Practical application of regression analysis to study load parameters of mine excavator equipment

V S Velikanov, N V Dyorina, E I Rabina and T Yu Zalavina
Nosov Magnitogorsk State Technical University, Magnitogorsk, 455000, Russia

E-mail: rizhik_00@mail.ru

Abstract. The main failure causes of the working equipment of mine excavators with gear rack pressure are: a high level of their dynamic loading, exceeding the permissible level; a construction form leading to a high concentration of stresses; consequences of repair actions; relatively low qualifications of drivers and other factors. Using regression analysis of statistical data from a sample of a computational experiment, the authors of the article obtained an original mathematical model in which heterogeneous factors are interconnected.

1. Introduction
A significant problem in setting the actual loads arising in the working equipment of mining excavators is the presence of a large number of operating factors that are difficult to describe by mathematical formulas, many of which are random. The study of loads in working equipment when managing a mining excavator, most often, defies to traditional methods of analysis and modeling [1].

2. Methodology
In the general case, when analysing the results of scientific research, there is a situation in which the quantitative change in the studied quantity (voltage in the handle - \( \sigma \)) depends not on one quantity, but on several factors, namely: density \( (x_1) \), dust \( (x_2) \), experience \( (x_3) \), illumination \( (x_4) \), learnability \( (x_5) \), manoeuvrability \( (x_6) \), mass \( (x_7) \), noise level \( (x_8) \), serviceability \( (x_9) \), technological ef \( (x_{10}) \), vibration \( (x_{11}) \), speed \( (x_{12}) \) [1].

The process of analysing pair dependencies is easier to implement by graphical interpretation of the results. In the study of multifactorial dependencies, this possibility is practically excluded due to the limited perceived dimension of space. All studies are actually carried out in \( p + 1 \) - dimensional space, where \( p \) is the number of influencing factors, \( p = 12 \):

\[
(\sigma) = f(speed, vibration, ..., x_{12})
\]  

(1)

In this case, the observation results are not two column vectors, but a matrix of dimension \( p+1, n \), where \( p \) is the number of factors, \( n \) is the number of experiments 130. In accordance with the obtained sample of a computational experiment, the results matrix has the form [1-3].
The essence of the method is as follows: find such values of the regression parameters at which the
regression model allows you to take into account any number of factors, practically this lacks necessity. The selection of factors is based on qualitative analysis. However, theoretical analysis often does not allow an unambiguous answer to the question of the quantitative relationship of the considered features and the advisability of including the factor in the model.

Thus, theoretically, the regression model allows you to take into account any number of factors, practically this lacks necessity. The selection of factors is based on qualitative analysis. However, theoretical analysis often does not allow an unambiguous answer to the question of the quantitative relationship of the considered features and the advisability of including the factor in the model.

Therefore, the selection of factors is usually carried out in two stages: at the first, factors are selected based on the nature of the problem; on the second, student statistics for regression parameters are determined on the basis of a matrix of correlation indices [4-6].

If \( f(x_1, x_2, \ldots, x_{12}) \) is a linear combination of factors, then the regression is linear and can be written as:

\[
\hat{y} = b_0 + \sum_{j=1}^{12} b_j x_j,
\]

where \( b_0, b_j \) are the regression parameters, \( j = 1, p \).

For a linear model, the interpretation of the parameters is as follows: \( b_0 \) - characterizes the magnitude of the influence of all unaccounted factors selected to describe the process; \( b_j \) is the linear rate of change of the dependent feature, which shows how many units the response function will increase on average if the corresponding factor \( j \) is increased by 1 unit; \( b_1 + b_2 + \cdots + b_p \) - shows how many units the response function will increase on average if each of the factors is increased by 1 unit. To determine the parameters, the least-squares technique is most often used.

The essence of the method is as follows: find such values of the regression parameters at which the sum of the squares of the deviation of the initial values of the response function and theoretical (calculated by regression) takes a minimum value.

