SELECTED PROBLEMS IN STATISTICAL MODELLING OF METALLURGICAL PROCESSES

Filip TOŠENOVSKÝ, Josef TOŠENOVSKÝ, Marta BLAŠTÍKOVÁ

VSB - Technical University of Ostrava, Faculty of Materials Science and Technology, Department of Quality Management, Ostrava, Czech Republic, josef.tosenovsky@vsb.cz

https://doi.org/10.37904/metal.2019.776

Abstract

The paper introduces modifications to the classical regression analysis that are deemed necessary when working with typical data from metallurgy, and foundry industry in particular. It specifically concerns the problem of multicollinearity, and shows both the practical application of the modified procedures and its consequences. The usefulness of the alternative procedures was verified for the metallurgical data that reflect existence of relation between selected mechanical properties of casts and their chemical composition.

1. INTRODUCTION

When modelling technological processes using the regression-based approach, the real-world data - and specifically those from metallurgical processes, as it turns out rather typically, - give rise to many pitfalls that may affect significantly the quality of a regression model. The input data in itself can cause problems, if the corresponding data matrix does not comply with the important prerequisite of having linearly independent columns (the so-called problem of multicollinearity). One methodology of dealing with this problem, simple at a glance, is to skip the dependent columns. This, however, is not understandably popular with practicians, as such a procedure may in the end result in problems with the interpretation of the model. If all the covariates or factors are to be kept in the model, another possibility is to use the so-called ridge regression. Yet another way of proceeding is to centralize the data, as described in chapter three, which also mitigates the consequences of wrong data. Since the ridge regression requires a constant parameter to be selected, in order for the method to be applied in practice, the paper also describes the technique of doing so, and relates the selection of the constant to the overall quality of the regression estimator. The ridge estimator reduces potentially the imprecision of the regression estimates by lowering their variance, but also introduces a bias to the estimates. Both these statistical properties combined define the quality of the estimator in the form of the mean squared error.

2. THE PROBLEM OF MULTICOLLINEARITY

Let us have the following problem [1]. A chemical composition of casts is controlled. If it is found out that the composition is inappropriate, the cast is isolated, and the outcome of mechanical tests is awaited. The mechanical tests control firmness characteristics, such as y_1 = yield strength (Re), y_2 = failure strength (Rm), y_3 = yield point (A). The tests are applied to scratch patterns, in line with the norm EN DIN 1563. The elements of the chemical composition observed are variables x_1 - x_9. **Table 1** shows some of the measurements that were taken in this respect. The entire data matrix contained thirty measurements.

The measurement results were used in the regression analysis when searching for a model that would describe dependencies of the y_i's characteristics on the elements x_i. In problems of this kind, a high-degree linear dependence among the x_i's in the data matrix frequently occurred, i.e. a multicollinearity turned up. Such dependency affects considerably the quality of the regression model found. To be more exact, the order of magnitude of the variance of the estimated regression coefficients increases quite a bit (even a hundred times). It is therefore convenient to check whether the input data matrix, containing the values of the xi's, is burdened with a linear dependence, and if so, then try to solve this problem. We shall show two ways of reducing the problem of multicollinearity and the consequences of not doing so.

Table 1 Selected measurements as a result of the control of casts [1]

x_1 - C	x_2 - Cr	x_3 - Cu	x_4 - Mg	x_5 - Mn	x_6 - P
3.62	0.019	0.021	0.056	0.23	0.025
3.65	0.051	0.030	0.035	0.218	0.022
3.65	0.017	0.026	0.043	0.252	0.023
3.66	0.022	0.021	0.045	0.227	0.031
3.62	0.025	0.037	0.056	0.212	0.023
3.6	0.037	0.033	0.063	0.242	0.022
3.69	0.030	0.005	0.063	0.281	0.024
3.71	0.028	0.026	0.042	0.211	0.022
3.66	0.029	0.038	0.058	0.258	0.023
etc.					

