6 FREQUENCY AND REGRESSION ANALYSIS OF HYDROLOGIC DATA

R.J. Oosterbaan

PART II: REGRESSION ANALYSIS

On web site www.waterlog.info

For free sofware on segmented regression analysis see: *www.waterlog.info/segreg.htm*

Chapter 6 in: H.P.Ritzema (Ed.), Drainage Principles and Applications, Publication 16, second revised edition, 1994, International Institute for Land Reclamation and Improvement (ILRI), Wageningen, The Netherlands. ISBN 90 70754 3 39

Table of contents

6 FREQUENCY AND REGRESSION ANALYSIS 3
6.1 Introduction 3
6.5 Regression Analysis 4
6.5.1 Introduction 4
6.5.2 The Ratio Method 5
6.5.3 Regression of y upon x 9
Confidence statements, regression of y upon x 11
Example 6.2 Regression y upon x 14
6.5.4 Linear Two-way Regression 16
Confidence belt of the intermediate regression line 18
6.5.5 Segmented Linear Regression 20
6.7 References 24

6 FREQUENCY AND REGRESSION ANALYSIS

6.1 Introduction

Frequency analysis, regression analysis, and screening of time series are the most common statistical methods of analysing hydrologic data.

Frequency analysis is used to predict how often certain values of a variable phenomenon may occur and to assess the reliability of the prediction. It is a tool for determining design rainfalls and design discharges for drainage works and drainage structures, especially in relation to their required hydraulic capacity.

Regression analysis is used to detect a relation between the values of two or more variables, of which at least one is subject to random variation, and to test whether such a relation, either assumed or calculated, is statistically significant. It is a tool for detecting relations between hydrologic parameters in different places, between the parameters of a hydrologic model, between hydraulic parameters and soil parameters, between crop growth and water table depth, and so on.

Screening of time series is used to check the consistency of time-dependent data, i.e. data that have been collected over a period of time. This precaution is necessary to avoid making incorrect hydrologic predictions (e.g. about the amount of annual rainfall or the rate of peak runoff).

6.5 Regression Analysis

6.5.1 Introduction

Regression analysis was developed to detect the presence of a mathematical relation between two or more variables subject to random variation, and to test if such a relation, whether assumed or calculated, is statistically significant. If one of these variables stands in causal relation to another, that variable is called the independent variable. The variable that is affected is called the dependent variable.

Often we do not know if a variable is affected directly by another, or if both variables are influenced by common causative factors that are unknown or that were not observed. We shall consider here relations with only one dependent and one independent variable. For this, we shall use a two-variable regression. For relations with several independent variables, a multivariate regression is used.

Linear two-variable regressions are made according to one of two methods. These are:

- The ratio method (Section 6.5.2);
- The "least squares" method (Section 6.5.3).

The ratio method is often used when the random variation increases or decreases with the values of the variables. If this is not the case, the least-squares method is used. The ratio method, as we use it here, consists of two steps, namely:

- Calculate the ratio $p = y/x$ of the two variables y and x;
- Calculate the average ratio p_{av} , its standard deviation $s_{p_{av}}$, and its upper and lower confidence limits p_u and $_v$, to obtain the expected range of in repeated samples.

The least squares method consists of finding a mathematical expression for the relation between two variables x and y, so that the sum of the squared deviations from the mathematical relation is minimized. This method can be used for three types of regressions:

- Regressions of y upon x;
- Regressions of x upon y;
- Two-way regressions.

Regressions of y upon x are made if y is causally influenced by x, or to predict the value of y from a given value of x. In these regressions, the sum of the squared deviations of y to the regression line, i.e. in the y-direction, are minimized.

Regressions of x upon y are made to predict the value of x from a given value of y. Except for the reversal of the variables, the procedure for making these regressions is identical to that for making regressions of y upon x. Hence here it is the sum of the squared deviations of x that are minimized.

Two-way regressions are made if no dependent variable can be selected. These are intermediate regressions that cover the territory between regressions of y upon x and of x upon y.

The relation between y and x need not be linear. It can be curved. To detect a non-linear relation, it is common practice to transform the values of y and x. If there is a linear relation between the transformed values, a back-transformation will then yield the desired non-linear relation.

The majority of these transformations is made by taking log-values of y and x, but other transformations are possible (e.g. square root functions, goniometric functions, polynomial functions, and so on). Curve fitting can be done conveniently nowadays with computer software packages.

Further discussion of transformations and non-linear regressions is limited to Example 6.3 of Section 6.5.4 and Example 6.4 of Section 6.5.5. For more details, refer to statistical handbooks (e.g. Snedecor and Cochran 1986).

6.5.2 The Ratio Method

If the variation in the data (x, y) tends to increase linearly, the ratio method can be applied. This reads

$$y = p.x + \varepsilon \quad \text{or} \quad \hat{y} = p.x$$

or

$$y/x = p + \varepsilon' \quad \text{or} \quad (\hat{y}/x) = p$$

where

p	= a constant (the ratio)
$\hat{y}$	= the expected value of y according to the ratio method
ε and ε'	= a random deviation
$(\hat{y}/x)$	= the expected value of the ratio y/x

Figure 6.17 suggests that there is a linear relation between y and x, with a linearly increasing variation. The envelopes show that the ratio method is applicable. In situations like this, it is best

to transform the pairs of data (y, x) into ratios p = y/x. The average ratio for n pairs is then calculated as

$$p_{av} = \Sigma p/n \tag{6.26}$$

Using Equation 6.13, we find the standard deviation of p from

$$s_p^2 = \Sigma(p-p_{av})/(n-1) = (\Sigma p^2 - \Sigma p/n)/(n-1) \tag{6.27}$$

and using Equation 6.19, we find the standard deviation of p_{av} from

$$s_{p_{av}} = s_p/\sqrt{n} \tag{6.28}$$

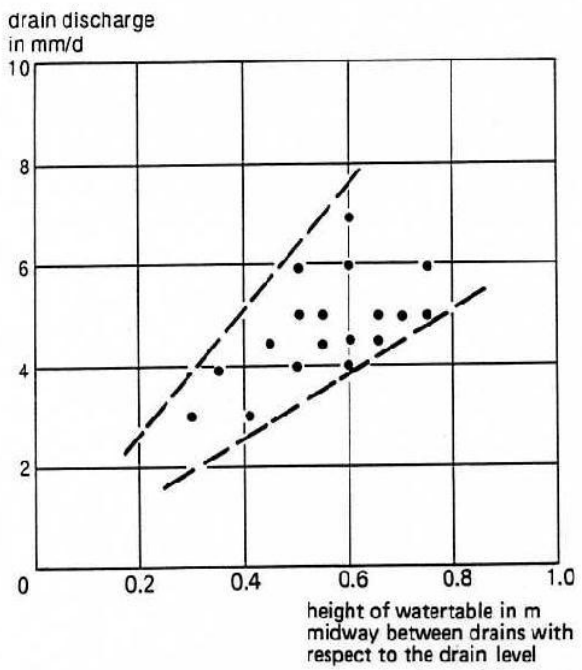

Figure 6.17 The ratio method. The variation of y increases with increasing x

The confidence interval of p_{av} , i.e. the expected range of p_{av} in repeated samples, is approximated by

