MTH225-13 In-class exercise 1: Logistic Regression

Names: (signatures only please, printed names will not be counted)

1.) 4.)

2.) 5.)

3.) 6.)

Overview

Previously we estimated the probability of success in a Bernoulli trial.

A closely related problem is one where we observe the number of "successes" in some number n of independent Bernoulli trials, each with the same probability of success p.

In this case, the number of successes has a binomial distribution. Usually the probability of success is the parameter of interest.

Models for this type of data are called **logistic regression** models.

The objective in this analysis is to run a Bayesian logistic regression.

Description of the data

Suppose we are interested in the survival probability p of tadpoles raised in captivity.

We have data from 48 different tanks that represents the original number of tadpoles in the tank, the number that survived, and a categorical variable for the size of the tank (big or small).

We are interested in estimating the survival probability and the precision of our estimate of that probabillity for each tank.

We are also interested in whether the size of the tank has any effect on the survival probability.

The logit transform and logistic regression model

Like any probability, we know p has to be in the interval $[0, 1]$. However, the computations will be more stable if we "stretch" the range of p to $-\infty, \infty)$. This is known as a *logit* transformation, and the formula for it is:

$$l(p) = \log\left(\frac{p}{1-p}\right) \quad 0 < p < 1$$

A bit of thought should convince you that

$$l(p) \to \infty \quad \text{as} \quad p \to 1$$

and

$$l(p) \to -\infty \quad \text{as} \quad p \to 0$$

The wider range for p makes the algorithms used to estimate it more stable numerically.

The model for logistic regression can be written as:

$$l(p) = X\beta + e$$

where X represents a 'design' matrix or matrix of predictors, and β is a vector of regression coefficients, one for each predictor.

The e term is a vector representing "noise" or measurement error in the response, and is usually assumed to be a vector of independent, identically distributed normal random variables with mean zero and a common standard deviation σ_e.

In this problem, we have two classification-type predictors:

- $a_{tank}[i]$ represents the effect of tank i on $l(p)$
- $a_{size}[j]$ represents the effect of tank size on $l(p)$ (j is 1 for small tanks, 2 for large tanks)

We could also describe this model as a two-way ANOVA with binomial data, but it is more common to lump this type of model into "logistic regression".

Logistic regression is very widely used in the sciences.

Running the analysis

Use the following files (downloaded from github) to run the logistic regression:

- `MTH225_week13_IC1a_example1.Rnw`
- `MTH225_week13_IC1a_example1.stan`

Interpreting the results

The STAN code will produce estimates for the following parameters:

- `a_tank[i]` Effect of tank i on $l(p)$
- `a_size[j]` Effect of size j on $l(p)$
- `lp[i]` Logit-transformed probability of survival for tank i
- `p[i]` Probability of survival for tank i

Use the STAN output to answer the following questions:

- Find the point estimate of the logit-transformed survival probality $l(p)$ for tank 1 and a 95% credible interval for it.
- Find the point estimate of the survival probability p for tank 2 and a 95% credible interval for it.
- Based on the 95% credible intervals, are the survival probabilities for tanks 1 and 2 different (i.e., do the credible intervals overlap? If so, by how much?)
- Does tank size appear to make a difference in the survival rate?

Note: The data used is from *Statistical Rethinking: A Bayesian Course with Examples in R and STAN* by Richard McElreath (CRC Press, December 2015)