An assessment of current software:

parameter estimate accuracy for Generalized Linear Mixed Models with binary outcome data

Ty Stanford[†] and Patty Solomon[†]

September 20, 2015

[†]School of Mathematical Sciences
The University of Adelaide

THE UNIVERSITY of ADELAIDE

Australian Government
Australian Research Council

Motivation

Research question:

- Identify ICUs with unusual performance

Data:

- Hierarchical

Model:

- Generalized Linear Mixed Model (GLMM)
- Binary response of mortality

ML is used to obtained the parameter estimates in GLMMs

- The (profiled) log-likehood function is approximated to a specified degree

Can we rely on the estimates produced?

Motivating data: ANZICS Adult Patient Database

mortality	hosp/icu	patid	APACHEIII	covariates
0	1	1	49	$x'_{1,1}$
1	1	2	88	$x'_{1,2}$
⋮	⋮	⋮	⋮	⋮
0	1	n_1	59	x'_{1,n_1}
1	2	1	91	$x'_{2,1}$
0	2	2	45	$x'_{2,2}$
⋮	⋮	⋮	⋮	⋮
0	2	n_2	94	x'_{1,n_2}
⋮	⋮	⋮	⋮	⋮
1	m	1	49	$x'_{m,1}$
1	m	2	147	$x'_{m,2}$
⋮	⋮	⋮	⋮	⋮
0	m	n_m	57	x'_{1,n_m}

2-level GLMM

- ICUs $i = 1, \ldots, m$
- Patients $j = 1, \ldots, n_i$ within ICU i
- Fixed effects $\boldsymbol{x}_{ij}$ (including APACHEIII)
- Random effects (population of ICUs)
  - Random intercept $u_{i0} \sim N(0, \tau_0)$
  - Random APACHEIII slope $u_{i1} \sim N(0, \tau_1)$
  - $\text{cor}(u_{i0}, u_{i1}) = \rho$
- Mortality $y_{ij} \in \{0, 1\}$
  - $(y_{ij} | \boldsymbol{x}_{ij}, u_{i0}, u_{i1}) \sim \text{Bernoulli}\,(\eta_{ij})$

$$\text{logit}\,(\eta_{ij}) = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \boldsymbol{z}_{ij}^T \boldsymbol{u}_i$$

Model of Zhang et al. (2011)

The GLMM:

$$\operatorname{logit}(\eta_{ij}) = \beta_0 + u_{i0} + x_{ij}(\beta_1 + u_{i1}),$$

with $i = 1, 2, \ldots, 500$ and $j = 1, 2, 3$.

Data were generated using:

- $\beta_0 = \beta_1 = 1$
- $u_{i0} \sim N\left(0, \tau_0^2 = 4\right)$
- $u_{i1} \sim N\left(0, \tau_1^2 = 4\right)$
- $\operatorname{cor}(u_{i0}, u_{i1}) = \rho = 0.25$
- $x_{ij} \sim N(0, 1)$
- $(y_{ij} | x_{ij}, \boldsymbol{u}_i) \sim \text{Bernoulli}(\eta_{ij})$

The model parameters are:

$$\boldsymbol{\lambda} = \{\boldsymbol{\theta}, \boldsymbol{u}\} = \{\{\beta_0, \beta_1, \tau_0, \tau_1, \rho\}, \{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_m\}\}$$

Profiled density

As the random effects $\boldsymbol{u}_i$ are nuisance parameters,[1] the profiled density can be used. The profiled density:

$$
\begin{aligned}
&\ell_p(\boldsymbol{\theta}; \boldsymbol{y}) \\
&= \ln \int_{\boldsymbol{U}} L(\boldsymbol{\theta}, \boldsymbol{u}; \boldsymbol{y})\ d\boldsymbol{U} \\
&= -m \ln(2\pi\tau_0\tau_1) + \sum_{i=1}^{m} y_{i\cdot}(\beta_0 + x_i\beta_1) + \\
&\ln \int \cdots \int \frac{\exp\left\{\sum_{i=1}^{m} y_{i\cdot}(u_{i0} + x_i u_{i1}) - \frac{u_{i0}^2}{2\tau_0^2} - \frac{u_{i1}^2}{2\tau_1^2}\right\}}{\prod_{i=1}^{m}\left[1 + \exp\{\beta_0 + u_{i0} + x_i(\beta_1 + u_{i1})\}\right]^{n_i}}\, d\boldsymbol{u}_1 \ldots d\boldsymbol{u}_m
\end{aligned}
$$

[1] ML relies on the assumption the number of model parameters is invariant to the number of observations. $\boldsymbol{u}_i$ are nuisance parameters as more hospitals/groups means the number of parameters increase. By marginalising these parameters, the ML assumption of fixed number of parameters (to be optimised) holds.

