Reviewing: Analysis of Website Visits Data
Advanced Methods III 140.753

1 Summary

The author analyzed the visits data to two websites —Tumblr and Wordpress— separately and also jointly. The goals were to calculate the expected fraction of visitors the websites retain from each spike and to identify any factors that influence this fraction. To do so, the author first analyzed three different peak finder methods: (1) top 5% of all visits data, (2) greater than the mean plus 2 standard error, or (3) inclusion in the largest group from a k-means clustering. After an exploratory data analysis the author decided to use only the first two methods. To attack the goals, the author then used three regression models:

1. linear piecewise spline regression on the date given the visits. This model is used to give a final justification of the peak finding method.

2. linear regression model on the number of visits at a spike given the difference in mean values from before and after the spike using non-spike values. The author claims that the coefficients give the fraction of interest.

3. linear regression model on the same difference in mean values as in the previous model given the fraction of interest.

The author finds that when using the first peak finding method and all the data in their second model, the fraction retained by the websites after a spike is 8.1% with a 90% CI of [-5%, 22.5%]. In addition, Tumblr performed better than Wordpress with estimates of 4.3% and 1.7% respectively. Finally, using the third model there is a positive correlation between the difference in the mean values and the retained fraction of visitors although the estimates are close to 0. Thus, the larger the difference in means, the larger the fraction of visitors retained.

2 Major revisions

- The exploratory data analysis part uses plots to justify which peak finding method to use. However, no plot is included. If a plot is part of the story, as in this case, then it should be included in the report.

- The results from table 1 and the text are incongruent. How come the 90% for Tumblr is [0.5%, 0.8%] but it doesn't include the estimate of 4.3%? Furthermore, why are the estimates for Tumblr and Wordpress lower (4.3% and 1.7%) than when using all the data (8.1%)? Could it be that the Wordpress data used is the one for all time under study instead of when SimplyStatistics moved to Wordpress?

- The description of the third statistical model is not clear in specifying how the fraction of visitors retained is calculated. Is it from the second model? If so, doesn't the second model give two coefficients? Or is the second model used for each spike individually (in this case, can you justify using linear regression)? This confusion can be solved by explained more explicitly the models.

3 Minor revisions

- There is no mention of which websites are being analyzed. It could be the whole of Tumblr and Wordpress or some specific blogs. The reader would not know that it's SimplyStatistics.org.
- Tables do not have numbers (table 1, table 2, etc) nor captions.
- Reproducibility can be improved. Sure, the code has some comments, but there is no need to have a single 855 line long file. Specially when reloading of the data is needed. For example, the code could be split in three files: one for each peak finding method. Then, code that is used more than once can be turned into functions and saved into a functions.R file that is sourced in the other three. Furthermore, some of the variable names are not descriptive enough. For instance: *aggregate* and *aggregate2*. Plus there are no comments describing what they are.
- Is the report itself reproducible? If not, maybe doing so can avoid issues such as the incongruence between table 1 and the confidence intervals.