Testing the statistical significance of the improvement of cubic regression compared to quadratic regression using analysis of variance (ANOVA)

R.J. Oosterbaan on www.waterlog.info , public domain.

Abstract
When performing a regression analysis, it is recommendable to test the statistical significance of the result by means of an analysis of variance (ANOVA). In a previous article, the significance of the improvement of segmented regression, as done by the SegReg calculator program, compared to simple linear regression was tested using analysis of variance (ANOVA). The significance of the improvement of a quadratic (2nd degree) regression over a simple linear regression is tested in the SegRegA program (A stands for "amplified") by means of ANOVA. Also, here the significance of the improvement of a cubic (3rd degree) over a simple linear regression is tested in the same way. In this article, the significance test of the improvement of a cubic regression over a quadratic regression will be dealt with.

Contents

1. Introduction

When performing a regression analysis, it is recommendable to test the statistical significance of the result by means of an analysis of variance (ANOVA).

In a previous article [Ref. 1], the significance of the improvement of segmented regression, as done by the SegReg program [Ref. 2], compared to simple linear regression was dealt with using analysis of variance. The significance of the improvement of a quadratic (2nd degree) regression over a simple linear regression is tested in the SegRegA program (the A stands for "amplified"), also by means of ANOVA [Ref. 3].

In an article [Ref. 4], the yield response of a potato variety was shown to be cubic (3rd degree, figure 1), but a comparison with a quadratic model was not shown. In this article the comparison will be made.

The potato variety is called "927" and was tested in the Salt Farm Texel [Ref. 5]

Figure 2 shows the quadratic regression for the same data. This picture show less curvature than figure 1.

