A MATHEMATICAL MODEL FOR PROCESS CYCLE TIME - THEORY AND CASE STUDY

FILIP TOŠENOVSKÝ

1 INTRODUCTION

It is a common practice to carry out an economic analysis, using mathematical model of the form $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$. The relation is an expression of dependence of the i-th level of random variable y on the i-th level of variables $x_1, \ldots, x_k$ that represent the most influential factors, and also its dependence on the random variable ε which represents the remaining factors. It is assumed that variables ε_i have zero expected value, they are uncorrelated and have a constant variance, i.e. $E(\varepsilon_i) = 0$ for any i, $var(\varepsilon_i) = \sigma^2$ for any i and $cor(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$. The least squares method and a data sample $(y_1, \ldots, y_n)$, $(x_{i1}, \ldots, x_{ik})$, $i = 1,2, \ldots, n$, are used to estimate the unknown parameters β_i. If the conditions imposed on ε_i hold, the least squares method yields the best linear unbiased estimates of the coefficients β_i. From a practical point of view, however, it is more important that we get estimates that do not differ much from the unknown coefficients β_i, with a probability that increases as the size of the data sample n increases (Greene, 1990). The importance of this statistical property lies in the fact that the estimates have this characteristic when any single data sample of size n is used, whereas the property of being an unbiased estimate relates to an average estimate computed from at least a large number of generally different data samples of size n, and theoretically from an infinite number of such data samples. Given that meeting the condition $E(\varepsilon_i) = 0$ is assumed a priori, it is obvious that no correlation, or more generally statistical independence, and the constant variance of the variables y_i is what we are interested in when building a model. If $var(y_i)$ depends on i, we may try to find a transformation T such that variance of the transformed variable $var(T(y_i))$ will be stable. Using the Box-Cox transformation is one of the ways how to find the function T. The Box-Cox transformation is useful in that it may also bring the probability distribution of y_i closer to normal distribution, apart from stabilizing the variance (Box, Cox 1964). The additional prerequisite of normality then ensures estimates of β_i are the best unbiased estimates of all conceivable estimates, not only of those that are linear in its nature.

In this article, we derive a suitable transformation T, and thus a regression model as well, to describe a dependence of working process cycle time y on relevant factors $x_1, \ldots, x_k$ that enter the process. We specifically deal with processes that are stable in a certain sense. At the end of the article, we also present a real industrial case study in which the derived model was used. Since we model process cycle time, we may assume that the variable y we work with is positive.

2 BOX-COX TRANSFORMATION

We use the Box-Cox transformation to find an appropriate function *T* that stabilizes the variance of a positive random variable y, and which may also bring the distribution of y closer to normality. The Box-Cox transformation is defined as

$$y_i^{(\lambda)} = \frac{y_i^{\lambda} - 1}{\lambda} \text{for } \lambda \neq 0,$$

where y_i is the i-th value of the original variable and $y_i^{(\lambda)}$ is the i-th value of the transformed variable. The transformation depends on parameter lambda. If the parameter is close to zero, the transformation $y_i^{(\lambda)} = ln(y_i)$ is used as $\lim_{\lambda\to 0}\{(y_i^{\lambda} - 1)/\lambda\} = ln(y_i)$. We set the following objective: finding lambda which stabilizes the variance $var(y_i^{(\lambda)})$, i.e. finding lambda such that the variance of the transformed variable is constant regardless of i. We shall find the lambda for a working process that is characterized by the fact that the expression $var(y_i)/E(y_i)^2$, or the squared coefficient of variation of y_i, is constant regardless of i.

