A REPORT COMPARING METHODS USED TO TEACH REGRESSION ANALYSIS TO NON-STATISTICIANS

LEE, Sophie and WADE, Angie
Centre for Applied Statistics Courses,
Institute of Child Health, UCL, London, UK
sophie.a.lee@ucl.ac.uk

This paper compares 3 methods of teaching introductory regression analysis to non-statisticians from diverse backgrounds. The Centre for Applied Statistics Courses at UCL has developed regression courses that are structured differently and use software to varying degrees. 'Logistic Regression' is taught on a single day in a traditional classroom, is completely theory-based and involves no software. 'Regressions with R' runs over two shorter days and allows delegates to apply the theory being taught in the course to datasets using R software. This course is held in a computer room where everyone has their own computer to apply the theory just taught. There is complete integration between the theoretical sessions and the immediate practical applications interspersed. 'Introduction to Regression' combines the two ideas, with one short day teaching the theory of regression, followed by a second day interpreting SPSS output from an example application. The dataset is provided for delegates to go away and practice replication if they wish to, with whatever package they may prefer. Feedback and questionnaire results from course delegates to each of the 3 types of course will be used to investigate which method appears more effective in providing them with useful tools for their own research.

INTRODUCTION

The Centre for Applied Statistics Courses (CASC), based in UCL's Institute of Child Health, comprises a small team that organise and teach a range of short-courses to non-statisticians. The audience for these courses is diverse, with delegates coming from a wide range of backgrounds in terms of discipline, profession, and seniority. CASC currently delivers 16 different short-courses, and delegates attending these courses are asked to complete a short feedback questionnaire to assess how suitable the course was to their needs, and how future courses could be improved.

The understanding of, and ability to build, regression models is a cornerstone for any quantitative researcher. Hence it is important that individuals whose research conclusions are based on the analysis of a collation of sampled data learn about the technique. Included in the programme of courses we deliver are a range of introductory regression courses; these courses were developed at different times and utilise slightly different approaches with respect to the breakdown of classroom-based theory and software application. This paper compares feedback provided by students who have attended one of three CASC introductory regression analysis courses: 'Introduction to Regression', 'Introduction to Logistic Regression' and 'Regressions with R'. Feedback from these courses is presented in the context of andragogic research in order to discuss the optimum method to teach regression analysis to non-statisticians, so that they may best comprehend the theory underpinning regression analysis, retain this information, and appropriately apply it in real-world contexts.

THE COURSES

'Introduction to Regression Analysis' was the first introductory course developed and delivered in 1998, before CASC was formally founded, in response to a large number of enquiries from researchers within the Institute of Child Health regarding the application of regression analysis. When this course was first designed, there was only a single computer lab within the institute which was often not available for teaching in its entirety and very few delegates would own a laptop, making hands-on practice of the methods infeasible. To overcome the lack of resources, the notes contained output and interpretation of an example analysis which had been carried out using the statistical package SPSS. Additional to the lack of resources, there was a belief that learning away from the computer screen may be more effective as software operation may detract from the course aims of instilling understanding of the technique and application.

The basic format of the 'Introduction to Regression Analysis' has not changed since its inception, although the content has been updated. It consists of two short days (7.5 teaching hours in

In H. MacGillivray, M. Martin and B. Phillips (Eds.). Proceedings of the Ninth Australian Conference on Teaching Statistics, December 2016, Canberra, Australia.

total) with the first day devoted to the theory of linear regression, including interpretation of model coefficients, assessing goodness-of-fit, prediction and validation. During the second day of the course, an example dataset is introduced and output presented from an analysis carried out using SPSS. Group discussion is used throughout day two to aid delegates' understanding of the methods taught on day one. The example dataset introduced during the course is made available for delegates after the course so that they are able to apply the taught methods using the software of their choice. There is no hands on demonstration or application of any software, even by the presenter.

'Introduction to Logistic Regression' was developed in 2005 as an additional one day (5 hours) course to complement the existing 'Introduction to Regression Analysis' course. Although this was 7 years after the first course was produced, computer facilities for teaching large groups were still somewhat limited, so hands-on practice of the models on the day was again discounted as an option. The Logistic Regression course also included SPSS outputs, but in contrast to the earlier course, these were interspersed in the notes and discussion of them integrated with the formal training of the theory. Delegates are taught how logistic regression can be applied to explain the variation in the probability of an outcome occurring and how to interpret and utilise model diagnostics produced by the computer software. The course also contains a supplementary section that explains how this method can be extended to deal with ordinal and nominal outcomes.

