Exploring Factors Affecting Non-response in Self-administered Survey Using Google Trends

Xiaoyi Deng

2019-12-13

Contents

Background

Almost all types of survey methods are experiencing a declining response rate, and low response rates have been a concern for survey researchers nowadays. Even though it is increasingly clear that low response rate does not necessarily associated with nonresponse bias of survey estimates(R. M. Groves 2006), response rates is still one of the most commonly used quality indicators in the field and it is held to be the most important criterion of survey quality with the exception of the sample size. Groves et al. (B. Groves R. M. 2008) proposed several alternative approaches as survey quality indicators at both survey level and estimate level that could outperform response rates, but response rate is still the most widely used survey quality indicator because of its simplicity and transparency.

$$Bias(\bar{y}_r) = \bar{Y}_r - \bar{Y}$$

$$Bias(\bar{y}_r) = \frac{M}{N}(\bar{Y}_r - \bar{Y}_m)$$

In details, from the non-response bias function, we learnt that bias exists only when respondents and non-respondents answer the questions differently, which means it is a waste to bring in more potential non-respondents if they will answer the same as respondents.

Moreover, from Total Survey Error perspective, measurement error and non-response error are the two major components of total error. Studies of total survey error provide some evidence that it is worthwhile to bring in more respondents to surveys. For example, according to Olson's research, bring in respondents with lower response propensity, or to say those who response with higher persuasion efforts is worthwhile in reduing non-response bias(Olson 2006). Sakshaug, Yan, & Tourangeau also found that for social undesirable questions, measurement error can be larger than non-response bias, and bringing in more respondents can also be worthwhile in terms of reduce non-response bias(Sakshaug 2010) .

To achieve higher reponse rates, we need a better understanding of factors affecting participation behavior. Groves et al. categorized factors that influence participation in self-administered surveys into societal level factors, characteristics of the sample person, and attributes of the survey design(C. Groves R. M. 1992). While lots of researchers studied factors in the last two categories, not enough attention has been paid to societal level factors. According to Keusch , Brick & William , potential societal level factors that affects participation include survey fatigue, culture, busyness, barrier to intrusion, social isolation, etc.(Keusch 2015)(Brick 2013). Among all societal level factors, I am especially interested in barrier to intrusion and privacy concerns given that barrier to intrusion is a direct cause of non-contact and it could be the result of survey fatigue, social isolation and also privacy concerns, while privacy concerns or to say fear of scam is another topic I am interested in. To improve our understanding of the influence of barrier to intrusion and privacy concerns on response rate in self-administered surveys. This study investigates the popularity of topics that represent barrier to intrusion and privacy concerns of different states from 2013 to 2017 using google trends, and explored the relationship between the popularity of topics and response rate of a federal large-scale national survey, American Community Survey (ACS). Hopefully, this study can help researchers and practitioners to learn more about the effect of barrier to intrusion on response rate, and also help make informed decisions to deal with low response rates.

Research hypothesis

Hypothesis 1: Response rate decreases over time in ACS; Refusal rate and noncontact rate increase over time in ACS.

Hypothesis 2: The higher popularity of topics that represent barrier to intrusion or privacy concern in google trends, the lower the response rate, higher refusal rate and non-contact rate.

Method

Dataset and data processing

American Community Survey(ACS): response rate, demographic characteristics

ACS is a federal national survey that used mixed mode data collection that start with a mail request to respond via email, a second mailing of paper questionnaire, a follow up computer assisted telephone interviewing (CATI) when a telephone number is available, if still unable to reach respondent, the address may be selected for computer-assisted personal interviewing (CAPI). Since the initial two contacts are using self-administered survey mode, use ACS as an example of self-administered survey is reasonable.

The ACS response rates by state from 2013 to 2017 were downloaded from the ACS website https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/response-rates/, detail demographic characteristics of each State is available via CensusAPI in R, both response rate and demographic characteristics of each state are weighted. As a result of the 2013 government shutdown, less follow up was done for 2013 October panel, the 2013 response rate is about 7 point lower than other years.

The ACS response rate data also included reasons for noninterviews, including refusal and serveral reasons for non-contact like unable to locate, no one home, temporarily absent. So in addition to reponse rate, I added refusal rate and non-contact rate as outcome variables. State level demographic variables include gender, race, ethnicity, education, income, native-born or foreign-born, age distribution by gender, to make sure they are comaprable across states, percentage instead of total number were used for variables except for age median and income median.

