

THE LOGISTIC REGRESSION MODEL

Generalized Linear Models (GLMs) are able to model non-normally distributed dependent variables, and thus overcome the problems of the assumptions of regular linear regression models (Venables and Ripley 2002). Quinn and Keough (2002) note that GLMs have three components: 1) a response (dependent) variable with a population distribution belonging to the exponential family, 2) the predictor (independent) variables, and 3) a 'link function' that links 1) and 2). For example, the logistic model (for multiple predictors) is:

$$\pi(x) = \frac{e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i}}{1 + e^{\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i}} \quad \text{Eq. 1.1}$$

where $\pi(x)$ is the probability that the response variable $y = 1$, α is the equation constant, and β_i is the coefficient of predictor variable x_i . Thus, the binary response variable is modelled as an odds ratio, i.e. the probability that y will be a member of one class relative to the other class (Trexler and Travis 1993). Rather than model Eq. 1.1 directly, the link function $g(x)$ allows the response variable to be modelled as:

$$g(x) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i \quad \text{Eq. 1.2}$$

For a binomial response variable the logistic (logit) link is the natural logarithm of the odds ratio:

$$g(x) = \ln \left[\frac{\pi(x)}{1 - \pi(x)} \right] \quad \text{Eq. 1.3}$$

Thus, the logistic regression model is more easily solved as:

$$\ln \left[\frac{\pi(x)}{1 - \pi(x)} \right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i \quad \text{Eq. 1.4}$$

The equation intercept (constant) α and variable coefficients β_i are estimated from calibration data using maximum likelihood techniques. Once these are known, the model can be applied to estimate future states based on an alternative data set considering the same variables.

This method can be extended for response variables with more than two categories by using the Multinomial Logit Model (MNL) as a probability model to estimate the category of the response variable given the predictor variables. Consider y as a dependent variable that can potentially take one of J nominal categories, and that these categories are numbered from 1 to J (but not assumed to be ordered). Let

$$X\Phi_m = \alpha + x_1\beta_1 + x_2\beta_2 + \dots x_i\beta_i \quad \text{Eq. 1.5}$$

where m is also a category of y . The probability of observing each category m can now be calculated (when $\Pr(y = m|X)$ is the probability of observing m given X , the set of predictor variables):

$$\Pr(y = m|X) = \frac{1}{1 + \sum_{j=2}^J e^{X\Phi_j}} \quad \text{for } m = 1 \quad \text{Eq. 1.6}$$

$$\Pr(y = m|X) = \frac{e^{X\Phi_m}}{1 + \sum_{j=2}^J e^{X\Phi_j}} \quad \text{for } m > 1 \quad \text{Eq. 1.7}$$

To statistically test the importance of predictor variables in these models, the Likelihood Ratio (LR) statistic is used. The LR statistic is calculated by comparing the Residual Deviance (RD) of the full model (containing all variables) against a reduced model (full model minus the variable in question):

$$\text{LR Statistic} = -2(\log\text{-likelihood}_{\text{reduced}} - \log\text{-likelihood}_{\text{full}}) \quad \text{Eq. 1.8}$$

The resulting statistic is compared to the χ^2 distribution, to examine whether the variable has a significant effect on the response variable (i.e. the variable coefficient is statistically different from zero), with:

$$\text{Degrees of Freedom} = (J - 1)(N_{X_{\text{full}}} - N_{X_{\text{reduced}}}) \quad \text{Eq. 1.9}$$

where N_X is the number of predictor variables.

References

- Quinn GP and Keough MJ. 2002. Experimental design and data analysis for biologists. Cambridge University Press: Cambridge.
- Trexler JC and Travis J. 1993. Nontraditional Regression Analyses. *Ecology* 74:1629-1637.
- Venables WN and Ripley BD. 2002. Modern applied statistics with S. Fourth edition. Springer: New York.