Generazione automatica di report con R e LaTeX

Maurizio W. Himmelmann*
Ufficio Statistica
Scuola Superiore Sant'Anna

Emiliano G. Vavassori†
Università di Pisa

Articolo presentato al secondo Convegno Nazionale su TeX, LaTeX e Tipografia Digitale

Pisa, 22 ottobre 2005

Sommario

Il presente articolo si propone di illustrare quanto facile sia integrare le potenzialità LaTeX con il programma di analisi statistica R. L'obiettivo è quello di realizzare un file sorgente ibrido contenente istruzioni relative sia alla stesura del testo (LaTeX) sia all'analisi dei dati (R) ed in grado di leggere i dati da un file esterno, elaborarli ed inserire i risultati all'interno di un report LaTeX.

Indice

*email: `himmel@sssup.it`
†email: `testina@sssup.it`

1 Introduzione

Molto spesso all'interno della consueta attività lavorativa ci si trova a dovere realizzare lavori ripetitivi con cadenza periodica che vengono rapidamente etichettati come un routine noiose e *time-consuming*. In questi casi è possibile usare dei programmi in grado di svolgere automaticamente questi compiti, limitando così al minimo l'intervento dell'operatore. L'uso sinergico di due programmi quali R e LaTeX permette di assolvere in modo egregio alla generazione automatica di report statistici.

All'interno di questo articolo è convenzionalmente utilizzato un `riquadro` per indicare un sorgente od una sua parte, il simbolo `>` per il prompt di R e infine `$` per il prompt dell'interprete dei comandi.

Sebbene non direttamente specificato nel corso del testo, tutto quanto riportato in quest'articolo è perfettamente applicabile anche al programma di analisi statistica S-PLUS.

2 Il programma statistico R

R più che un semplice software statistico può essere definito come un vero e proprio ambiente di programmazione orientato alla gestione ed all'analisi dei dati. Scritto nel 1996 da Ross Ihaka e Robert Gentelman, del Dipartimento di Statistica dell'Università di Auckland, R può essere considerato un dialetto di S, linguaggio di programmazione statistico sviluppato nel 1980 nei *Bell Labs* su cui è stato costruito il noto programma S-PLUS. A differenza di quest'ultimo tuttavia, R è un software *open-source*, è disponibile per tutti i principali sistemi operativi (Linux, Windows, MacOS) e viene gratuitamente distribuito con licenza GPL (General Public License).

Esso è inoltre estremamente flessibile e permette l'implementazione delle più svariate funzioni di calcolo e di rappresentazione grafica statistica, con standard di qualità e di precisione superiori a molti altri software di analoghe finalità. In particolare esso mette a disposizione:

- un efficace manipolatore di dati e un altrettanto efficace dispositivo di memorizzazione;
- un insieme di operatori per i calcoli su array, in particolare matrici;
- una grande, coerente, integrata raccolta di strumenti intermedi per l'analisi dei dati;
- risorse grafiche per l'analisi dei dati con visualizzazione direttamente su video o su carta attraverso stampante;
- un ben sviluppato, semplice ed efficace linguaggio di programmazione che include istruzioni condizionali, loop, funzioni ricorsive definite dall'utente e strumenti di input/output.

In modo del tutto analogo a quanto avvenuto per LaTeX, R è riuscito a catturare l'attenzione e l'interesse di un numero sempre maggiore di appassionati. La comunità formatasi ha contribuito in modo determinante ad aumentarne la

potenza di calcolo e le funzionalità di analisi attraverso la creazione di packages dedicati.

È possibile scaricare gratuitamente R da uno dei numerosi CRAN (Comprehensive R Archive Network), raggiungibili dalla pagina principale del Progetto: `http://www.r-project.org/`

2.1 Il pacchetto *Sweave*

Sweave è un pacchetto di R sviluppato da Friedrich Leisch[1] che permette di generare automaticamente report statistici di alto livello. Nelle ultime versioni di R questo pacchetto è già incluso nella dotazione di base del programma e non occore quindi effettuare alcuna installazione aggiuntiva.

