이론통계학 II

Project #9. Survival Analysis

발표일 2021. 11. 24

1조

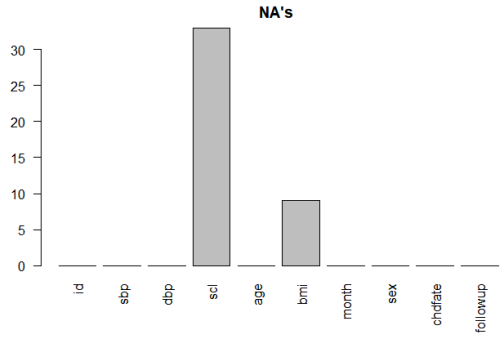
202STG26	박지윤
202STG27	이수현
212STG04	김이현
212STG12	박윤정

Part 0 : 데이터 변수 설명 및 전처리

Framingham Heart Study (1948~1998)

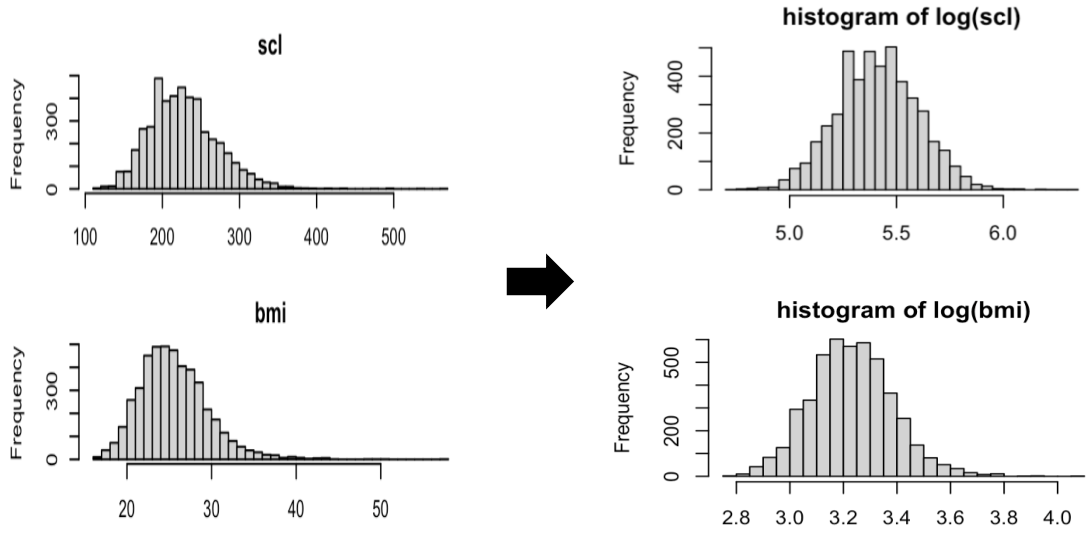
변수명	설명
sbp	수축기혈압
dbp	이완기혈압
age	나이
scl	혈청 콜레스테롤
bmi	비만도
sex	성별
month	측정 시작된 해의 달
id	환자 id
followup	측정이 시작된 후로 심장병 발생까지 일수
chdfate	심장병 발병 1=발병(uncensored) 0= 발병하지 않음(censored)

① NA(결측치)



분석할 데이터는 총 4699명의 환자에 대한 데이터이다. 이 중 41개의 관측치에 대해서 결측치가 존재하였다. 주로 scl, bmi 변수가 NA였다. 비만과 콜레스테롤은 개인에 따라 상이하고 심장병에 중요한 변수이다. 또한 전체 데이터의 약 1%이므로 다른 값으로 대체하기는 보다는 해당 41개의 관측치를 삭제하였다.

