모형 불확실성에 따른 random survival forest와 Cox 비례위험모형의 예측오차 비교

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
엄 지 윤

모형 불확실성에 따른 random survival forest와 Cox 비례위험모형의 예측오차 비교

지도 남 정 모 교수

이 논문을 석사 학위논문으로 제출함

2016년 12월 일

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
엄 지 윤

엄지윤의 석사 학위논문을 인준함

심사위원 남 정 모 인

심사위원 정 인 경 인

심사위원 송 기 준 인

연세대학교 대학원

2016년 12월 일

감사의 글

어느덧 2 년이라는 대학원 생활을 마무리하는 시기가 찾아왔습니다. 대학교를 갓 졸업한 제가 하나하나 배우며 성장하는 과정이 힘들고 바쁘기도 하였지만 얻는 것 또한 많았던 보람된 시간이었습니다. 많은 분들의 도움 덕분에 이 자리까지 올 수 있었습니다. 이 자리를 빌어 감사의 마음을 드리고자 합니다.

먼저 바쁘신 와중에도 꼼꼼히 논문을 지도해주시며 부족했던 저에게 진심 어린 조언과 애정을 주신 남정모 교수님, 항상 많은 것을 전해주려 하시고 온화한 미소로 맞아주시는 정인경 교수님, 제가 바르게 성장할 수 있도록 지식뿐 만 아니라 세세한 부분까지도 살펴주신 송기준 교수님, 모두 진심으로 감사 드립니다.

그리고 대학원 생활에 잘 적응할 수 있도록 이끌어주신 선배님들, 2 년이라는 시간을 동거동락하며 서로에게 힘이 되어주며 든든한 친구가 된 동기 명균오빠와 경민이, 새로운 길을 찾아가는 덕성오빠, 저를 잘 믿고 따라와주며 필요할 때 마다 많은 도움을 준 해린이와 지현이, 묵묵히 자신의 몫을 잘 해가고 있는 종수와 선민씨, 웃음꽃을 피워 준 하영과 하나, 많은 시간을 함께 하지 못해 아쉬운 근우씨와 호재씨에게도 감사의 마음 전합니다.

또한 옆에서 많은 조언과 도움을 주신 진우 선생님, 바쁜 저를 항상 이해해줘서 미안한 마음이 더 큰 나연, 민지, 지현, 각자 길을 찾아 열심히 살아가고 있을 원요, 수빈, 기태, 민혜, 송이, 저를 잊지 않고 찾아주는 아쎌 식구들에게도 감사합니다.

무엇보다 든든한 버팀목이 되어주고 아낌없는 지원과 사랑을 주시는 부모님, 멀리서도 응원해주는 평생 친구 동생 소연이. 사랑하고 감사하다는 말 전하고 싶습니다.

언급하진 못했지만 이외에도 저에게 힘찬 응원가 격려를 아끼지 않으시고 도움를 주셨던 많은 분들께도 감사의 마음를 드리고 싶습니다. 감사합니다.

2016 년 12 월

엄지윤 올림

차 례

표 차 례

국 문 요 약

모형 불확실성에 따른 random survival forest와 Cox 비례위험모형의 예측오차 비교

빅데이터 시대가 도래하면서 대용량의 정보들과 이를 처리하기 위한 데이터 마이닝 방법들이 임상 분야에 활발히 적용되고 있다. Random survival forest는 붓스트랩과 각 마디의 split 단계에서 무작위성을 부여하여 생존확률을 예측하는 마이닝 기법으로, 특별한 분포나 통계학적 가정이 전제되지 않아도 분석 가능하며 처리 속도가 빠르다는 장점 때문에 최근 각광받고 있다. 하지만 실제 임상 자료에서 random survival forest와 Cox 비례위험모형의 예측 정확도는 유사하였다. 본 연구의 목적은 두 방법의 예측력이 비슷한 이유를 다양한 자료 구조의 시뮬레이션 환경에서 찾고자 한다. 독립변수 내 상관성을 고려하여 여러 가지 경우의 참모형을 통해 생존 자료를 생성하고, 이차항과 교호작용, noise 변수 유무를 고려한 4가지 분석모형으로 random survival forest, Cox 비례위험모형, Lasso-Cox 모형을 비교하였다. 비교 척도로는 Harrell의 concordance index와 integrated brier score를 사용하였다.

모의실험 결과 분석모형이 참모형과 일치하는 경우 random survival forest는 Cox 비례위험 모형과 Lasso-Cox 모형에 비해 C-index가 낮고, IBS는 높아 예측력이 다소 떨어졌다. 반면, 이차항의 효과나 독립변수 간의 교호작용이 존재함에도 불구하고 이들을 포함하지 않고 독립변수의 주효과만으로 분석 모형에 적합하였을 경우, Cox 비례위험모형과 Lasso-Cox 모형는 Random survival forest보다 C-index가 낮았다. 특히 IBS는 2배 이상 높아져 예측 오차가 크게 증가하였다. 이에 반해 Random survival forest는 C-index와 IBS의 값이 크게 달라지지 않고 일정한 크기로 나타나 복잡한 참모형에서 분석모형이 부적합되더라도 어느 정도 예측력이 유지되

었다. noise 변수가 추가된 경우에도 동일한 양상을 보였다.

빅데이터에서 자료 구조가 복잡할 경우, 자료의 참모형을 파악하기 어렵기 때문에 분석모형이 부적합될 위험성이 증가한다. 따라서 random survival forest가 분석모형을 잘못 적합하여도 예측력을 일정하게 보존할 수 있다는 측면에서 안정적이고, 빅데이터 시대에서 유용하게 사용될 수 있는 방법이라 기대된다.

핵심되는 말 : Random survival forest, Cox 비례위험모형, Lasso-Cox 모형, noise

제 1장 서론

1.1 연구 배경 및 목적

최근 빅데이터(big data) 시대가 도래하면서 임상 분야에서는 이를 활용하여 더 나은 의사결정을 내리고자 한다. 기존에 가장 많이 사용되었던 분석방법은 Cox의 비례위험모형이다. 사건이 발생할 위험을 예측하고 사건에 유의한 영향을 주는 요인을 찾아 그 효과를 추정하기 용이하지만 비례위험 가정이 전제되어야 하고 시간에 대해서 비선형 효과가 존재할 경우, 모형의 변환이나 구조행렬(design matrix)를 확장시켜야 한다. 또한 빅데이터에서 모수를 파악하는 것은 어렵기 때문에 추정의 정확성은 감소할 수 밖에 없다. 이러한 이유로 대용량의 정보를 처리하기 위한 데이터 마이닝 방법들이 임상 연구에 활발히 적용되고 있다.

그 중 하나인 random survival forest(Ishwaran 외 2008)는 붓스트랩(bootstrap)과 각 마디(node)의 분리 단계에서 무작위성을 부여하여 생존 확률을 예측하는 마이닝 기법으로, 특별한 통계학적 가정이 전제되지 않아도 분석 가능하다는 장점이 있다. Random survival forest의 바탕인 random forest(Brieman 2001)는 여러 개의 모형을 선형 결합하는 앙상블 기법이다. 기본 예측모형인 tree는 각 마디에서 분리에 사용할 후보변수들을 무작위로 선택하고, 가지치기(pruning)를 수행하지 않는다는 점에서 기존의 CART(Brieman 외 1984)와는 다르다. 따라서 최대로 성장한 tree로 이루어진 random forest는 CART를 이용한 경우보다 분산과 오차가 감소한다(Ishwaran 외 2010). 주로 SNP나 microarray 자료에 적용되었던 random forest는 중도절단 자료에도 이용되기 시작했다. 생존 자료를 분석하기 위해 고안된 random survival forest는 반복적인 분할과 무작위성 부여를 통해 독립변수의 선형 효과는 물론이고 자료에 내포된 비선형 관계와 교호작용을 자동으로 처리하여 예측력이 향상된다. 또한 고차원 구조(high-dimensional)의 자료를 분석할 경우 변수 선택을 어떻게 할 것인지, 그리고 적정 마디 크기와 후보 분리변수 선정에 관하여 논의되

었다(Ishwaran 외 2011). 나아가 경쟁위험(competing risk)모형을 적용하기 위해 random survival forest를 확장하여 새로운 앙상블 추정치를 소개하였다(Ishwaran 외 2014).

실제 자료를 적용한 연구 결과에서 random survival forest와 Cox의 비례위험모형의 예측력 차이는 크지 않았다. 23개의 변수로 이루어진 279명의 유방암 환자의 자료(Ture, Tokatli, and Kurt 2008)를 여러 기준에 따라 마디를 분리한 4가지 경우의 random survival forest와 Cox의 비례위험모형의 오차는 비슷하였다(Imran 2009). 머리와 목의 편평 세포 암종 환자 1371명 자료를 위와 동일한 방법으로 비교한 결과도 다르지 않았으며(Datema 2012), 수축성 심부전증 환자 2231명으로 이루어진 코호트 자료를 분석한 연구에서도 마찬가지로 두 모형의 예측 정확도는 유사하였다(Hsich 2011).

