ECo 602 - Analysis of Environmental Data

The Limits of Group I Methods

Michael France Nelson

Eco 602 – University of Massachusetts, Amherst
Michael France Nelson

General Linear Model Limitations

Group 1: [General] Linear Models

Four key assumptions:

- Normality: normality refers to the model residuals
- Constant variance a.k.a homoskedasticity, a.k.a. homogeneity
- Independent observations
- Fixed x: no measurement error in our predictor variables

Group 1 requirements:

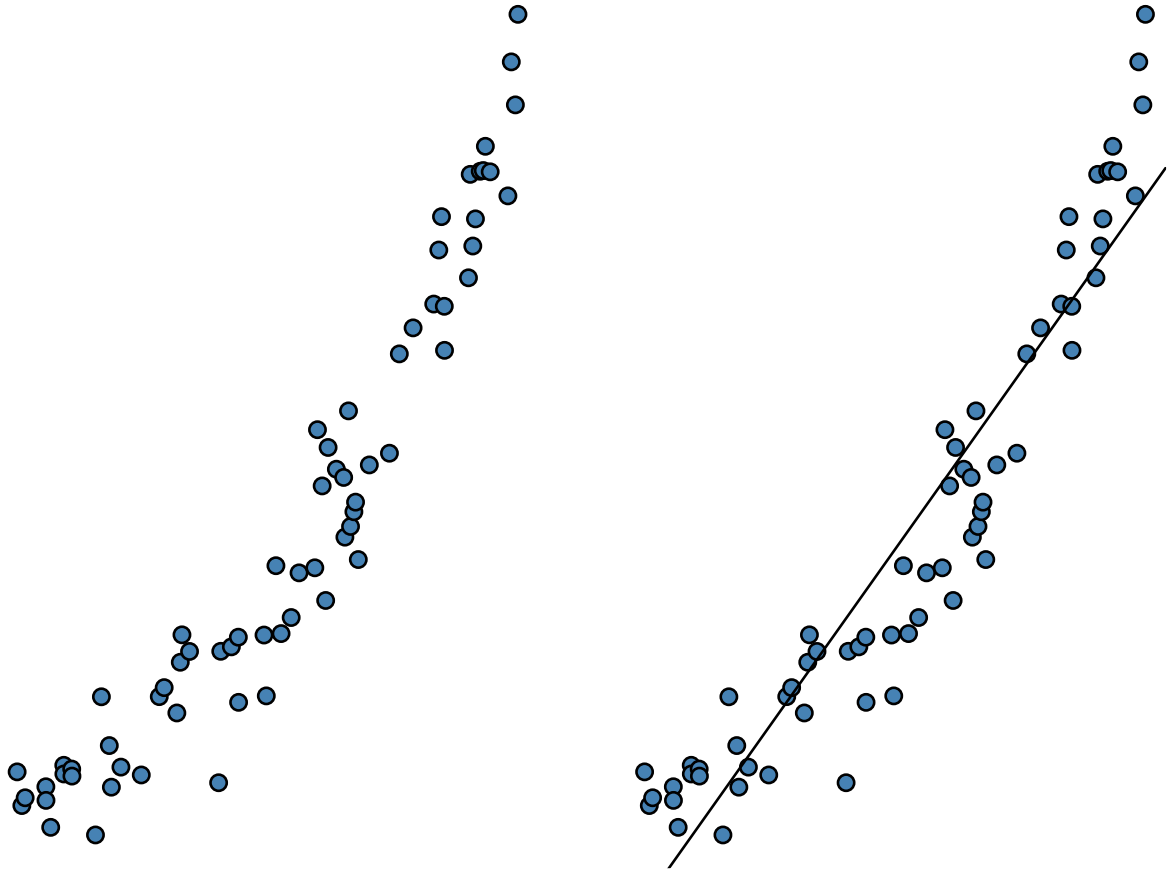
- Group 1 models are linear in the parameters
- Group 1 models have a single continuous response variable

Challenge 1: Non-Linearity

Linearity in the Parameters

Group 1 models have to be *linear in the parameters*, but they can still handle certain types of nonlinear relationships between predictors and responses.

- From previous deck: What does linear in the parameters mean?



Nonlinearity: Some Options You Can Try:

If the relationship between our data doesn't seem linear, there are a few options we can try:

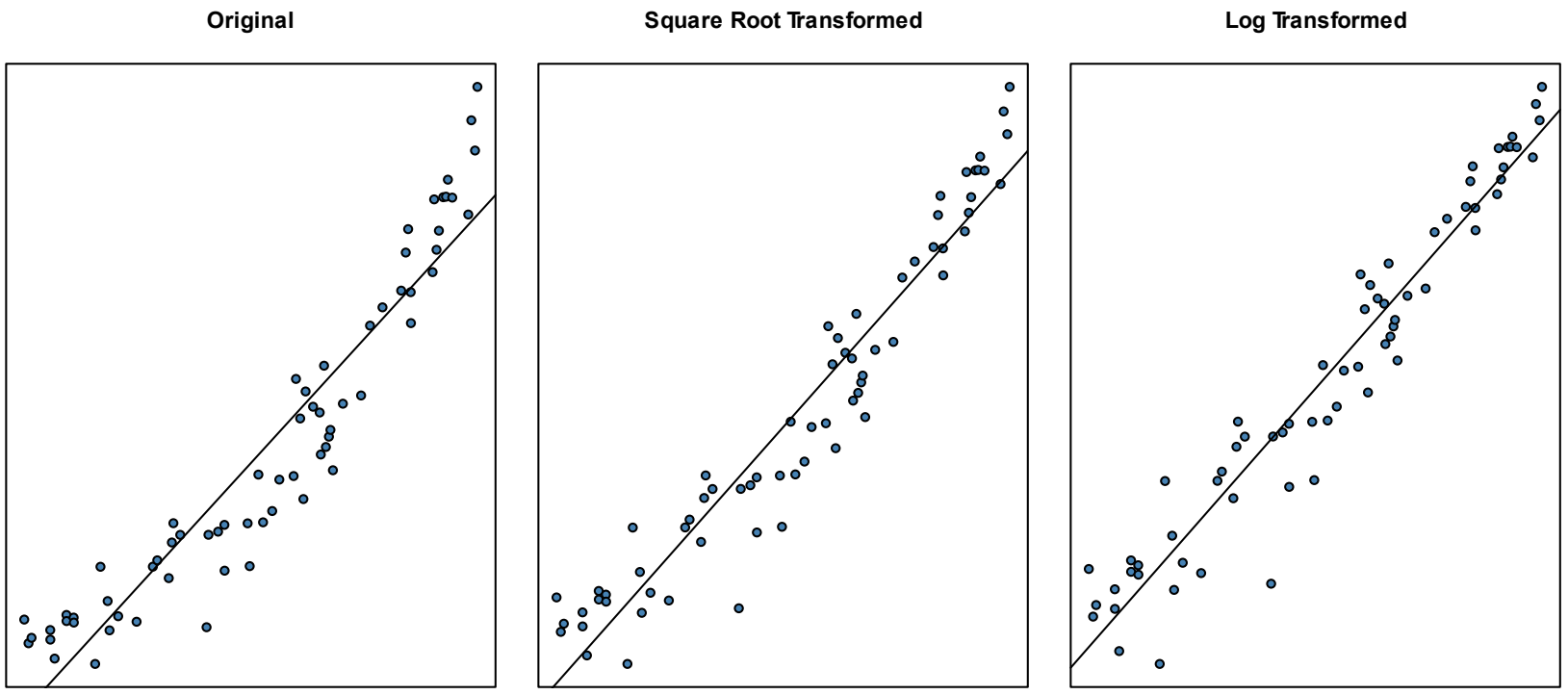
1. Data transformation (often the logarithm)
2. Adding polynomial or power terms
3. Adding interaction terms

Each option has pros and cons

Data transformations

Transforming the response can help with:

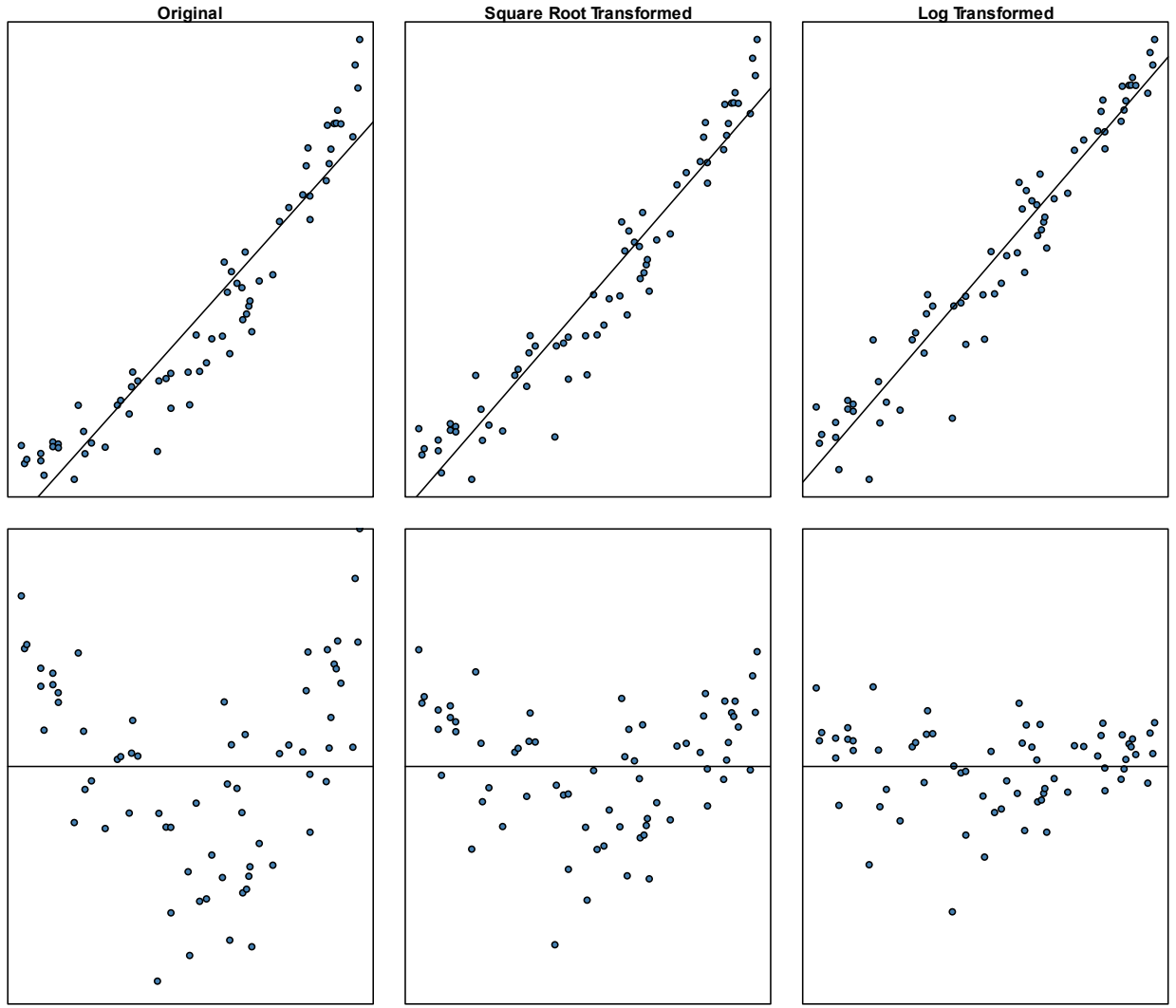
1. Stabilizing the variance: log transformations
2. Linearizing the relationship



