복합표본조사 데이터에 대한 통계분석

Analysis of Complex Sample Survey Data

이 기 재*

Ⅰ. 개 요

(1) 조사 데이터 분석에 두 가지 접근법

- 설계기반 접근법(Design based approach)
  - 표본설계 특성을 고려한 조사 데이터 분석 방법
  - 대규모 조사 데이터의 분석에 사용됨
- 모형기반 접근법(Model based approach)
  - 조사 데이터 분석에서 표본설계 특성을 반영하지 않은 분석 방법
  - 얻어진 조사 데이터에 대해서 서로 독립이고 동일한 분포를 갖는다고 가정
    ⇒ 표본이 단순임의추출법으로 추출된 것으로 간주하여 분석
  - 간단한 조사 데이터나 확률표본으로 간주할 수 없는 경우에 사용

(2) 조사 데이터 분석 기법

- 기술적 분석(Descriptive Analysis)
  - 조사 데이터에 대한 기초적인 분석으로 도수분포표, 중심위치 및 산포 측도 등의 수치요약, 히스토그램과 산점도 등의 그래프, 상관계수 등
- 추론적 분석(Inferential Analysis)
  - 가설검증을 하거나 통계모형을 적합하여 분석하는 것으로 모평균 검증, 회귀분석, 상관분석, 카이제곱 검증, 로그선형모형 분석, 로지스틱모형 등

* 한국방송통신대학교, 정보통계학과, kjlee@mail.knou.ac.kr

(3) 조사 데이터 분석 기법의 선택

연구주제, 변수의 수, 각 변수의 측정척도(명목형, 순서형, 연속형) 등을 고려하여 선택

〈표 1〉 자주 활용되는 조사 데이터 분석 기법 예시

일변량분석	이변량분석	다변량분석
1. 빈도표 분석 2. 모평균, 모총계, 모비율, 모분산 등의 추정 3. 히스토그램, 원그래프 등 그래프를 이용한 분석	1. 분할표 분석 2. 산점도 분석 3. 단순회귀분석 4. 상관분석 5. 두 모평균의 비교	1. 조건부 분할표 분석 2. 다중 및 부분 상관분석 3. 다중회귀분석 4. 로그선형모형 분석 5. 인자분석 6. 경로분석(path analysis)

(4) 조사 데이터의 2차 분석(secondary analysis of survey data)

- 과거 : 기술통계적 분석(descriptive analysis) 위주
- 현재 : 추론적 분석(analytic analysis) 연구에 중점

  ⇒ 조사 데이터를 학문적 목적의 2차 분석에서 널리 활용됨
- 조사 데이터의 2차 분석이란?
  - 데이터 수집 과정에 관여하지 않은 연구자들이 수집된 조사 데이터를 분석하는 것
  - 정치, 사회연구에 많이 활용되고 있음

    ⇒ 조사 데이터 분석에서 표본설계를 반영한 분석기법 적용

(5) 주요 강의 내용

- 복합표본설계란?
- 복합표본설계 사례
- 복합표본조사 데이터의 특징
- 복합표본조사 결과를 단순임의표본에서 얻은 결과인 것처럼 분석할 있는가? 단순임의표본으로 가정해서 분석할 경우에 생길 수 있는 문제는?
- 표본설계를 반영한 조사 데이터 분석법은?
  - 기술통계량 분석
  - 카이제곱 검증
  - 회귀분석
  - 로지스틱모형 분석

▪ 조사분석용 통계패키지 및 가능 분석법 소개

※ 조사 데이터를 분석할 때는 모수 추정과 추정량의 분산 계산 과정에서 표본설계에 대한 특성을 고려해야 한다.

Ⅱ. 복합표본조사(complex sample survey)

1. 복합표본설계

(1) 층화(stratification)

- 모집단의 대표성을 높이기 위해서 사용됨
  예) 지역, 성별, 연령, 학력 등이 층화 변수로 사용됨
- 일반적으로 추정의 정확도를 높여줌
- 층화를 무시하고 분석하면 추정 편향이 발생하고, 추정량의 분산도 커지게 됨

(2) 집락화(clustering)

- 모집단에 대한 최신 추출틀이 없는 경우 표본추출을 위한 현실적 방법
- 조사원의 업무 부담과 조사비용을 고려한 방법
- 지역적으로 인접한 가구나 사람들이 표본으로 추출됨
  ⇒ 같은 집락 내 표본들은 양의 상관관계가 예상됨
  ⇒ 추정의 정확도가 상대적으로 떨어지게 됨

(3) 불균등확률추출

- 주로 집단간의 비교를 목적으로 Oversampling하는 경우
  예) 도시와 농촌 어린이들의 건강상태 비교
- 표본추출 상황에서 불가피하게 발생하는 경우
  예) 전화조사 : 표본으로 추출될 확률은 해당 가정의 전화번호가 몇 개인가에 비례
- 불균등확률추출을 무시하고 분석하면 추정에 편향 발생
  ⇒ 가중치 사용

(4) 복합표본설계

- 실제 조사는 비용절감, 조사관리, 표본추출틀 마련 등의 필요에 따라서
  단순임의추출법의 적용이 곤란함
- 대규모 표본조사는 대부분 다단계추출법 이용
  특히, 층화, 집락추출, 불균등확률추출 등이 복합적으로 이용됨

• 단순임의추출법이외의 다른 방법으로 추출되는 모든 경우가 해당됨

2. 복합표본설계 사례

(1) 『사회통계조사』에 대한 표본설계

① 조사 목적 : 국민의 삶의 수준과 사회적 변동의 파악
사회개발 정책수립의 기초자료로 제공

② 조사대상 : 15세 이상의 모든 가구원(부분적으로 14세 이하도 조사대상으로 포함)

③ 주요 조사내용

- 기본사항 : 성별, 교육정도, 나이, 산업, 직업, 직위 등
- 부문별 주요 조사사항
  - 가족 : 가족생활만족도, 청소년 고민, 부모 부양, 결혼 및 이혼에 대한 태도 등
  - 소득과 소비 : 소득 만족도, 소비생활만족도, 저축 및 부채 등
  - 노동, 교육, 보건, 주거 및 교통, 정보와 통신, 환경, 복지, 문화와 여가 등

④ 표본설계

- 모집단 기초자료
  인구 주택 총조사 10% 표본조사 결과 자료 이용
- 층화
  - 7대 도시와 9개 도로 층화한 후에 9개 도에서는 동부 및 읍·면부로 나눔
    ⇒ 모두 25개 층으로 구성됨
  - 모집단의 대표성 제고와 시도별 독립적 추정 목적
- 표본규모
  - 시도별 목표정도를 충족시킬 수 있는 범위 내에서 조사가능 한 표본크기로 결정
  - 전국적으로 총 1,231개 조사구 추출
- 표본조사구 추출
  - 1단계(표본조사구 추출) : 각 층에서 확률계통비례추출법(PPS)으로 추출
  - 2단계(표본가구 추출) : 표본으로 선정된 표본조사구에서 24가구 추출
- 최종적으로 28,800가구가 표본으로 결정됨

⑤ 추정과정

- 단순기술통계량 산출 : 가중치 산정 후 계산

♣ 통계청에서 실시되고 있는 『경제활동인구조사』, 『정보화실태조사』, 『생활시간조사』등을 비롯한 대부분 가구 대상 통계조사는 다단계추출법을 이용하고 있다.

