Generating Automated Dataset Summaries

Agustin Calatroni and Herman Mitchell, Rho, Inc.

Overview

Every time a dataset is created, either for data management purposes or for statistical analyses, it is imperative that each variable be reviewed carefully. A summary report of the dataset should succinctly display critical information to enhance its ease of examination and allow dissemination among interested parties. As basic as this process appears, there is no straight forward procedure in SAS to produce such a report. It is of interest to develop a SAS macro that combines the capabilities of several programs to generate an automated summary report of dataset variables.

Motivation

- Show all the data; display many numbers in small spaces
- Reveal the data at several levels of detail, from broad overview to detailed structure
- Serve as a description, tabulation and exploration of the data

Objective

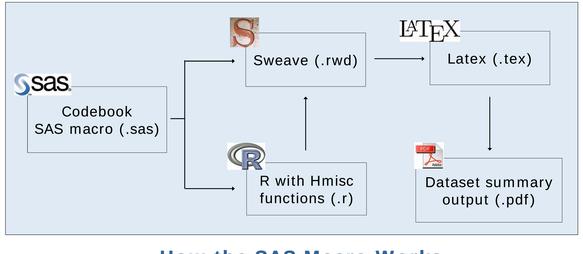
Create a concise statistical description of each variable in a dataset with the following characteristics:

- Output should be dictated by the variable type
- Output should be able to combine text and graphics (show all the data, including extreme values - hide nothing!)
- A statistical summary should be able to run in SAS using SAS datasets (SAS is the lingua franca for both pharma and CROs)
- A statistical summary should be automated so that it requires minimum input

Important Considerations

- Must include variable labels and formats
- Allow for special numeric missing values
- Display dates and times in a legible way
- Ability to subset variables and/or observations

Architecture



How the SAS Macro Works

The SAS macro works by interacting with a variety of programs to obtain the final dataset summary output. The macro creates two files:

- A Sweave file with:
  - Chunk to import the SAS dataset
  - Chunk to generate the latex descriptive statistics
- An R file with code to submit the code and generate the output

Guided Tour

We have a SAS dataset and we need to create a statistical summary for all the variables. The SAS macro insert shows all the current options in the macro. We specify the format library, pdf location and pdf name for the final output. The macro creates two files. The Sweave file contains the R code in chunks intertwined in Latex. You can distinguish the R code because it is enclosed between <<>>= and @. The second file is a set of R commands that will submit the Sweave file and create the pdf from the latex file. The SAS macro executes the set of R commands in batch mode using windows commands from within SAS (X command). From the analyst's standpoint none of these steps are visible, only the SAS macro call and the final pdf output.

A strength of the current implementation of this dataset summary generator is that the data displays depends upon the variable type. The final output will contain variable names, labels, formats, observations (n), frequency of missing values (nmiss), and number of unique observations for all variables. Special missing codes (._,.,.a,.b,...,.z) will be printed when appropriate. Then depending on the variable type: sum, mean, sd, quantile, frequency and the five lowest and highest values will be presented. When the variable is numeric and there are more than twenty unique values, output will also include a spike histogram showing all the data.

Installation Steps

Although the SAS macro is built to interact with other programming languages automatically, the installation requires some care:

1. Obtain the SAS macro codebook.sas
2. Download and install R under c:\
   - Add Hmisc packages by Frank Harrell Jr.
   - Obtain extra R functions from the author
3. Download and install Latex or Mitex if working under Windows (with add-on usepackages setspace, relsize and fancyhdr)
4. Call SAS macro codebook.sas

Summary

Enhancement

I'm currently working on two enhancements:

- The ability to run the additional applications from a server to alleviate the installation of additional programs.
- An option that will adapt the output to the case of repeated data

Pros

- Highly adaptable program, possibility to include a wide range of summary statistics , plots, etc.
- Attractive displays for data cleaning and distribution analysis
- Extremely user friendly macro once installed

