Journal of the Korean Data & Information Science Society
2009, **20**(1), 77–87

한국데이터정보과학회지

고객집단별 보험금에 대한 소지역 추정[†]

김영화[1] · 김기수[2]

[1]중앙대학교 자연과학대학 수학통계학부 · [2]중앙대학교 대학원 통계학과

접수 2008년 12월 29일, 수정 2009년 1월 10일, 게재확정 2009년 1월 15일

요 약

최근 들어 소지역 추정 문제를 해결하는데 베이지안 방법이 주목을 받고 있다. 본 논문에서는 고객집단별 보험금에 대한 실제 자료를 MCMC 기법을 통한 계층적 베이지안 모형과 일원분류, GLM-Normal, GLM-Gamma 모형으로 분석하여 그 결과를 비교하였다. 결론적으로 소지역 추정에 의하여 얻어진 보험금 추정량이 다른 방법으로부터 얻어진 추정량들과 비교하여 가장 합리적이고 좋은 추정량임을 보일 수 있었다. 특히, 표본 수가 적은 집단에 대하여 소지역 추정의 정확성이 현저하게 높음을 알 수 있었다.

주요용어: 계층적 베이즈, 보험금, 소지역 추정, 일반선형모형.

1. 서 론

소지역 (small area)이란 일반적으로 지리적, 사회적 또는 그 외의 다른 부집단으로 정의하며, 소지역의 전형적인 특징은 표본의 크기가 작다는 것이다. 표본조사에서 다양한 세부 기준에 따라 소지역을 정의하면 특정한 소지역에서는 이에 해당하는 케이스의 수가 매우 적은 경우가 발생하게 된다. 소지역에서 표본의 크기가 작은 이유는 대부분의 조사가 일반적으로 큰 모집단에 대한 통계를 작성하기 위해 설계되기 때문이다. 따라서 케이스의 수가 많은 지역에 대한 합이나 평균에 대해서는 신뢰할 수 있는 추정치를 구할 수 있으나, 케이스의 수가 적은 소지역에 해당되는 표본 데이터만 가지고 일반적으로 사용되는 직접 조사 추정량을 계산하면 소지역의 표본크기가 작기 때문에 표준오차나 변동계수가 커지게 되어 추정치가 신뢰성을 잃게 된다. 또한 표본이 전혀 없는 경우에는 직접 추정에 의한 추정 자체가 불가능해진다. 이를 해결하기 위한 방법으로 소지역추정 (small area estimation)이 사용되고 있으며, 소지역추정량을 정의하는 데 일반적으로 보조정보 (auxiliary information)를 이용한다. 여기서 보조정보란 모집단의 특성과 유사한 인근 소지역에 관한 각종 통계정보를 의미한다. 이와 같이 보조정보를 이용하는 추정 및 추측방법을 간접 또는 모형근거방법 (indirect or model-based methods)이라 한다. 보조정보를 이용하는 추정방법은 흔히 'borrowing strength'로 특징지어지는데, 이때 'borrowing strength'는 해당 소지역의 반응변수와 보조정보 간의 관계를 나타낸다.

이러한 모형근거방법에 관한 연구는 오랜 역사를 지니고 있지만, 이 방법이 소지역 추정에 직접 이용된 것은 불과 20-30년 전부터라고 할 수 있다. 이 모형기반 방법은 반응변수의 소지역 간 변동성을 단지

[†] 이 논문은 2007년도 중앙대학교 연구장학기금 지원에 의한 것임.
[1] 교신저자: (156-756) 서울특별시 동작구 흑석동 221, 중앙대학교 자연과학대학 수학통계학부, 부교수.
E-mail : gogators@cau.ac.kr
[2] (156-756) 서울특별시 동작구 흑석동 221, 중앙대학교 대학원 통계학과, 석사과정.

보조정보의 변동성에 의해서 설명하는 고정효과 모형 (fixed effect model)과 보조정보에 의하여 설명되지 않는 소지역 특유의 변동성까지를 포함시키는 혼합모형 (mixed model)으로 구분된다.

고정효과 모형에서는 반응변수에 있어서 소지역 간 변동을 단지 이미 알려져 있는 요인들에 의해서만 설명한다. 고정효과 모형은 소지역 추정 분야의 주류를 이루어 왔으며, 최근 경험적 베이즈 (Empirical Bayes) 및 계층적 베이즈 (Hierarchical Bayes) 추정방법 (Ghosh와 Rao, 1994)들도 소지역 추정에 이용되고 있다. HB 접근방법에서는 미지의 모수들이 사전분포로부터 추출되는 확률변수로 취급된다. 이 방법들의 공통된 주안점은 동시에 여러 소지역에 대한 보다 정밀한 추정치를 얻기 위해 연관된 지역들로부터 정보를 빌려오는 것이다.

