해외출장보고

JSM(Joint Statistical Meeting) 2004
통계회의 참가 보고

(Toronto, Canada August 8~12, 2004)

2004. 9

조 사 관 리 과

목 차

JSM 2004 통계회의 참가

(Joint Statistical Meeting 2004)

1. 출장 목적

○ 현재 연구중에 있는 소지역추정기법, 무응답처리기법 등의 통계적 추정기법과 관련하여 선진국의 최신동향 및 기법을 파악하여 개발업무에 활용하고자 함

2. 출장기간 및 수행사항

가. 출장기간 : 2004. 8. 7. ~ 8. 14.(8일간)

나. JSM(Joint Statistical Meeting) 통계 회의 참가

- 기간 : 8. 8. ~ 8. 12.(5일간)
- 장소 : 캐나다 토론토 컨벤션 센터

3. 출장자

○ 조사관리과: 허남거과장(4급), 윤연옥(4급)
○ 통계연구과: 정동명(5급)

4. JSM 통계회의 주최기관

○ American Statistical Association(ASA)
○ Statistical Society of Canada
○ International Biometric Society-Eastern North American Region(ENAR)

○ International Biometric Society-Western North American Region(WNAR)
○ Institute of Mathematical Statistics(IMS)

5. JSM 회의 개요

○ JSM은 북미에서 매년 8월에 개최되는 통계인들의 모임으로 약 4,000명이 참석
※2005년 JSM 회의는 미국 미네소타, 미네아폴리스(Minneapolis, Minnesota)에서 8월 7~11일에 개최됨

○ 2004년 회의는 전체 443개의 세션(Session)으로 구성되어있으며, 각 세션에서는 5~7개의 논문이 1시간 50분에 걸쳐서 진행
- 그 중 조사연구 및 정부통계와 관련된 세션은 약90개임
- 주요관심 분야
· 소지역추정기법(Small Area Estimation)
· 무응답처리기법(Imputation, Non-response Adjustments)
· 가중치 및 추정기법(Weighting and Estimation for Surveys)
· 조사분석(Survey Analysis)
· 분산추정(Variance Estimation)
· 정부통계조사(Government Surveys)
· 표본설계(Sample Designs)
· 통계조사의 새로운 기법들(New Technologies in Surveys)
· Data Quality and Data-Confidentiality of Microdata 등

○ JSM 회의는 ASA(American Statistical Association)회의에서 발전되었으므로 현재도 ASA 회의라고도 구두로 이야기 함.

6. 회의참가 소감

○ JSM 회의는 우선 규모면에서 ISI 회의와 비교하여 거의 2배 정도 되는 회의여서인지 그 내용 및 참가자도 훨씬 다양하였음. 북미 뿐만 아니라 중국, 인도 등 아시아권, 유럽 등에서도 많이 참석하여 세계 각국에서 온 사람들로 북적거렸음. JSM 회의는 공식통계와 관련된 연구 및 프로젝트는 물론이고 학문적인 면도 많이 발표되었음. 공식통계와 관련된 발표에는 많은 사람들이 참석하였는데, 어떤 세션은 발표자 모두가 Census Bureau에서 참석한 사람들로 구성되었음.

○ JSM 회의장에서 발표된 다양한 내용 중에서 소지역추정기법 및 무응답처리방법에 관한 내용은 많은 분야에서 진행되고 있었음. 특히 미국 인구주택총조사의 표본조사(longform)를 대치하기 위하여 개발된 ACS(American Community Survey) 조사의 소지역추정기법은 다방면으로 진행되고 있었으며, 소득관련된 조사인 SAIPE(Small Area Poverty and Income Estimates) 조사에서도 소지역추정기법이 많이 적용되고 있었음. 무응답에 관한 처리방법은 거의 모든 통계조사에서 다루고 있는 내용이었음.

○ 회의에 참석하여 발표되고 있는 내용을 모두 100% 이해 할 수는 없지만, 현재 어떠한 방향으로 진행되고 있으며 앞으로는 어떻게 흘러갈 것인가를 이해하고 우리의 현실에 적용할 수 있는 방향을 모색할 수 있는 것이 바로 회의참가에서 얻을 수 있는 것이 아닌가 생각됨.

- Rolling Census인 ACS 조사만 하여도 현재 우리가 실행하고 있는 인구주택총조사에 대한 앞으로의 개선점을 시사하고 있었음. 현재 우리나라에서는 5년마다 인구주택총조사를 실시하고 있는데 만약 Rolling Census를 실시하게 되면 10년마다 전수항목을 중심으로 한 총조사를 실시하여도 되지 않을까하는 생각을 갖게됨.

