APPLICATION OF THE STATISTICAL METHODS IN STUDY OF THE DEFORMATION OF THE GROUND SURFACE

George Valev[1], Penka Kastreva[2]

SUMMARY

Frequently in the nature as well as in the social life some interdependence between two or more accidental values is arising. These interrelations in most cases through empirically drawed formulas are expressing. Such cases happen very often in geodesy and cartography. One of more important means for creation of empiric functions is the least square method. The deduced empiric functions are using: at the interpolation, at study of deformations, as well as at prognostics of various events.
Both cases for drawing dependence between two quantities on the basis of experimental data are considered: regression analysis and least- squares method. These methods are applied in investigation of deformations in the area of Mirovo Salt Deposit near to Provadya, where geodetic measurements have been doing since 1985. 33 cycles of measurements on the well cameras are provided and the subsidence of the wells for every cycle has been carried out. In this case the pair of quantities are the time T (epoch of measurements) and the elevation H of the wells. For one well detail results from regression analysis and graphic are enclosed. For all 38 wells the summary (generalized) results are enclosed.

Key words: Method of least squares, Probabilistic - statistical approach correlation coefficient, subsidence of the well,

1. GENERAL PRINCIPLES

The case with two dependent variables is mostly seen [4, 5, and 7]. The relation between the two non-random variable values x and y is functional, where each value of x corresponds to a specific value of y.

$$y = f(x) \qquad (1)$$

However between two random variables x and y, there may be so-called stochastic or correlation dependence. In this case when one variable changes this leads to changes in the average value of the other variable. In other

[1] **Prof. DSc. Eng. George Valev,** georgvalev@abv.bg,
University "Ep. Konstantin Preslavski", Shumen, Bulgaria,
[2] **Assoc.Prof.Dr. Penka Kastreva,** penkakastreva@gmail.com,
South-West University "Neofit Rilski", Blagoevgrad, Bulgaria,

words, stochastic correlation between two quantities is such a relation in which each value of x corresponds to a distribution of the values of Y.
Different relations are used as empirical functions: polynomial, power, indicative, harmonious and others.
The relation between the two variables (if there are any) can also be linear and non-linear. In the regression analysis appears such empirical linear relation. Formally, there are two methods for the construction of these dependences: probabilistic - statistical method and the method of least squares (MLS). Each method displays so-called "Equation of the best straight" which finds a very wide application. Here we will make a theoretical and practical comparison between these two methods.

2. GRAPHICALLY DETERMINATION OF EMPIRICAL RELATIONSHIPS BETWEEN TWO QUANTITIES

n all cases when an empirical formula has to be built, firstly it's necessary to determine whether the relation between the two variables is linear. This is most easily done graphically, in order to get a notion about the pattern of the empirical formula. Points have to be plotted on the scheme with their rectangular (planar) coordinates and they have to be assessed whether they lie approximately in a straight line. The points of the graph give a distribution, i.e. random deviations from the assumed or visible relation dependent on unavoidable errors in measurements in each experiment or observations. In general, the graph is curved broken line.
If the location of the positions of points has obviously linear pattern, then in the simplest case as an empirical formula could be considered the equation of a straight line. If the pattern of the graph obviously deviates significantly from the straight line, this means that the relation is not linear.
Theoretically, it is assumed that between all of the n points, with coordinates Xi and Yi can always be passed curve which is expressed analytically by a polynomial of n-1 degree, so that it can pass through each of the points. Practically, such an approach usually doesn't lead to the aim, because the random distribution of the points on the graph is the reflection of the statistical distribution of the al results from the experiment. Using MLS , however, can decrease the irregular, random deviations and is the best way to express the general pattern of the dependence of y on x or vice versa. The same effect is obtained by regression analysis.

3. LINEAR RELATIONSHIP (FORWARD OR REVERSE)

The relation between two variables is expressed by the equation of a straight line

$y = a + b\,x$ or $x = c + dy$. (2)

It is called equation of the linear regression line or regression line. The coefficient b (respectively d), which represents the angular coefficient of the line is called coefficient of regression or a regression coefficient. As it is well known, this is the tangent of the angle which the straight line concludes with the x axis.

$b = tg(\alpha)$

$d = tg(\beta)$ (3)

The coefficient a (c), which is constant, is the so-called segment.

