회귀모형에 의한 소지역추정

최지영[1] 최기헌[2] 한근식[3]

〈차 례〉

요 약

표본의 크기가 작은 경우 추정치의 정도에 문제가 발생한다. 본 연구에서는 대규모 조사에서의 표본을 소지역 혹은 소도메인에 할당하였을 경우 발생하는 추정치의 문제점을 해결하는 방안으로서 회귀모형을 도입하였다. 회귀모형을 기계산업 표본설계 자료에 적용하여 소지역추정의 가능성을 확인하였으며, 고전적인 추정방법과의 비교도 함께 이루어졌다.

주요용어 : 소지역추정, 오차제곱합회귀모형, 조사회귀모형, 직접추정모형

1. 서론

전국규모의 표본조사 혹은 대규모 표본조사의 설계는 전국의 추정치, 시·도별 추정치 등에 정도를 맞추어 설계되므로 중·소도시의 추정치는 표본설계 당시의 정도를 만족하지 못하며 때로는 추정치가 의미를 부여받을 수 없는 상황에 이를 때도 있다. 즉 전국을 대상으로 한 표본의 크기는 정도를 고려하여 적절한 크기를 갖지만 이를 세부지역에 할당하게 되면 시·군지역의 표본의 크기가 작기 때문에 해당지역의 직접 추정치는 정도가 떨어진다. 그러나 지방자치제도의 정착과 사회의 복잡성등을 감안할때, 보다 세분류된 추정치의 제공이 절실히 요구되고 있다. 이런 요구에 적절한 추정량을 제공하는 추정방법이 소지역 추정이다. 여기에서는 기계산업 부품소재 표본설계에서 대분류에 할당한 표본의 크기를 바탕으로 세분류 도메인의 추정을 소지역추정법을 이용하여 추정하고 비교한다.

2. 부품소재산업 표본설계

부품소재산업표본설계의 주목적은 첫째, 부품소재산업 정책수립 등에 실질적으로 필요한 통계생산 및 기업동향을 파악하기 위한 것이며. 둘째, 부품소재특별법상 기본계획과 시행계획을 수립토록 규정, 실효성 및 현장성 있는 정책수립을 지원하기 위한 것이며, 셋째, 부품소재기업

1) 알투코리아부동산투자자문 데이터팀 대리
2) 덕성여자대학교 자연과학부 통계학과 교수
3) 한신대학교 정보과학대학 정보시스템공학과 교수

에 통계정보제공을 통한 경영전략수립 및 기업경쟁력을 제고하기 위한 것이며, 넷째, 부품소재산업 육성정책 추진실적 대국민 홍보 및 산업정책의 효과를 측정하기 위한 것이다.

부품소재산업의 모집단은 한국표준산업분류상 17, 21, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36 등이며 위와 같은 대분류는 4자리의 중분류로 확장되어 최종 8자리까지 세분류된다. 표본설계방법으로 절사(cut-off)법을 이용하였다. 표본의 크기는 대분류를 기준으로 정도에 맞게 추정하였으며 표본의 크기를 대분류 내의 중분류에 할당하였다. 중분류까지를 고려하였을 때 대다수의 중분류에 할당된 표본이 작았으며, 전혀 표본이 할당되지 않는 경우도 있었다.

본 연구에서는 특별히 표준산업분류, 33의 경우에 대해서 회귀추정을 통해 중분류상의 총계를 추정해보고자 한다. <표 1>은 모집단의 크기가 308인 대분류 33의 총계를 추정하기 위해 계산된 표본의 크기 30을 중분류에 할당한 것을 보여주고 있다. <표 1>은 3321을 제외한 중분류에 할당된 표본의 크기가 작아 전통적인 방법을 이용한 총계추정치에 의미를 부여할 수 없다는 것을 알 수 있으며, 추정된 MSE 값을 관측치에 따라 변동이 매우 클 것이라고 예상할 수 있다.

<표 1> 기계산업표본의 중분류 할당 결과

품목	표본의 크기	모집단의 크기	종사자수	생산액
3311	2	17	33	1421
			92	5759
3319	2	41	35	1549
			106	3434
3321	14	148	31	1963
			34	1011
			35	1126
			40	3358
			44	2355
			54	2194
			58	1802
			65	2003
			89	2882
			102	9223
			146	5638
			472	29773
			196	15943
			226	15941
3322	6	23	30	367
			55	920
			44	582
			80	2579
			80	4284
			290	15130
3332	4	67	43	1367
			71	5217
			100	6748
			272	16354
3340	2	12	31	1175
			53	2020

3. 오차제곱합모형 (Error variance component model)

<표 1>의 자료에서 종사자수를 독립변수로한 회귀모형은 다음과 같이 정의할 수 있다.

