"Developments in Statistical Computing Software"

Ian McLeod

November 13, 2014. Systems Design Engineering, University of Waterloo

http://www.stats.uwo.ca/faculty/aim/2014/SydSeminar/ LINK

In[1]:=
```
SetDirectory[NotebookDirectory[]]
```

Out[1]=
```
D:\DropBox\math\2014\misc\SYDE
```

Overview of R

http://cran.r-project.org/

R provides an advanced and sophisticated statistical computing environment.

R can be used in an interactive manner often called **repl** - Read-evaluate-print-loop. APL and Lisp were among the first widely used such computer languages and currently we have many others including MatLab, Maple, python, perl, etc. These computer languages are easy to learn and are very powerful. While the execution speeds don't usually compare favourable with expertly written programs in languages like C or C++ they are often good enough to get the job done. In fact for they are often much faster if programming time is taken into account.

At a higher level than repl, R supports scripts, functional programs, many data formats and packages.

CRAN

R has many thousands users world-wide and is contantly under development. Many researchers have published data and software on the R website CRAN. R is free. It has also been incorporated in proprietary software including Excel, SPSS, SAS and *Mathematica*. There is a vast enterprise of consulants, blogs, books and refereed journals that support R.

To see the scope of R checkout CRAN Views on the CRAN website, http://cran.r-project.org/web/views/ LINK.

Reproducible research

Knuth (1992) introduced the idea of **Literate Programming** that combines text and computer code into a single file with the concepts of tangle, to extract and run the computer code, and weave, to produce from the file a human readable text. Ramsey (1994) introduced the **noweb** paradigm for implementing the ideas of Knuth in a general and practical way. The noweb paradigm has been used in R to develop the Sweave software that is built-in to R and RStudio for producing beautiful dynamic documents.

Reproducible Research is important not only for the advancement of Science but also for teaching, education, proprietary research by industry and finance as well as for large research programs carried out by graduate students and professors.

Old concept and started to gain traction with the paper of Buckheit and Donoho (1995). Wavelab and reproducible research. More recently in the R community there is the paper, Gentleman and Lang (2007). Statistical Analyses and Reproducible Research.

R and time series analysis

With respect to time series, R is the most comprehensive and best computing environment for most purposes. My survey paper discusses many state-of-the-art computing developments in time series analysis that are available in R.

McLeod, A. I., H. Yu and E. Mahdi (2012). Time Series Analysis with R. In *Time Series Analysis: Methods and Applications*, Chapter 23 (pp. 661-712) in Handbook in Statistics, Volume 30, Edited by T. S. Rao, S. S. Rao and C. R. Rao. ISBN: 978-0-444-53858-1. Elsevier. http://www.sciencedirect.com/science/handbooks/01697161. Preprint & Online Appendix: http://www.stats.uwo.ca/faculty/aim/tsar/ LINK.

R blogosphere

Tal Galili. Provides central hub with content from many other blogs about R, sometimes called the R blogosphere.

```
http://www.r-bloggers.com/
```

R Packages

R packages are similar to libraries in C or better to toolkits in MatLab or packages in *Mathematica*. An R packages contains one more functions and possibly some relevant datasets. R functions may be interfaced to C, C++, Fortan or Java. Each function and dataset is documented. Additional documentation in the form of an overview, user's manual, research report and demonstrations may also be included. Encapsulating the R code in a package improves the reliability of the functions as well as making it easy to be used for later use. R functions may be uploading to CRAN and, if accepted, they are made available for Windows, Mac and Linux OS. My son Matthew published his first R package, mvrtn, on CRAN this summer!

Dynamic graphics and R

Purpose of graphics

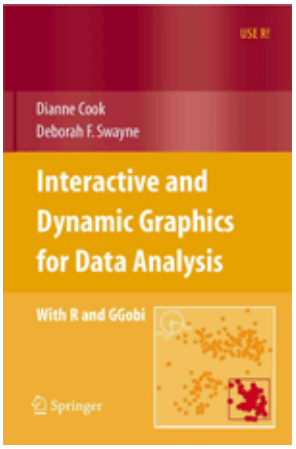
1. data analysis and discovery
2. presentation

GGobi and rggobi

Dynamic statistical graphics was pioneered at Bell Labs and Princeton starting in the 1980's with the XGobi software and its latest version is GGobi. GGobi can be used standalone or using an interface provided by the rgobi package on CRAN. I prefer to use the standalone version since it is more reliable.

A major idea in dynamic statistical graphics are interlinked plots that enable brushing. A small rectangle is moved over points to select them and they are indicated on various interlinked plots.

Another interesting ideas was “grand tours”, dynamic 3D plots of higher dimensional data projected into 3 dimensions and viewed as an animation. The algorithm attempts to automatically find all interesting projections.