4. THE CORRELATION COEFFICIENT

Of particular importance is to determine the strength or degree (narrowness) of the correlation, and the type of this relation, expressed by a formula that will allow us to calculate the average value of a variable with a given value of the other. As a measure and an important feature of the relation between two random variables is used the *correlation coefficient* r_{xy}, which actually expresses the strength of the correlation. It should be noted that the correlation coefficient is usually used as an indication of the relation between the two variables where this relation is linear.

The correlation coefficient has a value from -1 to +1. When the correlation coefficient is closer to 1 the relation between x and y is closer to functional one. If the correlation coefficient is 0, the relation is not linear, but a non-linear correlation or even functional relation could exist.
The reliability with which the coefficient of the correlation is determined depends on the number n of the values of x and y. When n> 50 the correlation coefficient (or the correlation) is considered to be reliably established,

$r_{xy} >= 3$ (4)

where σ_r is the standard of correlation coefficient .

5. PROBABILISTIC - STATISTICAL APPROACH

For probabilistic - statistical approach equation is presented as follows [1, 7]:

$Y = b\,X$, (5)

where X and Y are centered values, respectively,

$X = x - Xo$,

$Y = y - Yo$ ’ (6)

Xo and Yo are mean values

$$Xo = \frac{[X]}{n},$$

$$Yo = \frac{[Y]}{n} \quad (7)$$

and n is the number of pairs of values of the variables (x, y). Centered values could be regarded as coordinates in a coordinate system with initial point 0(Xo, Yo).

After that, the standards (σ_x and σ_y) and the covariance cov(x,y) are calculating:

$$\sigma_x = \sqrt{\frac{[X.X]}{n}},$$

$$\sigma_y = \sqrt{\frac{[Y.Y]}{n}}. \quad (8)$$

$$\mathrm{cov}(x, y) = \frac{[X.Y]}{n}$$

A covariance matrix is developed:

$$R = \begin{bmatrix} \sigma_x^2, \mathrm{cov}(x, y) \\ \mathrm{cov}(x, y), \sigma_y^2 \end{bmatrix} \quad (9)$$

The regression coefficient b is calculated:

$$b = \frac{[XY]}{[XX)} \quad (10)$$

The correlation coefficient **r** is calculated according to the next formula

$$r = \frac{\mathrm{cov}(x, y)}{\sigma_x . \sigma_y} = \frac{[XY]}{\sqrt{[XX].[YY]}} \quad (11)$$

The standard correlation coefficient can be calculated from the approximate formula

$$\sigma(r) = \frac{1 - r^2}{\sqrt{n}}. \quad (12)$$

The relation between both coefficients - regression b and correlation r is

$$r = b\frac{\sigma_x}{\sigma_y} \quad (13)$$

6. METHOD OF LEAST SQUARES (MLS)

Let (x, y) is a two-dimensional random variable and let us have a combination of corresponding values of the two variables x and y, which are supposed to be dependent, i.e. one can be represented as a linear function of the other. Then the equation will have the form (2):

$$y = a + bx. \qquad (14)$$

When the coefficients in the equation of the straight line are being calculated, then the aim is to find the maximum approximation that can be achieved by using the MLS, or to seek "best straight line" [2, 3, 6]. The parametric method of adjustment is usually applied. In this case, the equations of type adjustments are drawn.

$$v_i = 1a + x_i b - y_i \qquad (15)$$

The weight of each equation is usually taken for one. A normal system has to be specified

$$\begin{aligned} na + [x]a - [y] &= 0 \\ [x]a + [xx]b - [xy] &= 0 \end{aligned} \qquad (16)$$

and from the solution are derived coefficients a and b in the equation of the line.

The reverse (weighted) matrix will be:

$$Q = \begin{bmatrix} Q_{11}, Q_{12} \\ Q_{12}, Q_{22} \end{bmatrix} = \begin{bmatrix} n, [x] \\ [x], [xx] \end{bmatrix}^{-1} \qquad (17)$$

The residues or deviations v_i are calculated by using the formula (15).

As in every parametric adjustment mean square errors are calculated: for unit weight (m_e) and of parameters (m_a and m_b):

$$\begin{aligned} m_e &= \sqrt{\frac{[vv]}{n-2}}, \\ m_a &= m_e . \sqrt{Q_{11}}, \\ m_b &= m_e . \sqrt{Q_{22}} \end{aligned} \qquad (18)$$

The deviations v_i should not be more than $3m_e$ i.e. $V_{admissible} = 3m_e$.

With the estimated coefficients may be calculated so-called modeled values $y_i(\mathrm{mod}) = a + b.x_i$.