Estimation of coefficients

The optimisation problem ($\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ell_p$) using the profiled log likelihood has two parts

$$\begin{aligned} \ell_p(\boldsymbol{\theta}; \boldsymbol{y}) &= a(\boldsymbol{\theta}) + \ln \int \cdots \int g(\boldsymbol{\theta}, \boldsymbol{u})\, d\boldsymbol{u} \\ &\approx a(\boldsymbol{\theta}) + \ln b(\boldsymbol{\theta}) \end{aligned}$$

- ***Estimate*** $b(\boldsymbol{\theta})$, i.e. create approximate function of integral term using Laplace or aGHQ
- $a(\boldsymbol{\theta}) + \ln b(\boldsymbol{\theta})$ can then be ***optimised*** ($\arg\max_{\boldsymbol{\theta}}$)

Many optimisation algorithms can be employed

- Iterate until a minmum change threshold is met

Gauss-Hermite Quadrature (GHQ)

A univariate integral:

$$\int_{-\infty}^{\infty} g\left(x\right) dx \approx \sum_{q=1}^{Q} w_q g\left(x_q\right)$$

- Q is what is referred to as the number of quadrature points
- x_q and w_q are the nodes and weights
  - the x_q are the roots of the Q^{th}-order Hermite polynomial $H_Q\left(x\right)$
  - the w_q are values of the $\left(Q-1\right)^{\text{th}}$-order Hermite polynomial at the x_q: $H_{Q-1}\left(x_q\right)$
- Assumes, amongst other things, that the distribution is centred around zero

Adaptive GHQ (aGHQ)

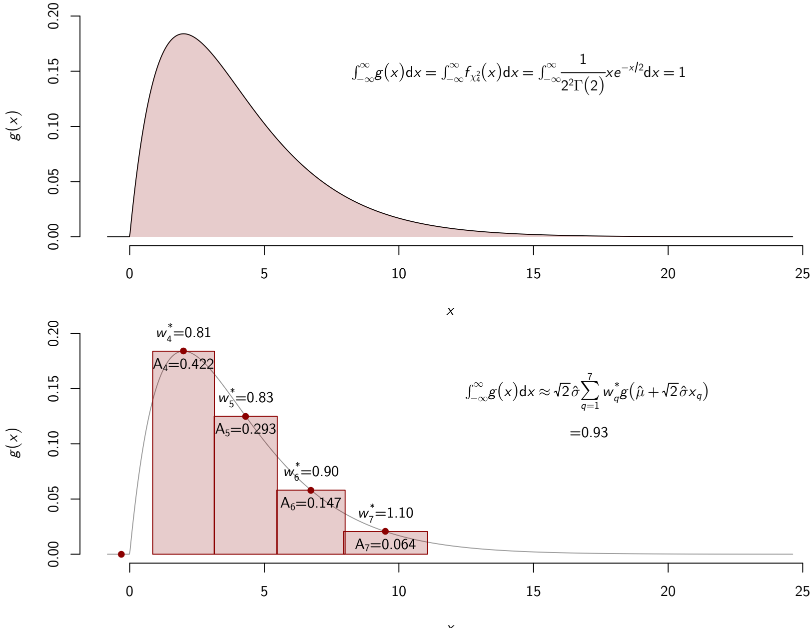
aGHQ simply means standard normal variate type tranformation takes place which alters the formula (Liu and Pierce, 1992):

$$\int_{-\infty}^{\infty} g\left(x\right) dx \approx \sqrt{2}\hat{\sigma} \sum_{q=1}^{Q} e^{-x_q^2} w_q g\left(\hat{\mu} + \sqrt{2}\hat{\sigma} x_q\right)$$

where

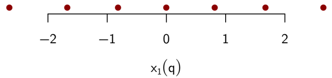
- $\hat{\mu} = \arg\max_x g\left(x\right)$, and
- $\hat{\sigma}$ is the Fisher Information at $\hat{\mu}$: $\hat{\sigma} = \frac{1}{\sqrt{-\frac{\partial^2}{\partial x^2} \ln g(x)\Big|_{x=\hat{\mu}}}}$

aGHQ example ($Q = 7$)

