

APPLICATION OF THE REGRESSION ANALYSIS IN THE WATER SUPPLY FORECAST

Marin Radkov and Anna Yordanova

Institute of Water Problems, Sofia, BULGARIA
e-mail: anisofia@bas.bg

Abstract

The necessity of a forecast in the water resources state and water supply is due to the high rates of the economy developing and its requirements to the water resources. In the last decades the regression analysis methods obtained a wide-spread in the tendency extrapolation among the formal methods, applied to the random phenomenon forecasting. An example of the power regression using for water supply is represented. The test of the statistical significant of the proposed forecasting function in three points is made.

Key words: *water supply, forecast, regression*

The necessity of a forecast in the water resources state and water supply is due to the high rates of the economy developing and its requirements to the water resources.

In the last decades the mathematic-statistical methods of extrapolation of the tendencies has a wide spread for expressing and forecasting of the random events.

The regression analysis is one of these methods and it is a logical extension of the correlation analysis. It develops and extends the concept for a correlation relation. While the correlation analysis expresses only the relation form, the regression analysis gives the mathematical equation to estimate some parameters, using other ones.

Regression equation is being obtained from the empirical data and their best approximation with a chosen theoretical function.

In the case of linear relation, it can be given on the coordinate system with a straight line.

$$(1) \quad y = a \cdot x + b,$$

where y is the value of the dependent variable; x - the factor, which influences on y ; a - the coefficient, which shows the relation degree between variables; b - the ordinate, which shows the distance from the beginning of the coordinate system.

The parameters of the equation (1) can be determined by the method of the least squares. This method determines line with two parameters so, that the sum of the square deviations of the empirical data from the line to be the least one. The fit of the empirical and theoretical curves can be tested by a Chi-square test.

When the nonlinear relation between the function and its variable arguments is established, the recommended equation to be used is:

$$(2) \quad y = a / x + b,$$

which presents hyperbola.

Since, as the water supply depends not only by one, but more independence variables, the best is the multiple regressions to be used for it forecast. The equation of that regression is represented by:

$$(3) \quad y = a_0 + a_1 x_1 + \dots + a_p x_p,$$

where x_1, x_2, \dots, x_{p-1} are known measurable variables; a_0, a_1, \dots, a_p are unknown parameters, which have to be estimated; p – the number of independence variables.

In the case of N measurements of the independence variables, the regression equation is:

$$(4) \quad y_i = a_0 + a_1 x_{1i} + \dots + a_p x_{pi}, \quad i = 1, 2, \dots, N$$

Regression coefficients a_i are estimated by relations:

$$(5) \quad a_i = b_i \frac{S_y}{S_i}, \quad i = 1, 2, \dots, p$$

where:

$$(6) \quad b_i = \sum_{k=1}^p \frac{r_{ky}}{r_{ki}};$$

r_{ky} - correlation between the k -th parameter and water supply y ;

r_{ki} - correlation between the k -th parameter and the i -th parameter;

S_y - standard deviation of the dependence variable y ;

S_i - standard deviation of the i -th parameter.

The free term is determined by the following way:

$$(7) \quad a_0 = \bar{y} - \sum_{j=1}^p a_j \cdot \bar{x}_j$$

where:

\bar{y} - mean value of the dependence variable y ;

\bar{x}_j - mean value of the i -th parameter.

The important step in the regression analysis is to prove the right choice of the parameters, i.e. to check at what extent the including parameters in the model determine the dependence variable value. For that purpose it is necessary to be made a test for a regression coefficients signification.

Statistical significant of the particular regression coefficients is estimated by t -test of Student. I.e. it is necessary to check the trough of the hypothesis:

$$(8) \quad H_0: a_i = 0, \quad i = 1, 2, \dots, p$$

For that purpose one calculates the criteria:

$$(9) \quad t_i = \frac{a_i}{S_{a_i}}, \quad \text{where } S_{a_i} \text{ is the standard deviation of the regression coefficient } a_i.$$

The last one is calculated by the formula:

$$(10) \quad S_{a_i} = \sqrt{\frac{1}{r_{iy} \cdot D_{yy}} \cdot \frac{RSS}{(N - p - 1)}},$$

where:

$$(11) \quad D_{yy} = \sum_{i=1}^N (y_i - \bar{y})^2 \text{ is the common deviation, and}$$

$$(12) \quad RSS = \sum_{i=1}^N (y_i - \bar{y})^2 \text{ is the sum of residuals.}$$

If $t_i < t_\alpha$, where α is a degree of significance (usually 0,95), than the hypothesis is accepted. I.e. the corresponding regression coefficient has to be excluded from the regression equation.

For illustration of the formulated relations, the example is given. It is on the basis of indicators for the fresh water supply – the production in nature, the production in cost and the part of the circulation water in the common fresh water supply for the 20-years period (table 1).

