STATISTICAL EVALUATION OF SOIL PROPERTIES

T. M. NAHHAS, PH. D.
Dean Of The Academic Affairs

M. M. EL-SHERIF, PH. D.
Assistant Prof.,
Faculty Of Technology, Abha

ABSTRACT

The multiple linear and nonlinear regression analysis is often used to investigate the relationship between metrically scaled random variables according to a preeffected estimation formula. The parameters of this formula are calculated by minimizing the squared residuals (least squares). The correlation between dependent and independent variables is valuated by coefficient of correlation (r^2). Applying this method; the relative and absolute references between the results and estimated values remain unconsidered. Therefore, the authors suggest four additional steps to optimize the estimation formula according to these aspects. For the evaluation of these results, the "coefficient of optimization" can be defined. These are approximately corresponding to r^2.

To test this suggested extended regression; the author carried out some investigations on various soil parameters gathered in the "Data-Bank" of the Institute of National Research, Cairo, Egypt. In this paper, a representative result of these investigations on the undrained shear strength C_u is presented. This example shows that there may be quite large residual errors although the coefficient of correlation r^2 may be more than (0.83). It is therefore advisable also to take into account these errors of estimation.

1. INTRODUCTION

Statistical and probablistical methods are more and more used in the field of soil mechanics and foundation engineering. The main purpose of these efforts is to estimate the real probability of failure of a construction instead of using the arbitrary of predetermination of a universal factor of safety. However, this is only possible if the engineer or the designer possesses sufficient experience about all influencing factors and their statistical and probablistical parameters. In many cases, the engineer has no possibility gathering enough data to carry out such an investigation. Therefore, it seems to be still more promising to use statistics for the investigation of soil parameters and in the same time improving the applied statistical methods.

Engng. Res. Jour., Vol. 62,pp. 134 -145, APRIL 1999
Helwan University, Faculty of Engng , Mataria Cairo

2. REGRESSION ANALYSIS

The multiple linear and nonlinear regression analyses were often used to investigate the relationship between metrically scaled variables. The main aim of these methods is to determine the regression parameters of the preelected estimation formula by minimizing the squared residuals. The measured y_i-values and the calculated $\overline{y_i}$-values will be minimized according to the following equation:

$$\sum_{i=1}^{n}\left(y_i-\bar{y}_i\right)^2 \Rightarrow \mathit{Minimum} \quad ----------------\rightarrow (1)$$

With:

$$\bar{y} = a_0 + a_1x_1 + a_2x_2 + \cdots + a_kx_k$$

$$a_i = \mathit{optimum\ \ regression\ \ parameters}$$

$$x_i = \mathit{independent\ variables}$$

The main advantage of this method of least squares is that the optimized regression coefficients of the preelected estimation formula $\bar{y}$ can be easily found by solving some linear equation systems.

The reliability of the chosen formula, that means the degree of correlation between the dependent and independent variables, is usually valuated by the multiple coefficients of correlation r^2. A coefficient of $r = +1$ indicates positive and $r = -1$ indicates negative functional correlations.

To get a statement about the magnitude of correlation the coefficient r^2 is often used. This coefficient may be interpreted as the percentage with which the total variance of the test results is explained by the derived estimation formula. The coefficient r^2 can be evidently written by applying the total variance of the test results (A_1) and residual undeclared variance (R_1) as follows:

$$r^2 = \frac{A_1 - R_1}{A_1} \quad --------------------\rightarrow (2)$$

Where:

$$A_1 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{\bar{y}}\right)^2$$

$$R_1 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2$$

With:

$$\bar{\bar{y}} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

However, the significance of r^2 depends on the number of the analyzed data. Therefore, it is at least necessary to give additionally the confidence limits (e.g. for the 95% confidence interval).

Using the optimization method, one has to consider that the values of the least squares remain quite inevident. Besides that the relative and absolute differences between measured test results (y_i) and estimated values ($\overline{y_i}$) remain unvaluated.

