User Manual

DamReg

Version 1.1
December 2004

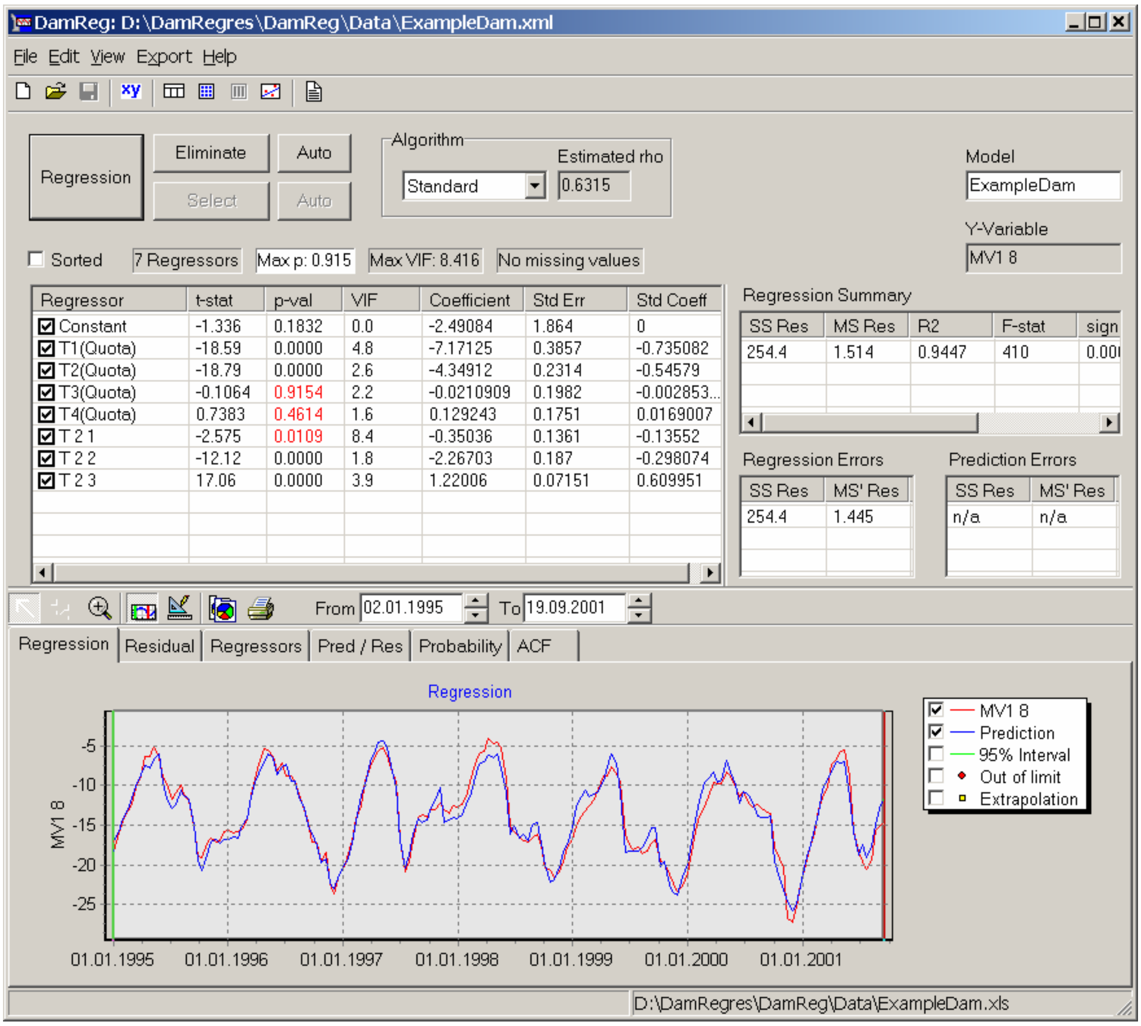
Benedikt Weber, Lennoxville, Québec, Canada
Federal Office for Water and Geology (FOWG) , Bienne, Switzerland

Contents

Getting started

DamReg is a software that performs statistical regression analysis. It has many built-in transformations commonly used for dam models, as well as custom transformations defined by formulas or tabular values (spline). Besides the standard regression algorithm, the program can also perform Ridge Regression, Principal Component Analysis, and the Prais-Winsten algorithm.

When the program is started, the main window shows up. In the main window you can perform the regression analysis, eliminate and select regressors manually or automatically. The results are shown in different tables and plots.



DamReg: D:\DamRegres\DamReg\Data\ExampleDam.xml

Algorithm: Standard

Estimated rho: 0.6315

Model: ExampleDam

Y-Variable: MV1 8

□ Sorted | 7 Regressors | Max p: 0.915 | Max VIF: 8.416 | No missing values

Regressor	t-stat	p-val	VIF	Coefficient	Std Err	Std Coeff
☑ Constant	-1.336	0.1832	0.0	-2.49084	1.864	0
☑ T1(Quota)	-18.59	0.0000	4.8	-7.17125	0.3857	-0.735082
☑ T2(Quota)	-18.79	0.0000	2.6	-4.34912	0.2314	-0.54579
☑ T3(Quota)	-0.1064	0.9154	2.2	-0.0210909	0.1982	-0.002853...
☑ T4(Quota)	0.7383	0.4614	1.6	0.129243	0.1751	0.0169007
☑ T 2 1	-2.575	0.0109	8.4	-0.35036	0.1361	-0.13552
☑ T 2 2	-12.12	0.0000	1.8	-2.26703	0.187	-0.298074
☑ T 2 3	17.06	0.0000	3.9	1.22006	0.07151	0.609951

Regression Summary

SS Res	MS Res	R2	F-stat	sign
254.4	1.514	0.9447	410	0.00

Regression Errors

SS Res	MS' Res
254.4	1.445

Prediction Errors

SS Res	MS' Res
n/a	n/a

From 02.01.1995 To 19.09.2001

D:\DamRegres\DamReg\Data\ExampleDam.xls

Main Window

Steps for a regression analysis

The most important steps you have to do when performing a regression analysis are:

- Importing data
- Defining regressors
- Performing the regression
- Modifying the regression model
- Exporting results and saving the model

