Regression Model Development and Yet Another Regression Function

Dr. Werner Stahel
Seminar für Statistik, ETH Zürich

31st March 2008

A strategy to develop a regression model involves many steps and decisions which are based on pertinent numeric tables and thorough analysis of residual plots. With the standard regression functions available in R, such an assessment consists of several function calls and informed settings of their arguments, depending also on the type of target variable (continuous, count, binary, multinomial, multivariate, ...). The examination of a logistic regression fit, e.g., involves calling glm, summary, drop1, influence, plot, and termplot, and selecting the useful information from what is obtained from them.

This contribution presents a user oriented function that sets the sensible choices for the different models. It produces an object which gives the useful information for judging the model fit when printed and plotted.

More specifically, the function accepts the same arguments as `lm` or `glm`, and some more. It also accepts ordered, multinomial, and multivariate responses. Of course, calculations are done by calling the available fitting functions.

The function stores results that are produced by the fitting function and by calling `summary` on the object, as well as some additional ones, like the leverage values. If printed, it gives a table that contains, for continuous or binary explanatory variables, the coefficients, their P-values, the collinearity measure R_j^2 and a new measure of significance that additionally characterizes the confidence interval. For factors, the P-value is given, since individual coefficients and their P-values are of limited information content. The coefficients of factor levels are reproduced separately. – The last part of the print output is very similar to the usual summary part of printing the `summary`, but includes, in the case of a `glm`, an overdispersion test if applicable.

The strength of the new function lies in its plotting method. All residual plots use a plotting scale that is not affected by outliers in the residuals, but outliers are still shown in a marginal region of the plot. Most plots are complemented by a smooth by default. In order to judge the significance of any curvature shown by this line, 19 such lines are simulated from random data corresponding to the model. Reference lines indicate contours of equal response values and help to identify suitable transformations of the explanatory variables.

In summary, the function `regr` and its printing and plotting methods have made my life much easier when developing regression models and have lead to higher quality of analyses obtained by students.