Assessing Partial Association between Mixed Data

Shaobo Li

School of Business
University of Kansas

(joint work with Jonathan Fan, Dungang Liu, Ivy Liu and Philip S. Morrison)

CBAR Seminar, University of Kansas

April 14, 2023

Outline

What is partial association

- Blood pressure vs. amount food eaten
  - What if control for weight, age, and hours of exercise?
- Firms credit rating vs. debt-to-equity ratio
  - What if control for stock volatility, net income to total asset?
- Voting vs. party identification
  - What if control for education, age and income?
- Wellbeing vs. Anxiety
  - What if control for financial strain and health condition

What is partial association

- Blood pressure vs. amount food eaten
  - What if control for weight, age, and hours of exercise?
- Firms credit rating vs. debt-to-equity ratio
  - What if control for stock volatility, net income to total asset?
- Voting vs. party identification
  - What if control for education, age and income?
- Wellbeing vs. Anxiety
  - What if control for financial strain and health condition

Note that:

— These associations do not have directions.

— Variables to be controlled for are often called confounding factors or moderators (loosely defined).

Why partial association analysis

Partial association analysis is a statistical procedure that can help

- revealing the hidden (true) association between two variables after removing potential confounders
- measuring the strength of this association
- testing if this association is statistically significant
- visualizing the shape of this association
- assessing confounding effects
  - quantification
  - testing significance

A real data example

- Data: college students wellbeing survey
- Outcome variables:
  - Y_1: Wellbeing score (continuous); Y_2: Anxiety (scale 1-5)
- Covariates: $\boldsymbol{X}$ including Financial strain, Healthiness, Loneliness, Accommodation, Age and Gender.
- Question:
  - What is the partial association between *Wellbeing* and *Anxiety*?
  - What is the change of association strength due to covariates, i.e., the confounding effect?

A traditional approach

- To answer question 1, fit model

$$Wellbeing = \alpha_1 + \beta_1 \times Anxiety + \gamma_1^T \boldsymbol{X} + \epsilon$$

- β_1 can be viewed as a measure of partial association

A traditional approach

- To answer question 1, fit model

$$Wellbeing = \alpha_1 + \beta_1 \times Anxiety + \gamma_1^T \boldsymbol{X} + \epsilon$$

- β_1 can be viewed as a measure of partial association

- To answer question 2, fit model

$$Wellbeing = \alpha_0 + \beta_0 \times Anxiety + \epsilon$$

- $\frac{\beta_1 - \beta_0}{\beta_0}$ can be used to quantify the confounding effect
- A rule of thumb: 10% indicates nonnegligible effect

A traditional approach

- To answer question 1, fit model

$$Wellbeing = \alpha_1 + \beta_1 \times Anxiety + \gamma_1^T \boldsymbol{X} + \epsilon$$

- β_1 can be viewed as a measure of partial association

- To answer question 2, fit model

$$Wellbeing = \alpha_0 + \beta_0 \times Anxiety + \epsilon$$

- $\frac{\beta_1 - \beta_0}{\beta_0}$ can be used to quantify the confounding effect
- A rule of thumb: 10% indicates nonnegligible effect

- However, there are some problems

A traditional approach

- To answer question 1, fit model

$$Wellbeing = \alpha_1 + \beta_1 \times Anxiety + \gamma_1^T \boldsymbol{X} + \epsilon$$

- β_1 can be viewed as a measure of partial association

- To answer question 2, fit model

$$Wellbeing = \alpha_0 + \beta_0 \times Anxiety + \epsilon$$

- $\frac{\beta_1 - \beta_0}{\beta_0}$ can be used to quantify the confounding effect
- A rule of thumb: 10% indicates nonnegligible effect

- However, there are some problems
  - The scale of β depends on the scale of all variables
  - Switching *Wellbeing* and *Anxiety* leads to different results
  - Multiple β's if *Anxiety* is treated as a categorical variable

Analysis results

Panel A. linear regression model with response variable: "wellbeing score"			
	$Y_W \sim Y_A$	$Y_W \sim Y_A + \boldsymbol{X}$	% change
anxiety (2)	-4.588^{**}	-2.180	-52.48
	(2.290)	(2.172)	
anxiety (3)	-11.039^{***}	-7.341^{***}	-33.50
	(2.251)	(2.150)	
anxiety (4)	-21.021^{***}	-15.526^{***}	-26.14
	(2.204)	(2.134)	
anxiety (5)	-31.623^{***}	-23.466^{***}	-25.80
	(2.600)	(2.549)	
Avg.			-34.48
Panel B. adjacent category logit model with response variable: "anxiety"			
	$Y_A \sim Y_W$	$Y_A \sim Y_W + \boldsymbol{X}$	% change
wellbeing score	-0.032^{***}	-0.029^{***}	-9.278
	(0.002)	(0.002)	
Panel C. stereotype model with response variable: "anxiety"			
wellbeing score	0.124^{***}	0.118^{***}	-4.631
	(0.012)	(0.011)	
Note:			$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Outline