② 변수 변환

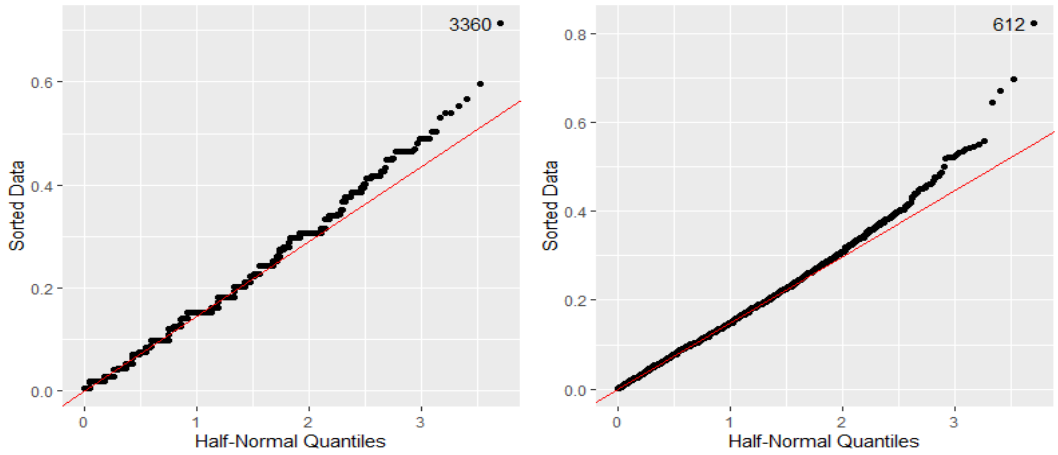


정량변수들의 히스토그램을 살펴보니 sbp, dbp, scl, bmi 4개의 변수들이 모두 오른쪽 꼬리가 긴 분포를 보였다. $1/x^2$, $1/x$, $log(x)$, x^2과 같이 다양한 변수변환을 고려한 결과 4변수 모두 로그변환 하였을 때 오른쪽 꼬리가 긴 문제가 완화되었다. 또한 변수마다 값의 범위가 다르기에 로그변환을 통해 모두 일의 단위로 맞추어 주었다.

Sex는 성별을 나타내는 더미변수이므로 factor로 변환해주었다. Month는 단순히 연속형보다는 범주형 변수로 보는 것이 적합하므로 factor로 변환해주었다. 추가적으로 age의 경우 해당 데이터에서 30세~68세까지 존재하였다. 해석에 용이함과 연령대별 효과를 확인하기위해 30대, 40대, …, 60대이상으로 범주화 시킨 후 factor로 변환하였다.

모형에 적합하기전 모든 정량변수들은 평균으로 빼주어 mean centering하였다.

③ Outlier(이상점)



R의 halfnorm 함수를 사용하여 각 연속형 변수(sbp, dbp, scl, bmi)의 이상점을 찾았다. 더 나은 예측을 위해 위 그래프와 같이 정규성에서 크게 벗어나는 총 7개의 행을 제거하였다.

Part 1 : Cox PHM 을 이용한 심장병발생시간 예측

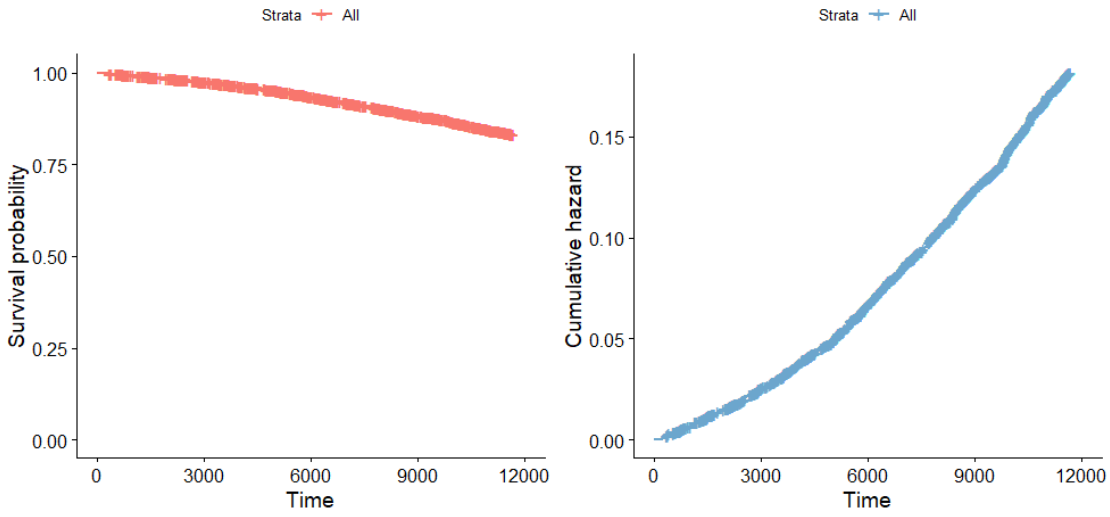
① Cox PHM

교호작용	변수 선택	선택된 변수 개수	AIC
X	X	19	22708.04
X	O	7	22698.88
O	X	132	22811.39
O	O	21	22688.9