Gtrends: popularity of topics that represent barrier to intrusion, privacy concern

Google Trends is a tool that allows us to explore the popularity of topics for different regions over time based on Google search queries. To represent privacy concerns, which mainly focused on fear of scam, topics including 'cybersecurity','fraud',and 'scam' were used. "spam","call blocking", "whitelist","blacklist","robocalls","donotcall","spoofing" were used to represent barriers to intrusion.

Popularity of the ten topics from 2013 to 2017 by state were downloaded from Gtrends. Those topics were generted by starting with key topics 'cybersecurity' and 'call blocking' and then including relevant topics suggested by google trends. 'donotcall' and 'call blocking' are excluded because there are too many missing values (>20%).

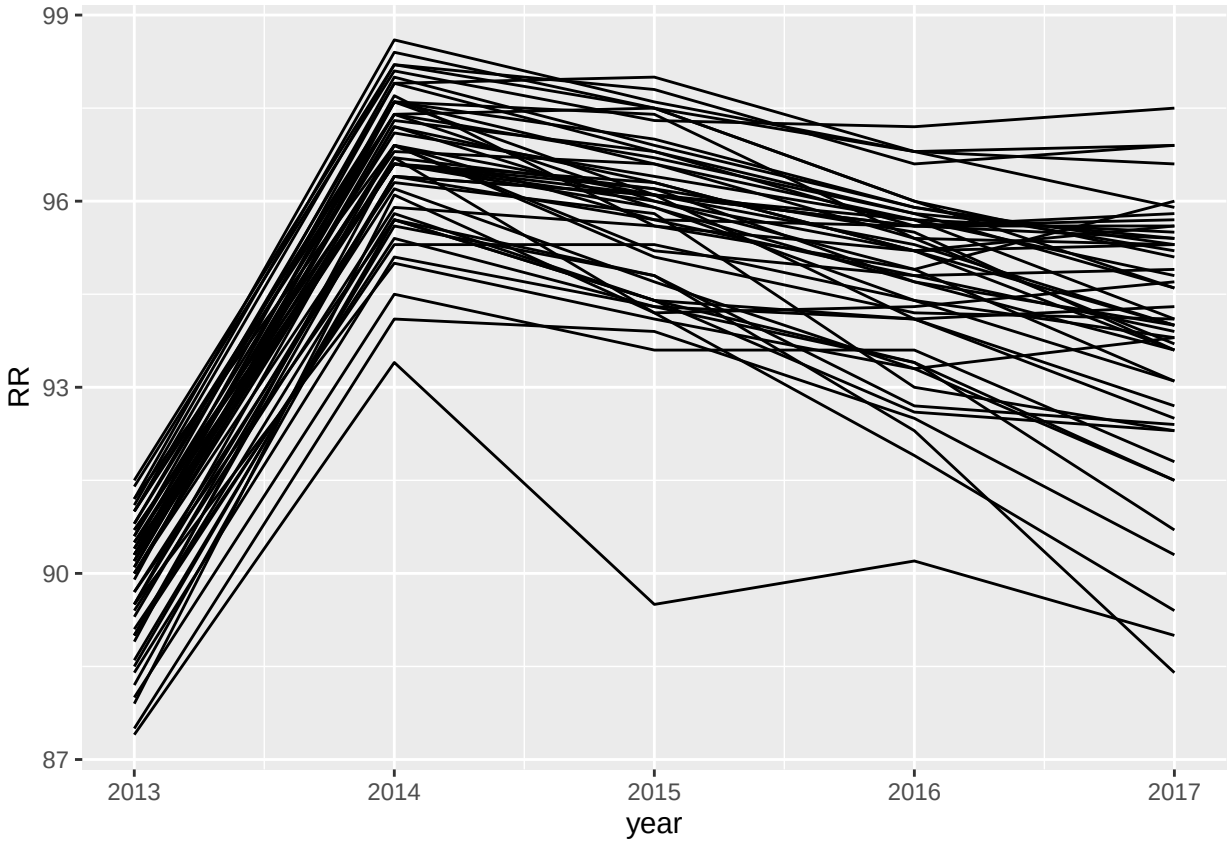
The following self-defined function was used to get Gtrends data.

