Review Article
doi:http://dx.doi.org/10.5397/CiSE.2013.16.1.63

대한견·주관절학회지 제16권 제1호
Clinics in Shoulder and Elbow
Volume 16, Number 1, June, 2013

임상의를 위한 다변량 분석의 실제

서울대학교 의과대학 정형외과학교실, 분당서울대학교병원 관절센터,
건국대학교 의학전문대학원 정형외과학교실*, 건국대학교병원 어깨팔꿈치관절센터

오주한 · 정석원*

Multivariate Analysis for Clinicians

Joo Han Oh, M.D., Ph.D., Seok Won Chung*, M.D.

Department of Orthopaedic Surgery, Seoul National University College of Medicine,
Seoul National University Bundang Hospital, Korea,
Konkuk University School of Medicine, Konkuk University Medical Center, Korea*

In medical research, multivariate analysis, especially multiple regression analysis, is used to analyze the influence of multiple variables on the result. Multiple regression analysis should include variables in the model and the problem of multi-collinearity as there are many variables as well as the basic assumption of regression analysis. The multiple regression model is expressed as the coefficient of determination, R^2 and the influence of independent variables on result as a regression coefficient, β. Multiple regression analysis can be divided into multiple linear regression analysis, multiple logistic regression analysis, and Cox regression analysis according to the type of dependent variables (continuous variable, categorical variable (binary logit), and state variable, respectively), and the influence of variables on the result is evaluated by regression coefficient β, odds ratio, and hazard ratio, respectively. The knowledge of multivariate analysis enables clinicians to analyze the result accurately and to design the further research efficiently.

Key Words: Multivariate analysis, Multiple regression analysis, Multiple linear regression analysis, Multiple logistic regression analysis, Cox regression analysis

서 론

임상의로서 연구를 수행하는 과정은 대체로 임상 경험을 바탕으로 아이디어를 얻어 연구의 계획을 세우고 이를 수행한 뒤 결과를 통계 분석하여 논문화하는 일련의 과정이라고 볼 수 있다. 이 과정에서 통계 분석의 과

※통신저자: 정 석 원
서울시 광진구 능동로 120-1
건국대학교 의과대학원 정형외과학교실, 건국대학교병원 어깨팔꿈치센터
Tel: 02) 2030-7604, Fax: 02) 2030-7748, E-mail: smilecsw@gmail.com
접수일: 2013년 5월 15일, 1차 심사완료일: 2013년 6월 17일, 게재 확정일: 2013일 6월 18일

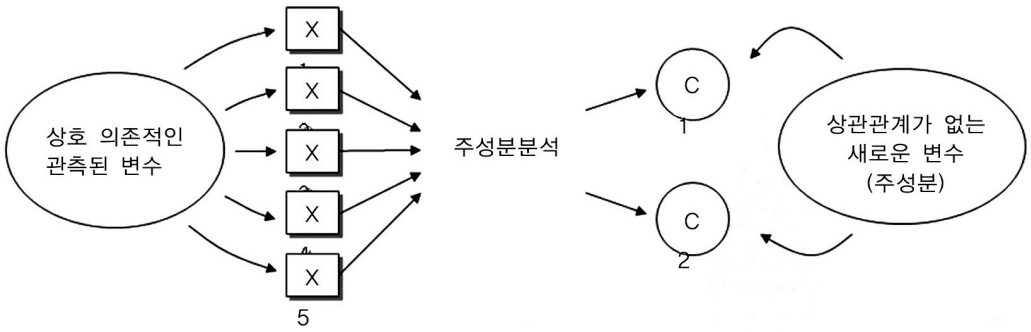

Fig. 1. Principle component analysis.

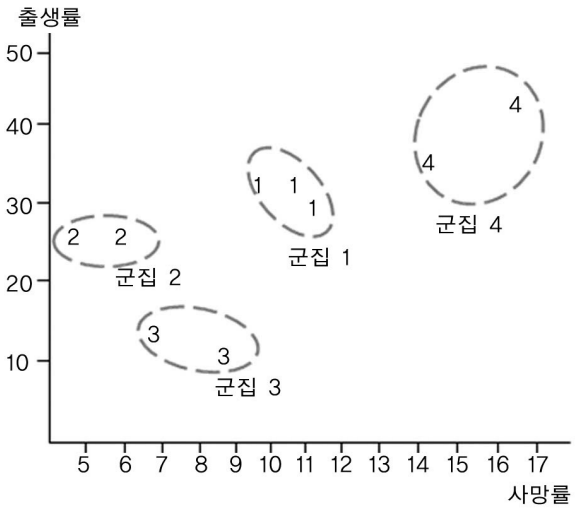

Fig. 2. Cluster analysis.

정은 수집된 자료로부터 합리적인 결론에 도달하는 의사 결정 과정이라고 할 수 있는데, 적절한 통계 방법의 선택과 이의 사용 및 해석 능력은 연구 결과의 정확한 분석을 가능하게 해 줄 뿐만 아니라, 연구의 계획, 대상 선정(표본 추출), 연구 수행, 데이터 입력, 데이터 분석, 결과 작성 및 해석에 이르는 연구의 전 단계에서 불필요한 시행 착오 및 오류, 데이터 손실을 줄여주어 보다 효율적이고 가치 있는 연구가 되도록 할 수 있다. 따라서, 연구를 수행하고자 하는 임상의로서 의학 통계에 대한 기본 소양은 필수적이라고 할 수 있다. 대부분의 의학 연구에서 결과를 평가하기 위해서는 다양한 위험 인자들, 혹은 많은 변수들이 분석되게 된다. 이 과정에서 분석의 최종 목표는 변수들간의 관계를 평가하기보다는 결과를 예측하거나 설명하기 위한 경우가 대부분이고, 이 때 변수들간의 교란 효과(confounding effect)에 의한 삐뚤림(bias)이 반드시 고려되어야 한다. 이러한 여러 변수들의 교란 효과를 통계적으로 적절히 처리하며 두 개 이상 여러 개의 변수를 동시에 분석하는 방법이 다변량 분석(multivariate analysis)이다.