The analytical expression of the method is as follows:

\[
S = \sum_{i=1}^{n} \left( y_i - \hat{y}(x_{1i}, x_{2i}, \ldots, x_{pi}) \right)^2 \rightarrow \min.
\]

For linear regression \( \hat{y} = b_0 + b_1 x_{1i} + \cdots + b_p x_{pi} \) the expression has the form:

\[
S = \sum_{i=1}^{n} \left( y_i - \hat{y}(x_{1i}, x_{2i}, \ldots, x_{pi}) \right)^2 = \sum_{i=1}^{n} \left( y_i - (b_0 + b_1 x_{1i} + \cdots + b_p x_{pi}) \right)^2.
\]

We got a function of many variables. In order to determine the minimum of the function, it is necessary to determine the partial derivatives with respect to all incoming parameters:

\[
\frac{\partial S}{\partial b_j} = 2 \cdot \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{1i} - \cdots - b_p x_{pi}) \cdot (x_{ji}), \text{npu } j = 1, p.
\]
The resulting expressions are equal to zero.
As a result, after the conversion, we obtain a system of normal equations:

\[
\begin{align*}
nb_0 + b_1 \sum_{i=1}^{n} x_{i1} + b_2 \sum_{i=1}^{n} x_{i2} + \ldots + b_p \sum_{i=1}^{n} x_{ip} &= \sum_{i=1}^{n} y_i, \\
b_0 \sum_{i=1}^{n} x_{i1} + b_1 \sum_{i=1}^{n} x_{i1}^2 + b_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \ldots + b_p \sum_{i=1}^{n} x_{i1}x_{ip} &= \sum_{i=1}^{n} y_i x_{i1}, \\
\cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots
\end{align*}
\]

or in matrix form

\[
(X^*X)B = X^*Y,
\]

where B is the column vector of the sought coefficients of the approximating function:

\[
B = \begin{pmatrix}
b_0 \\
b_1 \\
b_2 \\
\vdots \\
b_p
\end{pmatrix},
\]

\[
X = \begin{pmatrix}
1 & x_{11} & x_{21} & \ldots & x_{p1} \\
1 & x_{12} & x_{22} & \ldots & x_{p2} \\
1 & x_{13} & x_{23} & \ldots & x_{p3} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & x_{1n} & x_{2n} & \ldots & x_{pn}
\end{pmatrix}
\]

and

\[
Y = \begin{pmatrix}
y_1 \\
y_2 \\
y_3 \\
\vdots \\
y_n
\end{pmatrix}
\]

where X is the matrix of all values of the considered factors obtained during measurements or observations; (1) - a column vector defining the free term of the regression equation (this column consists of units in the source data matrix); Y is a column vector of experimental values of the studied quantity; X* - is the matrix transposed to the matrix X. For the coefficients of the system, we obtain the matrices X*X - the covariance matrix and X*Y - the column vector of free terms:

\[
X^*X = \begin{pmatrix}
\sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i}^2 & \cdots & \sum_{i=1}^{n} x_{1i}x_{pi} \\
\sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{2i} & \cdots & \sum_{i=1}^{n} x_{2i}x_{pi} \\
\vdots & \vdots & \ddots & \vdots \\
\sum_{i=1}^{n} x_{1i}x_{pi} & \sum_{i=1}^{n} x_{2i}x_{pi} & \cdots & \sum_{i=1}^{n} x_{pi}^2
\end{pmatrix},
\]

\[
X^*Y = \begin{pmatrix}
\sum_{i=1}^{n} y_i \\
\sum_{i=1}^{n} y_i x_{1i} \\
\vdots \\
\sum_{i=1}^{n} y_i x_{pi}
\end{pmatrix}
\]

In order to assess the applicability of the constructed empirical model for the loading parameters of a quarry excavator handle and for subsequent forecasting and control, three statistical indicators are used (F - Fisher statistics, T – Student’s statistics; A - average relative approximation error) and the results of the analysis of residuals (\( \varepsilon_i \) - residuals).

3. Implementation

For statistical data analysis, the Statistika program was used.

We assume that there is a linear relationship between predictor variables and voltages in the handle of a mining excavator (table 1).

P-value – probability to determine the significance of the regression coefficient. In cases where p-value > 0.05, the coefficient can be considered zero, which means that the corresponding independent variable practically does not affect the dependent variable, therefore the predictor variables \( x_2, x_4, x_5, x_9 \), \( x_{10} \) are excluded from the model (table 2).

The multiple correlation coefficient R=0.93 indicates a very strong relationship between the whole set of factors and the result.

The multiple determination coefficient \( R^2 \) characterizes the percentage of the constructed multiple regression model that explains the variation in the values of the productive variable relative to its average level, i.e., it shows the share of the total variance of the productive variable explained by the variation
of the factor variables included in the regression model. This means that at $R^2 = 0.87$ the variation in stress in the handle (σ) by 87% is explained by the variation of factor attributes: $x_1$ - density of the excavated rock mass; $x_2$ - the experience of the driver; $x_3$ - controllability; $x_4$ - the mass of the handle, ladle and excavated rock; $x_5$ - noise level in the cabin; $x_{12}$ - vibration; $x_{13}$ - ladle lift speeds. If $R^2$ lies in the range from 0.8 to 0.95, they speak of a satisfactory approximation (the model as a whole is adequate to the described phenomenon).