대개 자료 구조의 이해가 선행되지 않은 채 분석이 진행되고, 빅데이터에서 특히 독립변수가 많은 경우 생존 시간에 영향을 미치는 변수 선정과 관계를 파악하는 것은 어렵다. 또한 모형이 부적합(misspecification)하면 예측력이 감소하며 예측 결과는 불안정해질 수 밖에 없다. 본 연구의 목적은 실제 임상 자료를 분석한 random survival forest와 Cox 모형의 예측력이 비슷한 이유를 자료 구조가 복잡할 경우 random survival forest와 Cox의 비례위험모형, Lasso-Cox 모형의 예측력을 다양한 모의실험 환경에서 비교한 결과에서 찾고자 한다.

1.2 연구 내용 및 방법

본 연구에서는 독립변수 내 상관성을 고려하여 여러 가지 경우의 참모형을 통해 생존 자료를 생성한다. 이차항과 교호작용, noise 변수 유무를 고려한 4가지 분석모형으로 random survival forest, Cox 비례위험모형, Lasso-Cox 모형을 비교한다. Lasso-Cox 모형은 독립변수의 수가 많은 자료에 적용되는 방법이므로 비교모형에 포함하였다. 모형의 성능을 비교하기 위한 지표로 Harrell의 concordance index와

integrated brier score를 사용한다. 전자는 예측 확률을, 후자는 예측 오류를 평가하는 척도로써 두 지표 모두 random survival forest와 다른 모형 간의 예측력을 비교할 수 있다. 추가적으로 변수 선택과 관련하여 Cox의 비례위험모형과 Lasso-Cox 모형은 공변량의 회귀계수를 통하여 요인의 효과를 추정하고, random survival forest는 VIMP와 minimal depth로 요인의 중요성을 파악한다.

1.3 논문의 구성

제 1장에서는 연구의 배경 및 목적, 연구 내용과 방법을 소개하였고, 제 2장에서는 Cox의 비례위험모형, Lasso-Cox 모형과 random survival forest의 개념을, 제 3장에서는 제 2장에서 소개한 세 모형의 비교하기 위한 지표, Harrell의 concordance index와 integrated brier score에 대해 설명하였다. 제 4장은 모의실험 설계 방법과 random survival forest, 모수모형인 Cox의 비례위험모형과 Lasso-Cox 모형을 여러 상황에서 비교한 결과를 제시하였다. 마지막으로 제 5장은 모의실험 결과를 토대로 본 논문에 대한 결론 및 고찰에 대하여 논의하였다.

제 2장 이론적 배경

2.1 Cox의 비례위험모형 (Cox’s proportional hazard model)

관심 있는 사건이 발생하기까지의 시간을 T라 할 때, 임의의 시점 t에서 사건이 발생할 순간 위험률을 위험함수(hazard function)라 하고, 생존함수(survival function)과 확률밀도함수(probability density function)로 표현된다.

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \le T \le t + \Delta t \mid T \ge t)}{\Delta t} = \frac{f(t)}{S(t)}$$

누적위험함수(cumulative hazard function)는 $H(t) = \int_0^t h(u)du$로 정의되기 때문에 생존함수는 다음과 같다.

$$S(t) = \exp\left[-\int_0^t h(u)du\right] = \exp[-H(t)]$$

본 논문에서는 관심 있는 사건을 사망이라 가정하고, 생존 예측 모형은 생성하기 위해 다음의 기호를 따른다. n명의 생존 자료에서 개체 i의 생존 시간을 T_i, 우중도 절단 시간을 C_i라 하였을 때, 관찰된 생존 시간 $\tilde{T}_i$는 T_i와 C_i의 최솟값으로 정의할 수 있으며, $\Delta_i = \mathrm{I}\{T_i \le C_i\}$로 중도절단 지표이다. 중도 절단 되기 전 개체가 사망했다면 Δ_i는 1, 그렇지 않으면 0이다. $\boldsymbol{X}_i^{\mathrm{T}} = (\boldsymbol{X}_i^1, \cdots, \boldsymbol{X}_i^P)$는 P개의 독립변수들로 구성된 공변량 행벡터를 뜻하고 $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_P)^{\mathrm{T}}$는 해당 모수의 열벡터를 의미할 때, 개체 i의 위험함수는 아래와 같이 정의할 수 있다.

$$h(t|\boldsymbol{X}_i) = h_0(t)\exp(\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta})$$

위 식에서 $h_0(t)$는 독립변수 $\boldsymbol{X}_i$의 값이 모두 0일 때, t시점에서의 위험함수이고 이것을 기저위험함수(baseline hazard function)라 한다. Cox의 비례위험모형은 기저 분

포의 영향을 받지 않고 위험함수로 표현되는 회귀모형이며, 모든 독립변수는 시간이 변하더라도 공변량에 대한 위험률의 비가 항상 일정하다는 비례위험 가정을 만족해야 한다. 추정한 회귀계수는 $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1, \cdots, \widehat{\boldsymbol{\beta}}_P)^{\mathrm{T}}$로 나타낼 수 있고, 사건이 발생할, 즉 사망할 위험은 공변량 X_i^j가 한 단위 증가할 때마다 $\exp(\hat{\beta}_j)$만큼 증가하며, 시간과 기저위험함수 $h_0(t)$에 따라 변하지 않는다.

2.2 Lasso-Cox 모형

독립변수의 수가 많거나 변수 간 상관관계가 존재할 경우, 회귀모형의 계수가 과대추정 되는 것을 방지하기 위해 shrinkage 방법을 사용한다. 대표적인 방법이 Lasso penalty(Tibshirani 1996)이다. 본 논문에서는 shrinkage와 변수 선택이 모두 가능한 Lasso penalty를 부여하여 회귀계수를 추정한 Lasso-Cox 모형(Tibshirani 1997)을 비교 방법으로 추가하였다.

2.1절에서 정의한 생존 자료에서 Cox 비례위험모형의 공변량 모수 열벡터 $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_P)^{\mathrm{T}}$는 penalty로 하여금 과대 추정된 추정치를 축소할 수 있다. Cox 모형의 로그 편우도함수(log partial likelihood)를 $l(\boldsymbol{\beta})$라고 정의한다면,

$$\widehat{\boldsymbol{\beta}}^{lasso} = \operatorname{argmax}\left[l(\boldsymbol{\beta}) - \lambda P_\alpha(\boldsymbol{\beta})\right]$$

Lasso-Cox 모형의 회귀계수는 위의 식과 같이 추정된다. 여기서 $P_\alpha(\boldsymbol{\beta}) = \sum_{j=1}^{P}|\beta_j|$는 Lasso penalty이고, $\sum_{j=1}^{P}|\beta_j| \le s$를 만족해야 한다. s와 λ는 크기에 따라 추정치가 달라지는 조절 모수이므로 적절한 값으로 선택해야 한다. 만약 일부 변수의 회귀계수가 0으로 축소된다면 결과적으로 해당 변수는 제거가 되는 것이다. 따라서, Lasso-Cox 모형은 $\sum_{j=1}^{P}|\beta_j| \le s$ 하에서 Lasso penalty를 부여한 로그 편우도함수를 최대화하는 λ 중 최솟값을 채택하여 회귀계수를 추정한다(Tibshirani 1997).

2.3 Random Survival Forest

2.3.1 알고리즘

Random survival forest는 생존 자료를 분석하기 위해 random forest를 발전시킨 것으로 random forest의 알고리즘을 기본으로 한다(Ishwaran 외 2008).

단계 1. 주어진 자료로부터 $B(b = 1, \cdots, B)$개의 붓스트랩 표본을 생성한다. 생성한 각각의 붓스트랩 표본의 일부를 out-of-bag(OOB)라 칭하고 이들을 제외한 in-bag 붓스트랩 표본(in-bag-data)으로 모형을 생성한다.

단계 2. In-bag 붓스트랩 표본에서 survival tree를 성장시킨다. 각 마디에서 p개의 후보변수를 무작위를 골라, 이 중에서 자식 마디의 동질성이 최대가 되는 변수를 선택하여 최적의 분리가 발생되는 지점을 찾는다. 정지기준(stopping criterion)에 도달할 때까지 이 과정을 반복하며 마디를 분리해나간다.

단계 3. 마디가 더 이상 분리되지 않은 지점에 도달하면, 그 마디를 끝마디(terminal node)라 한다. 그리고 tree의 끝마디에서 얻은 정보를 결합하여 앙상블 예측모형을 얻는다.

마디 간 차이는 생존 시간이 다름을 의미하므로, 분리규칙은 log-rank 검정 통계량이 최대가 되는 변수와 지점을 찾아 생존 차이(survival difference)를 극대화하는 것이다. 마디의 불순도(impurity)는 생존 여부와 시간과 관련이 있으므로 자료에는 생존 시간과 우중도 절단 여부에 대한 정보가 반드시 포함되어야 한다. Survival tree는 위와 같은 분리규칙을 반복해서 적용하여 한 개의 마디를 두 개의 마디로 나누는 이진분리(binary split)를 따르며, CART와 달리 최대로 성장한다. 생존 차이를 최대화하여 마디를 분리하는 과정에서 유사하지 않은 생존 정보를 가진 개체들은 다른 마디로 할당된다. Tree는 최소 $d_0 > 0$의 사망 사건이 끝마디에 포함될 때까지 성장한다. 결과적으로 최대로 성장한 tree의 끝마디 간 이질성과 끝마디 내 동질성은 높아진다(Ishwaran 외 2008).

