2013 통계개발원
정책연구용역

『다양한 출처 자료 처리 및 통계 생산방안 연구』

2013. 12

한국조사연구학회

제출문

통계개발원원장 귀하

본 보고서를 “다양한 출처 자료 처리 및 통계 생산방안 연구” 과제의 최종연구 결과보고서로 제출합니다.

2013년 12월 18일

한국조사연구학회장 김 영 원

연구진 :

책임 연구원	:	변종석(한신대학교 응용통계학과 교수)
공동 연구원	:	박민규(고려대학교 통계학과 교수)
		박인호(부경대학교 통계학과 교수)
		임경은(통계개발원)
		최재혁(통계개발원)
연구 보조원	:	김대건(고려대학교 통계학과 석사과정)
		모은비(성균관대학교 통계학과 박사과정)
		우일상(부경대학교 통계학과 학부과정)
		박라나(통계개발원)
		풍진주(통계개발원)

목 차

세부과제 1 : 자료 연계 및 통합 기법 연구

세부과제 2 : 조사모드의 효과적 활용 및 추정방법의 향후과제

차 례

표 차 례

그 림 차 례

연구과제 : 다양한 출처 자료 처리 및 통계 생산방안 연구

세부과제 1 : 자료 연계 및 통합 기법 연구

변종석, 박민규

I. 서론

1. 연구 배경 및 목적

1) 연구배경

최근 국내외에서는 개인 정보 보호에 대한 인식이 높아지면서 조사 참여 거부 및 무응답이 증가하고 있으며, 1인 가구 증가, 재택 시간 감소 및 고급 아파트의 증가로 인해 표본과의 접촉이 어려워지는 등 조사 환경이 악화되고 있는 실정이다. 통상적인 표본조사에서는 동일한 표본으로부터 서로 연관된 많은 양의 정보를 수집하여 분석에 필요한 데이터셋(data set)을 만들어 제공하고 있다.

그러나 최근 통계 이용자 및 수요자의 다양한 요구 수준은 점점 복잡해지고 다양하기에 이를 만족하는 조사의 수행은 어려워지고 있다. 실제로 정부나 공공기관에서도 이를 반영하여 한 조사에서 가능한 많은 정보를 얻고자 하여 많은 비용과 시간을 들여 과거보다 많은 항목에 대해 설문을 작성하여 조사를 수행하지만 설문의 양이 증가할수록 응답에 대한 부담이 증가하기에 표본의 응답률도 낮아지고 있다. 하지만 조사 참여를 확대하고 응답률을 높이기 위해서는 설문의 양을 최소화하여 응답 부담을 감소시키는 등의 방안이 가능하지만 이는 연구에 필요한 서로 연관된 많은 양의 자료수집이 충분하지 않기 때문에 한 조사에서 동일한 표본으로부터 연계된 정보를 얻어 분석하는 기존의 조사설계관점에서는 아직 널리 수용하지 못하고 있는 상황이다. 그리하여 최근 조사 환경이 악화되고, 서로 연관된 많은 양의 설문을 확보하기 어려운 상황에서 통계 이용자 및 수요자의 다양하고 복잡한 수요를 해결하는 방안으로 센서스 자료, 행정 자료 및 표본조사 자료 등 다양한 출처의 기존 조사 자료들을 대상으로 자료 통합(data fusion) 및 자료 연계(data matching) 등을 이용한 2차 자료 생산에 많은 관심과 연구가 진행되고 있다(통계개발원, 2011).

국가통계발전전략을 살펴보더라도 통계 이용의 활성화를 위한 자료 연계 등의 2차 자료 생산 및 확대 계획을 수립하여 추진하고 있으며, 이를 위해 통계청에서도 표준화된 다양한 조사자료 확보를 위한 국가통계시스템을 구축하여 선진국 수준으로 올리기 위해 노력하고 있다. 또한 국가통계의 종단자료 생산에도 관심을 두고 표본의 종단화를 추진하고 있는 데, 종단 자료 생산을 위해서는 표본으로부터 최소한의 자료만을 수집하여 조사 참여를 확대하고 다양한 자료 생산을 위해 행정자료 및 연계 가능한 다른 표본조사와의 연계를 통해 표본조사 자료의 종단화하는 방안도 검토하고 있다. 이처럼 국가 통계 발전 관점에서도 조사자료와 행정자료 및 상이한 조사자료의 연계 등을 통해 결합 마이크로 자료(synthetic micro data)의 생산을 위한 연구들을 진행하고 있다. 그러나 통계법상 개인과 기업 등의 표본 조사 자료에 대한 비밀을 보호하기 위해 가구번호나 사업자 등록번호와 같은 식별 자료

에 대한 외부 제공을 금지하고 있는 상황에서 다양한 출처 자료의 연계는 표본의 지역, 성별, 연령대, 직업 및 가구 소득 등의 인구사회경제적 특성 변수나 지역, 산업, 업종, 규모 등의 기업 활동 변수를 이용하여 가능하지만 식별 정보가 제공되지 않는 상황에서 완벽한 연계가 이루어지지 않으므로 통합된 2차 자료의 분석 결과에 대한 신뢰도는 낮아지게 될 것이다. 통계 품질이 확보되도록 보다 신뢰도가 높은 2차 자료의 생산을 위해 제공되는 조사단위의 식별 정보를 활용한 자료 연계 및 통합 기법의 연구가 필요하다.

향후 다양한 출처 자료를 이용한 자료 연계 및 통합으로 생성된 2차 자료의 생산 과정에서 통계 생산단위 및 공표단위별 자료 연계 및 통합 기법의 연구를 통해 지역 통계 등 통계 이용자 및 수요자의 세분화되고 복잡한 요구수준에 부응하는 통계 생산에 활용이 가능하다.

그러므로 악화된 조사 환경에 대처하고, 응답 부담을 감소해 조사 참여를 확대하고, 국가통계발전 전략 관점에서도 다양한 출처 자료의 연계나 통합을 통한 2차 자료의 생산에 대한 이론적 검토 및 활용 방안에 대한 연구가 필요하다.

2) 연구목적

자료 연계 및 자료 통합은 서로 다른 둘 이상의 표본조사 자료들을 하나의 합성된 결합 마이크로 자료 혹은 파일(synthetic micro file)로 결합하는 방법론을 표현하는 과정이나 개념을 의미한다. 일반적으로 기록 연계(record linkage 혹은 정확 연계/exact linkage)는 서로 다른 조사 자료들의 동일한 개체(same entity)를 결합하지만 자료 연계(data matching) 및 통합은 기본 데이터셋에 존재하는 개체들의 추가적인 자료를 동일한 개체의 조사 자료로부터 얻어 결합하는 과정이 아니라 결합하려는 개체와 유사한 개체(similar entity) 자료와 결합하여 새로운 데이터셋을 제공함을 의미한다. 즉, 자료 연계 및 통합을 통해 기존 파일에 다른 조사자료를 연계 추가해 새로운 결합 마이크로 자료(synthetic micro dataset)를 제공하게 되는 것이다. 그래서 일부 연구자들은 자료 연계 혹은 통합을 대체(imputation)의 한 방법이라고 표현하기도 한다. 자료 연계 및 통합은 개체들의 부정확한 정보나 기록을 결합하는 정확 연계 혹은 기록 연계의 대안적인 방법으로 연구되고 있는 상황이다.

이런 의미로 본 연구에서는 표본의 조사 참여 거부 및 응답 거부, 표본과의 접촉 어려움 등 악화된 조사 환경에 대응하고, 복잡하고 다양한 통계 이용자 및 수요자의 요구에 부응하며, 추가 비용 및 응답자의 부담이 없으면서 각종 정책 수립에 필요한 새로운 통계 생산 수요에 충족 가능하도록 다양한 출처 자료의 연계 및 통합 기법에 대한 연구 및 2차 자료 생산에 적용 가능한 사례 발굴을 목적으로 한다.

본 연구에서는 자료 연계 및 통합 기법은 지금까지 연구된 자료 연계 및 통합 기법, 매칭 알고리즘, 결합 자료의 활용 사례 등에 대한 문헌 연구를 통해 새로운 방안을 탐색하기로 한다. 또한, 본 연구용역에서는 통계청에서 수행하는 센서스자료 및 조사통계자료의 자료 연계 및 통합을 통한 2차 자료 생산이 가능하도록 실제 조사자료를 이용하여 자료 연계 및 통합 기법을 검토하고, 시범적으로 통계청에서 수행하는 경제활동인구조사자료와 생활시간조사 자료를 이용하여 살펴보기로 한다.

구체적인 본 연구의 목적은 다음과 같다.

첫째, 다양한 출처 자료의 자료 연계 및 통합을 위한 기법뿐만 아니라 자료 연계 및 통합과정에서 수반되는 상이한 자료들 간의 차이를 보정하기 위한 표준화 방안, 자료 연계 및 통합 과정에서 나타나는 개인 정보 보호 등 정보 유출을 최소화하는 방안 등 자료의 품질을 확보하는 자료 연계 및 통합 기법을 살펴보고,

둘째, 통계 이용자 및 수요자의 다양한 요구 수준에 부응하면서 새로운 통계 수요를 충족하고, 다양한 출처 자료의 연계 및 통합을 통해 통계의 활용 효율성을 높이기 위하여 국가통계조사에서 적용 가능한 시범사례 발굴을 위해 통계청에서 수행되는 조사통계자료를 기반으로 자료 연계 및 통합 기법 및 방안을 살펴보고,

셋째, 자료 연계 및 통합을 통해 생성된 2차 자료(결합 마이크로 자료)에서의 모수 생산을 위한 가중치 산정 방안 및 추정에 미치는 영향에 대해 살펴본다.

2. 연구 내용과 연구 방법

본 연구에서는 요구되는 연구목적을 달성하기 위해 다음과 같은 연구 내용과 방법으로 연구를 수행한다.

◦ 문헌연구
- 통계청에서 수행된 현재까지의 연구 결과 검토
- 자료 연계 및 통합 기법을 적용한 사례 검토
- 자료 연계 및 통합 기법을 적용한 사례에서의 추정 기법 검토

◦ 이론연구
- 자료 연계 및 통합 기법에 대한 이론적 연구 결과 검토
- 자료 연계 및 통합을 위한 표준화 여부 및 방안 검토
- 자료 연계 및 통합된 2차 자료를 이용한 통계 생산 방안
- 자료 연계 및 통합된 2차 자료에서의 추정량 및 가중치 산출 방안

◦ 시범사례 발굴

- 연구결과를 토대로 적용 가능한 사례 방안 모색
- 통계청에서 수행되는 조사 중 자료 연계 및 통합을 활용한 2차 자료 생산이 가능한 시범 사례 발굴

II. 문헌 연구 및 이론적 고찰

1. 기본 개념

1) 자료 연계 및 통합의 개념

자료 연계(data matching) 및 자료 통합(data fusion or data integration)은 서로 다른 둘 이상의 표본조사 자료들을 하나의 합성된 결합 마이크로 자료 혹은 파일(synthetic micro file)로 결합하는 방법론을 표현하는 과정이나 개념을 의미한다.

기록 연계(record linkage)와 차이를 간단히 요약하면, 일반적으로 기록 연계(record linkage 혹은 정확 연계/exact linkage)는 서로 다른 조사 자료들의 동일한 개체(same entity)를 결합하지만 자료 연계(data matching)는 기본 데이터셋에 존재하는 개체들의 추가적인 자료를 동일한 개체의 조사 자료로부터 얻어 결합하는 과정이 아니라 결합하려는 개체와 유사한 개체(similar entity) 자료와 결합하여 새로운 데이터셋을 제공함을 의미한다(Moriarity 2009). 이러한 관점에서 일부 연구자들은 자료 연계 혹은 통합을 대체(imputation)의 한 방법이라고 표현하기도 한다(Rodgers 1984, Rubin 1986, Singh et al. 1993). 자료 연계 및 통합은 개체들의 부정확한 정보나 기록을 결합하는 정확 연계 혹은 기록 연계의 대안적인 방법으로 연구되고 있다.

자료 연계와 자료 통합의 차이를 살펴보면, 자료 연계(data matching)는 자료 관점에서 서로 다른 자료를 결합하는 과정을 의미하고, 자료 통합(data fusion or data integration)은 서로 다른 자료를 결합하여 새로운 데이터셋을 제공하기 위한 자료 연계 과정의 모든 활동을 포괄적으로 포함하는 개념으로 사용되기에 엄밀한 의미에서 차이가 있지만 일부 연구자들은 자료 연계와 자료 통합은 동일한 개념으로 사용하기도 한다(National Research Council 1992, Kamakura and Wedel 1997).

따라서 본 연구에서도 자료 연계 및 통합을 엄밀하게 구분하지 않고 동일한 개념으로 간주하여 설명하기로 한다.

2) 기본 용어

자료 연계 및 통합을 위한 기본적인 조사 혹은 자료의 형태를 보면, <그림 1>과 같이 서로 다른 설계 및 경로로 수행된 조사 A는 공통변수 X와 조사 A의 고유목적을 위한 변수 Y에 대한 자료를 수집하고, 조사 B는 공통변수 X와 조사 B의 고유목적을 위한 변수 Z에 대한 자료를 수집한다.

만일 공통 변수 X와 고유 조사 항목인 Y자료를 수집한 조사 A의 자료에 다른

조사인 조사 B의 조사 자료 Z를 통합해 결합 자료 파일(synthetic data file)을 생성하려고 한다고 가정한다. 여기서 자료 연계 및 통합에서 자료 연계 및 통합의 기본이 되는 조사 A를 기본 조사 혹은 수용자 파일(recipient file or base file)이라고 하며, 연계 자료를 제공하는 조사 B를 공여자 파일(doner file)이라고 한다.

자료 연계 및 통합을 위한 통계적 매칭을 통해 수여자 파일 A에 미관측된 자료 Z의 측정값으로 조사 B의 자료 Z를 추가하여 결합하게 되며, 수여자 파일 A에 미관측된 자료 Z으로 조사 B의 자료 Z를 추가하여 새로이 생성된 파일을 결합 마이크로 파일(matched micro file)이라고 한다.

<그림 1> 자료 통합의 기본 구조

Variables	**X**	**Y**	**Z**
Sample or Data A	**O**	**O**	
Sample or Data B	**O**		**O**
<자료 연계 및 통합 후 결과>			
Matched micro file	**O**	**O**	**O**

기본적으로 동일모집단에서 수행된 서로 다른 조사 자료들을 통계적 매칭으로 연계하여 새로운 결합 마이크로 자료를 생성하려는 이유는 수용자 파일에서 미관측된 변수의 자료를 공여자 자료로부터 제공받아 새롭게 생성한 결합 마이크로 자료를 이용하여 추가적인 통계를 생산하거나 혹은 관심변수들의 관계를 분석하는 연구에 필요한 기초 자료를 제공하려는 것으로 볼 수 있다.

2. 문헌연구

본 절에서는 자료 연계 및 통합을 위한 통계적 매칭 알고리즘, 연계 및 통합 기법에 대한 연구, 기존 자료를 연계해 새롭게 생성한 결합 자료를 이용한 활용 사례 연구 등의 기존 연구 결과를 요약해 정리한다.

1) 자료 연계 및 통합을 위한 통계적 매칭 알고리즘에 대한 연구

자료 연계 및 통합을 위해서는 수용자 자료의 개체와 정확 또는 유사한 개체를 공여자 자료에서 자료를 찾는 방법이 매우 중요하다. 기본적으로 수용자 자료의 개체와 정확하게 일치하거나 가장 근사한 개체와 연계하는 방안이 가장 바람직한 연계가 되는 것이다. 수용자 자료와 공여자 자료의 개체 간 근사성을 측정하여 연계하는 통계적 매칭에 대한 알고리즘의 연구도 중요한 연구 분야가 되는 것이다.

자료 연계 및 통합을 위한 통계적 매칭 알고리즘에 대한 연구 결과를 정리하면, 단계적 매칭 알고리즘(van Pelt 2001), K-최근접이웃 매칭 알고리즘(Van der Putton et al. 2002), 회귀분석 매칭 알고리즘(Ingram et al. 2000), 회귀분석과 k-최근접이웃 방법의 결합 매칭 알고리즘(정성석 외 2004), 랜덤 핫덱 방법 등이 연구되어 오고 있다. 매칭 기법에 대한 주요한 연구 결과와 매칭 기법별 근사성의 측정 방법과 매칭 과정 등 자료 연계를 위한 통계적 매칭 알고리즘 연구의 구체적인 과정과 설명은 통계개발원의 연구보고서(2007)를 참조하기 바란다.

참고로 여기서는 통계적 매칭 알고리즘의 개념을 간략하게 소개한다. 자료 연계 및 통합을 위한 통계적 매칭 알고리즘으로는 가장 널리 사용되는 최근접이웃 매칭 알고리즘은 가장 유사한 하나의 개체를 매칭에 사용하는 방법이며, K-최근접이웃 매칭 알고리즘은 유사한 K개의 개체를 선택하여 매칭에 사용하는 방법이다. 회귀분석 매칭 알고리즘은 회귀분석을 적용하여 매칭을 하는 방법으로 하나의 자료 파일에서 회귀모형을 추정한 후, 추정된 회귀모형을 이용하여 두 조사 자료의 예측치를 구해 예측치 간 차이가 가장 작은 개체를 연계하는 방법이다. 랜덤 핫덱 기법은 수용자 자료의 개체와 공여자 자료의 개체를 랜덤하게 선택하게 연계하는 방법이다.

2) 자료 연계 및 통합 기법에 관한 연구

(1) 통계조사자료와 행정자료간의 자료 매칭기법 연구(통계 개발원 2007)

연구는 두 조사 자료의 매칭 기법에 대한 개념, 매칭의 여러 종류, 매칭 수행과정, 매칭 알고리즘, 평가 방법 등 선행연구 결과를 소개하고, 실제 통계청의 조사자료와 행정자료와의 매칭 기법을 연구한 보고서이다.

연구보고서를 보면, 자료 연계 및 통합을 위한 통계적 매칭 알고리즘으로 단계적 매칭 알고리즘(van Pelt 2001), K-최근접이웃 매칭 알고리즘(Van der Putton et al. 2002), 회귀분석 매칭 알고리즘(Ingram et al. 2000), 회귀분석과 k-최근접이웃 방법의 결합 매칭 알고리즘(정성석 외 2004), 랜덤 핫덱 방법에 대해 개념과 연계 과정 및 근사성의 측정 방법들을 정리하였으며, 통계적 매칭의 알고리즘에 대한 자세한 설명은 통계개발원의 연구보고서(2007)를 참조하기 바란다. 추가하여 자료 연계 후 결합한 결합 자료의 평가 방법으로는 예측력, 대표성을 소개하고 있다. 또한 모의실험을 통해 매칭 후 자료의 성질과 독립성을 검토하였다.

연구보고서에서는 통계조사자료와 행정자료 사이의 자료 연계에 대한 연구로 서울지역으로 연계 자료의 대상 지역을 제한하여 통계청의 사업체기초조사를 수용자 자료로, 국민연금자료를 공여자 자료로 선정하여 정확 매칭과 통계적 매칭 결과를 적용하였다. 두 조사 자료는 사업체 기준의 자료이며, 연계 하고자 하는 서울지역의 자료 규모는 사업체기초조사의 자료가 741,229개, 국민연금자료가 223,186개이었다. 자료 매칭을 위한 공통변수는 사업자 등록번호, 대표자 성명, 법인 등록번호, 소재지, 사업자 형태, 업종을 고려하였고, 두 조사의 결합이 요구되는 관심 변수는 종사자 수(사업체기초조사), 가입자 수(국민연금자료)를 선정하여 매칭 방법별 결과를 검토하였다.

정확매칭은 사업자 등록번호와 대표자 성명, 법인등록번호 등의 공통변수를 선정하여 정확 매칭으로 자료의 연계를 시도하였고, 매칭된 새로운 파일은 수여자 자료가 공여자 자료보다 2배 이상 크기 때문에 자료 규모가 작은 공여자 자료에 의존하게 되는 결과를 보여주고 있다. 정확 매칭에서는 수용자 자료 관점에서 많은 자료의 결측이 존재하는 결과를 제공하였다.

통계적 매칭은 정확 매칭 자료만을 이용하여 소재지, 사업장 형태, 업종을 공통변수로 선정하여 자료의 연계를 검토하였으며, 통계적 매칭 결과에 대한 평가는 공여자 자료인 국민연금자료의 가입자 수 분포와 자료 연계해 생성된 결합 마이크로 자료의 가입자 수 분포를 비교해 매칭 결과의 타당성을 보이고 있다. 공여자 자료와 매칭 자료의 가입자 수 평균, 중위값, 표준편차와 MAE 등을 비교하였으며, 매칭 결과는 사업체의 가입자 수 자료의 실제값과 매칭자료 값이 정확하게 일치하는 비율이 31.94%, 차이가 5이하인 경우가 74.97%로 나타났다.

연구 내용 및 결과를 보면, 정확 매칭을 위한 공통변수 사이의 조건을 논의하고, 통계적 매칭을 위한 공여자 자료의 추가 변수 활용 방안을 통해 매칭의 효율성을 높일 수 있다고 언급하고 있다. 하지만 두 자료를 연계한 매칭 자료는 자료의 수가 절대적으로 부족한 국민연금자료의 크기를 기준으로 정확 매칭으로 연계된 자료를 제공하고 있으며, 통계적 매칭 기법의 검토도 정확 매칭 자료만을 이용해 연구함으로써 수용자 자료 관점에서의 통계적 매칭 기법을 살펴보는 데 한계가 있는 것으로 보인다.

(2) 데이터 보강을 위한 데이터 통합 기법에 관한 연구(정성석외 2004)

연구는 자료 수집에 대한 어려움을 자료의 보강(data enrichment)으로 해결하는 방안을 모색한 연구로 자료의 보강을 위해 자료 통합(data fusion) 기법을 사용하여 자료를 통합해 새로운 결합 마이크로 자료를 생성하는 방안을 논의하고 있다.

매칭 방안으로는 회귀분석 기법에 k-최근접이웃 기법을 적용해 상대적으로 유사한 개체 자료를 이용하여 정보의 손실을 감소시켜 자료 통합의 성능을 개선하는 방안을 제안하고 있다. 회귀분석(regression analysis)방법을 이용한 통계적 매칭(statistical matching)방법은 추정치의 거리가 가장 가까운 하나의 개체만을 사용함으로써 상대적으로 유사한 다른 개체들의 정보를 무시하게 된다. 본 연구에서 제시한 수정된 데이터 통합기법은 상대적으로 유사한 개체에 대한 정보 손실을 줄여 데이터 통합기법의 성능을 높이고자 회귀분석 기법에 k-최근접이웃(k-nearest neighbor)접근법을 결합하여 가장 가까운 하나의 개체가 아니라 k개의 개체를 이용하여 통합변수를 추가시키는 방법이다. 이 방법을 자세히 살펴보면 다음과 같다.

step1. 공여자 파일에서 유일변수 Z중 임의의 s번째 변수를 목표변수로 공통변수 X를 설명변수로 하여 회귀모형을 추정한다.
step2. 추정된 회귀모형을 수용자 파일과 공여자 파일에 적용하여 각 파일에서 s번째 유일변수 Z_s의 예측치를 구한다.
step3. 두 파일에서의 예측값을 이용하여 수용자 파일의 각 개체에 대해 모든 공여자 파일의 개체와의 거리를 구한다.
step4. 계산한 거리를 이용하여 공여자 파일의 각 개체에 가장 가까운 제공자 파일에 해당하는 k개의 개체를 선택한다.
step5. 선택된 공여자 파일의 k개 개체들의 유일변수 Z_s들의 평균이나 최빈값을 구한 후 이 값을 수용자 파일의 해당 개체에 추가한다. 이때, 유일변수가 연속형이면 k개 Z값의 평균(mean)을, 범주형이면 k개 Z값의 최빈값(mode)을 이용한다.

실험 과정은 먼저 하나의 데이터를 수용자 파일과 공여자 파일로 분리한 후 기존의 데이터 통합 알고리즘과 이를 개선시키고자 제안한 알고리즘을 사용하여 데이터를 통합한 다음, 실제값과 매칭 기법을 통해 통합된 값의 차이를 근거로 두 알고리즘의 성능을 비교하였다. 수용자 파일과 공여자 파일로 개체분리를 위한 데이터 비율은 Yoshizoe and Araki(1999)에서 사용한 60%대 40%로 하고 데이터의 분리는 단순임의(simple random)방법을 사용하였다. 수용자 파일과 공여자 파일로 분리된 데이터 파일을 통합하는 과정은 근본적으로 회귀분석 기법(k=1)을 사용하며 k-최근접이웃 접근법(k=3,5,7)을 적용하였다.

수행한 데이터 통합과정은 회귀모형으로 구해진 두 파일의 예측치 차이가 1 이

하인 개체만을 통합에 고려하기 위해 회귀모형의 반응변수가 공여자 파일의 고유변수 Z가 연속형인 경우에는 표준화시키고, 공여자 파일에서 회귀모형을 적합할 때, 통합에 사용될 회귀모형에 설명력 있는 공통변수만 포함되도록 단계적(stepwise) 변수 선택을 수행하였다. 정확도 평가는 연속형 변수에 대해서는 평균제곱오차(mean squared error : MSE)를, 범주형 변수에 대해서는 오분류율(false discovery rate)을 사용하였다.

결과를 보면, 연속형 변수를 연계시키고자 할 때는 k가 1일 때보다는 k가 3,5,7로 증가할수록 보다 정확한 데이터 통합 작업을 수행 가능하였으며 이때, 일반적으로 k가 1에서 3으로 증가할 때 가장 큰 MSE의 감소를 보였다. 이는 어느 정도 오류를 감수하더라도 계산량을 줄이기 원한다면 자료 통합시 고려할 개체 수를 3으로 정하는 것이 적절하다고 주장하고 있다. 연구 결과를 정리하면 다음과 같다.

매칭 기법	회귀방법 데이터 통합기법	수정된 데이터 통합기법
통합 방법	기존 회귀분석 접근법으로 데이터 통합	k-최근접 이웃방법의 아이디어를 고려한 데이터 통합
차이점	예측치가 가장 유사한 한 개의 개체만을 사용하여 통합이 이루어짐	상대적으로 유사한 여러 개체를 사용하여 통합이 이루어짐
결론	연속형 변수를 매칭 시키고자 할 때는 k가 1일 때보다는 k가 3,5,7로 증가할수록 보다 정확한 데이터 통합 작업을 수행 가능하며 일반적으로 k가 1에서 3으로 증가할 때 가장 큰 MSE의 감소를 보이므로 어느 정도 오류를 감수하더라도 계산량을 줄이기 원한다면 데이터 통합시 고려할 개체 수를 3으로 정하는 것이 적절하다.	

(3) 통계적 매칭을 이용한 데이터 통합에 관한 연구(고은애 2004)

본 논문에서는 연계하고자 하는 변수가 범주형 변수인 경우의 자료 통합 방안을 논의하고 있다. 범주형 변수의 범주가 두 개인 경우와 세 개인 경우에 대하여 다루었다. 범주가 두 개인 경우에는 로지스틱 회귀(logistic regression)방법과 로지스틱 회귀 방법에 k-최근접이웃 방법을 같이 적용한 통합 방법의 성능을 비교하였다. 범주가 세 개인 경우에는 다항로지스틱 회귀(polynomial regression)만을 고려한 방법과 k-최근접이웃 방법을 보완한 방법을 사용하여 통합 방법의 성능을 비교하였다. 또한, 매칭에 이용되는 유클리디안 거리(euclidean distance)를 구할 때 사용되는 확률을 구하는 방법으로 다항로지스틱회귀모형과 판별분석(discriminant) 방법을 이용하였을 때의 통합 성능을 비교하였다.

기존에 제안된 데이터 통합방법(k-최근접 이웃방법의 아이디어를 고려한 데이터 통합방법, 정성석 외)을 인용하여 로지스틱 회귀분석방법에 k-근접이웃 방법을 결합하여 가장 가까운 하나의 개체 뿐 아니라 가장 가까운 k-개의 개체를 이용하여

통합변수를 추가시키는 방법을 사용하였고, 두 자료의 통합 과정은 다음과 같다.

step1. 통합될 파일(공여자 자료)의 유일변수 Z중 임의의 i번째 변수를 종속변수로 하고, 공통변수 Y를 설명변수로 하여 회귀모형을 추정한다. step2. 추정된 회귀모형을 기존 파일(수용자 자료)과 통합될 파일에 적용하여 각 파일에서 i번째 유일 변수 Z_i의 예측값 $\hat{Z}_i$을 구한다. step3. 기존 파일과 통합될 파일에서의 예측값 $\hat{Z}_{H_{ij}}$와 $\hat{Z}_{D_{ik}}$의 거리를 구하여 기존 파일의 각 개체에 가장 가까운 통합될 파일의 개체 k개를 선택한다. 여기에서 k=1일 때가 회귀분석 방법이다. 개체 k개의 관측값 Z_i의 평균이나 최빈값을 구한 후 해당하는 기존 파일의 개체에 추가한다.

연구 결과를 보면, 범주가 두 개인 모든 범주형 변수의 경우 판별분석 방법을 사용하는 것이 로지스틱회귀 방법을 사용할 때보다 더 작은 오분류율을 나타냈으며, 범주가 세 개인 모든 범주형 변수에서는 다항로지스틱회귀 방법을 이용할 때가 판별분석 방법을 이용할 때보다 더 작은 오분류율을 나타났음을 보이고 있다. 또한 범주가 두 개일 때와 세 개일 때 모두 k=1일 때보다 k가 증가함에 따라 오분류율이 감소하며, 특히 k=1에서 3으로 증가할 때 오분류율이 가장 큰 폭으로 감소함을 보이고 있다. 따라서 가장 가까운 하나의 값을 사용하는 것보다 주변에 유사한 값을 사용하는 것이 더 좋은 예측력을 나타낸다고 하였다. 그러므로 실제 데이터 예측에 있어서 범주가 두 개인 범주형의 경우 판별분석 방법을 이용하여 매칭에 필요한 확률의 예측값을 구하고, 가장 가까운 개체 세 개를 고려한다면 비록 어느 정도의 오차는 감안하더라도 가장 좋은 통합 성능을 얻을 수 있음을 보이고 있다. 연구를 요약해 정리하면 다음과 같다.

범주의 수	두 개인 경우	세 개인 경우
통합방법	로지스틱회귀모형에 k-최근접 이웃방법을 같이 적용한 것이 더 작은 오분류율을 나타냄.	다항로지스틱회귀모형에 k-최근접 이웃방법을 같이 적용한 것이 더 작은 오분류율을 나타냄.
유클리디안 거리	판별분석 방법을 사용하는 것이 더 작은 오분류율을 나타냄.	다항로지스틱회귀방법을 이용할때가 더 작은 오분류율을 나타냄.
결론	k=1일 때보다 k가 증가함에 따라 오분류율이 감소하며 특히 k=1에서 3으로 증가할 때 오분류율이 가장 큰 폭으로 감소하므로 가장 가까운 한 값을 사용하는 것보다 주변에 유사한 값을 사용하는 것이 더 좋은 예측력을 나타낸다. 그러므로 실제 데이터 예측에 있어서 범주가 두 개인 범주형의 경우 판별분석 방법을, 범주가 세 개인 범주형의 경우 다항로지스틱모형을 이용하여 매칭에 필요한 확률의 예측값을 구하고 가장 가까운 개체 세 개를 고려한다면 비록 어느 정도의 오차는 감안하더라도 가장 좋은 통합 성능을 얻을 수 있다.	

(4) 혼합형 데이터의 통계적 결합에 관한 연구(안일호 2003)

안일호의 연구는 자료 연계를 위한 단계적 매칭 알고리즘에 대한 van Pelt(2001)의 연구에서 제안된 통계적 매칭 알고리즘을 이용하여 혼합형 자료에 적용한 연구이다.

van Pelt의 연구에서 제안된 매칭 알고리즘처럼 변수 특성을 고려하여 범주형 변수만을 결합하는 경우, 연속형 변수만을 결합하는 경우, 범주형 변수와 연속형 변수를 동시에 결합하는 경우로 나누어 통계적 결합을 수행하고 있다. 통계적 결합을 수행하는 방법으로는 변수별 중요도에 따라 단계적으로 근사성을 측정하고, 제공 파일의 개체를 한 번만 결합에 사용하는 제약이 있는 통계적 결합을 수행하고, 또한 관찰된 자료의 값은 평균을 취하는 등의 변형 없이 있는 그대로를 결합하는 방안을 제안하고 있다. 이 때, 결합하고자 하는 변수에 따라 자료 간 근사성을 측정한 방법을 자세히 살펴보면 다음과 같다.

가. 결합하려는 변수가 범주형인 경우

step1. 자료간의 근사성을 측정하는 방법으로 먼저 로지스틱 회귀분석의 결과를 이용하였다. 로지스틱 회귀분석을 하여 유의한 변수를 중요변수로 간주하고 첫 단계에서 그 변수들을 이용하여 근사성을 측정한다. 얻고자 하는 변수 Y의 값은 1 또는 2의 범주형 자료이므로 제공 파일에서 Y를 종속변수로 나머지 변수를 독립변수로 하는 로지스틱 회귀 분석의 결과로 얻어지는 회귀 계수를 각 변수의 가중치로 고려한다. 즉 수용 파일과 제공 파일의 각 개체를 추정된 회귀식에 적합시켜 얻은 값을 근사성을 측정하기 위한 점수로 한다. step2. 로지스틱 회귀분석 결과 회귀식에 포함되지 않은 범주형 범주들을 이용하여 두 번째로 근사성을 측정한다. step3. 표준화한 연속형 변수의 차이로 세 번째 근사성을 측정한다.

