NRES_798_11_201501

Generalized linear model

Logistic regression

The General Linear Model

In a general linear model

$$y_i = B_0 + \ B_1 x_i + \cdots + B_p X_p + \ \varepsilon_i$$

the response y_i is modelled by a linear function of **explanatory variables** x_i, plus an error term

General and Linear Model

Here **general** refers to the dependence on potentially more than one explanatory variables, v.s. the **simple linear model**:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

The model is linear in the coefficients,

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon_i$$
$$y_i = \beta_0 + \gamma_1 \delta_1 x_1 + \exp(\beta_2) x_2 + \epsilon_i$$

but not

$$y_i = \beta_0 + \beta_1 x_1^{\beta_2} + \epsilon_i$$
$$y_i = \beta_0 \exp(\beta_1 x_1) + \epsilon_i$$

Error structure

We assume that the errors ϵ_i are independent and identically distributed such that

$$E[\epsilon_i] = 0$$
$$\text{and } \operatorname{var}[\epsilon_i] = \sigma^2$$

Typically we assume

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

as a basis for inference,

Restrictions of Linear Models

Although a useful framework, there are some situations where general linear models are not appropriate

- the range of Y is restricted (e.g. binary, count)
- the variance of Y depends on the mean

Generalized linear models extend the general linear model framework to address both of these issues

GLM potential response variables

- Count data expressed as proportions
  - E.g. proportion male
- Count data that are not proportions
  - E.g. bounded population data (negative values meaningless)
- Binary response variables
  - e.g. present or absent, dead or alive
- Data on time to death where the variance increases faster than linearly with the mean

Generalized Linear Models (GLMs)

A **generalized linear model** is made up of three things:

- a **linear predictor**

$$n_i = B_0 + \ B_1 x_i + \cdots + B_p X_p$$

and two functions

- a **link** function that describes how the mean, E(Yi) = ui depends on the linear predictor

$$g(\mu_i) = n_i$$

- An **variance** function that describes how the variance, var(Yi) depends on the mean

$$g(Y_i) = \emptyset V(\mu)$$

where the **dispersion parameter** ϕ is a constant

Error structure

Normal General Linear Model as a Special Case

For the general linear model with $\epsilon \sim N(0, \sigma^2)$ we have the linear predictor

$$n_i = B_0 + B_1 x_i + \cdots + B_p X_p$$

the link function

$$g(\mu_i) = \mu_i$$

And the variance function

$$\mathrm{V}(\mu_i) = 1$$

Error structure

Error structure

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

- When standard normal error fails

- Errors are strongly skewed
- Errors are kurtotic
- Errors are strictly bounded
  - e.g. proportions (0, 1)
- Error that can’t lead to negative fitted values
  - e.g. counts

Possible GLM error distributions

- Poisson error: count data
- Binomial error: proportions data
- Gamma errors: data with a constant coefficient of variation
- Exponential errors: data on time of death (survival analysis)

Transformations vs. GLM

In some situations a response variable can be transformed to improve linearity and homogeneity of variance so that a general linear model can be applied.

This has some drawbacks

- response variable has changed!
- transformation must simultaneously improve linearity and homogeneity of variance
- transformation may not be defined on the boundaries of the sample space

The `glm` Function

Generalized linear models can be fitted in R using the `glm` function, which is similar to the `lm` function for fitting linear models.

