Course synthesis

Analysis of Ecological and Environmental Data

QERM 514

Mark Scheuerell

29 May 2020

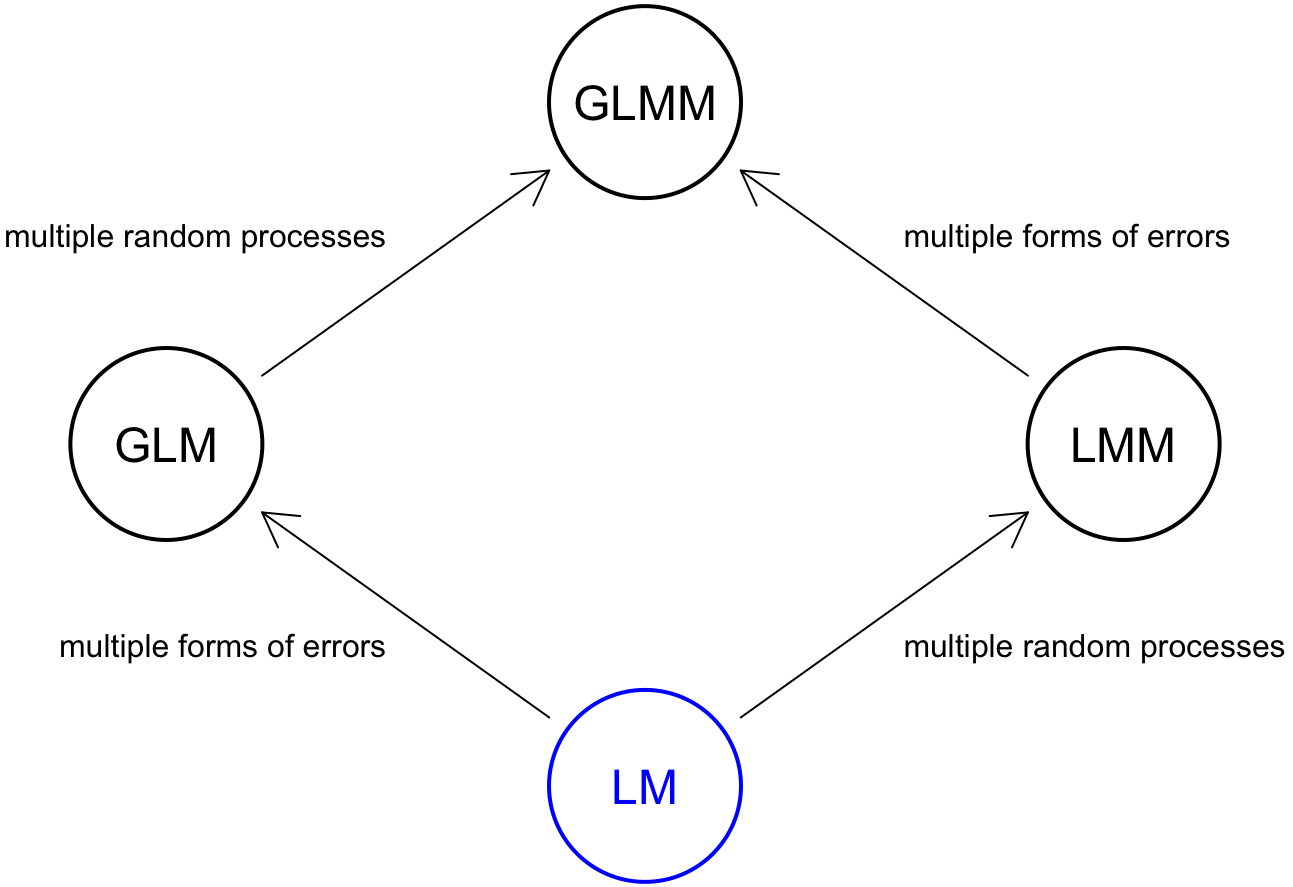
Goals for today

Sit back and reflect on how much you've learned

Learning objectives for the course

- Identify an appropriate statistical model based on the data and specific question
- Understand the assumptions behind a chosen statistical model
- Use **R** to fit a variety of linear models to data
- Evaluate data support for various models and select the most parsimonious model among them
- Use **R Markdown** to combine text, equations, code, tables, and figures into reports

