복합표본조사 데이터 분석을 위한 회귀모형 접근법의 비교

- 소규모사업체조사 데이터 분석을 중심으로 -

(Comparison of regression model approaches fitted to complex survey data)

이 기 재*
Kee Jae Lee

복합표본조사 데이터 분석에서 회귀모형 접근법은 크게 표본설계 기반 접근법(design-based approach)과 일반화 추정 방정식 접근법(generalized estimating equations approach)으로 구분된다. 본 논문은 이들 접근법과 모형기반 접근법을 비교하여 설명하고, 각 접근법에서 표본설계가 모수 추정에 미치는 영향을 설계효과와 가중치 효과 분석을 통해서 살펴보았다.

In this paper, we conducted an empirical study to investigate the design and weighting effects on descriptive and analytic statistics. We compared the regression models using the design-based approach and the generalized estimating equations(GEEs) approach with the model-based approach through the design and weighting effects analysis.

Ⅰ. 서 론

대부분의 국가조사 또는 대규모 표본조사는 층화, 집락추출, 불균등확률추출 등이 복합적으로 사용된 복합표본설계에 따라 표본을 추출하여 조사하고 있다. 이것은 단순임의추출법을 적용했을 때 엄청난 비용과 조사의 어려움이 수반되기 때문이다.

본 논문은 노동부에서 매년 실시하고 있는 소규모사업체근로실태조사 데이터를 분석하여 복합표본설계가 회귀모형에 미치는 영향을 실증적으로 분석한 것이다. 소규모사업체근로실태조사는 5인 미만의 사업체를 대상으로 임금, 정상 및 초과 근로시간, 근로일수 등의 근로실태를 조사하는 것을 목적으로 한다. 조사결과는 5인 미만 사업

* 한국방송통신대학교 정보통계학과 부교수

체에 대한 임금, 근로시간 등에 대한 정기적인 공식통계 작성을 위해 사용되고 있다. 또한 고용보험 적용 사업장의 확대에 따른 5인 미만 사업체의 근로조건에 대한 기초자료로 사용되고 있다(이기재 등, 1999). 이 조사는 층화1단계집락추출법(stratified one-stage cluster sampling)에 의해서 14,942개 표본 사업체를 추출하여 표본으로 추출된 사업체 내의 상용근로자들을 조사하였다. 본 논문에서 분석하는 것은 1998년 11월 기준으로 조사된 데이터이다.

현재 우리 나라에서 시행되고 있는 사업체 대상 조사는 산업대분류별, 직업별, 사업체 규모별, 성별, 학력별 등 다양한 하부영역에 대해서 추정치를 만드는 것이 필요하다. 이렇게 세부 영역별 추정치 생산을 위해서는 규모가 작은 영역에 대해서 상대적으로 높은 추출률로 표본을 추출하여 통계생산에 필요한 적정한 크기의 표본을 얻게 된다. 그런데 추출단위에 따라서 추출률을 달리 하면 추정과정에서 가중치의 배정이 필요하다. 가중치는 추출단계에서 추출률뿐만 아니라 무응답 조정, 사후층화 조정(post-stratification adjustment) 등을 보정하기 위해서도 필요하다. 만약 추정단계에서 가중치를 무시하고 분석하면 추정치에 심각한 편향(bias)이 발생할 수 있다.

일반적으로 복합표본조사 데이터를 분석하는 경우에 표준 통계 소프트웨어 패키지들은 가중분석을 지원하기 때문에 관심 모수에 대해서 비편향 점추정값을 얻는데는 어려움이 없다. 그러나 통계 소프트웨어 패키지로 가중치를 이용한 분석을 하는 경우에도 분산 추정에는 심각한 오류가 발생하게 된다(성내경, 2000).

소규모사업체근로실태조사는 층화(stratification)와 집락추출(cluster sampling) 등이 복합적으로 사용된 복합표본설계에 의해서 얻어진 표본을 대상으로 조사한다. 또한 소규모사업체근로실태조사는 표본의 모집단에 대한 대표성을 높이고, 추정량의 정도(精度)를 높이기 위해서 모집단(사업체기초통계조사)의 정보를 추정 단계에서 이용하는 사후층화(post-stratification)의 기법을 사용한다. 본 논문에서는 소규모사업체조사 데이터 분석을 위해서 복합표본조사 분석용 통계소프트웨어인 SUDAAN을 이용하였다.

