

The Role of Statistical Principles in Quantitative Biomedical Modeling

*Atlantic Coast Conference on Mathematics in the
Life and Biological Sciences*

Marie Davidian
Department of Statistics
North Carolina State University



<http://www.stat.ncsu.edu/~davidian>

Outline

1. Modeling in biomedical research
2. Sources of variation in data
3. Statistical models
4. Hierarchical statistical models for longitudinal biomedical data
5. Fitting statistical models and statistical inference
6. Using models in biomedical research
7. Closing remarks

1. Modeling in biomedical research

Increasingly: Recognition of the role of *mathematical models of biological systems* in biomedical research

- Mechanisms underlying disease and its progression
- Processes involved in the disposition of drugs
- Processes governing the action of treatments
- In general, *physiological mechanisms taking place within a subject over time*

Usefulness:

- Evaluation of *effects* of new treatments
- Identification of *appropriate doses* of drugs for further study
- Development of strategies for *administration of treatments over time*
- Design of *clinical studies*

1. Modeling in biomedical research

For example:

<http://www.fda.gov/oc/initiatives/criticalpath/>



1. Modeling in biomedical research

Most popular models: (Deterministic) systems of *ODEs* with s states

$$\dot{x}(t) = g\{t, x(t), \theta\}, \quad \text{initial conditions } x(0) = x_0, \quad \text{solution } x(t, \theta) \quad (s \times 1)$$

- *Forward simulation* of $x(t)$ given θ, x_0
- *Inverse problem* to determine θ given $x(t)$

Information available to facilitate modeling:

- *Established knowledge* and *hypotheses* about physiology at various levels (e.g., molecular, cellular, organs/tissues, system, ...)
- *Longitudinal data* on (*parts of*) $x(t, \theta)$ arising from *human subjects*

$$y_1, \dots, y_n \quad \text{at times} \quad 0 \leq t_1 < \dots < t_n$$

observed for *each subject*

1. Modeling in biomedical research

Usual features of data:

- y_1, \dots, y_n for any subject do not “*track*” exactly along $x(t, \theta)$
- Different *observed trajectories* for different subjects
- *Variation*

Statistics: “*The study of variation*”

- *Not* a branch of mathematics; more a *philosophy* for characterizing how *data arise* . . .
- . . . and for *drawing inferences* from data in the face of *variation*

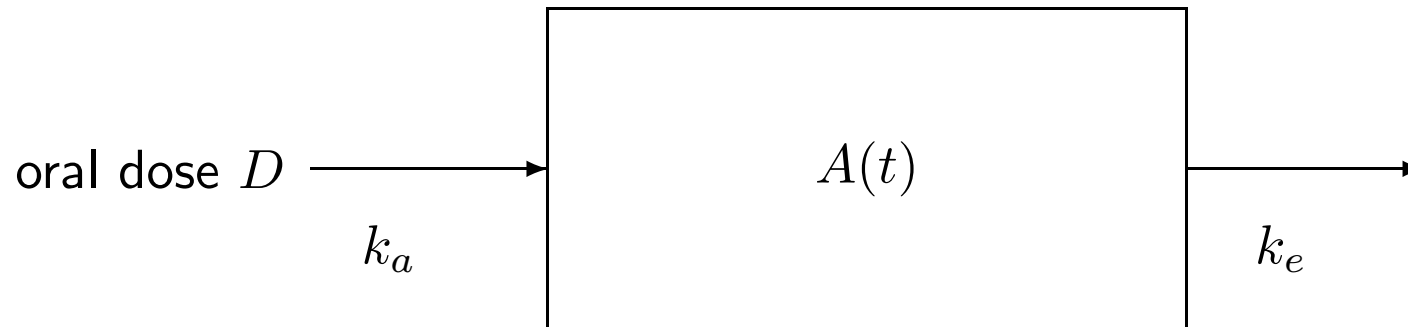
This talk: Describe how *statistical* principles *must* form the basis for

- Application of such models to *data*
- The *broader goals* on Slide 3

2. Sources of variation in data

Simple example: *Pharmacokinetics (PK) of theophylline*

- Understanding of PK (*absorption, distribution, elimination*) needed for *dose recommendations*
- *One compartment* model ($s = 2$)



- *Assume* constant relationship between concentration $C(t)$ and amount $A(t) \Rightarrow C(t) = A(t)/V$

$$C(t) = \frac{k_a F D}{V(k_a - k_e)} \{ \exp(-k_e t) - \exp(-k_a t) \}$$

2. Sources of variation in data

Thus: *Closed form solution* for observable state

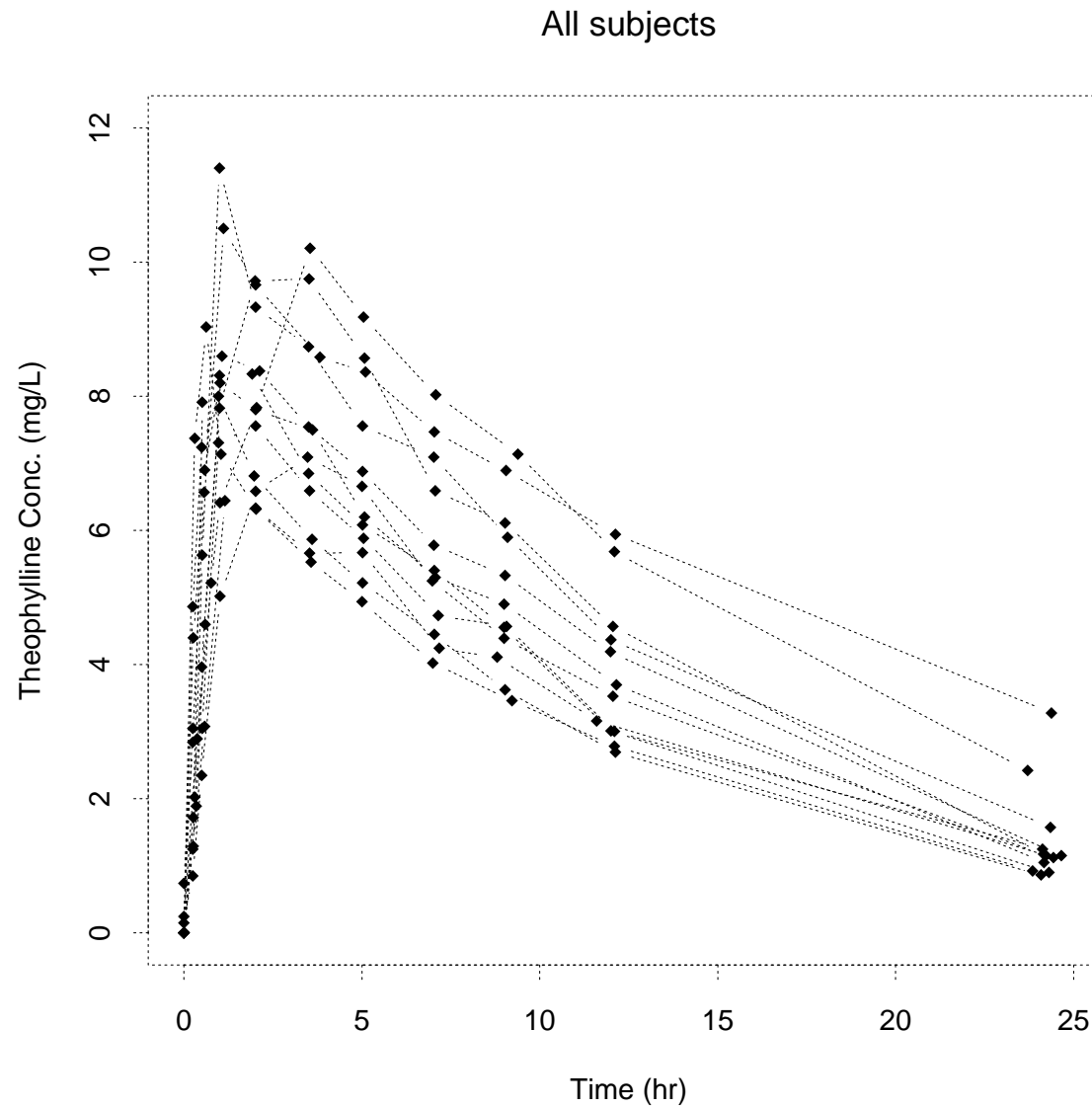
$$x_1(t, \theta) = \frac{k_a F D}{V(k_a - k_e)} \{ \exp(-k_e t) - \exp(-k_a t) \}$$

$$\theta = (k_a, V, k_e)^T$$

Typical intensive PK study:

