2008년 표본조사와 베이지안 통계(SSBS)회의
출장결과 보고

2008. 9.

통계개발원

2008년 표본조사와 베이지안 통계(SSBS)회의 출장결과 보고

Ⅰ. 출장개요

1. 출장목적

○ 사우스햄튼 통계연구소(S3RI: Southampton Statistical Sciences Research Institute)가 개최하는 "2008년 표본조사와 베이지안 통계(SSBS: Sample Survey and Bayesian Statistics) 회의"에 참석하여 표본조사에서의 표본설계, 무응답 조정, 복합 및 합성 추정 등에 광범위하게 응용되고 있는 베이지안 방법론에 대한 최근 연구 동향을 파악하고자 함

2. 출장기간

○ 2008. 8. 25. ~ 8. 31. (5박7일간)

3. 출장자 및 출장지역

소 속	직 급	성 명	출장지역
연구기획실	통계주사 통계주사	이 현 정 정 미 옥	영국 (사우스햄튼)

4. 주요일정

일 시	방 문 지	비고
- 8월 25일	서울(인천공항) → 영국	
- 8월 26일~8월 29일	회의참석 - 장소 : 사우스햄튼 대학	
- 8월 30일~8월 31일	영국 → 서울(인천공항) 도착	

Ⅱ. 회의개요

1. 회의명

○ 2008년 표본조사와 베이지안 통계(SSBS) 회의

2. 회의목적

○ 베이지안 추정과 관련한 기법 소개
○ 실제 자료를 활용한 무응답 적용사례 소개
○ 소지역 추정에 대한 사례 등

3. 회의일시 : 2008. 8. 26.~ 8. 29.(4일간), 영국 사우스햄튼대학교

4. 회의주기 : 매년 개최

5. 참가자 : 약 80명

6. 세부 주제별 발표내용

□ Topic 1 : 사례연구

1) 개요
- 다단계 모델 및 다단계 표본조사에 대한 각국의 사례 발표

2) 세부 발표주제
- 베트남에서의 다단계 모델과 삶의 표준
- 미국 동부에서 발전된 오존집중에 관련한 시공간 내일의 일기예보
- 일상 음식물의 중금속 섭취량의 다단계 표본조사 연구
- 조사 무응답과 기업체의 기대특성-기업경향조사에 근거를 둔 분석

□ Topic2 : 표본추출

1) 개요
- 베이지안의 표본이론과 복합 표본 디자인 등에 대한 논의

2) 세부 발표주제
- Bowley의 베이지안 표본이론과 내용
- 집합단계에서 다른 자료 소스로부터 정보를 결합하는 베이지안 틀
- 유한혼합모델을 통한 복합표본 디자인의 계정
- 사전정보 표본추출 계획하에서의 베이지안 예상 추론

□ Topic3 : 표본설계방법에 따른 베이지안 추정

1) 개요
- 표본조사에 의해 수집된 자료가 광범위하게 추론에 활용되고 있음
- 그러나 표본설계 특성들이 종종 무시되고 전통적 방법에 의해 표본조사 자료가 분석되고 있고 이는 추론에 심각한 오류를 가져올 수 있음
- 따라서 설계방법에 따른 표본추출편향(selection bias)을 조정하기 위한 논의가 요구됨

2) 세부 발표주제
- 사전정보(informative) 확률 표본설계 하에서의 베이지안 추정기법
  ・사전정보 확률 표본설계에서 횡단면(cross sectional)자료와 종단(longitudinal)자료에 따른 추정기법 및 모형 소개
- 베이지안 사후 층화(post-stratification)를 이용한 표본추출 편향 조정방법
- 불균등 확률 표본추출(unequal probability sampling)에서의 평균과 분위수에 대한 베이지안 추정기법
- Record linkage에 대한 계층적 베이지안 모형

□ Topic4 : 무응답 및 다중대체법

1) 개요
- 무응답을 대체하는 데 있어서 단일대체법이 일반적임
- 그러나 베이지안에서는 다중대체법의 장점이 많음

2) 세부 발표주제
- 부분적으로 종합적인 자료의 모델 선택
- 패널자료에서 축차회귀 다중대체법의 적용
- 사업체 조사에서 유한 Gaussian 혼합법을 통한 다중대체법
- 무응답을 위한 베이지안식 접근

