3. Outcome Variables in Multivariable Analysis
(다변수 분석에서의 결과변수)

2005. 9. 28.(수)
의료관리학교실
김부경

1. 어떤 종류의 다변수 분석을 선택할 것이냐에 영향을 미치는 결과변수의 특성

표 3.1에서 보듯이 다변수 분석의 선택은 가장 먼저 얻고자 하는 결과 변수의 종류에 의존한다. 간격변수(연속변수라고도 함)에서는 한 직선 상에서 각 단위(interval)는 같은(명목적으로) 양의 가치를 지닌다. 예를들어 간격변수에는 혈압, 체중, 체온과 같은 것이 있다. 이러한 예들에서 한 단위의 변화는 시작 시점과 관계없이 동일한 양의 수은주, 파운드(또는 킬로그램) 등의 변화를 의미한다.

이분변수(가장 간단한 종류의 변수)는 한 시점에서 두 개의 배타적인 가치(범주)를 가진다. 예를들어 생과 사, 암 발생의 유무와 같은 것이다. 이분적 상황이 발생한 시기는 어떤 사건이 있었던 것인데, 사망 시각 또는 암 발생 시기와 같은 것인데 일정 기간(예: 5년)동안의 발생한 것을 말한다.

표 3-1. 다변수 분석의 선택을 결정하는 결과 변수의 종류

결과 변수의 종류	결과 변수의 예	다변수 분석의 종류
간격	혈압, 체중, 체온	다중 선형 회기분석
이분	사망, 암, 중환자실 입원	다중 로지스틱 회기분석
이분적 상황의 발생 시기	사망 시간, 암 발생 시기	위험률 분석 (proportional hazards analysis)

정의: 간격변수(interval variable)에서는 한 직선 상에서 각 단위(interval)는 같은 양의 가치를 지닌다.

정의: 이분변수(dichotomous variable)는 두 개의 배타적인 가치를 가진다.

2. 결과변수가 서열이나 명목변수일 경우 어떻게 해야 하는가?

서열 변수는 순서를 정할 수 있는 다양한 범주를 가진다. 예를들어 뉴욕 심장 학회의 심장 기능의 기능적 분류(New York Heart Association's functional classification of cardiac function)와 같은 것이다. 4가지 등급(level)의 순서를 정할 수는 있지만, 1등급(신체적 활동에

제한이 없는 상태)과 2등급(신체적 활동에 약간의 제한이 있는 상태) 사이에 명목적인 양적 차이는 없다. 1등급과 2등급 사이의 차이는 3등급(신체적 활동에 현저한 제한이 있는 상태)과 4등급(불편함이 없이 어떤 신체적 활동도 할 수 없는 상태) 사이의 차이와 같지 않다.

명목변수는 순서를 정할 수 없는 다양한 범주를 가진다. 예를들어 사망 원인(암, 심장질환, 감염 등)과 같은 것이다. 서열변수와는 달리 사망 원인에 순서를 정할 수는 없다.

표 3-1에서 볼 수 있듯이, 서열변수나 명목변수는 이 책에서 논의되는 세가지 다변수 분석에서 사용되지 않는다.

서열변수나 명목변수를 사용할 수 있는 한가지 방법은 이것들을 이분적 결과로 변환시키는 것이다. 예를들어 뉴욕심장학회의 분류는 종종 1-2등급(약간의 호흡곤란)과 3-4등급(심한 호흡곤란)으로 구분된다. 이와 비슷하게 사망의 원인이 암인가의 예, 아니오로 될 수 있다. 그러나 분명히 이러한 그룹화 방법은 자료에 대한 정보를 소실하게 된다.

대신 자료들은 로지스틱 회귀분석의 변형을 이용하여 분석할 수 있다. 서열결과(ordinal outcomes)는 proportional odds logistic regression을 이용하여 분석할 수 있고, 명목 결과(nominal outcomes)는 polytomous logistic regression을 이용하여 분석할 수 있다. 이러한 기술들은 의학 연구에서 잘 사용하지 않기 때문에 이 책에서는 논의하지 않을 것이다. 그러나 독자들은 이러한 방법에 대한 정보를 다른 자료들을 이용하여 얻을 수 있다. 명목 결과를 사용할 수 있는 다른 기술은 discriminant function 분석이다. 이 방법은 여기서 논의되는 세가지 방법과 공통점과 차이점을 동시에 가지고 있다.

