https://gitlab.com/swejk12/esa-workshop

Generalized Linear Mixed Effects Modeling Deep Dive:

Selecting Conditional Distributions, Checking Assumptions, and Drawing Inference for Ecological Data

Leigh Ann Starcevich, Jared Swenson, and Laura Martinez-Steele

https://gitlab.com/swejk12/esa-workshop

Introductions

Leigh Ann Starcevich
Principal Statistician

Laura Martinez-Steele
Associate Statistician

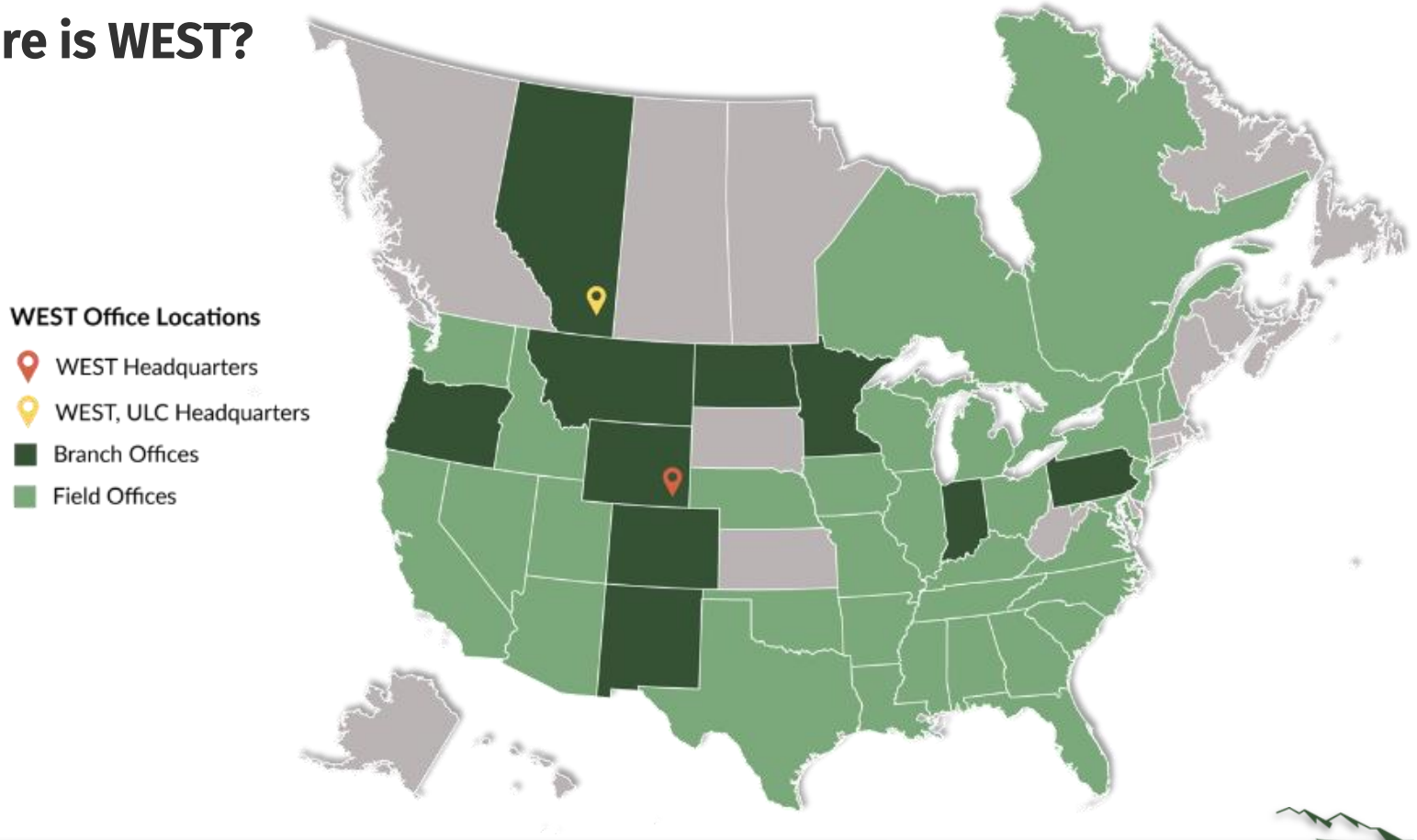
Jared Swenson
Associate Statistician

WEST

Who is WEST?

https://gitlab.com/swejk12/esa-workshop

Where is WEST?



WEST

https://gitlab.com/swejk12/esa-workshop

Our Expertise

- Aerial Surveys
- Bat and Avian Surveys
- Bird and Bat Conservation Plans
- Big Game and Mammal Surveys
- Collision Risk Modeling
- Ecosystems Study Design
- Statistical Analyses
- Estimation of Biological Size
- Resource Selection Studies
- Eagle Conservation Plans
- Evaluation of Mitigation Measures Effectiveness
- Expert Testimony
- Fatality Monitoring Studies
- Habitat Conservation Plans
- Impact Assessments and more...

WEST

https://gitlab.com/swejk12/esa-workshop

Outline

Objectives

Build

- Use graphical tools to determine *candidate* conditional distributions
- Specify random and fixed effects appropriately based on study design

Assess

- Use DHARMa package to assess residuals
- Understand application of REML and ML
- Understand uses of AIC/BIC

Draw Conclusions

- Use graphical tools to display results from a model
- Interpret results from a model to make biological conclusions.
- Understand variance structure and what it says about your system

WEST

Generalized Linear Mixed Effects Modeling Basics

Although selection procedures are helpful exploratory tools, the model-building process should utilize theory and common sense

- Agresti 2002

There is no consensus in the statistics community about what constitutes correct practice, and there likely never will be. On the other hand, there is broad consensus on what constitutes poor practice...

