

● Framework for Matching Mixed Covariates and Handling Missing Data Using Latent Variables

STA714 FINAL

2024020409 권휘준, 2024021609 정윤주, 2021150470 백소윤

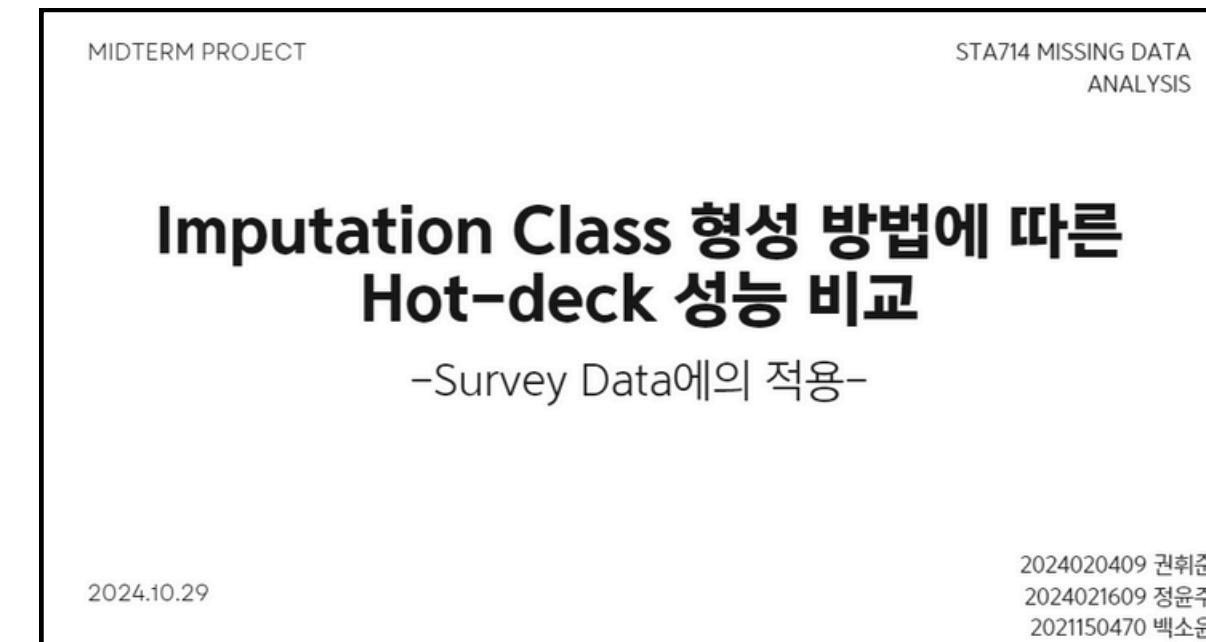
CONTENTS



- 01 Introduction
- 02 Paper Overview
- 03 Simulation
- 04 Apply to Data
- 05 Result
- 06 Reference

Final Project Idea

- 중간 프로젝트 데이터를 다시 활용해 보완하여 진행해보자



- Imputation within Class method에서 Class 생성 시 활용한 propensity score에 주목하여 진행해보자

“다양한 타입의 데이터를 모두 고려해 matching 시킬 수 있을까?”



반응변수 (Y)	독립변수 (X)						
	공복혈당 (HE_glu)	BMI (HE_BMI)	당화혈색소 (HE_HbA1c)	...	성별 (sex)	빈혈여부 (HE_anem)	...
94	26.507	5.6	...	2	0	...	4
84	27.152	5.3	...	1	0	...	3
84	21.308	5	...	2	0	...	4
...

DATA FEATURE

- 연속형, 범주형(이항 & 다항) 변수 모두 존재
 - Matching에서 distance를 활용하는 데 문제 존재
 - mahalanobis distance의 경우, 범주형에는 적용 할 수 없음
 - gower distance의 경우, weight 지정에 문제 존재
- 어떤 방식을 활용하면 distance를 기반으로 matching할 수 있을까?
 - 관련 논문을 찾아보자!

송 주 원 교수지도
석사학위논문

잠재변수를 이용한 혼합형 공변량을
포함한 자료에 대한 짹짓기 방법

이 논문을 통계학석사 학위논문으로 제출함.

2012년 12월 일

고려대학교 대학원

통계학과
윤형석



MAIN IDEA OF PAPER

1. latent variable을 통해 범주형(이항변수) 내의 연속형 분포를 찾아낸다
 - a. latent variable을 찾을 때 Gibbs sampler 활용
2. Matching에 latent variable을 활용
 - a. latent variable 기반 propensity score matching
 - b. latent variable을 활용하여 mahalanobis distance 계산 후, distance 기반 matching
 - c. propensity score가 유사한 그룹 내에서 mahalanobis distance 기준 matching

DETAILS - MATCHING

Only propensity score

[정규 근사를 위해 아래 식을 활용]

$$\hat{q}(x) = \log\left(\frac{\hat{e}(x)}{1 - \hat{e}(x)}\right)$$

1. Object Random Ordering

a. 개체별 공평한 짹짓기 기회 위함

2. 기준 group의 각 개체의 q값과 가장 가까운 값을 가진 group의 개체와 짹짓고, 해당 개체는 제거

3. 나머지 관찰치들에도 동일하게 진행

propensity & mahalanobis

[Mahalanobis Distance]

$$d(u_1 - u_2) = (u_1 - u_2)'Cov^{-1}(u_1 - u_2)$$

1. Object Random Ordering

a. 개체별 공평한 짹짓기 기회 위함

2. 기준 group의 각 개체와 mahalanobis 거리기준 가장 가까운 개체값과 짹짓고 해당 개체 제거

3. 나머지 관찰치들에도 동일하게 진행

propensity group & mahalanobis distance

[threshold]

$$c = a\sigma, 0 < a < 1$$

$$\sigma = ((\sigma_1^2 + \sigma_{0R}^2)/2)^{1/2}$$

1. Object Random Ordering

a. 개체별 공평한 짹짓기 기회 위함

2. 기준 group의 관찰치에서 어떤 상수 c보다 작은 q값을 갖는 다른 그룹의 관측치를 찾음

a. 해당 범위 내 관측치 없다면 q값 기준 가장 가까운 관측치 짹짓기

3. 모인 개체들 중 mahalanobis 거리 기준 짹짓고 해당 개체 제거

DETAILS - LATENT VARIABLE

Variables

$$X_g \sim N_p(\mu_g, \Sigma_g),$$

$$Z_g \sim N_l(X_g^T \beta, I_{p \times p}), \quad g = 1, 2,$$

$$Y_{g,i} = 1 \text{ if } Z_{g,i} > 0,$$

$$Y_{g,i} = 0 \text{ if } Z_{g,i} < 0, \quad i = 1, \dots, l,$$

$$Y_{g,i} \sim Bernoulli(P_{g,i}), \quad P_{g,i} = P(Y_{g,i} = 1) = \Phi(X_g^T \beta_i).$$

- p 개의 연속형 변수 X
- L개의 이항 변수 Y
- group = 1, 2

Gibbs sampling

[joint posterior distribution of Z and beta]

$$\pi(\beta, Z|y) = C\pi(\beta) \prod_{i=1}^l \{1(Z_i > 0)1(y_i = 1) + 1(Z_i \leq 0)1(y_i = 0)\} \times \Phi(Z_i; X_i^T \beta, 1)$$

Gibbs sampler using conditional distribution

1) beta conditional distribution

$$\pi(\beta|y, Z) = C\pi(\beta) \prod_{i=1}^N \Phi(Z_i; X_i^T \beta, 1)$$

2) Z conditional distribution

$Z_i|y, \beta$ 는 $y_i = 1$ 일 때 $N(X_i^T \beta, 1)$ 에서 0 좌측으로의 절단분포

$Z_i|y, \beta$ 는 $y_i = 0$ 일 때 $N(X_i^T \beta, 1)$ 에서 0 우측으로의 절단분포

process

1. beta's initial value = LSE = $(X'X)^{-1}X'y$
2. sampling Z using Z conditional
3. sampling beta using beta conditional

Setting of Simulation

$$X_g = \begin{bmatrix} X_{g1}^T \\ X_{g2}^T \\ X_{g3}^T \end{bmatrix} \sim N_3(\mu_g, \Sigma_g),$$

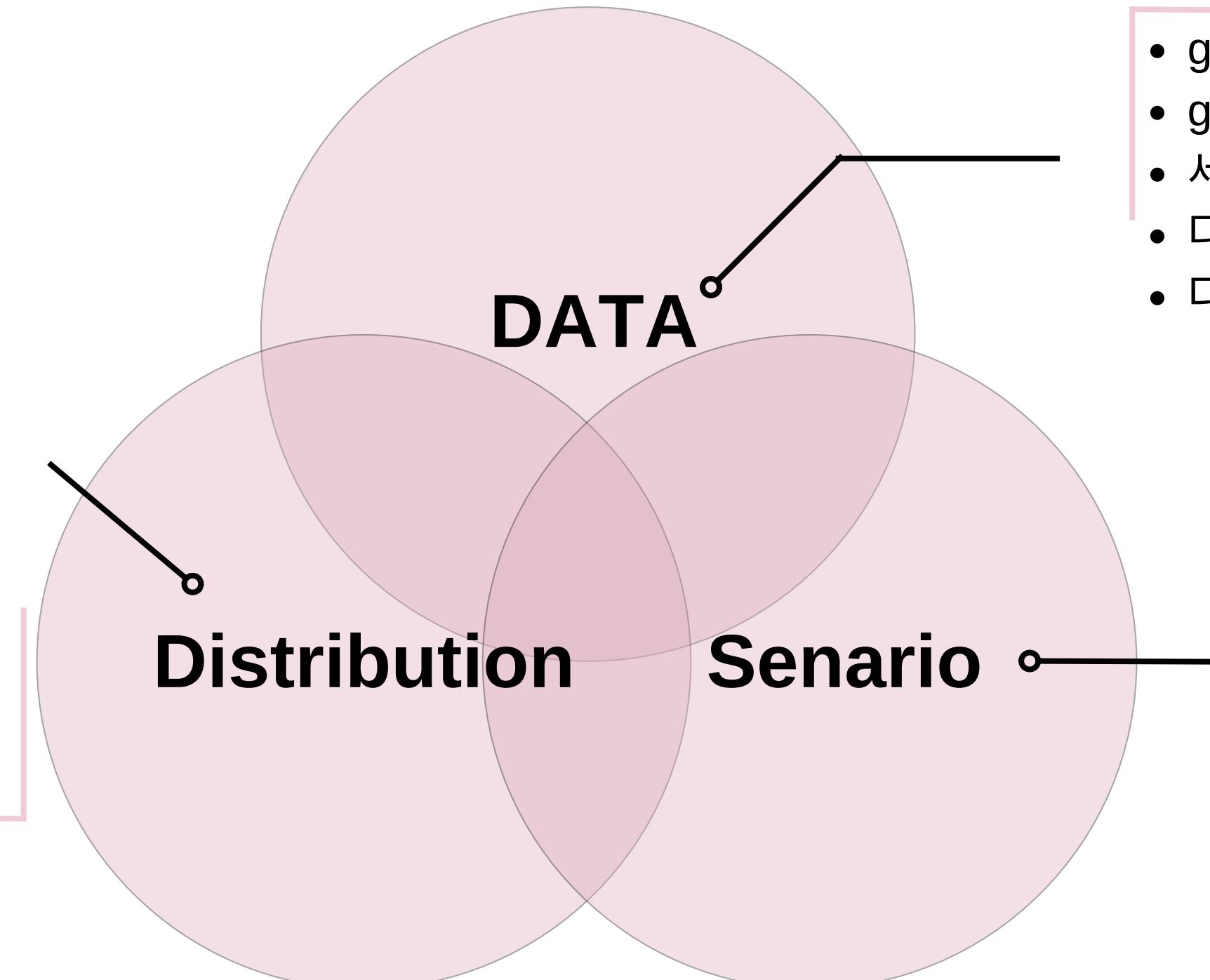
$$\mu_g = \begin{bmatrix} \mu_{g1} \\ \mu_{g2} \\ \mu_{g3} \end{bmatrix}, \quad \Sigma_g \begin{pmatrix} \sigma_{g,1}^2 & \sigma_{g,12} & \sigma_{g,13} \\ \sigma_{g,2}^2 & \sigma_{g,23} & \\ \sigma_{g,3}^2 & & \end{pmatrix}$$

$$Z_g \sim N(X_g^{*T} \beta)$$

$$Y_g \sim \text{Bernoulli}(\Phi(X_g^{*T} \beta))$$

$$X_g^* = \begin{bmatrix} 1_{1 \times n_g} \\ X_g \end{bmatrix}, \quad \beta = [\beta_1, \beta_2, \beta_3, \beta_4, \beta_5], \quad g = 1, 2$$

$$\beta = \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 \\ 0.5 & 0.7 & 0.3 & 0.5 & 0.5 \\ -0.1 & 0.1 & 0.2 & 0.15 & 0.3 \\ -0.2 & -0.1 & -0.2 & 0.25 & 0.7 \\ 0.2 & -0.5 & 0.15 & -0.15 & 0.5 \end{pmatrix}$$



- group = 1 : 100개
- group = 2 : 300개
- 세 개의 연속형 변수 X
- 다섯 개의 이항 변수 Y
- 다섯 개의 잠재 변수 Z

$$\mu_1 : (-1 \ 0 \ 1)^T \quad \mu_2 : (-2 \ 1 \ 2)^T \quad vs \quad \mu_1 : (-1 \ 0 \ 1)^T \quad \mu_2 : (-4 \ -3 \ 4)^T$$

$$\sigma_i^2 : 1 \quad vs \quad 10 \quad i = 1, \dots, n_g, \quad g = 1, 2$$

- 1) 두 집단의 위치 차이 크/작을 경우
- 2) 각 그룹의 X 변수들의 분산이 크/작을 경우

Simulation Results (in Paper)

상황	1집단의평균 분산/공분산	2집단의평균 분산/공분산	짝을 찾는 관찰치의 비율(%)
1	$\mu=(-1,0,1)$	$\mu=(-2,1,2)$	97%
	$\sigma(X_{ii})=1$	$\sigma(X_{ii})=1$	
	$\sigma(X_{ij})=0.7$	$\sigma(X_{ij})=0.7$	
2	$\mu=(-1,0,1)$	$\mu=(-2,1,2)$	92%
	$\sigma(X_{ii})=1$	$\sigma(X_{ii})=10$	
	$\sigma(X_{ij})=0.7$	$\sigma(X_{ij})=0.7$	
3	$\mu=(-1,0,1)$	$\mu=(-2,1,2)$	95%
	$\sigma(X_{ii})=10$	$\sigma(X_{ii})=1$	
	$\sigma(X_{ij})=0.7$	$\sigma(X_{ij})=0.7$	
4	$\mu=(-1,0,1)$	$\mu=(-2,1,2)$	95%
	$\sigma(X_{ii})=10$	$\sigma(X_{ii})=10$	
	$\sigma(X_{ij})=0.7$	$\sigma(X_{ij})=0.7$	
5	$\mu=(-1,0,1)$	$\mu=(-4,-3,4)$	39%
	$\sigma(X_{ii})=1$	$\sigma(X_{ii})=1$	
	$\sigma(X_{ij})=0.7$	$\sigma(X_{ij})=0.7$	
6	$\mu=(-1,0,1)$	$\mu=(-4,-3,4)$	64%
	$\sigma(X_{ii})=1$	$\sigma(X_{ii})=10$	
	$\sigma(X_{ij})=0.7$	$\sigma(X_{ij})=0.7$	
7	$\mu=(-1,0,1)$	$\mu=(-4,-3,4)$	42%
	$\sigma(X_{ii})=10$	$\sigma(X_{ii})=1$	
	$\sigma(X_{ij})=0.7$	$\sigma(X_{ij})=0.7$	
8	$\mu=(-1,0,1)$	$\mu=(-4,-3,4)$	76%
	$\sigma(X_{ii})=10$	$\sigma(X_{ii})=10$	
	$\sigma(X_{ij})=0.7$	$\sigma(X_{ij})=0.7$	

