Obtain Data from Internet

Shu Guo

Thursday, February 12, 2015

Introduction

Today, more and more data is available online and is ready for us to download. Of course, we can first download the data by hand to the local drive of our computer and write the data later. But a more efficient way for this work would be obtain the data directly using our data analysis tools, and this method is also helpful for our reproducible research.

Download Data

The following R code first check if there is a fold named "data" in the current working directory, and create it if it does not exist. Then download a data file from internet and check the file in the data file

```
DownloadDate <- date() # Record the date
bankURL <- "http://robjhyndman.com/tsdldata/finance/bankdata.dat"
download.file(bankURL, destfile = "bankdata.txt", method = "auto")
DownloadDate
```

```
## [1] "Thu Feb 12 22:24:02 2015"
```

```
list.files()
```

```
## [1] "1-1 Reading Data from Internet.Rmd"
## [2] "1-1_Reading_Data_from_Internet.pdf"
## [3] "1-1_Reading_Data_from_Internet.Rmd"
## [4] "1-1_Reading_Data_from_Internet_cache"
## [5] "bankdata.txt"
```

Now, we can read the data file from local drive. Another way of doing this is to write the online data to our statistical software. The following R and SAS code read the same data online and create a

```
## Read the online data using R
bankdata <- read.delim(file=bankURL, header=F, sep="\t")
```

And the SAS code read the same data is:

```
FILENAME bankdata URL "http://robjhyndman.com/tsdldata/finance/bankdata.dat";
DATA bankdata;
    INFILE bankdata ;
    INPUT Bank_v1 Bank_v2 Bank_v3;
RUN;
```

```
The following notes were printed on SAS log:
```

```
NOTE: The infile BANKDATA is:
      Filename=http://robjhyndman.com/tsdldata/finance/bankdata.dat,
      Local Host Name=Shu,
      ...

      Lrecl=32767,Recfm=Variable

NOTE: 60 records were read from the infile BANKDATA.
      The minimum record length was 25.
      The maximum record length was 25.
NOTE: The data set WORK.BANKDATA has 60 observations and 3 variables.
NOTE: DATA statement used (Total process time):
      real time           1.37 seconds
      cpu time            0.09 seconds
```