다변량 분석 (multivariate analysis)

다변량 분석은 변수 갯수와 개체 갯수가 많은 복잡한 자료에 쓰이는 분석 방법으로, 변수들 간의 상관 관계를 이용하여 변수를 축약하거나 개체들을 분류하는데 관련된 분석 방법과, 변수들(독립 변수와 종속 변수) 간의 인과 관계를 규명하는 데 관련된 분석 방법이 있다. 전자에 해당하는 다변량 분석의 종류에는 대표적으로 주성분 분석(principle component analysis), 군집 분석(cluster analysis) 및 판별 분석(discriminant analysis)이 있는데, 이 중 주성분 분석은 측정 변수들 사이의 복잡한 상호 의존 관계를 쉽게 설명할 목적으로, 상호 의존적인 여러 측정 변수들을 서로 독립인 몇 개의 새로운 변수(주성분)로 간단하게 만드는 것이다. 이 방법은 독립 변수와 종속 변수의 구분이 없이 여러 변수를 한꺼번에 고려하여 변수들 사이의 상호 의존적인 구조를 파악하는 방법이다(Fig. 1). 또한, 군집 분석은 연구 대상이 가지고 있는 다양한 특성을 고려하여 비슷한 특성을 가진 그룹으로 묶는 통계적 분석 방법이고(Fig. 2), 판별 분석은 두 그룹 이상으로 나누어진 상황에서 연구 대상이 어떠한 그룹에 속할 것인지를 판별식을 이용하여 판단할 수 있게 만드는 통계적 기법이다(Fig. 3). 그러나, 이러한 통계적 방법들은 대체로 사회 과학, 심리학 및 경영학 등에서 사용되는 방법들로 의학적 연구의 분석 방법과는 거리가 있다. 후자에 해당하는 방법, 즉 변수들간의 인과 관계를 규명하여 결과를 예측하거나 설명하기 위한 다변량 분석 방법이 임상의가 수행하게 되는 연구에 보다 적합하다 할 수 있다. 이러한 분석 방법에는 다중 회귀 분석(multiple regression analysis)과 다변량 분산 분석(multivariate analysis of variance)이 있는데, 다변량 분산 분석은 결과 해석이 용이하지 않아 많이 쓰이지 않는 분석 방법이다. 따라서, 여기서는 임상

의학 연구의 대표적 다변량 분석 방법인 다중 회귀 분석에 대해, 임상의가 직접 분석을 수행하는 과정의 실제적 측면을 고려하여 살펴보기로 한다.

다중 회귀 분석 (multiple regression analysis)

다중 회귀 분석이란 다변량이 미리 독립 변수(설명 변수) 여러 개와 종속 변수(결과 변수) 1개로 나뉘어져 있어서 전자에 의해 얻어진 정보에서 후자를 추정하려고 하는 분석법으로, 여러 개의 변수에 의한 "한꺼번에의 영향력"을 분석하기 위한 방법이라고 할 수 있다. 이를 위해서는 먼저 회귀 분석에 대한 이해가 필요하다 하겠다.

1. 회귀 분석(regression analysis)

회귀 분석이란 일반적으로 변수들간의 함수적 관련성(인과 관계)을 규명하기 위하여 측정된 변수들의 자료로부터 모형을 추정하고 분석하는 방법인데, 변수들 간의 인과 관계를 파악하기 위한 대표적인 기법이라 할 수 있다. 여기서 인과 관계가 성립되기 위해서는 다음의 세 가지 조건이 충족되어야 할 것이다. 첫째, 원인은 결과보다 시간적으로 앞서야 하고, 둘째, 원인과 결과는 공동으로 변화해야 하며, 셋째, 결과는 원인 변수에 의해서만 설명되어져야 하고, 다른 변수에 의한 설명 가능성은 배제되어야 한다. 즉, 다른 변수의 영향이 모두 제거되어도 추정된 원인과 결과의 관계가 유지되어야 한다. 회귀 분석은 인과 관계의 분석에서 이 세 가지 조건 중 특히 세 번째 조건을 충족시키기 위하여 유용한 방법이라고 할 수 있는데, 이러한 특성 때문에 인과 관계를 통해 결과를 해석하려는 임상 연구에서 널리 사용되고 있다. 즉, 변수들간의 관련성을 본다는 측면에서는 상관분석의 성격을 가지고 있지만, 변수들간의 인과 관계를 알 수 있고 이를 통해 한 변수로부터 다른 변수의 변화를 예측할 수 있다는 점이 회귀 분석의 특징이라고 하겠다. 예를 들어, 혈중 콜레스테롤과 체질량 지수(BMI)에 대한 연구에서 혈중 콜레스테롤과 체질량 지수의 상호 관련성만을 조사한다면 상관 분석을 시행하면 되지만(이 경우 상관 계수값 r과 유의성 척도인 p값을 제시하게 된다), 혈중 콜레스테롤이 체질량 지수에 미치는 영향력(설명력)을 알고 싶다면 회귀 분석을 시행하여야 할 것이다(회귀 계수값 β와 유의성 척도인 p값 제시).

