_j - f(t_j, \vec{\theta})]^2,$$

where our statistical model again has the form

$$Y_j = f(t_j, \vec{\theta}_0) + \epsilon_j, \quad j = 1, \ldots, n,$$

with $\{\epsilon_j\}_{j=1}^n$ independent and identically distributed, $E(\epsilon_j) = 0$ and constant variance $var(\epsilon_j) = \sigma^2$. As usual $\vec{\theta}_0$ is the "true" value of $\vec{\theta}$ which we assume to exist. As noted above, we use Θ to represent the set of all the admissible parameters $\vec{\theta}$ and assume that Θ is a compact subset of Euclidean space of R^p with $\vec{\theta}_0 \in \Theta$.

Let $\theta^n(\vec{Y}) = \theta^n_{OLS}(\vec{Y})$ be the OLS *estimator* using J_n with corresponding *estimate* $\hat{\theta}^n = \theta^n_{OLS}(\vec{y})$ for a realization $\vec{y} = \{y_j\}$. That is,

$$\theta^n(\vec{Y}) = \arg\min_{\vec{\theta} \in \Theta} J_n(\vec{Y}, \vec{\theta}) \quad \text{and} \quad \hat{\theta}^n = \arg\min_{\vec{\theta} \in \Theta} J_n(\vec{y}, \vec{\theta}).$$

Remarks: In most calculations, one actually uses an approximation f^N to f, often a numerical solution to the ODE or PDE for modeling the dynamical system. Here we tacitly assume f^N will converge to f as the approximation improves. There are also questions related to approximations of the set Θ when it is infinite dimensional (e.g., in the case of function space parameters such as time or spatially dependent parameters) by finite dimensional discretizations Θ^M. For extensive discussions related to these questions, see [8] as well as [6] where related assumptions on convergences $f^N \to f$ and $\Theta^M \to \Theta$ are given. We shall ignore these issues here, keeping in mind that these approximations will also be of importance in the methodology discussed below in most practical uses.

In many instances, including the motivating example given above, one is interested in using data to address the question whether or not the "true" parameter $\vec{\theta}_0$ can be found in a subset $\Theta_H \subset \Theta$ which we assume for discussions here is defined by

$$\Theta_H = \{\vec{\theta} \in \Theta | H\vec{\theta} = c\} \tag{75}$$

where H is an $r \times p$ matrix of full rank, and c is a known constant.

In this case we want to test the *null hypothesis* H_0: $\vec{\theta}_0 \in \Theta_H$.

Define then

$$\theta_H^n(\vec{Y}) = \arg\min_{\vec{\theta}\in\Theta_H} J_n(\vec{Y},\vec{\theta}) \quad \text{and} \quad \hat{\theta}_H^n = \arg\min_{\vec{\theta}\in\Theta_H} J_n(\vec{y},\vec{\theta})$$

and observe that $J_n(\vec{Y},\hat{\theta}_H^n) \geq J_n(\vec{Y},\hat{\theta}^n)$. We define the related non-negative test statistics and their realizations, respectively, by

$$T_n(\vec{Y}) = n(J_n(\vec{Y},\theta_H^n) - J_n(\vec{Y},\theta^n)) \quad \text{and} \quad \hat{T}_n = T_n(\vec{y}) = n(J_n(\vec{y},\hat{\theta}_H^n) - J_n(\vec{y},\hat{\theta}^n)).$$

One can establish asymptotic convergence results for the test statistics $T_n(\vec{Y})$, as given in detail in [6]. These results can, in turn, be used to establish a fundamental result about much more useful statistics for model comparison. We define these statistics by

$$U_n(\vec{Y}) = \frac{T_n(\vec{Y})}{J_n(\vec{Y},\theta_n)}, \tag{76}$$

with corresponding realizations $\hat{U}_n = U_n(\vec{y})$. We then have the asymptotic result that is the basis of our ANOVA–type tests.

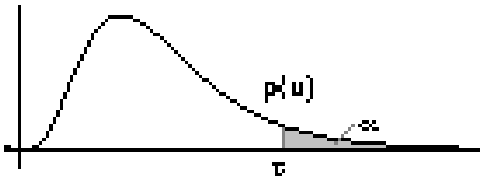
Under reasonable assumptions (very similar to those required in the asymptotic sampling distribution theory discussed in previous sections–see [6, 8, 31]) involving regularity and the manner in which samples are taken, one can prove [6]:

(a) We have the estimator convergence $\theta^n \longrightarrow \vec{\theta}_0$ as $n \to \infty$ with probability one;

(b) If H_0 is true, $U_n \xrightarrow{\mathcal{D}} U(r)$ as $n \to \infty$ where $U \sim \chi^2(r)$, a χ^2 distribution with r degrees of freedom.

