8장. 분석의 실행 (Performing the Analysis)

2005. 11. 2.
의료관리학교실, 김소영

*** 앞의 내용**

8.1 분석시에 변수 코딩번호 할당 방법
8.2 참고치 선정 방법
8.3 상호작용변수(interaction term) 처리방법
8.4 proportional hazards나 생존분석에서 시간(time)을 적용하는 방법

*** 지금부터 다룰 내용**

8.5 연구시작시점에서 outcome을 나타낸 연구대상자의 처리방법
- 진행이 빠른 질환은 시간(hour)단위로 처리
- 진단 후 당일 사망한 경우, 일반적으로 0.5일로 처리
- 한 달 단위로 진단, 사망을 기록함으로써 생존기간이 0으로 처리된 경우, 탈락자에 대한 특성을 별도로 분석하여 기술해주어야 함.

8.6 연구대상자의 생존기간이 짧은 경우 처리방법
- 미연에 방지하기 위하여 대상자 등록 전에 등록기준(pre-enrollment criteria)을 엄격히 정하여 선정
- 등록자 중에서 제외시켜야 할 대상자에 대한 배제기준(exclusion criteria)을 미리 선정

8.7 독립변수의 결측값(missing data) 처리방법
1. 결측값을 갖는 모든 case를 연구에서 제외시킴
2. 이변량 변수로 결측값을 표현
3. 추가 자료를 확보
4. 독립변수의 수를 감소시킴
5. 결측 cases의 값을 추정하여 결측값을 채움

*** 다음 주에 다룰 내용**

8.8 종속변수의 결측값(missing data) 처리방법
8.9~11 변수선정방법
8.12 Tolerance 선정
8.13~15 converge

8.5 연구시작시점에서 outcome이 이미 나타난 연구대상자를 어떻게 처리할 것인가?

○ 이 경우에 대상자의 연구진행시간은 0인데, 정의상으로 연구시작시점에서 연구대상자가 outcome을 보여서는 안 되므로, 이 대상자는 제외되어야 함. 이런 상황을 사전에 막을 수 있는 방법은?
○ 우선, 실제로 outcome이 연구시작시점에 발생한 경우와 그렇게 기록된 경우를 분별해야 함. 가령, 성인성호흡곤란증후군(adult respiratory distress syndrome, ARDS)과 같이 빠르게 진행하는 질병에

대한 병원생존(hospital survival)을 분석할 경우, 몇몇 환자는 병원입원 당일 날 사망할 수 있음. 이런 경우, 하루 단위로 생존율을 분석하면 연구진행시간이 0이 되어 해당 대상자는 탈락되지만, 한 시간 단위로 생존율을 분석하면 탈락을 막을 수 있음.
○ ARDS처럼 진행이 빠른 질병의 경우에는 시간을 분석단위로 쓰면 좋지만 대개 경우에 임상경과가 시간단위로 바뀌지는 않으므로 하루를 분석단위로 사용함.

○ 더 복잡한 경우를 생각해 보자.

○ 샌프란시스코의 AIDS 등록부는 하루 단위로 AIDS의 진단 및 사망을 기록하는데, AIDS로 진단받은 당일 사망한 경우에는 어떻게 처리할 것인가? 이런 경우의 예로 2가지 상황이 있을 수 있음. 하나는 AIDS로 진단받지는 않았으나 기존의 HIV 감염자였던 사람이 병원에 입원한 당일 AIDS 진단을 받고 사망한 경우임. 이 경우에는 ARDS와 마찬가지로 시간단위로 생존율을 분석해야 함. 그러나 의료기록에는 진단 날짜만 기입되어 있어서 시간을 알 수 없다는 문제가 발생함. 다른 하나는 부검관이 AIDS를 진단한 경우임. 이 경우는 실제로 생존기간을 알 수 없음. 그럼, 두 경우 모두 생존시간을 0으로 처리할 것인가?
○ 이에 첫 번째 경우는 진단 후에 사망한 것이므로 생존시간을 0.5일로 처리해주고, 두 번째 경우는 실제 생존기간을 알 수 없으므로 결측처리함.

