# **Advanced statistical modeling**

1. **A protocol for regression -> see badger example**

Regressions can be used to explain relationships between variables (often in ecology) or to predict.

1. First look at your data:
   * Check the data and clean it: 70% of the work!
     + Drop:
       - variables with too many missing values and observations with NAs
       - correlated variables (choose one), coef pearson > 0.5-0.7
       - Categorical variables with only a few levels represented (only 1 or 0 eg)
     + Transform categorical variables in factors and dates to dates
   * Exploration: basic graphs to look at relationships between variables, interactions, sample size and nb of obs per level, outliers, missing values…
   * Descriptive stats: to get info about the distribution of the variables (package data explorer)
2. Define candidate models based on scientific question, hypothesis/predictions and what is relevant biologically speaking, and then fit candidate models to data
3. Compare and find the best model: model selection
   * Select the most relevant random effects with REML including the max nb of fixed variables
   * Select the most relevant fixed effect with ML including the selected random effects
   * Refit the best model with REML
4. Check assumptions a posteriori: if the assumptions are not met, change the model structure
5. Repeat until you find the best option
6. Check summary and interpret coefficients if possible
7. Prediction plots

*NB: the more complex the model is, the more data you need and the less easy it is to interpret it*

**Assumptions of the linear model by order of importance:**

Checking model assumptions should be done after building and choosing the best model. You can plot each plot or use the DHARMa package in R.

* Validity: does it make sense to use a model with the data we have? (scientific protocol)
* Additivity and linearity: linear function of predictors: y = 0 + β1x1 + β2x2 +… -> linear pattern between explanatory and observed variables.
* If not: transformation of variables (square, polynomial, log…), GLM (link function), GAMM (smoothing splines) or non-linear models
* Independence of residuals (= error of the model, difference between the model predictions and observations) and observations. Non independence is suspected in time series and spatial data -> wave or any kind of pattern in the residuals plot + ACF plot (first bar is always big and = 1, the rest should be within the 2 dashed horizontal lines)
* Try to include variables that are not in the model to see if the pattern disappears. If not, try to interpret it and then correct (eg with GAMMS which are flexibles, or model the error structure to account for non-independence: generalized least squares GLS, but not used so much in ecology)

Chart, scatter chart

Description automatically generated Chart, histogram

Description automatically generated

* Homoscedasticity = Homogeneity of the variance of residuals (also observations!): the residuals should be dispersed on a band of same width around 0 -> Suspected if cone in the residuals plot or other pattern. Consequence: the standard errors of the estimators are biased
* Allow different variances with generalized least square (GLS, not really used anymore) OR adjust distribution and model structure by adding a random factor (GLMM)

Chart, scatter chart

Description automatically generated

* Normality of residuals -> QQ plot: the dots should follow the line. Deviations are common towards the extremes: you can ignore it (at least for estimations, more important for predictions!)
* If not met: adjust the distribution and model structure with GLM

Chart, line chart, scatter chart

Description automatically generated

*NB: Same assumptions for the GLM except for the normality of the residuals*

*NB: linear doesn’t mean straight line between Y and X but predictors = additive combination of parameters*

**Interpretation des coefficients (additifs):** increasing the explanatory variable Xi of 1 unit ⬄ increasing the response variable of + βi units

1. **GLM**

Linear model : Y ~ Normal (mean μ, variance σ2) with μi = β0 + βixi +…

Generalized linear model: Y ~ Probability\_distribution (centrality parameter μi) with

g(μi) = β0 + βi\*Xi with g() = link function

μi = g-1(β0 + βi\*Xi) with g-1() = inverse link function

**Probability distributions with GLM**:

Determine the distribution a priori based on the nature of your observations. To interpret the results: transform back to the original scale using the inverse link function.

* Chart, line chart

  Description automatically generated**Gaussian distribution** (continuous, positive and negative values). Biological parameters centered around their mean like growth, weight, height… Symmetrical so very easy to work with -> usually used as a first approximation.

Y ~ normal (μ, σ2), link function = identity

Mean (X) = μ and variance(X) = σ2 (estimated by CI of the estimates and also how far are the observations from the predicted values of the model)

*Interpretation coefficients (additive)*:increasing the explanatory variable Xi of 1 unit ⬄ increasing the response variable of + βi units

* *Chart, bar chart, histogram

  Description automatically generated***Poisson distribution:** for counts (positive discrete values). The bump occurs around small values, but some rare values can be big (long tail).

Y ~ Poisson (λ) with link function = ln()

Moments of the distribution: E(X)= variance(X) = λ

*Interpretation of coefficients (multiplicative)*:if βi>1, when the explanatory variable increases the response variable increases as well, and it decreases if βi<1. Increasing the explanatory variable Xi of 1 unit ⬄ multiplying the response variable by exp(βi)

* Chart, histogram

  Description automatically generated**Binomial distribution** repetition of success/failure or proportions (observations comprised between 0 and 1): Parameters: nb of trials (n) and probability of success (p)

Y ~ Binomial (n,p) with link function = logit() = ln(p/(1-p))

Moments of the distribution: E(X)=np and var(X) = np(1-p)

*Interpretation*: exp(βi) = odd ratio = mesure par combien la cote (p/1-p) de la variable à expliquer (succès repro) est multipliée lorsque la variable explicative correspondante (berry index) augmente d’une unité (de 0 à 1).