2. 다중 회귀 분석(multiple regression analysis)

회귀 분석은 독립 변수의 갯수에 따라 단순 선형 회귀 분석과 다중 회귀 분석으로 나눌 수 있다. 즉 독립 변수가 하나인 경우를 단순 선형 회귀 분석이라고 하며, $Y=\alpha+\beta X+\varepsilon$ (X=독립 변수, Y=종속 변수, α=절편, β=회귀 계수, ε=오차)인 모형으로 표현할 수 있다. 회귀 분석은 이러한 모형에서 절편인 α와 기울기인 β값을 주어진 자료로부터 추정하는 것을 의미한다. 그러나, 실제로 임상 연구에서 단일 요인에 의해 결과가 결정되는 경우는 매우 드물다. 대부분의 인과 관계에서 어떤 결과를 야기하는 원인들은 복수인 경우가 대부분이며, 이 원인들끼리도 서로 얽혀있기 마련이다. 따라서, 다수의 독립 변수를 모형에 포함시키는 회귀 분석이 불가피하고, 이를 다중 회귀 분석이라 한다. 즉, 다중 회귀 분석이란 종속 변수의 변화를 설명하고 예측하기 위하여 두 개 이상의 독립 변수가 사용되는 회귀 모형을 말한다. 다중 회귀 분석의 모형식은 $Y=\alpha+\beta_1X_1+\beta_2X_2+\beta_3X_3+\cdots+\varepsilon$ (X=독립 변수, Y=종속 변수, α=절편, β=회귀 계수, ε=오차) 형태로 표현될 수 있다. 이러한 다중 회귀 분석은 여러 가지 장점을 가지고 있는데, 특히 다른 독립 변수의 값을 통제한

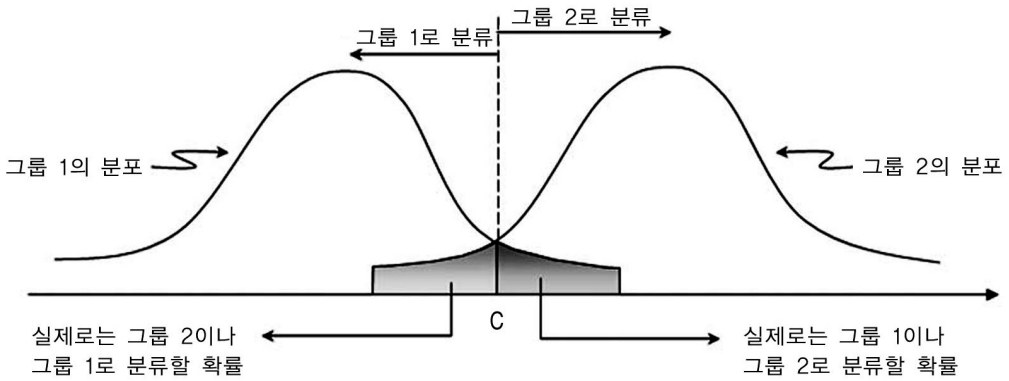

Fig. 3. Discriminant analysis.

상태에서 특정 독립 변수가 종속 변수에 독립적으로 행사하는 영향력을 측정할 수 있고, 이를 통해 각 독립 변수가 종속 변수에 미치는 효과의 상대적인 비교와 보다 정밀한 인과 관계의 분석이 가능하다는 장점이 있다. 다중 회귀 분석은 종속 변수의 형태에 따라 다중 선형 회귀 분석(multiple linear regression analysis), 다중 로지스틱 회귀 분석(multiple logistic regression analysis), 콕스 회귀 분석(Cox regression analysis)으로 나눌 수 있다. 이에 대해서는 뒤에 다시 다루기로 한다.

다중 회귀 분석에서의 고려점

1. 회귀 분석의 기본 가정

다중 회귀 분석을 위해서 기본적으로 고려할 사항들이 있는데, 우선 회귀 분석의 기본 가정을 만족해야 한다. 즉, 독립 변수와 종속 변수 사이에는 선형의 관계가 있고, 오차항(ε)의 평균이 0이고 정규 분포 및 동분산성을 만족하며, 오차항 사이에는 상관 관계가 존재하지 않아야 한다. 여기서, 오차항이란 종속 변수에 영향을 미치지만 회귀 모형에 포함되지 않은 모든 독립 변수들이 종속 변수에 미치는 영향의 합이라고 할 수 있는데, 회귀 분석의 기본 가정에서 보이듯 이러한 오차항이 회귀 모형을 결정하고 분석하는 데 영향을 미침을 알 수 있다. 오차항에 대한 기본 가정의 검정은, 오차항들의 히스토그램 및 정규 확률도를 그려서 정규 분포 여부를 확인하고, 산포도를 그려서 독립 변수의 값들과 무관하게 오차 정도가 균일함(0을 중심으로 직사각형의 분포)을 확인하며, Durbin-Watson의 d통계치를 통해 오차항 간의 자기 상관성을 확인할 수 있다. 중요한 독립 변수가 회귀식에서 빠진 경우나(이 경우 오차와 독립 변수 간의 상관 계수가 높아지게 된다.) 종속 변수가 하나 혹은 그 이상의 독립 변수의 원인이 되는 경우, 혹은 독립 변수의 측정이 오류인 경우 회귀 분석의 기본 가정이 충족되지 못하게 된다.

단순 회귀 분석과 달리 다중 회귀 분석의 경우에는 둘 이상의 독립 변수들이 모형에 포함되므로 단순 회귀 분석에서는 발생하지 않는 중요한 문제가 발생할 수 있는데, 이는 독립 변수들의 투입 순서 결정 및 다중 공선성(multi-colinearity)에 대한 문제이다.