Table 1 (**continued)** selected measurements [1]

x_7 - S	x_8 - C	x_9 - Si	y_1 - Re	y_2 - Rm	y_3 - A
0.005	1.015	2.191	266	401	22
0.006	1.026	2.223	267	406	23
0.007	0.016	2.098	259	396	22.5
0.005	0.017	2.071	268	404	25
0.003	1.02	2.25	274	413	23.5
0.005	0.989	1.945	259	405	23
0.003	1.036	2.199	269	412	24
0.004	1.031	2.082	273	416	22.5
0.005	1.02	2.12	263	403	23.5
etc.					

One of the ways how to detect multicollinearity is to calculate the correlation matrix for the pairs formed from the variables x_1 - x_9, and find the elements of the matrix of the greatest magnitude (in absolute value). In our case, the only pair that showed a significant linear dependence was the pair x_6 and x_8, in the case of which the correlation coefficient $R(x_6, x_8)$ = -0.55 with the p-value of 0.0017. By excluding one of these variables from the model, the multicollinearity is reduced. It does matter, however, which of the two variables is actually taken out of the model. One way of proceeding is as follows: We find the regression model $Y = b_0 + b_1x_1 + \ldots + b_9x_9$ and calculate the coefficient of determination R^2 =0.56. Subsequently, we find the model without

a) the factor x_6 or
b) the factor x_8

and calculate $R^2(Y, \mathbf{x})$, where **x** is the vector of the included variables. The factor whose elimination lowered the criterion R^2 more is considered more significant and is kept in the model. In our case:

a) R^2(without x_6) = 0.34,
b) R^2(without x_8) = 0.38.

Leaving out the factor x_6 resulted in a more profound drop in R^2, it is therefore kept in the model, and the factor x_8 is removed from the model. If we want to keep all the variables in the model, we may opt for the technique of ridge regression.

The model quality is judged, among other things, also based on the variability of the estimated coefficients b_i. Let us use the sum of the coefficient variances as a criterion of the model quality, the criterion being denoted here by the symbol $s^2(b)$. For the full model with all the factors, $s^2(b)$ = 961,284.26. For the model without the factor x_8, $s^2(b)$ = 773,683.58, i.e. it is lower, but not by orders. This way of tackling the problem of multicollinearity lies basically in altering the input data.

The same problem may, however, be also addressed by selecting a different general formula for the parameter estimation. Instead of the basic formula

$$b = \left(X^T X\right)^{-1} X^T Y \tag{1}$$

the ridge-method formula [2] can be used,

$$b(k) = \left(X^T X + k.I\right)^{-1} X^T Y, \quad k \geq 0, \tag{2}$$

where I is the unit matrix and k is a constant to be selected. For k = 0, we get (1). A higher k reduces the sum $s^2(b)$, also often denoted as g_1 (see **Table 2**). It can be calculated as [2]

$$g_1(k) = s^2\left[Trace\left(X^T X + k.I\right)^{-1} - k.Trace\left(X^T X + k.I\right)^{-2}\right]; s^2 = \frac{\sum_i e_i^2}{n-p} \tag{3}$$

Table 2 Progress of the criteria g_1 and g_2 (see below), depending on k

k	g_1	g_2	g_1+g_2
0	961,300.4	0.0	961,300.4
0.00001	690,909.5	91,223.2	**782,132.7**
0.00002	530,489.8	263,525.9	794,015.7
0.00005	30,1524.8	794,205.2	1,095,730.0
0.0005	46,870.3	2,736,531.7	2,783,402.0
0.001	26,940.1	3,040,870.0	3,067,810.1
0.005	4,839.1	3,336,366.2	3,341,205.3
0.01	2,004.2	3,379,806.4	3,381,810.6
0.05	258.5	3,417,101.4	3,417,359.9
0.1	79.5	3,421,875.0	3,421,954.5
0.5	18.3	3,426,528.5	3,426,546.8
0.7	12.7	3,426,905.2	3,426,917.9
1	8.6	3,427,203.6	3,427,212.3

The table shows a striking drop in the variability $s^2(b) = g_1$. At the same time, however, an estimator bias, an undesired property, is also introduced. The amount of bias is denoted g_2 and calculated [2] as

$$g_2(k) = k^2 b^T \left(X^T X + k.I\right)^{-2} b \tag{4}$$

When the ridge approach is used, one looks for a reasonable balance between the bias and variability of the estimator. This is why the criteria g_1 and g_2 are summed. Typically, the sum drops first and after reaching a minimum (here for k = 0.00001), it rises again. **Table 3** is another example that shows better the typical progress of the summation g_1+g_2 for other data. Here, the minimum of the sum is achieved at k = 100.

