Slovak University of Technology in Bratislava
Institute of Information Engineering, Automation, and Mathematics

PROCEEDINGS

17$^{\mathrm{th}}$ International Conference on Process Control 2009

Hotel Baník, Štrbské Pleso, Slovakia, June 9 – 12, 2009

ISBN 978-80-227-3081-5

http://www.kirp.chtf.stuba.sk/pc09

Editors: M. Fikar and M. Kvasnica

Javůrek, M.: Program for Analysis of Residua, Editors: Fikar, M., Kvasnica, M., In *Proceedings of the 17th International Conference on Process Control '09*, Štrbské Pleso, Slovakia, 528–533, 2009.

Full paper online: `http://www.kirp.chtf.stuba.sk/pc09/data/abstracts/029.html`

PROGRAM FOR ANALYSIS OF RESIDUA

M. Javůrek

University of Pardubice, Studentská 95, 532 10 Pardubice, Czech Republic
Fax: +420 466 037 068, e-mail: milan.javurek@upce.cz

Abstract: At present, the evaluation of experimental data by means of regression methods represents one of most frequently adopted procedures thanks to large expansion of computer technology. There exist a number of professional algorithms that perform regression calculations of various functions with application of many known methods. However, the calculation results need to be additionally verified, i.e. it must be stated whether or not the solution found is sufficiently correct and accurate. For this purpose a significant tool is the analysis of residua, but this analysis is often omitted even in commercial algorithms. Therefore, a subroutine REZID was assembled: it tests the obtained set of residua by numerical and some graphical methods. The subroutine was designed mainly for use with the program set SPONA [LUKŠAN, 1982 and 1987], which is why it was written in FORTRAN 77 language.

Keywords: analysis of residua, nonlinear regression

1 INTRODUCTION

One of the most significant and most frequently used tools for evaluation of quality of a model from both the standpoint of presumptions about random component and the standpoint of quality of selected observations is the analysis of set of residua [JAVŮREK, 2006]. These are presented in their normal form or transformed as normalised, studentized etc. It can be clearly stated that any lack of randomness found in residua indicates some imperfection of the model used. For instance, systematic changes of residua depending on the variables included indicate an insufficient number of terms in regression equation; the dependence on non-included terms expresses the necessity of these terms in regression equation. Systematic changes of variability of residua with changes of independent variables are signs of heteroscedascity. By means of analysis of residua it is possible to reveal extreme values of dependent variable or off-lying or too influential observations. This diagnostics usually combines numerical and graphical methods.

2 NUMERICAL METHODS

2.1 Classical Residua

Sets of classical values of residua are evaluated by means of current statistics, where the normality of their distribution is considered:

- arithmetic mean – it should approach zero
- dispersion (variance) and/or standard deviation as its square root should approach the instrumental error
- coefficient of skewness – it should approach zero
- coefficient of kurtosis – it should vary about the value of three

The relationships for calculating the above-mentioned estimates of these central moments can be found in any literature about mathematical statistics, e.g. [MELOUN, 1986].

An excellent characteristic is Hamilton's *R*-factor, which expresses goodness of fit:

$$R_F = \sqrt{\frac{n \, . \, s^2}{\sum_{i=1}^{n} y_i^2}} \qquad (1)$$

where s^2 means the dispersion (variance) of residua set, and n is number of measurements.

The value of R_F – factor is a dimensionless number; if it is less than one half of instrumental error, then the fit is excellent. If the value of Hamilton's factor is higher than instrumental error, then the fit is unsatisfactory.

The normality of distribution of residua is also tested by means of the χ^2-test, where the residua are divided according to their magnitude into classes, and their population in individual classes is tested. One of possible classifications is division into six classes: $\pm 1\sigma, \pm 2\sigma, \pm 3\sigma$ (σ is standard deviation). Application of classical residua is based on the presumption that their distribution corresponds with distribution of errors; hence their properties are identical with the properties of errors. Another presumption is, e.g., the presumption that a greater value of residuum indicates a more strongly influencing point which should be eliminated. However, the residuum $\hat{e}_i$ does not only represent the random component of errors, but it is a linear combination of all sorts of errors ε_i, which can be expressed as follows:

$$\hat{e}_i = (1 - h_{ii})\varepsilon_i - \sum_{j=i}^{n} h_{ij}\varepsilon_j \qquad (2)$$

where h_{ii} are diagonal elements of matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ (where $\boldsymbol{X}$ is matrix of independent variables)

Hence, the distribution of residua depends upon the distribution of errors, upon the elements of matrix of projection matrix *H*, and on the magnitude of selection *n*. In the case of smaller selections, the elements of projection matrix are large, and the predominating role will be played by the summation term in Eq. (2); hence, the distribution of residua will approach the normal distribution even when the distribution of errors is different. However, in the case of sufficiently large selections (when $1/n \approx 0$) it is $\hat{e}_i \approx \varepsilon_i$. The dispersion (variance) of residua is non-constant, as it can be seen from the following equation:

$$D(\hat{e}_i) = (1 - H_{ii})\hat{\sigma}^2 \qquad (3)$$

where $\hat{\sigma}$ is standard deviation of the basic selection.

