European Commission

CoBo 3.1

Control Board for Crop Yield Forecasting

Regression Analysis

Log File Contents

Joint Research Centre

Release	Issue	Date	Authors
3.1	1	December 2016	G.Ferrari//L.Nisini/ I.Cerrani/M.Van der Velde

Disclaimer

On any of the MARS pages you may find reference to a certain software package, a particular contractor, or group of contractors, the use of one or another sensor product, etc. In all cases, unless specifically stated, this does not indicate any preference of the Commission for that particular product, party or parties. When relevant, we include links to pages that give you more information about the references.

Feel free to contact us, in case you need additional explanations or information.

Contents

Log File Contents

This document is targeted to **Analysts** and **Supervisors** who use **CoBo Forecasts** to inspect, create, and manage crop yield forecasts.

The aim is to give more detailed statistical information by describing the formulas and the indicators that are used in the analysis and how these work.

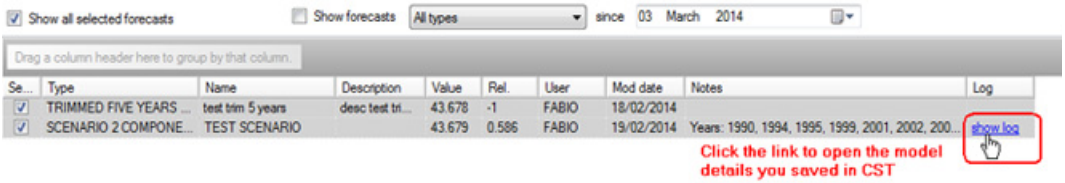
Tip:

Please refer to the **CoBo User Guide** for detailed information on how to use CoBo to create and manage crop yield forecasts.

The document is organized into the following sections:

Viewing the results (Log file)

Once your analysis is completed, CoBo provides a link to the log file (which is available in both the **CST Regression** and the **Regressions on Residuals** analysis). The following shows the CST Regression window:



Note:

For further information on how to create and manage the analysis, please refer to the **CoBo User Guide**, chapter **Creating and managing forecasts > Step 3 - Creating the preliminary forecasts > Create regressions**.

By clicking the **show log** link a file opens that includes detailed information on the model. You can locally save the file.

This topic describes in detail the contents of the log file that you can display after completing your analysis.

For further information:

- “Log File contents” on page 5

Log File contents

The following table shortly describes the main sections/graph in the log file and links to the detailed description where, besides the description, all the formulas are provided:

Table/Graph	Description
Description of regression model	The main parameters that have been set for the regression analysis (Start, End, and Target years, indicators, type of trend, and so forth).
Summary statistics	See “Summary Statistics” on page 6 for a detailed description of all indicators. The formulas are provided, as well.
Regression coefficients	See “Regression Coefficients” on page 12 for a detailed description of all coefficients.
Case statistics	The observed and fitted values of regression analysis, the ordinary residual, the leverage and the influence of the target year prediction. See “Case Statistics” on page 13
Diagnostic Plots: • Observed and Fitted values vs. Year • Residuals vs. Fitted • Residuals vs. Years • Leverage vs. Years • Influence on Prediction Years	See section “Diagnostic Plots” on page 14 for a full list of plots with the relevant description.

Summary Statistics

Indicator	Description
R-squared	The R-squared (R^2) is called *Coefficient of Determination* (Weisberg, 2005) and is defined as the percentage (over the total) of explained variance by the regression model. It provides information on the strength of the relationship between the output data (in this case the statistical yield) and the regressor(s). It's defined as: $$R^2 = \frac{SS_R}{SS_T}$$ **where**: SS_R = Sum of the squared differences between the yield predicted with the model and the mean yield for each year. SS_T = Sum of the squared differences between the observed yield and the mean yield for each year.
Adjusted R-squared	Sometimes denoted as $\bar{R}^2$ is the *adjusted percentage variance explained*. It is a modification of R^2 that adjusts for the number of explanatory terms in a model. Unlike R^2, the adjusted $\bar{R}^2$ increases only if the new term improves the model more than would be expected by chance. The adjusted $\bar{R}^2$ can be negative, and will always be less than or equal to R squared. The adjusted $\bar{R}^2$ is defined as: $$\bar{R}^2 = 1-(1-R^2)\times\left[\frac{n-1}{n-k-1}\right] = R^2-(1-R^2)\times\left[\frac{k}{n-k-1}\right]$$ **where**: n = Sample size k = Total number of explanatory variables in the model (not including the constant term).
Residual Standard Deviation	The *square root of the residual mean square* (or *mean squared error*). It is the standard deviation of the residuals (residuals = differences between observed and predicted values). It is calculated as follows: $$Sres = \sqrt{\frac{\Sigma(Y_{obs}-Y_{est})^2}{n-k-1}}$$ **where**: Y_{obs} = yield observed Y_{est} = yield estimated

