Linear regression

Menu:	QCExpert	Linear regression

Linear regression module is used to build and analyze linear regression models in their general form

$$G(y) = a_1F_1(\mathbf{x}) + a_2F_2(\mathbf{x}) + \ldots + a_mF_m(\mathbf{x}) + a_0, \qquad (1\text{-}1)$$

where y is a response variable, $\mathbf{x} = (x_1, x_2, \ldots x_p)$ are values of explanatory variables (written as a vector). p is the number of explanatory variables in the regression model. There are m parameters, $\mathbf{a} = (a_1, a_2, \ldots, a_m)$ in the model. a_0 is the intercept. $F_i(.)$, i=1,…m are arbitrary functions of explanatory variables which do not involve parameters. $G(.)$ is an arbitrary function of the response variable which does not involve parameters. Individual summands $F_i(\mathbf{x})$ on the right hand side of the model equation are sometimes called model terms. Ideally, **x** is assumed to be a deterministic, i.e. non-random vector, being either purportedly set to pre-specified values or its values are found out via an essentially error-free procedure. y depends on **x,** but the dependence is blurred by the presence of a random error ε. Vector of model parameters **a** can be estimated from data by various methods. Some methods are robust, some of them might not be. The (data, model, method) triplet is sometimes called the regression triplet. In order to get correct results, each of the triplet components should be given appropriate attention. Regression diagnostics and other tools offered by QC.Expert™ are useful in this context. There is also a wide choice of models available in the program. The user can select one of the three basic model types: simple linear model without transformation, polynomial model or general user-defined model. The selection takes place in the Linear regression dialog panel, particularly in the Transformation field:

No transformation: corresponds to a regression model of the form

$$y = a_1x_1 + a_2x_2 + \ldots + a_mx_m + a_0, \qquad (1\text{-}2)$$

For this model, the number of parameters, m is specified by the number of explanatory variables selected in the *Explanatory variable* window. The simplest example of such a model is a regression line, e.g.
[profit] = a_1 . [investment] + a_0,
another example, involving several explanatory variables is
[steel_strength] = a_1 . [Cr_concentration] + a_2 . [melting_time] + a_3 . [carbon_concentration] + a_0,

For instance:

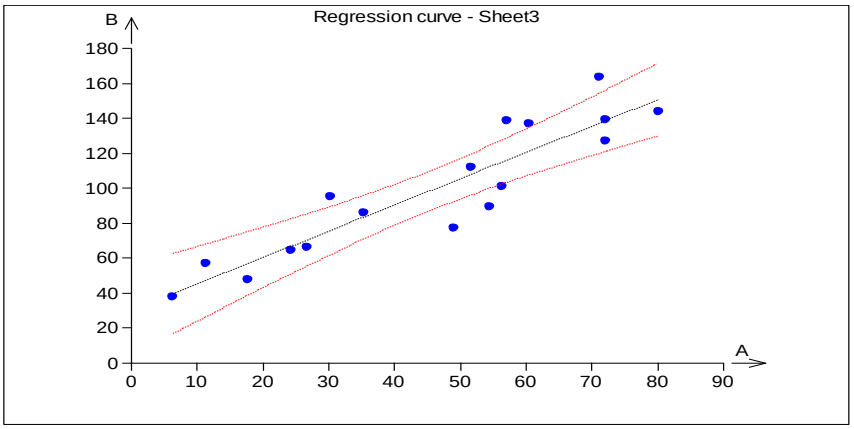


Fig. 1 Regression line

Polynomial is a model of the following form:

$$y = a_1 x + a_2 x^2 + \ldots + a_m x^m + a_0, \qquad (1\text{-}3)$$

m is degree of the polynomial. It is equal to the number of model parameters minus one. There is only one independent variable *x* in such a polynomial. All of its powers 1 through *m* appear in the model, however. Following quadratic (i.e. nonlinear) relationship can serve as an example.

[number_of_items_sold] = a_0 + a_1 . [advertisement_costs] + a_2 . [advertisement_costs] ^{2} + a_0.

When a model involving only some powers is desired instead of the full polynomial (e.g. a model with the first and third powers only), user transformation has to be used.

For instance:

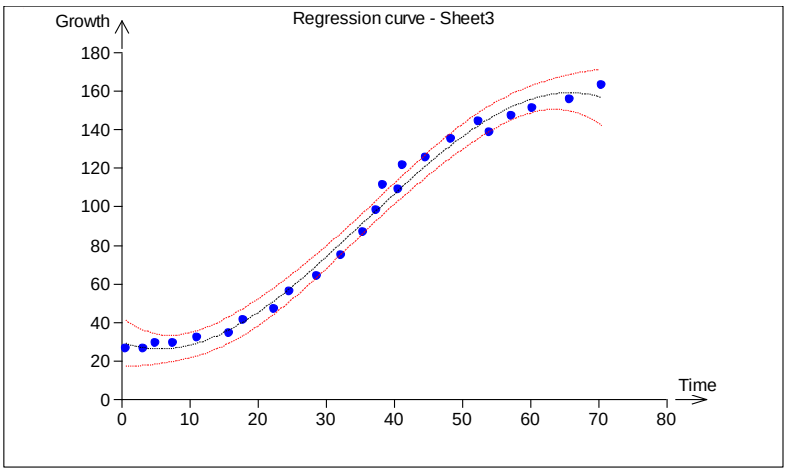


Fig. 2 Polynomial of the 3rd order

QC.Expert™ also allows polynomial transformation of several explanatory variables. In more than one independent variables are selected, polynomial transformation will allow for the full 2nd degree Taylor series, which is often used to fit and optimize response surfaces. Results n the protocol then include type of the stationary point (possibly optimum) and parameters of the regression model. Details are described below in paragraph 0.

User transformation: allows you to specify a linear model. It is a general formulation which includes the two special cases discussed previously (*without transformation* and *polynomial*). Earlier defined models can be selected using the appropriate selection window. A new model is specified upon choosing *Model…* after clicking the *User…* button. This action opens model specification dialog panel, see later). Individual transformation functions F_1, F_2, …, and/or *G*, see below can be specified there. User transformation can be used when linearizing the exponential model $y = A$. exp($B\,x$) to the the form ln (y) = $a + b\,x$, where a = ln A, $b = B$. G = ln(y), $F_1 = x$, in this case, . Another example is

1 / [consumption] = a_1 [X1] + a_2.[X1]$^{1/2}$ + a_3 [X1].[X2] + a_4 ln[X2] + a_0,

where
G = 1/[consumption],
F_1 = [X1],
F_2 = sqrt[X2],
F_3 = [X1][X2],
F_4 = ln [X2],
[consumption] is the response variable, [X1] and [X2] are explanatory variables.

It is important to keep in mind that the transformations can involve only explanatory/response variables, they must not involve parameters a_i. Models like $y = a_1 \,.\, x^{a2} + a_0$, or $y = a_0 + a_1 \,.\, \exp(a_2\, x)$ cannot be specified through such transformations.

One of various (robust or classical) estimation method can be selected. This should be done in accord with the general character of the data, error behavior, or other considerations. Individual data points can be weighted differently upon inputting user-supplied weights. QC.Expert™ allows you to inspect various potential models corresponding to all combinations of model terms by fitting all possible regression subsets. This can help you to find the most important variables, which should be included in the model, and/or their transformations.

Data and parameters

Unknown parameters $\mathbf{a} = (a_1, a_2, \ldots, a_m)$ are estimated from data in the current data sheet. Each of its columns corresponds to a variable. Column header contains a variable name. The way, in which variables are selected depends on a requested transformation.

