

# Generalized linear mixed models for ecologists and evolutionary biologists

Ben Bolker, University of Florida

Harvard Forest

27 February 2009

# Outline

- 1 Precursors
  - Generalized linear models
  - Mixed models (LMMs)
  
- 2 GLMMs
  - Estimation
  - Inference

# Outline

- 1 Precursors
  - Generalized linear models
  - Mixed models (LMMs)
- 2 GLMMs
  - Estimation
  - Inference

# Generalized linear models (GLMs)

- non-normal data, (some) nonlinear relationships
- presence/absence, alive/dead (binomial); count data (Poisson, negative binomial)
- nonlinearity via **link** function  $L$ : response is nonlinear, but  $L(\text{response})$  is linear (e.g. log/exponential, logit/logistic)
- modeling via **linear predictor** is very flexible:  
 $L(\text{response}) = a + b_i + cx \dots$
- stable, robust, efficient: typical applications are **logistic regression** (binomial/logit link), **Poisson regression** (Poisson/log link)

# Generalized linear models (GLMs)

- non-normal data, (some) nonlinear relationships
- presence/absence, alive/dead (binomial); count data (Poisson, negative binomial)
- nonlinearity via **link** function  $L$ : response is nonlinear, but  $L(\text{response})$  is linear (e.g. log/exponential, logit/logistic)
- modeling via **linear predictor** is very flexible:  
 $L(\text{response}) = a + b_i + cx \dots$
- stable, robust, efficient: typical applications are **logistic regression** (binomial/logit link), **Poisson regression** (Poisson/log link)

# Generalized linear models (GLMs)

- non-normal data, (some) nonlinear relationships
- presence/absence, alive/dead (binomial); count data (Poisson, negative binomial)
- nonlinearity via **link** function  $L$ : response is nonlinear, but  $L(\text{response})$  is linear (e.g. log/exponential, logit/logistic)
- modeling via **linear predictor** is very flexible:  
 $L(\text{response}) = a + b_i + cx \dots$
- stable, robust, efficient: typical applications are **logistic regression** (binomial/logit link), **Poisson regression** (Poisson/log link)

# Generalized linear models (GLMs)

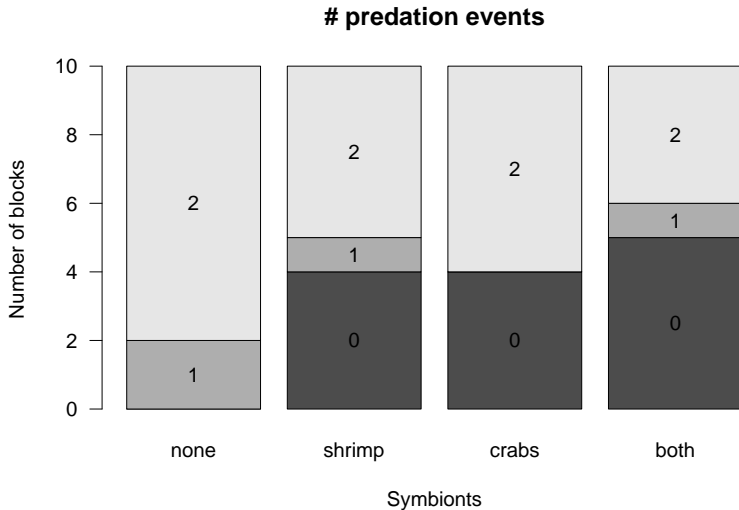
- non-normal data, (some) nonlinear relationships
- presence/absence, alive/dead (binomial); count data (Poisson, negative binomial)
- nonlinearity via **link** function  $L$ : response is nonlinear, but  $L(\text{response})$  is linear (e.g. log/exponential, logit/logistic)
- modeling via **linear predictor** is very flexible:  
$$L(\text{response}) = a + b_i + cx \dots$$
- stable, robust, efficient: typical applications are **logistic regression** (binomial/logit link), **Poisson regression** (Poisson/log link)

# Generalized linear models (GLMs)

- non-normal data, (some) nonlinear relationships
- presence/absence, alive/dead (binomial); count data (Poisson, negative binomial)
- nonlinearity via **link** function  $L$ : response is nonlinear, but  $L(\text{response})$  is linear (e.g. log/exponential, logit/logistic)
- modeling via **linear predictor** is very flexible:  
$$L(\text{response}) = a + b_i + cx \dots$$
- stable, robust, efficient: typical applications are **logistic regression** (binomial/logit link), **Poisson regression** (Poisson/log link)

