

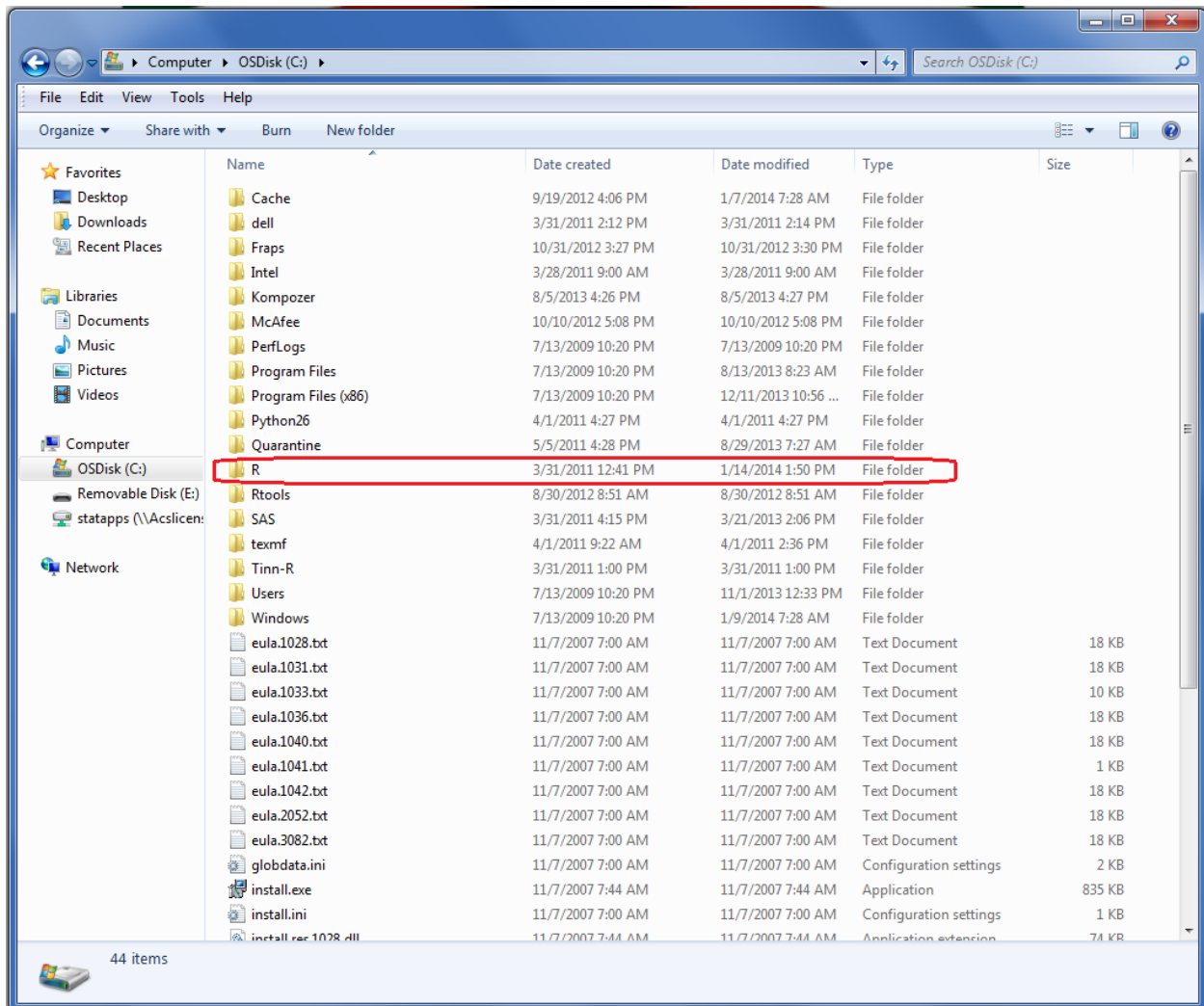
Your one-stop multiple missing value imputation shop: R 2.15.0 with the rrp package.

Dr. Jon Starkweather, Research and Statistical Support consultant.