Data transformations

Transformations: data and residuals

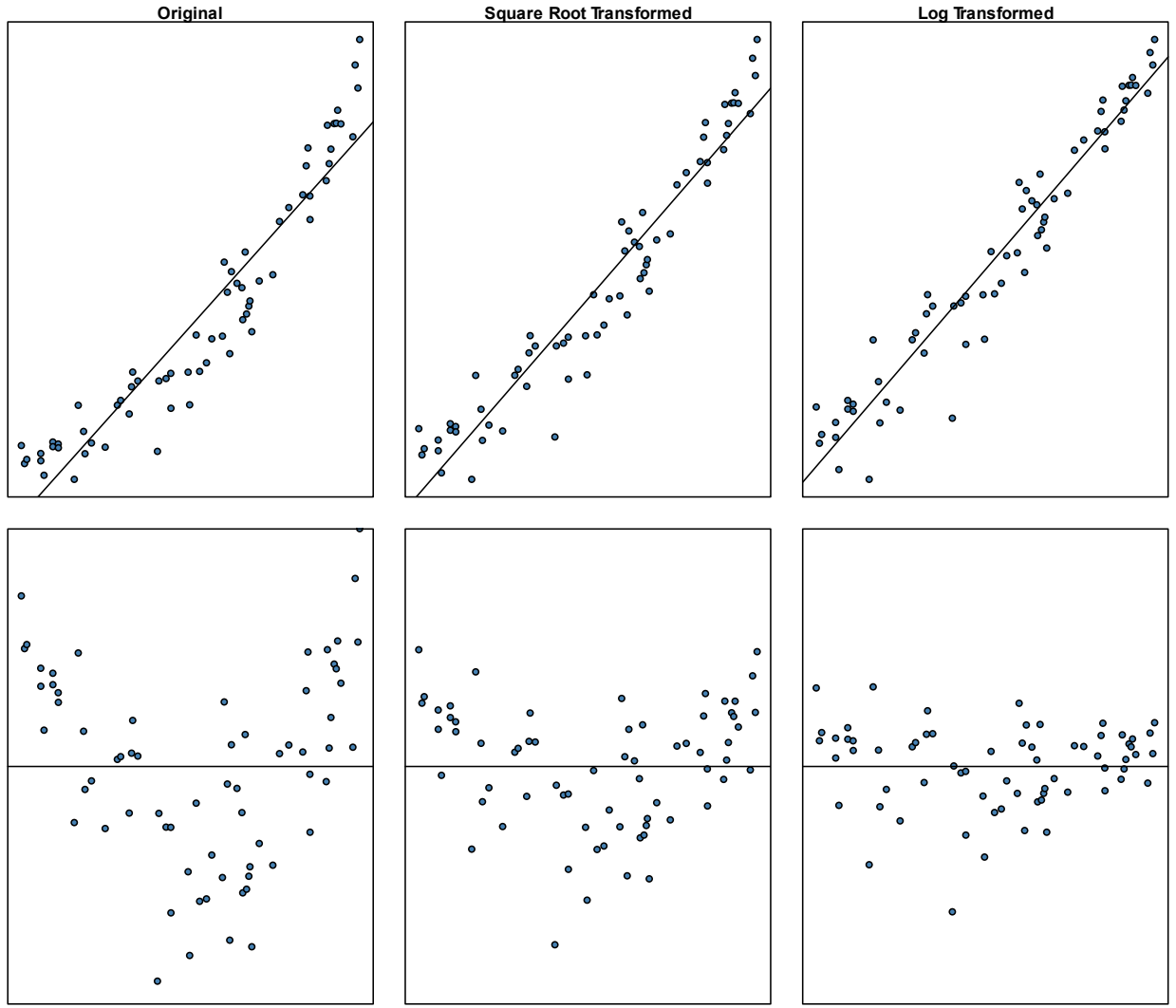
- Top row: data with linear fit
- Bottom row: model residuals



Data transformations: Linearizing

Transformations: data and residuals

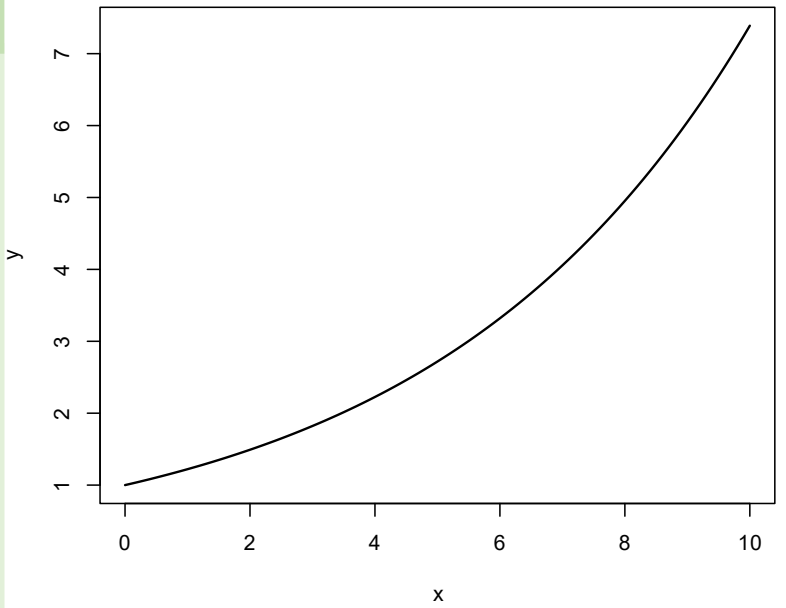
- Original data are nonlinear
- Square root transformation improves linearity a little
- Log-transformation makes the data linear.

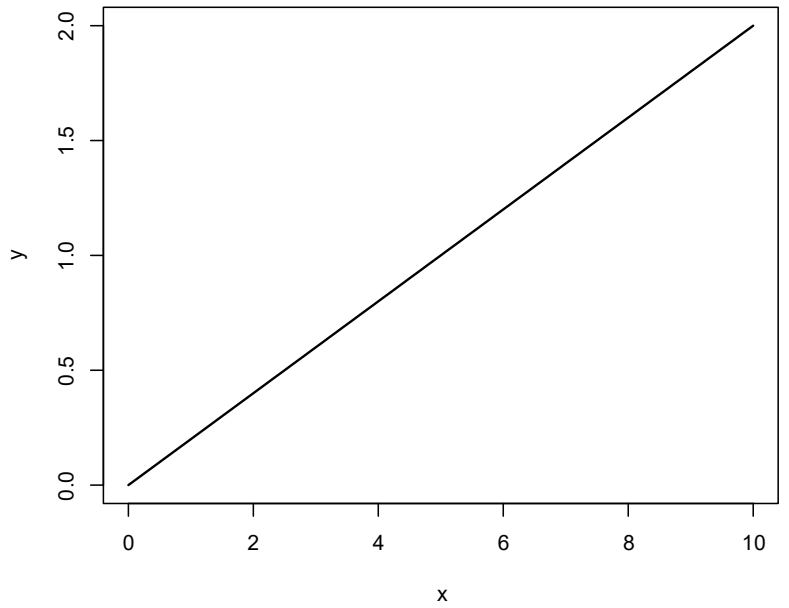


Log transformations

Effects of log-transforming the response:

- **Can make super-linear functions linear.**
  - 'concave up' from calculus
  - Functions whose slope increases with x, like the exponential
- Stabilizing the variance
  - Can help with the megaphone shape in the residuals





