Reproducible Research with R, \LaTeX, sweave, and knitr

Frank E Harrell Jr
Terri Scott

Department of Biostatistics
Vanderbilt University School of Medicine and
Vanderbilt Institute for Clinical & Translational Research

*useR*! 2012 Nashville 12 June 2012

Outline

1. Background
2. Scientific Methods Quality
3. Pre-Specification
4. Summary
5. Software
6. Sweave Approach
7. Enhancing Sweave Output
8. Example Enhanced Report Handout
9. knitr
Non-reproducible Research

- Misunderstanding statistics
- “Investigator” moving the target
- Lack of a blinded analytic plan
- Tweaking instrumentation / removing “outliers”
- Pre-statistician “normalization” of data and background subtraction
- Poorly studied high-dimensional feature selection

Non-reproducible Research, continued

- Programming errors
- Lack of documentation
- Failing to script multiple-step procedures
  - using spreadsheets and other interactive approaches for data manipulation
- Copying and pasting results into manuscripts
- Insufficient detail in scientific articles
- No audit trail
General Importance of Sound Methodology

- Hackam and Redelmeier [2006]: Translation of research evidence from animals to humans
- Screened articles having preventive or therapeutic intervention in in vivo animal model, > 500 citations
- 76 “positive” studies identified
- Median 14 years for potential translation
- 37 judged to have good methodological quality (flat over time)
- 28 of 76 replicated in human randomized trials; 34 remain untested
- ↑ 10% methodology score ↑ odds of replication × 1.28 (0.95 CL 0.97–1.69)
- Dose-response demonstrations: ↑ odds × 3.3 (1.1–10.1)

Note: The article misinterpreted $P$-values

BMJ 1994;308:283 (Published 29 January 1994)

Editorial

The scandal of poor medical research

D G Altman

We need less research, better research, and research done for the right reasons

What should we think about a doctor who uses the wrong treatment, either willfully or through ignorance, or who uses the right treatment wrongly (such as by giving the wrong dose of a drug)? Most people would agree that such behaviour was unprofessional, arguably unethical, and certainly unacceptable.

What, then, should we think about researchers who use the wrong techniques (either willfully or in ignorance), use the right techniques wrongly, misinterpret their results, report their results selectively, cite the literature selectively, and draw unjustified conclusions? We should be appalled. Yet numerous studies of the medical literature, in both general and specialist journals, have shown that all of the above phenomena are common. This is surely a scandal.
The Truth Wears Off

Prescribe drugs while they still work

Rhine and ESP: “the student’s extra-sensory perception ability has gone through a marked decline”

Regression to the mean

Floating definitions of X or Y: association between physical symmetry and mating behavior; acupuncture
The Truth Wears Off, continued

- Selective reporting and publication bias
- Journals seek confirming rather than conflicting data
- Damage caused by hypothesis tests and cutoffs
- Ioannidis: $\frac{1}{3}$ of articles in *Nature* never get cited, let alone replicated
- Biologic and lab variability
- Weak coupling ratio exhibited by decaying neutrons fell by 10 SDs from 1969–2001

“The decline effect is actually a decline of illusion”

---

The New York Times

April 16, 2012

A Sharp Rise in Retractions Prompts Calls for Reform

By CARL ZIMMER

In the fall of 2010, Dr. Ferric C. Fang made an unsettling discovery. Dr. Fang, who is editor in chief of the journal Infection andImmunity, found that one of his authors had doctored several papers.

It was a new experience for him. “Prior to that time,” he said in an interview, “Infection and Immunity had only retracted nine articles over a 40-year period.”

The journal wound up retracting six of the papers from the author, Naoki Mori of the University of the Ryukyus in Japan. And it soon became clear that Infection and Immunity was hardly the only victim of Dr. Mori’s misconduct. Since then, other scientific journals have retracted two dozen of his papers, according to the watchdog blog Retraction Watch.
Biomarker Discoveries

Izvestia (News) | Pravda (Truth)
---|---
Big Effects | Validated Effects

Strong Inference

16 October 1964, Volume 146, Number 3642

**Strong Inference**

Certain systematic methods of scientific thinking may produce much more rapid progress than others.

John R. Platt
Strong (Inductive) Inference, *continued*

- Devise alternative hypotheses
- Devise an experiment with alternative possible outcomes each of which will exclude a hypothesis
- Carry out the experiment
- Repeat
  - Regular, explicit use of alternative hypotheses & sharp exclusions → rapid & powerful progress
  - “Our conclusions . . . might be invalid if . . . (i) . . . (ii) . . . (iii) . . . We shall describe experiments which eliminate these alternatives.”

Platt [1964]

---

Science

*A theory which cannot be mortally endangered cannot be alive.*

> W. A. H. Rushton

*Religion is a culture of faith; science is a culture of doubt.*

> Richard Feynman

*Science is the belief in the ignorance of experts.*

> James Porter

*Fiction is about the suspension of disbelief; science is about the suspension of belief.*

> Matt Ridley

*A true scientist is bored by knowledge; it is the assault on ignorance that motivates him.*

> Matt Ridley
System Malfunctions

Reproducible Research with R, \LaTeX{}, sweave, and knitr

Background
Scientific Methods
Quality
Pre-Specification
Summary
Software
Sweave
Approach
Enhancing Sweave
Output
Enhanced sweave Report
knitr

References

System Cost of Investigating Research Malpractice

Published online 17 August 2010 | Nature | doi:10.1038/news.2010.414

High price to pay for misconduct investigations

A single investigation into research malpractice cost US$525,000.

Eugonio Samuel Reich

Investigations into research misconduct cost US institutions more than US$110 million per year, estimates a study published this week. But experts contacted by Nature question whether calculating the cost of investigation is the right way to measure the impact of research misconduct.

The research, published in PLoS Medicine, is based on the costs of a single recent case of research misconduct at the Roswell Park Cancer Institute in Buffalo, New York. In the case, a senior scientist was accused of fabricating data in at least one grant application, and an internal investigation reached a conclusion of research misconduct. As the work was partly funded by the US Department of Health and Human Services, the matter was referred to the department's Office of Research Integrity (ORI), which has yet to close the case or name the researcher involved. But the lengthy and convoluted nature of the investigation has raised questions about the validity of calculating the cost of a single investigation.
Pre-Specified Analytic Plans

- Long the norm in multi-center RCTs
- Needs to be so in all fields of research using data to draw inferences (Rubin [2007])
- Front-load planning with investigator
  - too many temptations later once see results (e.g., $P = 0.0501$)
- SAP is signed, dated, filed
- Pre-specification of reasons for exceptions, with exceptions documented (when, why, what)
- Becoming a policy in VU Biostatistics

What Do Methodologists Offer?

Biostatisticians and clinical epidemiologists play important roles in

- assessing the needed information content for a given problem complexity
- minimizing bias
- maximizing reproducibility

For more information see:

- ctspedia.org
- reproducibleresearch.net
- groups.google.com/group/reproducible-research
Some Random Thoughts

Kelvin’s curse: The unthinking and inappropriate worship of quantifiable information in medicine

Feinstein [1977]

... monetization of intellectual property appears to be a powerful force favoring methodological limitations and an excessive reductionism and fragmentation of biologic knowledge

Porta et al. [2007]

There is nothing wrong with cancer research that a little less money wouldn’t cure.

