보험해지모형 (Logistic, GAM)

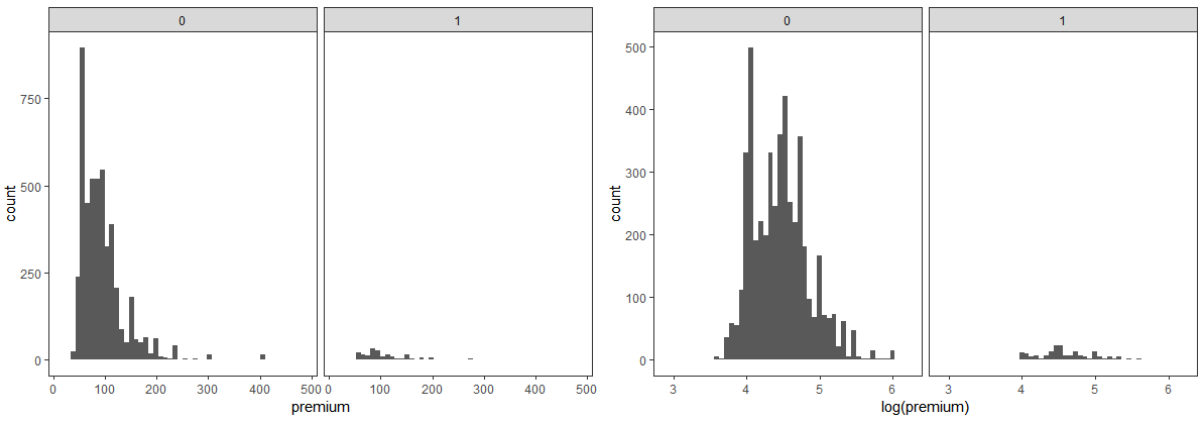
a) Logistic Regression을 이용한 방법 – GLM

* 변수 설정

구분	변수명	변수내용	Type	비고
타겟변수	delta	TARGET_YN		
기존변수	x1	가입연령	numeric	
기존변수	x2	납입방법	factor	
파생변수	x3	납입기간	numeric	월단위 총 납입기간 = 납입기간(연) * 12
파생변수	x4	수금방법	factor	X2 변수를 자동이체 or 나머지로 변환
파생변수	x5	보험료	numeric	납입방법에 따른 월납보험료 변환
기존변수	x6	부활유무	Factor	
기존변수	x7	계약일자	Date	
기존변수	x8	보험금지급만기일자	Date	
기존변수	x9	최종납입횟수	Numeric	
기존변수	x10	상품중분류	Factor	
기존변수	x11	상품소분류	Factor	
파생변수	x12	보험기간	numeric	지급만기일자 – 계약일자(x7)
파생변수	x13	차월	numeric	현재시점('2001-06-30') – 계약일자(x7)
파생변수	X14	연체	factor	1 (최종납입횟수<min(최종납입횟수, 월차)) 0 (최종납입횟수>min(최종납입횟수, 월차))
파생변수	X15	Percent	numeric	최종납입횟수 / min(최종납입횟수, 월차)
기존변수	X16	약관대출유무	factor	1 (대출 유) 2(대출 무)

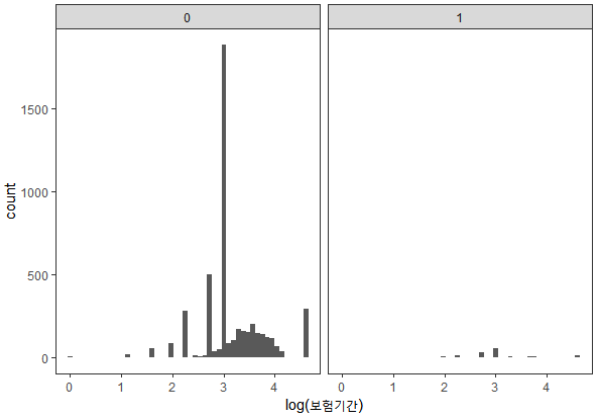
위의 변수 중 Numeric 변수에 대해서 그래프를 그려보았다.

① 보험료

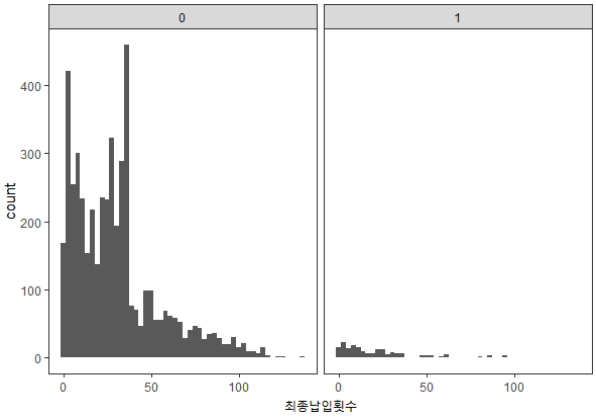


보험료 변수에 대한 histogram 을 그려본 결과 왼쪽으로 치우쳐 있는 형태를 보인다. 따라서 적절한 log 변환을 통해 종모양 분포로 변환하였다.

