보험금융연구 제28권 제1호 (2017. 2) pp. 3-32
DOI : https://doi.org/10.23842/jif.2017.28.1.001

Stopping Rule을 이용한 자동차보험 위험집단 구분 및 적정 손해 수준 추정

Risk Segmentation and Optimal Estimation Using Stopping Rule in Auto Insurance

김 명 준* · 이 상 준** · 김 영 화***
Myung Joon Kim · Sang Jun Lee · Yeong-hwa Kim

보험료를 결정하는 과정에서 성별과 같은 변수의 경우 구분 단위가 명확해 리스크를 추정하는 단위의 설정이 필요 없으나, 연령 변수와 같은 연속형 변수의 경우 구분 단위의 범위가 넓을 뿐만 아니라, 구분 단위를 결정하는 기준도 명확하지 않다. 1세 단위로 세분화해 손해 수준을 추정하는 경우, 해당 연령에 포함되는 고객의 수가 제한적이어서 추정된 리스크에 대한 신뢰성이 담보되기 어렵다. 또한 연령은 시간적 흐름에 따라 자연적으로 증가하는 속성을 가진 변수로 구분 단위를 결정하는 데 있어 순서의 개념이 고려되어야 하는 특징을 가지고 있다.

따라서, 본 논문에서는 연령 변수가 갖는 특징을 고려하여 이를 그룹화하는 효과적인 방법을 제안한다. 기존에 활용되고 있는 다양한 방법과 함께 연령의 순서적 개념을 고려하여 본 논문에서 새롭게 제안하는 'Stopping Rule'을 설명하고, 실제 데이터를 이용해 각 방법을 비교, 분석하였다. 실증 자료 분석을 통해 본 논문이 제안하는 Stopping Rule 방식의 적정성과 효율성을 증명한다.

국문 색인어: 그룹화, 리스크, 연령, 자동차보험, Stopping Rule
한국연구재단 분류 연구분야 코드: B051605, B051608, C030805

* 한남대학교 비즈니스 통계학과 조교수(mkim@hnu.kr), 주저자
** 중앙대학교 통계학과 대학원 석사(onlyki427@naver.com)
*** 중앙대학교 응용통계학과 교수(gogators@cau.ac.kr), 교신저자
논문 투고일: 2016. 02. 01, 논문 최종 수정일: 2016. 09. 19, 논문 게재 확정일: 2017. 02. 13

I. 서론

1. 연구 배경

현재 우리나라에 등록된 자동차의 수는 2015년 4월 기준으로 20,411,896대이다[1]. 행정자치부에 의하면 2015년 5월 기준으로 우리나라 인구수는 51,413,925명이고, 그 중에 운전이 가능한 20대 이상의 인구수는 40,948,788명이다[2]. 즉, 20대 이상의 인구 1인당 자동차 보유 대수는 0.5대로, 2명당 1대를 보유하고 있는 것으로 나타났다.

이러한 자동차를 목적물로 하는 자동차보험은 의무보험과 단기보험이라는 두 가지 큰 특징을 가지고 있다. 먼저 의무보험이라는 것은 약 2,000만대의 모든 자동차가 자동차보험에 가입했다는 것으로서, 이는 자동차보험에 관련된 자료의 양이 방대하다는 것을 의미한다. 즉, 보험회사별로 분석의 대상이 되는 자료가 많이 축적돼 있으며, 보험회사는 방대한 양의 자료를 바탕으로 다양한 통계적 분석 방식을 수월하게 적용할 수 있다. 다음으로 일부를 제외한 대부분의 자동차 보험은 보험기간이 1년인 단기보험으로서, 연구자가 1년 동안의 자료를 바탕으로 위험을 평가한 것을 1년 뒤에 검증, 수정, 보완, 재평가하는 반복적 과정이 가능하다는 것이다. 이러한 특징으로 인해 통계 분석 기법이 빈번하게 적용되는 동시에 가장 선진적인 기법들이 많이 쓰이고 있는 분야가 바로 자동차보험이다.

이러한 자동차보험에 대해 보험회사들이 가장 관심을 가지는 것은 소비자로부터 받은 보험료(insurance premium)와 사고로 지급된 보험금(insurance benefit)이다. 자동차 보험료의 가격책정 과정에 가장 큰 영향을 주는 것은 위험수준(risk)이며, 이러한 위험수준이 화폐의 개념으로 원가(true cost)에 해당한다고 할 수 있다. 따라서 동일한 원가에 대해 보험회사 입장과 고객 입장에서 각각 다른 관점으로 고려하게 된다. 보험회사의 입장에서는 원가에 적정한 회사 이익과 비용을 더해 책정된 보험 상품을 개발해 판매하려고 하는 반면, 고객의 입장에서는 사고 발생

1) 국토교통통계누리(http://stat.molit.go.kr/portal/main/portalMain.do).
2) 행정자치부(http://www.moi.go.kr/frt/a01/frtMain.do).

시 동일한 서비스가 제공된다고 하면 조금이라도 보험료가 저렴한 보험회사의 보험 상품을 선택하려고 할 것이다. 이러한 상충된 입장으로 인해 보험회사들 간의 자동차보험 가격 경쟁이 날로 심화되고 있다. 따라서 리스크의 합리적인 추정과 예측을 통해 보험 상품의 가격산정(pricing)을 하는 것이 매우 중요한 문제가 되고 있으며, 보험 상품의 가격책정은 다음과 같이 크게 두 부분으로 생각할 수 있다.

첫 번째는 기본 보험료 수준을 예측하는 것으로, 순보험료와 부가보험료로 구분되며, 보험회사에서는 전체적인 보험료의 수준에 대해 과거 데이터에 근거하여 예측하게 된다. 전체적인 보험료의 수준, 즉 기본 보험료라는 것은 연도별로 사고빈도(frequency)의 증감과 사고 건당 발생하는 피해금액인 심도(severity)의 증감에 대한 추세를 예측하는 것을 말한다. 함상호(1998)는 자동차보험 가격자유화 본격 시행과 관련해 보험회사의 경쟁전략 수립방향에 초점을 두어 현실적으로 도입 가능한 가격자유화 추진방향을 제시했고, 통계적으로 보다 정확한 예측을 위한 연구가 김명준(2013)에 의해 제안되었다. 보험료 수준은 물가에 민감한 영향을 주기 때문에 정부에서는 전체적인 수준의 보험료를 통제하게 되는데, 이러한 정부의 통제 하에서 보험회사가 자체적 경쟁력을 갖기 위한 노력을 하게 되는 것은 당연한 것이며 그 답은 바로 두 번째에 있다.

두 번째는 기본 보험료 수준이 동일한 상태에서 각각의 보험료를 차별화시키는 것으로, 고객의 특성에 맞게 개인별 또는 그룹별 위험수준을 추정해 적용하는 가입자 특성요율이다. 보험회사들은 보험업감독규정 시행세칙에 따라 통계적 근거에 의거해 고객들에게 신뢰를 줄 만한 요율을 산출해야한다. 예를 들어, 나이가 한 살 더 많아졌다고 해서 갑자기 보험료가 대폭 인상된다면 보험료 책정의 타당한 근거를 제시해도 고객의 입장에서는 납득할 수 없게 된다. 이는 보험요율은 과도하거나 과소해서는 안된다는 비과도성 원리에 부합하지 않으며, 요율 검증을 담당하는 금융 감독자는 물론 고객을 설득하는 것도 어렵게 된다. 따라서 고객에게 신뢰를 줄 수 있는 적정한 범위내의 수준을 예측하고 추정해야 한다. Kim and Kim(2013)은 보험료 산출에 처음으로 베이지안 프레임워크 하에서의 평가 방식을 적용했다.

위험수준 추정에 영향을 끼치는 변수를 위험 요인이라고 하며, 위험수준을 추정할 때 사용되는 변수의 그룹핑 방법에 따라 결과가 달라질 수 있다. 이러한 변수들의 종류에는 성별, 차량의 크기, 차량 옵션, 운전자 경력, 운전자 연령 등이 있으며, 이들 중 성별이나 차량의 크기 같이 그룹핑 기준이 명백하기 때문에, 이러한 기준에 따라 구분된 자료에는 현재의 통계 기법을 사용해서 추정을 하는데 큰 어려움이 없다.