□ Topic5 : 소지역 추정

1) 개요
- 영역(domain) 추정에 대한 모델링 기법 소개 및 소지역 추정에 대한 각국의 사례 발표

2) 세부 발표주제
- 종단 자료의 직접 추정량의 평균과 분산에 대한 결합 모형·소영역의 인구 예측에 대한 적용 결과 논의
- 의료수술 수요에 대한 소지역 추정
  · 의료서비스와 자원을 공정하게 분배하기 위한 소지역별 수요를 추정했던 영국의 사례 소개
- 네덜란드 실업률의 EBLUP과 베이지안 소지역 추정값 비교 결과 논의

□ Topic6 : 통계적 노출 위험 평가

1) 개요
- SDC(Statistical Disclosure Control) 방법론에서의 노출위험을 평가하는 방법에 대해 서로 다른 통계주체들이 가지고 있는 관점의 차이점 제기

· 두 부분의 조화를 위해 방법론과 기술적인 부분에 대한 보강 필요

2) 세부 발표주제

- 베이지안 측면에서 노출위험평가에 대한 기초 이론을 소개
- 통계작성 기관(agency)과 침입자(intruder) 관점의 차이를 비교

Ⅳ. 회의결과

1. 주요 주제별 발표내용

- 베트남에서 다단계 모델과 삶의 표준
  · 이 보고서는 두개의 다단계 모델에 의해 1인당 실제 가구 지출 연구의 관점에서 베트남의 삶의 표준을 조사하기 위해 제안함
  · 2002년 베트남 가구의 삶의 표준에 대한 지표 데이터를 이용하면, 첫 번째는 나이, 가구주의 교육정도 등과 같은 수많은 가구의 특성들과 같은 종속변수로서 1인당 실제 지출의 대수로 다단계 모델을 구성함
  · 모델은 다음과 같은 4가지 레벨을 포함함: 가구(level 1), 공동체(level 2), 지방(level 3), 주(level 4)
  · 지정학적인 위치에 기인하여 삶의 표준은 다른 추정치를 산출하는 방해에 있어서 무작위한 효과를 나타내고, 가족의 특성이 제한될 때, 더욱이 무작위한 효과는 분리된 다양한 도시와 시골에서 나타남
  · 이 문서는 다양한 레벨의 모델이 어떻게 이용되는가를 나타내고 있음. 공동체 레벨에서 1인당 가구 평균지출의 소지역 추정치를 얻기 위해 사용되어지는 방법을 제안함
  · 분류방법으로 어떤 주에 어떤 지역에 어떤 공동체에 2년에 각 가구의 중첩되는 구조를 교차분류하고, 분리된 다양한 인간관계를 분류함으로써 2002년, 2004년 가족구성원과 인터뷰한 보고서임

・이 작업은 Brown, Goldstein과 Rasbash(2001, 통계 모델링 1, 103-124)의 견해에 의한 것이고 그 모델은 베이지안의 방법에 의해 산출됨

- 미국 동부에서 발전된 오존집중에 관련한 시공간 내일의 일기예보
  ・정확한 공기의 질에 대한 정보와 일기예보를 일반대중이나 환경 위생에 대한 의사 결정자에게 알려주는 것은 정말 필요한 일임. 이 보고서는 하루 동안 순차적 시공간모델을 나타냄
  ・미국동부 중 8시간 최대 오존층이 집중되는 데이터에 대해 이 모델은 관찰데이터와 컴퓨터 시뮬레이션(NOAA, CMAQ)에서 산출된 예보를 융통성있게 결합시킨 것임
  ・그럼에도 불구하고 매우 빠르게 그리고 실시간 모드로 다음 날의 일기예보가 계산되어짐
  ・제안하는 모델은 CMAQ 일기예보에서 시공간의 치우침을 조정하고, 일정한 표준에 의해 산출되는 컴퓨터 모델의 일기예보가 아닌 적절한 모니터링으로 관찰되는 데이터의 혼합 세팅치에서 자주 발생하는 표준치가 정해지지 않는 점을 피하기 위한 것임
  ・제안하는 모델은 미국 동부에서 0$ ~ 3$ 패턴으로 개선된 일기예보를 보여주는 것과 비축한 방대한 양의 데이터로 그 정당성이 입증되어짐

