Model selection and parameter estimation with missing covariates in logistic regression models

Fabrizio Consentino and Gerda Claeskens

K.U. Leuven
ORSTAT and Leuven Statistics Research Center
Naamsestraat 69, 3000 Leuven
Fabrizio.Consentino@econ.kuleuven.be

Abstract

This research deals with logistic regression models under the presence of missing covariates on some observations, when the missing data mechanism is ignorable. We extend the model selection method introduced by Claeskens and Consentino (2008), to the t-distribution in order to choose and decide which distribution fits the missing covariates in a better way. Especially with outlying observations, the t-distribution would be advised. The latter method is based on the EM algorithm and it is computationally intensive. Our new method avoids the iterative approach, employing the method of Gao and Hui (1999) in the context of logistic regression models. Further we extend their parameter estimation method in two ways. First, instead of only working under normality, we assume a univariate t-distribution for the error of the univariate missing covariate. Second, we allow to have more than one covariate missing, under multivariate normal or t-distributions. The model/distribution selection method and estimation procedure are investigated via a simulation study and real data analysis.

References

Claeskens, G. and Consentino, F. (2008). Variable selection with incomplete covariate data, *Biometrics*, to appear.

Gao, S. and Hui, S. L. (1997). Logistic Regression Models with Missing Covariate Values for Complex Survey Data, *Statistics in Medicine*, **16**, 2419–2428.