Meta-analysis

Meta-analysis

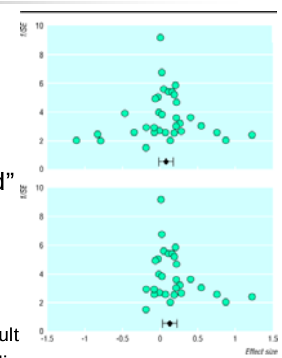
- The combination of many independent studies on one topic to:
  - Get an overall effect
  - See how effect varies with "meta-factors" like geographic location, type of study
- Focus on
  - Effect size/direction
  - Not on significance

The basic steps

1. Search of the literature
   - Commonly Web of Science or Google Scholar with keywords
2. Selection of literature
   - MUST have a priori selection criteria
     - Necessary data (p, r2, n) reported
     - Quality factors (control, double blind, etc)
3. Calculate effect sizes
4. Analyze effect sizes
   - Overall
   - Vs meta-factors

Publication bias

- A big problem
- Meta-analysis assumes all study results included
- In practice non-significant and contrary results often "file-drawered"
- Ways to deal with:
  - Measure
    - Funnel plot – n or 1/se vs. effect size
  - Estimates of non-significant results
    - E.g. research centers that record all result
  - Medicine leading to database of all studies



Measuring effect sizes

$$ES = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n_1+n_2-2}}} = \frac{\bar{X}_1 - \bar{X}_2}{s_{pooled}}$$

$$ES = t\sqrt{\frac{n_1+n_2}{n_1 n_2}}$$

$$ES = \sqrt{\frac{F(n_1+n_2)}{n_1 n_2}}$$

r

$$ES_{Zr} = .5\ln\left[\frac{1+r}{1-r}\right] \qquad r = \frac{e^{2ES_{Zr}}-1}{e^{2ES_{Zr}}+1}$$

See http://mason.gmu.edu/~dwilsonb/ma.html for more details

Adjust for size of study

- Standardized Mean Difference:

$$se = \sqrt{\frac{n_1+n_2}{n_1 n_2} + \frac{\overline{ES}_{sm}}{2(n_1+n_2)}} \qquad w = \frac{1}{se^2}$$

- Zr transformed Correlation Coefficient:

$$se = \sqrt{\frac{1}{n-3}} \qquad w = n-3$$

The Weighted Mean Effect Size

Study	ES	w	w*ES
1	-0.33	11.91	-3.93
2	0.32	28.57	9.14
3	0.39	58.82	22.94
4	0.31	29.41	9.12
5	0.17	13.89	2.36
6	0.64	8.55	5.47
7	-0.33	9.80	-3.24
8	0.15	10.75	1.61
9	-0.02	83.33	-1.67
10	0.00	14.93	0.00
		269.96	41.82

- Start with the effect size (ES) and inverse variance weight (w) for 10 studies.
- Next, multiply w by ES.
- Repeat for all effect sizes.
- Sum the columns, w and ES.
- Divide the sum of (w*ES) by the sum of (w).

$$\overline{ES} = \frac{\sum(w \times ES)}{\sum w} = \frac{41.82}{269.96} = 0.15$$

Mean, Standard Error, Z-test and Confidence Intervals

Mean ES

$$\overline{ES} = \frac{\sum(w \times ES)}{\sum w} = \frac{41.82}{269.96} = 0.15$$

SE of the Mean ES

$$se_{\overline{ES}} = \sqrt{\frac{1}{\sum w}} = \sqrt{\frac{1}{269.96}} = 0.061$$

Z-test for the Mean ES

$$Z = \frac{\overline{ES}}{se_{\overline{ES}}} = \frac{0.15}{0.061} = 2.46$$

95% Confidence Interval

$$Lower = \overline{ES} - 1.96(se_{\overline{ES}}) = 0.15 - 1.96(.061) = 0.03$$

$$Upper = \overline{ES} + 1.96(se_{\overline{ES}}) = 0.15 + 1.96(.061) = 0.27$$

Interpreting Effect Size Results

- Cohen's "Rules-of-Thumb"
  - standardized mean difference effect size
    - small = 0.20
    - medium = 0.50
    - large = 0.80
  - correlation coefficient
    - small = 0.10
    - medium = 0.25
    - large = 0.40
  - odds-ratio

Homogeneity Analysis (Q statistic)

