이화여자대학교 대학원
2008학년도
석사학위 청구논문

생존분석법을 이용한 생명보험 해약 연구

統 計 學 科
朴 英 蘭

2009

생존분석법을 이용한 생명보험 해약연구

이 論文을 碩士 學位 論文으로 提出함

2008년 12월

梨花女子大學校 大學院

統計學科 朴 英 蘭

朴英蘭의 碩士 學位 論文을 認准함

指導教授 蘇秉秀 _____

審査委員 吳滿淑 _____

車智煥 _____

蘇秉秀 _____

梨花女子大學校 大學院

목 차

표 목 차

그 림 목 차

<논문개요>

본 논문은 생존분석 기법인 Cox의 비례위험모형(Cox Proportional Hazards Model)과 가속화 고장시간 모형(Accelerated Failure Time Model)을 통하여 생명보험 고객이 상품을 계약한 시점부터 해약 시점까지 걸리는 시간의 분포를 찾아보고, 로지스틱 회귀 모형과 그 예측 성능을 비교한 논문이다. 최근 보험사의 CRM(Customer Relationship Management)측면과 RM(Risk Management)측면에서 해약을 예측하는 연구가 많이 이루어지고 있으나 기존의 이분류 예측기법은 해약 혹은 유지라는 종속변수의 이분류 예측에 머무르며, 해약할 때까지의 시간인 $\triangle t$가 고정되어있어 정보제공에 한계가 있다.

따라서 본 논문에서는 Cox의 비례위험 모형을 통하여 생명보험 해약에 영향을 끼치는 변수들을 찾아보고 가속화 고장시간 모형을 통하여 Hazard Function의 최적 기저분포를 확인하였다. 또한 그 예측력 평가를 통하여 세 모형의 성능을 비교해 보았다. 분석에 사용된 설명변수가 제한적이었고, 해약고객의 비율이 3~4% 정도밖에 되지 않아 예측률이 그다지 높지 못한 한계가 있으나 이분류 예측을 보완하여 해약율의 분포를 제시했다는 점에서 본 논문의 의의를 찾을 수 있다.

Ⅰ. 서론

2003년 8월 방카슈랑스의 도입으로 금융산업의 겸업화가 진전되고 종합금융 서비스 제공을 위한 합병 제휴 등이 활발히 진행되면서 보험산업은 이제 보험회사간의 경쟁일 뿐 아니라 다른 금융업과의 경쟁으로 확대되어 더욱 치열한 경쟁의 시대에 들어서게 되었다. 금융겸업화는 금융시장의 큰 흐름으로 겸업화 시장에서 여러 금융권역은 서로 치열한 경쟁을 해야만 하는 상황이다. 은행, 증권 등 타 금융권에서는 겸업화에 부응한 다양한 경영전략을 펼치고 있다. 은행은 경쟁력 확보를 위해 금융지주회사를 중심으로 여러 금융권역의 회사를 소유하는 조직 확대 전략을 실행하고 있고 증권계열에서는 2009년 2월 자본시장통합법 발효를 앞두고 향후 겸업화를 대비하여 조직규모 확대 및 다양한 상품개발을 준비 중이다.[1] 이와 같이 새로운 보험시장 환경이 전개되면서 국내 생명보험 회사들은 내실 위주의 경영체제로 돌입하지 않으면 안되며 생명보험 산업의 질적 수준을 평가할 수 있는 방법의 하나인 상품계약 유지율 개선에 만전을 기해야 한다.

따라서 본 논문에서는 생명보험 해약 예측모형 구축을 위한 로지스틱 회귀모형과 생존분석 기법 중 Cox의 비례위험모형(Cox Proportional Hazards Model)과 가속화 고장시간모형(Accelerated Failure Time Model)에 대한 소개와 함께 실증분석을 위해 국내 생명보험사의 고객 데이터를 사용해 분석법을 적용해 보고자 한다. 이를 통해 생명보험 산업의 질적 수준 개선과 경쟁력 제고에 기여 하고자 한다. 본 논문의 구성은 다음과 같다. 1장 서론에서는 연구의 배경과 목적에 대해 기술하였고, 제2장에서는 선행연구와 로지스틱 회귀분석, Cox의 비례위험모형, 가속화 고장시간모형에 대한 설명을 기술하였다. 제3장 실증분석에서는 앞서 설명한 방법론을 실제 자료에 적용하여 모형의 추정 및 예측을 수행하였고, 제4장 요약 및 결론에서는 연구모형에 대한 결과를 요약하고 연구의 한계점 및 향후 연구 방향에 대해 기술하였다.

1) 안철경,기승도, 「금융 겸업화에 대응한 보험회사의 채널전략」, 보험개발원 보험연구소, 2008

Ⅱ. 생명보험 해약 예측 모형의 이론적 배경

1. 선행 연구

생명보험 해약 예측 모형은 Black 과 Skipper(1994)에 의해 생명보험의 실효·해약 정도를 통하여 계약자들이 느끼는 보험상품의 질 및 서비스 수준에 대한 만족도를 평가할 수 있다는 연구가 촉매제가 되어 1990년 이후 보험회사의 높은 실효·해약 문제에 대한 해결책을 제시하는 연구가 많이 진행되어 왔다.

특히 이현우, 강중철(1999)은 생존분석 기법을 처음으로 생명보험 해약 연구에 활용하였다. 대수선형회귀모형을 통한 분석으로 생명보험 실효·해약 관련 요소들이 무엇인지 찾고, 실효·해약에 어떠한 영향을 주는지에 대한 분석에 초점을 맞추었다.

이호영(2005)은 생명보험 실효·해약 예측 모형 평가에 중점을 둔 연구로서 로지스틱 회귀모형과 의사결정나무, 인공신경망의 예측력을 평가하고 분석하였다. 이를 통해 어떤 한 가지 방법론을 통하여 예측한 모형보다 컴바인드 모형을 활용하였을 때 예측력이 향상되는 것을 보여주었다.

김수나(2008)는 로지스틱 회귀모형에서 시간변수를 추가하였을 때와 그렇지 않았을 때 각각 Cox의 비례위험모형과 예측력을 비교하여 더 나은 예측 모형을 제시하고자 하였다. 시간변수를 추가하였을 때 Cox의 비례위험모형과 거의 동일한 효과를 낼 수 있다는 것을 보여주었고 Cox의 비례위험모형을 활용하여 생명보험 해약 연구를 시도하였다는 점에서 의의를 찾을 수 있다.

2. 로지스틱 회귀분석

로지스틱 회귀분석은 중회귀분석, 판별분석과 함께 이진형의 종속 변수를 예측하기 위한 대표적인 다변량분석 통계기법 중 하나이다. 종속변수의 범위가 $(-\infty,\infty)$를 가지고 있는 일반 선형회귀분석과 달리 로지스틱 회귀분석의 경우 종속변수가 0또는 1의 값만을 갖게된다. 이를 일반 선형회귀모형으로 나타내면,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad Y_i = 0, 1 \quad (1)$$

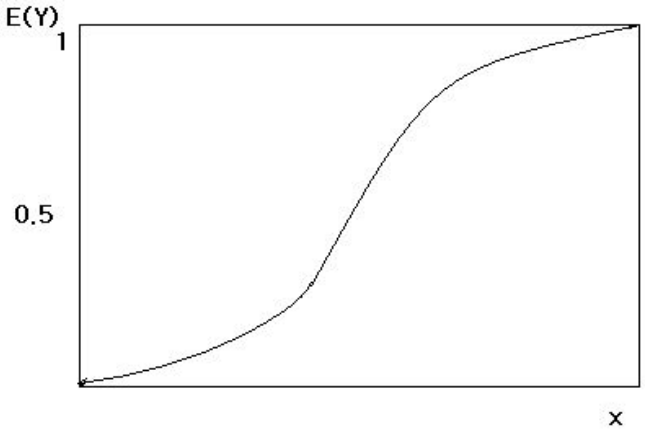
라는 식을 얻을 수 있고, 생명보험을 해약하는 경우를 $Y=1$, 유지하는 경우를 $Y=0$으로 할 때, 다음과 같은 확률분포를 가지게 된다.

Y	Probability
1	$p(Y_i=1)=\ \pi_i$
0	$p(Y_i=0)=\ 1-\pi_i$

위 확률분포를 통해 다음과 같은 식을 얻을 수 있다.

$$E(Y_i)=1\cdot\pi_i+0\cdot(1-\pi_i)=\pi_i \quad (2)$$

종속변수가 이진형인 로지스틱 회귀분석의 경우 반응함수의 그래프가 S자 곡선 형태를 보이고 감소함수의 경우에는 역S자 곡선의 형태를 띤다.



<그림1> 로지스틱 반응곡선

이는 독립변수의 값이 증가함에 따라 종속변수의 값이 0에서부터 천천히 증가하다가 일정 수준에서 급격히 증가하고 다시 종속변수의 값이 1에 가까워지면 서서히 1에 수렴하는 형태를 띠게 되고 이런 함수를 로지스틱 반응함수(logistic response function)라 부른다. 이를 모형화 하면 다음과 같다.

$$E(Y)=\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}=[1+\exp(-\beta_0-\beta_1 x)]^{-1} \quad (3)$$

(2)식과 (3)식을 통하여

$$E(Y)=\pi=\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)} \qquad (4)$$

$$\Rightarrow \quad \pi+\pi \bullet \exp(\beta_0+\beta_1 x)=\exp(\beta_0+\beta_1 x)$$

$$\Rightarrow \quad \pi=(1-\pi) \bullet \exp(\beta_0+\beta_1 x)$$

$$\Rightarrow \quad \exp(\beta_0+\beta_1 x)=\frac{\pi}{(1-\pi)}$$

$$\Rightarrow \quad \beta_0+\beta_1 x=\ln(\frac{\pi}{1-\pi})$$

라는 식을 얻을 수 있다. 마지막 식을 로짓반응함수(logit response function)라고 부르며 종속변수가 $(-\infty,\infty)$의 범위를 가지게 된다. 로지스틱 함수의 계수는 우도함수(likelihood function)를 최대화 하는 β_0,β_1을 추정함으로 구할 수 있다.

