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

PROCEEDINGS

of the 18th International Conference on Process Control

Hotel Titris, Tatranská Lomnica, Slovakia, June 14 – 17, 2011

ISBN 978-80-227-3517-9

http://www.kirp.chtf.stuba.sk/pc11

Editors: M. Fikar and M. Kvasnica

Javůrek, M., Taufer, I.: Tests of Various Types of Residuals in Regression Diagnostics, Editors: Fikar, M., Kvasnica, M., In *Proceedings of the 18th International Conference on Process Control*, Tatranská Lomnica, Slovakia, 373–377, 2011.

Full paper online: http://www.kirp.chtf.stuba.sk/pc11/data/abstracts/066.html

Tests of Various Types of Residuals in Regression Diagnostics

M. Javůrek I. Taufer

Department of Process Control, Faculty of Electrical Engineering and Informatics,
The University of Pardubice, Studentská 95, 532 10 Pardubice.
e-mail: Milan.Javurek@upce.cz; Ivan.Taufer@upce.cz;

Abstract: Approximation of experimental data by means of an analytical or general mathematical dependence is performed most frequently by the regression method using the least squares approach. The quality of curve fitting is evaluated on the basis of analysis of resulting set of residuals which, however, can be defined in various ways. This paper deals with suitability tests of the individual types from the standpoint of curve fitting quality of the regression dependence.

1. INTRODUCTION

Regression is one of most common and most favored approximation methods of experimental dependences. The principle consists in optimization (i.e. minimization) of the users function, most often in the form of the least squares, which expresses closeness of curve fitting of the regression and experimental dependence. The function fitted may be known in analytical form, where the parameters have direct physical meaning, or various types of mathematical dependences are used. The basic classification of regression methods is done according to the parameters of fitted dependences, i.e. linear regression and non-linear regression are known. While the linear regression is evaluated according to univocal formulas, the so-called normal equations, the course and results of non-linear regression are affected by a number of factors, such as initial assessment of parameters, the adopted optimizing method, interdependence of individual parameters etc. Therefore, the parameters found can be neither correct nor accurate, particularly in the cases where even their approximate values are unknown. The non-linear regression offers relatively few tools for verification of the found parameters. If several calculations are performed with different models, their comparison is carried out by means of Akaike Information Criterion (AIC), mean quadratic error of prediction (MEP), the value of users function (residual-square-sum, RSC) etc. (MELOUN, 2011). However, if we have only a single calculation, then one of the few available tools is the analysis of set of residuals. It is quite paradoxical that most PC programs (even the commercial ones. such as STATISTICA) do not include this analysis; only in algorithmic regime they provide the values of parameters with their standard deviations. No further verification is performed, and the curve fitting quality of regression dependence and experimental dependence cannot be evaluated.

If the conditions of application of regression method are fulfilled (the data do not exhibit heteroscedasticity, supernormality, multicollinearity, autocorrelation, outliers, the model is significant), then the set of residuals should exhibit normal distribution, which can be proved on the basis of calculated values of central moments, Pearson's test, sign tests and other tests. However, residuals can be defined in various ways, and their information abilities differ.

2. DEFINITION OF VARIOUS TYPES OF RESIDUALS

2.1 Classical Residuals $\hat{e}_i$ (MELOUN, 2011)

These residuals are defined as a difference between the calculated values and the experimental ones. They are correlated, do not have constant dispersion, and they need not correctly indicate deviated points.

2.2 Normalized Residuals $\hat{e}_{Ni}$ (MELOUN, 2011)

In this case the normalization consists in division of the value of classical residual by the value of standard deviation of the whole set. The set of residuals should have normal distribution with zero mean value and standard deviation is equal to one. The values higher/lower than the triple of standard deviations are considered as outliers.

However, the mathematical analysis shows that the dispersion $D(\hat{e}_{Ni}) = (1\text{-}H_{ii})$ is neither constant nor unit, so the recommended elimination of the values exceeding the interval of the triple of standard deviation need not be correct.