Nathan Mantel, NCI

Goals of Reproducible Analysis/Reporting

- Be able to reproduce your own results
- Allow others to reproduce your results
  
  Time turns each one of us into another person, and by making effort to communicate with strangers, we help ourselves to communicate with our future selves. (Schwab and Claerbout)

- Reproduce an entire report, manuscript, dissertation, book with a single system command when changes occur in:
  - operating system, stat software, graphics engines, source data, derived variables, analysis, interpretation

- Save time

- Provide the ultimate documentation of work done for a paper

http://biostat.mc.vanderbilt.edu/StatReport
History

- Donald Knuth found his own programming to be sub-optimal
- Reasons for programming attack not documented in code; code hard to read
- Invented **literate programming** in 1984
  - mix code with documentation in same file
  - “pretty printing” customized to each, using \TeX
  - not covered here: a new way of programming
- Knuth invented the *noweb* system for combining two types of information in one file
  - *weaving* to separate non-program code
  - *tangling* to separate program code

http://www.ctan.org/tex-archive/help/LitProg-FAQ

History, *continued*

- Leslie Lamport made \TeX\ easier to use with a comprehensive macro package LATEX in 1986
- Allows the writer to concern herself with structures of ideas, not typesetting
- LATEX is easily modifiable by users: new macros, variables, *if-then* structures, executing system commands (Perl, etc.), drawing commands, etc.
- S system created by Chambers, Becker, Wilks of Bell Labs, 1976
- R created by Ihaka and Gentleman in 1993, grew partly as a response to non-availability of S-Plus on Linux and Mac
- Friedrich Leisch developed Sweave in 2002
A Bad Alternative to Sweave

Sweave Approach

- Sweave is a function in the R tools package
- Uses noweb and an sweave style in \LaTeX
- **Insertions** are a major component
  - R printout after code chunk producing the output; plain tables
  - single pdf or postscript graphic after chunk, generates \LaTeX includegraphics command
  - direct insertion of \LaTeX code produced by R functions
  - computed values inserted outside of code chunks
- Major advantages over Microsoft Word: composition time, batch mode, easily maintained scripts, beauty
- Sweave produces self-documenting reports with nice graphics, to be given to clients
  - showing code demonstrates you are not doing “pushbutton” research

http://www.ci.tuwien.ac.at/~leisch/Sweave
Some Sweave Features

- R code set off by lines containing only `<<>>=`
- \LaTeX{} text starts with a line containing only `@`
- If the code fragment produces any graphs, the fragment is opened with `<<fig=t>>=` instead of `<<>>=`
- All other lines sent to \LaTeX{}, R code and output sent to \LaTeX{} by default but this can easily be overridden
- Including calculated variables directly in sentences, e.g. `And the final answer is \Sexpr{sqrt(9)}.` will produce “And the final answer is 3.”

Running Sweave from Command Line

```
R CMD Sweave my.Rnw produces my.tex with insertions
A useful Linux/Unix script if you use .Rnw as the suffix:

#!/bin/sh
R CMD Sweave $1.Rnw
# Add rmblines $1.tex to automatically suppress lines with #rm#
rm -f Rplots.*
```

Execute using Sweave my to run my.Rnw and produce my.tex etc., then run pdflatex my or latex my.

There are utility functions for extracting just the R output or just the \LaTeX{} text.
This lecture...

▷ Learning objectives:
  - To understand the concept & importance of reproducible research.
  - To understand the role of each software component in the automatic generation of statistical reports.
  - To understand how to generate a reproducible statistical report from scratch.

▷ Outline:
  - A common (flawed) approach for generating statistical reports.
  - A (better) alternative approach.
  - How to generate reproducible statistical reports using R, \LaTeX, & Sweave.
  - Some additional information.
Section I:

A common (flawed) approach for generating statistical reports

Typical steps leading up to the reporting

▷ FIRST,
  ■ Data entry & storage.
  ■ Data cleaning (including checking for, resolving, & correcting data entry errors).
  ■ Data preparation (including transforming/recoding variables, creating new variables, & creating necessary subsets).
  ■ Performing the proposed statistical analyses, including generating desired graphs.
  ■ Recording/saving the desired results/graphs.

▷ FINALLY,
  ■ Writing a results report, which may include documentation text, tables and/or graphs.
‘Common’ approach: write report around results

▷ First, **POINT & CLICK**
- Use Microsoft (MS) Excel for data entry/cleaning/preparation, & possibly statistical analyses.¹
- Possibly import the data into SPSS (point & click statistical software package) for data preparation & statistical analyses.
- Possibly use MS Excel to record/format the desired results & generate the desired graphs

▷ Then, **COPY & PASTE/TYPe BY HAND**
- Take advantage of pre-formatted tables & graphs generated by many statistical software packages, like SPSS.
- Copy & paste/type by hand desired results (text, tables, graphs) from data analysis system to a word processor (eg, MS Word).

¹ **BAD IDEA:** Handling of missing data; poor algorithms & unreliable results – see lecture. Okay for data entry.

Problems with ‘common’ approach

▷ **VIGNETTE 1**: You sit down to finish writing your manuscript. You realize that you need to clarify one result by running an additional analysis. You *first* re-run the primary analysis. Major problem: the primary results don’t match what you have in your paper.

▷ **VIGNETTE 2**: When you go to your project folder to run the additional analysis, you find multiple data files, multiple analysis files, & multiple results files. You can’t remember which ones are pertinent.

▷ **VIGNETTE 3**: You’ve just spent the week running your analysis & creating a results report (including tables & graphs) to present to your collaborators. You then receive an email from your PI asking you to regenerate the report based on a subset of the original data set & including an additional set of analyses – she would like it by tomorrow’s meeting.
Problems with ‘common’ approach, cont’d

- With point & click programs (eg, MS Excel or not using SPSS’s log), no way to record/save the steps performed that generated the documented results.

- Common to keep analysis code, results, & reports as separate files & to save various versions of each of these as separate files.
  - After several modifications of one or more of the files involved, becomes unclear which version of the files exactly correspond to the desired analysis & results.

- Every time analyses and/or results change, have to regenerate the results report by hand – very time consuming.

- Very easy for human error to creep into results report (eg, typing in results by hand, copying/pasting the wrong tables/graphs).

Section II:

A (better) alternative approach
Alternative to ‘common’ approach

▷ First, use R instead of Excel/SPSS for data cleaning/prep & statistical analyses (including graphs).
  - R is a programming language – removes point & click.
  - R is free to run, study, change, & improve.
  - R runs on Windows, MacOS, Linux & UNIX platforms.
  - R uses functions that are organized into packages.
    - Some packages are automatically installed when you install R, while other “contributed” (ie, add-on) packages are available to install if you need them.
  - R has publication quality graphing capabilities.
    - Able to generate typical statistical plots (eg, scatterplots, boxplots, & barplots).
    - Also allows you to create a plot ‘from scratch’ when no existing plot provides a sensible starting point.

Alternative to ‘common’ approach, cont’d

▷ Then, use LaTeX instead of MS Word for writing the report.
  - LaTeX is a document preparation system, not a word processor.
    - Rather than type words & then format them using drop-down menus, the formatting is part of the text (specified using commands).
    - Saves you time.
  - LaTeX contains features for
    - (1) automatic formatting of title pages, section headers, headers/footers, & bulleted/ enumerated lists;
    - (2) cross-referencing of sections, tables, & figures;
    - (3) typesetting of complex mathematical formulas;
    - (4) creating tables & inserting graphs; &
    - (5) automatic generation of bibliographies & indexes (eg, table of contents).
Alternative to ‘common’ approach, cont’d

▷ **A PROBLEM REMAINS:** Have removed point & click with R & have saved time spent formatting with \LaTeX, but still haven’t removed the need to copy & paste results and/or type them by hand.

▷ **BETTER APPROACH:** Embed the analysis into the report.
  
  - That is, embed the R code to clean/prep the data & to perform the desired statistical analysis into the \LaTeX document that contains the documentation text of the report.

▷ Possible using a tool called **Sweave**.
  
  - Actually, a *function* in R – part of the (base) *utils* package.
  - Utilizes a *sweave* style in \LaTeX.
  - Created by Friedrich Leisch, PhD.

