USE OF ROBUST CHARACTERISTICS FOR METALLURGICAL PROCESS MODELLING

Tošenovský,F., Tošenovský,J., Smajdorová,T., Growková,V.

VŠB-Technical University of Ostrava; Dept. of Quality Management.

VŠB-Technical University of Ostrava; Dept. of Quality Management.

Industrial Systems Management at VŠB-Technical University of Ostrava.

filip.tosenovsky@vsb.cz

Abstract

The paper deals with problems of modelling technological processes in cases when the available data sample is of small size and contains outliers. The essence of the method used in the paper lies in robust quantitative characteristics. For the purposes of the paper, data are simulated and processed both the standard and nonstandard way using robust characteristics. Both approaches are then compared. The characteristics that are modelled involve a measure of central tendency of a quality variable Y and a characteristic of variability of that variable.

Keywords: Design of experiments, regression function, robust characteristics

1. INTRODUCTION

When monitoring and measuring parameters of technological processes, it happens fairly often that the process values observed contain significant mistakes. Later, when the values are used again for modelling the processes, all subsequent and related calculations are affected by the mistakes.

The aim of this study is to try robust characteristics in seeking models for technological processes, and compare the quality of the models with those found by standard procedures. The motivation behind this study can be found in [1] – [6]. To be able to make the comparisons, the same type of regression function is selected, however, the model describing dependence of dispersion on process inputs will not be selected, but calculated from a regression function. We generated quintets of numbers from a normal distribution for the purpose of the study, the whole experiment being done once. The use of robust characteristics aims to verify resistance of analytical results based on small samples containing outliers. The knowledge of sought-after functions enables comparison with the functions found. The sum of least squares was used as a criterion for comparing standard and robust models.

2. EXPERIMENTAL PLAN

As part of the study, an experimental plan was constructed for a complete quadratic model with three regressors, all of which were observed at three levels. Thus, the plan has 3^3 = 27 runs (see Table 1).

Table 1 Experimental plan and generated data

	Process inputs			Process outputs				
Run	x_1	x_2	x_3	Y_1	Y_2	Y_3	Y_4	Y_5
1	-1	-1	-1	45,0120	71,5225	54,3656	68,1386	74,6080
2	0	-1	-1	71,2046	52,8104	58,2020	59,5089	52,5769
3	1	-1	-1	75,3637	75,2404	66,9159	64,8753	71,1094
4	-1	0	-1	61,9194	69,1429	59,3658	50,9689	56,1180
5	0	0	-1	55,4977	58,8705	60,1308	50,4295	49,3237
6	1	0	-1	54,7455	48,9226	55,7903	51,7093	56,1756
7	-1	1	-1	52,9605	48,5435	72,6555	71,2942	75,8098
8	0	1	-1	66,8004	50,3633	63,4062	62,3914	42,5675
9	1	1	-1	66,6478	60,1027	71,3065	69,3365	65,2846
10	-1	-1	0	64,7836	63,9936	67,3676	67,3512	55,7784
11	0	-1	0	61,3162	57,9483	58,3493	52,7401	53,6739
12	1	-1	0	55,1722	62,3910	38,7444	62,7431	61,4871
13	-1	0	0	51,2170	57,6184	55,5612	55,4515	60,0692
14	0	0	0	50,0000	50,0000	50,0000	50,0000	50,0000
15	1	0	0	54,6788	56,8304	62,8478	48,3562	54,6208
16	-1	1	0	66,5299	68,2293	65,6057	53,4811	56,4721
17	0	1	0	50,4406	53,5017	55,0014	52,6933	55,6412
18	1	1	0	71,5934	50,1567	69,2449	63,3159	59,9212
19	-1	-1	1	69,7902	50,3752	58,1274	65,3885	63,4676
20	0	-1	1	58,3471	51,5335	71,8556	58,8780	65,9212
21	1	-1	1	54,7596	61,4692	63,5439	61,0810	45,9344
22	-1	0	1	52,0973	56,4063	56,1415	56,4025	54,5731
23	0	0	1	62,9214	55,3193	52,7714	55,7263	63,8747
24	1	0	1	57,9675	61,5892	52,3193	62,6902	70,9372
25	-1	1	1	57,3997	60,7605	69,4416	64,3344	64,0586
26	0	1	1	56,2460	49,5037	60,1052	57,2257	61,8005
27	1	1	1	43,6813	65,2412	63,7289	79,5226	64,5544