Super linear

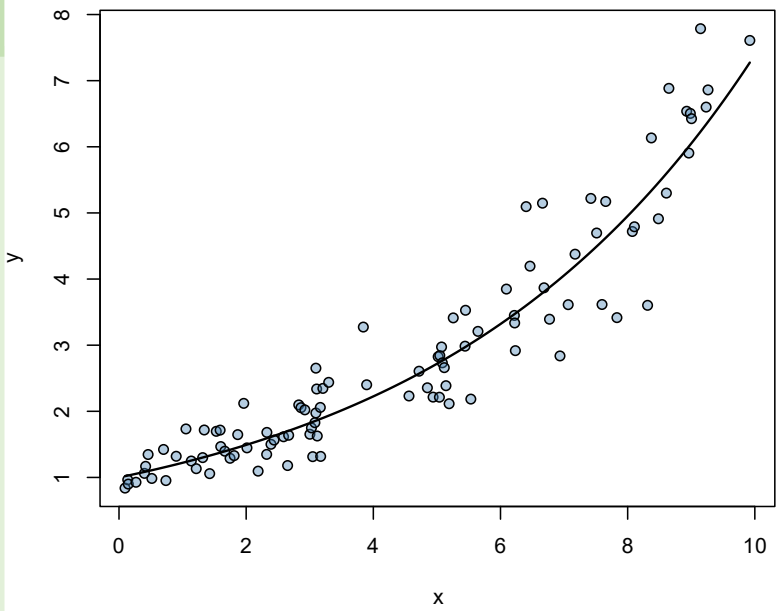
Log-transform

Linearized

Log transformations

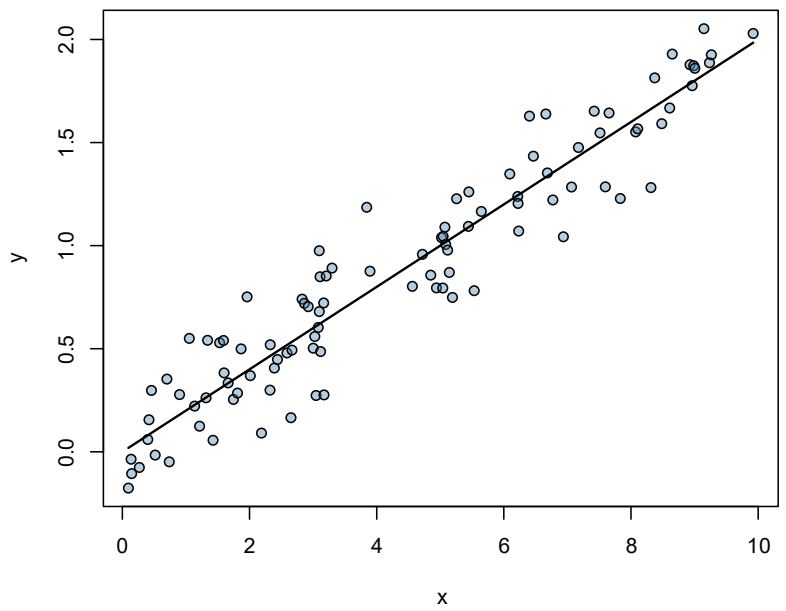
Effects of log-transforming the response:

- **Can make super-linear functions linear.**
  - 'concave up' from calculus
  - Functions whose slope increases with x – like the exponential
- Stabilizing the variance
  - Can help with the megaphone shape in the residuals



Super linear

Log-transform



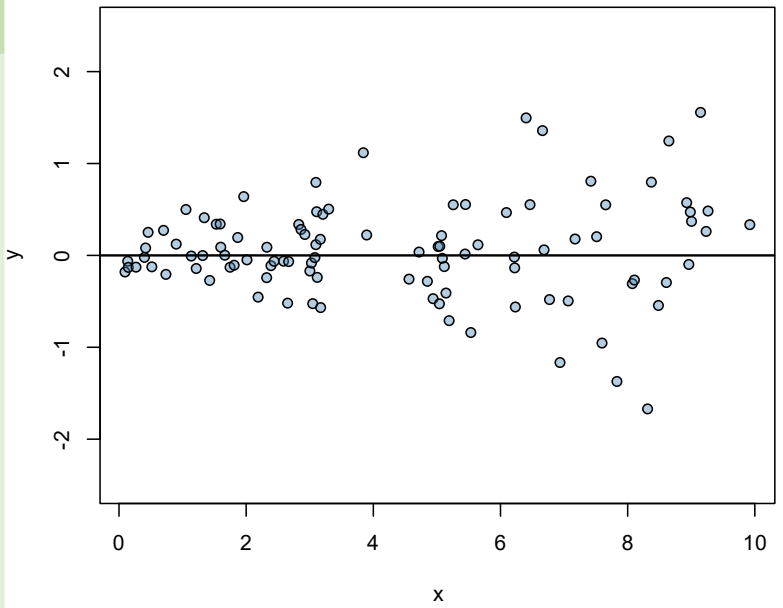
Linearized

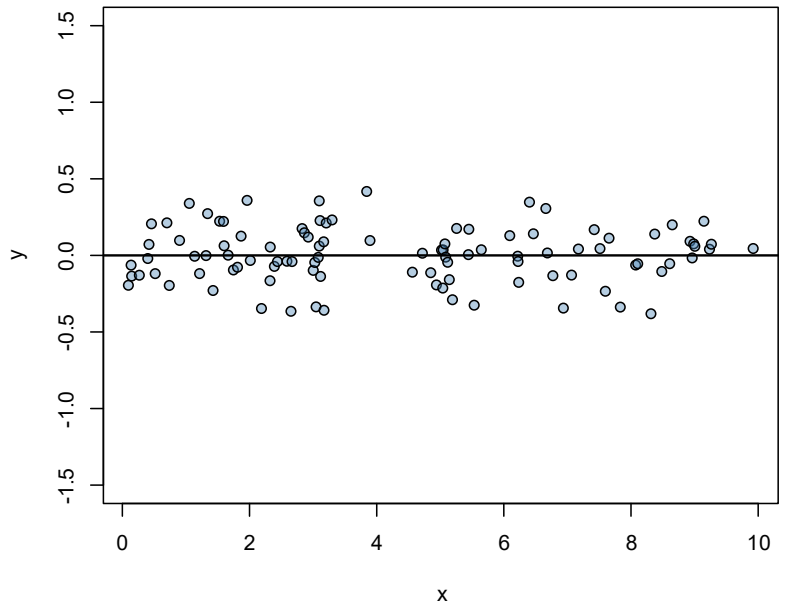
Log transformations

Effects of log-transforming the response:

- Can make super-linear functions linear.
  - ‘concave up’ from calculus
  - Functions whose slope increases with x – like the exponential
- **Stabilizing the variance**
  - Can help with the megaphone shape in the residuals

Residuals





Megaphone - heterogeneous

Log-transform

homogeneous

Log transformations: challenges

Transformations affect both the deterministic and stochastic model components

- Sometimes this helps: it often fixes non-constant variance
- Transformed model coefficients can be difficult to interpret or explain to others.
  - Coefficients are now in terms of proportional increases/decreases not constant amounts.
- It's not always straightforward to 'back-transform' coefficients.

Log-Transformed Coefficient Interpretations

Transformed variables are more difficult to interpret

Recall the linear slope coefficient interpretation:

- "Every 1% increase in survival was associated with 2 additional killed trees per hectare per year."

Log-transformed coefficient:

- "Within a stand, a 1% increase in beetle survival was associated with a 6% proportional increase in tree mortality rate over the mortality rate of the previous year."

Additional model terms

Polynomial Terms	Interaction Terms
Polynomial regression: raise the predictor variable to a constant power. • Nonlinear predictor/response relationship • You have to decide on the power – it's not optimized from the data. • Model parameters are still linear. Interaction Terms	

Interaction terms

Example model:

$$y_i = 1.3 + 2.0x_1 + 2.4x_2 + 2.3x_1x_2 + \epsilon$$

- 1-unit increase in predictor 1 associated with 2-unit increase in response.
- 1-unit increase in predictor 2 associated with 2.4-unit increase in response.
- What if we simultaneously increase predictor 1 and 2 by one unit?

Interaction terms

Interaction terms can be synergistic or inhibiting

Stochastic part = normal distribution
Intercept > predictors set to zero > alpha
Slope coefficients > beta

$$y_i = 1.3 + 2.0x_1 + 2.4x_2 + 2.3x_1x_2 + \epsilon$$

- What if we simultaneously increase predictor 1 and 2 by one unit?
- Without an interaction we would expect an increase of 4.4, the sum of β_1 and β_2.
- With the interaction we get an increase of 6.7 because the value of the interaction slope is 2.3!
  - This is a *superlinear* increase: synergistic

Nonlinearity: Group 2 approaches

Transforming data has some serious drawbacks.

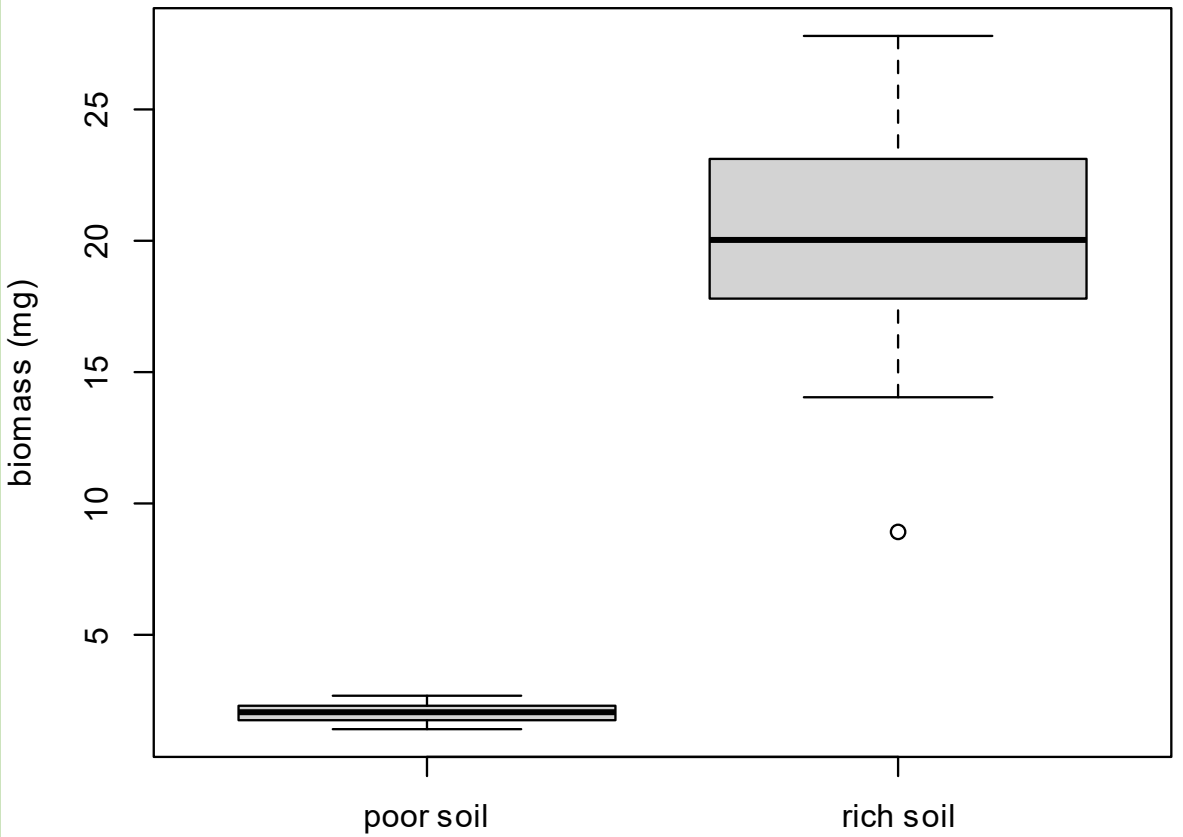
Some extend linear techniques can directly describe nonlinear relationships:

- Generalized [Nonlinear] Least Squares: GLS and GNLS
  - still require: independent observations, normal errors
- Generalized Linear Models: GLM
  - These can handle certain kinds of non-linearity.

