Regression assessment of the model based on the experimental planning matrix in composite materials’ analysis problems

Elita Balatkhanova¹*, Irina Erofeeva², and Victor Afonin²

¹Grozny State Oil Technical University named after acad. M.D. Millionshchikov, Grozny, Russia
²National Research Mordovian State University named after N.P. Ogarev, Saransk, Russia

Abstract. An automated coefficients calculation of the regression model based on the experimental data in the form of a planning matrix is considered. The calculations are based on polynomial regression with possible consideration of the interaction effects between its factors. The application of the Moore – Penrose pseudoinverse matrix for determining the regression equation coefficients is shown. The choice of the calculated planning matrix is carried out taking into account the determination coefficient value and the factors’ calculated matrix rank. Calculation verification is carried out using the rank correlation between the experimental and calculated response functions. The experimental part of the work is given to determine the dependence of fungal resistance and fungicidal properties of filled cement composites on the type and grain size composition of the filler.

1 Introduction

Composite materials, cement composites are widely studied in order to determine both the most preferable composition and to analyze their properties under various operating conditions [1–3]. Much attention is paid to cement composites containing fillers of various types [4–6]. In many cases, the problem of determining a regression model that reflects the investigated relationships between the properties of composite materials is posed and solved [7–9]. For this, the mathematical apparatus of the theory of experiment planning and statistical methods of processing experimental data are used [10–12]. To check the adequacy of the results obtained, such an indicator as the determination coefficient R² can be used [12], which gives an opportunity to compare the values’ closeness of the experimental and calculated response functions. In addition, the rank correlation coefficients can be used [13].

Experimental data analysis can be performed using regression analysis methods and experiment planning. To obtain a calculated adequate regression model, interaction effects can be introduced into the planning matrix, and various existing regression methods can be applied. This paper highlights the results of a computational planning matrix formation using polynomial regression, which lays down in its implementation the principle of expanding

* Corresponding author: elita.74@mail.ru
differentiable functions in a Taylor series. The number of terms in polynomial regression can be increased by introducing the real factors’ interaction effects and increasing polynomial degree of the regression model. It can be noted that in modern programming environments such as Python there are modules / classes to implement polynomial regression.

## 2 Materials and methods

An example of an experimental planning matrix presentation is presented in Table 1.

| No. | Working Factor Matrix X | Response function values |
|-----|-------------------------|-------------------------|
|     | X₁₁  X₂₁  X₃₁  ...  Xₖ₁ | Y₁                      |
| 1   | X₁₂  X₂₂  X₃₂  ...  Xₖ₂ | Y₂                      |
| 2   | X₁₃  X₂₃  X₃₃  ...  Xₖ₃ | Y₃                      |
| ... | ...  ...  ...  ...  ... | ...                     |
| n   | X₁ₙ  X₂ₙ  X₃ₙ  ...  Xₖₙ | Yₙ                      |

As it is possible to see from the Table 1, the working factor matrix (X) is generally rectangular, i.e., the number of its rows is not equal to the number of columns (n ≠ k). In this case, it is assumed that the rank of the matrix X is equal to the dimension of the response function Y, i.e., the rank (X) = n. In this case, it seems possible to resolve the regression equation relative to the vector of coefficients:

\[ Y = X \cdot C \]  

(1)

Where C — is the vector of coefficients.

To solve the equation (1) with respect to the vector of coefficients, we multiply it on the left by the pseudoinverse matrix \((X^+)\) Moore-Penrose [14]:

\[ X^+Y = X^+X \cdot C \]  

(2)

If the rank of the matrix X is n, then the matrix \(X^+X\) will be unit, which entails the desired solution with respect to the matrix of coefficients:

\[ C = X^+Y \]  

(3)

When the rank of the matrix X is less than the value n, then it is necessary to search for a conditionally extended matrix of factors \(X_p\) with the introduction of interaction effects and taking into account the degrees of factors in accordance with polynomial regression. Then equation (3) will have the form:

\[ C = X_p^+ \cdot Y \]  

(4)

Polynomial regression is quite fully represented in the Python toolbox based on the Polynomial Features module. Now the calculated response function \(Y_p\) will be defined as:

\[ Y_p = X_p \cdot C \]  

(5)
It is advisable to check the calculations results for adequacy in terms of determining the rank correlation between $Y$ and $Y_p$ using the methods of Pearson (Pearson's Linear Correlation Coefficient), Kendall's Tau Coefficient, Spearman's Rho. These methods are implemented, for example, in the MATLAB environment.