나. 결합하려는 변수가 연속형인 경우

step1. 자료간의 근사성을 측정하는 방법으로 먼저 일반 회귀분석의 결과를 이용하였다. 일반 회귀분석을 하여 유의한 변수를 중요변수로 간주하고 첫 번째 단계에서 그 변수들을 이용하여 근사성을 측정한다. 제공 파일에서 A5를 종속변수로 나머지 변수로 독립변수로 하는 일반 회귀 분석의 결과로 얻어지는 회귀계수를 각 변수의 가중치로 고려한다. 즉 수용 파일과 제공 파일의 각 개체를 추정된 회귀식에 적합시켜 얻은 값을 근사성을 측정하기 위한 점수로 한다.

step2. 로지스틱 회귀분석 결과 추정된 회귀식에 포함되지 않은 범주형 변수들을 이용하여 두 번째로 근사성을 측정한다. step3. 표준화한 연속형 변수의 차이로 세 번째 근사성을 측정한다.

다. 범주형 변수와 연속형 변수를 동시에 결합하는 경우

step1. 결합하려는 변수가 범주형인 경우와 연속형인 경우의 첫 번째 단계의 순위합으로 근사성을 측정한다. step2. 결합하려는 변수가 범주형인 경우와 연속형인 경우의 두 번째 단계의 순위합으로 근사성을 측정한다. step3. 결합하려는 변수가 범주형인 경우와 연속형인 경우의 세 번째 단계의 순위합으로 근사성을 측정한다.

연구자는 결합하려는 변수가 범주형인 경우와 연속형인 경우, 범주형의 경우는 로지스틱회귀분석결과로, 연속형의 경우에는 일반 회귀분석결과를 이용하여 추정된 회귀식에 포함되지 않은 범주형 변수들을 이용하여 두 번째로 근사성을 측정하고, 표준화한 연속형 변수의 차이로 세 번째 근사성을 측정하여 결합하고 있으며, 범주형과 연속형을 동시에 결합하는 경우에는 세 번의 단계에 걸쳐 순위합(rank sum)으로 근사성을 측정해 결합하는 방안을 제안하고 있다.

연구 결과를 보면, 범주형 자료가 있는 경우 통계적 결합에 있어서 변수의 중요도에 따라 단계적으로 근사성을 측정하는 것은 범주형 자료의 특징을 살려 결합할 수 있는 방법 중의 하나가 되며, 연속형 변수와 범주형 변수를 동시에 고려하는 경우에 있어서 각각의 변수에 대해 측정한 개체별 근사성 측정값의 순위합을 이용하여 결합하였을 때 결합의 정확성을 잃지 않고 유지할 수 있음을 보여주고 있다.

3) 자료 연계를 통합 결합 자료를 이용한 연구

통계청의 조사 자료 중 경제활동인구조사와 가계조사 자료를 이용하여 통계 생산 단위인 지역, 성별 등의 정보 등의 관점으로 연계한 결합 자료를 이용하여 분석한 연구 결과를 요약하기로 한다.

(1) 저소득 노동시장 분석(이병희 2008)

본 연구는 저소득 노동시장의 실태와 문제점을 정태적, 동태적인 측면에서 실증적으로 규명하고, 나아가 근로빈곤층, 저소득 미취업자, 저임금 근로자를 대상으로

하는 주요 정책에 대해 다양한 자료와 분석방법을 이용하여 그 효과를 실증적으로 분석함으로써 고용 변동의 구조, 사회적 불평등과 빈곤에 대한 이해를 높이고 정책적 함의를 풍부하게 하는 데 기여하고자 연구된 논문이다.

연구자는 연구 내용에 따라 서로 다른 두 조사를 연계하여 결합 자료를 이용하여 분석하고 있는 데, 연계된 결합 자료의 내용을 소개하면 다음과 같다.

첫째, 본 연구에서 저소득 노동시장의 실태와 동태적 분석을 위해 2003~2006년 「가계조사」와 「경제활동인구조사」의 매년 6월 결합자료를 이웃하는 연도별로 대응시켜 연결한 패널자료를 구성하고, 이를 통합한 자료(일종의 연계 자료)를 이용하여 저소득 노동시장에 대한 정태적, 동태적 분석을 통해 저소득 노동시장의 구조와 특징, 노동시장에서의 상향이동 가능성을 분석하고 있다.

둘째, 종횡단면 분석기법을 활용하여 최저임금의 고용효과를 분석하기 위해 고용자료와 임금 자료를 연계하여 분석하고 있다. 고용자료는 취업자, 실업자, 비경제활동인구 등 경제활동상태를 지역 및 연령계층별로 연간 월평균으로 집계한 통계청의 「경제활동인구조사」를 이용하고, 임금자료는 노동부의 「임금구조기본통계조사」 원자료의 지방노동관서 정보를 이용하여 지역,연령계층별로 구축하여 16개 시도의 2000～2006년 연간자료를 이용하여 최저임금의 고용효과를 15～24세 청소년층, 25～54세 청장년층, 55세 이상 중고령층으로 세분하여 분석하고 있다.

(2) 근로빈곤의 동태적 분석(김혜련 2009)

본 연구는 통계청의 가계동향조사 자료와 경제활동인구조사 월별 자료를 연계한 결합 자료를 이용하여 근로 빈곤 현황과 결정 요인, 근론 빈곤의 동태적 현황을 분석하고 있다.

본 연구에서는 취업빈곤에 대한 현황 파악만으로는 다양한 형태의 빈곤층에 대한 정부의 복지 정책과 노동시장 정책을 충족할 수 없으므로 확대된 개념의 근로빈곤층에 대한 현황 파악이 필요하고, 노동과 빈곤에 대한 관련연구를 위해 「가계동향조사」와 「경제활동인구조사」의 결합 자료를 구축하여 사용하고 잇으며, 또한 결합 자료를 원자료와 비교함으로써 결합자료의 대표성 등에 대한 검증 등 정확한 구조 및 특징을 파악하고 있다.

연구 내용 중 본 연구에서 연구하고자 하는 두 조사 자료의 연계 과정과 관련된 사항을 정리하면 다음과 같다.

첫째, 자료연계를 통한 결합 자료의 조사와 연계를 위한 공통변수, 매칭 방법을 보면, 2006년~2009년 2/4분기의 통계청「가계동향조사」월별자료와 통계청「경제활동인구조사 월별자료를 가구식별번호(ID)를 이용하여 개인별 소득과 고용정보를 결합하고 있다. 연계를 위한 공통 변수는 가구식별번호, 가구주와의 관계, 성, 연령 및 취업여부를 이용하고 있으며, 공통 변수를 이용하여 「경제활동인구조사」의 개인의 고용정보와 「가계동향조사」의 가구원 소득정보를 연계한 결합자료를 생성하였다. 「가계동향조사」와 「경제활동인구조사」는 매 5년마다 표본을 개편하므로 5년간 동일 가구 및 개인별로 패널자료 생성이 가능하므로 기간별 동태적 변화를 비교하기 위해 2006~2007년 및 2008~2009년의 두 기간 중 2/4분기 기준으로 연계한 결합 패널자료를 생성하였다. 이 때, 동일한 개인에 대한 2/4분기 결합자료의 가구식별번호, 가구원 성, 생년월일을 이용하여 패널자료를 형성하였고, 월 기준 결합 자료는 분기를 기준으로 연결하여 결합 패널자료를 구축하였다. 소득 및 지출 등 가계수지는 분기평균자료를, 가구원수, 취업자수 등의 가구특성 및 개인특성 변수는 2/4분기의 마지막 조사된 자료를 대표치로 사용하였으며, 빈곤 진입 및 탈출 결정 요인, 빈곤 지속 등에 대한 분석을 위해 동기간 동안 매년 2/4분기에 모두 조사된 개인만을 결합 자료로 선택하고 있다(balanced panel).

둘째, 두 조사 자료를 연계한 결합 자료의 기초 통계량을 비교한 결과를 보면, 연계하여 생성한 결합 자료의 2003년~2009년 상반기까지 월평균 개인수는 12,980명으로 「경제활동인구조사」의 월평균 70,408명의 18.3%에 해당된다. 월평균 가구수는 6,577가구로 동기간 「가계동향조사」의 월평균 7,349가구의 88.9%에 해당된다. 두 자료를 연계한 결합 자료의 결합률은 개인 기준 2003년 18.6%에서 2007년 18.0%, 2008년 17.9%에서 2009년 상반기 16.7%로 시간이 지날수록 감소하고 있다. 동태적 분석을 위해 구축된 2006~2007년과 2008~2009년 두 기간의 패널자료 결합률은 각각 평균 48.0%, 50.9%이었으며, 2인 이상 가구로 구성된 2003~2007년 5년간 지속된 자료는 715명으로 2003년 1/4분기 결합자료 기준 1.1%로 매우 낮은 수준이었다.

셋째, 원자료와 결합자료의 대표성을 살펴보기 위해 원자료와 결합 자료의 고용현황을 비교한 결과를 보면,

① 결합 자료는 비경제활동인구의 비중이 높고 취업자 중 임금근로자의 비중이 높으며, 비경제활동인구의 비중은 2006년~2008년 3개년 평균 42.1%로 「경제활동인구조사」의 원자료 38.5%보다 3.5%p 높았다. 임금근로자 중 상용직은 평균 23.3%로 「경제활동인구조사」의 원자료 22.3%보다 1.0%p 높으며, 비임금근로자 중 자영업자의 비중은 평균 10.6%로 「경제활동인구조사」의 원자료 11.2%보다 0.6%p 낮았다.

② 결합 자료의 가계소득을 보면, 2006~2008년 평균 가계소득은 2,967천원으

로 원자료의 가계소득 2,868천원보다 약 99천원 많으며, 결합 자료의 가계 지출도 2006~2008년 평균이 2,422천원으로 원자료 2,345천원보다 77천원 많았다.

③ 결합 자료의 개인별 근로소득 및 사업소득은 2006~2009년 상반기 평균 각각 1,147천원 및 349천원으로 원자료의 근로소득 평균 1,812천원, 사업소득 평균 585천 원보다 낮았다.

④ 가계조사 원자료와 결합 자료의 빈곤율은 유사한 추이를 나타내지만 결합 자료가 원자료보다 2인 이상 가구는 평균 0.3%p, 1인 이상 가구는 평균 0.8%p 낮으며, 개인단위 결합 자료의 빈곤율도 가구단위 결합 자료와 유사한 추이를 나타내지만 가구 단위 결합 자료의 빈곤율보다 낮았다.

3. 이론적 고찰

1) 자료 연계 및 통합을 위한 가정

자료 연계(data matching) 및 자료 통합(data fusion or data integration)은 서로 다른 둘 이상의 표본조사 자료들을 하나의 합성된 결합 마이크로 자료 혹은 파일(synthetic micro file)로 결합하는 방법론을 표현하는 개념이다. 자료 연계 및 자료 통합은 서로 다른 조사 자료들의 동일한 개체(same entity)를 결합하는 기록 연계(record linkage) 혹은 정확 연계(exact linkage)와는 달리 기본 데이터셋에 존재하는 개체들과 연계하는 추가 자료를 다른 조사 자료의 동일한 개체 자료로부터 연계하여 결합하는 것이 아니라 결합하려는 개체와 유사한 개체(similar entity) 자료와 연계해 새로운 결합 데이터셋(synthetic micro dataset)을 생성함을 의미한다(Moriarity 2009).

자료 연계 및 통합에서 자료 연계 및 통합의 기본이 되는 조사 A는 기본 조사 혹은 수용자 파일(recipient file or base file)로 공통 변수 X와 고유변수 Y를 측정하고, 연계 자료를 제공하는 조사 B는 공여자 파일(doner file)로 공통 변수 X와 고유변수 Z를 측정한다고 하자. 공통변수를 이용하여 자료 연계 및 통합을 통해 수여자 파일 A의 개체단위별 자료 (X, Y)에 조사 B의 유사한 개체의 자료 Z를 연계해 추가하게 된다. 이 때 고유변수들은 연속형 변수일 수도 있으며, 범주형 변수일 수도 있으며, 혼합형 변수일 수도 있다. 매칭 과정을 통해 수여자 파일 A의 개체단위별 자료에 공여자 파일의 자료가 추가적으로 연계하여 통합 자료 (X, Y, Z)가 생성되는 데, 새로이 생성된 파일을 결합 파일(matched file), 또는 결합 마이크로 자료/결합 마이크로 데이터셋(synthetic micro dataset)이라고 한다.

서로 다른 조사 자료들을 연계하여 새로운 결합 마이크로 자료를 생성하려는 이유는 연계해 결합한 파일을 이용하여 추가적인 통계를 생산하거나 혹은 관심변수들의 관계를 분석하는 연구에 필요한 기초 자료를 제공하려는 것으로 볼 수 있다.

본 절에서는 먼저 자료 연계 및 통합을 위한 통계적 매칭에서 요구되는 몇 가지의 중요한 가정을 살펴보고자 한다.

첫째, 서로 다른 조사들의 연계 및 통합을 위한 통계적 매칭에서 요구되는 기본 가정은 연계하고자 하는 원시자료들은 동일한 모집단으로부터 조사된 자료들이어야 한다는 점이다. 즉, 공여자 파일이 수용자 파일을 대표할 수 있어야 한다는 것이다(van der Putton et al. 2002). 그러나 반드시 두 자료가 동일한 모집단에서 조사된 자료일 필요는 없다. 통계적 매칭을 위한 이 가정은 매우 중요한 가정이지만 자료 연계 및 통합 과정에서 연계하려는 원시 자료를 매칭하기 이전에 필수적으로 모집단을 조정하면 된다. 예를 들어 가구자료에 세금 환급(tax return) 자료를 연계

해 새로운 결합 자료를 생성한다면, 세금 환급 가구원이 1명 이상인 가구도 존재하고 세금 환급이 전혀 없는 가구도 존재하기 때문에 두 자료를 연계하기 전에 모집단을 조정해 연계하면 된다.

두 번째 가정은 결합 파일에서는 연계된 (Y, Z)에 관한 유용한 보조정보는 존재하지 않는다고 가정한다. 수용자 파일 A는 개체단위별로 자료 (X, Y)만 존재하고, 공여자 파일 B는 개체단위별로 자료 (X, Z)만 존재하므로 연계하여 생성한 결합 파일에서 두 조사의 고유 변수인 (Y, Z)에 대해 어떠한 정보로도 파악할 수 없음을 의미하는 것이다. 그러나 두 조사를 연계하여 새로운 결합 마이크로 자료를 생성하려는 이유는 결합 파일을 이용하여 추가적인 통계 생산 혹은 연구에 필요한 기초자료를 제공하는 것이므로 자료의 연계 및 통합을 위해서 공통 변수 X가 주어졌을 때, 두 조사의 고유변수인 Y와 Z사이에 다음과 같은 조건부 독립관계가 성립되어야 한다고 가정한다. 이를 조건부 독립성 가정(CIA ; conditional independent assumption)이라고 한다.

$$P(Y,Z|X) = P(Y|X) \cdot P(Z|X)$$

이러한 조건부 독립성을 가정하는 이유는 수용자 파일과 공여자 파일로부터는 두 자료를 연계하여 결합된 ($\boldsymbol{X, Y, Z}$)의 결합확률함수 $f(\boldsymbol{x, y, z})$를 추정할 수 없기 때문이다. 즉, 수용자 파일에서는 Z의 자료가 없으므로 $f(x,y,z) = f(y,z|x)f(x)$을 추정하지 못하고, 또한 공여자 파일에서도 Y의 자료가 없으므로 $f(y,z|x)$를 추정하지 못하게 된다. 따라서 자료 연계 및 통합의 결합 자료에 대한 결합확률함수를 추정하기 위해 공통변수 X와 두 조사의 고유 조사 항목인 Y와 Z에 대한 조건부 독립성 가정이 필요하다.

만약 조건부 독립성 가정 $f(\boldsymbol{x, y|z}) = f_{\boldsymbol{Y|X}}(\boldsymbol{y|x}) f_{\boldsymbol{Z|X}}(\boldsymbol{z|x})$이 성립된다면 ($\boldsymbol{X, Y, Z}$)의 결합확률함수는 다음과 같이 추정할 수 있다.

$$f(\boldsymbol{x, y, z}) = f_{\boldsymbol{Y|X}}(\boldsymbol{y|x}) f_{\boldsymbol{Z|X}}(\boldsymbol{z|x}) f_{\boldsymbol{X}}(\boldsymbol{x})$$

여기서 $f_{\boldsymbol{Y|X}}(\boldsymbol{y|x})$는 수용자 파일로 부터 추정이 가능하고, $f_{\boldsymbol{Z|X}}(\boldsymbol{z|x})$는 공여자 파일로 부터 추정 가능하다. 그러면 $f(\boldsymbol{x, y, z})$가 추정 가능하므로 $f(\boldsymbol{y, z})$도 다음과 같이 추정 가능하다.

$$f(y,z) = \int_{-\infty}^{\infty} f(x,y,z)dx \quad \text{for continuous } x$$

이는 조건부 독립성이 만족되면 통계적 매칭 후 각각의 조사 자료로 부터는 추정할 수 없었던 Y와 Z의 관계를 파악할 수 있게 된다는 것을 의미한다.

또한 대부분의 통계패키지에서는 조건부 독립성 가정 하에서 통계적 매칭 프로그램을 제공하고 있다.

2) 자료 연계 및 통합을 위한 통계적 매칭 방법

자료 연계 및 통합해 새로운 결함 마이크로 파일을 생성하는 통계적 매칭 방법은 매칭 목적, 매칭의 접근 방법 및 공여자 자료의 사용 제한 조건 등에 따라 다양하게 구분하고 있다.

먼저, 매칭 목적에 따라 결합 마이크로 파일에서 ($\boldsymbol{X, Y, Z}$)의 결합확률함수를 추정하거나 결합 확률 분포의 중요한 특성(예, 상관계수)을 추정하기 위해 매칭하는 매크로 매칭(macro matching)과 조건부 독립성 가정(CIA) 하에서 서로 다른 조사 결과인 수용자 파일의 개체단위에 관측되지 않은 자료 Z값으로 공여자 파일의 개체단위 자료를 연계하여 추가한 ($\boldsymbol{X, Y, Z}$)에 대한 결합 마이크로 파일을 생성하기 위해 매칭하는 마이크로 매칭(micro matching)으로 구분한다.

매크로 매칭에서는 ($\boldsymbol{X, Y, Z}$)의 결합확률함수를 추정할 때, 평균과 같은 모수 θ를 추정하여 결합확률분포함수를 추정하는 모수적 방법(parametric setting)과 두 조사 A, B의 관측 자료를 이용한 경험 분포로부터 ($Y, \boldsymbol{Z}$)의 결합 확률분포를 추정하여 ($\boldsymbol{X, Y, Z}$)의 결합확률함수를 추정하는 비모수적 방법(nonparametric setting)으로 구분하기도 한다. 즉, 수용자 파일에서 관측되지 않은 변수를 예측하여 결합 마이크로 파일을 생성할 때 비모수적 방법은 특정 모형을 가정하지 않고 자료에 기초하여 통계적 결합을 수행하는 방법이고, 모수적 방법은 연계 자료의 특성을 잘 반영하는 특정 모형을 설정하여 비관측 자료를 예측하여 결합 마이크로 자료를 생성하는 방법으로 설명할 수 있다. 비모수적 방법은 특정모형을 가정하지 않는 방법으로 사전 준비 작업이 필요하지 않고 수행 작업이 쉽다는 장점이 있으나 계산 시간이 오래 소요되는 단점이 있으며, 모수적 방법은 특정 모형을 가정해 매칭하는 방법으로 일반화의 장점이 있지만 연계 자료의 수가 많은 경우 자료 형태가 복잡하므로 모형을 설정하기 어렵거나 모형 설정이 잘못되는 경우는 적절한 매칭 결과를 제공하지 못하는 단점이 나타날 수 있다. 마이크로 매칭에서도 모수적 접근과 비모수적 접근 모두 가능하다.

그리고 모수 모형을 추정하는 매크로 매칭과 마이크로 자료를 생성하기 위해 개별 자료를 연계하여 매칭하는 비모수적 마이크로 매칭을 혼합한 혼합 매칭도 가능하며, 통계적 매칭에서 식별 문제(the identification problem)를 해소하기 위해 고려되는 다중 대체(multiple imputation)에서 매우 유용한 베이지안 접근 방법으로

도 통계적 매칭이 가능하다.

공여자 자료의 사용 제한 조건에 따른 매칭 방법은 비모수적 마이크로 매칭에서 거리 함수(distance function)를 이용하여 수용자 파일의 대체 자료를 매칭할 때 비제한적 방법(unconstrained matching)과 제한적 방법(constrained matching)으로 구분한다. 비제한적 방법은 공여자 파일의 모든 자료를 사용하여 연계하지 않는 방법으로 결합 마이크로 파일에서는 공여자 자료의 원 결과와 달라지게 되는 단점이 있고, 제한적 방법은 연계하려는 공여자 자료의 모든 개체 단위를 사용하여 수용자 자료와 한 번 이상 연계하는 방법으로 결합 마이크로 파일에서도 공여자 자료의 결과와 동일하게 유지되지만 공여자 자료의 모든 개체들을 연계하므로 공통변수 X와의 차이가 크더라도 매칭하게 되는 단점이 나타나게 된다.

본 절에서는 앞으로 통계청에서 다양한 출처 자료를 이용한 통계적 매칭 방법으로 적용할 때 검토 가능한 몇 가지의 매칭 방법을 설명하고자 한다.

(1) 비제한적 매칭 방법과 제한적 매칭 방법

자료 연계 및 통합을 위한 통계적 매칭의 기본 개념을 이해하기 위해 이론적 연구 초기인 1960년와 1970년대에 연구되었던 통계적 매칭에 대한 기본 방향을 소개하기로 한다. 초기 통계적 매칭에 대한 연구는 비모수적 마이크로 매칭관점에서 주로 공통 변수 X의 함수로 표현되는 거리 함수(distance function)를 이용하여 매칭하였으며, 서로 다른 파일을 매칭할 때 비제한적 매칭(unconstrained matching)과 제한적 매칭(constrained matching)의 두 가지 방법으로 결합 마이크로 자료를 생성하였다. 일부 연구에서는 연계하고자 하는 두 조사에서 공통변수 X에 대해 교차-분류(cross-classify)하여 얻은 정보에 대해 확률적 혹은 결정론적 과정의 어떤 형태를 사용해 교차-분류 조건하에서 두 자료를 연계(Budd 1971, Ingram et al. 2000)하였는데, 이는 계층적 대체 방법(hierarchical imputation method)과 유사한 것으로 알려져 있다(Kalton and Kasprzyk 1986).

통계적 매칭의 기본 방안을 이해하기 위해 Rodgers(1984)의 연구 자료를 인용해 설명하기로 한다(재인용 : Moriarity 2009).

예제 자료는 조사 A의 데이터 파일에는 성별, 연령과 고유변수 Y, 가중치에 대한 8명의 자료가 있고, 조사 B의 데이터 파일에는 성별, 연령과 고유변수 Z, 가중치에 대한 6명의 자료가 있다. 서로 다른 조사이므로 조사의 가중치는 다르지만 가중치 합은 24로 동일하다. 이는 동일 모집단의 표본 조사임을 의미한다. 이 때, 조사 A는 수용자 자료이고, 조사 B는 공여자 파일(doner file)로 정의하여 조사 A자료와

조사 B 자료를 연계해 새로운 결합 마이크로 파일을 생성하려고 한다.

예제 자료에서 공통 변수 X는 성별과 연령으로 선정하며, 통계적 매칭을 통해 자료 A의 고유변수 Y와 자료 B의 고유변수 Z를 연계한 새로운 결합 자료를 생성하여 다변량분석을 실시하려고 한다고 하자.

통계적 매칭 과정을 설명하면, 통계적 매칭을 위한 개체들간의 근사성 함수로는 개체들의 연령간 절대 차이로 정의된 거리 함수를 이용하며, 거리 함수 값이 가장 작은 값을 이용하여 조사 A의 개체와 동일한 성별(일종의 블록 변수라고도 함)내에서 유사한 개체를 조사 B에서 찾아 조사 A의 개체 자료에 조사 B의 고유변수 Z를 연계하여 새로운 결합 마이크로 파일을 생성하면 된다.

File A records(수용자 자료):

Case #	Sex	Age	Y	Weight
A1	M	42	9.156	3
A2	M	35	9.149	3
A3	F	63	9.287	3
A4	M	55	9.512	3
A5	F	28	8.494	3
A6	F	53	8.891	3
A7	F	22	8.425	3
A8	M	25	8.867	3

- 연령 : 평균 40.475, 표준편차(비가중) : 15.3245
- Y : 평균 8.9726, 표준편차(비가중) : 0.3782

File B records(공여자 자료);

Case #	Sex	Age	Z	Weight
B1	F	33	6.932	4
B2	M	52	5.524	4
B3	M	28	4.223	4
B4	F	59	6.147	4
B5	M	41	7.243	4
B6	F	45	3.230	4

- 연령 : 평균 43.0, 표준편차(비가중) : 11.5758
- Z : 평균 5.5498, 표준편차(비가중) : 1.5669

(가) 비제한적 매칭(Unconstrained matching) ;

비제한적 연계 방법은 두 자료를 연계할 때 공여자 자료를 복원 추출로 연계하는 방법으로 자료 A의 공통 변수 X와 유사하거나 비교 기준 값이 가장 가까운 조사 B의 표본과 연계시키는 방법이다. 비제한적 연계방법에서 두 자료를 연계하기 위해 유사한 개체를 찾는 조건은 다음과 같은 목적함수를 이용하여 동일한 성별내에서

유사한 개체를 찾는다.

성별 k내에서 유사한 개체를 찾는 목적 함수 : $\min. d(i,j)_k = |age_{iA} - age_{jB}|_k$

예를 들어 수용자 파일의 A1 개체는 성별이 남자(M)이고, 연령이 42세 이므로 A1 개체와 대응 가능한 공여자 파일의 개체는 성별이 남자인 B2, B3, B5이다. A1 개체와 연령의 절대 차이를 계산하면, 각각 10, 14, 1이므로 A1 개체와 거리 함수 값이 가장 작은 B5와 연계하면 된다. A2 개체는 남자이고, 35세이므로 공여자 파일의 연계 가능한 개체는 A1 개체와 마찬가지로 성별이 남자인 B2, B3, B5이다. 개체 A2와 연령의 절대 차이를 계산하면, 각각 17, 7, 6이므로 최소 거리를 갖는 B5 개체를 A2 개체와 연계하면 된다.

이와 같이 비제한적 매칭은 공여자 파일에서 동일한 개체가 중복하여 연계 가능하다. 만약 거리 함수 값이 동일하다면 랜덤하게 하나의 개체를 선택해 연계하면 된다.

비제한적 연계 방법으로 연계하여 생성된 결합 마이크로 파일의 자료는 다음과 같다.

Matched Case #'s	Sex	Age File A	Age File B	Y	Z	Weight
A1,B5	M	42	41	9.156	7.243	3
A2,B5	M	35	41	9.149	7.243	3
A3,B4	F	63	59	9.287	6.147	3
A4,B2	M	55	52	9.512	5.524	3
A5,B1	F	28	33	8.494	6.932	3
A6,B4	F	53	59	8.891	6.147	3
A7,B1	F	22	33	8.425	6.932	3
A8,B3	M	25	28	8.867	4.223	3

비제한적 연계 방법으로 생성된 결합 마이크로 파일에서 연계된 공여자 파일 B 자료의 연령과 고유변수 Z의 기초 통계량을 살펴보면, 결합 마이크로 파일에서 연계된 공여자 자료의 연령은 평균 43.25(s.d.=12.1037), 고유변수 Z는 평균 6.2989(s.d.=1.0378)로 원자료와 차이가 있음을 알 수 있다.

따라서 비제한적 연계 결과를 정리하면, 1) B1, B4, B5가 조사 A의 개체들과 중복적으로 연계되어 있으며, 2) B6은 연계되지 않음을 볼 수 있으며, 3) 조사 B의 공통 변수인 연령의 평균이 연계 자료에서는 달라짐을 볼 수 있으며, 4) 가중치도 조사 A의 가중치를 사용하고 있음을 볼 수 있다.

(나) 제한적 매칭(Constrained matching) ;

제한적 연계 방법은 두 자료를 연계할 때 공여자 자료를 비복원추출로 연계하는 방법으로 조사 A와 조사 B의 가중치를 유지하면서 가중치를 이용한 공통요인의 차이를 최소로 하는 개체 혹은 표본 단위를 연계하는 방법이다.

제한적 연계 방법에서는 비제한적 연계 방법과는 달리 공여자 파일의 모든 개체 자료를 이용해 연계한다는 제한조건이 추가적으로 필요하다. 제한 조건은 연계하고자 하는 파일의 가중치를 이용해 다음과 같이 표현한다.

$$\sum_{j=1}^{m} w_{ij} = w_i (\text{for } i = 1, \cdots, n), \ \sum_{i=1}^{n} w_{ij} = w_j (\text{for } j = 1, \cdots, m)$$

여기서 w_i는 공여자 파일의 개체별 가중치를 의미하고, w_j는 수여자 파일의 개체별 가중치를 의미하며, w_{ij}는 두 자료를 연계한 결합 마이크로 파일의 개체별로 조정된 가중치를 의미한다.

제한적 연계방법에서는 개별 조사의 가중치를 이용하여 제한 조건을 정의하고 있는 데, 기본적으로 자료 연계를 위한 원시자료(수용자 조사, 공여자 조사의 자료를 의미함)들은 동일한 모집단으로부터 조사된 자료이므로 두 조사의 가중치 합은 동일해야 함을 내포하고 있다(Paass 1985). 즉, 두 조사자료의 가중치 합에 대해 추가되는 제한조건은 다음과 같다(Goel and Ramalingram 1989). 이 조건은 층에 대해서도 가중치 합이 일치해야 함을 내포하고 있다.

$$\sum_{i=1}^{n} w_i = \sum_{j=1}^{m} w_j$$

제한적 연계 방법에서는 거리 함수와 제한 조건을 이용한 목적함수를 다음과 같이 정의하고 있으며, 결합 마이크로 자료에서 가중치 w_{ij}를 재부여하는 것은 목적함수를 최소로 하기 위한 것이다(Barr and Tuner 1978).

$$\min. \sum_{i=1}^{n} \sum_{j=1}^{m} (d_{ij} \times w_{ij})$$

여기서 d_{ij}는 조사 A의 개체 (j)와 조사 B의 개체 (i) 사이의 거리 함수를 의미한다.

결합 마이크로 파일에서 연계 쌍들에 대해 새롭게 재부여되는 가중치는 연계하기

전 원시자료의 속성을 유지하기 위해서 기본적으로 원시자료의 원 가중치와 동일하게 유지해야 하며, 동시에 목적함수를 최소로 하면서 수용자 및 공여자의 가중치 합 조건을 만족해야 한다. 최종적인 연계 쌍의 결정은 원 가중치 조건을 만족하면서 목적 함수 값이 최소가 되는 연계 조합을 구축하는 쌍을 선택하거나 목적함수 값이 동일한 경우는 랜덤하게 선택하면 된다. 만일 예제자료처럼 수용자 개체 가중치보다 공여자 개체 가중치가 크거나 작다면 개체의 원 가중치를 유지하기 위해 공여자 개체는 수용자 개체와 중복 연계가 가능하다.

예제 자료에 대해 제한적 연계 방법으로 유사한 개체를 연계하는 방법을 설명하고자 한다. 먼저 공통변수인 연령에 대한 두 파일의 연계 기준은 다음과 같이 설정한다. 여기서 A(j)와 B(i)는 원시자료들의 공통변수 X에 대응되는 개체를 의미한다.

$$\begin{aligned} &\text{만일 } X_j < \min X_i \text{ 이면 } (A_j, B_i) \\ &\quad X_j > \max X_{i'} \text{ 이면 } (A_j, B_{i'}) \\ &\quad X_i < X_j < X_{i'} \text{ 이면 } (A_j, B_i) \text{ 혹은 } (A_j, B_{i'}) \end{aligned}$$

예제 자료를 보면, 공여자 파일의 남자 연령 범위는 $28 \leqq X_{iM} \leqq 52$이고, 여자 연령 범위는 $33 \leqq X_{iF} \leqq 59$이다. 수용자파일의 남자 연령 범위는 $25 \leqq X_{jM} \leqq 55$이고, 여자 연령 범위는 $22 \leqq X_{jF} \leqq 63$이다.

남자		여자	
수용자	공여자 개체 탐색	수용자	공여자 개체 탐색
A1(42)	41(B5)<A1<52(B2)	A3(63)	A3>59(B4)
A2(35)	28(B3)<A2<41(B5)	A5(28)	A5<33(B1)
A4(55)	A4>52(B2)	A6(53)	45(B6)<A6<59(B4)
A8(25)	A8<28(B3)	A7(22)	A7<33(B1)

남자의 경우 수용자 개체 중 나이가 가장 작은 A8 개체(25세)는 공여자 개체 중 나이가 가장 작은 B3 개체(28세)보다 나이가 작으므로 A8 개체는 B3 개체와 연계하고, 반대로 수용자 개체 중 나이가 가장 많은 A4 개체(55세)는 공여자 개체 중 나이가 가장 많은 B2 개체(52세)보다 나이가 많으므로 A4 개체는 B2 개체와 연계하면 된다. 그리고 수용자 자료 중 A1 개체(42세)는 공여자 개체 중 41세 보다 많고 52세 보다 작은 구간 내에 존재하므로 A1 개체는 B2(52세) 개체, B5(41세) 개체와 복원으로 연계한다. 공여자 B2 개체는 A1과 A4와 복원으로 연계되고 있다. 남자 자료의 연계 결과를 보면, (A1, B2), (A1, B5), (A2, B3), (A2, B5), (A4, B2), (A8, B3)으로 연계되어 있다. 결합 마이크로 파일에서 연계된 자료의 개체 가중치를 부여하면, 먼저 단독으로 연계된 (A4, B2)의 가중치는 A4의 원 가중치를 유지하기 위해 3을 부여하게 되며, 그러면 연계된 (A4, B2)에서 B2는 3의

가중치를 가지게 된다. 공여자 B2는 또한 (A1, B2)와도 연계되어 있으므로 여기서는 B2의 가중치가 1이 되어야 원 가중치 4를 유지하게 된다. 그러면 (A1, B2)에서 A1의 가중치는 1이 되고, A1의 원 가중치 3을 유지하기 위해 (A1, B5)는 2의 가중치를 가져야 A1이 원 가중치 3을 유지하게 된다. 이러한 과정으로 결합 마이크로 파일에서 가중치를 재부여하면 최소 조건을 만족하는 재부여된 가중치를 얻을 수 있다.