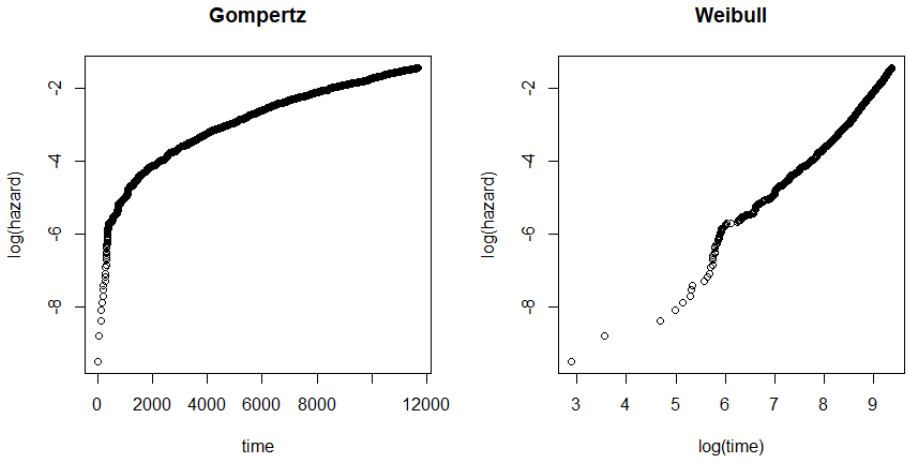
선택된 변수
sbp , dbp , scl , age40 , age50 , age60 , bmi , sex1 , sbp:dbp , sbp:scl , sbp:age40 , sbp:age50 , sbp:age60 , dbp:scl , dbp:age40 , dbp:age50 , dbp:age60 , dbp:bmi , scl:age40 , scl:age50 , scl:age60

모든 변수들의 교호작용 유무와 변수선택(stepwise)을 고려하여 Cox PHM 모형에 적합시켰다. 그 중 AIC를 기준으로 선택된 모형은 교호작용을 포함하여 변수선택을 진행한 모형이다. 해당 모형은 총 21개의 변수가 포함되었고 AIC는 22688이다. 기존의 7개의 변수는 month는 선택되지 않았다. 나머지 6개의 원변수는 모두 포함되었으며 그 중 몇몇의 변수들 간의 교호작용이 포함되었다.

② Cumulative hazard function & Survival function

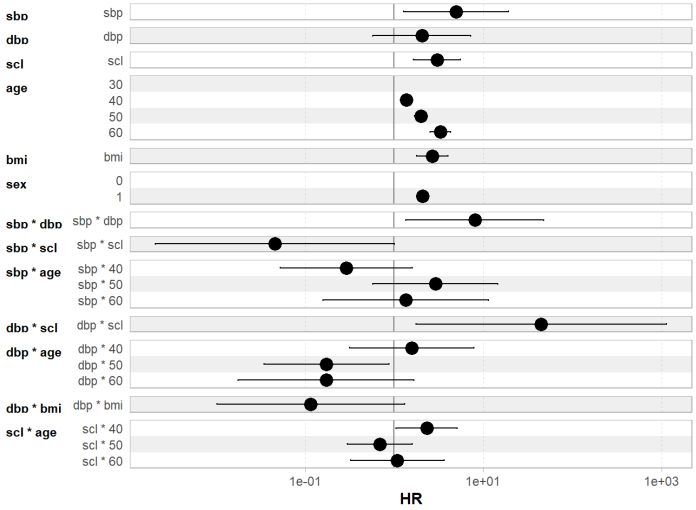


위에서 선택된 최종 Cox모형의 Survival function과 Cumulative hazard function이다. 일반적으로 Survival function은 감소하고 Hazard function은 증가한다는 사실과 일치하는 것을 확인할 수 있다. 즉, 시간이 지날수록 심장병에 걸릴 확률이 커진다.