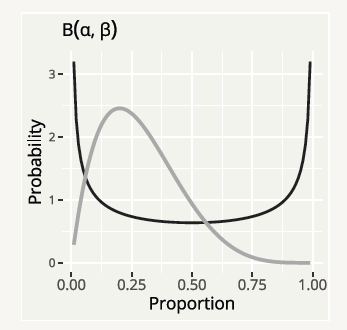
* Chart, line chart

  Description automatically generated**Gamma distribution** (continuous but strictly positive values): growth with ageing, or biomass. Parameters shape k, scale θ OR shape α = k and scale β = 1/θ -> difficult to interpret, better to avoid it…

Y ~ gamma (α, β), link function = inverse

(but you can use log link, easier to interpret)

Moments: Mean = α / β and variance = α / β 2

* **Beta distribution** (values comprised between 0 and 1, excluding 0 and 1), proportion of survivors. Parameters: shape α and beta β. Special package in R betareg

Y ~ Beta (α, β), link function = log function

Moments: E(X) = α/α+β and var(X) = αβ/(α+β)2(α+β+1)

**Classic problems**:

* Under and overdispersion (for Poisson distributions only, whose variances are linked to the mean): variance < and > than what it should be in theory -> negative binomial distribution
* Excess of zeros (for Poisson distributions) -> zero inflated models
* Non-linearity -> GAM
* Heterogeneity of variance -> GLMM (random factors)
* Temporal dependence -> Time Series / GAM
* Spatial dependence -> Spatial Model / GLMM
* Temporal and spatial dependence (repeated sampling over different study areas): add AR structure in GLMM using GLMMTMB
* Non-linearity and heteroscedasticity -> GAMM

1. **Overdispersion**

Over dispersion can in theory occur for any distribution for which the mean and variance are linked or constrained: Poisson, binomial, beta… (not normal!). But in practice, it usually happens for Poisson distributions in ecology (really rare for binomial and not many tools in R to do it anyway).

Poisson distribution (counts): V(X) = mean(X) = λ. But in reality, this relation is often not respected by your data: the dispersion differs from the ideal distribution.

* **Underdispersion:** var < E(X). Quite rare. Caused by overfitting (when the model fits outliers too well or if too many explanatory variables are included for the size of the dataset -> model too complex to be useful, no generalization possible).
* **Overdisperdion:** var > E(X). Common (so common that it is often not possible to use a poisson distribution in ecology).
  + **Apparent:** Mis-specified model
    - Missing important variables or interactions
    - Outliers
    - Non-linear effects
    - Predictors that need transformation
    - Inappropriate link function
  + **True:** Problem in the data: real excess in variation
    - Clustered, correlated response data -> diagnostic plot
    - Non – independence of observations
    - Success probability varies among trials (binomial data)
    - Many zeros
* Overdispersion underestimates the uncertainty of the estimates (small p-values, bigger AIC…) -> the effect is more precise that it should be or is detected (ie falsely significant) even if it doesn’t exist -> Selection of overly complex models which can lead to poor ecological inference.

**Detecting overdispersion (see zero inflates models)**:

* Before modelling: compare the histogram of response to the ideal distribution (long tails, lot of small values…)
* After modelling: DHARMa diagnostic plots (QQ plot and residual plot, see below for precisions)
* Summary: coefficient of overdispersion for GLMs: Deviance/df > 1 -> overdispersion
* Hypothesis test (not really used)

**How to fix it:**

* **Add a random effect (animal D)**
* **Quasi poisson GLM (not really used)**: E(X) = λ and V(X) = ϕ x v(X)

ϕ: dispersion parameter, overdispersion when ϕ > 1

Limit: Not a real distribution: quasi-likelihood instead of likelihood -> you can’t get an AIC (but a QAIC) with this method, so no possible to compare poisson models and quasi poison models

Packages: AICcmodavg::AIC or MuMIn::QAIC

* **Negative binomial distribution**: more flexible because real distribution (likelihood). In ecology, use the negative binomial 2 or ecological parametrization. Poisson-Gamma mixture (allows you to model a bump for small values and a long tail). Reflects unmeasured variability in the population.

Y ∼ NB(μ, k)

* + μ: mean number of failures or counts (mu in R)
  + k: overdispersion parameter (size in R). the smaller k = the more overdispersed AND

k>10: variance ~ mean and NB ~ Poisson

NB: package glmmTMB

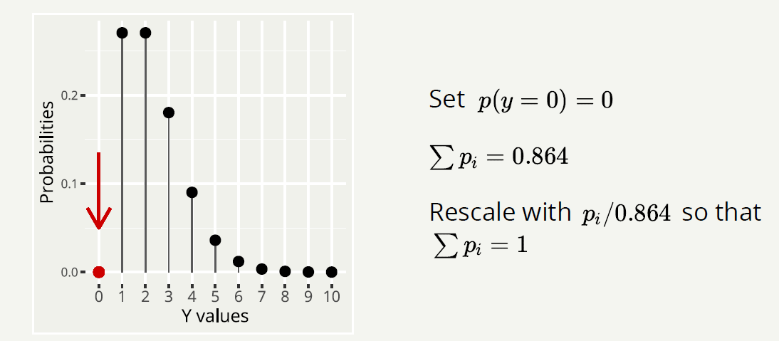
m5 <- glmmTMB(Count ~ Field + species + offset(log(Nrelease)), data = b1, family = nbinom2(),

ziformula = ~ .)