Simple linear models

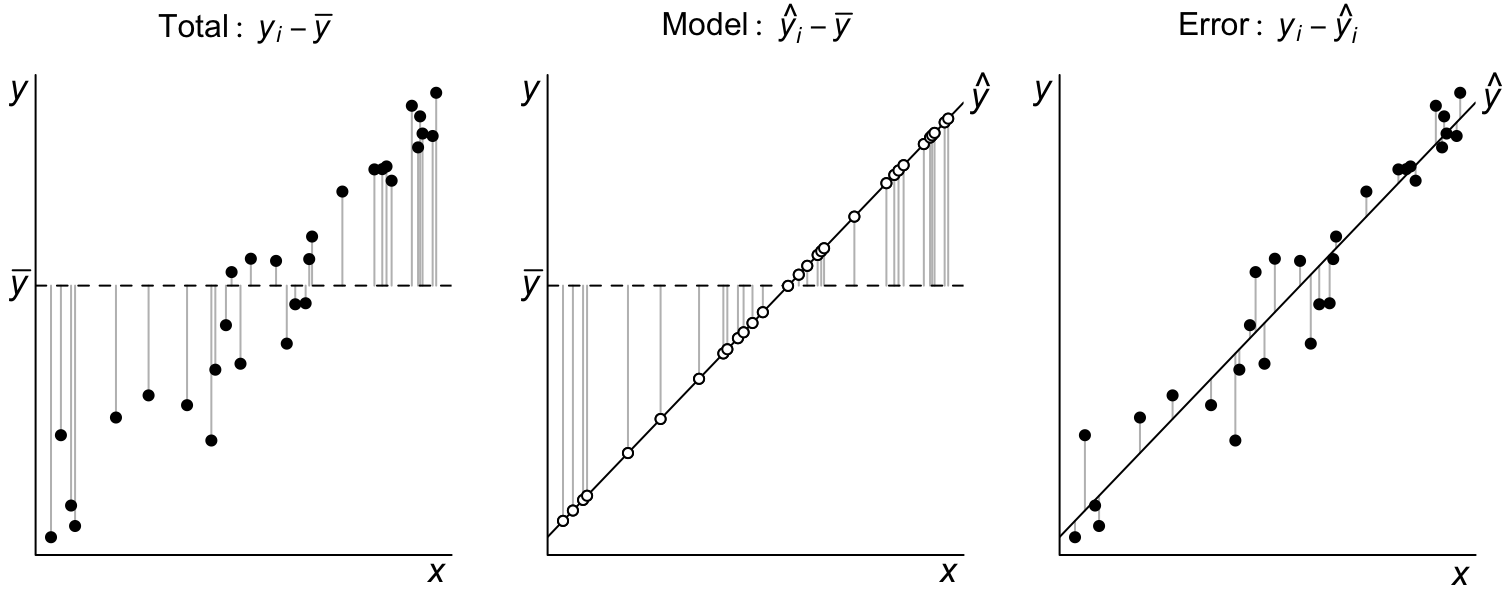


Partitioning total deviations

The total deviations in the data equal the sum of those for the model and errors

$$\underbrace{y_i - \bar{y}}_{\text{Total}} = \underbrace{\hat{y}_i - \bar{y}}_{\text{Model}} + \underbrace{y_i - \hat{y}_i}_{\text{Error}}$$

Partitioning total deviations



Partitioning sums-of-squares

The sums-of-squares have the same additive property as the deviations

$$\underbrace{\sum (y_i - \bar{y})^2}_{SSTO} = \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{SSR} + \underbrace{\sum (y_i - \hat{y}_i)^2}_{SSE}$$

Linear models in matrix form

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + e_i$$

$$\Downarrow$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} \\ 1 & x_{1,2} & x_{2,2} \\ \vdots & \vdots & \\ 1 & x_{1,N} & x_{2,N} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}$$

$$\Downarrow$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Ordinary least squares

Rewriting our model, we have

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}$$
$$\Downarrow$$
$$\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

so the sum of squared errors is

$$\mathbf{e}^\top \mathbf{e} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

Ordinary least squares

Minimizing the sum of squared errors leads to

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

$$\Downarrow$$

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

Variance estimates

Parameters

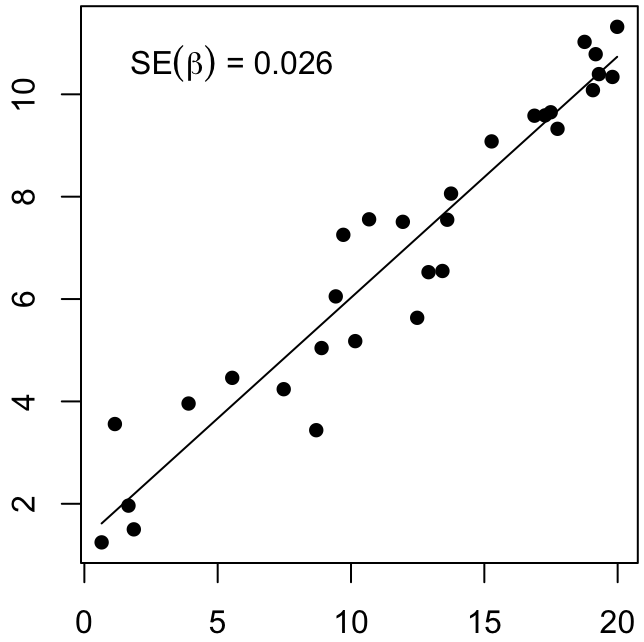
The variance of $\hat{\boldsymbol{\beta}}$ is given by

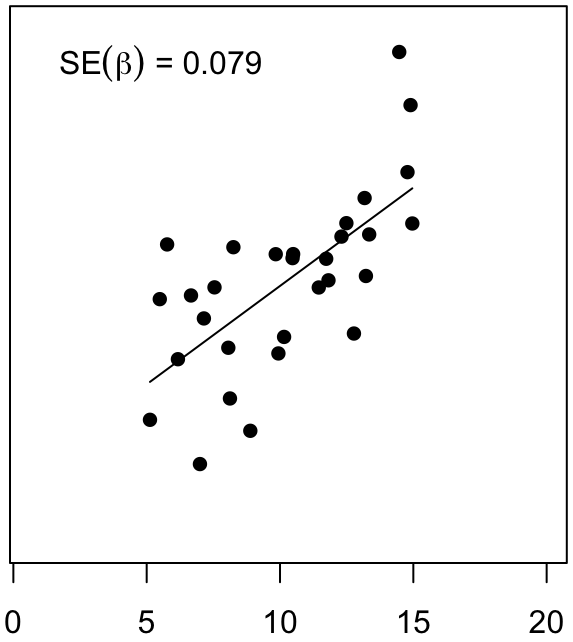
$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$$

This suggests that our confidence in our estimate increases with the spread in $\mathbf{X}$

Effect of $\mathbf{X}$ on parameter precision

Consider these two scenarios where the slope of the relationship is identical





CI for the mean response

A CI for the mean response is given by

$$\hat{\mathbf{y}}^* \pm t_{df}^{(\alpha/2)} \sigma \sqrt{\mathbf{X}^{*\top}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^*}$$