여자의 경우 공여자 개체 중 나이가 가장 작은 B1 개체(33세)는 수용자 개체 중 33세보다 작은 A7 개체(22세), A5 개체(28세)와 복원으로 연계하고, 반대로 수용자 개체 중 나이가 가장 많은 A3 개체(63세)는 공여자 개체 중 나이가 가장 많은 B4 개체(59세)보다 나이가 많으므로 A3 개체는 B4 개체와 연계한다. A6 개체(53세)는 공여자 개체 중 45세 보다 많고 59세 보다 작은 구간 내에 존재하므로 B6(45세), B4(59세)와 복원으로 연계된다.

그러나 여자 공여자 중 B6 개체(45세)는 원 가중치가 4인 데, A6개체(53세, 가중치 3)와만 연계되므로 가중치 4를 만족할 수 없는 상황이다. 원 가중치를 유지하면서 가중치 합의 제한조건을 만족하기 위해 B6은 수용자 개체와 추가적으로 중복 연계해야만 한다. 공여자 B6 개체와 추가 연계가 가능한 수용자 개체는 A3(63세), A5(28세), A7(22세)이며, 목적함수 값이 가장 작은 개체와 연계하면 된다. B6 개체와 나이 차이가 가장 작은 A5와 연계한다면, A5와 B6의 추가 연계로 영향을 받는 연계 쌍은 (A5, B6), (A5, B1), (A7, B1)이다. 각 연계 쌍들에 대해 가중치를 재부여하면, 단독 연계된 (A7, B1)의 가중치는 A7의 원 가중치를 유지하기 위해 3의 가중치가 재부여되므로 (A5, B1)은 B1의 원 가중치를 유지하기 위해 1의 가중치가 재부여되고, (A5, B6)는 A5의 원 가중치 3을 유지하기 위해 2의 가중치를 재부여한다. B6와의 추가 연계로 영향을 받는 3개 연계 쌍들의 목적함수 값은 72(=5x1+17x2 +11x3)로 나타나 연계 쌍에 부여된 재가중치는 목적함수 값을 최소로 하면서 원 가중치를 유지하는 가중치가 된다. 또한 B6 개체와 나이 차이가 가장 큰 A7과 연계한다면, A7과 B6의 추가 연계로 영향을 받는 연계 쌍은 (A7, B6), (A7, B1), (A5, B1)이다. 각 연계 쌍들에 대해 가중치를 재부여하면, 단독 연계된 (A5, B1)의 가중치는 A5의 원 가중치를 유지하기 위해 3의 가중치가 재부여되므로 (A7, B1)은 B1의 원 가중치를 유지하기 위해 1의 가중치가 재부여되고, (A7, B6)는 A7의 원 가중치 3을 유지하기 위해 2의 가중치를 재부여한다. B6와의 추가 연계로 영향을 받는 3개 연계 쌍들의 목적함수 값을 계산하면 72(=5x1+17x2+ 11x3)이 된다. B6의 추가 연계가 가능한 A5와 A7은 동일한 목적함수 값을 가지게 되므로 랜덤하게 하나의 연계 쌍을 선택하여 추가 연계하면 된다. 여기서는 B6 개체는 랜덤하게 선택된 A7과 연계하여 새로운 결합 마이크로 파일을 생성하였다.

제한적 연계 방법으로 연계된 결합 마이크로 자료들을 보면, 공여자 파일의 모든 개체를 수용자 개체와 연계하고 있으며, 연계 쌍에 대해서는 각 개체들의 원 가중치 일치를 위한 제한 조건과 목적함수를 최소로 하는 가중치를 재부여하여 결합 마이크로 파일을 생성하고 있다.

공여자의 모든 개체를 사용하는 제한적 연계 방법으로 조사 A와 조사 B자료를 연계하여 생성한 결합 마이크로 파일 자료이다.

Matched Case #'s	Sex	Age File A	Age File B	Y	Z	Weight
A1,B2	M	42	52	9.156	5.524	1
A1,B5	M	42	41	9.156	7.243	2
A2,B3	M	35	28	9.149	4.223	1
A2,B5	M	35	41	9.149	7.243	2
A3,B4	F	63	59	9.287	6.147	3
A4,B2	M	55	52	9.512	5.524	3
A5,B1	F	28	33	8.494	6.932	3
A6,B4	F	53	59	8.891	6.147	1
A6,B6	F	53	45	8.891	3.230	2
A7,B1	F	22	33	8.425	6.932	1
A7,B6	F	22	45	8.425	3.230	2
A8,B3	M	25	28	8.867	4.223	3

제한적 연계 방법으로 생성된 결합 마이크로 파일에서 연계된 공여자 파일 B자료의 연령과 고유변수 Z의 기초 통계량은 가중치를 부여해 계산하면 평균이 동일함을 확인할 수 있다.

공여자의 모든 제한적 연계 방법의 연계 결과를 보면, 1) 조사 A와 조사 B의 공통요인에 대한 평균이 연계 전의 평균과 일치하며, 2) 제공 파일인 조사 B의 모든 단위와 연계되어 있으며, 3) 조사 A와 조사 B의 가중치도 동일하게 유지되어 있음을 볼 수 있다. 그러나 제한적 연계 방법은 연계를 위한 계산 과정이 복잡하고 어렵다는 단점이 있다.

(2) 매크로 매칭과 마이크로 매칭

서로 다른 조사 자료들을 연계해 새롭게 결합 마이크로 자료를 생성하려는 목적에 따라 구분하면, 매크로 매칭(macro matching)은 서로 다른 조사 자료들을 연계해 결합된 (X, Y, Z)의 결합확률함수를 추정하거나 결합확률분포의 중요한 특성을 추정하기 위해 매칭하는 것이고, 마이크로 매칭(micro matching)은 각각의 조사에서 관측되지 않은 자료를 어떤 적절한 값으로 예측해 대체하거나 조건부 독립성 가

정(CIA) 하에서 서로 다른 조사 결과인 수용자 파일의 개체단위에 관측되지 않은 자료 Z값으로 공여자 파일의 개체단위 자료를 연계하여 대체함으로써 ($\boldsymbol{X, Y, Z}$)의 결합 마이크로 파일을 생성하는 것이다.

(가) 매크로 매칭

매크로 매칭은 수용자 파일의 개체에 서로 다른 조사 자료인 공여자 자료를 연계하여 결합된 결합 마이크로 자료인 ($\boldsymbol{X, Y, Z}$)의 결합확률함수를 추정하거나 결합확률분포의 중요한 특정 특성을 추정하는 것이 목적이다.

매크로 매칭에서는 ($\boldsymbol{X, Y, Z}$)의 결합확률함수를 추정할 때, 어떤 모수 θ를 조건으로 한 분포 함수를 추정하는 모수적 방법과 연계하고자 하는 조사 자료들의 경험분포를 이용하여 경험 결합확률함수를 추정하는 비모수적 방법으로 구분해 설명하고자 한다.

① 모수적 방법을 이용한 매크로 매칭(Macro approach in a parametric setting)

모수적 방법을 이용한 매크로 매칭의 목적은 어떤 특정한 모수 θ의 조건 하에서 ($\boldsymbol{X, Y, Z}$)의 결합확률함수를 추정하거나 특정 특성을 파악하는 것이 관심사항이다. 결합 마이크로 파일에서 어떤 특정한 모수 θ를 고려한 결합 마이크로 자료 ($\boldsymbol{X, Y, Z}$)의 결합확률함수는 다음과 같이 표현 가능하다.

$$f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}\,|\boldsymbol{\theta}) = f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}\,|\,\boldsymbol{x}\colon \boldsymbol{\theta}_{\boldsymbol{y}|\boldsymbol{x}}) f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}\,|\,\boldsymbol{x}\colon \boldsymbol{\theta}_{\boldsymbol{z}|\boldsymbol{x}}) f_{\boldsymbol{X}}(\boldsymbol{x}\colon \boldsymbol{\theta}_{\boldsymbol{x}})$$

모수적 방법을 이용한 매크로 매칭에서는 모수 $(\theta_X, \theta_{Y|X}, \theta_{Z|X})$를 추정하여 결합 마이크로 자료의 ($\boldsymbol{X, Y, Z}$)에 대한 결합확률함수를 추정하게 된다.

모수 $(\theta_X, \theta_{Y|X}, \theta_{Z|X})$는 연계하려는 두 조사의 전체 자료($A \cup B$)의 관측 우도 함수(the observed likelihood function)를 이용한 최대우도추정법으로 추정한다. Rubin(1974)은 일부 결측 자료가 존재하더라도 반복 과정 없이 어느정도의 적절한 완전 자료를 이용해 최대우도추정값을 직접 유도할 수 있음을 보였다. θ_X의 최대우도 추정은 연계하고자 하는 전체 자료($A \cup B$)로부터 계산 가능하고, $\theta_{Y|X}$와 $\theta_{Z|X}$는 각각 조사 A와 B 자료로부터 계산 가능하다. 자세한 사항은 Rubin(1974)의 연구 결과를 참조하기 바란다.

예를 들어 모수적 방법을 이용한 매크로 매칭을 설명하면, 두 조사 자료를 연계

한 결합 마이크로 파일에서 ($\boldsymbol{X, Y, Z}$)가 모수가 다음과 같은 3변량 정규분포를 따른다고 가정하자.

$$f(\boldsymbol{x, y, z}\,|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim MN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\text{여기서 } \theta = (\mu, \Sigma)' = \left[\begin{pmatrix}\mu_X\\ \mu_Y\\ \mu_Z\end{pmatrix}, \begin{pmatrix}\sigma_X^2 & \sigma_{XY} & \sigma_{XZ}\\ \sigma_{XY} & \sigma_Y^2 & \sigma_{YZ}\\ \sigma_{XZ} & \sigma_{YZ} & \sigma_Z^2\end{pmatrix}\right].$$

조건부 독립성 가정 하에서 결합 마이크로 파일의 ($\boldsymbol{X, Y, Z}$)에 대한 결합확률함수 $f(\boldsymbol{x, y, z}\,|\boldsymbol{\mu}, \boldsymbol{\Sigma})$는 분해 가능하며, 이를 통해 모수 $(\theta_X, \theta_{Y|X}, \theta_{Z|X})$를 추정할 수 있다. 예로써, 결합 마이크로 자료의 결합확률함수를 추정하기 위하여 조사 A에서 공통변수 X 조건에서 고유변수 Y에 대한 조건 분포 $f_{\boldsymbol{Y|X}}(\boldsymbol{y}\,|\,\boldsymbol{x}\colon \boldsymbol{\theta_{y|x}})$를 추정하면, 다음과 같은 모수 $\theta_{Y|X} = (\mu_{Y|X}, \sigma^2_{Y|X})$를 갖는 정규분포를 따르게 된다(Anderson 1984).

$$\mu_{Y|X} = \alpha_Y + \beta_{YX} \cdot X$$

$$\sigma^2_{Y|X} = \sigma^2_Y - \frac{\sigma^2_{XY}}{\sigma^2_X} = \sigma^2_Y - \beta^2_{YX}\sigma^2_X$$

$$\text{여기서} \quad \alpha_Y = \mu_Y - \beta_{YX}\mu_X,\ \beta_{YX} = \frac{\sigma_{XY}}{\sigma^2_X}.$$

그리고 이 결과를 이용하면 공통변수 X 조건에서 고유변수 Y에 대한 조건 분포 $f_{\boldsymbol{Y|X}}(\boldsymbol{y}\,|\,\boldsymbol{x}\colon \boldsymbol{\theta_{y|x}})$는 다음과 같은 회귀모형으로 표현할 수 있다.

$$Y = \mu_{Y|X} + \epsilon_{Y|X} = \alpha_Y + \beta_{YX}X + \epsilon_{Y|X} \quad \text{, 여기서 } \epsilon_{Y|X} \sim N(0, \sigma^2_{Y|X}).$$

이와 동일한 과정으로 조사 B를 이용하여 공통변수 X 조건에서 고유변수 Z에 대한 조건 분포 $f_{\boldsymbol{Z|X}}(\boldsymbol{z}\,|\,\boldsymbol{x}\colon \boldsymbol{\theta_{z|x}})$도 추정 가능하다.

이와 같이 조건부 독립성 가정 하에서 각각의 조사자료로부터 조건 분포를 추정하면 결합 마이크로 자료 ($\boldsymbol{X, Y, Z}$)에 대한 결합확률함수를 추정할 수 있게 된다. 결합 마이크로 자료 ($\boldsymbol{X, Y, Z}$)에 대한 다변량 정규분포의 구체적인 특성과 이론적 내용은 Anderson(1984)의 논문을 참조하기 바란다.

모수적 방법을 이용한 매크로 매칭 방법으로 조건부 독립성 가정 하에서 결합 마이크로 자료 ($\boldsymbol{X, Y, Z}$)에서 ($Y, \boldsymbol{Z}$)의 상관계수를 추정하기 위해 공분산은 다음과

같이 계산한다(D'Orazio et al. 2006, p.p. 14-19).

$$\sigma_{YZ} = \frac{\sigma_{XY}\sigma_{XZ}}{\sigma_X^2}$$

조건부 독립성 가정 하에서 공통변수 X 조건에서 ($Y, \boldsymbol{Z}$)의 상관계수는 $\rho_{YZ|X} = 0$ 이므로 ($Y, \boldsymbol{Z}$)의 상관계수는 다음과 같다.

$$\rho_{YZ} = \rho_{XY}\rho_{XZ}$$

② 비모수적 방법을 이용한 매크로 매칭(Macro approach in a nonparametric setting)

비모수적 방법을 이용한 매크로 매칭의 목적은 특정 모수 θ를 가정하지 않고 관측 자료의 경험 분포를 이용하여 ($\boldsymbol{X, Y, Z}$)의 결합확률함수를 추정하거나 특정 특성을 파악하는 것이 관심사항이다. 비모수적 방법에서는 경험 누적분포함수를 이용한 추정 방법과 경험 밀도함수를 직접 추정하는 방법으로 결합확률함수의 추정이 가능하다. 비모수적 방법을 이용한 매크로 매칭에 대한 구체적인 사항은 Wand and Jones(1995)의 연구 결과를 참조하기 바란다.

첫 번째로 경험 누적분포함수를 이용하는 방법을 설명하면, 결합 마이크로 파일에서 조건부 독립성 가정을 만족할 때 관측 자료를 이용한 결합 마이크로 자료 ($\boldsymbol{X, Y, Z}$)에 대한 누적 결합분포함수는 다음과 같이 표현 가능하다.

$$F_{YZ|X}(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{x}) = F_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y} \mid \boldsymbol{x}) F_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z} \mid \boldsymbol{x})$$

경험 누적분포함수를 이용한 비모수적 방법의 매크로 매칭은 공통변수 X 조건하에서 고유 변수 Y와 Z의 누적 분포함수를 조사 A와 B로부터 다음과 같이 각각 추정하면 결합 마이크로 자료 ($\boldsymbol{X, Y, Z}$)에 대한 누적 결합분포함수를 추정할 수 있다.

$$\hat{F}_{Y|X}(y|x) = \frac{\sum_{a=1}^{n_A} I(y_a \leqq y) I(x_a \leqq x)}{\sum_{a=1}^{n_A} I(x_a \leqq x)}, \quad \hat{F}_{Z|X}(z|x) = \frac{\sum_{b=1}^{n_B} I(z_b \leqq z) I(x_b \leqq x)}{\sum_{b=1}^{n_B} I(x_b \leqq x)}$$

두 번째로 경험 밀도함수를 직접 추정하는 방법을 설명하면, 결합 마이크로 파일에서 관측 자료를 이용한 결합 마이크로 자료 ($\boldsymbol{X, Y, Z}$)에 대한 결합밀도함수는 다음과 같이 표현 가능하다.

$$f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x})$$

만일 연계하려는 전체 표본자료($A \cup B$) $n_A + n_B$가 밀도함수 $f \in F$로부터 측정된 관측자료가 독립적이고 동일한 분포를 따른다면, 모수적 방법(①에서 설명)과 마찬가지로 전체 표본 자료($A \cup B$)의 어떤 적절한 완전 자료(complete data)로부터 추정이 가능하다. 모수적 방법과 다른 점은 커널 밀도 추정량(kernel density estimator)과 같은 비모수적 방법을 사용하는 점이 다른 것이다.

결합 마이크로 자료의 결합밀도함수를 추정하기 위하여 조사 A에서 공통변수 X 조건에서 고유변수 Y에 대한 조건 분포 $f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})$를 추정하면, 다음과 같은 커널 밀도 추정량을 이용해 추정 가능하다.

$$\hat{f}_{Y|X}(y|x) = \frac{\hat{f}_{XY}(x,y)}{\int_{-\infty}^{\infty} \hat{f}_{XY}(x,y)dy} = \frac{\sum_{a=1}^{n_A} K_2\left(\frac{x-x_a}{h_x}, \frac{y-y_a}{h_y}\right)}{\sum_{a=1}^{n_A} \int_{-\infty}^{\infty} K_2\left(\frac{x-x_a}{h_x}, y\right)dy}$$

여기서 $\hat{f}_{XY}(x,y) = \frac{1}{n_A} \sum_{a=1}^{n_A} \frac{1}{h_x h_y} K_2\left(\frac{x-x_a}{h_x}, \frac{y-y_a}{h_y}\right)$, $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} K_2(x,y)dxdy = 1$.

이와 유사한 과정으로 전체 표본자료로부터 $\hat{f}_{\boldsymbol{X}}(\boldsymbol{x})$를 추정하고, 조사 B 자료로부터 $\hat{f}_{Z|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x})$를 추정하여 결합밀도함수를 추정하면 된다. 이에 대한 자세한 사항은 Wand and Jones(1995)의 연구 결과를 참조하기 바란다(재인용; D'Orazio et al. 2006, p.p. 32-33).

비모수적 매크로 매칭에서는 커널 방법 대신에 최근접 이웃 방법(the nearest neighbor method)을 이용하거나 k-최근접 이웃 방법(the k-nearest neighbor method : kNN)을 이용해 통계적 매칭에 적용할 수 있다(Silverman 1986). 또한 커널 추정량 대신 비모수적 마이크로 매칭에서 널리 사용되는 k-최근접 이웃 방법을 이용한 비모수적 회귀 추정량을 대신 사용할 수도 있다(Eubank 1988).

(나) 마이크로 매칭

마이크로 매칭(micro matching)은 개별 조사에서 관측되지 않은 자료를 어떤 적절한 값으로 예측해 대체하거나 조건부 독립성 가정(CIA) 하에서 서로 다른 조사 결과인 수용자 파일의 개체단위에 관측되지 않은 자료 Z값으로 공여자 파일의 개체단위 자료를 연계하여 대체함으로써 (X, Y, Z)의 결합 마이크로 파일(synthetic micro file or synthetic micro complete data set)을 생성하는 것이 목적이다.

마이크로 매칭은 결합하고자 하는 자료를 개별 조사에서 결측 자료로 보고 이를 다른 조사 자료의 값으로 대체함으로써 결합 마이크로 자료를 생성하게 된다. 마이크로 매칭도 매크로 매칭과 마찬가지로 모수적 방법과 비모수적 방법으로 매칭이 가능하다.

① 모수적 방법을 이용한 마이크로 매칭

모수적 방법을 이용한 마이크로 매칭에서는 조사 A에서 결측 자료 Z를 예측하고, 조사 B에서는 결측 자료 Y를 예측하기 모수 모형(parametric model)이 추정되면 전체 표본자료 $A \cup B$의 결측 자료를 주어진 관측 변수의 분포와 대응하는 변수의 분포에서 적절한 값으로 교체(subtitution)하여 결합 자료를 생성하는 것이다.

모수적 방법을 이용한 마이크로 매칭은 관측되지 않은 결측 자료를 예측한 값으로 대체하기 때문에 예측적 접근(predictive approach)이라고도 한다. 즉, 모수적 방법을 이용한 마이크로매칭은 연계하려고 하는 자료를 결측 자료로 보고 관측 자료를 이용한 예측 모형으로부터 추정한 예측 값으로 교체(subtitution)함으로써 결합 마이크로 자료를 생성하게 된다. 이 방법은 어떤 명확한 모수 모형을 사용한 대체(imputation)의 한 방법으로 생각할 수 있다.

모수적 방법을 이용한 마이크로 매칭에서는 조건부 평균 매칭(conditional mean matching)과 조건부 예측 분포에 근거한 추출(draws based on a conditional predictive distribution)의 두 가지 방법이 널리 사용된다.

◦ 조건부 평균 매칭(conditional mean matching)

예측적 접근인 모수적 방법을 이용한 마이크로 매칭에서 가장 중요한 점은 개별적인 결측 자료를 관측된 자료들로부터 계산 가능한 결측 변수의 기댓값으로 교체하는 것이다. 이 의미는 개별 결측 자료의 대체값을 결측 변수의 기댓값으로 대체하기 때문에 공통변수 값이 동일한 개체는 대체값이 모두 동일하게 된다는 것이다.

이 방법은 조사 A와 B에서 미관측된 결측 자료를 다음의 조건을 만족하는 조건부 기댓값으로 교체한다.

$$\text{조사 A에서 결측된 Z값의 대체 : } \widetilde{z_a} = E(Z|X=x_a)$$
$$\text{조사 B에서 결측된 Y값의 대체 : } \widetilde{y_b} = E(Y|X=x_b)$$

만약 변수들이 다변량 정규분포를 따른다면, 조건부 기댓값을 계산하는 과정에서 미지의 모수 $\theta_{Z|X}$와 $\theta_{Y|X}$는 모수적 방법을 이용한 매크로 매칭의 최대우도 추정값을 사용할 수도 있다. 그러므로 대체값(imputed value)은 X에 대한 Z의 추정된 회귀모형이나 X에 대한 Y의 추정된 회귀모형으로 추정된 값이 된다. 연속형의 관측 자료들이 정규분포를 따를 때, 결측 자료를 대체하기 위한 조건부 평균 대체 방법으로는 회귀 대체(regression imputation)가 적절한 방법이다(Little and Rubin 2002, p. 62).

예를 들어 조사 A는 공통변수 X와 고유 변수 Y를 측정한 자료이고, 조사 B는 공통변수 X와 고유 변수 Z를 측정한 자료라고 가정하자. 조건부 평균 매칭으로 대체값을 구하는 과정을 설명하면, 조사 A로부터 추정한 공통변수 X에 대한 고유 변수 Y의 회귀모형식에 조사 B의 공통변수 X자료를 대입하여 예측된 Y값을 조사 B에서 미관측해 결측된 Y의 대체값으로 사용하고, 조사 B로부터 추정한 공통변수 X에 대한 고유 변수 Z의 회귀모형식에 조사 A의 공통변수 X자료를 대입하여 예측된 Z값을 조사 A에서 미관측해 결측된 Z의 대체값으로 사용한다. 즉, 이를 표현하면 다음과 같다.

$$\text{조사 A에서 결측된 Z값의 대체값(조사 B의 회귀식 이용) : } \tilde{z}_a^A = \hat{\alpha}_Z + \hat{\beta}_{ZX}\, x_a^A$$
$$\text{조사 B에서 결측된 Y값의 대체값(조사 A의 회귀식 이용) : } \tilde{y}_b^B = \hat{\alpha}_Y + \hat{\beta}_{YX}\, x_b^B$$

여기서 회귀계수는 매크로 매칭에서 설명한 최대우도 추정법으로 추정한다. 회귀대체를 이용한 조건부 평균 매칭 방법은 실제 관측 자료가 아니며, 또한 결측 자료의 대체값이 모두 회귀직선 상의 값으로 대체되기 때문에 공통 변수 X 조건에서 예측된 대체값은 어떠한 변동성을 나타내지 못하게 된다. 그리고 이 방법은 대체가 필요한 자료가 범주형 변수일 때 심각한 문제가 존재하게 된다.

조건부 평균 매칭 방법은 결측 자료를 변수의 기댓값으로 대체하기 때문에 2차 손실함수(quadratic loss function) 관점에서 최적의 추정값을 제공하는 것으로 알려져 있다. 그렇지만 1) 대체값은 실제 관측된 값이 아니고, 2) 결측 자료를 예측값으로 대체한 결합 마이크로 자료의 분포가 공통변수 X 조건 하에서 고유 변수 Y 혹은 Z의 기댓값이 집중화되는 경향을 보이는 단점들이 존재하여 좋은 매칭 방법이

라고 말할 수는 없다. 그럼에도 불구하고 이 방법은 널리 사용되고 있다.

◦ 조건부 예측 분포에 근거한 추출(draws based on a conditional predictive distribution)

앞서 설명한 조건부 평균 매칭 방법은 공통변수와 같은 매칭을 위한 조건 변수가 동일하면 대체 값이 어떤 변화를 줄 수 없기 때문에 모두 동일한 값으로 대체되는 단점이 있다.

Little and Rubin(2002, p. 66)은 MAR 매커니즘 조건 하에서 다변량 분포로부터 생성되는 자료가 예측값의 분포로부터 랜덤하게 추출하여 결측 자료를 대체하는 것이 더 낫다는 것으로 보였다. 이 방법은 조사 A의 결측 자료 Z는 $f_{\boldsymbol{Z}|\boldsymbol{X}}\left(\boldsymbol{z}\,|\,\boldsymbol{x_a};\,\hat{\boldsymbol{\theta}}_{z|x}\right)$ 분포로부터 랜덤하게 추출하고, 조사 B의 결측 자료 Y는 $f_{\boldsymbol{Y}|\boldsymbol{X}}\left(\boldsymbol{y}\,|\,\boldsymbol{x_b};\,\hat{\boldsymbol{\theta}}_{y|x}\right)$로부터 랜덤하게 추출하여 대체하는 것이다. 이 때, 랜덤 추출을 위해 사용되는 분포에서 미지의 모수 $\theta_{Z|X}$와 $\theta_{Y|X}$는 최대우도 추정값으로 대체해 얻을 수 있다.

예를 들어 설명하면, 관측 자료 X, Y, Z가 다변량 정규분포를 따를 때, 조건부 예측 분포에 근거한 추출의 매칭 방법은 확률적 회귀 대체(stochastic regression imputation) 방법으로 설명할 수 있다.

조건부 평균 매칭의 회귀 대체 모형식을 확률적 회귀 대체 방법에서 최대우도추정을 이용하여 추정한 회귀계수를 사용한 회귀모형으로 표현하면 다음과 같다.

조사 A에서 결측된 Z값의 대체값(조사 B의 회귀식 이용) :

$$\tilde{z}_a^A = \hat{\alpha}_Z + \hat{\beta}_{ZX}\, x_a^A + e_a$$

조사 B에서 결측된 Y값의 대체값(조사 A의 회귀식 이용) :

$$\tilde{y}_b^B = \hat{\alpha}_Y + \hat{\beta}_{YX}\, x_b^B + e_b$$

여기서 e_a는 $N(0,\hat{\sigma}^2_{Z|X})$의 분포로부터 랜덤하게 생성하고, e_b는 $N(0,\hat{\sigma}^2_{Y|X})$의 분포로부터 생성한다. $\hat{\sigma}^2_{Z|X}$과 $\hat{\sigma}^2_{Y|X}$는 모수적 방법을 이용한 매크로 매칭의 식을 이용해 추정한다.

② 비모수적 방법을 이용한 마이크로 매칭

비모수적 방법을 이용한 마이크로 매칭에서는 관심 변수에 대한 특정한 모수적 분포를 가정하지 않고 조사 A와 조사 B 자료의 합병(fusion)을 통해 결합 마이크로 자료를 생성하는 것이 목적이다(Okner 1972).

비모수적 방법을 이용한 마이크로 매칭은 두 가지 방법으로 결합 마이크로 자료를 생성하고 있다. 하나는 모수적 방법의 마이크로 매칭에서 조건부 예측 분포에 근거한 랜덤추출과 마찬가지로 비모수적 방법으로 결합 자료 ($\boldsymbol{X, Y, Z}$)의 분포를 추정하여 랜덤 추출하는 방법이고, 다른 하나는 모수적 방법의 마이크로 매칭에서 조건부 평균 매칭과 마찬가지로 관측된 변수로 대체하기 위해 비모수적 회귀 모형(nonparametric regression model)의 값으로 대체하는 조건부 평균 매칭 방법이다. 이 두 가지의 비모수적 마이크로 매칭 방법은 다양한 비모수적 추정 과정에 따라 여러 가지의 추가적인 접근으로 통계적 매칭이 가능하다. 가장 널리 사용되는 비모수적 대체 방법으로는 관측 자료를 대체 값으로 사용하는 핫덱 대체(hot-deck imputation) 방법이 있다. 핫덱 대체 방법은 분포에 대한 가정이 필요 없고, 어떤 특성의 분포를 추정할 필요가 없기 때문에 널리 사용되는 대체 방법이다. 하지만 어떤 분포나 조건부 평균 함수에 대한 추정을 내재적으로 가정하고 있다. 그래서 핫덱 대체 방법을 사용한 통계적 매칭에서는 특정한 틀(framework) 속에서 매칭하고 있다(Singh et al. 1993). 핫덱 대체 방법으로 통계적 매칭을 하기 위해서 한 조사는 수용자 파일로, 다른 하나는 수용자 파일의 결측 자료에 대해 대체 자료를 제공하는 공여자 파일로 정의해야 한다.

통계적 매칭에 사용되는 핫덱 대체 방법은 랜덤 핫덱(random hot deck), 순위 핫덱(rank hot deck), 거리 핫덱(distance hot deck)의 세 가지 방법이 주로 사용된다(Singh et al. 1993).

◦ 랜덤 핫덱(random hot deck)

랜덤 핫덱은 수용자 파일의 결측 자료를 대체하는 값으로 공여자 파일의 자료로부터 랜덤하게 선택하여 매칭하는 방법이다. 이 때, 매칭을 위한 공통 변수로는 각종 조사에서 주로 사용하는 지역, 성별, 연령 등의 일반적인 공통 특성에 따라 층화한 동질적인 그룹을 형성하고, 그룹 내에서 랜덤하게 추출해 대체하는 것이 적절하다. 이러한 그룹을 공여 계층(donation classes)이라고 한다. 일반적으로 공여 계층은 하나 이상의 범주형 변수를 사용한다.

범주형 공통변수 X로 층화한 공여 계층 내에서 예측하는 비모수적 매크로 매칭과 랜덤 핫덱의 비모수적 마이크로 매칭을 비교하면 조사 B에서 공통변수 X 조건하에서 Z의 조건 분포를 추정하고, 추정 결과로부터 랜덤하게 관측 값을 추출하는 과정은 동일하다. 즉, $F_{Z|X}$는 경험 누적분포함수로부터 추정하며, 계층 X내의 공여자 파일로부터 관측 값을 랜덤하게 추출하는 것은 추정된 $\hat{F}_{Z|X}$에서 값을 추출하는 것과 같다는 의미이다. 만일 Z가 범주형 변수라면 일치하게 된다.

공여 계층을 정의하지 않고 랜덤 핫덱으로 매칭하는 경우는 Z와 X가 독립이라는 가정이 필요하다. 이 경우는 추정된 $\hat{F}_{Z|X}$대신에 조사 B의 주변 경험분포(marginal empirical distribution)로부터 대체 값을 찾아 매칭하면 된다.

◦ 순위 핫덱(rank hot deck)

순위 핫덱은 공통 변수 X를 기준으로 순위 매칭(ordinal matching)할 때, 공통 변수 X 사이에 순위 관계를 활용하여 수용자 파일의 결측 자료를 대체하는 값을 공여자 파일의 자료로부터 선택해 매칭하는 방법이다(Singh et al. 1990).

순위 핫덱으로 매칭하기 위해서는 수용자 파일과 공여자 파일을 각각 공통변수 X를 기준으로 순위를 부여한다. 이 때, A가 수용자 파일이고, 공여자 파일의 자료가 $n_B = kn_A$(여기서 k는 정수)라고 하면, 자료가 같다면 수용자 파일과 공여자 파일에서 같은 순위의 자료를 매칭하면 된다. 그러나 두 파일의 자료 수가 다르다면 공통변수 X의 경험 누적분포함수를 고려하여 매칭을 하면 된다. 순위 핫덱을 위해 수용자 파일과 공여자 파일에서 X의 경험 누적분포함수는 다음과 같이 추정한다.

$$\text{수용자 파일 : } \hat{F}_X^A(x) = \frac{1}{n_A}\sum_{a=1}^{n_A} I(x_a \leqq x)$$

$$\text{공여자 파일 : } \hat{F}_X^B(x) = \frac{1}{n_B}\sum_{b=1}^{n_B} I(x_b \leqq x)$$

그리고 수용자 파일의 결측 자료를 대체하는 공여자 파일의 자료는 다음의 조건을 만족하는 b^*와 관련된 개체 자료를 선택해 매칭하면 된다. 이는 수용자 파일의 순위와 공여자 파일의 차이가 작은 순위에 해당하는 공여자 파일의 자료를 이용하여 매칭한다는 의미이다.

$$\left|\hat{F}_X^A(x_a^A) - \hat{F}_X^B(x_{b^*}^B)\right| = \min_{1 \leq b \leq n_B}\left|\hat{F}_X^A(x_a^A) - \hat{F}_X^B(x_b^B)\right|$$

◦ 거리 핫덱(distance hot deck)

거리 핫덱은 통계적 매칭의 초기 연구단계에 널리 사용되던 방법으로 매칭 기준 변수를 이용한 거리를 측정하여 수용자 파일의 개체 자료와 가장 근접한 공여자 파일의 자료를 매칭하는 방법이다(Okner 1972, Rodgers 1984).

