제25차 가구조사 무응답 국제워크숍 참가 결과 보고

2014. 10.

통계청 STATISTICS KOREA

통계정책국
표본과

목 차

Ⅰ 출장 개요

◇ 참가 회의: 제25차 가구조사 무응답 국제워크숍
(International Household Survey Non-response workshop)
◇ 개최 지역: 아이슬란드(레이캬비크)
◇ 여행 기간: 2014년 9월 1일 ~ 9월 6일(4박 6일)
◇ 참가자: 표본과 이정현 사무관, 한혜은 주무관
* 참가 국가: 미국, 캐나다/ 스웨덴 등 유럽 10개 국가/ 한국 (약 30여명)
* 참가 자격: 논문 또는 포스터 발표

□ 참가 목적

◦ 가구조사에서 무응답을 줄이기 위한 전략, 무응답 발생 시 처리방안 등에 대한 각국의 최근 동향을 파악하고 국내 통계작성에 활용

□ 수행 내용

◦ 포스터 세션에서 **가계동향조사의 무응답 조정 방법 발표**
- (질의) 무응답 조정 방법의 선택 이유와 진행 상황 및 연구 방향
- (답변) 변수의 영향을 검토하기 위해 기존의 방법을 적용하였으며, 제안한 방법의 적용을 위해서는 몇 가지 한계가 있어 보류 중임

◦ 패널 토론 및 주제별 발표 세션에 참가하여 **무응답 관련 연구 현황 파악**
- 발표자들이 현재 검토 중인 무응답 자료 처리 방법과 주요 관심 사항을 공유하고 참석자들에게 자문을 구하는 방식으로 진행

□ 업무 활용 방향

◦ 가구 조사의 무응답 조정 방법 개선에 활용
◦ 조사과정자료(paradata)를 활용한 추정방법 검토
◦ 표본연구회를 통해 청내 업무 관련자에게 지식 공유

Ⅱ 회의 개요

1. 회의 기간 : 2014년 9월 2일 ~ 9월 4일

2. 주관 기관 : 아이슬란드 통계청(Hagstofa Íslands)

◦ 가구조사 무응답 국제 워크숍(International Household Survey Non-response workshop)은 1990년 스웨덴 스톡홀름에서 Robert Groves, Lars Lyberg, Bob Barnes의 주관 하에 처음 개최되었고, 그 후 매년 9월 미국, 캐나다, 영국, 독일 등 16개국에서 개최됨

3. 회의 내용

□ 개최 목적

◦ 참가국의 통계청 및 통계작성기관의 실무자들이 가구조사에서의 **무응답 자료 처리 방법에 대한 자문을 구하고 경험을 공유하는 실무자 워크숍**

□ 주요 세션

◦ 특별 세션: 무응답 조정 가중치가 얼마나 효과적인가

- 다목적 조사에서 소수의 보조정보를 이용하여 무응답층을 구성하여 조정하는 것이 적절한가에 대한 질문을 시작으로 대표적인 무응답 조정 방법에 대하여 소개하고 각각의 방법을 적용하고 비교할 수 있는 방법과 한계점에 대하여 논함(Westat의 J. Michael Brick의 기조 연설)
- 무응답 메커니즘에 따른 추정 결과 비교(노르웨이)
- 표본의 균형도과 보조변수의 설명력이 추정에 미치는 영향 연구(스웨덴)
- 극단 가중치 절단의 영향에 대한 검토(벨기에)

◦ 주요 세션

- 가구 조사에서 응답률 제고 방안: 답례품과 성과급 영향력 검토 등(독일)
- Call Record 활용: 접촉횟수별 특성 및 응답률 분석, 표본 관리 등(독일, 캐나다)
- 패널조사의 무응답: Wave 비교와 직전 응답자료 활용 등(영국, 독일, 노르웨이)