Importing data

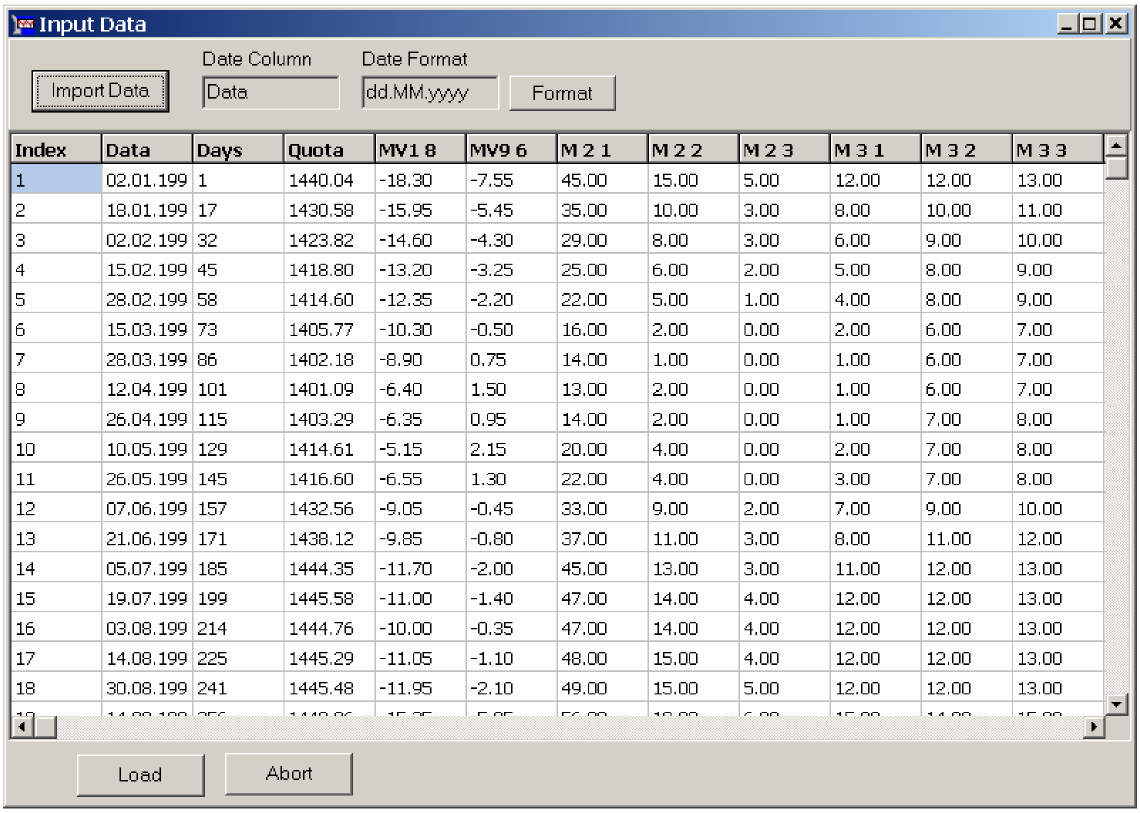
The first step for performing a regression analysis is to import measured data. The standard input format is an Excel file with a heading (variable name) for each data column. The row with the variable names may be preceded by several title lines that are ignored by the program. Tab-separated and comma-separated text files can also be imported.

From the menu, choose **File | New Model…** or click 🗋 in the menu bar. The **Input Data** window opens with an empty spread sheet. Press the **Import** button to open a **File Dialog.** Select any of the *.xls or *.txt or *.csv files provided in the data directory or it subdirectories. The data are shown in the spread sheet.

For Excel files, the format for displaying and reading the values is taken from Regional Options (Windows Settings | Control Panel). For text files, the input format is determined automatically as far as possible, but it might be necessary to adjust the format manually by pressing the **Format** button and changing the corresponding formats. Particularly, check the date format displayed at the top of the window and adjust it manually if not correct.

The first column is an index column generated automatically. The index will be used as a time scale, if no date column is provided. If a date column is provided, there is an additional Day column showing the number of days, starting with 1 at the first data row. The date column is indicated at the top of the window. To use the imported data, press the **Load** button. The **Input Data** window closes and the name of the data file is shown at the right in the **Status Bar** at the bottom of the window. If there are still reading errors, the window does not close, and the errors are indicated.

To view the **Input Data** window again, chose **View | Input Data...** from the menu or click ▥ in the main tool bar.



Date Column: Data

Date Format: dd.MM.yyyy

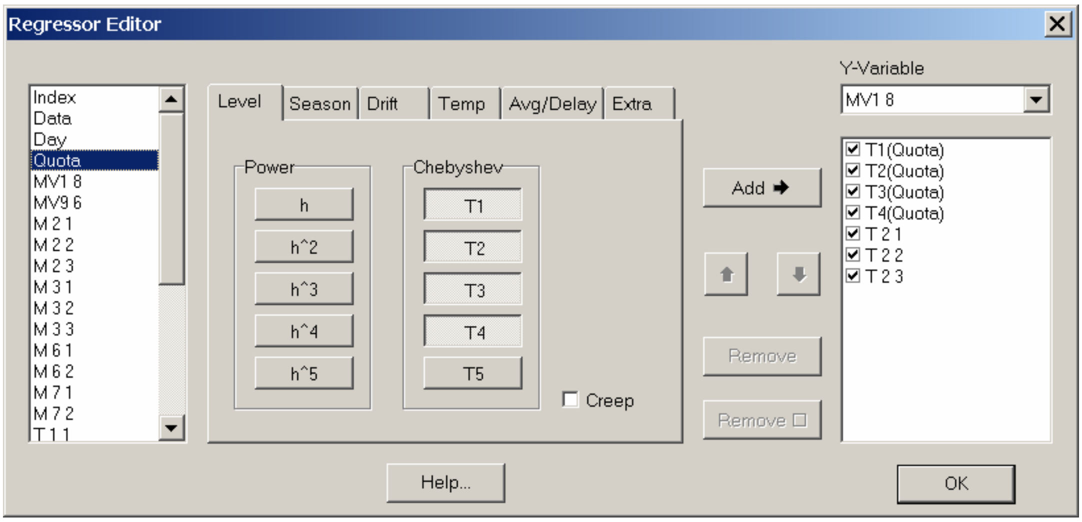
Index	Data	Days	Quota	MV1 8	MV9 6	M 2 1	M 2 2	M 2 3	M 3 1	M 3 2	M 3 3
1	02.01.199	1	1440.04	-18.30	-7.55	45.00	15.00	5.00	12.00	12.00	13.00
2	18.01.199	17	1430.58	-15.95	-5.45	35.00	10.00	3.00	8.00	10.00	11.00
3	02.02.199	32	1423.82	-14.60	-4.30	29.00	8.00	3.00	6.00	9.00	10.00
4	15.02.199	45	1418.80	-13.20	-3.25	25.00	6.00	2.00	5.00	8.00	9.00
5	28.02.199	58	1414.60	-12.35	-2.20	22.00	5.00	1.00	4.00	8.00	9.00
6	15.03.199	73	1405.77	-10.30	-0.50	16.00	2.00	0.00	2.00	6.00	7.00
7	28.03.199	86	1402.18	-8.90	0.75	14.00	1.00	0.00	1.00	6.00	7.00
8	12.04.199	101	1401.09	-6.40	1.50	13.00	2.00	0.00	1.00	6.00	7.00
9	26.04.199	115	1403.29	-6.35	0.95	14.00	2.00	0.00	1.00	7.00	8.00
10	10.05.199	129	1414.61	-5.15	2.15	20.00	4.00	0.00	2.00	7.00	8.00
11	26.05.199	145	1416.60	-6.55	1.30	22.00	4.00	0.00	3.00	7.00	8.00
12	07.06.199	157	1432.56	-9.05	-0.45	33.00	9.00	2.00	7.00	9.00	10.00
13	21.06.199	171	1438.12	-9.85	-0.80	37.00	11.00	3.00	8.00	11.00	12.00
14	05.07.199	185	1444.35	-11.70	-2.00	45.00	13.00	3.00	11.00	12.00	13.00
15	19.07.199	199	1445.58	-11.00	-1.40	47.00	14.00	4.00	12.00	12.00	13.00
16	03.08.199	214	1444.76	-10.00	-0.35	47.00	14.00	4.00	12.00	12.00	13.00
17	14.08.199	225	1445.29	-11.05	-1.10	48.00	15.00	4.00	12.00	12.00	13.00
18	30.08.199	241	1445.48	-11.95	-2.10	49.00	15.00	5.00	12.00	12.00	13.00

Input Data Window

Defining regressors

After importing the data, the next step is to define the regressors. From the menu select **Edit | Regressors…** or click xy in the menu bar. This opens the **Regressor Editor** that lets you define the regressors and the response (y-variable).

