An Imputation approach to correct measurement error based on changing survey mode

강원대학교 박승환*, 연세대학교 임종호

2021년 춘계 통계학회 조사통계연구회 기획세션

목 차

I. 서론

II. 제안 방법론

III. 가계금융복지 조사 적용

IV. 결론 및 향후 과제

I. 서론

- 코로나로 인하여 20년 이후 조사 환경이 급변.
- 대면으로 진행되었던 조사들이 비대면으로 전환 되고 있는 추세.
- 조사 모드: 조사원, 우편, 전화, 인터넷 등
  - 면접 조사 – 조사원(face to face), 전화
  - 자기 기입 조사 – 우편, 인터넷

I. 서론

- 혼합조사(mixed-mode survey): 하나의 조사에서 여러 가지 조사 모드를 사용하는 조사 기법
  - 장점 – 응답자에게 선택권을 줌으로써 참여도를 높이고 응답자의 편의를 존중함.
  - 단점 – 조사 모드의 차이를 고려하지 않은 단순한 집계 방식은 서로 다른 조사 모드로 인해 생기는 측정 오차로 인한 편향을 가져올 위험이 있음.
- 혼합조사의 정확성과 신뢰성을 제고하기 위한 조사 모드의 차이를 보정해주는 통계적 추정 방법이 필요함.

I. 서론 – 가계금융 복지 조사

- 전통적으로 면접조사로 수행됨.
- 20년 조사에서 대구.경북 지역은 비대면조사 원칙
  - 비대면조사: 자기기입, 인터넷, 전화
- 대구.경북 외의 지역에서는 자기기입식과 면접조사 동시 가능
- 과거 조사의 비대면 조사 비율이 4%정도 였던 것에 비하여 20년 조사에서는 약 20%정도 임.
- 비대면조사를 실시 함에 따른 조사 모드 간 응답특성의 차이 발생 예상.

Ⅱ. 제안 방법론

- 잠재 변수 (y): 측정 오차가 없는 관심변수 참값
- 보조 변수 (x): 성,연령,지역 등의 인구사회학 변수 및 과거 조사자료
- 관측 변수:
  - y_a: 대면조사 로 인해 얻어지는 관심변수 값
  - y_b: 비대면조사 로 인해 얻어지는 관심변수 값
- 혼합조사의 자료구조:

조사모드	X	Y_a	Y_b
대면조사(Sample A)	O	O	O or X
비대면조사(Sample B)	O	X	O

Ⅱ. 제안 방법론

- 관심 모수 : $\psi_N = N^{-1}\sum_{i=1}^{N} y_i$
- 대면 조사: $\hat{\psi}_{HT} = N^{-1}\sum_{i\in S_a} w_i\, y_i$
- 혼합 조사: $\hat{\psi}_{Naive} = N^{-1}\{\sum_{i\in S_a} w_i\, y_i + \sum_{i\in S_b} w_i\, y_{bi}\}$
- $\hat{\psi}_{Naive}$는 비편향 일까? Yes! 단, $E(y_{ai}) = E(y_{bi})$
- 편향 보정: $\hat{\psi}_{Cali} = N^{-1}\{\sum_{i\in S_a} w_i\, y_i + \sum_{i\in S_b} w_i\, E(y_i|y_{bi}, x_i)\}$
- 목표: 비대면 조사로 얻어진 자료를 사후 모형을 통해 대면조사로 얻어진 자료로 변환.

