Project 9

Survival Analysis Cox PHM/ALT 모형을 이용한 심장병 발생 예측

강아미, 고유정, 윤보인, 이혜린, 홍지원

Contents

Part 1.

Cox PHM을 이용한 심장병발생시간 예측

1. Cox PHM을 이용한 심장병발생시간 예측

a) 조사 시작시점의 공변량이 $(x_1, x_2, \dots, x_p)$인 환자의 t 시점이후의 순간 심장병 발생율 $\mathrm{h}(t|x_1, \dots, x_p)$에 대한 최적 Cox PHM

Cox Model	AIC
coxph(formula = Surv(followup, chdfate) ~ sbp + dbp + scl + age + bmi + sex + ages + ages2 + sbp:scl + sbp:age + sbp:ages + dbp:scl + dbp:ages2 + scl:ages + age:bmi + age:sex + bmi:ages2)	22738.23

Variable	변수 효과	Variable	변수 효과
bmi	1.827	age:bmi	-3.8
sex1	0.5026	sbp:scl	-7.912
sbp:age	10.28	dbp:scl	8.627

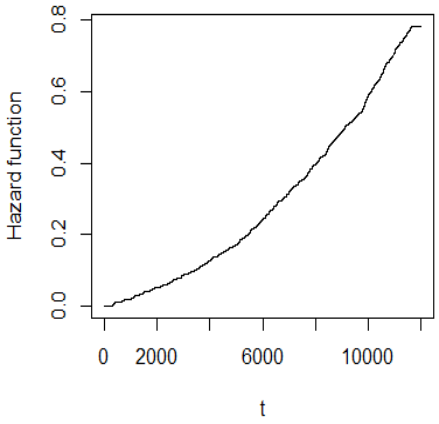
ages : 나이 변수를 10년 간격으로 그룹화 한 변수

ages2 : 나이 변수를 45세 기준으로 그룹화 한 변수

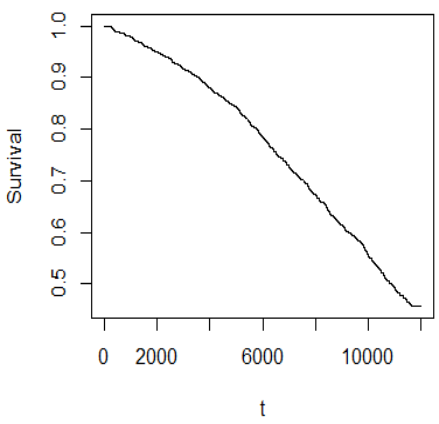
1. Cox PHM을 이용한 심장병발생시간 예측

b) 최적 Cox PHM에 대한 환자의 cumulative hazard function, survival function의 그래프를 그리고 Weibull, Gompertz 등 분포가 적합한지 Q-Q plot 등을 이용하여 검토하시오.

