Intermezzo: Generalized Linear Models

Jan van den Broek

2020

1 Definition

A generalized linear model is characterized by the following 3 points:

1. The distribution is a distribution from **the exponential family**.
2. There is a linear part of the model called the **linear predictor**.
3. There is a function that relates the mean of the distribution with the linear part. This function is called **the link function**.

Let's, for the ease of illustration, take the linear part as $\alpha + \beta x$ were x is an group indicator variable. Some special cases of a generalized linear model are

1.1 Normal distribution; the linear model

This distribution is used for continuous data. The probability of observing an observation y_i

$$f(y_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2}\right]$$

and so the log-likelihood is:

$$l(\mu, \sigma) = \sum_{i=1}^{n} \left[\ln(\sigma\sqrt{2\pi}) - \frac{(y_i - \mu)^2}{2\sigma^2}\right]$$

The mean of this distribution is μ (population mean) and the variance is σ^2. The link function is the identity: $\mu = \alpha + \beta x$. α is the mean of the group not indicated and β is the difference in group means. This is the case because the linear model is for the mean.
As can be seen from the formula for $l(\mu, \sigma)$ the log-likelihood is maximized for μ if $\sum(\frac{(y_i-\mu)^2}{2\sigma^2})$ is minimized. This is equivalent to the least-squares method.

1.2 Binomial distribution; the logistic regression model

The mean of this distribution is π (population fraction) and the variance is $\pi(1-\pi)$. The log-likelihood is:

$$l(\pi) = \ln[L(\pi)] = \sum_{i=1}^{98} \left[y_i \ln(\pi) + (1 - y_i) \ln(1 - \pi) \right]$$

The link function is the log-odds (logit): $\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x$. α is the log-odds of the group not indicated and β is the difference in log-odds between the groups (which is the log odds-ratio). This is the case because the linear model is for the log-odds.

1.3 Poison log-linear model

The mean of this distribution is μ (the population mean of the counts) and the variance is μ. The log-likelihood is:

$$l(\mu) = \sum_{i=1}^{24} \left[-\mu + y_i \ln(\mu) - \ln(y_i!) \right]$$

The link function is the log: $\log(\mu) = \alpha + \beta x$. α log of the mean of the group not indicated and β is the difference in log-means between the groups (which is the log-ratio of the means). This is the case because the linear model is for the log of the means.

1.4 Other example

Other examples of generalized models are, the exponential distribution with a link function $\frac{1}{\mu} = \alpha + \beta x$; The gamma distribution with a link function $\frac{1}{\mu} = \alpha + \beta x$ and the inverse-gauss distribution with a link function $\frac{1}{\mu^2} = \alpha + \beta x$

2 Limited Range of observations

From generalized linear models one can observe that in many cases, as soon as the range of the observations is limited (e.g. there are only zero's and one's or the observations are larger then zero) then:

1. the mean and the variance are related
2. the link function is not the identity

The last point means that the difference between groups is not a difference in means but a difference in transformed means.
This also shows that if the range of the observations is limited, the normal distribution is not appropriate except as an approximation.