Ⅱ. 제안 방법론

- 예측모형(Imputation model) = 구조오차 모형 + 측정오차 모형

$$f(y \mid y_b, x) = \frac{f(y|x)g(y_b|y)}{\int f(y|x)g(y_b|y)\,dy}$$

- 구조오차 모형: $f(y \mid x)$
- 측정오차 모형: $g(y_b \mid y)$
- 선택모형 : P($M=a \mid x, y$), 조사모드의 선택이 랜덤이 아닐 경우 선택모형 필요
- 모수추정 : Mote Carlo EM algorithm + parametric fractional imputation

Ⅱ. 제안 방법론

- 사후 모형의 모수 추정치가 얻어지면 그 사후 분포가 얻어지는 것임
- 비대면 조사에서 관측된 보조 정보 x와 y_b값을 바탕으로 대면조사에서의 관측값 y_a를 사후 분포로부터 발생시키는 것이 목표임.
- 사후 분포가 정규 분포인 경우에는 쉽게 발생시킬 수 있지만 그렇지 않은 일반적인 경우에는 Markov Chain Monte Carlo 기법을 사용하게 됨. 또는 Parametric fractional imputation 기법을 적용할 수도 있음.

Ⅱ. 제안 방법론

Suppose that θ, α and ϕ are the parameter of distributions $f(y_{ai}|\mathbf{x}_i;\theta)$, $g(y_{bi}|y_{ai};\alpha)$ and $P(m_i = a|\mathbf{x}_i, y_{ai};\phi)$, respectively. Then, the EM algorithm using the PFI method under nonignorable choice mechanism is computed by the following steps:

- [Step 1] Set $t = 0$. Calculate the estimate of the parameter θ of $f(y_{ai}|\mathbf{x}_i;\theta)$ with data S_a. Let the estimate, denoted as $\hat{\theta}^{(0)}$, be the initial value.
- [Step 2] For each unit $i \in S_b$, generate M imputed values, $y_{ai}^{*(1)}, \ldots, y_{ai}^{*(M)}$, from $f(y_{ai}|\mathbf{x}_i;\hat{\theta}^{(0)})$. Set $w_{ij(0)}^{*} = 1/M$.

Ⅱ. 제안 방법론

- [Step 3] Update $\hat{\theta}, \hat{\alpha}$ and $\hat{\phi}$ by solving the imputed score equations:

$$\sum_{i \in S_a} w_i S_1(\theta; \mathbf{x}_i, y_{ai}) + \sum_{i \in S_b} \sum_{j=1}^{M} w_i w^*_{ij(t)} S_1(\theta; \mathbf{x}_i, y^{*(j)}_{ai}) = 0$$

$$\sum_{i \in S_b} \sum_{j=1}^{M} w_i w^*_{ij(t)} S_2(\alpha; y^{*(j)}_{ai}, y_{bi}) = 0$$

$$\sum_{i \in S_a} w_i S_3(\phi; m_i, \mathbf{x}_i, y_{ai}) + \sum_{i \in S_b} \sum_{j=1}^{M} w_i w^*_{ij(t)} S_3(\phi; m_i, \mathbf{x}_i, y^{*(j)}_{ai}) = 0,$$

where $S_1(\theta; \mathbf{x}_i, y_{ai}) = \partial \log f(y_{ai} \mid \mathbf{x}_i; \theta)/\partial\theta$, $S_2(\alpha; y_{ai}, y_{bi}) = \partial \log g(y_{bi} \mid y_{ai}; \alpha)/\partial\alpha$ and $S_3(\phi; \mathbf{x}_i, y_{ai}) = \partial \left\{ \log I(m_i = a) \log \left(P_i/(1-P_i) \right) + \log(1-P_i) \right\} /\partial\phi$ with $P_i = P(m_i = a \mid \mathbf{x}_i, y_{ai}; \phi)$

Ⅱ. 제안 방법론

- [Step 4] Calculate weight w_{ij}^* for each $i \in S_b$,

$$w_{ij(t)}^* \propto g(y_{bi}|y_{ai}^{*(j)}; \hat{\alpha}^{(t)}) \frac{f(y_{ai}^{*(j)} \mid x_i; \hat{\theta}^{(t)})}{f(y_{ai}^{*(j)} \mid x_i; \hat{\theta}^{(0)})} P(m_i = b \mid \mathbf{x}_i, y_{ai}^{*(j)}; \hat{\phi}^{(t)})$$

  and $\sum_{j=1}^{M} w_{ij(t)}^* = 1$, where $\hat{\eta}^{(t)} = (\hat{\theta}^{(t)}, \hat{\alpha}^{(t)}, \hat{\phi}^{(t)})$ is the current estimate of $\eta = (\theta, \alpha, \phi)$.