변수 중에 그룹핑을 하는데 있어서 적용 기준이 모호하면서 보험회사 간 가장 주요 쟁점이 되고 있는 변수가 바로 연령이다. 연령으로 최적의 그룹핑을 하는 방법은 연령을 연속형 변수로 간주해 1세 단위로 분류하는 방법이다. 그러나 연령을 1세 단위로 분류하게 되면 표본수가 적은 연령이 존재할 가능성이 있는데, 이러한 경우 표본의 대표성과 신뢰도 부분에서 문제점이 생길 수 있다. 특히 중소규모 회사의 경우 저연령 구간에서 현상이 발생하며, 대형 회사에서도 1세 단위로 세분화하게 되는 경우 일부 연령 구간에서 동일한 현상이 발생한다. 이러한 문제점을 극복하고자 이창수(1997)는 보험자료의 충분성 평가와 보험요율의 조정을 위해 신뢰도 기법을 적용하는 방안을 제시했고, 김영화・이현수(2010)는 다양한 신뢰도 적용 방안을 제시했다. 또한, 근접한 연령그룹 간에 동일한 그룹이 돼 추정된 위험수준이 그룹 내 연령별로 유사한 것인지, 이질적인 것인지를 구분하는 명확한 기준이 존재하지 않는다는 문제점도 있다. 마지막으로 그룹별로 추정된 위험수준을 적용하는 경우, 특정 연령에서 급격한 변동이 발생하는 개연성이 존재하는 경우가 있어 고객의 입장에서는 쉽게 받아들일 수 없는 어려운 현실적인 문제도 존재한다. 이러한 여러 문제점 때문에 연령을 그룹핑하는 방식이 보험회사별 주요 이슈가 되고 있다.

2. Stopping Rule 제안 목적

보험요율에 과도한 변동이 발생하는 경우, 이를 검증하고 승인하는 감독 당국이나 보험 계약자를 설득하는 것은 어려운 문제이며 이는 위험수준의 참값을 추

정하는 문제와는 별개로 고려해야하는 사항이다. 따라서 현재 국내에서 적자를 감당하고 있는 보험회사들은 수지 상등의 원칙과 각 보험회사의 가격 경쟁 측면을 모두 고려해야하고, 적정 요율을 부과하지 못하는 계층에 대해서는 인수 방침에 따라 계약자의 유입을 차단하는 방안을 활용하고 있기도 하다.

구분 단위가 시간의 흐름에 따라 자연적으로 증가하는 연령 변수의 경우는 특히 고려해야하는 문제가 많은 변수 중 하나이다. 통계학에서 활용하는 군집(clustering) 개념의 적용은 인접하지 않은 연령 간에 하나의 군집이 돼 연령 증가에 따라 앞에서 지적한 요율의 급격한 변동이 발생할 개연성이 있으며, 인접한 연령 간에 동일한 연령 그룹이 되는 것과 그렇지 않은 것에 대해 적용되는 보험료의 변동이 발생하기 때문이다.

따라서 이를 해결하고자 요율 변동의 과격함을 완만하게 이어주는 평활법이 제안되어 사용되어 오고 있다. 그러나 이 또한 하나의 위험 함수의 오차가 최소화하는 방안을 고려하고 있을 뿐 연령별로 해당하는 위험의 참값에 맞는 요율을 적용해야한다는 대원칙에 부합하지 않는 문제가 발생하게 된다.

이러한 문제점을 해결하고자, 본 논문은 연령의 증가에 따라 동일 그룹의 포함 여부를 리스크의 추정값과 분산을 활용해 판단하고 결정하는 새로운 그룹핑 방식인 'Stopping Rule'을 제안한다. 이는 요율의 참값을 최대한 반영하면서 요율의 변동을 일부 억제할 수 있는 대안이 될 수 있다. 즉, 연령별 참값과 인접한 연령간의 차이를 검정해가며 그룹을 결정하는 방식으로 하나의 함수식으로 오차를 최소화하는 평활법의 단점을 보완하면서, 각 그룹의 참값을 그대로 반영할 수 있는 장점을 가지는 방식이라고 할 수 있다. 이의 효용성을 증명하고자 두 가지 고려 대상을 평가할 수 있는 지수를 설정하고, 실증 분석을 통해 기존의 방법과의 비교 분석을 실시하고자 한다. 이는 보험회사 입장에서 참값의 적용의 폭의 확대하고 감독당국과 보험 계약자를 설득할 수 있는 대안이 될 수 있으며, 또한 이를 바탕으로 산출된 조정율의 반영 비율을 조정해 타사의 경쟁을 위한 전략적 판단이 가능할 것으로도 판단된다.

3. 논문의 구성

본 논문은 보험료 책정에 대해 전체적인 수준 예측보다는 연속적인 변수로 간주할 수 있는 연령에 대해 최적의 그룹핑 방식을 제안하는 것에 중점을 두었다. 실제로 위험 그룹을 분류하는 방법이 현재 활발하게 논의되고 있지 않은 상태일 뿐만 아니라 보험회사들 간의 큰 현안 가운데 하나가 연령이기 때문에 이에 대한 그룹핑에 중점을 두었다. 그룹핑을 하는 방법에 있어서 기존에 있는 이론뿐만 아니라 새로운 'Stopping Rule'이라는 방법을 제시하고, 이를 토대로 적용한 실증 분석을 통해 연령 그룹핑 방식의 합리적인 대안을 제시하고자 하며, 논문의 구성은 다음과 같다.

서론에 이어 II장에서는 회귀모형의 분류 및 스플라인에 대해 설명하고, III장의 실증 자료 분석에서는 국내 보험회사의 데이터를 토대로 연령에 대한 기존의 6가지 방법과 본 논문이 제안하는 방법을 사용해 연령 그룹핑을 하고, 각각의 손해 수준을 추정했다. 이를 바탕으로 연령구간 세분화 측도를 나타내는 지수와 연령 변동 수준을 나타내는 지수를 산출했으며, 마지막 IV장에서는 III장에서 구한 여러 가지 방법의 결과들을 바탕으로 본 연구가 제안하는 방법의 효과성과 적정성에 대한 검증을 위해, 제시된 지수의 결과를 비교, 제시하였다.

II. 선행 연구

본 논문에서 사용한 대부분의 연령 그룹핑 방법들은 이론적으로 어렵지 않기 때문에 다음 장의 실증 자료 분석에서 설명하기로 한다. 다만 스플라인(spline) 방식의 경우는 이론적인 내용이 방대하고 난해하기 때문에 본 논문에서 사용하는 평활 스플라인(smoothing spline)에 대해서만 자세하게 다루고 나머지 부분은 간략하게 설명하기로 한다. 본 장에서 설명하는 이론적인 내용은 김충락 · 강근석(2010)을 참고했다.

1. 회귀모형의 분류

회귀모형은 크게 모수 회귀모형과 비모수 회귀모형으로 구분되며, 이에 따른 추정법도 다양하다. 회귀모형을 구분하는 기준은 회귀모형에서 표현되는 회귀함수의 설정 방법에 따라 달라진다. 회귀함수의 형태가 사전에 주어지는 경우가 모수 회귀모형에 속하며, 회귀함수가 미리 주어지지 않고, 특정한 조건을 만족하는 함수군에 속한다고 가정하는 경우가 비모수 회귀모형에 해당한다. 이 경우에는 무수히 많은 함수군에서 특정 조건을 만족하는 하나의 함수를 선택해야 하므로 함수 자체 또한 추정의 대상이 된다.

자동차보험에서는 일반적인 정규분포 가정이 어려우므로 GLM(Generalized Linear Model) 방식이 많이 활용되고 있으며, 이러한 방식을 적용하는 연구를 Jorgensen(1994), Murphy(2000) 등이 제시했고, 국내에서도 김영화·김명준(2009)에 의한 연구결과가 발표됐다. 또한 최우석·한상일(2008)은 보험요율 추정에 DGLM(Double Generalized Linear Model)을 이용해 정확도를 향상하는 대안을 제시했다.

자동차보험에 주로 사용되는 비모수 회귀모형의 추정 방식으로는 커널 추정, 급수 추정, 스플라인 추정 등이 있다. 이러한 방식 중 본 연구의 실증자료 분석에는 평활 스플라인(smoothing spline) 추정이 사용됐으며, 이는 회귀함수를 소볼레프 공간(Sobolev space)으로 축소시킨 다음 회귀함수의 적합도와 회귀함수의 평활성을 적절하게 조화시켜 추정치를 구하는 방법이다. 이와 관련해 이우동 외 3인(1998)이 연구결과를 제시했고, 보험분야에 적용하는 연구결과를 Vickor(2008) 등이 제안했다.

