MATHEMATICAL INSTITUTE

INTERNSHIP REPORT

STATISTICAL SCIENCE FOR THE LIFE AND BEHAVIOURAL SCIENCES

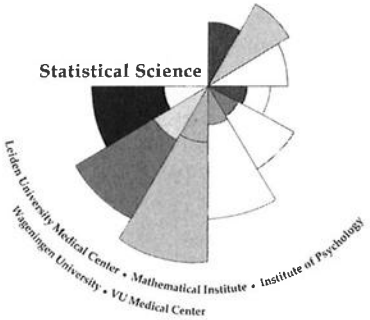
Improving and extending the R package "quint"

Author:
Jeanne van de Put, s0924954

Internal Supervisor:
Dr. Elise Dusseldorp
Affiliation:
Associate Professor, Institute of Psychology; Guest Lecturer at the Mathematical Institute

April, 2016

Universiteit Leiden
The Netherlands



Introduction

The chosen internship for the master Statistical Science had as main focus improving the R package `quint` 1.0 (Dusseldorp et al., 2013). This is a package for performing analyses with the QUalitative INteraction Trees (QUINT) method (Dusseldorp & van Mechelen, 2014; Dusseldorp et al., 2015). The internship was carried out at the Social Sciences Faculty of Leiden University under supervision of Dr. Elise Dusseldorp. Secondary tasks of the internship comprised of performing consulting sessions for master students of this faculty who are working on their master thesis research. A short description of the QUINT method will follow as introduction for the internship report.

QUINT is an unsupervised learning method based on recursive partitioning. It tries to find qualitative treatment-subgroup interactions in randomized controlled trial (RCT) data. Qualitative interactions regard subgroups of patients for whom one treatment A works better than B and subgroups for whom treatment B works better than A. Within QUINT a final subgroup can be distinguished for whom the treatment does not matter. These subgroups are referred to P1, P2 and P3 respectively. QUINT grows a binary tree with which it constructs the arrangement of subjects into the subgroup classes. The input model for QUINT comprises a continuous outcome variable, a treatment variable with two categories and a set of (baseline) variables. The latter set of variables include demographic variables and moderator variables measured at baseline (i.e., patient characteristics). Partitioning is done with splitting variables, their optimal thresholds and the partitioning criterion C. C should reflect the strength of a treatment-subgroup interaction given a certain variable and its best split point. C is determined by the difference in treatment outcome for treatments A and B within subgroups P1 and P2 and the number of patients in P1 and P2. With small values of either or both of these measures the procedure cannot deduce a treatment-subgroup interaction of practical significance. The splitting variable with split point is chosen if it maximizes C over all other split point possibilities. These three elements comprise an 'optimal tripet'. The split is only allowed when the partitioning criterion C found for the next step exceeds C of the current step. QUINT stops when this optimal triplet is not found, a larger C value for a next step cannot be found or when a maximum number of leaves (i.e., tree size) is reached. The end nodes (i.e., the leaves) represent a number of patients per treatment group from the data set assigned to the subgroups P1, P2 or P3. The amount of patients per treatment group, the difference in treatment effect and corresponding standard error is given as leaf information as well. The tree can be used for assigning treatments to patients based on the partitioning variables and corresponding split points. The analysis might also generate new hypotheses for follow-up research. Pruning of a QUINT tree to an optimal size is done based on a bias correction procedure using bootstrap samples. (Dusseldorp & van Mechelen, 2014; Dusseldorp et al., 2015).

Site Description

The internship was done at the Methodology & Statistics department of the Faculty of Social Sciences (FSW) from Leiden University. Hence this site is an academic environment. The staff mainly includes researchers, teachers and teaching assistants. The (researchers of the) department is/are specialised in statistical methods including e.g., fMRI or high dimensional data analysis, multidimensional scaling or unfolding, psychometrics and various statistical learning methods. The research is mainly focussed on applications within psychology. The department is also in charge the statistics courses for the Bachelor track Psychology and the Master track specialisation Methodology and Statistics in Psychology. Some researchers are guest lecturers at the mathematical institute for the Statistical Science Master track. The teaching staff is quite big and ranges from students assisting with work groups to people working full-time as university teachers. Furthermore the department has some employees who program their methods for IBM SPSS.

