응용통계연구 (2008)
21(4), 621–630

반복측정의 다가 반응자료에 대한 일반화된 주변 로짓모형

최재성[1]

[1]계명대학교 통계학과

(2008년 4월 접수, 2008년 6월 채택)

요 약

본 논문은 개체의 특성으로 다가의 명목형 반응변수가 반복측정 요인인 시간요인에 의해 주기적으로 반복측정 되었을 때, 자료를 분석하기 위한 모형으로 일반화된 주변 로짓모형을 논의하고 있다. 다가의 반응변수에 영향을 미치는 공변량중 일부가 처치로써 상대적으로 큰 크기의 실험단위에 배정되고 반복측정 요인인 시간요인의 수준들이 또한 처치요인으로 비확률화에 의해 상대적으로 작은 크기의 실험단위에 배정될 때 이를 고려한 모형구축과정과 예상되는 공분산 구조의 가정하에서 모수를 추정하기 위한 방법으로 가중최소제곱 방법을 이용할 수 있음을 제시하고 있다.

주요용어: 일반화된 주변로짓, 고정효과, 반복측정, 가중최소제곱법.

1. 서론

일반화된 로짓 선형모형에 관한 연구는 많은 문헌들에서 논의되고 있다. Agresti (1990)는 단일 반응변수가 다가의 명목형일 때, 자료를 분석하기 위한 일반화된 로짓 선형모형을 다루고 있다. 개체의 반응변수에 대한 관측이 셋 이상 다가의 형식적인 구분을 위한 범주들로 주어질 때, 다가의 명목형 변수로 정의한다. 다가의 명목범주들에 대한 로짓변환은 일반적으로 기준범주(baseline category)에 의한 로짓변환을 이용한다. 일반화된 로짓모형은 개체의 반응에 영향을 미치는 독립변수들의 유형과 표본을 추출하는 방법 또는 실험계획과 관련하여 다양한 유형의 일반화된 로짓모형을 논의할 수 있다. Mcfadden (1974)은 설명변수들로 상품의 선택특성들을 고려한 이산적인 선택모형으로 일반화된 로짓모형을 제안했다. Hosemer와 Lemeshow (2000)는 다가의 로지스틱 회귀모형에 관한 모형 구축 과정과 자료 분석방법을 논의하고 있다. 최재성 (2004)은 관심질병의 예방백신 효과를 알아보기 위한 비이환율 자료분석에 대한 모형으로 일반화된 혼합효과 모형을 다루고 있다. Koch와 Reinfurt (1971), Koch 등 (1977)은 처음으로 반복측정의 범주형 자료에 가중최소제곱법(weighted least squares)을 적용했다. Liang과 Zeger (1986)는 일반화된 선형모형으로 모형을 세울 수 있는 종속자료를 다루기 위한 방법으로 GEE(generalized estimating equations) 방법을 소개했다. 실험 또는 조사에서 개체의 관심특성을 나타내는 반응변수가 다가의 명목형 반응변수일 때, 관심은 일반적으로 반응범주들의 확률의 추론에 있게 된다. 개체의 반응변수를 Y라 둘 때, Y가 $h = 1, 2, \ldots, l$개의 명목범주 중 한 범주로 관측된다면, 관심범주들의 수는 $l-1$개 이다. 따라서, 관심확률들의 수도 $l-1$개 이다. 그러나, 반응범주들의 확률에 영향을 미치는 공변량들이 다수 존재할 때, 모형에 근거하여 이들 변수의 효과를 추론하기 위해서는 관측범주의 확률에 대한 다양한 변환함수들을 생각할 수 있다. 가능한 변환함수들과 관련된 모형들에 관한 연구가 McCullagh와 Nelder (1989) 그리고 Agresti (1990)에서 보여진다.

[1](704-701) 대구광역시 달서구 신당동 1000번지, 계명대학교 자연과학대학 통계학과, 교수.
E-mail: jschoi@kmu.ac.kr

