Analysis of SimplyStatistics' new visitors after popularity spikes
Advanced Methods III 140.753

1 Introduction

SimplyStatistics is a blog ran by three faculty members of the Johns Hopkins School of Public Health Biostatistics Department. It has been live since late 2011 and has been hosted under the Tumblr and WordPress platforms. The administrators of the blog have visitors data given by Google Analytics including information on referrals from Twitter. This blog, online at `http://simplystatistics.org/`, is one of the most popular statistics blog in terms of number of visits. These bloggers are interested in knowing what is the expected fraction of visitors they retain after a spike in the number of visitors. Additionally, they want to know if there are any factors that influence such fraction.

2 Pre-processing

The initial visitors data was given to us by Jeffrey Leek. To complement this information, I scrapped from their Tumblr and WordPress sites the author, date and title of their posts. Furthermore, I acquired the Google Analytics data for the blog Fellgernon Bit available at `http:://fellgernon.tumblr.com`. I also scrapped the date of the posts. However, since some posts were moved to `http://fellger.tumblr.com`, I also scraped the relevant post dates from that site.

The original SimplyStatistics data is split from their two sites, so for simplicity I merged their visitor information by date while still keeping the visits separate from the twitter information. I used the dates from the Tumblr site up until they moved to WordPress, and then switched to WordPress. Similarly for the authors and titles of the posts.

Finally, I ranked on a 1 to 5 scale the *controversy* level of the title of each of the 511 posts. I did so by permuting the titles to avoid any temporal bias and repeated the procedure once in an effort to be more consistent in the ranking. Note that the author name was blind when doing this process.

3 Exploratory data analysis

3.1 Fellgernon Bit

Using the data from Fellgernon Bit I was able to determine the minimum level that posting has in creating a spike in the number of visitors. The data from this blog is useful to answer this question because posts are made sporadically and we can consider the visitors on non-post days as mostly noise as shown in Figure 1. The difference in mean number of visitors on post days vs non post days is significant (two-sided T-test, p-value 1.5279×10^{-15}, 95% CI: (-12.465,-8.122)). Furthermore, the two sided t-Test for a difference in means for the number of visitors during non-post days versus non-post days excluding in addition the day after a post is published is non-significant (p-value 0.3171, 95% CI: (-0.269,0.828)). Therefore, at the very least, posting drives visitors into a blog the day the post is made.

3.2 SimplyStatistics

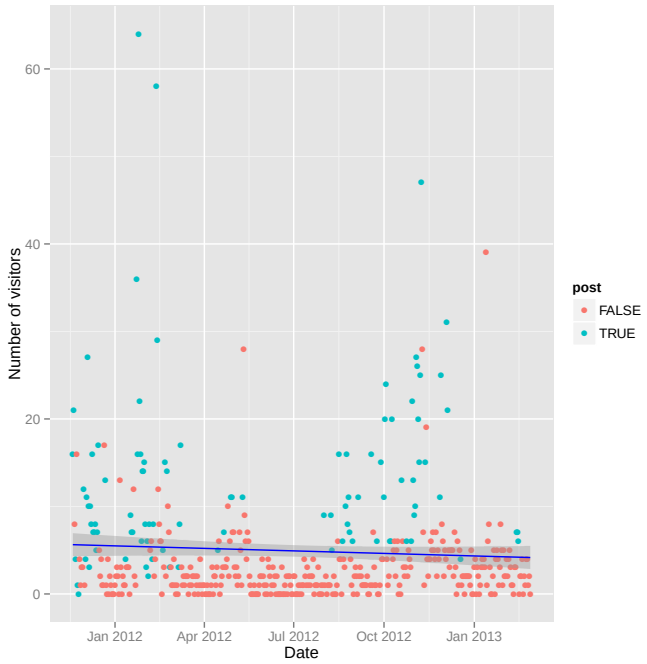

Figure 1: Number of visits to Fellgernon Bit per day. Points are colored according to whether a post was made that day or not. Linear model fit (blue line) of visits explained by date along with confidence bands.