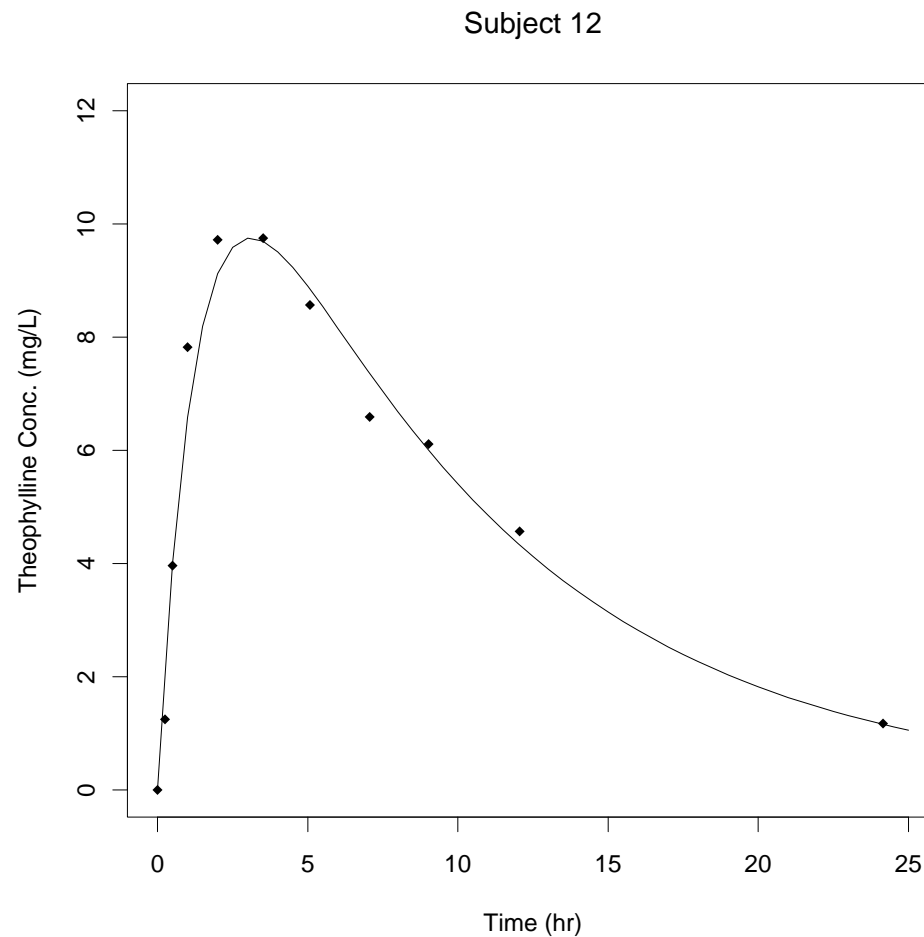
- N subjects given *same oral dose* D (mg/kg) at $t = 0$
- Blood samples at subsequent times assayed for concentration $\Rightarrow y_1, \dots, y_n$ at t_1, \dots, t_n for each subject
- *Main objective*: Learn extent to which absorption, distribution, elimination, i.e., $\theta = (k_a, V, k_e)^T$ *vary from subject to subject*...
- ...and use this knowledge to develop *dosing recommendations* for the *population*

2. Sources of variation in data



2. Sources of variation in data

Single subject: With “*fitted model*”



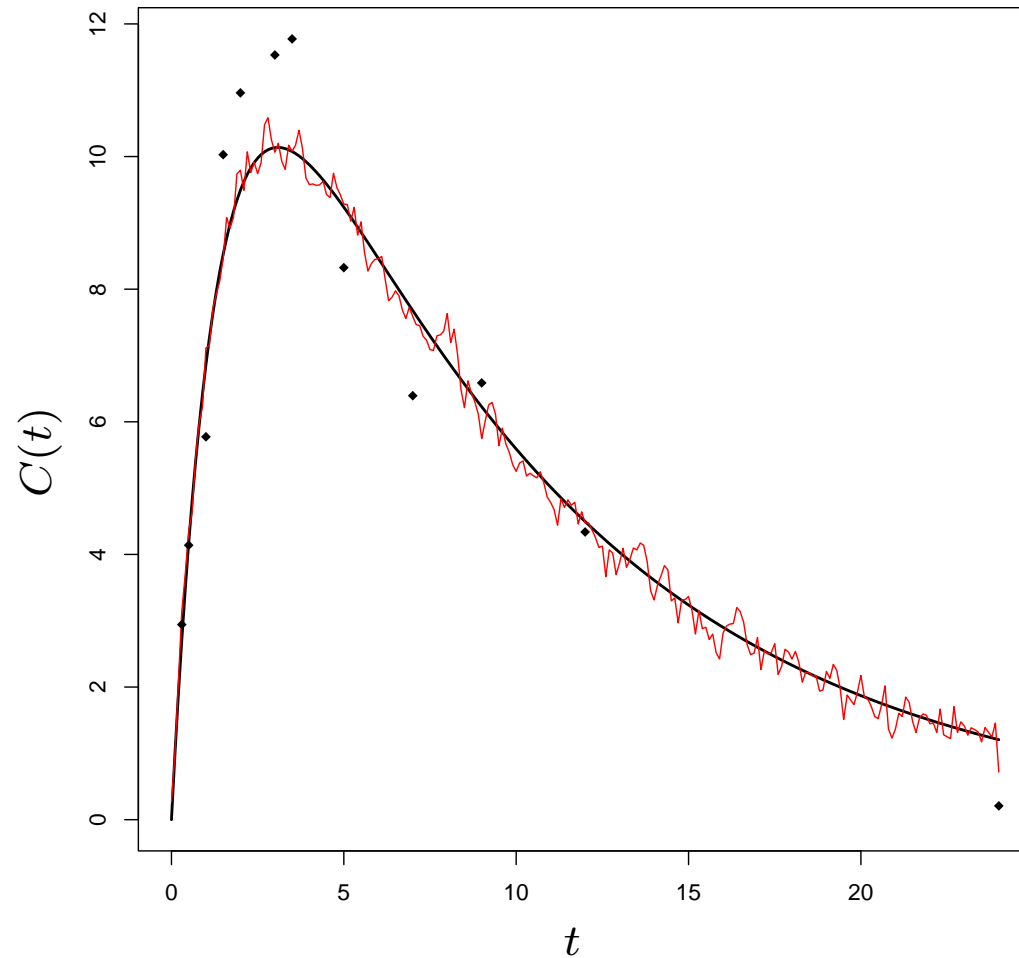
2. Sources of variation in data

Why don't the observed concentrations lie exactly on the fitted trajectory?

