Relationship of Independent Variables to One Another

2005.10.11
박재현

6.1 독립변수들이 서로 상관관계가 있으면 문제가 되는가?

1.1장에서 설명한 바와 같이 다변량 분석의 장점은 서로 상관관계를 갖고 있는 다수의 독립변수들이 종속변수와 어떻게 연관되어 있는지를 알아낼 수 있는데 있다. 즉, 다변량 분석은 각각의 독립변수들이 종속변수에 미치는 영향을 다른 변수의 영향을 보정한 상태에서 알아낼 수 있게 해준다.

하지만 만약 두개의 독립변수간에 상관관계가 매우 커서, 즉 한 변수를 알면 자동적으로 다른 변수를 알 수 있을 때에는 다변량 분석이라도 이 두개의 독립변수가 종속변수에 미치는 영향을 따로 분리해서 산출할 수는 없다. 이러한 문제를 바로 '다중 공선성(multicollinearity)'이라 한다.

예를 들어 폐렴 환자의 병원 재원일수에 영향을 미치는 요인을 찾는 연구에서, 의사에 대한 선호도를 고려하여 분석하기 위해 간호사가 환자의 체온을 화씨와 섭씨로 모두 기록했다고 하자. 만약 당신이 병원 재원일수를 종속변수로 넣고 화씨 체온과 섭씨 체온을 독립변수로 모두 넣었다고 하면 그 모델은 틀린 것이다. 즉, 당신은 에러 메시지나 예상치 못한 결과를 얻게 될 것이다. 그 이유는 섭씨 체온과 화씨 체온은 숫자는 다르지만 같은 변수이기 때문이다. 즉, 한 변수를 다른 변수로 수학적인 식을 써서 바꿀 수 있다(섭씨 = (화씨 - 32) × 0.56)).

섭씨 체온과 화씨 체온이 같은 변수이기 때문에 당신이 만든 모델은 독립변수가 종속변수에 미치는 영향을 찾아내지 못한다. 하지만 당신이 두 변수를 실수로 같이 모델에 넣지 않는다면 위와 같은 상황이 발생하지 않을 것이다. 이 보다 좀더 흔한 경우는 두 독립변수가 충분히 다르지 않을 경우이다. 예를 들어 Phibbs 등은 출산체중과 재태기간(gestational age)간의 상관성이 매우 높기 때문에 이 두 변수를 모두 신생아 사망을 예측하는 모델에 포함시키기는 어렵다고 하였다.

6.2 어떻게 독립변수들간에 다중 공선성이 있는지 알 수 있는가?

Person correlation 또는 r로 표현되는 상관계수(correlation coefficient)는 두개의 변수가 얼마나 강하게 서로 연관되어 있는지, 즉 상관성이 있는지 측정할 때 쓰인다. 상관계수는 두개의 변수간의 관계가 서로 선형 관계에 있다는 것을 가정한다. 상관계수는 -1에서 1사이의 값을 가진다. 만약 상관계수가 -1 또는 1이라면 이 두 변수는 정확히 같이 움직이는 것이다. 즉, 한 변수를 알면 정확하게 다른 변수를 알 수 있다. 단, 상관계수가 - 1이라는 것은 두 개의 변수가 서로 다른 방향으로 움직인다는 것을 의미한다. 즉, 한 변수가 커지면 다른 변수는 작아진다. 상관계수가 0이면 두 변수가 전혀 상관성이 없다는 것을 의미한다. 상관계수 r을 제곱한 값에 100을 곱하면 두 변수가 얼마나 서로의 정보를 공유하는지를 측정할 수 있으며 이는 0%에서 100%의 값을 갖는다.

앞의 예에서 화씨 온도와 섭씨 온도 사이의 상관계수는 1이다. 즉, 두개의 변수는 같은 정보를 100% 공유하는 것이다. 반대로 Figure 5.1(p 38)에서 보았듯이 예방접종 시행 후 비타민 B_{12}와 Pneumococcal antibody titer 사이의 상관계수는 0.55이다. 이것은 이 두 변수가 정보의 30%($0.55^2 \times 100=30$)를 공유함을 뜻한다.

