MODELLING OF TECHNOLOGICAL PROCESSES WITH QUALITATIVE VARIABLES

Josef TOŠENOVSKÝ [a], Filip TOŠENOVSKÝ [b], Tereza SMAJDOROVÁ [a]

[a] VSB - Technical University of Ostrava, Ostrava, Czech Republic, EU, josef.tosenovsky@vsb.cz

[b] Silesian University in Opava, Karvina, Czech Republic, EU, tosenovskyfilip@opf.slu.cz

Abstract

The paper deals with modelling processes whose output is affected by both continuous quantitative variables and qualitative variables. Three examples working with simulated data show how to perform regression involving one or two qualitative variables. The paper also shows a simple approach to assessing model correspondence, using a test of the absolute terms of regression functions.

Keywords: design of experiments, regression function, nominal and continuous variables

1. INTRODUCTION

Design, management and evaluation of technological processes face many problems of different nature. Therefore various methods and approaches must be applied, including those that observe processes from the financial point of view, such as those presented in [4], [5], for instance, robust methods of process management [6], unconventional management techniques [3], [1] or methods working with metrological aspects of management [2]. An important aspect of this complex approach is also process modelling which enables to study and experiment with processes at minimum costs. Building a model often presents a problem of incorporating entry parameters that cannot be measured (qualitative factors), but affect the process under scrutiny significantly. To give an example, such factors may include the technology used in the process, the type of material worked with, the staff attending the process and so on. These factors cannot be inserted in the regression model that describes the process, however, their influence on the process output is often crucial. It is therefore convenient to include these factors among the influential factors, find various mathematical models and construct their graphs, and make a judgment as to whether the models found differ significantly or not. For instance, if models reflecting different technological procedures are mutually consistent, it does not matter that much which of the procedures will be used. In this paper, we illustrate, using simulated data, how models with qualitative factors can be constructed and compared.

2. A MODEL WITH CONTINUOUS VARIABLE AND TWO-LEVEL QUALITATIVE VARIABLE

Let Y be a quality characteristic observed, which depends on two variables: a continuous variable x_1 and a nominal variable x_2, the latter taking on two values, and representing a type of technology used. We shall work with an experimental plan in which the two values are 1 and 0. The process under scrutiny was run 16 times: in 8 cases, $x_2 = 1$, and in the other 8 cases, $x_2 = 0$. The result of the experiment is shown in **Table 1**.

Table 1 Experimental plans for two different types of technology x_2

A)					B)			
	x_1	x_2	Y			x_1	x_2	Y
1	7	1	22.70704		1	7	1	22.70704
2	55	1	8.976949		2	55	1	8.976949
3	38	1	11.65757		3	38	1	11.65757
4	74	1	1.886896		4	74	1	1.886896
5	52	1	10.14272		5	52	1	10.14272
6	80	1	3.20113		6	80	1	3.20113
7	26	1	16.63271		7	26	1	16.63271
8	80	1	1.55513		8	80	1	1.55513
9	17	0	14.14009		9	7	0	17.14009
10	37	0	10.54875		10	55	0	5.14875
11	60	0	1.332736		11	38	0	7.932736
12	4	0	17.83611		12	74	0	-3.16389
13	58	0	3.396037		13	52	0	5.196037
14	35	0	8.3186		14	80	0	-5.1814
15	47	0	5.75086		15	26	0	12.05086
16	17	0	15.51993		16	80	0	-3.38007

The corresponding regression coefficients and their characteristics were found (see **Table 2**).