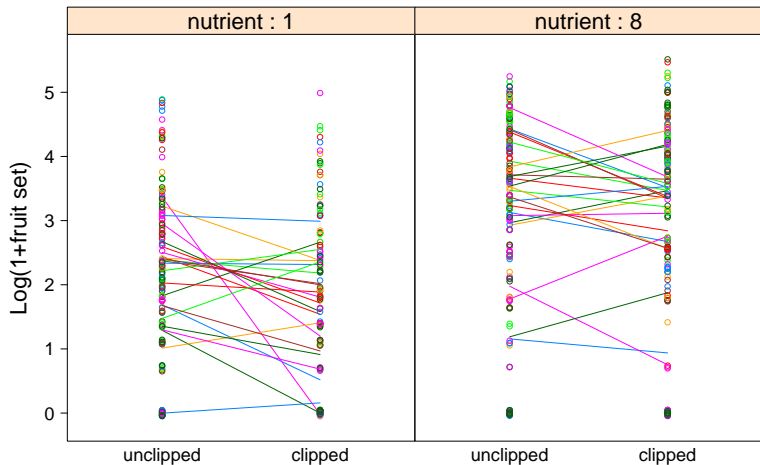


# Coral symbiont example



# Arabidopsis example

panel: nutrient, color: genotype



# Outline

- 1 Precursors
  - Generalized linear models
  - Mixed models (LMMs)
- 2 GLMMs
  - Estimation
  - Inference

# Random effects (RE)

- examples: experimental or observational “blocks” (temporal, spatial); species or genera; individuals; genotypes
- inference on **population** of units rather than individual units
- (?) units randomly selected from all possible units
- (?) reasonably large number of units

# Random effects (RE)

- examples: experimental or observational “blocks” (temporal, spatial); species or genera; individuals; genotypes
- inference on **population** of units rather than individual units
- (?) units randomly selected from all possible units
- (?) reasonably large number of units

# Random effects (RE)

- examples: experimental or observational “blocks” (temporal, spatial); species or genera; individuals; genotypes
- inference on **population** of units rather than individual units
- (?) units randomly selected from all possible units
- (?) reasonably large number of units

# Random effects (RE)

- examples: experimental or observational “blocks” (temporal, spatial); species or genera; individuals; genotypes
- inference on **population** of units rather than individual units
- (?) units randomly selected from all possible units
- (?) reasonably large number of units

# Mixed models: classical approach

- Partition sums of squares, calculate null expectations if fixed effect is 0 (all coefficients  $\beta_i = 0$ ) or RE variance=0
- Figure out numerator (model) & denominator (residual) sums of squares and degrees of freedom
  - **Model SSQ, df**: variability explained by the “effect” (difference between model with and without the effect) and number of parameters used
  - **Residual SSQ, df**: variability caused by finite sample size (number of observations minus number “used up” by the model)



# Mixed models: classical approach

- Partition sums of squares, calculate null expectations if fixed effect is 0 (all coefficients  $\beta_i = 0$ ) or RE variance=0
- Figure out numerator (model) & denominator (residual) sums of squares and degrees of freedom
  - **Model SSQ, df**: variability explained by the “effect” (difference between model with and without the effect) and number of parameters used
  - **Residual SSQ, df**: variability caused by finite sample size (number of observations minus number “used up” by the model)

# Classical LMM cont.

- Robust, practical
- OK if
  - data are normally distributed
  - design is (nearly) **balanced**
  - design not too complicated (single RE, or nested REs)  
(**crossed** REs: e.g. year effects that apply across all spatial blocks)

# Mixed models: modern approach

- Construct a **likelihood** for the data (Prob(observing data|parameters)) — in mixed models, requires integrating over possible values of REs
- e.g.:
  - likelihood of  $i^{\text{th}}$  obs. in block  $j$  is  $L_{\text{Normal}}(x_{ij}|\theta_j, \sigma_w^2)$
  - likelihood of a particular block mean  $\theta_j$  is  $L_{\text{Normal}}(\theta_j|0, \sigma_b^2)$
  - overall likelihood is  $\int L(x_{ij}|\theta_j, \sigma_w^2)L(\theta_j|0, \sigma_b^2) d\theta_j$
- Figure out how to do the integral

