Using Stata on the Notre Dame CRC Machines

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/

Last revised January 9, 2019

I recommend buying a personal copy of Stata if you want to work off campus. Try to avoid hand me down old versions of Stata (e.g. Stata 13) because they often contain bugs that Stata fixed years ago and lack some of the latest and greatest Stata features.

However, as an alternative, you may want to use Stata on Notre Dame's Center for Research Computing (CRC) machines. These use Unix and can be accessed remotely over the web. The advantages of CRC are

- It is free for Notre Dame faculty and students. This includes access to the most powerful version of Stata, Stata-MP
- For lightweight tasks, xstata can be used interactively. It has a GUI interface that looks pretty much the same as it does when Stata is running on your own machine.
- Bigger/monster tasks can be run by submitting a job script to the batch system. I have found this to be especially helpful when analyzing data sets with millions of records.
- If you wanted to, I suspect you could do pretty well with a modest laptop (maybe even a Chromebook) using Google Apps and CRC machines for most of your heavy duty work.

Potential downsides include

- A little more learning curve, but not too bad, at least for basic tasks.
- CRC may or may not have the most current updates of Stata (which often have bug fixes) installed. But if not and the old version is causing you problems, you can request an update. (This can be a concern with Lab and Classroom machines as well.)
- If you are working on different platforms (e.g. Win10 machines at school, CRC at home) you may have to copy your data and programs between them and change file locations in programs.
- Access is not as reliable as it is when running your own copy of Stata. CRC machines are down sometimes for maintenance. You need an Internet connection. CRC wants to keep the frontend machines available for interactive, non-computing intensive tasks, so if your interactive session uses more than 1 hour of CPU time, the CRC may kill it without warning and ask you to submit a job script to the batch system. If a bunch of people are using Stata at the same time there may not be enough licenses to go around. Your CRC account will leave Notre Dame when you do unless you make special arrangements.

Here are some things you need to know to use Stata with CRC.

The Essentials

Set up a free account. https://wiki.crc.nd.edu/w/index.php/How_to_Obtain_a_CRC_Account. You will get an email telling you what else you need to do to set up your account.

The CRC FAQ and the CRC Wiki will answer many questions. Refer to them as needed.
https://wiki.crc.nd.edu/w/index.php/FAQ
https://wiki.crc.nd.edu/w/index.php/Welcome

Install MobaXTerm. This will let you log on to the CRC machines and upload and download files. https://wiki.crc.nd.edu/w/index.php/MobaXterm

NOTE: The CRC web page suggests installing the portable version but I prefer the installer version (but you can get both if you want). Machines you can specify include *crcfe01.crc.nd.edu* and *crcfe02.crc.nd.edu*. The Social Sciences can also use *daccssfe.crc.nd.edu*. If off-campus, the latter requires that you have a VPN connection.

Install filezilla or cyberduck (optional). MobaXterm can upload and download files for you but I personally find an ftp program like filezilla easier to use. Plus it is free. https://filezilla-project.org/. When running filezilla, the host is the name of the CRC machine you want to connect to, e.g. crcfe01.crc.nd.edu. For port you should use 22. If you use some other ftp program you can tell it sftp should be used. I don't like it as well, but CRC recommends cyberduck (also free) https://wiki.crc.nd.edu/w/index.php/Cyberduck

Running xstata interactively

Once you get logged onto a CRC machine using MobaXTerm, it is easy to start running Stata interactively. Just give the commands

```
module load stata
xstata-se
```

If you need it you can load the more powerful `xstata-mp` (but if you just say `xstata` you get the weaker Stata IC). A Stata window will then pop up and you can use Stata like you would on other platforms.

Running Stata in batch (background) mode

Bigger Stata jobs should be run in batch mode. You will receive emails when the jobs start and finish. Your CRC account will have a file with the Stata output.