'Regressions with R' is the newest of the three regression courses currently offered by CASC, and was developed in 2014. By the time 'Regressions with R' was first run, teaching facilities with access to computers and ICT support was readily available making a software-based course feasible. These resources were utilised to create a teaching environment different to the other introductory courses, allowing delegates to not only learn the theory of the methods but also practice using software to apply them to real data. The course is taught over the same timescale as the 'Introduction to Regression Analysis' (two short days) in a computer room in which each delegate has their own computer and access to R software. The theory that is covered in the course is supplemented with examples of R code and output, which aim to explain key ideas and demonstrate how these methods can be applied to real-world problems. Linear, logistic, ordinal logistic, poisson, negative binomial, Cox proportional hazard and multilevel regression models are covered. Importantly, the course also includes instruction regarding interpretation of models, model diagnostics, and good practice when fitting regression models. Each type of regression is introduced using an example dataset on which the teacher demonstrates analyses using R software. Delegates are then asked to complete practical exercises, based on the theory just covered. Hence the computer interaction is interspersed throughout the course, with alternation between taught theory and hands-on practice.

COURSE FEEDBACK

Since 2010, following each course that is run by CASC, all delegates who attend are sent an online feedback request. The feedback form includes multiple questions regarding content, presentation, usefulness and the organisation of the course. For each question, delegates are asked to respond on a 5-point likert scale from (1) *not at all* to (5) *very much.* Respondents are also asked several open-ended questions to give them the opportunity to explain their scores and/or make more detailed suggestions about how they believe the course could be improved. These feedback forms are intended to inform future courses and give indications of when courses should be changed or are missing anything of importance. We present here selected quotes from all courses on which electronic feedback was available together with a more detailed quantitative assessment of those within the current year (2016).

Introduction to Regression Analysis

This flagship course has run 27 times since 2010 and the feedback was typically completed by approximately half of the delegates that attend this course. Comments frequently referred to the absence of practical application involved in the course. Some delegates suggest that the instructor could/should carry out the analysis as a live demonstration during the class:

"…*would be good to have time to practice example live on SPSS and then trouble shoot with trainers*"

In contrast, others implied that hands-on practical sessions would be preferable, where each delegate has access to their own individual computer:

"*A stats course must involve a practice component running analysis in a statistical package. The practice stuff involved only theoretical questions, print outs and slides". "I would have liked to use SPSS simultaneously when doing the regression analysis rather than follow handouts*".

Although most groups contained at least one comment regarding the lack of practical application within the course, the majority of comments were positive and the use of example outputs, even without hands-on demonstration of the software, was appreciated:

"*…the course was extremely informative. The SPSS examples were very useful and I was impressed with the depth that the course when into in relation to SPSS outputs - I have not experienced this on other courses*".

In 2016 the course has so far run 4 times with a total of 90 people attending of whom 59 (66%) provided feedback. The responses received were overwhelmingly positive: 45 (76%) of the delegates that responded to feedback rated the course overall 'very good' (5), 11 (19%) rated it 'good' (4), the remaining 3 (5%) rated the course 'average' (3).

Introduction to Logistic Regression

Since 2010 this course has been run 24 times with most feedback expressing satisfaction regarding the level at which the course was pitched, and the way in which it was structured. However, some delegates suggested that more practical sessions should be included to help consolidate their knowledge. Since 2015, the course has received an increase in the number of suggestions for the inclusion of a software element to the course. One delegate reported that he/she felt that "*it was a bit too abstract*", while another suggested changing the format of the course, and felt that running it "*as a 2-day course with some practice activities in SPSS would be better*".

However, the most recent feedback from Logistic Regression in July 2016 contained two individuals who felt that some parts of the course were irrelevant as they were based on SPSS:

"*I use STATA rather than SPSS so the SPSS parts weren't so relevant to me. Perhaps some written examples of how to run/interpret STATA outputs would have been helpful*".

One of these delegates had brought example STATA output to the course to interpret, and subsequently struggled to apply the interpretation of SPSS outputs to the STATA output:

"*Even with the STATA output in front of me, I wasn't sure exactly how this related to the SPSS outputs*".

The 'Introduction to Logistic regression' course has run three times so far in 2016 with 38 delegates attending in total; of those 38, 22 (58%) responded to feedback requests. Of the 22 who provided feedback, 12 (55%) rated the course overall as 'very good' (5), 7 (32%) rated the course 'good' (4) and the remaining 3 (13%) rated the course 'average' (3).

Regressions with R

This course, first run in 2014, has elicited electronic feedback for all of the 11 occasions on which it has been presented. While 'Regressions with R' incorporates the software exposure that delegates felt was missing from the other two regression courses, some delegates reported that they found it difficult to concentrate on both the theory being taught and the R code being explained at the same time:

"*I found myself trying too hard to keep up with typing the code and missing the real point*".

The feedback for this course suggested that combining software exercises and theory within the same session may not be the most effective structure for a course of this type; the implication of feedback being that some delegates struggle to understand the theory whilst also trying to understand the software application, and vice versa. Some delegates suggested spending one day of the course solely teaching the theory, and the second day applying this theory using computer programs.

"Thought [the teacher] went quite quickly through the R code, (which was fine if very familiar with R) but I am not so familiar. I feel a day could have been dedicated to the theory and a day dedicated to the exercises"

This course has a pre-requisite stated on the website that delegates must have a working knowledge and ability in R in order to attend this course; however, the feedback received suggests that this criteria is not always met.

