Inverse Problems: Mathematical and Statistical Methodology

H. T. Banks

Center for Research in Scientific Computation (CRSC)
Center for Quantitative Sciences in Biomedicine (CQSB)
North Carolina State University
Raleigh, NC 27695

Outline of Talk

- Motivation: Non-invasive Detection and Characterization of Occlusion and Cardiac Stenosis
- Inverse Problems and Parameter Estimation: MLE, OLS, GLS
- Computation of Σ, Standard Errors and Confidence Intervals
- Error Analysis and Residual Plots Techniques
- Model Comparison Techniques

References

[1] H.T. Banks and N. S. Luke, Simulations of propagating shear waves in biotissue employing an internal variable approach to dissipation , Tech. Rep. CRSC-TR06-28, NCSU, December, 2006; *Communications in Computational Physics*, **3** (2008), 603–640.

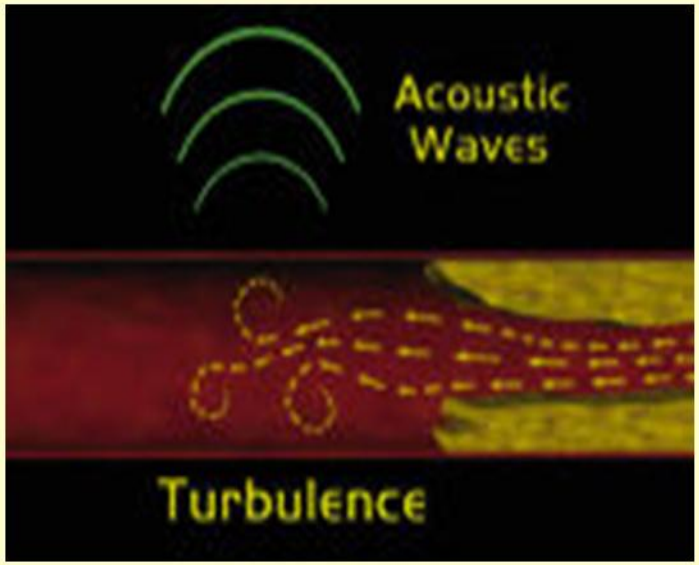
[2] H.T. Banks and J.R. Samuels, Jr.,Detection of cardiac occlusions using viscoelastic wave propagation, CRSC-TR08-23, December, 2008; *Advances in Applied Mathematics and Mechanics,* **1** (2009), 1–28.

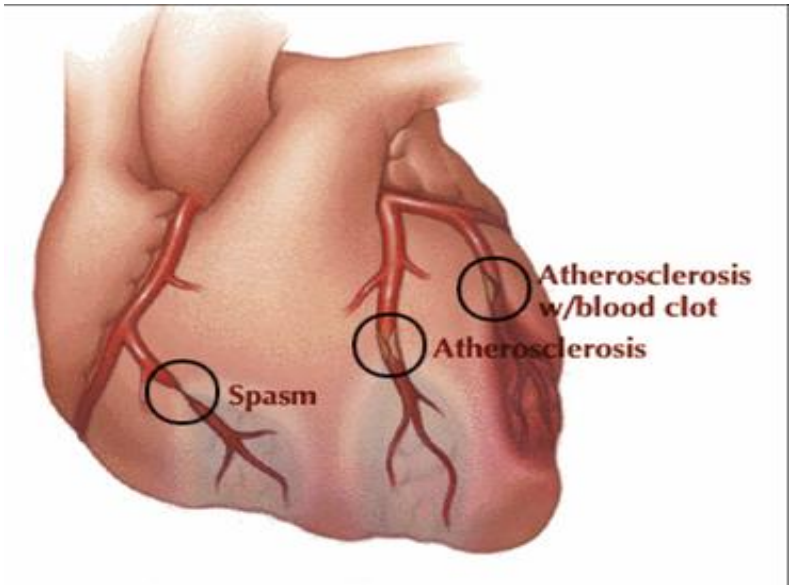
[3] H.T. Banks, M. Davidian, J.R. Samuels, Jr., and K.L. Sutton, An Inverse Problem Statistical Methodology Summary, *CRSC-TR08-01*, January, 2008; Chapter 11 in *Statistical Estimation Approaches in Epidemiology*, (edited by Gerardo

Chowell, Mac Hyman, Nick Hengartner, Luis M.A Bettencourt and Carlos Castillo-Chavez), Springer, Berlin Heidelberg New York, 2009, pp. 249–302.

Motivation

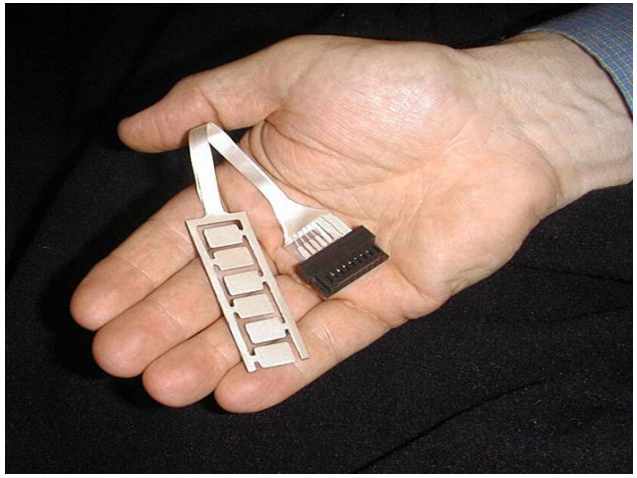
- Develop models and inverse problem techniques for detection and characterization of cardiac stenosis
- Stenosis induced turbulence in arteries produces normal forces on artery walls–propagated as shear waves to chest surface

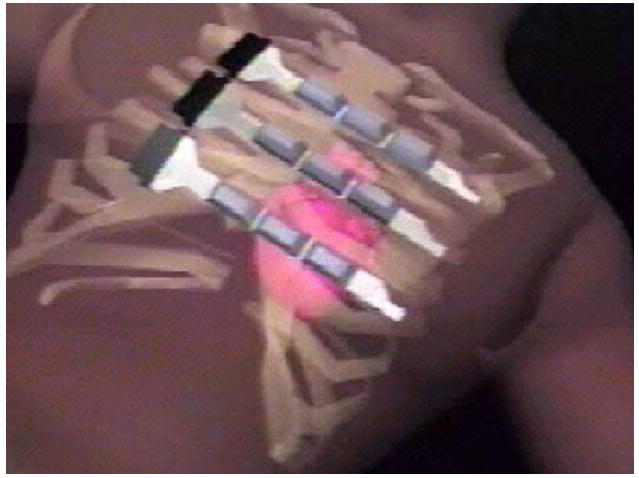




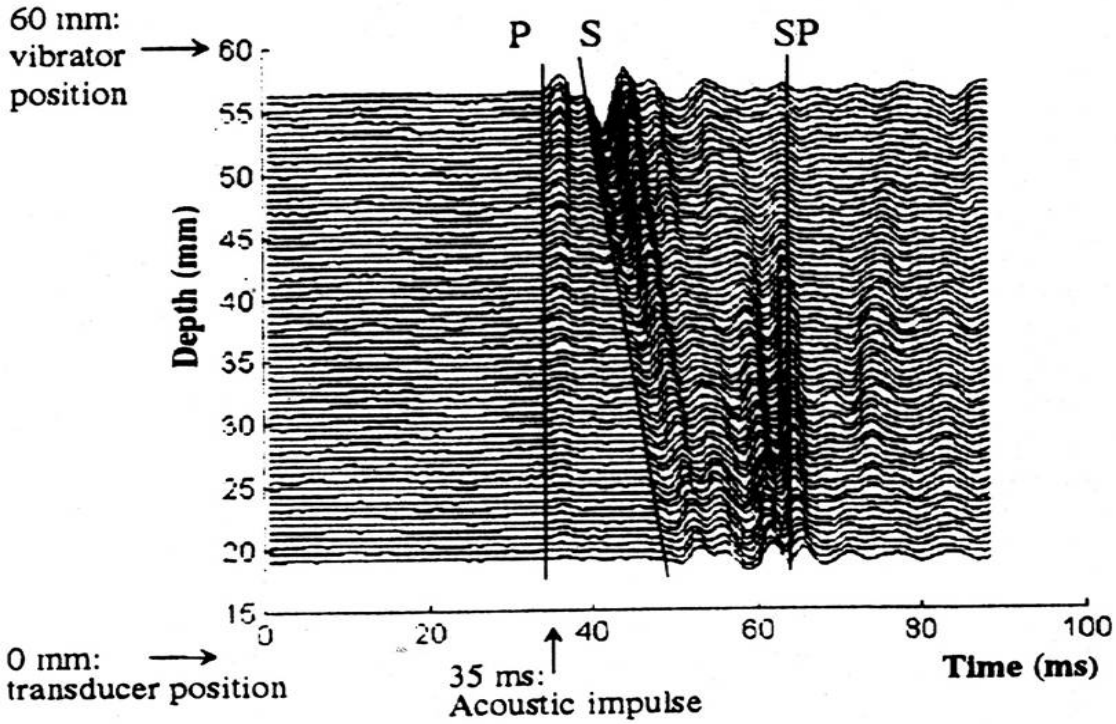
Stenosis or Arterial Occlusion

- Sensors developed at MedAcoustics measure shear waves–need to understand sensor capabilities, sensitivities, uncertainty in estimates produced by inverse algorithms–experiments to be designed with scientists, physicians at Brunel University (Shaw, Whiteman), Queen Mary Univ and Royal London Hospital (Greenwald), Barts/London NHS Trust (Birch)

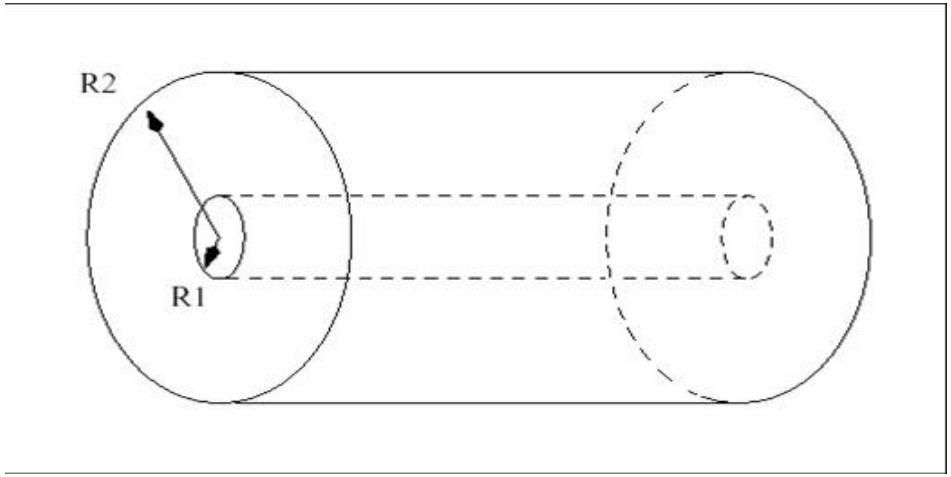


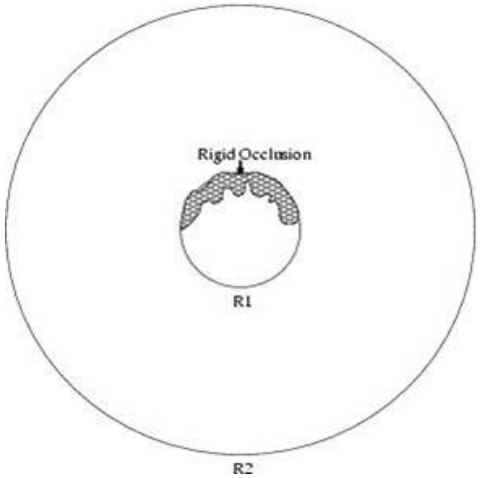


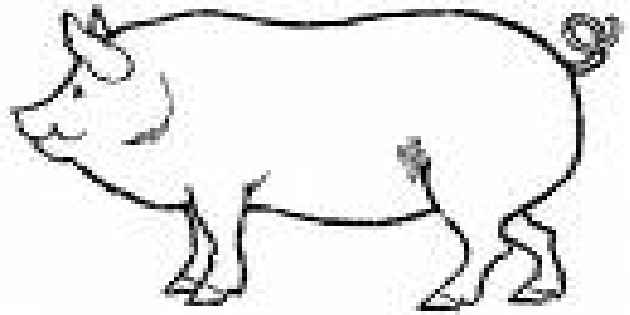
Piezoelectric Sensors


60 mm:
vibrator
position
P
S
SP
60
55
50
45
40
35
30
25
20
15
Depth (mm)
0
20
40
60
80
100
Time (ms)
0 mm:
transducer position
35 ms:
Acoustic impulse

