EXAMPLES: LINEAR, LONGITUDINAL MODELS AND COMPUTATION

(live longitudinally and prosper)

Don McLeish and Cyntha Struthers

University of Waterloo

Dec 5, 2015

Longitudinal models and regression

- Much of the analysis of longitudinal models is carried out using linear models or generalized linear models.
- We begin with methods for dealing with **missing responses** in a normal linear model.
- Then we look at methods for dealing with **missing covariates** in a normal linear model.
- Finally we look at methods for dealing with missing data in a longitudinal study.

Normal linear model: Responses only MAR

- Suppose $Y_i \sim N(\beta_0 + \beta_1 x_i + \beta_2 v_i, \sigma^2)$, $i = 1, \ldots, n$ ind.
- The score vector for one observation is

$$S(y_i|x_i, v_i, \beta) = \frac{\partial}{\partial \beta} \ln f_\beta(y_i|x_i, v_i)$$

$$= \frac{1}{\sigma^2} [y_i - (\beta_0 + \beta_1 x_i + \beta_2 v_i)] \begin{pmatrix} 1 \\ x_i \\ v_i \end{pmatrix}$$

- We want to estimate $\beta = (\beta_0, \beta_1, \beta_2)$ when some of the Y_i are MAR.
- Let $R_i = 1$ if Y_i is observed and $R_i = 0$ otherwise.

Normal linear model: Responses MAR

- For complete data, β is estimated by solving $\sum_{i=1}^{n} S(y_i|x_i, v_i, \beta) = 0.$
- Assume ALL data are independent (Y_i, R_i, x_i, v_i), $i = 1, \ldots, n$.
- Let $Y_{obs} = (R_i y_i, x_i, v_i)$, $i = 1, \ldots, n$ be observed data.
- Conditioning on Y_{obs}, we obtain the estimating function:

$$\sum_{i=1}^{n} \left[R_i S(y_i|x_i, v_i, \beta) + (1 - R_i) E_\beta \{ S(Y_i|X, V, \beta) | Y_{obs} \} \right].$$

- Assume Y is MAR, that is, $P(R = 1) = \pi(X, V)$ where π is a known function (up to vector of parameters γ).
- Because of MAR, the second term in the estimating function equals zero.

Normal linear model: Responses MAR

- The estimating function

$$\sum_{i=1}^{n} R_i S(y_i|x_i, v_i, \beta)$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} R_i \left[y_i - (\beta_0 + \beta_1 x_i + \beta_2 v_i)\right] \begin{pmatrix} 1 \\ x_i \\ v_i \end{pmatrix}$$

  gives the usual least squares estimator based on the observed values only: $R_i \times (y_i, x_i, v_i)$, $i = 1, \ldots, n$.

- **Conclusion:** *When the response is MAR, listwise deletion (complete-case analysis) is fully efficient for* **regression parameters** *relative to ML estimation based on* Y_{obs}.

Linear model: Covariates MAR

- Suppose (y_i, v_i) are observed but some x_i are MAR.
- Let $R_i = 1$ if x_i is observed and $R_i = 0$ otherwise. If observations are MAR, $P(R = 1) = \pi(Y, V)$ where π is a known function (up to vector of parameters γ).
- Assume ALL the data are independent observations (Y_i, R_i, x_i, v_i), $i = 1, \ldots, n$.
- For a normal linear model the ML estimating function is

$$\sum_{i=1}^{n} \left\{ R_i S(y_i | x_i, v_i, \beta) + (1 - R_i) E_\beta \left[S(Y | X, V, \beta) | y_i, v_i \right] \right\}.$$

- Second term in the estimating function uses information about x obtained from (y, v). LD is NOT consistent, or efficient relative to ML estimation based on the observed data.

Covariates missing: Parametric approach

- **Problem:** To evaluate $E_\beta[S(Y|X, V, \beta)|y_i, v_i]$ we need either $f_\beta(y, x|v)$ or $f(x|v)$ which are unknown.
- How do we estimate $f_\beta(y, x|v)$ or $f(x|v)$?
- **Option 1 - Parametric Model:**
- We have already assumed a parametric (normal) model for $f_\beta(y|x, v)$.
- If we also model $f(x|v)$ parametrically then we can obtain ML estimates based on the observed data using a method such as the EM algorithm.

Covariates missing: Semiparametric approach

- **Option 2 - Semiparametric Model:**
  Estimate $f(x|v)$ nonparametrically.
- For complete data, $E[S(y_i|X, v_i, \beta) f_\beta(y_i|X, v_i)|V = v_i]$ could be estimated using the average of the observed x_i's for a given $V = v_i$.
- This would require one or more complete observations for each observed value of v.
- Suppose for $V = v_i$ we have complete observations $(y_j, x_{j,} v_i)$, $j = 1, \ldots, J_i$.

Covariates missing: Semiparametric approach

- Estimate $E_\beta\left[S(Y|X,V,\beta)|y_i,v_i\right]$ using the weighted average

$$\frac{\sum_{j=1}^{J_i} S(y_i|x_j,v_i,\beta)w_{ij}}{\sum_{j=1}^{J_i} w_{ij}} \quad \text{where } w_{ij} = f_\beta(y_i|x_j,v_i)I(v_j = v_i).$$

- This approach is equivalent to estimating $f(x|v)$ (and conditional expectations given V) by averaging over the observed x's in the same stratum, where the stratum is defined by the common value of v.
- **Possible problem:** There may be too few observations in the stratum for a given value of v resulting in a poor estimate.

Semiparametric approach: Chatterjee, Chen and Breslow

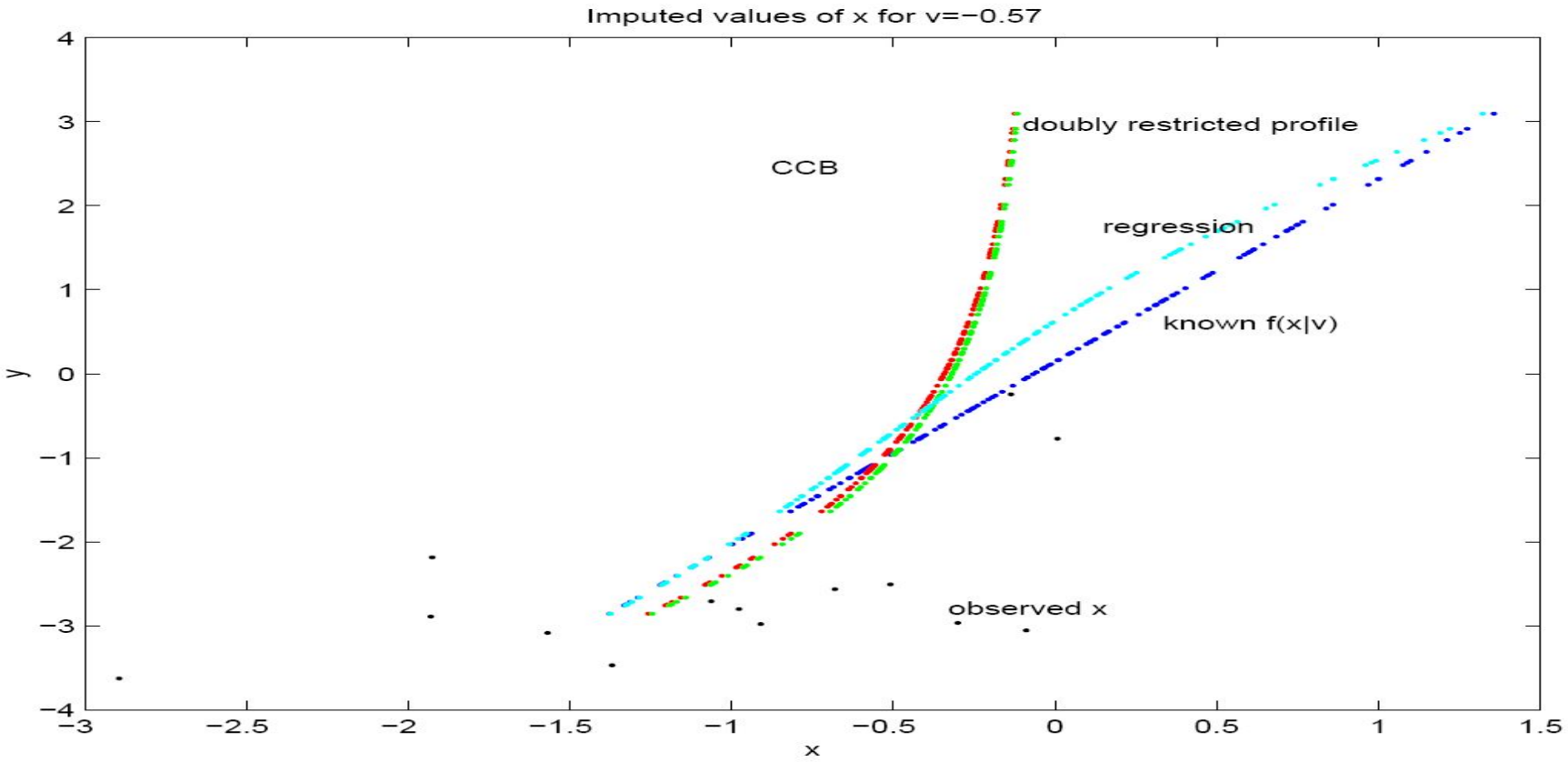
- Chatterjee, Chen and Breslow (2003) used inverse probability weighting to determine the weights w_{ij}.
- The probability a given point (y_j, x_j, v_j) is observed is $P(R_j = 1|y_j, v_j) = \pi(y_j, v_j)$.
- Let $E(R_j|x_j, v_j, \beta) = \eta(x_j, v_j; \beta)$.
- The Chatterjee, Chen and Breslow (CCB) weights are given by

$$w_{ij} = \frac{R_j}{\eta(x_j, v_j; \beta)} f_\beta(y_i|x_j, v_i) I(v_j = v_i).$$