전술한 바와 같이, 비모수 회귀모형에서는 무수히 많은 함수군에서 특정한 조건을 만족시키는 가장 좋은 하나의 함수를 선택해 추정하게 되는데, 주어진 자료를 직선과 같이 단순한 함수에 적합 시키는 경우 해석은 용이하나 오차가 커지는 경향이 나타나며, 반대로 복잡한 함수에 적합 시키는 경우 오차는 감소하나 회귀모형에 대한 해석이 어려워진다. 따라서 오차의 크기를 줄이는 동시에 함수의 단순성도 함께 고려할 수 있는 방안이 검토되어야 한다.

2. 평활모수(smoothing parameter)의 추정방식

최적의 함수는 오차의 크기가 작은 동시에 단순성도 가지고 있어야 한다. 즉, 오차 제곱 합은 작으면서 상대적으로 매끄러운(smooth) 곡선을 가진 함수를 좋은 함수라고 할 수 있다. 매끄러운 정도를 2차 미분 값에 근거한다면, 다음 두 가지 조건을 동시에 충족하는 함수를 좋은 함수라고 할 수 있다.

$$(1)\ \sum\{y_i - f(x_i)\}^2 \ :\ \text{적합도}$$

$$(2)\ \int_a^b f''(x)^2 dx \ :\ \text{매끄러운 정도}$$

그러나 위의 두 가지 조건은 동시에 충족할 수 없는 관계이다. 수식 (1)에 해당하는 적합도의 오차를 작게 하려면 차수가 큰 다항식이 필요하며 그런 경우에는 (2)의 매끄러운 정도가 커지게 된다. 반대로 (2)의 매끄러운 정도를 작게 하려고 차수가 작은 다항식을 사용하는 경우 수식 (1)에 해당하는 적합도의 오차는 커지게 된다. 따라서 (1)과 (2)를 균형 있게 조절할 수 있는 가중치 $q(0 < q < 1)$ 를 사용해 최적의 회귀함수를 선택해 추정하는 것이 적절한 방법이라고 할 수 있다. 즉, 가중치 $q(0 < q < 1)$ 에 대해 다음 식의 값을 최소로 하는 함수 f 를 구하는 과정으로 이해할 수 있다.

$$(1-q)\sum_{i=1}^{n}\{y_i - f(x_i)\}^2 + q\int_a^b \{f''(x)\}^2 dx$$

여기서 $\lambda = \frac{q}{1-q}$ 라고 하면 위 식은 다음과 같이 표현된다.

$$S(\lambda) = \sum_{i=1}^{n}\{y_i - f(x_i)\}^2 + \lambda\int_a^b \{f''(x)\}^2 dx$$

이를 최소로 하는 회귀함수를 f 라 하면 위 식은 다음과 같이 표현할 수 있다.

$$\hat{f_\lambda}(x) = \underset{f \in W_2^n[a,\,b]}{\arg\min} \sum_{i=1}^{n}\{Y_i - f(x_i)\}^2 + \lambda\int_a^b \{f''(x)\}^2 dx$$

여기서 λ는 적합도와 매끄러운 정도의 균형을 맞추는 값으로 평활모수(smoothing parameter)라고 하며, 평활모수 λ 의 값을 최소화가 적합도를 충족시키는 과정과 동

일하다고 볼 수 있다. 반대로 λ가 매우 클 경우에는 2차 미분 값이 최소가 되므로 직선에 가깝게 나타난다. 평활모수 λ를 추정하는 데 있어 교차확인(Cross Validation: CV) 또는 일반화 교차확인(Generalized Cross Validation: GCV) 방식이 일반적으로 활용으로 CV와 GCV의 값을 최소화하는 λ를 선택해 사용하게 된다.

따라서 이렇게 구해진 함수에서 설명 변수가 연령이 되고, 해당 연령에 해당하는 리스크가 종속 변수에 해당된다. 이 경우 1세 단위로 세분화된 그룹과 동일한 결과를 적용할 수 있는 장점이 있는 반면, 평활화로 발생하는 연령별 오차를 간과하는 문제점을 동시에 가지게 된다. 본 논문은 새로운 방식과 현재 연속형 형태의 리스크 추정에 가장 많이 활용하는 평활 방식을 비교 대상으로 활용한다. 평활 방식 이외에도 제안된 다양한 방식들이 있으나, 연구의 취지와 목적에 부합하도록 평활 방식을 비교 대안 중의 하나로 설정해 진행한다.

III. 실증 자료 분석

현재 국내 자동차보험회사는 고객들의 다양한 정보를 가지고 있으며, 보유한 정보를 이용해서 손해 수준(risk)을 예측하고 추정해 고객들이 납부할 보험료를 산정한다. 이때, 그룹핑 방법에 따라 보험회사의 손익과 고객이 부담해야할 보험료 규모가 달라지므로, 보험료 책정에서 그룹핑이 매우 중요한 절차가 된다. 본 장에서는 국내 자동차보험회사의 실제 데이터를 사용해 여러 가지 방법에 따라 연령 그룹핑을 실시하고, 각 그룹핑 하에서의 리스크를 추정해, 각 방식에 대한 적합성 여부를 비교해 보고자 한다.

1. 고객 데이터의 구성

본 연구의 분석에 사용된 데이터는 국내 자동차보험회사의 실제 데이터로, 특정 담보에서 사고가 발생한 45,466개의 표본으로 구성되며, 여러 변수 중 분석에 필요한 운전자의 연령과 사고 발생 손해액을 고려해 사고 심도에 대한 비교를 진행한다.

보험개발원이 제시하는 연령 그룹과 참조 요율이 존재하기는 하나, 통상적으로 보험회사별로는 이와 상이한 연령 그룹을 적용하고 있는 바, 국내 특정 보험회사에서 적용하고 있는 연령 그룹핑 기준을 준용했으며, 적용 기준은 〈Table 1〉과 같다.

〈Table 1〉 Age Grouping Criteria of Certain Insurance Company

Classification	Age
20	Below 20
23	Over 20 ~ Below 23
25	Over 23 ~ Below 25
⋮	⋮
73	Over67 ~ Below 73
99	Over73

〈Table 2〉 Result of Data Adjustment

Certain Company Criteria	
Classification	Sample (%)
~ 20	42 (0.09)
21 ~ 23	372 (0.82)
24 ~ 25	681 (1.50)
26 ~ 29	3,274 (7.20)
30 ~ 32	2,843 (6.25)
33 ~ 42	2,662 (5.85)
43 ~ 47	7,343 (16.15)
48 ~ 52	7,414 (16.31)
53 ~56	4,260 (9.37)
54 ~ 60	2,741 (6.03)
61 ~ 67	2,693 (5.92)
68 ~ 73	986 (2.17)
74 ~	349 (0.77)
Total	45,466 (100)

⇨

One Year Unit Classification			
Group	Age	Population	Sample
20	20	100%	42
23	21	33.229%	124
	22	33.918%	126
	23	32.853%	122
25	24	50.347%	343
	25	49.653%	338
29	26	24.783%	811
	27	24.461%	801
	28	24.876%	815
	29	25.879%	847
⋮	⋮	⋮	⋮
73	68	16.840%	166
	69	15.026%	148
	70	16.025%	158
	71	16.100%	159
	72	19.791%	195
	73	16.218%	160
99	74	10.627%	37
	⋮	⋮	⋮
	88	1.812%	6
	89	1.532%	5
	⋮	⋮	⋮
	99	0.099%	0

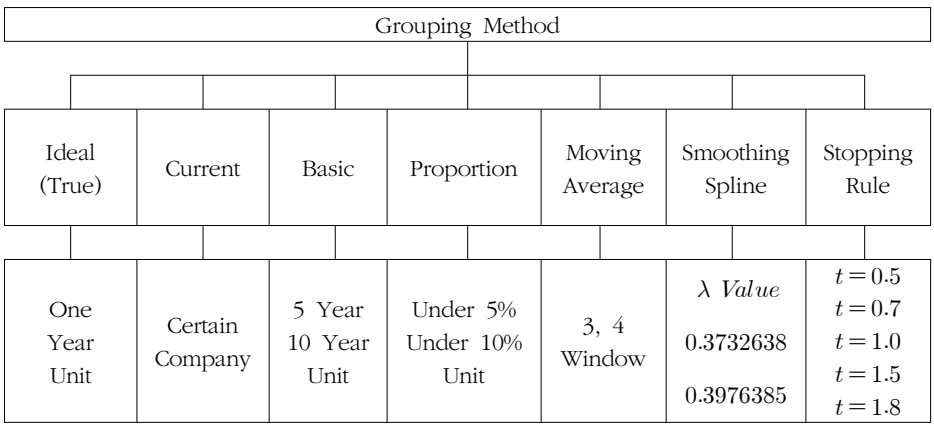
본 연구는 〈Table 1〉과 같은 연령 그룹핑 데이터와, 참값의 비교 기준이 되는 1세 단위의 데이터가 필요하다. 그러나 이러한 자료의 취득이 불가한 바, 〈Table 1〉과 같은 데이터를 현재 국가통계에서 조사된 연령의 인구비율에 맞도록 랜덤하게 조정했으며, 그 결과가 〈Table 2〉와 같다.