1. Introduction
2. Proposed method
3. Estimation, inference and visualization
4. Real data analysis

Basic idea – from classical statistics textbooks

- Fit marginal regression for each response variable

$$Y_1 \sim \beta_1^T X + \epsilon_1$$
$$Y_2 \sim \beta_2^T X + \epsilon_2$$

Basic idea – from classical statistics textbooks

- Fit marginal regression for each response variable

$$Y_1 \sim \beta_1^T X + \epsilon_1$$
$$Y_2 \sim \beta_2^T X + \epsilon_2$$

- Assess the association between ϵ_1 and ϵ_2

$$\phi = \phi(\epsilon_1, \epsilon_2)$$

Basic idea – from classical statistics textbooks

- Fit marginal regression for each response variable

$$
\begin{aligned}
Y_1 &\sim \beta_1^T X + \epsilon_1 \\
Y_2 &\sim \beta_2^T X + \epsilon_2
\end{aligned}
$$

- Assess the association between ϵ_1 and ϵ_2

$$\phi = \phi(\epsilon_1, \epsilon_2)$$

- Choices of $\phi(\cdot, \cdot)$ include
  - Pearson's correlation
  - Kendall's Tau
  - Schweizer-Wolff's Sigma (Copula based association measure)

Challenges for discrete data

- This study considers continuous, ordinal and binary response variables
- Challenges: how to obtain an appropriate residual ϵ?

Proposed method – based on a unified residual

- For a general parametric model

$$y|X \sim F(y, X, \beta),$$

we define a unified residual:

$$R(Y = y \mid \boldsymbol{x}, \boldsymbol{\beta}) = S(Y = y \mid \boldsymbol{x}, \boldsymbol{\beta}) - \mathbb{E}(S \mid \boldsymbol{x}, \boldsymbol{\beta}).$$

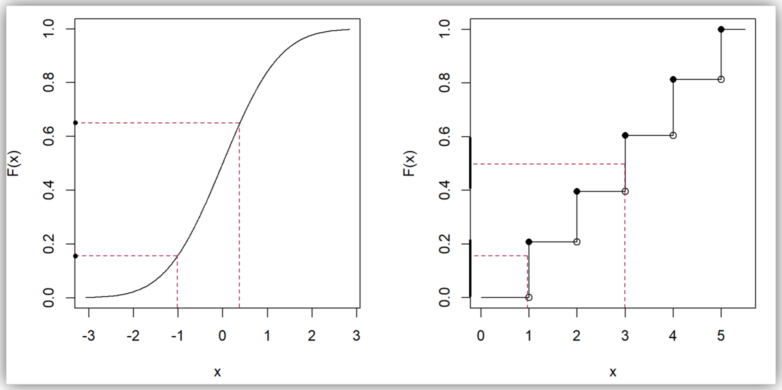
- S is a surrogate variable of Y, which is defined as

$$S(y; \boldsymbol{x}, \boldsymbol{\beta}) \sim U(F(y_-; \boldsymbol{x}, \boldsymbol{\beta}), F(y; \boldsymbol{x}, \boldsymbol{\beta})),$$

and $F(y_-; \boldsymbol{x}, \boldsymbol{\beta}) = \lim_{z \to y^-} F(z; \boldsymbol{x}, \boldsymbol{\beta})$ is the left limit on y.
- If $F(y, X, \beta)$ is correctly specified, $R \sim U(-1/2, 1/2)$.
- If Y is continuous, R is equivalent to the classical residual.

