韓國 OA學會 論文誌
第5卷 第2號, 2000. 6.

2000-5-2-4-2

범주형 데이터의 인과관계분석에 관한 기초적 연구

노 형 진*

A Study on the Analysis of Causal Relation about Categorical Data

Hyung-Jin Rho

요 약

질적 데이터의 수량화를 통하여 통계분석이 가능한 수량화이론 중 인과관계분석을 위한 수량화이론 I류와 II류에 대한 기초개념과 알고리즘을 소개한다. 또한 이들 두 기법을 Excel에 의해 처리할 수 있는 방법론을 제시함으로써 그 활용성을 시사하고자 한다.

Abstract

The purpose of this study is to suggest usefulness of causal relation analysis about categorical data to business administration and other social sciences. Some examples of their application are presented.

* 경기대학교 경영학부 교수 경영학박사

Ⅰ. 서론

수량화이론(quantification theory)이라고 불리는 것은 L. Guttman의 예측이론으로부터 출발하여 일본의 하야시(林知己夫)에 의해서 전개된 독창적인 이론이다. 결과적으로는 더미변수법과 극히 유사한 것인데, 더미변수가 아직 보급되어 있지 않았던 1950년대 초에 이미 오늘날의 체계적인 형태를 갖추고 있었다는 점에서 높이 평가할 만하다. 더미변수의 경우와 마찬가지로 양적변수를 질적변수와 나란히 함께 쓸 수가 있다.

수량화이론은 단지 질적인 것에 1, 2, 3, 4 등 수치를 기계적으로 범주 내용의 단계순으로 부여하여 해석하는 것이 아니라, 질적인 것에 어떻게 수치를 부여하면 보다 최적의 데이터 분석이 가능한 것인가에 대한 사고방식에 의거하여 개발되었다.

수량화이론의 처음 발단은 1947년 형무행정(刑務行政)에 통계를 활용한 가석방의 연구이며 그 후 심리학자, 사회학자 등에 의해 계승·발전되어 왔다. 수량화이론에는 Ⅰ류, Ⅱ류, Ⅲ류, Ⅳ류 등 네 종류가 있는데 Ⅰ~Ⅳ류 등으로 명명한 것은 하야시 본인이 아니고 그 후의 이용자측에세 이름붙여 현재에 이르고 있다.

특히 수량화이론 Ⅲ류에 대해서는 하야시의 흐름과는 독립적으로 프랑스에서는 J. P. Benz'e cri 및 그의 제자들이 대응분석(correspondence analysis)이라고 하는 수법을 개발하여 다방면에 걸쳐 응용하고 있다는 점에서 주목할 가치가 있다.

또한 1950년대부터 미국에서 발달해 오고 있고 현재도 전세계에서 사용되고 있는 다차원척도법(multidimensional scaling : MDS)의 목적과 똑같은 수량화이론 Ⅳ류는 다차원척도법의 선구적인 수법이라고도 할 수 있다.

본고에서는 수량화이론 중 인과관계분석을 위한 Ⅰ류와 Ⅱ류에 대한 기초개념과 알고리즘을 소개하기로 한다. 또한 이 두 기법을 Excel에 의해 처리할 수 있는 방법론을 제시하고자 한다. 이들 두 기법은 사회과학 분야의 인과관계분석에 널리 활용될 것으로 기대된다.

Ⅱ. 수량화이론 Ⅰ류

수량화이론 Ⅰ류는 명목척도의 카테고리 변수로 되는 몇 개의 항목에 의해서 설명변수의 조가 구성되는 경우에, 외적기준(목적변수) 변수에 대한 합성변수의 예측오차를 최소로 하는 가중치를 각 카테고리에 부여하는 방법이다. 형식적으로는 설명변수가 명목척도로 되는 더미변수의 데이터를 이용한 중회귀분석이라고 해도 좋다. 그러나 각 항목에 포함되는 모든 카테고리에 주어진 득점의 총합은 항상 1이 되므로 각열의 성분에 평균을 빼는 조작을 가하면 각 카테고리에 대응하는 열 벡터가 1차종속이 되어 버린다. 따라서 실제의 계산시에는 각 항목에서 하나씩의 카테고리를 제거하고 분석하며 제거한 카테고리 변수에는 가중치 0을 부여하면 된다.

