Segmented linear regressions of a dependent response variable on two independent (influential) variables and how the free SegRegA software accomplishes that.

R.J. Oosterbaan
May 2022 www.waterlog.info

Abstract
When data are available of the relation between one dependent variable (Y) and two independent (influential) variables (X and Z), while the independent variables may have a range of influential values plus a range of values that have no effect on Y, it may be possible first to analyze first the relations between Y and X and between Y and Z. Selecting the relation with the highest coefficient of explanation (either X or Z), one can then analyze the relation between the residuals (Yr, the deviations of the data from the segmented regression lines) with the remaining influential variable. After this, the complex relation between Y on the one hand and X and Z on the other, can be composed. When there is no correlation between X and Z, the result can be quite satisfying. However, when there is a strong correlation, the explanation given by the second independent variable may be disappointing. In this article, detailed explanations are given how the free SegReg model (or the amplified model SegRegA) handles such a situation which is illustrated with examples.

Contents
1. Introduction
2. Principles of SegReg for segmented (Y,X,Z) regressions
3. Examples
   3.1 Crop yield (Y), depth of water table (X) and soil salinity (Z)
   3.2 Wheat yield (Y), soil salinity (X) and number of irrigations (Z)
4. Discussion
5. References

1. Introduction

Figure 1 gives the equations that are normally used to solve the linear regression of one response variable on two independent (influential, explanatory, predicting) variables.

The equation for *a* with two independent variables is:

$$a = \bar{Y} - b_1 \bar{X}_1 - b_2 X_2$$

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

and

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

Figure 1. *Screen print. Solution of a linear regression equation of one dependent on two independent variables* [Ref. 1].

The first expression in figure 1 is usually written as: $Y = b_1X_1 + b_2X_2 + a$, whereas SegReg [Ref. 2] uses X for X_1 and Z for X_2. The factors b1 and b2 are called coefficients.

When the relations between Y and X as well as between Y and Z are not linear, the solution becomes more complex because first it must be detected first which type on non-linearity comes in the picture and then the calculation of the coefficients is more cumbersome,

SegReg approaches the non-linearity by segmented linear regression as illustrated in the next figure.

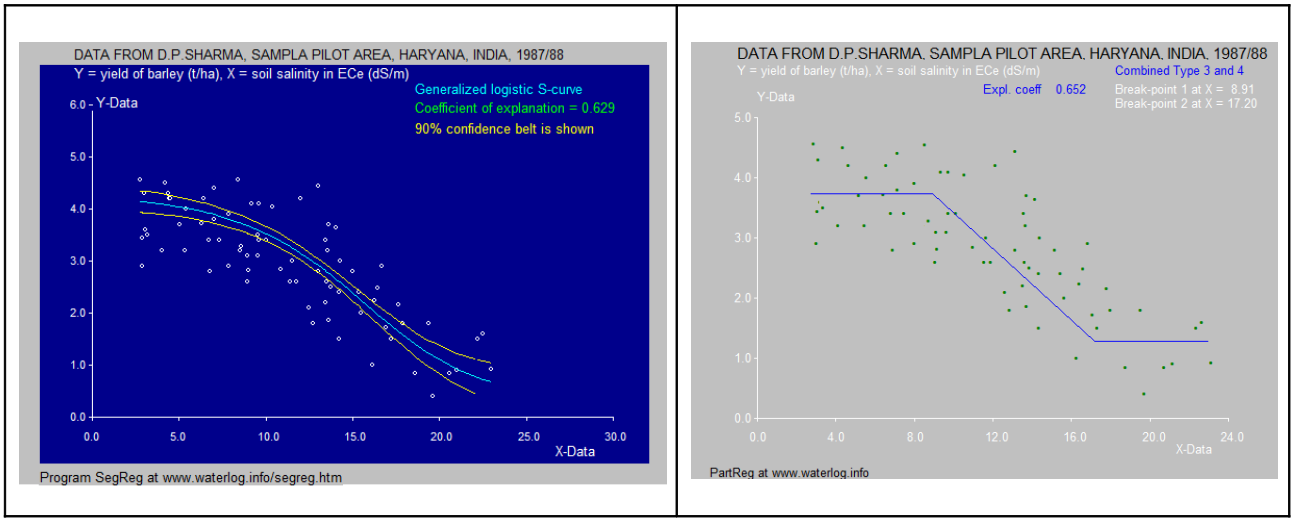


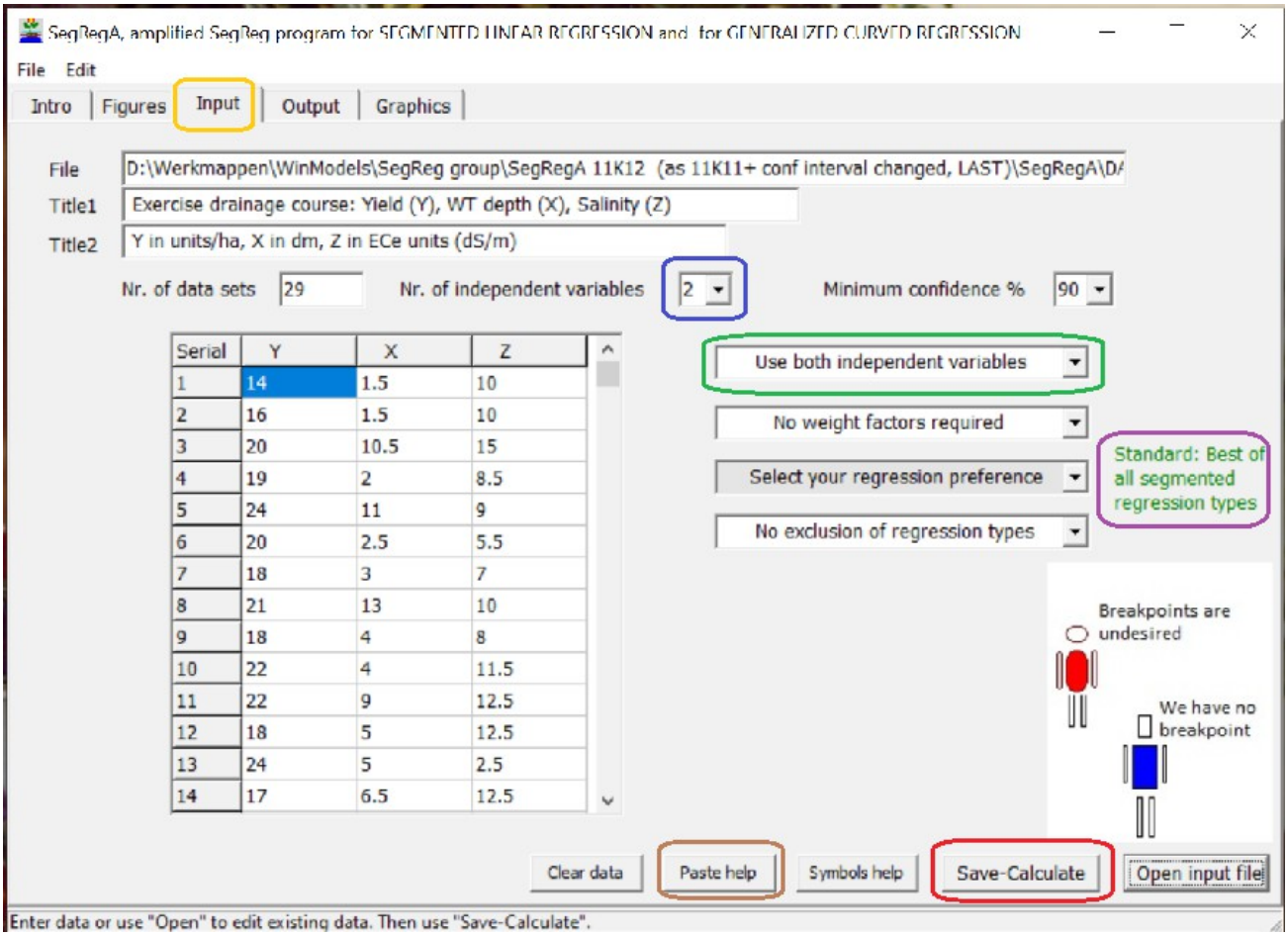
Figure 2. *The mirrored S-curve regression (left) is replaced by segmented linear regressions in z-shape (right). It concerns the yield of barley (Y, t/ha) versus soil salinity (X, ECe in dS/m) .*