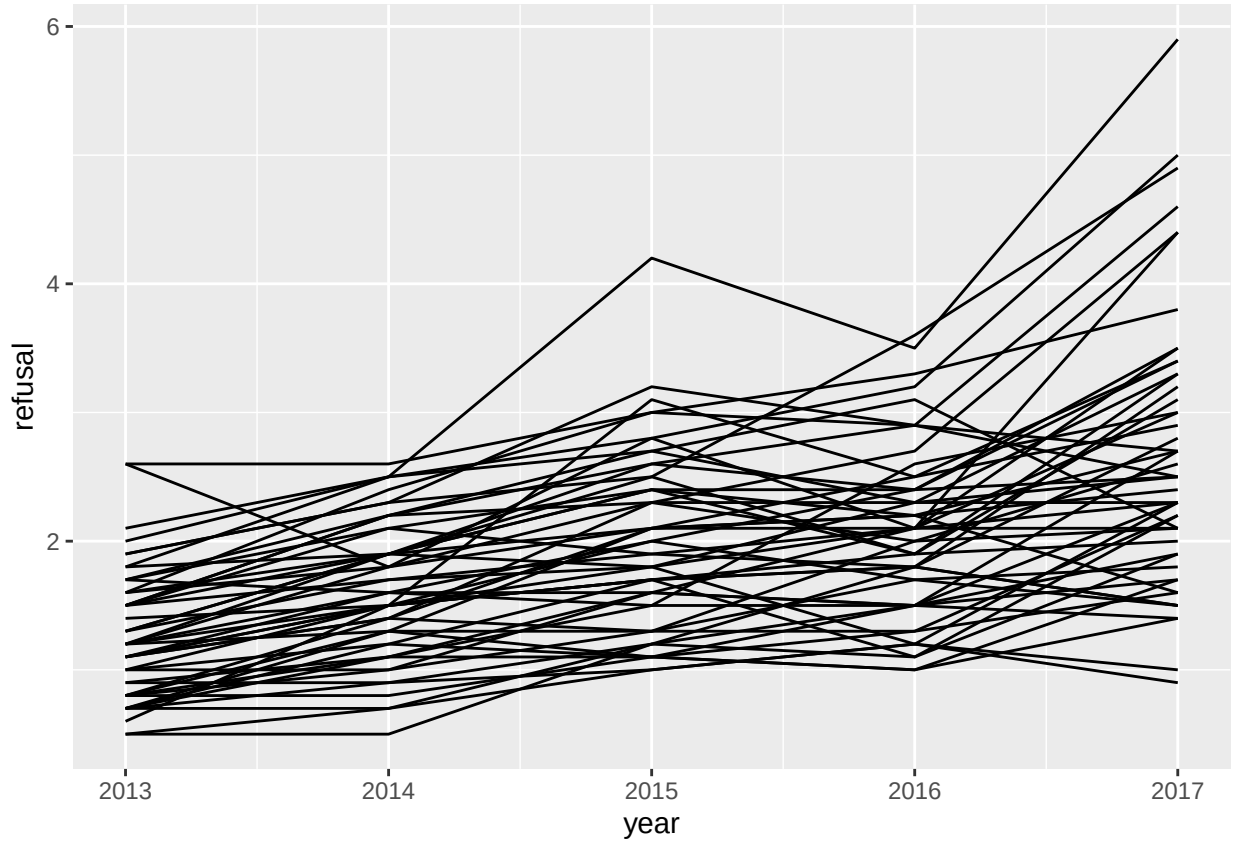
```
keywordhit<-function(i){
res<-gtrends(c("cybersecurity","fraud","scam","spam","call blocking"), geo="US",time=i,low_search_volume
res_region<-as.tibble(res$interest_by_region)
res_region_w<-spread(res_region, key = keyword, value = hits)
res2<-gtrends(c("whitelist","blacklist","robocalls","donotcall","spoofing"), geo="US",time=i,low_search_
res_region2<-as.tibble(res2$interest_by_region)
res_region_w2<-spread(res_region2, key = keyword, value = hits)
res_region_all<-cbind(res_region_w,res_region_w2)
return(res_region_all)
}
```