Shear Waves

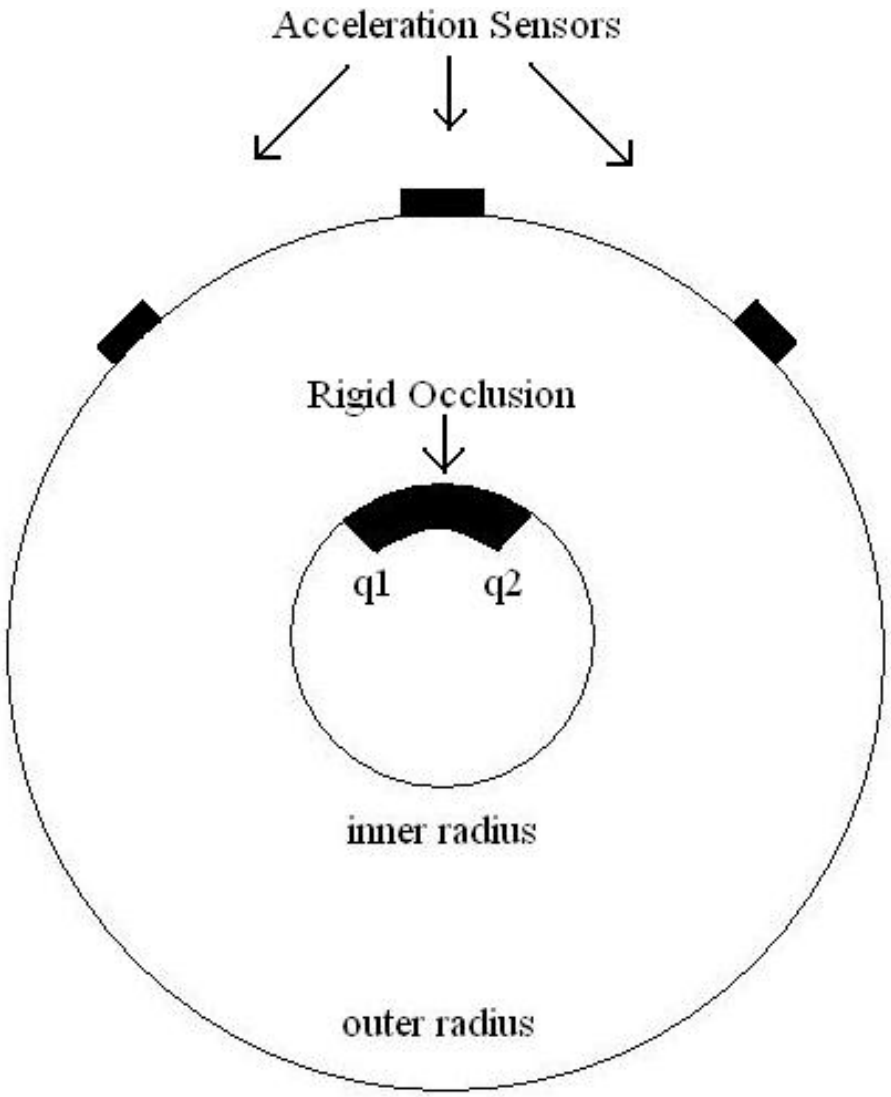






Model Geometry

gel phantoms, pigs, humans



2-D geometry with multiple chest sensors

2-D Model [BL] (internal variable based viscoelastic)

$$
\begin{aligned}
\rho\frac{\partial^2 u_1}{\partial t^2} &= \frac{\partial}{\partial r}\left(\epsilon_\lambda + \epsilon_\mu^{11}\right) + \frac{1}{r}\frac{\partial}{\partial \theta}\left(\epsilon_\mu^{21}\right) + \frac{1}{r}\left(\epsilon_\mu^{11} - \epsilon_\mu^{22}\right) \\
\rho\frac{\partial^2 u_2}{\partial t^2} &= \frac{\partial}{\partial r}\left(\epsilon_\mu^{12}\right) + \frac{1}{r}\frac{\partial}{\partial \theta}\left(\epsilon_\lambda + \epsilon_\mu^{22}\right) + \frac{1}{r}\left(\epsilon_\mu^{12} + \epsilon_\mu^{21}\right) \\
\frac{\partial \epsilon_{\lambda_k}}{\partial t} &= -\nu_{\lambda_k}\epsilon_{\lambda_k} + C_{\lambda_k}\frac{\partial}{\partial t}\left(S_{11}^{(\mathrm{e})} + S_{22}^{(\mathrm{e})}\right) \\
\frac{\partial \epsilon_{\mu_k}^{11}}{\partial t} &= -\nu_{\mu_k}\epsilon_{\mu_k}^{11} + C_{\mu_k}\frac{\partial}{\partial t}\left(S_{11}^{(\mathrm{e})}\right) \\
\frac{\partial \epsilon_{\mu_k}^{12}}{\partial t} &= -\nu_{\mu_k}\epsilon_{\mu_k}^{12} + C_{\mu_k}\frac{\partial}{\partial t}\left(S_{12}^{(\mathrm{e})}\right) \\
\frac{\partial \epsilon_{\mu_k}^{22}}{\partial t} &= -\nu_{\mu_k}\epsilon_{\mu_k}^{22} + C_{\mu_k}\frac{\partial}{\partial t}\left(S_{22}^{(\mathrm{e})}\right)
\end{aligned}
$$

Relevant model terms:

- radial (r) and tangential (θ) displacement of the shear wave (u_1 and u_2 respectively)
- "elastic" stress tensor $S_{ij}^{(\mathrm{e})}$ and internal strain variables ϵ_λ and ϵ_μ^{kl}
- parameters C_{λ_k} and C_{μ_k} for $k = 1, 2$ in approximating reduced relaxation parameter G in Fung's quasi-linear viscoelastic model
- occulsion parameters q_1 and q_2 which arise in the inner radius boundary conditions

Possible parameters to be estimated:

$$\vec{\theta} = (q_1, q_2, C_{\lambda_k}, C_{\mu_k}, \nu_{\lambda_k}, \nu_{\mu_k})^T$$

Inverse Problems and Parameter Estimation: MLE, OLS, and GLS

The Underlying Mathematical and Statistical Models

Inverse or parameter estimation problems in context of 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**

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

As usual assume a discrete form of observations–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 corresponding observations or data $\{\vec{y}_j\}$ will not be exactly $\vec{y}(t_j)$ — must treat this uncertainty pertaining to the observations with a statistical model for the observation process

Description of Statistical Model

We consider a **statistical model** of the form

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

where

- $\vec{f}(t_j, \vec{\theta}) = \mathcal{C}\vec{x}(t_j; \vec{\theta})$, $j = 1, \ldots, n$, corresponds to solution of mathematical model at the j^{th} covariate for a particular vector of parameters $\vec{\theta} \in R^p, \vec{x} \in R^N, \vec{f} \in R^m$,
- $\mathcal{C}$ is an $m \times N$ observation matrix
- $\vec{\theta}_0$ represents the "truth" or the parameters that generate the observations $\{\vec{Y}_j\}_{j=1}^n$
- The term $\vec{\mathcal{E}}_j$ 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 $\vec{f}(t, \vec{\theta})$

- Fluctuations are unknown to modeler—assume $\vec{\mathcal{E}}_j$ generated from a probability distribution that reflects the assumptions regarding these phenomena, e.g., in a statistical model for pharmacokinetics of drug in human blood samples, a natural distribution for $\vec{\mathcal{E}} = (\mathcal{E}_1, \ldots, \mathcal{E}_n)^T$ might be the multivariate normal distribution

- Needs: methodology related to the estimation of the true value of the parameters $\vec{\theta}_0$ from a set Θ of admissible parameters, the variance of the error $\mathrm{var}(\vec{\mathcal{E}}_j)$ and resulting uncertainties

- Here: two inverse problem methodologies for calculation of 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 distributions of the error process $\{\vec{\mathcal{E}}_j\}$ are known

MLE: Known error processes: Normally distributed error

- In statistical model, made no mention of the probability distribution that generates the error $\vec{\mathcal{E}}_j$
- In many situations one readily assumes that errors $\vec{\mathcal{E}}_j,\ j = 1, \ldots, n$, are **independent and identically distributed (iid)**
- **In some cases,** one is able to make **further assumptions on the error**, namely that the **distribution for $\vec{\mathcal{E}}_j$ is known**. In this case maximum likelihood techniques may be used.
- Discuss first for a **scalar** observation system, i.e., $m = 1$. Most common assumption: sufficient evidence to suspect the error is generated by a normal distribution then willing to assume $\mathcal{E}_j \sim \mathcal{N}(0, \sigma_0^2)$, and hence $Y_j \sim \mathcal{N}(f(t_j, \vec{\theta}_0), \sigma_0^2)$.

Can then obtain expression for determining $\vec{\theta}_0$ and σ_0 by seeking maximum over $(\vec{\theta}, \sigma^2) \in \Theta \times (0, \infty)$ of likelihood function for $\mathcal{E}_j = Y_j - f(t_j, \vec{\theta})$ defined by

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

- Resulting solutions θ_{MLE} and σ^2_{MLE} are the maximum likelihood **estimators** (MLEs) for $\vec{\theta}_0$ and σ_0^2, respectively
- Solutions $\theta_{\text{MLE}} = \theta_{\text{MLE}}(\vec{Y})$ and $\sigma^2_{\text{MLE}} = \sigma^2_{\text{MLE}}(\vec{Y})$ are *random variables* because $\vec{Y}$ is a random variable
- Corresponding maximum likelihood **estimates** are obtained by maximizing (5) with $\{Y_j\}$ replaced by a given realization $\vec{y} = \{y_j\}$—will be denoted by $\hat{\theta}_{\text{MLE}}$ and $\hat{\sigma}_{\text{MLE}}$, respectively.

Maximizing (5) equivalent to maximizing log likelihood

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

Can determine 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 $\frac{\partial}{\partial\vec{\theta}}\log L(\vec{Y}|\vec{\theta},\sigma^2) = 0$ which is equivalent to

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

Solving (7) is the **same** as the least squares optimization

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

Next fix $\vec{\theta}$ to be θ_{MLE} and solve $\frac{\partial}{\partial \sigma^2} \log L(\vec{Y}|\theta_{\text{MLE}}, \sigma^2) = 0$, 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—one that won't arise in more complicated formulations discussed later

The 2^{nd} derivative test verifies that expressions above for θ_{MLE} and σ^2_{MLE} do indeed maximize (6)

Vector Observations

For vector observations, for the j^{th} covariate t_j statistical model reformulated as

$$\vec{Y}_j = \vec{f}(t_j, \vec{\theta}_0) + \vec{\mathcal{E}}_j \tag{10}$$

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

$$V_0 = \text{var}(\vec{\mathcal{E}}_j) = \text{diag}(\sigma^2_{0,1}, \ldots, \sigma^2_{0,m}) \tag{11}$$

for $j = 1, \ldots, n$

- allows for possibility that observation coordinates Y^i_j may have different *constant* variances $\sigma^2_{0,i}$, i.e., $\sigma^2_{0,i}$ does not necessarily have to equal $\sigma^2_{0,k}$
- if (again) there is sufficient evidence to claim errors are *i.i.d.* and generated by a normal distribution then $\vec{\mathcal{E}}_j \sim \mathcal{N}_m(0, V_0)$

Can obtain the maximum likelihood estimators $\theta_{\rm MLE}(\{\vec{Y}_j\})$ and $V_{\rm MLE}(\{\vec{Y}_j\})$ for θ_0 and V_0 by determining the maximum of log of the likelihood function for $\vec{\mathcal{E}}_j = \vec{Y}_j - \vec{f}(t_j, \vec{\theta})$ defined by

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

$$= -\frac{n}{2}\sum_{i=1}^{m} \log \sigma_{0,i}^2 - \sum_{j=1}^{n}[\vec{Y}_j - \vec{f}(t_j, \vec{\theta})]^T V^{-1}[\vec{Y}_j - \vec{f}(t_j, \vec{\theta})].$$

Using arguments similar to those given for the scalar case, find 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**—requires some care when solving numerically–discuss issue further below

Unspecified Error Distributions and Asymptotic Theory

- Above examined the estimates of $\vec{\theta}_0$ and V_0 under the assumption *that the error is normally distributed and is constant longitudinally.*
- But what if it is suspected that the error is not normally distributed, or the error's distribution is completely unknown to the modeler (as in most applications)?
- How should we proceed in estimating $\vec{\theta}_0$ and σ_0 (or V_0) in these circumstances?
- Two popular estimation procedures for such situations: ordinary least squares (OLS) and generalized least squares (GLS)

Ordinary Least Squares (OLS)

Statistical model in the scalar case takes the form

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

where the variance $\mathrm{var}(\mathcal{E}_j) = \sigma_0^2$ is constant in longitudinal data (note that the error's distribution is **not** specified). Define

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

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

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

Note: $\theta_{\rm OLS}$ is a *random variable* ($\mathcal{E}_j = Y_j - f(t_j, \vec{\theta})$ 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

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

provides an realization for $\theta_{\rm OLS}$.