Semiparametric approach: The profile estimator

- Another approach to determining the weights is to estimate $f(x|v)$ using the non-parametric ML estimate.
- The support of $f(x|v)$ for a given v is NOT just the observed x for that value of v.
- There are pairs (x, v) for which no data are observed and yet including these points in the support of $f(x|v)$ increases the likelihood function for $f(x|v)$. (See: McLeish and Struthers (2006)).

Example



In this example only small x are observed for this value of v. The CCB weights perform poorly in this case. Similarly, any hot-deck. ML estimation performs better since, for this v value, it also includes large x in the support set for $f(x|v)$.

Doubly robust estimating function

Several authors[1] have proposed a slightly different estimating function: $\psi(\beta) =$

$$\sum_{i=1}^{n} \{ \frac{R_i}{\eta(x_j, v_j; \beta)} S(y_i | x_i, v_i, \beta) + [1 - \frac{R_i}{\eta(x_j, v_j; \beta)}] E_\beta \left[S(Y | X, V, \beta) | y_i, v_i \right] \}$$

$$= \sum_{i=1}^{n} \frac{R_i}{\eta(x_j, v_j; \beta)} \left\{ S(y_i | x_i, v_i, \beta) - E_\beta \left[S(Y | X, V, \beta) | y_i, v_i \right] \right\} + \sum_{i=1}^{n} E_\beta \left[S(Y | X, V, \beta) | y_i, v_i \right]$$

where $P(R_j = 1 | x_j, v_j) = \eta(x_j, v_j; \beta)$.

[1] e.g. Robbins, Rotnisky

Doubly robust estimating function

- If $\eta(x_j, v_j; \beta)$ is correctly specified, then $\psi(\beta) =$ $\sum_{i=1}^{n} \frac{R_i}{\eta(x_j, v_j; \beta)} \left\{ S(y_i | x_i, v_i, \beta) - E_\beta \left[S(Y | X, V, \beta) | y_i, v_i \right] \right\}$ $+ \sum_{i=1}^{n} E_\beta \left[S(Y | X, V, \beta) | y_i, v_i \right]$ has the same expected value as $\sum_{i=1}^{n} S(y_i | x_i, v_i, \beta)$.
- Therefore $\psi(\beta)$ is unbiased and the estimator obtained by solving $\psi(\beta) = 0$ is consistent.

Doubly robust estimating function

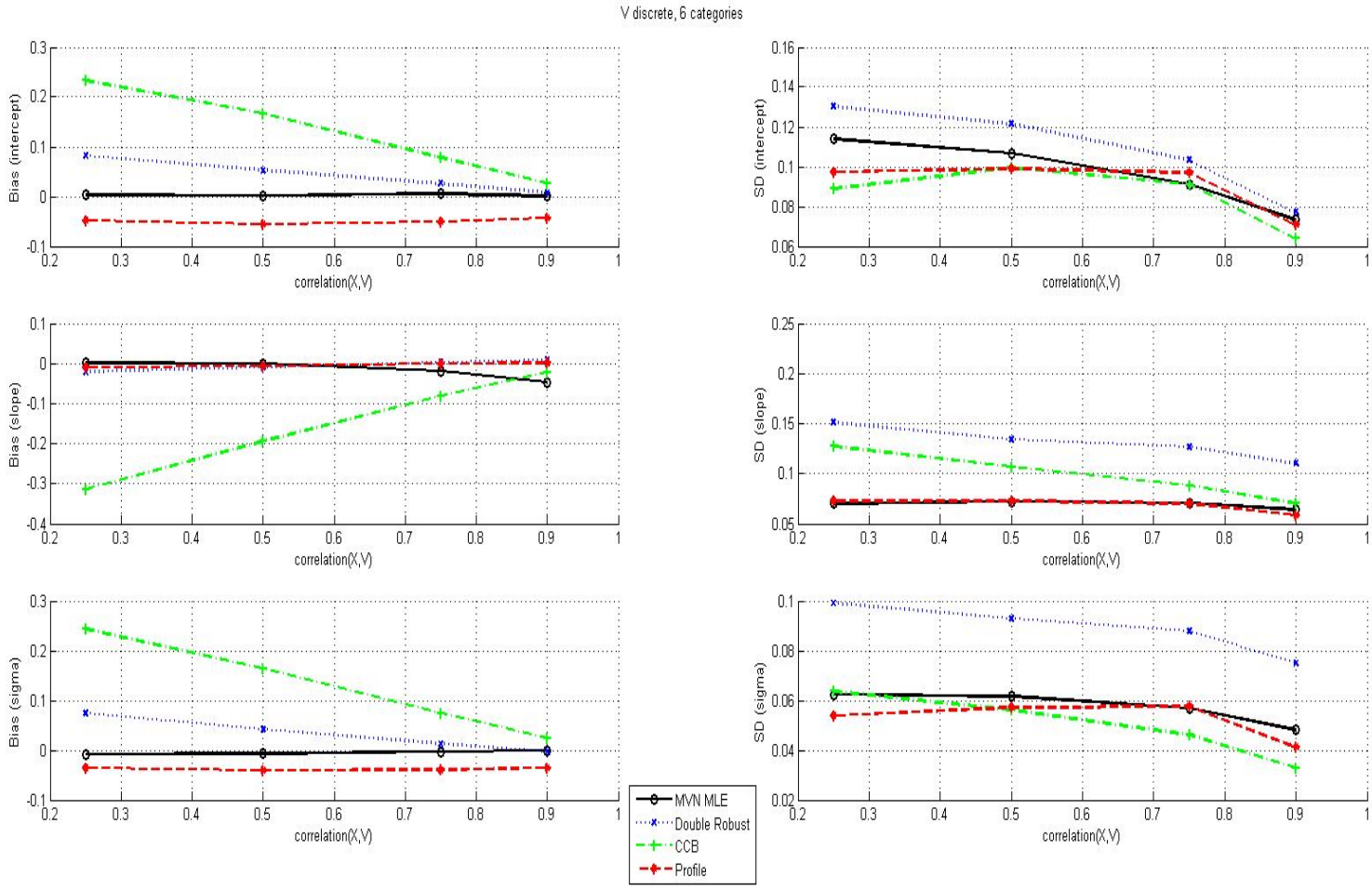
If $\eta(x_j, v_j; \beta)$ is incorrectly specified but $f(x|v)$ is correctly specified then $\left\{S(y_i|x_i, v_i, \beta) - E_\beta\left[S(Y|X, V, \beta)|y_i, v_i\right]\right\}$ and $E_\beta\left[S(Y|X, V, \beta)|y_i, v_i\right]$ both unbiased estimating functions.

- So is $\psi(\beta) = \sum_{i=1}^{n} \frac{R_i}{\eta(x_j, v_j; \beta)} \left\{S(y_i|x_i, v_i, \beta) - E_\beta\left[S(Y|X, V, \beta)|y_i, v_i\right]\right\} + \sum_{i=1}^{n} E_\beta\left[S(Y|X, V, \beta)|y_i, v_i\right]$. The estimators are consistent.
- Only **ONE** of η and $f(x|v)$ needs to be correctly specified for consistency. (called *double-robustness*[2] *property*). If both are correct, full semiparametric efficiency obtained.

[2] Han and Wang (2014) use *multiple robustness*. Given finite sets of models for π and $f(x|v)$, if any one is correct consistency is obtained.

Regression with missing covariates: Comparison of the MSE and bias of various estimators

Y, v always observed. Joint model for (y, x, v) is MVN.[3]



[3] v grouped into 6 categories, $\mathrm{P}(x \text{ observed}) \approx 0.15$, $n_{obs} \approx 150$.

