Reading SAS files

SAS stores its files in a number of different formats, depending on the platform (PC, Unix, etc.), version, etc. The `read.xport` function can read the SAS transport (XPORT) format. How you produce one of these files depends on which version of SAS you're using. In older versions, there was a `PROC XPORT`. Newer versions use a "library engine" called `sasv5xpt` to create them. From the SAS technical note referenced in the `?read.xport` help page, a line like

```
libname xxx sasv5xpt 'xxx.dat';
```

creates a library named `xxx` which lives in the physical file 'xxx.dat'. New datasets created in this library, e.g. with

```
data xxx.abc;
  set otherdata;
```

will be saved in the XPORT format.

To read them, use the `read.xport` function. For example, `read.xport('xxx.dat')` will return a list of data frames, one per dataset in the exported library. If there is only a single dataset, it will be returned directly, not in a list. The `lookup.xport` function gives information about the contents of the library.

Reading Minitab files

Minitab also stores its data files in several different formats. The **foreign** package can only read the "portable" format, the MTP files. In its current incarnation, only numeric data types are supported; in particular, dataset descriptions cannot be read.

The `read.mtp` function is used to read the files. It puts each column, matrix or constant into a separate component of a list.

Reading SPSS and Stata files

The **foreign** functions `read.spss` and `read.dta` can read the binary file formats of the SPSS and Stata packages respectively. The former reads data into a list, while the latter reads it into a data frame. The `write.dta` function is unique in the **foreign** package in being able to *export* data in the Stata binary format.

Caveats and conclusions

It's likely that all of the functions in the **foreign** package have limitations, but only some of them are documented. It's best to be very careful when transferring data from one package to another. If you can, use two different methods of transfer and compare the results; calculate summary statistics before and after the transfer; do anything you can to ensure that the data that arrives in R is the data that left the other package. If it's not, you'll be analyzing programming bugs instead of the data that you want to see.

But this is true of any data entry exercise: errors introduced in data entry aren't of much interest to your client!

Duncan Murdoch
University of Western Ontario
murdoch@stats.uwo.ca

Maximally Selected Rank Statistics in R

by Torsten Hothorn and Berthold Lausen

Introduction

The determination of two groups of observations with respect to a simple cutpoint of a predictor is a common problem in medical statistics. For example, the distinction of a low and high risk group of patients is of special interest. The selection of a cutpoint in the predictor leads to a multiple testing problem, cf. Figure 1. This has to be taken into account when the effect of the selected cutpoint is evaluated. Maximally selected rank statistics can be used for estimation as well as evaluation of a simple cutpoint model. We show how this problems can be treated with the **maxstat** package and illustrate the usage of the package by gene expression profiling data.

Maximally selected rank statistics

The functional relationship between a quantitative or ordered predictor X and a quantitative, ordered or censored response Y is unknown. As a simple model one can assume that an unknown cutpoint μ in X determines two groups of observations regarding the response Y: the first group with X-values less or equal μ and the second group with X-values greater μ. A measure of the difference between two groups with respect to μ is the absolute value of an appropriate standardized two-sample linear rank statistic of the responses. We give a short overview and follow the notation in Lausen and Schumacher (1992).

The hypothesis of independence of X and Y can be formulated as

$$H_0 : P(Y \leq y | X \leq \mu) = P(Y \leq y | X > \mu)$$

for all y and $\mu \in \mathbb{R}$. This hypothesis can be tested as follows. For every reasonable cutpoint μ in X (e.g., cutpoints that provide a reasonable sample size in both groups), the absolute value of the standardized two-sample linear rank statistic $|S_\mu|$ is computed. The maximum of the standardized statistics

$$M = \max_\mu |S_\mu|$$

of all possible cutpoints is used as a test statistic for the hypothesis of independence above. The cutpoint in X that provides the best separation of the responses into two groups, i.e., where the standardized statistics take their maximum, is used as an estimate of the unknown cutpoint.

