# Generalized Linear Models

2020-04-07

In this set of notes, you will learn about the broader class of models that binomial logistic regression belongs to, generalized linear models.

## Generalized Linear Models

Generalized linear models, introduced by Nelder & Wedderburn (1972), allow the modeling of data which have non-normal conditional distributions of the residuals. These models consist of three components:

- A **linear function describing the structure between the predictors**, $X_1, X_2, \ldots, X_k$. This structure can be additive (main-effects) or multiplicative (interactions) in nature, and can include transformations of the predictors, polynomial terms, dummy coded predictors, etc. For example,

$$\eta_i = \beta_0 + \beta_1(X_{i1}) + \beta_2(X_{i2}) + \ldots + \beta_k(X_{ik})$$

- A **link function** which transforms the mean of the outcome variable to the specified linear set of predictors. This function needs to be mathematically smooth (no gaps or jumps) and invertible (we can backtransform). For example if our link function is $g(\cdot)$, then

$$\begin{split}
g(\mu_i) &= \eta_i \\
&= \beta_0 + \beta_1(X_{i1}) + \beta_2(X_{i2}) + \ldots + \beta_k(X_{ik})
\end{split}$$

Since the link function is required to be invertible, we can also write the inverse of $g(\cdot)$,

$$\begin{split}
\mu_i &= g^{-1}(\eta_i) \\
&= g^{-1}\bigg(\beta_0 + \beta_1(X_{i1}) + \beta_2(X_{i2}) + \ldots + \beta_k(X_{ik})\bigg)
\end{split}$$

The inverse link function is, for obvious reasons, sometimes referred to as the *mean function*.

- A **random component** specifying the conditional distribution of the response variable, $Y_i$, given the predictors in the model. This distribution is either a member of the *exponential family* of distributions or from the *multivariate exponential family* of distributions.

In our binomial logistic regression example from the last set of notes, the first model we fitted was,

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1(\text{ACT Score}_i)$$

In this example, we used an additive linear function ($\eta_i = \beta_0 + \beta_1(\text{ACT Score}_i)$). We related this to the the mean (proportion of 1's in this example) via the logistic transformation ($\Lambda$; our link function). Lastly, we posited that the binomial distribution (a member of the exponential family of distributions) could be used to describe each of the conditional distributions of $Y_i$ at each ACT score.

The mean function can be computed by inverting the link function,

$$\begin{aligned}
\mu_i = \pi_i &= g^{-1}\bigg[\ln\left(\frac{\pi_i}{1-\pi_i}\right)\bigg] \\
&= g^{-1}\bigg[\beta_0 + \beta_1(\text{ACT Score}_i)\bigg] \\
&= \frac{1}{1+e^{-\big[\beta_0+\beta_1(\text{ACT Score}_i)\big]}}
\end{aligned}$$

# Exponential Family of Distributions

The conditional distribution of $Y$ need to be specified as a distribution which is a member of the exponential family of distributions. The general mathematical form of an exponential distribution is:

$$f(y_i|\theta,\phi) = \exp\left[\frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i,\phi)\right],$$

where:

- $y_i$ are elements of a random variable $Y$ which we can classically write as $y_i = \mu + \epsilon_i$
- $\theta_i$ is called the *canonical/natural parameter*, which is a function of the mean ($\mu$) of the distribution;
- $b(\theta)$ is a function of the canonical parameter, which is also a function of the mean (since $\theta$ is a function of the mean);
- $a(\phi)$ is a function of the *dispersion/scale parameter*, $\phi$, which plays a role in determining the variance of $y$; and
- $c(y_i,\phi)$ is a function of the observations $y$ and the dispersion parameter $\phi$.

It turns out that many of the common probability distributions used in statistics are actually members of the exponential family of distributions. For example the binomial distribution, the Poisson distribution, and even the normal (Gaussian) distribution are all members of the exponential family of distributions.

## Normal/Gaussian Distribution

Consider the equation for the probability density based on the normal probability distribution,

$$f(y_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y_i-\mu)^2}{2\sigma^2}\right]$$

Because the normal distribution is a member of the exponential family of distributions, we can re-express the probability density in a form consistent with the general form presented earlier for the exponential distribution. Doing so, we find,

$$f(y_i|\mu,\sigma^2) = \exp\left[\frac{y_i\mu - \mu^2/2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right]$$

where,

- $\theta = \mu$,
- $b(\theta) = \mu^2/2$,
- $a(\phi) = \sigma^2$, and
- $c(y, \phi) = -\frac{1}{2}\left[\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\right]$

The canonical parameter, $\theta$, also referred to as the **link function** for the normal probability distribution is $\mu$. Since the link function is equal to $\mu_i$ (no transformations needed to obtain $\mu_i$), this link is referred to as the *identity link*.

