Workshop 7: Generalized linear models

In this workshop, we will fit generallized linear models to data, implemented in the R command `glm`. A generalized linear model is useful when the response variable has a specified distribution other than the normal distribution, and when a data transformation is inadequate or impossible. Example situations include binary response data (0/1, alive/dead alive) or data that are counts (number of offspring, leaves, etc). The approach is also useful in the analysis of contingency tables.

Natural selection in song sparrows

The song sparrow population on the island of Mandarte has been studied for many years by Jamie Smith, Peter Arcese, and collaborators. The birds were measured and banded and their fates on the island have been recorded over many years. Here we will look for evidence of natural selection using the relationship between phenotype and survival.
The data file, located here, gives survival of young-of-the-year females over their first winter (0=died, 1=survived). The file includes measurements of beak and body dimensions: body mass (g), wing length, tarsus length, beak length, beak depth, beak width (all in mm), year of birth, and survival. These data were analyzed previously in Schluter & Smith (1986, *Evolution* 40: 221–231).

Read and examine the data

1. Read the data from the file and inspect the first few lines to make sure it was read correctly.

2. We'll be comparing survival probabilities among different years. To this end, make sure that year is a categorical variable in your data frame. Hint: you can use the function `str` to check the type of variable year is, and `as.character` to change it to a character.

3. Plot survival against tarsus length of female sparrows. Use a method to reduce the overlap of points (the response variable is 0 or 1) to see the patterns more clearly.

4. Examine the plot. Can you visualize a trend? Add the best fit line from a simple `lm` (note that this is not the appropriate model for this data).

Fit a generalized linear model

Let's start by ignoring the fact that the data are from multiple years. If time permits we can add year to the model to see how much difference it makes.

1. The response variable is binary. What probability distribution is appropriate to describe the error distribution around a model fit? What is an appropriate link function?

2. Fit a generalized linear model to the data on survival and tarsus length.

3. Use `visreg` to visualize the model fit. Remember to load the `visreg` package.

4. Obtain the estimated regression coefficients for the fitted model. What is the interpretation of these coefficients? On a piece of paper, write down the complete formula for the model shown in the `visreg` plot.

5. Use the coefficients to calculate the predicted survival probability of a song sparrow having tarsus length 20.5mm. Does the result agree with your plot of the fitted regression curve?

6. The ratio (-intercept/slope) estimates the point at which probability of survival is changing most rapidly. In toxicology this point is known as the LD_{50}. Calculate this value and compare it visually with the fitted curve. Does it agree? Finally, the slope of the curve at a given value for the explanatory variable x is

$$b * p(x) * (1–p(x))$$

   where b is the slope coefficient of the fitted logistic regression model and $p(x)$ is the predicted probability of survival at that x.

7. Calculate the likelihood-based 95% confidence interval for the logistic regression coefficients.

8. The `summary(out)` output for the regression coefficients also includes z values and P-values. What caution would you take when interpreting these P-values? Use `drop1` to test the null hypothesis of zero slope.

9. If time permits, add year to your logistic regression model as a categorical variable (ensure that year is a factor in your data set). Ignore the interaction between year and tarsus length. Plot the resulting curves using `visreg` with option `xvar='year'`. Is there any evidence of a difference among years in the relationship between survival and tarsus length in these data? Use `emmeans` to calculate the model estimates of mean survival in each year.

Crab satellites

The horseshoe crab, *Limulus polyphemus*, has two alternative male reproductive morphs. Some males attach to females with a special appendage. The females bring these males with them when they crawl onto beaches to dig a nest and lay eggs, which the male then fertilizes. Other males are satellites, which are unattached to females but crowd around nesting pairs and obtain fertilizations. What attributes of a female horseshoe crab determine the number of satellite males she attracts on the beaches?
The data here provide measurements of 173 female horseshoe crabs and record the number of satellites she attracted. The data were gathered by Brockman (1996. Satellite male groups in horseshoe crabs, Limulus polyphemus. *Ethology* 102:1-21) and were published by Agresti (2002, *Categorical data analysis*, 2nd ed. Wiley). The variables are female color, spine condition, carapace width (cm), mass (kg), and number of satellite males.

Read and examine the data

1. Read the data from the file. View the first few lines of data to make sure it was read correctly. Use the `str` command to see the variables and groups.

2. Plot the number of satellites against the width of the carapace, a measure of female body size. Fit a smooth curve to examine the trend.