Table 2 Regression analysis performed for the data in **Table 1**

	bi	**s(*bi*)**	**t stat**	**p-val.**
b_0	**19.28646**	0.568818	33.90624	4.5081E-14
b_1	**-0.28163**	0.012226	-23.0356	6.34294E-12
b_2	**4.812565**	0.581115	8.281602	**1.52805E-06**

The regression model is of the form $Y = 19.28 - 0.28\, x_1 + 4.81\, x_2$. Inserting a specific value in x_2, we get the equation for the corresponding technology x_2:

$x_2 = 1$ $\quad Y = 19.28 - 0.28\, x_1 + 4.81 = 24.09 - 0.28\, x_1$

$x_2 = 0$ $\quad Y = 19.28 - 0.28\, x_1$

We deal with two parallel lines whose slope is -0.28. Depending on whether $x_2 = 1$ or $x_2 = 0$, we use the appropriate equation to calculate the process output Y. The graph of the two lines is in **Fig. 1**.

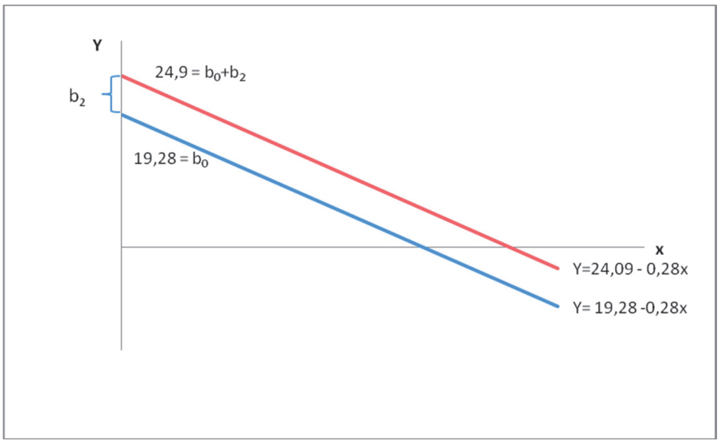

Fig. 1 Graphs of the models of the first type

Testing statistical significance of the coefficient b_2, we may determine if there is any significant difference in switching from one technology to another, i.e. when $x_2 = 1$ or $x_2 = 0$. If the coefficient is zero, no difference between the two technologies can be assumed. The p-value accompanying the test of significance of b_2 suggests that the hypothesis H_0: $\beta_2 = 0$ is rejected, so that the coefficient is not zero, and the two technologies make a difference.

In **Table 1** A), there are different values of x_1, depending on whether $x_2 = 1$ or $x_2 = 0$. If the plan was run in such a way that the values of x_1 were the same both when $x_2 = 1$ and $x_2 = 0$, we would get a very similar regression model $Y = 19.27 - 0.28\ x_1 + 5.12\ x_2$ (see **Table 1 B)**).

3. A MODEL WITH CONTINUOUS VARIABLE AND THREE-LEVEL QUALITATIVE VARIABLE

Let the quality characteristic Y depend now on the following two variables: a continuous variable x_1 and a three-level nominal variable x_2. To work with three levels of x_2, we introduce two auxiliary two-level variables z_1 and z_2. The necessary relations between the auxiliary variables on the nominal variable are shown in **Table 3**. Generally speaking, one can use *k-1* auxiliary variables to express a *k*-level nominal variable.

Table 3 Auxiliary variables

z_1	z_2	x_2
1	0	A
0	1	B
0	0	C

Table 4 presents a new experiment utilizing the idea of auxiliary variables.

Table 4 leads to a regression function $Y = 1.66 + 10\ x_1 - 2.56\ z_1 + 7.02\ z_2$.

If the process under scrutiny is run for the option x_2 = A or B or C, the corresponding regression model a) or b) or c) for such a process is obtained by inserting the appropriate values in the variables z_1 and z_2 of the regression function related to **Table 4**:

a) $Y = 1.66 + 10\ x_1 - 2.56 = -0.9 + 10\ x_1$

b) $Y = 1.66 + 10\ x_1 + 7.02 = 8.68 + 10\ x_1$

c) $Y = 1.66 + 10\ x_1$

Table 4 An experimental plan for the three-level nominal variable

	x_1	z_1	z_2	Y
1	707	1	0	7071.807
2	655	0	1	6561.477
3	638	0	0	6383.058
4	574	1	0	5741.087
5	552	0	1	5531.743
6	980	0	0	9807.201
7	926	1	0	9261.433
8	680	0	1	6811.555
9	597	0	0	5974.24
10	637	1	0	6373.649
11	660	0	1	6610.333
12	704	0	0	7044.036