최근 소지역 통계에 대한 수요는 공공부문과 민간부문에서 신뢰할 수 있는 소지역통계에 대한 요구와 필요에 의해 크게 증가하고 있는 추세이며, 베이지안 방법이 모형을 통해 소지역들을 체계적으로 연결하는데 적합하여 보다 신뢰성 있는 추정치를 얻을 수 있기 때문에 최근 들어 소지역 추정 문제에 베이지안 방법의 적용이 매우 활발하다. 특히 MCMC 계산의 발전으로 비록 모형이 복잡하더라도 베이지안 추정치와 이에 관련된 표준오차를 쉽게 계산할 수 있게 되어 소지역추정 문제에서 베이지안 분석이 유용한 것으로 평가를 받고 있다.

본 연구에서는 소지역 추정법의 적용 가능성과 효율성을 국내 화재보험사의 실제 자료 분석을 통하여 제시한다. 특히 본 논문은 처음으로 국내 보험사의 실제 자료를 분석 대상으로 한다는 것에 큰 의미를 부여할 수 있다고 사료된다. 국내의 경우, 성나영과 김영원 (2000)이 도소매업 사업체 조사를 통하여 소지역 추정의 적용 가능성을 제시한 이후, 박종태와 이상은 (2001)은 경기도 실업자 총계 추정문제에 소지역 추정 방법을 모의실험을 통해 다루었으며, Kim과 Jo (2004)는 소지역 추정기법으로 시군구 실업률을 산출하는 방법을 제시하였다. 특히 Kim (2007)은 소지역 추정을 보험 자료 분석에 적용하였으나 이는 미국 보건통계국의 자료를 사용한 것이다.

2. 실제 자료 분석

현재 국내 화재보험사에서는 다양한 고객의 정보를 이용하여 고객이 받는 보험금을 기준으로 고객이 납부하는 보험료를 책정한다. 따라서 고객 정보의 정확한 분석을 통한 고객 집단별 보험금의 정교한 추정은 보험사에 있어 가장 중요한 일이라 할 수 있다. 이 장에서는 실제 보험회사 자료에 근거하여 여러 가지 보험금 추정방법을 소개하고, 이들 방법과 소지역 추정 (small area estimation)방법에 의한 보험금 추정의 정확성을 비교하고자 한다.

2.1. 고객 데이터의 구성

분석에 사용된 자료는 국내 화재보험사인 S사의 실제 자료이다. 설명 변수로서 고객의 연령대, 성별, 운전경력, 차종, 운전자 한정에 따라 고객 집단을 나누며 사고 시 발생되는 손해액, 즉 보험사에서 지급하는 보험금을 집단별로 추정하는 것이 이 분석의 목적이다. 설명변수는 표 2.1과 같으며, 종속 변수는 보험금 (mul_los2)이다.

설명변수인 연령대, 운전자 한정, 성별, 운전경력, 차종에 따라 전체 고객은 192개의 집단으로 분류되며 ($3 \times 4 \times 2 \times 2 \times 4 = 192$), 전체 데이터를 모집단으로 가정한 후, 각 집단의 보험금 통계량을 구하여 각 방법의 추정치들을 각각 비교한다. 분석에서는 모수 추정에 역점을 두며, 단순한 유의성 검정보다 더 다양한 정보를 제공한다. 192개의 그룹 가운데 보험금 발생이 0 인 6개 그룹에 대해서는 지수 분포로 난수를 생성하여 그룹 평균이 0 인 것을 제거하였다.

표 2.1 설명 변수

변수명	범주수	범주
age (연령대)	3	① 50대 이상 ② 40대 ③ 30대 이하
drv (운전자 한정)	4	① 1인 ② 부부 ③ 가족 ④ 기본
sex (성별)	2	① 여 ② 남
exp (운전경력)	2	① 3년 이상 ② 2년 이하
car (차종)	4	① 다인승 ② 대형 ③ 중형 ④ 소형

2.2. One-way를 이용한 보험금 추정

이 방법은 각 요인 수준별 평균을 비교하여 상대 위험도를 산출하는 단순한 방법으로서, 일원배치분산분석을 사용하며 모형의 적합도보다 각 수준별 평균에 관심을 둔다. 각 요인 수준별 평균은 다음의 표로 나타난다.