$$p_u = p_{av} + t.s_{p_{av}} \tag{6.29}$$

$$p_v = p_{av} - t.s_{p_{av}} \tag{6.30}$$

Here, the subscripts u and v denote the upper and lower confidence limits. The letter t stands for the variate of Student's distribution (Table 6.9) at the frequency of exceedance f . If one wishes an interval with c% confidence, then one should take f = 0.5(100 - c)/100 (e.g. f = 0.05 when c = 90%). The value of t depends on the number (n) of observations. For large values of n, Student's distribution approaches the standard normal distribution. For any value of n, the t-distribution is symmetrical about t = 0.

Table 6.9 Values t of Student's distribution with d degrees of freedom and frequency of exceedance F_u

	Fu			
d	0.10	0.05	0.025	0.01
5	1.48	2.02	2.57	3.37
6	1.44	1.94	2.45	3.14
7	1.42	1.90	2.37	3.00
8	1.40	1.90	2.37	3.00
9	1.38	1.83	2.26	2.82
10	1.37	1.81	2.23	2.76
12	1.36	1.78	2.18	2.68
14	1.35	1.76	2.15	2.62
16	1.34	1.75	2.12	2.58
20	1.33	1.73	2.09	2.53
25	1.32	1.71	2.06	2.49
30	1.31	1.70	2.04	2.46
40	1.30	1.68	2.20	2.42
60	1.30	1.67	2.00	2.39
100	1.29	1.66	1.99	2.37
200	1.28	1.65	1.97	2.35
∞	1.28	1.65	1.96	2.33

If the confidence interval p_u - p_v contains a zero value, then p_{av} will not differ significantly from zero at the chosen confidence level c. Although the value of p_{av} is then called insignificant, this does not always mean that it is zero, or unimportant, but only that it cannot be distinguished from zero owing to a large scatter or to an insufficient number of data.

Example 6.1

A series of measurements of drain discharge and water table depth are available on an experimental area. The relation between these two variables is supposedly linear, and the variation of the data increases approximately linearly with the x and y values. We shall use the ratio method to find the relation. The data are tabulated in Table 6.10.

Table 6.10 Data used in Figure 6.17, where y = drain discharge (mm/d) and
x = height of the water table (m) midway between the drains,
ith respect to drain level

nr.	y	x	p=y/x	nr.	y	x	p=y/x
1	3.0	0.30	10.0	10	7.0	0.60	11.7
2	4.0	0.35	11.4	11	6.0	0.60	10.0
3	3.0	0.40	7.5	12	4.5	0.60	7.5
4	4.5	0.45	10.0	13	4.0	0.60	6.7
5	6.0	0.50	12.0	14	5.0	0.65	7.7
6	5.0	0.50	10.0	15	4.5	0.65	6.9
7	4.0	0.50	8.0	16	5.0	0.70	7.1
8	5.0	0.55	9.1	17	6.0	0.75	8.0
9	4.5	0.55	8.2	18	5.0	0.75	6.7

Ratio method : $p = y/x$, $\Sigma p = 158.5$, $\Sigma p^2 = 1448$, $n = 18$

Equation 6.26 : $p_{av} = 158.5/18 = 8.8$ (average p)

Equation 6.27 : $s_p = \sqrt{(1448-18 \times 8.8^2)/17} = 1.78$

Equation 6.28 : $s_{pav} = 1.78/\sqrt{18} = 0.42$

Table 6.9 : $F_u=0.05$, $d=17 \rightarrow t_{90\%} = 1.75$

Equation 6.29 : $p_u = 8.8 + 1.75 \times 0.42 = 9.5$

Equation 6.30 : $p_v = 8.8 - 1.75 \times 0.42 = 8.1$

The data of Table 6.10 show that parameter p is estimated as 8.8, the 90% confidence limits being $p_u = 9.5$ and $p_v = 8.1$. Hence the ratio p is significantly different from zero. In Chapter 12, the ratio is used to determine the hydraulic conductivity.

Figure 6.18 illustrates situations where y is not zero when x = 0. When this occurs, the ratio method can be used if $y - y_o$ is substituted for y, and if $x - x_o$ is substituted for x. In these cases, x_o and y_o should be determined first, either by eye or by mathematical optimisation.

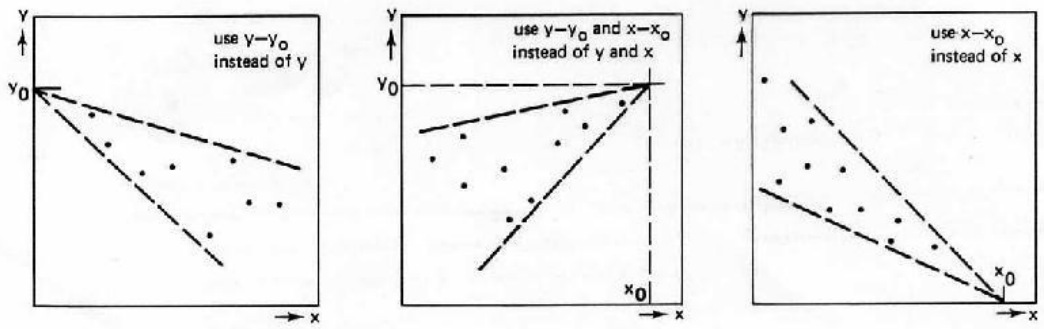

Figure 6.18 Adjustments of the ratio method when y and x are not zero

6.5.3 Regression of y upon x

The linear regression of y upon x is designed to detect a relation like the following

$$y = ax + b + \varepsilon, \text{ or } \hat{y} = ax + b \tag{6.31}$$

where

a =	the linear regression coefficient, giving the slope of the regression line
b =	the regression constant, giving the intercept of the regression line on the y axis
ε =	a random deviation of the y value from the regression line
$\hat{y}$ =	the expected value of y according to the regression ($\hat{y} = y - \varepsilon$).