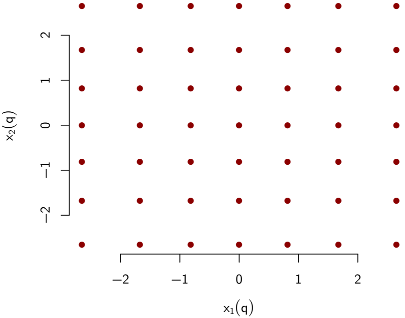


Multidimensional aGHQ grids

1-D ($Q = 7$):

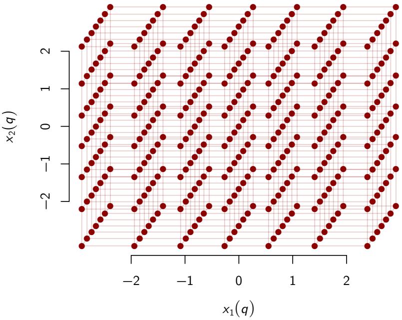


2-D ($Q = 7$):



Multidimensional aGHQ grids

3-D ($Q = 7$):



p-D:
The total number of quadrature points in the p-dimensional approximation is

$$Q^p$$

Laplace $\equiv (Q = 1)$

First, note that the Fisher information can be rearranged:

$$\hat{\sigma} = \frac{1}{\sqrt{-\frac{\partial^2}{\partial x^2} \ln g(x)\Big|_{x=\hat{\mu}}}} \Leftrightarrow \frac{\partial^2}{\partial x^2} \ln g(x)\Big|_{x=\hat{\mu}} = -\frac{1}{\hat{\sigma}^2}$$

If we take a 2^{nd}-order Taylor series of $g_*(x) = \ln g(x)$ around $\hat{\mu}$:

$$\begin{aligned} \ln g(x) = g_*(x) &\approx g_*(\hat{\mu}) + (x - \hat{\mu})\, g_*'(\hat{\mu}) + \frac{1}{2}(x - \hat{\mu})^2 g_*''(\hat{\mu}) \\ &\approx \ln g(\hat{\mu}) - \frac{(x - \hat{\mu})^2}{2\hat{\sigma}^2} \end{aligned}$$

$$\therefore \int_{-\infty}^{\infty} g(x)\, dx = \int_{-\infty}^{\infty} e^{\ln g(x)} dx \approx g(\hat{\mu}) \int_{-\infty}^{\infty} e^{-\frac{(x-\hat{\mu})^2}{2\hat{\sigma}^2}} dx \approx g(\hat{\mu}) \sqrt{2\pi}\hat{\sigma}$$

- Equivalent to a $Q = 1$ aGHQ ($x_1 = 0$ and $w_1 = \sqrt{\pi}$)

Penalised quasi-likelihood (PQL)

- Taylor series expansion of the likelihood function
- Biased, especially when Bernoulli trials low samples per cluster[2]
- Avoid using this method[3]

[2] W. W. Stroup. *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications.* Chapman & Hall, 2012.

[3] C. E. McCulloch, S. R. Searle, J. M. Neuhaus. *Generalized, Linear, and Mixed Models, 2nd Edition.* John Wiley & Sons, 2008.

Survey of available software

Software/package	Routine/function
Stata	`xtmelogit`
SAS	`NLMIXED`
SAS	`GLIMMIX`
ADMB	`ADMB-RE`
R/lme4	`glmer`
R/glmmADMB	`glmmADMB`
S-Plus	`nlme`
Matlab	`fitglme`
SPSS	`GENLINMIXED`

Notable absentees

Software	Routine/package	Comment
Julia	`GLM` or `MixedModels`	Neither seem to fit GLMMs as of yet
Python	`StatsModels`	Has linear mixed effect models and GEE GLMS but no GLMMs as of yet

Optimisers

There are 3 general choices:

- Hessian second order partial derivatives (e.g. Newton-Raphson, trust region)
- Gradient first order partial derivatives (e.g. Quasi-Newton)
- Non-gradient based (e.g. Nelder-Mead simplex)

Second order methods

- Requires more memory
- Non-positive-definite errors

Non-derivative methods

- More iterations required, less computation per iteration

Survey of available software

Function	Integral estimation	Default optimiser	Optimiser ∂ order
`xtmelogit`	aGHQ	Newton-Raphson	2
`NLMIXED`	aGHQ	Dual Quasi-Newton	1
`GLIMMIX`	aGHQ	Dual Quasi-Newton	1
`ADMB-RE`	aGHQ	Quasi-Newton	2
`glmer`	Laplace†	BOBYQA/Nelder-Mead	0
`glmmADMB`	Laplace	[ADMB's optimiser]	
`nlme`	Laplace	Newton-type	2
`fitglme`	Laplace	Quasi-Newton	1
`GENLINMIXED`	PQL?	???	

