Merge/Append using R
(draft)

Oscar Torres-Reyna
Data Consultant
otorres@princeton.edu

http://dss.princeton.edu/training/
**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**

```r
mydata <- merge(mydata1, mydata2, by=c("country", "year"))
edit(mydata)
```
MERGE – EXAMPLE 2 (one dataset missing a country)

mydata1

<table>
<thead>
<tr>
<th>country</th>
<th>year</th>
<th>y</th>
<th>y_bin</th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2000</td>
<td>1343</td>
<td>1</td>
<td>0.28</td>
<td>-1.11</td>
<td>0.28</td>
</tr>
<tr>
<td>A</td>
<td>2001</td>
<td>-1900</td>
<td>0</td>
<td>0.32</td>
<td>-0.95</td>
<td>0.49</td>
</tr>
<tr>
<td>A</td>
<td>2002</td>
<td>-11</td>
<td>0</td>
<td>0.36</td>
<td>-0.79</td>
<td>0.7</td>
</tr>
<tr>
<td>A</td>
<td>2003</td>
<td>2646</td>
<td>1</td>
<td>0.28</td>
<td>-0.89</td>
<td>-0.09</td>
</tr>
<tr>
<td>B</td>
<td>2000</td>
<td>-5935</td>
<td>0</td>
<td>-0.08</td>
<td>1.43</td>
<td>0.02</td>
</tr>
<tr>
<td>B</td>
<td>2001</td>
<td>-712</td>
<td>0</td>
<td>0.11</td>
<td>1.65</td>
<td>0.26</td>
</tr>
<tr>
<td>B</td>
<td>2002</td>
<td>-1533</td>
<td>0</td>
<td>0.35</td>
<td>1.59</td>
<td>-0.25</td>
</tr>
<tr>
<td>B</td>
<td>2003</td>
<td>3073</td>
<td>1</td>
<td>0.73</td>
<td>1.69</td>
<td>0.26</td>
</tr>
<tr>
<td>C</td>
<td>2000</td>
<td>-1292</td>
<td>0</td>
<td>1.31</td>
<td>-1.29</td>
<td>0.2</td>
</tr>
<tr>
<td>C</td>
<td>2001</td>
<td>-3416</td>
<td>0</td>
<td>1.18</td>
<td>-1.34</td>
<td>0.28</td>
</tr>
<tr>
<td>C</td>
<td>2002</td>
<td>-356</td>
<td>0</td>
<td>1.26</td>
<td>-1.26</td>
<td>0.37</td>
</tr>
<tr>
<td>C</td>
<td>2003</td>
<td>1225</td>
<td>1</td>
<td>1.42</td>
<td>-1.31</td>
<td>-0.38</td>
</tr>
</tbody>
</table>

mydata3

<table>
<thead>
<tr>
<th>country</th>
<th>year</th>
<th>x4</th>
<th>x5</th>
<th>x6</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2000</td>
<td>10</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>A</td>
<td>2001</td>
<td>7</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>A</td>
<td>2002</td>
<td>7</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>A</td>
<td>2003</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>B</td>
<td>2000</td>
<td>0</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>B</td>
<td>2001</td>
<td>5</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>B</td>
<td>2002</td>
<td>9</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>B</td>
<td>2003</td>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

Merge merges only common cases to both datasets

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

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

Adding the option “all=TRUE” includes all cases from both datasets.

