Representative Targeting

Reaching out to representative samples is a major issue within the context of opinion polls: biased samples lead to biased conclusions. In addition to this major issue, today, we are witnessing a shift in opinion polls. The methods used in the past (face-to-face and telephone surveys) become more disadvantageous compared to internet-based methods. Opinion polls conducted via the Internet are becoming more and more popular applications due to their low cost and high speed. However, these new applications lead to a dilemma in terms of reaching representative samples because social media users constitute a highly biased section of the society. Therefore, the persons reached do not reflect the general society. In this context, multilevel regression analysis has come to the fore in adjusting the data taken from social media to real society.

One of the ways to reach generalizable information from new data sources such as social media or CDR is to apply the amplified learning method (Salganik, 2019), that is combining a survey and a new data source. A groundbreaking example of this method is Blumenstock, Cadamuro, and On's (2015) study performed by combining CDR and survey data. To date, various studies have been carried out using social media to make society-wide generalizations. Extracting a representative sample of users is a critical phase in these studies.

This work has been put forward to help amplified learning studies, especially the ones using social media data. Due to the characteristics of social media data it has undermentioned arguments:

1. Activity Levels: Social media users vary in the degree of use of the relevant platform. For example, one user may have tweeted dozens of times in the last month, while another may not tweet at all in the last year. This affects the quality of the work to be done using the relevant users. Therefore, it is important to identify users at different activity levels according to the characteristics of the subject being studied. In this study, three different activity levels were defined. The algorithm tries to fill the quota in the relevant category by looking at three groups respectively. This can be used, for example, to retrieve users from active, semi-active and inactive groups, respectively.

2) gen_pop_filename: This argument refers to the frequency of the general population on the relevant variables.

3) count_column_name: This argument refers to the column indicating the number of people in the relevant categories.

4) desired_sample_size: This argument refers to the desired sample size.

5) relevant_variable: These three arguments represent the variables to be taken as indicators while creating a representative sample of the general population. Three variables have been defined.

References

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, *350*(6264), 1073-1076.

Salganik, M. J. (2019). *Bit by bit: Social research in the digital age*. Princeton University Press.