Journal of Korean
Data & Information Science Society
2002, Vol. 13, No.2 pp. 129 138

Logistic regression model for major separation rate[1)]

Jaesung Choi[2)]

Abstract

This paper deals with logistic regression models for analysing separation rates from majors. The model building procedure shows how to incoporate the effects of some factors causing from three-way nested sampling scheme and discusses what type of characteristics as independent variables directly affecting the rates should be considered.

Keywords: binary data, binomial data, nested design, random effects

1. 서론

대학에 학부제가 도입된 이후 이전의 학과별 신입생 선발때와는 달리 학부내 소속 전공별 학생수가 해당전공의 정원보다 초과하거나 정원미달현상을 보이는 등 학생수가 일정하지 않게 되고 더욱이 학부간의 전부가 용이함에 따라 학부간의 인원변동도 심하여 전공 또는 학부운영에 많은 문제점과 함께 바람직하지 않은 여러 가지 현상들을 드러내고 있다.

본 논문은 학부제 실시 이후 발생하는 여러 가지 현상들중 신입생들이 애초에 지원한 소속학부에서 배정받은 전공을 이탈하는 학생들의 변동율, 즉 전공이탈율과 관련된 자료분석모형을 제시하고자 한다. 전공이탈율에 대한 모형의 고려는 학부제를 다년간 실시함에 따라 학생들이 배정받은 전공을 떠나게 되는 주된 요인들이 무엇인가를 분석하기 위한 기본모형의 필요성과 이 모형에 근거하여 소속전공을 이탈하는 학생들의 비율을 체계적으로 규명할 수 있을 때 학과나 학부의 운영에 도움이 될 수 있으리라 사료되기 때문이다.

---

1. The present research has been conducted by the attached research institute Research Grant of Keimyung University in 2001.

2. Professor, Department of Statistics, Keimung University. Taegu, 704-701, Korea. E-mail : jschoi@kmu.ac.kr

이탈율은 개체의 이가반응에 대한 성공확률로 간주할 수 있으므로 이 성공확률에 영향을 미치는 독립변수들과의 관계를 분석하기 위한 모형은 일반화된 선형모형으로 기술할 수 있다. 개체의 관심반응에 대한 확률을 분석하기 위하여 이용되는 일반화된 선형모형은 여러 문헌에서 제시되고 있다. Agresti(1990)는 개체의 이가반응변수들에 대한 다양한 모형들을 제시하고 모형들의 특성 및 모수들의 추론방법에 대해 자세히 설명하고 있다. Hosmer and Lemeshow(2000)는 개체의 이가반응변수에 대한 분석모형으로 다양한 성격의 독립변수들을 포함하고 있는 로지스틱 회귀모형을 논의하고 있다. 그리고 Cox and Snell(1989)은 개체의 반응이 구조적으로 지분관계인 이가반응변수들에 대해 로지스틱 회귀모형을 다루고 있다. 본 논문은 이러한 문헌들에 근거하여 관심자료를 분석하기 위한 체계적이고 합리적인 모형을 설정한 후 모형내 미지모수들을 구체적으로 추론해 보고자한다.

2. 자료구조

전공이탈율을 분석하기 위한 모형설정을 위하여 자료구조를 생각해 보기로 한다. 먼저, 조사모집단은 학부제를 도입한 대학에서 조사시점에 학부내 전공을 배정받은 학생들을 대상으로 한다. 이때, 전공을 배정받은 학생은 전공이탈여부에 대한 관측개체이고 개체의 반응변수는 전공이탈여부이며, 졸업시 까지 전공을 이탈하지 않으면 0 이탈하면 1의 값을 취하는 변수 Y로 나타낼 수 있다. 전공이탈율을 π로 두면 반응변수 Y는 성공확률이 π인 베르누이 분포를 나타낸다. 전공이탈율에 영향을 미치는 독립변수들로 성별, 전공배정 이전까지의 평균평점, 입학시 교차지원여부, 그리고 배정받은 전공에서의 적성여부등을 고려할 수 있겠다. 전공이탈율과 고려된 독립변수들 간의 관련성을 판단하기 위하여 일반화 선형모형을 이용할 때, 모형은 다음과 같이 표현된다.