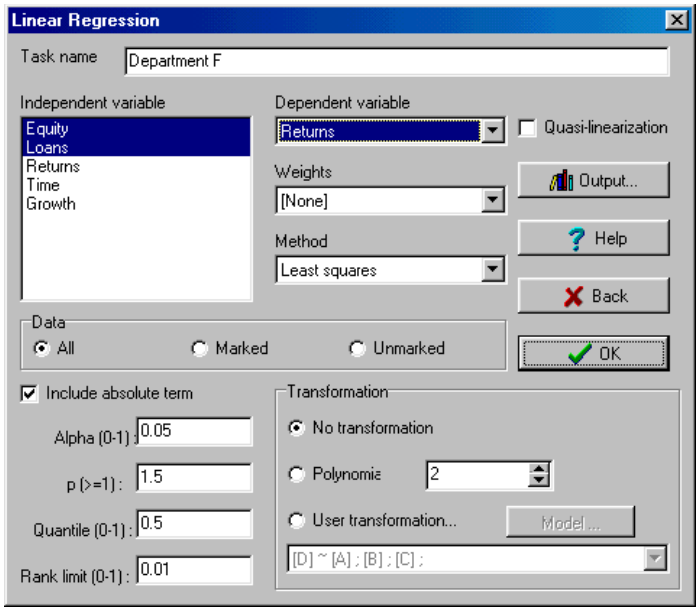


Fig. 3 Linear regression dialog panel

No transformation: when the linear regression dialog panel is open for the first time, the last column is automatically selected as the response variable column, any other data columns are selected as explanatory variables implicitly. Other choices can be made using mouse, Shift and Ctrl keys. Number of explanatory variables is not restricted at all. Exactly one response variable has to be selected. Any model specified in this way has a general form (2). The dependent variable column corresponds to y and the explanatory variables data columns correspond to x_1, x_2, …

Polynomial: requested polynomial transformation (3) amounts to fitting a polynomial curve through data. Such a choice involves only one response and one explanatory variable. Requested polynomial order p has to be specified. Lower order polynomials are strongly preferred in general. Higher order polynomial models can be numerically unstable. Their statistical properties might be bad as well, resulting into high variability of parameter estimates and poor prediction abilities. Choice of the polynomial degree might be guided by the APSR (all possible subsets regression) results - see later. When Polynomial transformation is selected, all powers 1 through p are forced into the model. When only some of the powers are to be included in the model (e.g. 1^{st} , 3^{rd}, 5^{th}), User transformation has to be selected, where each of the model terms is specified explicitly.

If more than one explanatory variable is selected, then the *Polynomial order* field stays inactive and a full quadratic model is constructed automatically. The model includes all pure terms terms of up to second order and all cross products of linear terms, so that its terms involve

$$x_1, x_2, \ldots, x_m, x_1x_2, x_1x_3, \ldots x_ix_j, \ldots, x_mx_{m-1}, x_1^2, x_2^2, \ldots, x_m^2 \quad (1\text{-}4)$$

When our variables are named A, B, C, the full quadratic model with intercept will be of the form

$y = a_0+a_1.A+a_2.B+ a_3.C+ a_4.AB+ a_5.AC+ a_6.BC+ a_7.A^2+ a_8.B^2+ a_9.C^2$

This model corresponds to an *m* dimensional quadratic surface. Such a surface can have one extreme point (minimum or maximum), corresponding to the minimum or maximum expected response. The model can be fitted in the Response surface module as well, although the output is much less detailed there (e.g. without diagnostics). One has to keep in mind that the number of data points should be larger (ideally much larger) than the number of model terms. With *m* different explanatory variables, the full quadratic model has $1.5m+m^2/2+1$ model terms. For instance, $m=10$ gives 66. Detailed description of the quadratic model output can be found in the *Response surface methodology* chapter.

User transformation: after selecting User transformation, the model specification panel is opened upon clicking the *Model…* button. If any models were specified previously, one of them can be selected without going through the *Model specification* panel. In such case, it is necessary to make sure that variable names in the model and in current data sheet agree. A new model is formulated using the *Model specification* panel. The current data sheet variables list is displayed in the left part of the *Model specification* panel. Only the listed variables can be used when specifying a regression model. Model input line appears in the upper right hand corner of the panel. The window located just below the input line lists dependent variable together with model terms included already.

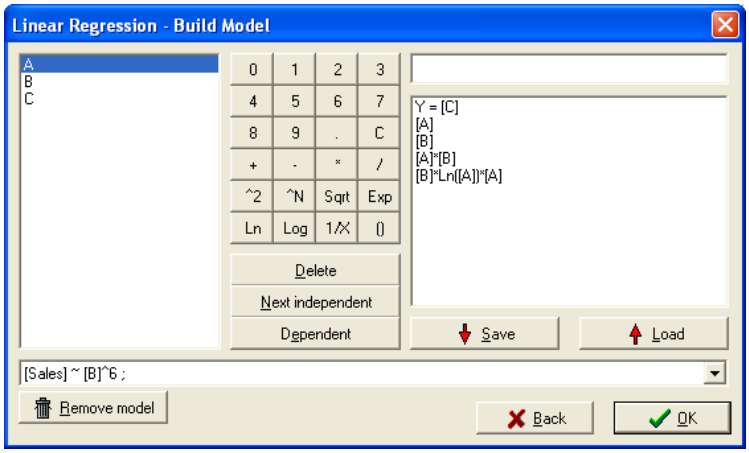


Fig. 4 Model specification dialog panel

Any function of one or more explanatory variables can be a model term. There has to be only one response variable. It is either directly the variable selected by the user from the current data sheet variables or any function of it. The response variable is denoted as Y=. For instance, let us define the model $\ln(y) = A.x + B.x^2 + C.x^{-1}$ model. The intercept can be included in the model either by checking the *Intercept*, or by including 1 (number one) as an explanatory variable during the model specification. Only one of the two possibilities should be used. (When both of them are used, an error caused by model over specification results.)

Model specification instructions:
Double click on a variable name in the available variables list copies the name to the model input line. Variable name is always enclosed in square brackets. Function buttons can be useful when specifying more complicated models. Highlight a part of the model input line and clicking a function button subsequently to apply the function on the highlighted part as an argument. For instance, expression

ln([*x*]+1) can be assembled in the following way: double click on the variable *x* (there has to be a column of this name in the current data sheet): **[x]**; write + 1 manually; highlight whole expression: **[x]+1**; click the *Ln* button, resulting into: **ln([x]+1)**. Application of *^2, ^A, Sqrt, Exp, Log, 1/X, ( )* is similar. The *C* button erases model input line. Other functions have to be inputted manually (writing their name in the model input line). Available functions are listed in the table below. After specifying a term completely, another term is added by clicking on the *Next explanatory* button. The response variable is included by clicking on the *Response* button. A highlighted model term is erased when clicking the *Erase* button. There has to be exactly one response variable in any model. When finished with specification, the model is saved by the *Save* button. Then, it automatically appears in the list of previously specified models located in the bottom part of the panel. The *Read* button reads in a model from the set of previously defined models. Its terms can be edited subsequently. The *Erase model* deletes a selected model from the model list. Warning: the operation is irreversible! *OK* button finishes model specification.

You can select previously specified models from the list directly in the *Linear regression* dialog panel without opening the *Model specification* panel, be careful however: variable names in the model and in the current data sheet have to agree.