(2) 『국민구강건강실태조사』에 대한 표본설계

① 조사목적

- 구강건강 지표와 구강진료 필요 및 구강보건의식 행태 파악
- 국민구강보건정책에 필요한 기초자료 확보

② 조사내용

- 구강보건에 대한 지식 및 태도, 행동 등
- 치아건강상태, 치주조직건강상태, 의치·보철상태

③ 표본설계 개요

- 1995년 인구주택총조사 결과 이용(시설단위 조사구 제외)
- 층화

  층 1 : 7대 시(서울, 부산, 대구, 인천 광주, 대전, 울산)의 동

  층 2 : 기타 시의 동

  층 3 : 읍·면 지역

  층 4 : 7대 시의 동 지역의 신축 아파트

  층 5 : 기타 시의 동 지역의 신축 아파트

  층 6 : 읍·면 지역의 신축 아파트
- 표본크기

  · 181개 표본조사구 : 1995년 인구주택총조사의 조사구 이용

  · 19개 표본조사구 : 신축 아파트에서 추출
- 표본 배분

  읍·면 지역에 상대적으로 많이 표본 배정
- 표본 조사구 및 가구 추출

  · 1차 추출 : 각 층에서 크기의 측도에 비례하는 확률비례계통추출

  · 2차 추출 : 각 표본 조사구에서 약 35 가구 추출

  18세 이상의 성인에 대해 면접 및 구강검사
- 응답자 현황 (구강건강상태조사)

  18세 이상 성인 8,927명 조사

(3) 복합표본조사 데이터의 특징

▪ 유한모집단 대상

▪ 표본추출 단계 : 복합표본설계에 의한 추출

- 층화, 집락추출 : 표본 단위들의 비독립성

  ⇒ 집락 내 관측단위들 간에 존재하는 상관관계

  ⇒ 분산 추정에 영향을 미침
- 불균등확률추출 : 서로 다른 추출확률

▪ 가중치 부여

- 불균등확률선택을 보정하기 위한 방법
- 무응답 처리 및 사후 보정을 위한 방법

♣ 복합표본조사 데이터는 일반적인 통계분석의 가정인 i.i.d 가정을 만족하지 못한다. 조사 데이터를 분석하여 얻은 결과를 일반화하기 위해서는 표본설계가 추론과정에 반영된 설계기반 접근법으로 분석되어야 한다. 조사 데이터 분석에서 복합표본설계는 모수 추정과 분산 계산에 영향을 미친다.

3. 가중치

(1) 가중치의 필요성

- 표본추출과정의 불균등 추출확률에 대한 보정
- w_i : i번째 표본 단위가 대표하는 모집단 내 단위의 수

(2) 조사 데이터 분석에 미치는 영향

- 추정과정에서 가중치 이용 : 예) $\bar{y} = \frac{\sum_{i \in S} w_i y_i}{\sum_{i \in S} w_i}$

  ⇒ 모수에 대한 비편향 추정(unbiased estimation) 가능
- 추정량의 분산이 커지게 됨

  ⇒ 추정량의 분산이 약 $1+CV^2$배 증가, CV : 가중치의 변이계수

♣ 조사 데이터를 분석할 때 가중치를 무시하면 심각한 편향(bias)이 발생할 수 있다. 한편, 가중치를 사용하면 분산이 커지게 된다. 대규모 표본조사에서는 표본크기가 크기 때문에 중요한 것은 추정량의 편향이다.

(3) 가중치를 사용하지 않을 경우 발생할 수 있는 분석 오류

① 기술통계분석의 경우 : 『국민구강건강실태조사』결과 분석 중 일부

- 우리 나라 18세 이상 성인의 영구치 우식경험치아수

〈표 2〉 영구치 우식경험치아수 평균 : 가중치 이용

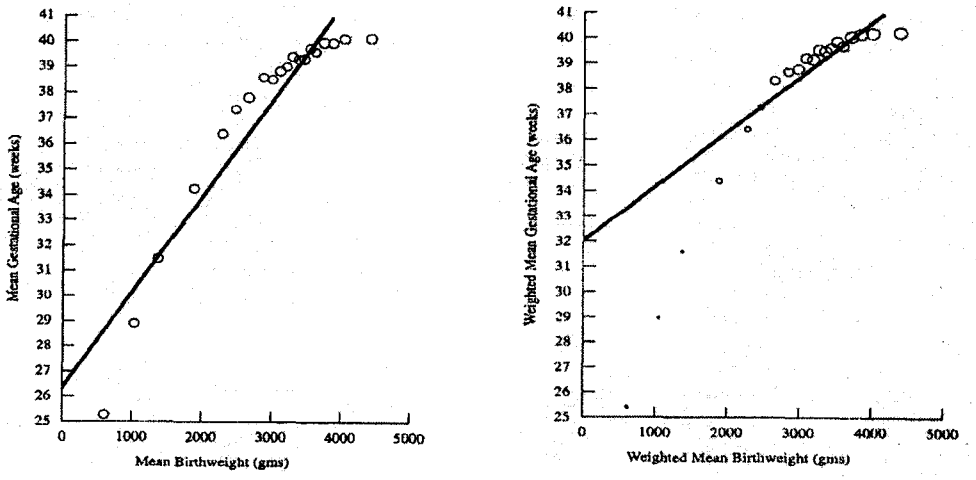
	전체	남	여
전국	6.99	5.58	8.30
대도시	6.76	5.46	7.97
중소도시	6.97	5.41	8.45
군지역	7.86	6.54	9.01

〈표 3〉 비가중 표본평균의 상대편향(%)

	전체	남	여
전국	14.5%	20.1%	7.1%
대도시	6.2%	8.4%	1.1%
중소도시	15.9%	22.2%	7.3%
군지역	11.8%	17.7%	6.0%

$$\text{상대편향} = \frac{\text{비가중평균} - \text{가중평균}}{\text{가중평균}} \times 100(\%)$$

② 회귀분석의 경우



(a) 가중치를 사용하지 않는 회귀직선 적합 (b) 가중치를 사용한 회귀직선 적합

출처 : Korn & Graubard(1995)

(4) 가중치 부여 과정

① 설계가중치 : 추출확률의 역수

- 추출단계의 불균등추출확률을 보정하여 추정 편향 제거

② 무응답 조정

- 응답자 속성에 따른 응답률의 차이로 인해서 발생한 편향 제거

③ 사후층화 조정

- 모집단과 표본의 구조를 유사하게 조정함으로써 편향 감소 효과

♣ 표본설계의 정보와 가중치가 조사 데이터에 함께 있어야 추정과 분산추정 과정에서 이용될 수 있다.