D. Cook and D. F. Swayne (2007). *Interactive and Dynamic Graphics for Data Analysis: With Examples Using R and GGobi*.

http://www.ggobi.org/ LINK

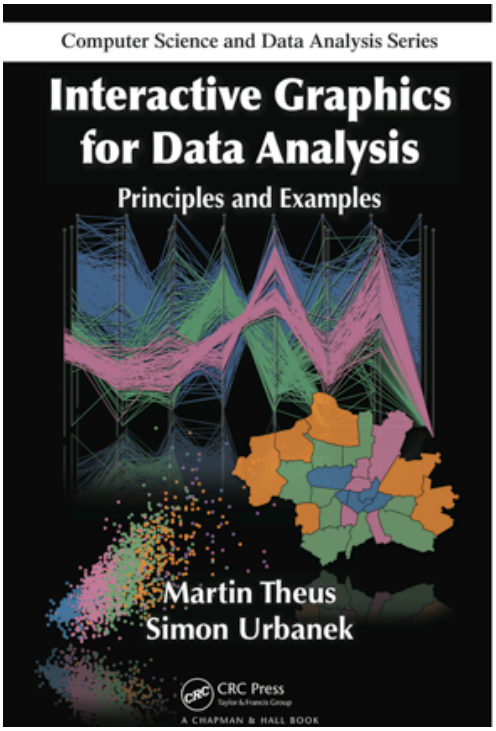
http://www.ggobi.org/rggobi/ LINK

Air Quality Data. This data is used to Ilustrate GGobi. It consists of 111 observations on successive days of the ground ozone with three dependent variables: temperature at noon, windspeed, and solar radiation.

Mondrian and iplots

http://www.rosuda.org/iplots/ LINK

http://www.theusrus.de/Mondrian/ LINK



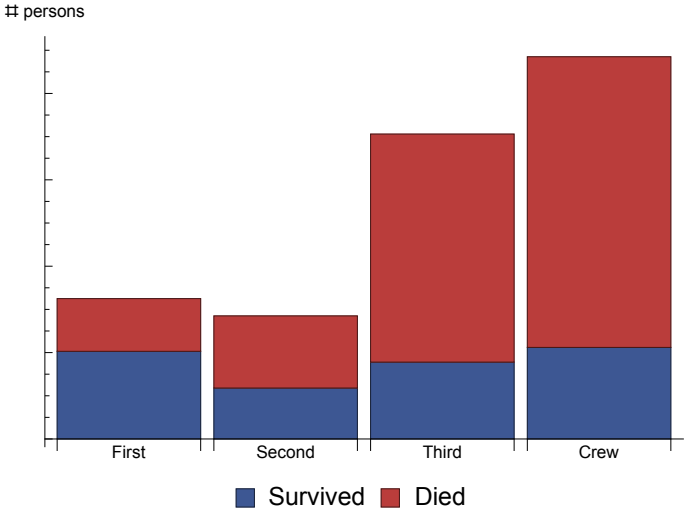
http://www.interactivegraphics.org/Home.html LINK

Mosaic Display. The mosaic display is comprised of rectangles that form a partition of a larger rectangle that represents the entire dataset. The subrectangles are created recursively so that the area of each subrectangle is proportion to the observed count or frequency in the original contingency table. The choice of color and spacing between rectangles can aid in the perception of patterns and insight.

	1st Class	2nd Class	3rd Class	Crew
Adult Female Survived	140	80	76	20
Adult Male Survived	57	14	75	192
Adult Female Died	4	13	89	3
Adult Male Died	118	154	387	670
Child Female Survived	1	13	14	0
Child Male Survived	5	11	13	0
Child Female Died	0	0	17	0
Child Male Died	0	0	35	0

Interactive stacked barchart of frequencies for 'Class' and 'Survived'

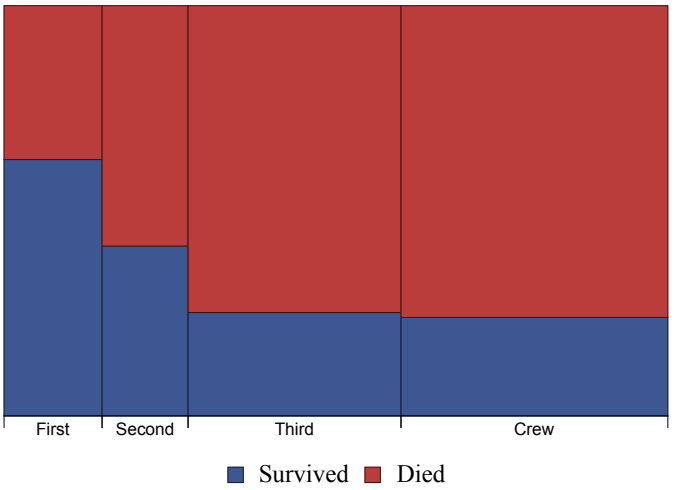
```
BarChart[Transpose[{ClassYes, ClassNo}],
 ChartLayout → "Stacked", ChartLabels → {ClassNames, None},
 ChartLegends → Placed[{"Survived", "Died"}, Below], ChartStyle → "DarkRainbow",
 (*PlotLabel→Style["Titanic, Interactive Stacked Barchart",
   "Title",16,Black],*)AxesLabel → {None, "# persons"}]
```



Spineplot with interaction of frequencies for 'Class' and 'Survived'