- Tredennick et al. 2021

Whereas GLMMs themselves are uncontroversial, describing how to use them to analyze data necessarily touches on controversial statistical issues. We acknowledge the difficulty while remaining agnostic.

.....different methods are appropriate for different problems .... how one analyzes data depends strongly on one's philosophical approach.

- Bolker et al. 2008

What is a Generalized Linear Mixed Model?

Mixed model
(incorporates fixed and random effects)

+

Generalized linear model
(non-normal data, link functions)

Why Random Effects?

VARIATION

STRUCTURE

temporal and spatial, account for replication, etc

Biological Conclusions and GLMM

Draw biological conclusions from estimates and confidence intervals to explain or understand relationships in ecological systems.

Estimate Parameters

- Covariate effects and interactions
- Treatment effects
- Quantify variation
- Useful in exploration, inference, and prediction

Exploration/Ecological Relationships

- Describe patterns and develop hypothesis about the natural world
- Trade-offs: Including variables of interest vs. spurious relationships
- Type-I errors (false discoveries)
- Make use of biological intuition

Inference

- Relationships/associations between response and covariates
- Evaluate strength of evidence for patterns
- Null-hypothesis testing
- Fewer models: lower risk of Type-I errors

Prediction

- Overlaps with exploration and inference
  - Model that best explains process should improve prediction, right?
- Focused on predicting the mean

Tredennick et al. 2021

https://gitlab.com/swejk12/esa-workshop

What Makes it a "Mixed" Model?

"nuisance variable", if it contributes a lot to variation, "unmeasured variation that needs to be accounted for"

	Fixed Effects	Random Effects
Use:	Test statistical significance or relationship	*Don't want to test statistical significance* Test another effect across all levels
Models:	Mean structure	Variance structure
Levels of a variable:	Fully represented in data Ex) *All measurements have an associated habitat or island of covariate that we can assess a relationship*	Treated as from a population effects *Think block design or repeat measures. Samples from a greater population*
Numbers of levels:	Finite	Many
Basis for inference:	Levels of the variable present in the data	Larger population

nesting of sites within islands, include island

Hierarchy and correlation of your variables

WEST

GLMM Components

Conditional Distribution

Link Function

Random Effect Structure

Fixed Effects Structure

Simple Linear Regression

- $y = \beta_0 + \beta_1 x + \varepsilon$
  - y is the dependent variable
  - β_0 and β_1 are regression coefficients define mean of y
  - x is the independent linear predictor
  - ε is the error term
- Fixed effects: β_0, β_1
  - Define the mean of y
- Random effect: ε, where $\varepsilon \sim N(0, \sigma^2)$ doesnt effect mean structure at all, but does impact variance
  - Defines the variance of y

Breaking Down Random Effects

transformed mean, E is the espected mean, conditioning on u, random

Simple Linear Regression: $y = \beta_0 + \beta_1 x + \varepsilon$

GLMM: $g\left[E\left(y_{ijk}|\boldsymbol{u}_{ijk}\right)\right] = \beta_0 + \beta_1 w_j + b_j + a_i + t_i w_j + c_{ijk}$

- Random year effect: $b_j \sim \mathrm{N}\left(0, \sigma_b^2\right)$ normal with mean 0, and year to year variation
- Random site effects: $\begin{pmatrix} a_i \\ t_i \end{pmatrix} \sim \mathrm{MVN}\begin{pmatrix} 0 & \sigma_a^2 & \sigma_{at} \\ 0' & \sigma_{at} & \sigma_t^2 \end{pmatrix}$ correlated site effects, t is random site slope giving trendline of each site
- Random site-by-year interaction effect: $c_{ijk} \sim \mathrm{N}(0, \sigma_c^2)$ same site within a year, variation accounted for
- $\sigma_b^2, \sigma_a^2, \sigma_t^2,$ and σ_c^2 are variance components

Piepho and Ogutu (2002)

WEST

GLMM Form

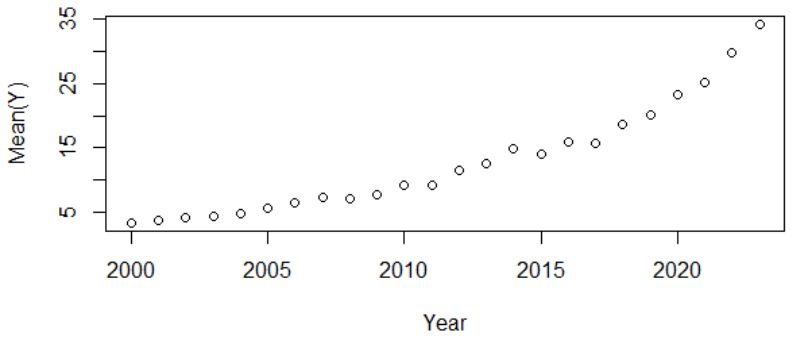
$$g[E(\boldsymbol{y}|\boldsymbol{u})] = \boldsymbol{\eta} = \boldsymbol{\beta X} + \boldsymbol{Zu}$$

