Conditional AIC For Mixed Effects Models

Florin Vaida

Division of Biostatistics and Bioinformatics, UCSD

Vienna Workshop on Model Selection

July 24, 2008

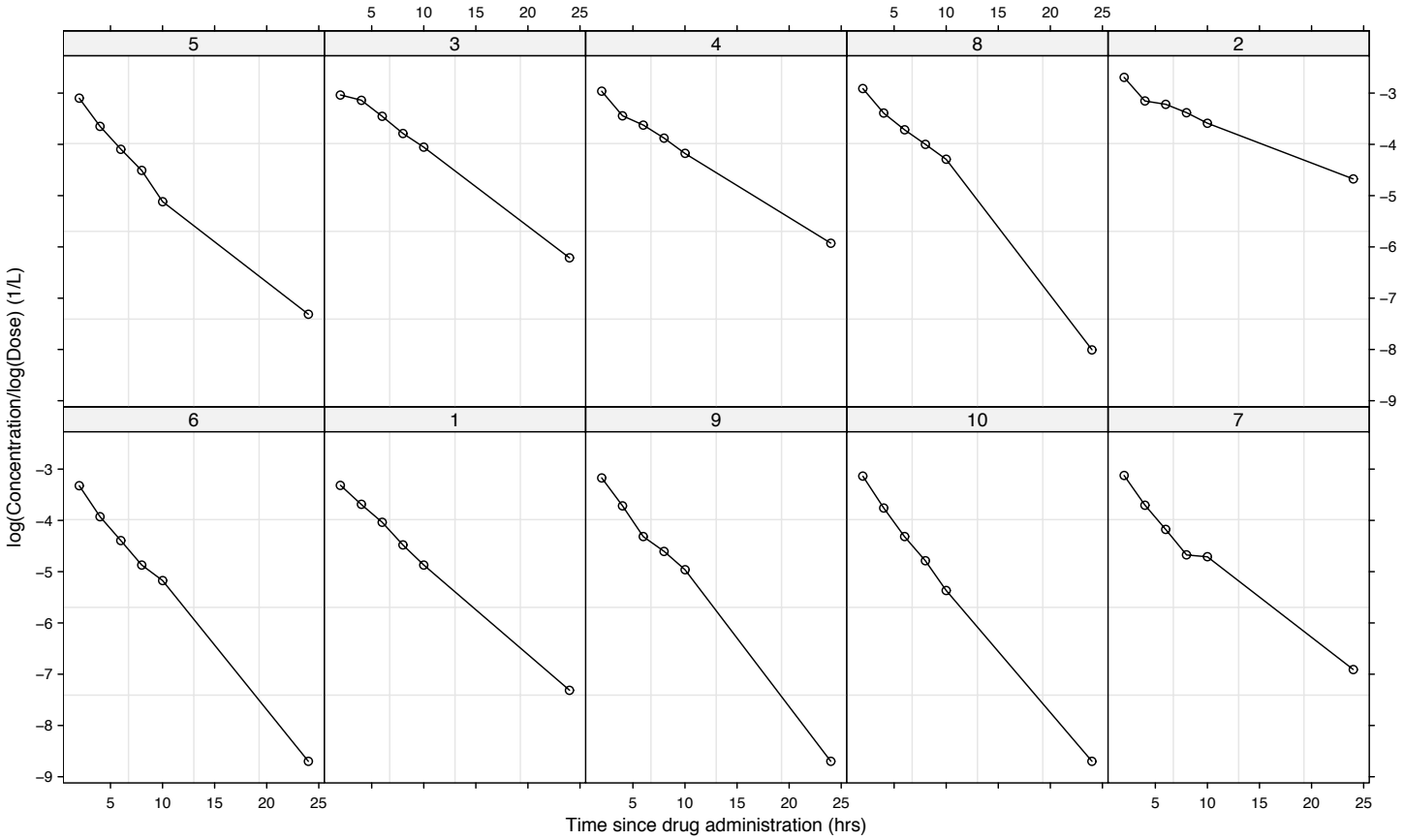
Model Selection for Mixed Effects Models

