이론통계학2-Project #7 생명보험해지-CRM

강아미, 고유정, 윤보인, 이혜린, 홍지원

Data: 생명보험계약 해지여부자료 : (excel 파일)
국내 A생명보험사의 2001년도 6월말에서 9월말까지 3개월 내에 보험계약 해지여부 및 계약자정보 중 일부를 표본추출(N=10,000)한 자료. 자료 시점을 [2001-09-30]으로 정해 분석을 진행한다.

Part 1) 고객 변수$(x_1, \cdots, x_{11})$을 활용하여 3개월 내 고객의 보험계약해지여부(δ_i)를 예측하고자 한다.

a) Logistic Regression을 이용한 방법 :

i) 정성변수들의 지시변수화 및 정량변수들의 추가 변수변환 검토

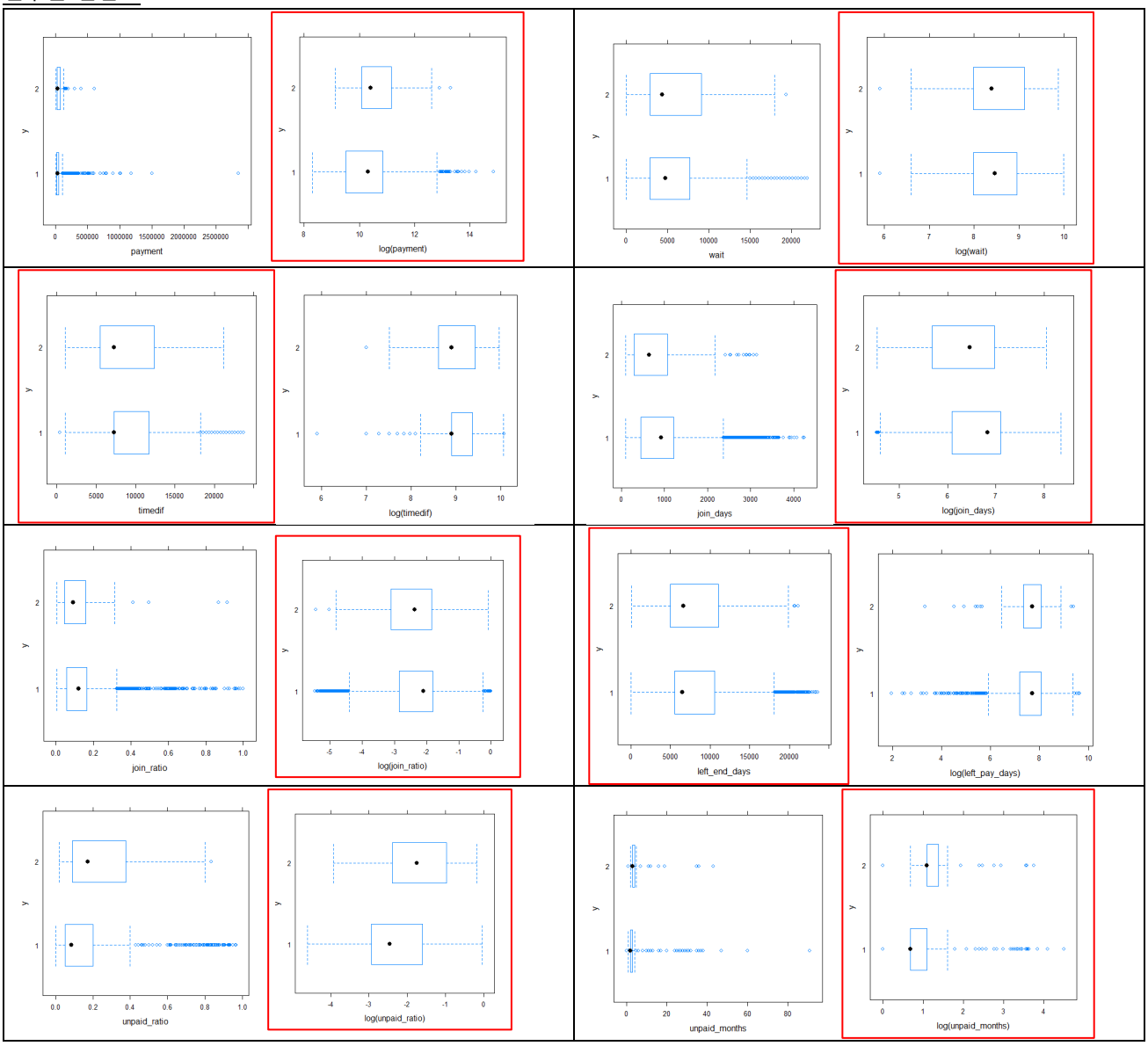
데이터 정리

반응변수(Y)				
	Variable	Description	Details	Others
δ	Y	이탈유무	0: 이탈하지 않음, 1: 이탈	
기존 설명변수(X)				
	Variable	Description	Details	Others
X1	age	가입연령	11세~85세	
X2	cycle	납입방법	1: 월납, 2: 3개월납, 3: 6개월납, 4: 연납	범주형 변수
X3	pd_y	납입기간	단위: 년	
X4	t_method	수금방법	1: 방문, 2: 자동이체, 3:지로, 4: 직납, 5: 카드납	범주형 변수
X5	payment	보험료		로그변환 검토
X6	revival	부활유무	1: 있음, 2: 없음	
X7	p_date	계약일자		
X8	expiration	지급만기일자	무만기=99990000 수명을 80세로 간주해 만기일 설정	
X9	number	최종납입횟수		
X10	type_m	상품 중분류	0-5	범주형 변수
X11	type_s	상품 소분류	0-9	범주형 변수
파생변수(X)				
	Variable	Description	Details	Others
X12	cycle_yr	연 납입 횟수	일년에 납입하는 횟수	범주형 변수
X13	age_L	나이 그룹화	10세 간격으로 범주화	범주형 변수
X14	p_date_y	가입연도		
X15	expiration_y	만기연도	수명을 80세로 간주해 만기일 설정	
X16	mm	가입월	가입 시 날짜와 만기 시 날짜 동일	범주형 변수
X17	timedif	총 계약기간	만기 날짜에서 계약일자를 뺀 일 수	
X18	no_exp	무만기여부	1: 무만기, 0: 만기	범주형 변수
X19	pay_end	납입기한 채운 여부	1: 납입 기한 만료, 0: 납입 중	범주형 변수
X20	pay_expiration	납입만료일자	보험 시점에서 납입기간을 합한 날짜	
X21	wait		지급만기일자에서 납입 만료 일자를 뺀 일수	로그변환 검토
X22	join_days	가입일수	계약일자로부터 자료시점까지의 일수	로그변환 검토
X23	join_months	가입월수	가입일수를 30.5로 나누어 구한 값	로그변환 검토

X24	left_end_days	계약 남은 일수	보험금 지급까지 남은 일수	
X25	left_pay_days	납입 남은 일수	보험금 납입 완료까지 남은 일수	로그변환 검토
X26	join_ratio		전체 계약기간에서 현재까지의 비율	로그변환 검토
X27	pay_join_ratio		납입 기간에서 현재까지의 비율	로그변환 검토
X28	pay_ratio		전체 납입 횟수 중 지금까지 납입한 비율	로그변환 검토