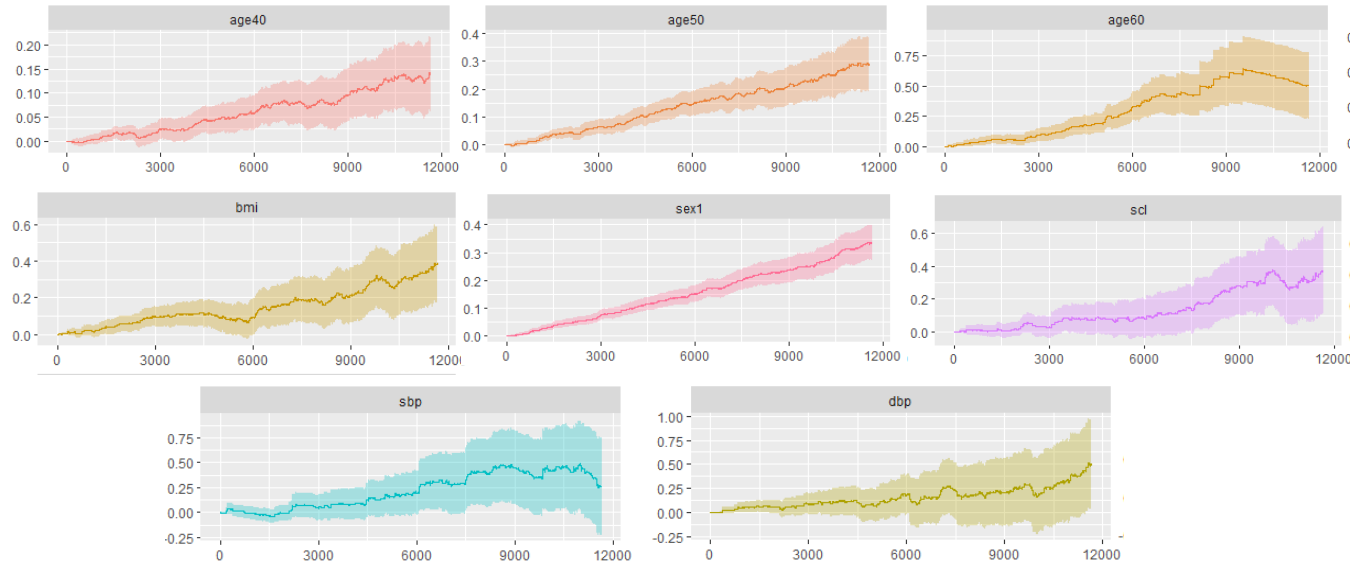
또한 baseline hazard rate function 추정을 위해 적절한 분포 Q-Q plot으로 살펴보았다. 주로 생존분석에 사용되는 Gompertz, Weibull 분포의 Q-Q plot 보면 Gompertz는 왼쪽 꼬리부분이 급격하게 떨어지면서 곡선의 형태를 보인다. 반면, Weibull은 왼쪽 꼬리 부분이 살짝 벗어나기는 하나 전체적으로 직선의 형태를 보인다. 따라서 해당 데이터의 baseline hazard rate function에는 Weibull모형이 적합하다.

③ 주요변수 효과



최종 Cox모형의 변수들의 coefficient를 나타낸 그래프이다. 점은 해당 변수의 exp(coef)를 의미하고 선은 신뢰구간을 의미한다. 그래프의 중간 실선은 1로 1을 기준으로 이보다 작으면 음의 영향, 크면 양의 영향이다. 1차항들을 살펴보면 sbp, dbp, scl, age, bmi, sex 모두 1보다 크므로 해당 변수값이 커지면 심장병이 발생할 위험이 증가한다.

아래 그래프는 변수별 시간에 따른 hazard rate function의 영향을 나타낸 것이다. 나이의 경우 30대보다 40대, 50대, 60대가 위험이 컸고 위험의 증가세는 연령대가 높을수록 더 컸다. 60대의 경우 시간이 9000이후에는 감소하는 형태를 보인다. 이는 약 24년 후이므로 측정 시점에 60대였던 사람들이 24년 이후에는 심장병이나 기저질환으로 인해 이미 사망하였을 가능성이 크다. 이러한 이유로 해당 시점이후에 감소하는 형태를 보이는 것으로 예상된다. 또한 아래 그래프에서 y축 값을 비교해보면 Bmi(비만도), scl(콜레스테롤)은 비슷한 수준의 영향을 미치고, 혈압과 관련된 sbp, dbp가 비슷한 수준의 영향을 미치는 것을 확인할 수 있다. 성별의 경우 여성에 비해 남성이 시간이 지남에 따라 더 위험하다. 해당 데이터는 2000년대 이전으로 그 시절에는 남성이 여성보다 많은 흡연과 음주, 사회생활로 인한 영향으로 예상된다.