(Battese, Harter, Fuller. 1988)

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + a_{ij} \tag{3.1}$$

여기에서 i는 한국표준산업분류에 의한 대분류이고, ($i = 17, 21, 24. \ldots , 35, 36$) j는 대분류 i내의 중분류를 의미한다. 한편 오차 a_{ij}는 다음과 같이 쓰여질 수 있다.

$$a_{ij} = \nu_i + e_{ij}$$

여기에서, ν_i는 대분류 i의 소지역효과를 나타내는 변수이고, e_{ij}는 대분류 i내의 특정 중분류 j의 표본 추출에 의한 오차이고, ν_i와 e_{ij}의 분포는 다음과 같이 정의된다.

$$\nu_i \sim iid\ N(0,\ \sigma_\nu^2)\ ,\quad e_{ij} \sim iid\ N(0,\ \sigma_e^2)$$

공분산은 벡터와 행렬을 이용하여 다음과 같이 표현할 수 있다.

$$\begin{aligned} E &= (\boldsymbol{a}\boldsymbol{a})' \\ &= V \\ &= \text{block dialog}\ (V_1, V_2, \ldots , V_m)' \end{aligned}$$

여기에서, V_i는 다음과 같이 주어진다.

$$V_i = J_i\, \sigma_\nu^2 + I_i\, \sigma_e^2 \tag{3.2}$$

<표 1>에서 대분류별 생산액과 종사자수의 표본평균은 다음과 같이 추정할 수 있다.

$$\bar{y}_{i.} = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}\ ,\quad \bar{x}_{i.} = n_i^{-1} \sum_{j=1}^{n_i} x_{ij}$$

위의 추정치를 이용하여 모형 (3.1)을 다음과 같이 정리 할 수 있다.

$$\bar{y}_{i.} = \beta_0 + \beta_1\, \bar{x}_{i.} + \nu_i + \bar{e}_{i.} \tag{3.3}$$

따라서, 대분류별 평균생산액은 다음과 같이 표현 할 수 있다.

$$Y_i = \overline{X}_i \beta + \nu_i \tag{3.4}$$

여기에서, $\overline{X}_i$는 다음과 같이 나타낼 수 있다.

$$\overline{X}_i = N_i^{-1} \sum_{j=1}^{N_i} x_{ij} = (1, X_{ij})$$

이제, 식 (3.2)의 추정을 생각할 때, 우선 확률변수 ν_i와 $\bar{a}_{i.}$는 다음과 같은 공분산을 갖는 이변량 정규분포를 따른다고 가정하자.

$$\begin{pmatrix} \sigma_\nu^2 & \sigma_\nu^2 \\ \sigma_\nu^2 & \sigma_\nu^2 + n_i^{-1}\, \sigma_e^2 \end{pmatrix}$$

$\bar{a}_{i.}$ 이 주어졌을 때, ν_i의 조건부 기대값은 다음과 같다.

$$E(\nu_i \mid \bar{a}_{i.}) = \bar{a}_{i.} g_i \tag{3.5}$$

이때, g_i는 다음과 같다.

$$g_i = m_i^{-1} \sigma_\nu^2$$

이제 식 (3.5)를 이용하면, MSE를 다음과 같이 추정할 수 있다.

$$\begin{aligned} E[(\nu_i - \bar{a}_{i.} g_{i.})^2] &= \sigma_\nu^2 (1 - g_i) \\ &= n_i^{-1} \sigma_e^2 - n_i^{-2} \sigma_e^2 m_i^{-1} \sigma_e^2 \end{aligned}$$

만약, σ_ν^2 과 σ_e^2 을 알고 있다면 식 (3.2)를 이용하여 β를 다음과 같이 추정한다.

$$\hat{\beta}_{GLS} = (X' V^{-1} X)^{-1} X' V^{-1} Y \tag{3.6}$$

대분류 효과를 나타내는 변수 ν_i는 다음과 같이 추정할 수 있다.

$$\hat{\nu}_i = n_i^{-1} \hat{u}_{i.} g_i \tag{3.7}$$

이때, $\hat{a}_{i.}$는 다음과 같고,

$$\hat{a}_{i.} = n_i^{-1} \sum_{j=1}^{n_i} \hat{a}_{ij}$$

$\hat{a}_{ij}$ 는 다음과 같다.

