Survey Engagement Analysis on Harvard CAS Dataset

Justin Weltz, Irene Ji, Keru Wu

Abstract

Harvard SPH College Alcohol Study (CAS) collected a multi-round survey that interviewed students about their alcohol use and other high risk behaviors. The survey contains multiple Likert scale questions, and we identified that a large number of students tend to give the same answer for these similar questions, indicating that they're less engaged in the survey. To address this latent survey engagement factor, we use structual equation model (SEM) to estimate response quality and survey engagement, finding out relationships between drinking behaviors and survey engagement.

1. Introduction

Harvard CAS contains four sets of survey responses from undergraduates in four years: 1993, 1997, 1999 and 2001. The survey mainly focus on students high risk behaviors (e.g. alcohol, tobacco and illicit drugs), including other information such as students' views on campus alcohol policies, personal background variables, etc.

These four surveys each consists of over 400 questions, which could take up to an hour for a student to complete. The survey also has many similar nested questions (e.g. Fig 1). It's probable that some students are less engaged in completing the survey, finally returning the survey with non-informative responses. In psychology, these nested similar questions are called Likert scale questions. Here's a standard five point Likert scale: (1) Strongly Disagree (2) Disagree (3) Undecided (4) Agree (5) Strongly Agree. With so many Likert scale questions in a single survey, we find out that a non-negligible proportion of students tend to give the same answer for these questions.

From this perspective, we are interested in exploring the latent survey engagement factor behind students. Specifically, we are interested in (1) quantifying and estimating latent survey engagement (2) exploring relationship between drinking behaviors and survey engagement. To make our analysis more detailed, our report focuses on the 1999 survey, but it's straightforward to use our approahces to analyze other surveys.

2. Materials and Methods

Survey Engagement paper: (Hess and Stathopoulos 2013)

SEM paper: (Ullman and Bentler 2003)

Model structure link: Fig 2

Simplified model link: Fig 3

Model 1 semplot link: Fig 4

Model 2 semplot link: Fig 5

3. Results

3.1 Exploratory Data Analysis & Data Preprocessing

Initial data exploration suggests that some students give the same answer for nested Likert scale questions (Fig 6). We can see a large number of students response with the first Likert scale (so their sum over 10 questions are 0), implying that engagement among students varies.

The survey contains a large number of missing data, and we consider different ways to deal with it: (1) use complete cases, (2) manually impute with reasonable values for questions need not to answer, (3) use MICE (Buuren and Groothuis-Oudshoorn 2010) to impute, (4) use DPMPM (Fig 7, Dunson and Xing (2009), Si and Reiter (2013)). We will later explain our sensitivity analysis through different missing data manipulations.

Instead of analyzing all questions in the survey, which is unnecessary and redundant, we select some important questions of interest (Fig 8). Specifically speaking, for personal characteristics, we include sex, age group, comments, drink_category (from codebook), drink_occ (question C8, C9), drink_num (question C8, C10), advice (question D4), complaint (question D5), perception (D3A, D3B), etc. To further address our concern on students' view on alchohol policies, we include questions related to their attitudes (question B1, B2, B3, B4, B5). For choice indicator Y, we will focus on question B15 (Fig 1). We preprocess these variables using different criteria and formulas, which is explicitly explained in our codebook for variables of interest (Fig 8). Among quesitons of interest, around 20% cases have missing data.

3.2 Main Results

Model 1 result link: Fig 9, 10

Model 2 result link: Fig 11, 12

3.3 Sensitivity Analysis

About 20% cases have missing data with respect to our variables of interest (Fig 8). We consider four ways to manipulate these missing data: (1) use complete cases, (2) manually impute with reasonable values for questions no need to answer, (3) use MICE (Buuren and Groothuis-Oudshoorn 2010) to impute, (4) use nonparamatric Bayesian imputation DPMPM (Fig 7, Dunson and Xing (2009), Si and Reiter (2013)). We would expect DPMPM to have the best performance since it's designed for large-scale categorical surveys.

Our models using four different ways above didn't show distinguishable difference. This is probably because proportion of missing data is relatively low among questions we focus on. Therefore we choose to use the original dataset in future work, and explore missing data manipulation further.

4. Discussion

Our Structural Equation Model successfully discovers latent survey engagement factors in Havard CAS dataset. We also discover relationships between drinking behaviors and survey engagement. However, we carry out a simplified version of original model due to constraint of lavaan package (Rosseel 2012). A direct future work is to implement the model jointly. To further account for uncertainty, we may implement it using Rstan or JAGS.