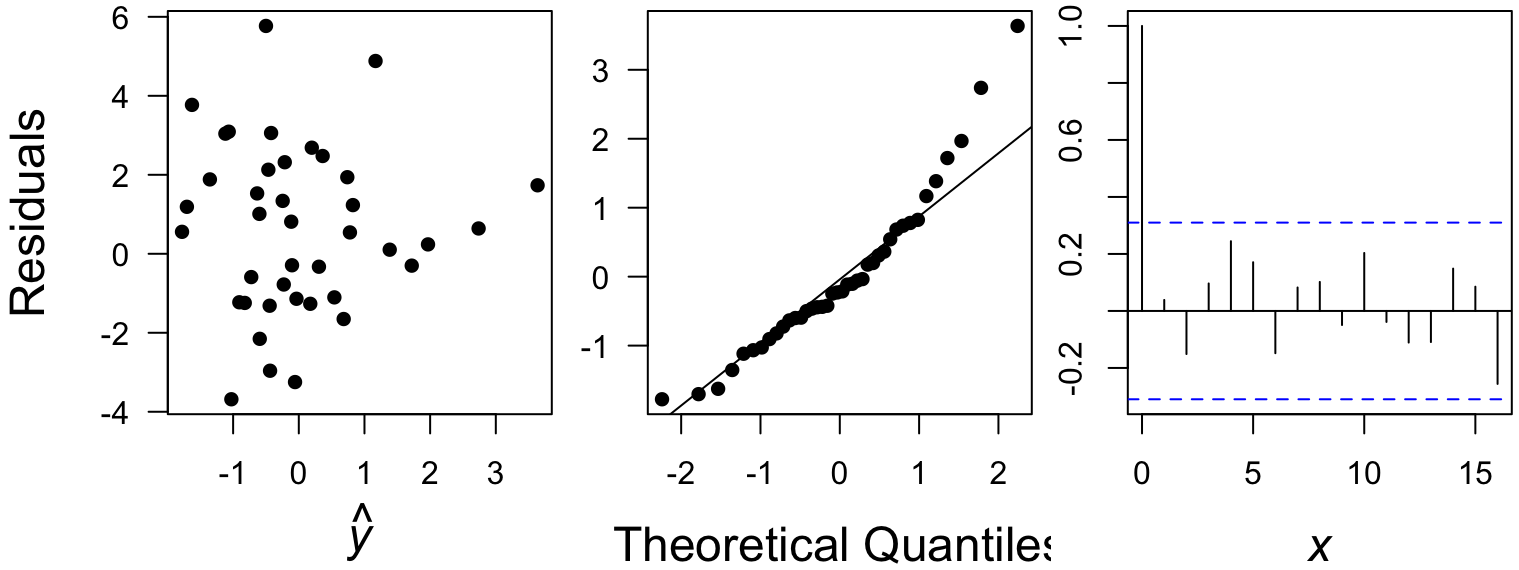
CI for a specific response

A CI on a new prediction is given by

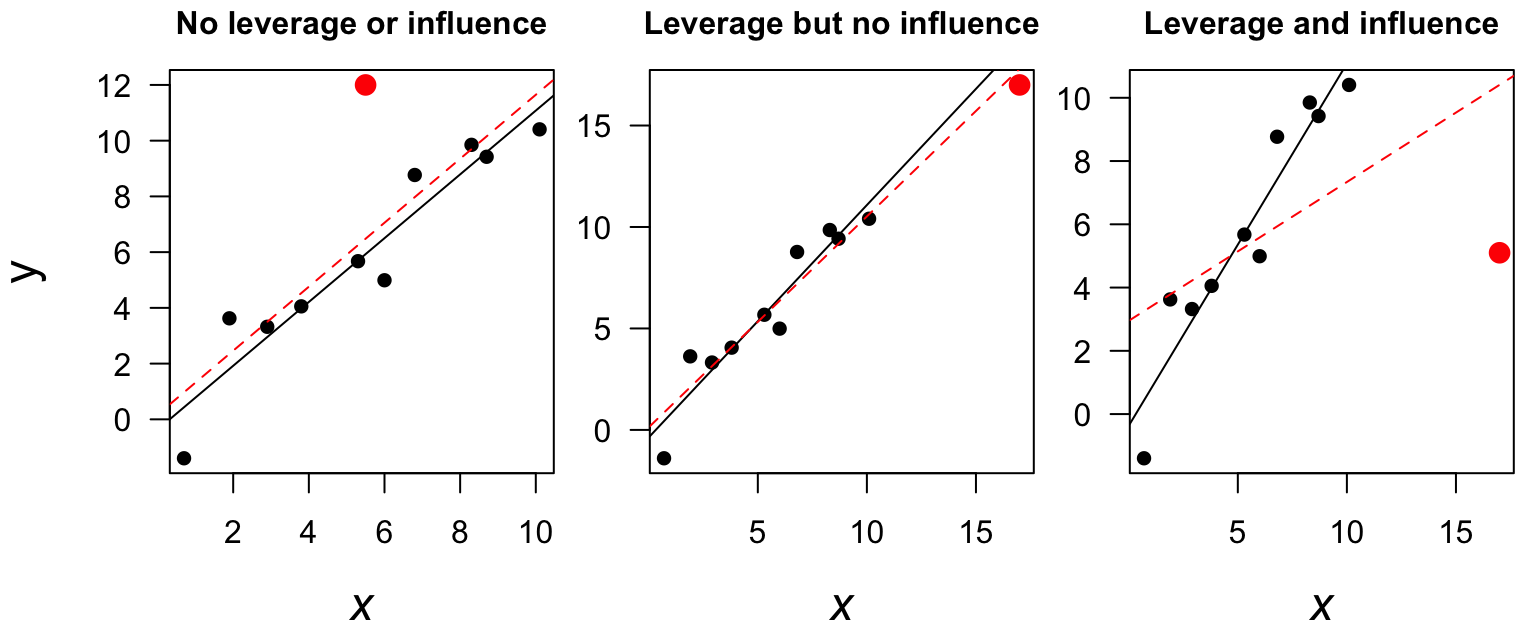
$$\hat{\mathbf{y}}^* \pm t_{df}^{(\alpha/2)} \sigma \sqrt{1 + \mathbf{X}^{*\top} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^*}$$