B개의 tree를 생성한 다음 단계는 앙상블 누적위험함수(ensemble cumulative hazard function)을 추정하는 것이다. 추정함수에서 사용하는 기호는 다음과 같다. b번째 in-bag 붓스트랩 표본에서 형성된 survival tree의 h번째 끝마디에 $N(h)$번의 사망 사건이 포함되어 있다면, 관측된 시점을 $\tilde{T}_{1,h} < \tilde{T}_{2,h} < \cdots < \tilde{T}_{N(h),h}$으로 표현한다. 이 때 $d_{s,h}$를 시점 s까지 사망이 발생한 횟수, $\tilde{Y}_{s,h}$를 시점 $t_{s,h}$에서 위험 상태에 있는 개체 수라 한다면, h번째 마디의 누적위험함수는 Nelson-Aalen의 추정치로 아래의 식과 같고, 동일한 마디에 포함된 개체들은 모두 같은 누적위험함수를 따른다.

$$\hat{H}_h(t) = \sum_{t_{s,h} \le t} \frac{d_{s,h}}{\tilde{Y}_{s,h}}$$

모형 구축에 사용된 모든 개체는 반드시 하나의 끝마디에 속하고, 개체 i는 P차원의 공변량 벡터 X_i로 설명된다. 따라서 벡터 $\boldsymbol{X}_i$는 반드시 1개의 끝마디에 속할 수 있으므로, 개체 i의 누적 위험함수는 아래와 같이 정의할 수 있다.

$$H(t|\boldsymbol{X}_i) = \hat{H}_h(t), \qquad \boldsymbol{X}_i \in h$$

In-bag 붓스트랩 표본을 이용한 앙상블 누적위험함수도 마찬가지로 Nelson-Aalen 추정량에 근거한다. b번째 in-bag 붓스트랩 표본에서 생성한 tree의 누적위험함수를 $\mathrm{H}_b^*(t|X)$라 하면, i의 앙상블 누적위험함수는 B개의 tree를 결합한 평균이다.

$$\hat{H}_e^*(t|\boldsymbol{X}_i) = \frac{1}{B}\sum_{b=1}^{B} \hat{H}_b^*(t|\boldsymbol{X}_i)$$

그리고 앙상블 생존함수는 위험함수와 생존함수의 관계에 따라 다음과 같이 표현한다.

$$\hat{S}^*(t|\boldsymbol{X}_i) = \exp(-\frac{1}{B}\sum_{b=1}^{B} \hat{H}_b^*(t|\boldsymbol{X}_i))$$

개체 i가 OOB에 속하면 $I_{i,b} = 1$, 그렇지 않은 경우를 $I_{i,b} = 0$라 정의할 때, OOB 자료를 이용한 i의 앙상블 누적위험함수는 아래의 식과 같다. 형성한 tree에 OOB 자료를

대입하여 i가 도달하는 끝마디의 누적위험함수를 구하는 것이다.

$$\hat{H}_e^{**}(t|\boldsymbol{X}_i) = \frac{\sum_{b=1}^{B} I_{i,b}\hat{H}_b^*(t|\boldsymbol{X}_i)}{\sum_{b=1}^{B} I_{i,b}}$$

Random survival forest의 알고리즘은 크게 in-bag 붓스트랩 표본으로부터 survival tree를 생성하고, 이를 합하여 앙상블 누적위험함수를 구축하는 두 단계로 구성된다. 마지막으로 모형을 평가하는 과정을 거치는데, OOB 자료를 적용하여 모형의 성능을 살펴본다.

2.3.2 변수 선택

Random survival forest는 VIMP(variable importance)와 minimal depth에 기초하여 생존 시간에 유의한 영향을 주는 독립변수를 평가한다. VIMP는 중요성을 확인하고자 하는 변수를 무작위로 "noise" 처리한 경우의 오차가 처리하지 않았을 때보다 얼만큼 증가 또는 감소하였는지를 의미한다. 다시 말하면, 오차 변화량을 뜻한다(Breiman 2001). Survival tree에서 임의의 변수 X를 noise 처리한다는 것은 분리변수로 선택되면, 생존 차이를 최대로 하는 지점 c를 찾는 것이 아니라 임의대로 두 개의 마디로 분할하는 것을 말한다. 이러한 방법으로 형성된 모형의 예측오차에서 noise 처리를 하지 않은 모형의 예측오차를 뺀 것이 변수 X의 VIMP가 된다(Ishwaran 2008). 다르게 정의하면 VIMP는 변수의 예측 능력이다. X를 noise 처리함으로써 0이나 음의 값이 얻어졌다면, 차이가 없거나 예측력이 향상된 것이라 볼 수 있다. 따라서 양의 VIMP를 갖는 변수에 대해서만 예측력에 영향을 주는 것으로 설명할 수 있다.

Minimal depth도 생존 시간에 미치는 영향력을 평가하는 지표이다(Ishwaran 외 2010). 뿌리 마디(root node)에서 변수 X가 처음으로 분리기준으로 사용된 마디까지의 거리를 변수 X의 minimal depth라 한다. 값이 작을수록 뿌리 마디와 가까운 곳에

서 분리변수로 선택되었음을 의미하며, 이는 곧 생존 시간의 차이를 극대화할 가능성이 높은 변수라는 것과 연결된다. 따라서 minimal depth의 크기가 작을수록 예측력이 높다고 할 수 있다. Minimal depth의 평균은 random survival forest에서 변수 선택 기준으로 사용되며, 평균보다 minimal depth가 작은 경우에 생존 시간에 영향을 주는 변수로 선정된다.

제 3장 모형 평가 지표

3.1 Harrell의 concordance index

모형을 평가하는 지표로 Harrell의 concordance index(c-index, Harrell 외 1982)를 사용한다. C-index는 ROC(receiver operating curve) 곡선 아래의 면적으로 두 개체를 임의로 선택하여 생존 시간이 짧은 개체의 위험도가 더 높게 예측된 경우의 확률을 말한다. 즉 모형의 예측력을 평가하는 지표이다. 본 논문에서는 OOB 자료로 모형의 예측력을 평가한다. $\sum_{l=1}^{m} \hat{H}_e^{**}(t_l \,|\boldsymbol{X}_i) > \sum_{l=1}^{m} \hat{H}_e^{**}(t_l \,|\boldsymbol{X}_j)$는 OOB 자료에서 사전에 지정한 m개의 사건 발생지점($t_1 < t_2 < \cdots < t_m$)에 대해 개체 i의 사망 위험이 j보다 높다는 것을 의미하며, 이를 바탕으로 c-index를 계산한다. C가 0.5이면 무작위로 예측한 것과 차이가 없고, 1에 가까울수록 모형의 분류 능력이 높다.

3.2 IBS(integrated brier score)

BS(brier score)는 모형의 예측오차를 평가하는 지표이다. Harrell의 c-index와 마찬가지로 OOB 자료를 이용하여 BS를 구한다. $\hat{S}^*$가 in-bag 붓스트랩 표본으로부터 생성된 모형이고 $\tilde{Y}_i(t) = I(\tilde{T}_i > t)$는 OOB에 속한 개체 i의 상태를 의미할 때, 개체 i의 BS는

$$\mathrm{BS}(t, \hat{S}) = \mathrm{E}(\tilde{Y}_i(t) - \hat{S}^*(t|\boldsymbol{X}_i))^2$$

이다. 일반적인 BS는 실제와 예측확률의 차이를 제곱하지만 중도절단 자료에서는 IPCW(inverse probability of censoring weights, Gerds and Schumacner 2006)를 부여한다.

$$\widehat{W}_i(t) = \frac{(1-\tilde{Y}_i(t))\Delta_i}{\hat{G}(\tilde{T}_i - |\boldsymbol{X}_i)} + \frac{\tilde{Y}_i(t)}{\hat{G}(t|\boldsymbol{X}_i)}$$

t시점 이전에 사망한 경우($\Delta_i = 1$)와 중도절단된 경우($\Delta_i = 0$) 각각 다른 가중치가 적용된다. 여기서 $\hat{G}(t\,|\boldsymbol{X}_i) \approx \mathrm{P}(C_i > t\,|\boldsymbol{X}_i = \mathrm{x})$는 중도절단 시간에 대한 조건부 생존함수의 추정량이다. IPCW 가중치를 부여한 BS를 재정의하면 아래와 같다.

$$\widehat{\mathrm{BS}}(t, \hat{S}) = \frac{1}{M}\sum_{i\in \mathrm{OOB}} \widehat{W}_i(t)\{\tilde{Y}_i(t) - \hat{S}^*(t\,|\boldsymbol{X}_i)\}^2$$

M은 OOB 자료의 크기이다. B개의 OOB 자료마다 1개의 BS가 도출되고, 앙상블 모형의 최종 BS는 아래와 같이 B개의 평균이다.

$$\mathrm{BS}_{bootcv}(t\,, \hat{S}) = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{M_b}\sum_{i\in \mathrm{OOB}} \widehat{W}_i(t)\,\{\tilde{Y}_i(t) - \hat{S}_b^*(t\,|\boldsymbol{X}_i)\}^2$$

앙상블 모형의 BS를 시간에 따라 나타내면 예측 오차 곡선이 얻어진다. 이 때, IBS는 최대 생존 시간까지의 누적 예측 오차 곡선(cumulative prediction error curve)이다.

$$\mathrm{IBS} = \frac{1}{\max(t_i)}\int_0^{\max(t_i)} \mathrm{BS}_{bootcv}(t)dt$$

제 4장 모의실험

제 4장에서는 제 2장에서 소개된 random survival forest와 Cox의 비례위험모형 및 Lasso-Cox 모형을 비교하고자 한다. 생존 시간에 대한 참모형을 다르게 고려한 4가지 경우에서 이차항과 교호작용, noise 변수 유무를 고려한 4가지 분석모형을 적합하여 Harrell의 c-index와 IBS로 세 방법의 예측력을 평가한다. 또한 독립변수 선정을 위해 random survival forest는 VIMP와 minimal depth를 이용하고, Cox의 비례위험모형과 Lasso-Cox 모형은 회귀계수로 확인한다.