- Setting: longitudinal data
- Subjects $i = 1 \ldots m$; Observations $j = 1 \ldots n_i$
- Model: Linear, Generalized Linear, Nonlinear Mixed Effects Models (LME, GLME, NLME)
- NLME: $y_{ij} = h(\eta_{ij}) + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, \sigma^2)$
- GLME: $E(y_{ij}) = h(\eta_{ij}), \qquad y_{ij} \sim$ exponential family
- Linear predictor $\eta_{ij} = x_{ij}^\top \beta + z_{ij}^\top b_i$
- $b_i \overset{iid}{\sim} N(0, G)$ random effects;

Model Selection For Cluster Focus

- Focus of inference/prediction on subjects (clusters) in the dataset, not on new subjects (clusters)
- Model selection:
  - what covariates to include?
  - what random effects to include?
  - should I fit subject effects as fixed or random?
- Vaida & Blanchard (2005): conditional Akaike information
- For LME cAIC $= -2\text{loglik} + 2K$
- conditional loglik and effective d.f. K
- Extend such a formula to GLME and NLME?

Example: PK of Cadralazine



$$y_{ij} = \beta_{0i} + \beta_{1i} \cdot t_{ij} + e_{ij}$$

Two models:

1. Linear regression model :

   β_{0i}, β_{1i} = fixed parameters, subject-specific

2. Random effects model:

$$\beta_{0i} = \beta_0 + b_{0i}, \qquad \beta_{1i} = \beta_1 + b_{1i}$$

$$(b_{0i}, b_{1i}) \overset{iid}{\sim} N(0, G)$$

   Which model is better?