Overall the FSW has more students and employees. A major difference is that this faculty has its own statistics department, while the Mathematical Institute does not have this. There are also many other departments within the FSW where of course a lot of research for various disciplines within social sciences is carried out. A big difference between the Methodology & Statistics department in which the internship was done and the Mathematical Institute is that statistics research done at this department is more focussed on

a specific area of application, in this case psychology. Also I have noticed that many people are especially appointed for teaching as their main jobs while in e.g., Statistical Science all teachers had to teach aside of their own work or research. Furthermore the FSW offers more services to its students, such as career counselling or the 'ScriptieAtelier' to help students who are writing their thesis. The ScriptieAtelier includes various services for students: it offers desks or work places besides the library, but it is also possible to receiving help on literature search, statistical counselling or other general help for writing a thesis. At the Mathematical Institute however students often have to work independently. For me it was good to do an internship in the social sciences building to experience working in a different setting with a different atmosphere and to get in touch with all kinds of new people working on something else to what I was used to.

Performed tasks during internship

Primary activities

The main goal of the internship was to improve version 1.0 of the R-package `quint` for the method QUINT and upload it to CRAN. The package has been upgraded to version 1.2. The most important differences between versions 1.0 and 1.2 are the ability to support categorical variables with more than two levels, the possibility to make predictions on (new) data with a quint object and addition of a function which estimates the bias of a (pruned) QUINT tree, from which the generalizability of the tree can be deduced. Previously the QUINT method only supported numerical variables or dichotomous categorical variables. Moreover the prediction and validation procedures were only described in papers (e.g. Dusseldorp & van Mechelen, 2014; Formanoy et al., submitted) and available in the internal package version 1.1 but there were no functions for these procedures available yet for users. Besides these major changes there were some bugs in version 1.0 that needed repairing and options for output display of the quint objects to be elaborated.

The first week comprised reading literature on recursive trees and the method QUINT. The literature included the main reference book on Classification and Regression Trees (CART; Breiman et al., 1984) and a technical description of the QUINT algorithm/method including a simulation study assessing its performance (Dusseldorp & van Mechelen, 2014). Also some papers focussing more on the application of QUINT were studied, such as a paper on the usage of the R package QUINT on data from a randomized controlled trial (Dusseldorp et al., 2015) and a QUINT analysis explanation specifically aimed at psychologists (Doove et al., 2015). The **References** section contains a complete overview of all literature. This period also included experimenting with version 1.0 of the QUINT package. For example, the analysis of the Dusseldorp et al. (2015) on breast cancer recovery patients was replicated. These data were also analysed with CART using the `rpart` package (Therneau et al., 2015), following an analysis procedure described in Faraway (2006). This was done to compare how the two different decision tree procedures are analysed in R. Another more theoretical reason to apply two methods on the same data set was to make a comparison between the algorithms, regarding parameters and criteria used for growing the tree, to compare the usage of the functions in R or to compare the output and interpretation of grown trees.

The focus of the second week was on the construction of R packages in general. This included reading a book on R packages (Wickham, 2015) and experimenting with examples from this book. Wickham describes in his book how an R package can be made with the package `devtools` (Wickham & Chang, 2015). An example R package with code from a previous course was built as a preparation for (re)building the existing `quint` package (which was previously built with the `package.skeleton()` function).

In the third week the `quint` package was (re)constructed with the existing functions. Also the first versions of new or improved (i.e., adjusted to support categorical variables) functions for this package were embedded. Some files were arranged in a slightly different way within the package such that related (sub)functions were clustered together. The help files of the old package were copied so the main functions could be exported to to the NAMESPACE. Then this package version was built in R to prepare for testing.