Better approach: using Sweave

▷ When the ‘weaved’ document is run through Sweave all of the data analysis output (including text, tables & graphs) is created on the fly & inserted into the \LaTeX report document.
  
  - No longer need to copy & paste results and/or type them by hand.

▷ The statistical report is now completely *reproducible*.
  
  - Allows for truly *reproducible research*.

▷ Also, the report is now *dynamic*.
  
  - Can be easily regenerated when the data or analyses change – all of the results/tables/figures are automatically updated.

▷ **BONUS:** Clients are very impressed with the professional looking report.
Section III:

How to generate reproducible statistical reports using R, \texttt{\LaTeX}, & Sweave

Diagram of process

1. Generate a 'noweb' (.nw) Sweave source file in a text editor
2. Run the 'noweb' (.nw) Sweave source file through the Sweave function in R
   - Generates a \LaTeX\ (.tex) file
3. Run the Latex (.tex) file through a \LaTeX\ compiler
   - Generates a DVI (.dvi), PostScript (.ps), or PDF (.pdf) file
‘Noweb’ (.nw) Sweave source file

- ‘Noweb’: Literate-programming tool; allows you to combine program source code & corresponding documentation into single file.

- **Sweave source file**: A text file which consists of a sequence of R code & LaTeX documentation segments called chunks:
  - **LaTeX documentation chunks** start with a line that has only an @ (‘at’) sign.
  - Default for the first chunk is documentation — no @ sign needed.
  - **R code chunks** start with a line that has only <<>>=.
    - <<>>= syntax can be modified to have additional control.
  - **IMPORTANT**: Because the Sweave source file is a pre-cursor to a LaTeX document it must also include the file structure items necessary for a LaTeX document.
  - Created in any text editor (eg, Notepad) & saved to relevant project folder/directory (eg, where data files are located).

---

Simple example: example.nw

```latex
\documentclass[12pt]{article}
\usepackage[margin=1.0in]{geometry}
\title{Sweave Example}
\author{Jane Doe, MS}
\begin{document}
\maketitle
\section{Analysis & Results}
The `mtcars` (Motor Trend Car Road Tests') data set is comprised of 11 aspects of automobile design and performance (columns) for 32 automobiles (rows). We wish to know if there is a significant difference in the quarter mile track times (qsec) between the different cylinder classes (cyl; 4, 6, and 8).