$$g(\pi(\boldsymbol{x})) = g(E(Y)) = \alpha + \boldsymbol{\beta}'\boldsymbol{x} \tag{2.1}$$

단, g는 연결함수(link function)이고 $\boldsymbol{x}$는 독립변수들의 벡타이며, $\boldsymbol{\beta}$는 모수벡이다.

g가 logit 연결함수일 때 로지스틱 회귀모형으로 정의된다. 다른 연결함수를 이용하는 모형에 비해 이 모형의 이점은 표본추출계획이 전향적이거나 후향적이건 간에 모수효과들이 추정될 수 있다는 점이다. 또한, 이 모형은 수리적 관점에서 활용이 다양하며 쉽게 이용할 수 있고 의미있는 해석이 가능하므로 이가 또는 이항반응자료 분석에 주로 이용된다. 모형의 형태는 다음과 같다.

$$\log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \alpha + \boldsymbol{\beta}'\boldsymbol{x} \tag{2.2}$$

전공이탈율을 분석하기 위한 로지스틱 회귀모형에서 모형식을 추정하고 모수추론을 위해서는 자료들을 얻기 위한 표본추출계획을 고려해야 한다. 연구시점에서 각 대

학에서의 전공배정후 이탈율에 대한 자료를 얻기위한 표본추출방법으로 삼원지분계획을 생각할 수 있다. 즉, 대학집단에서 일부대학을 임의로 추출하여 표본으로 취하고 표본으로 추출된 대학내 학부집단에서 임의로 일부 학부를 추출하여 그 학부내 모든 전공 또는 일부 전공에서 학생들을 표본으로 취할 수 있다. 이와 같이 표본을 취하여 필요한 자료를 얻을 때, 고려해야 할 점은 이용된 표본추출계획에 의해 반응변수의 관심확률에 영향을 미치는 요인들이 있는 가를 살펴볼 필요가 있고 그러한 요인들이 존재한다면 이들 요인들의 효과가 포함되도록 모형을 설정해야 한다.

삼원지분계획하의 표본추출방법은 반응의 관심확률인 이탈율에 영향을 미칠 수 있는 세 가지 요인들이 존재하게 된다. 첫째는 대학들의 집단에서 표본을 추출함으로써 발생하게 되는 대학간의 변이 또는 변동은 이탈율의 한 변동요인이 될 수 있고 둘째는 추출된 대학에서 임의로 일부 학부를 추출할 때 학부간의 변동이 예상되고 이들 학부간의 변동은 이탈율에 영향을 미치는 또 하나의 요인으로 간주될 수 있으며 마지막으로 학부내 전공간의 변동 또한 영향을 미치는 요인일 수 있다. 대학간의 변동을 나타내는 요인의 효과 즉, 선택된 대학효과는 대학집단에서 임의로 추출한 데 따른 효과이므로 확률효과로 취급할 수 있다. 마찬가지로 대학내 학부효과, 학부내 전공효과도 확률효과로 간주하게 된다. 이들 세 확률효과를 고려한 모형은 지분관계의 확률효과를 포함하고 있는 로지스틱 회귀모형이 되게 된다. 이외에도 중요하게 다루어야 할 점은 각 전공에서의 이탈율 π에 대한 정의이다. 전공배정후 이탈여부에 대한 개체의 관측반응은 이탈하거나 존속하는 이가의 반응을 나타내며 각 개체의 이탈여부에 대한 반응은 독립이고 이탈확률은 일정하다는 가정이 만족되어야 한다는 점이다. 이러한 가정은 여러 유형의 비율, 예를들면, 제품의 불량률, 특정정당의 지지율, 그리고 사망률과 같은 관심비율을 분석하는 데 필요한 조건들이다. 이 두 가지 가정들은 이항자료의 분석을 위한 이항분포를 적용하기 위해 필수적인 가정들이다.