- “*Observation error*”
- One obvious reason – *Assay (measurement) error*
- Model *misspecification* – true PK *more complex*
- Times/dose recorded *incorrectly*
- Etc. . .

2. Sources of variation in data

Hypothetically:



2. Sources of variation in data

Sources of variation: The (*deterministic*) model is a good representation of the *long-term* trajectory, but *observed concentrations* are subject to

- Intra-subject “*fluctuations*” about it
- *Measurement error*

E.g., measurement error: A particular sample has a “*true*” concentration

- *Ideally* all determinations of this “*true*” concentration should be the same, but measuring devices commit *errors*
- Measure *over and over* – a *different error* committed each time
- \Rightarrow *All possible* concentrations we might observe *vary* due the effect of measurement error

2. Sources of variation in data

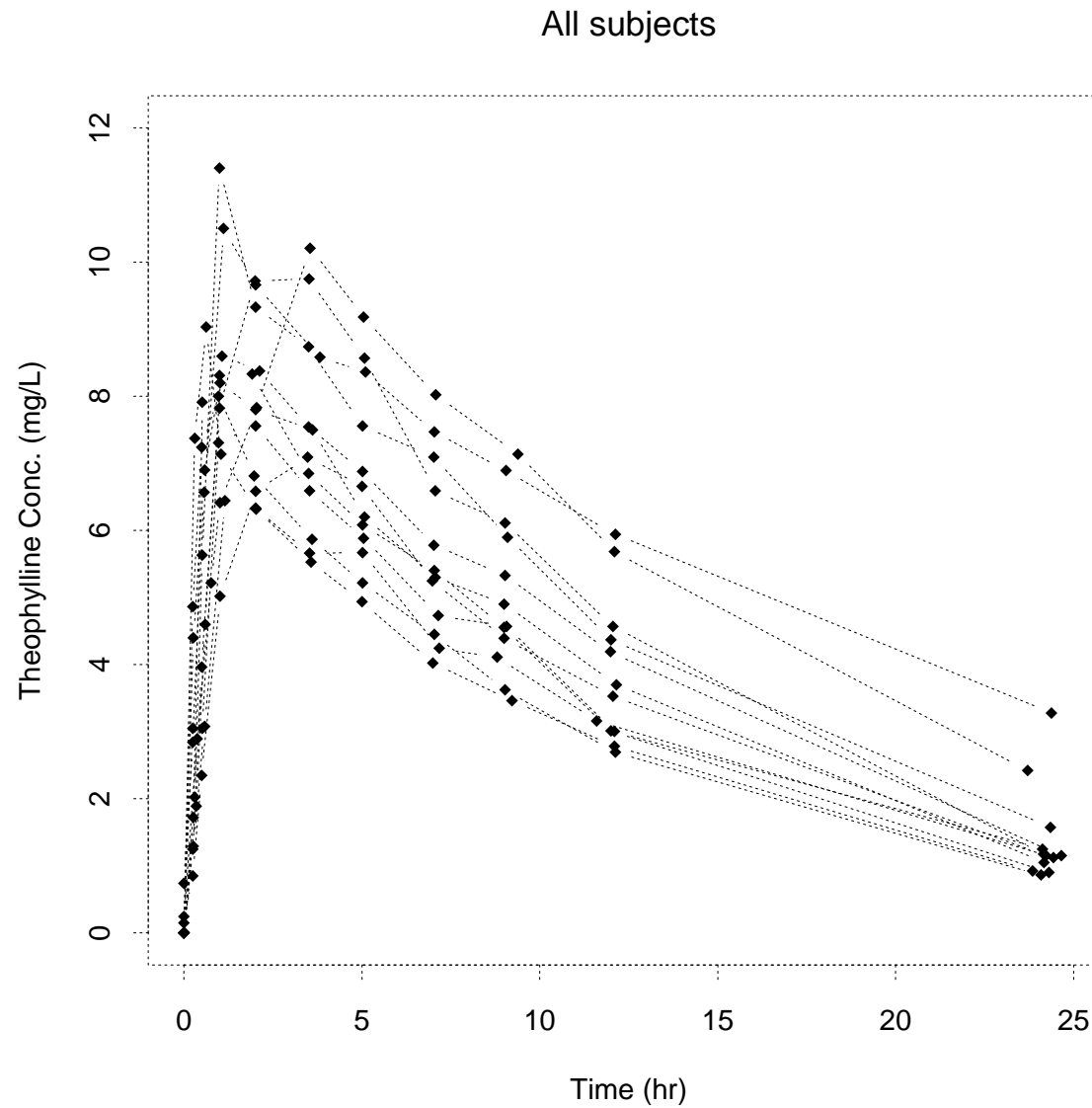
Result: If we wish to learn about θ for a *particular subject* based on *data*, we must take *appropriate account* of these sources of *within-subject variation*

Variation begets: *Uncertainty* in what we observe

- Because observations *could have turned out differently* due to *variation* in their possible values . . .
- . . . Any determination of θ from *data* is subject to *uncertainty*
- Uncertainty must be *characterized and quantified* to gauge how much *faith* to place in results

More variation: Variation due to *within-subject* phenomena is not the only issue. . .

2. Sources of variation in data



2. Sources of variation in data

Interpretation: Although the observed pattern is *consistent with* the model for all subjects. . .

- . . . there is *among-subject* variation
- \Rightarrow Each subject has his/her *own* θ dictating his/her PK

Result:

- *Objective restated* – characterize how θ values are *distributed* (and hence *vary*) in the *population of subjects* like these
- And use this characterization to develop *dosing strategies*
- In doing this based on *data*, we must take *appropriate account* of *both within-* and *among-subject* variation
- Have seen only a *sample* from this population \Rightarrow any attempt to do this is subject to *uncertainty*