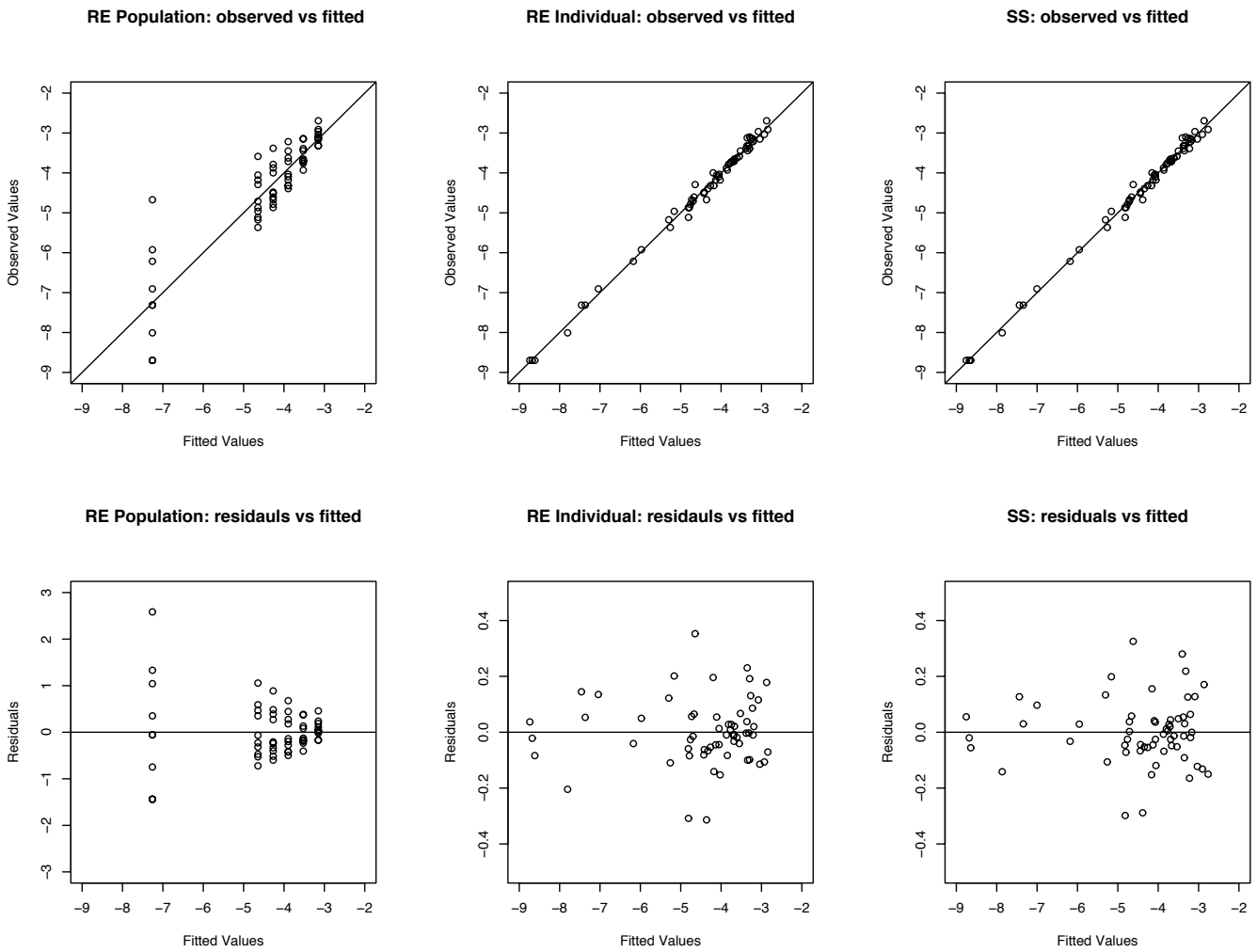

RE Population: observed vs fitted
RE Individual: observed vs fitted
SS: observed vs fitted
Observed Values
Fitted Values
RE Population: residauls vs fitted
RE Individual: residauls vs fitted
SS: residuals vs fitted
Residuals
Fitted Values

Figure 1:

Comparison of the two models

	Linear regression	Random effects
AIC	-47.1	11.0

AIC: small is beautiful

$|\Delta\text{AIC}| \; < 2 =$ similar fit $\qquad > 10 =$ overwhelming evidence

Why is AIC for linear regression model so much smaller?
Something wrong with AIC?

Effective Degrees of Freedom for LME

- LME:

$$y_i = X_i\beta + Z_i b_i + e_i, \qquad b_i \sim N(0, \sigma^2 D_0)$$

  Or in general

$$y = X\beta + Zb + e, \qquad b \sim N(0, G = \sigma^2 D), \quad e \sim N(0, \sigma^2 I)$$

- **Inference:** Maximum likelihood for β, σ^2, D (or REML)
- $\hat{b}_i = \arg\sup p(b_i|y, \hat{\beta}, \hat{G}) = \arg\sup p(y_i, b_i|\hat{\beta}, \hat{G})$ = BLUP, or Empirical Bayes
- Hodges and Sargent (2001): Effective degrees of freedom

$$\rho = \text{trace}(H) \qquad \text{where} \quad \hat{y} = Hy$$

  Count b_i as a *fraction* of a parameter

Effective Degrees of Freedom

- Henderson's "score" equations for (β, b):

$$\begin{aligned} X^\top y &= X^\top X\beta + X^\top Z b \\ Z^\top y &= Z^\top X\beta + (Z^\top Z + D^{-1})b \end{aligned}$$

- Corresponding to the *formal* linear model

$$\begin{pmatrix} y \\ 0 \end{pmatrix} = \begin{pmatrix} X & Z \\ 0 & -I \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} + \begin{pmatrix} e \\ b \end{pmatrix}$$

- $\hat{y} = Hy$,

$$H = \begin{pmatrix} X & Z \end{pmatrix} \begin{pmatrix} X^\top X & X^\top Z \\ Z^\top X & Z^\top Z + D^{-1} \end{pmatrix}^{-1} \begin{pmatrix} X & Z \end{pmatrix}^\top$$

$$
\begin{aligned}
\rho &= \text{trace}(H) \\
&= \text{trace}\left\{ \begin{pmatrix} X & Z \end{pmatrix} \begin{pmatrix} X^\top X & X^\top Z \\ Z^\top X & Z^\top Z + D^{-1} \end{pmatrix}^{-1} \begin{pmatrix} X & Z \end{pmatrix}^\top \right\} \\
&= \text{trace}\left\{ \begin{pmatrix} X^\top X & X^\top Z \\ Z^\top X & Z^\top Z + D^{-1} \end{pmatrix}^{-1} \begin{pmatrix} X^\top X & X^\top Z \\ Z^\top X & Z^\top Z \end{pmatrix} \right\}
\end{aligned}
$$

Inspired by Hastie and Tibshirani (1990) for GAM.