3. THEORETICAL QUANTITATIVE CHARACTERISTICS

Unlike the classical procedure, in which a regression function is explored that would describe properly the available data, the procedure to be presented is reversed: a regression function $m(y) = 50 + 5x_1^2 + 5x_2^2 + 5x_3^2$ was selected and corresponding data were generated. The reversed procedure enables one to compare the calculated, or estimated, and known model coefficients. We also calculated the corresponding function depicting the dependence of variance of Y on inputs $s^2(y) = 25x_1^2 + 25x_2^2 + 25x_3^2$. Now, replacing the variables x_1, x_2 and x_3 with their values from the experimental plan, theoretical averages m(y) and variances $s^2(y)$ of Y were obtained (Table 2).

Table 2 Theoretical quantitative characteristics for each experimental run

Run	1	2	3	4	5	6	7
m(y)	65	60	65	60	55	60	65
$s^2(y)$	75	50	75	50	25	50	75

Run	8	9	10	11	12	13	14
m(y)	60	65	60	55	60	55	50
$s^2(y)$	50	75	50	25	50	25	0

Run	15	16	17	18	19	20	21
m(y)	55	60	55	60	65	60	65
$s^2(y)$	25	50	25	50	75	50	75

Run	22	23	24	25	26	27
m(y)	60	55	60	65	60	65
$s^2(y)$	50	25	50	75	50	75

4. EMPIRICAL QUANTITATIVE CHARACTERISTICS

In the next step, five random numbers $Y_1, \dots, Y_5$ were generated from the distribution N[m(y), $s^2(y)$] for each pair of parameters m(y) and $s^2(y)$ from the table, and the following characteristics were calculated for the generated data: moment characteristics - mean and variance; robust quantile characteristics - median M, median of median deviations MAD and interquartile range QR. The resulting characteristics are in Table 3.

Table 3 Empirical quantitative characteristics

	Characteristics from generated data				
	Mean	Variance	Median	MAD	QR
1	62,7293	157,8847	68,1386	6,4694	17,1569
2	58,8606	57,3296	58,2020	5,3916	6,6985
3	70,7009	22,6949	71,1094	4,1935	8,3245
4	59,5030	45,7686	59,3658	3,2478	5,8014
5	54,8504	23,6379	55,4977	4,6331	8,4410
6	53,4687	9,5275	54,7455	1,4301	4,0810
7	64,2527	157,0123	71,2942	4,5156	19,6950
8	57,1058	104,6103	62,3914	4,4090	13,0429
9	66,5356	18,3916	66,6478	2,6887	4,0519
10	63,8549	22,6687	64,7836	2,5676	3,3576
11	56,8056	12,5925	57,9483	3,3679	4,6754
12	56,1076	103,7014	61,4871	1,2560	7,2188
13	55,9835	10,6367	55,5612	2,0572	2,1669
14	50,0000	0,0000	50,0000	0,0000	0,0000
15	55,4668	27,0590	54,6788	2,1516	2,2096
16	62,0636	43,8586	65,6057	2,6236	10,0578
17	53,4556	4,2099	53,5017	1,4997	2,3081

