Merge/Append using R

(draft)

Oscar Torres-Reyna

Data Consultant

otorres@princeton.edu

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 –`rbind`– function.

Appending two datasets require that both have variables with *exactly* the same name. 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).

MERGE – EXAMPLE 1

mydata1

Data Editor

File Edit Help

	country	year	y	y_bin	x1	x2	x3
1	A	2000	1343	1	0.28	-1.11	0.28
2	A	2001	-1900	0	0.32	-0.95	0.49
3	A	2002	-11	0	0.36	-0.79	0.7
4	A	2003	2646	1	0.25	-0.89	-0.09
5	B	2000	-5935	0	-0.08	1.43	0.02
6	B	2001	-712	0	0.11	1.65	0.26
7	B	2002	-1933	0	0.35	1.59	-0.23
8	B	2003	3073	1	0.73	1.69	0.26
9	C	2000	-1292	0	1.31	-1.29	0.2
10	C	2001	-3416	0	1.18	-1.34	0.28
11	C	2002	-356	0	1.26	-1.26	0.37
12	C	2003	1225	1	1.42	-1.31	-0.38

mydata2

Data Editor

File Edit Help

	country	year	x4	x5	x6
1	A	2000	10	1	9
2	A	2001	7	1	9
3	A	2002	7	9	4
4	A	2003	1	2	3
5	B	2000	0	5	6
6	B	2001	5	8	5
7	B	2002	9	4	5
8	B	2003	1	5	1
9	C	2000	4	5	4
10	C	2001	6	9	6
11	C	2002	6	5	3
12	C	2003	7	3	3

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

edit(mydata)
```

Data Editor

File Edit Help

	country	year	y	y_bin	x1	x2	x3	x4	x5	x6
1	A	2000	1343	1	0.28	-1.11	0.28	10	1	9
2	A	2001	-1900	0	0.32	-0.95	0.49	7	1	9
3	A	2002	-11	0	0.36	-0.79	0.7	7	9	4
4	A	2003	2646	1	0.25	-0.89	-0.09	1	2	3
5	B	2000	-5935	0	-0.08	1.43	0.02	0	5	6
6	B	2001	-712	0	0.11	1.65	0.26	5	8	5
7	B	2002	-1933	0	0.35	1.59	-0.23	9	4	5
8	B	2003	3073	1	0.73	1.69	0.26	1	5	1
9	C	2000	-1292	0	1.31	-1.29	0.2	4	5	4
10	C	2001	-3416	0	1.18	-1.34	0.28	6	9	6
11	C	2002	-356	0	1.26	-1.26	0.37	6	5	3
12	C	2003	1225	1	1.42	-1.31	-0.38	7	3	3

MERGE – EXAMPLE 2 (one dataset missing a country)

mydata1

Data Editor

File Edit Help

	country	year	y	y_bin	x1	x2	x3
1	A	2000	1343	1	0.28	-1.11	0.28
2	A	2001	-1900	0	0.32	-0.95	0.49
3	A	2002	-11	0	0.36	-0.79	0.7
4	A	2003	2646	1	0.25	-0.89	-0.09
5	B	2000	-5935	0	-0.08	1.43	0.02
6	B	2001	-712	0	0.11	1.65	0.26
7	B	2002	-1933	0	0.35	1.59	-0.23
8	B	2003	3073	1	0.73	1.69	0.26
9	C	2000	-1292	0	1.31	-1.29	0.2
10	C	2001	-3416	0	1.18	-1.34	0.28
11	C	2002	-356	0	1.26	-1.26	0.37
12	C	2003	1225	1	1.42	-1.31	-0.38

mydata3

Data Editor

File Edit Help

	country	year	x4	x5	x6
1	A	2000	10	1	9
2	A	2001	7	1	9
3	A	2002	7	9	4
4	A	2003	1	2	3
5	B	2000	0	5	6
6	B	2001	5	8	5
7	B	2002	9	4	5
8	B	2003	1	5	1

`Merge` merges only common cases to both datasets

