Applied Probabilistic Programming & Bayesian Machine Learning

Toy Datasets

MIT IAP 2017 Max W. Shen

All of these toy datasets were gathered from the Statistical Rethinking book by McElreath.

1 Tadpole Data

Keywords: Binomial, Clusters

Filename: data_tadpole.csv

Original Data Source: https://classes.warnercnr.colostate.edu/bayes-workshop/files/2013/04/Vonesh-Bolker-2004-Ecology.pdf

In Statistical Rethinking: Chapter 12.1

This dataset studied 48 different "ponds" each containing some density of tadpoles, represented as integers in the **density** column. The main measured quantity is the density of surviving tadpoles after some event, represented as integers in the **surv** column. It is sensible to model surv and density with a binomial distribution.

The **size** column represents the initial size of the tadpoles. Do small or big tadpoles have a better survival rate?

The **pred** column represents presence or absence of predators. Does the presence of predators have a negative effect on survival rate?

Is there an interaction between size and predation?

Summary via "str(dataset)" in R:

```
'data.frame': 48 obs. of 5 variables:
$ density : int 10 10 10 10 10 10 10 10 10 10 ...
$ pred : Factor w/ 2 levels "no","pred": 1 1 1 1 1 1 1 1 2 2 ...
$ size : Factor w/ 2 levels "big","small": 1 1 1 1 2 2 2 2 1 1 ...
$ surv : int 9 10 7 10 9 9 10 9 4 9 ...
$ propsurv: num 0.9 1 0.7 1 0.9 0.9 1 0.9 0.4 0.9 ...
```

1.1 Tadpole Data: Example of a Simple Model

$$\begin{aligned} surv_i &\sim binomial(density_i, p_i) \\ logit(p_i) &= \alpha_i \\ \alpha_i &\sim \text{weakly-informative}() \end{aligned} \tag{1}$$

2 Kline Data

Keywords: Poisson, Regression, GLM

Filename: data_tadpole.csv

In Statistical Rethinking: Chapter 10.2, 13.4

This dataset studied 10 island societies in Oceania to study incidents of tool development. The columns are self-explanatory. One could reason that the total number of tools excavated should depend somehow on the population of the societies as well as each societies' rate of contact with other societies.

Entire Dataset:

	culture	population	contact	total_tools
1	Malekula	1100	low	13
2	Tikopia	1500	low	22
3	Santa Cruz	3600	low	24
4	Yap	4791	high	43
5	Lau Fiji	7400	high	33
6	Trobriand	8000	high	19
7	Chuuk	9200	high	40
8	Manus	13000	low	28
9	Tonga	17500	high	55
10	Hawaii	275000	low	71

Guiding Questions: Can you predict total_tools using population and contact? Is population linearly to total_tools, or perhaps is log population linear to total_tools?

This is quite a small dataset - how sensitive are your results to your choice of priors?

A more interesting question: Is the impact of population and contact completely independent? If it is not, how can you include this in your model?

An optional additional matrix of distances between islands, in thousands of kilometers, can spice up your data analysis. The natural setup would be a Gaussian process regression.

	Ml	Ti	SC	Ya	Fi	Tr	Ch	Mn	To	Ha
Malekula	0.0	0.5	0.6	4.4	1.2	2.0	3.2	2.8	1.9	5.7
Tikopia	0.5	0.0	0.3	4.2	1.2	2.0	2.9	2.7	2.0	5.3
Santa Cruz	0.6	0.3	0.0	3.9	1.6	1.7	2.6	2.4	2.3	5.4
Yap	4.4	4.2	3.9	0.0	5.4	2.5	1.6	1.6	6.1	7.2
Lau Fiji	1.2	1.2	1.6	5.4	0.0	3.2	4.0	3.9	0.8	4.9
Trobriand	2.0	2.0	1.7	2.5	3.2	0.0	1.8	0.8	3.9	6.7
Chuuk	3.2	2.9	2.6	1.6	4.0	1.8	0.0	1.2	4.8	5.8
Manus	2.8	2.7	2.4	1.6	3.9	0.8	1.2	0.0	4.6	6.7
Tonga	1.9	2.0	2.3	6.1	0.8	3.9	4.8	4.6	0.0	5.0
Hawaii	5.7	5.3	5.4	7.2	4.9	6.7	5.8	6.7	5.0	0.0

2.1 Kline Data: Example of a Simple Model

$$\begin{aligned} tools_i &\sim poisson(\lambda_i) \\ \log \lambda_i &= c_w + w_{pop}\text{Pop}_i + w_{contact}\text{contact}_i \\ w, c_w &\sim \text{weakly-informative}() \end{aligned} \tag{2}$$