4.1 생존 시간의 생성

본 논문에서는 지수분포를 따르는 Cox의 비례위험모형으로부터 생존 시간을 생성하였다. $S_0(t) = e^{-\lambda t}$라 가정하였을 때, 생존 시간에 대한 참모형은 다음과 같이 표현된다.

$$T = -\frac{\log(U)}{\lambda \exp(X^T \beta)}$$

독립변수 X_1, X_2, X_3, X_4, X_5는 주효과 변수이며 5개의 변수 간의 상관계수의 크기를 0.5로 가정하여 평균이 0인 다변량 정규분포로부터 생성하였다. U는 균등분포에서 생성된 0과 1사이의 실수이고, 기저위험도 λ는 1로 가정하였다. 추가적으로 10개의 변수 $N_1, \cdots, N_{10}$를 생성하여 자료의 noise를 부여하였다. noise는 생존 시간에 영향을 주진 않지만 분석모형에는 포함되는 변수를 말하며(Ishwaran 외 2010), 평균이 0인 다변량 정규분포로부터 생성되었다. 모의실험은 자료의 복잡성을 고려하여 독립변수의 주효과와 noise 변수 내 상관성 유무에 따라 수행된다. X와 N 내 상관성이 있다면 상관계수 크기를 0.4, noise 내의 상관성은 0.3으로 가정하였고, X와 N이 서로 상관성이 없을 경우 10개의 noise는 서로 독립으로 가정하였다. 생성한 표본 크기는

300이고, 생존 시간의 상위 20%를 중도절단하여 type 2 censoring을 고려하였다 (Klein and Moeschberger 2003).

참모형은 4가지 상황을 가정한다. 첫 번째는 참모형에 주효과만 있는 경우, 두 번째는 참모형에 주효과와 이차항만 있는 경우, 세 번째는 참모형에 주효과와 교호작용이 있는 경우, 마지막은 주효과, 이차항, 교호작용 모두 참모형에 포함되는 경우이다. 주효과 X_1, X_2, X_3, X_4, X_5의 참 회귀계수는 $\boldsymbol{\beta}_X = (1.2, 1.2, 0.4, 0.4, 0)^{\mathrm{T}}$로 설정하였다. 이차항은 X_1, X_2, X_3, X_4, X_5를 제곱한 Z_1, Z_2, Z_3, Z_4, Z_5이며 이차항의 참 회귀계수는 $\boldsymbol{\beta}_Z = (1, 1, 1, 1, 0)^{\mathrm{T}}$으로 가정하였다. 교호작용은 서로 다른 2개의 독립변수의 곱이다($W_{pq} = X_p \times X_q,\ p \neq q$). 10개의 교호작용항이 있고, W_{12}, W_{13}, W_{14}의 참 회귀계수는 1, W_{23}, W_{24}, W_{34} 는 0.4로 가정하였다. 주효과 변수 X_5는 생존 시간에 영향을 주지 않기 때문에 X_5와 관련된 이차항 Z_5과 교호작용항 W_{15}, W_{25}, W_{35}, W_{45}의 참 회귀계수는 0으로 설정하였다.

4.2 모의실험 설계

본 연구의 모의실험에서는 하나의 자료에 4가지 분석모형을 적용한다. 분석모형1은 참모형과 일치하는 모형, 분석모형2는 주효과모형, 분석모형3은 주효과모형에 noise가 추가된 모형, 분석모형4는 참모형에 noise가 추가된 모형이다. 참모형에 주효과만 있는 경우, 분석모형1과 2가 그리고 분석모형 3과 4가 동일하다. 참모형에 따라 4가지 경우의 생존 자료를 생성하고, 각각을 4개의 분석모형으로 적합하여 random survival forest와 Cox의 비례위험모형, Lasso-Cox 모형을 비교한다. 모의실험은 총 1000번의 반복 수행을 한다. Random survival forest는 100개의 나무가 결합되어 구축된다. R 패키지 randomForestSRC에서 제공하는 기본 옵션에 따라 분리기준은 로그 순위(log-rank)방법, 후보 분리변수의 수(mtry)는 제곱근 p라 지정하였다. Tree는 끝마디에 최소 3번의 사망 사건이 속할 때까지 성장한다. Lasso-Cox 모형의 조절 모수 λ는 R 패키지 glmnet을 이용하여 최솟값으로 채택한다. Random survival forest와 Cox의 비례위험모형, Lasso-Cox 모형을 평가하기 위해 100번의

붓스트랩을 수행하였으며, 붓스트랩 표본 크기는 전체 표본의 63%(Efron 1977)인 189명으로 설정하였다.

4.3 모의실험 결과

표 1, 표 2, 표 3, 표 4는 random survival forest, Cox의 비례위험모형, Lasso-Cox 모형의 c-index와 IBS이다. 분석모형이 참모형과 동일(모형1)하거나 참모형에 noise가 추가된 모형(모형4)일 때, Cox 또는 Lasso-Cox 모형은 random survival forest보다 c-index가 크고 IBS는 작았다. 반면, 분석모형이 주효과모형(모형2)이거나 주효과모형에 noise를 추가한 모형(모형3)으로 분석한 경우에는 random survival forest의 c-index가 가장 높았고, IBS가 가장 낮았다. 이차항의 효과나 독립변수 간의 교호작용이 존재함에도 불구하고 이들을 포함하지 않고 독립변수의 주효과모형으로 적합하게 되면 세 방법의 IBS가 모두 증가하였다. 특히, Cox와 Lasso-Cox 모형은 IBS가 최대 2배 이상 변화하였으며, 주효과모형에 noise가 추가된 분석모형을 적합 하에서 오차는 더욱 커졌다. C-index 측면에서 random survival forest는 어떠한 분석모형이 적용되어도 변동성이 크지 않고 일정한 크기로 유지되었다. 독립변수의 주효과와 noise 내 상관성을 가정한 경우에도 동일한 양상을 보였다.

모의실험 결과 random survival forest로부터 독립변수의 VIMP와 minimal depth는 참 회귀계수의 크기와 비례하였다. 주효과 변수는 (X_1, X_2), (X_3, X_4), X_5 순으로 VIMP가 작아지고, minimal depth는 커졌다. 참모형에서 생존 시간에 유의한 영향을 주지 않는 변수 X_5, Z_5, W_{15}, W_{25}, W_{35}, W_{45}는 나머지 유의한 변수들보다 상대적으로 VIMP가 작고 minimal depth가 컸다. Noise는 대부분 VIMP가 0에 가깝고 minimal depth가 독립변수들보다 매우 컸다. Tree의 mean depth보다 작은 minimal depth를 갖는 변수를 자료에서 유의한 영향이 있는 것으로 보았을 때, 주효과, 이차항, 교호작용 대부분이 1000번의 모의실험에서 100%에 가깝게 선택되었으며 noise는 상대적으로 적었다. 주효과와 noise 간의 상관성이 없는 경우와 있는 경우 결과의 방향성은 일치하였다.

표 5-1과 표 5-2는 참모형이 주효과로만 이루어진 자료에서 생존 시간에 미치는 독립변수의 영향력을 평가한 결과로 각각 Cox 비례위험모형과 Lasso-Cox 모형의 회귀계수 추정치이다. Cox와 Lasso-Cox 모형의 회귀계수 추정치는 앞서 설정했던 참 회귀계수 $\boldsymbol{\beta}_X = (1.2, 1.2, 0.4, 0.4, 0)^{\mathrm{T}}$와 거의 일치하였다. 또한 독립변수와 noise의 상관성을 가정하면 가정하지 않았을 때보다 회귀계수 추정치의 표준오차는 증가하였다.