(5) 가중치 부여 사례 : 국민구강건강상태조사

- 설계가중치
  - 표본 조사구 추출확률의 역수와 가구조사 완료율의 역수의 곱
- 무응답 조정
  - 지역(3), 성별(2), 연령(11) 구분에 따라 66개 무응답 조정 셀 구성
- 사후층화 조정 : 2000년 주민등록통계 이용
  - 지역(3), 성별(2), 연령(11) 구분에 따라 모두 66개 층 구성

〈표 4〉 구강건강상태조사 데이터에 대한 가중치

	지역별 구분			성별 구분		전 체 (n=21,829)
	대도시 (n=8,312)	중소도시 (n=8,843)	군 지역 (n=4,674)	남자 (n=10,104)	여자 (n=11,725)	
최소값	0.17	0.20	0.10	0.10	0.11	0.10
1사분위수	1.05	0.48	0.33	0.53	0.45	0.48
중앙값	1.18	1.12	0.44	1.20	1.00	1.05
3사분위수	1.33	1.44	0.49	1.39	1.21	1.33
최대값	5.75	4.43	5.34	5.75	5.34	5.75
평균	1.26	1.03	0.49	1.11	0.91	1.00
CV	37.83	52.23	66.99	55.43	52.91	55.54

4. 복합표본설계가 추정에 미치는 영향

(1) 분석과정에서 복합표본설계를 무시할 때 발생하는 문제들

- 추정 편향 발생
- 추정량 분산의 과소 추정 (특히 집락표본에서 심각함)

⇒ 신뢰구간의 폭이 실제보다 작아짐

⇒ 가설검정에서 실제로는 통계적으로 有意하지 않지만 有意하다는 결론 도달 (제2종오류의 가능성이 높아짐)

(2) 설계효과(misspecification effect: meff)

- 층화 : 일반적으로 추정의 정도(精度)가 높아짐
- 집락추출 : 추정의 정도(精度)가 낮아짐

⇒ 이를 종합적으로 살펴보기 위한 것이 설계효과(meff)

- $meff(\hat{\theta}) = \frac{\text{표본설계를 반영하여 구한 } \hat{\theta}\text{의 분산}}{\text{단순임의표본으로 간주하여 구한 } \hat{\theta}\text{의 분산}}$
  - meff는 분석과정에서 표본설계를 무시함으로써 발생하는 추정량 분산의 과대 또는 과소 추정의 정도를 나타냄
  - meff가 1보다 크게 나타나는 경우
    단순임의표본으로 간주해서 분석하면 실제보다 추정량의 분산이 작게 계산됨

(3) 급내상관계수(intraclass correlation coefficient : ρ)

- 같은 집락 내 조사단위들의 유사성의 정도를 수치로 표현한 것
- 집락 내 모든 가능한 쌍들간의 상관계수를 의미
  ⇒ 집락내 단위이 랜덤하게 구성된 경우 : $\rho = 0$
  ⇒ 대부분의 자연적 집락 : 양의 급내상관계수를 나타냄

(4) 집락표본에서 설계효과

- $meff = 1 + (\bar{b} - 1)\rho$, $\bar{b}$: 집락의 평균 표본크기, ρ : 급내상관계수
- 대부분의 사회조사는 센서스 조사구를 1차추출단위로 사용
  ⇒ 조사구 내의 가구들은 동질적이어서 급내상관계수가 양수의 값을 갖게 됨
  ⇒ 단순임의표본으로 가정해서 분석하면 실제보다 분산을 과소하게 추정하게 됨
  ⇒ 일반적으로 급내상관계수는 인구학적 변수에 대해서는 다소 낮게 나타나지만, 사회·경제적 변수에 대해서는 상대적으로 크게 나타나는 경향이 있음

♣ 예를 들어 통계청에서 시행되고 있는『사회통계조사』는 지역 특성에 따라서 층화를 하고, 각 층에서 1차추출단위인 조사구를 추출하고, 각 표본조사구에서 몇 가구씩 추출하고 있다. 따라서 조사구 내 가구의 유사성 정도가 상당히 높을 것으로 예상된다. 일반적으로 추정량의 표집분산은 층화, 집락추출, 불균등 가중치 등의 종합적인 영향으로 나타나게 된다. 이러한 이유로 조사 데이터는 가중치, 층 식별 변수, 1차추출단위(PSU) 구분 등의 표본설계에 대한 정보를 함께 담고 있어야 한다.

5. 조사 데이터 분석에서 분산 추정

(1) 조사 데이터 분석에서 분산 추정

- 조사설계에 따라서 분산 추정 방식도 달라짐
- 불균등추출확률, 층화, 집락추출 등 적용한 표본추출
  ⇒ 전통적인 데이터 분석의 기본 가정을 따르지 않음
  ⇒ 표본설계를 반영한 분산 추정 필요
- 가중치를 이용한 비선형 추정량 이용

⇒ 근사적인 분산 추정법 필요

(2) 조사데이터 분석에서 분산 추정을 위한 일반적인 방법들

① 선형화 방법

- 조사 데이터 분석 소프트웨어들에서 기본적으로 지원하는 분산의 근사 추정법
- 복잡한 비선형 추정량를 선형함수로 근사하여 분산계산

② 잭나이프 방법

- 조사 데이터 분석 소프트웨어들에서 널리 지원되는 분산의 근사 추정법

③ 균형반분 방법

- 선형화 방법과 거의 유사한 결과를 줌

④ 붓스트랩 방법

- 얻어진 표본 데이터를 모집단으로 간주해서 재표집해서 추정량의 분산 계산

⑤ 랜덤그룹법

- 조사된 표본을 조사설계를 그대로 반영하며 중첩되지 않는 r개의 그룹으로 임의분할한 다음, 각 랜덤그룹마다 독립적으로 추정량을 계산하고, 또 이렇게 얻은 r 개의 추정값들의 표본분산으로 $\hat{\theta}$의 분산을 추정하는 방법