An illustration

Continuous (left) vs. ordinal (right)



Motivated by (Liu et al. 2021 JASA)

- Ordinal regression with cumulative link model

$$G^{-1}(\Pr\{Y \leq j\}) = \alpha_j - \boldsymbol{X}\beta, \quad j = 1, \ldots, J.$$

$G(\cdot)$ is link function, e.g., probit, logit, complementary log-log

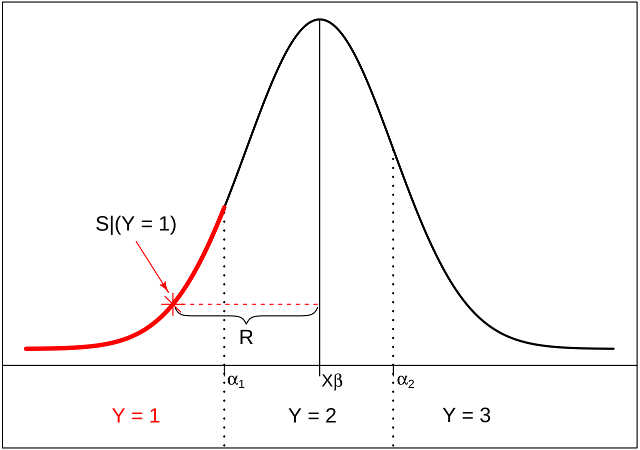
- Liu and Zhang (2018 JASA) proposed surrogate residual

$$R = S - E\{S \mid \boldsymbol{X}\} = S - E\{Z \mid \boldsymbol{X}\}$$

- S is a surrogate of latent variable Z

$$S \sim \begin{cases} Z \mid -\infty < Z \leq \alpha_1 & \text{if } Y = 1, \\ Z \mid \alpha_1 < Z \leq \alpha_2 & \text{if } Y = 2, \\ \cdots & \\ Z \mid \alpha_{J-1} < Z \leq +\infty & \text{if } Y = J. \end{cases}$$

An illustration



A key result

Theorem 1

Suppose that two outcome variables Y_1 and Y_2 follow general parametric models, $F_{Y_1|\boldsymbol{X}=\boldsymbol{x}}(y;\boldsymbol{x},\beta_1)$ and $F_{Y_2|\boldsymbol{X}=\boldsymbol{x}}(y;\boldsymbol{x},\beta_2)$. Let R_1 and R_2 be the unified residual. Then, for any given value $\boldsymbol{x}$, we have

$$(Y_1 \perp\!\!\!\perp Y_2)|(\boldsymbol{X}=\boldsymbol{x}) \Leftrightarrow (R_1 \perp\!\!\!\perp R_2)|(\boldsymbol{X}=\boldsymbol{x}). \qquad (1)$$

Furthermore, we have

$$(Y_1 \perp\!\!\!\perp Y_2)|(\boldsymbol{X}=\boldsymbol{x}) \text{ for all possible } \boldsymbol{x} \Rightarrow R_1 \perp\!\!\!\perp R_2. \qquad (2)$$

(2) implies that if $R_1 \not\perp\!\!\!\perp R_2$, then $(Y_1 \not\perp\!\!\!\perp Y_2)|(\boldsymbol{X}=\boldsymbol{x})$ for some $\boldsymbol{x}$.

Outline

A Kendall's τ-based measure

- We focus on a Kendall's τ-based measure

$$\mathcal{T}(Y_1, Y_2 : \boldsymbol{X}) = \tau(R_1, R_2).$$

Here, $\tau(\cdot, \cdot)$ is Kendall's tau, a rank-based correlation. Given n pairs of realizations $(\boldsymbol{r}_1, \boldsymbol{r}_2) = \{(r_{1i}, r_{2i})\}_{i=1}^{n}$, it is computed as

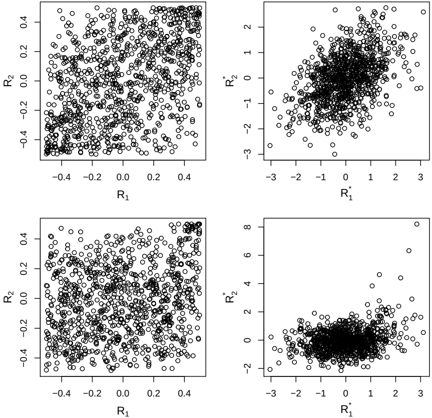
$$\hat{\mathcal{T}} = \hat{\tau}(\boldsymbol{r}_1, \boldsymbol{r}_2) = \binom{n}{2}^{-1} \sum_{i<j} \mathrm{sgn}(r_{1i} - r_{1j})\mathrm{sgn}(r_{2i} - r_{2j}).$$

- Other types of correlation measure can also be applied.

Key properties

- In the absence of covariates, the $\mathcal{T}$-measure is exactly the same as Kendall's tau, i.e., $\mathcal{T}(Y_1, Y_2) = \tau(Y_1, Y_2)$.
  - This property provides justification on quantifying confounding effect based on marginal and partial association.
- The $\mathcal{T}$-measure is invariant to monotonic transformations of either or both residual variables R_k's, i.e., $\mathcal{T}(Y_1, Y_2 : \boldsymbol{X}) = \tau(R_1, R_2) = \tau(h_1(R_1), h_2(R_2))$ where $h_k(\cdot)$'s are monotonic transformation functions.
  - This property allows a visualization tool, e.g., partial regression plot for $\Phi^{-1}(R_1)$ against $\Phi^{-1}(R_2)$.