You need two kinds of files to submit a job. These can be prepared and/or edited with an ASCII text editor. The file creation and text editing capabilities in MobaXTerm are easy to use but if you prefer you can prepare the files on your own machine and then transfer them to CRC (e.g. you can create a file with notepad or the Stata do file editor). The files you need to create are

- A Stata .do file with the commands you want to run. Do your best to debug before submission. For example, you might try running your code interactively on a 1% sample.
- A Unix Job Script. These have the commands needed to run the job in batch mode. I am always using the same few job scripts. So, I have created two script files: *mystatase* and *mystatamp*. When I run them I specify the name of the Stata do file. However, if you prefer, you can create a new job script for every job you run.

Here is a sample Stata do file. Call it logit01.do.

```
use https://www3.nd.edu/~rwilliam/statafiles/logist.dta, clear
sum
logit grade gpa tuce i.psi
```

Here is a job script that will run it. I call it *mystatase*. [NOTE: Where it says *yournetid*, substitute your netid, e.g. jsmith@nd.edu]

```
#!/bin/csh
# mystatase - Job Script for submitting Stata/SE jobs. Syntax:
# qsub mystatase StataDoFilename

#$ -M yournetid@nd.edu      # Email address for job notification
#$ -m bea                   # Send mail when job begins, ends and aborts
#$ -pe smp 1                # Stata/SE can only use 1 processor

# Load and execute Stata/SE. -s gives smcl output. If you prefer text,
# specify -b instead of -s
module load stata
stata-se -s do $1
```

The –pe command tells Stata to only use one processor. To submit the job, type

```
qsub mystatase logit01
```

I find that Stata/SE is fine for many tasks. Further, if you try to use more processors, your job gets a lower priority and you may wait longer for results. But, if your job can take advantage of the multiprocessor capabilities of Stata MP, you can instead run something like *mystatamp*:

```
#!/bin/csh
# mystatamp - Job Script for submitting Stata/SE jobs. Syntax:
# qsub mystatase StataDoFilename

#$ -M yournetid@nd.edu      # Email address for job notification
#$ -m bea                   # Send mail when job begins, ends and aborts
#$ -pe smp 4                # Stata/MP can use 4 processors

# Load and execute Stata/MP. -s gives smcl output. If you prefer text,
# specify -b instead of -s
module load stata
stata-mp -s do $1
```

To actually submit the job, from within an interactive CRC session give the command

```
qsub mystatamp logit01
```

To check what jobs you have running, you can type

```
qstat -u yournetid
```

Jobs can be canceled or killed using the qdel command. The most common form is "qdel JobID", e.g. `qdel 1111691`, which kills the job that matches the Job ID. The Job ID will be in the email that CRC sent to you and will also appear when you use the qstat command.

OPTIONAL/ADVANCED

CRC quick start guide. https://wiki.crc.nd.edu/w/index.php/CRC_Quick_Start_Guide

Unix/Linux Command Reference. https://files.fosswire.com/2007/08/fwunixref.pdf .

CRC Stata Information. These explain in more detail how to run Stata jobs (and other jobs) in the background and interactively.
https://wiki.crc.nd.edu/w/index.php/Stata
https://wiki.crc.nd.edu/w/index.php/Submitting_a_STATA_Job_to_SGE
https://wiki.crc.nd.edu/w/index.php/Submitting_Batch/UGE_jobs
https://wiki.crc.nd.edu/w/index.php/DACCS_Cluster

Web Browser Access. I'm not sure I like it myself, but you can access the CRC machines through a web browser without installing any software. It actually isn't too bad if all you want to do is run xstata interactively. https://wiki.crc.nd.edu/w/index.php/Fastx. Or, just go straight to https://crcfe01.crc.nd.edu/ and tell it to launch an xterm session.

Installed Software. Other neat packages that are installed, e.g. R, Matlab, Mathematica. You may need to first get permission to use some of them.
https://wiki.crc.nd.edu/w/index.php/Installed_Applications

Rclone. I tried this and didn't like it all, but others might. "Rclone, sometimes known as rsync for the cloud, is a tool which is used to transfer data to or from a computer and a cloud hosted data storage center. Rclone can be used on the CRC front ends to upload/download data from your Google Drive or other Cloud Hosted Data storage to your AFS or /scratch spaces."
https://wiki.crc.nd.edu/w/index.php/Rclone