

Analysis of Statistical Standard Errors in Inverse Problems

H.T. Banks, Stacey L. Ernstberger and Sarah L. Grove

Center for Research in Scientific Computation
North Carolina State University
Raleigh, North Carolina 27695-8205

July 25, 2007

NC STATE UNIVERSITY

*Center for Research
in Scientific Computation
North Carolina State University*

Outline

- Introduction
 - Statistical Review
- Analytical Derivation
 - Regional Partitions
- Implementation
- Results
 - Regional Partitions
 - Standard Errors
- Conclusions

Introduction

- Goal: To illustrate use and possible pitfalls in statistical inverse problem analysis of sensitivity equations and related standard errors.
- When data is sampled from a region near a steady state of a curve, the optimized parameters will not necessarily improve.
- The standard errors for estimated parameters may increase with more data points, contrary to intuition. To explain this, we will look at a specific example.

Statistical Review

We assume that n scalar longitudinal observations are represented by the statistical model

$$Y_j \equiv f_j(\beta_0) + \epsilon_j, \quad j = 1, 2, \dots, n.$$

Thus we seek to use data $\{y_j\}$ for the observation process $\{Y_j\}$ with the model to seek a value $\hat{\beta}^n$ that minimizes

$$J_n(\beta) = \sum_{j=1}^n |f_j(\beta) - y_j|^2.$$

Since Y_j is a random variable, we have that the estimator $\hat{\beta}_{OLS}^n$ is also a random variable with a distribution called the *sampling distribution*. For large n , the sampling distribution approximately satisfies

$$\hat{\beta}_{OLS}^n(Y) \sim \mathcal{N}_p(\beta_0, \sigma_0^2[\chi^T(\beta_0)\chi(\beta_0)]^{-1}) := \mathcal{N}_p(\beta_0, \Sigma_0),$$

where $\chi(\beta) = F_\beta(\beta)$ is the $n \times p$ sensitivity matrix with elements

$$\chi_{jk}(\beta) = \frac{\partial f_j(\beta)}{\partial \beta_k} \quad \text{and} \quad F_\beta(\beta) \equiv (f_{1\beta}(\beta), \dots, f_{n\beta}(\beta))^T.$$

$F_\beta(\beta)$ can be found using difference quotients, sensitivity equations, or direct analytic calculations (in some cases).

Since β_0 , σ_0 are not known, we follow standard practice and use the approximation

$$\Sigma_0 \approx \Sigma(\hat{\beta}^n) = \hat{\sigma}^2[\chi^T(\hat{\beta}^n)\chi(\hat{\beta}^n)]^{-1}$$

where $\hat{\beta}^n$ is the parameter estimate obtained, and the approximation $\hat{\sigma}^2$ to σ_0^2 is given by

$$\sigma_0^2 \approx \hat{\sigma}^2 = \frac{1}{n-p} \sum_{j=1}^n |f_j(\hat{\beta}^n) - y_j|^2.$$

Standard errors to be used in confidence interval calculations are thus given by $SE_k(\hat{\beta}^n) = \sqrt{\Sigma_{kk}(\hat{\beta}^n)}$, $k = 1, 2, \dots, p$ and consequently we define the confidence level parameters associated with the estimated parameters so that

$$P\{\hat{\beta}_k^n - t_{1-\alpha/2}SE_k(\hat{\beta}^n) < \beta_k^n < \hat{\beta}_k^n + t_{1-\alpha/2}SE_k(\hat{\beta}^n)\} = 1 - \alpha.$$

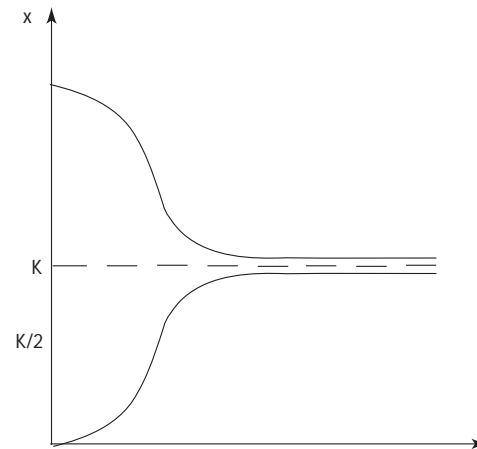
Example

Consider the logistic growth population model

$$\frac{dx}{dt} = rx \left(1 - \frac{x}{K}\right)$$

where K is the carrying capacity and r is the intrinsic growth rate. The solution is given by

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



Analytical Derivation

We will examine the problem

$$\dot{x} = ax - bx^2,$$

where

$$\begin{aligned} x(t) &= \frac{a/b}{1 + \left(\frac{a/b}{x_0} - 1\right)e^{-at}} \\ &= \frac{a}{b + ke^{-at}} \end{aligned}$$

as $k = \frac{a}{x_0} - b$. Here $x(t)$ has an asymptote at $\frac{a}{b} = K$.

We begin with an ordinary least squares problem for $\beta = (a, b, x_0)$ where

$$X = \frac{\partial x}{\partial \beta} = \begin{pmatrix} \frac{\partial x(t_1)}{\partial a} & \frac{\partial x(t_1)}{\partial b} & \frac{\partial x(t_1)}{\partial x_0} \\ \vdots & \vdots & \vdots \\ \frac{\partial x(t_n)}{\partial a} & \frac{\partial x(t_n)}{\partial b} & \frac{\partial x(t_n)}{\partial x_0} \end{pmatrix}$$

with the covariance matrix

$$\begin{aligned} \Sigma &= \hat{\sigma}^2 (X^T X)^{-1} \\ &= \frac{1}{n-3} \sum_{j=1}^n |y_j - f(t_j, \hat{\beta}_n)|^2 (X^T X)^{-1} \end{aligned}$$

and the estimate of the standard error

$$\text{SE}_k = \sqrt{\hat{\sigma}^2 (X^T X)^{-1}_{kk}}, \quad k = 1, 2, 3.$$

Let R_0 be the region where $t \in [0, \tau_1]$, R_1 be the region corresponding to $t \in [\tau_1, \tau_2]$, and R_2 where $t \in [\tau_2, \infty)$.

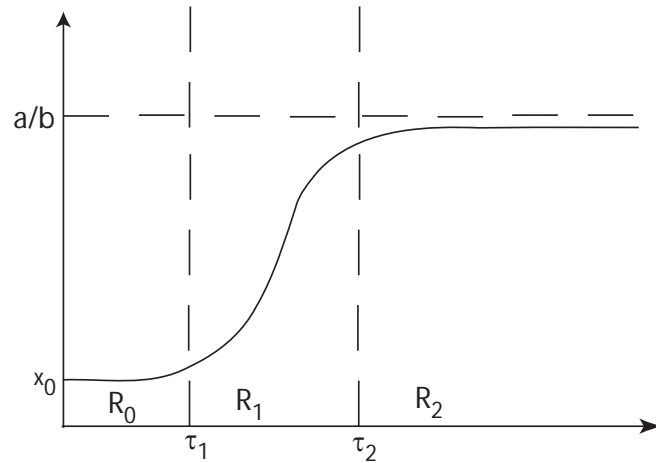


Figure 1: Partition of solution curve into distinct regions.

In order to divide the problem

$$\dot{x} = ax - bx^2$$

into separate regions, we consider the solution

$$x = \frac{a}{b + ke^{-at}}$$

and then examine the partial derivatives $\frac{\partial x}{\partial a}$, $\frac{\partial x}{\partial b}$, and $\frac{\partial x}{\partial x_0}$.

- As $t \rightarrow \infty$, which corresponds to region R_2 , we see that $x(t) \rightarrow \frac{a}{b}$.
- Also, $x(t) \rightarrow \frac{a}{b+k}$ as $t \rightarrow 0$, which corresponds to region R_0 .
- We will denote $x(0) = x_0 = \frac{a}{b+k}$.

We have the following partial derivatives

$$\frac{\partial x}{\partial a} = \frac{b + (atk - b)e^{-at}}{(b + ke^{-at})^2},$$

$$\frac{\partial x}{\partial b} = \frac{-a(1 - e^{-at})}{(b + ke^{-at})^2},$$

$$\frac{\partial x}{\partial x_0} = \frac{a^2 e^{-at}}{x_0^2 (b + ke^{-at})^2},$$

which are $n \times 1$ vectors for $t = t_1, \dots, t_n$. The matrix X is composed of these vectors and we obtain the following matrix,

$$X^T X = \sum_{j=1}^n \begin{pmatrix} \left(\frac{\partial x(t_j)}{\partial a} \right)^2 & \frac{\partial x(t_j)}{\partial a} \frac{\partial x(t_j)}{\partial b} & \frac{\partial x(t_j)}{\partial a} \frac{\partial x(t_j)}{\partial x_0} \\ \frac{\partial x(t_j)}{\partial b} \frac{\partial x(t_j)}{\partial a} & \left(\frac{\partial x(t_j)}{\partial b} \right)^2 & \frac{\partial x(t_j)}{\partial b} \frac{\partial x(t_j)}{\partial x_0} \\ \frac{\partial x(t_j)}{\partial x_0} \frac{\partial x(t_j)}{\partial a} & \frac{\partial x(t_j)}{\partial x_0} \frac{\partial x(t_j)}{\partial b} & \left(\frac{\partial x(t_j)}{\partial x_0} \right)^2 \end{pmatrix}.$$

If we sample data from R_0 , where $t_j < \tau_1$ for $j = 1 \dots n$, we have

$$\frac{\partial x(t_j)}{\partial a} \approx 0, \quad \frac{\partial x(t_j)}{\partial b} \approx 0, \quad \frac{\partial x(t_j)}{\partial x_0} \approx 1.$$

Also notice that

$$\sum_{j=1}^n \frac{\partial x(t_j)}{\partial x_0} \frac{\partial x(t_j)}{\partial x_0} = \sum_{j=1}^n 1 = n$$

for $t_j < \tau_1$ in R_0 . Hence, in this region,

$$X^T X \approx \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & n \end{pmatrix}.$$

Next, consider region R_2 , where $t_j > \tau_2$ for $j = 1 \dots n$. Notice that in R_2 ,

$$\frac{\partial x(t_j)}{\partial a} \approx \frac{1}{b}, \quad \frac{\partial x(t_j)}{\partial b} \approx -\frac{a}{b^2}, \quad \frac{\partial x(t_j)}{\partial x_0} \approx 0,$$

and hence,

$$X^T X \approx \sum_{j=1}^n \begin{pmatrix} \frac{\partial x(t_j)}{\partial a} & \frac{\partial x(t_j)}{\partial a} & \frac{\partial x(t_j)}{\partial a} & \frac{\partial x(t_j)}{\partial b} & 0 \\ \frac{\partial x(t_j)}{\partial b} & \frac{\partial x(t_j)}{\partial a} & \frac{\partial x(t_j)}{\partial b} & \frac{\partial x(t_j)}{\partial b} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} = n \begin{pmatrix} \frac{1}{b^2} & -\frac{a}{b^3} & 0 \\ -\frac{a}{b^3} & \frac{a^2}{b^4} & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Notice that the second column of $X^T X$ is a scalar multiple of the first column, which means that we should be able to estimate the ratio $\frac{a}{b}$ but not the individual parameters.

Implementation

We create a simulated data set, y_j , $j = 1, \dots, n$, using the analytical solution with a specific β_0 . We want to solve an inverse problem. The cost function uses *ode15s* to approximate the solution and returns

$$J(\beta) = \sum_{j=1}^n |y_j - f(t_j; \beta)|^2$$

where $f(t; \beta) = x(t; a, b, x_0)$ is the approximation to the solution. We use the MATLAB function *fminsearch* to optimize β in order to obtain the minimized cost $J(\hat{\beta}_n)$. Here, $\hat{\beta}_n$ represents the optimized value of β over n data points.

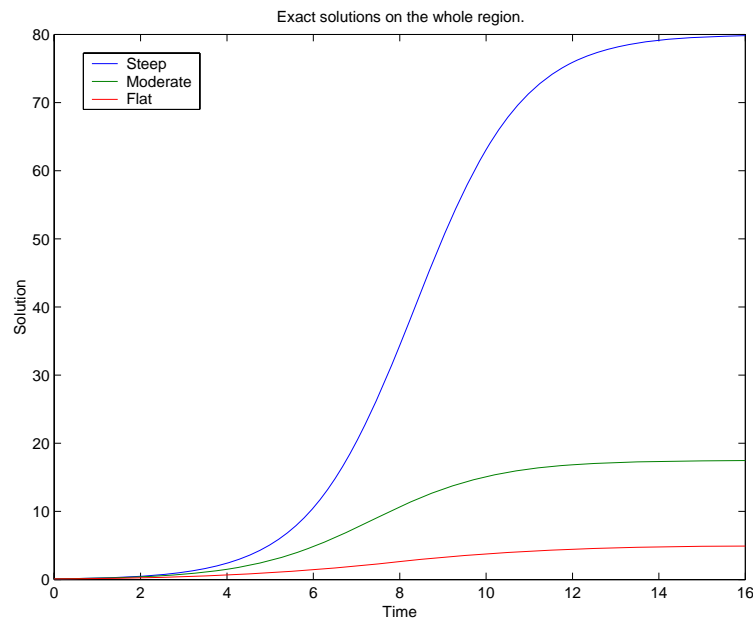


Figure 2: Three sets of simulated data: a relatively flat curve with $\beta_0 = (0.5, 0.1, 0.1)$, a moderately sloped curve with $\beta_0 = (0.7, 0.04, 0.1)$, and a steep curve with $\beta_0 = (0.8, 0.01, 0.1)$.

Define $R_0 \triangleq [0, 2]$, $R_1 \triangleq [2, 12]$, $R_2 \triangleq [12, 16]$.

Results

- Regional Partitions
 - Regions R_0, R_1, R_2
- Standard Errors
 - Regions $R_0 \cup R_1$ and $R_1 \cup R_2$
 - Simulated Noise

Region R_0

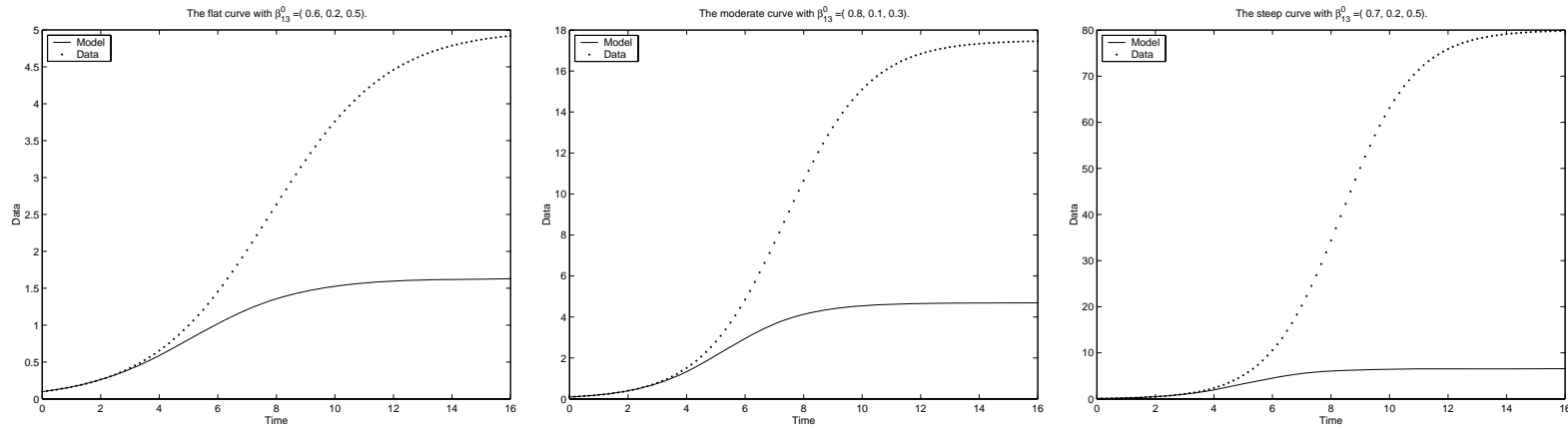


Figure 3: Simulated data plotted with the solution obtained from the estimated parameters in region R_0 for the a) flat curve with $\beta^0 = (0.6, 0.2, 0.5)$ b) moderate curve with $\beta^0 = (0.8, 0.1, 0.3)$ c) steep curve with $\beta^0 = (0.7, 0.2, 0.5)$.

Region R_1

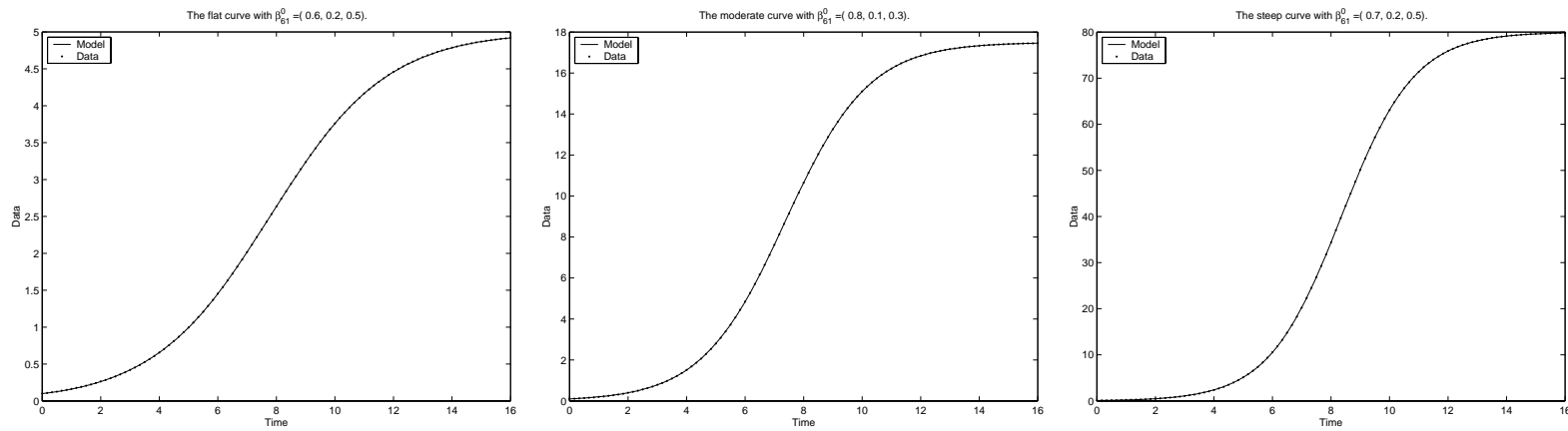


Figure 4: Simulated data plotted with the solution obtained from the estimated parameters in region R_1 for the a) flat curve with $\beta^0 = (0.6, 0.2, 0.5)$ b) moderate curve with $\beta^0 = (0.8, 0.1, 0.3)$ c) steep curve with $\beta^0 = (0.7, 0.2, 0.5)$.

