Survey Engagement Analysis on Harvard CAS Dataset

Justin Weltz, Irene Ji, Keru Wu

Introduction

- Data: Surveys of undergraduate drinking habits in 4 years.
- Goal:
  - Estimate response quality and survey engagement.
  - Find relationships between drinking behaviors and survey engagement.
- Model:
  - Structual Equation Model (SEM)

Likert Scale

- A typical psychometric response scale:
  - Five points: (1) Strongly disagree; (2) Disagree; (3) Neither agree nor disagree; (4) Agree; (5) Strongly agree
- The survey contains many nested Likert scale questions:
  - When a student is not engaged in the survey, it's likely that he/she tends to give the same answer for these questions.
  - Aim to estimate this effect

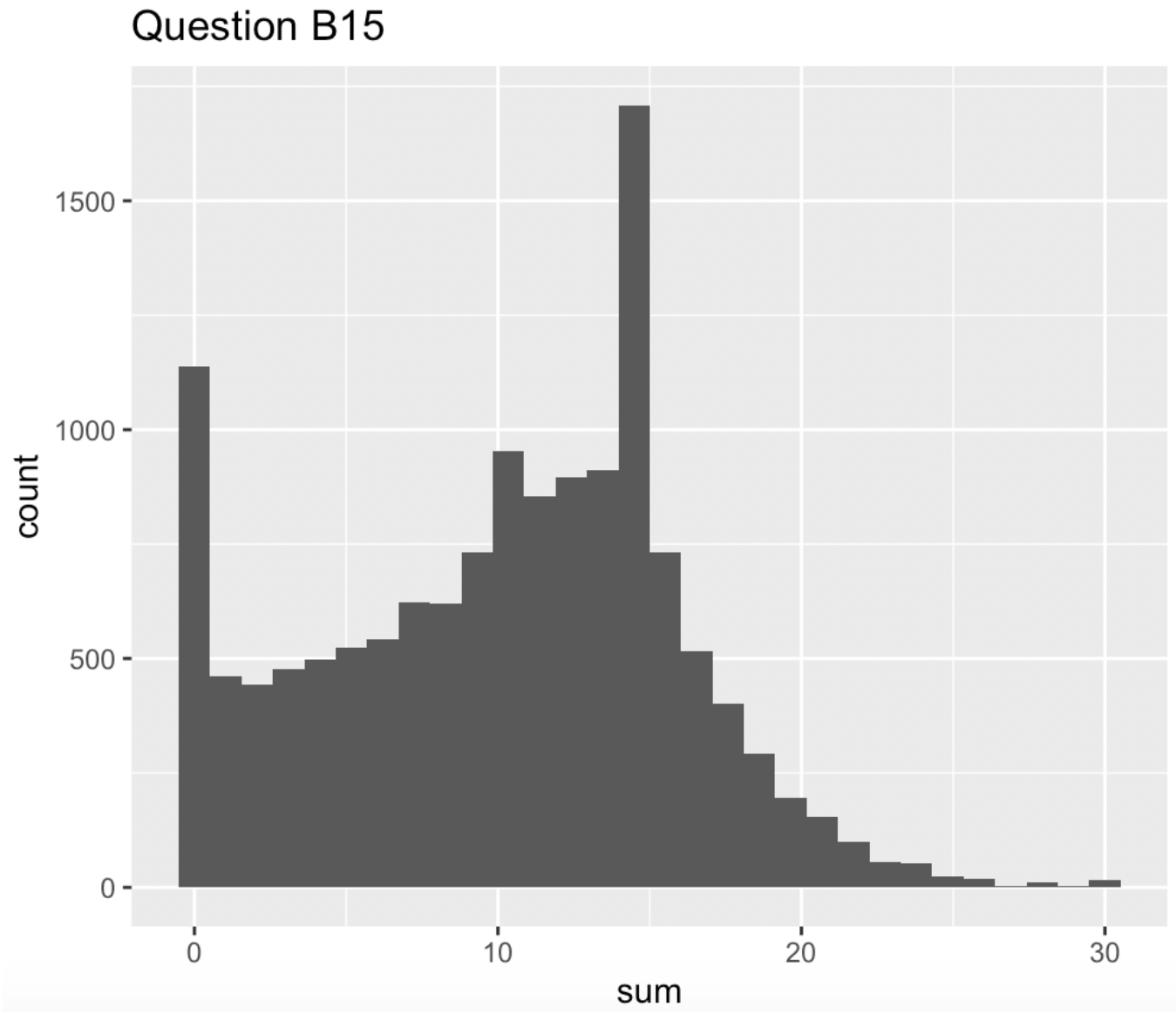
EDA - Example

- Question B15 in 1999 survey

B15. To what extent do you support or oppose the following possible school policies or procedures? (Choose one answer in each row.)	Strongly Support	Support	Oppose	Strongly Oppose
a. Prohibit kegs on campus	○	○	○	○
