Challenge for data science position

John Lee

15 February 2016

Summary

The objective of this analysis was to:

- Download data subjects in the Human Connectome Project; namely, statistical descriptions of MRI scans and demographic information. The data is stored on ConnectomeDB and an Amazon repository.
- While controlling for total brain volume, perform a regression in order to assess the affect of both age and gender on the volume of each of 45 brain structures reported in the dataset.
- Plot the beta weights of the resulting regression models for each factor.

This analysis is compiled using the knitr package and RStudio (Version 0.99.865) on a Unix machine. The R Markdown file ('.Rmd') to generate the pdf is located in the same directory. In order to run a number of prerequisites exist.

Essential information for working with the Human Connectome Project (HCP) data is the documentation and the wiki. Guides to many of the prerequisites listed below are located in these helpful guides.

The analysis

After loading the appropriate R packages the data is loaded.

When querying the Amazon database using the s3cmd tool the .stat files for each subject were not always found. They are listed in the output below.

```
## [1] "Stat file for Subject 104012 not found"
## [1] "Stat file for Subject 105923 not found"
## [1] "Stat file for Subject 111514 not found"
## [1] "Stat file for Subject 146129 not found"
## [1] "Stat file for Subject 153732 not found"
## [1] "Stat file for Subject 156334 not found"
## [1] "Stat file for Subject 175540 not found"
## [1] "Stat file for Subject 192641 not found"
## [1] "Stat file for Subject 287248 not found"
## [1] "Stat file for Subject 512835 not found"
## [1] "Stat file for Subject 660951 not found"
## [1] "Stat file for Subject 662551 not found"
## [1] "Stat file for Subject 715950 not found"
## [1] "Stat file for Subject 725751 not found"
## [1] "Stat file for Subject 783462 not found"
## [1] "Stat file for Subject 825048 not found"
```

The demographic data was downloaded as a csv file from the ConnectomeDB website and save to the local working directory. Some parsing failures occur on loading this data but they are not relevant to our analysis:

```
## Warning: 1340 parsing failures.
## row           col   expected   actual
```

```
##   1 PSQI_BedTime   valid date 09:00:00
##   1 PSQI_GetUpTime valid date 15:30:00
##   2 PSQI_BedTime   valid date 22:30:00
##   2 PSQI_GetUpTime valid date 06:00:00
##   3 PSQI_BedTime   valid date 22:00:00
## ... .............. .......... ........
## .See problems(...) for more details.
```

364 subjects were found in common in the two datasets. The subsequent analysis was performed on these subjects.