An example of the χ^2 density is depicted in Figure 21 where the density for $\chi^2(4)$ (χ^2 with $r = 4$ degrees of freedom) is graphed.

Figure 21: Example of $U \sim \chi^2(4)$ density



In this figure two parameters (τ, α) of interest are shown. For a given value τ, the value α is simply the probability that the random variable U will take on a value greater than α. That is, $Prob\{U > \tau\} = \alpha$ where in hypothesis testing, α is the *significance level* and τ is the *threshold*.

We wish to use this distribution to test the null hypothesis, H_0, where we approximate by $U_n \sim \chi^2(r)$. If the test statistic, $\hat{U}_n > \tau$, then we *reject* H_0 as false with confidence level $(1-\alpha)100\%$. Otherwise, we *do not reject* H_0 as true. For cat brain problem, we use a $\chi^2(1)$ table, which can be found in any elementary statistics text or online and is given here for illustrative purposes.

Table 8: $\chi^2(1)$

α	τ	confidence
.25	1.32	75%
.1	2.71	90%
.05	3.84	95%
.01	6.63	99%
.001	10.83	99.9%

7.1.1 p-values

The minimum value α^* of α at which H_0 can be rejected is called the *p-value*. Thus, the smaller the p-value, the stronger the evidence in the data in support of rejecting the null hypothesis and including the term in the model, i.e., the more likely the term should be in the model. We implement this as follows: Once we compute $\hat{U}_n = \bar{\tau}$, then $p = \alpha^*$ is the value that corresponds to $\bar{\tau}$ on a χ^2 graph and so, we reject the null hypothesis at any confidence level, c, such that $c < 1 - \alpha^*$. For example, if for a computed $\bar{\tau}$ we find $p = \alpha^* = .0182$, then we would reject H_0 at confidence level $(1 - \alpha^*)100\% = 98.18\%$ or lower. For more information, the reader can consult ANOVA discussions in any good statistics book.

7.1.2 Alternative statement

To test the null hypothesis H_0, we choose a significance level α and use χ^2 tables to obtain the corresponding threshold $\tau = \tau(\alpha)$ so that $P(\chi^2(r) > \tau) = \alpha$. We next compute $\hat{U}_n = \overline{\tau}$ and compare it to τ. If $\hat{U}_n > \tau$, then we *reject* H_0 as false; otherwise, we do not reject the null hypothesis H_0.

7.2 Revisiting the cat-brain problem

We summarize use of the above model comparison techniques outlined above by returning to the cat brain example discussed in detail in [7, 8]. There were *3 sets of experimental data* examined, under the null-hypothesis $H_0 : \mathcal{V} = 0$.

For the *Data Set 1*, we found after carrying out the inverse problems over Θ and Θ_H, respectively,

$$J_n(\hat{\theta}^n) = 106.15 \quad \text{and} \quad J_n(\hat{\theta}^n_H) = 180.1,$$

which gives us that $\hat{U}_n = 5.579$ (noting that $n = 8 \neq \infty$), for which $p = \alpha^* = .0182$. Thus, we reject H_0 in this case at *any* confidence level less than 98.18%. Thus, we should *reject* that $\mathcal{V} = 0$, which suggests convection is important in describing this data set.

For *Data Set 2*, we found

$$J_n(\hat{\theta}^n) = 14.68 \quad \text{and} \quad J_n(\hat{\theta}^n_H) = 15.35,$$

and thus, in this case, we have $\hat{U}_n = .365$, which implies we *do not reject* H_0 with *high degrees of confidence* (p-value very high). This suggests $\mathcal{V} = 0$, which is completely opposite to the findings for Data Set 1.

For the final set (*Data Set 3*) we found

$$J_n(\hat{\theta}^n) = 7.8 \quad \text{and} \quad J_n(\hat{\theta}^n_H) = 146.71,$$

which yields in this case, $\hat{U}_n = 15.28$. This, as in the case of the first data set, suggests (with $p < .001$) that $\mathcal{V} \neq 0$ is important in modeling the data.

