September 24, 2019

Partial population weighting for improving panel measurement

Anonymous Anonymous

Partial population weighting for improving panel measurement

Abstract

The present disclosure describes a system for utilizing a weighting scheme for ensuring that market research panels adequately represent segments of a population on whom large-scale data is available on a social media website. The system also ensures that market-sensitive information about social media users available at the social media website is not revealed to panel partners. A large number of weighting variables, held by the social media website, that have the greatest impact on answers are selected. A principal component analysis (PCA) algorithm is utilized on a dataset associated with all of the weighting variables for the full population of interest, in order to obtain principal components for the population. Each PC is a linear transformation of all of the weighting variables, where weights w_{ij} are assigned to each of the weighting variables for $i, j = 1, \dots d$ (d = the number of the weighting variables). Thus a value of each PC is computed for all matched panelists, being as they are members of the population of interest. The principal components are calculated using an eigen decomposition of a variance-covariance matrix of a d-dimensional weighting variables vector. The first few principal components are measured in terms of their ability to explain a total variance of the population and the number of PCs sent out to panel partners is then determined. Not all of the PCs are sent out to panel partners so as to preserve a privacy of the sensitive dataset. Partial of the PCs are sent to the panel partners, along with a mean value of each PC among the population as a whole. The panel partners then use these principal components to construct the weights, which use some of the information contained in the social media website's weighting variables, without the panel partner gaining direct access to that information. The panel partners' own weights are then augmented with the weights computed using the principal components.

Problem statement

It often happens that market research panels are representative of only a particular population e.g. internet users, all main shoppers, all adults 13+, etc. because of a recruitment method. This leads to underrepresentation and non-coverage of certain sections of the population. For instance, there may be too many men and not enough women completing a survey, or too many young people and not enough old people. In these scenarios, weighting of the panelists is required to ensure accurate representativeness. Otherwise, it might generate a biased dataset. It is also very important to choose weighting variables in an adequate manner. The weighting variables (e.g. demographic variables) that appear to have the greatest impact on the answers need to be identified. For example, if different age groups are giving similar answers, then choosing age as a weighting variable might not be a good idea. A variable can only be used for weighting if the panel partner has access to both individual values of that variable as well as information about a distribution of that variable in the population. Social media websites have access to both of these for their user bases, which in turn can comprise large segments of the population. However, the distribution of certain weighting variables, as well as individual values of those variables, may be sensitive information for the social media website and therefore not shareable to the panel partners. Meanwhile, the panel partners may not be comfortable in sharing their data with the social media website. So, an intermediate solution is required to address this issue.

System and working

The present disclosure solves the above problem by selecting a large number of weighting variables (for example, 10+) in the dataset held by a social media website. The dataset is created based on answers given by panelists to survey questions. Also, a panel partner contributes in eliminating underrepresentation and non-coverage of certain sections of population in the survey corresponding to data they have been able to collect from the panelists directly. The large number of weighting variables are chosen so that the chances of a single principal component (PC) being highly correlated with a single weighting variable are low. A principal component analysis (PCA) algorithm generates principal components, such that each PC is a linear transformation of all of the weighting variables (i.e. the weighting variables are input variables in the principal component analysis).

For each user in the full population of interest on the social media website,

- Let $X = (x_1, \dots x_d)^T$ denote a d-dimensional weighting variables vector with variance-covariance matrix Σ
- The goal of the PCA algorithm is to construct linear combinations

  $p_i = \sum_{j=i}^{d} w_{ij} * x_j$ for $i = 1, \dots . d$, where w_{ij} denote weights assigned to each of the weighting variables x_j for $j = 1, \dots . d$, in such a way that
  - the p_i's are orthogonal, so that $E[p_i, p_j] = 0$ for $i \neq j$, and

- the p_i's are ordered in such a way that $p_1$explains a largest percentage of a total variance of the population and each pi explains the largest percentage of the total variance that has not already been explained by $p_1,,,,p_{i-1}$.