3. 생존분석-Cox의 비례위험모형

생존분석기법은 생의학 분야에서 시간에 관련된 변수를 설명하기 위하여 개발되었으며 생존모형(Survival model)은 하나의 사상이 종료되기까지의 시간(예를 들면, 보험계약이 해약되기까지의 시간, 사람의 수명, 제품이 고장날 때까지 걸린 시간)인 확률변수 t의 분포를 규정한다. 발전 초기의 생존분석법은 표본 내의 모든 개체가 동일한 사건발생 확률을 가지고 있다는 동질성(同質性)의 가정 이라는 결정적인 한계를 가지고 있었다. 가령 남녀의 성별에 따라 거주지역에 따라 직업에 따라 사건발생 확률이 다를 수 있음에도 생존함수에 영향을 미치는 여러 요인을 평가하기가 어려웠다. 1972년에 이르러 영국의 통계학자 D.R. Cox는 다변수적 생존분석인 이른바 비례위험모형(Proportional Hazards Model)을 개발, 생존함수에 미치는 요인의 효과를 수량적으로 측정하여 생존분석법에 획기적인 기여를 하였다.

계약유지시간을 t라 하고 x를 이 계약유지기간에 영향을 주는 공변량 값이라 할 때, Cox의 비례위험모형(proportional hazards model)은

$$h(t;x_i)=\exp(\beta x_i)h(t;0) \qquad (5)$$

로 표현한다. Base hazard function인 $h(t;0)$는 공변량의 값이 모두 0인 고객의

hazard function 이다. 일반적으로 생존분포에 영향을 미치는 p개의 공변량 $x_1,\cdots x_p$ 을 고려할 때 Cox의 비례위험모형(proportional hazards model)은,

$$h(t;x_1,\cdots,x_p)=\exp(\beta_1x_1+\cdots+\beta_px_p)h(t;0,\cdots,0) \qquad (6)$$

로 풀어 쓸 수 있다. 위 식의 양변에 로그를 취하면,

$$\log h_i(t)=\alpha(t)+\beta_1x_{i1}+\cdots+\beta_kx_{ik} \quad , \ \alpha(t)=\log h(t;0) \quad (7)$$

이 된다. 위험함수는 $\alpha(t)=\alpha$일 때 위험함수 형태가 시간에 관계없이 일정한 지수분포 모형을, $\alpha(t)=\alpha t$일 때 곰페르츠 모형을, $\alpha(t)=\alpha\log t$일 때 위험함수 형태가 단조증가 혹은 단조감소하는 와이불분포 모형을 갖는다. 이들 분포와 함께 대표적인 위험함수 형태로 위험함수 형태가 증가하다가 감소하는 로그-정규분포가 있다.

Cox모형의 기초가 되는 위험함수(hazard function) h(t)는 t시점까지 가입을 유지한 고객이 t시점 바로 직후 순간적으로 해약할 조건부확률로 순간위험률이라고도 하며 다음과 같이 표현한다.

$$h(t)=\lim_{\Delta t\to 0}\frac{P(t\le T<t+\Delta t|T\ge t)}{\Delta t}=\frac{1}{P(T\ge t)}\lim_{\Delta t\to 0}\frac{P(t\le T<t+\Delta t)}{\Delta t}=\frac{f(t)}{S(t)} \quad (8)$$

여기서 S(t)는 생존함수로 t시점까지 해약이 일어나지 않고 고객이 계약을 유지할 확률을 의미한다. 고객의 계약 유지기간 T의 누적확률분포 F(t)가 $\int_0^t f(u)du$라 할 때, 생존함수 S(t)는 다음과 같이 나타낼 수 있다.

$$S(t)=p(T\ge t)=1-F(t) \qquad (9)$$

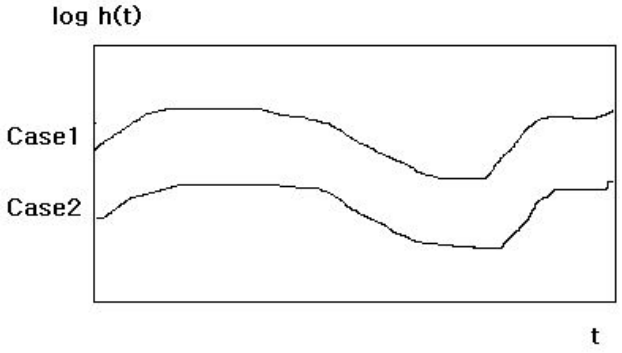
생존함수 S(t)와 위험함수 h(t)의 관계를 이용하여 누적위험함수 H(t)를 구해보면 다음과 같다.

$$H(t)=\int_0^t h(u)du=\int_0^t\frac{f(u)}{[1-F(u)]}du=-\log[1-F(t)]=-\log S(t) \quad (10)$$

Cox의 생존분석 모형을 비례위험 모형이라고 부르는 이유는 한 관측과 또 다른 관측과의 위험률 비가 고정되어 있기 때문이다. 임의의 고객 i와 j의 위험률 비를 나타내 보면 다음과 같다.

$$\frac{h_i(t)}{h_j(t)} = \exp\{\beta_1(x_{i1} - x_{j1}) + \cdots + \beta_k(x_{ik} - x_{jk})\} \quad (11)$$

위와 같이 기저위험 함수인 $h(t;0)$은 상쇄되어 사라져서 두 고객간 위험률 비는 시간에 관계없는 값을 갖게 된다.



<그림2> 비례위험모형에서 평행한 위험률 함수

따라서 Cox 모형에서는 부분우도함수(partial likelihood function)를 이용하여 부분우도를 최대로 하는 β값을 구할 수 있다. 전체우도함수(entire likelihood function)를 사용할 때 보다는 큰 표준오차를 갖지만, Efficiency의 손실이 매우 작으므로 기저위험 함수의 모양에 관계없이 고객의 해약확률의 순위를 구할 수 있다.

4. 생존분석-가속화 고장시간모형

가속화 고장시간 모형은 Cox의 비례위험 모형에서 위험함수의 비가 일정함을 가정한 것과 다르게 생존시간 간의 비례관계를 로그-선형(log-linear)의 형태로 모형화 한다.

$$\log T = \alpha + \beta x + \epsilon \quad (12)$$

모형에서 설명 변수인 공변량 값이 0인 경우의 계약유지기간(생존시간)에 대한 확률

변수를 T_0라 하고, 이는 사전적으로 어떤 기저분포(baseline distribution)를 갖는 것으로 가정한다. 흔히 쓰이는 기저분포로는 지수(exponential)분포, 로그-정규(log-normal)분포, 와이블(weibull)분포, 감마(gamma)분포 등이 있으며, 사후적으로 이렇게 가정된 분포의 적합성에 대한 가설 검정이 가능하다. 공변량의 값이 x인 경우에서의 계약유지기간(생존시간)을 T라하고 기저 계약유지기간(생존시간)인 T_0에 다음과 같이 비례한다고 가정하자.

$$T = \exp(\beta x) \bullet T_0 \quad (13)$$

이는 계약유지기간(생존시간) T가 x값에 따라 지수적으로 변화한다는 것을 의미한다. 양변에 로그를 취하면,

$$\log T = \beta x + \log(T_0) = \mu + \beta x + \log(T_0) - \mu \quad (14)$$

와 같은 식이 되고, $a = \mu$, $\epsilon = \log(T_0) - \mu$ 라 놓으면 (1)식의 형태가 된다. 여기서 μ는 $\log(T_0)$의 확률분포에서 위치모수를 나타낸다.