(3) 조사 데이터 분석 소프트웨어 패키지 및 가용한 분산 추정법

소프트웨어	운영체제	요구 시스템	분산추정법				
			선형화	BHS	잭나이프	랜덤그룹	붓스트랩
SUDAAN	Windows	(SAS 6.12, 8)	√	√	√		
SAS	Windows		√				
VPLX	Windows			√	√	√	
WesVarPC	Windows			√	√		√
Stata	Windows		√		√		√
IVEware	Windows	SAS 6.12	√		√		
GES	Windows	SAS 6.12	√		√		
CSPro	Windows		√				
PC CARP	DOS		√				
Epi Info	DOS		√				
CLUSTERS	DOS		√				

출처 : 성내경(2000) 자료 인용
참고 : http://www.fas.harvard.edu/~stats/survey-soft/survey-soft.html/

(4) 분산 추정법의 선택

- 조사 데이터 분석에서 주로 활용되고 있는 분산추정법들 : 선형화 방법, 잭나이프법, 균형반분표본법
- 선형화 방법 : 중앙값이나 백분위수 추정에 대한 분산을 계산할 수 없음
- 균형반분법 : 상대적으로 간단하지만 각 층에서 여러 개의 PSU를 추출하는 경우에 적합하지 않을 수 있음
- 잭나이프 방법 : 상대적으로 계산량이 많지만 무응답 조정이나 사후층화 조정을 반영해서 분산을 계산할 수 있는 방안으로 유용

Ⅲ. 조사 데이터 분석

1. 조사 데이터 분석을 위한 준비

(1) 예비 분석의 중요성(2차분석을 목적으로 하는 경우)

- 유익한 분석 결과를 제공할 수 있는가에 대해서 사전에 검토할 것
- 분석하고자 하는 각 부분집단별로 충분한 크기의 표본이 있는가 확인할 것
- 결측값과 극단값 등이 있는지 확인할 것
- 충분한 수의 1차추출단위(PSU)가 있는가를 검토할 것

♣일반적인 가이드라인으로 PSU의 수는 어떤 추정량에 대해서 PSU간의 분산 추정값을 계산할 수 있을 만큼 커야 한다. 연령과 성별 구분에 따라서 통계를 작성하고자 하는 경우를 생각해 보자. 이 때 다수 PSU에서 특정 연령과 성별에 대해서 관측값이 없는 경우라면 몇 개의 연령 구분을 묶어서 다시 정의하는 방안을 검토해야 한다. 우선 비가중 분할표를 이용해서 의미 있는 분석이 될 수 있을 정도로 PSU들이 분포하고 있는가를 살펴봐야 한다.

(2) 조사 데이터를 분석하는 데 필요한 정보들

- 가중치 : 추정치 계산과 분산 계산에 필수적인 정보
- 층, PSU, 추출단위 등에 대한 구분 변수들

⇒ 각 층에서는 적어도 두 개의 PSU가 추출되어야 함
그렇지 않은 경우에는 PSU들을 묶어 주는 작업이 필요하게 됨

(3) 사용 가능한 조사 데이터 분석용 소프트웨어들

① SUDAAN

- 집락표본에서 나오는 상관된 자료(correlated data)를 분석하는 소프트웨어로 개발
- 복합조사 데이터에 대한 다양한 분석 기능 제공

· 기술통계분석(DESCRIPTIVE) : 평균, 총계, 비율, 분위수 등의 추정과 분산 계산
· 교차표 분석(CROSSTAB) : 카이제곱 검증, 각종 연관성 측도 계산 등
· 선형회귀(REGRESS) : 선형 회귀모형을 적합하고, 모수에 대한 검증 수행
· 로지스틱 회귀(LOGISTIC) : 로지스틱모형을 적합하고 모수에 대한 검증 수행
· 로그선형모형(MULTILOG) : 로그선형모형 적합
· 비례위험모형(SURVIVAL)

- 다양한 분산 추정법 제공 : 선형화 방법, 균형반분법, 잭나이프법
- 거의 모든 조사설계에 대한 데이터 분석 가능

② Stata

- 전통적인 통계분석뿐만 아니라 조사 데이터 분석도 가능한 통계 소프트웨어
- 복합조사 데이터에 대한 다양한 분석 기능 제공
  기술통계, 교차표 분석, 선형회귀, 로지스틱 회귀, 로그선형모형, 비례위험모형 등
- 다양한 분산 추정법 사용 가능 : 선형화 방법, 잭나이프법, 붓스트랩법
- 거의 모든 조사설계에 대한 데이터 분석 가능

③ SAS

- SAS 8 판부터 표본조사 데이터 분석을 위한 절차들을 도입
- SURVEYSELECT : 표집방법, 표집틀 등 표본설계에 관련된 사항을 저정하면 확률표본을 추출 가능
- SURVEYMEANS : 조사설계를 반영한 각종 기술통계량들 산출
- SURVEYREG : 조사 데이터에 회귀모형 적합, 회귀계수에 대한 유의성 검증 등

2. 회귀분석

(1) 조사 데이터에 대한 회귀모형

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

여기서, β_0, β_1, $\cdots$, β_k : 추정될 회귀계수들, ε_i : 오차항

(2) 조사 데이터에 대한 회귀모형 적합법들

① 표본추출이 SRS안 경우

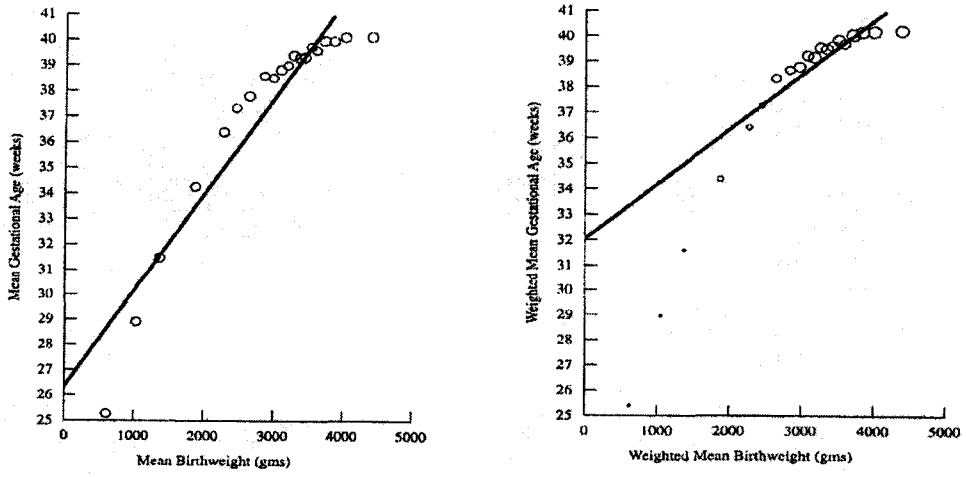
- 전통적인 통계기법(model based approach) 적용
- 회귀계수의 추정을 위해서 최소제곱법(Ordinary Least Squares: OLS) 이용