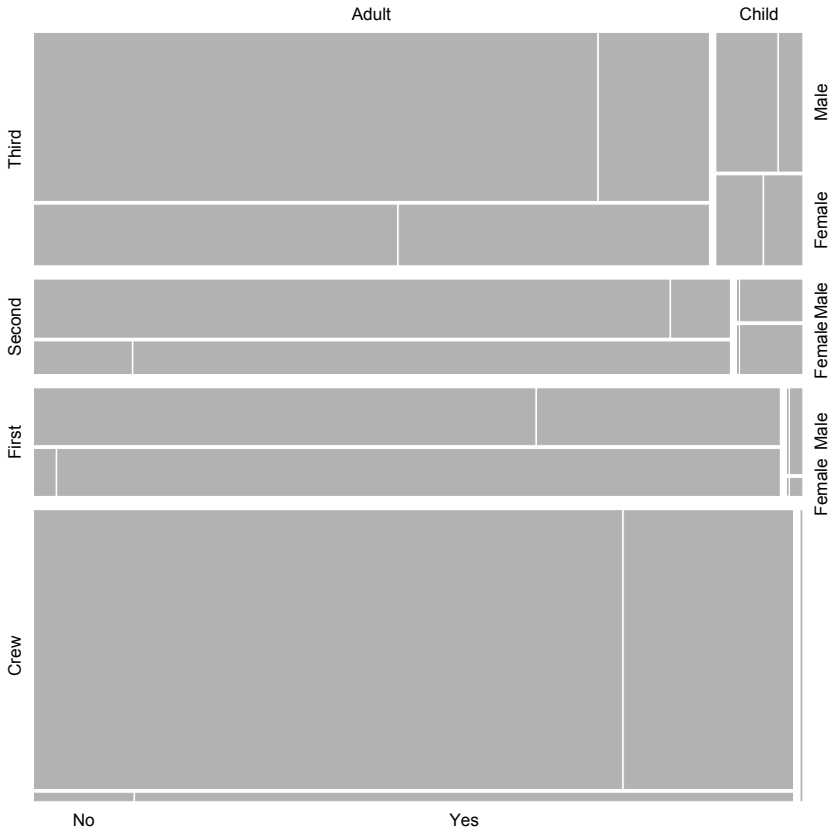
The spineplot is like a stacked barchart except we use width of the bar rather than its height to indicate the count or relative frequency for each category on the horizontal axis. We don't need the vertical axis scale to interpret the plot since we just look at the areas. This simpler and leads to the generalization to include more faactors.

The splineplot is basically a one-dimensional version of the mosaicplot. The total red and blue area represents the 2201 passengers.



Mosaic plot using Antonov's *Mathematica* Package, MosaicPlot.m

I obtained Antovov's package from his blog but the colour option is not implemented in the code he provided.



Sample Mondrian Session showing the Titanic Data. Martin Theus.
http://vimeo.com/71355383 LINK

googleVis

The R CRAN package googleVis provides an interface between R and the Google Charts API.

Perhaps the best known example of the Google Chart API is the motion chart, popularised by Hans Rosling in his 2006 TED talk. Hans Rosling: ***"The best stats you've ever seen"***

```
http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen  LINK
```

RStudio

http://www.rstudio.com/

LINK

RStudio is the newest and by far the best R “IDE” - Integrated Development Environment. The source editor has many nice features so it is simply the best editor for R.

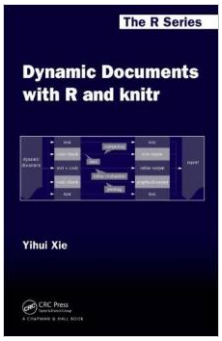
RStudio supports the concept of ***reproducible research*** by incorporating recently developed features in R for preparing beautiful interactive documents based tables and figures that you generate within the document. Reproducible research is becoming a requirement for publications in many scientific journals and is important in business, industry and education.

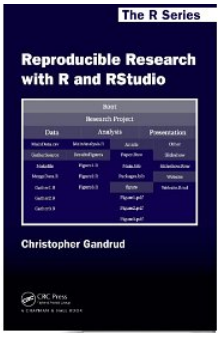
Dynamic documents may be generated in PDF, HTML or Word format. Documents are said to be dynamic if they are generated from a markup file that contains text , data and computer code. The computer code when run produces tables and figures. When processed, a beautiful or at minimum a very readable document is produced. The document is said to be dynamic because it may be produced on the fly and every step in creating the computational results, graphs and tables, is exactly reproducible.

Dynamic documents may contain ***dynamic graphics***, that is, graphics that we can interact with using the mouse and possibly some GUI widget, see Wikipedia “Graphical control element”.