표 2.2 요인 수준별 평균과 표준편차

		도수	평균	표준편차
age	1	25745	56179.01377	405015.9712
	2	36637	48050.7386	367632.6897
	3	37618	38992.14215	278510.526
drv	1	17615	43024.69082	314438.144
	2	39172	41446.02781	306060.5429
	3	27617	57779.62994	416916.5943
	4	15596	44656.72306	347318.0521
sex	1	18250	48303.11797	373857.1679
	2	81750	46385.7874	341724.2799
exp	1	80324	46469.87575	351443.8029
	2	19676	47820.88447	332564.519
car	1	20525	55634.44454	389095.5924
	2	15030	42814.49739	336841.7129
	3	25857	48833.19074	381436.7391
	4	38588	42124.27024	302023.3982
mul_lose2		100000	46735.70023	347808.8932

표 2.2의 조합별 평균을 이용하여 연령대별 상대 위험도는 다음과 같이 산출한다.

$$\text{age1\_rate} = \frac{\text{E}(\text{X}_{\text{age1}})}{\text{E}(\text{X}_{\text{age3}})}, \quad \text{age2\_rate} = \frac{\text{E}(\text{X}_{\text{age2}})}{\text{E}(\text{X}_{\text{age3}})}, \quad \text{age3\_rate} = \frac{\text{E}(\text{X}_{\text{age3}})}{\text{E}(\text{X}_{\text{age3}})} = 1$$

표 2.3은 위와 같은 방법으로 나머지 요인들의 수준별 상대 위험도를 위의 식처럼 계산하다. 다음은 각 요인의 수준별 상대 위험도를 나타낸 표이다.

표 2.3 요인 수준별 상대 위험도

age	rate	drv	rate	sex	rate	exp	rate	car	rate
1	1.4407778	1	0.9634538	1	1.0413344	1	0.9717485	1	1.3207218
2	1.2323185	2	0.9281027	2	1.0	2	1.0	2	1.0163854
3	1.0	3	1.2938618					3	1.1592649
		4	1.0					4	1.0

표 2.3에 따르면, 연령대가 높아질수록 위험도가 높으며 남자보다 여자가 위험도가 높음을 알 수 있다. 또한, 운전 경력이 낮을수록 위험도가 높다. 즉, 일반적으로 생각할 수 있는 수준별 위험도의 높고 낮음이 나타나고 있음을 알 수 있다.

요인 수준의 조합으로 분류된 192개의 고객 집단의 위험도는 각 요인 수준별 위험도의 값을 곱하여 구한다. 즉, 192개의 고객 집단의 상대 위험도 및 추정치를 산출하기 위해 각 수준 조합에 대하여 전체 평균을 곱하여 추정치를 산출한다. 예를 들어, 192개의 고객 집단 중 40대 (연령대②), 부부운전자 한정 (운전자한정②), 여자 (성별①), 경력 2년이하 (운전자경력②), 중형 (차종③)인 87번째 집단의 상대 위험도는 각 요인 수준의 위험도를 곱하여 구한다.

- 위험도$(\text{age}=2, \text{drv}=2, \text{sex}=1, \text{exp}=2, \text{car}=3) = \text{RiskRate}_{87}$
$= 1.2323185 \times 0.9281027 \times 1.04133444 \times 1.0 \times 1.1592649$
$= 1.3806766$
- 추정치$(\text{age}=2, \text{drv}=2, \text{sex}=1, \text{exp}=2, \text{car}=3) = \theta_{87}$
$= RiskRate_{87} \times Overall\_mean = 1.3806766 \times 46735.70023$
$= 64526.88769$

이렇게 계산된 값을 사용하여 모집단의 전체 합과 위험도를 고려한 추정치의 전체 합의 비율을 맞추고, 수정된 추정치의 값을 다음과 같이 구한다.

$$\theta_{adjusted87} = \theta_{87} \times \sum_{i=1}^{192} n_i\mu_i / \sum_{i=1}^{192} n_i\theta_i = 44198.14164 (i = 1, \cdots, 192)$$

여기서 n_i , μ_i , θ_i는 각각 i번째 그룹의 관측 도수, 모평균, 추정치이다.

