How to analyze binary responses?
Tradeoffs between reproduction and survival

Figure 1. Flowers and fruits of *Hypericum cumulicola*

Ecological theory hypothesizes that, given the restricted energy budget of organisms, there may be tradeoffs among different vital rates. The evidence for a complete energy partitioning is controversial, and it is not discussed in this demo. The purpose of this session is to demonstrate how to perform a logistic regression model in *R* and *Stan*. We will calculate a logistic regression to model the survival probability of *Hypericum cumulicola* plants with different number of reproductive structures. We will evaluate two model specifications that assume different type of survival change as function of number of reproductive structures.

For this demo you will need to download the script `Logistic_Regression2019` and the `hypericum_survival.txt` data file (both of which can be found in https://sciences.ucf.edu/biology/d4lab/methods-2/). You also need to have installed the packages `rethinking`, `rstan` and `bbmle`. We assume that you had already installed `stan`.

Part I. Preparing the data

Enter the following commands to load the *Hypericum cumulicola* dataset. This dataset contains two predictor variables (`height` and `rep_structures`) and a binary response variable (`survival`, where 0 = dead and 1 = alive). For this demo we will focus on the regression of number of reproductive structures vs. survival but feel free to explore more complex models featuring height as well.

```
orig_data <- read.table("Hypericum_survival.txt",header=T)
```

We will constrain the data to concentrate only on the fate of reproductive individuals:

```
dd  <- subset(orig_data, orig_data$rep_structures>0)
```

Part II. Model generation

The response variable *survival* consists of 0 and 1 values; therefore, a binary Generalized Linear Model can be used to analyze these data. A GLM consists of three parts (Zuur et al. 2015): (1) a likelihood distribution for the response, (2) a link function, and (3) a predictor function. For a logistic regression they are:

(1) The likelihood distribution is given by:

$$Survival_i \sim Bin(\pi_i)$$
$$E(survival_i) = \pi_i$$
$$\text{var}(survival_i) = \pi_i * (1 - \pi_i)$$

(2) The link function for the logistic model is the logit of π, the estimate for *p* is the probability of a variable as a function of x is given by:

Logit (π) = η

$$\pi_i = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

(3) The linear function η is a function of the covariates:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \eta = \beta_0+\beta_1 X$$

We will now build two Bayesian logistic regression models to evaluate the effect of number of reproductive structures on the probability of survival. We use the function `map2stan`. We call our first model `model1.b.stan`. This model assumes that survival changes as a linear function of the number of reproductive structures. We will examine the model and call its summary to inspect our estimates:

```
model1.b.stan <- map2stan(
alist(
survival ~ dbinom(1,p),
logit(p) <- a + b*rep_structures,
a ~ dnorm(0,10),
b ~ dnorm(0,10)
),
data = dd,chains=3,
start <- list(a= 0.0, b= 0.0)
)
```

We define diffuse prior distributions for both parameters. See the plot below to convince yourself that these are diffuse priors (in red). We also provide start values for this procedure. We use the function `precis` to call a summary of our estimates.

```
precis(model1.b.stan,digits=3)
```

	Mean	StdDev	lower 0.89	upper 0.89	n_eff	Rhat
a	1.058	0.115	0.875	1.239	605	1.001
b	-0.004	0.001	-0.005	-0.003	1119	1.002

Please see the code to plot in the script `Logistic_Regression2019 in stan.R`.

