Understanding Generalized Linear Models

Let's revisit the general linear model to understand how it differs from a generalized linear model.

The general linear model is mathematically represented as $Y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i$ where: Y_i is the i^{th} observed value for the response variable. x_{1i} through x_{ki} are the i^{th} values for the predictor variables x_1 through x_k. β_1 through β_k are the regression coefficients for the corresponding predictor variables x_1 through x_k, respectively, and ε_i is the i^{th} value of random errors.

You assume that the random errors, ε_i, are independently and identically distributed following a normal distribution with a mean of zero and a variance of σ^2. Since the mean of the error term is equal to zero and the variance is sigma squared, it follows that the expected value of the response is equal to the linear combination of the values for the predictor variables x_1 through x_k multiplied by their regression coefficients.

A generalized linear model extends the general linear model in three ways: First, no assumption of normality is required. The distribution of the observations can come from the family of exponential distributions, including the normal, gamma, Poisson, binomial, and negative binomial distributions. By allowing for distributions from the exponential family, we can model data in which the response or outcome variable is either a discrete (for example, binomial or Poisson) random variable or a continuous (normal or gamma) random variable.

Second, the variance of the response variable can be expressed as a function of its mean. In the case of the normal distribution, the relationship can be expressed as $\sigma^2 = \mu^0 \bullet \sigma^2$. Click the Information button for additional details on calculating the variance of the response variable.

Third, a link function, *g*, is used to fit the linear model. Note that we are not modeling the individual values of the response variable, Y, but the expected value or mean of Y. This expected value may or may not be able to be modeled by a linear combination of our predictor variables (Xβ). However, a function of the expected value might. This function is referred to as the link function.

The link function is any monotonic differentiable function g that relates the mean of y to a linear combination of predictor variables Xβ. Xβ is fit to a link function of E(y) suggested by the distribution of the observations.

In general, if we have a distribution for our outcome data and this distribution has certain restrictions on the parameter values, we create the link function to ensure that these restrictions are upheld. The link function must be monotonic, but it does not have to be an identity function, as is the case for general linear models. These three extensions of the generalized linear model over the general linear model indicate that it is applicable to a wider range of data analysis problems.

Close