Generalized analysis of variance

Menu:	QC.Expert	Anova	Multi-factor

Generalized analysis of variance (GANOVA) helps to assess extent to which a selected numeric response variable is influenced by (a) qualitative factors (as name of operator, work shift, number of production line, raw material supplier, etc.) and/or (b) quantitative numeric variables (as temperature, pressure, mixer revolutions). The response variable of interest may typically be a quality parameter or any process or experimental output. As such, this module is useful for spotting how to influence or stabilize an important output parameter, to identify and prove influence of certain factors on important output and consequently eliminate or stabilize the influential factors to stabilize or improve quality.

This module is a generalization of a linear regression model with so called dummy binary variables and with use of generalized Moore-Penrose pseudoinversion of the characteristic matrix. Linear regression model also allows to predict the response for a given values of predictors (a) and (b). The module analyses the influence of fixed-effect factors and continuous variables on the response. Observations Z_i at n_j different levels of predictor factor X_j and various values of the predictor variable Y_k can be described by linear regression model with unknown parameters

$$Z = \alpha_0 + \sum_j \boldsymbol{\alpha}_j X_j + \sum_k \beta_k Y_k + \varepsilon\,,$$

where α_0 is the absolute term (overall mean value), $\boldsymbol{\alpha}_j$ is an ($n_j \times 1$) vector of latent parameters for j-th factor and β_k is the regression coefficient for k-th variable. Random error ε_{ij} is assumed to have normal distribution with zero mean, $\varepsilon \sim N(0,\sigma^2)$. Latent parameters for factors have no direct meaning, but they can be used to test statistical significance of the factor using an F-test. The module *Anova – multi factor* computes $\mathbf{a}_j$, b_k, e_i, which are best estimates of the coefficients $\boldsymbol{\alpha}$, β, and the measurement errors ε.

Data and parameters

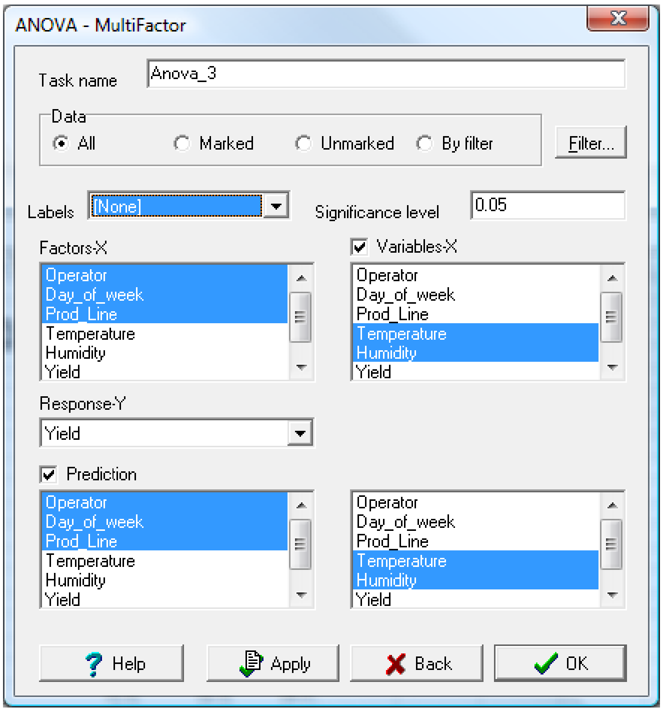
Data are organized in columns, in any order. Each column corresponds to one numerical predictor or one factor or the numerical response value. Data must contain at least one factor column and one column of the response variable. Factor columns are not numerical, factor levels are defined as texts like YES, NO, or A, B, C, D, etc. Number of different text strings define the number of levels of the factor. Each factor must have at least two levels. Variables are numerical values. Value in the response column must correspond to the combination of predictors in the respective row.

Table: Sample input. Predictors are two factors (production line and operator) and one numerical variable (temperature). For every combination of predictors there is one observed value of the response (yield).

Predictors			
Factors		*Variable*	*Response*
Line	**Operator**	**Temp**	**Yield**
A	Brown	13.3	14.6
B	Smith	16.3	17.4
C	Brown	18.7	13.3
A	Mitchel	14.5	12.6
B	Mitchel	11.1	17.5
C	Smith	16.0	14.9
.....	.....	.....	.....

If the data contain only numerical predictors, it is necessary to use linear regression module. If there is only one or two factors, it is more suitable to use the respective one- or two-factor Anova module.

In the „Anova-Multifactor“ dialog panel (see Fig. 1), select the factors and variables, select the response variable column. If there is no variable predictor, uncheck the checkbox *Variables-X*. In the field *Response-Y* select one column with the response values. Select the significance level (typically 0.05, or 5%). If you want to calculate predicted values, select columns for prediction in the field *Prediction* and check the *Prediction* checkbox. The prediction columns must have the same structure as the predictors. The predictors themselves may be selected for calculating prediction. After clicking on *OK*, the results are written in the *Protocol* and *Graphs* windows.



