Upworthy dataset evaluation: Length of a magazine's headline can affect the article's click-rate

YL3635

18/12/2020

Title

Upworthy dataset evaluation: Length of a magazine's headline can affect the article's click-rate

Author

Yifei Li

Date

December 20th, 2020

Code and data supporting this analysis is available at: https://github.com/amandayifei/sta304final (https://github.com/amandayifei/sta304final)

Abstract

The topic choice of this final report is option D, and this report aimed to test whether Length of a magazine's headline can affect the article's click-rate. After cleaning the original dataset, with creating new data columns including excerpt_ct,headline_ct,lede_ct,slug_ct and click_rate, this assignment intend to use Propensity Score Matching for the following analysis due to the property of the experiment: Observational Study. To reach this intent, a logistic regression model is built for the propensity score and a treated observation group(of those click_rate >= its 75% quantile) is built to make comparison with the original controlled group. Finally, a linear regression model shows a significant level of 'headline_ct'(length of the headline) influence to the click-rate of articles, which indicates a positive influence of headline length to the exposure(click-rate) of a article.

Keywords

A/B Test, Upworthy Research Archive, Observational Study, Propensity Score Matching

Introduction

While the Upworthy research archive from https://upworthy.natematias.com/index (https://upworthy.natematias.com/index) is a large A/B tests' open online dataset, and the dataset used for this assignment is requested from this Upworthy company. The dataset provide the imformation of the factors affecting the exposure rate(click_rate) and some other related imformation. The upworthy Research Archive is a dataset of headline A/B tests conducted by Upworthy between early 2013 and April 2015.

While this project aimed to find out whether the length of the headline can affect the click_rate of the article, so that it can be a useful imformation providing suggestions for reporters and editorial staffs when writing further articles.

The experiment is a observational study with A/B tests data. Therefore, direct regression model cannot apply. A Propensity Score Matching is used first to pair the alike outcome into group, and then Simple Linear Regression model is built to find out the significant level of the predictors.

The data was first cleaned by extracting the useful columns, change the char columns into length counting columns, and set up a mew variable click_rate=clicks/impressions. After that, A barrier of 0.020686 was counted for the whole data group to be put into the treated group, to compare with the original group and match pairs. Lastly, propensity score was calculated for each pair and a linear regression model was built to find out the treatment's significance. The variables to be used in this study are as follows: excerpt,headline,lede,slug,impression and clicks. new varaiable click_rate is set up(based on impression and clicks) as result variable, and the rest of the variable as predictors.

The result from the linear regression model shows a significant level of headline, slug and excerpt. Therefore, the result of the data analysis is positive. Headline length is affecting the click rate of a article.

Data

The dataset is downloaded from the Upworthy research archive(https://upworthy.natematias.com/index (https://upworthy.natematias.com/index)) which is a large A/B test open data platform shows investigation of A/B headline influences and headline related influences, such as excerpt and lede's. While this platform provide a chance for students and learners to do project learning and analysis on real data. However, the data is a bit not up-dated, therefore, after done the analysis, the useful further steps and suggestions the researchers purposed may cannot be considered in the great extent. In this project, the main purpose is to find out the whether there is a relationship between headline length and click rate exposure. The population of this experiment is 22666(which is also the population of exploratory dataset), viewing from the lines of the frame: raw_data. While the cleaned_data is the sample data with 20228 units. This experiment do not find respondent purposely, instead, all of the viewers from the internet who click the articles or is exposed to the articles will be the respondents. Therefore, the non-respondents are ignored and excluded from the experiment. The archive included aggregate results on the number of viewers who received a package and who click on the pavkage but does not include any individual information for the purpose of viewer differentiation(Data in the Upworthy Research Archieve).

During the cleaning process, the main things done are: set up 5 new variables: excerpt_ct,headline_ct,lede_ct,slug_ct,click_rate based on the original's. The first four variable is by counting the string length of the original variables: excerpt, headline, lede and slug. These four variables is four of the ways for upworthy team to change the articles. However, there is another variable called eyecatcher_id, which basically represent the article's picture's id. This variable is not counted in the following analysis due to it's variable type of string, which slow down the calculation in a great extent. However, this will be counted as a weakness since this variable may have casual inference or other relationship with the outcome. While the last vairable is based on the original variable clicks divided by the variable impressions. The view of the raw data is shown below, followed by a view of the cleaned data:

```
## # A tibble: 6 x 17
##      X1 created_at          updated_at          clickability_te… excerpt
##   <dbl> <dttm>              <dttm>              <chr>            <chr>
## 1     0 2014-11-20 06:43:16 2016-04-02 16:33:38 546d88fb84ad38b… Things…
## 2     1 2014-11-20 06:43:44 2016-04-02 16:25:54 546d88fb84ad38b… Things…
## 3     2 2014-11-20 06:44:59 2016-04-02 16:25:54 546d88fb84ad38b… Things…
## 4     3 2014-11-20 06:54:36 2016-04-02 16:25:54 546d902c26714c6… Things…
## 5     4 2014-11-20 06:54:57 2016-04-02 16:31:45 546d902c26714c6… Things…
## 6     5 2014-11-20 06:55:07 2016-04-02 16:25:54 546d902c26714c6… Things…
## # … with 12 more variables: headline <chr>, lede <chr>, slug <chr>,
## #   eyecatcher_id <chr>, impressions <dbl>, clicks <dbl>, significance <dbl>,
## #   first_place <lgl>, winner <lgl>, share_text <chr>, square <chr>,
## #   test_week <dbl>
```

```
## # A tibble: 6 x 5
##   excerpt_ct headline_ct lede_ct slug_ct click_rate
##        <int>       <int>   <int>   <int>      <dbl>
## 1         32          84     130      83     0.0491
## 2         32          84     130      83     0.0402
## 3         32          84     130      83     0.0356
## 4         32          43     294      50     0.0255
## 5         32          43     294      50     0.0342
## 6         32          43     294      50     0.0290
```

