Linear Regression

Description

Linear regression is a model of dependence between **input** and **output variables** with the linear link function.

Linear regression is one of the most frequently used algorithms in **machine learning**. This algorithm frequently produces good results even for small data sets.

Wide use of linear regression is explained by the fact that many real processes in science, economics and business can be described in terms of linear models. For example, linear regression enables to estimate anticipated sales volume depending on the established price.

The handler can be used to accomplish different **Data Mining** tasks, for example, such as **forecasting** and numerical prediction.

To get resulting data sets, it is required to provide preliminary training of the handler.

> **Important:** Input data must never contain missing values, output data must not contain missing values during training.

Ports

Input

- — Input data source (data table) — required port.
- — Control variables (variables) — optional port. It is possible to set values of the wizard parameters using variables.

Outputs

- — Regression output: the table that consists of the following fields: field regression of output data, source data set.
- — Regression model coefficients: data table.
- — Summary: variables.

Node Wizard

The wizard includes the following groups of parameters:

- Normalization settings.
- Partitioning.
- Linear regression configuring.
- Detailed settings.

Partitioning

The *Partitioning* page of the wizard enables to divide a set into the training and test ones:

- **Train**: the structured data set used for training of **analytical models**. Each record of the training set is a training example with the set input effect and correct output (target) result that corresponds to it.
- **Test**: the training sample subset that contains test examples, namely, the examples used not to train the model but to check its results.

Available Parameters:

- Size of training and test set in percentage terms or in rows. It can be set by means of variables.
- Method of partition to training and test sets. There are two partition methods:
  - Random method provides partitioning of records set to training and test sets.
  - Sequence: groups of sets rows (training, unused, test) are selected in a sequential order, namely, only the records that are included into the first set are selected first, then it is required to select the records that are included into the second set, etc. It is possible to change the order of sets (*Move up*, *Move down* buttons).
- Validation method that can take the following values:
  - No validation.
  - **K-fold cross validation** enables to select the *Method of sampling* and number of *Cross validation fold*.
  - **Monte Carlo** enables to select *Resampling iteration count* and set the size of training and **validation set**.

Linear Regression Configuring

A set of parameters to configure the linear regression can be grouped in the following packs:

Configure Method

- Auto setup:
  - Boolean value. Enabled by default.
  - It has an impact on usage of the following packs of parameters: if it is enabled, it is possible to configure *Auto setup priority* pack, if it is disabled, it is possible to select

the algorithm of **factors** selection and protection against overfitting, and it is also possible to set priorities.

- Auto setup priority:
  - It has an impact on selection of the particular method and its settings according to the Accuracy - Speed scale.
  - Integer type. It can take the following values:
    - Maximum accuracy.
    - Increased accuracy.
    - Average speed.
    - Increased speed.
    - Maximum speed.
- **Denormalize model coefficients**: denormalization is required for interpretation of results. As the model can work only with the normalized data, first, it is required to normalize data that has been sent to the model for its usage, and then denormalization must be performed to make data return the same kind it has had before normalization. It is a boolean value, enabled by default.

Configure Parameters

It is used if *Auto setup* checkbox is not selected, or it is set by means of variable.

- **Factor selection and protection against overfitting** - value of enumeration:
  - **Enter**: enter all set indicators into the regression model irrespective of the fact whether they have meaningful influence or not.
  - **Forward**: this method is based on the following principle: it is required to start from absence of indicators and gradually find the "best" ones that will be added to the subset.
  - **Backward**: this method is based on the following principle: it is required to start from all available indicators and exclude the "worst" ones by means of successive iterations.
  - **Stepwise**: modification of the *Forward* method except that at each step upon entering of the new variable into the model, other variables that have already been entered into it earlier are tested for significance.
  - **Ridge** is one of the methods used for **dimension decrease**. It is used to avoid data redundancy when independent variables **correlate** with each other (**multicollinearity**) that causes instability of estimates of linear regression coefficient.
  - **LASSO** is used to avoid data redundancy as *Ridge*.
  - **Elastic-Net**: regression model with two regularizers - *L1*, *L2*. *LASSO L1 = 0* and *Ridge* of *L2 = 0* regression are the models that represent special cases. Both regularizers help to improve generalization and errors of test as they protect the model against overfitting in connection with data noise:
    - **L1** implements it by selecting the most important factors that have the highest impact on the result.
    - **L2** prevents from the model overfitting by forbidding disproportionately large weighted coefficients.
- Accuracy/speed priority.
  - Integer type. It can take the following values:

        - Maximum accuracy.
        - Increased accuracy.
        - Average speed.
        - Increased speed.
        - Maximum speed.
- Exact/inexact data priority.
    - Integer type. It can take the following values:
        - Accurate data.
        - Increased accuracy.
        - Average accuracy.
        - Reduced accuracy.
        - Unreliable data.
- Less/more factors priority.
    - Integer type. It can take the following values:
        - Minimum factors.
        - Less factors.
        - Average number of factors.
        - More factors.
        - Maximum factors.

The following options are available for different methods:

Method	Accuracy/speed priority	Exact/inexact data priority	Less/more factors priority
Enter	•		
Forward	•		•
Backward	•		•
Stepwise	•		•
Ridge	•	•	
LASSO	•		•
Elastic-Net	•	•	•

- **Use detailed settings** provides more detailed configuration of linear regression (additional wizard page appears - pack of detailed settings). It is a boolean value, disabled by default.

> **Note:** All available parameters of the linear regression configuration can be set by means of variables.

Detailed Settings

They are used if the parameters configuration pack is enabled, and *Use detailed settings* checkbox is selected in it, or it can be set by means of variable.

Detailed settings are joined into the following packs of parameters:

Method Settings

Available Parameters:

- **Solution accuracy**: criterion of iterations stop. This setting enables to define the accuracy of definition of the error function minimum. It is a real value from 0 to 1. It is an editor with value change interval equal to 0.000001.
- **Include intercept into the model** adds the dependent variable to the model.

Statistics Calculation Settings

Available Parameters:

- Calculate **confidence interval**.
- % confidence interval.
- Statistics calculation mode:
  - Do not calculate.
  - For all models.
  - For the final model.

Regularization Settings

Available Parameters:

- **L1-regularization coefficient setup**: configuration of this parameter is possible only for *LASSO*, *Elastic-Net* algorithms.
- **L2-regularization coefficient setup**: configuration of this parameter is possible only for *Ridge*, *Elastic-Net* algorithms.

It is possible either to select auto setup of value for each of parameters, or to enter the required value in the manual way.

Factor Selection Settings

Available Parameters:

- **Factor selection criterion** enables to select one of the following information criteria:
  - **F-test**.
  - **Determination coefficient**.
  - **Adjusted determination coefficient**.
  - **Akaike information criterion**.

    - **Akaike information criterion corrected**.
    - **Bayesian information criterion**.
    - **Hannan-Quinn information criterion**.
- **Significance threshold** in the case of factor addition.
- **Significance threshold** in the case of factor exception.

> **Note:** All available parameters of the detailed settings can be set by means of variables.

Articles in Section:

- Report
- Regression Coefficients