The data of figure 2 stem from the Sampla pilot area, Haryana, India [Ref. 3]. It can be seen that at the soil salinity values below ECe = 9 dS/m the yield hardly affected by increasing salinity while beyond that value (the breakpoint) the yield goes down as the soil becomes too salty.

The segmentation can be used to overcome the difficulties in non-linear multivariate regression of Y upon X and Z.

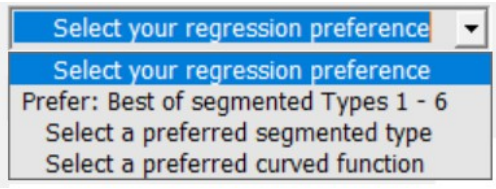
2. Principles of SegRegA for segmented (Y,X,Z) regressions

Figure 3 shows the input user interface of SegRegA (not SegReg).



Serial	Y	X	Z
1	14	1.5	10
2	16	1.5	10
3	20	10.5	15
4	19	2	8.5
5	24	11	9
6	20	2.5	5.5
7	18	3	7
8	21	13	10
9	18	4	8
10	22	4	11.5
11	22	9	12.5
12	18	5	12.5
13	24	5	2.5
14	17	6.5	12.5

Figure 3. *Input user interface of SegRegA (orange square). There are two independent variables (blue square) that will both be used (green square). Standard, the program the best of all segmented regression types (purple square). The input data can be pasted in the table (brown square). After completing the input specifications, one can click on the "Save-Calculate" button (red square) to obtain the results.*

In figure 3 it can also be seen that there is an option to select different regression procedures (gray block), but in this article the standard procedure (purple square) will be used. The various regression options can be seen at the right. For the segmentation type see reference 4 and for the curved functions see reference 5.	

After performing the calculations, SegReg or SegRegA shows the output tab sheet as depicted in figure 4.

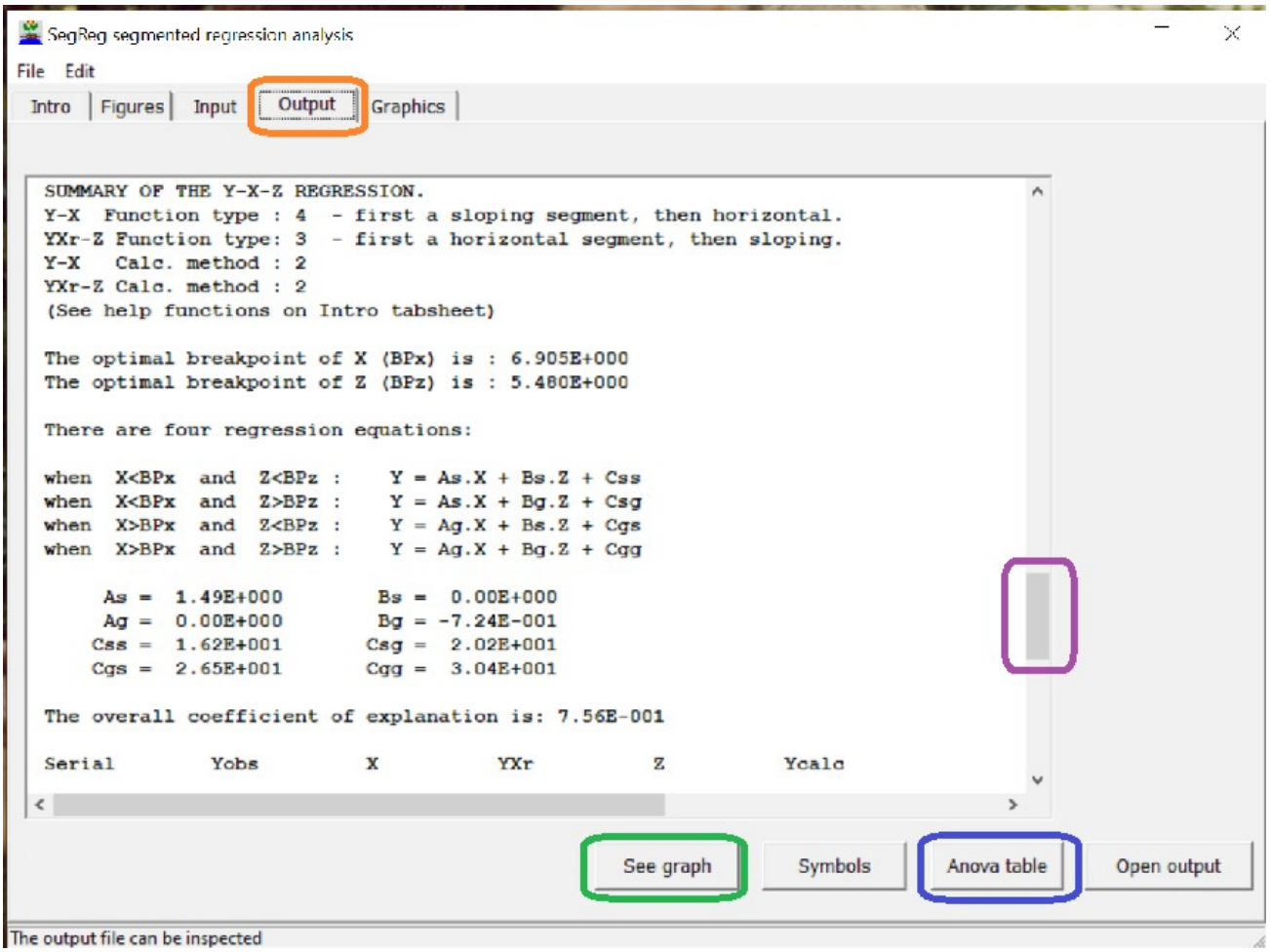


Figure 4. *Output tab sheet of SegReg or SegRegA (orange square) which can be inspected using the scroll bar (purple square). The program can also produce an Anova (analysis of variance) table (blue square) as well as graphic presentations (green square).*

In figure 4 it is seen that, amongst other, the output gives the types of regression equations, the breakpoint values of the influential X and Z variables, four regression equations with specification of their parameters and the value of the coefficient of explanation (determination), or R^2).

During the calculations, SegReg and SegRegA first determine segmented regressions of Y upon X and Y upon Z and it determines their coefficients of determination. When the independent variable X produces the larger coefficient, the program calculates the deviations (residuals) of the observed X values (YXr) from the segmented regression line and with these it carries out the segmented regression on Z. However, when Z produces the larger coefficient of explanation it finds the residuals of the observed Z values (YZr) and the algorithm implements the segmented regression of YZr on X. Assuming the first case is true then we have:

$$Y = As*X + M + YXr \qquad [X<BPx]$$
$$Y = Ag*X + N + YXr \qquad [X>BPx]$$

where BPx represent the break point of X (see figure 2 right-hand side).