During the two subsequent weeks the contents of the R package were tested, debugged and adjusted where necessary. An overview of all changes and scripts with analyses is given in the **Appendix** (sections A and

B). First analyses by Dusseldorp et al. (2015) were reproduced to check whether the all embedded adjusted functions of the package functioned properly. Although most runs produced (almost) identical results a few functions had some bugs. Some debugging procedures were performed to overcome these minor problems. To confirm whether the method could support categorical data with more than two levels the test package was used to reproduce the analysis on a drug use data set by Doove et al. (2015). Also the model formulas used for fitting a tree by Doove et al. were adjusted to force the tree to make splits based on mainly categorical variables. Most (sub) functions (re)produced the right results, although upon inspection the output of a quint object had to be adjusted such that it would display the actual split points levels (this was originally represented by an integer indexing something within the internal code). Furthermore the plot function had to be extended to display categorical split levels at the correct branches of the nodes.

In the sixth week the newly introduced validate and predict functions were tested and improved. The validate function has been used in Formanoy et al. (submitted) to analyse data on work stress recovery. This analysis has been reproduced to inspect the utilization of the function and to check whether it functions properly. The output objects of the validate function have been shortened to show by default the most relevant output. Furthermore most input arguments are left out, leaving only the input quint object and the number of bootstrap samples as arguments. This makes the application of the function more user-friendly (i.e., the user does not have to specify a lot of values). The code has been adjusted accordingly to compensate for the deletion of input arguments: most required values are now directly extracted from the input quint object. The predict function was tested on the breast cancer recovery and drug use dataset. The predict function has been adjusted such that it can give two variations on outputs. The outputs can either be predicted treatment subgroup classes per patient or the position of the patient within the quint tree (leaf and corresponding node number). Besides these changes to the validate and predict functions their help files with descriptions and examples were written from scratch.

Final result

The final week constituted of performing the final checks before uploading the upgraded `quint` package to CRAN. A series of checks have been performed in the R console. Also a NEWS file was written for the package to document the changes made for `quint` 1.2 with respect to version 1.0. Besides these final preparations the internship report was written. On the final day of the internship the package has been submitted to CRAN and it has been received. The final product can (soon) be downloaded via this link or in R (Studio): https://cran.r-project.org/web/packages/quint/index.html

Secondary activities

The secondary activities, concerning statistical consulting sessions, will be described in the **Project Example** section below.

Project Example

The R package improvement has been quite a large project that took up most time. However this was not the only task. The secondary goal of this internship was to practice statistical consulting. As a project example I would like to summarize the process of the consulting sessions and some (new) topics that were discussed.

As mentioned before the ScriptieAtelier offers more specific help on statistical analyses through consulting hours for students. One of the employees providing this service on a weekly basis is Dr. Elise Dusseldorp. In one afternoon about three students can come by for a session for 30 or 45 minutes. In the first three weeks I observed sessions done by Dr. Dusseldorp and after three weeks I gradually started leading consulting sessions myself together with or under supervision of Dr. Dusseldorp. A session starts by asking the student for a small introduction to the research. Some questions on details might follow, regarding e.g., the study design, outcome variable, predictors, measures or scales of the predictors. Then the student explains his/her

problem, after which the help is given. Sometimes a student needed some advice on how to analyse his/her data. Sometimes the student just needed additional explanation or help with interpretation of analyses already performed. Analyses were usually done in SPSS. In most cases the student got help on more than just the main problem but also with some other things like computing other supporting statistics, help with diagnostics, demonstration the usage of syntax in SPSS, recoding or centering variables and suggestions for further analyses. A session usually ended when either the problem was solved or when the time was up.