3 LAMBDA

It is suitable to express $var(y_i^{(\lambda)})$ as a function of lambda if we are to find the lambda stabilizing $var(y_i^{(\lambda)})$. However, such an expression is unknown exactly due to the fact that $y_i^{(\lambda)}$ is a nonlinear function of lambda. Therefore, we shall first approximate linearly y_i^{λ}, using the Taylor polynomial function of the first order (Jarník, 1984). Doing so from a positive point a, we get for *i* = 1,2,…,*n* an approximation

$$y_i^{\lambda} \cong a^{\lambda} + \lambda a^{\lambda-1}(y_i - a), \tag{1}$$

so that

$$var(y_i^{\lambda}) \cong var\{a^{\lambda} + \lambda a^{\lambda-1}(y_i - a)\} = \left(\lambda a^{\lambda-1}\right)^2 var(y_i). \tag{2}$$

Altogether, we have for $\lambda \neq 0$

$$var(y_i^{(\lambda)}) = var\left(\frac{y_i^{\lambda} - 1}{\lambda}\right) \cong \frac{1}{\lambda^2}\lambda^2 a^{2\lambda-2} var(y_i) = a^{2\lambda-2} var(y_i). \tag{3}$$

Approximation (1) can be used whenever the right-hand side and the left-hand side of the expression (1) make sense. This is certainly the case when the values y_i, a are positive as is our case in which the variables y_i represent a working process cycle time. There is no other problem in expression (1). The question is how accurate the approximation is. If the values y_i are close to the point a, the approximation is acceptable, otherwise it doesn't have to be acceptable. Since the variables y_i may come from probability distributions that have different parameters, it is not wise to approximate them from a single point a. It is more natural to use expected values $E(y_i)$ which characterize the location of y_i (Rényi, 1970). If we use the expected values, we get for i =1,2,…,n an approximation

$$var(y_i^{(\lambda)}) \cong E(y_i)^{2\lambda-2} var(y_i) = E(y_i)^{2\lambda} \frac{var(y_i)}{E(y_i)^2}. \tag{4}$$

If the ratio $var(y_i)/E(y_i)^2$ doesn't change too much with respect to i, then (4) implies that the variance of the transformed variables will be stable if lambda is close to zero. If the ratio is more or less equal to a constant C, we have for a very small lambda

$$var(y_i^{(\lambda)}) \cong E(y_i)^{2\lambda} \frac{var(y_i)}{E(y_i)^2} \cong E(y_i)^0.C = C. \tag{5}$$

Thus, a very small lambda used in the Box-Cox transformation stabilizes variance provided the second power of the coefficient of variation is stable. This, however, means logarithmic transformation is the transformation we were looking for. This leads to a model

$$y_i^{(\lambda)} = ln(y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_{i,} \tag{6}$$

or

$$E\{ln(y_i)\} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}, \tag{7}$$

where on the right-hand side we may as well work with a more general expression, but preferably with the one linear in parameters, i.e. with an expression of the form $\beta_0 + \beta_1 f_1(x_{i1}, \dots, x_{ik}) + \cdots + \beta_k f_k(x_{i1}, \dots, x_{ik})$. It is known that the statistical properties of coefficient estimates resulting from the least squares method hold for the more complex functional form as well.

4 MODEL

It would be better if we worked with the model $ln(E(y_i)) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$, which leads to equation $E(y_i) = \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})$, instead of (7). Expected value of logarithm is not equal to logarithm of expected value, but under certain conditions we may still perform this substitution (see below). The conclusion then is: if $var(y_i)/E(y_i)^2$ shows stability, a suitable model is of the form

$$E(y_i) = \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}) \tag{8}$$

It is obvious from the described procedure that we may restrict ourselves to a positive and small lambda as it has the same effect on variance as a negative and small lambda. Let us examine now the approximations we used in the process of deriving the model:

a) approximation (1);
b) approximation $E(y_i)^{2\lambda} \cong 1$ for a very small (and positive) lambda;
c) approximation $ln\{E(y_i)\} \cong E\{ln(y_i)\}$.