ANOVA - MultiFactor

Task name: Anova_3

Data: ● All ○ Marked ○ Unmarked ○ By filter

Labels: [None]

Significance level: 0.05

Factors-X: Operator, Day_of_week, Prod_Line (selected); Temperature, Humidity, Yield

☑ Variables-X: Operator, Day_of_week, Prod_Line; Temperature, Humidity (selected); Yield

Response-Y: Yield

☑ Prediction: Operator, Day_of_week, Prod_Line (selected); Temperature, Humidity, Yield | Operator, Day_of_week, Prod_Line; Temperature, Humidity (selected); Yield

Fig. 1 Anova-Multifactor dialog panel

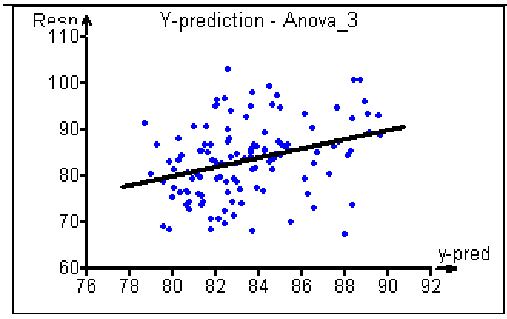
Protocol

Anova - Multifactor	
No of cases	Total number of cases, or rows in the data table.
Total no of predictors	Number of factors and variables.
No of factors	Number of factors.
No of variables	Number of variables.
Mean Y	Average of all responses.
Absolute term	Predicted response value with no influence of factors and zero value of all variable predictors.
Significance level	Chosen value of significance for statistical tests. Recommended value is 0.05.
No of levels	Numbers of levels for all factors in the model.
Overall ANOVA	Overall test if the predictors have any influence at the response.
Source	Source of the variability is assessed. If the model explained satisfactory portion of the response variability, then it may be assumed significant. Primary variability measure is the sum of squared residuals.

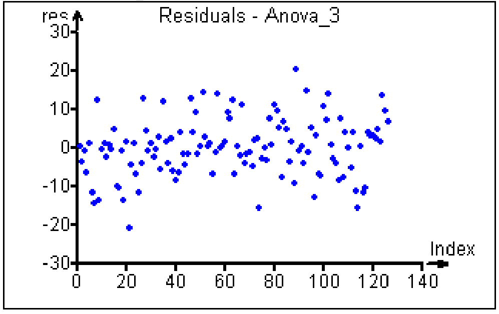
Degrees of freedom	Degrees of freedom for every factor or variable.
Sum of squares	Variability expressed as sum of squares.
Variance	Variability expressed as variance.
F-statistic	The ratio of the variance without and with the model.
p-value	If the p-value is less then the chosen significance value, then the factor is statistically significant.
Significance	Verbal result of the significance test.
Source	Variability sources
Total variability	Values for the total variability in terms of sum of squares $TSS = \sum_{i=1}^{n} \left(Z_i - \bar{Z} \right)^2$
Explained variability	*TSS – RSS*
Residual variability	Residual variability, [data] – [model] in terms of sum of squares, or $RSS = \sum_{i=1}^{n} \left[Z_i - \left(a_0 + \sum_j \mathbf{a}_j X_{ij} + \sum_k \mathbf{b}_k Y_{ik} \right) \right]$ *Note*: TSS stands for total sum of squares, RSS is residual sum of squares
ANOVA for individual factors	Amount of variance explained by the predictors contained in the model, factors and variables.
Predictor	Name of factor or variable.
Parameter	Estimated value of the variable coefficient b_k, For factors this field remains empty.
Sum of squares	Sum of squares explained by this predictor.
F-statistic	Corresponding F- quantile.
p-value	*p*-value. If *p*-value is less then the required significance level (usually 0.05) this predictor is significant (it significantly influences the value of response).
Significance	Verbal result of the test: Significant or Insignificant.
Prediction	When the checkbox *Prediction*, was checked, a table of predicted values of response is included.
Prediction	Calculated (predicted) values of the response

Graphs

Y-Prediction plot	Overall fit plot. Measured values (Resp) are plotted against the predicted values (y-pred). When the points lie close to the line y=x the prediction is successful and the model describes the data well. Here, the model is assessed as a whole, separate factors and variables are assessed in the partial prediction plots, see below. This plot corresponds to the overall explained variability of the model

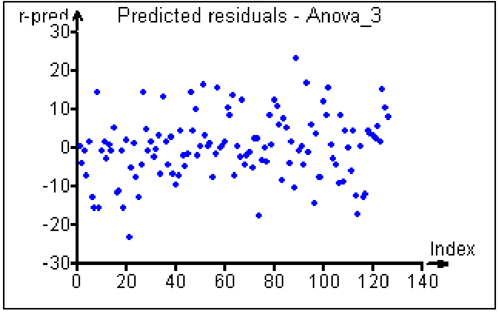


Residuals plot



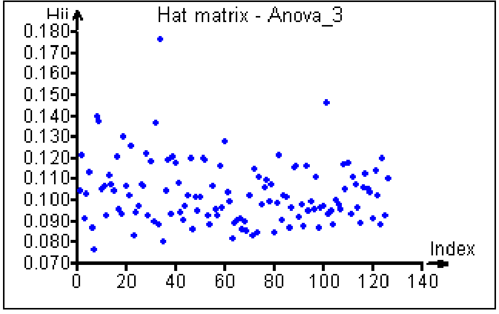
In the plot of residuals, the distances of the response from the model (or residuals) are plotted. Points that are far from the horizontal line are suspected outliers, possibly errors in the measured response of non-typical measurement.