Therefore, the authors suggest additional steps to optimize the estimation formula according to these aspects. For the evaluation of the results, "coefficient of optimization" can be defined. They are approximately corresponding to r^2. The suggested optimization criteria and respective "coefficients of optimization" are expressed as follows:

a) Average relative error of estimation:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{y}_i}{y_i}\right) \Rightarrow \textit{Minimum} \qquad ,y_i\neq 0 \qquad \dashrightarrow (3)$$

The respective "coefficient of optimization" can be defined as follows:

$$W_2=\frac{A_2-R_2}{A_2} \qquad \dashrightarrow (4)$$

Where:

$$A_2=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\bar{y}}}{y_i}\right)$$

$$R_2=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{y}_i}{y_i}\right)$$

A_2 = average relative deviation between test results and the mean value.

R_2 = average relative error of estimation.

b) Maximum relative error of estimation ⟹ minimum

$$i=1^{max},\ldots,n\left(\frac{y_i-\bar{y}_i}{y_i}\right) \Rightarrow \textit{Minimum} \qquad ,y_i\neq 0 \qquad \dashrightarrow (5)$$

With the respective "coefficient of optimization" as follows:

$$W_3=\frac{A_3-R_3}{A_3} \qquad \dashrightarrow (6)$$

Where:

$$A_3=1^{max},\ldots,n\left(\frac{y_i-\bar{\bar{y}}}{y_i}\right)$$

$$R_3=1^{max},\ldots,n\left(\frac{y_i-\bar{y}_i}{y_i}\right)$$

A_3= maximum of the relative deviation between test results and mean value.

R_3 = maximum relative error of estimation.

c) Average absolute error of estimation ⟹ minimum

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y}_i\right) \Rightarrow \mathit{Minimum}$$

$$W_4 = \frac{A_4 - R_4}{A_4} \quad \dashrightarrow (7)$$

Where:

$$A_4 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{\overline{y}}\right)$$

$$R_4 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y}_i\right)$$

$\boldsymbol{A_4}$= average absolute deviation between test results and the mean value.

$\boldsymbol{R_4}$ = average absolute error of estimation.

d) Maximum absolute error of estimation: ⟹ minimum

$$i = 1^{max},\ldots,n\ \left(y_i - \overline{y}_i\right) \Rightarrow \mathit{Minimum}$$

$$W_5 = \frac{A_5 - R_5}{A_5} \quad \dashrightarrow (8)$$

Where:

$$A_5 = 1^{max},\ldots,n\ \left(y_i - \overline{\overline{y}}\right)$$

$$R_5 = 1^{max},\ldots,n\ \left(y_i - \overline{y}_i\right)$$

$\boldsymbol{A_5}$= maximum of the absolute deviation between test results and the mean value.

$\boldsymbol{R_5}$ = maximum absolute error of estimation.

The "coefficient of optimization" W_2 to W_5 will normally have value between (0) and (+1.0) and can be interpreted in the similar way as the coefficient of correlation r^2. Having the exceptional case that $r^2 = 1.0$ all the other "coefficient of optimization" will reach the same value and every optimization will lead to the same identical estimation formula. However, usually r^2 and W_2 to W_5 will be less than (1.0). Therefore, one has to expect different parameters in each formula according to the different criteria of optimization. For further interpretation of the results, it is also possible to calculate the respective confidence intervals.

3. STEPS OF CALCULATION

A possible basis for the suggested additional optimization is the results of the ordinary regression analysis. By the coefficient of regression a_0 to a_k (eq. 1) and the coefficient of correlation r^2 (eq. 2) are derived. Besides that, the respective average and maximum relative error of estimation (eq. 3 and 5) as well as the average and maximum absolute errors of estimation (eq. 7 and 8) are calculated.