The second step is the segmented regression of YXr on Z as follows:

$$YXr = Bs*Z + P \qquad [Z<BPz]$$
$$YXr = Bg*Z + Q \qquad [Z>BPz]$$

where BPz represent the break point of Z.

Substituting the second set of equations into the first set gives:

$$Y = As*X + M + Bs*Z + P = As*X + Bs*Z + Css \qquad [X<BP, Z<BP]$$
$$Y = Ag*X + N + Bs*Z + P = Ag*X + Bs*Z + Cgs \qquad [X>BP, Z<BP]$$
$$Y = As*X + M + Bg*Z + Q = As*X + Bg*Z + Csg \qquad [X<BP, Z>BP]$$
$$Y = Ag*X + N + Bg*Z + Q = Ag*X + Bg*Z + Cgg \qquad [X>BP, Z>BP]$$

as shown in figure 4.

The kind of graphic presentations mentioned in the subscript of figure 4 are defined in the graph selection menu, see figure 5.

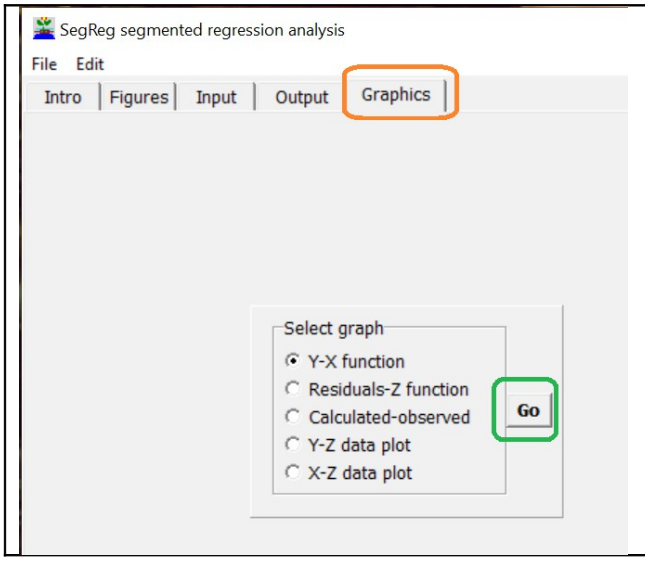


Figure 5.

The graphics tabsheet (orange square) depicts 5 graph type options. When clicking on the "Go" button (green square) the selected graph type (in this case the Y-X function) will be shown, see the examples furtheron.

3. Examples

3.1 Crop yield (Y), depth of water table (X) and soil salinity (Z)

The input data for example 3.1 are shown in figure 3. The output file can be seen in figure 4 together with the values of the parameters of the equation for segmented regression. The first regression is that of Y upon X as this gives a higher explanation the that of Y upon Z, so that the second regression will be done with Yxr upon Z. The breakpoint of X (depth water table) is 6.9 dm and for Z (soil salinity, ECe) it is 5.8 dS/m. The file also reveals that the overall coefficient of explanation equals 0.756 or 75.6 % (figure 8).

The first graph selection (see figure 5) is Y upon X and that graph is demonstrated in figure 6.

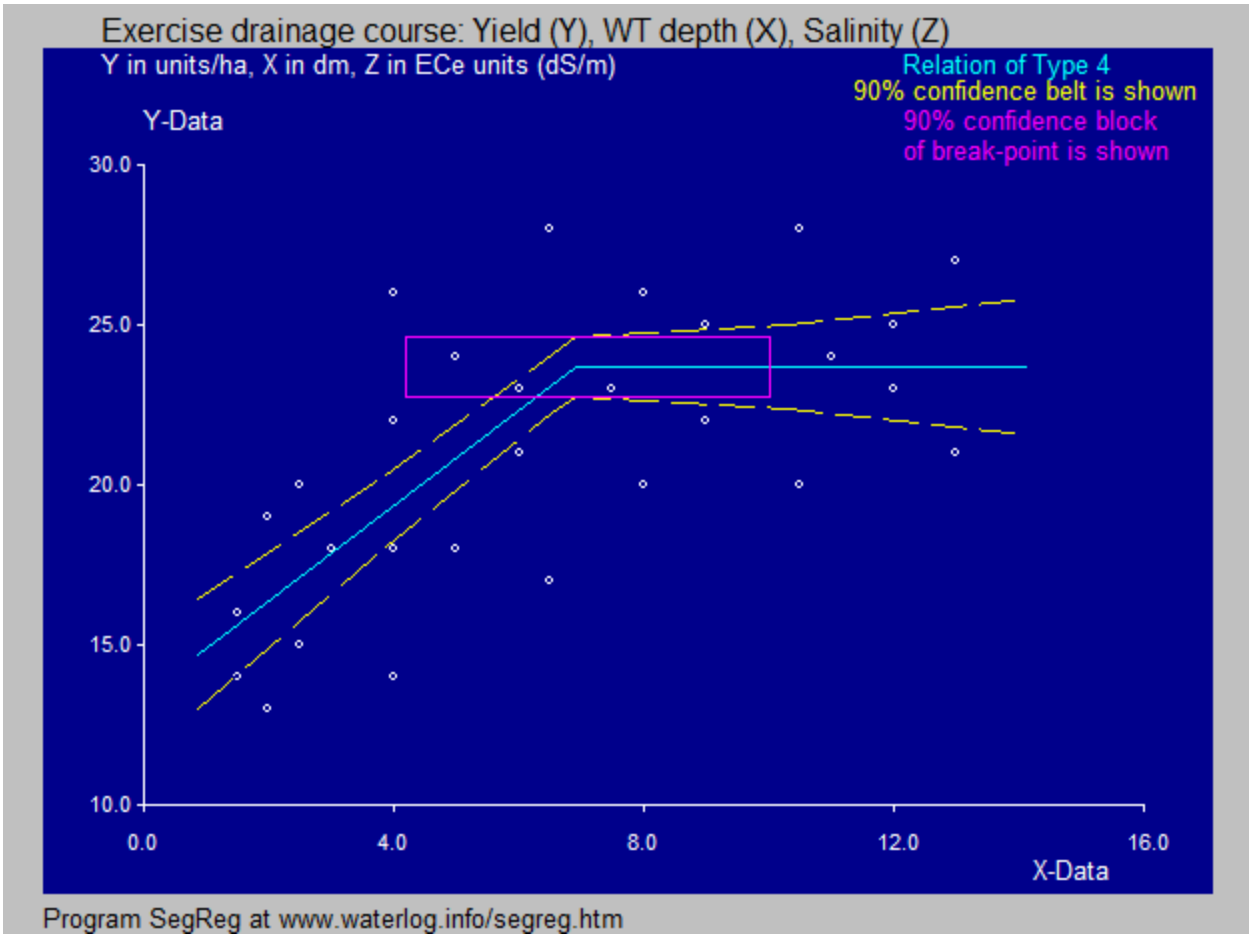


Figure 6. *Segmented regression of Type 4 giving the relation between crop yield (Y) and depth of the water table (X). When the water table is deeper than 6.9 dm the yield is not influenced, but when it is shallower the yield diminishes as the water table gets shallower. The coefficient of explanation equals 0.475 or 47.5 % which can be found in the output file.*

The second graph selection (see figure 6) is YXr (the deviations in figure 5) upon Z (soil salinity) and that graph is depicted in figure 7.