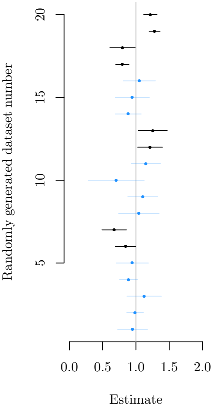
†aGHQ available for random intercept only models

Finally results

Firstly, a look at the fixed slope estimation, $\hat{\beta}_1$.

Results shown as 'spine plots'

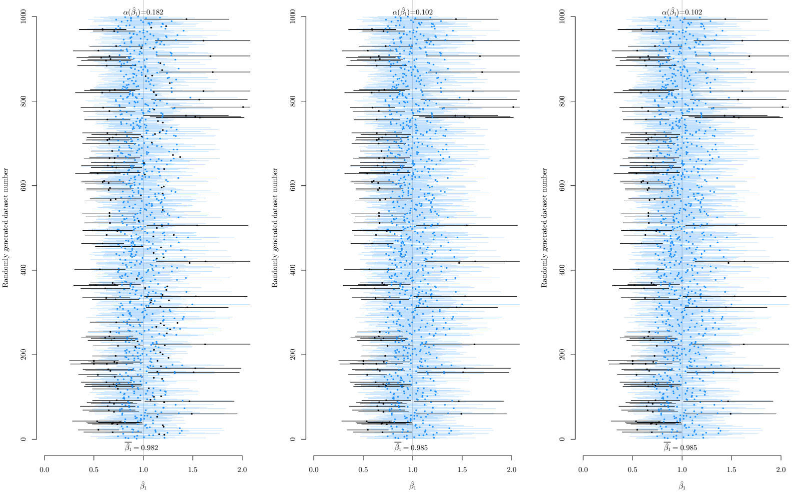
- Zhang et al. (2011) datasets were randomly generated 1000 times
- Each dataset's 95% CI is a horizontal line
- Spine is the true value which should be covered by 95%
- Horizontal lines that do not cover the true value are blackened



A table of summary statistics, the α error and the average value.

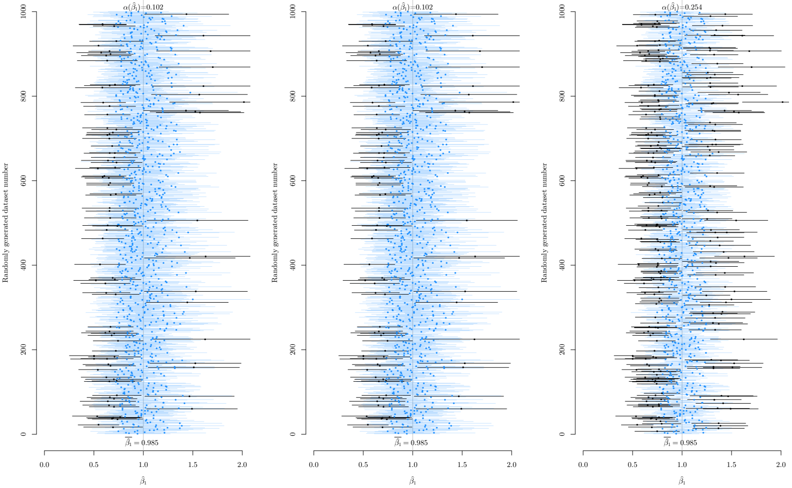
Fixed effects: Laplace

$\beta_1 = 1$	glmer	glmmADMB	ADMB	GLIMMIX	xtmelogit	fitglme	GENLINMIXED
$\alpha\left(\hat{\beta}_1\right)$	0.182	0.102	0.102	0.102	0.102	0.254	1.000
$\overline{\hat{\beta}}_1$	0.982	0.985	0.985	0.985	0.985	0.985	0.454



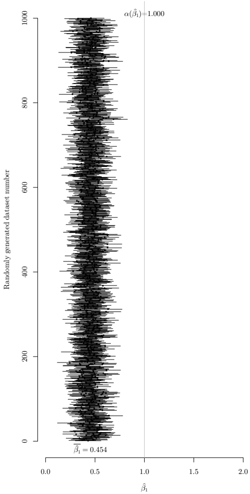
Fixed effects: Laplace

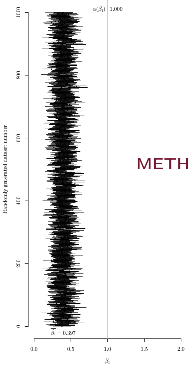
$\beta_1 = 1$	glmer	glmmADMB	ADMB	GLIMMIX	xtmelogit	fitglme	GENLINMIXED
$\alpha\left(\hat{\beta}_1\right)$	0.182	0.102	0.102	0.102	0.102	0.254	1.000
$\overline{\hat{\beta}_1}$	0.982	0.985	0.985	0.985	0.985	0.985	0.454



Fixed effects: PQL?