Region R_2

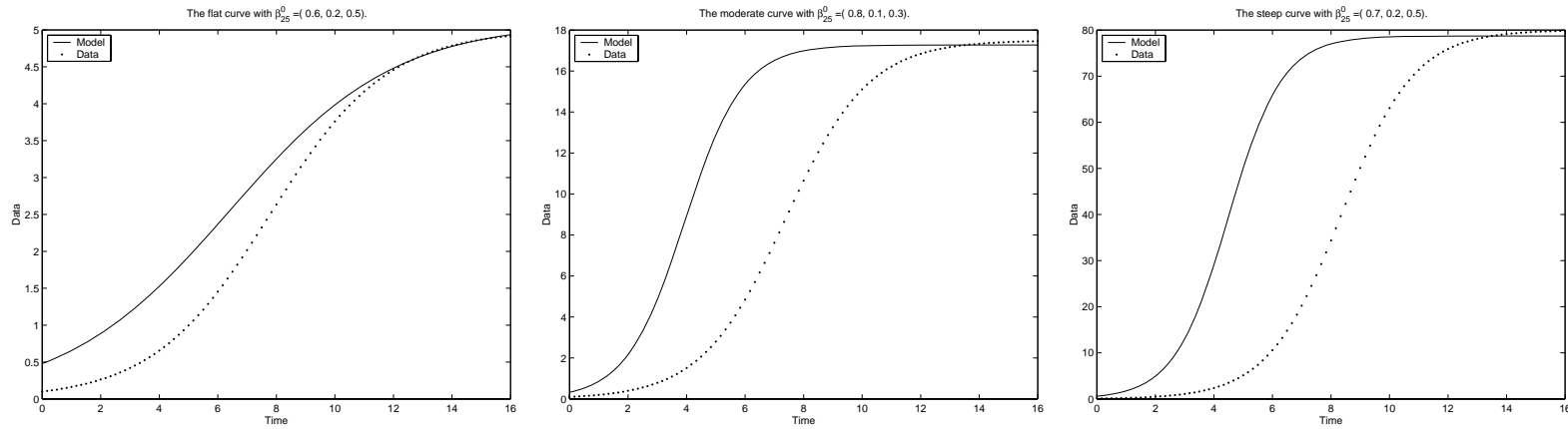


Figure 5: Simulated data plotted with the solution obtained from the estimated parameters in region R_2 for the a) flat curve with $\beta^0 = (0.6, 0.2, 0.5)$ b) moderate curve with $\beta^0 = (0.8, 0.1, 0.3)$ c) steep curve with $\beta^0 = (0.7, 0.2, 0.5)$.

Regions R_0 and R_1

The moderate curve with $\beta_0 = (0.7, 0.04, 0.1)$ using $\beta^0 = (0.8, 0.1, 0.3)$.

n_0	n_1	$\hat{\beta}_n$	Standard Errors
25	0	(0.7244,0.1414,0.0992)	(0.0049, 0.0011, 0.0025)
49	0	(0.7241,0.1401,0.0992)	(0.0094, 0.0020, 0.0049)
25	24	(0.6998,0.0397,0.1)	($2.2059e^{-4}$, $1.4424e^{-5}$, $1.5696e^{-4}$)

Table 1: The standard errors for a , b , and x_0 for the optimized $\hat{\beta}_n$.

Regions R_1 and R_2

The moderate curve with $\beta_0 = (0.7, 0.04, 0.1)$ using $\beta^0 = (0.8, 0.1, 0.3)$.

n_1	n_2	$\hat{\beta}_n$	Standard Errors
0	25	(1.0057, 0.0582, 0.3249)	(1.0794, 0.0663, 1.3714)
0	49	(1.0058, 0.0582, 0.3249)	(2.1739, 0.1335, 2.7618)
24	25	(0.6996, 0.04, 0.1003)	($5.0686e^{-3}$, $3.3366e^{-4}$, $3.6103e^{-3}$)

Table 2: The standard errors for a , b , and x_0 for the optimized $\hat{\beta}_n$.

Simulated Noise

We consider two sets of simulated noisy data for the moderate curve. One set corresponds to R_0 and is perturbed using $\sigma_0 = 0.005$ and the other corresponds to R_2 and has $\sigma_0 = 0.5$. These values represent an added noise level of approximately ten percent at the lowest bound in each region.

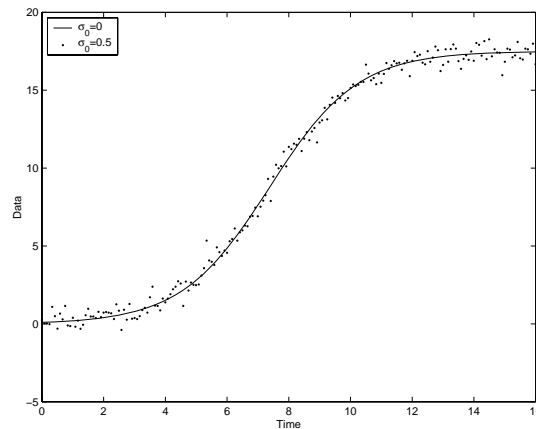


Figure 6: Noisy data using $\sigma_0 = 0.5$ where $\beta^0 = (0.8, 0.1, 0.3)$.

Regions R_0 and R_1 with Noise

The moderate curve with $\beta_0 = (0.7, 0.04, 0.1)$ using $\beta^0 = (0.8, 0.1, 0.3)$.

n_0	n_1	$\hat{\beta}_n$	Standard Errors
25	0	(0.7244,0.1455,0.0988)	(0.052, 0.0116, 0.0269)
49	0	(0.7245,0.1452,0.0988)	(0.1060, 0.0236, 0.0548)
25	24	(0.7086,0.0517,0.0988)	(0.0497, 0.0041, 0.033)
49	48	(0.703,0.0435,0.0994)	(0.091, 0.0065, 0.0631)
0	25	(0.7058,0.0491,0.0994)	(0.0133, 0.0011, 0.009)
0	49	(0.6958,0.0362,0.1008)	(0.0252, 0.0015, 0.0184)

Table 3: The standard errors for the optimized $\hat{\beta}_n$ with $\sigma = 0.005$ using data sampled from the interval $[0, 4]$.

Regions R_1 and R_2 with Noise

The moderate curve with $\beta_0 = (0.7, 0.04, 0.1)$ using $\beta^0 = (0.8, 0.1, 0.3)$.

n_1	n_2	$\hat{\beta}_n$	Standard Errors
0	25	(1.0058,0.0582,0.3249)	(2.7953, 0.1716, 3.5515)
0	49	(1.0057,0.0582,0.3249)	(6.3606, 0.3906, 8.0808)
24	25	(0.5286,0.0297,0.4397)	(3.7015, 0.2516, 10.375)
48	49	(0.673,0.0384,0.123)	(10.2462, 0.679, 8.8898)
25	0	(0.5175,0.0281,0.4307)	(1.0073, 0.0671, 2.8477)
49	0	(0.4713,0.0249,0.5921)	(2.1167, 0.1414, 8.0601)

Table 4: The standard errors for the optimized $\hat{\beta}_n$ with $\sigma = 0.5$ using data sampled from the interval $[8, 16]$.

Conclusions

- Sampling data from regions near steady states alone may not provide sufficient information to obtain all parameters with any degree of accuracy (lack of sensitivity).
- Including additional data from the same regions near steady states will only increase the standard errors without significant improvement to the optimized parameters (no new information from new sample points).
- When extra data points are sampled in a way that provides new information from a region where the solution is sensitive to the parameters, then the parameter estimation will be improved (i.e., region R_1 from this problem).

References

- [BN] H. T. Banks and H. K. Nguyen, Sensitivity of dynamical system to Banach space parameters, CRSC Tech Rep., CRSC-TR05-13, N.C. State University, February, 2005; *J. Math Anal. Appl.*, in press.
- [BEG] H.T. Banks, Stacey L. Ernstberger and Sarah L. Grove, Standard errors and confidence intervals in inverse problems: Sensitivity and associated pitfalls, CRSC-TR06-10, March, 2006; *J. Inverse and Ill-posed Problems*, **15** (2007), 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: 633-643, 1969.
- [K] M. Kot, *Elements of Mathematical Ecology*, Cambridge University Press, 2001, p. 7-9.
- [SeWi] G. A. F. Seber and C. J. Wild, *Nonlinear Regression*, John Wiley & Sons, Inc., New York, 1989.

Region R_1 with $\sigma_0 = 0.005$

The moderate curve with $\beta_0 = (0.7, 0.04, 0.1)$ using $\beta^0 = (0.8, 0.1, 0.3)$.

n_1	Interval	$\hat{\beta}_n$	Standard Errors
25	[2,4]	(0.7058,0.0491,0.0994)	(0.0133, 0.0011, 0.009)
49	[2,4]	(0.6958,0.0362,0.1008)	(0.0252, 0.0015, 0.0184)
49	[2,6]	(0.6981,0.0394,0.1006)	(0.0355, 0.0023, 0.0255)
73	[2,8]	(0.6999,0.04,0.1001)	(0.0794, 0.0052, 0.0564)

Table 5: The standard errors for the optimized $\hat{\beta}_n$ with $\sigma = 0.005$ using data sampled from various intervals within R_1 .

Region R_1 with $\sigma_0 = 0.5$

The moderate curve with $\beta_0 = (0.7, 0.04, 0.1)$ using $\beta^0 = (0.8, 0.1, 0.3)$.

n_1	Interval	$\hat{\beta}_n$	Standard Errors
25	[8,12]	(0.5175,0.0281,0.4307)	(1.0073, 0.0671, 2.8477)
49	[8,12]	(0.4713,0.0249,0.5921)	(2.1167, 0.1414, 8.0601)
49	[4,12]	(0.6720,0.0383,0.1254)	(5.1571, 0.3409, 4.5525)

Table 6: The standard errors for the optimized $\hat{\beta}_n$ with $\sigma = 0.5$ using data sampled from various intervals within R_1 .

STATISTICAL METHODOLOGY IN INVERSE PROBLEMS

from notes in the course

MA/ST 810Q, FALL 2002
INVERSE PROBLEM METHODOLOGY
IN COMPLEX STOCHASTIC MODELS

Instructors: H.T. Banks and M. Davidian

(htbanks@ncsu.edu and davidian@ncsu.edu)

SAMSI and North Carolina State University

Center for Research in Scientific Computation

and

Center for Quantitative Sciences in Biomedicine

Introduction to Statistical Modeling and Probability

Outline:

1. Mathematical models vs. statistical models
2. Review of probability distributions

1. Mathematical vs. Statistical Models

Mathematical (deterministic) models: Representation of an exact relationship

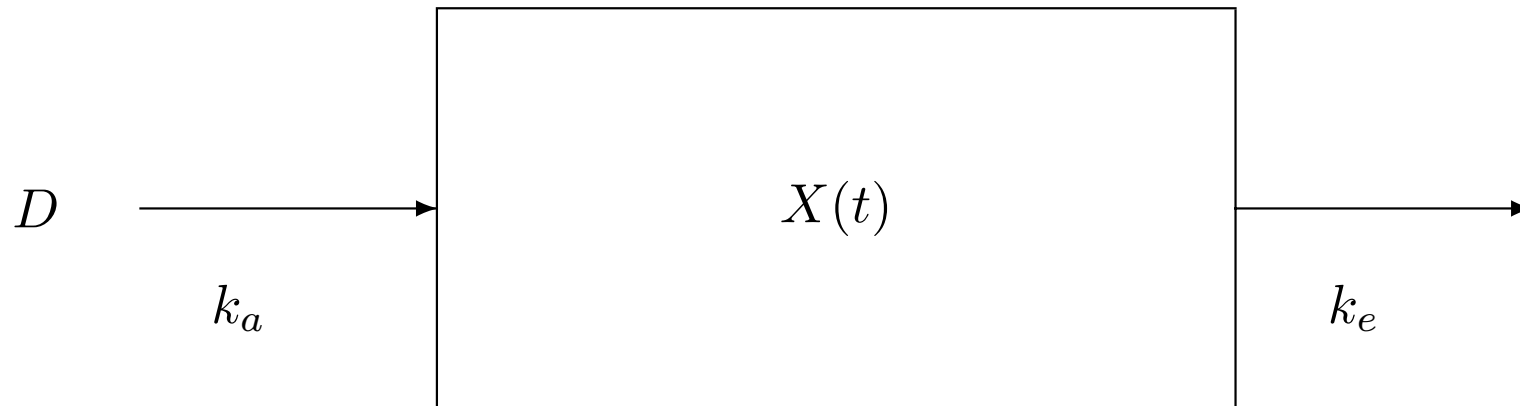
- *Dynamics* $\frac{dx}{dt}(t) = g(t, x(t), \theta)$
- *Output* $y(t) = Cx(t, \theta)$

Observations: Suppose we observe the system over time and record values y_1, \dots, y_n at times $0 \leq t_1 < \dots < t_n$

- *Commonly*, observations *do not* track exactly on the output curve $y(t) = Cx(t, \theta)$

Example: Pharmacokinetics of theophylline (anti-asthmatic agent)

- Understanding of processes of *absorption*, *distribution*, *elimination* important for developing dosing recommendations
- Common mathematical model: *One compartment model with first-order absorption and elimination* following oral dose D



- *Assumption:* $X(t) = \nu c(t)$ (constant relationship between drug concentration in plasma $c(t)$ and amount of drug in body $X(t)$ for all times t)

Mathematically: Letting $X_a(t)$ be the amount of drug at the absorption site at time t

$$\begin{aligned}\dot{X}(t) &= k_a X_a(t) - k_e X(t) \\ \dot{X}_a(t) &= -k_a X_a(t)\end{aligned}$$

with initial conditions $X_a(0) = X_{a0} = FD$, $X(0) = X_0 = 0$, where F is the fraction available.

$$\dot{x}(t) = g(t, x(t), \theta)$$

- $\dot{x}(t) = (\dot{X}(t), \dot{X}_a(t))^T$

$$g(t, x(t), \theta) = \begin{pmatrix} k_a X_a(t) - k_e X(t) \\ -k_a X_a(t) \end{pmatrix}, \quad \theta = (k_a, k_e)^T$$

Solution: May be found analytically in a *closed form*

- By the method of Laplace transforms

Laplace transform of $X(t)$: $\mathcal{L}X = \int_0^{\infty} e^{-st} X(t) dt$

$$s\mathcal{L}X - X_0 = k_a\mathcal{L}X_a - k_e\mathcal{L}X \quad (1)$$

$$s\mathcal{L}X_a - X_{a0} = -k_a\mathcal{L}X_a \quad (2)$$

- Solve (2) for $\mathcal{L}X_a$ and substitute in (1) to obtain

$$\mathcal{L}X = \frac{k_a F D}{(s + k_e)(s + k_a)}$$

- From a table of Laplace transforms, we find immediately that

$$X(t) = \frac{k_a F D}{k_a - k_e} \{e^{-k_e t} - e^{-k_a t}\}$$

so that (divide by ν)

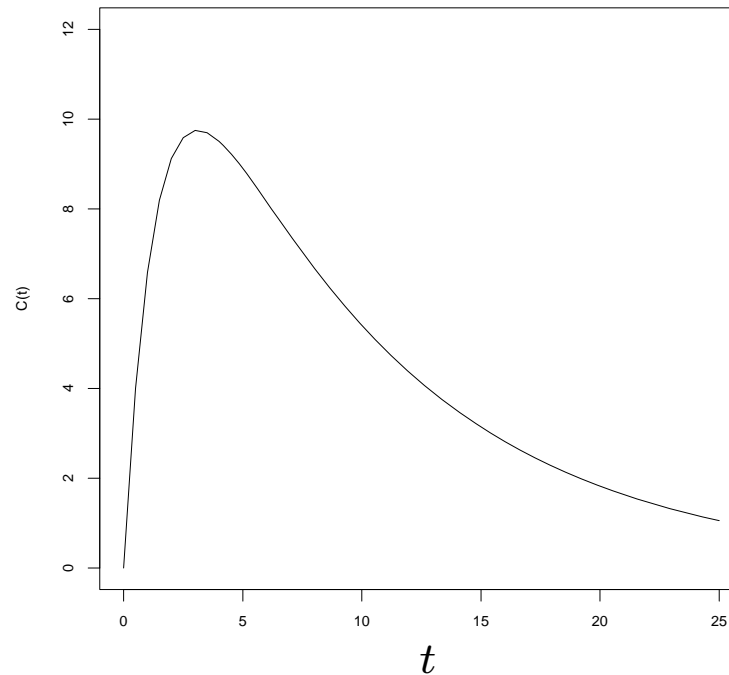
$$c(t) = \frac{k_a F D}{\nu(k_a - k_e)} \{e^{-k_e t} - e^{-k_a t}\}$$

Result: If the model is *perfectly correct*, the relationship

$$c(t) = \frac{k_a F D}{\nu(k_a - k_e)} \{e^{-k_e t} - e^{-k_a t}\}$$

should describe the concentration observed at time t

For example: With $k_a = 0.7$, $k_e = 1.1$, $\nu = 0.4$



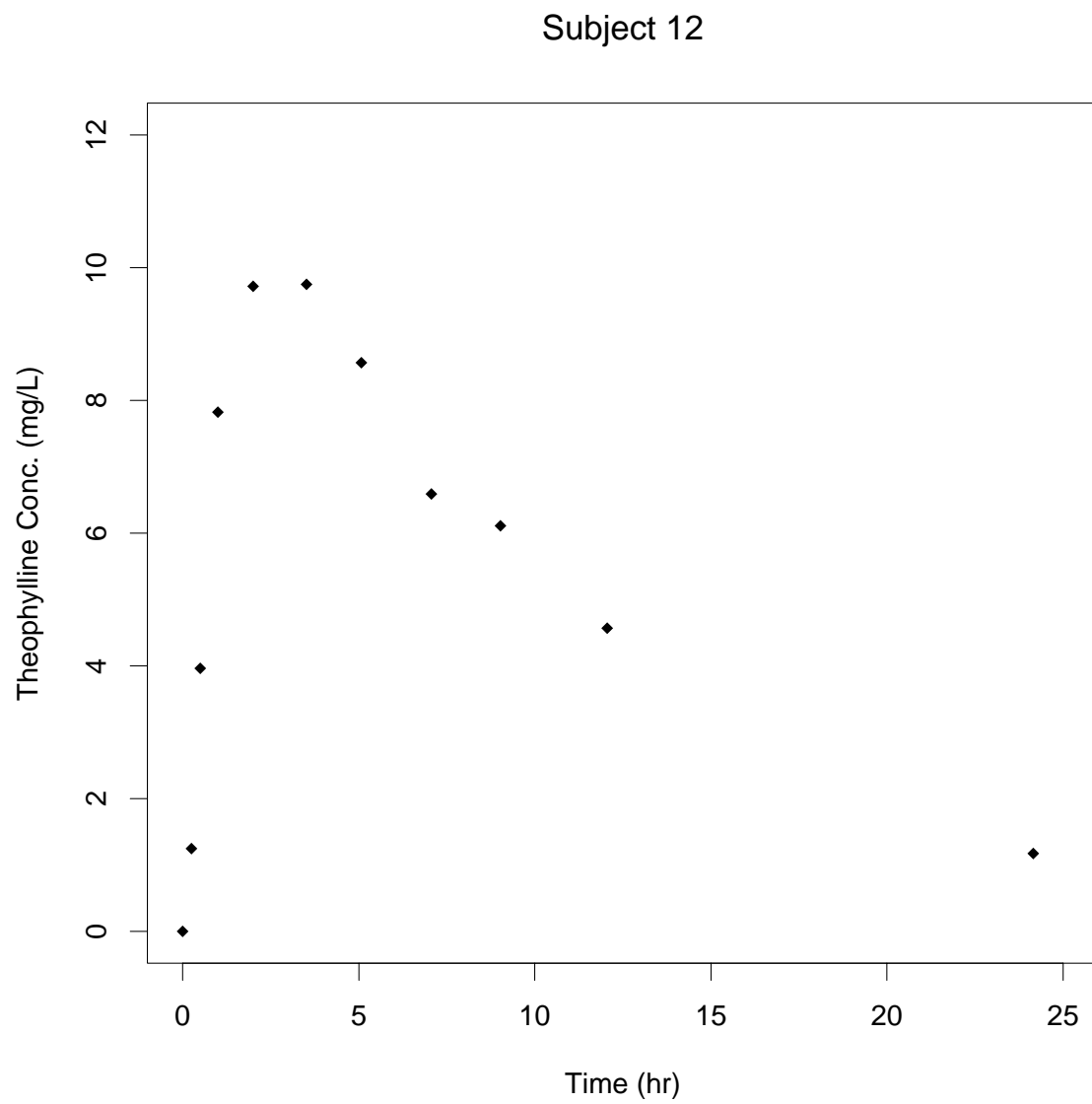
Experiment: PK in *humans* following oral dose