② 보험기간



③ 최종납입횟수



보험기간, 최종납입횟수에 대해 히스토그램을 그려본 결과는 위와 같다. 최종납입횟수는 월단위로 계산하였고, 0~50 구간에 대부분의 값이 몰려있고 오른쪽으로 꼬리가 긴 분포를 가진다.

$$\log\left(\frac{p}{1-p}\right) = x1 + x16 + x6 + x7 + x8 + x12 + x14 + x15 + log(x5) + x9 +$$
$$x1:x16 + x1:x6 + x16:x6 + x16:x8 + x16:x12 + x16:log(x5) + x16:x9 +$$
$$x6:x7 + x6:x12 + x7:x8 + x7:x14 + x7:x15 + x7:x9 + x8:x15 + x8:log(x5) +$$
$$x8:x9 + x12:log(x5) + x14:x15 + x14:log(x5) + x14:x9 + x15:x9 + log(x5):x9$$

AIC = 1157

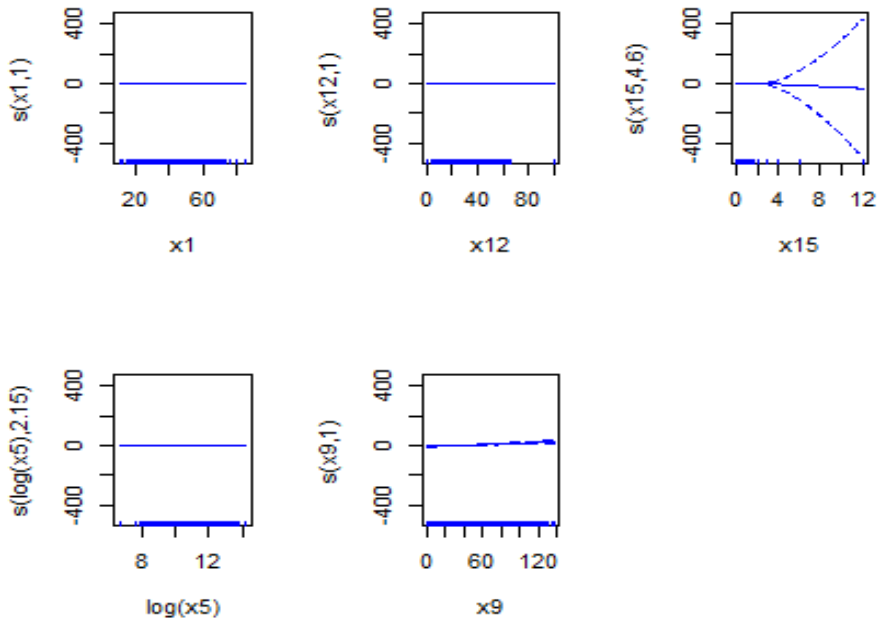
기존변수를 위 표와 같이 변환하여 파생변수를 생성하였다. 파생변수를 포함한 x1 ~ x21까지의 변수를 additive model 을 가정했을 때 y에 유의한 효과를 나타낸 x1, x3, x7, x8, x10, x15, x17, x20, x21 변수를 1차적으로 선택하였다. Alpha = 0.15일 때 Stepwise selection 방법을 사용하였고, link function에 대해 probit, logit, gompit 함수를 적용하여 본 결과 logit link를 적용한 경우 AIC = 1157.243로 모델의 예측력이 가장 좋았다. 따라서 위와 같은 모델을 최종 모델로 선택하였다.

b) GAM 을 이용한 방법

	장점	단점
GAM MODEL	- 설명변수들에 비선형함수를 적합할 수 있어 비선형관계를 자동으로 모델링 가능하게 한다 - 예측의 정확도가 높다 - y에 대한 x변수들의 각각의 영향을 개별적으로 표현가능하다	-모델이 가산적이어야하기 때문에 중요한 상호작용 효과를 놓칠 수 있다 -만약 교호작용이 꼭 반영되어야 하는 경우 새로운 함수를 찾아야한다

$$\log\left(\frac{p}{1-p}\right) = s(x1) + x16 + x6 + x7 + x8 + s(x12) + x14 + s(x15) + s(\log(x5)) + s(x9) +$$
$$x1{:}x16 + x1{:}x6 + x16{:}x6 + x16{:}x8 + x16{:}x12 + x16{:}\log(x5) + x16{:}x9 +$$
$$x6{:}x7 + x6{:}x12 + x7{:}x8 + x7{:}x14 + x7{:}x15 + x7{:}x9 + x8{:}x15 + x8{:}\log(x5) +$$
$$x8{:}x9 + x12{:}\log(x5) + x14{:}x15 + x14{:}\log(x5) + x14{:}x9 + x15{:}x9 + \log(x5){:}x9$$

AIC = 1212



Generalized additive model은 기존의 선형모델에서 가법성은 유지하면서 각 변수에 non-linear한 적합을 가능하게 하는 방법이다. 1차적으로 선택한 x1, x3, x7, x8, x10, x15, x17, x20, x21 변수 중 numeric type인 x1, x15, x17, x21에 최적함수를 찾기 위해 spline 함수를 적용한 결과 GAM 모델에서 AIC = 1212로 glm model이 더 예측력이 좋았다.

이 모델은 각 변수들의 영향을 개별적으로 볼 수 있는데, 현재 그래프에서는 각각의 변수가 다른 변수를 고정했을 때 해지확률에 유의한 영향을 미치지 않는다는 것을 알 수 있다. 이는 numeric 변수들의 spline 변환이 유의하지 않다는 것을 의미한다.