- From the list with all variables at the left select one or more variables you want to transform.
- In the middle, select one of the categories (Level, Season, Drift, Temperature, Average/Delay, and Extra) and select the transformations you want apply. For some categories, the most commonly used transformations are already selected, but you can always select different ones.
- Press the **Add** button to add the regressors to the list at the right. If you need to make changes, you can **Remove** regressors or change their order by pressing the **Up** and **Down Arrow** buttons.
- Select the **Y-Variable** (response variable).
- Click **OK** to close the Regressor Editor.



Regressor Editor

Index
Data
Day
Quota
MV1 8
MV9 6
M 2 1
M 2 2
M 2 3
M 3 1
M 3 2
M 3 3
M 6 1
M 6 2
M 7 1
M 7 2
T 1 1

Y-Variable: MV1 8

☑ T1(Quota)
☑ T2(Quota)
☑ T3(Quota)
☑ T4(Quota)
☑ T 2 1
☑ T 2 2
☑ T 2 3

□ Creep

Regressor Editor

Performing the regression

When you are back from the **Regressor Editor**, the **Regression Table** in the **Main Window** shows all regressors including a constant, and the **Plot Region** shows the response variable.

To perform the regression analysis press the **Regression** button.

The **Regression Table** shows the results for each regressor. The main result are the regression coefficients. More important for the model building, however, are the p-values and the VIF's (Variance Inflation Factors). Regressors with large p-values (> 1%) have a regression coefficient not significantly different from zero and should be eliminated. Regressors with high VIF's (>100) indicate a multicollinearity, that should be avoided.

Some information is displayed above the Regression Table: Number of regressors, Maximum p-value, Maximum VIF, Number of missing values. When the maximum p-value or the maximum VIF are above the limit, the background of the corresponding display is changed to white. The limit values are shown, when the cursor is moved over the respective display. The values can be changed in **the Edit | Options…** menu.

The **Summary Tables** to the right of the Regression Table show the summary of the regression and the errors for the regression and the prediction period, respectively. The most recent values are shown at the bottom of the tables. Numbers from a few previous calculations are shown for comparison.

You can view the different plots by choosing the tabs in the **Plot Region**. Within some of the plots you can also select different curves by checking or unchecking them in the legend. In the first three plots (Regression, Residual, Regressors) the regression period is shown by a light grey background. If you have regressors with a startup time, the overall startup time is shown by a dark grey background.

Modifying the Regression Model

Eliminating and selecting regressors

You can eliminate unwanted regressors manually by unselecting them and pressing the **Regression** button, or automatically by pressing the **Eliminate** button or the upper **Auto** button. **Eliminate** eliminates the regressor with the largest p-value (or equivalently the smallest absolute t-statistic). **Auto** performs several elimination steps until the largest p-value is below the limit value.

Analogously, there is a **Select** button and an **Auto** button for including regressors. For defining more regressors go back to the **Regressor Editor** (**Edit | Regressor…**).

Changing the regression period

The **Regression Period** may be changed by dragging the red and the green **Plot Cursors**. For exact values you can also select the dates in the two boxes above the plot region labelled **From** and **To**, respectively. After changing the regression period you have to update the regression by pressing the **Regression** button.

Exporting Results

There are different ways to view and export the results of a regression analysis. Tables with the regression results, the correlation matrix and the eigenvectors can be viewed by choosing **View | Regression…**, **View | Correlation…**, or **View | Eigenvectors…**, respectively in the menu or clicking the corresponding buttons 📈, ▦, or ▥ in the menu bar. The values in the tables can be selected and copied to the clipboard for pasting in other applications like Excel or Word.

If you want to create your own plots with Excel or some other charting tool, you can copy the necessary data to the clipboard or to a file by choosing **Export | Plot Data…** form the menu.

The most comprehensive information is given by the **Report**. Choose **View | Report...** from the menu or click 🗎 to generate the report.

Report

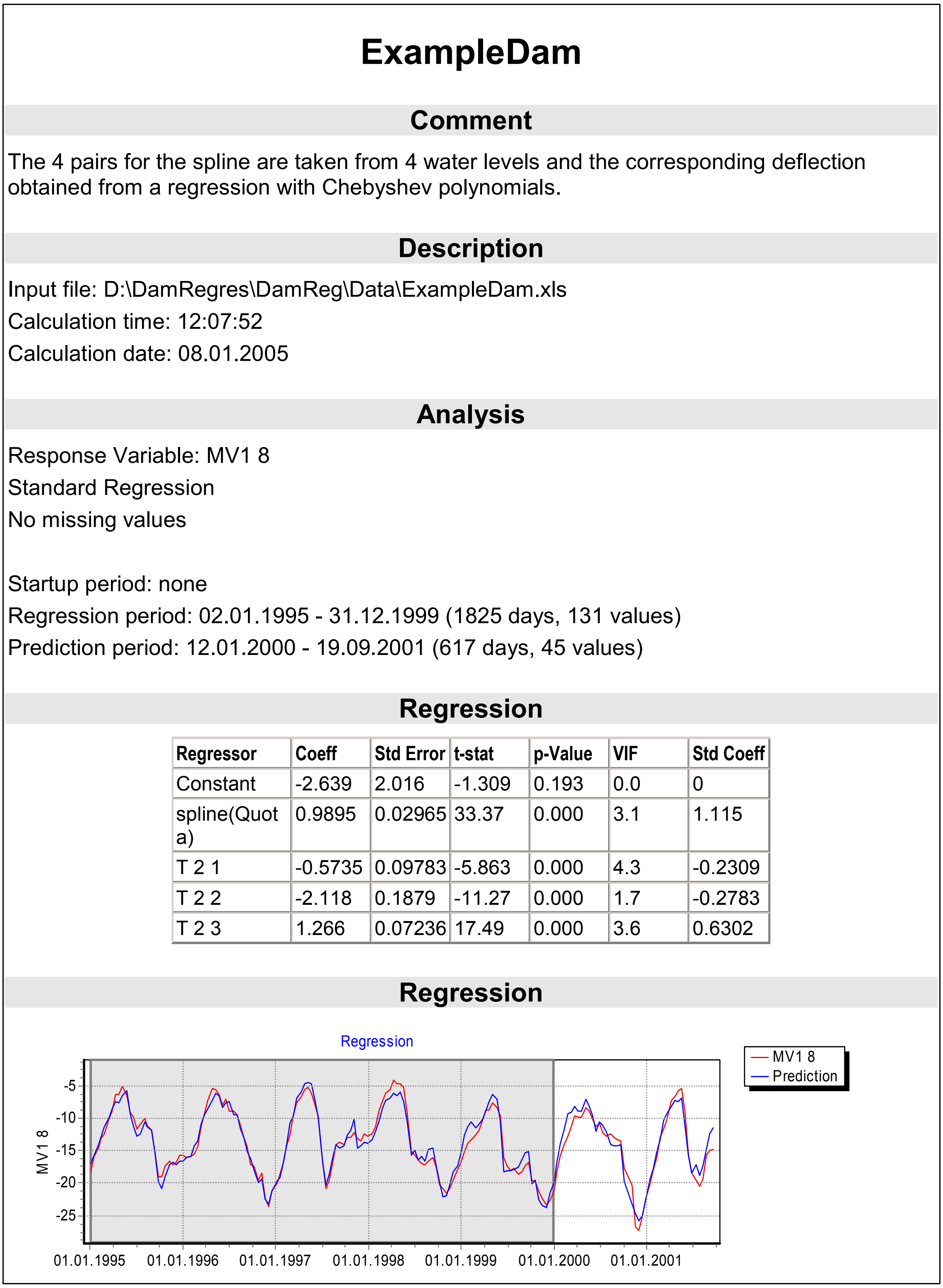
The report shows all relevant results including the plots. You can easily customise the report to your needs and save it as an RTF-file for importing in Word.