18	62,8464	71,8141	63,3159	5,9290	9,3237
19	61,4298	55,7076	63,4676	5,3402	7,2611
20	61,3071	60,6865	58,8780	7,0432	7,5741
21	57,3576	51,5696	61,0810	2,4629	6,7096
22	55,1241	3,4446	56,1415	0,2648	1,8294
23	58,1226	24,5878	55,7263	2,9549	7,6021
24	61,1007	46,6130	61,5892	3,6217	4,7227
25	63,1990	20,1441	64,0586	3,2981	3,5739
26	56,9762	22,3746	57,2257	2,8795	3,8592
27	63,3457	163,3954	64,5544	0,8255	1,5123

5. REGRESSION MODELS

Regression functions were found for the following characteristics of Table 3:

a) mean and median
b) variance and MAD (median of absolute deviations of the generated values Y_i from their median M)
c) variance and QR
d) mean and median from data containing outliers
e) variance and MAD from data containing outliers.

Pairs of models were compared using the criterion $S_e = \sum_i e_i^2$; also, the extent of concordance between the empirical and theoretical coefficients for the known functions m(y), $s^2(y)$ was analyzed.

Ad a) Characteristics of the regression function for
- *mean*

i	bi	s(bi)	Ti	p-val
b_o	51,03538	1,597271	31,95162	1,47E-20
b_{11}	4,450117	1,280666	3,474847	0,00205
b_{22}	5,09413	1,280666	3,977721	0,000595
b_{33}	2,93344	1,280666	2,290559	0,031487

- *median*

i	bi	s(bi)	Ti	p-val
bo	49,80312	1,453967	34,25328	3,05E-21
b_{11}	6,049094	1,165767	5,188941	2,92E-05
b_{22}	7,059944	1,165767	6,056052	3,55E-06
b_{33}	3,130561	1,165767	2,68541	0,013209

Comparison:

All the coefficients for the model of mean and median are nonzero, although the p-values are smaller, and thus more convincing for the latter model. In both cases, the estimate of the absolute term is close to the theoretical value bo = 50. Other estimated parameters are not far from their theoretical counterpart 5 either, given how small the generated sample is. Further, for the median, Se = 187,54, whereas it is slightly higher for the mean: Se = 226,33. Using the model for median doesn't bring a striking improvement.

Ad b) and c) Characteristics of the regression function for

- *variance* (Se = 42514,28)

i	bi	s(bi)	Ti	p-val
b_{11}	18,27161	14,52186	1,258214	0,220418
b_{22}	38,06396	14,52186	2,62115	0,01497
b_{33}	20,51964	14,52186	1,413018	0,170489

- *MAD* (Se = 9754,61)

i	bi	s(bi)	Ti	p-val
b_{11}	27,39329	6,956002	3,938079	0,000616
b_{22}	28,40414	6,956002	4,0834	0,000427
b_{33}	24,47475	6,956002	3,518509	0,00176

Comparison:

The criterion Se is much smaller for the MAD regression. Regarding the variance, only the coefficient b_{22} is nonzero, and the coefficient estimates deviate more from the expected value 25, as compared to the case of MAD. As for the MAD case, all coefficients are nonzero. Using MAD is much better here, compared to the model of variance.

- *interquartile range* QR = $X_{75} – X_{25}$ (Se = 443,15)

i	bi	s(bi)	Ti	p-val
b_{11}	1,395069	1,482632	0,940941	0,35611
b_{22}	4,286269	1,482632	2,890986	0,008028
b_{33}	3,542302	1,482632	2,389198	0,025096

Comparison:

The coefficients b_{22} and b_{33} are statistically significant (p-value is below 0,05), but they are a poor estimate of the corresponding theoretical value, although Se seems better for this case than for the case of MAD.

Ad d) Characteristics of the regression function for mean and median calculated from data with outliers.

The first value of Table 1 is Y_1 = 45,012 (red); for further calculations, the decimal point will be shifted by one order to create the outlier Y_1 = 450,12. The quality of regression functions for mean and median will then be compared.

