Merge/append data using R/RStudio

(v. 1.0)

Oscar Torres-Reyna

otorres@princeton.edu

January 2011

Intro

Merge – adds variables to a dataset. This document will use –`merge`– function.

Merging two datasets require that both have at least one variable in common (either string or numeric). If string make sure the categories have the same spelling (i.e. country names, etc.).

Explore each dataset separately before merging. Make sure to use all possible common variables (for example, if merging two panel datasets you will need country and years).

Append – adds cases/observations to a dataset. This document will use the –`smartbind`– function from the –`gtools`– package. Appending two datasets require that both have variables with exactly the same name and spelling.

If using categorical data make sure the categories on both datasets refer to exactly the same thing (i.e. 1 “Agree”, 2”Disagree”, 3 “DK” on both).

Appending

```
# Data from 1960 to 1989

mydata6080 = read.csv("http://www.princeton.edu/~otorres/mydata6080.csv",
                      header=TRUE,
                      stringsAsFactors = FALSE)

table(mydata6080$year)
```

```
1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989
  95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95
```

```
# Data from 1990 to 2013

mydata9020 = read.csv("http://www.princeton.edu/~otorres/mydata9020.csv",
                      header=TRUE,
                      stringsAsFactors = FALSE)

table(mydata9020$year)
```

```
1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013
  95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95
```

```
# Appending the two datafiles. Full data from 1960 to 2013

library(gtools)

mydata = smartbind(mydata6080, mydata9020)

table(mydata$year)
```

```
1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990
  95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95

1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013
  95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95   95
```

Merging

```
# Merge join variables from two datafiles
# To an existing file called mydata (created in the previous slides)

str(mydata)
'data.frame':       5130 obs. of  6 variables:
 $ year   : int  1989 1980 1989 1981 1988 1989 1989 1980 1989 1987 ...
 $ country: chr  "Estonia" "Thailand" "Hong Kong SAR, China" "Thailand" ...
 $ gdppc  : num  NA 882 16973 915 16767 ...
 $ unemp  : num  0.6 0.9 1.1 1.3 1.4 1.4 1.5 1.6 1.6 1.7 ...
 $ export : num  NA 9.92e+09 NA 1.08e+10 NA ...
 $ import : num  NA 1.66e+10 NA 1.67e+10 NA ...

# we are going to add one more variable from a dataset called mydatapol
# Notice the common variables 'country' and 'year' which we are going to use
# to match each row on both files

mydatapol = read.csv("http://www.princeton.edu/~otorres/mydatapol.csv",
                     header=TRUE,
                     stringsAsFactors = FALSE)
str(mydatapol)
'data.frame':       3655 obs. of  3 variables:
 $ year    : int  1996 1996 1996 1996 1996 1996 1996 1996 1996 1996 ...
 $ country : chr  "Afghanistan" "Albania" "Algeria" "American Samoa" ...
 $ politics: num  1.12 24.05 16.69 NA 89.51 ...

# Merging the two files

mydata <- merge(mydata, mydatapol, by=c("country","year"), all=TRUE)

mydata <- mydata[order(mydata$year,mydata$country),]  # Sorting data by year/country
```

(see next slide)

Merging

```
# The file mydata has now one extra variable

str(mydata)

'data.frame':    7170 obs. of  7 variables:
 $ country : chr  "Algeria" "Argentina" "Australia" "Austria" ...
 $ year    : int  1960 1960 1960 1960 1960 1960 1960 1960 1960 1960 ...
 $ gdppc   : num  1766 3732 13469 10862 NA ...
 $ unemp   : num  NA NA NA NA NA NA NA NA NA NA ...
 $ export  : num  2.24e+10 3.98e+09 1.02e+10 1.03e+10 NA ...
 $ import  : num  1.61e+10 5.43e+09 1.01e+10 1.14e+10 NA ...
 $ politics: num  NA NA NA NA NA NA NA NA NA NA ...


# Common variables do not need to have the same name, for example:

mydata <- merge(mydata1, mydata2,
                by.x=c("country","year"),
                by.y=c("nations","time"), all = TRUE)


# If one file has 'country/year', but the other only 'country',
# you can still merge.
# Values in the data with only 'country' will repeat in the 'country/year' data

mydata <- merge(mydata1, mydata2, by=c("country"), all=TRUE)
```

Data source

World Development Indicators (World Bank)
http://databank.worldbank.org/data/reports.aspx?source=world-development-indicators