Econ 899: Assignment 4 – Data, Model and Methods

Tim Schulz

Jul 8, 2016

1 Data

The data underlying this paper are hourly Wikimedia page view statistics.[1] This database contains the number of visits and the amount of data (in bytes) transmitted in response for every page of the various Wikimedia projects accessed within a given hour from 18:00 (UTC), December 9, 2007 onward. These statistics are categorised by language of the page, page or project type[2], page/article name, number of visits[3], and the response size.

Since the individual pages are not of interest in this analysis, the statistics are first aggregated to hourly language-type data by summing the number of visits and the amount of traffic within each language-type category. Additionally, the number of pages that are aggregated to one data point is recorded.

A note regarding the "mobile" type is in order: This category does not differentiate between pages. It is already an aggregation of all mobile Wikimedia pages. As a consequence, the "mobile" category of each language seemingly contains only one page after aggregation, meaning the total number of pages is only available for non-mobile categories. However, the number of visits and amount of traffic are still analogous to other categories after aggregation. Therefore, the different non-mobile types are added up to one category in order to retain comparability with "mobile". The resulting panel contains the hourly number of visits and the amount of data of all mobile and all non-mobile parts of the Wikimedia Project by language. There are (roughly) 70,000 hours and about 200 language for a total of 14 million observations per variable.[4] However, since hourly data is extremely sensitive and technological adoption is probably measured in weeks if not months or years rather than hours, the data is further aggregated to weekly frequency. This has the added benefit of not having to worry about time fixed effects (possibly by language) of frequencies higher than that.

Lastly, a time series for both the number of visits and the amount of traffic is calculated. This time series contains the ratio, R_{lt} of mobile activity relative to the sum of mobile and non-mobile activity per hour t and language l.

[1] The root directory of the data dumps can be found here:
`https://dumps.wikimedia.org/other/pagecounts-raw/`

[2] The types are Wikibooks, Wiktionary, Wikimedia, Wikipedia, mobile, Wikinews, Wikiquote, Wikisource, Wikiversity and Mediawiki.

[3] The number of visits is the total number, not unique visits. I.e. if a person accesses a Wikipedia article multiple times within hone hour, each of these events is counted. It is not how many *people* visited a page.

[4] At this point, these numbers are just estimates. I will only know the exact number once my data set is complete in another week or so.

The main explanatory variable is the date of introduction of new mobile technology. In the case of iPhones, this is relatively straight forward given the small number of devices and low frequency. In this case, the release date of a new iPhone is used. For Android phones, however, this is not feasible given the large number of devices and irregular release dates across manufacturers. Therefore, the release dates of major Android OS versions is used instead.

2 Model and Methods

Since the first smartphones only appeared towards the end of of the 2000s, the ratio of mobile to total activity will be equal to or close to zero in the beginning of the time series and can naturally not exceed 1. That is, $R_{lt} \in [0, 1]$. Given these restrictions, the development of R_{lt} is assumed to follow a logistic growth function which resembles the observed S-curve[5] and has the convenient property of a simple derivative which can be interpreted at the instantaneous growth rate at a given time.

Realistically, mobile traffic will probably never represent all traffic. In other words, the ratio of interest will never equal 1 and probably also never approach it. Instead, it makes more sense to assume convergence to some other upper bound K_l that can be different for each language.

$$\lim_{t \to \infty} R_{lt} = K_l < 1$$

This closely follows the initial work on technology S-curves done by Griliches (1957) where

$$R_{lt} = K_l \Lambda \left(\alpha_l + \beta_l t \right) + \epsilon_{lt}$$

and

$$\Lambda(x) = \frac{1}{1 + e^{-x}},$$

which allows for different "starting points" α (in terms of Griliches, the first time R exceeds 10%) and rates of acceptance β for each language.

Therefore, when fitting the logistic function to the use of mobile websites within one language, first K_l has to be estimated. Next, α_l and β_l are estimated. All of these are estimated using maximum likelihood based on the `nlstools` package for non-linear models in R.

If, indeed, the introduction of new mobile technology results in a temporary acceleration of mobile internet usage, one would expect a series of positive residuals after the introduction date. The size of these residuals can be estimated using dummy variables that indicate whether a week coincides with the introduction of new technology, or is one, two etc. weeks later. This way, the effect can be estimated as a function of the temporal distance from the introduction of new

[5] I should be able to show this curve once I have all the data.

technology. More formally, assuming introduction occurs at time $t^\star$,

$$\begin{aligned}
R_{lt} = K_t \Lambda \left(\alpha_t + \beta_t t\right) &+ \delta_0 \mathbb{1}\left\{t = t^\star\right\} \\
&+ \delta_1 \mathbb{1}\left\{t = t^\star + 1\right\} \\
&+ \delta_2 \mathbb{1}\left\{t = t^\star + 2\right\} \\
&\vdots \\
&+ \epsilon_{lt} \\
= K_t \Lambda \left(\alpha_t + \beta_t t\right) &+ \sum_{j=0}^{J} \delta_j \mathbb{1}\left\{t = t^\star + j\right\} + \epsilon_{lt}
\end{aligned}$$

where $\mathbb{1}\{.\}$ is the indicator function and δ_j measures the effect of new technology j periods after its introduction.[6]

With the introduction of several technologies over time at times $t^\star$, $t^{\star\star}$ etc., this can be extended to

$$R_{lt} = K_t \Lambda \left(\alpha_t + \beta_t t\right) + \sum_{j=0}^{J} \delta_j \mathbb{1}\left\{t \in \left\{t^\star + j, t^{\star\star} + j, \ldots\right\}\right\} + \epsilon_{lt}$$

with the caveat that this forces the effect of the first introduction j periods after the introduction date to be the same as the effect of latter introductions of new technology j periods after the fact.

Theoretically speaking, one could have $t_l^\star$ depend on the language. This works in cases where a language is almost exclusively spoken within one country (e.g. Danish) but not in cases spoken in several countries with possibly different release dates of new devices (e.g. English). However, usually devices are introduced worldwide if not on the same day, then in the same week. In the case of software, this is even less of a concern as the release usually happens globally at the same time (down to the second). Therefore, having $t^\star$ vary by language is not necessary.

[6] I still have to figure out what a sensible J is.