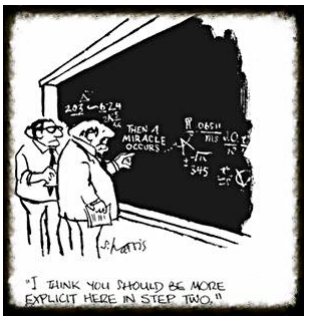
Generalized linear models



Lecture 11
Biological Statistics III
Ayco Tack

Laura's test exercise session is through Zoom

Room number: 460 700 0117

2

Outline

- *Logistic regression / logit model*
  - *Generalized linear models with binomial response*
  - *Using* `glm` *or* `glmer` *with continuous or qualitative x-variables*
- *Poisson regression*
  - *An example with Poisson distributed response, where we use* `lm`, `glm`, `lmer`, *or* `glmer`

3

Generalized linear models

Idea:

- Most of the analyses described thus far were based on linear models that assume equal variances and a normal distribution of the error terms from the fitted model
- While most models are robust to this assumption, it can be hard to ascertain the level of robustness
- We can use transformations to overcome deviations from the assumptions
- However, in some situations transformation is not effective in making errors normal
- We may wish to model the actual data rather than data that are transformed to meet the assumptions
- => we need a modelling technique that allows other types of distribution!

4

Generalized linear models

In generalized linear models, we have:

I. The **response variable,** which has a distribution from the exponential family, for example:
   I. Binomial distribution (for example, if the response is binary)
   II. A Poisson distribution (counts may be Poisson distributed)
   III. Gamma distribution
   IV. Negative binomial
   V. Normal distribution

II. We have qualitative and/or quantitative **predictors**

III. The response and predictor variables are linked by the **link function**

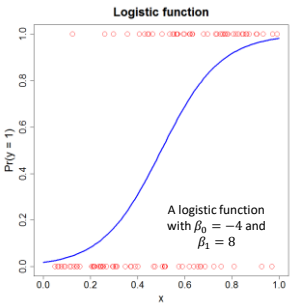
5

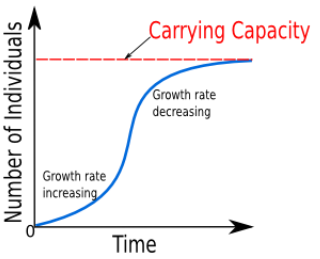
Simple logistic regression

- A binary response variable (categorical with two levels, which are coded as zero and one), for example:
  - presence/absence, dead / alive, blue / red wings
  - if you have more than two categories, you may consider the Dirichlet distribution (e.g. https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.13234)
- $\pi(x)$ is the probability that *y* equals 1 for a given value of *x*
- We usually fit the logistic regression model, which is a nonlinear model with a sigmoidal shape:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

, where β_0 (intercept) and β_1 (slope) are parameters to be estimated. We could use a nonlinear model to estimate the β_0 and β_1; but this approach is rather tedious (https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.12044)



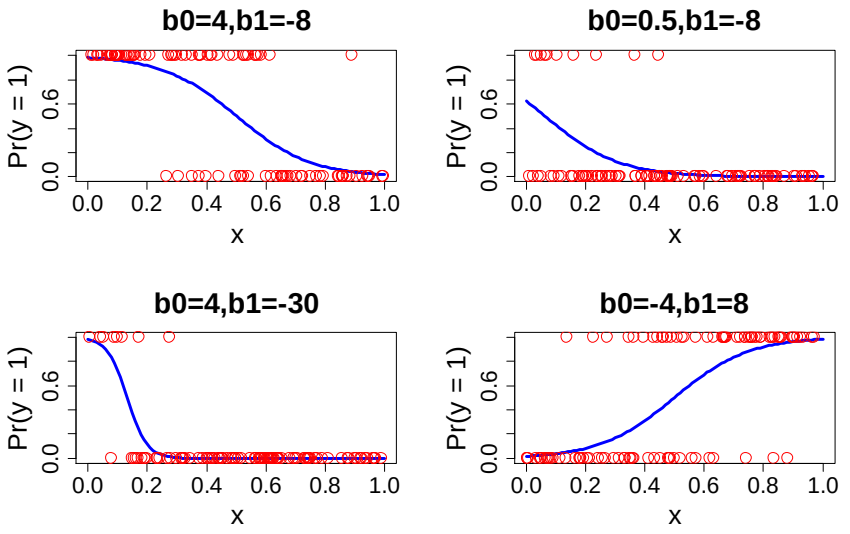


$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

$$P(t) = \frac{KP_0 e^{rt}}{K + P_0(e^{rt} - 1)}$$

6

Illustration of a logistic function



7

Simple logistic regression

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

, where β_0 (intercept) and β_1 (slope) are parameters to be estimated. We could use a nonlinear model to estimate the β_0 and β_1; but this approach is rather tedious

