Generalised Linear Models 2: Modelling Binary and Proportional Data

Michael Noonan

Biol 520C: Statistical modelling for biological data

Table of contents

1. Generalised Linear Models Review
2. Logistic Regression
3. Logistic Regression on Presence-Absence Data
4. Logistic Regression on Proportion Data
5. Where to from here?

Generalised Linear Models Review

GLM review

Last lecture we saw how GLMs offer a powerful framework for modelling data types that can not be assumed to have Gaussian distributed errors.

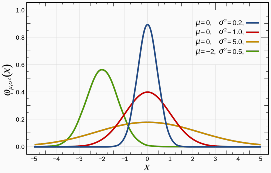
We then saw how, when fitting GLMs, we need to carry out 3 steps:

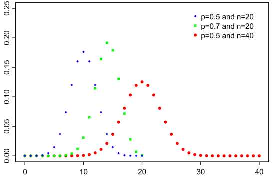
1. Make a distributional assumption on the response variable Y_i. This also defines the mean and variance of Y_i.
2. Specify the deterministic part of the model.
3. Formally specify the ‘link’ between the mean of Y_i and the deterministic part based on your distributional assumption.

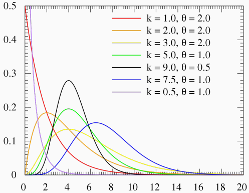
We then saw how to fit GLMs to count data in `R` using the `glm()` function.

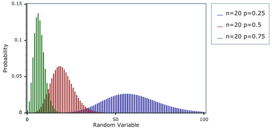
GLM review cont.

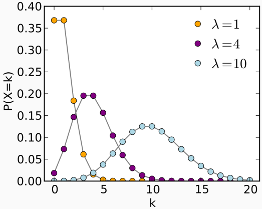
Because R functions streamline the process of fitting GLMs, the key step that's left in your hands is knowing when you will need to switch from a Gaussian model to a GLM, and identifying the correct distribution

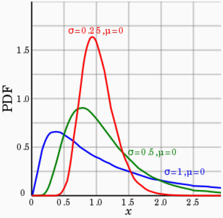












Today's lecture

Today we are going to learn how to extend what we covered last lecture about modelling count data to two other common data types in biology:

1. 0-1 data (presence-absence, infected or not, alive vs. dead, etc...).
2. Proportional percentage data (prop. of pop. infected, % forest cover, prop. of pop. with a mutation, etc...).

Applying GLMs to these data is also commonly referred to as 'logistic regression'.

The three step process we covered last lecture is identical for these data types, we just need to familiarise ourselves with a new set of distributions, and a new link function.

Logistic Regression

Logistic regression overview

Logistic regression is a method for fitting a regression curve, $y = f(x)$, when y consists of proportions, probabilities, or binary coded (0,1–failure,success) data (i.e., anything bound between 0 and 1).

The term 'logistic regression' comes from the fact that the link function we use fits a logistic curve to the relationship between x and y.

Assumptions:

1. The true conditional probabilities are a logistic function of the independent variables (i.e., correct model specification).
2. No important variables are omitted & no extra one included.
3. The independent variables are measured without error.
4. The observations are independent.
5. There is no collinearity in the independent variables.

Logistic regression overview cont.

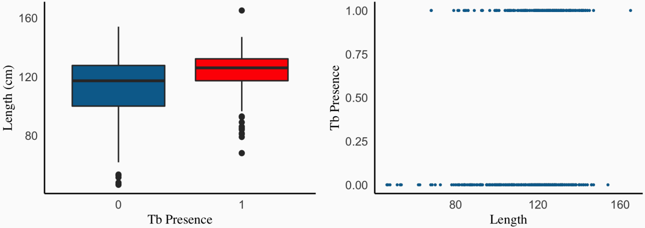
If you are going to be working with 0-1 data a lot in your career, you might consider reading a book focused entirely on this subject. A good place to start would be:

Agresti, A. (2018). An introduction to categorical data analysis. John Wiley & Sons.

