Statistical modelling

Dr. Vladimir Grosbois
vladimir.grosbois@cirad.fr

CIRAD
UR AGIRs

Statistical modelling, a step beyond patterns

Patterns

– Comparison:

- Is the mean of one variable in one population similar or different from the mean of same variable in another population
- Does the mean of one variable in one population differ from an expected theoretical value

– Association

- Independent or related variations of two variables in a population

Causality: what are the factors that drive the variation in a focal response variable ?

Two or more than two variables are measured on individuals in a population

–In association analysis, the two variables have a similar status/role regarding the statistical analysis

–In statistical modelling

- One variable is the variable which variation is explained
  - Response variable
  - Dependent variable
- The other variables drive the variation of the response/dependent variable
  - Explanatory variables
  - Independent variables

What is a statistical model

Statistical model: description of how variation in the response variable (Y) is generated by relationships with explanatory variables (Xs).

- Parameters of a statistical model describe the relationships between Y and the Xs
- A model is fitted to the data (*i.e.* it's parameters are estimated) by maximizing the fraction of the variation in the response variable explained by the model
- Usually, several candidate models can be defined for describing the variation in the same response variable
- An important step in statistical modelling is to select one or a few models among the set of candidate models. Two important properties considered:
  - A model should be adequate: explain a significant fraction of the variation in the response variable
  - It should also be minimal in the sense that it should include only important relationships

An important concept for in statistical modelling: parsimony

- Given a set of equally good explanations for a given phenomenon the correct explanation is the simplest explanation
  - models should be simplified until they *are minimal adequate*
  - a variable is retained in the model only if it's removal from the current model results in a significant decrease in the fraction of variation in the response variable explained by the model.

- For a same fraction of explained variation
  - simple explanations should be preferred to complex explanations

Important models

Model	Interpretation
Saturated model	•Includes one parameter by data point. •Describe perfectly the data but is useless for inferring the mechanisms that generate variation in the response variable.
Maximal model	•Includes the effects of all the potential explanatory variables and all their interactions. •Usually used as a starting point for the model selection process
Minimal adequate model	**•Includes only the effects of the potential variables and of the interactions which removal results in a significant decrease in the fraction of explained variation** **•The description of the response variable retained**
Null model	•Includes only one parameter which represent the estimation of the response variable under the hypothesis that it is homogeneous in the population (no variation). •A kind of baseline model: models that do not explain more variation can be considered as irrelevant.

The standard process for model selection

- **Build the maximal model**
  - If it includes too many parameters (more than number of data points/10) → use a model containing only explanatory variables and interactions suspected as the most influential

- **Remove all explanatory variables and interaction which removal doesn't result in a significant decrease of the fraction of variation explained by the model**
  - Start by trying to remove the interactions

- **The model obtained after the removal of all the interactions and variables that do not influence the response variable is the minimum adequate model**
  - If it is the null model → none of the potential explanatory variable influence the response variable.
  - If it includes explanatory variables → they influence the response variable

Type of model depend on types of response and explanatory variables

Response	Explanatory	Statistical model
Continuous	All continuous	Linear regression
Continuous	All categorical	Analysis of variance (ANOVA)
Continuous	Continuous and categorical	Analysis of covariance (ANCOVA)
Continuous	Any combination of continuous and/or categorical variables	Linear model (LM)
Categorical, Count, Probability, Proportion	Any combination of continuous and/or categorical variables	Generalized linear models (GLM)

The quantification of variation

- **An important step in modelling is to remove from the maximal models all the unnecessary terms (explanatory variables)**

- **These terms (explanatory variables) are those which removal doesn't affect the quality of the description of the data by the model**

- **In other word their removal does not reduce to a great extent the quantity of variation in the response variable explained by the model**

- **The variation is quantified with**
  - Variance in ANOVAs, regressions, linear models
  - Deviance in generalized linear models

- **So, statistics used to decide to keep or to remove a term are based on :**
  - Variance in ANOVAs, regressions, linear models
  - Deviance in generalized linear models

Model syntax in R

In R the model is usually defined with a function including a formula as one argument

`aov(formula,options)` Analysis of variance and covariance (ANOVA, ANCOVA)
`lm(formula,options)` Regressions and linear models
`glm(formula,options)` Generalized linear models

Formula syntax

~	separates response and explanatory variables
+	addition of an explanatory variable and b
a:b	interaction between a and b
a*b	equivalent to a+b+a:b
a/b	b is nest in a

What are interactions and nested variables

- An interaction between the explanatory variables A and B means:
  - The effect of A on the response variable depending on the level of B
    - A is treatment against a pest/ no treatment against a pest
    - B is either of two varieties of maize
    - The response variable is the yield of experimental maize fields

→
- If the treatment increases the yield for one variety and not for this other
- the effect of treatment depends on the level of the variable variety
- this is modelled with an interaction between treatment and variety

- B is nest in A means:
  - Any level of B is found in only one level of A
    - A is either of two distinct regions
    - B is either of four breeds of cattle
    - The response variable is the weekly milk production

→
- Breed is nested in region if in the breeds found in one region are different from the breeds found in the other region

Important quantities for the interpretation of a statistical models

- The estimations of the model coefficients
  - Direction
  - and magnitude
  - of the effects of the explanatory variables

- The fraction of variation in the response variable explained by the model
  - Tells you if the model describes adequately the data
  - Quantifies the predictive power of your model