Abstract

The first objective of this research is to compare the efficiency of the coefficients of determination (R_O^2, R_l^2) and adjusted coefficients of determination ($R_{O,adj,MS}^2$, $R_{O,adj,LM}^2$, $R_{l,adj,MS}^2$, $R_{l,adj,LM}^2$, $R_{l,adj,SAS_{DEV}}^2$, $R_{l,adj,SAS_{AIC}}^2$) for a logistic regression based on ordinary least squares (OLS) and maximum likelihood (ML). The second objective is to study the relative bias and root mean square error (RMSE) of the R^2 and R_{adj}^2 values with 3 types of misspecified models: (1) wrong functional form of the covariate; (2) misspecified link function; and (3) missing covariate. The true link function is logit. This research is a Monte Carlo simulation study conducted with different predictor numbers (1, 5, and 10) and different observation numbers (50, 100, 1000, and 2000). The evaluation of the performance of the studied R^2 and R_{adj}^2 measures are based on the relative bias, RMSE, and the percentage that the coefficient of determination estimate lies outside [0,1] as calculated from 1,000 runs.

Several conclusions were drawn from this research. The $R_{l,adj,SAS_{DEV}}^2$ value is the least efficient measure in all situations. The $R_{l,adj,LM}^2$ value is the most efficient measure of the proportion of variation of a dependent variable explained by covariates when the sample size is small and/or the number of explaining variables is large.

When the 3 types of misspecified models were studied, the adjusted coefficients of determination based on maximum likelihood yielded higher relative biases than that the adjusted coefficients of determination based on ordinary least squares. The $R_{l,adj,SAS_{AIC}}^2$ and $R_{l,adj,LM}^2$ values gave the highest relative biases in the first and second ranks, respectively. The complementary log-log link function induces a higher RMSE for $R_{l,adj}^2$ than $R_{O,adj}^2$.