- Homogeneity analysis tests whether the assumption that all of the effect sizes are estimating the same population mean is a reasonable assumption.
- If homogeneity is rejected, the distribution of effect sizes is assumed to be heterogeneous.
  - Single mean ES not a good descriptor of the distribution
  - There are real between study differences, that is, studies estimate different population mean effect sizes.
  - Two options:
    - model between study differences
    - fit a random effects model

Regression vs. meta-factors

- ES~continent+phylum+body_size
- Use Q analysis
- Basically partioning variance (ANOVA)

Outputs

- Effect Size ES
  - Cohen's "Rules-of-Thumb"
    - standardized mean difference effect size
      - small = 0.20
      - medium = 0.50
      - large = 0.80
    - correlation coefficient
      - small = 0.10
      - medium = 0.25
      - large = 0.40
- Q
  - Homogeneity
    - Variation in effect sizes explainable by one underlying effect
  - Heterogeneity
    - Need additional factors to explain variation in effect size – due Q-mode (ANOVA) analysis

Limitations of meta-analysis

- GIGO – garbage in-garbage out
  - Is lumping a lot of bad studies better than one really good study?
  - Should lots of bad studies weigh equally to good studies?
- If a meta-analysis is needed are effect sizes too small to be biologically significant

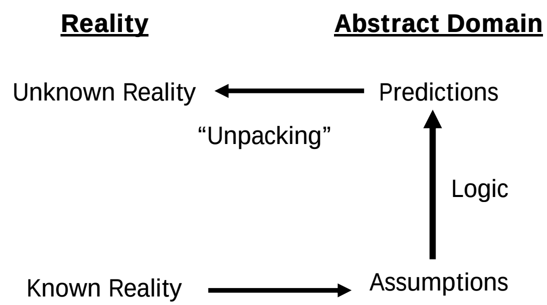
Modeling

What is a model?

A model is:

- An abstraction of the real world into a domain where logical inference (deduction) can be applied to transform explicit assumptions into new predictions

Modeling



Manfred Eigen

- A theory has only the alternative of being right or wrong. A model has a third possibility: it may be right, but irrelevant."
- Model has to accomplish something in the REAL world or it is irrelevant
  - Huge number of models only accomplish things in the mathematical world

Domains of logical inference (types of models)

- Mathematical
  - Algebraic formulas
  - Geometric
- Graphical (isocline analysis)
- Verbal/qualitative
- Computer algorithm
  - Deterministic
  - Monte Carlo (random)

Checkered history of modeling in ecology & evolution

- Sharon Kingsland (a historian of ecology)
  - tension between modelers vs field workers
  - closely related general results vs. specific
  - Central theme in ecology
- Many debates
  - Lotka Volterra 1920-1930's
  - Much other mathematicization in 1930's
  - 1950's Nicholson & Bailey (density dependence) vs Andrewartha & Birch
  - Late 1960's and 1970's MacArthur, May
  - 1980's backlash – emphasis on field experiments and advanced statistics
  - Restored balance now?

Evolution & Behavior

- Evolution has always been more friendly to modeling
  - Harder to do experiments
  - Mendel's laws give precise starting point
- Behavior
  - Idea of optimization is a major thread, fits well into models
  - Somewhat controversial

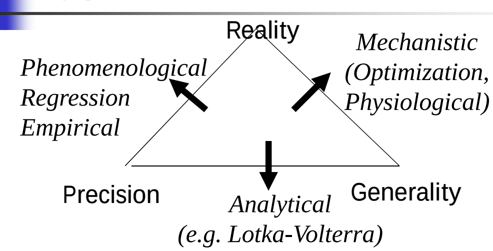
Model goals

- Purely predictive
  - Useful for management – will species X go extinct, will water levels rise in Everglades?
- Explanatory
  - Claims to capture (or elucidate) underlying mechanism
  - Gets at causality
  - ANOVA/regression have simple explanation in mind: "temperature" causes "abundance"
    - Phenomenological
  - More advanced models:
    - Fluid dynamics of stomach→digestion rates
    - Group size optimizes reproductive rate
    - Temperature affects physiology affects resource acquisition affects abundance

Types of parameters

- Mechanistic – have units, directly definable, measurable independent of model
  - Maximum running speed
  - Reproductive output in a group of 7
- Phenomenological – measured in context of model
  - Derived from regression
  - Lotka-Volterra competition parameters (α,K)
  - Parameters of Type II functional response
- Is r a mechanistic or phenomenological parameter