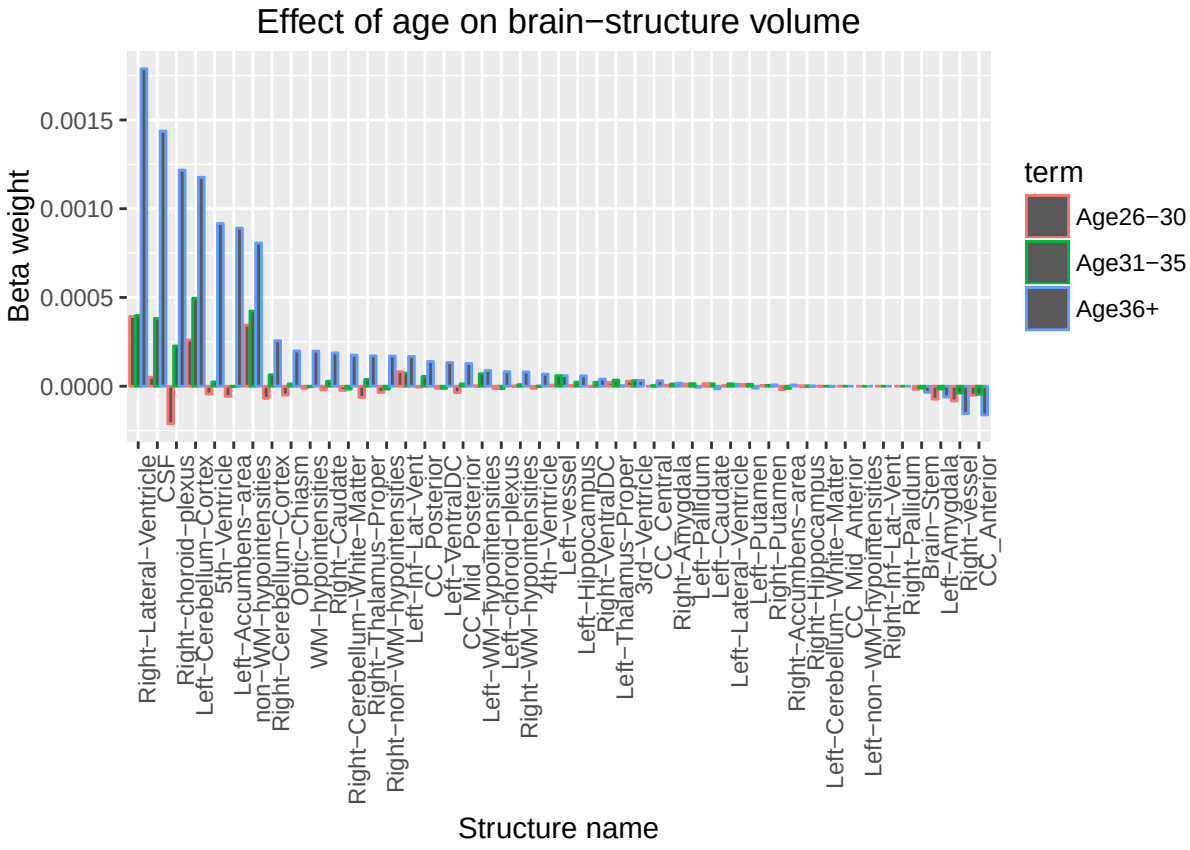
The linear regression models were constructed using the youngest age group and females as intercept terms. Below the regression coefficient estimates (beta weights) are listed for each other factor level where the p-value of the estimate is less than 0.01 :