```
mydata <- merge(mydata1, mydata3, by=c("country","year"))
```

```
edit(mydata)
```

Data Editor

File Edit Help

	country	year	y	y_bin	x1	x2	x3	x4	x5	x6
1	A	2000	1343	1	0.28	-1.11	0.28	10	1	9
2	A	2001	-1900	0	0.32	-0.95	0.49	7	1	9
3	A	2002	-11	0	0.36	-0.79	0.7	7	9	4
4	A	2003	2646	1	0.25	-0.89	-0.09	1	2	3
5	B	2000	-5935	0	-0.08	1.43	0.02	0	5	6
6	B	2001	-712	0	0.11	1.65	0.26	5	8	5
7	B	2002	-1933	0	0.35	1.59	-0.23	9	4	5
8	B	2003	3073	1	0.73	1.69	0.26	1	5	1

MERGE – EXAMPLE 2 (cont.) – including all data from both datasets

mydata1

Data Editor

File Edit Help

	country	year	y	y_bin	x1	x2	x3
1	A	2000	1343	1	0.28	-1.11	0.28
2	A	2001	-1900	0	0.32	-0.95	0.49
3	A	2002	-11	0	0.36	-0.79	0.7
4	A	2003	2646	1	0.25	-0.89	-0.09
5	B	2000	-5935	0	-0.08	1.43	0.02
6	B	2001	-712	0	0.11	1.65	0.26
7	B	2002	-1933	0	0.35	1.59	-0.23
8	B	2003	3073	1	0.73	1.69	0.26
9	C	2000	-1292	0	1.31	-1.29	0.2
10	C	2001	-3416	0	1.18	-1.34	0.28
11	C	2002	-356	0	1.26	-1.26	0.37
12	C	2003	1225	1	1.42	-1.31	-0.38

mydata3

Data Editor

File Edit Help

	country	year	x4	x5	x6
1	A	2000	10	1	9
2	A	2001	7	1	9
3	A	2002	7	9	4
4	A	2003	1	2	3
5	B	2000	0	5	6
6	B	2001	5	8	5
7	B	2002	9	4	5
8	B	2003	1	5	1

Adding the option "`all=TRUE`" includes all cases from both datasets.

```
mydata <- merge(mydata1, mydata3, by=c("country","year"), all=TRUE)

edit(mydata)
```

Data Editor

File Edit Help

	country	year	y	y_bin	x1	x2	x3	x4	x5	x6
1	A	2000	1343	1	0.28	-1.11	0.28	10	1	9
2	A	2001	-1900	0	0.32	-0.95	0.49	7	1	9
3	A	2002	-11	0	0.36	-0.79	0.7	7	9	4
4	A	2003	2646	1	0.25	-0.89	-0.09	1	2	3
5	B	2000	-5935	0	-0.08	1.43	0.02	0	5	6
6	B	2001	-712	0	0.11	1.65	0.26	5	8	5
7	B	2002	-1933	0	0.35	1.59	-0.23	9	4	5
8	B	2003	3073	1	0.73	1.69	0.26	1	5	1
9	C	2000	-1292	0	1.31	-1.29	0.2	NA	NA	NA
10	C	2001	-3416	0	1.18	-1.34	0.28	NA	NA	NA
11	C	2002	-356	0	1.26	-1.26	0.37	NA	NA	NA
12	C	2003	1225	1	1.42	-1.31	-0.38	NA	NA	NA

MERGE – EXAMPLE 3 (many to one)

mydata1

Data Editor

File Edit Help

	country	year	y	y_bin	x1	x2	x3
1	A	2000	1343	1	0.28	-1.11	0.28
2	A	2001	-1900	0	0.32	-0.95	0.49
3	A	2002	-11	0	0.36	-0.79	0.7
4	A	2003	2646	1	0.25	-0.89	-0.09
5	B	2000	-5935	0	-0.08	1.43	0.02
6	B	2001	-712	0	0.11	1.65	0.26
7	B	2002	-1933	0	0.35	1.59	-0.23
8	B	2003	3073	1	0.73	1.69	0.26
9	C	2000	-1292	0	1.31	-1.29	0.2
10	C	2001	-3416	0	1.18	-1.34	0.28
11	C	2002	-356	0	1.26	-1.26	0.37
12	C	2003	1225	1	1.42	-1.31	-0.38

mydata4

Data Editor

File Edit Help

	country	x7
1	A	100
2	B	200
3	C	300