To find nonlinear effects of latent factors, we may consider using autoencoder (Kramer 1991). Autoencoder can be viewed as a nonlinear factor analysis that primarily uses neural networks to first encode the input and later decode it. However, autoencoder may suffer from interpretability due to its complexity.

Another direction is to consider low-rank factorization of contingency tables. Sparse PARAFAC (@ Zhou et al. 2015) uses low rank tensor factorization together with parallel factor analysis, which is

helpful when the sample size is massively less than the number of cells. This model can be further used for latent class clustering.

Figure 1:   Question B15

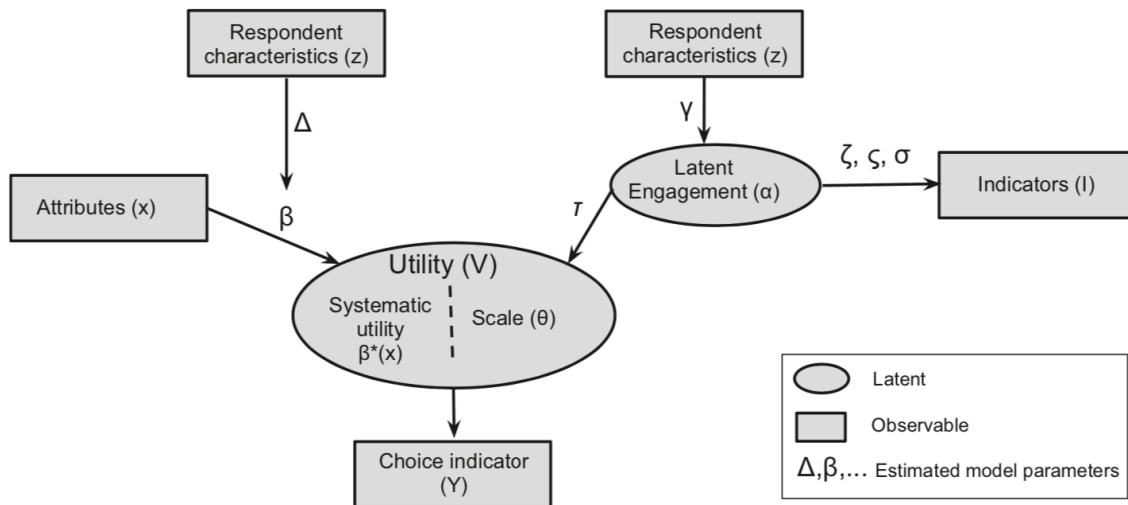

Figure 2:   Model Structure
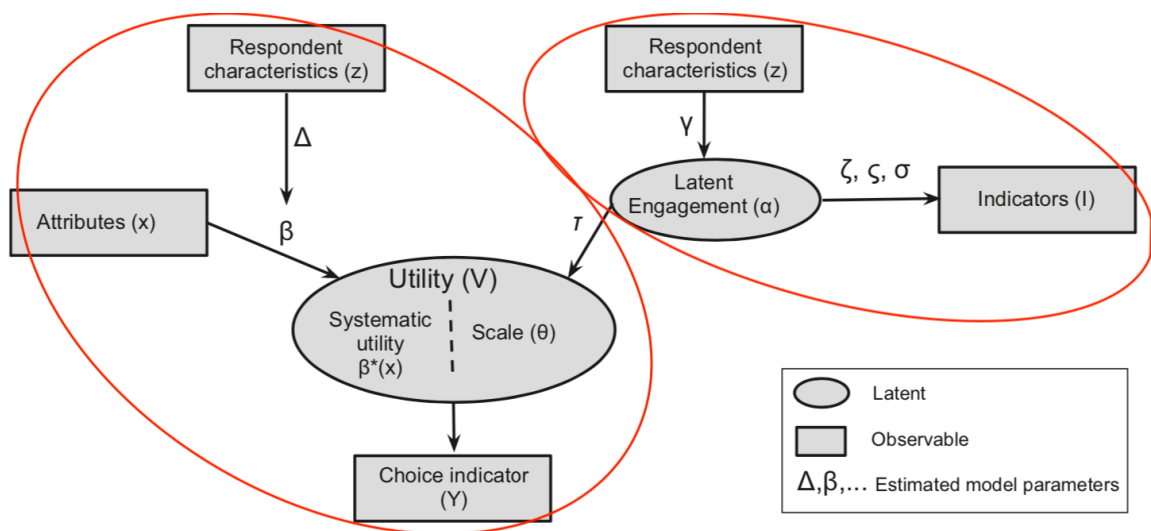
Figure 3: Simplified Model