표 6-1과 표 7-1, 표 8-1은 Cox의 비례위험모형으로부터 추정된 회귀계수 및 표준오차 결과이고, 표 6-2, 표 7-2, 표 8-2는 Lasso-Cox 모형의 추정치이다. 참모형이 주효과 외에 이차항 또는 교호작용이 있는 경우에서 Cox와 Lasso-Cox의 추정 결과는 거의 다르지 않았다. 참모형과 동일하게 분석모형과 noise가 추가된 참모형으로 적합한 결과 두 경우 모두 참 회귀계수와 근사하게 추정되었고, 참모형에 포함되지 않았던 noise의 회귀계수의 추청치는 0에 가까웠다. 그리고 분석모형이 주효과모형이거나 noise가 추가된 주효과모형인 경우, 주효과의 회귀계수는 실제보다 작게 추정되었다. Cox 모형에서 독립변수와 noise 내 상관성을 고려하면 회귀계수 추정치의 표준오차가 증가하였다. 반면, Lasso-Cox 모형은 참모형이 주효과만 있는 경우와 주효과와 교호작용이 함께 있는 경우에는 증가하지만, 나머지 경우에는 noise와의 상관관계를 가정하더라도 표준오차가 크게 달라지지 않았다.

표 1. 참모형이 주효과만 있는 경우에서 Harrell의 c-index와 IBS

상관계수[1]	모형[2]	C-index[†]			IBS[‡]		
		RSF[§]	CRA[‖]	Lasso-Cox[¶]	RSF	CRA	Lasso-Cox
0	모형 1	0.845	0.866	0.869	0.040	0.035	0.034
	모형 2						
	모형 3	0.840	0.869	0.868	0.054	0.036	0.033
	모형 4						
0.4	모형 1	0.843	0.864	0.864	0.041	0.035	0.034
	모형 2						
	모형 3	0.838	0.865	0.865	0.050	0.037	0.035
	모형 4						

[1] 주효과와 noise 의 상관계수, [2] 분석모형(모형 1: 참모형, 모형 2: 주효과모형, 모형 3: 주효과모형에 noise 가 추가된 모형, 모형 4: 참모형에 noise 가 추가된 모형), [†] Harrell 의 concordance index, [‡] Integrated brier score, [§] Random survival forest, [‖] Cox regression analysis, [¶] Lasso-Cox regression

표 2. 참모형이 주효과와 이차항만 있는 경우에서 Harrell의 c-index와 IBS

상관계수[1]	모형[2]	C-index[†]			IBS[‡]		
		RSF[§]	CRA[‖]	Lasso-Cox[¶]	RSF	CRA	Lasso-Cox
0	모형 1	0.846	0.883	0.881	0.084	0.069	0.069
	모형 2	0.840	0.705	0.703	0.083	0.187	0.182
	모형 3	0.827	0.708	0.641	0.108	0.191	0.180
	모형 4	0.843	0.885	0.882	0.097	0.072	0.063
0.4	모형 1	0.846	0.881	0.881	0.087	0.071	0.069
	모형 2	0.840	0.705	0.705	0.086	0.187	0.182
	모형 3	0.825	0.704	0.704	0.108	0.192	0.180
	모형 4	0.841	0.882	0.882	0.099	0.074	0.070

[1] 주효과와 noise 의 상관계수, [2] 분석모형(모형 1: 참모형, 모형 2: 주효과모형, 모형 3: 주효과모형에 noise 가 추가된 모형, 모형 4: 참모형에 noise 가 추가된 모형), [†] Harrell 의 concordance index, [‡] Integrated brier score, [§] Random survival forest, [‖] Cox regression analysis, [¶] Lasso-Cox regression

표 3. 참모형이 주효과와 교호작용만 있는 경우에서 Harrell의 c-index와 IBS

상관계수[1]	모형[2]	C-index[†]			IBS[‡]		
		RSF[§]	CRA[‖]	Lasso-Cox[¶]	RSF	CRA	Lasso-Cox
0	모형 1	0.805	0.868	0.866	0.133	0.074	0.072
	모형 2	0.790	0.709	0.708	0.139	0.186	0.180
	모형 3	0.756	0.711	0.706	0.154	0.189	0.177
	모형 4	0.795	0.871	0.867	0.140	0.078	0.073
0.4	모형 1	0.804	0.860	0.860	0.134	0.077	0.080
	모형 2	0.790	0.706	0.705	0.139	0.186	0.182
	모형 3	0.757	0.695	0.696	0.151	0.190	0.183
	모형 4	0.794	0.856	0.857	0.140	0.081	0.089

[1] 주효과와 noise 의 상관계수, [2] 분석모형(모형 1: 참모형, 모형 2: 주효과모형, 모형 3: 주효과모형에 noise 가 추가된 모형, 모형 4: 참모형에 noise 가 추가된 모형), [†] Harrell 의 concordance index, [‡] Integrated brier score, [§] Random survival forest, [‖] Cox regression analysis, [¶] Lasso-Cox regression

표 4. 참모형이 주효과, 이차항, 교호작용만 있는 경우에서 Harrell의 c-index와 IBS

상관계수[1]	모형[2]	C-index[†]			IBS[‡]		
		RSF[§]	CRA[‖]	Lasso-Cox[¶]	RSF	CRA	Lasso-Cox
0	모형 1	0.860	0.911	0.885	0.090	0.071	0.096
	모형 2	0.841	0.628	0.620	0.091	0.192	0.188
	모형 3	0.819	0.632	0.617	0.116	0.196	0.187
	모형 4	0.856	0.913	0.891	0.096	0.078	0.097
0.4	모형 1	0.844	0.893	0.862	0.097	0.075	0.089
	모형 2	0.827	0.636	0.636	0.097	0.185	0.179
	모형 3	0.813	0.648	0.630	0.113	0.180	0.188
	모형 4	0.847	0.908	0.909	0.100	0.073	0.064

[1] 주효과와 noise 의 상관계수, [2] 분석모형(모형 1: 참모형, 모형 2: 주효과모형, 모형 3: 주효과모형에 noise 가 추가된 모형, 모형 4: 참모형에 noise 가 추가된 모형), [†] Harrell 의 concordance index, [‡] Integrated brier score, [§] Random survival forest, [‖] Cox regression analysis, [¶] Lasso-Cox regression

표 5-1. 참모형이 주효과만 있는 경우에서 Cox 의 비례위험모형의 회귀계수 추정치

상관계수[1]	독립 변수	참 회귀계수	모형[2]			
			모형 1	모형 2	모형 3	모형 4
			β	SD	β	SD
0	X_1	1.2	1.223	0.105	1.251	0.109
	X_2	1.2	1.216	0.105	1.250	0.109
	X_3	0.4	0.409	0.090	0.423	0.092
	X_4	0.4	0.409	0.088	0.406	0.092
	X_5	0	−0.002	0.086	−0.004	0.089
0.4	X_1	1.2	1.215	0.134	1.265	0.147
	X_2	1.2	1.224	0.135	1.273	0.147
	X_3	0.4	0.413	0.114	0.425	0.126
	X_4	0.4	0.407	0.114	0.425	0.126
	X_5	0	0.006	0.110	0.010	0.122

[1] 주효과와 noise 의 상관계수, [2] 분석모형(모형 1: 참모형, 모형 2: 주효과모형, 모형 3: 주효과모형에 noise 가 추가된 모형, 모형 4: 참모형에 noise 가 추가된 모형)

표 5-2. 참모형이 주효과만 있는 경우에서 Lasso-Cox 모형의 회귀계수 추정치

상관계수[1]	독립 변수	참 회귀계수	모형[2]			
			모형 1	모형 2	모형 3	모형 4
			β	SD	β	SD
0	X_1	1.2	1.179	0.104	1.218	0.108
	X_2	1.2	1.178	0.104	1.216	0.108
	X_3	0.4	0.384	0.088	0.399	0.091
	X_4	0.4	0.382	0.088	0.396	0.091
	X_5	0	0.010	0.086	0.007	0.089
0.4	X_1	1.2	1.174	0.133	1.228	0.146
	X_2	1.2	1.183	0.134	1.237	0.146
	X_3	0.4	0.393	0.114	0.406	0.125
	X_4	0.4	0.388	0.114	0.406	0.126
	X_5	0	0.019	0.110	0.014	0.122

[1] 주효과와 noise 의 상관계수, [2] 분석모형(모형 1: 참모형, 모형 2: 주효과모형, 모형 3: 주효과모형에 noise 가 추가된 모형, 모형 4: 참모형에 noise 가 추가된 모형)

표 6-1. 참모형이 주효과와 이차항만 있는 경우에서 Cox의 비례위험모형의 회귀계수 추정치

상관계수[1]	독립변수	참 회귀계수	모형[2]							
			모형 1		모형 2		모형 3		모형 4	
			β	SD	β	SD	β	SD	β	SD
0	X_1	1.2	1.227	0.105	0.547	0.103	0.553	0.105	1.258	0.109
	X_2	1.2	1.228	0.105	0.545	0.103	0.552	0.104	1.257	0.109
	X_3	0.4	0.416	0.089	0.165	0.100	0.166	0.101	0.423	0.092
	X_4	0.4	0.413	0.088	0.162	0.100	0.162	0.102	0.421	0.092
	X_5	0	−0.003	0.090	−0.003	0.085	−0.004	0.087	−0.004	0.093
	Z_1	1	1.031	0.076	−	−	−	−	1.056	0.078
	Z_2	1	1.027	0.075	−	−	−	−	1.053	0.078
	Z_3	1	1.029	0.074	−	−	−	−	1.053	0.077
	Z_4	1	1.023	0.074	−	−	−	−	1.054	0.077
	Z_5	0	0.006	0.054	−	−	−	−	0.006	0.056
0.4	X_1	1.2	1.252	0.138	0.562	0.130	0.574	0.139	1.301	0.150
	X_2	1.2	1.249	0.138	0.557	0.130	0.570	0.139	1.296	0.151
	X_3	0.4	0.416	0.116	0.172	0.127	0.174	0.136	0.430	0.129
	X_4	0.4	0.413	0.116	0.163	0.127	0.164	0.136	0.425	0.129
	X_5	0	0.011	0.118	0.002	0.108	0.006	0.119	0.010	0.130
	Z_1	1	1.052	0.101	−	−	−	−	1.093	0.107
	Z_2	1	1.048	0.100	−	−	−	−	1.088	0.106
	Z_3	1	1.050	0.098	−	−	−	−	1.091	0.104
	Z_4	1	1.041	0.099	−	−	−	−	1.081	0.104
	Z_5	0	0.002	0.073	−	−	−	−	0.003	0.077