```r
<<>>= data(mtcars) names(mtcars) with(mtcars, tapply(X = qsec, INDEX = list(cyl), FUN = median, na.rm = TRUE)) with(mtcars, kruskal.test(qsec ~ cyl))$p.value @
```
At the R command line prompt (‘>’), execute the Sweave function by specifying a single argument – the name of the .nw file.

- Example: `Sweave("example.nw")`
  - File name specified in quotes & must include extension (.nw).
  - IMPORTANT: The R session’s ‘working directory’ must be the folder/directory in which the .nw file is located – see R lectures.
  - Will receive screen output: Writing to file example.tex
    Processing code chunks...

- If all goes well, will receive the screen output
  - You can now run LaTeX on ‘example.tex’ & a new command line prompt.
  - .tex LaTeX file is created in same folder/directory as .nw file.
  - If error occurs, will be told which code chunk error occurred in – referenced by number (1, 2, ...; <<>>= counted).

\documentclass[12pt]{article}
\usepackage[margin=1.0in]{geometry}
\title{Sweave Example}
\author{Jane Doe, MS}
\usepackage{.../Sweave}
\begin{document}
\maketitle
\section{Analysis & Results}
The \texttt{mtcars} (`Motor Trend Car Road Tests') data set is comprised of 11 aspects of automobile design and performance (columns) for 32 automobiles (rows). We wish to know if there is a significant difference in the quarter mile track times (\texttt{qsec}) between the different cylinder classes (\texttt{cyl}; 4, 6, and 8).

\begin{Schunk}
\begin{Sinput}
> data(mtcars)
> names(mtcars)
\end{Sinput}
\begin{Soutput}
[1] "mpg" "cyl" "disp" "hp" "drat" "wt" "qsec" "vs" "am" "gear" "carb"
\end{Soutput}
\begin{Sinput}
> with(mtcars, tapply(X = qsec, INDEX = list(cyl), FUN = median, na.rm = TRUE))
\end{Sinput}
\begin{Soutput}
4     6     8
18.900 18.300 17.175
\end{Soutput}
\begin{Sinput}
> with(mtcars, kruskal.test(qsec ~ cyl))$p.value
\end{Sinput}
\begin{Soutput}
[1] 0.006234986
\end{Soutput}
\end{Schunk}
\end{document}
Compiling the .tex file

▷ On a Linux/Unix machine:
  - Open a Terminal shell & using the cd command, move to the relevant working directory (where the .nw/.tex files are saved).
  - To create a PDF (.pdf) file, execute the pdflatex command at the command prompt – eg, pdflatex example
    - Do not need to specify .tex extension.
    - .pdf file created in same directory as the .nw/.tex files.
  - Often necessary to compile the .tex file twice – use &&
    - Example: latex example && latex example
  - If all goes well, will be returned to a new command prompt.
  - Other options: R’s system() function or a shell script (see Sweave manual FAQ A.3).

2 Use the latex command to create a DVI (.dvi) file.
3 For elements like a table of contents & cross-referencing (ie, section, table, & figure labeling)

Compiling the .tex file, cont’d

▷ On a Windows/Mac machine:
  - Use MikTeX (free @ http://miktex.org).
    - May have problem referencing Sweave style (.sty) file because of the space in the ‘Program Files’ folder name – see Sweave manual (FAQ A.12) for solution.
  - Can also use a text editor like WinEdt, which by default is already configured for MikTeX – point & click capabilities.
    - Free @ http://www.winedt.com/; MikTeX must be installed.
  - If no WinEdt:
    - Open a Terminal shell by clicking on ‘Run’ from the ‘Start’ menu & typing ‘C:/command’ (or ‘cmd’).
    - Using the cd command, move to the relevant working directory.
    - Use commands similar to latex and pdflatex.4

4 Usually not necessary to compile the .tex file twice – MikTeX compiles as many times as necessary.
Sweave Example

Jane Doe, MS
May 6, 2008

1 Analysis & Results

The \textit{mtcars} ("Motor Trend Car Road Tests") data set is comprised of 11 aspects of automobile design and performance (columns) for 32 automobiles (rows). We wish to know if there is a significant difference in the quarter mile track times (\textit{qsec}) between the different cylinder classes (\textit{cyl}; 4, 6, and 8).

\begin{verbatim}
> data(mtcars)
> names(mtcars)
[1] "mpg" "cyl" "disp" "hp" "drat" "wt" "qsec" "vs" "am" "gear"
[11] "carb"
> with(mtcars, tapply(X = qsec, INDEX = list(cyl), FUN = median, + na.rm = TRUE))
 4     6     8
18.900 18.300 17.175
> with(mtcars, kruskal.test(qsec ~ cyl))$p.value
[1] 0.006234986
\end{verbatim}

Modifying R code chunk output – \texttt{<<>>=} options

- Named ‘flags’ (separated by commas) can be specified within the \texttt{<<>>=} R code chunk header to pass options to Sweave, which control the final output.

  - \texttt{echo} flag: value indicating whether to include (\texttt{true}) or not include (\texttt{false}) the R \textit{code} (commands) in the output file.
  - \texttt{results} flag: value indicating whether to include (\texttt{verbatim}) or not include (\texttt{hide}) the results of the R code (ie, what is normally printed to the screen) in the output file.

  - When just \texttt{<<>>=} is specified, Sweave implements the \textit{default} values of the \texttt{echo} & \texttt{results} flags – as we saw, both the R code & its results are included in the output file.

  - \texttt{<<>>=} is equivalent to \texttt{<<echo = true, results = verbatim>>=}.

  - Often use \texttt{<<echo = false, results = hide>>=} for R code chunks that contain data input, cleaning, & preparation steps.
Can generate tables using a results = tex flag.

- R code chunk contains the code that generates the \LaTeX{} syntax to create a table.
  - \LaTeX{} syntax is inserted in the .tex file; the table is created when the .tex file is compiled.
  - \LaTeX{} syntax generating functions available from the Hmisc & xtable add-on packages\(^5\) – \texttt{latex()} & \texttt{xtable()}/\texttt{print.xtable()} functions, respectively.
    - Contain arguments to specify formatting of the table, table caption (for 'List of Tables'), & cross-referencing.
  - \texttt{\usepackage{\{} statements (additional \LaTeX{} file structure items) often needed – see additional example posted on website.

\(^5\) Must be installed & loaded – see R lectures

Can insert generated graphs using a fig = true flag.

- R code chunk contains the code that generates the graph.
  - IMPORTANT: R code must generate only one figure.
- An EPS & PDF file of the graph are created & saved (by default) to same folder/directory as .nw file.
  - Can be saved in a sub-folder/directory – see Sweave manual.
- An \texttt{\includegraphics{\}} statement is inserted in the .tex file, which inserts the saved file when the .tex file is compiled.
- By default, no caption is given to inserted graph – causes graph not to be listed in ‘List of Figures’.
  - Solution: Wrap R code chunk with \texttt{fig = true} flag with \texttt{\begin{figure}} & \texttt{\end{figure}} environment & a corresponding \texttt{\caption{}} statement.
Embedding R code in a \LaTeX{} sentence

▶ Often wish to incorporate a value calculated using R into a \LaTeX{} documentation sentence.

- Can do this using \texttt{\textbackslash Sexpr\{expr\}}, where \texttt{expr} is R code.
- Example: 'The mean quarter mile track time of the $N =$ \texttt{\textbackslash Sexpr\{nrow(mtcars)\}} cars included in the \texttt{mtcars} data set was \texttt{\textbackslash Sexpr\{round(mean(mtcars$qsec$, na.rm = TRUE), 1)\}} seconds.'
  
  evaluates to 'The mean quarter mile track time of the $N = 32$ cars included in the \texttt{mtcars} data set was 17.8 seconds.'

▶ The \texttt{\textbackslash Sexpr\{} cannot break over many lines & must not contain curly brackets (\{} \}).

- More complicated/lengthy expressions can be easily executed & assigned as an object in a \textit{hidden} code chunk & then the assigned object referenced inside the \texttt{\textbackslash Sexpr\{}.

Section IV:

Some additional information
What to do...

▷ When you get an error in the Sweave step: check R code chunks.
  - Recall, will be told in which code chunk the error occurred.
  - Check to make sure every R code chunk begins with a \texttt{<<\=} (with possible flags) & ends with an \texttt{@} sign.\textsuperscript{6}

▷ When you get an error in the \LaTeX{} compile step: check \LaTeX{} documentation chunks & .tex file.
  - Error could be caused by output inserted in the .tex file via a \texttt{\textbackslash Sexpr{}} expression or a \texttt{results = tex} flag.
  - Comment out \LaTeX{} documentation chunks and/or whole R code chunks (from \texttt{<<\=} to \texttt{@}) in .nw file using % signs.

▷ Whenever .nw file or data file changes: re-run Sweave step on the (modified & saved) .nw file & re-compile resulting .tex file.

\textsuperscript{6} Even though \texttt{@} sign is technically a header for a \LaTeX{} documentation chunk, think of it as a footer for an R code chunk.

Useful tips/recommendations

▷ Work out details of R code within an R session & then copy & paste correct code to an R code chunk within the .nw file.

▷ On a Windows machine, show all file extensions – uncheck the ‘Folder Option’ to ‘Hide file extensions for known file types’.

▷ On a Linux/Unix machine, use Kate or ESS (Emacs Speaks Statistics) as your text editor; on a Windows machine, use WinEdt.

▷ When you start a new R session,
  - (1) Use the \texttt{Stangle()} function to extract all of the R code chunks from the .nw file & write them to a .R code file.
  - (2) Use the \texttt{source()} function to read in the .R code file & execute the R code chunks.
  - Allows you to quickly execute all the R code chunks without having to copy/paste from .nw file to the R command prompt.
Another literate programming option in R

▷ The **brew** function:
  - Part of the (add-on) *brew* package.
  - Allows you to embed R code in HTML (and other text) documents.
    - If embedded in an HTML document, generates an HTML file.
  - Only a one-step process – no other software, like \LaTeX{}, needed.
  - R code chunks start with a line that has only `<%` and end with a line that has only `%>` – does not show R code in the output file.
  - Can embed R code in a sentence using `<% expr %>`, where *expr* is R code.
  - At the R command line prompt, executed by specifying a single argument – the name of the (text) file.
    - Example: `brew("brew_report.html")`

Resources & references

▷ Today’s material:
  - [http://biostat.mc.vanderbilt.edu/SweaveLatex](http://biostat.mc.vanderbilt.edu/SweaveLatex)
  - Includes *extended* Sweave example, Sweave and \LaTeX{} links, and some of Frank Harrell’s material for enhancing your report output.

▷ Additional reproducible research options in R:
  - The “Reproducible Research” CRAN Task View webpage.
    - [http://www.cran.r-project.org](http://www.cran.r-project.org) – “CRAN Task View” link on the “Packages” page.
Enhancing Output

- Graphics size and quality suitable for publication using SweaveHooks
- Customizing the \LaTeX\ Sweave.sty style macro
- Pretty printing of code and output, with shaded boxes
- Direct insertion of \LaTeX\ code created by R functions
  - Allows complex tables with micrographics
- Selectively suppressing parts of R output using Hmisc prselect function
- Comments in R code containing symbolic references to \LaTeX\ sections
- Auto-documenting R and package versions used
- Floating figures & captions: see bottom of template wiki below

See also http://biostat.mc.vanderbilt.edu/SweaveTemplate or the R SweaveListingUtils package

Sweavel.sty

- Uses listings and relsize \LaTeX\ packages with differently shaded boxes for R code and its output
- Save http://biostat.mc.vanderbilt.edu/wiki/pub/Main/SweaveTemplate/Sweavel.sty into Sweavel.sty where your \LaTeX\ installation can find it
- Comments inside Sweavel.sty or in the online template show how to change colors, darkness of gray scale, font sizes
- Add to \LaTeX\ preamble to preserve comments, use only pdf (in graphics dir.), set default graphic size:
  \usepackage{Sweavel}
  \SweaveOpts{keep.source=TRUE}
  \SweaveOpts{prefix.string=graphics/plot, eps = FALSE, pdf = TRUE}
  \SweaveOpts{width=5, height=3.5}
Example Enhanced Report

Frank E Harrell Jr
Department of Biostatistics
Vanderbilt University School of Medicine
January 23, 2012

1 Descriptive Statistics