Advantage: You can compare Poisson and negative binomial distribution models with AIC and estimates overdispersion with the k parameter. *NB: These models can still be overdisperesed! But usually, they can deal well with big overdispersions.*

1. **Zero truncation or inflation**

One of the reasons for overdispersion: lot of zeros in your data. Most of the time, negative binomial distribution is sufficient to fix it but not always… Some people say that we don’t need these models because in practice the negative binomial is very flexible.

* **Zero-truncated regression**: When the response variable can’t take the value 0 (age, weight…). The zero truncated regression takes into account that 0 is in theory impossible to obtain and then scale the rest of the probabilities so that sum(pi)=1 (by dividing them by 1-p(Y=0)). Not that common in ecology. If lambda (average count) is large (~100) the negative binomial is enough. For small lambda (often in ecology) -> biased estimates.

Different libraries in R: countreg (function zerotrunc), VGAM (function vglm)

* **Zero-inflated regression**: when there are more zeros counted than expected. Consequences: overdispersion, biased estimates. Cause:
  + **Real zeros (ecological process) = structural errors**: animal nor present because habitat is not suitable
  + **False zeros (sampling process)**:
    - Design errors (counting at the wrong time/place, or sampling for a to short period or too small area…)
    - Observer error: species difficult to detect
    - “Bird error”: habitat suitable but not used for unknown reasons

**How to deal with it**:

* + **Zero-altered models or hurdle models** (zeros vs non-zeros). If you don’t want to figure out if the zeros are true or false. Not really used in ecology. Combines 2 models:
    - *One model for all zeros*: binomial (presence or absence of zeros?). Includes intercept only or covariate (if you suspect that one variable could explain the presence of zeros)
    - *One model for the remaining counts (non-zeros)*: zero-altered Poisson (ZAP) or zero-altered negative binomial (ZANB, if the count part of the histogram is still overdispersed = long tail) + different covariates (that explains the count)

Packages: pscl, function hurdle (first part = count, second = zeros) + countreg, VGAM…

* + **Zero inflated models or mixture models** (true vs false zeros): most commonly used in ecology! If it is expected that zero counts can be biologically explained, and you can somewhat or want to differentiate between true zeros and false zeros. Splits the data in 2 groups: false zeros and counts which may produce true zeros (which zeros belong to group 1 or 2 is unknown):
    - *Model for false zeros*: binomial +/- covariates (that explain the false zeros).
    - *Model for the counts including true zeros*: poisson or negative binomial (if this pqrt is still overdisperesed on the histogram) +/- covariates (that explain the counts)

*NB: In practice, ecologists usually use only the intercept model (~ 1) for the false zeros or the same covariates for false zeros and counts.*

Packages: library glmmTMB

m3 <- glmmTMB(Count ~ Field + species + Field:species + offset(log(Nrelease)), data = b1, family = poisson(), ziformula = ~ .)

*NB: ziformula: model that tries to determine which variables can explain the presence of false zeros. “.” Means that we take the same variables a the model above. Don't spend too much time trying to fit the best model for ziformula because you want to account for them but not explain them. Test just for obvious variables*

**How to detect overdispersion and zero-inflation:**

* Before modelling: compare the histogram of response to the ideal density of distribution (long tails, lot of small values…)
* **Diagnostic plots with DHARMa** (deals with GLMs and mixed models)**:** Ideally, you want:
  + Left plot: dots follow a line (ie QQ plot)
  + Right plot: black parallel lines and the dots widespread with no pattern or blank space (otherwise the model doesn’t fit well for some obs)

In poisson models (discrete values), you usually have dots clustered in lines -> difficult to visually evaluate on the graph -> packages like DHARMa tells you if something is wrong by simulating a high nb of residuals (= dots) and then scale them to identify errors.

* + **If underdispersion** (less variance than expected = short tail or smaller range): residuals are clustered together, and the tail part (y=1) is empty (no big values)

Chart

Description automatically generated

* + **If overdispersion** (more variance than expected = long tail): residuals overspread but concentrated on the value zero (Y=0) and big values (tail = red stars) -> adjustment needed. The red stars correspond to simulated outliers (no present in your data, but if it detects a lot of them, you probably have some in your data as well)

Chart, scatter chart

Description automatically generated

* + **Zero inflation (particular case):** concentration of residuals around zeros(Y=0)

Chart

Description automatically generated

**Formal tests for dispersion:** testDispersion OR testZeroInflation functions: compare the variance of the model with the variance of the ideal poisson distribution. There is a significant difference in the variance if p < 0.05 and the red bar (fitted model) is not aligned with the bump of the ideal distribution (the further away, the bigger the difference):

* Underdispersion: red bar on the left of the histogram
* Overdispertion (or zero inflation): red bar on the right of the histogram

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

*NB:**some people are comparing all these models using AIC. Zero inflated poisson / binomial and poisson negative binomial is ok to compare, for the rest? Nevertheless, a lot of packages are doing it anyway*.