X29	paid_ratio		보험료를 빼먹지 않고 납입한 비율	
X30	paid_month		보험료를 납입한 개월수	로그변환 검토
X31	unpaid_month		보험료를 납입하지 않은 개월수	로그변환 검토
X32	unpaid_ratio		보험료를 납입하지 않은 비율	로그변환 검토
X33	total_payment	총 납입 보험료	보험료 * 최종납입횟수	로그변환 검토

wait 변수가 음수인 경우, unpaid_months가 음수, 납입 기한이 99년인 관측치는 이상치로 여기고 데이터에서 제외한다. 연속형 변수에 대해서는 로그변환을 통해 적절한 변수를 선택한다.

변수별 산점도



왜도가 높은 설명변수에 대해 로그 변환을 한 후 분석을 시작한다.

ii) 적절한 link 고려 및 AIC 기준을 활용해 최적 회귀모형 찾기

	Model	AIC
model1 (link=probit)	y ~ paid_ratio + timedif + payment_per + age + cycle + pay_join_ratio + pay_ratio + join_days + left_end_days + payment + paid_months + total_payment + unpaid_months + number + unpaid_ratio + timedif:join_days + total_payment:unpaid_months	1325.079
model2 (link=logit)	y ~ revival + pay_end + left_ratio + join_ratio + _months + payment_per + age + timedif + unpaid_ratio + cycle + paid_ratio + pay_join_ratio + pay_ratio + left_end_days + payment + paid_months + total_payment + number + revival:unpaid_ratio + pay_end:unpaid_ratio + join_ratio:unpaid_ratio + age:unpaid_ratio + timedif:unpaid_ratio + unpaid_ratio:pay_join_ratio + unpaid_ratio:pay_ratio + unpaid_ratio:left_end_days + unpaid_ratio:payment + unpaid_ratio:paid_months + unpaid_ratio:total_payment + unpaid_ratio:number	1357.824
model3 (link=cloglog)	y ~ age + cycle + revival + number + timedif + join_days + left_end_days + join_ratio + pay_join_ratio + left_ratio + pay_ratio + total_payment + paid_months + unpaid_months + unpaid_ratio + payment_per + pay_end + revival:left_end_days + revival:join_ratio + unpaid_months:unpaid_ratio + join_ratio:unpaid_ratio + join_days:unpaid_ratio + age:pay_end	1412.938