[1] 주효과와 noise 의 상관계수, [2] 분석모형(모형 1: 참모형, 모형 2: 주효과모형, 모형 3: 주효과모형에 noise 가 추가된 모형, 모형 4: 참모형에 noise 가 추가된 모형)

표 6-2. 참모형이 주효과와 이차항만 있는 경우에서 Lasso-Cox 모형의 회귀계수 추정치

상관계수[1]	독립변수	참 회귀계수	모형[2]							
			모형 1		모형 2		모형 3		모형 4	
			β	SD	β	SD	β	SD	β	SD
0	X_1	1.2	1.194	0.136	0.533	0.130	0.553	0.134	1.245	0.144
	X_2	1.2	1.200	0.135	0.536	0.130	0.547	0.134	1.249	0.143
	X_3	0.4	0.394	0.115	0.455	0.126	0.160	0.130	0.409	0.122
	X_4	0.4	0.387	0.115	0.146	0.126	0.151	0.130	0.406	0.121
	X_5	0	0.013	0.117	0.009	0.108	0.007	0.113	0.013	0.123
	Z_1	1	0.999	0.098	–	–	–	–	1.042	0.104
	Z_2	1	1.001	0.098	–	–	–	–	1.043	0.104
	Z_3	1	1.000	0.097	–	–	–	–	1.041	0.102
	Z_4	1	1.002	0.096	–	–	–	–	1.043	0.102
	Z_5	0	0.004	0.072	–	–	–	–	0.004	0.076
0.4	X_1	1.2	1.202	0.136	0.548	0.130	0.563	0.139	1.248	0.148
	X_2	1.2	1.199	0.136	0.543	0.130	0.558	0.139	1.244	0.149
	X_3	0.4	0.394	0.116	0.164	0.127	0.167	0.136	0.406	0.129
	X_4	0.4	0.392	0.115	0.155	0.127	0.157	0.136	0.401	0.128
	X_5	0	0.018	0.117	0.008	0.108	0.008	0.119	0.013	0.129
	Z_1	1	1.008	0.099	–	–	–	–	1.050	0.105
	Z_2	1	1.004	0.099	–	–	–	–	1.045	0.105
	Z_3	1	1.007	0.097	–	–	–	–	1.049	0.102
	Z_4	1	0.999	0.097	–	–	–	–	1.040	0.102
	Z_5	0	0.004	0.072	–	–	–	–	0.005	0.076

[1] 주효과와 noise 의 상관계수, [2] 분석모형(모형 1: 참모형, 모형 2: 주효과모형, 모형 3: 주효과모형에 noise 가 추가된 모형, 모형 4: 참모형에 noise 가 추가된 모형)

표 7-1. 참모형이 주효과와 교호작용만 있는 경우에서 Cox의 비례위험모형의 회귀계수 추정치

상관계수[1]	독립변수	참 회귀계수	모형[2]							
			모형 1		모형 2		모형 3		모형 4	
			β	SD	β	SD	β	SD	β	SD
0	X_1	1.2	1.244	0.118	0.570	0.092	0.578	0.094	1.279	0.121
	X_2	1.2	1.243	0.113	0.589	0.090	0.599	0.093	1.276	0.117
	X_3	0.4	0.415	0.098	0.185	0.086	0.185	0.089	0.427	0.101
	X_4	0.4	0.414	0.098	0.179	0.087	0.179	0.089	0.423	0.101
	X_5	0	0.005	0.092	0.002	0.085	−0.002	0.087	0.000*	0.095
	W_{12}	1	1.044	0.136	−	−	−	−	1.072	0.141
	W_{13}	1	1.041	0.132	−	−	−	−	1.066	0.136
	W_{14}	1	1.033	0.133	−	−	−	−	1.062	0.137
	W_{15}	0	0.010	0.122	−	−	−	−	0.002	0.125
	W_{23}	0.4	0.418	0.123	−	−	−	−	0.427	0.127
	W_{24}	0.4	0.413	0.123	−	−	−	−	0.420	0.121
	W_{25}	0	0.000*	0.119	−	−	−	−	0.004	0.123
	W_{34}	0.4	0.419	0.120	−	−	−	−	0.426	0.124
	W_{35}	0	−0.008	0.117	−	−	−	−	−0.005	0.120
	W_{45}	0	0.002	0.117	−	−	−	−	0.003	0.121

(계속)

표 7-1. 참모형이 주효과와 교호작용만 있는 경우에서 Cox의 비례위험모형의 회귀계수 추정치 (계속)

상관계수[1]	독립변수	참 회귀계수	모형[2]							
			모형 1		모형 2		모형 3		모형 4	
			β	SD	β	SD	β	SD	β	SD
0.4	X_1	1.2	1.325	0.223	0.569	0.156	0.587	0.177	1.430	0.252
	X_2	1.2	1.331	0.215	0.600	0.154	0.628	0.174	1.434	0.245
	X_3	0.4	0.437	0.186	0.179	0.148	0.175	0.168	0.470	0.213
	X_4	0.4	0.459	0.186	0.190	0.149	0.186	0.169	0.498	0.215
	X_5	0	0.000*	0.176	−0.001	0.144	−0.005	0.164	0.006	0.202
	W_{12}	1	1.118	0.268	−	−	−	−	1.208	0.294
	W_{13}	1	1.117	0.258	−	−	−	−	1.204	0.284
	W_{14}	1	1.114	0.260	−	−	−	−	1.206	0.287
	W_{15}	0	0.008	0.236	−	−	−	−	0.010	0.258
	W_{23}	0.4	0.445	0.243	−	−	−	−	0.482	0.267
	W_{24}	0.4	0.433	0.243	−	−	−	−	0.469	0.267
	W_{25}	0	0.001	0.235	−	−	−	−	−0.004	0.257
	W_{34}	0.4	0.467	0.240	−	−	−	−	0.507	0.265
	W_{35}	0	−0.006	0.229	−	−	−	−	−0.001	0.251
	W_{45}	0	−0.009	0.230	−	−	−	−	−0.001	0.253

[1] 주효과와 noise 의 상관계수, [2] 분석모형(모형 1: 참모형, 모형 2: 주효과모형, 모형 3: 주효과모형에 noise 가 추가된 모형, 모형 4: 참모형에 noise 가 추가된 모형), * 추정치가 0 에 근사

표 7-2. 참모형이 주효과와 교호작용만 있는 경우에서 Lasso-Cox 모형의 회귀계수 추정치

상관계수[1]	독립변수	참 회귀계수	모형[2]							
			모형 1		모형 2		모형 3		모형 4	
			β	SD	β	SD	β	SD	β	SD
0	X_1	1.2	1.224	0.154	0.554	0.117	0.565	0.123	1.275	0.162
	X_2	1.2	1.221	0.147	0.577	0.115	0.593	0.120	1.273	0.156
	X_3	0.4	0.402	0.128	0.172	0.110	0.174	0.115	0.418	0.135
	X_4	0.4	0.406	0.128	0.176	0.110	0.177	0.115	0.424	0.135
	X_5	0	0.009	0.122	0.008	0.108	0.008	0.113	0.009	0.128
	W_{12}	1	1.024	0.182	–	–	–	–	1.065	0.194
	W_{13}	1	1.015	0.176	–	–	–	–	1.058	0.185
	W_{14}	1	1.038	0.176	–	–	–	–	1.082	0.185
	W_{15}	0	0.006	0.164	–	–	–	–	0.007	0.172
	W_{23}	0.4	0.415	0.165	–	–	–	–	0.431	0.173
	W_{24}	0.4	0.388	0.164	–	–	–	–	0.405	0.172
	W_{25}	0	0.013	0.160	–	–	–	–	0.014	0.167
	W_{34}	0.4	0.400	0.161	–	–	–	–	0.415	0.168
	W_{35}	0	0.004	0.155	–	–	–	–	0.003	0.163
	W_{45}	0	0.008	0.156	–	–	–	–	0.008	0.164

(계속)

표 7-2. 참모형이 주효과와 교호작용만 있는 경우에서 Lasso-Cox 모형의 회귀계수 추정치 (계속)

