3.8 Practical 3

Dataset required: `insect.dta`

Introduction

The purpose of this session is to learn how to fit and interpret generalised linear models to binary data.

We will use data from a toxicological experiment. Eight groups of insects were exposed for five hours to gaseous carbon disulphide (CS_2) at different concentrations, with the purpose of investigating how the risk of death depends on the dose received.

The data are in `insect.dta`, with one record per group as defined by CS_2 dose, and three variables, as follows:

dose = CS_2 dose (mg/l);

r = number of insects killed;

n = number of insects in group.

You should have your own Do file and run the commands yourself, but you should discuss the results, interpretations and any queries within your Breakout Room.

Aims

The aims of the analysis are to answer the following questions:

1 Is there an association between dose level and the proportion of insects killed?

2 On what scale is this association best modelled: identity, log, logit?

3 What is the nature of this association: linear or quadratic?

Investigation into effect of CS_2 at different doses on insect survival

1 Open Stata, start a new Do file, and load the data. Explore the dataset. Are there any missing values? How many insects do you have data for?

2 Generate a new variable for the proportion of insects killed at each dose of CS2. List these proportions and plot them against the dose. What do you conclude? Why would simple linear regression not be appropriate for analysing the association between the proportion killed and dose?

3 Generate new variables for the

   (a) log

   (b) log odds

   of being killed at each dose. Plot these values against the dose. What do you conclude?

   Discussion: Looking at these plots, which link function is likely to best fit these data?

4 Write down algebraically an appropriate generalized linear model for these data, which can be used to investigate the association between dose received and risk of death. Your model should specify the distribution, the linear predictor and the link function.

5 Fit the model and obtain MLE's of the parameters using the `glm` command in Stata.

(a) According to the fitted model, what is the estimated probability of death at a dose of 55 mg/l CS_2?

(b) Calculate (using pen, paper and a calculator) the dose that, on the basis of this model, would lead to a 50% death rate (sometimes termed the LD50).

(c) Is there evidence of an increasing risk of death with increasing dose?

(d) Interpret the dose parameter in terms of an odds-ratio, and calculate its 95% CI. Check your calculation by using the eform option of the `glm` command.

6 Look at the fitted values from the model obtained using predict. What variable has been fitted? Plot both the fitted and observed proportions against the dose.

Discussion: Compare the fitted and observed proportions. What do you conclude?

7 Calculate a new variable for the square of the dose. Use this to investigate whether the inclusion of a quadratic dose term improves the fit of the linear dose model.

(a) Test whether there is evidence against linearity by performing a Wald test of the quadratic coefficient. What do you conclude?

(b) Compare the fitted proportions from this model with those you obtained from the previous model and also with the original data, for example by plotting a graph against dose. In what way has the fit been improved?

8 Now consider dose as a categorical variable, fitting generalized linear models with logit, log and identity links.

(a) Fit the models using the following series of commands, interpreting each of the parameter estimates obtained. What do you notice?

```
egen dosecat = rank(dose)

glm r i.dosecat, family(binomial n) link(logit)

glm r i.dosecat, family(binomial n) link(log)

glm r i.dosecat, family(binomial n) link(id)
```

(b) Fit a logistic regression model that includes dose as both categorical and continuous. Test the effect of the categorical dose variable.

```
glm r dose i.dosecat, family(binomial n) link(logit)

test 2.dosecat 3.dosecat 4.dosecat 5.dosecat 6.dosecat 7.dosecat
```

Discussion: Why are there estimates for only six (of eight) dose categories in this model? Discuss the interpretation of each parameter estimate and the test statistic computed by the final command above.

Discussion: Working together with one or more colleagues (in your Breakout Room if online), write a paragraph to answer the aims of the analysis. If online, one of you should post your group's paragraph in the Zoom chat.