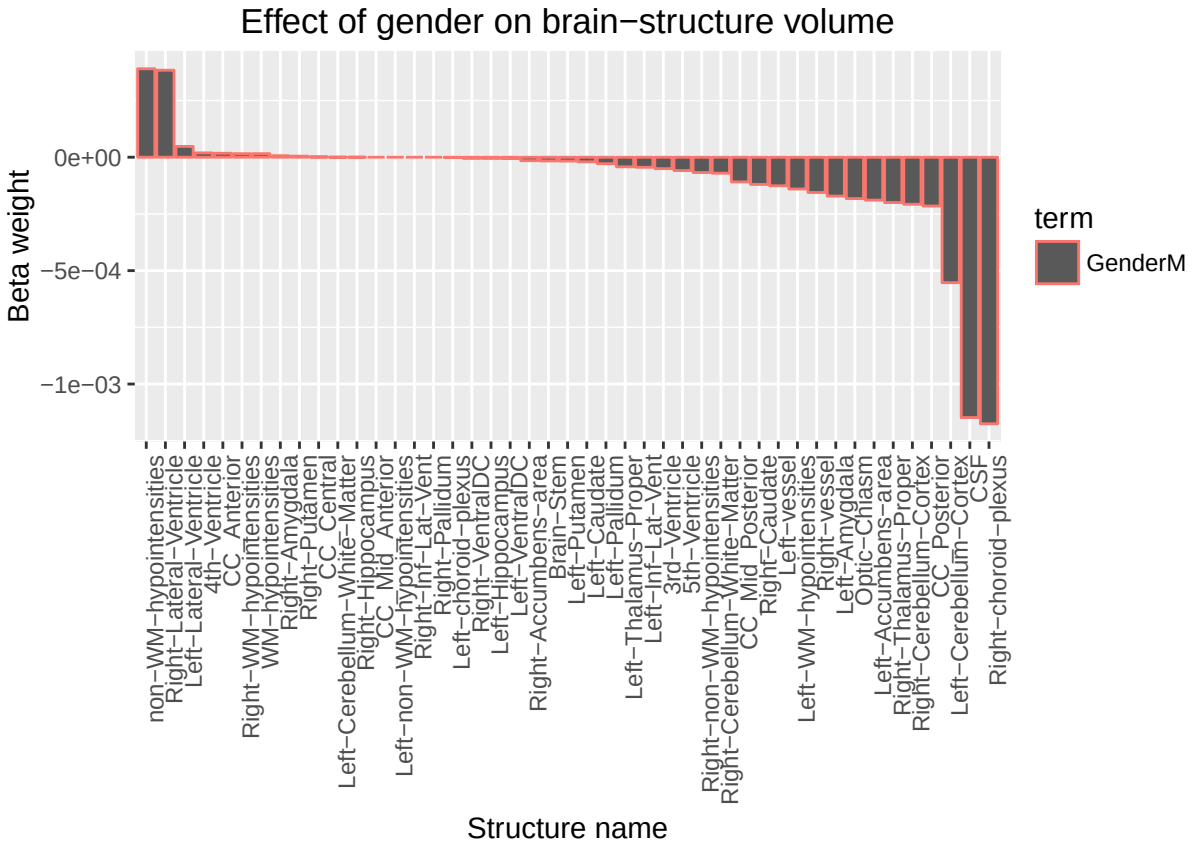
Structures where gender has an effect with a significance of $p<0.01$ are reported:

```
## Source: local data frame [20 x 4]
##
##                        StructName    term      estimate      p.value
##                            (fctr)   (chr)         (dbl)        (dbl)
## 1                  Left-Amygdala GenderM -1.713528e-04 1.860908e-07
## 2                   Right-vessel GenderM -1.554854e-04 2.477148e-06
## 3        Left-WM-hypointensities GenderM -1.399646e-04 2.615298e-06
## 4                  3rd-Ventricle GenderM -5.038816e-05 3.809125e-06
## 5   Right-non-WM-hypointensities GenderM -6.732362e-05 6.240131e-06
## 6           Left-Thalamus-Proper GenderM -4.225235e-05 2.591201e-05
## 7                    Left-vessel GenderM -1.260006e-04 4.860383e-05
## 8                   CC_Posterior GenderM -2.151112e-04 6.926817e-05
## 9                  Left-Pallidum GenderM -2.851652e-05 8.705825e-05
## 10 Right-Cerebellum-White-Matter GenderM -7.038933e-05 9.835125e-05
## 11       Right-Cerebellum-Cortex GenderM -2.073903e-04 2.328481e-04
## 12                 Right-Caudate GenderM -1.198034e-04 2.779326e-04
## 13                  Optic-Chiasm GenderM -1.824567e-04 2.846157e-04
## 14              CC_Mid_Posterior GenderM -1.088585e-04 5.191959e-04
## 15         Right-Thalamus-Proper GenderM -1.998438e-04 7.685924e-04
## 16        Left-Lateral-Ventricle GenderM  4.762258e-05 7.844089e-04
## 17        Left-Cerebellum-Cortex GenderM -5.529642e-04 1.103139e-03
## 18                           CSF GenderM -1.147832e-03 2.801905e-03
## 19          Right-choroid-plexus GenderM -1.175164e-03 3.315431e-03
## 20                 Left-Caudate GenderM -2.035439e-05 6.611725e-03
```

Structures where age has an effect with a significance of $p<0.05$ are reported:

```
## Source: local data frame [7 x 4]
##
##                      StructName     term      estimate    p.value
##                          (fctr)    (chr)         (dbl)      (dbl)
## 1            Right-Hippocampus Age26-30  1.626948e-06 0.01406029
## 2         Left-Thalamus-Proper Age31-35  3.444294e-05 0.01652720
## 3 Right-non-WM-hypointensities   Age36+  1.699954e-04 0.03514110
## 4               Right-Amygdala Age31-35  1.226150e-05 0.03729696
## 5       Left-Cerebellum-Cortex Age31-35  4.956014e-04 0.04171586
## 6         Right-Accumbens-area Age26-30 -1.925630e-05 0.04302877
## 7                3rd-Ventricle Age31-35  3.141709e-05 0.04319938
```