이와 같이 해서 얻어진 각 카테고리의 가중치의 최대값과 최소값의 차를 범위(range)라고 부르며, 각 항목이 외적기준 변수에 어느 정도 영향을 주고 있는지를 추정하는 하나의 기준을 부여한다. 그리고 설명변수가 연령, 수입 등의 간격척도인 경우에도 이들 변수가 예측의 대상으로 되어 있는 기준변수와 곡선적 상관을 갖는 경우에는 예를 들면 연령이라면,

① 10세 이하, ② 11~20세, ③ 21~30세,
④ 31~40세, ⑤ 41~50세, ⑥ 51세 이상

과 같은 여섯 개의 카테고리로 구분하고 각 카테고리에 외적기준과의 예측오차를 최소로 하는 최적의 가중치를 정할 수 있다.

〔사례 1〕

제약회사 K사는 신제품의 시판을 앞두고 TV 방송을 통한 광고를 위해서 13개의 TV 프로에 대한 시청률과 프로의 내용, 시간대를 조사했다. 다음의 데이터표(1)은 그 조사결과이다. 이 13개의 TV 프로는 월요일에서 토요일까지 방영되는 것이며, 여기에서 시청률을 외적기준(목적변수)으로 하고 프로의 내용과 시간대를 설명변수로 한다. 두 개의 설명변수는 다음과 같이 각각 네 개의 범주를 갖는다.

프로의 내용 : 1. 뉴스 2. 연속극 3. 만화 4. 가요·오락
시간대 : 1. 오후 7~8시, 2. 오후 8~9시, 3. 오후 9~10시, 4. 오후 10~11시

데이터 표(1)

No.	프로의 내용(x_1)	시간대(x_2)	시청률(y)
1	3	1	7.2
2	4	1	7.1
3	4	3	4.4
4	2	4	6.5
5	1	1	4.2
6	2	2	14.6
7	3	1	2.3
8	4	1	4.9
9	1	2	6.2
10	1	3	8.5
11	3	1	2.9
12	3	1	3.5
13	2	3	13.8

이와 같은 데이터가 얻어졌을 때에 프로의 내용(x_1)과 시간대(x_2)로부터 시청률을 예측하는 식을 만들고자 하는 것이 수량화이론 I류이다. 수량화이론의 세계에서는 프로의 내용(x_1)이나 시간대(x_2)를 아이템(item) 혹은 요인이라 부르고 있다. 그리고 선택지인 1, 2, 3, 4 등을 카테고리(category) 혹은 범주라고 부른다.

◆ (0, 1) 데이터

앞의 데이터표(1)을 다음의 데이터표(2)와 같이 바꿔 써 본다. 데이터표(2)에서는 각 아이템의 네 개의 카테고리 중 응답한 하나에만 1이 입력되고 나머지에는 0이 입력되어 있다.

이와 같이 (0, 1) 데이터를 사용해서 표현함으로써 질적 데이터를 수치화할 수 있다.

◆ (0, 1) 데이터의 중회귀분석

프로의 내용(x_1)과 시간대(x_2)로부터 시청률을 예측하는 식을 만들고자 하는 문제는 (0, 1) 데이터로 바꿔 쓴 데이터표(2)에 있어서 프로의 내용(x_1)과 시간대(x_2)에 대한 각각의 카테고리를 x_{11}, x_{12}, x_{13}, x_{14}, x_{21}, x_{22}, x_{23}, x_{24}로 바꿔 표현하면 x_{11}에서 x_{24}까지를 설명변수, 시청률 y를 목적변수로 하는 중회귀분석의 문제로 바꿔 놓을 수 있다.