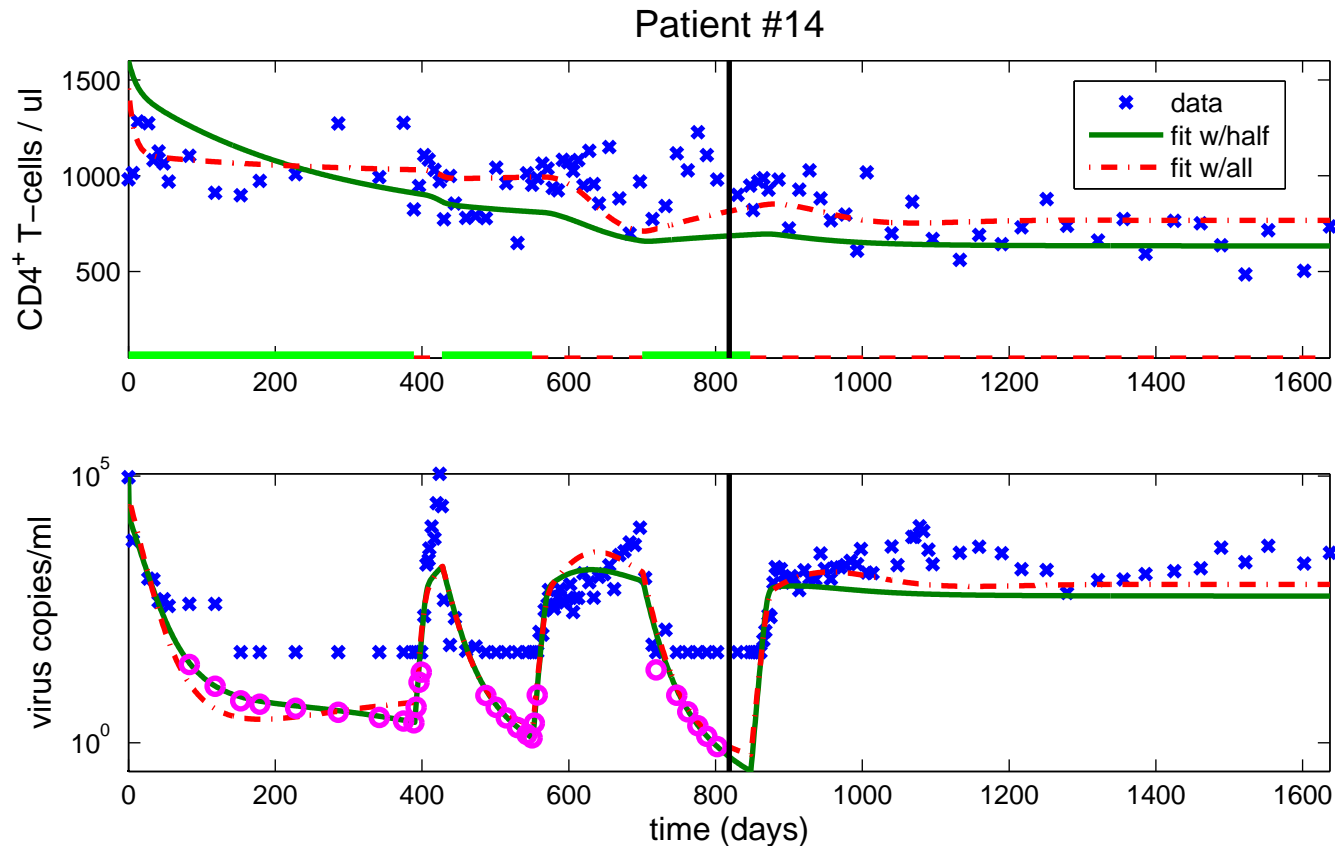
2. Sources of variation in data

To address the objectives: Need a *framework* in which to

- State and address the objectives in a *formal* way
- This will require characterizing *variation* (*within* and *among* subjects) ...
- ... and will provide a basis for *assessing the uncertainty* involved in using the *data*

2. Sources of variation in data

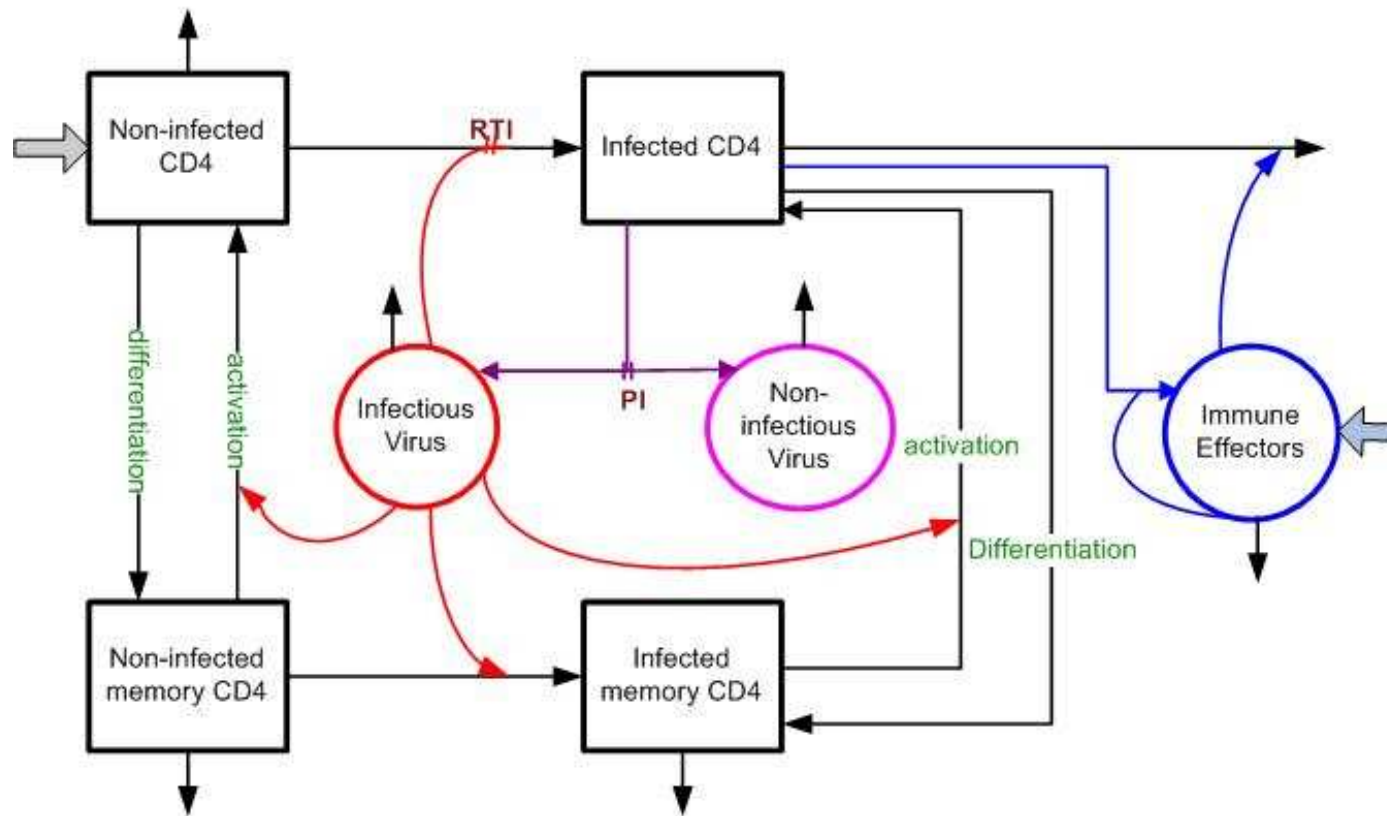
A fancier example: *HIV dynamics* under *antiretroviral therapy*



- *CD4⁺ T-cell count* – “*immunologic status*”
- *Viral load* – “*virologic status*”

2. Sources of variation in data

Model for within-subject dynamics:



2. Sources of variation in data

Model for within-subject dynamics: $s = 7$

$$\begin{aligned}\dot{T}_1 &= \lambda_1 - d_1 T_1 - \{1 - \epsilon_1 u(t)\} k_1 V_I T_1 \\ \dot{T}_2 &= \lambda_2 - d_2 T_2 - \{1 - f \epsilon_1 u(t)\} k_2 V_I T_2 \\ \dot{T}_1^* &= \{1 - \epsilon_1 u(t)\} k_1 V_I T_1 - \delta T_1^* - m_2 E T_1^* \\ \dot{T}_2^* &= \{1 - f \epsilon_1 u(t)\} k_2 V_I T_2 - \delta T_2^* - m_2 E T_2^* \\ \dot{V}_I &= \{1 - \epsilon_2 u(t)\} 10^3 N_T \delta (T_1^* + T_2^*) - c V_I - \{1 - \epsilon_1 u(t)\} \rho_1 10^3 k_1 T_1 V_I \\ &\quad - \{1 - f \epsilon_1 u(t)\} \rho_2 10^3 k_2 T_2 V_I \\ \dot{V}_{NI} &= \epsilon_2 u(t) 10^3 N_T \delta (T_1^* + T_2^*) - c V_{NI} \\ \dot{E} &= \lambda_E + \frac{b_E (T_1^* + T_2^*)}{(T_1^* + T_2^*) + K_b} E - \frac{d_E (T_1^* + T_2^*)}{(T_1^* + T_2^*) + K_d} E - \delta_E E\end{aligned}$$

- Plus initial conditions, $\theta = (\lambda_1, d_1, \epsilon_1, k_1, \dots)^T$
- Observable: *CD4 count* $= T_1 + T_1^*$, *viral load* $= V_I + V_{NI}$
- $u(t)$ = treatment input at t

2. Sources of variation in data

Objectives:

- What would CD4 and viral load progression look like *in the population of HIV-infected subjects* if *all* subjects followed a particular treatment pattern $u(t)$?
- Can we design “*good*,” “*realistic*” $u(t)$ that work well for the *population* (at least *on average*)?