Effect of age on brain−structure volume
Beta weight
0.0015
0.0010
0.0005
0.0000
Right−Lateral−Ventricle
CSF
Right−choroid−plexus
Left−Cerebellum−Cortex
5th−Ventricle
Left−Accumbens−area
non−WM−hypointensities
Right−Cerebellum−Cortex
Optic−Chiasm
WM−hypointensities
Right−Caudate
Right−Cerebellum−White−Matter
Right−Thalamus−Proper
Right−non−WM−hypointensities
Left−Inf−Lat−Vent
CC_Posterior
Left−VentralDC
CC_Mid_Posterior
Left−WM−hypointensities
Left−choroid−plexus
Right−WM−hypointensities
4th−Ventricle
Left−vessel
Left−Hippocampus
Right−VentralDC
Left−Thalamus−Proper
3rd−Ventricle
CC_Central
Right−Amygdala
Left−Pallidum
Left−Caudate
Left−Lateral−Ventricle
Left−Putamen
Right−Putamen
Right−Accumbens−area
Right−Hippocampus
Left−Cerebellum−White−Matter
CC_Mid_Anterior
Left−non−WM−hypointensities
Right−Inf−Lat−Vent
Right−Pallidum
Brain−Stem
Left−Amygdala
Right−vessel
CC_Anterior
Structure name
term
Age26−30
Age31−35
Age36+



Notes

- I struggled quite a bit with getting the data. I've documented my attempts at downloading from the AWS server below. Regarding the demographic data I had a quick look at ascp, using REST etc. but I didn't have time to figure out the api for ConnectomeDB and so resorted to the manual download strategy in order to allow presentation of my skills in statistics, visualization, and reporting.
- Reported p-values are not adjusted for multiple testing.

Prerequisites for successful compilation of the Rmd file.

Working with the analysis files

Extract the analysis from the zip and set the working directory to the base folder.

Create ConnectomeDB account

Follow the documentation

Generate AWS credentials

Follow the documentation

Install s3cmd utility.

This can be installed using python. It must be python 2.X. I manage versions with pyenv hence the first command below to switch the appropropriate python version on my system. To install:

```
pyenv shell 2.7.10
pip install --upgrade pip
pip install s3cmd
```

Configure s3cmd

Help found here. The configuration writes a config file to the home directory. And is accessed during queries of the AWS bucket 's3://hcp-openaccess/HCP'. Once this is done `s3cmd ls` on the command line should list the buckets in the AWS account.

A latex installation

Errors will occur upon knitting if packages from Tex-live are missing. An example is the framed package (not included in the the basic tex installation on OSX). Error message mentions missing style or template file.

Using the s3cmd tool for AWS bucket download

As described in the human connectome wiki the s3cmd tool is a useful tool for downloading from the AWS bucket. The original command I was attempting to use was:

```
s3cmd sync -rv --include 'stats/aseg.stats' --exclude '*' s3://hcp-openaccess/HCP/ ./HCP/
```

It would have been a nice solution to downloading the '.stat' files because it maintains the directory structure and additional directories can subsequently be populated as required. Additionally I would be more certain of the missing '.stat' files. In the current analysis I searched in the same directory for each subject: a reasonable approach but not immune to error.

Using the previously mentioned sync command elicited a warning and executed painfully slowly. The error is detailed as an issue at the github repository and also discussed here. From what I gathered, its not a particularly useful warning and can be caused by a number of issues.

Things I tried to no avail:

- Originally I installed version 1.6.0 using the homebrew package manager. I then used python to install the more recent 1.6.1.
- I changed explicitly defined the bucket location as "us-east-1" and set my locale to en_US.
- I tried the utility awscli. This had an issue in that it could not find the buckets. It required a url as input rather than the bucket name. I didn't think I could troubleshoot this rapidly enough

After this, I just resorted to writing a loop and searching for the file for each subject individually. When more restricted queries were used it took very little time to download the files.