Types of models



- Levins 1966 – can only do 2 of 3 goals
- May: strategic vs tactical models

Types of predictions

- Exact:
  - Abundance next year will 42 indiv/ha
  - Group size will be 7 individuals
- Qualitative
  - Abundance will go up
  - Groups will form
  - Evolution will increase body size
- Relational
  - Abundance goes up when temperature goes up
  - Exact relational N=23*T/(15+T)
- Rosenzweig's dipswitch
  - One qualitative prediction has a 50/50 chance of being right
  - Four qualitative predictions have a 2^{-4}=1/16=6% chance of being right

Sensitivity analysis

- Any parameter that is not known very well should have sensitivity analysis
- E.g.
  - If temperature goes from 20 to 22
  - Does abundance go from 100 to 101, 110, 150, 300?
- Sensitivity analysis allows us to:
  - Focus on most critical parameters
  - Put extra effort into estimating sensitive parameters
- Combinatorial explosion
  - 5 levels of 3 parameters is 125 cases to explore!
- Stochastic complexity
  - If use random numbers in model need replicates

Aviles 1993

- Used ANOVA
- Sexratio~ColonySize+ MigrationLevel+ ColonyMortality
- Several MonteCarlo replicates/cell
- Gave p-values
- Gave causality?
- Familiar to ecologists

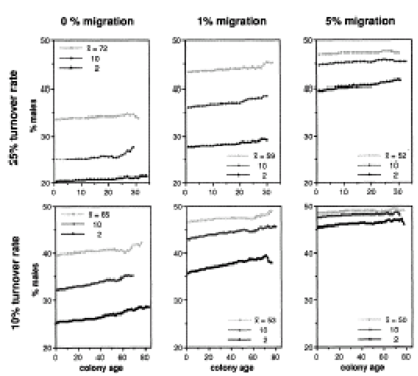


FIG. 5.—Average sex ratio of colonies of different age classes, averaged over the last 500 generations at equilibrium. Curves within each figure correspond to the three propagule sizes investigated (2, 10, and ~1/2 of the parental colony). The three migration rates (0%, 1%, and 5%) are shown in different graphs along the columns, and the two turnover rates (10% and 25%), along the rows. The threshold size for proliferation was set at 200 individuals, and there was no facultative control of the production of at least one male.

Calibration vs validation

- Calibration=estimation=curve fitting
  - Post hoc twiddling of model parameters to make it fit data
  - Is a good way to get parameter estimates
- Validation
  - A priori prediction (no part of model derived from data) tested against data not used for calibration
  - Can:
    - Get parameters from literature
    - Calibrate on half the data, validate on half
    - Calibrate on past, validate on future
    - Calibrate on one site, validate on another
- Very common fatal error – no validation!

A good model does (Aber):

- Gives an explicit list of parameters
  - Gives units of each parameter
  - Gives sources of estimates for parameters
  - Describes sensitivity analysis of parameters
- Calibrates the model
  - Estimates unknown parameters and demonstrates how much our understanding depends on them (sensitivity)
- Makes novel predictions
  - Predictions must be testable
- Validates the model
  - Tests predictions vs. ***INDEPENDENT*** data (not used in calibration)

The fundamental modelling decision

- How complex should the model be?
- Einstein Principle: "A scientific theory should be as simple as possible, but no simpler"
- Every model is missing real complexity
  - Not a valid criticism of a model
  - Must argue that missing a piece that substantially changes results
  - Only full model is the size of the universe!

Interpretation of results

The backend - interpretation

- I've done the experiment
- I've done the statistics
- Now I have to write the conclusion
- What does my experiment really tell me

Central issue

- Multiple causality
- Physics: one force explains 90%+
  - 2nd usually explains 8% more
- Ecology: A 28%, B 19%, C 12%, D 3%, miscellaneous 38%
- Easy for balance between forces to shift with context

Keep in mind

- When interpreting results keep in mind
  - Scale of experiment vs scale of conclusions
  - Scope of experiment vs scope of conclusions
  - Homogeneity of experiment vs homogeneity of conclusions
  - Sense vs. referential meaning
  - Hidden treatment effects
  - Correlation vs. causation