```
mydata <- merge(mydata1, mydata4, by=c("country"))

edit(mydata)
```

Data Editor

File Edit Help

	country	year	y	y_bin	x1	x2	x3	x7
1	A	2000	1343	1	0.28	-1.11	0.28	100
2	A	2001	-1900	0	0.32	-0.95	0.49	100
3	A	2002	-11	0	0.36	-0.79	0.7	100
4	A	2003	2646	1	0.25	-0.89	-0.09	100
5	B	2000	-5935	0	-0.08	1.43	0.02	200
6	B	2001	-712	0	0.11	1.65	0.26	200
7	B	2002	-1933	0	0.35	1.59	-0.23	200
8	B	2003	3073	1	0.73	1.69	0.26	200
9	C	2000	-1292	0	1.31	-1.29	0.2	300
10	C	2001	-3416	0	1.18	-1.34	0.28	300
11	C	2002	-356	0	1.26	-1.26	0.37	300
12	C	2003	1225	1	1.42	-1.31	-0.38	300

MERGE – EXAMPLE 4 (common ids have different name)

mydata1

Data Editor

File Edit Help

	country	year	y	y_bin	x1	x2	x3
1	A	2000	1343	1	0.28	-1.11	0.28
2	A	2001	-1900	0	0.32	-0.95	0.49
3	A	2002	-11	0	0.36	-0.79	0.7
4	A	2003	2646	1	0.25	-0.89	-0.09
5	B	2000	-5935	0	-0.08	1.43	0.02
6	B	2001	-712	0	0.11	1.65	0.26
7	B	2002	-1933	0	0.35	1.59	-0.23
8	B	2003	3073	1	0.73	1.69	0.26
9	C	2000	-1292	0	1.31	-1.29	0.2
10	C	2001	-3416	0	1.18	-1.34	0.28
11	C	2002	-356	0	1.26	-1.26	0.37
12	C	2003	1225	1	1.42	-1.31	-0.38

mydata5

Data Editor

File Edit Help

	nations	time	x4	x5	x6
1	A	2000	10	1	9
2	A	2001	7	1	9
3	A	2002	7	9	4
4	A	2003	1	2	3
5	B	2000	0	5	6
6	B	2001	5	8	5
7	B	2002	9	4	5
8	B	2003	1	5	1
9	C	2000	4	5	4
10	C	2001	6	9	6
11	C	2002	6	5	3
12	C	2003	7	3	3

When common ids have different names use `by.x` and `by.y` to match them. R will keep the name of the first dataset (`by.x`)

```
mydata <- merge(mydata1, mydata5, by.x=c("country","year"), by.y=c("nations","time"))

edit(mydata)
```

Data Editor

File Edit Help

	country	year	y	y_bin	x1	x2	x3	x4	x5	x6
1	A	2000	1343	1	0.28	-1.11	0.28	10	1	9
2	A	2001	-1900	0	0.32	-0.95	0.49	7	1	9
3	A	2002	-11	0	0.36	-0.79	0.7	7	9	4
4	A	2003	2646	1	0.25	-0.89	-0.09	1	2	3
5	B	2000	-5935	0	-0.08	1.43	0.02	0	5	6
6	B	2001	-712	0	0.11	1.65	0.26	5	8	5
7	B	2002	-1933	0	0.35	1.59	-0.23	9	4	5
8	B	2003	3073	1	0.73	1.69	0.26	1	5	1
9	C	2000	-1292	0	1.31	-1.29	0.2	4	5	4
10	C	2001	-3416	0	1.18	-1.34	0.28	6	9	6
11	C	2002	-356	0	1.26	-1.26	0.37	6	5	3
12	C	2003	1225	1	1.42	-1.31	-0.38	7	3	3

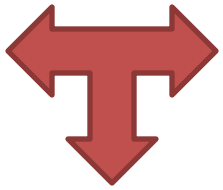
MERGE – EXAMPLE 5 (different variables, same name)

mydata1

Data Editor

File Edit Help

	country	year	y	y_bin	x1	x2	x3
1	A	2000	1343	1	0.28	-1.11	0.28
2	A	2001	-1900	0	0.32	-0.95	0.49
3	A	2002	-11	0	0.36	-0.79	0.7
4	A	2003	2646	1	0.25	-0.89	-0.09
5	B	2000	-5935	0	-0.08	1.43	0.02
6	B	2001	-712	0	0.11	1.65	0.26
7	B	2002	-1933	0	0.35	1.59	-0.23
8	B	2003	3073	1	0.73	1.69	0.26
9	C	2000	-1292	0	1.31	-1.29	0.2
10	C	2001	-3416	0	1.18	-1.34	0.28
11	C	2002	-356	0	1.26	-1.26	0.37
12	C	2003	1225	1	1.42	-1.31	-0.38



mydata6

Data Editor

File Edit Help

	nations	time	x2	x3	x4
1	A	2000	10	1	9
2	A	2001	7	1	9
3	A	2002	7	9	4
4	A	2003	1	2	3
5	B	2000	0	5	6
6	B	2001	5	8	5
7	B	2002	9	4	5
8	B	2003	1	5	1
9	C	2000	4	5	4
10	C	2001	6	9	6
11	C	2002	6	5	3
12	C	2003	7	3	3

When common ids have different names use `by.x` and `by.y` to match them. R will keep the name of the first dataset (`by.x`)

When different variables from two different dataset have the same name, R will assign a suffix `.x` or `.y` to make them unique and to identify which dataset they are coming from.

