regression diagnostic plots — Postestimation plots for regression

Postestimation commands

The following postestimation commands are of special interest after `h2oml gbregress` and `h2oml rfregress`:

Command	Description
`h2omlgraph rvfplot`	residual-versus-fitted plot
`h2omlgraph rvpplot`	residual-versus-predictor plot

h2omlgraph rvfplot

Description for h2omlgraph rvfplot

`h2omlgraph rvfplot` graphs a residual-versus-fitted plot, a graph of the residuals against the fitted values after `h2oml gbregress` and `h2oml rfregress`.

Menu for h2omlgraph rvfplot

Statistics > H2O machine learning

Syntax for h2omlgraph rvfplot

`h2omlgraph rvfplot` [, *rvfplot_options*]

rvfplot_options	Description
Plot	
marker_options	change look of markers (color, size, etc.)
marker_label_options	add marker labels; change look or position
Y axis, X axis, Titles, Legend, Overall	
twoway_options	any options other than `by()` documented in [G-3] ***twoway_options***
`train`	specify that residuals be reported using training results
`valid`	specify that residuals be reported using validation results
`test`	specify that residuals be computed using testing frame
`test(`*framename*`)`	specify that residuals be computed using data in testing frame *framename*
`frame(`*framename*`)`	specify that residuals be computed using data in H2O frame *framename*
`framelabel(`*string*`)`	label frame as *string* in the output

`train`, `valid`, `test`, `test()`, `frame()`, and `framelabel()` do not appear in the dialog box.

Options for h2omlgraph rvfplot

Plot

marker_options affect the rendition of markers drawn at the plotted points, including their shape, size, color, and outline; see [G-3] ***marker_options***.

marker_label_options specify if and how the markers are to be labeled; see [G-3] ***marker_label_options***.

Y axis, X axis, Titles, Legend, Overall

twoway_options are any of the options documented in [G-3] ***twoway_options***, excluding `by()`. These include options for titling the graph (see [G-3] ***title_options***) and for saving the graph to disk (see [G-3] ***saving_option***).

The following options are available with h2omlgraph rvfplot but are not shown in the dialog box:

train, valid, test, test(), and frame() specify the H2O frame for which residuals are reported. Only one of train, valid, test, test(), or frame() is allowed.

train specifies that residuals be reported using training results. This is the default when validation is not performed during estimation and when a postestimation frame has not been set with h2omlpostestframe.

valid specifies that residuals be reported using validation results. This is the default when validation is performed during estimation and when a postestimation frame has not been set with h2omlpostestframe. valid may be specified only when the validframe() option is specified with h2oml *gbm* or h2oml *rf*.

test specifies that residuals be computed on the testing frame specified with h2omlpostestframe. This is the default when a testing frame is specified with h2omlpostestframe. test may be specified only after a testing frame is set by using h2omlpostestframe. test is necessary only when a subsequent h2omlpostestframe command is used to set a default postestimation frame other than the testing frame.

test(*framename*) specifies that residuals be computed using data in testing frame *framename* and is rarely used. This option is most useful when running a single postestimation command on the named frame. If multiple postestimation commands are to be run on the same test frame, it is more computationally efficient and convenient to specify the testing frame by using h2omlpostestframe instead of specifying test(*framename*) with individual postestimation commands.

frame(*framename*) specifies that residuals be computed using the data in H2O frame *framename*.

framelabel(*string*) specifies the label to be used for the frame in the output.

h2omlgraph rvpplot

Description for h2omlgraph rvpplot

`h2omlgraph rvpplot` graphs a residual-versus-predictor plot (also known as an independent variable plot or a carrier plot), a graph of the residuals against the specified predictor after `h2oml gbregress` and `h2oml rfregress`.

Menu for h2omlgraph rvpplot

Statistics > H2O machine learning

Syntax for h2omlgraph rvpplot

`h2omlgraph rvpplot` *predictor* [, *rvpplot_options*]

rvpplot_options	Description
Plot	
marker_options	change look of markers (color, size, etc.)
marker_label_options	add marker labels; change look or position
Y axis, X axis, Titles, Legend, Overall	
twoway_options	any options other than `by()` documented in [G-3] ***twoway_options***
`train`	specify that residuals be reported using training results
`valid`	specify that residuals be reported using validation results
`test`	specify that residuals be computed using testing frame
`test(`*framename*`)`	specify that residuals be computed using data in testing frame *framename*
`frame(`*framename*`)`	specify that residuals be computed using data in H2O frame *framename*
`framelabel(`*string*`)`	label frame as *string* in the output

`train`, `valid`, `test`, `test()`, `frame()`, and `framelabel()` do not appear in the dialog box.

