Detecting unusual Responses in Social Media

Sayaji Hande, Ramesh Srinivasaraghavan, Vineet Gupta & Sandeep George
Adobe Research Lab, India

September 11, 2013

Abstract

Social Media platforms have not only made reactions possible, but they have reduced the gap between action and reaction. Before social media, traditional platforms of advertising such as TV and Newspapers, did not have a direct way to measure reactions - that is no longer the case. Sure, there are reactions to actions; but how does one measure or quantify these.

For example, a recent NYTimes article stated that the Pope who is an infrequent communicator in social media has more influence on society rather than the President of the United States of America. Clearly, it is important to asses not just the frequency; but its impact. Consider another example of an individual who has a tendency to complain most of the time. Now, if this individual actually expresses satisfaction, that response needs to be re-scored to a higher value.

It is a fact, that responses of an individual are often colored by his/her personality, even when he/she is reacting to specific situations or stimuli. There is no mechanism in statistics for readjusting response measurements to these kinds of situations, especially for categorical variables. If one aggregates the data without taking 'personality' factor, results could be misleading.

In this paper, we develop a mathematical model that uncovers the personalities of individuals based on how they respond to a set of questions. Although we formulate our model in the application context of surveys, this work can easily be extended to analysis of responses to marketing campaigns and behavior on social networks.

1 Methodology Overview

Let us suppose that there is an event E_i for $i = 1, 2, \ldots, k$. An event here could be a response to question in set of questions of a survey, or it could be a posting by company on social media page, etc. As highlighted in abstract, the measurements could have been tilted based on individual characteristics, which would get reflected across multiple events. Hence, effects are estimable. In continuous response models, *Fixed Effects*, *Random Effects*, models exists. However, there are no well established models for categorical data. At best, categorical variables get treated like continuous variables.

Here is basic conceptualization of the model. We assume people are either in one of the segment. For sake of simplicity, lets have three segments: Optimist, Realist & Pessimist. Let X_{ij} be a response of ith individual for jth event. In our model, we assume every individual responds at start with as a realist. Then, if they are not in *'Realist'* category, then, with certain probability they move (increase by one or decrease by one). The joint probability mass function would look like, if we assume n_1 are from category O, n_2 from category R and n_3 from category P, with $n = n_1 + n_2 + n_3$,

$$l(\vec{x}) \quad \propto \quad \frac{n_1!n_2!n_3!}{n!} \prod_j \prod_i \prod_r p_{ijr}^{x_{ijr}}$$

In above equation, i stands for ith event, r stands for a category of an individual, and x_{ijr} is # of rth category chose ith option for jth event. Of course, we don't know individual $xijr$'s, however we know $x_{ij} = xijP + xijR + xijO$, which we know. From standard multinomial distributions we know that *max* occurs when probability of that cell is equated to that of proportion of the count and in our case solving those equations.

Following is algo for estimation..

- Choose 'monotonous' and 'consistence' method of segmentation of population, deter, choose, n_1, n_2, n_3
- Estimate parameters and compute likelihood function
- As soon as likelihood function is converging, stop.

Our model evaluates the maximum likelihood estimates for 'r' possible ordinal responses in a campaign. This estimate when evaluated across different campaigns can help us understand the distribution of responders who are optimistic, pessimistic and normal. Once we have this understanding of the population, we can use this to normalize the response to remove the individual behavioral effect and get optimal response.

2 Mathematical Model

We explain our mathematical model in terms of how a set of N respondents choose to answer K questions in a survey, where each question seeks to get the users choice on an ordinal scale from 1 to L. Further assume that a value of 1 implies most negativity, dis-satisfaction or weakness; and conversely L implies the most positivity, satisfaction or strength.

We assume that there are various types of people who respond to a survey. These respondents can vary with respect to each other on multiple dimensions. However, for this discussion, we will pick one specific scale; that is, the continuum ranging from pessimism to optimism. Based on this, we can classify respondents into three types - Pessimists, Realists and Optimists. It must be noted here that type does not always have to be a value in the

Pessimist-Optimist scale. We can easily extend this notion to other scales of interest, such as Angry-Calm, Bored-Excited, Sad-Happy, etc.

We can now model pessimists and optimists in terms of realists. In relative terms, pessimists tend to choose a lower value than what they truthfully imply, whereas optimists tend to over rate their responses. We can assume every respondent regardless of type answers every question in the survey truthfully and then the bias of their type gets injected on their answer making them choose a higher or lower option in the choice continuum for some of the questions. We call this injection of bias as the *jump probability*, and for simplicity we assume that a pessimist or optimist may shift their answers by utmost one in the choice continuum. Let η and θ be the respective jump probabilities for pessimists and optimists.