Table 1 List of available functions

Function	Value, description, restrictions	Syntax
Basic binary operators		
+	Summation	x+y
–	Subtraction	x−y
*	Multiplication	x*y
/	Division; $y \neq 0$	x/y
^	Power; for a negative x, the INTPOWER function has to be used	x^y
DIV	Integer divisor; $y \neq 0$	x DIV y
MOD	Modulo; $y \neq 0$	x MOD y
Functions		
TAN	Tan; $x \neq n\pi+\pi/2$	tan(x)
SIN	Sine	sin(x)
COS	Cosine	cos(x)
SINH	Hyperbolic sine	sinh(x)
COSH	Hyperbolic cosine	cosh(x)
ARCTAN	Arc tan	arctan(x)
COTAN	Cotan; $x \neq n\pi$	cotan(x)
EXP	Exponential function, base e	exp(x)
LN	Natural logarithm; $x > 0$	ln(x)
LOG	Decadic logarithm; $x > 0$	log(x)
LOG2	Base 2 logarithm; $x > 0$	log2(x)
SQR	Square	Sqr(x)
SQRT	Square root; $x \geq 0$	Sqrt(x)
ABS	Absolute value (abs(0) = 0)	Abs(x)
TRUNC	Truncation	Trunc(x)
INT	Truncation	int(x)
CEIL	Ceiling	Ceil(x)
FLOOR	Floor	Floor(x)
HEAV	Heaviside function (indicator of a nonnegative argument, 0	Heav(x)

	for a negative argument, 1 else)	
SIGN	Sign (-1 for a negative argument, 0 for 0, 1 for a positive argument)	Sign(x)
ZERO	Indicator of zero (1 for zero argument, 0 else)	Zero(x)
RND	Random number from a uniform distribution on (0,x); x > 0	Rnd(100)
RANDOM	Random number from (0,1) uniform distribution. Even though it does not use any argument, a dummy argument has to be specified.	Random(0)
–	(Unary) minus before an expression	–x

Functions with two arguments		
MAX	Maximum	MAX(x,y)
MIN	Minimum	MIN(x,0)
INTPOWER	The first argument raised to the power specified by the second argument, the second argument is integer valued; it can be used even for a negative x	INTPOWER(x, –2)
LOGN	Logarithm of the first argument, using the second argument as base; x > 0, y > 1	Logn(x,3)

Relations		
GT	Greater than; if x>y then it returns 1, 0 else	GT(x,y)
LT	Less than; if x<y then it returns 1, 0 else	LT(x,y)
EQ	Equal; if x = y then it returns 1, 0 else	EQ(x,y)
NE	Not equal; if x ≠ y then it returns 1, 0 else	NE(x,y)
GE	Greater or equal; if x ≥ y 1, 0 else	GE(x,y)
LE	Less or equal; if x ≤ y 1, 0 else	LE(x,y)

Function names can be written in lowercase or uppercase letters. Relations result in 0 or 1, which can be used when specifying discontinuous functions, like le(x,0)*1+gt(x,0)*5, see also the Nonlinear regression chapter.

Further details on the Linear regression dialog panel.

Task name: A project identification (one line). It appears in the protocol and graphic output headers.

Independent variable: Select one or more explanatory variables. Use mouse (dragging, Shift-click or Ctrl-click) when selecting more than one variable. This item is not active when User transformation is selected – the variables are specified in the Model specification panel.

Dependent variable: Select one data column as a response variable. This item is not active when User transformation is selected – the variables are specified in the Model specification panel.

Intercept: When checking this option, intercept is included in the model. Do not use it when the intercept is already entered manually as the unit explanatory variable!

Alpha (0 – 1): Significance level, α which will be used for all tests and confidence intervals. It has to be larger than 0 and smaller than 1. α=0.05 is the default.

p (p ≥ 1): Coefficient *p* for L_p regression. The value is used only when the *Lp-regression* is selected (see later). *p*=1 corresponds to the least absolute differences method, *p*=2 corresponds to the least

squares, $p \to \infty$ ($p \approx 10$ is typically taken in practice) corresponds to minimization of maximum error (minimax). When $1 \le p < 2$ is selected, the resulting estimates are rather robust against outliers. $p=1.5$ is the default.

Quantile (0 – 1): Probability value specifying a particular quantile regression. It is used only when the *Quantile regression* is selected (see later). It has to be larger than 0 and smaller than 1. 0.5 is the default, corresponding to the least absolute differences method.

Rank limit (0 – 1): It is a restriction parameter related to the Rank correction method. Zero parameter value corresponds to the usual method of least squares (OLS). When a positive parameter is selected, the components related to small eigenvalues of the $\mathbf{X}^T\mathbf{X}$ matrix are suppressed, resulting into biased parameter estimates with smaller variance than usual estimates. Such estimates are less sensitive to an ill conditioned $\mathbf{X}^T\mathbf{X}$ matrix, which occurs typically e.g. when fitting a high degree polynomial models (see later). Value of at most 0.1 is recommended.

Quasi-linearization: When this selection is checked, quazilinearization is applied. It is useful when User transformation is selected and the response variable is nonlinearly related to one of the explanatory variables. This occurs for instance for the model **`ln(y) ~ [x]; [x]^2`**. Nonlinear transformation $G(y)$ linearizes the model, but it deforms error distribution and biases parameter estimates. The quasilinearization technique can eliminate the bias, to some extent. The quasilinearization is based on the idea of introducing weights $w_i=[\partial G(y)/\partial y]^{-1}$.

Weights: Select a data column w_i, you want to serve as a weighting variable. Alternatively, you van select one of the pre-specified weight types: [None], [Y], or [1/Y]. The [1/Y] weights are used when the relative error for the response is constant. The weights must not be negative. Zero weight results in dropping the corresponding line from the analysis. The default is [None] – all weights are equal to one. When variances of the response in different data rows are known, say $\mathbf{S} = \mathrm{diag}(w_1^{-2}, w_2^{-2}, \ldots, w_n^{-2})$ than the weights should be proportional to the square roots of reciprocal variances.

Method: Select one of computational methods. The selection should depend on the nature of the analyzed data.

Least squares: The basic and commonly used method. It works fine when errors are normally distributed, data are free from gross errors in both response and explanatory variables and the problem is not ill conditioned due to an unfavorable design matrix composition. The method may fail badly when some of these conditions are not satisfied.

Rational rank: A method commonly used for instance for higher order polynomials, full second order polynomials and other cases when collinearity is a problem (explanatory variables are "correlated"). Detected collinearity is indicated in the QC.Expert™ protocol (in the *Multicollinearity paragraph*). The extent to which the rank is corrected is given by the *Restriction* parameter (value of at most 0.1 is recommended). When a positive parameter is selected, the components related to small eigenvalues of the $\mathbf{X}^T\mathbf{X}$ matrix are suppressed, resulting into biased parameter estimates with smaller variance than usual estimates. Such estimates are less sensitive to an ill-conditioned $\mathbf{X}^T\mathbf{X}$ matrix.

Quantile regression: Quantile regression method, using the quantile α specified in the *Quantile* field. It corresponds to the model in which probability of the event (linear predictor<Y) is α. The method is advantageous when we are not interested in modeling changes in expected value as a function of explanatory variables, but rather in modeling changes of a more extreme tendency of the distribution, specified by a quantile. For instance, one might be interested in "minimal" strength and choose $\alpha=0.05$, or in "maximal" pollution and choose $\alpha=0.95$, etc. The computation method is iterative (weighted least squares method is used iteratively). Computation time depends on the number of data

points. Number of data points (*n)* should be larger for more extreme quantiles (i.e. for α close to 0 or 1). *n* should be larger than $5/\min(\alpha,1-\alpha)$. α=0.5 gives median regression, corresponding to the Lp regression for *p*=1, i.e. to the method of the least absolute differences. Generally, the returned solution is less precise for small or large α. In some cases, the solution might not be unique.