Once we have solved for $\theta_{\rm OLS}$ in (15), we can replace $\vec{\theta}_0$ in

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

by $\hat{\theta}_{\rm OLS}$ to obtain an estimate $\hat{\sigma}^2_{\rm OLS}$ for σ_0^2.

Even though the error's distribution not specified we can use **asymptotic theory** to approximate the mean and variance as well as distribution of the random variable θ_{OLS}—more detail below—as $n \to \infty$, find that

$$\theta_{\text{OLS}} \sim \mathcal{N}_p(\vec{\theta}_0, \sigma_0^2[\chi^T(\vec{\theta}_0)\chi(\vec{\theta}_0)]^{-1}) = \mathcal{N}_p(\vec{\theta}_0, \Sigma_0) \tag{19}$$

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

$$\chi_{jk}(\vec{\theta}) = \frac{\partial f(t_j, \vec{\theta})}{\partial \vec{\theta}_k}$$

Distribution of θ_{OLS} called **sampling distribution** and contains information about **uncertainty** in our **process** and the **estimates** it produces!!

However, $\vec{\theta}_0$ and σ_0^2 are **generally unknown**–usually instead use *realization* $\vec{y} = \{y_j\}_{j=1}^n$ of the random process $\vec{Y}$ to obtain estimate

$$\hat{\theta}_{\text{OLS}} = \arg\min_{\vec{\theta}\in\Theta} \sum_{j=1}^{n} [y_j - f(t_j, \vec{\theta})]^2 \tag{20}$$

and the *bias adjusted* estimate

$$\hat{\sigma}_{\text{OLS}}^2 = \frac{1}{n-p} \sum_{j=1}^{n} [y_j - f(t_j, \hat{\theta})]^2 \tag{21}$$

to use as an approximation in (19).

Note: (21) represents estimate for σ_0^2 of (18) with factor $\frac{1}{n}$ replaced by 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– Note: (18) true even in general nonlinear case–it does not rely on any asymptotic theories

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

$$\Sigma_0 \approx \hat{\Sigma} = \hat{\sigma}^2 [\chi^T(\hat{\theta})\chi(\hat{\theta})]^{-1}. \tag{22}$$

Can obtain the standard errors $SE(\hat{\theta}_{\text{OLS},k})$ (discussed in more detail later) for the k^{th} element of $\hat{\theta}_{\text{OLS}}$ by calculating

$$SE(\hat{\theta}_{\text{OLS},k}) \approx \sqrt{\hat{\Sigma}_{kk}}$$

Note similarity between the MLE equations (8) and (9), and the scalar OLS equations (20) and (21). That is, under a **normality assumption** for the error, **the MLE and OLS formulations are equivalent**

Vector Observation OLS

For vector observations for the j^{th} covariate t_j, assuming variance is still constant in longitudinal data, statistical model is reformulated as

$$\vec{Y}_j = \vec{f}(t_j, \vec{\theta}_0) + \vec{\mathcal{E}}_j \tag{23}$$

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

$$V_0 = \text{var}(\vec{\mathcal{E}}_j) = \text{diag}(\sigma^2_{0,1}, \ldots, \sigma^2_{0,m}) \tag{24}$$

for $j = 1, \ldots, n$. As in MLE case allow for possibility that observation coordinates Y^i_j may have different *constant* variances $\sigma^2_{0,i}$–Note: this formulation also can be used to treat case where V_0 is used to simply scale 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)).

Problem consists of finding minimizer

$$\theta_{\text{OLS}} = \arg\min_{\vec{\theta}\in\Theta} \sum_{j=1}^{n} [\vec{Y}_j - \vec{f}(t_j, \vec{\theta})]^T V_0^{-1} [\vec{Y}_j - \vec{f}(t_j, \vec{\theta})], \qquad (25)$$

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 (25) 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 $\vec{\mathcal{E}}_j = \vec{Y}_j - \vec{f}(t_j, \vec{\theta})$ is); hence if $\{\vec{y}_j\}_{j=1}^n$ is a realization of the *random process* $\{\vec{Y}_j\}_{j=1}^n$ then solving

$$\hat{\theta}_{\text{OLS}} = \arg\min_{\vec{\theta}\in\Theta} \sum_{j=1}^{n} [\vec{y}_j - \vec{f}(t_j, \vec{\theta})]^T V_0^{-1} [\vec{y}_j - \vec{f}(t_j, \vec{\theta})] \qquad (26)$$

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

$$\hat{V} = \text{diag}\left(\frac{1}{n-p}\sum_{j=1}^{n}[\vec{y}_j - \vec{f}(t_j, \hat{\theta})][\vec{y}_j - \vec{f}(t_j, \hat{\theta})]^T\right). \qquad (27)$$

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 $\hat{V}$:

$$\vec{\theta}_0 \approx \hat{\theta} = \arg\min_{\vec{\theta}\in\Theta} \sum_{j=1}^{n} [\vec{y}_j - \vec{f}(t_j, \vec{\theta})]^T \hat{V}^{-1} [\vec{y}_j - \vec{f}(t_j, \vec{\theta})] \tag{28}$$

$$V_0 \approx \hat{V} = \text{diag}\left(\frac{1}{n-p} \sum_{j=1}^{n} [\vec{y}_j - \vec{f}(t_j, \hat{\theta})][\vec{y}_j - \vec{f}(t_j, \hat{\theta})]^T \right) \tag{29}$$

Note: expressions for $\hat{\theta}$ and $\hat{V}$ constitute a coupled system of equations—-requires greater effort in implementing a numerical scheme

Just as in the scalar case we can determine the asymptotic properties of the OLS estimator (25). As $n \to \infty$, θ_{OLS} has the following asymptotic properties:

$$\theta_{\text{OLS}} \sim \mathcal{N}(\vec{\theta}_0, \Sigma_0) \tag{30}$$

where

$$\Sigma_0 = \left(\sum_{j=1}^{n} D_j^T(\vec{\theta}_0) V_0^{-1} D_j(\vec{\theta}_0) \right)^{-1} \tag{31}$$

and the $m \times p$ matrix $D_j(\vec{\theta})$ 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}:

$$\theta_{\text{OLS}} \sim \mathcal{N}_p(\vec{\theta}_0, \Sigma_0) \approx \mathcal{N}_p(\hat{\theta}, \hat{\Sigma}) \tag{32}$$

where

$$\Sigma_0 \approx \hat{\Sigma} = \left(\sum_{j=1}^{n} D_j^T(\hat{\theta}) \hat{V}^{-1} D_j(\hat{\theta}) \right)^{-1}. \tag{33}$$

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, note similarity between the MLE equations (12) and (13), and the OLS equations (28) and (29) for the vector statistical model (23).

Numerical Implementation: OLS

In scalar statistical model (14), the estimates $\hat{\theta}$ and $\hat{\sigma}$ can be solved for separately (also true of vector OLS) in the case $V_0 = \sigma_0^2 I_m$, where I_m is the $m \times m$ identity)—thus numerical implementation straightforward - first determine $\hat{\theta}_{\text{OLS}}$ by (20), then calculate $\hat{\sigma}^2_{\text{OLS}}$ according to (21)

The estimates $\hat{\theta}$ and $\hat{V}$ in the case of the vector statistical model (23), however, require more effort since they are coupled:

$$\hat{\theta} = \arg\min_{\vec{\theta}\in\Theta} \sum_{j=1}^{n} [\vec{y}_j - \vec{f}(t_j, \vec{\theta})]^T \hat{V}^{-1} [\vec{y}_j - \vec{f}(t_j, \vec{\theta})] \tag{34}$$

$$\hat{V} = \text{diag}\left(\frac{1}{n-p} \sum_{j=1}^{n} [\vec{y}_j - \vec{f}(t_j, \hat{\theta})][\vec{y}_j - \vec{f}(t_j, \hat{\theta})]^T \right) \tag{35}$$

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 (34). Set $k = 0$.
2. Use $\hat{\theta}^{(k)}$ to calculate $\hat{V}^{(k+1)}$ using (35).
3. Re-estimate $\vec{\theta}$ by solving (34) 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.

Generalized Least Squares (GLS)

In OLS formulation, error distribution remained unspecified–however required the error remain constant in variance in longitudinal data–assumption may not be appropriate for some data sets whose error depends on value of observation– A common relative error model that experimenters use (in this instance for the scalar observation case) is

$$Y_j = f(t_j, \vec{\theta}_0)\,(1 + \mathcal{E}_j) \tag{36}$$

where $E(Y_j) = f(t_j, \vec{\theta}_0)$ and $\mathrm{var}(Y_j) = \sigma_0^2 f^2(t_j, \vec{\theta}_0)$. 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* $\theta_{\rm GLS}$ the following equation must be solved for the estimator $\theta_{\rm GLS}$:

$$\sum_{j=1}^{n} w_j [Y_j - f(t_j, \theta_{\rm GLS})] \nabla f(t_j, \theta_{\rm GLS}) = 0, \tag{37}$$

where Y_j obeys (36) and $w_j = f^{-2}(t_j, \theta_{\rm GLS})$. The quantity $\theta_{\rm GLS}$ is a random variable, hence if $\{y_j\}_{j=1}^n$ is a *realization* of the random process Y_j then solving

$$\sum_{j=1}^{n} f^{-2}(t_j, \hat{\theta}) [y_j - f(t_j, \hat{\theta})] \nabla f(t_j, \hat{\theta}) = 0, \tag{38}$$

for $\hat{\theta}$ gives an estimate $\hat{\theta}_{\rm GLS}$ for $\theta_{\rm GLS}$

The GLS estimator has the following asymptotic properties:

$$\theta_{\text{GLS}} \sim \mathcal{N}_p(\vec{\theta}_0, \Sigma_0) \tag{39}$$

where

$$\Sigma_0 = \sigma_0^2 \left(F_{\vec{\theta}}^T(\vec{\theta}_0) W(\vec{\theta}_0) F_{\vec{\theta}}(\vec{\theta}_0) \right)^{-1}, \tag{40}$$

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

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 (40) to calculate

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

where we take the approximation

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

Then approximate standard errors of $\hat{\theta}_{\text{GLS}}$ by taking square roots of diagonal elements of $\hat{\Sigma}$. NOTE: solutions to (28) and (38) depend upon the numerical method used to find the minimum or root, and since Σ_0 depends upon $\vec{\theta}_0$ and hence its approx on the estimate for $\vec{\theta}_0$, standard errors are therefore affected by numerical method chosen.

GLS motivation

We note the similarity between (16) and (38). The GLS equation (38) can be motivated by examining the weighted least squares (WLS) estimator

$$\theta_{\text{WLS}} = \arg\min_{\vec{\theta}\in\Theta} \sum_{j=1}^{n} w_j [Y_j - f(t_j, \vec{\theta})]^2 \tag{41}$$

In many situations where the observation process is well understand, 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. If we differentiate the sum of squares in (41) 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 (38) 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 (38) can be given: if the data is 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 (38). 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 ($w_j = g^{-2}(t_j, \vec{\theta})$)

$$\sum_{j=1}^{n} g^{-2}(t_j, \vec{\theta})[Y_j - f(t_j, \vec{\theta})] \nabla f(t_j, \vec{\theta}) = 0 \quad ? \tag{42}$$

In their text, Carroll and Rupert briefly describe how distributions belonging to the exponential family of distributions generate maximum-likelihood estimating equations equivalent to (42).

Numerical Implementation of the GLS Procedure

Recall that an estimate $\hat{\theta}_{\text{GLS}}$ can either be solved directly according to (38) or iteratively using a procedure. An iterative procedure as 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

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

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

One finds in practice that the above procedure sometimes does not adequately estimate $\vec{\theta}_0$, so we instead outline a different numerical algorithm with which one often can achieve better results. Recall that the above iterative procedure was formulated by maximizing (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. Thus, an alternative iterative procedure involves completing the following steps:

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

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

   to obtain the $k+1$ estimate 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.

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 [**?**], then $\hat{w}_j \to f^{-2}(t_j, \hat{\theta}_{\text{GLS}})$.

Computation of Σ, Standard Errors and Confidence Intervals

Return to case of n scalar longitudinal observations, consider OLS case extension of ideas to vectors is completely straight-forward). Assume statistical model

$$Y_j \equiv f(t_j, \vec{\theta}_0) + \mathcal{E}_j, \;\; j = 1, 2, \ldots, n, \tag{43}$$

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). Further assume that the errors ϵ_j, $j = 1, 2, \ldots, n$, are $(i.i.d.)$ random variables with mean $E[\mathcal{E}_j] = 0$ and constant variance $var[\mathcal{E}_j] = \sigma_0^2$, 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 $var[Y_j] = \sigma_0^2$.