먼저, 2장에서는 소규모사업체조사의 전체적인 측면과 복합표본설계가 각종 통계량의 추정에 미치는 영향에 대해서 설명하고, 3장에서는 복합표본조사 데이터에 회귀모형을 적합하는 경우에 고려할 수 있는 여러 종류의 회귀모형들에 대해서 설명한다. 4장에서는 소규모사업체조사에 세 가지 접근법에 따라서 회귀모형을 적합해서 표본설계와 가중치를 추정과정에서 무시하는 모형기반 접근법(model-based approach), 표본설계 기반 접근법 (design-based approach), 일반화 추정 방정식(generalized estimating equations : GEEs)에 의한 접근법 등 복합표본조사 데이터를 분석할 때 회귀모형에 대한 세 가지 접근법을 비교한다. 마지막으로 간단한 결론을 덧붙인다.

Ⅱ. 소규모사업체조사의 표본설계

1. 조사개요

소규모사업체근로실태조사는 5인 미만의 사업체를 대상으로 임금, 정상 및 초과 근로시간, 근로일수 등의 근로실태를 조사하는 것을 목적으로 한다. 조사결과는 5인 미만 사업체에 대한 정기적인 임금, 근로시간 등의 공식통계 작성에 이용되고 있다. 조사항목은 크게 사업체에 관한 사항과 근로자에 대한 사항으로 구분되는데 각각에 대한 주요 조사항목은 다음과 같다.

- ㅇ 사업체에 관한 주요 조사항목 : 해당 사업체의 산업대분류 구분, 사업체의 형태, 임금인상 여부, 퇴직금 지급여부에 대한 사항, 휴가 실시에 대한 사항, 업무상 치료에 대한 사항, 성별 상용근로자 수 등
- ㅇ 근로자에 대한 주요 조사항목 : 성별, 연령, 학력, 입직경로, 근속년수, 경력년수, 직종, 출근일수, 실근로시간, 월급여액, 연간특별급여액 등

이 조사의 대상은 농업, 임업, 어업, 수렵업 등과 국가 또는 지방행정기관, 군·경찰 및 국·공립교육기관을 제외한 사업체노동실태조사 결과 중 상용근로자 5인 미만의 전 사업체이다. 이 조사는 층화1단계집락추출법(stratified one-stage cluster sampling)에 의해서 14,942개 사업체를 추출하여 표본 사업체 내의 모든 상용근로자들을 조사하였다.

표본으로 추출된 상용근로자에 부여되는 가중치는 특정 응답자가 전체 모집단에 대해 대표하는 정도를 나타내고, 모집단 특성치에 대한 비편향 추정량(unbiased estimator)을 얻기 위해서 사용된다. 일반적으로 가중치는 설계 가중치(design weight), 무응답 조정(non-response adjustment), 사후층화 조정(post-stratification adjustment) 요인의 곱으로 구해진다. 다음의 <표 1>은 표본으로 추출된 근로자에 대해서 부여된 가중치에 대한 기술통계량 요약이다.

<표 1> 가중치에 대한 기술통계량 요약

구 분	남자 (n=19580)	여자 (n=13536)	전체 (n=33116)
최소값	2.78	1.00	1.00
25%	8.79	10.76	8.79
중앙값	19.60	40.42	26.28
75%	75.12	111.47	89.01
최대값	154.13	123.23	154.13
평 균	41.40	56.52	41.58
CV(%)	98.95	81.34	91.89

2. 표본설계 및 가중치 효과

복합표본설계가 추정량에 미치는 효과는 설계효과를 통해서 평가될 수 있다. 설계효과(design effect: DEFF)는 복합표본설계에 따라 구해진 추정량의 분산과 같은 크기의 표본을 단순임의추출법에 따라서 추출되었다고 가정하여 구한 분산의 비(ratio)이다. 어떤 조사설계 D에 대해서 모수 θ의 추정량으로 $\hat{\theta}$을 사용할 때 설계효과(DEFF)는 다음과 같이 추정할 수 있다.