2. 독립 변수들의 투입 순서 결정

다중 회귀 분석에서 회귀 계수들을 추정할 때 독립 변수들을 모형에 포함시키는 방법은 전진 선택법(forward slection), 후진 제거법(backward elimination), 단계별 방법(stepwise method) 및 입력(enter) 방법이 있다. 먼저 전진 선택법은 독립 변수 후보들 중에서 종속 변수에 가장 큰 영향을 주는 변수부터 통계적 기준에 따라 선택하여 모형에 포함시키며, 더 이상 추가할 의미 있는 변수가 없을 때 변수 선택이 중단되는 방법이다. 전진 선택법에서는 일단 선택된 변수는 다른 변수에 의해 중요성이 상실되더라도 회귀 모형에서 빠져 나올 수 없다. 후진 제거법은 독립 변수 후보 모두를 포함시킨 모형에서 출발하여 통계적 기준에 따라 가장 적은 영향을 주는 변수부터 하나씩 제거하면서 더 이상 제거할 변수가 없을 때의 모형을 선택하는 방법이다. 또한, 단계별 방법은 전진 선택법과 후진 선택법을 개선한 방법으로써 독립 변수의 추가와 제거를 적절히 조합하여 변수를 선택하는 방법이다. 즉, 전진 선택법에 의하여 변수를 추가하면서 새롭게 추가된 변수에 기인하여 기존 변수가 그 중요도가 약화되면 해당 변수를 제거하는 등, 매 단계별로 추가 또는 제거되는 변수의 여부를 검토하여 최종 포함되는 변수를 결정하는 방법으로 가장 많이 사용되는 방법이다. 마지막으로 입력 방법은 연구자의 판단에 따라 선택한 독립 변수들을 강제로 모형에 투입하여 회귀 계수들을 추정하는 방법이다. 이들 중 어떤 방법을 사용할 지는 변수들의 특성 및 중요도, 결과 해석의 방향 및 연구자의 선호도 등에 따라 결정될 수 있을 것이다.

3. 다중 공선성(multi-colinearity)

다중 공선성이란 다중 회귀 분석에서 독립 변수들 사이에 높은 선형 상관 관계(일반적으로 $r>0.8$ 혹은 0.9)가 존재하는 현상을 일컫는다. 이러한 경우, 높은 상관 관계를 보이는 하나의 변수가 투입되었을 때 나머지 변수들이 갖는 고유한 설명력은 매우 작아져서, 회귀 계수는 독립 변수가 종속 변수에 미치는 독자적인 영향력을 제대로 반영하지 못하게 된다. 다중 공선성을 알아보기 위한 가장 간단한 방법은 독립 변수들간의 상관 관계를 조사하는 것이지만, 다중 공선성을 보다 엄격하게 점검하려면 공차 한계(tolerance)와 분산 팽창 계수(variance inflation factor, VIF)를 살펴봄으로써 알 수 있다. 이 두 가지 지표들은 한 독립 변수가 다른 모든 독립

변수들에 의해서 설명되는 정도를 알려 준다. 먼저 공차 한계의 수식은 1-R^2으로 표현되는데(R^2은 다중 회귀 분석의 결정 계수를 의미하며 이에 대해서는 뒤에 설명하기로 한다), 일반적으로 그 값이 0.4이하인 경우 다중 공선성을 의심할 수 있고 0.1이하인 경우 심각한 다중 공선성 상태를 의미한다. 분산 팽창 계수는 1/공차 한계, 즉 1/(1-R2) 이며, 2.5보다 크면 다중 공선성을 의심해 볼 수 있고 10보다 큰 경우 심각한 다중 공선성 상태를 의미한다. 다중 공선성의 발생을 방지하기 위하여 미리 변수들간의 상관 계수를 파악하여 상관 관계가 높은 변수들 중 하나 혹은 일부를 회귀 분석 모형에서 제거하거나, 단계적 회귀 분석 방법을 이용하여 상관 관계가 높은 변수들 중 가장 설명력이 있는 독립 변수만을 모형에 포함시킬 수 있다. 또는, 상관성이 높은 변수들을 서로 합하거나 평균값을 구하여 새로운 변수를 만들거나, 좀 더 많은 데이터를 수집하여 재분석해 볼 수 있을 것이다. 이러한 방법들이 모두 용이하지 않은 경우에는 능형 회귀(ridge regression), 주 성분 회귀(principle component regression), 잠복근 회귀(latent root regression) 등을 사용하여 회귀 모형을 적합할 수도 있다.

4. 기타 고려 사항

기타 고려 사항으로는 데이터의 극단치(outlier)를 꼭 확인해야 한다. 회귀 분석에서 극단치는 절편과 회귀 계수에 영향을 주기 때문에 회귀 모형의 예측력을 떨어뜨릴 수 있다. 또한, 회귀 모형에서 독립 변수가 서열 척도나 명목 척도로 측정된 질적 변수(qualitative variable)인 경우 이를 더미 변수(dummy variable)화하여 분석하여야 한다. 독립 변수 한 단위의 변화가 종속 변수에 미치는 영향을 설명할 때 독립 변수가 질적 변수인 경우 회귀 분석을 통한 양적 변화를 보여줄 수 없기 때문이다.

다중 회귀 분석의 해석

다중 회귀 분석을 통한 결과를 해석할 때 그 다중 회귀 모형의 설명력이 어느 정도나 되고 독립 변수들이 종속 변수에 미치는 영향력이 어느 정도인지를 평가하게 된다.

1. 다중 회귀 모형의 설명력

총 분산 중에서 회귀식으로 설명되는 분산의 비율을 의미하며 결정 계수 $R^2(0 \leq R^2 \leq 1)$으로 표현한다. R^2값이 1에 가까울수록 독립 변수의 설명력이 크고 추정된 회귀식의 적합도가 높은 것으로 평가되고, 반대로 0에 가까워질수록 설명력이 약화되고 적합도도 떨어지게 된다. 일반적으로 회귀식에 포함되는 독립 변수의 개수가 늘어나면 결정 계수 R^2값이 높아지는 단점이 있다. 이러한 단점을 보완하기 위해 수정된 결정 계수(adjusted R^2)값을 이용할 수 있다. 주의할 것은, 이러한 모형의 설명력이 높다고 모형의 유의성이 높은 것은 아니다. 모형의 유의성 검정은 F-test로 하게 된다.

2. 종속 변수에의 영향력

다중 회귀 분석에서 독립 변수가 종속 변수에 얼마나 영향력이 있는지(얼마나 결과에 미치는 영향이 큰 요인인지)를 보는 값이 회귀 계수인 β값이다. 회수 계수 β값은 독립 변수들간의 관련성이 낮은 경우 더 중요도가 높다고 할 수 있다. 그러나, 이러한 회귀 계수는 독립 변수의 측정 단위에 크게 영향을 받기 때문에 독립 변수들간의 회귀 계수를 직접 비교하는 것은 위험하다. 다른 독립 변수 값을 고정시킨 상태에서, 보고자 하는 독립 변수의 값이 1단위 증가할 때 결과적으로 초래되는 종속 변수 값의 변화량인 비 표준화 부분 회귀 계수(B)값을 통해 다른 독립 변수들의 영향을 배제한 영향력을 평가할 수 있다.