이렇게 하면 데이터표(2)를 그대로 중회귀분석에 적용하면 될 듯 싶은데 실은 그렇게 잘 되지 않는다. 왜냐하면 설명변수 사이에는

$$x_{11} + x_{12} + x_{13} + x_{14} = 1$$
$$x_{21} + x_{22} + x_{23} + x_{24} = 1$$

이라고 하는 선형관계가 성립하고 있어, 바로 다중공선성 때문에 해가 구해지지 않는 것이다. 그래서 x_{11}, x_{12}, x_{13}, x_{14}중의 하나(여기에서는 x_{14})를 삭제하고, x_{21}, x_{22}, x_{23}, x_{24} 중의 하나(여기에서는 x_{24})를 삭제한다. 데이터표(2)를 이와 같이 해서 바꿔 쓰면 데이터표(3)이 얻어진다. 이 데이터에 중회귀분석을 적용하면 되는 것이다.

표 2. 데이터

아이템	프로의 내용(x_1)				시간대(x_2)				시청률(y)
No.	1	2	3	4	1	2	3	4	%
1	0	0	1	0	1	0	0	0	7.2
2	0	0	0	1	1	0	0	0	7.1
3	0	0	0	1	0	0	1	0	4.4
4	0	1	0	0	0	0	0	1	6.5
5	1	0	0	0	1	0	0	0	4.2
6	0	1	0	0	0	1	0	0	14.6
7	0	0	1	0	1	0	0	0	2.3
8	0	0	0	1	1	0	0	0	4.9
9	1	0	0	0	0	1	0	0	6.2
10	1	0	0	0	0	0	1	0	8.5
11	0	0	1	0	1	0	0	0	2.9
12	0	0	1	0	1	0	0	0	3.5
13	0	1	0	0	0	0	1	0	13.8

표 3. 데이터

아이템	프로의 내용(x_1)			시간대(x_2)			시청률
No.	x_{11}	x_{12}	x_{13}	x_{21}	x_{22}	x_{23}	y
1	0	0	1	1	0	0	7.2
2	0	0	0	1	0	0	7.1
3	0	0	0	0	0	1	4.4
4	0	1	0	0	0	0	6.5
5	1	0	0	1	0	0	4.2
6	0	1	0	0	1	0	14.6
7	0	0	1	1	0	0	2.3
8	0	0	0	1	0	0	4.9
9	1	0	0	0	1	0	6.2
10	1	0	0	0	0	1	8.5
11	0	0	1	1	0	0	2.9
12	0	0	1	1	0	0	3.5
13	0	1	0	0	0	1	13.8

◆ 더미변수

중회귀분석에서는 더미변수(dummy variable)라고 하는 것을 사용하고 있다. 데이터표(3)은 데이터표(1)에 있어서 다음과 같은 더미변수를 도입하여 바꿔 쓴 것에 지나지 않는다.

	X11	X12	X13
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

이와 같이 실은 수량화이론 I류라고 하는 수법은 설명변수가 모두 더미변수로 되어 있는 중회귀분석이라고 생각하면 된다. 여기에서 카테고리 4를 기준으로 해서 (0, 0, 0)로 했지만 기준이 되는 카테고리는 임의로 정할 수 있다.

◆ Excel에 의한 해법

〈순서1〉 데이터의 입력
데이터표(3)을 워크시트에 입력하여 처리한다.

