DEVELOPING INTERACTIVE EBOOKS AND AN ANALYSIS ASSISTANT TO TEACH AND APPLY MODERN QUANTITATIVE METHODS

Richard M. A. Parker[1], Danius T. Michaelides[2], Christopher M. J. Charlton[1],
Luc Moreau[2] and William J. Browne[1]
[1] University of Bristol, United Kingdom
[2] University of Southampton, United Kingdom
richard.parker@bristol.ac.uk

Whilst a large range of valuable training resources are available to those interested in learning quantitative techniques, few discuss how statistics are conducted in practice by working analysts. Advances in technology, however, have opened up the possibility of using more interactive tools to develop such resources. We have conducted interviews with quantitative researchers from a variety of disciplines, and are collaborating with them to produce interactive eBooks based on a case study each of them have chosen. These are written using our Stat-JR package which can interoperate with a variety of other statistical software, and can thus allow users to explore how a given analysis might be conducted in a range of different packages. The resulting eBooks will form a library of case studies that those newer to the field can use as a learning tool, with the aim of elucidating and demystifying the quantitative research process.

BACKGROUND

The ability to understand and use quantitative methods has long been a valuable skill for researchers in a large range of applied disciplines. Indeed, the increasing availability of data has brought with it even more opportunities for understanding and exploration for those with the expertise to realise them. A large range of valuable training resources exist to support students and other researchers who wish to develop their statistical techniques and knowledge, but whilst many cover statistical concepts in detail, few discuss how statistics are conducted in practice by working analysts (e.g. Nolan & Lang, 2007).

Apprentices learning a new trade, such as plumbing, will often shadow an experienced practitioner alongside formally taught training. Most would agree that this is a sensible strategy when learning such a practical profession: if the apprentice were solely taught in the lecture theatre, he or she would soon become unstuck when trying to diagnose and resolve problems in a real house with its mix of old and newer pipes, curious angles, obsolete appliances, and so on. Whilst statistical analysts are not working in such a manual context, their everyday decisions also benefit from the experience and intuitions built up over the course of their career. However, unlike a plumber, the detail of a statistical analyst's everyday decisions and other practical aspects of their work are much less accessible.

Typically, much of the detail of the journey taken from a study's conception to its conclusion is lost from a manuscript by the time it is accepted for publication by a journal. Whilst this results in a more succinct article, the polished end product belies the challenges and difficult decisions made by research teams *en route*; it risks giving the impression that the analysts engaged in a somewhat mystical process, one which found the one true path to the perfect analytical outcome. Anyone who has analysed even modestly-challenging data, however, knows that in practice this is not the case.

Statistical analysts often have to take difficult decisions which have no obviously 'correct' resolution; there may have been alternative analytical avenues they may have reasonably taken instead; they might uncover mistakes late in the analytical process and have to backtrack and re-think; they might have to engage with secondary data which has uncertain provenance, or which was coded using flexible or unclear criteria; and so on. Engaging with everyday experiences such as these could help those learning about statistical analysis understand more deeply that 'real' data is often challenging, and that they too will encounter difficulties and doubts when engaging with it, but that with experience and engaged rational judgment they can nevertheless resolve many of these difficulties, or at least find a constructive path through.

How can such details be best presented and communicated? For obvious historical reasons, textbooks have been the predominant prescription for those wishing to learn more about

In: M.A. Sorto (Ed.), *Advances in statistics education: developments, experiences and assessments. Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),* July 2015, Rio de Janeiro, Brazil. iase-web.org/Conference_Proceedings.php

quantitative research concepts and techniques. Whilst such resources are hugely valuable, their linear and static nature may not, however, be best suited to allowing the user to more deeply explore the context in which an analysis is conducted. With technological advances over recent years has come considerable interest in the use of interactive tools to teach statistical theory and skills, with many valuable resources developed, and the field continuing to rapidly evolve (e.g. Chance et al. 2007). Furthermore, there is the potential for non-linear, interactive environments to provide the infrastructure which will allow users to engage more fruitfully with the decisions a quantitative research team take during a study, and the multiplicity of possible avenues which are open to them.

With this objective in mind, in this study we interview a range of researchers, asking each to choose a quantitative research project they have worked on, and try to capture the decisions they take and the context in which their research is undertaken. We are then working with them to translate their case study into an interactive eBook using our statistical software package Stat-JR. As such, we plan to create a library of case studies that those newer to the field can use as a learning tool, with the aim of elucidating and demystifying the quantitative research process.

