Bias correction in GLMs

Bicko, Jonathan & Ben

2021 Jun 21 (Mon)

Notation :

- x_f: a value of the focal predictor
- $x_{\{n\}}$: a vector of values of the non-focal predictors for a particular observation
- $\eta(x_f, x_{\{n\}}) = \beta_f x_f + \sum \beta_{\{n\}} x_{\{n\}}$ = linear predictor (e.g. prediction on the log-odds scale)
- $g^{-1}()$: inverse-link function (e.g. logistic)
- $D(x_{\{n\}}|x_f)$: distribution of the non-focal predictors conditional on a particular value of the focal predictor
- β_{fi}: the coefficient describing the interaction(s) of the focal and non-focal parameters

—

Population-based approach for bias correction

Binned non-focal linear predictor

To implement this:

- compute linear predictor of the non-focal predictors, $\eta_{\{n\}} = \sum \beta_{\{n\}} x_{\{n\}}$
- find a list of vectors of observations of $\eta_{\{n\}}$ associated with each value (bin) of the focal predictor, $\eta_{j\,\{n\}}$, $j = 1, 2, \cdots$
- for each $\eta_{j\,\{n\}}$:
  - compute $\hat{y}_j = \text{mean } g^{-1}\left(\beta_f x_{j_f} + \eta_{j\,\{n\}}\right)$

If we compute the individual back-transformed predictions for a poorly sampled/finely spaced set of focal values, we will get a noisy prediction line as the values of the non-focal predictors shift across the focal values. Simple example: suppose everyone below the median age has wealth index w_1, everyone above the median has w_2. Then the predicted value will have a discontinuity at the median age. We can deal with this by taking bigger bins (a form of smoothing), or by post-smoothing the results (by loess, for example). The principled form of this would be to assume/recognize that our uneven distribution of observed non-focal predictors actually represents a sample of a distribution that will vary *smoothly* as a function of the focal predictor.

Whole population non-focal linear predictor

Instead of binning non-focal associated with particular level of focal predictor, we add the overall contribution of the non-focal predictor to the corresponding value of the focal linear predictor. In particular:

- compute linear predictor of the non-focal predictors, $\eta_{\{n\}} = \sum \beta_{\{n\}} x_{\{n\}}$
- for every value of the focal predictor, x_{j_f}:
  - compute $\hat{y}_j = \text{mean } g^{-1}\left(\beta_f x_{j_f} + \eta_{\{n\}}\right)$

Delta method

Suppose that we have a response Y and the transformed response on the link scale, η and an inverse-link function $h(.) = g^{-1}(.)$. We consider η as a one-dimensional random variable with mean (μ_η) and standard deviation (σ_η). We are interested in the expected value of $h(\eta)$, i.e., the back-transformed values so that $Y = h(\eta)$. More specifically, based on a second-order Taylor expansion

$$Y \approx h(\mu_\eta) + h'(\eta)(\eta - \mu_\eta) + \frac{1}{2}h''(\eta)(\eta - \mu_\eta)^2$$

so that

$$E(Y) = h(\eta) + \frac{1}{2}h''(\mu_\eta)\sigma_\eta^2$$

McCulloch and Searle (2001)

$$E(Y) \approx h\left(\eta - \frac{0.5\sigma^2 \tanh(\eta(1 + 2\exp(-0.5\sigma^2)))}{6}\right)$$

Diggle et al. (2004)

$$E(Y) \approx h\left(\frac{\eta}{\sqrt{1 + \left(\frac{16\sqrt{3}}{15\pi}\right)^2 \sigma^2}}\right)$$