TEACHING OF REGRESSION MODELS BUILDING BY ADSTAT PACKAGE

Jirí Militký, Technical University, Czech Republic
Karel Kupka, Trilobyte Ltd. Pardubice, Czech Republic
Milan Meloun, Pardubice University, Czech Republic

The first part of this contribution is devoted to the description of graphically oriented strategy for regression models building. The level of presentation corresponds to the parts of course for graduate students already having completed the general statistics course but without deeper knowledge about regression. In the second part, the student's version of ADSTAT software, and its modules relevant for regression models building, is discussed. Third part contains a practical example from the textile branch.

INTRODUCTION

Regression type models building is a relatively specific discipline ranging across mathematical statistics, informatics and technical sciences. Multiple linear and nonlinear model building belongs generally to the most complex problems solved in practice. In many cases it is not possible to construct the mathematical form of model based on the information about the system under investigation. In these cases the interactive approach to regression type models building could be attractive.

In the proposed strategy of regression models building , the graphically oriented methods for estimation of model correctness and identification of spurious data are selected. These methods are based on the special projections enabling the investigation of partial dependencies of response on the selected exploratory variable. Classical ones are partial regression graphs. For identification of spurious data the so called LR graphs can be used as well. For evaluation of model quality the characteristics derived from predictive capability are used. Some statistical tools for realization of the above mentioned techniques are described in the book (Meloun, Militky, and Fornia, 1994).

For the teaching of regression models building at the graduate students level it is necessary to have software for simple and interactive data analysis by linear and nonlinear regression with extensions for the above mentioned graphically oriented strategy of model building and evaluation of their quality. An example of this software type is ADSTAT, which was built on the ground of the authors' long time experience with regression modelling and teaching of this topic at technical universities.

SUMMARY OF LINEAR REGRESSION

The standard linear model with n observations of m explanatory variables is assumed. For additive model of measurements errors the linear regression model has the form $\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta} + \varepsilon_i$ .The n x m matrix **X** contains the values of m explanatory (predictor) variables at each of n observations, $\boldsymbol{\beta}$ is the m x 1 vector of regression parameters and ε_i is n x 1 vector of experimental errors. The **y** is n x 1 vector of observed values of the dependent variable (response).

When some assumptions are valid (see Meloun et al, 1994), the parameter estimates **b** found by minimization of least squares criterion are best linear unbiased estimators (BLUE). The corresponding covariance matrix is $D(\mathbf{b}) = \sigma^2 (\mathbf{X}^T\mathbf{X})^{-1}$ (Meloun et al, 1994)

From geometrical point of view, columns of design matrix **X** define m-dimensional subspace L in n-dimensional Euclidean space E^n. The vector $\mathbf{X}\boldsymbol{\beta}$ and prediction vector $\mathbf{y}_P = \mathbf{X}\,\mathbf{b}$ lie in subspace L. The prediction vector is orthogonal projection of vector **y** to the subspace L. Projection matrix has the form $\mathbf{H} = \mathbf{X}\,(\mathbf{X}^T\,\mathbf{X})^{-1}\,\mathbf{X}^{T.}$ Residual vector $\mathbf{e} = \mathbf{y} - \mathbf{y}_P$ is orthogonal to subspace L and has the minimal length. Variance matrix corresponding to prediction vector $\mathbf{y}_P$ has the form $D(\mathbf{y}_P) = \sigma^2\,\mathbf{H}$ and variance matrix for residuals is $D(\mathbf{e}) = \sigma^2\,(\mathbf{E} - \mathbf{H})$.

Statistical analysis related to least squares is based on normality of estimates **b**. Quality of regression is often (not correctly) described by the multiple correlation coefficient R (Meloun et al, 1994). For model building the multiple correlation coefficient is not suitable. It is non decreasing function of number of predictors and therefore the overdefined model often results.