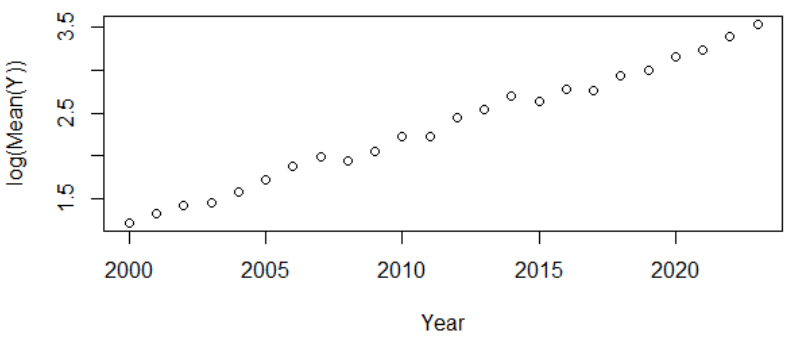
- $g(.)$ is the link function (*e.g.* log, logit) modeling the mean of y,
- E is the expectation for the conditional distribution of $\boldsymbol{y}$
- $\boldsymbol{\eta}$ is the linear predictor point of link is to get linear relationships for this model
- Fixed effects and random effects

If cant get linear, its time for GAM

Link Function

- Transforms the expected value of the response variable so that relationships with predictors are **linear**
- Common link functions:
  - Gaussian: identity (LMM)
  - Counts, non-negative continuous data: log
  - Proportions: logit, probit, log-log, clog-log





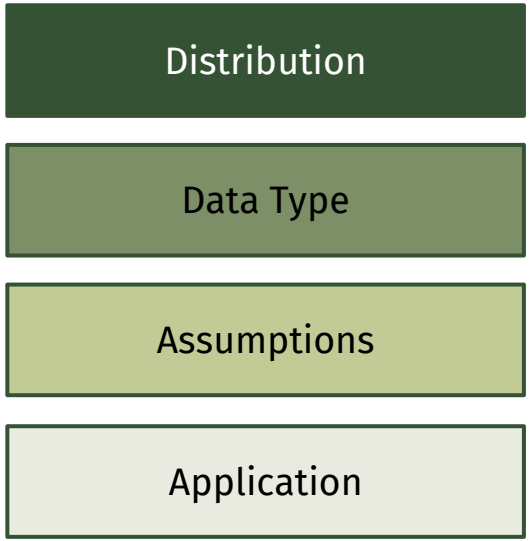
GLMM Components

Conditional Distribution	Link Function
Based on data structure Exponential family	Transforms the expected value of the response variable so that relationships are linear
Random Effect Structure	Fixed Effects Structure

Conditional Distributions of the Response

- Also called “error distributions”
- Exponential family
  - Poisson
  - Negative binomial (two forms)
  - Gamma
  - Tweedie “amazing”
  - Binomial
  - Beta

- glmmTMB package also includes mixture and truncated distributions
  - Zero inflation (excess zeros)
  - Zero truncation (no zeros)



Conditional Distributions of the Response

Binomial	Beta	Beta-Binomial
Binary, presence-absence, proportions	Proportion: not based on Bernouilli trials	Overdispersed or correlated binomial data: presence/absence
Independent trials with a binary outcome	Two scale parameters	Binomial probability is a Beta random variable
Family = binomial **Link = logit, probit, log-log, clog-log**	Family = beta **Link = logit, probit, log-log, clog-log**	Family = betabinomial **Link = logit, probit, log-log, clog-log**

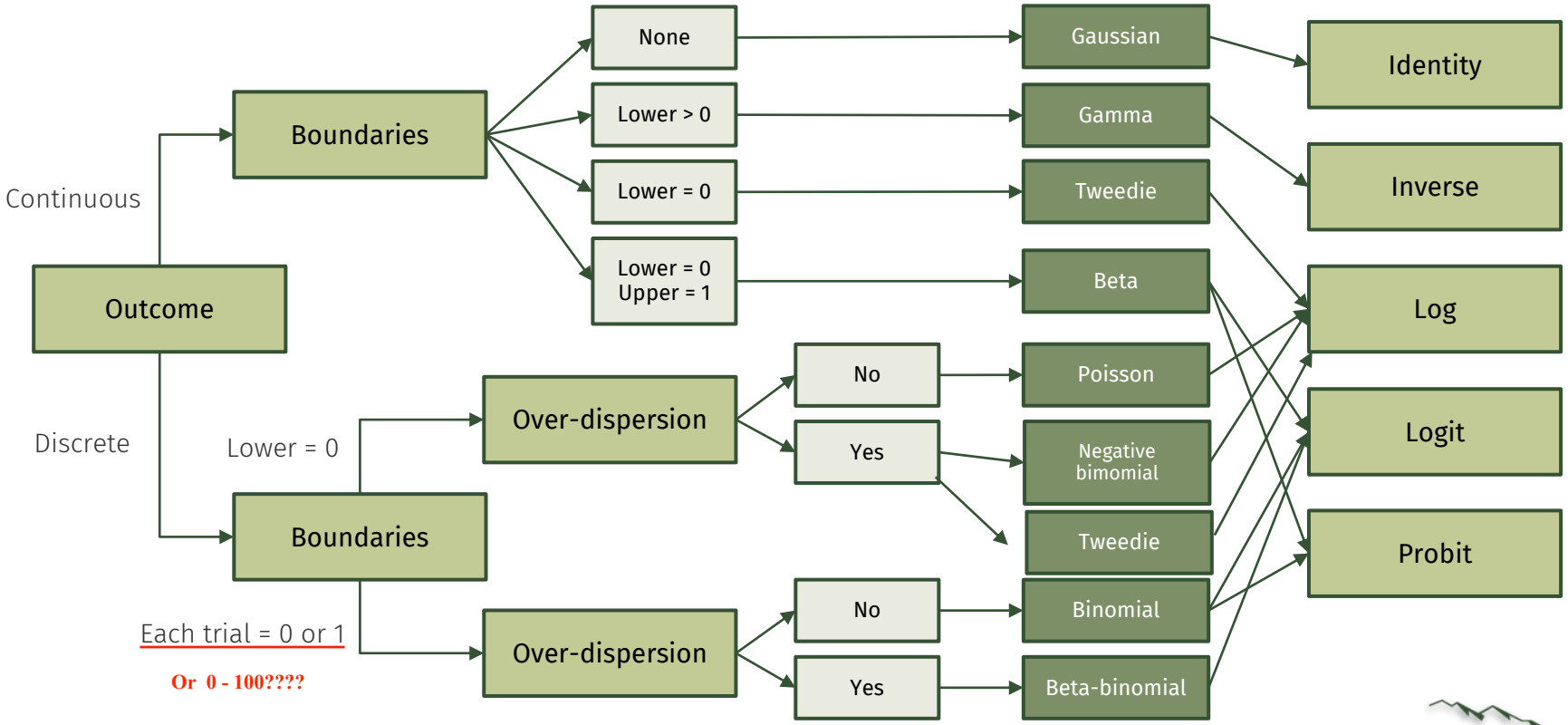
Conditional Distributions of the Response