2.2 Kline Data: Example of a Model using Distances

$$\begin{aligned} tools_i &\sim poisson(\lambda_i) \\ \log \lambda_i &= \alpha + \gamma_i + w_{pop} \log \text{Pop}_i \\ \gamma &\sim MVNormal(\text{mean} = (0, ..., 0), K) \\ K_{ij} &= \eta^2 \exp(-\rho^2 D_{ij}^2) \\ \alpha, w_{pop}, \eta^2, \rho^2 &\sim \text{weakly-informative}() \end{aligned} \tag{3}$$

Here, D is the matrix of distances. K is the covariance matrix, which scales inversely with distance - the farther two societies are, the less similar we should expect them to be. η^2 acts as the maximum possible covariance between any two societies, and ρ^2 is a coefficient acting as the rate of decline with increasing distance.

3 WaffleDivorce Data

Keywords: Multivariate Regression, Clusters

Filename: data_waffledivorce.csv

In Statistical Rethinking: Chapter 5.1, 14.1

This dataset consists of a variety of statistics for each of the 50 U.S. states.

```
'data.frame':   50 obs. of  13 variables:
 $ Location         : Factor w/ 50 levels "Alabama","Alaska",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ Loc              : Factor w/ 50 levels "AK","AL","AR",..: 2 1 4 3 5 6 7 9 8 10 ...
 $ Population       : num  4.78 0.71 6.33 2.92 37.25 ...
 $ MedianAgeMarriage: num  25.3 25.2 25.8 24.3 26.8 25.7 27.6 26.6 29.7 26.4 ...
 $ Marriage         : num  20.2 26 20.3 26.4 19.1 23.5 17.1 23.1 17.7 17 ...
 $ Marriage.SE      : num  1.27 2.93 0.98 1.7 0.39 1.24 1.06 2.89 2.53 0.58 ...
 $ Divorce          : num  12.7 12.5 10.8 13.5 8 11.6 6.7 8.9 6.3 8.5 ...
 $ Divorce.SE       : num  0.79 2.05 0.74 1.22 0.24 0.94 0.77 1.39 1.89 0.32 ...
 $ WaffleHouses     : int  128 0 18 41 0 11 0 3 0 133 ...
 $ South            : int  1 0 0 1 0 0 0 0 0 1 ...
 $ Slaves1860       : int  435080 0 0 111115 0 0 0 1798 0 61745 ...
 $ Population1860   : int  964201 0 0 435450 379994 34277 460147 112216 75080 140424 ...
 $ PropSlaves1860   : num  0.45 0 0 0.26 0 0 0 0.016 0 0.44 ...
```

In the dataset, there is a spurious correlation between rate of marriage and rate of divorce across states. If you include information about the median age of marriage, it will go away.

The Marriage.SE and Divorce.SE are standard error values across states. Can you include this in your analysis?

3.1 WaffleDivorce Data: Example of a Simple Model

We suggest building a regression model to try to predict some quantity of interest.

4 Bangladesh Data

Keywords: Ordered Outcomes, Clusters

Filename: data_bangladesh.csv

In Statistical Rethinking: Chapter 12.6

This dataset comes from a 1988 fertility survey on women in Bangladesh, at a time when contraceptives were available but many families chose not to use them.

```
'data.frame':   1934 obs. of  7 variables:
 $ woman            : int  1 2 3 4 5 6 7 8 9 10 ...
 $ district         : int  1 1 1 1 1 1 1 1 1 1 ...
 $ use.contraception: int  0 0 0 0 0 0 0 0 0 0 ...
 $ living.children  : int  4 1 3 4 1 1 4 4 2 4 ...
 $ age.centered     : num  18.44 -5.56 1.44 8.44 -13.56 ...
 $ urban            : int  1 1 1 1 1 1 1 1 1 1 ...
```

The data consists of 1934 survey responses from women across 60 different districts. Usage of contraception is a binary variable, and the number of living children in the dataset ranges from 1 to 4. Age.centered is a centered (mean = 0) continuous variable, and lastly urban is also a binary variable.

What is the effect of contraception on the number of living children, controlling for urban/rural? Is there a varying effect across districts? What about age?

4.1 Bangladesh Data: Example of a Simple Model

Left as an exercise.