상관계수[1]	독립변수	참 회귀계수	모형[2]							
			모형 1		모형 2		모형 3		모형 4	
			β	SD	β	SD	β	SD	β	SD
0.4	X_1	1.2	1.274	0.219	0.552	0.156	0.573	0.177	1.371	0.248
	X_2	1.2	1.282	0.211	0.586	0.153	0.614	0.174	1.381	0.240
	X_3	0.4	0.416	0.184	0.171	0.147	0.168	0.168	0.444	0.211
	X_4	0.4	0.467	0.185	0.182	0.148	0.179	0.169	0.471	0.212
	X_5	0	0.008	0.174	0.006	0.144	−0.002	0.163	0.008	0.200
	W_{12}	1	1.073	0.264	−	−	−	−	0.006	0.168
	W_{13}	1	1.072	0.255	−	−	−	−	0.003	0.166
	W_{14}	1	1.069	0.257	−	−	−	−	0.006	0.168
	W_{15}	0	0.013	0.233	−	−	−	−	−0.003	0.169
	W_{23}	0.4	0.423	0.240	−	−	−	−	0.005	0.167
	W_{24}	0.4	0.413	0.241	−	−	−	−	−0.001	0.167
	W_{25}	0	0.005	0.233	−	−	−	−	−0.001	0.168
	W_{34}	0.4	0.447	0.238	−	−	−	−	0.001	0.167
	W_{35}	0	−0.001	0.227	−	−	−	−	0.007	0.168
	W_{45}	0	−0.005	0.228	−	−	−	−	0.003	0.168

[1] 주효과와 noise의 상관계수, [2] 분석모형(모형1: 참모형, 모형2: 주효과모형, 모형3: 주효과모형에 noise가 추가된 모형, 모형4: 참모형에 noise가 추가된 모형)

표 8-1. 참모형이 주효과, 이차항, 교호작용만 있는 경우에서 Cox 의 비례위험모형의 회귀계수 추정치

상관계수[1]	독립변수	참 회귀계수	모형[2]							
			모형 1		모형 2		모형 3		모형 4	
			β	SD	β	SD	β	SD	β	SD
0	X_1	1.2	1.260	0.115	0.352	0.096	0.035	0.098	1.293	0.119
	X_2	1.2	1.264	0.113	0.372	0.096	0.374	0.098	1.296	0.117
	X_3	0.4	0.424	0.096	0.083	0.095	0.082	0.096	0.436	0.099
	X_4	0.4	0.415	0.095	0.078	0.095	0.077	0.097	0.415	0.099
	X_5	0	−0.004	0.096	−0.003	0.085	−0.003	0.087	−0.004	0.100
	Z_1	1	1.066	0.105	−	−	−	−	1.094	0.109
	Z_2	1	1.039	0.138	−	−	−	−	1.090	0.105
	Z_3	1	0.418	0.126	−	−	−	−	1.090	0.106
	Z_4	1	0.414	0.125	−	−	−	−	1.094	0.105
	Z_5	0	−0.005	0.126	−	−	−	−	0.017	0.094
	W_{12}	1	1.063	0.102	−	−	−	−	1.066	0.142
	W_{13}	1	1.063	0.102	−	−	−	−	1.070	0.143
	W_{14}	1	1.065	0.101	−	−	−	−	1.068	0.142
	W_{15}	0	0.016	0.091	−	−	−	−	−0.011	0.134
	W_{23}	0.4	1.042	0.138	−	−	−	−	0.429	0.131
	W_{24}	0.4	1.041	0.137	−	−	−	−	0.423	0.130
	W_{25}	0	−0.009	0.129	−	−	−	−	−0.003	0.131
	W_{34}	0.4	0.412	0.125	−	−	−	−	0.426	0.130
	W_{35}	0	−0.003	0.126	−	−	−	−	−0.006	0.131
	W_{45}	0	−0.008	0.126	−	−	−	−	−0.007	0.131

(계속)

표 8-1. 참모형이 주효과, 이차항, 교호작용만 있는 경우에서 Cox 의 비례위험모형의 회귀계수 추정치 (계속)

상관계수[1]	독립변수	참 회귀계수	모형[2]							
			모형 1		모형 2		모형 3		모형 4	
			β	SD	β	SD	β	SD	β	SD
0.4	X_1	1.2	1.191	0.151	0.347	0.121	0.384	0.129	1.348	0.169
	X_2	1.2	1.194	0.149	0.368	0.122	0.414	0.129	1.362	0.168
	X_3	0.4	0.409	0.130	0.087	0.120	0.100	0.127	0.459	0.144
	X_4	0.4	0.400	0.129	0.084	0.121	0.102	0.130	0.453	0.144
	X_5	0	−0.003	0.130	0.008	0.108	0.004	0.118	−0.003	0.144
	Z_1	1	0.931	0.141	−	−	−	−	1.150	0.162
	Z_2	1	0.891	0.187	−	−	−	−	1.149	0.145
	Z_3	1	0.353	0.173	−	−	−	−	1.160	0.144
	Z_4	1	0.445	0.173	−	−	−	−	1.158	0.146
	Z_5	0	0.067	0.175	−	−	−	−	0.014	0.126
	W_{12}	1	0.925	0.137	−	−	−	−	1.122	0.202
	W_{13}	1	0.938	0.136	−	−	−	−	1.131	0.191
	W_{14}	1	1.026	0.138	−	−	−	−	1.130	0.201
	W_{15}	0	0.015	0.124	−	−	−	−	−0.005	0.179
	W_{23}	0.4	0.907	0.185	−	−	−	−	0.437	0.171
	W_{24}	0.4	0.903	0.186	−	−	−	−	0.441	0.171
	W_{25}	0	−0.003	0.176	−	−	−	−	−0.001	0.171
	W_{34}	0.4	0.357	0.173	−	−	−	−	0.441	0.171
	W_{35}	0	0.003	0.175	−	−	−	−	−0.007	0.170
	W_{45}	0	0.081	0.174	−	−	−	−	−0.016	0.172

[1] 주효과와 noise의 상관계수, [2] 분석모형(모형1: 참모형, 모형2: 주효과모형, 모형3: 주효과모형에 noise가 추가된 모형, 모형4: 참모형에 noise가 추가된 모형)

표 8-2. 참모형이 주효과, 이차항, 교호작용만 있는 경우에서 Lasso-Cox 모형의 회귀계수 추정치

상관계수[1]	독립변수	참 회귀계수	모형[2]							
			모형 1		모형 2		모형 3		모형 4	
			β	SD	β	SD	β	SD	β	SD
0	X_1	1.2	1.225	0.176	0.336	0.121	0.341	0.125	1.279	0.160
	X_2	1.2	1.228	0.178	0.359	0.121	0.362	0.125	1.282	0.158
	X_3	0.4	0.407	0.161	0.072	0.120	0.073	0.124	0.423	0.136
	X_4	0.4	0.400	0.157	0.073	0.120	0.071	0.124	0.417	0.136
	X_5	0	0.014	0.155	0.011	0.107	0.010	0.112	0.015	0.136
	Z_1	1	1.044	0.168	–	–	–	–	1.089	0.148
	Z_2	1	1.047	0.231	–	–	–	–	1.089	0.145
	Z_3	1	1.052	0.237	–	–	–	–	1.097	0.144
	Z_4	1	1.042	0.234	–	–	–	–	1.085	0.145
	Z_5	0	0.018	0.217	–	–	–	–	0.018	0.131
	W_{12}	1	1.009	0.172	–	–	–	–	1.053	0.195
	W_{13}	1	1.013	0.175	–	–	–	–	1.058	0.194
	W_{14}	1	1.011	0.170	–	–	–	–	1.056	0.197
	W_{15}	0	−0.006	0.153	–	–	–	–	−0.006	0.184
	W_{23}	0.4	0.392	0.226	–	–	–	–	0.411	0.181
	W_{24}	0.4	0.395	0.223	–	–	–	–	0.414	0.182
	W_{25}	0	−0.011	0.216	–	–	–	–	−0.010	0.183
	W_{34}	0.4	0.401	0.219	–	–	–	–	0.417	0.181
	W_{35}	0	−0.008	0.219	–	–	–	–	−0.006	0.182
	W_{45}	0	0.008	0.221	–	–	–	–	0.006	0.184

(계속)

표 8-2. 참모형이 주효과, 이차항, 교호작용만 있는 경우에서 Lasso-Cox 모형의 회귀계수 추정치 (계속)