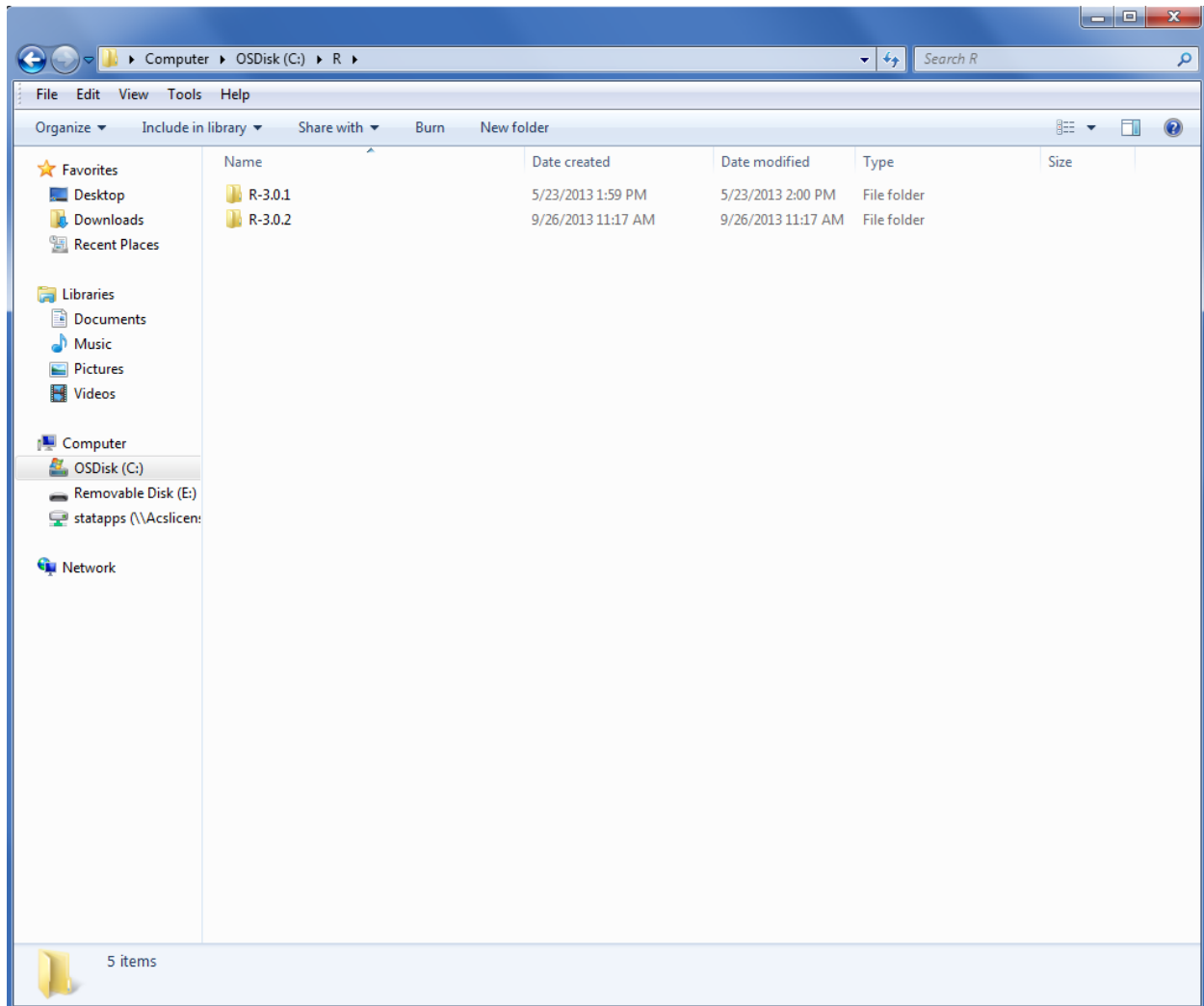
This month we provide a recommendation for dealing with multiple missing value imputation. Of course, every researcher must deal with missing values at some point. The first key issue when dealing with missing values is attempting to determine if the values are missing at random or if there is some discernable pattern to the missing-ness. For a thorough treatment of that issue, please see Little and Rubin (1987). If there is no discernable pattern among the missing values and a decision to impute, or estimate, those missing values has been taken; a choice must be made among the many techniques available for imputation of missing values. Some methods for identifying, displaying, and imputing missing values have been previously discussed in this column (see: Starkweather, 2010). However, the present article will deal exclusively with the use of the `rrp.impute` function of the `rrp` package (Iacus, 2012). The *rrp* stands for Random Recursive Partitioning (Iacus & Porro, 2009 & 2007). However, the `rrp` package is only available from R-Forge for older versions of R (e.g. R version 2.15.0). Therefore, this article will provide instructions for downloading and installing R version 2.15.0, as well as the installation of the `rrp` package into R version 2.15.0. We will be using a Windows 7 PC (please note: you must have administrator privileges in order to install software). The article will then proceed to show how to use the `rrp.impute` function in order to impute multiple missing values with a simulated data set.