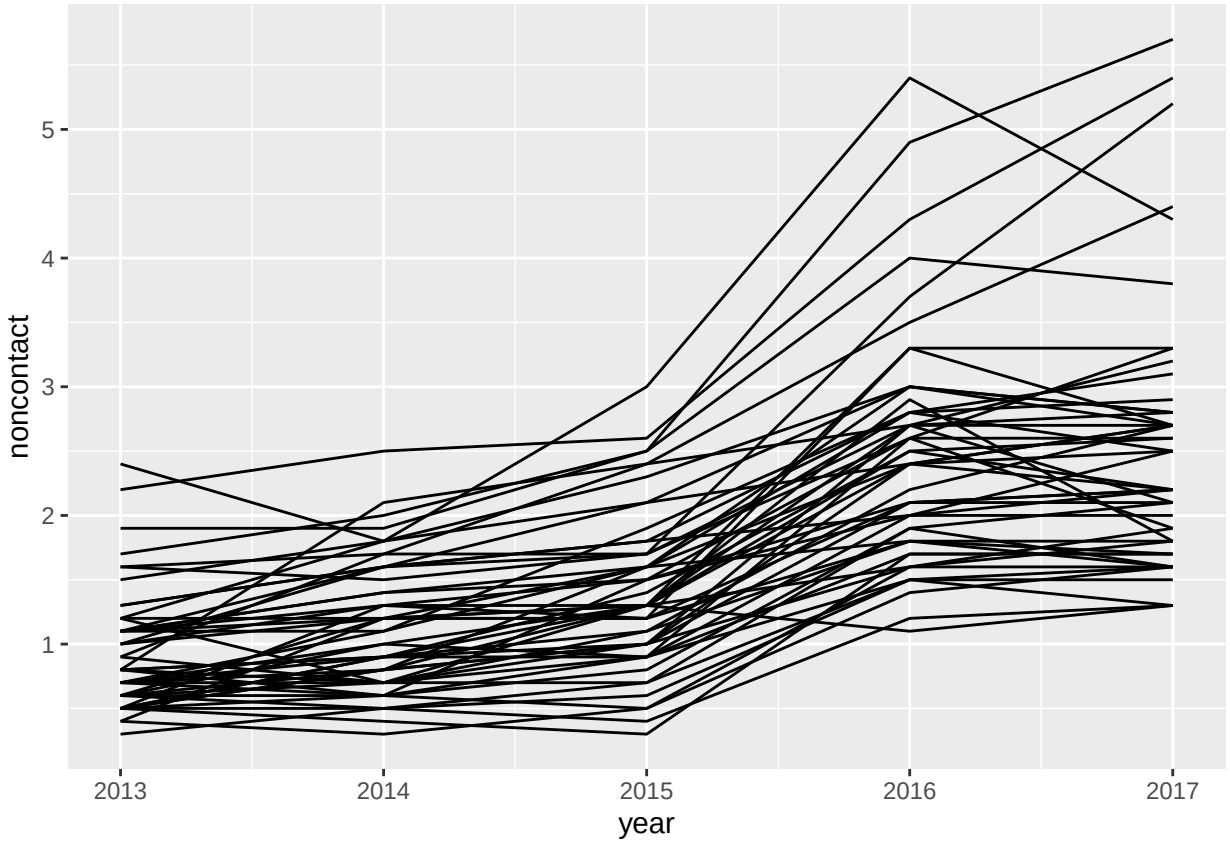
Analysis

To explore the change of response rate by state over time, and factors that could help explain the differences. I fitted the data with several mixed effect models given that each state have different characteristics. The first set of models I fitted are between time and response rate indicators to test hypothesis 1. After that, I checked the correlation between popularity of topics with response rate indicators and add those topics that are significant into the first set of models to test hypotheis 2. I also added demographic characteristic variables into mixed models to see if it can help explain the difference.

Results




6
4
2
refusal
2013
2014
2015
2016
2017
year



From the three spaghetti plots of response rate indicators and year, we can observe that refusal rate and noncontact rate increase overtime for almost all states from 2013 to 2017, and different states could have different intercept and slope. Meanwhile, reponse rate declines from 2014, extremely low response rate in 2013 is a result of the 2013 government shutdown, to aviod analysis being skewed by 2013 response rate data, response rate were excluded from outcome variables for further analysis.

```
model1_refusal<-lmer(refusal ~ time + (1|State), REML = T, data = alldata13_17)
summary(model1_refusal)
rand(model1_refusal)
model1_noncontact<-lmer(noncontact ~ time + (1|State), REML = T, data = alldata13_17)
summary(model1_noncontact)
rand(model1_noncontact)
```

The first set of mixed models indicated that refusal rate and noncontact rate did increase over time. In 2013,the average refusal rate across all states is 0.95 percent, and average non-contact rate across all states on is 0.33 percent; each year refusal rate increases by 0.32 percent, noncontact rate increases by 0.45 percent.

```
model2_refusal<-lmer(refusal ~ time+ cybersecurity +fraud +spoofing + whitelist+ (1|State), REML = T, da
summary(model2_refusal)
rand(model2_refusal)

model2_noncontact<-lmer(noncontact ~ time+ cybersecurity +fraud +spoofing +robocalls + (1|State), REML =
summary(model2_noncontact)
rand(model2_noncontact)
```