Observations: y_1, \dots, y_n at t_1, \dots, t_n for each subject, y_j is (2×1) vector of observed CD4 and viral load

- *Within-subject variation*: *Fluctuations* from *smooth trajectories* for $T_1(t) + T_1^*(t)$ and $V_I(t) + V_{NI}(t)$ and *assay error*
- *Among-subject variation*: Each subject has his/her *own subject-specific dynamic parameters* θ
- *Distribution of* θ in population dictates *progression in the population*

2. Sources of variation in data

Objectives: Learn about *distribution of* θ in population

- Dictates *distribution of* CD4 and viral load progression in the population under different $u(t)$

Further complication: The viral load assay has a *lower limit of quantification* L

- Viral load measurements known only to lie below $L \Rightarrow$ *censored data*
- *Disregarding* censored data or setting equal to L , $L/2$, etc, *compromises* determination (*estimation*) of θ

2. Sources of variation in data

Again: Need a *formal framework* to characterize the *sources of variation* and the *uncertainty* involved

- Need to *embed* the mathematical model in a *statistical model*
- *Formal statement* of objectives
- Will dictate how to “*fit*” the mathematical model to the data (*and* how to *assess uncertainty* in doing this) in pursuit of the objectives
- Also provides a basis for *correct* handling of *censored observations*

3. Statistical models

Perspective: *Data* that are observed are envisioned as *realizations* of *random variables* (*vectors*), e.g.,

- *PK study*: For subject i , $i = 1, \dots, N$, observed at n_i times t_{i1}, \dots, t_{in_i} , Y_{ij} = drug concentration at time $t_{ij} \Rightarrow$

$$Y_i = (Y_{i1}, \dots, Y_{in_i})^T$$

- *HIV study*: For subject i , $i = 1, \dots, N$, observed at n_i times t_{i1}, \dots, t_{in_i}

$$Y_{ij} = (Y_{ij}^{CD4}, Y_{ij}^{VL})^T \text{ at time } t_{ij} \Rightarrow Y_i = (Y_{i1}, \dots, Y_{in_i})^T$$

- Actual *data* on i are realizations y_{i1}, \dots, y_{in_i} summarized by y_i
- For *all subjects* $i = 1, \dots, N$, we have Y_1, \dots, Y_N

3. Statistical models

Additional data collected:

- *Conditions* under which i is observed, e.g., single dose, on/off treatment pattern over time $\Rightarrow U_i$
- *Subject characteristics*, e.g., demographic, physiologic information (age, weight, renal function, prior treatment history, IV drug use, etc) $\Rightarrow A_i$

Full set of observed data: *Realizations* of random vector triplets

$$Z_i = (Y_i, U_i, A_i), \quad i = 1, \dots, N$$

- Z_1, \dots, Z_N , ordinarily assumed to be *independent random vectors* (from *unrelated* subjects)

3. Statistical models

Statistical model: Aka *probability model*

- The class of *probability distributions* for the Z_i , $i = 1, \dots, N$, that we believe could have generated the *data* (i.e., *realizations* of Z_1, \dots, Z_N) we saw
- I.e., such a *probability distribution* describes the way in which Z_1, \dots, Z_N corresponding to N subjects drawn from a *population of interest* would take on their values
- Should embody *realistic assumptions* about the *sources of variation*

For simplicity: Assume for now *no* A_i collected and U_i are *fixed* and focus on a *probability model for* the Y_i

3. Statistical models

For longitudinal data: It is standard to *build up* a statistical model in a *hierarchy of stages*

1. The probability mechanism by which Y_i would take on its values for a *single subject* $i \Rightarrow$ involves *within-subject* variation
2. The probability mechanism that describes *variation among subjects* in how Y_1, \dots, Y_N take on their values

Mathematical model: s states

$$\dot{x}(t) = g\{t, x(t), \theta\}, \text{ solution } x(t, \theta) \text{ (} s \times 1 \text{)}$$

- Observations *not available* on all s states
- $\bar{x} = \mathcal{O}x$ for *observation operator* \mathcal{O}

4. Hierarchical statistical models

First stage: Consider a *single subject* i

- θ_i denotes the parameters dictating i 's PK/HIV dynamics/etc

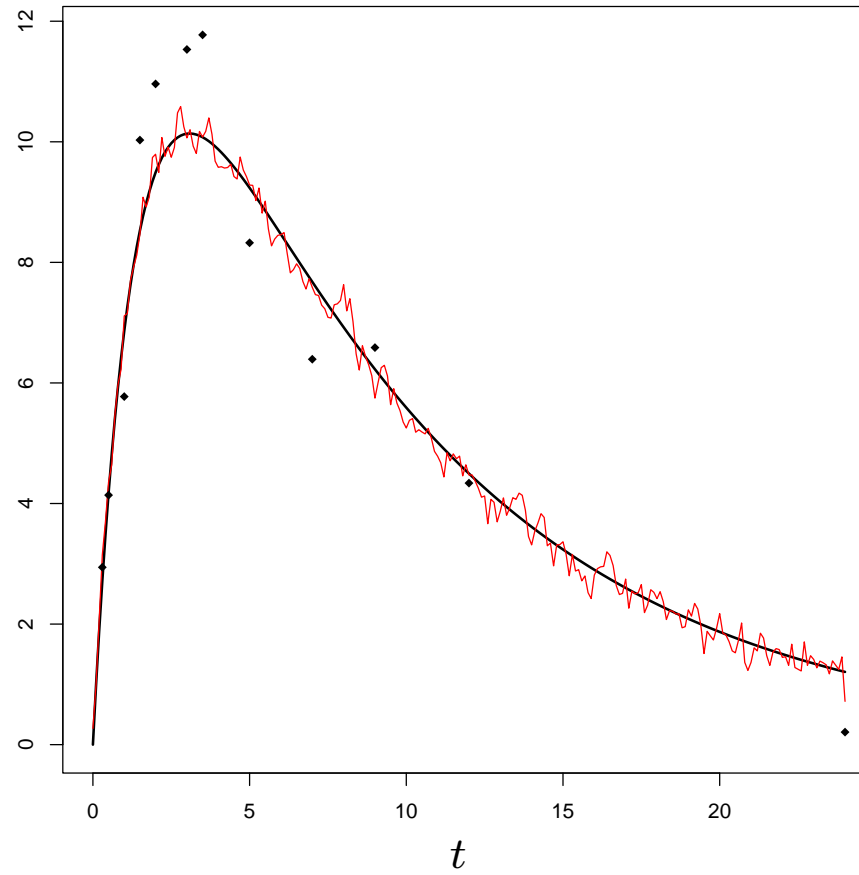
Approach: Embed \bar{x} in a *subject-specific stochastic process*

$$Y_i(t, U_i) = \bar{x}(t, U_i, \theta_i) + e_i(t, U_i)$$

- $e_i(t, U_i)$ is the *deviation process* describing how *realizations* of $Y_i(t, U_i)$ would *deviate* from the deterministic trajectory $\bar{x}(t, U_i, \theta_i)$ due to *fluctuations* and *measurement error*
- $E\{e_i(t, U_i)|U_i, \theta_i\} = 0 \Rightarrow$ over *all possible realizations*, the deviations *average out to zero*
- $\Rightarrow E\{Y_i(t, U_i)|U_i, \theta_i\} = \bar{x}(t, U_i, \theta_i)$, i.e., interpret $\bar{x}(t, U_i, \theta_i)$ as the *average* trajectory over *all possible realizations* we could see on subject i under conditions U_i

