Generalized linear mixed models for biologists

Ben Bolker, University of Florida

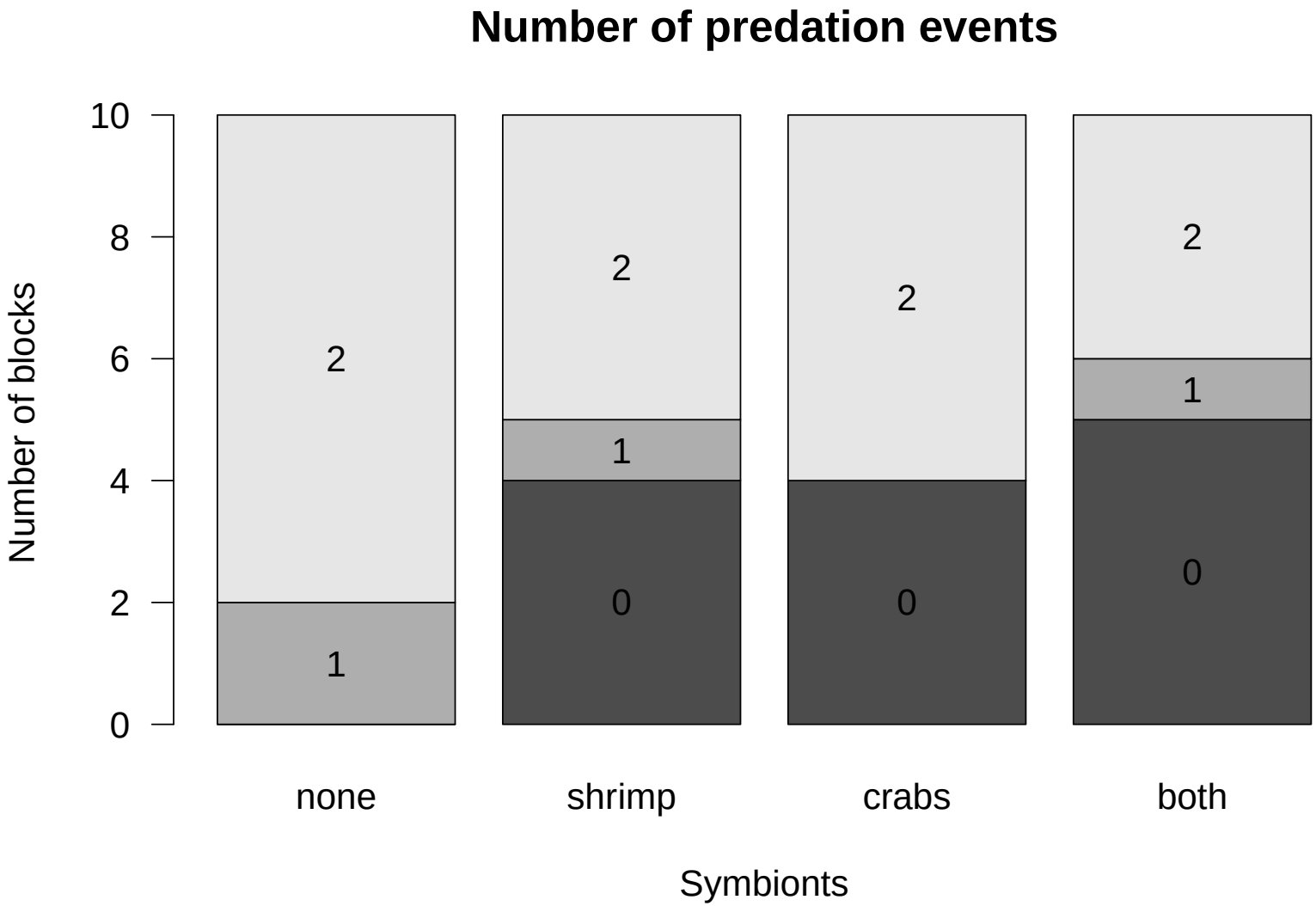
McMaster University

7 May 2009

Outline

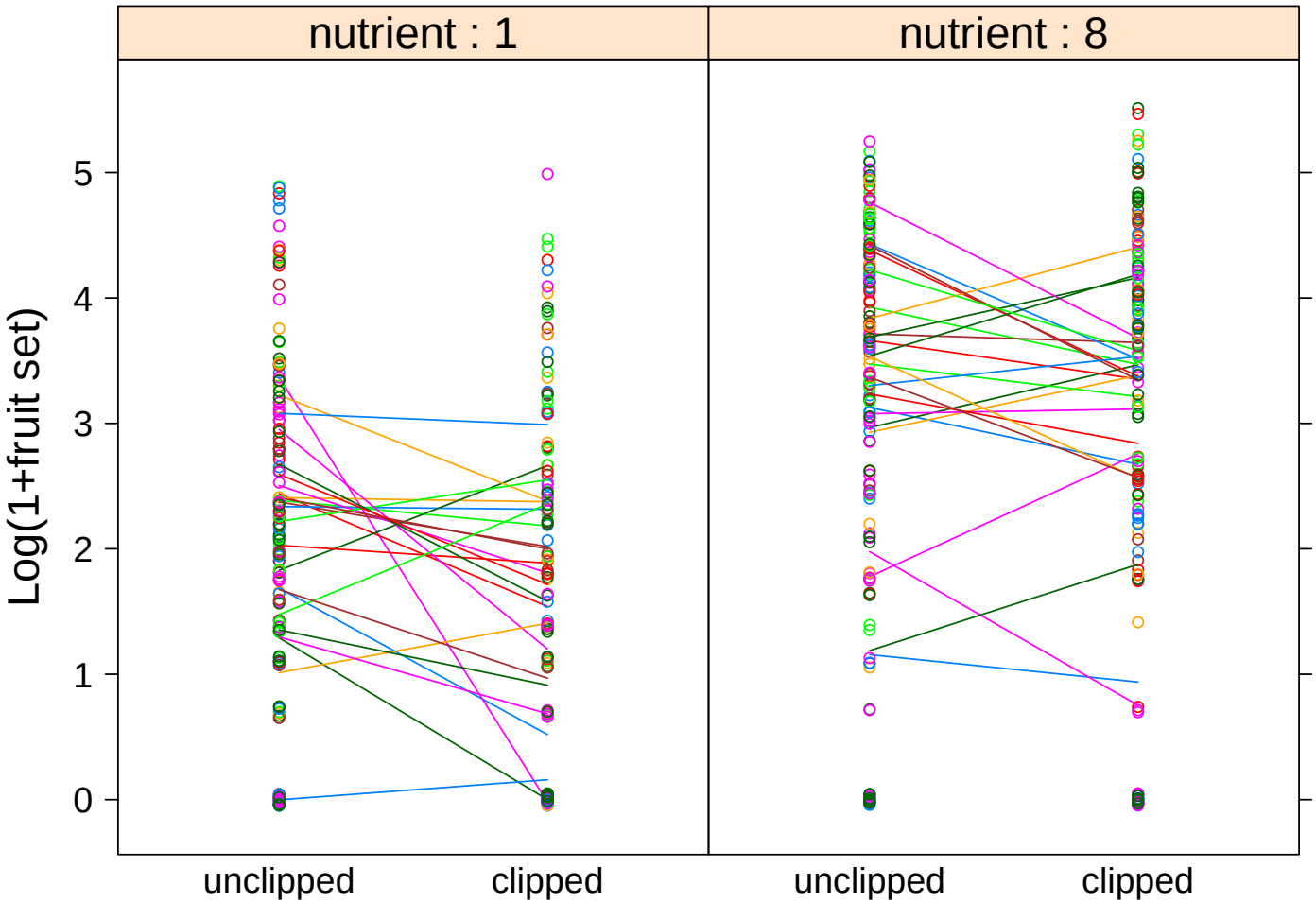
Outline

Coral protection by symbionts

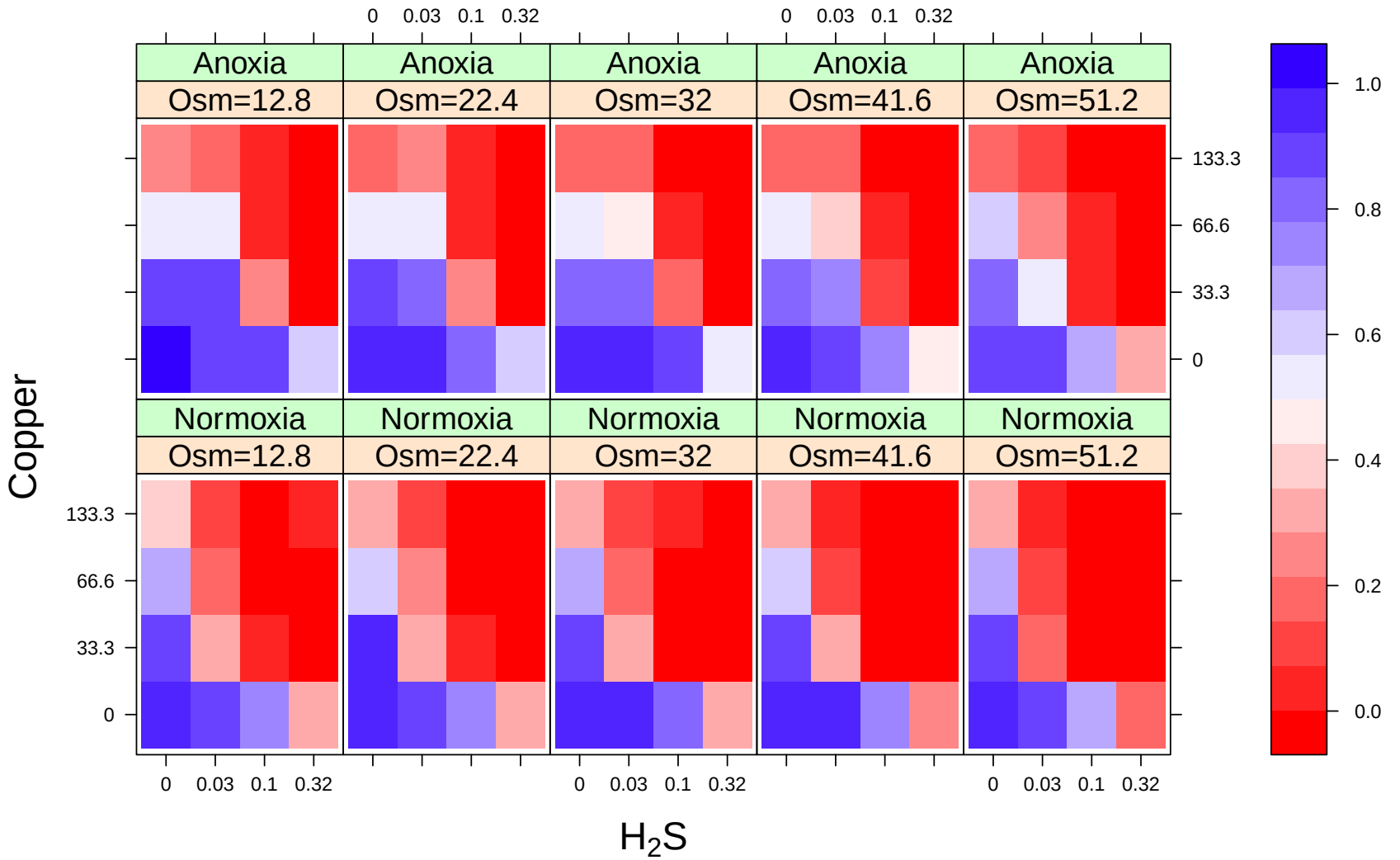


Arabidopsis response to fertilization & clipping

panel: nutrient, color: genotype



Environmental stress: *Glycera* cell survival



Outline

Generalized linear models (GLMs)

- non-normal data, (some) nonlinear relationships; modeling via **linear predictor**
  - presence/absence, alive/dead (binomial)
  - count data (Poisson, negative binomial)
- typical applications: **logistic regression** (binomial/logistic), **Poisson regression** (Poisson/exponential)

Generalized linear models (GLMs)

- non-normal data, (some) nonlinear relationships; modeling via **linear predictor**
-
  - presence/absence, alive/dead (binomial)
  - count data (Poisson, negative binomial)
- typical applications: **logistic regression** (binomial/logistic), **Poisson regression** (Poisson/exponential)

Generalized linear models (GLMs)

- non-normal data, (some) nonlinear relationships; modeling via **linear predictor**
-
  - presence/absence, alive/dead (binomial)
  - count data (Poisson, negative binomial)
- typical applications: **logistic regression** (binomial/logistic), **Poisson regression** (Poisson/exponential)

Outline

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 **Normal**
  - 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 (**marginal likelihood**)
- e.g.:
  - likelihood of i^{th} obs. in block j is $L_{\text{Normal}}(x_{ij}|\theta_i, \sigma_w^2)$
  - likelihood of a particular block mean θ_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