Installing R 2.15.0 and rrp

The first thing we need to do is determine where on your machine you want to install the old version of R; from here on we will refer to this version as R 2.15.0. Generally, RSS personnel recommend creating a specific directory (i.e. folder) on your machine's hard drive for all R installations. The file path location of such a directory should look something like:



However, we recognize some people have R installed in the default location (inside the Program Files directory); in which case, your R directory will be located inside the Program Files directory. Inside the R directory there should be at least one installation of R, typically the most recent version; which as of this writing is R 3.0.2 which can be seen in the image below.



The location shown above will be referred to as the R directory; in which we will install R 2.15.0 (and which should contain the latest version of R [e.g. R 3.0.2]). Next, we need to retrieve R 2.15.0 from the CRAN archives and download it to the R directory on our machine (the location shown in the image above). Old versions of R can be accessed from CRAN (<http://cran.us.r-project.org/>) by clicking on the R Binaries link on the left side of the main CRAN page (see image below with binaries link marked with the red rectangle).

The Comprehensive R Archive Network - Mozilla Firefox
File Edit View History Bookmarks Tools Help
The Comprehensive R Archive Network
cran.us.r-project.org
UNT--RSS My Home Page R_SC -- Jon FrontRange UNT UNT Library CRAN Gelman's Blog R-bloggers TWC Weather Now



CRAN

- [Mirrors](#)
- [What's new?](#)
- [Task Views](#)
- [Search](#)

About R

- [R Homepage](#)
- [The R Journal](#)

Software

- [R Sources](#)
- [R Binaries](#)
- [Packages](#)
- [Other](#)

Documentation

- [Manuals](#)
- [FAQs](#)
- [Contributed](#)

Download and Install R

Precompiled binary distributions of the

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, y

Source Code for all Platform

Windows and Mac users most likely wa
you do not know what this means, you 1

- The latest release (2013-09-25, F1
- Sources of [R alpha and beta relea](#)
- Daily snapshots of current patche reports.
- Source code of older versions of 1
- Contributed extension [packages](#)

Questions About R

- If you have questions about R lik

Once you click on the R Binaries link, you will then select the operating system in which you want to install (“windows” marked below with a red rectangle);

The screenshot shows the CRAN website in a Mozilla Firefox browser. The page title is "Index of /bin". On the left side, there is a navigation menu with links for "CRAN", "Mirrors", "What's new?", "Task Views", "Search", "About R", "R Homepage", "The R Journal", "Software", "R Sources", "R Binaries", and "Packages". The main content area displays a directory listing table with columns for "Name", "Last modified", and "Size Description". The "windows/" directory is highlighted with a red rectangle. Below the table, it says "Apache/2.2.22 (Debian) Server at cran.r-project.org Port 80".

<u>Name</u>	<u>Last modified</u>	<u>Size</u>	<u>Description</u>
Parent Directory		-	
linux/	23-Jan-2008 19:47	-	
macos/	19-Apr-2005 09:45	-	
macosx/	25-Sep-2013 14:13	-	
windows/	24-Feb-2012 18:41	-	

then click “base” distribution from the Subdirectories as show below;

The screenshot shows the CRAN website in a Mozilla Firefox browser. The page title is "Subdirectories:". On the left side, there is a navigation menu with links for "CRAN", "Mirrors", "What's new?", "Task Views", "Search", "About R", "R Homepage", "The R Journal", "Software", and "R Sources". The main content area displays a list of subdirectories: "base", "contrib", and "Rtools". The "base" subdirectory is highlighted with a red rectangle. Below the list, there is a note: "Please do not submit binaries to CRAN. Package developers might want to contribute to the CRAN Binaries Infrastructure (CBI). You may also want to read the R FAQ and R for Windows FAQ. Note: CRAN does some checks on these binaries for viruses, but cannot give guarantees.".

Subdirectories:

- [base](#) Binaries for base distribution (manage binaries, build, and install)
- [contrib](#) Binaries of contributed packages (manage environment and make variables)
- [Rtools](#) Tools to build R and R packages (manage environment and make variables)

Please do not submit binaries to CRAN. Package developers might want to contribute to the CRAN Binaries Infrastructure (CBI).

You may also want to read the [R FAQ](#) and [R for Windows FAQ](#).

Note: CRAN does some checks on these binaries for viruses, but cannot give guarantees.

then click “Previous releases” (marked with the red rectangle in the image below).