◦ 포스터 세션: 가계동향조사의 무응답 조정 방법 등

4. 참가자 현황 : 13개 국가, 30여명 참가

◦ 참가국 : 미국, 캐나다/ 스웨덴, 네덜란드, 영국, 독일, 헝가리, 노르웨이, 벨기에, 스위스, 아이슬란드, 폴란드/ 한국

◦ 기관 : 통계청, 통계작성기관, 대학, 연구소

- 미국(U.S. Census Bureau, U.S. Bureau of Labor Statistics, RTI International, Westat), 캐나다(Statistics Canada), 스웨덴(Statistics Sweden), 네덜란드(Statistics Netherlands, SCP, Utrecht University), 영국(UK Office for National Statistics, University of Southampton), 독일(IAB, GESIS, University of Munich, University of Mannheim, University of Maryland), 헝가리(Hungarian Central Statistical Office), 노르웨이(Statistics Norway), 벨기에(Centre for Sociological Research KU Leuven, Katholieke Universiteit Leuven), 스위스(University of Lausanne), 아이슬란드(Statistics Iceland), 폴란드(University of Lodz)

Ⅲ 향후 계획

1. 활용 방향

◦ 가구조사의 무응답 조정 방법 개선에 활용

- 현재 적용하고 있는 무응답 처리 기법을 점검하고 개선 사항 발굴

◦ 추정방법 개선에 활용

- 극단 가중치 처리 방법 검토
- 접촉 횟수 등의 조사과정자료를 무응답 조정이나 추정에 활용하는 방안 검토

◦ 「가구 표본 관리 지침」 보완 및 조사과정 자료 수집

- 불응·불능 가구의 처리 방안 및 표본 조사구 및 가구 관리 지침 보완
- 무응답 자료 처리를 위한 가구 명부 및 조사 과정 자료 항목 검토

◦ 답례품 및 성과급의 영향력 분석 등을 통한 응답률 제고 방안 검토

2. 향후 계획

◦ 매년 워크숍에 참가하여 한국의 무응답 조정 사례 발표('15년, 벨기에 루벤)

부록 주요 발표내용(요약)

1. An Examination of Current Nonresponse Adjustment Method in The Household Income and Expenditure Survey

(Hyeeun Han, Statistics Korea)

가. 개요

○ 현재 적용하고 있는 가계동향조사의 무응답 조정방법과 소득을 고려한 무응답 조정방법 검토

나. 무응답 조정 방법

○ 현재 모형

- 지역, 가구구분(가구주 직업), 가구원수, 거처유형으로 50개 무응답층 구성
  - 지역: 서울, 부산/광주/인천/대구, 대전/울산/경기동부, 도지역동부(경기제외), 도지역 군부
  - 가구구분: 사무직, 생산직, 자영자, 무직
  - 거처유형: 아파트, 아파트외
- 표본가구의 분포에 맞게 매월 무응답 보정
- 고소득층* 가구 비율: 6.81%

  * 고소득층: 가구 내 소득이 가장 많은 가구원의 **평균 소득이 400만원 이상**

○ 모형 1 : 가구구분 세분화 및 점유형태 고려하여 무응답층 구성

- 고소득 여부에 영향을 주는 가구구분, 가구원수, 거처유형, 점유형태, 지역을 고려하여 30개 무응답층 구성
  - 가구구분: 관리자 및 전문가, 사무직, 생산직, 자영자, 무직
  - 거처유형: 아파트, 아파트외
  - 점유형태: 자가, 전세, 월세, 무상
  - 지역: 서울/울산/경기, 그 외 지역
- 표본가구의 분포에 맞게 매월 무응답 보정
- 고소득층 가구 비율: 7.23%

○ 모형 2 : 전용면적 고려 & 인총 10% 표본 분포로 Calibration

- 모형 1의 변수에 전용면적을 추가로 고려하여 32개 무응답층 구성
- 2010 인총의 10% 표본가구의 무응답층별 분포에 맞게 매월 무응답 보정
- 고소득층 가구 비율: 10.1%

다. 검토 결과

- 월평균 가구소득은 현재 모형과 모형1, 2의 차이가 거의 없음
- 모형 2는 소득 400만원 초과 계층의 비율이 높게 나타남
  · 조정 전 13.8%에서 16.2%로 2.4%p 증가

<소득구간별 분포 비교>

(단위 : %)