Group 1,2간 평균차이가 크지 않는 경우

거의 대부분 개체들이 짹지어짐을 확인 가능

Group 1,2간 평균 차이가 큰 경우

X의 평균이 멀어질수록 짹지어진 비율이 줄어듬을
확인 가능

Simulation Results (in Paper)

상황	1집단의평균 분산/공분산	2집단의평균 분산/공분산	짝을 찾는 관찰치의 비율(%)
1	$\mu=(-1,0,1)$	$\mu=(-2,1,2)$	97%
	$\sigma(X_{ii})=1$	$\sigma(X_{ii})=1$	
	$\sigma(X_{ij})=0.7$	$\sigma(X_{ij})=0.7$	
2	$\mu=(-1,0,1)$	$\mu=(-2,1,2)$	92%
	$\sigma(X_{ii})=1$	$\sigma(X_{ii})=10$	
	$\sigma(X_{ij})=0.7$	$\sigma(X_{ij})=0.7$	
3	$\mu=(-1,0,1)$	$\mu=(-2,1,2)$	95%
	$\sigma(X_{ii})=10$	$\sigma(X_{ii})=1$	
	$\sigma(X_{ij})=0.7$	$\sigma(X_{ij})=0.7$	
4	$\mu=(-1,0,1)$	$\mu=(-2,1,2)$	95%
	$\sigma(X_{ii})=10$	$\sigma(X_{ii})=10$	
	$\sigma(X_{ij})=0.7$	$\sigma(X_{ij})=0.7$	
5	$\mu=(-1,0,1)$	$\mu=(-4,-3,4)$	39%
	$\sigma(X_{ii})=1$	$\sigma(X_{ii})=1$	
	$\sigma(X_{ij})=0.7$	$\sigma(X_{ij})=0.7$	
6	$\mu=(-1,0,1)$	$\mu=(-4,-3,4)$	64%
	$\sigma(X_{ii})=1$	$\sigma(X_{ii})=10$	
	$\sigma(X_{ij})=0.7$	$\sigma(X_{ij})=0.7$	
7	$\mu=(-1,0,1)$	$\mu=(-4,-3,4)$	42%
	$\sigma(X_{ii})=10$	$\sigma(X_{ii})=1$	
	$\sigma(X_{ij})=0.7$	$\sigma(X_{ij})=0.7$	
8	$\mu=(-1,0,1)$	$\mu=(-4,-3,4)$	76%
	$\sigma(X_{ii})=10$	$\sigma(X_{ii})=10$	
	$\sigma(X_{ij})=0.7$	$\sigma(X_{ij})=0.7$	

Variance 측면 (평균 차이가 클 때)

두 집단 중 어느 하나의 집단에서 분산이 작아지는 경우, 짹짓는 비율이 그렇지 않은 경우에 비해 줄어듬을 확인 가능

“즉, 두 집단 간 평균차이가 크지 않거나,
크더라도 각 집단의 분산이 큰 경우 짹짓기가 적절할 것”

Simulation Results (in Paper)

matching metric

[매칭된 개체들의 X 평균 차이의 표준화]

$$(\overline{X_1} - \overline{X_{0M}}) / [(S_1^{2+} S_{0R}^2) / 2]^{1/2}$$

- OM : 짹지어진 집단을 의미(평균)
- OR : 짹지어진 집단을 의미(분산)

[매칭된 개체들의 성향점수 평균 차이의 표준화]

$$(\overline{\hat{q}(x)_1} - \overline{\hat{q}_{0M}}) / [(S_1^{2+} S_{0R}^2) / 2]^{1/2}$$

[Y의 일치율]

- 각 개체가 Y에서 얼마나 일치하는지를 비교

method name

[방법 0]

random matching

[방법 1*]

X, Y만을 이용해 성향점수 추정, 성향점수 기준
가장 가까운 값을 matching

[방법 1]

잠재변수 Z를 추정한 후 X,Z를 이용해 성향점수
추정, 이 성향점수 기준 가장 가까운 값을
matching

[방법 2]

X, Z, 성향점수를 이용해 측정한 mahalanobis
거리가 짧은 개체들을 짹짓기

[방법 3]

성향점수 차이가 일정 값보다 작은 개체 중
mahalanobis 거리가 가장 짧은 개체들을 짹짓기

03 Simulation

Simulation Results (in Paper)

	1집단의 평균 분산/공분산	2집단의 평균 분산/공분산	평균	무작위 짝짓기 방법	방법1*	방법1	방법2	방법3
1	$\mu=(-1,0,1)$	$\mu=(-2,1,2)$	X의 거리	0.648	0.521	0.480	0.698	0.479
	$\sigma(X_{ii})=1$	$\sigma(X_{ii})=1$	$q(x)$ 의 거리	3.719	2.112	3.438	3.499	3.438
	$\sigma(X_{ij})=0.7$	$\sigma(X_{ij})=0.7$	Y의 일치율	58.2%	53.0%	55.0%	79.8%	60.4%
2	$\mu=(-1,0,1)$	$\mu=(-2,1,2)$	X의 거리	0.381	0.167	0.228	0.185	0.228
	$\sigma(X_{ii})=1$	$\sigma(X_{ii})=10$	$q(x)$ 의 거리	4.240	1.571	3.162	3.269	3.162
	$\sigma(X_{ij})=0.7$	$\sigma(X_{ij})=0.7$	Y의 일치율	54.6%	53.4%	57.6%	82.6%	64.0%
3	$\mu=(-1,0,1)$	$\mu=(-2,1,2)$	X의 거리	0.317	0.408	0.316	0.287	0.314
	$\sigma(X_{ii})=10$	$\sigma(X_{ii})=1$	$q(x)$ 의 거리	3.868	2.065	3.635	3.635	3.635
	$\sigma(X_{ij})=0.7$	$\sigma(X_{ij})=0.7$	Y의 일치율	54.8%	54.2%	51.6%	71.2%	53.4%
4	$\mu=(-1,0,1)$	$\mu=(-2,1,2)$	X의 거리	0.261	0.059	0.220	0.103	0.104
	$\sigma(X_{ii})=10$	$\sigma(X_{ii})=10$	$q(x)$ 의 거리	1.275	0.605	0.974	1.087	1.050
	$\sigma(X_{ij})=0.7$	$\sigma(X_{ij})=0.7$	Y의 일치율	53.6%	55.8%	58.0%	81.2%	80.8%

	1집단의 평균 분산/공분산	2집단의 평균 분산/공분산	평균	무작위 짝짓기 방법	방법1*	방법1	방법2	방법3
5	$\mu=(-1,0,1)$	$\mu=(-4,-3,4)$	X의 거리	2.538	2.446	2.423	2.301	2.423
	$\sigma(X_{ii})=1$	$\sigma(X_{ii})=1$	$q(x)$ 의 거리	13.902	10.187	13.647	13.814	13.647
	$\sigma(X_{ij})=0.7$	$\sigma(X_{ij})=0.7$	Y의 일치율	53.0%	56.0%	52.2%	64.0%	52.2%
6	$\mu=(-1,0,1)$	$\mu=(-4,-3,4)$	X의 거리	1.196	0.623	0.768	1.194	0.764
	$\sigma(X_{ii})=1$	$\sigma(X_{ii})=10$	$q(x)$ 의 거리	4.352	4.133	3.724	3.934	3.725
	$\sigma(X_{ij})=0.7$	$\sigma(X_{ij})=0.7$	Y의 일치율	58.4%	51.4%	57.2%	75.6%	60.6%
7	$\mu=(-1,0,1)$	$\mu=(-4,-3,4)$	X의 거리	0.866	1.033	0.855	0.808	0.857
	$\sigma(X_{ii})=10$	$\sigma(X_{ii})=1$	$q(x)$ 의 거리	5.117	6.711	4.988	5.001	4.988
	$\sigma(X_{ij})=0.7$	$\sigma(X_{ij})=0.7$	Y의 일치율	57.2%	52.6%	55.8%	68.6%	57.8%
8	$\mu=(-1,0,1)$	$\mu=(-4,-3,4)$	X의 거리	0.805	0.402	0.302	0.569	0.407
	$\sigma(X_{ii})=10$	$\sigma(X_{ii})=10$	$q(x)$ 의 거리	2.248	1.227	1.421	1.746	1.544
	$\sigma(X_{ij})=0.7$	$\sigma(X_{ij})=0.7$	Y의 일치율	54.0%	52.8%	56.0%	71.4%	65.8%

Simulation Results (in Paper)

1) 성향점수만을 활용(방법1*) vs 성향점수 group 후 mahalanobis distance 활용
 : 다른 방법들에 비해 좋은 성능 (두 방법간 차이는 크지 않음)

2) 두 자료의 평균이 상대적으로 멀었던 5번 상황

: 다섯 방법 모두 성향 점수 차이가 컸음

: 따라서, 성향점수를 활용한 방법1*, 방법3의 결과가 좋지 않았음

3) 4가지 방법 모두 무작위 matching보다는 좋은 성능

4) X, Y만을 활용한 방법1*의 경우, 평균 차이가 컼을 때, 일치율 낮음(무작위 제외)

: 이항형 활용한 성향점수보다는 잠재변수 변환 후 성향점수에 활용이 더 좋음

5) Mahalanobis distance만을 활용하는 방법2, 대부분 다른 방법보다 Y일치율 및 공변량 거리 관점 좋은 성능

: Mahalanobis distance를 활용한 방법이 다른 방법들보다 공변량이 비슷한 성격을 가지므로, 좋은 방식

*X, Y가 독립일 때에도 비슷한 결과임

	1집단의 평균 분산/공분산	2집단의 평균 분산/공분산	평균 분산/공분산	무작위 짹짓기 방법	방법1*	방법1	방법2	방법3
1	$\mu=(-1,0,1)$ $\sigma(X_{ii})=1$ $\sigma(X_{ij})=0.7$	$\mu=(-2,1,2)$ $\sigma(X_{ii})=1$ $\sigma(X_{ij})=0.7$	$\mu=(-1,0,1)$ $\sigma(X_{ii})=1$ $\sigma(X_{ij})=0.7$	$q(x)$ 의 거리 53.0% 55.0%	2.538 13.902 53.0%	2.446 10.187 56.0%	2.423 13.647 52.2%	2.301 13.814 64.0%
2	$\mu=(-1,0,1)$ $\sigma(X_{ii})=1$ $\sigma(X_{ij})=0.7$	$\mu=(-2,1,2)$ $\sigma(X_{ii})=1$ $\sigma(X_{ij})=0.7$	$\mu=(-1,0,1)$ $\sigma(X_{ii})=1$ $\sigma(X_{ij})=0.7$	$q(x)$ 의 거리 53.0% 55.0%	1.196 4.352 58.4%	0.623 4.133 51.4%	0.768 3.724 57.2%	1.194 3.934 75.6%
3	$\mu=(-1,0,1)$ $\sigma(X_{ii})=10$ $\sigma(X_{ij})=0.7$	$\mu=(-2,1,2)$ $\sigma(X_{ii})=10$ $\sigma(X_{ij})=0.7$	$\mu=(-1,0,1)$ $\sigma(X_{ii})=10$ $\sigma(X_{ij})=0.7$	$q(x)$ 의 거리 54.8% 54.2%	0.866 5.117 0.866	1.033 6.711 1.033	0.855 4.988 0.855	0.808 5.001 0.808
4	$\mu=(-1,0,1)$ $\sigma(X_{ii})=10$ $\sigma(X_{ij})=0.7$	$\mu=(-2,1,2)$ $\sigma(X_{ii})=10$ $\sigma(X_{ij})=0.7$	$\mu=(-1,0,1)$ $\sigma(X_{ii})=10$ $\sigma(X_{ij})=0.7$	Y 의 일치율 58.0% 58.0%	2.248 2.248 54.0%	1.227 1.421 52.8%	1.421 1.746 56.0%	1.544 1.544 71.4%
5	$\mu=(-1,0,1)$ $\sigma(X_{ii})=1$ $\sigma(X_{ij})=0.7$	$\mu=(-4,-3,4)$ $\sigma(X_{ii})=1$ $\sigma(X_{ij})=0.7$	$\mu=(-4,-3,4)$ $\sigma(X_{ii})=10$ $\sigma(X_{ij})=0.7$	Y 의 일치율 53.0% 56.0%	2.538 13.902 53.0%	2.446 10.187 56.0%	2.423 13.647 52.2%	2.301 13.814 64.0%
8	$\mu=(-1,0,1)$ $\sigma(X_{ii})=10$ $\sigma(X_{ij})=0.7$	$\mu=(-4,-3,4)$ $\sigma(X_{ii})=10$ $\sigma(X_{ij})=0.7$	$\mu=(-1,0,1)$ $\sigma(X_{ii})=10$ $\sigma(X_{ij})=0.7$	Y 의 일치율 54.0% 52.8%	2.538 13.902 53.0%	2.446 10.187 56.0%	2.423 13.647 52.2%	2.301 13.814 64.0%

Simulation Code

Bayesian Analysis of Binary and Polychotomous Response Data

JAMES H. ALBERT and SIDDHARTHA CHIB*

A vast literature in statistics, biometrics, and econometrics is concerned with the analysis of binary and polychotomous response data. The classical approach fits a categorical response regression model using maximum likelihood, and inferences about the model are based on the associated asymptotic theory. The accuracy of classical confidence statements is questionable for small sample sizes. In this article, exact Bayesian methods for modeling categorical response data are developed using the idea of data augmentation. The general approach can be summarized as follows. The probit regression model for binary outcomes is seen to have an underlying normal regression structure on latent continuous data. Values of the latent data can be simulated from suitable truncated normal distributions. If the latent data are known, then the posterior distribution of the parameters can be computed using standard results for normal linear models. Draws from this posterior are used to sample new latent data, and the process is iterated with Gibbs sampling. This data augmentation approach provides a general framework for analyzing binary regression models. It leads to the same simplification achieved earlier for censored regression models. Under the proposed framework, the class of probit regression models can be enlarged by using mixtures of normal distributions to model the latent data. In this normal mixture class, one can investigate the sensitivity of the parameter estimates to the choice of “link function,” which relates the linear regression estimate to the fitted probabilities. In addition, this approach allows one to easily fit Bayesian hierarchical models. One specific model considered here reflects the belief that the vector of regression coefficients lies on a smaller dimension linear subspace. The methods can also be generalized to multinomial response models with $J > 2$ categories. In the ordered multinomial model, the J categories are ordered and a model is written linking the cumulative response probabilities with the linear regression structure. In the unordered multinomial model, the latent variables have a multivariate normal distribution with unknown variance-covariance matrix. For both multinomial models, the data augmentation method combined with Gibbs sampling is outlined. This approach is especially attractive for the multivariate probit model, where calculating the likelihood can be difficult.