각 분석의 추정치와 소지역 추정의 추정치를 비교하기 위해서는 비교의 기준이 필요하다. 이를 위하여, 추정치 $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_{192})^T$와 모집단의 평균 $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_{192})^T$에 대해서 다음을 정의한다.

- 평균 절대편의 (average relative bias) $= \dfrac{1}{192}\sum_{i=1}^{192} |\mu_i - \theta_i|$
- 과대과소 지급비율 $= \sum_{i=1}^{192} \left(\dfrac{\theta_i}{\mu_i} - 1 \right)$

과대과소 지급비율은 각 고객 집단에서 추정한 보험금이 모집단의 보험금에 비해 얼마나 더 많게 또는 적게 지급되었는지를 합산하여 전체적인 손망실을 평가하는 척도라 할 수 있다. 비교의 지표로서 평균 절대편의와 함께 과대과소 지급비율을 사용하여, One-way에 의한 추정, 일반화 선형모형 (Normal, Gamma)에 의한 추정, 소지역 추정의 결과를 평가한다. 표 2.4는 One-way 추정 결과에 대한 평가이다.

표 2.4 One-way 추정 결과 평가

One-way	
평균 절대편의	과대과소 지급비율
20708.45295	40.50%

2.3. 일반화 선형 모형을 이용한 보험금 추정

일반화 선형 모형의 공통된 세 가지 요소는 랜덤성분, 체계적 성분, 연결함수 (link function)이다. 확률변수 Y의 기댓값을 $\mu = E(Y)$로 나타낸다면, GLM에서 Y의 기댓값이 설명변수의 수준에 따라 다양하게 나타난다. 만약 age= i, drv= j, sex= k, exp= l, car= m이라면, 이 경우의 체계적 성분은 다음과 같은 선형식으로 표현된다.

$$\alpha + \beta_i + \gamma_j + \delta_k + \zeta_i + \eta_m$$

연결함수는 랜덤 성분과 체계적 성분을 연결하는 역할을 하며, 이것은 선형 예측식에 있는 설명변수와 $\mu = E(Y)$가 어떻게 관련되어 있는지를 설명한다. 연결함수를 $g(\cdot)$이라 하면, age= i, drv= j, sex= k, exp= l, car= m인 경우의 연결함수와 종속변수의 관측값 y_{ijklm}는 다음과 같이 표현되며

$$g(\mu) = \alpha + \beta_i + \gamma_j + \delta_k + \zeta_i + \eta_m$$
$$y_{ijklm} = g^{-1}(\alpha + \beta_i + \gamma_j + \delta_k + \zeta_i + \eta_m) + error$$

연결함수를 $g(X) = \ln(X)$라 하면 다음 식을 얻게 된다.

$$y_{ijklm} = \exp(\alpha)\exp(\beta_i)\exp(\gamma_j)\exp(\delta_k)\exp(\zeta_i)\exp(\eta_m) + error$$
$$E(y_{ijklm}) = \exp(\alpha)\exp(\beta_i)\exp(\gamma_j)\exp(\delta_k)\exp(\zeta_i)\exp(\eta_m)$$

연결함수를 로그함수로 하여 고객 집단별 추정치를 One-way 방법에서와 같이 위험도를 곱하여 나타낼 수 있다. 고객 집단별 보험금 자료의 종속변수인 보험금은 연속형자료이므로 Normal 랜덤 성분을 가정하여 분석을 실시한다. 표 2.5는 GLM-Normal 인 경우의 요인 수준별 상대 위험도이며, 표2.6은 분석 모형 및 추정 결과이다.

표 2.5 요인 수준별 상대 위험도 (GLM-Normal)

age	rate	drv	rate	sex	rate	exp	rate	car	rate
1	1.2986353	1	0.9520666	1	1.0397735	1	0.9628683	1	1.3383272
2	1.2281440	2	0.9447977	2	1.0	2	1.0	2	0.9943285
3	1.0	3	1.2195871					3	1.1468629
		4	1.0					4	1.0

표 2.5의 요인 수준별 위험도를 이용하여 각 수준의 곱으로서 다음과 같이 고객 집단별 추정치를 산출한다.

$$E(y_{ijklm}) = \exp(\alpha)\exp(\beta_i)\exp(\gamma_j)\exp(\delta_k)\exp(\zeta_i)\exp(\eta_m)$$

모집단의 보험금과 GLM-Normal 랜덤 가정과 비교하여 평균 절대편의와 과대과소 지급비율은 다음과 같다.