여기서 간격변수와 서열변수를 구별하는 것을 설명했지만 실제로는 gray area가 존재한다. 임상 연구자들, 특히 행동 과학에 관심이 있는 사람들은 간격변수의 정의에 정확히 맞지 않는 결과 변수들을 분석하기 위해 종종 다중 선형 회귀분석을 사용한다. 예를들어 환자의 만족도, 환자 자신의 건강 지각, 통증이나 스트레스의 정도와 같은 것이다. 이러한 변수들은 전형적으로 1에서4, 1에서5, 1에서100과 같이 정한 임의의 명목척도로 응답하도록 만든다. 이러한 척도는 1=매우 좋음, 2=좋음, 3=보통, 4=나쁨, 5=매우 나쁨, 또는, 1=매우 찬성, 2=찬성, 3=잘 모름, 4=반대, 5=매우 반대와 같은 단서들이 제시된다. 이러한 척도들은 실제로 등간이 아니다. 매우 좋음과 좋음 차이는 보통과 나쁨의 차이와 같을 필요가 없기 때문이다. 그럼에도 불구하고 이러한 변수들을 다중 선형 회귀분석의 종속변수로 사용하고자 한다면, 독립변수와의 관계가 다중 선형 회귀분석의 가정(assumption)을 만족시키는 한 사용할 수 있다.(Section 5.1-5.4)

정의: 서열변수(ordinal variable)는 순서를 정할 수 있는 다양한 범주를 가진다.

정의: 명목변수(nominal variable)는 순서를 정할 수 없는 다양한 범주를 가진다.

3. 어느 한 시점에서 이분적 사건의 단순 축적 결과를 사용하는 것보다 사건의 발생 시간을 사용했을 때의 장점은 무엇인가?

단면연구와 이분적 결과가 있을 때 보통은 다중 로지스틱 회귀분석을 사용한다. 여기서 이

분적 결과의 시간을 측정하기 위해 위험률 분석을 사용할 가능성은 없다. 단면연구에서는 독립변수들과 결과가 같은 시점에서 측정되기 때문이다.

그러나 종적연구(longitudinal syudy)에서 결과는 독립변수를 측정한 후에 발생한다. 이분적 결과를 측정하는 것에서 자료를 어떻게 분석할 것인가의 두가지 선택이 있다.

TIP: 임상의학은 치유(cure)보다는 치료(treatment)일 경우가 더 많다.

특히 특정시점에서는 단순 축적 결과를 사용할 수 있다. 3년 동안 대상자들 중 심장 발작을 경험한 사람의 비율과 같은 것이다. 이 경우에 단순 축적결과 대신 심장 발작의 시간을 사용할 수 있다.

특정 시점에서의 축적 결과를 제시하는 것이 더 간단한데 왜 많은 임상 연구자들은 결과 시간을 사용할까?

한가지 중요한 이유는 임상 의학은 치유보다는 치료가 더 많다는 것이다. 이렇게 주어졌을 때 문제가 되는 것은 질병이 얼마나 일찍 발생했느냐와 생존율이 얼마나 증가했는가이다.

그림 3.1은 이것을 나타내고 있다. 두개의 선을 볼 수 있는데, 치료를 받은 그룹과 받지 않은 그룹이다. 두 그룹에서 모두 2년 동안 95%의 환자가 사망하였다. 그러나 1년째에는 두 그룹에서의 사망률이 현저히 다르다. 치료군에서 환자의 사망률은 서서히 상승한다. 1년째에서는 치료하지 않은 그룹의 사망률은 48%인 반면, 치료한 그룹에서는 5%밖에 사망하지 않았다.

두 번째 이유는 발생 시간 모델은 분석 대상의 추적관찰(follow-up) 기간의 차이를 제공한다. 종적연구(longitudinal study)에서 추적관찰 기간의 차이가 발생하는 이유는 여러 가지가 있는데, 연구 중 탈락되거나, 연구가 중단되거나, 연구 대상자가 증가하기 때문이다. 만약 3년간의 심장 발작 환자들의 단순 축적 결과를 사용한다면, 2년 반 동안에 탈락된 대상자까지 포함하게 되고, 이 대상은 분석에서 제외해야 할 것이다. 그러나 발생 시간 모델을 사용할 경우 이 대상자들은 사건이 없는 2년 반 동안에도 기여를 할 수 있다.

대상자들의 다양한 추적관찰 기간을 통합할 수 있는 방법으로 censoring이 있다. Censoring은 연구에 완전하지 못한 대상을 제외하는 것보다 통계적으로 더 큰 설득력이 있다. 이 방법은 7.3-7.5장에서 더 상세히 기술된다.

