Generalized Linear Models

October 22, 2017

THE UNIVERSITY OF MEMPHIS

Generalized Linear Models (GLMs)

These models are useful when 1) variance is not constant or 2) errors are not normally distributed.

Use GLMs when the response variable is

- count data expressed as proportions - logistic regression
- count data not proportions - log-linear models
- binary response variables
- data on time to death where the variance increases faster than linearly with the mean

Components of GLMs

A generalized linear model has three important components:

- the error structure
- the linear predictor
- the link function

Error Structure

Many kinds of errors are not normally distributed

- errors that are strongly skewed
- errors that have high kurtosis
- errors that are strictly bounded (as in proportions)
- errors that cannot lead to negative values (such as counts)

Without GLMs our only tools in these cases were transformations or non-parametric methods

Error Structure

A GLM allows the specification of different error distributions:

- Poisson errors, useful with count data
- binomial errors, useful with data on proportions
- gamma errors, useful with data showing a constant coefficient of variation
- exponential errors, useful with survival data

Linear Predictor

The linear predictor η is a linear sum of the effects of one or more explanatory variables

$$\eta_i = \sum_{j=1}^{p} X_{ij}\beta_j$$

where X_{ij} is the jth covariate variable measured on the ith individual and the β's are the unknown parameters.
In multiple regression or ANOVA the linear predictor, η is assumed to be the mean of the response. In GLMs that is not necessarily the case.

Link Function

If μ is the mean of the response variable, Y, the link function relates the mean to the linear predictor:

$$\eta = g(\mu)$$

In the regression model and the ANOVA model, the link function is the identity link.

How do we pick the link function?

Example

Suppose we have a random sample from a Poisson variable with pdf

$$P(Y = y) = \frac{\mu^y e^{-\mu}}{y!}$$

With mean $\mu > 0$.

Link Function

Suppose we have linear predictor

$$\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

In this case, η can be any real number depending on the values of X and parameters β. We can't use an identity function because μ must be positive. We need to choose a link function so that it maps real numbers into $(0, \infty)$. For example if we let

$$\mu = e^{\eta}$$

then μ will be greater than zero. The inverse transformation gives us the link function

$$\eta = \log \mu$$

called the log-linear model.

Binary Data

Binary data arises when there are only two possible outcomes we can label 0 or 1. Let $\pi = P(Y = 1)$ so that $1 - \pi = P(Y = 0)$. Assume as before we have a linear predictor $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ which can be any real number. The mean of the binary response variable is $\pi \in (0, 1)$. We need a transformation that takes the real number line into the interval $(0, 1)$. Some commonly used link functions are:

- logit (logistic) $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$
- Probit $g(\pi) = \Phi^{-1}(\pi)$ where $\Phi(x)$ is the cumulative distribution function of a standard normal random variable
- log-log $g(\pi) = \log(-\log(\pi))$

Canonical Link Functions

Certain special link functions can be derived from the form of the distribution of the assumed error structure. These are called canonical links. The canonical links for some distributions are given in the chart below. In R these links are the default when you select an error structure

Error	Canonical Link
normal	identity
poisson	log
binomial	logit
Gamma	reciprocal

overdispersion - Misspecified Error Structurr

If we choose a Poisson model for count data we are specifying not only the mean but the variance as well. For the Poisson distribution the mean equals the variance. Most of the time with real data the variance is greater (overdisperson) or less (underdispersion) than that of the chosen model.

Misspecified Link Function

It is always possible that a different link function may give a better fit to a dataset. An alternative to changing the link function is to transform the values of the response variable. The aim of the transformation is to

- constant error variance
- approximately normal errors
- additivity
- a linear relationship between the response variables and explanatory variables
- straightforward scientific interpretation

Deviance: Measuring the Goodness of Fit of a GLM

For GLMs it is customary to use the deviance to assess the GOF. The deviance is defined to be -2 times the difference in log-likelihood between the fitted model and a saturated model (i.e. a model that fits the data perfectly. Minimizing the deviance is equivalent to maximizing the likelihood. A model is better if it has a smaller deviance.