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

edit(mydata)
MERGE – EXAMPLE 3 (many to one)

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

```
mydata
```

```
mydata1
```

```
mydata4
```
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`)

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

edit(mydata)
```
MERGE – EXAMPLE 5 (different variables, same name)

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 datasets 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.

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

```
edit(mydata)
```
APPEND
### APPEND—EXAMPLE 1

**mydata7**

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

```
# mydata7

country | year | y  | y_bin | x1   | x2   | x3  
--------|------|----|-------|------|------|-----
       1 | A    | 2000 | 1345  | 0.28 | -1.11| 0.28
       2 | A    | 2001 | -1900 | 0.32 | -0.95| 0.49
       3 | B    | 2000 | -5935 | -0.08| 1.43 | 0.02
       4 | B    | 2001 | -712  | 0.11 | 1.65 | 0.26
       5 | C    | 2000 | -1292 | 1.31 | -1.29| 0.2
       6 | C    | 2001 | -3416 | 1.18 | -1.34| 0.28
```

**mydata8**

```
edit(mydata)
```

```
# mydata8

country | year | y  | y_bin | x1   | x2   | x3  
--------|------|----|-------|------|------|-----
       7 | A    | 2002 | -11   | 0.36 | -0.79| 0.7
       8 | A    | 2003 | 2646  | 0.25 | -0.89| -0.09
       9 | B    | 2002 | -1933 | 0.35 | 1.59 | -0.23
      10 | B    | 2003 | 3073  | 0.73 | 1.69 | 0.26
      11 | C    | 2002 | -356  | 1.26 | -1.26| 0.37
      12 | C    | 2003 | 1225  | 1.42 | -1.31| -0.38
```
Notice the square brackets and parenthesis

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

mydata_sorted

<table>
<thead>
<tr>
<th>row.names</th>
<th>country</th>
<th>year</th>
<th>y</th>
<th>y_bin</th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>2000</td>
<td>1343</td>
<td>1</td>
<td>0.28</td>
<td>-1.11</td>
<td>0.28</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>2001</td>
<td>-1900</td>
<td>0</td>
<td>0.32</td>
<td>-0.95</td>
<td>0.49</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>2002</td>
<td>-11</td>
<td>0</td>
<td>0.36</td>
<td>-0.79</td>
<td>0.7</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>2003</td>
<td>2646</td>
<td>1</td>
<td>0.25</td>
<td>-0.89</td>
<td>-0.09</td>
</tr>
<tr>
<td>5</td>
<td>B</td>
<td>2000</td>
<td>-5935</td>
<td>0</td>
<td>-0.08</td>
<td>1.43</td>
<td>0.02</td>
</tr>
<tr>
<td>6</td>
<td>B</td>
<td>2001</td>
<td>-712</td>
<td>0</td>
<td>0.11</td>
<td>1.65</td>
<td>0.26</td>
</tr>
<tr>
<td>7</td>
<td>B</td>
<td>2002</td>
<td>-1933</td>
<td>0</td>
<td>0.35</td>
<td>1.59</td>
<td>-0.23</td>
</tr>
<tr>
<td>8</td>
<td>B</td>
<td>2003</td>
<td>3073</td>
<td>1</td>
<td>0.73</td>
<td>1.69</td>
<td>0.26</td>
</tr>
<tr>
<td>9</td>
<td>C</td>
<td>2000</td>
<td>-1292</td>
<td>0</td>
<td>1.31</td>
<td>-1.29</td>
<td>0.2</td>
</tr>
<tr>
<td>10</td>
<td>C</td>
<td>2001</td>
<td>-3416</td>
<td>0</td>
<td>1.18</td>
<td>-1.34</td>
<td>0.28</td>
</tr>
<tr>
<td>11</td>
<td>C</td>
<td>2002</td>
<td>-356</td>
<td>0</td>
<td>1.26</td>
<td>-1.26</td>
<td>0.37</td>
</tr>
<tr>
<td>12</td>
<td>C</td>
<td>2003</td>
<td>1225</td>
<td>1</td>
<td>1.42</td>
<td>-1.31</td>
<td>-0.38</td>
</tr>
</tbody>
</table>
If one variable is missing in one dataset you will get an error message

```r
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)

```r
mydata7$x3 <- NULL
```

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

```r
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

- *Data Manipulation with R* / Phil Spector, Springer, 2008
- *Statistics with Stata (updated for version 9)* / Lawrence Hamilton, Thomson Books/Cole, 2006