- great text editor for R
- full IDE with support for debugging using breakpoints
- markdown and sweave for dynamic documents
- projects, new method for working with data, R scripts and other resources
- creating R packages
- configured to work with Git/GitHub
- Create dynamic graphics and webpages with Shiny and Google Charts

Germane RStudio Resources





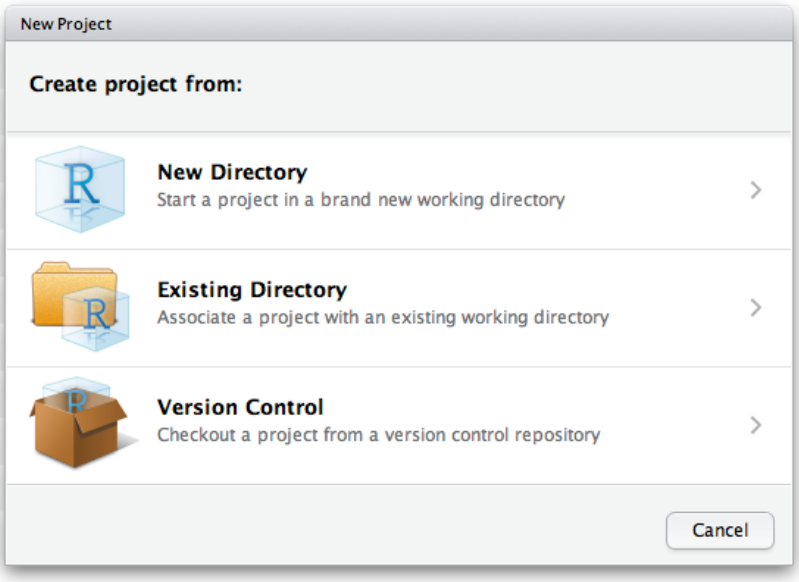
Yihui Xie (2013). Dynamic Documents with R and knitr Paperback
C. Gandrud (2014). Reproducible Research with R and RStudio.

http://www.rstudio.com/resources/training/online-learning/ LINK

https://support.rstudio.com/hc/en-us/categories/200035113-Documentation LINK

RStudio projects

Includes folders and subfolders as described above but in addition RStudio greatly enhances the organization. An RStudio project contains a folder and subfolders plus a history and the state at which you left the project when you last worked on it. RStudio provides the capability to work with these projects on GitHub using Git version control.



Markdown

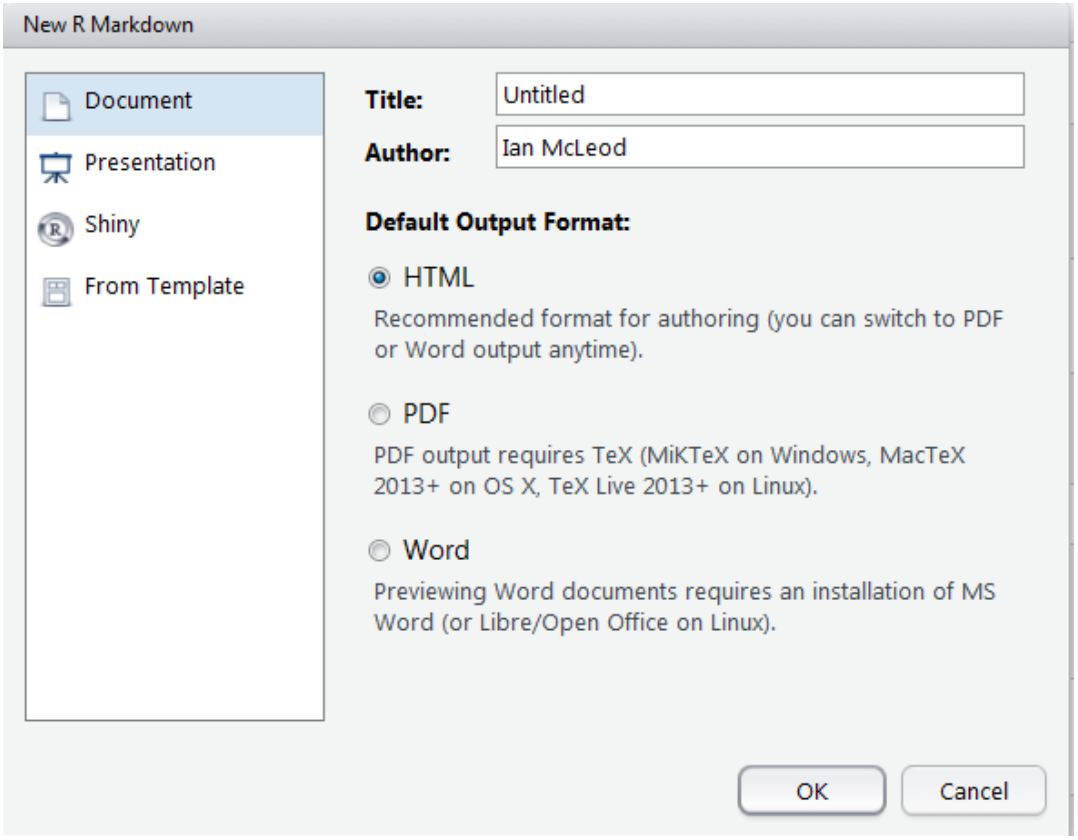
Markdown (see Wikipedia, LINK) is a simple markup language for producing HTML that contains graphics, tables and equations. The original source: http://daringfireball.net/projects/markdown/ LINK. RStudio makes it easy to create HTML output for your statistical analysis. I used this in a recent paper to provide details to the interested reader,