The Comprehensive R Archive Network - Mozilla Firefox

File Edit View History Bookmarks Tools Help

The Comprehensive R Archive Network

cran.r-project.org

UNT--RSS My Home Page R_SC -- Jon FrontRange UNT UNT Library CRAN Gelman's Blog R-bloggers Weather Now Google Maps Wiki CityData PayPal



R-3.0.2 for

[Download R 3.0.2 for Windows](#) (52 megabytes, 32/64 bit)

[Installation and other instructions](#)

[New features in this version](#)

If you want to double-check that the package you have downloaded exactly matches the package distributed by [graphical](#) and [command line versions](#) are available.

Frequently asked questions

- [How do I install R when using Windows Vista?](#)
- [How do I update packages in my previous version of R?](#)
- [Should I run 32-bit or 64-bit R?](#)

Please see the [R FAQ](#) for general information about R and the [R Windows FAQ](#) for Windows-specific informa

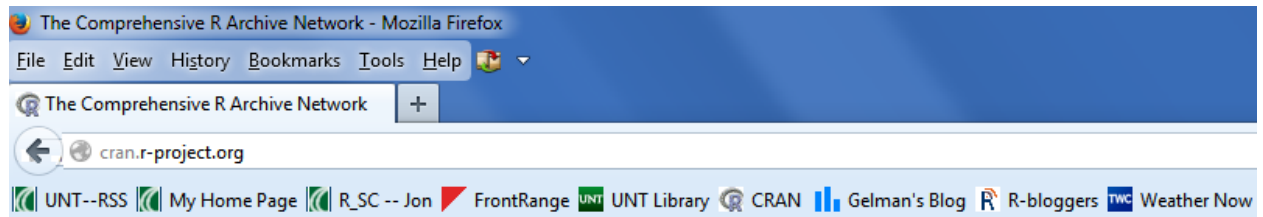
Other builds

- Patches to this release are incorporated in the [r-patched snapshot build](#).
- A build of the development version (which will eventually become the next major release of R) is availa
- [Previous releases](#)

Note to webmasters: A stable link which will redirect to the current Windows binary release is <CRAN MIRROR>/bin/windows/base/release.htm.

Last change: 2013-09-25, by Duncan Murdoch

Then click on R 2.15.0 (marked with the red rectangle in the image below).



CRAN
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

About R
[R Homepage](#)
[The R Journal](#)

Software
[R Sources](#)
[R Binaries](#)
[Packages](#)
[Other](#)

Documentation
[Manuals](#)
[FAQs](#)
[Contributed](#)

This directory contains previous binary releases of R to run on V

The current release, and links to development snapshots, are ava

In this directory:

- [R 3.0.1](#) (May, 2013)
- [R 3.0.0](#) (April, 2013)
- [R 2.15.3](#) (March, 2013)
- [R 2.15.2](#) (October, 2012)
- [R 2.15.1](#) (June, 2012)
- [R 2.15.0](#) (March, 2012)
- [R 2.14.2](#) (February, 2012)
- [R 2.14.1](#) (December, 2011)
- [R 2.14.0](#) (November, 2011)
- [R 2.13.2](#) (September, 2011)
- [R 2.13.1](#) (July, 2011)
- [R 2.13.0](#) (April, 2011)
- [R 2.12.2](#) (February, 2011)
- [R 2.12.1](#) (December, 2010)
- [R 2.12.0](#) (October, 2010)
- [R 2.11.1](#) (May, 2010)
- [R 2.11.0](#) (April, 2010)
- [R 2.10.1](#) (December, 2009)
- [R 2.10.0](#) (October, 2009)

Then click on “Download R 2.15.0 for Windows” (marked with the red rectangle in the image below). This will allow you to save the installation, or executable, file to the R directory on your machine as located and discussed above.

The Comprehensive R Archive Network - Mozilla Firefox


File Edit View History Bookmarks Tools Help

The Comprehensive R Archive Network

cran.r-project.org

UNT--RSS My Home Page R_SC -- Jon FrontRange UNT UNT Library CRAN Gelman's Blog R-bloggers TWC Weather Now Google Maps Wiki CityD

R-2



CRAN

[Mirrors](#)

[What's new?](#)

[Task Views](#)

[Search](#)

About R

[R Homepage](#)

[The R Journal](#)

Software

[R Sources](#)

[R Binaries](#)

[Packages](#)

[Download R 2.15.0 for Windows](#) 47 megabytes, 32/64 bit

[Installation and other instructions](#)

New features in this version: [Windows specific, all platforms.](#)

