

Robust Box-Cox response transformations based on optimal prediction

A. Marazzi¹ and V. Yohai²

¹ University of Lausanne, Institut de médecine sociale et préventive, Bugnon 17, CH 1005 Lausanne, Switzerland

² University of Buenos Aires, Departamento de Matematica, Ciudad Universitaria, Pabellon 1, 1428 Buenos Aires, Argentina

Keywords: Robust regression, Box-Cox transformations, conditional M-expectation, smearing estimate.

1 Abstract

Response transformations have become a widely used tool to make data conform to a linear regression model. By far the most common example is the Box-Cox transformation. The transformed response is usually assumed to be linearly related to the covariates and the errors normally distributed with constant variance. The regression coefficients, as well as the parameter λ defining the transformation, are generally estimated by maximum likelihood (ML). Unfortunately, near normality and homoscedasticity are hard to attain simultaneously with a single transformation. In addition, the ML-estimate is not consistent under non-normal or heteroscedastic errors and it is not robust.

Various semiparametric and nonparametric approaches to relax the parametric structure of the response distribution have been studied. However, these procedures do not provide effective protection against heavy contamination and heteroscedasticity. A first proposal of robust Box-Cox transformations for simple regression, which are robust and consistent even if the assumptions of normality and homoscedasticity do not hold, is given in Marazzi and Yohai (2004).

This paper presents new estimates based on optimization of the prediction error. Our multiple regression model does not specify a parametric form of the error distribution. In order to develop a new nonparametric criterion, we introduce the basic concept of conditional M-expectation (CME), a robust version of the classical conditional expectation of the response for a given covariate vector. The CME minimizes a M-scale in place of the classical mean squared error. We then consider the CME of the transformed response as a function of λ , the coefficients being estimated using a robust (e.g., MM-) estimator. The optimal prediction property of the CME provides a criterion to define the CME-estimate of λ , as well as some modifications of this estimate. An efficient resampling algorithm to compute the new estimates is described and asymptotic properties are investigated. Since the conditional mean of the response on the original scale is often the parameter of interest, we provide a robust version of the smearing estimate (Duan, 1983) which is consistent for the CME. Monte Carlo results show that the new estimators perform better than other available methods. Applications concerning modeling of hospital cost of stay with the help of covariates such as length of stay and admission type are presented.

2 References

Duan N. (1983). Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association*, 78, 605-610.

Marazzi A., Yohai V.J. (2004), Robust Box-Cox transformations for simple regression. *Theory and Applications of Recent Robust Methods*, edited by M. Hubert, G. Pison, A. Struyf and S. Van Aelst, Series: Statistics for Industry and Technology, Birkhauser, Basel. In press.