Analytical Uncertainty Quantification for Multilevel Mediation Analysis

August 19, 2024

1 Notes to Self

- Summarize problem
- Describe MSEP formula and decomposition
- Give conditional variance formula
  - Explain derivation
  - Explain components of formula
- Justify neglecting correction term

This note summarizes the work of Booth and Hobert (1998) as it pertains to Rowin's thesis.

Our statistical model is a GLM with random effects. We assume that our response, Y, has an exponential family distribution with the canonical parameterization. That is,

$$f(y) = \exp\left([y\theta - b(\theta)] + c(y)\right) \tag{1}$$

If Y is Bernoulli(p), then $\theta = \log[p/(1-p)]$, $b(\theta) = \log(1+e^{\theta}) = -\log(1-p)$, and $c(y) = 0$. We model the parameter θ as depending on two sets of covariates, X and Z, through some fixed effects, β, and random effects, U, respectively. We assume that the observed responses occur in groups, or clusters. Write Y_{ij} for the jth observation from group i. We further assume that each group has its own level for the random effect, U_i, and that these random effects are independent and identically distributed from some common distribution, usually Normal$(0, \Sigma)$.

Since we're doing regression, we will focus on the conditional mean of Y given covariates. Since we're doing mixed-effects modelling, we also want to condition on the random effects, U. To this end, write $\mu_{ij} = \mathbb{E}(Y|U = U_i, X = X_{ij}, Z = Z_{ij})$ for the mean of observation j from group i, when the random effects and covariates are held fixed. Following the usual GLM setup, we model $g(\mu_{ij}) = \eta_{ij} = X_{ij}\beta + Z_{ij}U_i$.

As an aside, I've always had trouble keeping all the different scales which arise in GLMs straight in my head. There's μ, the mean of Y, η, the linear predictor, and θ, the canonical parameter[1]. There are then functions which map between these different scales: the link function, $g(\mu) = \eta$, the mean function, $b'(\theta) = \mu$, and the "variance function", $b''(\theta)$. This last one bothers me because if you include nuisance (or overdispersion) parameters in your exponential family then the variance of Y is not equal to the variance function. If you choose the canonical link function though, then $g(\mu) = \theta$ (this is the definition of the canonical link), which really muddies things for me. Ramble over.

Our current goal is the estimate/predict the values of the random effects within each group. It is common in the literature to use the word "predict" here, because the random effects aren't parameters to be estimated, but instead random variables, which we would usually talk about "predicting". Note that the different groups in our problem are independent, so we can focus on prediction of a single U_i, then apply our method individually to the other groups. Booth and Hobert recommend predicting U_i with its conditional mean given the data from group i, Y_i. Note that this conditional mean depends on the model parameters, β and Σ, which in practice must be estimated. A lot of software (e.g. the `lme4` package in `R`) actually uses the conditional mode instead of the conditional mean, since there is a nice algorithm to compute the mode, but the mean is hard in general.

Booth and Hobert discuss prediction of the linear predictor, η_{ij}, but we can just think of this as U_i.

References

James G. Booth and James P. Hobert. Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association*, 93(441), 1998.

[1] Note that this canonical parameter often differs from the parameters we're used to thinking about for the distribution in question. See, e.g., the Bernoulli distribution above, or try expressing the Normal(μ, σ) as an exponential family distribution.