관측 또는 실험에서 개체의 반응에 영향을 미칠 수 있는 공변량들이 자료수집을 위한 표본추출 방법, 또는 실험계획으로 인해 서로 다른 크기의 실험단위들에서 측정되는 경우를 생각해 보기로 한다. 실험계획으로는 분할구 실험계획을 생각할 수 있다. 분할구 계획에서 처치들이 요인들의 결합수준들로 주어지고 일부 요인들의 수준은 큰 단위의 실험단위들에 임의로 배정되고 다른 요인들의 수준들은 상대적으로 작은 실험단위에서 관측되는 불완전 블록 설계구조하의 처치구조를 갖게 된다. 표본추출 방법으로는 집락추출 방법을 생각할 수 있다. 이원 집락추출을 이용할 때, 일차의 추출단위들은 큰 단위의 실험단위 또는 개체로 간주되고 이차의 추출단위들은 상대적으로 크기가 작은 실험단위 또는 분할된 개체들로 간주된다. 이와 같은 경우에 반응변수의 관측범주들은 상대적으로 크기가 작은 실험단위 또는 분할된 개체들에서 관측된다. 따라서, 처치로 주어지는 관심요인들에 대한 추론은 관심요인들의 수준들이 어느 크기의 실험단위들에서 관측되는 가에 따라 해당하는 실험단위들의 변이에 따른 처치간의 비교가 이루어져야 한다. 본 연구에서는 처치들에 서로 다른 크기의 실험단위를 고려하고 있다. 자료가 연속형의 자료일 때, 이와 관련한 모형에 관한 논의는 Milliken과 Johnson (1984)에서 볼 수 있다.

개체에 대한 반응이 시간 또는 다른 반복측정 요인에 의해 반복적으로 측정될 때, 동일 개체의 반복측정으로 인해 반응간에 종속성을 가정할 수 있게 된다. 이때, 반복측정 요인이 시간과 같은 연구자가 임의로 조정할 수 없는 요인일 때, 동일 개체에 대해 반복측정으로 얻어지는 반응값들 간의 공분산 구조는 개체내 변이의 특성을 감안한 공분산 구조가 고려되어야 한다. 주어진 공분산 구조하에서 범주형 자료를 일반화된 로짓 선형모형의 가정하에 분석하는 방법으로 최대우도법, 가중최소제곱법, GEE 방법 등이 이용된다. 이들 방법들은 때로는 개체의 반응함수에 대한 선택과 고려된 독립변수들의 유형에 따라 분석에 제한적일 수 있다. 즉, 선택된 반응함수에 적용할 수 있는 분포에 대한 가정이 성립하지 않는다면 최대우도법을 이용할 수 없게 되고, 공변량중 연속변수가 존재하거나 결측치가 있을 때 GEE 방법을 이용할 수 있다. 그러나, 본 연구에서는 일반화된 주변 로짓모형의 가정하에 가중최소제곱 방법을 이용하여 자료를 분석하는 방법을 논의해 보기로 한다.

2. 모형의 가정

연구자의 관심모집단에서 개체의 특성을 나타내는 반응변수를 Y라 둔다. 반응변수 Y는 $h = 1, 2, \ldots, l$ 중 하나의 범주로 관측되는 다가의 명목형 반응변수라 가정한다. 반응범주 h의 확률을 π_h라 둔다. π_h에 영향을 미치는 공변량들로 두 명목형 변수 A, B와 시간요인 T를 가정한다. 명목형 변수 A는 a개의 수준, $i = 1, 2, \ldots, a$을 갖는 요인이라 가정한다. 명목형 변수 B는 b개의 수준, $j = 1, 2, \ldots, b$를 갖는 요인으로 정의한다. 요인 A의 수준 i와 요인 B의 수준 j의 결합수준을 (i, j)라 둘 때, ab개의 모든 결합수준들이 처치로써 전체 크기의 개체에 임의로 배정된다고 가정한다. 시간요인 T는 T의 g개 수준, $k = 1, 2, \ldots, g$들이 작은 단위의 분할된 개체들에 비확률화로 배정되는 반복측정 요인이라 가정한다. 동일 개체가 시간요인 T의 g개 시점에서 반복측정 되므로 단일 반응변수에 대한 관측벡타는 다변량 벡타로 주어진다. 반응변수 벡타 $\mathbf{Y}$를 $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_g)'$로 정의한다. 동일 개체의 주기적인 반복측정으로 인한 반응변수들 간의 종속성을 가정한다. 따라서, 벡타 $\mathbf{Y}$는 반응들 간의 종속성으로 인한 공분산구조를 갖는 다변량의 결합확률 분포를 갖게 된다. 고려된 세 변수들의 수준결합에서 관측되는 반응변수 Y_k가 범주 h로 반응할 확률을 $\pi_h(k; ij)$라 둔다. 따라서, 개체단위의 처치 (i, j)에서 시간요인 T의 시점 k에서 관측되는 반응변수 Y_k의 l개 반응범주들에 대한 확률벡타를 $\boldsymbol{\pi}_{(k;ij)}$라 두면, $\boldsymbol{\pi}_{(k;ij)} = [\pi_{1(k;ij)}, \pi_{2(k;ij)}, \ldots, \pi_{l(k;ij)}]'$이다. 요인 A의 수준 i와 요인 B의 수준 j의 수준결합 (i, j)에서 시간요인 T의 g개 시점에서 관측되는 범주들의 확률벡타를 $\boldsymbol{\pi}_{(ij)}$라 두면, $\boldsymbol{\pi}_{(ij)} = [\boldsymbol{\pi}_{(1;ij)}{}', \boldsymbol{\pi}_{(2;ij)}{}', \ldots, \boldsymbol{\pi}_{(g;ij)}{}']'$로 주어진다.