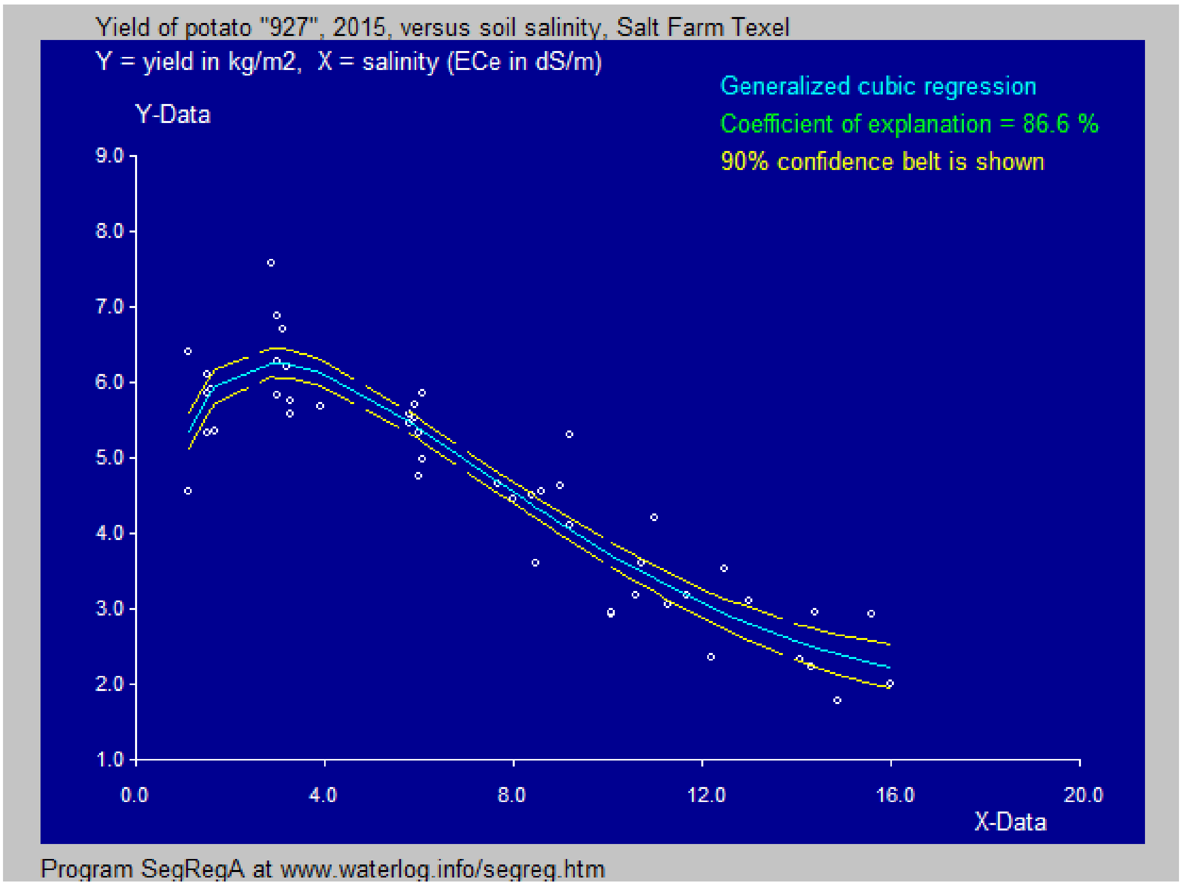

Figure 1. Cubic regression using the "927" data of the Salt Farm Texel [Ref. 4].
*The regression equation is Y = 0.537*Z^3 – 4.70*Z^2 + 11.2*Z – 1.84, where Z = X^0.49, the exponent 0.49 effectuating a generalization of the cubic regression (in other words the X values are raised to the power 0.49 before the cubic regression is done, this to increase the goodness of fit). The coefficient of explanation (also called Rsquared) equals 86.6%*

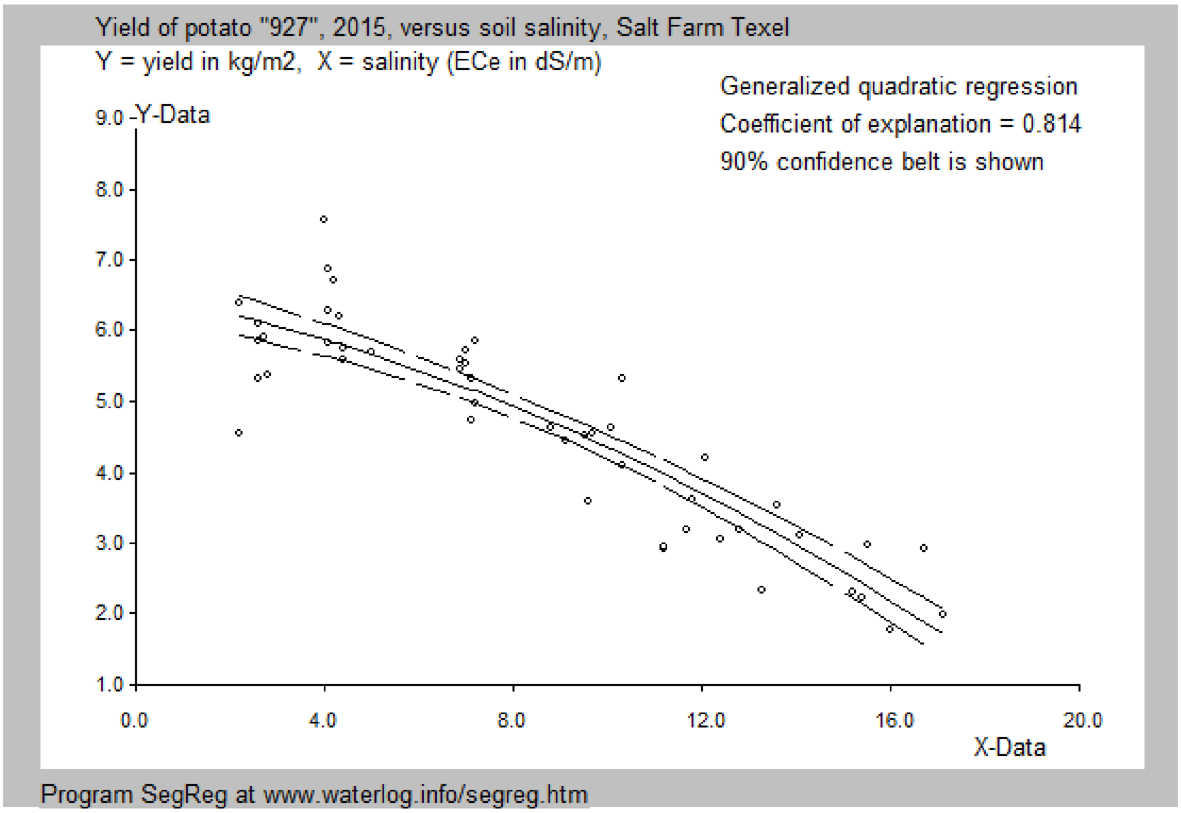

*Figure 2. Quadratic regression using the "927" data of the Salt Farm Texel [Ref. 4]. The regression equation is Y = –0.000907 * X^2 – 0.146*X + 6. The coefficient of explanation (also called Rsquared) equals 81.4%, which is less than the 86.6% for cubic regression.*