To customize the Report select entries from **Edit | Show Text**, **Edit | Show Table,** and **Edit | Show Plot.** To change the title or the comment select **Edit | Title and Comment…** This is the same menu as in the **Main Window.**

The Format menu lets you select different font sizes (**Format | Text Font** and **Format | Table Font Size**). You can also select a height to width ratio for the plots (**Format | Plot Aspect Ratio**) and turn the gray backgrounds on and off (**Format | Regression Background…**).

The File menu lets you setup the printer and print the report. The width of the plots is adjusted to the paper width. You can get larger plots by using a wider paper size or landscape mode.

To export the Report to RTF (for importing in Word) or HTML choose **File | Export…** Another possibility is to select part or all of the Report and copy it to the clipboard (**Edit | Copy** or **Ctrl-C**). When selecting tables, select the table as part of the global text flow, not the individual cells. Individual cells can be selected but are not copied. Plots are selected by double-clicking.



ExampleDam

Comment

The 4 pairs for the spline are taken from 4 water levels and the corresponding deflection obtained from a regression with Chebyshev polynomials.

Description

Input file: D:\DamRegres\DamReg\Data\ExampleDam.xls

Calculation time: 12:07:52

Calculation date: 08.01.2005

Analysis

Response Variable: MV1 8

Standard Regression

No missing values

Startup period: none

Regression period: 02.01.1995 - 31.12.1999 (1825 days, 131 values)

Prediction period: 12.01.2000 - 19.09.2001 (617 days, 45 values)

Regression

Regressor	Coeff	Std Error	t-stat	p-Value	VIF	Std Coeff
Constant	-2.639	2.016	-1.309	0.193	0.0	0
spline(Quota)	0.9895	0.02965	33.37	0.000	3.1	1.115
T 2 1	-0.5735	0.09783	-5.863	0.000	4.3	-0.2309
T 2 2	-2.118	0.1879	-11.27	0.000	1.7	-0.2783
T 2 3	1.266	0.07236	17.49	0.000	3.6	0.6302

Regression

Report

Saving the model

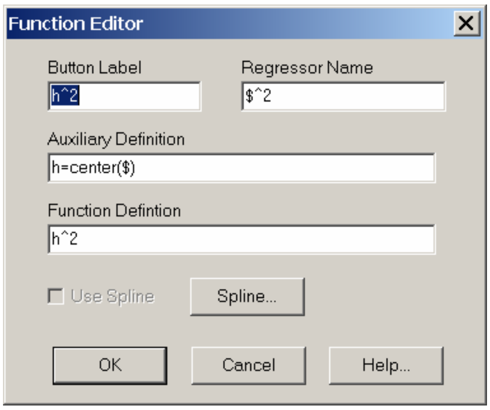
You can save the model by choosing **File | Save** or **File | Save As…** from the menu. This saves the name of your input data file, all modifications in the Regressor Editor, the regressors with their transformation functions, and the regression period. Also saved are the regression coefficients. When the model is opened later, the input data file is read in, the regressors are recalculated and the regression is performed. The calculated regression coefficients are compared with the saved coefficients and a message is shown if there are any differences.

Modifying transformations

Although most functions commonly used are defined in the **Regression Editor** there might be a need to use different functions. To modify a function, right-click the corresponding button and select **Edit**. Depending on the kind of function, a corresponding function editor opens. Many functions may be defined by the general **Function Editor.** More specialized functions, such as temperature transformations, have their own specialized function editors.

In all function editors you can define the label to appear on the button, the name for the regressor as it appears in the list at the right of the **Regressor Editor** and in the **Regression Table.** In the general **Function Editor** you can define the function as a formula, in the specialized function editors you can change only parameters of the corresponding function.

For more information on how to use the function editors, press the corresponding **Help…**



Button Label: h^2

Regressor Name: $^2

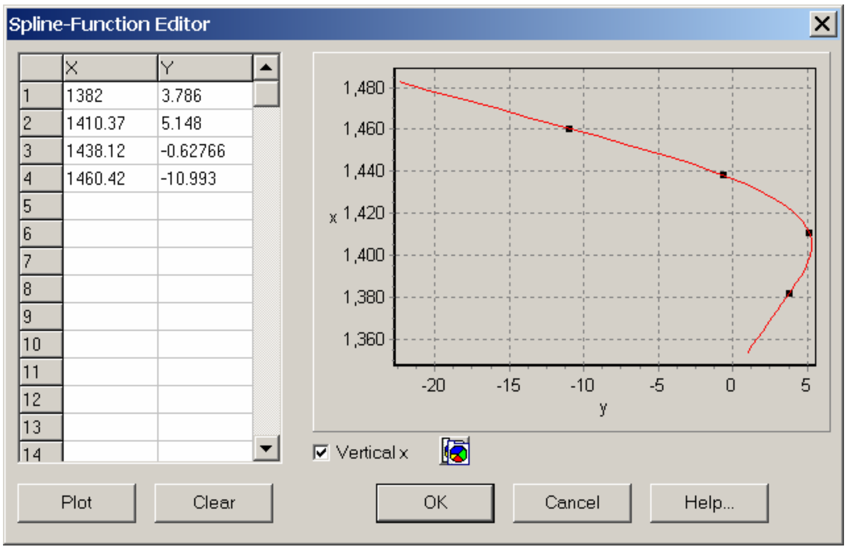
Auxiliary Definition: h=center($)

Function Defintion: h^2

☐ Use Spline

General Function Editor

Instead of defining the function by a formula, it can also be defined by a spline function. Press the **Spline…** button to invoke the **Spline Editor**. To define a spline function, fill in some points in the **X-Y-Table** and press the **Plot** button. The x-values must be in increasing order and there must be at least 3 points. To use a vertical x-axis (e.g. if x is the water level), check the **Vertical x** box.



Spline-Function Editor

	X	Y
1	1382	3.786
2	1410.37	5.148
3	1438.12	-0.62766
4	1460.42	-10.993
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		

☑ Vertical x

Spline Editor

Main Window

Menus

The menus with icons are also accessible through the corresponding tool buttons.

Menu File

New Model… Create new model and open data file.
Open Model… Open existing model.
Save: Save model.
Save As… Save model with a different name.
Exit: Exit application.