**Summary**

* Histogram: does it look like the ideal distribution?
  + Yes -> Poisson distribution and then diagnostic with DHARMa
  + No (lot of zeros, long tail…) -> negative binomial distribution directly, then diagnostic
    - if not enough (still overdispersal, too many zeros)-> zero inflated models

1. **GAMs**

* See Generalized Additive Models - Michael Clark

**Dealing with non-linearity:** Sometimes the combination of variable (predictor) is non additive -> no possible to use linear models. In this case different options:

* Include more explanatory variables
* Include interactions
* Include quadratic effect, like polynomial regression (a + b1x + b2x2) -> still linear model
* Transform the response variable (ln…) : usually not efficient
* GLMs : transform the response variable with link function
* Use a smoother with the exp. variables -> polynomial regression, segmented regression, splines, GAMs

*Former smoothing techniques to deal with non-linearity:*

* Polynomial regression (μ = a + b1x + b2x2). Problem: the curve shape is constraint (rigid structure), limited for complex relationships + difficult to specify to obtain something satisfying. Usually non representative of the true biological relationships

 Chart, scatter chart

Description automatically generated

* Segmented regression: break X in segments and fit a linear model in each segment. Rarely seen in ecology

Text, letter

Description automatically generated Chart, line chart

Description automatically generated

* Splines / smoothing: split X in regions and fit a polynomial regression for each segment (more flexible). Predictions is made for each of these bins that are then connected to obtain a smoothing curve. Possible to add a penalty. Different methods:

LOESS: the simplest one. A window is chosen around each value = knot and the mean response for this target value is predicted using a linear or polynomial model. Operation repeated for each window to obtain a curve. *NB: method used for geom\_smooth in ggplot2*

Problem: what happens at the edge of each window: can be difficult to connect the segments if predictions are really different at the edge…. Be careful with interpretation

Text

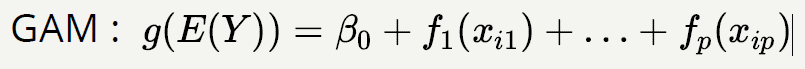
Description automatically generated Chart, scatter chart

Description automatically generated

**GAMs**

Packages in R: mgcv, function gam.

Extension of GLMs (different variables’ distribution can be modelled: gaussian, poisson, binomial...), that allows non-linear relationship between predictors and response variable via the use of smooth functions while maintaining additivity. In mgcv, predictions are penalized to reduce wiggliness ie to get something representative (prevents overfitting) but as simple as possible (penalized likelihood).



*NB: smoothers can only be used for qualitative variables, but you can include qualitative variables in the model*

**Non-linear smooth function** (argument s(Var)) have a common basis (bs parameter in s()), but in mgcv each function is automatically adjusted for each predictor based on the lowest GCV. Basis can be of various nature (In practice, small differences except for very specific data):

* *Thin plate smoother* (tp): generally quite good (but not for large dataset), first option and default
* *Shrinkage smoother* (ts): penalty that brings the coefficient toward 0 (towards absence of effect) if the predictor is not necessary. Allows you to select only the necessary variables (like AIC). Can be used if you are not sure if you should drop a variable or not.
* *Cubic regression spline* (cr): for cyclic predictors, when you want to connect the far left to the far right of your curve, eg if X represents the months
* …

*NB: if the smoother is in more than 2 dimensions (like latitude, longitude…), don’t use s() but ti() or te()*

**The nb of knots (k parameter)** can be set to get your curve more or less wiggly (= defines the max degree of freedom). Need to be big enough to allow your curve to be flexible but not too big because of computational efficiency. General recommendations: <50 obs: 3 knots, > 100 observations: 5 knots. Automatically selected by the gam function and defaults are usually reasonable, but can be changed if not appropriate (for instance a huge amount of data…). Need to be checked afterwards:

* Check if the effective degrees of freedom EDF in the summary of the final model is close to the chosen k. If so, k might be too low -> increase k and if EDF changes a lot, k was too low
* With the gam.check function:
  + K-index: The further below 1 the k-index is, the more likely it is that k is too low.
  + P-values: Low p-values may indicate that the basis dimension k has been set too low

**Advantage**: GAMs allow you to mix different kind of variables:

* **Response**: different type of distributions (like GLM’s -> link function). Argument “family”
* **Explanatory**: Quantitative (smoothed or not), qualitative (not), interaction between a smoother and a factor (ie different smoother for each level of the factor): y ~ s(X, by=int Factor) + int Factor
* **Random effect:** s(random factor, bs =”re”) for random intercept and s(CollarID, bs = “re”) + s(CollarID, OLA, bs = “re”) for random intercept and slope
* **Interaction (not often used):** s() <-> + te() <-> : ti() <-> \*

*g*(*μ*) = *α* + *βVi* + *f*1(*Xi, by=Wi)* + *factor*(*Wi*) + *f*2(*Zi*, *bs=”re”)*