표 2.7에 따르면, One-way 방법과 비교했을 때, 평균 절대편의는 조금 감소하였으나 과대과소 지급비율은 오히려 더 증가하였음을 알 수 있다. 즉, 보험금이 0에서 ∞의 값을 가지므로 Normal 을 가정하는 것이 타당하지 않으며 Normal 이외에 조건을 만족하는 다른 분포의 랜덤 가정이 필요하다는 것이므로, 0에서 ∞의 값을 갖는 손해액의 경우 Gamma 랜덤 가정을 하여 분석을 실시하였다. 표 2.8은 Gamma 랜덤 가정에 대한 추정결과이다.

Gamma 랜덤 가정시 일반화 선형 모형의 각 요인 수준별 상대 위험도는 다음과 같다.

표 2.6 분석 모형 및 추정 결과

구분	비고	
종속변수	보험금	
분포가정	Normal	
연결함수	LOG	
Parameter	DF	Estimate	exp (estimate)
Intercept	1	10.4918	36018.93
age1	1	0.2613	1.298617
age2	1	0.2055	1.228139
age3	0	0	1
drv1	1	-0.0491	0.952086
drv2	1	-0.0568	0.944783
drv3	1	0.1985	1.219572
drv4	0	0	1
sex1	1	0.039	1.03977
sex2	0	0	1
exp1	1	-0.0378	0.962906
exp2	0	0	1
car1	1	0.2914	1.3383
car2	1	-0.0057	0.994316
car3	1	0.137	1.146828
car4	0	0	1

표 2.7 GLM-Normal 평가

GLM-Normal	
평균 절대편의	과대과소 지급비율
20524.68228	44.08%

표 2.9의 요인 수준별 상대위험도를 이용하여 GLM-Normal 랜덤가정과 동일한 방법으로 각 수준의 곱으로서 다음과 같이 고객 집단별 추정치를 산출한다. 모집단의 보험금과 GLM-Gamma 랜덤 가정과 비교하여 평균 절대편의와 과대과소 지급비율은 다음과 같다.

Gamma 랜덤가정의 경우, Normal 랜덤 가정의 경우에 비하여 평균 절대편의에 대한 개선은 없었지만 과대과소 지급비율이 낮아졌다. 우도비를 최대로 해주면서 반복적으로 계수의 추정치를 얻는 반복의 수가 Gamma 랜덤 가정 모형이 적고 Normal 랜덤 가정, Gamma 랜덤 가정의 두 모형의 적합도면에서도 Gamma 랜덤 가정의 모형이 더 적합하였다.

2.4. 소지역 추정

선형 혼합 모형에서 연결모형과 표본모형을 통하여 모수들이 서로 연관되어 있으므로 이를 계층적 베이지안을 통해 $f(\theta|y)$를 구할 수 있다. 그러나 다차원에서의 적분이 어렵기 때문에 깁스 샘플러를 이용하여 $f(\theta|y)$를 구한다. 깁스 샘플러는 마르코프 연쇄의 원리를 이용하는 몬테카를로 적분 기법이다. 이 기법은 S.Geman과 D.Geman (1984)에 의해 제안되었으며 Gelfand와 Smith (1990)에 의하여 베이지안 문제에 적용되었다. 실제 소지역 추정의 효과를 알아보기 위하여 One-way, GLM에서 사용된 동일

표 2.8 분석 모형 및 추정 결과

구분	비고
종속변수	손해액
분포가정	Gamma
연결함수	LOG

Parameter	DF	Estimate	exp (estimate)
Intercept	1	13.4253	676914.3
age1	1	0.0299	1.030351
age2	1	0.1114	1.117842
age3	0	0	1
drv1	1	-0.0936	0.910647
drv2	1	-0.1013	0.903662
drv3	1	0.0563	1.057915
drv4	0	0	1
sex1	1	-0.076	0.926816
sex2	0	0	1
exp1	1	0.0917	1.096036
exp2	0	0	1
car1	1	0.1688	1.183883
car2	1	0.1824	1.200094
car3	1	0.1679	1.182818
car4	0	0	1

표 2.9 요인 수준별 상대 위험도 (GLM-Gamma)

age	rate	drv	rate	sex	rate	exp	rate	car	rate
1	1.0303900	1	0.9106248	1	0.9268208	1	1.0959818	1	1.1838281
2	1.1178369	2	0.9036222	2	1.0	2	1.0	2	1.2000533
3	1.0	3	1.0579475					3	1.1828287
		4	1.0					4	1.0