Menu Edit

Regressors… Open regressor editor for defining regressors and response variable.
Title and Comment… Open window for entering title and comment.
Options… Change options such as confidence level or maximum p-value.
Select all Regressors: Select all regressors in the main table. Also available as popup menu.
Deselect all Regressors: Deselect all regressors in the main table. Also available as popup menu.

Menu View

Data… Open window with table showing input data.
Correlation… Open window showing the correlation matrix.
Eigenvectors… Open window showing the singular values and eigenvectors.
Regression… Open window showing the regression results.
Report… Open window showing the Report.

The correlation matrix, the eigenvectors, and the regression results are also available in the Report.

Menu Export

Plot Data… Export plot data to tab-separated file or to clipboard. Plot data can be used to generate plots with other programs such as Excel.

Menu Help

Contents… Show Help.
About… Information about the software.

Regression Table

The **Regression Table** shows the following regression results:

- **Regressor:** Regressor name and checkbox for selecting the regressor. The first regressor is always the constant and should be selected.
- **t-stat:** t-statistic, equals Coefficient / Std Err.
  For inactive regressors, this value is shown in parenthesis and represents the value that would be there if the regressor were selected.

- **p-val:** p-value corresponding to the t-statistic.
  For inactive regressors, this value is shown in parenthesis and represents the value that would be there if the regressor were selected.
- **VIF:** Variance inflation factor
- **Coefficient:** Regression coefficient
- **Std Err:** Standard error of the regression coefficient
- **Std Coeff:** Standardized regression coefficient, regression coefficient for standardized regressors and response

When **Sorted** is checked, the regressors are sorted by ascending p-values. The Constant is not included in sorting but always shown at the top. Sorted regressors are useful for manual elimination.

Summary Tables

Regression Summary

- **SS Res:** Residual sum of squares
- **MS Res:** Residual mean square, equals SS Res / number of degrees of freedom, estimate for variance (standard deviation square)
- **R2:** Coefficient of multiple determination
- **F-stat:** F-statistic
- **signif F:** Significance of F-statistic, p-value corresponding to the F-statistic
- **DW-stat:** Durbin-Watson-statistic
- **Observ:** Number of observations, needed also for the interpretation of the Durbin-Watson-statistic

Regression Errors

- **SS Res:** Residual sum of squares for regression period
- **MS'Res:** SS Res / number of observations (slightly different from MS Res)

Predictions Errors

- **SS Res:** Residual sum of squares for prediction period
- **MS'Res:** SS Res / number of observations

Plotting Control

Zooming and Scrolling

Reset to normal view (no zoom and no scroll).

Scroll the plot by periods. Cycles through the following views (if available): 1. All periods, 2. Regression and prediction periods, 3 Regression period, 4. Prediction period.

Change to zoom mode. In zoom mode you can select a zoom window by dragging the cursor with the left mouse button from left to right. Dragging from right to left resets the zoom. Scroll horizontally and vertically with the right mouse button.

Changing Appearance

Toggle the visibility of the plot cursors in the regression plot (first tab).

Invoke the plot editor. The plot editor lets you modify the appearance of the plot in many ways. Changes are, however, only temporary and are lost when the plot is updated due to a new regression calculation.

Copying and printing

Copy the plot (as metafile) to clipboard.

Open the print dialog for the plot.

Copying and printing can also be done from the Report.

Selecting regression period

Move the red and green plot cursors in the Regression plot or select From and To date for small time steps.

Exporting Plot Data

The menu **Export | Plot Data…** lets you export plot data to a tab-separated file (**Save…**) or to the clipboard (**Copy**). These data can be used to generate plots with other programs such as Excel. You can select what data to save by checking the corresponding boxes. Y-values, prediction and residuals are always saved.

Pre-Regression: Y-values before the regression period
Confidence: Prediction interval, extrapolation and exceedance for prediction and residuals
Regressors: Regressors, either individually or grouped by categories depending on the selection in **Edit | Options…**
Statistics: Distribution of residuals, autocorrelation

The confidence level is the same as the one selected in **Edit | Options…** but may be changed temporarily for exporting the data.

Report Window

The report contains all results of a regression analysis. It can be customized to include only selected items, can be printed, exported to RTF and HTML and copied to the clipboard.

Menu File

The menu File lets you preview, print, and export the report. The following items are available:

Print Preview…
Print…
Print Setup…
Export…
Close

The report can be exported to RTF (for import in Word) and to HTML.

Changing the paper size or orientation (Print Setup…) also influences the size of the plots.

Menu Edit

The menu Edit lets you customize the content of the report. You can modify the title and the comment and select various text items, tables and plots. Copy and Select All let you copy the report to the clipboard. The following items are available:

Copy
Select All
Title and Comment…
Show Text >Title, Comment, Description, Analysis
Show Table > Correlation, Eigenvectors, Regression, Definition, Summary, Error Summary
Show Plot > Regression, Residual, Regressors, Probability, Pred/Res, ACF

Menu Format

The menu Format lets you select an overall size for text, tables and plots. The following items are available:

Text Font Size > Large, Medium, Small
Table Font Size > Large, Medium, Small
Plot Aspect Ratio > As in Main Window, 0.2, 0.3, 0.4, 0.5
Regression Background

Table font size is relative to text font size. Small table font size is useful to fit large tables to the page width (many regressors).

Plots are scaled to the width of the paper (with 2.5 cm left and right margin), so changing the paper size or orientation (**File | Print Setup…**) changes the size of the plots. The height of the plots is determined by the aspect ratio selected in this menu.

The startup and regression periods in the plots (regression, residual and regressor) are plotted with a dark grey and light grey background, respectively. If the background is unselected in this menu, only a grey frame is plotted to indicate the regression period.

Selecting tables and plots with cursor

Tables should be selected as a whole, i.e. by dragging the cursor over the whole table (starting to the left or above the table). Selecting all individual cell is possible but the selection cannot be copied to the clipboard. If you need to copy partial tables you can use the tables available in the menu **View** of he **Main Window**.

Plots are selected either by double-clicking or by dragging the cursor from the left of or from above the plot.

Defining Regressors

Regressor Editor

Defines regressors and the response variable.

At the left, there is a list of all variables read from the data file, including the generated variables "Index" and "Days" (if a date column is present). Select a single variable or several variables (shift-click for consecutive items, ctrl-click for non-consecutive items).

The middle part of the regressor editor contains several buttons for selecting transformation functions grouped by categories Level, Season, Drift, Temperature, Average/Delay, and Extra. Choose a category and select the buttons you want to use. The functions associated with the different buttons can be modified by right-clicking them. For information on individual categories see next section (Regressor Transformations).

Push the **Add** button to actually include the regressors in the list at the right. Each variable selected is transformed by each function selected. You can change the order of the regressors using the **Up** and **Down Arrows**, remove selected regressors by pressing the **Remove** button, and remove unchecked regressors by pressing the **Remove □** button.

To modify a regressor, remove it and add a new one. Modifying the definition of a function does not modify regressors already added.

To define the response variable, select it from the **Y-Variable** list.

Regressor Transformations (Categories)

The following categories are available in the Regressor Editor:

Level: Contains functions for transforming the water level. Predefined functions are polynomials and Chebyshev polynomials. Check the **Creep** box to apply creep effects for all selected transformations. The Creep function can also be modified by right-clicking the **Creep** checkbox**.**

Season: Contains functions modelling seasonal changes. Predefined functions are Sine and Cosine functions.

