Crucial Problems in Regression Model Building

Milan Meloun

Department of Analytical Chemistry, Faculty of Chemical Technology, Pardubice University, CZ-532 10 Pardubice, Czech Republic, email: milan.meloun@upce.cz,

The main part of regression model building is the careful analysis of the *regression triplet*. This analysis is facilitated by an examination of *the data* (influential points, i. e. outliers and high leverages), *the model proposed* and *the method* for parameter estimation. The various types of residuals indicate *suspicious points* which should then be tested for influence. On the basis of an analysis of the residuals (ordinary, normalized, standardized, jackknife, predicted and recursive) and diagonal elements of the projection matrix, five diagnostic plots for *influential points* indication (the graph of predicted residuals, Pregibon graph, Williams graph, McCulloh-Meeter graph and Gray's L-R graph) may be used. Index graphs of vector and scalar influence measures (the diagonal elements of the hat matrix and modified hat matrix, Cook measure, Atkinson measure, Belsey's DFFITS measure, Anders-Pregibon measure, Cook-Weisberg likelihood measures are elucidated. The six most efficient diagnostic plots were selected to give a reliable indication of influential points, and four of these are able to separate influential points into outliers and leverages. A critical survey of 22 influence diagnostics is compared on an illustrative example.

The main problems in regression are often caused by multicollinearity. When multicollinearity in data occurs, the ordinary least squares estimates of regression parameters are still unbiased but their variances are large so their significance is lower. With strong multicollinearity the parameter estimates and hypotheses tests are affected more by linear "links" between indepen-dent variables than by the regression model itself. The classical t-test of significance is highly inflated owing to large variances in regression parameter estimates and the results of statistical analysis are often unacceptable. One way of variances reduction is utilization of biased regression estimate. By adding a degree of bias to the regression estimates, *the Generalized Principal Component Regression* reduces variances which avoids problems with the "jump" in regression results due to neglecting small principal components. The results of regression are continuously changed in dependence on precision parameter P. Several statistical criteria for the selection of suitable bias (the coefficient of determination R^2, the predicted coefficient of determination R^2_p and the Akaike information criterion AIC) on the problem of parameter estimation in a polynomial regression model should be considered together. One of the most suitable seems to be the *mean quadratic error of prediction MEP*.

References:

1. M. Meloun, J. Militký and M. Forina, *Chemometrics for Analytical Chemistry, Vol. 2. PC-Aided Regression and Related Methods*, Horwood, Chichester, 1994.

2. M. Meloun, J. Militký, M. Hill, R. Brereton: The Analyst 127 (2002) 433 - 450.

3. M. Meloun, J. Militký, K. Kupka, R. Brereton: Talanta 57 (2002) 721 - 740.