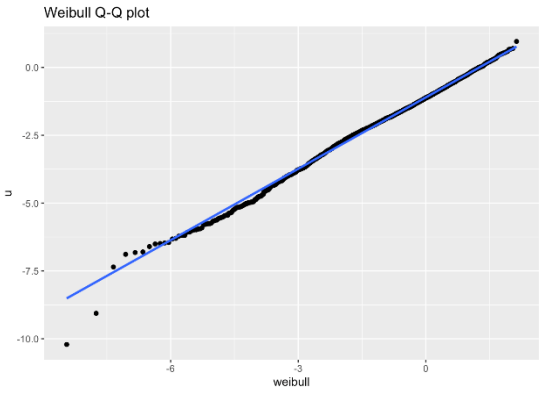


Part 2 : ALT 를 이용한 심장병발생시간 예측

① ALT

분포	교호작용	선택된 변수 개수	AIC
Weibull	X	9	31676.78
	O	23	31667.84
Log-logistic	X	9	31685.83
	O	23	31675.53
Log-normal	X	9	31803.48
	O	23	31785.15
Logistic	X	9	32057.24
	O	22	32042.48

선택된 변수
sbp, dbp, scl, age40, age50, age60, bmi, sex1, sbp:dbp, sbp:scl, sbp:age40, sbp:age50, sbp:age60, dbp:scl, dbp:age40, dbp:age50, dbp:age60, dbp:bmi, scl:age40, scl:age50, scl:age60

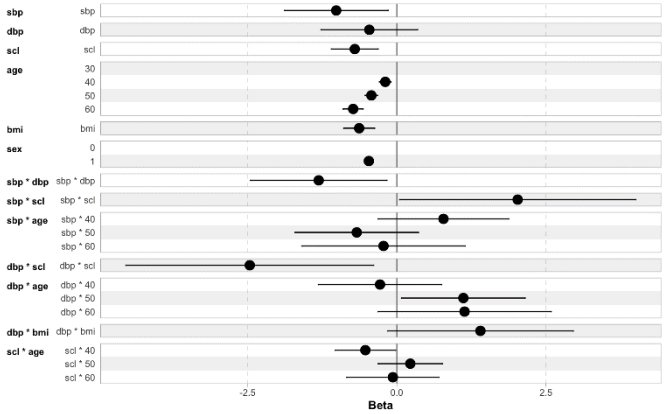


모든 변수들의 교호작용 유무와 변수선택(stepwise)을 고려하여 4가지 분포에 대하여 ALT를 적합시켰다. 그 중 AIC를 기준으로 선택된 모형은 교호작용을 포함하여 변수선택을 진행한 Weibull 모형이다. 해당 모형은 총 23개의 변수가 포함되었고 AIC는 31667이다. 기존의 7개의 변수는 month는 선택되지 않았다. 나머지 6개의 원변수는 모두 포함되었으며 그 중 몇몇의 변수들 간의 교호작용이 포함되었다. 또한 선택된 변수들은 Cox PHM과 동일하였다.

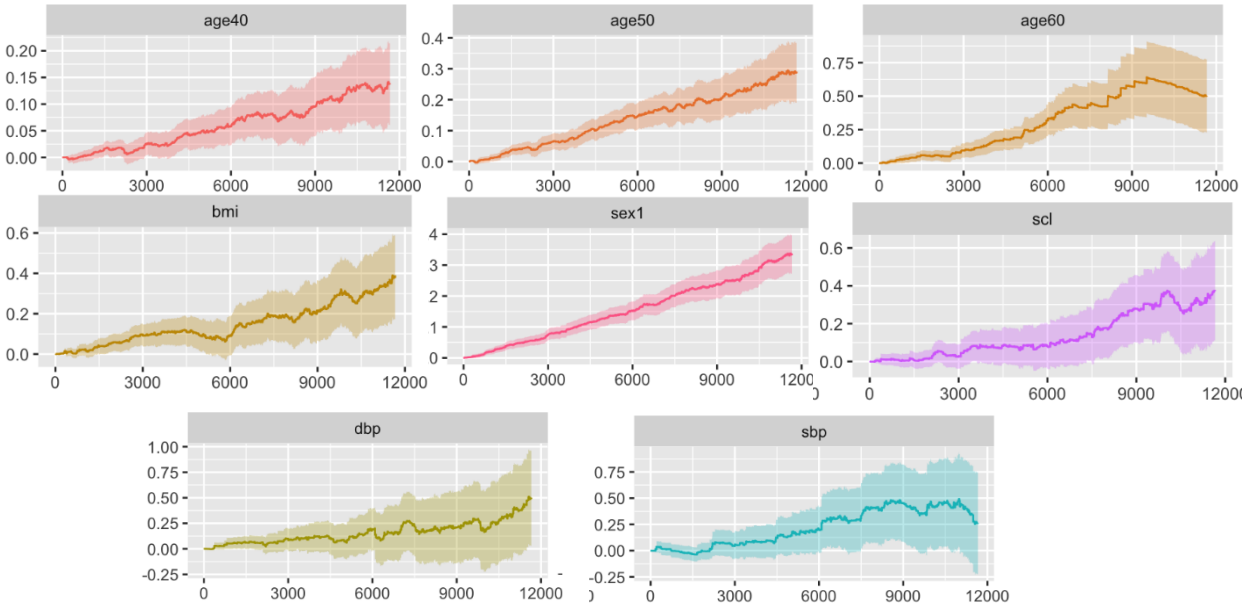
왼쪽은 최종 모형의 Q-Q plot을 그린 것이다. 매우 높은 적합도로 직선에 가깝다. 따라서 최종 모형이 타당함을 확인할 수 있다.