Variable	Coefficient	Variable	Coefficient
(Intercept)	-1.5932e+02	log(join_days)	3.4745e+01
log(paid_ratio)	-1.1469e+01	left_end_days	1.7240e-02
timedif	-1.7369e-02	log(payment)	4.05634e+01
log(payment_per)	2.6053e+01	log(paid_months)	1.4511e+01
age	-1.6748e-02	log(total_payment)	-4.2440e+01
cycle3	-2.5667e+01	log(unpaid_months)	3.7764e+00
cycle6	-5.0753e+01	number	4.8669e-01
cycle12	-6.2261e+01	log(unpaid_ratio)	4.2595e+00
log(pay_join_ratio)	-3.1356e+01	timedif:log(join_days)	1.7318e-05
log(pay_ratio)	3.1257e+01	log(total_payment):log(unpaid_months)	4.0290e-01
			p =20개, AIC = 1325.079

b) GAM을 이용한 방법

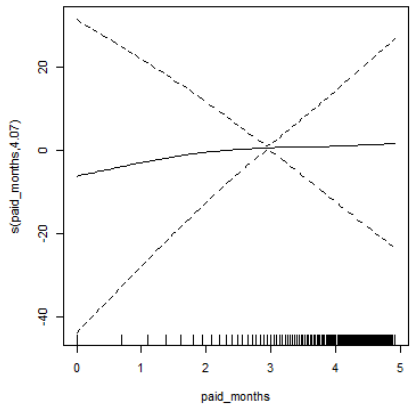
i) 고객의 향후 3개월 이내 보험계약 해지확률 $p(x_1,\cdots,x_p)$을 예측하는 적절한 GAM 모형을 찾으시오. (독립변수 중 정성변수에 대해서는 GLM을 정량변수들에 대해 GAM을 적용함)

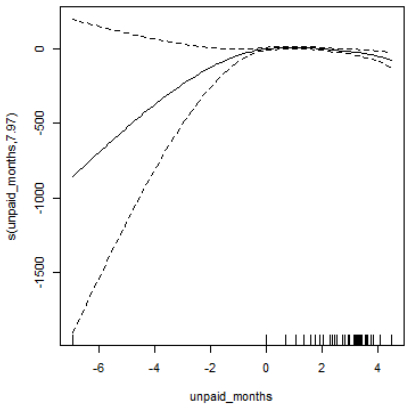
$p(x_1,\cdots,x_p) = p_r(\delta = 1|x_1,\cdots,x_p) = G(\beta_0 + \sum_{i=1}^{q}\beta_i x_i + \sum_{j=q+1}^{p} f_j(x_j))$

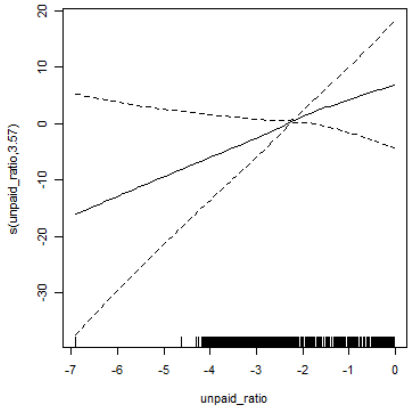
	Model	AIC	p
model1 (link=probit)	y ~ age + cycle + payment + number + timedif + no_exp + pay_end + join_days + left_end_days + pay_join_ratio + pay_ratio + paid_ratio + total_payment + payment_per + *s(paid_months) + s(unpaid_months) + s(unpaid_ratio)*	1212.416	32.60611
model2 (link=logit)	y ~ age + s(payment) + number + s(timedif) + join_days + left_end_days + join_ratio + pay_join_ratio + s(left_ratio) + pay_ratio + paid_ratio + total_payment + paid_months + unpaid_months + unpaid_ratio + payment_per	1340.133	19.01489
model3 (link=cloglog)	y ~ age + cycle + payment + number + timedif + no_exp +join_days + left_end_days + pay_join_ratio + pay_ratio + s(paid_ratio) + total_payment + s(unpaid_months) + unpaid_ratio + payment_per	1223.124	30.86864