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 $\hat{\theta}^n_{\rm OLS}$ where

$$\hat{\theta}^n_{\rm OLS} = \arg\min J_n(\vec{\theta}) = \sum_{j=1}^{n} [y_j - f(t_j, \vec{\theta})]^2. \tag{44}$$

Since Y_j is a random variable, the corresponding estimator $\theta^n = \theta^n_{\rm OLS}$ (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 $\{y_j\}$.

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

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

That is, for n large, the sampling distribution approximately satisfies

$$\theta^n_{\text{OLS}}(Y) \sim \mathcal{N}_p(\vec{\theta}_0, \sigma_0^2[\chi^T(\vec{\theta}_0)\chi(\vec{\theta}_0)]^{-1}) := \mathcal{N}_p(\vec{\theta}_0, \Sigma_0) \qquad (45)$$

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 $\dot{\vec{x}} = \vec{g}(t, \vec{x}(t), \vec{\theta})$, then one can use the $N \times p$ matrix **sensitivity equations**

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

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

$$\Sigma_0 = \sigma_0^2[\chi^T(\vec{\theta}_0)\chi(\vec{\theta}_0)]^{-1} \approx \hat{\Sigma}(\hat{\theta}^n_{\text{OLS}}) = \hat{\sigma}^2[\chi^T(\hat{\theta}^n_{\text{OLS}})\chi(\hat{\theta}^n_{\text{OLS}})]^{-1}. \quad (46)$$

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

$$\sigma_0^2 \approx \hat{\sigma}^2 = \frac{1}{n-p}\sum_{j=1}^{n}[y_j - f(t_j, \hat{\theta}^n_{\text{OLS}})]^2. \quad (47)$$

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

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

$$P\{\hat{\theta}_k^n - t_{1-\alpha/2} SE_k(\hat{\theta}^n) < \theta_k^n < \hat{\theta}_k^n + t_{1-\alpha/2} SE_k(\hat{\theta}^n)\} = 1 - \alpha, \quad (48)$$

where $\alpha \in [0, 1]$ and $t_{1-\alpha/2} \in \mathbb{R}_+$. For small α (e.g., $\alpha = .05$ for 95% confidence intervals), the critical value $t_{1-\alpha/2}$ is computed from 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 $T \sim t^{n-p}$.

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 (e,g., how taken) of the sampling times $\{t_j\}$. Moreover, it is the matrix $[\chi^T \chi]^{-1}$ in (46) and the parameter $\hat{\sigma}^2$ in (47) that ultimately determine the SE and CI. At first investigation of (47), it appears that an increased number n of samples will 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 (47).

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 [SeWi]) is

$$\frac{1}{n}\chi^T(\vec{\theta})\chi(\vec{\theta}) \to \mathcal{X}(\vec{\theta}) \quad \text{as n} \to \infty$$

for some **nonsingular** matrix $\mathcal{X}(\vec{\theta}_0)$. 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 [BEG]).

All of the above theory readily generalizes to vector systems with partial, non-scalar observations. Suppose we have a vector system with partial vector observations: we have m coordinate observations where $m \leq N$. In this case, we have

$$\frac{d\vec{x}}{dt}(t) = \vec{g}(t, \vec{x}(t), \vec{\theta}) \tag{49}$$

and

$$\vec{y}_j = \vec{f}(t_j, \vec{\theta}_0) + \vec{\epsilon}_j = \mathcal{C}\vec{x}(t_j, \vec{\theta}_0) + \vec{\epsilon}_j, \tag{50}$$

where $\mathcal{C}$ is an $m \times N$ matrix and $\vec{f} \in R^m, \vec{x} \in R^N$. Assume that different observation coordinates f_i may have different variances σ_i^2 associated with different coordinates of the errors $\mathcal{E}_j$, then we have

$$\vec{\mathcal{E}}_j \sim \mathcal{N}_m(\vec{0}, V_0)$$

where $V_0 = diag(\sigma_{0,1}^2, ..., \sigma_{0,m}^2)$. May follow similar asymptotic theory to calculate approximate covariances, SE and CI.

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 $\mathcal{E}_j$, $j = 1, 2, \ldots, n$, are independent identically distributed (*i.i.d.*) random variables with mean $E[\mathcal{E}_j] = 0$ and constant variance $var[\mathcal{E}_j] = \sigma_0^2$.

- After carrying out the estimation procedures, one can readily plot the *residuals* $r_j = y_j - f(t_j, \hat{\theta}^n_{OLS})$ *vs. time* t_j and the *residuals vs. the resulting estimated model/ observation* $f(t_j, \hat{\theta}^n_{OLS})$ *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.

Investigation of Statistical Assumptions

Form of data dictates which method

- OLS (for constant variance obs. $Y_j = f(t_j, \vec{\theta}_0) + \mathcal{E}_j$)
- GLS (for nonconstant variance obs. $Y_j = f(t_j, \vec{\theta}_0)(1 + \mathcal{E}_j)$)

should be used.

Note that in order to obtain the correct standard errors for a set of data 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 the correct asymptotic statistics were used or not.

Residual Plots

One can carry out simulation studies with a proposed mathematical model to assist in understanding the behavior of the model in IP with different types of date with respect to mis-specification of the statistical model. For example, we consider a statistical model with constant variance noise

$$Y_j = f(t_j, \vec{\theta}_0) + \frac{\alpha}{100}\mathcal{E}_j, \qquad \mathrm{Var}(Y_j) = \frac{\alpha^2}{10000}\sigma^2,$$

and nonconstant variance noise

$$Y_j = f(t_j, \vec{\theta}_0)(1 + \frac{\alpha}{100}\mathcal{E}_j), \qquad \mathrm{Var}(Y_j) = \frac{\alpha^2}{10000}\sigma^2\, f^2(t_j, \vec{\theta}_0).$$

We can obtain a data set by considering a *realization* $\{y_j\}_{j=1}^n$ of the random process $\{Y_j\}_{j=1}^n$ 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

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

Since it is assumed that the error $\mathcal{E}_j$ 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.

What should we expect if this residual test is applied to a data set that has nonconstant variance generated error? That is, what if the data is incorrectly assumed to have constant variance error when in fact it has nonconstant variance error? Since in the nonconstant variance example, $R_j = Y_j - f(t_j, \vec{\theta}_0) = f(t_j, \vec{\theta}_0)\,\mathcal{E}_j$ 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.

If a data set has nonconstant variance generated data then

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

If the distributions $\mathcal{E}_j$ are *i.i.d.* then a plot of the modified residuals $r_j^m = (y_j - f(t_j, \hat{\theta}))/f(t_j, \hat{\theta})$ vs. t_j should be random in the 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 is incorrectly assumed to have nonconstant variance error when in fact it has constant variance error? Since $Y_j - f(t_j, \vec{\theta}_0) = \mathcal{E}_j$ in the constant variance case, we should expect that a plot of $r_j^m = (y_j - f(t_j, \hat{\theta}))/f(t_j, \hat{\theta})$ vs. t_j as well as that for $r_j^m = (y_j - f(t_j, \hat{\theta}))/f(t_j, \hat{\theta})$ vs. $f(t_j, \hat{\theta})$ should possess some distinct pattern.

Two further issues re 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. 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 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.

Example using Residual Plots

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

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

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

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

The left plot in Figure 1 depicts the solution of the logistic model for $K = 17.5$, $r = .7$ and $x_0 = 1$ for $0 \leq t \leq 25$. 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).

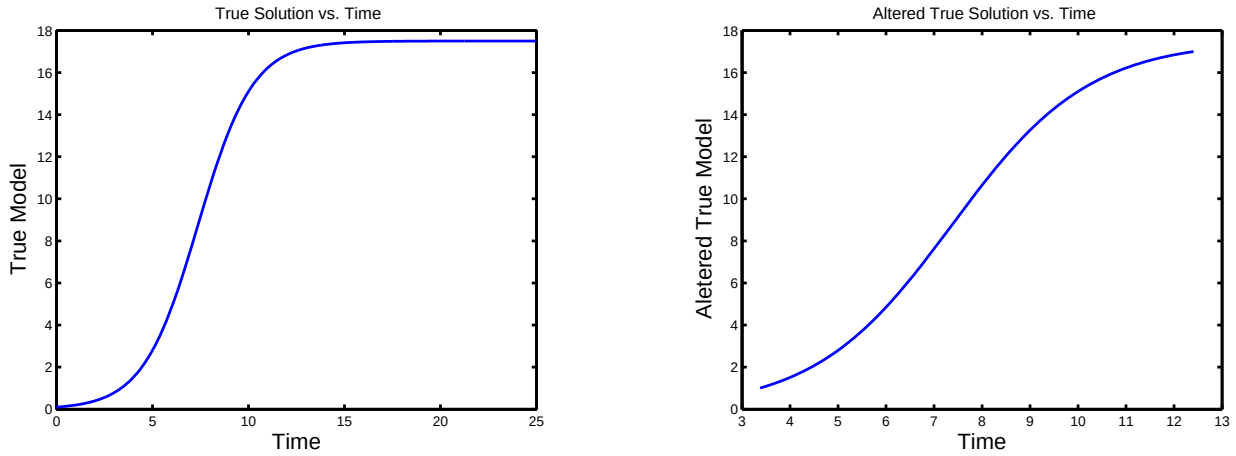


Figure 1: Original and truncated logistic curve with $K = 17.5$, $r = .7$ and $x_0 = .1$.

For this example we generated both constant variance and nonconstant variance noisy data 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 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.

α	$\vec{\theta}_{\text{init}}$	$\vec{\theta}_0$	$\hat{\theta}^{\text{cv}}_{\text{OLS}}$	$\text{SE}(\hat{\theta}^{\text{cv}}_{\text{OLS}})$	$\hat{\theta}^{\text{tcv}}_{\text{OLS}}$	$\text{SE}(\hat{\theta}^{\text{tcv}}_{\text{OLS}})$
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 1: Estimation using the OLS procedure with constant variance data for $\alpha = 5$

α	$\vec{\theta}_{\text{init}}$	$\vec{\theta}_0$	$\hat{\theta}^{\text{cv}}_{\text{GLS}}$	$\text{SE}(\hat{\theta}^{\text{cv}}_{\text{GLS}})$	$\hat{\theta}^{\text{tcv}}_{\text{GLS}}$	$\text{SE}(\hat{\theta}^{\text{tcv}}_{\text{GLS}})$
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 2: Estimation using the GLS procedure with constant variance data for $\alpha = 5$

α	$\vec{\theta}_{\text{init}}$	$\vec{\theta}_0$	$\hat{\theta}^{\text{ncv}}_{\text{OLS}}$	$\text{SE}(\hat{\theta}^{\text{ncv}}_{\text{OLS}})$	$\hat{\theta}^{\text{tncv}}_{\text{OLS}}$	$\text{SE}(\hat{\theta}^{\text{tncv}}_{\text{OLS}})$
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 3: Estimation using the OLS procedure with nonconstant variance data for $\alpha = 5$

α	$\vec{\theta}_{\text{init}}$	$\vec{\theta}_0$	$\hat{\theta}^{\text{ncv}}_{\text{GLS}}$	$\text{SE}(\hat{\theta}^{\text{ncv}}_{\text{GLS}})$	$\hat{\theta}^{\text{tncv}}_{\text{GLS}}$	$\text{SE}(\hat{\theta}^{\text{tncv}}_{\text{GLS}})$
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

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

When the OLS method was applied to nonconstant variance data and the GLS method was applied to constant variance data, the residual plots 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 both 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.