$$\text{deff}(\hat{\theta}) = \frac{\hat{V}_D(\hat{\theta})}{\hat{V}_{SRS}(\hat{\theta})} \quad (1)$$

대부분의 경우 집락추출법이 적용되면 설계효과는 1보다 크게 나타나는데, 그 의미는 우리가 고려하고 있는 조사설계가 같은 표본크기의 단순임의추출법에 비해서 추정량의 분산이 커짐을 의미한다.

Kish와 Frankel(1974)은 경험적 분석을 통해서 모평균, 모비율, 선형 회귀계수 등에 대해서 복합표본설계의 설계효과를 평가하였다. 일반적으로 설계효과는 다양하게 사용될 수 있는데, 예를 들어 복합표본조사의 유효 표본크기(effective sample size)를 구하는데 사용될 수 있고, 통상적인 방법으로 추정량의 분산을 계산할 수 없는 경우에 이를 근사적으로 추정하는 방법으로 사용될 수 있다.

일반적으로 추정과정에서 가중치를 무시하고 분석하면 모수 추정에 심각한 편향(bias)이 발생할 수 있다. 반면에 추정과정에서 가중치를 이용할 경우에는 일반적으로 추정량의 표집오차(sampling error)가 증가한다. 불균등 가중치를 사용함으로써 발생하는 표집오차의 증가분에 대한 추정방법은 Kish(1965), Korn과 Graubard(1995) 등에 의해서 제안되었다. 한편 이기재(2001)는 모의실험을 통해서 Kish에 의해 제안된 방법이 약간 과대 추정하는 경향이 있지만 합리적인 기준이라는 점을 보여주었다. 모의실험 결과 불균등 가중치 사용에 따른 추정량의 분산 증가분은 약 46%로 나타났다.

다음의 <표 2>는 주요 관심변수에 대하여 모평균 추정을 위해서 가중 평균을 사용

하는 경우의 설계효과와 비가중 평균을 사용할 때 발생하는 편향의 크기를 정리한 것이다.

<표 2> 가중평균에 대한 설계효과와 비가중 평균에 대한 편향

변 수	가중 평균	설계효과	비가중 평균	편향 (%)
월급여액	880.15	2.75	905.32	2.86
ln(월급여액)	6.70	2.86	6.73	0.45
정상근로시간	212.08	3.78	208.22	-1.82
초과근로시간	21.04	3.57	16.49	-21.63
나이	33.69	2.38	34.57	2.61

즉, 복합조사 데이터를 분석할 때 층화, 집락추출, 가중치 등의 조사설계를 분석단계에서 반영하지 않으면 모수의 점추정치에 심각한 편향이 발생할 수 있고, 추정량의 분산이 과소평가되어 문제가 된다.

Ⅲ. 복합표본조사 데이터 분석을 위한 회귀모형 접근법들

1. 모형 기반 접근법

일반적으로 복합표본조사를 통해서 얻어진 데이터에 회귀모형을 적합하는 방법은 (ⅰ)모형 기반 접근법, (ⅱ) 설계 기반 접근법, (ⅲ) 일반화 추정 방정식 접근법 등으로 구분할 수 있다. 모형 기반 접근법(model-based approach)은 추정 과정에서 표본설계나 가중치를 고려하지 않으며 복합표본조사 이외의 분야에서 통상적으로 사용되고 있는 익숙한 방법으로 최소제곱법(ordinary least square method: OLS)을 이용하여 회귀계수를 추정한다. 그러나 복합표본조사 데이터에 대해 모형 기반 접근법으로 회귀모형을 적합할 때 추정치에 편향이 발생하고, 추정량의 분산이 과소평가된다는 점은 널리 알려진 사실이다.