○ 뉴욕의 AIDS 등록부처럼 한달 단위로 AIDS의 진단 및 사망을 기록하는 경우는 더 복잡한 문제가 발생함. 한 달 동안에 진단과 사망이 동시에 발생한 경우, 생존시간은 0으로 처리되고 많은 대상자가 탈락되는데다 생존시간이 짧은 경우들이 모두 탈락되어 분석결과도 달라짐. 이 경우에 연구자는 탈락된 대상자들이 다른 참여자와 다른 특징이 있지 않은지를 우선 살펴야 함. 실제로 탈락된 대상자들은 여성, 유색인종, 약물복용자들이 많았고 이 점은 또 다른 중요한 부분이므로 기술해주어야 함.

8.6 대상자의 생존기간이 일반적인 경우보다 짧을 경우, 어떻게 처리할 것인가?

○ 대상자가 일반적인 자연경과보다 빨리 사망한 경우 진행경과가 늦은 것으로 알려진 질병에 대한 해석을 어렵게 만듦. 예를 들어, 암발생률 연구 대상자가 등록된 후 1주일 만에 폐암으로 진단받은 경우 어떻게 처리할 것인가? 암세포가 분화하기까지는 수년이 걸린다는 것은 이미 알고 있는 사실이므로 이 경우를 탈락시킨다고 한다면, 1달 만에 진단된 경우나 1년 만에 진단된 경우는 또 어떻게 처리할 것인가?

○ 우선 이런 상황을 미연에 방지하기 위하여 등록기준을 엄격하게 적용해야 함. 폐암의 예에서는 등록 전에 호흡기 증상을 검사하고 등록 전에 흉부 방사선 촬영을 실시함.
○ 그러나, 어떤 질병은 매우 침습적인 검사를 해야만 진단이 되는 경우가 있음. 예를 들어, HIV 감염자에서 잘 발생하는 Pneumocystis carinii 폐렴(PCP)의 경우, 증상이 거의 없고 정상 흉부 방사선 소견을 보임. PCP 예방연구에 앞서서 PCP 발생유무를 진단하려면, 기관지경 검사를 실시해야함. 그러나 이것은 침습적인 시술이어서 무증상환자를 등록에 앞서 실시하기에는 무리가 많음.

○ 또한 PCP의 경우 진단이 어려울 뿐만 아니라 질병진행경과도 느려서 잠복기가 1주일 이상임. 따라서 PCP 예방연구의 대상자가 예방치료를 받고 1주일 만에 PCP로 진단 받은 경우, 치료실패로 볼 것인가 아니면 분석에서 제외시킬 것인가?
○ 대부분의 연구자들은 등록 후 28일 이내에 PCP가 발생하면 연구대상에서 제외시키고, 28일 이후에 발생하면 치료실패로 간주함.
○ RCT의 경우에 믿기 어려울 정도로 초기에 발생한 결과들은 연구 분석 시에 전혀 다른 위치에 고르게 분포하는 양상을 보이기 때문에 치료실패율을 높임에도 불구하고 분석에서 bias로 작용하지는 않음. 그러나 관찰연구의 경우에는 믿기 어려울 정도로 초기에 발생한 결과들이 분석에 포함되었을 경우 RCT와 같은 규칙적인 분포양상을 보이지 않아서 연구의 bias로 작용할 수 있음.
○ 분석에 앞서서 배제기준(exclusion criteria) 또한 설정해야 함.

8.7 독립변수의 결측값은 어떻게 처리할 것인가?

○ 결측값을 처리하는 일은 모든 분석에서 늘 문제가 되는 부분인데 이변량 분석에 비해 다변량 분석에서는 더 큰 문제임. 왜냐하면 각각의 대상자들이 각각 다른 독립변수에서 결측값을 보일 것이기 때문임. 예를 들어, 300명의 대상자와 10개의 독립변수가 있고 각 변수 당 10개의 결측대상자를 갖는 경우에 이변량 분석에서는 표본의 크기가 290이고 이 수는 전체표본 수의 97%임.
○ 그러나 다변량 분석에서는 결측대상자는 10명이상으로 커짐. 극단적으로 10개의 독립변수들이 각기 다른 대상에서 결측값을 보이는 경우에 변수 당 10개씩을 제외시키게 되므로 총 100개의 사례를 잃게 됨. 그런 경우 총 대상자는 200명으로 처음의 66%에 그침. 더더구나 남은 대상이 결측 대상과 체계적으로 다른 경우에 연구결과를 일반화할 수도 없음.
○ 그러나 일반적으로 하나의 변수에 결측값을 보이는 대상자는 다른 변수에도 결측값을 보이는 경우가 많음. 또한 제외되는 사례는 100개 이하여야 함.