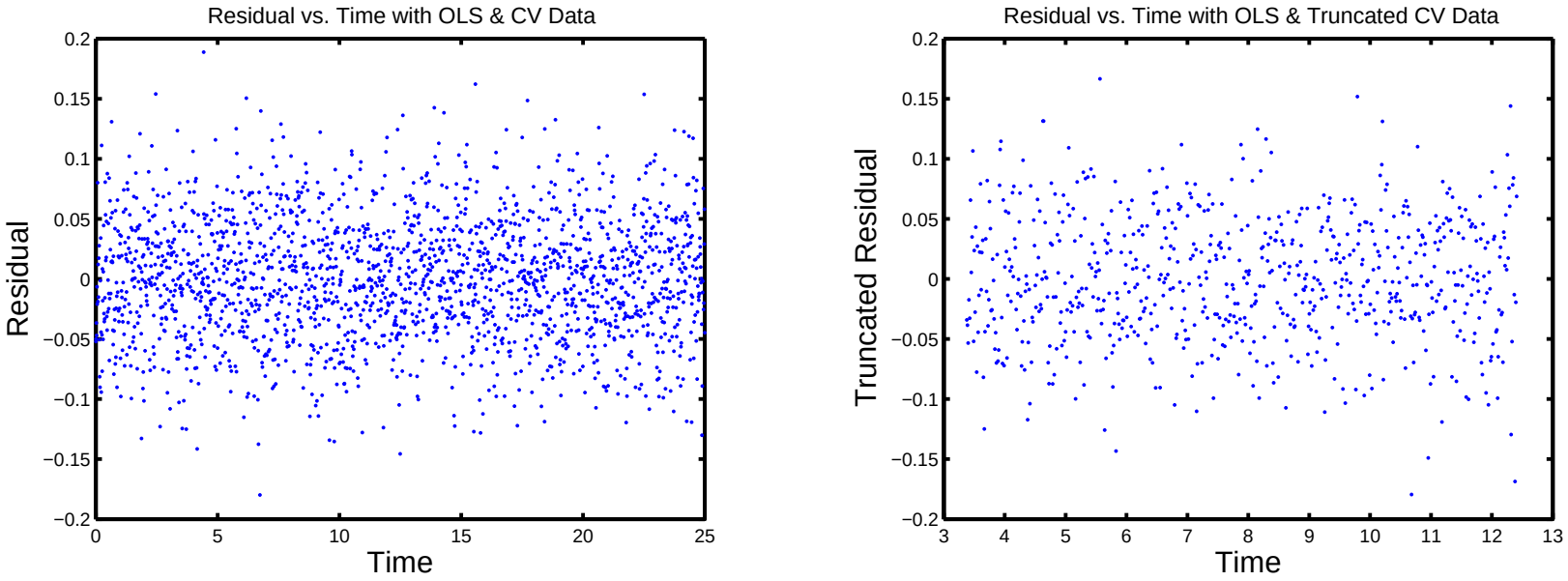


Figure 2: Original and truncated logistic curve for $\hat{\theta}^{cv}_{\text{OLS}}$ with $\alpha = 5$.

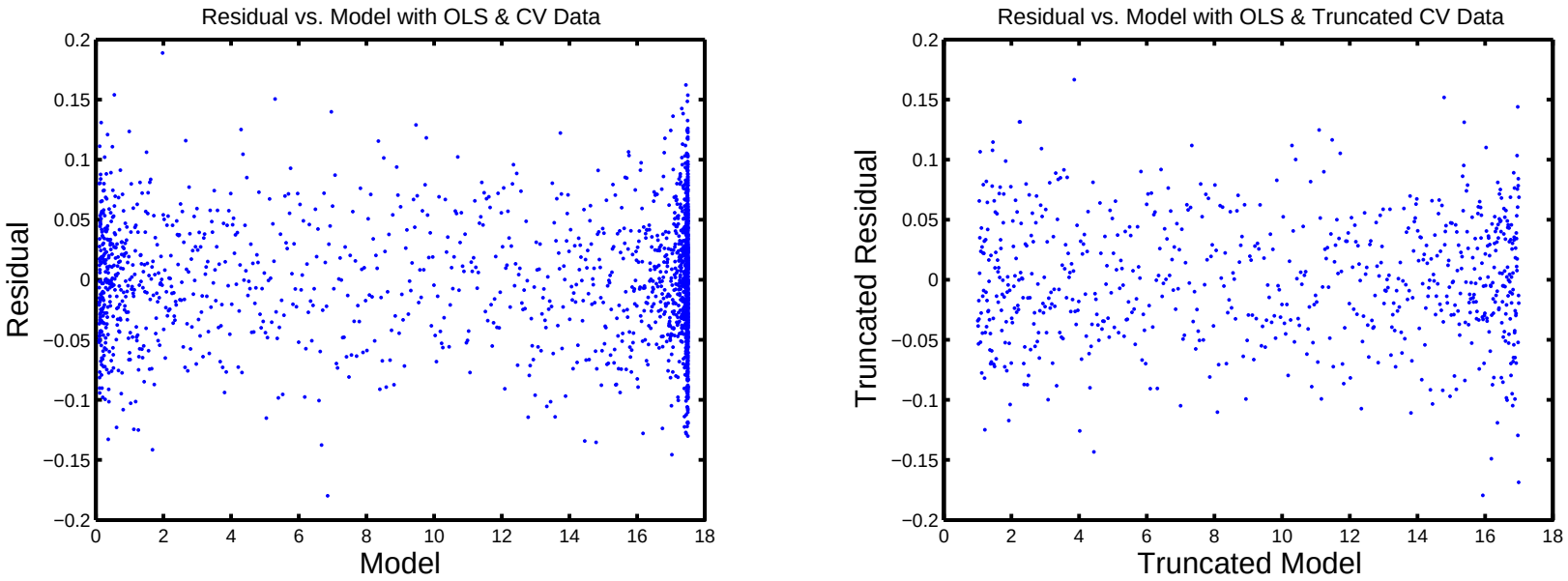


Figure 3: Original and truncated logistic curve for $\hat{\theta}^{cv}_{\text{OLS}}$ with $\alpha = 5$.

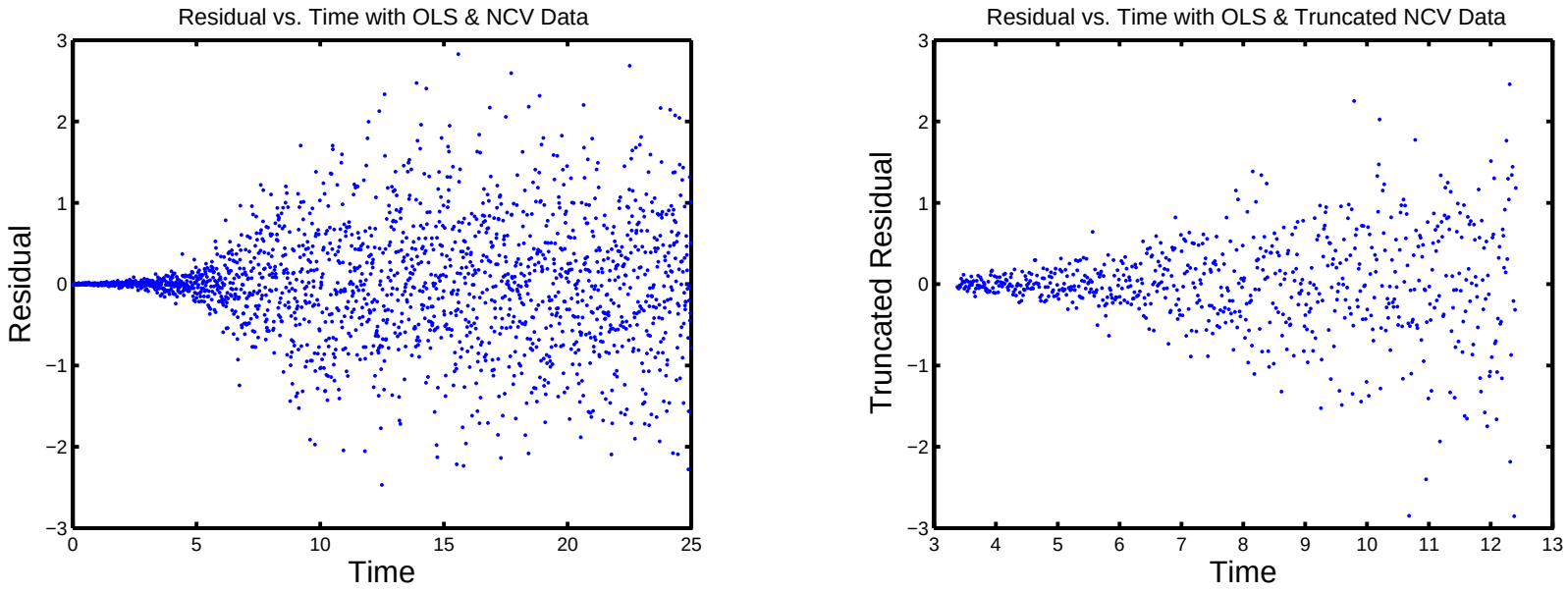


Figure 4: Original and truncated logistic curve for $\hat{\theta}^{ncv}_{\text{OLS}}$ with $\alpha = 5$.

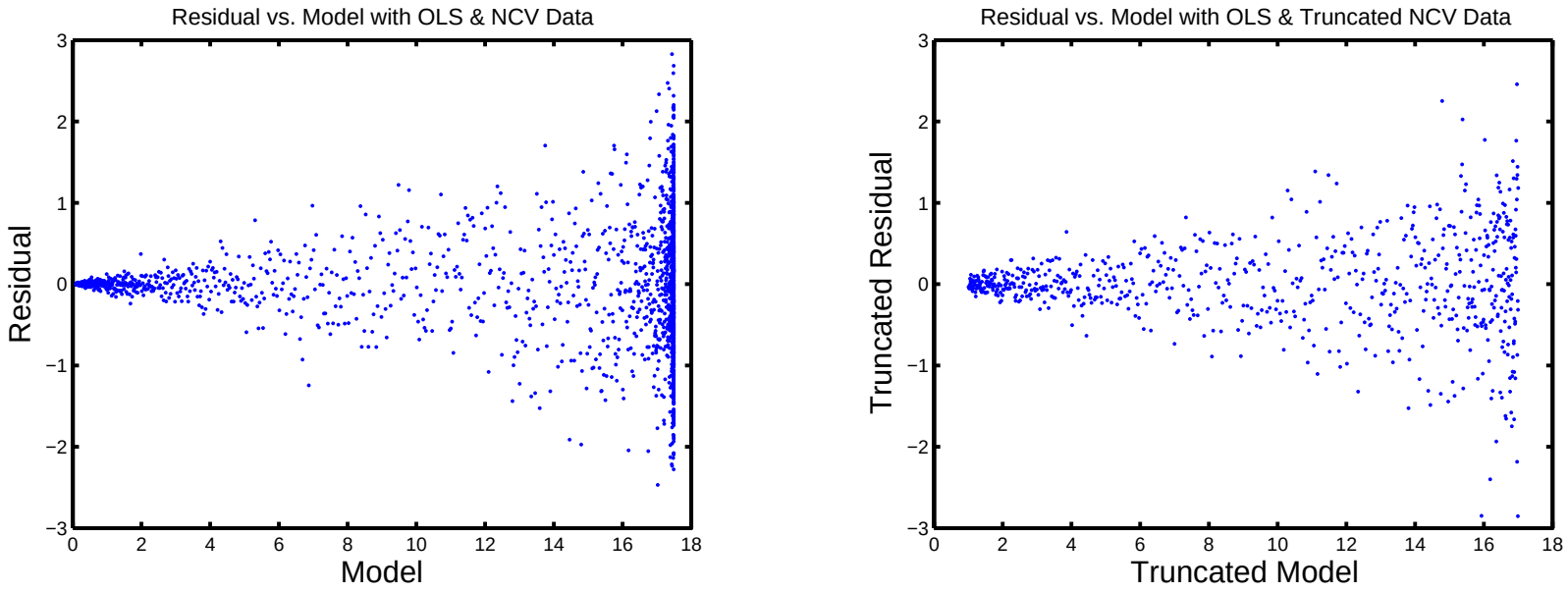


Figure 5: Original and truncated logistic curve for $\hat{\theta}^{ncv}_{\text{OLS}}$ with $\alpha = 5$.

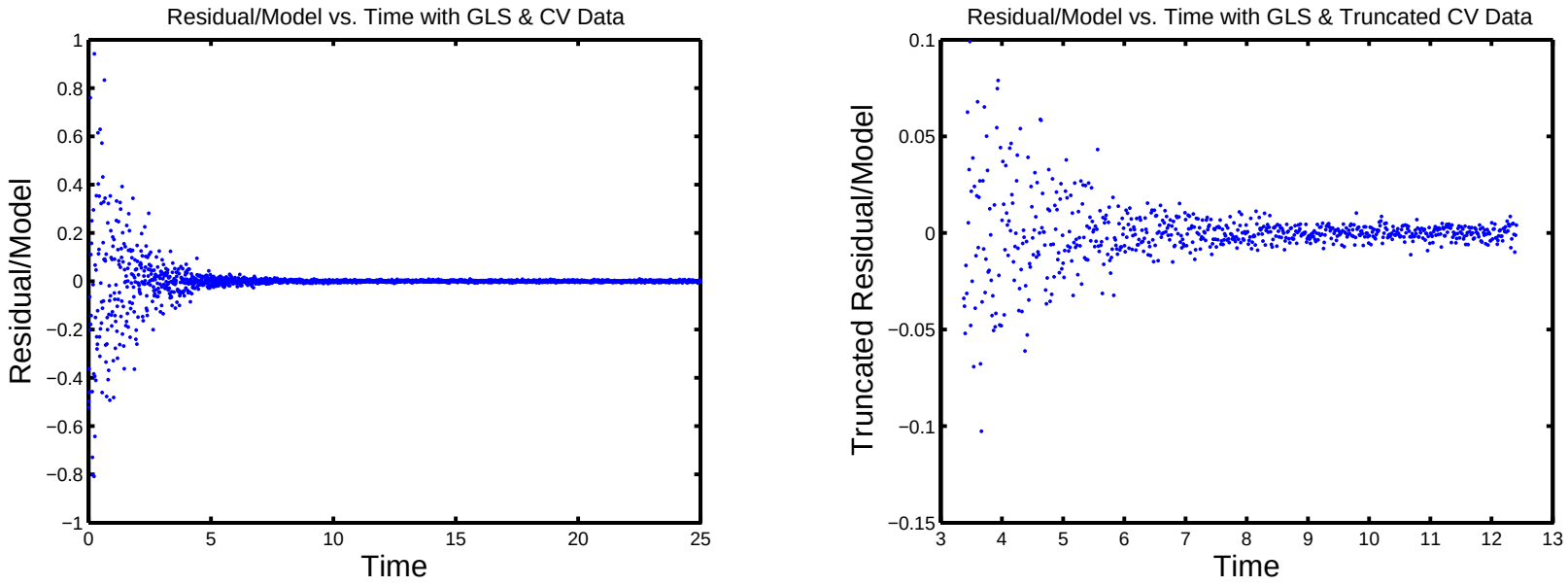


Figure 6: Original and truncated logistic curve for $\hat{\theta}^{cv}_{\text{GLS}}$ with $\alpha = 5$.