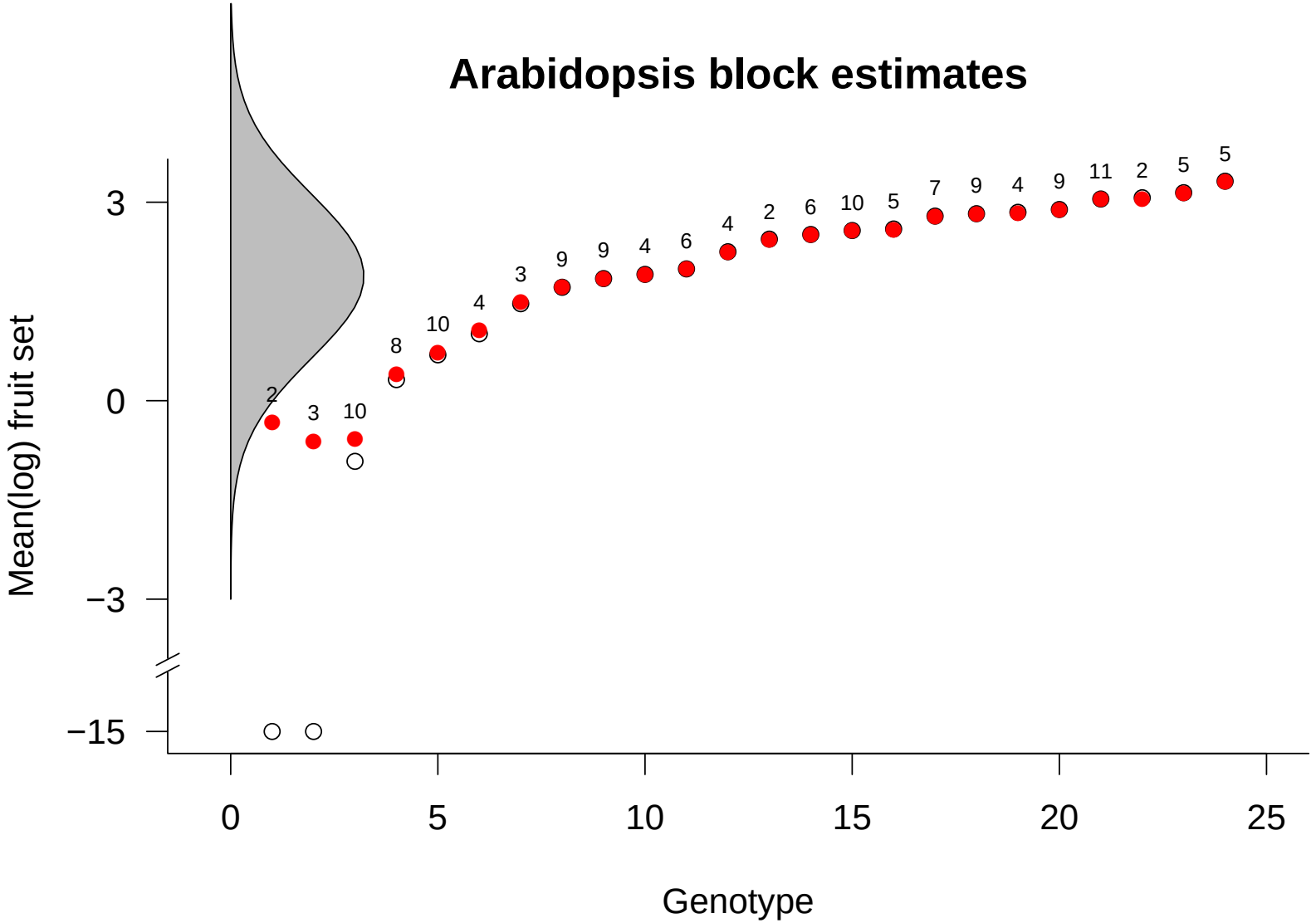
Mixed models: modern approach

- Construct a **likelihood** for the data (Prob(observing data|parameters)) — in mixed models, requires integrating over possible values of REs (**marginal likelihood**)
- e.g.:
  - likelihood of i^{th} obs. in block j is $L_{\text{Normal}}(x_{ij}|\theta_i, \sigma_w^2)$
  - likelihood of a particular block mean θ_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 (**marginal likelihood**)
- e.g.:
  - likelihood of i^{th} obs. in block j is $L_{\text{Normal}}(x_{ij}|\theta_i, \sigma_w^2)$
  - likelihood of a particular block mean θ_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


Arabidopsis block estimates
Mean(log) fruit set
3
0
−3
−15
0
5
10
15
20
25
Genotype
2 3 10 8 10 4 3 9 9 4 6 4 2 6 10 5 7 9 4 9 11 2 5 5

RE examples

- Coral symbionts: simple experimental blocks, RE affects intercept (overall probability of predation in block)
- *Glycera*: applied to cells from 10 individuals, RE again affects intercept (cell survival prob.)
- *Arabidopsis*: region (3 levels, treated as fixed) / population / genotype: affects intercept (overall fruit set) as well as treatment effects (nutrients, herbivory, interaction)

Outline

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)
- **biased** for small unit samples (e.g. counts < 5, binary or low-survival data) (Breslow, 2004)
- nevertheless, **widely** used: SAS `PROC GLIMMIX`, R `glmmPQL`: in $\approx 90\%$ of small-unit-sample cases

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)
- **biased** for small unit samples (e.g. counts < 5, binary or low-survival data) (Breslow, 2004)
- nevertheless, **widely** used: SAS `PROC GLIMMIX`, R `glmmPQL`: in $\approx 90\%$ of small-unit-sample cases

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)
- **biased** for small unit samples (e.g. counts < 5, binary or low-survival data) (Breslow, 2004)
- nevertheless, **widely** used: SAS `PROC GLIMMIX`, R `glmmPQL`: in $\approx 90\%$ of small-unit-sample cases

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)
- **biased** for small unit samples (e.g. counts < 5, binary or low-survival data) (Breslow, 2004)
- nevertheless, **widely** used: SAS `PROC GLIMMIX`, R `glmmPQL`: in $\approx 90\%$ of small-unit-sample cases

Better methods

- **Laplace approximation**
  - approximate marginal likelihood
  - considerably more accurate than PQL
  - reasonably fast and flexible
- **adaptive Gauss-Hermite quadrature** (AGQ)
  - 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 marginal likelihood
  - considerably more accurate than PQL
  - reasonably fast and flexible
- **adaptive Gauss-Hermite quadrature** (AGQ)
  - compute additional terms in the integral
  - most accurate
  - slowest, hence not flexible (2–3 RE at most, maybe only 1)