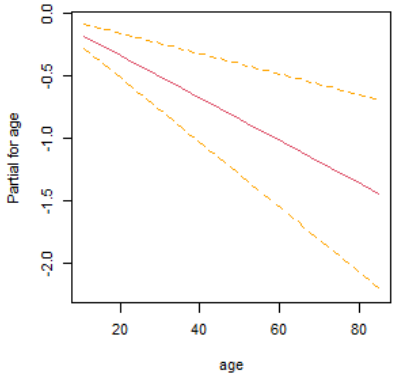
Variable	Coefficient	Variable	Coefficient
(Intercept)	-3.589897e+02	log(join_days)	9.182035e+01
age	-1.698847e-02	left_end_days	5.887087e-03
cycle3	-8.161032e+01	pay_join_ratio	-8.276335e+01
cycle6	-3.959573e+02	log(pay_ratio)	8.281348e+01
cycle12	-1.780422e+02	paid_ratio	-1.124737e+01
log(payment)	8.462376e+01	log(total_payment)	-8.639576e+01
number	1.508842e-01	log(payment_per)	2.024514e+01
timedif	-5.906692e-03	s(paid_months)	edf 4.067
no_exp1	3.455821e-01	s(unpaid_months)	edf 7.970
pay_end1	-2.879286e+00	s(unpaid_ratio)	edf 3.569
			p =32.606개, AIC = 1212.416

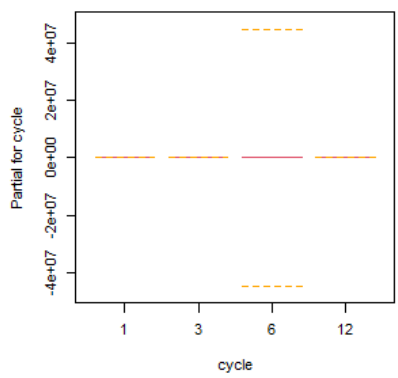
ii) 각 정량변수에 대한 최적 함수: f_j (x_j),j=q+1,⋯,p의 그래프를 그리고 그 의미를 해지확률과 연관하여 설명하시오.