대부분의 경우 통계모형을 적용할 때 중요한 것은 가능한 단순한 모형을 통해서 가장 효율적인 방법으로 모수를 추정하는 것이다. 이 경우에 표본 크기는 작거나 약간 큰 정도이다. 반면에 대부분의 복합표본조사에서 표본의 크기는 상당히 크고, 이러한 이유로 데이터의 분석에서 중요시 되는 것은 추정의 효율성보다는 모형 가정이 성립하지 않는 경우에 대한 강건성(robustness)이다.

2. 설계 기반 접근법

설계 기반 접근법(design-based approach)은 표본이 추출된 모집단이 있다는 기본적인 가정에서 출발한다. 그래서 모형의 회귀계수 벡터 β를 추정하는 대신에 모집단 특성치인 $B=(X'X)^{-1}X'y$를 추정하는 것을 목적으로 한다. 여기서 y는 반응변수(종속변수)에 대한 유한모집단의 전체 값들을 뜻하고, X는 유한 모집단에 대한 설명변수 행렬이다. 결과적으로 설계 기반 접근법은 $B=(X'X)^{-1}X'y$ 자체를 하나의 모수로 간주해서 추정하는 것이다.

일반적으로 회귀모형을 적합하는 가장 중요한 이유는 현재의 데이터를 생성하였다고 볼 수 있는 회귀모형을 적합하여 그 계수를 추정하는 것이라고 할 수 있다. 설계 기반 접근법은 이러한 적합된 회귀모형의 의미를 약하게 하는 측면이 있지만 전체적으로 회귀모형 접근법의 확장된 개념으로 볼 수 있다(Kott, 1991).

일반적으로 가중치를 이용한 회귀계수 추정방법은 일부 설명변수가 분석에서 누락되는 경우에도 상당히 강건한(robust) 추정 방법이다(Pfeffermann과 Holmes, 1985). 따라서 회귀모형의 설명변수들 중에 누락된 설명변수가 있을 것으로 판단되면 설계 기반 접근법에 의한 가중최소제곱법이 알맞다. 한편 회귀계수 추정량에 대해서 복합표본조사 분석용 통계소프트웨어에서 널리 사용되고 있는 선형화 방법을 적용하면 오차항에 대한 통상적인 가정인 독립성과 등분산성 등이 위배되어도 근사적인 비편향(unbiased) 추정이 가능한 장점이 있다.

3. 일반화 추정방정식 접근법

일반화 추정방정식 접근법(generalized estimating equations approach)은 일반화 선형모형(generalized linear model)에서 관측치 간에 서로 상관관계가 있는 경우에 이를 반영하기 위한 방법으로 소개되었다. 앞으로 일반화 추정방정식 접근법은 GEE 접근법으로 표기하도록 한다. GEE 접근법은 관측치가 이산형과 연속형 모두의 경우에 적용될 수 있는 일반적인 접근법이지만 본 논문에서는 회귀모형인 경우로 국한한다.

GEE 접근법은 관측치에 대한 엄밀한 분포가정을 필요로 하지는 않는다. 다만 모평균에 대한 함수를 설명변수의 선형함수로 정의하고, 관측치 간의 의존성(dependence)이 존재하는 경우에도 분석할 수 있다는 것이 특징이다. 또한 관측치 간에 가정된 상관관계 모형이 실제와 다른 경우에도 GEE 접근법으로 추정된 결과는 일치성을 만족하고, 추정량의 분포가 점근적으로 정규분포를 따르게 된다(Liang과 Zeger, 1986).

집락추출을 통해서 얻어진 데이터의 분석에서 같은 집락에 소속된 서로 다른 두 조

사단위는 서로 같은 정도로 상관관계를 갖고, 서로 다른 집락에 속한 조사단위는 서로 독립적이라고 가정하는 것이 합리적이다(Bieler와 Williams, 1995). 소규모사업체조사의 경우에 같은 사업체에서 조사되는 근로자들은 이와 같은 가정을 만족한다고 할 수 있다. 이 경우에 종속변수로 사용되고 있는 월급여액의 로그 변환값에 대한 급내상관계수는 $\rho_Y=0.612$로 나타나서 집락내에서 근로자들이 받는 월급여액은 상당히 유사성이 있다고 할 수 있다.