Challenge 2: Heterogeneity: Non Constant Variance

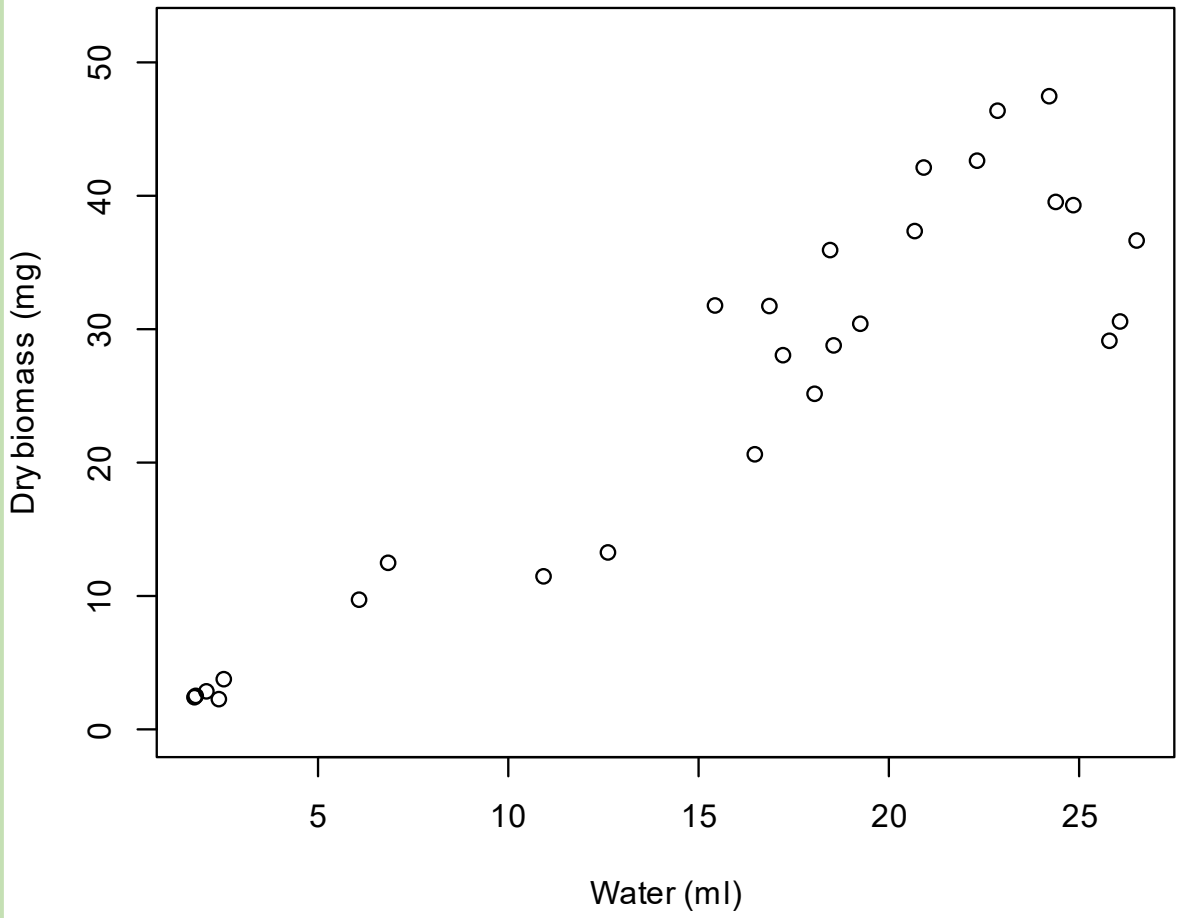
What is heterogeneity?

- Group 1 requires the variance to be the same at every value of the predictors!
- This is often unrealistic: hypothetical plant growth
  - Plants grown in poor soil have 2.0 grams biomass (on average)
  - Plants grown in rich soil have 20.0 grams biomass (on average)
  - Do you expect the magnitude of variation to be the same in each group?



What is heterogeneity?

- Group 1 requires the variance to be the same at every value of the predictors!
- This is often unrealistic: hypothetical plant growth
  - Smaller plants will likely have less absolute variability



Dealing With Heterogeneity

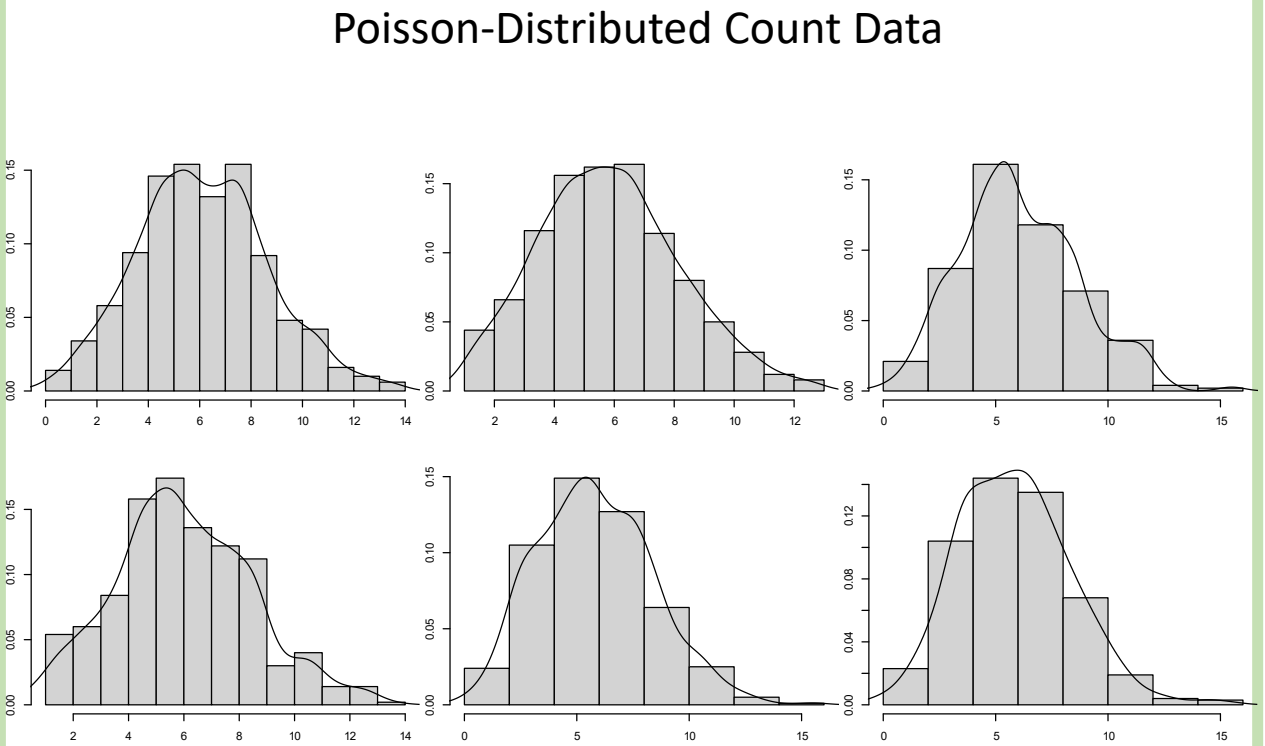
- Log transformations often help!
- Weighted Least Squares
- Regression using a variance/covariance matrix (instead of a single error term)
  - Model variance as a function of a predictor
- Simulation: bootstrapping and MC
  - Obtaining null and alternative distributions for confidence/significance estimates
- Adjusting standard errors for non-constant variance: standard error is a function of the predictor variable.
  - This can work well for continuous predictors.
- Random effects and multi-level models.
  - These can work well for categorical predictors.

Challenge 3: non-normal errors

Models of count data won't have normally-distributed errors by definition:

- Generalized Linear Models can accommodate some types of non-normal errors.
  - Especially useful for binary or count data

Data transformations can sometimes fix non-normal errors.



Challenge 3: Non-independent observations/errors

Non-independent observations result in data with lower information content.

- This seems like a strange statement.
- Can we reason out why this might be?

Violations of the assumption of independent, randomized sampling affect our estimates of *significance*.

- Hierarchical structure
- Nearby observations (in space or time)
  - Autocorrelation
- Repeated measurements/time series

Autocorrelation

Does the value of your current observation help you guess what you will observe next?

- Observations nearby in space or time might be more similar than expected due to chance alone.
- Walter Tobler's 1st law of Geography: "Everything is related to everything else, but near things are more related than distant things."

Temporal dependence

Does knowing what happened yesterday give you any info about what might happen today?

- Can we guess the high temperature on July 28th, 2000 if we know the high temperature on July 27th, 2000?
- Can we guess the high temperature on July 28th 2012?

Autoregressive order 1: AR1 - assumes that the current observation is related to the immediately previous observation.