KEY WORDS: Binary probit; Data augmentation; Gibbs sampling; Hierarchical Bayes modeling; Latent data; Logit model; Multinomial model; Bayesian analysis; Student-t link function

PAPER GITHUB LINK



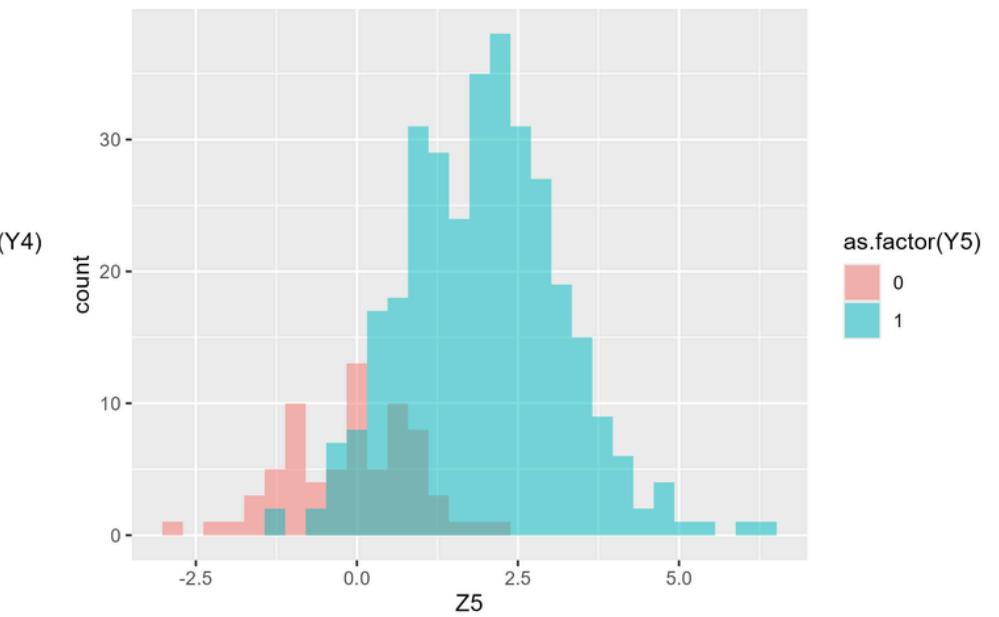
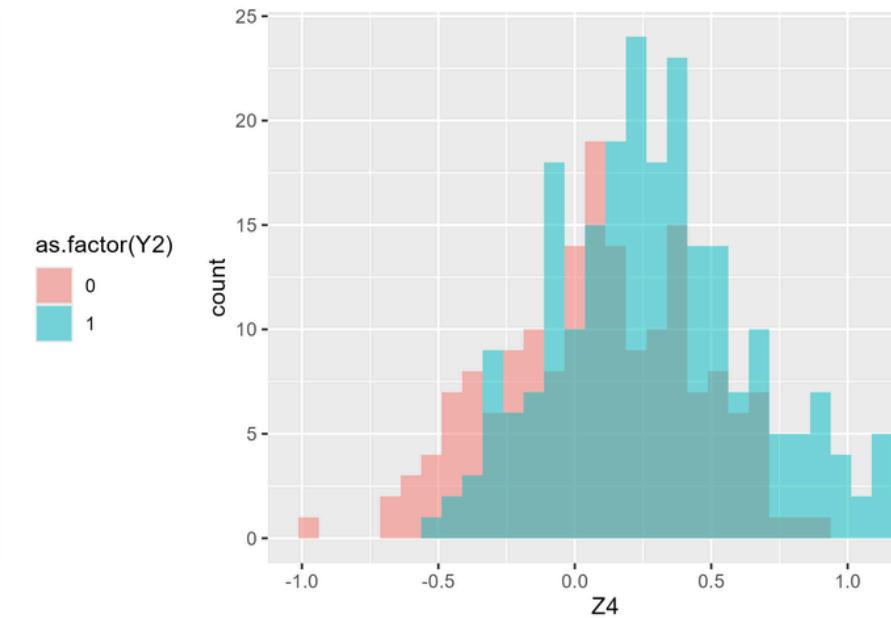
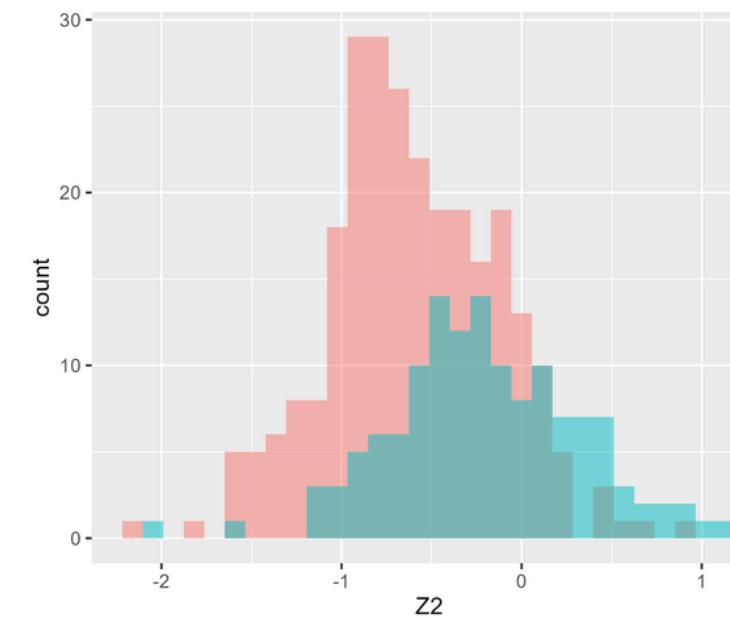
[PACKAGE FUNCTIONS]

Binary_Gibbs_Functions.R	Included package dependency. stats.
Binary_Gibbs_Functions_Prediction.R	Included package dependency. stats.
Binary_Gibbs_PosteriorDensityPlots.R	Worked on documentation of the functions. .
Binary_Gibbs_TraceplotsPlots.R	Included and documented plotting functions
Binary_Gibbs_TrainingAccuracy.R	Included package dependency. stats.
Poly_Gibbs_BetaPosterior.R	Included and documented plotting functions
Poly_Gibbs_BetaTrace.R	Included and documented plotting functions
Poly_Gibbs_GammaPosterior.R	Included and documented plotting functions
Poly_Gibbs_GammaTrace.R	Finished the first draft of the vignette.
Poly_Gibbs_ModelFitting.R	Included and documented MultinomGibbs_p
Poly_Gibbs_Prediction.R	Worked on function documentation.
Poly_Gibbs_TestAccuracy.R	Worked on function documentation.

Reproduce

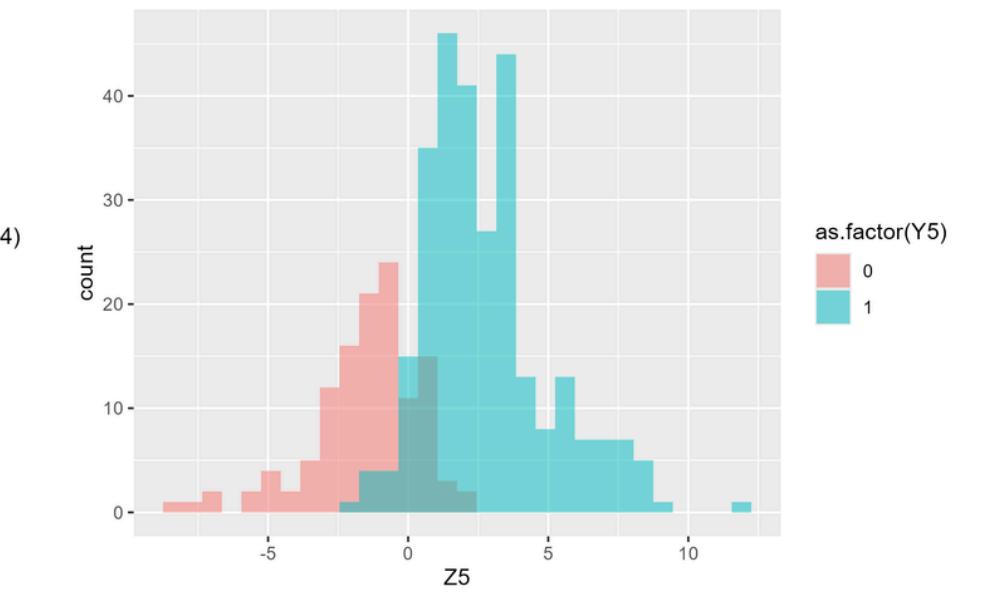
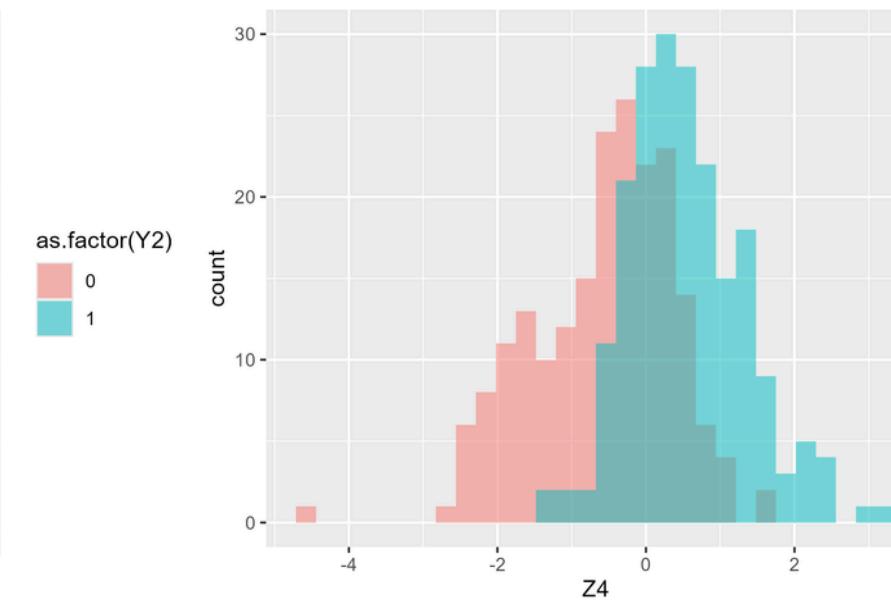
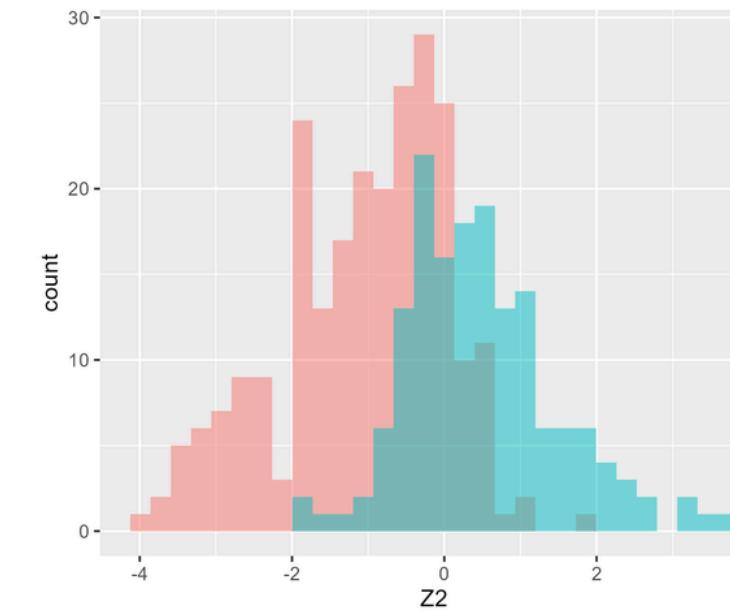
Situation 1	
GROUP1	GROUP2
$\mu = (-1, 0, 1)$	$\mu = (-2, 1, 2)$
$\sigma_{ii} = 1$	$\sigma_{ii} = 1$

$\sigma_{ij} = 0.7$



Situation 2	
GROUP1	GROUP2
$\mu = (-1, 0, 1)$	$\mu = (-2, 1, 2)$
$\sigma_{ii} = 1$	$\sigma_{ii} = 10$

$\sigma_{ij} = 0.7$



*more graphs : appendix

Reproduce

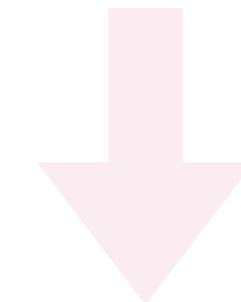
Situation 1	
GROUP1	GROUP2
mu = (-1,0,1)	mu = (-2,1,2)
sigma_ii = 1	sigma_ii = 1
sigma_ij = 0.7	sigma_ij = 0.7

result	method1*	method1	method2
x의 표준화 거리	0.63062	0.63085	0.69243
q의 표준화 거리	2.37601	2.35974	2.3301
y의 일치도	0.574	0.566	0.63

Situation 2	
GROUP1	GROUP2
mu = (-1,0,1)	mu = (-2,1,2)
sigma_ii = 1	sigma_ii = 10
sigma_ij = 0.7	sigma_ij = 0.7

result	method1*	method1	method2
x의 표준화 거리	0.08617	0.14746	0.31486
q의 표준화 거리	0.17828	0.68964	0.7585
y의 일치도	0.57	0.56	0.644

- 논문에서 metric 계산식에 모호한 부분이 존재하기에, 완전히 똑같은 결과는 아님
- 그러나, latent variable와 mahalanobis distance를 모두 활용한 method2에서 좋은 y 일치율을 보임을 확인가능
- X의 표준화 거리 기준, method1*이 X간 거리가 매우 작음을 확인가능
- 전반적인 성능 차이도 논문 결과와 유사함



simulation code와 동일한 logic으로 데이터에 적용!

PAPER CONCLUSION

Q : 왜 latent variable을 활용하는가?

논문 결과, 이항 변수를 적용한 성향 점수보다는, latent variable을 활용한 성향 점수가 matching에 있어 더 좋은 성능을 보였음

Q : Mahalanobis distance를 활용하는 이유?

논문에서 활용한 5가지 방법 중 공통적으로 mahalanobis distance를 활용한 방법이 대체적으로 좋은 성능을 보였음

Q : Matching과 Imputation의 관계?

Hotdeck within class를 적용할 때, 공변량 관점에서 유사한 class를 생성할 때, matching을 활용할 수 있다고 판단함

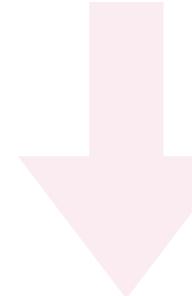
Q : 실제 데이터 적용 시의 point?

1. one-to-one matching의 경우, 하나의 값으로 대체 되며, greedy matching issue가 있을 수 있다고 판단
2. with replacement, 일대다 matching을 활용하면 imputation 관점에서 더 좋은 성능을 보일 것이라 생각하여 해당 방법 두 가지를 비교하고자 함.

03 Simulation

IMPUTATION with SIMULATION

Situation 1	
GROUP1	GROUP2
mu = (-1,0,1)	mu = (-2,1,2)
sigma_ii = 1	sigma_ii = 1
sigma_jj = 0.7	sigma_jj = 0.7



X1 : missing generating
MAR10 (survey data와 동일 logic)

m : missing indicator

```
> res
# A tibble: 400 × 19
   X2    X3    Y1    Y2    Y3    Y4    Y5    Z1    Z2    Z3    Z4    Z5    m    X1_1   X1_2   X1_3   X1_4   X1_5   X1
   <dbl> <dbl>
1 0.440  0.960  1     0     1     0     1  0.856 -0.0424  0.303  0.470  1.22  0  -0.766 -0.766 -0.766 -0.766 -0.766
2 0.763  2.04   0     0     0     1     1  0.982 -0.482   0.432  0.479  2.18  0  -0.436 -0.436 -0.436 -0.436 -0.436
3 1.23   1.36   1     0     1     1     1  0.677 -0.300   0.311  0.868  2.33  0  0.204  0.204  0.204  0.204  0.204
4 1.71   3.80   1     0     1     0     1  1.106 -1.27   0.591  0.706  4.17  0  0.504  0.504  0.504  0.504  0.504
5 0.620  2.29   1     0     1     1     1  0.961 -0.521   0.673  0.583  2.54  0  0.272  0.272  0.272  0.272  0.272
6 -0.586 -0.435  1     1     0     1     0  0.906  0.611  0.158  0.0859 -0.728  0  -2.02   -2.02   -2.02   -2.02   -2.02
7 0.515  0.729  1     0     0     1     1  0.611  0.0783  0.533  0.868  1.69  0  0.575  0.575  0.575  0.575  0.575
8 0.897  3.48   1     0     0     1     1  1.116 -1.01   0.767  0.483  3.41  0  0.290  0.290  0.290  0.290  0.290
9 0.659  2.36   1     0     1     1     1  1.112 -0.595   0.430  0.300  2.14  0  -0.834 -0.834 -0.834 -0.834 -0.834
10 0.191  1.42   1     0     0     0     1  1.107 -0.183   0.338  0.195  1.12  0  -1.31   -1.31   -1.31   -1.31   -1.31
```

multiple imputation D=5, X1은 MI값들의 mean

MAR10				COMPLETE	
MEAN	BIAS	TOTAL VARIANCE	FMI	MEAN	VARIANCE
-1.645	-0.0006	1.1063	6.948e-05	-1.644	1.112

missing에서 잘 적용됨 ! 동일한 방식으로 data에 적용하고자 함.