〈순서 2〉 중회귀분석의 실행과 결과

	J	K	L	M	N	O	P	Q	R
1	요약 출력								
2									
3	회귀분석 통계량								
4	다중 상관계수	0.9183							
5	결정계수	0.8434							
6	조정된 결정계수	0.6867							
7	표준 오차	2.1406							
8	관측수	13							
9									
10	분산 분석								
11		자유도	제곱합	제곱 평균	F 비	유의한 F			
12	회귀	6	148.0113	24.6686	5.3838	0.0300			
13	잔차	6	27.4917	4.5820					
14	계	12	175.5031						
15									
16		계수	표준 오차	t 통계량	P-값	하위 95%	상위 95%	하위 95.0%	상위 95.0%
17	Y 절편	-1.7038	3.2245	-0.5284	0.6162	-9.5940	6.1863	-9.5940	6.1863
18	x11	0.5654	1.9237	0.2939	0.7787	-4.1419	5.2726	-4.1419	5.2726
19	x12	8.2038	2.4115	3.4019	0.0145	2.3030	14.1047	2.3030	14.1047
20	x13	-1.2365	1.7561	-0.7041	0.5077	-5.5336	3.0606	-5.5336	3.0606
21	x21	6.9154	3.2517	2.1267	0.0776	-1.0413	14.8721	-1.0413	14.8721
22	x22	7.7192	2.8161	2.7411	0.0337	0.8285	14.6099	0.8285	14.6099
23	x23	7.6808	2.8161	2.7275	0.0343	0.7901	14.5715	0.7901	14.5715

Sheet1 Sheet2 Sheet3

Excel에 있는 〔분석 도구〕의 대화상자 중에서 〔회귀분석〕을 선택하여 실행한다. 실행순서는 중회귀분석의 경우와 같으며, 그 실행결과는 앞의 〈순서2〉와 같다.

위의 실행결과로부터

$$y = 0.5654x_{11} + 8.2038x_{12} - 1.2365x_{13} + 6.9154x_{21} + 7.7192x_{22} + 7.6808x_{23} - 1.7038$$

라고 하는 회귀식이 얻어졌다.

여기에서 만일 새로운 프로를 택하여 조사해 본 결과 프로의 내용은 2. 연속극, 시간대는 4. 오후 10~11시라고 응답했을 때의 시청률은

$$y = 0.5654x_{11} + 8.2038x_{12} - 1.2365x_{13} + 6.9154x_{21} + 7.7192x_{22} + 7.6808x_{23} - 1.7038$$
$$= 0.5654 \times 0 + 8.2038 \times 1 - 1.2365 \times 0 + 6.9154 \times 0 + 7.7192 \times 0 + 7.6808 \times 0 - 1.7038$$
$$= 8.2038 - 1.7038 = 6.5(\%)$$

로 예측할 수 있다. 여기에 수량화이론 I류의 의미가 있다.

그리고 결정계수는

$R^2 = 0.8434$

이므로 설명변수의 변동으로 설명할 수 있는 정보는 84.34%로 비교적 위의 회귀식이 잘 들어맞고 있음을 알 수 있다.

분산분석표에서

유의한 $F = 0.03 <$ 유의수준 $\alpha = 0.05$

이므로 회귀식에는 의미가 있다고 할 수 있다.

Ⅲ. 수량화이론 II류

수량화이론 I류에서는 설명변수는 질적변수, 목적변수는 양적변수라고 하는 형식의 데이터를 다루었다. 이에 비해서 설명변수는 질적변수, 목적변수도 질적변수라고 하는 형식의 데이터를 다루는 것이 수량화이론 II류이다. 즉 수량화이론 I류의 경우와 마찬가지로 명목척도의 카테고리 변수로 되는 몇 개의 항목에 의해서 설명변수의 조가 구성되어 있는 경우, 각 카테고리에 가중치를 부여해서 각 그룹간의 차를 최대로 하는 합성변수를 정하는 수법이다. 형식적으로는 설명변수가 명목척도인 질적 데이터를 이용한 판별분석의 방법이라고 해도 무방하다. 그룹의 수가 r개 있는 경우에는 중판별분석의 경우와 마찬가지로 이론적으로는 전부해서 (r-1)종류의 합성변수가 얻어지게 되지만, 실용적으로는 상관비의 값이 작은 합성변수는 의미가 없다. 또 얻어진 각 카테고리의 가중치는 일의적으로 정해지지 않으므로 수량화이론 I류에서 설명되고 있는 순서에 따라서 각 카테고리의 가중치를 정하지 않으면 안 된다.