2. ANOVA symbols used

For the analysis of variance the following symbols are used:

Y	value of dependent variable
η	average value of Y (mean)
r	correlation coefficient
δ	residual after regression, also called deviation from the regression: $\delta = \ddot{Y} - Y$, with $\ddot{Y}$ being the expected value of Y according to the regression
R	overall coefficient of explanation (determination) $R = 1 - \Sigma\delta^2 / \Sigma(Y - \eta)^2$ in simple linear regression $R = r^2$ otherwise $R > r^2$
df	degrees of freedom
N	number of (X,Y) data sets
X	independent variable
SSD	sum of squares of deviations. The SSD of any variable Z equals $\Sigma\,(Z - Z_{av})^2$, Z_{av} signifying the average value of Z
Var	variance or "mean of deviations squared", it is the square value of the standard error (Var = SSD/df)

The term $\Sigma\,(Y - \eta)^2$ stands for "sum of squares of all reduced data", briefly "reduced sum of squares". With St.Dev.Y being the standard deviation of Y, one finds:

$$\Sigma\,(Y - \eta)^2 = [(\text{St.Dev.Y}) * (N - 1)]^2$$

In the following $\Sigma\,(Y - \eta)^2$ will be called SSD_0

Further we will have:

$$SSD_2 = \Sigma\,(Y - Y_{Lin})^2$$

where Y_{Lin} = Y value calculated by linear regression, i. e. the SSD value remaining after the linear regression

$SSD_1 = SSD_0 - SSD_2$, the SSD explained by linear regression

$$SSD_4 = \Sigma\,(Y - Y_{Poly})^2$$

where Y_{Poly} = Y value calculated by quadratic or cubic regression, i. e. the SSD value remaining after the polynomial regression

$SSD_3 = SSD_2 - SSD_4$, the SSD explained by polynomial regression

3. Explanatory examples

<u>A. Quadratic regression</u>

The linear regression equation reads:

$$Y_L = A_L.X + C_L$$

Where

$A_L = -2.91E\text{-}001 \quad C_L = 6.77E{+}000$

The quadratic regression equation reads:

$$\ddot{Y} = Yq = Aq.X^2 + Bq.X + Cq$$

where

$Aq = -9.07E\text{-}003 \quad Bq = -1.46E\text{-}001 \quad Cq = 6.38E{+}000$

This regression uses 3 parameters, which need to be discounted from the total number of degrees of freedom.

The general data in the example of potato "927" are as follows:

$\Sigma\,(Y - \eta)^2 = 99.200$ (total sum of squares of deviations, SSD_0)
Total nr. of data N = 48
Degrees of freedom df = 47 (one df is lost by using the parameter η)

For the quadratic regression in the example of potato "927", from the analysis of variance, it is found that (table 1):

Table 1. ANOVA analysis for quadratic regression
The corresponding SSD and Var values are printed in the same color while the corresponding df values are shown in bold.

Sum of squares of deviations (SSD)	Degrees of freedom (df)	Variance (Var)	F-test variable	Probability or Signifance or Reliability (%)
Explained by linear regression: **79.500** (SSD_1)	**1** (the slope Aq uses 1 more df)	**79.500 / 1** = **79.500** (Var1 = SSD_1 / **1**)	F (**1**, **46**) = **79.500** / **0.428** = **185.635** (Var1 / Var2)	> 99.9 % highly significant
Remaining unexplained: 99.200 - **79.500** = **19.700** (SSD_2 = SSD_0- SSD_1	47 – 1 = **46**	**19.700 / 46** = **0.428** (Var2 = SSD_2 / **46**)		
Extra explained by quadratic regression: **1.276** (SSD_3)	**1** (one more parameter is used: Bq)	**1.276 / 1** = **1.276** (Var3 = SSD_3 / **1**	F(**1**,**45**) = **1.276** / **0.409** = **3.117** (Var3 / Var4)	91.6 % (>90%, <95%) just significant
Remaining unexplained **19.700** – **1.276** = **18.424** (SSD4 = SSD_2- SSD_3)	46 -1 = **45**	**18.424 / 45** = **0.409** (Var4 = SSD4 / **45**)		

Conclusion: The quadratic regression may be applied as its superiority over linear regression is significant, though not highly significant.