This regression is used when the ε values are independent of the values of y and x. It is used to predict the value of y from a value of x, regardless of whether they have a causal relation.

Figure 6.19 illustrates a linear regression line that corresponds to 8 numbered points on a graph. A regression line always passes through the central point of the data (χ, $\breve{y}$, sometimes indicated by $\bar{x}$ and $\bar{y}$). A straight line through the point (χ, $\breve{y}$) can be represented by

$$(y - \breve{y}) = a(x - \chi) \tag{6.32}$$

where a is the tangent of the angle α in the figure, $\breve{y}$ and χ are the arithmetic mean values of x and y respectively (sometimes indicated by $\bar{y}$ and $\bar{x}$).

Normally, the data (x, y) do not coincide with the line, so a correct representation of the regression is

$$(y - \breve{y}) = a(x - \chi) + \varepsilon \tag{6.33}$$

where ε is a vertical distance of the point (x, y) to the regression line. The sum of all the ε values equals zero. The difference y-ε gives a y value on the regression line, $\hat{y}$. Substitution of $\hat{y} = y - \varepsilon$ in Equation 6.32 gives

$$(\hat{y} - \breve{y}) = a(x - \chi) \tag{6.34}$$

where a is called the regression coefficient of y upon x.

Equation 6.34 can also be written as

$$\hat{y} = a\,x + \breve{y} - a\chi \tag{6.35}$$

By substituting $b = \breve{y} - a\chi$ we get Equation 6.31.

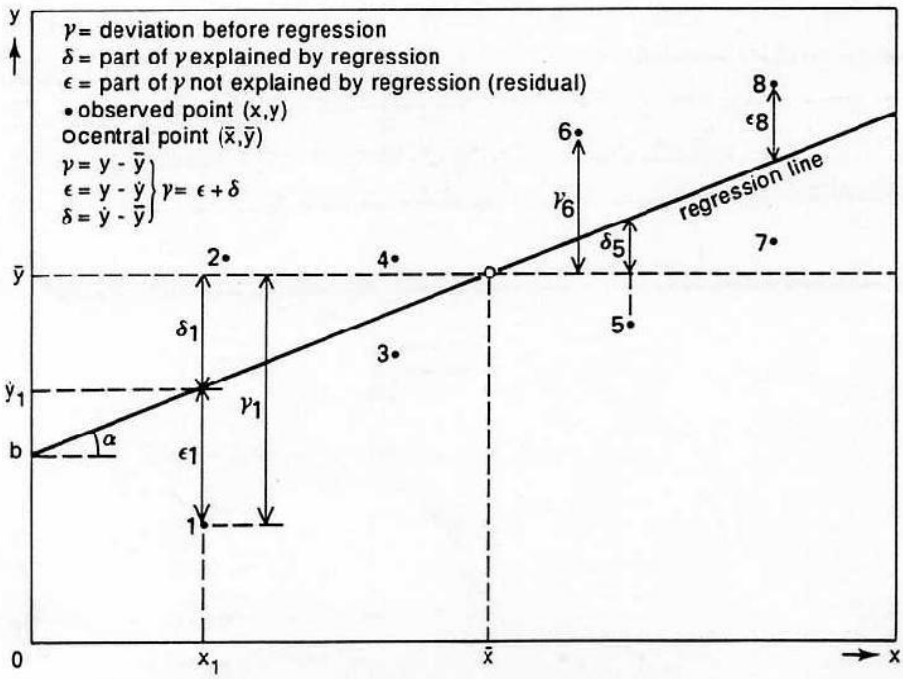

Figure 6.19 Variations in the y-direction

To determine the best possible regression coefficient, one must minimize the $\Sigma\varepsilon^2$ (the least squares method). In other words the choice of the slope of the line and the intercept must fit the points as well as possible. To meet this condition we must take

$$a = \Sigma' yx \,/\, \Sigma' x^2 \tag{6.36}$$

where

$$\Sigma' yx = \Sigma(y-\breve{y})(x-\chi) \tag{6.37}$$
$$\Sigma' x^2 = \Sigma(x-\chi)^2 \tag{6.38}$$
$$\Sigma' y^2 = \Sigma(y-\breve{y})^2 \tag{6.39}$$

in which the symbol Σ' means "reduced sum". Equation 6.39 was included for use in the ensuing confidence statements.

The coefficient a can be directly calculated from the (x, y) pairs of data. If a is positive, the regression line slopes upward, and an increase in x causes an increase in y, and vice versa. If a is negative, the regression line slopes downward. If the regression coefficient a is zero, then there is no linear relation between y and x, and the line is horizontal.
The following equations give additional definitions (see Equation 6.13 also)

$$s_x^2 = \Sigma' x^2/(n-1) = \Sigma(x-\chi)^2/(n-1) = \{\Sigma x^2 - (\Sigma x)^2 /n\}/(n-1) \tag{6.40}$$

where s_x^2 is called the variance of x

$$s_y^2 = \Sigma' y^2/(n-1) = \Sigma(y-\breve{y})^2/(n-1) = \{\Sigma x^2 - (\Sigma x)^2/n\}/(n-1) \tag{6.41}$$

where s_y^2 is called the variance of y

$$s_{xy}^2 = \Sigma' xy/(n-1) = \Sigma(y-\breve{y})(x-\chi)/(n-1) = \{\Sigma xy - (\Sigma x \Sigma y)/n\}/(n-1) \tag{6.42}$$

where s_{xy}^2 is called the covariance of x and y.
Therefore, we can also write for Equation 6.36

$$a = s_{xy} / s_x^2 \tag{6.43}$$

Confidence statements, regression of y upon x

The sum of the squares of the deviations ($\Sigma\varepsilon^2$) is minimum, but it can still be fairly large, indicating that the regression is not very successful. In an unsuccessful regression, the regression coefficient a is zero, meaning that variations of x do not explain the variation in y, and $\Sigma\varepsilon^2 = \Sigma(y-\breve{y})^2 = \Sigma' y^2$ (compare with Equation 6.39). But if the coefficient a is different from zero, part of the y-variation is explained by regression, and the residual variation drops below the original variation: $\Sigma\varepsilon^2 < \Sigma' y^2$. In other words, the residual deviations with regression are smaller than the deviations without regression. The smaller the non-explained variation $\Sigma\varepsilon^2$ becomes, the more successful the regression is. The ratio $\Sigma\varepsilon^2/\Sigma' y^2$ equals $1-R^2$, in which R^2 is the coefficient of determination, which is a measure of the success of the regression.