- 일상 음식물의 중금속 섭취량의 다단계 표본조사 연구
  ・음식물조사에 있어서 좋은 표본추출이 필수적인 것과 같이 음식물의 구성이 중요함
  ・효과적으로 단계적인 표본추출계획을 세워야 하고 그것의 정확성을 판단하기 위해, 우리는 계절과 지역, 개인적인 표본에 따라 그 구성이 얼마나 다양한지 연구해야 함
  ・우리는 세가지 기준에서 일상 음식물 표본을 모았고 그 기준들은 계절과 도시, 두가지 보관방법임
  ・일상 음식물 표본 중 습기, 카드늄, 질산염의 농도로 정하고 그 결과를 변량효과 모델을 이용하여 분석함

· 우리는 엄격하고 정확하게 단계적으로 표본추출 시나리오를 시뮬레이션하고 데이터의 대표값을 개선하기 위해 아주 많은 표본들을 시험했음

- 사전정보적인 표본추출계획 하에서의 베이지안 예상 추론
  · 알려진 covariates를 포함하기 위해 Sudgen과 Smith(Biometrika, 1984)의 접근법이 확대되었음
  · 변수(부분적으로 혹은 전체적으로 알려지지 않은) 디자인의 역할은 일반적인 경우에 사용되어짐
  · 간단한 예들은 어떻게 부분적인 디자인 정보가 응답 변수들의 비-표본화된 값들에게 베이지안 예상 추론이 어떻게 영향을 주는지 설명되어 있음

- Record linkage에 대한 계층적 베이지안 모형
  · 동일한 유한 모집단으로부터 독립적으로 뽑힌 두 표본들이 있고, 그 두 표본들이 몇 가지 변수를 공유한다고 가정하면, 이 두 자료를 통합하는 것은 공유된 변수들에 의한 모집단 모형의 추정방법을 개선할 수 있음
  · 공식통계에서는 정의가 어려운 모집단의 크기에 관심이 있을 때 record linkage는 필수적이고 사전적인 단계로 활용되고 있지만 다음과 같은 문제가 있을 수 있음
  · 일치 레코드들의 측정값들 사이에 독립이라는 강하고 무거운 가정이 있어야 하고 record linkage과정을 수행함에 있어서 가장 어려운 점은 양쪽 표본들에 영향을 줄 수 있는 측정오차(measurement error)가 존재할 수 있다는 것임
  · 이에 따라 그 대안으로 각 표본에 대해 관측된 키(key) 변수들은 비관측된 실제 값들에 조건부로 모델링되는 계층적 베이지안 모형을 제안하였고, record linkage가 MCMC알고리즘에 의한 반복을 통해 효율적으로 수행될 수 있다고 강조하였음

- 부분적으로 종합적인 자료의 모델 선택
  - 여러 통계청은 공개적인 데이터 사용에 있어서 응답자들의 신원이나 혹은 민감한 부분이 밝혀지는 위험을 최소화하기 위해 다중대체법을 사용함
  - 예를 들면, 통계청은 노출위험성이 있는 민감한 값들이나 식별성이 있는 키(key)값과 같은 값들로 조사되어 구성한 원본 단위들을 구성하면서 다중대체법으로 바꾸면서 부분적으로 종합적인 데이터셋을 배포함
  - 다변량의 추정치와 스칼라 추계를 얻는 방법은 여러 종류의 복합적으로 추계되어진 일련의 데이터셋 뿐만 아니라 부분적으로 종합적인 데이터셋으로 발전되어 왔음
  - 이러한 시나리오에서 종합적인 데이터셋들과 모델을 완성하는 베이지안 접근법과 BIC와 유사한 베이즈 요인 추정 모형선택의 접근은 유추되고 설명되어 질 것임
  - 우리는 이러한 과정이 간단한 경우를 넘어서서 어떻게 일반화되고 결측치에 관한 다중대체법이 어떻게 확장되는지를 고려할 수 있을 것임

