

Statweave: a step on the path towards literate programming

Nicholas J. Horton

Smith College, Northampton, MA, USA

June 10, 2008

nhorton@email.smith.edu

<http://www.math.smith.edu/~nhorton/statweave>

Plan for talk

- background on literate programming
- goals and motivation
- history of application to statistics
- example
- more details
- conclusions and future work (Statdocs as the holy grail)

Background on literate programming

- integrate code and documentation
- automate documentation generation
- (particularly) useful to statisticians as reproducible statistical analysis

Background on literate programming

Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to humans what we want the computer to do. (Donald E. Knuth, 1984).

Background on literate programming

The purpose of Sweave is to create dynamic reports, which can be updated automatically if data or analysis change. Instead of inserting a prefabricated graph or table into the report, the master document contains the R code necessary to obtain it. When run through R, all data analysis output (tables, graphs, . . .) is created on the fly and inserted into a final LATEX document. The report can be automatically updated if data or analysis change, which allows for truly reproducible research. (Leisch, R-News 2002)

Reproducible statistical analysis

- document details of derived variables and subsetting
- facilitate re-running analyses with updated datasets
- goal: avoid cut and paste errors
- goal: single document (compendium) to generate different views for different audiences (Gentleman and Temple Lang, 2004)
- useful for collaborative work as well as teaching (Nolan and Temple Lang, 2007)

History of reproducible statistical analysis

- noweb (Ramsay, 1994) [support for programming language]
- R and XML (Temple Lang, 2001) [specialized]
- Sweave (Leisch, 2002) [S-plus and R only and \LaTeX , quite popular]
- odfWeave (Kuhn, 2007) [extends Sweave to open document format, .odt]
- SASweave (Lenth and Hojsgaard, 2007) [added Sweave-like support for SAS]

StatWeave

- builds and generalizes prior systems
- supports multiple output formats (tex, dvi, pdf, odt)
- supports many languages/engines (R, S-plus, Stata, SAS, IML, Maple, Linux)
- plans for support for Matlab, Mathematica, GenStat
- runs under Windows, Mac OS X and Linux
- installed on `stat12.stat.auckland.ac.nz`
- downloadable from
`http://www.stat.uiowa.edu/~rlenth/StatWeave`
- sample files in
`http://www.math.smith.edu/~nhorton/statweave`

Running StatWeave

- code chunks for each engine are collected
- code file are run (in order)
- output is collected and embedded

Example

suitable header, then

```
\begin{Rcode}
lmres <- lm(logfev~age+loght+initage+loginiht, data=ds)
summary(lmres)
\end{Rcode}
```

Output of regression

```
R> lmres <- lm(logfev~age+loght+initage+loginiht, data=ds)
R> summary(lmres)
```

Call:

```
lm(formula = logfev ~ age + loght + initage + loginiht, data = ds)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.46501	-0.07267	0.00192	0.07922	0.37385

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.33289	0.02071	-16.07	< 2e-16
age	0.02868	0.00208	13.81	< 2e-16
loght	2.04343	0.06880	29.70	< 2e-16
initage	-0.01498	0.00396	-3.78	0.00016
loginiht	0.39223	0.08263	4.75	2.2e-06

Residual standard error: 0.114 on 1988 degrees of freedom

Multiple R-squared: 0.88, Adjusted R-squared: 0.88

F-statistic: 3.64e+03 on 4 and 1988 DF, p-value: <2e-16

Execute a block of code

At this point we can access particular values, for example, the age of the subject in row `\Rexpr{i}` is `\Rexpr{age[i]}`.

generates:

At this point we can access particular values, for example, the age of the subject in row 4 is 12.46.

Format regression results

```
\begin{Rcode}{results=tex}  
library(xtable)  
lmtab <- xtable(lmres,digits=c(0,3,3,2,4),  
               caption="Better formatted results")  
print(lmtab)  
\end{Rcode}
```

Formatted output of regression

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.333	0.021	-16.07	0.0000
age	0.029	0.002	13.81	0.0000
loght	2.043	0.069	29.70	0.0000
initage	-0.015	0.004	-3.78	0.0002
loginiht	0.392	0.083	4.75	0.0000

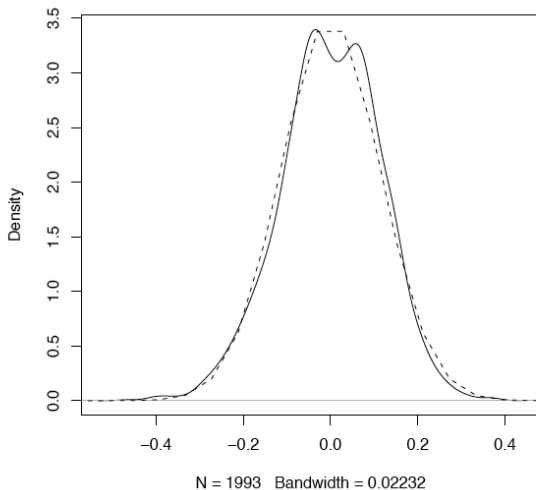
Table 1: Better formatted results

Create a plot

Figure `\ref{bplot}` displays the empirical density of the residuals.

```
\begin{figure}[tbph]
\caption{Empirical density of residuals}
\label{bplot}
\begin{center}
\begin{Rcode}{fig,label=bplot}
plot(density(resid),main="")
x <- seq(from=-3,to=3,length=100)
lines(x,dnorm(x,0,sd(resid)),lty=2)
\end{Rcode}
\end{center}
\end{figure}
```

Plot of residuals



Open Office (.odt)

- slightly different syntax
- examples on my website
- can be saved as Microsoft Word (.doc) format

Open Office

```
R:  
options(digits=3)  
options(show.signif.stars=FALSE)  
ds <- read.csv("http://www.math.smith.edu/~nhorton/statweave/fev.csv")  
attach(ds)  
i <- 4  
ds[i,]
```

At this point we can access particular values, for example, the age of the subject in row R

Open Office

```
R> options(digits=3)
R> options(show.signif.stars=FALSE)
R> ds <- read.csv("http://www.math.smith.edu/~nhorton/statw
R> attach(ds)
R> i <- 4
R> ds[i,]
```

```
  id height  age  inihth  initage  logfev  loght  loginihth
4  1   1.42 12.5    1.2    9.34  0.751 0.351    0.182
```

Collaborative research project: repeatability of distortion product otoacoustic emissions

- data merging (subject information, session information, distortion product measurements)
- data cleaning
- data summaries
- data analysis

Summary

- reproducible statistical analyses are a (very) good thing
- integrating documentation and output is helpful
- ability to rerun analyses with different datasets or analytic decisions is a big win
- StatWeave is portable and extensible

Future work and closing thoughts

- Statweave creates static documents, which can be extremely helpful
- particularly useful for statisticians in training or starting to work on collaborative research
- Creative people are working on extensions (DynDocs and StatDocs, Temple Lang and Nolan) that have even more potential