3 Scrittura del file ibrido

L'obiettivo di *Sweave* [2] è quello leggere un file ibrido, contenente comandi R all'interno del corpo di un documento LaTeX, procedendo a sostituire le istruzioni, relative all'analisi dei dati (lettura, elaborazione, tabelle e grafici), con un output interpretabile da LaTeX e che rappresentano i risultati di tale analisi. Per fare questo è sufficiente realizzare il sorgente ibrido, ottenuto combinando insieme codice LaTeX e codice R, precompilare questo sorgente con R e successivamente compilare con LaTeX l'output ottenuto.

D'ora in avanti, in accordo con la letteratura esistente, ci riferiremo alle porzioni di codice R nel documento ibrido come *chunks* (porzioni); è necessario definire delle stringhe *delimitatrici* che indichino ad R le porzioni di codice da prendere in considerazione. *Sweave* è in grado di interpretare due possibili delimitazioni, analoghe per utilizzo e finalità ma leggermente diverse per sintassi: *Noweb* e `SweaveSyntaxLatex`.

Essendo le funzionalità dei due sistemi di delimitazione del tutto equivalenti, in questo articolo verrà trattato in maniera più approfondita solo il secondo caso, ricordando che qualsiasi opzione o comando di seguito definito sarà comunque valido per entrambi i sistemi.

3.1 I *chunks* di tipo *Noweb*

Noweb [7] è un semplice programma che permette di combinare insieme codice sorgente di un programma e la corrispondente documentazione all'interno di un singolo file. *Noweb* è lo strumento usato di default da *Sweave* per discernere il codice R dal codice LaTeX. Tali *chunks* sono solitamente definiti come:

```
<<>>=
...
@
```

Ogni *chunk* di codice R viene pertanto definito dal comando di apertura "`<<>>=`" e dal comando di chiusura "`@`"; tali delimitatori sono posizionati all'inizio della riga. Tradizionalmente, un file ibrido con *chunks Noweb* avrà

[1] Institut für Statistik und Wahrscheinlichkeitstheorie - Technische Universität Wien

estensione `.nw`, ma sono anche molto diffuse le estensioni `.rnw`, `.Rnw`, `.snw` e `.Snw`.

3.2 I *chunks* di tipo `SweaveSyntaxLatex`

Le ultime versioni di *Sweave* sono in grado di interpretare anche dei *chunks* definiti in un linguaggio più familiare agli utenti LaTeX. In questo caso i *chunks* assumono quest'aspetto:

```
\begin{Scode}{<opzione1>, <opzione2>}
...
\end{Scode}
```

Per comunicare a *Sweave* che sarà usato il sistema "*LaTeX friendly*", a cui si è soliti riferirsi con il nome di `SweaveSyntaxLatex`, è sufficiente modificare l'estensione del file ibrido in `.Rtex` o `.Stex`.

3.3 Gestione dei *chunks* `SweaveSyntaxLatex`

L'ambiente `\begin{Scode}` permette l'uso di numerose opzioni, le principali sono:

`label` essa permetta di richiamare il risultato prodotto dal *chunk* anche in altre parti del documento;

`echo=FALSE` permette di disabilitare la scrittura del codice computato da R sul documento LaTeX (default `=TRUE`);

`fig=TRUE` impone la generazione di una figura in formato `.eps` e `.pdf`. Tali file sono creati all'interno della cartella corrente e saranno poi richiamati durante la successiva fase di compilazione con LaTeX (default `=FALSE`);

`results=tex` impone che l'output sia prodotto nel font usato nel documento e non con font `verbatim` come di default.

`results=hide` nasconde qualsiasi risultato prodotto all'interno del *chunk*.

Inoltre sono anche a disposizione i comandi `\Sexpr{<comando R>}`, che permette di eseguire una singola istruzione R ottenendo l'output nello stesso font del documento, ed il comando `\SweaveOpts{<opzione1>, <opzione2>}` che modifica le opzioni di default dalla sua occorrenza fino alla fine del documento.

3.4 Esempio

Di seguito è riportato il codice del file `esempio.Rtex`, scritto utilizzando il sistema `SweaveSyntaxLatex`, che prenderà in esame i dati dell'esempio `airquality` già presenti[2] a scopo didattico in tutte le versioni di R.

