BETTER INFERENCE FOR LOGISTIC REGRESSION IN SPARSE DATA

S.B. Bull[1,2], S.S.F. Lee [1,2], J.P. Lewinger[1,3]

[1]*Samuel Lunenfeld Research Institute of Mount Sinai Hospital, Toronto, Canada*
[2]*University of Toronto, Toronto, Canada*
[3]*University of Southern California, Los Angeles, USA*

Email: *bull@mshri.on.ca*

Logistic regression is a widely used regression model in practice, but conventional maximum likelihood-based inference does not perform well in small or sparse samples, due to failure of the quadratic approximation to the log-likelihood. Modification of the logistic regression score function to remove first order bias is equivalent to penalizing the likelihood by the Jeffreys prior, and yields penalized likelihood estimates (PLEs) that always exist, even in samples in which maximum likelihood estimates (MLEs) are infinite. PLEs are less biased and have smaller MSE than MLEs in small to moderate-sized samples, and their profile likelihood confidence intervals (CIs) have better coverage. However, both MLE and PLE Wald-type tests perform poorly. We present penalized likelihood ratio (LR) test statistics for the multinomial logistic model, including single parameter tests and joint tests of parameters across regressions, and apply the methods in sparse datasets with infinite estimates. Using simulations, we compare test size and power of PLE and MLE LR statistics in binomial and trinomial regressions with both binary and continuous covariates, and evaluate a hybrid testing strategy in which test statistic choice depends on the occurrence of infinite estimates in a dataset. In settings where finite sample bias and infinite estimates are likely to occur, we recommend wider reliance on inference using PLE profile CIs and LR tests in preference to the corresponding MLE methods.