$\beta_1 = 1$	glmer	glmmADMB	ADMB	GLIMMIX	xtmelogit	fitglme	GENLINMIXED
$\alpha\left(\hat{\beta}_1\right)$	0.182	0.102	0.102	0.102	0.102	0.254	1.000
$\bar{\hat{\beta}}_1$	0.982	0.985	0.985	0.985	0.985	0.985	0.454

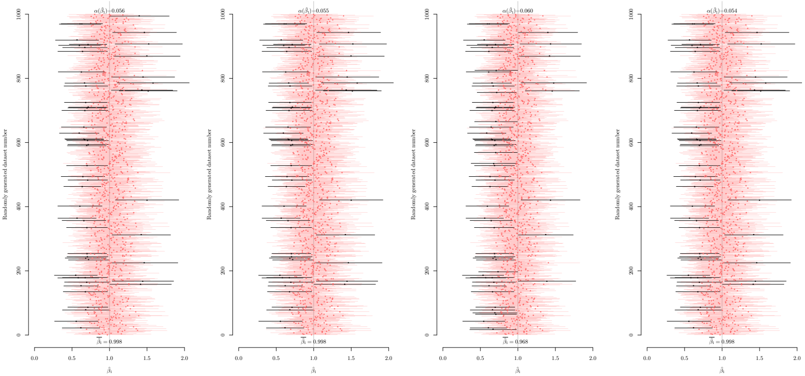




GLIMMIX
METHOD=MMPL
(PQL)

Fixed effects: aGHQ=7

$\beta_1 = 1$	ADMB	GLIMMIX	NLMIXED	xtmelogit
$\alpha\left(\hat{\beta}_1\right)$	0.056	0.055	0.060	0.054
$\overline{\hat{\beta}_1}$	0.998	0.998	0.968	0.998



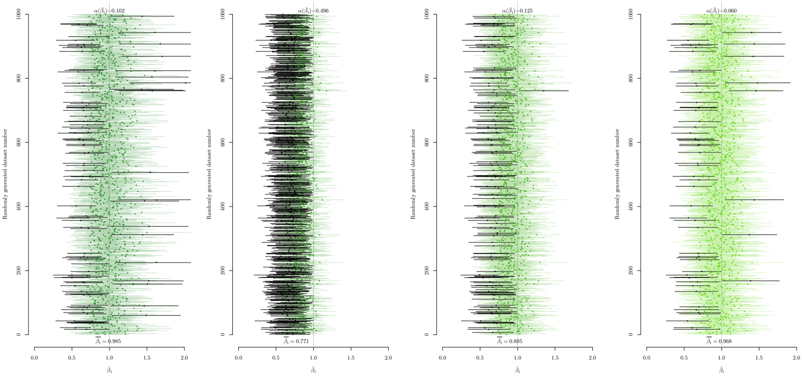
Stata results: increasing Q

- Effect of increasing Q on estimation of
  - β_1 $(= 1)$
  - τ_0 $(= 2)$
  - τ_1 $(= 2)$
  - ρ $(= 0.25)$

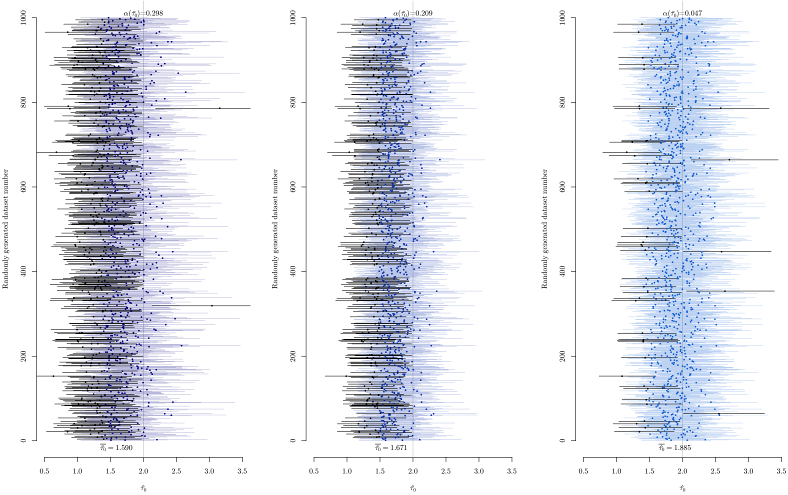
Note: `xtmelogit` calculates the variance components on the scales $\ln(\tau_0)$, $\ln(\tau_1)$, $\tanh^{-1}(\rho)$ because the sampling distribution