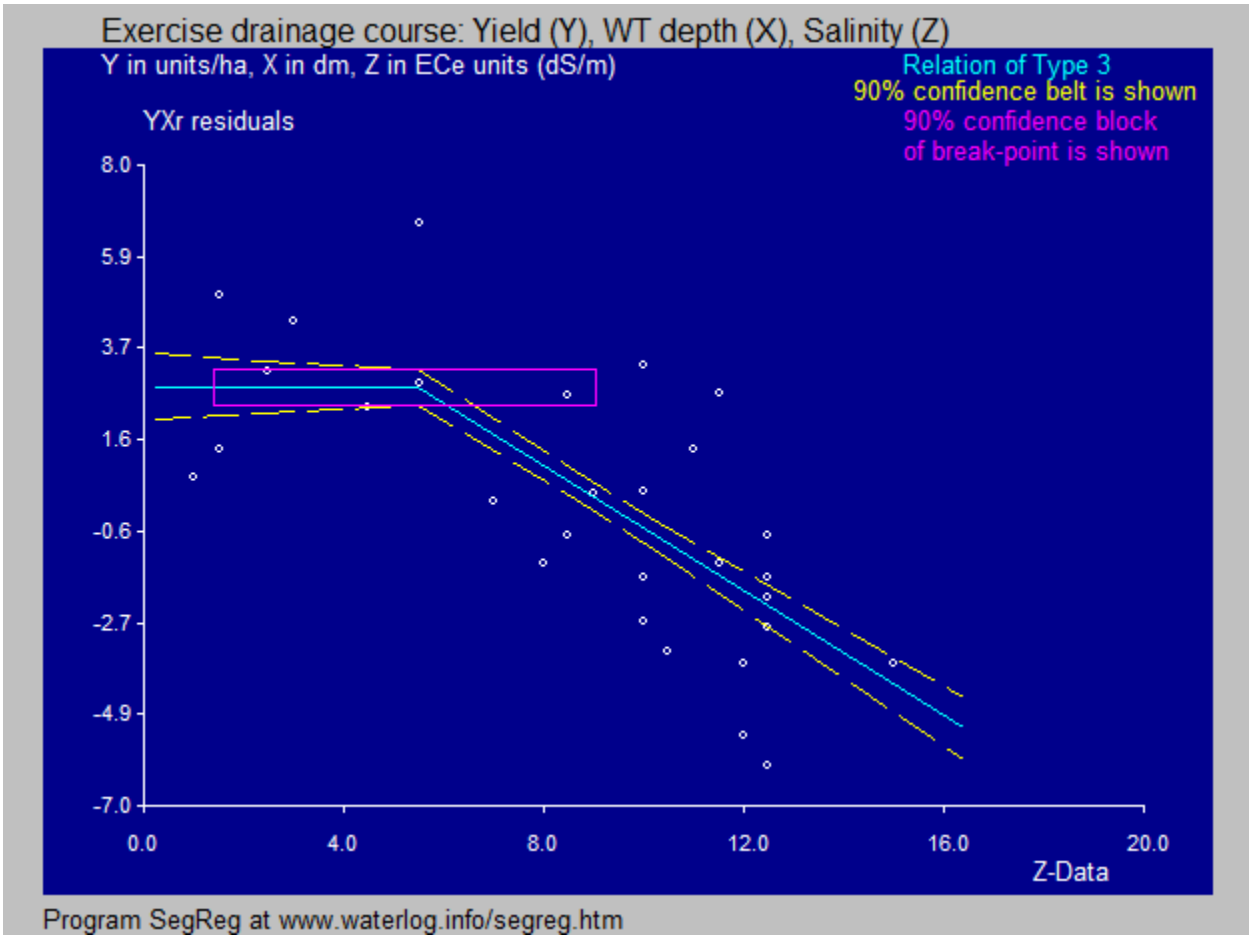


Figure 7. *Segmented regression of Type 3 giving the relation between the YXr residuals in figure 6 and soil salinity (Z). When the soil salinity is lower than ECe=5.8 dS/m the residual yield is not influenced, but when it is higher the residuals diminish as the salinity gets higher.*

A graph showing the relation between the observed yield values and the calculated ones according to the four equations given at the end of the previous section 2 is presented in figure 8.

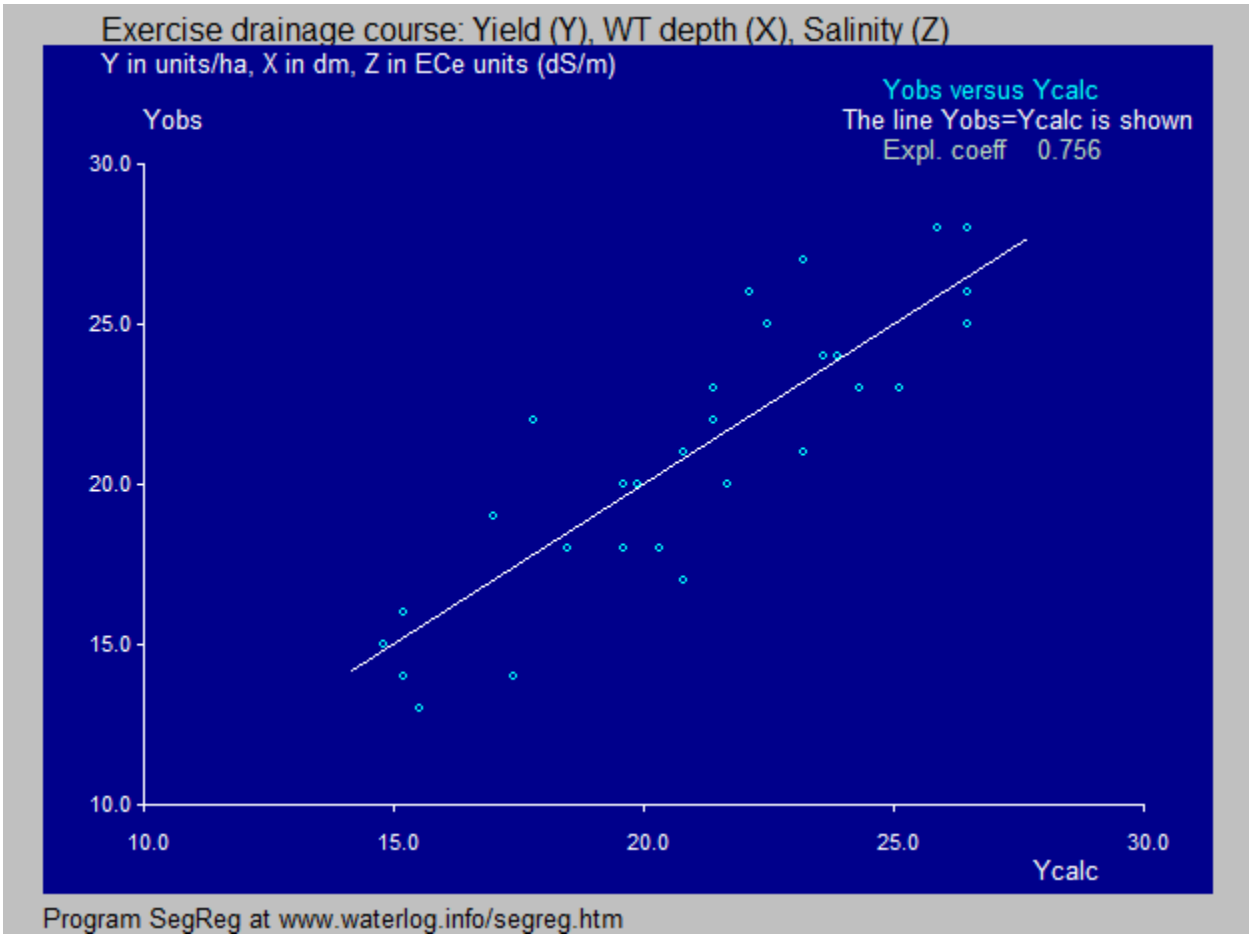


Figure 8. *Relation between the observed yield values (Yobs) and the calculated ones (Ycalc) according to the four equations given at the end of the previous section 2.The coefficient of explanation (determination) or R2 is quite high (0.756 or 75.6 %) which is a considerable improvement over the coefficient 47.5 % mentioned in the subscript of figure 6.*

When, with SegRegA (not SegReg), changing the option “use both independent variables” (see figure 3) into “use second independent variable only”, the result will be as in figure 9.