○ 결측값을 어떻게 처리할 지를 결정하기에 앞서 결측값을 갖는 자료의 개수를 파악하는 것이 좋음. 개수를 파악하는 방법은 독립변수 하나라도 결측인 경우는 1로 코딩하고 결측이 전혀 없는 경우는 0으로 코딩한 후에 단순빈도(simple frequency)를 측정하면 됨.
○ 결측값을 갖는 자료의 개수를 파악한 후에는 표 8.5와 같은 방법을 사용하여 결측치를 다룸.

표 8.5 다변량 분석에서 결측값을 처리하는 방법

1. 결측값을 갖는 모든 case를 연구에서 제외시킴
2. 이변량 변수로 결측값을 표현
3. 추가 자료를 확보
4. 독립변수의 수를 감소시킴
5. 결측 cases의 값을 추정하여 결측값을 채움

가. 결측값을 갖는 모든 case를 연구에서 제외시키는 방법

○ 결측값을 갖는 모든 case를 연구에서 제외시키는 방법은 임상연구에서 가장 흔히 사용하는 간단한 방법임. 그러나 제외시킨 사례의 특성이 남아있는 자료와 다를 경우 분석에 bias로 작용함. 그 예로는 결측값이 있는 case에서 조사에 덜 순응적이었거나, 조사자에 대한 신뢰도가 더 낮거나, 인지장애가 있었던 경우 등이 있음. 따라서 중요한 독립변수에 대해서는 결측자료와 그렇지 않은 자료를 갖는 대상자를 비교하고 나서 중요한 차이가 있을 경우에는 기록해둠. 차이가 없는 경우에는 결측자료를 기꺼이 제외시킴. 그러나 미처 고려하지 못해 놓치는 중요한 요인이 있을 수도 있음을 늘 인지해야 함.
○ 다변량 분석에서 case를 제외시키기 위한 계획을 세울 때는 우선 단변량 또는 이변량 분석에서는 그 case를 어떻게 처리할 지에 대해서 결정해야 함. 두 가지 방법이 있는데, 하나는 단변량 분석을 시작하면서부터 결측치를 갖는 case를 제거하는 것이고, 다른 하나는 다변량 분석을 시작할 때까지는 일단 남겨두는 것임. 임상연구에서는 대부분 전자를 선호함. 그러나 전자의 방법은 단변량 분석에서 제거한 자료가 다변량 분석에서는 유의미한 자료일 수 있어서 초점을 잃게 될 수도 있다는 단점이 있음. 그럼에도 불구하고 각각의 분석마다 표본크기가 변하면 분석논문을 쓰기가 어려움.
○ 결측자료가 여러 변수에서 산재되어 나타날 경우에는 결측값을 갖는 case를 제거하는 것이 합리적임. 그러나 하나 또는 두개 정도의 변수에서만 대부분의 결측값을 갖는 경우에는 모든 분석방법에서 동일한 표본의 크기를 유지하기 위하여 case를 줄일 필요는 없음. 단, 논문에서 각 분석마다 표본의 크기가 다름을 꼼꼼하게 기술해야 함.

나. 이변량변수로 결측값을 표현하는 방법

○ 이 방법은 신장이식실패의 결정요인에 대한 연구에서 결측자료를 처리할 때 사용됨. 변수를 cold ischemia time에 따라서 6개의 이분변수로 코딩함: 9-16시간(예/아니오), 17-24시간(예/아니오), 25-36시간(예/아니오), 37-48시간(예/아니오), 48시간 이상(예/아니오), 결측값(예/아니오). 참조치는 0-8시간임.
○ 이 방법의 장점은 첫째, 모든 대상자가 분석에 포함된다는 점임. 둘째, 결측자료로 인해 발생한 bias를 분석할 수 있음. 셋째, 연구자는 결측자료가 포함하는 이변량변수로 구성된 모델과 결측자료를 갖는 모든 case를 제거한 모델이 크게 다르지 않음을 보고할 수 있음.

다. 추가자료를 확보하는 방법

○ 이 책은 일차적으로 자료를 분석하는 것에 대하여 기술하고 있으므로 추가자료를 확보하는 방법도 전략으로 언급될 수 있음. 물론 자료를 획득하는 가장 적절하고 효과적인 시기는 자료를 수집하는 단계이지만 자료를 분석하기 시작했다고 해서 자료를 획득할 시기가 이미 지난 것은 아님. 따라서 결측자료가 중요한 것일 경우, 추가적으로 결측값을 채울 수 있고 이는 수고스러움에도 불구하고 중요한 작업임.