위 식은 공변량이 하나인 경우이나 만일 생존시간에 영향을 주는 공변량이 p개 존재하고, 척도모수(scale parameter,σ)가 포함되면 일반적인 AFT모형이 된다.

$$T = \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p) \bullet T_0^{\sigma} \quad (15)$$

위 식의 양변에 로그를 취하면 다음과 같은 식이 된다.

$$\log T = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \sigma\epsilon \quad (16)$$

Survival function은 Cox의 비례위험 모형에서도 살펴보았듯이 $S(t) = p(T \geq t) = 1 - F(t)$이고 이를 AFT 모형에서 살펴보면 다음과 같다.

$$S(t|x_1, \cdots, x_p) = S(\ln t - (\alpha + \sum_{j=1}^{p} \beta_j x_j)/\sigma) \quad (17)$$

기저분포가 Lognormal 분포의 경우 분포함수가 $F(t) = \Phi\{(\ln t - \mu)/\sigma\}$ 이므로 생존함수(Survival function)은

$$S(t) = 1 - \Phi[(\ln t - \alpha - \sum_{j=1}^{p} \beta_j x_j)/\sigma] \quad (18)$$

이고, 이 때 보험가입 후 t시점에 보험계약을 유지하고 있는 고객$(x_1, \cdots, x_p)$이 향후 $\triangle t$기간 이내에 보험을 해지할 확률은

$$\begin{aligned} p(t; x_1, \cdots, x_p) &= p(t < T < t + \triangle t | T > t; x_1, \cdots, x_p) \quad (19) \\ &= 1 - \frac{S(t + \triangle t | x_1, \cdots, x_p)}{S(t | x_1, \cdots, x_p)} \\ &= h(t | x_1, \cdots, x_p) \triangle t \end{aligned}$$

이므로, lognormal 분포의 경우 다음과 같은 식으로 나타낼 수 있다.

$$p(t; x_1, \cdots, x_p) = 1 - \frac{1 - \Phi[\{\ln(t + \triangle t) - \mu_i\}/\sigma]}{1 - \Phi[(\ln t - \mu_i)/\sigma]} \quad (20)$$

본 논문에서는 분포에 대한 적합도 검정을 고려하여 기저 분포를 가정하였으며, 이에 따른 AFT모형을 추정하여 Cox의 비례위험모형과 로지스틱 회귀 모형과 예측결과를 비교하였다. AFT모형은 생존시간 자체에 대한 설명변수의 효과를 모형화 하기 때문에 Cox의 비례위험 모형보다 예측 목적에 적합하다.

Ⅲ. 실증분석

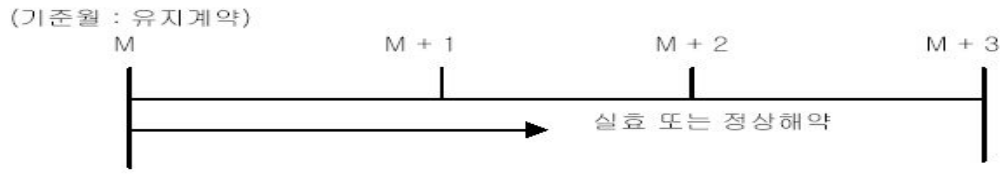
1.자료의 구성

앞서 소개한 방법론들을 실제 자료들을 이용하여 검증해 보았다. 실증분석에 사용된 자료는 국내 생명보험회사의 데이터로 2001년 6월과 2001년 9월에 표본으로 선정된 20000명의 고객의 개인정보 데이터이다.

자료의 분석과 검증을 위해 데이터를 분석용과 평가용으로 구분하였다. 데이터 분할 시 분석용과 평가용을 분류하는 방법에는 다양한 방법이 있지만 본 논문에서는 향후 예측력을 알아본다는 의미에서 기간으로 구분하여 사용하였다. 따라서 2001년 6월 데이터로 고객의 계약유지 기간을 분석 한 후 2001년 9월 고객 데이터로 예측력을 평가하는 방법을 사용하였다.

분석용 데이터는 2001년 6월 1일 가입유지 상태에 있는 고객으로 3개월 이내 (즉 2001년 9월 1일 이내)에 실효 또는 해약한 고객이 373명, 3개월 후에도 계약을 정상적으로 유지하고 있는 고객이 9627명으로 구성된 데이터이다. 373명 중 1명은 보험만기일을 3개월 이내로 남겨두고 해약한 것으로 나타나 전산 상 입력 착오로 간주하고 "가입유지"상태로 변경하여 실효 또는 해약 고객이 372명 가입 유지 고객이 9628명이 되었다.

평가용으로 사용된 데이터는 2001년 9월 1일 가입유지 상태에 있는 고객으로 3개월 이내 (즉 2001년 12월 1일 이내)에 실효 또는 해약한 고객이 413명, 가입을 정상적으로 유지한 고객이 9587명으로 구성된 데이터이다.



<그림3> 실효 또는 해약의 정의

<표1> 데이터의 구성

분석용 (2001.06)	가입 유지	실효 또는 해약	합계
	9628	372	10000
평가용 (2001.09)	가입 유지	실효 또는 해약	합계
	9587	413	10000

2. 종속변수 및 독립변수

분석을 위해 본 논문에서 사용한 변수는 크게 고객의 3개월 이내 해약여부 변수, 고객이 계약한 보험 상품 관련 변수, 계약자 관련 변수, 피보험자 관련변수, 고객의 계약유지기간(t) 변수가 있다.

고객의 해약여부 변수는 2001년 6월 1일부터 2001년 9월 1일까지 실효 또는 해약한 고객의 경우 '1'의 값을 가지고 계약을 유지한 고객은 '0'의 값을 가져 로지스틱 회귀분석의 종속변수로, Cox의 비례위험모형, 가속화 고장시간모형에서 절단여부를 지시해 주는 변수로 사용된다.

보험상품 관련 변수는 고객의 해약 또는 계약유지의 대상이 되는 상품에 관련된 변수이다. 월납/3월납/6월납/년납 으로 구분된 '납입방법'에 관한 변수, 보험료 수금 방법을 나타내는 방문/자동이체/지로/카드납 으로 구분된 '수금방법' 변수, 일회 납입 보험료를 납입 방법에 따라 월납 보험료로 계산해 준 '한달납입보험료'변수, 주계약과 함께 가입 가능한 부가적인 보험 상품인 특약의 가입유무를 보여주는 '특약유무'변수가 있다. 약관대출이란 계약자가 가입한 보험 해약환급금의 70~80%의 범위에서 수시로 대출 받을 수 있는 제도이다. 이와 관련된 '약관대출유무' 변수는 약관대출 경험이 있는 고객은 값이 1 없는 고객은 값이 0인 변수이다. '약관대출상환여부' 변수는 약관대출을 미상환 한 고객이 1의 값을 갖고 상환하여 대출상태가 아닌 고객이 0의 값을 갖는다. '약관대출잔고'는 약관대출 미상환 고객의 남아있는 대출 잔고를 말해주는 변수이다. 보험료를 2개월 째 미납하고 3개월에 접어들면 실효 상태가 되는데 실효일로부터 2년 이내(만기이전) 계약의 효력을 다시 발생시키는 것을 부활이라 한다. 부활 경험이 있는 고객이 1의 값을 갖고 부활 경험이 없는 고객이 0의 값을 갖는 '부활'변수가 있다. '보험금지급만기유무' 변수는 보험금지급 만기가 있는 고객은 0, 만기가 없는 고객은 1의 값을 갖는 변수이다. '가입경로' 변수는 개인영업에 의해 가입한 고객은 1의 값을 Tele-Marketing에 의해 가입한 고객은 0의 값을 갖는다. 상품군을 5가지로 나누어서 분류한 변수가 '상품중분류', 상품군을 9가지로 분류한 변수가 '상품소분류'이다.

계약자 관련 변수는 보험상품을 계약 당사자로서 보험료 납입의 의무를 가진 계약자의 정보를 가지고 있다. 계약자가 가입 당시 몇 세 이었는지를 나타내는 '계약자가입연령'변수, 계약자가 기혼자인지 미혼자인지를 나타내는 '계약자결혼유무'변수, '계약자성별'변수, 계약자의 직업군을 나누어 그룹화한 '계약자직업'변수로 구성되어 있다.

피보험자는 보험사고 발생의 객체가 되는 사람을 말하며, '피보험자의 가입당시 연령'과 '피보험자 성별' 변수로 구성되어 있다.

생존분석의 종속변수로 사용된 고객의 계약유지기간(t)를 구하기 위해서 분석용 데이

터와 평가용 데이터 모두 최종납입횟수를 계산하였다. 납입 방법에 따라 월납의 경우 원본데이터의 최종납입횟수를 그대로 사용하였고, 그 외의 경우는 납입방법에 따라 3월납의 경우 (최종납입횟수)*3을 하였고, 6월납의 경우 (최종납입횟수)*6, 연납의 경우 (최종납입횟수)*12를 하여 계약유지기간의 단위를 월단위로 수정하여 이를 종속변수로 사용하였다.

<표2> 선별된 생명보험 관련 요소

NO	구분	변수명	구분방식
1	보험상품관련	납입방법	월납/3월납/6월납/년납
2		수금방법	방문/자동이체/지로/카드납
3		한달납입보험료	(단위:원)
4		특약유무	유/무
5		약관대출유무	유/무
6		약관대출상환여부	상환/미상환
7		약관대출잔고	(단위:만원)
8		부활유무	유/무
9		보험금지급만기유무	유/무
10		가입경로	개인영업/TM
11		상품중분류	상품1/상품2/상품3/상품4/상품5
12		상품소분류	상품1/상품2/상품3/상품4/상품5/ 상품6/상품7/상품8/상품9
13	계약자관련	계약자가입연령	(단위:세)
14		계약자결혼유무	유/무
15		계약자성별	남/여
16		계약자직업	직업1/직업2/직업3/직업4/ 직업5/직업6/직업7/직업8/직업9
17	피보험자관련	피보험자가입연령	(단위:세)
18		피보험자성별	남/여

분석용 자료와 평가용 자료의 실효 · 해약관련 요소별 빈도를 살펴봄을 통하여 데이터의 구성에 대해 알아보았다. 이를 통해 데이터의 구성에 이상이 있는 변수들을 제거하였는데 원래 데이터에서 '계약자주거형태'변수와 '계약자자녀수'변수가 분류상 알 수 없는 군에 해당하는 데이터가 62%가 넘으므로 삭제되었다. 또한 수금방법에서 직납에 해당하는 고객이 0명, 상품중분류에서 0번에 해당하는 고객이 0명, 상품소분류에

서 0번에 해당하는 고객이 0명이므로 이를 삭제한 후 더미변수를 생성하였다. 이러한 수정을 거치고 난 후의 변수들의 분포는 표<3> - 표<6>와 같다.

