H2O setup — Prepare data for H2O analysis in Stata[+]

[+]These features are part of StataNow.

Description Remarks and examples Also see

Description

In this entry, we provide an introduction to the H2O integration with Stata. We introduce commands for initiating H2O and working with data frames in H2O, both of which are necessary before you can use `h2oml` commands described in [H2OML] **h2oml** and throughout this manual.
stata.com

Remarks and examples

Remarks are presented under the following headings:

What is H2O?

H2O is a scalable and distributed machine learning and predictive platform. It is an open-source platform, and its core code is written in Java. Stata uses H2O's REST API to connect to H2O. You can perform in-memory data analysis and machine learning using this framework. More information about the H2O framework can be found on the H2O website at https://docs.h2o.ai/. We also refer you to H2O's *User Guide*.

We separate H2O related commands in Stata into two categories:

1. Commands to establish connection with H2O and work with H2O frames. For details, see [P] **H2O intro** and https://www.stata.com/h2o/.

2. Commands for machine learning (`h2oml`). For the Stata examples, see [H2OML] **h2oml**.

How does H2O work from Stata?

You can either start a new H2O cluster or connect to an existing H2O cluster from within Stata. Then you use the suite of Stata commands (`h2o`, `_h2oframe`, and `h2oml`) to interact with the H2O cluster.

Start a local H2O cluster

You can start a local H2O cluster by typing in Stata

```
. h2o init
```

`h2o init` will look for the existence of an `h2o.jar` file, a Java Archive (JAR) file that is used to start H2O. This file is distributed by H2O. Stata does not distribute `h2o.jar` with its installation.

Downloading and placing an h2o.jar

To download the h2o.jar file and place it in the local directory so that Stata can locate it, you can follow the steps below. Note that these steps need to be completed only once.

You can obtain the h2o.jar file from H2O's download page.

1. Go to https://h2o.ai/resources/download/.
2. Click on the tab **H2O Open Source Platform**.
3. Go to **Latest Stable Release** or **Prior Releases**. Stata's H2OML documentation is written using **Version 3.46.0.6**.
4. Click on **Download H2O**.
5. After downloading the file (for example, h2o-3.46.0.6.zip), unzip it and look for the h2o.jar file. This is the only file from within the zip file that you will need.

After downloading the h2o.jar file, place the file in a directory included in Stata's system directories (ado-path). To view directories on the ado-path, you can use the adopath command. For details, see [P] **sysdir**. For example, the following is a typical Stata output on a Windows computer:

```
. adopath
  [1]  (BASE)      "C:\Program Files\Stata18\ado\base"
  [2]  (SITE)      "C:\Program Files\Stata18\ado\site"
  [3]              "."
  [4]  (PERSONAL)  "C:\ado\personal"
  [5]  (PLUS)      "C:\ado\plus"
  [6]  (OLDPLACE)  "C:\ado"
```

We recommend using the SITE, PERSONAL, or PLUS directory. When h2o.jar is placed along the ado-path, h2o init will use it directly to start a new local H2O cluster. If multiple copies of h2o.jar exist along the ado-path, Stata will prioritize based on the order that the adopath command presents and will use the first h2o.jar it locates. Because we are looking for a .jar file, h2o init can locate h2o.jar if it is placed in a jar/ subdirectory. Please create the jar/ subdirectory if it does not exist in any of the defined ado-path locations. If h2o.jar cannot be located, h2o init will produce an error.

After h2o.jar is located, h2o init will determine whether a cluster is already running on your local machine.

When the cluster has been successfully initialized, Stata will automatically connect to this cluster, and a summary of the H2O cluster status similar to the following will be displayed:

```
. h2o init
Connecting to the H2O cluster running at http://127.0.0.1:54321.....not found.
Starting a new cluster running at http://127.0.0.1:54321.
Connecting to the H2O cluster running at http://127.0.0.1:54321..... Successful.
------------------------------------------------------------------------------
H2O cluster uptime:         1 sec
H2O cluster timezone:       America/Chicago
H2O data parsing timezone: UTC
H2O cluster version:        3.46.0.6
H2O cluster version age:    3 months and 7 days
H2O cluster total nodes:    1
H2O cluster free memory:    15.73 Gb
H2O cluster total cores:    32
H2O cluster allowed cores: 32
H2O cluster status:         accepting new members, healthy
H2O connection url:         http://127.0.0.1:54321
------------------------------------------------------------------------------
```

`h2o init` allows some options for customizing the initialization of the H2O cluster. For example, we can specify the `nthreads()` option to set the maximum number of parallel threads to use when launching the H2O cluster. For details, see https://www.stata.com/h2o/h2o18/h2o.html.

❑ Technical note

`h2o init` uses the address of **localhost:54321**, where the IP of localhost is **127.0.0.1** and the port is **54321**. If a cluster is not already running, `h2o init` will attempt to create one at this location, and by default, the new cluster will allow connections only from the local machine.