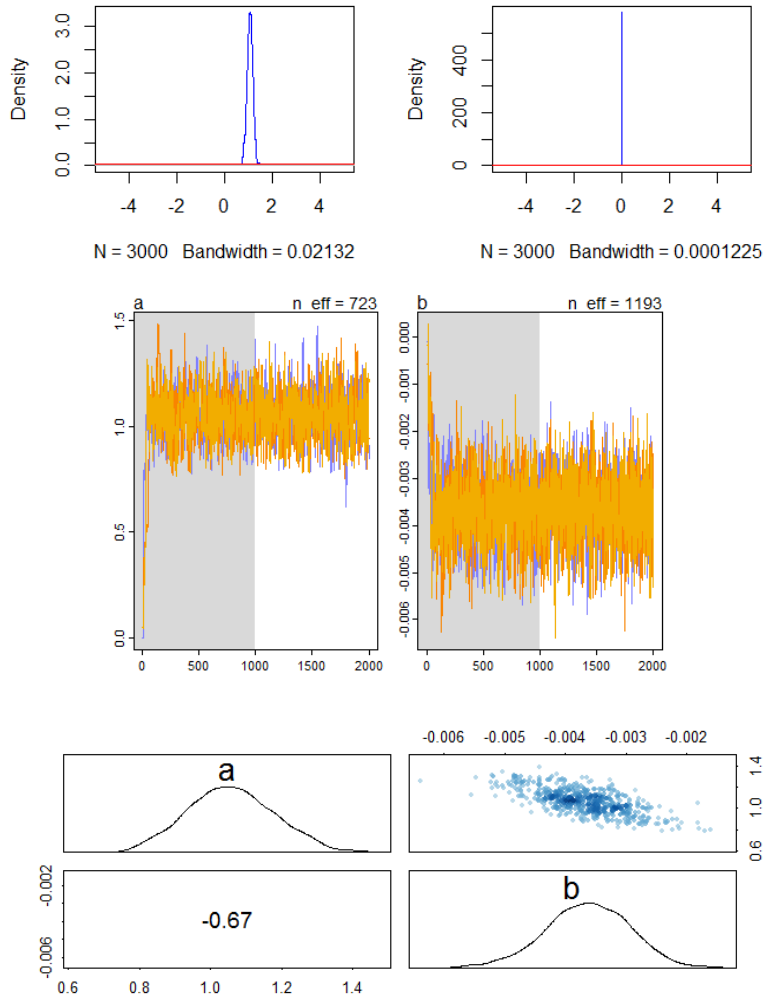


Figure 2. Plots of a model of survival as a linear function of number of fruits for *H. cumulicola* measured at Archbold Biological Station. On top the line in red indicates the distribution of the priors and the blue that of the posterior distributions of the parameters. In the middle are the sequence of the generating chains and at the bottom the distribution and the correlation of the parameters of the model.

Inspecting the shape of the sequence of the chains in the model doesn't reveal any problems with the generation of these estimates.

However, we are not satisfied with the distribution of the residuals in this model. The scarce information for plants with many reproductive structures has too much leverage and is pulling the model down, predicting extremely low survival values for plants with > 400 reproductive structures, even when plants with intermediate numbers of reproductive structures had survival probabilities around 0.5 (see figure below). This model predicts a probability of survival of 0.29 for plants with 544 reproductive structures. This is not

consistent with the observed survival proportions for plants with intermediate and large number of fruits.

Next, we try a model where the number of reproductive structures has been log-transformed to see if that improves the fit and assumptions. This model assumes that the change is now proportional to the number of fruits. We call our second model `model2.b.stan`.

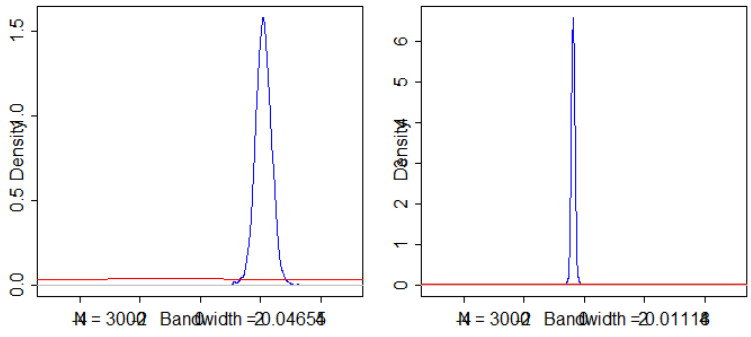
```
dd$rep <- log(dd$rep_structures)

model2.b.stan <- map2stan(
  alist(
    survival ~ dbinom(1,p),
    logit(p) <- a + b*rep,
    a ~ dnorm(0,10),
    b ~ dnorm(0,10)
  ),
data = dd,chains=3,
start <- list(a= 0.0, b= 0.0))
```

We also define diffuse priors for this model. In this occasion the default start values work fine. We use the function `precis` to call a summary of our estimates.

```
precis(model2.b.stan)
```

	Mean	StdDev	lower 0.89	upper 0.89	n_eff	Rhat
a	2.09	0.27	1.71	2.53	757	1
b	-0.37	0.06	-0.47	-0.28	739	1

After inspecting this model, we do not identify any areas of concern.