Fit a generalized linear model

1. What type of variable is the number of satellites? What probability distribution might be appropriate to describe the error distribution around a model fit? What is the appropriate link function?

2. Fit a generalized linear model to the relationship between number of satellite males and female carapace width.

3. Use `visreg` to examine the relationship on the transformed scale, including confidence bands. This plot reminds us that on the transformed scale, `glm()` is fitting a straight line relationship. Don't worry about the points - they aren't the transformed data, but rather are "working values" for the response variable from the last iteration of model fitting, which `glm()` uses behind the scenes to fit the model on the transformed scale.

4. Extract the estimated regression coefficients from your model object. What is the interpretation of these coefficients? On a piece of paper, write down the complete formula for your fitted model.

5. Plot the data on the original scale, and add the `glm()` model fit to your plot. Note that it is not linear.

6. Calculate the likelihood-based 95% confidence interval for the regression coefficients. The most useful estimate is that for the slope (`width.cm`): the quantity `exp(x)` (where x is the slope) corresponds to the magnitude of expected change in the response variable accompanying a 1-unit change in the explanatory variable.

7. Test the null hypothesis of no relationship between number of satellite males and female carapace width. Notice how small the P-value is for the null hypothesis test for the slope. I'm afraid that this is a little optimistic. Why? Read on.

8. When you extracted the regression coefficients from your model object, you probably saw the following line of output: "(Dispersion parameter for poisson family taken to be 1)". What are we really assuming here?

9. If you did not want to rely on this assumption (or you wanted to estimate the dispersion parameter), what option is available to you? Refit a generalized linear model without making the assumption that the dispersion parameter is 1. Save the results in a new `glm` object so that you can compare your results with the previous fit.

10. Extract and examine the coefficients of the new `glm` model object. Examine the estimated dispersion parameter. Is it close to 1? On this basis, which of the two `glm` fits to the same data would you regard as the more reliable?

11. How do the regression coefficients of this new fit compare with the estimates from the earlier model fit? How do the standard errors compare? Why are they larger this time?

12. Compare the visreg plot of the current model to that of the earlier fit. What difference do you notice?

13. Redo the test of significance for the slope of the relationship between number of satellite mates and female carapace width. Use the F-test, rather than the likelihood ratio test in the `drop1` command (because we are using a quasi-distribution). How do the results compare with those from the previous fit?

Prion resistance not futile

This last exercise is to demonstrate the use of `glm()` to model frequencies of different combinations of two (or more) variables in a contingency table. The presence of an interaction between the variables indicates that the relative frequencies of different categories for one variable differ between categories of the other variable. In other words, the two variables are then not independent.
Kuru is a prion disease (similar to Creutzfeldt–Jakob disease) of the Fore people of highland New Guinea. It was once transmitted by the consumption of deceased relatives at mortuary feasts, a ritual that was ended by about 1960. Using archived tissue samples, Mead et al. (2009, *New England Journal of Medicine* 361: 2056-2065) investigated genetic variants that might confer resistance to kuru. The data in the accompanying table are genotypes at codon 129 of the prion protein gene of young and elderly individuals all having the disease. Since the elderly individuals have survived long exposure to kuru, unusually common genotypes in this group might indicate resistant genotypes. The data are here.

Read and examine the data

1. Read the data from the file. View the first few lines of data to make sure it was read correctly. Use the `str` command to see the variables and groups.

2. Create a contingency table comparing the frequency of the three genotypes at codon 129 of the prion protein gene of young and elderly individuals (all having the disease). Notice any pattern? Hint: You can use the function `table`. By comparing the frequencies between young people and older people, which genotype is likely to be more resistant to the disease?

3. Optional: create a grouped bar graph illustrating the relative frequencies of the three genotypes between afflicted individuals in the two age categories.

Fit a generalized linear model

1. Model the frequencies in the contingency table with a generalized linear model. You will first need to convert the contingency table to a "flat" frequency table using `data.frame()`.

2. To begin, fit the additive model (i.e., a `glm()` model lacking an interaction between the two variables genotype and age).

3. Examine the fit of the additive model to the frequency data using `visreg`. Notice how the additive model is constrained from fitting the exact frequencies in each category.

4. Repeat the model fitting but include the interaction term as well. Visually compare the fit of the model to the data. Notice how this "full" model really is full - it fits the frequencies exactly.

5. Using the "full" model, which includes the interaction term, test whether the relative frequencies of the three genotypes differs between the two age groups.