그리고 설명변수의 각 카테고리가 순서척도로 되고 게다가 각 카테고리에 주어지는 가중치 $a_1, a_2, \cdots, a_r$이 $a_1 \geq a_2 \geq \cdots \geq a_r$라고 하는 순서를 만족시키지 않으면 안 된다고 생각하는 경우에 있어서 선형계획법(linear programming)의 수법을 이용해서 위의 제한하에 군간의 분리를 최대로 하는 합성변수를 발견할 수 있다.

〔사례 2〕

다음과 같은 제품검사 항목이 있다. (항목 1)과 (항목 2)는 중간검사 항목이고 (항목 3)은 최종검사 항목으로 합격 여부를 가리는 것이다. 데이터표(1)은 부품 20개에 대한 검사성적표를 나타내고 있다.

(항목 1) 부품의 생산방식은 어떻게 이루어졌습니까?
(가) 수작업 (나) 자동생산

(항목 2) 부품의 조립과정에서 어느 점이 가장 중요합니까?
(가) 공차 (나) 도금 (다) 조립

(항목 3) 조립 완제품의 합격 여부는?
(가) 합격 (나) 불합격

(항목 1)과 (항목 2)의 응답으로부터 (항목 3)의 응답인 합격 여부를 예측(판별)하고 싶다. 이 문제에 대해서는 수량화이론 II류를 적용하여 해결할 수 있다. 수량화이론 II류는 정준판별분석에 있어서 모든 설명변수가 질적변수인 경우에 해당된다.

데이터 표(1)

아이템	항목 1		항목 2			항목 3
No.	가	나	가	나	다	y
1	1	0	1	0	0	가
2	1	0	1	0	0	가
3	1	0	1	0	0	가
4	1	0	1	0	0	가
5	1	0	0	1	0	가
6	1	0	0	1	0	가
7	1	0	0	1	0	가
8	1	0	0	1	0	가
9	0	1	1	0	0	가
10	0	1	1	0	0	가
11	1	0	0	1	0	나
12	1	0	0	0	1	나
13	0	1	0	1	0	나
14	0	1	0	1	0	나
15	0	1	0	1	0	나
16	0	1	0	0	1	나
17	0	1	0	0	1	나
18	0	1	0	0	1	나
19	0	1	0	0	1	나
20	0	1	0	0	1	나

수량화이론 II류는 2군의 판별과 3군 이상의 다군판별(多群判別)로 나누어서 생각할 수 있는데, 여기에서는 2군의 판별만 다루기로 한다.

◆ 수량화이론 II류에서 중회귀분석으로

수량화이론 II류를 2군의 판별문제로 국한해서 생각하면, 중회귀분석을 이용해서 해결할 수 있다. 즉, 설명변수 (항목 1)과 (항목 2)는 수량화이론 I류에서와 마찬가지로 (0, 1) 데이터로 바꿔 쓰고 질적인 목적변수 (항목 3)은 수치화함으로써 중회귀분석을 적용할 수 있다. 먼저 데이터표(1)을 데이터표(2)로 바꿔 쓴다.

데이터표(2)

No.	항목 1	항목 2	항목 3
1	가	가	가
2	가	가	가
3	가	가	가
4	가	가	가
5	가	나	가
6	가	나	가
7	가	나	가
8	가	나	가
9	나	가	가
10	나	가	가
11	가	나	나
12	가	다	나
13	나	나	나
14	나	나	나
15	나	나	나
16	나	다	나
17	나	다	나
18	나	다	나
19	나	다	나
20	나	다	나

◆ (0, 1) 데이터의 중회귀분석

(항목 1)(x_1)과 (항목 2)(x_2)로부터 (항목 3)(y)을 예측하는 식을 만들고자 하는 문제는 (0, 1) 데이터로 바꿔 쓴 데이터표(2)에 있어서 (항목 1)과 (항목 2)에 대한 각각의 카테고리를 x_{11}, x_{12}, x_{21}, x_{22}, x_{23}로 바꿔 표현하면 x_{11}에서 x_{23}까지를 설명변수, (항목 3)인 y를 목적변수로 하는 중회귀분석의 문제로 바꿔 놓을 수 있다.