In addition to that, the residua are inter-correlated even if the errors ε_j and ε_i are independent, which is shown in the following relationship:

$$r_{ij} = \frac{-H_{ij}}{\sqrt{(1 - H_{ii})(1 - H_{jj})}} \qquad (4)$$

For extreme points the diagonal elements of projection matrix *H* approach one, whereas the non-diagonal elements approach zero; hence, when entering values into Eq. (2) we get the estimate of residuum equal to zero regardless of the magnitude of dependent variable.

Thus the classical residua are correlated, have non-constant dispersion, and need not indicate strongly off-lying points. That is why several types of variously transformed residua are introduced, which increases the reliability of analysis of residua [MELOUN, 2004].

2.2 Normalized Residua

The normalization of residua is carried out by dividing them by the estimate of standard deviation of the basic selection. Then the residua $\hat{e}_{Ni}$ would have normal distribution with the mean value equal to zero and dispersion equal to one. Their application is based on the presumption that the values higher than triple the standard deviation are considered to be off-lying. However, from Eq. (2) it follows that the dispersion is neither constant nor equal to one.

2.3 Standardized Residua

These residua have constant dispersion. They are defined as follows:

$$\hat{e}_{Si} = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - H_{ii}}} \qquad (5)$$

As a rule, these residua are sufficient for identification of heteroscedasticity.

2.4 Jackknife Residua

If in Eq. (5) the estimate of standard deviation $\hat{\sigma}$ is replaced by the modified standard deviation $\hat{\sigma}_{(-i)}$ obtained by leaving out the i-th point, the result is fully studentized residua, i.e. jackknife residua $\hat{e}_{Ji}$:

$$\hat{e}_{Ji} = \hat{e}_{Si}\sqrt{\frac{n - m - 1}{n - m - \hat{e}_{Si}^{\,2}}} \qquad (6)$$

where *n – m – 1* is the number of degrees of freedom.

These residua possess the Student distribution and serve—instead of the classical residua—for identification of outliers points. However, they need not be reliable in the case of extremes.

2.5 Predicted Residua

They are defined by the following relationship:

$$\hat{e}_{Pi} = y_i - x_i\mathbf{b}_{(i)} = \frac{\hat{e}_i}{1 - H_{ii}} \qquad (7)$$

where $\mathbf{b}_{(i)}$ are estimates of parameters from all points except the i-th point. These residua are used also for identification of outliers points.

2.6 Recursive Residua

All the residua described so far are inter-correlated. For obtaining the non-inter-correlated residua it is possible to use the recursive method of least squares, i.e. estimates of parameters performed iteratively: one point (one row of matrix *X*) is added in each iteration step. These residua can identify non-stability of the model in time – hence the auto-correlation. The recursive residua are calculated from the following relationship:

$$\hat{e}_{Ri} = 0, \quad for\ i = 1,\dots,m \tag{8}$$

$$\hat{e}_{Ri} = \frac{y_i - x_i \mathbf{b}_{i-1}}{\sqrt{1 + x_i (X_{i-1}^T X_{i-1})^{-1} x_i^T}}, \quad for\ i = m+1,\dots,n \tag{9}$$

where $\mathbf{b}_{i-1}$ are estimates obtained from the first (i – 1) points

X_{i-1} is the matrix containing the first (i - 1) rows of matrix *X*

y_i values of dependent variable

These recursive residua are independent and possess constant dispersion. They are used in tests of normality or stability of regression coefficients.

2.7 Other Characteristics

The elements of matrix ***H*** are very significant in evaluation of influence of individual observations. The diagonal elements h_{ii} have the average magnitude k/n; it varies from zero to one. That is why just this ratio is used for evaluation of the influence of the i-th observation. The observations having h_{ii} more than twice or three times as large as the k/n ratio are much too influential. The Cook distance is used for evaluation of influence of the i-th observation:

$$D_i = \frac{u_i^2 h_{ii}}{k\,(1 - h_{ii})} \tag{10}$$

The quantities D_i have approximately the distribution F with *k* and $n - k$ degrees of freedom. Even more sensitive quantity for finding off-lying observations is the following one:

$$T_i = \sqrt{\frac{(n-k)\,h_{ii}}{k\,(1-h_{ii})}}\ |t_i| \tag{11}$$

which has F distribution with *k* and $n - k$ degrees of freedom.