- 12 “*healthy volunteers*” each given dose D (mg/kg) at time $t = 0$
- Blood samples drawn at 10 subsequent time points over the next 25 hours for each subject
- Samples *assayed* for theophylline concentration
- *Observe* y_1, \dots, y_{10} at times t_1, \dots, t_{10}

Objectives:

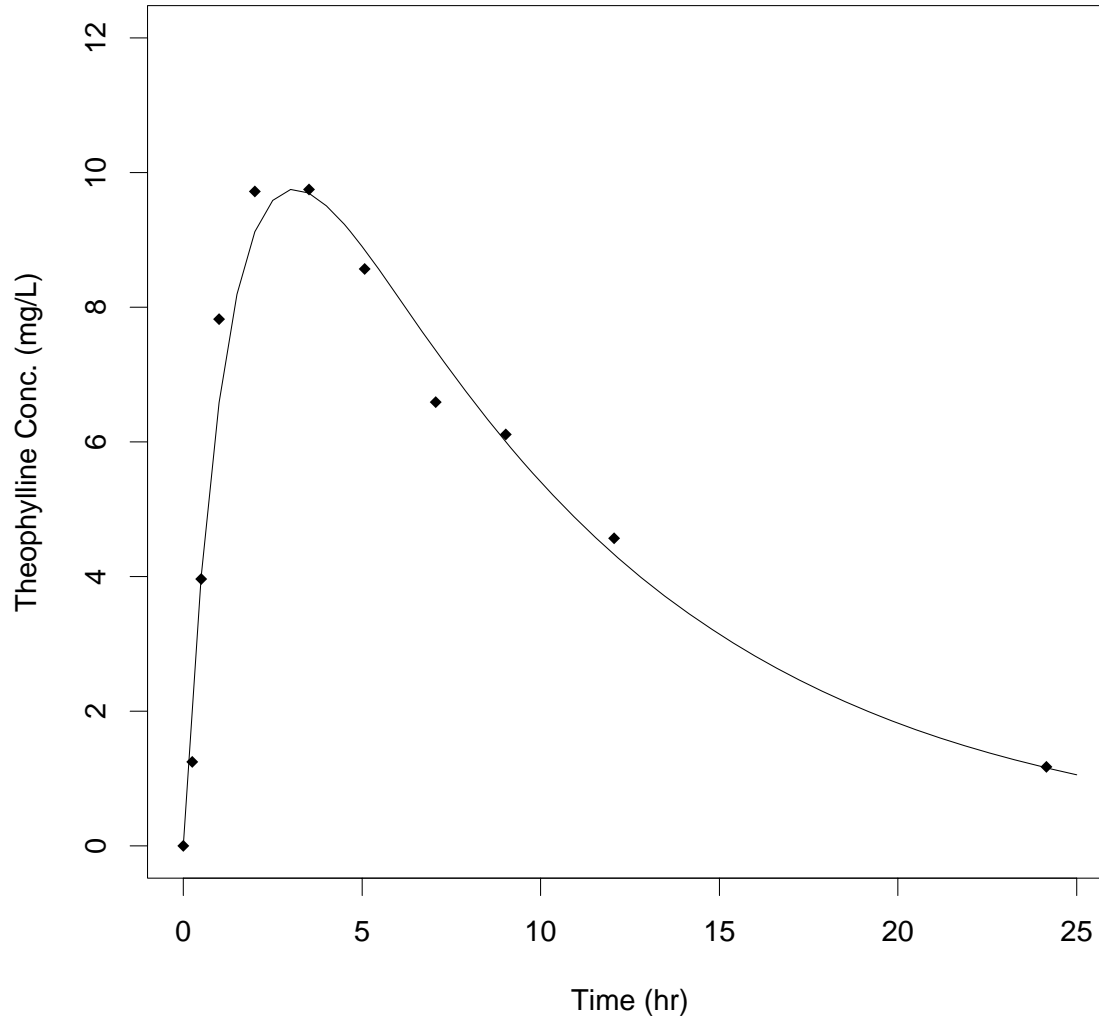
- For a *specific subject*, learn about absorption, elimination, distribution by determining $k_a, k_e, \nu \Rightarrow$ dosing recommendations for this subject
- Learn about how absorption, elimination, and distribution differ from subject to subject \Rightarrow dosing recommendations for the *population* of likely subjects

Data for subject 12: Plot of concentration vs. time



Data for subject 12: With “fitted model” superimposed

Subject 12



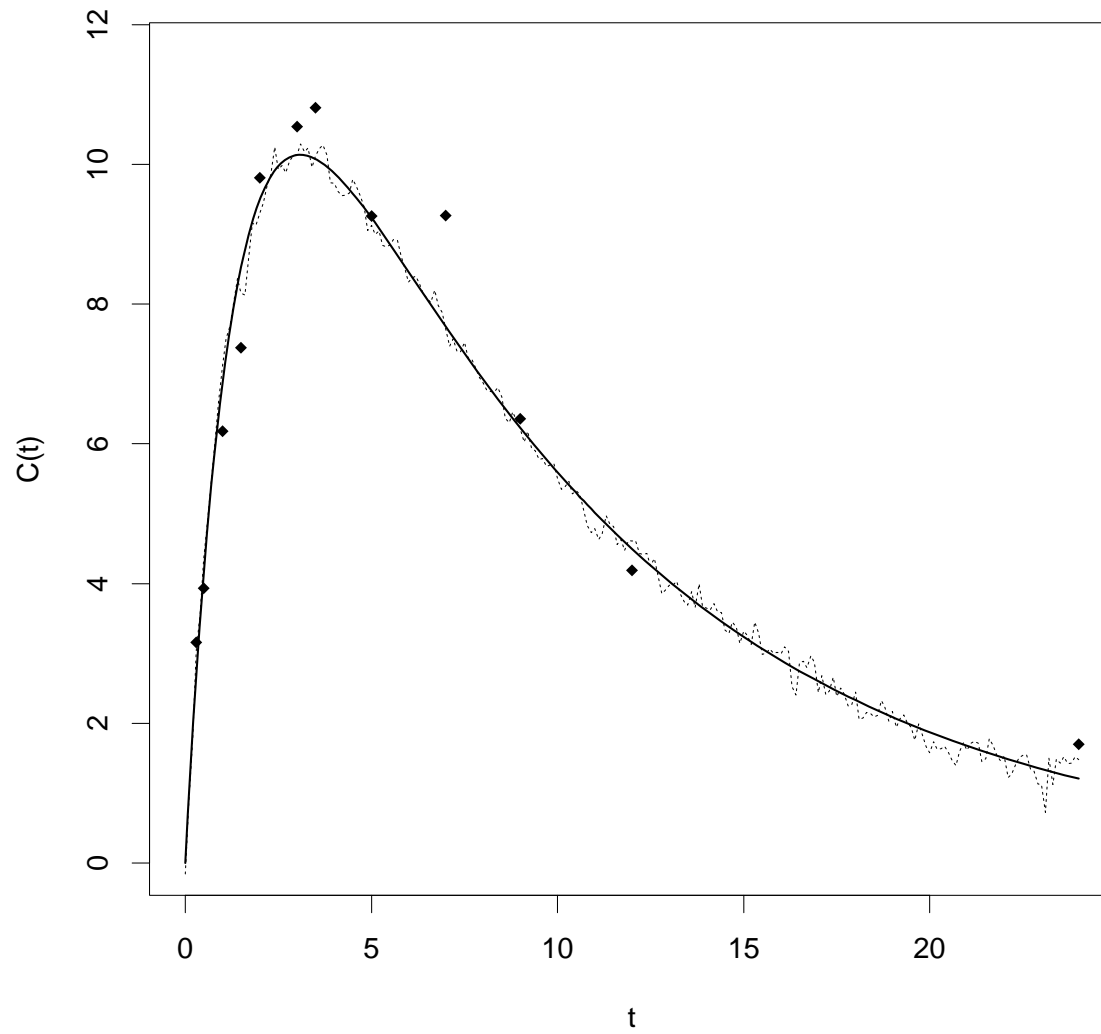
Remarks:

- Observed concentrations appear to trace out a pattern over time quite similar to that dictated by the one compartment model
- But they do not lie *exactly* on a smooth trajectory
- “*Observation error*”

Why?

- One obvious reason: Assay is not perfect, cannot measure concentration *exactly* (measurement error)
- Other reasons?

Approximation: Model is an *idealized* representation of a more complicated biological process



Thus: Can think of what we *observe* as

$$y_j = f(t_j, \theta) + \epsilon_j$$

- $f(t, \theta) = c(t)$, a function of $\theta = (k_a, k_e, \nu)^T$
- ϵ_j is the *deviation* between what the (mathematical) model dictates we would see at t_j and what we actually observe
- Here, ϵ_j represents deviations from $f(t_j, \theta)$ due to *measurement error*, “*biological fluctuations*”

$$\epsilon_j = \epsilon_{1j} + \epsilon_{2j}$$

Overall deviation Measurement Error “Fluctuation”

The diagram illustrates the decomposition of the overall deviation ϵ_j into two components: measurement error and fluctuation. The equation $\epsilon_j = \epsilon_{1j} + \epsilon_{2j}$ is shown at the top. Below it, three labels are positioned: 'Overall deviation' on the left, 'Measurement Error' in the middle, and '“Fluctuation”' on the right. Red curved arrows point from 'Overall deviation' to ϵ_j , from 'Measurement Error' to ϵ_{1j} , and from '“Fluctuation”' to ϵ_{2j} .

Thought experiment: Consider measurement error

- A particular blood sample has a “*true*” concentration of theophylline
- When we measure this concentration, an error is committed, which causes “*observed*” to deviate from “*true*” by an amount that is negative or positive
- Suppose we were to measure the same sample *over and over* (zillions of times) – each time, a possibly different error is committed
- So all such observations would turn out *differently*, even though, *ideally*, they should be all the *same* (measuring the *same thing*)

Result: Measurement error is a *source of variation* that leads to *uncertainty* in what we *observe*. Conceptually

- In actuality, we measure the concentration only *once*
- The error that results may be thought of as drawn from a “*population*” of *all possible errors* that could be committed when measuring concentration
- \Rightarrow *UNCERTAINTY* – the observation could have turned out *differently!!!!*
- Errors, and hence, observations, are *variable*

Recall: Would like to determine θ from the pairs (y_j, t_j) , $j = 1, \dots, n$

- Any determination of θ we try to make from these observations will be subject to *uncertainty*
- That is, if we *estimate* θ from data subject to measurement error (and other sources of variation), the estimate *could have turned out differently*

In these lectures: We will see that

- Failure to acknowledge this can lead to *erroneous conclusions*
- Acknowledging this requires a *formal* way to describe and assess uncertainty, and thus limitations of what can be learned from *data*

Mathematical models: Representation of the “ideal relationship”

$$\dot{x}(t) = g(t, x(t), \theta), \quad y(t) = \mathcal{C}x(t, \theta) \quad (3)$$

Statistical models: Representation of the “actual relationship”

- Incorporate sources of *uncertainty* (variation)
- Framework for *formalizing* assumptions about the effects of variation
- Main tool: *probability*

Statistical model for observed theophylline concentrations:

$$Y_j = f(t_j, \theta) + \epsilon_j, \quad j = 1, \dots, n$$

- Think of ϵ_j as a *random variable* with a *probability distribution* that characterizes “*populations*” of possible values of phenomena like measurement errors, fluctuations that might occur at t_j

Statistical model for observed theophylline concentrations:

$$Y_j = f(t_j, \theta) + \epsilon_j, \quad j = 1, \dots, n$$

- If ϵ_j is a random variable, then so is what we observe
- $\Rightarrow Y_j$ is a random variable with a *probability distribution* that characterizes how observations at t_j may vary because of measurement error, fluctuations, etc.
- The statistical model describes pairs (Y_j, t_j) , $j = 1, \dots, n$, we might see; i.e, the model describes the *mechanism* by which data are thought to arise
- *Data* we observe are realizations of Y_j , $j = 1, \dots, n$: y_1, \dots, y_n
- The *mechanism* is characterized by assumptions on the *probability distribution* of ϵ_j (so, equivalently, on that of Y_j)

2. Review of Probability Distributions

Probability distributions in statistical models:

- Are used to formalize assumptions on model components
- Arise in formalizing assessments of uncertainty based on statistical models (*inference*)

Here: Review

- Basics of probability theory*
- Some important, specific probability distributions

* A comprehensive introduction at a moderate level is given in **Casella and Berger (2002)**, *Statistical Inference, Second Edition*

(Statistical) experiment: Formal conceptualization

- One toss of a coin
- Choose a person from a population of size N at random
- Observe concentration in a blood sample

Sample space Ω : Set of *all possible outcomes* of an experiment

Examples: Countable or uncountable

- One toss of a coin: $\Omega = \{H, T\}$
- Choose a person from a population of size N : $\Omega = \{\omega_1, \dots, \omega_N\}$
- Observe concentration: $\Omega = (0, \infty)$
- Observe error committed: $\Omega = (-\infty, \infty)$

Event: A *collection of possible outcomes* of an experiment; i.e., any *subset* A of Ω

- Events $A \subset \Omega$ obey usual set-theoretic rules; e.g., union, intersection

Probability: For each $A \subset \Omega$, assign a *number between 0 and 1*, denoted by $P(A)$

- Technically, not that simple
- \mathcal{B} = collection of subsets of Ω that includes \emptyset , closed under complementation, closed under countable unions (σ -algebra), (Ω, \mathcal{B}) is a σ -field

Probability function or measure: For Ω with associated \mathcal{B} , P is a *probability function* with domain \mathcal{B} and range $[0, 1]$ if

- $P(A) \geq 0$ for $A \in \mathcal{B}$
- $P(\Omega) = 1$
- $A_1, A_2, \dots \in \mathcal{B}, A_i \cap A_j = \emptyset \Rightarrow P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Probability space or triple: $(\Omega, \mathcal{B}, \mathcal{P})$

Properties: For $A, B \in \mathcal{B}$

- $P(\emptyset) = 0$
- $P(A) \leq 1$
- $P(A^c) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A) \leq P(B)$ if $A \subset B$
- $P(A) = \sum_{i=1}^{\infty} P(A \cap B_i)$, B_i disjoint partition of Ω

Random variable: A *function* from Ω into the real numbers
(assigns a real number to each element of the sample space)

- Mapping from *original sample space* Ω to *new sample space* \mathcal{X}
- Often denoted by capital letters, e.g., X , Y

Example: Toss a coin two times

- $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\} = \{HH, HT, TH, TT\}$,
 $X(\omega) = \#$ of heads in two tosses taking values in $\mathcal{X} = \{0, 1, 2\}$

Example: Sample a person from a population of size N and observe survival time

- $\Omega = \{\omega_1, \dots, \omega_N\}$,
 $X(\omega) =$ survival time taking values in $\mathcal{X} = (0, \infty)$

Example: Measurement error (uncountable)

- $\Omega = \{ \text{all possible conditions of measurement} \}$,
 $\epsilon(\omega) =$ error committed taking values in $\mathcal{X} = \mathfrak{R}$

Probability function for random variable X :

- *Countable* Ω and \mathcal{X} : $X = x_i \in \mathcal{X}$ iff ω_j is such that $X(\omega_j) = x_i$

$$P_X(X = x_i) = P(\{\omega_j \in \Omega : X(\omega_j) = x_i\})$$

- *Uncountable* \mathcal{X} : for $A \in \mathcal{X}$ (actually, in a certain σ -algebra (measurable sets) of subsets of \mathcal{X})

$$P_X(X \in A) = P(\{\omega \in \Omega : X(\omega) \in A\})$$

- Customary to discuss probability with respect to *random variables* and suppress X subscript
- Write X for the random variable (the function) and x for its possible values (realizations, elements of \mathcal{X})
- “*Probability distribution*”

Cumulative distribution function (cdf): For random variable X

$$F_X(x) = F(x) = P(X \leq x) \quad \text{for all } x$$

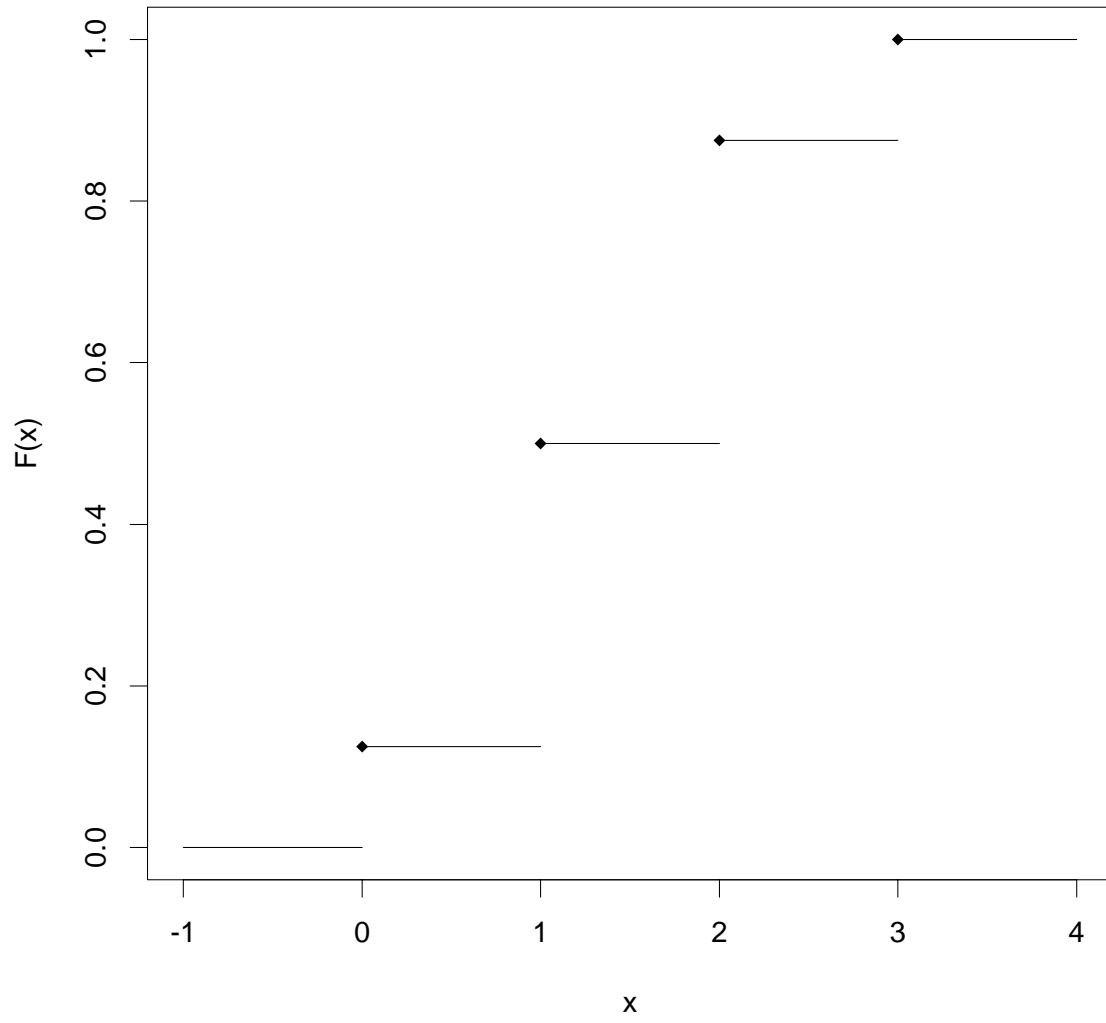
(not just $x \in \mathcal{X}$)