예를 들면, 어떤 연속형의 변수 X를 기준으로 가장 단순하게 거리를 계산하여 수용자 파일의 결측 자료를 대체하는 공여자 파일의 자료는 다음의 조건을 만족하는 b^*와 관련된 개체 자료를 선택해 매칭하면 된다. 이는 수용자 파일의 기준 변수 값과 공여자 파일의 기준 변수 값이 가장 근사한 공여자 파일의 자료를 이용하여 매칭한다는 의미이다.

$$d_{ab^*} = \left|x_a^A - x_{b^*}^B\right| = \min_{1 \le b \le n_B}\left|x_a^A - x_b^B\right|$$

만일 수용자 파일의 자료와 차이가 같은 공여자 파일의 대체 자료가 여러 개 존재한다면 랜덤하게 하나의 자료를 선택해 매칭하면 된다.

거리 함수를 이용한 통계적 매칭의 초기 연구단계에서는 공여자 파일의 개별 개체 자료를 한 번이상 대체값으로 매칭할 수 있는 비제한적 거리 핫덱 방법(unconstrined distance hot deck)과 오직 한 번만 매칭하도록 제한하는 제한적 거리 핫덱(constrined distance hot deck)으로 구분하여 매칭하였다.

일반적으로 제한적 거리 핫덱 방법에서는 공여자 파일의 자료 수가 수용자 파일의 자료 수보다 같거나 큰 조건($n_A \leqq n_B$)이 필요하다. 제한적 거리 핫덱 방법은 결합 마이크로 파일의 대체 변수 결과가 공여자 파일의 변수 결과와 동일하게 유지된다는 점이 장점이다. 반면, 매칭을 위한 기준변수 X에 대한 공여자 파일과 수용자 파일의 평균 거리가 비제한적 방법보다 더 커지게 되는 단점이 있다. 이는 제한적 방법은 수용자 파일의 자료와 공여자 파일의 모든 개체를 반드시 한번만 매칭해야 하므로 거리 차이가 큼에도 매칭해야 하는 상황이 존재하기 때문이다.

그러나 공여자 파일의 자료 수가 작은 경우도 가능하며, 이에 대한 예와 거리 함수를 이용한 비제한적 방법과 제한적 방법의 구체적인 매칭 과정에 대한 설명 및 특성은 본 연구보고서에 정리한 비제한적 방법과 제한적 방법을 설명한 부분(3절 이론적 고찰)을 참조하기 바란다.

(3) 혼합 매칭

혼합 매칭(mixed matching)은 두 단계의 과정으로 매칭하는 방법으로 일부는 모수적 방법으로 매크로 매칭을 수행하고 나머지 부분은 비모수적 방법으로 마이크로 매칭을 수행하는 방법이다. 실제로 서로 다른 조사 자료를 연계하는 대부분의 통계적 매칭에서는 혼합 매칭 방법으로 매칭하고 있다.

혼합 매칭 과정을 간단히 설명하면 다음과 같다.

첫 번째 단계에서는 모수 모형의 모수 추정 과정으로 모수적 매크로 매칭 방법에서 설명한 바와 같이 이용하고자 하는 모수 모형을 추정하는 단계이고,

두 번째 단계에서는 매칭 과정으로 핫덱 방법 같은 비모수적 방법 마이크로 매칭 방법으로 결합 마이크로 파일을 생성하는 단계이다.

모수 모형의 모수 추정은 앞서 설명한 바와 같이 모수적 방법 혹은 비모수적 방법을 이용하여 최대우도추정 방법을 모형을 추정하고, 매칭 방법은 수용자 파일의 대체값으로 거리 함수를 이용하여 예측값 혹은 실제 관측값을 연계하거나 제한적 매칭으로 연계 하는 등의 다양한 방법으로 매칭이 가능하므로 혼합 매칭은 모수 추

정과 매칭 과정의 다양한 결합으로 통계적 매칭을 수행할 수 있다. 이에 대한 자세한 설명은 D'Orazio et al.(2006, 2.5절)의 책을 참조하기 바란다.

혼합 매칭은 모수적 모형이 비모수적 모형보다 더 보수적인 점과 비모수 방법이 모형 오류에 대해 더 로버스트한 점을 활용하는 장점이 있다. 또한 비모수 방법의 핫덱 매칭 방법은 예측 값을 대체값으로 제공하는 모수적 마이크로 매칭의 단점을 극복하여 실제 관측 값을 대체값으로 제공해 매칭하는 점을 이용한 매칭 방법이다.

3) 통계적 매칭 결과에 대한 평가 방법

특정 조사 A(수용자 파일)에서 미관측한 자료를 결측 자료로 간주하여 동일 모집단의 다른 조사 B(공여자 파일)에서 측정된 자료와 연계해 수용자 파일의 개체와 유사한 개체들의 관측 자료를 결측 자료에 대한 대체자료로 통합하여 결합 마이크로 자료를 생성하는 통계적 매칭 결과에 대한 검토 및 평가는 다양하게 제안되고 있다.

서로 다른 자료를 매칭하여 결합 마이크로 자료를 생성하는 통계적 매칭에서는 조사되지 않은 미관측된 자료를 MCAR의 결측 매커니즘을 따르는 결측 자료로 간주한다. 그리하여 자료 통합을 위한 관심 변수의 결합 정보를 알 수 없는 관측 자료가 일부 존재하고, 결합 마이크로 자료를 생성하기 위한 모형에 대한 몇 가지 기본적인 가정을 확인할 필요가 있다.

일반적으로 서로 다른 자료을 연계하여 새로운 결합 마이크로 자료를 생성하는 통계적 매칭의 과정과 결과를 검토하기 위해서 다음 네 가지 관점을 확인해야 한다(D'Orazio et al. 2002).

(a) 결합 모형 (X,Y,Z)를 위한 고려해야 하는 가정
(b) 결합확률함수 $f(x,y,z)$의 추정량(매크로 매칭)
(c) 결측 변수를 대체하는 적절한 대체값의 생성 방법(마이크로 매칭)
(d) 결합 마이크로 파일에서의 추론 과정

여기서 (a)와 (b)는 매크로 매칭과 관련된 사항이고, (a), (b), (c)는 마이크로 매칭과 관련되어 있으며, 4가지 모든 사항은 통계적 매칭 결과의 정확성을 살펴볼 수 있는 마이크로 매칭 자료의 분석과 관련된 사항이다.

◦ 모형 가정에 대한 검토

서로 다른 조사 자료로부터 관측되지 않은 자료를 결측 자료로 간주하여 관측된 다른 조사 자료를 연계하여 예측값 혹은 유사 개체의 관측값으로 대체하여 결합 마

이크로 자료를 생성하는 통계적 매칭에 대한 가정의 검토는 서로 다른 조사자료를 통합한 결합 마이크로 자료 (X,Y,Z)가 (Y,Z)에 대한 충분한 정보를 가지지 못하므로 확인하기가 쉽지 않다.

◦ 추정량의 정확성 검토

추정량의 정확성을 검토하는 관점은 매크로 매칭과 관련된 사항이므로 매크로 매칭의 결과가 적절한지 확인하고 평가하는 관점이다.

매크로 매칭은 특정한 모형을 가정하고 결합 자료 (X,Y,Z)에 대한 결합확률함수 혹은 특정한 특성을 추정하기 때문에 기본적으로 결합 자료 (X,Y,Z)의 결합확률함수 모형이나 모형 설정을 위한 기본 모수에 대한 신뢰가 요구된다. 따라서 매크로 매칭에서는 결합확률함수 $f(x,y,z)$에 대한 추정량의 정확성으로 매크로 매칭의 정확성을 평가하게 된다.

추정량의 정확성을 평가하는 측도로는 평균제곱오차(MSE)를 주로 사용한다. 모수적 방법과 비모수적 방법 모두 편의(bias)와 분산의 계산이 가능하므로 대체 자료의 MSE를 계산하여 정확성을 평가하면 된다. 비모수적 방법에서는 추정량의 정확성을 계산하는 대안적인 방법으로 결과의 일치성(consistency)으로 확인이 가능하다. 이를 위해 오분류행렬 혹은 오분류율을 사용해 측정하기도 한다.

특히 모수적 방법에서는 MSE를 최소화하기 위해 최대우도 추정량을 사용하는 것이다.

◦ 결합 마이크로 파일의 대표성 검토

일반적으로 통계적 매칭 과정의 정확성을 평가하는 방법으로 결합 마이크로 파일의 대표성을 널리 사용한다. 이는 마이크로 매칭과 관련된 사항이다.

통계적 매칭에서 결합 마이크로 파일을 생성하는 대부분의 접근 관점은 다음과 같으며, 이는 정확성 평가 과정의 네 가지 관점과도 일치한다(Rassler 2002).

첫째, 결합 자료는 관측하지 않았지만 실측 자료의 참값과 일치한다.

둘째, 모든 변수들의 결합 분포는 통계적으로 결합 파일에 반영되어 있다.

셋째, 변수들의 상관 구조는 유지된다.

넷째, 원시 파일에서 모든 변수들의 주변 분포와 결합 분포는 결합 파일에서도 유지된다.

첫 번째 관점은 이상적인 사항으로 확인하기 어렵지만, 논리적 혹은 수리적인 기준으로 확인 가능하다. 세 번째와 네 번째 관점은 최종적인 결합 마이크로 자료를 이용한 추론이 적절한지 판단해 주지는 못하지만 결합 자료가 어떤 주요한 특성을

가지고 있는지 확인해 줄 수 있다. 두 번째 관점이 만족된다면, 결합 마이크로 자료가 (X,Y,Z)의 결합 분포로부터 생성된 표본으로 간주할 수 있다. 이러한 결합 마이크로 자료는 $f(x,y,z)$의 전형적인 표본이 생각하게 되고, 표본의 특성을 추론하기 위한 일반적인 유의 표본(genernal purposive sample)으로 사용하게 된다.

모형으로부터 생성한 실제 자료와 결합 마이크로 자료의 기본 모형과의 불일치를 매칭 잡음(matching noise)이라고 한다.

이와 같이 결합 마이크로 파일의 대체자료가 적절한지를 평가하는 관점은 매칭 결과가 원시 파일의 성질을 그대로 유지하는 지 살펴보는 것과 같으며, 이를 대표성(representation)이라고 한다. 결합 마이크로 자료에서 원 표본의 고유변수와 평균, 분산 등의 모수가 같은 지를 살펴보거나 변수들의 관계를 측정하는 상관관계, 공분산 및 분포 등을 살펴봄으로써 동일한 속성을 유지하는지 평가할 수 있다.

◦ 결합 마이크로 파일를 활용한 추정량의 정확성 검토

이 관점은 마이크로 매칭 방법의 비판적인 내용이다. 결합 마이크로 파일의 자료가 $f(x,y,z)$의 분포로부터 생성된 표본으로 간주한다면, 결측 자료가 없는 완전 자료를 사용하여 추정하는 것은 적절하다. 하지만 서로 다른 조사 자료를 연계해 생성된 결합 마이크로 자료는 $f(x,y,z)$의 분포로부터 생성된 표본이 아니기 때문에 매칭 잡음을 감소하는 것이 매우 중요한 과정이 된다.

기본적으로 결합 마이크로 자료를 생성한 모형 관점에서 추정량은 불편성 및 일치성과 같은 추론 성질을 잘 유지하고 있다.

결합 마이크로 파일을 활용한 분석을 통해 추정한 결과는 매칭 잡음의 크기가 작아야 추론 성질이 잘 유지된 결과를 제공하게 되므로 매칭 잡음의 크기를 이용해 평가할 수 있다.

4) 가중치 산출 방법

서로 다른 조사 자료들에 대한 조건부 독립성 가정과 보조 정보를 이용한 통계적 매칭으로 생성한 결합 마이크로 파일은 복합 조사 설계에 의해 유한 모집단으로부터 추출된 표본으로 생각할 수 있다.

Renssen(1998)은 표본 A와 B로부터 얻는 추정치를 융합할 목적으로 자료를 통합을 시도하였으며, 이를 위해 유한 모집단으로부터 추출된 표본의 통계적 방법을 사용하였다. Rubin(1986)은 서로 연결해 구성한 단일 표본 $A \cup B$의 구조에 초점을 두고 자료의 통합을 시도하였으며, 두 표본조사를 결합하여 분석할 때 표본가중치의 타당성에 관심을 두었다.

비모수적 접근 방법은 주로 마이크로 매칭에 사용된다. 이 때 조건부 독립성 가

정 하에서 매칭하거나 보조 정보를 이용하여 매칭하기도 하는 데, 이 두 방법의 차이는 서로 다른 표본설계에 의해 유추되는 표본 가중치의 처리 방법이 다르다는 점이다.

본 보고서에서는 결합 마이크로 자료의 가중치에 대해 Renssen(1998)이 제안한 유한 모집단 상황에서의 가중치 산출 방안과 Rubin(1986)이 제안한 두 조사 파일을 연결하여 구성한 파일을 분석할 때 필요한 가중치 산출 방안을 설명하기로 한다.

(1) 매크로 매칭

매크로 매칭에서 사용 가능한 가중치는 범주형 변수들의 결합 분포를 추정하기 위해 사용된다.

매크로 매칭을 위한 가중치를 산출하기 위해 범주형 변수 U와 V의 모집단 분포는 존재하며, V에 대한 진의 모집단 분포(true population distribution)를 알고 있다고 가정한다.

매크로 매칭의 가중치를 부여하는 과정은 다음과 같다.

1단계 : 알려진 모집단 분포에 대한 캘리브레이션 가중치(calibration weight)를 구한다.

2단계 : 표본 A와 B를 이용하여 U에 대한 모집단 분포를 유도한다.

3단계 : 알고 있는 모집단 분포와 추정된 모집단 분포에 대해 조정된 최종 캘리브레이션 가중치를 산출한다.

4단계 : 3단계에서 얻은 가중치를 이용하여 Y와 Z의 결합 분포를 정의한다.

(2) 마이크로 매칭

마이크로 매칭의 가중치는 매크로 매칭과 마차가지로 캘리브레이션 가중치를 사용하거나 거리 함수를 계산하는 데 이용한 변수를 이용하여 가중치를 산출한다.

(3) 파일 연결을 위한 가중치 부여 방안(File Concatenation)

Rubin(1986)은 통계적 매칭을 수행할 때 통계적 분석 논리에 의해 전제 표본 $A \bigcup B$을 고려하는 것이 목적이었다. 서로 다른 두 표본 A와 B로부터 단일 표본 $A \bigcup B$의 구성할 때, 유한 모집단 관점에서 작업은 쉽지 않은 작업이다. 이유는

$A \bigcup B$에 대한 결합 표본 설계는 개별 조사인 A와 B의 표본 설계로부터 유도할 수 있어야 하기 때문이다.

Rubin은 조사 A로부터 가중치 w_a^{AB}와 조사 B로부터 가중치 w_b^{AB}를 새로운 동일한 방법으로 계산하여 두 조사를 연결한 결합 단일 표본 $A \bigcup B$에 대한 가중치를 산출하는 방안을 제안하였다. 새로운 가중치 산출 방안은 연결해 결합한 단일 표본 $A \bigcup B$의 가중치를 계산하기 위해 개별 단위들의 두 표본에서 모두에서 관측될 확률 $\pi_s^{A \cup B} (= \pi_s^A + \pi_s^B)$를 계산하여 각 조사에서의 새로운 가중치를 산출하고 있다.

- 조사 A에서 개별 단위에 대한 새로운 가중치

$$w_a^{AB} = \frac{1}{\pi_a^{A \cup B}} = \frac{1}{1/w_a^A + 1/w_a^B}$$

- 조사 B에서 개별 단위에 대한 새로운 가중치

$$w_b^{AB} = \frac{1}{\pi_b^{A \cup B}} = \frac{1}{1/w_b^A + 1/w_b^B}$$

캘리브레이션과 같은 추가적인 방법을 사용하여 최종 가중치를 산출하는 방안을 제안하였다. Rubin은 제안된 가중치를 다중 대체 방법에 적용하였다.

III. 사례연구 : 2009년 경제활동인구조사와 생활시간조사 자료의 매칭

본 장에서는 II장에서 설명된 여러 매칭 방안들을 2009년 경제활동인구조사 자료와 생활시간조사 자료에 적용하고 그 결과를 살펴보았다. 자료 매칭의 목적이 경제활동인구조사 변수를 활용한 생활시간 조사 자료 분석의 다양성을 확보하는 것에 있음으로 두 자료 중 상대적으로 그 크기가 큰 경제활동인구조사 자료를 제공(donor) 파일로 그리고 생활시간조사 자료를 수용(recipient) 파일로 사용하여 매칭을 수행하였다. 두 자료의 매칭을 위해서는 *R* 프로그램의 *StatMatch* 패키지를 사용하였다. *StatMatch* 패키지는 D'Orazio(2012)에 의하여 작성되었으며 패키지에 대한 보다 세부적인 내용은 D'Orazio, Di Zio 그리고 Scanu (2006)과 D'Orazio(2012)을 참조하면 된다.

가. 매칭을 위해 사용된 자료에 대한 설명

매칭을 위해 고려된 자료는 2009년 경제활동인구조사 자료와 생활시간조사 자료이다. 경제활동인구조사의 경우 만 15세 이상의 성인남녀가 모집단으로 정의되며 생활시간조사의 경우는 이러한 연령 제약이 없는 전국단위 모집단이 분석의 대상이 된다. 연령이외에도 두 조사의 조사 시점 상에 약간의 차이가 있을 수 있으나 본 연구에서는 두 조사의 목표모집단(target population)이 연령 기준을 제외하고 동일한 것으로 간주하였다. 매칭 방안들의 비교가 본 연구의 목적이기 때문에 분석의 편의를 위하여 전체 자료가 아닌 서울지역의 자료만이 분석을 위해 사용되었다. 서울지역의 조사된 개인의 수는 경제활동조사의 경우 109,712명, 생활시간조사의 경우 5,220명으로 생활시간조사의 표본 수가 경제활동인구조사 표본 수의 약 4.7%이다.

서울지역의 자료 중 자료 값의 오류가 있는 것으로 판단되는 두 조사에서의 개체들은 매칭에서 제외되었다. 경제활동인구조사 자료 값의 오류는 가구주와의 관계, 성별, 혼인상태 그리고 직업 변수에서 해당하는 범주 값이 기록되지 않은 오류로서 가구주와의 관계, 성별, 혼인상태 그리고 직업이 0으로 기록된 개인들이다. 생활시간조사의 경우 만 15세 이하의 개인도 조사대상인 관계로 15세 이상만을 조사하는 경제활동인구조사와의 매칭을 위해서 생활시간조사 자료 중 만 15세 미만인 개인은 매칭을 위한 자료에서 삭제하였다. 두 자료의 매칭을 위하여 사용하는 변수의 값이 결측인 경우 논리적으로 타당성을 갖는 범위 내에서 결측을 0으로 처리하였다. 매칭 방안에 따라 결측치의 처리방법이 다르나 고려된 변수 중 교육정도의 경우, 결측치와 관측치 사이의 거리를 다른 값과의 거리와 비교하여 크게 만들어 주는 개념으로 0을 부여하였다.

자료의 조정 후 서울지역 경제활동조사 자료의 크기는 109,232명이며 생활시간조사 자료는 4,712명을 포함하게 되었다. <표 1>은 매칭매개변수의 각 조사 별 분포를 나타내고 있다. 부업시간을 제외한 모든 공통변수에 있어서 두 조사 자료간의 차이는 거의 없는 것으로 파악된다. 통계청 표본과에 문의한 결과 실제로 생활시간조사 표본은 경제활동인구조사 표본의 부분집합이고 표본설계를 위해 사용된 변수 기준으로 두 표본의 분포 차이는 거의 없는 것으로 확인되었다. 따라서 매칭을 위한 매개변수들 중 표본설계 변수와 관계가 있는 변수들은 두 조사에서 거의 유사한 분포를 나타내는 것으로 파악된다.

<표 1> 매칭매개변수의 분포

공통변수		경제활동인구조사	생활시간조사
성	남	50,916(46.61%)	2,212(46.94%)
	여	58,316(53.39%)	2,500(53.06%)
혼인유무	미혼	33,282(30.47%)	1,328(28.18%)
	배우자있음, 사별, 이혼	75,950(69.53%)	3,384(71.82%)
교육정도	무학, 초등학교, 중학교	20,488(18.76%)	834(17.7%)
	고등학교	38,191(34.96%)	1,666(35.36%)
	대학: 4년제 미만, 4년제 이상	44,775(40.99%)	1,948(41.34%)
	대학원 석사과정, 박사과정	5,778(5.29%)	264(5.6%)
연령	평균	42.99	43.06
	표준편차	16.55	16.85
	최솟값	15.00	15.00
	1사분위수	30.00	30.00
	중위수	42.00	42.00
	3사분위수	54.00	54.00
	최댓값	99.00	93.00
주업시간	평균	26.54	26.83
	표준편차	25.71	26.83
	최솟값	0.00	0.00
	1사분위수	0.00	0.00
	중위수	33.00	28.00
	3사분위수	48.00	50.00
	최댓값	109.00	105.00
부업시간	평균	0.04	0.23
	표준편차	0.81	2.16
	최솟값	0.00	0.00
	1사분위수	0.00	0.00
	중위수	0.00	0.00
	3사분위수	0.00	0.00
	최댓값	47.00	48.00

R 프로그램을 사용하기 위해서는 자료의 크기를 조절할 필요가 있다. 이는 R 프로그램의 실행이 램에서 이루어지고 따라서 자료의 크기가 큰 경우 *StatMatch* 패키지의 실행을 할 수 없기 때문이다. *R* 프로그램의 사용에 있어서 자료의 크기는 관측치의 수와 변수의 수에 의하여 결정되나 본 연구에서 고려된 두 자료 모두 변수의 수는 그리 많지 않아 관측치의 수가 자료의 크기를 결정하였다. 고려한 두 자료 중 제공자 파일인 경제활동인구조사 자료의 크기로 인하여 패키지를 사용하는

부분에 제약이 있었다. 여러 경우를 확인한 결과 제공자 파일의 관측치의 수가 20,000개 정도에 이르자 패키지를 실행하는데 문제가 발생하였다. 이러한 문제를 해결하기 위하여 본 연구에서는 매칭매개변수 중 범주형 변수를 이용하여 자료를 분할한 후 각 분할된 자료에서 매칭을 수행하였다. 자료의 분할을 위해서 성별(남/여), 혼인유무(미혼/기타), 교육정도(무학, 초등학교, 중학교/ 고등학교/대학(4년제 미만, 4년제 이상)/대학원 석사, 박사과정 이상)를 고려하여 총 16개의 그룹을 고려하였다. 따라서 그룹을 나누기 위하여 사용된 변수의 값은 같은 그룹 내에 속한 모든 개체에 대하여 동일한 값을 갖는다. 이 후 각 그룹 내에서의 매칭을 위해서는 연령, 주업시간 및 부업시간의 세 변수만이 사용되었다. 이는 매칭을 위한 변수들 중 그룹을 구성하기 위해 사용된 변수들이 상대적으로 다른 변수에 비하여 중요하다고 생각하고 이 값들이 같은 제공자들 중에서 매칭을 위한 최종 개체를 선택한 것과 동일하다.

나. 로지스틱 회귀분석을 통한 매칭매개변수의 설명력 검토

본 예에서는 취업여부 및 실업여부와 같이 경제활동인구조사에서 얻어진 경제활동 관련 변수를 생활시간조사에서 활용하는 것이 목적이며 따라서 매칭매개변수의 경제활동 관련 변수에 대한 설명력을 살펴보는 것이 중요하다. 즉 경제활동 관련 변수가 매칭매개변수에 의하여 충분히 설명되어야만 매칭매개변수를 이용하여 매칭된 자료의 분석이 그 타당성을 얻을 수 있기 때문이며 이와 더불어 이 경우 조건부 독립성 조건이 근사적으로라도 만족되기 때문이다. 본 절에서는 따라서 경제활동인구조사 자료를 이용하여 매칭매개변수들과 실업여부 및 취업여부 관계를 설명하기 위한 로지스틱 회귀분석을 실시하여 매칭매개변수들의 예측력을 평가하였다.

실업률과 매칭매개변수와의 관계를 살펴보기 위하여 경제활동인구조사의 실업여부 변수를 종속변수로 그리고 성, 혼인유무, 교육정도, 연령, 주업시간 및 부업시간을 설명변수로 갖는 로지스틱 회귀분석을 실시하였다. 실업률이 경제활동인구(취업자+실업자) 중 실업자의 비율을 나타내기 때문에 로지스틱 회귀모형의 적합을 위해서는 경제활동인구만을 고려하였다. 또한 경제활동인구조사 자료에 주어진 가중치를 활용한 경우와 적용하지 않은 경우 모두를 살펴보았다. <표 2>는 모형 적합 통계와 모형의 설명력을 나타내고 있다.

<표 2> 실업여부의 로지스틱 회귀모형의 적합 통계

	가중치 미적용		가중치 적용	
Criterion	평균 모형	로지스틱 회귀모형	평균모형	로지스틱 회귀모형
AIC	23,570.093	3,606.658	22,252,392	3,293,839.1
SC	23,579.183	3,688.471	22,252,401	3,293,920.9
−2 Log L	23,568.093	3,588.658	22,252,390	3,293,821.1
R^2	0.2628		1.0000	
$Adj.\ R^2$	0.8699		1.0000	

위의 R^2는 Cox & Snell(1989)이 제안한 일반화 선형 모형에 대한 결정계수로다음과 같다.

$$R^2 = 1 - \left(\frac{L(0)}{L(\hat{\beta})} \right)^{\frac{2}{n}}$$

여기서 $L(0)$는 절편항만 있는 모형, 즉 평균모형(intercept−only model)의 우도함수이고 $L(\hat{\beta})$는 고려한 로지스틱 회귀모형 우도함수이며, n은 표본크기이다. R^2의 최댓값은 $R^2_{max} = 1 - \{L(0)\}^{\frac{2}{n}}$ 로 실제 모형의 설명력을 과소 추정하는 경향이 있는 것으로 알려져 있다. 이를 조정하기 위한 R^2값이 $Adj.\ R^2$ 로 Nagelkerke (1991)이 제안한 조정된 계수(adjusted coefficient)이며 이는 $\tilde{R}^2 = \left[R^2_{\max} \right]^{-1} R^2$ 로 정의된다. 가중치를 적용한 결과를 얻기 위해서는 SAS의 SURVEYLOGISTIC 프로시저를 사용하였고 WEIGHT문만을 적용하였다. 적합결과를 나타내는 통계량들을 볼 때 고려된 매칭매개변수들이 실업여부를 유의하게 설명하고 있는 것을 확인할 수 있다. 각 매칭매개변수의 설명력은 <표 3>의 결과와 같다. 가중치를 적용한 경우 모든 매칭매개변수들이 실업 여부를 설명함에 있어 유의함을 확인할 수 있다.

<표 3> Type III Analysis of Effects.

	가중치 미적용			가중치 적용		
Effect	DF	Wald χ^2	P −value	DF	Wald χ^2	P −value
성별	1	179.7276	<.0001	1	187.2060	<.0001
결혼유무	1	249.0324	<.0001	1	233.0779	<.0001
교육정도	3	162.2417	<.0001	3	176.2861	<.0001
만 나이	1	24.3779	<.0001	1	23.5629	<.0001
주업시간	1	266.8737	<.0001	1	48.8819	<.0001
부업시간	1	0.0007	0.9791	1	15.8428	<.0001

<표 4>와 <표 5>는 취업여부를 종속변수로 그리고 매칭매개변수를 설명변수로 고려한 로지스틱 회귀모형의 적합결과를 나타낸다. 취업률은 만 15세 이상의 인구 중 취업자의 비율로 정의되기 때문에 취업여부의 로지스틱 회귀분석을 위해서는 모든 경제활동인구조사 자료가 사용되었다. 취업여부를 종속변수로 정의한 로지스틱 회귀모형에서도 모든 매칭변수가 높은 설명력을 갖는 것을 확인하였다. 즉 본 예에서 고려하고 있는 매칭매개변수들은 자료 매칭 후 경제활동인구조사 자료 중 주 관심이 되는 경제활동여부 관련변수와 밀접한 관계를 나타내고 있고 따라서 이 매칭매개변수를 이용한 매칭결과 자료를 활용한 분석들은 적절하리라 여겨진다.

<표 4> 취업여부의 로지스틱 회귀모형의 적합 통계

	가중치 미적용		가중치 적용	
Criterion	평균모형	로지스틱 회귀모형	평균모형	로지스틱 회귀모형
AIC	149,052.8	11796.73	135,239,435	10,593,804
SC	149,062.4	11883.14	135,239,444	10,593,890
−2 Log L	149,050.76	11778.73	135,239,433	10,593,786
R^2		0.7154		1.0000
$Adj.\ R^2$		0.9609		1.0000

<표 5> Type III Analysis of Effects.

	가중치 미적용			가중치 적용		
Effect	DF	Wald χ^2	P −value	DF	Wald χ^2	P −value
성별	1	150.7944	<.0001	1	126.6315	<.0001
결혼유무	1	333.4071	<.0001	1	259.8938	<.0001
교육정도	3	401.1424	<.0001	3	420.7656	<.0001
만 나이	1	163.7736	<.0001	1	169.5724	<.0001
주업시간	1	3338.182	<.0001	1	825.8757	<.0001
부업시간	1	0.0043	0.9476	1	115.0552	<.0001

다. 매칭 방안

매칭 방안으로는 크게 최근접 이웃 핫덱(Nearest neighbor distance hot deck)방안과 랜덤 핫덱(random hot deck)방안을 고려하였다. 최근접 이웃 핫덱 방안은 주어진 거리함수를 바탕으로 수여자와 가장 가까운 거리에 있는 제공자 정보를 매칭하는 방법이며 랜덤 핫덱(random hot deck)은 주어진 거리함수를 바탕으로 주어진 규칙에 따라 수여자와 가까운 거리에 있는 제공자의 그룹을 정의하고 그로부터 매칭을 위한 제공자를 랜덤하게 추출하는 방안이다. 이 이외에도 모형에 근거한 매칭 방안들이 있으나 본 연구에서는 위의 두 비모수적 방법을 통한 매칭방안만을 비교를 위해 고려하였다. 각 방안 별 설명은 다음과 같다.

1) 최근접 이웃 핫덱 방안과 랜덤핫덱 방안에 사용되는 거리함수

공통변수 중 실제 매칭을 위해 사용되는 변수인 매칭매개변수를 $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{in})$ 로 정의하자. StatMatch 패키지에서 사용할 수 있는 거리함수 $d(i,j)$ 중 다음의 함수들을 본 연구에서는 고려하였다.

• 맨하튼(Manhattan) 거리 함수

$$d(i,j) = \sum_{k=1}^{n} |x_{ik} - x_{jk}|$$

• 유클리디안(Euclidean) 거리 함수

$$d(i,j) = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2}$$

• 마할라노비스(Mahalanobis) 거리 함수

$$d(i,j) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{S}^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_j)}.$$

여기서 $\boldsymbol{S}$ 는 분산 공분산 행렬의 추정량을 나타낸다.

• 정확(Exact) 거리 함수

$$d(i,j) = \sum_{k=1}^{n} D_{(ij)k},$$

$$D_{(ij)k} = \begin{cases} 1, & x_{ik} = x_{jk} \\ 0, & x_{ik} \neq x_{jk} \end{cases},$$

• 비유사성 지수를 이용한 Gower 거리 함수

$$\text{dissimilarity } d(i,j) = \frac{\sum_{k=1}^{n} \delta_{ijk} d_{ijk}}{\sum_{k=1}^{n} \delta_{ijk}}.$$

여기서 δ_{ijk}는

$$\delta_{ijk} = \begin{cases} 0, & \text{if } x_{ik} = Missing \text{ or } x_{jk} = Missing, \\ 0, & \text{if } x \text{ is binary variable and } x_{ik} = x_{jk} = 0 \text{ or } x_{ik} \neq x_{jk}, \\ 1, & otherwise. \end{cases}$$

이며 d_{ijk}는 다음 표의 정의와 같다.

범주형 변수(nominal scale):	$d_{ijk} = \begin{cases} 0, & \text{if } x_{ik} = x_{jk}, \\ 1, & otherwise. \end{cases}$
연속형 변수(interval scale):	$d_{ijk} = \frac{\lvert x_{ik} - x_{jk} \rvert}{R_k}$, R_k는 k번째 변수의 범위
순서형 변수(ordinal scale):	$z_{ik} = \frac{(r_{ik} - 1)}{max(r_{ik}) - 1}$, r_{ik}는 k번째 변수기준 관측치 i의 상대적 위치. 새로운 변수 z_{ik}를 연속형으로 처리하여 거리를 정의.

• Minimax 거리함수

$$d(i,j) = max(|x_{i1} - x_{j1}|, |x_{i2} - x_{j2}|, \cdots, |x_{in} - x_{jn}|)$$

맨하튼 거리함수와 유클리디안 거리함수는 각 매칭 매개변수의 변량을 고려하지 않은 거리함수로 각 변수의 분산이 서로 상이하게 차이가 나는 경우 그 측정단위 즉 관측 값이 큰 변수가 거리의 결정에 큰 역할을 하게 되는 문제가 발생할 수 있다. 이에 반해 마할라노비스 거리함수는 각 변수의 변량이 서로 다를 때에도 분산-공분산 행렬을 거리의 정의를 위해 사용함으로 각 변수의 측정단위에 의존하지 않는 장점이 있다. 정확 거리함수는 모든 공통변수를 범주형 변수로 간주하고 두 개체의 값이 동일하지 않을 때 1의 일종의 벌칙을 받는 거리함수이다. 매칭 매개변수가 범주형 변수인 경우 정확 거리함수 이외의 거리함수의 사용에는 제약이 따른다. 비유사성 지수를 이용한 거리함수는 변수의 측정 스케일에 따라 적절한 거리함수를 자동적으로 적용하기 때문에 매칭 매개변수가 연속형 변수와 범주형 변수 모두를 포함하고 있을 때 사용할 수 있는 거리함수이다. Minimax 거리함수는 각 변수 별 두 관측치 간의 거리의 최댓값을 거리함수로 사용하고 있다.