- 패널 데이터에서 축차회귀 다중대체법의 적용
  - 삶의 질과 개발에 관한 통계 연구(PSLSD)와 KwaZulu-Natal 소득 원동력 연구(KIDS)는 축차회귀 다중대체법(SRMI)이 적용된 패널 데이터를 제공함
  - 축차회귀 다중대체법 절차는 결측 데이터를 위해 대체되어지는 값들의 완성된 데이터셋을 창출하고, 잘못된 구성요소 모델추정절차와 루빈의 규칙, 패널 데이터가 형성된 추론을 이용함. 이러한 데이터에 대한 축차회귀 다중대체법의 적용가능성과 완성된 회귀분석의 결과에 대해서 결론이 도출됨

- 사업체 조사에서 유한 Gaussian Mixtures를 통한 다중대체법
  - 다중대체법은 불완전한 데이터를 추론하기 위한 하나의 접근임. 다중대체법은 불완전한 데이터셋의 다른 대체를 행하기 위해 필수적으로 존재하고, 그 결과 완전한 데이터셋을 얻음
  - 그 다른 완전한 데이터셋은 표준방법으로 독립적으로 분석되어지고 그 결과들은 바람직한 추정치를 얻기 위해서 조합됨
  - 만약, 그 대체방법이 어떤 특성을 가진다면 루빈 조합규칙은 분석가들이 대상인구에 대한 임의의 매개변수를 이끌어 낼 다양한 추론을 얻음
  - 특히, 추정치의 다양성은 무응답과 관련한 불확실성을 고려하는데 사용될 수 있음. 이 양상은 하나하나의 대체방법에 대하여 가중 중요한 다중대체법의 이점을 표현함
  - 사실상, 조사 통계학에 있어서는 모든 데이터들이 실제로 조사되어진 것처럼 단일대체법으로 데이터를 분석하는 것은 흔히 일반적임. 이것은 오차가 적은 신뢰성 있는 부분에서나 혹은 매우 적은 p-value에 있어서 이러한 것들의 추정치의 정확성은 과대평가로 되어버릴 수도 있음
  - 다중대체 방법론은 베이지안 틀에서는 쉽게 정당화될 수 있으며 또한 종종 조사자료에서 무작위에 근거를 둔 추론으로 쓰일 수 있음. 분석을 하기 위한 변수들이 연속성일때, 데이터는 Gaussian 확률분포를 통해서 자주 모델화 되어짐
  - 사실상 정규성 가정은 데이터를 다양하게 대체하는 것을 쉽게 하고, 많은 컴퓨터 프로그램은 이러한 목적들을 완성시켜 왔음. 그러나 많은 실생활에서는 특히 사업체 조사에서 조사된 데이터 분포를 대칭화하기 위해서 데이터가 우선적으로 변경되어질지라도 정규성 가정을 정당화하는 것은 어려움
  - 이러한 문제점을 극복하기 위해 많은 접근방법들이 제안되어 왔음. 정규성 자정을 완화시키기 위해서, 일련의 방법론들이 모수적 혹은 비모수적 기술을 혼합하는 것에 근거를 둠

- 네덜란드 실업률의 EBLUP과 베이지안 소지역 추정값 비교
  · 네덜란드 통계청에 의해 수행되고 있는 노동력조사에서는 매년 실업률에 대해 국가과 광역지역(regional)수준 보다 좀 더 세분화된 425개 municipality 수준에서의 통계에 대한 수요가 증가하고 있음
  · 이에 따라 소지역 추정이 도입되었고, EBLUP(Empirical Best Linear Unbiased Predictor)와 HB(Hierarchical Bayes)의 두 가지 추정량에 대해서 unit level 모형과 area level 모형으로 비교 연구가 수행되었음
  · 비교결과, unit level 모형이 정확성에서 더 나은 면을 보였고, EBLUP과 HB 추정량에는 큰 차이가 없어 좀 더 간편한 형태의 EBLUP을 선택함
  · 주목할 점은 네덜란드의 경우 행정자료가 잘 정비되어 있기 때문에 등록된 실업률이 매우 좋은 설명변수로 이용 가능했음
  · 또한 이외의 다른 covariates나 모형에 대한 시뮬레이션 스터디와 연구가 지속되어야 한다고 밝혔음