$$\hat{a}_{ij} = y_{ij} - x_{ij} \hat{\beta}$$

대분류별 생산액 평균은 다음과 같이 추정된다.

$$\hat{y}_i = \bar{x}_{i(b)} \hat{\beta} + \hat{\nu}_{i.} \tag{3.8}$$

그리고, 표본오차 $\hat{\sigma}_e^2$은 다음과 같이 표현 할 수 있다.(Kackar, Hanvile. 1986)

$$\hat{\sigma}_e^2 = \hat{e}' \hat{e} \left[\sum_{i=1}^{T} (n_i - 1) - 2 \right]^{-1} \tag{3.11}$$

여기에서, 우리는 $\hat{e}' \hat{e}$은 y에 대한 회귀분석의 잔차 제곱합에 의해서 추정을 할 수 있고, $\hat{\sigma}_e^2$ 은 σ_e^2의 불편 추정치이다. 그렇다면, 대분류별 도메인에 영향을 주는 보조변수의 분산을 추정할 때, i번째 대분류 도메인의 최소제곱오차(least-squares residual)을 이용해 보기로 하겠다.

$$\hat{a}_{i.} = \bar{y}_{i.} - \bar{x}_{i.} (X' X)^{-1} X' Y$$

그리고, 기대값은 다음과 같이 나타낼 수 있다.

$$E[\ \hat{a}_{i.}^{\ 2}] = b_i\,\sigma_\nu^2 + d_i\,\sigma_e^2$$

여기에서, b_i는 다음과 같고,

$$b_i = 1 - 2n_i\,\bar{x}_{i.}\,(X'X)^{-1}\,\bar{x}_{i.}' + \bar{x}_{i.}\,(X'X)^{-1}\Big(\sum_{j=1}^{T} n_j^2\,\bar{x}_{j.}'\,\bar{x}_{j.}\Big) \times (X'X)^{-1}\,\bar{x}_{i.}'$$

d_i는 다음과 같다.

$$d_i = n_i^{-1}[1 - n_i\,\bar{x}_{i.}\,(X'X)^{-1}\,\bar{x}_{i.}']$$

그리고, 평균잔차(average of residual)의 가중제곱합(weighted sum of squares)은 다음과 같이 표현 될 수 있다.

$$m_{..} = \Big(\sum_{i=1}^{T} n_i b_i\Big)^{-1}\Big(\sum_{i=1}^{T} n_i\,\hat{u}_i^2\Big)$$

그리고, 기대값은 다음과 같다.

$$E[m_{..}] = m_{..} = \sigma_\nu^2 + c\,\sigma_e^2$$

이때, c는 다음과 같고,

$$c = \Big(\sum_{i=1}^{T} n_i b_i\Big)^{-1}\Big(\sum_{i=1}^{T} n_i d_i\Big)$$

$\hat{m}_{..}$은 $\hat{\sigma}_e^{\ 2}$과 독립일 때, σ_ν^2의 기대값은 다음과 같다.(Battese, Harter, Fuller. 1988, Kackar, Hanvile 1986)

$$\hat{\sigma}_\nu^{\ 2} = \max\{\ \hat{m}_{..} - c\ \hat{\sigma}_e^{\ 2},\ 0\} \tag{3.12}$$

그리고, g_i의 추정치는 다음과 같다.

$$\hat{g}_i = (\ \hat{\sigma}_\nu^{\ 2} + n_i^{-1}\ \hat{\sigma}_e^{\ 2})^{-1}\ \hat{\sigma}_\nu^{\ 2}$$

이를 이용하여서 평균생산액의 추정치는 다음과 같다.

$$\hat{y}_i = \bar{x}_{i(\beta)}\hat{\beta} + \hat{u}_{i.}\ \hat{g}_{i.}$$

β의 추정은 일반화선형제곱으로 추정이 가능하고, $\hat{V}$는 $\hat{\sigma}_\nu^{\ 2}$와 $\hat{\sigma}_e^{\ 2}$을 공분산 행렬식에 대입해서 추정을 할 수 있다. 지금까지의 모형과 모수 추정을 위해 종사자수를 이용하여서 생산액을 알고자 했다. 그래서, 위의 식으로 추정된 총 생산액은 1,671,955 이다. 식 (3.7)과 식 (3.11)을 이용한 분산의 추정치는 다음과 같다.

