What is REML? Why Does it Work? And How Do We Extend it to the Generalized Linear Mixed Model?

Christopher M. Gotwalt, Elizabeth A. Claassen, Walter Stroup

JMP Division, SAS
and
Department of Statistics
University of Nebraska-Lincoln

Outline

Introduction

- REML is the standard method of estimating parameters in linear mixed models.
- It better bias and inference properties than ML.
- None of the existing derivations of REML fully explain its nice properties.
- We have a new derivation that explains precisely why REML works well.
- The method can be generalized to GLMM's where there are similar bias problems.
- We show some promising results for the one and two way simple random effects logistic model.

Notation

$$\begin{aligned} y &\sim Normal(\mathbf{X}\beta, \mathbf{V}) \\ V &= \mathbf{ZG}(\sigma^2)\mathbf{Z}^T + \mathbf{R}(\theta) \end{aligned}$$

- $\mathbf{X}$ is the $n \times p$ full rank fixed effects design matrix
- β are the fixed effects parameters
- $\mathbf{Z}$ is the $n \times n_r$ random effects design matrix
- σ^2 are the variance components and random coefficient covariances.
- θ is the other variance parameters, AR(1), etc.
- G is a linear matix function of σ^2 and is the covariance matrix of the random effects.
- For notational simplicity, $\xi = (\sigma^2, \theta)$ is all the variance parameters.

Linear Mixed Model Likelihood

$$
\begin{aligned}
L(\beta, \xi | y) &= \frac{1}{2}(y - \mathbf{X}\beta)^T \mathbf{V}^{-1}(y - \mathbf{X}\beta) + \frac{1}{2}\log|\mathbf{V}| + \frac{n}{2}\log(2\pi) \\
\tilde{\beta}(\sigma^2, \xi) &= \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}y
\end{aligned}
$$

- $L(\beta, \sigma, \theta | y)$ is the negative loglikelihood
- $\tilde{\beta}(\xi)$ is the conditional MLE of β given ξ.

Distribution of estimates

- Asymptotically we have,

$$\hat{\theta}_{ML} \approx Normal(\theta, Var(\hat{\theta}_{ML})).$$

Where,

$$Var(\hat{\theta}_{ML})^{-1} \approx \begin{pmatrix} \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X} & 0 \\ 0 & \frac{1}{2}\{tr(\mathbf{V}^{-1}\dot{\mathbf{V}}_{\xi_i}\mathbf{V}^{-1}\dot{\mathbf{V}}_{\xi_j})\}_{i,j=0}^{k} \end{pmatrix}$$

Maximum Likelihood Properties

- Very general in that ML can be applied to any variance structure
- If the sample size is large enough $\hat{\beta}_{ML}$ and $\hat{\sigma}^2_{ML}$ are consistent and asymptotically optimal
- In small samples variance ML estimates of variance components are biased downward.

Introducing REML

- The REML objective function is the classical likelihood with an extra term.

$$L_{REML}(\beta, \xi | y) = L_{ML}(\beta, \xi | y) - \frac{1}{2} \log |\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}|$$

REML as a Maximum A Posteriori Estimator

- -$\frac{1}{2}\log|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|$ is the negative log Jeffrey's prior of the fixed effects.
- REML is an uninformative prior Bayes estimator: Jeffrey's prior on fixed effects, flat prior on variance parameters.
- By itself, this says little about REML's frequentist properties.

REML as a Penalized Estimator

$$
\begin{aligned}
L_{REML}(\theta) &= L_{ML} - \frac{1}{2}\log|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}| \\
&= L_{ML} + \frac{1}{2} log|Var(\hat{\beta})|
\end{aligned}
$$

- REML is penalizing values of ξ that lead to estimates of β with higher variance.
- Interesting, but not clear how lower variance fixed effects would lead to lower bias variance parameters.