예를들어 비교하고자 하는 처치가 둘 있고, 두 처치집단에서의 성공비율을 π_1, π_2라 둘 때, 두 집단에서의 성공비율에 대한 처치효과를 알아 보기 위한 모형으로 로지스틱 회귀모형을 고려한다면 다음과 같다.

$$\log it(\pi) = \alpha + \beta z \tag{2.3}$$

단, z는 처치의 두 수준을 나타내는 가변수이다.

두 처치의 성공비율을 모형으로 표현할 때, 각 처치집단에서의 성공확률에 영향을 미치는 변수는 각 집단에서 행해진 처치의 고정효과를 가정한 모형이다. 이 예에서 논의의 초점은 처치를 관심집단의 특성으로 기술하고 있는 점에 주목할 필요가 있다.

본 논문에서의 전공이탈율에 대한 분석의 가정으로 각 전공에서 이탈율은 전공만을 고려할 때 전공들은 비교하고자하는 처치들로 고려되고 따라서 전공이탈율간의 차는 전공효과들을 감안한 모형으로부터 분석될 수 있음을 의미하며 전공내에서 이탈율은 일정하고 개체의 반응은 독립적으로 주어진다는 가정하에 자료분석에 대한 논의가 행해지고 있음을 내포하고 있다. 두번 째는 자료수집을 위한 표본추출방법에서 집락추출에 의한 삼원지분계획을 고려할 때 이탈율에 영향을 미칠 수 있는 세 변동요인들의 효과들이 확률효과로 간주되기 때문에 이들 효과에 대한 확률분포를 가정해야 한다. 확률효과에 대한 분포의 가정은 일반적으로 평균이 0 이고 분산이 일정한 상수분

산을 갖는 정규분포를 가정한다.

표본추출계획과 관련된 로지스틱 회귀모형을 구체적으로 기술하기 위해 자료의 구조를 살펴보기로 한다. 자료의 구조를 단순화하기 위해 임의로 추출된 대학이 3개, 대학내 임의로 추출된 학부가 2개이며 학부내 전공이 3개 전공이라 하자. 이 경우 자료구조는 <표2.1>과 같다. <표2.1>은 단순히 표본추출방법으로 삼원지분계획을 이용하였을 때 주어지는 자료들 만을 나타내고 있다.

그러나 관심확률에 영향을 미칠 수 있는 독립변수들을 고려할 때, 수집되는 자료의 구조는 달리 표현되리라는 것을 예상할 수 있다. 관심확률에 영향을 미칠 수 있는 독립변수들을 파악하기 위해서 우선 각 전공의 이탈확률에 변이를 줄 수 있는 변수들이 무엇인 가를 고려해야 한다. 독립변수들로 전공이탈율에 영향을 미칠 수 있는 변수들은 전공간에 변동이 예상되는 전공특성들을 생각할 수 있다. 예를들면, 각 전공에서의 교차지원한 학생들의 수, 그 전공의 난이도에 대한 측정으로 학과평균성적에 미달하는 학생들의 수, 취업률, 전공선택에 만족하지 못하고 있는 학생들의 수, 그리고 전공교수들의 관심도등을 생각할 수 있다.

<표2.1: 삼원지분계획하의 자료구조를 나타내는 표>

대학	학부	전공	이탈여부	
			이탈	존속
1	1	1	y_{111}	$n_{111} - y_{111}$
		2	y_{112}	$n_{112} - y_{112}$
		3	y_{113}	$n_{113} - y_{113}$
	2	1	y_{121}	$n_{121} - y_{121}$
		2	y_{122}	$n_{122} - y_{122}$
		3	y_{123}	$n_{123} - y_{123}$
⋮	⋮	⋮	⋮	⋮
3	1	1	y_{311}	$n_{311} - y_{311}$
		2	y_{312}	$n_{312} - y_{312}$
		3	y_{313}	$n_{313} - y_{313}$
	2	1	y_{321}	$n_{321} - y_{321}$
		2	y_{322}	$n_{322} - y_{322}$
		3	y_{323}	$n_{323} - y_{323}$

