Infographic: Google Search Prevalence by State

Chitika, Inc.

August 9, 2013

Abstract

This paper explains the statistical procedure used by Chitika Insights in producing the recent infographic and article on Google Search Prevalence by State. Google is the dominant search engine across all browsers. Moreover, unlike other search browsers, Google's presence is largely consistent across all browsers, and hence, different types of users. At Chitika, during the months of June and July, we observed that between states, Google usage rates were positively correlated with median household income, rate of job growth, level of education and median age. We utilized the Generalized Linear Model (GLM) to quantify the relationship between the variables.

1 Introduction

Every week Chitika produces a series of reports recognizing unique trends in internet usage. These trends can help better identify consumer behavior. For the article on Google search prevalence, we wanted to understand how Google usage varied by state in order to provide additional clarity for marketers and advertisers.

Previous research on Google usage rates indicates that Google is the go-to search engine for users across all browsers. Even on Internet Explorer, where the default engine is Bing, Google accounted for more than half of all search engine traffic as of August 2012. At that time, Yahoo was the second most popular engine, directing a distant 12.3 percent of search engine traffic compared to Google's 74.7 percent [2]. These figures highlight the consistency of Google's dominance. Understanding the wider range of usage rates between different states can help marketers better target regionalized campaigns based upon the broader preferences of users within particular states.

It is imperative to clarify that the model and the analysis below are not comprehensive. The data on both the dependent and the independent variables are at the state level, thus removing a great deal of granularity from the analysis. Moreover, there are several other independent variables which could be included. The model being used, however, provides a good framework for further research.

2 Data Sources

Chitika is an advertising network that serves billions of online advertisements globally per month. The company collects data from user agent information provided by web

browsers when it serves an ad impression. It maintains a database internally that is used to identify the search engine used by the user agent, along with information about the state where the search query originated. This database is constantly being refined as new software is released [1].

For the Google search prevalence report, the data was collected from June 26 to June 30, 2013 and July 20 to July 24, 2013. The data for all the independent variables were collected from the most recent Census[1], the American Community Survey (ACS)[2] [3] and the Bureau of Labor Statistics[4].

3 Methodology

The classical linear model and the majority of the minimum bias procedures are special cases of Generalized Linear Models (GLMs). The theoretical framework for GLMs allows for explicit assumptions to be made about the nature of the data and its relationship to the predictive variables. For example, unlike the traditional linear models, GLMs do not penalize the model if the error term does not have a Gaussian distribution or if the response variables are discrete.

After collecting the data on search engine usage rates by various users, the data was organized by state. Additionally, search engine usage was further grouped based on searches originating from Google or other search engines (other search engines included AOL, Yahoo, Bing and Ask). This was done to categorize the response variable into a matrix, where the first column is the number of 'successes' as measured by Google search usage rate and the second column is the number of 'failures' as measured by other search engines usage rate (Google search engine usage rate + other search engine usage rate = total search engine usage rate).

The assumed structure of the GLM can be specified as:

$$E[\mathbf{Y}] = g^{-1}(\sum \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \tag{1}$$

where $\mathbf{Y}$ is a vector of responses, $g(x)$ is the link function, $\mathbf{X}$ is a matrix produced from the factors, $\boldsymbol{\beta}$ is a vector of model parameters, which is to be estimated and $\boldsymbol{\epsilon}$ is a vector of known effects or 'offsets'. For our model, $\mathbf{Y}$ is a matrix of proportions for Google usage rates and non-Google usage rates, $g(x)$ is a logit function, $\mathbf{X}$ includes state-level data on Median Household Income, Rate of Job Growth, Level of Education, and Median Age, and $\boldsymbol{\epsilon}$ is a null set, since we are not including any priors in our model.

Logistic functions are useful when the response variable is binomial and the predictor variables are continuous, as in our case. Logistic regression models how the probability of an event might be affected by predictive variables (independent variables). In the traditional linear regression, both the dependent and the independent variables are

[1] http://www.census.gov/prod/2012pubs/acsbr11-02.pdf/

[2] http://www.census.gov/prod/cen2010/briefs/c2010br-03.pdf

[3] http://www.census.gov/compendia/statab/2012/tables/12s0233.pdf

[4] http://wpcarey.asu.edu/bluechip/jobgrowth/jgu_states.cfm

continuous, and if applied to a model with binary dependent variable can provide probabilities outside the bound of [0,1]. Logistic regression provides a simple solution to this problem by transforming the probabilities.