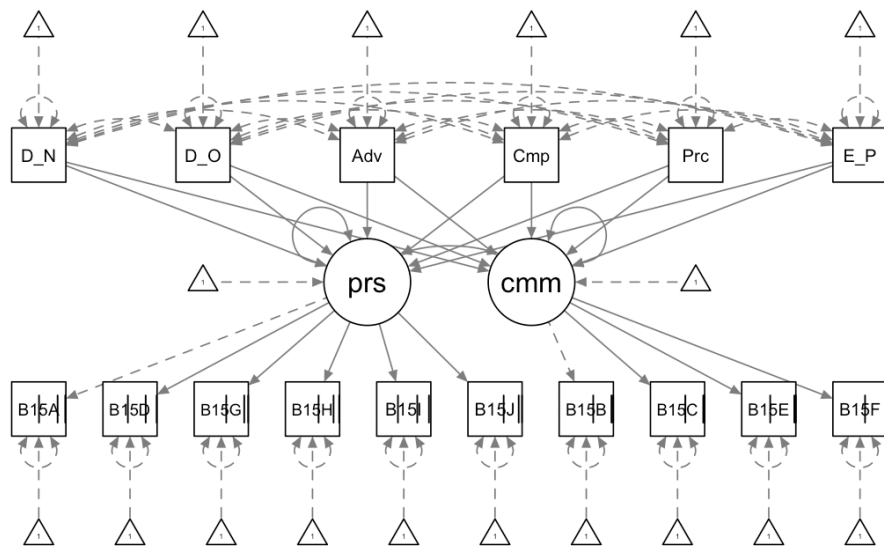
Figure 4: Model 1 SEM plot

4
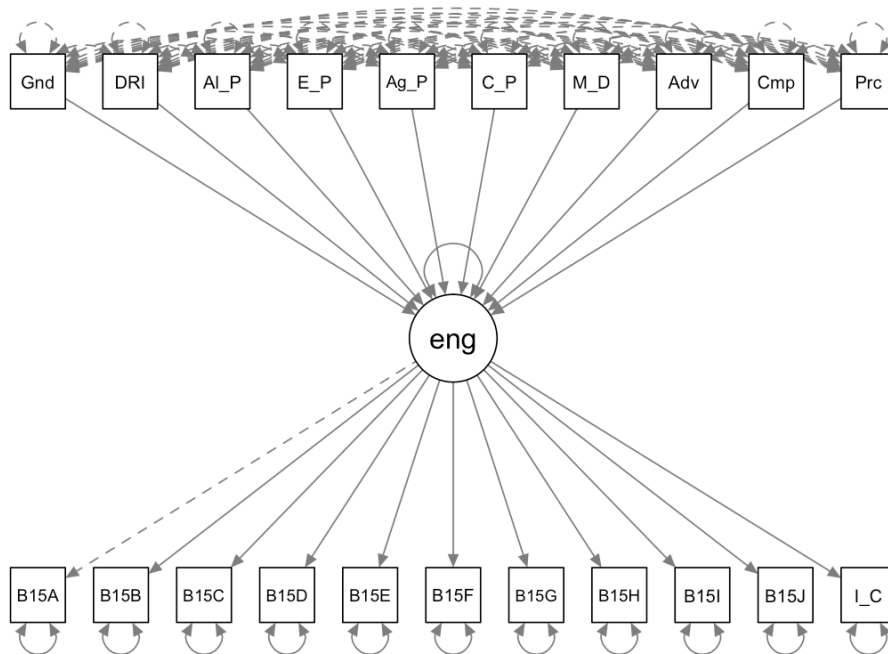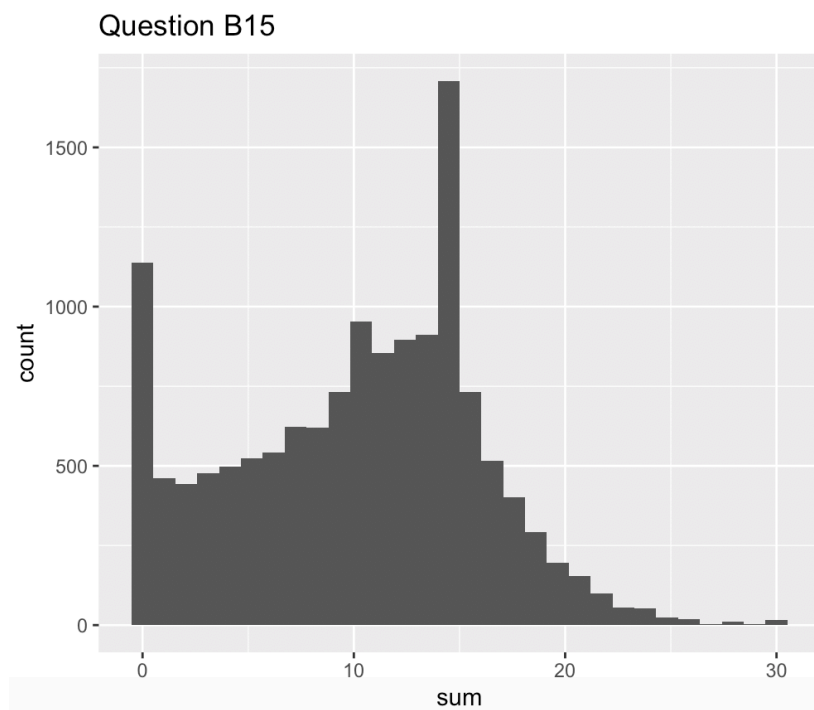
Figure 5: Model 2 SEM plot