04 Apply to Data

DATA SELECTION

MIDTERM

반응변수 (Y)	독립변수 (X1, ..., X33)				
	공복 혈당 (HE_glu)	성별 (sex)	나이 (age)	요크레아티닌 (HE_Ucrea)	... (HE_BMI)
94	2	56	84.6	...	26.507
84	1	30	54.3	...	27.152
87	2	25	192.4	...	21.308
...	

FINAL

반응변수 (Y)	독립변수 (X)				
	공복혈당 (HE_glu)	BMI (HE_BMI)	당화혈색소 (HE_HbA1c)	... (HE_anem)	... (HE_DMfh3)
94	26.507	5.6	...	0	...
84	27.152	5.3	...	0	...
84	21.308	5	...	0	...
...

반응변수 : 공복 혈당 (HE_glu)

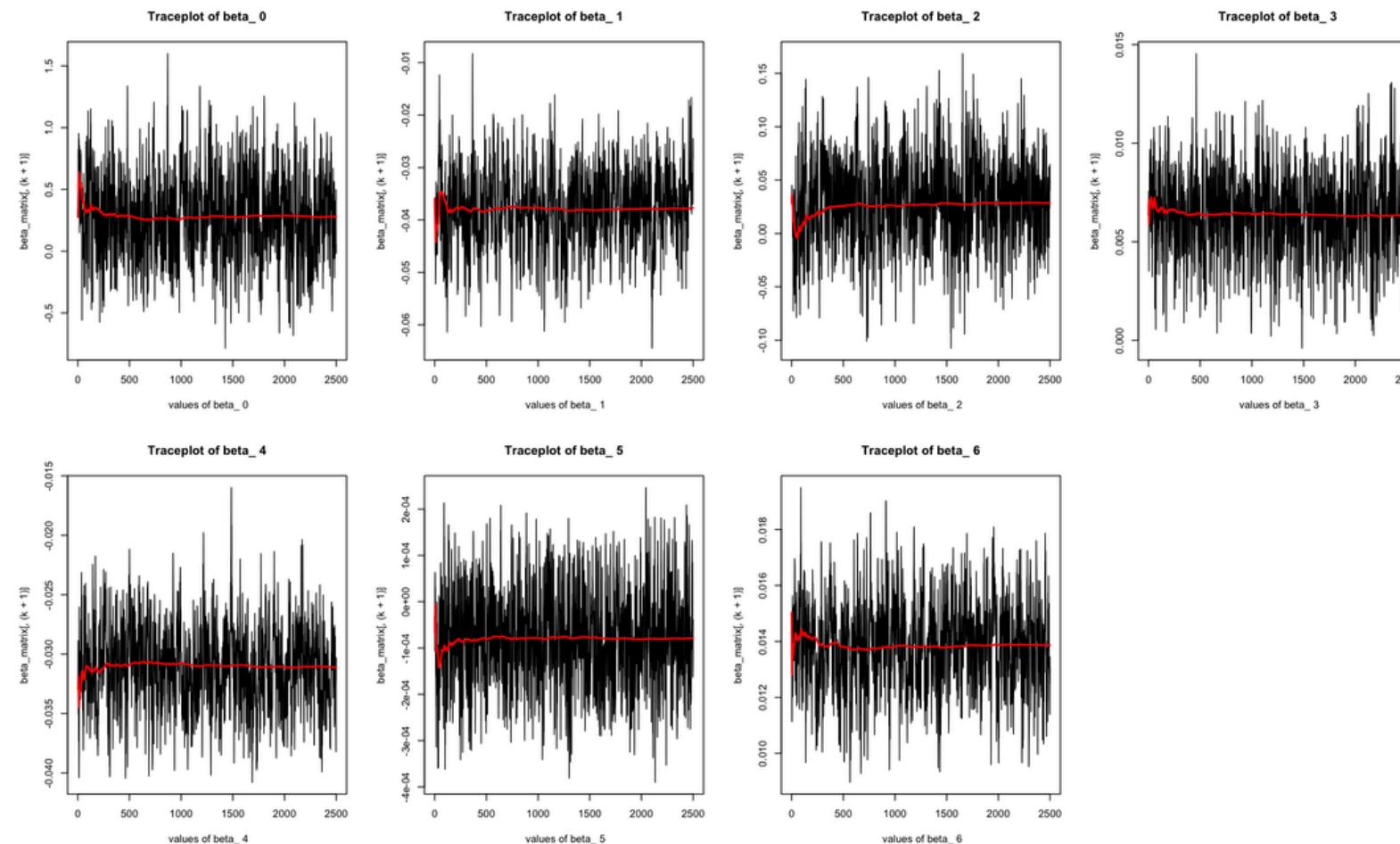
독립변수 : 33개의 변수

- 인적 사항(성별, 나이)
- 혈액 검사 결과 중 일부
- 신체 검사 결과 중 일부

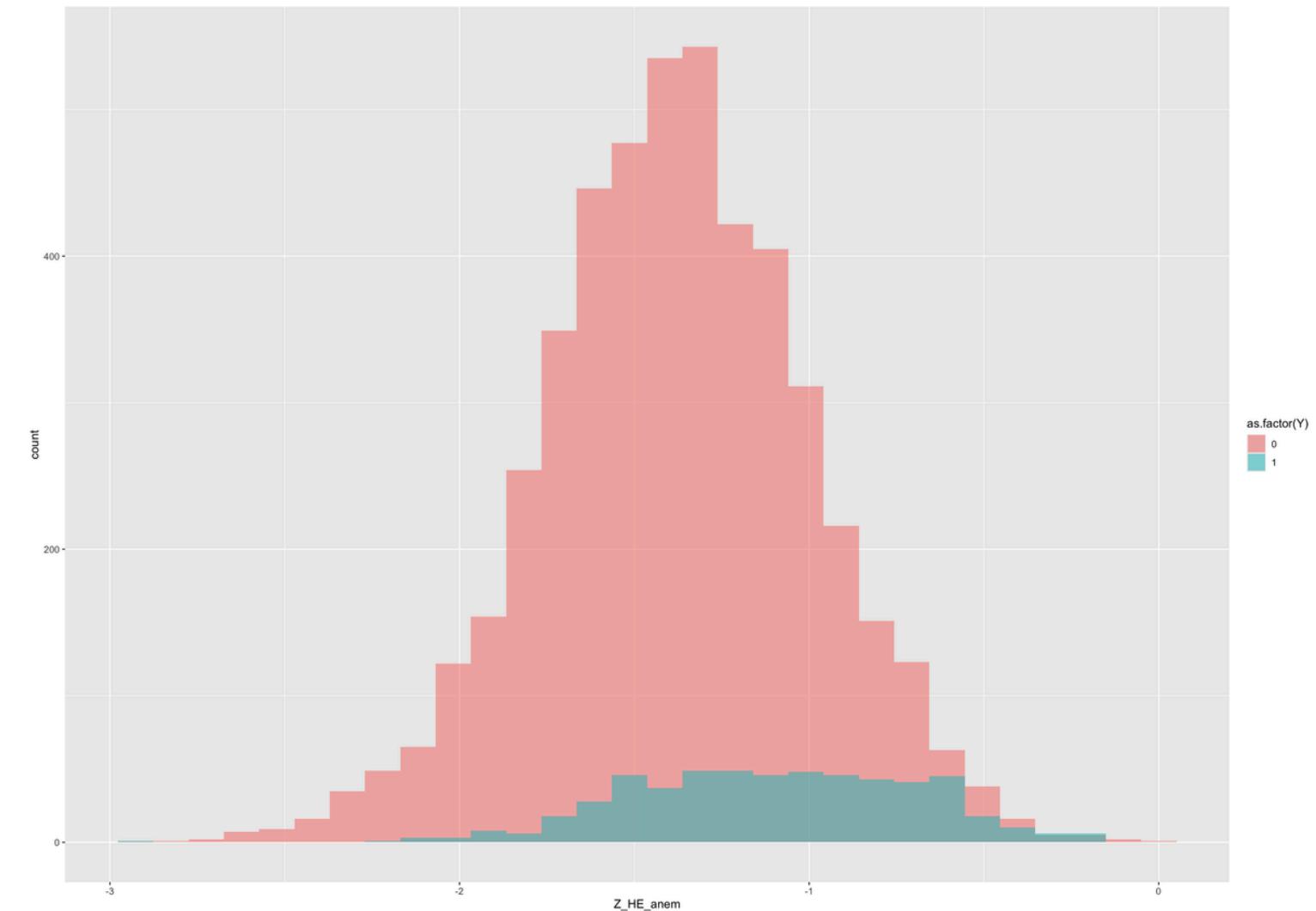
- 잠재변수 활용에 용이한 변수 선정
HE_anem(빈혈유병여부), HE_DMfh3(가족력)
- 잠재변수 추정 및 HE_glu의 missing 잘 설명하는 변수 추가
HE_BMI, HE_HbA1c, HE_sbp, HE_dbp, DE1_ag, age
(continuous)

GENERATING LATENT VARIABLES

HE_anem (MAR10)



beta 수렴이 잘 됨을 확인 가능

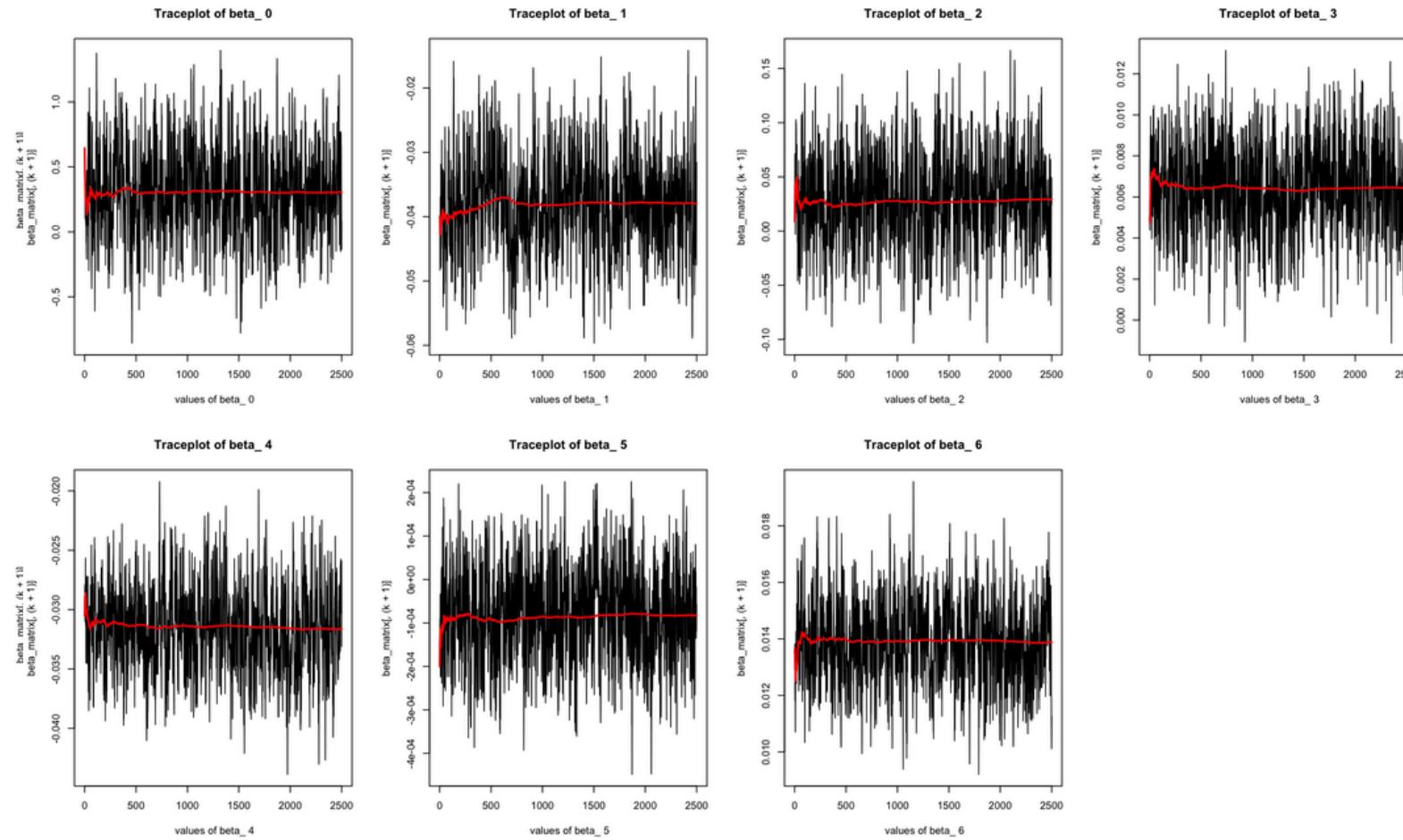


latent variable이 매우 잘 추정되었다고 보기는 어려움
그러나 Y별 Z의 중심점이 다름을 확인할 수 있음

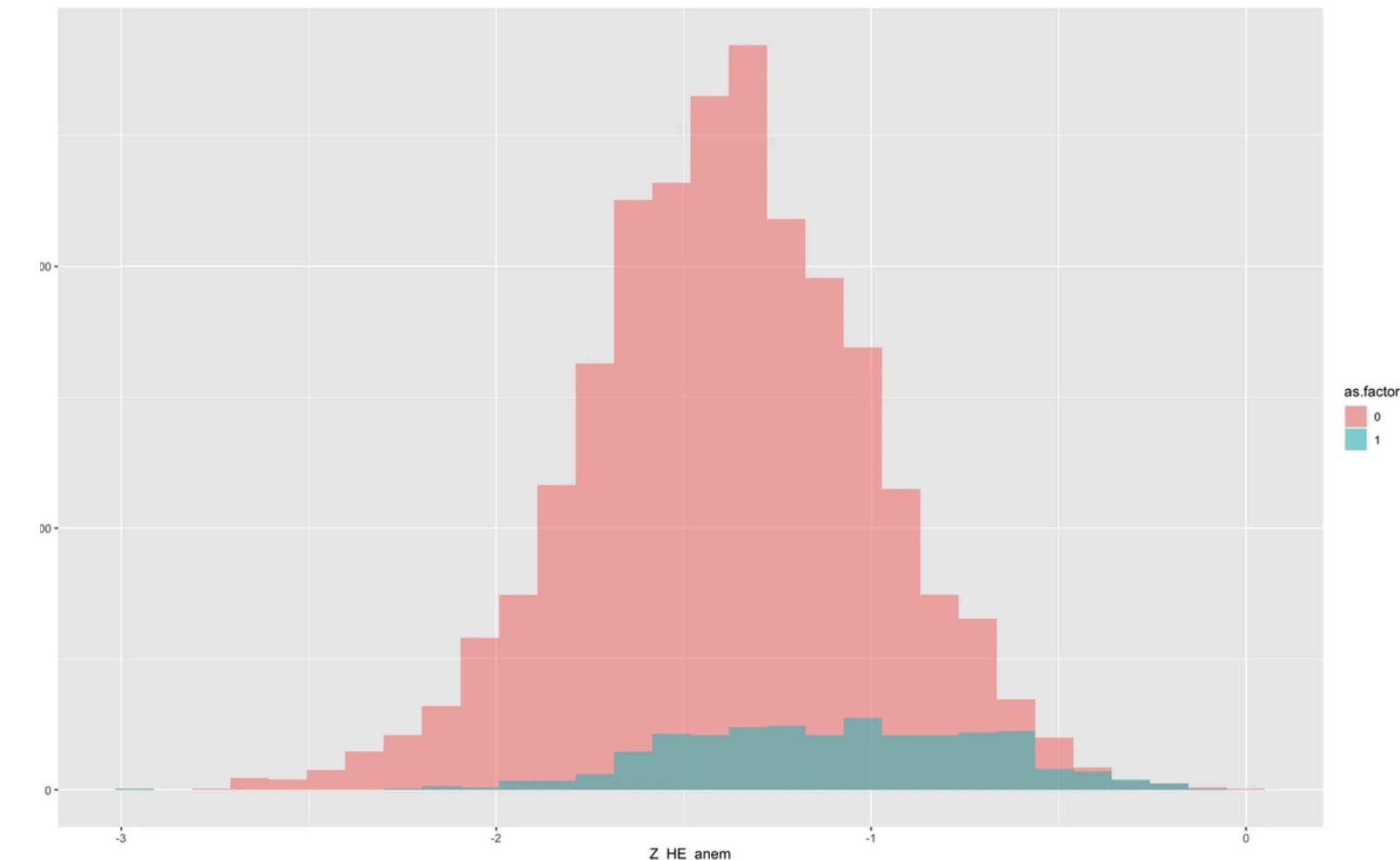
*해당 변수와 관련있는 변수가 전 데이터셋에서 많지 않았기 때문이라 판단