Lp-regression: This method is based on minimization of the sum $\Sigma|e_i|^p$, amounting to a generalization of the least squares method based on Σe_i^2 minimization. Parameter *p* is entered in the *p* field *(p ≥ 1).* *p*=1 gives median regression, i.e. the method of the least absolute differences. It is very useful for data whose distribution is similar to the Laplace distribution. *p*=2 corresponds to the least squares regression, $p \to \infty$ (*p*≈10 is typically selected in practice) corresponds to minimization of maximum error (minimax). It is very sensitive to outliers and it should be used only when the errors are uniformly distributed. When $1 \le p < 2$ is selected, the resulting estimates are rather robust against outliers. *p*=1.5 is the default. Solution to the Lp regression might not be unique. Iterative randomized simplex optimization method is used for computations.

Least median: A modern, highly robust regression (often called LMS) method based on minimization of the median of squared differences. Iterative randomized simplex optimization method is used for computations.

IRWLS exp(-e): A robust regression method producing M-type estimates. It is based on iterative minimization of sum of squared standardized residuals $w(e_{ni})$, using weights $w(e) = \exp(-e)$. *Iteratively Re-Weighted Least Squares* are used for computations.

M-estimates, Welsch: A robust regression method producing M-type estimates. It is based on iterative minimization of sum of squared standardized residuals $w(e_{ni})$, using weights $w(e) = \exp(-e^2)$. *Iteratively Re-Weighted Least Squares* are used for computations.

BIR: Bounded influence regression. This method is robust not only against response variable outliers but also against influential observations (influence is connected to dependent variables values). It is this second robustness feature that distinguishes the method from previously discussed robust techniques. It might be useful for polynomial models when trying to suppress influence of extreme **x** points (low and high) on the fit. *Iteratively Re-Weighted Least Squares* are again used for computations.

Stepwise All: All possible subsets regression (*APSR*). This method is a useful tool for selection of important variables to be included in a regression model. The models can be compared by one of the following three measures: F-statistic (FIS), Akaike's information criterion (AIC) and mean squared error of prediction (MEP). When *APSR* is invoked, QC.Expert™ explores all combinations of variables from the set of potential explanatory variables (model terms) supplied by the user. A regression model is fitted for each of the combinations. The results are outputted both to the protocol and to a special output data sheet *APSR* (the sheet is created automatically). The text output is further enhanced by three plots in the graph window. Warning! Maximum number of model terms allowed is 12 without the intercept, or 13, including the intercept. The restriction is common for polynomial, full quadratic and general models. Since the results are outputted to a data sheet, the restriction comes from the maximum number of data sheet rows allowed by the QC.Expert™. The number of all possible models gets large very quickly. For *m* potential model terms (including the intercept), there are $2^m - 1$ possible regressions. Ordinary least squares method is used for all computations. For further details, see the Protocol and Graphical output paragraphs.

Data: Here, you can specify which part of data you want to use in computations. You can specify all data rows, selected rows only, or the rows which are not selected.

Transformation: Data transformation is defined here, see the previous paragraphs discussing model specification.

Output: Invokes a panel allowing you to customize some of the output features, see the next paragraph for details.

Help: Invokes help screen.

Cancel: Cancels immediately preceding operation.
OK: Runs the computations.

Output

The panel is invoked by the *Output* button in the Linear regression panel. Some of the output features can be customized here, specifying text and/or graphical items requested. There are three lists in the panel: *Protocols* (protocol items), *Graphical output* (plots or groups of plots), *Prediction* (predictions are requested for variables selected here). The *Prediction* list variables are used only when the *Prediction* item is checked. Shortcut buttons *Minimal*, *Standard*, *Extended*, *Complete*, *All*, *None* are available.

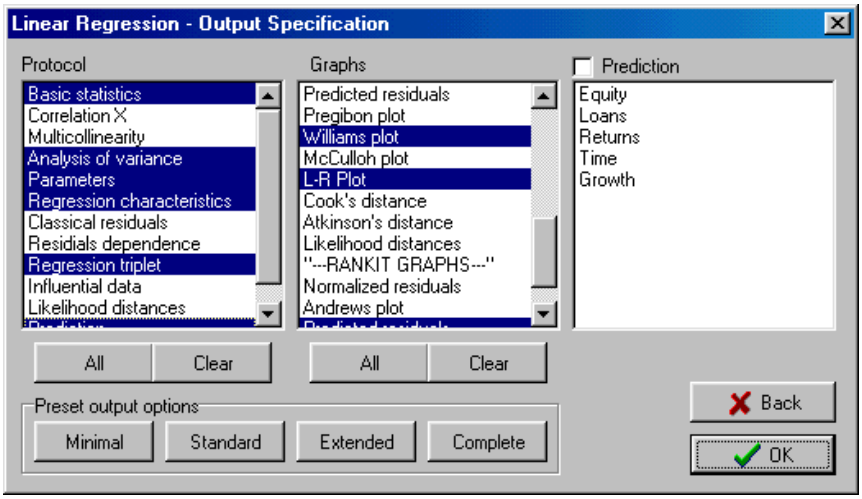


Fig. 5 Output dialog panel

Size of some output items depends on the number of data points. Keep in mind that the output can become rather large and difficult to read when all items are requested for large datasets. Next, we will describe contents of various individual output items, both text and graphical.

Protocol field
Summary statistics: Basic summary statistics: mean, standard deviation. Correlation coefficient and result of its test are produced for all pairs response-explanatory variable;
Correlation X: Pairwise correlation coefficients and results of their tests for all possible explanatory variables pairs;
Multicollinearity: Eigenvalues related to the design matrix (matrix of explanatory variables), condition number κ, variance inflation factor (VIF), multiple correlation coefficients;
ANOVA: overall (arithmetic) mean of the response variable, sums of squares, mean squares for the following variability sources: (corrected) total, model, residuals (error). Results of the overall F-test for the model, observed F-statistic value, $F(1-\alpha, m-1, n-m)$ quantile;
Parameters: regression parameters estimates, followed by estimates of their standard errors, individual confidence intervals and results of their tests;
Characteristics: Multiple correlation coefficient R, coefficient of determination R^2, R_p, mean squared prediction error (*MEP)*, Akaike information criterion (*AIC)*;

Residuals: observed *Y*, predicted *Y*, standard deviation of Y, residual standard deviation, residual variance, residual sum of squares, residuals, weights, mean of absolute residuals, skewness and kurtosis computed from residuals;
Residual dependence: Wald test for autocorrelation, Durbin-Watson test for autocorrelation, and sign test for lack of residual independence;
Regression triplet: Fisher-Snedecor test for the model, Scott's multicollinearity criterion, Cook-Weisberg test for heteroscedasticity, Jarque-Berr test for normality, tests for dependence;
Influential data: standard residuals, jackknife residuals, predicted residuals, projection matrix (i.e. hat matrix, **H**) diagonal, extended hat matrix ($\mathbf{H}^*$) diagonal, Cook's distance, Atkinson's distance, Andrews-Pregibon statistic, assessment of individual data points influence upon prediction, parameter estimates *LD*(*b*), variance *LD*(*s*), total influence *LD*(*b*,*s*);
Likelihood-related influence measure: assessment of individual data points influence upon parameter estimates *LD*(*b*), variance *LD*(*s*), total influence *LD*(*b*,*s*);
Prediction: Predictor values. Predictions and their confidence intervals.
Graphical output field
There are five groups of items in this field:
Regression curve;
Residuals: *Y-predicted values, Residuals vs. Predicted, Abs. residuals, Squared residuals, residual QQ-plot, Autocorrelation, Heteroscedasticity, Jackknife residuals, Predicted residuals;*
Partial regression plots: *Partial regression plots, Partial residual plots;*
Influential data: *Projection matrix, Predicted residuals, Pregibon, Williams, McCulloh, L-R Plot, Cook's D, Atkinson's distance;*
Q-Q plots: *Standardized residuals, Andrews plot, Predicted residuals, Jackknife residuals.*