```
mydata <- merge(mydata1, mydata6, by.x=c("country","year"), by.y=c("nations","time"))
```

```
edit(mydata)
```

Data Editor

File Edit Help

	country	year	y	y_bin	x1	x2.x	x3.x	x2.y	x3.y	x4
1	A	2000	1343	1	0.28	-1.11	0.28	10	1	9
2	A	2001	-1900	0	0.32	-0.95	0.49	7	1	9
3	A	2002	-11	0	0.36	-0.79	0.7	7	9	4
4	A	2003	2646	1	0.25	-0.89	-0.09	1	2	3
5	B	2000	-5935	0	-0.08	1.43	0.02	0	5	6
6	B	2001	-712	0	0.11	1.65	0.26	5	8	5
7	B	2002	-1933	0	0.35	1.59	-0.23	9	4	5
8	B	2003	3073	1	0.73	1.69	0.26	1	5	1
9	C	2000	-1292	0	1.31	-1.29	0.2	4	5	4
10	C	2001	-3416	0	1.18	-1.34	0.28	6	9	6
11	C	2002	-356	0	1.26	-1.26	0.37	6	5	3
12	C	2003	1225	1	1.42	-1.31	-0.38	7	3	3

APPEND

APPEND– EXAMPLE 1

mydata7

Data Editor

File Edit Help

	country	year	y	y_bin	x1	x2	x3
1	A	2000	1343	1	0.28	-1.11	0.28
2	A	2001	-1900	0	0.32	-0.95	0.49
3	B	2000	-5935	0	-0.08	1.43	0.02
4	B	2001	-712	0	0.11	1.65	0.26
5	C	2000	-1292	0	1.31	-1.29	0.2
6	C	2001	-3416	0	1.18	-1.34	0.28

mydata8

Data Editor

File Edit Help

	country	year	y	y_bin	x1	x2	x3
1	A	2002	-11	0	0.36	-0.79	0.7
2	A	2003	2646	1	0.25	-0.89	-0.09
3	B	2002	-1933	0	0.35	1.59	-0.23
4	B	2003	3073	1	0.73	1.69	0.26
5	C	2002	-356	0	1.26	-1.26	0.37
6	C	2003	1225	1	1.42	-1.31	-0.38

```
mydata <- rbind(mydata7, mydata8)

edit(mydata)
```

Data Editor

File Edit Help

	country	year	y	y_bin	x1	x2	x3
1	A	2000	1343	1	0.28	-1.11	0.28
2	A	2001	-1900	0	0.32	-0.95	0.49
3	B	2000	-5935	0	-0.08	1.43	0.02
4	B	2001	-712	0	0.11	1.65	0.26
5	C	2000	-1292	0	1.31	-1.29	0.2
6	C	2001	-3416	0	1.18	-1.34	0.28
7	A	2002	-11	0	0.36	-0.79	0.7
8	A	2003	2646	1	0.25	-0.89	-0.09
9	B	2002	-1933	0	0.35	1.59	-0.23
10	B	2003	3073	1	0.73	1.69	0.26
11	C	2002	-356	0	1.26	-1.26	0.37
12	C	2003	1225	1	1.42	-1.31	-0.38

APPEND – EXAMPLE 1 (cont.) – sorting by country/year

Notice the square brackets and parenthesis