Approximation b) is apparently of no harm given the continuity of exponential function. If lambda is small and y_i is not too small, approximation a) is of no harm either because x^λ as a function of x has the derivative $dx^\lambda/dx = (1/x)\,\lambda x^\lambda$, and this derivative will be close to zero for a small lambda and not too small x. Thus, the function x^λ will be very flat (see an example of such a function in figure 1). In this case, approximation of such a function by a line will be suitable. Approximation c) will be the major source of inaccuracy. Using the Taylor polynomial function of the first order, we have for the third approximation $E\{ln(y_i)\} = ln\{E(y_i)\} - (1/2\xi^2)var(y_i)$, where $E(y_i) < \xi < y_i$, or $E(y_i) > \xi > y_i$. This implies that approximation c) will be more accurate in case variables y_i achieve higher values and/or have a small variance.

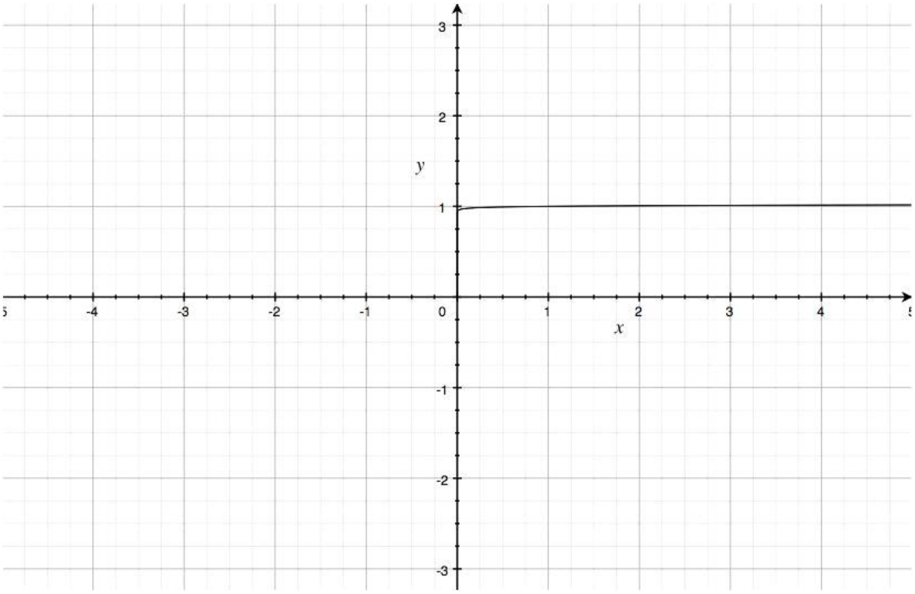

Figure 1- function $h(x) = x^{0{,}01}$

5 A CASE STUDY FROM THE CZECH REPUBLIC

In 2009 a working process in the testing laboratory of a major Moravia-based ironmaker was scrutinized. The laboratory struggled to perform its tests of quality of the ironmaker's products in the shortest time possible, which is a distinctive feature of testing laboratories in general. The reason is that the ironmaker, a part of which the laboratory is, funds its activities in part through bank credits just like any other bigger company. Thus, it has to pay off debts and interests, and so it tries to function as fast as possible so that the interests do not accumulate over time. The time factor also plays its role due to the fact that the ironmaker pays České dráhy railway company for expedition of iron products. Any delays on the rail track are therefore unwanted. The reality is that situations occur when the final product is more or less loaded up on train wagons, and the ironmaker awaits only the results of laboratory tests. In such cases, the laboratory is perceived as a hurdle by the ironmaker, although a necessary one. Therefore it was an imperative to solve the problem of process cycle time in the laboratory. The analysis of the problem focused on finding a suitable mathematical model which would express dependence of the working process cycle time of a problematic laboratory test on relevant factors. Conceivable factors were the number of technicians working in the laboratory and amount of work of different kind in the lab. However, since the number of technicians was constant during the time period of the analysis, it was not placed in the model as a variable, and so the factor influencing the process cycle time was the number of raw material samples the laboratory had to test. Specifically, there were two types of raw material samples that had the influence on the problematic test, so two independent variables occurred in the model. All the data required for finding the model were available as the laboratory recorded everything it did in its own computer system, including completed tests, results of the tests, names of technicians who carried out the tests, and the time it took them to perform the tests. The original data are shown in table 1. The data include working process cycle times for the problematic test (in hours), depending on the amount of raw material samples of two kinds x_1, x_2 that were used during the test. Different combinations of x_1, x_2 that ocurred in the test are recorded at the top of table 1. Table 2 contains sample characteristics obtained from table 1. These are: average process cycle times $\bar{x}$ of the problematic test, sample variances of these times s^2, coefficients of variation $s/\bar{x}$, and a comparison of the values $\overline{ln(y)}$ and $ln(\bar{y})$, which points to the extent of inaccuracy embedded in the approximation c).