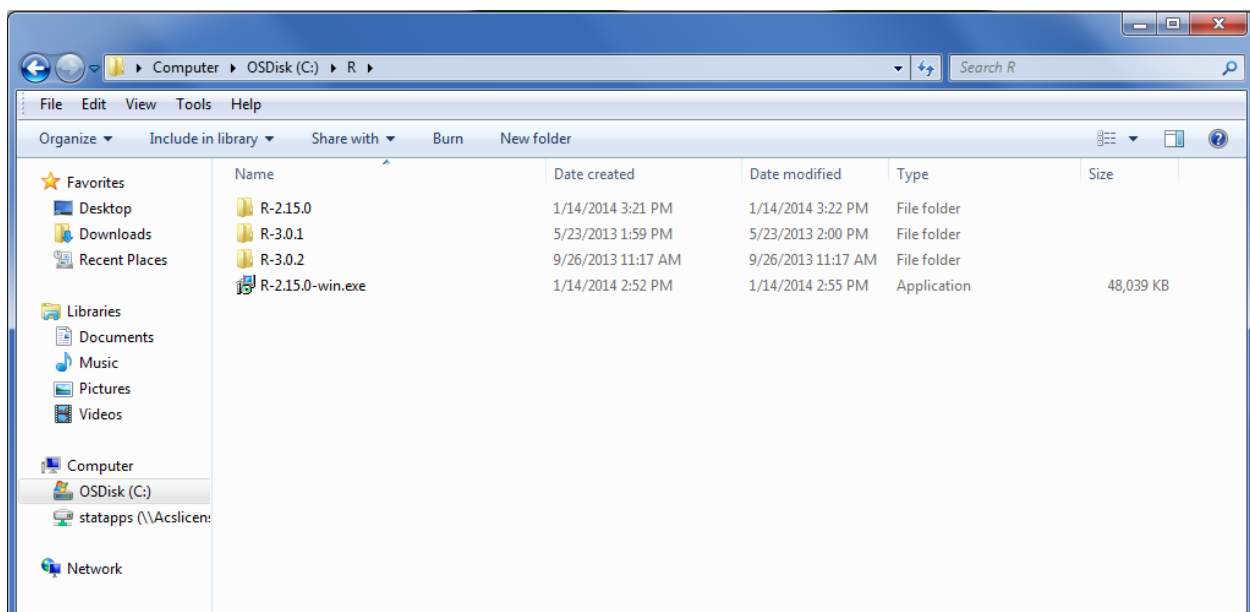
If you want to double-check that the package you have downloaded exactly matches the package [graphical](#) and [command line versions](#) are available.

Frequently asked questions

Please see the [R FAQ](#) for general information about R and the [R Windows FAQ](#) for Windows-s

Last change: 2012-03-30, by Duncan Murdoch

Then, you simply navigate to your R directory and double click the installation file to install R 2.15.0. Once you have finished installing R 2.15.0, your R directory should look something like what is below.