2.3 Standardized Residuals $\hat{e}_{Si}$ (MELOUN, 2011)

They also should exhibit normal distribution with constant dispersion; they are defined as follows:

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

where $\hat{\sigma}$ stands for standard deviation

H_{ii} are diagonal elements of projection matrix

Their properties are almost identical with the classical ones.

2.4 Jack-Knife Residuals $\hat{e}_{Ji}$ (MELOUN, 2011)

If in Eq. (1) we use, instead the overall standard deviation, its estimate obtained with omitting of the i-th point:

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

where n stands for the number of measurements

m is the number of parameters determined

Under the assumption of normality of errors, these residuals exhibit Student distribution with n-m-1 degrees of freedom. These residuals are used for identification of outliers points.

2.5 Predicted Residuals $\hat{e}_{Pi}$ (MELOUN, 2011)

These are defined as follows:

$$\hat{e}_{Pi} = y_i - x_i\,\mathbf{b}_{(i)} = \frac{\hat{e}_i}{1-H_{ii}} \qquad (3)$$

where x is/are independent variable(s) and y is the dependent variable quantity

$\mathbf{b}_{(i)}$ are estimates of parameters obtained by the least squares method from all points except the i-th point

3. OTHER DIAGNOSTIC TOOLS

3.1 Cook's Distance D_i (MELOUN, 2011)

This, in fact, is the Euclidean distance between the vector of prediction of independent variable obtained by the least squares method and the same vector obtained with elimination of the i-th point. Cook's distance expresses the effects of the i-th point upon the estimates of parameters only. It is defined as follows:

$$D_i = \frac{\hat{e}_{Si}}{m}\frac{H_{ii}}{1-H_{ii}} \qquad (4)$$

3.2 Atkinson's Distance A_i (MELOUN, 2011)

This is used in order to increase the sensitivity of regression to extreme points. It is defined as follows:

$$A_i = |\hat{e}_{Ji}|\sqrt{\frac{n-m}{m}\frac{H_{ii}}{1-H_{ii}}} \qquad (5)$$

3.3 Distances of Likelihood

This quantity is the difference of logarithms of credibility function using all point and that obtained with elimination of the i-th point. If its value is higher than the quantile $\chi^2_{1-\alpha}(m+1)$ of distribution, then the given point is considered as influential.

3.4 Summary Characteristics of Properties of Whole Set of Residuals

The following characteristics have also been used for determination of validity of the basic presumptions of application of the least squares method from the whole set of residuals (the classical residuals were always used):

- Cook–Weisberg's test of heteroscedasticity
- Jarque–Berr's test of normality
- Wald's test of autocorrelation
- Sign test

Description of these tests is somewhat complicated and time-consuming; therefore, see Ref. (MELOUN, 2011).

Original data			
Type of residuals	Classical	Normalized	Standard.
Users function	0.0030	1.7373	29.9600
Arithm. mean	0.0000	-0.0430	0.3816
Stand. deviation	0.0101	0.2368	0.9236
Mom. coeff. of skew.	-0.2084	-3.0775	-1.2577
Mom. coeff. of curt.	0.0298	11.6552	2.3077
R-factor	0.0006	0.0136	0.0563
Type of residuals	Jack-Knife	Predicted	
Users function	32.2588	0.0035	
Arithm. mean	-0.0058	0.0000	
Stand. deviation	1.0369	0.0108	
Mom. coeff. of skew.	-0.2908	-0.1922	
Mom. coeff. of curt.	0.3215	-0.0634	
R-factor	0.0584	0.0006	
Other criteria			
AIC	-227.0000		
MEP	0.0001		
Heteroscedasticity	yes		
Normality	yes		
Autocorrelation	not		
Sign test	negat.		