The next figure depicts the result of the F-test demonstrated in table 1 using the F-tester mentioned in [Ref. 6]

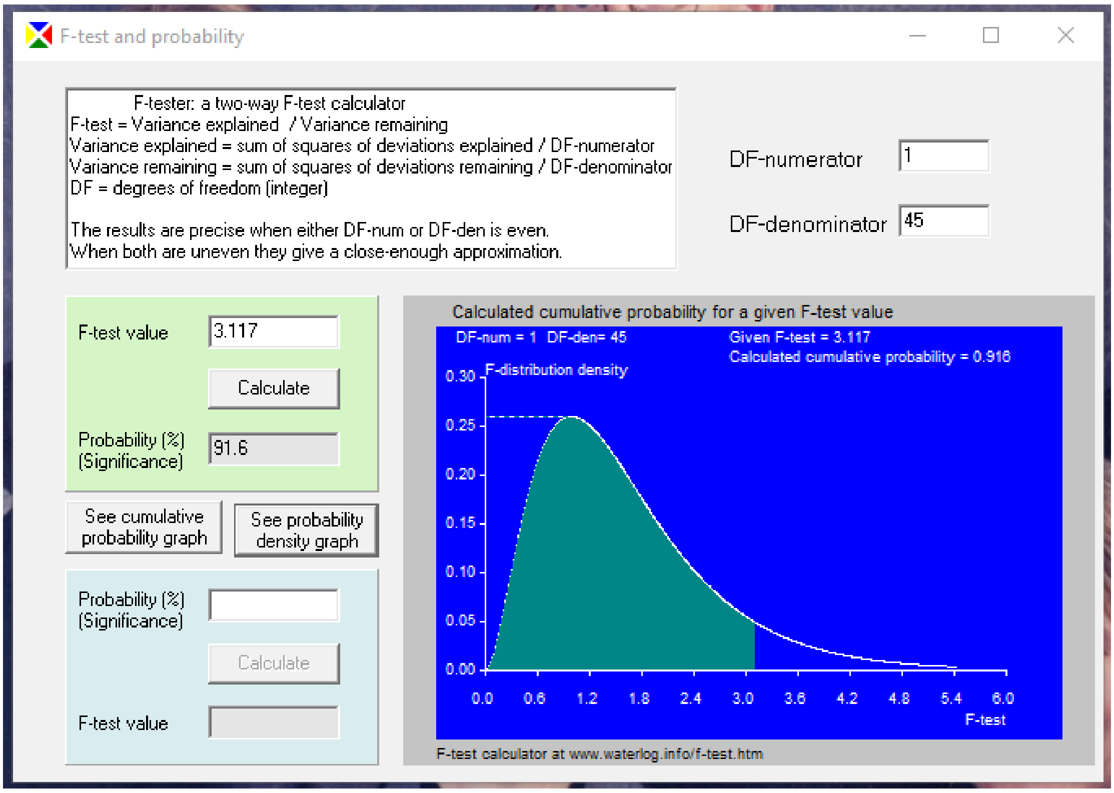

Figure 3. result of the F-test demonstrated in table 1 using the F-tester mentioned in [Ref. 6]

B. Cubic regression

The linear regression equation reads, like for the quadratic case:

$$Y_L = A_L.X + C_L$$

Where

$A_L = -2.91E\text{-}001 \quad C_L = 6.77E+000$

The generalized cubic regression equation reads:

$$\ddot{Y} = Yc = Ac.W^3 + Bc.W^2 + Dc.W + Cc$$

where

$Ac = 5.37E\text{-}001 \quad Bc = -4.70E+000 \quad Dc = 1.12E+001 \quad Cc = -1.84E+000$

$W = X^E$ with $E = 0.49$

Here the X data are raised to the exponent (power) E to obtain an the best possible fit of the cubic curve to the data points. For this reason, the cubic regression is called “generalized”.
In total 5 parameters are used that have to be discounted from the initial number of degrees of freedom.

The general data in the example of potato "927" are as follows:

$\Sigma (Y - \eta)^2 = 99.200$ (total sum of squares of deviations, SSD) Total nr. of data N = 48 Degrees of freedom df = 47 (one df is lost by using the parameter η)

These values are the same as used for the quadratic regression.

For the cubic regression in the example of potato "927", the analysis of variance is shown in Table 2

Table 2. ANOVA analysis for the generalized cubic regression
The corresponding SSD and Var values are printed in the same color while the corresponding df values are shown in bold.

