로지스틱 회귀모형을 분석하기 위한 SPSS, SAS, STATA의 비교분석

김 순귀[1], 정 동빈[2]

요약

최근 여러 분야에서 로지스틱 회귀에 대한 필요성과 그 응용이 급증하면서 이를 분석하기 위한 통계패키지가 많이 개발되어 사용되고 있다. 이 논문에서는 자료의 유형에 따라 활용할 수 있는 여러 형태의 로지스틱 회귀모형을 간단히 살펴보고, SPSS, SAS, STATA, MINITAB과 같은 통계패키지를 사용하여 로지스틱 회귀모형에 적용할 때 각각 다룰 수 있는 범위와 그 특징에 대해 다룬다.

주요용어 : 공변량 패턴, 로지스틱 회귀모형, 조건 로지스틱 회귀모형, ROC 곡선

1. 서론

우리 주위에는 독립변수와 종속변수사이의 함수관계를 설명하려는 다양한 많은 문제들이 산적해 있으며, 이런 관계를 자료를 분석하여 알아낼 수 있다면 많은 값진 정보를 얻을 수 있을 것이다. 회귀분석은 종속변수와 하나 또는 그 이상의 독립변수들 사이의 함수관계를 설명하려는 통계적인 기법이지만, 종속변수의 척도가 연속형이 아니라 명목척도 또는 서열척도인 범주형으로 측정된 경우에는 적절하게 사용할 수 없게 된다. 이 경우 회귀분석과 같이 하나의 종속변수와 하나 이상의 독립변수 사이의 관계를 표현하기 위해, 가장 잘 적합되고 모수의 수를 절약한 모형을 찾는 것이 바로 로지스틱 회귀분석의 목표이다. 독립변수는 종종 공변량(covariate)이라고도 하며 회귀모형과의 유일한 차이점은 고려된 종속변수의 형태가 범주형이어야 한다는 것이다. 이러한 특성 때문에 데이터 마이닝의 판별분야에 자주 사용되는 기법이기도 하다.

실생활에서 로지스틱 회귀분석을 행할 수 있는 예를 들어보면 다음과 같다.

- 어떤 시민들은 선거에 참여하고 다른 사람들은 그렇지 않은가?
- 어떤 사람에게는 관상심장병이 생기고 다른 사람에게는 그렇지 아니한가?
- 어떤 사업은 성공하고 또 다른 사업은 실패하는가?
- 추석 때 고향 방문 시 고속도로와 국도 중 어느 곳을 선택하여야 할 것인가?
- 바둑에서 패를 써야 할 것인가 또는 말아야 할 것인가?

1) (210-702) 강원도 지변동 123번지 강릉대학교 정보통계학과 교수
2) (210-702) 강원도 지변동 123번지 강릉대학교 정보통계학과 조교수

본 논문에서는 종속변수가 명목척도로 측정되었을 때, 수준의 수가 두 개인 경우(이분형 로지스틱 회귀모형), 세 개 이상인 경우(다항 로지스틱 회귀모형)와 종속변수가 순위척도로 측정되었을 때(순서형 로지스틱 회귀모형)로 구분하여 이에 관련된 여러 통계프로그램들을 살펴보고 그 특성을 비교해 보고자 한다. 또한 실제 응용에서 많이 다루는 조건 로지스틱 회귀모형을 추가시키고자 한다.

그 동안 통계이론을 통계패키지에 연관시켜 패키지의 선택과 활용에 관한 비교 연구가 발표되었다. 이에 관련된 논문으로 EDA기능에 관한 패키지 비교 연구(허명회, 장진환, 1990), 시계열 분석에 관한 패키지 비교 연구(김수화, 김승희, 조신섭, 1994), 공정관리를 위한 통계패키지의 비교에 관한 연구(조신섭, 신봉섭, 1997), 반복측정 자료를 분석하기 위한 통계패키지의 고찰(최은숙, 박태성, 문경미, 1998) 등이 있다. 참고로 로지스틱 회귀분석에 관한 저서로는 김순귀 외(2002), 성웅현(2001)등이 있으며 회귀분석에 관한 여러 저서에서 부분적으로 다루어지고 있다고 생각한다.

