GENERALIZED LINEAR MODELING

H. Geys
Limburgs Universitair Centrum, Hasselt University, Belgium

Keywords: Exponential family, random component, systematic component, quasi-likelihood, generalized estimating equations

Contents

Summary

Starting from linear regression and the exponential family, generalized linear modeling is introduced. Standard estimation procedures, based on the likelihood, are introduced, as well as alternatives such as quasi-likelihood. The extension to generalized estimating equations is discussed briefly.

1. Introduction

For several decades normal linear models of the form

$$Y = X\beta + \varepsilon \qquad (1)$$

where ε is assumed to be normally distributed with mean zero and variance σ^2 have formed the basis of most analyses on continuous data. Recent advances in statistical theory and computer software allow us to use methods analogous to those developed for linear models in the following situations.

1. The response variables have distributions other than the normal distribution; they may even be categorical rather than continuous.
2. The relationship between the response and explanatory variables need not be of the simple linear form in (1).

One of these advances has been the recognition that many of the nice properties of the normal distribution are shared by a wider class of distributions called the *exponential family of distributions*. We will come back to this family later in Section 2. A second

advance is the extension of the numerical methods for estimating parameters, from linear combinations such as $X\beta$ to functions of linear combinations $g(X\beta)$.

In this chapter we focus on "Generalized Linear Models" (GLM) which refer to a family of regression models described by Nelder and Wedderburn, which provide a unified approach to many of the most common statistical approaches. To summarize the basic ideas, the generalized linear model differs from the general linear model (of which, for example, multiple regression is a special case) in two major respects: First, the distribution of the dependent or response variable can be (explicitly) non-normal, and does not have to be continuous, i.e., it can be binomial, multinomial, or ordinal multinomial (i.e., contain information on ranks only); second, the dependent variable values are predicted from a linear combination of predictor variables, which are "connected" to the dependent variable via a link function. The general linear model for a single dependent variable can be considered a special case of the generalized linear model: In the general linear model the dependent variable values are expected to follow the normal distribution, and the link function is a simple identity function (i.e., the linear combination of values for the predictor variables is not transformed).

2. A Corner Stone: the Exponential Family of Distributions

The exponential family of distributions forms a corner stone in the development of generalized linear models. The probability function for the canonical form of the exponential family for the *i*th observation is:

$$f(y_i;\theta_i,\phi,w_i) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi,w_i)} + c(y_i,\phi,w_i)\right], \quad (i=1,\ldots,N) \tag{2}$$

where the parameter of interest for the *i*th observation is θ_i, ϕ is a scale or dispersion parameter and w_i is a weighting constant. The function $b(\cdot)$ satisfies following properties:

1. $b'(\theta_i) = \mathrm{E}(Y_i) = \mu_i$,
2. $b''(\theta)_i = \mathrm{Var}(Y_i) = v(\mu_i)$.

The expression $b'(\theta_i) = \mathrm{E}(Y_i) = \mu_i$ implies the *natural* or *canonical link* for that distribution so that $\theta_i = g(\mu_i)$. Let us give some well-known examples. When data are normally distributed, the density of the observations can be written as

$$f(y_i;\mu_i,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-(y_i-\mu_i)^2/2\sigma^2)$$
$$= \exp\left\{\left[y_i\mu_i - \frac{\mu_i^2}{2}\right]\frac{1}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\}.$$

Therefore, $b(\theta_i) = \theta_i^2 / 2$ so that $\mu_i = b'(\theta_i) = \theta_i$ and $\phi = \sigma^2$. Hence, the canonical link for the normal distribution is the identity link.

In case of a Bernoulli logistic model the density of the observations is given by

$$f(y_i, \mu_i) = \mu_i^{y_i}(1-\mu_i)^{1-y_i}$$
$$= \exp\left\{ y_i \ln\left(\frac{\mu_i}{1-\mu_i}\right) + \ln(1-\mu_i) \right\}.$$

Therefore, $b(\theta) = \ln\{1+\exp(\theta)\}$ so that $g(\mu_i) = \ln\left(\frac{\mu_i}{1-\mu_i}\right)$, $v(\mu_i) = \mu_i(1-\mu_i)$ and $\phi = 1$. Hence, the canonical link is the logit link.

In summary, for the most common applications, the canonical links are:

Distribution	$\mu = b'(\theta)$	Canonical link, $g(\mu)$
Normal	θ	Identity
Binomial	$\frac{\exp(\theta)}{1-\exp(\theta)}$	logit
Poisson	$\exp(\theta)$	Log

One could use any differentiable link function with any error distribution. However, problems may arise in the fitting of the model by Newton-Raphson iteration. For example, one could use the log link with a binomial error distribution in lieu of the usual logistic regression model with logit link. However, this model does not ensure that estimated probabilities $\pi(x) = \mu(x)$ are bounded by (0,1), and the iterative solution of the coefficients and the estimated information may fail unless a method for constrained optimization is used to fit the model.

