Identifying Irregular Respondents Through Response Time

Changrong Xiao, Yuxin Wu, Yunlu Chen and Zhiyong Zhang
Department of Psychology, University of Notre Dame, IN 46556

UNIVERSITY OF NOTRE DAME

Identifying and controlling for poor respondents in online surveys is a difficult and non-trivial task in study design and data collection.

Data Sources:

- Our data came from an online survey consisting of six blocks of questions. We mainly focused on the friendship block which recorded relationships among the students.
- Information we collected included the answer of each question as well as the timestamp when the student clicked a select option.
- Eleven ids were already confirmed to be problematic, and they were required to take the survey again.

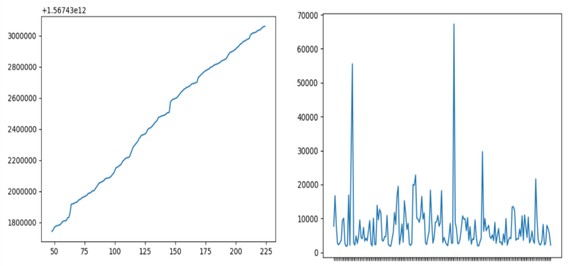
Objectives:

- To identify problematic responses in this online survey.
- To inform future data collection.

Data Preprocessing

In order to figure out how much time spent to answer each question for each student, we extracted the timestamps, **took differences**, and obtained the response time series.

Also, there were some dirty data we need to deal with, for example, the NA values, the negative response time and the outliers. We simply replaced them with **average** response time of the question, which proved to be reasonable.



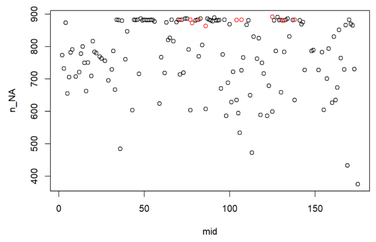
Figures 1-2: timestamps of each question (left) and response time of each question (right)

For the answer data, we only filled the NA data with average values, since it was cleaner than the response time data. Based on these two types of preprocessed data, we were able to analyze the respondent quality.

Methods and Results

Exploratory Data Analysis

In the friendship block, we intuitively found a pattern that those careless respondents tended to keep selecting the option ("*I didn't hear of this student*") when they were given a student's name so that they didn't need to answer the remaining 5 questions about this student. In this case, many missing response times were found in the friendship block.



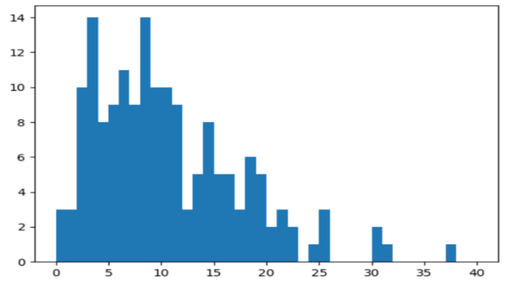
Figures 3: missing values of each id in the friendship block (red dots: 11 confirmed problematic ids)

Through this finding, we assumed that each respondent knew at least 10 other students and based on this criterion, we identified 43 engaged respondents.

Logical Consistency In Responses

Inspired by the property of this network study, we analyzed data quality from the perspective of the network's adjacency matrix. In the normal cases, if student A knows student B, then student B should also know student A. We called this situation "**consistent**".

So we counted the number of inconsistent cases for each student. In other words, compared the differences between the rows and columns of the binary adjacency matrix. If the number of inconsistent cases is greater than **20**, we regarded this id as "abnormal".



Figures 4: distribution of the number of inconsistent cases

Clustering Using Dynamic Time Warping

Unsupervised method **clustering** was also applied to further explore problematic respondents. The key of clustering method is to determine a proper similarity measure function to divide the data into clusters. We used a common criterion in time series clustering field, **dynamic time warping**, as the similarity function.

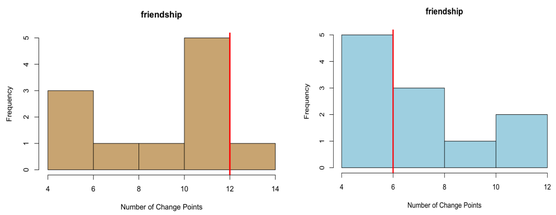
$$W = w_1, \ldots, w_K, \max(m,n) \le K < M + n - 1$$

$$DTW(Q,C) = \min\{(\sqrt{\textstyle\sum_{k=1}^{K} w_k})/K\}$$

Change Point Analysis

We used **change point analysis** to determine the change points in the response time series and estimated the time of each change.

We initially hypothesized that bad respondents may have fewer change points detected due to their invalidity. However, contrary to our hypothesis, we found that those engaged respondents had fewer change points, also, after re-submitting the survey, fewer change points were detected in the response time series of 11 confirmed irregular respondents.

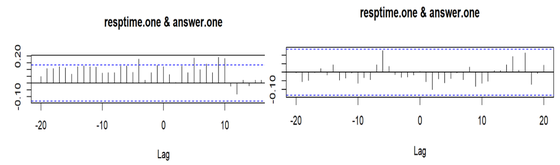


Figures 5-6: number of change points in the response time series of 11 confirmed irregular respondents (left: before; right: after)

Cross Correlation Analysis

Cross correlation function (CCF) is a measure of similarity between two series. In CCF, the lag refers to how far the series are offset, and its sign determines which series is shifted. In this study, we checked the cross correlation between response time series and answer series.

We found that the values of cross correlation were more likely to be negative for disengaged respondents.

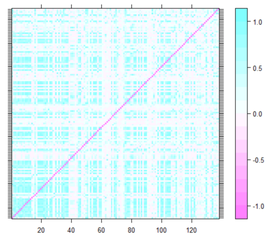


Figures 7-8: cross correlation function between response time series and answer series (left: engaged respondents; right: disengaged respondents)

Partial Correlation Analysis

Partial correlation measures the degree of association between two variables, ruling out the effect of confounding variables. We calculated the correlation between two response time series but wanted to remove the effect of their answer series.

We got the correlation matrix visualizing the correlations between two response time series. Moreover, we set a threshold (>0.9) for the correlations so that we could use the confirmed problematic respondents to identify more irregular respondents.



Figures 9: correlation matrix for the response time series

Future Work

- More work should be done to validate any thresholds that are developed.
- Incorporate different methods we used to identify the irregular respondents.

References

1. James Soland, Steven L. Wise & Lingyun Gao (2019) Identifying Disengaged Survey Responses: New Evidence Using Response Time Metadata, Applied Measurement in Education, 32:2, 151-165

Acknowledgements

Lab For BIG DATA Methodology
UNIVERSITY OF NOTRE DAME