- Not correlated with observations more than 1 time-lag behind.
- Includes a model prediction term for the t - 1 observation

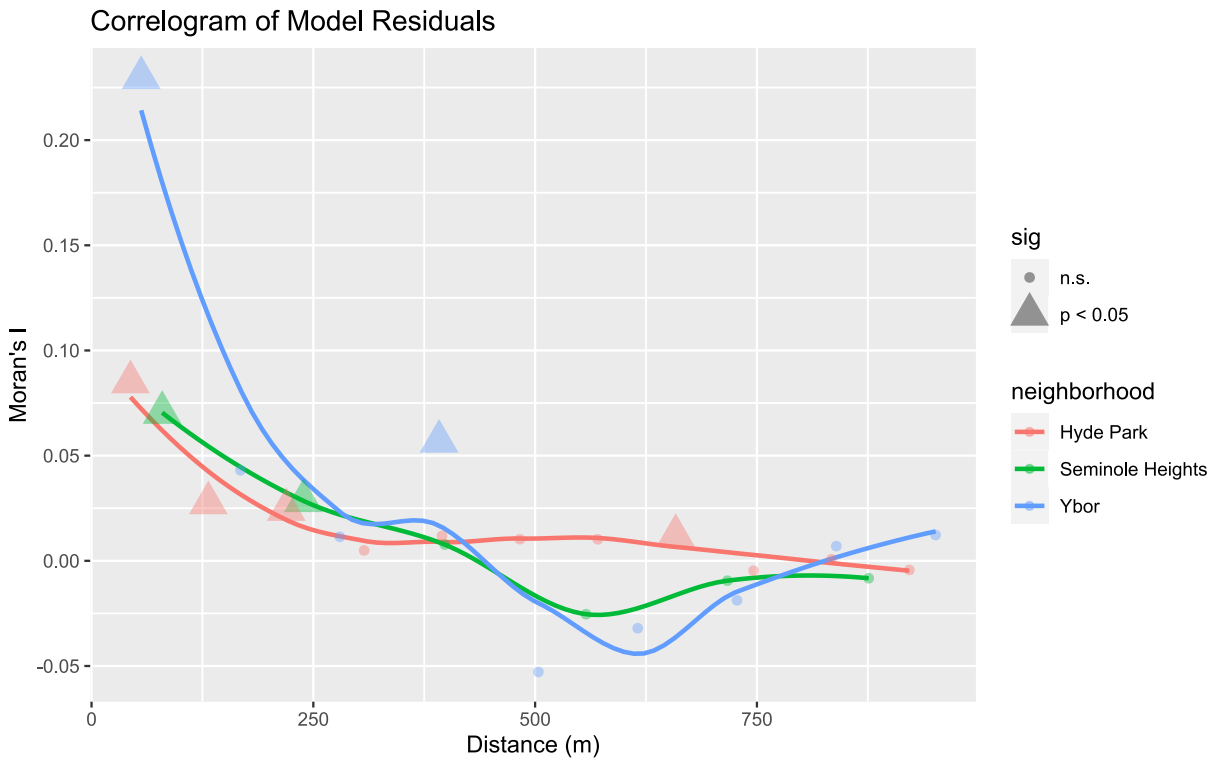
Autoregressive order n: AR(n) - assumes that the current observation is related to the n immediately previous observations.

- AR(n)

Spatial Dependence

Correlation among observations might decrease with increasing distance:

- Nearby observations are more similar than observations separated by large distances.
- Points within a *characteristic scale* are correlated.



Regression with autocorrelated errors

Custom models with custom variance/covariance structures for heterogeneity,

- A difficult (but not impossible) field!
- Zuur 2009 has some good descriptions and examples.

Key Concepts

- What are the Group 1 limitations?
- What are the important Group 1 extensions?

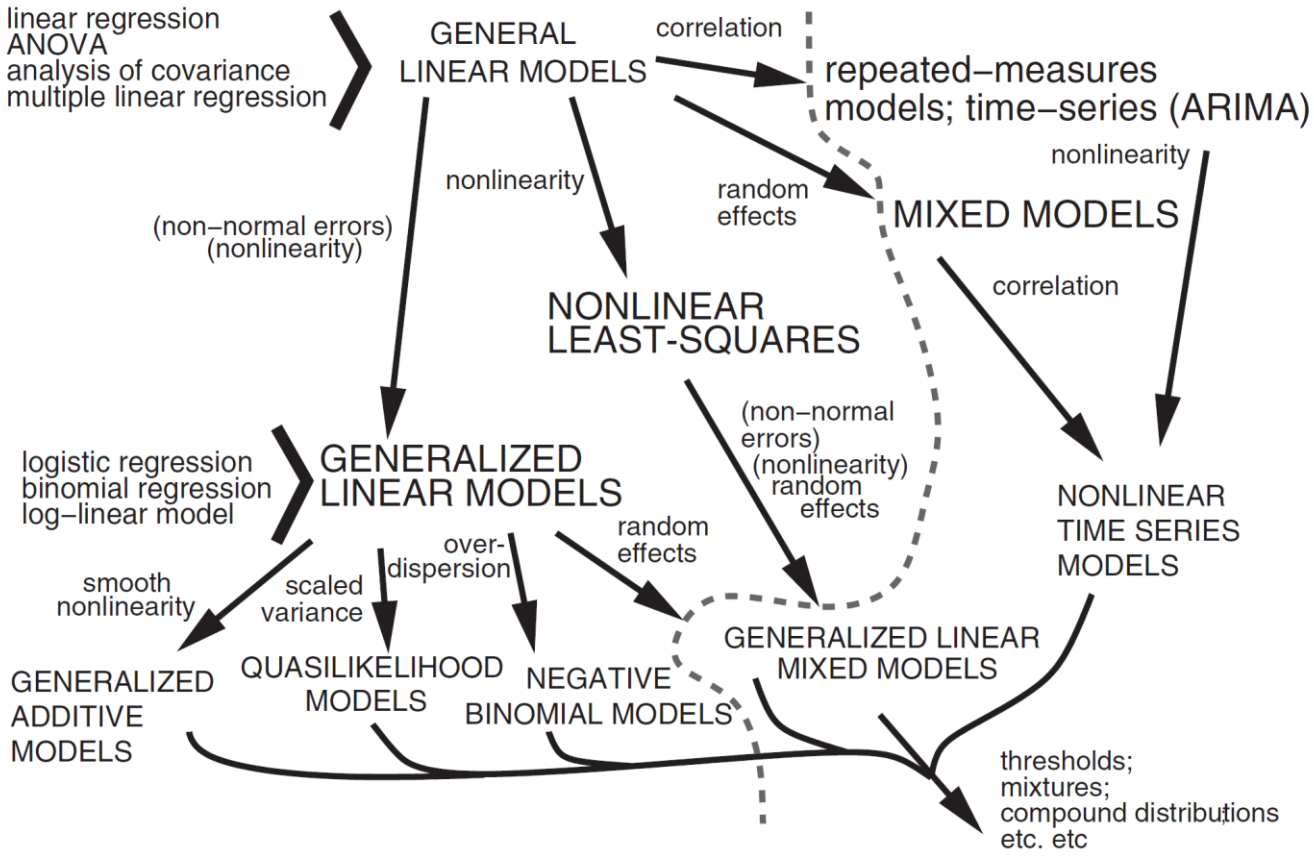
The Constellation of Models

Name Collisions

As with the unfortunate similarity in the names of sample standard error, sample standard deviation, population standard deviation, etc. the regression world has many names and acronyms of that are confusingly similar.

- I have attempted to group methods in a logical way, and use terminology that helps highlight the differences, but there is no way to avoid having to learn the differences between terms and acronyms like general linear models, generalized linear models, GLM, GLMM, GLS, GAM, etc.

The Constellation

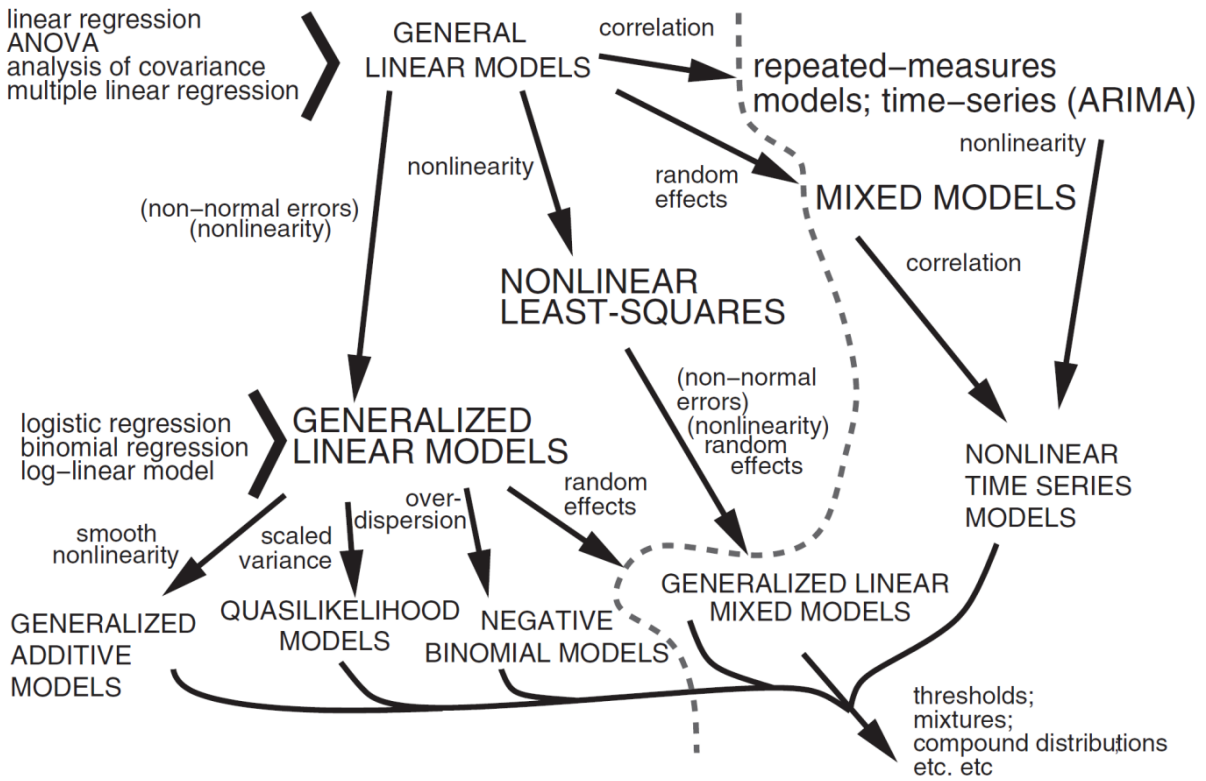


Bolker: Ecological Models and Data in R, Figure 9.2

Constellation of Methods: Groups

I propose a grouping of model types:

- Group 1: Linear methods
- Group 2: Extended linear methods
- Group 3: Random Effects



Groups 1 - 3

How are the groups different?