Indicator	Description
Root mean squared error for prediction (RMSE)	Defines e_i as the difference between the i^{th} observation and the predicted value for the i^{th} observation: $$e_i = Y_{obs,\,i} - \hat{Y}est(i),-i$$ **where**: $\hat{Y}_{est(i),\,i}$ is based on a model fit to the remaining observations (i.e., without the i^{th} observation). Then, it is calculated as: $$RMSE = \sqrt{\frac{\Sigma(e_i)^2}{n-k-1}}$$ This is sometimes called the PRESS residual or the *leave one out* residual. The root mean squared error for prediction is the *root of the mean value of all the squared estimated errors* e_i.

Indicator	Description
VIF (Variance Inflation Factors) and Maximum of VIF	It's possible that a large correlation among the predictor variables used in the regression analysis is detected. This characteristic is called *multicollinearity* that can have effects on our regression analyses and subsequent conclusions (i.e., the variance of the estimated regression coefficient is inflated by the multicollinearity). In order to check if the multicollinearity is present in the data submitted to regression, the *VIF (Variance Inflation Factors)* is used. If the follow linear regression model with X_1, X_2,... correlated predictors is considered: $$Y_{obs} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$$ the *Variance Inflation Factor* for the i^{th} predictor is: $$VIF_i = \frac{1}{1 - R_i^2}$$ **where**: R_i^2 is the R^2-value obtained by regressing the i^{th} predictor on the remaining predictors: $$X_i = \beta_0 + \beta_1 X_1 + \beta_{i-1} X_{i-1} + \beta_{i+1} X_{i+1} + \ldots + \varepsilon$$ Note that a variance inflation factor exists for *each of the i predictors* in a multiple regression model. The VIF is a measure of how much the variance of the estimated regression coefficient is inflated by the correlation among the predictor variables in the model. If $VIF_i = 1$ means that there isn't correlation among the i^{th} predictors and the remaining predictor variables, so the variance of β_i is not inflated at all. The *maximum of VIF* is the maximum of the inflation factors for the indicators (i.e., the inflation factors for the constant and the time trend effects are not taken into account). If it is high (more than 5), it indicates that 2 or more indicators are linearly dependent. It means that a model with a high maximum VIF is probably a "bad" model that overfits the observations.

Indicator	Description
Standard Error of Prediction (Mean):	Is the standard error of the mean yield prediction μ_x for the target year $X = X_0$. Indeed, the estimation of the coefficients for various indicators in a regression model produce uncertainty, also in the estimated mean yield. The Standard Error for Mean is given by: $$SE(\hat{\mu}_X) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)S_X^2}}$$ **where**: n = number of observations, Y_i the i^{th} observation, $\hat{Y}_i$ the predicted value for the i^{th} observation and: • *Residual Standard Deviation (1)*: $$\hat{\sigma} = Sres = \sqrt{\frac{1}{n-2}\Sigma(Y_i - \hat{Y}_i)^2}$$ • *Sample Variance of X's (2)*: $$S_X^2 = \frac{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}{n-1}$$