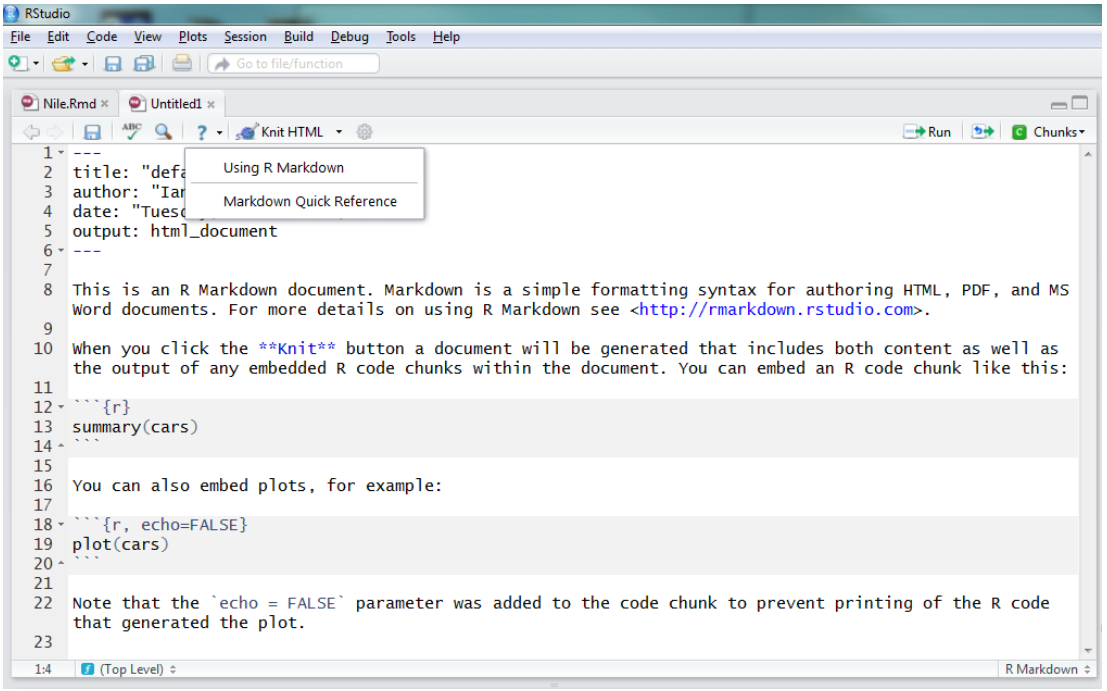
McLeod with 6 others, (October 2014). Road safety impact of Ontario street racing and stunt driving law. Accident Analysis & Prevention. LINK to paper. UWO link.

See the RStudio tutorials on Markdown, https://support.rstudio.com/hc/en-us/sections/200149716-R-Markdown LINK

Getting started

Select from Menu: **`New File  -> R Markdown`**





Examples

In our examples we make extensive use of chunk options, discussed here, http://yihui.name/knitr/options#chunk_options LINK

Some R packages print warnings and other information that you don't want to include in your report, for example the CRAN packages stargazer and wavethresh. To attach these package without producing any messages you can use the following method:

```
```{r LOADstargazer, results='hide', echo=FALSE, warning=FALSE}
#attach stargazer library but suppress all messages
```
```

```
capture.output(suppressMessages(require("stargazer", quietly=TRUE)))
```

Example Rmd files

- Abalone.Rmd
- ShuttleChallenger.Rmd
- Nile.Rmd

This source files are available from my webpage, http://www.stats.uwo.ca/faculty/aim/2014/3859/-Data/ LINK. The output has been published on Rpubs:

- Nile, http://rpubs.com/AIM/40091 LINK
- Abalone, http://rpubs.com/AIM/40093 LINK
- Shuttle Challenger, http://rpubs.com/AIM/40094 LINK

Annual nile riverflow intervention and wavelet analysis

Using the text and data below, I can quickly build a report using markdown.

Text for Rmd file

In this report an analysis of the average annual riverflow, Nile at Aswan, 1870-1945. The intervention analysis for this data was discussed by Hipel et al. (1975) and in the textbook of Hipel and McLeod (1994). The data prior to 1901, corresponding to the first 32 observations are unrelated flows in cms. The data after 1902 are the regulated flows. Both are downstream from the dam. Due to evaporation and percolation we might expect lower annual flows.
Part of the initial data analysis (IDA) is to plot the data to check basic features. R provides many excellent graphics. We use the lattice time series plot since we can easily control the aspect-ratio. It is desirable to choose an aspect-ratio so the average “slope” is about 45 ° or for many stationary time series an aspect-ratio of about 0.25 is a reasonable choice.

The intervention model that we consider may be written,

$$z_t = \mu + \omega S_t + \frac{a_t}{1-\phi(B)}$$

where μ is the overall mean, ω is the step intervention parameter, and ϕ is the AR(1) parameters. It is assumed that $a_t \sim \text{NID}(0, \sigma_a^2)$.