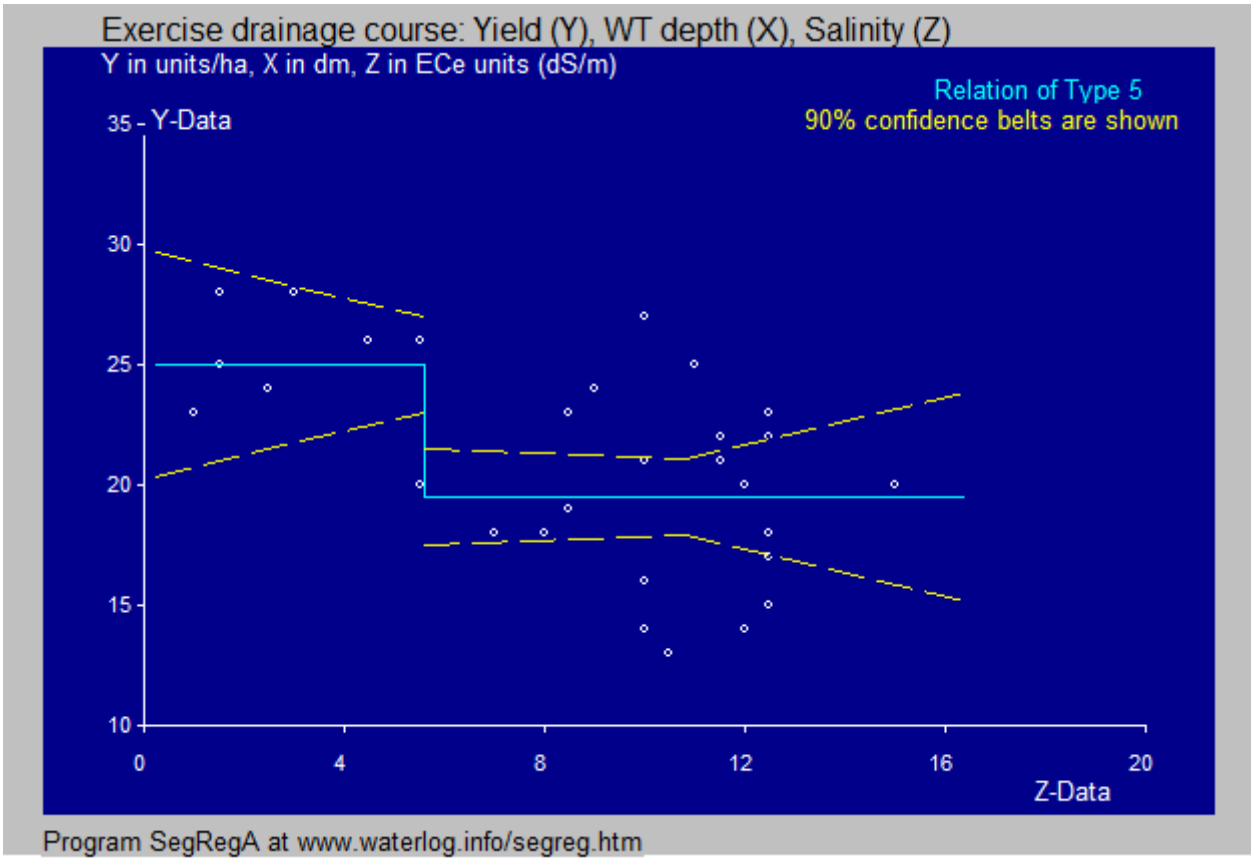


Figure 8. *Relation of type 5 between the crop yield (Y) and the soil salinity (Z). The scatter of the points is quite high and SegReg or SegRegA can only detect a jump at Z=ECe=5.6 dS/m.*

Thus far we have seen relations of Type 3, 4 and 5 of the many identified in reference 4. The types 3, 4, and 5, according to that reference are actually types 3hd, 4ah and 5ud.

As a final observation, the relation between the independent variables X and Z is shown in figure 9. There is hardly any correlation between them so they do not influence each other and their subsequent use can lead to success.