Indicator	Description
Standard Error of Prediction (New)	The standard error of the individual yield prediction for the target year (i.e., given the values for the indicators in the target year). Indeed, it is possible estimate an individual's response for the target year at $X = X_0$ and also the uncertainty, in terms of Standard Error, can be estimate. The Standard Error for prediction Y at $X = X_0$ is given by: $$SE(\hat{Y}_x) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)S_X^2}}$$ where $\hat{\sigma}$ and S_X^2 are reported in **(1) - Residual Standard Deviation** and **(2) - Sample Variance of X's** in the previous page. The following equality holds:] $$SE(\hat{Y}_x)^2 = SE(\hat{\mu}_x)^2 + S_{res}{}^2$$ (i.e., Standard Error of Prediction (New)^2 = Standard Error of Prediction (Mean)^2 + ResidualStandardDeviation^2). This formula tells us that the uncertainty in the prediction for an individual year is the sum of the uncertainty in the regression line and the uncertainty in the observations done in the target year.
Residual Degrees of Freedom	The number of degrees of freedom for residual (i.e., the number of observations minus the *number of constraints* - i*n this case no. of estimated coefficients* β_i *, including the estimated constant* β_0) If k = number of indicators (regressors) in the model and n = number of observations, the residual degrees of freedom is: $$v = n - k - 1$$

Indicator	Description
F Statistic	This statistic is part of the output in a F-test and can be used to compare regression models. The formula for this statistics is calculated dividing the variance explained by the model and the variance unexplained as follows: $$F = \frac{\frac{SS_R}{p}}{\frac{SS_E}{v}} = \frac{MS_R}{MS_E}$$ The quantity SS_R (*Regression Sum of Squares*) is the variance due to the model. It's calculated as sum of the squared differences between the predicted yield and the mean yield for each year. The number of degrees of freedom associated with SS_R is $p = (k-1)$. The MS_R (*Regression Mean Square*) is obtained by dividing the regression sum of squares by the respective degrees of freedom as follows: $$MS_R = \frac{SS_R}{1} = SS_R$$ The quantity SS_E (*Residual Sum of Squares of Error Sum of Squares*) is the variance that is not explained by the model. It's calculated as sum of the squared differences between the observed yield and predicted yield for each year. The number of degrees of freedom associated to SS_E is $\upsilon = (n-k)$. The MS_E (*Error or Residual Mean Square*) is obtained by dividing the sum of squares by the respective degrees of freedom: $$MS_E = \frac{SS_E}{\upsilon}$$ Finally, the quantity SS_T (*Total Sum of Squares*) is the total variance of the observations. It's calculated as sum of the squared differences between the observed yield and the mean yield for each year. The value *q = n - 1* is the number of degree of freedom associated with the total variance.

Indicator	Description
	The MS_T (*Total Mean Square*) is obtained by dividing the sum of squares by the respective degrees of freedom. The formula is as follows: $$MS_T = \frac{SS_T}{q}$$ The total variance of the observation can be written as: $$SS_T = SS_R + SS_E$$ and $$q = n - 1 = p + v$$
F-Probability	Linked to the F-value there is an F probability (p - value). Lower is the p - value and higher is the probability that the values of the estimated coefficients are true (i.e. that a relation does exist between the statistical yield and its predictors). According to the literature the threshold for the significance of p-value is 0,05 or 0,01 (i.e., if a calculated p-value < 0,05 (0,01) does exist the relation between the variables involved in the regression analysis).

See also:

- "Regression Coefficients" on page 12
- "Case Statistics" on page 13
- "Diagnostic Plots" on page 14

Regression Coefficients

The multiple regression coefficients β_0 (*intercept*), $\beta_1, \beta_2 \ldots$ are estimated using the least square criterion.

For each estimated coefficient also the estimation of *Standard Error* is calculated.

The **Standard Error for the Intercept** (β_0) is given by:

$$SE(\beta_0) = \hat{\sigma}\sqrt{\frac{1}{2} + \frac{\bar{X}^2}{(n-1)S_X^2}}$$

Where $\hat{\sigma}$ is the *Residual Standard Deviation* and S_X^2 is the *Sample Variance of X*.

The Standard Error of the Regression Coefficient (β_k) is a measure of the amount of sampling error in the estimation of the coefficient. If (b_k) is a point estimate of (β_k), the formula is given by:

$$s_{b_k} = \frac{s_e}{\sqrt{\left(1 - R^2_{X_k,G_k}\right) * S^2_{X_k} * (n-1)}}$$

See also:

- "Summary Statistics" on page 6
- "Case Statistics" on page 13
- "Diagnostic Plots" on page 14

Case Statistics

This area shows the observed and fitted values of the regression analysis, the ordinary residual, the leverage and the influence of the target year prediction.