② 복합표본조사 데이터의 경우

- 표본설계가 SRS 가정에서 크게 벗어남
  ⇒ 최소제곱법(OLS)을 이용하면 추정치의 편향과 가설검증의 잘못된 결과 초래

- 가중최소제곱법(Weighted Least Squares: WLS)을 이용하여 회귀계수 추정
  ⇒ 추정에서 조사 데이터의 가중치 이용
  ⇒ 회귀계수에 대한 비편향 추정 가능
- 추정치의 분산 계산 :
  분산 계산 과정에서 PSU 내 관측치간의 공분산 구조를 반영 필요
- 설명변수의 누락과 오차항에 대한 가정 위배에 대해서 강건성(robustness) 유지

♣각 관측값들은 서로 다른 가중치를 갖는 경우에 가중치가 주변수값 y_i와 관련이 있으면, 회귀계수의 추정치에 심각한 편향이 발생할 수 있다.



(a) 가중치를 사용하지 않는 회귀직선 적합　　(b) 가중치를 사용한 회귀직선 적합

출처 : Korn & Graubard(1995)

♣ 예를 들어 학교를 대상으로 조사하는 경우에 학급을 PSU로 하면 같은 학급 내의 학생들은 대개 어떤 학교 관련 문제에 대해서 비슷하게 응답하는 경향이 있다. 조사 데이터 분석에서 오차항 가정과 관련된 사항을 무시하면 가중치를 사용함으로 회귀계수의 추정에는 큰 문제가 없다. 그러나 회귀계수 추정량의 분산을 과소 평가함으로써 가설검증에 문제가 되어 제2종오류가 커진다. 이러한 오차항의 가정과 관련된 문제, 가중치, 표본설계 등을 모두 고려한 분산 추정법은 조사 데이터 분석용 소프트웨어에서 제공하는 선형화 방법, 잭나이프 방법, 균형반분법 등을 이용해야 한다.

(3) 회귀계수의 추정

▪ 회귀계수의 표현

$$B_1 = \frac{\sum_{i=1}^{N} x_i y_i - \frac{\left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{N}}{\left(\sum_{i=1}^{N} x_i^2\right) - \left(\sum_{i=1}^{N} x_i\right)^2 / N}, \quad B_0 = \frac{\left(\sum_{i=1}^{N} y_i\right) - B_1\left(\sum_{i=1}^{N} x_i\right)}{N}$$

▪ 회귀계수의 추정

- 가중치 w_i를 이용한 가중최소제곱법 이용

$$- \quad \hat{B}_1 = \frac{\sum_{i\in S} w_i x_i y_i - \frac{(\sum_{i\in S} w_i x_i)(\sum_{i\in S} w_i y_i)}{\sum_{i\in S} w_i}}{\sum_{i\in S} w_i x_i^2 - (\sum_{i\in S} w_i x_i)^2 / \sum_{i\in S} w_i},$$

$$\hat{B}_0 = \frac{\sum_{i\in S} w_i y_i - \hat{B}_1 \sum_{i\in S} w_i x_i}{\sum_{i\in S} w_i}$$

▪ 추정량의 분산 추정

- 선형화 방법 또는 잭나이프 방법 이용

<참고> ① 선형화 방법 적용 : $\hat{V}_L(\hat{B}_1) = \frac{n\sum_{i\in S}(x_i - \bar{x}_S)^2(y_i - \hat{B}_0 - \hat{B}_1 x_i)^2}{(n-1)\left[\sum_{i\in S}(x_i - \bar{x}_S)^2\right]^2}$

② 전통적인 방법 적용 : $\hat{V}_M(\hat{\beta}_1) = \frac{n\sum_{i\in S}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{(n-2)\left[\sum_{i\in S}(x_i - \bar{x}_S)^2\right]^2}$

(4) 조사 데이터의 회귀분석에서 표본설계를 반영해서 분석해야 하는 경우

▪ 공공 목적으로 사용되는 공식통계(official statistics)를 만들고자 하는 경우

▪ 표본이 확률추출법에 의해서 추출되었고, 표본의 크기가 충분히 큰 경우

♣ 만약 확률추출법에 의해서 추출되지 않았거나 표본의 크기가 작다면 전통적인 통계분석 방법에 따라서 분석해야 한다. 과학적 이론이나 과거 연구로부터 충분히 타당성이 입증된 모형이 있다면 이를 이용할 수 있다.

(5) 우리 나라 소규모 사업체 조사에 대한 회귀분석 적용 사례

▪ 전통적인 추정법(비가중 추정, 표본설계 무시) 적용

⇒ 추정량의 편향, 분산 과소평가 문제

▪ <부록 1>의 내용 참고

- 비가중 회귀계수 추정량의 편향(bias)는 29.2%-91.7%으로 나타남

- 회귀계수 추정량의 설계효과(meff)는 1.39-5.15로 나타남

3. 카이제곱 검증

참고문헌 : Lohr(1999) Chap 10, Lee et al.(1989) Chap 7 내용

(1) 간단한 사례

① SRS로 추출된 500쌍의 부부 조사

(i) 가정에 개인용 PC가 있는가?, (ii) 케이블 TV에 가입했는가?

② 조사 결과

		Computer?		
		Yes	No	
Cable?	Yes	119	188	307
	No	88	105	193
		207	293	500

③ 독립성 가정에서의 기대 도수

		Computer?		
		Yes	No	
Cable?	Yes	127.1	179.9	307
	No	79.9	113.1	193
		207	293	500

$$\hat{m}_{ij} = n\hat{p}_{i+}\,\hat{p}_{+j} = n\frac{x_{i+}}{n}\,\frac{x_{+j}}{n}$$

④ 피어슨의 카이제곱 검증통계량

$$X^2 = \sum_{\text{모든 셀}} \frac{(\text{관측도수} - \text{기대도수})^2}{\text{기대도수}} = 2.281$$

$$G^2 = 2\sum_{\text{모든 셀}} \text{관측도수} \ln\left(\frac{\text{관측도수}}{\text{기대도수}}\right) = 2.275$$