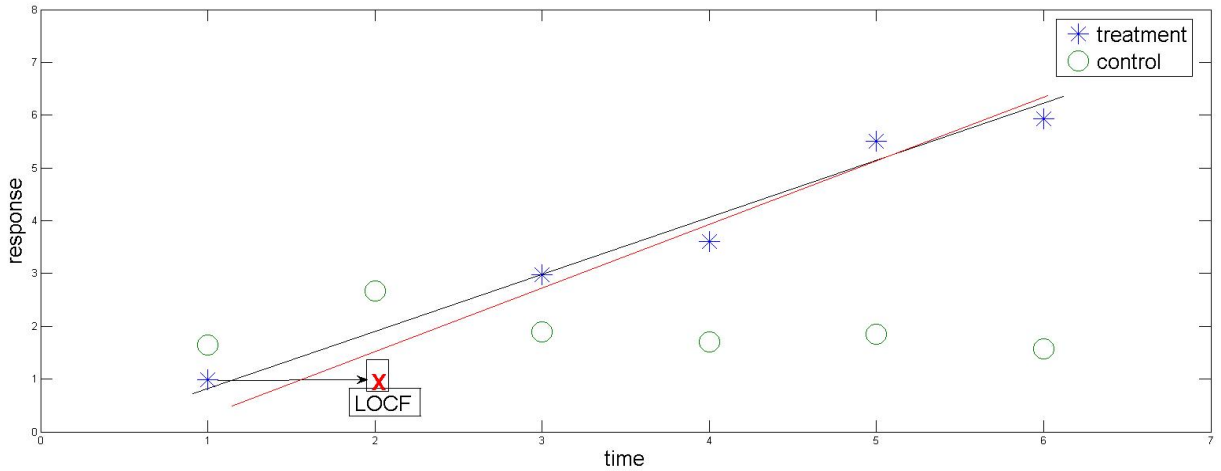
National Research Council Special Report on Missing Data in Clinical Trials

Recommendation 13 of the N.R.C. Special Report: Weighted generalized estimating equations should be more widely used in settings when missing at random can be well justified and a stable weight model can be determined, as a possibly useful alternative to parametric modeling.

Longitudinal data

- In longitudinal models, individuals are measured repeatedly over time.
  - For example, in a study on a treatment for high blood pressure a patient's blood pressure is measured repeatedly over time since the treatment effect may vary over time.
- Why are longitudinal studies more difficult to analyze?
  - Models for longitudinal data must taken in to account the correlation between measurements over time for a given individual. (Involves a large number of parameters unless simplifying assumptions are made about the correlation structure.).
  - Dropout may occur or individuals may be censored (e.g. removed from treatment due to adverse side effects).

Longitudinal data: Naive single imputation



A simple but naive solution to the problem of missing data is to replace a missing observation with *Last Observation Carried Forward* (LOCF) or *Baseline Observation Carried Forward* (BOCF).

- Neither reflect MAR assumptions.
- Both may lead to bias[4]. The bias is not necessarily conservative.

[4] Mohlenbergs and Kenward (2007,2009)

Longitudinal data: Naive single imputation

- Panel on Handling Missing Data in Clinical Trials recommends against LOCF and BOCF use without justifying assumptions.
- Conservative approach:
  - Use best possible outcome for missing outcome in the control group.
  - Use worst possible outcome for missing outcome in the treatment group.
- This approach is often used in a sensitivity analysis to determine if the imputation method affects conclusions.

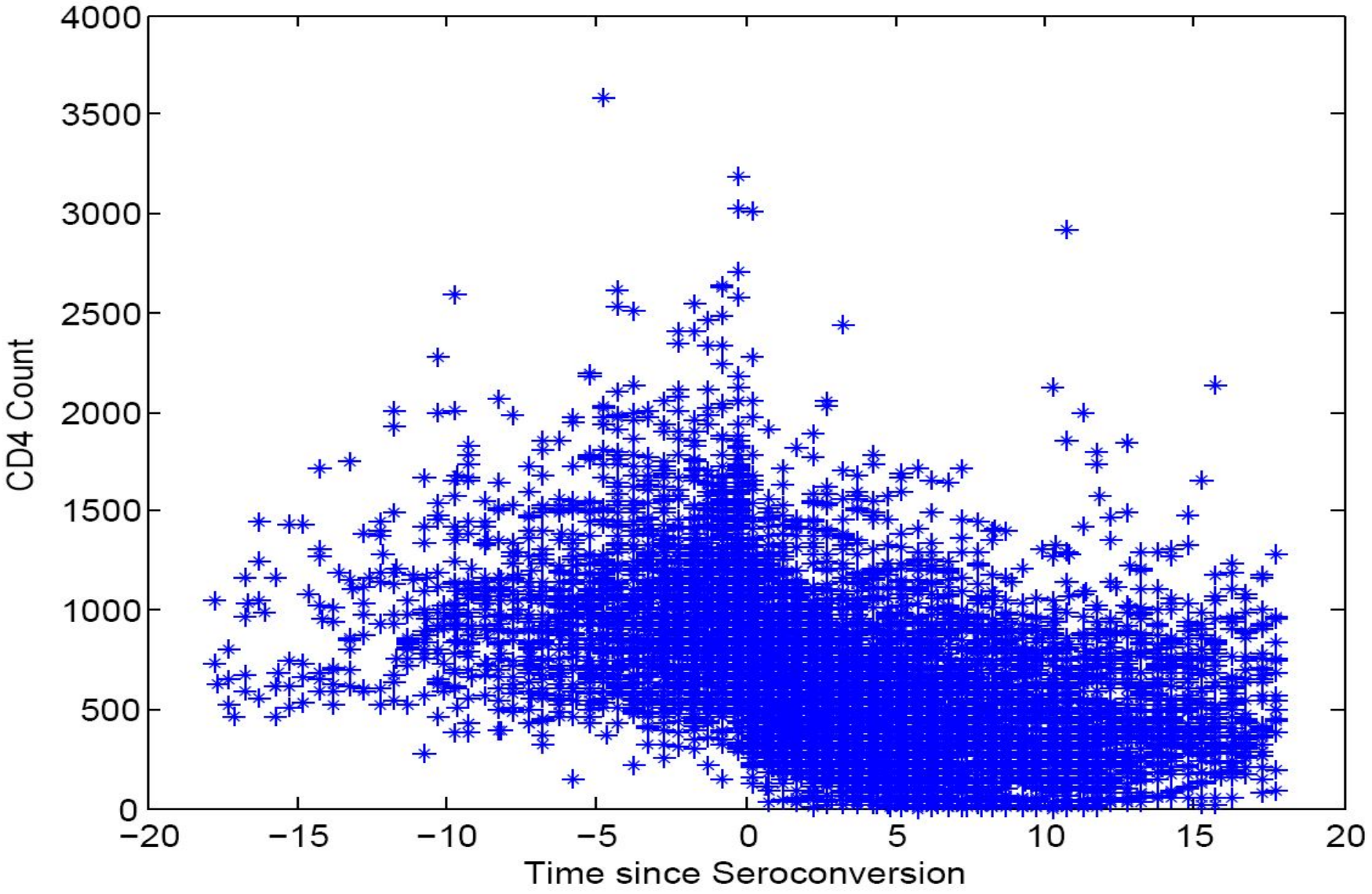
Longitudinal data motivating example: Multicenter AIDS Cohort Study (MACS)

- Diggle et al. (2002) analyzed a subset of the 2007 release of the Multicenter AIDS Cohort Study (MACS)
- Public data set available from http://statepi.jhsph.edu/macs/macs.html
- The data cover the years 1984 - 2002.
- Participants were followed up every six months (maximum number visits $= 37$).
- Information on 5622 gay and bisexual men.

Motivating example: Multicenter AIDS Cohort Study (MACS)

- The public data set includes participants who seroconverted from HIV- to HIV+ during the study.
- Data set also includes information on CD4 count at each visit, date of last negative visit, and date of first positive visit. These data are sometimes missing.
- There were 595 seroconverters, of whom 29 had missing seroconversion intervals, and 126 had intervals of seroconversion greater than 6 months.
- The data consist of 11743 visits for the 595 seroconverters.
- Diggle et al. analyzed the data for participants for whom the seroconversion window was 6 months or less.

CD4 count versus time since conversion for seroconverters with observed seroconversion window



The data

- **Response:**
  - CD4 count
- **Covariates:**
  - time of seroconversion from HIV- to HIV+
  - cigarette smoking (no, $\leq \frac{1}{2}$ packs, $> \frac{1}{2}$ packs/day)
  - recreational drug use (yes/no)
  - depressive symptoms (scale $[-7, 53]$)
  - number of sexual partners $(0, 1, > 1)$
  - use of drugs to fight AIDS (yes/no).