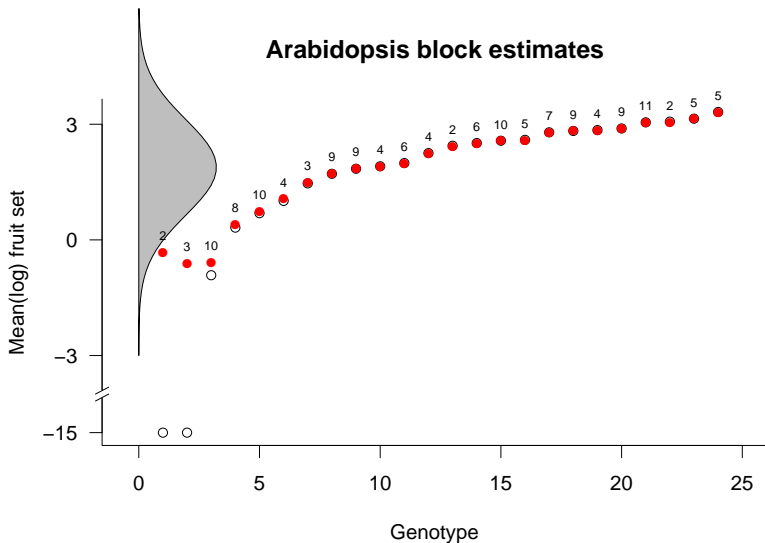
# Mixed models: modern approach

- Construct a **likelihood** for the data  
(Prob(observing data|parameters)) — in mixed models,  
requires integrating over possible values of REs
- e.g.:
  - likelihood of  $i^{\text{th}}$  obs. in block  $j$  is  $L_{\text{Normal}}(x_{ij}|\theta_j, \sigma_w^2)$
  - likelihood of a particular block mean  $\theta_j$  is  $L_{\text{Normal}}(\theta_j|0, \sigma_b^2)$
  - overall likelihood is  $\int L(x_{ij}|\theta_j, \sigma_w^2)L(\theta_j|0, \sigma_b^2) d\theta_j$
- Figure out how to do the integral

# Mixed models: modern approach

- Construct a **likelihood** for the data  
(Prob(observing data|parameters)) — in mixed models,  
requires integrating over possible values of REs
- e.g.:
  - likelihood of  $i^{\text{th}}$  obs. in block  $j$  is  $L_{\text{Normal}}(x_{ij}|\theta_i, \sigma_w^2)$
  - likelihood of a particular block mean  $\theta_j$  is  $L_{\text{Normal}}(\theta_j|0, \sigma_b^2)$
  - overall likelihood is  $\int L(x_{ij}|\theta_j, \sigma_w^2)L(\theta_j|0, \sigma_b^2) d\theta_j$
- Figure out how to do the integral

# Shrinkage



# RE examples

- Coral symbionts: simple experimental blocks, RE affects intercept (overall probability of predation in block)
- *Arabidopsis*: region (3 levels, treated as fixed) / population / genotype: affects intercept (overall fruit set) as well as treatment effects (nutrients, herbivory, interaction)

# Outline

- 1 Precursors
  - Generalized linear models
  - Mixed models (LMMs)
- 2 GLMMs
  - Estimation
  - Inference



# Penalized quasi-likelihood (PQL)

- alternate steps of estimating GLM given known block variances; estimate LMMs given GLM fit
- flexible (allows spatial, temporal correlations, crossed REs, etc.)
- but: “quasi” likelihood only (inference problems?)
- **biased** for small unit samples (e.g. counts  $< 5$ , binomial with  $\min(\text{success}, \text{failure}) < 5$ ) (!! (Breslow, 2004)
- nevertheless, **widely** used: SAS PROC GLIMMIX, R `glmmPQL`: “95% of analyses of binary responses ( $n = 205$ ), 92% of Poisson responses with means less than 5 ( $n = 48$ ) and 89% of binomial responses with fewer than 5 successes per group ( $n = 38$ )”

# Penalized quasi-likelihood (PQL)

- alternate steps of estimating GLM given known block variances; estimate LMMs given GLM fit
- flexible (allows spatial, temporal correlations, crossed REs, etc.)
- but: “quasi” likelihood only (inference problems?)
- **biased** for small unit samples (e.g. counts  $< 5$ , binomial with  $\min(\text{success}, \text{failure}) < 5$ ) (!! (Breslow, 2004)
- nevertheless, **widely** used: SAS PROC GLIMMIX, R `glmmPQL`: “95% of analyses of binary responses ( $n = 205$ ), 92% of Poisson responses with means less than 5 ( $n = 48$ ) and 89% of binomial responses with fewer than 5 successes per group ( $n = 38$ )”