분석용 자료의 정성변수에서 납입방법변수의 경우 월납이 차지하는 비중이 97.79%로 압도적으로 많은 부분을 차지하고 있었고, 수금 방법에 있어서 자동이체가 가장 많은 비중을 카드납이 그 다음을 차지하고 있었다. 특약을 신청한 고객이 82.18%로 신청하지 않은 고객보다 많았고, 약관대출을 하지 않은 고객이 약관대출을 한 고객보다 많았다. 대출을 상환하지 않은 고객은 전체의 4.63%에 해당하였고, 부활경험이 있는 고객은 6.57%에 해당하였다. 만기가 없는 고객이 6.25%에 해당하여 만기가 있는 고객 보다 적은 수치였고, 개인영업에 의한 가입자가, 계약자 중 미혼자가 눈에 띄게 많은 비중을 차지하고 있었다. 그 이외의 계약자와 피보험자 성별, 상품 중분류, 상품 소분류, 계약자 직업에 해당하는 변수들은 비교적 고르게 분포함을 알 수 있다.

<표3> 분석용 자료의 실효 · 해약관련 요소별 빈도분포_정성변수

(단위 : 건, %)

요소		관측치수	비중
납입방법	월납	9779	97.79
	3월납	49	0.49
	6월납	25	0.25
	년납	147	1.47
수금방법	방문	106	1.06
	자동이체	9293	92.93
	지로	72	0.72
	카드납	529	5.29
특약유무	특약무	1782	17.82
	특약유	8218	82.18
약관대출유무	약관대출무	9401	94.01
	약관대출유	599	5.99
약관대출상환여부	상환	9537	95.37
	미상환	463	4.63
부활유무	부활무	9343	93.43
	부활유	657	6.57
보험지급만기유무	만기무	625	6.25
	만기유	9375	93.75
가입경로	개인영업	8743	87.43
	TM	1257	12.57
상품중분류	상품1	4050	40.50
	상품2	3760	37.60
	상품3	200	2.00

	상품4	1593	15.93
	상품5	397	3.97
상품소분류	상품1	2339	23.39
	상품2	773	7.73
	상품3	3760	37.60
	상품4	200	2.00
	상품5	397	3.97
	상품6	60	0.60
	상품7	1048	10.48
	상품8	153	1.53
	상품9	1270	12.70
계약자결혼유무	결혼무	9813	98.13
	결혼유	187	1.87
계약자성별	남	4317	43.17
	여	5683	56.83
계약자직업	직업1	1602	16.02
	직업2	717	7.17
	직업3	1703	17.03
	직업4	2680	26.80
	직업5	151	1.51
	직업6	12	0.12
	직업7	206	2.06
	직업8	473	4.73
	직업9	2456	24.56
피보험자성별	남	5055	50.55
	여	4945	49.45

분석용 자료의 정량변수에서 한달 납입 보험료는 816.667원이 가장 적었는데 이는 3월납, 6월납, 연납의 고객의 일회 납입 금액이 월납 보험료로 수정된 것을 알 수 있었다. 최대 납입 보험료는 210만원 이었다. 약관 대출 잔고 변수에서 중위수가 0이고 평균값은 8만5천원 인 것으로 보아 0에 해당하는 고객이 상당히 많음에도 약관대출 중인 고객의 대출 금액이 크다는 것을 알 수 있었다. 계약자와 피보험자 가입연령은 평균값에 많은 고객이 몰려있고 좌.우로 치우치지 않은 분포를 하고 있음을 예상할 수 있다.

<표4> 분석용 자료의 실효 · 해약관련 요소별 빈도분포_정량변수

(단위: 원, 만원, 세)

요소	최소값	최대값	중위수	평균값	관측치수
한달납입보험료*	816.667	2100000	30000	49408	10000
약관대출잔고*	0	3000	0	8.508	10000
계약자가입연령*	3	85	37	37.8881	10000
피보험자가입연령*	0	89	34	32.8371	10000

*표시된 변수는 값의 편차가 크므로 로그를 취한 변수를 추가하여 분석하였다.

평가용 자료는 모형적합을 통해 선별된 변수들만을 따로 뽑아 빈도를 살펴보았다. 그 결과 아래와 같이 분석용 자료와 그 추이가 비슷함을 알 수 있었다.

<표5> 평가용 자료의 선택된 변수의 빈도분포_정성변수

(단위 : 건, %)

요소		관측치수	비중
수금방법	방문	105	1.05
	자동이체	9322	93.22
	지로	57	0.57
	카드납	516	5.16
약관대출유무	약관대출무	9461	94.61
	약관대출유	539	5.39
보험지급만기유무	만기무	717	7.17
	만기유	9283	92.83
상품중분류	상품1	4219	42.19
	상품2	3538	35.38
	상품3	356	3.56
	상품4	1560	15.60
	상품5	327	3.27
상품소분류	상품1	2549	25.49
	상품2	740	7.40
	상품3	3538	35.38
	상품4	356	3.56
	상품5	327	3.27
	상품6	48	0.48
	상품7	1051	10.51
	상품8	145	1.45
	상품9	1246	12.46
계약자직업	직업1	1716	17.16
	직업2	732	7.32

	직업3	1677	16.77
	직업4	2888	28.88
	직업5	177	1.77
	직업6	10	0.10
	직업7	202	2.02
	직업8	486	4.86
	직업9	2112	21.12
피보험자성별	남	5055	50.55
	여	4945	49.45
가입경로	개인영업	8693	86.93
	TM	1307	13.07
부활유무	부활무	9443	94.43
	부활유	557	5.57

<표6> 평가용 자료의 선택된 변수의 빈도분포_정량변수

(단위: 원, 세)

요소	최소값	최대값	중위수	평균값	관측치수
한달납입보험료*	362.5	3000000	30925	51711.92	10000
계약자가입연령*	0	79	37	38.0179	10000

3. 분석결과

3.1 로지스틱 회귀분석

위의 과정을 통해 선별된 독립변수들을 이용하여 로지스틱 회귀분석을 한 결과는 다음과 같다. step-wise selection에 의해 변수 선택을 하였고, SAS 통계 패키지를 이용하였다.

<표7> 로지스틱회귀모형 분석결과

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
AIC	3180.823	3053.068
SBC	3188.033	3117.961
-2 LOG L	3178.823	3035.068

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	143.7551	8	<.0001
Score	166.5463	8	<.0001

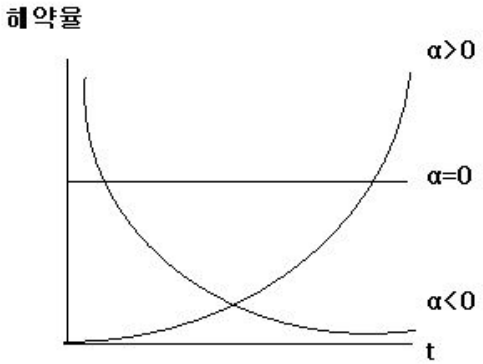
Wald	153.6626	8	<.0001

Analysis of Maximum Likelihood Estimates			
변수	추정회귀계수	표준오차	Pr > ChiSq
Intercept	-0.7648	0.9463	0.4190
lnt	-0.3275	0.0463	<.0001
부활유무	0.4110	0.2073	0.0474
로그계약자가입연령	-1.0030	0.1874	<.0001
로그한달납입보험료	0.1708	0.0641	0.0077
약관대출유무	0.7448	0.1992	0.0002
수금방법_방문	1.2836	0.3076	<.0001
계약자직업_3	0.2670	0.1327	0.0443
상품소분류_2	0.4587	0.1738	0.0083

계약유지기간을 나타내는 변수(T)와 로그-계약유지기간을 나타내는 변수(lnT)중 lnT가 유의하게 선택되었다. T가 유의하게 선택된 경우 보험 해약율의 분포가 Gompertz 분포를 따르고 lnT가 유의하게 선택된 경우 Weibull 분포를 따른다고 볼 수 있다.

본 논문의 분석 결과 lnT만이 유의하게 선택되어 생명보험 가입자의 시간에 따른 보험 해약율의 분포는 Weibull분포를 따른다는 것을 알 수 있다. Weibull분포의 경우 α에 따라 크게 다음 세가지의 해약율 형태를 갖는다.

$$\text{Weibull Distribution:} \quad h(t) = t^{\alpha} \quad ,\alpha > 0\,; = 0;\ < 0$$



<그림4> Weibull분포의 hazards function

α>0인 경우 시간이 증가함에 따라 해약율이 높아지고, α=0인 경우 해약율은 시간과 무관하게 되고, α<0인 경우 시간이 증가할수록 해약율이 낮아진다.

선택된 lnT의 계수가 양수일 때는 보험 가입 후 계약유지 기간이 길어질수록 해약율이 증가하고, 음수일 때는 보험 가입 후 계약유지 기간이 길어질수록 해약 확률이 줄어든다. 이번 실증분석에서는 lnT의 계수가 음수가 나왔다. 따라서 계약유지 기간이 길어질수록 해약율이 감소한다고 예측할 수 있다.

그 외에 부활 경험이 있는 고객일수록, 계약자가입연령이 작을수록, 월납보험료가 클수록, 약관대출경험이 있을 때, 수금방법이 방문일 때, 계약자 직업이 3군에 속하고, 상품소분류 중 2군에 속할 때 해약율이 높게 나타난다.