The following shows an example:

Year	Yield	Fitted	Residual	Leverage	Influence
1994	2.08	2.265	-0.18482	0.08694	-0.00356
1995	2.06	2.213	-0.15322	0.06396	-0.00268
1996	2.24	2.215	0.02533	0.03738	0.00094
1997	2.28	2.27	0.00998	0.11867	-0.00144
1998	2.19	2.204	-0.01415	0.09058	0.00208
1999	2.34	2.364	-0.0243	0.32475	-0.00426
2000	2.52	2.225	0.29466	0.06796	0.00701
2001	2.24	2.335	-0.09496	0.08496	-0.00324
2002	2.43	2.339	0.0909	0.10593	-0.00353
2003	2.18	2.237	-0.05734	0.24872	-0.00023
2004	2.37	2.347	0.02294	0.08027	-0.00199
2005	2.34	2.293	0.04696	0.19239	0.00524
2006	2.23	2.311	-0.08085	0.06769	0.00042
2007	2.52	2.343	0.17728	0.05007	-0.00262
2008	2.55	2.377	0.17303	0.06039	-0.00082
2009	2.35	2.358	-0.00757	0.05491	0.00028
2010	2.36	2.343	0.01651	0.08054	0.00187
2011	2.54	2.419	0.12077	0.07368	-0.00293
2012	2.32	2.36	-0.04016	0.07619	0.00155
2013	2.05	2.371	-0.32099	0.03403	0.00682

See also:

- “Summary Statistics” on page 6
- “Regression Coefficients” on page 12

Diagnostic Plots

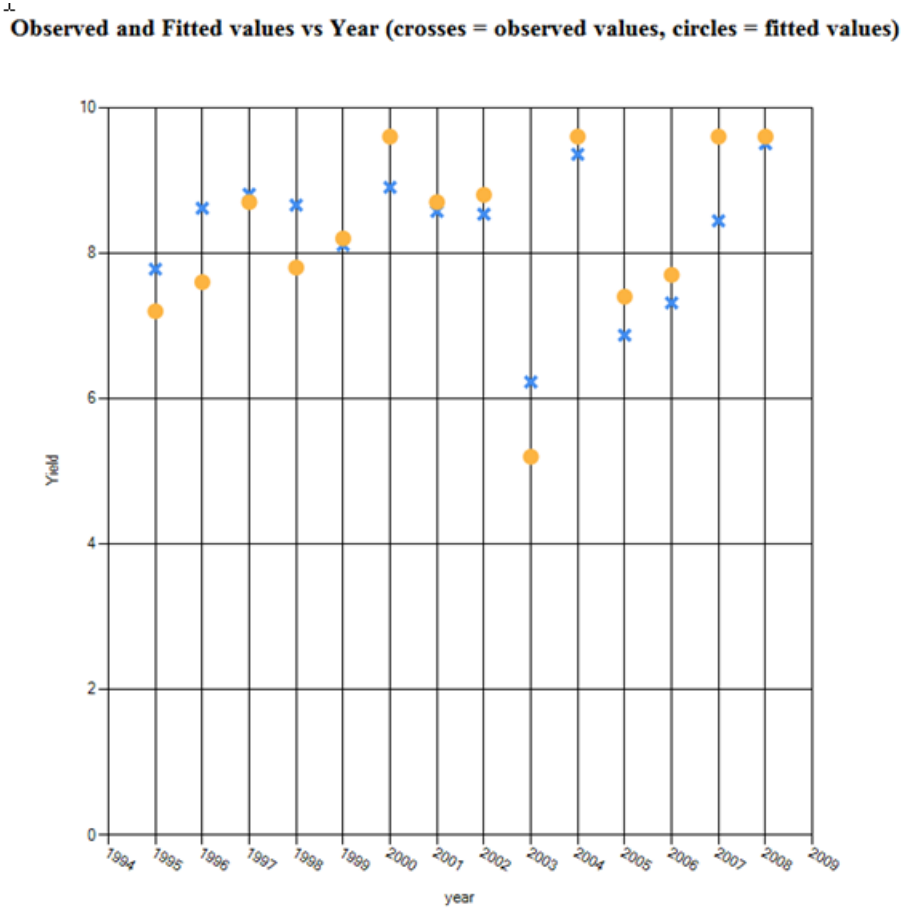
Set of plots for checking the quality of the regression analysis results.