4. Hierarchical statistical models

Conceptually:



— $\bar{x}(t, U_i)$

4. Hierarchical statistical models

$$Y_i(t, U_i) = \bar{x}(t, U_i, \theta_i) + e_i(t, U_i)$$

- $\bar{x}(t, U_i, \theta_i)$ is “*inherent tendency*” for i 's system to evolve over time
- Depends on θ_i , viewed as an “*inherent characteristic*” of i dictating this tendency
- Can think of $e_i(t, U_i) = e_{i,F}(t, U_i) + e_{i,M}(t, U_i)$
- “*Fluctuations*” and “*measurement error*”

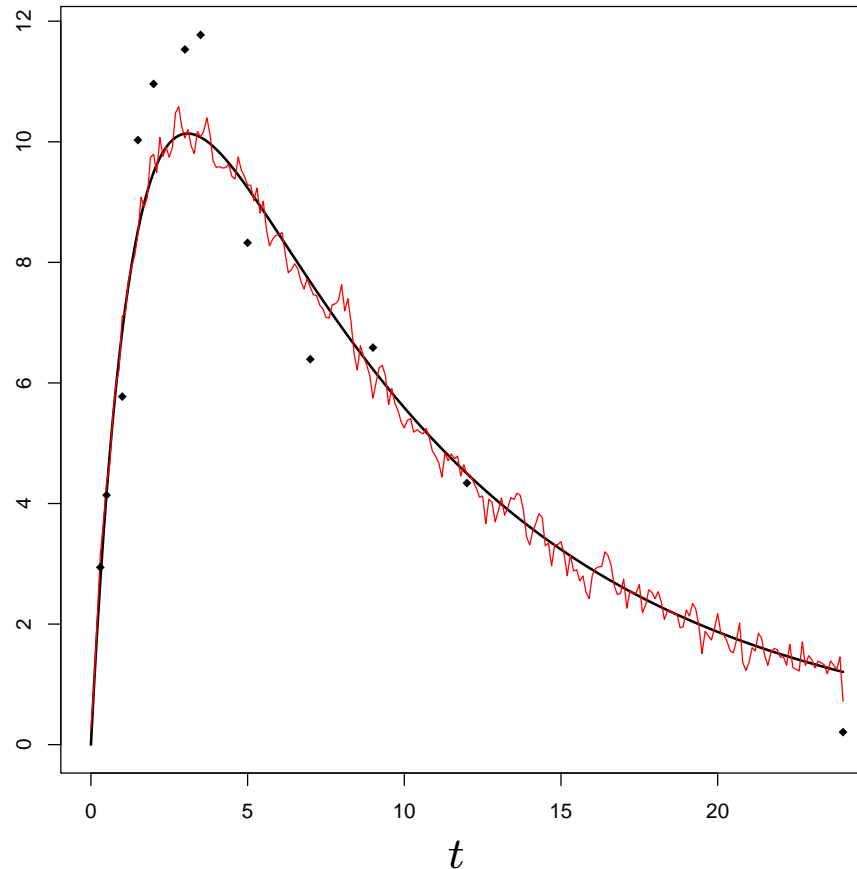
Intermittent observations on $Y_i(t, U_i)$: At times t_{i1}, \dots, t_{in_i}

- $Y_{ij} = Y_i(t_{ij}, U_i), e_{ij} = e_i(t_{ij}, U_i)$

$$Y_{ij} = \bar{x}(t_{ij}, \theta_i) + e_{ij}, \quad j = 1, \dots, n_i$$

4. Hierarchical statistical models

Conceptually:



— $\bar{x}(t, U_i)$, — $\bar{x}(t, U_i) + e_{i,F}(t, U_i)$, • $Y_{ij} = Y(t_{ij}, U_i)$

4. Hierarchical statistical models

To complete the first stage model: Must make *assumptions* on

$$e_i(t, U_i) = e_{i,F}(t, U_i) + e_{i,M}(t, U_i) \quad \text{and hence on} \quad e_{ij} = e_{ij,F} + e_{ij,M}$$

Measurement error: Typical assumptions on $e_{i,M}(t, U_i)$

- *Uncorrelated* or *independent* across time \Rightarrow devices commit *haphazard errors*
- At particular time t , elements of $e_{i,M}(t, U_i)$ *may or may not* be correlated \Rightarrow *separate assays* for CD4, viral load but common *sample preparation*
- *Normal probability distribution* for each element
- *Variances* of each component are *different*, *may or may not* be *constant*

4. Hierarchical statistical models

To complete the first stage model: Must make *assumptions* on

$$e_i(t, U_i) = e_{i,F}(t, U_i) + e_{i,M}(t, U_i) \quad \text{and hence on} \quad e_{ij} = e_{ij,F} + e_{ij,M}$$

Fluctuation process: Typical assumptions on $e_{i,F}(t, U_i)$

- Fluctuations “*close together*” in time tend to be “*similar*” \Rightarrow $e_{i,F}(t, U_i)$ and $e_{i,F}(s, U_i)$ (*auto*)*correlated* depending on $|t - s|$
- If times t_{i1}, \dots, t_{in_i} *sufficiently intermittent*, treat autocorrelation as *negligible*
- Elements of $e_{i,F}(t, U_i)$ at a particular time t *may or may not* be *correlated*
- *Normal probability distribution* for each element
- *Variances* of each element *different*, *may or may not* be *constant*

4. Hierarchical statistical models

Transformation: Assumptions like *normal distribution* may make more sense on a *transformed scale*, e.g.

$$\log\{Y_i(t, U_i)\} = \log\{\bar{x}(t, \theta_i)\} + e_i(t, U_i)$$

Result: *Probability model* for Y_i , e.g., for HIV dynamic example

- At each time t_{ij} , $j = 1, \dots, n_i$

$$\log(Y_{ij}) | U_i, \theta_i \sim \mathcal{N}\left(\log\{\bar{x}(t, \theta_i)\}, \Sigma\right), \quad \Sigma = \begin{pmatrix} \sigma_{CD4}^2 & 0 \\ 0 & \sigma_{VL}^2 \end{pmatrix}.$$

- Y_{ij} *uncorrelated* (*independent*) across $j = 1, \dots, n_i$
- *Whatever* assumptions are made determine a *probability density* for Y_i (*given* U_i, θ_i)

$$p(y_i | u_i, \theta_i; \Sigma)$$

4. Hierarchical statistical models

Interest in subject i : If we wish to learn *only* about subject i 's PK or dynamics

- Want to *estimate* θ_i based on observing a realization of Y_i
- How this is accomplished *should be based on* $p(y_i | u_i, \theta_i; \Sigma)$
- *More later...*