For the mean, (Se = 24660,32)

i	bi	s(bi)	Ti	p-val
b_o	59,86993	16,67257	3,590924	0,001544
b_{11}	5,324207	13,3678	0,398286	0,694092
b_{22}	5,96822	13,3678	0,446463	0,659438
b_{33}	3,80753	13,3678	0,284829	0,778324

and for the median, (Se = 210,072)

i	bi	s(bi)	Ti	p-val
b_o	49,50159	1,538818	32,16857	1,26E-20
b_{11}	6,275244	1,233799	5,086114	3,77E-05
b_{22}	7,286094	1,233799	5,905413	5,09E-06
b_{33}	3,356711	1,233799	2,72063	0,012194

Comparison:

For the mean, the coefficients b_{ii} are statistically insignificant, and so the model cannot be used. For the median, all the coefficients b_{ii} are nonzero and quite close to the theoretical values (given the small size of the sample). The use of the median gives better results than the mean.

Ad e) Characteristics of the regression function for variance and MAD calculated from data containing outliers.

As in d), the first value Y_1 = 45,012 was adjusted for the subsequent calculations: the decimal point was shifted by an order to create the outlier Y_1 = 450,12. The quality of the regression functions for variance and MAD shall be compared now.

For the variance, we have

i	bi	s(bi)	Ti	p-val
b_{11}	714,314	1989,993	0,358953	0,722769
b_{22}	734,1063	1989,993	0,368899	0,715437
b_{33}	716,562	1989,993	0,360083	0,721935

and for the MAD,

i	bi	s(bi)	Ti	p-val
b_{11}	27,47386	6,930467	3,964214	0,000577
b_{22}	28,48471	6,930467	4,11007	0,000399
b_{33}	24,55532	6,930467	3,543098	0,001656

Comparison:

For the case of variance, all coefficients are statistically insignificant (p-value is around 0,7), whereas for MAD, all the coefficients are significant (p-value is close to zero) and close to the theoretical values b_{ii} = 25. Se = 9683,126 for MAD; for variance, Se is far higher. MAD seems much better for modelling purposes.

CONCLUSION

The aim of the paper was to examine how to use selected robust characteristics when searching for regression functions for small data samples containing outliers. For a known regression function describing behaviour of a variable Y and its variability, a small-size sample was simulated. The data were also burdened with outliers, and a regression model for the central tendency of Y and its variability was searched. The classical procedure and the procedure with robust characteristics were compared, using the sum of squares, as well as the extent of concordance between the regression coefficients estimated and known. The simulation has shown that some robust characteristics seem better for the modelling purposes than their standard counterparts.

AKNOWLEDGEMENT

This paper was prepared under the specific research project No. SP2015/112 conducted at the Faculty of Metallurgy and Materials Engineering, VŠB-TU Ostrava with the support of Ministry of Education of the Czech Republic.

REFERENCES

[1] Chanseok, P., Byung, R. C.: Development of robust design under contaminated and non-normal data.QE Vol.15, No.3, 2003.

[2] Myers,R.H., Montgomery,D.C.: Response Surface Methodology,2nd Ed., Wiley,NY 2002.

[3] Zgodavová,K., Bober,P.: An Innovative Approach to the integrated Management System Development: SIMPRO-IMS Web-Based Environment, Quality, Innovation, Prosperity, Vol. XVI/2-2012, p. 59-70, ISSN: 1335-1745, DOI: 10.12776/qip.v16i2.69.

[4] Caridad, J.M., Caridad,D.L.: Estadística básica e introducción a la econometría. UCO Córdoba 2005. ISBN 84-95723-07-7.

[5] Zgodavova, K., Slimak, I. 2011: Focus on Success. Quality Innovation Prosperity. Vol 15, No 1, pp 1-4. DOI: 10.12776/qip.v15i1.36.

[6] Zgodavova, K.: Complexity of Entities and its Metrological Implications. In: Proceedings of the 21st International DAAAM Symposium, ISBN 978-3-901509-73-5, ISSN 1726- 9679, pp. 0183, Editor B. Katalinic, Publisher by DAAAM International, Vienna, Austria 2010.