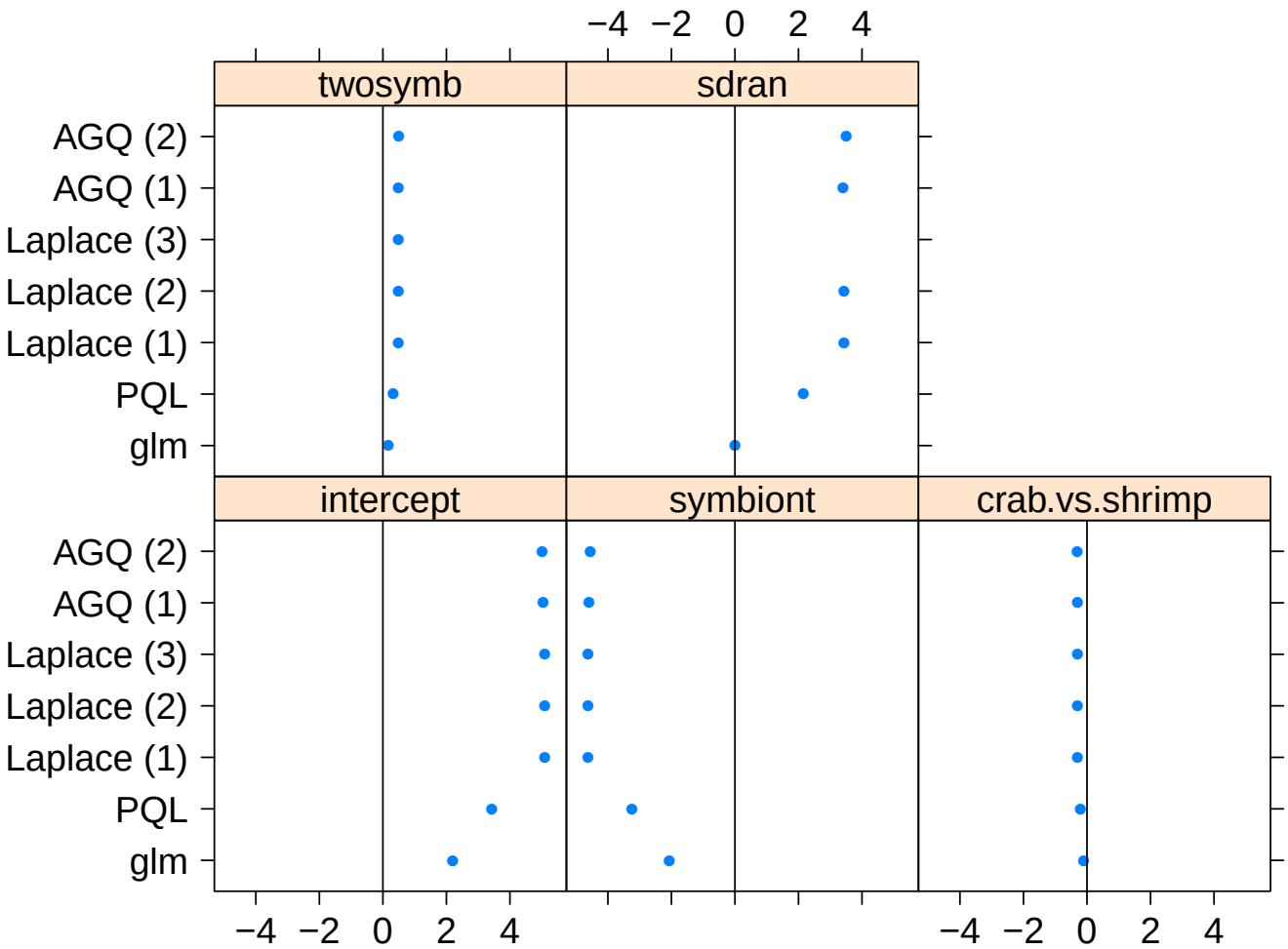
Becoming available: R `lme4`, SAS `PROC NLMIXED`, `PROC GLIMMIX` (v. 9.2), Genstat `GLMM`

Better methods

- **Laplace approximation**
  - approximate marginal likelihood
  - considerably more accurate than PQL
  - reasonably fast and flexible
- **adaptive Gauss-Hermite quadrature** (AGQ)
  - 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`

Comparison of coral symbiont results



Outline

General issues: testing RE significance

- Counting "model" df for REs
  - how many parameters does a RE require? Somewhere between 1 and n . . . Hard to compute, and depends on the **level of focus** (Vaida and Blanchard, 2005)
- Boundary effects for RE testing
  - most tests depend on null hypothesis being **within** the parameter's feasible range (Molenberghs and Verbeke, 2007): **violated** by $H_0 : \sigma^2 = 0$
  - REs may count for < 1 df (typically ≈ 0.5)
  - if ignored, tests are conservative

General issues: testing RE significance

- Counting "model" df for REs
  - how many parameters does a RE require? Somewhere between 1 and n . . . Hard to compute, and depends on the **level of focus** (Vaida and Blanchard, 2005)
- Boundary effects for RE testing
  - most tests depend on null hypothesis being **within** the parameter's feasible range (Molenberghs and Verbeke, 2007): **violated** by $H_0 : \sigma^2 = 0$
  - REs may count for < 1 df (typically ≈ 0.5)
  - if ignored, tests are conservative

General issues: finite-sample issues (!)

How far are we from "asymptopia"?

- Many standard procedures are **asymptotic**
- "Sample size" may refer the number of RE **units** — often far more restricted than total number of data points
- Hard to count degrees of freedom for complex designs: **Kenward-Roger correction**

General issues: finite-sample issues (!)

How far are we from "asymptopia"?

- Many standard procedures are **asymptotic**
- "Sample size" may refer the number of RE **units** — often far more restricted than total number of data points
- Hard to count degrees of freedom for complex designs: **Kenward-Roger correction**

General issues: finite-sample issues (!)

How far are we from "asymptopia"?

- Many standard procedures are **asymptotic**
- "Sample size" may refer the number of RE **units** — often far more restricted than total number of data points
- Hard to count degrees of freedom for complex designs: **Kenward-Roger correction**

Specific procedures

- Likelihood Ratio Test:
  need large sample size (= large # of RE units!)
- Wald (Z, χ^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 (properties unknown)
  - could use AIC_c, but ? need residual df