〈Table 2〉에서 74세부터 99세까지는 다른 연령대에 비해서 빈도가 낮게 나타나 신뢰도의 문제의 개연성이 존재하는 바, 1세 단위일 때의 분석에만 포함시키고, 기타 방법의 그룹핑 결과와 비교하는 과정에서는 제외시켜 실제적인 분석 과정에 포함되는 연령은 20세부터 73세까지가 된다.

2. 연령 그룹핑 방법

본 논문에서 가장 중점을 두고 있는 부분은 연령 변수에 대한 최적의 그룹핑을 도출하는 것이다. 본 연구에서는 그룹핑의 다양한 방법 중 다음과 같은 7가지 방법을 사용해서 연령에 대한 그룹핑을 진행하고 그 결과를 비교한다.

〈Figure 1〉 Method of Age Grouping



첫 번째 방법은 이상적이지만 현실적인 제약이 많아서 현업에서 적용하기 힘든 1세 단위의 구분이다. 두 번째는 현재 국내 특정 보험회사에서 사용하고 있는 그룹핑이며, 세 번째는 가장 간단하면서 기본적인 방법으로 연령을 5세 단위와 10세

단위로 구분하는 방법이다. 네 번째는 구성 분포를 고려하는 방식으로 전체 데이터의 개수에 대해서 5%이내, 10%이내로 구분하는 방법이다. 다섯 번째는 이동평균법을 사용하는 것으로, 창(window)의 크기를 각각 3개씩, 4개씩 이웃돼 있는 연령끼리 묶는 방법이다. 여섯 번째는 평활 스플라인(smoothing spline)을 사용하는 방법으로 λ값에 따라 그룹핑 하는 방식이다. 마지막으로는 'Stopping Rule'은 본 논문이 새롭게 제시하는 그룹핑 방법이다. 각각의 방법으로 그룹핑을 실시한 후, 리스크의 추정 값인 상대도를 구하고 각 그룹핑에 대한 장단점을 확인하기 위해 연령 세분화와 연령별 변동을 측정하는 두 가지 기준의 지수 값을 정의하고 도출해 구분 방식에 대한 효용성을 검증한다.

3. 이상적인 기준의 연령 그룹핑

1세 단위 그룹핑은 다른 방법의 결과들과 비교하기 위해서 필요한 부분이다. 본 논문은 다양한 방법으로 그룹핑을 진행한 후, 최종적으로 두 가지 지수 값을 구하고, 1세 단위의 결과에서 구해지는 지수 값의 결과와 비교해 적정성을 검증하게 된다.

가. 1세 단위별 연령에 대한 상대도(R_i)

본 실증자료 분석에서 참값으로 고려하는 1세 단위별 연령 상대도(R_i)를 구하는 공식은 식 (1)과 같으며, 추정된 상대도는 〈Table 3〉과 같다. 여기서 $\overline{X}$ 는 전체 손해액 데이터의 평균을 의미하며, $\overline{X_i}$는 각 연령에 해당하는 손해액의 평균을 나타낸다. 따라서 전체 평균과 해당 연령의 평균이 동일한 경우, 즉 $\overline{X}=\overline{X_i}$ 인 경우 상대도는 1이 되고, $\overline{X}<\overline{X_i}$ 인 경우는 전체적인 위험수준보다 해당 연령의 리스크가 높은 경우이므로 상대도는 1보다 크게 나타나며, 반대의 경우 1보다 작은 상대도를 가지게 된다.

$$R_i=\frac{\overline{X_i}}{\overline{X}}=\frac{\overline{X_i}}{908{,}037} \tag{1}$$

나. 연령 그룹핑 방법 비교를 위한 지수

가격의 결정은 참값(true cost)에 수렴하도록 책정하려는 이론적 관점과 보험료 변동이 상식적인 수준에서의 이루어져야 한다는 현실적 관점이 모두 고려돼야 한다. 이러한 이유 때문에 본 논문은 세분화 수준을 반영하는 지수와 변동 수준을 나타내는 지수를 각각 구해 비교하고자 하며, 전자의 지수 값을 I_1, 후자의 지수 값을 I_2 로 정의한다.

〈Table 3〉 Relativity(R_i) of One Year Unit

Age	R_i	Age	R_i	Age	R_i	Age	R_i
20	0.63058	40	0.93898	60	0.99478	80	0.88626
21	1.20288	41	0.96740	61	0.86493	81	4.24561
22	1.17785	42	1.04753	62	0.94186	82	0.62163
23	1.48116	43	0.95680	63	1.22720	83	0.95114
24	1.07216	44	0.98545	64	1.17493	84	0.87953
25	1.12568	45	0.95094	65	1.05880	85	1.01434
26	1.14753	46	1.02223	66	1.00092	86	0.76949
27	0.87603	47	0.98068	67	0.95586	87	1.27244
28	1.10115	48	0.99283	68	1.00938	88	0.75030
29	0.96554	49	1.01729	69	1.28883	89	0.48007
30	0.97491	50	0.94826	70	1.19702	90	0.53227
31	0.90977	51	0.97950	71	1.18056	91	0.52936
32	0.94542	52	0.98816	72	1.07501	92	0.47004
33	1.08494	53	0.92239	73	0.79335	93	1.11625
34	0.93051	54	1.03016	74	1.13653	94	0.31932
35	0.95581	55	0.97549	75	0.87011	95	0.66077
36	0.97235	56	1.06937	76	1.15869	96	0.54912
37	0.98467	57	1.07682	77	1.07034	97	3.55762
38	1.01255	58	1.01500	78	0.80970	98	0.00000
39	0.98762	59	1.13293	79	1.15485	99	0.00000

(1) 연령 구간 세분화 측도의 오차 수준 지수

각 그룹별 손해액의 평균을 이용해 각 그룹별 상대도(C_i)를 구하고 제시된 수식을 사용해 오차 수준을 나타내는 지수 I_1의 값을 구한다. 여기서 R_i는 1세 단위 적용 기준의 상대도를 의미하며, C_i는 해당 그룹핑 방식에 따라 구해진 각 그룹의 상대도를 의미한다. n은 연령그룹의 수로 오차제곱을 자유도로 나누는 것을 의미하므로 그룹핑 방식에 따라 n값은 다르게 나타난다.

$$I_1 = \sum_{i=1}^{n} (R_i - C_i)^2 / (n-1) \tag{2}$$

식 (2)를 사용해 각각의 연령 그룹핑 방식에서 도출된 지수 I_1 의 값과 이상적인 구분 방식으로 설정한 1세 단위 연령 그룹핑에서 도출된 지수의 값을 비교한다. 오차 수준 지수 I_1 는 R_i와 C_i 간의 편차 개념으로 설정했으며, 따라서, I_1의 값이 작을수록 이상적이라 할 수 있다.

(2) 연령 변동 수준 반영 지수

고객은 가격 변동의 영향을 받는 입장으로 연령 증가에 따른 보험료의 변동이 최소화돼 계획된 또는 과거와 유사한 보험료를 납부하기를 바란다. 각 연령별 보험료의 변동률이 작을수록 보험료의 가격이 변함이 없다는 것을 의미하게 되는데, 이러한 영향도를 나타내는 지수 I_2의 값을 변동 지수라 정의하고, 이 지수를 합리적 그룹핑의 두 번째 기준으로 설정했다.

$$I_2 = \sum_{i=1}^{n} \frac{(C_i - C_{i+1})^2}{C_i} / (n-1) \tag{3}$$

변동 지수도 수치가 작을수록 보험료의 변동이 최소화된다는 의미이며, 이는 고객 관점에서 합리적으로 수용 가능한 결과일 것이다. 다양한 방법으로 상대도와 두 가지 지수 값을 구하고, 그룹핑 방식별 결과를 비교해 정의된 지수를 최소화하는 최적의 그룹핑 방법을 도출하게 된다.