Scale

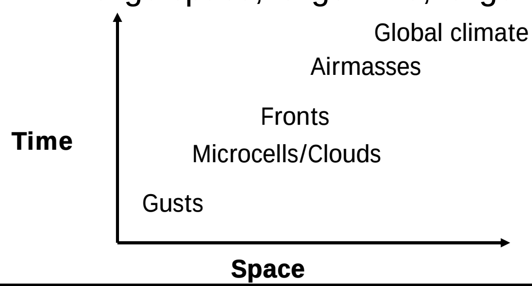
- Ecology/evolution exists on many scales
  - 5-6 orders of magnitude
- Scale of time: minutes-geologic eras
- Scale of space: mm^2-globe
- Scale of taxon: individuals/populations/species/genera/etc
- Compare with cell biology
  - Spatial scale fairly constant, time scale fairly small range (2-3 orders magnitude), taxon irrelevant
- Compare with physics
  - Laws hold across enormous scales (e.g. law of gravity)
  - Only two scale breaks – three scale domains
    - Subatomic (quantum mechanics), Cosmological (General relativity), In between (Newtonian)

Scale and process

- Just as in physics, critical processes change with scale
- But in EEBB critical processes change with every order of magnitude in scale
- What controls productivity
  - Square meter – what plants growing there
  - Hectare – soil & drainage, disturbance (e.g. fire)
  - Km^2 – topography, slope aspect, rivers
  - 1000 km^2 – climate
  - Globe – CO_2 levels, size of ice cap
- What controls population level over time
  - Less than a generation – stochastic deaths
  - A generation – controls on fecundity (food), predation
  - Many generations – climate, productivity

Scales often change in synch

- Processes important to small space are often short time, limited taxon
- Large space, large time, large taxon



Tradeoff

- No a priori “best” scale
- Arguably scales of most importance to humans are large (conservation, global warming)
- Yet experiments become harder, more expensive as increase scale
  - How often do you run experiment on 20 species just to see if it generalizes?
  - Krebs 1 km^2 enclosures - $7,000,000
  - How often are experiments done on 1-100 m^2
- Can we extrapolate to km^2’s from m^2’s?

Transmutation

- Dominant processes change with scale, but even the same process takes on different forms as you aggregate
- O’Neill 1979 – Transmutation principle
- If aggregate across individuals, changes
- Minimum food threshold for individual to mature
- Threshold varies normally
- Mortality vs food for whole population looks logistic

Closure and scale

- Most theories assume the system is closed from external influences (immigration, influx of nutrients)
  - Systems become less closed at smaller scales
- Most theories assume system at equilibrium
  - True most of the time for long time scales, large spatial scales (averaging)

Implications of scale I

- Unproven that you can generalize results beyond the spatial, temporal, taxonomic scale of experiment
- If you want to make a general claim burden of proof is on you to justify extrapolation or “transmutation”
- Sometimes appropriate
  - Energetics and life history of mammals are very similar (albeit some well known scalings with body size)

Implications of scale

- Larger scales
  - Highly relevant to humans
  - Often more appropriate for theories
- Growing trend to large scale ecology
  - Doesn’t replace, just supplements smaller scales
- Yet hard to experiment
- Techniques for large scales
  - Metaanalysis – statistics on top of dozens of papers with statistics
  - Comparative biology – study of data on many species (usually from many papers)
  - Ecoinformatics – ecology + databases
  - GIS

Scope

- What piece of scale
- If I do an experiment at Mt St Hillaire station on deer, what other space/taxonomic units does it apply to?
- Spatial
  - Mt St Hillaire, S. Quebec, Canada, Globe
- Taxonomic
  - Your species, your genus, all mammals, all organisms?
  - Model organism?

Homogeneity

- Dominant processes can change with amount of environmental heterogeneity
- Egs
  - Highly disturbed environments colonization controls community composition, but competition/predation control in static environments
  - Variability on scale > generation → phenotypic plasticity, variability on scale<<generation → physiological homeostasis
- Contradiction
  - Experiments try to minimize variance
  - Understanding real world often needs heterogeneity

Microcosms

- Laboratory work, bottle experiments, microcosms
- Famous examples
  - Gause's paramecium
  - Park's *Tribolium* beetles
  - Cushing et al Tribolium
  - Drake – metacommunities
  - Lenski, Bohannan, Bell - evolution
  - Davis et al – Drosophila in growth chambers for species ranges and global warming
- Often criticized for applicability to field
- My opinion – highly generalizable except need special attention to homogeneity issue
  - eg. competitive exclusion in Gause, Park would never happen with spatial heterogeneity
  - Is this a good thing or bad thing?

