XX IMEKO World Congress
Metrology for Green Growth
September 9–14, 2012, Busan, Republic of Korea

INFLUENCE OF MEASUREMENT ACCURACY ON THE RELIABILITY OF REGRESSION MODEL ESTIMATION FOR D-OPTIMAL PLANS EVALUATION

Ing. Martin Motyčka[1], *doc. Ing. Olga Tůmová, CSc.*[2]

[1]Department of technology and measurement, Faculty of Electrical engineering, University of West Bohemia, Pilsen, Czech Republic, mmotycka@ket.zcu.cz

[2]Department of technology and measurement, Faculty of Electrical engineering, University of West Bohemia, Pilsen, Czech Republic, tumova@ket.zcu.cz

Abstract: The main objective of this paper is to describe the dependence of the reliability of estimates of the regression model due to measurement accuracy. The main focus is on the experiments with reduced number of the tests, especially on the D-optimal plans, because these types of experiments are nowadays very important in the industrial praxis.

Keywords: Reliability of regression model, d-optimal plans, method of least squares

1. INTRODUCTION

Experimental work is an important part of development of every product. Each experimental work has several stages that can be summarized in the following chain: planning – modeling – preparation – experimentation – analysis – conclusion. This paper aims to the analysis of experiment. It is important to have high quality of data for analysis, because the reliability of regression models is highly sensitive to the measurement error. This sensitivity is increasing with decreasing volume of analyzed data and data quality plays an important role especially in experiment with reduced number of tests like Taguchi design and D-optimal plans.

2. FACTORIAL DESIGNS

The full factorial designs are one of the classic methods of DOE. All factors have the same weight in these types of experiments, so it is possible to change the order of test before analysis of the results. The disadvantage of the full factorial design is the considerable rising of the number of experiments with rising number of factors, its levels and repetition.

The randomized block design, the balanced incomplete block design, the Latin square design or Greco-Latin square design belongs to factorial design. These experiments are one of the methods with reduced number of tests.

The hierarchical design is one of the newer methods of factorial design. In this experiment, each value of a factor occurs in conjunction with only one level of another factor, and therefore it cannot be arbitrarily interchangeable. These methods are mainly used to study the influence of sources of variability that can occur over time.

For example: For measurement systems, where sources of variability have not yet been investigated, it is recommended the three level DOE. This model is especially recommended for calibration and verification of measurement system and for determination of measurement uncertainty.

- 1st level – is the lowest level. The measurements are done in the short time period (during one day, or one shift) and it is observed in particular to repeatability of measurements.
- 2nd level – The measurement are done during few days (or similar time period)
- 3rd level – is the highest level. The iterations are separated by the months. This level is the long time period.

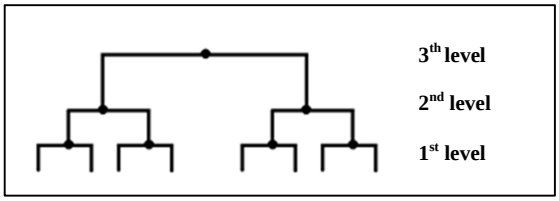

Fig. 1: The balanced hierarchical design

3. TAGUCHI APPROACH OF DOE

It's obvious, that in the common usage it isn't often acceptable to perform large number of tests in the experiment. The main idea of the Taguchi approach is the reduction of number of experiments. So called *Orthogonal*

Arrays are used for this purpose. These arrays define the setting of particular factors for each experiment. These orthogonal arrays are the basic knowledge of Taguchi approach of DOE.

We can demonstrate the application of orthogonal arrays on L-9 array for 9 tests in the experiment. This array is determined for 4 factors, 3 levels each. If we would use the full factorial design, it will be necessary to carry out 3^4 (81) tests.

The key aspect of this method is the correct choice of the orthogonal array. For the simple situations there are the tables or modifications of these basic tables to according our requirements that we can find in the expert literature [4].