GENERATING LATENT VARIABLES

HE_anem (MAR40)



beta 수렴이 잘 됨을 확인 가능

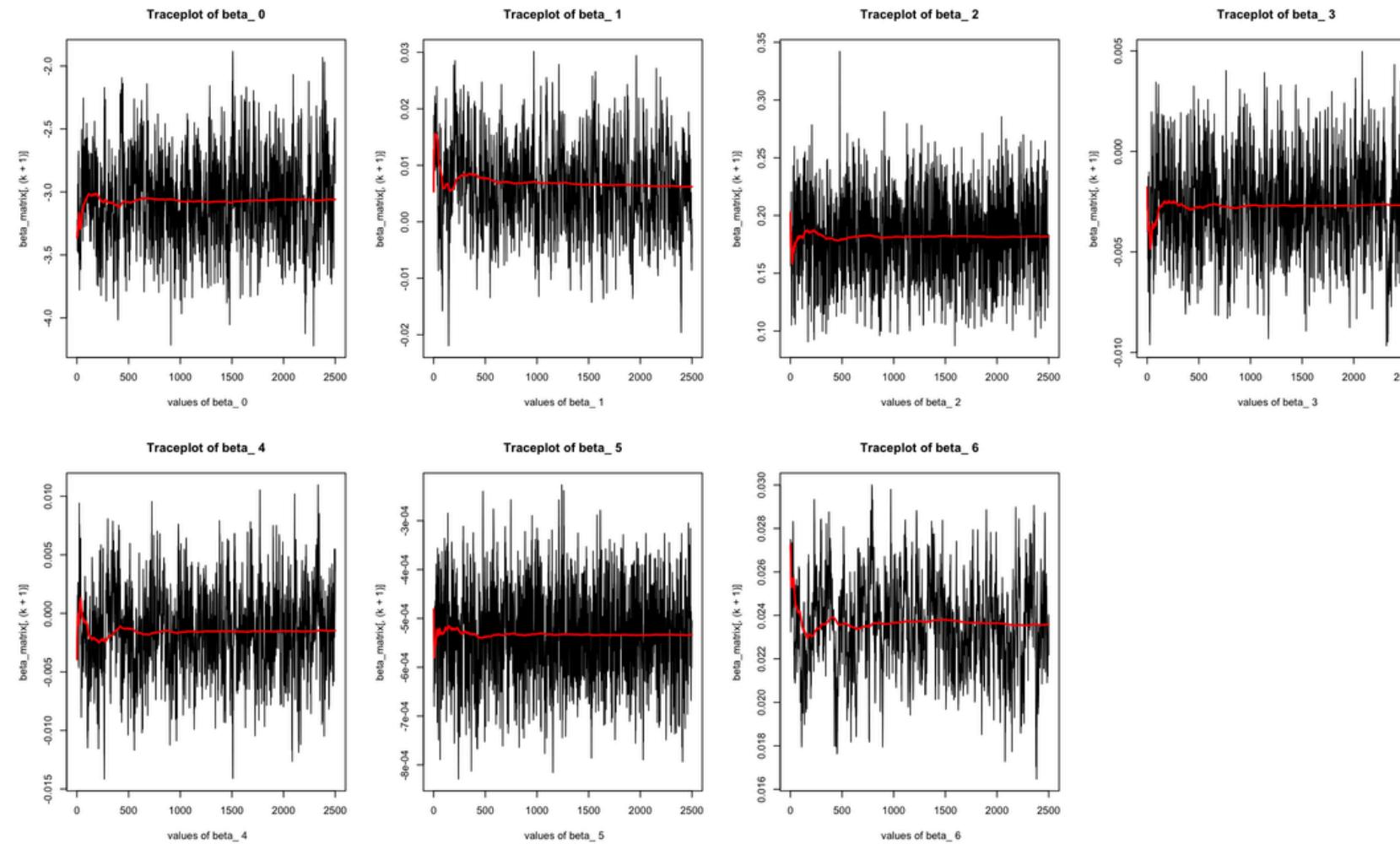


latent variable이 매우 잘 추정되었다고 보기는 어려움
그러나 Y별 Z의 중심점이 다름을 확인할 수 있음

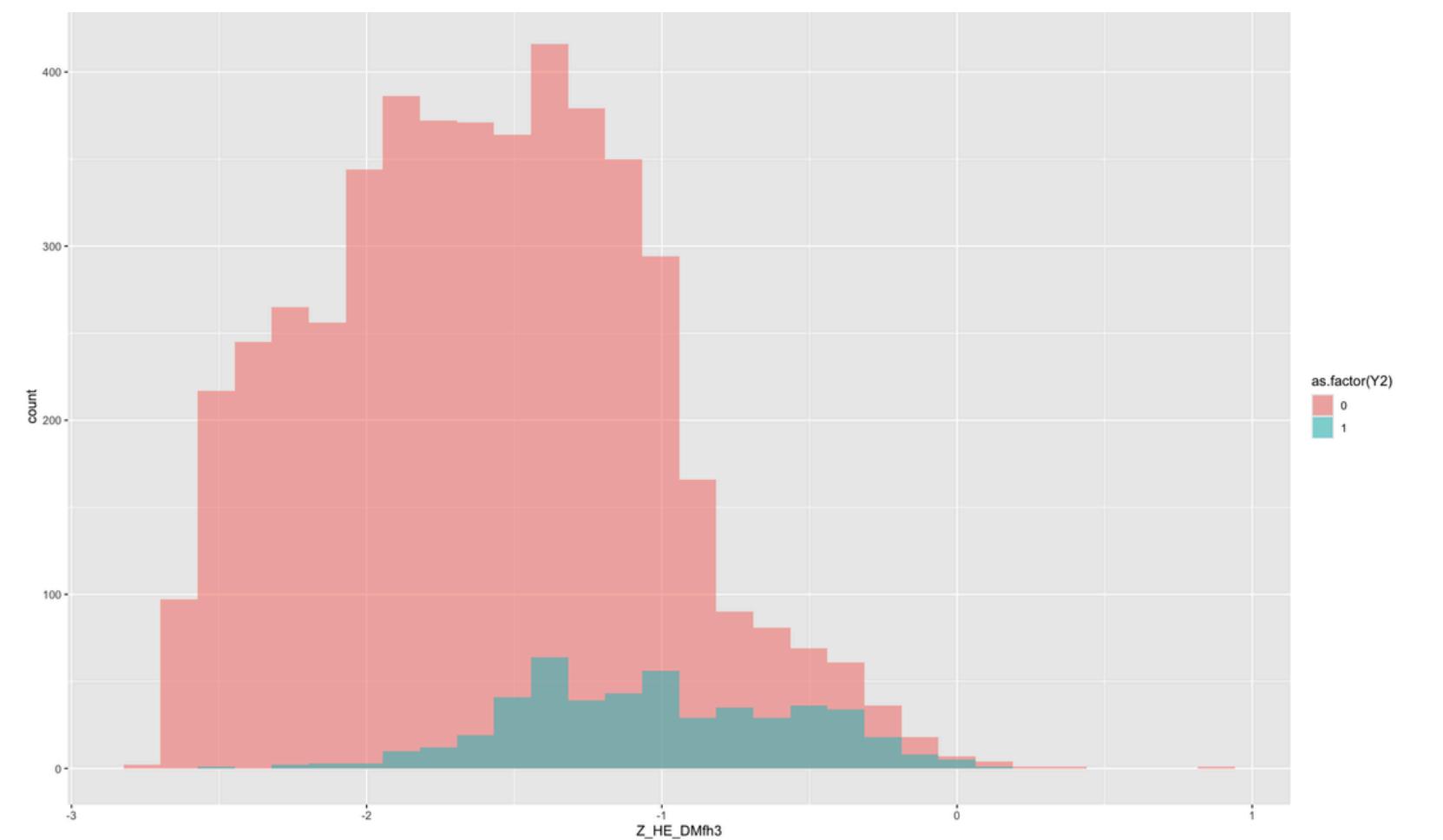
*해당 변수와 관련있는 변수가 전 데이터셋에서 많지 않았기 때문이라 판단

GENERATING LATENT VARIABLES

HE_DMfh3 (MAR10)



beta 수렴이 잘 됨을 확인 가능

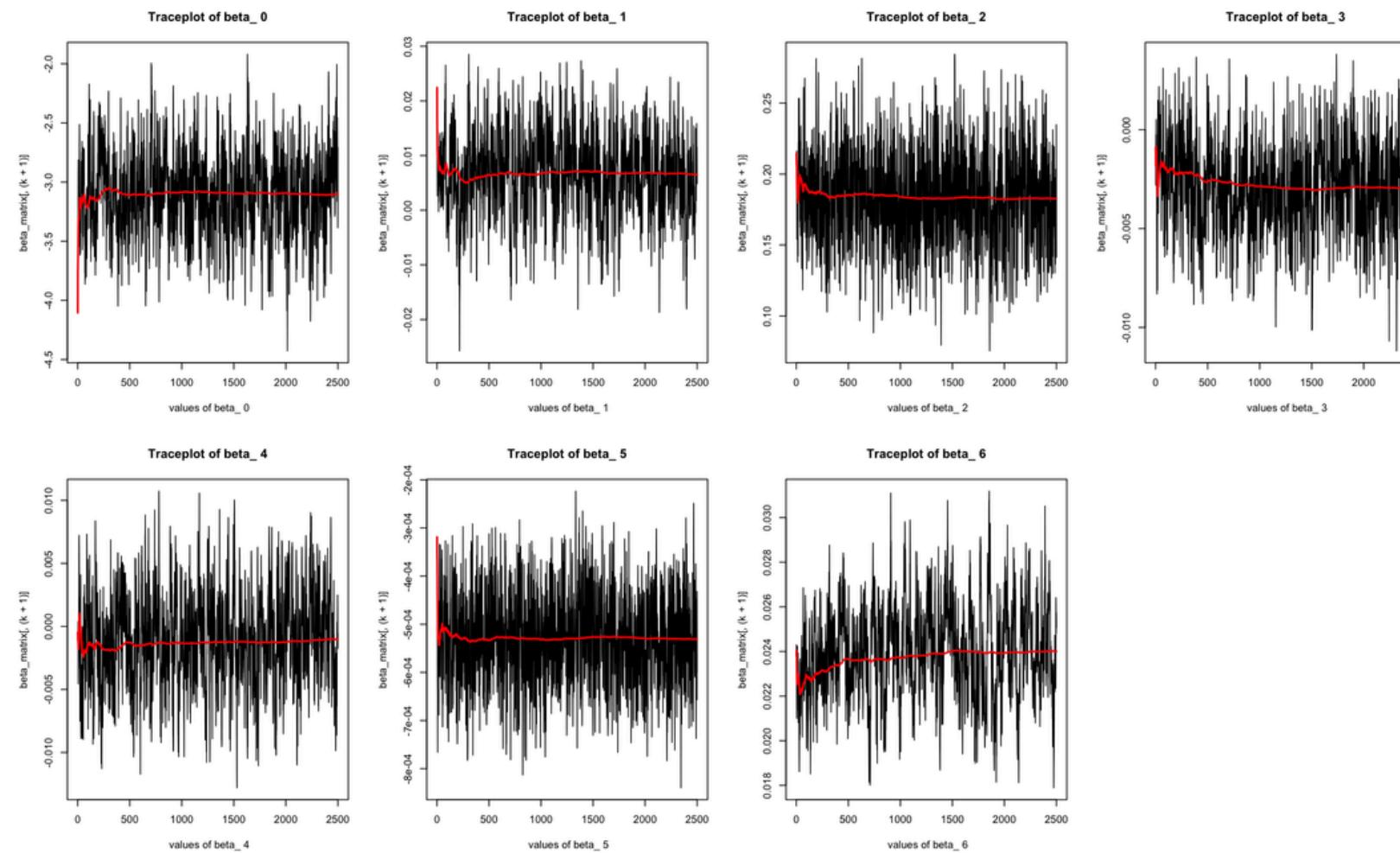


latent variable이 매우 잘 추정되었다고 보기는 어려움
그러나 Y별 Z의 중심점이 다름을 확인할 수 있음

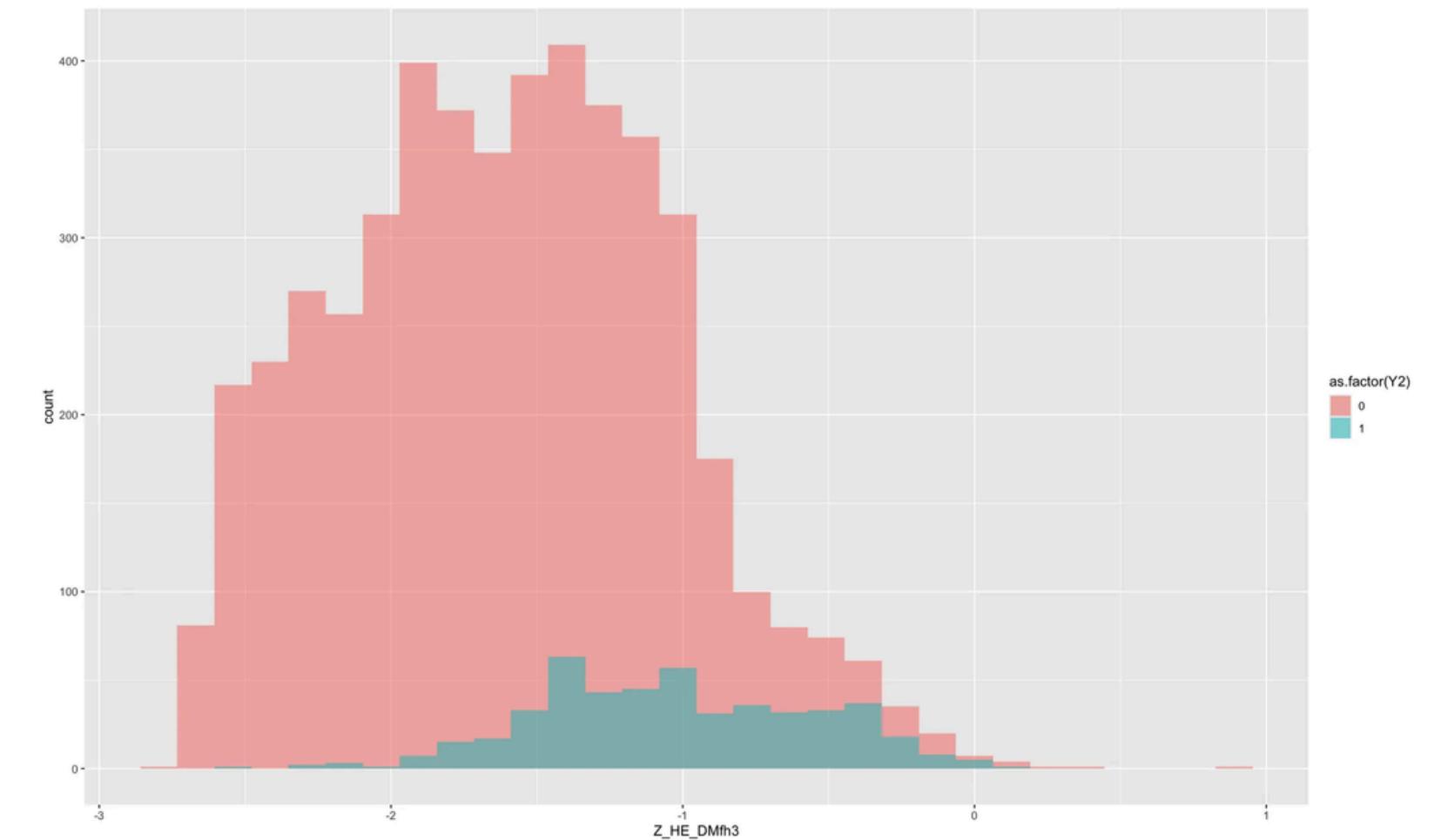
*해당 변수와 관련있는 변수가 전 데이터셋에서 많지 않았기 때문이라 판단

