**MULTIVARIABLE PROGNOSTIC MODELS: ISSUES IN DEVELOPING MODELS, EVALUATING ASSUMPTIONS AND ADEQUACY, AND MEASURING AND REDUCING ERRORS**

**Summary**

(Multivariable regression models can use categorical and continuous variables)

예측된 차이의 인덱스(order) + Assessing calibration (구간 측정) 를 쉽게 해석하는 방법에 대해서 얘기해보고자 한다.

여러 모델에서 사용될 수 있다(Cox regression)

1. **Intro**
2. **Prelimnary steps**

Missing data가 자주 있는 문제 -> simple + Kuhufeld + Harrell

지킬 수 있는 데이터는 지키는게 좋다 하지만 필요없는 데이터 지우면 overfitting 방지하고 일반화 잘 시킬 수 있다.

Predicting clinical outcome시 중요한 가정

1. Interactions between treatment and the severity of disease being treated. Patients with little disease have little opportunity to receive benefit.

2. 나이 먹었으면 그만큼 앞으로도 risk 없이 견딜 가능성이 크다

3. 어떤 질병은 나이가 risk로 작용할 수 있다

4. 측정값과 측정값을 얻는 동안의 상태의 관련성(좌심실 예시)

5. 시간이 지남에 따라 치료법의 발전

6. 얼마나 증상이 심한지

1. **Data reduction**

가정에 위배되거나 환자 샘플링 잘못하면 모델 성능 안 좋다

1. **Verifying model assumptions**
2. **Quantifying predictive accuracy**

예측 쓰이는 3가지

1) 예측 인자 같은 거 확인할 때

2) over(lack )fitting 확인하는 것

3) 방법이나 모델 비교할 때

[5.1]

1. Expected squared error -> average square of (predict – observed) => 보통 MSE로사용된다
2. Calibration -> bias의 정도를 나타냄(평균이 일치하는 정도)
3. Discrimination -> 다른 반응 갖는 환자 구분하는 능력(비가 일년에 얼만큼 오는지뿐만 아니라 언제가 건기인지까지 알아낼 수 있는 사람)

* 평균만 맞아서는 의미가 없다.

[5.2] uncensored outcomes

Discrimination은 mse와 관련이 있다.



R^2 = 1, the model is perfectly able to separate all patient responses based on the predictor variables, and MSE = 0.

[5.3] Discrete or censored outcomes

결과값이 이분적이고 event에 대한 확률이라면 calibration과 discrimination이 mse 하나만으로도 의미 있는 정확도를 말해줄 것이다.

\*calibration 계산하는 방법

환자 하위 그룹 만들고, 예측한거랑 관측한 거에 대한 bias 확인하기

+ 스무싱

[5.5]

Discrimination은 formation of subgroups 나 smoothing이 필요하지 않고, correlation만 알면 얻어낼 수 있다. (**Rank correlation** -> 결과쌍 비교)

Censored 된 dependent variable 처리할 때 mse 사용하지 않음 -> c index사용함

근데 이 discrimination을 예측하는 c index는 예측결과랑 순위 매겨서 하는거랑 관련있다.

* The ***c*** index is defined as the proportion of all usable patient pairs in which the predictions and outcomes are concordant.

Predictor variables의 집합을 통해서 예측 정보 얻는 것

-둘 중 하나는 사망한 모든 환자 쌍에 대해서 계산함

-예측된 생존 시간이 한 명이 길면 그 쌍은 ‘concordant’ 하다고 함

-비교 대상 중 A 죽고 B살면, B가 더 오래 산 것

-만약 예측 생존이 동일한 경우, concordant한 쌍의 수의 분자에 1이 아닌 1/2를 더함 / 분모에는 1 더함 -> 여기까지는 **usable**하다고 여김

-근데 (둘다 동일한 시간에 죽기) or (한 명 죽고, 나머지가 살아 있지만 죽은 사람보다 더 짧은 시간 살아있기) -> **unusable**

0.5 -> random / 1 -> perfect

근데 (예측,결과) 쌍에서 (0.01, 0) / (0.9, 1) 이것과 (0.05, 0), (0.8,1) 같이 미세한 것을 구분하지 못함 (비유하자면, 정성적으로는 가능하나 정량적으로는 구분할 수 없다는 걸로 이해)

1. **Model validation methods**
2. **Summary of modelling strategy -> 차례대로 정리해놓음**