다가의 범주형 자료분석의 목적이 관심 범주들의 확률에 대한 추론일 수도 있으나 때로는 관심 범주들의 주변확률의 변화추이에 대한 경향을 알아보는데 그 목적이 있을 수 있다. 예를 들면, 신세대들의 취업경향에 대한 추세변이를 알아보기 위한 조사에서 다범주의 취업군에서 선호하는 직업들을 여러 시점에서 관측함으로써 경향에 대한 정보를 파악해 볼 수 있다. 이러한 목적의 자료분석에 이용되는 반응함수로는 주변확률, 주변확률의 변환함수인 주변로짓, 반응범주들이 일정한 순서를 갖는 다가의 범주형 변수일 때는 누적확률 로짓을 이용할 수 있다.

본 연구에서는 다가의 범주형 변수가 명목형임을 가정하므로 반응함수의 변환함수로 기준범주의 주변로짓변환을 가정한다. 따라서, 가정된 실험환경에서 고려된 공변량들의 효과를 알아 보기위한 일반화된 선형모형으로 기준범주의 주변로짓을 이용한 주변 로짓모형을 논의한다. 실험 또는 관측에서 이용되는 실험단위나 개체는 독립적인 개체들로 관측되므로 개체간의 관심 반응변수의 결과들에 대해 변이에 대한 독립성을 가정한다. 그러나 개체내 시간요인의 수준변화에 따른 관측범주들의 변이는 동일 개체에 대해 반복측정 되므로 종속성을 가정한다. 특히, 반복측정 요인으로 시간을 고려하고 있으므로 시간수준의 변화에 따른 동일 개체의 반응범주들 간의 종속성은 분할구 실험과는 달리 연구자 임의로 배정할 수 없는 요인의 수준들로 인한 공분산 구조를 염두에 둔다. 가정된 실험환경을 만족하고 있는 타당한 모형설정을 논의하고 모형내 모수의 추론과 함께 자료분석을 위한 방법으로 가중최소제곱 방법을 논의해 보기로 한다.

3. 모형에 관한 논의

이가 또는 다가의 범주형 자료를 분석하기 위한 일반화된 선형모형에 관한 연구에 비해 반복측정의 다가자료를 분석하기 위해 일반화된 로짓(generalized logits)을 이용한 주변 로짓모형에 관한 연구는 흔치 않다. 따라서, 2절의 실험환경의 가정하에 다가의 반복측정된 명목형 자료를 분석하기 위해서 기준범주의 일반화된 주변 로짓모형을 논의해 보기로 한다. 우선, 반응함수로 주변확률의 변환인 주변로짓을 고려하고 있기 때문에 다가의 범주형 변수에 대한 주변확률을 생각해 보기로 한다.

요인 A의 수준 i와 요인 B의 수준 j의 수준결합 (i, j)에서 시간요인 T의 시점 k에서 주어지는 관측벡타는 $\boldsymbol{\pi}_{(k;ij)} = [\pi_{1(k;ij)}, \pi_{2(k;ij)}, \ldots, \pi_{l(k;ij)}]'$이다. 그리고 요인 A의 수준 i와 요인 B의 수준 j의 수준결합 (i, j)에서 시간요인 T의 g개의 모든 시점에서 관측될 수 있는 결합 반응범주들에 대한 확률벡타는 $\boldsymbol{\pi}_{(ij)}$이다. 요인 A의 수준 i와 요인 B의 수준 j의 수준결합 (i, j)에서 시간요인 T의 g개 시점에서 관측되는 g개 반응변수 Y들의 결합범주들의 수는 l^g개 이다. 주변범주 h에 대한 주변확률을 $\phi_h(k; ij)$라 두자. $\phi_h(k; ij)$를 구하기 위한 계수벡타를 $\boldsymbol{x}_{h(k;ij)}$라 둘 때, $\boldsymbol{x}_{h(k;ij)}$=[0, 0, ..., 1, ..., 0, ..., 0, 0 ..., 1, ..., 0]로 주어진다. $\boldsymbol{x}_{h(k;ij)}$에서 1은 시점 k에서 반응변수 Y가 범주 h로 관측되는 경우를 나타낸다. $\boldsymbol{x}_{h(ij)}$를 요인 A의 수준 i와 요인 B의 수준 j의 수준결합 (i, j)에서 시간요인 T의 g개 시점에서 관측되는 l^g개의 결합범주중 범주 h와 관련된 계수벡타라 두자. 이때, $\boldsymbol{x}_{(ij)} = [\boldsymbol{x}_{h(1;ij)}', \boldsymbol{x}_{h(2;ij)}', \ldots, \boldsymbol{x}_{h(l;ij)}']'$로 주어진다. 따라서, 범주 h의 주변확률을 $\phi_h(ij)$라 두면, $\phi_h(ij)$는 $\boldsymbol{x}_{(ij)}$와 $\boldsymbol{\pi}_{(ij)}$의 내적으로 주어진다.