```
attach(mydata)
mydata_sorted <- mydata[order(country, year),]
detach(mydata)
edit(mydata_sorted)
```

mydata_sorted

Data Editor

File Edit Help

	row.names	country	year	y	y_bin	x1	x2	x3
1	1	A	2000	1343	1	0.28	-1.11	0.28
2	2	A	2001	-1900	0	0.32	-0.95	0.49
3	7	A	2002	-11	0	0.36	-0.79	0.7
4	8	A	2003	2646	1	0.25	-0.89	-0.09
5	3	B	2000	-5935	0	-0.08	1.43	0.02
6	4	B	2001	-712	0	0.11	1.65	0.26
7	9	B	2002	-1933	0	0.35	1.59	-0.23
8	10	B	2003	3073	1	0.73	1.69	0.26
9	5	C	2000	-1292	0	1.31	-1.29	0.2
10	6	C	2001	-3416	0	1.18	-1.34	0.28
11	11	C	2002	-356	0	1.26	-1.26	0.37
12	12	C	2003	1225	1	1.42	-1.31	-0.38

APPEND– EXAMPLE 2 – one dataset missing one variable

mydata7

Data Editor

File Edit Help

	country	year	y	y_bin	x1	x2	x3
1	A	2000	1343	1	0.28	-1.11	0.28
2	A	2001	-1900	0	0.32	-0.95	0.49
3	B	2000	-5935	0	-0.08	1.43	0.02
4	B	2001	-712	0	0.11	1.65	0.26
5	C	2000	-1292	0	1.31	-1.29	0.2
6	C	2001	-3416	0	1.18	-1.34	0.28

mydata9

Data Editor

File Edit Help

	country	year	y	y_bin	x1	x2
1	A	2002	-11	0	0.36	-0.79
2	A	2003	2646	1	0.25	-0.89
3	B	2002	-1933	0	0.35	1.59
4	B	2003	3073	1	0.73	1.69
5	C	2002	-356	0	1.26	-1.26
6	C	2003	1225	1	1.42	-1.31

If one variable is missing in one dataset you will get an error message

mydata <- rbind(mydata7, mydata9)

```
Error in rbind(deparse.level, ...) :
   numbers of columns of arguments do not match
```

Possible solutions:

Option A) Drop the extra variable from one of the datasets (in this case mydata7)

mydata7$x3 <- NULL

Option B) Create the variable with missing values in the incomplete dataset (in this case mydata9)

mydata9$x3 <- NA

Run the –`rbind`-- function again.

References/Useful links

- Main references for this document:
  - UCLA R class notes: http://www.ats.ucla.edu/stat/r/notes/managing.htm
  - Quick-R: http://www.statmethods.net/management/merging.html

- DSS Online Training Section http://dss.princeton.edu/training/
- Princeton DSS Libguides http://libguides.princeton.edu/dss
- John Fox's site http://socserv.mcmaster.ca/jfox/
- Quick-R http://www.statmethods.net/
- UCLA Resources to learn and use R http://www.ats.ucla.edu/stat/R/
- DSS - R http://dss.princeton.edu/online_help/stats_packages/r

References/Recommended books

- *An R Companion to Applied Regression*, Second Edition / John Fox , Sanford Weisberg, Sage Publications, 2011
- *Data Manipulation with R* / Phil Spector, Springer, 2008
- *Applied Econometrics with R* / Christian Kleiber, Achim Zeileis, Springer, 2008
- *Introductory Statistics with R* / Peter Dalgaard, Springer, 2008
- *Complex Surveys. A guide to Analysis Using R* / Thomas Lumley, Wiley, 2010
- *Applied Regression Analysis and Generalized Linear Models* / John Fox, Sage, 2008
- *R for Stata Users* / Robert A. Muenchen, Joseph Hilbe, Springer, 2010
- *Introduction to econometrics* / James H. Stock, Mark W. Watson. 2nd ed., Boston: Pearson Addison Wesley, 2007.
- *Data analysis using regression and multilevel/hierarchical models* / Andrew Gelman, Jennifer Hill. Cambridge ; New York : Cambridge University Press, 2007.
- *Econometric analysis* / William H. Greene. 6th ed., Upper Saddle River, N.J. : Prentice Hall, 2008.
- *Designing Social Inquiry: Scientific Inference in Qualitative Research* / *Gary King, Robert O. Keohane, Sidney Verba, Princeton University Press*, 1994.
- *Unifying Political Methodology: The Likelihood Theory of Statistical Inference* / Gary King, Cambridge University Press, 1989
- *Statistical Analysis: an interdisciplinary introduction to univariate & multivariate methods /* Sam Kachigan, New York : Radius Press, c1986
- *Statistics with Stata (updated for version 9) /* Lawrence Hamilton, Thomson Books/Cole, 2006