② 주요변수 효과

최종 ALT모형의 변수들의 coefficient를 나타낸 그래프이다. 점은 해당 변수의 coefficient를 의미하고 선은 신뢰구간을 의미한다. 그래프의 중간 실선은 0이다. Coefficient가 음수이면 $u = lnt - (\beta_0 + \sum \beta_j x_j)/\sigma$이므로 u가 커져 양의 영향의 미친다. 1차항들을 살펴보면 sbp, dbp, scl, age, bmi, sex 모두 음수로 해당 변수값이 커지면 심장병이 발생할 위험이 증가한다.

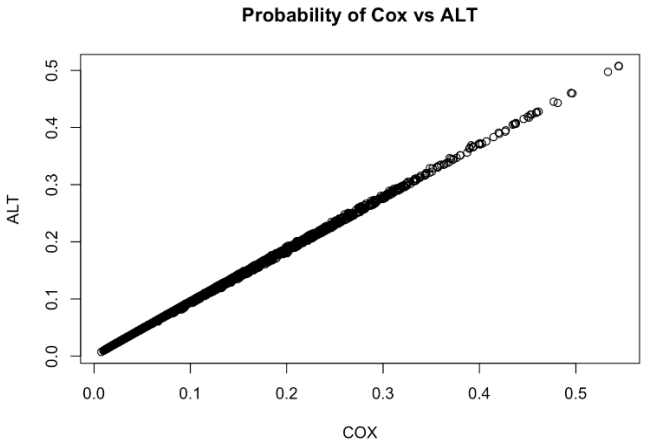


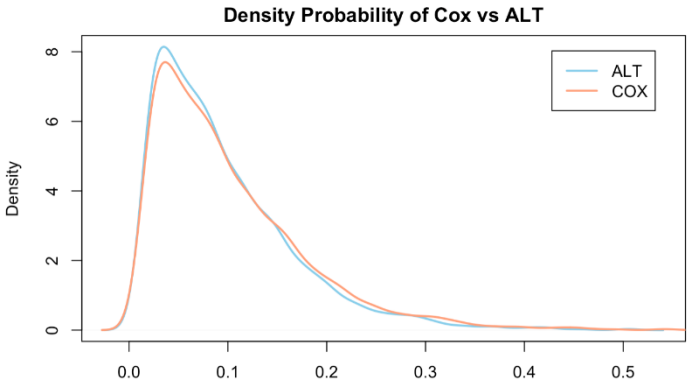
아래 그래프는 변수별 시간에 따른 hazard rate function의 영향을 나타낸 것이다. 나이의 경우 30대보다 40대, 50대, 60대가 위험이 컸고 위험의 증가세는 연령대가 높을수록 더 컸다. Cox에서와 동일하게 60대의 경우 시간이 9000이후에는 감소하는 형태를 보인다. 또한 아래 그래프에서 서로 관련 있는 bmi(비만도), scl(콜레스테롤)이 비슷한 수준의 영향을 미치고, 혈압과 관련된 sbp, dbp가 비슷한 수준의 영향을 미치는 것을 확인할 수 있다. 성별의 경우 여성에 비해 남성이 시간이 지남에 따라 더 위험하다. 전반적으로 Cox에서의 변수별 효과와 매우 비슷한 양상을 보인다.



Part 3 : Cox / ALT 모형을 이용한 10년 내 심장병 발생 가능성 예측

① 10년 내에 심장병 발생 확률





앞에서 적합한 Cox와 ALT 모형으로 환자들이 10년(3650일) 내 심장병 발생 확률을 예측하였다. 왼쪽은 Cox와 ALT를 이용했을 때 각각의 예측된 확률의 산점도이다. 산점도에서 점들은 y=x 위에 존재하므로 두 예측값이 거의 동일함을 확인할 수 있다. 또한, 오른쪽은 두 모형에서의 예측 확률의 분포를 나타낸 것이다. 두 density가 거의 겹쳐져 있다. 따라서 Cox와 ALT 모형을 이용했을 때 최종 모형에 사용된 변수도 동일하고 이를 이용하여 예측한 확률도 동일하다.