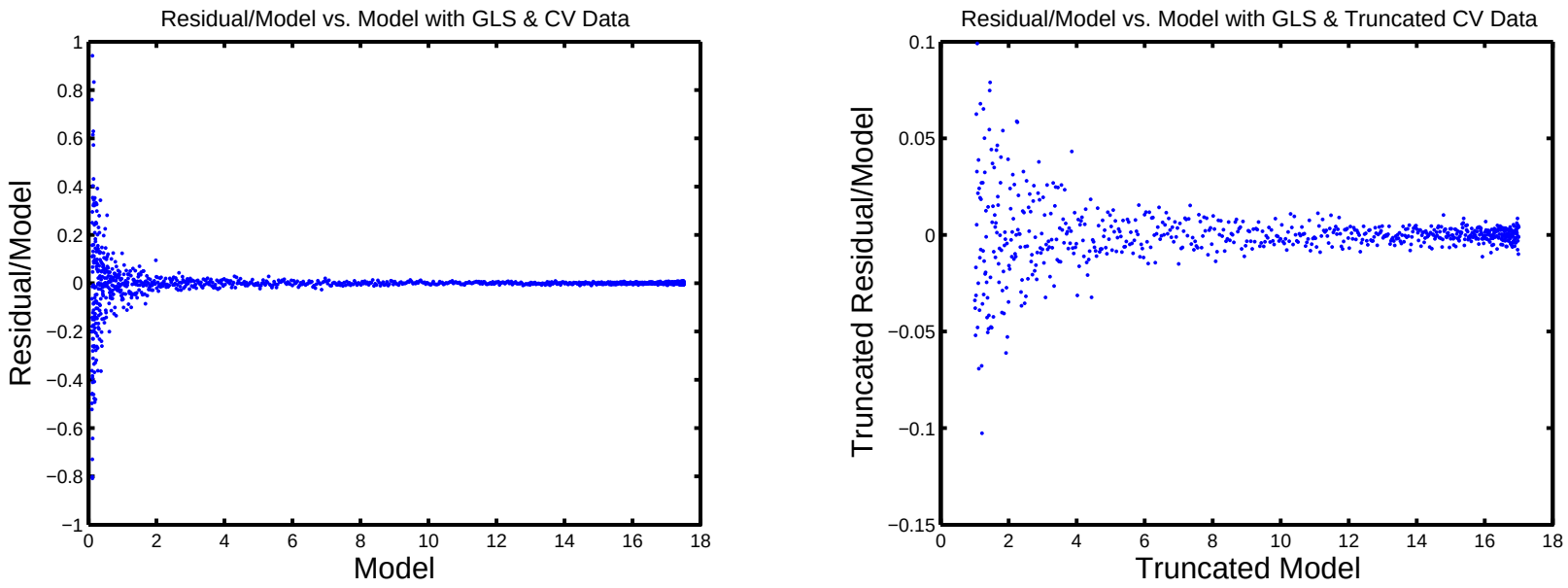


Figure 7: Original and truncated logistic curve for $\hat{\theta}^{cv}_{\text{GLS}}$ with $\alpha = 5$.

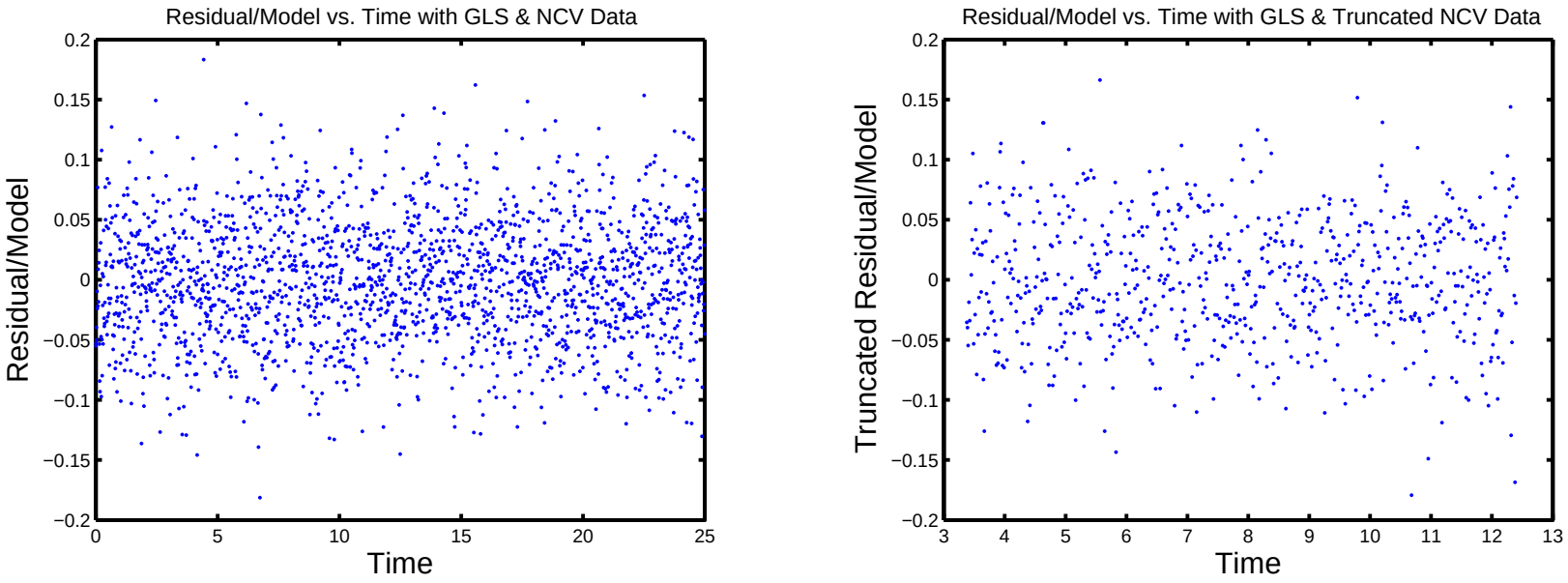


Figure 8: Original and truncated logistic curve for $\hat{\theta}^{ncv}_{\text{GLS}}$ with $\alpha = 5$.

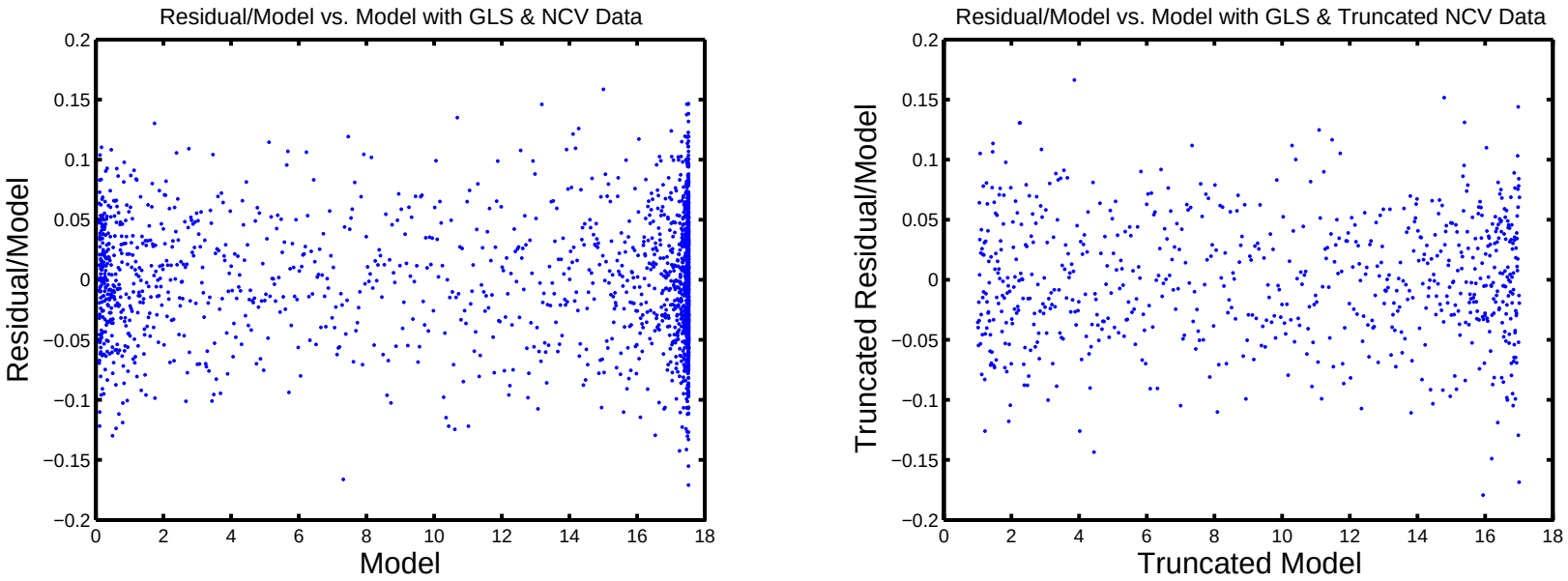
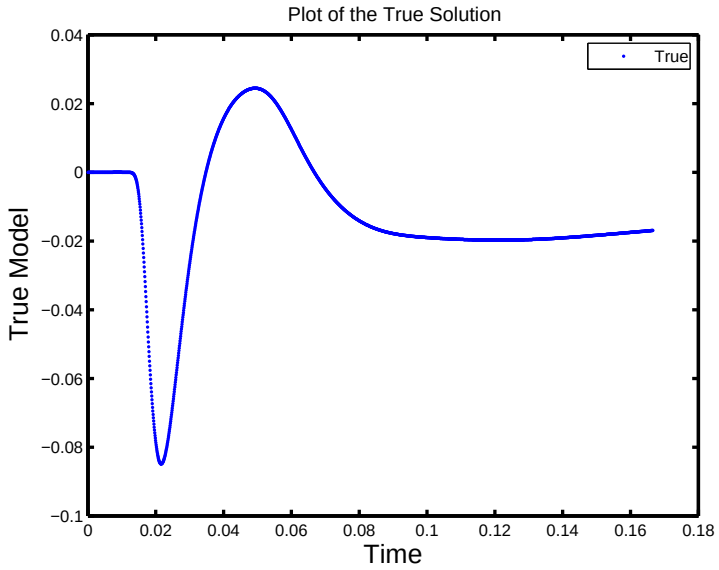


Figure 9: Original and truncated logistic curve for $\hat{\theta}^{ncv}_{\text{GLS}}$ with $\alpha = 5$.

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 [BEG]. Doing so reduces the quality of the estimate 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 stenosis 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.

Shear Wave Model Residual Plots

The residual tests presented worked very well on the logistic model example. We also applied these tests to the shear wave propagation model outlined earlier. The setup for the tests is exactly the same as was performed earlier - the OLS and GLS methods were used to estimate $\vec{\theta}_0$ on data sets with constant and nonconstant variance noise. Also, to avoid the dense vertical lines and division by zero we also considered the truncated data sets as well. Looking at Figure 10 there are high frequency repeated values at $f(t_j, \vec{\theta}_0) = 0$ and $f(t_j, \vec{\theta}_0) = -.02$; thus the truncated data set will not include model values that are near zero or $-.02$ (which simultaneously takes care of dividing by zero).