❑

Connect to an existing H2O cluster

Another way to interact with H2O is to connect to an existing H2O cluster by using the `h2o connect` command. For example, an existing H2O cluster can be a cluster previously started by `h2o init`. For details, see https://www.stata.com/h2o/h2o18/h2o.html.

To connect to an existing H2O cluster, we can type `h2o connect` in Stata. If the connection is built successfully, Stata will report a summary of the cluster status similar to the following:

```
. h2o connect
Connecting to the H2O cluster running at http://localhost:54321. Successful.
------------------------------------------------------------------------------
H2O cluster uptime:         29 mins 58 secs
H2O cluster timezone:       America/Chicago
H2O data parsing timezone: UTC
H2O cluster version:        3.46.0.6
H2O cluster version age:    3 months and 7 days
H2O cluster total nodes:    1
H2O cluster free memory:    15.70 Gb
H2O cluster total cores:    32
H2O cluster allowed cores: 32
H2O cluster status:         locked, healthy
H2O connection url:         http://localhost:54321
------------------------------------------------------------------------------
```

You can also connect to an H2O cluster running on a remote machine by specifying its IP and port in the `ip()` and `port()` options in the `h2o connect` command. For details, see *Options for h2o connect*.

❑ Technical note

By default, `h2o connect` will attempt to connect to a cluster running at **localhost:54321** on your local machine; if you started a local cluster with `h2o init`, then credentials will automatically be used.

❑

When you connect to an existing H2O cluster, a new Stata H2O session is created between Stata (the client) and the H2O cluster. Multiple clients can be connecting to the H2O cluster at the same time, and they will all share its resources, such as the data and models within the cluster.

Interact with the H2O cluster

Once a connection with an H2O cluster has been established, you can interact with it directly from within Stata.

For example, you can import data from the local drive to the cluster as an H2O frame or put data currently in Stata into an H2O frame. The following code will load the `iris` dataset to the cluster into an H2O frame `h2oiris`. For details, see https://www.stata.com/h2o/h2o18/.

```
. use https://www.stata-press.com/data/r18/iris
(Iris data)
. _h2oframe put, into(h2oiris)
```

To load a subset of the data, you can specify *varlist* and the `if` and `in` qualifiers. For more details, see https://www.stata.com/h2o/h2o18/h2oframe_put.html.

You can type `_h2oframe dir` to list all H2O frames in the cluster, along with the dimensions of the data and the amount of memory the data consume in the cluster.

```
. _h2oframe dir
Name                                    Rows        Cols        Size
h2oiris                                  150           5    1.773 Kb
Total: 1
```

For more information about H2O frames, see https://www.stata.com/h2o/h2o18/h2oframe.html.

You can set or change to the `h2oiris` frame as the current working H2O frame by using the `_h2oframe change` command. Then to perform, for instance, gradient boosting multiclass classification using the dataset on this frame, type

```
. _h2oframe change h2oiris
. h2oml gbmulticlass iris seplen sepwid petlen petwid
  (output omitted)
```

Instead of separate `_h2oframe put` and `_h2oframe change` commands, it is often convenient to put data into an H2O frame and make that frame current in a single step by typing, for instance,

```
_h2oframe put, into(h2oiris) current
```

Close and disconnect the H2O cluster

Once you have finished the analysis on the H2O cluster, you can type

```
. h2o disconnect
```

to close the connection from the H2O session between Stata and the cluster or

```
. h2o shutdown
```

to shut down the cluster.

The `h2o disconnect` command will close the H2O connection between Stata and the cluster, leaving the H2O cluster running. Later in the same Stata session, you can type `h2o connect` to rebuild the connection to it and reaccess the resources it contains.

The `h2o shutdown` command will destroy the cluster you are currently connected to along with all its resources. By default, `h2o shutdown` will exit with an error and give a warning about its destructive nature. To override this warning and actually shut down the cluster, use the `force` option. This will force the cluster to shut down, and everything in the cluster will be destroyed regardless of whether the cluster was created from Stata or outside of Stata.

Note that if the cluster was created by Stata using the `h2o init` command, then by exiting a Stata session, it will be automatically shut down. We recommend to ensure that all the necessary resources within the cluster are saved before exiting. To prevent a cluster that Stata created from automatically getting shut down, use `h2o disconnect` before closing Stata. If the cluster was created outside of Stata and a connection was made using `h2o connect`, then exiting Stata will close only the connection, leaving all resources within the cluster intact.

The table below summarizes the alternatives to close or disconnect an H2O frame.

Option	Cluster created by Stata	Cluster created outside of Stata
`h2o disconnect`	close H2O session without loss of information	close H2O session without loss of information
`h2o shutdown, force`	close H2O session and discard information in the cluster	close H2O session and discard information in the cluster
Exit Stata session	same as `h2o shutdown, force`	same as `h2o disconnect`

In practice, if you are certain that all necessary results have been saved, it is preferable to use `h2o shutdown` to shut down the H2O cluster. Putting all H2O-related commands between `h2o init` and `h2o shutdown, force` is the recommended practice.

Also see

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

[P] **H2O intro** — Introduction to integration with H2O