GENERATING LATENT VARIABLES

HE_DMfh3 (MAR40)



beta 수렴이 잘 됨을 확인 가능



latent variable이 매우 잘 추정되었다고 보기는 어려움
그러나 Y별 Z의 중심점이 다름을 확인할 수 있음

*해당 변수와 관련있는 변수가 전 데이터셋에서 많지 않았기 때문이라 판단

MATCHING

*MatchIt library의 matchit function을 활용

PS model with covariates

- 연속형 변수 + 이항형 변수 활용
- distance = “glm”
- replace=TRUE
 - randomness로 인해 유사하지 않음
에도 matching될 경우를 제외하기
위함

1.1:1 matching (method1)

2.1:10 matching (method2)

a. ratio = 10

PS model with latent variables

- 연속형 변수 + 잠재변수 활용
- distance = “glm”
- replace=TRUE
 - randomness로 인해 유사하지 않음
에도 matching될 경우를 제외하기
위함

1.1:1 matching (method3)

2.1:10 matching (method4)

a. ratio = 10

PS model with latent variables using mahalanobis

- 연속형 변수 + 잠재변수 + 성향점수 활용
- distance = “mahalanobis”
- replace=TRUE
 - randomness로 인해 유사하지 않음
에도 matching될 경우를 제외하기
위함

1.1:1 matching (method5)

2.1:10 matching (method6)

a. ratio = 10

총 6가지 방법에 대해 비교할 것

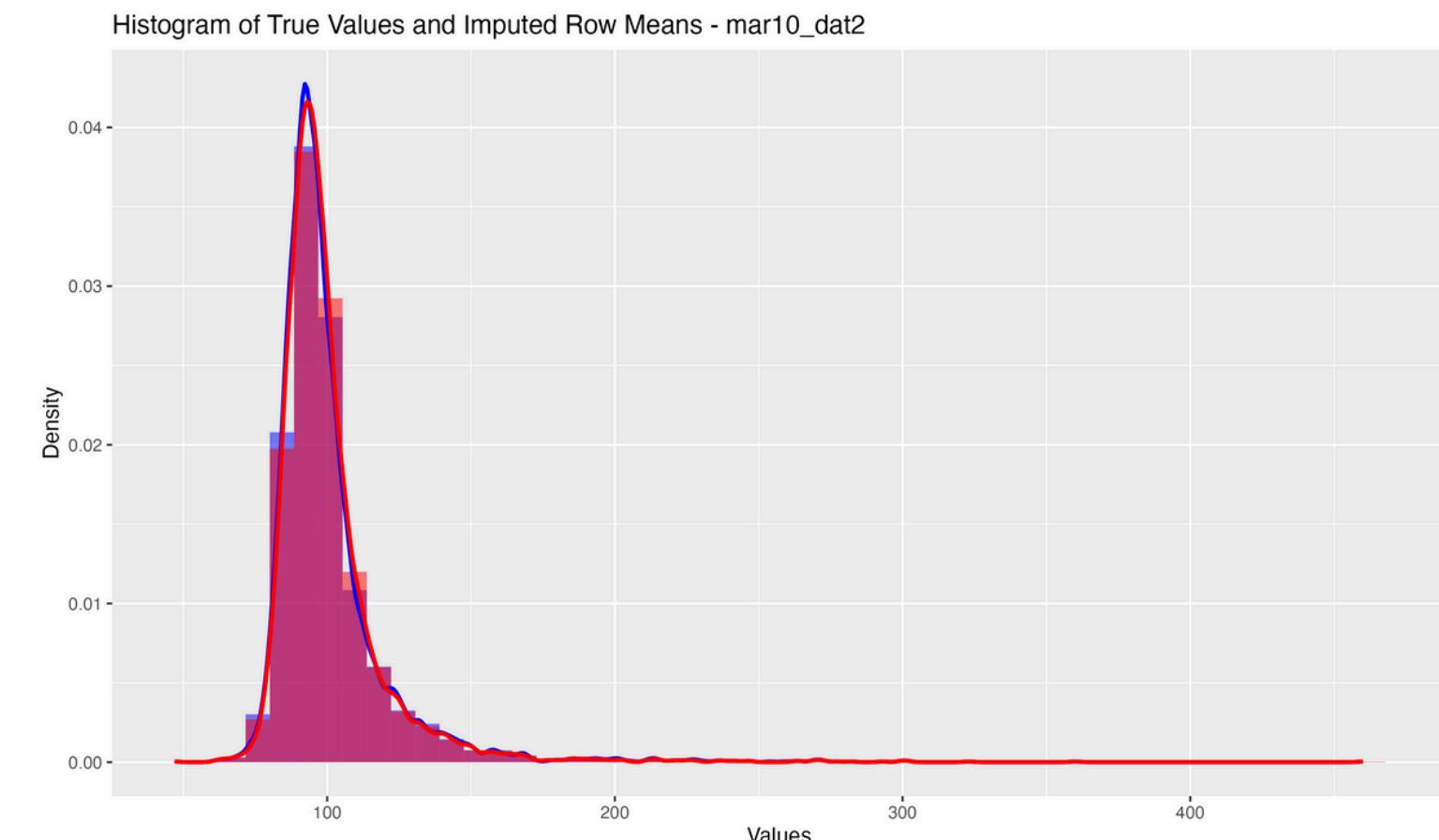
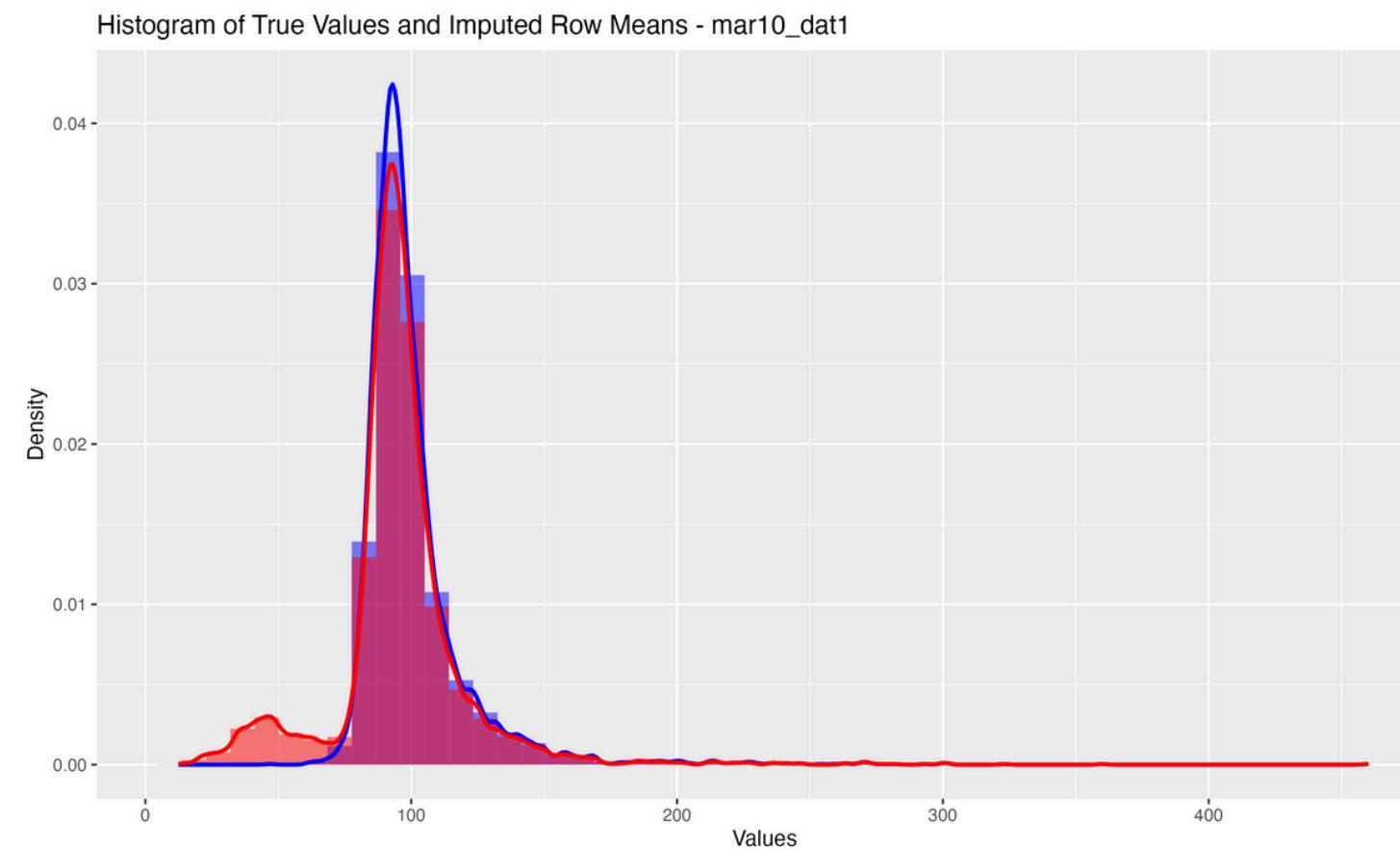
RESULT MAR10

DATA	MAR10				COMPLETE	
	TYPE	MEAN	BIAS	TOTAL VARIANCE	FMI	MEAN
method1	96.2369	-4.551431	745.6803	2.30013E-05	100.7883 476.8308	100.7883 476.8308
method2	100.7957	0.00735923	483.0284	4.67E-05		
method3	96.31218	-4.476148	774.0583	3.3435E-05		
method4	100.829	0.04066159	486.1553	1.50135E-05		
method5	95.94567	-4.842663	731.4412	5.097E-05		
method6	100.6508	-0.1375209	478.8945	5.30363E-06		

RESULT MAR10

*blue : true / red : imputed

one-to-one matching



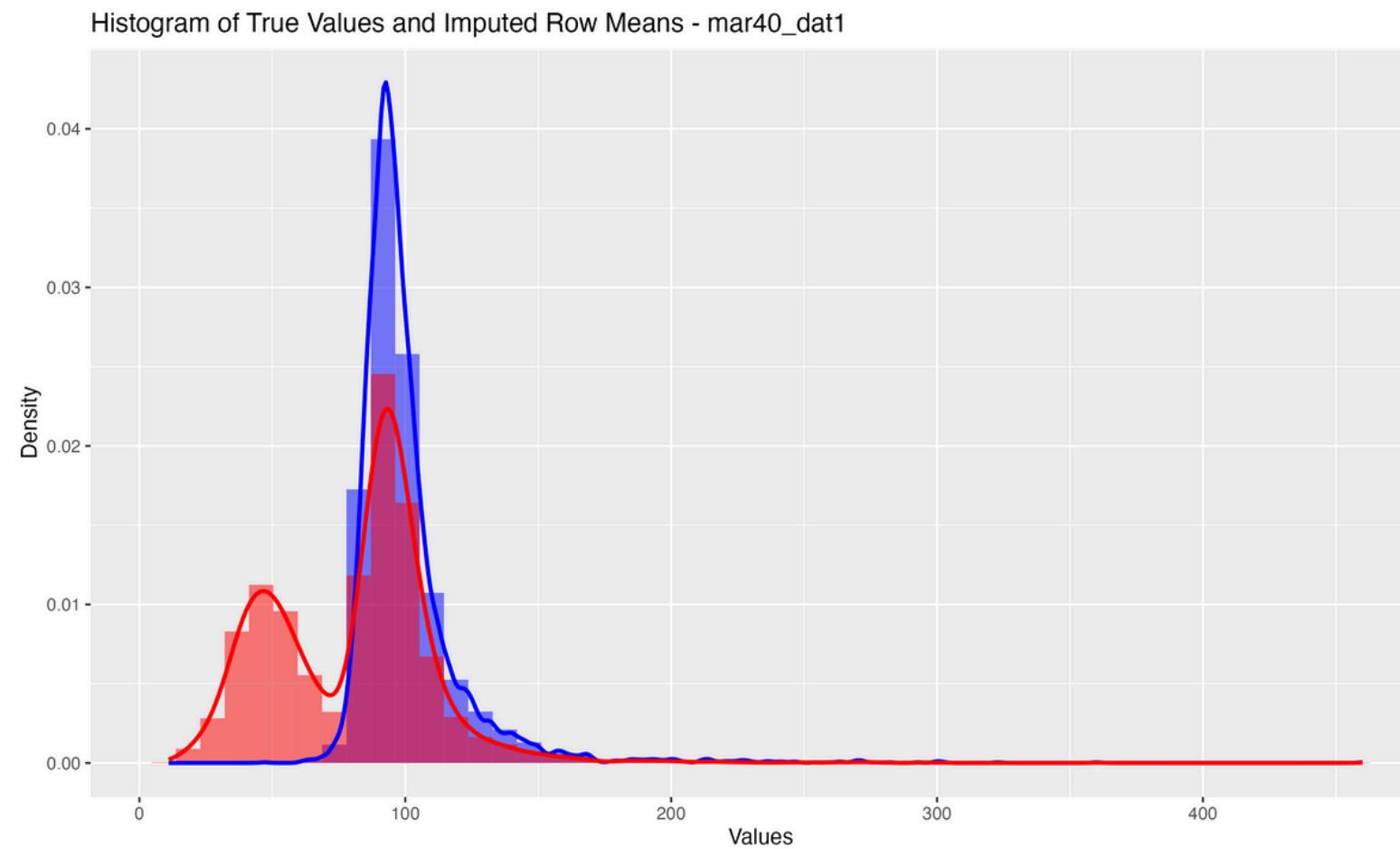
전체 method에서 one-to-many matching이 더 적절하게 impute되고 있음을 확인할 수 있음

