Sock-throwing and logistic models

Overview

In wildlife biology, we are often dealing with binary variables – survived/dead, detected/not detected, occupied/not occupied, etc. – and we estimate the probability of "success", ie. survival, detection, occupancy. Often we want to investigate a relationship between the binary variable (the 'response') and other variables known as explanatory variables, predictors, or just covariates. The most common way to do this uses logistic modelling.

In this exercise, you try throwing the sock into a box from different distances and with either hand. We want to know if the probability of success depends on distance or which hand you use. We'll also record the gender of the thrower, as that may also have an effect. And we'll record the score on a die, which we would not expect to affect the result. We'll analyse the data in R with the `glm` function.

Objective

To learn how to set up a logistic model to understand how distance, hand and gender, and the score on a die affect success in getting the sock into the box, and to make predictions about success when we know the distance, hand, gender and die score.

Collecting the data

You will try throwing the sock into the box placed against the wall from distances of 2, 3, 4 and 5m, both with your "good" hand – the one you write with – and your "bad" hand. For each throw, record whether you succeeded or not (you're allowed to bounce the sock off the wall). After each throw, roll a 10-sided die and record the result (0 to 9).

	A	B	C	D	E	F
1	Name	result	gender.male	distance	hand.good	die
14	John	0	1	2	1	3
15	John	1	1	2	0	4
16	John	1	1	3	1	7
17	John	1	1	3	0	9
18	John	0	1	4	1	4
19	John	1	1	4	0	9
20	John	0	1	5	1	7
21	John	0	1	5	0	8
22	Amanda	1	0	2	1	9
23	Amanda	1	0	2	0	5
24	Amanda	0	0	3	1	3

Record the results in a spreadsheet as shown on the right. Result = 1 means the sock went in; gender.male = 1 for males; hand.good = 1 if you were using your good hand.

We'll gather the results of all the groups into one spreadsheet and distribute it as a .csv file with a name like "sox_in_box.csv".

Logistic models

The simplest model connecting a response variable (y) to covariates (x_1, x_2, …) is a **linear model**:

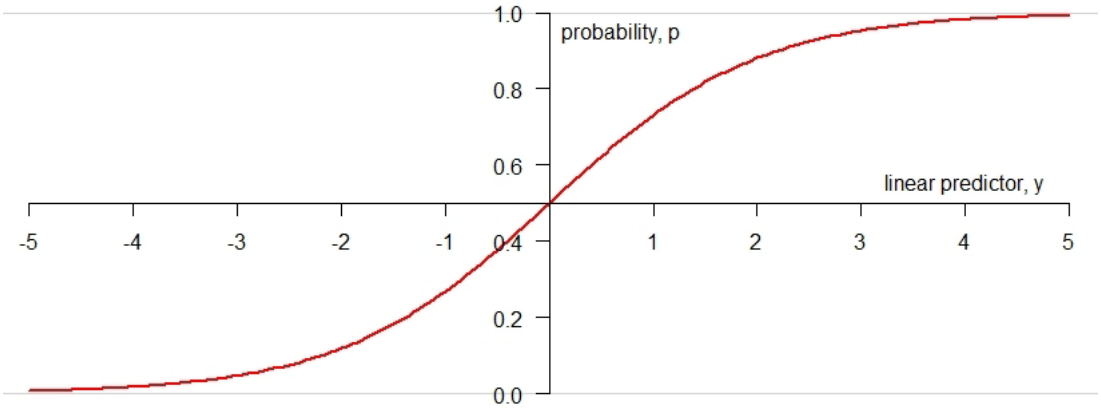
$$y = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \ldots$$

In our case, a simple model would be:

$$y = \beta_0 + \beta_1 \times distance$$

(We'll investigate the effect of hand and gender later.)

We want to estimate the parameters (β_0, β_1, etc) from the data using maximum likelihood. The problem with this model is that y can take any value between $-\infty$ and $+\infty$, while the variable we are interested in, the probability of getting the sock in the box, p, has to be between 0 and 1. So we usually use a logistic link (or "logit link") to connect p to y:



With this link, when $y = \infty$, $p = 1$, and when $y = -\infty$, $p = 0$. When $y = 0$, $p = 0.5$.

The formulae, if you need them, are: $p = \frac{e^y}{1+e^y} = \frac{1}{1+e^{-y}}$ and $y = \log\left(\frac{p}{1-p}\right)$

Now the expression $y = \beta_0 + \beta_1 \times distance$ is called the **linear predictor** for p.

Analysis in R

Import the data into R : Go to the Modelling folder and double-click on the **R** icon to start R. Go to File > Open script…" and open "Sox_in_Box.R".

Follow the code in the script to import the data from the .csv file.

Take a look at the data and plot *result* vs *distance*. Since *result* only has values of 0 and 1, the dots will be on top of each other. One way to get around this is to add a small random value to *result*, which we can do with `jitter`.

Running the models : Logistic models are a type of Generalized Linear Model (GLM), which we can run with the `glm` function in R. Putting `family='binomial'` in the call to `glm` specifies a logistic model.

You specify the response and the covariates with the tilde (~), eg.

```
result ~ hand.good + distance
```

specifies a model with *result* as the response and *hand* and *distance* as covariates.

Run a model with *distance* as the covariate and look at the parameter estimates. What do they mean?

Similarly, run models with *hand* and both *distance* and *hand* as covariates. Make sure you understand what the parameters mean in each model.

Also run a model with no covariates, called the "null model"; do that with `result ~ 1`. Then compare the value of AIC for all the models. Which is the best model? Is there uncertainty about which model is best?

Key points

- With binary data, we want to estimate the probability of "success" – survival, detection, etc.
- If we have covariates (aka predictors or explanatory variables), we put these into a **linear predictor** and link this to the probability of success with a **link function**, usually the logistic (or "logit") link.
- We can then obtain maximum likelihood estimates for the parameters of the model, calculate AIC and compare different models.

- In R we can specify models with the ~ convention; R converts this to a model matrix which it uses to calculate the linear predictor. In PRESENCE and MARK we work with the model matrix (aka design matrix).

Additional exercises

- Add *gender* as a covariate to each of the models we ran in R. Does including gender improve the model? Is there evidence of a gender effect?
- The file *sox_In_Box_All.csv* has data from past workshops where we have done this activity. Load this file in R and run the different models and compare them. Why do some effects show up with this data set but not with the data from your group?

References

Morrell, C H; R E Auer. 2007. Trashball: A logistic regression classroom activity. *J Statistics Education* **15**:1. On-line at www.amstat.org/publications/jse/v15n1/morrell.html