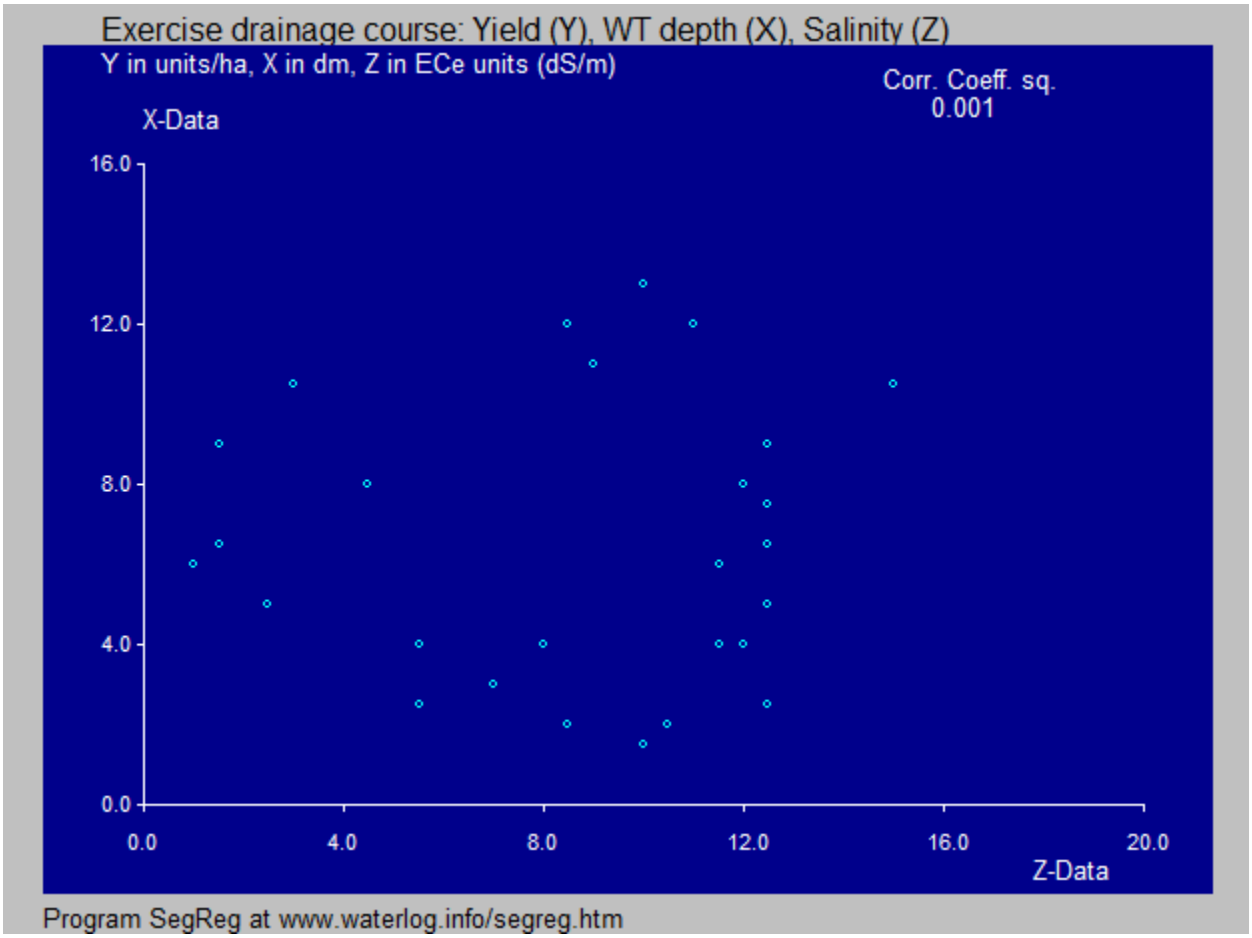


Figure 9. *The independent variables X (depth to water table) and Z (soil salinity) have hardly any correlation so they do not influence each other and their subsequent use can lead to success. Apparently the depth of the water table does not influence the soil salinity*.

3.2 Wheat yield (Y), soil salinity (X) and number of irrigations (Z)

The input data were provided by O.P. Singh (personal communication) and stem from the Gohana area, Haryana.

The first graph selection (see figure 10) is the wheat yield Y upon soil salinity X and that graph is demonstrated in figure 10.

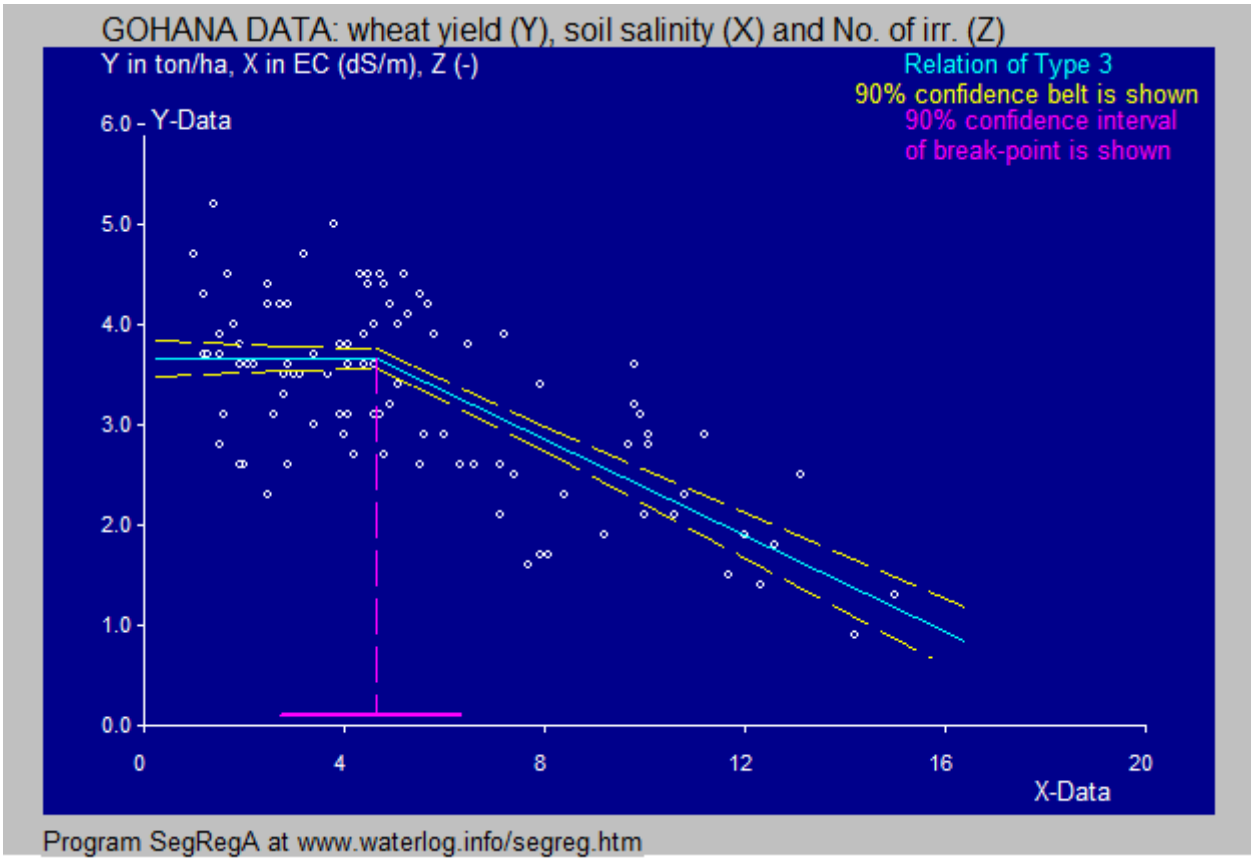


Figure 10. *Segmented regression of Type 3 giving the relation between wheat yield (Y) and soil salinity (X). The coefficient of explanation equals 0.447 or 44.7 % which can be found in the output file.*

The second graph selection (see figure 11) is YXr (the deviations in figure 510) upon Z (number of irrigations) and that graph is depicted in figure 11.