Cox의 경우 분포 없이 적합과 예측이 가능하다는 장점이 있다. ALT는 데이터에 적절한 분포를 찾아주어야 한다. 가정하는 분포가 적합해야 예측도 정확할 것이다. 해당 분석에서는 Cox와 Weibull 분포를 사용한 ALT가 동일한 결과를 보이는 것을 통해 데이터에 적합한 분포를 사용하였고 분석이 타당했음을 알 수 있다.

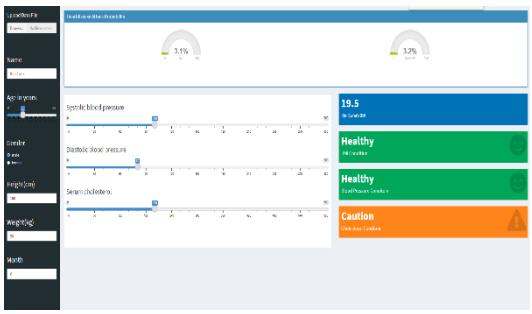
② AIC 비교

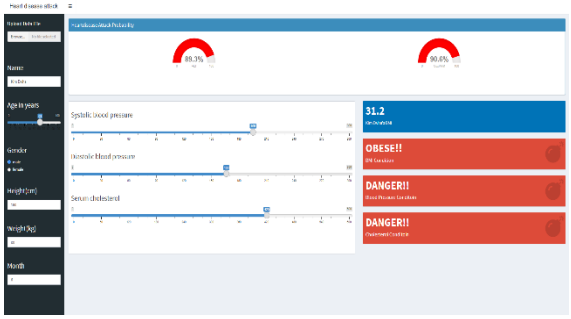
	COX	ALT
변수 개수	21	23
AIC	22688.9	31667.84

최종 Cox 모형에서 선택된 변수 개수는 21개이고 AIC는 22688.9이다. 최종 ALT 모형에서 선택된 변수 개수 또한 동일하게 21개이며 ALT의 경우 intercept(β_0)와 scale(σ)을 추가적으로 추정하므로 p=23이다. ALT에서 AIC는 31667.84이다.

③ R Shiny

https://soohyeonlee.shinyapps.io/Heart_Disease/





성별, 키, 몸무게, 나이, 혈압, 콜레스테롤을 입력하면 COX와 ALT를 이용하여 10년 내에 심장병 발생 확률을 보여주는 어플을 만들었다.

Part 4 : 새 반응변수

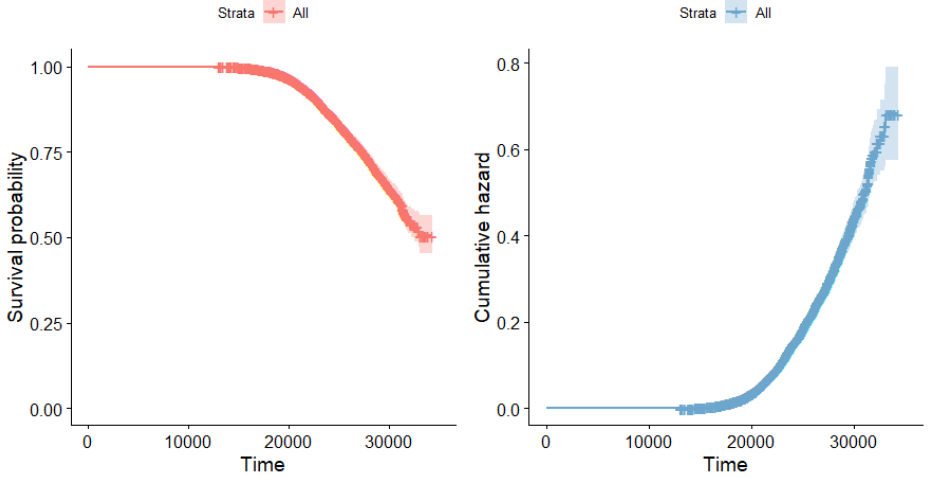
앞서 분석에서는 반응변수를 관측 시점부터 심장병이 발생한 날까지의 일수인 'followup'으로 설정하였다. 반응변수를 다음과 같이 age와 followup을 이용하여 탄생 후 심장병이 발생한 날로 새롭게 정의하여 분석해보았다.

$$T_i^* = age \times 365.25 + followup$$

① COX

교호작용	변수 선택	선택된 변수 개수	AIC
X	X	16	22372.74
X	O	5	22361.86
O	X	81	22450.63
O	O	9	22355.24

Age를 반응변수로 활용하고 age를 제외한 sbp, dbp, scl, bmi, month, sex 6개 변수를 설명변수를 하여 적합하였다. AIC 기준으로 선택된 최종 모형은 교호작용과 변수선택을 진행한 모형이다. 주요 변수 효과로는 성별과 bmi가 영향이 컸다. 남성이 여성보다 위험율이 컸고 bmi가 클수록 위험율이 커졌다. Survival function과 Cumulative hazard function은 다음 그림과 같았다. 탄생 후 약 16000일, 44년 이후 심장병 발생 위험이 급격하게 증가한다.