예로써, 주변확률의 정의를 위해 $g = 2$인 경우에 다가 반응변수 Y_i가 세 개의 범주를 가질 때, 6개의 주변확률을 구해보기로 한다. 각 시점에서 반응변수 Y_i ($i = 1, 2$)는 세 개의 범주중 하나로 관측되는 다가반응이므로 두 시점에서 3^2개의 가능한 결과들이 존재한다. 세 개의 가능한 범주들을 각기 f, o, p로 나타내기로 할 때, 모든 가능한 결과들은 $\{ff, fo, fp, of, oo, op, pf, po, pp\}$이다. 이들 결과들의 확률벡타를 π라 두면, $\boldsymbol{\pi} = [\pi_{ff}, \pi_{fo}, \pi_{fp}, \pi_{of}, \pi_{oo}, \pi_{op}, \pi_{pf}, \pi_{po}, \pi_{pp}]'$이다. $g = 2$개의 시점에서 9개

결합범주의 6개 주변확률들은

$$\begin{pmatrix} \phi_f(k=1;ij) \\ \phi_o(k=1;ij) \\ \phi_p(k=1;ij) \\ \phi_f(k=2;ij) \\ \phi_o(k=2;ij) \\ \phi_p(k=2;ij) \end{pmatrix} = \begin{pmatrix} 1\,1\,1\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,1\,1\,1\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,1\,1\,1 \\ 1\,0\,0\,1\,0\,0\,1\,0\,0 \\ 0\,1\,0\,0\,1\,0\,0\,1\,0 \\ 0\,0\,1\,0\,0\,1\,0\,0\,1 \end{pmatrix} \begin{pmatrix} \pi_{ff} \\ \pi_{fo} \\ \pi_{fp} \\ \pi_{of} \\ \pi_{oo} \\ \pi_{op} \\ \pi_{pf} \\ \pi_{po} \\ \pi_{pp} \end{pmatrix} \tag{3.1}$$

로 주어진다.

$$\boldsymbol{\phi}_1(ij) = \begin{pmatrix} \phi_f(k=1;ij) \\ \phi_o(k=1;ij) \\ \phi_p(k=1;ij) \\ \phi_f(k=2;ij) \\ \phi_o(k=2;ij) \\ \phi_p(k=2;ij) \end{pmatrix}$$

라 두고,

$$\mathbf{B}_1 = \begin{pmatrix} 1\,1\,1\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,1\,1\,1\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,1\,1\,1 \\ 1\,0\,0\,1\,0\,0\,1\,0\,0 \\ 0\,1\,0\,0\,1\,0\,0\,1\,0 \\ 0\,0\,1\,0\,0\,1\,0\,0\,1 \end{pmatrix}$$

라 두면, 식 (3.1)은

$$\boldsymbol{\phi_1}(ij) = \boldsymbol{B}_1\boldsymbol{\pi} \tag{3.2}$$

로 표현된다. 선형변환 행렬 $\mathbf{B}_1$에 의한 주변확률들의 합은 1이 되므로 두 개의 개별시점에서 선형적으로 독립인 주변확률들의 수는 둘이다. 범주 p를 기준범주로 둘 때, 각 시점에서 두 개의 선형적으로 독립인 주변확률을 얻을 수 있다. 즉, $\phi_f(k;ij)$와 $\phi_o(k;ij)$이다. 따라서, 선형적으로 독립인 주변확률 벡타를 $\boldsymbol{\phi}_2(ij)$, 관련된 선형변환 행렬을 $\mathbf{B}_2$라 두면,

$$\boldsymbol{\phi}_2(ij) = \begin{pmatrix} \phi_f(k=1;ij) \\ \phi_o(k=1;ij) \\ \phi_f(k=2;ij) \\ \phi_o(k=2;ij) \end{pmatrix}$$

이고,

$$\mathbf{B}_2 = \begin{pmatrix} 1\,1\,1\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,1\,1\,1\,0\,0\,0 \\ 1\,0\,0\,1\,0\,0\,1\,0\,0 \\ 0\,1\,0\,0\,1\,0\,0\,1\,0 \end{pmatrix} \tag{3.3}$$

이다.

