On the choice of a linear model for regression or time series analysis

Anthony Atkinson[1] **and Marco Riani**[2]

[1] *London School of Economics, UK*
[2] *University of Parma, Italy*

Abstract

We combine the selection of a statistical model with the robust parameter estimation and diagnostic properties of the Forward Search. As a result we obtain procedures that select the best model in the presence of outliers. We derive distributional properties of our method and illustrate it on data. The effect of outliers on the choice of a model is revealed.

There is a vast literature on methods for model selection. The basic idea of Akaike (1974) is that in choosing between non-nested statistical models there needs to be a trade-off between the improved fit of a larger model and the increased number of parameters. Akaikes elegant solution uses an information criterion to penalize twice the maximized log-likelihood by two times the number of parameters in the model. In his honour this criterion is known as AIC.

There is no restriction on the form of the models to which AIC can be applied. A separate development in regression led to the C_p statistic (Mallows 1973, named in honour of Cuthbert Daniel) in which the residual sum of squares for each model is penalized by twice the number of parameters in the linear model. The nuisance parameter σ^2 is estimated from a full model containing sufficiently many terms to provide an unbiased estimate. We exploit the close relationship between the two procedures in the case of Gaussian models.

In regression the null distribution of C_p is F. For time series data we use state-space modelling and the Kalman filter (Harvey 1989) to derive a similar statistic of known distribution for the selection of a time series model, with or without regressor variables.

In both cases these criteria for choice of a model are based on aggregate statistics and so are subject to potential distortion due to the presence of outliers. We use the forward search (Atkinson and Riani 2000) to provide efficient and robust criteria that exhibit the effect of each observation on the process of model choice. Although C_p and AIC are then based on truncated samples, we obtain useful distributional results for C_p during the forward search. Finally we stress that we do not view use of these criteria as an automatic way of choosing a model. We show how the indications of AIC and

C_p can be confirmed by combining a forward search with standard methods of regression analysis.

Keywords

AIC, C_p, Forward search, Outliers.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*, 716–723.

Atkinson, A. C. and M. Riani (2000). *Robust Diagnostic Regression Analysis*. New York: SpringerVerlag.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics 15*, 661–675.