We can transform $\pi(x)$ so that the logistic model closely resembles a familiar linear model:
1. We first calculate the odds for an event to occur, i.e. the probability of $y_i = 1$ relative to $y_i = 0$

$$\frac{\pi(x)}{1 - \pi(x)}$$

, where we note that the odds are >1 when the probability of $y_i = 1$ is greater than the probability for $y_i = 0$

2. We then take the natural logarithm:

$$\ln\left[\frac{\pi(x)}{1 - \pi(x)}\right]$$

, which is the logit transformation or link function and ranges from $-\infty$ to $+\infty$ (rather than the probability, which is constrained between 0 and 1). This broad range is more appropriate for a linear model. We can model this term against $\beta_0 + \beta_1 x_i$

If we solve for $\pi(x)$, we get the equation $\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$

3. We will call this term $g(x)$, which can be modelled against the predictor as:

$$g(x) = \beta_0 + \beta_1 x_i$$

8

http://logisticregressionanalysis.com/910-how-to-solve-the-logistic-regression-equation-for-the-probability-p/

The link function

Idea:

- The link function links the expected value of Y to the predictors by the function:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X + \ldots.$$

, where $g(\mu)$ is the link function and we want to estimate the parameters β_0, β_1 and others

Three link functions are commonly used:

1. **Identity link**, which models the mean or expected value of Y:

$$g(\mu) = \mu$$

2. **Log link**, which models the log of the mean and is used for count data (which cannot be negative):

$$g(\mu) = \log(\mu)$$

3. **Logit link**, which is used to model binary data:

$$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$$

9

Fitting logistic regression with maximum likelihood:

https://www.youtube.com/watch?v=BfKanl1aSG0

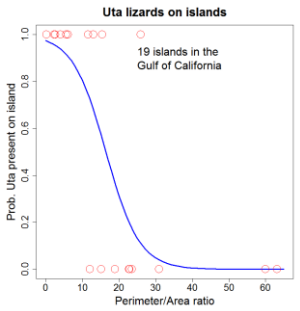
10

Some further remarks on GLM

- GLMs are parametric models, as we specify a probability distribution for the response variable (and hence for the error terms from the model)
- GLMs are linear models, as we can describe the response variable by a linear combination of predictors
- We use maximum likelihood estimation of the parameters, which is based on iterative reweighted least squares algorithms like Newton-Raphson
- For the Poisson and binomial distribution the variance is related to the mean; we may use the more flexible alternative that estimates the dispersion parameter from the data (rather than constraining it by the value inherent to the chosen distribution). This is called the quasi-likelihood models (with the quasi-Poisson and the quasi-binomial), or, for count data, there is also the negative binomial (e.g. in package glmmTMB).
  - In biology, it is particularly common to have count data that is overdispersed
- If we use an identity link, we simply model the expected value (the mean of *y*) of the response variable; the OLS (ordinary least squares) estimates will be very similar to the ML estimates from fitting the GLM

11

Presence/absence of lizards on islands



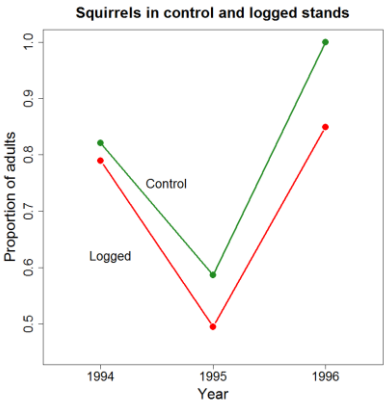
Quinn & Keough, Box 13.1

- We will model $g(x) = \beta_0 + \beta_1 (P/A\ ratio)_i$
  , where $g(x)$ is the natural log of the odds that Uta is present on an island (relative to being absent)
- H_0: $\beta_1 = 0$, which means the absence of a relationship between Uta presence and P/A ratio
- We can judge the improvement in fit using the log-likelihood statistic (deviance)
- If the explained deviance is too large, we reject the hypothesis of no effect of x on the distribution of y
- A logistic regression gives the estimates $b_0 = 3.6061$ and $b_1 = -0.2196$ for the parameters of the logistic function. b_1 is an estimate of how much the natural log of the odds changes for a change in one unit of x; the negative sign of b_1 indicates that the natural log of the odds of Uta being present decreases with the P/A ratio
- For the test of the null hypothesis that the true $\beta_1 = 0$, we get a log-likelihood X^2 of 12.066 with $df = 1$ and $p = 0.0005$. We conclude that there is an effect of island perimeter/area ratio on the probability of finding Uta lizards on an island.