Poisson	Negative Binomial
Counts: abundance	Counts with overdispersion: abundance
Variance is equal to the mean	Relationship between variance and mean: *nbinom1*: linear *nbinom2*: quadratic
Family = poisson Link = log, identity	Family = nbinom1, nbinom2 Link = log, identity

Conditional Distributions of the Response

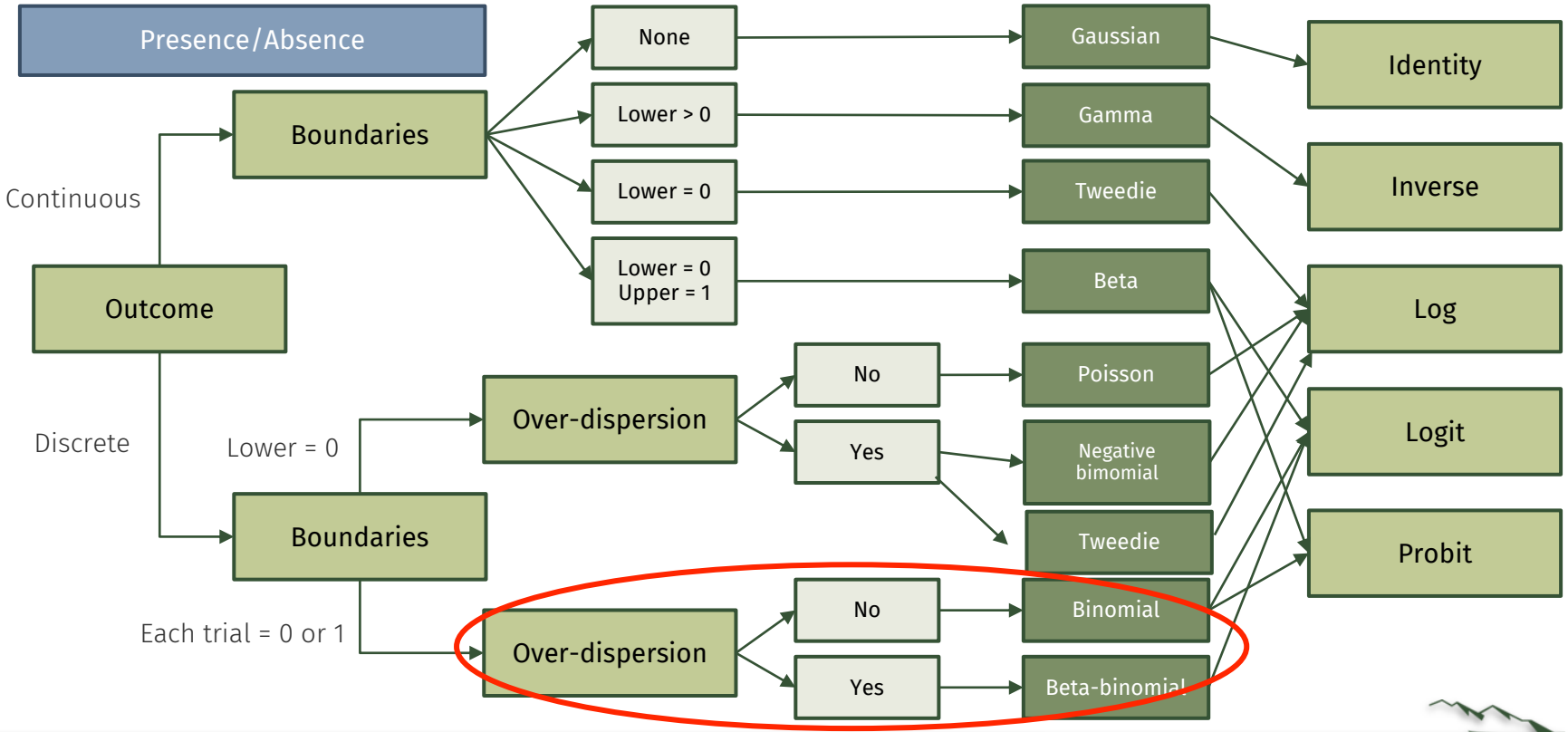
Gaussian	Gamma	Tweedie
Continuous metric: temperature, time	Positive continuous data: time	Continuous or count, may contain zeros: abundance, weight, ...
Linear mixed model Variance is not a function of the mean	Mean is linear in scale parameter, variance is quadratic in scale parameter	Useful for continuous data with zeros
Family = gaussian Link = identity	Family = gamma Link = inverse	Family = tweedie Link = log