# Penalized quasi-likelihood (PQL)

- alternate steps of estimating GLM given known block variances; estimate LMMs given GLM fit
- flexible (allows spatial, temporal correlations, crossed REs, etc.)
- but: “quasi” likelihood only (inference problems?)
- **biased** for small unit samples (e.g. counts  $< 5$ , binomial with  $\min(\text{success}, \text{failure}) < 5$ ) (!! (Breslow, 2004)
- nevertheless, **widely** used: SAS PROC GLIMMIX, R `glmmPQL`: “95% of analyses of binary responses ( $n = 205$ ), 92% of Poisson responses with means less than 5 ( $n = 48$ ) and 89% of binomial responses with fewer than 5 successes per group ( $n = 38$ )”

# Penalized quasi-likelihood (PQL)

- alternate steps of estimating GLM given known block variances; estimate LMMs given GLM fit
- flexible (allows spatial, temporal correlations, crossed REs, etc.)
- but: “quasi” likelihood only (inference problems?)
- **biased** for small unit samples (e.g. counts  $< 5$ , binomial with  $\min(\text{success}, \text{failure}) < 5$ ) (!! (Breslow, 2004)
- nevertheless, **widely** used: SAS PROC GLIMMIX, R `glmmPQL`: “95% of analyses of binary responses ( $n = 205$ ), 92% of Poisson responses with means less than 5 ( $n = 48$ ) and 89% of binomial responses with fewer than 5 successes per group ( $n = 38$ )”

# Penalized quasi-likelihood (PQL)

- alternate steps of estimating GLM given known block variances; estimate LMMs given GLM fit
- flexible (allows spatial, temporal correlations, crossed REs, etc.)
- but: “quasi” likelihood only (inference problems?)
- **biased** for small unit samples (e.g. counts  $< 5$ , binomial with  $\min(\text{success}, \text{failure}) < 5$ ) (!! ) (Breslow, 2004)
- nevertheless, **widely** used: SAS PROC GLIMMIX, R `glmmPQL`: “95% of analyses of binary responses ( $n = 205$ ), 92% of Poisson responses with means less than 5 ( $n = 48$ ) and 89% of binomial responses with fewer than 5 successes per group ( $n = 38$ )”

# Better methods

- **Laplace approximation**

- approximate integral ( $\int L(\text{data}|\text{block})L(\text{block}|\text{block variance})$ ) by Taylor expansion
- considerably more accurate than PQL
- reasonably fast and flexible

- **adaptive Gauss-Hermite quadrature** ([A]G[H]Q)

- compute additional terms in the integral
- most accurate
- slowest, hence not flexible (2–3 RE at most, maybe only 1)

Becoming available: R lme4, SAS PROC NLMIXED, PROC GLIMMIX (v. 9.2), Genstat GLMM

# Better methods

- **Laplace approximation**
  - approximate integral ( $\int L(\text{data}|\text{block})L(\text{block}|\text{block variance})$ ) by Taylor expansion
  - considerably more accurate than PQL
  - reasonably fast and flexible
- **adaptive Gauss-Hermite quadrature** ([A]G[H]Q)
  - compute additional terms in the integral
  - most accurate
  - slowest, hence not flexible (2–3 RE at most, maybe only 1)

Becoming available: R lme4, SAS PROC NLMIXED, PROC GLIMMIX (v. 9.2), Genstat GLMM

# Better methods

- **Laplace approximation**

- approximate integral ( $\int L(\text{data}|\text{block})L(\text{block}|\text{block variance})$ ) by Taylor expansion
- considerably more accurate than PQL
- reasonably fast and flexible