12

Effect of logging on squirrel age structure

The effect of logging on the demography of southern flying squirrels was investigated over three years in southern Arkansas (Quinn & Keough, Table 14.3b)



- *The squirrels were caught in traps and classified as juveniles or adults*

13

Analysis of squirrel age structure

Using a GLM:

- First fit a glm model with binomial response (adult vs. juvenile) and with Treatment, Year and their interaction as explanatory variables.

ANOVA table (type II)

Source	df	LR Chisq	p
Treatment	1	2.244	0.13
Year	2	44.389	**<0.0001**
Treatment x Year	2	1.882	0.39

- The interaction is not statistically significant, so we fit a model without the interaction term

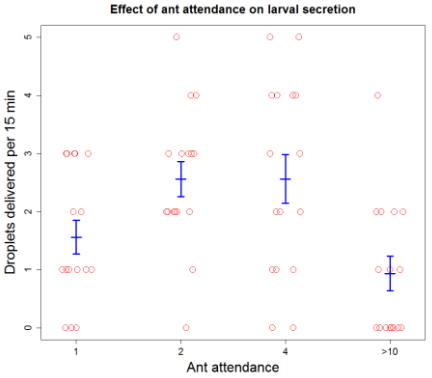
ANOVA table (type II)

Source	df	LR Chisq	p
Treatment	1	2.244	0.13
Year	2	44.389	**<0.0001**

- We conclude that the treatment (logging) did not have a statistically significant effect on squirrel age structure, but the age-structure differed between years

14

Common blue butterfly-ant mutualism



Example: Mutualism between lycaenid butterfly larvae and ants. Larvae ($n = 16$) of the common blue (*Polyommatus icarus*) were exposed to different numbers of attending ants (*Lasius flavus*), and the number of droplets delivered by a larva was measured

Approaches:

- These data could be analysed using the nonparametric Friedman test
- Here we will try several parametric approaches
  - If we consider the process of secretion of droplets, it seems possible that the number of droplets in 15 min is Poisson distributed. We might then try a transformation (square root is a possibility)
  - Alternatively, we could use the Poisson family for `glm` or `glmer`
  - Finally, since we have several cases of 0 droplets, we could reduce the observations to a binary (Yes/No) variable

15

Analysis of butterfly-ant mutualism with `lm`

R-command:

```
fm1 = lm(sqrt(Drops+ 1) ~ AntTreatm + Larva, data=dat)
```

we get, from `Anova(fm1)`

ANOVA table (Type II tests)

	df	F value	p
AntTreatm	3	9.564	**<0.0001**
Larva	15	3.351	**0.0008**

Some thoughts:

- Examining diagnostic plots suggests that this analysis is basically OK, although one may doubt whether the residuals are normally distributed. In any case, the conclusion from the analysis is that the effect of ant treatment is statistically significant.
- Another issue is the randomized complete block design of the study (the individual larvae are the blocks), where the blocks naturally correspond to a random effect (the larvae were randomly picked from a population). This suggests that a mixed model analysis would be appropriate.

16

Analysis of butterfly-ant mutualism with lmer

R-command:

```
fm2 = lmer(sqrt(Drops+ 1) ~ AntTreatm + (1|Larva), data=dat)
```

we get, from `Anova(fm2)`

Analysis of Deviance table (Type II Wald chisquare tests)

	df	Chisq	p
AntTreatm	3	28.692	**<0.0001**

Some thoughts:

- The conclusion is again that the effect of treatment is statistically significant. From the output from `summary(fm2)`, we also find that the variance components associated with larva and residual have standard deviations 0.243 and 0.318.
- The estimates of the variance components might be one reason to prefer this mixed-model analysis to the previous one.
- Even so, for these data it seems very natural to try to fit a generalized linear model with Poisson response variable.

17

The log link function for Poisson glm or glmer

For generalized linear models with Poisson response variable, the log is the default link function. For a two-way ANOVA without replication, this means that the logarithm of the expected value is modelled as a linear function of the treatment (α_i) and block (β_i) effects:

$$\log(\mu_{ij}) = intercept + \alpha_i + \beta_j$$

The expected value is thus modelled as a product of independent treatment and block contributions

$$\mu_{ij} = e^{intercept + \alpha_i + \beta_j} = constant * e^{\alpha_i} * e^{\beta_j}$$

The idea of the modelling is that the response should be Poisson distributed with mean μ_{ij}. If that is true, the variance should be equal to μ_{ij}. It can sometimes happen that the variance is bigger than this, which is called overdispersion. We can then use the quasi-Poisson distribution or the negative binomial (the latter in package glmmTMB).