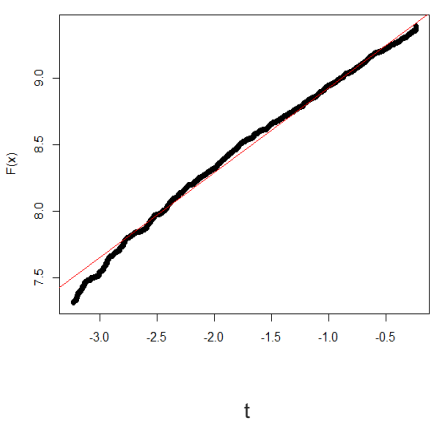
t vs Cumulative Hazard Risk



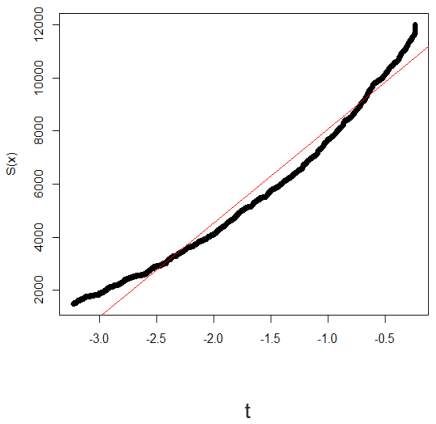
t vs Survival



Weibull Q-Q plot



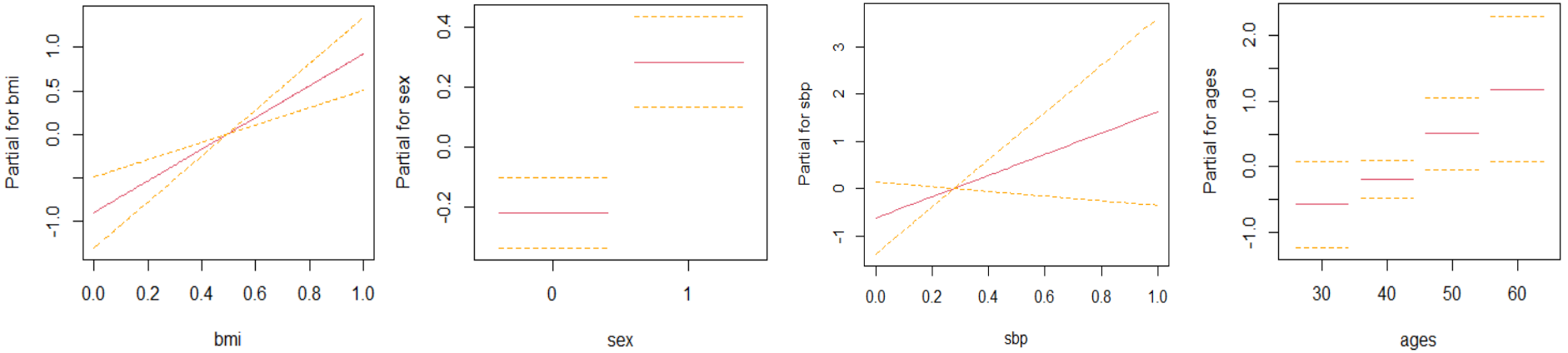
Gompertz Q-Q plot



Gompertz 분포보다 Weibull 분포가 더 적합해보인다.

1. Cox PHM을 이용한 심장병발생시간 예측

c) 위에서 추정된 최적 Cox PHM의 모수 $\widehat{\beta_1}, \ldots, \widehat{\beta_p}$ 및 $\widehat{S}(\,.\,)$를 이용하여 심장병 발생에 유의한 영향을 주는 주요 변수들의 효과를 각각 그래프로 그리고 그 의미를 설명하시오.



sbp(수축기혈압)와 bmi, 그리고 환자 개인의 연령대를 의미하는 ages가 커질수록 심장병 발생 확률이 증가한다.

sex는 1(=male)일 때가 0(=female)일 때보다 심장병 발생 확률이 높다.

→ 수축기혈압과 BMI지수, 그리고 연령대가 높을수록 심장병 발생확률이 증가한다.

→ 남성이 여성보다 심장병 발생확률이 높음을 관찰할 수 있다.

Part 2.

ALT 모형을 이용한 심장병 발생시간 예측

2. ALT 모형을 이용한 심장병 발생시간 예측

a) 정성변수에 대한 가변수 도입; 정량변수에 대한 변수변환 $(1/x^2, \frac{1}{x}, ln\,x, x, x^2)$ 추가 여부; 교호작용 추가 여부; 적절한 분포 $G(t)$선택, 변수선택 등을 적절히 고려한 환자의 심장병 발생시간 T의 누적분포함수에 대한 최적 ALT