The arguments to a glm call are as follows

```
glm(formula, family = gaussian, data, weights, subset,
    na.action, start = NULL, etastart, mustart, offset,
    control = glm.control(...), model = TRUE,
    method = "glm.fit", x = FALSE, y = TRUE,
    contrasts = NULL, ...)
```

Formula Argument

The formula is specified to glm as, e.g.

```
y ~ x1 + x2
```

where `x1, x2` are the names of

- numeric vectors (continuous variables)
- factors (categorical variables)

Other symbols that can be used in the formula include

- `a:b` for an interaction between `a` and `b`
- `a*b` which expands to `a + b + a:b`
- `.` for first order terms of all variables in `data`
- `-` to exclude a term or terms
- `1` to include an intercept (included by default)
- `0` to exclude an intercept

Family Argument

The `family` argument takes (the name of) a family function which specifies

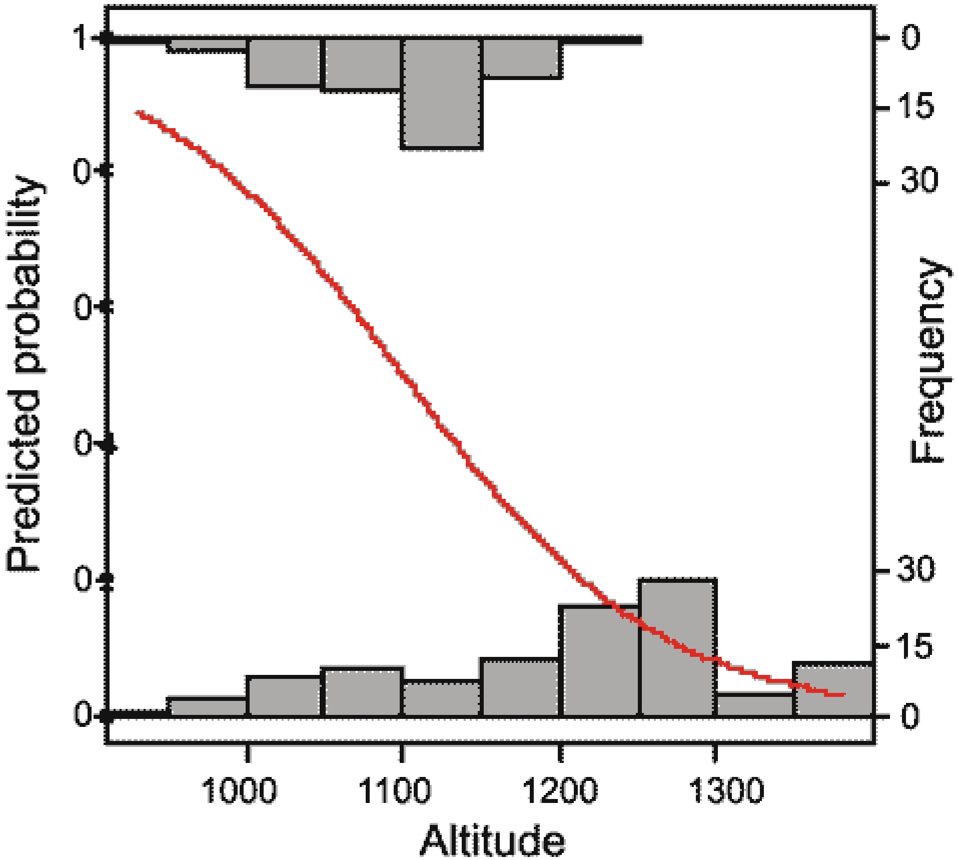
- the link function
- the variance function
- various related objects used by `glm`, e.g. `linkinv`