The difference in conclusions between the first and last sets and that of the second set is interesting and perhaps at first puzzling. However, when discussed with the doctors who provided the data, it was discovered that the first and last set were taken from the *white matter* of the brain, while the other was taken from the *grey matter*. This later finding was consistent with observed microscopic tests on the various matter (micro channels in white matter that promote convective "flow"). Thus, it can be suggested with a reasonably high degree of confidence, that white matter exhibits convective transport, while grey matter does not.

8 Epi Model Comparison

We return to the previously introduced epidemiological model as another example of a way in which the model comparison statistic may be used. Here we apply this statistic to determine whether a more sophisticated model is appropriate based on the surveillance data from the Australian NNDS website. Here we introduce the modified model and describe the test statistic for this example. We then present the results from the least squares estimation procedure in both the cases of the simplified and more complex model, and finally, interpret the conclusions indicated by the test statistic.

So far, in our model of invasive pneumococcal disease dynamics, we have considered the progression of individuals from a colonized to an infected state by a constant linear per capita rate. However, this is a gross simplification of more complex physiological processes, many of which occur within the individual and would likely require more sophisticated mathematical representations. But it is also possible that at a population level, this linear term may sufficiently capture the dynamics of the infections when the model solutions are compared with observed data. One specific mechanism that we can explicitly consider is 'exogenous reinfection', that is, the establishment of an infection within a colonized individual through repeated exposure to *S. pneumoniae* via contacts with other individuals harboring the bacteria. The inclusion of this mechanism results in the following modified model equations

$$\frac{dS}{dt} = \lambda - \beta_1 S\frac{E+E_V+I+I_V}{N} + \alpha E + \gamma I - \phi S + \rho S_V - \mu S \tag{77}$$

$$\frac{dE}{dt} = \beta_1 S\frac{E+E_V+I+I_V}{N} - \alpha E - l\kappa(t)E - \phi E + \rho E_V - \mu E - l\beta_2 E\frac{E+E_V+I+I_V}{N} \tag{78}$$

$$\frac{dS_V}{dt} = \phi S - \epsilon\beta_1 S_V\frac{E+E_V+I+I_V}{N} + \alpha E_V + \gamma I_V - \rho S_V - \mu S_V \tag{79}$$

$$\frac{dE_V}{dt} = \epsilon\beta_1 S_V\frac{E+E_V+I+I_V}{N} - \alpha E_V + \phi E - \rho E_V - \delta\kappa(t)E_V - \mu E_V - \delta\beta_2\frac{E+E_V+I+I_V}{N} \tag{80}$$

$$\frac{dI}{dt} = l\kappa(t)E + l\beta_2 E\frac{E+E_V+I+I_V}{N} - (\gamma+\eta+\mu)I \tag{81}$$

$$\frac{dI_V}{dt} = \delta\kappa(t)E_V + \delta\beta_2\frac{E+E_V+I+I_V}{N} - (\gamma+\eta+\mu)I_V. \tag{82}$$

8.1 Surveillance Data

Our interpretation of the case notification data must also be modified to reflect the additional infection mechanism, so that the number of new cases is now

$$Y_j \sim f(t_j, \vec{\theta}) = \int_{t_j}^{t_{j+1}} [l\kappa E + l\beta_2 E \frac{E + E_V + I + I_V}{N} + \delta\kappa E_V + \delta\beta_2 E_V \frac{E + E_V + I + I_V}{N}] ds,$$

where $j = 1, ..., 36$. We estimate parameters $\vec{\theta} = (\beta_1, \kappa_0, \kappa_1, \delta, \beta_2)^T$ now from these 36 monthly cases, and from the corresponding annual reports of which of these cases were vaccinated or unvaccinated. These data are represented by

$$Y_i \sim f(t_i, \vec{\theta}) = \int_{t_i}^{t_{i+1}} \left[l\kappa E + l\beta_2 E \frac{E + E_V + I + I_V}{N} \right] ds,$$

and

$$Y_i \sim f(t_i, \vec{\theta}) = \int_{t_i}^{t_{i+1}} \left[\delta\kappa E_V + \delta\beta_2 E_V \frac{E + E_V + I + I_V}{N} \right] ds,$$

for $i = 1, 2, 3$, and $t_i = 1, 13, 25, 37$ months for $i = 1, \ldots, 4$. Again, we assume the statistical model $Y_j = f(t_j, \vec{\theta_0}) + \epsilon_j$ where $E[\epsilon_j] = 0$, $var(\epsilon_j) = \sigma_1^2$ for all $j = 1, ..., 36$, and $Y_i = f(t_i, \vec{\theta_0}) + \epsilon_i$ where $E[\epsilon_i] = 0$, $var(\epsilon_i) = \sigma_2^2$ for all $i = 1, 2, 3$. Thus, we have assumed that the variance is constant longitudinally, but not equivalent across types of observations. That is, it is likely that there is more variation in the annually reported observations than those reported on a monthly basis. The least squares estimation procedure is described in more detail in [32].