Prediction field

You can select variables to be used as predictors of the response. Names of the predictors are arbitrary, but their number and order in which they appear must respect the regression model specification. When User transformation is selected, the *Variable association* panel is invoked (Fig. 6). There, you must associate selected predictor names listed on right to the model explanatory variable names listed on left. Predictors can have arbitrary number of rows (corresponding to points in which the predictions are requested). Explanatory variables used in model fitting can be used as predictors.

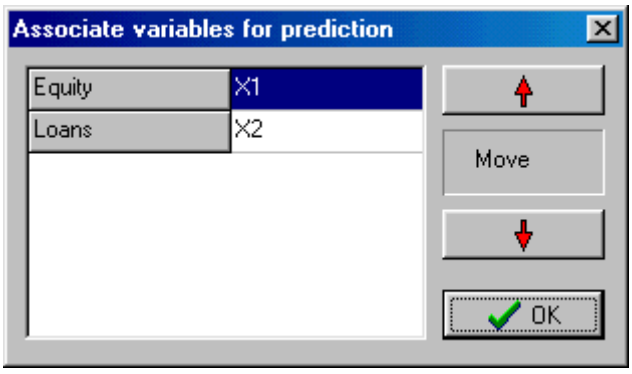


Fig. 6 Variable association panel

All: Selects all items
Nothing: Cancels previous selection
Minimal, Standard, Extended, Complete: Selects protocol and graphical output items according to the rules listed in the following table.

Table 2 Automatic protocol item selection

Item	Minimal	Standard	Extended	Complete
Summary statistics		o	o	o
Correlation X			o	o

Multicollinearity			o	o
ANOVA		o	o	o
Parameters	o	o	o	o
Characteristics	o	o	o	o
Residuals			o*	o*
Residual dependence				o
Regression triplet		o	o	o
Influential data			o*	o*
Likelihood related influence measure				o*
Prediction	o**	o**	o**	o**

* Size of this item depends on the number of data points!

** Depends on how the *Prediction* item is set

Table 3 Automatic graphical output item selection

Item	Minimal	Standard	Extended	Complete
Regression curve	o	o	o	o
Y-prediction		o	o	o
Residuals vs. Predicted	o	o	o	o
Abs. Residuals			o	o
Squared residuals				o
Residual QQ-plot		o	o	o
Autocorrelation			o	o
Heteroscedasticity			o	o
Jackknife residuals				o
Predicted residuals				o
Partial regression plots			o	o
Partial residual plots				o
Projection matrix	o	o	o	o
Predicted residuals				o
Pregibon				o
Williams		o	o	o
McCulloh				o
L-R plot		o	o	o
Cook’s D				o
Atkinson’s distance				o
Standardized residuals				o
Andrews plot			o	o
Predicted residuals		o	o	o
Jackknife residuals				o

Protocol

Task name	Project name, as inputted in the dialog panel.
Significance level	Inputted in the dialog panel. The level is used for all tests and confidence intervals.
Quantile t(1-alpha/2,n-m)	t-distribution quantile.
Quantile F(1-alpha,m,n-m)	F-distribution quantile.
Intercept	Is intercept included in the model?
Number of data rows	Number of complete data rows containing values for all model variables.
Number of parameters	Number of model terms, including intercept and terms created by

	transformations. For instance, for the 3rd order polynomial, the number of terms is 4.
Method	Computation method selected by the user.
Columns used in the model	List of variables used in the regression model.
Transformation	Transformation type selected by the user.
Summary statistics	
Variable characteristics	
Variable	Explanatory variable name.
Mean	Arithmetic average.
Std. deviation	Standard deviation.
Correlation with Y	Correlation between the response variable and the explanatory variable.
Significance	p-value from the correlation coefficient test.
Paired correlations (Xi, Xj)	Paired correlation coefficients for all explanatory variables pairs.
Multicollinearity indication	
Variable	Name of the variable related to the last column, where multiple correlations are listed (it has no relation to the other part of the output since eigenvalues cannot be, in general, directly related to individual variables).
Eigenvalues	Eigenvalues of the explanatory variables correlation matrix.
Condition number, kappa	Condition number (κ_{max}) is the ratio of largest and smallest eigenvalues (it is the maximum of condition index; l-th condition index is defined as the ratio of largest eigenvalue and the l-th eigenvalue). $\kappa_{max} > 1000$ indicates a strong multicollinearity.
VIF	Variance inflation factor, VIF > 10 indicates a strong multicollinearity.
Multiple correlation	Multiple correlation coefficient between the response and all explanatory variables.
ANOVA	
Overall Y mean	Arithmetic average of the response.
Source	Source of variability in the ANOVA table.
(Corrected) total	Response variability related to the model Y = Mean of(Y).
Model	[Total] – [Error].
Error	Residual variability, not explained by the model (i.e. the error variability).
F	F-statistic for the model. It should be larger than an appropriate theoretical F quantile. If it is larger, the actual model is significantly better than the null model Y=Mean of (Y).
Quantile F (1-alpha, m-1, n-m)	F-distribution quantile.
P-value	p-value for the test, if it is smaller than a specified significance level, the model is claimed to be significantly better than the null model.
Conclusion	Result of the test, stated in words.
Parameter estimates	
Variable	Variable name.
Estimate	Estimate of the regression coefficient associated with the explanatory variable.
Std. error	Standard error of the regression coefficient.
Conclusion	Result of the regression coefficient test, stated in words.
P-value	p-value for the regression coefficient test. If it is smaller than a specified

	significance level, significance is claimed.
Lower limit	Lower limit of the confidence interval computed with the pre-specified confidence level.
Upper limit	Upper limit of the confidence interval computed with the pre-specified confidence level. If zero is included in the interval, the regression coefficient is not significantly different from zero.
Characteristics of the model fit	
Multiple correlation coefficient, R	Multiple correlation coefficient characterizes how closely the model fits the data. It does not necessarily express how good the model is. R cannot decrease when a new variable is included in the model (it usually increases whenever a new variable is added)!
Coefficient of determination R^2	Square of the multiple correlation coefficient.
Predicted correlation coefficient, Rp	Predicted correlation coefficient, useful in the context of data containing outliers.
Mean square error of prediction, MEP	The i^{th} error is the difference between actual value of the i^{th} observation and its prediction. The prediction comes from the model based on data with the i^{th} row omitted. MEP is a sensitive indicator of some problems, like multicollinearity and outliers. It is an important characteristics of the regression model quality.
Akaike information criterion	AIC in the regression context is related to the residual sum of squares, penalized by the model size (number of explanatory variables).
Residual analysis	
Characteristic	
Y observed	Observed response value, as it appears in the current data sheet.
Y predicted	Predicted response value.
Std. error of Y	Estimated standard error of the prediction.
Raw residual	Difference between observed and predicted response value.
Residual [%Y]	Relative residual, raw residual divided by the response value.
Weights	Weights for individual observations as inputted by the user.
Residual sum of squares	Residual sum of squares cannot decrease when a new variable is included in the model (usually, it increases).
Mean of absolute residuals	Mean of absolute residuals.
Residual standard deviation	Standard deviation estimated from residuals.
Residual variance	Variance estimated from residuals.
Residual skewness	Skewness estimated from residuals.
Residual kurtosis	Kurtosis estimated from residuals.
Regression triplet testing	
Fisher-Snedecor overall test	Tests whether the actual model is better than the null model including only the overall mean.
F	Computed value of the F test stastistic.
Quantile F (1-alpha, m-1, n-m)	F-distribution quantile.
P-value	p-value for the test, if it is smaller than a specified significance level, the model is claimed to be significantly better than the null model.
Conclusion	Result of the test, stated in words.