Logistic Regression on Presence-Absence Data

Tuberculosis in Wild Boar

We're going to work with a dataset collected by Vicente et al. (2006). They analyzed the distribution on tuberculosis-like lesions in wild boar (*Sus scrofa*) for potential importance of persistence of tuberculosis in south central Spain.

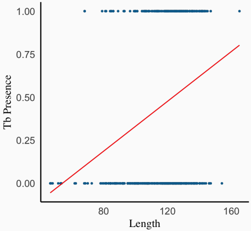


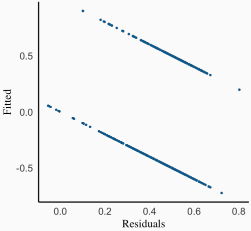
With these data we're interested in knowing whether body size is related to Tb prevalence.

Linear regression on 0-1 data

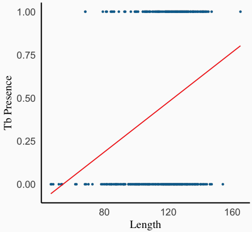
We already know a Gaussian model isn't a great choice, but let's see what that would look like.

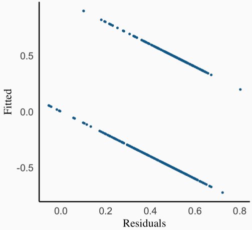
```
library(nlme)

Fit_Linear <- gls(Tb ~ Length,
                  data = data,
                  method = "ML")

summary(Fit_Linear)

Generalized least squares fit by maximum likelihood
  Model: Tb ~ Length
  Data: data
       AIC      BIC    logLik
  680.8635 693.4711 -337.4317

Coefficients:
                 Value  Std.Error   t-value p-value
(Intercept) -0.3912856 0.13589236 -2.879379  0.0042
Length       0.0072337 0.00114342  6.326359  0.0000
```





Linear regression on 0-1 data





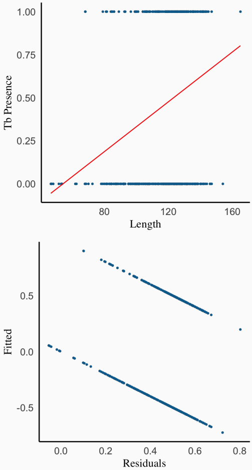
We found a significant relationship between body length and the presence/absence of Tb, but the residuals look terrible (no amount of transformations will help us here).

Our fitted model is:
$\mathrm{Tb}_i = -0.39 + 0.007 \times \mathrm{Length}_i + \varepsilon_i$

Which predicts that an animal that is 120cm long will have ∼0.45 Tb.

Can a boar have 0.45 Tb?

Linear regression on 0-1 data



Animals are either infected or not and there is no in between, so we need to redefine what our fitted values mean.

Instead we re-define our response as π_i, where π_i represents the probability of have Tb, so a 120cm long boar will have a $\sim$0.45 chance of having Tb.

But our model also says that a 50cm boar has a -0.04 chance of having Tb...

We need a deterministic function that maps the values between 0 and 1, and a dist. that makes more sense.

Step 1: Distributional assumption

The first step is to make a distributional assumption on our Tb prevalence data.

Can you think of a good candidate for 0,1 data?

Binomial distribution

The binomial distribution describes the probability of obtaining k yes/no successes in a sample of size n, or in other words, the distribution of the number of successful trials among a defined number of trials.

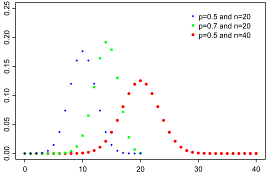
Parameters: n and p

Type: Discrete

Biological scenarios: Mark recapture data, live vs dead survival data, killed by a predator or not, yes/no behavioural outcomes, anything with a discrete yes/no outcome.