Nason (2008) provides a general introduction to wavelet methods in statistics, including smoothing and multiscale time series analysis. The figure shows the denoised annual Nile riverflows using the universal threshold with hard thresholding and Haar wavelets. The fitted step intervention is represented by the three line segments while the denoised flows are represented by the jaggedcurve.

K.W. Hipel and A.I. McLeod (1994). Time Series Modelling of Water Resources and Environmental Systems. http://www.stats.uwo.ca/faculty/aim/1994Book/default.htm
Hipel, K.W., Lennox, W.C., Unny, T.E. & McLeod, A.I. (1975). Intervention analysis in water resources. Water Resources Research, V.11, pp.855--861.
Guy Nason (2013-10-21) wavethresh: Wavelets statistics and transforms. R package version 4.6.6. http://CRAN.R-project.org/package=wavethresh
Guy Nason (2008). Wavelet Methods in Statistics with R. Springer-Verlag.

LaTeX for Rmd file

```
$$z_t = \mu +\omega  S_t + \frac{a_t}{1-B \phi}$$
$a_t \sim {\rm NID}(0, \sigma_a^2)$
```

R script for Rmd file

```
#install.packages("waveslim")
require("waveslim")
require("lattice")
z<-c(3958.043,3369.694,3485.242,3437.691,3702.352,3817.610
,2875.578,3054.686,4724.150,3834.007,3076.773,2965.759
,3461.708,3141.010,3371.237,2988.425,3607.541,2946.083
,2709.200,3294.848,3556.615,3653.934,3846.064,3713.637
,4252.313,3657.503,3639.370,3197.722,3112.749,2353.684
,2843.652,2194.926,2689.428,2950.906,2247.877,2628.279
,2491.126,2792.630,3321.469,3058.062,2889.853,2495.273
,1648.823,1981.963,2411.072,3035.203,3556.133,3261.959
,2377.893,2394.964,2499.999,2610.242,2743.633,2744.116
,2338.637,2494.984,2474.440,2446.373,2963.059,2732.252
,2205.150,2681.808,2580.535,2954.378,3025.944,2902.777
,2642.457,2860.242,2665.412,2306.905,1848.090,2569.540
,2503.954,2438.753,2211.130)
z <- ts(z, start=1870, freq=1)
#fit step-intervention model
IV<-c(rep(0,32),rep(1,75-32))
out<-arima(x=z, order=c(1,0,0), xreg=IV )
zFit <- coef(out)[2] + IV*coef(out)[3]
lines(as.vector(time(z)), zFit, col="black", lwd=3, lty=1)
#wavelet analysis
wc <- modwt(z, wf = "haar", n.levels = 5, boundary = "periodic")
ws <- universal.thresh.modwt(wc, max.level = 4, hard = TRUE)
zs <- imodwt(ws)
zs <- ts(zs, start=1870, freq=1)
```

```
> xyplot(z, xlab="year", ylab="flow", panel=function(x,y){
+     panel.xyplot(x,y,type="o")
+     panel.grid(h=-1, v=-1, col=rgb(0.5,0.5,0.5,0.5))
+ })
```

```
plot(zs, lwd=3, col="red", ylim=c(1600, 4800), xlab="year", ylab="flow (cms)")
points(as.vector(time(z)), as.vector(z), cex=1, pch=16, col="blue")
```

K.W. Hipel and A.I. McLeod (1994). Time Series Modelling of Water Resources and Environmental Systems. http://www.stats.uwo.ca/faculty/aim/1994Book/default.htm
Hipel, K.W., Lennox, W.C., Unny, T.E. & McLeod, A.I. (1975). Intervention analysis in water resources. Water Resources Research, V.11, pp.855--861.
Guy Nason (2013-10-21) wavethresh: Wavelets statistics and transforms. R package version 4.6.6. http://CRAN.R-project.org/package=wavethresh
Guy Nason (2008). Wavelet Methods in Statistics with R. Springer-Verlag.

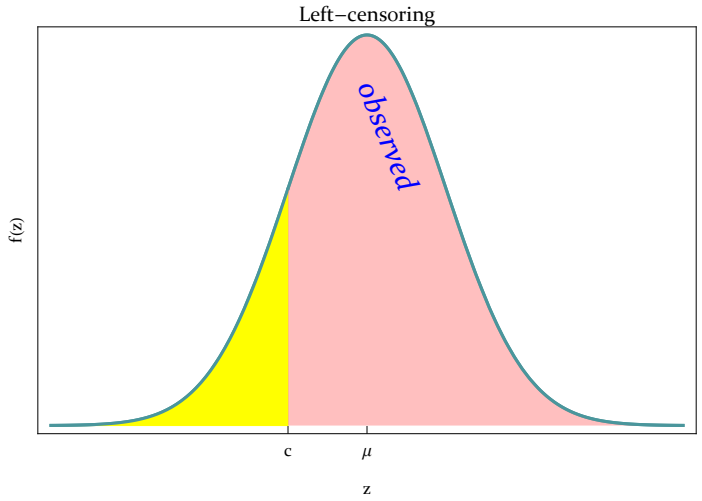
Logistic regression vs LS demo

Demo-LogisticVsLS.nb LINKS: notebook or browser

Sample size computation with censored normal samples

Mohammad, Nagham (2014), Censored Time Series Analysis. University of Western Ontario - Electronic Thesis and Dissertation Repository. Paper 2489. http://ir.lib.uwo.ca/etd/2489

