Generalized Linear Models

Recall that under the General Linear Model (or GLM) the mean of a vector of observations **y** is a linear function of the predictor variables (summarized in the design matrix **X**) and the to-be-estimated model parameters $\boldsymbol{\beta}$, so that E[**y**] = **X** $\boldsymbol{\beta}$, i.e.,

$$E[\, y_i \,] = \mu + \sum_{k=1}^{n} \beta_k x_{ik} \tag{1}$$

The **Generalized Linear Model** (note the **ized** ending) takes this a step further by assuming for some monotonic function g, that

$$E[\, y_i \,] = g\left(\mu + \sum_{k=1}^{n} \beta_k x_{ik} \right) \tag{2}$$

In particular, taking the inverse g^{-1} of the function g returns a linear model, with

$$g^{-1}\left(E[\, y_i \,] \right) = \mu + \sum_{k=1}^{n} \beta_k x_{ik} \tag{3}$$

The function f with the property that expresses the expected value of the response variable as a linear function of the predictor variables, i.e.,

$$f\left(E[\, y_i \,] \right) = \mu + \sum_{k=1}^{n} \beta_k x_{ik}$$

is called the **link function** of the particular generalized linear model.

Binary Logistic Regressions

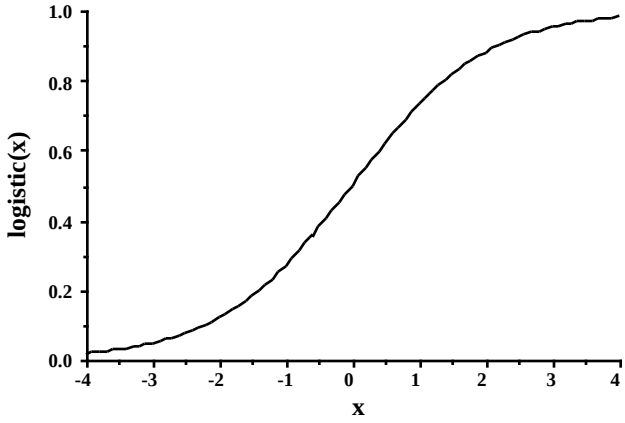
The classic example of a generalized linear model is when our response data y is **binary**, so that we can code it as zero/one. For example, one has the disease/does not have the disease, lives/dies, the device fails/device does not fail. To model binary data, a quite reasonable and very general approach is to use the predictor variables (the x's) to estimate the probability p that $y = 1$. For example, suppose we wish to model the presence/absence of a disease as a function of age x. A first (naive) attempt is to model $p(x)$, the probability of the disease given the age is x, as a simple linear function of age

$$p(x) = \alpha + \beta x$$

The problem with this approach is that there is no guarantee that p will lie between zero and one for all the x values in our study. What we would like is a function that is zero for very small x and one for large x. One such widely-used function is the **logistic**,

$$l(x) = \frac{1}{1 + \exp(-x)} \tag{4}$$

For $x < -4$, this is essentially zero, while for $x > 4$, this is essentially one, as the following figure shows:



The logistic function is the basis for **binary logistic regression**. Note that if y takes on value one with probability p and is zero otherwise, then $E[Y] = p$. The logistic regression of y on x is given by

$$p = \Pr(Y = 1 \,|\, x) = l(\alpha + \beta x) = \frac{1}{1 + \exp(-\alpha - \beta x)} \tag{4}$$

More generally, with n predictor variables $x_1, \cdots x_n$, the logistic regression becomes

$$\Pr(Y = 1 \,|\, \mathbf{X}) = \frac{1}{1 + \exp(-\mathbf{X}\boldsymbol{\beta})} = \left(1 + \exp\left[-\left(\mu + \sum_{k=1}^{n} \beta_k x_{ik}\right)\right]\right)^{-1} \tag{5}$$

The inverse of the logistic function is given by the **logit** function, so that if $p = l(x)$, then given p we can solve for x = logit(p), where

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) \tag{6}$$

The logit function is thus the link function for the binary logistic regression model. The logit function is also referred to as the **log of the odds**, the log of the ratio of a success ($Y = 1$) to a failure ($Y = 0$). Note that under the binary regression model,

$$\text{logit}(Y = 1|\mathbf{X}) = \text{logit}(p|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta} \tag{7}$$

The logit funcion transforms the data to a standard general linear model. Note that the logistic regression is a probability model, returning the value p for a given value of the response variables. Hence, the residuals for a given value of p follow a binomial distribution with variance $p(1-p)$, i.e., the residuals are non-normal and heteroscedastic.

Ordinal Logistic Regression

Often the data is discrete, but not binary. However, if we can order (ordinate) the data, for example, a disease might have states normal, mild, strong, and terminal, we can extend the logistic regression to ordinal data as follows. Denote the ordered states of the response variable by $0, 1, \cdots, k$. The ordinal logistic regression is given by

$$\begin{aligned}\Pr(Y \geq j \,|\, \mathbf{X}) &= \frac{1}{1+\exp[-(\alpha_j + \mathbf{X}\boldsymbol{\beta})]} \\ &= \left(1 + \exp\left[-\left(\alpha_j + \mu + \sum_{k=1}^{n} \beta_k x_{ik}\right)\right]\right)^{-1} \end{aligned} \tag{8}$$

In particular, $\Pr(Y = j) = \Pr(Y \geq j-1) - \Pr(Y \geq j)$, or

$$\Pr(Y = j \,|\, \mathbf{X}) = \frac{1}{1+\exp[-(\alpha_{j-1} + \mathbf{X}\boldsymbol{\beta})]} - \frac{1}{1+\exp[-(\alpha_j + \mathbf{X}\boldsymbol{\beta})]}$$

Log-linear Models for Discrete Data

Binary and logistic regression deal with proportions, or ratios of discrete (count) data. Often, we deal directly with the count data, such as the number of insects in a sample. If the number if sufficiently large, one can often use a normal approximation, assuming the count data is drawn from a normal. However, if the data is very discrete (for example, a number of the observations are zero or one), the normal model is not appropriate.

One model for discrete (count) data is the **Poisson distribution**, where the probability we observed k counts is given by

$$\Pr(k) = e^{\lambda}\,\frac{\lambda^k}{k!} \tag{9}$$

Under this model, the mean is λ. Hence, if the response variable Y returns count data, one model is to assume it follows a Poisson distribution where λ is a function of the predictor variables (the x_i's), so that

$$\lambda(\mathbf{x}) = f\left(\sum_{i=1}^{n} \beta_i x_i\right) \tag{10}$$

Again, the obvious model is to assume a completely linear function,

$$\lambda(\mathbf{x}) = \sum_{i=1}^{n} \beta_i x_i \tag{11}$$

The problem is that the mean count number λ must be positive, and this is not guaranteed by using a strictly linear model. However, one model that always returns a positive value is to use the exponential function,

$$\lambda(\mathbf{x}) = \exp\left(\sum_{i=1}^{n} \beta_i x_i\right) = \prod_{i=1}^{n} \exp(\beta_i x_i) \tag{12}$$

This model thus assumes that the variables interact in a multiplicative fashion. The link function for this model is the log, with

$$\ln(\lambda \,|\, \mathbf{x}\,) = \sum_{i=1}^{n} \beta_i x_i \tag{13}$$

Hence, if μ_i denotes the mean value expected for the ith observation response variable and $\mathbf{x}_i$ the vector of response variables for the ith observation, then the log-linear model is of the form

$$E[\,y_i\,] = \mu_i = \boldsymbol{\beta}^t \mathbf{x}_i$$