PMF: $\binom{n}{k} p^k (1-p)^{n-k}$
where
$\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Range: discrete $(0 \leq x \leq n)$
Mean: np
Variance: $np(1-p)$



Source: Wikipedia

Step 2: Specify the deterministic model

The second step is to specify the deterministic model (same as always)

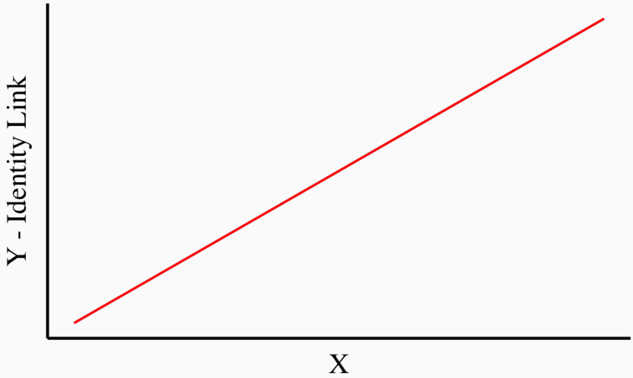
$$\pi_i = \beta_0 + \beta_1 \times \mathrm{Length}$$

The last step is to specify a link function that maps the values between 0 and 1.

The logit link

Standard linear regr. with an ‘identity link‘ maps values between $-\infty, \infty$.

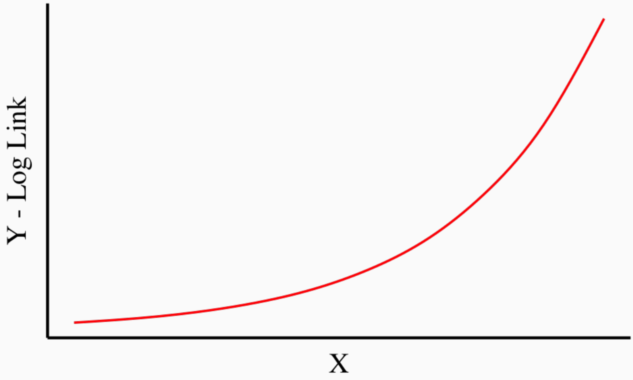
$$\mu = \beta_0 + \beta_1 X$$



The logit link

Last lecture we saw how a ‘log link‘ maps values between $0, \infty$.

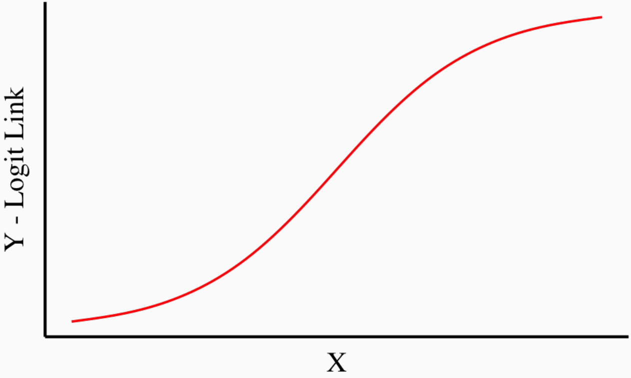
$$\mu = e^{\beta_0 + \beta_1 X}$$



The logit link

The ‘logit link‘ is a link function that maps values between $0, 1$. (How?)

$$\mu = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$



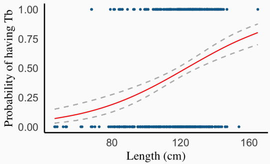
Logistic regression on 0-1 data

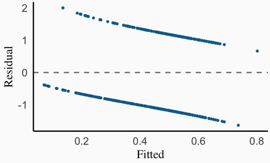
Now we have all the pieces we need for fitting our GLM:

$$Y_i \sim Binomial(1, \pi_i) \qquad E(Y_i) = \pi_i \quad \text{and} \quad (Y_i) = \pi_i \times (1 - \pi_i)$$