From correlation statistics among refusal rate and non-contact rate, I found the popularity of several topics correlated with refusal rate and noncontact rate, spam is postiviely correlated with both, which is opposite to the hypothesis, the reason is that 'spam' not only represents email filter, but also represents a brand of

canned meat product, so the result is not reliable. Other topics that are negatively correlated with refusal rate are cybersecurity(r=0.4), fraud(r=0.3), spoofing(r=0.24), and whitelist(r=0.14,). After adding the fixed effect of the four topic into the mixed model, it turns out only cybersecurity can help exaplain the increasing refusal rate, and also as cybersecurity gets more popular the refusal rate increases. As for noncontact rate, cybersecurity(r=0.43), fraud(r=0.29), spoofing(r=0.22),and robocalls(r=0.28) are postively correlated with it. After adding the fixed effect of the four topic into the mixed model, I found that both cybersecurity and spoofing can partially exaplain increasing refusal rate, and the more popular these topic gets, the higher noncontact rate.

```
model3_refusal<-lmer(refusal ~ time+ hispanic+youngadults+bachelor_above+foreign_born+income13_16+under
+(1|State), REML = T, data = alldata13_17)
summary(model3_refusal)
rand(model3_refusal)
model4_refusal<-lmer(refusal ~ time+income13_16+log(incomemedian)
+(1|State), REML = T, data = alldata13_17)
summary(model4_refusal)
rand(model4_refusal)
```

```
model3_noncontact<-lmer(noncontact ~ time+ black+youngadults+bachelor_above+foreign_born+income13_16+un
summary(model3_noncontact)
rand(model3_noncontact)
```

Most of the selected demographic characteristics significantly correlates with non-contact rate and refusal rate. In details, refusal rate postively correlates with 'income_median'(r=.32), 'Hispanic'(r=.18),'maleyoungadults'(r=.24) and 'femaleyoungadults'(r=.28),'bachelor_above'(r=.32), 'foreign_born'(r=.23), 'income13_16'(r=.37); and negatively correlated with 'maleunder18'(r=-.15) and 'femaleunder18'(r=-.15), 'White'(r=-.13), 'income1_4'(r=-.19),'income5_8'(r=-.36),'income9_12'(r=-.29). Before adding these variables into mixed models I combined the age distribution by gender variables and got 'under18' which is negatively correlated with refusal rate, and 'youngadults' which postively correlated with refusal rate. To aviod singular model, if the total of two probability variable equals one like nonhispanic and hispanic, only one variable enters into the model. The results of the mixed model found that only income variables have significant fixed effects, the higher percentage of people with household income higher than 75,000(income13_17) the higer refusal rate, the higher odds ratio of income median, the lower refusal rate.

The correlation statistics for noncontact rate is very similar to that of refusal rate, the only difference is that 'Black' replaced 'Hispanic' to be the only race percentage variable that postively correlated with noncontact rate. But the result of the mixed model found that none of these variables have a significant fixed effect.

Conclusion

In summary, we confirmed that response rate decreases, refusal rate and noncontact rate increase over time in ACS. Soceital-level factor like privacy concerns could potentialy affect refusal rate, which means refusal rate can be higher with more people having strong privacy concerns in a state. Some state-level characteristics like percentage of people with high income can also predict response rate. I also developed one bubble plots to demostrate the relationship between percentage of high income and refusal rate, and two bubble plots to show the relationship between societal level factors and refusal non-contact rate. (bubble plots only viewable in Rmd)

Limitation

Findings from this study probably cannot be generated to other self-administered surveys since that ACS is a federal survey that have higher reponse rate comparing to other surveys. Introducing one more data from non-federal survey might help address this issue. The devolpment of topics representing societal level factors do not have enough literature support or advices from expert, there could be better options.

References

Brick, & Williams, J. M. 2013. "Explaining Rising Nonresponse Rates in Cross-Sectional Surveys." The ANNALS of the American academy of political; social science, 645(1), 36-59.

Groves, Brick, R. M. 2008. "Issues Facing the Field: Alternative Practical Measures of Representativeness of Survey Respondent Pools." Survey Practice, 1(3).

Groves, Cialdini, R. M. 1992. "Understanding the Decision to Participate in a Survey." Public opinion quarterly, 56(4), 475-495.

Groves, R. M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." Public opinion quarterly, 70(5), 646-675.

Keusch, F. 2015. "Why Do People Participate in Web Surveys? Applying Survey Participation Theory to Internet Survey Data Collection." Management review quarterly, 65(3), 183-216.

Olson, K. 2006. "Survey Participation, Nonresponse Bias, Measurement Error Bias, and Total Bias." International Journal of Public Opinion Quarterly, 70(5), 737-758.

Sakshaug, Yan, J. W. 2010. "Nonresponse Error, Measurement Error, and Mode of Data Collection: Tradeoffs in a Multi-Mode Survey of Sensitive and Non-Sensitive Items." Public Opinion Quarterly, 74(5), 907-933.