Generalized Linear Models

Matt Farr

Models in Science

- Science is too complex to understand → must reduce complexity

- Describe and predict systems/processes

- Model is set of assumptions about scientific systems/processes

- Simplification of reality

Mathematical Models

- Description of a system or process composed of variables, typically written in algebra

- Algebra: explicit and great transparency

- Simple example is the equation for a line: $y = mx + b$

- Format for linear relationships, where the expected response (y) can be treated as the result of explanatory variables (x) whose effects (mx + b) are additive.

Statistical Models

- Stochasticity or randomness
- Statistical models: a description of a system/process composed of variables but where one or more **random variables** are related to other variables
- Explicitly acknowledge stochasticity in system/processes

$$\text{Response} = \text{deterministic part} + \text{stochastic part}$$

- Parametric statistical models (generalized linear models): a description of a system/process using probability distributions thought to have produced the data

GENERALIZED LINEAR MODEL

Response Data = Deterministic Part + Stochastic Part

$y=\alpha+\beta\cdot x+\epsilon$

$\epsilon\sim Normal(0,\sigma^2)$

Response Data ~ Distribution(Parameters)

$y \sim Normal(\mu,\sigma^2)$

Link(Mean Response) = Linear Model

$\mathrm{l}(\mu)=\alpha+\beta\cdot x$

$\mu=\alpha+\beta\cdot x$

DATA TYPES

- Types of responses (y)
  - Continuous (body mass)
  - Binary (heads/tails; dead/survived)
  - Categorical (rolling a die, nationality; geographic location)
  - Counts (number of birds in a quadrant)

Frequent Distributions

- Continuous (body mass)
  - Random part: $y \sim Normal(\mu, \sigma^2)$
  - Deterministic part: $\mathrm{I}(\mu) = \alpha + \beta \cdot x$
- Binary: Binomial distribution (coin flip; sex; alive/dead)
  - Random part: $y \sim Binomial(p, N)$
  - Deterministic part: $logit(p) = \alpha + \beta \cdot x$
- Categorical (rolling a die; nationality; age)
  - Random part: $y \sim Multinomial(\boldsymbol{\pi}, N)$
  - Deterministic part: $logit(\boldsymbol{\pi}) = \alpha + \beta \cdot x$
- Counts (number of birds in a quadrant)
  - Random part: $y \sim Poisson(\lambda)$
  - Deterministic part: $log(\lambda) = \alpha + \beta \cdot x$

Design Matrix

- Deterministic part of the linear model
- A matrix of explanatory variables
- Manipulation of the design matrix that leads to classical statistical models
- Column for each predictor variable
- Means parameterization or an effects parameterization (intercept)
- Row for each data point (individuals, sites, time period, etc.)

Design Matrix

$$y_i \sim Normal(\mu_i, \sigma^2)$$

$$I(\mu_i) = \beta_F \cdot females_i + \beta_M \cdot males_i$$

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} * \begin{pmatrix} \beta_F \\ \beta_M \end{pmatrix}$$

Design Matrix

$$y_i \sim Normal(\mu_i, \sigma^2)$$

$$\mathrm{I}(\mu_i) = \beta_F \cdot females_i + \beta_M \cdot males_i$$

$$\mu_1 = \beta_F \cdot 1 + \beta_M \cdot 0$$

$$\mu_2 = \beta_F \cdot 1 + \beta_M \cdot 0$$

$$\mu_3 = \beta_F \cdot 1 + \beta_M \cdot 0$$

$$\mu_4 = \beta_F \cdot 0 + \beta_M \cdot 1$$

$$\mu_5 = \beta_F \cdot 0 + \beta_M \cdot 1$$

$$\mu_6 = \beta_F \cdot 0 + \beta_M \cdot 1$$

Design Matrix

$$y_i \sim Normal(\mu_i, \sigma^2)$$

$$I(\mu_i) = \beta_F \cdot females_i + \beta_M \cdot males_i$$

$$\mu_1 = \beta_F$$

$$\mu_2 = \beta_F$$

$$\mu_3 = \beta_F$$

$$\mu_4 = \beta_M$$

$$\mu_5 = \beta_M$$

$$\mu_6 = \beta_M$$

$\beta_F = mean\ of\ females$

$\beta_M = mean\ of\ males$

Design Matrix

$$y_i = \beta_F \cdot females_i + \beta_M \cdot males_i + \epsilon_i$$

$$\epsilon_i \sim Normal(0, \sigma^2)$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} * \begin{pmatrix} \beta_F \\ \beta_M \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{pmatrix}$$

Model of the Mean

$$y_i \sim Normal(\mu, \sigma^2)$$

$$y_i = \mu + \epsilon_i$$

$$\epsilon_i \sim Normal(0, \sigma^2)$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} * (\mu) + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{pmatrix}$$

Model Analysis

ZIPKIN QUANTITATIVE ECOLOGY LAB

- Parameters are unknown!!
- Frequentist analysis
  - Least-squares (LM function in R)
  - Maximum likelihood
  - Method of moments
- Bayesian analysis

Modeling Steps

- Data Exploration
- Response explanation
- Model structure & development
- Data formatting
- Model analysis
- Model results & interpretation

Exercise:

https://raw.githubusercontent.com/farrmt/Rworkshop/master/GLMfun.csv

References

- Kéry, Marc. 2010. Introduction to WinBUGS for ecologists: A Bayesian approach to regression, ANOVA, mixed models, and related analyses. Academic Press, Boston.