HOME

MedicalBiostatistics.com

Generalized linear models

Generalized estimating equations

# GENERAL LINEAR MODELS

General linear model is the name given to the method that unifies the ordinary **regression**, the **analysis of variance** and the **analysis of covariance**. In all these setups, the response variable, also called the **dependent**, should be continuous and is expected to follow nearly a **Gaussian distribution**. The values must be independent and not correlated. This condition in effect means that the observations must belong to separate units or different persons and not belong to one family, one person at different points in time or different body sites, or any other affinity group that have tendency to be similar to one another.

To fix ideas, denote the response of the *i*th person by $y_i$ and his or her explanatory variables by $x_{1i}, x_{2i}, \ldots, x_{Ki}$. If you are trying to explain a particular kidney function by the person's weight, age, sex, water intake per day, fiber content in the diet, etc., this notation is like saying that 3rd person ($i$ = 3) in our study has kidney function $y_3$ = 54 mg/dL and his weight $x_{13}$ = 62 kg, age $x_{23}$ = 45 years, sex $x_{33}$ = 0 (where 0 is the notation for females and 1 for males), water intake per day $x_{43}$ = 3.4 liters, fiber content $x_{53}$ = 340 g. Note that $i$ = 3 in all these *x*s. Under these notations, a general linear model is given by

$$y_{\mathrm{i}} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_K x_{Ki} + \varepsilon_i,$$

where $y_i$ is the response or dependent, and $\varepsilon_i$ is the error that has Gaussian distribution with mean zero and any standard deviation (SD) σ. This is written as $\varepsilon_i \sim N(0, \sigma)$. N stands for normal distribution, which we like to call Gaussian since the term normal has different meaning in health and medicine. Note that the SD is same for each *i*—the condition popularly known as homogeneity of variances or **homoscedasticity**.

The *x*s are the explanatory or the **independent** variables and βs are the **regression coefficients**. These coefficients are the **parameters** of the model. The expression $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_K x_{Ki}$ is the mean of all the persons in the target population whose explanatory values are $x_{1i}, x_{2i}, \ldots, x_{Ki}$. This can be denoted by $\mu_i$. This and the previous explanation regarding $\varepsilon_i$ imply that $y_i \sim N(\mu_i, \sigma)$. In our example, $\mu_i$ is the population mean of those persons whose weight is 62 kg, age 54 years, etc. If there are 16 persons in the population with exactly same values of all *x*s, they will most likely have different kidney function and $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_K x_{Ki}$ is the notation of their mean. Thus general linear models are for means and not for individual values—a fact that many forget while interpreting a model. You can see that the mean response will change if any *x*-value changes, for example, if age is different or weight is different. Now it should be clear that $\varepsilon_i$ is the deviation of the response of the *i*th person in the population from its mean of similar people, thus $\varepsilon_i$ has zero mean. This is what is being called error. If the model is a good fit (which really means that the estimates of βs are such that the residuals are minimal—see next for the term residuals), their SD will be small. A large SD will indicate that the model is not a good fit. This SD is estimated by the square root of **mean square error (MSE)**, which is the sum of the squares of the residuals divided by its **degrees of freedom**. Residuals are explained later on in this section.

The model is linear so long as the coefficients are linear. That is, there is no $\beta^2$, $e^\beta$ or logβ type of coefficient. This means that when any $x_k$ increases by one, the response $y$ is supposed to increase by its coefficient $\beta_k$. The regression coefficient $\beta_k$ is generally interpreted as the *net contribution* or net effect of the variable $x_k$ ($k = 1, 2, \ldots, K$) but the word net is too strong. For this to be really net contribution, the model must include all possible variables that can affect the response. This is a tall order first because it is generally not feasible to include *all* the variables, and second because only those variables can be included that are known or suspected. Many are unknown and this **epistemic uncertainty** is many times forgotten. An example given later in this section will clarify one aspect of the interpretation of the regression coefficient. In case the relationship is not linear, general linear model will limit itself to whatever is the linear part and ignore the rest. In this case, the model will not be a good fit to the data.