The exponential family functions available in R are

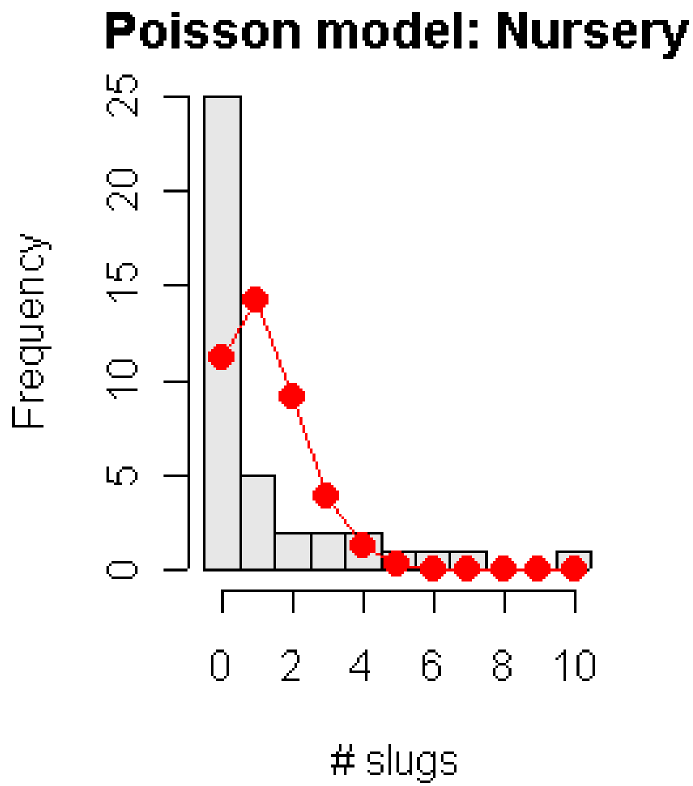
- `binomial(link = "logit")`
- `gaussian(link = "identity")`
- `Gamma(link = "inverse")`
- `inverse.gaussian(link = "1/mu^2")`
- `poisson(link = "log")`

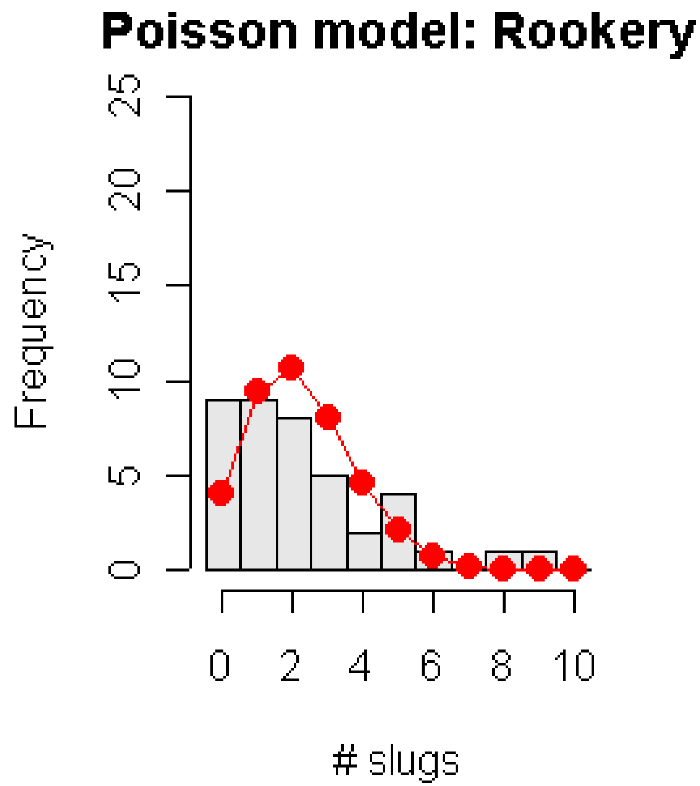
GLM: Logistic regression, binomial family