Let τ_i be a score that represents the typical choice of the i^{th} individual. For example, a simple measure could be the sum of how the individual's choice deviates from the median choice across all questions. Since we have divided the population into three segments along the Pessimist-Optimist continuum, we can now cluster the individuals into three groups based on this score.

Let N_0, N_r and N_p be the number of the optimistic, realistic and pessimistic segments of the respondents, respectively.

The probability that an individual of type t chooses the l^{th} option for given a question k, can readily be estimated as

$$\hat{P}^t_{kl} \quad = \quad n^t_{kl}/N_t \tag{1}$$

where n^t_{kl} is the number of times that the population segment of type t, chose the l^{th} option for the k^{th} question, and N_t is the cardinality of the population whose type is t

We now derive the probability with which each type of individual (realists, optimists and pessimists) chooses the l^{th} option for the k^{th} question.

$$P(\mathit{Realist} = l) \quad = \quad P^r_{kl} \tag{2}$$

$$P(\mathit{Optimist} = l) \quad = \quad P^r_{kl}\theta + P^r_{kl}(1-\theta) \ \forall \ l{>}1 \text{ and } P_{k,l-1} = 0 \text{ when } l = 1 \tag{3}$$

$$P(\mathit{Pessimist} = l) \quad = \quad P^r_{k,l+1}\eta + P^r_{kl}(1-\eta) \ \forall \ l{<}l \text{ and } P_{k,l+1} = 0 \text{ when } l = L \tag{4}$$

The above model is discussed in following section in further detail and possible solution is arrived at.

2.1 Model Formulation & Solution Discussion

As stated thoughts in abstract. We assume, every individual self - a normal realist. There may be some subjects that wear a personality cover. One may be pessimistic and other may be optimistic. We are specifically talking about ordinal data. Individual personality estimation is not focus of discussion, but, how one would aggregate the responses for a posting or a champaign.

In the model stated above - its kind of additive of normal self and its personality. For example, normal self may choose a particular response and if personality is 'Optimistic', the model says (s)he is likely to jump to right, with certain degree of unknown probability. Similar formulation is done for 'Pessimist' as well.

Here, its easy to note what are the unknown parameters.

- Proportions of three clusters
- Positive Tilt in Optimist
- Negative Tilt in Pessimist
- Distribution of responses for all L campaigns (questions, postings)

Estimability of the parameters is a question. For example, if one does not assume a lower bound for a 'Tilt', then creation of clusters would be difficult. Similarly if one does not assume some distribution on personality in population, 'Tilt' parameters would be difficult.

To start with, we assume proportions of the cluster is known to the subject matter experts. This paper provides this leaver to the expert who would provide this as an input, see the estimates and so on. Following are the solution steps when you have ordinal data of the nature discussed above.

Following is the process of analyzing data:

- **Inputs:** Data & Take the cluster sizes (proportions)
- **Process Steps:**
  1. Score individual for his Pessimism as well as Optimism.
  2. Rank order individuals individual and put them in appropriate clusters in accordance with size prescribed.
  3. Estimate 'Tilt' parameters for individual clusters
- Check the 'tilt' estimates, change proportions, if desired and follow the process again.
- Estimate multinomial probabilities for three clusters.
- Utilize 'Tilt' parameter estimates to come-up with combined estimates

In above we have stated that one would need to score every individual towards Pessimism as well as Optimism. This of course, can be done on basis of cross comparing with rest of the population responses. For a k campaigns, in an individual response could be compared with rest. Since k is large enough (say above 35) the individual treats would get reflected.

The folks with higher 'tilt' will have high score. For example, one can measure how many times one is scoring above median $+1$. In simulation model, we have seen that accuracy of this is above 90%.

3 Simulation Study: Steps

The 1st step of checking validation of model is assume the parameters, create scenario (data) and estimate the parameters. What are the parameters involved? The size of clusters, the 'tilt' (positive as well as negative) for respective clusters, the naked probabilities for categories of each campaign. Assume there are K campaigns. Let for each of campaign, there be L responses. Hence, one would need to assume parameter of matrix, $K \times L$ where row sums are 1. Following is what we stated.