3. Generalized Linear Modeling

The generalized linear model generalizes the basic normal linear model in two ways. First, the distribution of the observations can be any member of the exponential family. In many practical applications for example, the outcomes will be binary or the number of successes out of a certain number of trials. In that case one might assume a binomial distribution for the random component. In other situations one might encounter nonnegative counts. We could then assume a Poisson distribution for the random component. If observations are continuous, such as a person's height or weight, a normal random component is often assumed. All of the above examples belong to the exponential family of distributions. Second, the link between the expectation μ and the linear predictor can be any monotone, differentiable function. Thus a generalized linear

model has three components.

1. The ***random component*** identifies the response variable *Y* and assumes a probability distribution for it that belongs to the exponential family of distributions. The random component specification implies a specific relationship between the mean and the variance. Sometimes it is also necessary to incorporate a scale or dispersion factor into the model, designated as ϕ.

2. The ***systematic component*** specifies the explanatory variables used as predictors on the right hand side of the model equation:

   $$\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p.$$

3. The ***link*** describes the functional relationship between the systematic component and the expected value or mean, $\mu = \mathrm{E}(Y)$, of the random component. For a general link function $g(\cdot)$, we have:

   $$\eta = g(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p.$$

The best known link function is the so-called *identity link* $g(\mu) = \mu$, which specifies a linear model for the mean response:

$$\mu = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p.$$

This is the form used in general linear models for continuous responses. Other links permit the mean to be nonlinearly related to the predictors. Specifically, for the binomial distribution we have that $0 < \mu < 1$. Hence, a link should satisfy the condition that it maps the interval (0,1) onto the entire real line. The best known link functions for this situation are:

a. the logit link

   $$\eta = \log(\mu/(1-\mu)).$$

A GLM that uses the logit link is called a *logit model* and constitutes the basis of logistic regression analyses.

b. the probit link

   $$\eta = \Phi^{-1}(\mu).$$

A GLM that uses the probit link is called a *probit model* and forms the basis of probit regression analyses.

Similarly, when we are dealing with nonnegative counts and the distribution is Poisson we have that $\mu > 0$. Therefore a suitable link function in this situation is the log link

$\eta = \log(\mu)$
A GLM that uses the log link is called a *loglinear model*.

As pointed out in Section 2 each distribution for the random component has its own special function of the mean that is called its natural parameter. For the normal distribution it is the mean itself, for the Poisson it is the log of the mean, and for the binomial it is the logit of the mean. The accompanying link function, i.e., the identity, the log and the logit functions, respectively, are called the canonical links. These canonical links are the most commonly used. However, alternative link functions may also be applied.

-
-
-

Bibliography

Agresti, A. (1990) *Categorical Data Analysis*, New York: John Wiley. [A standard text on categorical data, including generalized linear model.]

Agresti, A. (1996) *An Introduction to Categorical Data Analysis*, New York: John Wiley. [Comprehensive text on categorical data.]

Diggle, P.J., Heagerty, P., Liang, K.-Y., and Zeger, S.L. (2002) *Analysis of Longitudinal Data* (2nd ed). Oxford Science Publications. Oxford: Clarendon Press. [Text on repeated measures, including generalized estimating equations.]

Dobson, A.J. (1990) *An introduction to Generalized Linear Models*, London: Chapman and Hall. [Comprehensive introduction to generalized linear models.]

Lachin, J.M. (2000) *Biostatistical Methods, the assessment of relative risks*. New York: John Wiley. [Focus on categorical data for biostatistics and epidemiology.]

Liang, K.Y. and Zeger,S. (1986) Longitudinal Data Analysis using Generalized Linear Models. *Biometrika*, **73**, 13-22. [Seminal paper in which generalized estimating equations were introduced.]

Lindsey, J.K. (1997) *Applying Generalized Linear Models*. New York: Springer-Verlag. [Text on generalized linear models.]

Lipsitz, S.R., Laird, N.M. and Harrington, D.P. (1991) Generalized Estimating Equations for Correlated Binary Data: Using the Odds Ratio as a Measure of Association. *Biometrika*, **78**, 153-160. [Key paper in which generalized estimating equations with odds ratios are introduced.]

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*. New York: Chapman and Hall. [Classic text on generalized linear models.]

Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized linear models. *Journal of the Royal Statistical Society, Series B*, **19**, 92-100. [Key paper in which the concept of generalized linear models was introduced.]

Biographical Sketch

Helena Geys holds a B.S. degree in mathematics from the University of Antwerp and Master of Science and Ph.D. degrees in Biostatistics form the Limburgs Universitair Centrum (LUC, Belgium). Her research focuses on clustered data in the context of toxicological experiments, on pseudo-likelihood methodology, on spatial methodology in an epidemiological context, including cancer registration, and on high dimensional longitudinal data in neuroscience applications. She is co-author on a book on clustered data methodology and faculty in the M.Sc. in Biostatistics Programme at LUC.