8.2 Test Statistic

Here we describe the application of a test statistic to this example to compare the modified model to the comparably simpler model. The statistic will provide a basis from which to decide whether the observed data warrants the additional complexity incorporated in the above model.

From the $n = 42$ observations Y_j approximated by the model quantities $f(t_j, \vec{\theta})$, we seek to estimate parameters $\vec{\theta} = (\beta_1, \kappa_0, \kappa_1, \delta, \beta_2)^T$. We obtain these estimates via a least squares estimation process in which our estimate for $\vec{\theta}$ minimizes an objective functional $J_n(\vec{\theta})$. When $f(t_j, \vec{\theta})$ is that for the more sophisticated model above,

$$\hat{\theta} = \arg\min_{\vec{\theta} \in \Theta} J_n(\vec{\theta}),$$

where $\Theta \subset \mathbb{R}^5_+$ is a (compact) feasible parameter set. The constraint operator $H : \mathbb{R}^5 \to \mathbb{R}^1$ of (75) for our example is then the 1×5 vector $H = (0, 0, 0, 0, 1)$ and $c = 0$.
Thus the reduced parameter space (in the case of the reduced model, where $\beta_2 = 0$), is

Table 9: Parameter estimates without and with 'exogenous reinfection'.

$\vec{\theta}$	$\hat{\theta}_H$ ($\beta_2 = 0$)	$SE(\hat{\theta}_H)$	$\hat{\theta}$ ($\beta_2 \neq 0$)	$SE(\hat{\theta})$
β	1.52175	0.02	1.52287	0.0029
κ_0	$1.3656e^{-3}$	$1.3e^{-4}$	$1.3604e^{-3}$	0.0012
κ_1	0.56444	0.04	0.5632	0.52
δ	0.7197	0.06	0.71125	0.38
β_2	N/A	N/A	$2.2209e^{-14}$	0.01

$$\Theta_H = \{\vec{\theta} \in \Theta : H\vec{\theta} = 0\} = \{\vec{\theta} \in \Theta : \beta_2 = 0\}.$$

The estimate for $\vec{\theta}$ over Θ_H is denoted by $\hat{\theta}_H$, and is found by minimizing the same objective functional over the smaller parameter space Θ_H, i.e.,

$$\hat{\theta}_H = \arg\min_{\vec{\theta} \in \Theta_H} J_n(\vec{\theta}).$$

We use the test statistic

$$U_n = n\frac{J_n(\hat{\theta}_H) - J_n(\hat{\theta})}{J_n(\hat{\theta})}$$

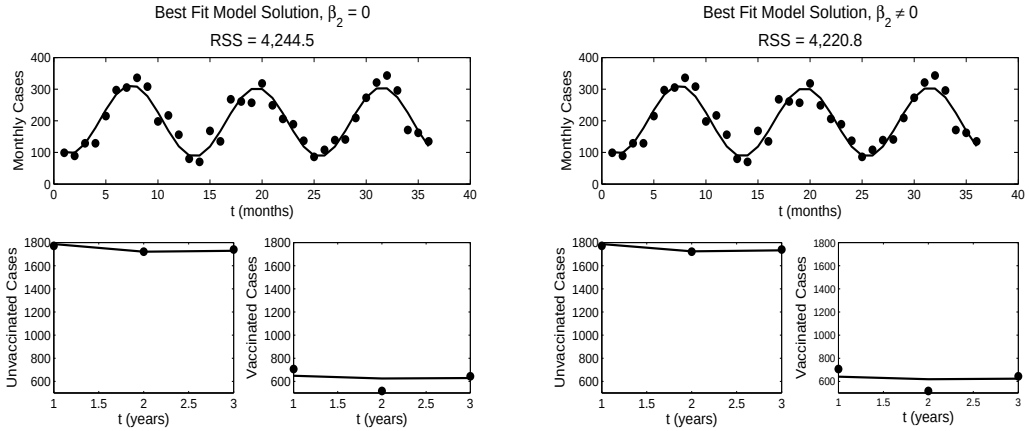
which, under reasonable conditions, converges to a $\chi^2(1)$ distribution. For a significance level α, we would find a threshold τ such that $Pr\{\chi^2(1) > \tau\} = \alpha$. Then if $U_n > \tau$, we reject our null hypothesis as false, otherwise we do not reject. Our null hypothesis in this case is $H_0 : K\vec{\theta}_0 = 0$, or that the true $\beta_2 = 0$.