표 2.10 GLM-Gamma 평가

GLM-Gamma	
평균 절대편의	과대과소 지급비율
20982.4286	37.36%

표 2.11 Gamma, Normal 랜덤 가정 모형의 적합도 비교

Gamma 랜덤 가정				Normal 랜덤 가정			
Criteria For Assessing Goodness Of Fit				Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF	Criterion	DF	Value	Value/DF
Deviance	5742	5593.89	0.9742	Deviance	1.00E+05	1.21E+16	1.21E+11
Chi-Square	5742	12295.40	2.1413	Chi-Square	1.00E+05	1.21E+16	1.21E+11
Iteration	5			Iteration	14		

한 데이터를 사용하고 변수들에 대한 설명은 다음과 같다.

$$y_i = \text{고객 집단별 모집단 평균}, \quad V_i = \text{고객 집단별 모집단 분산}$$

$$x_{1i} = age, x_{2i} = drv, x_{3i} = sex, x_{4i} = exp, x_{4i} = car$$

$$\boldsymbol{x}_i = (1, x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i})^T, \quad \boldsymbol{b} = (b_0, b_1, b_2, b_3, b_4, b_5)^T$$

고객 집단별 보험금 추정을 위한 소지역 모형은 다음과 같다.

$$y_i = \boldsymbol{x}_i^T \boldsymbol{b} + \mu_i + e_i, i = 1, \cdots, m$$

여기서 e_i 는 서로 독립이며 평균이 0, 알려진 분산이 V_i 인 정규 분포를 따른다고 가정하고 μ_i는 서로 독립이며 평균이 0, 분산이 τ^2인 정규분포를 따른다고 가정한다. 계층적 베이지안 형태는 다음과 같이 표현되며

$$\begin{aligned}
&y_i|\boldsymbol{\theta}, \boldsymbol{b}, \tau^2 \sim indep.N(\theta_i, V_i), i = 1, \cdots, m \\
&\theta_i|\boldsymbol{b}, \tau^2 \sim indep.N(\boldsymbol{x}_i^T \boldsymbol{b}, \tau^2), i = 1, \cdots, m \\
&\boldsymbol{b}\text{와 } \tau^2\text{은 서로 독립이고 } \boldsymbol{b} \sim Uniform(R^5), \pi(\tau^2) \propto 1
\end{aligned}$$

깁스 샘플러를 생성하기 위해 필요한 조건부 확률분포를 구하면 다음과 같다.

$$\begin{aligned}
&\boldsymbol{b}|\boldsymbol{\theta}, \tau^2, \boldsymbol{y} \sim N_P\left((X^T X)^{-1} X^T \boldsymbol{\theta}, \tau^2 (X^T X)^{-1}\right) \\
&\tau^2|\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{y} \sim IG\left(\frac{m-2}{2}, \frac{1}{2}\sum_{i=1}^{m}(\theta_i - x_i^T b)^2\right) \\
&\theta_i|\boldsymbol{b}, \tau^2, \boldsymbol{y} \sim indep.N\left((V_i^{-1} + \tau^{-2})^{-1}(V_i^{-1} y_i + \tau^{-2} \boldsymbol{x}_i^T \boldsymbol{b}), (V_i^{-1} + \tau^{-2})^{-1}\right)
\end{aligned}$$

여기서 $\boldsymbol{y} = (y_1, \cdots, y_m)^T$, $\boldsymbol{X}^T = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m)$, $rank(\boldsymbol{X}) = p + 1$이다.

θ의 초기값으로 정규분포의 난수를 192개 생성하였으며, $(\tau^2)^{-1}$의 분포가 역감마분포이므로 τ^2의 초기값을 Gamma (0.1, 0.1)으로 하였다. 분석 결과, 모집단의 보험금과 소지역 추정의 사후평균인 $E(\theta_i|y)$와 비교하여 다음과 같은 평균 절대편의와 과대과소 지급비율을 얻었다.

표 2.12 소지역 추정 결과 평가
Small Area Estimate

평균 절대편의	과대과소 지급비율
18594.90092	11.75%

표 2.12로부터 One-way, GLM의 경우와 비교하여 소지역 추정의 평균 절대편의가 작으며 과대과소 지급비율도 현저하게 개선되었음을 알 수 있다.