5 Trolley Data

Keywords: Ordered Outcomes, Clusters

Filename: data_trolley.csv

In Statistical Rethinking: Chapter 11.1

This dataset is from an experiment in philosophy regarding moral intuition. 331 unique individuals are included in the experiment, with 9930 rows.

```
'data.frame':   9930 obs. of  12 variables:
 $ case     : Factor w/ 30 levels "cfaqu","cfbur",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ response : int  4 3 4 3 3 3 5 4 4 4 ...
 $ order    : int  2 31 16 32 4 9 29 12 23 22 ...
 $ id       : Factor w/ 331 levels "96;434","96;445",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ age      : int  14 14 14 14 14 14 14 14 14 14 ...
 $ male     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ edu      : Factor w/ 8 levels "Bachelor's Degree",..: 6 6 6 6 6 6 6 6 6 6 ...
 $ action   : int  0 0 0 0 0 0 1 1 1 1 ...
 $ intention: int  0 0 0 1 1 1 0 0 0 0 ...
 $ contact  : int  1 1 1 1 1 1 0 0 0 0 ...
 $ story    : Factor w/ 12 levels "aqu","boa","box",..: 1 4 8 3 4 11 1 2 3 4 ...
 $ action2  : int  1 1 1 1 1 1 1 1 1 1 ...
```

Demographic information such as age, gender, and education are included in the data. The main crux of the experiment revolves around presenting "stories" to the participants, while varying the three following aspects:

(1) The action principle: Harm caused by action is morally worse than equivalent harm caused by omission.

(2) The intention principle: Harm intended as the means to a goal is morally worse than equivalent harm foreseen as the side effect of a goal.

(3) The contact principle: Using physical contact to cause harm to a victim is morally worse than causing equivalent harm to a victim without using physical contact

For example, in the traditional trolley problem, the participant has the choice of pulling a lever to save 5 people's lives from a train, but you will kill 1 person as a side effect. This story involves the action principle (since there is a choice between action and omission) but not the intention principle (since there's only one side effect) or contact principle (since the participant uses a lever). This story can be modified using a "fat man" setup to involve the contact principle, where you can push a fat man in front of a train, killing him but saving 5 people.

After seeing a story, the participant is asked to rate the moral permissibility of an action - in the trolley example, pulling the lever - on a scale from 1 (not permissible) to 7 (very permissible).

The experiment presented a variety of stories, each involving some subset of the 3 principles. The overall question of the experiment is: what effect do these principles have on moral permissibility?

5.1 Trolley Data: Example of a Simple Model

$$
\begin{aligned}
\text{response}_i &\sim \text{ordered\_logistic}(\eta, c) \\
\eta &= w_{action}\text{action}_i + w_{intent}\text{intent}_i + w_{contact}\text{contact}_i \\
w &\sim \text{weakly-informative}() \\
c &\sim \text{weakly-informative}()
\end{aligned}
\tag{4}
$$

6 Milk Data

Keywords: Missing Data, Collinearity, Clusters

Filename: data_milk.csv

In Statistical Rethinking: Chapter 5.2, 14.2

This dataset compares different species' milk. One hypothesis of interest is that animals with a greater neocortex percent (neocortex.perc) produce more milk with higher energy content (kcal.per.g).

Another question could be, do animals that have a high neocortex percent *relative to their body mass* produce milk with higher energy content?

In addition, some data is missing - can you impute them for a more complete analysis?

```
'data.frame':   29 obs. of  8 variables:
 $ clade         : Factor w/ 4 levels "Ape","New World Monkey",..: 4 4 4 4 4 2 2 2 2 2 ...
 $ species       : Factor w/ 29 levels "Alouatta seniculus",..: 11 8 9 10 16 1 2 6 27 28 ...
 $ kcal.per.g    : num  0.49 0.51 0.46 0.48 0.6 0.47 0.56 0.89 0.91 0.92 ...
 $ perc.fat      : num  16.6 19.3 14.1 14.9 27.3 ...
 $ perc.protein  : num  15.4 16.9 16.9 13.2 19.5 ...
 $ perc.lactose  : num  68 63.8 69 71.9 53.2 ...
 $ mass          : num  1.95 2.09 2.51 1.62 2.19 5.25 5.37 2.51 0.71 0.68 ...
 $ neocortex.perc: num  55.2 NA NA NA NA ...
```

6.1 Milk Data: Example of a Simple Model

Left as an exercise.

7 Bibliography and Additional Resources

All of these toy datasets were gathered from the Statistical Rethinking book by McElreath.

The Stan reference is available here: http://mc-stan.org/documentation/