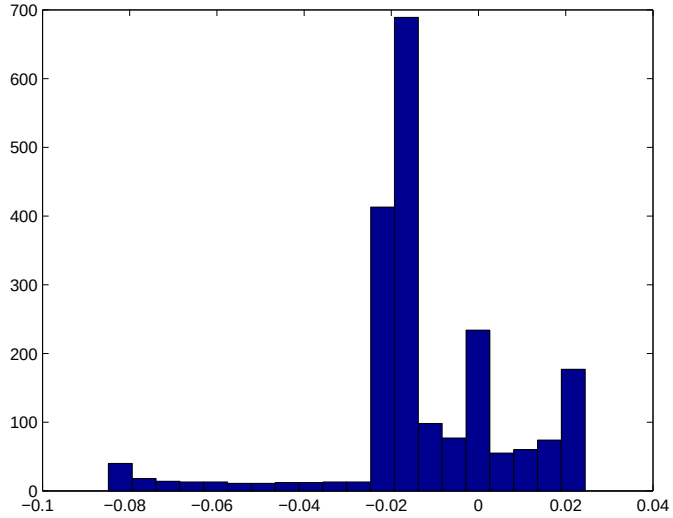


Figure 10: 8^{th} sensor's true plot, and a histogram of the model's values with 20 bins for $\alpha = 3$.

As in the logistic model studied earlier, both methods do a good job at estimating $\vec{\theta}_0$ in the shear wave propagation model, however the error structure was not always correctly specified. That is, the OLS method was applied to nonconstant variance data and the GLS method was applied to constant variance data. Just as in the logistic model example, the residual plots reveal that the error structure was misspecified. For instance, the plot of the residuals for $\hat{\theta}^{\text{ncv}}_{\text{OLS}}$ given in Figure 13 reveals 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 Figure 15 reveals the residuals have a deterministic structure, which indicates the nonconstant variance assumption is suspect. As expected, when the correct error structure is assumed the i.i.d. test and the model dependence test both display a random pattern (Figures 11 and 17).

Figures 12, 14, 16 and 18 also display the residual and modified residual plots with the truncated data sets. Using these removed the dense vertical lines in the plots with $f(t_j, \hat{\theta})$ along the x-axis. More importantly, though, the modified residuals no longer blowup in the truncated data set, making the conclusions regarding the error structure straightforward.

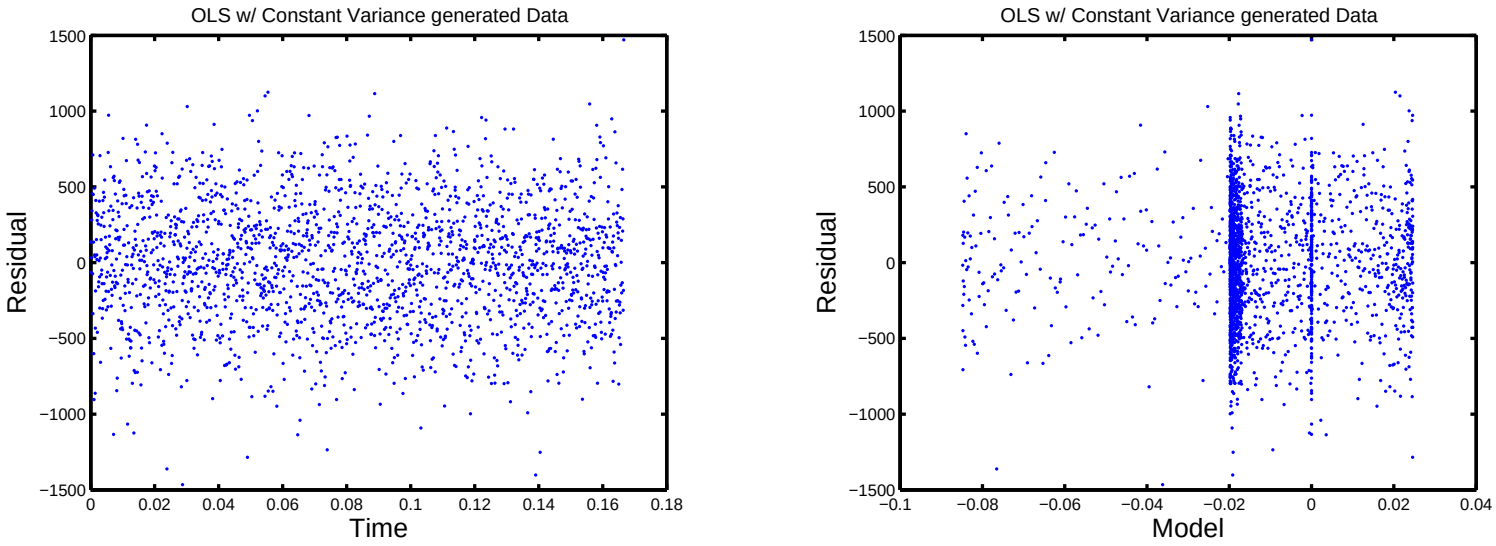


Figure 11: 8^{th} sensor's residual plots for $\hat{\theta}^{cv}_{\text{OLS}}$ with $\alpha = 3$.

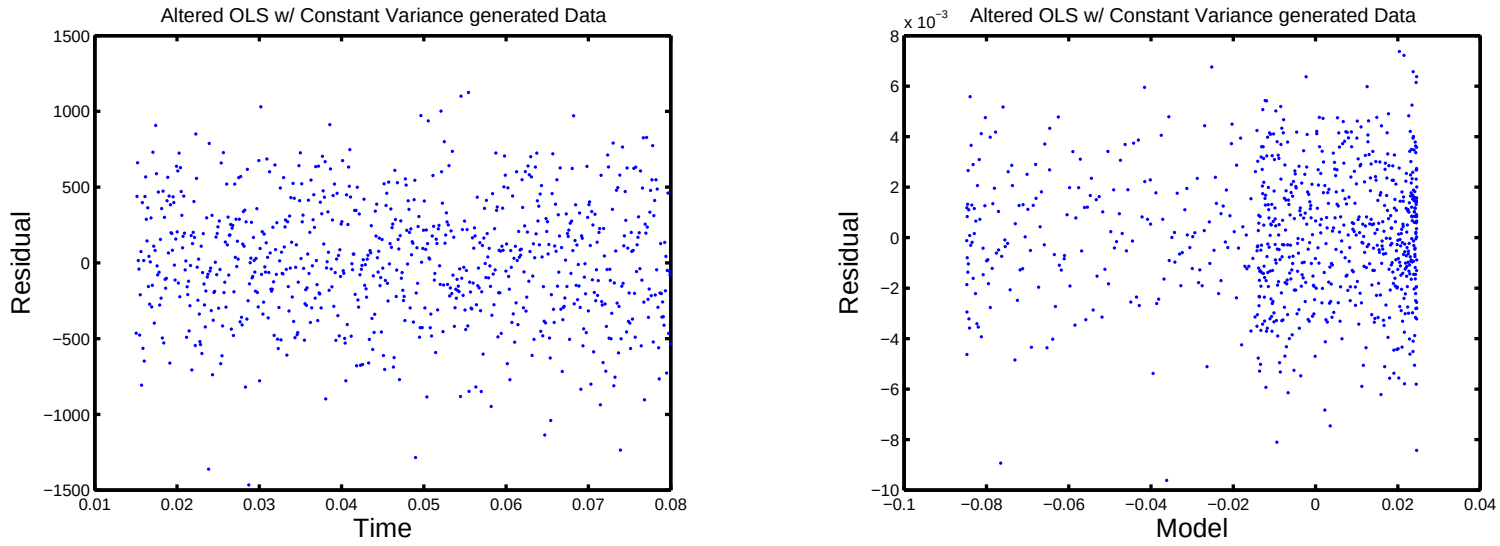


Figure 12: 8^{th} sensor's residual plots for $\hat{\theta}^{cv}_{\text{OLS}}$ with $\alpha = 3$ and truncated data.

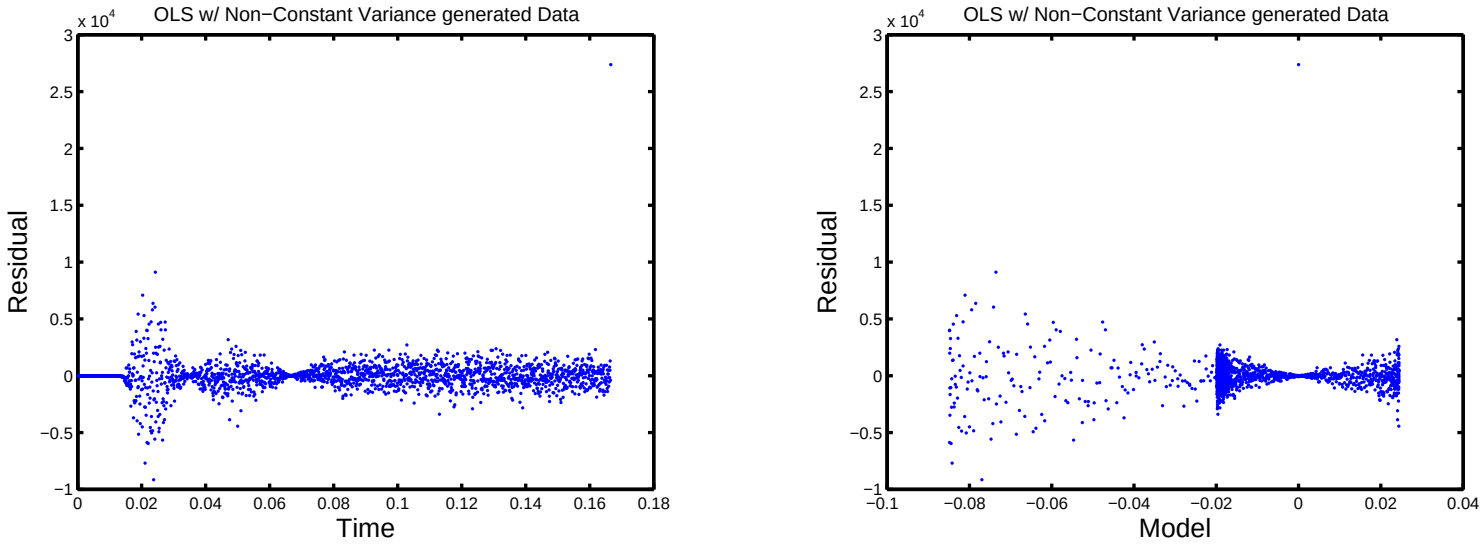


Figure 13: 8^{th} sensor's residual plots for $\hat{\theta}^{ncv}_{\text{OLS}}$ with $\alpha = 3$.

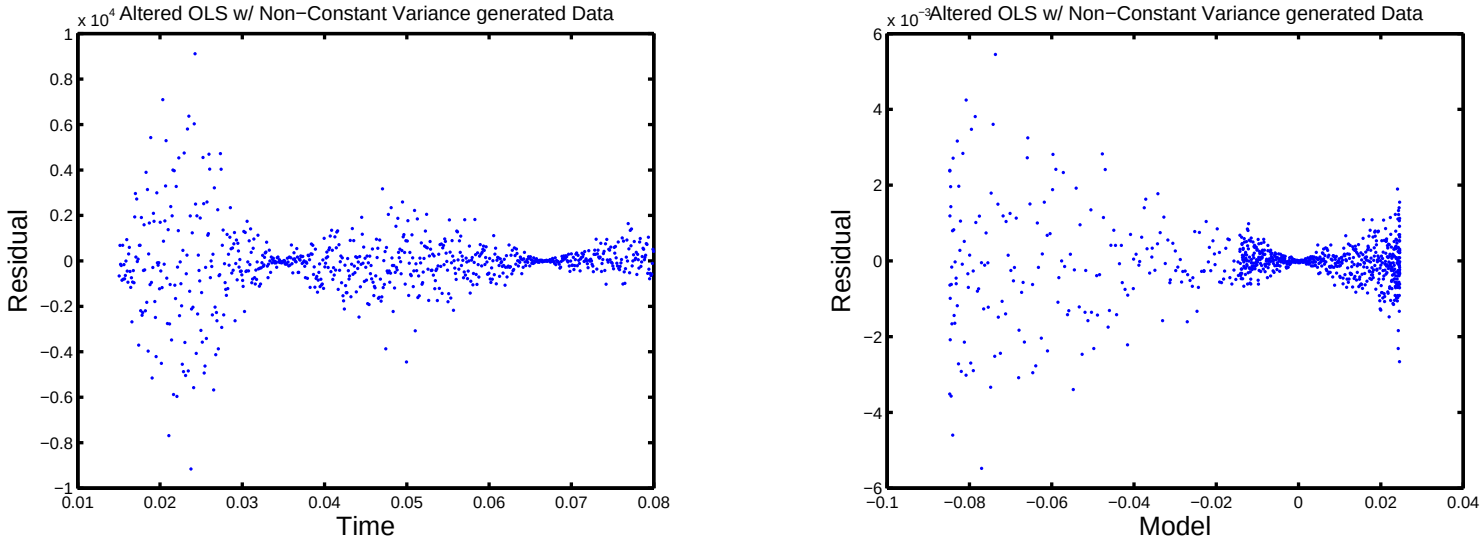


Figure 14: 8^{th} sensor's residual plots for $\hat{\theta}^{ncv}_{\text{OLS}}$ with $\alpha = 3$ and truncated data.