Table 3 A typical development of the criterion $g_1 + g_2$

k	g_1	g_2	$g_2 + g_2$
0	9,301,666.044	0	9,301,666.044
0.2	8,904,236.44	862.73	8,905,099.175
0.4	8,531,744.49	3,306.57	8,535,051.069
0.6	8,182,146.65	7,134.94	8,189,281.597
0.8	7,853,604.48	12,175.02	7,865,779.514
1	7,544,460.47	18,274.64	7,562,735.121
10	2,101,890.47	509,129.64	2,611,020.121
100	64,218.87	1,555,142.31	**1,619,361.18**
150	30,199.67	1,644,990.82	1,675,190.49
200	17,488.8	1,692,863.54	1,710,352.34

The regression model for the data in **Table 1** results in quite different estimated coefficients, if they are calculated according to (2), as shown in **Table 4**.

Table 4 Coefficients calculated by LS and ridge

i	b_i	$b(0.5)$	$b(0.00001)$
0	262.89786	13.8580715	239.699419
1	-24.0293289	47.4426684	-19.1624544
2	-47.7566488	0.45006282	-34.0053101
3	55.8123267	-0.10095729	36.1541954
4	-162.993893	0.68916138	-125.623855
5	5.9176775	2.79804419	3.77908707
6	697.117097	0.50974087	659.081333
7	-1,685.90698	-0.12801234	-1,390.55369
8	2.50380999	3.06643996	2.36655636
9	40.8065705	34.9346125	42.5952223

Table 5 compares the empirical and fitted values of the dependent variable, based on the estimated coefficients **b(k)**.

Table 5 Fitted and empirical values of the modelled variable

i	Fitted Y_i		Y_i
	k = 0.5	k = 0.00001	
1	265.9554	269.5311767	266
2	268.4941	268.8075543	267
3	261.1163	260.4965191	259
4	260.5891	266.5144258	268
5	267.9818	273.825484	274
6	256.3766	256.3849634	259
7	269.7695	270.6900701	269
8	266.4057	264.1765056	273
9	265.4696	264.5632228	263
etc.			

The vector of all the fitted values is calculated by multiplying the matrix of regressors **X** and the vector of estimated coefficients ***b(k)***: **Fitted Y = X·b(k)**.

3. MULTICOLLINEARITY CURED BY VARIABLE TRANSFORMATION

Multicollinearity can also be treated via a suitable variable transformation, if the regression model is selected in such a way that allows the transformation to be successfully exploited. Let us consider the model of the form $y = b_o + b_1x + b_2x^2 +$ noise. **Table 6** contains a real data sample to be used for this illustration.

Table 6 A data sample with multicollinearity [3]

Y	x	x^2
252	78	6,084
259	73	5,329
256	77	5,929
256	68	4,624
267	85	7,225
270	86	7,396
261	83	6,889
231	76	5,776
etc.		

Calculating the correlation between x and x^2, we get the following correlation matrix (see **Table 7**).

Table 7 Correlation matrix for the pair x and x^2

	x	x^2
x	1	0.998949
x^2	0.998949	1

The extent of the dependence can be reduced in this situation by centralizing the x variable. The mean of x is 80.4, therefore the new, centralized variable is $x_1 = x - 80.4$. The correlation matrix now is shown in **Table 8**.

Table 8 Correlation matrix for x_1 and x_1^2

	x_1	x_1^2
x_1	1	-0.50537
x_1^2	-0.50537	1

In absolute terms, the correlation dropped by half. Comparing the estimated coefficients by the sum $s^2(b)$ for the (non)centralized regressor x, we get the following results (see **Table 9**).

Table 9 Regression with the noncentralized x

	b_i	$s(b_i)$	t Stat	p-value
b_o	530.6425	184.135	2.881813	0.006546
b_1	-8.60404	4.720552	-1.82268	0.076442
b_2	0.064905	0.030114	2.155332	0.037711

The regression model is $Y = 530.6425 - 8.60404x + 0.064905x^2$, and

sum s^2 (b) = $184.135^2 + 4.720552^2 + 0.030114^2 = 33927.97$.