REML as a Marginal Likelihood

$$
\begin{aligned}
\exp(-L_{REML}(\sigma^2, \theta)) &= exp(-L_{ML}(\tilde{\beta}(\xi), \xi|y))|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|^{\frac{1}{2}} \\
&= \int exp(-L_{ML}(\beta, \xi|y)d\beta
\end{aligned}
$$

- Resulting marginal likelihood after putting a flat prior on the fixed effects parameters and integrating them out.
- This is the original, Patterson and Thompson (1971) motivation.
- Marginal likelihoods are useful for handling nuisance parameters.
- By itself it doesn't explain the bias reducing properties.
- Note that it is a function of $\tilde{\beta}(\xi)$. Not a unified approach to estimating all the parameters.

REML as the Likelihood of the **RE**siduals

$$P_x = \mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}$$

- REML is the likelihood of the residuals, $\varepsilon = (I - P_x)y$, after projecting off the column space of **X**.
- The likelihood function after plugging in $\tilde{\beta}(\xi)$ is singular Gaussian.
- In balanced cases, we can show analytically that the denominator are correct: e.g. the error denominator is $n - p$.
- Doesn't lead to a unified approach for estimating β and σ^2.

Where Does Bias Come From?

- Let $S(\theta) = \frac{\partial}{\partial\theta} L(\theta|y)$ be the score equation.

$$
\begin{aligned}
E_{\theta_0}(S(\theta_0)|y) &= 0 \\
\hat{\theta}_{MLE} &= S^{-1}(\mathbf{0}|y)
\end{aligned}
$$

- At θ_0, the expected value of the score is zero.
- The MLE is the inverse function of the score evaluated at zero.
- If f is a nonlinear function,

$$
E(f(y)) \neq f(E(y)).
$$

- Unbiasedness of the score equation *induces bias in the MLE when $S(\theta)$ is nonlinear in parameters.*

Firth Bias Adjusted Estimating Equations

- Firth (1993) developed a general preventative method for reducing the bias of an MLE.
- Most bias reduction techniques are corrective in nature:
  - Derive the expectation and apply an additive or multiplicative correction.
  - Do a simulation to estimate the bias and adjust (Bootstrap)
  - Use the Jacknnife.
- Firth used the asymptotic expansion of the MLE bias and identified an additive correction to the score equation that annihilates the highest order bias term.

The General Firth Adjusted Estimating Equation

- Firth added bias to the score equation to annihilate the first term of the MLE's bias.

$$\begin{aligned} S_{Firth}(\theta) &= S(\theta) + A(\theta) \\ A(\theta) &= -\frac{1}{2}\sum_j \sum_m \kappa^{j,m}(\kappa_{i,j,m} - \kappa_{i,jm}) \end{aligned}$$

- The κ's are likelihood tensors. $\kappa^{j,m}$ is the inverse Fisher information.

$$\begin{aligned} \kappa^{i,j} &= \left(E_\theta \left(\frac{\partial L(\theta)}{\partial \theta_i} \frac{\partial L(\theta)}{\partial \theta_j} \right) \right)^{-1} \\ \kappa_{i,j,m} &= E_\theta \left(\frac{\partial L(\theta)}{\partial \theta_i} \frac{\partial L(\theta)}{\partial \theta_j} \frac{\partial L(\theta)}{\partial \theta_m} \right) \\ \kappa_{i,jm} &= E_\theta \left(\frac{\partial L(\theta)}{\theta_i} \frac{\partial^2 L(\theta)}{\partial \theta_j \partial \theta_m} \right) \end{aligned}$$

Linear Mixed Model Likelihood Derivatives

- Let p, q, r be indices for fixed effect parameters, s, t, u are variance indices.
- x_p are fixed effect design matrix columns.
- Let $\varepsilon = y - \mathbf{X}\beta$ be the residual.

