Command Reference: FillRegression()

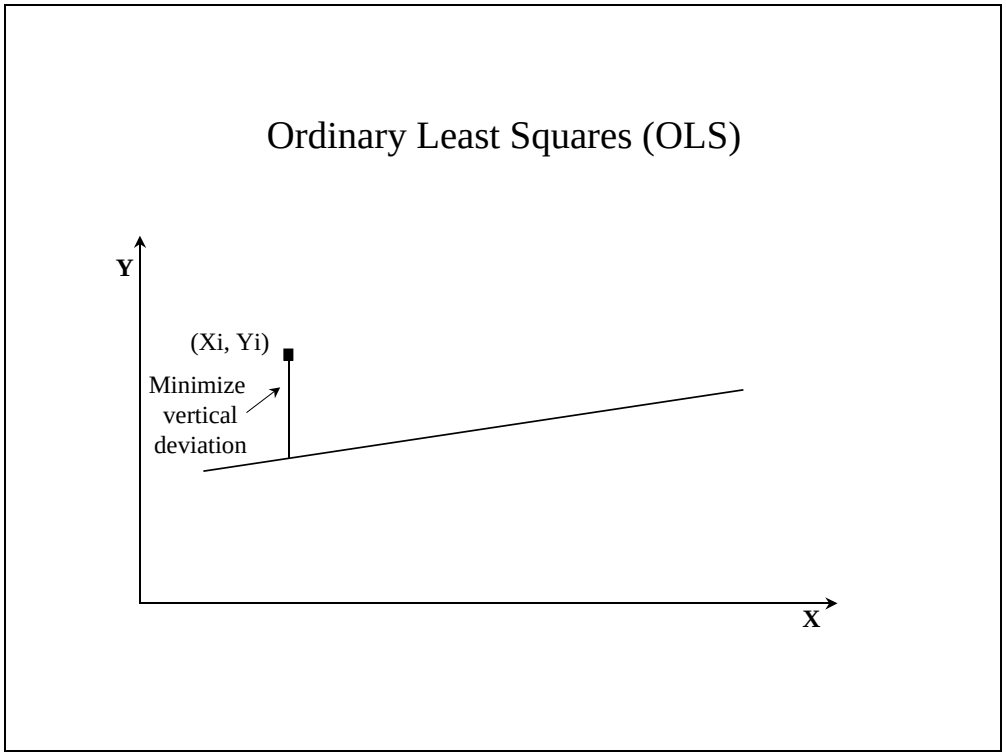
Fill missing time series data using ordinary least squares regression

Version 10.01.00, 2011-12-05

The `FillRegression()` command fills missing data in a time series using ordinary least squares (OLS) regression (see also the `FillMOVE2()` and `FillMixedStation()` commands). Regression can be applied only to regular interval time series. The first time series selected (dependent time series) will be filled using the other selected time series (independent time series). The periods of record and output period for the time series should be verified to make sure that the time series periods overlap sufficiently. The ***Results…Graph - XY-Scatter*** is a useful tool for reviewing data. Regression relationships are developed using the analysis period for the time series and are applied to the fill period. Refer to the output table, log file, and time series properties for analysis details. Several parameters are available to ensure that filling uses reasonable relationships.

Important: TSTool does allow filled values to be flagged. However, other commands do not exclude these values from computations when determining relationships for subsequent fill steps. Therefore, it is important to perform regression data filling as early in processing as possible so that earlier manipulation does not introduce derived values and bias.

In OLS regression, the vertical distance from the data point to the regression line is minimized. OLS regression provides the minimum-variance estimate for a single value or observation. However, if an ensemble of points is estimated from OLS regression, the estimated values will have lesser variability than the true values.



The following OLS equation is used to estimate values for the dependent time series from the independent time series:

$$Y_i = \overline{Y_1} + r\frac{S_{y1}}{S_{x1}}\left[X_i - \overline{X_1}\right]$$

or

$$Y_i = a + bX_i$$

where

N_1 = concurrent or overlapping period of record

$\overline{X}_1$ = mean for independent variable for N_1 years

$\overline{Y}_1$ = mean for dependent variable for N_1 years

S_{y1} = standard deviation for N_1 years

S_{x1} = standard deviation for N_1 years

$b = r\frac{S_{y1}}{S_{x1}}$, r = correlation coefficient

$a = \overline{Y}_1 - b\overline{X}_1$

Note that the correlation coefficient, r, is used to compute the slope, *b*, of the line.

A number of statistics are computed and are available for output to a table, as described below (see the `TableID` and related parameters). In the following descriptions, the statistic for one equation has a name like "Mean" and monthly statistics correspondingly have a name like "Mean_1", where 1=January.