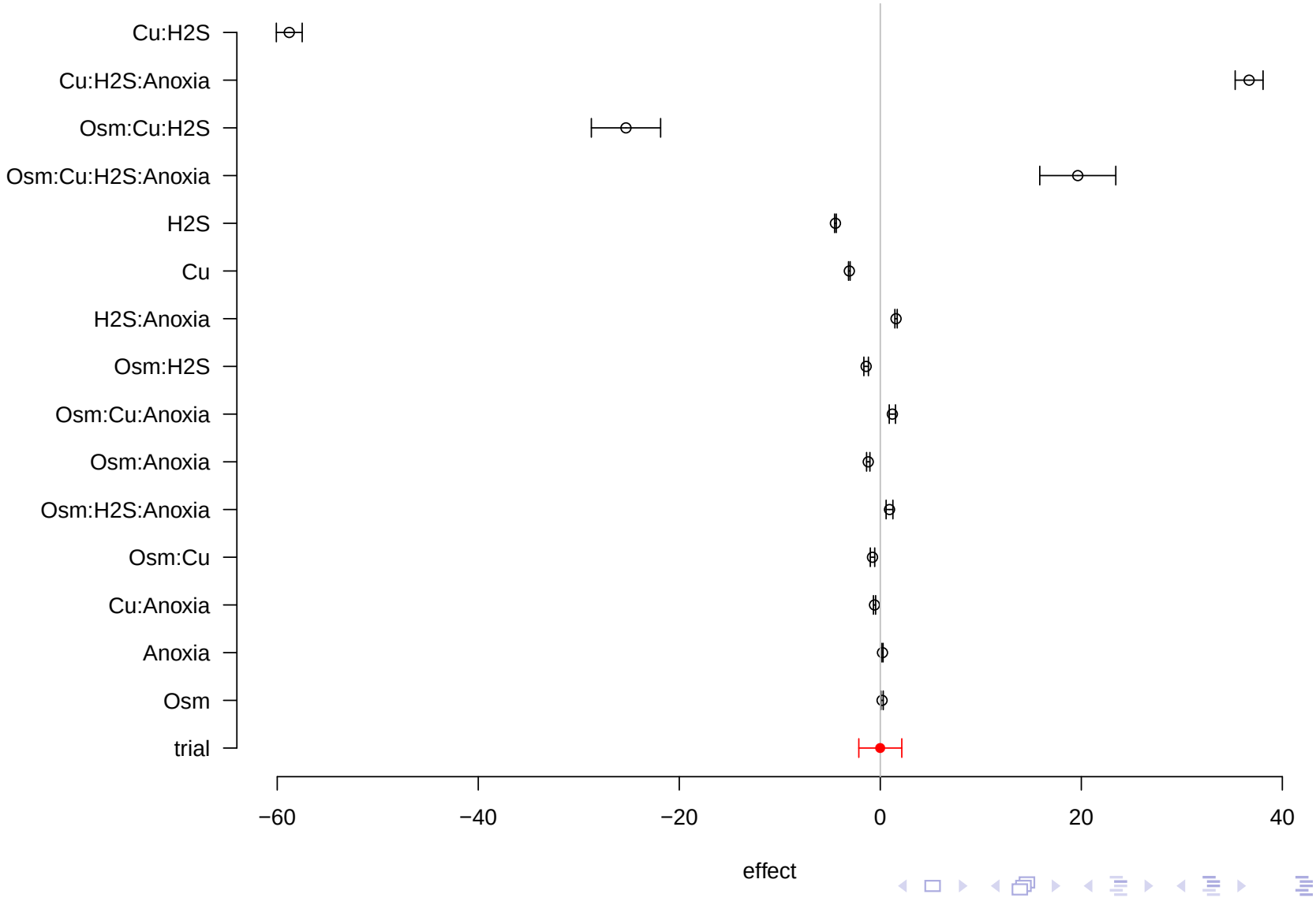
Specific procedures

- Likelihood Ratio Test:
  need large sample size (= large # of RE units!)
- Wald (Z, χ^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 (properties unknown)
  - could use AIC_c, but ? need residual df