3.2 생존분석-Cox 비례위험모형

Cox의 비례위험 모형은 독립변수가 시간에 의존하는 경우와 그렇지 않은 두 가지 경우로 분석할 수 있으나 본 논문에서는 시간에 의존하지 않는 경우만을 분석하였다. Cox의 비례위험 모형의 step-wise selection을 통해 유의한 변수들을 선택한 결과는 다음과 같다.

<표8> Cox 비례위험모형 분석결과

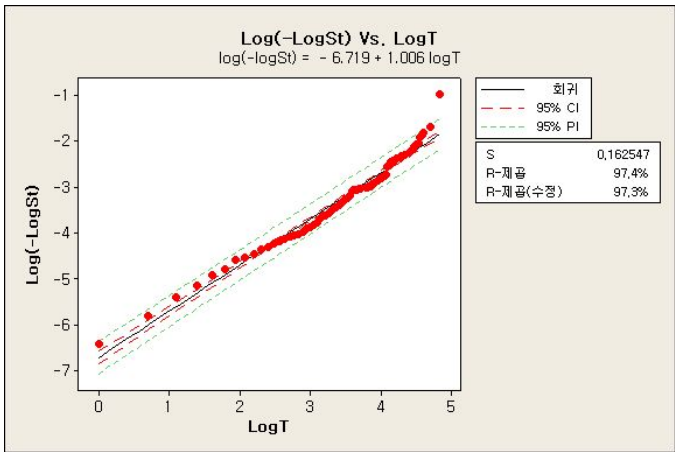
Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	244.6998	14	<.0001
Score	294.7780	14	<.0001
Wald	261.4278	14	<.0001

Model Fit Statistics		
Criterion	**Without Covariates**	**With Covariates**
AIC	6294.247	6077.548
SBC	6294.247	6132.412
-2 LOG L	6294.247	6049.548

Analysis of Maximum Likelihood Estimates			
변수	**추정회귀계수**	**Pr > ChiSq**	**Hazard Ratio**
수금방법_방문	1.04328	0.0003	2.839
로그한달납입보험료	0.42173	<.0001	1.525
만기유무	-1.00914	<.0001	0.365
가입경로	-1.09963	<.0001	0.333
상품중분류_3	2.02719	<.0001	7.593
상품소분류_1	-0.42275	0.0029	0.655
상품소분류_2	0.91444	<.0001	2.495
로그계약자가입연령	-0.97980	<.0001	0.375
계약자직업_1	0.81418	<.0001	2.257

계약자직업_2	0.80453	0.0004	2.236
계약자직업_3	1.11875	<.0001	3.061
계약자직업_4	1.05659	<.0001	2.877
계약자직업_5	0.75633	0.0439	2.130
계약자직업_8	1.09742	<.0001	2.996

해약률 함수가 Gompertz분포를 따를 때, $\log(-\log S(t))$와 T가 선형관계이며, Weibull분포를 따를 때는 $\log(-\log S(t))$와 $\log T$의 선형관계가 성립한다. 이를 확인하기 위해 LLS 그래프를 그려본 결과 다음과 같았다.



<그림5> Cox의 비례위험 모형의 LLS그래프

$\log(-\log S(t))$와 $\log T$의 그래프가 직선 형태를 이루는 것을 보아 위험률 함수는 Weibull분포를 따른다는 것을 알 수 있고 회귀식은 $\log(-\log S(T)) = -6.719 + 1.006 \log T$로 적합되었다. Weibull 분포에서 α=1 일 때, 지수분포와 같다는 것을 위에서 살펴보았다. 적합시킨 α값이 1.006으로 해약율이 지수분포를 따르며 시간에 무관한 것을 알 수 있다.

3.3 생존분석-가속화 고장시간모형

AFT모형은 위험함수를 이용한 생존분석과 같은 범주이므로 Cox의 비례위험 모형의 step-wise selection결과 선별된 변수와 동일한 변수를 설명변수로 사용하였다. 기저분포를 가정하기 위해, Weibull, inverse-Weibull, Exponential, inverse-Exponential, Gamma, inverse-Gamma, LogLogistic, inverse-Loglogistic, Logistic, inverse-Logistic, Lognormal, inverse-Lognormal, Normal, inverse-Normal의 log-likelihood값을 비교한

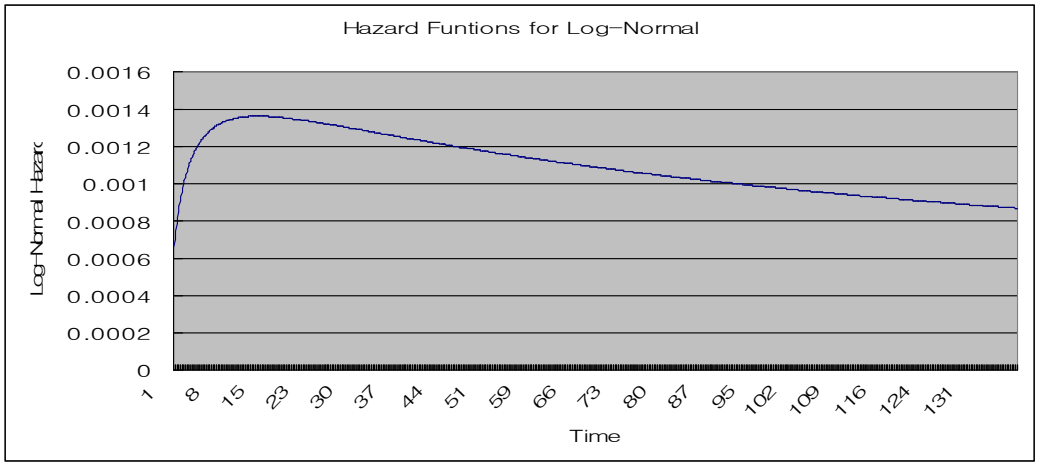
결과 Lognormal 분포가 베스트로 선택되어 기저 분포로 가정하였다.

<표9> 가속화 고장시간모형(AFT)분석 결과

Name of Distribution	Lognormal
Log Likelihood	-1773.9222

Analysis of Maximum Likelihood Estimates			
변수	추정회귀계수	표준오차	Pr > ChiSq
Intercept	8.2349	1.0133	<.0001
수금방법_방문	-1.3971	0.3509	<.0001
로그한달납입보험료	-0.4245	0.0729	<.0001
만기유무	0.8463	0.2546	0.0009
가입경로	1.1075	0.1684	<.0001
상품중분류_3	-2.0032	0.4513	<.0001
상품소분류_1	0.2586	0.1384	0.0617
상품소분류_2	-1.0146	0.1956	<.0001
로그계약자가입연령	0.9042	0.2123	<.0001
계약자직업_1	-0.9467	0.1798	<.0001
계약자직업_2	-0.8907	0.2293	0.0001
계약자직업_3	-1.2360	0.1740	<.0001
계약자직업_4	-1.1385	0.1619	<.0001
계약자직업_5	-0.8712	0.4126	0.0347
계약자직업_8	-1.2024	0.2472	<.0001
Scale	2.1641		

Cox의 비례위험 모형이 위험률을 예측하는 것과 달리 AFT모형은 계약유지시간 자체를 예측하므로 Cox의 비례위험 모형의 계수와 AFT모형의 계수의 부호가 반대이다. Lognormal분포의 특성상 시간이 지남에 따라 어느 순간 까지는 해약율이 증가하다가 일정 순간 t에서 피크를 이루고 그 후로 해약율이 감소하는 형태를 띈다. 다음 그래프는 공변량의 값이 모두 평균값인 한 고객을 가정하고 그 고객의 해약율을 그려본 것이다. 그래프를 보면 시간이 지남에 따라 생명보험의 해약율이 증가하다가 어느 순간 peak에 이르고 그 후 감소하고 있음을 알 수 있다. SAS 패키지의 옵션을 이용해 Hazard function을 통해 peak 점을 계산해 보았더니 t=14.5였다. 계약유지기간이 14개월과 15개월 사이에 이르렀을 때, 해약율이 peak에 이르고 서서히 감소하는 것이다.



<그림6> AFT 모형의 Lognormal Hazard 그래프

4. 모형평가

최적의 분석도구나 모형을 얻기 위해서는 여러 분석도구에서 나온 모형을 비교 평가해야 하고, 이를 통해 어떤 분석도구가 효율적인지를 판단 할 수 있다. 일반적인 모형평가의 기준으로는 모형이 얼마나 효과적으로 구축 되었는가(즉, 얼마나 간단한 모형인가)의 문제나 혹은 같은 모집단 내의 다른 데이터에 적용하는 경우 얼마나 안정적인 결과를 제공해 주는가. 즉, 일반화의 가능성 등 여러 각도에서 생각할 수 있다.

그러나 무엇보다도 우선적으로 고려되어야 할 사항은 구축된 모형이 얼마나 예측과 분류에서 뛰어난 성능을 보이는가이다. 이는 아무리 안정적이고 효과적인 모형도 실제 문제를 적용했을 경우 빗나간 결과만 양산한다면 아무런 의미가 없기 때문이다. 따라서 모형평가란 예측을 위해서 만든 모형이 임의의 모형모다 우수한지, 고려된 서로다른 모형들 중 어느것이 가장 우수한 예측력을 보유하고 있는지를 비교, 분석하는 과정이라 할 수 있다.