Scott‘s multicollinearity criterion	Assessment of multicollinearity („dependence“) among explanatory variables. Severe collinearity can inflate regression coefficient variances substantially.
SC criterion	Computed test statistic.
Conclusion	Result of the test, stated in words.
Cook-Weisberg test for heteroscedasticity	Tests whether the error variance is constant across values of the explanatory variables. When the heteroscedasticity is detected, use of appropriate weights should be considered.
CW criterion	Computed test statistics.
Quantile Chi^2(1-alpha,1)	χ^2-distribution quantile.
P-value	p-value for the test, if it is smaller than a pre-specified significance level, significance is claimed.
Conclusion	Result of the test, stated in words.
Jarque-Berra test for normality	Test for error normality based on residuals.
JB criterion	Computed test statistic.
Quantile Chi^2(1-alpha,2)	χ^2-distribution quantile.
P-value	p-value for the test, if it is smaller than a pre-specified significance level, significance is claimed.
Conclusion	Result of the test, stated in words.
Wald test for autocorrelation	Test for autocorrelation among errors. It is based on residuals
WA criterion	Computed test statistic.
Quantile Chi^2(1-alfa,1)	χ^2-distribution quantile.
P-value	p-value for the test, if it is smaller than a pre-specified significance level, significance is claimed.
Conclusion	Result of the test, stated in words.
Durbin-Watson test for autocorrelation	Test for autocorrelation among errors.
DW criterion	Computed test statistic.
Conclusion	Result of the test, stated in words.
Sign test	A nonparametric test for residual dependence. It can detect some of the model inadequacies.
Sg criterion	Computed test statistic.
Quantile N(1-alpha/2)	Normal distribution quantile.
P-value	p-value for the test, if it is smaller than a pre-specified significance level, significance is claimed.
Conclusion	Result of the test, stated in words.
Influence measures	
A. Residual analysis	
Characteristic	
Standardized	It is sometimes called the studentized residual. Raw residual divided by its standard error s_r.sqrt(1-H_{ii}). s_r is the residual standard deviation.
Jackknife	Jackknife residual. It is similar to the Standardized residual. Instead of s_r, the residual standard deviation for the model based on data with i-th row

	deleted is used for the i-th residual. This type of residual is more sensitive to outliers.
Predicted	Predicted residual, difference between the *i*-th response value and prediction obtained from the model based on data with the i-th row deleted. This type of residual is more sensitive to outliers.
Diag(Hii)	Diagonal elements of the projection matrix. A large value indicates a data point that can potentially have a high influence upon the regression estimates. Sum of the H_{ii}'s is equal to the number of parameters in the model. Potentially influential points are marked in red.
Diag(H*ii)	Diagonal elements of the H^* matrix. The matrix is obtained when the design matrix (i.e. the matrix containing explanatory variables columns) is augmented with the response variable column. A large value indicates a data point that can potentially have a high influence upon the regression estimates. Sum of the H^*_{ii}'s is equal to the number of parameters in the model plus one. Potentially influential points are marked in red.
Cook's D	Cook's distance measures influence of the i-th data point upon the regression estimates. It combines measure of potential influence with the assessment of whether the point is actually an outlier. Influential points are marked in red.
B. Influence analysis Characteristic	
Atkinson's statistic	Atkinson's modification of Cook's D (1985), both characteristics yield similar results usually. Influential points are marked in red.
Andrews-Pregibon statistic	Andrews-Pregibon statistic measures influence that individual data points have on the variance of the regression parameters (volume of the confidence ellipsoid). Influential points are marked in red.
Y^ influence	Relative influence of individual data points upon prediction. Influential points are marked in red.
Parameter influence, LD(b)	Relative influence of individual data points upon parameter estimates. Influential points are marked in red.
Variance influence, LD(s)	Relative influence of individual data points upon residual variance. Influential points are marked in red.
Total influence, LD(b,s)	Simultaneous influence of individual data points upon parameter estimates and variance. Influential points are marked in red.
Prediction	
Predictor value	Values of all model terms. The intercept is represented by the column of ones.
Prediction	Predicted value based on the fitted model.
Lower limit	Lower limit of the confidence interval for the predicted mean, computed for a pre-specified confidence coefficient α.
Upper limit	Upper limit of the confidence interval for the predicted mean, computed for a pre-specified confidence coefficient α.

APSR regression protocol

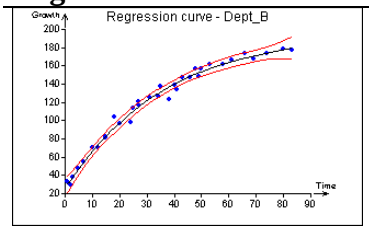
APSR (*all possible subsets regression*) helps to find the best model according to one of the following three criteria: F-statistic, Akaike's information criterion (AIC) or MEP (mean squared prediction error). The APSR procedure fits all possible model terms combinations. The results are outputted both to the protocol and to a special output data sheet *APSR* (created automatically). For each possible combination of model terms, the protocol contains a paragraph indicating which terms

were actually used and values of the three criteria. To save space, each of the model terms is coded by a short alphanumeric code (instead of its actual name which can be rather long and complicated). These codes are then used for each of the subsets description. The model which is the best in terms of a particular criterion can be found easily by sorting the *APSR* data sheet rows according to the criterion. Before sorting, all columns of the *APSR* sheet have to be selected, see *QC.Expert – Sort*. Alternatively, the point with the best value of a particular criterion can be found on the plot (part of the output, see the next paragraph, Graphical output) and selected there. A good model should have large value of F, small AIC value and small MEP value. Each of the criteria can favor different models. It is generally recommended to explore several models corresponding to very good values of a particular criteria (not only the model selected as the best). One should also keep in mind a somewhat different nature of the three criteria when interpreting *APSR* results. F is the F statistic involved in the usual F test, Akaike's criterion $AIC = n.\ln(RSS/n) + 2.m$ judges residual sum of squares together with a model size (the number of model terms) penalization. It was derived under much more general circumstances from information theory principles. *MEP* judges model's prediction abilities. There is no universally „best" model. Selection of the model should be led by the purposes which it is intended for and subject matter knowledge of the modeled situation.

Selected columns	Variables which are considered as potential model terms. Each of them is assigned a simple code to save space and keep the output easily readable.
Model comparison	A copy of this table is saved to an automatically created data *APSR* sheet. The sheet output can be sorted according to various criteria ((*Menu – QCExpert – Sort*). Various models can be also selected graphically. The *Protocol* window output cannot be manipulated with. Output contains columns with values of the F, AIC, MEP criteria, as well as the residual sum of squares (SSE). Warning: SSE might not directly express how good the model is! The largest model has always the smallest SSE.