[2]Per la compilazione del file riportato come esempio è stato caricato anche il pacchetto `xtable` per la realizzazione di tabelle direttamente in codice LaTeX. Quest'ultimo, non disponibile nella dotazione standard di R è gratuitamente scaricabile dall'indirizzo `http://cran.at.r-project.org/src/contrib/Descriptions/xtable.html`.

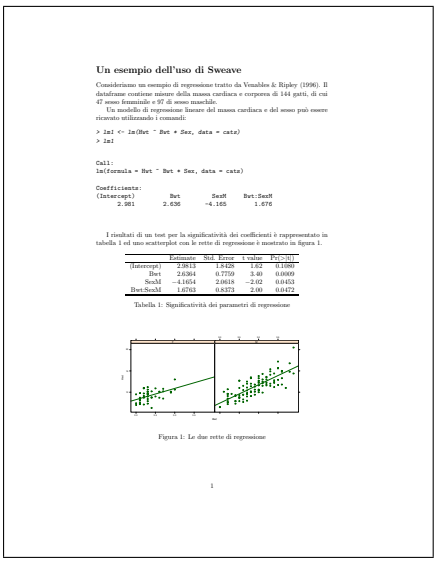
```
\documentclass[a4paper]{article}
\usepackage[italian]{babel}
\usepackage[utf8x]{inputenc}
\begin{document}

\section*{Un esempio dell'uso di Sweave}

\begin{Scode}{echo=FALSE}
 library(lattice)
 library(xtable)
 data(cats, package="MASS")
\end{Scode}

Consideriamo un esempio di regressione tratto da Venables \& Ripley
(1996). Il dataframe contiene misure della massa cardiaca e corporea
di \Sexpr{nrow(cats)} gatti, di cui \Sexpr{sum(cats$Sex=="F")} sesso
femminile e \Sexpr{sum(cats$Sex=="M")} di sesso maschile.

Un modello di regressione lineare del massa cardiaca e del sesso
può essere ricavato utilizzando i comandi:

\begin{Scode}
 lm1 <- lm(Hwt~Bwt*Sex, data=cats)
 lm1
\end{Scode}

I risultati di un test per la significatività dei coefficienti è
rappresentato in tabella~\ref{tab:coeff} ed uno scatterplot con
le rette di regressione è mostrato in figura~\ref{fig:cats}.

\SweaveOpts{echo=false}

\begin{Scode}{results=tex}
 xtable(lm1,
  caption="Significatività dei parametri di regressione",
  label="tab:coeff")
\end{Scode}

\begin{figure}[ht]\centering
 \begin{Scode}{fig=TRUE, width=12, height=6}
  trellis.par.set(col.whitebg())
  print(xyplot(Hwt~Bwt|Sex, data=cats, type=c("p","r")))
 \end{Scode}
 \caption{Le due rette di regressione}
 \label{fig:cats}
\end{figure}

\end{document}
```

Per compilare il file è sufficiente lanciare dalla console di R il comando:

```
> Sweave("<path>/esempio.Rtex")
```

che produce il file `esempio.tex` nella cartella di lavoro corrente.

Un successivo step di compilazione con LaTeX permette a questo punto di ottenere il risultato mostrato in figura 1:



Un esempio dell'uso di Sweave

Consideriamo un esempio di regressione tratto da Venables & Ripley (1996). Il dataframe contiene misure della massa cardiaca e corporea di 144 gatti, di cui 47 sesso femminile e 97 di sesso maschile.

Un modello di regressione lineare del massa cardiaca e del sesso può essere ricavato utilizzando i comandi:

```
> lm1 <- lm(Bwt ~ Bwt * Sex, data = cats)
> lm1

Call:
lm(formula = Bwt ~ Bwt * Sex, data = cats)

Coefficients:
(Intercept)          Bwt         SexM     Bwt:SexM
      2.981        2.636      -4.165        1.676
```

I risultati di un test per la significatività dei coefficienti è rappresentato in tabella 1 ed uno scatterplot con le rette di regressione è mostrato in figura 1.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.9813	1.8428	1.62	0.1089
Bwt	2.6364	0.7759	3.40	0.0009
SexM	−4.1654	2.0618	−2.02	0.0453
Bwt:SexM	1.6763	0.8373	2.00	0.0472

Tabella 1: Significatività dei parametri di regressione

Figura 1: Le due rette di regressione

1

Figura 1: Documento compilato a partire dal file `esempio.Rtex`.