Drift: Contains functions for long-time behaviour such as inelastic deformations of the foundation.

Temperature: Contains functions for calculating the temperature in the dam from measured temperatures at the surface. A simple one-dimensional heat conduction model is used. Points near the surface respond to high-frequency variations and can only be calculated if the time steps are small enough. E.g. for temperatures at a depth of 2 meters the time step should be not more one week, for time steps at 4 meters it should be not more a month. The program roughly checks the resolution and gives a warning if it is not adequate.

Average/Delay: Contains functions for time averaging and time shifting of variables.

Extra: Can be used for any special effects not belonging to any other category.

General Function Editor

Define a function either by a formula or by a spline passing through several user defined points. The general function editor is used to define polynomials, seasonal functions, drift functions and extra functions.

Button Label

Text displayed on the button of the regression editor. Should be short enough to fit in the button.

Name

Text used as the name for the regressor. Should be a relatively short name to identify the regressor in the regression table. The letter $ is substituted by the actual variable name.

Auxiliary Definition

Formula to define an auxiliary function for use in the function definition. May be empty.

Examples: h=center($), h= ($-1200)/200, t=$/365

Function Definition

Formula to define the regressor.

Examples: h^2, sin(t), where h or t are given by the Auxiliary Definition.

Formulas

Use $ as a placeholder for the variable. Undefined variables and functions are treated as zero.

To define a function (Auxiliary Definition) use f= sin, or f= sin($), to evaluate it use f($). Do not type f($)= sin($). The parser will try to evaluate f($) and give an error.

Formulas may contain the following constants, operators and functions:

- Constant: e, pi
- Operators: +, -, *, /, ^
- Functions: sin, cos, exp, ln, sqrt, abs, ramp, step, center

The step function (step) is the function 1, if x ≥ 0 and 0 if x < 0.
Similarly, the ramp function (ramp) is x, if x ≥ 0 and 0 if x < 0.

The center function centers the variable with respect to its mean value.

Restriction: Only functions of one variable can be defined.

Use Spline

Check for using a spline function instead of formulas.

Def Spline...

Press to define the spline function.

Chebyshev-Polynomial Editor

Define a Chebyshev polynomial. The variable is normalized to take values between -1 and +1. Chebyshev polynomials are particularly suited to capture the influence of the water level. In contrast to ordinary polynomials, Chebyshev polynomials are almost orthogonal and do no lead to high VIF's.

Button Label

Text displayed on the button of the regression editor. Should be short enough to fit in the button.

Name

Text used as the name for the regressor. This should be a relatively short name to identify the regressor in the regression table. The letter $ is substituted by the actual variable name.

Order

The order of the Chebyshev polynomial.

Use Spline

Check for using a spline function instead of formulas.

Def Spline...

Press to define the spline function.

Creep-Function Editor

Define a function that models the influences of creep. The function is evaluated by a convolution integral. The input variable for the creep function is typically a polynomial that models the elastic behaviour. If **Creep** is checked in the Level tab of the **Regressor Editor,** the creep transformation is applied to all selected Transformations.

Button Label

Text displayed on the button of the regression editor. Should be short enough to fit in the button.

Name

Text used as the name for the regressor. This should be a relatively short name to identify the regressor in the regression table. The letter $ is substituted by the actual variable name.

Creep Parameter

Parameter used for the creep calculation. A typical value for concrete is 0.01/day.

Startup Time

Time period which has to precede the regression to make sure the influence of initial conditions is negligible. Taking a large startup time is generally not desirable because it reduces the regression period.

Convolution Time

Longest history that is considered in the convolution integral. If the value is set to 0, the whole past is always included. Considering only a certain time period of the past can considerably speed up the calculation.

Temperature-Function Editor

Define a function that evaluates the temperature at a certain depth from the temperature time history at the surface. The calculation is performed using heat conduction in a semi-infinite domain. The function is evaluated as a convolution integral.

Button Label

Text displayed on the button of the regression editor. Should be short enough to fit in the button.

Name

Text used as the name for the regressor. This should be a relatively short name to identify the regressor in the regression table. The letter $ is substituted by the actual variable name.

Depth

Depth at which the temperature is evaluated (meters).

Diffusivity

A typical value for concrete is 0.1 m^2/day.

Startup Time

Time period which has to precede the regression to make sure the influence of initial conditions is negligible. Taking a large startup time is generally not desirable because it reduces the regression interval.

Convolution Time

Longest history that is considered in the convolution integral. If the value is set to 0, the whole past is always included. Considering only a certain duration of the past can considerably speedup the calculation.

Average-Function Editor

Define a function for averaging the variable over some time in the past. This is a simple way to include effects of past history.

Button Label

Text displayed on the button of the regression editor. Should be short enough to fit in the button.

Name

Text used as the name for the regressor. This should be a relatively short name to identify the regressor in the regression table. The letter $ is substituted by the actual variable name.

Averaging Time

Time used for averaging. This is a minimum value. For non-uniform time steps, the actual time might be larger.

Delay-Function Editor

Define a time shift function. Each value is shifted by the selected number of time steps. The first value is repeated to fill the gap behind the shifting values.

Button Label

Text displayed on the button of the regression editor. Should be short enough to fit in the button.

Name

Text used as the name for the regressor. This should be a relatively short name to identify the regressor in the regression table. The letter $ is substituted by the actual variable name.

Delay Steps

Number of time steps used. The time shift is given by the number of time steps rather that the actual delay time.

Spline-Function Editor

Define a spline function by a number of data points. The spline function is a cubic polynomial between data points and has continuous slope and curvature at the data points. Outside of the boundary points, the function is defined by a linear function. The extrapolation range is restricted to an interval of the same size as the interval at the boundary.

X-Y Table

Define the data points as x-y pairs. X-Values must be increasing and there must be a least 3 data points. The table is increased dynamically as you type in.

Plot

Press the Plot button to see the spline defined by the values in the table.

Vertical x

Check to plot the x-variable on the vertical axis. This representation is often used if x is the height of the water level.

Clear

Press the Clear button to delete the data in the table. This is useful for restarting or if you don't need the spline anymore and don't want to save it.

Copy to Clipboard

Press 📋 to copy the Plot to clipboard.

Additional Reference

Data and Model Files

Input Data

Input files may be either Excel files (*.xls), or text files with tab-separated values (*.txt) or comma-separated (*.csv) values. If an Excel file contains more that one worksheet, a dialog shows up, where you have to select the desired worksheet.

Each data column has to have a text on top, describing the variable. Optionally, several title lines can precede the row with the variable descriptions. These lines are ignored by the program. The program generates an index column starting with 1 at the first data row. One column should contain the dates and optionally the time for each row. If no date column is found, the generated index column is used for the time. If more than one column contains date values, the first one is selected automatically. For each column containing date values, the program generates a column with the number of days, starting with one at the first data row.

The directory DamReg\Data\InputTest contains several example files.

A simple algorithm is used to determine the first data line and the date column. The program looks for the first date value and the first floating point value within the first 20 lines. If a date is found, the corresponding row and column are selected as the first data row and the date column, respectively. If no date is found, the row with the first floating point number is selected as the first data row. The first data line and the date column can be changed by the user by pressing the **Format** button and modifying the corresponding values.