설계 기반 접근법은 기본적으로 오차항에 대한 가정을 필요로 하지 않기 때문에 급내상관계수(intracluster correlation)가 존재하는 경우에도 사용될 수 있다. 하지만 모수 추정 단계에서 급내상관계수가 사용되는 것은 아니다. 집락 내의 관측치들이 서로 독립이라고 가정할 수 있는 경우에 GEE 접근법과 설계 기반 방법은 근사적으로 같다. SUDAAN에서 설계 기반 접근법은 GEE 접근법의 특수한 경우(급내상관계수 $\rho=0$인 경우)로 간주하여 추정할 수 있다(Shah 등, 1997).

Ⅳ. 소규모사업체조사에 대한 회귀모형의 비교

1. 회귀모형 적합

본 연구에서는 좀 더 신뢰성이 있는 회귀계수와 그 분산 추정을 위해서 표본의 크기가 작은 광업과 제조업을 합치고 가스, 수도 및 전기업과 건설업을 합쳐서 분석하였다. 제시되는 결과 중에서 통상적 최소제곱법(OLS)에 의한 결과는 SAS의 PROC REG 절차를 통해서 얻어졌고, 설계 기반 접근법과 GEE 접근법에 의한 분석결과는 SUDAAN을 이용해서 구해진 것이다. 여기서 산업대분류, 지역 구분, 사업체 크기 등은 사업체 단위의 변수이고, 직종, 학력, 근속년수, 성별, 근로시간 등은 근로자 단위의 변수이다. 회귀모형의 종속변수는 월평균 임금총액에 로그를 취한 값이다. 다음의 <표 3>은 회귀모형에서 사용된 독립변수와 종속변수를 정리한 것이다.

<표 3> 회귀모형의 적합에 사용된 독립변수와 종속변수 목록

변수 이름	변수 종류	가변수 명	비고
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
지 역	가변수(3)	서울, 광역시, [시·군지역]	
근로자 수	연속형		
직 종	가변수(8)	관리자 및 입법자, 전문가, 기술공 및 준전문가, 사무직원, 판매원, 기능 근로자, 장치 조작원, [단순 노무직 근로자]	근로자 단위 변수
학 력	가변수(4)	중학교 졸업 이하, 고등학교 졸업, 전문대 졸업, [대학 졸업 이상]	
성 별	가변수(2)	남성, [여성]	
종사기간	연속형		
연 령	연속형		

2. 설계 및 가중치 효과 분석

이 절에서는 부록에 제시되어 있는 설계기반 접근법과 GEE 접근법에 의한 회귀계수 추정량의 분산에 대해서 살펴본다. 일반적으로 GEE 접근법은 설계기반 접근법에 비해서 효율적이다. GEE 방법이 효율적인 것은 급내상관계수(intracluster correlation)가 존재하는 경우에 모수 추정 단계에서 이 정보가 사용되기 때문이다. 일반적으로 모형 기반 접근법은 모형이 정확하게 규정되었을 때 설계 기반의 방법에 비해서 효율적이다. 설계 기반 방법은 대부분의 복합조사에서 표본의 크기가 대단히 크기 때문에 분산 추정의 효율보다는 모형 가정이 어긋날 경우에 대한 강건성이 주요 관심이다. 일반적으로 설계 기반 접근법에 의한 회귀계수 추정은 모형 기반 접근법에 비해서 모형 가정에 대한 위배에 대해서 강건하지만 효율이 떨어진다는 점은 널리 알려져 있다.

일반적으로 OLS에 의한 회귀계수 및 그 분산 추정법을 복합조사 데이터에 적용하면 회귀추정량에 편향이 생기고, 분산 추정에 문제가 발생한다. 본 연구에서도 회귀모형을 적합시킨 결과 통상적인 모형 기반 방법은 상당한 편향이 발생하는 것으로 나타났다. 특히 직종 변수에 대한 회귀계수 추정에 대한 편향이 심각해서 29.2%-91.7%로 나타났다. 모형 기반 접근법에 대한 각 회귀계수의 편향은 부록에 수록된 <표 A-1>에서 계산될 수 있다. 다음의 <표 4>는 설계 기반 접근법과 GEE 접근법에 대해서 32개 회귀계수 추정량에 대한 설계효과를 정리하여 요약한 것이다.