For an example of each and the relevant description, see:

- “Observed and Fitted values vs. Year” on page 15
- “Residuals vs. Fitted” on page 16

Observed and Fitted values vs. Year

The following shows an example:

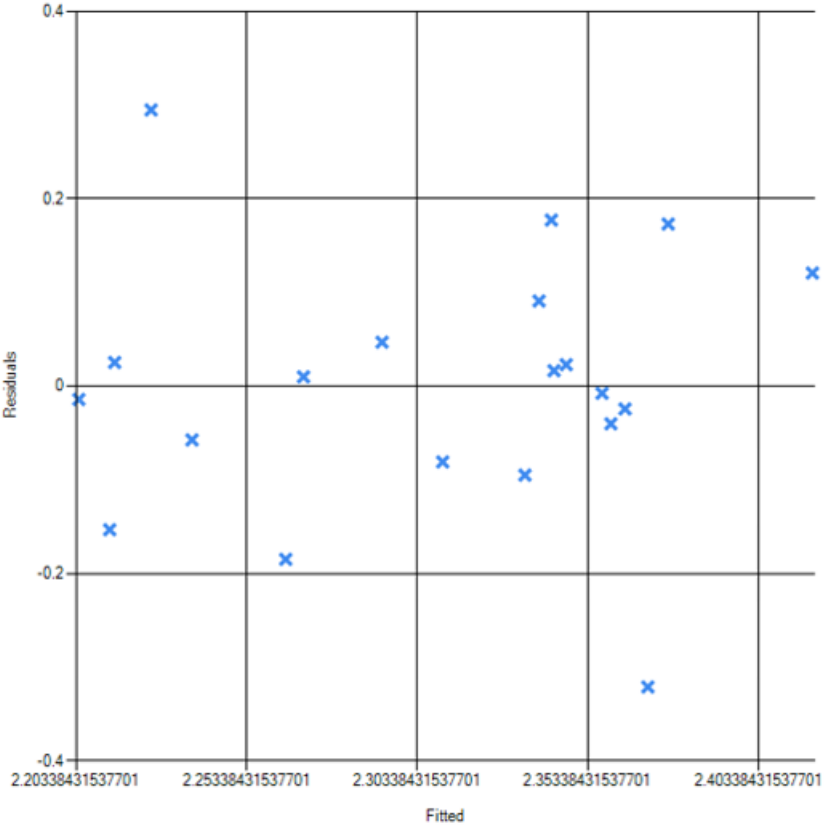


The plot allows checking the observed and fitted yields.

Note that only the predicted value is displayed for the target year: therefore, you can compare the prediction with the observed yields only in the previous years.

Residuals vs. Fitted

The following shows an example:



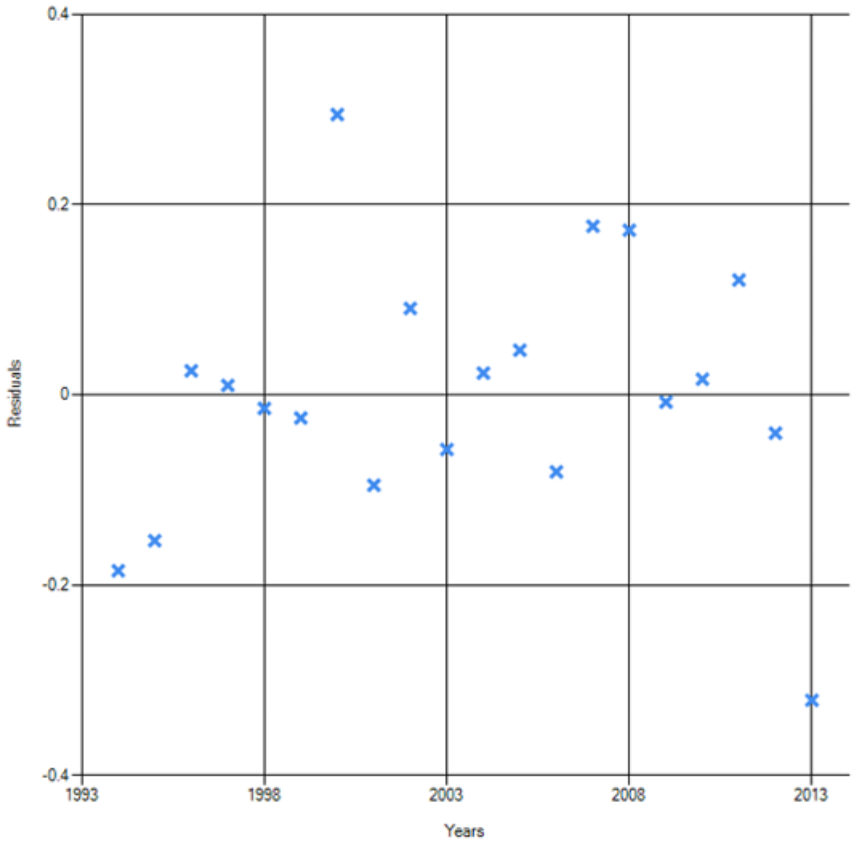
The plot shows whether or not the data have a linear pattern. If the residuals are symmetrically distributed and are homogeneous around a horizontal line without distinct patterns, this means that there is not a non-linear relationship. Otherwise, if there are some patterns in the residuals, it means that there is a non-linear relationship.

Moreover, if the residuals get larger as the prediction moves from small to large (or from large to small), it means that the variance of the observations is not homogeneous (heteroscedasticity). See page 140 of the document indicated in "(*) Sources:" on page 20).

The residual plot is useful for the outliers' detection, as well.

Residuals vs. Years

The following shows an example:

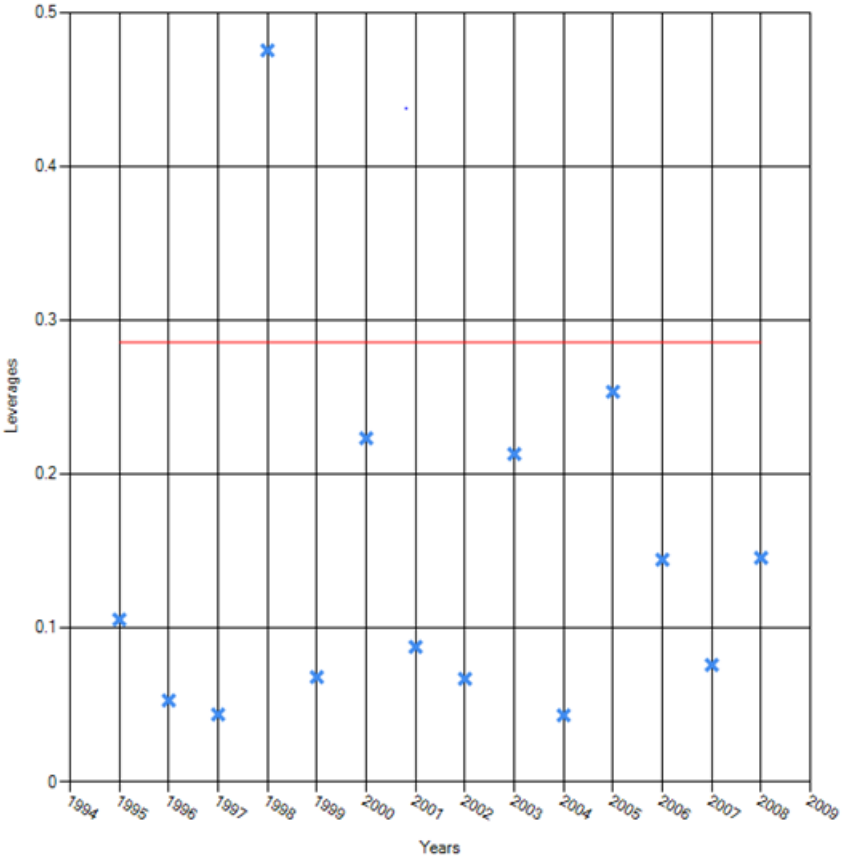


The ordinary residuals plot versus year shows if the residuals have a relationship with year. A linear or quadratic plot might indicate that a time trend should be added to the model. The plot can also reveal that

the variance is changing with time, or that residuals are correlated. (See page 149 of the document indicated in "(*) Sources:" on page 20).

