root

Go Up

Name	LinearRegression
Version	3.0.0
Description	Linear Regression Algorithm Bundle
License	http://www.apache.org/licenses/LICENSE-2.0
Copyright	Copyright (C) 2017 HPCC Systems
Authors	HPCCSystems
DependsOn	ML_Core, PBblas
Platform	6.2.0

Table of Contents

OLS

Go Up

IMPORTS

ML_Core | ML_Core.Types | PBblas | PBblas.Types | PBblas.Converted | PBblas.MatUtils | ML_Core.Math |

DESCRIPTIONS

MODULE OLS

	OLS
	`(DATASET(NumericField) X=empty_data, DATASET(NumericField) Y=empty_data)`

Ordinary Least Squares (OLS) Linear Regression aka Ordinary Linear Regression.

Regression learns a function that maps a set of input data (independents) to one or more output variables (dependents). The resulting learned function is known as the model. That model can then be used repetitively to predict (i.e. estimate) the output value(s) based on new input data. Two major use cases are supported:

1. Learn and return a model.
2. Use an existing (e.g. persisted) model to predict new values for Y.

Of course, both can be done in a single run. Alternatively, the model can be persisted and used indefinitely for prediction of Y values, as long as the record format has not changed, and the original training data remains representative of the population.

OLS supports any number of independent variables (Multiple Regression) and multiple dependent variables (Multivariate Regression). In this way, multiple variables' values can be predicted from the same input (i.e. independent) data.

Training data is presented as parameters to this module. When using a previously persisted model (use case 2 above), these parameters should be omitted.

This module provides a rich set of analytics to assess the usefulness of the resulting linear regression model, and to determine the best subset of independent variables to include in the model. These include:

- For the whole model:
  - Analysis of Variance (ANOVA)
  - R-squared
  - Adjusted R-squared
  - F-Test
  - Akaike Information Criterion (AIC)
- For each coefficient:
  - Standard Error (SE)
  - T-statistic
  - P-value
  - Confidence Interval

PARAMETER **X** ||| TABLE (NumericField) — The independent variable training data in DATASET(NumericField) format. Each observation (e.g. record) is identified by 'id', and each feature is identified by field number (i.e. 'number'). Omit this parameter when predicting from a persisted model.

PARAMETER **Y** ||| TABLE (NumericField) — The dependent variable training data in DATASET(NumericField) format. Each observation (e.g. record) is identified by 'id', and each feature is identified by field number. Omit this parameter when predicting from a persisted model.

RETURN — The instantiated model.

SEE ML_Core.Types.NumericField

PARENT **ML_Core.Interfaces.IRegression**
</home/tetrapod/pcsource/ML_Core/Interfaces/IRegression.ecl>

Children

1. GetModel : Return the learned model that maps Independent X variables to Dependent variable(s) Y
2. Betas :

   Extract Beta values from the model
3. Predict : Predict the dependent variable values (Y) for any set of independent variables (X)
4. RSquared : Calculate the R-Squared Metric used to assess the fit of the regression line to the training data
5. AnovaRec : Record layout for the information returned from calcAnova
6. Anova : ANOVA (Analysis of Variance) report
7. SE : Compute the Standard Error of the Regression Coefficients
8. TStat : Compute the T-Statistic
9. AdjRSquared : Calculate Adjusted R Squared, a normalized version of R Squared that does not arbitrarily increase with the number of features
10. AICRec : Record layout for return from AIC function
11. AIC : Calculate the Akaike Information Criterion (AIC)
12. pVal : Calculate the P-value for each coefficient, which is the probability that the coefficient is insignificant (i.e
13. ConfintRec : Record layout for the return from ConfInt
14. ConfInt : Compute the Confidence Interval for each coefficient
15. FTestRec : The record layout for the return from Ftest
16. FTest : Perform an F-test

ATTRIBUTE GetModel

OLS \

DATASET(Layout_Model)	**GetModel**

Return the learned model that maps Independent X variables to Dependent variable(s) Y.

In the case of OLS, the model represents a set of Betas which are the coefficients of the linear model: Beta0 * 1 + Beta1 * Field1 + Beta2 * Field2 . . . The ID of each model record specifies to which Y variable the coefficient applies.

The Field Number ('number') indicates to which field of X the beta is to be applied. Field number 1 provides the intercept portion of the linear model and is always multiplied by 1.

Note that if multiple work-items are provided within X and Y, there will be multiple models returned in one dataset. The models can be separated by their work item id (i.e. 'wi'). A single model can be extracted from a myriad model by using e.g., model(wi=myWI_id).