Conditional Distributions and Link Functions



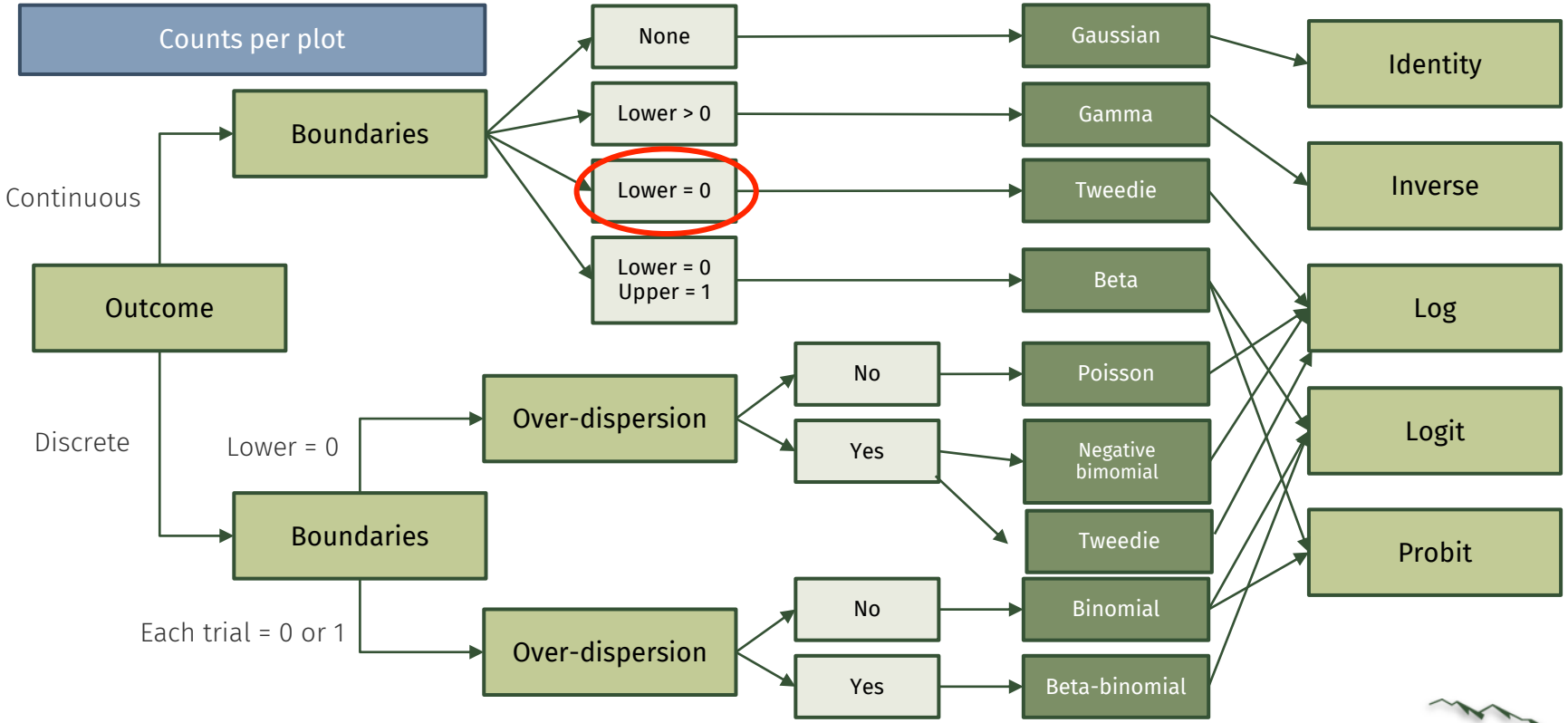
https://gitlab.com/swejk12/esa-workshop

Conditional Distributions and Link Functions



WEST

Conditional Distributions and Link Functions



WEST

GLMM Components

Conditional Distribution	Link Function
Based on data structure Exponential family	Transforms the expected value of the response variable so that relationships are linear
Random Effect Structure	Fixed Effects Structure
Reflects study design, correlation structure	Predictors of interest