Options for h2omlgraph rvpplot

Plot

marker_options affect the rendition of markers drawn at the plotted points, including their shape, size, color, and outline; see [G-3] ***marker_options***.

marker_label_options specify if and how the markers are to be labeled; see [G-3] ***marker_label_options***.

Y axis, X axis, Titles, Legend, Overall

twoway_options are any of the options documented in [G-3] ***twoway_options***, excluding `by()`. These include options for titling the graph (see [G-3] ***title_options***) and for saving the graph to disk (see [G-3] ***saving_option***).

The following options are available with h2omlgraph rvpplot but are not shown in the dialog box:

train, valid, test, test(), and frame() specify the H2O frame for which residuals are reported. Only one of train, valid, test, test(), or frame() is allowed.

train specifies that residuals be reported using training results. This is the default when validation is not performed during estimation and when a postestimation frame has not been set with h2omlpostestframe.

valid specifies that residuals be reported using validation results. This is the default when validation is performed during estimation and when a postestimation frame has not been set with h2omlpostestframe. valid may be specified only when the validframe() option is specified with h2oml *gbm* or h2oml *rf*.

test specifies that residuals be computed on the testing frame specified with h2omlpostestframe. This is the default when a testing frame is specified with h2omlpostestframe. test may be specified only after a testing frame is set by using h2omlpostestframe. test is necessary only when a subsequent h2omlpostestframe command is used to set a default postestimation frame other than the testing frame.

test(*framename*) specifies that residuals be computed using data in testing frame *framename* and is rarely used. This option is most useful when running a single postestimation command on the named frame. If multiple postestimation commands are to be run on the same test frame, it is more computationally efficient and convenient to specify the testing frame by using h2omlpostestframe instead of specifying test(*framename*) with individual postestimation commands.

frame(*framename*) specifies that residuals be computed using the data in H2O frame *framename*.

framelabel(*string*) specifies the label to be used for the frame in the output.

Remarks and examples

Remarks and examples are presented under the following headings:

h2omlgraph rvfplot
h2omlgraph rvpplot

h2omlgraph rvfplot

h2omlgraph rvfplot graphs the residuals against the fitted values. Residual plots tend to be less informative for machine learning models than for ordinary least squares. However, they can still be useful for examining the behavior of the residuals. In general, for a well-fitted model, we expect the residuals to show no pattern. The presence of a pattern may indicate underfitting or overfitting. Residual plots can also give us an idea of the size of the residuals. For example, large residuals for certain observations may suggest that the model is struggling to capture their behavior (Biecek and Burzykowski 2021).

▷ Example 1

Using auto.dta described in **[U] 1.2.2 Example datasets**, we will use h2oml gbregress to fit a gradient boosting regression model of price on weight, mpg, foreign, and length.

We start by opening the dataset in Stata and then putting it into an H2O frame. Recall that `h2o init` initiates an H2O cluster, `_h2oframe put` loads the current Stata dataset into an H2O frame, and `_h2oframe change` makes the specified frame the current H2O frame. For details, see *Prepare your data for H2O machine learning in Stata* in [H2OML] **h2oml** and [H2OML] **H2O setup**.

```
. use https://www.stata-press.com/data/r19/auto
(1978 automobile data)
. h2o init
. _h2oframe put, into(auto)
Progress (%): 0 100
. _h2oframe change auto
```