### 3 Research results

As an object of research, we used the results from [15], some of which are shown in Table 2.

**Table 2.** Test results of testing compositions with fillers

| Experiment no. | Planning matrix in natural values, wt. h. | Relative indicators compositions of cement composites (response functions) |
|----------------|------------------------------------------|---------------------------------------------------------------------|
|                | X<sub>1</sub> X<sub>2</sub> X<sub>3</sub> | with river limestone with mountain dolomite with organic limestone |
| 1              | 1 0 0                                      | K<sub>g</sub> (Y<sub>1</sub>) 1.00 K<sub>c</sub> (Y<sub>2</sub>) 1.00 K<sub>g</sub> (Y<sub>3</sub>) 1.00 |
| 2              | 0 1 0                                      | 1.00 1.00 1.00 0.80 0.83 0.92 |
| 3              | 0 0 1                                      | 1.00 1.00 1.00 0.73 0.75 0.83 |
| 4              | 1/3 2/3 0                                 | 1.12 0.92 1.00 0.73 0.83 0.83 |
| 5              | 1/3 0 2/3                                | 0.75 0.92 0.33 0.73 0.92 0.92 |
| 6              | 0 1/3 2/3                                | 0.87 1.00 1.00 0.80 0.67 1.00 |
| 7              | 2/3 1/3 0                                | 0.62 1.00 1.33 0.80 0.50 1.08 |
| 8              | 2/3 0 1/3                                | 0.75 1.00 1.33 0.73 0.50 0.83 |
| 9              | 0 2/3 1/3                                | 1.00 1.00 1.00 0.73 0.67 1.17 |
| 10             | 1/3 1/3 1/3                              | 0.74 0.92 0.33 0.87 0.75 0.92 |

The values $X_1$, $X_2$, $X_3$ are the column values of planning working matrix in Table 2. Obviously, the rank of such a matrix will not exceed the number three, which is less than the number of experiments $n = 10$. Therefore, the research was carried out to search for a working planning matrix with the interaction effects and the corresponding degree of the polynomial, with which the coefficient of determination is practically equal to one and the rank of the matrix is 10. In accordance with the formulas (1) - (5), the calculated values of the regression model coefficients were obtained and an estimate was made with the correspondence of the calculated values of the response function $Y_p$ with the experimental values $Y_1$, $Y_2$, $Y_3$, $Y_4$, $Y_5$, $Y_6$. In all cases, the calculated value of the coefficient of determination $R^2 = 1$. Table 3 shows the results of rank correlation.

**Table 3.** Rank correlation results

| Comparison | Pearson | Kendall | Spearman |
|------------|---------|---------|----------|
|            | rho     | pval    | rho      | pval    |
| (Y<sub>1</sub> ↔ Y<sub>3</sub>) | 0.894427 | 9.758788e-04 | 0.956636 | 1.468026e-05 |
| (Y<sub>2</sub> ↔ Y<sub>5</sub>) | 0.715678 | 2.103526e-02 | 0.807573 | 4.721383e-03 |
| (Y<sub>3</sub> ↔ Y<sub>4</sub>) | 0.797724 | 5.366337e-03 | 0.883452 | 6.997192e-04 |
| (Y<sub>4</sub> ↔ Y<sub>6</sub>) | 1.701606e-01 | 8.432744e-03 | 0.884516 | 1.302404e-04 |
| (Y<sub>5</sub> ↔ Y<sub>6</sub>) | 0.954521 | 2.765139e-04 | 0.987804 | 9.537128e-08 |
| (Y<sub>6</sub> ↔ Y<sub>6</sub>) | 6.646897e-01 | 6.135913e-04 | 0.972345 | 2.475112e-06 |

Table 3 the following notation is used:
rho – is the correlation coefficient, which can vary from -1 to +1; a value of +1 corresponds to an almost direct functional relationship; 

pval – is the probability that indicates rejection of the hypothesis that there is no correlation between Y and Yp;

symbol ↔ means the operation of calculating the rank correlation between two arrays. 