Table1. The Dynamic Of The Production And Water Supply
In One Factory for 20-Years Period

years	Water supply, 10^6 m^3	Production		Part of circulation water, %
		nature, 10^6 tons	cost, 10^6 lv.	
1.	60,47	1,42	48,18	77,50
2.	62,99	1,49	43,21	77,60
3.	63,22	1,52	44,08	77,63
4.	67,27	1,61	46,69	77,84
5.	69,25	1,69	49,01	77,96
6.	71,10	1,72	55,04	78,15
7.	67,16	1,63	52,16	77,66
8.	77,01	1,70	54,40	78,02
9.	73,10	1,78	56,96	80,31
10.	76,70	1,88	60,16	82,82
11.	78,18	1,94	62,08	83,61
12.	78,12	1,96	62,72	83,154
13.	84,62	2,12	72,08	87,44
14.	87,80	2,21	75,14	88,75
15.	85,34	2,14	72,76	89,60
16.	83,20	2,15	74,34	90,19
17.	82,04	2,16	75,16	92,32
18.	80,60	2,15	75,04	92,61
19.	78,80	2,16	76,08	92,82
20.	75,40	2,18	76,42	93,05

The data are worked up by a computer program STATAN, made in the institute, using the represented method. The obtained results are given in a table 2.

Table2. The Determined Regression Equation

Variant	Independent parameters	<i>t</i> -test (of Student)	$r_{k y}$	Regression Equation
1	x_1 – nature	12,45	0,95	$y = 21,2 + 28,5 x_1$
2	x_2 – cost	9,01	0,91	$y = 37,6 + 0,60 x_2$
3	x_3 – circulation water	6,06	0,82	$y = -13,8 + 1,07 x_3$
4	x_1 – nature x_2 – cost	4,18 -1,66	0,95	$y = 12,3 + 46,7 x_1 - 0,41 x_2$
5	x_2 – cost x_3 – circulation water	4,72 -2,16	0,93	$y = 89,3 + 1,08 x_2 - 0,97 x_3$
6	x_1 – nature x_3 – circulation water	9,34 -3,99	0,97	$y = 59,1 + 47,7 x_1 - 0,88 x_3$
7	x_1 – nature x_2 – cost x_3 – circulation water	5,15 0,26 -3,28	0,97	$y = 62,3 + 45,9 x_1 + 0,06 x_2 - 0,93 x_3$

From the results one can see, that the coefficient $r_{k y}$ is significant for all variants. Therefore the seven variant equations are satisfactory to be applied in the water supply forecasting. The coefficient $r_{k y}$ increases, when number of parameters increases, which is regularly.

The comparison of variant 6 to a variant 7 shows that the including of the independent parameter x_2 , production cost, doesn't influent significant on the final results and it can be ignored.

The 6-th variant is outlined as a best one. Its expression is:

$$(13) \quad (\text{water supply}) = 59,1 + 47,7 \times (\text{nature}) - 0,88 \times (\text{part of circulation water})$$

The graphical area, which the equation (13) represents, visually is shown of figure 1. From it one can determines the water supply for a given nature and a part of the circulation water.

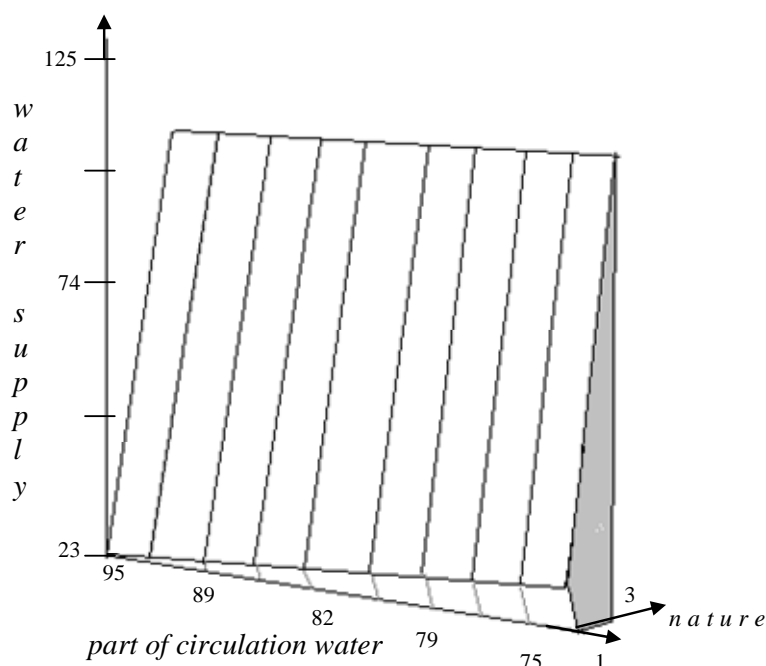


Figure1. An area of the water supply

It is made a test for the equation (13) accuracy as a determination of water supply forecast for 1-st, 9-th and 18-th years from the data table 1. For the first year the real water supply is $60,47 \times 10^6 \text{ m}^3$ and the theoretical, read from the equation, is $50,50 \times 10^6 \text{ m}^3$, i.e. the deviation is 3,3%. For 9-th year the real water supply is $73,10 \times 10^6 \text{ m}^3$ and the theoretical is $73,20 \times 10^6 \text{ m}^3$, i.e. the deviation is 1%. For 18-th year the real water supply is $80,60 \times 10^6 \text{ m}^3$ and the theoretical is $80,00 \times 10^6 \text{ m}^3$, i.e. the deviation is 1%.

These results give a reason to consider the regression analysis as efficient tool in water supply investigations.

References

- Draper N. and G. Smith, 1973, Applied Regression Analysis, *Statistics*, Moscow.
 Radkov M. ,1990, Industrial Water Supply – Forecasting and Losses in Water Feed Restraint, Sofia.