The relationship between the method of least squares and the statistical method can be seen if we apply MLS, as follows. First the variables x and y have to be centered in relation to the average variables.

$$X_0 = \frac{\sum x}{n}, \qquad Y_0 = \frac{\sum y}{n} \tag{19}$$

$$X = x - X_0 \qquad Y = y - Y_0 \tag{20}$$

Then the normal equation is only one:

$$[XX]b - [XY] = 0 \tag{21}$$

Because in the normal system (16) [X] = 0, [Y] = 0.

This equation corresponds to adjustment equations of the type

$$V = bX - Y\,. \tag{22}$$

Although the equation (21) differs from equation (16), the values of the adjustments (или correction) v and V are the same. In fact, the equation (21) is once reduced system (16).

$$[xx.1].b - [xy.1] = 0\,, \tag{23}$$

from which the regression coefficient b could be estimated.

$$b = \frac{[X.Y]}{[X.X)} \tag{24}$$

The segment **a** could be calculated as follows:

$$a = Y_0 - X_0 b \tag{25}$$

So if we substitute in equation (22) the values $X = x - Xo$ and $Y = y - Yo$ we will obtain

$$V = b\,(X - Xo) - (y - Yo) = (Yo - b.Xo) + bx - y\,. \tag{26}$$

If, in the equation above, the expression (25) is replaced, equation (26) is converted into equation (15). The correlation coefficient can be calculated by the formulas:

$$r = \frac{1}{\sqrt{1 + \left(\frac{m_b}{b}\right)^2 (n-2)}} = \frac{1}{\sqrt{1 + \frac{[vv]}{b^2} Q_{22}}} \tag{27}$$

It is easy to determine the relationship between formula (27) and formula (11), taking into account that

$$[vv] = [YY] - [X.Y]b \tag{28}$$

$$b=\frac{[XY]}{[XX]}, Q_{22}=\frac{1}{[X.X]} \tag{29}$$

$$m=\sqrt{\frac{[vv]}{n-2}}, m_b=m.\sqrt{Q_{22}}$$

Substituting these expressions in formula (27), we obtain formula (11):

$$r=\frac{1}{\sqrt{1+\frac{b^2.[XX]-2b[XY]+[YY]}{b^2.[XX]}}}=\frac{b.\sqrt{[XX]}}{\sqrt{2b^2.[XX]-2b[XY]+[YY]}}= \tag{30}$$

$$=\frac{b.\sqrt{[XX]}}{\sqrt{[YY]}}=\frac{\frac{[XY]}{[XX]}.\sqrt{[XX]}}{\sqrt{[YY]}}==\frac{[XY]}{\sqrt{[XX].[YY]}}.$$

The residual dispersion of the magnitude y with respect to x will be:

$$M(y-a-bx)^2=\sigma_y(1-r^2)=[vv]=\min \tag{31}$$

If the covariance matrix is $R=\begin{bmatrix}\sigma_x^2, \operatorname{cov}(x,y)\\ \operatorname{cov}(x,y), \sigma_y^2\end{bmatrix}=\begin{bmatrix}\sigma_x^2, \sigma_x^2.\sigma_y^2.r\\ \sigma_x^2.\sigma_y^2.r, \sigma_y^2\end{bmatrix}$ (32)

Inverse matrix Q will be

$$A=R^{-1}=\frac{1}{1-r^2}\begin{bmatrix}\frac{1}{\sigma_{1x}^2}, -\frac{r}{\sigma_x.\sigma_y}\\ -\frac{r}{\sigma_x.\sigma_y}, \frac{1}{\sigma_y^2}\end{bmatrix} \tag{33}$$

determinant of which is

$$\det(A)=\frac{1}{\sigma_x^2.\sigma_y^2.(1-r^2)}. \tag{34}$$

Correlation coefficient between values obtained for adjustment by method of least squares (MLS)

Correlation coefficient between the unknown parameters in parametric adjustment when we have two parameters a and b

$$r_{ab}=\frac{Q_{12}}{\sqrt{Q_{11}.Q_{22}}} \tag{35}$$

7. EXPERIMENTAL CALCULATIONS

Here we will give as an example the results of such a regression analysis of subsidence of wells in a region of the Mirovo salt deposit field near the town Provadia where geodesic measurements are made since 1985. To present, 35 cycles of measurements of all wells were done on each and the subsidence of each well for every cycle were estimated. In this case, the pair variables are the time (age) T_i and elevation H_i. In the following *Table 1* are listed the results for well No3. In the first column is the number of the measurement cycle, the second - the relevant measurement period in years, in the third – derived elevations in meters and the fourth - the calculated subsidence, related to the first cycle.