⑤ X^2과 G^2은 귀무가설 하에서 자유도 $(r-1)(c-1)$인 카이제곱분포를 따름

$\chi^2_{(1,\,0.05)} = 3.84$

(2) 단순임의표본에 대한 독립성 검증

① 자료구조

		C			
		1	2	⋯	c	
R	1	p_{11}	p_{12}	⋯	p_{1c}	p_{1+}
	2	p_{21}	p_{22}	⋯	p_{2c}	p_{2+}
	⋮	⋮	⋮		⋮	⋮
	r	p_{r1}	p_{r2}	⋯	p_{rc}	p_{r+}
		p_{+1}	p_{+2}	⋯	p_{+c}	1

② 귀무가설과 기대도수 계산

- $H_0: p_{ij} = p_{i+}\,p_{+j}$ for $i = 1, \dots, r$ and $j = 1, \dots, c$.
- $\hat{m}_{ij} = n\hat{p}_{i+}\,\hat{p}_{+j} = n\frac{x_{i+}}{n}\,\frac{x_{+j}}{n}$
- $\hat{p}_{ij} = x_{ij}/n, \quad \hat{p}_{i+} = \sum_{j=1}^{c}\hat{p}_{ij}, \quad \hat{p}_{+j} = \sum_{i=1}^{r}\hat{p}_{ij}$

③ 검증통계량

- $X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(x_{ij}-\hat{m}_{ij})^2}{\hat{m}_{ij}} = n\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(\hat{p}_{ij}-\hat{p}_{i+}\hat{p}_{+j})^2}{\hat{p}_{i+}\hat{p}_{+j}}$
- $G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c}x_{ij}\ln\left(\frac{x_{ij}}{\hat{m}_{ij}}\right) = 2n\sum_{i=1}^{r}\sum_{j=1}^{c}\hat{p}_{ij}\ln\left(\frac{\hat{p}_{ij}}{\hat{p}_{i+}\hat{p}_{+j}}\right)$

(3) 복합표본조사 데이터에 대한 독립성 검증

① 귀무가설 하에서 X^2과 G^2은 자유도 $(r-1)(c-1)$인 χ^2분포를 따르지 않음

⇒ 계산되는 p-값이 보통 실제보다 작게 계산됨

⇒ 실제로는 통계적으로 有意하지 않은 현상을 有意한 것으로 판정 내릴 수 있음

⇒ 특히 집락추출인 경우에 집락내 상관계수가 양수인 경우에 문제가 됨

② 독립성 검정에 대한 귀무가설

$H_0: p_{ij} = p_{i+}\,p_{+j}$ for $i = 1, \dots, r$ and $j = 1, \dots, c$

$\Leftrightarrow H_0: \theta_{11}=0,\ \theta_{12}=0, \cdots, \theta_{r-1,c-1}=0,$

여기서, $\theta_{ij} = p_{ij} - p_{i+}p_{+j}$로 정의

③ Wald 통계량 : 2×2 분할표의 경우 예시

- $\theta = p_{11} - p_{1+}p_{+1} = p_{11}p_{22} - p_{12}p_{21}$, $\hat{\theta} = \hat{p}_{11}\ \hat{p}_{22} - \hat{p}_{12}\ \hat{p}_{21}$로 정의
- $H_0 : p_{ij} = p_{i+}\ p_{+j} \Leftrightarrow H_0 : \theta = 0$
- 표본크기가 충분히 크게 되면 $H_0 : \theta = 0$ 하에서

$$\frac{\hat{\theta}}{\sqrt{\hat{V}(\hat{\theta})}} \sim N(0,\ 1),$$

$$X_w^2 = \frac{\hat{\theta}^2}{\hat{V}(\hat{\theta})} \sim \chi^2(1) \ : \text{Wald 통계량}$$

④ Wald 통계량 : 일반적인 경우

- $H_0 : p_{ij} = p_{i+}\ p_{+j}$ for $i = 1, \ldots, r$ and $j = 1, \ldots, c$.

  $\Leftrightarrow H_0 : \theta = 0,\ \theta = [\theta_{11}, \theta_{12}, \cdots, \theta_{r-1,c-1}]^T$

  $\Leftrightarrow H_0 : \theta_{11} = 0,\ \theta_{12} = 0,\ \cdots,\ \theta_{r-1,c-1} = 0$
- $X_w^2 = \hat{\theta}^T\ \hat{V}(\hat{\theta})^{-1}\hat{\theta} \sim \chi^2((r-1)(c-1))$

  표본설계를 반영한 분산 추정방법에 의해서 $\hat{V}(\hat{\theta})$ 을 구한다.
- Wald 통계량 검증은 범주 수가 많은 경우에는 좋지 못함

⑤ Bonferroni 방법

- 독립성 검정에 대한 귀무가설

  $H_0 : \theta_{11} = 0,\ \theta_{12} = 0,\ \cdots,\ \theta_{r-1,c-1} = 0$
- 귀무가설을 구성하는 각 요소에 대해서 유의수준 α/m 의 가설검정 시행

  ($m = (r-1) \times (c-1)$)

$$\frac{|\hat{\theta}_{ij}|}{\sqrt{\hat{V}(\hat{\theta}_{ij})}} > t_k\left(\frac{\alpha}{2m}\right) \text{ 이면 귀무가설 기각}$$

⑥ Rao and Scott의 방법

<참고> ① 가중치를 이용해서 각 셀의 비율을 추정해서 대입하는 경우

- $\hat{p}_{ij} = \dfrac{\sum_{k \in S} w_k y_{kij}}{\sum_{k \in S} w_k}$ 이용

여기서, $y_{kij} = \begin{cases} 1 & \text{만약 } k\text{번째 단위가 셀 } (i, j)\text{에 속하면} \\ 0 & \text{그렇지 않은 경우} \end{cases}$

표본설계에서 사용된 가중치를 이용해서 각 셀의 비율이나 도수를 추정해서 대입하는 경우는

표본의 크기가 마치 $\sum_i w_i$인 단순임의표본으로 간주하여 계산하는 것으로 가설검증에서 오류의 가능성이 높다.

② 층화만 사용되고 집락추출은 사용되지 않은 경우

설계효과(meff)가 1보다 작게 나타나며, 이에 따라 실제 값보다 p값이 크게 계산된다.

③ 집락추출이 사용된 경우

$\hat{p}_{ij}$에 대한 설계효과(meff)는 1보다 크게 나타나며, 집락추출을 무시하고 단순임의표본인 것처럼 분석하면 실제보다 검증통계량 X^2과 G^2의 값을 크게 하게 된다. p값이 작게 계산된다. 결과적으로 실제는 통계적으로 유의하지 않지만 유의한 것으로 나타나게 된다.