- cannot be assumed symmetric, and
- are constrained

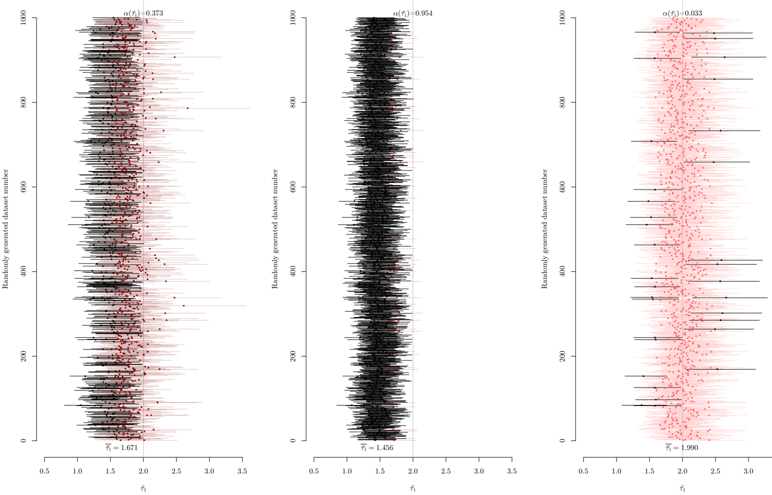
$\beta_1 = 1$	$Q = 1$	$Q = 2$	$Q = 3$	$Q = 4$	$Q = 5$	$Q = 6$	$Q = 7$
$\alpha(\hat{\beta}_1)$	0.102	0.496	0.125	0.057	0.060	0.052	0.055
$\bar{\hat{\beta}}_1$	0.985	0.771	0.895	0.990	0.968	1.024	0.998



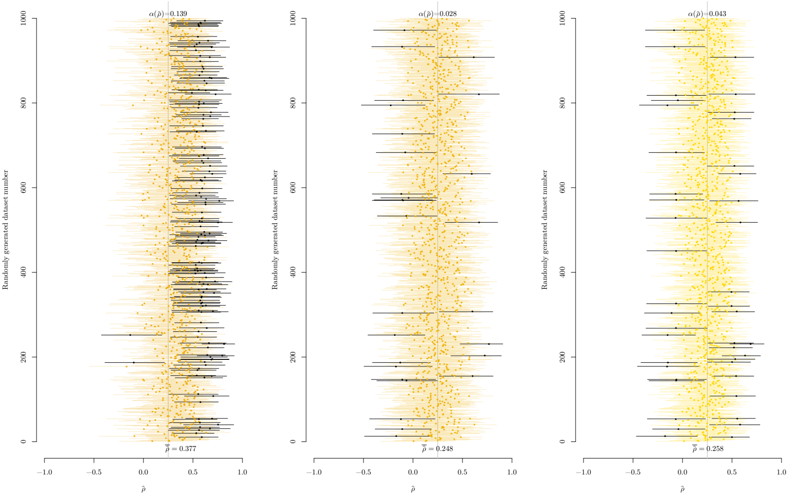
$\tau_0 = 2$	$Q = 1$	$Q = 2$	$Q = 3$	$Q = 4$	$Q = 5$	$Q = 6$	$Q = 7$
$\alpha(\hat{\tau}_0)$	0.298	0.993	0.209	0.038	0.047	0.060	0.040
$\bar{\hat{\tau}}_0$	1.590	1.201	1.671	1.930	1.885	2.069	1.982



$\tau_1 = 2$	$Q = 1$	$Q = 2$	$Q = 3$	$Q = 4$	$Q = 5$	$Q = 6$	$Q = 7$
$\alpha(\hat{\tau}_1)$	0.373	0.954	0.204	0.035	0.037	0.040	0.033
$\bar{\hat{\tau}}_1$	1.671	1.456	1.757	1.949	1.927	2.033	1.990



$\rho = 0.25$	$Q = 1$	$Q = 2$	$Q = 3$	$Q = 4$	$Q = 5$	$Q = 6$	$Q = 7$
$\alpha(\hat{\rho})$	0.139	0.028	0.028	0.049	0.036	0.046	0.043
$\bar{\hat{\rho}}$	0.377	0.248	0.243	0.248	0.259	0.251	0.258