RESULT MAR40

DATA	MAR40				COMPLETE	
	TYPE	MEAN	BIAS	TOTAL VARIANCE	FMI	MEAN
method1	86.87487	-13.91346	1196.638	2.7203E-05	100.7883	476.8308
method2	101.0351	0.2467571	534.077	0.00018278		
method3	86.65451	-14.13382	1145.982	1.98961E-05		
method4	101.0065	0.2182122	529.383	6.07163E-05		
method5	86.82581	-13.96252	1174.976	9.16038E-05		
method6	100.7075	-0.0808028	477.3921	6.24714E-06		

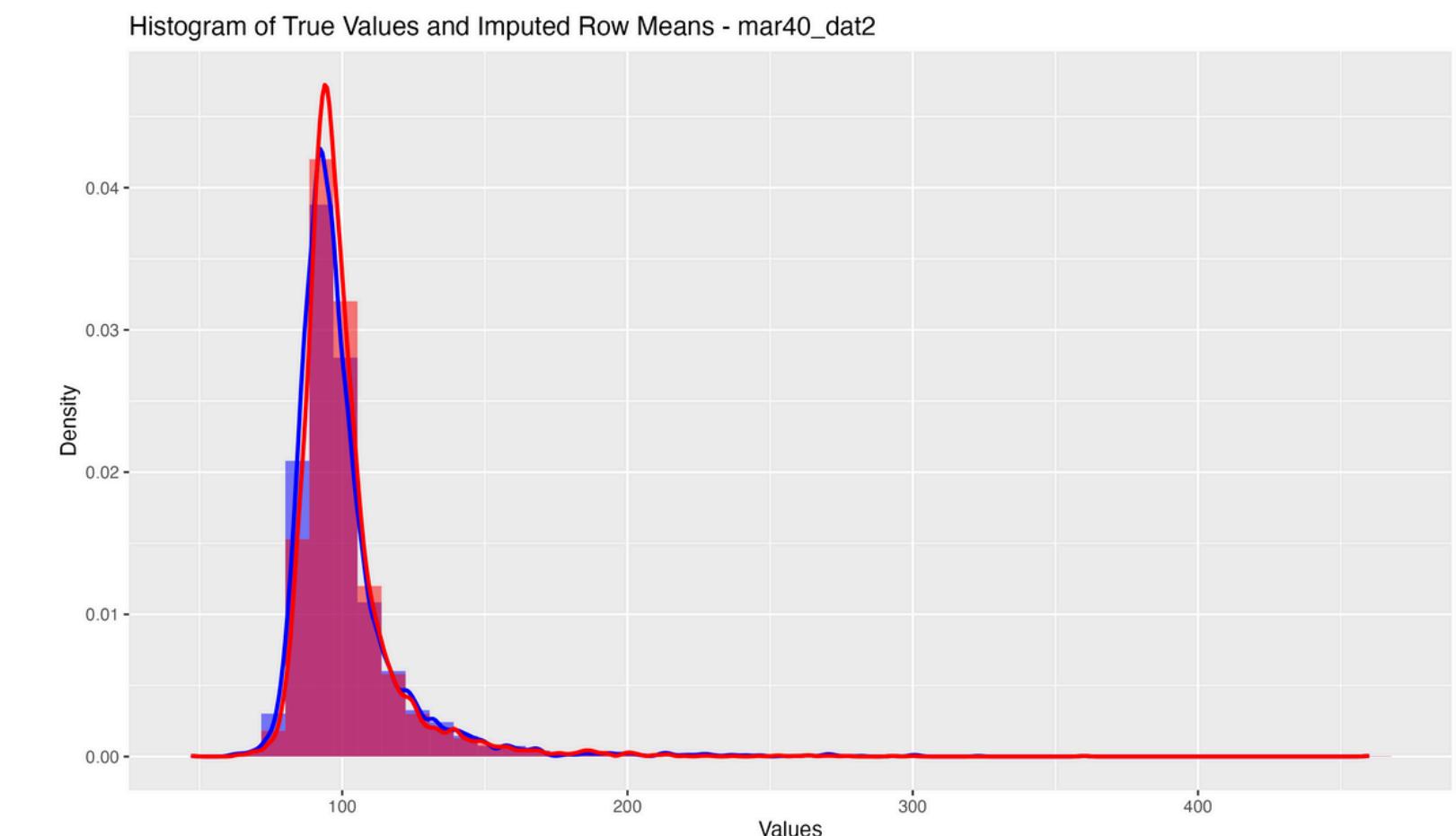
RESULT MAR40

one-to-one matching



*blue : true / red : imputed

one-to-many matching



전체 method에서 one-to-many matching이 더 적절하게 impute되고 있음을 확인할 수 있음

CONCLUSION

- **One-to-one vs One-to-many 매칭 비교:**
 - One-to-many 매칭은 실제 평균에 더 근접한 추정치를 제공하여 Bias가 작고, 분산을 잘 추정
 - 추정량의 분산을 제공하여 불확실성 반영 측면에서도 One-to-many 매칭의 장점이 더 많음.
- **잠재변수의 추정이 뛰어나지 않더라도, 잠재변수를 활용해 Propensity, Mahalanobis를 계산하는 것이 더 좋은 성능을 보임**
- **종합적으로, One-to-many 매칭과 Propensity, Mahalanobis 활용한 대체(Imputation)가 보다 신뢰성 있고 안정적인 결과를 제공함**

연구 의의

이산형 공변량의 처리

- 연속형과 이항형 형태의 혼재된 상황에서 전통적 매칭 방법이 갖는 한계를 극복
- 이항형 자료를 잠재변수로 모델링하여 연속형 변수처럼 처리할 수 있음
- 잠재변수를 도입함으로써, 공변량간 관계 구조를 풍부하게 반영할 수 있음.
- 단순 이항형 혹은 범주형 변수를 그대로 사용 할 때보다 matching 성능 향상 가능

imputation class 형성

- matching으로 적절한 class 형성을 통한 hot deck imputation 성능 향상을 도모
- latent variable을 활용한 matching 방법을 적용 하여, 이항형 변수가 많은, high dimension 상황에서 cell 간 sparseness 문제 해결 가능
- one-to-one matching이 아닌 one-to-many matching을 활용해 hot deck에서의 randomness 반영 가능

연구 한계

DATA 측면

- 불균형한 이항형 변수
 - 대부분의 이항변수가 극심한 불균형을 가짐
 - 한 범주에 치우쳐 분포하는 경우 잠재 변수 추정 안정성이 저하됨.
 - 이로 인해 잠재변수를 통한 성향점수 추정 및 매칭 품질이 떨어질 수 있음.
- 다항형 변수에 대한 확장
 - modeling code issue로 다항형으로 확장 못함 (result가 좋지 않았음, appendix 참고)
 - 다양하고 복잡한 공변량 구조를 온전히 반영하는 데 있어 한계

METHOD 측면

- matching 방법을 적용하기에 앞서 random ordering을 여러번 적용해야함
 - computational cost와 시간적 한계로 해당 과정을 구현하지 못함
 - 그렇기에, one-to-one matching에서의 성능이 더 좋지 않았을 수 있다고 판단
- 다른 imputation class 사용과의 비교
 - 다른 imputation class 형성 방법 후 hot deck imputation 과정과의 비교 부재
 - matching을 class로 활용했을 때, 다른 방법과 비교한 성능을 알 수 없기에 추가적인 연구 필요

설계 연구에서의 성향점수

- 국민건강영양조사데이터는 설계된 자료이므로 비관측연구의 성향점수 활용 이유와 다를 수 있음.
- 표본추출 단계에서 통제가 이루어져 있어, 성향 점수를 통한 추가적인 균형화가 필수적이지 않을 수 있음.
- 그럼에도 불구하고 일부 특성별 동질성 확보나 결측 치환 후 그룹 비교 시 성향점수 개념을 참고할 수 있음.

Reference

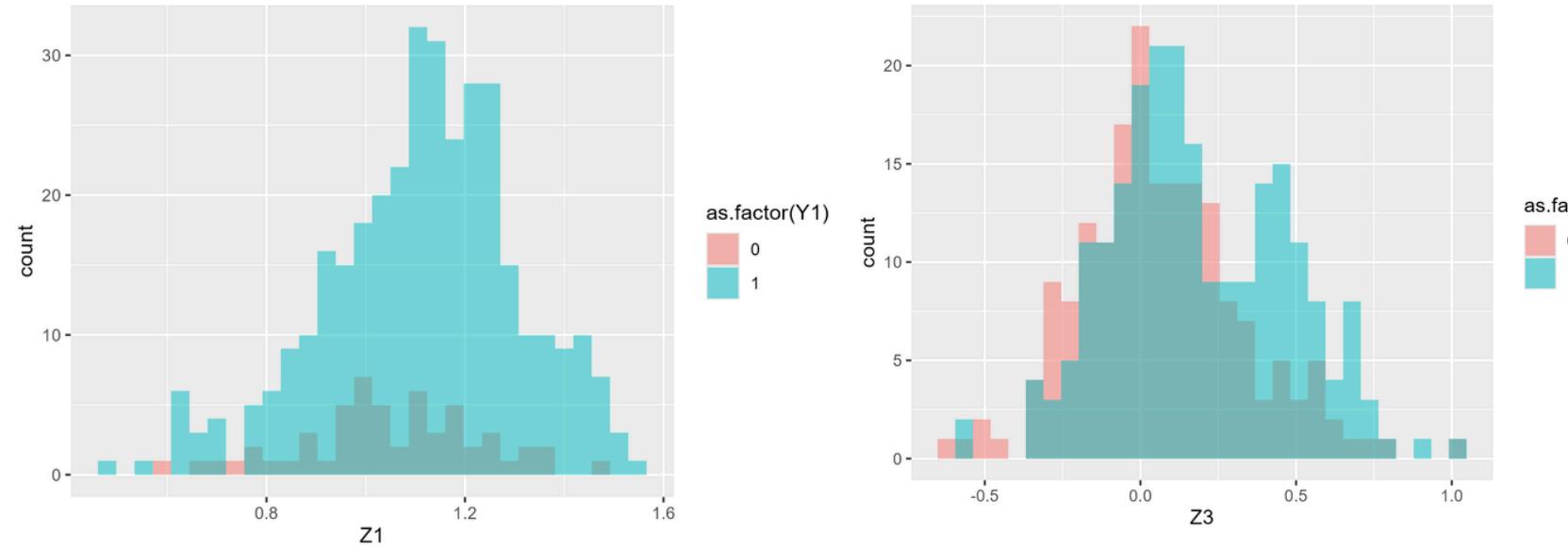
윤형석. (2012). 잠재변수를 이용한 혼합형 공변량을 포함한 자료에 대한 짹짓기 방법 (석사학위논문). 고려대학교 대학원, 서울.

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422), 669–679.

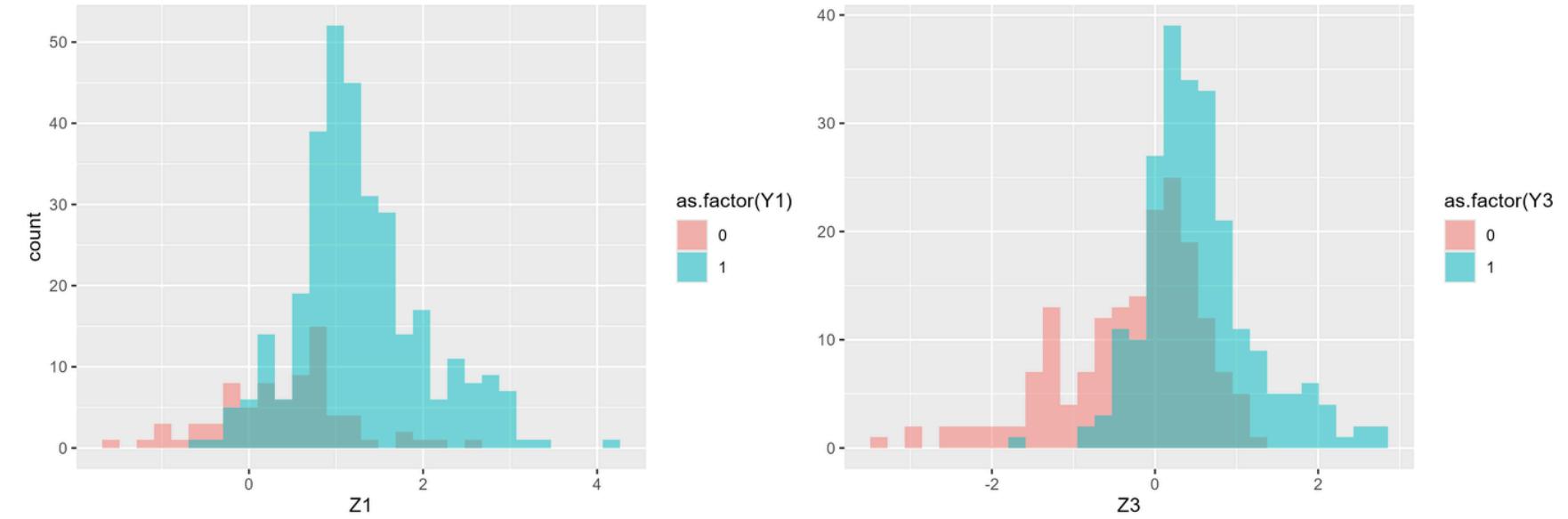
Little, R. J. A., & Rubin, D. B. (2020). *Statistical analysis with missing data* (3rd ed.). John Wiley & Sons.