4. 현업에서 활용하는 연령 그룹핑과 분석 결과

〈Table 1〉은 현재 특정 보험회사에서 실제로 사용 중인 연령 그룹핑 방식이며, 이에 근거해 분석을 적용했다. 연령을 20~73세까지 고려했을 때, 그룹의 개수는 15개이며, 구분한 그룹의 상대도(C_i) 결과는 다음 〈Table 4〉와 같다.

〈Table 4〉 Age Grouping and Relativity(C_i) of Certain Company Criteria

Age	C_i	Age	C_i	Age	C_i
20	0.63058	33~35	0.99067	48~52	0.98541
21~23	1.28567	36~37	0.97874	53~56	0.99711
24~25	1.09872	38~40	0.97817	57~60	1.05793
26~29	1.02248	41~42	1.00749	61~67	1.02315
30~32	0.94221	43~47	0.97913	68~73	1.08692
Number of Group		I_1		I_2	
15		0.02950		0.05233	

추정된 상대도(C_i) 결과를 식 (2)와 (3)에 적용, 도출한 지수의 결과도 〈Table 4〉에 포함했다.

5. 기본적인 방법을 이용한 연령 그룹핑과 분석 결과

연령 그룹핑을 하는데 있어 가장 기본적인 방법은 연령을 5세 단위 또는 10세 단위로 분류하는 것이다. 이러한 기본적인 방법을 이용하면 그룹핑 작업이 수월하고, 신뢰도가 보장되는 표본 수 확보가 가능하다는 장점을 가지고 있다. 그러나 포함 연령이 많아질수록 그룹 수가 감소하고, 일부 고연령 그룹에서는 표본수가 부족할 뿐만 아니라 동일그룹 내 고객들 사이에 불만이 발생할 개연성이 있고, 분류기준이 모호하다는 단점을 가지고 있다.

기본적인 방법을 이용해 5세 단위와 10세 단위로 구분해 추정한 각 그룹의 상대도(C_i)와 지수 I_1, I_2 값을 구한 결과가 다음 〈Table 5〉, 〈Table 6〉과 같다.

〈Table 5〉 Age Grouping and Relativity(C_i) of 5 Year Unit

Age	C_i	Age	C_i	Age	C_i
20~24	1.15258	40~44	0.97827	60~64	1.02437
25~29	1.03214	45~49	0.99307	65~69	1.03542
30~34	0.97111	50~54	0.97363	70~73	1.06552
35~39	0.98346	55~59	1.04889		
Number of Group		I_1		I_2	
11		0.07729		0.00244	

〈Table 6〉 Age Grouping and Relativity(C_i) of 10 Year Unit

Age	C_i	Age	C_i	Age	C_i
20~29	1.05301	40~49	0.9858	60~69	1.0287
30~39	0.97787	50~59	1.00236	70~73	1.05513
Number of Group		I_1		I_2	
6		0.16053		0.00141	

6. 비율을 이용한 연령 그룹핑과 분석 결과

비율을 이용한 연령 그룹핑은 전체 데이터 수에서 5%이내 또는 10%이내에 포함된 데이터 수에 맞게 연령을 그룹핑 하는 것으로 그룹별로 일정 수준의 데이터를 확보하는 방식이다. 비율 방식을 이용하면 각 그룹에 포함된 데이터의 수가 유사하다는 장점을 가지고 있지만 비율에 대한 명확한 분류 기준이 없고, 전체 데이터 수가 달라질 때마다 새롭게 계산을 해야 할 뿐만 아니라 특정 연령대에서 비율을 넘는 데이터 수를 보유하고 있는 경우, 그룹핑 방법에 어긋난다는 단점을 가지고 있다. 실제 데이터에 비율 방식을 적용하는 방식은 다음과 같이 나타낼 수 있다.

$$Grouping\ Criteria \leq\ 45{,}466 \times \frac{x}{100}$$

단, 특정 연령에서 분할기준을 넘는 현상이 발생할 경우, 하나의 연령을 2개로 구분할 수 없으므로 하나의 그룹으로 간주하도록 한다. 얘를 들어 기준이 5%인 경우에는 $45{,}466 \times 5\% = 2{,}273.3$ 으로, 각 연령에 포함된 데이터 수를 더해 2,273개

이하에 속하면 하나의 그룹으로 간주해 그룹핑을 하게 되고 10% 기준인 경우 4,546개 이하에 속하면 하나의 그룹으로 간주해 그룹핑을 진행한다.

이러한 방식으로 연령을 그룹핑한 후, 상대도(C_i)와 지수 값을 도출한 결과가 다음의 〈Table 7〉, 〈Table 8〉과 같다.

〈Table 7〉 Age Grouping and Relativity(C_i) of 5% Unit

Age	Sample	C_i	Age	Sample	C_i	Age	Sample	C_i
20~26	1,906	1.14566	40	1,422	0.93898	50	1480	0.94826
27~28	1,616	0.98957	41	1,330	0.96740	51	1408	0.97950
29~30	1,689	0.97021	42	1,332	1.04753	52	1565	0.98816
31~32	2,001	0.92844	43	1,514	0.95680	53~54	2257	0.97653
33~34	2,236	1.00787	44	1474	0.98545	55~56	2003	1.0203
35	1,103	0.95581	45	1483	0.95094	57~59	2141	1.07563
36	1,190	0.97235	46	1488	1.02223	60~64	2195	1.02437
37	1,284	0.98467	47	1384	0.98068	65~73	2375	1.0421
38	1,262	1.01255	48	1449	0.99283			
39	1,309	0.98762	49	1512	1.01729			
Number of Group			I_1			I_2		
28			0.02657			0.00282		

〈Table 8〉 Age Grouping and Relativity(C_i) of 10% Unit

Age	Sample	C_i	Age	Sample	C_i	Age	Sample	C_i
20~29	4,369	1.0530	40~42	4,084	0.9836	52~54	3,822	0.9813
30~33	3,963	0.9825	43~45	4,471	0.9643	55~59	4,144	1.0489
34~36	3,409	0.9533	46~48	4,321	0.9991	60~73	4,539	1.0455
37~39	3,855	0.9948	49~51	4,400	0.9820			
Number of Group			I_1			I_2		
11			0.07773			0.00141		

7. 이동평균법에 의한 연령 그룹핑과 분석 결과

이동평균법은 평균의 계산 기간을 순차적으로 한 개항씩 이동시키면서 기간별 평균을 계산해 경향치를 구하는 방법이다. 이동평균법의 장점은 그룹 분류가 간

편하고, 1세 단위와 비교 했을 때 표본수가 적다는 단점을 보완할 수 있으나, 정확한 분류 기준이 없어 연구자마다 창의 크기가 다를 수 있고, 그룹수가 상당히 많아질 뿐만 아니라 맨 처음과 마지막 그룹의 표본수가 다른 그룹의 표본 수에 비해 현저히 작아진다는 단점이 있다. 〈Table 9〉, 〈Table 10〉는 이동평균법에 따른 연령 그룹핑과 상대도(C_i)와 지수를 창의 크기에 따라 구한 결과이다.