라. 독립변수의 수를 감소하는 방법

○ 분석에 무리가 없는 수준에서 독립변수의 수를 감소하는 것도 하나의 방법임. 7.2장에서 표본의 크기가 변수의 수에 비해 적을 경우에 변수의 수를 줄이는 것과 마찬가지로 결측자료가 있는 경우에도 마찬가지로 적용할 수 있음.
○ 첫째, 특정 변수에서 결측값이 두드러지게 많은 경우에 그 변수를 삭제할 수 있음. 둘째, 두 변수

가 연관성을 갖는 경우에 결측값이 더 많은 변수를 삭제할 수 있음. 예를 들어, 교육과 수입은 높은 상관관계를 가지므로 교육은 결측값이 없고 수입은 결측률이 25%인 경우에 수입을 변수에서 제거함. 그러나 수입을 분석에서 제외할 경우, 교육과 수입이 동일한 것은 아니므로 수입에 대한 내용을 보고할 수 없음. 따라서 결측자료를 포함함으로써 잃는 게 많은지 아니면 변수를 제거함으로써 잃는 게 더 많을지에 대해서는 연구자 스스로 결정해야 함.

○ 8.9에서 설명할 변수선택방법을 살펴보면 도움이 될 것임. 변수를 선택한 후에 선정한 변수만으로 통계를 돌리고, 결측자료가 많은 변수를 하나씩 제거하면서 통계를 돌려서 어떤 변수가 적합한지를 찾음.

마. 결측 case의 값을 측정하는 방법

○ 이 방법은 가장 만족스러우면서도 가장 위험한 방법임. 가장 만족스러운 이유는 자료손실이 없기 때문이고 가장 위험한 이유는 분석결과의 bias를 예측하기가 쉽지 않기 때문임. 표 8.6은 단면연구나 종단연구에서 결측치를 측정하는 몇 가지 방법임.

표 8.6 결측치 측정 방법

단면연구
표본평균을 할당
subgroup에 따라 조건부 평균(conditional mean)을 할당
Simple imputation : 공변량(covariate)을 사용하여 결측값을 채움
Multiple imputation : 공변량을 사용하여 결측값을 채우고 무작위 요소(random component)를 추가함
종적연구(대상자를 반복 측정하는 경우)
마지막(가장 최근의) 관찰값으로 결측값을 채움
일련의 값들에 근거하여 결측값을 만듦

(1) 표본평균을 쓰는 방법

○ 독립변수에 결측값을 할당하는 가장 단순한 방법은 결측값을 갖는 변수에 표본평균을 할당하는 것임. 만약 분포가 치우쳐져 있는 경우에는 중앙값을 선택. 결측값이 무작위로 일어나는 경우에 평균/중앙값을 할당하는 것은 가장 좋은 방법임. 이 방법의 장점은 결측값을 모두 분석에서 잃지 않는다는 점임.

○ 그러나 이 방법은 대상자가 하나 또는 두개의 독립변수에 대해서만 결측치를 가질 때에만 분별력이 있음. 예를 들어, 10개의 독립변수 중에서 하나의 변수에서만 15개의 case에서 결측값을 보인 경우에, 15개의 결측값에 모두 평균 또는 중앙값을 할당 할 수 있음. 이 경우 다른 9개의 변수는 실제값이므로 유용함.

(2) subgroup에 따라 조건부 평균을 쓰는 방법

○ 예를 들어, 수입에 대해서만 결측값을 갖는 경우에 수입은 교육 및 직업과 상관관계를 가지므로 표본전체의 평균/중앙값을 할당하기 보다는 동일한 교육수준과 직업상태인 대상의 평균/중앙값을 할당하는 것이 더 적절함.

(3) Simple imputation : 공변량을 사용하여 결측값을 채움

○ 다중선형 또는 로짓회귀분석을 이용하여 다른 독립변수를 분석함으로써 결측값을 예측하는 것. 이 방법은 결측값을 좀 더 정확하게 예측할 수 있으나, 변수가 상호관련 되어 발생하는 오차를 과소평가할 수 있다는 문제점이 있음. 왜냐하면 일단 회귀분석에 근거하여 결측값을 채우면, 컴퓨터는 관찰값과 회귀분석에 근거하여 채운 결측값을 구분하지 못하기 때문임.