<표 4> 각 접근법에 대한 회귀계수 추정에 대한 설계효과 요약통계량

구 분	설계기반 접근법	GEE 접근법 (exchangeable)
최소값	1.39	1.64
25%	2.55	2.10
중앙값	3.19	2.39
75%	3.53	3.18
최대값	5.05	4.12
평 균	3.16	2.63

잘 알려진 대로 가중치와 조사설계를 무시하고 구한 회귀계수 추정치의 표준오차는 설계 기반 접근법과 GEE 접근법에 비해서 대단히 과소평가되었다. 전체적으로 GEE 접근법에 의한 표준오차 추정값은 설계 기반 접근법에서 구한 표준오차와 크게 차이 나지는 않았다. 다만 세부적으로 살펴보면 집락 단위(사업체 단위)의 독립변수에 대한 회귀계수의 표준오차는 그 차이가 작고, 근로자 단위 독립변수에 대한 회귀계수의 표준오차에서 GEE 접근법은 설계 기반 접근법에 비해서 효율적인 것으로 나타났다. Lipsitz 등(1994), Bieler와 Williams(1995) 등은 집락추출법에 의해서 추출된 표본을 조사하여 얻은 데이터를 분석하면서 이와 같은 현상을 보고하고 있다.

다음의 <표 5>는 설계기반 방법과 GEE 접근법 각각에 대해서 가중치 효과를 요약하여 나타낸 것이다. 여기서 모의실험 결과는 원 복합조사 데이터로부터 100개의 동일 추출확률 부차표본을 추출하여 각각의 부차표본에 대해서 회귀모형을 적합하여 설계효과를 계산하였다. 자세한 모의실험 방법에 대해서는 이기재(2001)을 참고하기 바란다. 전체적으로 Kish(1965), Korn과 Graubard(1995)가 모평균 추정에서 제안한 가중치 효과 추정법이 회귀계수 추정의 경우에도 근사적으로 사용할 수 있음을 알 수 있다.

<표 5> 가중치의 사용에 따른 회귀계수 추정량의 비효율성 정도 요약표

구 분	Kish의 방법	K & G의 방법	모의실험 결과
설계기반 접근법	45.8%	42.9%[a] 44.7%[b]	43.4%[a] 45.4%[b]
GEE 접근법 (exchangeable correlation)		45.2%[a] 46.6%[b]	43.7%[a] 43.5%[b]

a : 32개 회귀계수 추정치에서 구한 비효율 정도에 대한 평균값
b : 32개 회귀계수 추정치에서 구한 비효율 정도에 대한 중앙값

설계기반 접근법과 GEE 접근법은 대단히 비슷한 결과를 제공하며, 다만 근로자 단위에서 조사된 설명변수에 대한 회귀계수의 추정에서 GEE 방법이 효율적인 것으로

나타났다. 한편 GEE 접근법을 이용하는 경우에도 추정단계에서 가중치를 사용하지 않는 경우에는 심각한 편향이 발생한다는 점을 알 수 있었다.

회귀모형은 서로 다른 설명변수 값의 분포를 갖는 두 그룹 사이의 비교를 위해서 사용될 수 있다. 예를 들어 월평균 임금총액의 산업대분류별 차이를 알고자 하는 경우를 생각해 보자. 각 산업대분류에 따라서 근로자 특성의 분포에 차이가 있을 것이다. 어떤 산업은 상대적으로 학력수준이 낮거나 근속년수가 낮은 근로자들이 주로 분포되어 있다면 상대적으로 낮은 임금을 받는 것으로 나타날 것이다. 이 경우에 단순하게 각 산업대분류별로 평균 임금을 구한다면 각 산업 내에서 근로자들의 특성 분포의 차이로 발생하는 차이를 설명할 수 없게 되어 문제다. 회귀모형을 적합하는 의미는 임금에 대한 모형을 찾는 것뿐만 아니라 적합된 모형을 통해서 산업대분류별, 직종별, 성별, 학력별 등 다양한 구분에 대한 임금을 비교할 목적으로 사용될 수 있다.