Visualization example



Left: R_1 vs. R_2; right: $\Phi^{-1}(R_1)$ vs. $\Phi^{-1}(R_2)$. Top: linear association; bottom: nonlinear but monotone association

Estimation

- $\hat{\mathcal{T}}(Y_1, Y_2)$ is consistent under mild conditions
- To reduce variability due to randomness of R

$$\hat{\mathcal{T}}_M = \frac{1}{M}\sum_{m=1}^{M}\hat{\mathcal{T}}^{(m)},$$

  where $\hat{\mathcal{T}}^{(m)}$ is an estimate using the m-th simulation of R.
- Practically, $M = 30$ is sufficient.

Bootstrap-based inference

- For B bootstrap samples, obtain $\{\hat{\phi}_1^{(M)}, \hat{\phi}_2^{(M)}, \ldots, \hat{\phi}_B^{(M)}\}$ that forms a bootstrap distribution $\hat{F}_B(\phi)$
- $100(1-\alpha)\%$ confidence interval is

$$(\hat{F}_B^{(-1)}(\alpha/2), \hat{F}_B^{(-1)}(1-\alpha/2))$$

- Testing $H_0 : \phi = 0$, the p-value is

$$2\min(\hat{F}_B(0), 1-\hat{F}_B(0))$$

- Testing composite hypothesis $H_0 : |\phi| \leq \delta$, p-value is

$$2\min(\hat{F}_B(\delta), 1-\hat{F}_B(-\delta))$$

Outline

YOU Survey data – college student wellbeing study

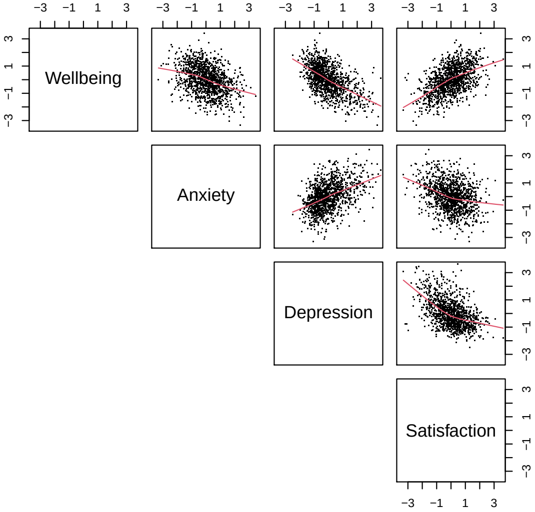
- Ongoing project starting from 2019
- Our sample contains responses from freshmen in 2019 and 2020
- Response variables: wellbeing (continuous), anxiety (ordinal 1-5), depression (continuous), and satisfaction (ordinal 1-7)?
- Two questions we attempt to answer using the proposed method:
  1. How do the covariates (risk factors) confound (moderate) the association between the response variables?
  2. How does COVID-19 affect those confounding effects?

Summary statistics

Summary statistics for numerical variables stratified by cohort

	2019					2020				
Variable	Min	Med.	Max	Mean	Std.	Min	Med.	Max	Mean	Std.
Wellbeing	8	56	100	53.25	17.03	4	52	100	51.67	18.24
Anxiety	1	3	5	3.34	1.03	1	3	5	3.22	1.04
Depression	0	29	100	32.91	21.40	0	29	100	32.20	20.28
Satisfaction	1	6	7	5.11	1.47	1	6	7	5.11	1.44
Financial strain	1	2	5	2.50	1.15	1	2	5	2.26	1.09
Healthiness	1	4	5	3.48	0.84	1	4	5	3.67	0.82
Loneliness	1	2	5	2.54	1.05	1	2	5	2.25	1.05
Age	15	18	35	18.73	2.29	15	18	35	18.91	2.64

Pairwise partial regression plot (2020 student cohort)



Marginal and partial associations (2020 student cohort)

Shown are the estimates of the association measure $\mathcal{T}$ and their standard errors (in the parenthesis).