Missingness in the data

Missingness

Variable	Number of Visits Missing	Percent
CD4 Counts	1190	10.1%
Cigarette Smoking	114	1%
Recreational Drugs	237	2%
Antivirals AIDS Drugs	112	1%
Non-Antivirals AIDS Drugs	111	1%
Depressive Symptoms	666	5.7%
Number of Male Sexual Partners	302	2.6%

Missingness in the data

- If we exclude any visit for which there is a missing CD4 count or covariate we have 9997/11743 or 85% of the data.
- If we also exclude visits corresponding to seroconverters whose interval of seroconversion is greater than 6 months then we have 8005/11743 or 68% of the data.
- In this assessment of missingness we have not considered visits which were completely missed. (For example some participants had missing visits for several years in the middle of the study and then started attending visits again, often after seroconverting.)

Simple linear model for CD4 counts

Consider the simple model

$$Y_{ij} = \mu(t_{ij}) + \beta_3^T \mathbf{x}_i(t_{ij}) + \varepsilon_{ij}$$

where

$$\begin{aligned} Y_{ij} &= \sqrt{CD4} \text{ for individual } i \text{ measured at time } t_{ij},\ j = 1, ..., n_i \\ t_{ij} &= \text{time since seroconversion for subject } i \text{ at visit } j \\ \mathbf{x}_i(t_{ij}) &= \text{values of covariates for subject } i \text{ at visit } j \\ \varepsilon_{ij} &\sim N(0, \sigma^2) \text{ random errors} \end{aligned}$$

$$\mu(t) = \begin{cases} \beta_0 + \beta_1 t + \beta_2 t^2, & t > 0 \\ \beta_0 & t \leq 0 \end{cases}$$

and $\beta_0, \beta_1, \beta_2, \beta_3^T$ are regression coefficients to be estimated.

Simple linear model for CD4 counts

If the Y_{ij} are independent this is just a simple linear model and the least squares estimate of $\beta = \left(\beta_0, \beta_1, \beta_2, \beta_3^T\right)$ is found by solving the estimating equation

$$\sum_{i=1}^{n} \mathbf{X}_i^T (\mathbf{Y}_i - \mathbf{X}_i \beta) = 0$$

where

$$\mathbf{X}_i = \begin{bmatrix} 1 & t_{i1}^+ & t_{i1}^{+2} & [\mathbf{x}_i(t_{i1})]^T \\ 1 & t_{i2}^+ & t_{i2}^{+2} & [\mathbf{x}_i(t_{i2})]^T \\ \vdots & & & \\ 1 & t_{in_i}^+ & t_{in_i}^{+2} & [\mathbf{x}_i(t_{in_i})]^T \end{bmatrix} \quad \text{and} \quad \mathbf{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{bmatrix}$$

where $t_{ij}^+ = t_{ij}$ if $t_{ij} > 0$, and 0 otherwise.

A more realistic model

Observations on the same individual at adjacent times t_{ij}, $t_{i(j+1)}$, if close, should be positively correlated.

- There should be larger correlation within individuals than between them (individual effect)

$$Cov(Y_{ij}, Y_{ik}) > Cov(Y_{ij}, Y_{mk}), \text{ if } m \neq i$$

- The estimate of β is obtained by solving

$$\sum_{i=1}^{n} \mathbf{X}_i^T \left[Var\left(\mathbf{Y}_i\right)\right]^{-1} \left(\mathbf{Y}_i - \mathbf{X}_i\beta\right) = 0.$$

- The estimates are the ML estimates for a MVN model if $Var\left(\mathbf{Y}_i\right)$ is correctly specified.
- Consistent estimators of β are obtained even if $Var\left(\mathbf{Y}_i\right)$ is not correctly specified.

Semiparametric approach

- In a semiparametric approach to the estimation of β it is assumed that $E(\mathbf{Y}_i) = \mu_i(\beta) = g^{-1}\left(x_i^T\beta\right)$ or $x_i^T\beta = g(\mu_i(\beta))$ where g is called a link function.
- Possible choices for g are the identity function, the log function, the logit function, etc.
- The estimate of β is then determined by solving the equation

$$\sum_{i=1}^{n}\left(\frac{\partial\mu_i}{\partial\beta}\right)^T [Var(\mathbf{Y}_i)]^{-1}[\mathbf{Y}_i - \mu_i(\beta)] = 0$$

  which is called a generalized estimating equation (GEE) or quasi-score estimating equation.
- The weighted least squares equation is a special case of a quasi-score estimating equation.

Quasi-Score function with missing response

- Let $R_{ij} = 1$ if Y_{ij} observed and $R_{ij} = 0$ otherwise. Suppose also that $P(R_{ij} = 1) = p_{ij} > 0$.
- Let $D_i = diag(R_{ij}/p_{ij})$.
- An unbiased estimating equation obtained from the quasi-score function is the inverse probability weighted (IPW) estimating equation given by

$$\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta} \right)^T [Var(\mathbf{Y}_i)]^{-1} D_i [\mathbf{Y}_i - \mu_i(\beta)] = 0$$

Semiparametric approach

- In the quasi-score estimating equation

$$\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta} \right)^T \left[Var\left(\mathbf{Y}_i \right) \right]^{-1} \left[\mathbf{Y}_i - \mu_i(\beta) \right] = 0$$

  the term $Var\left(\mathbf{Y}_i \right)$ may be replaced by a "working" correlation matrix which does not depend on the regression parameters β but may depend on correlation parameters α.
- If $Var\left(\mathbf{Y}_i \right)$ is replaced by a consistent estimator then there is no loss of efficiency in the estimation of β.
- If the working correlation matrix is not correctly specified the estimator of β is still consistent.

Example of working correlation matrices

- **Independence** (components uncorrelated)
$$\begin{pmatrix} 1 & 0 & 0. \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$
- **Exchangeable** (compound symmetry) $\begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$
- **AutoRegressive Order 1** (AR 1) $\begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}$
- **General** (unstructured) $\begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix}$

Quasi-Likelihood: Justification of the quasi-score

Among linear combinations of the functions $Y_i - \mu_i(\beta)$, (that is, all estimating functions linear in Y_i) the quasi-score is the "best", in the sense of having the smallest asymptotic variance (maximum Godambe information) among all such functions.
This is called *semi-parametric efficiency*.
This is "best in show"....but how good is the competition?

Modelling Var(Y): Are all men created equal?

Nearly every longitudinal model needs to accommodate an individual effect. For example, some individuals have higher CD4 counts on average than others.

- A "random effect" model accommodates individual effects using a "hidden" factor U_i that depends on the individual.
  - Assume $\mathbf{Y}_i$ given U_i is $MV\mathbf{N}(\mu_i(\beta), V_i)$ where U_i has some distribution. Often $U_i \sim N(0, \sigma^2_{re})$ is assumed.
  - Assume also that $\mathbf{Y}_i = U_i\mathbf{1} + \mu_i(\beta) + \varepsilon_i$ where $\mathbf{1}$ is a vector of 1's. Then

$$Var\left(\mathbf{Y}_i\right) = Var(U_i)\mathbf{1}\mathbf{1}^T + Var(\varepsilon_i).$$

  - A random effect (intercept) adds a constant $Var(U_i) = \sigma^2_{re}$ to every element of the covariance matrix $Var(\mathbf{Y}_i)$ but does not affect the covariances $Cov(Y_{ij}, Y_{mk})$ where $m \neq i$.

Modelling Var(Y)

- If we assume $\mathbf{Y}_i$ given U_i is $MV\mathbf{N}(\mu_i(\beta), V_i)$ then the simplest structure for $\mathbf{V}_i$ is a diagonal matrix $\sigma^2_{me}\mathbf{I}$ where $\mathbf{I}$ is the identity matrix.
- This structure assumes independence of the values Y_{ij}, $j = 1, 2, \ldots$ given U_i and models measurement error.

Modelling Var(Y)

- It is also common to assume exponential decay over time of the correlation, that is $Cov\left(Y_{ij}, Y_{ik}\right) = e^{-\lambda|t_{ij}-t_{ik}|}$.
- For example, for 4 equally spaced observation times, this is equivalent to assuming $\mathbf{V}_i$ has the form

$$\mathbf{V}(r) = \begin{bmatrix} 1 & r & r^2 & r^3 \\ r & 1 & r & r^2 \\ r^2 & r & 1 & r \\ r^3 & r^2 & r & 1 \end{bmatrix}$$

where $|r| < 1$.

Modelling Var(Y)

If we combine

measurement error (me)
+ serial correlation
+ random effect (re)

then the assumed form of $Var\left(\mathbf{Y}_i\right)$ is

$$\sigma^2_{me}\mathbf{I} + \mathbf{V}(r) + \sigma^2_{re}\mathbf{1}\mathbf{1}^T$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{1}$ is a vector of 1's.

Back to CD4 count's and approach of Diggle et al.