○ 앞으로 우리 통계청이 통계선진국이 되기 위해서 가장 신경써야 될 부분이 있다면 추정기법과 관련된 내용이라고 생각된다. 이용자에게 만족을 주는 통계, 신뢰도 높은 통계, 바로 그런 통계를 생산하기 위해서는 추정기법과 관련된 새로운 방법이 연구되어야 하고, 그러기 위해서는 국제통계회의, 세미나, 워크샵 등 에 많이 참석해서 정보를 가져오고, 가져온 정보를 연구하여 적용하여야 할 것임.

○ 현재 연구중인 소지역추정기법 및 무응답처리기법 관련하여 발표된 논문의 자세한 내용은 저자에게 직접 메일로 연락하거나 2005년 1월에 발간되는 논문집을 참조하여 업무에 도움이 되도록 할 계획임

7. 주요 발표논문 요약

○ 발표된 논문들은 2005년 1월에 CD-Rom 으로 발간 예정
○ 참가한 주요 관심분야 발표논문은 다음과 같음

□ Weighting Alternatives to Compensate for Longitudinal Nonresponse in the Survey of Income and Program Participation
(SIPP 소득조사에서 무응답자료을 대치하기위한 가중치 조정법)
- Leroy Bailey

SIPP(Survey of Income and Program Participation)소득조사의 횡단면적 추정에서 무응답에 대한 처리를 위하여 승수조정 칸을 이용한 방법을 현재 사용하고 있다. 이 논문에서는 SIPP 횡단면 추정에서 선택된 무응답에 대한 영향을 연구하였다. 이 논문에서 중점적으로 다룬면은 무응답 모형에 관한 것이며, SIPP 주요 항목에 대한 자료를 바탕으로 승수조정에 대한 일반화된 방법들, 경험적인 비교 및 절차에 평가 등이 이루어졌다.

☐ Classification of Address Register Coverage Rates- A Field Study

(주소등록대장부의 포함률에 대한 연구)

- Gavin Tompson, C. Turmelle

매 10년마다. 캐나다 노동력조사(Canadian Labour Force Survey, CLFS)는 표본 재설계가 이루어진다. 이번에는 캐나다 주소 등록대장(Canadian Address Register(CAR))을 표본추출 등록대장으로 대치하는 것을 고려하였다. 등록대장을 만들기 위해서는 비용이 많이 든다. 캐나다 노동력조사는 2단계 디자인으로 설계되었다. 1단계 표본추출단위인 PSU(Primary Sampling Unit)는 모든 거주 가구들이 있는 자료로부터 선택되며, 최종 표본 추출을 추출하기 위해서 재정리되어야 한다. PSU 리스트는 일반적으로 조사원들이 PSU 내 전체를 돌면서 각 거주가구에 대해 정리한다. 주소등록부를 사용하게 되면 PSU 내 거주가구를 정리하는 단계를 없애주거나 거주가구 리스트 작성과정에서 보다 정확한 리스트를 만드는데 도와준다. 어떤 PSU는 주소등록부를 사용하기에 너무 자료가 부실한 경우도 있다. 5개 시도의 55,000 가구를 대상으로 2003년 12월에 현장 조사하였다. 이 논문에서는 현장조사를 어떻게 하였으며, 주소 등록부의 과대 및 과소 비율을 검토하고 두 방법에 의한 결과를 비교하였다.

☐ An Alternative to the Principal Person Method for Weighting in the American Community Survey

(ACS 조사의 다른 가중치에 대한 연구)

- Keith Albright, A. Navarro, M. E. Asiala

American Community Survey(ACS)조사는 미국 센서스 국에서 매월 실행되는 인구주택 센서스조사의 표본조사(census long form)를 대치하는 조사이다. ACS 조사에 대한 검증은 1996년에 시작되었으며 현재 36개 카운티에서 실행되고 있다. ACS 추정에 관해 제기되는 관심은 현재 거주하고 있는 주택에 대한 추계가 주택소유로부터 추계한 수치와 일치하지 않는 다는 것이고, 결혼한 남자의 추정값이 결혼한 여자의 추정값과 일치하지 않는 다는 것이다. 그 이유는 ACS 조사에서 사용하는 승수(weighting)는 principal person method 이다. ACS 승수의 또다른 방법은 뉴욕시 빈주택 조사(New York City Housing Vacancy Survey, NYCHVS)에서 사용된 승수조정법이다. 이방법은 실제 거주하는 주택에 대한 추계치와 주택 소유자들에서 추계한 수치와 일치한다. 또한 기혼여자의 추정치와 기혼남자의 추정치가 일치한다. 이 논문에서는 NYCHVS 승수조정방법을 다른 ACS 주택유형 및 인구추정에 대한 영향을 조사하였다.