$$\frac{\partial L}{\partial \beta_p} = -x_p^T \mathbf{V}^{-1} \varepsilon \qquad \frac{\partial L}{\partial \xi_s} = -\frac{1}{2}\varepsilon^T \mathbf{V}^{-1} \dot{\mathbf{V}}_{\xi_s} \mathbf{V}^{-1} \varepsilon + \frac{1}{2} tr(\mathbf{V}^{-1} \dot{\mathbf{V}}_{\xi_s})$$

$$\frac{\partial^2 L}{\partial \beta_p \partial \beta_q} = x_p^T \mathbf{V}^{-1} x_q \qquad \frac{\partial^2 L}{\partial \beta_q \partial \xi_s} = -x_p^T \mathbf{V}^{-1} \dot{\mathbf{V}}_{\xi_s} \mathbf{V}^{-1} \varepsilon$$

$$\begin{aligned}\frac{\partial^2 L}{\partial \xi_s \partial \xi_t} &= \varepsilon^T \mathbf{V}^{-1} (\dot{\mathbf{V}}_{\xi_s} \mathbf{V}^{-1} \dot{\mathbf{V}}_{\xi_t} - \frac{1}{2} \ddot{\mathbf{V}}_{\xi_s, \xi_t}) \mathbf{V}^{-1} \varepsilon \\ &\quad + \frac{1}{2} tr(\mathbf{V}^{-1} \ddot{\mathbf{V}}_{\xi_s, \xi_t} - \mathbf{V}^{-1} \dot{\mathbf{V}}_{\xi_s} \mathbf{V}^{-1} \dot{\mathbf{V}}_{\xi_t})\end{aligned}$$

Linear and Quadratic Form Product Expectations

- Let $\varepsilon \sim Gaussian(0, V)$.
- All odd order moments are zero:

$$
\begin{aligned}
E(a^T\varepsilon) &= 0 \\
E((\varepsilon^T A\varepsilon)(b^T\varepsilon)) &= 0 \\
E((\varepsilon^T A\varepsilon)(\varepsilon^T B\varepsilon)(c^T\varepsilon)) &= 0
\end{aligned}
$$

- The even order moments:

$$
\begin{aligned}
E(\varepsilon^T A\varepsilon) &= tr(AV) \\
E((\varepsilon^T A\varepsilon)(\varepsilon^T B\varepsilon)) &= tr(AV)tr(BV) + 2tr(AVBV) \\
E((\varepsilon^T A\varepsilon)(\varepsilon^T B\varepsilon)(\varepsilon^T C\varepsilon)) &= tr(AV)tr(BV)tr(CV) + 8tr(AVBVCV) \\
&\quad +2tr(AVBV)tr(CV) + 2tr(AVCV)tr(BV) \\
&\quad +2tr(BVCV)tr(AV)
\end{aligned}
$$

Expectations of Products of Likelihood Derivatives

- Let p, q, r be indices for fixed effect parameters, s, t, u are variance indices.
- x_p are fixed effect design matrix columns.

$$\kappa_{p,q} = x_p^T \mathbf{V}^{-1} x_q \qquad \kappa_{p,s} = 0$$

$$\kappa_{s,t} = \frac{1}{2} tr(\mathbf{V}^{-1}\dot{\mathbf{V}}_{\xi_s}\mathbf{V}^{-1}\dot{\mathbf{V}}_{\xi_t})$$