```r
require(rms)  # Get access to rms and Hmisc packages
getHdata(support)  # Use Hmisc/getHdata to get dataset from VU DataSets wiki
d <- subset(support, select=c(age, sex, race, edu, income, hospdead, slos, dzgroup, meanbp, hrt))
latex(describe(d), file='')
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>N</th>
<th>Missing</th>
<th>Unique</th>
<th>Mean</th>
<th>.05</th>
<th>.10</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
<th>.90</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>Age</td>
<td>1000</td>
<td>0</td>
<td>970</td>
<td>62.47</td>
<td>33.76</td>
<td>38.91</td>
<td>51.81</td>
<td>64.90</td>
<td>74.50</td>
<td>81.87</td>
<td>86.00</td>
</tr>
<tr>
<td>sex</td>
<td></td>
<td>1000</td>
<td>0</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>female</td>
<td>1000</td>
<td>0</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>1000</td>
<td>0</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>race</td>
<td></td>
<td>995</td>
<td>5</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>white</td>
<td>781</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>black</td>
<td>157</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>asian</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>other</td>
<td>12</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>hispanic</td>
<td>36</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>edu</td>
<td>Years of Education</td>
<td>798</td>
<td>25</td>
<td>11.78</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>14</td>
<td>16</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>income</td>
<td></td>
<td>651</td>
<td>349</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>under $11k</td>
<td>309</td>
<td>0</td>
<td>47</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$11-$25k</td>
<td>161</td>
<td>0</td>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$25-$50k</td>
<td>106</td>
<td>0</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$50k+</td>
<td>75</td>
<td>0</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hospdead</td>
<td>Death in Hospital</td>
<td>1000</td>
<td>0</td>
<td>253</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>slos</td>
<td>Days from Study Entry to Discharge</td>
<td>1000</td>
<td>0</td>
<td>88</td>
<td>17.86</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>11</td>
<td>20</td>
<td>37</td>
<td>53</td>
</tr>
</tbody>
</table>

lowest: 18.04 18.41 19.76 20.30 20.31
highest: 95.51 96.02 96.71 100.13 101.85

lowest: 3 4 5 6 7, highest: 145 164 202 236 241
**DESCRIPTIVE STATISTICS**

### dzgroup

<table>
<thead>
<tr>
<th>n missing unique</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 0 8</td>
</tr>
</tbody>
</table>

**ARF/MOSF w/Sepsis COPD CHF Cirrhosis Cona Colon Cancer Lung Cancer**

<table>
<thead>
<tr>
<th>Frequency</th>
<th>ARF/MOSF w/Sepsis COPD CHF Cirrhosis Cona Colon Cancer Lung Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>391</td>
<td>116 143 55 60 49 100</td>
</tr>
<tr>
<td>%</td>
<td>39 12 14 6 6 5 10</td>
</tr>
</tbody>
</table>

**MOSF w/Malig**

<table>
<thead>
<tr>
<th>Frequency</th>
<th>MOSF w/Malig</th>
</tr>
</thead>
<tbody>
<tr>
<td>86</td>
<td></td>
</tr>
<tr>
<td>%</td>
<td>9</td>
</tr>
</tbody>
</table>

#### meanbp : Mean Arterial Blood Pressure Day 3

<table>
<thead>
<tr>
<th>n missing unique Mean .05 .10 .25 .50 .75 .90 .95</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 0 122 84.98 47.00 64.75 78.00 107.00 120.00 128.05</td>
</tr>
</tbody>
</table>

**lowest :** 0 20 27 30 32, **highest:** 155 158 161 162 180

#### hrt : Heart Rate Day 3

<table>
<thead>
<tr>
<th>n missing unique Mean .05 .10 .25 .50 .75 .90 .95</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 0 124 97.87 54.0 60.0 72.0 100.0 120.0 135.0 146.1</td>
</tr>
</tbody>
</table>

**lowest :** 0 11 30 35 36, **highest:** 189 193 199 232 300

Race is reduced to three levels (white, black, OTHER) because of low frequencies in other levels (minimum relative frequency set to 0.05).

```
d <- transform(d, race = combine.levels(race, minlev = 0.05))
```

Summaries of variables stratified by sex are below.

```
látex(summary(sex ~ ., method='reverse', data=d, test=TRUE), npct='both', dotchart=TRUE, file='', landscape=TRUE, round=1)
```
Table 1: Descriptive Statistics by sex

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>female</th>
<th>male</th>
<th>Test Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td>1000</td>
<td>N = 438</td>
<td>1000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>51.5 64.9 75.9</td>
<td>52.1 64.9 72.7</td>
<td></td>
</tr>
<tr>
<td>race</td>
<td>995</td>
<td>N = 562</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OTHER</td>
<td></td>
<td>6% 44%</td>
<td>5% 49%</td>
<td></td>
</tr>
<tr>
<td>white</td>
<td></td>
<td>76% 34%</td>
<td>81% 29%</td>
<td></td>
</tr>
<tr>
<td>black</td>
<td></td>
<td>18% 82%</td>
<td>14% 86%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>798</td>
<td>10 12 14</td>
<td>9 12 14</td>
<td></td>
</tr>
<tr>
<td>Years of Education</td>
<td>651</td>
<td>N = 562</td>
<td>N = 438</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>12 14 16</td>
<td>12 14 16</td>
<td></td>
</tr>
<tr>
<td>income</td>
<td></td>
<td>under $11k</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>54% 37%</td>
<td>42% 34%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$11-$25k</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>21% 28%</td>
<td>16% 16%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$25-$50k</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>16% 16%</td>
<td>9% 14%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;$50k</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>9% 14%</td>
<td>9% 14%</td>
<td></td>
</tr>
<tr>
<td>Death in Hospital</td>
<td>1000</td>
<td>25% 26%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Days from Study Entry to Discharge</td>
<td>1000</td>
<td>7 12 21</td>
<td>6 10 19</td>
<td></td>
</tr>
<tr>
<td>dzgroup</td>
<td>1000</td>
<td>N = 562</td>
<td>N = 438</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ARF/MOSF w/Sepsis</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>41% 33%</td>
<td>37% 32%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>COPD</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>14% 10%</td>
<td>16% 10%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CHF</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>11% 17%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cirrhosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5% 6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Coma</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6% 6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Colon Cancer</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5% 5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lung Cancer</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>9% 11%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>MOSF w/Malig</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10% 8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Arterial Blood Pressure Day 3</td>
<td>1000</td>
<td>64 77 107</td>
<td>65 79 107</td>
<td></td>
</tr>
<tr>
<td>Heart Rate Day 3</td>
<td>1000</td>
<td>74 105 122</td>
<td>71 100 118</td>
<td></td>
</tr>
</tbody>
</table>

\[ a \ b \ c \] represent the lower quartile \( a \), the median \( b \), and the upper quartile \( c \) for continuous variables.

\( N \) is the number of non–missing values.

Tests used:

\(^1\)Wilcoxon test; \(^2\)Pearson test
2 Redundancy Analysis and Variable Interrelationships

```
v ← varclus(~., data=d)
plot(v)
redun(~ age + sex + race + edu + income + dzgroup + meanbp + hrt, data=d)
```

Redundancy Analysis

```
redun(formula = ~age + sex + race + edu + income + dzgroup +
       meanbp + hrt, data = d)
```

n: 617 p: 8 nk: 3

Number of NAs: 383

Frequencies of Missing Values Due To Each Variable

<table>
<thead>
<tr>
<th></th>
<th>age</th>
<th>sex</th>
<th>race</th>
<th>edu</th>
<th>income</th>
<th>dzgroup</th>
<th>meanbp</th>
<th>hrt</th>
</tr>
</thead>
<tbody>
<tr>
<td>counts</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>202</td>
<td>349</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Transformation of target variables forced to be linear

$R^2$ cutoff: 0.9 Type: ordinary

$R^2$ with which each variable can be predicted from all other variables:

<table>
<thead>
<tr>
<th></th>
<th>age</th>
<th>sex</th>
<th>race</th>
<th>edu</th>
<th>income</th>
<th>dzgroup</th>
<th>meanbp</th>
<th>hrt</th>
</tr>
</thead>
<tbody>
<tr>
<td>values</td>
<td>0.196</td>
<td>0.088</td>
<td>0.120</td>
<td>0.284</td>
<td>0.339</td>
<td>0.253</td>
<td>0.067</td>
<td>0.163</td>
</tr>
</tbody>
</table>

No redundant variables

```
# Alternative: redun(~., data=subset(d, select=-c(hospdead, slos)))
```

Note that the clustering of black with white is not interesting; this just means that these are mutually exclusive higher frequency categories, causing them to be negatively correlated.