4. A CASE OF CONTINUOUS VARIABLE AND TWO TWO-LEVEL QUALITATIVE VARIABLES

Let the observed quality characteristic Y depend on one continuous variable x and two qualitative variables z_1 and z_2, both of which can be at two levels: a lower level represented by 0 and an upper level represented by 1. A part of the corresponding experimental plan is in **Table 5**.

Table 5 An experimental plan for two qualitative variables

x	z_1	z_2	Y
707	0	0	1414.807
655	1	1	1312.477
638	1	0	1272.058
574	0	1	1152.087
552	0	0	1105.743
980	1	1	1964.201
926	1	0	1849.433
680	0	1	1365.555
etc.			

The regression function related to the data of **Table 5** is

$$Y = 0.55 + 2\,x - 2.74\,z_1 + 4.07\,z_2 \qquad (1)$$

and generally it is of the form $Y = b_0 + b_1\,x + c_1\,z_1 + c_2\,z_2$.

For various settings of the variables z_1, z_2, we get the regression equations given in **Table 6**.

Table 6 Different models for different levels of z_1 and z_2

Model	z_1	z_2		
M(0,0)	0	0	$Y = 0.55 + 2\,x$	Generally Y = $b_0 + b_1\,x$
M(1,1)	1	1	$Y = 1.88 + 2\,x$	Generally Y = $b_0 + b_1\,x + c_1 + c_2$
M(1,0)	1	0	$Y = -2.19 + 2\,x$	Generally Y = $b_0 + b_1\,x + c_1$
M(0,1)	0	1	$Y = 4.62 + 2\,x$	Generally Y = $b_0 + b_1\,x + c_2$

Graphs of the four functions are depicted in **Fig. 2**.

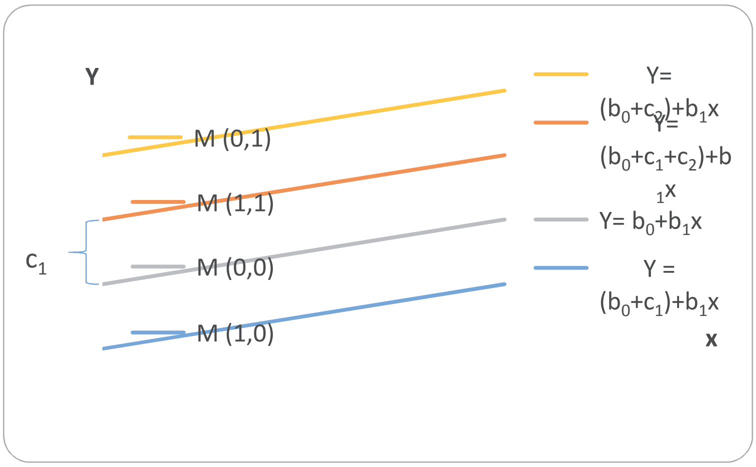

Fig. 2 Graphs of the models from **Table 6**

If we want to determine whether the various process settings differ significantly, all we need to do is test significance of the differences of the regression absolute terms. For instance, if we want to compare the first and second setting of z_1 and z_2 from **Table 6**, we test statistically the difference $(b_o + c_1 + c_2) - b_o = c_1 + c_2$.

To test the theoretical value of $c_1 + c_2$, we need to estimate the standard deviation $s\,(c_1 + c_2)$ of $c_1 + c_2$. For the variance or dispersion D of $c_1 + c_2$, we have

$$D(c_1 + c_2) = D(c_1) + D(c_2) + 2\mathrm{cov}(c_1, c_2). \qquad (2)$$

All these characteristics are contained in the variance matrix of the vector of regression coefficients, denoted var. The matrix for our case is in **Table 7**. Generally, for a vector of regression coefficients $\vec{b}$,

$$\mathrm{var}(\vec{b}) = \frac{\sum e_i^2}{n-k}(X^T X)^{-1}, \qquad (3)$$

where X is the matrix of regressors, e's are residuals from the regression, n is the number of observations of Y and k is the number of regressors on the right-hand side of the regression equation.