*1. Linear fixed effect, 2. Nonlinear effect (smooth function) with interaction, 3. Factor that interacts with the previous exp variable, 4. Random effect*

**Interpretation of GAMs:**

* GCV: to compare models like AIC (AIC is more commonly used)
* Deviance explained ~ R square: percentage of variation explained by the model
* Edf (effective degree of freedom): overall contribution of the smoother to explaining the response (the bigger, the more important contribution to the response variable). If equals to 1 -> linear coefficient is sufficient, no need of smoother
* Don’t look at p-values!
* No possible interpretation of coefficients (because it changes depending of the Xi value)
* Look at the prediction graph and interpret visually

**Visualization of each predictor**:

* Package mgcv function plot or package gratia function draw -> only deals with smoothers, linear components are not plotted. If some parts are too wiggly -> overvitting: check the k parameter with gam check

Chart, line chart

Description automatically generated

*NB: vertical dashes ⬄ were are the observations in the dataset. If no observations, interpolation (straight line) and bigger variance*

* **Diagnostic plot to check assumptions****(residuals plots gam.check)**:

Chart

Description automatically generated Chart, histogram

Description automatically generatedChart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

* 1: QQ-plot (should follow a line)
* 2: compare histogram to theorical density of distribution. In case of gaussian distribution, the residuals are still expected to be normally distributed.
* 3: Homogeneity of variance. More variance in the biggest values: heteroscedasticity -> dealt with mixed models
* 4: how good is the prediction? More variability for big values but the model predicts it well. If too perfect, might be an indication of overfitting

**Other problems to check**

* Independence (autocorrelation, especially for time and space series) - ACF
* Concurvity (=collinearity for GAMs)
* Nested data (hierarchical), voir clarks

**Prediction based on model:** create a data set based on the original one where one variable varies while other ones are held constant. Fix the other variables to their mean (or to one level for categorical variable), and for the variable that varies, keep the range of the initial observed data or choose the values for which you would like to predict (outside the range for ex). Function predict.

Graphical user interface, text, application

Description automatically generated

*NB: here we fix all variables, if you want to see more than 2D relationships, you can plot 3D graphs in which 2 variables vary at the same time. But can be difficult to read and doesn’t take into account missing values. Another possibility, use te instead of s and then draw.*

*Package mgcv function vis.gam and also in gratia*

**Present your results**: Include the parameters of the summary and then prediction plots

**Advantages of Gams**:

* Allow non-linear effects
* Very flexible: Different smoother possible for different variables, hybrid models with non-smoothed variables, mixed effects…
* Additive: interpretable as the same way: effect of Xi while holding other predictors constants
* Extensions with random effects, autocorrelation structures…

**Disadvantages:**

* Tend to overfit due to wiggliness
* Sometimes add more complexity than necessary
* May be a bit tricky to interpret

1. **GLS**

Assumption of LM: residuals should be normally distributed, independent and homogeneity of the variance (F and t-tests). To deal with residuals that are not respecting these assumptions (eg heterogeneity of variance, nested data, temporal correlation, spatial correlation, random noise…), we can (1) directly specify the structure (variance and covariance) of the residuals (LS/GLS) or (2) add random components in the model to create groups and estimate directly the structure of residuals(GLMM).

*NB: GLS are not used often in ecology: most people use GLMMs*

**The variance covariance matrix** M specifies the structure of the variance (diagonal) and covariance (off-diagonal).

If the assumptions of LM are respected:

M = variance x Identity matrix (1 on the diagonal and 0 elsewhere)

This matrix can be modified to model changes in the structure:

* **Weighted least square** (WLS): to model heterogeneity of variance of residuals (variances change along the diagonal)
* **Generalized least square** (GLS): to model heterogeneity of variance AND correlation of residuals (covariances change off-diagonal)

1. ***Dealing with heterogeneity of variance***

Chart, scatter chart

Description automatically generatedIn ecology, it is interesting to understand what causes the heterogeneity of variance.

**How to detect and deal with heteroscedasticity**: Look for strong patterns in residual plot (ie residuals not distributed in a band of width 2x sigma). If so, look at plot of residuals against each variable to determine which variable is responsible for this (numeric or categorical). Then change the diagonal of the matrix by weighing each observation using GLS. Different possible variance structure (see ppt)

1. ***Dealing with dependence***

Non-independence is suspected in case of temporal /repeating measurements, time series) and spatial data (multiple spatial locations, clusters). Things close to each other (temporally or spatially) are more likely to be similar -> Think of it beforehand when designing the protocol!

1. Chart, histogram

   Description automatically generated**Temporal dependence**

**How to detect temporal dependence:**

* ACF function
* DHARMa: function testTemporalAutocorrelation. If you have several observations per time step, e.g. because you have several locations, you will have to aggregate with recalculateResiduals.

res = recalculateResiduals(res, group = b1$Date)

testTemporalAutocorrelation(res, time = unique(b1$Date))

**How to deal with temporal dependence**: use a correlation structure which make sense biologically speaking (lagged effect of one year for instance). Frequent correction functions used:

* Autocorrelation: simplest example of dependence -> the value of X is closer to X+1 than X+2…
* Compound symmetry structure
* Auto regressive model of order 1 (AR1): auto residuals at time s are modelled as a function of the residuals at time s-1 with noise (when 2nd bar of ACF is over the threshold)
* AutoRegressiveMovingAvergae (ARMA): residuals at s are a function of residuals at p previous pints with noise (generalization of the previous). If you have more than one bar above 95% interval in the ACF plot. Not often see in ecology, more in economy.