Table 4 shows the calculated values of the regression models’ coefficients.

| Coefficients’ designations in | Values of the regression models’ coefficients |
|-------------------------------|---------------------------------------------|
|                              | Assessment Y1                               | Assessment Y2 |
| a0                            | 0.485027                                    | 0.621758    |
| a1                            | -0.186539                                   | 0.250110    |
| a2                            | 0.596140                                    | 0.185824    |
| a3                            | 0.075426                                    | 0.185824    |
| a4                            | -0.344176                                   | 0.498461    |
| a5                            | -0.171896                                   | -0.124176   |
| a6                            | 0.329533                                    | -0.124176   |
| a7                            | 1.171360                                    | 0.125604    |
| a8                            | -0.403324                                   | 0.184396    |
| a9                            | 0.149217                                    | 0.125604    |
| a10                           | 1.045687                                    | -0.307330   |
| a11                           | -2.12074                                    | 0.434396    |
| a12                           | 0.732211                                    | 0.434396    |
| a13                           | 2.329711                                    | -0.208462   |
| a14                           | -0.379533                                   | -0.350110   |
| a15                           | -0.023146                                   | -0.208462   |
| a16                           | -1.252527                                   | 0.066813    |
| a17                           | 0.094176                                    | 0.267253    |
| a18                           | -0.117967                                   | 0.267253    |
| a19                           | 0.290329                                    | 0.066813    |

In accordance with the coefficients’ designations in Table. 4 the general view of the calculated regression equation will be as follows:

\[ Y_p = a_0 + a_1 X_1 + a_2 X_2 + a_3 X_3 + a_4 X_1^2 + a_5 X_2^2 + a_6 X_3^2 + a_7 X_1 X_2 + \ldots + a_{19} X_1 X_2 X_3. \]

Figures 1 - 6 show the graphs of the fungal resistance and fungicidal properties’ dependences of filled cement composites on the type and grain size composition of the filler in the form of dependences on the readings of the experimental functions and approximation.
In the form of dependences on the readings of the experimental functions and approximation dependences of filled cement compositions, the correlation between the calculated regression equation will be as follows:

\[ Y_p = a_0 + a_1 X_1 + a_2 X_2 + a_3 X_3 + a_4 X_4 + a_5 X_5 + a_6 X_6 + a_7 X_7 + a_8 X_8 + a_9 X_9 + a_{10} X_{10} + a_{11} X_{11} + a_{12} X_{12} + a_{13} X_{13} + a_{14} X_{14} + a_{15} X_{15} + a_{16} X_{16} + a_{17} X_{17} + a_{18} X_{18} + a_{19} X_{19} \]

The coefficients \( a_0 \) to \( a_{19} \) are determined by calculating the rank correlation between two arrays and the correlation coefficient, which can vary from 0 to 1. Table 4 shows the calculated values of the regression models' coefficients:

| \( \rho \) | \( p \) | \( T \) |
|------------|-------|-------|
| 0.117967   | -0.403324 | 2.122074 |
| 0.290329   | 0.094176  | 0.434396 |
| 2.329711   | 1.045687  | 0.149217 |
| 0.732211   | 0.117967   | 0.184396 |

Coefficient \( \rho \) stands for the correlation coefficient, which can vary from 0 to 1. The probability \( p \) is the value of the Student's t-test and is used to determine the significance of the correlation. The \( T \) value is the Student's t-test and is used to determine the significance of the correlation. The symbols \( \rho \) and \( p \) are introduced in Table 4, which shows the calculated values of the regression models' coefficients.