3.5 Personalizzare il proprio report

Sweave richiama automaticamente all'interno del sorgente LaTeX il pacchetto `Sweave.sty`; tale pacchetto non è presente su CTAN e normalmente non si trova negli alberi della distribuzione LaTeX in quanto è installato congiuntamente con R (nei sistemi GNU/Linux ad esempio si trova in `/usr/lib/R/share/texmf/`). Questo pacchetto definisce ambienti che sono necessari a LaTeX per interpretare l'output di *Sweave* e fornisce dei valori di default che possono essere personalizzati.

Per modificare la larghezza delle figure inserite (che è indipendente dalla dimensione dei file `.eps` o `.pdf` generati da R) è necessario specificare la seguente istruzione subito `\begin{document}`:

```
\setkeys{Gin}{width=<width>}
```

dove `<width>` è il parametro da modificare, di default pari a `0.8\textwidth`.

4 Come semplificarsi ulteriormente la vita

4.1 Automazioni con `make`

In ambiente GNU/Linux è possibile utilizzare un semplice script di bash per velocizzare ulteriormente la procedura di compilazione. Quello che segue è un esempio minimale che procede in modo automatico a tutti gli step di compilazione del documento ed alla successiva, eventuale, rimozione dei file ausiliari.

```
nomefile="esempio"

pdf: compileR
 pdflatex ${nomefile}.tex;
 pdflatex ${nomefile}.tex;

compileR:
 echo -e "Sweave(\"${nomefile}.Rtex\")" | R CMD BATCH -

clean:
 rm -f *.aux *.log *.Rout;
```

Il comando completo per generare un report `.pdf` è il seguente:

```
make nomefile="<nome>" pdf
```

dove `<nome>` è il nome del file ibrido scritto omettendo l'estensione.

4.2 *Emacs* ed ESS

Il pacchetto ESS (Emacs Speak Statistics) permette di disporre di un ambiente di lavoro perfettamente integrato fra R e LaTeX. Ulteriori informazioni sono disponibili consultando il manuale di ESS [8].

4.3 Il pacchetto `Rtangle`

`Rtangle` (o `Stangle`) estrae da un file *Sweave* il codice R in esso contenuto, trasferendolo su un nuovo file con estensione "`.R`". Questo consente di compilare separatamente la parte relativa all'elaborazione dei dati senza bisogno di dovere necessariamente ottenere il report finito per visualizzare i risultati delle elaborazioni.

Riferimenti bibliografici

[1] Peter Dalgard. *Introductory Statistics with R*. Spinger 2002

[2] Friedrich Leisch, *Sweave Dynamic generation of statistical reports using literate data analisys*. Compstat 2002 – Proceedings in Computational Statistics. 2005

[3] Friedrich Leisch, *Sweave User Manual*, Vienna 2005
`http://www.ci.tuwien.ac.at/~leisch/Sweave`

[4] Friedrich Leisch, *Sweave and Beyond: Computation on Text documents*. Proceedings of the 3rd International Workshop on Distributed Statistical Computing. 2003.
`http://www.ci.tuwien.ac.at/~leisch/Conferences/DSC-2003/`

[5] Angelo M. Mineo, *Una guida all'utilizzo dell'ambiente statistico R*. 2003
`http://cran.r-project.org/doc/contrib/Mineo-dispensaR.pdf`

[6] Vito M. R. Muggeo, *Il linguaggio R: concetti introduttivi ed esempi*. 2002
`http://cran.r-project.org/doc/contrib/nozioniR.pdf`

[7] Norman Ramsey. *Literate programming simplified*. IEEE Software. 1994
`http://www.eecs.harvard.edu/~nr/noweb/`

[8] Antony J. Rossini, Richard M. Heiberger, Kurt Hornik, Martin Maechler, Rodney A. Sparapani e Stephen J. Eglen. *ESS: Emacs Speaks Statistics*.
`http://ess.r-project.org/`

[9] William N. Venables e Brian D. Ripley. *S Programming*. Springer 2000