Prediction ability of regression model can be characterized by quadratic error of prediction (MEP) Optimal model has minimal value of MEP. The MEP can be used for definition of the predicted multiple correlation coefficient PR (see, Meloun et al, 1994).

GRAPHICAL AIDS FOR MODEL CREATION

In multiple regression one usually starts with the assumption that response y is linearly related to each of the predictors. The aim of graphical analysis is to evaluate the type of non-linearity due to function of predictors describing well the experimental data.

Several diagnostic plots have been proposed for detection of curve between y and x_j (Berk and Booth, 1995). Very useful for designed experiments without marked colinearities is partial regression plot (PRL). This plot uses the residuals from the regression of y on the predictor x_j, graphed against the residuals from the regression of x_j on the other predictors. If the term $\mathbf{x}_j$ is correctly specified the partial regression graph forms straight line. Systematic nonlinearity is indication of incorrect specification of $\mathbf{x}_j$. Random pattern shows unimportance of $\mathbf{x}_j$ for explaining the variability of **y**. The partial regression graph (PRL) has the following properties:

1. The slope c in PRL is identical with estimate b_j in a full model.
2. The correlation coefficient in PRL is equal to the partial correlation coefficient R_{yxj}.
3. Residuals in PRL are identical with residuals for full model.
4. The influential points, nonlinearities and violations of least squares assumptions are markedly visualized.

The PRL graph can be constructed without recalculating of least squares (Meloun et al, 1994).

STRUCTURE OF ADSTAT PACKAGE

Modular statistical system ADSTAT consists of relatively independent modules, each of them containing a data editor, and several programs concerning one kind of statistical problems.

Data can be entered and pretreated in a spreadsheet like data editor. This editor enables the block oriented operations as a copying, deletion, filling by number and user defined functional transformation. Data can be read from or stored into ASCII character files.

All results of computation are stored in special results file. This file can be revised, edited on screen or printed. Hard copy of graphs from screen can be printed or stored in TIFF format. The ADSTAT is controlled by a user-friendly interactive menu-driven system combined with user’s panels. The basic control element is a lined menu offering data input, method choice, setting of initial conditions and outputs in table or graphical form. The menu system has a general tree like (hierarchical) structure. The user’s panels contain initially default values for all initial information which will be appreciated by students and beginners. For models building linear and nonlinear regression modules are created:

Linear Regression

This module serves for creation of linear and linearized regression models, estimation of their parameters and corresponding statistical analysis. For linear regression, special algorithms have been implemented. The least squares method is the only special case among a series of biased parameter estimation, controlled by a single parameter. Before computing, the data can be transformed to a polynomial form, the Taylor expansion (up to quadratic terms) or generally (any variable is transformed by a user function). A powerful SVD based algorithm is used (Meloun et al, 1994).Variety of regression characteristics including partial regression graphs can be computed. For diagnostic purposes, the program with over 40 graphs for proving the assumptions about data, models and least squares criterion have been included (Meloun et al, 1994).

Nonlinear Regression

For parameter estimation and statistical analysis of user defined nonlinear models the least squares criterion is used. Iterative minimization is based on stable and reliable algorithm MINOPT. The regression model containing up to nine independent variables and up to nine regression parameters is written in the usual algebraic form. Some regression parameters can be held constant during minimization. Process of parameters estimation can be interrupted or aborted by user. This leads to the flexible control of convergence, by restarting etc.

In the student version (limited to 100 points and without saving graphs in TIFF format) are modules independent and therefore can be used separately. Each module can be operated from single 5.25" diskette.

Professional version uses extensively text windows, pull down menu and interactive panels. The rules of their use are simple, direct and generally known from other programs. System is controlled with keyboard and/or mouse. The program allows to work interactively or to analyze data files in the batch operating mode.

EXAMPLE

This is the typical example used in a course of empirical models building for textile students. The aim is the description of PET/cotton yarn tenacity (response) in dependence on the following parameters (rotor diameter (x_1), rotor speed (x_3), yarn fineness (x_2) ,PET fibres length (x_5) and fineness (x_4)). Details of yarn creation and tenacity measurements are given in work (El Shahat, 1994).