In linear regression, the coefficient R equals the absolute value of the correlation coefficient r. In addition, $r^2\Sigma' y^2$ equals the linearly explained variation and $(1-r^2)\Sigma' y^2$ is the residual variation, $\Sigma\varepsilon^2$. The value of r can be calculated from

$$r = \frac{\Sigma' xy}{\Sigma' x^2\, \Sigma' y^2} = \frac{s_{xy}}{s_x\, s_y} \tag{6.44}$$

This correlation coefficient is an indicator of the tendency of the y variable to increase (or decrease) with an increase in the x variable. The magnitude of the increase is given by the coefficient a. Both are related below as

$$r = a\, s_x / s_y \tag{6.45}$$

The correlation coefficient r can assume values of between -1 and +1. If $r > 0$, the a coefficient a is also positive. If $r = 1$ there is a perfect match of the regression line with the (x, y) data. If $r < 0$, the coefficient a is also negative, and if $r = -1$, there is also a perfect match, although y increases as x decreases and vice versa. If $r = 0$, the coefficient a is also zero, the regression line is parallel to the x-axis, i.e. horizontal, and the y variable has no linear relation with x.

In non-linear relations, the r coefficient is not a useful instrument for judging a relation. The coefficient of determination (or explanation) $R^2 = 1-\Sigma\varepsilon^2/\Sigma'y^2$ is then much better (Figure 6.20).

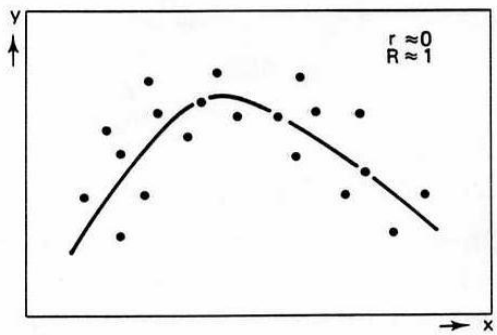

Figure 6.20 A clear relation between y and x, although $r \approx 0$

Because the coefficient a was determined with data of a certain random variation, it is unlikely that all of its values will be the same if it is determined again with new sets of data. This means that the coefficient a is subject to variation and that a confidence interval will have to be determined for it. For this purpose, one can say that it is c% probable that the value of a in repeated experiments will be expected in the range delimited by

$$a_u = a + ts_a \tag{6.46}$$

$$a_v = a - ts_a \tag{6.47}$$

with

$$s_a^2 = \frac{\Sigma\varepsilon^2}{(n-2)\,\Sigma'x^2} = \frac{(1-r^2)\,\Sigma'y^2}{(n-2)\,\Sigma'x^2} \tag{6.48}$$

where

a_u and a_v are the upper and lower confidence limits of a

t = a variable following Student's distribution, with $d = n - 2$ degrees of freedom (Table 6.9) and $f = 0.5(100-c)/100$ is the frequency with which the t value is exceeded (the uncertainty)

s_a = the standard deviation of the coefficient a

Theoretically, this statement is valid only if the ε deviations are normally distributed and independent of x. But for most practical purposes, the confidence interval thus determined gives a fair idea of the possible variation of the regression coefficient.

One can also say that, in repeated experiments, there is a c% probability that the $\hat{y}$ value found by regression ($\hat{y}$, Equation 6.34) for a given x value, will be in the range limited by

$$\hat{y}_u = \hat{y} + t\, s_{\hat{y}} \tag{6.49}$$

$$\hat{y}_v = \hat{y} - t\, s_{\hat{y}} \tag{6.50}$$

where $\hat{y}_u$ and $\hat{y}_v$ are the upper and lower confidence limits of $\hat{y}$ and $s_{\hat{y}}$ is the standard deviation of $\hat{y}$ equal to

$$s_{\hat{y}} = \sqrt{\{s_Y^2 + (x-\chi)^2 s_a^2\}} \tag{6.51}$$

Here, s_Y is the standard deviation of Y, which is the average value of $\varepsilon = \hat{y} - y$:

$$s_Y = \sqrt{\{\Sigma\varepsilon^2 / (n-2)n\}} \tag{6.52}$$

By varying the x value, one obtains a series of and $\hat{y}_u$ and $\hat{y}_v$ values, from which the confidence belt of the regression line can be constructed. Taking x = 0, one obtains the confidence limits of the regression constant b. In this case, the value of s_Y^2 is often relatively small, and so Equation 6.51 can be simplified to

$$s_b = \chi\ s_a \tag{6.53}$$

and the upper and lower confidence limits of b are

$$b_u = b + t\, s_b = b + t\, \chi\, s_a \tag{6.54}$$

$$b_v = b - t\, s_b = b - t\, \chi\, s_a \tag{6.55}$$

Note that the above confidence intervals are valid for the regression line. The intervals of individual values are wider. To calculate the confidence interval of the y value calculated from a certain x value one may use, in similarity to Equations 6.49, 6.50 and 6.51, $y_u = \hat{y} + t\, s_Z$ and $y_v = \hat{y} - t\, s_Z$ where $s_Z = \sqrt{\{s_Y^2 + (x-\chi)^2 s_a^2\}}$.

With a pocket calculator, it is relatively simple to compute a linear 2-variable regression analysis and the corresponding confidence statements because all the calculations can be done knowing

only n, Σx, Σy, $\Sigma(xy)$, Σx^2, Σy^2. This is illustrated in the following example. Nowadays, personal computers are making regressions even easier, and general software packages like spreadsheets can be conveniently used.

Example 6.2 Regression y upon x

The data from Table 6.11 were used to do a linear regression of y upon x to determine the dependence of crop yield (y) on water table depth (x): y = ax + b. The result is shown in Figure 6.21.