(4) Multiple imputation : 공변량을 사용하여 결측값을 채우고 무작위 요소(random component)를 추가함

○ simple imputation의 문제점을 극복하기 위하여, multiple imputation에서는 무작위 요소(random component)를 추가함. 결측값을 갖는 변수와 가장 상관성을 갖으면서도 결측치가 없는 변수의 대상자를 이용하여 다중회귀 또는 로지스틱 분석을 실시함. 우선, 적합한 모델을 이용하여 결측값의 평균과 분산을 측정함. 다음으로 간격변수는 정규분포를 보이고 이변량 변수는 이항분포를 보인다는 가정 하에, 각 결측치에 대하여 임의번호생성기(random number generator)를 사용하여 관측치를 예측함. 그 다음으로 결측값을 채워서 만든 데이터셋을 사용하여 분석하는 과정을 최소한 10회 이상 시행하여 적당한 값을 채움. 이러한 반복과정에서 결측값 대신에 채워 넣는 값이 변하기 때문에 표준오차를 구할 수 있게 됨.

(5) 마지막(가장 최근의) 관찰값으로 결측값을 채움

○ 종적연구에서 대상자를 반복 측정하는 경우, 가장 최근에 선행된 관찰값이 가장 합당한 추정치라는 가정 하에 가장 최근의 관찰값으로 결측값을 대신 채울 수 있음.

(6) 일련의 값들에 근거하여 결측값을 만듦

○ 앞의 방법보다 좀 더 복잡한 방법으로 대상자를 반복 측정하다가 발생한 결측값을 채우는 방법은 prior, subsequent, 또는 prior and subsequent 값을 결측값으로 채우는 것임.
○ 예를 들어, HIV 감염자에 대한 연구에서는 주로 CD4 림파구의 수를 연속적으로 측정하게 되는데 대상자에서 CD4 림파구 값이 한 번 결측 된 경우에, 항레트로바이러스로 치료를 하면 CD4 림파구의 수가 시간경과에 따라 감소하는 점을 고려하여 실제관찰값을 사용하여 회귀선을 그려서 결측값을 예측함.

○ 이제 까지 언급한 방법들을 혼합하여 사용할 수도 있음. 예를 들어 Halfon 등이 실시한 라틴계 아

동의 보건의료접근성 연구에서 수입변수의 결측률이 전체 표본의 13%에 달했음. 이에 수입변수의 결측값은 표본 평균값으로 대치함과 동시에 수입변수 중 결측값을 보인 case를 이분변수로 표현함. 이 방법으로 표본 손실하지 않을 뿐만 아니라 결측값을 갖는 case와 그렇지 않은 case의 차이를 확인할 수 있었음.

○ 다음은 복잡한 경우들에 대한 필자의 guideline임.

1. 결측값이 최소화 되도록 자료를 수집
2. 독립변수 당 결측률을 평가
3. 다른 변수들에 비해서 결측률이 높은 하나 또는 두개의 독립변수의 경우, 아무리 중요한 변수일지라도 결측값이 많으면 bias 될 가능성이 많으므로 삭제를 고려
4. #1과 #3번 과정을 통해 결측자료를 최소화 한 이후에 결측값을 갖는 case의 수를 체크함. 모든 분석에서는 표본수가 동일한 것이 논문을 쓰기 수월하므로 결측값을 갖는 case 수가 적은 경우에는 결측값을 갖는 case를 삭제하고 분석에 들어감.
5. 결측값을 갖는 case 수가 많은 경우에는, 결측값을 갖는 case가 그렇지 않은 case와 다르지는 않는지 조사
6. 만약 두 경우가 다르지 않다면, 표본 평균이나 조건부 평균을 할당할 수 있음. 그에 앞서, 모든 case는 변수의 반 이상이 관찰값을 가져야 함. 만일 case에 포함되는 대부분의 독립변수가 결측값을 갖는 경우에는, 해당 case를 삭제함. 그러나 통계적인 근거가 타당한 경우에는 multiple imputation을 사용하여 결측값을 채워넣는 것도 고려해봄.
7. 두 경우가 다른 경우에도 #6번과 같은 방식을 사용함. 하지만 결측값을 갖는 case와 그렇지 않은 case의 다르기 때문에 해석이 쉽지 않음을 유의해야 함.
8. 가능한 결측값을 처리하는 방법은 1가지 이상을 사용하는게 좋음.
9. 결측값을 다룰 때는 이론과 실제에 있어서 신중을 기해야 함.