8.3 Inverse Problem Results

In this section we compare the results of the least squares estimation procedure with and without the 'exogenous reinfection' term. The same set of surveillance data, described in Section 8.1, is used in both cases. The parameter estimates and corresponding standard errors are shown in Table 9.

The parameter estimates themselves, $\hat{\theta}_H$ and $\hat{\theta}$, do not differ significantly. Although, the standard errors indicate that our ability to estimate β and $\kappa(t)$ does change drastically depending on whether or not the two mechanisms of infection are considered. When the reinfection term is considered, we see that the standard error for this particular parameter indicates that our data do not provide a significant amount of information on this process. However, the smaller residual, RSS, when the objective functional is minimized over a larger parameter space (when $\beta_2 \neq 0$), might indicate that including the extra term provides a better fit. To resolve these two seemingly contrasting pieces of information, we turn to the

Figure 22: Best fit model solutions to monthly case notifications with constant variance error.



test statistic to determine if the difference in residuals is enough to justify the inclusion of this extra infection rate.

8.4 Model Comparison

The test statistic can be calculated as

$$U_n = n\frac{J_n(\hat{\theta}_H) - J_n(\hat{\theta})}{J_n(\hat{\theta})} = 42 \times \frac{4,244.5 - 4,220.8}{4,220.8} = 0.236.$$

Note that the residual sum of squares is the value of the objective function, so that $RSS = J_n(\hat{\theta})$. We compare this to a $\chi^2(1)$ table (see Table 8) and see that even at a significance level of only 75% we cannot reject our null hypothesis. That is, the difference in residuals, and hence the improvement of the model fits to this data (with $n = 42$), is not sufficient to warrant including the additional infection mechanism. This does not mean that reinfection does not occur, but it does suggest that to accurately capture the dynamics of the population, as evidenced by this surveillance data, it is reasonable to neglect this term. Therefore, we conclude that "reinfection" is not sufficiently present in this data to argue for inclusion of this term in population level models of the infection dynamics.

9 Concluding Remarks

As might be expected, mathematical and statistical models cannot fully represent reality in most scientific situations. The best that one can hope is that models can approximate reality as presented by data from experiments sufficiently well to be useful in promoting basic understanding as well as prediction. We have in this presentation outlined some techniques for evaluation of assumptions regarding statistical models as well as comparison techniques for mathematical models under the assumption that the statistical model assumed is correct. The RSS based techniques discussed represent just one (which happens to enjoy a rigorous theoretical foundation!) of many model comparison/selection techniques available in a large literature. For example, among a wide class of so-called "model selection" methods (some of which are heuristic in nature) are those based on Kullbeck-Leibler information loss. Among the best known of these is the Akaike's Information Criterion (AIC) selection procedure and its numerous variations (AIC_c,TIC, etc.) [13, 14, 15, 16, 17, 28] as well as Bayesian model selection (e.g., BIC) procedures. While these are important modeling tools, space limitations prohibit their discussion here.

Finally, we have also limited our discussions to estimation problems based on OLS and GLS with appropriate corresponding data noise assumptions of constant variance and non-constant variance (relative error), respectively. There are many other important approaches (e.g., regularization, asymptotic embedding, perturbation, equation error, adaptive filtering and identification, and numerous Bayesian based techniques–see [8, 12, 20, 23, 27, 33, 35] and the references therein) which again we ignore because of space limitations.

Acknowledgments

This research was supported in part by the National Institute of Allergy and Infectious Disease under grant 9R01AI071915-05, in part by the U.S. Air Force Office of Scientific Research under grant AFOSR-FA9550-04-1-0220, and in part by the Statistical and Applied Mathematical Sciences Institute, which is funded by NSF under grant DMS-0112069.

References

[1] 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..

[2] H. T. Banks, S. Dediu and S.E. Ernstberger, Sensitivity functions and their uses in inverse problems, CRSC-TR07-12, July, 2007; *J. Inverse and Ill-posed Problems*, to appear.

[3] H. T. Banks, S. Dediu and H.K. Nguyen, Sensitivity of dynamical systems to parameters in a convex subset of a topological vector space, CRSC-TR06-25, September, 2006; *Math. Biosci. and Engineering,* **4** (2007), 403–430.