3 GRAPHICAL METHODS

However, numerical statistical analysis of residua does not possess full information value: numerical characteristics can be acceptable, but they do not reflect, e.g., trends in curve fitting. Therefore, it is sometimes advantageous to use other, graphical tools, where it can be seen at first sight how the residua are dislocated. These tools do not provide exact results, their evaluation is a subjective matter depending on experience, but they are very useful for evaluating the quality of curve fitting and especially in the phase of looking for suitable model.

3.1 Graph of Distribution of Residua

Their magnitudes are plotted in a graph being dislocated around the mean value (i.e. the zero line). Their potential trend can be seen at first sight. It is shown on Fig. 1.

3.2 Graph of Curve Fitting

Even a simple graph of both dependences – the experimental one and the calculated one – possesses high information value: we can judge how the two dependences "match" each other. It is shown on Fig. 3.

3.3 Graphs of Shape of Users Function

In this case it is possible to model the shape of users function either for one parameter or for a selected combination of two parameters (see above) and to judge the possibility of determination of individual parameters. It is shown on Fig. 4.
A similar tool is the so-called contours, ***contour lines of users function*** projected on top view, which also show the shape of the minimum found after completed optimisation. These contours should be circular in shape; if they are elliptical in shape (with a large difference between the axes), then it is clear that the shape of purpose function is distorted due to different interdependences of parameters. It is shown on Fig. 2.

In the next part are presented:

- Graph of Dependence of Residua $\hat{e}_i$ upon Index i

- Graph of Dependence of Residua $\hat{e}_i$ upon Variable x_i

-. Graph of Dependence of Residua $\hat{e}_i$ upon prediction $\hat{y}_i$

If these graphs represent a "cloud" of points, then everything is correct. A shape of sector indicates heteroscedasticity in data. A shape of band in the first two graphs indicates incorrect calculation or an error in model.

3.7 Graphs of Identification of Influential Points [MELOUN, 2004]

Identification of various types of influential points makes use of a number of graphs, which combine various types of residua with elements of projection matrix H_{ii}:

- *graph of predicted residua* (classical residua $\hat{e}_i$ vs. predicted residua $\hat{e}_{Pi}$)

In this graph the extremes lie outside the straight line $y = x$.

- *Williams's graph* (elements of projection matrix H_{ii}:vs. jackknife residua $\hat{e}_{Ji}$).

Here the limit lines for off-lying points are drawn, $y = t_{0,95}(n\text{-}m\text{-}1)$ – which is 95% quantile of Student's distribution with $n - m - 1$ degrees of freedom, and limit lines for extremes x = $2\ m/n$.

- *Pregibon's graph* (elements of projection matrix H_{ii}:vs. squares of normalized residua $\hat{e}_{Ni}$)

In this case, two limit straight lines are drawn in the graph:

$$\begin{aligned} y &= -x + 2(m+1)/n \\ y &= -x + 3(m+1)/n \end{aligned} \tag{12}$$

A point is considerably influential if it lies above the upper straight line, and it is only influential (or also an extreme or outliers point) if it lies between the two straight lines.

- *McCulloh–Meeter's graph* (ln[H$_{ii}$/(m(1-H$_{ii}$)] vs. ln $\hat{e}_{Si}^{\ 2}$)

The limiting line for extremes is:

$$x = \ln\frac{2}{n - 2m} \tag{13}$$

and for outliers points it is:

$$y = \ln[(n-m)t_{0,95}^2(t_{0,95}^2 + n - m)] \tag{14}$$

where $t_{0,95}$ is 95% quantile of Student's distribution with $n -\ m - 1$ degrees of freedom.

Besides the characteristics of residua, most algorithms of non-linear regression provide the values of parameters found together with their respective errors calculated at the end of optimisation. Even the mutual ratio between the parameter value and the error magnitude can easily show the quality of determination of individual parameters, i.e. the parameter error should be lower than the parameter value by at least one order of magnitude.

4 EXAMPLES OF RESULTS

Input of procedure REZID are number of points, matrixes of independent, dependent variables and residuals.
This section shows examples of results obtained from procedure REZID: individual statistical characteristics and some graphical representations in Figs 1-3. Figure 4 has been obtained from the EXCEL program; changing of selected pair of parameters, their varying in the chosen range, and reproduction of three-dimensional picture cannot be included in the procedure (input for EXCEL program are values of parametres only and selected interval their variation).