이렇게 하면 데이터표(2)를 그대로 중회귀분석에 적용하면 될 듯 싶은데 실은 그렇게 잘 되지 않는다. 왜냐하면 설명변수 사이에는

$$x_{11} + x_{12} = 1$$

$$x_{21} + x_{22} + x_{23} = 1$$

이라고 하는 선형관계가 성립하고 있어, 바로 다중공선성 때문에 해가 구해지지 않는 것이다. 그래서 x_{11}, x_{12}중의 하나(여기에서는 x_{12})를 삭제하고, x_{21}, x_{22}, x_{23}중의 하나(여기에서는 x_{23})를 삭제한다. 데이터표(2)를 이와 같이 해서 바꿔 쓰면 데이터표(3)이 얻어진다.

수량화이론 I류에서도 더미변수(dummy variable)라고 하는 것을 소개한 바 있다. 데이터표(3)은 데이터표(1)에 있어서 다음과 같은 더미변수를 도입하여 바꿔 쓴 것에 지나지 않는다.

	X21	X22
가	1	0
나	0	1
다	0	0

여기에서 '다'를 기준으로 해서 (0, 0)로 했지만 기준이 되는 카테고리는 임의로 정할 수 있다.

그런데 데이터표(2)에서는 목적변수 y도 질적변수이기 때문에 수치화하지 않으면 안 된다. '가' 그룹(합격)에 속하는 대상의 수를 n_1, '나' 그룹에 속하는 대상의 수를 n_2라고 할 때,

'가' 그룹 → y = 1

'나' 그룹 → $y = -\frac{n_1}{n_2}$

로 수치화한다. 본 예제에서는 $n_1 = n_2 = 10$이므로,

'가' 그룹 → y = 1

'나' 그룹 → y = -1

로 수치화하면 된다. 이와 같이 해서 바꿔 쓴 데이터표(3)에 중회귀분석을 적용하면 되는 것이다.

데이터 표(3)

아이템	항목 1	항목 2		항목 3
No.	x_{11}	x_{21}	x_{22}	y
1	1	1	0	1
2	1	1	0	1
3	1	1	0	1
4	1	1	0	1
5	1	0	1	1
6	1	0	1	1
7	1	0	1	1
8	1	0	1	1
9	0	1	0	1
10	0	1	0	1
11	1	0	1	-1
12	1	0	0	-1
13	0	0	1	-1
14	0	0	1	-1
15	0	0	1	-1
16	0	0	0	-1
17	0	0	0	-1
18	0	0	0	-1
19	0	0	0	-1
20	0	0	0	-1

◆ Excel에 의한 해법

데이터표(3)을 워크시트에 입력하고, 〔분석 도구〕의 대화상자 중에서 〔회귀분석〕을 선택하여 실행한다. 실행순서는 중회귀분석의 경우와 같으며, 그 실행결과는 다음과 같다.

위의 실행결과로부터

$$y = 0.7423x_{11} + 1.6289x_{21}$$

$$+ 0.6598\ x_{22} - 1.1237$$

라고 하는 회귀식(판별식)이 얻어졌다. 이 식을 선형 판별식으로 사용하여,

$y > 0$이면 검사대상은 '가' 그룹(합격)에 속한다

$y < 0$이면 검사대상은 '나' 그룹(불합격)에 속한다

고 판별하면 된다.

분산분석표에서

유의한 F = 0.0001 〈 유의수준 α = 0.05

이므로 판별식에는 의미가 있다고 할 수 있다.

그리고 결정계수는 R^2=0.7113 이므로 설명변수의 변동으로 설명할 수 있는 정보는 71.13%로 비교적 위의 판별식이 잘 들어맞고 있다고 생각할 수 있으나, 이 경우에는 결정계수보다는 오판별률을 계산해 볼 필요가 있다.