여기서 가능한 질문은 만약 대략적인 예후가 같다면, 특히 가격의식(cost-conscious)적인 이 기간동안 몇 개월의 생존율 증가가 얼마나 중요한가이다. 이 질문에 대한 대합은 통계적이라기보다는 철학적이다. 일반적으로 결과의 시간은 질병이 경한 경우보다는 중증인 경우에서 더 문제가 된다. 생명의 위협을 주는 질환을 가진 환자일 경우 며칠 혹은 몇 달이라는 시간은 가치를 지니는데, 특히 그 시간들 속에서 자녀의 대학 졸업을 볼 수 있다거나 손녀의 첫 걸음마를 볼 수 있다면 더욱 그러하다.

반면, 경증인 경우 시간의 연장은 임상적으로 근 의미가 없다. 예를들어 수두에 걸린 어린이를 acyclovir로 치료하는 것은 플라시보에 비해 열이 내리는 시간과, 수포가 감소되는 시간을 단지 하루 감소시킬 뿐이다. 만약 연구자의 결과 관찰 시점이 7일째라면 acyclovir의 아무런 효과도 관찰 할 수 없을 것이다. 면역능력이 있는 어린이라면 7일째에는 치료를 받은 것과 안받은 것의 차이가 없기 때문이다. 증상을 하루 감소시키기 위해 acyclovir에 투자하는

것이 가치가 있을까? 그것이 문제를 일으키면서까지(이제 걸음마를 하는 아기에게 하루에 네 번 약 먹이기를 시도해 본 사람은 이 말의 의미를 알 것이다) 그렇게 할만한 가치가 있는가? 분명 이것은 통계적 의문은 아니다. 실제로 acyclovir의 결과가 통계적으로 중요하다고 할지라도, 대부분의 소아과의사들은 면역능력이 있는 아이에게 이 약을 처방하지 않을 것이다.

결과 시기의 약간의 향상은 실제로 그것이 환자에게 미치는 이득이 거의 없다고 할지라도 과학적 발전과정을 격려할 수 있다. 의학적 진보는 증가하고 있다. 특정 전략을 제공하는 것은 아주 조금이라도 생존율을 증가시킬 수 있고, 그것은 더 나은 치료를 이끌어낼 수 있다.

요약하면 결과 시간(time to outcome)은 특정시점에서의 축적 결과(cumulative outcome)보다 더 민감한 방법이다. 이 방법은 대상자들의 다양한 추적관찰 기간도 포함한다. 만약 상대적으로 적은 결과들을 가지고 있다면 추적관찰 기간이 상대적으로 짧고, 대상자의 탈락도 적을 것이다. 이럴 경우 로지스틱 회귀분석이나 위험률 분석을 사용하는 것이 비슷한 결과를 얻을 수 있다. 이것은 coronary angiography를 시행하고 coronary revascularization이 필요한 것으로 판단되는 환자들의 사망률 연구에서 나타난다. 671명의 환자들 중 70명(10.4%)은 이미 사망하였고, 추적관찰 기간의 중앙값은 797일이며, 모든 대상자의 사망 자료를 사용하였다. 다중 로지스틱 회귀 분석에서, 가능성이 있는 교란변수들을 보정한 후, 1년째에서 coronary revascularization을 받은 환자들의 사망률의 odds가 시술을 받지 않은 환자들보다 중요하게 낮음을 볼 수 있다(OR = 0.49; 95% CI = 0.30-0.84). 위험률 분석에서는 교란변수들을 보정한 후, revascularization이 중요하게 사망 위험률을 감소시킨다는 것을 알 수 있다(relative hazard(RH) = 0.59; 95% CI = 0.36-0.97).

TIP: 발생시간 모델은 대상자의 서로 다른 추적관찰(follow-up)기간을 포함한다.

정의: Censoring은 대상자의 추적관찰 기간을 통합하는 방법이다.

TIP: 발생시간은 경증보다 중증일 경우 더 중요한 문제가 된다.

4. Independent Variables in Multivariable Analysis
(다변수 분석에서의 독립변수)

1. 다변수 분석에서는 어떤 종류의 독립변수를 사용할 수 있는가?

간격과 이분 독립변수는 세가지 다변수 분석 모두에서 사용할 수 있다.(표4.1) 서열과 명목변수는 다른 변수로 변환하지 않고는 이 중 어떤 기술에서도 사용할 수 없다.

2. 서열변수와 명목변수일 경우 어떻게 해야 하는가?