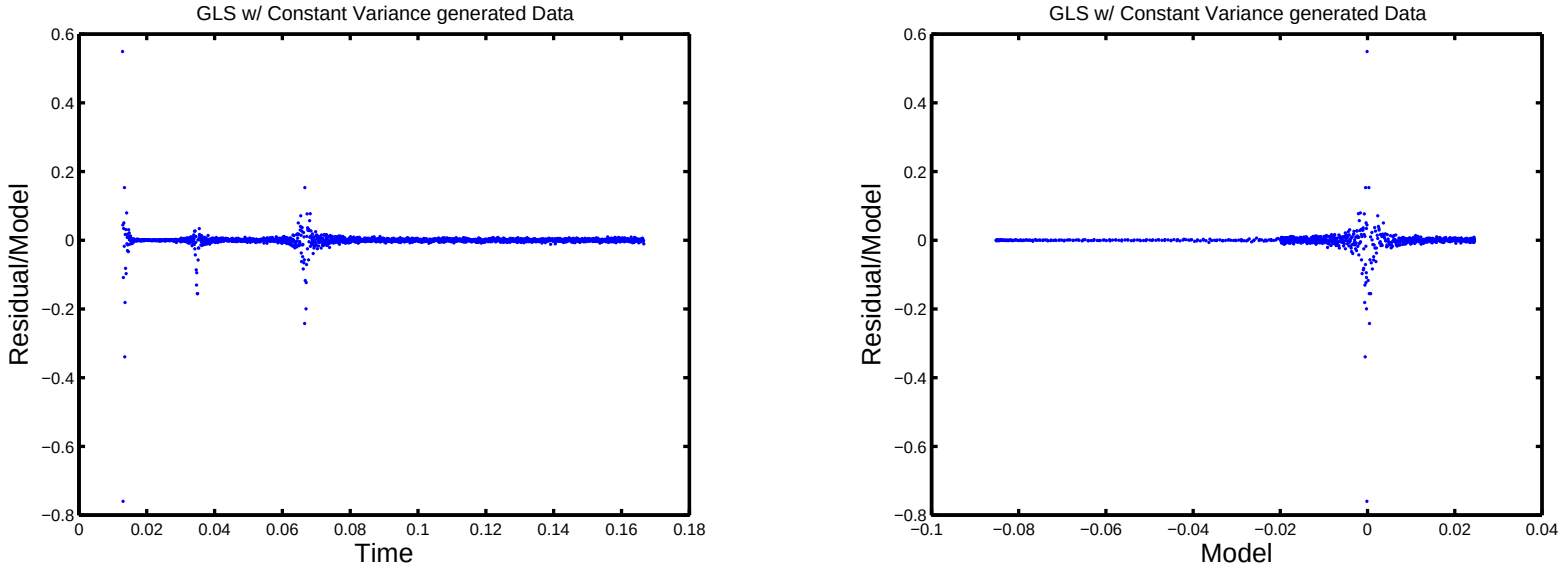


Figure 15: 8^{th} sensor's residual plots for $\hat{\theta}^{cv}_{\text{GLS}}$ with $\alpha = 3$.

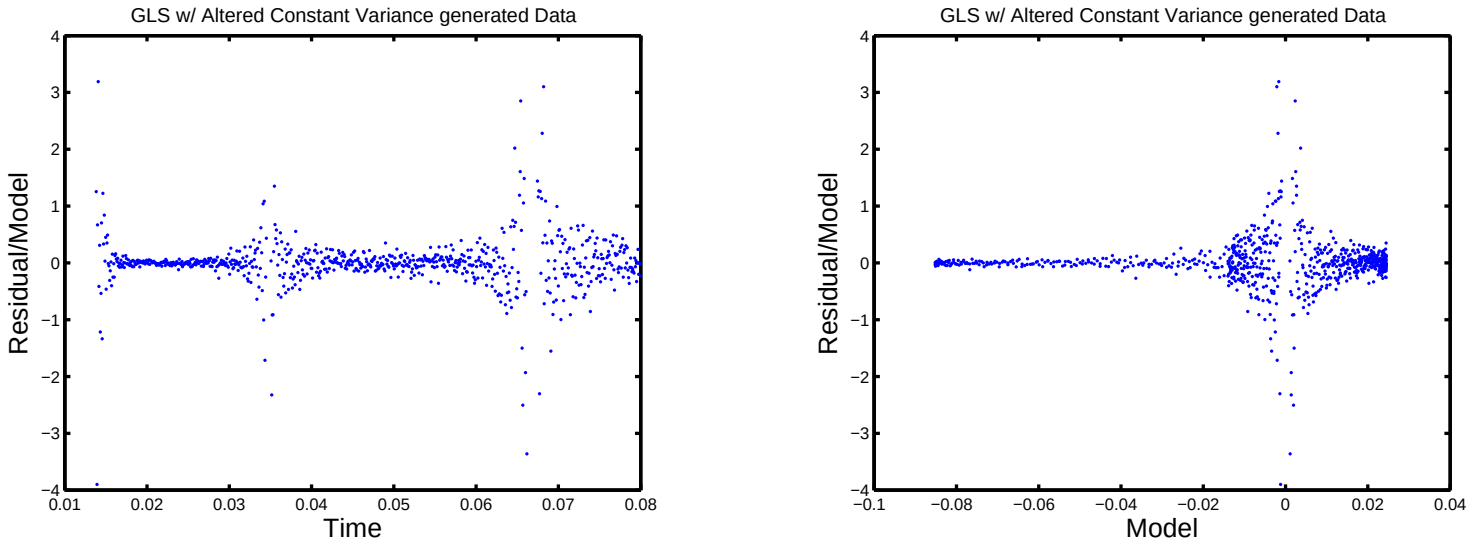


Figure 16: 8^{th} sensor's residual plots for $\hat{\theta}^{cv}_{\text{GLS}}$ with $\alpha = 3$ and truncated data.

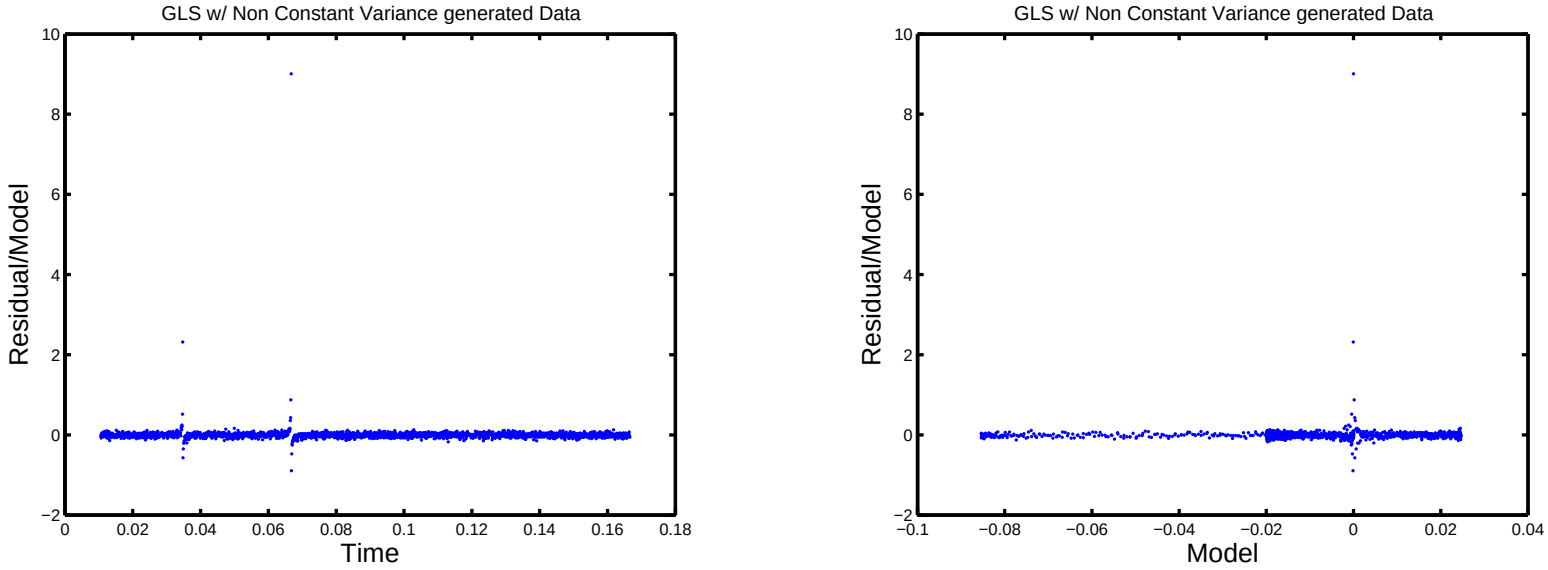


Figure 17: 8^{th} sensor's residual plots for $\hat{\theta}^{ncv}_{\text{GLS}}$ with $\alpha = 3$.

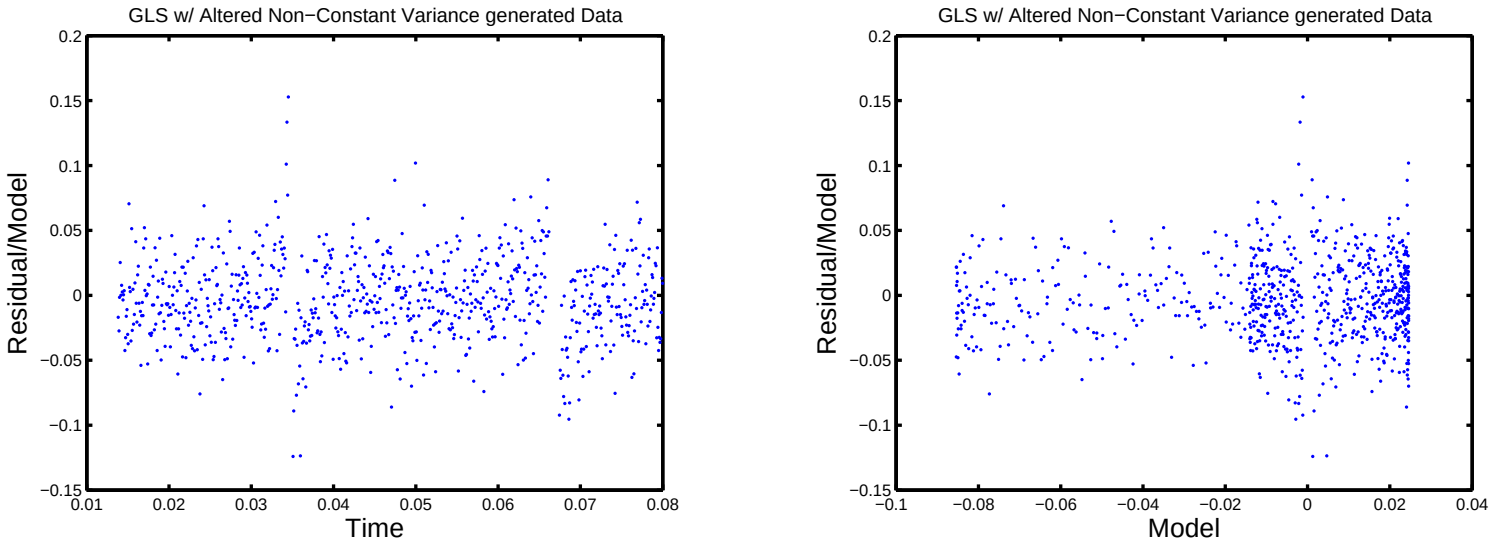


Figure 18: 8^{th} sensor's residual plots for $\hat{\theta}^{ncv}_{\text{GLS}}$ with $\alpha = 3$ and truncated data.

References

[BD] H. T. Banks and M. Davidian, SAMSI MA/ST 810 Notes.

[BEG] HTB, S.L. Ernstberger and S.L.Grove, Standard errors and confidence intervals in inverse problems: sensitivity and associated pitfalls, *J. Inv. Ill-posed Problems* **15** (2006), 1–18.

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

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

[G] A. R. Gallant, *Nonlinear Statistical Models*, John Wiley & Sons, Inc., New York, 1987.

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

[SeWi] G. A. F. Seber and C. J. Wild, *Nonlinear Regression*, John Wiley & Sons, Inc., New York, 1989.