ALT Model	AIC
survreg(formula = Surv(followup, chdfate) ~ sbp + age + bmi + sex + ages + sbp:dbp + sbp:scl + sbp:age + sbp:ages + dbp:scl + ages:scl + ages:bmi + age:sex)	31754.48

ages : 나이 변수를 10년 간격으로 그룹화 한 변수

ages2 : 나이 변수를 45세 기준으로 그룹화 한 변수

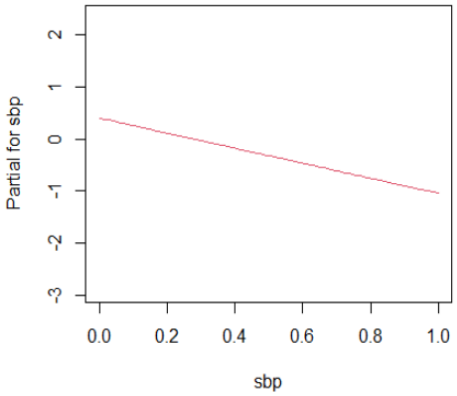
2. ALT 모형을 이용한 심장병 발생시간 예측

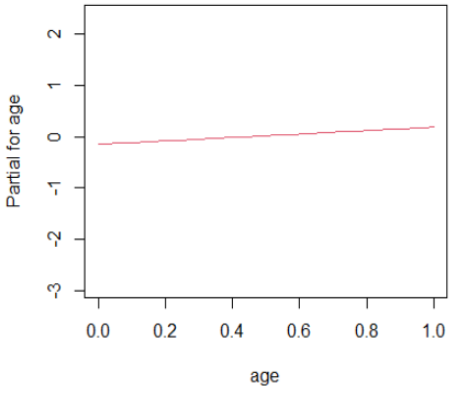
a) 정성변수에 대한 가변수 도입; 정량변수에 대한 변수변환 $(1/x^2, \frac{1}{x}, ln\,x, x, x^2)$ 추가 여부; 교호작용 추가 여부; 적절한 분포 $G(t)$선택, 변수선택 등을 적절히 고려한 환자의 심장병 발생시간 T의 누적분포함수에 대한 최적 ALT

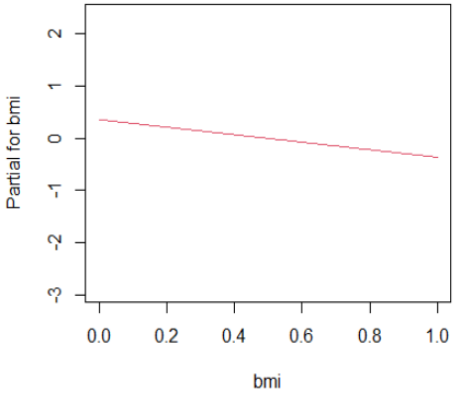
Variance	Coefficient	Variance	Coefficient
sbp	-1.45	sbp:age	-4.65
bmi	-0.72	sbp:ages40	2.04
sex1	-0.33	sbp:ages50	2.40
ages50	-0.79	sbp:ages60	3.75
ages60	-1.23	dbp:scl	-3.20
sbp:dbp	0.95	ages40:scl	-1.28
sbp:scl	1.05		

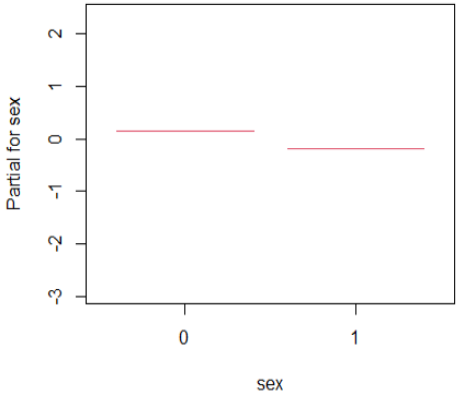
2. ALT 모형을 이용한 심장병 발생시간 예측

b) 위에서 추정된 최적 ALT 모형의 모수 $\widehat{\beta_1}, \ldots, \widehat{\beta_p}, \hat{\sigma}$ 및 분포 $G(\,.\,)$를 이용하여 심장병 발생에 유의한 영향을 주는 주요변수들의 효과를 각각 그래프로 그리고 그 의미를 설명하시오.









sbp와 bmi의 값이 커질수록 심장병 발생까지 걸리는 시간이 짧아졌다.

age의 값이 커질수록 심장병 발생까지 걸리는 시간이 미미하게 증가했다.

sex는 1(=male)일 때가 0(=female)일 때보다 심장병 발생까지 걸리는 시간이 짧았다.