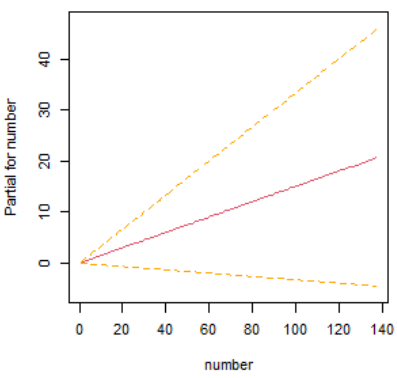


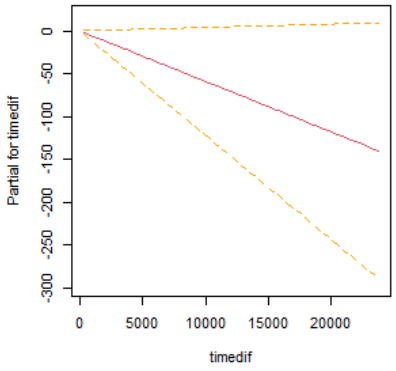


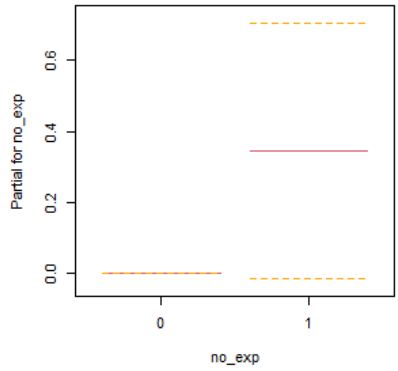


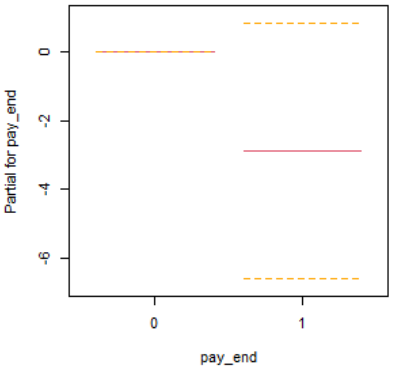


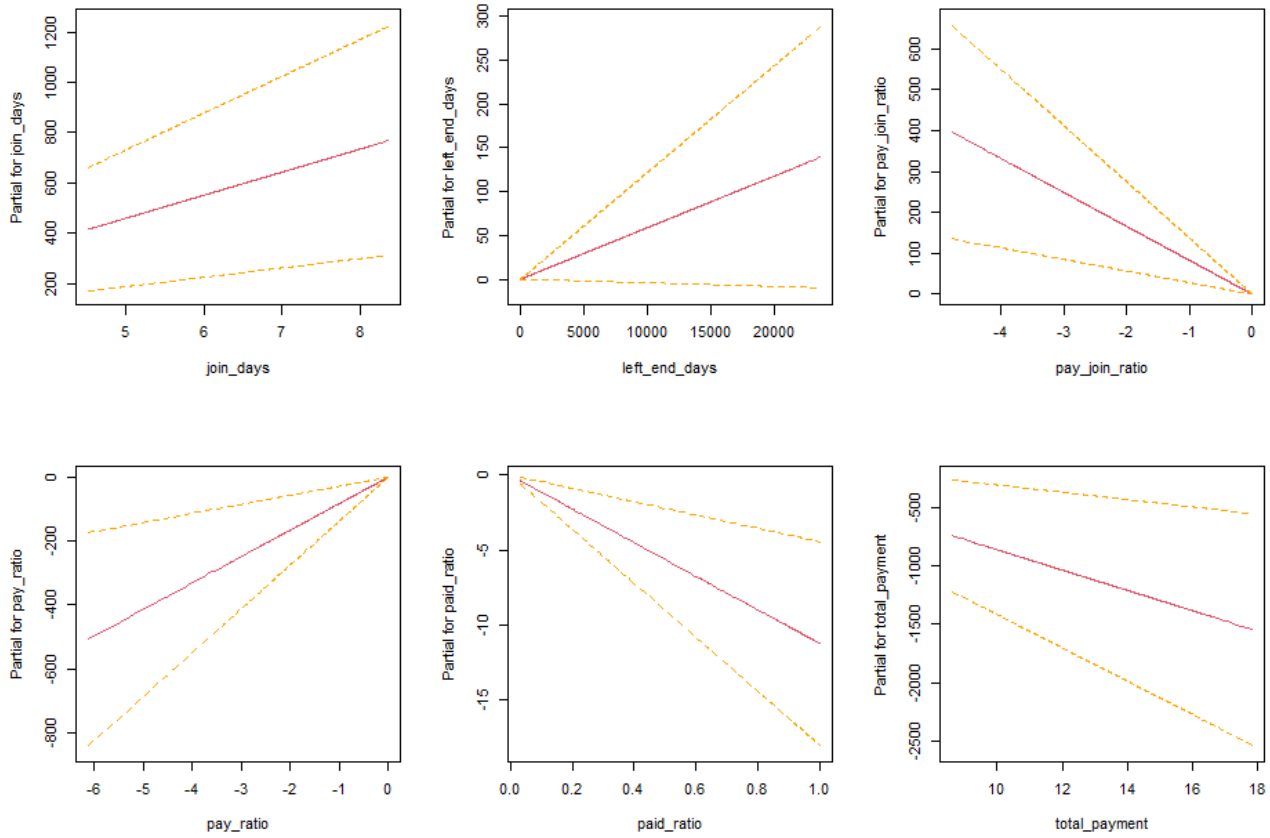












paid_months, unpaid_months, unpaid_ratio 는 해지 확률에 대해 비선형적 관계를 가지고, 나이가 적을수록 계약 만기까지 남은 날이 많을수록, 지금까지 납입한 비율이 작을수록, total_payment 가 작을수록 해지 확률이 높다.

Part 2) Survival Analysis (생존분석기법을 이용한 해지시점 예측방법)

생명보험을 가입한 각 고객에 대하여

T_i = 보험가입고객이 보험을 해지할 때까지 걸리는 시간 $(life\ time\ variable)$

C_i = 중도절단 변수 $(censoring\ variable)$

$y_i = \min(T_i, C_i)$ = 최종납입기간 = 최종납입횟수 $*$ 납입간격 (단위 개월로 환산)

$\delta_i = I(T_i < C_i)\ censoring\ indicator$ (완전 $= 1$, 중도절단 $= 0$)

$i = 1, \cdots, n$

로 각각 정의할 때 중도 절단된 자료 $(y_i; \delta_i, x_1, \cdots, x_p); i = 1, \cdots, n$을 이용하여

a) 순간 해지확률 $h(t|x_1, \cdots, x_p)$에 대한 최적 Cox PHM (Proportional Hazard Model)을 찾아보시오.

Cox PHM : $h(t|x_1, \cdots, x_p) = h(t) \cdot \exp(\sum_{j=1}^{p} \beta_j x_j)$

	Model	AIC
Cox PHM	Surv(join_months, y) ~ age + payment + timedif + join_days + left_end_days + pay_join_ratio + pay_ratio + paid_ratio + total_payment + unpaid_ratio + payment_per + paid_months + unpaid_months + unpaid_ratio:paid_months + unpaid_ratio:unpaid_months + paid_months:unpaid_months	1422.081

Variable	Coefficient	Variable	Coefficient
age	-2.818e-02	total_payment	-3.124e+01
payment	2.498e+01	unpaid_ratio	5.346e+01
timedif	-6.545e-02	payment_per	6.950e+01
join_days	5.341e+01	paid_months	-1.507e+01
left_end_days	6.543e-02	unpaid_months	-5.280e+01
pay_join_ratio	-4.301e+01	unpaid_ratio:paid_months	-6.125e+00
pay_ratio	4.297e+01	unpaid_ratio:unpaid_months	-2.837e+00
paid_ratio	-3.941e+01	paid_months:unpaid_months	5.293e+00
			p =16개, AIC = 1422.081

b) 보험가입 후 t시점에 보험계약을 유지하고 있는 고객$(x_1,\cdots,x_p)$이 향후 $(\Delta t = 3$개월$= 0.25$년$)$기간 이내에 보험을 해지할 확률은