	Marginal association			Partial association		
	anxiety	*depression*	*satisfaction*	*anxiety*	*depression*	*satisfaction*
wellbeing	-0.374	-0.472	0.466	-0.247	-0.383	0.341
	(0.019)	*(0.015)*	*(0.017)*	*(0.017)*	*(0.016)*	*(0.017)*
anxiety		0.421	-0.341		0.294	-0.207
		(0.018)	*(0.021)*		*(0.016)*	*(0.017)*
depression			-0.421			-0.310
			(0.018)			*(0.017)*

Confounding effect (2020 student cohort)

The confounding effect of the risk factors (presented using the percentage change of association after adjusting for risk factors). In the parenthesis are the standard errors.

	anxiety	*depression*	*satisfaction*
wellbeing	-34.0%	-18.9%	-26.8%
	(2.6%)	*(2.1%)*	*(2.0%)*
anxiety		-30.2%	-39.3%
		(2.2%)	*(2.7%)*
depression			-26.4%
			(2.3%)

Impact of COVID-19

Comparing confounding effect of individual risk factor for the association between Wellbeing and Anxiety, before and after COVID-19.

	physical healthiness	*loneliness*	*accommodation*	*financial strain*
	Confounding effects			
2019 cohort	-0.23 *(0.02)*	-0.17 *(0.01)*	-0.17 *(0.01)*	-0.21 *(0.01)*
2020 cohort	-0.31 *(0.02)*	-0.23 *(0.02)*	-0.21 *(0.01)*	-0.23 *(0.02)*
	Changes			
Numerical change	0.08	0.06	0.04	0.02
Percentage change	35%	35%	24%	10%
p-value	0.03	< 0.01	< 0.01	0.58

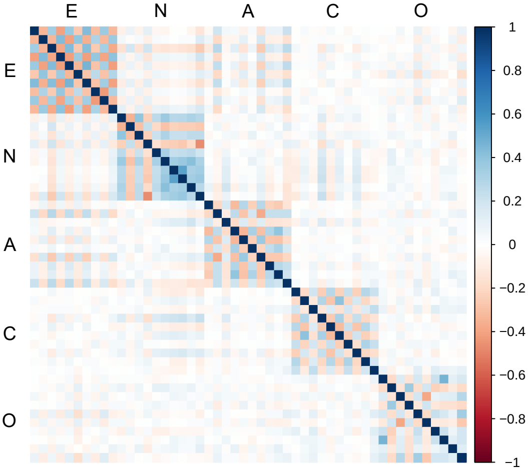
Conclusions from the real data analysis

- The associations between all pairs of the outcome variables (mental health) are significantly confounded by the set of covariates.
- The partial associations remain significant after controlling for covariates.
- Except for financial strain, the confounding effects of all other risk factors are increased in early stage of COVID-19.
- A potential implication: Under public health disruption, university administrators may develop programs that help improving students physical health, social connections and learning and living environment.

Big-Five personality traits

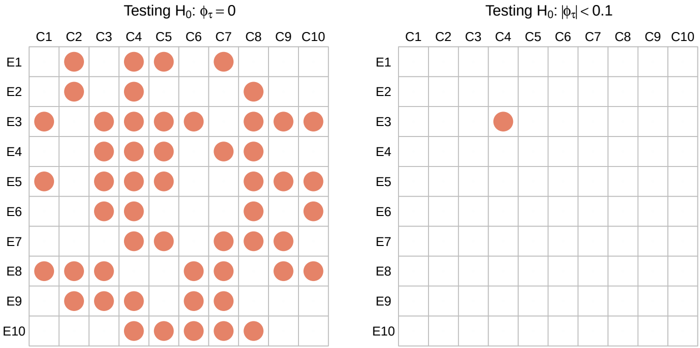
- Big Five include: Extraversion (E), Neuroticism (N), Agreeableness (A), Conscientiousness (C), and Openness to experience (O)
- Each trait is measured by 10 instruments (survey questions)
- 50 ordinal variables in total
- Potential confounding factors include: *age*, *gender*, *engnat* (is English native language), *hand* (what hand does participant use to write with), and *source* (how the participant came to the test)
- Data: `https://openpsychometrics.org/_rawdata/`

Partial association matrix in color scale



A composite hypothesis test

Testing partial independence between Extraversion and Conscientiousness (orange dot indicates p-value < 0.05)



Summary

- A unifying framework for assessing partial association between mixed data: continuous, ordinal and binary
  - The measure
  - Graphical tools
  - Hypothesis testing
- The framework relies on the a unified residual
- The R package, PAsso is available on CRAN.
- Future studies:
  - Extending the method for nominal data
  - Feature selection (identify confounders)
  - Causal discovery?
  - Comparing results for 2021 and 2022 cohort data
  - May worth to study the direct effect of risk factor to the wellbeing (not sure how our method can help).

Thank You!