당신이 분석하고자 하는 모든 변수들끼리의 상관성을 보기 위해서는 상관계수 교차표(correlation coefficient matrix)를 만들어야 할 것이다. 일반적으로 두개의 변수들끼리의 상관계수가 0.9 이상이면 분석에 문제가 발생한다. 만약 0.8 이하이면 문제가 없으며, 0.8에서 0.9 사이라면 문제가 발생할 가능성이 있다.

상관계수 교차표의 문제점은 다른 변수들을 보정하지 않은 체 두개의 변수들 사이의 상관성을 분석한다는 것이다. 이러한 이유로 대부분의 다변량 분석 프로그램은 각각의 회귀계수(parameter estimates)에 대한 상관성 교차표를 만들어준다. 이러한 방법은 다른 변수의 영향을 보정하기 때문에 위에서 설명한 단순한 상관성 분석보다 다중 공선성의 문제를 찾아내는데 더 좋은 방법이다. 이때에도 단순한 상관성 분석의 경우와 마찬가지로 0.9 이상이면 문제가 발생하며, 0.8에서 0.9 사이는 gray area라고 할 수 있다.

예리한 독자들은 이러한 두 가지 방법 모두 한 개의 독립변수가 두 번째 독립변수와 매우 높은 상관성을 가질 때에만 쓰는 방법이라는 것을 알 수 있을 것이다. 그런데 만약 독립변수들의 조합이 나머지 한 개의 독립변수와 연관되어 있을 경우는 어떻게 해야 하는가? 통계학적으로 이러한 경우 또한 앞서 보았던 두개의 독립변수가 서로 높은 상관성을 가지는 경우와 마찬가지로 문제가 된다. 하지만 이러한 경우는 밝혀내기가 힘들다.

우리는 4.2장에서 명목변수를 여러 개의 이분변수(dummy variable)로 전환하는 것을 보았다. Table 4.2 (p 35)를 다시 보자. 당신은 백인의 경우 따로 이분변수로 만들 필요가 없었다는 것을 기억할 것이다. 왜냐하면 다른 인종이 모두 0의 값을 가질 경우 그 변수는 자동적으로 백인이라는 것을 뜻하기 때문이다. 만약 당신이 이러한 사실을 모르고 백인까지 포함한 모든 인종을 이분변수(예/아니로)로 넣는다면 어떠한 일이 발생하겠는가? 그렇게 되면 여러 개의 독립변수들의 조합이 다른 한 변수를 완벽하게 설명하는 상황이 발생할 것이다. 다음 두 질문에 대해서 대답을 하게 되면 당신은 왜 그렇게 되는지 이해하게 될 것이다.

1. 만약 백인을 제외한 다섯 변수들 중 하나라도 '예'라는 값을 가진다면, 백인이 가지는 값은 '예'인가 아니면 '아니오'인가?
2. 만약 백인을 제외한 다섯 변수들이 모두 '아니오'라는 값을 가진다면, 백인이 가지는 값은 '예'인가 아니면 '아니오'인가?

첫 번째 질문에 대한 답은 '아니오'이며, 두 번째 질문에 대한 답은 '예'라는 것을 알 수 있을 것이다. 이러한 경우가 여러 개의 독립변수들의 조합이 다른 한 변수를 완벽하게 설명하는 상황인 것이다. 만약 당신이 백인을 포함하여 모든 인종을 다변량 분석 모델에 넣는다면 당신은 잘못된 결과를 얻을 것이다.