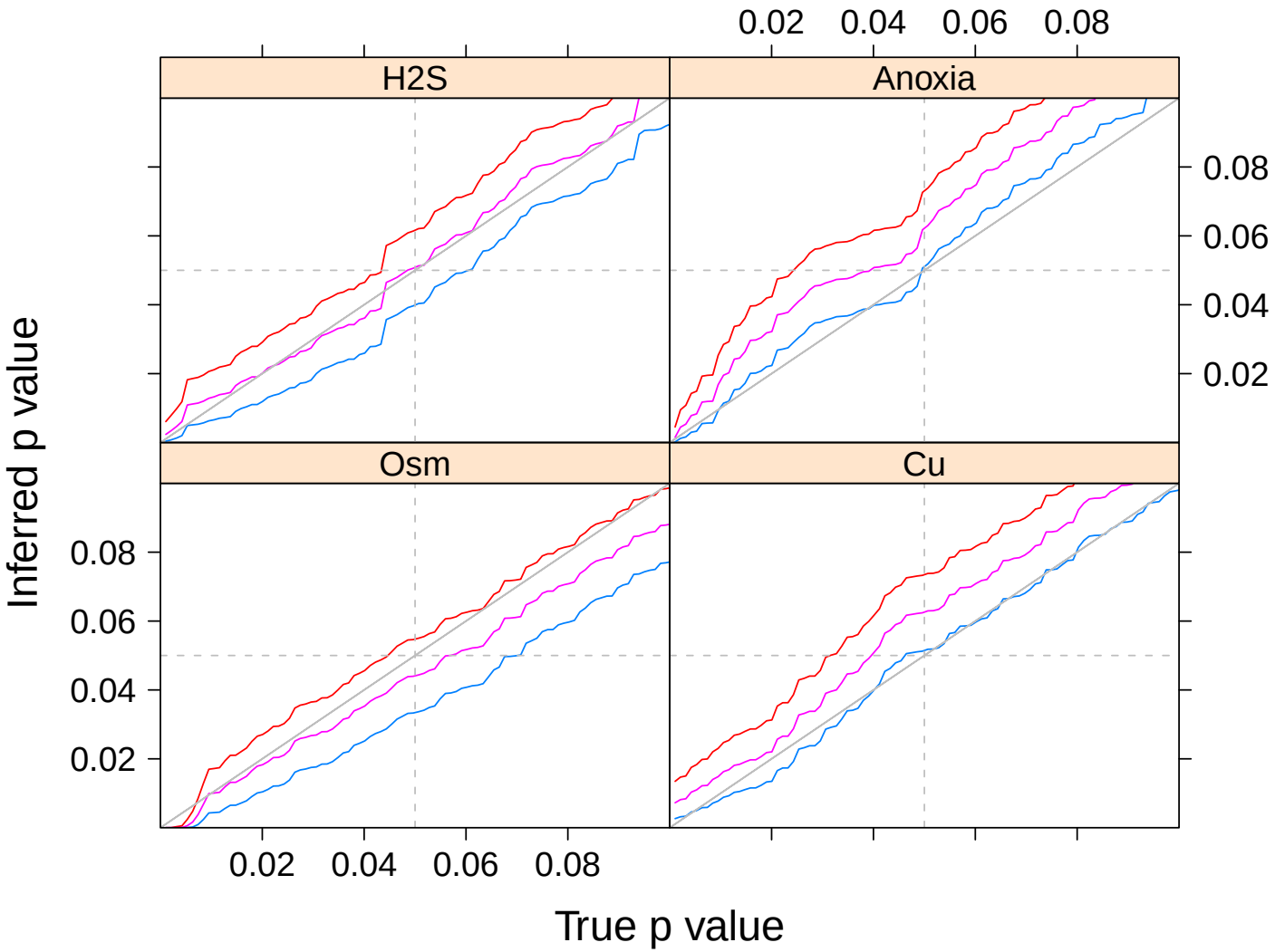
Specific procedures

- Likelihood Ratio Test:
  need large sample size (= large # of RE units!)
- Wald (Z, χ^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 (properties unknown)
  - could use AIC_c, but ? need residual df