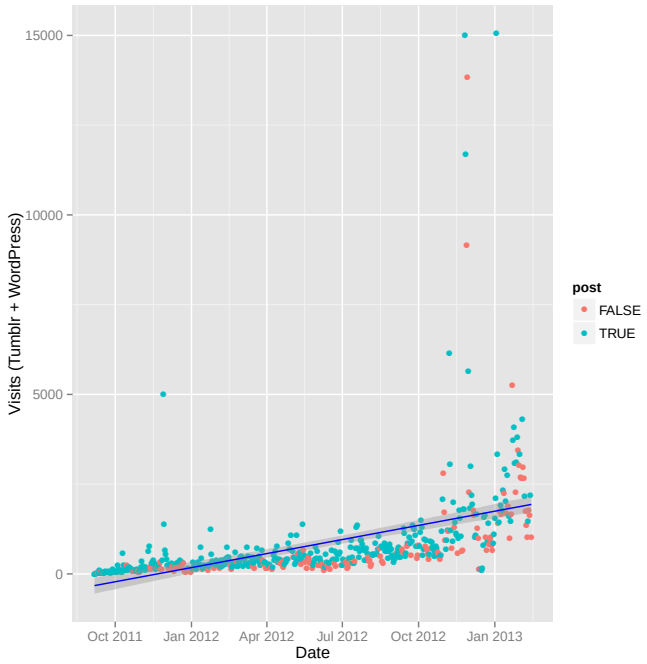

Figure 2: Number of visits to Simply Statistics per day. Points are colored according to whether a post was made that day or not. Linear model fit (blue line) of visits explained by date along with confidence bands.

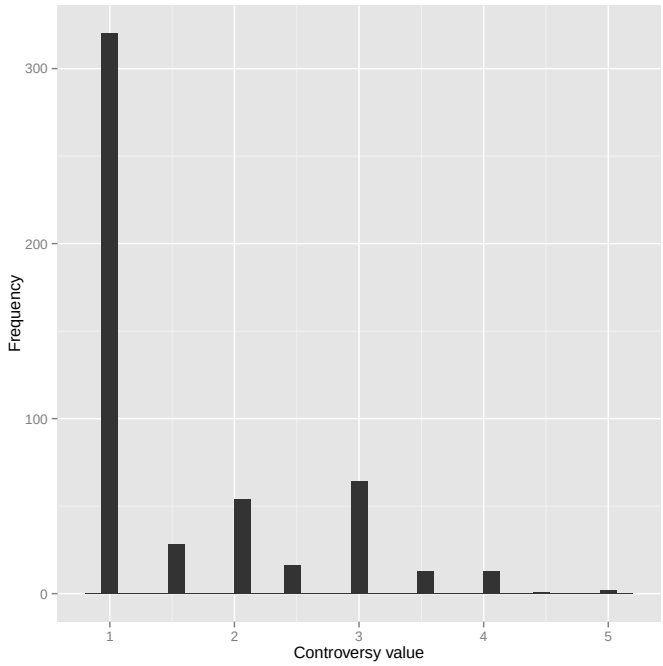

Figure 3: Histogram of the controversy value of the SimplyStatitics post titles. Data used is the mean from two replicates of the rankings where 5 is a highly controversial and 1 is not considered controversial at all.

In Figure 2 there are three observations to be made. First, that the overall trend os the number of visitors is positive as shown by the linear regression fit[1]. Secondly, there are posts made the majority of the days in the time under study. So a more elaborated way of choosing the peaks is needed rather than just select days with posts made. Third, there is some data missing at the end of 2012. To fix the five days where Google Analytics failed for WordPress, I considered using the mean from the previous days of the month and the data from Tumblr.

Figure 3 explores the results from the *controversy* of the titles. Most of the titles do not seem to be controversial and only a handful are highly controversial. However, the interesting part is that the mean controversy scores for the posts are 2.2188, 1.6385, 1.3312 and for Rafael Irizarry, Jeffrey Leek, and Roger Peng respectively. All the pairwise differences are statistically significant (two-sided t-Tests p-values: Roger vs Rafa 5.9887×10^{-10}, Roger vs Jeff 1.2197×10^{-4}, Rafa vs Jeff 5.2213×10^{-5}). This could be further improved by looking at the posts themselves since many of them do not have a title.

4 Results

4.1 Finding the peaks

To find the peaks I used a method based on running means. First, I calculated the running mean for each day $\pm$ a window of size w. Then, I calculated the value at a given day minus it's corresponding running mean. From this difference, I then chose those beyond a quantile q as possible peaks. This is shown in Figure 4 (top panel). The candidates seem reasonable under various w and q (data only shown for $w = 20, q = 0.9$). We can see this in Figure 4 (bottom panel) because the colors don't

[1]It is not completely appropriate since it can go into negative values. But it serves it's illustrative purpose.