This is typically referred to as the *prediction interval*

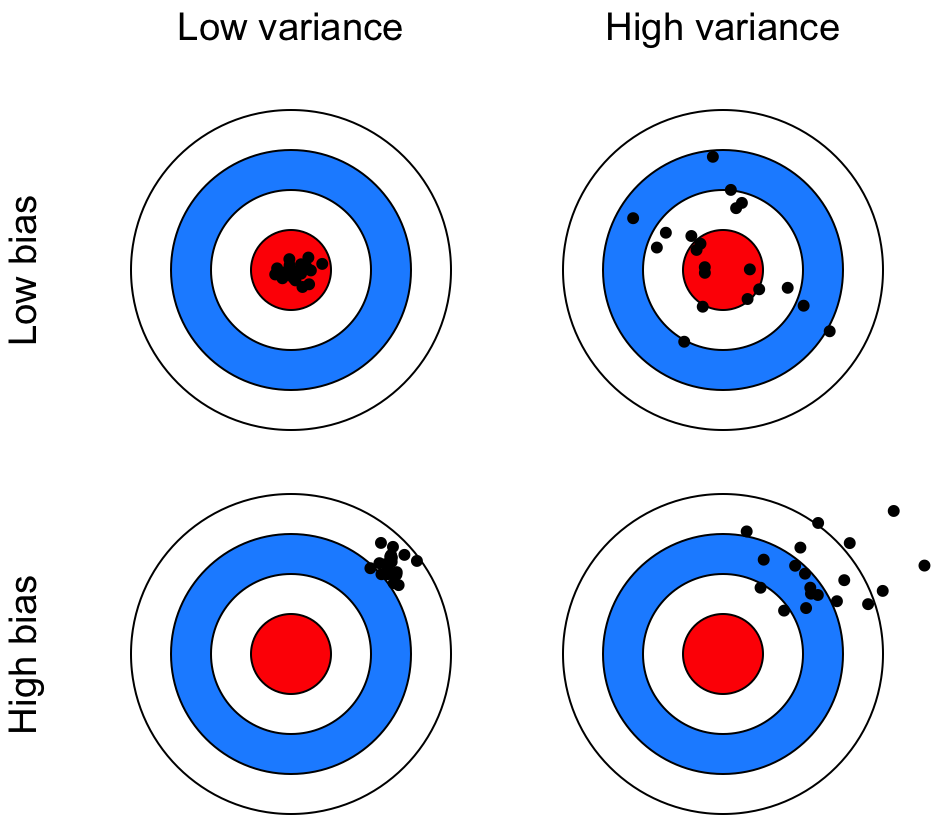
Diagnostics



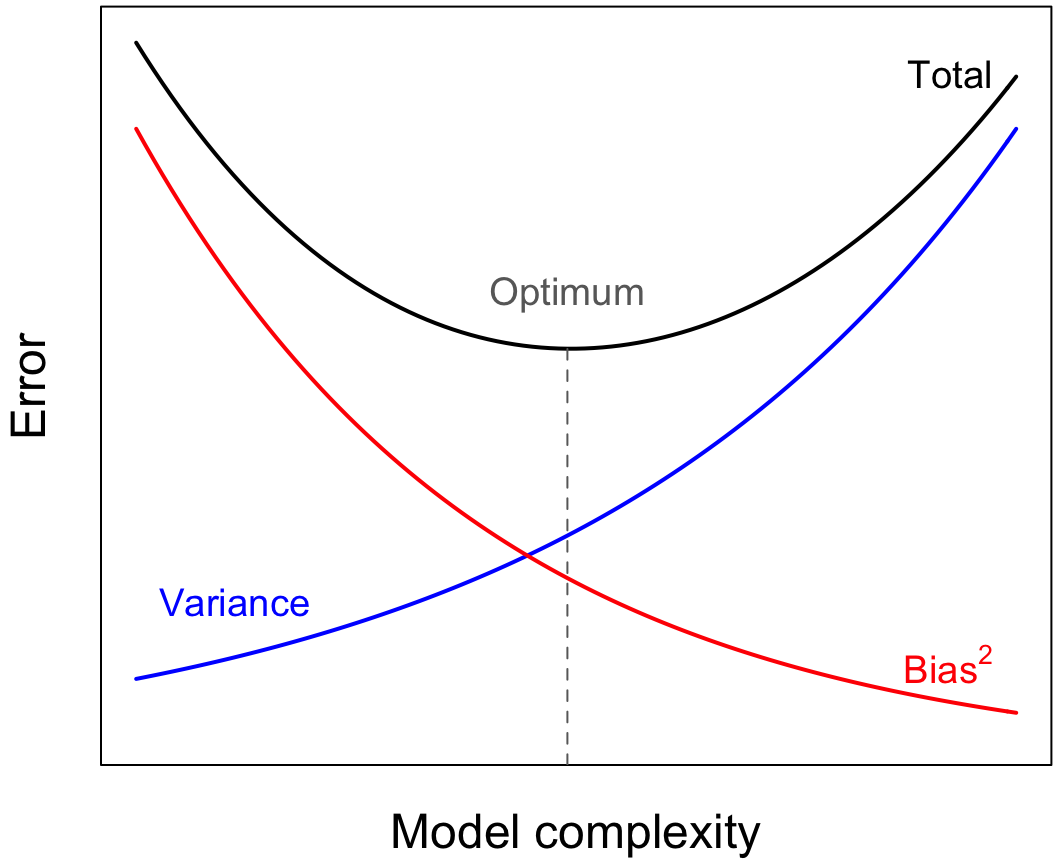
Unusual observations



Bias versus variance



Bias-variance trade-off



Model selection

In-sample

Null hypothesis testing

- F test, likelihood ratio test, χ^2 test