정확 거리함수 혹은 비유사성 거리함수를 사용하지 않는 경우에 범주형 변수가 순서형인 경우 각 범주의 순위를 이용하거나 혹은 각 범주 간 차이 값을 정의하고 이를 사용하여 거리함수를 적용하는 방안을 고려할 수 있을 것이다. 그러나 명목형 변수가 매칭을 위해 사용되는 경우 변수가 갖는 값이 그 크기를 나타내는 것이 아니므로 위의 거리함수를 직접적으로 사용하기에는 무리가 있다. 이 경우 가능한 방안 중 하나는 각 명목형 범주별로 제공자 파일과 수여자 파일을 분할 한 후 각 분

할된 범주 안에서 연속형 혹은 순서형 변수를 이용하여 매칭을 시행하는 것이다.

본 연구에서는 R 프로그램이 수행할 수 있는 자료의 크기가 한정되어 있고 매칭 매개변수로 사용되는 범주형 변수가 매칭에 있어 매우 중요한 변수임을 감안하여 명목형 매개변수의 범주 별로 경제활동인구조사 파일과 생활시간조사 파일을 분할한 후 각 분할된 파일 별로 매칭을 수행하였다.

2) 제약조건을 고려한 방법(Constrained method)

고려된 거리함수들을 이용하여 아래의 목적함수(object function)가 정의되고 이 목적함수를 최소화하는 쌍 (i,j)을 매칭하게 된다.

$$\sum_{i=1}^{n_A}\sum_{j=1}^{n_B}[d(i,j),\ w_{ij}]$$

여기서 $w_{ij}\in\{0,1\}$의 값을 갖는 이항변수이다. 동일한 제공자가 여러 수여자에게 매칭되는 것을 방지하기 위하여 이항변수 w_{ij}에 대한 제약조건이 부여될 수 있다. 제공자 파일과 수여자 파일의 크기에 따라 다음의 두 경우의 제약조건을 고려할 수 있다. 이 두 제약조건은 특별히 제공자 파일의 크기가 수여자 파일보다 그 크기가 매우 크지 않은 경우 즉 한 제공자가 여러 수여자와 매칭이 될 가능성이 높을 때 사용할 수 있는 방법이다. 본 연구에서는 제공자 파일의 크기가 매우 큰 관계로 아래의 제약 조건은 사용하지 않았다.

i) $n_A = n_B$

$$\sum_{j=1}^{n_B} w_{ij} = 1\ ,\ \sum_{i=1}^{n_A} w_{ij} = 1\ ,\ i=1,\cdots,n_A,\ j=1,\cdots,n_B$$

$$w_{ij} = \begin{cases} 1\ , if\ the\ pair\ (i,j)\ is\ matched, \\ 0\ , if\ not. \end{cases}$$

ii) $n_A < n_B$

$$\sum_{j=1}^{n_B} w_{ij} = 1\ ,\ \sum_{i=1}^{n_A} w_{ij} \le 1\ ,\ i=1,\cdots,n_A,\ j=1,\cdots,n_B$$

$$w_{ij} = \begin{cases} 1\ , if\ the\ pair\ (i,j)\ is\ matched, \\ 0\ , if\ not. \end{cases}$$

3) 랜덤 핫덱의 그룹을 형성하는 방안

랜덤 핫덱 방안은 수여자와 가까운 거리에 있다고 판단되는 그룹을 구성하고 그 그룹 내에서 하나의 제공자를 무작위로 추출하는 방법으로 그룹을 구성하는 방안이 결과에 큰 영향을 미치게 된다. 아래는 R의 StatMatch 패키지에서 사용할 수 있는 그룹 생성 옵션을 설명하고 있다. rot은 그룹의 크기를 제공자의 제곱근의 수에 1을 더한 값으로 정의한다. exact은 그룹의 크기를 직접, 그리고 span은 비율을 통해 정하고 있으며 k.dist는 수여자와 각 제공자의 거리를 이용하여 그룹의 크기를 결정한다. 본 연구에서는 span 옵션과 exact 옵션을 고려하였다.

- rot : 그룹의 크기 $=\sqrt{n_D}+1$, n_D 는 가능한 제공자의 수
- span : 그룹의 크기 $= n_D \times k$, $0 < k \leq 1$
- exact : 그룹의 크기 $= k$, $0 < k \leq n_D$
- k.dist : 그룹의 크기 = 거리가 k 미만인 제공자의 수

라. 매칭 결과

본 연구에서는 위에서 살펴본 매칭방안들의 여러 옵션들을 고려하여 총 9가지의 매칭방안을 수행하였다. 수행된 9가지 방안들의 정의는 <표 6>와 같다. 랜덤 핫덱 방안 사용 시, 대략적으로 그룹의 크기를 5로 정하였고 따라서 Exact 옵션을 사용할 시 k=5로 지정하였으며, span 옵션을 위해서는 사용된 비율 k의 범위는 대략적으로 0.03%~0.88%이다. 매칭 결과는 크게 두 부분에 걸쳐 평가하였다. 그 첫 번째는 매칭이 이루어진 생활시간조사 결과를 이용한 경제활동 관련 변수의 분포와 원 경제활동인구조사 자료의 분포를 비교하는 것이다. 단순히 몇 변수들의 분포 비교만으로 여러 방안들의 직접적인 평가는 어려우나 매칭매개변수들의 분포가 두 자료에서 유사한 점을 고려하여 매칭된 결과와 경제활동인구조사 결과와는 유사하게 나타나는 것이 바람직하다고 판단된다. 두 번째로는 경제활동조사 뿐 아니라 생활시간조사에서도 관측은 되었으나 매칭을 위해 사용되지 않은 변수들의 값을 비교함으로서 실제 관측된 값과 매칭된 값과의 일치 정도를 검토하려 한다. 주 조사에서 모두 관측오차가 없거나 무시할 정도의 수준이라며 이 두 결과가 유사하게 나타나야 할 것이다.

<표 6> 고려된 9가지 매칭 방안

ID	이름	설명
1	NND_man	최근접 이웃 핫덱, 맨하튼 거리 함수
2	NND_maha	최근접 이웃 핫덱, 마할라노비스 거리 함수
3	NND_gower	최근접 이웃 핫덱, 비유사성 지수
4	RND_man_E	랜덤 핫덱, 맨하튼 거리 함수, Exact 옵션
5	RND_maha_E	랜덤 핫덱, 마할라노비스 거리 함수, Exact 옵션
6	RND_gower_E	랜덤 핫덱, 비유사성 지수, Exact 옵션
7	RND_man_S	랜덤 핫덱, 맨하튼 거리 함수, Span 옵션
8	RND_maha_S	랜덤 핫덱, 마할라노비스 거리 함수, Span 옵션
9	RND_gower_S	랜덤 핫덱, 비유사성 지수, Span 옵션

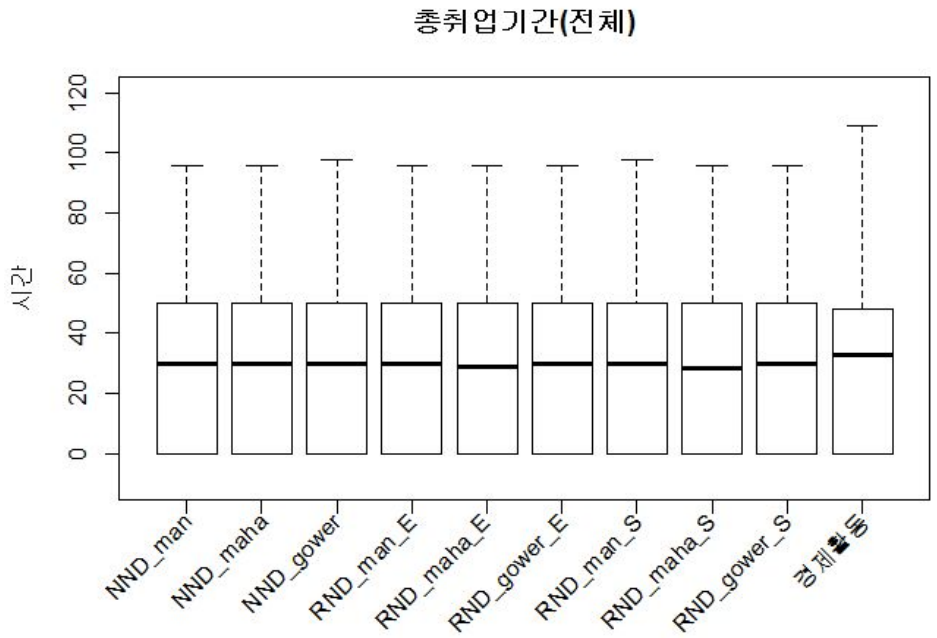
<표 7>~<표 9> 그리고 <그림 1>~<그림 3>은 매칭이 이루어진 생활시간조사 자료를 이용하여 경제활동인구조사로부터 조사된 주요 변수들의 분포를 나타내고 있다. 각 변수 별 분포들은 제공자 파일인 경제활동인구조사 자료의 분포와 비교할 수 있다. 실제 두 파일간의 공통변수 기준 차이가 거의 나타나지 않기 때문에 각 방법 별 비교 혹은 경제활동인구조사 결과와의 비교는 각 방안 별 차이를 반영한다고 판단할 수 있다.

고려된 변수들 중 총 취업기간에 경우, 고려된 모든 9가지 방안들이 비슷한 결과를 제공하고 있고 그 결과 또한 경제활동인구조사 결과와 유사함을 확인할 수 있다. 구직활동기간을 살펴보면 취업기간과는 달리 랜덤 핫덱을 사용할 경우 방안 별 차이가 최근접 이웃 핫덱의 방안 별 차이보다 더 크게 나타남을 알 수 있다. 이러한 차이는 랜덤 핫덱의 방안 별 차이일 수도 있으나 랜덤 핫덱이 가지고 있는 무작위성 성질에 기인할 수도 있다. 즉 이 변수에 대해서는 수여자와 가까운 거리에 있다고 판단되는 제공자들 사이에 상당한 차이가 있음을 말해 주는 것이라고 판단된다.

실업 관련 통계를 살펴보면 실업자 수에 있어서는 최근접 이웃 핫덱의 방안이 그리고 실업률에 있어서는 랜덤 핫덱의 방안이 경제활동인구조사 결과와 보다 유사한 결과를 제공하고 있다. 또한 두 경우 모두 가능한 방안들 사이의 차이는 구직활동기간에 비하여 적게 나타남을 확인할 수 있다.

위의 결과들을 볼 때 최근접 이웃 방안은 랜덤 추출이 실행되는 핫덱 랜덤 그룹 내의 변수들의 변동이 적을 때는 그 결과가 랜덤 핫덱 결과와 비슷하며 그렇지 않을 경우 즉 그룹 내 관심변수의 변동이 클 경우에는 그 차이가 크게 나타남을 알 수 있다. 매칭된 자료를 이용한 분석 결과의 편향이 커지는 위험을 최소화하기 위해서는 랜덤 핫덱 방안을 그리고 분석 결과의 변동을 줄이기 위해서는 최근접 이웃 방안을 사용하는 것이 적절하리라 판단된다. 그러나 랜덤 핫덱이 적용되는 경우에는 단 한 번의 랜덤추출 대신 여러 번의 추출을 통해 그 분산을 예측하며 여러 매칭 값들의 평균을 추정량으로 사용함으로 분산을 줄이고 가능한 오차를 줄일 수 있을 것으로 기대된다.

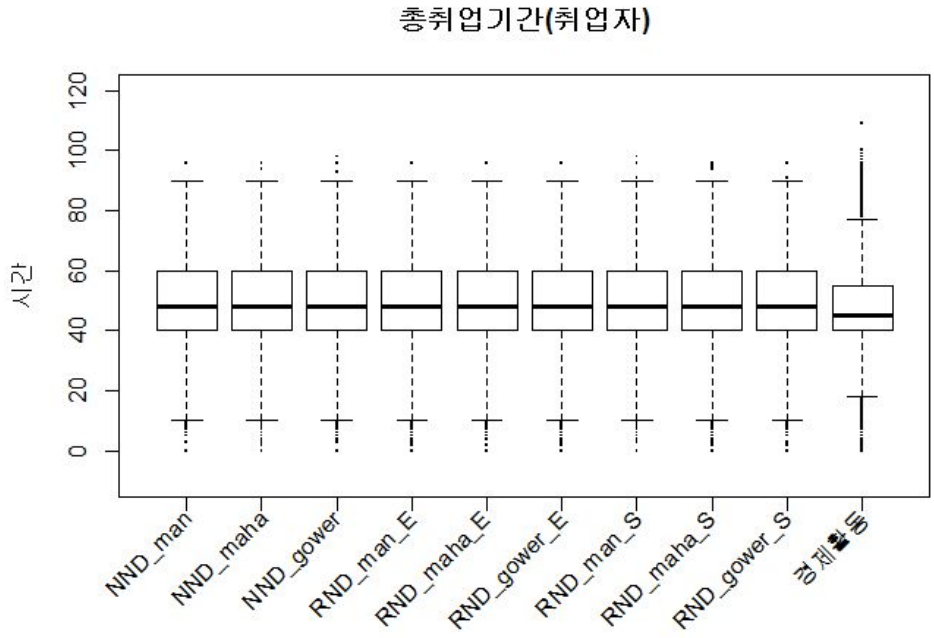
<그림 1.> 매칭결과 비교 : 총취업기간(전체)



<표 7.> 매칭결과 비교 : 총취업기간(전체)

	mean	sd	min	Q1	median	Q3	max
NND_man	26.91	26.87	0.00	0.00	30.00	50.00	96.00
NND_maha	27.04	26.90	0.00	0.00	30.00	50.00	96.00
NND_gower	27.04	26.90	0.00	0.00	30.00	50.00	98.00
RND_man_E	26.87	26.83	0.00	0.00	30.00	50.00	96.00
RND_maha_E	26.98	26.82	0.00	0.00	29.00	50.00	96.00
RND_gower_E	27.01	26.87	0.00	0.00	30.00	50.00	96.00
RND_man_S	26.87	26.82	0.00	0.00	30.00	50.00	98.00
RND_maha_S	26.99	26.85	0.00	0.00	28.50	50.00	96.00
RND_gower_S	27.02	26.87	0.00	0.00	30.00	50.00	96.00
경제활동	26.58	25.74	0.00	0.00	33.00	48.00	109.00

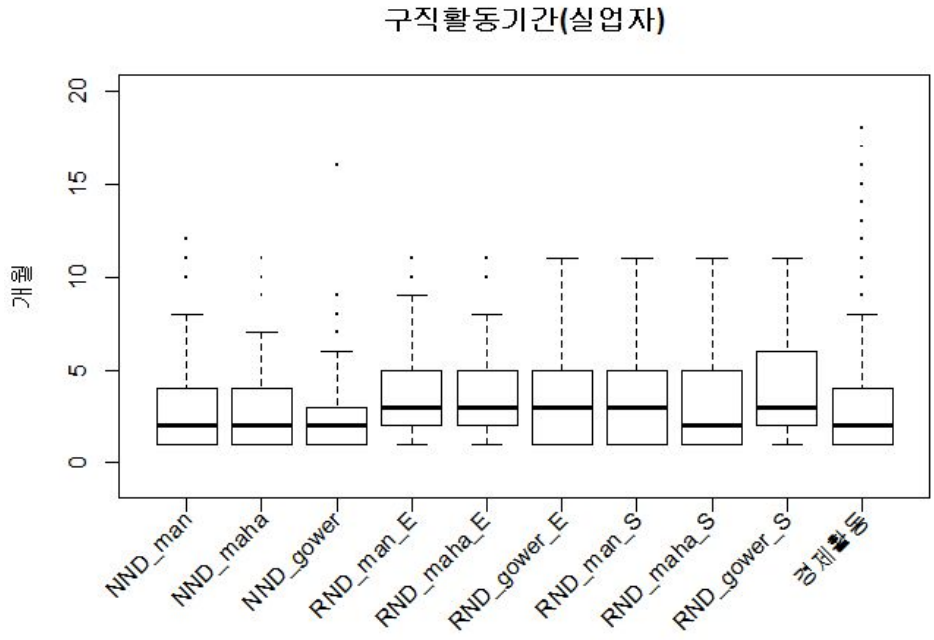
<그림 2.> 매칭결과 비교 : 총취업기간(취업자)



<표 8> 매칭결과 비교 : 총취업기간(취업자)

	mean	sd	min	Q1	median	Q3	max
NND_man	47.37	17.44	0.00	40.00	48.00	60.00	96.00
NND_maha	47.77	17.02	0.00	40.00	48.00	60.00	96.00
NND_gower	47.69	17.12	0.00	40.00	48.00	60.00	98.00
RND_man_E	46.82	17.94	0.00	40.00	48.00	60.00	96.00
RND_maha_E	46.81	18.03	0.00	40.00	48.00	60.00	96.00
RND_gower_E	47.01	17.85	0.00	40.00	48.00	60.00	96.00
RND_man_S	46.74	18.03	0.00	40.00	48.00	60.00	98.00
RND_maha_S	47.23	17.56	0.00	40.00	48.00	60.00	96.00
RND_gower_S	46.96	17.94	0.00	40.00	48.00	60.00	96.00
경제활동	46.30	15.57	0.00	40.00	45.00	55.00	109.00

<그림 3.> 매칭결과 비교 : 구직활동(실업자)



<표 9> 매칭결과 비교 : 구직활동(실업자)

	mean	sd	min	Q1	median	Q3	max
NND_man	2.91	2.59	1.00	1.00	2.00	4.00	12.00
NND_maha	2.74	2.12	1.00	1.00	2.00	4.00	11.00
NND_gower	2.83	2.75	1.00	1.00	2.00	3.00	16.00
RND_man_E	3.79	2.64	1.00	2.00	3.00	5.00	11.00
RND_maha_E	3.88	2.76	1.00	2.00	3.00	5.00	11.00
RND_gower_E	3.48	2.74	1.00	1.00	2.00	5.00	11.00
RND_man_S	3.39	2.60	1.00	1.00	3.00	5.00	11.00
RND_maha_S	3.11	2.43	1.00	1.00	2.00	5.00	11.00
RND_gower_S	3.82	2.83	1.00	2.00	3.00	5.75	11.00
경제활동	3.15	2.59	1.00	1.00	2.00	4.00	18.00

<표 9.> 매칭결과 비교 : 취업자, 실업자, 비경활자

	취업자	실업자	비경활자	만15세이상인구	실업률	취업률	경제활동참가율
NND_man	2,674 (56.75%)	96 (2.04%)	1,942 (41.21%)	4,712	3.47%	56.75%	58.79%
NND_maha	2,663 (56.52%)	97 (2.06%)	1,952 (41.43%)	4,712	3.51%	56.52%	58.57%
NND_gower	2,667 (56.60%)	97 (2.06%)	1,948 (41.34%)	4,712	3.51%	56.60%	58.66%
RND_man_E	2,701 (57.32%)	70 (1.49%)	1,941 (41.19%)	4,712	2.53%	57.32%	58.81%
RND_maha_E	2,716 (57.64%)	65 (1.38%)	1,931 (40.98%)	4,712	2.34%	57.64%	59.02%
RND_gower_E	2,704 (57.39%)	80 (1.70%)	1,928 (40.92%)	4,712	2.87%	57.39%	59.08%
RND_man_S	2,705 (57.41%)	65 (1.38%)	1,942 (41.21%)	4,712	2.35%	57.41%	58.79%
RND_maha_S	2,688 (57.05%)	73 (1.55%)	1,951 (41.40%)	4,712	2.64%	57.05%	58.60%
RND_gower_S	2,706 (57.43%)	62 (1.32%)	1,944 (41.26%)	4,712	2.24%	57.43%	58.74%
경제활동	62,658 (57.36%)	2,870 (2.63%)	43,704 (40.01%)	109,232	4.38%	57.36%	59.99%

경제활동인구조사와 생활시간조사 모두에서 관측되었으나 매칭에서 고려되지 않은 변수들을 통해 매칭 방안들의 결과를 비교하였다. 분석을 위해 고려된 변수들은 “지난 1 주일간 수입을 목적으로 일을 했는가?”와 “종사상 지위”이다. <표 10>과 <표 11>은 이 두 변수를 이용하여 9가지의 가능한 결과들을 비교한 것이다. <표 10>과 <표 11>의 생활시간에 대응하는 행은 생활시간조사 결과의 단순집계이며 생활시간(가구원 가중치)는 생활시간조사에 부여된 가중치를 활용하여 추정된 결과이다. 이 두 행은 생활시간조사 자료를 그대로 사용하였기 때문에 9가지 방안사이에 동일한 값을 갖는다. 경제활동인구조사의 “지난 1 주일간 수입을 목적으로 일을 했는가?”에 대한 답변 중 결측이 존재하며 확인 결과 해당 개체의 주업시간이 모두 1시간 이상이기 때문에 생활시간조사의 “일을 했음”에 해당하는 것으로 간주하였다. 매칭 후 추정으로 표기 된 행은 매칭 후에 생활시간조사 가중치를 이용하여 추정된 결과를 나타낸다. 마지막으로 일치 여부로 표기된 행은 두 조사에서 나온 각 개인별 결과들이 일치하는지를 평가하는 것이다. 두 조사 모두 해당 변수에 대한 관측오차가 없거나 무시할 정도로 매우 적다면 이 일치 비율은 9가지 방안의 비교에 직접적으로 사용될 수 있을 것이다.

고려된 두 변수 모두에서 9가지 방안들의 차이는 거의 존재하지 않고 있다. 다만 두 변수 모두에서 비유사성 지수를 활용한 최근접 이웃 핫덱 방안이 조금 더 나은 일치비율을 보이고 있으나 이 결과를 다른 경우로 확대 해석하기는 어렵다고 판단된다. 비유사성 지수를 활용한 방안은 매칭변수에 형태에 따라서 근사성을 정의하며 또한 연속형 변수에 대해서는 각 변수의 변이를 고려하였음으로 특별히 매칭을 위한 변수의 형태가 다양하고 각 변수들의 변동 혹은 분산이 매우 다른 경우 효율적인 방안이 될 것으로 파악된다. 다만 자료의 수가 큰 경우 알고리즘이 수행에 시간이 걸리는 단점이 있다.

<표 10.> 매칭결과 비교 : 지난 1 주일간 1시간 이상 노동 여부

		NND_man	NND_maha	NND_gower	RND_man_E	RND_maha_E	RND_gower_E	RND_man_S	RND_maha_S	RND_gower_S
생활시간	0									
	일을 했음	2,650 (56.24%)	2,650 (56.24%)	2,650 (56.24%)	2,650 (56.24%)	2,650 (56.24%)	2,650 (56.24%)	2,650 (56.24%)	2,650 (56.24%)	2,650 (56.24%)
	일을 하지 않았음	2,062 (43.76%)	2,062 (43.76%)	2,062 (43.76%)	2,062 (43.76%)	2,062 (43.76%)	2,062 (43.76%)	2,062 (43.76%)	2,062 (43.76%)	2,062 (43.76%)
생활시간 (가구원 가중치)	0									
	일을 했음	4,959,064 (58.71%)	4,959,064 (58.71%)	4,959,064 (58.71%)	4,959,064 (58.71%)	4,959,064 (58.71%)	4,959,064 (58.71%)	4,959,064 (58.71%)	4,959,064 (58.71%)	4,959,064 (58.71%)
	일을 하지 않았음	3,487,302 (41.29%)	3,487,302 (41.29%)	3,487,302 (41.29%)	3,487,302 (41.29%)	3,487,302 (41.29%)	3,487,302 (41.29%)	3,487,302 (41.29%)	3,487,302 (41.29%)	3,487,302 (41.29%)
매칭 후, 추정	0	2,500 (53.06%)	2,514 (53.35%)	2,510 (53.27%)	2,502 (53.10%)	2,528 (53.65%)	2,520 (53.48%)	2,509 (53.25%)	2,514 (53.35%)	2,511 (53.29%)
	일을 했음	138 (2.93%)	132 (2.80%)	136 (2.89%)	136 (2.89%)	121 (2.57%)	121 (2.57%)	129 (2.74%)	132 (2.80%)	132 (2.80%)
	일을 하지 않았음	2,074 (44.02%)	2,066 (43.85%)	2,066 (43.85%)	2,074 (44.02%)	2,063 (43.78%)	2,071 (43.95%)	2,074 (44.02%)	2,066 (43.85%)	2,069 (43.91%)
매칭 후, 추정 (가구원 가중치)	0	4,682,329 (55.44%)	4,706,133 (55.72%)	4,696,596 (55.6%)	4,664,881 (55.23%)	4,731,436 (56.02%)	4,711,336 (55.78%)	4,686,172 (55.48%)	4,701,371 (55.66%)	4,697,645 (55.62%)
	일을 했음	250,919 (2.97%)	244,526 (2.90%)	254,063 (3.01%)	268,367 (3.18%)	226,118 (2.68%)	229,318 (2.71%)	247,076 (2.93%)	251,617 (2.98%)	246,789 (2.92%)
	일을 하지 않았음	3,513,118 (41.59%)	3,495,706 (41.39%)	3,495,706 (41.39%)	3,513,118 (41.59%)	3,488,812 (41.31%)	3,505,712 (41.51%)	3,513,118 (41.59%)	3,493,379 (41.36%)	3,501,932 (41.46%)
일치여부	일치	4700 (99.75%)	4704 (99.83%)	4704 (99.83%)	4700 (99.75%)	4701 (99.77%)	4703 (99.81%)	4700 (99.75%)	4700 (99.75%)	4705 (99.85%)
	불일치	12 (0.25%)	8 (0.17%)	8 (0.17%)	12 (0.25%)	11 (0.23%)	9 (0.19%)	12 (0.25%)	12 (0.25%)	7 (0.15%)

<표 11.> 매칭결과 비교 : 종사상 지위

		NND_man	NND_maha	NND_gower	RND_man_E	RND_maha_E	RND_gower_E	RND_man_S	RND_maha_S	RND_gower_S
생활시간	무직자	2,062 (43.76%)	2,062 (43.76%)	2,062 (43.76%)	2,062 (43.76%)	2,062 (43.76%)	2,062 (43.76%)	2,062 (43.76%)	2,062 (43.76%)	2,062 (43.76%)
	임금근로자	1,982 (42.06%)	1,982 (42.06%)	1,982 (42.06%)	1,982 (42.06%)	1,982 (42.06%)	1,982 (42.06%)	1,982 (42.06%)	1,982 (42.06%)	1,982 (42.06%)
	고용주	162 (3.44%)	162 (3.44%)	162 (3.44%)	162 (3.44%)	162 (3.44%)	162 (3.44%)	162 (3.44%)	162 (3.44%)	162 (3.44%)
	자영자	424 (9.00%)	424 (9.00%)	424 (9.00%)	424 (9.00%)	424 (9.00%)	424 (9.00%)	424 (9.00%)	424 (9.00%)	424 (9.00%)
	무급가족종사자	82 (1.74%)	82 (1.74%)	82 (1.74%)	82 (1.74%)	82 (1.74%)	82 (1.74%)	82 (1.74%)	82 (1.74%)	82 (1.74%)
생활시간 (가구원 가중치)	무직자	3,487,302 (41.29%)	3,487,302 (41.29%)	3,487,302 (41.29%)	3,487,302 (41.29%)	3,487,302 (41.29%)	3,487,302 (41.29%)	3,487,302 (41.29%)	3,487,302 (41.29%)	3,487,302 (41.29%)
	임금근로자	3,780,834 (44.76%)	3,780,834 (44.76%)	3,780,834 (44.76%)	3,780,834 (44.76%)	3,780,834 (44.76%)	3,780,834 (44.76%)	3,780,834 (44.76%)	3,780,834 (44.76%)	3,780,834 (44.76%)
	고용주	290,330 (3.44%)	290,330 (3.44%)	290,330 (3.44%)	290,330 (3.44%)	290,330 (3.44%)	290,330 (3.44%)	290,330 (3.44%)	290,330 (3.44%)	290,330 (3.44%)
	자영자	749,086 (8.87%)	749,086 (8.87%)	749,086 (8.87%)	749,086 (8.87%)	749,086 (8.87%)	749,086 (8.87%)	749,086 (8.87%)	749,086 (8.87%)	749,086 (8.87%)
	무급가족종사자	138,814 (1.64%)	138,814 (1.64%)	138,814 (1.64%)	138,814 (1.64%)	138,814 (1.64%)	138,814 (1.64%)	138,814 (1.64%)	138,814 (1.64%)	138,814 (1.64%)
매칭 후, 추정	무직자	1,708 (36.25%)	1,683 (35.72%)	1,715 (36.40%)	1,673 (35.51%)	1,656 (35.14%)	1,649 (35.00%)	1,647 (34.95%)	1,669 (35.42%)	1,665 (35.34%)
	임금근로자	2,271 (48.20%)	2,271 (48.20%)	2,259 (47.94%)	2,260 (47.96%)	2,243 (47.60%)	2,223 (47.18%)	2,242 (47.58%)	2,242 (47.58%)	2,264 (48.05%)
	고용주	203 (4.31%)	212 (4.50%)	207 (4.39%)	244 (5.18%)	256 (5.43%)	256 (5.43%)	274 (5.81%)	257 (5.45%)	242 (5.14%)
	자영자	415 (8.81%)	430 (9.13%)	429 (9.10%)	407 (8.64%)	441 (9.36%)	455 (9.66%)	434 (9.21%)	437 (9.27%)	424 (9.00%)
	무급가족종사자	115 (2.44%)	116 (2.46%)	102 (2.16%)	128 (2.72%)	116 (2.46%)	129 (2.74%)	115 (2.44%)	107 (2.27%)	117 (2.48%)

		NND_man	NND_maha	NND_gower	RND_man_E	RND_maha_E	RND_gower_E	RND_man_S	RND_maha_S	RND_gower_S
매칭 후, 추정 (가구원 가중치)	무직자	2,852,890 (33.78%)	2,805,426 (33.21%)	2,851,977 (33.77%)	2,825,987 (33.46%)	2,776,661 (32.87%)	2,773,981 (32.84%)	2,752,819 (32.59%)	2,799,689 (33.15%)	2,804,689 (33.21%)
	임금근로자	4,272,859 (50.59%)	4,261,240 (50.45%)	4,241,863 (50.22%)	4,226,443 (50.04%)	4,200,554 (49.73%)	4,162,873 (49.29%)	4,229,991 (50.08%)	4,225,969 (50.03%)	4,240,505 (50.21%)
	고용주	380,623 (4.51%)	417,556 (4.94%)	376,825 (4.46%)	437,110 (5.18%)	468,876 (5.55%)	490,310 (5.80%)	502,875 (5.95%)	469,559 (5.56%)	452,387 (5.36%)
	자영자	737,547 (8.73%)	766,390 (9.07%)	791,903 (9.38%)	736,303 (8.72%)	798,081 (9.45%)	805,745 (9.54%)	762,449 (9.03%)	775,543 (9.18%)	750,377 (8.88%)
	무급가족종사자	202,447 (2.40%)	195,754 (2.32%)	183,799 (2.18%)	220,523 (2.61%)	202,194 (2.39%)	213,458 (2.53%)	198,232 (2.35%)	175,606 (2.08%)	198,408 (2.35%)
일치여부	일치	3,375 (71.63%)	3,316 (70.37%)	3,376 (71.65%)	3,318 (70.42%)	3,274 (69.48%)	3,268 (69.35%)	3,289 (69.80%)	3,294 (69.91%)	3,302 (70.08%)
	불일치	1,337 (28.37%)	1,396 (29.63%)	1,336 (28.35%)	1,394 (29.58%)	1,438 (30.52%)	1,444 (30.65%)	1,423 (30.20%)	1,418 (30.09%)	1,410 (29.92%)

마. 토의 사항

두 자료의 매칭은 또 다른 관점에서는 무응답 대체(imputation)로 간주할 수 있다. 다만 무응답 대체의 경우 동일 서베이에서 응답이 이루어진 개체의 항목을 대체하는 것과는 달리 매칭 문제에서는 다른 조사에서 가장 유사한 개체를 선택하여 그 값을 대체하게 된다. 따라서 통계적으로 최적 매칭 방안을 찾는 것은 최적의 무응답 방안을 찾는 것과 거의 유사한 과정으로 이해할 수 있다. 이의 관점에서 볼 때 자료 매칭 시 주의해야 하는 부분은 매칭 후 수용자 파일에 추가된 변수에 대하여 극단치가 발생할 수 있다는 것이다. 특별히 랜덤 매칭이 이루어진 경우 무작위 선택을 위해 생성된 제공자 그룹 내 추가변수의 변동이 심할 경우 자료 매칭 후 이를 살펴보고 이에 대한 적절한 조정, 예를 들어 극단값을 나타내는 그룹에 한하여 평균 대체를 적용하는 등의 방법을 고려해야 할 것이다. 그러나 이러한 극단값의 발생에 따른 문제 해결은 자료 매칭의 특별한 통계적 기법에 의존한 처리보다는 자료에 대한 이해가 높은 해당 분야 전문가의 의견에 기초하여 수정을 해야 할 것으로 판단된다.

본 장에서는 경제활동인구조사 자료와 생활시간조사 자료를 이용하여 총 9가지의 매칭방안을 검토하였다. 9가지 매칭방안은 랜덤여부와 거리함수의 구별을 통해 생성되었다. 고려된 예에서는 이 9가지 방안 사이에 큰 차이가 없는 것으로 판단되며 고려된 두 자료의 매칭을 위해서는 비유사성 지수를 활용한 최근접 이웃 핫덱 방안이 다른 방안 보다 조금 나은 결과를 보여주고 있다. 그러나 이러한 결과는 본 예에 국한된 것이며 또 다른 자료의 매칭을 위해서는 매칭의 목적 그리고 이용 가능한 매칭매개변수 등을 고려하여 여러 가능한 매칭 방안 중 최적의 방안을 선택해야 할 것이다.

