STA 442/2101F Notes on generalized linear models.   November, 2010

We can think of the linear model

$$y_i = x_i^T \beta + \epsilon_i, \quad \epsilon_i \overset{ind}{\sim} (0, \sigma^2), \tag{1}$$

in two parts. First we specify the mean structure

$$E(y_i \mid x_i) = x_i^T \beta,$$

and then we specify the distribution of Y_i:

$$Y_i \sim N(x_i^T \beta, \sigma^2), \tag{2}$$

where we may not always need to make the normality assumption, but this is implicit if we base inference on t-tests and F-tests.

One possibility that is not very often addressed in textbooks is to relax the normality assumption but keep model (1); for example to assume that ϵ follows a t-distribution with an unknown number of degrees of freedom. This is touched on in Davison, §8.4.

The covariates in (1) may be of many types: measurements of continuous variables, transformations of continuous variables, sets of functions of these variables, such as cubic polynomials, and sets of dummy variables to index levels of a factor or categorical covariates. The response variable y is typically continuous, as the model would not usually be suitable for a response variable that was categorical.

One way to generalize (1) is to allow the mean structure to be nonlinear in β, leading to solution by nonlinear least squares. This is addressed in Davison §10.1. So, for example, we might have

$$E(y_i \mid x_i) = \beta_0 \{1 - \exp(-x_i/\beta_1)\}$$

and $y_i \sim N\{E(y_i \mid x_i), \sigma^2\}$. Equivalently of course we can write $y_i = E(y_i \mid x_i) + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$. It is usual in this model to estimate β by least squares, which is maximum likelihood under the normality assumption: iterative solution will be needed. These types of intrinsically non-linear models are most natural when there is a fairly well-understood mechanism that converts x to y, such as in modelling the uptake of drugs in biological systems, or the growth of populations in some ecological models.

Another generalization of (1) is to extend the distribution of response to accommodate a wider variety of types of response. For example, if y_i is a count variable, that takes values $0, 1, \ldots$, a simple starting point is the Poisson distribution:

$$f(y_i) = \lambda_i^{y_i} e^{-\lambda_i}/y_i!, \quad y_i = 0, 1, \ldots; \lambda_i > 0. \tag{3}$$

We then might link our n observations using a set of p covariates by modelling the mean function as

$$E(y_i) = \lambda_i = \exp(x_i^T \beta).$$

1

The exponential form is chosen because this ensures $\lambda_i > 0$. Note that there is no ϵ_i in model (2), i.e. no additive error. The randomness in the response is generated from the Poisson distribution, not from measurement (or other types of) error. The final model for a sample of size n is

$$L(\beta; \underline{y}) \propto \prod_{i=1}^{n} \{\exp(x_i^T \beta)\}^{y_i} e^{-\exp(x_i^T \beta)}, \tag{4}$$

where constants not involving β have been dropped.

When the response is a number of or a proportion of successes in independent trials, then a binomial distribution seems a plausible starting point. Thus

$$f(y_i) = \binom{m_i}{y_i} p_i^{y_i} (1 - p_i)^{m_i - y_i}, \, y_i = 0, 1, \ldots, m_i; 0 \leq p_i \leq 1. \tag{5}$$

The logistic model for p_i assumes

$$\log \left(\frac{p_i}{1 - p_i} \right) = x_i^T \beta,$$

or

$$p_i = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}.$$

The likelihood function is then

$$L(\beta; \underline{y}) \propto \prod_{i=1}^{n} \exp[(x_i^T \beta) y_i - m_i \log\{1 + \exp(x_i^T \beta)\}]. \tag{6}$$

Models (4) and (6), and in fact also (2), are all examples of *generalized linear models*; the generalization is through the mean structure, which is not linear in β, but is related to a linear function of β via a link function. In R generalized linear models are fit using `glm(formula = ...  , family = ...  , data = ...  )`. If the `family` is binomial, then the `formula` has as its left hand side either a vector of 1's and 0's, or a two-column matrix of successes and failures. (The second is just a way to tell R what the m_i's are.)

Generalized linear models are fit by maximum likelihood, and we use general asymptotic results from the theory of likelihood inference to construct confidence intervals and compute p-values.