SPSS 프로그램을 이용한 다중 회귀 분석의 사례 고찰

앞서 설명한 다중 회귀 분석이 실제 통계 프로그램을 통해 어떻게 이용될 수 있는지 실제 사례를 통해 살펴보기로 한다. 저자들은 회전근 개 파열의 관절경적 복원술 후 삶의 질에 미치는 인자를 조사하고자 하였다.[5] 이 때 수술 후 삶의 질이 결과에 해당하는 종속 변수이고 이에 영향을 미치는 인자들이 독립 변수들이 될 것이다. 임상 경험 및 문헌 고찰을 통해 회전근 개 복원술의 기능적 결과와 관련이 있을 수 있는 가능한 모든 인자들, 즉 나이, 성별, 증상 기간, 우세수 여부, 외상 여부, 당뇨, 혈압, 스포츠 활동 정도, 일의 강도, 술 전 견관절 강직, 술 전 통증 정도, 파열의 크기, 각 회전근 개 근육의 지방변성 정도, 수술 방법, 재파열 여부에 대한 데이터를 수집하였고, 술 후 최소 1년 후 대표적 삶의 질 평가 방법인 SF-36 평가 설문을 통해 얻은 데이터와 함께 분석을 수행하였다. 먼저 각각의 인자와 결과 사이에서 t-

test, ANOVA test 및 chi-square test 등 단변량 분석(univariate analysis)을 시행하여 유의한 결과가 나온 인자들을 선별하였다. 여기서는 나이, 성별, 당뇨, 스포츠 활동 정도, 술 전 통증 정도, 극하근과 견갑하근의 지방 변성, 그리고 재파열 여부가 삶의 질에 유의하게 영향을 미치는 인자였다. 다음 단계로, 이들 인자들 중 어떤 인자가 독립적으로 수술 후 삶의 질에 영향을 주는 인자이고 또 그들의 영향력은 어느 정도인지를 평가하였고, 이 때 사용한 통계적 방법이 다중 회귀 분석이다. 통계 프로그램으로 SPSS 15.0을 사용하였다. SPSS 프로그램 상단의 "분석"을 클릭하고 이어 "회귀 분석", 그리고 "선형"을 클릭하면(Fig. 4), "선형 회귀 분석"이라는 작업창이 생성된다(Fig. 5). 이 예제의 경우 종속 변수인 SF-36 점수가 연속 변수이기 때문에 "선형"을 선택하였다. 만약, 종속 변수가 '좋거나 나쁨'과 같이 이분형 변수라면 "선형" 대신에 "이분형 로지스틱"이라는 항목을 클릭하면 이후 동일한 과정을 거쳐 분석 결과를 얻을 수 있다. 이러한 과정을 거쳐 생성된 회귀 분석 작업창의 왼쪽 박스에는 SPSS 프로그램에 입력된 모든 변수 항목들이 보여지게 된다(Fig. 5). 이 항목들 중 SF-36 점수 항목을 "종속 변수" 박스로 옮기고, 단변량 분석을 통해 선별된 인자들을 "독립 변수" 박스에 옮겨 넣는다. "방법" 항은 독립 변수들의 모형에의 투입 방법을 의미하며 여기서는 가장 흔히 사용되는 단계 선택 방법을 선택하였다. 이후 하단 좌측의 "통계량"을 클릭하면 다시 새로운 작업창이 생성되고(Fig. 6), 여기서 좌측의 회귀계수 "추정값"과 "신뢰 구간" 그리고 우측의 "모형 적합"과 "공선성 진단"을 클릭하여 다중 회귀 분석을 시행하였다. 마지막으로 "계속"과 이어 "확인"을 클릭하면 결과값이 생성되게 된다. 먼저 모형에 진입/제거된 변수에 대한 정보를 확인할 수 있다(Fig. 7). 이 예제에서는 세 가지 모형이 생성되었으며(Fig. 8), 진입된 변수는 모형 1의 경우 DM 여부, 모형 2의 경우 DM 여부 및 성별, 그리고 모형 3의 경우는 DM 여부, 성별 및 나이 변수가 진입되었다.(나이 변수는 분석의 편의를 위해 〈55세, 55~60세, 60~65세, 〉65세의 4단계로 분류되었다.) 이 중 가장 많은 변수를 포함하고 있는 모형 3을 선택하였고, 이 때 수정된 R^2값은 0.109로 이 모형의 설명력은 10.9%였다. 모형의 유의성은 F-test를 통해 확인할 수 있으며, 모형 3의 경우 F=8.678, 유의 확률은 〈0.001로 모형이 유의미하다고 판단할 수 있다(Fig. 9). 앞서 통계량 선택에서 "공선성 진단"을 클릭하였고, 이 결과는 공차 한계와 분산 팽창 계수(VIF) 값으로 표현된다(Fig. 10). 모형 3의 경우, 모든 변수의 공차 한계가 0.4보다

Fig. 4. Example of multiple regression analysis using SPSS. Multiple linear regression analysis.

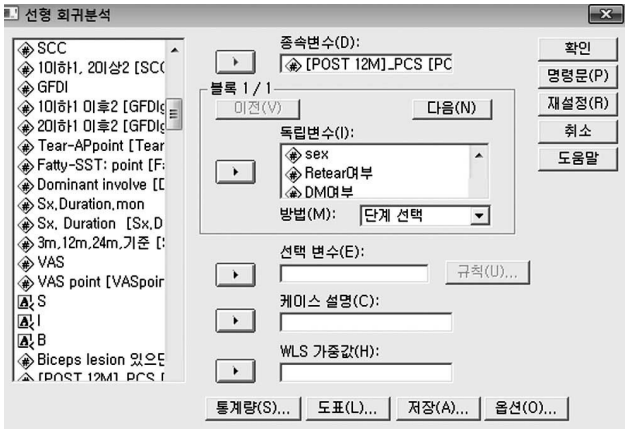

Fig. 5. Example of multiple regression analysis using SPSS. Dependent variable, independent variable, and the choice of variable selection method.