표본수가 적은 고객 집단은 표준오차나 변동계수가 커지게 되어 추정치가 신뢰성을 잃게 된다. 다음의 표2.13은 표본수가 100개 이하인 고객 집단의 각 추정 방법에 대한 과대과소 지급비율이다. 표2.13의 6, 24, 51, 65, 66, 71, 101, 110, 118, 129, 182, 183 집단 (음영부분)과 같이 One-way, GLM (Normal, Gamma)에서 특히 과대과소 지급비율이 높은 집단에서 소지역 추정의 과대 과소 지급비율이 현저히 떨어지는 것을 볼 수 있다.

3. 결론

본 논문에서는 소지역 추정 (small area estimation)을 이용한 보험금 추정과 그 효과를 확인하였다. 선형 모형과 계층적 모형을 고려하여 관측값, 즉 과거 자료 보험금이 주어졌을 때, 고객 집단별 보험금이 얼마나 산출되는지를 깁스 샘플러를 이용하여 추정하였다. 또한, One-way, GLM (Normal, Gamma), 소지역 추정의 추정치를 평균 절대편의와 과대 과소 지급비율을 이용하여 각 모형들을 비교

표 2.13 표본 수 100개 이하 집단의 과대과소 지급비율

모집단			One-way	Normal	Gamma	소지역 추정	요인 수준				
집단	N	평균	error_rate	error_rate	error_rate	error_rate	age	drv	sex	exp	car
5	6	32425.22	88.48%	91.15%	23.78%	-37.29%	1	1	1	2	1
6	7	23497.3	100.16%	95.98%	73.15%	26.45%	1	1	1	2	2
7	15	38474.19	39.43%	38.05%	4.23%	16.88%	1	1	1	2	3
70	19	30708.42	31.00%	41.82%	43.73%	46.12%	2	1	1	2	2
134	19	43560.18	-25.06%	-18.60%	-9.35%	-1.82%	3	1	1	2	2
53	20	44132.64	43.73%	47.51%	-0.13%	-11.19%	1	4	1	2	1
21	23	39773.91	48.02%	54.64%	0.13%	-15.24%	1	2	1	2	1
56	24	114319.17	-57.99%	-57.45%	-67.43%	-72.31%	1	4	1	2	4
55	26	29807.69	86.79%	87.16%	47.74%	9.09%	1	4	1	2	3
24	27	6111.11	629.42%	652.05%	450.51%	80.96%	1	2	1	2	4
22	29	62410.39	-27.41%	-26.78%	-35.31%	-23.94%	1	2	1	2	2
1	30	27456.33	131.79%	117.36%	60.21%	32.88%	1	1	1	1	1
54	30	34793	40.30%	39.02%	28.41%	-27.59%	1	4	1	2	2
69	30	110397.33	-52.65%	-46.90%	-60.56%	-74.06%	2	1	1	2	1
133	30	89975.67	-52.86%	-46.96%	-56.71%	-61.67%	3	1	1	2	1
23	34	27497.94	87.92%	91.68%	44.71%	74.67%	1	2	1	2	3
8	41	86687.32	-46.62%	-46.58%	-60.89%	-44.30%	1	1	1	2	4
130	47	134069.57	-74.65%	-74.53%	-67.72%	-79.94%	3	1	1	1	2
2	48	298208.33	-83.58%	-85.13%	-85.05%	-86.57%	1	1	1	1	2
166	51	63804.9	-31.29%	-28.81%	-28.10%	-28.00%	3	3	1	2	2
118	52	6713.65	521.91%	581.33%	621.97%	297.34%	2	4	1	2	2
182	56	12094.64	180.14%	207.95%	258.51%	29.77%	3	4	1	2	2
129	59	13669.49	223.13%	236.19%	212.30%	118.11%	3	1	1	1	1
71	64	7804.22	487.91%	543.63%	457.45%	103.15%	2	1	1	2	3
101	64	37458.91	87.40%	100.45%	35.04%	14.50%	2	3	1	2	1
65	65	15015.85	262.50%	275.87%	217.80%	84.70%	2	1	1	1	1
102	65	55338.46	-2.38%	0.81%	-7.34%	-33.14%	2	3	1	2	2
119	65	44180.15	7.79%	19.42%	8.14%	-3.04%	2	4	1	2	3
165	66	86294.24	-33.99%	-29.15%	-47.56%	-59.04%	3	3	1	2	1
135	68	26602.21	39.96%	53.74%	46.30%	-5.23%	3	1	1	2	3
29	72	76382.92	-25.98%	-22.56%	-43.74%	-43.40%	1	2	2	2	1
66	77	11092.86	277.63%	278.01%	336.09%	158.52%	2	1	1	1	2
110	78	22515.38	130.41%	138.29%	145.73%	67.54%	2	3	2	2	2
162	78	49092.56	-7.01%	-10.91%	2.41%	1.29%	3	3	1	1	2
37	79	80065.82	2.51%	-0.84%	-41.76%	-40.17%	1	3	1	2	1
117	79	27898.73	94.47%	120.68%	71.39%	55.37%	2	4	1	2	1
183	79	15833.42	144.07%	171.32%	169.93%	114.50%	3	4	1	2	3
167	81	89877.16	-44.37%	-41.71%	-49.69%	-73.27%	3	3	1	2	3
38	83	33939.16	86.10%	73.81%	39.27%	68.02%	1	3	1	2	2
62	84	29207.02	60.50%	59.27%	65.05%	22.49%	1	4	2	2	2
17	85	51828.82	18.28%	14.27%	-15.78%	3.00%	1	2	1	1	1
61	90	151204.11	-59.71%	-58.59%	-68.55%	-80.91%	1	4	2	2	1
51	94	5088.3	1039.46%	955.68%	848.53%	53.75%	1	4	1	1	3
50	98	61941.02	-17.93%	-24.81%	-20.95%	-63.46%	1	4	1	1	2
181	98	76486.02	-42.44%	-34.46%	-44.08%	-39.03%	3	4	1	2	1