Some differences among models in the groups:

- Assumptions
- Response data types
- Linearity of parameters
- Constant variance
- Number of response variables
- Stochastic model
- Independence of observations

Four key assumptions

Group I imposes four key assumptions:

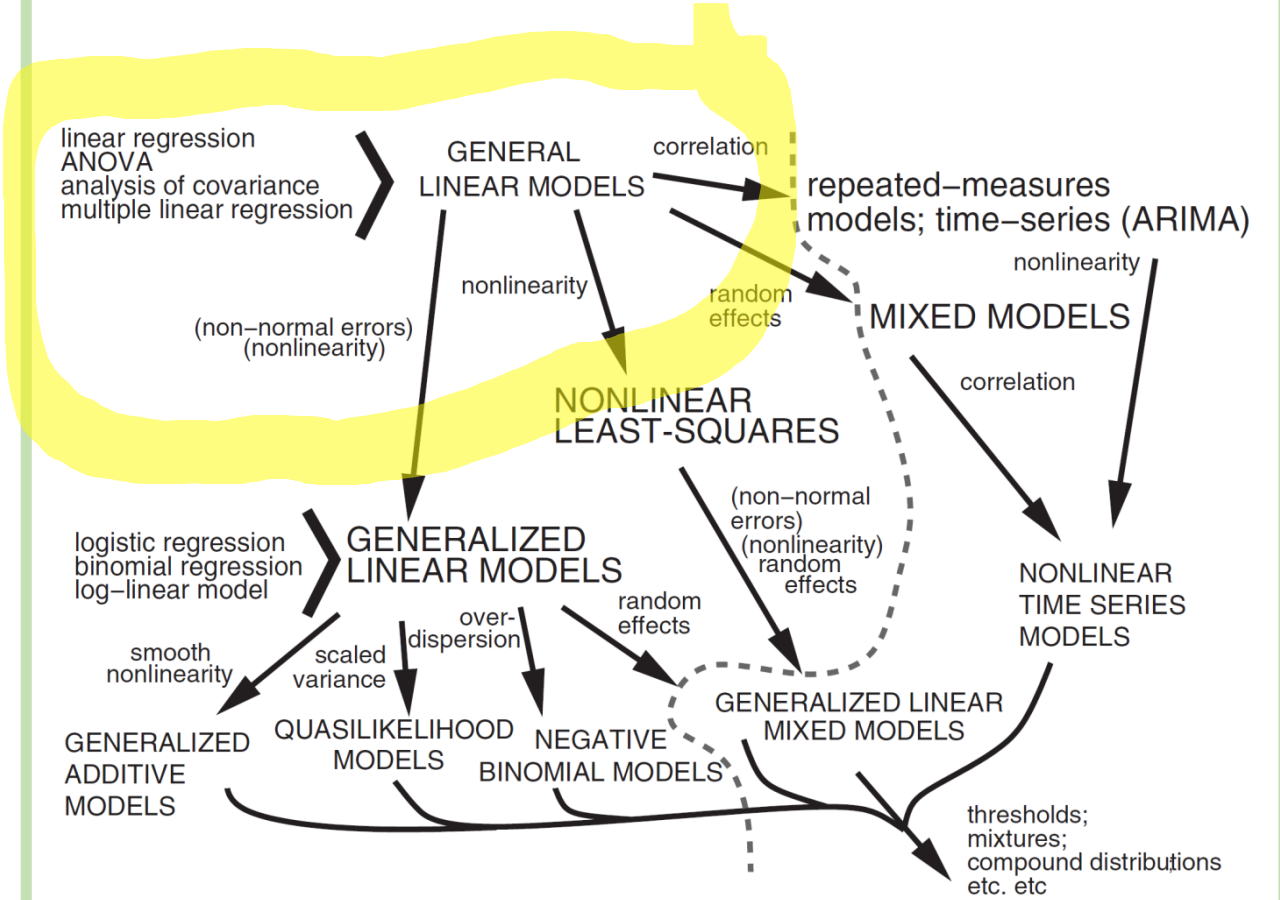
1. Independent observations
2. Constant variance a.k.a homoskedasticity, a.k.a. homogeneity
3. Fixed x: no measurement error in our predictor variables
4. Normality: normality refers to the model residuals

In addition, Group I requires that our models be *linear in the parameters* and have a response on a continuous scale.

The different Group 2 models can deal with different violations of these assumptions and requirements.

Group 1: General Linear Methods

- Single continuous response variable
- One or more predictor variables
  - They may be continuous or categorical
- Deterministic model must be *linear in the parameters*.
- Stochastic model is the Normal distribution.



Group 1: terms and coefficients

- response: Y
  - Also called the dependent variable
- predictor(s): X
  - Also called the independent variable(s)
- intercept(s): α. Sometimes symbolized as β_0
  - This is the expected value of the response when all of the predictors are equal to zero.
- slope(s): β_i: The regression slope is the rate of change in the response variable for each one-unit increase in the predictor variable.

Group 2: Violations of Assumptions

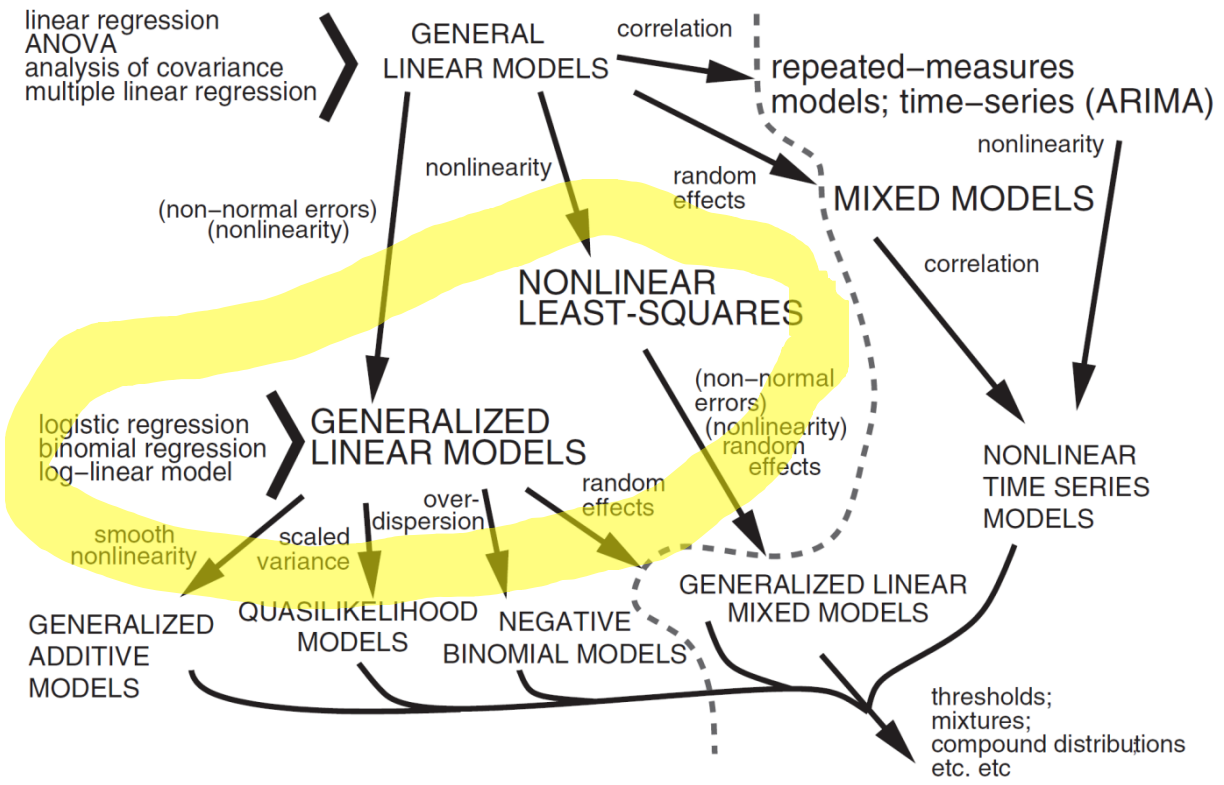
When not meeting assumptions > Group 2
>> Nonlinear least squares or generalised linear models

Recall the key assumptions:

- Independent observations
- Normality (of residuals)
- Constant variance
- Fixed x: no measurement error

Violations of some of these assumptions are more difficult to deal with than others!