2. 시사점 및 향후 과제

  · 베이지안 접근법은 소지역에 대해 기존 방법에 비해 효율적인 결과를 보여 소지역 추정기법을 위한 방법론으로 많이 인식이 되고 있음
  · 그러나 복잡한 표본설계에 의존하는 기존 방법의 대안으로도 활용 가능한 기법이며 우리의 연동표본제와 같이 복잡한 표본 구조를 바탕으로 하는 표본조사의 추정기법 연구에도 적용 가능
  · 모형기반 소지역 추정방법에 대한 기법 연구 및 사례연구
  · 복합 및 합성 추정기법을 연구과제에 적용
  · 다중무응답 대체를 적용할 수 있는 방법에 대한 연구

〔별첨 1〕 베이지안 추정기법 소개

□ Module 1 : 표본조사에서 추론에 대한 접근방법

- 고전적(classical) 추론방법
  · 관측된 데이터에 관한 모형만을 사용하여 모형이 가지는 미지의 모수를 추론
  · design-based라고도 부르며 아직까지 많은 분야에서 널리 활용되고 있으며 추정량의 선택에 따라서 모형이 종종 이용됨
  · 불편 추정량(unbiased estimator)과 대표본(large sample) 이론에 근거함

- 베이지안 추론방법
  · 관측된 데이터와 모수 모두에 확률 모형을 사용하는 방법으로 데이터로부터 얻은 모수에 관한 정보뿐만 아니라 모수에 관한 과거의 경험이나 사전 지식 같은 주관적 견해를 수량화한 모수의 특성을 결합시켜 보다 정확한 추론을 하고자 하는 데 있음
  · model-based라고도 부르며 추정량, 표준오차, 구간추정 등 추론의 전 영역에 사용됨
  · 이 접근법이 좀 더 통합적이나, 층화나 군집과 같은 표본설계 특성에 맞는 세심한 모형이 필요함

- 베이지안 추론의 장점
  · 베이지안 방법은 해석하기 쉬움
  · 베이지안 추론은 우도원리(likelihood principle)를 따름
  · 베이지안 추론은 모호한 결과를 제공하지 않음
  · 베이지안 추론은 대표본 이론을 요구하지 않음. 따라서 소표본 베이지안 추론을 대표본의 경우에도 같은 방법으로 수행함

□ Module 2 : 베이지안 추론의 기초

- 유한모집단의 모수를 위한 점추정
  · 점추정은 미지의 Q에 대한 하나의 요약된 "최상의"값으로 알려져 있음
  · 이러한 선택들은 Q의 사후분포의 평균, 최빈값, 중위수이다.
  · 대칭적인 분포를 위해서 직관적인 선택은 대칭의 중심이다.
  · 비대칭적인 분포를 위한 선택은 명백하지 않음. 그것은 "loss" 함수에 의존함

- 유한모집단의 모수를 위한 구간추정
  · 더욱 더 좋은 요약은 구간추정임
  · 커버리지 비율 1-α를 미리 고정하고 합계한 1-α사후확률 Q의 대부분의 값들을 포함하기 위하여 가장 높은 사후 밀도 지역 C를 결정
  · 미리 Q_0의 값을 고정하고 Q_0보다 더 좋은 Q값의 수집에 의해 C를 결정하고 C의 사후확률로서 1-α범위를 계산함

- 단순랜덤표본의 모델
  · $Y_i \sim iidN(\mu, \sigma^2), I = 1, 2, ..., n$
  · $\pi(\mu, \sigma^2) \propto \sigma^{-2}$
  · 단순랜덤표본은 결과가 $Y_{inc} = (y_1, ..., y_n)$임
  · $Q = \overline{Y} = \frac{n\bar{y} + (N-n)\overline{Y}_{exc}}{N} = f \times \bar{y} + (1-f) \times \overline{Y}_{exc}$
  · Q의 사후분포를 끌어냄

- 베이지안 비모수 방법
  · 모집단: $Y_1, Y_2, Y_3, ..., Y_N$
  · 모든 가능한 명백한 값: $d_1, d_2, ..., d_K$
  · 모델: $Pr(Y = d_k) = \theta_k$

· 사전분포: $\pi(\theta_1, \theta_2, ..., \theta_k) \propto \prod_k \theta_k^{-1}$ if $\sum_k \theta_k = 1$