Computational time (minutes)[1]

$$\gamma(\eta_{ij}) = \overbrace{\beta_0 + \beta_1 x_{1ij}}^{\text{fixed}} + \overbrace{\sum_{k=2}^{21} \beta_k x_{kij}}^{\text{fixed/noise}} + \overbrace{u_{i0} + u_{i1} x_{1ij}}^{\text{random}}, \qquad \begin{aligned} \beta_2 = \ldots = \beta_{21} = 0 \\ i = 1, 2, \ldots, 200 \\ j = 1, 2, \ldots, n_i \end{aligned}$$

Method	Software	$n_i = 10$	$n_i = 100$	$n_i = 1000$
Laplace	xtmelogit	2	5	32
	NLMIXED††	2	21	187
	GLIMMIX†	0	0	3
	ADMB-RE	2	14	N/A
	glmer	2	3	16
	fitglme	0	1	17
aGHQ ($Q = 7$)	xtmelogit	6	12	90
	NLMIXED††	18	203	4299
	GLIMMIX†	0	1	17
	ADMB-RE	2	17	N/A

[1] Mac Pro (2010): 2 × 2.93GHz 6-Core Intel Xeon, 32GB DDR3, SSD)

Forward step-wise model selection using AIC

$$\gamma\left(\eta_{ij}\right) = \overbrace{\beta_0 + \sum_{k=1}^{5} \beta_k x_{kij}}^{\text{fixed/signal}} + \overbrace{\sum_{k=6}^{25} \beta_k x_{kij}}^{\text{fixed/noise}} + \overbrace{u_{i0} + u_{i1} x_{1ij}}^{\text{random}},$$

$$\beta_6 = \ldots = \beta_{25} = 0$$
$$i = 1, 2, \ldots, 20$$
$$j = 1, 2, \ldots, n_i$$

$n_i = 10$ (run 100 times)	Laplace	$Q = 7$
Correctly identified covariates (/5)	4.76	4.76
Incorrectly identified covariates (/20)	3.54	3.55
Random slope identified (/1)	0.69	0.67

Forward step-wise model selection using AIC

$$\gamma\left(\eta_{ij}\right) =$$

$$\overbrace{\beta_0 + \sum_{k=1}^{5} \beta_k x_{kij}}^{\text{fixed/signal}} + \overbrace{\sum_{k=6}^{25} \beta_k x_{kij}}^{\text{fixed/noise}} + \overbrace{\sum_{k=1}^{24} \sum_{k'=k+1}^{25} \beta_{kk'} x_{kij} x_{k'ij}}^{\text{fixed/interactions}} + \overbrace{u_{i0} + u_{i1} x_{1ij}}^{\text{random}},$$

$$\text{where} \quad \begin{aligned} \beta_{12}, \beta_{13}, \beta_{45} &\neq 0 \\ \text{all other } \beta_{kk'} &= 0 \end{aligned}$$

$n_i = 30$ (run 10 times)	Laplace	$Q = 7$
Correctly identified covariates (/5)	4.9	4.9
Incorrectly identified covariates (/20)	3.0	3.0
Correctly identified interactions (/3)	2.8	2.8
Incorrectly identified interactions (/297)	2.5	2.5
Random slope identified (/1)	0.9	0.9

Summary

For GLMMs with random effects actually generated from the assumed distribution (Gaussian) AND binary outcome data:

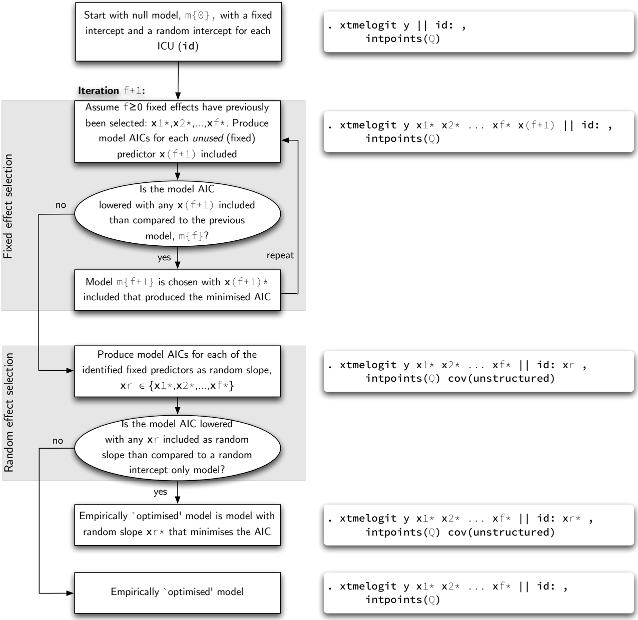
- Don't use $Q = 2$
- $Q = 1$ is insufficient to estimate variance components
- $Q \geq 7$ gives reasonably accurate results
- SAS was the fastest package for aGHQ
- Model building using AIC is the same irrespective of $Q = 1$ or $Q = 7$ (AIC overfits)

