Generalized Linear Models

36-290 – Statistical Research Methodology

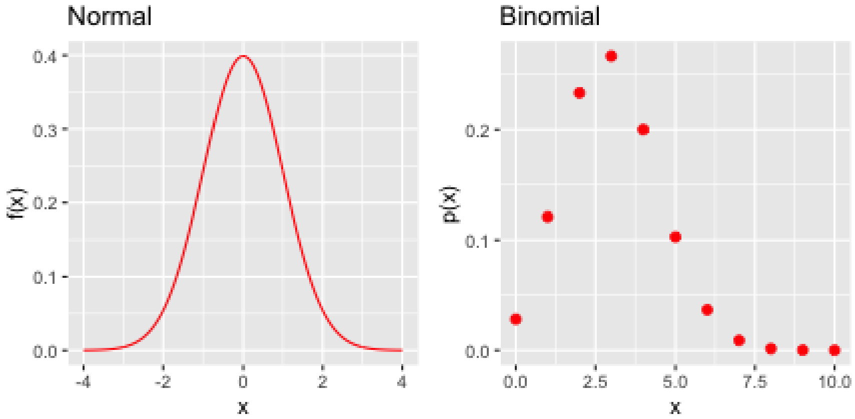
Week 5 Thursday – Fall 2021

Probability Distributions

A probability distribution is a mathematical function $f(y|\theta)$ where

- y may take on continuous values or discrete values;
- θ is a set of parameters governing the shape of the distribution (e.g., $\theta = \{\mu, \sigma^2\}$ for a normal);
- $f(y|\theta) \geq 0$ for all y; and
- $\sum_y f(y|\theta) = 1$ or $\int_y f(y|\theta) = 1$.

In practice, if y is continuously valued we often use f to denote the distribution (and call f a probability density function, or pdf), and if y is discretely valued we often use p (and call it a probability mass function, or pmf).



Probability Distributions and Regression

We are discussing distributions because in parameterized regression we make assumptions about how the response variable is distributed around the true regression line.

For instance, for linear regression, we assume that for every $\mathbf{x}$…

- the distribution governing the possible values of Y is a *normal* distribution;
- the mean of the normal distribution is $E[Y|\mathbf{x}] = \beta_0 + \sum \beta_i x_i$; and
- the variance of the normal distribution is σ^2, which is a constant (i.e., does not vary with x).

What if we cannot (or more to the point, should not) assume that $Y|\mathbf{x} \sim \mathcal{N}(\beta_0 + \sum \beta_i x_i, \sigma^2)$?

This is where we enter the realm of the *generalized linear model*.

Generalization: Example

In typical linear regression, the domain of $Y|\mathbf{x}$ is assumed to be $(-\infty, \infty)$, matching the domain of the mean μ of a normal distribution.

What, however, happens if we observe that the response variable is, e.g., discretely valued, with possible values 0, 1, 2, …?

The normal distribution isn't the correct one to assume here. What would be the right distribution to assume? In practice, there will be many possibilities and we might not know which one is right, but any assumption we make *should* be consistent with how the response is empirically distributed.

A suitable distribution in this case might be the *Poisson* distribution, which has a single parameter λ, which helpfully is the mean of the distribution (as well as the parameter which governs the distribution's shape).

So, when we apply generalized linear regression in this context, we would identify the distribution family as Poisson.

But there's another step in generalization…

Generalization: Example

Let's keep things in the realm of one predictor. The linear function is

$$\beta_0 + \beta_1 x \, .$$

As noted above, the range of this function is $(-\infty,\infty)$. But in our Poisson regression example, we know that the mean λ cannot be negative, *so we need to transform the linear function* so that the domain of the transformed function is, in this case, $[0,\infty)$. In other words, we need to find a function that maps the line defined above to $\lambda \vert x > 0$. (Remember, we are modeling the mean...while the data are discrete, the mean is continuous as a function of x.)

There is usually no unique transformation, but rather ones that are conventionally used. For this case:

$$g(\lambda \vert x) = \log(\lambda \vert x) = \beta_0 + \beta_1 x \, ,$$

or

$$\lambda \vert x = e^{\beta_0 + \beta_1 x} \, .$$

$g(\cdot)$ is dubbed the *link* function.

Note that with a GLM, we still can perform statistical inference, even with the transformation, because we can still determine how the predicted response varies as a function of x.

Finding the Optimal Coefficient Values

Once the link function is set, we use numerical optimization to estimate β_0 and β_1 via maximization of the likelihood function:

$$\mathcal{L} = \prod_{i=1}^{n_{\rm train}} p(Y_i \vert \lambda_i = e^{\beta_0+\beta_1x_i}) \,,$$

where $n_{\rm train}$ is the number of (training set) data. (Note: we don't use numerical optimization in typical multiple linear regression because the optimal values are computed via formula. That's why you are only see the likelihood function now. Also note that what really gets computed is $L = \log \mathcal{L}$, the "log-likelihood.")

Leaving many details under the rug: the maximum is the point at which the derivative of the likelihood function is zero. (You don't need to check the second derivative: wherever the derivative equals zero, it's a maximum value, not a minimum value.)

Numerical optimization can be slow. So, for instance, Poisson regression can take significantly longer to run than normal regression when datasets are large.

Generalization: Exercises

What family might be appropriate for…

1. $Y|\mathbf{x}$ continuous, but bounded between 0 and 1?
2. $Y|\mathbf{x}$ continuous, but bounded between 0 and ∞?
3. $Y|\mathbf{x}$ discrete, but can only take on the values 0 and 1?
4. $Y|\mathbf{x}$ discrete, but can only take on the values 0, 1, 2, …, n?

Case 3 is special, as you will soon see.