관심결과의 주변로짓을 이용한 모형설정을 고려하고 있으므로 로짓변환을 생각해 보기로 한다. 관심결과의 주변로짓은 관심결과의 주변확률을 기준범주 p의 주변확률의 비로써 주어진다. 먼저, 기준범주의 주변확률을 구해보기로 한다. 요인 A의 i번째 수준, 요인 B의 j번째 수준에서 반복측정 요인 T의 k번째 수준에서 기준범주 p의 주변확률을 $\phi_p(k;\, ij)$라 두자. $g = 2$개의 시점에서 기준범주 p의 주변확률들은

$$\begin{pmatrix} \phi_p(k=1;\, ij) \\ \phi_p(k=2;\, ij) \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \pi_{ff} \\ \pi_{fo} \\ \pi_{fp} \\ \pi_{of} \\ \pi_{oo} \\ \pi_{op} \\ \pi_{pf} \\ \pi_{po} \\ \pi_{pp} \end{pmatrix} \tag{3.4}$$

로 계산된다.

$$\boldsymbol{\phi}_p(ij) = \begin{pmatrix} \phi_p(k=1;\, ij) \\ \phi_p(k=2;\, ij) \end{pmatrix}$$

라 두고,

$$\mathbf{B}_p = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$

라 두면, 식 (3.4)는

$$\boldsymbol{\phi}_p(ij) = \mathbf{B}_p \boldsymbol{\pi} \tag{3.5}$$

로 표현된다.

따라서, 요인 A의 i번째 수준, 요인 B의 j번째 수준 그리고 반복측정 요인 T의 k번째 수준에서 관심범주 h의 주변로짓을 $L_h(k;\, ij)$라 둘 때,

$$L_h(k; ij) = \frac{\phi_h(k; ij)}{\phi_p(k; ij)}, h = f, o.$$

로 정의된다.

개체단위에서 관측되는 두 요인 A와 B의 한 결합수준에서 시간요인 T의 g개 시점에서 관측되는 g개 다가 반응변수들의 결합범주들의 수는 l^g개다. 시간요인 T의 g개 시점에서 관측되는 l^g개 범주들에 대한 확률벡타를 $\boldsymbol{\pi}_{(ij)}$라 두면

$$\boldsymbol{\pi}_{(ij)} = [\pi_1(1; ij), \pi_2(1; ij), \ldots, \pi_l(1; ij), \ldots, \pi_1(g; ij), \pi_2(g; ij), \ldots, \pi_l(g; ij)]'$$

로 주어진다. 두 요인 A와 B의 결합수준 $(i,\, j)$에서의 범주 h의 주변확률을 $\phi_h(ij)$라 두게 되면, 계수벡타 $\mathbf{x}_{(ij)}$와 $\boldsymbol{\pi}_{(ij)}$의 내적은 $\phi_h(ij) = [0, 0, \ldots, 1, \ldots, 0, \ldots, 0, 0 \ldots, 1, \ldots, 0,]\boldsymbol{\pi}_{(ij)}$로 주어진다. 따라

서, 반응변수 Y의 l개 범주들의 주변확률 벡타를 $\mathbf{F}(ij)$, 주변로짓 벡타를 $\mathbf{L}(ij)$라 두면,

$$\boldsymbol{F}(ij)=\begin{pmatrix}\phi_1(k=1;ij)\\ \phi_2(k=1;ij)\\ \cdots\\ \phi_{l-1}(k=1;ij)\\ \cdots\\ \phi_1(k=g;ij)\\ \phi_2(k=g;ij)\\ \cdots\\ \phi_{l-1}(k=g;ij)\end{pmatrix}=\boldsymbol{A}_1\boldsymbol{\pi}_{(ij)}$$

이고,

$$\mathbf{L}(ij)=\begin{pmatrix}\phi_1(k=1;ij)/\phi_l(k=1;ij)\\ \phi_2(k=1;ij)/\phi_l(k=1;ij)\\ \cdots\\ \phi_{l-1}(k=1;ij)/\phi_l(k=1;ij)\\ \cdots\\ \phi_1(k=g;ij)/\phi_l(k=g;ij)\\ \phi_2(k=g;ij)/\phi_l(k=g;ij)\\ \cdots\\ \phi_{l-1}(k=g;ij)/\phi_l(k=g;ij)\end{pmatrix}$$