(4) 실제 분석 사례 : 『국민구강건강실태조사』중 일부

- 65세 이상 노인을 대상으로 성별과 부정구강진료 여부의 독립성 검정
- 조사 데이터

	부정구강진료 경험 여부		계
	있 음	없 음	
남 자	129	322	451
여 자	261	488	749
계	390	810	1200

- SAS를 이용하여 분석한 결과 : 비가중, 단순임의표본 가정

검증통계량	자유도	통계량값	p값
Chi-Square	1	5.002	0.025
Likelihood Ratio Chi-Square	1	5.050	0.025

- SUDAAN을 이용하여 분석한 결과 : 표본설계와 가중치 이용

검증통계량	자유도	통계량값	p값
Chi-Square	1	3.980	0.0474

4. 로지스틱모형 분석

(1) 간단한 사례

- 반응변수 : 부정구강진료여부(부정구강진료 경험 : 1, 무경험 : 0)
- 설명변수 : 성별, 지역, 학력, 연령, 소득, 직업 등
- 부정구강진료 경험 여부를 설명변수들을 이용해서 어떻게 모형화할 것인가?

(2) 로지스틱모형

- $\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$

  " 로짓(logit)을 설명변수의 선형함수로 모형화 "

(3) SRS 표본의 경우

- 최대가능도추정법(Maximum Likelihood Method)를 이용하여 β_0, $\beta_1, \cdots, \beta_k$ 추정

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}, \text{ 여기서 } p_i = \frac{\exp(x_i{}^T\beta)}{1+\exp(x_i{}^T\beta)}$$

- Newton-Raphson 방법 이용
- SAS CATMOD 사용

(4) 복합표본조사 데이터에 대한 로지스틱 모형

- 가중최대가능도함수를 이용하여 추정

$$\hat{L}(\beta) = \prod_{i\in S} p_i^{w_i y_i}(1-p_i)^{w_i(1-y_i)}, \text{ 여기서 } p_i = \frac{\exp(x_i{}^T\beta)}{1+\exp(x_i{}^T\beta)}$$

- Newton-Raphson 방법 이용
- 선형화 방법에 의한 분산 추정 : Binder(1983)
  잭나이프 방법이나 BRR 방법에 의해서 분산 추정

(5) 실제 조사 데이터 분석 사례 : 『국민구강건강실태조사』중 일부 65세 이상 노인 대상

- 전체적으로 계수 추정에 대한 설계효과(meff)는 1.04-1.74로 나타남
- 전통적인 추정법에 따라서 추정할 경우에 발생하는 편향은 상당히 크게 나타남

〈표 5〉 65세 이상 노인의 부정구강진료(경험 : 1, 무경험 : 0)에 대한 모형 적합 결과

변 수		회귀계수	p-값	편향(%)	설계효과(meff)
Intercept		-1.91 ±0.54	0.00	-32.46	1.12
성별	남자	-0.19 ±0.19	0.32	78.95	1.47
	여자	0.00	.	.	.
지역 구분	대도시	0.70 ±0.27	0.01	-31.43	1.02
	중소도시	0.13 ±0.28	0.63	-184.62	1.04
	군지역	0.00	.	.	.

교육 정도	무학	1.22 ±0.47	0.01	-33.61	1.24
	초등졸	1.14 ±0.44	0.01	-14.91	1.07
	중졸	0.23 ±0.57	0.68	78.26	1.29
	고졸	0.97 ±0.48	0.04	-30.93	1.19
	전문대졸 이상	0.00	.	.	.
직업	1(전문,사무,서비스)	-0.33 ±0.53	0.54	-115.15	1.30
	2(농업)	0.38 ±0.21	0.08	5.26	1.05
	3(근로자)	0.87 ±0.32	0.01	-50.57	1.51
	4(주부)	-0.22 ±0.25	0.39	68.18	1.38
	5(기타)	0.00	.	.	.
가구 월수입		-0.14 ±0.06	0.02	7.14	1.74
평 균				-15.05	1.26

5. 복합표본조사 데이터 분석용 소프트웨어 이용

(1) SUDAAN

<예시 프로그램>

```
PROC REGRESS DATA="c:\\den_data\\su_data" FILETYPE=SAS DESIGN=WOR;
NEST stratum emu_sn;
WEIGHT wgt_n;
TOTCNT ps_psu_n ss_psu_n;
SUBPOPN age>17;
SUBGROUP gender gubun2 n_tb_c a1 a3 s1 ciga_i o_h_d s2;
LEVELS 2 3 4 3 3 3 5 2 4;
REFLEVEL gender=2 ciga_i=1 n_tb_c=4 a1=3 o_h_d=2;
MODEL gum_idx=sex age_c1 fam_in edu_c ciga_i a1 n_tb_c o_h_d;
SETENV COLWIDTH=15;
SETENV DECWIDTH=4;
```

(2) Stata

<예시 프로그램>

```
 .use c:\den_data\test.dta
.svyset strata STRATUM
```

.svyset pweight WEIGHT

.svyset psu PSU

.svymean GUM_IDX

.svylogit GUM_IDX GENDER AGE FAM_IN EDU_C

Ⅳ. 결 론

1. 조사 데이터는 표본설계와 가중치를 반영하여 분석해야 한다. 실제적으로는 조사 데이터 분석을 위해서는 조사 데이터 분석용 패키지를 이용해야 한다.
2. 조사 데이터 분석용 패키지를 사용하지 않고, 분석하는 경우에 나타날 수 있는 문제들은 다음과 같다.
   (1) 추정에 편향이 생길 수 있다.
   (2) 표집분산의 과소 추정이 있을 수 있음
      ⇒ 가설검증에서 2종 오류가 크게 나타나게 됨
         통계적으로 유의하지 않은 사실을 유의한 것으로 나타남
      ⇒ 신뢰구간의 폭이 지나치게 작아짐
3. 조사 데이터를 분석할 때 요구되는 절차들 : Levy and Lemeshow(1999)
   (1) 표본설계에 대한 다음의 사항을 명확히 한다.
      - 층화 구분 변수
      - 집락 구분 변수
      - 유한모집단 수정항을 위한 모집단 크기
   (2) 위에 제시된 정보에 기초해서 설계가중치를 산정한다.
   (3) 무응답 보정과 사후층화 보정 단계를 거쳐서 최종 가중치를 산정한다.
   (4) 조사 데이터가 층화, 집락, 모집단 크기 등에 대한 정보를 담고 있는지 확인한다.
   (5) 알맞은 조사 분석용 소프트웨어를 선택해서 분석하고, 그 결과를 해석한다.

참고문헌

성내경(2000), "조사 데이터 분석용 소프트웨어 패키지", 조사연구, 1권 1호., pp. 109-123.

이기재(2001), "Design and Weight effect in small firm survey in Korea.", 한국통계학회 춘계학술발표대회 논문집.