4. D – OPTIMNAL PLANS

The D – optimal plans are based on the full factorial design that is organized into the *matrix of candidate points* $\boldsymbol{\xi_N}$, where N is the overall number of test in the experiment. From this matrix of candidate points is compiled the design matrix **X**. For this compilation is necessary to know the mathematical model of experiment. In the practical usage of D – optimal plans it is mostly a linear model that might look like this:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + e, \qquad (1)$$

where $\boldsymbol{y}$ is a response, $\boldsymbol{\beta_i}$ is a coefficient of the model for each variable, $\boldsymbol{x_i}$ is an independent variable and $\boldsymbol{e}$ is a random error. The size of this random error has a key influence to the reliability of regression models (if the mathematical model is correctly chosen).

The minimum number of tests is determined by the number of elements in the model. The number of experiments depends only on our judgement. There is no definition, how many tests should be performed. It depends on the consideration of the experimenter according to the complexity in terms of time, finance or required time. The number of experiments is crucial how many design matrices can we pick from the matrix of candidate points.

From these matrices adjusted according the mathematical model we choose the best design matrix $\mathbf{X}^*$ called *optimal design*. For optimality determination there are several criteria. For D – optimal plans it is used D – criterion:

$$det(\boldsymbol{X}^{*T}\boldsymbol{X}^{*}) = max\ det(\boldsymbol{X}^{T}\boldsymbol{X}), \qquad (2)$$

where **X** is any design matrix, $\mathbf{X}^T$ is transposition of this matrix, $\mathbf{X}^*$ is optimal design matrix and $\mathbf{X}^{*T}$ is transposition of this matrix.

The product of $\mathbf{X}^T\mathbf{X}$ is called the information matrix and then the experiment is called D – optimal, if the determinant of this information matrix is maximal. One of the disadvantages of this method is very large computational complexity. E.g. for 4 factor, 3 levels each and 25 tests it is $5{,}25 \cdot 10^{20}$ combinations of design matrix. For that reason are different numerical methods used for D – Optimal planning but this is beyond this paper.

5. REGRESSION ANALYSIS FOR EVALUATING D – OPTIMAL PLANS

The relationship between single depend variables (or response) and several independent variables is characterized by a mathematical model. This model is called regression model. This model is in most cases unknown and the aim of regression analysis is to find this approximated model that will fit on the set of experimental data with the highest reliability.

The method of least squares is the most used method for estimating the regression coefficient in a multiple linear regression model. The method of least squares has these assumptions that must be followed:

I. Parameters of regression model $\boldsymbol{\beta}$ (Eq. 1) can take arbitrary values. In practice, this assumption is quite difficult to keep because of limitation of model parameters. These restrictions are mostly based on the physical meaning of this model.
II. Regression model is a linear model.
III. No two columns of the matrix **X** are not collinear vectors, i.e. parallel vectors. It follows that the information matrix $\mathbf{X}^T\mathbf{X}$ is a regular symmetric matrix, to which there is an inverse matrix with determinant larger than zero. This assumption requires that between independent variables wasn't any functional linear dependence, i.e. in the matrix **X** must not be linear dependent columns. The number of independent variables must not be higher than number of observations. In practice the number of observation should be significantly higher.
IV. The random errors have a zero mean value $E(\varepsilon) = 0$, the variance is clearly intended $D(\varepsilon) = \sigma^2$. There is an assumption of normal distribution of probability $N\ (0;\ \sigma^2)$ and that the errors are mutually uncorrelated and independent.