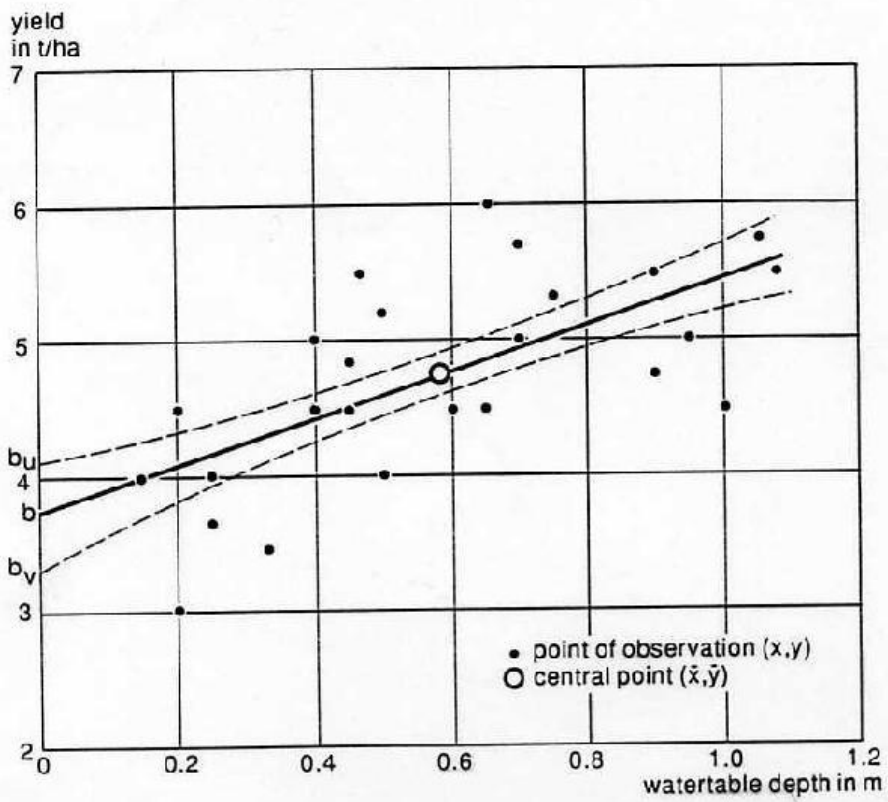

Figure 6.21 Linear regression of y upon x (Example 6.2)

Table 6.11 (y, x) data used in Figure 6.21, with y = crop yield (t/ha) and x = seasonal average depth f the water table

nr	y	x	nr	y	x
1	4.0	0.15	14	4.0	0.50
2	4.5	0.20	15	4.5	0.60
3	3.0	0.20	16	6.0	0.65
4	4.0	0.25	17	4.5	0.65
5	3.7	0.25	18	5.7	0.70
6	3.5	0.32	19	5.0	0.70
7	5.0	0.40	20	5.3	0.75
8	4.5	0.40	21	5.5	0.90
9	4.5	0.40	22	4.7	0.90
10	4.8	0.45	23	5.0	0.91
11	4.5	0.45	24	4.5	1.00
12	5.5	0.47	25	5.7	1.05
13	5.2	0.50	26	5.5	1.08

$\Sigma x = 14.87$ $\Sigma x^2 = 10.47$ $\Sigma xy = 73.46$

$\Sigma y = 122.60$ $\Sigma y^2 = 591.68$ $n = 26$ $n-2 = 24$

$\chi = \Sigma x/n = 14.87/26 = 0.57$ (average x)

$\breve{y} = \Sigma y/n = 122.60/26 = 4.7$ (average y)

Equation 6.38 : $\Sigma' x^2 = 10.47 - 14.87^2/26 = 1.97$

Equation 6.39 : $\Sigma' y^2 = 591.68 - 122.60^2/26 = 1357$

Equation 6.37 : $\Sigma' xy = 73.46 - 14.87 \times 122.60/26 = 3.34$

Equation 6.36 : $a = 3.34/1.97 = 1.70$

Equation 6.35 : $b = 4.7 - 1.70 \times 0.57 = 3.73$

Equation 6.44 : $r = 3.34/\sqrt{1.97 \times 13.57} = 0.65 \rightarrow r^2 = 0.42$

Equation 6.48 : $\Sigma\varepsilon^2 = (1-0.42)\ 13.57 = 7.87$

Equation 6.48 : $s_a = \sqrt{\{787/(24 \times 1.97)\}} = 0.42$

Table 6.9 : $F_u = 0.005$, $d = 24 \rightarrow t_{90\%} = 1.71$

Equation 6.46 : $a_u = 1.70 + 1.71 \times 0.41 = 2.4$

Equation 6.47 : $a_v = 1.70 + 1.71 \times 0.23 = 1.0$

Equation 6.53 : $s_b = 0.57 \times 0.41 = 0.23$

Equation 6.54 : $b_u = 3.73 + 1.71 \times 0.23 = 4.1$

Equation 6.55 : $b_v = 3.73 - 1.71 \times 0.23 = 3.3$

$x = \chi \rightarrow$

Equation 6.51 : $s_{\hat{y}} = s_Y = \sqrt{\{7.87/(24 \times 26)\}} = 0.11$

Equation 6.49 : $\hat{y}_u = 4.7 + 1.71 \times 0.11 = 4.9$

Equation 6.50 : $\hat{y}_u = 4.7 - 1.71 \times 0.11 = 4.5$

From the table, we see that the confidence limits of the regression coefficient (a = 1.70) are a_u = 2.4 and a_v = 1.0. Hence, although the coefficient is significant, its range is wide. Because r^2 = 0.42, we know that the regression explains 42% of the squared variations in y. As the regression equation (Equation 6.41), we get

$$(\hat{y} - 4.7) = a(x - 0.57)$$

With the calculated b, the regression result can also be written as

$$\hat{y} = a\,x + 3.73 \quad (n = 18,\ r = 0.65)$$

According to this, every 0.10 m that the water table rises results in an average crop yield increase of 0.17 t/ha (using a = 1.70), with a maximum of 0.24 t/ha (using a_u = 2.4) and a minimum of 0.10 t/ha (using a_v = 1.0).

6.5.4 Linear Two-way Regression

Linear two-way regression is based on a simultaneous regression of y upon x and of x upon y. It is used to estimate the parameters (regression coefficient a and intercept b) of linear relations between x and y, which do not have a causal relation.