GetModel should not be called when predicting using a previously persisted model (i.e. when training data was not passed to the module).

Example:

```
myModel := OLS(X, Y).GetModel;
```

RETURN **TABLE ({ UNSIGNED2 wi , UNSIGNED8 id , UNSIGNED4 number , REAL8 value })** — Model in DATASET(Layout_Model) format.

SEE ML_core/Types.Layout_Model

OVERRIDE

FUNCTION Betas

OLS \

DATASET(NumericField)	**Betas**
(DATASET(Layout_Model) model=GetModel)	

Extract Beta values from the model. Can be used during training and prediction phases.

For use during training phase, the 'model' parameter can be omitted. GetModel will be called to retrieve the model based on the training data.

For use during prediction phase, a previously persisted model should be provided.

The 'number' field of the returned NumericField records specifies to which Y the coefficient applies.

The 'id' field of the returned record indicates the position of the Beta value. ID = 1 provides the Beta for the constant term (i.e. the Y intercept) while subsequent values reflect the Beta for each correspondingly numbered X feature. Feature 1 corresponds to Beta with 'id' = 2 and so on.

If 'model' contains multiple work-items, Separate sets of Betas will be returned for each of the 'myriad' models (distinguished by 'wi').

PARAMETER **model** ||| TABLE (Layout_Model) — Optional parameter provides a model that was previously retrieved using GetModel. If omitted, GetModel will be used as the model.

RETURN **TABLE ({ UNSIGNED2 wi , UNSIGNED8 id , UNSIGNED4 number , REAL8 value })** — DATASET(NumericField) containing the Beta values.

SEE ML_Core.Tyeps.NumericField

FUNCTION Predict

OLS \

DATASET(NumericField)	**Predict**
(DATASET(NumericField) newX, DATASET(Layout_Model) model=GetModel)	

Predict the dependent variable values (Y) for any set of independent variables (X). Returns a predicted Y values for each observation (i.e. record) of X.

This attribute supports the 'myriad' style interface in that multiple independent work items may be present in 'newX', and multiple independent models may be provided in 'model'. The resulting predicted values will also be separable by work item (i.e. wi).

PARAMETER **newX** ||| TABLE (NumericField) — The set of observations of independent variables in DATASET(NumericField) format.

PARAMETER **model** ||| TABLE (Layout_Model) — Optional. A model that was previously returned from GetModel (above). Note that a model from a previous run will only be valid if the field numbers in X are the same as when the model was learned. If this parameter is omitted, the current model will be used.

RETURN **TABLE ({ UNSIGNED2 wi , UNSIGNED8 id , UNSIGNED4 number , REAL8 value })** — An estimation of the corresponding Y value for each observation of newX. Returned in DATASET(NumericField) format with field number (i.e. 'number') indicating the dependent variable that is predicted.

OVERRIDE

ATTRIBUTE RSquared

OLS \

DATASET(R2Rec)	**RSquared**

Calculate the R-Squared Metric used to assess the fit of the regression line to the training data.

Since the regression has chosen the best (i.e. least squared error) line matching the data, this can be thought of as a measurement of the linearity of the training data.

R Squared generally varies between 0 and 1, with 1 indicating an exact linear fit, and 0 indicating that a linear fit will have no predictive power. Negative values are possible under certain conditions, and indicate that the mean(Y) will be more predictive than any linear fit.

Moderate values of R squared (e.g. .5) may indicate that the relationship of X -> Y is non-linear, or that the measurement error is high relative to the linear correlation (e.g. many outliers). In the former case, increasing the dimensionality of X, such as by using polynomial variants of the features, may yield a better fit.

R squared always increases when additional independent variables are added, so it should not be used to determine the optimal set of X variables to include. For that purpose, use Adjusted R Squared (below) which penalizes larger numbers of variables.

Note that the result of this call is only meaningful during training phase (use case 1 above) as it is an analysis based on the training data which is not provided during a prediction-only phase.

RETURN **TABLE ({ UNSIGNED2 wi , UNSIGNED4 number , REAL8 RSquared })** — DATASET(R2Rec) with one record per dependent variable, per work-item. The number field indicates the dependent variable and coresponds to the number field of the dependent (Y) variable to which it applies.

RECORD AnovaRec

OLS \

	AnovaRec

Record layout for the information returned from calcAnova.

FIELD **Model_SS** ||| REAL8 — The Sum-of-Squares variance explained by the model.

FIELD **Model_F** ||| REAL8 — The F-test statistic.

FIELD **Error_DF** ||| UNSIGNED8 — Degrees of freedom of the error.