Correlation vs. causation

- Correlation does not demonstrate causality
  - Could go either direction
  - Could be co-correlated with 3rd factor
- Often cited as reason to do experiments
- Other ways around
  - Test of a priori hypotheses
  - Observational controls
  - Natural experiments

Hidden treatment effects

- Huston
- Treatment can incorporate many effects
  - Can you really isolate one process if treatment changes many processes
- Simple treatments (eg. +Nitrogen)
  - One sprayer/manure spreader spreads contaminants, bacteria, water, etc
  - Process controls & interspersion/no psuedo-replication both help
- Complex treatments much subtler
  - Enclosure effects
- Stability increases with diversity
  - Tilman: plots seeded with random subsets of species with S increasing →true
  - Huston: probability of including basic fundamental groups (nitrogen fixing, C3 vs C4, etc) goes up
  - Current consensus: Huston right but not clear if this is a criticism or mechanism supporting, other factors also at work

Sense vs. referential

- Psychology
- Sense – we know what something means even if abstract concept (eg "happy")
- Referential – we don't measure happy, we measure score on a test battery, hormone level, etc
- Need to think carefully about fact experiments deal with referential measurments
- Is ecology different?
  - Stability? Diversity? – dozens of measures of each

The ultimate goal

- To be able to:
  - Identify the list of forces
  - To assign relative strength (% variance) to each
  - To explain how the strengths vary with:
    - Scale
    - Scope
- "[one should erect a] two- or three- way classification of organisms and their geometrical and temporal environments, this classification consuming most of the creative energy of ecologists. The future principles of the ecology of coexistence will then be of the form 'for organisms of type A, in environments of structure B, such and such relations will hold' " Robert MacArthur 1972

Wrapping up ...

Scientific inference

- Deduction (Logic - Aristotle) vs Induction (Empirical data - Bacon)
- Falsification (can't prove, only disprove – Popper) vs. Accumulation of evidence (Lakatos)
- Objective (Bacon,Popper) vs Social process (Kuhn, Lakatos)
- Instant disproof (Popper) vs. gradual (Lakatos) vs gradual with instantaneous bursts (Kuhn)
- Experiment (better) vs. observation (better than nothing, first step)

Statistics

- The study of conclusions about probabilistic data
- Necessary because humans are bad at probability
- Four approaches:
  - Frequentist (null hypothesis)
  - Liklelihood (and AIC & model comparison)
  - Data randomization/Boostrap/Reshuffle
  - Bayesian (likelihood+informative? prior)

GLM (OLS)

- Regression/ANOVA/ANCOVA
  - Discrete variables are "coded" (0/1 w/ n-1 columns)
  - Interactions with discrete variables if slopes change
- $Y = X*b + N(0, \sigma I)$
- Assumptions:
  - Linearity
  - Errors are normal
  - Errors are homoscedastic
  - X is not collinear
  - Errors are independent
- OLS is fairly robust to most violations
- All the rest of statistics are about breaking one or more of these assumptions

3 modes

- Frequentist
- Bayesian
- Monte Carlo/boostrap

- Separate calculation from inference mode
  - Inference modes: p vs null, confidence intervals on parameters, model selection

GLIM

- Nonnormal errors with nonlinear relationships
  - Only certain combinations
  - Poisson/log
  - Binomial/logistic
  - Gamma/log or 1/x
- Uses likelihood/AIC

GLS & LMM

- GLS
  - Correlation in the errors
  - Covariance matrix instead of σI
  - Can't estimate n*n/2 parameters
  - Need a structure with a few parameters
    - Correlation: block, AR1, exp(-kd), phylogeney
    - Covariance: WLS (heteroscedastic)
- LMM
  - Data has hierarchical structure
  - Experimental design=split plot
  - Observational=hierchical models
  - At a minimum the errors to avoid pseudoreplication
  - More interestingly: predictors at multiple levels
- Bayesian hierarchical models

Experimental design

- Thinks about the types of variables and types of statistics before starting!
- Think about hypotheses before starting!
- Replicates more important than # of treatments
- Calculate power!
- Blocking & repeated measures increase power by isolating known, uninteresting sources of variance