V. 결 론

본 연구에서는 추정과정에서 조사설계를 무시하는 모형 기반 접근법의 회귀계수 추정치는 설계 기반 접근법이나 GEE 접근법에 비해서 추정치의 편향이 크고, 그 표준오차가 대단히 과소평가된다는 점을 확인할 수 있었다. 또한 전체적으로 GEE 접근법에 의한 표준오차 추정값은 설계 기반 접근법에서 구한 표준오차와 큰 차이를 보이지는 않았지만, 세부적으로 살펴보면 집락 단위(사업체 단위)의 독립변수에 대한 회귀계수의 표준오차는 그 차이가 작고, 근로자 단위의 독립변수에 대한 회귀계수의 표준오차에서 GEE 접근법이 효율적인 것으로 나타났다.

표본조사 데이터를 제대로 분석하기 위해서는 반드시 조사 설계를 반영할 수 있도록 고안된 전문적인 소프트웨어 패키지를 활용해야 한다. 특히 관심 통계량의 분산을 추정하고자 할 때 표준 통계 소프트웨어 패키지의 분석 결과는 대부분의 경우 과소평가하기 때문에 신뢰구간 작성과 가설 검정에 심각한 문제를 불러일으킨다.

참 고 문 헌

1. 성내경. 2000. "조사 데이터 분석용 소프트웨어 패키지." 《조사연구》 1(1): 109-123
2. 이기재., 임금숙., 류제복. 1999. "소규모사업체근로실태조사를 위한 표본설계 연구.", 한국통계학회 추계학술발표대회 논문집.
3. 이기재. 2001. "Design and Weight effect in small firm survey in Korea.", 한국통계학회 춘계학술발표대회 논문집.
4. Bieler, G. S., and Williams, R. L. 1995. "Cluster Sampling Techniques in Quantal Response Teratology and Developmental Toxicity Studies.", Biometrics, 51, 764-776.
5. Kish, L., and Frankel, M. R. 1974. "Inference From Complex Samples.", Journal of the Royal Statistical Society, Ser. B, 36, 1-37.
6. Korn, E. L. and Graubard, B. I. 1995. "Analysis of Large Health Surveys: Accounting for the Sampling Design.", Journal of the Royal Statistical Society, Ser. A, 158, 263-295.
7. Kott, P. S. 1991. "A Model-Based Look at Linear Regression with Survey Data." American Statistician, 45(2), 107-112.
8. Liang, K. and Zeger, S. 1986. "Longitudinal data analysis using generalized linear model.", Biometriaka 73, 13-22.
9. Lipsitz, S. R., Fitzmaurice, G. M., Orav, E. J., and Laird, N. M. 1994. "Performance of Generalized Estimating Equations in Practical Situations.", Biometrics, 50, 270-278.
10. Pfeffermann, D., and Holmes, D. J. 1985. "Robustness Considerations in the Choice of Methods of Inference for Regression Analysis of Survey Data.", Journal of the Royal Statistical Society, Ser. A, 148, 268-278.
11. Shah, B. V., Barnwell, B. G., and Bieler, G. S. 1997. *SUDAAN Users Manual, Release 7.5.* Research Triangle Park, NC: Research Triangle Institute.