Predicted residuals plot



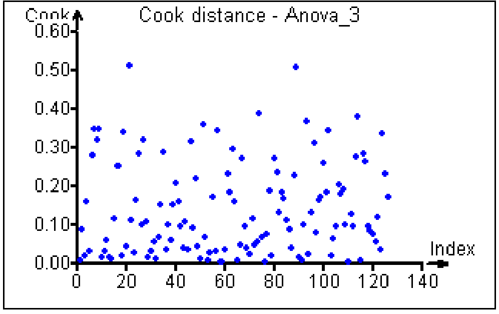
Predicted residuals are similar to the residuals in the previous plot. Here, each i-th residual value is computed from data with dropped i-th measurement, so the possible outliers are usually more visible. Points far from the line y=0 are suspected outliers, or gross errors.

Projection Hat-matrix plot



Plot of the diagonal elements of the projection matrix (or so-called hat matrix) $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Points with the high values are usually untypical in the predictor values (such as very high or low variable values or non-typical combination of factor levels). Such points are highly influential and should be paid special attention, as wrong response values could result in biased or unreliable estimates of the model parameters. On the other hand however precisely measured response in such points will significantly improve statistical properties of the model.

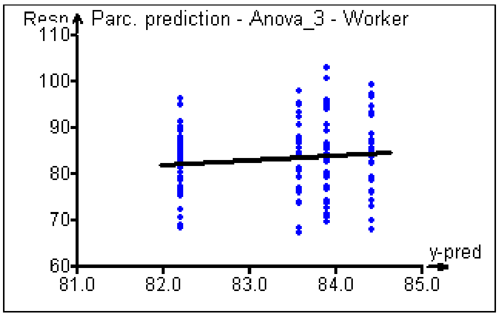
Cook distance plot



Value of Cook distances is another measure of influence of the measurements. Interpretation is similar as in the previous plot.

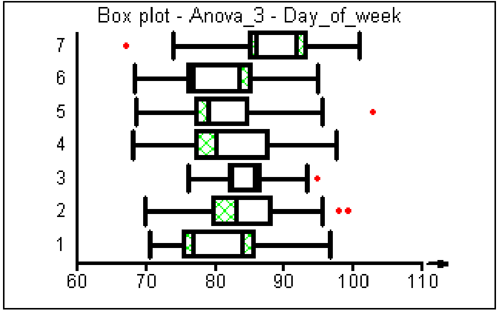
Partial prediction plot for a factor

Partial prediction plot shows how separate factor levels influence the response. Strong steep dependence suggests



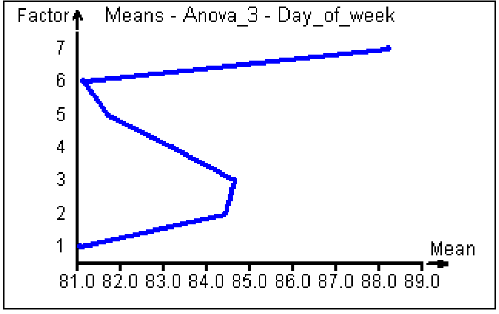
that the respective factor has strong influence on the response. This plot corresponds to explained variability for the factor and to test of significance of the factor.

Box plot for a factor



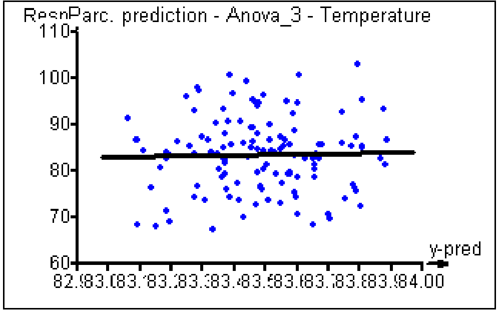
Box plot shows median and quantile characteristics of the response for every level of a factor. Limits of the box are the lower and upper quartiles (25% and 75% quantiles). Separate red points are suspected response outliers The morphology of the box plot is further described in Basic statistics.

Averages plot



This is another representation of means of response for every level of the factor. The values on the y-axis are averages of the response for the given factor level.

Partial prediction plot for a variable



This plot shows how much the prediction depends on a given variable. Interpretation of this plot is analogical to the partial plot for a factor.