Table 1. Characteristics of set of residuals of starting data

The above-mentioned diagnostic tools (except for 3.4) are ordinarily used only for detection of significant (extreme, outliers) points, which from the standpoint of overall view of the quality of curve fitting of regression model is of not very high significance. If the model is not suitable, then the search for these points has no meaning either. In most cases, the closeness of curve fitting is evaluated by means of classical residuals, but their information effectiveness is very small; the calculation of statistical moments expresses neither the outliers points not the trends in curve fitting. In this case, the tests given in section 3.4 come in useful and, furthermore, testing of suitability of the whole model by means of F-test, or Student's test of significance of the individual parameters. The two last mentioned tests have not been taken into account in this paper, since the model was known. Also Hamilton's R-factor has very good information ability:

$$R = \sqrt{\frac{RSC}{\sum y_i^2}} \qquad (6)$$

Classical residuals			
First point			
Multiple of characteristics of original set	1. point +1 s	1. point + 2s	1. point + 3s
Users function	0.9596	3.1891	11.1415
Arithmetic mean	–0.8662	2.4045	5.1910
St. deviation	0.9796	1.7858	3.3903
M. coeff. of skew.	1.2973	–12.1966	–19.1828
M. coeff. of curt.	4.7594	350.2452	655.3301
R-factor	0.9796	1.7858	3.3378
Other criteria			
AIC	–273.2000	–237.2000	–199.7000
MEP	0.0001	0.0004	0.0014
Heteroscedasticity	yes	yes	yes
Normality	yes	yes	not
Autocorrelation	not	not	not
Sign test	negat.	negat.	negat.
Second point			
Multiple of characteristics of original set	1. point +1s, 2. point -1s	1. point +2s, 2. point -2s	1. point +3s, 2. point -3s
Users function	1.0754	3.6223	12.1170
Arithmetic mean	0.0000	2.8989	0.6816
St. deviation	1.0370	1.9032	3.4809
M. coeff. of skew.	2.0439	–9.8283	–17.4348
M. coeff. of curt.	7.9841	317.6662	606.9347
R-factor	1.0370	1.9032	3.4809
Other criteria			
AIC	–269.8000	–233.4000	–197.1000
MEP	0.0001	0.0005	0.0016
Heteroscedasticity	yes	yes	yes
Normality	yes	yes	not
Autocorrelation	not	not	not
Sign test	negat.	negat.	negat.

Table 2. Various variants of calculation for classical residuals

Normalized residuals			
First point			
Multiple of characteristics of original set	1. point +1 s	1. point + 2s	1. point + 3s
Users function	0.4125	29.8402	112.5996
Arithmetic mean	0.4781	–3.6503	–7.2397
St. deviation	0.6469	5.5122	10.8850
M. coeff. of skew.	0.7344	–1.6724	–1.6877
M. coeff. of curt.	0.8122	2.3804	2.4561
R-factor	0.6422	5.4626	10.6110
Other criteria			
AIC	–273.2000	–237.2000	–199.7000
MEP	0.0001	0.0004	0.0014
Heteroscedasticity	yes	yes	yes
Normality	yes	yes	not
Autocorrelation	not	not	not
Sign test	negat.	negat.	negat.
Second point			
Multiple of characteristics of original set	1. point +1s, 2. point -1s	1. point +2s, 2. point -2s	1. point +3s, 2. point -3s
Users function	0.8572	33.3538	121.3035
Arithmetic mean	0.6204	–3.3425	–6.7443
St. deviation	0.9342	5.8382	11.1266
M. coeff. of skew.	1.2930	–1.5192	–1.6224
M. coeff. of curt.	1.6985	2.1565	2.3071
R-factor	0.9259	5.7752	11.0136
Other criteria			
AIC	–269.8000	–233.4000	–197.1000
MEP	0.0001	0.0005	0.0016
Heteroscedasticity	yes	yes	yes
Normality	yes	yes	not
Autocorrelation	not	not	not
Sign test	negat.	negat.	negat.

Table 3. Various variants of calculation for normalized residuals

If the value of R-factor is not higher than the uncertainty of measurement, then the curve fitting can be considered as a good one. However, the use of classical residuals could be replaced by another type, which is more sensitive and better reflects the deviation in the fitting of regression dependence. Hence in the subsequent text we will try to select the more suitable type out of the available types of residuals.