Students use these data in two runs. In the first run the linear regression model is created by using of ADSTAT. The partial regression graph for rotor speed is in Figure 1.

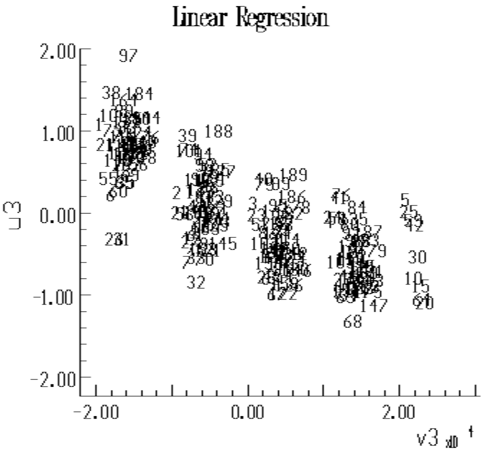

Figure1. Partial regression graph for rotor speed (run 1)

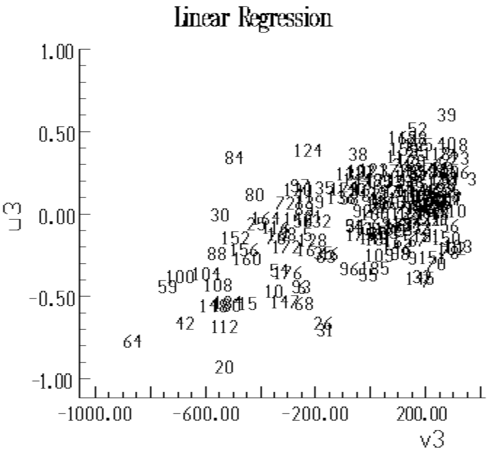

Figure 2. Partial regression graph for rotor speed (run 2)

Formally, the linear regression model is fully acceptable. All regression parameters are significant on the significance level 0.05 and multiple correlation coefficient is equal to R= 0.9425. Other statistical characteristics are: predicted correlation coefficient equal to PR= 0.93862 and mean quadratic error of prediction MEP =0.19738. From partial regression graphs is clear that there are some nonlinearities in variables x_2 and x_3 mainly. The negative sign of coefficient for rotor speed variable(b3 = -4.14e-5 ± 2.6e-6) is not acceptable from the point of view of the practical interpretation.

In the second run the model equivalent to the Taylor expansion of the unknown function to the quadratic terms is used. Partial regression graph for rotor speed is on the figure 2. It is clear that x_3 (rotor speed, see Figure 2) is now significantly linear and has the right positive sign (b3 = 5.869e-4 ± 6.63e-5). The significance of the other variables is hidden in the interactions or quadratic terms. The multiple correlation coefficient is equal to the R= 0.98455, predicted correlation coefficient PR= 0.98451 and mean quadratic error of prediction MEP =0.05097. This model has therefore better predictive ability than linear one and physical interpretation of this model is now without problems.

CONCLUSION

The utilization of partial regression graphs and suitable criterion expressing the predictive ability is very useful for building of statistical models especially based on the experimental design arrangements. Once students have mastered the creation and the interpretation of partial regression graphs, they are able to build the empirical regression models by the interactive application of ADSTAT.

ACKNOWLEDGEMENT

This work was supported by the Czech Ministry of Education Grant No. VS 97084.

REFERENCES

Berk, K. N., Booth, D. E. (1995). Seeing a curve in Multiple Regression, *Technometrics 37,* 385.

Meloun, M, Militký, J and Forina, M. (1994) Chemometrics in Analytical Chemistry vol. II, Interactive Model Building and Testing on IBM PC, Ellis Horwood, Chichester.

El Shahat I. (1994). Diploma Thesis, Mansoura University, Mansoura.