Generalized Linear Models

A generalized linear model (GLM) consists of three parts.

i) The first part is a random variable giving the conditional distribution of a response Y_i given the values of a set of covariates X_{ij}. In the original work on GLM's by Nelder and Wedderburn (1972) this random variable was a member of an exponential family, but later work has extended beyond this class of random variables.

ii) The second part is a linear predictor, $\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$.

iii) The third part is a smooth and invertible link function $g(.)$ which transforms the expected value of the response variable, $\mu_i = E(Y_i)$, and is equal to the linear predictor:

$$g(\mu_i) = \eta_i = \alpha + \beta_1 X_{i1} \ + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}.$$

As shown in Tables 15.1 and 15.2, both the general linear model that we have studied extensively and the logistic regression model from Chapter 14 are special cases of this model.

One property of members of the exponential family of distributions is that the conditional variance of the response is a function of its mean, $\upsilon(\mu)$, and possibly a dispersion parameter ϕ. The expressions for the variance functions for common members of the exponential family are shown in Table 15.2. Also, for each distribution there is a so-called canonical link function, which simplifies some of the GLM calculations, which is also shown in Table 15.2.

Estimation and Testing for GLMs

Parameter estimation in GLMs is conducted by the method of maximum likelihood. As with logistic regression models from the last chapter, the generalization of the residual sums of squares from the general linear model is the residual deviance,

$$D_m \equiv 2(\log L_s - \log L_m),$$

where L_m is the maximized likelihood for the model of interest, and L_s is the maximized likelihood for a saturated model, which has one parameter per observation and fits the data as well as possible. If the dispersion parameter ϕ is fixed at 1, then testing Models 0 and 1 where Model 0 has

$k_0 + 1$ coefficients and is nested within Model 1 which has $k_1 + 1$ coefficients, is conducted with

$$G_0^2 = D_0 - D_1,$$

which is just the likelihood-ratio statistic, having an asymptotic chi-squared distribution with $k_1 - k_0$ degrees of freedom. Confidence intervals can be obtained as described in the text by inverting the likelihood-ratio test. When the dispersion parameter is estimated, then two nested models 0 and 1 can be compared with an F test,

$$F_0 = \frac{\frac{D_0 - D_1}{k_1 - k_0}}{\widetilde{\phi}},$$

where the estimated dispersion parameter $\widetilde{\phi}$ is from the largest model, just as we have seen for ANOVA models in earlier chapters. If the largest model has $k + 1$ coefficients, then under H_0, F_0 follows an F distribution with $k_1 - k_0$ and $n - k - 1$ degrees of freedom. We can also define an R^2-like quantity with

$$R^2 = 1 - \frac{D_1}{D_0},$$

where D_0 is a model containing only the term α, and D_1 is the model of interest.