- $F(x)$ is *nondecreasing* and *right continuous*
- $\lim_{x \rightarrow -\infty} F(x) = 0$, $\lim_{x \rightarrow \infty} F(x) = 1$

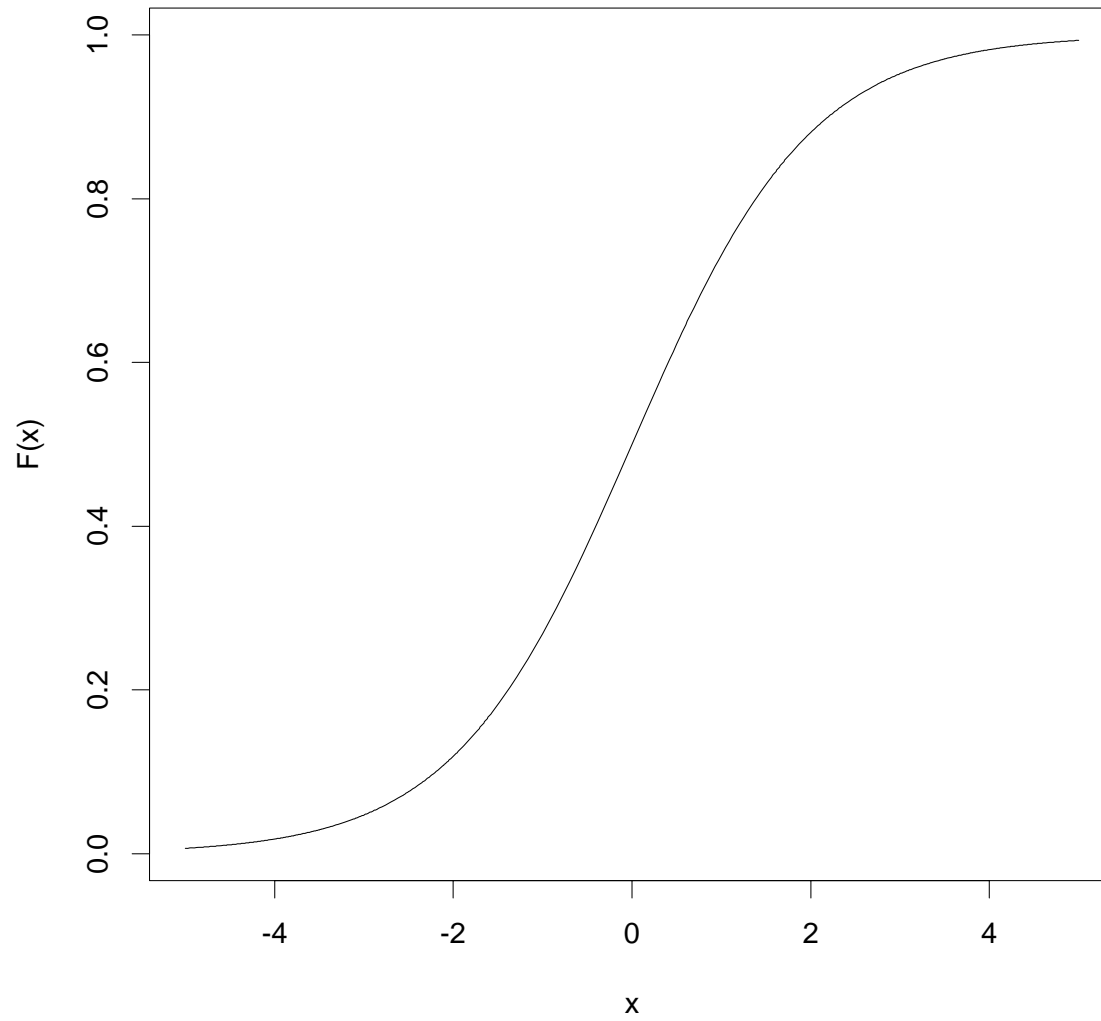
Example: Toss a coin three times, $X = \#$ heads

x	$P(X = x)$
0	$\frac{1}{8}$
1	$\frac{3}{8}$
2	$\frac{3}{8}$
3	$\frac{1}{8}$

$$F(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ \frac{1}{8} & \text{if } 0 \leq x < 1 \\ \frac{4}{8} & \text{if } 1 \leq x < 2 \\ \frac{7}{8} & \text{if } 2 \leq x < 3 \\ 1 & \text{if } 3 \leq x < \infty \end{cases}$$



$$F(x) = \frac{1}{1 + e^{-x}}$$



Discrete and continuous random variables: A random variable X is said to be

- *Continuous* if $F(x)$ is an (absolutely) continuous function of x
- *Discrete* if $F(x)$ is a step function of x

Probability mass and density functions: Concerned with “point probabilities” of random variables

- Discrete: *probability mass function*
- Continuous: *probability density function*

Probability mass function (pmf):

$$f(x) = P(X = x) \quad \text{for all } x$$

- Thus, $P(X \leq x) = F(x) = \sum_{u:u \leq x} f(u)$

Example: Binomial probability mass function

- *Bernoulli trial*: experiment with 2 possible outcomes
- X has a *Bernoulli* probability distribution with $\mathcal{X} = \{0, 1\}$, if

$$X = \begin{cases} 1 & \text{with probability } p & \text{“success”} \\ 0 & \text{with probability } 1 - p & \text{“failure”} \end{cases} \quad 0 \leq p \leq 1$$

- *Binomial distribution* – for n identical Bernoulli trials, let $Y =$ total # successes with sample space $\mathcal{Y} = \{0, 1, \dots, n\}$

$$f(y) = P(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y}, \quad y \in \mathcal{Y}, \quad = 0 \text{ otherwise}$$

Probability density function (pdf): Must be a little more careful when X is *(absolutely) continuous*

- $\{X = x\} \subset \{x - \epsilon < X < x\}$ for all $\epsilon > 0 \Rightarrow$

$$0 \leq P(X = x) \leq P(x - \epsilon < X \leq x) = F(x) - F(x - \epsilon)$$

Thus, by continuity of $F(\cdot)$,

$$0 \leq P(X = x) \leq \lim_{\epsilon \downarrow 0} \{F(x) - F(x - \epsilon)\} = 0$$

- By analogy to discrete pmf, for continuous $f(\cdot)$

$$P(X \leq x) = F(x) = \int_{-\infty}^x f(u) du \quad \text{for all } x$$

- $\Rightarrow d/dx F(x) = f(x)$ i.e., F is absolutely continuous with derivative which, when integrated, returns F
- Note: May have **mixed distributions** (A.C. + discrete jumps)– won't discuss here!

Probability mass and density functions satisfy:

- $f(x) \geq 0$ for all x
- $\sum_x f(x) = 1$ (pmf) or $\int_{-\infty}^{\infty} f(x) dx = 1$ (pdf)

Notation in these lectures: To avoid confusion with our use of f to denote the solution of a system as before

- We will often use $P(x)$ to denote the *cdf* of a random variable X and $p(x)$ to denote the *pmf* or *pdf* as appropriate
- We may add *subscripts* when speaking simultaneously of several random variables; e.g., $p_\epsilon(\epsilon)$ and $p_X(x)$
- We will use “ \sim ” to denote “*distributed as*”

Transformations of random variables: If X is a random variable with cdf $F_X(x)$, then a function $Y = g(X)$ is *also a random variable* with new sample space \mathcal{Y} with elements of form $y = g(x)$

$$P(Y \in A) = P\{g(X) \in A\} = P\{x \in \mathcal{X} : g(x) \in A\} = P\{X \in g^{-1}(A)\}$$

where g^{-1} is inverse mapping from \mathcal{Y} to \mathcal{X} .

- The distribution of Y depends on that of X
- In particular, $F_Y(y)$, $f_Y(y)$ are *related* to $F_X(x)$, $f_X(x)$

Several random variables at once: p -dimensional *random vector*

$(X_1, \dots, X_p)^T$ is a function from Ω into \mathfrak{R}^p . Consider $p = 2$

- All components *discrete – joint pmf*

$$f(x_1, x_2) = P(X_1 = x_1, X_2 = x_2)$$

Satisfies $\sum_{x_1, x_2} f(x_1, x_2) = 1$

- All components *continuous – joint pdf* $f(x_1, x_2)$ from \mathfrak{R}^2 into \mathfrak{R} satisfies

$$P\{X_1, X_2\} \in A\} = \int \int_A f(x_1, x_2) dx_1 dx_2, \quad \int \int f(x_1, x_2) dx_1 dx_2 = 1$$

- *Marginal* pmf and pdf: E.g., X_1

$$f_{X_1}(x_1) = \sum_{x_2} f(x_1, x_2) \text{ or } f_{X_1}(x_1) = \int f(x_1, x_2) dx_2$$

Independent random variables: X_1 and X_2 are *independent* if

$$f(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2), \quad X_1 \underline{\underline{\parallel}} X_2$$

Expectation of a random variable: The “*average*” value of a random variable

- “*weighted*” according to the probability distribution
- Measure of “*center*”

Expected value or mean: For random variable X , the expected value of $g(X)$ is

$$E\{g(X)\} = \begin{cases} \int_{-\infty}^{\infty} g(x)f(x) dx & X \text{ continuous} \\ \sum_x g(x)f(x) = \sum_x g(x)P(X = x) & X \text{ discrete} \end{cases}$$

$$g(x) = x \Rightarrow E\{X\} = \text{mean } \mu \text{ of } X$$

Higher moments: For random variable X and integer k

- The k th moment of X is $E(X^k)$
- The k th *central moment* is $E\left[\{X - E(X)\}^k\right]$

Variance: Second central moment

$$\text{var}(X) = E\left[\{X - E(X)\}^2\right]$$

- Measure of degree of “*spread*” of distribution about its mean
- Standard deviation = $\sqrt{\text{var}(X)}$ on same scale of X
- Quantifies *variation*

Random vectors: Element-by-element using marginal pmf/pdf

Covariance and correlation: Measures of “*degree of association*”

– For any two random variables

- *Covariance* between X_1 and X_2 is defined as

$$\text{cov}(X_1, X_2) = E\left[\{X_1 - E(X_1)\}\{X_2 - E(X_2)\}\right]$$

- Will be > 0 if $X_1 > E(X_1)$ and $X_2 > E(X_2)$ or $X_1 < E(X_1)$ and $X_2 < E(X_2)$ tend to happen together
- Will be < 0 if $X_1 > E(X_1)$ and $X_2 < E(X_2)$ or $X_1 < E(X_1)$ and $X_2 > E(X_2)$ tend to happen together
- Will be $= 0$ if X_1 and X_2 are ||
- *Correlation* is covariance put on a unitless basis

$$\rho_{X_1 X_2} = \text{corr}(X_1, X_2) = \frac{\text{cov}(X_1, X_2)}{\sqrt{\text{var}(X_1)\text{var}(X_2)}}$$

- $-1 \leq \rho_{X_1 X_2} \leq 1$; $\rho_{X_1, X_2} = -1$ or 1 iff $X_1 = a + bX_2$

Discrete probability distributions:

- $X \sim \text{Binomial}(n, p)$

$$f(x) = P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, 1, \dots, n$$

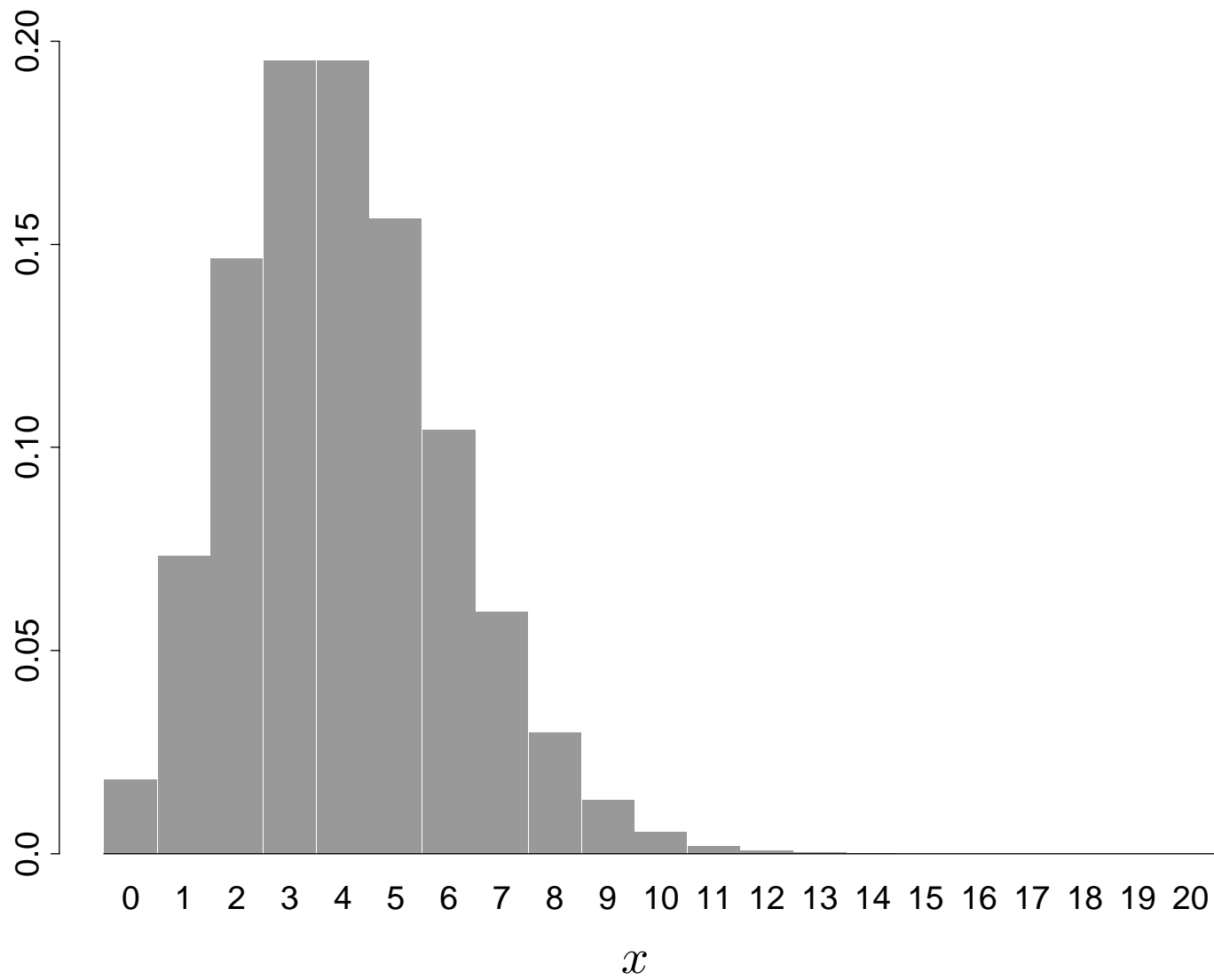
$$E(X) = np, \quad \text{var}(X) = np(1 - p)$$

- $X \sim \text{Poisson}(\lambda)$ – a model for *counts*

$$f(x) = P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$

$$E(X) = \lambda, \quad \text{var}(X) = \lambda$$

Poisson pmf with $\lambda = 4$:



Continuous probability distributions:

- *Normal* or *Gaussian* distribution:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

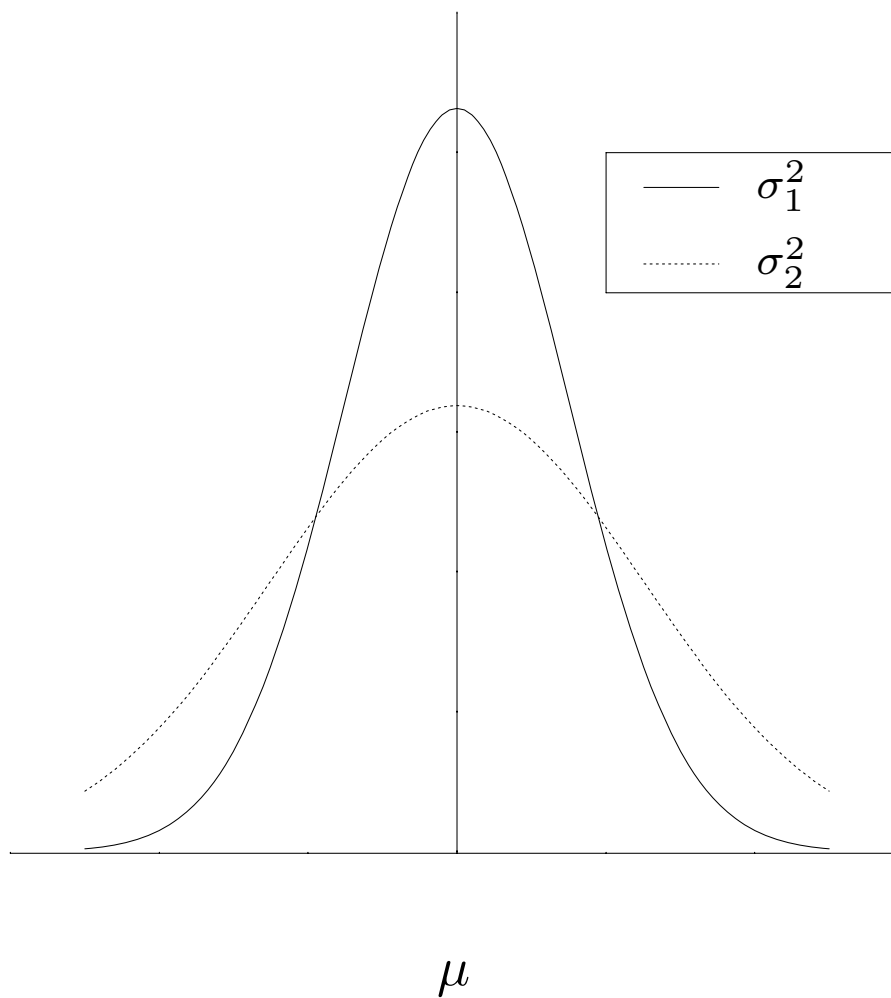
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\}, \quad -\infty < x < \infty$$

$$E(X) = \mu, \text{var}(X) = \sigma^2, \sigma > 0$$

- *Symmetric* about its mean
- $Z = (X - \mu)/\sigma \sim \mathcal{N}(0, 1)$ *standard normal*
- A (the most) popular model for phenomena such as measurement errors, observations on biological, physical phenomena
- Plays a central role in approximate methods of *statistical inference* for complex models