Modern regression

- Robust regression
  - nonnormality & outliers
- Quantile regression or envelope regression (95 percentile line)
  - One way to address multiple causality
- Local regression or smoothing
  - Puts a line through the data with no underlying model
  - Subjective degree of smoothing
- Nonlinear regression
  - If models are nonlinear

Modern regression multivariate

- Predict y as function of many x
- Local
- GAM – sum of local regression in each x
- Neural Nets
- CART
  - Tree
  - Balance of interpretability and generality
  - Beware circular testing

Multivariate

- Many continuous x variables, linear
- Extensions of GLM
  - Multiple Y's
- Data reduction methods
  - Visualization
  - PCA (and MDS/PCoA)
    - Finds major axes of variation, collinear redundancy – reduces data dimensionality
  - Clustering – lumps data into groups

Time and space

- GLS – error terms are correlated
- Key assumption correlation is a function of distance between points
  - Normally assume exponential decay of correlation to zero
- Decomposition Trend/Periodicity/Residuals

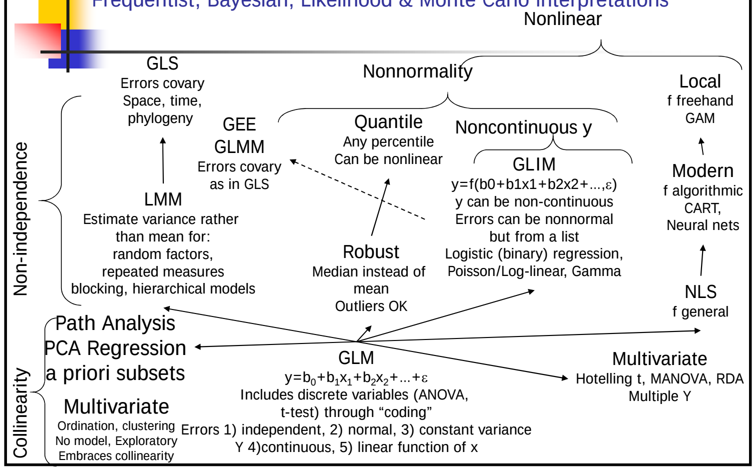
Generalizability & modeling

- For a given result how general can we conclude:
  - Think about scales (time, space, taxon)
  - Think about scope
  - Mechanism important?
- Modeling
  - Stepping into the abstract domain where we can use deduction to make predictions
  - Distinction between calibration and validation

A model y=f(x,ε):

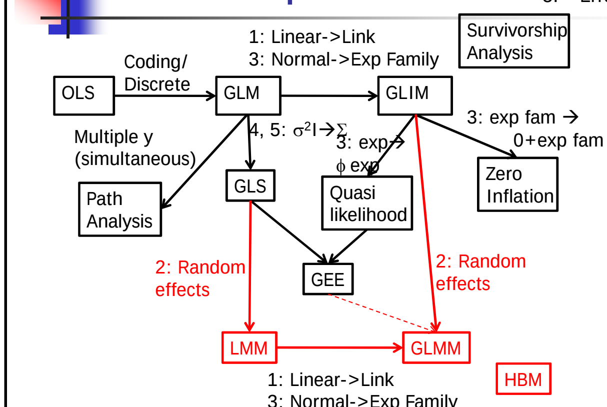
prediction, hypothesis test, exploration