		가계동향조사		
		현재 모형	모형 1	모형 2
월평균 소득		336만원	336만원	335만원
소득구간	<100	23.5	23.5	24.0
	100 ~ 200	30.3	30.3	28.2
	200 ~ 300	20.2	20.1	19.1
	300 ~ 400	12.2	12.3	12.5
	400 ~ 500	6.9	6.9	7.6
	500 ~ 650	4.4	4.4	5.2
	650<	2.5	2.5	3.4

라. 결과의 요약 및 향후 과제

- 가구 소득은 직업, 전용면적, 거처유형, 가구원수, 지역에 따라 차이를 보임
  · 가구주의 직업이 '관리자/전문가'인 경우와
  · 전용면적이 넓을수록(83㎡ 초과) 높아지는 것으로 나타났고,
  · 그 외 거처유형, 가구원수, 지역에 따라 소득에 차이가 나타남

- 무응답 조정 모형의 지속적이고 일관된 적용 한계
  · 2011년 가계동향조사 결과의 무응답 조정을 위해 '10년 인총 정보를 활용하는 것은 시의성에 문제가 없으나,
  · 인총 10% 표본 분포는 시간이 경과함에 따라 신뢰도가 낮아짐
  · 또한, 다가구주택의 경우 가구별 전용면적에 대한 추정 필요

- 무응답 조정을 위한 최신의 신뢰성 있는 정보 수집
  · 가구명부 작성 시 표본가구의 '가구주 직업'과 '전용면적'에 대한 정확한 정보 수집 필요

2. Estimating the proportion of smokers using three response models

(Magnar Lillegård and Ib Thomsen, Statistics Norway)

가. 개요

○ 무응답 메커니즘에 따른 세 가지 응답모형 하에서의 흡연률 추정 결과($\hat{p}_y$)를 비교하고 분석결과와 방향에 대하여 논의

나. 응답 모형

○ 유효 표본 크기($R=1$)

	Y		
X	1	0	Total
1	n_{11}	n_{10}	$n_{1.}$
0	n_{01}	n_{00}	$n_{0.}$
Total	$n_{.1}$	$n_{.0}$	$n_{..}$

- Y : 이산형 관심변수
- X : 이산형 보조변수, $q=P(X=1)$,
- R : 응답 지시변수
- 응답 성향 : $P(R)=P(R=1)$,

○ 무응답 메커니즘에 따른 추정량 (Zhang et al.(2013), Särndal et al(1992))

메커니즘	비율 추정량	(근사)분산 추정량
완전임의결측 (MCAR)	$\hat{p}_y = \frac{n_{.1}}{n_{..}}$	$\widehat{Var}(\hat{p}_y) = \frac{n_{.1}n_{.0}}{n_{..}^3}$
임의결측 (MAR)	$p_y^* = q\frac{n_{11}}{n_{1.}} + (1-q)\frac{n_{01}}{n_{0.}}$	$\widehat{Var}(p_y^*) = q^2\frac{n_{11}n_{10}}{n_{1.}^3} + (1-q)^2\frac{n_{01}n_{00}}{n_{0.}^3}$
비임의결측 (MNAR)	$\tilde{p}_y = \frac{q-\frac{n_{10}}{n_{.0}}}{\frac{n_{11}}{n_{.1}}-\frac{n_{10}}{n_{.0}}}$	$\widehat{Var}(\tilde{p}_y) = \left(\frac{1-\tilde{p}_y}{\frac{n_{11}}{n_{.1}}-q}\right)^2$ 괄호 $\tilde{p}_y^2\frac{n_{11}n_{10}}{n_{.1}^3} + (1-\tilde{p}_y)^2\frac{n_{01}n_{00}}{n_{.0}^3}$ 괄호

- 비율 추정량은 교차표와 전확률공식을 통해서 유도
- 근사 분산추정량은 Y의 표본 분포가 주어진 조건 하에서 유도하였으며, 다른 추정량과 비교할 때 위의 추정량으로 구한 표준오차는 약 30% 정도 과대 추정하는 경향이 있으며, MNAR 가정 하에서의 표준오차는 MCAR이나 MAR 가정 하에서의 표준오차 보다 3배 정도 커짐

○ $\hat{\rho}$: X와 Y의 상관계수, $\hat{\rho}^2 = \frac{p_y^* - \hat{p}_y}{\tilde{p}_y - \hat{p}_y}$