Two normal pdfs with same mean μ , different variances

$\sigma_1^2 < \sigma_2^2$:



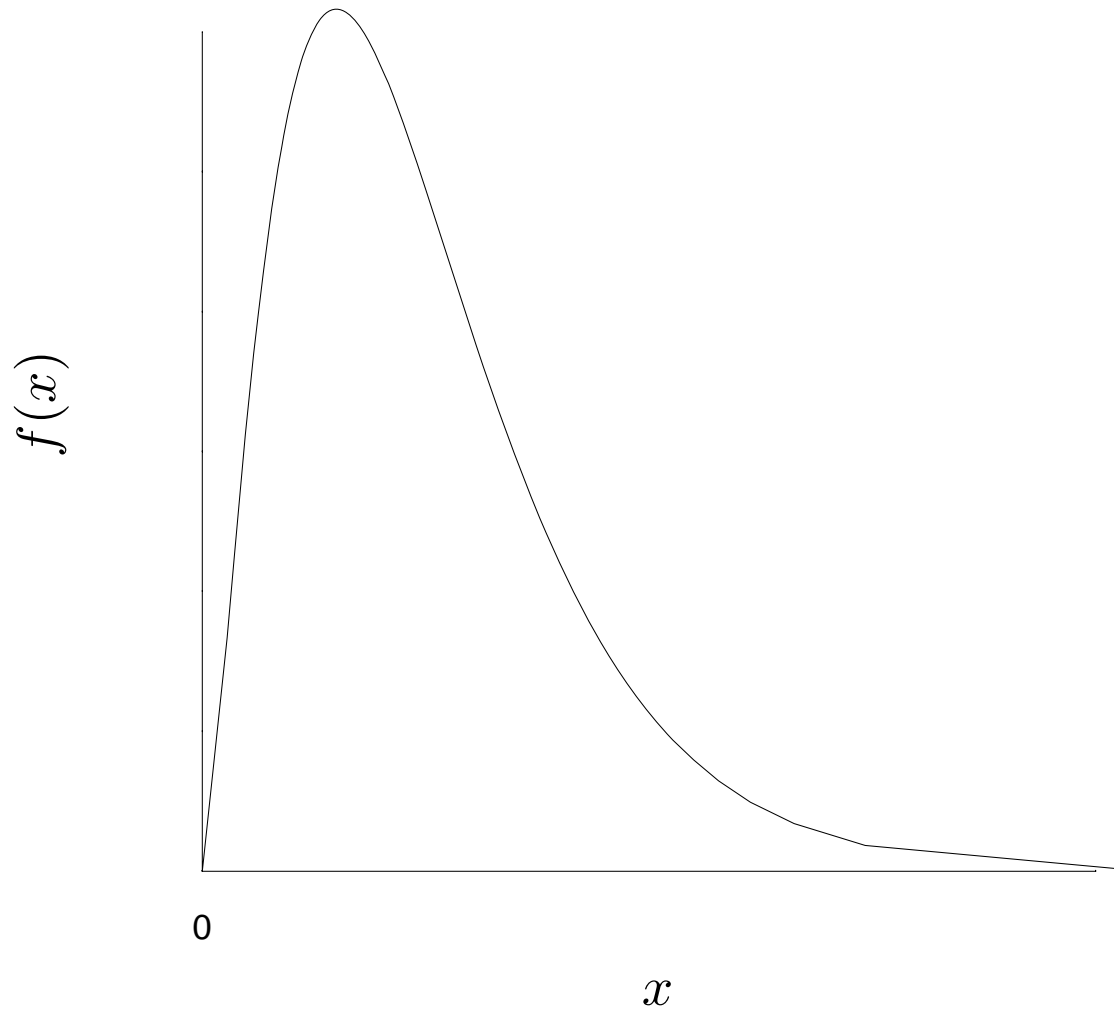
Continuous probability distributions:

- *Lognormal* distribution: If $\log X \sim \mathcal{N}(\mu, \sigma^2)$, then X has a lognormal distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \frac{1}{x} \exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\}, \quad 0 < x < \infty$$

$$E(X) = e^{\mu + \sigma^2/2}, \quad \text{var}(X) = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2} \propto \{E(X)\}^2$$

- Constant *coefficient of variation (CV)* = $\sqrt{\text{var}(X)}/E(X)$ (“*noise-to-signal*”) – does not depend on $E(X)$
- A common model for biological phenomena
- Skewed (asymmetric) with “*long right tail*”
- Looks more and more symmetric as $\sigma \rightarrow 0$



Continuous probability distributions:

- *Gamma* distribution

$$f(x) = \frac{1}{\Gamma(a)b^a} x^{a-1} \exp(-x/b), \quad 0 < x < \infty, \quad a, b > 0$$

$$E(X) = ab, \quad \text{var}(X) = ab^2$$

- Constant CV = $a^{-1/2}$
- Similar in shape to lognormal
- Looks more and more symmetric as $a \rightarrow \infty$
- Special case 1: *Exponential distribution* $a = 1$
- Special case 2: *Chi squared (χ^2) distribution with k degrees of freedom*

For integer $k > 0$, set $a = k/2$, $b = k \Rightarrow$ important in *statistical inference*

Continuous probability distributions: These two are also important in *statistical inference*

- *Student's t distribution with k degrees of freedom*

If $U \sim \mathcal{N}(0, 1)$, $V \sim \chi_k^2$ are ||, then $X = U/\sqrt{V/k} \sim t_k$ with pdf

$$f(x) = \frac{\Gamma\{(k+1)/2\}}{\Gamma(k/2)} \frac{1}{\sqrt{k\pi}} \frac{1}{(1+x^2/k)^{(k+1)/2}}, \quad -\infty < x < \infty$$

$E(X) = 0$ if $k > 1$, $var(X) = k/(k-2)$ if $k > 2$

- Symmetric like normal, with “*heavier tails*,” becomes normal as $k \rightarrow \infty$
- *F distribution with k_1, k_2 degrees of freedom*

If $U \sim \chi_{k_1}^2$, $V \sim \chi_{k_2}^2$ are ||, then $X = (U/k_1)/(V/k_2) \sim \mathcal{F}_{k_1, k_2}$ with pdf

$$f(x) = \frac{\Gamma\{(k_1 + k_2)/2\}}{\Gamma(k_1/2)\Gamma(k_2/2)} \left(\frac{k_1}{k_2}\right)^{k_1/2} \frac{x^{k_1/2-1}}{\{1 + (k_1/k_2)x\}^{(k_1+k_2)/2}}, \quad 0 < x < \infty$$

Multivariate normal distribution: *Random vector*

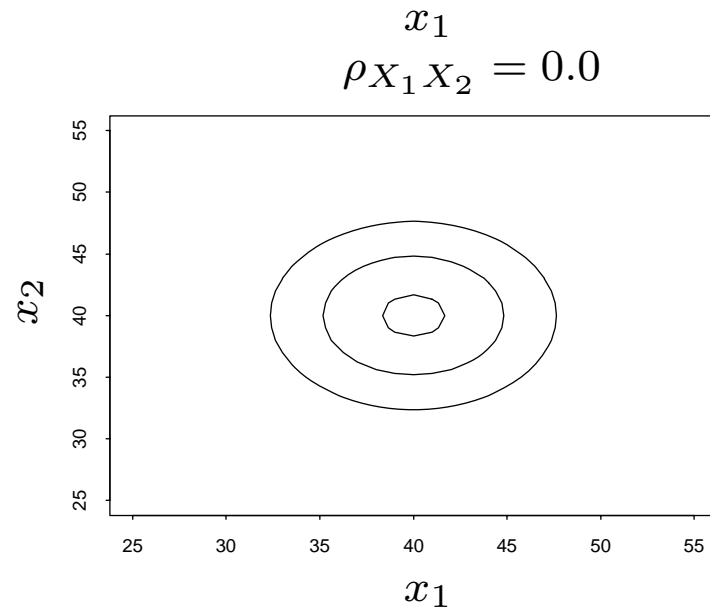
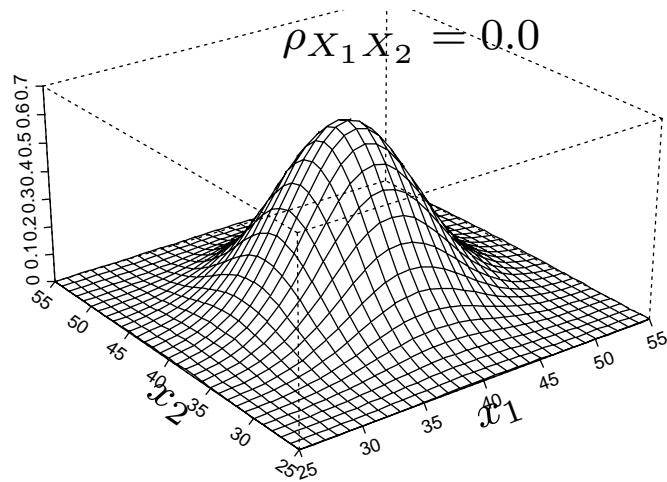
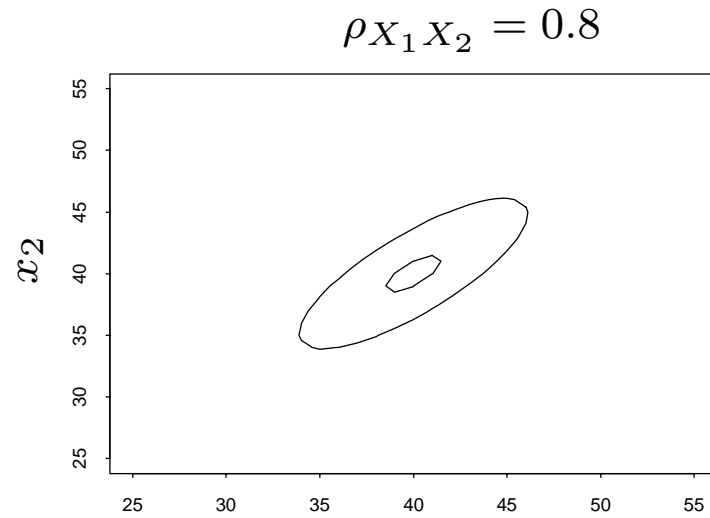
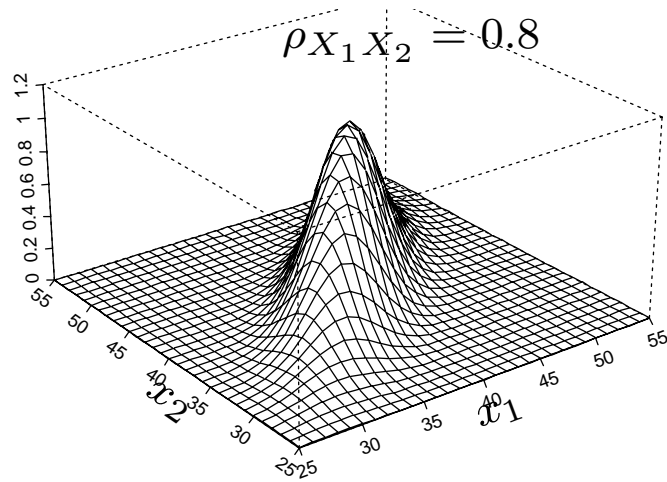
$X = (X_1, \dots, X_p)^T$ has a multivariate (p -variate) normal distribution if $\alpha^T X \sim \text{normal} \forall \alpha \in \mathfrak{R}^p$

$$f(x) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\{-(x - \mu)^T \Sigma^{-1} (x - \mu)/2\},$$

for $x = (x_1, \dots, x_p)^T \in \mathfrak{R}^p$

- $E(X) = \mu = (\mu_1, \dots, \mu_p)^T = \{E(X_1), \dots, E(X_p)\}^T$
- Σ ($p \times p$) is such that $\Sigma_{jj} = \text{var}(X_j)$, $\Sigma_{jk} = \Sigma_{kj} = \text{cov}(X_j, X_k)$
- $\Sigma = E\{(x - \mu)(x - \mu)^T\}$ is the *covariance matrix*
- The *marginal* pdfs are *univariate* normal
- Incredibly important in statistical *modeling* and *inference*

Two bivariate ($p = 2$) normal pdfs:



Statistical Models and Inference

Lecture 2

H.T.Banks and Marie Davidian

Outline:

1. Statistical models and sources of variation
2. Statistical inference (frequentist)

1. Statistical models and sources of variation

Key point: The notions of *random variables* and *probability distributions* are the building blocks of *statistical models*

- A *statistical model* is a representation of the *mechanism* by which *data* are assumed to arise
- Phenomena that are *subject to variation* and hence give rise to *uncertainty* in the way data may “*turn out*” are represented by *random variables*
- Assumptions on the nature of *probability distributions* for random variables in statistical models represent assumptions on the nature and extent of such variation
- Return to the *theophylline example* for a demonstration...

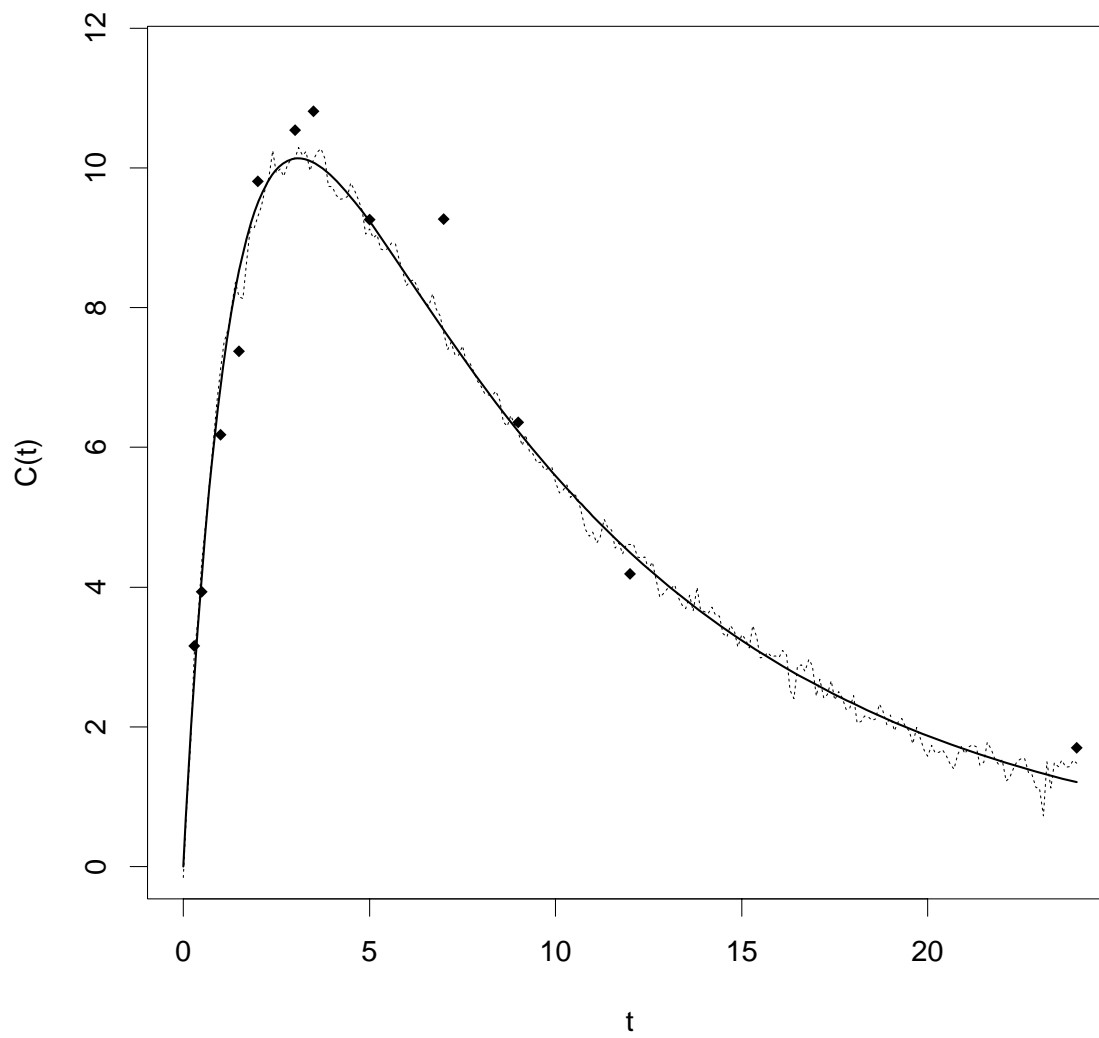
Recall: We wrote a statistical model for pharmacokinetics of theophylline for a given subject as

$$Y_j = f(t_j, \theta) + \epsilon_j, \quad j = 1, \dots, n$$

$$f(t, \theta) = \frac{k_a F D}{\nu(k_a - k_e)} \{e^{-k_e t} - e^{-k_a t}\}, \quad \theta = (k_a, k_e, \nu)^T$$

- $f(t, \theta)$ is the smooth function of t derived from the *(deterministic) mathematical* compartment model
- ϵ_j represents *deviation* that causes observations to not fall exactly on the smooth path $f(t, \theta)$
- Aggregate effects of *measurement error*, “*biological fluctuations*,” other phenomena
- $\Rightarrow \epsilon_j$ is a *random variable* whose *probability distribution* reflects assumed features of these phenomena
- And Y_j is also a *random variable* (transformation of ϵ_j)

Conceptual representation:



More formal representation: In principle, the pharmacokinetic process could be observed at *any point* in time

- Let $Y(t)$ be the observed concentration value that would be seen at time t and $\epsilon(t)$ be the corresponding deviation

$$Y(t) = f(t, \theta) + \epsilon(t), \quad t \geq 0$$

- This represents the assumed data generating mechanism for *any* point in time
- $Y(t)$ and $\epsilon(t)$ are *stochastic processes* – a *random function* of time with sample space of possible values (functions), e.g., $y(t)$, $t \geq 0$ (*sample paths*)
- For a *fixed set* of times $t_1 < \dots < t_n$ we may write

$$Y(t_j) = f(t_j, \theta) + \epsilon(t_j)$$

to represent observations that would be seen at these times

- $Y_j = Y(t_j)$, $\epsilon_j = \epsilon(t_j)$

$$Y(t_j) = f(t_j, \theta) + \epsilon(t_j)$$

Thus:

- $\{\epsilon(t_1), \epsilon(t_2), \dots, \epsilon(t_n)\}^T = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)^T$ and $\{Y(t_1), Y(t_2), \dots, Y(t_n)\}^T = (Y_1, Y_2, \dots, Y_n)^T$ are *random vectors*
- Can consider the *joint probability distribution* of $(\epsilon_1, \epsilon_2, \dots, \epsilon_n)^T$ [and thus of $(Y_1, Y_2, \dots, Y_n)^T$]
- *Technical point:* Probability distribution for a stochastic process arises from thinking of *all possible* such vectors and their joint distributions for all n (infinitely many)

$$Y(t_j) = f(t_j, \theta) + \epsilon(t_j)$$

Nature of $\epsilon(t)$: “*Biological fluctuations,*” *measurement error*

$$\epsilon(t) = \epsilon_1(t) + \epsilon_2(t)$$

- $\epsilon_1(t_j) = \epsilon_{1j}$ represents *measurement error* that could be committed at fixed time t_j
- $\epsilon_2(t_j) = \epsilon_{2j}$ represents “*fluctuation*” that might occur at t
- These random variables are *continuous* – concentrations *in principle* can take on *any value* (although we may be limited in what we may actually observe due to limits on resolution of measurement)
- Write $\{\epsilon_1(t_1), \dots, \epsilon_1(t_n)\}^T = (\epsilon_{11}, \dots, \epsilon_{1n})^T$ and $\{\epsilon_2(t_1), \dots, \epsilon_2(t_n)\}^T = (\epsilon_{21}, \dots, \epsilon_{2n})^T$ – *random vectors*