4.1 정오분류표

본 연구에서는 로지스틱 회귀분석과 Cox의 비례위험모형, 가속화 고장시간 모형(AFT)의 효율을 비교하기 위하여 우선 정오분류표를 사용하였다. 정오분류표는 종속변수의 실제 분류와 예측 분류 사이의 관계를 나타내는 표로 실제 계약을 유지한 계약자가 계약유지로 예측되는지, 실제 이탈한 계약자가 이탈로 예측되는지에 관한 빈도를 표로 나타낸 것이다. 이를 이용해 정분류율, 오분류율, 민감도, 특이도를 구할 수

있다.

<표10>정오분류표

	예측유지	예측이탈
실제유지	실제 계약 유지 고객이 유지로 예측된 빈도	실제 계약 유지 고객이 이탈로 예측된 빈도
실제이탈	실제 이탈한 고객이 유지로 예측된 빈도	실제 이탈한 고객이 이탈로 예측된 빈도

$$\text{정분류율} = \frac{(\text{실제유지}, \text{예측유지})\text{의 빈도} + (\text{실제이탈}, \text{예측이탈})\text{의 빈도}}{\text{전체빈도}}$$

$$\text{오분류율} = \frac{(\text{실제유지}, \text{예측이탈})\text{의 빈도} + (\text{실제이탈}, \text{예측유지})\text{의 빈도}}{\text{전체빈도}}$$

$$\text{민감도}(Sensitivity) = \frac{(\text{실제이탈}, \text{예측이탈})\text{의 빈도}}{\text{실제 이탈 고객의 빈도}}$$

$$\text{특이도}(Specificity) = \frac{(\text{실제유지}, \text{예측유지})\text{의 빈도}}{\text{실제 유지 고객의 빈도}}$$

실제 계약을 유지한 고객을 계약유지로, 실제 이탈한 고객을 이탈로 예측한 비율을 구하는 정분류율과 실제 계약을 유지한 고객을 이탈로, 실제 이탈한 고객을 계약유지로 예측한 비율을 구하는 오분류율, 실제 이탈한 고객 중 이탈이라고 예측하는 조건부 확률을 의미하는 민감도와 실제 계약 유지 고객 중 유지라고 예측하는 조건부 확률인 특이도를 구하는 식은 위과 같으며 민감도와 특이도가 높은 경우 예측력이 뛰어나다고 할 수 있으며 고객 관리 측면에서는 민감도가 더욱 중요한 평가 기준이 된다.

<표11> 로지스틱회귀모형의 정오분류표

실제 \ 예측	유지	이탈	합계
유지	5852	3735	9587
이탈	148	265	413
합계	6000	4000	10000

로지스틱 회귀분석의 정오분류표를 이용해 정분류율, 오분류율, 민감도, 특이도를 구

해 정분류율 61.17%, 오분류율 38.83%, 민감도 64.16%, 특이도 61.04%임을 알 수 있다.

<표12> Cox 비례위험모형의 정오분류표

실제 \ 예측	유지	이탈	합계
유지	5804	3783	9587
이탈	196	217	413
합계	6000	4000	10000

Cox의 비례위험모형의 정오분류표를 이용해 정분류율, 오분류율, 민감도, 특이도를 구해 정분류율 60.21%, 오분류율 39.79%, 민감도 52.54%, 특이도 60.54%임을 알 수 있다.

<표13> 가속화 고장시간모형(AFT)의 정오분류표

실제 \ 예측	유지	이탈	합계
유지	5812	3775	9587
이탈	188	225	413
합계	6000	4000	10000

가속화 고장시간 모형(AFT)의 정오분류표를 이용해 정분류율, 오분류율, 민감도, 특이도를 구해 정분류율 60.37%, 오분류율 39.63%, 민감도 54.48%, 특이도 60.62%임을 알 수 있다.

예측 결과 세 모형 모두 판별력에 큰 차이는 없었지만 로지스틱 회귀분석이 계약유지 고객 분류에 뛰어나고, 로지스틱 회귀분석, 가속화 고장시간 모형, Cox의 비례위험 모형의 순서로 이탈자 분류에 더 뛰어남을 알 수 있다.

4.2 리프트 차트

리프트 차트는 모형평가에 있어 다음과 같은 과정을 통해 만들어진다.

1. 모형설정을 통해 사후확률을 구한다.
2. 사후확률의 순서에 따라 전체 데이터 셋을 정렬한다.
3. 정렬이 끝난 데이터 셋을 균일하게 N등분 한다.
4. N등분의 각 등급에서 목표 변수의 특정 범주에 대한 빈도를 구한다.
5. N등분의 각 등급에서 %Response 및 Lift 통계량을 다음과 같이 계산한다.

$$\%Response = \frac{\text{해당 등급에서 목표변수의 특정 범주 빈도}}{\text{해당 등급에서 전체 빈도}}$$

$$\text{Base Line \%response} = \frac{\text{전체에서 목표변수의 특정 범주 빈도}}{\text{전체 빈도}} \times 100$$

$$\text{Lift} = \text{해당 등급의 } \%Response / \text{Base Line Lift}$$

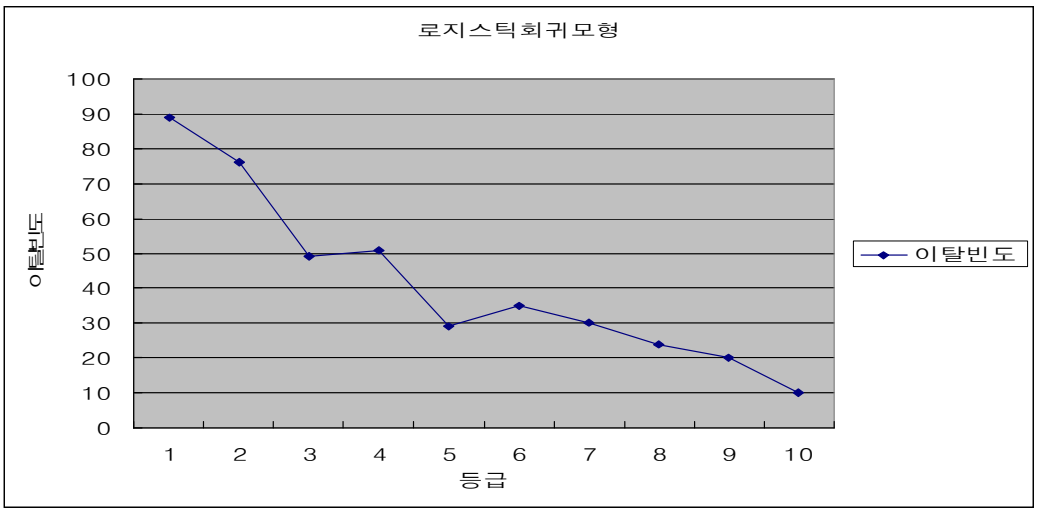
6. 수평축에는 N등분의 등급을 수직 축에는 위에서 구한 통계량을 이용해 그래프를 그린다.

총 10000개의 관측을 10구간으로 나누고 사후확률을 구해 각 등급의 %Response와 Lift값을 구해 보았다.

<표14> 로지스틱회귀모형의 Lift 값

등급	보험계약자이탈빈도	해당등급의 빈도	%Response	Lift
1	89	1000	8.9	2.154964
2	76	1000	7.6	1.840194
3	49	1000	4.9	1.186441
4	51	1000	5.1	1.234867
5	29	1000	2.9	0.702179
6	35	1000	3.5	0.847458
7	30	1000	3.0	0.726392
8	24	1000	2.4	0.581114
9	20	1000	2.0	0.484262
10	10	1000	1.0	0.242131

Base Line %response=413/10000*100=4.13



<그림7> 로지스틱 회귀모형의 등급에 따른 이탈 빈도

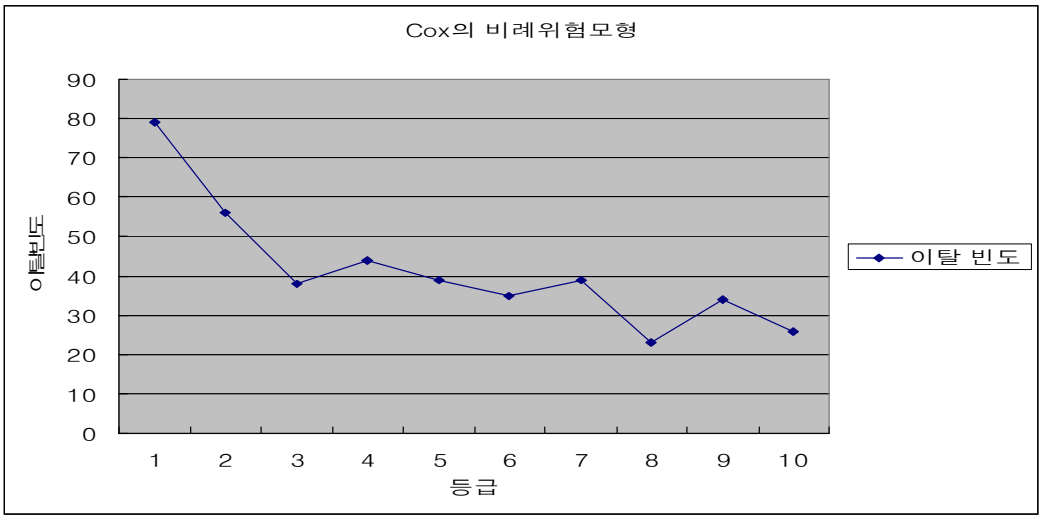
위의 <표14>와 <그림7>을 통해서 로지스틱 회귀모형은 1등급(상위 10%)에서 89명의 고객의 이탈을 예측하는데 성공하였으며 이는 평가용 데이터 셋 전체 이탈 고객 중 약 22%를 사전에 예측할 수 있다는 것을 의미한다.