Depending on the aim of the analysis these above-mentioned errors of estimation will be optimized iteratively in additional calculations using a comprehensive computer program. Each step of calculation will lead to respective coefficients a_0 to a_k of the preelected estimation formula and to the "coefficient of optimization" W_2 to W_5.

4. EXAMPLE

The basis of the authors' investigations is the comprehensive "Data-Bank" of the Institute of National Research, Cairo, Egypt, in which up to now the laboratory test results of more than 7000 soil samples are collected and accurately processed. This "Data-Bank" consists mainly of metrically scaled test results. To guarantee a faultless data procession some FORTAN computer programs with overlapping control mechanisms and plausibility tests were developed.

To test this extended way of analysis and for demonstrating the advantages of this method, the authors carried out some investigations on various soil parameters gathered in the "Data-Bank". At the same time, it was intended to develop formulas in which easy practicable test results are used to estimate soil properties which normally need expensive and time consuming test procedures.

Because of the lack of space it is only possible to give one representative example. For this purpose, the authors choose the results of the investigation on the undrained shear strength C_u.

This analysis was carried out on approximately 200 test results ranging from $C_u = 40\ t/m^2$ to $1000\ t/m^2$ (mean value $C_u = 200\ t/m^2$). Some tests with the regression analysis in beforehand lead to the following nonlinear multiple regression formula:

$$C_u = a_0 + a_1 \exp^{d_{0.02}} + a_2\, t^{1/3} + a_3 \exp^{d_{0.06}} + a_4\, d_{0.2} + a_5 \exp^{d_{0.2}} \longrightarrow (9)$$

Where: $d_{0.06}$ = retained material on the 0.06 mm sieve (%).

$d_{0.02}$ = retained material on the 0.02 mm sieve (%).

$d_{0.2}$ = retained material on the 0.2 mm sieve (%).

t = depth of sampling (m).

This formula is then optimized according to (eq. 1, 3, 5, 7 and 8) and results in the respective coefficients a_0 to a_5 (table 1).

***Table* 1: Optimization coefficients a_0 to a_5 of eq. (9) to the optimum criteria in eq. (1) to (8)**

Coefficients a_0 to a_5	Regression analysis eq(1)	Average relative error of estimation min. eq(3)	Maximum relative error of estimation min. eq(5)	Average absolute error of estimation min. eq(7)	Maximum absolute error of estimation min. eq(8)
a_0	500.81	416.29	425.68	528.50	575.93
a_1	1.33×10^{-12}	1.41×10^{-12}	1.14×10^{-12}	1.34×10^{-12}	1.14×10^{-12}
a_2	- 951.56	- 788.10	- 808.61	- 1004.14	- 1094.27
a_3	2.20×10^{-4}	2.88×10^{-4}	1.85×10^{-4}	2.55×10^{-4}	1.85×10^{-4}
a_4	6.40	4.96	5.43	5.05	7.32
a_5	1.16×10^{-12}	-8.07×10^{-13}	-1.32×10^{-12}	-9.99×10^{-13}	-1.32×10^{-12}

For the evaluation of the results of the analysis, the regression coefficient r^2 and the "coefficients of optimization" W_2 to W_5 are calculated. See table (2).

***Table* 2: Coefficient of correlation r^2 and coefficients of optimization W_2 to W_5.**

Coefficients r^2 and W_2 to W_5	Regression analysis eq(1)	Average relative error of estimation min. eq(3)	Maximum relative error of estimation min. eq(5)	Average absolute error of estimation min. eq(7)	Maximum absolute error of estimation min. eq(8)
r^2	0.839 (*)	0.793	0.733	0.826	0.798
W_2	0.568	0.625 (*)	0.592	0.598	0.442
W_3	0.511	0.656	0.703 (*)	0.537	0.391
W_4	0.540	0.527	0.459	0.558 (*)	0.441
W_5	0.712	0.624	0.642	0.678	0.773 (*)

The values indicated by (*) are the optimum values of r^2 or W_2 to W_5. They can be interpreted as a percentage of how much the variation of the test results are explained by the respective estimation formula optimized according to eq. (1) to eq. (8). The fact that the other values in table (2) not far from their own optimum underlines good quality of the estimation formulas of table (1).