3. 자료분석 모형

전공이탈율에 대한 자료가 삼원지분계획하의 표본으로부터 수집된 자료일 때, 자

료분석을 위한 모형을 구체적으로 논의해 보기로 한다. 먼저, 대학집단에서 임의로 추출된 대학을 i (단, $i = 1, 2, \cdots, a$), 대학 i 에서 임의로 추출된 학부를 j (단, $j = 1, 2, \cdots, b$), 그리고 학부내 임의전공을 k (단, $k = 1, 2, \cdots, c$)라 두자. 독립변수들의 벡타를 x라 둘 때, 가정할 수 있는 모형의 형태는

$$\log it(\pi_{ijk}) = \alpha + \boldsymbol{\beta}'\boldsymbol{x} + a_i + b_{j(i)} + c_{k(ij)} \tag{3.1}$$

이다. 단, a_i 는 대학 i 의 확률효과, $b_{j(i)}$는 대학 i 에서 임의추출된 학부 j의 확률효과이며 $c_{k(ij)}$는 대학 i , 학부 j에서 임의추출된 전공 k의 확률효과를 나타낸다. 이들 확률효과에 대한 분포는 각기 $N(0, \sigma_a^2)$, $N(0, \sigma_b^2)$, 그리고 $N(0, \sigma_c^2)$인 분포를 따른다고 가정한다.

지분계획하의 확률효과들의 특성에 의해 지분관계에 있는 효과들은 교호작용이 존재하지 않기 때문에 이를 고려할 필요가 없다. 그러나, 지분관계의 효과들이 모두 확률변수로 간주될 수 없는 경우에 일부 효과는 고정효과로 취급하여 추론을 해야한다. 예를들면, 임의추출된 학부내 모든 전공이 선택되었다면 전공효과는 더 이상 확률효과가 아닌 고정효과로 취급된다. 이때의 관련된 로지스틱 회귀모형의 한 형태는

$$\log it(\pi_{ijk}) = \alpha + \boldsymbol{\beta}'\boldsymbol{x} + \boldsymbol{\gamma}'\boldsymbol{z} + a_i + b_{(ij)} \tag{3.2}$$

이다. 단, $\boldsymbol{z} = (z_1, z_2, \cdots, z_{(c_{(ij)} - 1)})'$는 대학 i, 학부 j내 소속전공들의 수가 $c_{(ij)}$개일 때, $(c_{(ij)} - 1)$개의 가변수들의 관측벡타를 나타낸다. 일반적으로 대학마다 학부내 소속전공의 수는 동일하지 않을 뿐만 아니라 대학간에 동일 명칭의 학부라 하더라도 다양한 전공들이 소속되어 있기 때문이다.

본 논문에서의 주안점은 학부제 실시이후 학부제를 시행하고 있는 대학에서 배정받은 전공을 이탈하여 동일학부내 다른 전공 또는 다른 학부로 이동하는 학생들의 비율, 즉 전공이탈율을 추론하는 데 있다. 다시말하면, 배정받은 전공에서 벗어나는 개별학생의 이탈가능성에 대한 관심보다는 배정받은 전공에서의 이탈율에 대한 분석을 하고자 한다. 이는 전공간에 이탈율의 비교에 대한 추론을 위해서 필요하다.

구체적으로 설명하기 위한 예로써 자연과학부내 통계학 전공에서 이탈율을 π라 두자. 이때, π에 대한 추론을 위해 통계학을 전공하는 학생들의 모집단으로부터 크기 n인 확률표본을 취한 후 표본내 이탈한 학생들의 수에 대한 확률분포로 이항분포를 정의할 수 있을 때, π에 대한 추론이 행해지게 된다. 그러나, 전공이탈율에 영향을 미치는 독립변수들이 존재할 때, 이들의 효과를 모형에 근거하여 추론하는 것은 모형의 구조형태의 결정과 함께 체계적이고 효율적인 분석방법의 토대가 되기 때문이다. 전공이탈율에 영향을 미치는 변수들로 양적인 변수들과 질적인 변수들의 두 유형을 생각할 수 있다. 이 경우 두 유형의 변수들은 집단화된 자료의 특성을 알아볼 수 있는 변수들이어야 함에 유의할 필요가 있다. 예를들면, 질적변수들로 소속학부, 소속대학