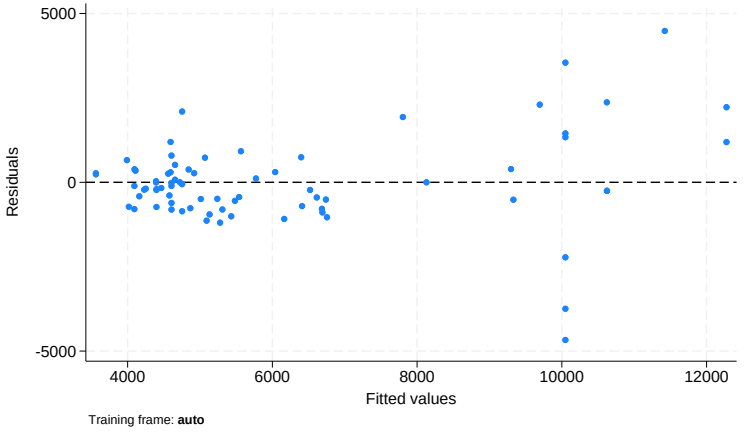
We fit gradient boosting regression with the default hyperparameter values.

```
. h2oml gbregress price mpg foreign length weight, h2orseed(19)
Progress (%): 0 100
Gradient boosting regression using H2O

Response: price
Loss:     Gaussian
Frame:                                  Number of observations:
  Training: auto                                  Training =     74

Model parameters

Number of trees      =  50              Learning rate       =     .1
             actual =  50               Learning rate decay =      1
Tree depth:                             Pred. sampling rate =      1
           Input max =   5              Sampling rate       =      1
                 min =   3              No. of bins cat.    =  1,024
                 avg = 4.0              No. of bins root    =  1,024
                 max =   5              No. of bins cont.   =     20
Min. obs. leaf split =  10              Min. split thresh.  = .00001

Metric summary
```

Metric	Training
Deviance	1680799
MSE	1680799
RMSE	1296.456
RMSLE	.168809
MAE	860.1429
R-squared	.8041476

We now use the `h2omlgraph rvfplot` command to graph the residuals against the fitted values:

```
. h2omlgraph rvfplot, yline(0)
```



All the diagnostic plot commands allow the `graph twoway` and `graph twoway scatter` options; we specified `yline(0)` to draw a line across the graph at $y = 0$; see [G-2] **graph twoway scatter**.

In a well-fitted model, we expect no discernible pattern in the residuals when plotted against the fitted values. If the model is correctly specified, the residuals should appear randomly scattered as in the plot above. Any systematic pattern in the plot suggests potential issues that require further scrutiny and adjustment. Here it seems the residual variance is larger for the more expensive cars.

◁

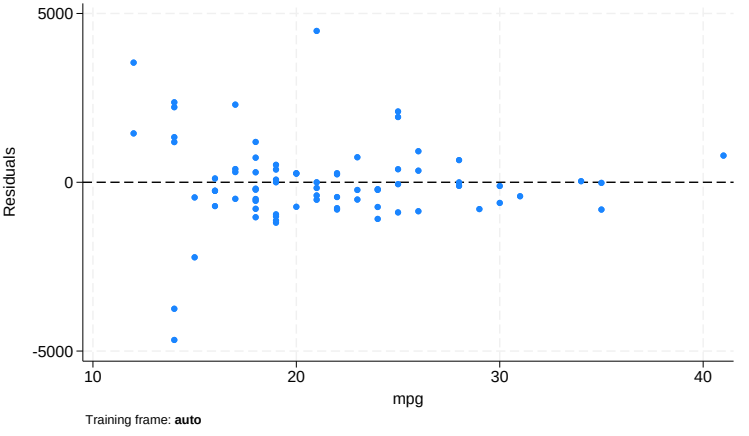
h2omlgraph rvpplot

The residual-versus-predictor plot is a simple way to look for violations of the regression assumptions. If the assumptions are correct, there should be no pattern on the graph.

▷ Example 2

Let's use our estimation results from example 1, and plot the residual-versus-predictor plot for the predictor `mpg`,

```
. h2omlgraph rvpplot, yline(0)
```



We see higher residual variance for smaller values of `mpg`.

◁

Reference

Biecek, P., and T. Burzykowski. 2021. *Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models*. Boca Raton, FL: CRC Press.

Also see

[H2OML] **h2oml** — Introduction to commands for Stata integration with H2O machine learning

[H2OML] **h2oml postestimation** — Postestimation tools for h2oml gbm and h2oml rf

[H2OML] **h2oml gbregress** — Gradient boosting regression

[H2OML] **h2oml rfregress** — Random forest regression