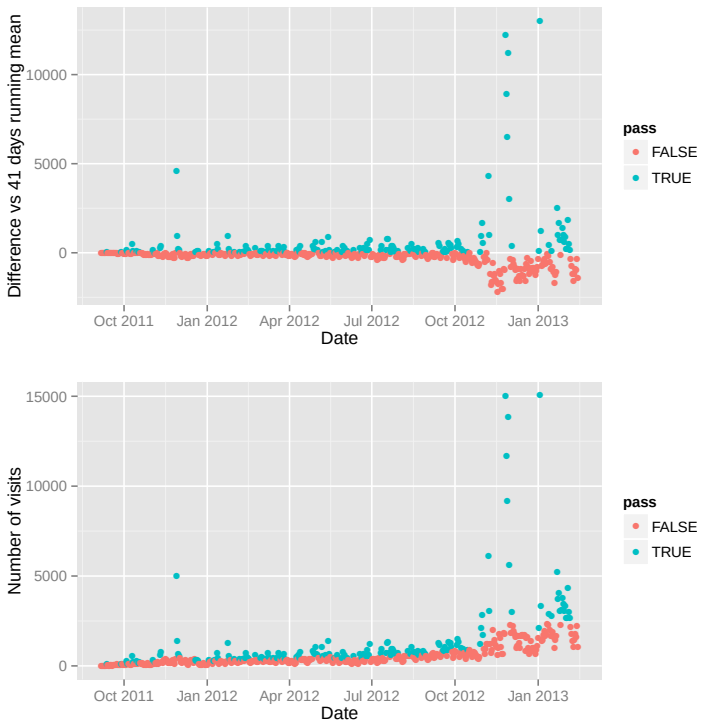

Figure 4: (Top) Number of visits minus the running mean by date. (Bottom) Number of visits by date. In both, the color is given by whether the observation it passes as a candidate peak or not.

mix and it's only the high values that are shown as candidate peaks. Next, I merged the candidate peaks that were within d days of each other (Figure 4 uses $d = 10$). In this case, the procedure reduced from 159 candidate peaks to 17 peaks. Choosing d is tricky since it can lead to wide peaks if using a large value, but a low value can fail to detect the peaks specially under highly variable periods such as the data for 2013.

4.2 Outcome of interest

The main outcome of interest is the expected fraction of visitors that SimplyStatistics retains after a spike in the number of visitors. For each peak, I calculated this fraction by

$$Y_p = 100 * \frac{\bar{c_p} - \bar{a_p}}{\max_j \{b_{pj}\} - \bar{a_p}}. \quad (1)$$

Where a_{pi} is the number of visitors for day i where i is in the days before the peak p up to the previous peak $p-1$ (not included). b_{pj} is the number of visitors for day j where j is in the days from the peak p. Finally, c_{pk} is the number of visitors for day k where k is in the days after the peak p up to the next peak $p+1$(not included).

Think of a, b, c as before, during and after the peak. Thus, the numerator of Y_p is the difference between the mean after the peak and before the peak. The denominator is the difference of the maximum height of the peak versus the mean before the peak. Thus, Y_p is the percent of visitors at a pinnacle of the peak immediately retained after a peak.

Using $w = 20, q = 0.9, d = 10$ (option 1), the mean $\bar{y}$ of the random sample is 6.1068 with standard deviation 8.1207. The null hypothesis that $\bar{y} = 0$ is rejected using a two-sided t-Test (p-value

0.0069, and 95% CI (1.932,10.282)). Using $w = 50, q = 0.7, d = 7$ (option 2) the results are similar with $\bar{y} = 8.2854$ and significantly different from 0 (p-value 0.0172, and 95% CI (1.676,14.895)). Further analysis is needed for determining which values of w, q, d to use and it's impact on the peaks' width.

4.3 Important factors

I calculated several metrics for each peak such as the number of posts by author during the peak, the maximum controversial title ranking of a post during the peak, the difference between the maximum number of twitter visitors during the peak versus the maximum number of twitter visitors immediately before the peak, among others. However, none of these factors was significantly associated with the outcome of interest Y_p using linear regression with the peaks from option 1.

Using the peaks from option 2 did yield one significant association. It's between the outcome of interest and the most controversial title post made during that peak[2] as shown in Table 1 and illustrated in Figure 5. It's the model

$$Y = \beta_0 + \beta_1(peak.posts.cont.max) + \epsilon, \quad \epsilon \sim N(0, \sigma^2). \tag{2}$$

Option 2 results in wider peaks than option 1 and thus polls more information during the peaks. This can potentially explain the difference between the results. Nevertheless, other factors were not significantly associated with the outcome of interest.