At this point, we can open the (R 2.15.0) console in preparation of installing the rrp package (Iacus, 2012).


```
R Console (64-bit)
File Edit Misc Packages Windows Help

R version 2.15.0 (2012-03-30)
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-pc-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

Next, we need to point our favorite browser to the rrp package page of R-Forge (https://r-forge.r-project.org/R/?group_id=1480). Once on that page (displayed below), you will need to copy the installation script line (marked below with a red rectangle) and paste it into your R 2.15.0 console in order to install the rrp package (as shown further below).

R-Forge: rrp: R Development Page - Mozilla Firefox

File Edit View History Bookmarks Tools Help

R-Forge: rrp: R Development Page

https://r-forge.r-project.org/R/?group_id=1480

UNT--RSS My Home Page R_SC -- Jon FrontRange UNT UNT Library CRAN Gelman's Blog R R-bloggers Weather Now Google Maps Wiki CityData PayPal eBay

R-Forge Search the entire project Search

Home My Page

Summary Activity Lists

R Development Page

Contributed R Packages

Below is a list of all packages provided by project **rrp**.

Important note for package binaries: R-Forge provides these binaries only for the most recent version of R, but not for older versions. In order to successful alternatively, install from the package sources (.tar.gz).

Packages	
rrp	<p>Random Recursive Partitioning</p> <p>Random Recursive Partitioning and Rank-based proximities for data matching, missing data imputation an</p> <p>Version: 2.94 Last change: 2012-10-19 17:29:02+02 Rev.: 4</p> <p>Download:  (.tar.gz)  (.zip) Build status: Current</p> <p>R install command: <code>install.packages("rrp", repos="http://R-Forge.R-project.org")</code></p> <p>Show/Hide extra info</p>

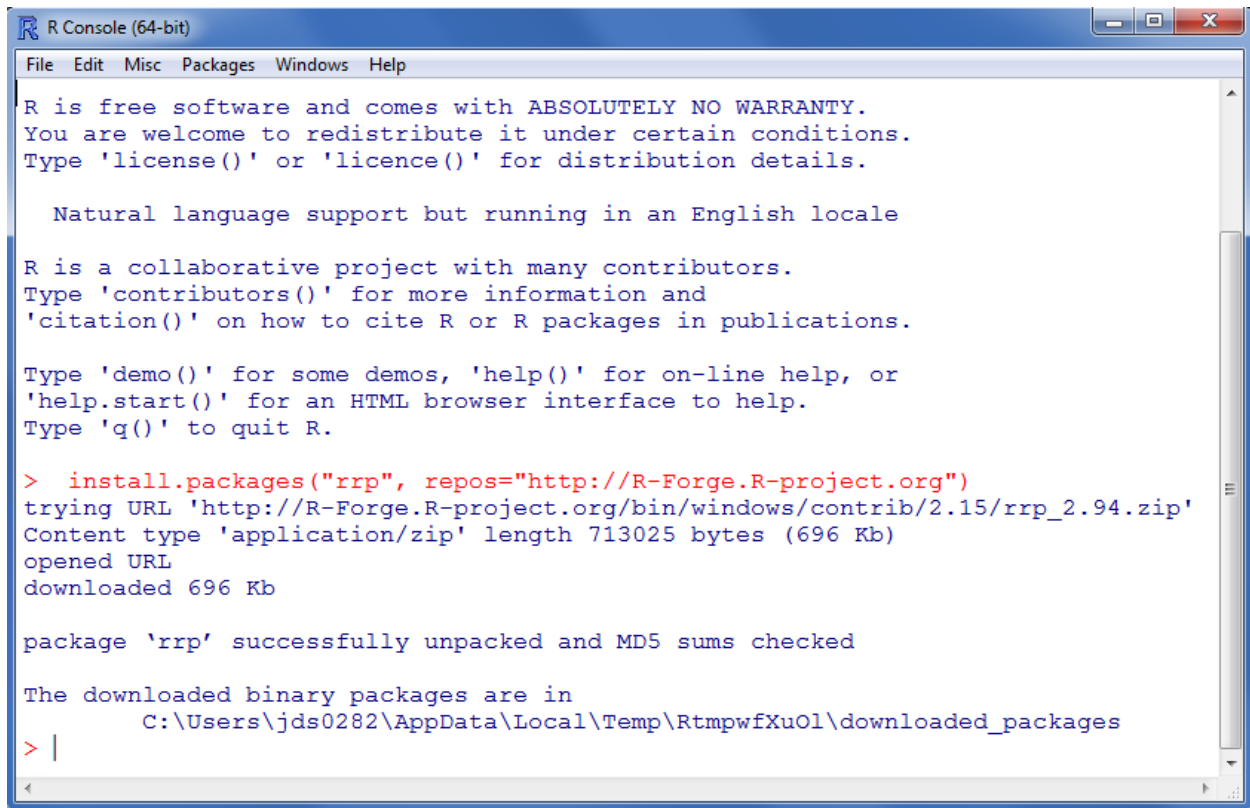
Build status codes

- 0 - Current: the package is available for download. The corresponding package passed checks on the Linux and Windows platform without ERRORS.
- 1 - Scheduled for build: the package has been recognized by the build system and provided in the staging area.
- 2 - Building: the package has been sent to the build machines. It will be built and checked using the latest patched version of R. Note that it is included in a bat
- 3 - Failed to build: the package failed to build or did not pass the checks on the Linux and/or Windows platform. It is not made available since it does not meet
- 4 - Conflicts: two or more packages of the same name exist. None of them will be built. Maintainers are asked to negotiate further actions.
- 5 - Offline: the package is not available. The build system may be offline or the package maintainer did not trigger a rebuild (done e.g., via committing to the pa

If your package is not shown on this page or not building, then check the [build system status report](#).

Thanks to:





```
R Console (64-bit)
File Edit Misc Packages Windows Help

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> install.packages("rrp", repos="http://R-Forge.R-project.org")
trying URL 'http://R-Forge.R-project.org/bin/windows/contrib/2.15/rrp_2.94.zip'
Content type 'application/zip' length 713025 bytes (696 Kb)
opened URL
downloaded 696 Kb

package 'rrp' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:\Users\jds0282\AppData\Local\Temp\RtmpwFXu01\downloaded_packages
> |
```

It is very important that you never update this version of R and that you never ‘update packages’ associated with this version of R (if you do, you’ll need to uninstall R 2.15.0 and start over). This way, you can have this old version of R for the dedicated purpose of multiple missing value imputation – and this version consume only a small amount of space on your hard drive because it should only have the rrp package installed. The latest version of R will continue to be the version you should use for all other operations.

Using rrp to impute missing values

First thing we need to do is import our simulated data (rrp.ex.data.txt) from the RSS webserver and get a summary of it. We name the data “data.1” for this example and we notice from the summary the data contains 158 cases ($n = 158$) and 8 columns: id, sex, age, Q1, Q2, Q3, Q4, Q5. We also notice from the summary there are missing values among the responses to the sex, age, Q2, and Q4 variables.

```

R Console (64-bit)
File Edit Misc Packages Windows Help

R version 2.15.0 (2012-03-30)
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-pc-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> data.1 <- read.table(
+   "http://www.unt.edu/rss/class/Jon/Benchmarks/rrp.ex.data.txt",
+   header = TRUE, sep = ",", na.strings = "NA",
+   dec = ".", strip.white = TRUE)
> summary(data.1)
      id      sex      age      Q1      Q2      Q3
Min.   : 1.00  female:67  Min.   :19.00  Min.   :1.000  Min.   :1.000  Min.   :1.000
1st Qu.: 40.25  male  :86  1st Qu.:26.00  1st Qu.:2.000  1st Qu.:2.000  1st Qu.:2.000
Median : 79.50  NA's  : 5  Median :30.00  Median :2.000  Median :2.000  Median :3.000
Mean   : 79.50                Mean   :35.15  Mean   :2.424  Mean   :2.788  Mean   :3.082
3rd Qu.:118.75                3rd Qu.:44.00  3rd Qu.:3.000  3rd Qu.:4.000  3rd Qu.:4.000
Max.   :158.00                Max.   :83.00  Max.   :7.000  Max.   :7.000  Max.   :7.000
      NA's      :8
      Q4      Q5
Min.   :1.000  Min.   :1.000
1st Qu.:2.000  1st Qu.:2.000
Median :2.000  Median :3.000
Mean   :2.658  Mean   :3.437
3rd Qu.:3.000  3rd Qu.:5.000
Max.   :7.000  Max.   :7.000
NA's   :9
> |

```

Next, we remove the (arbitrary) identification column (“id”) because it contains no meaningful information (i.e. it is not related at all to any of the other columns of data).

```

R Console (64-bit)
File Edit Misc Packages Windows Help

> data.2 <- data.1[,-1]
> summary(data.2)
      sex      age      Q1      Q2      Q3      Q4
female:67  Min.   :19.00  Min.   :1.000  Min.   :1.000  Min.   :1.000  Min.   :1.000
male  :86  1st Qu.:26.00  1st Qu.:2.000  1st Qu.:2.000  1st Qu.:2.000  1st Qu.:2.000
NA's   : 5  Median :30.00  Median :2.000  Median :2.000  Median :3.000  Median :2.000
      Mean   :35.15  Mean   :2.424  Mean   :2.788  Mean   :3.082  Mean   :2.658
      3rd Qu.:44.00  3rd Qu.:3.000  3rd Qu.:4.000  3rd Qu.:4.000  3rd Qu.:3.000
      Max.   :83.00  Max.   :7.000  Max.   :7.000  Max.   :7.000  Max.   :7.000
      NA's   : 8      NA's   : 7      NA's   : 9
      Q5
Min.   :1.000
1st Qu.:2.000
Median :3.000
Mean   :3.437
3rd Qu.:5.000
Max.   :7.000
> |

```

Next, we need to load the package (rrp) which contains the imputation function (rrp.impute). We also need to set the seed (set.seed) so that we can replicate exactly the resulting imputations we get. Notice below, we simply use the 8-digit date (at the time of writing) for the seed number (2014 Jan. 14th = 20140114). Then, we can submit this data (data.2) to the 'rrp.impute' function. Notice below, we have a "\$new.data" tacked onto the end of the function – this allows us to return just the imputed data frame (rather than the two object list which the function naturally returns). We assign the imputed data frame to a new object (data.3). You can gain a better understanding of the arguments of the 'rrp.impute' function by referring to the help files and / or package documentation (Iacus, 2009). See also Iacus, and Porro (2009); Iacus, & Porro (2007) listed at the bottom of this document.