3 Logistic Regression Model

Here we fit a tentative binary logistic regression model. The coefficients are not very useful so they are not printed (\ldots is printed in their place).
```
log <- datadist(d); options(datadist='log')
f <- lrm(hospdead ~ rcs(age,4) + sex + race + dzgroup + rcs(meanbp,5), data=d) # see Section 1 for descriptive statistics
f
```

**Logistic Regression Model**

```r
lrm(formula = hospdead ~ rcs(age, 4) + sex + race + dzgroup + rcs(meanbp, 5), data = d)
```

**Frequencies of Missing Values Due to Each Variable**

<table>
<thead>
<tr>
<th></th>
<th>hospdead</th>
<th>age</th>
<th>sex</th>
<th>race</th>
<th>dzgroup</th>
<th>meanbp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs</td>
<td>995</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>744</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>251</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>max</td>
<td>deriv</td>
<td>1e−09</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs</th>
<th>995</th>
<th>LR χ² 245.83</th>
<th>R² 0.323</th>
<th>C 0.800</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>744</td>
<td>d.f. 17</td>
<td>g 1.605</td>
<td>Dxy 0.601</td>
</tr>
<tr>
<td>1</td>
<td>251</td>
<td>Pr(&gt; χ²) &lt; 0.0001</td>
<td>gr 4.980</td>
<td>gamma 0.602</td>
</tr>
<tr>
<td>max</td>
<td></td>
<td></td>
<td>gp 0.228</td>
<td>τa-a 0.227</td>
</tr>
<tr>
<td></td>
<td></td>
<td>max</td>
<td>deriv</td>
<td>1e−09</td>
</tr>
</tbody>
</table>

Better: Output model statistics \LaTeX\ markup, automatically suppressing coefficients.

```
print(f, latex=TRUE, coefs=FALSE)
```

**Logistic Regression Model**

```r
lrm(formula = hospdead ~ rcs(age, 4) + sex + race + dzgroup + rcs(meanbp, 5), data = d)
```

**Frequencies of Missing Values Due to Each Variable**

<table>
<thead>
<tr>
<th></th>
<th>hospdead</th>
<th>age</th>
<th>sex</th>
<th>race</th>
<th>dzgroup</th>
<th>meanbp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs</td>
<td>995</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>744</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>251</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>max</td>
<td>deriv</td>
<td>1e−09</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The mean arterial blood pressure effect is shown below, on the probability scale. **Note:** Here we use the figure environment, with a caption. The \texttt{rm}lines shell script is run to remove lines containing \texttt{rm} surrounded by sharp signs.

```
# Lattice graphics require print() to render
p <- Predict(f, meanbp, fun=pllogis)
print(plot(p, ylab='Prob[hospital death]', adj.subtitle=FALSE))
# Figure 1
```

The likelihood ratio \(\chi^2\) statistic is 245.83 on 17 d.f. The fitted model in algebraic form is found below.

```
lm(f, file='')
```

\[
\Pr\{\text{hospdead} = 1\} = \frac{1}{1 + \exp(-X\beta)}, \quad \text{where}
\]
Figure 1: Partial effect of mean arterial blood pressure adjusted to age=64.9 sex=male race=white dzgroup=ARF/MOSF w/Sepsis.

Table 2: Wald Statistics for hospdead

<table>
<thead>
<tr>
<th></th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>7.12</td>
<td>3</td>
<td>0.0683</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>2.91</td>
<td>2</td>
<td>0.2338</td>
</tr>
<tr>
<td>sex</td>
<td>2.16</td>
<td>1</td>
<td>0.1413</td>
</tr>
<tr>
<td>race</td>
<td>1.38</td>
<td>2</td>
<td>0.5005</td>
</tr>
<tr>
<td>dzgroup</td>
<td>78.77</td>
<td>7</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>meanbp</td>
<td>65.62</td>
<td>4</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>48.11</td>
<td>3</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>TOTAL NONLINEAR</td>
<td>50.15</td>
<td>5</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>TOTAL</td>
<td>151.71</td>
<td>17</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>

$X\hat{\beta} =$

\[ 6.246868 - 0.01527011\text{age} + 1.926558 \times 10^{-5}(\text{age} - 33.76177)^3_+ \\
- 7.948748 \times 10^{-5}(\text{age} - 58.26838)^3_+ + 7.531077 \times 10^{-5}(\text{age} - 70.09373)^3_+ \\
- 1.508887 \times 10^{-5}(\text{age} - 86.00023)^3_+ \\
+ 0.2538355\text{male} \\
- 0.4126359\text{white} - 0.3369259\text{black} \\
- 0.9740300\text{COPD} - 2.3997310\text{CHF} + 0.3506404\text{Cirrhosis} + 1.4043122\text{Coma} \\
- 1.7956574\text{Colon Cancer} - 0.4113406\text{Lung Cancer} + 0.7656912\text{MOSF w/Malig} \\
- 0.1063267\text{meanbp} + 3.831943 \times 10^{-5}(\text{meanbp} - 47)^3_+ \\
- 5.483953 \times 10^{-5}(\text{meanbp} - 65.725)^3_+ - 3.595399 \times 10^{-6}(\text{meanbp} - 78)^3_+ \\
+ 2.231445 \times 10^{-5}(\text{meanbp} - 106)^3_+ - 2.198948 \times 10^{-6}(\text{meanbp} - 128.05)^3_+ \\
\]

and $\{c\} = 1$ if subject is in group $c$, 0 otherwise; $(x)_+ = x$ if $x > 0$, 0 otherwise.
4 Computing Environment

These analyses were done using the following versions of $\mathbb{R}^1$, the operating system, and add-on packages $\text{Hmisc}^2$, $\text{rms}^3$, and others:

- R version 2.14.1 (2011-12-22), x86_64-pc-linux-gnu
- Base packages: base, datasets, graphics, grDevices, grid, methods, splines, stats, utils
- Other packages: Hmisc 3.9-1, lattice 0.20-0, rms 3.3-2, survival 2.36-9
- Loaded via a namespace (and not attached): cluster 1.14.1, tools 2.14.1

References


5 Source Code for This Report

\verb|\documentclass{article}|
\verb|\usepackage{relsize,setspace} % used by latex(describe( ))|
\verb|\usepackage[url] % used in bibliography|
\verb|\usepackage[superscript,nomove]{cite} % use if \cite is used and superscripts wanted|
\verb|\usepackage{lscap} % for landscape mode tables|
\verb|\usepackage[calc,epic,color]{moreverb} % used for latex(..., dotchart=TRUE)|
\verb|\usepackage{fancyhdr} % this and next line are for fancy headers/footers|
\verb|\pagestyle{fancy} % define the following only if you put figures in a figure environment|
\verb|\newcommand{\bc}{\begin{center}} % abbreviate|
\verb|\newcommand{\ec}{\end{center}}|
\verb|\newcommand{\code}[3]{\begin{verbatim}|\verb|\leavevmode|\verb|\begin{verbatim}|
\verb|\end{verbatim}|\verb|\end{verbatim}|\verb|\leavevmode|\verb|\end{verbatim}}|
\verb|\newcommand{\R}{{\normalfont\textsf{R}}}|
\verb|\fg{basefilename}{label}{caption}|
\verb|\fg[htbp]{basefilename}{label}{caption}|
\caption{\smaller #3}\label{#2}\end{figure}}

\usepackage{Sweavel}
% Uncomment some of the following to use some alternatives:
% \def\Sweavesize{\normalsize} (changes size of typeset R code and output)
% \def\Rcolor{\color{black}}
% \def\Routcolor{\color{green}}
% \def\Rcommentcolor{\color{red}}
% To change background color or R code and/or output, use e.g.:
% \def\RbackgroundColor{\color{white}}
% \def\RoutbackgroundColor{\color{white}}
% To use rgb specifications use \color[rgb]{ , , }
% To use gray scale use e.g. \color[gray]{0.5}
% If you change any of these after the first chunk is produced, the
% changes will have effect only for the next chunk.