포기할 필요는 없다. 서열과 명목척도는 다변수 모델에 적합하도록 여러개의 이분 변수로 전환시킬 수 있다. 이 과정은 보통 “dummying"이라고 하고, 역학자들과 생물통계학자들이 사용한다. 그러나 ‘dummying"과 ”dummy variables"는 전문적 속어이다. 이 과정을 사용하여 다양한 범주의 변수들을 만들 수 있다. (만약 dummying을 사용한다면 내가 그러했던 것처럼 당신도 당신의 논문을 읽는 독자들로부터 많은 문의를 받을 수 있다.)

임상연구에서 인종은 대표적인 명목 변수이다. 인종들 간에는 수적인 순서도 없고, 고정된 간격도 없다. 그러므로 인종은 이분변수가 될 수 있는데(예: 백인, 백인아님), 이는 다변수 분석에서 몇 가지 이분 변수로 나타난다. 아래에 인종을 다섯가지 이분변수로 나타낸 예가 있다.

- 아프리카계 미국인(예/아니오)
- 라틴/히스페닉(예/아니오)
- 아시아/태평양 섬(예/아니오)
- 미국 원주민(예/아니오)
- 다른 유색인종(예/아니오)

어떤 사람이 백인/코카시안인 경우 어떻게 할까? 명목변수를 다변수 분석에 사용하기 위하여 여러 가지 이분변수로 만들 때 필요한 변수보다 하나의 변수가 덜 필요하다. 왜그런가라는 질문의 대답은 컴퓨터의 관점으로 생각해보아야 한다. 만약 5개의 이분 변수를 만들었다면 그것들은 모두 1(예) 또는 0(아니오)에 해당하게 되고, 컴퓨터는 표 4.2에서와 같은 6가지 패턴을 보인다.

이 경우 백인/코카시안이라는 변수는 만들지 않는데, 이것은 다른 5가지 변수들에 의해 결정되기 때문(5개 변수 모두가 0일 경우)이다. 다변수 분석에서 이것을 참조변수라고 한다.

인종을 명목척도의 예로 드는 것은 신중히 선택한 것인데, 이것을 어떻게 코딩하느냐는 연구 대상인구에 따라서 결정되기 때문이다. 예를들어 미국의 북동쪽에서 작은 규모로 시행된 임상연구에서는 미국 원주민이나 아시아/태평양 섬에 해당하는 사람들이 거의 없을지 모른다. 만약 그 그룹이 표본 전체의 5% 이하일 경우 해당 변수를 만드는 것은 통계적으로 중요한 정보를 제공하지 않는다. 이런 경우 조금 더 큰 그룹으로 묶어 “기타”라는 변수로 만들 수 있다. 예를들면 이렇게 나누는 것이다. 아프리카계 미국인(예/아니오), 라틴/히스페닉(예/

아니오), 다른 유색인종(예/아니오), 그리고 백인은 참조그룹이다.

비록 그룹 수를 감소시키는 것이 작은 정보를 전달하는 이분변수를 갖게 되는 것을 막아준다하더라도, 다른 인종들을 한 그룹으로 묶는 것은 자료를 정확히 반영하는 것은 아니다. 심지어 "아시아/태평양 섬"이라는 카테고리를 남겨둔다 하더라도 기억해야 할 것은 이 카테고리 안에도 고유한 언어, 관습, 유전 인자를 가진 십여가지 문화들이 있고, 그러한 요소들은 질병 발생에 모두 영향을 미치게 된다. 다변수 분석의 모든 어려운 문제들처럼 인종을 어떻게 코딩할 것이냐는 통계적인 질문이 아니다. 인종과 같은 명목 변수를 그룹화하는 가장 최선의 길은 연구의 질문이 무엇이냐, 명목변수의 분포(얼마나 많은 사람이 각각의 그룹에 포함되는가), 명목변수와 결과의 범주들 간의 관계와 같은 것들에 달려있다.

다양한 이분변수를 만드는 과정은 다른 방법들에 비해서 서열변수나 명목변수를 통합하는데 유용하다. 이것은 간격변수에 있어서도 동일한데, 변화되지 않은 간격-독립변수와 결과는 모델에 적합하지 않다. 이것은 5.5장에서 더 상세히 기술할 것이다.

3.2장에서 보았듯이, 어떤 변수들은 서열이 있으나 간격변수와 같이 조작한다. 어떤 서열변수를 다중선형회귀분석에서 종속변수로 사용하듯이, 이러한 것들이 모델의 가정에 적합하다면, 이러한 변수들을 세가지 종류의 분석에서 모두 사용할 수 있다.

TIP: 명목변수를 그룹화하는 최선의 방법은 해당 연구의 질문, 명목변수의 분포, 명목변수와 결과간의 이변량 관계에 의해 결정된다.