Several approximations for the distribution of the maximum of the standardized statistics S_μ have been suggested. Lausen and Schumacher (1992) show that the limiting distribution is the distribution of the supremum of the absolute value of a standardized Brownian bridge and consequently the approximation of Miller and Siegmund (1982) can be used. An approximation based on an improved Bonferroni inequality is given by Lausen et al. (1994). For small sample sizes, Hothorn and Lausen (2001) derive an lower bound on the distribution function based on the exact distribution of simple linear rank statistics. The algorithm by Streitberg and Röhmel (1986) is used for the computations. The exact distribution of a maximally selected Gauß statistic can be computed using the algorithms by Genz (1992). Because simple linear rank statistics are asymptotically normal, the results can be applied to approximate the distribution of maximally selected rank statistics (see Hothorn and Lausen, 2001).

The maxstat package

The package **maxstat** implements both cutpoint estimation and the test procedure above with several P-value approximations as well as plotting of the empirical process of the standardized statistics. It depends on the packages **exactRankTests** for the computation of the distribution of linear rank statistics (Hothorn, 2001) and **mvtnorm** for the computation of the multivariate normal distribution (Hothorn et al., 2001). All packages are available at CRAN. The generic method `maxstat.test` provides a formula interface for the specification of predictor and response. An object of class `"maxtest"` is returned. The methods `print.maxtest` and `plot.maxtest` are available for inspection of the results.

Gene expression profiling

The distinction of two types of diffuse large B-cell lymphoma by gene expression profiling is studied by Alizadeh et al. (2000). Hothorn and Lausen (2001) suggest the mean gene expression (MGE) as quantitative factor for the discrimination between two groups of patients with respect to overall survival time. The dataset `DLBCL` is included in the package. The maximally selected log-rank statistic for cutpoints between the 10% and 90% quantile of MGE using the upper bound of the P-value by Hothorn and Lausen (2001) can be computed by

```
> data(DLBCL)
> maxstat.test(Surv(time, cens) ~ MGE,
               data=DLBCL, smethod="LogRank",
               pmethod="HL")

        LogRank using HL

data:  Surv(time, cens) by MGE
M = 3.171, p-value = 0.02220
sample estimates:
estimated cutpoint
         0.1860526
```

For censored responses, the formula interface is similar to the one used in package **survival**: `time` specifies the time until an event and `cens` is the status indicator (`dead=1`). For quantitative responses y, the formula is of the form '`y ~ x`'. Currently it is not possible to specify more than one predictor `x`. `smethod` allows the selection of the statistics to be used: `Gauss`, `Wilcoxon`, `Median`, `NormalQuantil` and `LogRank` are available. `pmethod` defines which kind of P-value approximation is computed: `Lau92` means the limiting distribution, `Lau94` the approximation based on the improved Bonferroni inequality, `exactGauss` the distribution of a maximally selected Gauß statistic and `HL` is the upper bound of the P-value by Hothorn and Lausen (2001). All implemented approximations are known to be conservative and therefore their minimum P-value is available by choosing `pmethod="min"`.

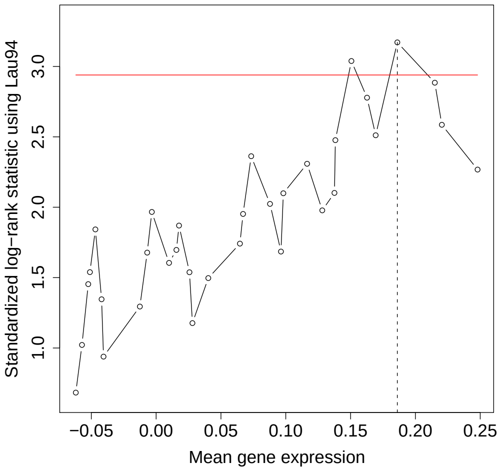

Figure 1: Absolute standardized log-rank statistics and significance bound based on the improved Bonferroni inequality.