회귀분석에 관한 통계함수 TREND를 이용해서 목적변수의 값을 예측하고 그 값의 양수·음수 부호에 의해서 그룹의 판별을 실시한다. 즉,

y의 검사결과 × y의 예측치 〈 0 → 오판별

y의 검사결과 × y의 예측치 ≥0 → 정판별

으로 판별한다.

〈순서 2〉의 화면에서,

〔셀의 입력내용〕

F3:

=TREND($E$3:$E$22,$B$3:$D$22,B3:D3, TRUE) (F3를 F4에서 F22까지 복사)

G3: =IF(E3*F3〈0, "오판별", "OK ") (G3를 G4에서 G22까지 복사)

〈순서1〉 중회귀분석의 실행결과

	G	H	I	J	K	L	M	N	O
1	요약 출력								
2									
3	회귀분석 통계량								
4	다중 상관계수	0.8434							
5	결정계수	0.7113							
6	조정된 결정계수	0.6572							
7	표준 오차	0.6007							
8	관측수	20							
9									
10	분산 분석								
11		자유도	제곱합	제곱 평균	F 비	유의한 F			
12	회귀	3	14.2268	4.7423	13.1429	0.0001			
13	잔차	16	5.7732	0.3608					
14	계	19	20.0000						
15									
16		계수	표준 오차	t 통계량	P-값	하위 95%	상위 95%	하위 95.0%	상위 95.0%
17	Y 절편	-1.1237	0.2502	-4.4906	0.0004	-1.6542	-0.5932	-1.6542	-0.5932
18	x11	0.7423	0.2988	2.4842	0.0244	0.1089	1.3757	0.1089	1.3757
19	x21	1.6289	0.3776	4.3135	0.0005	0.8284	2.4294	0.8284	2.4294
20	x22	0.6598	0.3521	1.8737	0.0794	-0.0867	1.4063	-0.0867	1.4063
21									
22									

Sheet1 / Sheet2 / Sheet3 / Sheet4 / Sheet5 /

〈순서 2〉 판별의 실행과 오판별률의 계산

G3 = =IF(E3+F3<0, "오판별", "OK")

	A	B	C	D	E	F	G
1	아이템	항목 1	항목 2		항목 3		
2	No.	x_{11}	x_{21}	x_{22}	y	예측치	판별
3	1	1	1	0	1	1.2474	OK
4	2	1	1	0	1	1.2474	OK
5	3	1	1	0	1	1.2474	OK
6	4	1	1	0	1	1.2474	OK
7	5	1	0	1	1	0.2784	OK
8	6	1	0	1	1	0.2784	OK
9	7	1	0	1	1	0.2784	OK
10	8	1	0	1	1	0.2784	OK
11	9	0	1	0	1	0.5052	OK
12	10	0	1	0	1	0.5052	OK
13	11	1	0	1	-1	0.2784	오판별
14	12	1	0	0	-1	-0.3814	OK
15	13	0	0	1	-1	-0.4639	OK
16	14	0	0	1	-1	-0.4639	OK
17	15	0	0	1	-1	-0.4639	OK
18	16	0	0	0	-1	-1.1237	OK
19	17	0	0	0	-1	-1.1237	OK
20	18	0	0	0	-1	-1.1237	OK
21	19	0	0	0	-1	-1.1237	OK
22	20	0	0	0	-1	-1.1237	OK

Sheet1 / Sheet2 / Sheet3 / Sheet4 / Sheet5

오판별률의 계산

No. 11의 판별대상이 실제는 불합격인데 합격이라고 잘못 판별되어 있다. 따라서 오판별률은 다음과 같이 계산된다.

$$\text{오판별률} = \frac{1}{20} = 0.05$$

Ⅳ. 결론

본고에서 다룬 선형모형(중회귀분석)은 양적 데이터를 다변량으로 분석하는 기법인데, 이를 이용해서 양적 데이터 뿐만 아니라 질적 데이터도 목적에 따라서 적당한 수량으로 변환하여 분석하고자 하는 방법이 수량화이론 Ⅰ류와 Ⅱ류이다.