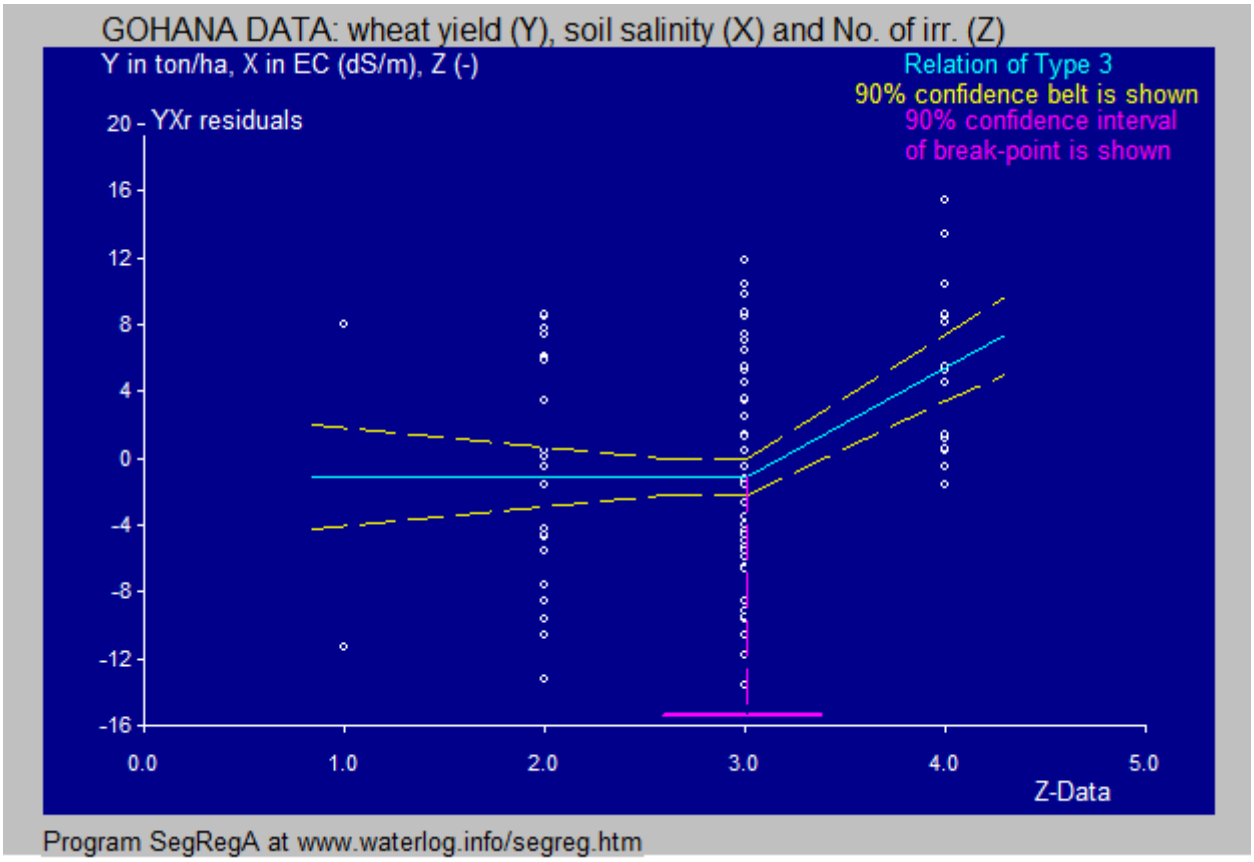


Figure 11. *Segmented regression of Type 3 (actually Type 3ha instead of 3hd in figure 7) giving the relation between the YXr residuals in figure 6 and soil number of irrigations (Z). When the number of irrigation is 3 or less the residual yield is not influenced, but when it is higher the residuals increase.*

A graph showing the relation between the observed yield values and the calculated ones according to the 4 equations given at the end of the previous section 2 is presented in figure 12.

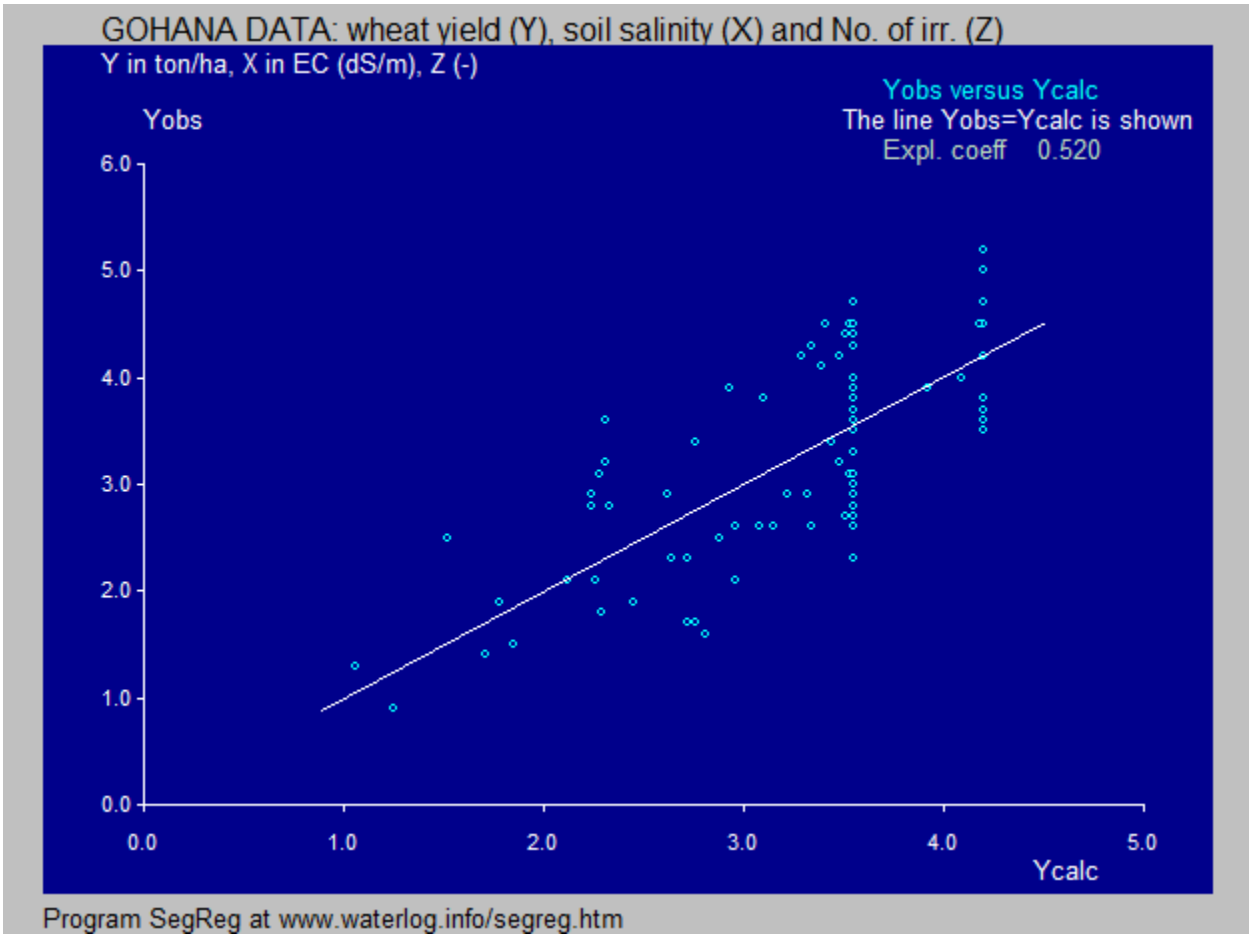


Figure 12. *Relation between the observed yield values of wheat (Yobs) and the calculated ones (Ycalc) according to the four equations given at the end of the previous section 2. The coefficient of explanation (determination) or* R^2 *is 0.52 or 52 %\ which is a reasonable but not impressive improvement over the coefficient 44.7 % mentioned in the subscript of figure 10. Apparently there is a scarcity of irrigation water in the Gohana area so that the influential number of irrigation of 4 or more is seldom achieved.*

When, with SegRegA (not SegReg), changing the option “use both independent variables” (see figure 3) into “use second independent variable only”, the result will be as in figure 13.