The variance function of a generalized linear model, which describes how the conditional variance is computed, is the second derivative of $b(\theta)$ (w.r.t. to $\theta$) multiplied by $a(\phi)$. For the normal distribution the variance function is simply computed as $\sigma_i^2$. Thus, the conditional variance is constant, it is $\sigma^2$ across all values of $i$.

## Binomial Distribution

The binomial distribution is another member of the exponential family of distributions. The common mathematic form for the binomial probability distribution is,

$$f(y_i|\pi_i) = \binom{n_i}{y_i}\pi_i^{y_i}(1- \ pi_i)^{n_i - y_i}$$

where $\pi_i = \frac{\mu_i}{n_i}$.

Again, since it is a member of the exponential family of distributions we can re-express this using the form presented earlier. This is,

$$f(y_i|\pi_i) = \exp\left[y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \ln(1-\pi_i) + \ln\binom{n_i}{y_i}\right]$$

where,

- $\theta = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$,
- $b(\theta) = -n_i \ln(1-\pi_i)$,
- $a(\phi) = 1$, and
- $c(y, \phi) = \ln\binom{n_i}{y_i}$

The link function, $\theta$, is the logistic transformation (or logit). The variance function is $n_i\pi_i(1-\pi_i)$. This means that the conditional variance in a binomial distribution is a function of the conditional means (or probabilitieis).

## Poisson Distribution

The Poisson distribution is another member of the exponential family of distributions. The common mathematic form for the Poisson probability distribution is,

$$f(y_i|\mu_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y!}$$

Again, since the Poisson distribution is a member of the exponential family of distributions we can re-express this using the form presented earlier as,

$$f(y_i|\mu_i) = \exp\left[y_i \ln(\mu_i) - \mu_i - \ln(y_i!)\right]$$

where,

- $\theta = \ln(\mu_i)$,
- $b(\theta) = \mu_i$,
- $a(\phi) = 1$, and
- $c(y, \phi) = -\ln(y_i!)$

The link function, $\theta$, is the log transformation. The variance function is $e^{\theta_i} = \mu_i$. Thus the Poisson distribution has a conditional variance that is equal to its conditional mean.

In general, in the exponential family of distributions, the conditional variance will be a function of the conditional mean. The only exception to this is the normal distribution which has a constant variance (it is not a function of the conditional mean). In regression this implies that the variation in $Y$ changes depending on the $\hat{Y}$ value (which itself is dependent on the predictors). When we posit different error distributions on the random components, we need to be sure that the conditional variance follows the pattern described by the particular distribution. For example, a Poisson distribution would imply that the conditional variance and mean are the same, so if we observed a positive association in the data between $X$ and $Y$ ($\hat{Y}$ is getting larger for higher values of $X$), we would also need to see an increase in variance at each subsequent $X$ value as well.

## Take Home for Applied Researchers

For applied researchers, what is important in all of this is that the conditional distribution of the outcome variable and the link function relating the linear set of predictors to the mean of the outcome need to be specified. We specify this in the arguments of the `glm()` function. The syntax to fit a generalized linear model follows the this form:

$$\texttt{glm(y} \sim 1 + \texttt{x}, \texttt{data} = \text{dataframe}, \texttt{family} = \texttt{distribution(link} = \texttt{"link\_function"})$$

The `family=` argument is where we specify the conditional distribution of the outome variable. For example, when fitting the binomial logisitc regression model we used `family = binomial()`. To fit the model with a normal distribution we would use `family = gaussian()`.

We also specify the link function by using `link=` in the name of the distribution function we used in `family=`. For the logit link, we used `family = binomial(link = "logit")`. For the normal distribution we would use the identity link, `family = gaussian(link = "identity")`.

The table below gives some of the common modeling scenarios, conditional distributions and link functions for several situations faced by applied researchers.

| **Typical Use** | **Distribution** | **Link Function** |
|---|---|---|
| Modeling counts/proportions of successes in $k = 2$ categories | Binomial | Logit |
| Modeling average Y in continuous data | Normal | Identity function |
| Modeling count/proportion of occurrences in fixed amount of time/space | Poisson | Log |
| Modeling counts/proportions of successes in $k > 2$ categories | Multinomial | Logit<br>(multiple equations) |
| Modeling time to event occurrence | Gamma | Inverse |

In R, you can use `help(family)` to obtain more information on how to specify each of these distributions. Running `help(binomial)` or `help()` on any of the distribution names will also bring up additional information about alternative link functions, etc.

## References

Nelder, J. A., & Wedderburn, T. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, *135*, 370–384.