Regression of y upon x yields a regression coefficient a. If the regression of x upon y yields a regression coefficient a', we get, analogous to Equation 6.34

$$(x_e -) = a'(y - \check{y}) \tag{6.56}$$

Normally, one would expect that $a' = 1/a$. With regression, however, this is only true if the correlation coefficient r = 1, because

$$a'.a = r^2 \tag{6.57}$$

The intermediate regression coefficient a^* becomes

$$a^* = \sqrt{(a/a')} = s_y/s_x \tag{6.58}$$

which gives the geometric mean of the coefficients a and $1/a'$. The expression of the intermediate regression line then becomes:

$$(y^* - \breve{y}) = a^*(x^* - \chi) \tag{6.59}$$

or

$$y^* = a^* x^* + b^* \tag{6.60}$$

where

$$b^* = y^* - a^* \chi \tag{6.61}$$

The symbols y* and x* are used to indicate the y and x values on the intermediate regression line.

Because the intermediate regression coefficient a* results from the regression of y upon x and of x upon y, one speaks here of a two-way regression.

The intermediate regression line is, approximately, the bisectrix of the angle formed by the regression lines of y upon x and of x upon y in the central point (χ, $\breve{y}$).

*Confidence interval of the coefficient a**

In conformity with Equations 6.46 and 6.47, the confidence limits of the intermediate regression coefficient a* are given by

$$a^*_u = a^* + t\, s_{a^*} \tag{6.62}$$

$$a^*_v = a^* - t\, s_{a^*} \tag{6.63}$$

where the standard deviation s_{a^*} of a* is found from

$$s_{a^*} = a^*\, s_a/a = a^*\, s_{a'}/a' \tag{6.64}$$

This shows that the relative standard deviation s_{a^*}/a^* is considered equal to the relative standard deviation s_a/a and $s_{a'}/a'$. In general, the relative standard deviations of all regression coefficients are equal

$$s_{a^*}/a^* = S_a/a = S_{a'}/a' = s_{1/a'}/(1/a') = a' s_{1/a'}$$

Confidence belt of the intermediate regression line

The confidence belt of the intermediate regression line can be constructed from the confidence intervals of y^* or x^*. We shall limit ourselves here to the confidence intervals of y^*.

In conformity with, Equations 6.49, 6.50, and 6.51 we can write

$$y_{u*} = y^* + t\, s_{y*} \tag{6.65}$$

$$y_{v*} = y^* - t\, s_{y*} \tag{6.67}$$

where

$$s_{y*} = \sqrt{(s_{\breve{y}2} + (x^* - \chi)^2 s_{a*})} \tag{6.68}$$

And in conformity with Equations 6.53, 6.54, and 6.55 we get

$$s_{b*} = \chi\ s_{a*} \tag{6.69}$$

$$b^*_u = b^* + t\, s_{b*} \tag{6.70}$$

$$b^*_v = b^* - t\, s_{b*} \tag{6.71}$$

An example of how to use these equations follows.

Example 6.3 Two-way regression

Let us assume that we wish to determine the hydraulic conductivity of a soil with two different layers. We have observations on drain discharge (q) and hydraulic head (h), and we know that q/h and h are linearly related: $q/h = a^*h + b^*$. The hydraulic conductivity can be determined from the parameters a* and b* (Chapter 12), whose values can be found from a two-way regression.

In Table 6.12 one finds the two-way regression calculations, made according to the equations above, in which h replaces x and z = q/h replaces y. Although the values of both a* and b* are significantly different from zero, we can see that they are not very accurate. This is partly owing to the high scatter of the data and partly to their limited number (Figure 6.22).

Figure 6.22 shows the confidence intervals of the regression line, which are based on the confidence intervals of b* and a* that were calculated in Table 6.11. Despite the fairly high

correlation coefficient (r = 0.83), the confidence intervals are quite wide. This problem can be reduced if we increase the number of observations.

Table 6.12 Values of the hydraulic head (h), the discharge (q) and their ratio (z=q/h) in a drainage experimental field

nr	q	h	z = q/h
1	0.0009	0.17	0.0053
2	0.0011	0.19	0.0058
3	0.0022	0.28	0.0079
4	0.0020	0.30	0.0066
5	0.0034	0.40	0.0085
6	0.0032	0.40	0.0080
7	0.0031	0.42	0.0074
8	0.0035	0.45	0.0078
9	0.0044	0.48	0.0092
10	0.0042	0.51	0.0082
11	0.0057	0.66	0.0086

$\Sigma h = 4.26$ $\Sigma h^2 = 1.86$ $\Sigma hz = 73.46$

$\Sigma z = 0.0833$ $\Sigma z^2 = 0.000645$ $n = 11$ $n-2 = 9$

$h_{av} = \Sigma h/n = 0.387$ $= 0.57$ (average h)

$z_{av} = \Sigma z/n = 0.00757$ $= 4.7$ (average z)