하였다. One-way, GLM, 소지역 추정으로 갈수록 과대과소 지급비율은 작아지며, 이는 고객 집단별로 정확하게 추정함을 의미한다. 적정한 보험료 책정이 화재 보험사의 중요한 업무라는 측면에서 보험금 추정은 보다 정확히 이루어져야 하며, 다른 방법과 비교하여 상대적으로 정확한 추정의 결과를 보여주는 소지역 추정은 화재 보험사의 실제 업무에 매우 유용할 것으로 판단된다. 고객 집단별 보험금 추정 분야

의 향후 연구과제로는, 고객 특성치인 설명 변수들의 합리적이고 통계적인 구분의 필요성, 보험금 기록이 전혀 없는 집단의 처리문제, 현실성 있는 오차항 분포의 사용 등을 고려할 수 있다.

참고문헌

박종태, 이상은 (2001). 소지역 추정법에 관한 비교연구. <한국데이터정보과학회지>, **12**, 47-55.

성나영, 김영원 (2000). 소지역 통계 생산을 위한 추정방법. <한국데이터정보과학회지>, **11**, 111-126.

Kim, D. H. (2007). Small domain estimation of the proportion using survey weights. *Journal of the Korean Data & Information Science Society*, **18**, 1179-1189.

Kim, Y. and Jo, R. (2004). Small area estimation of unemployment rate for the economically active population survey. *Journal of the Korean Data & Information Science Society* , **15**, 1-10.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of image. *IEEE Transaction in Pattern Analysis and Machine Intelligence*, **6**, 73-90.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398-409.

Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: An Appraisal. *Statistical Science*, **9**, 55-99.

Journal of the Korean Data & Information Science Society
2009, **20**(1), 77–87

한국데이터정보과학회지

Small area estimation of the insurance benefit for customer segmentations[†]

Yeong-Hwa Kim[1] · Ki Su Kim[2]

[12]Department of Statistics, Chung-Ang University

Received 29 December 2008, revised 10 January 2009, accepted 15 January 2009

Abstract

Bayesian methods have been focused in recent years for solving small area estimation problems. In this paper, the hierarchical Bayes procedure is implemented via MCMC techniques and compared with the results of One-way, GLM-Normal, and GLM-Gamma cases by analyzing real data of insurance benefit for customer segmentations. After analyzing insurance benefit real data for customer segmentations, we can conclude that the insurance benefit estimator through the small area estimation is more efficient than the estimators by other methods. In addition, we found that the small area estimation gave accurate estimation result for the small number domains.

Keywords: GLM, hierarchical Bayes, small area estimation.

[†] This research was supported by 2007 Research Grant for graduate student of Chung-Ang University.
[1] Corresponding Author: Associate Professor, Department of Statistics, Chung-Ang University, 221 Heuksuk-dong, Dongjak-gu, Seoul 156-756, Korea. E-mail : gogators@cau.ac.kr
[2] Graduate student, Department of Statistics, Chung-Ang University, 221 Heuksuk-dong, Dongjak-gu, Seoul 156-756, Korea.