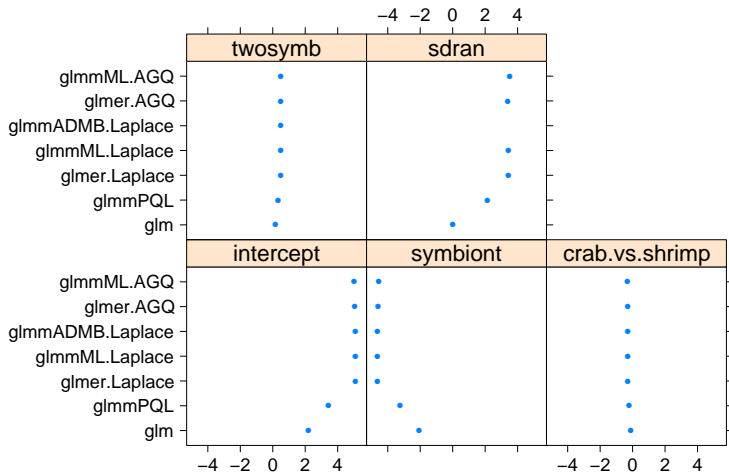
- **adaptive Gauss-Hermite quadrature** ([A]G[H]Q)

- compute additional terms in the integral
- most accurate
- slowest, hence not flexible (2–3 RE at most, maybe only 1)

Becoming available: R `lme4`, SAS PROC NL MIXED, PROC GLIMMIX (v. 9.2), Genstat GLMM



# Comparison of coral symbiont results



# Outline

- 1 Precursors
  - Generalized linear models
  - Mixed models (LMMs)
- 2 GLMMs
  - Estimation
  - Inference

# General issues: testing RE significance

- Counting “model” df for REs
  - how many parameters does a RE require? 1 (variance)? Or somewhere between 1 and  $n$  (shrinkage estimator)? Hard to compute, and depends on the **level of focus** (Vaida and Blanchard, 2005)
- Boundary effects for RE testing
  - most derivations of null distributions depend on the null-hypothesis value of the parameter being **within** its feasible range (i.e. not on the boundary) (Molenberghs and Verbeke, 2007)
  - REs may count for  $< 1$  df (typically  $\approx 0.5$ )
  - if ignored, tests are (slightly) conservative

# General issues: testing RE significance

- Counting “model” df for REs
  - how many parameters does a RE require? 1 (variance)? Or somewhere between 1 and  $n$  (shrinkage estimator)? Hard to compute, and depends on the **level of focus** (Vaida and Blanchard, 2005)
- Boundary effects for RE testing
  - most derivations of null distributions depend on the null-hypothesis value of the parameter being **within** its feasible range (i.e. not on the boundary) (Molenberghs and Verbeke, 2007)
  - REs may count for  $< 1$  df (typically  $\approx 0.5$ )
  - if ignored, tests are (slightly) conservative

# General issues: finite-sample issues (!)

How far are we from “asymptopia”?

- Many standard procedures are **asymptotic**
  - Likelihood Ratio Test
  - AIC
- “Sample size” may refer the number of RE **units** — often far more restricted than total number of data points, so finite-sample problems are more pressing
- Degree of freedom counting is hard for complex designs:  
**Kenward-Roger correction**

# General issues: finite-sample issues (!)

How far are we from “asymptopia”?

- Many standard procedures are **asymptotic**
  - Likelihood Ratio Test
  - AIC
- “Sample size” may refer the number of RE **units** — often far more restricted than total number of data points, so finite-sample problems are more pressing
- Degree of freedom counting is hard for complex designs:  
**Kenward-Roger correction**

# General issues: finite-sample issues (!)

How far are we from “asymptopia”?

- Many standard procedures are **asymptotic**
  - Likelihood Ratio Test
  - AIC
- “Sample size” may refer the number of RE **units** — often far more restricted than total number of data points, so finite-sample problems are more pressing
- Degree of freedom counting is hard for complex designs:  
**Kenward-Roger correction**

# Specific procedures

- **Likelihood Ratio Test:**  
unreliable for fixed effects except for large data sets (= large # of RE units!)
- **Wald** ( $Z$ ,  $\chi^2$ ,  $t$  or  $F$ ) tests
  - crude approximation
  - asymptotic (for non-overdispersed data?) or ...
  - ...how do we count residual df?
  - don't know if null distributions are correct
- AIC
  - asymptotic
  - could use  $AIC_c$ , but ? need residual df



# Specific procedures

- **Likelihood Ratio Test:**  
unreliable for fixed effects except for large data sets (= large # of RE units!)
- **Wald** ( $Z$ ,  $\chi^2$ ,  $t$  or  $F$ ) tests
  - crude approximation
  - asymptotic (for non-overdispersed data?) or ...
  - ... how do we count residual df?
  - don't know if null distributions are correct
- AIC
  - asymptotic
  - could use  $AIC_c$ , but ? need residual df