6. TEST FOR SIGNIFICANCE OF REGRESSION

These significance tests are helpful for measuring the estimated model usefulness. The test for significance of regression model is determining whether a linear regression exist a relationship between the response variable *y* and a subset of independent variables $x_1, x_2, \ldots x_k$. The appropriate hypotheses for this test are:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0,$$

$$H_1 : \beta_j \neq 0, \quad \text{for at least one } j. \qquad (3)$$

The rejection of H_0 hypothesis implies that in the independent variables is at least one that significantly

contributes to the model. The test procedure involves an analysis of the variance partitioning to the total sum of the squares SS_T into a sum of a squares due to the regression model and a sum of squares due to residual error:

$$SS_T = SS_R + SS_E. \quad (4)$$

If the null hypothesis H_0 is true, then SS_R/σ^2 is distributed as χ_k^2 with number of degrees of freedom equal to the number of independent variables in the model. The test procedure for H_0 is:

$$F_0 = \frac{SS_R/k}{SS_E/(n-k-1)} = \frac{MS_R}{MS_E}. \quad (5)$$

Tab. 1 ANOVA for Significance of Multiple Regressions

Source of variation	Sum of squares	Degrees of freedom	Mean Square	F_0
Regression	SS_R	k	MS_R	MS_R/MS_E
Error of residual	SS_E	n - k - 1	MS_E	
Total	SS_T	n - 1		

The general matrix notation of linear model is:

$$\boldsymbol{y = X\beta + \varepsilon}. \quad (6)$$

The least square estimator of **β** is:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}. \quad (7)$$

The sums of squares for ANOVA in tab. 1 are:

$$SS_R = \hat{\boldsymbol{\beta}}'\boldsymbol{X'y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}, \quad (8)$$

$$SS_E = \boldsymbol{y'y} + \hat{\boldsymbol{\beta}}'\boldsymbol{X'y}, \quad (9)$$

$$SS_T = \boldsymbol{y'y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}. \quad (10)$$

These tests are mostly performed with statistical regression software. In our case it is *DESTRA* from *Q-DAS GmbH*.

The coefficient of multiple determinations R^2 measures the amount of reduction in the variability of *y* obtained by using independent variables. The large value of R^2 does not necessary means that the regression model is the best. Adding a variable into the model (regardless if it is statistically significant or not) will always increase R^2.

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}. \quad (11)$$

Because R^2 is always increasing sometimes it is preferred usage of an *adjusted R^2 statistic*:

$$R_{adj}^2 = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)}. \quad (12)$$

The usage of adjusted is better, because when we added the statistically insignificant variables, the adjusted coefficient will decrease.

7. THE INFLUENCE OF MEASUREMENT ACCURACY

The method of least squares is quite sensitive to the size of random error. This we can demonstrate on the simple model example. In our model experiment we have 4 factors with 2, 3, 4 and 2 levels. In this experiment we will demonstrate the influence of measurement accuracy with generated errors that will increment the response of the experiment. We will generate the random errors. These errors have the normal distribution of probability N $(0;\sigma^2)$.

The mathematical model of our exemplary experiment is:

$$y = 2{,}3 + 1{,}5X_3 + 1{,}15X_4 + 0{,}35X_1X_2 + e,$$

where *A, B, C, D* represents the individual factors and *e* represents the random error. In our case this represents the measurement error.

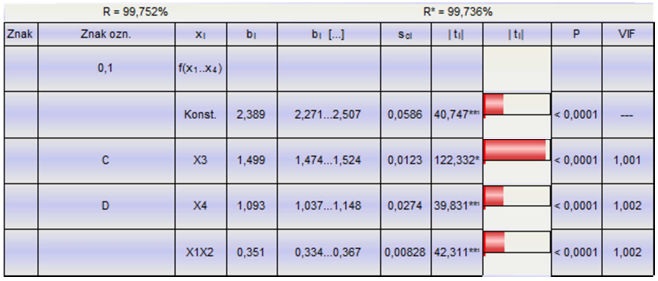

Fig. 2: The output of the statistical software for error distribution N (0;0,1)

At the Figure 1 we can see that with the small measurement error we can expect very precise estimates of the regression model coefficient.

On the second figure is situation, where the error that represents the measurement accuracy is little bit larger. Though R^2 = 98,91% we can see that the estimated regression model is very different from the exemplary experiment.