\SweaveOpts{keep.source=TRUE}
% To produce both postscript and pdf graphics, remove the eps and pdf
% parameters in the next line. Set default plot size to 5 x 3.5 in.
\SweaveOpts{prefix.string=graphics/plot, eps = FALSE, pdf = TRUE}
\SweaveOpts{width=5, height=3.5}
% To omit code and its output throughout, add \SweaveOpts{echo=F, results=hide}

\title{Example Enhanced Report}
\author{Frank E Harrell Jr\smaller Department of Biostatistics\smaller Vanderbilt University School of Medicine}
\begin{document}
\maketitle
% Use the following 3 lines for long reports needing navigation
%\tableofcontents
%\listoftables
%\listoffigures %not used unless figure environments used
<<echo=F>>=
# For more publication-ready graphics
spar <- function(mar=c(3.25+bot-.45*multi,3.5+left,.5+top+.25*multi,.5+rt),
  lwd = if(multi)1 else 1.75,
  mgp = if(multi) c(1.5, .365, 0) else c(2.4-.4, 0.475, 0),
  tcl = if(multi)-0.25 else -0.4,
  bot=0, left=0, top=0, rt=0, ps=14,
  mfrow=NULL, ...)
{
  multi <- length(mfrow) > 0
  par(mar=mar, lwd=lwd, mgp=mgp, tcl=tcl, ...)
  if(multi) par(mfrow=mfrow)
}
options(SweaveHooks=list(fig=spar)) # run spar() before every plot
options(prompt=' ',continue=' ') # remove prompt characters at start of lines

# Include the following only if taken control of figures (e.g, figure env.)
ppdf <- function(file, w=4.5, h=3, ...) # set your own default height and width
{
  pdf(paste('graphics/', substitute(file),'.pdf',sep=''), width=w, height=h)
  spar(...)
}
doff <- function() invisible(dev.off()) # invisible to prevent R output
@
\section{Descriptive Statistics}\label{descStats}
\<<results=tex>>=
require(rms) # Get access to rms and Hmisc packages
getHdata(support) # Use Hmisc/getHdata to get dataset from VU DataSets wiki
d <- subset(support, select=c(age,sex,race,edu,income,hospdead,slos,dzgroup,
meanbp,hrt))
latex(describe(d), file='')
@
Race is reduced to three levels (white, black, OTHER) because of low
frequencies in other levels (minimum relative frequency set to 0.05).
\<<results=tex>>=
d <- transform(d, race = combine.levels(race, minlev = 0.05))
@
Summaries of variables stratified by sex are below.
\<<results=tex>>=
l latex(summary(sex ~ ., method='reverse', data=d, test=TRUE),
    npct='both', dotchart=TRUE, file='', landscape=TRUE, round=1)
@
\section{Redundancy Analysis and Variable Interrelationships}
\bc
% Note: giving a chunk name to each code chunk that produces a figure
% makes it easy to know which plots to send to a collaborator, and
% will not allow numbered orphan plots to be left when code chunks are
% inserted into the file. The default in Sweave is for plots to be
% numbered by the chunks producing them.
<<vc,fig=T>>=
v <- varclus(~., data=d)
plot(v)
redun(~ age + sex + race + edu + income + dzgroup + meanbp + hrt, data=d)
# Alternative: redun(~., data=subset(d, select=-c(hospdead,slos)))
@
\ec
Note that the clustering of black with white is not interesting; this
just means that these are mutually exclusive higher frequency
categories, causing them to be negatively correlated.
\section{Logistic Regression Model}
Here we fit a tentative binary logistic regression model. The
coefficients are not very useful so they are not printed (\dots is
printed in their place).
<<z,eval=F,echo=T>>=
dd <- datadist(d); options(datadist='dd')
f <- lrm(hospdead ~ rcs(age,4) + sex + race + dzgroup + rcs(meanbp,5),
data=d) # see Section (*\ref{descStats}*) for descriptive statistics
SOURCE CODE FOR THIS REPORT

f
<<echo=F>>=
z <- capture.output( {
<<z>>>
  prselect(z, 'S.E.') # keep only summary stats; or:
  # prselect(z, stop='S.E.', j=-1) # keep only coefficients
}
)  
Better: Output model statistics \LaTeX\ markup, automatically
suppressing coefficients.
<<results=tex>>=
print(f, latex=TRUE, coefs=FALSE)
@

The mean arterial blood pressure effect is shown below, on the
probability scale.  \textbf{Note}: Here we use the figure
environment, with a caption. The \code{rmlines} shell script is run
to remove lines containing \code{rm} surrounded by sharp signs.
<<>>=
ppdf(meanbp)  #rm#
# Lattice graphics require print() to render
p <- Predict(f, meanbp, fun=plogis)
print(plot(p, ylab='Prob[hospital death]', adj.subtitle=FALSE))
# Figure (*\ref{fig:meanbp}* )
doff()  #rm#
@
\fg{meanbp}{fig:meanbp}{Partial effect of mean arterial blood pressure
adjusted to \Sexpr{attr(p, 'info')$adjust}.} 
<<results=tex>>=
latex(anova(f), where='h', file='') # can also try where='htbp'
@
The likelihood ratio $\chi^2$ statistic is
\Sexpr{round(f$stats['Model L.R.'],2)} on \Sexpr{f$stats['d.f.']} d.f.
The fitted model in algebraic form is found below.
<<results=tex>>=
latex(f, file='')
@
\section{Computing Environment}
These analyses were done using the following versions of \R\cite{Rsystem}, the
operating system, and add-on packages \code{Hmisc}\cite{Hmisc},
\code{rms}\cite{rrms}, and others:
<<echo=F, results=tex>>=
toLatex(sessionInfo(), locale=FALSE)
@
% Note: Rsystem reference is defined inside feh.bib. It is a slightly
% edited version of the output of citation().
\bibliography{/home/harrelfe/bib/feh.bib}
\bibliographystyle{unsrt}
% Use \bibliographystyle{abbrv} if want references alphabetized

\section{Source Code for This Report}
verbatimtabinput{sweaveEx.Rnw}
\section{Sweavel.sty}

\verbatimtabinput{/home/harrelfe/doc/latex/texinput/Sweavel.sty}

\end{document}

6 Sweavel.sty

% Usage: \usepackage{Sweavel}
% To change size of R code and output, use e.g.: \def\Sweavesize{\normalsize}
% To change just the size of output, use e.g.: \def\Routsize{\smaller[2]}
% To change colors of R code, output, and commands, use e.g.:
% \def\Rcolor{\color{black}}
% \def\Routcolor{\color{green}}
% \def\Rcommentcolor{\color{red}}
% To change background color or R code and/or output, use e.g.:
% \def\Rbackground{\color{white}}
% \def\Routbackground{\color{white}}
% To use rgb specifications use \color[rgb]{ , , }
% To use gray scale use e.g. \color[gray]{0.5}
% If you change any of these after the first chunk is produced, the
% changes will have effect only for the next chunk.