$$\begin{aligned}
\kappa_{p,q,r} &= 0 & \kappa_{p,qr} &= 0 \\
\kappa_{p,q,s} &= -x_p^T \mathbf{V}^{-1}\dot{\mathbf{V}}_{\xi_s}\mathbf{V}^{-1} x_q & \kappa_{p,qs} &= -x_p^T \mathbf{V}^{-1}\dot{\mathbf{V}}_{\xi_s}\mathbf{V}^{-1} x_q \\
\kappa_{p,s,t} &= 0 & \kappa_{p,st} &= 0 \\
\kappa_{s,p,q} &= x_p^T \mathbf{V}^{-1}\dot{\mathbf{V}}_{\xi_s}\mathbf{V}^{-1} x_q & \kappa_{s,pq} &= 0 \\
\kappa_{s,p,t} &= 0 & \kappa_{s,pt} &= 0 \\
\kappa_{s,t,u} &= -tr(\mathbf{V}^{-1}\dot{\mathbf{V}}_{\xi_s}\mathbf{V}^{-1}\dot{\mathbf{V}}_{\xi_t}\mathbf{V}^{-1}\dot{\mathbf{V}}_{\xi_u}) & \kappa_{s,tu} &= -tr(\mathbf{V}^{-1}\dot{\mathbf{V}}_{\xi_s}\mathbf{V}^{-1}\dot{\mathbf{V}}_{\xi_t}\mathbf{V}^{-1}\dot{\mathbf{V}}_{\xi_u}) \\
& & &\quad +\frac{1}{2} tr(\mathbf{V}^{-1}\dot{\mathbf{V}}_{\xi_s}\mathbf{V}^{-1}\ddot{\mathbf{V}}_{\xi_t,\xi_u})
\end{aligned}$$

Firth and the Linear Mixed Model

- Expected Fisher Information Matrix:

$$\mathbf{F} = \begin{pmatrix} \mathbf{X}\mathbf{V}^{-1}\mathbf{X} & 0 \\ 0 & \{\frac{1}{2}tr(\mathbf{V}^{-1}\dot{\mathbf{V}}_{\xi_t}\mathbf{V}^{-1}\dot{\mathbf{V}}_{\xi_u})\}_{t,u} \end{pmatrix}$$

- Fixed effects score equation adjustment:

$$A_\beta = \frac{1}{2}tr\left(\mathbf{F}^{-1}\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}\right) = 0$$

- Variance parameter score equation adjustment:

$$A_{\xi_s} = \frac{1}{2}tr\left(\mathbf{F}^{-1}\begin{pmatrix} \mathbf{X}^T\mathbf{V}^{-1}\dot{\mathbf{V}}_{\xi_s}\mathbf{V}^{-1}\mathbf{X} & 0 \\ 0 & \{\frac{1}{2}tr(\mathbf{V}^{-1}\dot{\mathbf{V}}_{\xi_s}\mathbf{V}^{-1}\ddot{\mathbf{V}}_{\xi_t,\xi_u})\}_{t,u} \end{pmatrix}\right)$$

Linear Mixed Models Firth Observations

$$
\begin{aligned}
A_\beta(\beta,\xi) &= 0 \\
A_{\xi_k}(\beta,\xi) &= -\frac{1}{2} tr((\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{V}^{-1}\dot{\mathbf{V}}_{\xi_k}\mathbf{V}^{-1}\mathbf{X})) \\
&\quad +\frac{1}{2} tr(\mathbf{F}_{\xi\xi}^{-1}\mathbf{B}_s)
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{F}_{\xi\xi} &= \frac{1}{2}\left\{ tr(\mathbf{V}^{-1}\dot{\mathbf{V}}_{\xi_t}\mathbf{V}^{-1}\dot{\mathbf{V}}_{\xi_u})\right\}_{tu} \\
\mathbf{B}_k &= \frac{1}{2}\left\{ tr(\mathbf{V}^{-1}\dot{\mathbf{V}}_{\xi_s}\mathbf{V}^{-1}\ddot{\mathbf{V}}_{\xi_t,\xi_u})\right\}_{tu}
\end{aligned}
$$

- The first A_{ξ_s} term is the derivative of the REML penalty!
- If $\mathbf{V}$ is linear in ξ, the second A_{ξ_s} term is zero!
- REML *is* a Firth adjusted likelihood estimator for pure variance component/random coefficient models.