4. Hierarchical statistical models

Second stage: Consider the *population of subjects*

- θ_i dictates PK/HIV dynamics/etc *specific to subject i*
- *Different* subjects in the *population* have *different* θ values leading to *variation in “inherent trajectories” among subjects*
- *Conceptualize* the population as *all possible* θ values \Rightarrow a *probability distribution*
- $\theta_1, \dots, \theta_N$ for a *sample* of N subjects are *independent* random vectors taking their values according to this distribution
- E.g., $\theta_i \sim \mathcal{N}(\theta_*, D)$ or $\log(\theta_i) \sim \mathcal{N}(\theta_*, D)$, $i = 1, \dots, N$
- θ_* is the *average value* of θ in the population
- D is a *covariance matrix* whose diagonal elements are the *variances* of elements of θ across the population

4. Hierarchical statistical models

Result: Assumed *probability model* for θ_i , $i = 1, \dots, N$, determines a *probability density*

$$p(\theta_1, \dots, \theta_N; \theta_*, D) = \prod_{i=1}^N p(\theta_i; \theta_*, D)$$

Objectives: *Distribution* of possible θ values in population

- *Estimate* θ_* and D (*more in a moment*)

Refinement: If A_i , $i = 1, \dots, N$, also available, can develop a *probability model* for θ_i and A_i *jointly*

- Values of A_i determine “*subpopulations*”
- Allows *different probability distributions* for θ depending on values of A_i (different subpopulations)

4. Hierarchical statistical models

Putting together: Full *statistical model* for Y_1, \dots, Y_N

- *Probability density* for Y_1, \dots, Y_N
- Y_i and θ_i assumed *independent* across i

$$\begin{aligned} p(y_1, \dots, y_N; \psi) &= \int p(y_1, \dots, y_N, \theta_1, \dots, \theta_N) d\theta_1 \cdots d\theta_N \\ &= \int \prod_{i=1}^N p(y_i, \theta_i) d\theta_i \quad \text{by } \textit{independence} \\ &= \prod_{i=1}^N \int p(y_i | \theta_i) p(\theta_i) d\theta_i \\ &= \prod_{i=1}^N \int p(y_i | u_i, \theta_i; \Sigma) p(\theta_i; \theta_*, D) d\theta_i \end{aligned}$$

- *Depends on* $\psi = (\Sigma, \theta_*, D)$

5. Fitting and inference

$$p(y_1, \dots, y_N; \psi) = \prod_{i=1}^N \int p(y_i | u_i, \theta_i; \Sigma) p(\theta_i; \theta_*, D) d\theta_i, \quad \psi = (\Sigma, \theta_*, D)$$

How is all of this used? The *statistical model* provides a formal framework for “*fitting*” the mathematical model to data

From statistical point of view: Under our assumptions

- *Probability distributions* that could have led to the *realization* y_1, \dots, y_N (*data*) we saw are specified by different values of ψ
- *Which value of ψ truly governs the data generating mechanism?*
- \Rightarrow How do we *estimate* ψ from a potential set of data, i.e., Y_1, \dots, Y_N ?
- Given we can't see the whole population, how *uncertain* will we be?

5. Fitting and inference

Estimator: A *function* of Y_1, \dots, Y_N that, if evaluated at a *particular realization* y_1, \dots, y_N yields a numerical value (the *estimate*) that gives information on the *true value of ψ*

- Write $\hat{\psi} = \hat{\psi}(Y_1, \dots, Y_N)$

Sampling distribution of an estimator: An *estimator* has a *probability distribution* characterizing how it takes on its possible values (depending on Y_1, \dots, Y_N and hence ψ)

- *All possible realizations* $y_1, \dots, y_N \Rightarrow$ *all possible values* $\hat{\psi}$ can take on (we observe *only one realization*)
- These values *vary* across realizations
- *Large variance* – another realization might give a *very different* estimate \Rightarrow *lots of uncertainty*
- *Small variance* – similar estimate from another realization \Rightarrow *mild uncertainty*

5. Fitting and inference

Result: If we *estimate* ψ based on *data*, we must also report some estimate of the *variance of the sampling distribution* of $\hat{\psi}$

- Sampling distribution often *approximated* based on *large sample theory*
- *Standard error* – an estimate of $\sqrt{\text{variance}}$ based on this
- *Quantifies uncertainty*

How do we get estimators and approximations to their sampling distributions?

- *Least squares* often *DOES NOT* yield an appropriate *estimator*...
- Even if we are only interested in *part of ψ* (e.g., θ_* and D), we often must estimate *all of ψ*

5. Fitting and inference

Standard approach: *Maximum likelihood estimation*

- *Likelihood function* $L(\psi|y_1, \dots, y_N) = p(y_1, \dots, y_N; \psi)$
- For each possible realization y_1, \dots, y_N there is a corresponding value $\hat{\psi}(y_1, \dots, y_n)$ maximizing $L(\psi|y_1, \dots, y_N)$ – define the *maximum likelihood estimator (MLE)* to be the corresponding function $\hat{\psi}(Y_1, \dots, Y_N)$
- MLE is the “*value of ψ most likely to have led to the observed data*”
- *Optimality properties* – “*most precise*” (\approx smallest sampling variance in “large samples”)
- General *large sample theory* yields approximate *sampling distribution* \Rightarrow *standard errors*
- *In general* – for *complex statistical models*, $L(\psi|y_1, \dots, y_N)$ is a *complex* function of ψ

5. Fitting and inference

Individual inference: Interested *only* in subject i 's HIV dynamics

- MLE for θ_i based on the *statistical model* $p(y_i | u_i, \theta_i; \Sigma)$ such that at each time t_{ij} , $j = 1, \dots, n_i$

$$Y_{ij} | U_i, \theta_i \sim \mathcal{N}\left(\bar{x}(t, \theta_i), \Sigma\right), \quad \Sigma = \begin{pmatrix} \sigma_{CD4}^2 & 0 \\ 0 & \sigma_{VL}^2 \end{pmatrix}.$$

$$\hat{\theta}_i = \arg \min_{\theta} \left\{ \hat{\sigma}_{CD4}^{-2} \sum_{i=1}^N |y_{ij}^{CD4} - \bar{x}_1(t_{ij}, \theta)|^2 + \hat{\sigma}_{VL}^{-2} \sum_{i=1}^N |y_{ij}^{VL} - \bar{x}_2(t_{ij}, \theta)|^2 \right\}$$

$$\hat{\sigma}_{CD4}^2 = N^{-1} \sum_{i=1}^N |y_{ij}^{CD4} - \bar{x}_1(t_{ij}, \hat{\theta}_i)|^2, \quad \hat{\sigma}_{VL}^2 = N^{-1} \sum_{i=1}^N |y_{ij}^{VL} - \bar{x}_2(t_{ij}, \hat{\theta}_i)|^2$$

- If *censoring* is much more complicated*...
- Σ must be estimated, too

*See, for example, Adams et al. (2007)

5. Fitting and inference

Inference on the population: MLE for $\psi = (\Sigma, \theta_*, D)$ maximizes

$$L(\psi|y_1, \dots, y_N) = \prod_{i=1}^N \int p(y_i|u_i, \theta_i; \Sigma) p(\theta_i; \theta_*, D) d\theta_i$$

- *Computational complications* – complex likelihood surface, intractable integration, etc.
- *Focus* is on θ_* and D , but must estimate Σ , too

Alternative: *Bayesian* inference

- Analogous *modeling*, *computational issues*

6. Using models

Recall:

- Evaluation of *effects* of new treatments
- Identification of *appropriate doses* of drugs for further study
- Development of strategies for *administration of treatments over time*
- Design of *clinical studies*

Great current interest: Use *mathematical models* embedded in an appropriate *statistical model* as the basis for *simulation* to address question like these

- E.g., “*Modeling and Simulation*” initiatives at most large pharmaceutical companies and FDA
- *Basis for simulation* – the *statistical framework*