$Y_{ij} = \mu(t_{ij}) + \beta_3^T \mathbf{x}_i(t_{ij}) + \varepsilon_{ij}$ where $Y_{ij} = \sqrt{CD4}$ for individual i measured at time t_{ij}, $j = 1, ..., n_i$

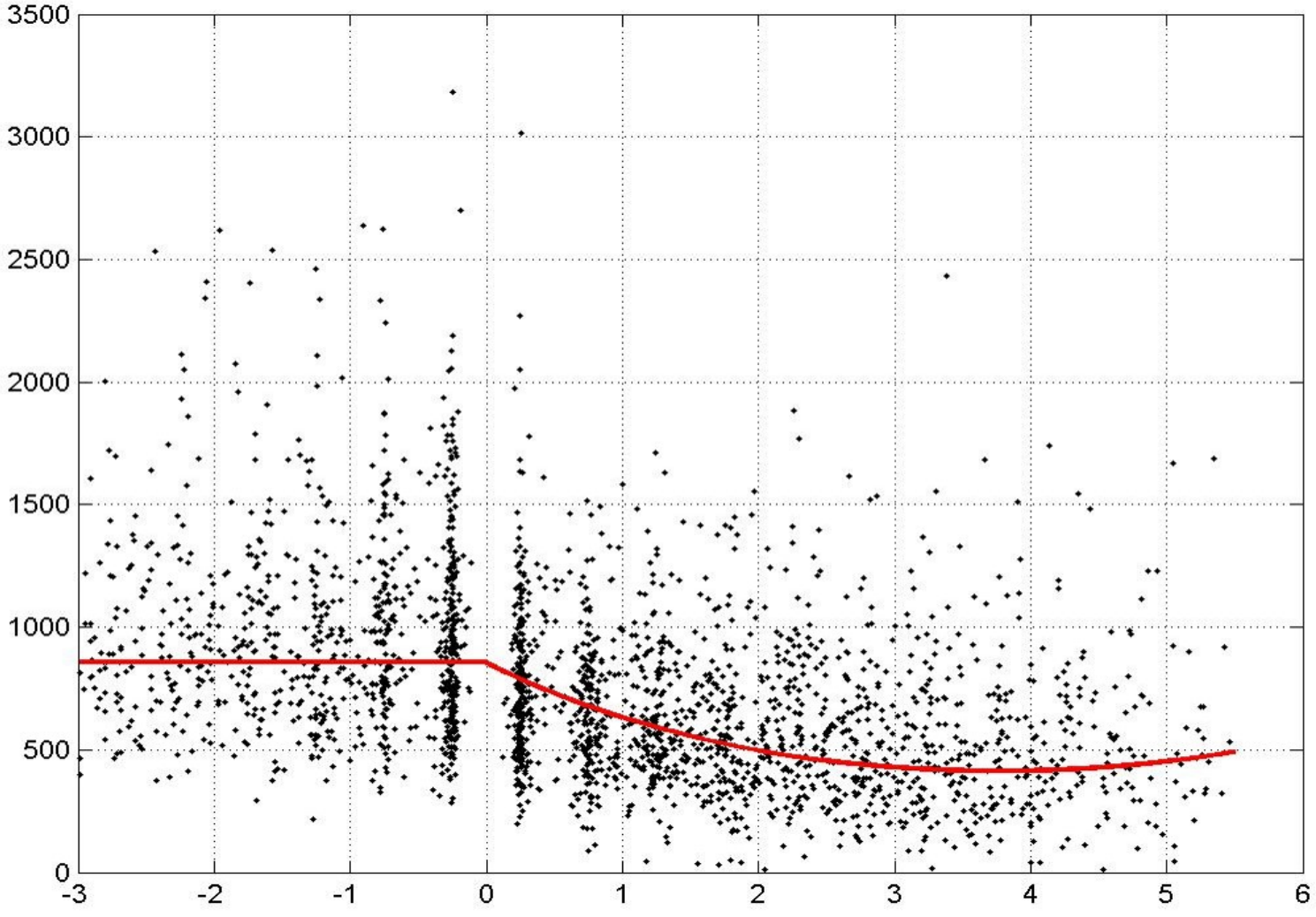
$t_{ij} =$ time since seroconversion for subject i at visit j

$\mathbf{x}_i(t_{ij}) =$ values of covariates for subject i at visit j

$$\mu(t) = \begin{cases} \beta_0 + \beta_1 t + \beta_2 t^2, & t > 0 \\ \beta_0 & t \leq 0 \end{cases}$$

- $Var\,(\mathbf{Y}_i)$ is assumed to be of the form: measurement error + serial correlation + random effects.
- Analysis is based only on complete cases, t_{ij} assumed to be the time since the midpoint of the seroconversion interval.

Diggle et al. model

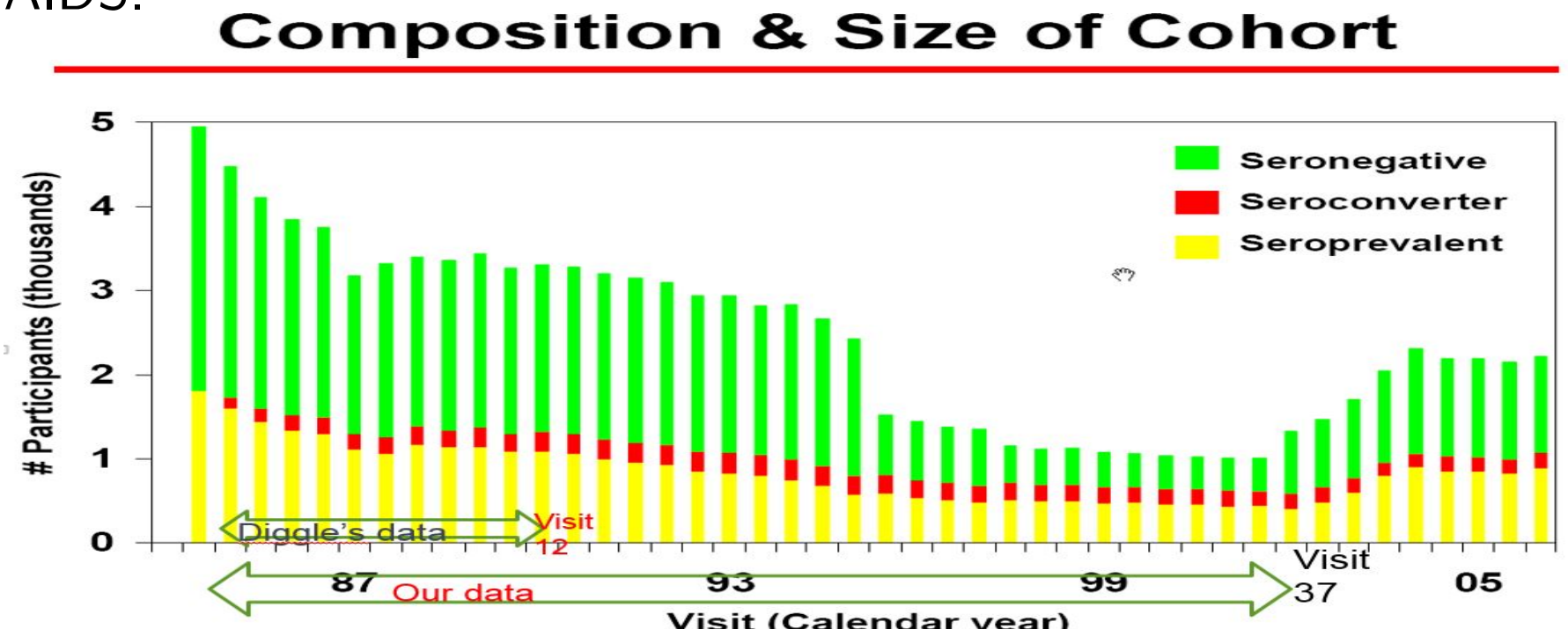


Deficiencies with the above model

- There is no attempt to deal with any of the missing data.
- The time of seroconversion is "guesstimated" by the midpoint of the interval between last HIV- visit and first HIV+ visit.
- CD4 count, like temperature or blood pressure, should be modeled with a process that has
  - a continuous-time extension
  - a stationary limit as $t \to \infty$ (quadratic model increases).

The data

Include information on response: CD4 counts (11102 visits), on covariates: smoking, recreational drug use, AIDS drugs and number of sexual partners (11644 visits and 3332 columns on 1948 variables), on depressive scale (11157 visits), on year of AIDS diagnosis and year of death. 212 seroconverters died of AIDS.



Are the data really missing?

- It is easy to identify a completely missed visit. In other cases it is not so easy.
- Number of packs smoked per day is missing for 7970/11644 (68.4%) visits.
  - Visits 1-7: "Do you smoke cigarettes now?"
  - Visits 8-37: "Have you ever smoked?"
  - Actually need to look at the questionnaires (available on the MACS website) and how they changed from visit to visit.
  - "No. of packs smoked/day" actually missing for only 15 visits.

