evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Tento projekt je spolufinancován Evropským sociálním fondem a Státním rozpočtem ČR
InoBio – CZ.1.07/2.2.00/28.0018

Statistical Analysis in Ecology using R

Linear Models/GLM

Ing. Daniel Volařík, Ph.D.

13. 11. 2013

european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitiveness
INVESTMENTS IN EDUCATION DEVELOPMENT

Recap

- Basics in R, exploring data (mean, median, correlation , …)
- Graphs
- Linear models
  - ANOVA
  - Linear regression
  - Ancova

Limitation of linear models

Assumptions of linear models:

- Normality – observation are normaly distributed for each value of X
- Homogeneity – constant variance
- Independence – value of one observation doesn‘t influence value of other, violated intime series data and spatial data.

Limitation of linear models

What to do when violating the assumptions?

- Normality – we can use GLM
- Homogeneity – GLM, GLS
- Independence – LME (nested data, random effect), GLS (spatial, temporal autocorrelations)
- Normality + independence – GLMM

Generalized linear models (GLM)

- Generalisation of linear models that allows other than normal distribution of errors.
- For distributions from exponential family – binomial, Poisson, gamma, negative binomial, quasi-binomial, quasi-poisson
- Concept proposed by Nelder & Wedderburn (1972)

- Other possibilities how to deal with non-normal distribution of errors:
  - Transformations (square root, log)

GLM components

- Linear predictor.
- Error distribution
- Link function

Linear predictor

- The same as in linear models
- Combination of explanatory variables
- e.g. a + bx; a + bx + cx
- interactions and quadratic terms could be included like in linear models

Error distributions

- (or distribution for error terms, distribution of the response variable)
- To model variance
- Normal, binomial, Poisson, gamma, negative binomial

Normal distribution

- Continuous two-parameter distribution:
- the mean, μ (mu), and the
- standard deviation, σ (sigma).
- Bell shaped probability distribution
- variance and mean are independent
- Could have both negative and positive values

$$X \sim N\left(\mu, \sigma^2\right)$$

Gamma distribution

- For continuous response variable with strictly positive values ($Y > 0$)
- Two parameter – mean μ and ν
- Variance is defined as μ^2/ν; ν^{-1} denotes to dispersion
- Depending on μ and ν, it could have various shapes

Poisson distribution

- Response variable has to have only integer values
- One parameter, the mean number of succeses, μ (mu). (μ can be non-integer)
- Variance is equal to mean
- Typically used for count data
- Allows heterogeneity
- in ecology it is quite often that the variance is even larger than mean – overdispersion – quasi-Poisson

Binomial distribution

- A sequence of independent Bernoulli trials (like tossing a coin).
- A two-parameter distribution: the number of trials, N, and the probability of success, p, in any given trial.
- Mean is given by $N \times p$
- Variance by $N \times p \times (1 - p)$
- Assumption: probability of success does not change from trial to trial.
- In ecology it can be presence/absence of mistletoe on the oak tree, presence/absence of some species on particular sites.

Link function

- Link between the mean of response variable and the systematic part (GLM alternative to data transformation).

Distribution	Default link	Alternative link
Normal	identity	log, inverse
Binomial	logit	probit, cauchit, log, complementary log-log
Poisson	log	identity, square root
Gamma	reciprocal	log, identity, square root

Maximum likelihood

- How to estimate parameters?
- In linear models – ordinary least squares
- GLM – maximum likelihood estimation
  - iterative approximation using method iteratively reweighted least squares

Analogies between Normal Least Squares and generalized Linear Models

Normal Least Squares	Generalized Linear Models
Sums of squares (SS)	Deviance
Normal least squares	Maximum likelihood iterative (re-)weighted least squares
Normal error distribution	Exponential family of distributions
Transform data	Link function
Variance ratio F-tests	Chi-squared tests of deviance (F tests of mean deviance ratios)

How to fit GLM in R

- Function glm() with parameters:
  - Formula (for linear predictor – the same as in lm()
  - Family – try `?family` for possible families
  - Link function specified by family(link ="possible_link")
- Example:

```
glm(y ~ x + z, data = data,
 family = Gamma(link = "inverse"))
```

GLM in R

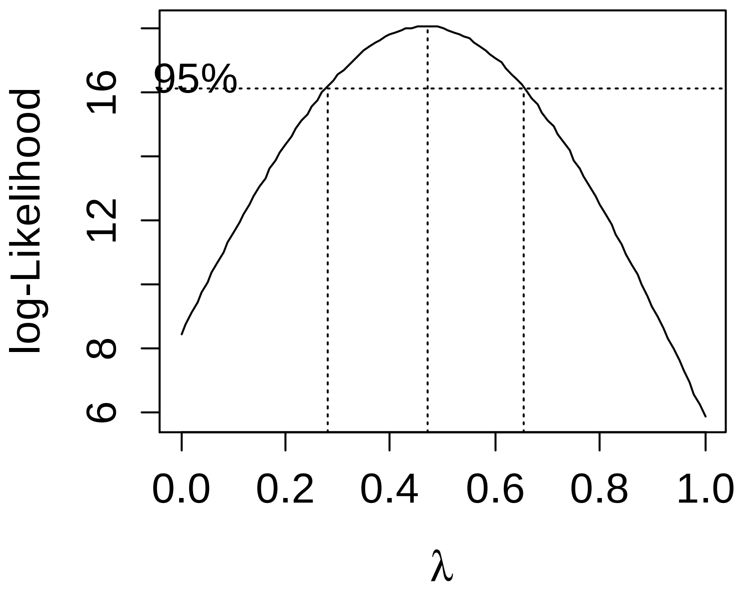
```
Call: glm(formula = Hardness ~ Density, family = Gamma(link =
"inverse"), data = Janka)

Deviance Residuals:
Min 1Q Median 3Q Max -0.55341 -0.18060 0.05147 0.14486 0.38080
Coefficients:
             Estimate     Std. Error    t value Pr(>|t|)
(Intercept)  1.998e-03    1.131e-04     17.67  < 2e-16 ***
Density      -2.498e-05   1.859e-06     -13.44 3.7e-15 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for Gamma family taken to be 0.05049825)

Null deviance: 11.268 on 35 degrees of freedom Residual
deviance: 1.902 on 34 degrees of freedom AIC: 516.42 Number of
Fisher Scoring iterations: 4
```

Transformations in R

```
> lm(log(Y) ~ X, data = data)
```

- To find the best transformation, we can use boxcox() function in MASS package
  - The Box-Cox family of transformations
  - Tries a whole series of transformations and indicates which one performs best

Transformations in R – boxcox function



- Lambda indicate the type of tranformation
- Lambda = 1, the data is untransformed
- Lambda = 2, the data is squared
- Lambda = 0.5, square-root transformed and so on.
- Lambda = 0, defined as the natural log transformation
- Doesn't work with zeros in the data

Exercise on Janka dataset

- Janka Timber Hardness Data
- From Williams (1959) via Venables (2000)
- Dataset is in the file DataRegressionTimberVenables.txt
- Just 2 variables – timber density and timber hardness
- We are interested in how timber hardness depends on timber density