Table 1. The results of the regression analysis (stage 1).

| N=130 | Regression Summary for Dependent Variable: stress $R=0.93575440$, $R^2=0.87563630$ Adjusted $R^2=0.86288105$, F(12,117)=68.649 p<0.0000 Std. Error of estimate: 17.990 |
|-------|-------------------------------------------------------------------------------------------------|
|       | $b^*$ | Std.Err. of $b^*$ | $b$ | Std.Err. of $b$ | t(122) | p-value |
| Intercept | 23.91919 | 0.38421 | 3.66264 | 0.631856 |
| density ($x_1$) | 0.211884 | 0.057850 | 1.4072 | 0.000376 |
| dust ($x_2$) | -0.041644 | 0.054148 | -1.0065 | 0.443934 |
| experience ($x_3$) | -0.120731 | 0.042356 | -1.0765 | 0.005162 |
| illumination ($x_4$) | -0.035696 | 0.069763 | -0.0717 | 0.609847 |
| learnability ($x_5$) | -0.083045 | 0.043327 | -4.4204 | 0.007714 |
| manoeuvrability ($x_6$) | 0.152774 | 0.065282 | 6.0300 | 0.009676 |
| mass ($x_7$) | 0.172789 | 0.055914 | 1.4733 | 0.002499 |
| noise_level ($x_8$) | 0.353814 | 0.066144 | 0.7156 | 0.349120 |
| serviceability ($x_9$) | -0.003207 | 0.063161 | -0.1184 | 0.304527 |
| technological_ef ($x_{10}$) | -0.051850 | 0.050276 | -1.6344 | 0.000000 |
| vibration ($x_{11}$) | 0.360483 | 0.048136 | 3.0342 | 0.000000 |
| speed ($x_{12}$) | 0.317336 | 0.058045 | 80.7493 | 0.000000 |

Table 2. The final results of the statistical analysis of the data sample.

| N=130 | Regression Summary for Dependent Variable: stress $R=0.93097639$, $R^2=0.86671704$, Adjusted $R^2=0.85906965$, F(7,122)=113.34 p<0.0000 Std. Error of estimate: 18.239 |
|-------|-------------------------------------------------------------------------------------------------|
|       | $b^*$ | Std.Err. of $b^*$ | $b$ | Std.Err. of $b$ | t(122) | p-value |
| Intercept | 15.6998 | 0.34977 | 5.11643 | 0.011844 |
| density ($x_1$) | 0.26945 | 0.052664 | 1.7896 | 0.000001 |
| experience ($x_3$) | -0.14138 | 0.040322 | -1.2607 | 0.000637 |
| manoeuvrability ($x_6$) | 0.141613 | 0.053275 | 5.5895 | 0.008909 |
| mass ($x_7$) | 0.175005 | 0.054958 | 1.4922 | 0.001842 |
| noise_level ($x_8$) | 0.37219 | 0.04103 | 0.7528 | 0.097110 |
| vibration ($x_{11}$) | 0.346477 | 0.045311 | 2.9163 | 0.000000 |
| speed ($x_{12}$) | 0.289182 | 0.056711 | 73.5853 | 0.000000 |
An assessment of the regression equation reliability as a whole and the coupling tightness index is given by the F - Fisher criterion. In our study, the calculated Fisher coefficient is many times greater than the critical. This indicates that the resulting equation is statistically significant, stable and reliable.

Then the linear multiple regression equation takes the following form:

\[ (\sigma) = -40.12 + 1.79 \cdot x_1 - 1.26 \cdot x_3 + 5.59 \cdot x_6 + 1.49 \cdot x_7 + 0.75 \cdot x_8 + 2.92 \cdot x_{11} + 73.59 \cdot x_{12} \]  \hspace{1cm} (10)

However, the “pure” regression coefficients in the formula (10) cannot allow determining the influence of each factor individually on the result. For this, the regression equation is constructed on a standardized scale:

\[ t_{x_a} = \beta_1 t_{x_1} + \beta_3 t_{x_3} + \beta_6 t_{x_6} + \beta_7 t_{x_7} + \beta_8 t_{x_8} + \beta_{11} t_{x_{11}} + \beta_{12} t_{x_{12}}, \]  \hspace{1cm} (11)

where \( t_{x_a}, t_{x_1}, t_{x_3}, t_{x_6}, t_{x_7}, t_{x_8}, t_{x_{11}}, t_{x_{12}} \) - standardized variables;

\( \beta_1, \beta_3, \beta_6, \beta_7, \beta_8, \beta_{11}, \beta_{12} \) - standardized regression coefficients (table 2, column - b *).