등은 전공에서의 이탈확률에 영향을 미치는 범주형 변수들이다. 양적변수들로는 교차지원한 학생수, 취업률, 그리고 전공에 부정적인 학생들의 수 등을 생각할 수 있다. 여기에서 개체별 특성을 고려하지 않은 이유로는 학생들의 개인별 전공이탈율에 관심을 두기 보다는 전공별 전공이탈율에 대한 분석에 관심을 두고 있다는 점이다. 따라서, 분석하고자 하는 관심확률이 어떤 유형의 확률인가에 따라 적절한 독립변수들을 고려해야 한다는 점에 유의해야 함을 나타내고 있다.

이러한 논의의 주안점은 개체에 대한 관측반응이 이가반응으로 주어지는 경우에 관심확률은 모집단의 전체적 특성을 나타내는 확률이거나 개체의 개별적 특성과 관련된 확률을 의미할 수 있다는 점이다. 예를들면, 어떤 회사에서 생산된 제품의 불량율, 어느 지역사회 집단에서의 실업률등은 그 회사 또는 그 집단에서의 개체들의 전체적 특성을 나타내는 모수이나, 어느 특정학생이 지원한 대학에 합격할 확률, 한 특정인이 신용불량자로 인정될 확률은 개별적 특성을 나타내는 모수이다. 따라서, 어떤 유형의 관심확률을 분석하느냐에 따라서 관련된 설명변수들의 성격도 명확히 정의되고 분리되어 고려되어야 한다는 것이다.

모집단의 전체적 특성을 나타내는 확률을 일반화 선형모형의 근거하에 분석하고자 하는 경우에 관련된 독립변수들은 확률표본을 취한 다음 수집된 이가자료를 집단화된 이가자료(grouped binary data) 또는 이항자료(binomial data)로 분류할 수 있는 변수들을 고려해야 함이 마땅하다. 본 논문의 전공이탈율도 전공집단의 전체적 특성과 관련된 모수로 정의하고 있다. 다시말하면, 학부제 시행이후 전공배정된 학생들이 전과 또는 전부의 시기를 통하여 비인기전공 또는 기초학문의 전공으로부터 소위 인기전공 또는 응용학문의 전공으로 옮겨가는 추세를 여러 가지 요인들로 진단해 보고 주된 요인들이 어떤 것인 가를 분석하기 위한 모형을 고려하고 있다.

4. 자료분석 예

다음은 어느 대학의 전공이탈율을 분석하기 위해 인위적으로 생성한 자료이다. 실제 자료의 수집은 대학마다 학생들의 개별 신상자료와 학적자료의 누출이 대학간의 교육정책과 운영상의 문제점을 야기할 수 있기 때문에 제공받기가 쉽지 않은 점이 있다. 그러나, 일반적인 사실에 근거한 생성자료의 분석예는 실제 자료를 어떻게 분석해야 하는 가에 대한 방법을 제시하므로 자료분석에 도움이 될 수 있다고 본다.

3절의 자료분석 모형에서는 학부제를 시행하고 있는 모든 대학의 전공집단에서 이탈율의 분석모형에 대한 논의를 했으나 이절에서는 한 특정대학의 일부 선정된 학부의 전공이탈율에 대한 모형을 생각해 보기로 한다. 자료의 구조가 <표4.1>과 같다고 하자. <표4.1>에서 고려되고 있는 학부는 인문분야의 기초학문을 수용하고 있는 인문

