

Running a two-way unbalanced ANOVA with interactions

[demoANO2.xls](#)

Dataset for a two-way unbalanced ANOVA with interactions

An Excel sheet with both the data and the results can be downloaded by clicking [here](#).

The data correspond to an experiment in which four different methods for growing crops were tested on four different types of fields (same soil but different light exposure).

The yield was measured after the harvest. Because the 3rd method was not tested on the 4th type of field (because of a lack of seeds), and the 2nd method on the 4th type of field (because of a hail storm), the experiment is a typical example of an unbalanced ANOVA. We have performed an ANOVA with interactions in order to determine the interactions between the types of methods used and the types of fields.

With XLSTAT's ANOVA function, we can find out if the growing method has a significant effect on the yield by controlling the type of field and the interaction between the method and the type of field.

Setting up a two-way unbalanced ANOVA with interactions

After opening XLSTAT, select the **XLSTAT / Modeling data / ANOVA** command, or click on the corresponding button of the "Modeling data" toolbar (see below).

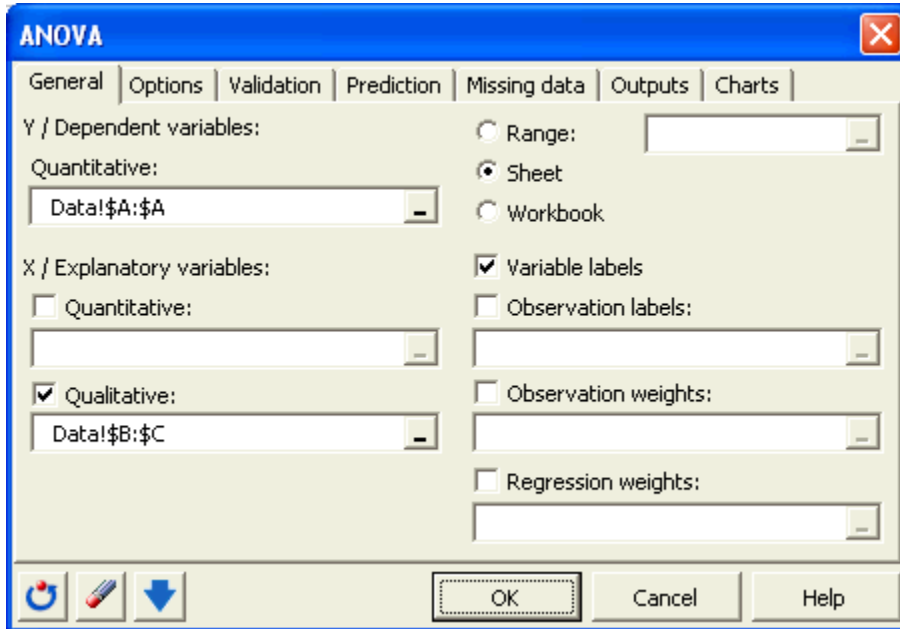


Once you've clicked on the button, the ANOVA dialog box appears.

Select the data on the Excel sheet. The **Dependent variable** (or variable to model) is here the "Yield".

Our aim is to determine the effect of the method, the type of field and the interaction between the two on the variability of the yield.

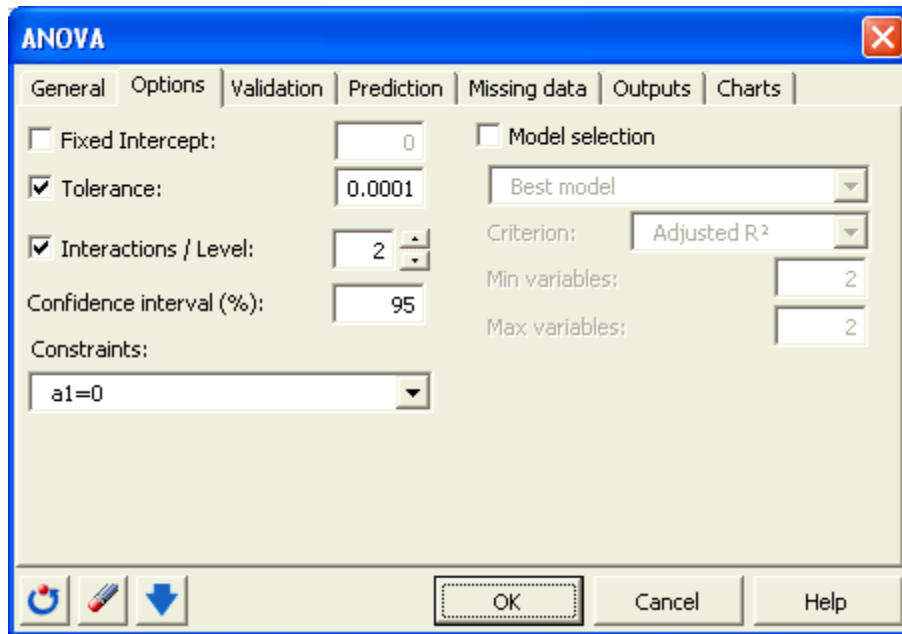
As we selected the column title for the variables, we left the option **Variable labels** activated.