가. 창(window)의 크기가 3개인 경우

〈Table 9〉 Age Grouping and Relativity(C_i) of 3 Windows

Age	C_i	Age	C_i	Age	C_i	Age	C_i
20	0.63058	32~34	0.98794	46~48	0.99906	60~62	0.94133
20~21	1.05808	33~35	0.99067	47~49	0.99747	61~63	0.99442
20~22	1.10976	34~36	0.95330	48~50	0.98631	62~64	1.09707
21~23	1.28567	35~37	0.97167	49~51	0.98198	63~65	1.15178
22~24	1.17912	36~38	0.99016	50~52	0.97216	64~66	1.07807
23~25	1.15683	37~39	0.99480	51~53	0.96715	65~67	1.00520
24~26	1.12526	38~40	0.97817	52~54	0.98130	66~68	0.98405
25~27	1.03222	39~41	1.08771	53~55	0.97620	67~69	1.04127
26~28	1.04235	40~42	0.98364	54~56	1.02386	68~70	0.79919
27~29	0.98131	41~43	0.98911	55~57	1.03577	69~71	1.22061
28~30	1.01283	42~44	0.99455	56~58	1.05624	70~72	1.14544
29~31	0.94841	43~45	0.96430	57~59	1.07563	71~73	1.01998
30~32	0.94221	44~46	0.98625	58~60	1.05076	72~73	0.94806
31~33	0.98460	45~47	0.98475	59~61	1.01936	73	0.79335

Number of Group	I_1	I_2
56	0.01434	0.01351

나. 창(window)의 크기가 4개인 경우

〈Table 10〉 Age Grouping and Relativity(C_i) of 4 Windows

Age	C_i	Age	C_i	Age	C_i	Age	C_i
20	0.63058	32~35	0.97986	47~50	0.98497	62~65	1.08794
20~21	1.05808	33~36	0.98586	48~51	0.98467	63~66	1.11351
20~22	1.10976	34~37	0.96189	49~52	0.98360	64~67	1.04737
20~23	1.21921	35~38	0.98233	50~53	0.96214	65~68	1.00575
21~24	1.18324	36~39	0.98950	51~54	0.98081	66~69	1.02721
22~25	1.15968	37~40	0.97976	52~55	0.98005	67~70	0.86772
23~26	1.15216	38~41	0.97548	53~56	0.99711	68~71	0.89529
24~27	1.03820	39~42	1.07685	54~57	1.03414	69~72	1.17759
25~28	1.05254	40~43	0.97638	55~58	1.03166	70~73	1.06161
26~29	1.02248	41~44	0.98816	56~59	1.07370	71~73	1.10180
27~30	0.97968	42~45	0.98340	57~60	1.05793	72~73	1.15965
28~31	0.98442	43~46	0.97877	58~61	1.01813	73	1.50904
29~32	0.94756	44~47	0.98492	59~62	1.00284		
30~33	0.98254	45~48	0.98676	60~63	0.99454		
31~34	0.97036	46~49	1.00379	61~64	1.03551		

Number of Group	I_1	I_2
57	0.02922	0.01066

8. 평활 스플라인을 이용한 그룹핑과 분석 결과

본 절에서는 2장 선행연구에서 다룬 내용을 바탕으로 평활모수(smoothing parameter)의 추정치를 구하는 방식을 다룬다. 평활 스플라인을 사용해 연령 그룹핑 하는 경우의 장점은 앞의 방법들과는 달리 분류 기준이 있고, 그룹수가 동일하다는 것이다. 하지만 계산과정이 복잡하고, 교차확인값(CV)과 일반화 교차확인값(GCV)을 통해 추정한 평활모수의 값이 상이하고 그룹수가 많아진다는 단점을 가진다.

각각의 평활모수에 대한 교차확인값과 일반화 교차확인값은 다음 〈Table 11〉과 같으며, 굵은 체로 처리돼 있는 부분이 교차확인과 일반화 교차확인의 값을 최소화하는 λ에 해당한다.

〈Table 11〉 Cross Validation Values for Smoothing Parameters

Smoothing Parameter(λ)	Cross Validation	Smoothing Parameter(λ)	Generalized Cross Validation
0.1	2,488,4461,886	0.1	25,979,769,304
0.2	19,413,940,841	0.2	13,848,841,143
0.3	14,754,019,218	0.3	10,944,547,189
0.3732638	13,486,984,210	**0.3976385**	9,587,350,701
0.4	13,655,123,515	0.4	9,588,246,968
0.5	16,250,600,726	0.5	10,966,593,031
0.6	18,295,441,919	0.6	13,056,847,746
0.7	18,209,665,159	0.7	14,169,833,532
0.8	16,919,861,769	0.8	14,199,672,451
0.9	15,585,972,047	0.9	13,859,700,611

〈Table 11〉에서 구한 결과를 바탕으로 각각의 평활모수 λ에 해당하는 상대도(C_i)와 지수 값을 구하게 된다. 즉, 평활 모수 수준에 따라 리스크를 대표하는 최적의 함수 곡선이 도출되고 각 연령별로 해당 함수에 해당하는 리스크의 값이 도출된다. 다음 〈Table 12〉, 〈Table 13〉은 각 연령별로 도출된 리스크 상대도 값과 적용된 결과에 따라 계산된 두 가지 지수를 보여준다. 연령 그룹의 수는 평활화된 함수로 적용되는 바, 각 연령이 개별 그룹으로 간주된다.

〈Table 12〉 Age Grouping and Relativity(C_i) of Cross Validation

Age	C_i	Age	C_i	Age	C_i	Age	C_i	Age	C_i
20	0.71979	31	0.94629	42	0.99686	53	0.96608	64	1.14496
21	1.07137	32	0.97381	43	0.98541	54	0.98919	65	1.07821
22	1.26497	33	0.98896	44	0.97634	55	1.01318	66	1.00637
23	1.31451	34	0.98091	45	0.97874	56	1.04730	67	0.99294
24	1.21382	35	0.96404	46	0.98845	57	1.06636	68	1.07085
25	1.12835	36	0.97094	47	0.99658	58	1.06910	69	1.18883
26	1.06579	37	0.98646	48	0.99690	59	1.04344	70	1.22393
27	1.00624	38	0.99373	49	0.99201	60	0.99023	71	1.17523
28	1.00593	39	0.98096	50	0.97621	61	0.94318	72	1.03583
29	0.98627	40	0.96772	51	0.97092	62	1.00609	73	0.83165
30	0.95888	41	0.98010	52	0.96701	63	1.12228		

Number of Group	I_1	I_2
54	0.00368	0.00656

〈Table 13〉 Age Grouping and Relativity of Generalized Cross Validation

Age	C_i	Age	C_i	Age	C_i	Age	C_i	Age	C_i
20	0.73914	31	0.95088	42	0.99328	53	0.96818	64	1.13144
21	1.06140	32	0.97236	43	0.98517	54	0.98964	65	1.07643
22	1.25070	33	0.98497	44	0.97801	55	1.01462	66	1.01491
23	1.30276	34	0.97990	45	0.97998	56	1.04664	67	1.00615
24	1.22050	35	0.96734	46	0.98838	57	1.06487	68	1.07715
25	1.13547	36	0.97260	47	0.99571	58	1.06515	69	1.18117
26	1.06828	37	0.98509	48	0.99616	59	1.03828	70	1.21564
27	1.01138	38	0.99087	49	0.99095	60	0.99182	71	1.16791
28	1.00216	39	0.98112	50	0.97708	61	0.95785	72	1.03367
29	0.98310	40	0.97123	51	0.97066	62	1.01329	73	0.84161
30	0.96024	41	0.98061	52	0.96687	63	1.10996		