```

R Console (64-bit)
File Edit Misc Packages Windows Help
> library(rrp)
Loading required package: rpart
Loading required package: MASS
Warning message:
package 'rrp' was built under R version 2.15.1
> set.seed(20140114)
> data.3 <- rrp.impute(data.2, k = 5, msplit = 10, Rep = 1000,
+                      cut.in = 15)$new.data
> |

```

We can see the missing (NA) have been imputed by comparing the summaries of each data frame.

```

R Console (64-bit)
File Edit Misc Packages Windows Help
> summary(data.2)
  sex      age      Q1      Q2      Q3      Q4
female:67  Min.   :19.00  Min.   :1.000  Min.   :1.000  Min.   :1.000
male   :86  1st Qu.:26.00  1st Qu.:2.000  1st Qu.:2.000  1st Qu.:2.000
NA's   : 5  Median :30.00  Median :2.000  Median :3.000  Median :2.000
      Mean :35.15  Mean  :2.424  Mean  :2.788  Mean  :3.082  Mean  :2.658
      3rd Qu.:44.00  3rd Qu.:3.000  3rd Qu.:4.000  3rd Qu.:4.000  3rd Qu.:3.000
      Max.  :83.00  Max.  :7.000  Max.  :7.000  Max.  :7.000
      NA's   : 8
      Q5
Min.   :1.000
1st Qu.:2.000
Median :3.000
Mean   :3.437
3rd Qu.:5.000
Max.   :7.000

> summary(data.3)
  sex      age      Q1      Q2      Q3      Q4
female:67  Min.   :19.00  Min.   :1.000  Min.   :1.000  Min.   :1.000
male   :91  1st Qu.:27.00  1st Qu.:2.000  1st Qu.:2.000  1st Qu.:2.000
      Median :31.00  Median :2.000  Median :3.000  Median :2.000
      Mean   :35.14  Mean   :2.815  Mean   :3.082  Mean   :2.629
      3rd Qu.:41.79  3rd Qu.:3.000  3rd Qu.:4.000  3rd Qu.:4.000  3rd Qu.:3.000
      Max.   :83.00  Max.   :7.000  Max.   :7.000  Max.   :7.000
      Q5
Min.   :1.000
1st Qu.:2.000
Median :3.000
Mean   :3.437
3rd Qu.:5.000
Max.   :7.000
> |

```

Conclusions

As you may have noticed above, we did not need to restrict the ‘rrp.impute’ function to only the numeric vectors (i.e. columns) of the data. This is one reason why RSS personnel recommend going through the (minor) trouble of having an old version of R installed on our machines. Having the old version (R 2.15.0) and the rrp package installed allows us to impute missing data quickly because ‘rrp.impute’ is the only function we are aware of which allows us to impute both numeric and categorical variables with one run of a function. The other main reason we recommend using ‘rrp.impute’ is because we have run simulations to compare the performance (in terms of bias & variability of estimated / imputed values) of ‘rrp.impute’ to a number of other highly recommended imputation strategies (e.g. maximum likelihood multiple imputation [package norm, package Amelia], Iterative Robust Model-based Imputation [package VIM], & Sequential *k* nearest neighbors [package SeqKnn, package rrcovNA]). Our results suggest the random recursive partitioning (rrp) method provides estimates with very low bias and low variability – approximately the same amounts one would get from applying the maximum likelihood method; and of the methods tested, ‘rrp.impute’ is the only one which imputes both numeric and categorical values. Keep in mind; all these methods assume the missing values are missing at random (i.e. no discernable pattern to the missing values).

For more information on what R can do, please visit the Research and Statistical Support [Do-It-Yourself Introduction to R](#) course website. An Adobe.pdf version of this article can be found [here](#).

Until next time; *make sure you get a retainer and keep your ear to the grindstone...*

References / Resources

Iacus, S. M. (2012). Package rrp. Available at: https://r-forge.r-project.org/R/?group_id=1480

Iacus, S. M. (2009). Package rrp manual. Available at: http://www.unt.edu/rss/class/Jon/R_SC/Module4/rrp.pdf

Iacus, S. M., & Porro, G. (2009). Random Recursive Partitioning: A matching method for the estimation of the average treatment effect. *Journal of Applied Econometrics*, 24, 163—185.

Iacus, S.M., & Porro, G. (2007). Missing data imputation, matching and other applications of random recursive partitioning. *Computational Statistics and Data Analysis*, 52(2), 773—789.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons.

Starkweather, J. (2010). How to identify and impute multiple missing values using R. Benchmarks Online, November 2010. Available at: <http://web3.unt.edu/benchmarks/issues/2010/11/rss-matters>