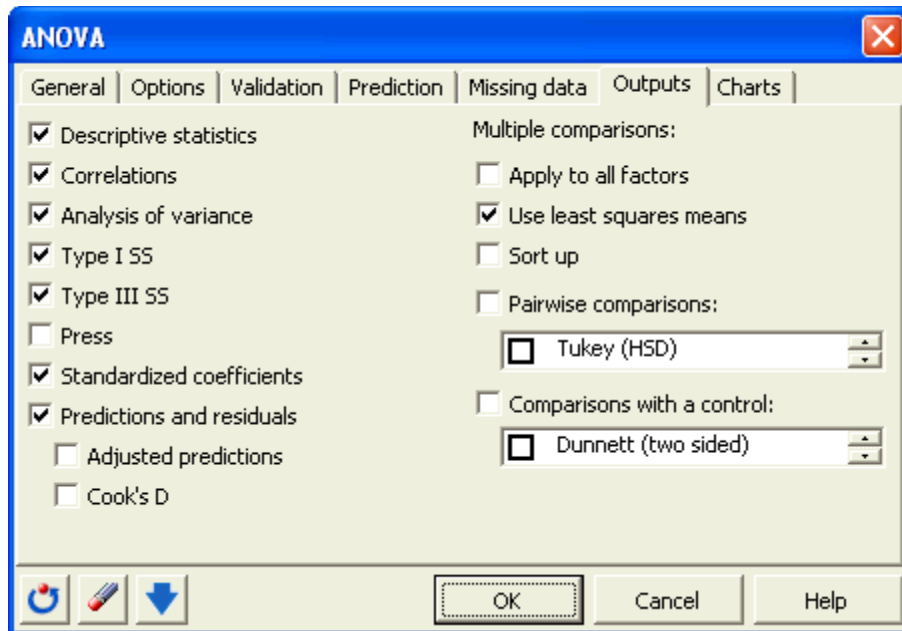
The **interactions** option is activated on the options tab, and the maximum level of interaction is set to 2.

We left the constraint option at **a1=0**, meaning that we want the model to be built on the assumption that the method "1" has the standard effect on the yield.

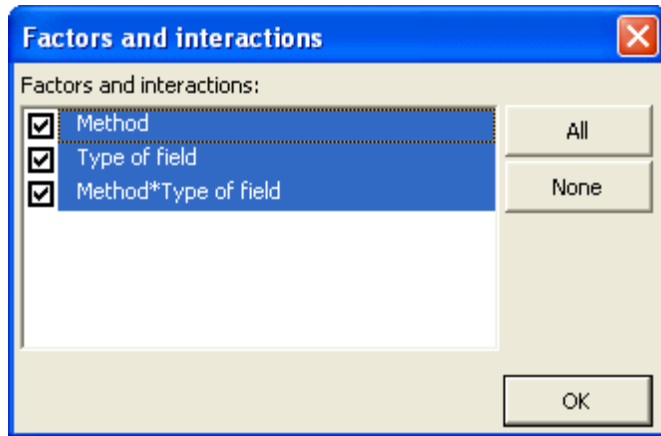
Although you have to apply a constraint to the model in ANOVA for theoretical reasons, it will not affect the results (goodness of fit, predictions). The only difference it makes is in the actual writing of the model.



In the **Outputs** tab, the **Type I SS** and **Type III SS** options were activated because we want the model to take the interactions into account, and because we want to analyze the F values given in the Type I SS, and Type III SS tables (SS stands for sum of squares).



The computations begin once you have clicked on the **OK** button. A dialog box is displayed so that you can confirm which factors have to be taken into account in the model.



The results will then be displayed.

Interpreting the results of a two-way unbalanced ANOVA with interactions

The first results displayed by XLSTAT are the goodness of fit coefficients including the R^2 (coefficient of determination), and the adjusted R^2 .

The coefficient of determination (0.92) gives us a fair idea of the extent to which the variability of the modeled variable (the yield) can be explained by the explanatory variables (the method, the type of field, and their interaction). In this particular example, 92% of the variability is explained. The remaining 8 percent are hidden in other variables, which the model classifies as "random effects."