STAT-JR

Stat-JR (Charlton et al., 2014) was written by a collaborative team from the Centre for Multilevel Modelling at the University of Bristol and Electronics and Computer Science at the University of Southampton, with funding from the UK's Economic and Social Research Council (ESRC). It consists of a modular system of templates, each defining a certain function (or suite of functions). Users choose a template to use in conjunction with their dataset of interest: some templates fit models, others plot charts or produce data summaries, and so on. Stat-JR is primarily written in the Python language (Python Software Foundation, 2010), and it can perform statistical operations in-house, using its own 'eStat' MCMC engine or via functions provided by the scientific Python packages, but it can also interoperate with third-party statistical software packages, including R (R Core Team, 2014), Stata (StataCorp, 2013), MLwiN (Rasbash et al., 2009), WinBUGS (Lunn et al., 2000), OpenBUGS (Thomas et al., 2006), SPSS (IBM Corp., 2013) and SAS (SAS Institute Inc., 2013). Depending on the template executed, Stat-JR's outputs include results tables, graphs (as scalable vector graphics), equations (as snippets of LaTeX), scripts, macros, point-and-click instructions (detailing how to perform the operation via direct operation of another package's graphical user interface (GUI)), and so on. See Figure 1 for an overview.

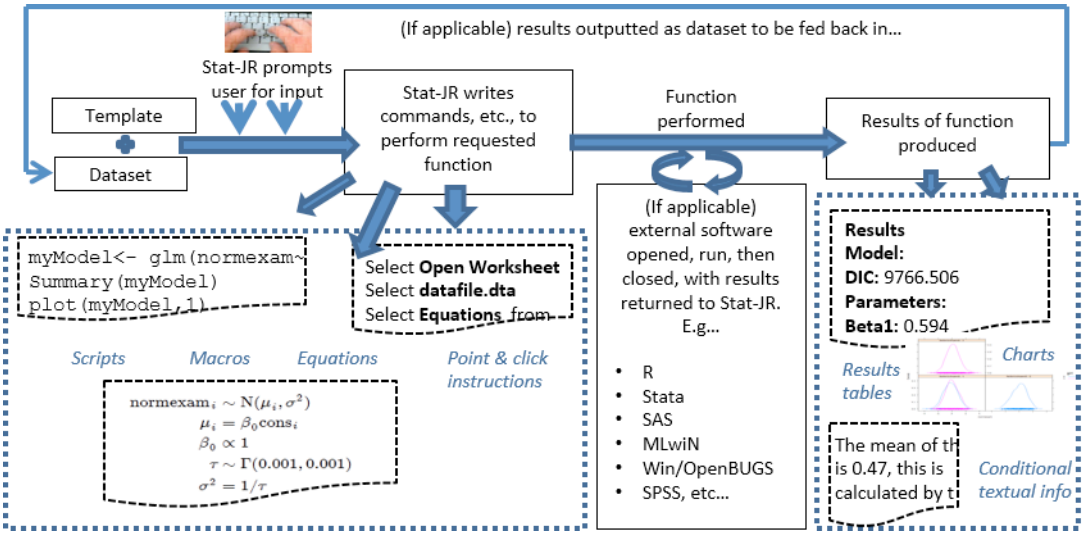

Figure 1. Overview of working with Stat-JR.

Stat-JR can be operated via the following two web-based interfaces (as well as via the command-line):

- TREE (Template Reading and Execution Environment), is a menu-driven point-and-click interface, hosted on a web browser. It is a flexible environment in which templates are paired up with datasets, and then users specify inputs to perform desired functions.
- DEEP (Documents with Embedded Execution and Provenance), on the other hand, is an interactive eBook-reading interface; it also uses templates and datasets and can fully-exploit Stat-JR's statistical functionality, but allows more contextual information to be provided by the eBook author, and is a more tailored environment. eBook authors can choose which outputs from a template execution to present embedded in the eBook for the reader (although all outputs are accessible behind the scenes). Results can be inserted in surrounding text, the content of which can itself be returned conditional on the results. See Figure 2 for an example screenshot.