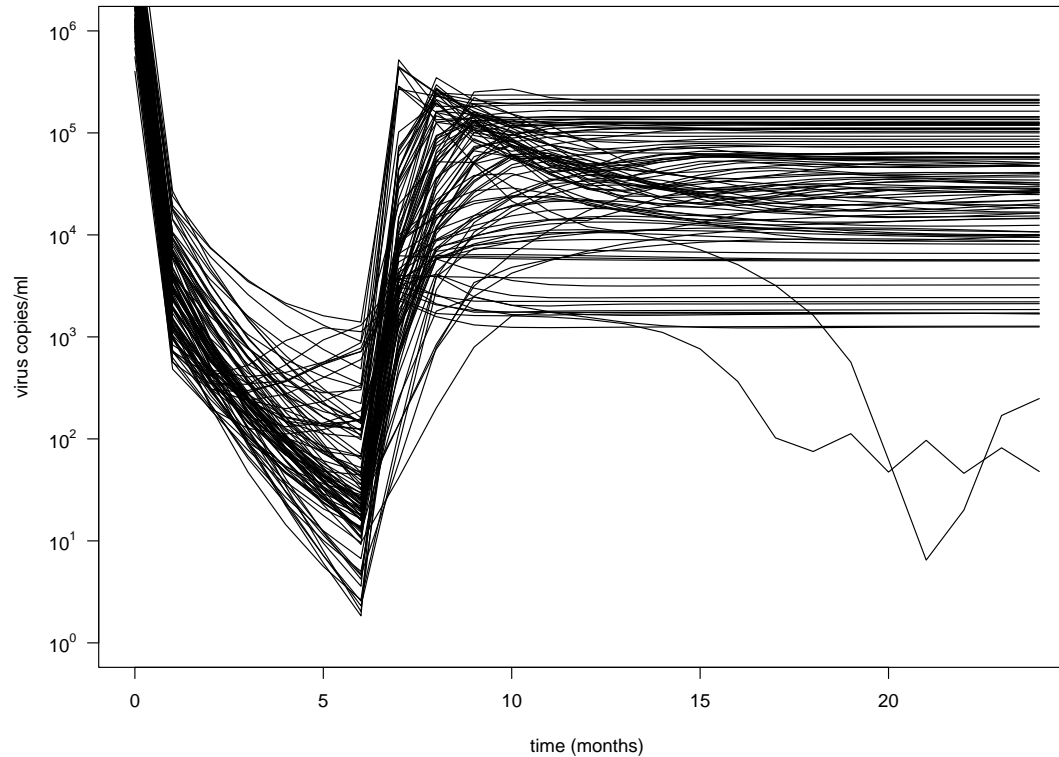
6. Using models

Approach: *Simulation* of the *population* based on a *fitted hierarchical statistical model*

- Generate N_{sim} “*virtual subjects*” by generating θ_i^* , $i = 1, \dots, N_{sim}$, from $p(\theta_i; \hat{\theta}_*, \hat{D})$
- Generate “*inherent trajectories*” $x(t, \theta_i^*)$ under different conditions U_i , e.g. input $u(t)$
- Can add within-subject *deviations* to obtain “*virtual data*”
- Can run “*virtual clinical studies*” with different *sample sizes*, *subject populations*, *doses*, etc, to predict outcome

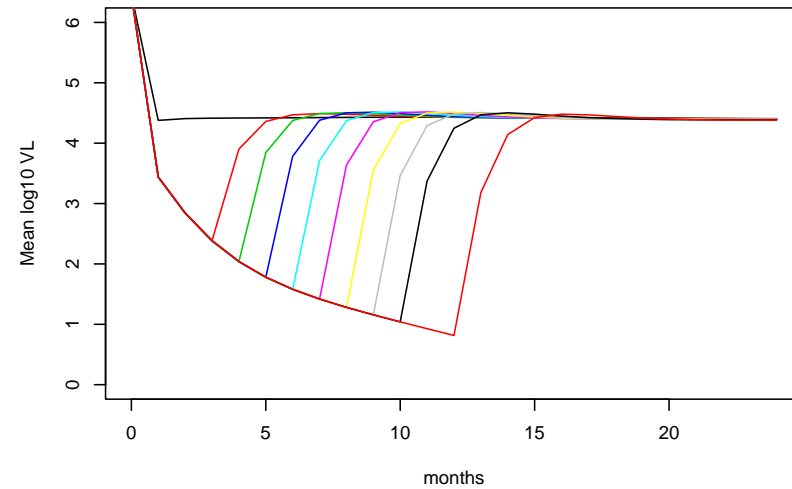
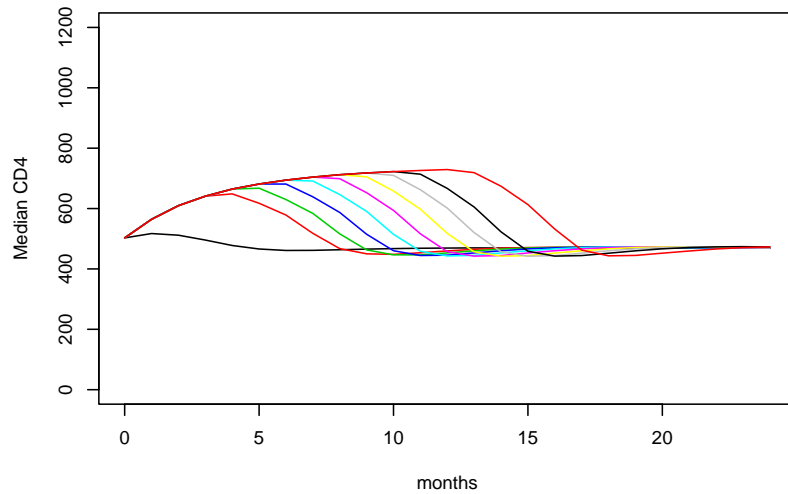
6. Using models

HIV dynamics: 100 “*virtual*” “*inherent*” viral load trajectories with *antiretroviral therapy terminated at 6 months*, i.e., $u(t) = 1, 0 \leq t \leq 6$, $u(t) = 0, t > 6$



6. Using models

HIV dynamics: Means of 15,000 “*virtual*” CD4 and viral load data profiles with $u(t) = 1, 0 \leq t \leq \tau, u(t) = 0, t > \tau, \tau = 0, 3, 4, \dots, 12$ months



7. Closing remarks

- For *application to data* mathematical models must be embedded in a *statistical model* that represents *sources of variation* in data
- *Inverse problem* should be regarded as an *statistical estimation/inference* problem
- *Estimation* should be accompanied by *assessment of uncertainty*
- Such *mathematical-statistical models* will be an increasingly important tool in biomedical research

References

- Adams, B.M., Banks, H.T., Davidian, M., and Rosenberg, E.S. (2007) Model fitting and prediction with HIV treatment interruption data. *Bulletin of Mathematical Biology* **69**, 563–584.
- Davidian, M. and Giltinan, D.M. (1995) *Nonlinear Models for Repeated Measurement Data*. London: Chapman & Hall.
- Davidian, M. and Giltinan, D.M. (2003) Nonlinear models for repeated measures data: An overview and update. Editor's invited paper, *Journal of Agricultural, Biological, and Environmental Statistics* **8**, 387–419.
- Rosenberg, E.S., Davidian, M., and Banks, H.T. (2007) Using mathematical modeling and control to develop structured treatment interruption strategies for HIV infection. *Drug and Alcohol Dependence* special supplement issue on “Customizing Treatment to the Patient: Adaptive Treatment Strategies” **88S**, S41-S51.