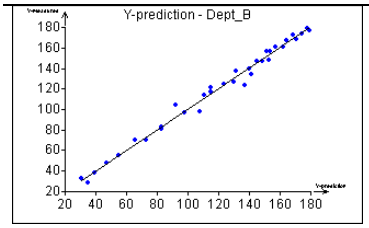
Graphs

Regression curve

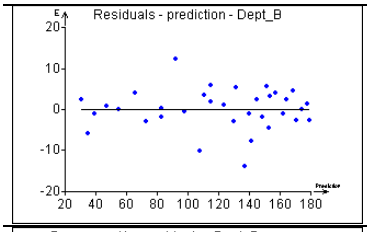
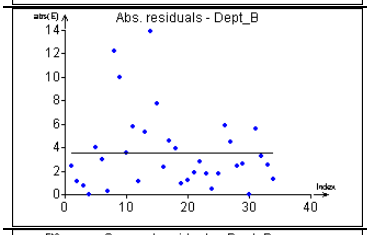
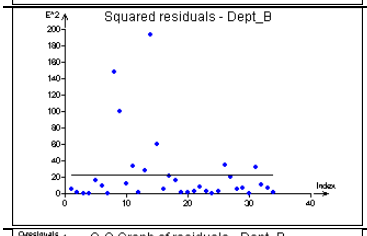
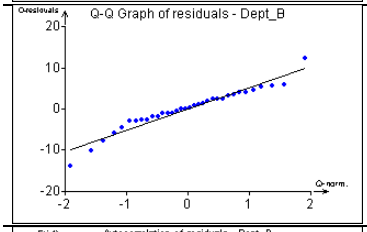
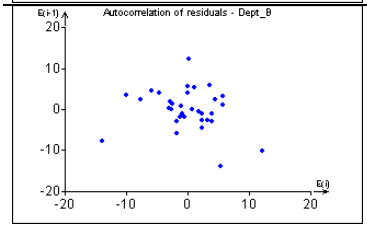
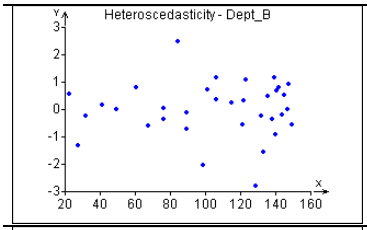
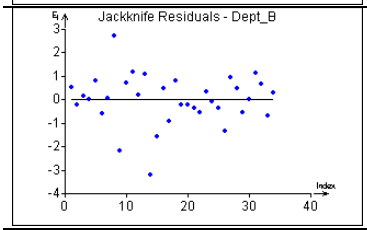
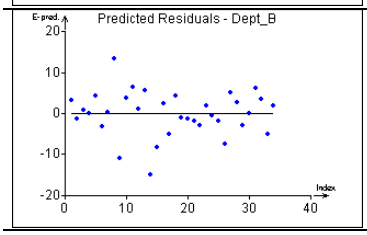


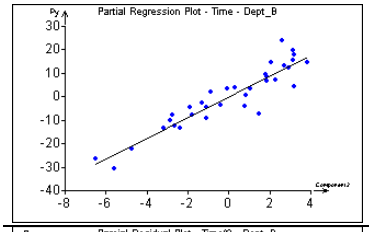
This plot is *not* produced when the model contains more than one explanatory variable. When only one explanatory variable appears in the model, the plot displays the regression curve. Red curves show the confidence band around the regression curve, computed for a pre-specified confidence coefficient. It should be noted that the confidence band is realistic only when the fitted model is (approximately) correct. This is even more important when predictions further from bulk of available data points are considered. Details of the plot can be inspected upon zooming part of it. The regression curve can be inspected even outside of the interval containing explanatory variable values actually used in model fitting by inverse zooming.

Residuals

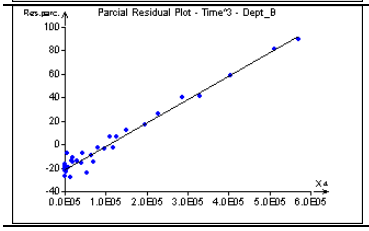


The plot shows how closely the model fits data. Predicted response values are plotted on the *X* axis, while observed response values are plotted on the *Y* axis. Vertical difference between a point and the line corresponds to a residual.

	Standardized residuals plot. Predicted response is plotted on the *X* axis, while the standardized residuals are plotted on the *Y* axis. Horizontal line corresponds to the mean of residuals. When ordinary least squares are used to fit a model including intercept, the residual mean is necessarily zero.
	Absolute residuals. The order in which a particular data point appears in the dataset is plotted on the *X* axis. The horizontal line corresponds to the mean absolute residual.
	Squared residuals. The order in which a particular data point appears in the dataset is plotted on the *X* axis. The horizontal line corresponds to the mean squared residual (i.e. mean squared error estimate).
	Q-Q plot for residual normality check. Approximately normally distributed (Gaussian) residuals should plot close to the line. Note that the ordinary least squares tends to enhance normal appearance of the residuals (so called supernormality effect). When in doubt, one should check also the residual Q-Q plot based on some robust method.
	Graphical check for the first order autocorrelation in residuals. The *i* -th residual is plotted on the *X* axis, while the (*i*-)-th is plotted on the *Y* axis. When the point cloud suggests a positive slope, positive 1-st order autocorrelation is suspected. Negative slope suggests negative autocorrelation. An autocorrelation in the residuals might not always be connected to the autocorrelation in errors. Residuals tend to be somewhat correlated even if the true errors are not.
	Graphical check for heteroscedasticity (error variance depends on explanatory variable(s)). A non-rectangular shape of the point cloud suggests a heteroscedasticity (e.g. a fan shape).
	Jackknife residuals (see the Protocol paragraph) are much more sensitive to outliers in the response variable than raw residuals. Even the jackknifed residuals may fail to detect a cluster of several outliers (they mask each other).
	Predicted residuals are much more sensitive to outliers than the raw residuals. Even the predicted residuals may fail to detect a cluster of several outliers (they mask each other).
Partial regression plots	

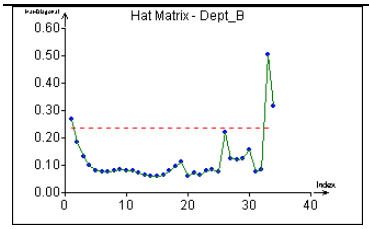


Partial regression plot displays relationship between the response and a given explanatory variable (a single model term) after the relationship has been cleared for a possible confounding caused by other variables in the model. Slope of the line corresponds to the regression coefficient for the variable in the complete model. Closeness of the linear fit on the plot is related to the significance test in the complete model.

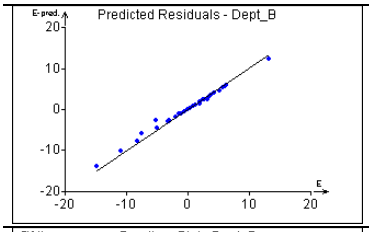


Partial residual plot. It is a modification of the partial regression plot. Nonlinear nature of the plot suggests that a term that is nonlinear in the variable just explored should be added to the model (e.g. a higher power of the variable might be tried).