→ 수축기혈압과 BMI지수가 높을수록 여성보다 남성이 심장병 발생까지 걸리는 시간이 짧다.

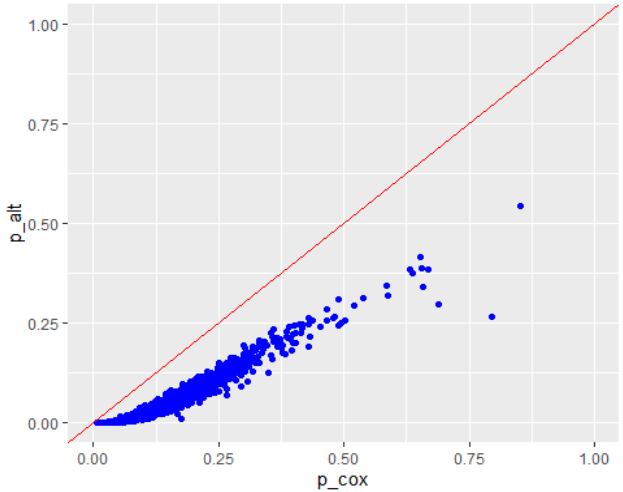
Part 3.

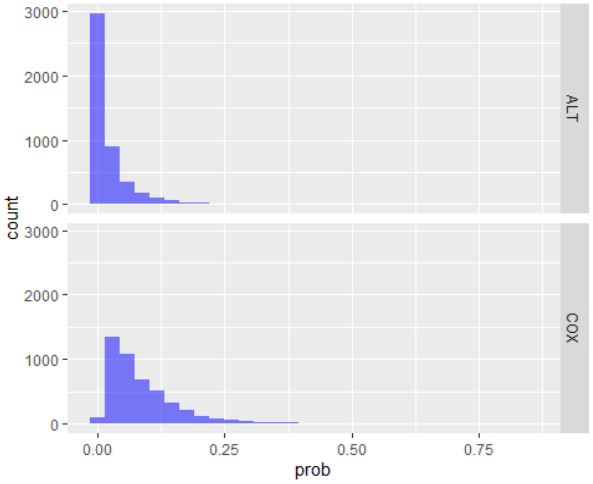
Cox PHM/ ALT 모형을 이용한 10년 내 심장병 발생가능성 예측

3. Cox PHM/ ALT 모형을 이용한 10년내 심장병 발생가능성 예측

위 두가지 예측모형을 이용하여 환자가 10년 내에 심장병이 걸릴 확률을 각각 구하여 차이를 비교해 보고 두 방법의 장단점을 검토하시오.

Cox PHM; ALT 모형





Cox는 ALT에 비해 데이터가 고르고 넓게 분포했고 0값이 출력되는 경우가 적었다.

→ ALT 장점) COX PHD에 비해 과대적합을 방지한다.

→ COX PHD 장점) ALT에 비해 0값과 같은 극값의 출현 빈도가 적다.

3. Cox PHM/ ALT 모형을 이용한 10년내 심장병 발생가능성 예측

위 두가지 예측모형을 이용하여 환자가 10년 내에 심장병이 걸릴 확률을 각각 구하여 차이를 비교해 보고 두 방법의 장단점을 검토하시오.

