Generalized linear madness

Richard Torkar

April 2021

We seek a measure of uncertainty that satisfies three criteria:

1. the measure should be continuous;
2. it should increase as the number of possible events increases; and
3. it should be additive.

The resulting unique measure of the uncertainty of a probability distribution p with probabilities p_i for each possible event i turns out to be just the average log-probability:

$$H(p) = -\sum_i p_i \text{log} p_i$$

This function is known as *information entropy*. When we want to choose between distributions (as part of our likelihood) we want to maximize the information entropy, i.e., maximum entropy,

> The distribution that can happen the most ways is also the distribution with the biggest information entropy. The distribution with the biggest entropy is the most conservative distribution that obeys its constraints.

In short, the distribution that can happen the greatest number of ways is the most plausible distribution. Call this distribution the *maximum entropy distribution*.

There are many distributions to pick from.[1] First, we have the Normal distribution (i.e., Gaussian). If all you know about a bunch of continuous values is its mean and variance, then Gaussian is the maximum entropy distribution. The Gaussian distribution has two parameters we want to estimate, i.e., the mean μ and the standard deviation σ (this you should know already by now).

Normal(μ, σ)

Second, we have the Binomial distribution (also called logistic regression when used), which is maximum entropy if only two things can happen (yes/no, true/false, alive/dead), and there's a constant chance p of each across n trials. The Binomial distribution has one parameter we want to estimate, i.e., p. However, since we are now using generalized linear models we also employ a link function to translate from log-odds to probability space. In the case of the Binomial it's the logit link function.

Binomial(n, p)

The Bernoulli distribution we use when $n = 1$.

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

If for some reason the Binomial cannot be used (because there is not a constant chance p) then we need to model that separately using instead a Beta-Binomial distribution where we estimate two parameters, p and θ. The Beta-Binomial uses a logit just as the Binomial.

Beta-Binomial(N, p, θ)

[1] S. A. Frank. The common patterns of nature. *Journal of Evolutionary Biology*, 22(8):1563–1585, 2009. doi: 10.1111/j.1420-9101.2009.01775.x

Third, the Exponential distribution is constrained to be zero or positive. It's often used when we're dealing with distance or duration. The Exponential distribution has maximum entropy among all non-negative continuous distributions with the same average displacement. The shape λ, is what we estimate. The Exponential distribution uses a log link.

Exponential(λ)

Fourth, the Gamma distribution is also constrained to be zero or positive. The difference is that Gamma can have a peak above zero. The Gamma distribution often uses a log link.

Gamma(λ, k)

Fifth, the Poisson distribution is also a count, like the Binomial's 0/1. If the number of trials n is very large (and usually unknown) and the probability of a success p is very small, then a binomial distribution converges to a Poisson distribution with an expected rate of events per unit time of $\lambda = np$. In short, when we have a count going from $0 \to \infty$ Poisson is a good first choice.

Poisson(λ)

But the Poisson is picky. We only estimate λ, which represents both the 'mean' and the variance so your outcome variable should have equal variance and mean. Very often that is not the case and instead you fall back to Gamma-Poisson, or Negative-Binomial as it is more commonly known as. The Poisson (and Gamma-Poisson) use a log link function. In the case of the Gamma-Poisson we model two parameters—we model variance separately, i.e., λ and the variance ϕ.

Gamma-Poisson(λ, ϕ) or more commonly Negative-Binomial(r, p)

Sixth, if you have categories you want to model (e.g., red, white, blue) then a Multinomial distribution is commonly used with a softmax link function.

a.k.a. multinomial logistic regression, which often is simply written as Categorical(p)

Seventh, and last, we have the case of *ordered* categorical outcomes (and predictors!), e.g., Likert scale.

The convention differs a lot but Ordered-Logit(ϕ, κ) is used in the course literature. While for predictors we set a prior using the Dirichlet(α) distribution.

LINK FUNCTIONS are something we must live with when working with generalized linear models. One must know when to use a logit and when to use a log (the two most common ones). Also, we must understand how to reverse them using inverse-logit and exponentiation.

So, we might now know what the point is with link functions, but there's more to it. The examples we see show that it is *very* important to do prior predictive checks before running your models. When our priors go through a link function we simply have a hard time seeing what they imply on the outcome. Repeat after me: Always be plotting your priors!

A common prior for β parameters is Normal$(0, 1)$. If one uses such a prior with a log link function then you are often on your way to mayhem...