Table (3) shows the deviations A_1 to A_5 according to the first components of eq. 1 to 8 and the residual errors of estimation R_1 to R_5. The values indicated by (*) are the optimum values of the undeclared errors of estimation.

Table 3: Deviations A_1 to A_5 and residual errors of estimation R_1 to R_5

A_1 to A_5 R_1 to R_5	Regression analysis eq(1)	Average relative error of estimation min. eq(3)	Maximum relative error of estimation min. eq(5)	Average absolute error of estimation min. eq(7)	Maximum absolute error of estimation min. eq(8)
A_1	27652.43				
R_1	4722.44 (*)	5710.24	6282.41	4842.63	5577.19
A_2	0.745				
R_2	0.322	0.279 (*)	0.304	0.299	0.416
A_3					
R_3	1.555	1.092	0.943 (*)	1.478	1.937
A_4	102.21 t/m^2				
R_4	47.12 t/m^2	48.80 t/m^2	55.40 t/m^2	45.50 t/m^2(*)	57.20 t/m^2
A_5	810.06 t/m^2				
R_5	233.60 t/m^2	304.50 t/m^2	291.75 t/m^2	260.10 t/m^2	195.60 t/m^2(*)

CONCLUSION

The values of the variance A_1 and the residual errors of estimation R_1 are quite inevident. On the contrary, the values of A_2 to A_5 and R_2 to R_5 can easily be interpreted. For instance, the average relative deviation between y_i and $\overline{y}(A_2)$ is 0.746 that means 74.6 %. When applying the ordinary regression analysis, this value will be reduced to the average relative error of estimation R_2 of 32.4 %. Using the optimization criterion according to eq. (3) R_2 results in the optimization of 27.9 %. The other values contained in table (3) can be interpreted in similar ways. They also show that the values of R_3 to R_5 can be minimized when applying the suggested steps of optimization.

As the test results of C_u have a large range of results of the optimization criteria according to eq. (7) and (8) are less significant.

But there are results of investigation of soil parameters for instance the Plastic Limit (P.L.) which do not scatter so much. In this case, the optimum average and maximum absolute errors of estimation R_4 and R_5 have also good signification and are very helpful valuating the results of the suggested extended regression analysis.

REFERENCES

CLAUS, G. and H. EBNER (1974). "Grund Lagen der Statistik", Verlage Volk und Wissen, Berlin, Germany.

ELNIMR, A. and V. RIZKALLAH (1975). "Regression Estimation of the Coefficient of Compressibility from Grouped Observations", Proce. Of 2nd ICASP, Aachen.

EZEKIEL, M. and K. FOX (1959). "Methods of Correlation Analysis", J. Weley & Sons Inc, New York.

GANSE, R. (1977). "Statistical Interpretation of Test Results of Distributed Soil Samples.", Belgin Road Research Center, Brussels.

HOLTZ, R and J. HRODE (1979). "Statistical Evaluation of Soil Test Data, Factor Analysis.", West, Lafayette, Indiana, USA.

JACK, R. and C. CORNEL (1970). "Probability, Statistics and Decision for Civil Engineers", Mc Graw-Hill, New York.

LUMB, P. (1970). "Safety Factors and Probability Distribution of Soil Strength", Canadian Geotechnical Journal, vol.7 No.3

RIZKALLAH, V. and G. MASCHWITZ (1979). "Estimation of Becring Capacity of Large Bored Piles in Cohesive Soils using Statistical Methods", Proce. Of 3rd ICASP, Sydney.

WU, T. (1974). "Uncertainty, Safety and Decision in Soil Engineering.", Proce. Of ASCE, vol.93, No.5.