b. Offer alcohol-free dorms	○	○	○	○
c. Require non-alcoholic beverages be available when alcohol is served at campus events	○	○	○	○
d. Ban advertisements of alcohol availability at campus events and parties	○	○	○	○
e. Provide more alcohol-free recreational and cultural opportunities such as movies, dances, sports, and lectures	○	○	○	○
f. Make the alcohol rules more clear	○	○	○	○
g. Enforce the alcohol rules more strictly	○	○	○	○
h. Crack down on drinking at sororities and fraternities	○	○	○	○
i. Hold hosts responsible for problems arising from alcohol use	○	○	○	○
j. Crack down on under-age drinking	○	○	○	○

EDA - Example

- Histogram of sum over B15 questions



Data preprocessing

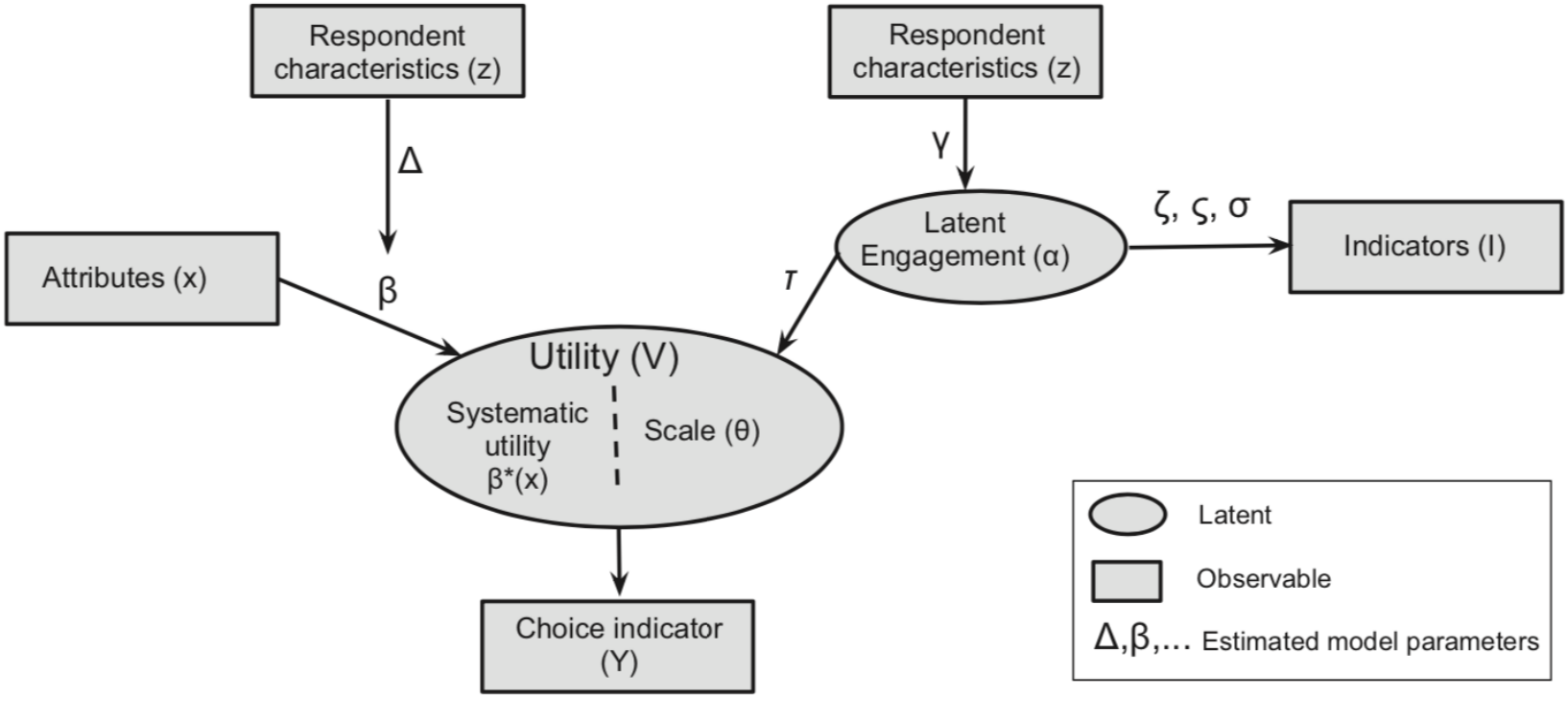
- Missing data
  - Among variables of interest, around 2000 cases have missing data.
- Different ways to manipulate missing data
  - (1). Use complete case for analysis
  - (2). Impute with reasonable values
  - (3). Nonparametric Bayesian Imputation (DPMPM):