Specifying Random Effects in R

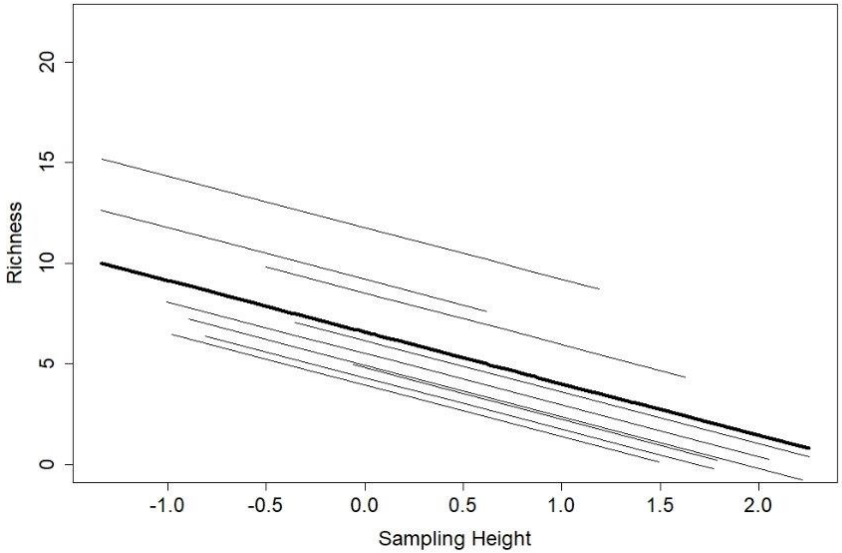
Random intercept	Nested random effects
(1|Site)	(1|Site/Island)
Correlated random intercept (site) and slope (year)	Uncorrelated random intercept and Year slope
(1+WYear|Site)	(1+WYear||Site) *or* (1|Site) + (-1+WYear|Site)

https://gitlab.com/swejk12/esa-workshop

Random Intercept

Richness ~ Sampling Height + (1|Beach)

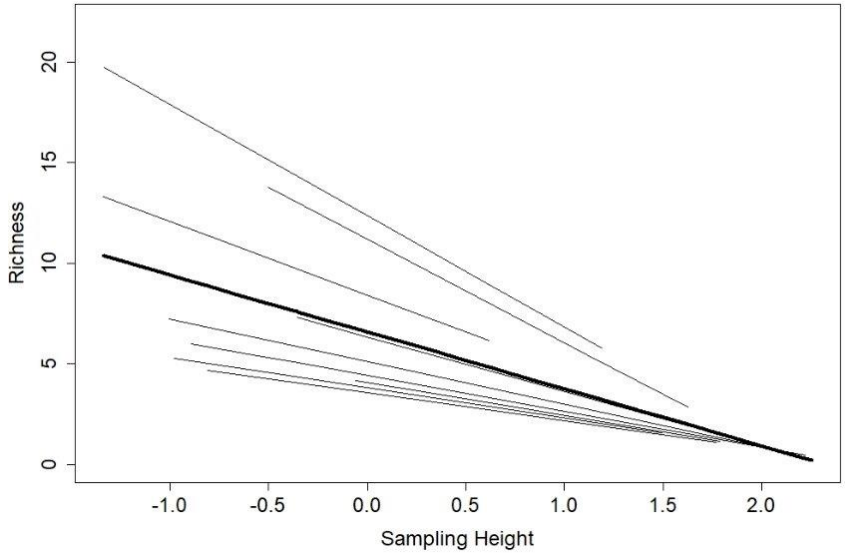
intercept varies slope does not



Random Slope and Intercept

Richness ~ Sampling Height + (1 + Sampling Height|Beach)

random slope by height and beach



Zuur et al. (2009)

WEST

Build and Assess

Conditional Distribution and Link

Determine family and link
Check for linearity & homogenous variance
Identify outliers

Check Model Assumptions

Specify and Fit the Full Model

Finalize Model

https://gitlab.com/swejk12/esa-workshop

Exercise 1:

Pick the *candidate* conditional distributions

Explore relationships and consider fixed and random effects for model fitting

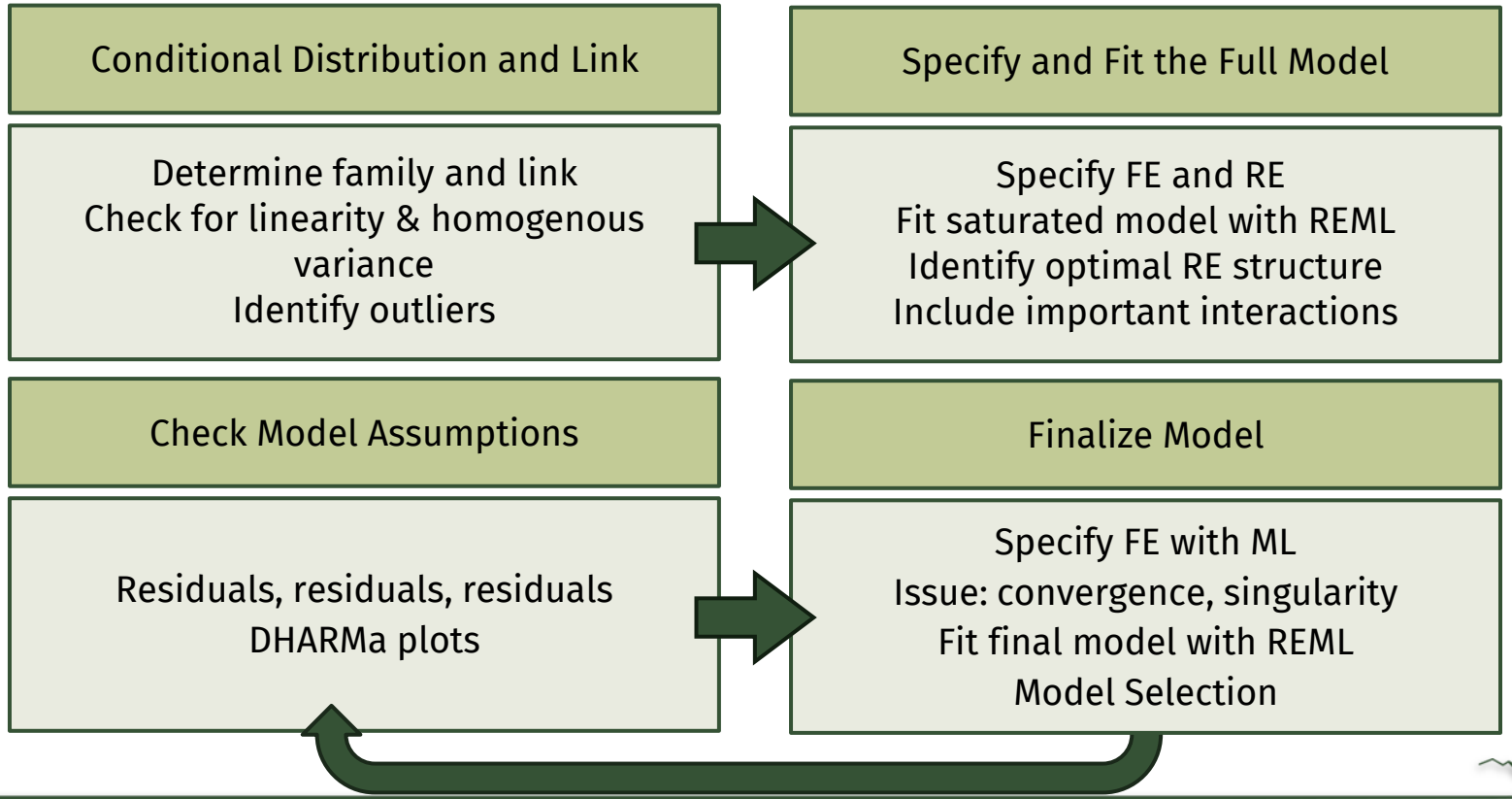
Objectives:

Use graphical tools to determine *candidate* conditional distributions

Examine data and variance structures to think about possible random effect structure

WEST

Build and Assess



Residual Diagnostics - "Goodness of fit" - DHARMa

Kolmogorov-Smirnov Test	Overdispersion	Underdispersion
Test of uniformity of residuals Does the assumed distribution match the data?	More variation than expected based on distribution ➡ **misspecified model** Underestimate FE SEs Inflated type-I errors	Less variation than expected based on distribution Low power to detect relationship

Visual detection or test (e.g. Pearson chi-squared)

Zero inflation	Heteroscedasticity	Residual correlation
Common cause of overdispersion Compare ZI family to alternative	Patterns in the residuals Check quantile deviations plot Some variation not accounted for in structure Missing predictors?	Spatial or temporal

Check Model Assumptions

Distribution and Link	Independence
Check linear relationship with link-transformed means	Partial autocorrelation plots of residuals
RE are Independent of Residuals	Normality of Random Effects
Pearson's correlation tests	Quantile-quantile plots and histograms of random effects

https://gitlab.com/swejk12/esa-workshop

Exercise 2:

Fit a GLMM that we expect to have variation across space and time.

Objectives:

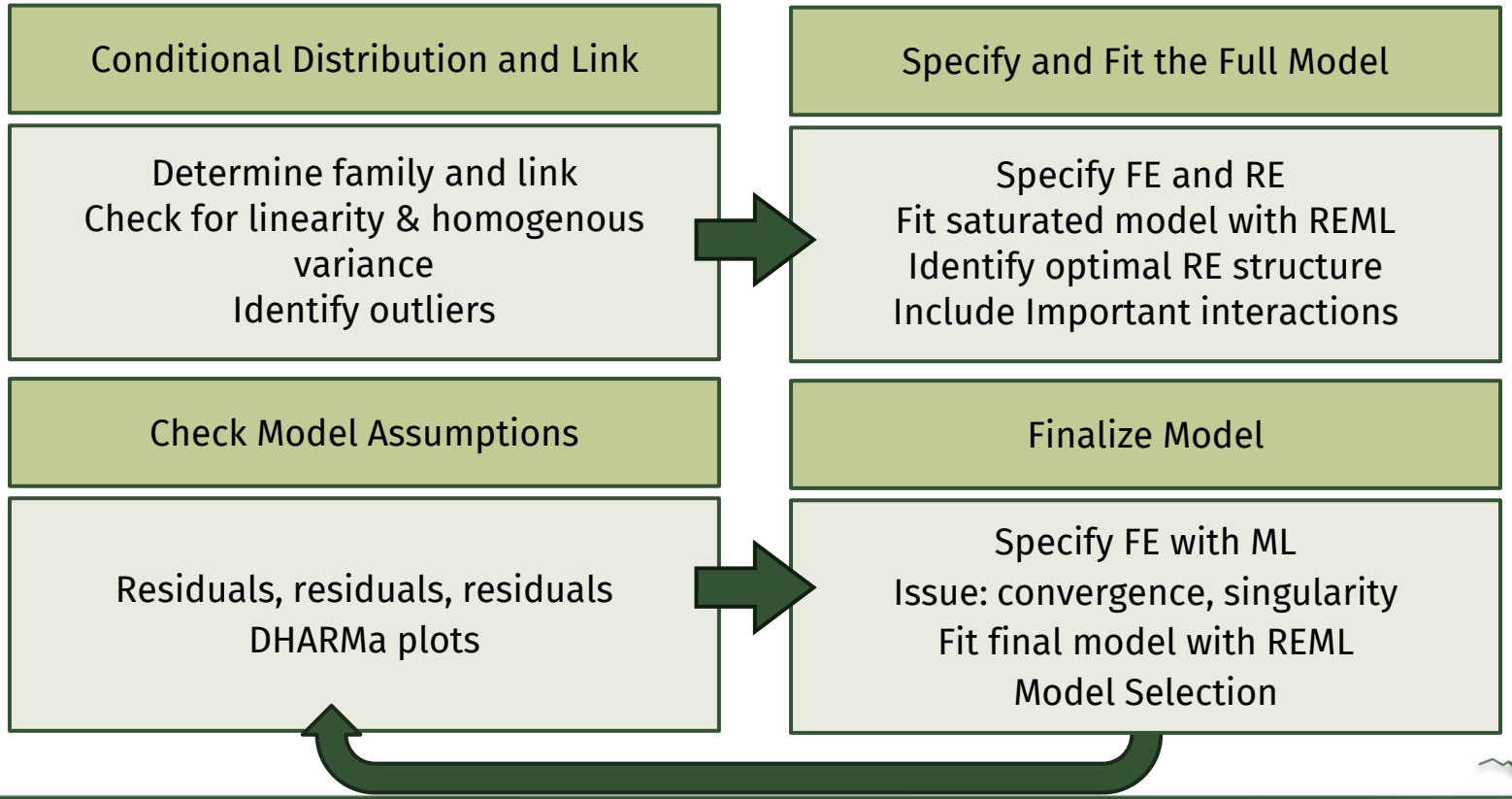
Identify the conditional distribution.