가 된다. 각 시점에서 선형적으로 독립인 주변확률들의 수는 $l-1$개이므로 g개 시점에서의 주변확률들의 수는 $g(l-1)$개 이다. 여기서 $\mathbf{A}_1$는 주변확률을 구하기 위한 선형변환 행렬을 나타낸다. 개체요인들로 고려된 두 요인 A와 B의 a개 수준과 b개의 수준결합들이 개체에 대한 처치로 간주될 때, $abg(l-1)$개의 선형적으로 독립인 주변확률들을 구할 수 있다. $abg(l-1)$개의 주변확률들의 벡타를 $\mathbf{F}(\boldsymbol{\pi})$라 두자. 선형변환 행렬을 $\mathbf{A}$, abl^g개의 반응범주를 나타내는 반응 프로필에서의 확률벡타를 $\boldsymbol{\pi}$라 두면, $\mathbf{F}(\boldsymbol{\pi})=\mathbf{A}\boldsymbol{\pi}$가 된다. 주변확률 벡타 $\mathbf{F}(\boldsymbol{\pi})$를 선형변환 행렬에 의해 구한 다음 각 시점에서 구해진 기준범주의 주변확률로 나누면 일반화된 주변로짓 벡타 $\mathbf{L}(\boldsymbol{\pi})$를 얻게 된다. 요인 A의 수준 i, 요인 B의 수준 j의 결합수준내 시간요인 T의 시점 k에서 관심범주 h의 기준범주에 의한 주변로짓을 $\mathrm{L}_h(k;ij)$라 둘 때, $\mathrm{L}_h(k;ij)$는 $\mathbf{L}(\boldsymbol{\pi})$내 한 원소이다. 따라서, 관심결과의 주변로짓을 이용한 로짓 고정효과 모형은 다음과 같이 나타낼 수 있다.

$$L_h(k;ij)=\mu+\alpha_i+\beta_j+(\alpha\beta)_{ij}+\tau_k+(\alpha\tau)_{ik}+(\beta\tau)_{jk}+(\alpha\beta\tau)_{ijk}, \tag{3.6}$$

단, $h=1,2,\ldots,l,\ \ k=1,2,\ldots,g,\ \ i=1,2,\ldots,a$이며 $j=1,2,\ldots,b$이다

여기서 μ는 주변로짓의 절편을 나타낸다. α_i는 요인 A의 i번째 수준효과이고 β_j는 요인 B의 j번째 수준효과를 나타낸다. $(\alpha\beta)_{ij}$는 요인 A와 요인 B의 결합수준 (i,j)에서 관측되는 교호작용을 나타낸다. τ_k는 요인 A와 요인 B의 결합수준 (i,j)에서 관측되는 시간요인 T의 수준 k에서의 효과이며, $(\alpha\tau)_{ik}$, $(\beta\tau)_{jk}$는 각기 요인 A와 T, 요인 B와 T의 두 요인 교호작용이다. $(\alpha\beta\tau)_{(ijk)}$는 세 요인의 결합수준 (i,j,k)에서 주어지는 세 요인 교호작용을 나타낸다. 식 (3.6)의 가정은 요인 A와 B의 결합수준이 처치로써 상대적으로 큰 크기의 실험단위인 개체들에 임의로 배정됨을 가정하고 있다. 반면에 시간요인의 수준들은 작은 단위의 분할된 개체들에 비확률화에 의해 배정됨을 나타낸다. 따라서, 모형내 모수들에

관한 추론은 해당하는 실험단위들에서의 공분산구조하에서 행해짐을 의미하고 있다. 식 (3.6)에서

$$\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

는 모형의 개체단위에서 주어지는 모수들로 개체간의 변이에 따른 공분산 구조하에서 추론되어야 할 부분이다. 한편,

$$\tau_k + (\alpha\beta)_{ik} + (\beta\tau)_{jk} + (\alpha\beta\tau)_{ijk}$$

는 시간구간의 소단위에서 주어지는 모수들로 개체내 변이 즉, 변수들 간의 종속성을 나타내는 공분산 구조하에서 추론되어야 할 부분이다. 가정된 실험환경 속에서 제안된 주변확률 모형의 가정하에 자료분석하는 방법을 구체적인 예로써 살펴보기로 한다. 식 (3.6)은 시간요인 T의 수준변화를 고려할 때,

$$L_h(k;ij) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \delta\nu_k + \delta_i\nu_k + \delta_j\nu_k + \delta_{ij}\nu_k \tag{3.7}$$