· 평균과 분산:

$$E(Y_i|\theta) = \mu = \sum_k d_k \theta_k$$

$$Var(Y_i|\theta) = \sigma^2 = \sum d_k^2 \theta_k - \mu^2$$

□ Module 3 : 복합표본설계를 위한 모델

- 데이터 수집과정을 무시(ignoring)
  · 데이터 수집과정을 무시하는 우도(likelihood)는 Y의 우도함수에 대한 모형에 근거함

$$L(\theta|X, Y_{inc}) \propto p(Y_{inc}|X, \theta) = \int p(Y|X, \theta) dY_{exc}$$

  · θ에 대한 사후분포: $p(\theta|Y_{inc}, X) \propto p(\theta) L(\theta|Y_{inc})$
  · 전체적인 사후분포가 이러한 단순한 사후분포로 감소할 때, 자료 수집 메카니즘은 베이지안 추론 θ에 대하여 무시할만 하다(ignorable)라고 함
  · 데이터 수집 메카니즘을 무시하는 데는 다음과 같은 두가지 일반적인 조건이 가정되어야 함
    i) 추출이 임의적이어야 함(Random)
    ii) 베이지안 특성(모형)이 동반되어야 함
  · 이러한 경우 데이터 수집 메커니즘이 모수 θ에 대한 추론에 영향을 미치지 않게 됨

- 층화추출에 대한 모형
  · 이 모형을 위한 베이즈 추론은 층화랜덤표본으로부터 모집단 평균에 관한 표준화된 고전적인 추론과 동일함
  · posterior mean은 포함확률의 역수에 가중함:

$$\overline{y_{st}} = N^{-1}\sum_{j=1}^{J} N_j \overline{y_j} = N^{-1}\sum_{j=1}^{J}\sum_{i:x_i=j} y_i/\pi_j$$

여기서 $\pi_j = n_j/N_j$ = j번째 층 내 추출 확률

- 2단 집락추출에 대한 모형
  - 1단계에서 C개의 집락으로부터 c개의 집락을 추출하고, 2단계에서 선택된 집락으로부터 k_i개의 표본을 선택하는 추출에서
  - population mean 은 다음과 같이 분해(decomposed)될 수 있음

$$NQ = \sum_{i=1}^{c}[K_i\overline{y}_i + (K_i - k_i)\,\overline{Y}_{i,\mathrm{exc}}] + \sum_{i=c+1}^{C} K_i\overline{Y}_i$$

  - posterior mean given Y_{inc}:

$$E[NQ_i \mid Y_{\mathrm{inc}}] = \sum_{i=1}^{c} k_i\overline{y}_i + (K_i - k_i)E(\mu_i \mid Y_{\mathrm{inc}})] + \sum_{i=c+1}^{C} K_i E(\Theta \mid Y_{\mathrm{inc}})$$

  - posterior variance :

$$Var(NQ \mid Y_{\mathrm{inc}}) = \sum_{i=1}^{c}(K_i - k_i)(\sigma^2 + (K_i - k_i)\tau^2) + \sum_{i=c+1}^{C} K_i(\sigma^2 + K_i\tau^2)$$

□ Module 4 : 조사무응답

- 다중대체법의 장점
  - 대체 모델은 분석모델과 다를 수 있음
    - ··최종분석에서 포함되어지지 않는 변수를 초함
    - ··다중분석을 통해 결측치의 처리의 일치성을 구함
    - ··다중대체법의 조합된 규칙은 완전한 자료의 추론이 베이지안(예: design based 조사의 추론)이 아닐 때 또한 적용되어질 수 있음
  - 이용자들이 고정되는 공공적인 이용 자료는 다중대체법에 의해 제공될 수 있음

- 항목무응답, 다중 대체법
  - 항목무응답은 일반적으로 “swiss cheese”의 패턴처럼 복잡함

· 가중하는 방법은 자료가 단조로운 패턴일 때 가능함. 그러나 일반적인 패턴에서 개발하는 것은 매우 어려움
· 모델-기반의 다중 대체 방법은 Module5에서 이용 가능할 수 있음
· 모든 관측된 자료에서 전체적으로 조건을 준다면 이러한 방법들은 MAR 가정은 약함