Number of Group	I_1	I_2
54	0.00418	0.00551

9. Stopping Rule 방식에 의한 그룹핑 방법

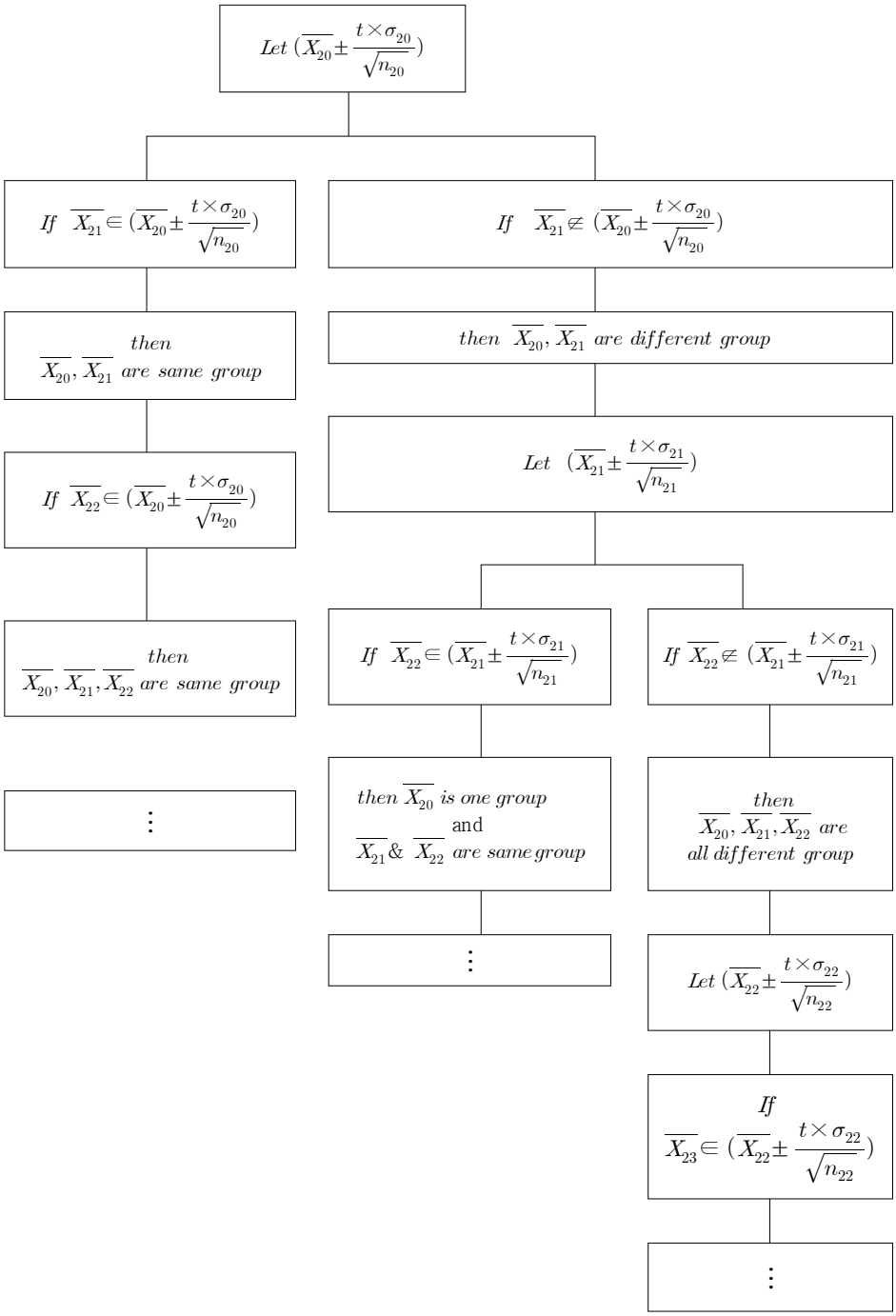
본 절에서는 본 논문이 새롭게 제안하는 연령 그룹핑 방법인 Stopping Rule을 살펴본다. 본 연구에서 제안하는 Stopping Rule의 기준은 식 (4)와 같다.

$$\overline{X_i} \pm \frac{t \times \sigma_i}{\sqrt{n_i}} \equiv (\overline{X_i} - \frac{t \times \sigma_i}{\sqrt{n_i}},\ \overline{X_i} + \frac{t \times \sigma_i}{\sqrt{n_i}}) \tag{4}$$

식 (4)에서 n_i는 i번째 연령에 해당하는 데이터 수를, t는 임의의 양의 실수를 가지는 변수로 그룹핑의 기준이 되는 값으로 이해할 수 있다. Stopping rule의 형태는 신뢰구간과 비슷한 형태이며, 본 논문이 제안하는 Stopping Rule에 따라 어떤 특정 구간에 각 연령의 상대도가 속하게 되면 해당 연령은 같은 그룹으로 포함된다. 상대도가 특정 구간에 속하지 않게 되는 경우, 해당 연령에서 다시 Stopping Rule에 의해 새로운 구간을 만드는 과정을 반복한다. 이해를 돕기 위해서 반복적인 수행 방법을 도표화하면 다음 〈Figure 2〉와 같다.

〈Table 14〉~〈Table 18〉은 본 논문이 제안하는 Stopping Rule을 적용해 추정한 상대도(C_i)와 지수를 t 값의 변화에 따라 반복적으로 산출해 나타난 결과이다.

〈Figure 2〉 Process of Stopping Rule



〈Table 14〉 Age Grouping and Relativity(C_i) of $t = 0.5$

Age	C_i	Age	C_i	Age	C_i
20	0.63058	41	0.96740	58	1.01500
21~22	1.19037	42	1.04753	59	1.13293
23	1.48116	43	0.95680	60	0.99478
24	1.07216	44	0.98545	61	0.86493
25~26	1.13661	45	0.95094	62	0.94186
27	0.87603	46	1.02223	63~64	1.20107
28	1.10115	47~48	0.98675	65	1.05880
29~30	0.97023	49	1.01729	66	1.00092
31	0.90977	50	0.94826	67	0.95586
32	0.94542	51~52	0.98383	68	1.00938
33	1.08494	53	0.92239	69	1.28883
34	0.93051	54	1.03016	70~71	1.18879
35~36	0.96408	55	0.97549	72	1.07501
37~39	0.99494	56	1.06937	73	0.79335
40	0.93898	57	1.07682		
Number of Group		I_1		I_2	
44		0.00007		0.02917	

〈Table 15〉 Age Grouping and Relativity(C_i) of $t = 0.7$

Age	C_i	Age	C_i	Age	C_i
20	0.63058	40	0.93898	56~57	1.07310
21~22	1.19037	41	0.96740	58	1.01500
23	1.48116	42	1.04753	59	1.13293
24~25	1.09892	43	0.95680	60	0.99478
26	1.14753	44	0.98545	61	0.86493
27	0.87603	45	0.95094	62	0.94186
28	1.10115	46	1.02223	63~64	1.20107
29~30	0.97023	47~48	0.98675	65	1.05880
31	0.90977	49	1.01729	66~68	0.98872
32	0.94542	50	0.94826	69~70	1.24292
33	1.08494	51~52	0.98383	71~72	1.12778
34~35	0.94316	53	0.92239	73	0.79335
36~37	0.97851	54	1.03016		
38~39	1.00008	55	0.97549		
Number of Group		I_1		I_2	
40		0.00040		0.03184	

〈Table 16〉 Age Grouping and Relativity(C_i) of $t = 1.0$

Age	C_i	Age	C_i	Age	C_i
20	0.63058	33	1.08494	62	0.94186
21~22	1.19037	34~36	0.95289	63~64	1.20107
23	1.48116	37~55	0.98363	65~66	1.02986
24~26	1.11513	56~58	1.05373	67~68	0.98262
27	0.87603	59	1.13293	69~71	1.22214
28	1.10115	60	0.99478	72	1.07501
29~32	0.94891	61	0.86493	73	0.79335
Number of Group		I_1		I_2	
21		0.00196		0.05623	

〈Table 17〉 Age Grouping and Relativity(C_i) of $t = 1.5$

Age	C_i	Age	C_i	Age	C_i
20	0.63058	33	1.08494	62	0.94186
21~22	1.19037	34~37	0.96084	63~65	1.15364
23	1.48116	38~52	0.98508	66~68	0.98872
24~26	1.11513	53	0.92239	69~72	1.18535
27	0.87603	54~58	1.03337	73	0.79335
28	1.10115	59~60	0.99755		
29~32	0.94891	61	0.86493		
Number of Group		I_1		I_2	
19		0.00473		0.06144	

〈Table 18〉 Age Grouping and Relativity(C_i) of $t = 1.8$

Age	C_i	Age	C_i	Age	C_i
20	0.63058	34~41	0.96874	63~65	1.15364
21~26	1.20121	42~52	0.98815	66~68	0.98872
27	0.87603	53	0.92239	69~72	1.18535
28	1.10115	54~58	1.03337	73	0.79335
29~32	0.94891	59~60	1.06386		
33	1.08494	61~62	0.90339		
Number of Group		I_1		I_2	
16		0.01202		0.06801	

IV. 결론

실증 자료 분석을 통해서 얻은 결과를 그룹핑 방법별로 정리하면 다음의 〈Table 19〉와 같다.