Equation 6.38 : $\Sigma' h^2 = 1.86 - 4.26^2/11$ $= 0.209$

Equation 6.39 : $\Sigma' z^2 = 0.00645 - 0.0333^2/11$ $= 0.0000145$

Equation 6.37 : $\Sigma' hz = 73.46 - 4.26 \times 0.0833/11$ $= 0.00144$

Equation 6.44 : $r = 0.00144/\sqrt{(0.209 \times 0.0000145} = 0.83$

$r^2 = 0.83^2$ $= 0.69$

Equation 6.51 : $a' = 0.69/0.0069$ $= 100$

Equation 6.53 : $a^* = \sqrt{(0.0069 \times 0.0100)}$ $= 0.0083$

Equation 6.48 : $\Sigma\varepsilon^2 = (1-0.69) \times 0.0000415$ $= 0.00000450$

Equation 6.48 : $s_a = \sqrt{\{0.00000450/(9 \times 0.209)\}}$ $= 0.00155$

Equation 6.53 : $s_{a*} = 0.0083 \times 0.00155/0.0069$ $= 0.00186$

Table 6.9 : $F_u = 0.005$, $d = 9$ → $t_{90\%} = 1.83$

Equation 6.57 : $a^*_u = 0.0083 + 1.83 \times 0.00186$ $= 0.0117$

Equation 6.47 : $a^*_v = 0.0083 - 1.83 \times 0.00186$ $= 0.0049$

Equation 6.61 : $b^* = 0.00757 - 0.0083 \times 0.0378$ $= 0.0044$

Equation 6.53 : $s_{b*} = 0.387 \times 0.0019$ $= 0.00074$

Equation 6.69 : $b^*_u = 0.0044 + 1.83 \times 0.00074$ $= 0.0058$

Equation 6.69 : $b^*_v = 0.0044 - 1.83 \times 0.00074$ $= 0.0030$

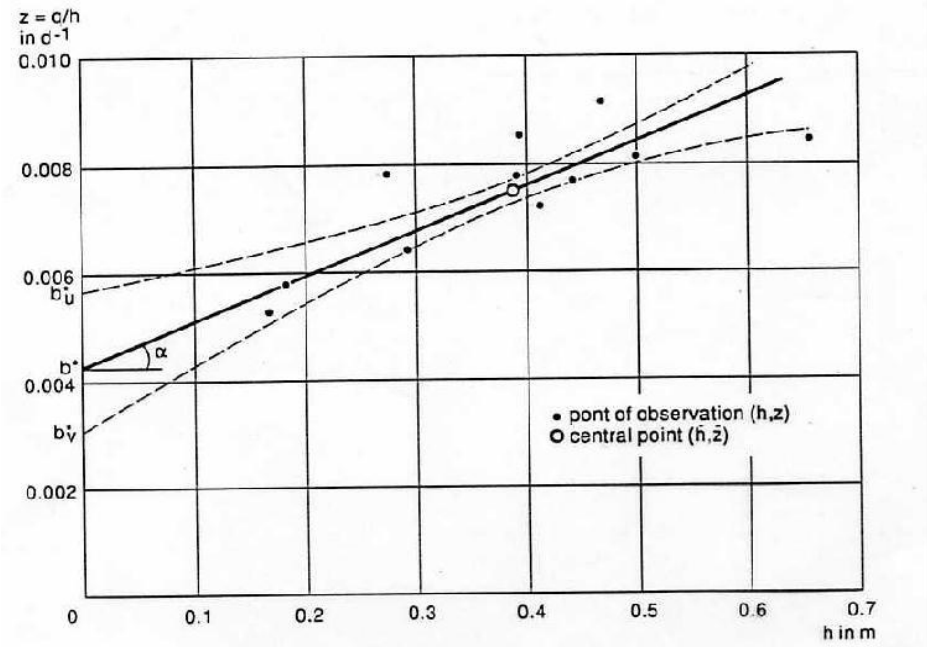

Figure 6.22 Two-way regression with the data of Table 6.12

(in this figure the symbols $\bar{h}$ and $\bar{z}$ are used instead of h_{av} and z_{av})

6.5.5 Segmented Linear Regression

In agriculture, crops will often react to a production factor x within a certain range of x, but not outside this range. One might consider using curvilinear regression to calculate the relation between crop yield (y) and x, but the linear regression theory, in the form of segmented linear regression, can also be used to calculate the relation.

Segmented linear regression applies linear regression to (x, y) data that do not have a linear relation. It introduces one or more breakpoints, whereupon separate linear regressions are made for the linear segments. Thus, the non-linear relation is approximated by linear segments. Nijland and El Guindy (1986) used it to calculate a multi-variate regression. A critical element is the locating of the breakpoint. Oosterbaan et al. (1990) have presented a method for calculating confidence intervals of the breakpoints so that the breakpoint with the smallest interval i.e. the optimum break point, can be selected.

Example 6.4 Segmented linear regression with one breakpoint

Segmented linearization (or broken-line regression) will be illustrated with the data from Figure 6.21 as shown again in Figure 6.23. In this example the optimum breakpoint was at x = 0.55 m. The subsequent calculations are presented in Table 6.13.

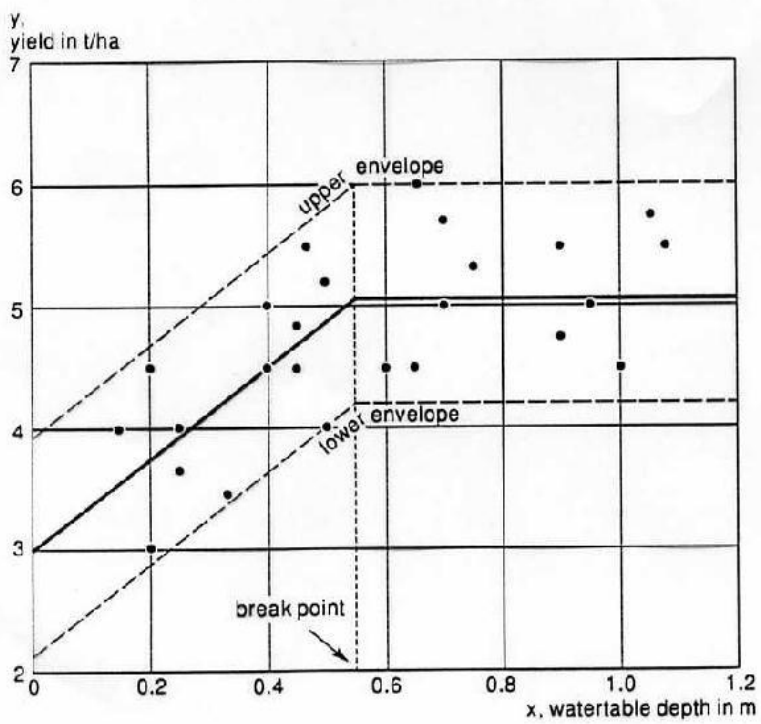

Figure 6.23 Segmented linear regression with the same data as in Figure 6.21

Table 6.13 Segmented linear regression calculations with the data of Table 6.11

1) Segment with x < 0.55m

$\Sigma x = 4.94$ $\Sigma x^2 = 1.94$ $\Sigma xy = 22.12$
$\Sigma y = 60.7$ $\Sigma y^2 = 269.26$ $n = 14$ $n-2 = 12$
$\chi = \Sigma x/n = 4.94/14 = 0.35$ m (average x)
$\breve{y} = \Sigma y/n = 60.7/14 = 4.3$ t/ha (average y)

Equation 6.38 : $\Sigma' x^2 = 1.94 - 4.94^2/14 = 0.19$
Equation 6.39 : $\Sigma' y^2 = 269.26 - 60.7^2/14 = 6.09$
Equation 6.37 : $\Sigma' xy = 22.12 - 4.94 \times 60.7/14 = 0.70$
Equation 6.36 : $a = 0.70/0.19 = 3.62$
Equation 6.35 : $b = 4.3 - 3.62 \times 0.35 = 3.06$
Equation 6.44 : $r = 0.70/\sqrt{(0.19 \times 6.09)} = 0.64 \rightarrow r^2 = 0.41$
Equation 6.48 : $\Sigma\varepsilon^2 = (1-0.41) \times 6.09 = 3.57$
Equation 6.48 : $s_a = \sqrt{\{3.57/(14 \times 0.19)\}} = 0.42$
Table 6.9 : $F_u = 0.005$, $d = 12 \rightarrow t_{90\%} = 1.78$
Equation 6.46 : $a_u = 3.62 + 1.78 \times 0.42 = 5.83$
Equation 6.47 : $a_v = 3.62 - 1.78 \times 0.42 = 1.40$
$x = \chi \rightarrow$
Equation 6.51 : $s_{\hat{y}} = s_{Y'} = \sqrt{\{3.57/(14 \times 12)\}} = 0.15$
Equation 6.49 : $x=\chi \rightarrow \hat{y}_u = 4.3+1.78 \times 0.15 = 4.6$ t/ha
Equation 6.50 : $x=\chi \rightarrow \hat{y}_u = 4.3 - 1.78 \times 0.15 = 4.0$ t/ha