\NeedsTeXFormat{LaTeX2e}
\ProvidesPackage{Sweavel}{} % substitute for Sweave.sty using
% listings package with relsize
\RequirePackage{listings,fancyvrb,color,relsize,ae}
\RequirePackage[T1]{fontenc}
\IfFileExists{upquote.sty}{\RequirePackage{upquote}}{}
\providecommand{\Sweavesize}{\smaller}
\providecommand{\Routsize}{\Sweavesize}
\providecommand{\Rcolor}{\color[rgb]{0, 0.5, 0.5}}
\providecommand{\Routcolor}{\color[rgb]{0.461, 0.039, 0.102}}
\providecommand{\Rcommentcolor}{\color[rgb]{0.101, 0.043, 0.432}}
\providecommand{\Rbackground}{\color[gray]{0.91}}
\providecommand{\Routbackground}{\color[gray]{0.935}}
% Can specify \color[gray]{1} for white background or just \color{white}

\lstdefinestyle{Rstyle}{fancyvrb=false,escapechar=`,language=R,%
% basicstyle=\Rcolor{\Sweavesize},%
% backgroundcolor=\Rbackground,%
% showstringspaces=false,%
% keywordstyle=\Rcolor,%
% commentstyle=\Rcommentcolor\ttfamily\itshape,%
% literate={\leftarrow}{{$\leftarrow$}}2{\twoheadleftarrow}{{$\twoheadleftarrow$}}2{\sim}{{$\sim$}}1{\leq}{{$\leq$}}2{\geq}{{$\geq$}}2{\wedge}{{$^\wedge$}}1,%
% alsoother=, %
% alsoletter=\leftarrow,%
% otherkeywords={!,!,\+,\-,\&,,\%,\%,\%,\%,<<,<<,},%
% escapeinside={(*}{*)}%
% Other options of interest:
\lstdefinestyle{Routstyle}{fancyvrb=false,literate={\sim}{{$\sim$}}1{R^2}{{$R^{2}$}}2{\$}{{$\$$}}2{\wedge}{{$\wedge$}}1{R-squared}{{$R^{2}$}}2,frame=single,framerule=0.2pt,framesep=1pt,basicstyle=\Routcolor\Routsize,backgroundcolor=\Routbackground}

\newenvironment{Schunk}{}{}
\lstnewenvironment{Sinput}{\lstset{style=Rstyle}}{}
\lstnewenvironment{Scode}{\lstset{style=Rstyle}}{}
\lstnewenvironment{Soutput}{\lstset{style=Routstyle}}{}
\lstnewenvironment{Sinputsmall}{\lstset{style=Rstyle,basicstyle={\small}}}{ }
\lstnewenvironment{Sinputsmaller}{\lstset{style=Rstyle,basicstyle={\smaller}}}{ }
\lstset{style=Routstyle}
\endinput

sudo cp ~/doc/latex/texinput/Sweavel.sty /usr/share/R/share/texmf/.
sudo mktexlsr
Better handling of graphics; no more `print(xyplot())`
Simplified interface to tikz graphics
Simplified implementation of caching
More automatic pretty-printing; support for \LaTeX\ listings package built-in
Can specify figure captions in chunk headers along with R graphics parameters
Easy to include animations in pdf reports
Chunks can produce multiple plots

http://yihui.github.com/knitr
http://cran.r-project.org/web/packages/knitr
http://biostat.mc.vanderbilt.edu/KnitrHowto

---

```
spar <- function(mar=if(!axes)
  c(2.25+bot-.45*multi,2+left,.5+top+.25*multi,.5+rt) else
  c(3.25+bot-.45*multi,3.5+left,.5+top+.25*multi,.5+rt),
  lwd = if(multi)1 else 1.75,
  mgp = if(!axes) mgp=c(.75, .1, 0) else
  if(multi) c(1.5, .365, 0) else c(2.4-.4, 0.475, 0),
  tcl = if(multi)-0.25 else -0.4,
  bot=0, left=0, top=0, rt=0, ps=if(multi) 14 else 10,
  mfrow=NULL, axes=TRUE, ...)
{
  multi <- length(mfrow) > 0
  par(mar=mar, lwd=lwd, mgp=mgp, tcl=tcl, ps=ps, ...)
  if(multi) par(mfrow=mfrow)
}

render_listings()
unlink('messages.txt') # Start fresh with each run
hook_log = function(x, options) cat(x, file='messages.txt', append=TRUE)
knit_hooks$set(warning = hook_log, message = hook_log)
knit_hooks$set(par=function(before, options, envir)
  if(before & & options$fig.show != 'none')
  {
    p <- c('bty','mfrow','ps','bot','top','left','rt','lwd',
    'mgp','tcl', 'axes')
    pars <- opts_current$get(p)
    pars <- pars[!is.na(names(pars))]
    if(length(pars)) do.call('spar', pars) else spar()
  })
```
Setup Code, continued

```r
# Set short aliases for names of commonly used parameters
opts_knit$set(aliases=c(h='fig.height', w='fig.width',
cap='fig.cap', scap='fig.scap'))
opts_knit$set(eval.after = c('fig.cap','fig.scap'))
## see http://yihui.name/knitr/options#package_options
## Use caption package options to control caption font size

Code for Beginning of Report or Chapter

```r
<<echo=FALSE>>=
source('...file listed above...')
\SweaveOpts{fig.path='plot-', fig.align='center', w=4.5, h=3.5,
fig.show='hold', fig.pos='htbp', par=TRUE, tidy=FALSE}
@
Code for a Chunk

<<bigplot,h=7,w=7,cap='A \textbf{caption} for the figure'>>=
# need to double backslashes to escape them
<<example2, cap=paste('Survival curves for study', study_name)>>=
<<this, results='tex'>>=
# need to put character values in quotes with knitr, unlike Sweave
<<that, ps=6, mfrow=c(2,2)>>=
plot(something)  # Figure (*\ref{fig:xxx-that}*)
[symbolic reference from R to LaTeX]

Linux Shell Script for Running in Batch Mode

rm -f messages.txt
xterm -hold -e R --no-save --no-restore -e \
"require(knitr); knit('$1.Rnw')"
echo PDF graphics produced:
ls -lgt *.pdf
This work used only free software

\LaTeX

Flowchart from Google+ Technics

References


Much of research that uses data analysis is not reproducible. This can be for a variety of reasons, the most major one being poor design and poor science. Other causes include tweaking of instrumentation, the use of poorly studied high-dimensional feature selection algorithms, programming errors, lack of adequate documentation of what was done, too much copy and paste of results into manuscripts, and the use of spreadsheets and other interactive data manipulation and analysis tools that do not provide a usable audit trail of how results were obtained. Even when a research journal allows the authors the “luxury” of having space to describe their methods, such text can never be specific enough for readers to exactly reproduce what was done. All too often, the authors themselves are not able to reproduce their own results. Being able to reproduce an entire report or manuscript by issuing a single operating system command when any element of the data change, the statistical computing system is updated, graphics engines are improved, or the approach to analysis is improved, is also a major time saver.

It has been said that the analysis code provides the ultimate documentation of the “what, when, and how” for data analyses. Eminent computer scientist Donald Knuth invented literate programming in 1984 to provide programmers with the ability to mix code with documentation in the same file, with “pretty printing” customized to each. Lamport’s \LaTeX, an offshoot of Knuth’s T\TeX typesetting system, became a prime tool for printing beautiful program documentation and manuals. When Friedrich Leisch developed Sweave in 2002, Knuth’s literate programming model exploded onto the statistical computing scene with a highly functional and easy to use coding standard using R and \LaTeX and for which the Emacs text editor has special dual editing modes using ESS. This approach has now been extended to other computing systems and to word processors. Using R with \LaTeX to construct reproducible statistical reports remains the most flexible approach and yields the most beautiful reports, while using only free software. One of the advantages of this platform is that there are many high-level R functions for producing \LaTeX markup code directly, and the output of these functions are easily directly to the \LaTeX output stream created by Sweave.

See ctspedia.org, reproducibleresearch.net, groups.google.com/group/reproducible-research, and biostat.mc.vanderbilt.edu/SweaveLatex for more information.