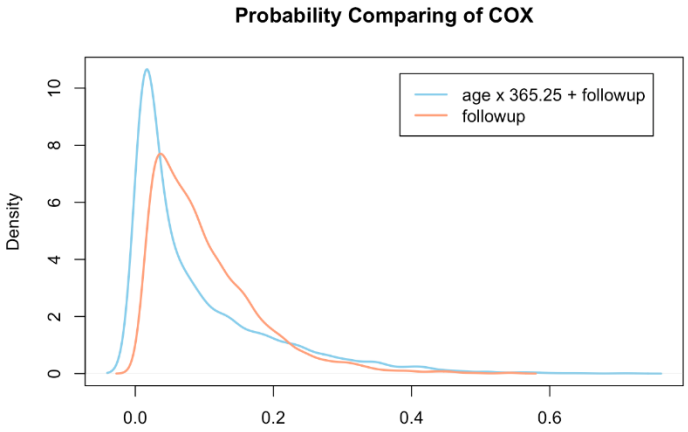
② ALT

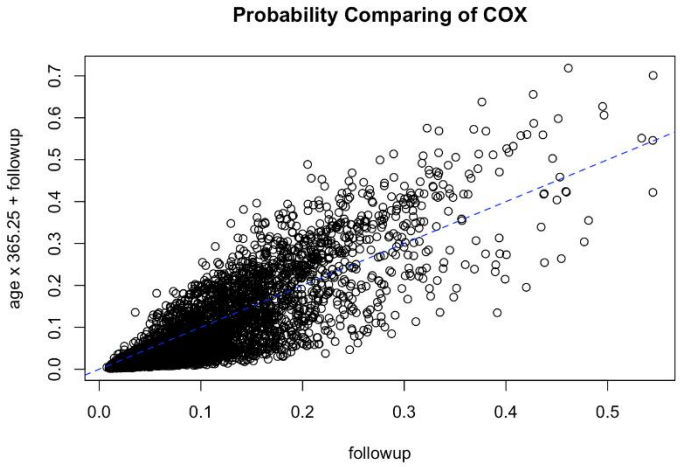
분포	교호작용	AIC
Weibull	X	32137.02
	O	32128.77
Log-logistic	X	32057.44
	O	32044.81
Log-normal	X	32006.75
	O	31991.20
Logistic	X	32193.71
	O	32180.79

Cox와 동일하게 새로운 반응 변수로 ALT를 적합하였다. AIC 기준으로 선택된 최종 모형은 교호작용을 포함한 log-normal 분포를 이용한 모형이다.

③ 10년 내에 심장병 발생 확률

Cox 모형에서 Baseline hazard rate function을 추정하기 위해 Weibull과 Gompertz Q-Q plot을 살펴보았다. Weibull 분포가 비교적 직선에 가까웠으므로 Weibull 분포를 이용하였다. 추정한 baseline hazard rate function을 이용하여 10년 심장병 발생할 확률을 구하였다. 이 결과를 앞서 followup을 반응변수를 하였을 때의 예측 결과와 비교해보았다. 예측 확률의 분포를 보면 Followup을 반응변수보다 새로운 반응변수가 왼쪽(0)에 더 치우쳐 있다. 오른쪽의 산점도를 보면 확률이 약 0.2 이상에서는 두 모형의 추정치가 y=x 근방에 분포한다. 하지만 확률이 작은 경우 y=x 밑에 점들이 더 많이 분포하므로 확률이 작을 때는 새로운 반응 변수가 더 작게 발생 확률을 추정한다.





ALT 모형에서 10년 이내 심장병이 발생할 확률을 추정하였다. 앞서 진행한 followup을 반응변수로 한 결과와 새로운 반응변수에서의 결과를 비교해보았다. Cox에서와 마찬가지로 새로운 반응변수를 이용한 모형의 예측 확률 분포가 왼쪽으로 치우쳐 있다. 이 차이는 Cox보다는 적었으며 산점도에서도 비교적 y=x 근방에 대칭적으로 분포하였다.