Frequentist, Bayesian, Likelihood & Monte Carlo interpretations

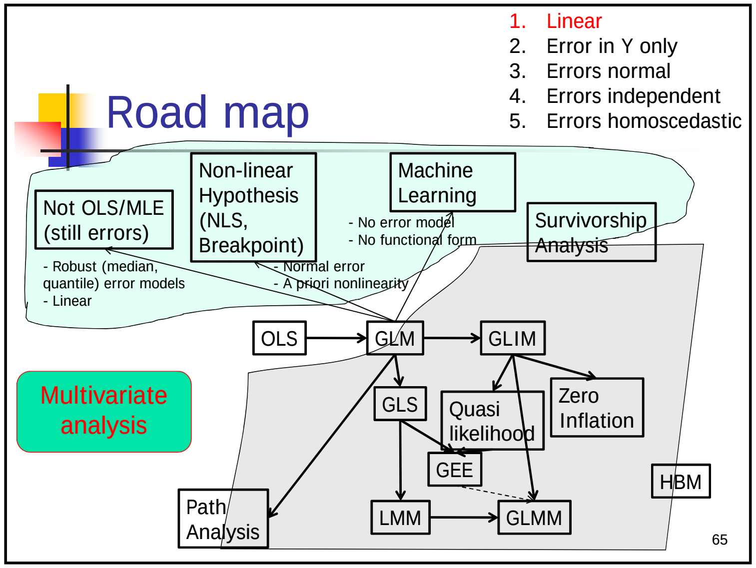


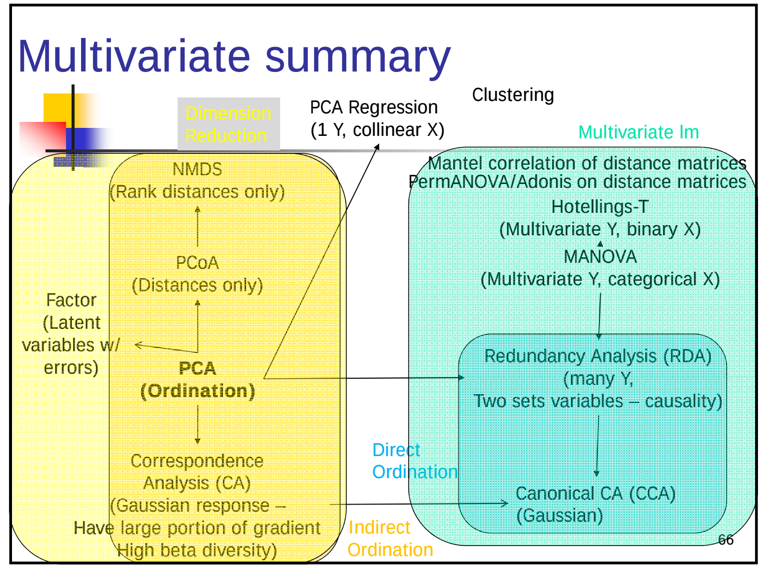
Road map

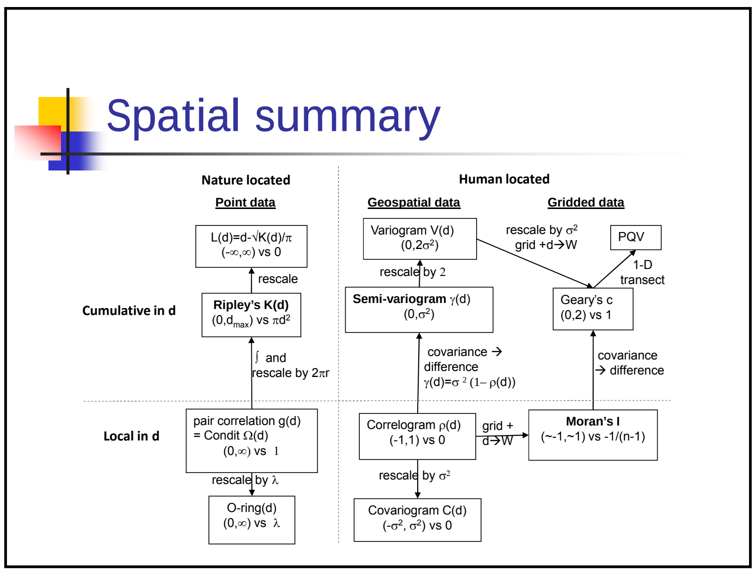
1. Linear
2. Error in Y only
3. Errors normal
4. Errors independent
5. Errors homoscedastic



64







7 things to remember

1. There are a lot of ways to do science
   - Usually a balance of approaches is best
2. Statistics does not have one right way
   - You have to think & justify
3. You know more techniques than 90% of ecologists
   - Balance cutting edge vs. familiarity
4. Prepare for your experiment
   - A priori hypotheses, what statistics, power, interspersion & replication
5. Find strong tests
   - Avoid circular reasoning (calibration w/o validation)
   - Use some form of alternative hypotheses (model comparison or null)
6. Likelihood & boostrapping can deal with more complex cases
7. Our field is complex enough you may need collaborations
8. “There are worse things for a scientist than to be wrong. One of them is to be trivial.”