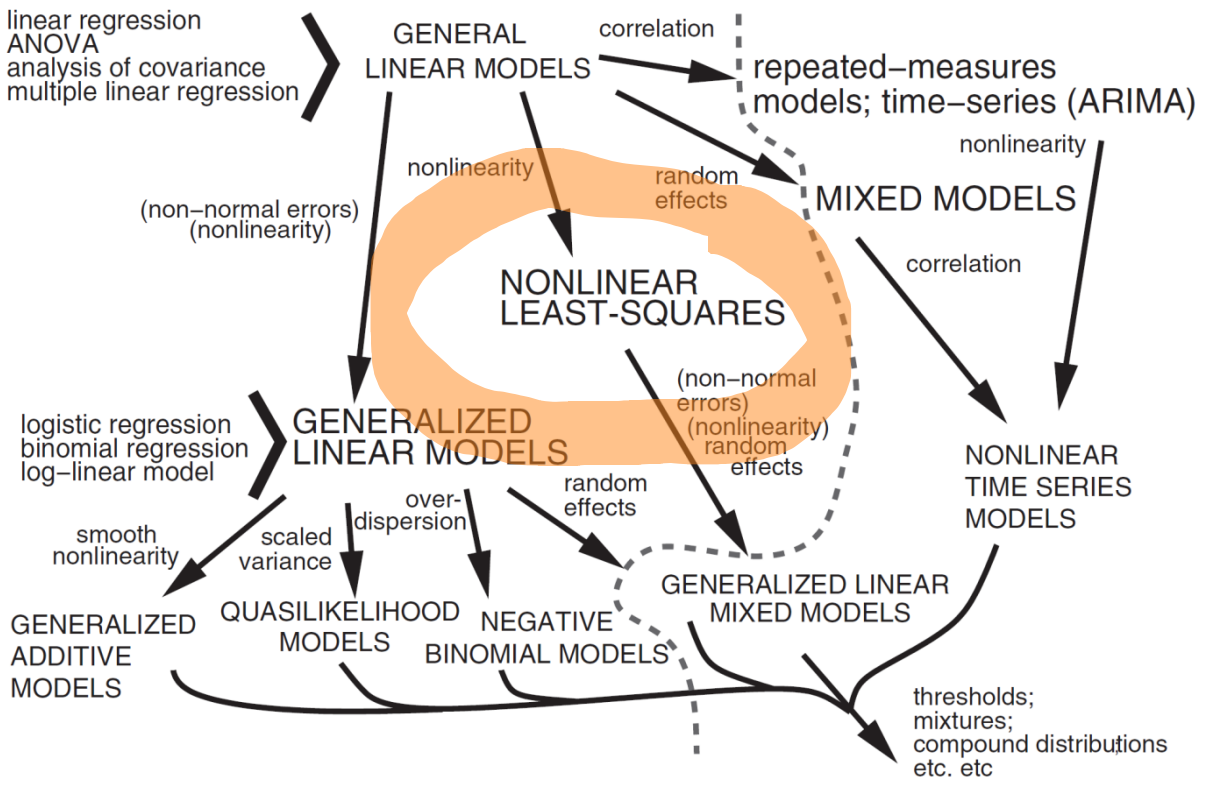
Group 1 methods also required *linearity in the parameters*.



Group 2: Nonlinear Least Squares

NLS requires all of the 4 assumptions, but does not require *linearity in parameters*.

- Useful with nonlinear functions such as Ricker, logistic, any other nonlinear mechanistic function we can propose.
- Uses the least squares optimization criterion
- Find model parameter values that minimize the sum of squared residuals



Nonlinear Least Squares

Challenge: Parameter Estimation

Needs numerical methods to estimate parameters

1. Relies on initial guesses for parameter values
2. Poor guesses can converge to local maxima – these may not be the global 'best' parameters.
3. Uses squared errors (like Group 1 methods)
   - Very sensitive to outliers

Advantage: Flexibility

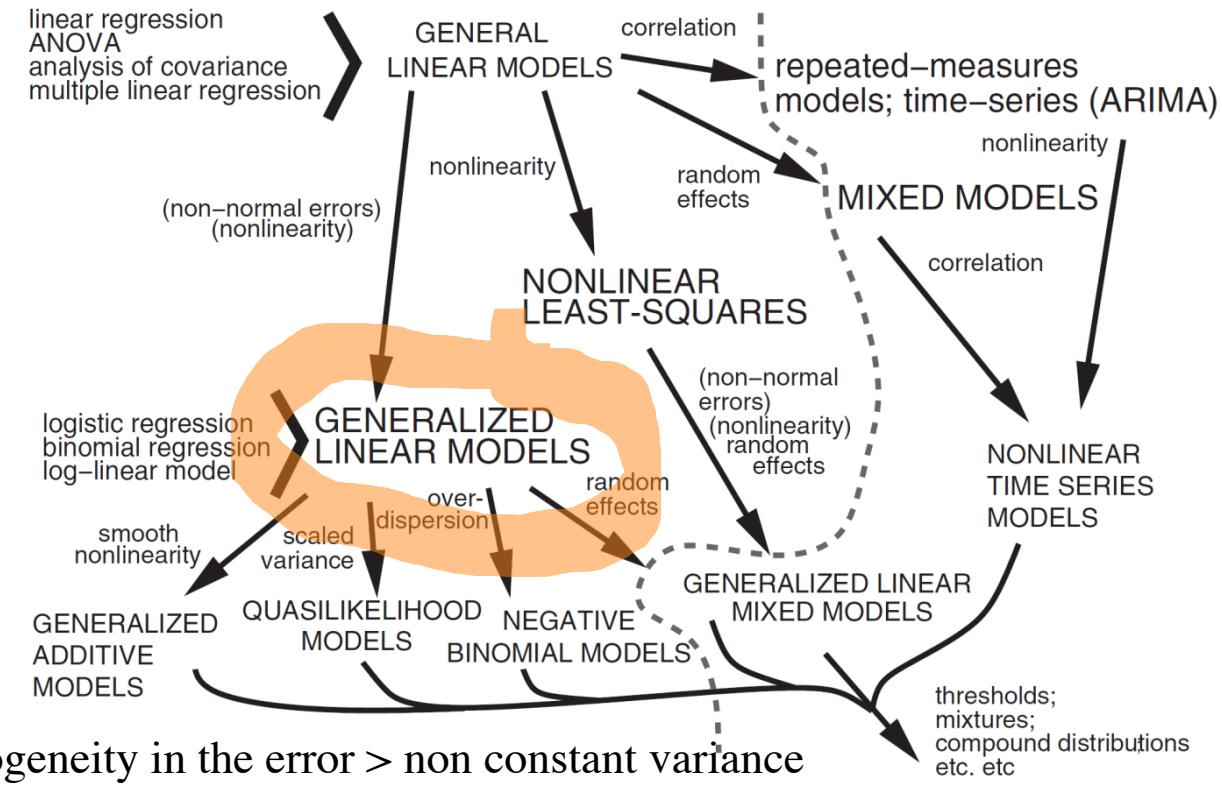
Nonlinear least > doesn't require linear parameters > can have exponent, polynomial function
Still requires assumptions
Do the same
But can't use algebra technique cos not linear
Find model parameter that minimise sum of residuals

- Can fit highly nonlinear functions:
  - Ricker, other mechanistic models
  - Periodic functions

Group 2: *Generalized* Linear Models

The name of this class of models is confusingly similar to *General Linear models*.

- Can sometimes handle heterogeneity in the errors
- Extremely useful for binary and count data: logistic and Poisson regression
- Useful with certain kinds of non-linearity and non-normal errors



Heterogeneity in the error > non constant variance
Normal distribution independent
Allows us to deal with heterogeneity
Deal with categorical or discrete response variables > binary or count ,
> Group one response continuous

Group 2: *Generalized* Linear Models

GLMs *generalize* general linear models by using a *linearizing link function* that can accommodate certain common types of non-normal errors.

- GLMs work with stochastic models that can be specified by a *exponential family* distribution.
  - Many common distributions belong to this family.

GLMs are good with discrete data:

- Counts
- Presence/Absence

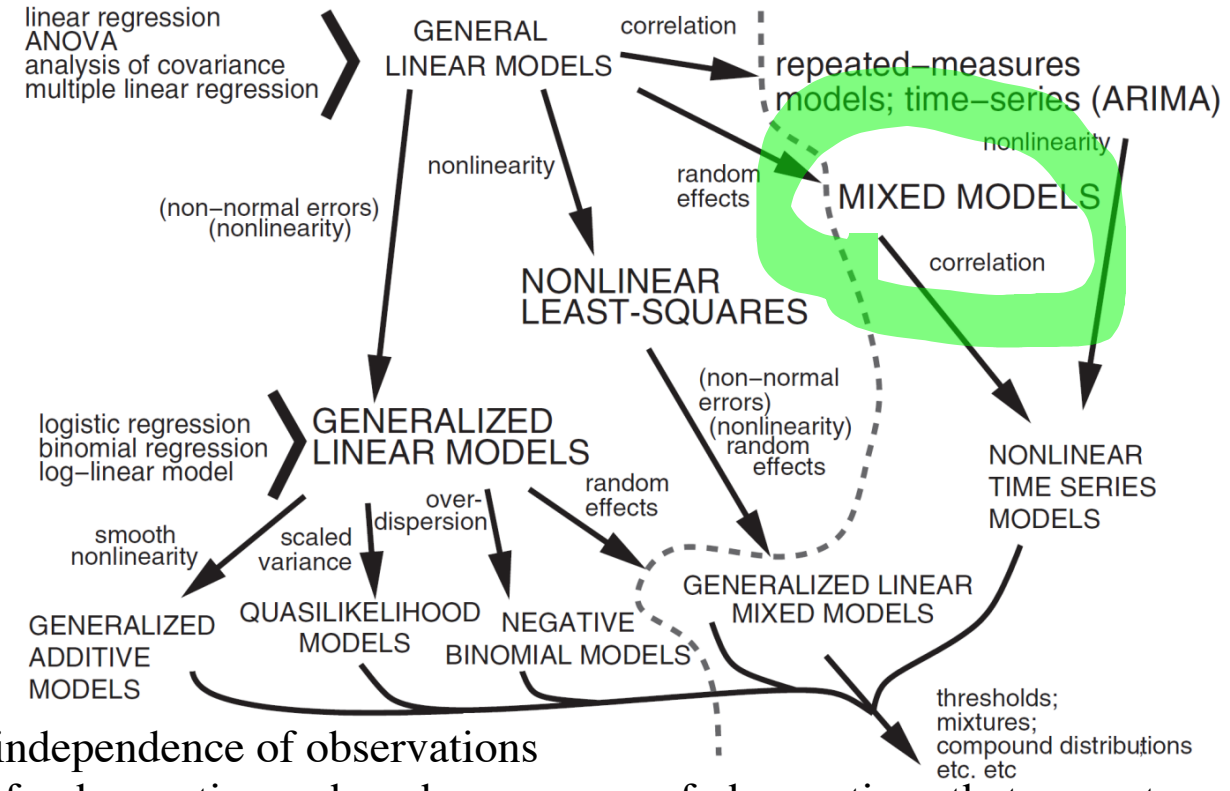
Link Funktion to linear function
We can change error from normal to exponential
Not doing regression to linear function
Can be transformed through link function
GLM counts and presence absence data

Group 3: Mixed Models

What if your experiment has a hierarchical structure?