로 변형하여 이용할 수 있게 된다. 여기서 δ는 양적인 시간(ν_k)의 변화에 따른 계수를 나타낸다. 가정된 모형하에서 모수를 추론하기 위한 가중최소제곱 방법을 살펴보기로 한다. 표본을 취한후 abl^g개의 반응범주를 나타내는 반응 프로필에서 관측되는 비율벡타 $\boldsymbol{p}$는 $E(\boldsymbol{p}) = \boldsymbol{\pi}$인 표본 비율벡타라 두자. 표본 비율벡타를 가정된 주변 로짓모형에 적합시킬 때, $\mathbf{L}(\boldsymbol{p}) = \mathbf{C}\log(\mathbf{A}\boldsymbol{p}) = \boldsymbol{X\beta}$가 된다. 여기서 $\boldsymbol{X}$는 모형행렬을 나타낸다. 행렬 $\boldsymbol{C}$는 주변 비율벡타 $\mathbf{F}(\boldsymbol{p})$를 기준범주에 의한 로짓변환과 관련된 행렬을 나타낸다. $\mathbf{L}(\boldsymbol{p})$의 공분산 행렬을 $\boldsymbol{V_L}$라 두면, $\boldsymbol{V_L} = \boldsymbol{QV}(\boldsymbol{p})\boldsymbol{Q'}$의 형태가 된다. $\boldsymbol{Q}$는 $\mathbf{L}(\boldsymbol{p})$의 공분산 행렬과 관련된 행렬이다. 가중최소제곱법에 의한 모수의 추정량과 공분산 행렬은 다음과 같이 주어진다. $\hat{\boldsymbol{\beta}} = (\boldsymbol{X'V_L}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X'V_L}^{-1}\boldsymbol{L}(\boldsymbol{p})$이고 $\boldsymbol{V}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X'V_L}^{-1}\boldsymbol{X})^{-1}$이다.

4. 선거자료 예

유력한 세명의 대선후보자들에 대해 유권자들이 선호하고 있는 후보자의 지지비율이 선거기간에 발생하고 있는 선거요인들로 인해 어떻게 변화하고 있는가를 판단하기 위해 두 번($g = 2$)에 걸쳐 여론조사를 한다고 하자. 세 명의 유력한 대선후보(Y)를 m, j, k라 하자. 지역요인(A)의 세 수준으로 수도권(a_1), 호남권(a_2), 영남권(a_3)의 세 지역을 선택한다. 선거요인(B)의 두 수준으로 경제여건(b_1)과 정부의 실책건(b_2)을 생각한다. 다음은 선거자료를 나타내고 있는 생성자료표이다.

위 자료를 분석하기 위한 모형으로 식 (3.6)을 적용하여 보기로 한다. 모형내 모수들을 추정하기 위해 가중최소제곱 방법을 이용할 때, 모형 (3.6)내 모수들의 추정값과 추정오차는 다음과 같이 주어진다.

$$\hat{\mu} = 0.0086(0.0932),\ \hat{\alpha_1} = 0.5324(0.1348),\ \hat{\alpha_2} = 1.1776(0.1718),\ \hat{\beta_1} = 0.0376(0.1340),$$
$$(\widehat{\alpha\beta})_{11} = 0.8181(0.1915),\ (\widehat{\alpha\beta})_{21} = -0.1495(0.2339),\ \hat{\tau_1} = 1.0618(0.1358),$$
$$(\widehat{\alpha\tau})_{11} = -0.7381(0.1857),\ (\widehat{\alpha\tau})_{21} = -1.7474(0.2209),\ (\widehat{\beta\tau})_{21} = -1.0076(0.1732),$$
$$(\widehat{\alpha\beta\tau})_{111} = -0.0841(0.2505),\ (\widehat{\alpha\beta\tau})_{211} = 1.2348(0.2283).$$

개체간의 변이에 따른 공분산 구조하의 요인 A의 두 수준 a_1과 a_2의 수준효과 간의 차는 $\widehat{\alpha_1 - \alpha_2} = -0.6451(0.1741)$로 구해진다. 모형내 포함된 모수들을 추론하기 위해 이용된 주변로짓들의 관측벡타를 $\mathbf{L}(\mathbf{p})$라 두면,

$$\begin{aligned}\boldsymbol{L}(\boldsymbol{p}) = [&0.5108, 0.7000, 1.4155, 1.2381, 0.9584, 0.7162, 0.6516, 0.4466,\\ &0.6737, 0.5318, 0.6376, 1.2981, 0.6308, 0.3643, 0.5898, 1.4385,\\ &0.1676, 0.0526, 0.0387, 0.1060, 1.2040, 0.9190, 0.1490, -0.1133]'\end{aligned}$$