5등급에서 6등급으로 넘어가는 시점에 Lift값이 증가하지만 등급이 낮아짐에 따라 비교적 안정적으로 Lift값이 감소하는 것을 확인할 수 있다.

<표15> Cox 비례위험모형의 Lift 값

등급	보험계약자이탈빈도	해당등급의 빈도	%Response	Lift
1	79	1000	7.9	1.912833
2	56	1000	5.6	1.355932
3	38	1000	3.8	0.920097
4	44	1000	4.4	1.065375
5	39	1000	3.9	0.94431
6	35	1000	3.5	0.847458
7	39	1000	3.9	0.94431
8	23	1000	2.3	0.556901
9	34	1000	3.4	0.823245
10	26	1000	2.6	0.62954

Base Line %response=413/10000*100=4.13



<그림8> Cox의 비례위험모형의 등급에 따른 이탈 빈도

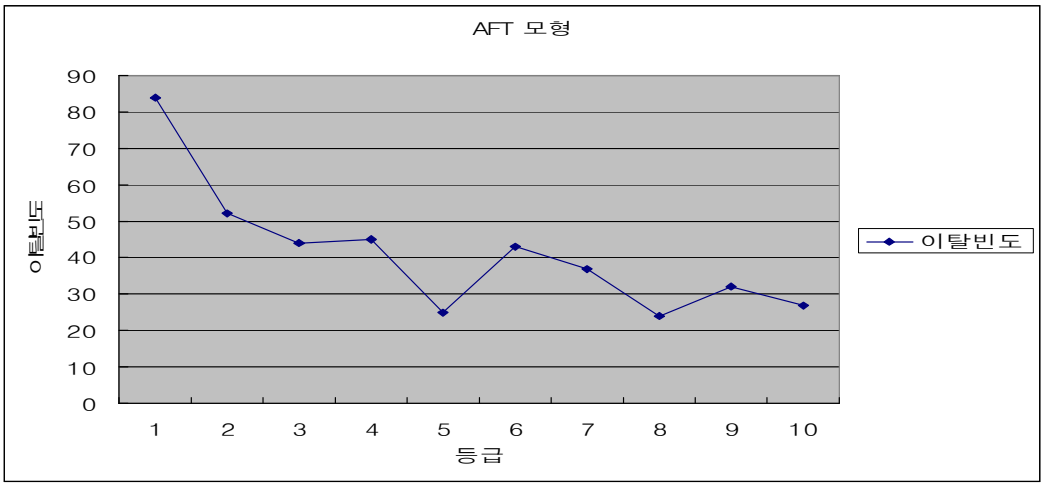
위의 <표15>와 <그림8>을 통해서 Cox의 비례위험모형은 1등급(상위 10%)에서 79명의 고객의 이탈을 예측하는데 성공하였으며 이는 평가용 데이터 셋 전체 이탈 고객 중 약 19%를 사전에 예측할 수 있다는 것을 의미한다.

Cox의 비례위험모형 역시 등급이 낮아짐에 따라 Lift값이 감소하는 것을 확인할 수 있으나 3등급에서 4등급으로 넘어가는 시점과 6등급에서 7등급으로, 8등급에서 9등급으로 넘어가면서 Lift값이 증가하여 다소 불안정한 것을 확인할 수 있다.

<표16> 가속화 고장시간모형(AFT)의 Lift 값

등급	보험계약자이탈빈도	해당등급의 빈도	%Response	Lift
1	84	1000	8.4	2.033898305
2	52	1000	5.2	1.259079903
3	44	1000	4.4	1.065375303
4	45	1000	4.5	1.089588378
5	25	1000	2.5	0.605326877
6	43	1000	4.3	1.041162228
7	37	1000	3.7	0.895883777
8	24	1000	2.4	0.581113801
9	32	1000	3.2	0.774818402
10	27	1000	2.7	0.653753027

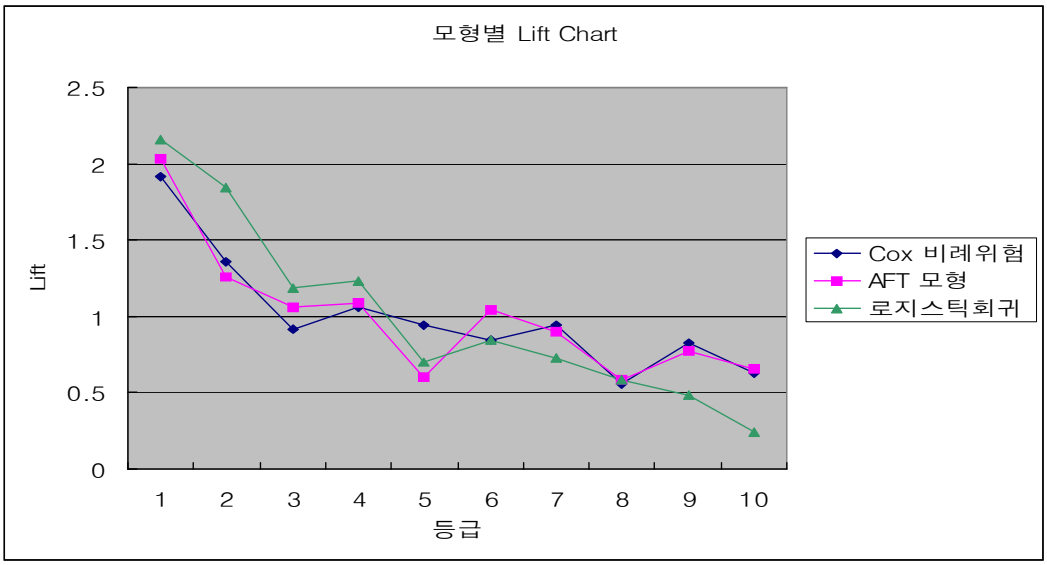
Base Line %response=413/10000*100=4.13



<그림9> AFT모형의 등급에 따른 이탈 빈도

위의 <표16>과 <그림9>를 통해서 가속화 고장시간모형은 1등급(상위 10%)에서 84명의 고객의 이탈을 예측하는데 성공하였으며 이는 평가용 데이터 셋 전체 이탈 고객 중 약 20%를 사전에 예측할 수 있다는 것을 의미한다.

가속화 고장시간모형 역시 등급이 낮아짐에 따라 Lift값이 감소함을 확인할 수 있다. 그러나 5에서 6등급 시점에서 Lift값이 큰 폭으로 상승하고 8에서 9등급 시점 역시 값이 상승하여 다소 불안정한 형태를 보이고 있다.



<그림10> 모형별 Lift Chart

Lift값을 이용해 Lift차트를 그려보았다. Lift차트는 전체 집단 대비 해당 등급의 예측력 정도를 나타내며 등급이 하위로 내려갈수록 Lift가 감소하면 모형의 예측력이 적절하다는 것을 나타낸다. 구해진 Lift차트를 보면 세 모형 모두 그래프가 감소하고 있어 예측력은 적절한 것을 알 수 있으나 Base Line값 자체가 4.13으로 매우 낮아 Lift값 자체는 그리 크지 못함을 알 수 있다. 로지스틱 회귀모형, 가속화 고장시간모형, Cox의 비례위험모형 순서로 예측력이 뛰어남을 확인할 수 있다.

Ⅳ. 요약 및 결론

지난 수십년 동안 우리나라의 경제 발전과 함께 생명보험산업 또한 비약적인 발전을 이루어 왔다. 방카슈랑스와 자본시장통합법과 같은 시장 환경의 변화 속에서 생명보험산업이 효율적인 경영을 해 나가기 위하여 논의 되는 것 중 하나는 높은 실효·해약율에 대한 방안을 마련하여 내실 위주의 경영을 해 나가는 것이다. 따라서 본 연구에서는 생존분석기법(Survival Analysis)중 Cox의 비례위험모형과 가속화 고장시간모형, 이분류 예측모델 중 로지스틱회귀분석을 소개하고 국내 생명보험사의 고객 데이터 20000개의 관측치를 통하여 고객의 실효·해약에 유의한 영향을 주는 변수를 찾아 예측모델을 만들어보고 그 성과를 비교해 보았다.

모형 평가 결과 정분류율이 로지스틱 회귀분석 61.17%, Cox의 비례위험모형 60.21%, 가속화 고장시간모형 60.37%로 계약유지 고객을 계약 유지로 이탈 고객을 이탈로 예측하는 성과에서는 거의 동일한 예측력을 갖고 있었다. 실제 계약 유지고객 중 계약 유지로 예측할 확률을 나타내는 특이도의 경우에도 로지스틱 회귀분석이 61.04%, Cox의 비례위험모형이 60.54%, 가속화 고장시간모형이 60.62%로 거의 동일한 예측력을 보였다. 그러나 실제 이탈 고객을 이탈로 예측할 확률을 나타내는 민감도의 경우 로지스틱 회귀모형이 64.16% Cox의 비례위험모형이 52.54%, 가속화 고장시간모형이 54.48%로 로지스틱 회귀모형, 가속화 고장시간모형, Cox의 비례위험모형 순서로 예측성능이 좋은 것으로 관측되었다.