- [Step 5] Set $t = t + 1$ and go to Step 3. Continue until convergence.

- The parametric fractional imputation estimator of the finite population mean is computed by

$$\hat{\psi}_{PFI} = N^{-1} \left\{ \sum_{i \in S_a} w_i y_{ai} + \sum_{i \in S_b} w_i \sum_{j=1}^{M} w_{ij}^* y_{ai}^{*(j)} \right\}.$$

Ⅲ. 가계금융복지조사 적용

- 19년 조사 자료의 비대면(자기기입) 비율은 약 3.7%.
- 20년 조사에서 비대면(자기기입, 인터넷, 전화) 비율은 약 22.6%.
- 19년 비대면 응답 가구의 20년 비대면 응답 비율은 19년 대면 응답 가구의 20년 비대면 응답 비율보다 약 3배 이상 높게 나타남
- 과거 조사모드에 대한 선택 성향이 20년 조사모드 선택에 영향을 미침

19년 조사모드	20년 조사모드				비대면 비율
	면접	자기기입	인터넷	전화	
면접	79.1%	12.5%	0.7%	7.8%	20.9%
자기기입	34.3%	58.3%	1.1%	6.3%	65.7%
전체	77.4%	14.2%	0.7%	7.8%	22.6%

Ⅲ. 가계금융복지조사 적용

- 시도별 20년 조사모드의 분포 현황 분석

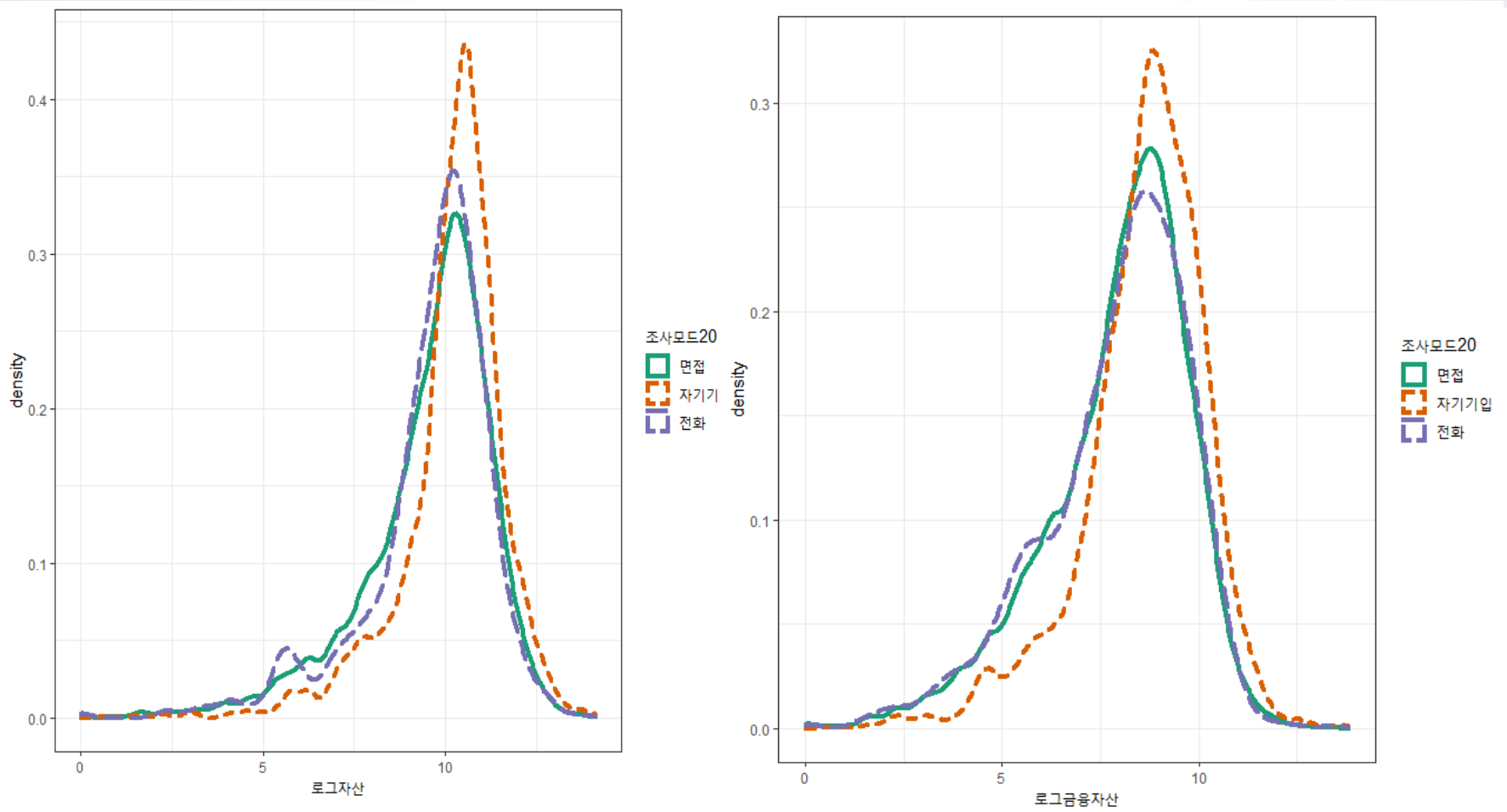
시도		면접	자기기입	인터넷	전화	비대면비율
19년 전체	서울	86.3%	11.8%	1.0%	0.8%	13.7%