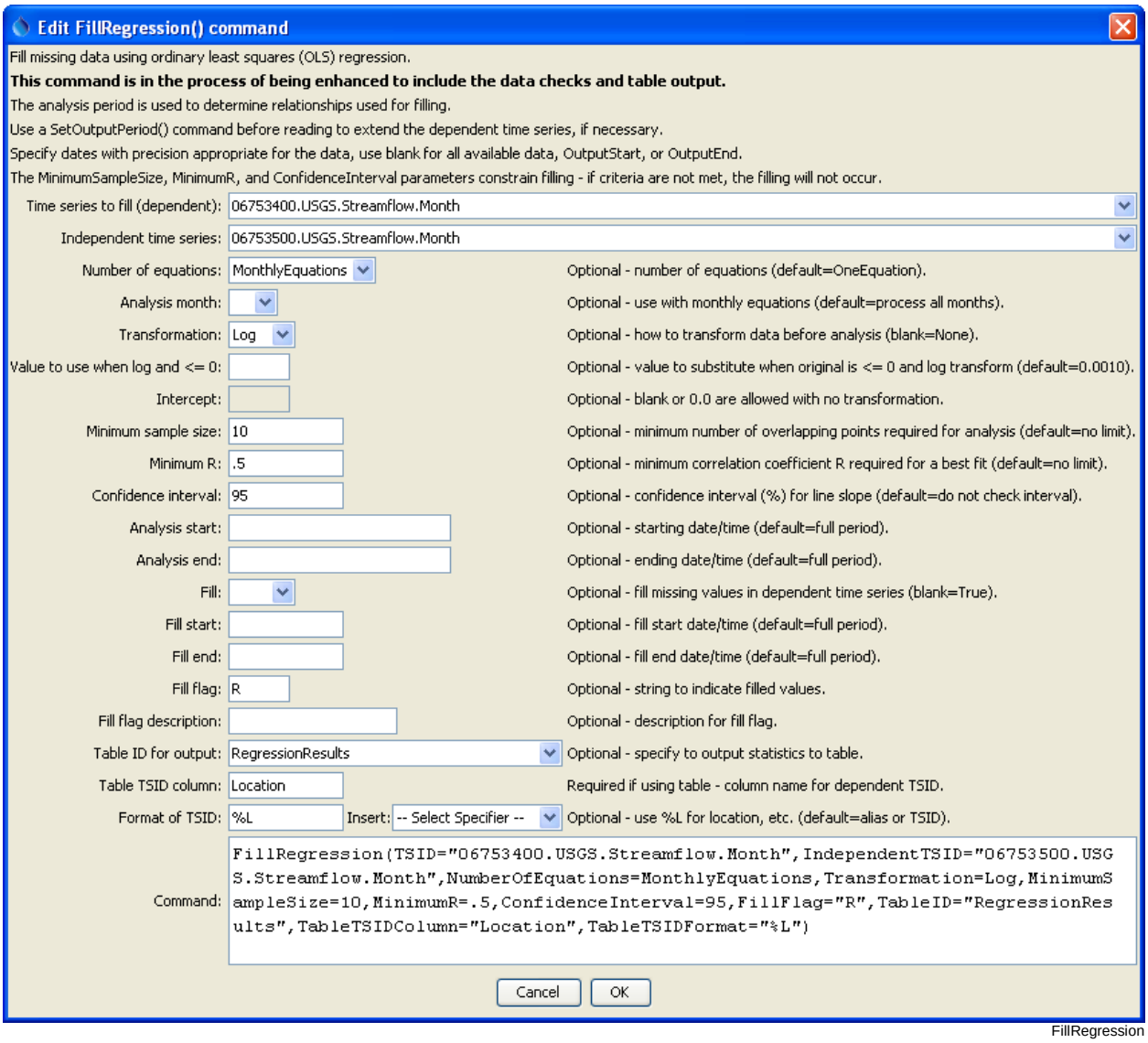
Statistics From Regression Analysis

Statistic (Table Column Name)	Dependent, Independent, or Both	Description
`N1`	Both	The number (count) of non-missing data values overlapping in the dependent and independent time series.
`MeanX1`	Independent	The mean of the independent N1 data values.
`SX1`	Independent	The standard deviation of the independent N1 values.
`N2`	Independent	The number (count) of non-missing independent values outside of N1.
`MeanX2`	Independent	The mean of the independent N2 values.
`SX2`	Independent	The standard deviation of the independent N2 values.
`MeanY1`	Dependent	The mean of the dependent N1 values.
`SY1`	Dependent	The standard deviation of the dependent N1 values.
`NY`	Dependent	The total number of non-missing dependent values.
`MeanY`	Dependent	The mean of the dependent NY values.
`SY`	Dependent	The standard deviation of the dependent NY values.