〈Table 19〉 Number of Group and Relativity for Each method

Grouping Method		Number of Group	I_1	I_2
True Value	One Year	54	-	-
Company	Current	15	0.02950	0.05233
Basic	5 Year	11	0.07729	0.00244
	10 Year	6	0.16053	0.00141
Proportion	5%	28	0.02657	0.00282
	10%	11	0.07773	0.00141
Moving Average	3 Windows	56	0.01434	0.01351
	4 Windows	57	0.02922	0.01066
Smoothing Spline	$\lambda = 0.3732638$	54	0.00368	0.00656
	$\lambda = 0.3976385$	54	0.00418	0.00551
Stopping rule	$t = 0.5$	44	0.00007	0.02917
	$t = 0.7$	40	0.00040	0.03184
	$t = 1.0$	21	0.00196	0.05623
	$t = 1.5$	19	0.00473	0.06144
	$t = 1.8$	16	0.01202	0.06801

본 논문은 1세 단위에서 구한 상대도(R_i)를 기준 값으로 설정한 다음 각각의 방법에 대한 상대도(C_i)와 두 가지 지수 값을 구했다. 이는 참값으로 수렴하는 정도의 측도로 활용될 수 있으며, 최적의 방안을 바탕으로 보험회사는 조정 수준의 반영 비율 고려를 통해 정책적인 의사결정을 내릴 수 있다. 실증 자료 분석을 통해 얻어진 결과와 현업에서 활용하는 연령 그룹핑을 통해 도출된 두 지수 값을 비교해보면 다음과 같다.

먼저, 연령구간 세분화 측도를 나타내는 I_1 값의 경우, 현재 특정 보험회사에서 사용 중인 방법의 I_1 값이 0.02950이며, 이는 단순 그룹핑 방식과 비율 방식보다 유사하거나 우수한 수치를 보이나, 이동평균법, 평활 스플라인 방식보다는 높게

나타나므로 비효율적인 그룹핑 방식임을 확인할 수 있다. 또한 본 논문이 제안한 Stopping Rule의 I_1 결과가 오차 한계를 크게 적용한 하나의 방식 이외에서는 모두 작게 시현되는 것을 확인할 수 있었으며, 이는 상대도 산출의 정교함에 있어 Stopping rule 방식의 적용이 보다 효율적이고 세분화의 적정성이 일정 수준 반영됐다고 할 수 있다. 특히, $t = 0.5$ 인 경우, I_1의 값이 제일 작게 시현됐지만 그룹수가 44개로 분류돼, 현업에서 분류해서 사용하는 것에 비해 상당히 증가했다. 따라서 Stopping Rule에서 I_1의 값이 작으면서 그룹수가 현업과 비슷한 경우는 1.8σ을 적용했을 때였으며, 16개의 그룹수를 가진다. 그러나 1.0σ부터 1.8σ까지 그룹수가 비슷하고, I_1값도 현업에서 사용하는 방식의 수치보다 작게 시현되므로 어떤 값을 선택해도 적절한 대안이 될 수 있다.

다음으로 연령 변동 수준을 반영하는 I_2 값의 경우, 전체적으로 결과들이 현업에서 사용하는 방법보다 작게 나왔음을 확인할 수 있다. 현업에서 쓰이는 방법의 지수보다 작을 경우 만족하는 결과를 얻었다 고 할 수 있으며, Stopping Rule의 경우에는 현업보다 크게 나온 값들도 존재하나 그 차이가 크지 않기 때문에 고객이 느끼는 체험적인 측면에서는 만족스러운 결과라 할 수 있다.

종합적으로 볼 때, 본 연구가 제시한 Stopping Rule이 다른 연령 그룹핑 방법보다 우수한 결과가 도출, 적용 방식의 적정성이 증명됐다. 마지막으로 회사에서는 두 가지 요소를 동시에 고려를 해야 하는 상황이기 때문에 가중치를 고려한 종합적 지수를 설정해 상황에 맞도록 운영할 수 있으며, 이를 수식으로 표현하면 다음과 같다.

$$I = \omega \times I_1 + (1 - \omega) \times I_2$$

따라서 현업에서는 제시된 수식의 가중치(w)를 조정해 고객의 불만을 최소화하면서 동시에 합리적인 가격을 제시할 수 있는 적정 수준을 고려할 수 있으며, 회사별 특성을 고려해 다양한 방식이 적용 가능할 것으로 기대된다. 다만 1세 단위의 연령 분포를 국가통계 비율로 고려한 바, 각 회사가 보유한 데이터를 1세 단위로 파악하지 못하는 한계점이 있음을 밝히는 바이다.

또한 σ의 계수인 t 값을 그룹별로 상이하게 적용할 수 있는 일반화된 기준에 대한 연구와 리스크 산출에 사용한 산술 평균 이외의 통계량에 대한 활용 가능성, 즉 최적해를 추정하는 방식에 대한 연구를 추후 과제로 제안한다.

참고문헌

김명준, "적정 보험료 수준 예측을 위한 심도 빈도의 추세 분석에 관한 연구", **계리학연구**, 제5권 제2호, 2013.

(Translated in English) Myung Joon Kim, "A Study on Trend Analysis of Severity and Frequency for Predicting the Proper Premium", *The Journal of Actuarial Science*, Vol. 5(2), 2013.

김영화·김명준, "다양한 모형화를 통한 자동차보험가격 산출", **한국데이터정보과학회지**, 제20권 제3호, 2009.

(Translated in English) Y. Kim and M. Kim, "Various modeling approaches in auto insurance pricing", *Journal of the Korean Data & Information Science Society*, Vol. 20(3), 2013.

김영화·이현수, "신뢰도 적용방법에 따른 자동차보험 가격산출", **Communications for Statistical Applications and Methods**, 제17권 제5호, 2010.

(Translated in English) Y. Kim and H. Lee, "A Comparison Study for the Pricing of Automobile Insurance Premium Based on Credibility", *Communications for Statistical Applications and Methods*, Vol. 17(5), 2010.

김충락·강근석, **회귀분석**, 제2판, 교우사, 2010.

(Translated in English) C. Kim and K. Kang, *Regression Analysis, 2nd ed.*, Kyowoo-sa, 2010.

이우동·강상길·윤용화·김종태, "단순 스무딩 스플라인 함수 추정", **기초과학**, 제2권 제1호, 1998.

(Translated in English) W. Lee, S. Kang, Y. Youn and J. Kim, "An Estimation of Simple Smoothing Spline Function", *Journal of Basic Science*, Vol. 2(1), 1998.

이창수, "신뢰도 기법을 이용한 자료의 충분성 평가와 보험요율의 조정", **보험개발연구**, 제21호, 1997.

(Translated in English) C. Lee, "Data sufficiency evaluation and rating adjustment using credibility method", *Journal of Insurance and Finance,* Vol. 21, 1997.

함상호, "자동차보험 가격자유화 도입방향과 손해보험회사의 경영전략에 대한 고찰", **보험개발연구**, 제24호, 1998.

(Translated in English) S. Ham, "A Study on the Auto Insurance Price Liberalization and the Management Strategy of the Property Insurance Company", *Journal of Insurance and Finance*, Vol. 24, 1998.

최우석·한상일, "이중일반선형모형(DGLM)을 이용한 자동차 보험요율 추정", **보험개발연구**, 제19호, 2008.

(Translated in English) W. Choi and S. Han, "Estimating the Rate of Motor Insurance Premium by Double Generalized Linear Model", *Journal of Insurance and Finance,* Vol. 19, 1998.

Jorgensen, B. and Paesde Souza, "Fitting Tweedie's compound model to insurance claims data", *Scandinavian Actuarial Journal*, Vol. 1, 1994, pp.69-93

Murphy, K.P., Brockman, M.J. and Lee,P.K.W., "Using generalized linear models to build dynamic pricing systems", *Casualty Actuarial Forum, Winter*, 2000

Viktor Grgic, "Smoothing splines in non-life insurance pricing", *Mathematical Statistics Stockholm University, Examensarbete*, 2008:3

Y. Kim and M. Kim, "Constrained Bayes and Empirical Bayes Estimator Applications in Insurance Pricing", *Communications for Statistical Applications and Methods*, Vol. 4, 2013, pp.321-327.

Abstract

For the insurance pricing, variable selection which has different risk pattern and grouping for risk estimation should be considered first. Since a variable, such as gender, is an obvious classification measure, it is not appropriate for considering one of grouping criteria. However, an age variable has a wide range and the criteria for its grouping is ambiguous. Considering each age for risk estimation makes a credibility issue due to the number of customer in each cell. Moreover, the age variable has its unique characteristics that is depending on time and the order for the grouping should be reflected.

In this research, the most effective way for the age variable grouping is proposed by considering the variable characteristics. More precisely, various grouping methods currently applied and new method 'The Stopping Rule' will be introduced. Using real insurance data, analysis results are given to compare the performance and also be shown that the properness and effectiveness of the proposed grouping method, 'The Stopping Rule'.

※ Key words: auto insurance, risk grouping, stopping rule, true cost