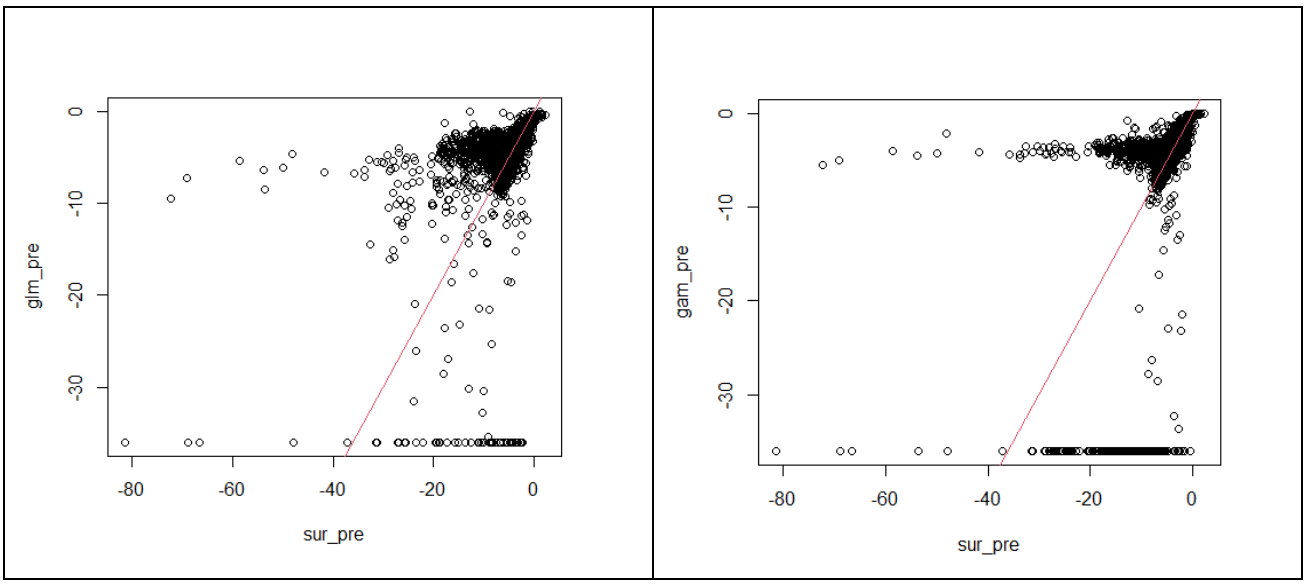
$$p^*(x_1,\cdots,x_p) = p(t < T < t+\Delta t | T > t;\ x_1,\cdots,x_p)$$
$$= 1 - S(t+\Delta t|x_1,\cdots,x_p)/S(t|x_1,\cdots,x_p)$$
$$= 1 - [S(t+\Delta t)/S(t)]^{\hat{\mu}(x_1,\cdots,x_p)}$$

$$x^*{}_j = x_j - \bar{x_j}; j = 1,\cdots,p, \hat{\mu}(x_1,\cdots,x_p) = \exp\left(\sum_{j=1}^{p}\hat{\beta_j}x_j{}^*\right), S(t) = e^{-H(t)}, H(t) = \int_0^t h(s)ds$$

이다. 따라서 Part 1, 2) 방법에서 구한 연체확률 사이에

$$\begin{cases}\ln[-ln(1-p^*)] \\ \ln[p^*/(1-p^*)]\end{cases} \cong lnp^* \cong lnh(t|x_1,\cdots,x_p) + ln\Delta t$$
$$= lnh(t) + \sum_{j=1}^{p}\beta_j x_j + ln\Delta t$$

의 관계가 근사적으로 성립한다. 위 Part 1, 2)의 분석결과를 이용해 위의 관계가 성립하는지 검토해보고 그 의미를 설명하시오.



$lnp^* \cong lnh(t|x_1,\cdots,x_p) + ln\Delta t$ 관계를 확인하기 위해 Cox-ph로 예측한 결과와 glm, gam을 통해 예측한 결과에 각각 log를 취해 plot을 그려보았다. 오른쪽 아래에 깔린 점들을 제외하고, 대부분의 점들의 해지확률이 비슷한 경향을 갖도록 예측되었다.

Part 3) 해지확률계산 및 Lift Chart를 이용한 다양한 방법의 예측력 비교

a) 위 Part 1,2) 에서 구한 최적 예측모형을 이용하여 추정용 자료(estimation data)에서 예측변수들을 이용하여 각 고객별로 I. Logistic/ GLM 해지확률: $p(x_1, \dots, x_p)$, II. GAM 해지확률, III. Cox PHM 해지확률, IV. 각 score가 상위 10%에 해당하는 상위 10%에 해당하는 500명 고객들의 지시변수를 구하시오.