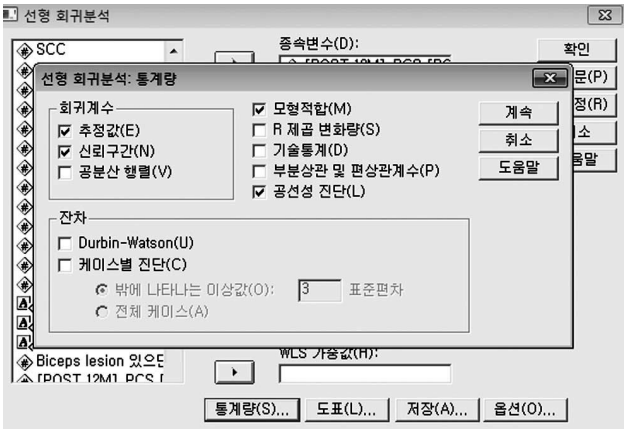

Fig. 6. Example of multiple regression analysis using SPSS. Regression coefficient, fitness of model, and diagnosis of multicolinearity.

크고 VIF 값도 2.5보다 작으므로 다중 공선성의 문제는 존재하지 않는다고 평가할 수 있다. 모형에 진입한 변수들이 종속 변수에 미치는 영향력은 비 표준화 계수 값(B)를 통해 확인할 수 있는데, 모형 3의 경우 B값은 당뇨, 성별, 나이 각각에서 −5.048, −2.865, −1.151이었고, 모든 변수에서 유의 확률이 0.05이하였다(Fig. 11). 유의 확률 결과로부터 모든 변수가 유의미하다는 것을 알 수 있고, 다른 변수의 영향을 배제한 상태에서, 당뇨가 있으면 수술 후 삶의 질이 5.048배 감소하고, 여성이 남성에 비해 수술 후 삶의 질이 2.865배 감소하며, 나이가 한 단계 증가할수록 삶의 질이 1.151배 감소했다고 분석할 수 있다.

진입/제거된 변수[a]

모형	진입된 변수	제거된 변수	방법
1	DM여부	.	단계선택 (기준: 입력할 F의 확률 <=, 050, 제거할 F의 확률 >=, 100).
2	sex	.	단계선택 (기준: 입력할 F의 확률 <=, 050, 제거할 F의 확률 >=, 100).
3	ageclass4 단계	.	단계선택 (기준: 입력할 F의 확률 <=, 050, 제거할 F의 확률 >=, 100).

a. 종속변수: [POST 12M]_PCS

Fig. 7. Example of multiple regression analysis using SPSS. Entry and removal of variables.

다중 회귀 분석의 종류에 따른 결과의 제시 및 해석

다중 회귀 분석은 종속 변수의 형태에 따라 다중 선형 회귀 분석, 다중로지스틱 회귀 분석 및 콕스 회귀 분석으로 나눌 수 있고, 이에 따라 분석법과 결과의 해석이 다르다. 종속 변수가 연속 변수(점수화 변수, 실측값)인 경우는 다중 선형 회귀 분석, 종속 변수가 범주형 변수(1, 0)인 경우는 다중 로지스틱 회귀 분석, 그리고 종속 변수가 시간의 영향을 고려한 상태 변수인 경우에는 콕스 회귀 분석을 시행해야 한다. 다중 선형 회귀 분석 결과의 해석은 앞서 설명한 바와 같이 다른 독립 변수를

모형 요약

모형	R	R 제곱	수정된 R 제곱	추정값의 표준오차
1	.240[a]	.058	.052	7.9844
2	.312[b]	.097	.088	7.8345
3	.350[c]	.123	.109	7.7442

a. 예측값: (상수), DM여부
b. 예측값: (상수), DM여부, sex
c. 예측값: (상수), DM여부, sex, ageclass4단계

Fig. 8. Example of multiple regression analysis using SPSS. Selection of model and explanation power of model.

분산분석[d]

모형		제곱합	자유도	평균제곱	F	유의확률
1	선형회귀분석	731.206	1	731.206	11.470	.001[a]
	잔차	11984.972	188	63.750		
	합계	12716.178	189			
2	선형회귀분석	1238.349	2	619.175	10.088	.000[b]
	잔차	11477.829	187	61.379		
	합계	12716.178	189			
3	선형회귀분석	1561.349	3	520.450	8.678	.000[c]
	잔차	11154.829	186	59.972		
	합계	12716.178	189			

a. 예측값: (상수), DM여부
b. 예측값: (상수), DM여부, sex
c. 예측값: (상수), DM여부, sex, ageclass4단계
d. 종속변수: [POST 12M]_PCS

Fig. 9. Example of multiple regression analysis using SPSS. Significance of model.