다. 실증분석

○ 자료 : 노르웨이 흡연 자료('11년 2분기 ~ '12년 3분기)

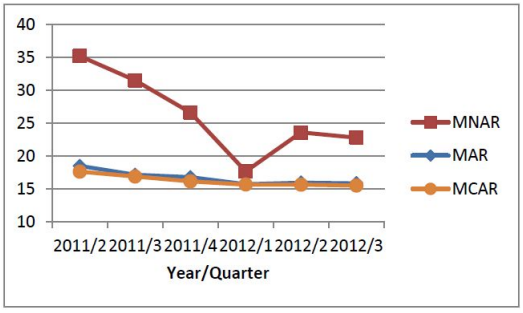
- 표본규모 2,000명, 평균 응답률 50~60%
- Y : 흡연 여부(1-흡연, 0-비흡연)
- X : 연령과 교육정도를 결합한 이산형 보조변수

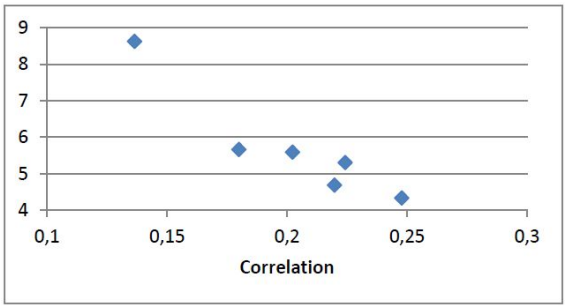
X	Unknown	Low	Medium	High
16-24	0	0	0	0
25-44	0	1	1	0
45-66	0	1	1	0
67+	1	1	0	0

	Y		
X	1	0	Total
1	759	2,353	3,112
0	357	3,402	3,759
Total	1,116	5,755	6,871

○ 주요 결과

- 흡연률에 대한 모형 비교 결과 MNAR을 가정한 모형의 추정값이 높게 나타났고,
- 흡연자의 응답 성향이 비흡연자보다 낮음(예, 2011년 2분기-각각 29%, 73%)
- MNAR 하에서 응답자 그룹의 X와 Y의 상관계수와 추정량의 표준오차는 반비례





라. 논의사항 및 향후과제

○ MAR 가정 하에서의 추정은 잘못된 결론에 이르게 할 수도 있음, 특히, MAR 모형과 MNAR 모형의 추정값의 차이가 큼

○ 조사 초기에 무응답한 표본에 대한 연구 방법

○ 무시할 수 있는 무응답인지 아닌지 확인할 수 있는 연구 방법

○ MNAR 모형의 추정값은 불편성은 만족하지만 분산이 크고, MAR 모형의 추정값은 편향되어 있지만 분산이 작아 모형의 신뢰성에 대한 문제가 제기될 수도 있는데 이런 경우, 두 추정값을 결합한 새로운 추정값을 사용하는 방안

3. Managed data collection and accuracy of estimates: The effects of degree of imbalance and degree of explanation

(Peter Lundquist(Statistics Sweden), Carl-Erik Särndal)

가. 서론

○ 연구 동기

- 연구변수 및 응답여부와 연관성이 높은 보조변수를 활용하면 조사단계에서는 균형된 응답표본을 얻을 수 있을 뿐 아니라 불편추정량을 구할 수 있음
- 추정의 정도를 높이기 위하여 활용하는 보조변수는 매우 중요하므로 보조변수의 영향력을 설명력(degree of explanation)과 균형(balance)의 관점에서 검토하고자 함

○ 연구 주제

- 연구변수에 대한 보조변수의 설명력이 추정량의 편향에 얼마나 영향을 주는가
- 보조변수에 대하여 균형된 자료를 얻는 것이 추정에 영향을 주는가
- 반응설계(Responsive design)를 응용한 방법으로 자료수집 과정을 관리하면 편향을 감소시킬 수 있는가
- 최종적으로 얻게 되는 응답자료의 불균형(Imbalance)을 얼마나 감소시킬 수 있는가
- 두 요인이 추정량의 정도에 영향을 미치는 결합효과 평가