$$
\begin{aligned}
X_{ij}|z_i, \phi &\sim \text{Multinomial}(\phi_{z_i,j1}, ..., \phi_{z_i jd_j}) \\
z_i \pi &\sim \text{Multinomial}(\pi_1, ..., \pi_\infty) \\
pi_h &= V_h \prod_{g<h} (1 - V_g), \quad h = 1, ..., \infty \\
V_h &\sim \text{Beta}(1, \alpha) \\
\alpha &\sim Gamma(a_\alpha, b_\alpha) \\
\phi_{hj} &= (\phi_{hj1}, ..., \phi_{hjd_j}) \sim \text{Multinomial}(a_{j1}, ..., a_{jd_j})
\end{aligned}
$$

Variables of Interest

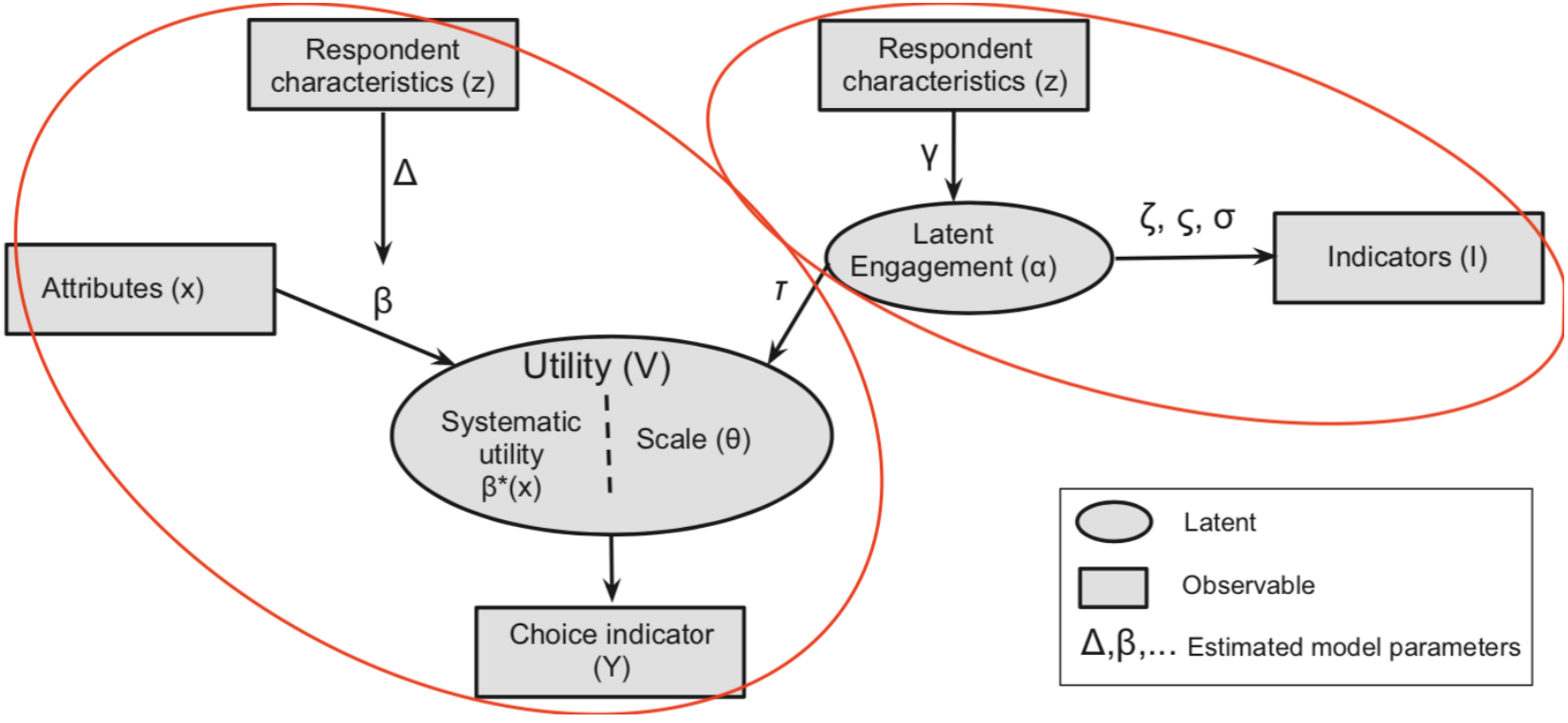
Ind_Comment	COMMENTS	1 (yes); 2 (no)	1 (yes); 0 (no)
Gender	SEX	0 (female); 1 (male)	-
Age_Group	AGEGROUP	1: <21; 2: 21-23; 3: >23	1: <21; 0: >=21
DRINKCAT	DRINKCAT	1,2,3 (codebook)	-
Alc_Problem	B1	1:major; 2:minor; 3:yes; 4: no	1: yes; 0: no
AP_all	B2	1 (all)	1: yes; 0: no
AP_stu	B2	2 (all students)	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