Figure 6: Histogram of sum of individual answers for question B15

$$X_{ij}|z_i, \phi \sim \text{Multinomial}(\phi_{z_i, j1}, ..., \phi_{z_i j d_j})$$

$$z_i \pi \sim \text{Multinomial}(\pi_1, ..., \pi_\infty)$$

$$pi_h = V_h \prod_{g<h}(1 - V_g), \quad h = 1, ..., \infty$$

$$V_h \sim \text{Beta}(1, \alpha)$$

$$\alpha \sim Gamma(a_\alpha, b_\alpha)$$

$$\phi_{hj} = (\phi_{hj1}, ..., \phi_{hjd_j}) \sim \text{Multinomial}(a_{j1}, ..., a_{jd_j})$$

Figure 7: Infinite Mixture of Product of Multinomials

Ind_Comment	COMMENTS	1 (yes); 2 (no)	1 (yes); 0 (no)
Gender	SEX	0 (female); 1 (male)	-
Age_Group	AGEGROUP	1: <21; 2: 21-23; 3: >23	1: <21; 0: >=21
DRINKCAT	DRINKCAT	1,2,3 (codebook)	-
Alc_Problem	B1	1:major; 2:minor; 3:yes; 4: no	1: yes; 0: no
AP_all	B2	1 (all)	1: yes; 0: no
AP_stu	B2	2 (all students)	1: yes; 0: no
AP_all21	B2	3 (all <21)	1: yes; 0: no
AP_stu21	B2	4 (all student <21)	1: yes; 0: no
AP_no	B2	5 (no policy)	1: yes; 0: no
AP_notknow	B2	6 (don't know)	1: yes; 0: no
Enforce_Pol	B3	1-3: enforced; 4-5: not enforced/don't know	1: enforced; 0: no
Agree_Pol	B4	1-2: agree; 3-4: disagree	1: agree; 0: disagree
Change_Pol	B5	2-3: change; 1: not change; 4: don't know	1: change; 0: others
Min_Drink_Age	B13	1-4: below 21; 5: 21	1: <21; 0: 21
Drink_Occ	C8, C9	C9=1: none<=30days; 2-7; C8:1,2,3: no drink<=30days	If C8=1,2,3 -> Drink_Occ=1; else follow C9
Drink_Num	C8, C10	C10=0: none<=30days; 1-9; C8:1,2,3: no drink<=30days	If C8=1,2,3 -> Drink_Num=0; else follow C10
Advice	D4	1,2,3,4	1 (no), 2,3,4
Complaint	D5	1,2,3,4	1 (no), 2,3,4
Perception	D3A, D3B	D3A: All students; D3B: Your friends	D3B/D3A

Figure 8: Variables of Interest

```
Latent Variables:
                     Estimate   Std.Err   z-value   P(>|z|)
  personal =~
    B15A               1.000
    B15D               1.023     0.013    81.215     0.000
    B15G               1.408     0.014    99.281     0.000
    B15H               1.204     0.013    93.422     0.000
    B15I               0.932     0.013    73.739     0.000
    B15J               1.314     0.013    97.315     0.000
  communal =~
    B15B               1.000
    B15C               0.879     0.016    53.845     0.000
    B15E               1.082     0.016    67.255     0.000
    B15F               1.218     0.018    67.855     0.000
```

