MedicalBiostatistics.com

General linear models

Generalized estimating equations

GENERALIZED LINEAR MODELS

Generalized linear model (GLM) is an extension of the **general linear model** to the setup where the response variable may have a **distribution** far from **Gaussian**. The response can be **continuous** (with Gaussian or nonGaussian distribution) or **discrete** (proportion or count). Other conditions remain the same as in general linear models. In case you are not familiar with general linear models, we recommend that you familiarize with the general linear models before trying to grasp the essentials of the generalized linear models. A brief description of these models is available at this site under the term General linear models.

As in case of general linear models, there is no restriction on the **independent** or explanatory variables in the generalized linear models. These variables could be continuous or discrete, may pertain to **fixed effects** or **random effects** (or may be mixed). When they are mixed, the model would be called **generalized linear mixed model**. But these explanatory variables must affect the response through linear coefficients although the variables themselves could be square or log or any such function. If the effect is not linear, the GLM will study only its linear part. Equation wise the GLM is

$$y_{\mathrm{i}} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_K x_K + \varepsilon_i,$$

where y_i is the response of the *i*th person or unit, the regression coefficients βs are linear (i.e., no β^2, e^β, etc.) but x_2 could be x_1^2, etc.; and ε_i may *not* follow a Gaussian distribution in GLM. For general linear model, these errors are required to follow a Gaussian pattern, or nearly so if *n* is large. In the case of GLM, these errors may follow any of the wide spectrum of distributions of, what is known as, the *exponential family*. This family includes **binomial**, **Poisson**, **exponential**, **Weibull**, **beta**, **gamma**, etc. Notice that these distributions can be highly **skewed** and can belong to **discrete variables**.

You may be aware that general linear model is the combination of the **regression**, the **analysis of variance** and the **analysis of covariance** but the response variable must be continuous in all these setups. **Gaussian distribution** is required in general linear models for building up **confidence intervals** and for **tests of hypotheses** on the model parameters although **point estimates** can be obtained with **least square method** even when the distribution is not Gaussian. These restrictions are dispensed with in GLM method. GLM uses, what is called a **link function**, which converts the response to a form that has a relatively easily analyzable distribution. For example, proportions that follow a **binomial distribution** are converted to **logits**—this transforms the probability between 0 and 1 to values that can be positive, negative or zero. Similarly counts (e.g., number of patients coming to a clinic) that follow a **Poisson** distribution are transformed to logarithms to yield to nearly a Gaussian pattern. Logit and log are the respective link functions in these situations. There are other link functions for other setups. No transformation is required if the response variable is already Gaussian. This is called *identity link*. However, in the GLM, just as in general linear models, various values of the response variable must not be correlated, that is they must belong to separate persons or units who do not affect each other and not, for example, belong to the same family who are likely to provide similar values at least to some extent. When the values are correlated, use Generalized estimating equations **(GEE)**.

Estimates of the regression coefficients βs are obtained such that the likelihood of the sample coming from the distribution postulated by the link is maximum. These are popularly called **maximum likelihood estimates (MLEs)**. For Gaussian distribution, these MLEs are well known and can be easily derived, but many other distributions admissible under GLM require an iterative weighted least square procedure. Iteration in effect means that a start is made with some plausible estimates such as mean, checked if the model fits well to the observed data, and the estimates are revised according to the discrepancies found. This can go on for several iterations till such time that the updated estimates by two successive iterations are nearly the same. (This is called *convergence* – there may be situations where the estimates do not converge, in which case we say that we are not able to obtain the plausible estimates). Statistical packages are well trained to do these iterations for you, and you would not get wrong estimates if a standard package is used. These packages will give you the estimates of the βs, their standard errors (SEs) and will also test the statistical significance of each regression coefficient. You can then decide which of the explanatories is worth retaining and which ones can be discarded.

As in the case of general linear models, **standardization** by subtracting mean and dividing by the standard deviation (SD) is recommended for explanatory variables, particularly for continuous variables, so that each gets similar importance. If standardization is not done, the variable with large values such as cholesterol level compared with hemoglobin level sways the estimates and the statistical tests. In statistical terms, the cholesterol level will get disproportionately large weight in calculations relative to the hemoglobin level in the absence of standardization.

Goodness of fit of the model and statistical significance of the contribution of each or a set of explanatory variables is tested by deviance.

The GLM method is due originally to Nelder and Wedderburn [1]. Further details of the method are available in Dobson and Barnett [2].

[1] Nelder JA, Wedderburn RWM. Generalized linear models. *J Royal Statistical Soc, Ser A* 1972;135;370-84.http://biecek.pl/MIMUW/uploads/Nelder_GLM.pdf

[2] Dobson AJ, Barnett A. *An Introduction to Generalized Linear Models*, Third Edition. Chapman & Hall/ CRC Press, 2008.