Data cleaning....it's a dirty job but somebody ...

- After all the data were cleaned up it consisted of 11743 visits for the 595 seroconverters.
- "Complete Cases"
  - If we exclude any visit for which there is a missing CD4 count or covariate or the interval of seroconversion is greater than 6 months then we observe 68% of the data.

Our objectives

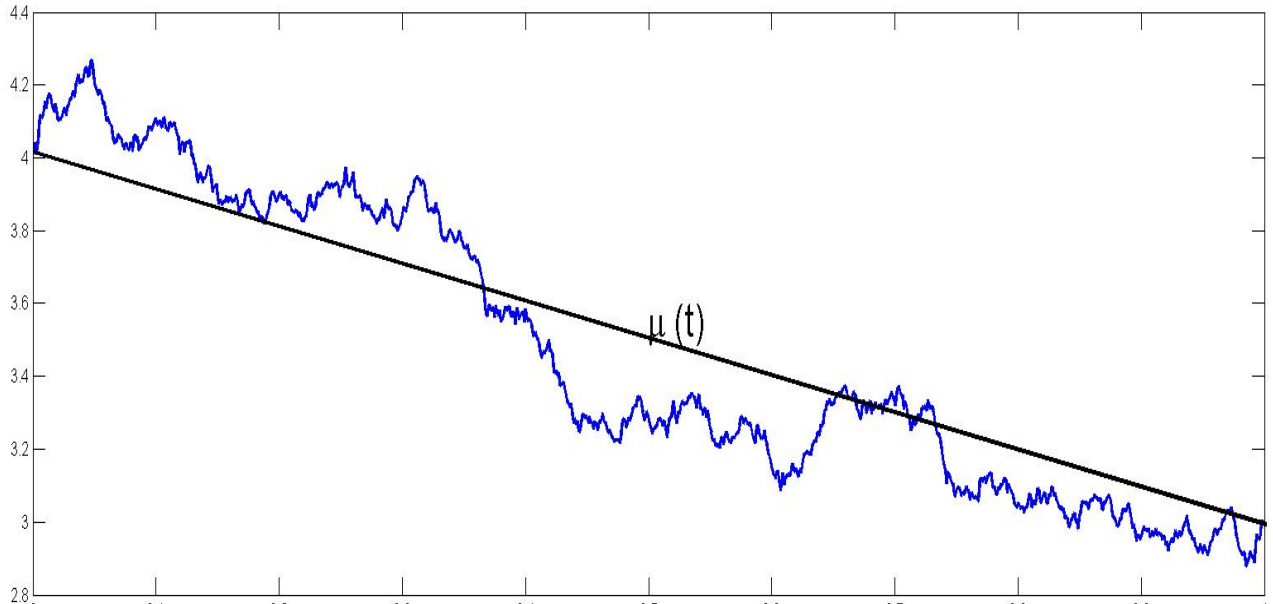
- Fit natural continuous-time (diffusion) model to data in which the covariates affect the parameters of the diffusion. (See Struthers and McLeish (2011).)
- Investigate the effect of a more conscientious treatment of missing data with respect to following objectives (see Diggle et al.).
  1. Estimate the average time of CD4 cell depletion.
  2. Estimate course for individuals with measurement error.
  3. Identify factors which predict CD4 cell changes.
  4. Characterize relationship between CD4 and progression to death due to AIDS.

A model for continuous time longitudinal data: Gaussian Ornstein-Uhlenbeck process with measurement error

$$dY_t = \kappa\left[\mu(t) - Y_t\right] + \sigma dW_t$$

$Y_t = (\text{True } CD4)^{0.4}$. $\mu(t) =$ real underlying process value at time t which depends on the covariates.

- Y_t tends *in the direction* of an unobservable "target", the current value of $\mu(t)$ which varies with covariates.

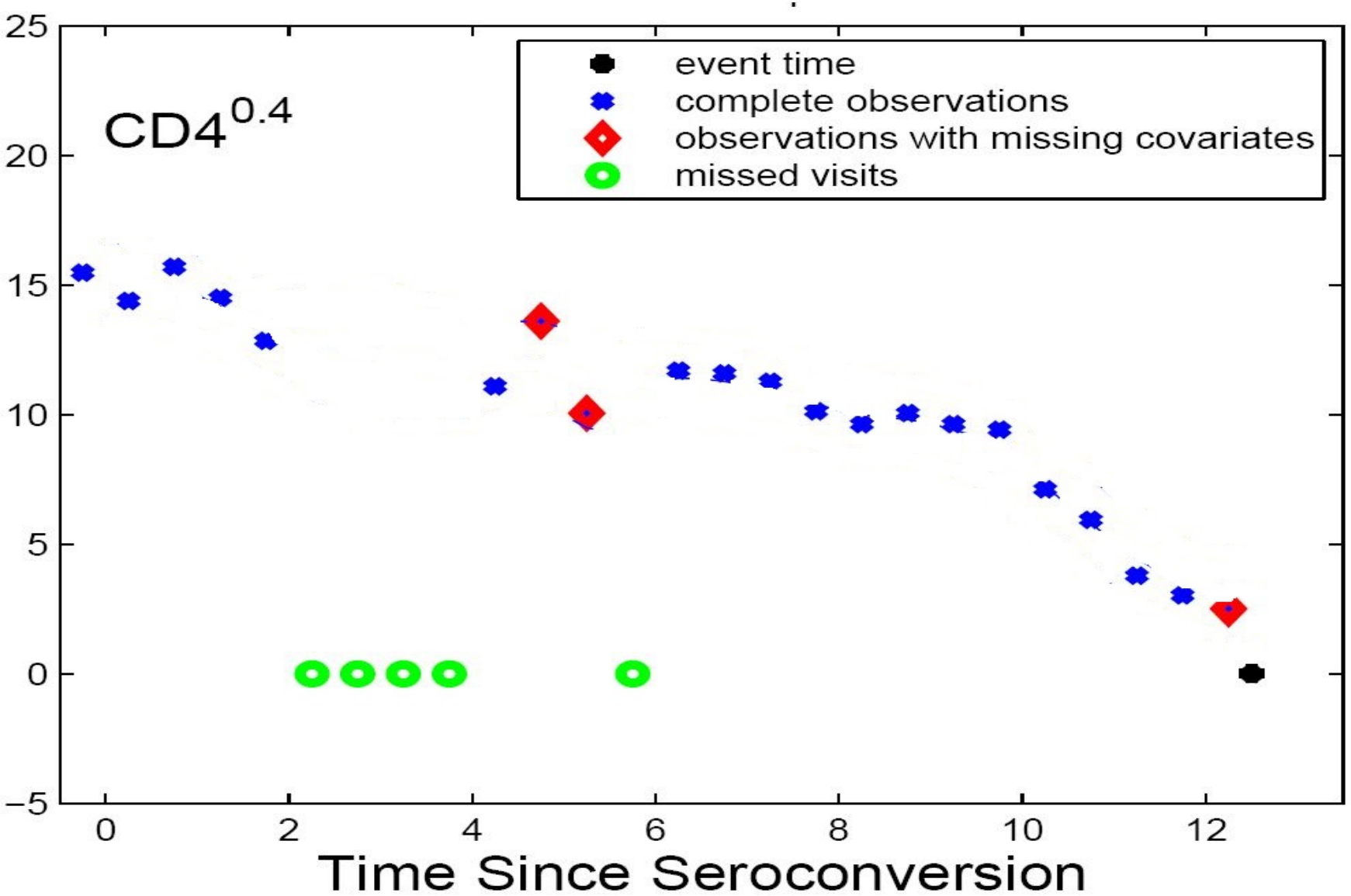


A model for continuous time longitudinal data: Gaussian Ornstein-Uhlenbeck process with measurement error

$$dY_t = \kappa \left[\mu(t) - Y_t\right] + \sigma dW_t$$

- κ determines how fast process moves towards $\mu(t)$.
- σ controls how much random movement.
- We observe Y_t+ measurement error at certain times.
- One objective: construct a model for $\mu(t)$ and determine covariate effects.