I. Logistic/ GLM 해지확률: $p(x_1, \dots, x_p)$

	LINK: PROBIT	LINK: LOGIT	LINK :LOGLOG
1	$3.38 * 10^{-5}$	0.0078	0.002
2	$2.12 * 10^{-4}$	0.0127	0.006
3	$1.85 * 10^{-3}$	0.0252	0.003
4	$2.62 * 10^{-5}$	0.0690	0.009
5	$3.35 * 10^{-3}$	0.0107	0.001

II. GAM 해지확률: $p(x_1, \dots, x_p)$

	LINK: PROBIT	LINK: LOGIT
1	$4.27 * 10^{-3}$	0.0004
2	$6.43 * 10^{-3}$	0.0024
3	$1.83 * 10^{-2}$	0.0077
4	$2.22 * 10^{-16}$	0.0155
5	$1.09 * 10^{-2}$	0.0016

III. Cox PHM 해지확률:

	MODEL 1	MODEL 2
1	-170.6639	-170.6639
2	-166.4758	-166.4758
3	-157.7503	-157.7503
4	-150.9881	-150.9881
5	-155.2731	-155.2731

IV. (excel 파일 첨부)

Logistic GLM			**GAM**			**Cox PHM**		
INDEX	**해지확률**	$\hat{y}$	**INDEX**	**해지확률**	$\hat{y}$	**해지확률**	**Predict**	$\hat{y}$
890	1.0	1	890	1	1	4719	2.10	1
347	0.99	1	4660	0.998	1	4436	1.607	1
468	0.96	1	2692	0.986	1	4531	1.555	1
3599	0.82	1	4229	0.979	1	4437	1.525	1
4299	0.77	1	4719	0.968	1	4692	1.481	1

b) 위에서 구한 GLM/ GAM/ CoxPHM 확률 및 점수값의 크기순으로 전체 5000명의 고객을 10개 구간으로 나눈 후 분석용 자료를 이용하여 각 구간별 실제 해지고객의 백분율 및 Lift Chart를 서로 겹쳐서 그려보고 각 방법의 장단점을 서로 비교 검토하시오.

[GLM]

	MIN	Q1	MEDIAN	MEAN	Q3	MAX
구간1	0.08628	0.09942	0.1231	0.19079	0.21065	1
구간2	0.05648	0.06213	0.0688	0.06972	0.07697	0.08605
구간3	0.04039	0.04398	0.04768	0.04808	0.05197	0.05648
구간4	0.0271	0.03022	0.03317	0.03339	0.03657	0.04036
구간5	0.01813	0.02002	0.02204	0.02229	0.02452	0.0271
구간6	0.01161	0.01295	0.01443	0.01462	0.0163	0.01813
구간7	0.007276	0.008267	0.009313	0.009354	0.010464	0.011594
구간8	0.003959	0.004803	0.005533	0.005597	0.006435	0.007268
구간9	0.001704	0.002284	0.002864	0.002845	0.003402	0.003957
구간10	0.00E+00	3.49E-05	5.69E-04	6.54E-04	1.17E-03	1.69E-03

[GAM]

	MIN	Q1	MEDIAN	MEAN	Q3	MAX
구간1	0.05716	0.07305	0.13407	0.23446	0.29983	1
구간2	0.03641	0.03981	0.04387	0.04495	0.04962	0.05702
구간3	0.02817	0.03	0.03185	0.03193	0.03385	0.0364
구간4	0.02236	0.02359	0.02487	0.02499	0.02634	0.02817
구간5	0.0176	0.0188	0.01984	0.01993	0.02094	0.02236
구간6	0.01358	0.01458	0.01562	0.01561	0.0166	0.0176
구간7	0.009791	0.010763	0.011773	0.011755	0.012693	0.013577
구간8	0.0062	0.007011	0.007823	0.007903	0.008749	0.009789
구간9	0.002446	0.003492	0.004429	0.004364	0.005238	0.006195
구간10	0.00E+00	0.00E+00	0.00E+00	6.62E-04	1.42E-03	2.44E-03

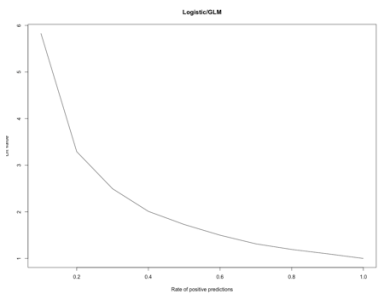
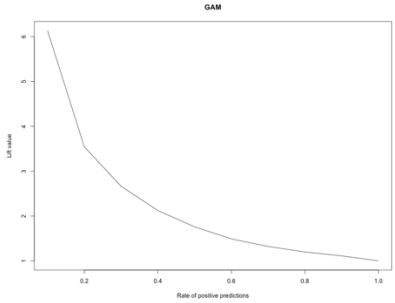
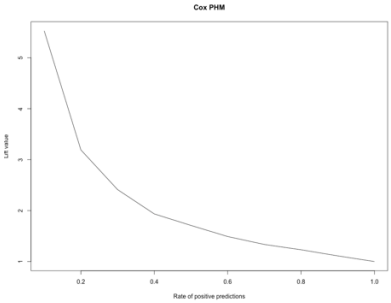
[Cox PHM]