Model

The software used to run this whole process is R studio, In order to analyze this A/B test observational experiment, propensity score matching method is used overall in the model build up process. While first construct a logistic regression model, followed by a forecast of the dataset, and create matches for the alike pairs. All of the predictors, as well as the outcome in these propensity score model are all in numeric, since numeric data is easier to be fitted in propensity score model than char data(the original data variables before cleaning). While for the variable that cannot be fixed into a numeric: eyecatcher_id is already excluded from the predictors since the data amount of eyecatcher_id is too large for a glm model to analyze.

In order to set up a 'treated' group, there has to be a barrier for the qualified click_rate to be considered as relatively high click rate among the group. While belows is a summary of the 'click_rate' column. As we can see from the summary: median of click_rate is 0.012837, 3rd quantile of click_rate is 0.020686. So 0.020686 will be set as the barrier for putting units into the treated group in order to make matches.

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000000 0.007651 0.012845 0.016105 0.020708 0.136063
```

A view of the matched dataset is shown below:

```
##   excerpt_ct headline_ct lede_ct slug_ct  click_rate     .fitted     .se.fit
## 1         96          40     287      95 0.022305654 0.003217741 0.002662894
## 2         32          59     191      90 0.007550336 0.005531791 0.002455258
## 3         32          66     294      87 0.030349014 0.008197256 0.002373903
## 4         32          74     267      94 0.021818182 0.008524770 0.002304114
## 5         32          83      40      97 0.003956624 0.009647223 0.001854892
## 6         32          64     337      80 0.015468608 0.009821807 0.002162749
##   cnts
## 1    1
## 2    1
## 3    1
## 4    1
## 5    1
## 6    1
```

Propensity score linear regression model is built, and the result is shown below:

	(1)
(Intercept)	0.019 ***
	(0.001)
excerpt_ct	0.000 ***
	(0.000)
headline_ct	0.000 ***
	(0.000)
lede_ct	0.000
	(0.000)

slug_ct	-0.000 ***
	(0.000)
N	10134
R2	0.007
logLik	28372.810
AIC	-56733.619

*** p < 0.001; ** p < 0.01; * p < 0.05.

The alternative model of this propensityscore matching model, might be a direct linear regression model. However, the amount of A/B test data for a linear regression model can be too large to be analyzed. Further, a Simple Linear Regression model cannot exclude the potential of correlation effects between variables.

Results

From the above process of establishing and analyzing the model, we can find that: first of all, the data collection process and hypothesis of this set of upworthy are all without problems, which are in line with the rules. Because there is no obvious correlation between the data, or there is no great degree of confusion in the data. At the same time, we can see that this group of data is very complete at the time of collection. Because when Na value is excluded, most of the rows of data are not removed.

Further, the summary table of the linear regression model is shown as following:

```
##
## Call:
## lm(formula = click_rate ~ excerpt_ct + headline_ct + lede_ct +
##     slug_ct, data = data_matched)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.026545 -0.011080 -0.001694  0.006747  0.111371
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.877e-02  9.114e-04  20.597  < 2e-16 ***
## excerpt_ct   3.697e-05  5.459e-06   6.774 1.33e-11 ***
## headline_ct  2.014e-04  4.213e-05   4.781 1.77e-06 ***
## lede_ct      1.019e-06  7.342e-07   1.388    0.165
## slug_ct     -1.817e-04  4.079e-05  -4.454 8.53e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01472 on 10129 degrees of freedom
## Multiple R-squared:  0.006781,   Adjusted R-squared:  0.006389
## F-statistic: 17.29 on 4 and 10129 DF,  p-value: 3.824e-14
```

From this linear regression model, we can observe that β_0 with <2e-16, β_1 with 1.05e-13, β_2 with 6.75e-06, β_4 with 1.90e-05 all shows a significant level of p-value, which represent that the intercept, excerpt_ct, headline_ct and slug_ct all shows influences to the outcome: click_rate. For β_0, β_1, β_2, β_4, the H_0 will be rejected and for β_3, the p-value of 0.165 > 0.05 shows a failure to reject H_0, which is saying, excerpt, headline and slug shows a significant level of influences to click-rate. However, the influences of slug_ct is not that obvious.

Discussion

The whole project does something as follows. Download AB test data about headline for article exposure from upworth website. Then, a model is established to analyze whether the length of the headline will affect the exposure rate of the article. According to the properties of AB test and observational data, the selected model of this project is the linear expression model processed by the project score.

The most significant weakness from this project is eyecatcher column was deleted. eyecatcher_ ID is a very important variable in the original dataset. However, because of its own nature is a string with many characters, it will not be processed in the model because of too much data, so it has to be deleted. But it is undeniable that its lack will make the assessment of the whole project vulnerable: it can not be ruled out that it has correlation with other variables or other types of relationships. So it affects the analysis of the whole experiment.

References

Matias, J. Nathan, et al. The Upworthy Research Archive. upworthy.natematias.com/index.

Data in the Upworthy Research Archive. (n.d.). Retrieved December 23, 2020, from https://upworthy.natematias.com/about-the-archive (https://upworthy.natematias.com/about-the-archive)