Counting parameters for NLME

- Our idea: take NLME

  $y_{ij} = h(\eta_{ij}) + \epsilon_{ij}, \qquad \eta_{ij} = x_{ij}^\top \beta + z_{ij}^\top b_i$

- Linearization:

$$\begin{aligned} h(\eta_{ij}) &\approx h(\hat{\eta}_{ij}) + h'(\hat{\eta}_{ij})(\eta_{ij} - \hat{\eta}_{ij}) \\ &= \{h'(\hat{\eta}_{ij})x_{ij}\}\beta + \{h'(\hat{\eta}_{ij})z_{ij}\}b_i + \text{const} \end{aligned}$$

- $w_{ij} = s_{ij}^\top \beta + t_{ij}^\top b_i + \epsilon_{ij}$
- Where $w_{ij} = y_{ij} - h(\hat{\eta}_{ij}) + h'(\hat{\eta}_{ij})\hat{\eta}_{ij}$

  $s_i = h'(\hat{\eta}_{ij})x_{ij}, \qquad t_{ij} = h'(\hat{\eta}_{ij})z_{ij}$

- Compute ρ for the linearized mixed effects model in w_{ij}
- Call ρ effective d.f. for NLME

Effective DF for GLME/NLME

- Lu, Carlin and Hodges (2007), for GLME:
- For a GLMM with linear predictor $\eta_{ij} = x_{ij}^\top \beta + z_{ij}^\top b_i$ expand loglik

$$\begin{aligned} l(\eta_{ij}) &= l(\hat{\eta}_{ij}) + l'(\hat{\eta}_{ij})(\eta_{ij} - \hat{\eta}_{ij}) + \frac{1}{2} l''(\hat{\eta}_{ij})(\eta_{ij} - \hat{\eta}_{ij})^2 \\ &= -\frac{1}{2\sigma_{ij}^2}(u_{ij} - \eta_{ij})^2 + \text{const}, \\ \sigma_{ij}^2 &= -1/E\{l''(\hat{\eta}_{ij}\} = \sigma^2/\{h'(\eta_{ij}^*)^2\} \end{aligned}$$

- $u_{ij} \approx \eta_{ij} + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, \sigma_{ij}^2)$

Formal linear model

$$\begin{pmatrix} u \\ 0 \end{pmatrix} = \begin{pmatrix} X & Z \\ 0 & -I \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} + \begin{pmatrix} e \\ b \end{pmatrix}$$

$$\mathrm{Var}(\epsilon) = \sigma^2 W^{-1} \qquad W = \mathrm{diag}[\{h'(\eta_{ij})\}^2]$$

$$\begin{aligned} \rho &= \mathrm{trace}(H) \\ &= \mathrm{trace}\left\{ \begin{pmatrix} X & Z \end{pmatrix} \begin{pmatrix} X^\top W X & X^\top W Z \\ Z^\top W X & Z^\top W Z + D^{-1} \end{pmatrix}^{-1} \begin{pmatrix} X & Z \end{pmatrix}^\top W \right\} \\ &= \mathrm{trace}\left\{ \begin{pmatrix} X^\top W X & X^\top W Z \\ Z^\top W X & Z^\top W Z + D^{-1} \end{pmatrix}^{-1} \begin{pmatrix} X^\top W X & X^\top W Z \\ Z^\top W X & Z^\top W Z \end{pmatrix} \right\} \end{aligned}$$

ρ = effective degrees of freedom of GLME/NLME

Effective DF for NLME

- **Result:** The two definitions of ρ for NLME are equivalent. They are based on different linearizations: on the scale of y_{ij} and on the scale of η_{ij}, respectively.
- For NLME/GLME ρ depends also on η_{ij} through $W = \text{diag}[\{h'(\eta_{ij})\}^2]$.
- Relevant values: true ρ^*, using "true" η^*_{ij}, W^* and estimated $\hat{\rho}$, using $\hat{\eta}_{ij}$, $\hat{W}$.
- H, ρ correspond to the score equations for (β, b):

$$\begin{aligned} X^\top W y &= X^\top W X \beta + X^\top W Z b \\ Z^\top W y &= Z^\top W X \beta + (Z^\top W Z + D^{-1}) b \end{aligned}$$

  which are the PQL equations of Breslow and Clayton (1993) for GLME.

Model selection using Akaike information

$$AI = E_{f(y)} \left\{ -2E_{f(y^*)} \log g(y^*|\hat{\theta}(y)) \right\}$$

How good is model $g(\cdot|\theta)$ at *predicting* new data y^* from model $f(\cdot)$, based on the sample y from $f(\cdot)$?

Akaike information is *not* about finding the "true model".