자료매칭은 일종의 자료 증대(data augmentation)로 생각할 수 있다. 이 때 추가되는 자료는 개체수가 아니라 변수의 수이다. 다른 자료 증대의 경우와 마찬가지로 자료 매칭에 있어서 역시 많은 주의가 필요하다. 이는 자료 매칭 후 자료는 관측된 실제 값이 아닌 비록 주어진 기준에 따라 부여되었지만 여전히 임의의 랜덤 프로세스를 통해 부여된 값이기 때문이다. 따라서 사용자에게 최소한 매칭 혹은 대체가 이루어진 값 혹은 변수 여부를 나타내는 또 다른 변수를 제공하여야 하고 매칭을 위해 사용된 알고리즘이 제공되어야 할 것이다.

본 연구에서 사용한 StatMatch 패키지의 사용에 있어서는 R 프로그램이 가지고 있는 제약으로 인하여 큰 자료의 자료 매칭에 어려움이 있다. 연구자의 제한적인 지식으로는 전문적으로 매칭을 수행하기 위한 패키지가 제공되는 다른 통계 분석 프로그램은 없는 것으로 알고 있다. 대형 자료의 매칭을 일괄적으로 처리하기 위해서는 매칭을 위한 기존의 알

고리즘을 이용하여 C와 SAS같은 언어를 이용하여 직접 프로그램을 작성해야 할 것으로 생각된다. 다만 자료 매칭을 위한 매개변수 중 특별히 중요한 범주형 변수가 있는 경우, 이 변수를 이용하여 1차적인 매칭 즉 자료 분할을 실시한 후 각 분할된 자료에 StatMatch 패키지를 적용하는 방법을 고려할 수 있을 것이다.

Reference

고은애 (2004), 통계적 매칭을 이용한 데이터 통합에 관한 연구, 석사학위논문, 동국대학교 대학원.

안일호(2003), 혼합형 데이터의 통계적 결합에 관한 연구, 석사학위논문, 고려대학교 대학원.

정성석, 김순영, 김현진 (2004), 데이터 보강을 위한 데이터 통합기법에 관한 연구, *응용통계연구,* 제 17권 3호, 605-617.

Anderson, T.W. (1984), *An Introduction to Multivariate Statistical Analysis*, 2nd Edition, New York: Wiley.

Barr, R.S. and Turner, J.S. (1978), A New, Linear Programming Approach to Microdata File Merging, *1978 Compendium of Tax Research*, U.S. Department of the Treasury, 131-149.

Budd, E.C. (1971), The Creation of a Microdata File for Estimating the Size Distribution of Income. Review of Income and Wealth, **17**, 317-333.

Cox, D.R. and Snell, E.J. (1989), *The Analysis of Binary Data*, 2nd Edition, London: Chapman & Hall.

D'Orazio, M., Di Zio, M. and Scanu, M. (2002), Statistical matching and official statistics, *Rivista di Statistica Ufficiale*, **2002/1**, 5-24.

D'Orazio, M., Di Zio, M. and Scanu, M. (2006), *Statistical Matching Theory and Practice*, Wiley.

Goel, P.K. and Ramalingam, T. (1989), *The Matching Methodology: Some Statistical Properties*, New York: Springer-Verlag.

Ingram, D., O'Hare, J., Scheuren, F. and Turek, J. (2000), Statistical matching: a new validation case study, *Proceedings of the survey Research Methods Section, American Statistical Association.*

Kamakura, W.A. and Wedel, M. (1997), Statistical Data Fusion for Cross-Tabulation. *Journal of Marketing Research*, **34**, 485-498.

Kalton, G. and Kasprzyk, D. (1986), The Treatment of Missing Survey data. Survey Methodology, **12**, 1-16.

Little, R.J.A. and Rubin, D.B. (2002), *Statistical Analysis with Missing Data*, 2nd Edition. Hoboken, NJ: Wiley.

Moriarity, C. (2009), Statistical Properties of Statistical Matching : Data Fusion Algorithm, VDM Verlag Dr, Muller.

Nagelkerke, N.J.D. (1991), A Note on a General Definition of the Coefficient of Determination, *Biometrika*, 78, 691-692

National Research Council (1992), *Combining Information: Statistical Issues and Opportunities for Research*. Washington, D.C.: National Academy Press.

National Statistics (2003), National Statistics code of Practice-Protocol on Data Matching, London:TSO.

Okner, B.A. (1972), Constructing a new data base from existing microdata sets: the 1966 merge file, *Annals of Economic and Social Measurement*, 1(3), 325-342.

Paass, G. (1985), Statistical Record Linkage Methodology: State of the Art and Future Prospects. *Invited Papers to Joint ISI-IASS Meetings*, 1, International Statistical Institute, Meetings, 9, 33-48.

Rässler, S. (2002), *Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*. New York: Springer-Verlag.

Rässler, S. (2004). Data fusion: identification problem, validity, and multiple imputation. *Austrian Journal of Statistics* 33(1-2), 153-171

Rodgers, W.L. (1984), An Evaluation of Statistical Matching. Journal of Business and Economic Statistics, 2, 91-102.

Renssen, R.H. (1998), Use of statistical matching techniques in calibration estimation, *Survey Methodology*, 24, 171-183.

Rubin, D.B. (1974), Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American Statistical Association*, 69, 467-474.

Rubin, D.B. (1986), Statistical Matching using file concatenation with adjusted weights and multiple imputations, *Journal of Business and Economic Statistics*, 4, 87-94.

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, London:Chapman & Hall.

Singh, A.C., Mantel, H., Kinack, M. and Rowe, G. (1990), On methods of

statistical matching with and without auxiliary information, Technical Report SSMD-90-016E, Methodology Branch, Statistics Canada.

Singh, A.C., Mantel, H.J., Kinack, M.D., and Rowe, G. (1993), Statistical Matching : Use of auxiliary information as an alternative to the conditional independence assumption, *Survey Methodology*, 19, 59-79.

U.S. Department of Commerce, (1980). Report on exact and statistical matching techniques, Statistical Policy Working Paper 5. Washington, DC: Federal Committee on Statistical Methodology.

Van Der Putten, Peter , Kok, Joost N., and Gupta, Amar (2002), Why the information explosion can be bad for data mining, and how data fusion provides a way out, *Second SIAM International Conference on Data Mining*, Arlington, April, 11-13.

Van Der Putten, Peter , Kok, Joost N., and Gupta, Amar (2002). Data Fusion through Statistical Matching, *Technical Paper 185, Center for eBusiness@MIT,* MITSloan (ebusiness.mit.edu)

van Pelt, X. (2001), The Fusion Factory: A Constrained Data Fusion Approach. Master of Science. Thesis, Leiden Institute of Advanced Computer Science, The Netherlands.

Wand, M. and Jones, C. (1995), *Kernel Smoothing*, London: Chapman & Hall.

연구과제 : 다양한 출처 자료 처리 및 통계 생산방안 연구

세부과제 **2** : 조사모드의 효과적 활용 및 추정방법의 향후과제

2013. 12.

연구기관 : 한국조사연구학회

차 례

표 차 례

그 림 차 례

1. 서론

조사환경의 악화로 인해 응답률은 점차 낮아지고 있으며 조사비용도 크게 증가하고 있다. 이를 극복하기 위해 최근 들어 혼합모드(mixed mode)를 조사에 채택하는 사례가 증가하고 있다. 통계청에서도 2000년 전후로부터 경제활동인구조사, 인구주택총조사, 사교육비조사 등의 많은 조사에서 전통적 조사모드인 면접원조사와 유치조사는 물론 전화조사와 인터넷조사 등을 혼합하여 채택하는 혼합모드조사를 실시하고 있다. 2010 인구주택총조사에서는 인터넷 조사의 적극적 도입으로 인건비, 인쇄 및 자료입력비 등을 포함하여 약 204억원의 예산을 절감하였다(임경은·박라나, 2013).

하지만, 혼합모드를 사용하게 되면 모드특성에 따른 차이로 인해 자료의 비교성이 떨어질 수 있게 된다. 이는 조사모드에 따라 접근할 수 있는 대상, 응답계층, 응답값 등이 달라질 수 있기 때문이다. 따라서, 조사수행에 있어서 혼합모드방식을 선택할 때에는 이로 인한 조사오류와 비용간의 균형에 대한 명확한 이해와 평가가 수반되어야 할 것이다.

본 연구는 혼합모드와 관련한 조사방법론 및 추정방법론에 대해 살펴보고, 통계청의 혼합모드조사에 대한 조사설계 및 현황에 대한 평가를 통해 보다 나은 혼합모드의 활용가능성을 제공할 수 있는 표본조사에 대한 실험설계 측면을 제시하고자 한다. 2절에서는 기존문헌에서 다루고 있는 조사모드에 대한 개략적 측면을 살펴본다. 3절에서는 무응답 정의 및 무응답 축소를 위한 접근들에 대해 살펴본다. 4절에서는 혼합모드 사용에 따른 모드효과를 측정하기 위한 기존 연구들의 접근방식과 방법론들을 논의하고자 한다. 5절에서는 기존 연구들이 제시하는 접근방법이 통계청의 가구조사 자료에 적용될 수 있는지에 대한 가능성과 한계점에 대해서 논의하고자 한다.

2. 조사모드 비교

2.1 조사모드

표본조사에는 전통적으로 면접조사, 우편조사, 전화조사를 많이 사용하고 있다. 최근 들어 컴퓨터와 인터넷 기술의 발달로 인해 기존 조사방식에 전산화 기술을 접목하거나 인터넷을 활용한 조사방법의 사용도 늘어나고 있는 추세이다. 본 절에서는 흔히 사용되는 대표적인 조사모드들에 대해서 간단히 살펴보고 장단점을 논의한다.

면접조사(face-to-face survey)란 조사원(interviewer)이 조사대상자를 직접 방문하여 설문지(questionnaire)를 함께 작성하는 방법을 일컫는다. 면접조사에서는 조사원이 조사대상자와 같은 장소에 있게 되므로 조사에 대한 참여 설득이 매우 용이하다. 또한 조사 진행 중에 설문문항들에 대한 추가적 정보제공이 가능하게 된다. 따라서 조사대상자의 설문문항에 대한 이해를 도울 수 있어 항목무응답(item non- response)을 주게 되므로 조사품질 향상에 매우 유리한 조사방법이라 할 수 있다. 이러한 점은 전화조사나 우편조사에서는 구현하기 힘든 면접조사가 갖는 장점이라 할 것이다. 따라서 다른 조사들에 비해 상대적으로 높은 응답률을 얻을 수 있다(Hox and de Leeuw, 1994). 하지만 조사대상지역이 여러 곳으로 분산되어 있는 경우에는 방문에 따른 노력과 비용이 전화조사나 우편조사에 비해 훨씬 많은 소요된다.

전화조사(telephone survey)란 조사원이 조사대상자에게 전화를 걸어 조사를 진행하는 자료수집방법이다. 면접조사와 비슷하게 조사원이 자료수집에 참여하여 조사대상자의 조사 참여를 설득할 수 있고 질문에 대한 추가적인 설명과 아울러 적절한 답을 찾는데 도움을 제공할 수도 있어서 항목무응답을 줄여줄 수 있다. 따라서 비교적 높은 응답률을 얻을 수 있다. 전화번호와 함께 해당 주소가 주어진다면 이를 이용하여 조사 전에 안내문 발송(advance letter)을 통해 조사의 실시계획과 필요성을 알림으로 선제적으로 조사 참여의 협조를 구할 수 있다. 하지만 휴대전화 보급 및 전화사용 패턴의 변화로 인해 기존 전화번호부의 포함률이 급격히 낮아지고 최근의 과도한 마케팅조사 및 사생활 보호의식의 강화로 인한 낯선 전화번호에 대한 수신거부가 증가하고 있어 전화조사의 효율성이 점차 떨어지고 있는 추세이

다.

우편조사(mail survey)란 조사대상자에게 우편으로 설문지를 발송하고 동봉한 반송용 봉투를 이용하여 응답을 받아내는 자기기입식(self-administration)의 조사방법을 말한다. 우편조사는 조사자와 조사대상자가 직접 상면하지 않기 때문에 조사 전에 안내우편을 발송하거나 설문지와 함께 안내장을 동봉하여 조사협조를 구하는 것이 좋다. 조사대상자가 조사 내용을 쉽게 이해할 수 있도록 설문지 디자인은 물론 용어나 표현을 명확하고 단순하게 해야 한다. 적은 비용과 노력으로 광범위한 지역과 대상을 조사할 수 있고 접근하기 쉽지 않던 대상도 포함할 수 있으며 조사의 익명성으로 인해 조사대상자가 충분한 시간적 여유를 갖고 답변할 수 있는 가능성도 있다. 반면 회수율이 낮고 회수시간이 많이 걸리며 응답자에 대한 통제가 불가능하고 응답자의 이해능력에 따라 자료의 질이 달라질 수도 있다.

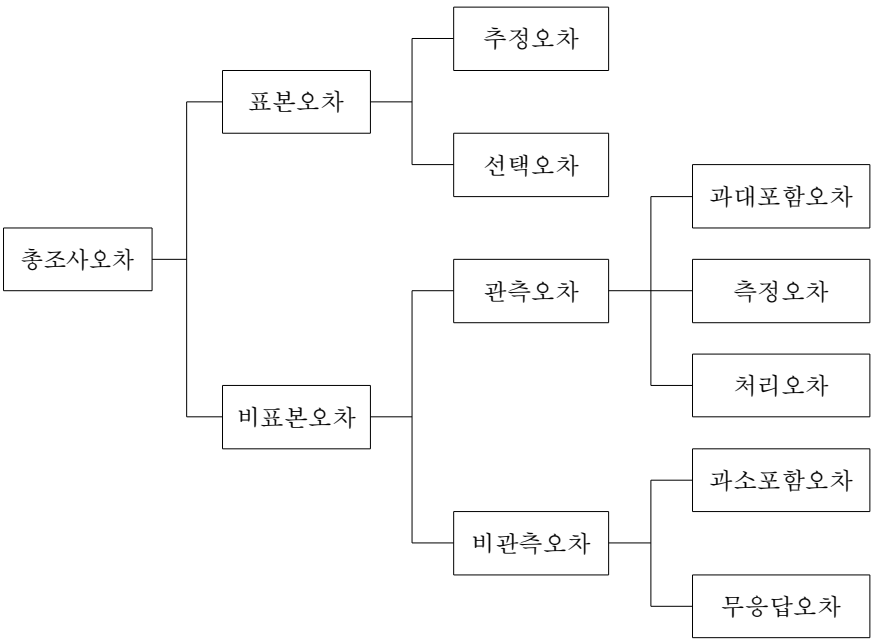
유치조사(leaving method) 혹은 배포조사란 조사원이 조사대상을 방문하거나 우송의 방법으로 조사표를 조사대상자에게 전달한 다음, 수일 내에 회수하는 조사방법을 말한다. 조사대상자가 부재중이라도 다시 방문해야 하는 수고를 덜 수는 있지만, 조사대상자에게 조사내용을 통일적으로 이해시키기 어렵고, 조사대상 이외의 사람이 기입하거나, 타인의 의견이나 참고서를 참고하여 회답하는 등의 우려가 있다. 하지만 이 방법은 우편조사에 비해 조사원이 적극적으로 조사를 관리·감독할 수 있다.

인터넷조사(internet survey)는 조사대상자로 하여금 특정 웹사이트를 통해 제공되는 설문지에 자기기입방식으로 진행하는 사이버 공간에서 이루어지는 조사를 일컫는다. 웹조사(web survey), 온라인조사(on-line survey), 컴퓨터를 이용한 웹조사(computer-assisted web interviewing, CAWI)이라고도 칭한다. 이는 최근에 급격히 발달되어 보급된 조사방식으로 초기 프로그램 개발비와 서버 유지비만을 필요로 하는 저비용 조사인데, 다양한 영상기술을 통해 조사대상자의 응답을 도울 수 있다는 장점을 갖는다. 반면, 조사모집단의 구성원 중 인터넷 사용이 불가능한 특정계층이 존재할 수 있으므로 대표성을 갖는 표본설계가 불가능할 수도 있다.

2.2 조사오차

조사추정에는 다양한 형태의 조사오류(survey error)가 따르게 된다. 조사오류란 궁극적으로 조사추정치(survey estimate)와 모수(parameter)간의 차이를 말한다. 조사품질의 관점에서 이러한 차이를 총조사오차(total survey error)라 칭하는데, 그 원인들은 다양한 방식으로 구분할 수 있다. Bethlehem *et al.* (2011, 7-9쪽)의 분류에 따르면 총조사오차는 먼저 표본오차(sampling error)와 비표본오차(non-sampling error)로 나뉜다(<그림 1>).

<그림 1> 총조사오차 분해도



[출처: Bethlehem *et al.* (2009, 7-8쪽)]

표본오차란 모수추정이 모집단 전체의 자료가 아닌 일부 표본만으로 이루어짐에 따라 발생하는 오차를 말한다. 표본이 무작위로 선택되었다면 표본오차는 추정오차(estimation error)로 분류될 수 있고, 이는 표본설계를 통해 오차 규모를 조정될 수 있다. 반면 부정확한 선택확률이 추정절차에 적용된다면 표본오차는 선택오차(selection error)로 구분할 수 있다. 정확하지 않은 선택확률의 예로는 표본개체가 표본추출틀(sampling frame)에 하나 이상 존재하는 경우를 들 수 있다.

비표본오차는 표본오차 이외의 모든 오차를 일컫는데, 이는 모집단내 모든 개체를 조사하더라도 발생할 수 있는 오차로 이는 다시 관측오차(observation error)와 비관측오차(nonobservation error)로 구분될 수 있다. 관측오차란 조사를 통해 답변을 얻고 기록하는 과정에서 발생한다. 이는 조사포함개체가 목표모집단(target population)에 속하지 않기 때문에 발생하는 과대포함오차(over-coverage error)와 설문의 참값과 조사 처리되고 기록된 값의 차이에 의해 발생하는 측정오차(measurement error), 그리고 자료처리과정에서 발생하는 처리오차(processing error)로 나뉠 수 있다. 이때 측정오차에는 응답자가 질문을 잘못 이해하였거나 바른 답변을 하지 않은 경우는 물론 조사자가 응답자의 답변을 잘못 기록할 때 발생할 수도 있다. 비관측오차란 조사에서 의도한대로 측정되지 않았을 때 발생하게 된다. 목표모집단의 개체가 표본추출틀에 포함되지 않는 과소포함오차(under-coverage error)와 표본개체가 요구된 조사값을 제공하지 않았을 때 발생하는 무응답오차(nonresponse error)로 나뉠 수 있다.

2.3 조사응답절차와 측정오차

조사응답절차에 대한 인지모형(cognitive models)을 살펴보면 조사모드가 측정에 어떠한 영향을 주는지 이해하는데 많은 도움을 줄 수 있다. Groves *et al.* (2009, 7장2절)는 조사대상자에게 질문이 주어질 때 발생할 수 있는 심리작용(mental process)에 관해 다음의 네 가지 절차를 포함한 조사응답절차의 모형을 논하고 있는데, 이를 도식화하면 <그림 2>와 같이 표현될 수 있다.

[조사응답 인지모형]
절차 1: 응답자가 질문을 해석하는 “이해(comprehension)” 절차
절차 2: 질문에 응답하기 위해 필요한 정보를 상기하는 “검색(retrieval)” 절차
절차 3: 상기한 정보들을 취합하고 정리하는 “판단(judgement)” 절차
절차 4: 답변을 주어진 형식의 틀에 맞추는 “보고(reporting)” 절차

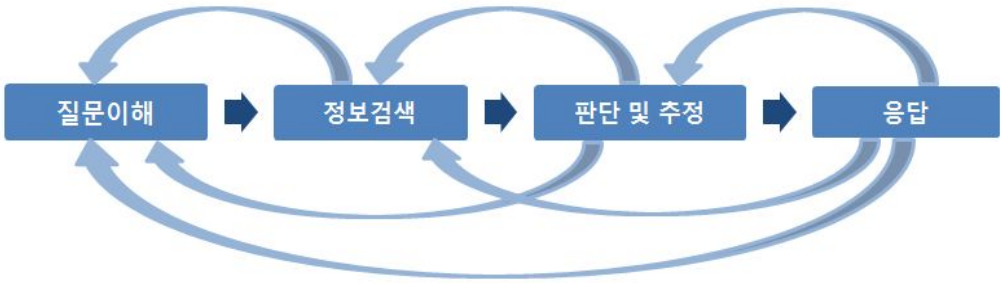
한편 Cannell *et al.* (1981)은 조사응답절차와 관련해서 조사대상자가 두 종류

의 노선(tracks)을 취할 수 있음을 가정한다. 한 노선 하에서는 앞서 기술한 절차에 따라 정확하거나 적어도 적절한 응답을 제공하려 한다. 다른 노선 하에서는 조사를 빨리 끝내기 위해 절차를 따르기 보다는 손쉬운 방식을 취하거나 정확한 답변을 하지 않고자 한다. 예를 들면, 조사환경에 따라 감지되는 프라이버시, 조사 타당성, 면접원과의 친밀성 등에 의해 정확하고 솔직한 답변보다는 바람직한 답변을 하고자 한다. 이를 묵인효과(acquiescence effect) 혹은 소망효과(social desirability effects)라 칭할 수 있다. 과제 난이도, 응답자 능력, 동기부여 등이 상호작용을 일으켜 조사대상자의 조사항목에 대한 (응답을 위한 노력의 정도를 정하는) 응답방식이 결정될 수 있다. 이때 적당히 조사에 응답하려는 만족효과(satisficing effect)와 좀 더 응답을 정확히 하려는 헌신효과(sacrficing effects)의 두 가지 측면을 고려할 수도 있다.

그 외에도 제시된 정보가 나중에 들어온 정보보다 기억에 훨씬 더 큰 영향을 미치는 초두효과(primacy effect)나 자유회상에서 목록의 끝 부분에 있는 항목들이 목록 중간에 있는 항목들보다 잘 회상되는 현상(recency effect) 등과 같은 심리적 현상에 대한 모형 등도 있다.

de Leeuw(2005)는 앞서 논의한 것과 유사한 자극·반응모형(stimulus-response model)을 통해 인지 및 반응에 영향을 주는 특성으로 청각·시각적 프리젠테이션 형태, 자기기입 혹은 조사원 (기입 혹은) 관여식 조사운영형태, (컴퓨터를 활용한) 역동적 미디어에 의한 설문 혹은 (종이를 이용한) 수동적 설문방식 등으로 분류하고 있다.

<그림 2> 조사응답과정에 대한 인지모형



[출처: Groves *et al.* (2009, p. 218)]

3. 무응답

3.1 무응답 이해

3.1.1 무응답 개요

무응답(nonresponse)은 추출된 표본개체가 조사에서 요구하는 값을 제공하지 않음으로 생기는데, 이는 대부분의 표본조사에서는 발생하게 된다. 무응답에는 표본개체가 조사에 전혀 응하지 않는 단위무응답(unit nonresponse)과 일부 항목만 응답하는 항목무응답(item nonresponse)으로 나뉠 수 있다. 반면, 과소포함(under-coverage)은 조사에 대한 적격대상(eligible)이지만 표본추출틀(sampling frame)에 포함되지 않아 표본으로 선택될 수 없어 해당 개체에 대한 정보를 얻을 수 없게 될 때 발생한다. 따라서 개체에 대한 정보부재의 측면에서는 무응답과 동일하지만 주의해서 구분되어야 한다. 과소포함 하의 조사결과에 대한 해석은 목표모집단 전체를 대상으로 하기보다는 표본추출틀, 다시 말해 추출틀모집단(frame population)을 대상으로 제한하여야 할 것이다.

단위무응답 발생요인은 일반적으로 크게 두 가지로 구분할 수 있다. 먼저, 표본개체에 대한 접촉실패 혹은 접근불가이다. 예로, 조사원이 방문하거나 혹은 전화를 걸었을 때 표본가구원이 가구에 없거나, 조사기간의 제약으로 인해 더 이상 조사시도를 할 수 없을 때, 혹은 주소나 전화번호 정보가 부정확하거나 오래되어 쓸모없게 되는 등의 이유이다. 두 번째로는, 표본개체에 대한 조사 설득의 실패이다. 예로, 조사대상자가 응답할 의사가 없거나, 주택출입관리자가 조사원의 접근을 막거나, 전화조사에서 최초 수신자가 표본가구원을 바꾸어주지 않는 경우 등이다. 이러한 단위무응답 발생요인에 대한 구분에 따라 무응답 효과나 처리방식이 달라질 수 있기 때문이다(Brick, 2013).

이 외에서 Platek(1977), Groves (1989) 등은 무응답을 발생시키는 요인들을 좀 더 상세히 구분하기도 한다. <표 1>은 Groves *et al.* (2009)가 제시하는 단위무응답률을 줄일 수 있는 도구목록을 정리하고 있다.

<표 1> 단위무응답 축소를 위한 조사단계별 도구목록

단계	도구목록
접촉 (contact)	• 방문횟수와 시기 (number and timing of calls) • 자료수집기간 (length of data collection period) • 조사원 업무부담 (interviewer workload) • 조사원 관측 (interviewer observations)
초기결정 (initial decision)	• 조사원 행동 (interviewer behavior) • 후원 (sponsorship) • 사전공지 (pre-notification) • 사례 (incentives) • 부담 (burden) • 응답규칙 (respondent rule) • 가구/면접원 매치 (householder/interviewer match)
최종결정 (final decision)	• 이중추출기법의 사용 (two-phase sampling) • 면접원교체 (interviewer switch) • 조사모드교체 (mode switch) • 조사설득편지 (persuation letters) • 사후조정 (postsurvey adjustment)

[출처: Groves *et al.* (2009, p. 190)]

3.1.2 무응답 모형

무응답이 모수추정에 미치는 영향을 살펴보기 위해서는 무응답 현상을 표본이론에 접목하여 고려하는 것이 필요할 것이다. 무응답 현상은 결정적(deterministic)이거나 혹은 확률적(stochastic)인 것으로 생각될 수 있다(Bethlehem *et al.*, 2011, Valliant *et al.*, 2013).

고정응답모형(fixed response model, FRM)에서는 개별 개체의 표본선택에 따른 응답여부가 이미 정해져 있다고 가정한다. 즉, 전체 표본은 응답개체 s_R과 무응답개체 s_{NR}으로 나눠어 질 수 있다. 만약, 설계가중치(design weight)가 $d_{0i}=1/\pi_i$으로 주어지고 모평균 $\overline{Y}$를 다음의 Hajek 평균추정량 $\hat{\overline{y}}_\pi$으로 추정한다고 가정하자.

$$\hat{\bar{y}}_{\pi} = \frac{\sum_{i \in s_R} d_{0i} y_i}{\sum_{i \in s_R} d_{0i}}$$

고정응답모형 하에서 평균추정량 $\hat{\bar{y}}_{\pi}$의 편향(bias)은 다음과 같이 정의될 수 있다.

$$bias_{FRM}(\hat{\bar{y}}_{\pi}) = \frac{N_{NR}}{N}\left(\overline{Y}_R - \overline{Y}_{NR}\right) \quad (1)$$

여기서 $N = N_R + N_{NR}$은 모집단 크기, N_R와 N_{NR}은 각각 모집단내 총 응답 및 무응답 개체 수, $\overline{Y}_R$와 $\overline{Y}_{NR}$은 각각 응답 및 무응답 개체의 모평균을 나타낸다. 다시 말해, 무응답 현상이 고정응답모형을 따른다면, 무응답으로 인한 (응답자) 평균추정량 $\hat{\bar{y}}_{\pi}$의 편향은 (i) 응답자와 무응답자의 평균 간의 차이가 크거나, (ii) 무응답률 (N_{NR}/N)이 클수록 커질 수 있음을 나타낸다.

반면, 확률응답모형(random response model, RRM)은 각 개체별로 영이 아닌 응답확률을 갖게 되는데, 이에 따라 조사요청 시 협조여부가 결정된다고 가정한다. I_i는 개체 i의 표본(포함)지시자를 나타내고, R_i는 표본개체 i의 응답지시자를 각각 나타낸다고 하자. 그러면, 포함확률(inclusion probability)은 $\pi_i = \Pr(I_i = 1)$이고 표본포함에 따른 개체 i의 응답확률(response probability)은 다음과 같이 정의될 수 있다.

$$\phi_i = \Pr(R_i = 1 | I_i = 1) \quad (2)$$

Rosenbaum and Rubin (1983)은 ϕ_i를 성향점수(propensity score)라고 칭하였다. 확률응답모형을 이용한 무응답 편향연구에서는 표본추출과 응답여부가 모두 확률적으로 결정된다는 "준확률(quasi-randomization)" 가정을 한다. 이러한 가정 하에서 Hajek 평균추정량 $\hat{\bar{y}}_{\pi}$의 편향은 다음과 같이 유도된다(Kalton and Maligalig, 1991).

$$bias_{RRM}(\hat{\bar{y}}_{\pi}) \doteq \frac{1}{\bar{\phi}N}\sum_{i\in s}\left(y_i - \bar{Y}_U\right)\left(\phi_i - \bar{\phi}\right) \tag{3}$$

여기서 $\bar{\phi}= N^{-1}\Sigma_U \phi_i$는 모집단 개체들의 평균응답확률을 나타낸다. 다시 말해, 무응답 편향은 조사변수 y_i와 응답확률 ϕ_i의 공분산과 평균응답확률 $\bar{\phi}$의 역수에 비례한다. 따라서 무응답현상이 확률응답모형을 따른다면 y_i와 ϕ_i 간에 상관관계가 없을때 무응답 편차도 없고 평균추정량 $\hat{\bar{y}}_{\pi}$은 모평균의 불편추정량이다.

Bethlehem *et al.* (2011)는 앞서의 확률응답모형 하에서 응답자 단순평균 $\bar{y}_R= \Sigma_{s_R} y_i / n_R$의 (근사적) 편향을 다음과 같이 제시하고 있다.

$$bias_{RRM}(\bar{y}_R) \doteq \tilde{Y} - \bar{Y} \doteq \frac{1}{\bar{\phi}}\rho_{\phi y} S_{\phi} S_y$$

여기서 $\tilde{Y}= N^{-1}\Sigma_U(\phi_i/\bar{\phi})y_i$, $E(\bar{y}_R) \approx \tilde{Y}$, $\rho_{\phi y}$는 응답확률 ϕ_i와 조사변수 y_i 간의 상관계수, S_{ϕ}와 S_y는 각각 응답확률과 조사변수의 표준편차를 나타낸다. 따라서 응답확률과 조사변수 간에 상관관계가 없다면 편향도 없으며, 반대로 상관관계가 높거나 평균응답확률이 작다면 편향도 매우 커짐을 나타낸다.

무응답 현상으로 인한 추정에 미치는 영향을 좀 더 잘 이해하기 위해서 Little and Rubin(2002)이 제시한 3가지의 무응답 매카니즘(nonresponse mechanism)을 고려할 수 있다. 관련한 논의는 앞서의 확률변수 I_i와 R_i에 더불어 조사변수 Y_i에 대한 확률모형을 추가적으로 고려한다. 이때 X_i는 표본개체 i에 대해 알 수 있는 보조변수(벡터)이고 표본설계에 사용된 정보도 포함한다고 가정하자.

첫 번째는 응답확률 ϕ_i가 조사변수 Y_i와 보조변수 X_i에 좌우되지 않는 형태의 무응답이다. 이는 완전임의무응답(missing completely at random, 이하 MCAR)이라 칭한다. 이러한 무응답은 표본으로부터 확률적으로 추출된 것으로 생각할 수도 있어 무응답의 영향은 무시할 수 있는 정도라고 할 수 있다. 무응답 현상이 완전임의무응답이라고 할 수 있다면, 응답자 s_R는 선택된 표본개체 s를 잘 대표할 수 있다고 할 것이다.

두 번째는 응답확률 ϕ_i이 조사변수 Y_i에 좌우되지는 않지만 보조변수 X_i의 일부

혹은 전부에 좌우되는 무응답현상이다. 이는 임의무응답(missing at random, MAR) 혹은 무시가능무응답(ignorable nonresponse)이라 부른다. 이러한 가정 하에서는, 보조변수 X_i에 의존하는 응답모형을 사용하여 무응답 현상을 설명할 수 있게 된다. 이는 응답여부와 관계없이 모든 표본대상 개체에 대해 보조변수 X_i를 알 수 있기 때문이다. 임의무응답에 대한 응답지시자의 조건부 분포함수는 $P(R_i|Y_i,X_i)=P(R_i|X_i)$으로 표현될 수 있다. 다시 말해, 보조변수 X_i가 주어졌을 때, 조사변수와 응답지시자는 서로 독립적, 즉 $Y_i \perp R_i|X_i$이라고 표현할 수 있다.

마지막으로 응답확률 ϕ_i이 하나 혹은 그 이상의 조사변수 Y_i에 좌우되는 무응답 형태이다. 이는 비임의무응답(NMAR, not missing at random)이라 칭한다. 실질적인 자료 분석에 있어서 이러한 형태의 무응답은 (정확히) 평가할 방법이 없다. 따라서 이러한 경우라면 무응답 매커니즘에 대한 조건부 분포함수는 더 이상 단순화되지 않게 되며, 무응답 추정연구를 제외하고는 무응답 패턴을 평가할 방법이 없게 된다.