The line chart of subsidence of the well is shown below *(Figure 1)*. The scale of time (epochs) in months is situated on the horizontal axis and the subsidence (in mm) is on the vertical one. It can be seen that the graph is actually very close to a straight line because of it was chosen precisely as an empirical function.

The linear relation in this case could be presented in the following form:

$$Hi = A_0 + A_1.(T_i - T_0) \qquad (36)$$

As an initial epoch T_0 was adopted in 1980. In this case, A_1 is the annual velocity of subsidence, and A_0 is the elevation in the initial period. Below we show detailed results of the regression analysis of one of the wells with number 3.

Table 1

No	Epoch (years)	Elevation (м)	Subsidence (мм)
1	1983.5	24.2090	**0.0**
2	1986.5	24.1330	**-76.0**
3	1987.5	24.1100	**-99.0**
4	1988.5	24.0830	**-126.0**
5	1989.5	24.0570	**-152.0**
6	1990.5	24.0317	**-177.3**
7	1991.4	24.0163	**-192.7**
8	1991.8	24.0045	**-204.5**
9	1992.4	23.9836	**-225.4**
10	1992.8	23.9769	**-232.1**
11	1993.4	23.9601	**-248.9**
12	1993.8	23.9542	**-254.8**

13	1994.4	23.9370	**-272.0**
14	1994.8	23.9275	**-281.5**
15	1995.4	23.9021	**-306.9**
16	1995.8	23.8947	**-314.3**
17	1996.4	23.8738	**-335.2**
18	1997.7	23.8503	**-358.7**
19	1998.5	23.8174	**-391.6**
20	1999.4	23.7931	**-415.9**
21	2000.4	23.7717	**-437.3**
22	2001.4	23.7425	**-466.5**
23	2002.4	23.7194	**-489.6**
24	2003.3	23.7041	**-504.9**
25	2004.3	23.6680	**-541.0**
26	2005.3	23.6472	**-561.8**
27	2006.3	23.6122	**-596.8**
28	2007.3	23.5961	**-612.9**
29	2008.3	23.5739	**-635.1**
30	2009.3	23.5519	**-657.1**
31	2010.4	23.5228	**-686.2**
32	2010.8	23.5135	**-695.5**
33	**2011.3**	**23.5074**	**-701.6**

Table 2

No	Epoch (years)	Modeling elevation (m)	Modeling subsidence (mm)	Residual deviations V (mm)
1	1983.5	24.2126	3.6	**3.6**
2	1986.5	24.1351	-73.9	**2.1**
3	1987.5	24.1093	-99.7	**-0.7**
4	1988.5	24.0835	-125.5	**0.5**
5	1989.5	24.0577	-151.3	**0.7**
6	1990.5	24.0319	-177.1	**0.2**
7	1991.4	24.0087	-200.3	**-7.6**
8	1991.8	23.9984	-210.6	**-6.1**
9	1992.4	23.9829	-226.1	**-0.7**
10	1992.8	23.9726	-236.4	**-4.3**
11	1993.4	23.9571	-251.9	**-3.0**
12	1993.8	23.9468	-262.2	**-7.4**
13	1994.4	23.9313	-277.7	**-5.7**
14	1994.8	23.9210	-288.0	**-6.5**
15	1995.4	23.9055	-303.5	**3.4**
16	1995.8	23.8952	-313.8	**0.5**
17	1996.4	23.8797	-329.3	**5.9**
18	1997.7	23.8461	-362.9	**-4.2**

19	1998.5	23.8255	-383.5	**8.1**
20	1999.4	23.8023	-406.7	**9.2**
21	2000.4	23.7764	-432.6	**4.7**
22	2001.4	23.7506	-458.4	**8.1**
23	2002.4	23.7248	-484.2	**5.4**
24	2003.3	23.7016	-507.4	**-2.5**
25	2004.3	23.6758	-533.2	**7.8**
26	2005.3	23.6500	-559.0	**2.8**
27	2006.3	23.6242	-584.8	**12.0**
28	2007.3	23.5984	-610.6	**2.3**
29	2008.3	23.5726	-636.4	**-1.3**
30	2009.3	23.5468	-662.2	**-5.1**
31	2010.4	23.5184	-690.6	**-4.4**
32	2010.8	23.5081	-700.9	**-5.4**
33	**2011.3**	**23.4952**	**-713.8**	**-12.2**