나. 연구 모형

○ 표현식

- 모집단 : $U=\{1, \ldots, k, \ldots, N\}$
- 포함확률 : $\pi_k = \Pr(k \in s) > 0$; 설계 가중치 : $d_k = 1/\pi_k$
- 연구변수 : y_k ; 관심모수 : $Y=\sum_U y_k$
- Horvitz-Thompson estimator : $\hat{Y}_{FUL} = \sum_s d_k y_k$
- 응답 지시변수 : $I_k = 1$ for $k \in r$, $I_k = 0$ for $k \in s-r$
- (가중)응답률 : $P = \sum_s d_k I_k / \sum_s d_k = \sum_r d_k / \sum_s d_k$
- 보조변수 벡터 : x_k 는 표본에 대해서는 모두 알고 있고 모집단에서도 알려짐

○ 보조변수 벡터 x가 주어졌을 때, 표본 s에 대한 x의 가중평균은 $\bar{x}_s = \sum_s d_k x_k / \sum_s d_k$ 이고, 응답표본 r의 가중평균은 $\bar{x}_r = \sum_r d_k x_k / \sum_r d_k$ 라 할 때, $\bar{x}_r = \bar{x}_s$이면 응답표본은 x에 대하여 "균형(Balanced)"이라고 함

- x의 불균형(Imbalance)에 대한 측도

$$IMB = P^2 (\bar{x}_r - \bar{x}_s)' \Sigma_s^{-1} (\bar{x}_r - \bar{x}_s), \quad \Sigma_s = (\sum_s d_k x_k x_k')/(\sum_s d_k)$$

$\bar{x}_r = \bar{x}_s$이면 $IMB = 0$ 이 되고, $0 \le IMB \le P(1-P)$

○ 무응답이 없다는 가정 하에 평균에 대한 y의 불평추정량은 $\bar{y}_s = \sum_s d_k y_k / \sum_s d_k$ 이고, 응답표본 r에서의 가중평균은 $\bar{y}_r = \sum_r d_k y_k / \sum_r d_k$라 할 때, $\bar{y}_r = \bar{y}_s$이면 응답표본은 완벽한 "균형(Balanced)"이 됨

- $\bar{y}_s$는 알 수 없으므로 불균형을 측정하기 위해 회귀계수를 활용

$$b_r = (\sum_r d_k x_k x_k')^{-1}(\sum_r d_k x_k y_k') \; ; \quad b_s = (\sum_s d_k x_k x_k')^{-1}(\sum_s d_k x_k y_k')$$

$$\bar{y}_r - \bar{y}_s = (\bar{x}_r - \bar{x}_s)' b_r + (b_r - b_s)' \bar{x}_s$$

- 전체 표본과 응답 표본의 차이 $\bar{x}_r - \bar{x}_s$는 불균형에 의해 발생하고, 회귀계수의 차이
  $b_r - b_s$는 회귀모형의 불일치에 의해 발생됨

○ 총계 추정량은 $\hat{N} = \sum_s d_k$을 곱하여 나타낼 수 있음

$$\hat{Y}_{EXP} = \hat{N} \sum_r d_k y_k / \sum_r d_k = \hat{N} \bar{y}_r$$

$$\hat{Y}_{CAL} = \sum_r d_k m_k y_k = \hat{N} \bar{x}_s' b_r \; ; \quad m_k = (\sum_s d_k x_k)' (\sum_s d_k x_k x_k')^{-1} x_k$$

- $\hat{Y}_{EXP}$는 응답자료만 이용하므로 편향이 발생하는 반면, $\hat{Y}_{CAL}$는 보조정보를 이용하여 응답자료와 표본정보가 일치($\sum_r d_k m_k x_k = \sum_s d_k x_k$)하도록 보정하여 편향을 감소시킬 수 있고, 모집단과 일치($\sum_r d_k m_k x_k = \sum_U x_k$)하도록 보정하면 편향 감소 효과가 커짐

$$\hat{Y}_{EXP} - \hat{Y}_{FUL} = (\hat{Y}_{EXP} - \hat{Y}_{CAL}) + (\hat{Y}_{CAL} - \hat{Y}_{FUL})$$

$$\hat{Y}_{EXP} - \hat{Y}_{CAL} = \hat{N}(\bar{x}_r - \bar{x}_s)' b \; ; \quad \hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N}(b_r - b_s)' \bar{x}_s$$