Measurement error: Some “*reasonable*” assumptions on aspects of the *joint probability distribution* of $(\epsilon_{11}, \dots, \epsilon_{1n})^T$

- Measuring device is *unbiased* – does not systematically err in a particular direction \Rightarrow

$$E(\epsilon_{1j}) = 0 \quad \text{for each } j = 1 \dots, n$$

(All possible errors for measuring concentration for the sample taken at any t_j “average out” to zero)

- In fact, negative or positive errors are *equally likely* \Rightarrow the *marginal probability density* of ϵ_{1j} is *symmetric* for each j
- Measurement errors at any two times $t_j, t_{j'}$ are “*unrelated*”

$$\epsilon_{1j} \perp \epsilon_{1j'} \quad \Rightarrow \quad \text{cov}(\epsilon_{1j}, \epsilon_{1j'}) = 0$$

- *Variation* among all errors that might occur at any t_j is the *same* \Rightarrow

$$\text{var}(\epsilon_{1j}) = \sigma_1^2$$

for all j (unaffected by time or “actual concentration” in the sample at t_j) – *is this realistic?*

“Biological fluctuations”: Some “*reasonable*” assumptions on aspects of the *joint probability distribution* of $(\epsilon_{21}, \dots, \epsilon_{2n})^T$

- Fluctuations tend to “track” the smooth trajectory $f(t, \theta)$ over time (sample path) but can be “above” or “below” at any point in time \Rightarrow

$$E(\epsilon_{2j}) = 0$$

(All possible fluctuations at any particular time “average out” to zero)

- In fact, negative or positive fluctuations at a particular time are *equally likely* \Rightarrow the *marginal probability density* of ϵ_{2j} is *symmetric*
- *Variation* among fluctuations that might occur at any t_j is *same* \Rightarrow

$$\text{var}(\epsilon_{2j}) = \sigma_2^2$$

- Fluctuations “*close together*” in time (at times $t_j, t_{j'}$) tend to behave “*similarly*,” with extent of “*similarity*” decreasing as $|t_j - t_{j'}| \uparrow$

$$\text{cov}(\epsilon_{2j}, \epsilon_{2j'}) = C(|t_j - t_{j'}|) \Rightarrow \text{corr}(\epsilon_{2j}, \epsilon_{2j'}) = c(|t_j - t_{j'}|)$$

for decreasing functions $C(\cdot), c(\cdot)$ with $C(0) = \sigma_2^2$ and $c(0) = 1$

“Biological fluctuations,” continued:

- E.g., for $\text{corr}(\epsilon_{2j}, \epsilon_{2j'}) = c(|t_j - t_{j'}|)$,

$$c(u) = \exp(-\phi u^2)$$

(so correlation between fluctuations at two times is *nonnegative*, reflecting “*similarity*”)

- Extent and direction of measurement error at any time t_j *unrelated* to fluctuations at t_j or any other time \Rightarrow

$$\epsilon_{1j} \perp \epsilon_{2j'}$$

for any $t_j, t_{j'}, j, j' = 1, \dots, n$

Remarks:

- The foregoing assumptions are not the *only* assumptions one could make, but exemplify the considerations involved
- The *normal probability distribution* is a natural choice to represent the assumption of *symmetry*

Aside: Fun facts for joint probability distributions

- For *random variables* X_1 and X_2 and constants a and b ,
 - $E(aX_1 + bX_2) = aE(X_1) + bE(X_2)$
 - $var(aX_1 + bX_2) = a^2var(X_1) + b^2var(X_2) + 2ab cov(X_1, X_2)$with

$$var(aX_1 + bX_2) = a^2var(X_1) + b^2var(X_2) \quad \text{if } X_1 \perp X_2$$

- For a $(n \times 1)$ *random vector* $X = (X_1, \dots, X_n)$ the *covariance matrix* $E\left[\{X - E(X)\}\{X - E(X)\}^T\right]$ has
 - diagonal elements $var(X_j)$, $j = 1, \dots, n$
 - off-diagonal elements $cov(X_j, X_{j'})$
- If X_1 and X_2 are two *independent*, $(n \times 1)$ *random vectors*, each with a *multivariate normal probability distribution*, and A and B are conformable constant matrices, then the probability distribution of $AX_1 + BX_2$ is *also multivariate normal*

Aside: Fun facts for joint probability distributions

- In fact, for two $(n \times 1)$ *random vectors* X_1 and X_2 with covariance matrices Σ_1 and Σ_2 and conformable constant matrices A and B

- $E(AX_1 + BX_2) = AE(X_1) + BE(X_2)$

- The *covariance matrix* of $AX_1 + BX_2$ is

$$A\Sigma_1A^T + B\Sigma_2B^T + AE\left[\{X_1 - E(X_1)\}\{X_2 - E(X_2)\}\right]B^T \\ + BE\left[\{X_2 - E(X_2)\}\{X_1 - E(X_1)\}\right]A^T,$$

which equals

$$A\Sigma_1A^T + B\Sigma_2B^T \quad \text{if } X_1 \perp X_2$$

(all elements of X_1 are *independent* of all elements of X_2)

Recapping the assumptions: $\epsilon_j = \epsilon_{1j} + \epsilon_{2j}$

- $E(\epsilon_{1j}) = 0, E(\epsilon_{2j}) = 0 \Rightarrow E(\epsilon_j) = 0$
- $var(\epsilon_{1j}) = \sigma_1^2, var(\epsilon_{2j}) = \sigma_2^2$, and $\epsilon_{1j} \perp \epsilon_{2j}$ for all j
 $\Rightarrow var(\epsilon_j) = \sigma_1^2 + \sigma_2^2$
- $cov(\epsilon_{1j}, \epsilon_{1j'}) = 0, cov(\epsilon_{2j}, \epsilon_{2j'}) = \sigma_2^2 c(|t_j - t_{j'}|) = \sigma_2^2 e^{-\phi(t_j - t_{j'})^2}$
- $(\epsilon_{11}, \dots, \epsilon_{1n})^T$ has *mean vector 0* and *covariance matrix*

$$\begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_1^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_1^2 \end{pmatrix} = \sigma_1^2 I_n$$

and has a *multivariate normal distribution*

Recapping the assumptions: $\epsilon_j = \epsilon_{1j} + \epsilon_{2j}$

- $(\epsilon_{21}, \dots, \epsilon_{2n})^T$ has *mean vector 0* and *covariance matrix*

$$\sigma_2^2 \begin{pmatrix} 1 & e^{-\phi(t_1-t_2)^2} & \dots & e^{-\phi(t_1-t_n)^2} \\ e^{-\phi(t_1-t_2)^2} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & e^{-\phi(t_{n-1}-t_n)^2} \\ e^{-\phi(t_1-t_n)^2} & \dots & e^{-\phi(t_{n-1}-t_n)^2} & 1 \end{pmatrix} = \sigma_2^2 \Gamma$$

and has a *multivariate normal distribution*

- So $(\epsilon_1, \dots, \epsilon_n)^T = (\epsilon_{11}, \dots, \epsilon_{1n})^T + (\epsilon_{21}, \dots, \epsilon_{2n})^T$ has mean vector 0 and covariance matrix

$$\sigma_1^2 I_n + \sigma_2^2 \Gamma$$

Thus: $(Y_1, \dots, Y_n)^T$

- $E(Y_j) = f(t_j, \theta) + E(\epsilon_j) = f(t_j, \theta)$
- Thus, may think of $f(t, \theta)$ as the result of averaging across all possible *sample paths of the fluctuation process* and *measurement errors*, so representing the “*inherent trajectory*” for subject 12
- $\text{var}(Y_j) = \text{var}(\epsilon_j) = \sigma_1^2 + \sigma_2^2$
- $\text{cov}(Y_j, Y_{j'}) = \text{cov}(\epsilon_j, \epsilon_{j'}) = \sigma_2^2 \exp\{-\phi(t_j - t_{j'})^2\}$
- Y_j is normally distributed
- The *random vector* $Y = (Y_1, \dots, Y_n)^T$ has a *multivariate normal distribution* with *mean vector* and *covariance matrix*

$$f(\theta) = \{f(t_1, \theta), \dots, f(t_n, \theta)\}^T \quad \text{and} \quad \sigma_1^2 I_n + \sigma_2^2 \Gamma$$

More succinctly: Write

$$Y \sim \mathcal{N}_n\{f(\theta), \sigma_1^2 I_n + \sigma_2^2 \Gamma\} \quad (1)$$

- Each *marginal* is a normal density, e.g. $Y_j \sim \mathcal{N}\{f(t_j, \theta), \sigma_1^2 + \sigma_2^2\}$

Simplifications: We may be willing to make *simplifying assumptions*

- If the t_j are *far apart in time*, $|t_j - t_{j'}|$ may be *large*, and hence $\exp\{-\phi(t_j - t_{j'})^2\}$ *close to zero* \Rightarrow “correlation among fluctuations at t_1, \dots, t_n is *negligible*”
- *Approximate* by assuming $\epsilon_{2j} \perp \epsilon_{2j'} \Rightarrow \text{cov}(\epsilon_{2j}, \epsilon_{2j'}) = 0$ and thus $\Gamma = I_n$, which implies

$$Y_j \perp Y_{j'} \Rightarrow \text{cov}(Y_j, Y_{j'}) = 0,$$

and $\text{var}(Y_j) = \sigma^2 = \sigma_1^2 + \sigma_2^2$

- The *statistical model* becomes

$$Y \sim \mathcal{N}_n\{f(\theta), \sigma^2 I_n\}, \quad \psi = (\theta^T, \sigma^2)^T \quad (2)$$

Key point: A *statistical model* like (1) or (2) thus describes *all possible probability distributions* for *random vector* Y representing the *data generating mechanism* for observations we might see at t_1, \dots, t_n

- E.g., for (1), *possible probability distributions* are specified by different values of the *parameter* $\psi = (\theta^T, \sigma_1^2, \sigma_2^2, \phi)^T \in \Psi$
- *The big question:* Which value of ψ truly governs the mechanism?
- In particular, we are interested in $\theta - (\sigma_1^2, \sigma_2^2, \phi)$ are required to describe things fully, but are a *nuisance* ... more later)

Objective: If we *collect data* [so observe *a single realization* of $Y = (Y_1, \dots, Y_n)^T$], what can we learn about ψ ?

- ...and how can we account for the fact that things could have turned out *differently* (i.e., a *different realization*)?

2. Statistical inference

Consider: A *statistical model* for a *data generating mechanism* like
(2)

$$Y_j = f(t_j, \theta) + \epsilon_j$$

- Y_1, \dots, Y_n are *independent* with

$$Y = (Y_1, \dots, Y_n)^T \sim \mathcal{N}_n\{f(\theta), \sigma^2 I_n\}, \quad \psi = (\theta^T, \sigma^2)^T$$

so that each Y_j satisfies

$$Y_j \sim \mathcal{N}\{f(t_j, \theta), \sigma^2\}$$

- That is, Y_j , $j = 1, \dots, n$, may be regarded as representing the result of a potential “*draw*” from the *normal probability distribution* describing how such observations would “*turn out*” at t_j (given phenomena like measurement error and “biological fluctuation”)

Conceptually:

- Think of the statistical model as a *formal representation* of the “*population*” of *all possible realizations* of Y_1, \dots, Y_n we would ever see
- When we collect data, we observe a *sample* from the *population*; i.e., we get to see a *single realization* of $Y_1, \dots, Y_n, y_1, \dots, y_n$

Objective, restated: What can we learn about the “*true value*” of ψ (which determines the nature of the *population*) from a *sample*?

- We do not get to observe the entire population (or else we’d know ψ)
- How *uncertain* will we be?

Statistical inference (loosely speaking): Making statements about a *population* of interest on the basis of only a *sample* from the population

Statistic: Any *function* of a random variable(s)

Parameter (point) estimation: Construct a *function* of Y_1, \dots, Y_n that, if evaluated at a *particular realization* y_1, \dots, y_n , yields a numerical value that gives information on the *true value* of ψ

- *Estimator*: The function itself
- *Estimate*: The numerical value based on a particular realization
- *Estimation*: Used both to denote the procedure (*estimator*) and actual calculation of a numerical value (*estimate*)

Example: Suppose that $f(t, \theta) = \theta_1 + \theta_2 t$, $\theta = (\theta_1, \theta_2)^T$

- *Simple linear regression model*
- *Usual assumption* – $Y_j \sim \mathcal{N}(\theta_1 + \theta_2 t_j, \sigma^2)$, $Y_1, \dots, Y_n \perp$

Standard OLS estimator and estimate for θ –

- The *ordinary least squares estimator* is defined by

$$\hat{\theta}_{OLS}(Y) = \hat{\theta}(Y) = \arg \min_{\theta \in \Theta} \sum_{j=1}^n (Y_j - \theta_1 - \theta_2 t_j)^2$$

- For a particular data set $\{y_j\}$, a realization of Y , the estimate of θ is defined by

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \sum_{j=1}^n (y_j - \theta_1 - \theta_2 t_j)^2$$

- Both $\hat{\theta}(Y) = \hat{\theta}^n(Y)$ and $\hat{\theta} = \hat{\theta}^n$ depend on n —**IMPORTANT** later!

Remark: The notational distinction between *estimator* and *estimate* is often subject to *abuse* in the literature

- Meaning is often clear from the context

Closed form solution: Defining

$$X = \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_n \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

and minimizing $|y - X\theta|^2$ yields the *estimator* and *estimate*

$$\hat{\theta}(Y) = (X^T X)^{-1} X^T Y \quad \text{and} \quad \hat{\theta} = (X^T X)^{-1} X^T y$$

- *Convention* – “ $\hat{}$ ” means *estimator* or *estimate* (for θ here)
- *Emphasis* that the *estimator* $\hat{\theta}(Y)$ is a *function* of Y_1, \dots, Y_n is usually (by statisticians) suppressed

Motivation for ordinary least squares: Coming up

- Separate estimation of σ^2 :
$$\hat{\sigma}^2 = (n - 2)^{-1} \sum_{j=1}^n (Y_j - \hat{\theta}_1 - \hat{\theta}_2 t_j)^2$$
- $\hat{\psi} = (\hat{\theta}^T, \hat{\sigma}^2)^T$

Question: How “*good*” is using $\hat{\theta} = \hat{\theta}(Y)$ as an *estimator* for the *true value* of θ ?

- A question about the *procedure*
- What do we mean by “*good*?” Can we characterize the *extent of uncertainty*?

Key idea:

- An *estimator* is a *function* of random variables that represent the *data generating mechanism*
- Thus, for any value of ψ , the *estimator* has *probability distribution* (that depends on that of Y_1, \dots, Y_n and hence on ψ)

Conceptually:

- *Each possible realization* of Y_1, \dots, Y_n would yield a (numerical) value of $\hat{\theta}$; i.e., the *random vector* $\hat{\theta}(Y)$ has *sample space* consisting of all these values
- We may thus think of the *probability distribution* of $\hat{\theta}(Y)$ as representing probabilities for values of $\hat{\theta}$ that would be observed across “*all possible data sets*” we could “*end up with*”
- When we *collect data*, we end up with *only one* of these
- If this distribution has *large variance*, estimates *vary a lot* across possible data sets, and our one estimate tells us little about the true value of ψ (another data set might have yielded a very *different* numerical value) \Rightarrow *lots of uncertainty*
- On the other hand, if this distribution has *small variance*, estimates *vary little* across possible data sets, and our one estimate may be quite informative about the true value of ψ (another data set might have yielded a *similar* value) \Rightarrow *only mild uncertainty*

Sampling distribution: The probability distribution of an estimator

- *Properties* of the sampling distribution characterize the *uncertainty* in the estimation procedure (*estimator*)

Unbiasedness: Intuitively, for a particular ψ (or its components)

- Some values in the sample space are larger than ψ , some smaller
- All possible values of *estimator* for ψ should “*average out*” to ψ
- Formally, writing $E_{\psi}(\cdot)$ to denote expectation with respect to the distribution of Y_1, \dots, Y_n under ψ

$$E_{\psi}(\hat{\psi}) = \psi \quad (1)$$

- An *estimator* that satisfies (1) is called an unbiased estimator

Sampling variance and standard error: Quantify *variation*

across possible values of ψ

- The *sampling covariance matrix* is the covariance matrix of the *sampling distribution* of $\hat{\psi}(Y)$

$$\text{var}_{\psi}(\hat{\psi})$$

- Diagonal elements of $\text{var}_{\psi}(\hat{\psi})$ are the *sampling variances* of the components of $\hat{\psi}(Y)$, e.g., $\text{var}(\hat{\psi}_k)$ for the k th component
- The *standard error* is the *standard deviation* of the k th component of $\hat{\psi}(Y)$, $\sqrt{\text{var}(\hat{\psi}_k)}$ (on the same scale as ψ)

Key factors determining magnitude of sampling variance:

- Variance of *original probability distribution* of Y (very well *out of our control*)
- *Sample size* n (often *under our control*)

Example, continued: *Sampling distribution of ordinary least squares estimator* (using fun facts)

- The estimator is *unbiased*: $E(Y) = X\theta$

$$\Rightarrow E(\hat{\theta}) = E\{(X^T X)^{-1} X^T Y\} = (X^T X)^{-1} X^T X\theta = \theta$$

- The *sampling variance*: $\text{var}(Y) = \sigma^2 I_n$

$$\begin{aligned}\Rightarrow \text{var}(\hat{\theta}) &= \text{var}\{(X^T X)^{-1} X^T Y\} = (X^T X)^{-1} X^T \text{var}(Y) X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1}\end{aligned}$$

- The *sampling distribution* itself is *multivariate normal*:

$$\hat{\theta} \sim \mathcal{N}_2\{\theta, \sigma^2 (X^T X)^{-1}\}$$

- *Sampling variance* depends on σ^2 (variance of Y_j , original probability distribution)
- Writing $(X^T X)^{-1} = n^{-1} \{n^{-1} X^T X\}^{-1}$, *sampling variance decreases* as n *increases* (or so one might intuitively expect!!!)

Absolute necessity: Whenever an estimate based on data is reported, it should be accompanied by an assessment of uncertainty based on the sampling distribution

- *Natural approach*: look at the *standard error*; here, for the k component of $\hat{\theta}$, $\sqrt{\text{var}(\hat{\theta}_k)} = \sqrt{\sigma^2(X^T X)_{kk}^{-1}}$
- σ^2 is NOT *known*, so provide an *estimate* of the *standard error* of the k th component (usually *itself* called “*standard error*”)
- In the example, $SE(\hat{\theta}_k) = \sqrt{\hat{\sigma}^2(X^T X)_{kk}^{-1}}$, $k = 1, 2$
- How to *interpret*?
 - (i) If $\hat{\theta}_2 = 2.0$ and $SE(\hat{\theta}_2) = 3.0 \Rightarrow$ *very uncertain*
 - (ii) If $\hat{\theta}_2 = 2.0$ and $SE(\hat{\theta}_2) = 0.03 \Rightarrow$ *feeling pretty good!*
- Ways to *improve precision* (*reduce uncertainty*) if case (i)?