There can be empty fields (separated by tabs or another delimiter) for missing values. Input data may have non-uniform time steps, but some time series calculations, like the autocorrelation function, assume more or less equally spaced data. Non-uniform time steps are also introduced by missing values.

Input Format

For Excel input files, dates, times and floating point values are format-free and are shown in the Input Data Window in the format specified by the Windows Regional Options. Since the same formats are used for reading, no problems are expected in this case. For text input files, the formats may differ from the format specified by Windows regional settings. The program attempts to extract date, time, and decimal separators from the text and use them for the input format.

To determine the date separator, the input is scanned for the first occurrence of the form <n1><d><n2><d><n3>, where <n1>, <n2>, <n3> are any numbers and <d> are two identical single-character date separators (for example 31.12.2004). Optionally, this pattern may be followed by <n4><t><n5> or <n4><t><n5><t><n6>, where <n4>, <n5>, <n6> are any numbers and <t> are the time separators (for example 12:30 or 12:30:00). The date format, that is the order of the day, month, and year cannot be determined in this way but is taken from the Windows Regional Options. The selected format is indicated at the top of the Input Data window (for example dd.MM.yyyy).

To determine the floating point separator, the first occurrence of the pattern <n1><p><n2> is searched, where <n1> and <n2> are any numbers and <d> is any single-character decimal separator (for example 10.2).

Separators that cannot be determined in this way are set according to the Windows Regional Options. The simple scheme should work most of the time, except if the day, month, and year are in a different order. In this case the user can change the format by pressing the **Format** button and modifying the corresponding values. Note that changes to the input format affect only the reading of the input data, not the formatting of the output.

Saving a Model

File names appearing in the save dialog are composed of the model name as appearing at the top right of the main window and the response variable. To use only the model name without the response variable, uncheck "File names with variable" in the **Edit | Options** menu.

As long as neither the model name nor the response variable have changed, the current file name as appearing in the main window title is used for saving without prompt. When the model name is changed, or when the variable name is changed and the option "File names with variable" is checked, you will be prompted for a new file name upon saving.

Models are saved in XML-format. You can view the content of an XML-file with any text editor or with Internet Explorer. Double clicking XML-files opens Internet Explorer and not DamReg.

Opening a Model

Model are stored in XML files.

When a model is opened, the data are read from the original data file, regressors are transformed and the regression is calculated. To make sure the input data are the same as the original data, the first date, the names of the variables and the regression coefficients are checked. A message is shown, if there are any differences.

When the model's data file cannot be found (e.g. if its directory has changed), the user is prompted to select the file manually. You can also enforce manual selection of the data file by choosing Files of Type: "Model (*.txt), **Ask for Data File**" in the opening dialog for the model.

Time Periods

Use the green and red **Plot Cursors** available in the Regression Plot to select the **Regression Period.** For exact values you can also select the dates in the two boxes above the Plot Region labelled **From** and **To**, respectively. Note that the begin of the Regression Period is restricted by the startup time of the active regressors.

Startup Period

Some transformations depend on values from previous time steps (i.e. temperatures calculated by heat conduction). A corresponding regressor is therefore only correct after a certain time from the beginning. This time is called startup time and is defined for all functions used for transformations. For most functions it is zero, for other functions it is defined implicitly or it can be defined by the user.

The Startup Period is the period before the Regression Period where regressors are calculated but not used in the regression. The Startup Period is chosen by the program and its length is at least the maximum startup time of all active regressors. If possible, it also covers the startup times of the inactive regressors.

Typically you start with a model that includes all regressors and the Startup Period is chosen based on the maximum startup time of all regressors. When you eliminate some regressors, the Startup Period is not changed, even if this would be possible. This leads to the correct p-values for the inactive regressors and saves some computations, because the regressors don't have to be recalculated. However, the user can selected an earlier start of the Regression Period, if the startup time of the active regressors allows that. The regressors are recalculated with a shorter Startup Period in this case. The startup time of the inactive regressors is not respected, but the p-values still give an indication of their significance. If the regressors are selected again, the Startup Period is adjusted accordingly if necessary, and the regressors are recalculated.

In the plots, the Startup Period is identified by a dark grey background.

Regression Period

This is the period used for the regression. The begin of the Regression Period is restricted by the startup times of the active regressors. In the plots, the Regression Period is identified by a light grey background.

Prediction Period

This is the period after the Regression Period until the end of input data.

Missing Values

The input file can contain missing values, represented by a blank or empty field between tab characters. In the calculations, missing values are not treated as zero but are ignored. When transforming variables or combining variables, a missing value results if any of the values is missing for a particular case (time).

For functions involving time shifts, the rule above is somewhat modified. There are two possibilities when performing a time shift: the gap can be moved with the values or be fixed at its position. Example:

Moving gap		Fixed gap	
1		1	
2	1	2	1
–	2	–	–
4	–	4	2
5	4	5	4
	5		5

The difference is only local. The moving gap seems to be more consistent, but results in two missing values when the original and the shifted sequence are involved. The fixed gap, on the other hand, results in only one missing value for the combination.

For the Durbin-Watson statistic, the fixed gap scheme is recommended in the literature and also used in the program. For the autocorrelation function, the moving gap scheme is used, because it also gives a value for time lags corresponding to a missing values. For the delay function, the fixed gap is used, mainly because it better fits the general implementation of transformation functions. The choice is somewhat arbitrary but should not make a large difference, when only a few values are missing.

For the regression analysis, cases with missing values are deleted, i.e. if any of the regressors has a missing value at a particular time, all values at that time are neglected. This scheme also involves inactive regressors, because a p-value is calculated for them. When the final model is selected, unnecessary deletions can be avoided, by removing the inactive regressors in the Regressor Editor (**Remove** □).

The same rules as for the regression analysis also apply to the correlation matrix. (Active and inactive regressors are included in the correlation matrix.)

The number of cases deleted due to missing values is shown as number of Missing Values above the Regression Table and in the Report.

In plots, missing values are shown as gaps. Because lines are plotted between points, a single missing value results in a gap of two time steps. Also, a single value between missing values is not plotted (no connecting line).

Algorithms

Standard Regression

Use the standard regression as the first algorithm for an analysis. Only if there are problems with a strong multicollinearity (indicated by high VIF's) use ridge regression or principal component analysis. If the Durbin-Watson statistic (DW) or the autocorrelation function (ACF) plot indicate an autocorrelation of the residuals, use the Prais-Winsten algorithm.

The standard regression is implemented with the QR-algorithm. Regressors are not normalized.

Ridge Regression

Ridge regression is useful, when there is a strong multicollinearity, indicated by high VIF's. In this case, the linear system of equations is ill-conditioned and should be regularized. While the QR-algorithm is very tolerant with respect to ill-conditioning, the undesirable effect is that the step-wise elimination of regressors does not necessarily lead to a good model.

This behaviour can be improved with the ridge regression. The results of step-wise elimination of regressors may still depend on the biasing parameter k, so it's a good idea to try several values. The values should be small in any case, at the order of 0.001 to 0.1.