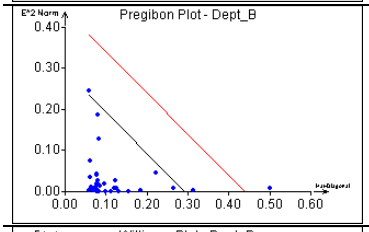
Influence



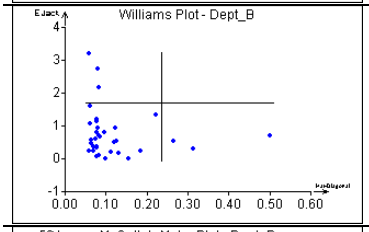
Plot of the projection matrix $\mathbf{H}=\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ diagonal elements. ($\mathbf{X}$ is the design matrix, i.e. the matrix containing explanatory variables as columns.) The element sizes are related to potential influence that the individual data points might have upon the regression results. The points plotted above the red horizontal line are considered to be potentially influential.



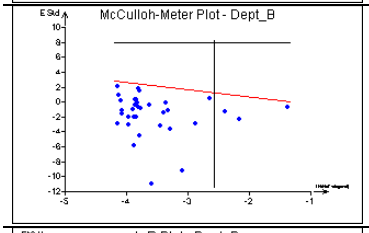
Predicted versus raw residuals plot. A large deviation from the line suggests that the corresponding observation is an outlier. The plot is very good in detection of isolated outliers. It is less sensitive to clusters of outliers which „mask“ each other.



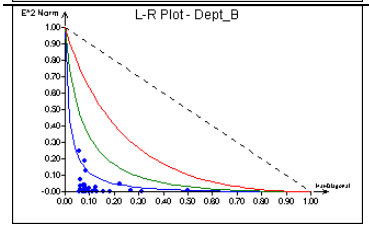
Pregibon's plot for simultaneous assessment of outliers and influence. The points above the lower (black) line are considered to be potentially influential, while the points above the upper (red) line are considered to be either substantially influential or outliers. Such data points should be checked carefully.



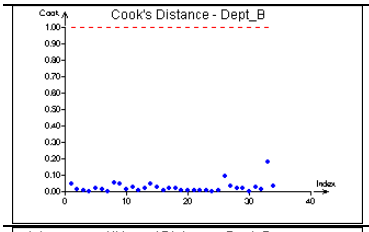
Williams' plot for simultaneous assessment of outliers and influence. The points located right from the vertical line are potentially influential, while the points above the horizontal line are suspected outliers.



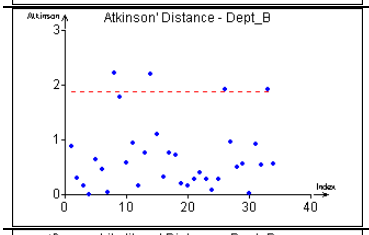
McCulloh-Meter plot for simultaneous assessment of outliers and influence. The points located right from the vertical line are potentially influential, while the points above the horizontal line are suspected outliers. The points above the red line are suspect either because they are influential or because they are outliers.



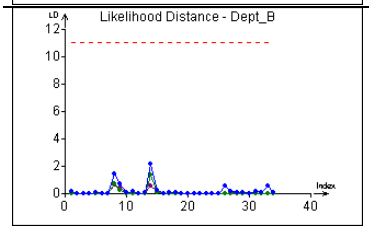
L-R plot for influence assessment. Hyperbolic curves are influence contours (connecting the points having the same influence). According to the location with respect to the three colored curves, data points can be classified as moderately influential, influential and substantially influential. The plot is most useful for smaller datasets.



Cook‘s distance is related to the influence data have upon magnitude (not variance) of the regression coefficients.

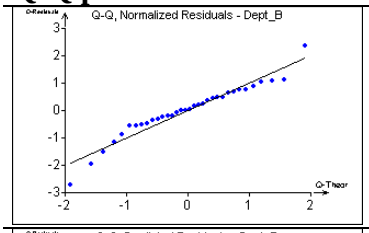


Atkinson‘s distance was derived as modification of the Cook‘s distance. Usually, the two yield similar results. Data points plotted above the horizontal line are considered to be influential.

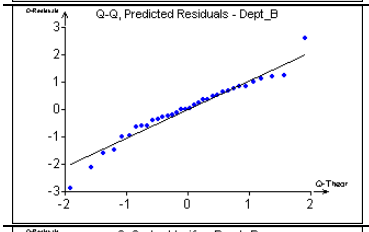


Likelihood related influence measure plot. The blue points express simultaneously the influence upon parameters and model predictions. Violet points express influence upon parameters, green points express the influence upon model predictions separately.

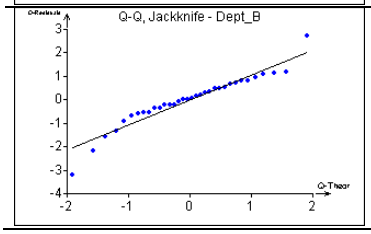
Q-Q plots



Q-Q plot of standardized residuals. It is used to assess residual normality. Approximately normally distributed residuals should plot close to the line.

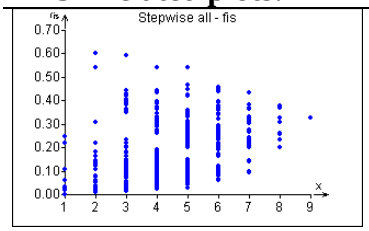


Q-Q plot of predicted residuals. It is used to assess residual normality. Approximately normally distributed residuals should plot close to the line.

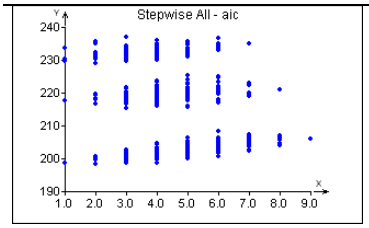


Q-Q plot of jackknife residuals. It is used to assess residual normality. Approximately normally distributed residuals should plot close to the line.

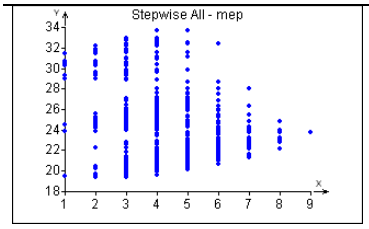
APSR related plots:



The plot is generated by the *APSR* (all possible subsets regression) procedure. It is useful when looking for the best models in terms of the F criterion. Number of variables included in the model is plotted on the *X* axis, while the F value is plotted on the *Y* axis. A good model should have a large F value. The best points (corresponding to models) can be selected interactively for further exploration (their detailed description can be found in the *APSR* data sheet, where they are selected automatically, once they are marked on the plot). It is highly recommended to choose several good looking models, explore them and select among them manually, using some subject matter knowledge.



The plot is generated by the *APSR* (all possible subsets regression) procedure. It is useful when looking for the best models in terms of the Akaike's criterion (AIC). Number of variables included in the model is plotted on the *X* axis, while the AIC value is plotted on the *Y* axis. A good model should have a small AIC value. The best points (corresponding to models) can be selected interactively for further exploration (their detailed description can be found in the *APSR* data sheet, where they are selected automatically, once they are marked on the plot). It is highly recommended to choose several good looking models, explore them and select among them manually, using some subject matter knowledge. Bands appear on the plot when there is a highly significant term among the potential model terms (much more important than other potential terms).



The plot is generated by the *APSR* (all possible subsets regression) procedure. It is useful when looking for the best models in terms of the mean squared error of prediction (MEP). Number of variables included in the model is plotted on the *X* axis, while the MEP value is plotted on the *Y* axis. A good model should have a small MEP value. The best points (corresponding to models) can be selected interactively for further exploration (their detailed description can be found in the *APSR* data sheet, where they are selected automatically, once they are marked on the plot). It is highly recommended to choose several good looking models, explore them and select among them manually, using some subject matter knowledge.