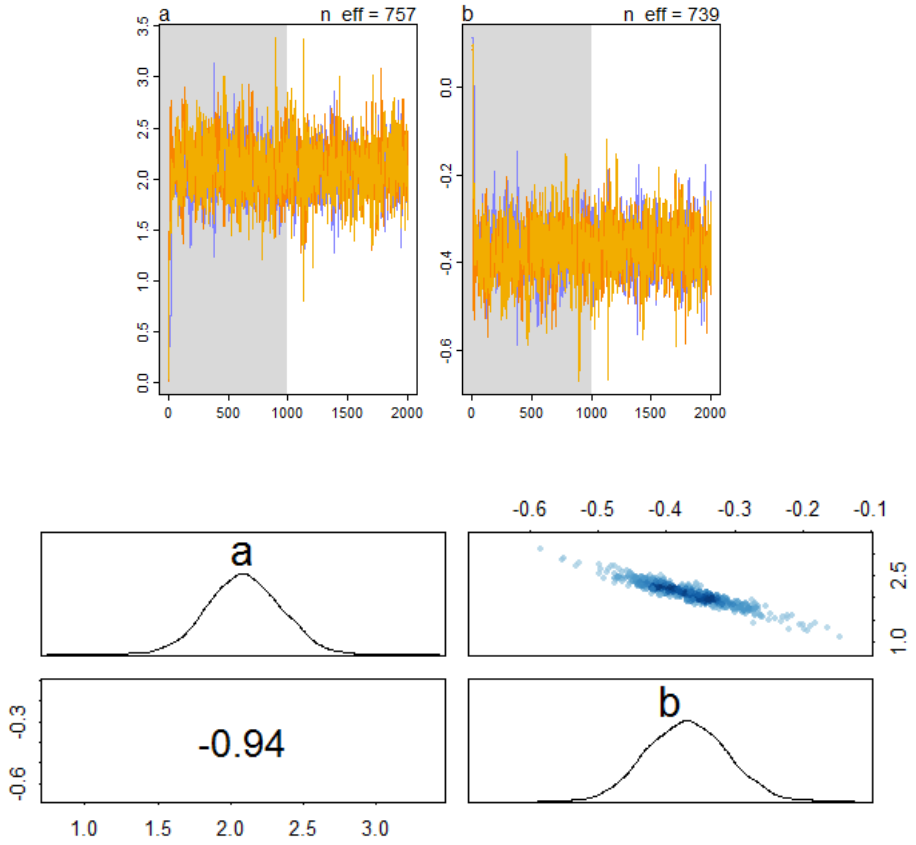


Figure 3. Plots of a model of survival as a logarithmic function of number of fruits for *H. cumulicola* measured at Archbold Biological Station. On top The line in red indicates the distribution of the priors and the blue that of the posterior distributions of the parameters. In the middle are the sequence of the generating chains and at the bottom the distribution and the correlation of the parameters of the model.

The new model has better distribution of residuals and allows us to conclude that the decline in survival with number of reproductive structures is larger for plants with fewer reproductive structures, and becomes less steep with increasing number of reproductive structures. We compare the models using the function `compare` and using information theory we confirm that the second model is most likely to explain these data.

	WAIC	pWAIC	dWAIC	weight	SE	dSE
model2.b.stan	702.5	2.0	0.0	0.76	18.38	NA
model1.b.stan	704.8	2.1	2.3	0.24	17.88	7.65

Model 2 has a weight of 0.76 vs a 0.24 for the first model.

We can visualize the differences between the two models even better by plotting them together (code included in the R script, model 1 in blue, model 2 in red). We incorporate 95% credibility intervals to depict the uncertainty on these parameters.

We challenge you to recalculate these models in the frequentist framework using the functions you learned in Methods I. We have code ready if you need it. In other demos we will present convincing arguments on why to use the Bayesian approaches. Completing this challenge will convince you that the Bayesian models in this demo, with diffuse priors, are completely commensurate with the frequentist procedures you learned before.