Main Model



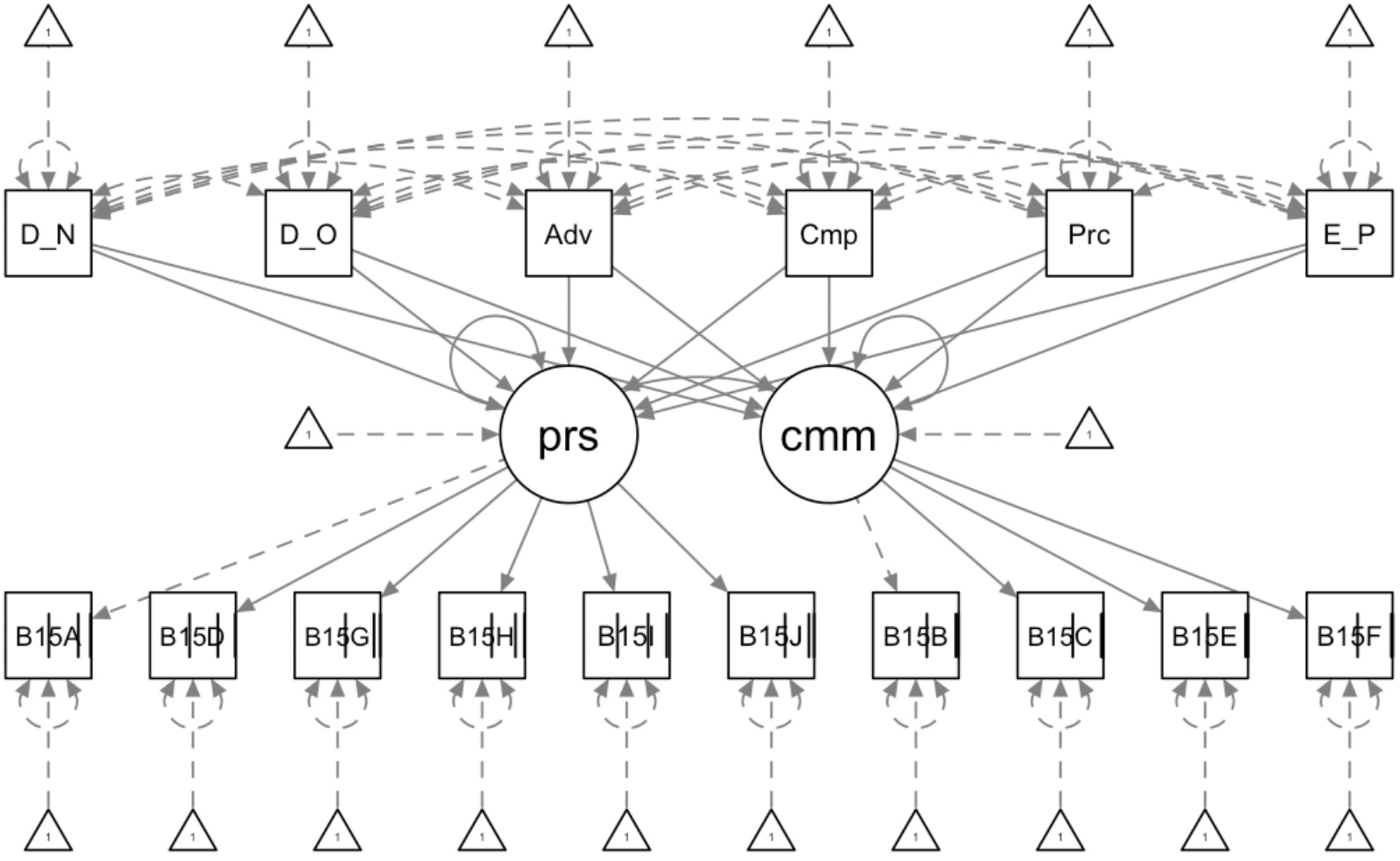
- Ordered logit for indicators I
- Random scale model for Utility $V = e^{\tau \alpha_n} \beta^T x_n$
- Likelihood $L = \sum_{n=1}^{N} ln \int_\beta \int_\alpha p(y_n|\cdot) p(I_n|\cdot) p(\alpha) p(\beta|\Omega) d\alpha d\beta$