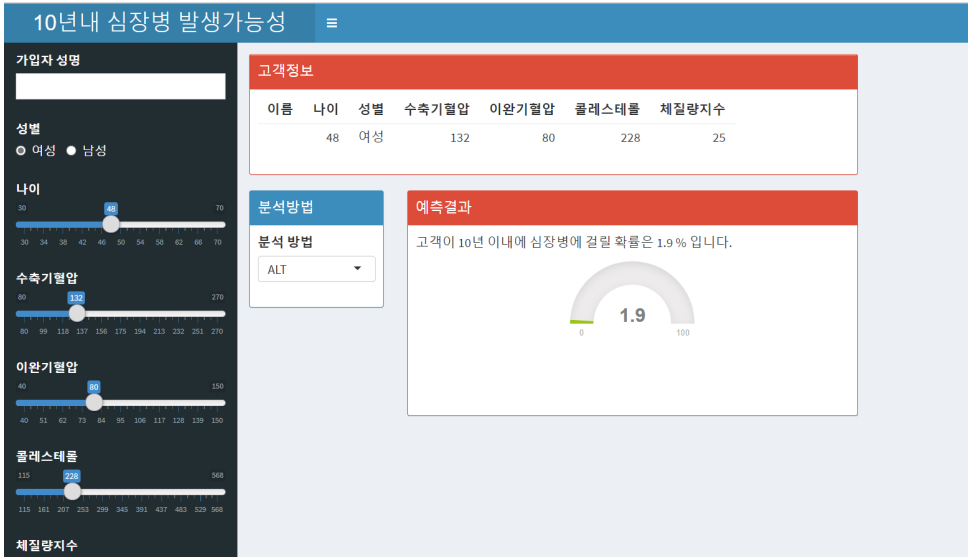
Part 1, 2에서 선택된 최적 PHM/ALT 모형의 AIC

PHM 모델	ALT 모델
22738.23	31754.48

3. Cox PHM/ ALT 모형을 이용한 10년내 심장병 발생가능성 예측

위 두가지 예측모형을 이용하여 환자가 10년 내에 심장병이 걸릴 확률을 각각 구하여 차이를 비교해 보고 두 방법의 장단점을 검토하시오.

위에서 구한 최적 모형을 이용하여 10년내 심장변 발생 화률을 계산하는 앱 개발(링크 첨부)



Part 4.

새 반응변수에 대한 Cox-PHM, ALT 최적모형 비교분석

4. 새 반응변수에 대한 Cox-PHM, ALT 최적모형 비교분석

a) 각 환자에 대하여 탄생 후 심장병 발생할 때까지 걸리는 시간 $T_i^* = age * 365.25 + followup$
새 반응변수에 대한 Cox PHM, ALT 최적 모형을 구하고 이를 Part1,2에서 구한 최적 모형과 비교하여 차이점 및 장단점을 검토하시오.

Cox Model	AIC
coxph(formula = Surv(t, chdfate) ~ sbp + dbp + scl + bmi + sex + ages + ages2 + sbp:scl + sbp:ages + sbp:ages2 + dbp:scl + dbp:sex + dbp:ages2 + scl:ages + bmi:ages + bmi:ages2 + sex:ages2)	22770.8

Cox PHM			
Variable	변수 효과	Variable	변수 효과
sbp	3.77	sbp:scl	-7.47
bmi	0.9477	dbp:scl	8.34
Sex	0.9074	sex:ages	0.24

4. 새 반응변수에 대한 Cox-PHM, ALT 최적모형 비교분석

a) 각 환자에 대하여 탄생 후 심장병 발생할 때까지 걸리는 시간 $T_i^* = age * 365.25 + followup$
새 반응변수에 대한 Cox PHM, ALT 최적 모형을 구하고 이를 Part1,2에서 구한 최적 모형과 비교하여 차이점 및 장단점을 검토하시오.