즉 수량화이론 Ⅰ류는 명목척도의 범주변량으로 되어 있는 몇 개의 요인에 의해서 설명변수의 조가 구성되는 경우에, 외적 기준에 대한 범주변량의 예측오차를 최소로 하는 가중치를 각 범주에 부여하는 방법이다. 형식적으로는 설명변수가 명목척도로 되어 있는 더미변수의 데이터를 이용한 중회귀분석이라고 해도 무방하다. 그러나 각 요인에 포함되는 모든 범주에 주어진 범주변량의 총합은 항상 1이 되므로, 실제의 계산에서는 각 요인에서 1개씩의 범주를 제거하고 분석을 실시하며, 제거한 범주변량에는 가중치 0을 부여하면 된다. 한편 수량화이론 Ⅱ류는 형식적으로는 설명변수가 명목척도인 질적 데이터를 이용한 판별분석과 같은 방법이다. 즉 수량화이론 Ⅰ류의 경우와 마찬가지로 명목척도의 범주변량으로 되는 몇 개의 요인에 의해서 설명변수의 조가 구성되어 있는 경우, 각 범주에 가중치를 부여해서 각 군간의 차가 최대가 되도록 하는 판별함수를 정하는 기법이다. 군의 수가 G개 있는 경우에는 중판별분석의 경우와 마찬가지로 이론적으로는 전부해서 (G-1) 종류의 판별함수가 얻어지게 되는데, 실용적으로는 상관비의 값이 작은 판별함수는 의미가 없다. 수량화이론 Ⅱ류의 분석목적은 다음과 같다.

(1) 요인, 범주 및 표본을 수량화함으로써 분류를 수량적으로 실시하는 것
(2) 분류에 대한 각 요인의 기여도를 수량적으로 표현하는 것
(3) 데이터로서 얻어져 있지 않은 새로운 유형의 데이터가 어느 군에 가장 가까운가를 수량적으로 표시하고자 하는 것

참고문헌

〔1〕 노형진, 다변량해석-질적 데이터의 수량화-, 석정, 1990.

〔2〕 노형진·한상도·장명복, Excel을 활용한 경영자료분석, 청암미디어, 1998.

〔3〕 노형진, Excel에 의한 조사방법 및 통계분석, 법문사, 1998.

〔4〕 노형진, Excel을 활용한 품질경영, 청암미디어, 1999.

〔5〕 노형진, Excel에 의한 통계적 조사방법, 형설출판사, 2000.

〔6〕 內田治, すぐわかるEXCELによる統計解析, 東京圖書, 1996.

〔7〕 內田治, すぐわかるEXCELによる多變量解析, 東京圖書, 1996.

〔8〕 奧野·芳賀·久米·古澤, 多變量解析法, 日科技連出版, 1971.

〔9〕 小林龍一,相關·回歸分析入門, 日科技連出版, 1982.

〔10〕 小林龍一, 數量化入門, 日科技連出版, 1981.

〔11〕 林知己夫, 數量化の方法, 東洋經濟新報社, 1974.

〔12〕 林知己夫, データ解析の方法, 東洋經濟新報社, 1974.

〔13〕 ラッヘンブルック(鈴木·三宅譯), 判別分析, 現代數學社, 1979.

〔14〕 Kendall, M. G. : *Multivariate Analysis*, Charles Griffin, 1975.

〔15〕 Lachenbruch, P. A. : *Discriminant Analysis*, Hafner, 1975.

〔16〕 Seber, G. A. F. : *Linear Statistical Analysis*, John Wiley & Sons, 1977.

저자소개

노 형 진

서울대학교 공과대학 졸업(공학사)
고려대학교 대학원 수료(경영학박사)
일본 쓰쿠바대학 대학원 수료
(경영공학 박사과정)
일본 동경대학 객원교수
현재, 경기대학교 경영학부 교수
관심분야 : 품질경영, 다변량분석, 6시그마, Single PPM품질혁신