Left-censoring



One-sample problem and time series: data y_t, $t = 1, \ldots, n$ and known censor points c_t, $t = 1, \ldots, n$. Latent process or sample z_t, $t = 1, \ldots, n$ so $y_t = \min(z_t, c_t)$. Simplest case $c_t = c$ and z_t is NID(μ, σ^2). Let $m = \#\{y_t > c\}$ so our sample size of n contains full information on m observations and partial information on $n - m$ obsersvations. The censor rate is $r = (n - m)/n$. If r is not small, statistical inferences ignoring censoring will be misleading.

EM algorithm MLE in censored NID(μ, σ^2)

$\in_{\hat{\mu}^{(j)}, \hat{\sigma}^{(j)}, c}\{Z\}$ denotes the expectaton for a right-truncated normal distribution with truncation point c and parameters $\hat{\mu}^{(j)}$, $\hat{\sigma}^{(j)}$. An explicit expression was obtained using *Mathematica*.

Start with initial estimates $\hat{\mu}^{(0)}$ and $(\hat{\sigma}^2)^{(0)}$ and $j \longleftarrow 0$

1. Compute $\tilde{\mu}_z \longleftarrow \in_{\hat{\mu}^{(j)}, \hat{\sigma}^{(j)}, c}\{Z\}$
2. $\hat{\mu}^{(j+1)} \longleftarrow (m/n)\,\overline{y} + (n - m)/n\,\tilde{\mu}_z$
3. Compute $\tilde{\sigma}_z^2 \longleftarrow \in_{\hat{\mu}^{(j)}, \hat{\sigma}^{(j)}, c}\{(Z - \mu)^2\}$
3. $(\hat{\sigma}^2)^{(j+1)} \longleftarrow n^{-1}\left\{\sum_{i=1}^{m}(y_i - \hat{\mu})^2 + (n - m)\,\tilde{\sigma}_z^2\right\}$
4. Test for convergence of the estimates ...

Implemented in *Mathematica* and R.

Details - via *Mathematica*

Z is a right-truncated $N(\mu, \sigma^2)$

$$\in_{\mu, \sigma, c}\{Z\} = \left(\mu\,\text{erf}\left(\frac{c - \mu}{\sqrt{2}\,\sigma}\right) - \sqrt{\frac{2}{\pi}}\,\sigma\, e^{-\frac{(c-\mu)^2}{2\sigma^2}} + \mu\right) \Big/ \text{erfc}\left(\frac{\mu - c}{\sqrt{2}\,\sigma}\right)$$

$$\in_{\mu, \sigma, c}(Z - \hat{\mu})^2 = \left(-e^{-\frac{(c-\mu)^2}{2\sigma^2}}\sqrt{\frac{2}{\pi}}\,\sigma\,(c + \mu - 2\hat{\mu}) + \left(1 + \text{Erf}\left[\frac{c - \mu}{\sqrt{2}\,\sigma}\right]\right)\left(\mu^2 + \sigma^2 - 2\,\mu\,\hat{\mu} + \hat{\mu}^2\right)\right) \Big/ \text{Erfc}\left[\frac{-c + \mu}{\sqrt{2}\,\sigma}\right]$$

Information matrix in censored normal random samples

The maximum likelihood estimators, $\hat{\mu}$ and $\hat{\sigma}$, have large-sample distribution that is normal with mean $(\mu,\ \sigma)$ and covariance matrix $n^{-1}\,\mathcal{I}_c(\mu,\,\sigma)^{-1}$, where n is the sample size. The (1,1)-entry in $\mathcal{I}_c(\mu,\sigma)$ is given by,

$$\in i_{1,1} = \sigma^{-2} - (1 - \,(1 - \Phi(c;\ \mu,\ \sigma)))\ \sigma^{-2}\ \partial_\mu \in_{\{\mu,\sigma,c\}} \{Z\}. \tag{1}$$

where

$$\partial_\mu \in_{\{\mu,\sigma,c\}} \{Z\} = \frac{1}{\pi\,\sigma\,\text{erfc}\left(\frac{\mu-c}{\sqrt{2}\ \sigma}\right)^2}$$

$$e^{-\frac{(c-\mu)^2}{\sigma^2}}\left(\sqrt{2\,\pi}\ e^{\frac{(c-\mu)^2}{2\,\sigma^2}}\left(\mu\,\text{erf}\left(\frac{c-\mu}{\sqrt{2}\ \sigma}\right) - c\,\text{erfc}\left(\frac{\mu-c}{\sqrt{2}\ \sigma}\right) + \mu\right) + \pi\,\sigma\,e^{\frac{(c-\mu)^2}{\sigma^2}}\left(\text{erf}\left(\frac{c-\mu}{\sqrt{2}\ \sigma}\right)+1\right)^2 - 2\,\sigma\right) \tag{2}$$

Similarly for the (1,2) and (2,2) entries.