A typical path: Plot of path for participant 4070



Question: www.socrative.com (calgary)

Problem

The missing CD4 counts (response) corresponding to missing visits can likely be assumed to be

A. *MCAR*
B. *MAR*
C. *MNAR*

Question: www.socrative.com (calgary)

Problem

The missing values for the "number of packs smoked per day" covariate can be assumed to be

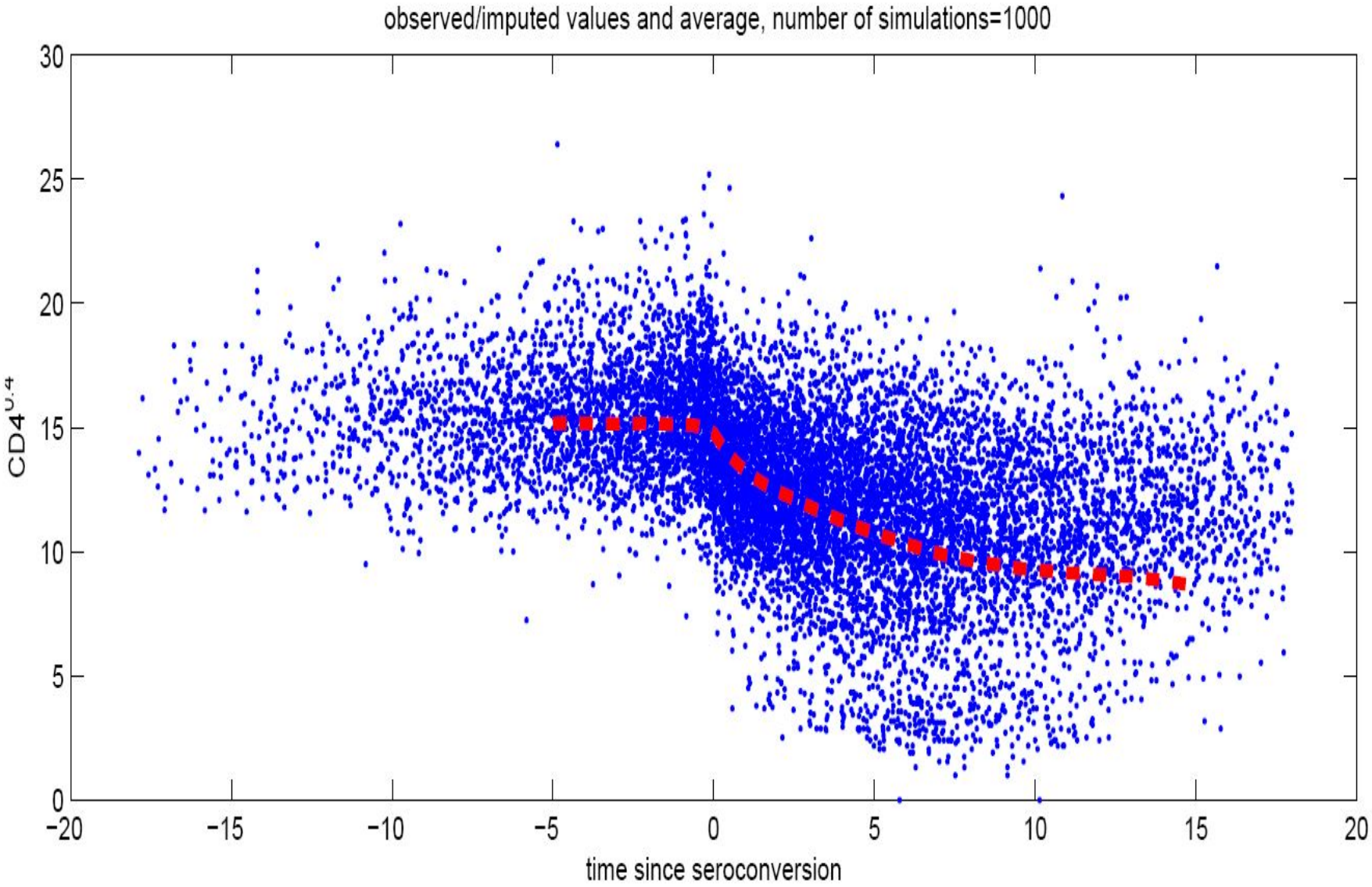
A. *MCAR*
B. *MAR*
C. *MNAR*

Prior distributions and generating from complete data posterior

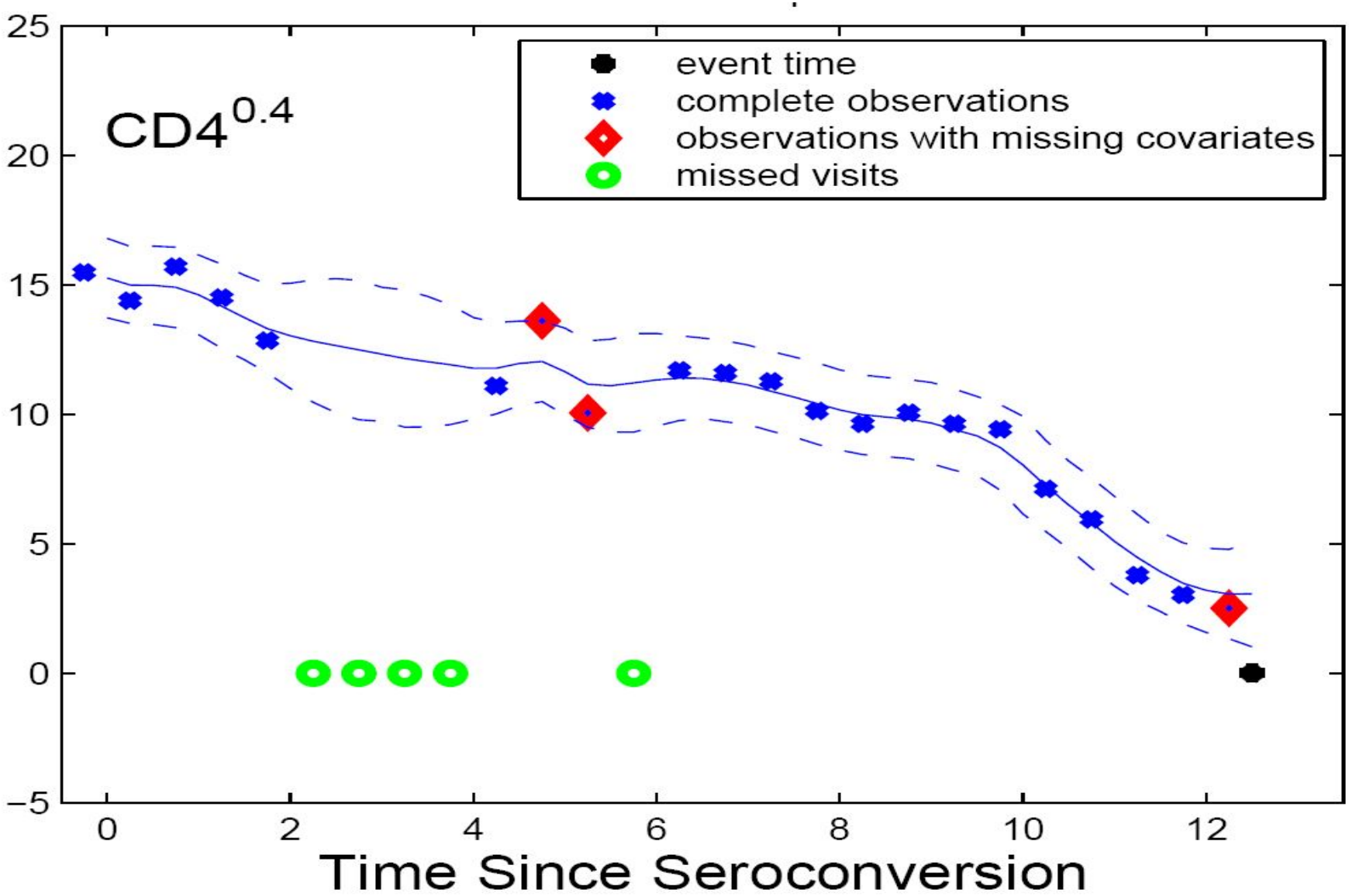
1. For regression parameters β and missing CD4 counts we used improper uniform priors so the posteriors are MVN.
2. For time varying covariate values we used a discrete state Markov chain with estimated infinitesimal generator as prior.
3. For κ and variance parameters, we used a uniform prior on suitable interval.

The posteriors for 2 and 3 are generated using acceptance-rejection, i.e. generate candidate parameter and accept with probability that depends on the prior.