In fact: We can do more: The *entire sampling distribution* can help us to make a *probability statement* to better characterize *uncertainty*

- For example, if the probability distribution of Y is indexed by ψ ,

$$Z = \frac{\hat{\theta}_k - \theta_k}{\sqrt{\sigma^2 (X^T X)^{-1}_{kk}}} \sim \mathcal{N}(0, 1) \text{ “standard normal distribution,” } k = 1, 2$$

- When σ^2 is replaced by $\hat{\sigma}^2$, $T = \frac{\hat{\theta}_k - \theta_k}{SE(\hat{\theta}_k)} \sim t_{n-2}$

“*Student’s t distribution with $n - 2$ degrees of freedom*”

- Let $\alpha =$ some small probability, so $1 - \alpha$ is large; e.g., $\alpha = 0.05$

- Let $t_{1-\alpha/2}$ satisfy $P(\{T \geq t_{1-\alpha/2}\}) = \alpha/2$, so, by *symmetry*,
 $P\{T \leq -t_{1-\alpha/2}\} = \alpha/2 \Rightarrow P\{-t_{1-\alpha/2} \leq T \leq t_{1-\alpha/2}\} = 1 - \alpha$

$$P\{\hat{\theta}_2 - t_{1-\alpha/2}SE(\hat{\theta}_2) \leq \theta_2 \leq \hat{\theta}_2 + t_{1-\alpha/2}SE(\hat{\theta}_2)\} = 1 - \alpha$$

- A probability pertaining to the *sampling distribution* of $\hat{\theta}_2$: For “*all possible realizations of Y of size n ,*” probability is $1 - \alpha$ that the *endpoints* of the *interval* $[\hat{\theta}_2 - t_{1-\alpha/2}SE(\hat{\theta}_2), \hat{\theta}_2 + t_{1-\alpha/2}SE(\hat{\theta}_2)]$ will include (the *fixed value*) θ_2

Confidence interval: A probability statement about the *procedure* by which an *estimator* is constructed from a realization of Y

- NOT a probability statement about θ_2 (*fixed*)
- *Interpretation:* “For all possible realizations of Y of size n , if we were to calculate the interval according to the *estimation procedure*, $(1 - \alpha)\%$ of such intervals would ‘*cover*’ θ_2 ”
- Provides *more information* than just *standard error* alone – how “large” or “small” SE must be *relative to* $\hat{\theta}_2$ to feel “*confident*” that the procedure of data generation and estimation provides a reliable understanding (“*confident*” quantified by α)
- α is chosen by the *analyst*
- For the example, $\alpha = 0.05$, n large $\Rightarrow t_{1-\alpha/2} \approx 1.96$
 - (i) $\hat{\theta}_2 = 2.0$, $SE(\hat{\theta}_2) = 3.0$ gives $[-3.88, 7.88] \Rightarrow$ *no confidence*
 - (ii) $\hat{\theta}_2 = 2.0$, $SE(\hat{\theta}_2) = 0.03$ gives $[1.94, 2.09] \Rightarrow$ *feeling pretty confident!*

Warning: The *numerical values themselves* are *meaningless* except for the *impression* they give about the *quality* of the *procedure*

- *Once a realization of data is in hand*, the interval either “*covers*” θ_2 or it *doesn't*
- *Wrong interpretation:* The probability is $1 - \alpha$ that θ_2 is between -3.88 and 7.88 .
- *What we CAN say:* We are $(1 - \alpha)\%$ “*confident*” that intervals constructed this way would “*cover*” $\theta_2 \Rightarrow$ the numerical values give only *a sense of the faith* we should attach to results of gathering data the way we did

Inference for the Nonlinear System Estimation Problem:

We consider the estimation problem with general nonlinear system

$$\frac{d\bar{x}}{dt}(t) = \bar{g}(t, \bar{x}(t), \theta)$$

with observations

$$\bar{Y}(t) = \bar{f}(t, \theta) + \bar{\epsilon}(t) = C\bar{x}(t, \theta) + \bar{\epsilon}(t)$$

where $\bar{x} \in R^N$, $\bar{f} \in R^m$ and $\theta \in R^p$.

We first discuss the case for $m = 1$, i.e., the observation system is scalar and $f(t, \theta) = C\bar{x}(t, \theta)$ where C is a $1 \times N$ array. As usual, we assume n scalar longitudinal observations

$$Y_j = f(t_j, \theta_0) + \epsilon_j, \quad j = 1, 2, \dots, n,$$

where

$$E[Y_j] = f(t_j, \theta_0), \quad \text{var}[Y_j] = \sigma_0^2.$$

Asymptotic distribution theory and standard error analysis:

As $n \rightarrow \infty$, we have that

$$\hat{\theta}_{OLS}^n(Y) \approx \sim \mathcal{N}_p \left(\theta_0, \sigma_0^2 [\chi^T(\theta_0)\chi(\theta_0)]^{-1} \right) = \mathcal{N}_p(\theta_0, \Sigma_0) \quad (2)$$

where $\chi(\hat{\theta}) = \frac{\partial F}{\partial \theta}(\theta) = F_\theta(\theta)$ is the $n \times p$ matrix with elements

$$\chi_{jk}(\theta) = \frac{\partial f(t_j, \theta)}{\partial \theta_k}. \quad (3)$$

Here and later we use the notation $F(\theta) = (f(t_1, \theta), \dots, f(t_n, \theta))$ and

$$F_\theta(\theta) = \begin{pmatrix} f_\theta(t_1, \theta) \\ \vdots \\ f_\theta(t_n, \theta) \end{pmatrix}.$$

In expression (1), θ_0 denotes the theoretical “true” parameter value and σ_0 is the true variance for the system which is being observed (both of which are generally unknown).

Approximating the covariance:

Since θ_0 is unknown, we approximate our covariance matrix using our estimate

$$\theta_0 \approx \hat{\theta} = \hat{\theta}_{OLS}(y_1, \dots, y_n) = \arg \min \sum_{j=1}^n (y_j - f(t_j, \theta))^2$$

and

$$\sigma_0^2 \approx \hat{\sigma}^2 = \frac{1}{n-p} \sum_{j=1}^n \left(y_j - f(t_j, \hat{\theta}) \right)^2$$

Then we have

$$\Sigma_0 \approx \Sigma = \hat{\sigma}^2 \left[\chi^T(\hat{\theta}) \chi(\hat{\theta}) \right]^{-1},$$

and the standard errors for $\hat{\theta}_k$ are then given by $SE(\hat{\theta}_k) = \sqrt{\Sigma_{kk}}$.

Computing the covariance: We can easily compute $\hat{\sigma}^2$ by multiplying our cost criterion or residual at $\hat{\theta}$ by an appropriate conversion factor. Computing χ is somewhat more involved in the general nonlinear system case. There are several ways to do this:

- **Forward differencing:**

$$\chi_{jk}(\theta) = \frac{\partial f(t_j, \theta)}{\partial \theta_k} \approx \frac{f(t_j, \theta + h_k) - f(t_j, \theta)}{h_k}. \quad (4)$$

But question: how to choose h_k ???

- **Sensitivity equations:** Recall that

$$f_{\theta}(t, \theta) = C \frac{\partial \bar{x}}{\partial \theta}(t, \theta)$$

where $\bar{z}(t, \theta) = \frac{\partial \bar{x}}{\partial \theta}(t, \theta)$ satisfies

$$\frac{d\bar{z}}{dt}(t) = \frac{\partial \bar{g}}{\partial \bar{x}} \bar{z}(t) + \frac{\partial \bar{g}}{\partial \theta}. \quad (5)$$

Vector systems with partial observations: Suppose now we have a vector system (dimension N) with partial observations, say m coordinate observations where $m \leq N$. In this case, we have

$$\frac{d\bar{x}}{dt}(t) = \bar{g}(t, \bar{x}(t), \theta) \quad (6)$$

and

$$\bar{y}_j = \bar{f}(t_j, \theta) + \bar{\epsilon}_j = \mathcal{C}\bar{x}(t_j, \theta) + \bar{\epsilon}_j, \quad (7)$$

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

$$\bar{\epsilon}_j \sim \mathcal{N}_m(\bar{0}, V)$$

where $V = \text{diag}(\sigma_1^2, \dots, \sigma_m^2)$. As usual, we have assumed that the errors $\bar{\epsilon}_j, j = 1, 2, \dots, n$ are independently distributed.

The corresponding OLS problem consists of minimizing

$$\sum_{j=1}^n (\bar{y}_j - \bar{f}(t_j, \theta))^T V^{-1} (\bar{y}_j - \bar{f}(t_j, \theta)).$$

As $n \rightarrow \infty$, one has

$$\hat{\theta}_{OLS} \sim \mathcal{N}_p(\theta_0, \Sigma_0)$$

where

$$\Sigma_0 = \left(\sum_{j=1}^n D_j^T(\theta_0) V^{-1} D_j(\theta_0) \right)^{-1}$$

and the $m \times p$ matrix $D_j(\theta)$ is given by

$$D_j(\theta) = \frac{\partial \bar{f}}{\partial \theta}(t_j, \theta). \quad (8)$$

The entries in D_j can be computed as in the scalar case using either forward differencing or the appropriate sensitivity equations.

Warning: For general nonlinear (in the parameters) systems, the asymptotic distribution results are based on a *linearization* of the output or solution. Thus, in the case that $\theta \rightarrow \bar{f}(t, \theta)$ is nonlinear (which occurs even in general *linear* dynamical systems

$$\frac{dx}{dt}(t) = A(\theta)x(t)$$

with parameter dependent coefficient), the asymptotic distributional results are *only* an **APPROXIMATION**. To see this, consider again (5),(6) under the assumption $\bar{\epsilon}_j \sim \mathcal{N}_m(\bar{0}, V)$ so that the OLS problem corresponds to minimizing

$$J(\theta) = \sum_{j=1}^n (\bar{y}_j - \bar{f}(t_j, \theta))^T V^{-1} (\bar{y}_j - \bar{f}(t_j, \theta)).$$

Let θ_0 be the true value of θ and approximate \bar{f} by linearization about θ_0 . That is, we let

$$\bar{f}(t_j, \theta) \approx \bar{f}(t_j, \theta_0) + D_j(\theta_0)\delta\theta,$$

where $\delta\theta = \theta - \theta_0$ and $D_j(\theta_0) = \frac{\partial \bar{f}}{\partial \theta}(t_j, \theta_0)$. Then the OLS functional associated with the estimator $\hat{\delta\theta}_{OLS}$ is given by

$$J(\delta\theta) = \sum_{j=1}^n (\bar{z}_j - D_j(\theta_0)\delta\theta)^T V^{-1} (\bar{z}_j - D_j(\theta_0)\delta\theta),$$

where $\bar{z}_j = \bar{y}_j - \bar{f}(t_j, \theta_0)$. The corresponding random variable $Z_j = Y_j - \bar{f}(t_j, \theta_0)$ is simply a translate of Y_j that satisfies

$$E[Z_j] = 0, \quad \text{var}[Z_j] = V.$$

The minimizing conditions for $\delta\theta$ are then

$$\sum_{j=1}^n X_j^T V^{-1} (\bar{z}_j - X_j \delta\theta) = 0,$$

where $X_j \equiv D_j(\theta_0)$ is the $m \times p$ sensitivity matrix.

Thus we have for $Z = (Z_1, Z_2, \dots, Z_n)$

$$\widehat{\delta\theta}(Z) = \left(\sum_{j=1}^n X_j^T V^{-1} X_j \right)^{-1} \sum_{j=1}^n X_j^T V^{-1} Z_j = \Sigma_0 \sum_{j=1}^n X_j^T V^{-1} Z_j,$$

with $E[\widehat{\delta\theta}] = 0$ (since $E[Z_j] = 0$) and $\Sigma_0 \equiv \{ \sum_{j=1}^n X_j^T V^{-1} X_j \}^{-1}$.

The corresponding covariance matrix

$$\text{cov}[\widehat{\delta\theta}] = E[\widehat{\delta\theta} \widehat{\delta\theta}^T]$$

is readily computed as

$$\begin{aligned} E[\widehat{\delta\theta} \widehat{\delta\theta}^T] &= E[\Sigma_0 \left(\sum_{j=1}^n X_j^T V^{-1} Z_j \right) \left(\sum_{k=1}^n Z_k^T V^{-1} X_k \right) \Sigma_0^T] \\ &= \Sigma_0 \left(\sum_{j=1}^n X_j^T V^{-1} V V^{-1} X_j \right) \Sigma_0^T = \Sigma_0 \Sigma_0^{-1} \Sigma_0^T = \Sigma_0, \end{aligned}$$

since $E[Z_j Z_k^T] = V$ with V symmetric.

Thus, as $n \rightarrow \infty$, we have

$$\widehat{\delta\theta}_{OLS} \sim \mathcal{N}_p(0, \Sigma_0)$$

and hence $\widehat{\theta}_{OLS} \approx \theta_0 + \widehat{\delta\theta}_{OLS}$ has approximate asymptotic distribution

$$\widehat{\theta}_{OLS} \sim \mathcal{N}_p(\theta_0, \Sigma_0).$$

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

- **Remark 1:** 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 $f(t_j, \theta) = \mathcal{C}x(t_j, \theta)$ as mentioned above. That is, the **sensitivity matrix**

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

depends on the number n and **nature** (e.g., how taken) of the **sampling times** $\{t_j\}$. Moreover, it is the matrix $[\chi^T \chi]^{-1}$ and the parameter $\hat{\sigma}^2$ that ultimately determine the SE and CI. At first investigation of

$$\sigma_0^2 \approx \hat{\sigma}^2 = \frac{1}{n-p} \sum_{j=1}^n \left(y_j - f(t_j, \hat{\theta}) \right)^2,$$

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

$$\frac{1}{n} \chi^T(\theta) \chi(\theta) \rightarrow \Upsilon(\theta) \quad \text{as } n \rightarrow \infty$$

for some **nonsingular** matrix $\Upsilon(\theta_0)$. 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] 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. Inverse Ill-posed Problems* **15** (2007), 1–18 and [BDE] H.T. Banks, S. Dediu and S.E.

Ernstberger, Sensitivity functions and their uses in inverse problems, CRSC-TR07-??, July, 2007; *J. Inverse and Ill-posed Problems*, submitted.)

- **Remark 2:** Since the computations for standard errors and confidence intervals (and also those for the *model comparison tests* outlined in the next section) 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 $\epsilon_j, j = 1, 2, \dots, n$, are *independent identically distributed (i.i.d.)* random variables with mean $E[\epsilon_j] = 0$ and *constant variance* $var[\epsilon_j] = \sigma_0^2$. After carrying out the estimation procedures, one can readily **plot** the *residuals vs. time* and the *residuals vs. the resulting estimated model (output or observation f) values*. A random pattern for

the first is strong support for validity of the independence assumption while a non increasing, random pattern for the latter suggests the assumption of constant variance may be reasonable for the data (measurements) used in the inverse problem calculations. The underlying assumption that the sampling size n must be large (recall the theory is asymptotic in that it holds as $n \rightarrow \infty$) is not so readily “verified” and is often ignored (albeit at the user’s peril in regard to the quality of the uncertainty findings). Indeed the asymptotic theories are often used in a very heuristic underlying manner to give [a loose feeling for the uncertainty](#) involved in the estimates and the level of parametrization used in approximating the underlying mathematical model. [It is often the case that the asymptotic results provide remarkably good approximations to the true sampling distributions for finite \$n\$.](#) However, in practice one has no way to ascertain whether this holds for a specific example of interest.

MODEL COMPARISON TECHNIQUES

H.T. Banks

Center for Research in Scientific Computation

and

Center for Quantitative Sciences in Biomedicine

North Carolina State University

Raleigh, NC 27695-8205

NC STATE UNIVERSITY

*Center for Research
in Scientific Computation
North Carolina State University*

Motivation

- Frequent QUESTION in modeling studies [BKa, BK]: can mathematical model be **improved** by **more detail** and/or **further refinement**
- **EXAMPLE:** More **detail** in a given mechanism (constant rate vs. time or spatially dependent rate—REF:[BBDS] **mortality rates** during sub-lethal damage in insect populations exposed to various levels of pesticides)
- **EXAMPLE:** Does an **additional mechanism** in the model produce a better fit to data—REF:[BF1, BF2, BKa] **diffusion alone** or **diffusion plus convection** in cat brain transport in grey vs. white matter

Important Remarks

- In model comparison results outlined below, there are really two models being compared: the **math model** and the **statistical model**.
- If one embeds the math model in the **wrong statistical model** (for example, assumes constant variance when it really isn't true), then the math model comparison results will be **invalid** (e.g., **worthless**).
- The **key** to all this is that you must have the math model you want to simplify or improve (e.g., test $\mathcal{V} = 0$ in the example below) embedded in the **correct statistical model**, so that the comparison really is **only with regard to the math model**.

EXAMPLE 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. The cat's brain contains grey and white matter[BKa]. In general, the transport of substance in cat's brains can be described by a PDE describing change in time and space. This model, which is widely discussed in the applied mathematics and engineering literature, has the form

$$\frac{\partial u}{\partial t} + \mathcal{V} \frac{\partial u}{\partial x} = \mathcal{D} \frac{\partial^2 u}{\partial x^2}. \quad (1)$$

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 (1) 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 $\{u(t_i, x_j)\}$, 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** from statistics.

ANOVA Type Statistical Tests

In general, assume we have an inverse problem $f(t, \vec{\theta})$ and are given n observations. We define

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

where our **statistical model** again has the form

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

. Here, $\vec{\theta}_0$ is the “true” value of $\vec{\theta}$ which we assume to exist. We again use Θ to represent the set of all the admissible parameters $\vec{\theta}$.

We take the standard statistical assumptions;

- A1) $\{\epsilon_j\}_{j=1}^{\infty}$ are identical independent distributed with $E(\epsilon_j) = 0$ and $\text{var}(\epsilon_j) = \sigma^2$.

Among other important hypotheses are

- A2) Θ is a compact subset of Euclidian space of R^p and $f(t, \vec{\theta})$ is continuous on $[0, T] \times \Theta$.
- A3) Observations are at $\{t_j\}_{j=1}^n$ in $[0, T]$. For some finite measure μ on $[0, T]$,

$$\frac{1}{n} \sum_{j=1}^n h(t_j) \longrightarrow \int_0^T h(t) d\mu(t)$$

as $n \rightarrow \infty$, for continuous functions h .

- A4) $J_0(\vec{\theta}) = \int_0^T (f(t, \vec{\theta}_0) - f(t, \vec{\theta}))^2 d\mu(t) = \sigma^2$ has a unique minimizer in Θ at $\vec{\theta}_0$.

Let $\theta^n = \theta_{OLS}^n(\vec{Y})$ be the OLS estimator for J_n with corresponding estimate

$$\hat{\theta}^n = \theta_{OLS}^n(\{y_j\})$$

for a realization $\vec{y} = \{y_j\}$. That is,

$$\theta^n(\vec{Y}) = \arg \min_{\vec{\theta} \in \Theta} J_n(\vec{Y}, \vec{\theta})$$

and

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

One can then establish a series of useful results (see [BF2] for detailed proofs).

- **Result 1:**

Under A1) to A4), $\theta^n \longrightarrow \vec{\theta}_0$ as $n \rightarrow \infty$ with probability 1.