	MIN	Q1	MEDIAN	MEAN	Q3	MAX
구간1	-1.5828	-1.3297	-0.9788	-0.7885	-0.3731	2.1044
구간2	-2.171	-2.047	-1.929	-1.909	-1.778	-1.585
구간3	-2.607	-2.509	-2.404	-2.4	-2.292	-2.171
구간4	-2.988	-2.88	-2.791	-2.796	-2.708	-2.61
구간5	-3.4	-3.307	-3.196	-3.199	-3.096	-2.99
구간6	-3.852	-3.736	-3.63	-3.627	-3.521	-3.401
구간7	-4.379	-4.233	-4.108	-4.109	-3.977	-3.855
구간8	-5.162	-4.916	-4.714	-4.729	-4.535	-4.38
구간9	-7.431	-6.308	-5.822	-5.95	-5.462	-5.167
구간10	-80.488	-15.936	-11.589	-14.329	-9.258	-7.437

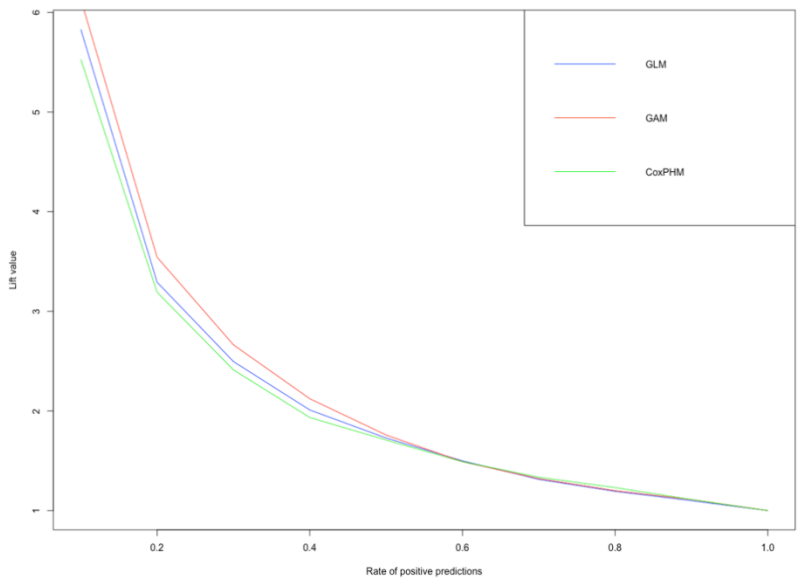
구간별 백분율 값

	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
GLM	**23.2**	**3.0**	**3.6**	**2.2**	**2.4**	**1.4**	**0.8**	**1.4**	**1.4**	**0.4**
GAM	**24.4**	**3.8**	**3.6**	**2.0**	**1.2**	**0.6**	**1.2**	**1.4**	**1.6**	**0.0**
CoxPHM	**22.0**	**3.4**	**3.4**	**2.0**	**3.2**	**1.6**	**1.6**	**2.0**	**0.6**	**0.0**

구간별 lift 차트

[Logistic]	GAM	Cox PHM
		

c) 위에서 구한 표를 그래프로 그린 Lift Chart를 서로 겹쳐서 그려보고 각 방법의 장단점을 서로 비교 검토하시오.



d) Part 1,2)에서 선택한 GLM/GAM/Cox PHM 예측 모형의 AIC값을 각각 제출하시오.

GLM	GAM	Cox PHM
1325.079	**1212.416**	1422.081

e) 검정용 자료 (test data ; 5001 – 10000)에 대하여 Logistic GLM/GAM 해지확률 및 COX PHM score가 상위 10% 해당하는 고객의 지시변수를 excel file로 제출하시오.

f) 위에서 사용한 3가지 방법 외에 아래의 다양한 Data Mining 기법을 이용한 분석을 추 가하여 Lift Chart를 서로 겹쳐서 그려보고 각 방법들의 장단점을 서로 비교 검토해 보시오.

g) 검정용 자료 (test data ; 5001-10000)에 대하여 LDA, k-NN, SVM, Random Forest, Tensor Flow 등을 이용한 해지확률이 상위 10% 해당하는 고객들의 지시변수 들을 excel file 로 추가로 제출하시오.

Part 4) (생명보험 해지확률 자동계산 Application 개발)