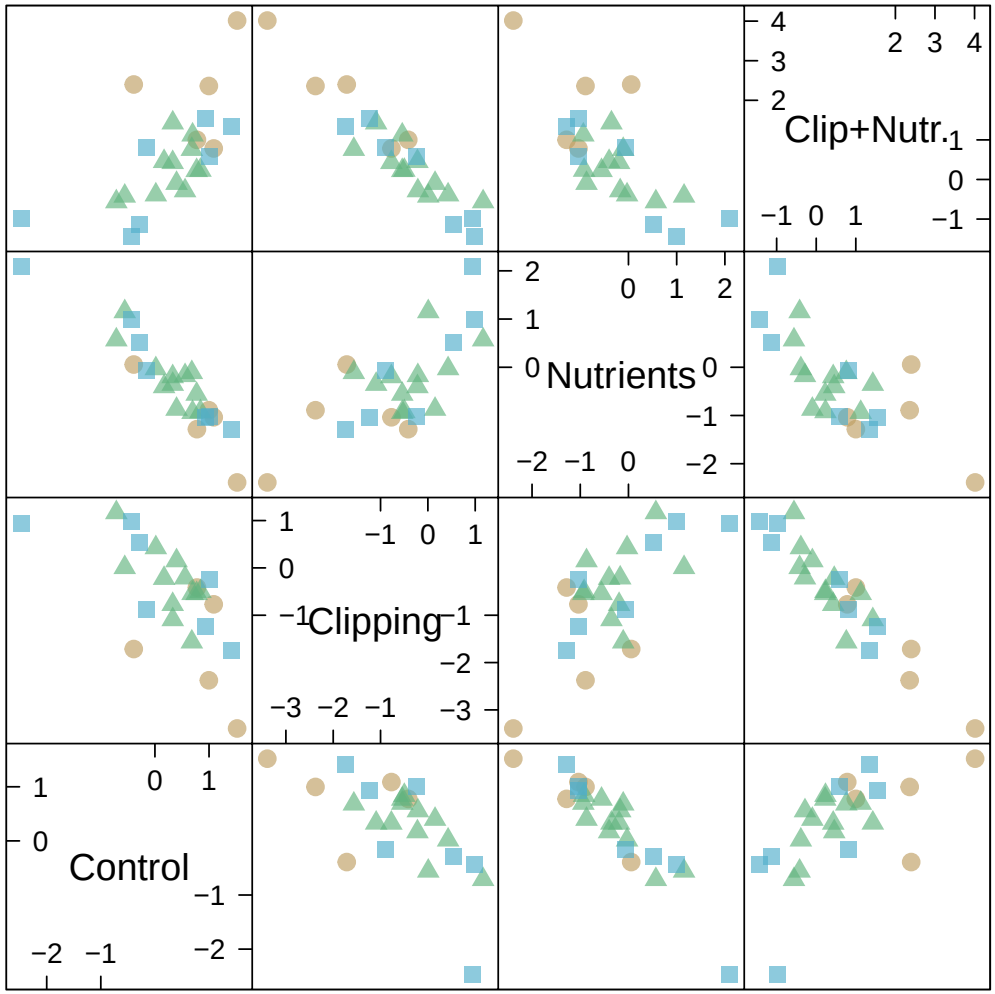
Glycera results



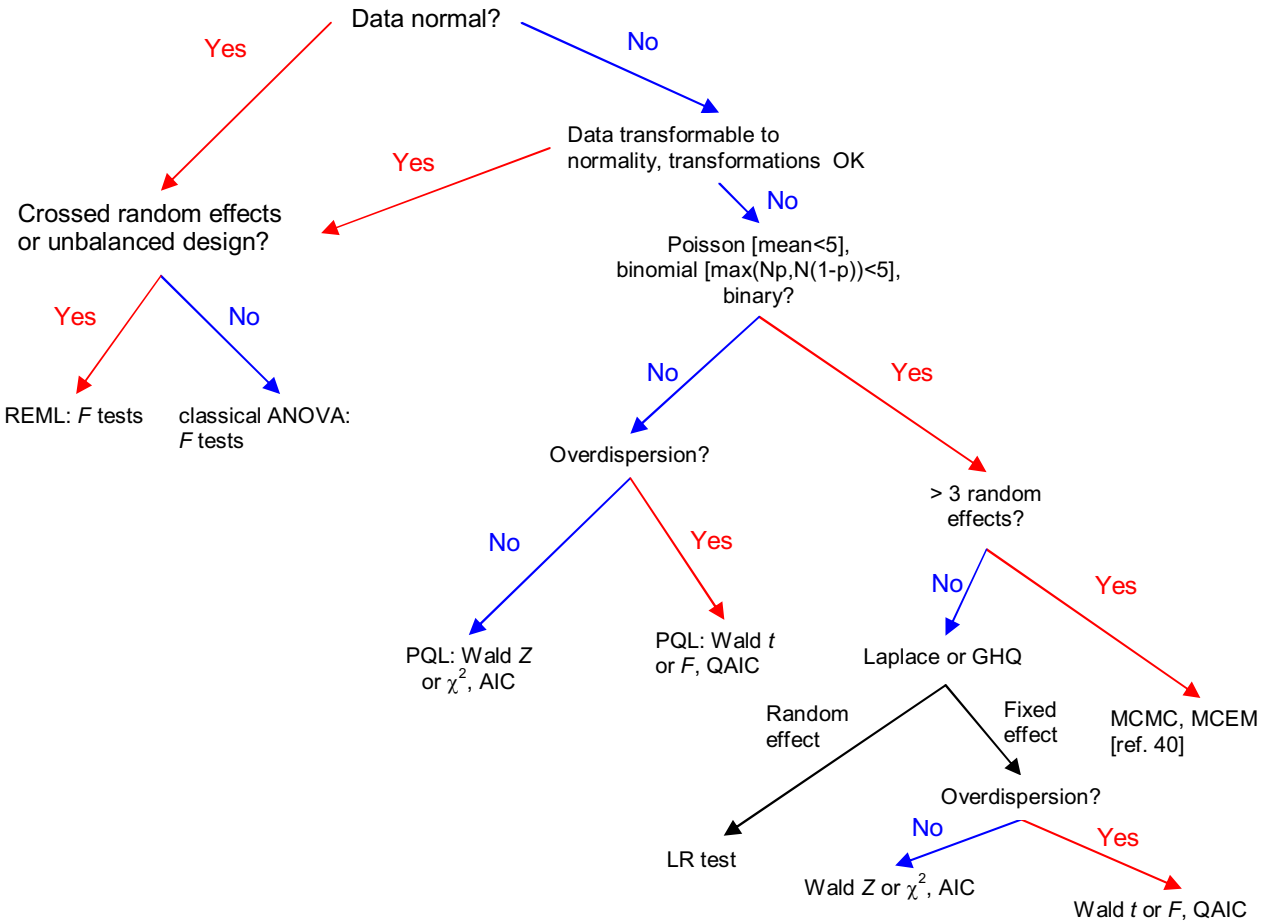
Testing assumptions



Arabidopsis genotype effects



Where are we?


Data normal?
Yes
No
Data transformable to normality, transformations OK
Yes
No
Crossed random effects or unbalanced design?
Poisson [mean<5], binomial [max(Np,N(1-p))<5], binary?
Yes
No
No
Yes
REML: F tests
classical ANOVA: F tests
Overdispersion?
> 3 random effects?
No
Yes
No
Yes
PQL: Wald Z or χ^2, AIC
PQL: Wald t or F, QAIC
Laplace or GHQ
MCMC, MCEM [ref. 40]
Random effect
Fixed effect
Overdispersion?
LR test
No
Yes
Wald Z or χ^2, AIC
Wald t or F, QAIC

Now what?

- MCMC (finicky, slow, dangerous, we have to "go Bayesian": specialized procedures for GLMMs, or WinBUGS translators? (`glmmBUGS`, `MCMCglmm`)
- quasi-Bayes `mcmcsamp` 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
  - ? analogue for confidence intervals?
- challenges depend on data: total size, # REs, # RE units, overdispersion, design complexity ...

More info: `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 `mcmcsamp` 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
  - ? analogue for confidence intervals?
- challenges depend on data: total size, # REs, # RE units, overdispersion, design complexity ...

More info: `glmm.wikidot.com`

Acknowledgements

- Data: Josh Banta and Massimo Pigliucci (*Arabidopsis*); Adrian Stier and Sea McKeon (coral symbionts); Courtney Kagan, Jocelynn Ortega, David Julian (*Glycera*);
- 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.