The correlation is R = 0.77 with p-value = 0.000 in the related ANOVA. The Durbin-Watson statistic equals 2.36 (p-value = 0.10).

For the centralized x, the results are in **Table 10**.

Table 10 Regression with the centralized x

	b_i	$s(b_i)$	t Stat	p-value
b_o	258.4356	1.781643	145.0546	1.37E-52
b_1	1.832723	0.250721	7.309819	1.1E-08
b_2	0.064905	0.030114	2.155332	0.037711

The sum $s^2(b)$ =3.24, far lower. The regression model is $Y = 258.44 + 1.83x_1 + 0.06x_1^2$, with the correlation coefficient R = 0.77 and p-value in the ANOVA. The Durbin-Watson statistic = 2.36 with the p-value 0.10.

The regression estimates of the two approaches are different. Regressing one variable on another, centralized variable brings better results also in the case of having input data with errors. For instance, by changing Y_1 = 252 erroneously to Y_1 = 2.52, we get the following results (see **Table 11**).

Table 11 Effect of erroneous data on the sum $s^2(b)$

	Data and method	b_i	Sum $s^2(b)$
1	Wrong Y_1, noncentralized data, LS	1,311.263	795153.3
		-29.2589	
		0.199275	
2	Wrong Y_1, centralized data, LS	246.9931	75.88
		2.784485	
		0.199275	
3	Correct Y_1, centralized data, LS	258.4356	3.23
		1.832723	
		0.064905	

It is worth noting the value of the coefficient b_o = 1,311.263 in approach (1). It is very different from the same coefficient obtained from the correct data. By contrast, approach (2) is not affected much by the data errors. The summation criterion $s^2(b)$ is worst by far in (1). A more systematic approach, however, to estimate regression coefficients from erroneous data is the robust-regression approach (RLS).

4. CONCLUSION

Except for the problem of multicollinearity, in metallurgical practice, imprecisely measured or wrongly stored data occur frequently, which damages regression models severely. To avoid these problem, it is either possible to try to detect and remove the errors in the data, or use robust methods, which are less sensitive to erroneous data inputs. Using specific data from the foundry industry (see **Table 1**), which contained wrong information of a hundred - times imprecision order about the amount of carbon presence (x_1), several approaches to modelling the yield strength (y_1) were verified. The model was searched with the Huber and Busquar method, and was compared to the classical regression approach.

Another problem, repeatedly encountered in metallurgical data and presented to the conference participants, though not further pursued in this paper, concerns heteroscedasticity (nonconstant variance of model residuals). Its consequence is an increased instability, or variance, of the estimated regression models employed in metallurgy. In this case, the classical least-squared (LS) model estimation was compared to the generalized least-squared (GLS) procedure. The metallurgical regression models working with GLS showed a lower - in - order variability of the estimates: for the case of the data from the foundry industry, the sum of the variances of the estimated coefficients was 45 times lower than that for the case of the GLS method.

We believe that in the foundry industry, and more generally in the metallurgical practice, it is an absolute imperative to pay due and extra attention to the data worked with when seeking a regression model, and opt for the corresponding (robust) regression procedures, otherwise solid model - based conclusions in this industrial branch are far from guaranteed.

ACKNOWLEDGEMENTS

This paper was prepared under the specific research projects No. SP2019/62 and SP2019/129, conducted at the Faculty of Materials Science and Technology, VSB - Technical University of Ostrava with the support of The Ministry of Education, Youth and Sports of the Czech Republic.

REFERENCES

[1] MAĎORANOVÁ, M. Hodnocení technologického procesu pomocí vícerozměrné ztrátové funkce. In *Mezinárodní vědecká konference Technické vedy a výrobný manažment*. Tatranská Štrba, 2015.

[2] HOERL, A.E., KENNARD, R.W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 2000, vol. 42, no. 1. pp. 80-85.

[3] KUTNER, M.H., NACHSTEIM, C.J., NETER, J. *Applied Linear Regression Models*. New York: Mc Graw-Hill, 2004.

[4] CIPRA, T. *Ekonometrie*. Praha: SPN, 1984.