Linear and Generalised Linear Models

A **statistical model** describing the relationship of a collection of predictor variables to a response variable is an equation that relates values of those variables to each other in a probabilistic way. Generally a model for a relationship between variables assumes also a **model for the response variable** from some common family of distributions.

The most common models used in statistics are **general linear models**; these generalise simple linear models such as linear regression and one-way ANOVA. General linear models assume a normal model for the response variable.

Generalised linear models are a further extension of general linear models and can be used for a variety of different response variable models.

The three basic (two variable) linear models/main effect terms

The following simple regression models are important by themselves, but also show the types of main effect terms (terms representing the effect of a single predictor on the response) that can arise in a general or generalised linear model.

1. Simple linear regression is a linear equation describing the relationship between two quantitative variables X and Y. It is assumed there is a relationship of the form

$$Y = m + bX + \varepsilon,$$

where m and b are constants and ε is a normally distributed random variable with mean equal to 0.

2. One-way ANOVA is a linear equation describing the relationship between a (**fixed effect**) nominal predictor variable, X with k levels, and a quantitative response variable, Y. It is assumed that there is a relationship of the form:

$$Y = m + c_1 X_1 + \ldots + c_k X_k + \varepsilon.$$

Here the variables X_1 to X_k are "dummy variables" coding the various levels of the predictor X. The dummy variable X_i is equal to 1 for datapoints at the ith level of X, and 0 for all others. The constants are m and c_1 to c_k, chosen so that

$$c_1 + \ldots + c_k = 0.$$

As before, ε is a normally distributed random variable with mean=0.

3. A simple random effects model is a third basic linear model that is not often considered by itself, but which is useful to discuss. It relates a (**random effect**) nominal predictor variable X to a quantitative response variable, Y. In is assumed that there is a relationship of the form:

$$Y = m + U + \varepsilon,$$

where m is a constant, U is a normally distributed random variable with mean 0 that assigns an effect value to each level of Y, and ε is a normally distributed random variable with mean 0 that assigns a random error to each unit in the population.

Interaction terms

If there are two or more predictor variables in a study, then in addition to terms representing the effect of each predictor independently on the response, the model may include interaction terms, which describe the way the two predictors interact in their effect on the response (for more on this, see material on interaction plot in the glossary).

Interaction terms involving two variables are of one of the following forms, depending on the types of variables involved:

First variable, W	Second variable, X	Interaction term
Quantitative	Quantitative	aWX
Quantitative	Fixed effect nominal with k levels	$a_1WX_1 + \ldots + a_kWX_k$
Quantitative	Random effect nominal modelled by U	aWU
Fixed effect nominal with j levels	Fixed effect nominal with k levels	$a_{11}W_1X_1 + \ldots + a_{1k}W_1X_k + \ldots + a_{j1}W_jX_1 + \ldots + a_{jk}W_jX_k$

Fixed effect nominal with k levels	Random effect nominal modelled by U	$a_1W_1U+\ldots+a_kW_kU$
Random effect nominal modelled by U	Random effect nominal modelled by V	aUV

Interaction terms can involve any number of variables, and the formulas generalise naturally. However, generally models only include terms involving two predictors at a time.

It is important to note that including an interaction term also forces you to include the corresponding main effect terms in your model. That is, you can't have a two-way interaction term if you don't have the single variable main effect terms also. You can't have a three-way interaction without each subsidiary two-way interaction, etc.

General linear models

A **general linear model** (the result of multiple linear regression) is a linear equation relating several predictor variables, which may be a combination of quantitative variables (X_1 to X_n), fixed effect nominal variables (W_1 to W_m) and random effect nominal variables (Z_1 to Z_l), to a quantitative response variable, Y. A model involving both fixed and random effect nominal variables is sometimes called a **mixed model**.

In a general linear model, the equation is of the form y=rhs, where y is the quantitative response variable and the right hand side is a sum of terms representing

- Main effects, which can be quantitative, fixed effect nominal or random effect nominal,
- Interaction effects, which can involve any number of predictors, but generally only involve two at a time,
- A random error term ε, which is a normally distributed random variable.

Generalised linear models

A **generalised linear model** is of a very similar form to a general linear model—the right hand side expression involving the predictor variables is again a linear expression, but for a

generalised linear model, there is no “error term” ε. The left hand side of a generalised linear model is an expression involving a parameter in the assumed distribution for the response variable, often the mean of this distribution.

For example, in the case of logistic regression, used when the response variable is a Bernoulli trial (a nominal variable with two levels, such as gender), the left side of the regression expression is

$$\ln(p/1-p),$$

where p is the probability of success in the Bernoulli response variable.

Other examples of generalised linear models include:

1. Poisson regression models
2. Negative binomial regression models
3. Multinomial regression models
4. Ordinal regression models
5. Exponential regression models
6. Weibull regression models
7. Cox proportional hazards models
8. Accelerated failure time models.

Assumptions of general and generalised linear models

In a general or generalised linear model there are two main assumptions aside from the standard assumption that the sample is random.

1. The first assumption corresponds to **homoscedasticity of variances**, which is the assumption in a general linear model that the variance of the error term ε is independent of the values of the predictor variables. This assumption arises because the model of the response variable in multiple linear regression is as a **normally distributed variable**. Normal distributions have two **parameters**, mean and variance. Regression is only done on the mean; variance is assumed to be constant.

In any model of the response variable that has more than one parameter, the assumption of the corresponding generalised linear model is that all other parameters are independent of the values of the predictors. So, for instance, in a negative binomial model, regression is only done on either the mean or the spread parameter, and the other is assumed to be constant.

2. The second assumption is that the effect of a change in the predictor variable is linear on the left (response) side of the equation. Sometimes this is not a good assumption, and the relationship is more complicated. This can sometimes be corrected by a transformation of the predictor variable. Otherwise, it may be necessary to build a model where the right side is a more general function of the predictor variables. This is best done approximating by splines. This sort of modelling is quite involved, and it is advisable both to consult the literature in your area before using such a model, and to discuss your data with a consultant.