Details for the derivation are given in InformationMatrix.nb LINK

Using *Mathematica* to generate C code - see DerivationI11.nb LINK

Sample size in NID samples

We use margin of error in confidence interval inference but details very similar if we consider statistical power in an hypothesis testing approach.

The 95% confidence in a large sample is approximately $\bar{z} \pm \text{MOE}$, where $\text{MOE} = 1.96 \times \sigma \big/ \sqrt{n}$

where n = sample size, σ = normal standard deviation.
Given a preliminary estimate of σ the required sample size to achieve a desired MOE is

$$n\ =\ (1.96 \times \sigma / \text{MOE})^2 \tag{3}$$

A 95% confidence interval implies the corresponding 5% two-sided test has 95% power.

Sample size in censored NID samples

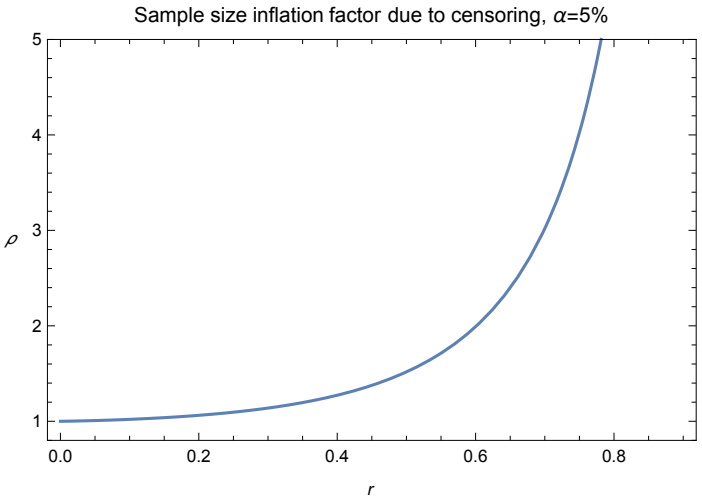
The 95% confidence is approximately $\bar{z} \pm \text{MOE}$, where $\text{MOE} = 1.96 \times \sigma_\mu \big/ n$ where σ_μ is the square-root (1,1) entry $n^{-1}\,\mathcal{I}_c(\mu,\,\sigma)^{-1}$. Given preliminary estimates of μ and σ the required sample size to achieve a desired m is

$$n_c = \left(1.96 \times \sigma_\mu \big/ \text{MOE}\right)^2 \tag{4}$$

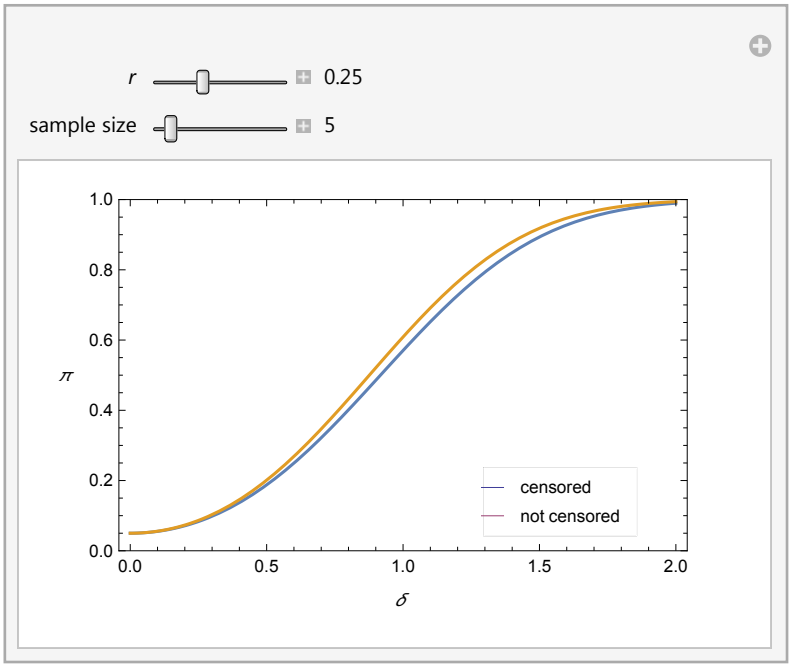
The sample size inflation factor is

$\rho = n_c / n$

ρ does not depend on μ, σ or MOE but only on the censor rate, $r = \Phi^{-1}(c)$.



Power function



Pb concentration in blood of herons in Virgina

Data discussed in Helsel and available in R package NADA. Units microgram/gram.

Left-censored at 0.02. With $n = 27$, $m = 12$ so $r = 15/27 \approx 56\ \%$.

The MLE of the mean and sd are respectively 0.03779965 and 0.09449939.

Our software gives $\sigma_\mu = 0.108616$ where σ_μ is the square-root (1,1) entry $\mathcal{I}_c(\mu, \sigma)^{-1}$

To estimate the mean with MOE = 0.02 would require

$n_c = \left(1.96 \times \sigma_\mu / 0.02\right)^2 = 113.302$