FIELD **Error_MS** ||| REAL8 — Mean square residual error.

FIELD **Model_DF** ||| UNSIGNED8 — Degrees of freedom of the model.

FIELD **Error_SS** ||| REAL8 — The Sum-of-Squares variance not explained by the model (aka Residual Sum-of-Squares).

FIELD **Model_MS** ||| REAL8 — Mean square variance of the model.

FIELD **Total_SS** ||| REAL8 — The total Sum-of-Squares variance.

FIELD **Total_MS** ||| REAL8 — Mean square variance of the data.

FIELD **wi** ||| UNSIGNED2 — The work-item number

FIELD **Total_DF** ||| UNSIGNED8 — Degrees of freedom in the data.

FIELD **number** ||| UNSIGNED4 — The dependent field (i.e. regressor) number.

ATTRIBUTE Anova

OLS \

	Anova

ANOVA (Analysis of Variance) report.

Analyze the sources of variance.

Basic ANOVA equality: Model + Error = Total

Determines how much of the variance of Y is explained by the regression model, versus how much is due to the error term (i.e. unexplained variance).

This attribute is only meaningful during the training phase.

Provides one record per work-item. Each record provides the following statistics:

- Total_SS – Total Sum of Squares (SS) variance of the dependent data.
- Model_SS – The SS variance represented within the model.

- Error_SS – The SS variance not reflected by the model (i.e. Total_SS - Error_SS).
- Total_DF – The total degrees of freedom within the dependent data.
- Model_DF – Degrees of freedom of the model.
- Error_DF – Degrees of freedom of the error component.
- Total_MS – The Mean Square (MS) variance of the dependent data.
- Model_MS – The Mean Square (MS) variance represented within the model.
- Error_MS – The MS variance not reflected by the model.
- Model_F – The F-Test statistic: Model_MS / Error_MS.

RETURN **TABLE ({ UNSIGNED2 wi , UNSIGNED4 number , REAL8 Total_SS , REAL8 Model_SS , REAL8 Error_SS , UNSIGNED8 Total_DF , UNSIGNED8 Model_DF , UNSIGNED8 Error_DF , REAL8 Total_MS , REAL8 Model_MS , REAL8 Error_MS , REAL8 Model_F })** — DATASET(AnovaRec), one per work-item per dependent (Y) variable The number field indicates the dependent variable to which the analysis applies.

SEE AnovaRec

ATTRIBUTE SE

OLS \

DATASET(NumericField)	**SE**

Compute the Standard Error of the Regression Coefficients.

Describes the variability of the regression error for each coefficient.

Only meaningful during the training phase.

RETURN **TABLE ({ UNSIGNED2 wi , UNSIGNED8 id , UNSIGNED4 number , REAL8 value })** — DATASET(NumericField), one record per Beta coefficient per dependent variable per work-item. The 'id' field is the coefficient number, with 1 being the Y intercept, 2 being the coefficient for the first feature, etc. The 'number' field indicates the dependent variable to which the coefficient applies.

ATTRIBUTE TStat

OLS \

DATASET(NumericField)	**TStat**

Compute the T-Statistic.

The T-statistic identifies the significance of the value of each regression coefficient. Its calculation is simply the value of the coefficient divided by the Standard Error of the coefficient. A larger absolute value of the T-statistic indicates that the coefficient is more significant.

Only meaningful during the training phase.

RETURN **TABLE ({ UNSIGNED2 wi , UNSIGNED8 id , UNSIGNED4 number , REAL8 value })** — DATSET(NumericField), one record per Beta coefficient per dependent variable per work-item. The 'id' field is the coefficient number, with 1 being the Y intercept, 2 being the coefficient for the first feature, etc. The number field indicates the dependent variable to which the coefficient applies.

ATTRIBUTE AdjRSquared

OLS \

DATASET(R2Rec)	**AdjRSquared**

Calculate Adjusted R Squared, a normalized version of R Squared that does not arbitrarily increase with the number of features.

Adjusted R2, rather than R2 should always be used when trying to determine the best set of features to include in a model. When adding features, R2 will always increase, whether or not it improves the predictive power of the model. Adjusted R2, however, will only increase with the predictive power of the model.

RETURN **TABLE ({ UNSIGNED2 wi , UNSIGNED4 number , REAL8 RSquared })** — DATASET(R2Rec), one record per dependent variable per work-item. The number field indicates the dependent variable and corresponds to the number field of the dependent (Y) variable to which it applies.

RECORD AICRec

OLS \

	AICRec

Record layout for return from AIC function.

FIELD **number** ||| UNSIGNED4 — The field-number of the dependent variable (i.e. regressor).