1. **Spatial dependence**

**How to detect spatial dependence:**

* Bubble plot: to observe the residuals of the model along latitude and longitude. Should be well dispersed and without any patche of one color. If not respected: spatial autocorrelation.
* DHARMa package: testSpatialAutocorrelation function. If you have several observations per location, create first a group with unique values for each location and then use the recalculateResiduals function to aggregate residuals per location,
* Variogram: observations taken by pairs and all pairs should cancel out each other. The gamma value indicates spatial dependance when it is low. It also indicates which function should be used, but the plot is complicated to read … -> see Evelyn workshop

Chart, scatter chart, bubble chart

Description automatically generated 

Graphical user interface

Description automatically generated with medium confidence**How to deal with spatial dependence:** Correlation structure based on shape of variogram

Generalization GLS to other distribution than gaussian -> GEE: Non-normal data, non-diagonal correlation matrix, possible to include random effects. BUT Cannot estimate random effect, Quasi-likelihood no AIC nd does not handle hierarchical clustering

* **Ccl: Gee considers the variance heterogeneity as nuisance and accounts for it but don’t estimate it whereas it is often something we are interested in in ecology. GLMM gives an approximation of the variance of the random effect (ie the dispersion around the mean).**

**Use GLMM instead of GLS/GEE!**

1. **Mixed effect models** (see glmm faq on internet)

Mixed models not only include fixed terms (or pop level effects, Xi) and error term (residuals = remaining unexplained noise, ei) but also random terms (or group effect, Zi).

A picture containing text, clock

Description automatically generated

*NB: in matrices (written in bold), each column = one variable and each row is an observations.*

* **Fixed terms**: deterministic components you chose to include in your model in relation to your scientific question -> effect across groups at the pop level
* **Random effects**: categorical variable that is not directly related to the scientific question but imposed by the sampling method -> effect vary within groups

Random variables are modelled with a specific probability distribution, and you can get the parameters describing this distribution (usually the variance) from the model summary. In GLMMs the random distribution used is Gaussian centered on 0 (mean) with a variance D. Usually sufficient in ecology, but if you need to use another distribution, go Bayesian.

*Text

Description automatically generated with medium confidence*

**What categorical variables should go in fixed/random effects:**

* Fixed effects:
  + Categorical variables for which all possible levels are represented in the data
  + Categorical variable of direct interest (related to the scientific question) for which we want to estimate the effect
  + Usually, you need 10 observations for each parameter (parameter = what is estimated in the summary, coefficients and variance for GLMM. Intercept doesn’t count) to prevent overfitting. In simple models, usually equivalent to 10 obs for each fixed variable
* Random effects:
  + Levels randomly sampled from a pop (not all levels are represented in the data)
  + Categorical variable that we are not interested in directly but that explains some deviation from what is described by the fixed effects (overD, clustering…)
  + Usually more than 6 obs per random levels are required to be able to model a random effect (otherwise doesn’t make sense to estimate a variance with so few points)
* Not always easy to define, no strict definitions apart from these guidelines. In Bayesian: all variables are considered random!

*NB: Including random effects in the model induces a correlation structure in residuals. But with GLMMs, no need to build variance-covariance matrices with a specified structure like in GLS, the model estimates it directly from the data*

**Different types of mixed models** (package lme4 function lmer)

* *Random intercept*: each level starts at a different intercept but has the same slop (ie same trajectory = parallel curves). Often in ecology, Ex: when modelling different plots, sampling sites … In R: (1|random factor)

Summary:

* Fixed effects: estimate of pop level intercept and slope
* Random effects: estimation of the variance of the intercept for each level of the random factor (ie deviation from the pop level)

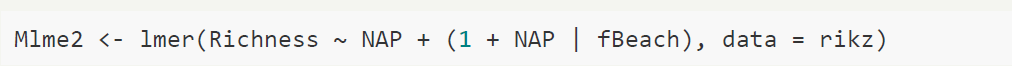
Text

Description automatically generated with low confidence

* *Random intercept and slope*: each level has a different intercept and slope (ie relationship and trajectory). Requires a lot of data because more complex! Rarely enough data in ecology…

In R: (1 + fix var that varies depending on random level | random var)

Summary: …



* Random slope only or only random effects (ie no fix effects) make little sense

**Why using GLMM is better**: To account for the variation among experimental units (eg tanks) you can either add a fix effect or a random effect:

* NO EFFECT: complete pooling -> the model is underfitting because doesn’t take into account the variation across tanks
* ADD FIX EFFECT: the prediction for one tank is estimated without taking into account other tanks -> overfitting, because outliers are pulling the estimated pop level towards them and therefore have a big impact on the predictions
* ADD A RANDOM EFFECT: Shrinkage and partial pulling -> the model pulls predictions towards the population intercept using information about other tanks. Each tank is weighted according to its value and the nb of observations available, and therefore the biggest change is observed for extreme values.