☐ Obtaining Stratum Breaks in Skewed Populations Using a Simple Method

(기울어진 모집단에서 층 분할에 방법 연구)

- Patricia M. Gunning, J. M. Horgan, G. Keogh

Dalenius(1950)은 연속변수에서 층화를 하고자 할 때, 추정분산을 최소화할 수 있는 지점에서 경계선을 결정하는 최적분배에 관한 수식을 유도하였다. 이 수식에 대한 정확한 해답은 존재하지 않고 최적층화에 대한 실행은 일반적으로 근사적인 방법으로 해결하고 있다. 이 논문에서는 오른쪽으로 기울어진 모집단에서 층화 경계선에 대한 새로운 알고리듬을 개발하였는데 이 방법은 현재 방법보다 실제 적용하기기 훨씬 쉽다. 이 알고리듬은 코크란(Cochran, 1961)이 주장한 최적 경계점을 가지면 변이계수는 모든 층에서 거의 같게된다는 이론에 기본을 두고 있다. 모집단이 오른쪽으로 기울어졌을 때 , 각 층의 경계점은 기학학적인 접근으로 얻을 수 있다. 이 방법을 4개의 실제 자료에 적용하여 일반적으로 많이 적용하는 Dalenius and Hodges(1957)에 의한 누적도수에 루트를 취하는 방법 및 Lavallee and Hidiroglou 에 의한 기울어진 분포의 평균과 총계를 이용하는 방법과 비교하여 좋은 결과를 나타내었다.

☐ Does Weighting for Nonresponse Increase the Variance of Survey Means?

(무응답 가중치 조정방법은 분산 증가를 시키는지?)

- Sonya L. Vartivarian, Roderick J. Little

무응답 승수조정은 통계조사에서 단위무응답를 처리하는 일반적인 방법이다. 일반적으로 알려진 견해는 승수조정방법은 무응답 바이아스는 줄여지나 분산은 증가하는 효과가 있다는 것이다. 그래서 승수조정법의 효율은 바이아스와 분산의 거래라고 이야기 할 수 있다. 이것으로 결론내릴 수 있는 것은 무응답 승수조정법은 바이어스 뿐만 아니라 분산에서도 감소를 유도할 수 있다는 것이다. 승수조정법의 공변수(covariate)는 무응답 바이어스를 감소시키기 위한 두가지 특성을 가지고 있다. 이것은 응답확률 및 조사결과와 연관되어있다. 만약 조사결과와 연관되어있으면, 승수는 표본분산도 줄일 수 있다. 바이어스와 분산에 대한 자세한 분석과 시뮬레이션은 수정칸에 기초한 평균추정방법을 이용하여 분석되었다. 분석된 내용에 의하면 승수조정에 포함되는 가장 중요한 변수의 성질은 조사결과의 예측이고, 두 번째는 응답예측이다.

☐ Two-stage Nonparametric Approach for Small-area Estimation

(소지역추정에 대한 2단계 비모수적 접근법)

- Pushpal Mukhopadhyay, Tapabrata Maiti

소지역추정은 일반적으로 연관되는 다른 지역들로부터 정보를 이용한다. 이러한 간접적인 추정량들은 보충자료를 통해 소지역을 관련시키는 모형을 사용한다. 다양한 단위수준과 지역수준에서의 소지역 모형들이 제시되지만, 이러한 모형들은 대부분 소지역의 평균이 부가적인 정보와 선형관계가 있다고 가정을 한다.

이 논문에서는 소지역평균에서 Nadaraya-Watson kernel에 기초한 지역수준 비모수적 회귀추정량을 제안한다. 또한 Prasad와 Rao(1990)에 의해 제안된 2단계 추정법을 채택한다. 제안된 추정량의 점근적인 성질들이 연구되고, 2단계 추정량의 2차수 평균제곱예측오차(MSPE)와 MSPE의 추정량이 정규성하에서 얻어진다. 마지막으로 제안된 추정량의 우수성을 보여주기 위해 모의실험을 실시한다.