In Figure 3 we can see that with increasing standard deviation which represents the measurement accuracy significantly decreases the accuracy of estimates of the regression model. When the measurement accuracy is too low than the estimated regression models is not vary precise.

R = 99,043%					R² = 98,910%				
Znak	Znak ozn.	x_i	b_i	b_i [...]	s_{ci}	$\|t_i\|$	$\|t_i\|$	P	VIF
	0,25	$f(x_1..x_7)$							
		Konst.	1,696	1,220...2,173	0,236	7,184***		< 0,0001	---
	A	X1	0,570	0,216...0,924	0,176	3,244**		0,00228	9,495
	C	X3	1,691	1,529...1,853	0,0802	21,084***		< 0,0001	9,897
	D	X4	0,980	0,744...1,217	0,117	8,374***		< 0,0001	4,212
		X1X2	0,189	0,087...0,291	0,0507	3,724***		0,00056	8,663
		X1X3	-0,107	-0,210...-0,004	0,0512	2,093*		0,0423	15,32
		X2X4	0,131	0,031...0,232	0,0500	2,627*		0,0119	8,927

Fig. 3: : The output of the statistical software for error distribution N (0;0,25)

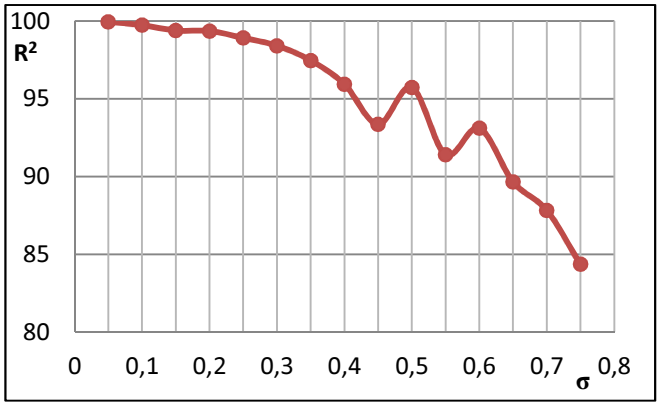

Fig. 4: Dependence of adjusted R^2 statistic on the standard deviation of random error

8. CONCLUISON

This paper presents the influence of accuracy on the reliability of estimates of regression coefficients. When performing experiments with a reduced number of experiments, this dependence is even greater. This should be kept in mind when measuring system is selected for the experiment. The accuracy of this measurement system should be maximized, because as has been shown, this accuracy has quite a significant impact on the reliability of the regression analysis.

9. ACKNOWLEDGEMENTS

This research was funded by the Ministry of Education, Youth and Sports of the Czech Republic, MSM 4977751310 – Diagnostics of Interactive Processes in Electrical Engineering and SGS-2012-026.

10. REFERENCES

[1] MONTGOMERY, Douglas C. *Design and analysis of experiments*. 7th ed. Hoboken: John Wiley , 2009, 656 p. ISBN 978-047-0398-821

[2] DE AGUIAR, P. F., et al. *D-optimal designs*. Chemometrics and Intelligent Laboratory Systems . 1995, vol. 30, Issue 2, p. 199-210.

[3] MITCHELL, T. J. *An Algorithm for the Construction of D-Optimal Experimental Designs.* Technometrics. May 1974, Vol. 16, No. 2, p. 203-210.

[4] ROY, Ranjit K. *Design of Experiments Using The Taguchi Approach*. Toronto: John Wiley and Sons, 2001. 538 s. ISBN 0-471-36101-1

[5] MELOUN, M. Compendium of statistical data processing. 1st ed. Prague: : Academia, 2002, 764 p. ISBN 80-200-1008-4

[6] TŮMOVÁ, O. *Metrology and process evaluation*. 1st ed. Prague: BEN 2009, 232p. ISBN 978-80-7300-249-7