*NB: it can lead to convergence problems if you don’t have enough observations in some levels of the random factors. If only one group is poor in data, not that important because the model infer the predictions from other groups and pop levels. If not enough data, you get error messages from R.*

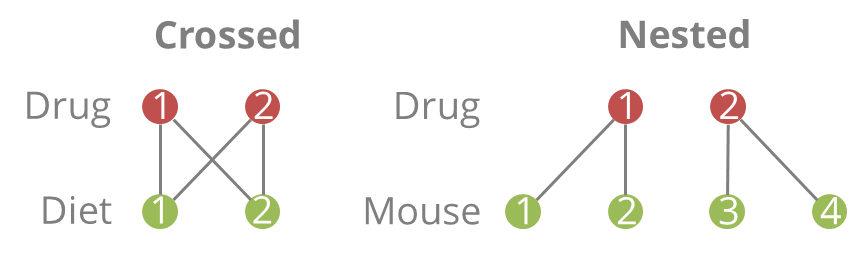
**REML (restricted maximum likelihood)** is better at estimated the variance of random factors than maximum likelihood (because biased estimates of variance). Set the method = ”REML” for GLMMs and GAMMs for model selection of random effects and in your final model. Not possible to compare models fitted with REML and ML with AIC!

**Model selection guideline (especially if you have a lot of variables to test):**

* Start with a model that includes as many fixed variables as possible and the possible random components (including interactions/nested)
* Do the model selection on random components only to determine which one you should keep (using REML)
* Then do the model selection on the fixed components (using ML) with models including the random components selected in the first step
* Refit the top model with REML (to better estimate the variance of random effects)
* Check diagnostic plot for assumptions validation

**Crossed vs nested data**

* **Crossed:** when each cluster contains the same possible values of another cluster. Ex: When the treatment A and B are tested with the diet 1 and 2
* **Nested**: when clusters are nested within other clusters. Ex: cities within counties, and counties within states.



Whether random effects are nested or crossed is a property of the data, not the model (ie there is nothing special done for models with random effects for nested factors and exactly the same computational methods are used). However, when fitting the model, effects can be included as either nested or crossed.

In R:

* Crossed: + (1 | hospital) + (1 | ward)
* Nested: + (1 | hospital) + (1 | hospital:ward),

1. **Bayesian analysis**

**INTRODUCTION**

The Bayesian approach to data analysis requires a different way of thinking about things, but its implementation can be seen as an extension of traditional approaches. The key difference regards the notion of probability, which is different than frequentist statistics but actually more intuitive. Furthermore, p-values and CI intervals will have the interpretation that many applied researchers incorrectly think their current methods provide. On top of this, one gets a very flexible toolbox that can handle many complex analyses. In short, the reason to engage in Bayesian analysis is that it has a lot to offer and can potentially handle whatever you throw at it. As we will see shortly, one must also get used to thinking about distributions rather than fixed points. With Bayesian analysis, we are not so much as making guesses about specific values as in the traditional setting, but more so trying to understand the limits of our knowledge and getting a healthy sense of the uncertainty of those guesses.

1. **What is a probabiblity ?**

* **Frequentist probability**: In traditional statistics, the probability = observed frequency of counted events on a nb of (hypothetical) trials (usually big) -> p-value: p(X > xobs | H0 is true) proba to observe at least as many observations as what we see in our dataset assuming the null hypothesis. We reject H0 if p <0.05 -> observed data unlikely to occur if H0 is true. In theory, this method is objective and conditioned on a hypothesis. But it assumes that:
  + the null hypothesis actually exists or is possible
  + Infinite rep of experiment is possible
  + Compute probability of unobserved events

*Problems of p-value*: no reproductible because depends on the sample size of your dataset + if H0 is not accepted, we usually assume that H1 fits better but in reality, we can’t conclude anything. And statistical significances doesn’t mean biological significance.

* **Bayesian probability:** intuitive probability of a hypothesis based on prior knowledge/belief and evidence at hand (data) + quantification of the certainty. Subjective but conditioned on data:

**Bayesian theorem**

Graphical user interface, application

Description automatically generated with medium confidence

1. **Estimating parameters**

* **Frequentist:** Based on Maximum likelihood estimation = try find the coefficients (alpha (intercept), beta (coefficient) and sigma (sd)) that maximize the likelihood for the observed dataset and the chosen distribution (normal, Poisson, binomial…) -> difficult to obtain based on one dataset.

*Output***:** point estimates and associated p-values / confidence interval

* **Bayesian:** What we end up with in the Bayesian context is not a specific value for all parameters that would make the data most likely, but a probability of distribution for the parameters that serves as a weighted combination of the likelihood and prior. Ie the prior belief on a coefficient b (expressed as a probability of distribution, eg normal centered on some value μ and with some variance σ2) is updated in light of the current dataset. All uncertainty is quantified and propagated (iterative).