☐ New Developments in SAIPE County Median Household Income Models

(SAIPE 카운티 가구소득 중앙값추정의 새로운 방법)

- Geoffrey M. Gee

미국 Census Bureau 소지역 빈곤 및 소득추정(Small Area Poverty and Income Estimates, SAIPE) 프로그램은 주, 카운티 및 학교구역의 빈곤 및 소득 정보를 제공하고 있다. 최근의 SAIPE 카운티 가구소득 중앙값추정(MHI, Median Household Income)에 관한 연구는 두 영역에 중점을 두고 있다. SAIPE 카운티 MHI 모델의 종속변수는 제공되는 해를 중심으로 3년간의 CPS(Current Population Survey) 자료를 쓰고 있다. 3년간에 걸친 자료를 사용하는 근본이유는 대부분의 카운티에서 한해의 CPS 표본규모는 아주 작기 때문이다. 인근 해의 자료을 추가하여 분석하는 것은 분산추정은 개선하나 시점별로 소득이 다름으로 인한 측정할 수 없는 에라를 포함하고 있다. 주별 어린이건강 보험 프로그램(State Children's Health Insurance Porgram, SCHIP)에 대한 대응으로써 CPS 표본규모가 크게 증가되었는데, 이는 CPS 한해 자료를 바탕으로 하여 SAIPE 카운티 MHI 추정을 하는 것은 불안정할 수 있다는 가능성 때문이다. 이 논문에서는 개발된 2개의 대체모형에 대하여 설명되어있다.

☐ A Study of Mass Imputation in Small-area Estimation

(소지역추정에서의 Mass 대체법에 대한 연구)

- Nancy Robbins, Richard Moore

매 5년마다 실시되는 사업소유주의 조사(SBO)는 흑인이나, 라틴계인, 아시아-태평양 지역 섬주민, 미국 인디언 토착민들과 여성들이 소유한 사업체에 대한 기본적인 경제 통계를 제공하는 가장 포괄적인 조사이다. 이 조사는 총 사업체 수, 영수증, 급료 지불 명부에 대한 정보와 소수의 자영업자, 합명(자)회사, 법인회사 등의 고용에 대한 정보를 제공한다. SBO가 2자리 숫자의 표준산업분류에 의해 인종에 대한 믿을만한 추정치를 제공하기 위해 설계된 반면, 좀더 세부적인 지리적, 산업적 수준에서 추정을 위한 것들이 요구되어진다. 조사에 의한 직접추정량은 표본이 이 수준들에서 잘 대표하지 않기 때문에 적절하지 않다.

이 논문에서는 무응답이나 비표본인 경우 완전한 모집단이 구성되도록 자료가 대체되어지는 Mass 대체의 방법을 제안했다. 이 방법에 의한 추정치는 소지역에서의 직접적인 추정치와 비교했을 때, 그 결과들은 고무적이다.

☐ Multiple Imputation of Missing Income Data in the National Health Interview Survey

(국민건강면접조사에서 결측 수입자료의 다중대체법)

- Nathaniel Schenker, Diane Makuc, et al.

국민건강면접조사(NHIS)는 건강상태와 건강주의증대와 활용을 연구하기 위한 주요한 자료를 제공하는 가구조사이다. NHIS에서 항목에 대한 무응답은 비록 낮지만 많은 분석에서 주요한 변수로 사용되는 연간 가구수입에 대해서는 높은 편이다. 예를 들어 2001년에는 29% 정도의 가구들이 가구수입을 보고하지 않았다. 1997년 NHIS부터는 조사되지 않은 가구수입과 개인소득에 대해서 다중대체법이 실시되고 있다. 대체를 위한 5개의 대체값 집합들은 축차회귀다중대체법의 사용으로 매년마다 만들어지고 있다. 1997년부터 2001년까지 수입에 대해 대체된 값을 포함하여 NHIS에 대한 응용이 언급되어 있다.

이 논문에서는 결측된 수입자료의 유형과 1997년부터 2001년까지 NHIS의 수입에 대한 다중대체법을 설명한다. 건강상태와 건강주의증대에 대해 다중대체법으로 분석한 것을 단일대체법에 의한 분석과 응답하지 않은 자료는 제외하고 분석한 것과 서로 비교한다.

☐ Response Rates and Nonresponse in BLS and Census Bureau Establishment Surveys.

(BLS와 CB의 사회시설조사에서의 응답률과 무응답)

- Rita J. Petroni, Stephan Cohen, et al.

Bureau of Labor Statistics(BLS)와 Census Bureau(CB)는 조사과정의 여러 단계에서 응답수준을 측정하거나 응답자로부터 얻은 자료와 행정기록 등과 같은 검증된 대안 자료를 합하여 추정한 결과의 정도를 측정하는 의미있는 방법들로부터 조사응답률을 정의하는데 노력하고 있다. 이들 기관내 그리고 기관간의 의미있고 상당히 정의된 응답률의 추세에 대한 평가는 개선이 되어졌고 또한 되어질 필요가 있는 정확한 목표를 정하는데 도움을 줄 수 있다.