Specify random and fixed effects.

Compare a random intercept model to a random slope and intercept model.

Assess residuals using DHARMa.

Build and Assess



Exercise 3:

Incorporate design variables into random effects structure

Objectives:

Specify design variable in our model.

Assess residuals using DHARMa.

Specifying Random Effects in R (reminder)

Random intercept	Nested random effects
(1|Site)	(1|Site/Island)
Correlated random intercept and Year slope	Uncorrelated random intercept and Year slope
(1+WYear|Site)	(1+WYear||Site) *or* (1|Site) + (-1+WYear|Site)

Importance of Variance Components

- Decomposing the error into variance components tells us something about the response variable
    - High variation from year-to-year ⇒ Not a useful metric to monitor over time, requires long time periods to overcome noise in data and detect signal
    - High variation from site to site ⇒ Adequate sample size of sites needed
    - High variation among trend lines at a site ⇒ Possible subpopulations responding differently to stressors
    - High site-by-year interaction variance ⇒ need replication within a site and year (e.g., quarterly surveys), possible crew variability, may need to focus on a specific time period each year

AIC and BIC

- Both are used in model selection to assess model fit and complexity
- Fit based on likelihood
- Complexity is based on the number of parameters
- BIC also accounts for number of observations and penalizes more complexity
- AIC/BIC use:
  - X REML <--> ML
  - ✓ REML <--> REML
  - ✓ ML <--> ML

do not compare aics between reml and ml

- AIC is best for prediction because it might be more forgiving of spurious correlations than NHST. More covariates typically improve prediction (Tredennick et al. 2021)
- AIC can be adjusted for overdispersion (QAIC)
- Model selection with LRT could be an abuse of hypothesis testing (Bolker et al. 2008)
- Bayes factor can be used as LRT alternative. Outcome similar to recommendations from BIC

REML and ML

- REML handles variance better than ML
- ML produces more biased variance estimates
- Balanced vs. unbalanced
- Slope and intercept may be similar across models, but variance should differ
- AIC/BIC use:
  - ✗ REML <--> ML
  - ✓ REML <--> REML
  - ✓ ML <--> ML
- REML should be used for finalizing random effects
- ML should be used to finalize fixed effects
- Final model should be presented using REML.

Covariate and Model Selection

- Depends on the goal
- Adds interactions that are biologically meaningful
- Too many variables:
  - Remove variables that have correlation coefficient of >= 0.3
  - Biologically informed
  - Question/Objective informed
  - LASSO
- AIC/BIC/QAIC can be used for comparison of model
  - REML for Random Effects, ML for fixed effects
- Top-down approach for covariate selection
  - Fit full model first
- stats::drop1() function
  - Compares models by dropping covariates one at a time
  - Should use correction for multiple hypothesis tests

References

Agresti, A. 2011. Categorical Data Analysis (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.
Bolker, B. 2016. Getting started with the glmmTMB package. Vienna, Austria: R Foundation for Statistical Computing. Software.
Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. and White, J.S.S., 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in ecology & evolution, 24(3), pp.127-135.
Hartig, F., 2020. DHARMa: residual diagnostics for hierarchical (multi-level/mixed) regression models. R package version 0.3. https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMa.html.
McCulloch, C.E. and Searle, S.R., 2004. Generalized, linear, and mixed models. John Wiley & Sons.
Pinheiro, J.C. and Bates, D.M., 2000. Linear mixed-effects models: basic concepts and examples. Mixed-effects models in S and S-Plus, pp.3-56.
Tredennick, A.T., Hooker, G., Ellner, S.P. and Adler, P.B., 2021. A practical guide to selecting models for exploration, inference, and prediction in ecology. Ecology, 102(6), p.e03336.
Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A. and Smith, G.M., 2009. Mixed effects models and extensions in ecology with R (Vol. 574, p. 574). New York: Springer.

https://gitlab.com/swejk12/esa-workshop

lstarcevich@west-inc.com

jswenson@west-inc.com

lsteele@west-inc.com

WEST

west-inc.com

WEST Headquarters | 415 West 17th Street, Suite 200, Cheyenne, Wyoming 82001 | 307-634-1756

WEST, ULC Headquarters | Suite S138, 6715 8 Street NE, Calgary, Alberta T2E 7H7 | 587-432-3015