*Output***:** posterior distribution of estimates (mean, median, 95% credible intervals)…

Text

Description automatically generated with medium confidence Text, letter

Description automatically generated with medium confidence

*NB: theta = model parameters: alpha, beta and sigma*

* + **Prior distribution:** distribution of the parameter that you assume given your experimental design and previous knowledge (initial belief)
  + **Likelihood:** current evidence in your data. Probability of obtaining the data with the current prior distribution -> partly know because you know that some kind of data are best explained by some distributions (poisson (counts) or binomial (frequencies))
  + **Posterior distribution:** reevalution/update of the distribution after combing your initial knowledge (prior) and your data -> outcome
  + **Normalizing constant = ugly beast**: can’t be calculated. But proportional to a ratio, or possible to calculate a bayes factor. If not possible to calculate, can be simulated by drawing random values from posterior distribution by iterative and stochastic simulation -> Monte Carlo Markov Chain MCMC. Various algorithms (metropolis\_hastings, BUGS/JAGS, Stan).

**Principle of MCMC sampling to approximate posterior distribution:** iterative process starting with a random value x and that determines the value y of the posterior distribution for this value x. Then jumps to another value based on the previous one and on a decision rule to determine where to explore next. Process repeated many times to produce simulated draws from the posterior distribution until a stationary distribution (gets less jumpy).

* Start with a random value and the nb of total iterations you want (iter)
* Warmup draws (warmup): give time to the process to settle down from the initial random starting point (which might be way off, and thus the subsequent estimates will also be way off for the first few iterations) -> these values are not taken into account to trace the posterior distribution
* Post-warmup draws: Iterative process until the stationary distribution. You can specify how many draws you want to keep to trace the posterior distribution (thin= 4, every fourth draws)
* Run the whole thing multiple times (chains), to check if you get the same result. If multiple chains get to the same place in the end, we can feel more confident about our results.

1. **How to do it in practice**

* Set prior distributions for each parameter you want to estimate based on your initial believes or prior data (possible priors: uniform, betha, triangular…) -> you get a probability for a given value of θ according to this distribution. Do not use your current dataset to make these assumptions!
* Compute the likelihood of the data given some values of θ (p(y|θ))
* Compute the posterior distribution via Bayes theorem and MCMC based on prior distribution and likelihood (reallocate credibility across possibilities)
* Evaluate the assumptions (does the new distribution fits well the data?) with traceplot for each parameter estimated. What we are looking for after the warmup period is a “fat hairy caterpillar”

(the estimates from each chain find their way from the starting point to a more or less steady state quite rapidly and all chains are mixing well after the warming up period)

* Specify which parameters you want and summarize the posterior distribution (mean of posterior draws, 95% credibility intervals, and other distribution parameters)
* Make predictions based on new datasets or replicate the observed data based on the parameters θ and compare the result to the observed data to see how well they match

In R: brms (more flexible but models not precompiled, more difficult to code) and rstanarm (model precompiled so easy to write regression models only, just add stan\_toglm) packages

1. **Advantage of Bayesian**:

* The concept is simple and more straight forward
* Can calculate probabilities of events that have not happened yet
* no H0, alpha levels, p-values, power analysis, small sample correction, post-hoc comparison.
* Unbiased for small samples and handles missing data (consider it as a prediction)
* Use prior information, explicit assumptions (guides the model in the right direction).
* Embrace uncertainty and supports complex decision making under uncertainty
* Active model building, very flexible to build hierarchical models, but computer efficiency ++
* You will always get a solution however complex is your model (you will not get stuck but can take weeks). Give you the posterior distribution rather than point estimate
* Easy to combine data from different sources (conceptually but difficult to write in R)

1. **Inconvenients**

* Learning code is steep, coding is more difficult in Sta (but more and more easy in R)
* Can be computationally intensive and slow (several weeks)
* Difficult model selection (but better now, almost as easy as with the AIC)
* Bad assumptions lead to wrong conclusions

1. **Where is it used? (see exemples canvas)**

* Spam filters
* Insurance companies (are you going to cost them money)
* More and more in ecology

**Summary**

The Bayesian approach provides an updated belief of the state of the world as a weighted combination of prior beliefs regarding that state (expressed as a mathematical model such as a (generalized) linear model) and the currently available evidence. In addition, there is the possibility of the current evidence overwhelming prior beliefs, or prior beliefs remaining largely intact in the face of scant evidence.

*NB:  if we add more observations to the data, we give more weight to the likelihood, which is what we want as scientists, ie we’d want the evidence/data to eventually outweigh our prior beliefs.*

|  |  |  |
| --- | --- | --- |
|  | **Frequentists (objective)** | **Bayesians (subjective)** |
| **Study and dataset** | Assumes infinite repetitions of unobserved events (study is repeatable) | Fixed (you work with your dataset) |
| **Parameters** | Fixed (one point estimate) | Random (posterior distribution) |
| **Model credibiliy** | How likely is the data given the values of parameters? | What is the credibility of the model given your data? |
| **Outcome** | p-values : proba d’obtenir au moins autant d’évènements obtenus dans le dataset sous l’hypothèse nulle.  Confidence interval: if we were able to repeat the experiment a 100 times, in 95 cases the CI will contain the estimate | Mean of posterior distribution for all parameters  Credibility interval: for these data there is 95 % chance that this estimate is in the credibility interval |