Exploration of calibration intercept (non-white)

Qi Wang

October 11, 2023

1 Data generation process

We generate 500 datasets for cross-validation. The generation process is as follows:

- The sample size of each dataset is $N = 50,000$.
- We consider three continuous predictors and two binary predictors: $(X_1, X_2, X_3) \sim \mathcal{N}\left(\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{bmatrix}1 & 0.5 & 0.1\\0.5 & 1 & 0.3\\0.1 & 0.3 & 1\end{bmatrix}\right)$, $V_1 \sim \text{Bernoulli}(0.7), V_2 \sim \text{Bernoulli}(0.4)$.
- For the stratifying variable, race/ethnicity group, we assume it is binary (white/non-white) and follows $W \sim \text{Bernoulli}(p_W)$, where p_W denotes the proportion of white people. Furthermore, $p_W = 0.9$.
- With the predictors (X, V) and the stratifying variable W, the binary disease outcome Y is generated by the logistic regression model: $P(Y = 1) = \frac{\exp\left(\alpha+\beta_X^\top X+\beta_V^\top V+\beta_W W\right)}{1+\exp\left(\alpha+\beta_X^\top X+\beta_V^\top V+\beta_W W\right)}$. $\beta_X, \beta_V, \beta_W$ are all fixed where $\beta_X = (\beta_{X_1}, \beta_{X_2}, \beta_{X_3}) = (0.3, 0.5, 0.7), \beta_V = (\beta_{V_1}, \beta_{V_2}) = (0.4, 0.6), \beta_W = 0.8$. We select the value of α such that the prevalence of cases in the population is $p_{case} = 0.05$.

2 Stratified CV

- When the number of CV folds is 5
  - Formula-based SE is around 100 times larger than the MC-based SE if we use average method (0.0932 vs 0.00084)
  - Formula-based SE is around 100 times larger than the MC-based SE if we use aggregate method (0.0932 vs 0.00075)
  - Formula-based SE is around 3.6 times larger than the MC-based SE if we only focus on 1 fold (0.208 vs 0.057)
- When the numer of CV folds is 2
  - Formula-based SE is around 44.1 times larger than the MC-based SE if we use average method (0.0923 vs 0.0021)
  - Formula-based SE is around 41.0 times larger than the MC-based SE if we use aggregate method (0.0923 vs 0.0023)
  - Formula-based SE is around 3.1 times larger than the MC-based SE if we only focus on 1 fold (0.131 vs 0.0424)

  - Formula-based SE is very close to the MC-based SE if we mimic average method when doing external validation (0.131 vs 0.136). Specifically, in the ith round of CV, we train our model on the train set except for the ith fold and assess model performance on the ith fold of the external data.

Now, we use 2-fold CV to examine:

- If we only conduct hypothesis using the first round of CV, the rejection rate is 0
- If we only conduct hypothesis using the second round of CV, the rejection rate is 0
- Averaging the two rejection rates yields the rejection rate of 0
- The MC-based correlation between the estimated calibration intercepts based on the first and second rounds of CV is -0.994

3 Non-stratified CV

- When the number of CV folds is 5
  - Formula-based SE is around 17 times larger than the MC-based SE if we use average method (0.061 vs 0.0036)
  - Formula-based SE is around 40 times larger than the MC-based SE if we use aggregate method (0.061 vs 0.0016)
  - Formula-based SE is almost the same as the MC-based SE if we only focus on 1 fold (0.137 vs 0.149)
- When the numer of CV folds is 2
  - Formula-based SE is around 42.5 times larger than the MC-based SE if we use average method (0.0926 vs 0.0022)
  - Formula-based SE is around 19.67 times larger than the MC-based SE if we use aggregate method (0.0924 vs 0.0047)
  - Formula-based SE is around 0.68 times that of the MC-based SE if we only focus on 1 fold (0.131 vs 0.193)
  - Formula-based SE is 0.70 times that of the MC-based SE if we mimic average method when doing external validation (0.092 vs 0.133). Specifically, in the ith round of CV, we train our model on the train set except for the ith fold and assess model performance on the ith fold of the external data.

Now, we use 2-fold CV to examine:

- If we only conduct hypothesis using the first round of CV, the rejection rate is 0
- If we only conduct hypothesis using the second round of CV, the rejection rate is 0
- Averaging the two rejection rates yields the rejection rate of 0
- The MC-based correlation between the estimated calibration intercepts based on the first and second rounds of CV is -0.9997