The link function $g(x)$ in the case of logistic regression is

$$g(\mathbf{X}) = \ln\left(\frac{\pi(\mathbf{X})}{1-\pi(\mathbf{X})}\right), \tag{2}$$

which is the inverse of the logistic function

$$\pi(\mathbf{X}) = \frac{e^{\boldsymbol{\beta}\mathbf{X}}}{e^{\boldsymbol{\beta}\mathbf{X}}+1}. \tag{3}$$

3.1 Analysis

The analysis for this model was performed in R[5]. The model can be fit in R by using a response vector $\mathbf{Y}$, and corresponding covariate vectors in $\mathbf{X}$. After importing the data in R and assigning column names, the following command was used to perform the GLM regression:

```
model <-  glm(cbind(y1, y2) ~ x1 + x2 + x3 + x4,
              data = data, family = binomial(link = "logit"))
summary(model)
```

Since the response variable is binomial (proportion of searches originating from Google and proportion of searches not originating from Google), the binomial distribution was used. The data was collected for two separate time periods and the same model was run for both the time periods. Additionally, the model was also run by combining the data from both the time periods, as traditional statistical theory would suggest.

Recall that the response variables for the GLM model are log odds since 'logit' is used as the link function, and the coefficients have to be transformed by converting the raw coefficients using the exponential to obtain the odds. The following command in R achieves that:

```
exp(coefficients(model))
```

To test for the validity of our model, we used k-fold cross validation. In this technique, data is randomly permutated, split into k-folds and then divided into test and training sets [3]. The idea behind this powerful technique is to be able to predict how well the model fits to a hypothetical test set when an explicit test set is not available. In R, the DAAG package provides the cross validation function [4].

```
cross_validation <- cv.glm(data,  model,  K = 5)
```

The cross-validation misclassification error is used to summarize the fit of the model. For our model, we can expect to be inaccurate in our predictions 0.4 percent of the time with the new data.

[5] http://www.r-project.org/

4 Results

The GLM model for Google usage rates establishes a positive correlation between the Google search usage and the four independent variables. A summary of the model output is given below:

Model	Model 1	Model 2	Model 3
Median Household Income	1.00	1.00	1.00
Rate of Job Growth	1.00	1.00	1.00
Bachelor's Degree of Higher	17.30	17.70	17.77
Median Age	1.00	1.00	1.00

Model 1 is for the date range of June 26 to June 30, 2013, Model 2 is for the date range of July 20 to July 24, 2013 and Model 3 combines the data from both the date ranges. The odds ratio corresponding to Median Household Income in Model 1 is 1.00. This implies that increasing the Median Household Income by 1 unit will increase the odds of Google usage rate by 1.00, *ceterus peribus*. All of the coefficients are statistically significant. It needs to be stated implicitly, however, that the data lacks granularity (as the data is at the state level). Additionally, there may be some other important independent variables which can included.

The data was collected from two separate time periods. Specifically, the data was collected for five days in June and for five days in July. This was done to ensure that the data being analyzed was not an anomaly, i.e., that certain trends were not being observed due to the time period for which the data was being collected. Our initial validation checks for the raw data confirmed our belief that there were no irregularities in the data.

5 Conclusion

Our analysis points towards state-level evidence on the factors affecting Google usage rates. Google search usage rates are positively correlated with median income, rate of job growth, level of education and median age at the state-level. Moreover, there are further potential benefits to improve the understanding of Google usage rates between and within states.

Future research should expand on the model we have considered by including other variables which can affect Google usage rates, such as percent of population that is urban versus percent of population that is rural in each state. We strongly believe that a more systematic study of factors affecting Google usage rate, and additionally building a non-heuristic profile of a user more likely to use Google, will enable marketers and advertisers to monetize their investments more efficiently.

This analysis provides justification for the theoretical approaches used at Chitika to understand shifts in the online world.

References

[1] Chitika, Inc. Analyzing tablet usage share during the december 2012 holiday season. `http://chitika.com/files/Holiday_Share_White_Paper.pdf`, 2013. [Online; accessed 06-February-2013].

[2] Chitika, Inc. Google usage rates vary across browsers. http://chitika.com/insights/2012/google-usage-rates-vary-across-browsers/, 2013. [Online; accessed 28-August-2013].

[3] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, volume 14, pages 1137–1145, 1995.

[4] Brian Ripley. boot: Bootstrap Functions. http://cran.r-project.org/web/packages/boot/index.html/, 2013. [Online; accessed 03-March-2013].