그렇다면 어떻게 여러 개의 독립변수들의 조합이 다른 한 변수를 설명하는 것을 발견해낼 수 있을까? 이러한 경우 'tolerance'와 이와 상보적인 'variance inflation factor'를 이용하여 다중 공선성을 측정할 수 있다. 이 두개는 모두 각각의 변수들의 회귀계수가 다른 변수들로부터 얼마나 설명 가능한지를 나타내준다. Tolerance값이 0.25보다 작으면 문제가 있을 수 있고(worrisome), 0.10보다 작으면 심각한 문제가 있는 것이다(serious). 이와 상보적인 'variance inflation factor는 4보다 크

면 문제가 있을 수 있고(worrisome), 10보다 크면 심각한 문제가 있는 것이다(serious).

대부분의 선형 회귀분석 프로그램은 모든 독립변수들에 대해 tolerance와 variance inflation factor 값을 산출해준다. 만약 어떤 변수가 문제가 있다고 나왔을 때, 당신은 문제가 되는 변수와 상관성이 높은 변수를 추가분석을 통해 찾아낼 수 있다. 즉, 문제가 되는 변수를 종속변수로 넣고 다른 변수들을 독립변수로 넣은 다음 회귀분석을 시행하면 된다. 이렇게 함으로써 어떠한 변수가 문제가 되는 변수와 상관성이 가장 높은지 알아낼 수 있으며, 따라서 어떠한 변수를 모델에 포함시킬지 결정할 수 있다.

다중회귀분석과 proportional hazards analysis에서 보통 연구자들은 상관계수 교차표에 의지하여 다중 공선성을 찾아낸다. 하지만 tolerance가 모델에 넣을 변수를 찾아내는데 사용하는 표준 도구이다. 매우 낮은 tolerance 값을 갖는 변수는 모델에 넣으면 안된다(Section 8.12).

※ 참고 - tolerance와 variance inflation factor[1)]

Tolerance는 공선성(collinearity)를 측정하는데 사용되는 변수이다. 이것은 다른 독립변수들로 설명되지 않는 변수의 variance이다. Tolerance 값을 산출하기 위해서 각각의 독립변수들을 차례로 종속변수로 넣고 나머지 다른 변수들은 독립변수로 넣은 다음 회귀분석을 시행한다. 만약 높은 다중 상관성(multiple correlation)이 발견된다면 그 종속변수는 다른 독립변수들의 조합과 높은 상관성을 가진다고 볼 수 있다. 만약 R^2가 1인 경우 이 변수는 다른 독립변수들과 완벽하게 상관성을 가지게 된다. Tolerance는 간단히 표현하서 $1-R^2$인데, 따라서 R^2가 1이라면 tolerance는 0 (1-1=0)이 된다. 이것은 완벽한 상관성을 나타내는 것이다. 보통 Tolerance는 회귀분석 결과의 한 부분으로 산출된다.

variance inflation factor는 tolerance와 상보적인 관계를 갖는다. 따라서 tolerance가 높은 값을 가지면 variance inflation factor는 낮은 값을 가진다.

다음 그림은 SPSS 결과인데, 분석에서 공선성 분석을 추가하면 다음과 같은 결과를 얻는다. Tolerance 값은 최저 0.657에서 최고 0.906의 값을 갖는다. Tolerance = $1-R^2$ 이므로, 'quality of life in the past month'라는 변수의 tolerance 값이 0.657이라는 의미는 이 변수가 갖는 variance의 34.3%(1-0.657=0.343)를 다른 변수들과 공유한다는 것을 의미한다(즉, R^2=0.343). 다음 표에서 다른 변수들의 tolerance 값은 더 크므로, 이 경우 다중 공선성은 문제가 되지 않는다고 볼 수 있다.

1) Barbara HM. Statistical Methods for Health Care Research. Lippincott. 1997

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	137.078	11.048		12.410	.000	115.272	158.884		
	Smoking history	-.179	2.187	-.004	-.082	.935	-4.497	4.139	.906	1.104
	Depressed state of mind	-21.750	2.700	-.492	-8.057	.000	-27.080	-16.420	.664	1.505
	Overall state of health	3.303	.907	.220	3.642	.000	1.513	5.093	.679	1.472
	Quality of life in past month	5.579	1.651	.208	3.379	.001	2.320	8.839	.657	1.522

a. Dependent Variable: TOTAL IPPA

FIGURE 12-7
Collinearity statistics.