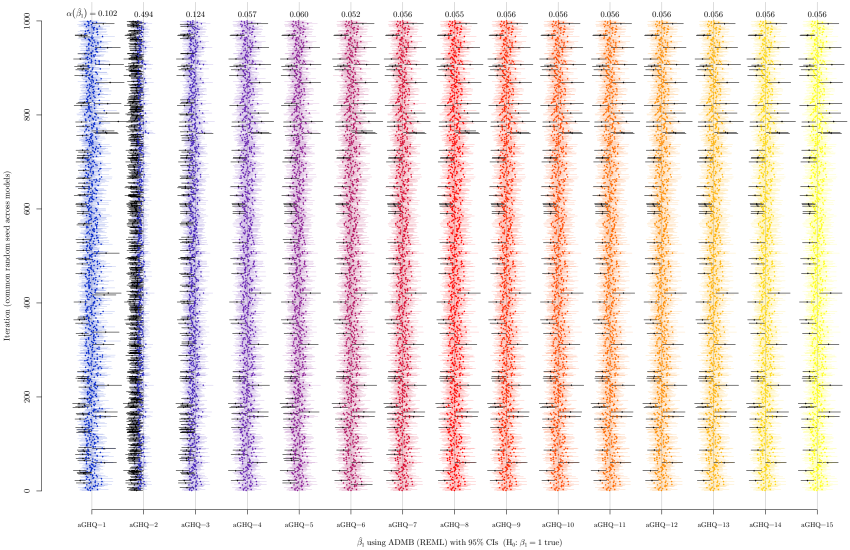
Acknowledgements

Thank you to The University of Adelaide and my co-author Professor Patty Solomon.

This research was supported under Australian Research Council's Discovery Projects funding scheme. The views expressed are those of the authors and are not necessarily those of the Australian Research Council.

Much of the computation was made feasible using the command line parallel computing utility: `GNU Parallel`. Please see `http://www.gnu.org/s/parallel` or the *;login: The USENIX Magazine* article (O. Tange; 2011) for more details.


Start with null model, m{0}, with a fixed intercept and a random intercept for each ICU (id)
. xtmelogit y || id: , intpoints(Q)
Iteration f+1:
Fixed effect selection
Assume f≥0 fixed effects have previously been selected: x1*,x2*,...,xf*. Produce model AICs for each unused (fixed) predictor x(f+1) included
. xtmelogit y x1* x2* ... xf* x(f+1) || id: , intpoints(Q)
Is the model AIC lowered with any x(f+1) included than compared to the previous model, m{f}?
no
yes
repeat
Model m{f+1} is chosen with x(f+1)* included that produced the minimised AIC
Random effect selection
Produce model AICs for each of the identified fixed predictors as random slope, xr ∈ {x1*,x2*,...,xf*}
. xtmelogit y x1* x2* ... xf* || id: xr , intpoints(Q) cov(unstructured)
Is the model AIC lowered with any xr included as random slope than compared to a random intercept only model?
no
yes
Empirically `optimised' model is model with random slope xr* that minimises the AIC
. xtmelogit y x1* x2* ... xf* || id: xr* , intpoints(Q) cov(unstructured)
Empirically `optimised' model
. xtmelogit y x1* x2* ... xf* || id: , intpoints(Q)


$\alpha(\hat{\beta}_1) = 0.102$
0.494
0.124
0.057
0.060
0.052
0.056
0.055
0.056
0.056
0.056
0.056
0.056
0.056
0.056
Iteration (common random seed across models)
0
200
400
600
800
1000
aGHQ=1
aGHQ=2
aGHQ=3
aGHQ=4
aGHQ=5
aGHQ=6
aGHQ=7
aGHQ=8
aGHQ=9
aGHQ=10
aGHQ=11
aGHQ=12
aGHQ=13
aGHQ=14
aGHQ=15
$\hat{\beta}_1$ using ADMB (REML) with 95% CIs (H_0: $\beta_1 = 1$ true)