Summary

- REML estimates are second order unbiased, *for linear variance structures*.
- This is a general and rigorous explanation that justifies the use of REML for LMMs.
- Existing GLMM methods underestimate variance components causing higher than nominal error rates
- Firth leads to *lower bias and variance*(!!) in logistic regression. Heinz and Schemper (2002)
- Suggests Firth is promising for GLMM's

Applying Firth to GLMMs

- Develop Firth for one and two way simple random effects binary logistic model
- Compare bias to ML
- Investigate Type I error rate and power

Generalized Linear Mixed Model Notation

$$\begin{aligned}
\mathbf{y}|\mathbf{x},\mathbf{b} &\sim f(\mu) \\
\mu &= g^{-1}(\boldsymbol{\eta}) \\
\boldsymbol{\eta} &= \mathbf{X}\beta + \mathbf{Z}\mathbf{b} \\
\mathbf{b} &\sim N(0,\mathbf{G})
\end{aligned}$$

- $\mathbf{y}|\mathbf{x},\mathbf{b}$ is f distributed with link function g.

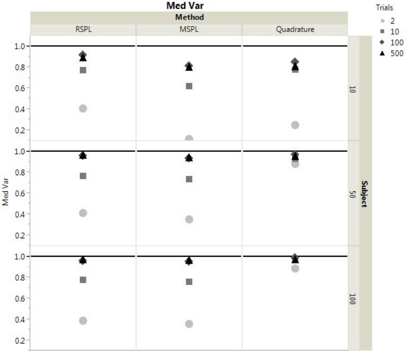
Existing Methods

- Breslow and Clayton (1993) and Wolfinger and O'Connell (1993)
- Use linearization of GLMM for RSPL (REML) & MSPL (ML)
- Pinheiro and Bates (1995)
- Uses integral approximation techniques such as Laplace or adaptive Gaussian quadrature

Previous Studies

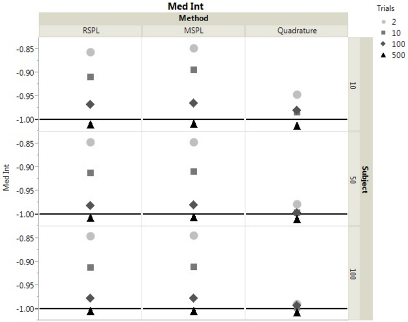
- McCulloch (1997)
  - Compared several maximum likelihood estimation methods for $\ln(\frac{\pi_{ij}}{1-\pi_{ij}}) = \beta x_{ij} + u_j$ with $u_j \sim N(0, \sigma^2)$ and y_{ij} iid $Bernoulli(\pi_i)$.
  - Methods consistently underestimated σ^2.
- Stroup (2013)
  - Compared modeling & estimation options for a beta-binomial with Gaussian random blocks scenario for Type I error and power.
  - With quadrature (the ML estimation method) the rejection rate exceeded 15% for a nominal 5%.

- Model: $\ln(\frac{\pi_i}{1-\pi_i}) = \eta + b_i$ with $y_{ij}|b_i$ iid $Bernoulli(\pi_i)$, $b_i \sim N(0,1)$ and $\eta = -1$



Random Intercept Binary Response Model

- How the variance estimate affects the intercept estimate



Which Method?

		nTrials	
		Small	Large
nSubjects	Small	??	PL
	Large	Quadrature	Either

- Quadrature implies biased variance components
- Can we improve?

Implementation Outline

- Implement MLE first. Compare against GLMMIX.
- Develop programs that implement the likelihood, gradient, and Hessian for the oneway random logit model.
  - Program for solving for the BLUPS (necessary for Laplace approximation).
  - Use Hermite quadrature assisted Laplace approximation to integrate out BLUPs.
- Early numerical experiments showed that analytic derivatives were a must.
- Firth correction consists of products of gradient & Hessian integrating out the data **y**.