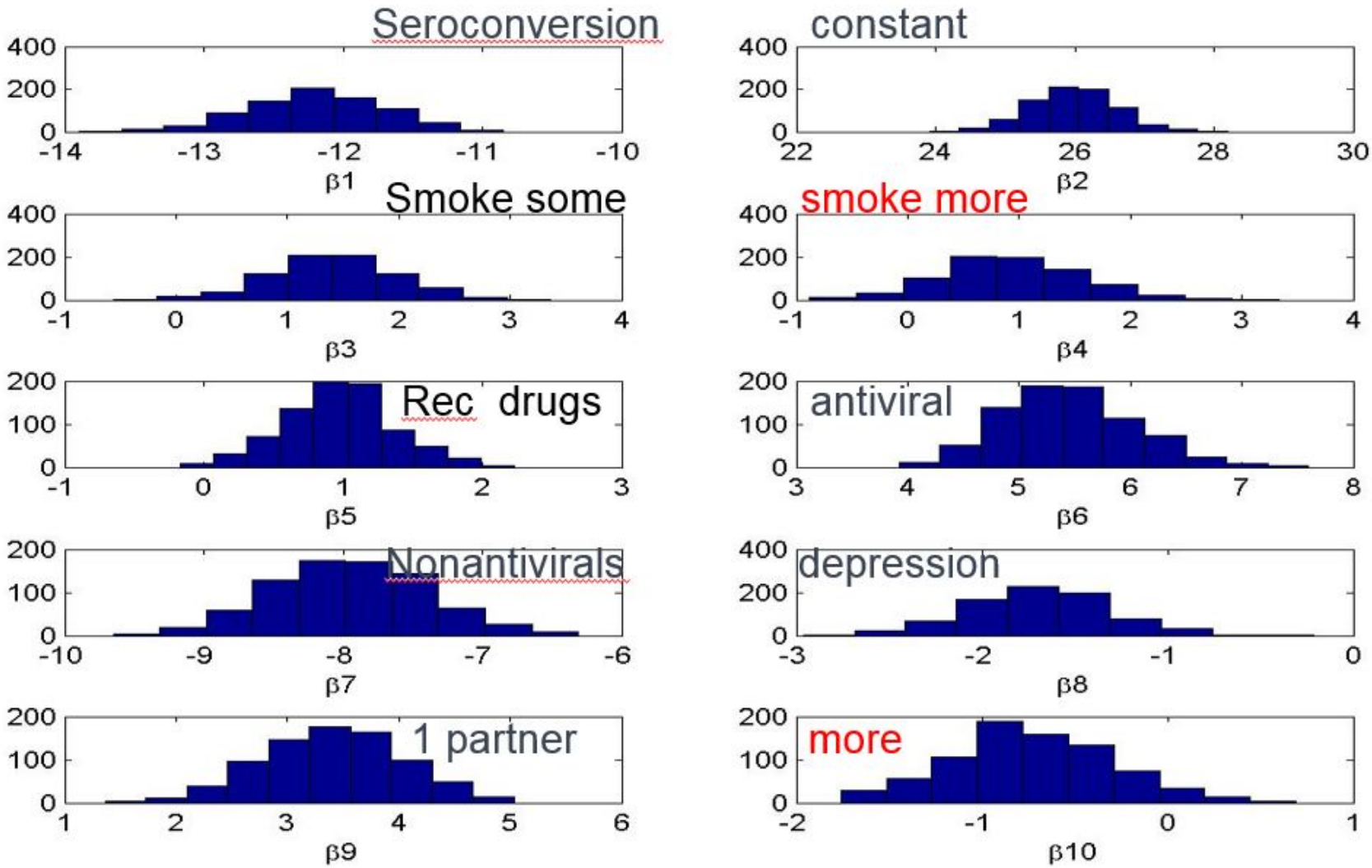
Results: Our fitted model for the mean



Results: Fitted expected CD4 count and credible interval for participant 4870



Results: Posterior distributions for regression coefficients



Red=not significant

Monotone, and the living is easy....

- Suppose Y is a $d-$dimensional vector corresponding to responses from an individual at consecutive times $t_1, \ldots, t_d$.
- If an individual "drops out" after time t_k, $k < d$, then the responses $Y_{k+1}, \ldots, Y_d$ are missing.
- This is an example of a **monotone missing pattern**:
  - indices for $Y_1 \subset$ missing indices for $Y_2 \cdots \subset$ missing indices for Y_d.
- For MVN and a monotone missing pattern, both ML estimates and multiple imputations are straightforward.

Likelihood for MVN and monotone missing pattern

Suppose we write the joint likelihood as $L(\theta_1, \theta_2, \ldots, \theta_d)$

$= f_{\theta_1}(y_1) f_{\theta_2}(y_2|y_1) f_{\theta_3}(y_3|y_1, y_2) \cdots f_{\theta_d}(y_d|y_{d-1}, y_{d-2}, \ldots, y_1)$

where θ_k are the parameters of $f_{\theta_k}(y_k|y_1, y_2, \ldots, y_{k-1})$.

- If Y is MVN then parameter θ_1 of $f_{\theta_1}(y_1)$ estimated from the marginal distribution of the observed y_1 values.
- θ_2 is estimated by fitting a linear regression of y_2 on y_1 using the completely observed values of (y_1, y_2).
- For $k = 1, 2, \ldots, d$, θ_k is estimated by fitting a linear regression of y_k on $y_1, y_2, \ldots, y_{k-1}$ using the completely observed values of $(y_1, y_2, \ldots, y_{k-1}, y_k)$.

These conditional distributions provide the ML estimates of parameters in the joint MVN model.

Confidence intervals are constructed using the estimated variances provided by the regressions and the delta method.

Monotone missing and imputations

For a monotone missing pattern the posterior distribution $\pi(\theta_1, \theta_2, \ldots, \theta_d | y_1, y_2, \ldots, y_k)$ can also be factored:

$$cf_{\theta_1}(y_1)\pi(\theta_1)f_{\theta_2}(y_2|y_1)\pi(\theta_2)\cdots f_{\theta_d}(y_d|y_{d-1}, y_{d-2}, \ldots, y_1)\pi(\theta_d)$$
$$= \pi(\theta_1|y_1)\pi(\theta_2|y_1, y_2)\cdots\pi(\theta_d|y_1, y_2, \ldots, y_d)$$

This factorization means the missing $y's$ can be imputed as follows:

- Generate θ_1 from the posterior $\pi(\theta_1|y_1^{obs})$.
- Draw $\hat{y}_1^{mis}$ from $P(y_1^{mis}|\theta_1, y_1^{obs})$. Let $\hat{y}_1 = (y_1^{obs}, \hat{y}_1^{mis})$.
- Generate θ_2 from the posterior distribution $\pi(\theta_2|\hat{y}_1, y_2^{obs})$.
- Draw $\hat{y}_2^{mis}$ from $P(y_2^{mis}|\theta_2, \hat{y}_1, y_2^{obs})$. Let $\hat{y}_2 = (y_2^{obs}, \hat{y}_2^{mis})$...etc.

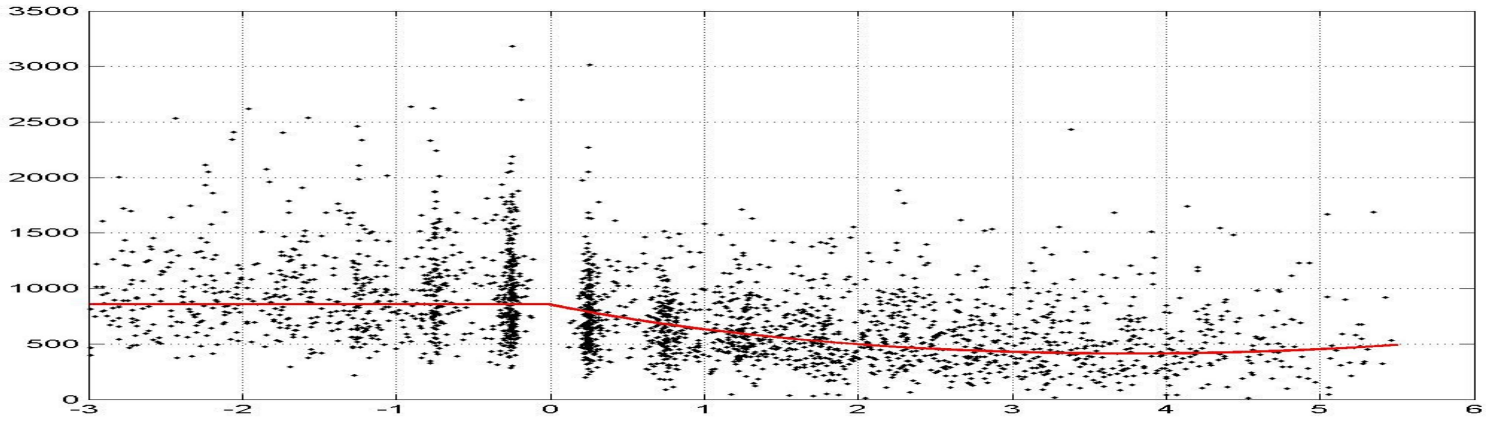
Question: www.socrative.com (calgary)

Problem

Individuals with a lower value of CD4 are at higher risk of dropping out of a population. A treatment is provided to maintain or increase CD4 in the population. The population average value of CD4 will

A. increase over time whether or not the treatment works.
B. increase over time only if the treatment works.
C. remain about the same over time.

Dealing with the apparent increase in CD4 counts



- Why does the average CD4 count increase towards the end? Are people getting healthier?
- Two ways to deal with this:
  - survivorship bias (next session)
  - use event time data (e.g. death from AIDS) and use joint likelihood. (See Struthers and McLeish (2011).)