이기재(2001), "복합표본조사 데이터 분석을 위한 회귀모형 접근법의 비교", 조사연구, 2권 1호, pp. 73-86.

이기재(2002), "국민구강건강실태조사에 대한 설계 및 가중치 효과 분석 ", 한국통계학회춘계학술발표대회 논문집

Korn, E.L. and Graubard, B. I. (1995), "Examples of Differing Weighted and Unweighted Estimates from a Sample Survey," The American Statistician, vol. 49, pp. 291-295.

Korn, E.L. and Graubard, B. I. (1999), *Analysis of Health Surveys*, John Wiley & Sons, Inc.

Lee, E.S., Forthofer, R.N. and Lorimor, R.J. (1989), *Analyzing Complex Survey Data*, Sage University Papers 71: Quantitative Applications in the Social Sciences.

Levy, P.S. and Lemeshow, S. (1999), *Sampling of Populations*, 3rd ed, John Wiley & Sons, Inc.

Lohr, S.L. (1999), *Sampling: Design and Analysis*, Duxbury Press

〈부록 1〉『소규모사업체근로실태조사』에 대한 표본설계

① 조사 목적

- 임금, 근로시간(정상, 초과), 근로일수 등의 근로실태와 임금결정요인의 분석

② 조사 내용

- 사업체 관련 사항 : 근로자 수, 지역, 산업, 임금인상 여부, 퇴직금 지급 등
- 근로자 관련 사항 : 성별, 나이, 학력, 입직경로, 근속년수, 직종, 근로시간, 월급여액, 연간 특별급여액 등

③ 표본설계 개요

- 조사 대상 : 근로자 1-4인 전 사업체(농·임·어업 부분 제외)
- 표본설계 : 층화집락추출법
  - 층화변수 : 산업중분류(52)
  - 1단계 추출 : 사업체
    2단계 추출 : 표본 사업체 내의 전체 상용근로자
  - 표본크기 : 조사사업체 수는 14,920개소, 조사근로자수는 33,116명

♣ 노동부에서 실시되고 있는『임금구조기본통계조사』,『비정규근로자실태조사』 등은 모두 1차 추출단위로 사업체, 2차추출단위로 근로자로 하는 통계조사들이다.

〈부록 2〉 소규모 사업체조사에 대한 회귀모형 적합 결과

<표 A-1> 회귀모형의 적합에 사용된 독립변수와 종속변수 목록

변수 이름	변수 종류	가변수 명	비고
ln(월평균 임금총액)	연속형		종속변수
산업대분류	가변수(10)	광업+제조업, 수도·가스·전기+건설업, 도·소매업, 음식·숙박업, 운수·창고·통신업, 금융·보험업, 부동산업, 교육 서비스업, 건강·사회 서비스업, [기타 서비스업]	사업체 단위 변수
지역	가변수(3)	서울, 광역시, [시·군지역]	
근로자 수	연속형		
직종	가변수(8)	관리자 및 입법자, 전문가, 기술공 및 준전문가, 사무직원, 판매원, 기능 근로자, 장치 조작원, [단순 노무직 근로자]	근로자 단위 변수
학력	가변수(4)	중학교 졸업 이하, 고등학교 졸업, 전문대 졸업, [대학 졸업 이상]	
성별	가변수(2)	남성, [여성]	
종사기간	연속형		
연령	연속형		

<표 A-2> 월평균 임금총액에 대한 회귀계수 추정치(종속변수 : ln(월평균 임금총액))

변수명	OLS 방법		가중치, 표본설계 반영		
	회귀계수	(s.e)	회귀계수	(s.e)	meff
광업+제조업[a]	0.129[c]	(0.0069[d])	0.154[c]	(0.0130[d])	3.31[e]
수도, 가스, 전기+건설업[b]	0.127	(0.0110)	0.107	(0.0198)	4.37
도소매업	0.133	(0.0073)	0.147	(0.0128)	4.73
음식 숙박업	0.138	(0.0090)	0.122	(0.0153)	4.52
운수, 창고, 통신업	0.066	(0.0076)	0.084	(0.0152)	1.63
금융·보험업	0.221	(0.0088)	0.267	(0.0150)	1.39
부동산업	0.118	(0.0077)	0.144	(0.0151)	3.13
교육 서비스업	0.010	(0.0121)	0.030	(0.0181)	3.59
건강, 사회 서비스업	0.207	(0.0125)	0.216	(0.0175)	3.54
기타 서비스업	0[f]		0[f]		
서울	0.073	(0.0039)	0.078	(0.0071)	3.32
광역시	-0.012	(0.0040)	-0.016	(0.0073)	3.34
시·군지역	0[f]		0[f]		
사업체 내의 근로자 수	0.023	(0.0016)	0.025	(0.0029)	3.45
관리자 및 입법자	0.312	(0.0111)	0.241	(0.0238)	4.09
전문가	0.169	(0.0125)	0.103	(0.0206)	3.16
기술공 및 준전문가	0.177	(0.0091)	0.122	(0.0163)	3.21
사무직원	0.143	(0.0076)	0.097	(0.0135)	2.96
판매원	0.091	(0.0083)	0.048	(0.0139)	3.52
기능 근로자	0.105	(0.0080)	0.060	(0.0133)	2.52
장치 조작원	0.130	(0.0084)	0.074	(0.0146)	2.48
단순 노무직 근로자	0[f]		0[f]		
중학교 졸업 이하	-0.152	(0.0072)	-0.166	(0.0128)	3.05
고등학교 졸업	-0.088	(0.0052)	-0.100	(0.0092)	3.08
전문대학교 졸업	-0.076	(0.0064)	-0.091	(0.0101)	2.58
대학 졸업 이상	0[f]		0[f]		
남성	0.286	(0.0039)	0.261	(0.0061)	2.49
1년 이하	-0.242	(0.0062)	-0.228	(0.0103)	2.81
1-3년	-0.166	(0.0056)	-0.154	(0.0093)	2.68
3-4년	-0.125	(0.0064)	-0.115	(0.0100)	2.35
4-5년	-0.108	(0.0067)	-0.100	(0.0103)	2.24
5-10년	-0.065	(0.0053)	-0.067	(0.0087)	2.45
근로자 연령	0.046	(0.0012)	0.043	(0.0021)	3.25
연령의 제곱	-0.001	(0.0000)	-0.000	(0.0000)	3.34
총 근로시간(월 단위)	0.001	(0.0000)	0.001	(0.0001)	5.05
절편	5.308	(0.0274)	5.364	(0.0493)	3.55
R	0.464		0.449		

단, a : 광업+제조업, b : 수도·전기·가스+건설업, c : 회귀계수, d : 표준오차, e : 설계효과,
f : 기준범주(Reference category)