Maximum Likelihood

$\ln(\frac{\pi_i}{1-\pi_i}) = \eta + b_i$ with $y_{ij}|b_i$ iid $Bernoulli(\pi_i)$

$$\begin{aligned}\ell(\boldsymbol{\theta}|\mathbf{y},\mathbf{z}) &= \sum_{i=1}^{r} \ln\binom{n_i}{y_i} + y_i(\eta+\sigma z_i) - n_i(\eta+\sigma z_i) \\ &\quad -n_i \ln\left(1+e^{-(\eta+\sigma z_i)}\right) - \ln\sqrt{2\pi} - \frac{z_i^2}{2}\end{aligned}$$

- Integrate out BLUPs with Gauss-Hermite quadrature corrected Laplace approximation

$$\ell(\boldsymbol{\theta}|\mathbf{y}) = \frac{\sqrt{2\pi}e^{-\ell_i(\hat{\mathbf{z}})}}{\sqrt{\ell''(\hat{\mathbf{z}}_\mathbf{i})}} \sum_i e^{-\ell_i\left(\frac{x_i}{\sqrt{\ell_i''(\hat{\mathbf{z}}_\mathbf{i})}}+\hat{\mathbf{z}}_\mathbf{i}\right)} e^{\ell_i(\hat{\mathbf{z}}_\mathbf{i})} e^{\frac{x_i^2}{2}} w_i$$

- $\hat{z}$ are the BLUPs.
- x_i and w_i are the quadrature weights and abscissas.
- Newton-Raphson used to optimize likelihood, and obtain BLUPs.

Firth Translated into Matrix Notation

- Firth Estimating Equation
  - $\mathbf{S}^*(\theta) = \mathbf{S}(\theta) + \mathbf{A}(\theta)$
- Firth Adjustment
  - $\mathbf{A}_{\theta j} = -\frac{1}{2}\,\mathrm{tr}\left(\mathbf{F}^{-1}\mathrm{E}_y\left[\mathbf{S}_{\theta j}\left(\mathbf{H} - \mathbf{S}\mathbf{S}^T\right)\right]\right)$

$$
\begin{aligned}
\mathbf{S} &= \frac{\partial^2 \ell(y|\theta)}{\partial\theta} \\
\mathbf{H} &= \frac{\partial^2 \ell(y|\theta)}{\partial\theta\partial\theta^T} \\
\mathbf{F} &= E_y(\mathbf{H})
\end{aligned}
$$

Analytic Gradient and Hessian

$$\mathbf{S} = \frac{\partial L(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}} = -(2\pi)^{1/2} \sum_i w_i \left(\frac{\partial}{\partial \boldsymbol{\theta}} D_i(\hat{z}_i(\boldsymbol{\theta}), \boldsymbol{\theta}) \right) e^{-D_i(\hat{z}_i(\boldsymbol{\theta}),\boldsymbol{\theta})} e^{\frac{x_i^2}{2}}$$

$$\mathbf{H} = \frac{\partial^2 L(\boldsymbol{\theta}|\mathbf{y})}{\partial \theta_j \partial \theta_k} = (2\pi)^{1/2} \sum_i w_i \left(\frac{\partial D_i}{\partial \theta_j} \frac{\partial D_i}{\partial \theta_k} - \frac{\partial^2 D_i}{\partial \theta_j \partial \theta_k} \right) e^{-D_i(\hat{z}_i(\boldsymbol{\theta}),\boldsymbol{\theta})} e^{\frac{x_i^2}{2}}$$

where

$$D_i(\hat{z}_i(\boldsymbol{\theta}), \boldsymbol{\theta}) = \ell_i \left(\hat{z}_i(\boldsymbol{\theta}) + \frac{x_i}{\sqrt{\ell''(\hat{z}_i(\boldsymbol{\theta}), \boldsymbol{\theta})}}, \boldsymbol{\theta} \right) + \frac{1}{2} \ln \ell_i''(\hat{z}_i(\boldsymbol{\theta}), \boldsymbol{\theta})$$