6.3 다중 공선성이 있는 변수들은 어떻게 처리해야 하는가?

상관성이 높은 변수는 다음과 같이 처리할 수 있다.

1. 그 변수를 모델에서 제외한다.
2. “and/or" 조합을 이용한다.
3. 척도(scale)를 만든다.

첫 번째 방법을 쓴다면, 즉 만약 당신이 어떤 변수를 제외하고자 한다면, 제외할 변수를 어떻게 선택할 수 있겠는가? 그 답은 missing data가 많거나 측정의 오류가 더 많거나 다른 이유로 인해 만족스럽지 못한 변수를 제외하는 것이다. 위에서 예로 든 신생아 사망률에 대한 연구의 경우에서 연구자는 재태기간을 제외시키고 대신 출생체중을 모델에 넣었다. 그 이유는 재태기간은 출생체중에 비해 missing data가 더 많고 코딩에서도 신뢰성이 떨어졌기 때문이다. 아이러니하게 이론적으로는 재태기간이 출생체중보다는 신생아 사망률을 결정하는데 더 중요한 변수이다. 따라서 연구자는 Small for gestational age(yes/no), large for gestational age(yes/no)라는 두개의 변수를 모델에 추가로 넣어 출생체중이 재태기간을 정확하게 반영하지 못하는 것을 보정하였다.

두 번째 방법인 "and/or" 조합을 이용하는 것은 변수들이 같은 과정을 공유할 때 가장 적합한 방법이다. 예를 들어 만약 어떤 폐렴 환자에 대해 발한(땀을 흘리는 것) 또는 떠는 증상(shaking)이 있었는지를 물어본다고 했을 때 두 변수는 매우 큰 상관성이 있다는 것을 예상할 수 있다. 왜냐하면 떠는 증상은 발한이 더 심해졌을 때 나타나는 증상이기 때문이다. 하지만 떠는 현상이 있었던 환자 중 몇몇은 초기에 발한이 있었다는 것을 인식하지 못할 수도 있다. 또한 어떤 사람은

발한이 있어 아스피린을 먹어서 떠는 증상이 나타나지 않았을 수도 있다. 이때는 "발한 and/or 떠는 증상"으로 새로운 변수를 만들 수 있다. 즉, 두 증상 중 어느 하나라도 나타나면 "예"로, 두 증상 모두 없었을 경우를 "아니로"로 하는 변수를 만들 수 있다.

세 번째 방법은 척도(scale)를 만드는 것인데, 이 방법은 심리학이나 사회학 자료를 다룰 때 많이 쓰인다. 이 방법은 각각의 변수들의 의미를 모두 종합하는 하나의 변수를 만드는 것인데, 이때 여러 개의 변수들의 값을 합하여 하나의 변수로 만들고 척도를 부여하게 된다. 연구자는 의도적으로 대상자의 응답의 신뢰성을 검증하기 위해 여러 개의 연관된 변수를 만들어 질문을 하기도 한다(예를 들어 두개의 비슷한 질문을 하면서 말을 조금 바꾼다. 이때 대상자는 같은 답을 하는가를 보는 것이다). 이러한 경우 연구자는 보통 조사전에 어떤 질문을 가지고 척도를 만들 것인지 계획하게 된다. 또는 어떤 질문들이 비슷한 정보를 제공하는지 분석하기 위해 요인분석(factor analysis)을 시행하기도 한다.

다중 공선성을 다루는 위의 세 가지 방법은 분석 대상자 수가 적어서 분석에 넣을 독립변수의 수를 줄어야 할 때도 사용할 수 있다. 하지만 이 변수들이 서로 상관성이 높지 않다면 대상자 수가 적다는 이유로 독립변수를 모델에서 제외해서는 안 된다. 모델에 포함할 독립변수를 줄이는 여러 가지 방법들은 7.2장에서 자세히 다룰 것이다.