It is interesting to note that the association shown in Figure 5 has a negative slope. The immediate interpretation is that SimplyStatistics losses visitors after a peak that was driven by controversial posts. However, further analysis are needed to confirm or disregard this claim.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	55.3547	12.0286	4.6019	0.0003
peak.posts.cont.max	-12.6012	3.1637	-3.9831	0.0012

Table 1: Linear regression summary results for model (2).

5 Conclusions

The number of visitors for SimplyStatistics has been growing over time. Having very active bloggers is most likely helping keep this trend positive. Over the course of it's existence, SimplyStatistics has had several peaks in the number of visitors. Using two sets of parameters to find the peaks resulted in significant results for the outcome of interest Y_p defined in (1). The 95% intervals are (1.932,10.282) and (1.676,14.895) with mean 6.1068 and 8.2854 respectively. Thus, depending on the peaks used, the expected number of visitors that give rise to a peak and that then return to the site is approximately 6% or 8%.

It is difficult to determine if any factor explains the outcome of interest Y_p. However, ranking the controversiality level of the post titles at random and blind from the author's name is potentially associated with Y_p. This result has to be interpreted carefully since the person doing the ranking has read nearly all of the posts of SimplyStatistics and is thus extracting more information from the title than someone who is seeing the titles for the first time. Nevertheless, this finding is interesting. Specially because more controversial titles can be detrimental to SimplyStatistics' interest of increasing it's fan base.

[2] Abbreviated as *peak.posts.cont.max*.

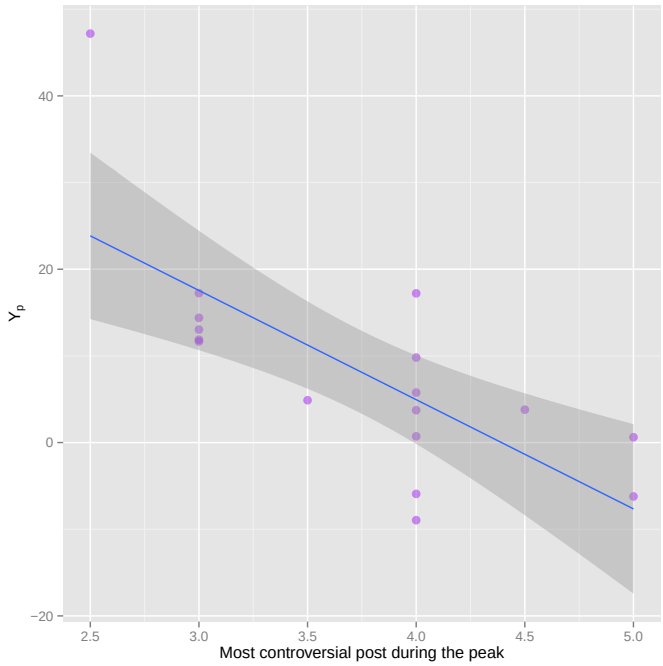

Figure 5: Outcome of interest vs the controversial rating of the most controversial post during the peak. Observations are shown in purple with alpha 1/2. Linear regression curve is shown in blue along confidence bands.

6 Acknowledgements

Discussed the general idea of how to analyze the data with Jiawei Bai and James Pringle. I tried to emulate the peak finder Jiawei described.

7 Reproducibility

The code, data, and report is available at `https://github.com/lcolladotor/lcollado753/tree/master/hw/data-analysis-02` where the file *README.md* describes the steps in which to run the scripts in order to reproduce nearly all the work. The part that is not reproducible is the *controversy* ranking, although you can try your own.

This report was generated using the following R packages.

- R version 2.15.2 (2012-10-26), `x86_64-apple-darwin9.8.0`
- Locale: `en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8`
- Base packages: base, datasets, graphics, grDevices, grid, methods, stats, utils
- Other packages: ggplot2 0.9.3, gridExtra 0.9.1, knitr 1.0.5, xtable 1.7-0
- Loaded via a namespace (and not attached): colorspace 1.2-1, dichromat 2.0-0, digest 0.6.1, evaluate 0.4.3, formatR 0.7, gtable 0.1.2, labeling 0.1, MASS 7.3-23, munsell 0.4, plyr 1.8, proto 0.3-10, RColorBrewer 1.0-5, reshape2 1.2.2, scales 0.2.3, stringr 0.6.2, tools 2.15.2