A skewed model for imbalanced health data

November 8 2022

Illustration depicting the asymmetric model developed by KAUST scientists that addresses the problem of imbalanced health data. Credit: KAUST; Xavier Pita

An asymmetric statistical model provides a better fit for imbalanced data with rare "positives," such as longitudinal health datasets.

Sometimes, a more complex but more accurate model is needed when the standard off-the-shelf models just do not cut it. That is the message from researchers from KAUST's Statistics Program.

One interesting example is for large health datasets that contain the

occurrence of rare diseases. Particularly in longitudinal studies that track many patients over many years, searching out the few instances of a disease in a large data set poses challenges for standard statistical approaches.

"In longitudinal studies, we might want to find the relationship between a certain disease and several potentially influential factors," says Zhongwei Zhang, a Ph.D. student with Raphael Huser. "To do so, we might collect data over time from hundreds of subjects. The resulting response data would be binary—either disease or no disease—and the responses for the same subject are correlated because they are collected from the same person."

For such correlated binary response data, the state-of-the-art model is the multivariate probit model. However, this model might not be suitable when the data are not distributed symmetrically or are not balanced, with roughly as many positives as negatives.

"The multivariate probit model might not always provide the best fit for highly imbalanced data because of this symmetric link model, possibly resulting in substantial bias in the estimation of the mean response," explains Zhang. "There is a need to develop flexible asymmetric link models for this type of data. In this study, we developed a novel multivariate skew-elliptical link model that can explain the data better."

The skew-elliptical link model is a flexible model that is able to capture the imbalance in the data, such as cases when the majority of the results are zero, but a small and significant portion is equal to one. With the multivariate probit model embedded as a special case, this model's mathematical flexibility allows it to be used for both balanced and imbalanced data.

The new model, developed by Zhang with KAUST professors Marc

Genton and Huser, was shown to provide a better fit to a highly imbalanced COVID-19 dataset from a region of California in the United States.

"There is often a tradeoff between flexibility and parsimony," Zhang says. "If you are looking for easily interpretable models with efficient inference, then go for the parsimonious models at hand. But if you are looking for models with the best performance according to certain criterion, there might exist more complicated models that are more suitable."

The research was published in *Biometrics*.

More information: Zhongwei Zhang et al, Tractable Bayes of skew-elliptical link models for correlated binary data, *Biometrics* (2022). DOI: 10.1111/biom.13731

Provided by King Abdullah University of Science and Technology