"Many participants seemed to have pretty advanced knowledge of R and could keep up, but for others (like me!) it was difficult to keep up."

Introducing a test of some sort, or clearer guidelines regarding who this course is appropriate for, may improve the feedback for this course and reduce the number of delegates who feel overwhelmed by the combination of theory and software application.

This course ran twice in 2016 with 33 attendees in total; of which, 21 responded to a request for feedback. Of the 21 respondents, 9 (43%) rated the course overall 'very good' (5), 10 (48%) rated the course 'good' (4) and the remaining 2 (9%) rated the course 'average' (3). Although this feedback was still overwhelmingly positive, the 'Regressions with R' course received fewer 'very good' (5) responses than the other two courses despite containing the software exposure that delegates felt were missing from the other courses; 75% of delegates that attended 'Introduction to Regression Analysis' in 2016 rated the course 'very good' (5) overall whereas just 43% of delegates attending 'Regressions with R' 'very good' meaning almost a third less (32.1%) rated the 'Regressions with R' course 5.

DISCUSSION

Active learning is a method of teaching which encourages students to actively take part in the lesson and gain hands-on experience of the subject matter being taught; active learning is "*anything that involves students in doing things and thinking about the things they are doing*" (Bonwell & Eison, 1991, p. 2). Since they require motivation, and are said to learn better when they feel the subject matter is important to them, active learning may be an effective tool for adult learners, as it can be used to demonstrate how the theory is applied in practice. There are many ways in which active learning can be incorporated into courses, such as use of practical exercises and discussion sessions within teaching. Problem based learning (PBL) is one such method of active learning, which provides delegates with hands-on experience of analysis. PBL requires the teacher to introduce a problem to students as well as guidelines about how the problem should be approached; the teacher acts as a facilitator whilst students take the lead and aim to solve the problem either as a group or alone (Hmelo-Silver, 2004). PBL encourages delegates to critically think and engage with the subject area.

All three regression analysis courses offered by CASC contain elements of active learning, but adopt different methods with regards to hands-on use of statistical software. We have here described these courses, the differences between them with regards to software usage, and contrasted student feedback. Through this initial look at our data, we have identified some potentially useful insights to inform our future direction.

Although all three courses received favourable feedback, the response to software components (or lack thereof) was rather mixed. It appeared that some delegates prefer to learn the theory behind the methods before moving on to discussing how those methods can be applied, whilst others commented that separating theory and application made it more difficult to retain the information taught in theory sessions without practical applications. Since the 'Introduction to

Regression Analysis' and 'Introduction to Logistic Regression' courses were first designed, computers have become substantially more accessible and powerful; many delegates now have access to one or more statistical packages and will have used them in their day-to-day work. Based on the number of comments noting the lack of software element in both courses, it is clear that delegates feel software exposure is something that they find necessary to any contemporary statistical courses. Additionally, the increased exposure to and use of computer based applications, does mean that usage in the classroom is less novel, more expected, and less likely to involve unnecessary disruption of the flow by unrelated computer concerns.

Another problem faced when combining a course with software use, is the choice of package. Dealing with delegates from diverse backgrounds, some may be familiar with more complex packages such as R, while others may have used a simpler package such as SPSS; similarly, some may have extensive experience, while other may only ever have used a statistical software package briefly or not at all. The risk of incorporating a day spent on either R or SPSS to 'Introduction to Regression Analysis' or 'Logistic Regression' is that some delegates may feel the course is no longer suitable to their needs.

The overall view from the feedback received is that, although there is no perfect method to teach regression analysis, a software component is desirable. This is a common theme across all three courses as reflected by their course evaluation comments. However, there is a danger, if these two components run concurrently (as they currently do in 'Regressions with R'), that delegates will focus too much on understanding the intricacies of the software program rather the methods being demonstrated.

Based on the evidence from course feedback, I feel that the effectiveness of methods for teaching regression analysis to non-statistician could be improved in one of two ways. Firstly, the current method used to teach Introduction to Regression Analysis could be extended to allow delegates access to computers in order for them to carry out analyses on the second day, with the help of the teacher. Alternatively, the method currently used to teach Regressions with R could be adopted, but with stricter guidelines regarding the level of software knowledge required for the course, and a greater allowance of time is given for explanations of theory. Both of these methods would provide exposure to software for delegates, would demonstrate the usefulness of methods taught in a real-world setting, and would include a PBL component to challenge and motivate adult learners.

REFERENCES

Bonwell, C. C., & Eison, J. A. (1991). Active Learning: Creating Excitement in the Classroom. 1991 ASHE-ERIC Higher Education Reports. ERIC Clearinghouse on Higher Education, The George Washington University, One Dupont Circle, Suite 630, Washington, DC 20036-1183.

Rubenson, K. (2011). Adult learning and education. (Ed.). Academic Press.

Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? Educational psychology review, 16(3), 235-266.