Lift Chart에 의한 평가 결과 세 모형 모두 등급이 낮아짐에 따라 그래프가 감소하는 형태를 보여 예측력이 적절함을 확인할 수 있었다. 그러나 Cox의 비례위험모형과 가속화 고장시간모형의 Lift값이 등급이 낮아지며 한 두 차례 증가하는 불안정한 모습을 보이기도 하여 로지스틱 회귀모형이 다소 안정적인 것으로 나타났다.

전체적으로 로지스틱 회귀모형은 이탈 쪽으로 Cox의 비례위험모형과 가속화 고장시간모형은 유지 쪽으로 편향된 예측을 하는 경향을 나타내었다. 예측력 평가 결과 로지스틱 회귀모형이 더 나은 예측력을 보였지만 유지 혹은 이탈이라는 이분류의 결과만을 제공하는 한계를 가지고 있다. Cox의 비례위험모형과 가속화 고장시간모형은 개별 고객의 평균 생존시간을 예측함으로써 생명보험사의 CRM측면 뿐 아니라 리스크 관리와 상품개발측면에서도 보다 유용한 결과를 제공한다고 할 수 있다.

본 연구에서 사용된 설명변수가 보험상품, 계약자, 피보험자와 관련한 제한된 정보로였다는 점, FP, 계약자 배당금 수준, 실제 경기 상황 등을 반영하지 못하였고 고객의 재무상태를 반영하는 변수가 적었다는 점에 그 첫 번째 한계가 있다. 또한 분석용 데

이터와 평가용 데이터의 실효·해약한 관측치가 적어 정확한 예측력 평가가 어려웠다는데 두 번째 한계가 있다. 이를 보완한 연구는 향후 과제로 남긴다. 본 연구는 로지스틱회귀분석이 가진 이분류 예측의 단점을 보완할 수 있는 생존분석법을 통해 생명보험가입자의 계약 유지 수명을 예측해 보고, 실효·해약 모형을 제시하였다는데 그 의의를 찾을 수 있다.

<참고문헌>

1. 논문

김수나, 「로지스틱과 생존분석을 이용한 생명보험 해약 예측모형연구」, 석사학위논문, 이화여자대학교, 2008.
배효원, 「기업의 수명 자료를 이용한 기업 부도 예측 모형」, 석사학위논문, 이화여자대학교, 2008.
안철경,기승도, 「금융 겸업화에 대응한 보험회사의 채널전략」, 보험개발원 보험연구소, 2008
윤미래, 「데이터 마이닝을 이용한 부도예측모델」, 석사학위논문, 이화여자대학교, 2004.
이현우, 강중철 「생존분석법을 이용한 생명보험 실효·해약분석」, 통계청 통계분석연구, 1999.
이호영, 「생명보험회사 고객관계관리(CRM)를 위한 고객이탈자 예측 모형에 관한 연구」,박사학위논문, 한국외국어대, 2005.
임지영, 「생명보험상품의 유지율분석을 통한 생보산업의 질적 발전에 관한 연구」,석사학위논문, 서강대학교, 2004.
정연식, 송상규, 최기주 「생존분석 기법을 이용한 고속도로 교통사고 지속시간 예측 모형」,대한교통학회지 제25권 제5호, 2007

2. 단행본

박성현, 『회귀분석』, 민영사, 1998.
박재빈, 『생존분석 이론과 실제』, 신광출판사, 2006
성내경, 『기본 SAS소프트웨어』, 자유아카데미, 2002
송혜양, 정갑도, 이원철, 『생존분석』, 청문각, 2001
허명회, 박미라, 『SAS와 NCSS를 이용한 생존분석』, 자유아카데미, 1991.
Brian Everitt, Graham Dunn 『Applied Multivariate Data Analysis』,Oxford Univ. Press, 2001.
Chatterjee,『Regression Analysis by Example』,wiley, 2006.
David Stirzaker,『Probability and random variables』, Cambridge Univ. Press, 1999.
Hogg, McKean, Craig,『Introduction to Mathematical Statistics』,Prentice Hall, 2005
Paul D. Allison, 『Survival Analysis Using the SAS System』, SAS, 2001.
Richard Scheaffer,『Introduction to probability and its applications』,Duxbury Pres. 1995.
Sheldon Ross,『A first course in probability』, Prentice Hall, 2005

3. 인터넷 사이트

http://www.boost.org/libs 2008.11.5

<Abstract>

A Study on Application of Survival Analysis Methods to Life Insurance

Park, Young-ran
Department of Statistics
The Graduate school
Ewha womans Univ.

Using Cox's Proportional Hazards Model (PHM) and Accelerated Failure Time (AFT) Model, we investigate the distribution of retention time from the start of an insurance contract to the possible termination of a contract. We also compares Survival Analysis with Logistic Regression in order to evaluate prediction performances. Recently, a number of researches on insurance cancellation prediction focused on the customer relationship management (CRM) and the risk management (RM). Nevertheless, these studies have the limitation in using prediction techniques based on dichotomous classification which only predict cancellation or retention of insurance. Moreover, due to the fixed Δt ,

such techniques provide limited information.

Therefore, we applied Cox's PHM to find significant variables having influence on cancellation of insurance and AFT Model to investigate optimal baseline hazard rate function. This thesis is not without limitations due to restricted number of explanatory variables and 3-4% of insurance cancellation rate. Despite these limitations, this thesis seems to make some contribution to the problem of identifying the distribution of retention time which complements conventional dichotomous prediction techniques.

감사의 글

먼저, 늘 함께하여 주시는 사랑과 은혜의 하나님 아버지께 감사와 영광을 돌립니다.

경영학을 주 전공으로 하고 있었던 제가 어려운 통계학을 복수전공하고 대학원에 진학하기까지 공부하게 된 것은 교수님들의 훌륭한 가르침과 노고 덕분이었습니다. 부족한 저를 늘 격려해주시고 진리에 대한 열정을 삶으로 가르쳐 주신 존경하는 소병수 교수님, 수업 시간에는 카리스마 있으시지만 개인적으로 질문 드릴 때면 자상하고 따뜻하셨던 이외숙 교수님, 통계학에 대한 흥미를 잃지 않을 수 있도록 알아듣기 쉽고 재미있게 수업을 진행해 주셨던 오만숙 교수님, 회귀분석에 대한 개념을 확실히 알 수 있게 지도해 주신 임용빈 교수님, 시계열 분석을 이해하기 쉽게 가르쳐 주실 뿐 아니라 사진과 음악을 통해 학생들에게도 기쁨과 추억을 남겨주시는 신동완 교수님, R에 친숙해 지도록 도와주시고 데이터 마이닝을 하는 자세에 대해 큰 가르침을 주셨던 송종우 교수님, 늘 자상하게 학생들을 배려해 주시는 젠틀한 이용희 교수님, 생존분석의 개념을 가르쳐 주셨던 강승호 교수님, 아무것도 모르던 학부 3학년 때, 처음으로 통계학을 가르쳐 주셔서 통계학에 관심을 갖도록 도와주신 문보영 선생님, 수업을 통해 리스크 관리에 대한 흥미를 갖게 해주시고, 늘 도움을 주려 하시는 경영학과 김진호 교수님 모두 감사드립니다.

늘 연구하는 자세에 대해 가르쳐 주시고 때론 아버지처럼 함께 식사할 때는 재미있는 선배처럼 대해 주시는 완소 김병연 박사님, 편안하게 공부하고 일할 수 있도록 배려해 주시는 이병관 과장님, 처음 KIF에 들어왔을 때 적응할 수 있도록 도와주신 송재만 연구원님, 털털한 성격이시지만 귀여운 캐릭터 좋아하는 지윤언니(많이 가르쳐 주시고 실수해도 쿨하게 이해해 주셔서 고마워요), 나와 해림이를 신우회로 이끌어 주신 참 좋은 은해언니, 우리끼리 있을 때의 실질적 리더이신 매력적인 지선언니, 학교 선배를 만나서 좋았던 기타 연주 실력 겸비하신 정화언니, 나의 멘토가 되어주신 멋있는 유진언니, 늘 칭찬해 주시고 격려해 주시는 과장님과 우리에게 영과 육의 양식을 주시는 실장님, 우리의 귀여니 스타 해림이(너랑 함께해서 4개월의 시간이 기쁨과 감사로 가득 채워졌던 거 같아.)

중학교 때부터 나의 든든한 동역자가 되어준 남현이와 지혜와 상민이, 신광에서의 추억을 간직한 은미, 진영이, 태옥이, 우리 셋이 모이면 항상 즐거웠던 혜민이 민선이, 김교열 목사님, 소영언니 주현언니, 수정언니, 기현오빠, 인영언니, 유미언니, 희진언니, IVF에서 동고동락했던 사랑하는 동생들(주애, 은실이, 은진이, 한나, 슬기, 은경이, 성은이)과 언니와 친구들 특히 논문 막바지라 바쁜데도 시간 쪼개서 큰 도움을 준 정윤이, 대학원의 시간 동안 함께여서 기뻤던 조현언니, 윤주언니, 희주를 비롯한 우리

학기 친구와 언니들, 늘 옆에서 든든하고 멋있는 동역자가 되어준 진욱오빠, 사랑과 헌신으로 키워주신 존경하는 부모님, 이기적인 동생을 당근과 채찍으로 바로잡아주는 사랑하는 오빠 모두 감사합니다.