□ Module 5 : 소지역 추정기법과 간접 추정기법 소개

- 소지역 추정의 정의 및 배경
· 소지역(small area)이란 county, municipality나 census division 과 같이 작은 지리적 지역을 일컬음
· 소영역(small domain)이란 넓은 지리적 지역 내의 특정한 나이, 성별, 인종 그룹과 같이 작은 부차집단(subpopulation)을 일컬음
· 소지역 추정에 대한 모형 기반(model-based) 추론은 다음과 같은 장점을 가지고 있음

(1) 가정된 모형 하에서 최적(Optimal)의 추정량이 도출될 수 있음
(2) 불확실한 측정값이 각 추정량과 결합될 수 있음
(3) 모형이 표본 데이터로부터 검증될 수 있음
(4) 복잡한 자료구조들과 반응 변수들이 성질에 따라 다양한 모형들이 이용될 수 있음

- 합성과 복합 추정기법 소개
· 회귀 추정값을 종종 합성 추정값이라 하는 반면, 가중 평균 형태의 추정값을 복합 추정값이라 함
· y_{ij}를 i번째 지역에 있는 j번째 단위에 대한 관심특성이라 하고
· x_{ij}를 i번째 지역에 있는 j번째 단위에 대한 보조정보라 하면

· 보조정보(auxiliary information)가 없을 때

추정기법	직접(direct)	합성(synthetic)	복합(composite)
표본평균 추정값 형태	$\overline{y_{is}}=\frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$	$\overline{y_s}=\frac{\sum_{i=1}^{m} n_i \overline{y_{is}}}{\sum_{i=1}^{m} n_i}$	$w_i \overline{y_{is}}+(1-w_i)\overline{y_s}$

· 보조정보(auxiliary information)가 있을 때

추정기법	직접(direct) (비추정량)	합성(synthetic)	복합(composite)
표본평균 추정값 형태	$\left(\frac{\overline{y_{is}}}{\overline{x_{is}}}\right)\overline{X_i}$	$\hat{\overline{y}}_i^{RS}=\frac{\overline{y_s}}{\overline{x_s}}X_i$	$\frac{n_i}{N_i}\overline{y_{is}}+(1-\frac{n_i}{N_i})\frac{\overline{y_s}}{\overline{x_s}}\overline{X_i}$

□ Module 6 : 경험적 베이지안(Empirical Bayesian) 추정기법

(i) $y_i,\cdots,y_n|\Theta_1,\cdots,\Theta_n,\lambda\sim p(y_i|\Theta_i)$
(ii) $\Theta_i,\cdots,\Theta_n|\lambda\sim\pi(\Theta_i|\lambda)$
(iii) $y_1,\cdots y_n$의 주변확률분포에 근거하여 초모수 λ를 추정한 후 이를 사후분포에 대입하여 구한 추정된 $\Theta_i,\cdots,\Theta_n$의 사후분포를 추론에 사용

- EB 추정은 직접 추정에 비해 좁은 지역에 대해서는 명백한 강점을 가지고 있고, 단지 표본정보에 기초하는 직접추정과 대조적으로 표본과 사전정보 모두를 이용하게 됨
- EB 절차는 알려지지 않은 초모수를 최우추정법, 정률법 혹은 UMVUE 등을 사용하여 추정함으로써 추정된 사전분포를 사용함
- 그러나 원시적(naive) EB는 표준오차를 과소 추정하는 것도 사실임

□ Module 7 : 계층적 베이지안 추정기법

(i) $y_i, \cdots, y_n | \Theta_1, \cdots, \Theta_n, \lambda \sim f(y_i | \Theta_i)$
(ii) $\Theta_i, \cdots, \Theta_n | \lambda \sim \pi(\Theta_i | \lambda)$
(iii) $\lambda \sim \pi(\lambda)$ 이와 같이 초모수에 분포를 설정하여 $\Theta_i, \cdots, \Theta_n$의 사후분포를 구함

- HB 절차는 사전분포가 가지는 초모수에 초사전분포를 고려하여 사전정보에서의 불확실성을 모형화 함
- EB와 HB 추정 모두 사전정보에서의 불확실성을 인식하여 사전분포를 고려함
- EB와 HB 추정 모두 점추정 면에서는 비교적 비슷한 결과를 제공함
- 그러나 HB 기법은 표준오차를 측정함에 있어서 원시적(naive) EB에 비해 명백히 우세함