[0,1]	[1,3]	[2,4]	[5,5]	[13,3]	[12,8]
22,9681	67,0595	71,84	81,1366	164,951	139,581
21,2965	50,5145	48,86	131,786	108,723	134,005
13,9083	69,324	48,95	95,7887	113,281	156,931
26,7113	80,4345	61,29	79,1434	151,387	186,156
28,5072	58,7454	52,26	104,573	104,73	148,873
26,4103	50,4678	51,92	72,6111	142,309	138,589

28,1765	59,906	57,02	86,2439	152,438	177,769
27,5524	53,4843	54,5	80,9638	145,522	158,37
32,9603	55,65	52,22	94,0361	104,289	159,901
20,193	57,9979	47,57	67,6957	144,329	148,341
30,3088	66,5641	40,19	82,4358	126,033	187,887
22,2892	60,757	34,52	80,433	154,254	165,222
36,7975	65,6812	69,09	82,4811	131,847	138,204
27,4256	57,5055	52,34	76,0038	80,529	166,423
32,5214	73,9333	62,08	92,8184	155,013	148,945
22,4586	98,9978	41,33	67,2483	116,453	229,357
19,0731	87,9221	43,18	63,9346	99,3141	146,857
22,3774	66,2063	41,5	83,145	91,4711	230,198
30,7776	49,6289	50,48	131,054	142,426	173,117
33,5733	67,7669	46,56	80,7934	114,519	212,137
25,4427	71,8023	-	-	129,325	226,896
30,4058	70,8647	-	-	113,367	158,778
23,0087	81,1306	-	-	-	118,447
-	51,6516	-	-	-	123,431
-	78,7595	-	-	-	161,073
-	-	-	-	-	185,652
-	-	-	-	-	202,319

Table 1 - Lab test times for a given number of testing samples

x_1	x_2	$\bar{x}$	s^2	$s/\bar{x}$	$\overline{ln(y)}$	$ln(\bar{y})$
0	1	26,31	29,43	0,21	3,248	3,233
1	3	66,11	154,92	0,19	4,175	4,191
2	4	51,39	90,28	0,18	3,923	3,894
5	5	86,72	332,00	0,21	4,444	4,416
13	3	126,66	548,94	0,18	4,824	4,799
12	8	166,20	999,98	0,19	5,105	5,113

Table 2 - Statistical characteristics calculated from the data in table 1

Characteristics in table 2 suggest the coefficient of variation is rather stable across the groups of process cycle times, and it is close to 0,2. The last two columns of table 2 also imply that the inaccuracy of approximation c) is not significant. Therefore the equation (8) was used as a model describing dependence of working process cycle time of the problematic laboratory test on the amount of work the laboratory was burdened with.

Using the least squares method to estimate the parameters of model (8), the resulting regression model is

$$\hat{y} = \exp(3{,}44 + 0{,}075x_1 + 0{,}11x_2). \qquad (9)$$

Figure 2 compares empirical data from table 1 and theoretical values resulting from (9).