보정한 상태에서 어떤 요인이 1단위 늘거나 주는데 따라 얼마만큼 종속 변수 값이 증가하거나 감소하는지로 해석하며 이 때 변화 정도는 회귀 계수인 β값을 사용한다. 또한, 다중 로지스틱 회귀 분석은 다른 독립 변수를 보정한 상태에서 어떤 요인이 1단계 늘거나 주는데 따라 종속 변수량이 몇 배 증가하거나 감소하는지로 해석하고, 변화 정도는 교차비(odds ratio)로 표현한다. 콕스 회귀 분석의 경우는 시간의 개념이 들어가며 다른 독립 변수를 보정한 상태에서 어떤 요인이 1단계 늘거나 주는데 어떤 사건의 발생율이 몇 배 증가하거나 감소하는지로 해석하고, 변화 정도는 위험비(hazard ratio)로 표현한다. 각각의 다중 회귀 분석 방법에 대한 결과의 제시와 그 해석에 대해 예를 들어 살펴 보기로 한다. 회귀 분석에서 결과의 제시는 모형에 포함된 각 독립 변수들이 종속 변수에 미치는 영향력을 표현하는 회귀 계수 β(다중 선형 회귀 분석), 교차비(다중 로지스틱 회귀 분석), 혹은 위험비(콕스 회귀 분석)와 함께 95% 신뢰 구간(95% CI) 및 유의 확률(p)을 제시해야 한다. Table 1은 소아의 신장, 체중, 성별이 수축기 혈압(연속 변수)에 미치는 영향에 대한 다중 선형 회귀 분석 결과표로, 이로부터 각 변수 서로의 영향을 보정한 상태에서 체중이 1 kg 증가할수록 수축기 혈압이 1.18단위 증가하고, 남자일 경우 수축기 혈압이 4.23단위 증가하였으며, 신장은 혈압과 무관하다고 해석할 수 있다. Table 2는 회전근 개 복원술 후 유합 실패(이분형 변수)에 영향을 미치는 인자에 대한 다중 로지스틱 회귀 분석 결과표로, 모형에 진입한 변수와 제거된 변수를 모두 제시하였다.[4] 이 경우, 각 변수 서로의 영향을 보정한 상태에서 회전근 개 퇴축이 1 cm 증가할수록 유합 실패 확률이 1.98배 증가하고, 극하근의 지방 변성이 심할수록(Goutallier 등급 0 혹은 1에 비해 등급 3 혹은 4) 유합 실패 확률이 8.13배 증가하였으며, 골질이 좋을수록(T score 1단위 증가) 유합 실패

계수[a]

모형		비표준화 계수		표준화 계수	t	유의확률	B에 대한 95% 신뢰구간		공선성 통계량	
		B	표준오차	베타			하한값	상한값	공차한계	VIF
1	(상수)	47.565	.629		75.590	.000	46.324	48.807		
	DM여부	-5.455	1.611	-.240	-3.387	.001	-8.632	-2.278	1.000	1.000
2	(상수)	52.889	1.952		27.089	.000	49.038	56.741		
	DM여부	-5.373	1.581	-.236	-3.399	.001	-8.492	-2.255	1.000	1.000
	sex	-3.335	1.160	-.200	-2.874	.005	-5.625	-1.046	1.000	1.000
3	(상수)	54.977	2.129		25.819	.000	50.776	59.178		
	DM여부	-5.048	1.569	-.222	-3.218	.002	-8.143	-1.953	.992	1.008
	sex	-2.865	1.165	-.172	-2.460	.015	-5.163	-.567	.969	1.032
	ageclass4단계	-1.151	.496	-.163	-2.321	.021	-2.130	-.173	.962	1.040

a. 종속변수: [POST 12M]_PCS

Fig. 10. Example of multiple regression analysis using SPSS. Diagnosis of multicolinearlity.

계수[a]

모형		비표준화 계수		표준화 계수	t	유의확률	B에 대한 95% 신뢰구간		공선성 통계량	
		B	표준오차	베타			하한값	상한값	공차한계	VIF
1	(상수)	47.565	.629		75.590	.000	46.324	48.807		
	DM여부	-5.455	1.611	-.240	-3.387	.001	-8.632	-2.278	1.000	1.000
2	(상수)	52.889	1.952		27.089	.000	49.038	56.741		
	DM여부	-5.373	1.581	-.236	-3.399	.001	-8.492	-2.255	1.000	1.000
	sex	-3.335	1.160	-.200	-2.874	.005	-5.625	-1.046	1.000	1.000
3	(상수)	54.977	2.129		25.819	.000	50.776	59.178		
	DM여부	-5.048	1.569	-.222	-3.218	.002	-8.143	-1.953	.992	1.008
	sex	-2.865	1.165	-.172	-2.460	.015	-5.163	-.567	.969	1.032
	ageclass4단계	-1.151	.496	-.163	-2.321	.021	-2.130	-.173	.962	1.040

a. 종속변수: [POST 12M]_PCS

Fig. 11. Example of multiple regression analysis using SPSS. Analysis of influence on dependent variable.

Table 1. The Effect of Height, Weight, and Gender on Systolic Blood Pressure in Childhood (The Result of Multiple Linear Regression Analysis)

변수	β	SE	95% CI	p value
intercept	79.439	17.118	(45.89, 112.99)	<0.001
height	-0.031	0.171	(-0.37, 0.31)	0.857
weight	1.179	0.261	(0.67, 1.69)	<0.001
gender	4.229	1.610	(1.07, 7.39)	0.010

β: estimated regression coefficient, SE: standard error, CI: confidence interval

확률이 0.53배로 감소했고, 기타 다른 요인은 유합 실패와 무관하였다고 해석할 수 있다. Table 3의 경우는 대장암으로 대장 절제술을 시행한 환자에서 국소 재발(사건의 발생율)에 영향을 미치는 인자에 대한 콕스 회귀 분석표이다. 이 경우, 각 변수 서로의 영향을 보정한 상태에서 수술만을 한 경우가 병행요법을 한 경우에 비해 3.4배 재발 확률이 높았고, 항문에서 거리가 가까울수록(10.1~15 cm에 비해 5.1~10 cm가 2.13배, ≤5 cm가 2.78배) 그리고 TNM stage가 높을수록(stage I에 비해 II가 3.44배, III가 9.69배, IV가 16.20배) 재발이 잘 일어나며, 수술법은 재발과 무관하였다고 해석할 수 있겠다.