Regularization

- AIC, QAIC, BIC

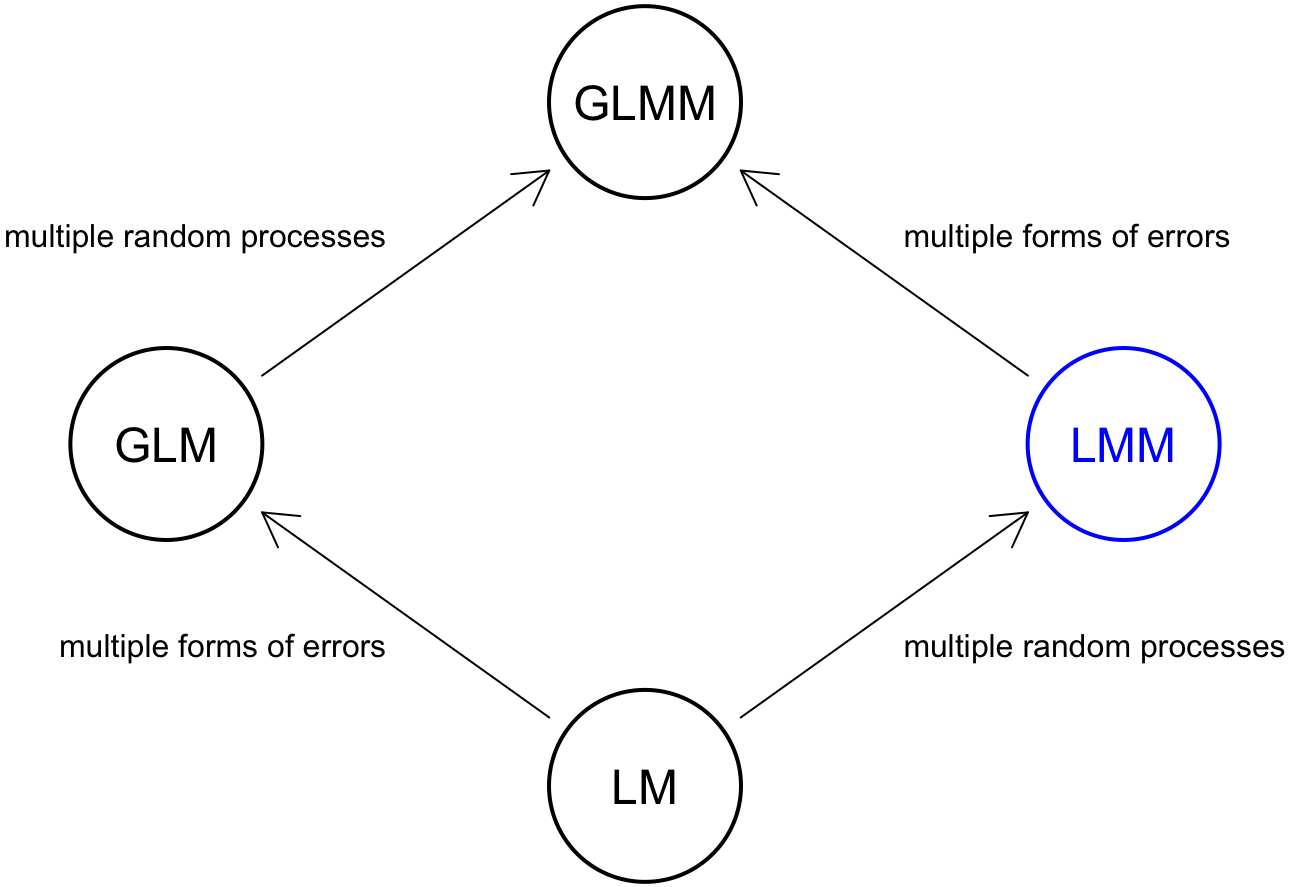
Model selection

Out-of-sample

Cross validation

- leave-k-out

Linear mixed models

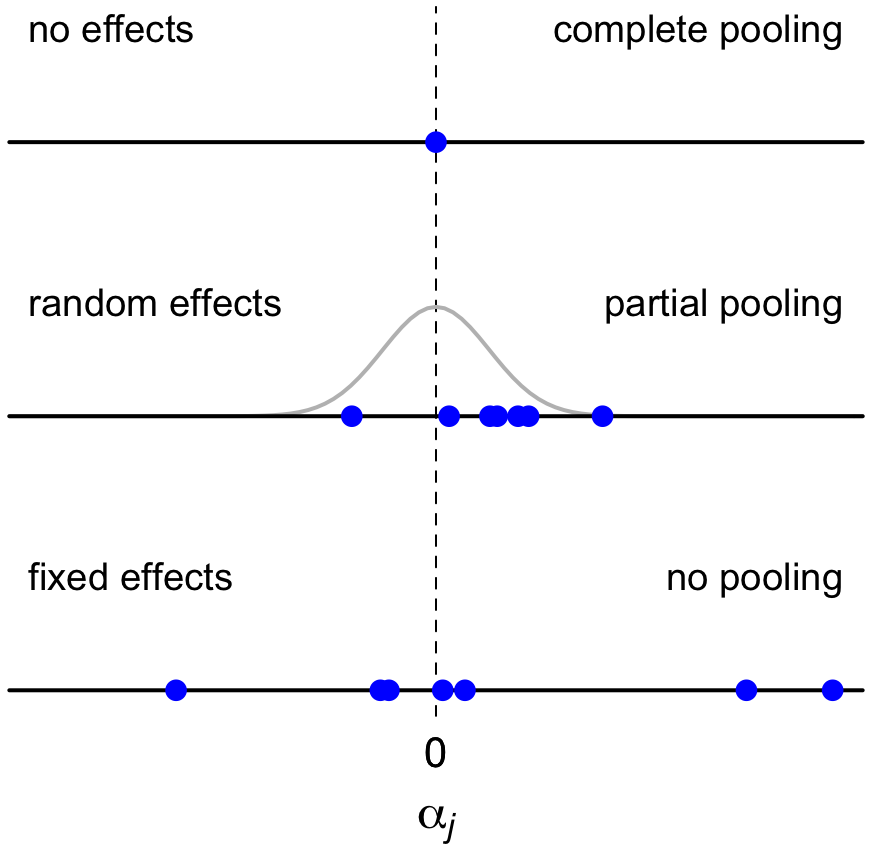


Fixed vs random effects

Fixed effects describe *specific levels* of factors that are *not* part of a larger group