학부의 4개 전공과 자연과학부의 9개 전공에 대한 자료이다. 학부제 실시이후 학부내 각 전공은 학부제 실시이전의 전공정원을 갖고 있으나 전공지원인원에 비례해 인원이 배정되기 때문에 전공정원을 초과한 인원이 배정되기도 하고 모자라는 인원이 배정될 수도 있다. 따라서 표의 배정인원은 실제 그 전공에 배정받은 인원수를 나타내고 있다. 이 대학의 전공배정기준은 1학년말에 학생들의 평균평점을 기준으로 전공을 배정한다. 학생들은 학부내 원하는 전공을 3개까지 순위별로 지망하게 되고 지원자가 많은 전공은 전공정원보다 많은 120%까지 선발할 수 있다. 자료표에서 배정인원과 각 전공에서 이탈자의 수는 모대학의 2001년도에 조사된 실제자료를 나타내고 있다. 교차지원 인원과 전공이 적성에 맞지 않은 인원은 가상자료로 주어졌다. 전공정원은 학부제 실시이전 전공별 정원을 나타내고 있으나 1학년말 학생들이 받은 평균평점과 기타기준에 의해서 실제로 지원하는 전공에 배정된 인원을 살펴보면 일부전공만이 정원을 채우고 있음을 알 수 있다. 전공배정을 받은 학생들이 2학년 후학기에 실시된 전과 또는 전부시기를 통하여 이탈한 학생들은 기초학문을 다루고 있는 두 학부의 모든 전공에서 많은 학생들이 다른 학부로 옮겨 가고 있음을 보여 주고 있다. 이러한 현상은 학부제 실시 이후 기초학문을 다루는 모든 전공에서 학생유치의 어려움을 보여주는 한 단면이기도 하다. 기초학문 분야를 이탈하는 현상에 영향을 미치는 요인들로 대학진학 때의 희망전공과는 상관없이 성적 때문에 다른 계열에 지원한 가를 알아보기 위한 교차지원여부, 졸업후 취업가능성 그리고 전공적성등의 변수들을 고려할 수 있다. 여기서는 여러 가능한 변수들의 자료수집의 어려움 때문에 교차지원여부와 전공적성의 두 독립변수들과 관련된 자료분석을 생각해 보기로 한다.

<표4.1>의 자료를 분석하기 위한 로지스틱 회귀모형을 생각해 보자. 특정대학의 2개학부의 전공만을 고려하였기 때문에 학부효과와 전공효과들은 고정효과로 간주된다. 양적변수로 전공특성을 나타내는 두 변수 즉, 교차지원 인원 과 전공이 적성에 맞지 않는 인원을 고려하지 않을 때의 모형은 다음과 같다.

$$\log it(\pi_{ij}) = \mu + \alpha_i + \beta_{j(i)} \tag{4.1}$$

단, μ는 절편이고 α_i는 i번째 학부효과이고 $\beta_{j(i)}$는 학부 i내 j번째 전공효과를 나타낸다.

<표4.1: 모대학의 2001년도 전공별 이탈자의 수를 나타내는 자료>

학부	전공	전공정원	배정인원	교차지원 인원	전공이 적성에 맞지않는 인원	이탈자의 수	이탈자 추정인원
자연과학부	수학	50	50	4	15	12	14.92
	통계학	40	40	2	8	5	7.75
	화학	60	50	8	25	21	21.90
	물리학	40	11	1	7	5	2.22
	생물학	40	40	2	8	7	7.75
	미생물학	40	40	4	7	4	5.09
	공중보건학	50	45	6	14	11	9.23
	식품가공학	40	36	5	5	5	3.06
인문학부	신학	40	6	2	2	1	0.58
	사학	40	40	3	6	4	5.31
	철학	40	31	2	25	23	21.72
	국민윤리학	40	40	5	15	12	10.46

식(4.1) 모형은 관측치가 13개이고 모형내 모수의 수가 13이므로 포화모형(saturated model)을 나타낸다. $\beta_{j(i)}$ 대신 각 전공에서 측정된 두 개의 양적변수를 이용한 모형의 형태는 다음과 같이 표현할 수 있다.

$$\log it(\pi_{ij}) = \mu + \alpha_i + \beta_i x_{1j(i)} + \gamma_i x_{2j(i)} \tag{4.2}$$

단, β_i는 학부 i에서 양적변수 x_1과 관련된 회귀변수이고, γ_i는 x_2와 관련된 회귀변수이다. 여기서 x_1은 교차지원한 인원 수이고, x_2는 전공이 적성에 맞지 않는 인원을 나타낸다. 이경우에 추정해야 할 모수의 수는 네 개이다. $\beta_{j(i)}$의 고정효과 대신에 두 양적변수를 이용한 모형은 다음과 같이 단순한 형태로 나타낼 수 있다.

$$\log it(\pi_{ij}) = \mu + \alpha_i + \beta_1 x_{1j(i)} + \beta_2 x_{2j(i)} \tag{4.3}$$

단, β_1과 β_2는 두 양적변수와 관련된 회귀변수들이다.