The data of the students mainly concerned randomized (clinical) trials, cohort studies, case-control studies or surveys. Although these types of datasets have been discussed in some Statistical Science courses, I did not have a lot of experience with most of these dataset types. Furthermore some statistical methods used for analysing psychology-related data were not really familiar to me. Hence I sometimes had to prepare a session by reading some summaries on these topics. One example of a relatively unfamiliar method encountered is the MANOVA (multivariate ANOVA). Instead of several separate ANOVA analyses (with a multiple testing correction) for every outcome it is possible to test all outcomes at once with this test. The output is extended to tables summarizing if there is any significant difference between groups given the multiple outcomes (the multivariate test), a separate ANOVA table summarizing for which outcomes the groups differ (between-subject effects) and a final table with multiple comparisons of group differences per outcome. Another example of a newly encountered method is hierarchical linear regression. The most important set of variables is specified in the upper level and its effects on the outcome are computed after correcting for the remaining variables in the lower levels. Furthermore I have learned that effect sizes can be used to select the most relevant interactions or variables in a model with many observations (e.g., $N > 10,000$) and many variables. Also I have become more familiar with moderator testing. So not only did the student learn some new things during the sessions but I did as well.

Acquired and strengthened skills

Among the first things to mention in this section is that I have gained more knowledge on decision tree methods. Although the focus was on QUINT, I have also studied CART and compared these two methods. The former is an unsupervised learning method and the latter a supervised learning method. Other important differences between the two algorithms concern their growing or stopping criteria, parameters to be specified, their pruning algorithms and the interpretation of the final trees. Another new procedure I have learned about is bootstrap-based validation (applied in QUINT) for estimating bias as an alternative to cross-validation (used in e.g., CART) for estimating bias. Furthermore I have compared the way how the input formulas for these methods should be specified and how these methods can be applied in R. There was another added value of studying decision tree algorithms during my internship: it has prepared me to some extent for my thesis. This thesis will be with be about CART and Random Forests, supervised by Elise Dusseldorp and Mark Bouts (both from the Social Sciences faculty).

Apart from papers I also learned about QUINT by studying the way it was programmed in the R package and how it is presented to the user. During the Statistical Science master we have mainly learned about some of the mathematics behind statistical methods and how methods are applied. We have also learned how to do statistical or mathematical computations in R. But learning how an entire statistical method can actually be programmed was a new experience. It gave me a lot of insight in the way a statistical method is 'translated' from theory to an algorithm and code.

Obviously generally strengthened and acquired skills of this internship are respectively programming and constructing R packages. This will also be a useful skill for in the future, if it ever becomes necessary to program a (new) method in R to share it with others. Also, I have gained more experience with debugging codes and solving problems. I have also learned how to program functions that are used by others in such a way that the usage is simple and the output returned is interpretable. Newly writing the help files for the two new main functions helped me learn how to present functions in an understandable way to the user.

Other than learning about decision tree methods and programming methods in R I have gained more knowledge about analyses of Randomized Controlled Trial (RCT) data and treatment effects. Treatment-subgroup interactions were a relatively unknown concept for me, let alone the distinction between quantitative and

qualitative treatment-subgroup interactions. Before starting this internship analyses with RCT data would tempt me to focus mainly on the overall treatment effect and to conclude whether or not the treatment it is better than an alternative. However inference regarding the treatment effect would then mainly be on the averages within the entire group. This might mask the possibility that the treatment works better or does not work well at all for certain subgroups of patients. Also I would be less likely to consider whether the alternative treatment might actually benefit some people. QUINT is a tool which could be a useful addition to a straightforward formal analysis like e.g., linear regression. This would allow for more detailed inference and could generate new ideas for future trials.

Although it was a necessity to read research papers I certainly regard studying them as a valuable asset, especially since hardly any research papers are discussed or read during the Statistical Science track. This is particularly useful for people who are considering a career in academia. I think this is important to read scientific papers as they link (the application of) statistics with scientific disciplines. It helps to develop some sense of reading, interpreting, reporting and communicating research. Also research papers might inspire to apply certain statistical methods in future analyses. Moreover learning how to read papers is important to understand novel or specific theory, since mainly the basics are taught in lectures. Other than all new theory studied and programming skills gained I have gained useful experiences with the statistical consulting sessions regarding for example solving various problems, applying various statistical methods and communicating with others. Helping with the consulting sessions has given me the chance to discover research data of yet another application field, namely psychology. I am not really familiar with this area, as I have a background in biology and I did not follow the Psychometrics course. Although I did have some experience with psychology-related data from other Statistical Science courses, this experience has been extended. I have seen various types of datasets and many kinds of statistical methods as tools for their analyses. Some statistical methods were familiar to me and some were not. These sessions also helped me to listen to students and to reformulate and understand their problems. The sessions have trained me to solve problems in a relatively short time and to switch quickly from one subject to the next between sessions. This all had to be done keeping in mind that everything should be explained in such a way that the student understands. One final skill strengthened is having more experience in SPSS. Most students from the Social Science faculty are only familiar with this type of software. As I mainly work in R it is important to know more about performing analyses in SPSS to stay flexible as a statistical consultant for coming towards the students.