- For example: You observe 5 locations on each of 3 beaches
  - Sites on the same beach have similar environments
  - You may not care about the effects of the e *specific beaches*
  - The beaches are a *random* collection of all the possible beaches you could have observed.

Mixed models work with fixed and random effects.



Non independence of observations
Specify observations where have groups of observations that are not independent
Individuals within lake are more similar than btw lakes
Mixed effects > adjust df > less info in data >
Hierarchical structure > mixed and random effects

Fixed and Random Effects

Fixed > categorical or numerical predictor > what magnitude of relationship of fixed effects
Glipper lengt
Sex and species as categorical predictors > magnitude of difference btw sex and species in length
> different island > one island not independent > don't care about island effect > can do island as grouping factor
Do i want to do inference on categorical predictor > no > can exclude as random effect
Yes or continuos > fixed effects

Fixed effects can be categorical or numerical.

- We usually want to do inference on fixed effects.

Random effects are always categorical/grouping variables.

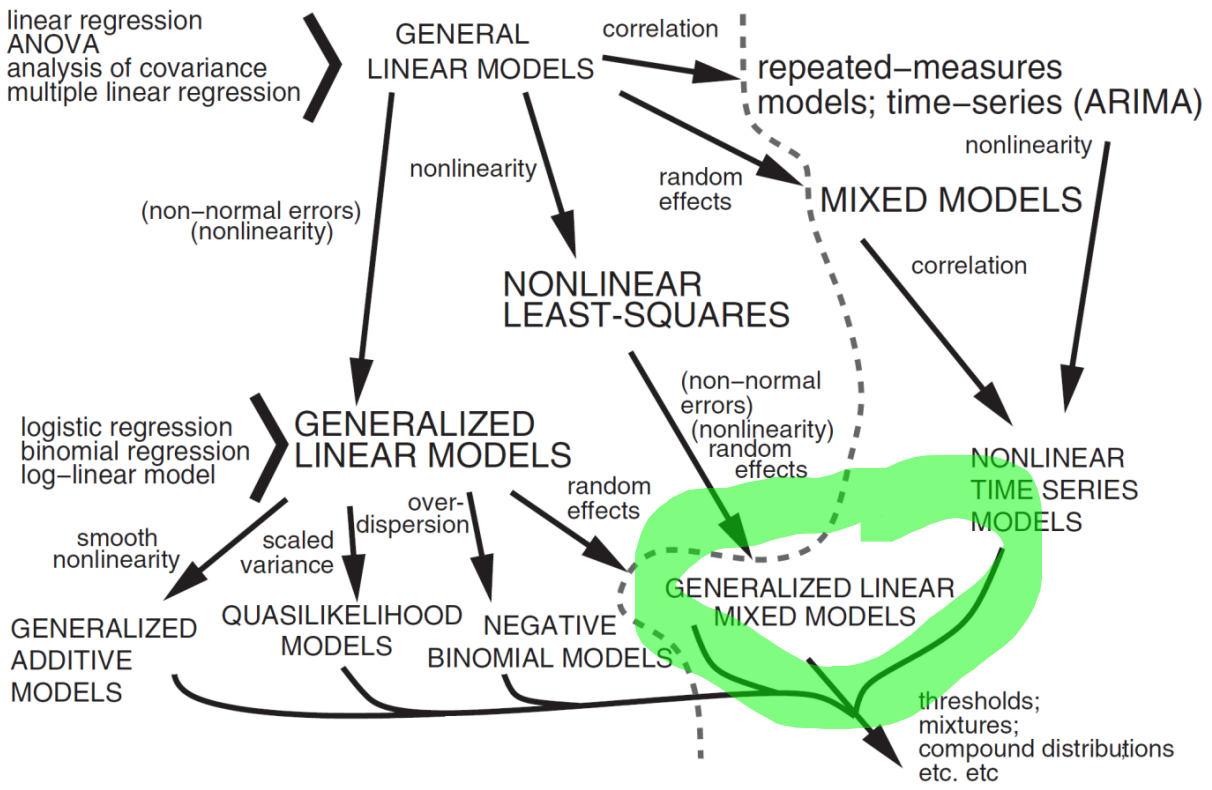
- They are usually not the focus of the experiment:
  - In a greenhouse experiment I care about *specific* water and fertilizer treatments.
  - I probably don't care about the effect of table #3 in greenhouse room #12

Group 3: Generalized Linear Mixed Models GLMM

Heterogeneity in the residuals + categorical responses
Different lakes and blocking structures
Link functions and how gonna model hierarchical structure

These models extend the generalized linear models to work with random effects.

- Censuses nested within beaches
- Presence/absence at sites on mountain ranges



Group 2: [General] Additive [Mixed] Models

usually fit linear function that describes whole data
Sometimes not appropriate
> additive models > not one single global function > Fit local functions

Additive models do not attempt to fit a *global* function.

- Local regressions
- Smoother functions
- Additive models are often considered descriptive or phenomenological.

Mulivariate Models
Instead of having 1 response and set of predictors
Compare multiple respones at one time
Lots of columns > dimensionality reduction
> see which of the variables are contributing most to the viability we observe
Acting together > reduce the dimensionality to two or three dimensions
> are there meaningful groups
> clusters of oberservations that share characteristics
> used for clustering > to see to what extent observations create meaningful groups

Multivariate Models

Multivariate models consider *more than one response* variable.

Common uses:

- Classification, assigning individuals to groups, cluster analysis
- Dimensionality reduction: combining many variables into just a few axes.

Multivariate statistics is a rich and complex field. We won't cover any details in this course.

Machine Learning Techniques: A Fourth Group?

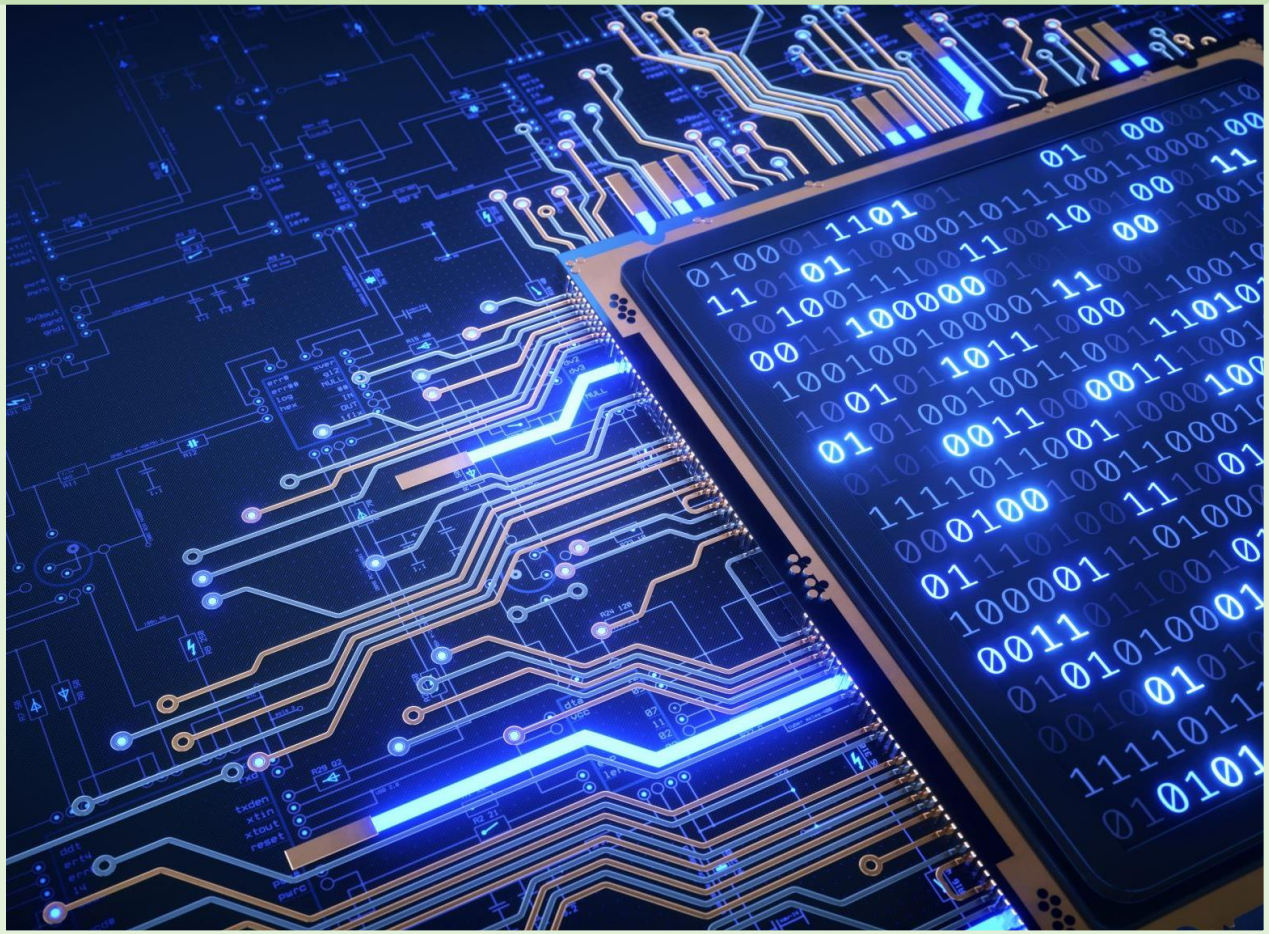
Machine learning
Applying algorithms > what patterns emerge
Classification
Regression
Classification> Categories > which category is it gonna fall in
Regression > categorical and continuous responses >
Random forests

There are two main types of ML: Classification and Regression

Machine learning techniques: 'training' an algorithm or data structure on data.

Machine learning methods include:

- Decision tree methods
- Support Vector Machines (SVM)
- Neural networks



The Constellation

Groups of model types

1. General Linear Models
2. Generalized + NLS
3. Mixed Models

Violations of Assumptions/Requirements

1. Linearity: LNS, GLM
2. Normality of errors: GLM
3. Independence: Mixed Models, time series