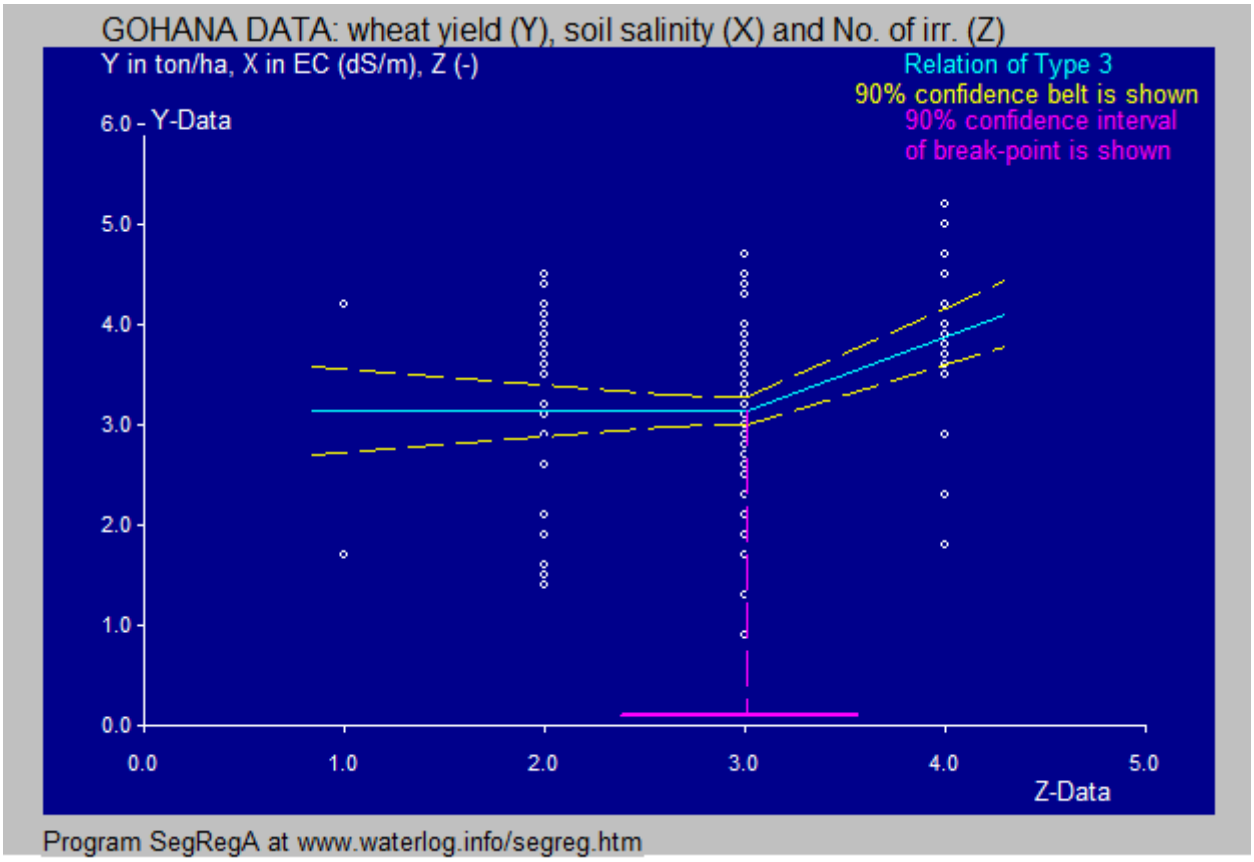


Figure 13. *Relation of type 3 between the crop yield (Y) and the number of irrigations (Z). The scatter of the points is quite high. 0.095 or 9.5% only.*

Thus far we have seen relations of Type 3, 4 and 5 of the many identified in reference 4. The types 3, 4, and 5, according to that reference are actually types 3hd, 4ah and 5ud.

As a final observation, the relation between the independent variables X and Z is shown in figure 14. There is no correlation between them so they do not influence each other and their subsequent use can lead some success.

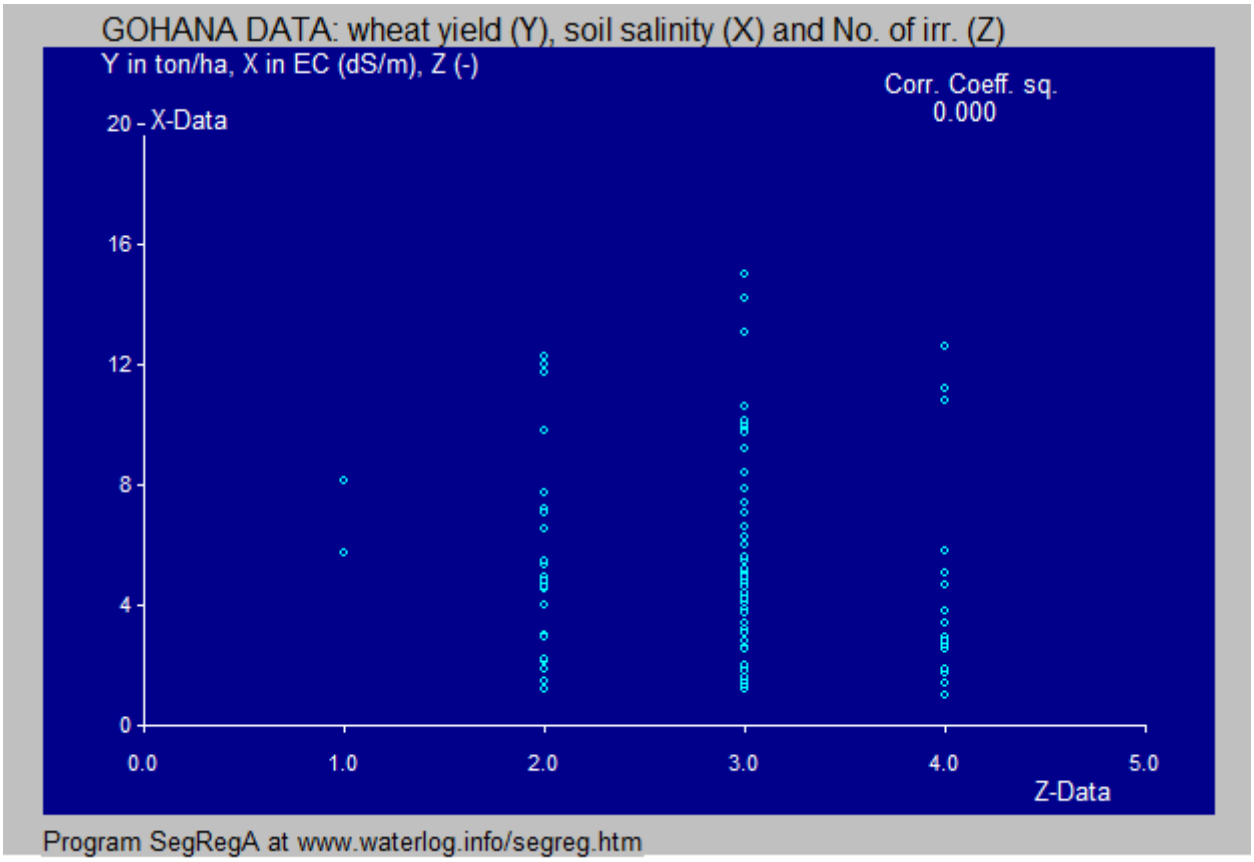


Figure 14. The independent variables X (soil salinity) and Z (number of irrigations) have hardly any correlation so they do not influence each other and their subsequent use can lead to some success. Apparently the number of irrigations has no influence on the soil salinity.

4. Conclusions

When data are available on one dependent (response) variable Y and two independent (influential) variables X and Z, employing the segmented regression analysis instead of curved regressions can lead to success, the more so since the procedure with the curved regressions is more complex.

The segmented regressions need to be done sequentially by first executing a segmented regression of Y on X or Z, depending on which of the two gives the highest coefficient of explanation, followed by a segmented regression of the residuals (the deviations of the observed Y values from the calculated ones) on the other influential variable (Z or X respectively).

It helps when X and Z are not strongly correlated.

5. References

[**Ref. 1**]
Michael T. Brannick. *Course Materials and Research Website*. On line:
http://faculty.cas.usf.edu/mbrannick/regression/Part3/Reg2.html

[**Ref. 2**]
SegReg or SegRegA, free software calculator for segmented linear regression. On line:
https://www.waterlog.info/segreg.htm

[**Ref. 3**]
Free calculator for the determination of positive and inverted S-curves for the response function of influential treatments or conditions with examples of crop yield versus soil salinity and depth of the water table

[**Ref. 4**]
Types of segmented regression and their statistical determination

[**Ref. 5**]
Analysis of Different Curved Regressions using Free Software and Selection of the Appropriate Type Based on Statistical Tests for Goodness of Fit and Analysis of Variance. In: International Journal of Mathematical and Computational Methods Volume 6, 2021. On line:
https://www.iaras.org/iaras/filedownloads/ijmcm/2021/001-0007(2021).pdf