

Exercise

Exercise: Binomial (Proportions) GLM - mites survival

The data for this exercise were collected for an experiment to test the toxicity of different chemicals at various concentrations to a species of mite. The aim was to determine whether the proportion of mites surviving was related to the concentration of the chemical and whether this relationship depended on the type of chemical the mites were exposed to. We will be using a binomial GLM again for this example, but this time we will be using it to model proportion data, from counts of “successes” (= mite survival, unless you don’t like mites), and “failures” (= mite death).

- Variables:
 - ID index of the observations
 - **Concentration**: Concentration of the chemical
 - **Toxic**: Chemical type
 - **Dead_mites**: Number of dead mites in assay
 - **Total**: Initial number of mites in assay
 - **Proportion**= **Dead_mites** / **Total**

As in previous exercises, either create a new R script (perhaps call it GLM_BinomProps) or continue with your previous R script in your RStudio Project. Again, make sure you include any metadata you feel is appropriate (title, description of task, date of creation etc) and don’t forget to comment out your metadata with a # at the beginning of the line.

1. Import the data file ‘DrugsMites.txt’ into R. We want to model the variable Toxic as a categorical predictor with 4 levels, so create a new variable with Toxic as a factor.
2. Perform the usual graphical data exploration, looking for outliers, relationships between predictors, and between response and predictors etc. You can use the variable **Proportion** in the data frame for these plots.
3. In order to model the proportion of mites surviving, create a new variable (called something creative like **Living_mites** for example) representing the number of surviving mites by differencing the two variables **Dead_mites** and **Total**.

4. We can now use this new variable when specifying a binomial GLM. Recall from the lecture that the response variable should be a data frame consisting of two columns, `cbind(Living_mites, Dead_mites)`. Ask if in doubt. If you hate mites you could also swap the order of the two columns: you would then be modelling the proportion that die.
5. Obtain summaries of the model output using the `summary()` function. Make sure you understand the mathematical and biological interpretation of the model, by writing down the complete model on paper (with distribution and link function). What biological hypothesis does each term imply, qualitatively?
6. Do you need to check for overdispersion? If so, how do you do it?
7. Do you need to perform model selection? What is the final model?
8. Perform model validation: are you satisfied with the model?
9. Obtain the fitted values from the model on the scale of the response, and plot to aid model interpretation. How do you interpret the results?
10. (Optional) Include the 95 % CI on the plot above. You will need to obtain the fitted values and SE on the scale of the link function, calculate the CI and then back-transform.

End of the Binomial (Proportions) GLM - mites survival