Supporting Information

Lean on your statistics: The generalization and simplification of the balance intercept problem

Boyi Guo, Jacqueline Rudolph

The marginal mean (probability of event for binary outcome), $\mathbb{E}_Y(Y)$ can be expressed as a double expectation of the covariates $\boldsymbol{X}$

$$\begin{aligned}\mathbb{E}_Y(Y) &= \mathbb{E}_{\boldsymbol{X}}(\mathbb{E}_Y(Y|\boldsymbol{X})) \\ &= \mathbb{E}_{\boldsymbol{X}}(g^{-1}(\beta_0 + \boldsymbol{\beta}_1\boldsymbol{X})),\end{aligned}$$

where g^{-1} is the inverse function of the link function g, β_0 is the balance intercept of interest, $\boldsymbol{\beta}_1$ is the coefficient vector for the covariates $\boldsymbol{X}$ which can be the enumeration of a categorical variable or multiple continuous variable.

Balance Intercept Calculation on the Linear Predictor Scale

Only when $g^{-1}(\mathbb{E}(\cdot)) = \mathbb{E}(g^{-1}(\cdot))$ (e.g. g^{-1} is a linear function), we can accurately calculate the balance intercept on the linear predictor scale, as

$$\begin{aligned}\mathbb{E}_Y(Y) &= g^{-1}\{\mathbb{E}_{\boldsymbol{X}}(\beta_0 + \boldsymbol{\beta}_1\boldsymbol{X})\} \\ g\{\mathbb{E}_Y(Y)\} &= \mathbb{E}_{\boldsymbol{X}}(\beta_0 + \boldsymbol{\beta}_1\boldsymbol{X}) \\ g\{\mathbb{E}_Y(Y)\} &= \beta_0 + \mathbb{E}_{\boldsymbol{X}}(\boldsymbol{\beta}_1\boldsymbol{X}) \\ \beta_0 &= g\{\mathbb{E}_Y(Y)\} - \mathbb{E}_{\boldsymbol{X}}(\boldsymbol{\beta}_1\boldsymbol{X}). \qquad (1)\end{aligned}$$

Equation (1) further simplifies to the analytic approximation in Rudolph et al. (2021) when $\boldsymbol{X}$ are pairwise independent,

$$\beta_0 = g\{\mathbb{E}_Y(Y)\} - \sum_{j=1}^{p} \beta_j \mathbb{E}_{X_j}(X_j).$$

Balance Intercept Calculation on the Response Scale

When $g^{-1}(\mathbb{E}(\cdot)) \neq \mathbb{E}(g^{-1}(\cdot))$, it would be only sensible to conduct the calculation on the response scale. And the calculation can be complicated as the complexity of the link function and the number of predictors increase. We show how to derive the balance intercept when the link function is the logarithm function, i.e. $g(x) = \log(x)$.

$$\begin{aligned}\mathbb{E}_Y(Y) &= \mathbb{E}_{\boldsymbol{X}}(\exp(\beta_0 + \boldsymbol{\beta}_1\boldsymbol{X})) \\ &= \mathbb{E}_{\boldsymbol{X}}(\exp(\beta_0) * \exp(\boldsymbol{\beta}_1\boldsymbol{X})) \\ &= \exp(\beta_0)\mathbb{E}_{\boldsymbol{X}}(\exp(\boldsymbol{\beta}_1\boldsymbol{X})) \\ \exp(\beta_0) &= \mathbb{E}_Y(Y)/\mathbb{E}_{\boldsymbol{X}}(\exp(\boldsymbol{\beta}_1\boldsymbol{X}) \\ \beta_0 &= \log\{\mathbb{E}_Y(Y)/\mathbb{E}_{\boldsymbol{X}}(\exp(\boldsymbol{\beta}_1\boldsymbol{X})\} \\ \beta_0 &= \log(\mathbb{E}_Y(Y)) - \log(\mathbb{E}_{\boldsymbol{X}}(\exp(\boldsymbol{\beta}_1\boldsymbol{X})). \qquad (2)\end{aligned}$$

When $\boldsymbol{X}$ are pairwise independent, Equation (2) simplifies to

$$\beta_0 = \log\{\mathbb{E}_Y(Y)\} - \sum_{j=1}^{p} \log(\mathbb{E}_{X_j}(\exp(\beta_j X_j))).$$

If the moment generating function $M_X(t)$ is known, it can be used to simplify the calculation for $M_X(\beta) = \mathbb{E}_X(\exp(\beta X))$ for individual variables or $M_{\boldsymbol{X}}(\boldsymbol{\beta}_1) = \mathbb{E}_{\boldsymbol{X}}(\exp(\boldsymbol{\beta}_1 \boldsymbol{X})$ for the variable vector.

For some more complicated link function, e.g. the logistic function $g(x) = log\frac{x}{1-x}$, or the distribution of covariate X is unknown, it is very difficult, if not impossible, to derive the closed-form solution of the balance intercept. Hence, it would be preferred to use the numeric solution.

Balance Intercept with Effect Coding

When the mean statistics (e.g. the prevalence of a disease) of each level/group is given, it is more convenient to simulate the outcomes out of each level/group and aggregate the group data to compose the overall data set. For a 2-level example, we can verify the simulation procedure works in this case, as follows

$$Pr(Y=1) = \frac{n_1 Pr(Y=1|X=0) + n_2 Pr(Y=1|X=1)}{n_1+n_2} = \frac{n_1 Pr(Y=1|X=0) + n_2 (Pr(Y=1|X=0) + RD}{n_1+n_2}$$

However, this case rarely happens even when we can calculate the conditional probability for each group from the original estimand (e.g. risk ratio and odds ratio with the reference group statistics) as continuous covariates could not be considered in this case unless use some form of categorization.

calculate

the balance intercept and hence the construct the simulation procedure using the effect coding system, following its definition. As previously introduced, the intercept term in the effect coding regression model is simply the grand mean