Sum of squares of deviations (SSD)	Degrees of freedom (df)	Variance (Var)	F-test variable	Probability or Signifance or Reliability (%)
Explained by linear regression **79.500** (SSD_1)	**1** (the slope uses 1 df)	**79.500** / **1** = **79.500** (Var1 = SSD1 / **1**)	F (**1, 46**) = **79.500** / **0.428** = 185.635 (Var1 / Var2)	> 99.9 % highly significant
Remaining unexplained 99.200-79.500 = **19.700** (SSD_2 = SSD_0- SSD_1	47 – 1 = **46**	**19.700** / **46** = **0.428** (Var2 = SSD2 / **46**)		
Extra explained by cubic regression = **6.393** (SSD_3)	**3** (three more parameters are used)	**6.393** / **3** = **2.131** (Var3 = SSD3 / **1**	F (**3, 43**) = **2.131** / **0.309** = 6.900 (Var3 / Var4)	> 99.9 highly significant
Remaining unexplained 19.700 – **6.393** = **13.307** (SSD_4)	46 – 3 = **43**	**13.307** / **43** = **0.309** (Var4 = SSD4 / **45**)		
Total explained by cubic regression 99.20 –**13.307** = **85.893** (SSD_5 = SSD_0- SSD_3	**4**	**85.893** / **4** = **21.473** (Var5 = SSD5 / **4**)	F (**4, 43**) = **21.473** / **0.309** = 72.61 (Var5 / Var4)	> 99.9 highly significant

Conclusion: The cubic regression could certainly be applied as its superiority over linear regression is highly significant.

C. Comparing quadratic and cubic regression

Making use of the data featuring in the tables 1 and 2, the following table (table 3) can be prepared.

Table 3. ANOVA analysis for the generalized cubic regression compared with the quadratic regression. The corresponding SSD and Var values are printed in the same color while the corresponding df values are shown in bold.

Sum of squares of deviations (SSD)	Degrees of freedom (df)	Variance (Var)	F-test variable	Probability or Signifance or Reliability (%)
Explained by quadratic regression 80.776				
Remaining unexplained 99.200-80.776 = 18.424	45 = 48-3			
Extra explained By cubic regression **5.117**	**2** (two more parameters are used)	**5.117** / **2** = **2.559** (**Var3**)	F (**2**, **43**) = **2.559** / **0.309** = 8.280 (**Var3** / **Var4**)	> 99.9 highly significant very reliable
Remaining unexplained 18.424 – 5.117 = **13.307**	45 -2 = **43**	**13.307** / **43** = **0.309** (**Var4**)		

Conclusion: The cubic regression, in this case, shows a highly significant improvement over the quadratic regression.

4. Summary

When performing a regression analysis, it is recommendable to test the statistical significance of the result by means of an analysis of variance (ANOVA). The F-test calculator discussed in [Ref. 6] may be helpful in this respect.

In the previous examples referring to the relation between soil salinity and crop yield of the potato variety ”927” it was seen that a linear regression has a statistically highly significant result, but a quadratic regression add a significant improvement while the cubic regression adds a significance that is still considerably higher.

Therefore, in this case, the cubic regression is highly recommendable.

5. References

[Ref 1.] Statistical significance of segmented linear regression with break-point using variance analysis and F-tests. On line: https://www.waterlog.info/pdf/ANOVA.pdf

[Ref. 2] SegReg, free calculator software for segmented regression. Download from: https://www.waterlog.info/segreg.htm

[Ref. 3] SegRegA, free calculator software for segmented and polynomial regression. Download from:
https://www.waterlog.info/segreg.htm

[Ref 4.] The potato variety "927" tested at the Salt Farm Texel,The Netherlands, proved to be highly salt tolerant. On line:
https://www.researchgate.net/publication/335789831_The_potato_variety_927_tested_at_the_Salt_Farm_TexelThe_Netherlands_proved_to_be_highly_salt_tolerant

[Ref. 5] A. de Vos et al. 2016. Crop salt tolerance under controlled field conditions in The Netherlands, based on trials conducted at Salt Farm Texel. On line:
https://library.wur.nl/WebQuery/wurpubs/fulltext/409817

[Ref.6] Free calculatotor software for Fisher's F-test. Download from:
https://www.waterlog.info/f-test.htm

List of publications in which SegReg is used:

https://www.waterlog.info/pdf/segreglist.pdf