Probability of Norway spruce occurrence along an altitudinal gradient

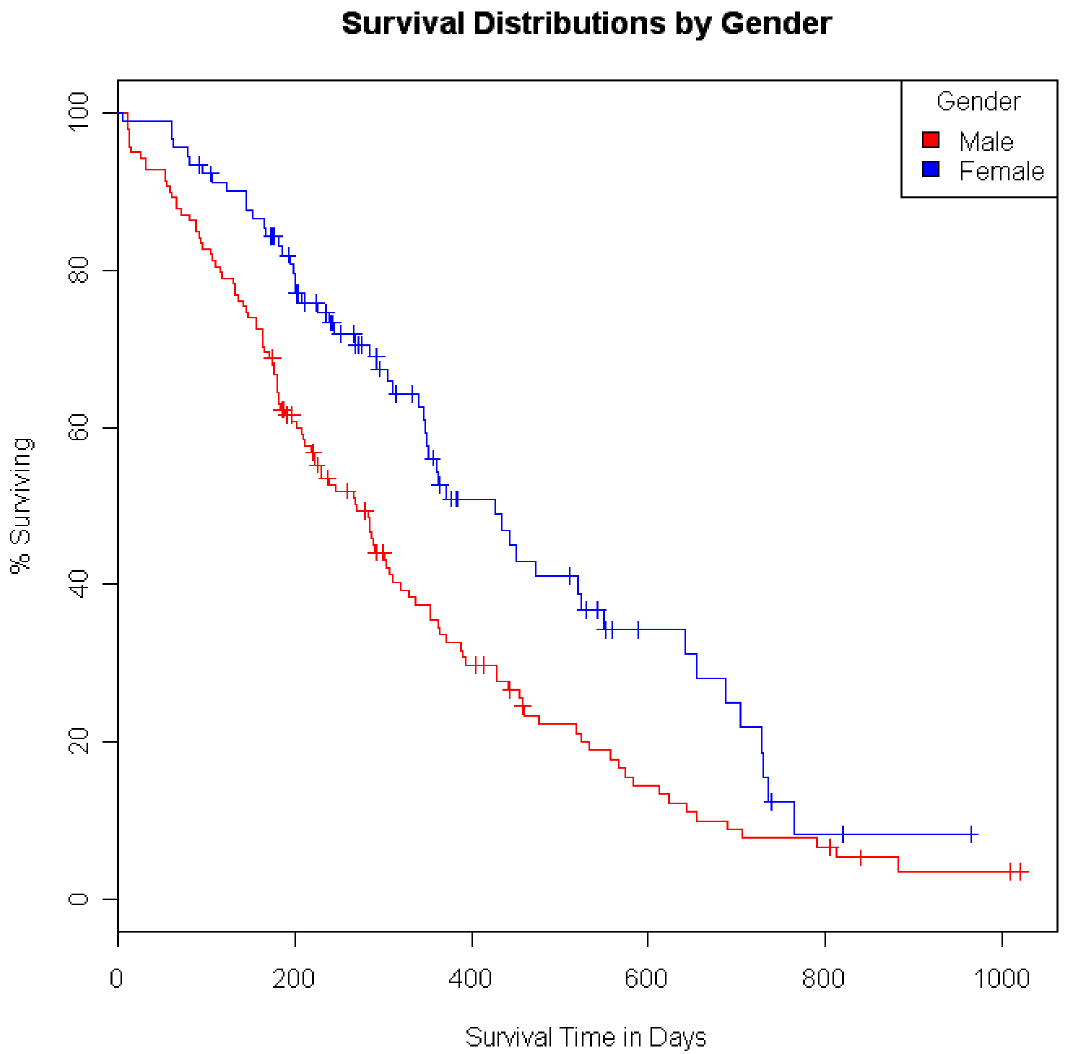


GLM: Poisson regression, Poisson family





GLM: Survival analysis



Exponential distribution
Weibull distribution
Gamma distribution

Overview

- **Logistic Regression**
- Logistic regression is useful when you are predicting a binary outcome from a set of continuous predictor variables.

- **Poisson Regression**
- Poisson regression is useful when predicting an outcome variable representing counts from a set of continuous predictor variables.

- **Survival Analysis**
- Survival analysis (also called event history analysis or reliability analysis) covers a set of techniques for modeling the time to an event.
- Data may be **right censored** - the event may not have occurred by the end of the study or we may have incomplete information on an observation but know that up to a certain time the event had not occurred (e.g. the participant dropped out of study in week 10 but was alive at that time).