Confidence intervals for scenario means and their differences and ratios

Roger Newson

28 August, 2005

1 Formulas

Assume that we fit to a set of data a generalized linear model with P parameters $\beta = (\beta_1, \ldots, \beta_P)^T$ to a set of data with N data points. We will denote the $Y-$value for the ith data point as Y_i, we will denote the jth $X-$value for the ith data point as X_{ij}, we will denote the conditional mean for the ith data point as μ_i, and we will denote its link function as η_i. The overall mean is defined as

$$M = N^{-1} \sum_{i=1}^{N} \mu_i \, , \tag{1}$$

and its derivative with respect to the jth parameter is

$$G_j = \frac{\partial M}{\partial \beta_j} = N^{-1} \sum_{i=1}^{N} \frac{\partial \mu_i}{\partial \eta_i} X_{ij} \, , \tag{2}$$

and the derivative of its log with respect to the jth parameter is

$$\Gamma_j = \frac{\partial}{\partial \beta_j} \log(M) = \left(\sum_{i=1}^{N} \mu_i \right)^{-1} \sum_{i=1}^{N} \frac{\partial \mu_i}{\partial \eta_i} X_{ij} \, . \tag{3}$$

Note that, except in the trivial case of a linear link function (such as the familiar identity link), these gradients are *not* the gradients arrived at by setting all the X-variates to their sample means.

To define confidence intervals for M and $\log(M)$, we define the P-column row vectors $\mathbf{G}$ and $\mathbf{\Gamma}$ by (2) and (3) respectively, denote by $\mathrm{Cov}(\beta)$ the covariance matrix of the vector parameter β, and we then have the estimates

$$\mathrm{Var}\,(M) = \mathbf{G}\mathrm{Cov}(\beta)\mathbf{G}^T, \quad \mathrm{Var}\,[\log(M)] = \mathbf{\Gamma}\mathrm{Cov}(\beta)\mathbf{\Gamma}^T \, , \tag{4}$$

and calculate standard errors and symmetric confidence limits in the usual manner, possibly exponentiating these confidence limits in the case of $\log(M)$ to derive asymmetric confidence limits for M.

To compare expected overall means under different scenarios, we usually want to estimate either their differences or their ratios. Using out–of–sample prediction, we can fantasize that, under "Scenario $*$", we have a sample of N^* observations, and their hypothesized X–values are denoted X_{ij}^* for the jth X–variate in the ith observation, and their hypothesized expected Y–values and their link functions (assuming the same β as before) are denoted μ_i^* and η_i^*, respectively, for the ith observation. The overall scenario mean is then

$$M^* = N^{*-1} \sum_{i=1}^{N^*} \mu_i^* \,, \tag{5}$$

and we can define vectors $\mathbf{G}^*$ and $\mathbf{\Gamma}^*$ analogously to (2) and (3), respectively, and define confidence intervals for M^* and its log using formulas similar to (4). For a second scenario, denoted "Scenario $**$", we might similarly assume a sample size of N^{**}, define X–values X_{ij}^{**}, expected Y–values μ_i^{**}, link functions η_i^{**}, an overall scenario mean M^{**}, and gradient vectors $\mathbf{G}^{**}$ and $\mathbf{\Gamma}^{**}$. The difference $M^* - M^{**}$ between the expected overall means under the two scenarios has a variance estimated as

$$\mathrm{Var}\left(M^* - M^{**}\right) \quad = \quad \left(\mathbf{G}^* - \mathbf{G}^{**}\right) \mathrm{Cov}(\beta) \left(\mathbf{G}^* - \mathbf{G}^{**}\right)^T \,, \tag{6}$$

and the corresponding log ratio $\log(M^*/M^{**})$ has a variance estimated as

$$\mathrm{Var}\left[\log\left(M^*/M^{**}\right)\right] \quad = \quad \left(\mathbf{\Gamma}^* - \mathbf{\Gamma}^{**}\right) \mathrm{Cov}(\beta) \left(\mathbf{\Gamma}^* - \mathbf{\Gamma}^{**}\right)^T \,. \tag{7}$$

We can therefore calculate standard errors and confidence limits for the scenario difference $M^* - M^{**}$, and for the log scenario ratio $\log(M^*/M^{**})$, in the usual manner, and define asymmetric confidence limits for M^*/M^{**}.

An important special case of the scenario ratio is the population unattributable fraction, which is subtracted from one to define the population attributable fraction. In the case of a cohort study, "Scenario *" might represent a hypothetical version of our cohort if they were all non–smokers and were the same in all other respects, and "Scenario **" might represent the cohort we actually have. In the case of a case–control study, "Scenario **" might represent the controls in our sample (assumed to represent the population at large because of the rare-disease assumption), and "Scenario *" might represent a hypothetical sample who are all non–smokers, but who are like the controls in our sample in all other respects. For further information on these examples, see Bruzzi *et al.* (1985) and Greenland and Drescher (1993).

2 References

Bruzzi P, Green SB, Byar DP, Brinton LA, Schairer C. Estimating the population attributable risk for multiple risk factors using case–control data. *American Journal of Epidemiology* 1985; **122(5)**: 904–914.

Greenland S, Drescher K. Maximum likelihood estimation of the attributable fraction from logistic models. *Biometrics* 1993; **49**: 865–872.