표 **4.1.** 선거자료의 생성표

A											
(a_1)				(a_2)				(a_3)			
(B)				(B)				(B)			
(b_1)		(b_2)		(b_1)		(b_2)		(b_1)		(b_2)	
(T)		(T)		(T)		(T)		(T)		(T)	
1	2	도수	1 2	도수	1 2	도수	1 2	도수	1 2	도수	1 2	도수
m	m	215	m m	125	m m	95	m m	65	m m	105	m m	156
m	j	125	m j	112	m j	89	m j	86	m j	52	m j	37
m	k	20	m k	42	m k	20	m k	20	m k	18	m k	57
j	m	148	j m	95	j m	23	j m	14	j m	14	j m	27
j	j	200	j j	77	j j	137	j j	102	j j	95	j j	95
j	k	87	j k	47	j k	17	j k	15	j k	47	j k	66
k	m	123	k m	39	k m	22	k m	22	k m	39	k m	12
k	j	82	k j	221	k j	45	k j	48	k j	22	k j	18
k	k	11	k k	46	k k	37	k k	21	k k	87	k k	45

로 관측된다. 따라서, 모형에 대한 잔차의 자유도는 12이다. 세 요인의 교호작용이 유의수준 $\alpha = 0.01$에서 유의함을 나타내므로 이에 대한 추정값 및 추정오차가 주어져 있다. 주어진 반복측정 자료가 위에서와 같이 명목형의 설명변수들로 주어질 때, 가중최소제곱 방법이 유용함을 알 수 있다.

5. 결론

본 논문은 개체 또는 실험단위의 반응변수가 다가의 명목형 변수일 때, 관심반응범주들의 확률에 영향을 미치는 세 요인 중 두 요인이 개체단위에서 처치로 행해지고 시간요인은 반응의 반복측정 요인으로 소단위의 분할된 개체에서 행해지는 처치요인으로 가정하고 있다. 그리고 자료를 분석하기 위한 모형으로 기준범주를 이용한 일반화된 주변 로짓모형을 가정하고 모형구축과정을 논의하고 있다. 또한, 동일 개체의 반복측정에 따른 반응변수간의 종속성을 감안한 공분산구조로 분할구 실험계획에서 예상할 수 있는 공분산 구조와는 다른 구조임을 가정한다. 왜냐하면, 작은 실험단위에서의 시간요인의 수준들은 연구자가 확률화의 원리에 의한 임의배정을 할 수 없기 때문이다. 관측값들 간에 내재된 공분산 구조하에서 자료를 분석하는 방법으로 가중최소제곱법을 이용할 수 있음을 보여주고 있다.

참고문헌

최재성 (2004). A mixed model for ordered response categories, <한국데이터정보과학회지>, **15**, 339–345.

Agresti, A. (1990). *Categorical Data Analysis*, John Wiley & Sons, New York.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression,* (2nd edition), John Wiley & Sons, New York.

Koch, G. G. and Reinfurt, D. W. (1971). The analysis of categorical data from mixed models, *Biometrics*, **27**, 157–173

Koch, G. G., Landis, J. R., Freeman, J. L., Freeman, D. H. and Lehnen, R. G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data, *Biometrics*, **33**, 133–158.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika*, **73**, 13–22.

Mcfadden, D. (1974). Conditional logit analysis of qualitative choice behavior, pp. 105–142, In *Frontiers in Econometrics*, Editted by P. Zarembka, Academic Press, New York.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, (2nd edition), John Wiley & Sons, New York.

Milliken, A. G. and Johnson, E. D. (1984). *Analysis of Messy Data*, Van Nostrand Reinhold, New York.

A Generalized Marginal Logit Model for Repeated Polytomous Response Data

Jaesung Choi[1]

[1]Dept. of Statistics, Keimyung University

(Received April 2008; accepted June 2008)

Abstract

This paper discusses how to construct a generalized marginal logit model for analyzing repeated polytomous response data when some factors are applied to larger experimental units as treatments and time to a smaller experimental unit as a repeated measures factor. So, two different experimental sizes are considered. Weighted least squares(WLS) methods are used for estimating fixed effects in the suggested model.

Keywords: Generalized logits, fixed effects, polytomous data, repeated measures, weighted least squares.

[1]Professor, Dept. of Statistics, Keimyung University, 1000 Shindang-Dong, Dalseo-Gu, Daegu 704-701, Korea. E-mail: jschoi@kmu.ac.kr