4. TESTING

The testing of individual types of residuals was performed on a simulated example of linear dependence. The values of dependent variable were calculated for the given values of independent variable and given parameters of straight line. The calculated values were loaded with errors exhibiting normal distribution. The obtained data were evaluated by linear regression, and the central moments of the set of residuals were calculated. In the next step, the value of dependent variable of the first point was increased stepwise by adding one, two and three multiples of standard deviation from the first set, and again the individual sets of data were evaluated by linear regression, and central moments of the sets of residuals were calculated. These values were referenced to the values of central moments of the original set – i.e. to find out to what extent the change of one point will make itself felt in the change of characteristics of the set of residuals. In subsequent step, also changes of the second point were added to the changes of the first point in the opposite direction as compared with the first point.

Apart from the tests described, we also carried out calculations with the sets in which the described changes had been realized in central part of the dependence; however, this

Standardized residuals			
First point			
Multiple of characteristics of original set	1. point +1 s	1. point + 2s	1. point + 3s
Users function	17.2142	1.0464	1.0595
Arithmetic mean	–0.0003	0.0145	0.0165
St. deviation	1.0810	1.1068	1.1312
M. coeff. of skew.	0.2081	–2.0917	–3.2168
M. coeff. of curt.	0.0453	4.7456	8.6059
R-factor	0.9991	1.0229	1.0293
Other criteria			
AIC	–273.2000	–237.2000	–199.7000
MEP	0.0001	0.0004	0.0014
Heteroscedasticity	yes	yes	yes
Normality	yes	yes	not
Autocorrelation	not	not	not
Sign test	negat.	negat.	negat.
Second point			
Multiple of characteristics of original set	1. point +1s, 2. point -1s	1. point +2s, 2. point -2s	1. point +3s, 2. point -3s
Users function	1.0034	1.0487	1.0606
Arithmetic mean	–0.0024	0.0112	0.0139
St. deviation	1.0838	1.1080	1.1143
M. coeff. of skew.	0.3467	–1.6840	–2.9227
M. coeff. of curt.	0.1070	4.2868	7.9575
R-factor	1.0017	1.0241	1.0298
Other criteria			
AIC	–269.8000	–233.4000	–197.1000
MEP	0.0001	0.0005	0.0016
Heteroscedasticity	yes	yes	yes
Normality	yes	yes	not
Autocorrelation	not	not	not
Sign test	negat.	negat.	negat.

Table 4. Various variants of calculation for standardized residuals

Jack-Knife residuals			
First point			
Multiple of characteristics of original set	1. point +1 s	1. point + 2s	1. point + 3s
Users function	1.0026	2.3131	9.0801
Arithmetic mean	1.0198	–21.0248	–70.0348
St. deviation	1.0013	1.5164	3.0392
M. coeff. of skew.	1.2606	–14.4374	–18.0003
M. coeff. of curt.	1.4778	64.8356	89.5911
R-factor	1.0013	1.5209	3.0133
Other criteria			
AIC	–273.2000	–237.2000	–199.7000
MEP	0.0001	0.0004	0.0014
Heteroscedasticity	yes	yes	yes
Normality	yes	yes	not
Autocorrelation	not	not	not
Second point			
Sign test	negat.	negat.	negat.
Multiple of characteristics of original set	1. point +1s, 2. point -1s	1. point +2s, 2. point -2s	1. point +3s, 2. point -3s
Users function	1.0122	2.0193	6.2412
Arithmetic mean	1.7396	–16.2757	–52.5175
St. deviation	1.0060	1.4182	2.4812
M. coeff. of skew.	1.8302	–12.6178	–17.8036
M. coeff. of curt.	1.6265	56.6179	86.6017
R-factor	1.0061	1.4210	2.4982
Other criteria			
AIC	–269.8000	–233.4000	–197.1000
MEP	0.0001	0.0005	0.0016
Heteroscedasticity	yes	yes	yes
Normality	yes	yes	not
Autocorrelation	not	not	not
Sign test	negat.	negat.	negat.

Table 5. Various variants of calculation for Jack-Knife residuals

almost did not make itself felt in the evaluation of the regression process, hence it was not tested any more. This fact is connected with different interdependence of parameters throughout the course of regression dependence, e.g., see Ref. (MELOUN, 1984).