결 론

대부분의 의학 연구에서 단일 요인에 의해 결과가 결정되는 경우는 매우 드물고 대부분 다양한 요인들에 의

Table 2. Factors Affecting Failure of Healing after Rotator Cuff Repair (The Result of Multiple Logistic Regression Analysis)

Variable	OR	95% CI	*p* value
Amount of retraction*	1.98	1.08- 3.65	0.027
FI of the infraspinatus* (reference)	1		
FI of the infraspinatus (1)	3.81	1.26-11.57	0.018
FI of the infraspinatus (2)	8.13	1.43-46.07	0.018
Bone mineral density (BMD)*	0.53	0.36- 0.78	0.001
Gender†			0.851
Age†			0.084
Diabetes†			0.551
FI of the supraspinatus†			0.164
FI of the subscapularis†			0.113
Tear size of AP dimension†			0.689
Acromiohumeral distance†			0.752
Biceps procedure†			0.524

* : statistically significant,

† : variables not significant in multivariate analysis (those exempted from the equation)

OR: odds ratio, CI: confidence interval, FI: fatty infiltration, AP: anteroposterior

FI of the infraspinatus (1) means FI grade 2 according to Goutallier' s classification, and FI of the infraspinatus (2) means FI grade 3 and 4. FI grade 0 and 1 was regarded as the standard reference for the comparison with the other grades.

Table 3. Factors Affecting Local Recurrence of Colon Cancer after Colectomy (The Result of Cox-regression Analysis)

Variable	HR (95% CI)	*p* value
Treatment group		<0.001
Radiotherapy and surgery	1.00	
Surgery alone	3.41 (2.05-5.70)	
Distance of tumor from anal verge		0.03
10.1~15 cm	1.00	
5.1~10 cm	2.13 (1.13-4.01)	0.02
≤5 cm	2.78 (1.22-6.31)	0.02
Type of resection		0.90
Low anterior	1.00	
Abdominoperineal	1.15 (0.59-2.24)	0.68
Hartmann†	1.16 (0.42-3.25)	0.78
TNM stage		<0.001
I	1.00	
II	3.44 (1.26-9.36)	0.02
III	9.69 (3.89-24.2)	<0.001
IV (distant metastasis)	16.20 (5.40-48.6)	<0.001

HR: hazard ratio, TNM stage: T describes the size or extent of the primary tumor, N describes the degree of spread to regional lymph nodes, and M describes the presence of distance metastasis.

해 어떤 결과가 야기되며, 이 원인들끼리고 서로 얽혀있게 마련이다. 따라서, 다양한 변수를 분석하고 변수들간의 인과 관계를 규명하여 결과를 예측하거나 설명하기 위한 다변량 분석, 특히 다중 회귀 분석에 대한 이해와 사용 및 이의 해석 능력은 의학 연구를 계획하고 분석하고자 하는 임상 의사에게 있어 필수적인 소양이라고 할 수 있다. 이러한 다변량 분석에 대한 이해를 통해 적절한 연구 계획의 수립 및 통계 방법의 선택과 정확한 연구 결과의 분석이 가능하고, 나아가 보다 효율적이고 가치 있는 연구가 되도록 할 수 있을 것이다.

REFERENCES

1) **Song KI, Choi JS.** *The analysis of clinical data using SPSS 15. 1st ed. Seoul: Hannarae academy; 2009. 78-257.*
2) **Ahn JE, Yoo GY.** *Statistical analysis of medical health science. 2nd ed. Seoul: Hannarae academy; 2010. 401-605.*
3) **Amit Choudhury.** *Multiple Regression Analysis Page [Inter-net]. Guwahati (India):University of Gauhati, Department of Statistics; 2009 Jun 18[updated 2013 Jul 02]. Available from: http://explorable.com/multiple-regression-analysis/.*
4) **Chung SW, Oh JH, Gong HS, Kim JY, Kim SH.** *Factors affecting rotator cuff healing after arthroscopic repair: osteoporosis as one of the independent risk factors. Am J Sports Med. 2011;39:2099-107.*
5) **Chung SW, Park JS, Kim SH, Shin SH, Oh JH.** *Quality of life after arthroscopic rotator cuff repair: evaluation using SF-36 and an analysis of affecting clinical factors. Am J Sports Med. 2012;40:631-9.*
6) **Johnson RA, Wichern DW.** *Applied multivariate statistical analysis. 5th ed. New Jersey: Prentice-Hall; 2002. 1-788.*
7) **Mitchell H.** *Multivariate Analysis. A practical guide for clinicians. 2nd ed. Cambridge: Cambridge University Press; 2006. 1-203.*

초 록

임상 의학의 연구에 사용되는 대표적 다변량 분석 방법은 다중 회귀 분석 방법인데, 이는 인과 관계를 토대로 여러 개의 변수에 의한 한꺼번에의 영향력을 분석하기 위한 방법이다. 다중 회귀 분석은 기본적으로 회귀 분석의 기본 가정을 만족해야 함은 물론, 여러 개의 독립 변수들이 포함되기 때문에 변수들을 모형에 포함시키는 방법 및 다중 공선성 문제에 대한 고려가 필요하다. 다중 회귀 분석 모형의 설명력은 결정 계수 R^2으로 표현되어 1에 가까울수록 설명력이 크며, 각 독립 변수들의 결과에의 영향력은 회귀 계수인 β값으로 표현된다. 다중 회귀 분석은 종속 변수의 형태에 따라 다중 선형 회귀 분석, 다중 로지스틱 회귀 분석, 콕스 회귀 분석으로 나눌 수 있다. 종속 변수가 연속 변수인 경우 다중 선형 회귀 분석, 범주형 변수인 경우 다중 로지스틱 회귀 분석, 시간의 영향을 고려한 상태 변수인 경우는 콕스 회귀 분석을 시행해야 하며, 각각 결과에의 영향력은 회귀 계수 β, 교차비, 위험비로 평가한다. 이러한 다변량 분석에 대한 이해는 연구를 계획하고 결과를 분석하고자 하는 임상 의사에게 있어 보다 효율적인 연구를 위해 필수적인 소양이라고 할 수 있다.

색인 단어: 다변량 분석, 다중 회귀 분석, 다중 선형 회귀 분석, 다중 로지스틱 회귀 분석, 콕스 회귀 분석