Big Entropy & Generalized Linear Models

Max Entropy

Recall that entropy, for a discrete probability distribution, p, can be written as the average log-probability

$$H(p) = -\sum_i p_i \log p_i$$

The idea of maximum entropy is to use distributions that maximize the entropy given certain assumptions on the allowable distributions. In modelling, this amounts to defaulting to using distributions that maximize the entropy for our assumptions.

The main priniciple is

> The distribution that can happen the most ways is also the distribuiton with the biggest entropy. The distribution with the biggest entropy is the most conservative distribtuion obeying the constraints.

If we will only assume that a continuous variable will have a finite variance, the distribution that maximizes the entropy is the **Gaussian**.

If all we are willing to assume for a variable with only 2 possible outcomes that the probabilities of the events are constant over every draw, then the maxent distribution is the **Binomial**.

Generalzied Linear Models

Recall that we can achieve a linear model by assuming that the outcomes are distributed as Gaussians with the mean defined as a linear function over predictors. This gave us

$$\begin{aligned} y_i &\sim \text{Normal}(\mu_i, \sigma) \\ \mu_i &= \alpha + \beta x_i \end{aligned}$$

For a continuous y_i and far from some theoretical bounds, then this model is a maxent model.

Once more stringent assumptions are layered in, then we might start to struggle with using a linear model (get negative numbers for data that is constrained to be positive, real estimates for integer based outcomes, etc.).

If we know some basic facts about the data, it isn't hard to use maxent to choose better distributions to use. The idea then is to choose the maxent distribution for our assumption and replace a parameter describing the shape of this distribution by a linear model. This gives rise to generalized linear model. Below is a good example of a glm:

$$y_i \sim \text{Binomial}(n, p_i)$$
$$f(p_i) = \alpha + \beta(x_i - \bar{x})$$

A new addition to our model is the addition of a function f at the start of the linear model part. The reason behind this is that for most distributions we won't be lucky and there won't be a neat parameter that controls the mean. The mean will typically be a combination of multiple parameters. So, we choose to only model a single parameter. Usually there is a clear choice of which parameter to use. For the binomial, we typically know the number of trials, so we model the unknown probability, p_i. However, p_i is a probability is a probability, so it must be within the range $[0, 1]$. The linear model will never have a guarantee to only predict values in this level, so we use a **link** function to transform our linear model to be constrained within this range.

Exponential Family

A common set of probability distributions used for modelling is the exponential family. These have multiple properties that make them desirable, one of them being that each is a maxent distribution for some set of constraints. The 2 most commonly used distributions in this family are the Gaussian and the binomial distributions.

The **exponential** distribution is a distribution that is best for measurements of distance or duration, a measurement with respect to some point of reference. If the event is constant in time or space, then it will converge towards an exponential. It is the maxent distribution for all non-negative distributions with the same average displacement.

The **Gamma** distribution is also a key distribution for modelling displacement measurements. However, it can peak above 0. It is related to the exponential distribution, as for integers, n$, it can be seen to model the time after n exponential events. It is the maxent for distributions with the same mean and average logarithm. It is controlled by 2 parameters, however there are 3 common parametrizations which can complicate dealing with it.

The **Poisson** distribution is a count distribution. It can be seen to be a limiting distribution of the binomial if $n \to \infty$ and p is small. It is a maxent under similar scenarios as the Binomial, but will work better when there is no obvious maximum in the data.

Link Functions

A link function's job is to map the unconstrained space of outcomes from $\alpha+\beta x$ to a constrained space of a parameter. There are 2 common link functions: logit and log.

The **logit** link function maps a parameter that is defined as some probability mass, so that it is constrained to be in $[0, 1]$. It has the following form

$$\text{logit}(p_i) = \log \frac{p_i}{1 - p_i}$$

These are the **log-odds**; the odds of an event are the probability it happens divided by the probability it doesn't. This implies the following function for the probability

$$p_i = \frac{\exp[\alpha + \beta x_i]}{1 + \exp[\alpha + \beta x_i]}$$

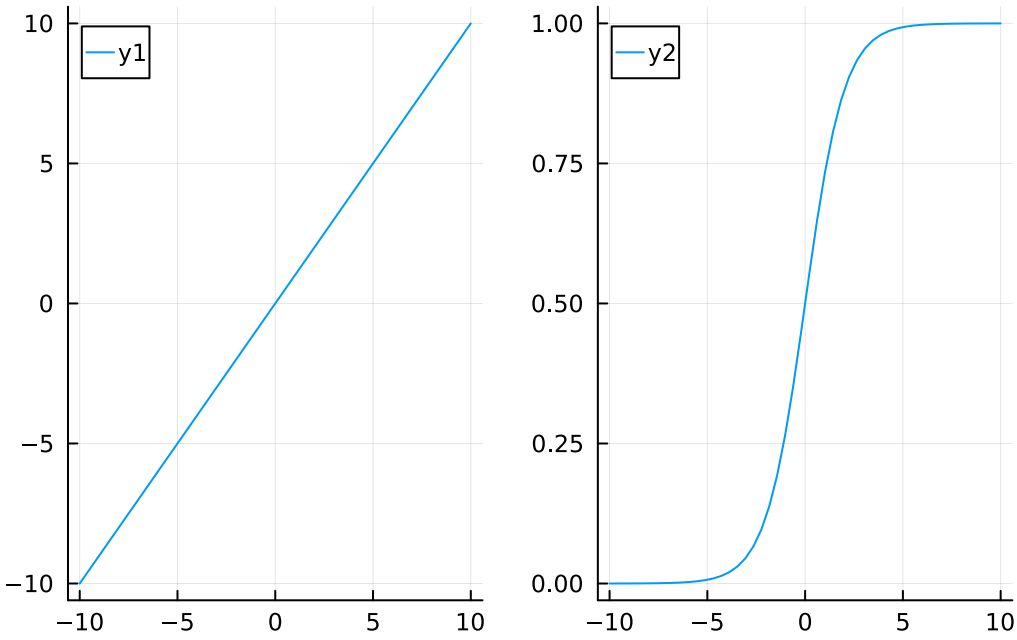
This is the logistic function.

```
using Plots

function logistic(y)
  1 / (1 + exp(-y))
end

alpha = 0
beta = 1
x = range(-10, 10, 50)
y = alpha .+ beta * x
p = logistic.(y)

plot(x, [y, p], layout = (1, 2))
```



Something that becomes clear here is that you can no longer get a clean estimate of the impact of a change in the input varaible. The impact of increasing x changes as you get farther from 0. This can be seen as an interaction with the parameter and tself.

The **log** link function maps the real line to the positive reals. An example might be modelling the standard deviation of a Gaussian:

$$\begin{aligned} y_i &\sim \text{Normal}(\mu, \sigma_i) \\ \log \sigma &= \alpha + \beta(x_i - \bar{x}) \end{aligned}$$

The inverse is the exponential function, so

$$\sigma_i = \exp[\alpha + \beta x_i]$$