다. 실증 분석

○ 응답 표본의 균형이 추정의 편향 감소에 효과가 있는지 검토하기 위해 스웨덴의 주거환경조사(LCS 2009)와 정당 선호도 조사(PPS 2012) 자료 분석

○ 불균형을 감소시키기 위한 전략

- 스웨덴 통계청은 조사과정을 관리하고 관련 기준을 만들기 위해 모든 표본에 대하여 접촉시도 기록
- 실제 조사자료와 비교할 수 있는 균형자료를 만들기 위해 표본특성에 따른 응답성향점수를 활용하여 점수가 표본에 대해서는 컨택 중단(반응설계 응용)
- 컨택을 중단하게 되는 응답성향점수를 "Intervention point(기준점수)"라고 정의
- 주거환경조사의 기준점수 $x_{k,LCS}$는 0.65, 0.60, 0.55
- 정당 선호도 조사의 기준점수 $x_{k,PPS}$는 0.65, 0.60, 0.55

* (참고) Threshold method, Equal proportions method(Särndal & Lundquist, 2014)

○ 연구변수의 설명력 정도

- 국세(등록)자료의 소득(y)를 활용하여 모든 표본에 대하여 새로운 연구변수 y_F 생성
- 등록자료를 이용하여 평균 $\bar{y}_s$와 분산 $S_y^2 = \sum_s d_k (\hat{y}_k - \bar{y}_s)^2 / \sum_s d_k$을 구할 수 있으며, 보조변수 x_k에 대한 회귀모형을 통해 표본 $k \in s$에 대하여 예측값 $\hat{y}_k = x_k' b_s$을 구함
- 회귀모형의 결정계수 $\rho^2 = \sum_s d_k (\hat{y}_k - \bar{y}_s)^2 / \sum_s d_k (y_k - \bar{y}_s)^2$

$$y_{Fk} = \bar{y}_s + S_y \left\{ F \times \frac{\hat{y}_k - \bar{y}_s}{S_{\hat{y}}} + (1 - F^2)^{1/2} \times \frac{y_k - \hat{y}_s}{S_{y-\hat{y}}} \right\}$$

$S_{\hat{y}} = \rho S_y$, $S_{y-\hat{y}} = (1-\rho^2)^{1/2} S_y$ 표준편차

$(\hat{y}_k - \bar{y}_s)/S_{\hat{y}}$, $(y_k - \hat{y}_k)/S_{y-\hat{y}}$ 는 평균이 0, 분산이 1이며 독립임

- 보조변수의 설명력의 정도를 측정하기 위해 다음의 통계량 검토

$$RDCAL = 100 \times (\hat{Y}_{CAL} - \hat{Y}_{FUL}) / \hat{Y}_{FUL}$$

$$RADJ = 100 \times (\hat{Y}_{EXP} - \hat{Y}_{CAL}) / \hat{Y}_{FUL}$$

$$RDEXP = RADJ + RDCAL$$

- y_F와 y의 평균과 분산이 동일하고 F^2가 보조변수 x_k에 대한 y_F의 회귀모형의 결정계수라 할 때, F=0.1, 0.3, 0.5, 0.7, 0.9로 가정하고 y_F의 총계에 대한 CAL 추정량과 EXP 추정량을 비교하였고, 실제 자료와 응답성향점수를 기준으로 보조 변수의 불균형도를 낮추었을 때의 결과를 비교

라. 주요 결과

- 보조변수의 설명력이 높을수록 Calibration을 통해 보정되는 크기가 커지며($RADJ\Uparrow$),
- 응답한 표본에서 보조변수에 대한 균형을 이룬다면, Calibration 후에도 남아있는 편향을 감소시킬 수 있음
- 자료수집 과정에서 균형적인 표본을 얻을 수 있는 방법을 적용하는 것은 추가적인 비용이 발생할 수 있는 반면 정확성을 제고할 수 있음
- 향후에는 자료수집 방법과 추정 방법을 동시에 고려하는 연구가 중요함