Figure 9:   Model 1 Latent Factors

```
Regressions:
                     Estimate   Std.Err   z-value   P(>|z|)
  personal ~
    Drink_Num          0.080     0.004    22.805     0.000
    Drink_Occ          0.188     0.006    30.606     0.000
    Advice             0.024     0.007     3.607     0.000
    Complaint         -0.145     0.014   -10.081     0.000
    Perception         0.144     0.012    12.484     0.000
    Enforce_Pol        0.044     0.017     2.638     0.008
  communal ~
    Drink_Num          0.061     0.004    14.955     0.000
    Drink_Occ          0.142     0.007    20.209     0.000
    Advice            -0.023     0.007    -3.044     0.002
    Complaint         -0.098     0.016    -5.992     0.000
    Perception         0.092     0.014     6.563     0.000
    Enforce_Pol        0.009     0.019     0.485     0.627
```

Figure 10:   Model 1 Regression results

```
Latent Variables:
                  Estimate  Std.Err  z-value  P(>|z|)
    engagement =~
      B15A_Res2       0.021    0.004    5.466    0.000
      B15B_Res2       0.083    0.003   24.244    0.000
      B15C_Res2       0.111    0.003   31.970    0.000
      B15D_Res2       0.038    0.004   10.173    0.000
      B15E_Res2       0.126    0.003   39.163    0.000
      B15F_Res2       0.187    0.004   51.104    0.000
      B15G_Res2       0.180    0.004   51.213    0.000
      B15H_Res2       0.053    0.003   16.390    0.000
      B15I_Res2       0.020    0.004    4.835    0.000
      B15J_Res2       0.121    0.003   37.507    0.000
      Ind_Comment     0.004    0.003    1.330    0.183
```

Figure 11: Model 2 Latent Factors

```
Regressions:
                  Estimate  Std.Err  z-value  P(>|z|)
    engagement ~
      Gender         -0.292    0.030   -9.593    0.000
      DRINKCAT       -0.961    0.023  -42.539    0.000
      Alc_Problem     0.330    0.040    8.288    0.000
      Enforce_Pol     0.044    0.037    1.197    0.231
      Agree_Pol       0.063    0.037    1.727    0.084
      Change_Pol      0.162    0.036    4.566    0.000
      Min_Drink_Age  -1.090    0.036  -30.471    0.000
      Advice         -0.002    0.015   -0.164    0.869
      Complaint       0.416    0.033   12.742    0.000
      Perception     -0.349    0.030  -11.721    0.000
```

Figure 12: Model 2 Regression results

References

Buuren, S van, and Karin Groothuis-Oudshoorn. 2010. "Mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software.* University of California, Los Angeles, 1–68.

Dunson, David B, and Chuanhua Xing. 2009. "Nonparametric Bayes Modeling of Multivariate Categorical Data." *Journal of the American Statistical Association* 104 (487). Taylor & Francis: 1042–51.

Hess, Stephane, and Amanda Stathopoulos. 2013. "Linking Response Quality to Survey Engagement: A Combined Random Scale and Latent Variable Approach." *Journal of Choice Modelling* 7. Elsevier:

1–12.

Kramer, Mark A. 1991. "Nonlinear Principal Component Analysis Using Autoassociative Neural Networks." *AIChE Journal* 37 (2). Wiley Online Library: 233–43.

Rosseel, Yves. 2012. "Lavaan: An R Package for Structural Equation Modeling and More. Version 0.5–12 (Beta)." *Journal of Statistical Software* 48 (2): 1–36.

Si, Yajuan, and Jerome P Reiter. 2013. "Nonparametric Bayesian Multiple Imputation for Incomplete Categorical Variables in Large-Scale Assessment Surveys." *Journal of Educational and Behavioral Statistics* 38 (5). Sage Publications Sage CA: Los Angeles, CA: 499–521.

Ullman, Jodie B, and Peter M Bentler. 2003. "Structural Equation Modeling." *Handbook of Psychology.* Wiley Online Library, 607–34.

Zhou, Jing, Anirban Bhattacharya, Amy H Herring, and David B Dunson. 2015. "Bayesian Factorizations of Big Sparse Tensors." *Journal of the American Statistical Association* 110 (512). Taylor & Francis: 1562–76.