Standardized regression coefficients show the influence degree of factors on the resulting indicator. The standardized regression coefficients are calculated by the formula:

\[ \beta_i = \frac{b_i \cdot \delta x_i}{\delta y}, \]  \hspace{1cm} (12)

where \( b_i \) – multiple regression coefficients;

\[ \delta_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}}, \]  \hspace{1cm} (13)

\[ \delta x_i = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x}_i)^2}{n}}. \]  \hspace{1cm} (14)

The linear regression equation on a standardized scale is:

\[ t_{x_a} = 0.27 t_{x_1(0.022)} + 0.14 t_{x_3(0.040)} + 0.14 t_{x_4(0.053)} + 0.18 t_{x_6(0.055)} + 0.37 t_{x_7(0.041)} + 0.35 t_{x_{11(0.040)}} + 0.29 t_{x_{12(0.057)}} \]  \hspace{1cm} (15)

where the standard errors of parameter calculation are indicated as a subscript in the record of the obtained equation [1].

The interpretation of the obtained model begins with a statistical evaluation of the regression equation as a whole and an assessment of the dependence of the factor attributes included in the model, i.e. ascertaining how they affect the value of a productive attribute. The greater the magnitude of the regression coefficient, the greater the influence of this trait on the modeled one. In this case, the sign in front of the regression coefficient is of particular importance, respectively: the sign (+) - with the increase of this factor, the productive sign increases and if the factor sign has the sign (-), then with its increase the productive sign decreases.

Taking into account the obtained dependence and on the basis of a statistical analysis of a sample of modeling data, the influence of the most significant factors on the stresses in the handle of a mine excavator when digging rock is established, the results are presented in table 3.
Table 3. Summary dependencies of the most significant factors affecting the stresses in the handle of a mine excavator.

| Predictor variables, $x_i$ | Dependence | Correlation ratio $R^2$ |
|---------------------------|------------|------------------------|
| driver experience         | $(\sigma) = 426.03e^{-0.106x_3}$ | 0.97                   |
| ladle lift speed          | $(\sigma) = 67.879e^{1.7465x_{12}}$ | 0.95                   |
| weight (ladle + rock)     | $(\sigma) = 5E-07x^2 - 0.01x_7 + 103.15$ | 0.96                   |
| excavated rock density    | $(\sigma) = 29.26e^{0.0006x_1}$ | 0.98                   |

4. Conclusions
A formalized mathematical model has been created that combines a single variety of factors affecting the efficiency of a mining excavator, as well as allowing us to build a qualitative and quantitative forecast for the operation of an excavator, taking into account the risks of failures and associated uncertainties. The linear multiple regression equation is obtained on a standardized scale:

$$t_\sigma = 0.27t_{v_{(0.052)}} - 0.14t_{v_{(0.040)}} + 0.14t_{v_{(0.055)}} + 0.18t_{v_{(0.055)}} + 0.37t_{v_{(0.041)}} + 0.35t_{v_{(0.040)}} + 0.29t_{v_{(0.011)}}$$

The determination coefficient is 86.7%.

It has been established that with an increase in the ladle lifting speed, determined by the experience of the driver, the tension in the handle increases exponentially $(\sigma) = 67.879e^{1.7465x_{12}} (R^2 = 95\%)$, which proves the feasibility of work to improve the system of in-house training of personnel on the basis of mining enterprises.

References
[1] Dyorina N V, Velikanov V S 2019 Development of theory bases for minimizing failures while driving career excavators 67
[2] Kiefer J 1959 J of the Royal Statistical Society, Sor. B 21
[3] Fedorov V V and Malyutov M B 1972 Math. Operationsforsch and Statist 3
[4] Logunova O S, Ilyina E A and Pavlov V V 2011 Theory and practice of processing experimental data on a computer: textbook. allowance for students Publishing House of Nosov Magnitogorsk State Technical University 300
[5] Logunova O S, Filippov E G, Pavlov V V et al. 2011 Software statistical systems: textbook. allowance for students Publishing Center "Academy" 240
[6] Logunova O S, Lednov A V and Koroleva V V 2014 Bulletin of Nosov Magnitogorsk State Technical University 3