Table 7 Variance matrix

	b_0	b_1	c_1	c_2
b_0	1.743727	-0.002621292	0.236958	-0.005839
b_1	-0.002621	4.25383E-06	-0.000523	-0.000129
c_1	0.236958	-0.000523487	**0.235672**	**0.015934**
c_2	-0.005839	-0.000129476	0.015934	**0.175191**

$D(c_1) = 0.235672$; $D(c_2) = 0.175191$; $\mathrm{cov}(c_1,c_2) = 0.015934$. The coefficients c_1 and c_2 are found in model (1). Equation (2) then gives the estimate of the variance of $c_1 + c_2$

$$D(c_1 + c_2) = 0.235672 + 0.175191 + 2\ 0.015934 = 0.44273. \quad (4)$$

From here, we obtain $s(c_1 + c_2)$ as a square root of $D(c_1 + c_2)$.

The test criterion for testing significance of $c_1 + c_2$ is

$$T = \frac{c_1 + c_2}{s(c_1 + c_2)} = \frac{-2,74 + 4,07}{0,66538} = 1,998859. \quad (5)$$

The critical value of the test is $t_{28}(0.05) = 2.048$, and it is not exceeded by T. Therefore, the models M(0,0) and M(1,1) do not differ significantly. This implies that using the regime $z_1 = 0$, $z_2 = 0$ or the other regime $z_1 = 1$, $z_2 = 1$ makes no difference.

CONCLUSION

This paper studied two frequent problems connected with regression modelling of technological processes. Simulated data showed three different situations in which the problem of working with qualitative variables in a model was resolved. The qualitative variables represented the type of technology used in a process. Also, a way of comparing different models by testing significance of their absolute regression coefficients was illustrated. This paper loosely follows up on the cited literature source [6], [2] and [1].

AKNOWLEDGEMENT

This paper was elaborated within the frame of the specific research project No. SP2014/81 which has been solved at the faculty of Metallurgy and Material Engineering, VSB***-TU Ostrava with the support of the Ministry of Education, Youth and Sports in the Czech republic.***

REFERENCES

[1] TOŠENOVSKÝ, J., TOŠENOVSKÝ, F., MONSPORTOVÁ, L. Application of Multivariate Loss Function for the Assesment of Quality Kontrol of Technological process. In Metal 2012: 21st International Conference on Metallurgy and Materials. Ostrava: Tanger, 2012.

[2] TOŠENOVSKÝ, J., TOŠENOVSKÝ, F. Possibilities of Production Process Financial Assessment. In Conference Financial Management of Firms and Financial Institutions, Ostrava: VSB - Technical University Ostrava, pp. 982-987.

[3] TOŠENOVSKÝ, F. A Mathematical Model for Process Cycle Time - Theory and Case Study. Quality, Innovation, Prosperity. Trenčín, 2010, pp. 64-71.

[4] TOŠENOVSKÝ, J., TOŠENOVSKÝ, F., KUDĚLKA, O. Analysis of robust technology design methods. In Metal 2013: 22nd International Conference on Metallurgy and Materials, Ostrava: TANGER, 2013, pp. 1662-1668.

[5] ZGODAVOVÁ, K., BOBER, P. An innovative approach to the integrated management system development SIMPRO-IMS web base environment. Quality Innovation Prosperity, 2012, Vol. XVI/2 - 2012, pp. 59.

[6] ZGODAVOVÁ, K. Complexity of Entities and its Metrological Implications. Ing Conference Proceedings of the 21st International DAAAM Symposium, Zadar: DAAAM International, 2010, pp. 365-367.