$$\hat{\sigma}_e^{\ 2} = 2867180\ ,\quad \hat{\sigma}_\nu^{\ 2} = 100732.73\ .$$

위의 분산 추정치와 식 (3.6)을 이용하여 β를 추정한 회귀모형은 다음과 같다

$$\hat{y}_{ij} = -1111.8 + 65.23x_{ij}$$

4. 조사회귀모형(Survey regression model)

표본 조사시 회귀 추정치는 다음과 같은 모형을 가진다.

$$y = \bar{x}_i \hat{\beta} + (\bar{y}_{i.} - \bar{x}_{i.} \hat{\beta})$$

이때, $\hat{y}_{i.}$는 다음과 같다.

$$\hat{y}_{i.} = \mu_{xi} \hat{\beta}_{OLS} + (\bar{y}_{i.} - \bar{x}_{i.} \hat{\beta}_{OLS})$$

일 때, $\beta_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$이고, MSE는 다음과 같다.

$$\begin{aligned} V(\hat{y}_{i.} - y_{i.}) &= V[\bar{e}_{i.} + (\mu_{xi} - \bar{x}_{i.})(\hat{\beta}_{OLS} - \beta)] \\ &= n_i^{-1}\sigma_e^2 + (\mu_{xi} - \bar{x}_{i.}) V(\hat{\beta}_{OLS} - \beta) \times (\mu_{xi} - \bar{x}_{i.})' \end{aligned}$$

5. 결과 및 결론

3장과 4장의 추정치의 결과는 다음과 같이 정리 할 수 있다.

<표 2> MSE의 비교

층의 총합	표본의 크기	오차제곱합 모형의 총계 추정	MSE의 추정		
			오차제곱합 회귀 모형	조사회귀 모형	직접 추정 모형
$Y_{1.}$	2	30,255	113,194	553,157	9,409,122
$Y_{2.}$	2	135,908	192,379	481,307	1,776,613
$Y_{3.}$	14	946,227	100,119	477,867	50,750,544
$Y_{4.}$	6	114,418	118,811	477,969	2,403,017
$Y_{5.}$	4	463,729	151,984	478,523	2,838,152
$Y_{6.}$	2	19,478	102,963	495,732	357,013

<표 2>는 β의 추정방법에 따른 차이를 보여준다. 오차제곱합회귀모형은 β를 GLS의 방법으로 추정하여 각각의 분산을 추정하고, 조사회귀모형은 β를 OLS의 방법으로 추정한 방법으로 <표 2>의 결과를 살펴보면 직접추정의 경우 MSE가 가장 크고, 다음으로 조사회귀모형에 의한 추정치의 MSE가 크고, 오차제곱합회귀모형에 의한 MSE가 가장 작다는 것을 볼 수 있다. 이는 각각의 오차의 분산을 추정한 차이일수도 있으나, 가장 큰 차이는 오차제곱합회귀모형에서 보조변수 ν_i의 상관관계가 클수록 MSE가 더욱 감소하게 된다. 그리고, 당연한 결과지만 표본의 크기가 증가함에 따라 각 추정치에 의한 MSE 역시 감소함을 볼 수 있다. 지금까지, 소지역추정에서 모수의 정보에 의한 총계와 MSE의 변화를 살펴보았다. 이는 소지역추정에 있어서 소지역의 보조정보가 중요함을 의미하며 각각의 추정방법에 따른 MSE의 비교로 인해 소지역추정의 효율성을 보였다. 그러나, 소지역추정은 추정변수와 상관관계가 큰
보조변수를 찾아야하는 제약조건이 있음에도 불구하고 지역 혹은 도메인 내의 보조정보에 의해 직접추

정치보다 안정된 추정치를 제공할 수 있으므로 소지역 혹은 소도메인에 널리 이용될 수 있다.

참 고 문 헌

[1] George E. Battese, Rachel M. Harter, Wayne A. Fuller, (1988), An error components model for prediction of county crop area using survey and satellite data, *Journal of the American Statistical Association* , 83 , 28-36

[2] Kackar, R. N, Hanvile, D. A, (1986), Approximations for standard error of estimation of fixed and random effects in mixed linear models, *Journal of the American Statistical Association* , 79 , 853-862

[3] Samdal, C. E, (1984), Design consistent versus model dependent estimation for small domain, *Journal of the American Statistical Association* , 79, 624-631