이 논문에서는 두 기관의 사회시설조사의 응답률의 정의와 응답률의 추세, 대행자의 응답률과 응답을 장려하는 방법간의 차이에 대한 설명, 그리고 무응답을 줄이는 연구 등이 설명되어 있다.

☐ Imputation by Propensity Matching

(Propensity matching에 의한 대체법)

- Murthy N. Mittinty, Easaw Chacko

결측자료는 흔히 발생되는 현상이다. 많은 조사기관들은 결측자료를 다루기 위해 최근방 이웃대체(NNI)와 같은 단일 대체법을 사용한다. NNI의 장점은 공변량의 정보를 이용할 수 있고, 또한 NNI에 의해 대체된 자료로부터 얻어진 점 추정량은 편향이 줄어든다는 것이다. 다변량에서 NN을 찾는 것은 match시키기 위해 모든 변수가 필요하기 때문에 복잡하다.

이 논문에서는 관찰의 연구에 대해 Rosenbaum과 Rubin(1983)이 소개한, 자료가 다변량일 때 기증자를 찾기 위한 propensity matching의 사용에 대해 연구한다. Propensity score에 의한 NNI(NNPS)는 선형과 볼록곡선인 경우 임의로 결측되는 자료를 이용한 모의실험을 사용하여 조사되어진다. NNPS를 사용함으로써, Propensity score가 주어진 공변량의 조건부분포는 응답자과 무응답자에 대해 같다는 것을 확신한다. 또한 NNPS를 Murthy 등이 소개한 차이행렬(NNDM)을 이용해 NNI와 비교한다. 주변분포를 유지하기 위해서 NNDM을 사용한다. 그 결과들은 NNPS에 의해 대체된 자료에서 구한 추정값들은 흔히 NNDM의 추정값들과 거의 같으며 차수의 유해성이 줄어든다는 것을 나타내고 있다.

☐ A Study of an Optimization Method Used for the Planning of a Complex Multivariate Survey

(복잡한 다변량조사의 기획에서 최적방법사용의 연구)

- Anders Holmberg

조사의 기획 단계에서 매우 중요한 것은 표본설계의 선택이다. 이 논문에서는 조사가 여러 가지의 목적을 가지고 있고 조사모집단이 층화되어 있을 때, 최적의 표본설계를 찾기 위한 방법의 예비적인 연구를 보여주고 있다. 고전적인 기획 문제를 해결하기 위해, 즉, 주어진 비용으로 추정량의 정도를 가급적 높게 하는 층을 찾기 위해, 비선형 최적문제가 공식화된다. 최적에 대한 문제는 조사단위의 포함확률을 결정하고, 어떠한 정도의 제약하에 추정량의 분산함수를 최소화하도록 층별로 표본을 배분하는 것이다. 여기서는 층내나 층간에서 어떤 변수들의 제약이 있을 때, 그 방법을 설명하고 이들의 성질을 연구한다.

☐ Responsive Design for Household Surveys.

(가구조사에 대한 응답 설계)

- Steven G. Heeringa, Robert Groves

70년 이상 동안 표본추출 기술의 개발과 가구조사에 대한 자료수집 방법은 비용과 오차를 감소시키는 것에 초점을 맞춰왔다. 역사적으로, 조사 설계자들은 새로운 설계에서 오차와 비용구조를 모형으로 만들기 위해 이전의 조사에서 얻은 경험에 많이 의존해 왔다. 정보가 조사기획과정에서 나온다는 장점에도 불구하고, 많은 조사 설계들은 기본적 조사조건에서 불확실을 예상하는 특징을 포함하지 않는다. 아주 적은 경우에만 조사 설계 모수나 또는 자료수집 비용을 추정하는데 있어 주요 실패요인이 되는 위험한 응답을 가능한 제외한 실시간자료에 응답하기위해 설계되어 있다. 자료수집과 조사관리시스템을 컴퓨터로 처리하기위한 변환은 조사 설계자와 처음 시작하는 자에게 조사과정에서 새로운 실시간 자료의 가치로 보여준다. 이러한 자료는 작업과정에서의 자세한 자료와 서로 다른 추출단위, 개별 면접요원, 그리고 심지어 개별 표본에 대한 관련된 비용들이 포함되어 있지 있다. 이 논문에서는 조사설계와 관리에 대한 새로운 모형을 보여준다.