1. Assume parameters and create simulated data (K, L, %age of cluster, N- total population of respondents, tilt - jump)
2. Create jumps based on assumed parameters and adjust the data for the same.
3. Estimate $K * L$ matrix, where we have K number of questions with L number of options

3.1 Detailed steps - Creation of Simulation Data

Assume we have to generate data for K questions with L choices

Input data for simulation

1. K - The number of questions that we need for the study
2. L - The number of options for each K questions in the study
3. N - The number of people taking part in the study
4. O_p - The percentage of people in the community assumed to be optimistic
5. P_p - The percentage of people in the community assumed to be pessimistic
6. R_p - The percentage of people in the community assumed to be realistic
7. θ - The jump probability of an optimistic individual in the study to shift to the right by 1.
8. η - The jump probability of a pessimistic individual in the study to shift to the left by 1.

3.1.1 Simulating Data

Creating the Choice-Probability Matrix
For each k question, we have, we should create a vector of length L with simulated probabilities that are likely to be the way people would answer to a question. We should ensure that the sum of all the probabilities of each k is 1. Here the key aim is to simulate the differences between questions which may be across categories that may yield different response pattern. We now have a KxL matrix with the probabilities in each row. The Row sum in all rows of this matrix is 1.

Creating the Choice-Response Matrix
With the $Choice - Probability$ matrix we will create another matrix: $Choice - Response$ matrix, with N rows and K columns.Using the previously generated matrix, we estimate a value between 1 and L for each of the cell in K x L matrix.
The filled N x K matrix now contains the simulated responses based on the $Choice - Probability$ matrix.

Injecting the biases
Using the input values O_p,R_p,P_p, we randomly identify individuals in the $Choice-Response$ matrix and mark them as *Optimist*,*Realist* or *Pessimist*. Now since we understand the biased individuals we can inject the bias into their responses.
For an individual marked as *Optimist* we evaluate every response they have made and use the jump probability θ defined earlier to create a right shift in their response.
For an individual marked as *Pessimist* we evaluate every response they have made and use the jump probability η defined earlier to create a left shift in their response.
We leave those individuals marked as a *Realist* without any change.

4 Appendix: Use Cases

- Case 1:
  Inputs: Responses of N individuals to a set of multiple campaigns, where the response is an ordinal number one a scale of J K (where J and K are arbitrary numbers)
  Output: Clustering of N individuals as Pessimists, Optimists and Realists
  Applications:

  1. We can now think of targeting each of these clusters differently. For example, you may use three different messaging to suit the psychometric profile of the person
  2. An descriptive statistics that one may use to describe the response of the entire population can now be subjected to stratified analysis. For example, you can compare the mean/median of pessimists with that of realists/optimists; and that may tell the true story. This is important because in the unstratified analysis the pessimists and the optimists may cancel each other out, hiding the naked truth.
  3. We can detect unusual responses

   (a) If we know that a person is optimistic, and he now rates something negatively this is unusual. Even if it may be a one-off instance, this is ripe for deeper analysis. Perhaps it may indicate impending customer churn? Or it could just be that the person had a bad hair day !

   (b) If, on the other hand, a pessimist rates something positively then that is unusual too and may have some signal that requires deeper analysis. Maybe we are doing something really well

- Case 2
  Inputs:

  1. Responses of N individuals to a set of multiple campaigns, where the response is an ordinal number one a scale of 1 L
  2. A new campaign and responses of the N individuals to this new campaign

  Output: Quality of the new campaign

  Applications:

  1. Assume quality of campaign is mostly based on the content. Therefore we can now estimate the quality of the new content; and we can compare this with the quality of prior content that you published
  2. What we are saying is that Actual Response is a function of True Response and Individuals Bias. Assuming a linear model this can be expressed as Actual Response = Alpha*TrueResponse + Beta*Pessi-OptiScore+ Epsilon for all individuals Here we know Pessi-OptiScore and actual response for each individual. We can estimate Alpha, True Response and Epsilon
  3. Here Alpha can be thought of as indicating the quality of the campaign.
  4. We can compute Alpha for each new campaign and this can be used for comparing all your campaigns.
  5. As a social marketer you may want to know whether your Alpha is improving or worsening
  6. You may want to find Alpha for Pessimists, Optimists and Realists separately
  7. Or we can also think of Alpha for various kinds of campaigns/posts (with/withoutURL; with/without pic/video; with/without humour, etc.)

  Here we know Pessi-OptiScore and actual response for each individual. We can estimate Alpha, True Response and Epsilon

  1. Here Alpha can be thought of as indicating the quality of the campaign.
  2. We can compute Alpha for each new campaign and this can be used for comparing all your campaigns.
  3. As a social marketer you may want to know whether your Alpha is improving or worsening

4. You may want to find Alpha for Pessimists, Optimists and Realists separately
5. Or we can also think of Alpha for various kinds of campaigns/posts (with/withoutURL; with/without pic/video; with/without humour, etc.)