Firth Implementation Details

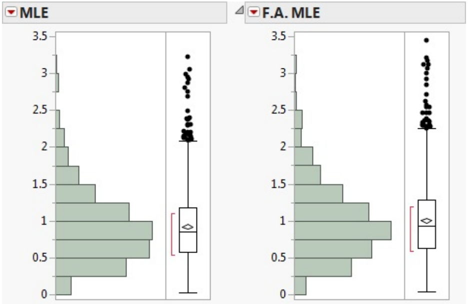
- Used GLMMIX ML starting values
- Laplace-17pt Hermite quadrature to integrate random effects
- Summed over support of binomial distribution for E_y, with convergence check in the tails for speed.

GLMM Firth is an estimating equation **NOT** a likelihood!

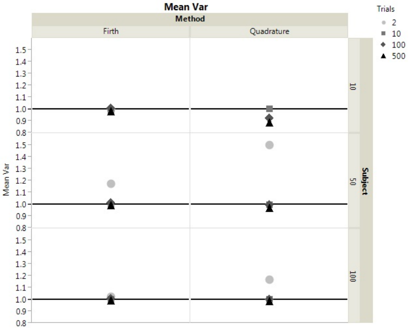
- Solve estimating equation with Broyden's quasi-Newton method
- Observed information to for standard errors and inference.
    - Opportunity for improvement here.

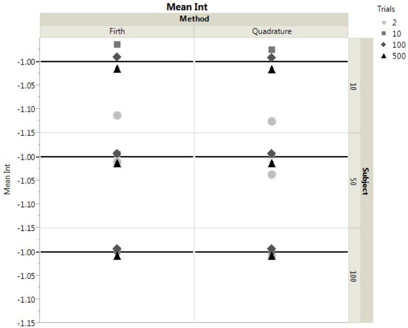
Firth Adjustment Balanced Data Results

- Should have similar bias reduction with less variance of the estimate.
- 1000 simulated experiments, all converged.
- MLE mean: .92; standard deviation: .493
- Firth MLE mean: 1.00; standard deviation: .526



Firth Simulation Results





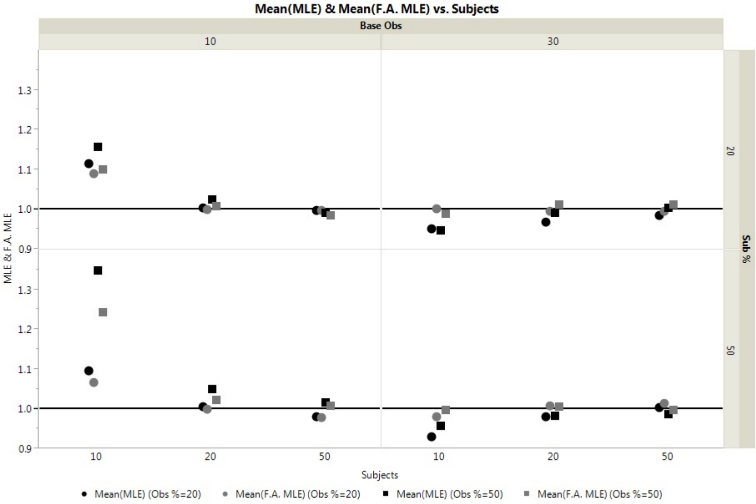
Balanced Data Implementation Observations

- Data are balanced, so each subject contributes equally
- Calculate expectations for one subject
- Multiply by number of subjects for full expectation - large time savings!
- Quadrature essential
- Early simulations showed that Laplace by itself is not enough.
- Firth requires a real likelihood to work.