	부산	91.9%	7.1%	0.2%	0.8%	8.1%
	대구	1.1%	68.0%	2.3%	28.6%	98.9%
	인천	89.8%	7.3%	0.4%	2.5%	10.2%
	광주	97.2%	2.8%	0.0%	0.0%	2.8%
	대전	79.7%	15.0%	0.6%	4.7%	20.3%
	울산	95.3%	2.7%	0.0%	2.0%	4.7%
	세종	75.5%	24.2%	0.0%	0.3%	24.5%
	경기	72.9%	16.8%	1.1%	9.2%	27.1%
	강원	86.3%	11.9%	0.0%	1.8%	13.7%
	충북	79.9%	6.3%	0.2%	13.6%	20.1%
	충남	80.4%	8.8%	1.3%	9.5%	19.6%
	전북	92.2%	3.1%	0.0%	4.7%	7.8%
	전남	90.1%	3.3%	0.5%	6.2%	9.9%
	경북	33.1%	23.1%	1.1%	42.7%	66.9%
	경남	89.5%	9.6%	0.2%	0.6%	10.5%
	제주	92.5%	5.1%	0.5%	1.9%	7.5%
	전체	77.4%	14.2%	0.7%	7.8%	22.6%

Ⅲ. 가계금융복지조사 적용

- 4인 가구 이상, 가구주 대학졸업이상, 가구주 연령 50세 미만, 가구 소득 5천만원 이상 가구에서 비대면 응답 비율이 높게 나타남.
- 자기기입 응답 비율은 가구원수/교육정도/가구소득 증가함에 따라 증가함.
- 전화 응답 비율은 1인가구에서 높게 나타남. 교육정도, 가구소득과 무관함.

Ⅲ. 가계금융복지조사 적용

- 자산, 금융자산등 소득 변수는 20년 비대면 조사 가구에서 20년 대면 조사 가구보다 중위수, 평균, 가중평균 등에서 전반적으로 크게 나타남.