[4] H.T. Banks, S.L. Ernstberger and S.L.Grove, Standard errors and confidence intervals in inverse problems: sensitivity and associated pitfalls, CRSC-TR06-10, March, 2006; *J. Inv. Ill-posed Problems,* **15** (2006), 1–18.

[5] 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.

[6] 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.

[7] H. T. Banks and P. Kareiva, Parameter estimation techniques for transport equations with application to population dispersal and tissue bulk flow models, LCDS Report #82-13, July 1982, Brown University; *J. Math. Biol.*, **17** (1983), 253–273.

[8] H. T. Banks and K. Kunsich, *Estimation Techniques for Distributed Parameter Systems*, Birkhauser, Boston, 1989.

[9] H. T. Banks and H. K. Nguyen, Sensitivity of dynamical system to Banach space parameters, CRSC-TR05-13, February, 2005; *J. Math. Analysis and Applications*, **323** (2006), 146–161.

[10] P. Bai, H. T. Banks, S. Dediu, A. Y. Govan, M. Last, A. Loyd, H. K. Nguyen, M. S. Olufsen, G. Rempala, and B. D. Slenning, Stochastic and deterministic models for agricultural production networks, CRSC-TR07-06, February, 2007; *Math. Biosci. and Engineering,* **4** (2007), 373–402.

[11] J. J. Batzel, F. Kappel, D. Schneditz and H.T. Tran, *Cardiovascular and Respiratory Systems: Modeling, Analysis and Control*, SIAM, Philadelphia, 2006.

[12] J. Baumeister, *Stable Solution of Inverse Problems*, Vieweg, Braunschweig, 1987.

[13] E.J. Bedrick and C.L. Tsai, Model selection for multivariate regression in small samples, *Biometrics*, **50** (1994), 226–231.

[14] H. Bozdogan, Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions, *Psychometrika*, **52** (1987), 345–370.

[15] H. Bozdogan, Akaike's Information Criterion and recent developments in information complexity, *Journal of Mathematical Psychology*, **44** (2000), 62–91.

[16] K. P. Burnham and D.R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, Springer, Berlin Heidelberg New York, 2002.

[17] K. P. Burnham and D.R. Anderson, Multimodel inference: Understanding AIC and BIC in model selection, *Sociological Methods and Research,* **33** (2004), 261–304.

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

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

[20] B. Chalmond, *Modeling and Inverse Problems in Image Analysis*, Springer, Berlin Heidelberg New York, 2003.

[21] J. B. Cruz, ed., *System Sensitivity Analysis*, Dowden, Hutchinson & Ross, Stroudsberg, PA, 1973.

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

[23] H. W. Engl, M. Hanke and A. Neubauer, *Regularization of Inverse Problems*, Kluwer, Dordrecht, 1996.

[24] M. Eslami, *Theory of Sensitivity in Dynamic Systems: An Introduction*, Springer, Berlin Heidelberg New York, 1994.

[25] P.M. Frank, *Introduction to System Sensitivity Theory*, Academic, New York, 1978.

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

[27] A. Gelb, ed., *Applied Optimal Estimation*, MIT Press, Cambridge, 1979.

[28] C.M. Hurvich and C.L. Tsai, Regression and time series model selection in small samples, *Biometrika*, **76** (1989), 297–307.

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

[30] A. Saltelli, K. Chan and E.M. Scott, eds., *Sensitivity Analysis*, Wiley, New York, 2000.

[31] G. A. F. Seber and C. J. Wild, *Nonlinear Regression,* Wiley, New York, 1989.

[32] K. L. Sutton, H.T. Banks, and C. Castillo-Chávez, Estimation of invasive pneumococcal disease dynamic parameters and the impact of conjugate vaccination in Australia, CRSC-TR07-15, August, 2007; *Mathematical Biosciences and Engineering*, **5** (2008), 175–204.

[33] A. Tarantola, *Inverse Problem Theory and Methods for Model Parameter Estimation*, SIAM, Philadelphia, 2004.

[34] K. Thomaseth and C. Cobelli, Generalized sensitivity functions in physiological system identification., *Ann Biomed Eng.*, **27(5)** (1999), 607–616.

[35] C. R. Vogel, *Computational Methods for Inverse Problems*, SIAM, Philadelphia, 2002.

[36] D. D. Wackerly, W. Mendenhall III, and R. L. Scheaffer, *Mathematical Statistics with Applications*, Duxbury Thompson Learning, USA, 2002.