Conclusion

My internship had a strong focus on programming in R and constructing R packages. I have become more skilful with programming, testing and solving problems in the code. Other than programming it was also important to adapt the new main functions and write their help files in such a way that they are relatively user-friendly and their output is interpretable. For me personally it was interesting to have an internship with a relatively strong focus on programming; it was something I enjoyed working on. It was an excellent way to apply many things I have learned during the programming courses of the Statistical Science master.

Besides these practical assets I have learned about a method with which I hardly had any prior experience. In the Statistical Learning course decision trees and CART have been discussed yet QUINT differs from what I have learned previously about a common tree algorithm like CART. Now I have seen how a basic concept like decision trees can be extensively altered with the aim of exploring something more specific than just an outcome variable with a set of predictors. Other than studying QUINT with papers I have been able to study the algorithm in depth by working with the R package. It made me realize that even though the user only sees the 'surface' of the method and function in papers or software, programming an entire statistical method is really complex. Although programming the method already comprises a great part of a job in developing a new method, the a part of the remaining challenges is to think of how to present it to scientists in e.g., the help files of the package and in publications.

The statistical consulting sessions have been really valuable in gaining experience with helping people, solving their problems and to learn about applying statistics in a field of research that was relatively unknown to me. This internship at the Social Sciences faculty has made me more familiar with terminology and methods

applied in this field of research. It has also sparked my interest in Social Sciences. My skills in communicating statistics has been strengthened, just as being flexible in working with various people or statistical problems as well as working with different statistical software. I might also enjoy doing a job including some statistical consulting in the future.

References

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J., 1984. Classification And Regression Trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Doove, L.L., van Deun, K., Dusseldorp, E. and Van Mechelen, I., 2015. QUINT: A tool to detect qualitative treatment-subgroup interactions in randomized controlled trials.

Dusseldorp, E., Doove, L. L., and Van Mechelen, I., 2013. Quint: Qualitative interaction trees. R Package Version 1.0. Retrieved from http://cran.r-project.org/package=quint

Dusseldorp, E., Doove, L., and Van Mechelen, I. Quint: An R package for the identification of subgroups of clients who differ in which treatment alternative is best for them. Behavior research methods (ahead-of-print), 1-14. DOI 10.3758/s13428-015-0594-z.

Dusseldorp, E., and Van Mechelen, I., 2014. Qualitative interaction trees: A tool to identify qualitative treatment-subgroup interactions. Statistics in Medicine, 33, 219-237. doi:10.1002/sim.5933

Faraway, J.J., 2006. Extending the Linear Model with R. London, UK: Chapman & Hall.

Formanoy, M.A.G., Dusseldorp, E., Coffeng, J.K., van Mechelen, I., Boot, C.R.L., Hendriksen, I.J.M., and Tak, E.C.P.M. (submitted). Physical activity and relaxation in the work setting to reduce the Need for Recovery: what works for whom?

Therneau, T., Atkinson, B. and Ripley, B., 2015. rpart: Recursive Partitioning and Regression Trees. R Package Version 4.1-10. Retrieved from https://cran.r-project.org/web/packages/rpart/index.html

Wickham, H., 2015. R Packages. Sebastopol, USA: O'Reilly Media, Inc.

Wickham, H. and Chang, W., 2016. Devtools: Tools to make Developing R Packages Easier. R Package Version 1.11.0. Retrieved from https://cran.r-project.org/web/packages/devtools/index.html