Random effects describe *varying levels* of factors drawn from a larger group

Model for means



Linear mixed model

We can extend the general linear model to include both of fixed and random effects

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{e}$$

$$\mathbf{e} \sim \text{MVN}(\mathbf{0}, \sigma^2\mathbf{I})$$

$$\boldsymbol{\alpha} \sim \text{MVN}(\mathbf{0}, \sigma^2\mathbf{D})$$

Restricted maximum likelihood

Estimating the parameters in a mixed effects model requires *restricted maximum likelihood* (REML)

REML works by

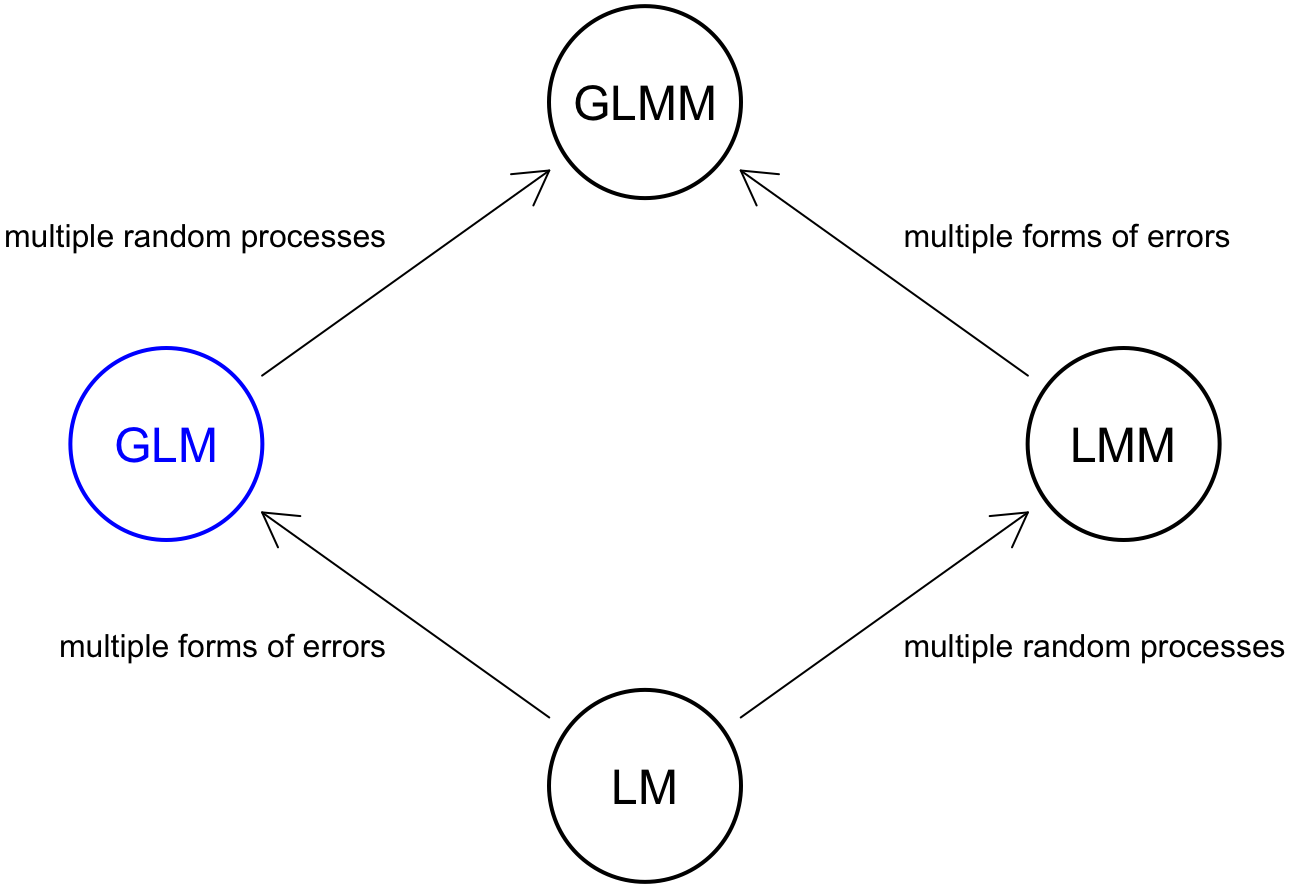
1. estimating the fixed effects ($\hat{\boldsymbol{\beta}}$) via ML
2. using the $\hat{\boldsymbol{\beta}}$ to estimate the $\hat{\boldsymbol{\alpha}}$

Model selection

To use AIC, we can follow these steps

1. Fit a model with *all* of the possible fixed-effects included
2. Keep the fixed effects constant and search for random effects
3. Keep random effects as is and fit different fixed effects

Generalized linear models



Generalized linear models (GLMs)

Three important components

1. Distribution of the data $y \sim f_\theta(y)$
2. Link function $g(\eta)$
3. Linear predictor $\eta = \mathbf{X}\boldsymbol{\beta}$

Common link functions

Distribution	Link function	Mean function
Identity	$1(\mu) = \mathbf{X}\boldsymbol{\beta}$	$\mu = \mathbf{X}\boldsymbol{\beta}$
Log	$\log(\mu) = \mathbf{X}\boldsymbol{\beta}$	$\mu = \exp(\mathbf{X}\boldsymbol{\beta})$
Logit	$\log\left(\frac{\mu}{1-\mu}\right) = \mathbf{X}\boldsymbol{\beta}$	$\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})}$

Logistic regression for binary response

We need 3 things to specify our GLM

1. Distribution of the data: $y \sim \text{Bernoulli}(p)$
2. Link function: $\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \eta$
3. Linear predictor: $\eta = \mathbf{X}\boldsymbol{\beta}$