In accordance with the assigned designations in Table 4, the graphs of the regression models' coefficients are shown in Figures 1-3:

**Fig. 1.** Composition indicator \( K_3 (Y_1) \) as a function of its countings

**Fig. 2.** Composition indicator \( K_c (Y_2) \) as a function of its counts

**Fig. 3.** Composition indicator \( K_d (Y_3) \) as a function of its counts
**Fig. 4.** Composition indicator $K_c (Y_4)$ as a function of its counts

**Fig. 5.** Composition indicator $K_g (Y_5)$ as a function of its counts

**Fig. 6.** Composition indicator $K_c (Y_6)$ as a function of its counts
4 Conclusion

An engineering technique that allows, to a certain extent, to optimize and automate the process of determining the coefficients of the regression equation, which reflects the calculated response function is considered and proposed. To solve the problem, a pseudo inverse Moore – Penrose matrix and a sequential study of the computational planning matrix dimension were used, in which the effects of real factors’ interaction can be taken into account. To check the adequacy of the obtained regression models, the methods of rank correlation of Pearson, Kendall, Spearman were used, as well as the calculation of the values of the determination coefficient $R^2$. The check was carried out on the study of the fungal resistance and fungicidal properties’ dependence of filled cement composites on the type and grain size composition of the filler.

References

1. V.T. Erofeev, I.V. Smirnov, P.V. Voronov, V.V. Afonin, E.N. Kablov, O.V. Startsev, V.O. Startsev, I.M. Medvedev, Fundamental research 11–5, 911–923 (2016).
2. V.G. Soloviev, E.A. Shuvalova, V.V. Shalamov, Bulletin of higher educational institutions. Construction 1, 78–86 (2020). DOI: 10.32683/0536-1052-2020-733-1-78-86.
3. D.V. Emelyanov, E.M. Balatkanova, I.V. Erofeeva, L.M. Oshkina, A.S. Bogdanov, S.V. Nadorov, P.A. Efremov, Cement composites with the addition of dolomite filler, New in architecture, design of building structures and reconstruction: materials of the IV International (X All-Russian) conference NASKR-2018. Publishing house of Chuvash, university, Cheboksary, 2018.
4. V.I. Kalashnikov, V.T. Erofeev, O. V. Tarakanov, Bulletin of higher educational institutions. Construction 4, 35–37.
5. V. Erofeev, A. Dergunova, A. Bogatov, N. Letkina, V. Smirnov, Materials Science Forum 974, 305–311 (2020).
6. T. Elchishcheva, V. Erofeev, Materials Science Forum 1011, 179–187 (2020).
7. I.N. Maksimova, N.I. Makridin, V.T. Erofeev, Regional architecture and construction 1 (46), 28–34 (2021).
8. V.T. Erofeev, V.F. Smirnov, D.A. Svetlov [et.al.], Privolzhsky scientific journal 2, 41–51 (2014).
9. E.N. Kablov V.T. Erofeev, A.V. Dergunova, E.V. Deraeva, D.A. Svetlov, Journal of Physics: Conference Series. "International Conference on Engineering Systems 2020" 012029 (2020).
10. V.B. Tikhomirov, Planning and analysis of the experiment (Light industry, Moscow, 1974).
11. V.A. Lunev, Mathematical modeling and experiment planning: textbook. (SPb., Publishing house of Polytechnic university, 2012).
12. A.M. Dubrov, V.S. Mkhitaryan, L.I. Troshin, Multivariate statistical methods: Textbook (Finance and statistics, Moscow, 2003).
13. J.D. Gibbons, S. Chakraborty, Nonparametric Statistical Inference (N.Y. e.a.: Marcel Dekker, 1992).
14. D.V. Beklemishev, Additional chapters in linear algebra (Science, Moscow, 1983).
15. I.V. Erofeeva, Physical and mechanical properties, biological and climatic resistance of powder-activated concrete. PhD Thesis, Penza, 2018.