Ridge regression is implemented with the QR-algorithm. Regressors are normalized to zero mean and unit variance. The biasing parameter is considered in the classical way by introducing additional rows in the X-matrix and the y-vector. This has the effect that t-statistics before and after the selection of a regressor are not identical. For a biasing parameter k=0, the results should be identical to the standard regression.

Principal Component Analysis

Principal component analysis is an algorithm, targeted to remove a strong multicollinearity (indicated by high VIF's). The Matrix X'X is basically decomposed into it's eigenvectors. Eigenvectors with small eigenvalues contribute to the multicollinearity and the high VIF's. Not including these eigenvectors improves the behaviour of the analysis. Each neglected eigenvector can be considered as an additional constraint to the system of linear equations.

Experience has shown that for step-wise elimination, the principal component analysis works not as well as expected. However, it is useful to remove a multicollinearity at the final stage of model selection.

Principal Component Analysis is implemented with the singular value decomposition. Regressors are normalized to zero mean and unit variance. The singular values and the corresponding eigenvectors can be viewed in a separate table (menu **View | Eigenvectors**). Singular values are the square roots of the eigenvalues.

Prais-Winsten Algorithm

The Prais-Winsten algorithm should be used if the residuals are autocorrelated. Autocorrelation can be discovered by the Durbin-Watson statistic and by the plot of the autocorrelation function (ACF). If the autocorrelation is of first order, it can be removed by a differencing scheme applied to the regressors and to the response variable. Note that not all types of autocorrelation can removed this way, only the first-order autocorrelation. For other kinds of autocorrelation, more advanced algorithms are needed which, however, are not implemented in this version.

The parameter entering the differencing scheme is ρ. It can take a value $|\rho| < 1$ (strictly less than 1). For a positive autocorrelation, ρ is close to one and for a negative autocorrelation it is close to minus one. The procedure calculates an estimate for ρ and the best value has to evaluated iteratively until convergence (by pressing the Regression button several times). The parameter ρ estimated in the standard algorithm is a good starting value.

Results obtained from the Prais-Winsten algorithm have to be interpreted somewhat differently than those for other algorithms. All regression results in the main table and the summary table relate to the transformed regressors. However, regression coefficients, t-statistics and p-values are also the valid for the original regressors. The coefficient of multiple determination R^2, the residual sum of squares SS_{Res}, and the residual mean square MS_{Res} relate only to the transformed problem. Equally, the Durbin-Watson statistic, the plot of the residuals vs. the predicted values, the plot of the probability distribution of the residuals, and the plot of the autocorrelation function are calculated with the residuals of the transformed problem.

To the right of the algorithm selection, the program also shows a value for MS_{Res} which is an estimate for the residual mean square of the original problem before differencing. This values is also used to calculate the confidence interval in the plots.

Statistics

t-Statistic and p-Value

The t-statistic and its corresponding p-value determines whether a regression coefficient is significantly different from zero. If the p-value is above some limit value (usually 1%), a regressor has a regression coefficient not significantly different from zero and should, therefore, be eliminated from the model.

p-values are determined from the t-statistic using the Student distribution. Small absolute values of the t-statistic correspond to high p-values and vice versa. The t-statistic is also an indication of the increase of the residual sum of squares SS_{Res} when the regressor is eliminated. A regressor with a small t-statistic is not only insignificant from a statistical point of view, but also leads to a small increase of the residuals when eliminated.

The program also calculates t-statistics and p-values for inactive regressors. These numbers are shown in parenthesis and give the values that would be obtained when the regressor were included in the model.

Eliminating or selecting regressors based on the p-values is the basic technique for finding an adequate model. The selection can either be performed manually or automatically using the **Eliminate**, **Select** and **Auto** buttons.

If the residuals are autocorrelated, the p-values can be wrong in the statistical sense. In this case, regressors have often small p-values, although they are not statistically significant. However, the relative importance is still maintained and the t-statistic still indicates the increase in the residual sum of squares when the regressor is eliminated.

Variance Inflation Factor VIF

The variance inflation factor VIF is the single most important number for detecting a multicollinearity. Values above a certain limit indicate a strong multicollinearity. The literature often gives 10 as the limit, although examples with dams show that values of 100 are still acceptable.

A multicollinearity means that two or several regressors are close to linearly dependent. The resulting regression coefficients are not very dependable, although the prediction might still be valid, as long as the collinearity is the same in the prediction period. Also the process of successive elimination of insignificant regressors does often not lead to the optimal model in the presence of a multicollinearity.

Often, a multicollinearity is caused by similar measurements (temperatures) and can be removed by neglecting some of the regressors. Other regressors that often cause a multicollinearity are regressors obtained from the same variable, e.g. polynomials of the water level. The multicollinearity obtained by polynomials can easily be avoided by centring (transform to zero mean) the variables before calculating the polynomials, or even more effectively by using Chebyshev polynomials.

Often a multicollinearity occurs if there are too many regressors. Many times it is not present anymore in the final reduced model. The process of successive elimination can be stabilized by introducing a small biasing parameter in the ridge regression. Several trials with different biasing parameters should be performed to obtain the best model.

Prediction Interval and Extrapolation

The plots also show a prediction interval and indicate points of extrapolation beyond the region containing the original observations.

The prediction interval is similar to the often used interval defined by 2 or 3 times the standard deviation (95% or 99.7% confidence). Additionally to this variation, the prediction interval considers the variation of the regression coefficients.

For several regressors, a prediction point is an extrapolation, if the regressor point is outside of an ellipsoid containing all regressor points used in the regression. An extrapolation can occur even if all regressors are in the range of their original values, but the combination is outside of the original combinations.

Durbin-Watson Test

The Durbin-Watson test indicates whether the residuals are autocorrelated. Autocorrelation of residuals violates the fundamental assumption of the errors being uncorrelated. The calculated regression coefficients are no longer best linear unbiased estimators for the real regression coefficients. The p-values and t-statistics are not valid any more in a statistical sense but can still be used as relative values for the model selection.

The Durbin-Watson statistic takes values from 0 to 4. If the residuals are not autocorrelated, the value is close to 2, for a positive correlation it is close to 0 and for a negative correlation close to 4. Roughly speaking, if the Durbin-Watson statistic is between 1.5 and 2.5 the autocorrelation is acceptable. The exact limits, however, depend on the number of data and the number of regressors. The exact procedure and tables with the distributions are given in many statistics books.

References

Montgomery, Douglas C., Elizabeth A. Peck, G. Geoffrey Vining (2001), *Introduction to Linear Regression Analysis*, Third Edition, John Wiley.

Pindyck, Robert S. and Daniel L. Rubinfeld (1998), *Econometric Models and Economic Forecasts*, Fourth Ed., McGraw-Hill.

Kmenta, Jan (1997), *Elements of Econometrics*, 2nd Edition, The University of Michigan Press.

Benedikt Weber (2002), *Vorhersage des Verhaltens von Talsperren mit Hilfe des Soll-Ist-Vergleichs – Statistischer Teil,* Report on behalf of the Federal Office for Water and Geology (FOWG), Bienne, Switzerland.

Benedikt Weber (2002), *Details zur Implementierung*: *Nachtrag zum Bericht Vorhersage des Verhaltens von Talsperren mit Hilfe des Soll-Ist-Vergleichs – Statistischer Teil,* Report on behalf of the Federal Office for Water and Geology (FOWG), Bienne, Switzerland.