2) Segment with x < 0.55m

$\Sigma x = 9.39$ $\Sigma x^2 = 8.54$ $\Sigma xy = 51.35$
$\Sigma y = 61.9$ $\Sigma y^2 = 322.41$ $n = 12$ $n-2 = 10$
$\chi = \Sigma x/n = 9.39/12 = 0.83$ m (average x)
$\breve{y} = \Sigma y/n = 60.7/26 = 5.2$ t/ha (average y)

Equation 6.38 : $\Sigma' x^2 = 8.54 - 9.39^2/12 = 0.32$
Equation 6.39 : $\Sigma' y^2 = 322.41 - 61.9^2/12 = 3.11$
Equation 6.37 : $\Sigma' xy = 51.35 - 9.39 \times 61.9/12 = 0.12$
Equation 6.36 : $a = 0.12/0.32 = 0.38$
Equation 6.35 : $b = 5.2 - 0.38 \times 0.83 = 4.48$
Equation 6.44 : $r = 0.12/\sqrt{(0.32 \times 3.11)} = 0.14 \rightarrow r^2 = 0.02$
Equation 6.48 : $\Sigma\varepsilon^2 = (1-0.02) \times 3.11 = 3.06$
Equation 6.48 : $s_a = \sqrt{\{3.06/(12 \times 0.32)\}} = 0.89$
Table 6.9 : $F_u = 0.005$, $d = 10 \rightarrow t_{90\%} = 1.81$
Equation 6.46 : $a_u = 0.38 + 1.81 \times 0.89 = 2.15$
Equation 6.47 : $a_v = 0.38 - 1.81 \times 0.89 = -1.38$
$x = \chi \rightarrow$
Equation 6.51 : $s_{\hat{y}} = s_{Y'} = \sqrt{\{3.06(12 \times 10)]} = 0.16$
Equation 6.49 : $x=\chi \rightarrow \hat{y}_u = 5.2+1.81 \times 0.16 = 5.5$ t/ha
Equation 6.50 : $x=\chi \rightarrow \hat{y}_u = 5.2 - 1.81 \times 0.16 = 4.9$t/ha

Discussion

The total $\Sigma\varepsilon^2 = 3.57 + 3.06 = 6.63$ in Table 6.13 is lower than the $\Sigma\varepsilon^2 = 7.87$ of Example 6.2, which represents the linear regression using all the data without a breakpoint. This means that the segmented regression gives a better explanation of the effect of water table depth on crop yield than does the un-segmented regression. One can test whether this improvement is significant at a certain confidence level by comparing the reduction in $\Sigma\varepsilon^2$ with the residual variation after segmented linear regression. One then checks the variance ratio using an F-test, a procedure that is not discussed here. In this example, the improvement is not statistically significant. This difficulty could be obviated, however, by the collection of more data.

The regression coefficient ($a = 0.38$) for the data with $x > 0.55$ is very small and insignificant at the 90% confidence level because $a_v < 0 < a_u$, meaning that no influence of x upon y can be established for that segment.

On the other hand, the regression coefficient ($a = 3.62$) for the data with $x < 0.55$ is significant at the chosen confidence. Hence, the yield (y) is significantly affected by water tables (x) shallower than 0.55 m.

In accordance with Equation 6.31, the regression equations become

$$\hat{y} = \breve{y} = 5.2 \qquad [x > 0.55 \text{ m}]$$

$$\hat{y} = 3.62\,(x-0.35) + 4.3 = 3.62\,x + 3.1 \qquad [x < 0.55 \text{ m}]$$

The intersection point of the two lines need not coincide exactly with the breakpoint; when the segmented regression is significant, the difference is almost negligible.

Using n_v = number of data with $x < 0.55$ and n_t = total number of data, and assuming that the points in Figure 6.23 represent fields in a planned drainage area, one could say that $n_v/n_t = 14/26 = 54\%$ of the fields would benefit from drainage to bring the water tabledepth x to a value of at least 0.55 m, and that 46% would not. An indication of the average yield increase for the project area could be obtained as follows, with χ being the average water table depth in the segment $x < 0.55$

$$\Delta y = a(0.55-\chi)n_v/n_t = 3.62(0.55-0.35)0.54 = 0.4 \text{ t/ha}$$

with confidence limits $\Delta y_u = 0.6$ and $\Delta y_v = 0.2$, which are calculated with $a_u = 5.83$ and $a_v = 1.40$ instead of $a = 3.62$. From Example 6.2, we know that the average current yield is $\breve{y} = 4.7$ t/ha. Accordingly, we have a relative yield increase of $0.4/4.7 = 9\%$, with 90% confidence limits of $0.6/4.7 = 13\%$ and $0.2/4.7 = 4\%$.

NOTE

The computer program SegReg (on www.waterlog.info) has been designed to automatically detect breakpoints in linear regression models and present the necessary confidence intervals.

6.7 References

Nijland, H.J. and S. El Guindy 1986. Crop production and topsoil/ surface-water salinity in farmer's rice-fields, the Nile Delta. In: Smith, K.V.H. and D.W. Rycroft (eds.), Hydraulic Design in Water Resources Engineering: Land Drainage. Proceedings of the 2nd International Conference, Southampton University. Springer Verlag, Berlin. pp. 75-84.

Oosterbaan R.J., D.P. Sharma and K.N. Singh 1990. Crop production and soil salinity: Evaluation of field data from India by segmented linear regression. Symposium on Land Drainage for Salinity Control in Arid and Semi-Arid Regions, Vol. 3, Cairo, pp. 373-382.

Snedecor, G.W. and W.G Cochran 1986. Statistical methods. Iowa State University Press. 8th ed., 593 p.