Logistic regression for proportions

We need 3 things to specify our GLM

1. Distribution of the data: $y \sim \text{Binomial}(N, p)$
2. Link function: $\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \eta$
3. Linear predictor: $\eta = \mathbf{X}\boldsymbol{\beta}$

Overdispersion

The variance is larger than expected

Overdispersion generally arises in 2 ways related to IID errors

1. trials occur in groups & p is not constant among groups
2. trials are not independent

Overdispersion

We can estimate the dispersion c from the deviance D as

$$\hat{c} = \frac{D}{n-k}$$

or from Pearson's χ^2 as

$$\hat{c} = \frac{X^2}{n-k}$$

Effects on parameter estimates

The estimate of $\hat{\boldsymbol{\beta}}$ is *not* affected by overdispersion…

but the variance of $\hat{\boldsymbol{\beta}}$ *is* affected, such that

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \hat{c}\left(\mathbf{X}^\top \hat{\mathbf{W}} \mathbf{X}\right)^{-1}$$

Options for overdispersed proportions

Model	Pros	Cons
binomial	Easy	Underestimates variance
binomial with VIF	Easy; estimate of variance	Ad hoc
quasi-binomial	Easy; estimate of variance	No distribution for inference
beta-binomial	Strong foundation	Somewhat hard to implement

Counts vs proportions

With count data, we only know the *frequency of occurrence*

That is, how often something occurred without knowing how often it *did not occur*

Poisson regression

Counts (y_i) as a function of covariates

$$\text{data distribution: } y_i \sim \text{Poisson}(\lambda_i)$$

$$\text{link function: } \log(\lambda_i) = \mu_i$$

$$\text{linear predictor: } \mu_i = \mathbf{X}\boldsymbol{\beta}$$

General variance for count data

We can consider the possibility that the variance scales linearly with the mean

$$\text{Var}(y) = c\lambda$$

If c = 1 then $y \sim \text{Poisson}(\lambda)$

If c > 1 the data are *overdispersed*

Overdispersed regression

Counts (y_i) as a function of covariates

$$\text{data distribution: } y_i \sim \text{negBin}(r, \mu_i)$$

$$\text{link function: } \log(\mu_i) = \eta_i$$

$$\text{linear predictor: } \eta_i = \mathbf{X}\boldsymbol{\beta}$$

Zero-truncated data

Zero-truncated data cannot take a value of 0

Although somewhat rare in ecological studies, examples include

- time a whale is at the surface before diving
- herd size in elk
- number of fin rays on a fish

Poisson for zero-truncated data

The probability that $y_i = 0$ and $y_i \neq 0$

$$f(y_i = 0; \lambda_i) = \exp(-\lambda_i)$$
$$\Downarrow$$
$$f(y_i \neq 0; \lambda_i) = 1 - \exp(-\lambda_i)$$

Poisson for zero-truncated data

We can exclude the probability that $y_i = 0$ by dividing the pmf by the probability that $y_i \neq 0$

$$f(y_i; \lambda_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}$$

$$\Downarrow$$

$$f(y_i; \lambda_i | y_i > 0) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!} \cdot \frac{1}{1 - \exp(-\lambda_i)}$$

$$\Downarrow$$

$$\log \mathcal{L} = (y_i \log \lambda_i - \lambda_i) - (1 - \exp(-\lambda_i))$$

Neg binomial for zero-truncated data

We can exclude the probability that $y_i = 0$ by dividing the pmf by the probability that $y_i \neq 0$

$$f(y; r, \mu) = \frac{(y + r - 1)!}{(r - 1)!y!}\left(\frac{r}{\mu + r}\right)^r\left(\frac{\mu}{\mu + r}\right)^y$$

$$\Downarrow$$

$$f(y_i; \lambda_i | y_i > 0) = \frac{\frac{(y+r-1)!}{(r-1)!y!}\left(\frac{r}{\mu+r}\right)^r\left(\frac{\mu}{\mu+r}\right)^y}{1 - \left(\frac{r}{\mu+r}\right)^r}$$

$$\Downarrow$$

$$\log \mathcal{L} = \log \mathcal{L}(\text{NB}) - \log\left(1 - \left(\frac{r}{\mu + r}\right)^r\right)$$

Zeros in ecological data

Lots of count data are *zero-inflated*

The data contain more zeros than would be expected under a Poisson or negative binomial distribution

Sources of zeros

In general, there are 4 different types of errors that cause zeros

1. Structural (animal absent because the habitat is unsuitable)
2. Design (sampling is limited temporally or spatially)
3. Observer error (inexperience or difficult circumstances)
4. Process error (habitat is suitable but unused)

Approaches to zero-inflated data

There are 2 general approaches for dealing with zero-inflated data

1. Zero-altered (“hurdle”) models
2. Zero-inflated (“mixture”) models

Hurdle models

Hurdle models do not discriminate among the 4 types of zeros

The data are treated as 2 distinct groups:

1. Zeros
2. Non-zero counts

Hurdle models

Hurdle models consist of 2 parts

1. Use a binomial model to determine the probability of a zero
2. If non-zero (“over the hurdle”), use a truncated Poisson or negative binomial to model the positive counts

Zero-inflated (mixture) models

Zero-inflated (mixture) models treat the zeros as coming from 2 sources

1. observation errors (missed detections)
2. ecological (function of environment)