Leverage vs. Years

The following shows an example:



This plot can be used to identify observations that are potentially influential. The observations with high value of leverage are potentially influential because these observations have indicators with high values if compared to the rest of the observations.

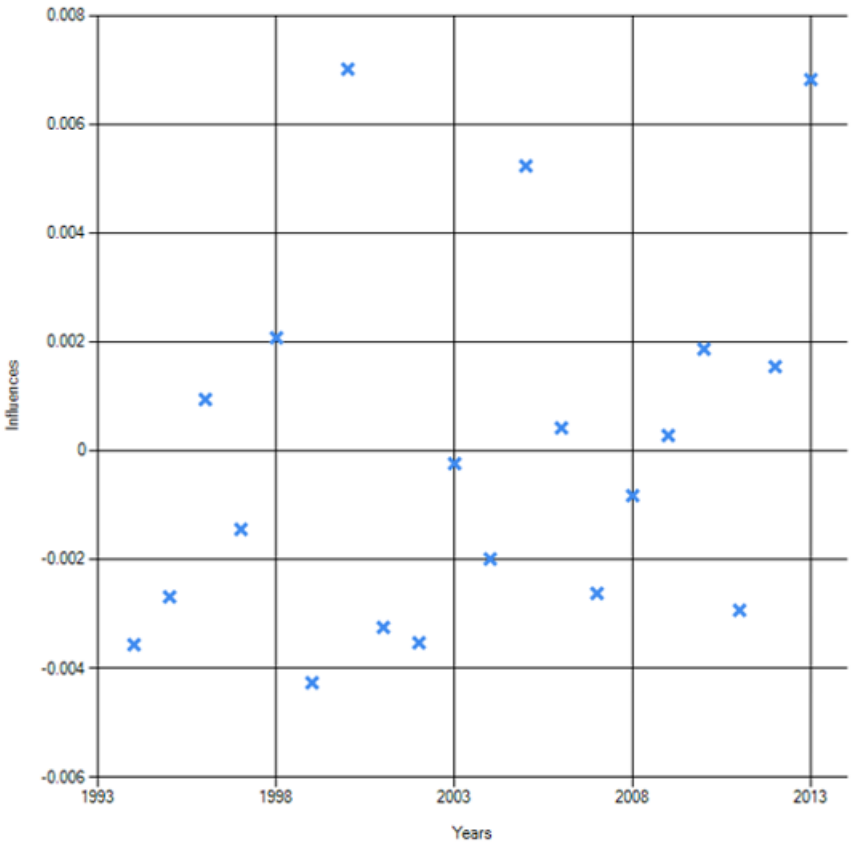
Leverages larger than 2p/n (where **p** is the number of regression coefficients and **n** is the number of observations) are generally

considered as points with high values of leverage. (See page 209 of the document indicated in “(*) Sources:” on page 20).

A leverage points has remote values for the indicators, but the regression line does not change much by deletion of the leverage point.

Influence on Prediction Years

The following shows an example:



The influence on the target prediction versus year plot shows the effect of deleting individual observations on the predicted value for the target year. A point can be influential on the regression model as a whole, but have a small target prediction residual.

***(*) Sources*:**

McCullagh, P. and Nelder, J.A. (1989) - Generalized Linear Models, second edition. Champang and Hall. London

Related topics:

- "Observed and Fitted values vs. Year" on page 15
- "Residuals vs. Fitted" on page 16
- "Residuals vs. Years" on page 17
- "Leverage vs. Years" on page 17
- "Influence on Prediction Years" on page 19