3.1.3 성향점수

확률응답모형에서는 모집단내 개체별 (미지의) 응답확률, 즉 성향점수 ϕ_i $(i=1,...,N)$가 정해져있고, 표본에 포함될 때 이에 따라 응답이 결정된다고 가정하였다. Rosenbaum and Rubin (1983)은 모든 표본개체 i에 대한 보조변수 x_i가 주어진다면 이를 통해 다음과 같이 정의되는 성향점수를 추정하는 기법을 소개하였다.

$$\phi(x_i)=P(R_i=1|I_i=1,X_i=x_i)$$

이 같은 성향점수방식의 접근에서는 보조변수 X값이 같은 개체들은 동일한 응답확률을 갖는 임의무응답(MAR)을 가정한다. 성향점수는 종종 로지스틱회귀(logistic regression) 모형이나 프로빗(probit) 모형으로 각각 다음과 같이 정의된다.

$$\phi(x_i) = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}} \quad \text{혹은} \quad \phi(x_i) = \Phi(x_i'\beta)$$

여기서 x_i는 $p-$개 값으로 이루어진 보조변수 값의 벡터이고 β 또한 $p\times 1$ 벡터이며, $\Phi(u) = \int_{-\infty}^{u} (2\pi)^{-1/2} e^{-u^2/2} du$ $(-\infty < u < \infty)$은 표준정규의 누적분포함수를 나타낸다.

3.1.4 무응답 조정

무응답에 대한 조정은 무응답 현상에 대해 얼마나 잘 이해하고 있는가에 따라 달라질 수 있을 것이다. 만약, 표본개체를 응답자와 무응답자의 특성이 같은, 즉, $\overline{Y}_R \approx \overline{Y}_{NR}$인 여러 개의 계급(class 혹은 cell)으로 나눌 수 있다면, 계급별로 무응답 편향(non-response bias)은 제거할 수 있다. 이러한 고려가 반영된 무응답 조정을 가중치 계층조정(weighting class adjustment)이라 할 수 있다(Valliant *et al.*, 2013). 이 방법은 가중치 조정계층(weighting adjustment cells) 내 응답개체와 무응답개체의 설계가중치 $d_{0i} = 1/\pi_i$에 가중응답률 rr_{wc}의 역수와 0을 각각 곱하여 무응답조정 가중치 w_i^*를 얻게 된다. 즉, 가중치 조정계층 c의 응답개체 i의 무응답조정계수는 다음과 같이 표현된다.

$$w_i^{FRM} = (\pi_i \times rr_{wc})^{-1}$$

여기서 조정계층별 가중응답률 rr_{wc}은 다음과 같이 정의된다.

$$rr_{wc} = \Sigma_{s_R} d_{0i} / (\Sigma_{s_R} d_{0i} + \Sigma_{s_{NR}} d_{0i}).$$

이때 가중치 조정계층(weighting adjustment class)을 결정할 때 조사변수 y는 무응답 개체들에 대해서는 알 수 없으므로 실질적으로는 표본추출틀에 존재하는 개체

들의 주요한 보조변수 x를 기준으로 삼게 된다.

일반적인 경우에 있어서 무응답에 의한 편향을 줄이거나 제거하기 위해, 표본설계 및 무응답 매커니즘(non-response mechanism)을 함께 고려한 가중치 조정 방식을 고려한다. w_i*을 무응답 조정가중치이고 단순 총합추정량 $\hat{Y}=\Sigma_{i\in s_R} w_i^* y_i$으로 모총합 $Y=\Sigma_U y_i$를 추정한다고 가정하자. 표본설계와 응답 매커니즘 하에서 $\hat{Y}$가 불편추정량이 되기 위해서는 응답개체의 무응답조정가중치는 다음과 같이 유도된다(Valliant *et al.*, 2013).

$$w_i^{RRM}=(\pi_i\times\phi_i)^{-1}$$

무응답 조정을 위해 고려할 수 있는 보조변수 x는 다음과 같은 세 가지의 특성을 갖추고 있는 것이 좋다. 첫째, 무응답성향(response behavior)을 잘 설명할 수 있어야 한다. 둘째, 조사가 목표로 하는 주요 관심변수(target survey variables)를 잘 설명할 수 있어야 한다. 셋째, 조사통계작성을 위해 고려하는 가장 중요한 영역을 잘 나타내 주어야 한다. 보조변수의 선택에 대한 상세한 논의는 Bethlehem *et al.* (2011, 9장)을 참고할 수 있다.

3.1.5 응답률 평가

단위무응답의 주요 발생요인으로 3.1.1절에서 언급하였듯이 접촉과 조사설득으로 나뉠 수 있다. 이와 관련한 지수들은 조사결과에 대한 조사품질기준으로 종종 고려된다. 조사결과 지표들은 조사결과분류(disposition)를 바탕(<표 2> 참조)으로 정의될 수 있는데, 미국여론조사학회(AAPOR)의 지침서(AAPOR, 2011)에서 제시하는 비를 고려하고자 한다.

접촉률(contact rate, CR)은 표본개체 중 접촉에 성공한 개체수의 비로 정의되고, <표 2>의 구분자를 이용하면 다음과 같이 나타낼 수 있다.

$$CR = \frac{\text{접촉성공 적격개체수}}{\text{적격개체수}}$$

$$= \frac{I + R + R + O}{I + P + R + NC + U + O}$$

접촉률 CR은 AAPOR의 $CON1$에 해당한다. AAPOR는 $CON1$ 이외에도 두 종류의 추가적인 접촉률을 제시하는데 이는 적격불명 개체에 대한 처리방식에 따라 달리 정의된다.

<표 2> 조사평가를 위한 조사결과분류

구분자	분류	조사적격성
I	조사완료	적격
P	부분완료	적격
R	거절/중단	적격
NC	접촉불가	적격
U	적격불명	적격불명
NE	부적격	부적격
O	기타	적격

응답률(response rate, RR)은 조사적격개체 중 응답개체의 비로 정의된다. <표 2>의 분류를 이용하면 다음과 같이 표시된다.

$$RR = \frac{\text{응답가구}}{\text{조사적격가구}}$$

$$= \frac{I + P}{I + P + R + NC + O + (e \times U)}$$

응답률 RR은 AAPOR의 네 번째 응답률 $RR4$로 CASRO[1] 응답률이라고도 칭하는데 가장 널리 사용된다. 여기서 e는 적격성 여부를 아는 개체들의 비율로 추정되는 적격불명 개체들에 대한 적격률(eligibility rate)의 추정치가 된다. 따라서 응답률의 분모는 확인된 조사적격 개체수가 아닌 추정된 조사적격 개체수에 해당한다.

1) CASRO는 the Council of American Survey Research Organizations의 줄임말인데, RR은 CASRO에서 추천하는 응답률 정의이다.

접촉률과 응답률 외에도 접촉정보를 얻은 개체의 비율인 추적률(location rate), 조사협조 개체의 비율인 협조률(cooperation rate) 등의 다양한 조사결과 비율들이 정의될 수 있다.

오랫동안 응답률과 무응답 편향은 동일시 여겨지거나 무응답률(즉, $NR=1-RR$)이 높은 경우 무응답 편향이 매우 커질 수 있는 것으로 인식되어져 왔다. 이러한 논리의 근거는 고정응답모형에 의한 표본평균의 무응답 편향을 나타내는 식 (1)을 통해서도 알 수 있다. 따라서 무응답을 가능한 줄여줌으로 무응답에 의한 편향을 축소하고자 노력하였다. 하지만 최근의 경험적 연구에서는 응답률과 무응답 편향 간에 상관관계가 약할 수 있음이 지적되었다(Curtin *et al.*, 2000).

응답률이 무응답 편향에 대한 평가를 제공하지 못할 수는 있지만 응답률의 저하는 무응답 편향에 대한 많은 우려를 낳았다. Schouten *et al.* (2009)은 응답률을 대체할 무응답 편향에 대한 지표로 R-지표(R-indicator)를 다음과 같이 제시하였다.

$$R(\phi(X))=1-2S(\phi(X))$$

여기서 $S(\phi(X))$는 응답성향점수 $\phi(X)$의 표준오차이고 X는 모든 표본개체에 대해 알고 있는 보조정보를 나타낸다. $R(\phi(X))$는 전체 표본개체에 비교하여 응답개체의 응답성향점수가 유사한가를 측정한다고 할 수 있다. $R(\phi(X))$이 1의 값에 가까울수록 응답개체의 대표성이 양호한 것을 뜻하며 이는 낮은 무응답 편향을 의미한다.

3.2 무응답 축소를 위한 표본설계

Hansen과 Hurwitz (1943)은 무응답 추적조사를 위해 이중추출 기법(nonresponse follow-up study 혹은 double sampling for nonresponse)을 사용할 것을 고려하였다. 이중추출이란 첫 번째 단계(first-phase)에서 선택된 n개의 표본개체는 n_R개의 응답개체와 n_{NR} 무응답개체로 나뉘고, 무응답개체에 대해 두 번째 단계(two-phase)에서는 무응답의 일부인 $100\upsilon\%$를 부가표본(subsample)의 형태로 추출하여 비록 많은 비용이 들더라도 완벽한 무응답 추적을 만들어 낼 수 있

도록 하는 조사방식이다. 이중추출 기법을 적용할 때는 주로 초기조사에서는 우편조사와 같은 저렴한 조사모드를 적용하고 추적조사에서는 면접원 방문조사와 같은 고비용의 조사모드를 적용한다(Lohr, 2010). 따라서 이러한 이중추출 기법은 4장2절에서 다룰 순차적 혼합모드를 통한 추론의 특수한 경우에 해당하는데, 이러한 접근은 다음의 세 가지 가정을 근거로 한다.

- 결정적 무응답(deterministic nonresponse),
- 완전한 무응답 추적(full follow-up response),
- 모드간 측정차이 부재(no measurement difference between modes)

4. 혼합모드조사

4.1 개요

조사모드는 각각의 장단점을 갖고 있어 특정한 조사모드를 선택할 때 조사의 질과 비용을 절충하여 고려하여야 한다. 예를 들면, 양질의 조사결과는 물론 높은 응답률도 얻고자 한다면 면접조사가 최선의 선택일 수 있지만 많은 비용이 들게 됨에 따라 충분한 표본규모를 확보할 수 없는 제약이 따르게 된다. 따라서 조사결과의 질적 감소를 감수할 수 있다면 저비용의 우편조사 혹은 인터넷조사도 함께 고려할 수 있을 것이다.

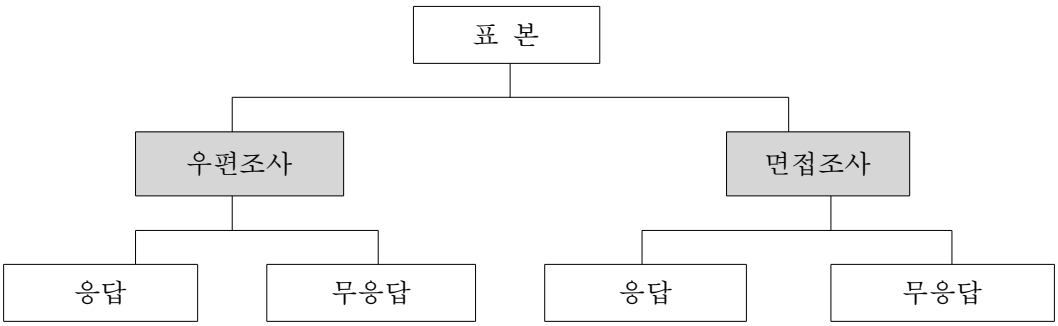
혼합모드는 적응방식에 따라 동시에 적용하는 병행적 혼합모드(concurrent mixed mode), 차례대로 적용하는 순차적 혼합모드(sequential mixed mode), 조사 진행에 따른 실시간 평가를 통한 적절한 모드를 선택하는 적응적·반응적 혼합모드(adaptive 혹은 responsive mixed-mode) 방식 등으로 구분할 수 있다.

객관적 정보에 근거한 혼합모드의 선택이 되기 위해서는 해당모드가 조사품질에 어떠한 영향을 미치는 지에 대해 적절한 평가가 가능하여야 할 것이다. 또한 자료품질에 부정적인 영향을 줄 수 있는 원인들을 최소화하는 혼합모드 적용의 조사방식을 설계하는 것을 계획한다면, 모드효과의 원인들을 이해할 수 있어야 한다. 이러한 노력을 통해 모드효과에 대한 적절한 조정도 가능해 질 수 있을 것이다.

4.2 적용방식

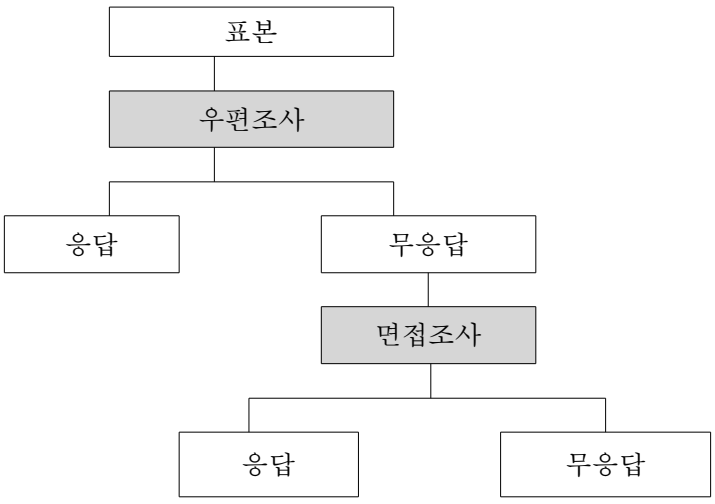
혼합모드의 병행적 적용(concurrent administration)은 다음과 같이 기술할 수 있다. 우선 전체 표본 s를 G개의 부표본 $s^{(1)},s^{(2)},...,s^{(G)}$으로 분리시킨 후, 부표본별로 각기 다른 모드를 적용하여 동시에 조사를 수행하는 방식이다. <그림 3>는 우편조사와 면접조사의 병행적 적용을 도식화하고 있다.

<그림 3> 병행적 혼합모드 적용



반면, 혼합모드의 순차적 적용(sequential administration)은 표본개체 모두에 대해 동일한 모드를 적용하여 조사를 먼저 수행하고 무응답자에 대해 다른 모드를 적용하여 추적조사를 수행하는 방식이다. 이러한 적용은 여러 차례 반복될 수 있지만 동일한 조사시기에서 동일한 방법을 채택하게 된다. <그림 4>는 우편조사와 면접조사를 순차적으로 적용하는 경우를 도식화하고 있다.

<그림 4> 순차적 혼합모드 적용



적응・반응적 적용(adaptive・responsive administration)은 표본조사에 따른 조사대상자의 반응을 실시간으로 평가하여 조사비용 절감 및 무응답편향의 축소를 지향하는 조사방식의 변경을 허용한다. 예로, 미국 질병관리본부의 가구성장조사

(National Surveys of Family Growth, NSFG)는 적응방식을 채택하고 있는데, 적응답식 채택의 목적을 다음과 같이 기술하고 있다 (Lepkowski *et al.*, 2013): "조사 현장업무(field work)를 관리하고, 조사비용을 감독하며, 특정 영역에 대한 과대추출을 적용하는 한편, 그 결과로 얻게 되는 표본의 편향을 줄일 수 있도록 하고자 한다."

4.3 장단점

혼합모드의 장점은 다음과 같다. 첫째, 모드별로 효율적인 접촉가능 계층을 다양하게 구성할 수 있어 단일모드조사와 비교할 때 포함오차(coverage error)를 줄일 수 있다. 둘째, 다양한 조사모드의 선택을 통해 무응답률과 무응답 편향을 축소시킬 수 있는 가능성을 증대시킨다. 마지막으로, 저비용의 조사모드를 선택하여 조사비용 절감이 가능해진다. 예를 들면, 기성세대가 방문조사를 선호하는 반면 젊은 세대는 웹조사를 선호한다면, 웹조사와 방문조사를 혼합한 조사가 이 둘 중 한 가지 조사모드만을 채택한 단일모드조사에 비해 연령계층별로 조사참여률을 높일 수 있을 뿐만 아니라 웹조사의 사용으로 인해 전체적인 조사비용을 절감하는 효과도 동시에 누릴 수 있게 된다.

하지만 혼합모드는 단점도 갖게 된다. 서로 다른 부류의 응답자가 상이한 조사모드를 선택하여 조사에 응하는 선택효과(selection effect)와 동일한 설문이 서로 다른 조사모드로 주어질 때 이로 인해 응답간의 차이가 발생되는 측정효과(measurement effect) 등의 모드효과가 생겨 혼합형태의 편향을 초래할 수 있다. 모드효과에 대한 적절한 평가와 조정 없이는 자료비교성이 떨어질 수 있는데 다음 절에서 이에 대해 보다 자세히 다루고자 한다.

4.4 모드효과 이해, 실험설계 및 평가

4.4.1 모드효과 이해

모드효과(mode effects)란 동일한 질문에 대해서 모드를 달리할 때 상이한 결과를 얻는 것을 일컫는다. Biemer and Lyberg (2003)은 모드효과를 순수모드효

과(pure mode effect)와 모드시스템효과(mode system effect)로 분류하고 있다. 순수모드효과란 자료수집모드의 선택에 따른 상이한 조사결과를 초래하는 것을 말한다. 예를 들면, 동일한 질문으로 동일ㄴ 특성을 측정하되 동일한 시기에 조사하여 순수한 모드 이외의 변동이 반영되지 않는 차이를 말한다. 반면, 모드시스템효과란 조사설계 상 모드선택에 따른 조사시스템 혹은 설계요소들이 복합적으로 상의한 조사결과를 초래하는 것을 말한다. 예를 들어, CAPI 조사와 우편조사는 2.1절에서 살펴본 것처럼 설문지(전자 대 종이), 응답률, 응답자 구성 등에 있어서 차이를 보일 수 있다. 순수모드효과는 모드시스템효과에 비해 평가하기 매우 어렵기 때문에 (Bethlehem *et al.*, 2011), 본 논의에서는 모드시스템효과를 모드효과로 고려한다. Roberts (2007)는 상이한 모드선택으로 (즉, 모드효과로) 인해 영향을 줄 수 있는 세 가지의 비표본오차(non-sampling error)로 포함오차, 무응답편차, 측정오차를 지적하고 있다.

모드효과는 선택효과(selection effects)와 측정효과(measurement effects)로도 구분된다(Vannieuwenhuyze *et al.*, 2010). 선택효과란 조사대상자 유형에 따른 모드선택이 상이할 때 발생하는 효과를 일컫는다. 다양한 유형의 조사대상자가 특정모드에 응하지 않고 다른 모드를 선택하는 일종의 무응답오차의 형태이다. 혼합모드를 채택하면 선택효과로 인해 단일모드로는 참여하지 않는 유형의 대상자들을 조사에 참여시킬 수 있는 장점이 있다. 반면, 측정효과는 조사대상자가 적용된 조사모드에 따라 상이한 응답을 하게 되는 것을 말한다. 다시말해, 측정오차(measurement error)로 인해 발생하는 효과를 뜻한다. 측정오차는 조사항목의 (병행적 혹은 순차적) 제공방식, 면접원 효과(interviewer effect)와 면접 혹은 자기기입조사에서의 소망효과(social desirability), 전화조사에서의 초두효과(primacy effect)나 최근효과(recency effect), 회상편의(recall bias), 묵인(acquiescence)으로 발생하는 차이로부터 기인한다. 모드선택에 따른 인지·반응에 영향을 주는 요소들은 2.3절을 참고할 수 있다. 혼합모드적용의 단계별 시스템들과 적용근거 및 조사품질에 대한 영향에 대한 상세한 논의는 de Leeuw(2005, 238쪽)을 참고할 수 있다.

모드효과는 선택효과와 측정효과로 구분할 수는 있지만 두 효과가 완전히 뒤섞여 있어 개별효과를 분리하여 평가하기가 쉽지 않다. de Leeuw (2005, 249쪽)는 이 두 효과를 분리하여 평가하기 위해서 조사대상자를 모드에 덜 민감하다고 판단

되는 변수 (예, 나이와 교육)로 매칭시킨 후 매칭그룹 내 조사값의 차이를 살펴봄으로써 혼합모드적용에 의한 선택효과를 제거한 측정효과를 평가하도록 하는 방식을 고려하였다. 물론 매칭변수는 모드와 관계없이 측정(예, 등록증 혹은 표본추출틀) 되어야 하며 또한 이러한 방법이 잘 계획된 실험설계에 의한 평가보다는 매우 취약할 수 있음을 언급하고 있다.

4.4.3 모드효과 평가를 위한 실험설계 – J□ckle *et al.* (2010)

J□ckle *et al.* (2010)은 측정에 대한 모드효과, 즉 측정효과에 대한 평가와 관련한 연구에서 2003년과 2005년도 유럽사회조사(European Social Survey, ESS)에서 수행한 두 가지 실험설계를 소개하고 있다.

실험설계 I (Phase I)는 회의장조사(hall test)[2]의 형태로 이루어졌는데, 조사참여자들을 나이, 성별, 교육 등의 특성으로 특정국가의 도시인구를 대변할 수 있도록 할당표집(quota sample)에 의해 선택하였다.[3] 이들은 여러 종류의 조사모드[4]로 정해진 실험조건(treatment condition) 중의 하나에 임의로 할당(random allocation)되어 조사에 참여하였다. 이들은 다시 동일한 질문에 대해 실험에서 고려된 다른 각각 모드로 조사를 진행하여 조사모드 간 문항은 물론 조사 참여자 간의 비교 분석이 가능토록 하였다.

실험설계 I에서는 모집된 참여자들을 실험조건들에 임의로 할당함으로써 조사모드에 따른 상이할 수 있는 응답성향 혹은 응답률을 통제하였다. 하지만 이러한 접근은 실질적인 조사현실과는 다를 수 있어 평가결과에 대한 실질적 적용은 용이하지 않을 수 있다는 단점을 갖는다.

실험설계 II (Phase II)는 기존 조사의 조사모드인 면대면조사와 대안으로 고려된 전화조사 간의 비교를 위해 수행되었다.[5] 조사비용의 절감을 위해 실험조사는 해당 국가의 수도에서만 수행되었고, 상이한 모드 사용에 의한 표본추출 및 포함률과 관련한 조사 조건 간의 차이를 줄이기 위해 전화번호와 함께 주소가 있는 대상

2) 회의장조사(hall test)는 시장조사의 한 기법으로 응답자로 하여금 실험장소(hall)에 모이게 하여 시제품, 광고 카피 등에 대한 반응을 테스트하는 것이다.
3) 실험설계는 2003년 5월과 6월에 헝가리 조사에서 수행되었다.
4) 4종류 조사모드 실험조건은 면대면조사, 전화조사, 자기기입식 종이조사, 웹기반 조사 등이다.
5) 실험설계는 2005년 7월에 헝가리와 포르투갈에서 수행되었다.

만을 실험조사의 표본으로 선택하였다. 이렇게 선택된 표본가구는 접촉을 시도하여 조사 의향을 먼저 확인한 후, 세 가지의 조사모드의 (처리)조건[6]으로 무작위로 배정되고 최근생일법(last birthday method)에 의해 가구 내 대상인 15세 이상의 구성원을 추출하였다. 실험설계에서는 모두 3가지의 실험조건을 고려하였는데 이는 각각 (i) 설문지(showcard)을 이용한 면접조사, (ii) 설문지를 사용하지 않은 면접조사, (iii) 설문지를 사용하지 않되 조건 (ii)와 동일한 질문으로 진행하는 전화조사이다.

실험설계 II에서는 조사모드에 따른 표본추출과 포함오차에 의한 측정효과에로의 교략을 제거하기 위해 표본선택을 주소와 함께 전화번호가 주어진 가구들로 제한하였다. 또한 세 가지의 실험조건들은 측정효과에 대한 평가에 있어서 조사모드의 특성, 즉 설문지 사용여부나 조사시 면접원의 존재여부 등에 대한 영향력을 통제할 수 있게 하였다.

4.4.4 모드효과 평가를 위한 실험설계 – Vannieuwenhuyze *et al.* (2010)

Vannieuwenhuyze *et al.* (2010)는 2008년도 유럽사회조사 중 네덜란드에서 시행된 모드효과 평가를 위한 실험설계를 소개하고 있다. 본 연구에서는 기존의 본조사와 동일한 질문을 사용하여 방문조사(CAPI), 전화조사(CATI), 인터넷조사(CAWI)를 함께 고려한 혼합모드를 병행하여 조사하였다.

본 연구를 위해 네덜란드 표본틀내 개체 중 전화번호 정보가 있는 표본명단에서 본 조사 대상(main ESS)으로 2,674명을 임의로 선택하였고, 또한 독립적으로 878명을 임의로 추출하였다. 본 조사의 대상자들은 총 10회 이하의 조사원 방문을 통해 접촉하였고, 혼합조사의 조사 대상자들(MM experiment)[7]은 최초 접촉에서 세 가지의 조사모드 중 하나를 선택할 수 있도록 하는 병행적 혼합모드방식으로 수행하였다. 혼합모드조사는 먼저 전체 조사대상자들에게 총 14번의 전화통화로 최초 접촉이 시도되었다. 접촉이 성공된 대상자들 중 조사에 응하고자 하는 이들에게는

6) 실험에 고려된 조사모드는 두 가지로 조사대상자의 집을 방문하여 진행하는 면접조사와 유선 혹은 무선전화에 의한 전화조사이고, 실험조건은 세 가지로 설문지(showcards)를 이용한 면접조사, 설문지를 사용하지 않은 면접조사, 설문지를 사용하지 않은 전화조사로 구성되었다.

7) 네덜란드 본조사와 실험조사에서 모두 가구 내에서 16세 이상의 구성원 중 한명을 임의로 선택하였다.

세 가지 조사모드 중 원하는 모드를 선택할 수 있게 하였다. 조사모드를 참여하기로 동의한 대상자들이 실제로 응답하지 않은 경우, 선택된 조사모드와 다른 모드를 선호할 때는 기존 선택을 바꿀 수 있도록 허용하였다. 단, 방문조사의 경우에는 추적조사를 수행하지 않았다. 또한 최초 접촉시도가 성공적이지 못했던 대상자들은 조사원들의 방문을 통해 조사를 수행하였는데 앞서와 마찬가지로 대상자가 다른 조사모드를 선호할 경우, 조사모드의 변경을 허용하였다. <표 3>은 조사대상자들에 대한 응답건수 및 응답률 통계를 포함하고 있다. 혼합조사와 본조사의 자료들은 모집단에 대한 대표성(population representativeness)을 향상시키기 위해 각각 인구·사회학적 변수(나이x성별, 도시화 구분, 가구크기)를 기준으로 래이킹 비 조정(raking ratio weight adjustment)을 고려하였다.

유럽사회조사의 조사항목 중 “정치적 관심”이라는 질문은 모드효과가 존재할 것으로 예상되는데, 이는 면접원과의 대면조사에서 응답자가 갖는 소망효과로 인해 측정오차가 발생할 수 있다. 또한 무응답자는 일반적으로 정치에 무관심한 것으로 알려져 있는데, 혼합모드조사에서 면접조사(CAPI)의 많은 부분이 비면접조사의 무응답자들로 구성되어 있기 때문에 “정치적 관심”이란 질문에 대해 선택오차를 발생시킬 수도 있을 것이다.

<표 3> 유럽사회조사 네덜란드 실험조사 응답건수 및 응답률 통계

	혼합조사	본조사
CAWI	160	-
CATI	88	-
CAPI	104	1294
전체응답	352	1294
부분응답	15	72
무응답	313	1022
비접촉	108	125
부적격	90	161
총표본	878	2674
응답률	44.7%	51.5%

출처: Vannieuwenhuyze *et al.* (2010, 1035쪽의 표1)

<표 4>는 정치적 관심을 묻는 질문에 대해 관측된 응답수준별 표본비율과 평균을 혼합모드의 조사모드와 본 조사 전체에 대해 각각 나타내고 있다. <표 4>는 정치적 관심에 대한 응답수준별과 전체 평균의 측정효과와 선택효과의 추정치 및 정도수준을 보여주고 있다. “거의 없음”의 응답비율은 CATI/CAWI에서 CAPI 보다 더 높게 나타난 반면, “꽤 있음”의 응답비율은 CATI/CAWI에서 CAPI 보다 더 낮게 나타나고 있다. 두 응답수준에 대해서는 모두 통계적으로 유의적인 측정효과가 있는 것으로 판단되나 선택효과에 있어서는 유의성이 없는 것으로 나타난다. 하지만 항목평균에 대해서는 측정효과는 물론 선택효과 모두에서 통계적으로 유의함을 알 수 있다.

<표 4> 정치적 관심 문항의 응답수준별 표본비율 및 모드효과 평가

표본비율	**혼합조사**		**본조사**	
	CATI/CAWI	**CAPI**	**(CAPI)**	
전혀 없음	0.084	0.033	0.067	
거의 없음	0.330	0.188	0.224	
꽤 있음	0.488	0.679	0.607	
매우 있음	0.098	0.100	0.101	
평균	2.600	2.846	2.743	
	효과	**표본오차**	**양측검정**	**단측검정**
측정효과				
전혀 없음	0.005	0.021	0.823	0.412
거의 없음	0.093	0.037	0.012	0.006
꽤 있음	-0.094	0.041	0.023	0.012
매우 있음	-0.004	0.025	0.877	0.439
평균	-0.107	0.062	0.086	0.043
선택효과				
전혀 없음	-0.046	0.028	0.100	0.050
거의 없음	-0.049	0.060	0.420	0.210
꽤 있음	0.097	0.072	0.178	0.089
매우 있음	-0.002	0.046	0.964	0.482
평균	0.139	0.098	0.154	0.077

[출처: Vannieuwenhuyze *et al.* (2010, 표2와 표3)]

응답수준별 측정효과와 선택효과의 평가는 다음과 같은 이론적 근거를 따른다.

먼저 혼합모드조사는 조사대상자의 선택에 따라 두 종류의 조사모드 A 혹은 B를 통해 조사가 진행되고 조사변수 Y는 J개의 범주를 갖는 범주형 변수라고 가정한다. 조사대상자의 모드선택에 따라 조사모드 A인 경우에는 Y_a으로 조사모드 B인 경우에는 Y_b으로 표기하자.[8] 범주 j의 모비율(π_j)에 대한 선택효과(selection effect)는 혼합모드 하에서 서로 다른 모드를 선택할 계층들에게서 특정한 모드(예, A)를 통해 측정될 수 있는 두 모비율 간의 차이로 다음과 같이 정의된다.

$$\sigma_a(\pi_j) = P(Y_a = j|M = a) - P(Y_a = j|M = b) \quad (4)$$

모비율(π_j)에 대한 측정효과(selection effect)는 혼합모드 하에서 특정모드(예, B)를 선택할 계층에 대해 서로 다른 모드로 측정될 수 있는 두 모비율 간의 차이로 다음과 같이 정의된다.

$$\mu_b(\pi_j) = P(Y_b = j|M = b) - P(Y_a = j|M = b) \quad (5)$$

위의 두 모드효과의 정의에서 첫 번째 확률은 선택모드와 관측모드가 동일하기 때문에 혼합모드조사에서 추정이 가능하지만 두 번째 확률 $P(Y_a = j|M = b)$은 조사현실상 선택모드와 관측모드를 달리 할 수 없기 때문에 관측이 불가능하다. 따라서 후자는 독립적인 비교 가능한 단일모드조사를 수행하고 이를 총 확률식 $P(Y_a = j)$[9]을 이용한 조건부 확률의 도출을 통해 간접적으로 추정할 수 있게 된다.

모드효과 측정은 "대표성 가정(representation assumption)"을 만족함을 전제로 한다. 두 조사를 통해 실현된 응답자(realized sample), 즉, 혼합모드와 단일모드조사의 두 조사를 위해 추출된 표본추출과 실제 응답자들이 동일한 모집단을 대표해야 한다는 암묵적 가정이다. 다시 말해, 두 표본(응답자)은 모집단 포함범위와 무응답오류에 있어서 차이가 없어야 함을 가정한다. 이러한 가정은 두 가지의

8) 만약 개별대상자에 대해 Jäckle *et al.* (2010)의 실험설계에서 처럼 Y_a와 Y_b 모두를 측정할 수 있다면 두 값 간의 차이가 측정오차가 될 것이다. 하지만, 동일인으로부터 같은 질문에 대해 두 가지 모드로 답변을 받는다면 학습효과 등과 같은 간섭작용 내지 교효작용이 발생할 수 있어 얻어지는 조사 값이 엄밀한 의미에서의 참값이라고 할 수 없음으로 해석상의 주의가 필요할 수도 있을 것이다.

9) 총확률식은 $P(Y_a = j) = P(M = a)P(Y_a = j|M = a) + P(M = b)P(Y_a = j|M = b)$인데, 여기서 확률변수 M은 조사대상자의 모드선택을 나타내는 지시변수이다.

논거를 통해 입증할 수 있을 것이다. 첫 번째는 만약 두 표본이 비교 가능한 응답자들로 구성된다면 혼합모드에서 A 혹은 B 중 어떠한 모드를 선택하던 A만이 주어지는 단일모드에서도 응답하였을 것이다. 따라서 두 표본의 응답률 간의 차이도 없을 것이다. 두 번째는 두 표본이 비교 가능한 응답자들로 구성된다면 조사모드에 민감하지 않은 인구·사회학적 특성에 있어서 서로 간의 차이를 보이지 않을 것이다.

하지만 앞서 살펴본 유럽사회조사의 네덜란드 실험조사의 자료분석에서처럼 혼합모드에서 두 개 보다 많은 조사모드를 선택한다면 이들을 분리할 수 없게 되며 또한 함께 묶어 모드효과를 분석한다면 모드간의 교호작용이 존재할 수 있어 주의가 요망된다.

4.5 모드효과 조정

논의의 단순화를 위해 <표 3>과 같이 두 가지 모드를 동시에 사용하는 병행적 적용을 가정한다. 혼합모드의 무응답조정을 위한 가장 단순한 방법으로는 모드에 관계없이 전체 응답과 무응답을 나누어 표본가중치에 무응답 조정방법(3.1.4절 참고)을 적용하는 것이다. 이 방법은 모드별 무응답의 차별성을 고려하지 않기 때문에 조정된 가중치를 사용한 분석은 (응답) 선택오차를 적절히 조정하지 못하고 무응답 편향 (혹은 선택효과)을 가질 수 있게 된다. 따라서 대안적 방법으로 조사대상자의 모드선택 변수를 정의하고 이를 다른 보조변수들과 함께 무응답 조정을 수행하는 것이다. 물론 모드와 응답성향간에 상호작용이 존재한다면 이를 적절히 반영하여 무응답 조정을 고려하여야 할 것이다. 예로, 모드별로 무응답 조정계층을 달리 나누고 가중치를 조정할 수 있을 것이다.