$$\pi_i = \frac{e^{\beta_0 + \beta_1 \times \text{Length}}}{1 + e^{\beta_0 + \beta_1 \times \text{Length}}}$$

```
Fit_Logistic <- glm(Tb ~ Length,
                    family=binomial(link="logit"),
                    data = data)

summary(Fit_Logistic)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.137107   0.695381  -5.949 2.69e-09
Length       0.033531   0.005767   5.814 6.09e-09
---

(Dispersion parameter for binomial family taken to be
       1)

    Null deviance: 681.25  on 493  degrees of freedom
Residual deviance: 641.23  on 492  degrees of freedom
AIC: 645.23
```





Are the residuals normally distributed? Should they be?

Logistic regression on 0-1 data

Now we have all the pieces we need for our GLM:

$$Y_i \sim Binomial(1, \pi_i) \qquad E(Y_i) = \pi_i \quad \text{and} \quad (Y_i) = \pi_i \times (1 - \pi_i)$$

$$\pi_i = \frac{e^{\beta_0 + \beta_1 \times \text{Length}}}{1 + e^{\beta_0 + \beta_1 \times \text{Length}}}$$

```
Fit_Logistic <- glm(Tb ~ Length,
                    family=binomial(link="logit"),
                    data = data)

summary(Fit_Logistic)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.137107   0.695381  -5.949 2.69e-09
Length       0.033531   0.005767   5.814 6.09e-09
---

(Dispersion parameter for binomial family taken to be
       1)

    Null deviance: 681.25  on 493  degrees of freedom
Residual deviance: 641.23  on 492  degrees of freedom
AIC: 645.23
```

We can try and calculate a Pseudo-R^2 just like we did when we modelled count data:

$$R^2 = \frac{\text{Null dev.} - \text{Resid. dev.}}{\text{Null dev.}}$$

$$R^2 = \frac{681.25 - 641.23}{681.25} \sim 6\%$$

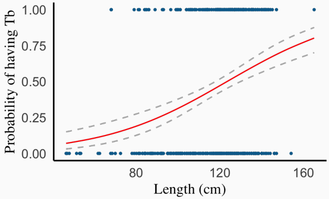
In logistic regression Pseudo-R^2 are almost always going to be low. Not very informative!

Logistic regression on 0-1 data

If we want to get a feel for how our model is performing we can use cross-validation.

Cross-validation splits up a dataset into two pieces one used to fit the model and a second used to test the model's ability to make predictions.

If the model is performing well, it should predict values correctly on average.

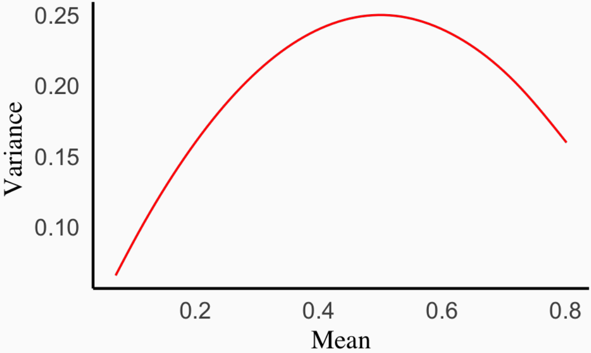


```
library(DAAG)

CVbinary(Fit_Logistic)

Fold:  3 8 4 5 9 10 2 6 7 1
Internal estimate of accuracy = 0.615
Cross-validation estimate of accuracy = 0.615
```

Mean and Variance

For our model the mean and variance are given by:

$$E(Y_i) = \pi_i \qquad \text{and} \quad (Y_i) = \pi_i \times (1 - \pi_i)$$



Variance is largest for intermediate values of π_i.

Logistic Regression on Proportion Data

Proportion data

As many as $\sim$15% of papers in ecology include some kind of proportional data.