<표 A-1> 월평균 임금총액에 대한 회귀계수 추정치 (종속변수 : ln(월평균 임금))

변수명	OLS 방법	표본설계 기반		GEE (exchangeable)	
산업대분류 구분					
광업+제조업[a]	0.129[c] (0.0069[d])	0.154[c] (0.0130[d])	3.31[e]	0.156 (0.0133)	2.12
수도,가스,전기+건설업[b]	0.127 (0.0110)	0.107 (0.0198)	4.37	0.121 (0.0195)	2.17
도소매업	0.133 (0.0073)	0.147 (0.0128)	4.73	0.155 (0.0131)	2.89
음식 숙박업	0.138 (0.0090)	0.122 (0.0153)	4.52	0.111 (0.0155)	2.93
운수, 창고, 통신업	0.066 (0.0076)	0.084 (0.0152)	1.63	0.086 (0.0150)	0.87
금융 · 보험업	0.221 (0.0088)	0.267 (0.0150)	1.39	0.291 (0.0154)	0.79
부동산업	0.118 (0.0077)	0.144 (0.0151)	3.13	0.139 (0.0156)	1.75
교육 서비스업	0.010 (0.0121)	0.030 (0.0181)	3.59	0.035 (0.0178)	2.19
건강, 사회 서비스업	0.207 (0.0125)	0.216 (0.0175)	3.54	0.209 (0.0181)	2.20
기타 서비스업	0[f]		0[f]		0[f]
지역					
서울	0.073 (0.0039)	0.078 (0.0071)	3.32	0.082 (0.0069)	1.72
광역시	-0.012 (0.0040)	-0.016 (0.0073)	3.34	-0.010 (0.0073)	1.88
시 · 군지역	0[f]		0[f]		0[f]
사업체 내의 근로자 수	0.023 (0.0016)	0.025 (0.0029)	3.45	0.024 (0.0029)	1.91
관리자 및 입법자	0.312 (0.0111)	0.241 (0.0238)	4.09	0.275 (0.0194)	3.50
전문가	0.169 (0.0125)	0.103 (0.0206)	3.16	0.166 (0.0197)	2.95
기술공 및 준전문가	0.177 (0.0091)	0.122 (0.0163)	3.21	0.156 (0.0140)	2.51
사무직원	0.143 (0.0076)	0.097 (0.0135)	2.96	0.096 (0.0114)	2.25
판매원	0.091 (0.0083)	0.048 (0.0139)	3.52	0.070 (0.0115)	2.40
기능 근로자	0.105 (0.0080)	0.060 (0.0138)	2.52	0.084 (0.0124)	2.02
장치 조작원	0.130 (0.0084)	0.074 (0.0146)	2.48	0.090 (0.0126)	1.94
단순 노무직 근로자	0[f]		0[f]		0[f]
교육수준					
중학교 졸업 이하	-0.152 (0.0072)	-0.166 (0.0128)	3.05	-0.141 (0.0108)	2.68
고등학교 졸업	-0.088 (0.0052)	-0.100 (0.0092)	3.08	-0.084 (0.0077)	2.83
전문대학교 졸업	-0.076 (0.0064)	-0.091 (0.0101)	2.58	-0.074 (0.0082)	2.29
대학 졸업 이상	0[f]		0[f]		0[f]
성별					
남성	0.286 (0.0039)	0.261 (0.0061)	2.49	0.261 (0.0054)	2.71
종사기간					
1년 이하	-0.242 (0.0062)	-0.228 (0.0103)	2.81	-0.260 (0.0091)	2.58
1-3년	-0.166 (0.0056)	-0.154 (0.0093)	2.68	-0.179 (0.0081)	2.51
3-4년	-0.125 (0.0064)	-0.115 (0.0100)	2.35	-0.131 (0.0084)	2.26
4-5년	-0.108 (0.0067)	-0.100 (0.0103)	2.24	-0.115 (0.0088)	2.28
5-10년	-0.065 (0.0053)	-0.067 (0.0087)	2.45	-0.073 (0.0072)	2.40
10년 이상	0[f]		0[f]		0[f]
근로자 연령	0.046 (0.0012)	0.043 (0.0021)	3.25	0.040 (0.0017)	3.08
연령의 제곱	-0.001 (0.0000)	-0.000 (0.0000)	3.34	-0.000 (0.0000)	3.28
총 근로시간(월 단위)	0.001 (0.0000)	0.001 (0.0001)	5.05	0.001 (0.0001)	4.12
절편	5.308 (0.0274)	5.364 (0.0493)	3.55	5.329 (0.0429)	2.99
R^2	0.464		0.449		0.446

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