Mixture models

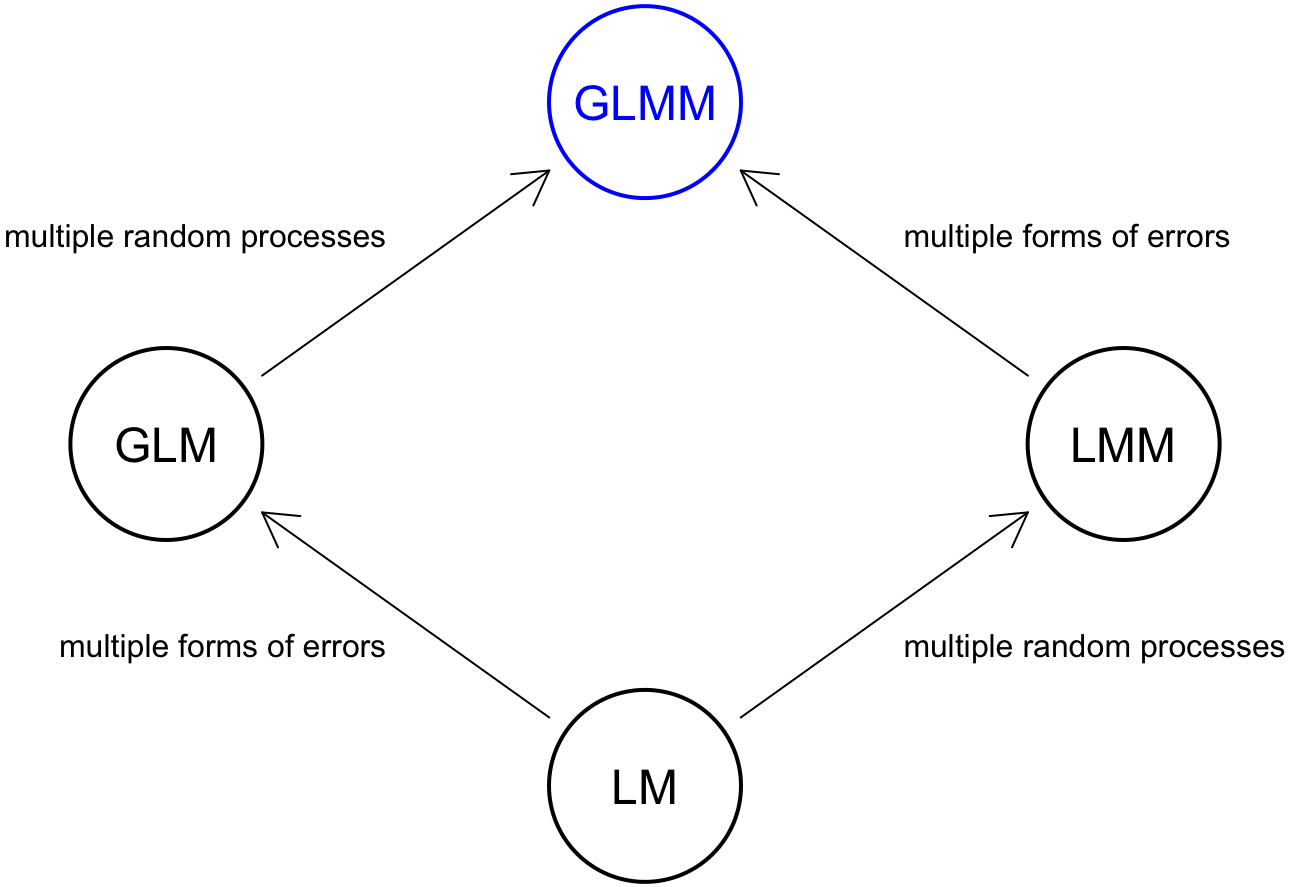
Zero-inflated (mixture) models consist of 2 parts

1. Use a binomial model to determine the probability of a zero
2. Use a Poisson or negative binomial to model counts, which can include zeros

Sources of zeros and approaches

Source	Reason	Over-dispersion	Zero inflation	Approach
Random	Sampling variability	No	No	Poisson
		Yes	No	Neg binomial
Structural	Outside count process	No	Yes	ZAP or ZIP
		Yes	Yes	ZANB or ZINB

Generalized linear mixed models



Generalized linear mixed model

GLMMs combine the flexibility of non-normal distributions (GLMs) with the ability to address correlations among observations and nested data structures (LMMs)

Generalized linear mixed model

Good news

- these extensions follow similar methods to GLMs and LMMs

Bad news

- these models are on the frontier of statistical research
- existing documentation is rather technical
- multiple approaches for fitting models; some with different results

Generalized linear mixed model

Just like GLMs, GLMMs have three components:

1. Distribution of the data $f(y; \theta)$
2. Link function $g(\eta)$
3. Linear predictor η

Linear predictor for a GLMM

For GLMMs, our linear predictor also includes random effects

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \alpha_0 + \alpha_1 z_1 + \cdots + \alpha_l z_l$$
$$\Downarrow$$
$$\eta = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha}$$

where the β_i are fixed effects of the covariates x_i

Generalized linear mixed model

$$
\begin{aligned}
\text{data distribution:} \quad & y_{i,j} \sim \text{Binomial}(N_{i,j}, s_{i,j}) \\
\text{link function:} \quad & \text{logit}(s_{i,j}) = \log\left(\frac{s_{i,j}}{1 - s_{i,j}}\right) = \mu_{i,j} \\
\text{linear model:} \quad & \mu_{i,j} = (\beta_0 + \alpha_j) + (\beta_1 + \delta_j) x_{i,j} \\
& \alpha_j \sim \text{N}(0, \sigma_\alpha^2) \\
& \delta_j \sim \text{N}(0, \sigma_\delta^2)
\end{aligned}
$$

Summary of GLMM methods

Method	Advantages	Disadvantages	R functions
Penalized quasi-likelihood	Flexible, widely implemented	inference may be inappropriate; potentially biased	`MASS::glmmPQL`

Laplace approximation	More accurate than PQL	Slower and less flexible than PQL	`lme4::glmer`
			`glmmsr::glmm`
			`glmmML::glmmML`

Gauss-Hermite quadrature	More accurate than Laplace	Slower than Laplace; limited random effects	`lme4::glmer`
			`glmmsr::glmm`
			`glmmML::glmmML`

THE FUTURE

Some things we didn't cover

Generalized additive models (QERM 514 in a different year)

Occupancy models (SEFS 590)

Capture-Mark-Recapture models (SEFS 590)

Multivariate response models (FISH 560)

Time series models (FISH 507)

Spatio-temporal models (FISH 556)

Bayesian methods (FISH 558, FISH 559)