식(4.3)을 위 자료에 적합시켰을 때 모형내 모수들의 추정값은 다음과 같다.

$\hat{\mu}$=-2.19(0.3196), $\widehat{\alpha_1}$=0.16(0.2865), $\widehat{\beta_1}$=-0.17(0.0727) 이고 $\widehat{\beta_2}$=0.13(0.0210) 이다.

우도비 검정통계량의 p-value는 0.3695이나, 학부효과가 0 이라는 귀무가설에 대한 p-value가 0.5745이므로 전공이탈율에 영향을 미치는 변수라고 생각할 수 없다. 따라서, 학부효과를 제외한 단순모형은 다음과 같이 주어진다.

$$\log it(\pi_{ij}) = \mu + \beta_1 x_{1j(i)} + \beta_2 x_{2j(i)} \tag{4.4}$$

식(4.4)를 자료에 적합시켰을 때 주어진 모수추정값은 다음과 같다.

$\hat{\mu}$=-2.13(0.2983), $\widehat{\beta_1}$=-0.18(0.0666) 이고 $\widehat{\beta_2}$=0.13(0.0200) 이다. 두 양적변수의 유의성에 대한 p-value는 0.01이하이고 모형에 대한 적합도 검정에서 우도비 검정통계량의 관측값에 대한 p-value는 0.4376으로 모형이 적합함을 나타내고 있다. 따라서, 자료분석에 적합한 모형으로 식(4.4)의 이용은 타당하다고 할 수 있겠다. <표4.1>의 자료에서 마지막열의 이탈자 추정인원은 식(4.4)를 자료에 적합시킨후 추정된 모형식으로 부터 구한 이탈자의 수에 대한 추정값을 나타낸 열이다.

5. 결론

본 논문은 학부제 실시이후 나타나는 여러 현상들중 전공이탈율과 관련된 자료분석 모형을 제시하고 있다. 자료분석을 위한 로지스틱 회귀모형과 관련하여 논의된 주안점들을 다음 몇가지로 요약해 볼 수 있겠다. 우선 전공이탈율은 개체의 이가반응변수인 전공이탈여부에 대한 이탈확률로 정의하고 이 확률은 개체간에 일정하고 독립적으로 관측된다는 가정하에 분석이 행해지고 있다는 점이다. 이 가정이 만족될 때, 표본으로 추출된 개체들의 이가반응에 대한 관측값들의 모임인 이가자료는 반응범주의 도수로 주어진 이항자료로 간주될 수 있다. 따라서, 독립변수들의 각 수준결합에서 수집된 이항자료의 확률분포로 이항분포의 적용이 가능하게 되고 최우추정법을 이용하여 적합된 모형내 모수추정값을 구할 수 있게 된다. 두번 째로 전공이탈율을 분석하기 위한 표본추출계획과 관련된 요인들이 관심확률에 영향을 미칠 수 있는 변수들 임을 논의하고 있다. 세번 째로 개체의 관심반응확률인 전공이탈율에 영향을 미치는 독립변수들은 개체간의 전공이탈율에 변이를 줄 수 있는 변수들이 고려되지 않아야 한다는 점에 유의해야 함을 논하고 있다.

참고문헌

1. Agresti, Alan. (1990). Categorical data analysis, John Wiley and Sons, Inc., New York.
2. Cox, D.R.. and Snell, E.J. (1989). Analysis of binary data (2nd edition), Chapman and Hall, London.
3. Hosmer, W. David, and Lemeshow, Stanley. (2000). Applied logistic regression (2nd edition), John Wiley and Sons, Inc., New York.

[2002년 8월 접수, 2002년 9월 채택]