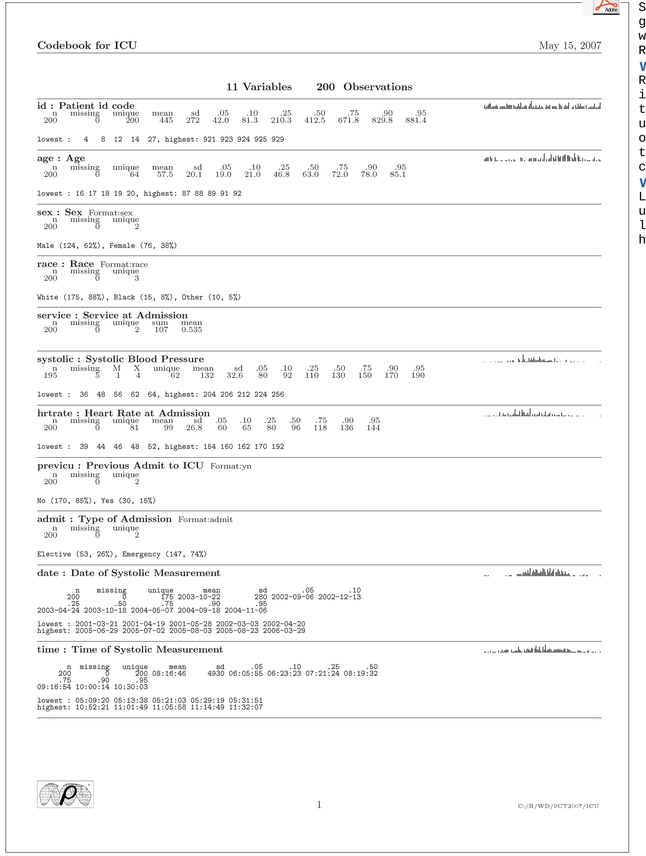
Cons

- Installation of additional programs requires extra care
- SAS is used as an interface to other programming languages
- Further customization requires complex programming in R and Latex

Conclusion

Although the current installation may be slightly intricate, the resulting output greatly simplifies data cleaning and analysis.

Final dataset summary output



Codebook for ICU — May 15, 2007

11 Variables 200 Observations

id : Patient id code

n	missing	unique	mean	sd	.05	.10	.25	.50	.75	.90	.95
200	0	200	445	272	42.0	81.3	210.3	412.5	671.8	829.8	881.4

lowest : 4 8 12 14 27, highest: 921 923 924 925 929

age : Age

n	missing	unique	mean	sd	.05	.10	.25	.50	.75	.90	.95
200	0	64	57.5	20.1	10.0	21.0	46.8	63.0	72.0	78.0	85.1

lowest : 16 17 18 19 20, highest: 87 88 89 91 92

sex : Sex Format:sex

n	missing	unique
200	0	2

Male (124, 62%), Female (76, 38%)

race : Race Format:race

n	missing	unique
200	0	3

White (175, 88%), Black (15, 8%), Other (10, 5%)

service : Service at Admission

n	missing	unique	sum	mean
200	0	2	107	0.535

systolic : Systolic Blood Pressure

n	missing	M	X	unique	mean	sd	.05	.10	.25	.50	.75	.90	.95
195	5	1	4	62	132	32.6	80	92	110	130	150	170	190

lowest : 36 48 56 62 64, highest: 204 206 212 224 256

hrtrate : Heart Rate at Admission

n	missing	unique	mean	sd	.05	.10	.25	.50	.75	.90	.95
200	0	81	99	26.8	60	65	80	96	118	136	144

lowest : 39 44 46 48 52, highest: 154 160 162 170 192

previcu : Previous Admit to ICU Format:yn

n	missing	unique
200	0	2

No (170, 85%), Yes (30, 15%)

admit : Type of Admission Format:admit

n	missing	unique
200	0	2

Elective (53, 26%), Emergency (147, 74%)