For the overall survival time, the estimated cutpoint is 0.186 mean gene expression, the maximum of the log-rank statistics is $M = 3.171$. The probability that, under the null hypothesis, the maximally selected log-rank statistic is greater $M = 3.171$ is less then than 0.022. The empirical process of the standardized statistics together with the α-quantile of the null distribution can be plotted using `plot.maxtest`.

```
> data(DLBCL)
> mod <-
    maxstat.test(Surv(time, cens) ~ MGE,
                 data=DLBCL, smethod="LogRank",
                 pmethod="Lau94", alpha=0.05)
> plot(mod, xlab="Mean gene expression")
```

If the significance level `alpha` is specified, the corresponding quantile is computed and drawn as a horizonal red line. The estimated cutpoint is plotted as vertical dashed line, see Figure 1.

The difference in overall survival time between the two groups determined by a cutpoint of 0.186 mean gene expression is plotted in Figure 2. No event was observed for patients with mean gene expression greater 0.186.

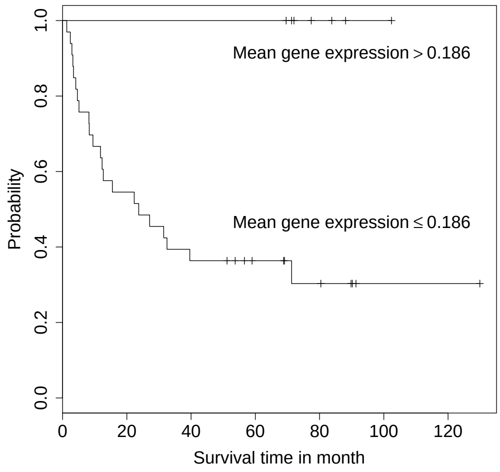

Figure 2: Kaplan-Meier curves of two groups of DLBCL patients separated by the cutpoint 0.186 mean gene expression.

Summary

The package **maxstat** provides a user-friendly interface and implements standard methods as well as recent suggestions for the approximation of the null distribution of maximally selected rank statistics.

Acknowledgement. Support from Deutsche Forschungsgemeinschaft SFB 539-A4 is gratefully acknowledged.

Bibliography

Ash A. Alizadeh, Michael B. Eisen, and Davis, R. E. et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000. 4

Alan Genz. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1:141–149, 1992. 4

Torsten Hothorn. On exact rank tests in R. *R News*, 1 (1):11–12, 2001. 4

Torsten Hothorn, Frank Bretz, and Alan Genz. On multivariate *t* and Gauss probabilities in R. *R News*, 1(2):27–29, 2001. 4

Torsten Hothorn and Berthold Lausen. On the exact distribution of maximally selected rank statistics. *Preprint, Universität Erlangen-Nürnberg, submitted*, 2001. 4

Berthold Lausen, Wilhelm Sauerbrei, and Martin Schumacher. Classification and regression trees (CART) used for the exploration of prognostic factors measured on different scales. In P. Dirschedl and R. Ostermann, editors, *Computational Statistics*, pages 483–496, Heidelberg, 1994. Physica-Verlag. 4

Berthold Lausen and Martin Schumacher. Maximally selected rank statistics. *Biometrics*, 48:73–85, 1992. 3, 4

Rupert Miller and David Siegmund. Maximally selected chi square statistics. *Biometrics*, 38:1011–1016, 1982. 4

Bernd Streitberg and Joachim Röhmel. Exact distributions for permutations and rank tests: An introduction to some recently published algorithms. *Statistical Software Newsletters*, 12(1):10–17, 1986. 4

Friedrich-Alexander-Universität Erlangen-Nürnberg,
Institut für Medizininformatik, Biometrie und
Epidemiologie, Waldstraße 6, D-91054 Erlangen
Torsten.Hothorn@rzmail.uni-erlangen.de
Berthold.Lausen@rzmail.uni-erlangen.de