Simplified version



- Fit two models sperately:
  - First build up a SEM for the choice model.
  - Plug the residuals into the second SEM to find latent engagement factors.

Model 1 - SEM plot



Model 1 - Latent Variables

Latent Variables:

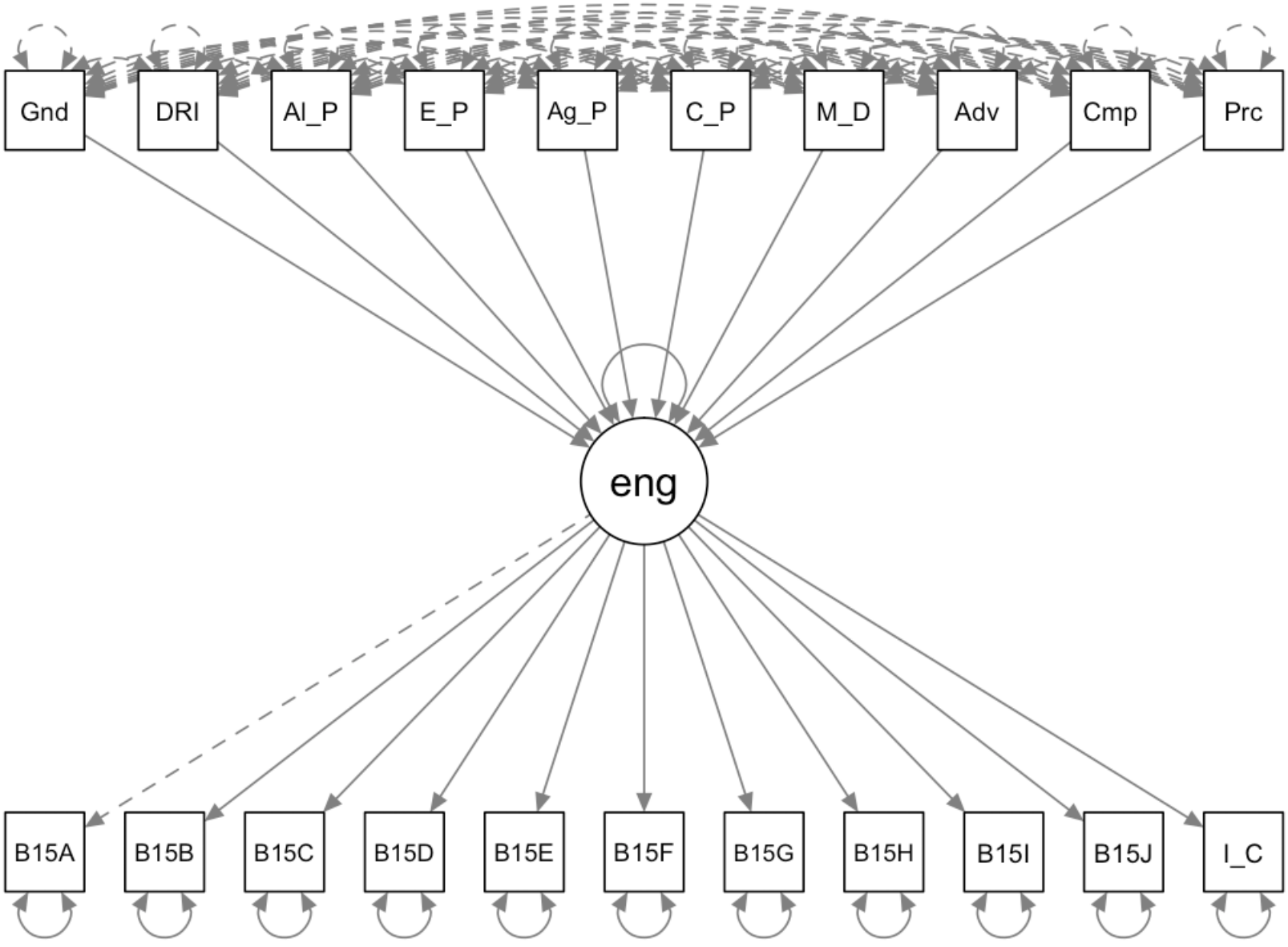
	Estimate	Std.Err	z-value	P(>|z|)
personal =~				
B15A	1.000			
B15D	1.023	0.013	81.215	0.000
B15G	1.408	0.014	99.281	0.000
B15H	1.204	0.013	93.422	0.000
B15I	0.932	0.013	73.739	0.000
B15J	1.314	0.013	97.315	0.000
communal =~				
B15B	1.000			
B15C	0.879	0.016	53.845	0.000
B15E	1.082	0.016	67.255	0.000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

Model 1 - Regression results

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

Model 2 - SEM plot



Model 2 - Latent Variables

Latent Variables:

	Estimate	Std.Err	z-value	P(>|z|)
engagement =~				
B15A_Res2	1.000			
B15B_Res2	-0.889	0.034	-26.000	0.000
B15C_Res2	-0.515	0.025	-20.779	0.000
B15D_Res2	1.227	0.045	27.298	0.000
B15E_Res2	-0.771	0.030	-25.876	0.000