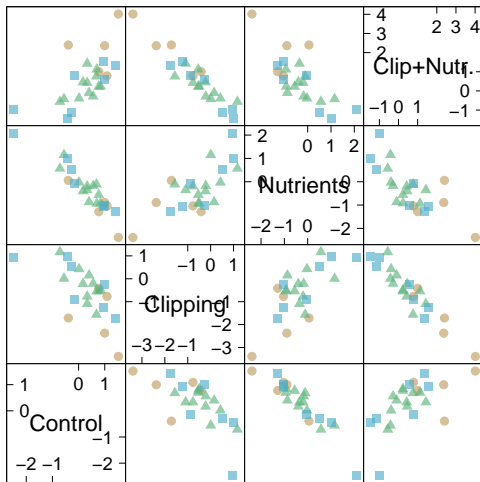
# Specific procedures

- **Likelihood Ratio Test:**  
unreliable for fixed effects except for large data sets (= large # of RE units!)
- **Wald** ( $Z$ ,  $\chi^2$ ,  $t$  or  $F$ ) tests
  - crude approximation
  - asymptotic (for non-overdispersed data?) or ...
  - ... how do we count residual df?
  - don't know if null distributions are correct
- AIC
  - asymptotic
  - could use  $AIC_c$ , but ? need residual df

# *Arabidopsis* results (sort of)

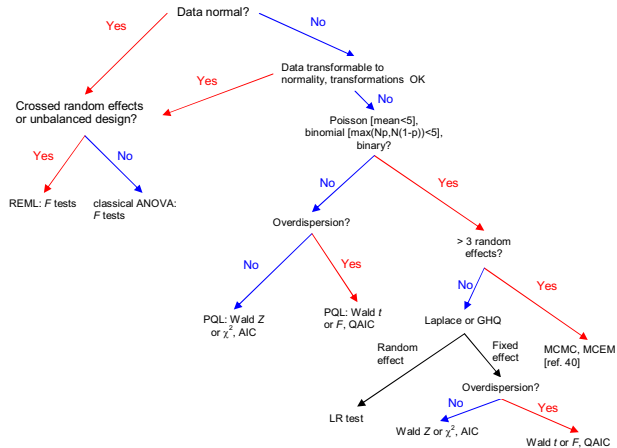
- RE model selection by (Q)AIC: intercept difference among populations, genotypes vary in all parameters (1+nutrient+clipping+ nutrient  $\times$  clipping)
- fixed effects: nutrient, others weak

# *Arabidopsis* genotype effects



Scatter Plot Matrix

## Where are we?



# Now what?

- MCMC (finicky, slow, dangerous, we have to “go Bayesian”: specialized procedures for GLMMs, or WinBUGS translators? (`glmmBUGS`, `MCMCglmm`)
- quasi-Bayes `mcmc` in `lme4` (unfinished!)
- parametric bootstrapping:
  - fit null model to data
  - simulate “data” from null model
  - fit null and working model, compute likelihood diff.
  - repeat to estimate null distribution

More info: [glmm.wikidot.com](http://glmm.wikidot.com)

# Now what?

- MCMC (finicky, slow, dangerous, we have to “go Bayesian”:  
specialized procedures for GLMMs, or WinBUGS translators?  
(`glmmBUGS`, `MCMCglmm`)
- quasi-Bayes `mcmc` in `lme4` (unfinished!)
- parametric bootstrapping:
  - fit null model to data
  - simulate “data” from null model
  - fit null and working model, compute likelihood diff.
  - repeat to estimate null distribution

More info: [glmm.wikidot.com](http://glmm.wikidot.com)

# Acknowledgements

- Data: Josh Banta and Massimo Pigliucci (*Arabidopsis*); Adrian Stier and Sea McKeon (coral symbionts)
- Co-authors: Mollie Brooks, Connie Clark, Shane Geange, John Poulsen, Hank Stevens, Jada White



# References

- Breslow, N.E., 2004. In D.Y. Lin and P.J. Heagerty, editors, *Proceedings of the second Seattle symposium in biostatistics: Analysis of correlated data*, pages 1–22. Springer. ISBN 0387208623.
- Molenberghs, G. and Verbeke, G., 2007. *The American Statistician*, 61(1):22–27. doi:10.1198/000313007X171322.
- Vaida, F. and Blanchard, S., 2005. *Biometrika*, 92(2):351–370. doi:10.1093/biomet/92.2.351.