However, there is no restriction on the *x*s. The type of *x*s distinguishes among ordinary regression, analysis of variance and the analysis of covariance. In ordinary regression, all *x*s have to be quantitative, in analysis of variance all *x*s have to be discrete—mostly defining the groups through **indicator variables**, and in the analysis of covariance these are mixed. It is easy to unify all these into one theoretical framework because basically all *x*s are considered fixed in this setup and the value of *y* is estimated for *given* values of the *x*s. The requirement of Gaussian distribution, independence and homoscedasticity is handy to pursue, what is called the **maximum likelihood estimates (MLEs)**. This method finds those estimates of the parameters that make the observed values most likely. These are denoted by *b*s. Any standard statistical software will easily obtain these estimates for you when the model is properly specified. Proper specification means that you correctly tell the computer program which variable is to be treated as continuous, which **categorical**, etc. When these estimates are used, the model can be written as

$$y_\mathrm{i} = b_0 + b_1x_{1i} + \ b_2x_{2i}+ \ldots + b_Kx_{Ki} + e_i,$$

and the estimated value of *y* for the *i*th person is $\hat{y}_i = b_0 + b_1x_{1i} + \ b_2x_{2i}+ \ldots + b_Kx_{Ki}$. Note that the error term is now denoted by $e_i$ and called **residual** in the context of the sample. This is the difference between the *mean* response in the sample for the persons with values $x_{1i}$, $x_{2i,} \ldots x_{Ki}$ of the explanatory variables and the actual observed response for the *i*th person. If there is only one person with these values of the *x*s, $e_i$ is the difference between the observed $y_i$ and the estimated $\hat{y}_i$.

Under Gaussian distribution of the response variable, the estimates *b*s of the regression coefficients βs also have Gaussian distribution. This allows to easily find the **confidence intervals (CIs)** for estimates of βs and to **test hypotheses** on them by using the property, for example, that the estimate ± 1.96*(estimated SE) is the 95% CI and (estimate – its mean)/(its estimated SE) has **Student *t*** distribution. This SE is estimated by the MSE as just explained.

A simple example of a general linear model is the work of Ainslie et al. [1]. They examined how blood flow velocity in the middle cerebral artery (MCAv) in healthy humans is affected by physical activity, body mass index (BMI), blood pressure (BP) and age. Physical activity was assessed as active and inactive. Thus this is categorical. The response variable is MCAv, which is quantitative and it must have nearly same variance for different age and different physical activity, etc. for general linear models to be applicable. The authors found that BMI and BP did not have statistically significant contribution while age and physical activity were important for MCAv. They have reported separate model for the active and the inactive persons but these combine into the following.

MCAv (in cm/sec) = 87.8 – 0.73*Age (in years) + 9.2*Activity – 0.03*Activity*Age,

where Activity = 1 for physically active persons and Activity = 0 for inactive persons. When these values of Activity are substituted, the models become

MCAv (in cm/sec) = 87.8 – 0.73*Age (in years) for inactive persons, and

MCAv (in cm/sec) = 97.0 – 0.76*Age (in years) for active persons.

This model means that MCAv reduced on average by 0.73 cm/sec for each year increase in age in inactive persons but by 0.76 cm/sec in active persons although the baseline for active persons was high (97.0 vs. 87.8). Gaussian distribution is not a requirement for getting these equations because the estimates of the regression coefficients can be obtained by the **least square method**. But the Gaussian distribution is needed to work out the CI. The authors also reported CI on these regression coefficients. Those who are aware will realize that this model is the same as the analysis of covariance where Age is the covariate. Analysis of covariance is the most generalized of the general linear models since it has both continuous and discrete independent variables.

Adequacy of a general linear model is assessed by an ***F*-test**, which is obtained as the mean **sum of squares** due to the model and the mean sum of squares due to error (MSE). Statistical significance of each regression coefficient can also be tested. Appropriate statistical software will do it for you but the model must be properly specified. Many modifications of the model can be done to test other kinds of **null hypotheses**. For a complete description of the general linear models, their strength and weakness, see Vik [2].

When the distribution of the response variable is far from Gaussian, we need to fall back on the Generalized linear models **(GLM)** and if the responses are correlated we take help of the Generalized estimating equations **(GEE)**.

[1] Ainslie PN, Cotter JD, George KP, Lucas S, Murrell C, Shave R, Thomas KN, Williams MJ, Atkinson G. Elevation in cerebral blood flow velocity with aerobic fitness throughout healthy human ageing. *J Physiol* 2008 Aug 15;586(16):4005-10. doi: 10.1113/jphysiol.2008.158279. Epub 2008 Jul 17. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2538930/

[2] Vik PW. *Regression, ANOVA, and the General Linear Model: A Statistics Primer*. Sage, 2013.