Unbalanced Data Simulation Study Plan

- 24 scenarios
- 10, 20 or 50 random subjects
- 10 or 30 planned observations per subject
- 20% or 50% subjects with observations missing
- 20% or 50% missing observations within a subject

Unbalanced Data Results



Two-treatment CRD

- Build a two-treatment solver from two one-treatment likelihoods.
- $L\left(\left(\begin{array}{c}\eta_1\\ \eta_2\\ \sigma^2\end{array}\right) \mid \left(\begin{array}{c}\mathbf{y}_1\\ \mathbf{y}_2\end{array}\right)\right) = L_1\left(\left(\begin{array}{c}\eta_1\\ \sigma^2\end{array}\right)|\mathbf{y}_1\right) + L_2\left(\left(\begin{array}{c}\eta_2\\ \sigma^2\end{array}\right)|\mathbf{y}_2\right)$
- $\left(\begin{array}{c}\frac{\partial L_1}{\partial \eta_1}\\ \frac{\partial L_2}{\partial \eta_2}\\ \frac{\partial L_1}{\partial \sigma^2} + \frac{\partial L_2}{\partial \sigma^2}\end{array}\right)$
- $\left(\begin{array}{ccc}\frac{\partial^2 L_1}{\partial \eta_1 \partial \eta_1} & 0 & \frac{\partial^2 L_1}{\partial \eta_1 \partial \sigma^2}\\ 0 & \frac{\partial^2 L_2}{\partial \eta_2 \partial \eta_2} & \frac{\partial^2 L_2}{\partial \eta_2 \partial \sigma^2}\\ \frac{\partial^2 L_1}{\partial \sigma^2 \partial \eta_1} & \frac{\partial^2 L_2}{\partial \sigma^2 \partial \eta_2} & \frac{\partial^2 L_1}{\partial \sigma^2 \partial \sigma^2} + \frac{\partial^2 L_2}{\partial \sigma^2 \partial \sigma^2}\end{array}\right)$

Simulation Study Design

- 5 subjects per treatment, 100 observations per subject; 1000 simulated experiments
- $\ln\left(\frac{p_i}{1-p_i}\right) = -1$ or $p_i \approx .27$
- $\ln\left(\frac{p_1}{1-p_1}\right) = -1$ or $p_1 \approx .27$ and $\ln\left(\frac{p_2}{1-p_2}\right) = -3.125$ or $p_2 \approx .04$
- In theory 80% power to detect difference

Type I Error

- Testing difference of log odds between treatments
- Quadrature MLE rejection rate 7.7%
- Firth-adjusted MLE rejection rate 5.8%

	$\hat{\eta}_{1,MLE}$	$\hat{\eta}_{2,MLE}$	$\hat{\eta}_{1,FAMLE}$	$\hat{\eta}_{2,FAMLE}$
Std. Dev. of Sampling Distribution	0.4629	0.4636	0.4639	0.4646
Average Standard Error	0.3990	0.3990	0.4391	0.4391
Estimated coverage probability	93.1%	91.5%	94.7%	94.5%

Power

- Firth-adjusted MLE rejection rate testing difference of log odds is 78.4%

	$\hat{\eta}_{1,MLE}$	$\hat{\eta}_{2,MLE}$	$\hat{\eta}_{1,FAMLE}$	$\hat{\eta}_{2,FAMLE}$
Std. Dev. of Sampling Distribution	0.5237	0.4637	0.5187	0.4633
Average Standard Error	0.4559	0.3905	0.4985	0.4356
Estimated coverage probability	94.7%	91.2%	96.9%	94.3%

Conclusions

- Early results are very promising!
- Firth estimator demonstrates lower bias, better Type I error rates in almost all situations.
- Power was as high as we had hoped as well.
- There is still much work to be done!

Firth could be way that all GLMMs should be analyzed as REML is the standard for LMMs

Future Work

- Extension to RCBD and other multiple treatments per block designs
- Other distributions
    - Poisson and negative binomial for counts
    - Gamma for time-to-event
    - Models with censored observations
    - Beta for continuous proportions
- More general covariance structures
    - Random coefficient models
    - More that two levels of experimental units, e.g. split split plot