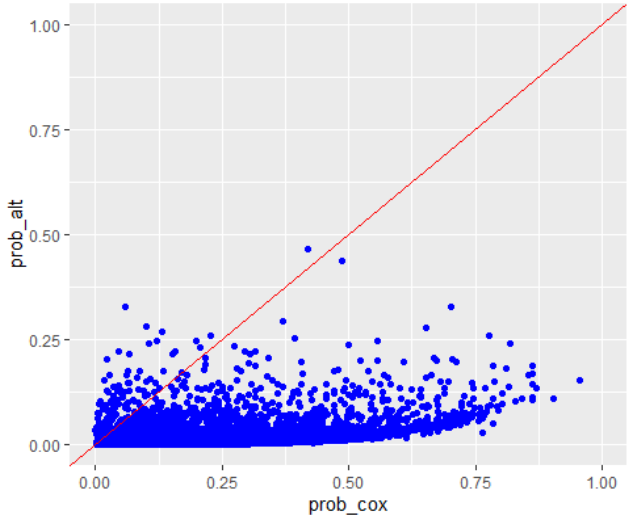
ALT	AIC
survreg(formula = Surv(age * 365.25 + followup, chdfate) ~ sbp + bmi + sex + ages2 + sbp:scl + sbp:ages + sbp:ages2 + dbp:scl + dbp:sex + dbp:ages2 + scl:ages + bmi:ages + bmi:ages2 + sex:ages2	31765.52

Cox PHM			
Variable	Coefficient	Variable	Coefficient
sbp	-2.12	sex1:dbp	0.47
bmi	-0.48	ages22:dbp	1.55
Sex1	-0.56	scl:ages40	-1.09
ages22	-0.45	scl:ages60	-1.48
sbp:scl	5.58	sex1:ages22	-0.15
scl:dbp	-6.17		

Part1, 2에서 도출한 결과와 유사한 것을 확인할 수 있다.

4. 새 반응변수에 대한 Cox-PHM, ALT 최적모형 비교분석

b) 새 반응변수 T_i^*에 대한 Cox PHM/ ALT 모형을 이용한 10년 내 심장병 발생확률을 구하고 이를 Part 3 결과와 비교 검토하시오.

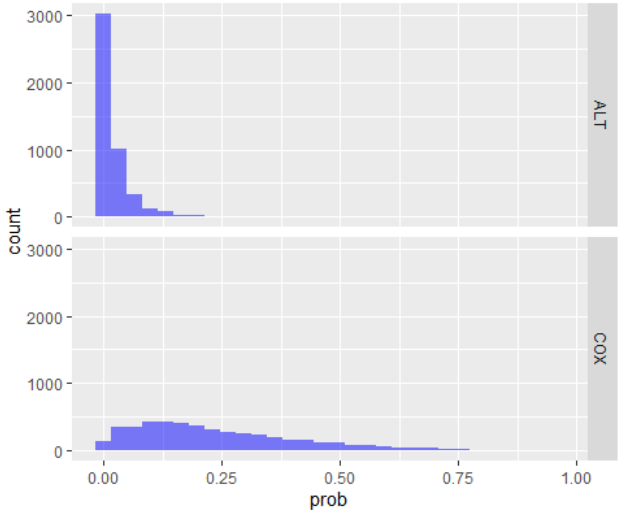


Part3와 마찬가지로 COX PHD의 확률값이 ALT에 비해 매우 높은 경향을 보인다.

ALT의 확률값은 대체로 0.25이하의 값을 가지는 등 part3의 경우보다 낮았다.

4. 새 반응변수에 대한 Cox-PHM, ALT 최적모형 비교분석

b) 새 반응변수 T_i^*에 대한 Cox PHM/ ALT 모형을 이용한 10년 내 심장병 발생확률을 구하고 이를 Part 3 결과와 비교 검토하시오.



Cox는 ALT에 비해 데이터가 고르고 넓게 분포했고 0값이 출력되는 경우가 적었다.

→ ALT 장점) COX PHD에 비해 과대적합을 방지한다.

→ COX PHD 장점) ALT에 비해 0값과 같은 극값의 출현 빈도가 적다.


Thank you