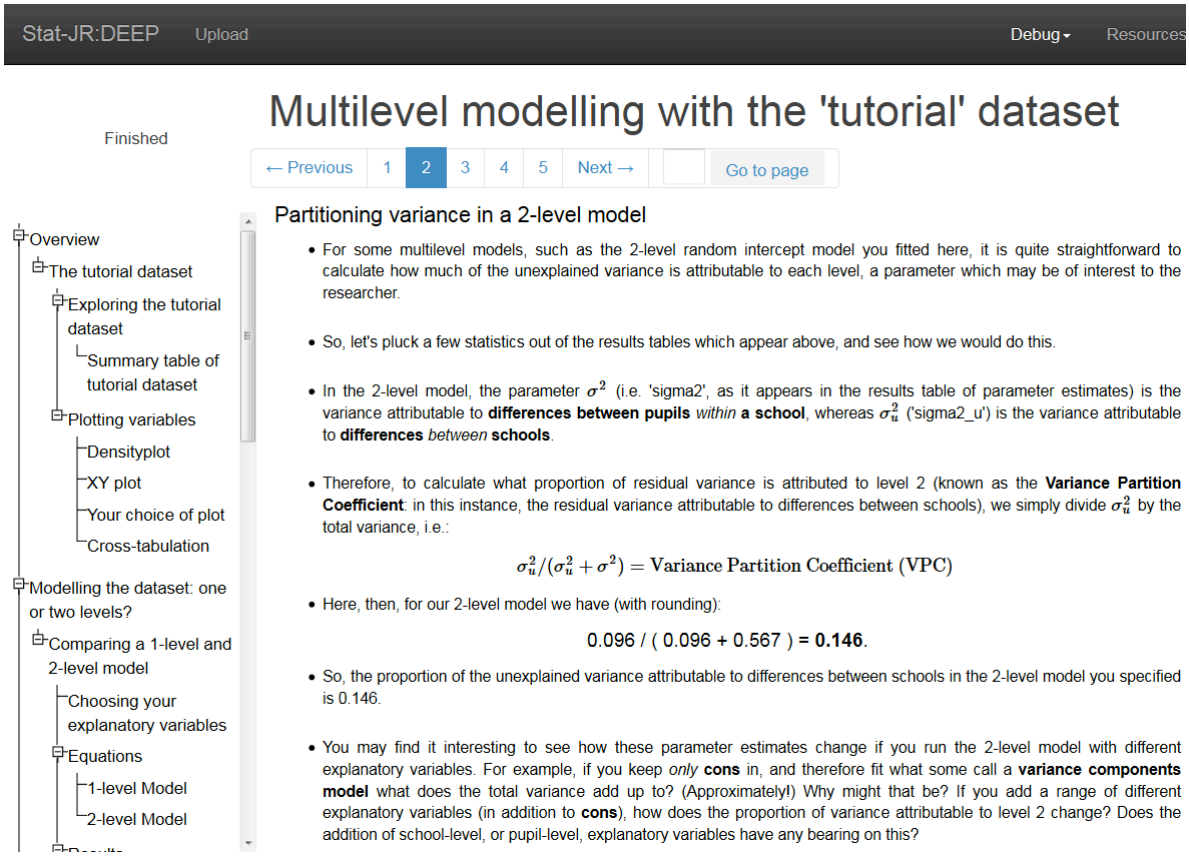

Figure 2. Example screenshot of a prototype eBook from Stat-JR's DEEP interface.

The software is continuing to evolve as part of a current ESRC-funded grant, including the development of a workflow system. It is currently distributed with the multilevel modelling software package MLwiN; funding from the ESRC has allowed both to be currently provided free to UK academics and they are otherwise available for purchase (www.bristol.ac.uk/cmm/software/).

INTERVIEWING RESEARCHERS & DEVELOPING EBOOKS

Arranging interviews

We have invited a number of research experts to take part in the study. As well as providing them with information outlining the aims of the project and the nature of their potential participation, they are sent a consent form indicating that they are free to withdraw from the study at any stage, thus ensuring the resulting eBook will only be published if all parties are satisfied with it.

If they agree to participate we then ask them to choose a quantitative research study they have worked on with a mind to translating it into an interactive eBook. Once a case study has been

chosen, we conduct an initial interview with them in which we try to capture the essence of how they approached their research question(s), and analysed the data.

Case studies & eBook development

At the time of writing, interviews have been conducted with eight quantitative researchers who have chosen case studies covering aspects of criminology, social geography, education, health statistics, social network analysis, and the analysis and presentation of official statistics amongst others. We are combining the information from the interviews with statistical software scripts, dataset(s) and other available contextual information to feed into the collaborative construction of interactive eBooks; these are being written using the Stat-JR package (described above).