Table 1 - Classical Statistical Moments:

–0.46132E–04	ARITHMETIC MEAN
0.37509E–02	MEAN DEVIATION
0.52115E–02	STANDARD DEVIATION
0.27159E–04	VARIANCE
0.13886E+01	MOM. COEFF. OF SKEW.
0.42511E+01	MOM. COEFF. OF KURT.

Table 2 - Pearson's χ^2 goodness-of-fit test:

	CLASS LIMITS		PROBABIL.		FREQ.		PART.
	LOW.	HIGH.	CALC	OBS	CALC	OBS	χ^2
1	–0.10E+31	–0.59E–02	0.125	0.03	3.8	1	2.017
2	–0.59E–02	–0.35E–02	0.125	0.26	3.8	8	4.817
3	–0.35E–02	–0.16E–02	0.125	0.16	3.8	5	0.417
4	–0.16E–02	0.00E+00	0.125	0.16	3.8	5	0.417
5	0.00E+00	0.16E–02	0.125	0.16	3.8	5	0.417
6	0.16E–02	0.35E–02	0.125	0.03	3.8	1	2.017
7	0.35E–02	0.59E–02	0.125	0.03	3.8	1	2.017
8	0.59E–02	0.10E+31	0.125	0.13	3.8	4	0.017

OBSERVED CHI-SQUARE IS	12.13
CHI SQUARE (6,0,95) SHOULD BE	12.60

Table 3 - Hamilton's R-factor:

R FACTOR= 0.008546

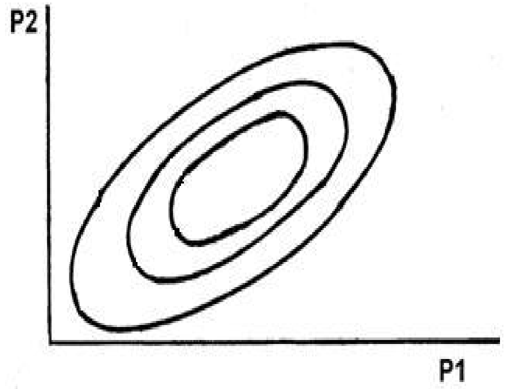

Fig. 2 – Contour lines of purpose function near the minimum

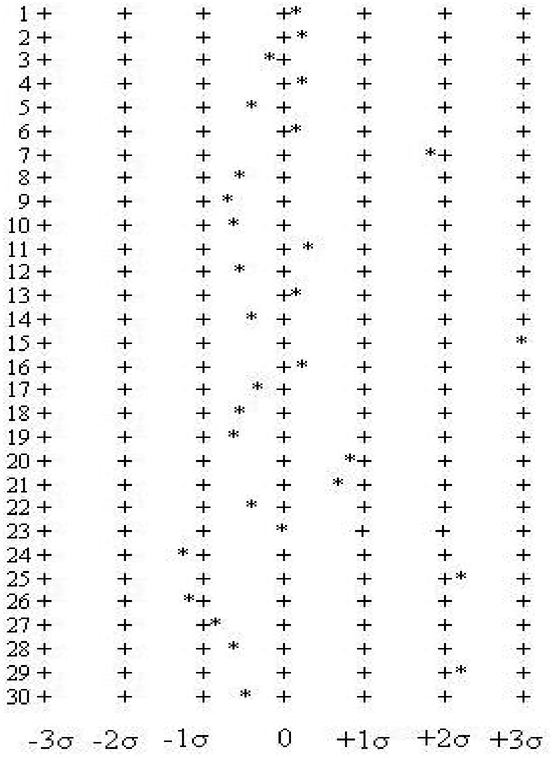

Fig. 1 – Map of residua around zero value

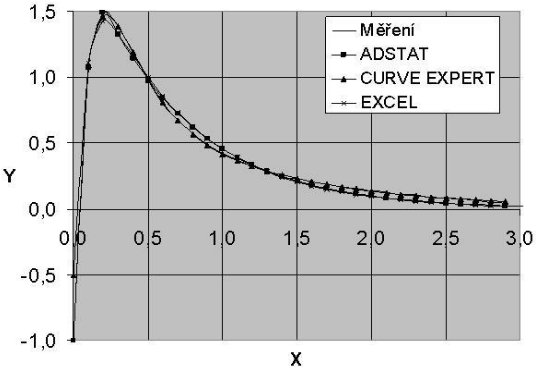

Fig. 3 – Graph of fitting of calculated vs. experimental dependence

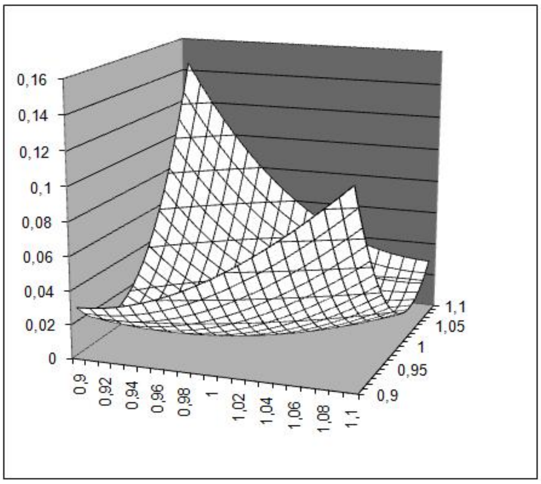

Fig. 4 – Shape of users function near the minimum (selected parameters were varied in the range of 10%)

Table 4 – INDICATION OF INFLUENTIAL POINTS:
(underlined number indicates outliers or influential point)

Point	Standardiz.	Jackknife	Predicted	Diag.
	residuum	residuum	residuum	elements
i	$e_S[i]$	$e_J[i]$	$e_P[i]$	H[i,i]
1	-0.020	-0.019	-0.810	0.088
2	-0.362	-0.354	-14.958	0.086
3	-0.356	-0.349	-14.719	0.085
4	-0.049	-0.048	-2.028	0.084
5	0.165	0.161	6.806	0.082
6	-0.364	-0.356	-14.953	0.075
7	0.277	0.271	11.382	0.075
8	0.022	0.022	0.914	0.071
9	0.038	0.037	1.569	0.069
10	-0.406	-0.398	-16.595	0.065
11	-0.530	-0.521	-21.666	0.063
12	-0.320	-0.313	-12.991	0.052
13	-0.785	-0.778	-31.738	0.042
14	0.831	0.824	33.556	0.042
15	0.235	0.230	9.534	0.047
16	0.272	0.267	11.038	0.048
17	-0.441	-0.433	-17.883	0.048
18	2.771	3.355	114.020	0.077
19	0.982	0.981	40.711	0.090
20	1.926	2.063	80.330	0.101
21	-0.533	-0.525	-22.515	0.122
22	-0.393	-0.386	-16.657	0.128
23	-0.044	-0.043	-1.862	0.132
24	-3.170	-4.201	-142.380	0.225

5 CONCLUSION

A procedure REZID has been assembled, which can be included into source programs written in FORTRAN 77 language. It performs a complex analysis of set of residua, and is often omitted in original programs. Analysis of residua is very important not only for judging adequacy of regression model suggested but also for verification of accuracy and correctness of regression calculations. Transfer of data between the modules is carried out in the form of formal parameters. After modification, the procedure could be used also with the newest version of SPONA program [LUKŠAN, 2008].

It is intended to rewrite the procedure in some of the more modern programming languages (Visual Basic) as an independent program, which will be able to perform statistical analysis of any external data set.

The problem was dealt with in the framework of research project MŠM 0021627505 "Control, optimizing and diagnostics of complex systems".

6 REFERENCES

JAVŮREK, M.; TAUFER, I. (2006). *Don't be afraid of non-linear regression* (in Czech) (1), *CHEMagazín*, 2 (XVI), pp. 27-29, ISSN 1210-7409.

LUKŠAN, L.; ZAMAZAL, M.; KOČKOVÁ, S. (1983). *Research report V-125: SPONA-82, user's description of set of programs for optimisation and non-linear approximation* (in Czech). Prague: SVT ČSAV, 125 p.

LUKŠAN, L. (1987). *Research report V-276: Notes on application of set of optimisation programs SPONA* (in Czech). Prague: SVT ČSAV, , 40 p.

LUKŠAN, L. et. al. (2008). *Research Report V-1040: UFO 2008 – Interactive System for Universal Function Optimisation.*, Prague: Czech Academy for Sciences, , 274 pp.

MELOUN, M.; JAVŮREK, M. (1986), *Chemometrics I. –Applications of computers in analytical and physical chemistry* (in Czech). Pardubice: VŠCHT, 290 p.

MELOUN, M.; MILITKÝ, J. (2004). *Statistical analysis of experimental data* (in Czech). Prague: Academia, 953 p., ISBN 80-200-1254-0.