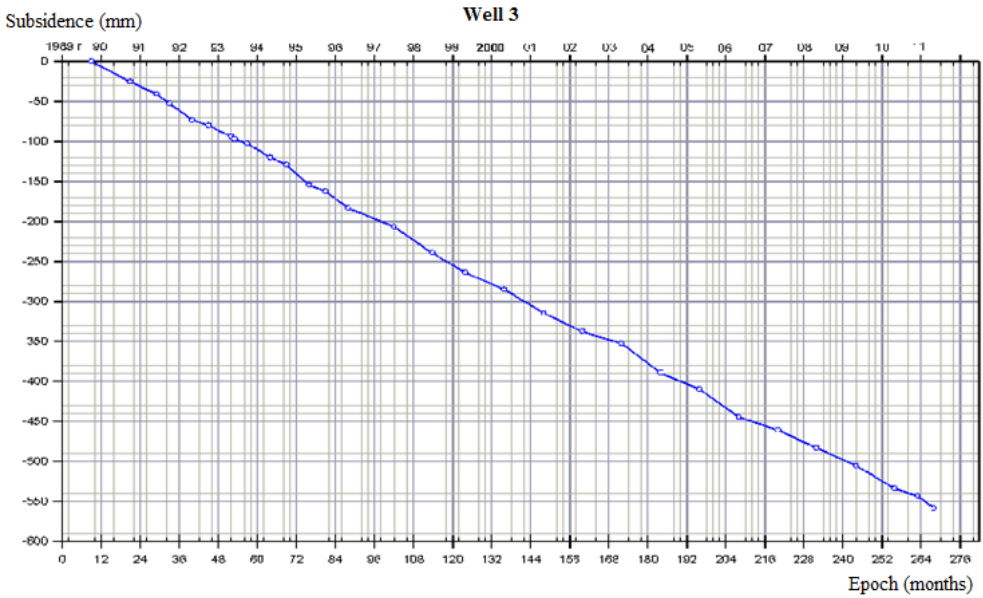

Fig.1. Graphic of the subsidence of the well 3

The normal system has the following numerical form:

$$33.00\ Ao + 594.80A_1 + 12251.10 = 0$$

$$594.80\ Ao + 12652.56\ A_1 + 270,666.30 = 0$$

The solution of the normal system gives us the following parameters:

Segment A_0 = 24.3029 m (+93.9 mm)

Regression coefficient A_1 = -25.805485 mm / yrear

The following *Table 2* shows the modelled elevation and subsidence, calculated by the empirical function (14) and the residual deviations.

The inverse matrix is

$$\begin{bmatrix} Q_{11}, Q_{12} \\ Q_{12}, Q_{22} \end{bmatrix} = \begin{bmatrix} 0.198479 & -0.009331 \\ -0.009331 & 0.000518 \end{bmatrix}$$

Mean errors: Me = 5.8 mm, MAo = 2.6 мм, MA1 = 0.000132 mm / yr

Correlation coefficient r = 0.9996.

Stochastic approach calculations give the following results:

XX] = 1931.7406060606, [YY] = 1.293470109090908,

[XY] = -49.968985454

A0 = 24.3037, A1 = -0.025867

MX,	MY,	COV
7.650986480385944	0.1979799583590362	-1.514211680440769

Correlation coefficient = 0.9996

[VV]	[V]	ME
9.0551759D-04	-4.91377771D-14	5.4046503D-03

$Q_{XX} = Q_{22}$ = 0.00051766

Covariance matrix

It can be seen that the results are the same as those of the least squares method.

$$R = \begin{bmatrix} \sigma_x^2, \mathrm{cov}(x,y) \\ \mathrm{cov}(x,y), \sigma_y^2 \end{bmatrix} = \begin{bmatrix} 7.6510^2, & -1.5142 \\ -1.5142, & 0.1980^2 \end{bmatrix}$$

Table 3

Wells No	Mean derivation Me (m)	Coefficient Ao(m) (Height)	Coefficient of regression A1(m/year) (Year velocity)	Mean error of the regression coefficient MA1	Coefficient of correlation K
3	0.0058	24.3029	-0.025805	0.000133	0.9996
4	0.0040	24.9046	-0.029889	0.000093	0.9998
5	0.0038	25.4740	-0.030435	0.000114	0.9998
6	0.0064	22.2652	-0.029402	0.000147	0.9996
7	0.0045	26.6757	-0.023944	0.000104	0.9997
8	0.0065	22.8887	-0.027985	0.000151	0.9996