<표 1>에서 살펴본 바와 같이 무응답은 조사단계별 원인(형태)과 순차적 발생이라는 특성을 갖는다. 따라서 이러한 특성을 반영하여 무응답 조정을 고려한다면 무응답 편향을 크게 줄일 수 있을 것이다. 조사대상자가 표본조사에 참여하지 않을 것을 선택한다면 무응답이 발생하고 따라서 대상자의 특성치를 알 수 없게 된다. 이러한 무응답의 결정과정을 표본선택(sample selection)이라고 칭하는 데, Cobben(2009, 9.5.1.3절)은 이러한 무응답 발생의 모형, 즉, 선택모형(selection

model)을 이용하여 혼합모드의 무응답 발생특성을 조정하는 모형에 대해 논하였다.

응답과정을 접촉(contact)과 참여(contact)의 두 단계로 구분하고 관심조사변수 Y는 연속형 변수라고 가정하자. γ_{ki}^{*}와 ρ_{ki}^{*}는 각각 i번째 개체에 대한 모드 k의 잠재접촉확률(latent contact probability)과 잠재참여확률(latent participation probability)을 나타낸다고 하자. 또한 C_{ki}와 P_{ki}는 각각 접촉과 참여여부를 나타내는 지시변수(indicator variable)라고 하자. 병행적 혼합모드에서는 부표본별로 할당된 모드로 병행적 조사를 진행한다. 표본개체의 두 지시변수가 $C_{ki}=1$이고 $P_{ki}=1$ 이면 조사항목 Y_i를 관측할 수 있게 된다. 표본개체의 지시변수가 $C_{ki}=0$이거나 $C_{ki}=1$ 이고 혹은 $P_{ki}=0$인 경우에는 조사항목을 관측할 수 없게 된다. 만약, 앞서의 두 잠재변수와 잠재조사변수 Y_i^{*}를 다음과 같이 정의한다면

$$\begin{aligned}
\gamma_{ki}^{*} &= X_{ki}^{C}\beta^{C}+\epsilon_{ki}^{C} && \text{(접촉잠재변수)}\\
\rho_{ki}^{*} &= X_{ki}^{R}\beta^{R}+\epsilon_{ki}^{R} && \text{(참여잠재변수)}\\
Y_i^{*} &= X_i^{Y}\beta^{Y}+\epsilon_i^{Y} \quad (i=1,\ldots,n) && \text{(조사잠재변수)}
\end{aligned} \tag{4}$$

조사관측변수 Y_i는 다음과 같이 정의될 수 있다.

$$Y_i=\begin{cases} Y_i^{*} & (C_{ki}=1,P_{ki}=1)\\ \text{결측} & (C_{ki}=1,P_{ki}=0)\text{이거나}(C_{ki}=0)\end{cases} \tag{5}$$

여기서 X_{ki}^{m}와 β^{m}는 각각 접촉($m=C$), 참여($m=R$), 조사($m=Y$)와 관련된 보조변수와 회귀계수를 나타낸다.

위의 표본선택모형을 통해 혼합모드의 무응답 편향 (즉, 선택오차)을 평가 (혹은 추정)할 수 있다. 또한 적합된 모형에 의해 예측된 조사변수의 기댓값을 이용한 가중표본추정을 통해 혼합모드의 무응답 편향을 조정할 수 있다. 이때 모형의 추정은 식 (4)의 오차항들에 대해 다변량 프로빗 모형을 가정하여 모드, 응답단계, 조사항목의 상관관계를 반영한 추정을 가능하게 한다. 물론 추정결과는 실제 표본조사 자료에 대한 모형 적합도에 따라 효율성이 결정될 수 있다.

혼합모드의 순차적 적용을 채택한 조사설계에 대해서는 모드적용방식에 따른 응

답진로(mode response path)를 나타내는 또 하나의 지시변수 M_i^T을 정의하고 응답원의 내포화(nesting) 구조를 반영하여 앞서 논의한 표본선택모형에 추가하여 무응답 편향을 평가하고 조정할 수 있다. 선택모형을 이용한 혼합모드의 무응답조정에 관한 상세한 논의는 Cobben (2009, 9장)을 참고할 수 있다.

그 외의 방법으로 Buelens – Van den Brakel(2011)은 특정모드의 사용여부를 한 변수로 포함하여 칼리브레이션 추정방식을 통해 모드비율을 고정비율에 보정시키는 방법을 제시하였다. 이 방법은 편향(bias)을 제거하지 않고, 표본평균들의 변화추정(change estimate)에 있어서의 편향을 제거하는 것을 지향한다.

Suzer-Gurtekin *et al.*, (2012)은 베이지안의 다중대체방식을 이용한 모드효과의 조정을 다음과 같은 5단계의 일반적 접근으로 기술하고 있다.

(1) 다중대체모형을 통해 특정한 자료수집 단계(data collection wave)에서의 응답자들을 전체표본(full sample)로 확대한다.
(2) 개별 모드에 대한 완성자료(관측치+대체값)로부터 평균을 추정. 이때 다중대체방법을 이용하여 대체에 따른 불확실성(uncertainty)을 반영한다.[10)]
(3) 각 모드별 추정량을 비교하여 모드효과 추정한다.
(4) 제안된 방법과 다른 모드들로부터의 결과를 가용한 기준값(benchmark values)과 비교하여 다른 모드 사용에 따른 편향이 존재하는지 평가한다.
(5) 유의한 모드효과가 확인되면 가용한 기준값과 모드혼합의 대체적 방식(alter- native method)에 따른 추정값을 비교한다.

베이지안 다중대체방식은 가능성만을 갖고는 있지만 Suzer-Gurtekin *et al.*의 실제자료에 대한 적용에서 그리 만족할 만한 결과를 보이지 못하고 있어, 추가적인 연구가 필요해 보인다.

10) Rubin, D.B. (1987). Multiple imputation for nonresponse in surveys. Hoboken, New Jersey: Wiley Classics Library.

5. 효과적 혼합모드 사용에 대한 제언

5.1 통계청 사례

5.1.1 경제활동인구조사

경제활동인구조사에서는 1999년에 휴대용 컴퓨터를 이용한 전자조사가 처음 도입되었고, 2004년에는 PDA로 교체하였으며, 2008년에는 인터넷을 이용한 CASI (computer-assisted self interview) 조사, 2009년에는 CATI(computer-assisted tele- phone interview)를 각각 도입하는 등, 다양한 모드를 혼합하여 조사하는 혼합모드 조사방식을 채택하고 있다(박시내 외 2인, 2014). 경제활동인구조사의 혼합모드방식의 특징으로는 조사표 구성이 주로 종이 조사표를 중심으로 진행하되 일부 항목은 CATI 방식에 적합하도록 수정되어 적용되고 있다.

2005년부터 경제활동인구조사의 표본설계방식은 반복조사(repeated survey)에서 연동조사(rotation survey)의 방식으로 변경되었다. 연동조사는 표본가구를 36개월 간 연속해서 조사한 뒤 다른 표본가구로 대체하여 조사한다. 최초 조사부터는 면접조사를 사용하되 6개월 이상 연속적으로 조사에 참여한 가구들 중 변동이 크지 않고 안정적이며 충분한 가구사항을 파악한 경우에는 CATI와 CASI방식으로 조사모드를 변경한다. 하지만 6개월 미만 응답한 가구라도 단독·맞벌이 등 면접이 곤란한 가구이거나 조사에 협조적이고 CATI나 CASI방식의 조사를 희망하는 경우에도 조사모드를 변경할 수 있게 허용하였다. 이같이 차수에 따른 조사방식 변경방식은 해외 주요국가들이 적용하는 방식과 어느 정도 유사하다. 하지만 우리나라 경제활동인구조사의 경우 주요국의 노동력조사가 채택하는 4~8달의 짧은 주기가 아닌 36개월의 긴 연동주기를 갖는 면에서 다르다. 또한 CATI와 CASI 방식으로 조사수행을 하는 경우에도 2개월에 1회 이상 조사대상가구를 방문하여 면접하거나 관리하도록 하고 있다.

2011년 기준의 경제활동인구조사에 적용된 조사모드는 면접조사, 전화조사, 자기기입식조사, e-mail 조사, FAX 조사 및 간접조사가 있으며, 이들은 종이조사와 전자조사 등과 접목하여 실시되었다. 혼합모드로 조사된 자료는 총 4가지의 서로 다른 자료입력방법이 적용되었다. 전자조사표로 진행된 면접조사(CAPI), 자기기입

조사(CASI), 전화조사(CATI)는 조사종류별로 전산 입력되었고, 종이조사표를 이용한 면접조사(PAPI), 일반전화조사(PATI), 전자우편조사(e-mail), 팩스(FAX)조사는 모두 가구부문통합관리시스템(이하, HIMS)으로 입력된다. 하지만 HIMS에 조사모드에 대해 얼마나 상세히 기록하고 있는지 분명치 않다.[11] 조사모드의 선택권과 관련해서 CASI는 응답자가 선택하나, PAPI(FAX, email 포함)와 CAPI의 경우에는 조사원이 선택할 수 있다. <표 5>는 경제활동인구조사의 조사모드별 자료수집비율을 나타내고 있다. 경제활동인구조사에서는 종이조사는 총 63.9%로 전체자료수집에서 대다수를 유지하고 있음을 알 수 있으며, CAPI는 23.9%로 전자조사의 대부분을 이루고 있다. <표 6>는 경제활동인구조사에 적용된 조사방식을 조사방식과 전자화 여부로 분류하여 나타내고 있다.

박시내 외 2인 (2014)의 혼합모드 사용과 관련된 기술로부터 판단하면 경제활동인구조사의 모드효과 분석이 다소 모호할 수 있을 것으로 보인다. 먼저, 기본모드 CAPI 이외의 다른 모드선택이 다소 불분명하다. CATI나 CASI 방식을 적용받기 위해서는 (i) 6개월 이상 응답가구 중 가구변동이 크지 않고 다소 안정적이며 가구상황이 충분하거나, (ii) 6개월 미만의 응답가구이더라도 단독·맞벌이 등의 면접이 곤란한 가구나 조사에 협조적이고 모드의 변동을 원하여야 한다. 이러한 기준은 다양한 조건들로 이루어져 있고 주관적인 측면도 많다. 이러한 기준이 제시되는 이유는 조사현장에서의 조사수행의 편리성을 배려한 것으로 판단된다.

접촉, 응답, 무응답 전환(nonresponse conversion) 등의 조사과정별 조사현황을 나타내는 자료를 알 수 있다면, 조사 진행에 따른 선택효과 및 측정효과를 보다 정확히 알 수 있을 것이다. 하지만, 박시내 외 2인 (2014)에서는 최종 응답모드별 자료수집비율만이 보고되어 있고 조사모드에 대한 정확한 기록은 이루어지지 않은 듯 보인다.[12] 따라서 모드효과에 대한 분석을 하기 위해서는 조사관리에 대한 상세한 정보(paradata)가 기록되고 보고되어야 할 것이다.

---

11) 박시내 외 2인 (2014, III장1절 참고)에 따르면 'HIMS에는 종이조사표를 이용한 대면면접 외에 일반전화를 이용한 전화조사 ..... 팩스(FAX)조사 등의 자료수집방법이 포함된다. 그러나 HIMS에서 일반전화를 이용한 전화조사 등은 통계청 현장조사 운영지침 및 경제활동인구조사 지침서에 따라 면접조사나 전자조사를 할 수 없는 경우에 허용되기 때문에 규모파악을 정확히 할 수는 없지만, 그 규모가 많지 않을 것으로 판단되고, 따라서 없는 것으로 간주한다.'이라 기술되어 있다.

12) 예로, 각주 11 참조.

<표 5> 경제활동인구조사 조사모드별 자료수집비율

조사모드		자료수집비율
종이조사		63.9%
전자조사	CAPI	23.9%
	CATI	8.7%
	CASI	3.6%
	소계	36.2%
총계		100.0%

<표 6> 경제활동인구조사에 적용된 조사방식 비교

조사방식	종이조사	전자조사
면접원조사	PAPI	CAPI
전화조사	PATI	CATI
웹조사	—	CASI
이메일조사	e-mail	—
FAX조사	FAX	—

5.1.2 인구주택총조사 시험조사

2010 인구주택총조사에서는 조사 참여의 편의성을 제공하기 위해 전통적인 면대면 조사 이외에도 다른 조사모드를 통하여 다양한 응답기회를 갖게 함은 물론 조사비용의 절감 효과도 모색하였다. 이러한 노력의 일환으로 본 조사에 앞서 두 차례에 걸친 시험조사를 수행하였다. 2007년의 1차 시험조사에서는 인터넷조사와 우편조사를 사용하였고, 2008년의 2차 시험조사(이하, 시험조사 II)에서는 자동응답시스템(automatic response system, 이하 ARS)과 CATI를 추가적으로 고려하였다. 박영실·정남수(2008)은 시험조사 II의 결과를 토대로 자료수집방법에 대한 비교분석을 실시하였다.

박영실·정남수(2008)에 따르면 시험조사 II는 선택된 일부지역[13]에 대해 3단계

13) 시험조사지역으로는 부산광역시, 경기도, 강원도이고 각각 두 곳의 행정동 혹은 읍이 포함되었다. 박영실 정남수(2008)의 표<2-12>를 참조할 수 있다.

에 걸쳐 혼합모드방식으로 진행하였다(<그림 3>). 1단계에서는 준비기간을 포함하여 총 12일간 인터넷·우편·ARS·CATI 등 네 가지 비방문 조사방식을 고려하였다. 인터넷조사는 조사 홈페이지에 접속한 후, 실명인증을 통해 조사표를 입력하도록 하였다. 우편조사는 준비기간 중 조사원이 배부한 조사표를 일반조사구와 우편조사구의 구분에 따라 각각 전통적인 우편배달방식과 관리사무소 혹은 경비실 등 공동주택 내 설치된 우편함을 통한 회수하는 방식으로 조사를 실시하였다.[14] ARS와 CATI조사도 동시에 실시되었는데 전화조사 시스템으로 직접 전화를 걸어 응답을 받는 방식을 택하였다.[15] 1단계의 비방문조사가 완료되고 이틀간 조사현황을 정리하여 응답과 무응답 가구명부를 정리하여 조사원별 방문조사의 업무량을 배정하였다. 비방문조사 시 응답하지 않은 가구를 대상으로 조사원 방문에 의한 면접조사를 실사하였다. 만일 조사대상가구가 자기기입방식을 원할 경우에는 조사표를 배포한 후, 추후에 회수하는 배포조사 (혹은 유치조사)의 방법을 동시에 적용하였다. 2단계 조사에서 2번 이상의 방문에도 불구하고 면접을 완료하지 못한 가구에 대해서는 전화조사를 통해 추적조사를 실시하였다.[16] 한편 비방문조사 자료수집방법의 참여율을 높이고자 인센티브를 제공하였는데, 인센티브의 제공은 조사 후 추첨을 통한 경품제공의 형식을 취하였고 홍보효과의 평가를 위해 홍보지역과 홍보하지 않은 대조군지역으로 나누었다.

시험조사 II를 조사품질 및 모드효과의 측면에서 측면을 살펴보면 다음과 같다. 최종 추적조사의 기간이 명시되어 있지 않지만, 시험조사 II의 1단계와 2단계의 조사 일수만을 고려하면 총 28일로 매우 짧은 것으로 보인다.[17] 비록 본 조사의 사

---

14) 박영실· 정남수 (2008)에는 '우편회수함을 설치한 경우라도 응답자의 선호에 따라서 우편배달방식을 택할 수도 있으므로 엄밀히 말해서 아파트조사구라고 해서 모두 우편회수함을 통해서 조사표를 회수하였다고 단정지을 수는 없다. 마찬가지로 일반조사구의 아파트에서도 회수함이 설치되었을 수 있다. 다만, 아파트조사구인 경우가 일반조사구인 경우에 비해서 회수함에 의한 조사표 소거의 양이 많을 것이라는 가정 하에 이후의 분석이 진행되었다'라고 기술하고 있다. 앞서의 기술내용으로 보았을 때 우편조사에서 조사표는 우편으로 발송되었을 가능성이 있고 작성된 조사표는 주거지역 내 우편함 상황에 따라 회송되었을 것으로 추측할 수 있다. 하지만 <그림 3>에서 1단계의 조사모드 적용에 대한 특별한 기술이 없어 조사표가 우편으로 발송되었다고 확신할 수는 없다.

15) 전화조사의 경우, 박영실· 정남수 (2008)의 기술을 토대로 판단할 때, 모두 전자조사 방식으로 수행되었을 것으로 판단되며 앞서의 비방문 방식의 조사들과 더불어 어떻게 조사방식이 조사대상자들에게 할당되었는지와 전화번호 정보의 출처의 정보는 주어지지 않았다.

16) 박영실· 정남수 (2008)는 "방문조사기간에도 응답자의 선호에 따라서 인터넷이나 우편조사에 의한 응답이 이루지기는 하였으나 그 사례는 매우 적은 편이다"라고 밝히고 있다. 이는 3단계의 무응답 최종추적에서도 조사대상가구에서 조사모드를 선택할 수 있게 허락함을 알 수 있고 또한 사례는 매우 작더라도 전화조사로 분류된 것으로 보인다. 또한 박영실· 정남수 (2008)은 최종단계의 전화조사가 CATI를 통해 수행되었는지 언급하고 있지 않는데 <표 2-14>를 토대로 판단할 때 조사원에 의해 전화조사가 직접 수행된 비전자식 전화조사(PATI)의 형태인 것으로 추측된다.

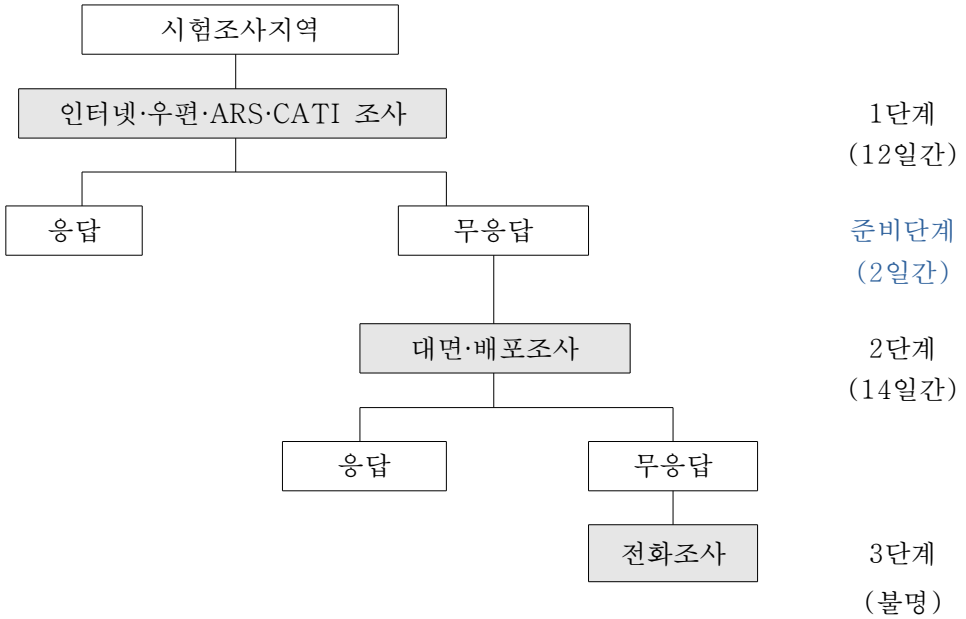
17) Lohr (2010, 8.2절)은 조사설계와 무응답 추적조사에 할애되는 시간이 부족할 때 조사의 질이 좋지 않을 수

전평가를 위한 시험조사이지만 좀 더 긴 시간을 할애할 수 있다면, 보다 충실한 추적조사를 가능케 하여 양질의 조사가 가능할 수 있을 것이다.

표본명단 내 조사대상자들에 대한 우편주소, 이메일주소, 전화번호 등의 접촉정보 실태에 대해 상세한 보고를 추가된다면 조사모드의 차이에 따른 포함범위에 대해 보다 잘 알 수 있을 것이다. 또한 최초 접촉에 사용된 모드와 단계별 모드선택의 주체나 방식, 그리고 모드별 무응답 규모 및 특성을 알 수 있다면, 혼합모드의 사용에 따른 선택효과를 파악할 수 있을 것이다. 더 나아가 모드별 응답자들의 인구·사회학적 특성을 통해서도 선택효과를 파악할 수 있을 것이다.

측정효과의 측면에서는 박영실·정남수 (2008)에서 제시된 정도의 자료 하에서는 조사과정과 심리작용에 따른 세부적인 영향력들을 제거하지 못하고 응답모드에 따른 특성치 비교를 통해 다소 불안전한 분석만이 가능할 것으로 판단된다.[18)]

<그림 5> 2010 인구주택총조사 2차 시험조사의 단계별 조사모드방식



[참고: 박영실·정남수(2008)의 표 2-13을 재구성함.]

있음을 지적하고 있다.

18) 각주 8참고.

<표 7>은 단계·조사모드별 (응답)가구수 및 가구비율을 나타내고 있다. 먼저, 단계별 가구비율을 살펴보면 1단계 비방문 조사에서는 27.0%, 2단계 조사원 방문 조사에서는 68.8%, 3단계 추적조사에서는 1.6%이었다. 단계별 응답(전환)률(conversion rate)로 환산해 보면 2단계와 3단계는 각각 94.2%와 61.5%으로 직전단계의 무응답가구에 대한 추적조사임에도 조사원방문조사는 물론 전화추적에 의한 전화 및 기타 조사모드에서도 매우 높은 응답률을 보여주고 있다. 이는 방문 혹은 전화를 통한 적극적인 조사원의 조사개입이 높은 조사의 성공으로 이어짐을 나타낸다고 할 수 있다. 단계내 조사모드별로 조사 참여 비율을 살펴보면, 1단계 비방문조사에서는 우편조사 82.8%, 인터넷조사 14.1%, CATI 조사 1.7%, ARS조사 1.4%의 순으로 나타났다. 2단계 조사원방문조사의 조사 참여 비율은 대면조사 95.8%, 배포조사 4.1%의 순이고, 3단계 추적조사에서의 전화조사와 기타 조사모드별 조사 참여 비율에 대한 자료는 주어지지 않았다.

<표 7> 단계·조사모드별 (응답)가구수 및 가구비율

단계	조사 모드	(응답)가구수			(응답)가구비율(%)		
		전체	전수	표본	전체	전수	표본
1단계 비방문 조사	소계	7,575	6,894	681	27.0	27.3	24.3
	인터넷	1,097	971	126	3.9	3.8	4.5
	우편	6,260	5,710	550	22.3	22.6	19.6
	ARS	99	99	-	0.4	0.4	-
	CATI	119	114	5	0.4	0.5	0.2
2단계 조사원 방문	소계	19,323	17,296	2,027	68.8	68.4	72.4
	대면	18,550	16,572	1,978	66.1	65.6	70.0
	배포	773	724	49	2.8	2.9	1.8
3단계 전화추적 조사	소계	1,179	1,093	91	4.2	4.3	3.3
	전화(기타)	723	675	53	2.6	2.7	1.9
	무응답	456	418	38	1.6	1.7	1.4
총계		28,082	25,283	2,799	100.0	100.0	100.0

[참고: 박영실·정남수(2008)의 표 2-15와 표 2-18을 재구성함.[

<표 8> 단계별 조사모드 응답가구수 구성비 및 응답전환률

단계	조사 모드	단계별 구성비(%)			응답전환(무응답)률(%)		
		전체	전수	표본	전체	전수	표본
1단계 비방문 조사	소계	100.0	100.0	100.0	27.0	27.3	24.3
	인터넷	14.5	14.1	18.5	-	-	-
	우편	82.6	82.8	80.8	-	-	-
	ARS	1.3	1.4	-	-	-	-
	CATI	1.6	1.7	0.7	-	-	-
2단계 조사원 방문	소계	100.0	100.0	100.0	94.2	94.1	95.7
	대면	96.0	95.8	97.6	-	-	-
	배포	4.0	4.2	2.4	-	-	-
3단계 전화추적 조사	소계	100.0	100.0	100.0	-	-	-
	전화(기타)	61.3	61.8	58.2	61.3	61.8	58.2
	무응답	38.7	38.2	41.8	(38.7)	(38.2)	(41.8)

[참고: 박영실·정남수(2008)의 표 2-15와 표 2-18을 재구성함.]

5.2 효율적 혼합모드의 사용을 위한 실험설계 방향 논의

조사모드의 선택은 2장에서 살펴본 것 같이 비표본오차의 다양한 측면에 영향을 준다. 어떠한 모드를 선택하느냐에 따라 목표모집단에 대한 포함정도와 접근할 수 있는 표본개체들의 응답률에 영향을 주게 되며, 궁극적으로는 최종 응답개체가 갖는 모집단에 대한 대표성을 결정짓게 된다. 또한 조사모드의 종류에 따라 응답자료에도 영향을 끼친다. 이러한 결과들은 4장에서 살펴본 선택효과와 측정효과에 기인한다고 할 수 있다. 혼합모드를 사용하면 모드 종류에 따라 포함정도, 응답률, 응답자 구성이 달라지며, 이는 무응답 편향의 형태로 인식될 수 있을 것이다. 또한 혼합모드로 조사할 때 각 모드별로 상이하게 발생하는 응답자의 심리작용으로 말미암아 동일한 질문에서 조차도 모드별로 상이한 값을 관측하게 되는 관측오차가 발생할 수 있다. 따라서 혼합모드의 사용으로 인해 발생할 수 있는 모드효과, 즉 선택효과와 측정효과를 잘 이해할 수 있다면, 보다 더 효과적인 혼합모드의 사용이 가능할 것이다.

외국의 국가통계기관들에서는 조사절차의 변경이 응답률 혹은 모수추정 등에 어떠한 영향을 미치는지를 평가하기 위해 무작위실험(randomized experiment)을 종종 고려한다. 4.4절에서 살펴본 유럽사회조사의 실험설계에 의한 조사들도 이러한 노력의 일환으로, 기존의 조사원면접 방식의 조사모드에서 전화조사와 인터넷조사를 추가로 사용하는 혼합모드방법으로 변경할 때 발생하는 모드효과에 대한 평가를 위해 고려되었다. J□ckle *et al.* (2010)의 첫 번째 모드실험은 측정효과를 평가하기 위해 표본추출 및 포함오차에 의한 교락작용을 통제하기 위하여 인구·사회학적 특성에 대한 인구비례에 맞추어 참여자를 선별하였고 이들을 모드(별) 실험군으로 할당시킨 후 동일한 질문으로 실험실(회의장)에서 조사를 수행하였다. 이와 같은 적극적 실험조건에 대한 통제는 관련한 효과를 평가하는데 외부적 영향력을 제거하여 좋을 수는 있으나 현실적 상황이 반영되지 않는 단점을 갖는다.

J□ckle *et al.* (2010)의 두 번째 모드실험은 표본선택 대상자의 포함범위를 통제하기 위해 고려되는 모드에서 동일하게 접근할 수 있는 대상자 (예, 주소와 전화번호가 동시에 있는 가구)들 만을 선택대상으로 통제한 뒤, 선택된 대상자를 무작위로 모드(별) 실험군에 할당하여 조사를 수행하였다. 따라서 첫 번째 모드실험에 비해 실험결과에 표본조사의 현실적 상황을 반영할 수 있도록 하기 위해 실험계획의 요인통제를 완화하였다. 하지만 두 가지의 모드실험은 모두 유럽사회조사 본조사에 내장된(embeded) 형태로 수행하지는 않았다.

Vannieuwenhyuze *et al.* (2010)의 모드실험은 J□ckle *et al.* (2010)의 두 번째 모드실험과 매우 유사한 형태로 진행되었다. 본 조사와 더불어 비교가능한 단일모드조사를 독립적으로 추가하여 모드효과를 좀 더 효과적으로 분해하고자 하였다.

간단히 정리하면, 비록 현실성을 덜 반영하지만 외부효과를 적극적으로 통제하도록 하여 실험실적 조사결과로부터 해당 효과를 정확히 평가할 수 있는 모드실험을 고려할 수도 있고, 좀 더 현실성을 반영하고자 소극적 통제만을 고려한 본래의 표본설계에 접목한 실험설계를 고려할 수도 있을 것이다. 어떠한 접근을 채택하더라도 가중치 조정을 통한 무응답 혹은 대상자 선택편향으로 인한 모집단에 대한 대표성의 강화를 위해 모드에 덜 민감한 인구·사회학적 변수도 함께 수집하여야 할 것이다.

더불어 조사 현실성을 감안한 모드실험의 경우에는 특히 조사응답과정에 대한 인지모형에 근거하여 최초 접촉과 이후 응답과정을 파악하고 조정할 수 있도록 접

촉률, 응답률, 적격률 등의 조사평가는 물론 무응답 특성의 파악이 가능하도록 자료 수집과정자료(paradata)를 수집하고 적절한 모형적 분석도 수반되어야 할 것이다

참고문헌

박영실·정남수 (2008). 2010 인구주택총조사 자료수집방법 비교분석 – 대면·인터넷·우편조사를 중심으로 –, 통계개발원 연구보고서 1--101.

박영실·정남수 (2009). 경제활동인구조사의 자료수집방법 비교분석, 통계개발원 연구보고서 1--40.

임경은·박라나 (2013). 혼합모드조사 추정방법 실무적용방안 검토 – 사교육조사를 중심으로–. 통계개발원 연구보고서. 1--63.

박시내·최유성·한승훈 (2014). 경제활동인구조사의 자료수집방법 별 효과 분석: 응답자의 자료수집방법 선택효과를 중심으로. 채택.

AAPOR. (2011). Standard definitions: Final dispositions of case codes and outcome rates for surveys, 7th edn. Tech. rep., The American Association for Public Opinion Research, Deerfield, IL, URL http://www.aapor.org/pdfs/standarddefs_4.pdf.

Bethlehem, J., Cobben, F., Schouten, B. (2011). Handbook of Nonresponse in Household Surveys. Wiley, New York.

Biemer, P.P. and Lyberg, L.E. (2003). *Introduction to Survey Quality*. Wiley, New York.

Brick, J.M. (2013). Unit Nonresponse and Weighting Adjustments: A Critical Review, *Journal of Official Statistics* **29**, 329—353.

Buelens, B., Van den Brakel, J. (2011). Inference in Surveys with Sequential Mixed-Mode Data Collecition. Discussion Paper. Statistical Netherlands.

Cannell, C., Miller, P., and Oksenberg, L. (1981). Research on Interviewing Techniques. in Leinhardt, S. (ed.), Sociological Methodology 1981, 389—437, San Francisco: Jossey-Bass.

Curtin, R., Presser, S., and Singer, E. (2000). The Effects of Response Rate Changes on the index of Consumer Sentiment. Public Opinion Quarterly, **64**, 413--428.

Cobben, F. (2009). Nonresponse in Sample Surveys Methods for Analysis and Adjustment. Ph.D. thesis. Universiteit van Amsterdam.

de Leeuw, D.D. (2005). To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics* **21**, 233--255.

Groves, R.M., Lowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., and Tourangeau, R. (2009). Survey Methodology, 2nd Ed., Wiley: New Jersey.

Groves, R.M. (1989). *Survey Errors and Survey Cost*. New York: Wiley.

Hansen, M.H., Hurwitz, W.H. (1943). On the Theory of Sampling from Finite Populations. *Annals of Mathematical Statistics,* **41**, 517—529.

Hox, J.J., and De Leeuw, E.D. (1994). A Comparison of Nonresponse in Mail, Telphone and Face-to-fact Survey. Applying Multilevel Modeling to Meta-analysis. Quality and Quantity, **28**, 329—344.

J□ckle, A., Roberts, C., Lynn, P. (2010). Assessing the Effect of Data Collection Mode on Measurement. *International Statistical Review*, **78**, 3--20.

Kalton, G., and Maligalig, D.S. (1991). A Comparison of Methods of Weighting Adjustment for Nonresponse. *Proceedings of the US Bureau of the Census Annual Research Conference*, 409--428.

Lepkowski, J.M, Mosher, W.D., Groves, R.M., West, B.T., Wagner, J., Gu, H., (2013). Responsive design, weighting, and variance estimation in the 2006-2010 National Survey of Family Growth. National Center for Health Statistics. Vital Health Stat 2(158).

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). Hoboken, New Jersey: Wiley Series in Probability and Statistics.

Platek, R. (1977). "Some Factors Affecting Non-Response," *Survey Methodology,* **3**, 191-214.

Roberts, C. (2007). *Mixing Modes of Data Collection in Surveys: A Methodological Review.* NCRM review paper, 8. ESRC National Centre for Research Methods, Swindon, UK.

Rosenbaum, P., Rubin, D.B. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects", Biometrika, **70**, 41-55.

Schouten, B., Cobben, F., Bethlehem, J. (2009). Measures for Representativeness of Survey Response. *Survey Methodology*, 35, 101—113.

Suzer-Gurtekin, Z. T., Heeringa, S. G., Valliant, R. (2012). Investigating the Bias of Alternative Statistical Inference Methods in Sequential Mixed-Mode Surveys. Proceedings of the Survey Research Methods Section, American Statistical Association.

Valliant, R., Dever, J.A., Kreuter, F. (2012). *Practical Tools for Designing and Weighting Survey Samples*. Springer: New York.

Vannieuwenhuyze, J., Loosveldt, G., Molenberghs, G. (2010) A Method for

Evaluating Mode Effects in Mixed-Mode Surveys. *Public Opinion Quarterly*, 74, 1027--1045.

Vannieuwenhuyze, J., Loosveldt, G., Molenberghs, G. (2012) A Method to Evaluate Mode Effects on the Mean and Variance of a Continuous Variable in Mixed-Mode Surveys. *International Statistical Review*, 80, 306--322.