Goodness of fit statistics:	
Observatio	21,000
Sum of we	21,000
DF	10,000
R^2	0,918
Adjusted F	0,837
MSE	111,300
RMSE	10,550
MAPE	27,020
DW	3,235
Cp	11,000
AIC	105,376
SBC	116,866
PC	0,261

It is important to examine the results of the analysis of variance table (see below). The results enable us to determine whether or not the explanatory variables bring significant information (null hypothesis H_0) to the model. In other words, it's a way of asking yourself whether it is valid

to use the mean to describe the whole population, or whether the information brought by the explanatory variables is of value or not.

Analysis of variance:					
Source	DF	Sum of squares	Mean square	F	Pr > F
Model	10	12515,667	1251,567	11,245	0,0003
Error	10	1113,000	111,300		
Corrected	20	13628,667			
<i>Computed against model Y=Mean(Y)</i>					

Given that the probability corresponding to the Fisher's F is lower than 0.0003, it means that we would be taking a 0.03% risk in assuming that the null hypothesis (no effect of the two explanatory variables and their interaction) is wrong.

Therefore, we can conclude with confidence that the two variables and their interaction do have a significant effect. We also want to find out if the two variables, and their interaction, provide the same amount of information. To do this, we have to examine the Type I SS and Type III SS tables.

Type I Sum of Squares analysis:					
Source	DF	Sum of squares	Mean square	F	Pr > F
Method	2	9499,619	4749,810	42,676	< 0,0001
Type of field	3	1032,411	344,137	3,092	0,076
Method*Ty	5	1983,636	396,727	3,564	0,041
Type III Sum of Squares analysis:					
Source	DF	Sum of squares	Mean square	F	Pr > F
Method	2	8969,889	4484,944	40,296	< 0,0001
Type of field	3	1042,515	347,505	3,122	0,075
Method*Ty	5	1983,636	396,727	3,564	0,041

The Type I SS table is constructed by adding variables in the model one by one, and by evaluating the impact of each on the model sum of squares (Model SS). In consequence, in Type I SS, the order in which the variables are selected will influence the results.

The Type III SS table is computed by removing one variable of the model at a time to evaluate its impact on the quality of the model. This means that the order in which the variables are selected will not have any effect on the values in the Type III SS. The Type III SS is generally the best method to use to interpret results when an interaction is part of the model.

Note: the higher the Model SS, the lower the Residual SS, and therefore the greater the influence of the variable.

From the results displayed in the Type III SS table, we can see that the "Method" variable is the one that has the highest impact on the model.

When we look at the model parameters (see below), we can see that methods 2 and 3 have a positive impact on the yield. The "Type of field" has a low effect on the yield, but the influence of the interaction between the type of field and the method should not be overlooked (the confidence range is 95 percent, meaning there is a 5 percent risk factor).

Model parameters:						
Source	Value	Standard error	t	Pr > t	Lower bound (95%)	Upper bound (95%)
Intercept	13,500	7,460	1,810	0,100	-3,122	30,122
Method-1	0,000	0,000				
Method-2	30,000	10,550	2,844	0,017	6,493	53,507
Method-3	39,500	10,550	3,744	0,004	15,993	63,007
Type of field	0,000	0,000				
Type of field	14,500	10,550	1,374	0,199	-9,007	38,007
Type of field	10,000	10,550	0,948	0,366	-13,507	33,507
Type of field	-4,500	10,550	-0,427	0,679	-28,007	19,007
Method-1*	0,000	0,000				
Method-1*	0,000	0,000				
Method-1*	0,000	0,000				
Method-1*	0,000	0,000				
Method-2*	0,000	0,000				
Method-2*	-29,000	14,920	-1,944	0,081	-62,243	4,243
Method-2*	12,000	14,920	0,804	0,440	-21,243	45,243
Method-2*	25,000	16,681	1,499	0,165	-12,167	62,167
Method-3*	0,000	0,000				
Method-3*	14,000	14,920	0,938	0,370	-19,243	47,243
Method-3*	13,500	14,920	0,905	0,387	-19,743	46,743
Method-3*	0,000	0,000				

The table depicted above can be used to analyze the impact of the explanatory variables on the yield and/or to predict the average yield in a situation not yet covered by the experiment, such as the 3rd method and 4th type of field. In this particular example, the average yield would be 48.5, given the fact that the influence of the interaction is unknown.

We can also look at the standardized residuals. These are residuals that, given the assumptions of the ANOVA model, should be normally distributed; i.e., 95 percent of the residuals should be in the interval [-1.96, 1.96]. All values outside this interval are potential outliers, or might suggest that the normality assumption is wrong. It appears here that there is no outlier, as all values are in the one [-1.96, 1.96] range.