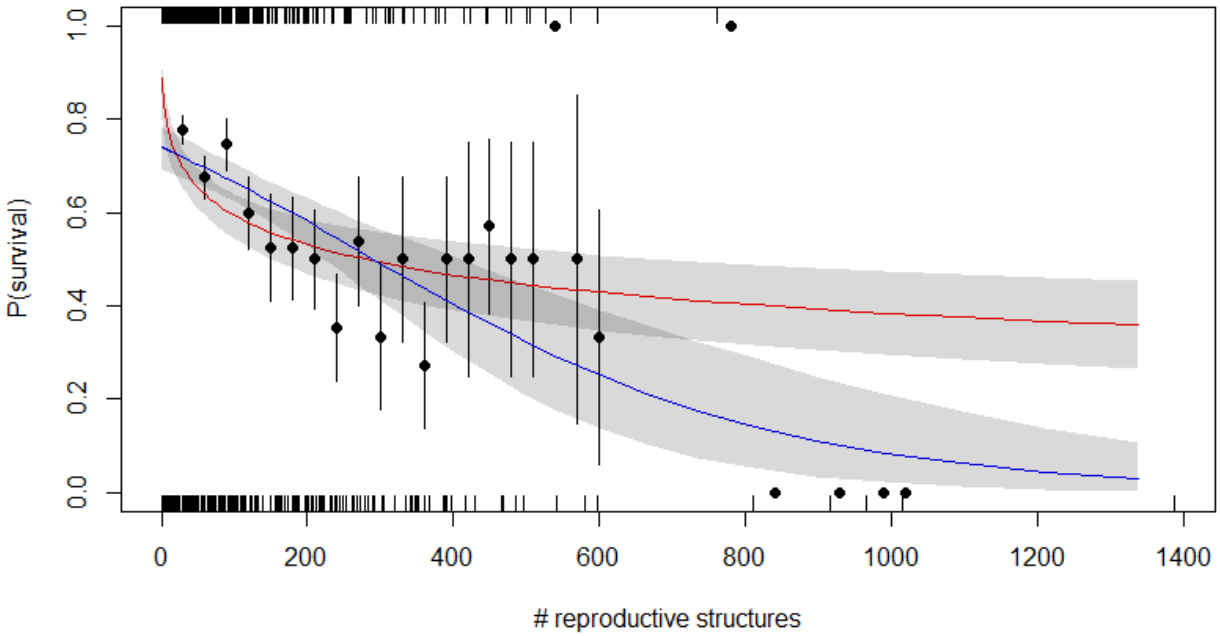


Figure 4: Survival probability as a function of number of reproductive structures. Closed circles are observed proportions for arbitrary bins with their 95% confidence intervals. The blue line is the average predictions for the first model. In red are the predictions for the second model. This is the model we prefer from the pair. Polygons are 95 credibility intervals.

NOTE: all the materials for this demo can be found at:

https://sciences.ucf.edu/biology/d4lab/methods-2/

References

McElreath, R.M. 2016. Statistical Rethinking: a Bayesian course with examples in R and Stan. Chapman and Hall.

Quintana-Ascencio, P. F., E. S. Menges, and C. Weekley. 2003. A fire-explicit population viability analysis of Hypericum cumulicola in Florida rosemary scrub. Conservation Biology, 17: 433-449.

Quintana-Ascencio, P.F. Koontz, S., Smith, V., David, A., Sclater, V. L. & E. S Menges. 2018. Predicting landscape-level distribution and abundance: Integrating demography, fire, elevation, and landscape habitat configuration. Journal of Ecology, 106: 2395-2408

Quintana-Ascencio, P.F. Koontz, S.M., Ochocki, B., Sclater, V. L., López-Borghesi, F., Li, H. & E. S Menges. 2019. Assessing the roles of seed bank, seed dispersal and historical disturbances for metapopulation persistence of a pyrogenic herb. Journal of Ecology, 107: 2760-2771.

Zuur, A, J.M. Hilbe and E N. Leno. 2015. A beginner's guide to GLM and GLMM with R. Highland Statistics, Ltd.

.