5. CONCLUSION

The tests performed unambiguously show that the normalized residuals (Tab. 3) are most suitable for evaluation of closeness of fit. The performed changes in analyzed data are most clearly reflected by the statistical characteristics of the sets of normalized residuals. Of course, these changes apply to such criteria as are the value of users function, the first and the second central moment and the R-factor. However, the parameters characterizing the form of probability distribution do not substantially change, which again speaks in favor of this type of residuals.

Similarly it is possible to evaluate the predicted residuals (Tab. 6), but in this case the variability is lower than that in the case of normalized residuals.

The Jack-Knife residuals (Tab. 5) can be placed behind the predicted residuals: the variability was still lower here. However, this type is very useful for guessing of significant points.

In the case of classical residuals (Tab. 2) the first group of criteria changes only little – hence it is problematic to evaluate changes of fitting – and the second group (the 3rd and the 4th central moments) are changed very markedly, which means that this type is absolutely unsuitable for evaluation of the quality of fitting.

Predicted residuals			
First point			
Multiple of characteristics of original set	1. point +1 s	1. point + 2s	1. point + 3s
Users function	0.9561	3.4940	12.5057
Arithmetic mean	0.0956	–8.0159	–17.0309
St. deviation	0.9778	1.8691	3.5919
M. coeff. of skew.	1.3179	–14.1426	–21.2923
M. coeff. of curt.	–1.0926	–181.0766	–318.7822
R-factor	0.9778	1.8692	3.5363
Other criteria			
AIC	–273.2000	–237.2000	–199.7000
MEP	0.0001	0.0004	0.0014
Heteroscedasticity	yes	yes	yes
Normality	yes	yes	not
Autocorrelation	not	not	not
Sign test	negat.	negat.	negat.
Second point			
Multiple of characteristics of original set	1. point +1s, 2. point -1s	1. point +2s, 2. point -2s	1. point +3s, 2. point -3s
Users function	1.0829	3.9852	13.6282
Arithmetic mean	0.8011	–6.6097	–14.9265
St. deviation	1.0406	1.9962	3.6915
M. coeff. of skew.	2.3294	–11.3779	–19.3415
M. coeff. of curt.	–4.1581	–162.9381	–294.3271
R-factor	1.0406	1.9963	3.6916
Other criteria			
AIC	–269.8000	–233.4000	–197.1000
MEP	0.0001	0.0005	0.0016
Heteroscedasticity	yes	yes	yes
Normality	yes	yes	not
Autocorrelation	not	not	not
Sign test	negat.	negat.	negat.

Table 6. Various variants of calculation for predicted residuals

Variability was almost absent in the case of standardized residuals (Tab. 4): hence their application is utterly meaningless.

The values of criteria for comparison of the quality of ftting between the individual sets (i.e. AIC and MEP) clearly reproduce the worsening conditions of the calculation. However, this is the only piece of information obtained from these characteristics. Interestingly, the values of both AIC and MEP are better for the first two variants of calculation as compared with the basic set – this is due to the fact that the first point in the basic set has a lower experimental value of dependent variable as compared with the predicted value; therefore, during the changes the regression improves at the beginning.

The remaining characteristics (the tests of heteroscedasticity, distribution normality, autocorrelation, and the sign test) possess relatively low information ability, and their application can only be tentative. They can be successfully replaced by a map of distribution of residuals around the zero value, where the trends such as heteroscedasticity, normality, and autocorrelation or sign alternation can be evaluated much more objectively by mere inspection.

The problem has been dealt with in the framework of the research project MŠM 0021627505 „Control, optimization and diagnostics of complex systems“.

REFERENCES

Meloun, M., Javůrek, M. Multiparametric Curve Fitting VIII. The Reliability of Dissociation Constants Estimated by Analysis of Absorbance-pH Curves. *Talanta 32(10),* (1985), pp. 973-986, ISSN 0039-9140

Meloun, M., Militký, J. *Statistical Data Analysis.*Cambridge, UK, Woodhead Publishing, Ltd. 2011, ISBN 978-0-85709-109-3, 900p.