상관계수[1]	독립변수	참 회귀계수	모형[2]							
			모형 1		모형 2		모형 3		모형 4	
			β	SD	β	SD	β	SD	β	SD
0.4	X_1	1.2	1.226	0.176	0.351	0.120	0.378	0.131	1.275	0.165
	X_2	1.2	1.235	0.178	0.375	0.120	0.403	0.131	1.289	0.164
	X_3	0.4	0.413	0.159	0.094	0.118	0.091	0.129	0.426	0.143
	X_4	0.4	0.405	0.156	0.095	0.121	0.091	0.132	0.420	0.142
	X_5	0	0.005	0.155	0.007	0.108	0.001	0.120	0.002	0.142
	Z_1	1	1.042	0.170	-	-	-	-	1.088	0.159
	Z_2	1	1.041	0.230	-	-	-	-	1.089	0.142
	Z_3	1	1.049	0.213	-	-	-	-	1.099	0.141
	Z_4	1	1.051	0.233	-	-	-	-	1.098	0.143
	Z_5	0	0.008	0.215	-	-	-	-	0.009	0.125
	W_{12}	1	1.018	0.172	-	-	-	-	1.061	0.199
	W_{13}	1	1.024	0.172	-	-	-	-	1.071	0.188
	W_{14}	1	1.018	0.169	-	-	-	-	1.068	0.198
	W_{15}	0	0.002	0.151	-	-	-	-	0.001	0.177
	W_{23}	0.4	0.393	0.224	-	-	-	-	0.411	0.169
	W_{24}	0.4	0.397	0.221	-	-	-	-	0.414	0.170
	W_{25}	0	0.002	0.213	-	-	-	-	0.002	0.169
	W_{34}	0.4	0.399	0.218	-	-	-	-	0.414	0.169
	W_{35}	0	-0.001	0.215	-	-	-	-	-0.002	0.169
	W_{45}	0	-0.008	0.215	-	-	-	-	-0.009	0.170

[1] 주효과와 noise의 상관계수, [2] 분석모형(모형1: 참모형, 모형2: 주효과모형, 모형3: 주효과모형에 noise가 추가된 모형, 모형4: 참모형에 noise가 추가된 모형)

제 5장 결론 및 고찰

Random survival forest는 이차항의 효과나 독립 변수간의 교호작용을 자동으로 처리할 수 있고, 고차원 자료에서 Cox 모형보다 우수하다고 알려져 있다. 하지만 실제 임상 자료를 분석한 결과 random survival forest와 Cox 비례위험모형의 예측오차가 비슷한 것을 여러 논문을 통해 확인하였다. 따라서 본 논문에서는 위와 같은 결과가 나타나는 원인을 자료 구조에서 찾아보고자 다양한 시뮬레이션 환경에서 두 방법과 Lasso-Cox 모형을 비교하였다. 독립변수 내 상관성을 가정하여 여러 가지 경우의 참모형을 통해 생존 자료를 생성하고, 이차항과 교호작용, noise 변수 유무, 독립변수와 noise의 상관성 유무를 고려하여 random survival forest, Cox의 비례위험모형, Lasso-Cox 모형을 평가하였다.

모의실험을 통해 분석모형이 참모형과 일치하는 경우, random survival forest는 Cox의 비례위험 모형과 Lasso-Cox 모형에 비해 c-index가 낮고 IBS는 높아 예측력이 다소 떨어졌다. 그리고 이차항의 효과나 독립변수 간의 교호작용이 존재함에도 불구하고 이들을 분석모형에 포함하지 않고 주효과만으로 적합하였을 경우, Cox와 Lasso-Cox 모형은 random survival forest보다 c-index가 낮았다. 특히 IBS는 2배 이상 높아져 예측오차가 크게 증가하였다. 반면, random survival forest는 c-index와 IBS가 크게 달라지지 않고 일정한 값으로 나타나 복잡한 참모형에서 분석모형이 부적합하더라도 예측력이 어느 정도 유지되었다. noise가 분석모형에 추가된 경우에도 동일한 양상을 보였다. Cox와 Lasso-Cox 모형에서 독립변수와 noise 내 상관성을 고려하면 회귀계수 추정치의 표준오차는 다소 증가하였다. 또한, 분석모형이 부적합하면 이차항이나 교호작용의 효과가 분석모형에 반영되지 않아 회귀계수가 과소 추정되었다. Random survival forest에서 변수의 minimal depth가 나무의 mean depth보다 작은 경우를 최종 변수로 선정할 시 noise를 제외한 모든 변수들이 100%에 가까운 빈도로 선택되었다.

위의 내용들을 종합하면, Cox의 비례위험모형이나 Lasso-Cox 모형은 분석모형

이참모형과 다른 모형으로 부적합되면 오차의 변화량이 크고 예측력이 다소 떨어진다. 하지만 random survival forest는 오차의 변동이 작아 예측력이 일정하게 유지된다. 이를 통해 실제 자료를 분석한 결과에서 random survival forest와 Cox 모형의 성능이 비슷한 원인을 복잡한 자료 구조, 또는 분석모형이 생존 시간의 참모형과 유사할 가능성이 높다는 것에서 찾을 수 있다. 빅데이터에서 자료 구조가 복잡할 경우, 자료의 참모형을 파악하기 어렵고, 주효과 외에 이차항 효과나 교호작용 또는 변수 간 상관성을 간과할 수 있다. 이 때 잘못된 모형으로 부적합하여 분석할 위험성은 높아질 수 밖에 없다. 따라서 random survival forest가 분석모형을 잘못 적합하여도 예측력을 일정하게 보존할 수 있다는 측면에서 안정적이고, 빅데이터 시대에서 유용하게 사용될 수 있는 방법이라 기대된다.

참고 문헌

Ishwaran, H., Kogalur, U. B., Blackstone, E.H., and Lauer, M. S. (2008). "Random Survival Forests". *The Annuals of Applied Statistics*, 2, 841-860

Breiman. (2001). Random forests. Machine Learning 45 5-32

Mogenses, U. B., Ishwaran, H., and Gerds, T. A. (2012). "Evaluating Random Forests for Survival Analysis Using Prediction Error Curves". *Journal of Statistical Software*, 50(11), 1-23

Ishwaran, H., Gerds, T. A., Kogalur, U. B., Moore, R. D., Gange, S. J., and Lau, B. M. (2014). "Random survival forests for competing risks". *Biostatistics*, 15(4), 757-773

Brieman et al. (1984). Classification and Regression Trees. Wadsworth, Belmont, California. MR0726392

Ishwaran, H., Kogalur, U. B., Chen, X. and Minn, A. J. (2011). "Random survival forests for high-dimensional data". *Statistical Analysis and Data Mining*, 4, 115-132

Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J. and Lauer, M. S. (2010). "High-dimensional variable selection for survival data". *Journal of the American Statistical Association*, 105(489), 205-217

Imran., K. O., Mevlut, T., Fusun, T. (2009). "The comparisons of random survival forests and Cox regression analysis with simulation and an application related to breast cancer". *Expert Systems with Applications*, 34(4), 8582-8588

Datema, F. R., Moya, A., Krause, P., et al. (2012). "Novel head and neck cancer survival analysis approach: random survival forests versus cox proportional hazards regression". *Head Neck*. 34(1), 50-58

Hsich, E., Gorodeski, E. Z., Blackstone, E. H., Ishwaran, H., Lauer, M. S. (2011). "Identifying important risk factors for survival in patient with systolic heart failure using random survival forests". *Cir Cardiovasc Qual Outcomes*, 4, 39-45

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso". *Journal of the Royal Statistical Society. Series B,* 58(1), 267-288

Tibshirani, R. (1997). "The Lasso Method for Variable Selection in the Cox Model". *Statistics in Medicine*, 16, 385-395

Harrell, F. E., Jr., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). "Evaluating the yield of medical tests". *JAMA*, 247(18), 2543-6

Gerds, T. A., and Schumacher, M. (2006). "Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times". *Biometrical Journal*, 48(6), 1029-1040

Efron, B., and Tibshirani, R. (1977). "Improvements on Cross-Validation: The .632+ Bootstrap Method". *Journal of the American Statistical Association*, pp. 548-560

김양진. 2013. 생존분석. 파주: 자유아카데미

Klein, J. P., Moeschberger, M. L. 2003. Survival Analysis: Techniques for Censored and Truncated Data Second edition. Springer

ABSTRACT

Prediction Error of Random Survival Forest and Cox Regression Analysis with model misspecification

Oum, Chi Yoon

Dept. of Biostatistics and Computing

The Graduate School

Yonsei University

As being the new era of big data, data mining methods are applied to clinical research. Random survival forest is ensemble learning for analysis of right-censored survival data. It draws bootstrap samples randomly and chooses candidate variables randomly for splitting at each node of the tree. By giving randomization to the base learning process, random survival forest improves the performance of predicting survival probability. However, applications of random survival forest in real data showed little difference of prediction error between random survival forest and Cox regression analysis. The purpose of this article is to find the reason why two methods have similar performance. Various survival data with the correlation of covariates are generated, and true models are specified according to data. Also, analysis model are presented considering quadratic effect, interaction effect, and noise variables. They are fitted in random survival forest, Cox regression model, and Lasso-Cox model. After that, we compare three methods. The performance measures are Harrell's c-index and IBS.

When the analysis model was the same with true model, random survival forest had the lowest c-index and the highest IBS. As quadratic terms or interaction terms had significant effects on true model, but not included in analysis model, c-index of random survival forest was the highest and IBS was the lowest among three methods in the misspecified analysis model. Especially, IBS increased by more than twice in the case of Cox analysis and Lasso-Cox model. The simulation results demonstrated that, for random survival forest, the prediction error of variation was quite smaller than the others and the prediction capability was maintained even if model was misspecified.

In the complex structure of big data, it is difficult to clarify true model. This gives rise to model misspecification risk. From this point of view, random survival forest can be useful methods in clinical research.

Key words : Random survival forest, Cox regression analysis, Lasso-Cox model, Noise