APPENDIX

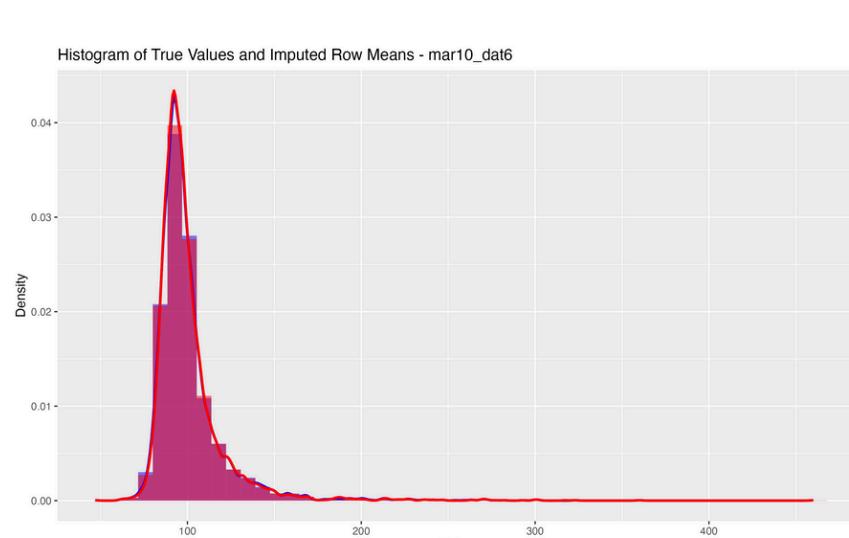
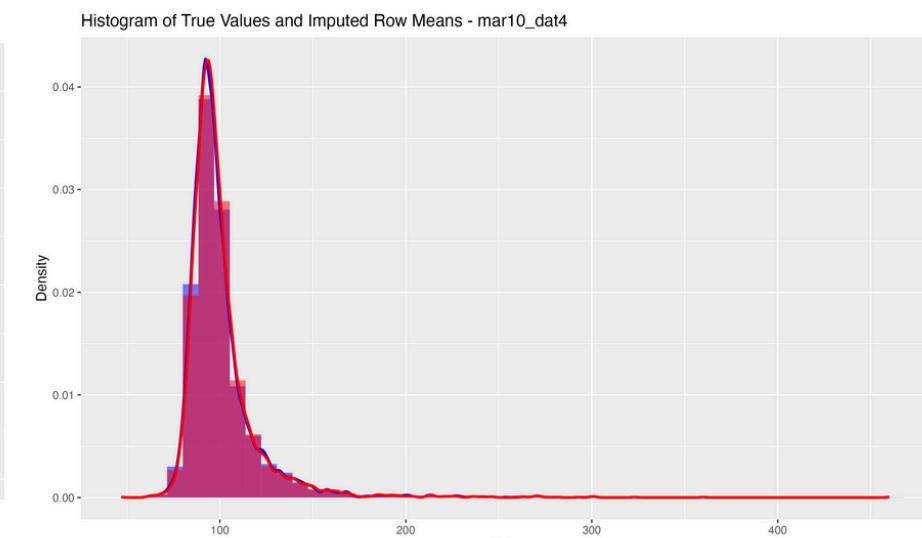
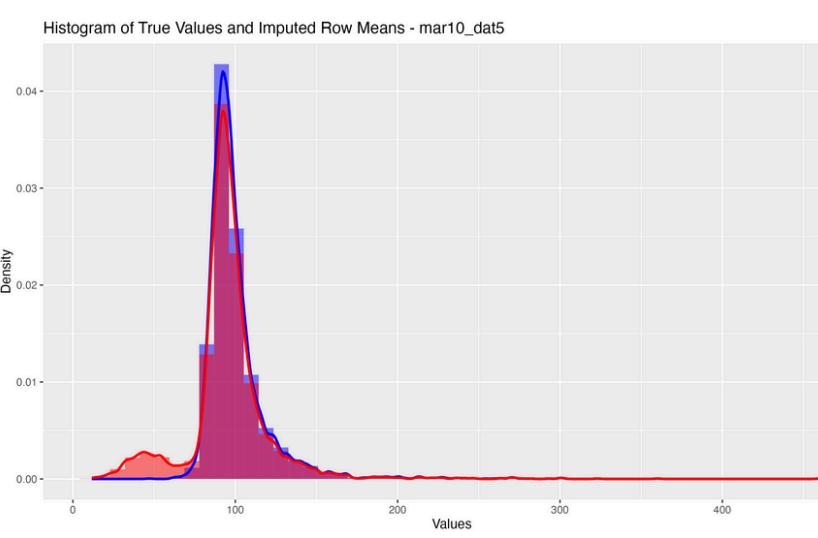
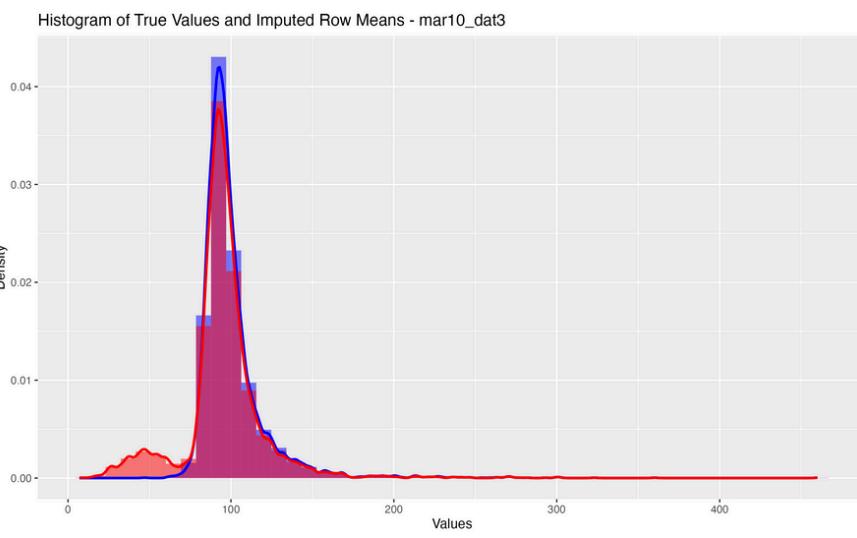
paper simulation beta convergence results = situation1



paper simulation beta convergence results = situation2



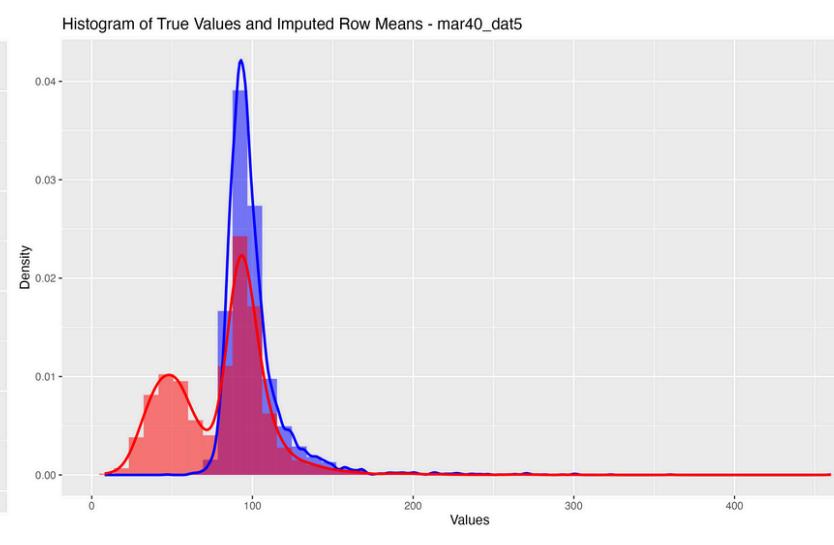
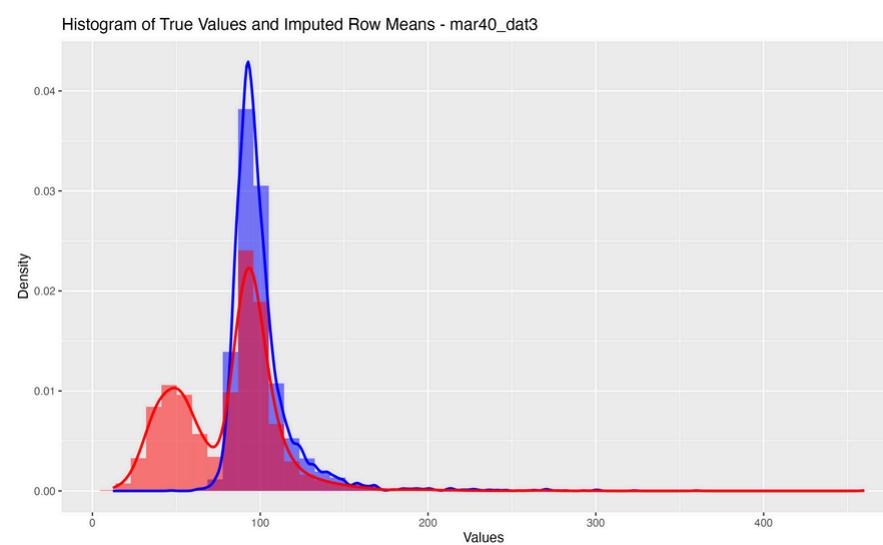
data application mar10 one-to-one matching



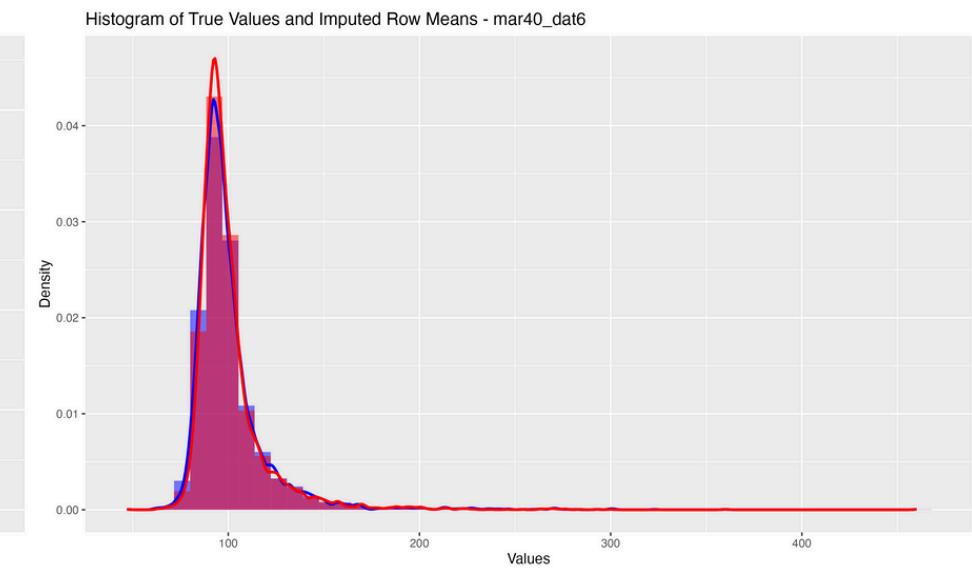
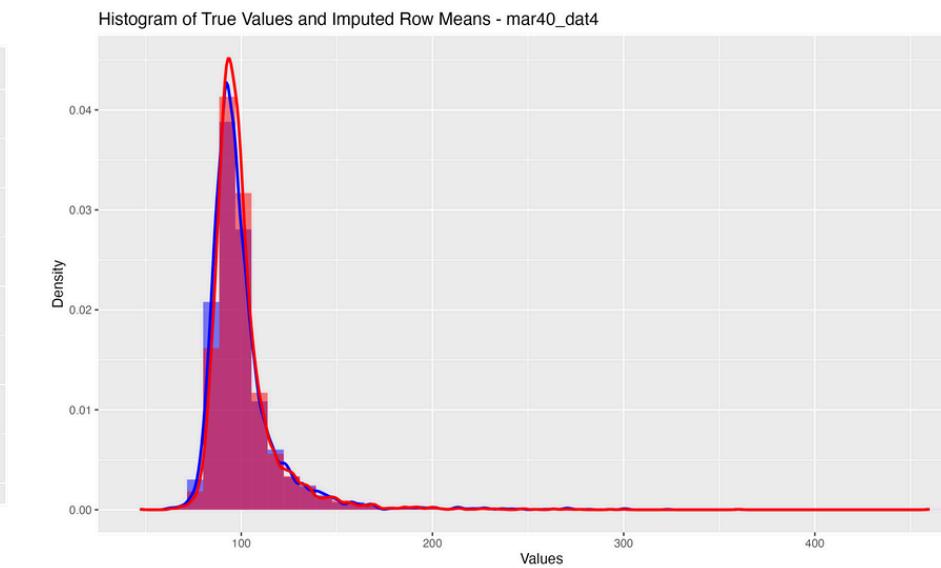
data application mar10 one-to-many matching

APPENDIX

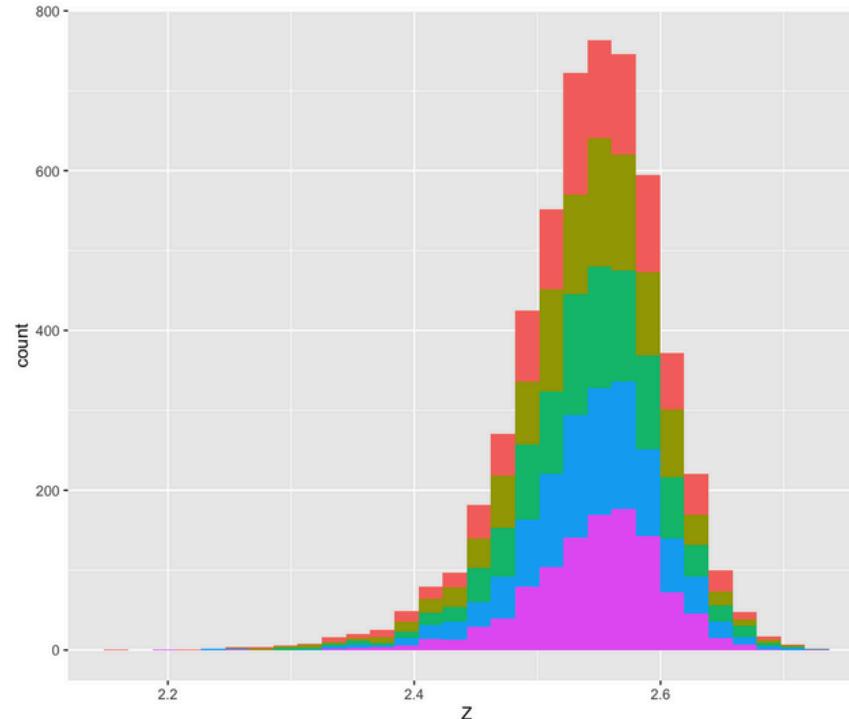
data application mar40 one-to-one matching

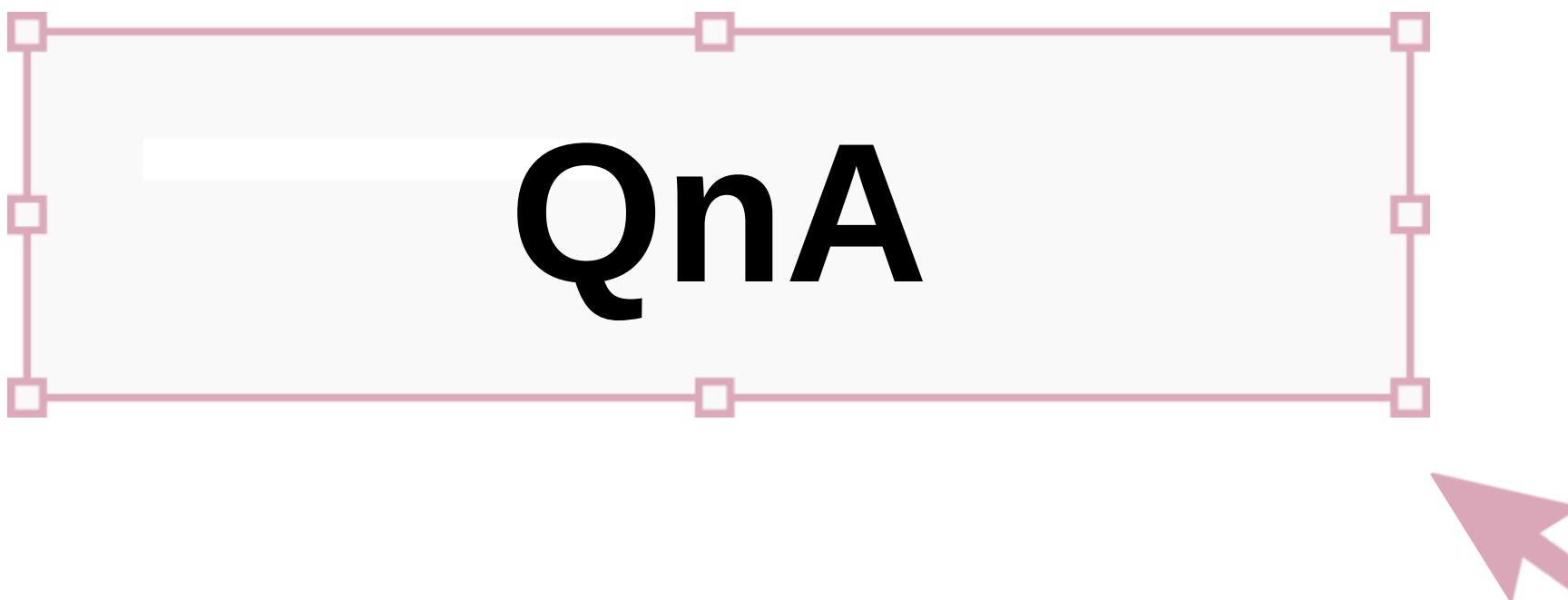


data application mar40 one-to-many matching



simulation multinomial version latent variables





QnA