- **Remarks:** In most calculations, we actually use an approximation f^N to f , often a numerical solution to the ODE or PDE for modeling our 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 dependent parameters) by finite dimensional discretizations Θ^M . For extensive discussions related to these questions, see [BK] as well as [BF2] where related assumptions A5), A6) on convergences $f^N \rightarrow f$ and $\Theta^M \rightarrow \Theta$ are given. We will ignore these issues here, keeping in mind that these approximations will also be of importance in the methodology discussed below in most practical uses.

We will need further assumptions to precede (these will be denoted by A7)–A11) to facilitate reference to [BF2]). These include:

- A7) Θ is finite dimensional in R^p and $\vec{\theta}_0 \in \Theta$.
- A8) $f : \Theta \rightarrow C[0, T]$ is C^2 function.
- A10) $\mathcal{J} = \frac{\partial^2 J_0}{\partial \vec{\theta}^2}(\vec{\theta}_0)$ is positive definite.
- A11) $\Theta_H = \{\vec{\theta} \in \Theta | H\vec{\theta} = c\}$ where H is an $r \times p$ matrix of full rank, and c is a known constant.

In many instances, including the motivating example given above, one is interested in using data to questioning whether the 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 the constraints of assumption A11).

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

Define then

$$\theta_H^n(\vec{Y}) = \arg \min_{\vec{\theta} \in \Theta_H} J_n(\vec{Y}, \vec{\theta})$$

and

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

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

$$T_n(\vec{Y}) = n(J_n(\vec{Y}, \theta_H^n) - J_n(\vec{Y}, \theta^n))$$

and

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

One can establish asymptotic convergence results for the test statistics $T_n(\vec{Y})$, as given in detail in [BF2]. 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

$$U_n(\vec{Y}) = \frac{T_n(\vec{Y})}{J_n(\vec{Y}, \theta_n)}, \quad (2)$$

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

BIG Result :

Under the assumptions A1)–A11) above and the assumption that H_0 is true,

$$U_n \xrightarrow{\mathcal{D}} \Upsilon(r)$$

as $n \rightarrow \infty$ where $\Upsilon \sim \chi^2(r)$, a χ^2 with r degrees of freedom.

An example of this the χ^2 density is depicted in Figure 1 where the density for $\chi^2(4)$ (χ^2 with $r = 4$ degrees of freedom) is graphed.

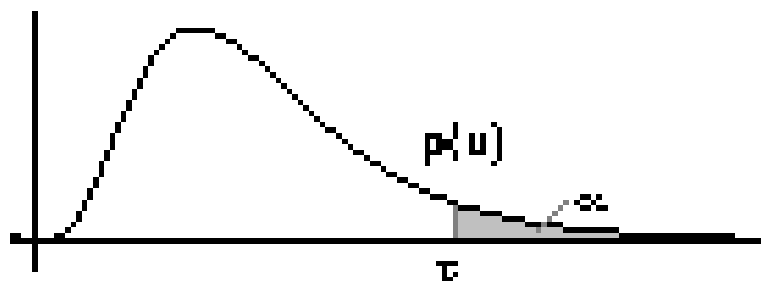


Figure 1: Example of $U \sim \chi^2(4)$ density

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 , for $U_n \sim \chi^2(r)$. If the test statistic, $\hat{U}_n > \tau$, then we *reject* H_0 as false with *confidence level* $(1 - \alpha)100\%$. Otherwise, we *accept* 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.

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

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

p-value:

The minimum value of α at which H_0 can be rejected, α^* , is called the *p-value*. Thus, the smaller the p-value, the greater the significance of the additional parameters/mechanisms, i.e., the more likely the term should be in the model.

Implementation: Once we compute $\hat{U}_n = \bar{\tau}$, then $p = \alpha^*$ is the value that corresponds to $\bar{\tau}$ on 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 reject H_0 at confidence level $(1 - \alpha^*)100\% = 98.18\%$ or lower. For more information, see ANOVA in any good statistics book.

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 $P(\chi^2(s) > \tau) = \alpha$. We next compute $\hat{U}_n = \bar{\tau}$ and compare it to τ . If $\hat{U}_n > \tau$, then we **reject** H_0 as false; otherwise, we accept the null hypothesis H_0 .

Revisiting the cat-brain problem: 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,

$$J_n(\hat{\theta}^n) = 106.15$$

$$J_n(\hat{\theta}_H^n) = 180.17,$$

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

$$J_n(\hat{\theta}^n) = 14.68$$

$$J_n(\hat{\theta}_H^n) = 15.35,$$

thus, in this case, we have $\hat{U}_n = .365$, which implies we **accept** 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

$$J_n(\hat{\theta}^n) = 7.8$$

$$J_n(\hat{\theta}_H^n) = 146.71,$$

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.

CONCLUSIONS

The difference in conclusions between the first and last sets and that of the second set is interesting.

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 conducive to observed microscopic tests on the various matter (micro channels in white matter that promotes convective “flow”). Thus, it can be concluded with a reasonably high degree of confidence, that white matter has convective properties, while grey matter does not.

References

- [BBDS] 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*, to appear.
- [BF1] 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.
- [BF2] 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.

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

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

An Inverse Problem Statistical Methodology Summary

H.T. Banks, M. Davidian and J.R. Samuels, Jr.
Center for Research in Scientific Computation
and
Center for Quantitative Sciences in Biomedicine
North Carolina State University
Raleigh, NC 27695-8205

July 20, 2007

!!DRAFT!!

Outline

1. Parameter Estimation: MLE, OLS, GLS
2. Computation of Σ , Standard Errors and Confidence Intervals
3. Model Comparison Techniques

1 Parameter Estimation: MLE, OLS, and GLS

1.1 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**

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

with **observation process**

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

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

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

In general the corresponding observations or data $\{\vec{y}_j\}$ will not be exactly $\vec{y}(t_j)$ and hence we choose to treat this uncertainty pertaining to the observations with a statistical model for the observation process.

1.2 Description of Statistical Model

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

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

where $\vec{f}(t_j, \vec{\theta}) = \mathcal{C}\vec{x}(t_j; \vec{\theta})$, $j = 1, \dots, n$, 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 term $\vec{\epsilon}_j$ 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 $\vec{f}(t, \vec{\theta})$. Since these fluctuations are unknown to the modeler, we will assume $\vec{\epsilon}_j$ is generated from a probability distribution 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 $\vec{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^T$ might be the multivariate normal distribution [1].

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 the variance of the error $\text{var}(\vec{\epsilon}_j)$. 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.

1.3 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 $\vec{\epsilon}_j = 1, \dots, n$, are independent and identically distributed. 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

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

The resulting solutions θ_{MLE} and σ_{MLE}^2 are the maximum likelihood **estimators** (MLEs) for $\vec{\theta}_0$ and σ_0^2 , respectively. We point out that these solutions $\theta_{\text{MLE}} = \theta_{\text{MLE}}(\vec{Y})$ and $\sigma_{\text{MLE}}^2 = \sigma_{\text{MLE}}^2(\vec{Y})$ 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 $\{Y_j\}$ replaced by a given realization $\vec{y} = \{y_j\}$ and will be denote by $\hat{\theta}_{\text{MLE}}$ and $\hat{\sigma}_{\text{MLE}}^2$ respectively.

Maximizing (5) is equivalent to maximizing the log likelihood

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

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

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

We see that solving (7) is the same as the least squares optimization

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

We 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

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

Note that we can solve for θ_{MLE} and σ_{MLE}^2 separately – a desirable feature, but one that won't 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_{MLE}^2 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

$$\vec{Y}_j = \vec{f}(t_j, \vec{\theta}_0) + \vec{\epsilon}_j \quad (10)$$

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

$$V_0 = \text{var}(\vec{\epsilon}_j) = \text{diag}(\sigma_{0,1}^2, \dots, \sigma_{0,m}^2) \quad (11)$$

for $j = 1, \dots, n$. 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 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 log of the likelihood function for $\vec{\epsilon}_j = \vec{Y}_j - \vec{f}(t_j, \vec{\theta})$ defined by

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

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

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

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

Unfortunately, this is a coupled system, which requires some care when solving numerically. We will discuss this issue further in Sections 1.4.2 and 1.4.5 below.

1.4 Unspecified Error Distributions and Asymptotic Theory

In section 1.3 we 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? In this section we will review two estimation procedures for such situations: ordinary least squares (OLS) and generalized least squares (GLS).

1.4.1 Ordinary Least Squares (OLS)

The statistical model in the scalar case takes the form

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

where the variance $\text{var}(\epsilon_j) = \sigma_0^2$ is constant in longitudinal data (note that the error's distribution is not specified). If we define

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

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 [13] to solving for $\vec{\theta}$ in

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

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

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

provides an realization for θ_{OLS} .

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

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

by $\hat{\theta}_{\text{OLS}}$ to obtain an estimate $\hat{\sigma}_{\text{OLS}}^2$ 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} [20]. As will be explained in more detail below, as $n \rightarrow \infty$, we have that

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

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

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

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

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

and the *bias adjusted* estimate

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

to use as an approximation in (19).

We note that (21) represents the estimate for σ_0^2 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 [20] and p. 63 of [13]). We remark that (18) is true even in the general nonlinear case (it does not rely on any asymptotic theories).

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

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

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

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

$$\vec{Y}_j = \vec{f}(t_j, \vec{\theta}_0) + \vec{\epsilon}_j \quad (23)$$

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

$$V_0 = \text{var}(\vec{\epsilon}_j) = \text{diag}(\sigma_{0,1}^2, \dots, \sigma_{0,m}^2) \quad (24)$$

for $j = 1, \dots, n$. 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., $V_0 = \text{diag}(v_1, \dots, v_m)$ 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

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

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

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

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

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

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} :

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

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

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 (25). As $n \rightarrow \infty$, θ_{OLS} has the following asymptotic properties [1]:

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

where

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

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

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

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 least squares estimator θ_{OLS} :

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

where

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

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 (28) and (29) for the vector statistical model (23).

1.4.2 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 (20) and then calculate $\hat{\sigma}_{\text{OLS}}^2$ 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:

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

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

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.

1.4.3 Generalized Least Squares (GLS)

Although in Section 1.4.1 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 [13] is

$$Y_j = f(t_j, \vec{\theta}_0) (1 + \epsilon_j) \quad (36)$$

where $E(Y_j) = f(t_j, \vec{\theta}_0)$ and $\text{var}(Y_j) = \sigma_0^2 f^2(t_j, \vec{\theta}_0)$. 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} :

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

where Y_j obeys (36) 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

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

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

The GLS estimator has the following asymptotic properties [1]:

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

where

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

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

and $W^{-1}(\vec{\theta}) = \text{diag}(f^2(t_1, \vec{\theta}), \dots, f^2(t_n, \vec{\theta}))$. Note that because $\vec{\theta}_0$ and σ_0^2 are unknown, estimates $\hat{\theta} = \hat{\theta}_{\text{GLS}}$ and $\hat{\sigma}^2 = \hat{\sigma}_{\text{GLS}}^2$ for will be used in (40) to calculate

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

. where [13] we take the approximation

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

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 (28) and (38) 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.

1.4.4 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

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

We note that 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 [1]. 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 found in [10]. In the text the authors claim that 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 ($w_j = g^{-2}(t_j, \vec{\theta})$)

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

In their text, Carroll and Rupert [10] briefly describe how distributions belonging to the exponential family of distributions generate maximum-likelihood estimating equations equivalent to (42).

1.4.5 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 the procedure outlined in Section 1.4.3. The iterative procedure as described in [13] 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 $\vec{\theta} \in \Theta$)

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

and then updating the weights $w_j = f^{-2}(t_j, \vec{\theta})$ 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

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

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

2 Computation of Σ , Standard Errors and Confidence Intervals

We return to the case of n scalar longitudinal observations and consider the OLS case of Section 1.4.1 (the extension of these ideas to vectors is completely straight-forward) where we recall these observations are represented by the statistical model

$$Y_j \equiv f(t_j, \vec{\theta}_0) + \epsilon_j, \quad j = 1, 2, \dots, n, \quad (43)$$

where $f(t_j, \vec{\theta}_0)$ is the model for the observations in terms of the state variables and $\theta_0 \in \mathbb{R}^p$ is a set of theoretical “true” parameter values (assumed to exist in a standard statistical approach). Recall we only assume for our statistical model of the observation or measurement process (43) that the errors ϵ_j , $j = 1, 2, \dots, n$, 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 data $\{y_j\}$ for the observation process $\{Y_j\}$ with the model to seek a value $\hat{\theta}^n$ that minimizes

$$J_n(\theta) = \sum_{j=1}^n [y_j - f(t_j, \vec{\theta})]^2. \quad (44)$$

Since Y_j is a random variable, we have that the estimator $\theta^n = \theta_{\text{OLS}}^n$ (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 $\{y_j\}$ (i.e., a realization of $\{Y_j\}$) when minimizing $J_n(\theta)$.

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 $n \rightarrow \infty$), the standard nonlinear regression approximation theory ([14, 17, 18], and Chapter 12 of [20]) for asymptotic (as $n \rightarrow \infty$) distributions can be invoked. As stated above, this theory yields that the sampling distribution for the estimator $\theta^n(Y)$, where $Y = \{Y_j\}_{j=1}^n$, is approximately a p -multivariate Gaussian with mean $E[\theta^n(Y)] \approx \theta_0$ and covariance matrix $\text{cov}[\theta^n(Y)] \approx \Sigma_0 = \sigma_0^2 [\chi^T(\theta_0) \chi(\theta_0)]^{-1}$. Here $\chi(\vec{\theta}) = F_{\vec{\theta}}(\vec{\theta})$ is the $n \times p$ sensitivity matrix with elements

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

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

$$\theta_{\text{OLS}}^n(Y) \sim \mathcal{N}_p(\theta_0, \sigma_0^2 [\chi^T(\vec{\theta}_0) \chi(\vec{\theta}_0)]^{-1}) := \mathcal{N}_p(\vec{\theta}_0, \Sigma_0). \quad (45)$$

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 $\dot{\vec{x}} = \vec{g}(t, \vec{x}(t), \vec{\theta})$ as in (1), then one can use the $N \times p$ matrix **sensitivity equations** (see [3, 7] and the references therein)

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

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 $\vec{\theta}_0, \sigma_0$ are unknown, we must approximate them in

$$\Sigma_0 = \sigma_0^2 [\chi^T(\vec{\theta}_0) \chi(\vec{\theta}_0)]^{-1}.$$

For this we follow standard practice and use the approximation

$$\Sigma_0 \approx \Sigma(\hat{\theta}^n) = \hat{\sigma}^2 [\chi^T(\hat{\theta}^n) \chi(\hat{\theta}^n)]^{-1}, \quad (46)$$

where $\hat{\theta}^n$ is the parameter estimate obtained, and the approximation $\hat{\sigma}^2$ to σ_0^2 , as discussed earlier, is given by

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

Standard errors to be used in confidence interval calculations are thus given by $SE_k(\hat{\theta}^n) = \sqrt{\Sigma_{kk}(\hat{\theta}^n)}$, $k = 1, 2, \dots, p$ (see [11]).

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

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

where $\alpha \in [0, 1]$ and $t_{1-\alpha/2} \in \mathbb{R}_+$. 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 $T \sim t^{n-p}$.

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 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 [20]) is

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

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 [3]).

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 (3), that is, we have m coordinate observations where $m \leq N$. In this case, we recall that we have

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

and

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

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 1.4.1, 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

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

where $V_0 = \text{diag}(\sigma_{0,1}^2, \dots, \sigma_{0,m}^2)$ and we may follow 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 the *model comparison tests* outlined in the next section) 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 , $j = 1, 2, \dots, n$, are independent identically distributed (*i.i.d.*) random variables with mean $E[\epsilon_j] = 0$ and constant variance $var[\epsilon_j] = \sigma_0^2$. After carrying out the estimation procedures, one can readily plot the *residuals vs. time* and the *residuals vs. the resulting estimated model (output or observation f) values*. A random pattern for the first is strong support for validity of the independence assumption while a non increasing, random pattern for the latter suggests the assumption of constant variance may be reasonable for the data (measurements) used in the inverse problem calculations. The underlying assumption that the sampling size n must be large (recall the theory is asymptotic in that it holds as $n \rightarrow \infty$) is not so readily “verified” and is often ignored (albeit at the user’s peril in regard to the quality of the uncertainty findings). Indeed the asymptotic theories are often used in a very heuristic underlying manner to give a loose feeling for the uncertainty involved in the estimates and the level of parametrization used in approximating the underlying mathematical model. It is often the case that the asymptotic results provide remarkably good approximations to the true sampling distributions for finite n . However, in practice one has no way to ascertain whether this holds for a specific example of interest.

References

- [1] H. T. Banks and M. Davidian, SAMS MA/ST 810 Notes.
- [2] H. T. Banks, S. Dediu and H.K. Nguyen, Sensitivity of dynamical systems to parameters in a convex subset of a topological vector space, Center for Research in Scientific Computation Report, CRSC-TR06-25, September, 2006, North Carolina State University; *Math. Biosci. and Engineering*, **4** (2007), 403–430.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] H. T. Banks and K. Kunsich, *Estimation Techniques for Distributed Parameter Systems*, Birkhauser, Boston, 1989.

- [7] H. T. Banks and H. K. Nguyen, Sensitivity of dynamical system to Banach space parameters, CRSC Tech Rep. CRSC-TR05-13, N.C. State University, February, 2005; *J. Math. Analysis and Applications*, **323** (2006), 146–161.
- [8] 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.
- [9] J. J. Batzel, F. Kappel, D. Schneditz and H.T. Tran, *Cardiovascular and Respiratory Systems: Modeling, Analysis and Control*, SIAM Frontiers in Applied Math, SIAM, Philadelphia, 2006.
- [10] R.J. Carroll and D. Ruppert. Transformation and Weighting in Regression. Chapman and Hall, New York, 1988.
- [11] G. Casella and R. L. Berger, *Statistical Inference*, Duxbury, California, 2002.
- [12] J. B. Cruz, ed., *System Sensitivity Analysis*, Dowden, Hutchinson & Ross, Inc., Stroudsburg, PA, 1973.
- [13] M. Davidian, Class Notes ST762, NCSU.
- [14] M. Davidian and D. Giltinan, *Nonlinear Models for Repeated Measurement Data*, Chapman & Hall, London, 1998.
- [15] M. Eslami, *Theory of Sensitivity in Dynamic Systems: An Introduction*, Springer-Verlag, Berlin, 1994.
- [16] P.M. Frank, *Introduction to System Sensitivity Theory*, Academic Press, Inc., New York, NY, 1978.
- [17] A. R. Gallant, *Nonlinear Statistical Models*, John Wiley & Sons, Inc., New York, 1987.
- [18] R. I. Jennrich, Asymptotic properties of non-linear least squares estimators., *Ann. Math. Statist.*, **40** (1969), 633–643.
- [19] A. Saltelli, K. Chan and E.M. Scott, eds., *Sensitivity Analysis*, Wiley Series in Probability and Statistics, John Wiley & Sons, New York, NY, 2000.
- [20] G. A. F. Seber and C. J. Wild, *Nonlinear Regression*, John Wiley & Sons, Inc., New York, 1989.
- [21] K. Thomaseth and C. Cobelli. Generalized sensitivity functions in physiological system identification., *Ann Biomed Eng.*, **27(5)** (1999), 607–616.
- [22] D. D. Wackerly, W. Mendenhall III, and R. L. Scheaffer, *Mathematical Statistics with Applications*, Duxbury Thompson Learning, USA, 2002.