초록 번호	I - 3 - 3

제 목	국 문	역학연구에서 폭로요인간의 다중공선성의 진단과 해결방법에 관한 연구
	영 문	A Study on the Diagonostics and Proposed Methods of Multicollinearity between Exposure Variables in Epidemiology
저 자 및 소 속	국 문	안윤옥, 김동현, 신명희, 배종면, 박은식 서울대학교 의과대학 예방의학교실
	영 문	Ahn Yoon Ok, Kim Dong Hyun, Shin Myung Hee, Park Eun Sik *Seoul National University College of Med. Dept of Prev. Med.*

분 야	역학연구 방법론	발 표 자	박 은 식 (일반회원)
발표 형식	구 연	발표 시간	15분
진행 상황	연구완료 ( ), 연구중 ( ) → 완료 예정 시기 : 년 월		

1. 연구 목적

우리가 흔히 사용하는 회귀계수 추정방법은 최소 제곱법이다. 그러나 독립변수들간에 상 관 관계가 커지면 최소제곱법의 맹목적인 사용에는 심각한 문제점이 발생하게 된다. 본 연구는 독립변수들간에 상관관계가 있을 때 상관관계를 진단하는 통계량과 그 해결방법을 제시하고자 한다.

독립변수가 x(1), ... ,x(k)이 있다고 하자. 이때 x(i) 를 종속 변수로 하고 나머지 k개의 독립변수에 대하여 회귀 방정식을 적합시킬 때 얻어지는 결정 계수를 R2(i) 라 하자. R2(i)=1이거나 1에 가까우면 완전에 가까운 다중공선성의 관계가 있다고 하겠다. 분산 팽창 계수 VIF(i)는 1/(1-R2(i))로 정의되며,R2(i)의 값이 1에 가까우면 물론 VIF(i)는 커지며 보통 VIF(i)의 값이 10이상이면 다중공선성이 있다고 말한다.

X를 절편항과 k개의 독립변수들로 이루어진 행렬이라 하자. 다중공선성이 있으면 회귀계수를 추정할때 사용되는 (X'X)의 역행렬이 구하기 어려워 회귀계수의 추정이 어렵다. (X'X)의 고유값이 0이거나 0에 가까우면 다중공선성이 있다고 말한다. 위의 생각에 기초한 통계량이 condition index이다. condition index는 고유값이 0에 가까우면 커지는 성질을 지니고 있으며, 그 값이 5-10이면 약한 상관관계가 존재하고, 30-100이면 강한 상관관계가 있다고 말한다.다중공선성의 문제를 해결하는 방법으로는 다중공선성이 강한 한 변수를 제거하거나, 다른 변수들에 의해 설명되는 부분을 제거한 잔차와 원래변수의 평균을 합한것을 새로운 변수로 정의하여 원래변수 대신에 이용하거나, 능형 회귀를 이용하는 방법등이 가장 많이 이용된다. 능형 회귀 추정은 편의 추정 방법에 해당하며, (X'X) 행렬 대신에 X'X+kI 행렬이 (I:Identity Matrix)라는 편의된 행렬을 이용하여 추정한다. 위의 경우 k를 추정하는 방법에 대해 많은 연구가 이루어지고 있으며, 그 추정방법에 따라 k의 값이 달라질 수 있다.

2. 연구 방법

'한국인 질병 예방을 위한 연구'의 연구 대상 중에서 그 일부인 493명을 대상으로 하여, cholesterol과 유의한 상관관계를 갖는 식이 변수중에서 떠먹는 요구르트의 섭취횟수와 섭취량의 상관계수가 0.75로서 가장 강하였다.또한, BMI를 quintile로 나눈 변수인 BMI1과 BMI2, BMI2와 BMI3, BMI3와 BMI4간의 상관계수도 0.6보다 컸다. 위의 여섯개의 변수와 연령을 독립 변수로 하고, cholesterol을 반응 변수로 하는 회귀모형을 적합해보았다.

3. 연구 결과

위의 회귀모형의 condition index는 1,3,5,8,10,15,18,35 이고,VIF(i)는 1--3의 값을 가지고 있다. 그러므로, 우리는 독립변수들간에 다중공선성이 존재한다는 판단을 내릴수 있다. 위의 다중공선성은 연령과 절편항간, BMI1, BMI2, BMI3, BMI4간, 요구르트 섭취 횟수와 섭취량간의 상관관계에 기인한 것이다. 요구르트 섭취횟수와 섭취량간의 상관관계를 제거한 모형을 연구 목적에서 제시한 세 가지 방법에 근거해서 적합시켜 보았다. BMI간의 다중공선성, 연령과 절편항간의 다중공선성도 아래의 모형과 같은 방법으로 제거할 수 있다.

	Model1	Model2	Model3	Model4
*요구르트 섭취횟수	0.09 (0.09)	0.17 (0.08)	0.15 (0.08)	0.09 ( ):표준오차
**요구르트 섭취량	3.34 (1.84)	4.05 (1.70)	4.03 (1.70)	3.33

위의 표의 회귀계수는 연령,BMI1,BMI2,BMI3,BMI4에 의해 보정된 회귀계수이다.

Model1: *,**변수가 동시에 들어간 모형
Model2: *,**변수가 각각 들어간 모형
Model3: *,**변수가 각각 보정된 경우의 모형
Model4: *,**변수가 동시에 들어간 능형회귀 모형

위의 표로부터 최소제곱법에 의해 추정한 Model1의 회귀계수와 다중공선성을 제거한 Model2,Model3,Model4의 그것은 다르며, 또한 후자의 경우가 표준오차도 다소 작음을 알수있다.

4. 고찰

독립 변수들간에 존재하는 다중공선성은 회귀계수의 추정을 어렵게 하고, 회귀계수의 분산을 크게 하여 그 정도를 떨어 뜨리므로,최소제곱법으로는 cholesterol에 미치는 각각의 독립변수의 영향을 올바로 추정해 낼 수가 없게 된다. 연구자의 분석 방법 뿐만 아니라 다중 경우에도 위에 열거한 내용에 기초한 확장이 가능하다.