It is a common practice to apply the PCA algorithm to the normalized weighting variables so that $E[x_i] = 0$; $Var(x_i) = 1$. This is achieved by subtracting means from the original weighting variables and dividing by their standard deviations to ensure that no single component of the X can influence the analysis by virtue of measurement units of that component.

The calculation of the principal components is performed using an eigen decomposition of the variance-covariance matrix Σ, which is a square matrix. Since variance-covariance matrix Σ is symmetric, the eigen decomposition implies that any symmetric matrix Σ can be written as:

$$\Sigma = \Gamma\Delta\Gamma^T$$

where,

- Δ is a diagonal matrix, $diag(\lambda_1,,,,\lambda_d)$, of eigen values of Σ without loss of generality ordered so that $\lambda_1 \geq \lambda_2 \geq \cdots . \geq \lambda_d$, and
- Γ is an orthogonal matrix with i^{th} column of Γcontaining i^{th} standardized eigen-vector, γ_i of Σ, whereas
    - "standardized" means ${\gamma_i}^T\gamma_i = 1$
    - orthogonality of Γ implies $\Gamma\Gamma^T = \Gamma^T\Gamma = I_d$ and $\Gamma^T = \Gamma^{-1}$

Further, the variance-covariance matrix Σ is a positive semi-definiteness matrix, which implies that $\lambda_i \geq 0$ for all $i = 1, \ldots d$. The principal components of the X are given by $P = (p_1, \ldots . p_d)$, which satisfies $P = \Gamma^T X$.

We can measure the ability of the first few principal components to explain the total variance:

$$\sum_{i=1}^{d} Var(p_i) = \sum_{i=1}^{d} \lambda_i = trace(\Sigma) = \sum_{i=1}^{d} var(x_i)$$

if we take $\sum_{i=1}^{d} Var(p_i) = \sum_{i=1}^{d} var(x_i)$ to measure the total variance, then by above equation we can interpret $\frac{\sum_{i=1}^{k} \lambda i}{\sum_{i=1}^{d} \lambda i}$ as a percentage of the total variance explained by first k principal components.

After the above calculation process, for each matched panelist between social media website and the panel partner, the social media website will send over values of the first k principal components $(p_i, \ldots p_k)$

together with mean of each of the first k principal components (whereas, $E[P] = 0$, since $E[X] = 0$) to the panel partner.

The panel partners construct weights for each matched panelist to balance their first k principal components towards the mean (0) of the principal components among the whole population of interest. This is achieved by an entropy-balancing method. The entropy balancing method is a modern ML-based approach to balance sample and target population by maximizing an entropy of the weights under some pre-specified balancing constraints. The end result is similar to a traditional propensity score modelling in which a weight is assigned to each unit in the sample population, thereby making them representative with respect to the weighting variables (here the principal components) of units in the target population (here the social media website population of interest). However, the panel partners do not know precise values of the weighting variables in the population. The panel partners' own weights, computed using the entropy-balancing method, are augmented to the weights computed using the principal components. This way, the panel partners run analysis on a weighted panel set, which is representative of the social media population without knowing market-sensitive information about the social media users.

Additional Embodiments

Each PC is a score that is correlated with the weighting variables. The panel partners could construct something close to actual population value of a weighting variable if a PC is highly correlated (>0.8) with that weighting variable. To mitigate this risk, each PC is checked for correlation with each weighting variable before sending to the panel partners. The number of weighting variables is increased further, if a PC is highly correlated with a single weighting variable. A probability of a PC being highly correlated with a single weighting variable decreases as the number of weighting variables increases.

Conclusion

Weights are required to correct for imperfections in a sample that might lead to bias and other departures between the sample and the reference population. Some of the examples of the imperfections include non-coverage and underrepresentation of the population, non-response, unequal probabilities of selection. The present disclosure provides procedures for developing and augmenting the weights to compensate for these unavoidable problems of surveys and thus improving the quality of observations in the survey. The advent of fast-speed computers and affordable statistical software should make the use of weights a routine aspect of the analysis of survey data.