Statistic (Table Column Name)	Dependent, Independent, or Both	Description
a	Both	The intercept for the relationship equation.
b	Both	The slope of the relationship equation.
R	Both	The correlation coefficient for N1 values.
R2	Both	R-squared, coefficient of determination for N1 values.
MeanY1est	Dependent	The mean for N1 values computed from the relationship (estimate the dependent values where values were previously known).
SY1est	Dependent	The standard deviation for N1 values computed from the relationship (estimate the dependent at locations where values are known).
RMSE	Dependent	The "room mean squared error" for N1 overlapping values, calculated as: $RMSE = \sqrt{\frac{\sum (Y_i - Y_i')^2}{N_1}}$ where Y_i is the original dependent value and Y_i' is the value estimated with the regression relationship.
SEE	Dependent	The standard error of estimate for N1 overlapping values, calculated as: $SEE = \sqrt{\frac{\sum (Y_i - Y_i')^2}{N_1 - 2}}$ where Y_i is the original dependent value and Y_i' is the value estimated with the regression relationship.
SEP	Both	The standard error of prediction for N1 overlapping values, calculated as: $SEP = \sqrt{1 + \frac{1}{N_1} + \frac{(X_i - \overline{X})^2}{\sum (X_i - \overline{X})^2}} * SEE$ where X_i is the original independent value and $\overline{X}$ is the average of the N1 independent values.
SESlope	Both	The standard error (SE) of the slope (b) for N1 overlapping values, calculated as: $SE = \frac{\sqrt{\frac{\sum (Y_i - Y'_i)^2}{N_1 - 2}}}{\sqrt{\sum (X_i - \overline{X})^2}}$

Statistic (Table Column Name)	Dependent, Independent, or Both	Description
		where X_i is the original independent value and $\overline{X}$ is the average of the N1 independent values; Y_i is the original dependent value and Y_i' is the value estimated with the regression relationship.
`TestScore`	Both	b/SE
`Test Quantile`		From the Student's T-test, which is a function of the confidence interval and degrees of freedom (DF), where DF is the degrees of freedom equal to N1 – 2.
`Test Related`	Both	Will be `Yes` if `TestScore < TestQuantile`, indicating that the data are related, and `No` if `TestScore >= TestQuantile`, indicating that the data are not related. If the data are not related, then the relationship between the dependent and independent time series will not be used for filling.

The following dialog is used to edit the command and illustrates the syntax of the command:



FillRegression

FillRegression() Command Editor

The command syntax is as follows:

```
FillRegression(Parameter=Value,…)
```

Command Parameters

Parameter	Description	Default
`TSID`	The time series identifier or alias for the time series to be filled.	None – must be specified.
`Independent TSID`	The time series identifier or alias for the independent time series.	None – must be specified.
`NumberOf`	The number of equations to use for the analysis:	`OneEquation`

`Equations`	`OneEquation` or `MonthlyEquations`.	
`AnalysisMonth`	Indicate the month to process when using monthly equations. Currently only a single month can be specified.	Process all months.
`Transformation`	Indicates how to transform the data before analyzing. Specify as `None` (previously `Linear`) or `Log` (for Log_{10}). If the `Log` option is used, zero and negative values are replaced with the value specified by the `LEZeroLogValue` parameter value for analysis (missing data values are ignored in the analysis).	`None` (no transformation).
`Intercept`	Specify as `0` to force the intercept of the best-fit line through the origin (not available for log transformation).	Parameter is optional and if specified the default is to not force the intercept through zero.
`Minimum SampleSize`	The minimum number of overlapping values required to use a relationship for filling.	No limit, other than imposed by calculation of statistics.
`MinimumR`	The minimum correlation coefficient required to use a relationship for filling.	No check is performed.
`Confidence Interval`	A confidence interval in percent (e.g., 95) required for the slope of the relationship. The T-test is performed to ensure that the independent and dependent time series are related.	The T-test is not performed to evaluate the confidence interval.
`AnalysisStart`	The date/time to start the analysis – use to focus on only a period appropriate from analysis. For example specify the unregulated period for streamflow.	Analyze the full period.
`AnalysisEnd`	The date/time to end the analysis – use to focus on only a period appropriate from analysis.	Analyze the full period.
`FillStart`	The date/time to start filling, if other than the full time series period.	Fill the full period.
`FillEnd`	The date/time to end filling, if other than the full time series period.	Fill the full period.
`FillFlag`	A single character that will be used to flag filled data.	Filled values will not be flagged.
`FillFlagDesc`	Description for the fill flag, used in reports.	Automatically generated.
`TableID`	A table identifier for a table to receive output of the regression analysis (statistics are described above).	Statistics are not written to the table. Refer to the log file for information.
`TableTSIDColumn`	The name of the column in the table that contains time series identifier information. This is used to match the table with time series being analyzed so that statistics can be written to the correct row.	Required if `TableID` is specified.
`TableTSIDFormat`	The specifier used to format the time series identifier in the TableTSIDColumn. The location part of the TSID, or the time series alias is typically used.	The alias will be used if available, or otherwise the full TSID will be used.
`Fill`	Indicate whether fill should occur (`True`) or just analyze to compute statistics (`False`). The latter is useful for testing combinations of fill parameters	`True`

	prior to actually performing filling.	

A sample command file to fill time series from the State of Colorado's HydroBase is as follows:

```
# 06753400 - LONETREE CREEK AT CARR, CO.
06753400.USGS.Streamflow.Month~HydroBase
# 06753500 - LONETREE CREEK NEAR NUNN, CO.
06753500.USGS.Streamflow.Month~HydroBase
FillRegression(TSID="06753400.USGS.Streamflow.Month",
IndependentTSID="06753500.USGS.Streamflow.Month")
```

This page is intentionally blank.