date : Date of Systolic Measurement

n	missing	unique	mean	sd	.05	.10
200	0	175	2003-10-22	280	2002-09-06	2002-12-13

.25	.50	.75	.90	.95
2003-04-24	2003-10-18	2004-05-07	2004-09-18	2004-11-06

lowest : 2001-03-21 2001-04-19 2001-05-28 2002-03-03 2002-04-20
highest: 2005-06-29 2005-07-02 2005-08-03 2005-08-23 2006-03-29

time : Time of Systolic Measurement

n	missing	unique	mean	sd	.05	.10	.25	.50
200	0	200	08:16:46	4930	06:05:55	06:23:23	07:21:24	08:19:32

.75	.90	.95
09:16:54	10:00:14	10:30:03

lowest : 05:09:20 05:13:38 05:21:03 05:29:19 05:31:51
highest: 10:52:21 11:01:49 11:05:58 11:14:49 11:32:07

1 — C:/R/WD/SCT2007/ICU

Terminology and Sample Code

What is Sweave?

Sweave is a flexible framework for mixing text and R code for automatic document generation. A single source file contains both documentation text and R code, which are then intertwined into a final document containing the combined text and R code and/or the output of the code (text, graphs).

What is R?

R is a language and environment for statistical computing and graphics. R is an implementation of the S programming language. R provides a wide variety of statistical and graphical techniques. The capabilities of R are extended through user-submitted packages, which allow specialized tasks. The Hmisc package developed by Frank Harrell Jr. is one such addition. It offers specialized functions to convert SAS dataset to R objects, create concise statistical summaries and conversion to Latex.

What is Latex?

Latex is a document preparation system for high-quality typesetting. It is often used for medium to large technical or scientific documents because of its excellence at managing equations, figures, tables and indices. As with R, Latex is highly extensible through add-ons.

SAS macro

```
%codebook(
  file=demo.ICU,              /* file location using two part name */
  formats=demo,               /* libname with location of format library */
  pdfname=ICU,                /* PDF name with final output */
  pdfloc=%str(C:/R/WD/SCT2007/ICU), /* PDF location with output */
  clean=F,                    /* Remove intermediary files when finished  */
  save=F,                     /* Save the SAS file as an R object */
  alphabetical=F,             /* alphabetic order by variable names */
  variables=Cs(),             /* Subset of variables for final output */
  specialmiss=T,              /* Should output display special missing */
  rsource=%str(C:/R/WD/SCT2007),
  sasexe=%str(C:/Program Files/SAS/SAS 9.1/sas.exe),
  rexe=%str(C:/R/R-2.4.1/bin/R.exe));
  /* Location of extra R files, sasexe and rexe commands */
```

Sweave exert

```
\documentclass{report}
\usepackage{verbatim,setspace,relsize,fancyhdr}
[…]
\begin{document}
\SweaveOpts{eps=FALSE}
<<echo=F,results=hide>>=
library(Hmisc,warn.conflicts=FALSE)
source('C:/R/WD/SCT2007/sas.get.R')
[…]
icu <- sas.get('C:/R/WD/SCT2007/ICU','icu',
               format.library='C:/R/WD/SCT2007/ICU',
               variables=Cs(),
               sasprog='"C:/Program Files/SAS/SAS 9.1/sas.exe"',
               keep.log=FALSE,clean.up=TRUE,recode=TRUE,special.miss=T)
@
<<echo=F,results=tex>>=
latex(describe(icu,digits=3,size='small'),file='')
@
\end{document}
```

R commands

```
setwd('c:/r/wd/sct2007/icu')
Sweave('icu.rnw')
system('pdflatex -quiet icu.tex')
```

SAS X commands

```
options xwait xsync;
X "C:/R/R-2.4.1/bin/R.exe --vanilla
 <c:/r/wd/sct2007/icu/icu.r
 >c:/r/wd/sct2007/icu/icu.out";
```

Agustin Calatroni
acalatro@rhoworld.com