9	0.0079	24.9003	-0.021828	0.000181	0.9989
10	0.0064	23.8018	-0.024298	0.000148	0.9996
11	0.0044	24.0289	-0.014399	0.000122	0.9991
12	0.0089	23.1547	-0.028610	0.000205	0.9992
13	0.0082	23.7223	-0.016619	0.000189	0.9980
14	0.0090	22.0924	-0.011611	0.000208	0.9950
15	0.0119	24.2209	-0.021046	0.000275	0.9974
16	0.0056	23.4436	-0.017269	0.000128	0.9991
17	0.0072	26.3017	-0.011812	0.000159	0.9974
18	0.0057	74.4174	-0.017160	0.000131	0.9991
19	0.0025	83.6592	-0.019733	0.000095	0.9998
20	0.0062	88.3968	-0.027191	0.000143	0.9996
21	0.0041	77.8600	-0.017643	0.000094	0.9996
23	0.0072	66.0768	-0.016031	0.000165	0.9984
24	0.0048	65.0702	-0.019495	0.000112	0.9995
25	0.0048	55.2182	-0.020973	0.000110	0.9996
26	0.0066	48.3501	-0.013749	0.000151	0.9981
27	0.0091	21.6310	-0.013941	0.000210	0.9965
28	0.0103	21.9697	-0.010690	0.000237	0.9925
29	0.0081	22.8830	-0.012315	0.000244	0.9953
30	0.0139	31.2320	-0.018464	0.000321	0.9953
31	0.0083	23.6828	-0.020771	0.000193	0.9987
32	0.0092	23.7403	-0.021945	0.000212	0.9986
33	0.0081	25.5014	-0.016236	0.000188	0.9979
35	0.0055	35.9139	-0.013219	0.000131	0.9985
37	0.0075	89.7026	-0.015686	0.000178	0.9981
38	0.0024	82.9751	-0.010502	0.000072	0.9994
42	0.0052	21.3391	-0.008748	0.000144	0.9963
43	0.0102	23.3343	-0.010168	0.000246	0.9913
45	0.0089	38.0236	-0.011504	0.000265	0.9934
46	0.0121	34.6587	-0.017526	0.000280	0.9961
50	0.0032	21.7093	-0.009395	0.000422	0.9930

8. CONCLUSION

The measuring cycles for all wells are the same - the measurements of all the wells were done simultaneously.

It is obviously that the correlation coefficient R_{xy} varies from 0.9913 to 0.9996, which means that the tread of the subsidence is most probably (almost 100%) linear function. This could be easily seen from the line chart of subsidence of the wells No.3 which is almost a straight line. The line charts of the other wells are similar to this one. The regression coefficients

A_1 or the angular coefficients of the lines are different and the biggest are those of the wells in the central part of the salt deposit.

The mean errors of both M_{A0} and M_{A1} coefficients are small. This also means that these coefficients are exactly determinate. The standard deviations M_e are also small – they are less than 10mm.This means that the subsidence of the wells could be prognosticated with such accuracy. Such prognosis is made for each of the wells for future epochs 2020, 2030, 2040 and 2050. The results of the regression analysis and respective prognostic data (which are not included here) were used for an estimation of the deformation state in the region of the deposit and for taking the relevant actions for safe exploitation of the region and its equipments.

The summary of the results from the regression analysis of the subsidence of all wells is presented in the following Table 3.

REFERENCES

1. Kovalenko I. Filippova A. Theory of probabilities and mathematical statistics. Moskow,1973.
2. Mazmishvily A. Method of least squares. "Nedra". Moskow, 1968.
3. Peevski V. Least-squares adjustment. "Technica". Sofia, 1973.
4. Russev B.,Atanasov St. Manual on Least-squares adjustment. "Technica". Sofia, 1975.
5. Tomova P., P.Bakalov, K.Kostadinov, B.Banov, V.Valchinov. Manual on Least-squares method. "Technica". Sofia, 1986.
6. Chebotarev A.S. Method of least-squares with theory of probability. Moskow, 1958.
7. Alan Julian Izenman. Modern Multivariate Statistical Techniques.Regression, Classification and Manifold. Learning. Springer Science + Business Media, LLC 2008.