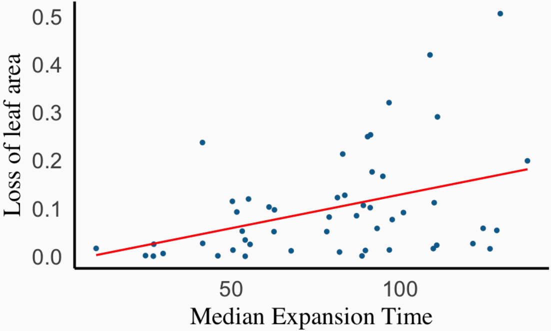
Proportions scale between 0-1, but can take any value between these limits.

Most of the time, ecologists model proportion data using an $\arcsin(\sqrt{p})$ transformation, but this is not an ideal solution:

Warton, D. I., & Hui, F. K. (2011). The arcsine is asinine: the analysis of proportions in ecology. Ecology, 92(1), 3-10.

Leaf data

This example forms part of a paper asking whether plant species with small leaves have shorter expansion times than large leaved counterparts (Moles & Westoby 2000). The data we're going to work with are the percentage loss of leaf area in relation to median expansion time.



Note A linear regression shows a significant relationship

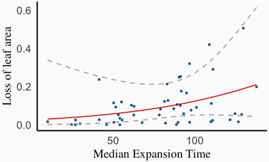
Logistic Regression on Proportion Data

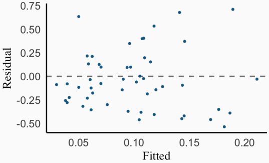
All the pieces we need for out GLM:

$$\text{Leaf Loss}_i \sim \mathit{Binomial}(1, \pi_i) \qquad E(Y_i) = \pi_i \quad \text{and} \quad (Y_i) = \pi_i \times (1 - \pi_i)$$

$$\pi_i = \frac{e^{\beta_0 + \beta_1 \times \text{Expansion Time}}}{1 + e^{\beta_0 + \beta_1 \times \text{Expansion Time}}}$$

```
Fit_Logistic <- glm(loss ~ expansion,
                    family=binomial(link="logit"),
                    data = data)

summary(Fit_Logistic)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.64244    1.58926  -2.292   0.0219
expansion    0.01687    0.01667   1.012   0.3116
---

(Dispersion parameter for binomial family taken to be
       1)

    Null deviance: 6.1273  on 50  degrees of freedom
Residual deviance: 5.0196  on 49  degrees of freedom
AIC: 17.703
```





What to write in a paper

If you end up using GLMs in a paper, it's good practice to lay out the pieces of your model.

For example, for our boar data we would say that we assumed our data were binomially distributed and that we modelled them using a logit link:

$$Y_i \sim Binomial(1, \pi_i) \qquad \text{with} \qquad \pi_i = \frac{e^{\beta_0+\beta_1 \times \text{Length}}}{1+e^{\beta_0+\beta_1 \times \text{Length}}}$$

Present all of your model outputs in a table, put diagnostic plots in supp. material, if you performed model selection make it clear how you got from A to B.

Where to from here?

Where to now?

Generalised Linear Models offer a powerful extension of Gaussian linear regression.

We covered standard GLMs, but this framework can be extended to handle nested data structures just like mixed effects models did for Gaussian linear regression.

Good resource for GLMMs:
http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html

We covered log links and logit links, but there are a number of different link functions that you can use when fitting GLMs. The general idea stays the same, use the one that maps the response variable onto the right scale (e.g., don't use a log-link for a binomial GLM).

Where to now? cont.

We also saw how if our models are of the form:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$$

modifying the off-diagonals of the correlation matrix can correct for various forms of autocorrelation.

$$V = \sigma^2 \underbrace{\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}}_{\text{correlation matrix}}$$

... but because we're working with different distributions now those approaches don't translate cleanly.

Where to now? cont.

If you find yourself with zero-inflated data you might need to use mixture models or hurdle models that are comprised of combinations of different distributions. See Zuur et al. (2009) Ch. 11.