Ⅲ. 가계금융복지조사 적용 - 시뮬레이션

- 가상 데이터 생성 – 지역*20년 조사모드를 사용 19년 자료 그룹화
- h그룹에서 $y_{h,20} = y_{h,19} \cdot r_h$ 를 통하여 20년 가상 변수 생성
- 18년, 19년 조사 자료 이용하여 증감률 r_h 추정 (중위수)
- 20년에 모두 대면으로 응답하였을 경우에 대한 가상 변수 생성

그룹	지역	20년 조사모드	자기기입 추정 증감률	전화 추정 증감률
1	경북 외 지역	대면	1.016	1.006
2	경북 외 지역	비대면	1.050	1.082
3	대구, 경북	대면	1.001	1.016
4	대구, 경북	비대면	1.256	1.041

Ⅲ. 가계금융복지조사 적용 - 시뮬레이션

- 예측모형: x_1 – 19년 응답값, x_2 – 19년 조사모드

$$f(y_{대면}|y_{비대면},x,모드=비대면)\propto f(y_{대면}|x)g(y_{비대면}|y_{대면})\Pr(모드=비대면|y_{대면},x),$$

구조오차 모형 – $f(y_{대면}|x)$: $y_{대면}=\beta_0+x_1\beta_1+x_2\beta_2+e_1,\quad e_1\sim N(0,\sigma_1^2)$

측정오차 모형 – $g(y_{비대면}|y_{대면})$: $y_{비대면}=\gamma_0+y_{대면}\gamma_1+e_2,\quad e_2\sim N(0,\sigma_2^2)$

선택 모형 – $\Pr(모드=비대면|y_{대면},x)$: $\log\frac{p}{1-p}=\psi_0+x_1\psi_1+\psi_2 y_{대면}$

- 오차 보정 관심 변수 평균 :

$$\widehat{Y^*}=\frac{1}{n}\left(\sum_{i\in S_{대면}} y_{i,대면}+\sum_{i\in s_{비대면}}\sum_{j=1}^{M} w_{ij}^{*}y_{ij}^{*}\right),$$

Ⅲ. 가계금융복지조사 적용 - 시뮬레이션

- 자산: 모수 추정결과 – 자기기입

구조오차 모형			측정오차 모형		선택 모형		
β_0	β_1	β_2	γ_0	γ_1	ψ_0	ψ_1	ψ_2
0.015	−0.040	1.000	0.248	0.986	−3.913	2.234	0.210

- 자산: 모수 추정결과 – 전화

구조오차 모형			측정오차 모형		선택 모형		
β_0	β_1	β_2	γ_0	γ_1	ψ_0	ψ_1	ψ_2
0.015	−0.009	1.000	0.040	1.000	−2.296	0.611	−0.002

- 20년 자산 오차보정 결과

조사모드	평균			중앙값		
	비대면	가상_대면	보정	비대면	가상_대면	보정
비대면	52929.2	48899.8	48706.6	31621.3	29187.8	29001.2
자기기입	61202.0	55976.6	55630.0	37943.5	34305.7	34388.9
전화	37855.4	36005.4	36091.6	22013.8	20987.2	21129.0

Ⅲ. 가계금융복지조사 적용 - 시뮬레이션

- 선택모형의 모수추정 결과 19년에 비대면 조사를 선택할 사람 일수록, 20년 자산의 값이 클수록 비대면 조사를 선택할 확률이 높아짐
- 비대면 표본의 보정 평균을 보게 되면 조사모드로 인한 오차 보정은 비대면 자산 자료의 크기를 줄여주는 효과
- 비대면 표본의 가구 특성 분석에서 비대면 조사 가구들의 자산이 대면 조사보다 크게 나오는 점에 비추어 볼 때 타당한 보정 방향이라고 보임.

Ⅳ. 결론 및 향후 과제

- 측정 오차 모형(measurement error model)을 통한 조사모드간 차이보정(calibration) 수행.
- 베이즈 정리를 이용한 사후 모형식 계산
- 사후 모형식으로 부터 예측치 (imputation) 발생
- 비대면 조사를 참값으로 오차 보정을 수행할 수 있음.
- 분산 추정 방법론 개발
- 실제 조사에 적용하기 위해서는 보다 정교한 모형 개발이 필요.

감사합니다