Whilst the form and emphasis of each eBook depends on the particular case study, in general the eBooks seek to describe the contextual background of the research, the hypotheses, how the dataset was sourced, and all the steps taken in drawing the conclusions, including data exploration and analysis. Interactive elements will invite the reader to engage with the eBook in a dynamic way to enhance, and test, their understanding, and to investigate other avenues the analyst may not have chosen. For example, some of the case studies emphasise the sensitivity of inferences to the choice (and any transformation) of variables, others to the selection of model type, others still to key decisions when choosing how to graphically communicate data.

The researchers interviewed have used a range of statistical software in their research, including R, Stata, SAS, MLwiN and JAGS and thus the eBooks will outline the manner in which they undertook their research in those packages. However, Stat-JR's interoperability also allows us to explore opportunities to convey how the analysis might have been conducted using other statistical software as well.

Developing a statistical analysis assistant

A parallel objective of the current project is to design a statistical analysis assistant (SAA) which will integrate the eBook system with software designed to run analysis workflows. The insights and experiences of our participating analysts, together with that of the project's co-investigators and other surveyed material will inform the structure of the SAA.

CONCLUSIONS

As opportunities for quantitative exploration and analysis become increasingly available through the proliferation of data, technology can offer new tools to address the parallel and important need to provide up-to-date, effective training to help researchers realise those opportunities. To date, statistical teaching resources have paid relatively little attention to conveying what working analysts do in practice – the everyday decisions and choices they make, despite the fact that such insights would help demystify and elucidate the quantitative research process. However, whilst not without challenges (e.g. Chance et al., 2007; Magnus & Morgan, 1999; Nolan & Lang, 2007; Velleman, 2014) the ability to develop interactive and non-linear software environments does open the door for such information to be fruitfully embedded and explored, allowing users to investigate decisions and details often lost or obscured prior to publication in a journal.

ACKNOWLEDGEMENTS

We thank the participating analysts for their positive and constructive collaboration, and the UK's Economic and Social Research Council (ESRC) for funding the current research project (ES/K007246/1). The development of Stat-JR was funded by the ESRC under the e-Stat node of the Digital Social Research programme and the LEMMA II and LEMMA III nodes of the National Centre for Research Methods.

REFERENCES

Chance, B., Ben-Zvi, D., Garfield, J., & Medina, E. (2007) The Role of Technology in Improving Student Learning of Statistics. *Technology Innovations in Statistics Education*, *1*(1). Retrieved from: http://escholarship.org/uc/item/8sd2t4rr

Charlton, C. M. J., Michaelides, D. T., Cameron, B., Szmaragd, C., Parker, R. M. A., Yang, H., Zhang, Z., & Browne, W. J. (2014). Stat-JR Version 1.0.2. UK: Center for Multilevel Modelling, University of Bristol and Electronics and Computer Science, University of Southampton. Available at http://www.bristol.ac.uk/cmm/software/statjr/

IBM Corp. (2013) IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.

Lunn, D. J., Thomas, A., Best, N. & Spiegelhalter, D. (2000). WinBUGS -- a Bayesian Modelling Framework: Concepts, Structure, and Extensibility. *Statistics and Computing, 10*, 325-337.

Magnus, J.R., & Morgan, M.S. (1999). Lessons from the tacit knowledge experiment. In: J.R. Magnus & M.S. Morgan (Eds.) *Methodology & Tacit Knowledge*. Chichester, England: John Wiley & Sons Ltd.

Nolan, D., & Lang, D.T. (2007). Dynamic, Interactive Documents for Teaching Statistical Practice. *International Statistical Review*, *75*(3), 295-321.

Python Software Foundation (2010). Python Language Reference, version 2.7. Available at http://www.python.org

R Core Team (2014). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. Available at http://www.R-project.org

Rasbash, J., Charlton, C. M. J., Browne, W. J., Healy, M., & Cameron, B. (2009). *MLwiN Version 2.1*. UK: Centre for Multilevel Modelling, University of Bristol. Available at: http://www.bristol.ac.uk/cmm/software/mlwin/

SAS Institute Inc. (2013) SAS 9.4. Cary NC: SAS Institute Inc. Available at http://www.sas.com

StataCorp. (2013) *Stata Statistical Software: Release 13*. College Station, TX: StataCorp LP. Available at http://www.stata.com

Thomas, A., O'Hara, B., Ligges, U., & Sturtz, S. (2006). Making BUGS Open. *R News, 6*, 12-17.

Velleman, P. (2013). Comment: Let's All Write and Teach with e-Books! *Technology Innovations in Statistics Education, 7*(3). uclastat_cts_tise_20090. Retrieved from: https://escholarship.org/uc/item/40v8s9nf