18

Analysis of butterfly-ant mutualism with glm

R-command:

```
fm3 = glm(Drops ~ AntTreatm + Larva, family=poisson, data=dat1)
```

we get, from `Anova(fm3)`

Analysis of Deviance table (Type II test)

	df	LR Chisq	p
AntTreatm	3	17.286	**0.0006**
Larva	15	32.130	**0.0062**

Some thoughts:

- We get some additional information from the output of `summary(fm3)`. A traditional way of checking for overdispersion is to compare the residual deviance (39.775) with the residual degrees of freedom (45). If the residual deviance is much bigger than the degrees of freedom, there may be overdispersion. There seems to be no reason to worry about overdispersion for this example.
- The conclusion is again that there is a statistically significant effect of ant treatment on larval droplet delivery.

19

Analysis of butterfly-ant mutualism with glmer

R-command:

```
fm4 = glmer(Drops ~ AntTreatm + (1|Larva), family=poisson, data=dat)
```

we get, from `Anova(fm4)`

Analysis of Deviance table (Type II Wald chisquare tests)

	df	Chisq	p
AntTreatm	3	15.321	**0.0016**

- The conclusion is again that there is a statistically significant effect of ant treatment on larval droplet delivery.

20

How to test assumptions of generalized linear (mixed) models

INVESTIGATING THE RESIDUALS

sjPlot works well with linear mixed models. But interpreting the residuals from GLMM is very difficult, as the residuals from the Poisson or binomial distribution do not match those from linear models with a Gaussian distribution. So we lack intuition on how the residuals from nice-fitting models are supposed to look like.

You may use the DHARMa package:

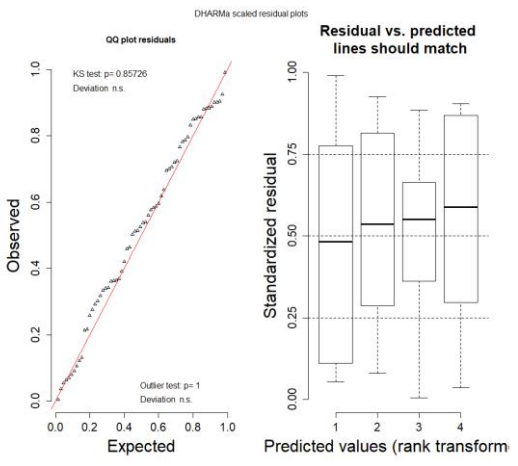
- The DHARMa package simulates residuals, which are more intuitive, and allow for easier diagnostics.
- The vignette is (relatively) easy to read: https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMa.html
- You can create diagnostic plots, but you have to get used to the new guidelines in the standard plots

```
simulationOutput <- simulateResiduals(fittedModel = fittedModel, plot = T)
plot(simulationOutput)
```

- In the left figure, we have the expectation of a classic QQ-plot for linear models
- In the right figure, we expect a uniform distribution in the y-direction for all values on the x-axis
- There is a separate test for overdispersion also: `testDispersion(simulationOutput)`
- Also good to inspect plots of the residuals against all predictors

21

Diagnostic plots from the glmer model from the DHARMa package for the butterfly model



- The QQ-plot looks nice
- The residuals seem reasonably well distributed in each of the four treatment levels
- A separate test for overdispersion does not indicate any problems

Code used:

```
library(DHARMa)
simulationOutput <-
simulateResiduals(fittedModel = fm4,
plot = T)
plot(simulationOutput)
testDispersion(simulationOutput)
```

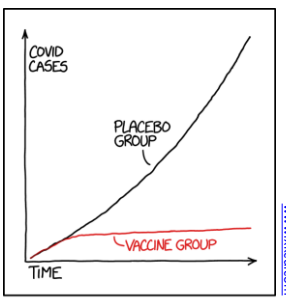
22

Laura's test exercise session is through Zoom

Room number: 460 700 0117

23

Related reading and information



STATISTICS TIP: ALWAYS TRY TO GET DATA THAT'S GOOD ENOUGH THAT YOU DON'T NEED TO DO STATISTICS ON IT

- **Quinn & Keough:** Sections 14.1 - 14.2
- **Crawley:** Sections 15.1 - 15.4, 17, 17.1 and 17.2

24