Estimator:

$$AI \approx \text{AIC} = -2 \log g(y|\hat{\theta}(y)) + 2K$$

$K = df = \#$ parameters in the model

Asymptotically, AIC $\approx$ unbiased for AI

Conditional AIC

- Assume truth $f(|b_0)$ and model $g(|\theta, b)$ GLMM/NLMM/LMM
- $f(y|b_0)$ = conditional distribution, b_0 true value of random eff.
- **Definition:** Conditional Akaike Info (V & B 2005)

$$\text{cAI} = -2E_{f(y,b_0)}E_{f(y^*|b_0)} \log g(y^*|\hat{\beta}(y), \hat{b}(y))$$

  where y^* iid with y, conditional on *same* b_0

- cAI is appropriate for comparing models at subject-specific level
- E.g., when chosing between *fixed* and *random* subject effects

Theorem: Conditional AIC for LME

Assume that $y \sim$ LME and model class g contains the operational model f; σ^2, D are known. Then

$$\text{cAIC} = -2 \log g(y|\hat{\beta}(y), \hat{b}(y)) + 2\rho$$

is an unbiased estimator of the conditional Akaike information.

- $g(y|\hat{\beta}, \hat{b})$ is the *conditional* distribution
- $\rho =$ effective d.f.
- Unknown σ^2, correction $= \rho + 1$ asymptotically, small sample correction available
- No correction needed for unknown D asymptotically

Back to Cadralazine data:

	Random effects model		Linear regression model
	mAIC	cAIC	AIC
Asymptotic	11.0	−44.5	−47.1
Finite-sample corrected	12.6	−42.3	−22.8
REML	–	−43.7	−40.6

Theorem: Conditional AIC for GLMM/NLMM

Assume that $y \sim$ GLMM or NLMM and σ^2, D are known. Then, under regularity conditions

$$\text{cAIC} = -2 \log g(y|\hat{\beta}, \hat{b}) + 2\rho$$

is an asymptotically unbiased estimator of the conditional Akaike information.

- Regularity conditions include $m/N \to 0$ as $N \to \infty$; $\min |Z_i| \geq c_0$ (Jiang, Jia and Chen, 2001). They ensure consistency of β, b_i.
- No results yet for unknown σ^2; correction $= \rho + 1$?

Simulation study

Bias of cAIC as an estimator of cAI $= -2$ cond loglik $+ 2(\rho + 1)$

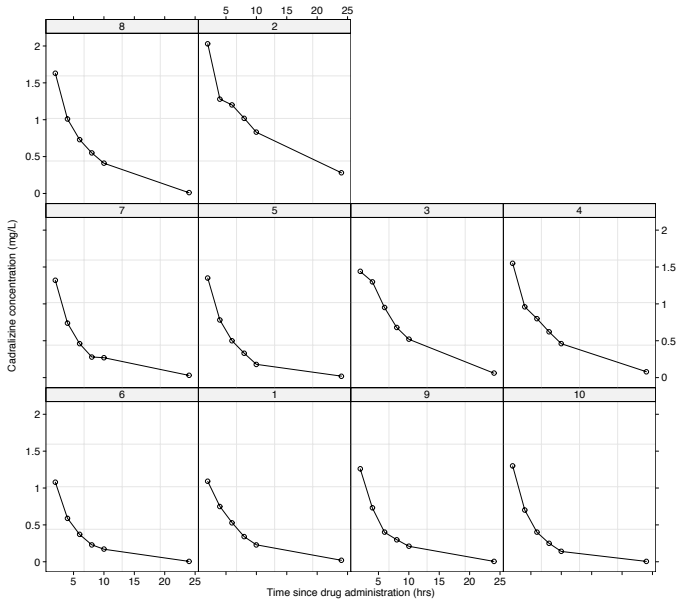
	σ		
n_i	.50	.25	.125
24	1.7	0.5	0.0
12	3.0	1.0	0.4
6	4.9	3.0	1.9
3	9.5	9.7	8.7

10 clusters, n_i observations each; One-compartment PK model

Bias reduces with increasing n_i and decreasing $\sigma/||D||$.

Bias includes effects of unknown D and model non-linearity

Cadralazine Data



Mean	random	var	ρ	cAIC
$\exp\{\beta_{1i} + \beta_2 t_{ij}\}$	β_{1i}	σ^2	10.23	-473.88
$\exp\{\beta_{1i} + \beta_{2i} t_{ij}\}$	β_{1i}, β_{2i}	σ^2	10.23	-473.88
$\exp\{\beta_{1i} + \beta_{2i} t_{ij}\}$	none	σ_i^2	30	-577.18