B15F_Res2	-1.903	0.067	-28.395	0.000
B15G_Res2	3.445	0.146	23.554	0.000
B15H_Res2	0.724	0.029	24.593	0.000
B15I_Res2	0.959	0.039	24.696	0.000
B15J_Res2	0.927	0.034	27.026	0.000
Ind_Comment	-0.098	0.016	-6.278	0.000

Model 2 - Regression results

Regressions:

	Estimate	Std.Err	z-value	P(>|z|)
engagement ~				
Gender	-0.004	0.001	-4.324	0.000
DRINKCAT	-0.009	0.001	-15.722	0.000
Alc_Problem	-0.007	0.001	-5.336	0.000
Enforce_Pol	-0.006	0.001	-4.683	0.000
Agree_Pol	-0.001	0.001	-0.827	0.408
Change_Pol	0.003	0.001	2.924	0.003
Min_Drink_Age	0.013	0.001	11.798	0.000
Advice	-0.008	0.001	-13.951	0.000
Complaint	0.008	0.001	7.361	0.000
Perception	-0.008	0.001	-8.317	0.000

Conclusions:

- Students with drinking behaviors tend to be more engaged in the survey
- Male are engaged in this alcohol study
- etc..

Discussion

- Implement Bayesian version to account for uncertainty
- Autoencoder: find nonlinear relationship
  - May lose interpretability
- Low-rank tensor factorization (e.g. sparse PARAFAC)