FIELD **AIC** ||| REAL8 — The Akaike Information Criteria value.

FIELD **wi** ||| UNSIGNED2 — The work-item number.

ATTRIBUTE AIC

OLS \

DATASET(AICRec)	AIC

Calculate the Akaike Information Criterion (AIC).

AIC is an Information Theory based criterion for assessing Goodness of Fit (GoF).

Lower values mean better fit.

RETURN **TABLE ({ UNSIGNED2 wi , UNSIGNED4 number , REAL8 AIC })** — DATASET(AICRec), one record per dependent variable per work-item. The number field indicates the dependent variable and corresponds to the number field of the dependent (Y) variable to which it applies.

ATTRIBUTE pVal

OLS \

	pVal

Calculate the P-value for each coefficient, which is the probability that the coefficient is insignificant (i.e. actually zero).

A low P-value (e.g. .05) provides evidence that the coefficient is significant in the model.

A high P-value indicates that the coefficient value should, in fact, be zero.

P-value is related to the T-Statistic, and can be thought of as a normalized version of the T-Statistic.

Only meaningful during the training phase.

RETURN **TABLE ({ UNSIGNED2 wi , UNSIGNED8 id , UNSIGNED4 number , REAL8 value })** — DATSET(NumericField), one record per Beta coefficient per dependent variable per work-item. The 'id' field is the coefficient number, with 1 being the Y intercept, 2 being the coefficient for the first feature, etc. The number field indicates the dependent variable and corresponds to the number field of the dependent (Y) variable to which it applies.

RECORD ConfintRec

OLS \

	ConfintRec

Record layout for the return from ConfInt.

FIELD **LowerInt** ||| REAL8 — The lower bound of the interval.

FIELD **number** ||| UNSIGNED4 — The dependent field number (i.e. regressor).

FIELD **id** ||| UNSIGNED8 — The coefficient number (1 is the constant term).

FIELD **UpperInt** ||| REAL8 — The upper bound of the interval.

FIELD **wi** ||| UNSIGNED2 — The work-item number.

FUNCTION ConfInt

OLS \

	ConfInt
	(Types.t_fieldReal level)

Compute the Confidence Interval for each coefficient.

The Confidence Interval determines the upper and lower bounds of each estimated coefficient given a confidence level (level) that is required.

For example, one could say that there is a 95% probability (level) that the coefficient of the first independent variable is between 2.05 and 3.62. This allows error margins to be determined with the desired confidence level. If the confidence interval spans zero, it implies that the coefficient may not be significant at the specified confidence level.

PARAMETER **level** ||| REAL8 — The level of confidence required, expressed as a percentage from 0.0 to 100.0.

RETURN **TABLE ({ UNSIGNED2 wi , UNSIGNED8 id , UNSIGNED4 number , REAL8 LowerInt , REAL8 UpperInt })** — DATASET(ConfintRec) with one record per coefficient per dependent variable per work-item. The 'id' field is the coefficient number, with 1 being the Y intercept, 2 being the coefficient for the first feature, etc. The number field indicates the dependent variable and corresponds to the number field of the dependent (Y) variable to which it applies.

RECORD FTestRec

OLS \

	FTestRec

The record layout for the return from Ftest

FIELD **pValue** ||| REAL8 — The p-value for the full distribution.

FIELD **number** ||| UNSIGNED4 — The dependent field (i.e. regressor) number.

FIELD **Model_F** ||| REAL8 — The F-value.

FIELD **wi** ||| UNSIGNED2 — The work-item number.

ATTRIBUTE FTest

OLS \

DATASET(FTestRec)	**FTest**

Perform an F-test.

Calculate the P-value for the full regression, which is the probability that all of the coefficients are insignificant (i.e. actually zero).

A low P-value (e.g. .05) provides evidence that at least one coefficient is significant. A high P-value indicates that all the coefficient values should in fact be zero, implying that the regression has no statistically significant predictive power.

P-value is related to the ANOVA F-Statistic, and can be thought of as a standardized version of the ANOVA F-Statistic.

The F-Test and T-Test are similar, except that the T-test is used to test the significance of each coefficient, while the F-Test is used to test the significance of the entire regression. For simple linear regression (i.e. only one independent variable, the T-Test and F-Test are equivalent.

RETURN **TABLE ({ UNSIGNED2 wi , UNSIGNED4 number , REAL8 Model_F , REAL8 pValue })** — DATASET(FTestRec), one record per dependent variable per work-item. The number field indicates the dependent variable and corresponds to the number field of the dependent (Y) variable to which it applies.