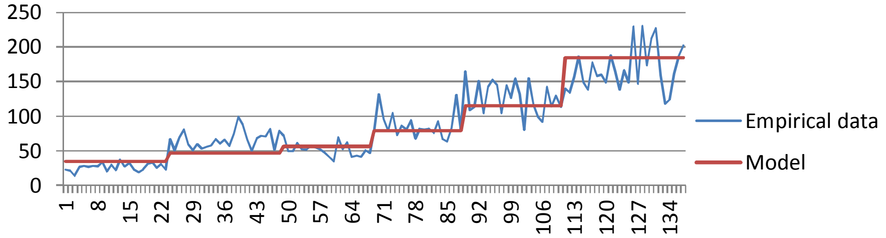

Figure 2-Model (9)

We may improve the model a bit as it doesn't fit well enough empirical values in the range of 20 to 45 and 109 to 134. If we use the polynomial function of the second order on the right-hand side of (8): $E(y_i) = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}x_{i2} + \beta_4 x_{i1}^2 + \beta_5 x_{i2}^2)$, the least squares method gives a model

$$\hat{y} = \exp(1{,}66 - 0{,}36x_1 + 2x_2 + 0{,}22x_1x_2 - 0{,}017x_1^2 - 0{,}42x_2^2). \qquad (10)$$

Figure 3 indicates that model (10) fits the data better than model (9).

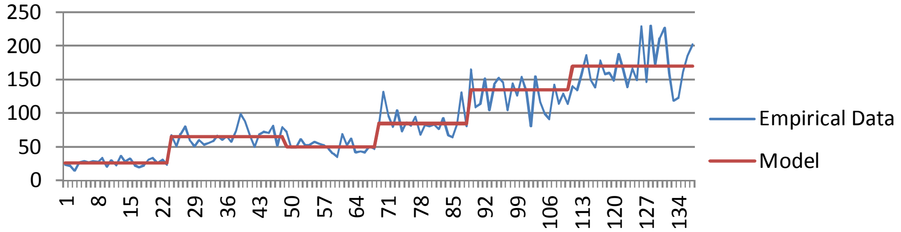

Figure 3-Model (10)

Model of the form (9) can be used, for instance, for the following purposes:

To monitor work productivity as the model shows how long it should approximately take to test two kinds of raw material samples sized x_1 and x_2.

To establish how many workers are necessary so that all the samples were tested in time. This use of the model is valid in case one of the variables x_i in the model represents the number of employees.

Since the model is exponential, we may assume that it is more convenient to process a greater number of batches containing a smaller number of raw material samples rather than the opposite. In the latter case, the time to process the samples increases disproportionately – by $exp(b_1)$ if the number of samples of one kind x_1 rises by a single unit (ceteris paribus), or by $exp(b_2)$ if the number of samples of the second kind x_2 rises by one unit.

To establish the amount of work the lab is able to accept if it is to handle the work by the time T at the latest. The requirement means that the acceptable raw material sample sizes x_1, x_2 must satisfy the inequality
$\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2) \leq T$.

6 CONCLUSION

The article dealt with a regression model that would describe dependence of working process cycle time on a group of relevant factors $x_1, \dots, x_k$. The analyzed working process was specific in that groups of cycle times corresponding to concrete levels of the factors $x_1, \dots, x_k$ had the same coefficients of variation. The Box-Cox transformation then implies that exponential model is a suitable regression model for such a situation. The reason behind it is the fact that taking a logarithm of measured times stabilizes their variance, and thus ensures that the least squares method gives estimates of the regression model with good statistical properties. Logarithm may also bring the probability distribution of times closer to normal distribution. The derived exponential model was applied to real-life data which showed that such a model is reasonable (the last two columns of table 2 and figure 3).

REFERENCES

Box, G., Cox D. (1964), "An Analysis of Transformations", Journal of the Royal

Greene, W.H. (1990), Econometric Analysis, Macmillan Publishing.

Jarník, V. (1984), Diferenciální počet I, Academia Praha.

Rényi, A. (1970), Probability Theory, North-Holland Publishing Company. Statistical Society, Series B, pp. 211-264.

ABOUT THE AUTHOR

Ing. Filip Tošenovský, Ph.D., The School of Business Administration in Karviná, Silesian University, Univerzitní náměstí 1934/3, 733 40 Karviná, Czech Republic, e-mail: tosenovskyfilip@opf.slu.cz.