본 논문의 구성은 다음과 같다. 2절에서는 앞에서 제시한 여러 로지스틱 회귀모형들(이분형, 다분형, 순서형)에 대해 간단히 소개하고, 3절에서는 여러 통계패키지들 중에서 SAS, SPSS, STATA, MINITAB을 중심으로 로지스틱 회귀분석을 처리할 수 있는 한계와 그 특성에 대해 살펴본다. 마지막으로 실제적인 상황에 필요한 추가기능과 향후 개선여지에 대해 토론하고자 한다.

2. 로지스틱 회귀분석을 위한 통계모형

2.1 이분형 로지스틱 회귀모형

종속변수 Y의 수준이 0, 1인 두 개인 경우 이분형 로지스틱 회귀모형을 사용함은 이미 앞 절에서 언급한 바가 있다. 표기법을 단순화하기 위해 $\pi(x) = E(Y|x)$로 나타내면, 공변량이 단지 하나인 경우 이분형 로지스틱 회귀모형은 다음과 같이 표현할 수 있다.

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{2.1}$$

또는

$$\pi(x) = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 x)\}}. \tag{2.2}$$

여기에서 β_0와 β_1은 추정될 모수이고, x는 공변량을 나타낸다.

로지스틱 회귀모형의 중요성은 $\pi(x)$의 로짓변환(logit transformation)에 있다고 할 수 있겠다. 로짓변환 g를 다음과 같이 정의하여 보자.

$$g(x) = \ln\left[\frac{\pi(x)}{1 - \pi(x)}\right] = \beta_0 + \beta_1 x. \tag{2.3}$$

여기에서 ln은 자연대수를 나타낸다. 로짓변환을 하게 되면 기존의 회귀모형이 가지고 있는 몇 가지 성질을 그대로 갖게 된다. 또한 x에 대한 종속변수 y의 관계식을 $y = \pi(x) + \varepsilon$으로 표

현할 수 있으며, ε은 y의 값에 따라 평균이 0, 분산이 $\pi(x)[1-\pi(x)]$인 이항분포를 따르게 된다. 오차항이 이항분포를 할 때 로지스틱 회귀모형의 모수 추정방법으로는 최대가능도법(method of maximum likelihood)을 사용하게 된다.

모형(2.1)을 확장시켜 p개의 독립변수가 존재할 때, 다음과 같은 다중(multiple) 로지스틱 회귀모형을 고려하여 보자.

$$\Pr(y=1|x_1,\cdots,x_p) = \pi(x_1,\cdots,x_p)$$

$$= \frac{\exp(\beta_0+\beta_1 x+\cdots+\beta_p x_p)}{1+\exp(\beta_0+\beta_1 x+\cdots+\beta_p x_p)}. \tag{2.4}$$

여기에서 $\beta_0,\beta_1,\cdots,\beta_p$는 추정할 모수들이다.

다중 로지스틱 회귀모형의 로짓은

$$g(x_1,\cdots,x_p) = \beta_0+\beta_1 x_1+\cdots+\beta_p x_p$$

가 된다. 만일 독립변수들 중 명목척도로 측정된 변수(예 : 인종, 성별 등)가 포함되어 있으면 회귀분석에서와 같이 가변수(dummy variable)로 취급하여 다룬다. 이 모형은 대부분의 통계패키이지에서 로지스틱 모형을 다루는 프로그램을 이용하여 분석할 수 있다. 이 모형에 관한 자세한 설명은 Hosmer & Lemeshow(2000) 또는 Kleinbaum(1994)을 참조하기 바란다.

2.2 다항 로지스틱 회귀모형

다항 로지스틱 회귀모형이란 서론에서 언급하였듯이 종속변수(Y)가 세 수준이상의 명목척도로 측정되었을 사용하는 모형이다. 간단한 예로, 종속변수(Y)의 범주가 0, 1, 2 형태의 명목척도로 측정되었다고 하자. 이분형 로지스틱 회귀모형에서는 $y=1$대 $y=0$의 로짓함수의 형태로 모수화시켰지만, 종속변수의 범주가 세 개인 로지스틱 모형에서는 두 개의 로짓함수를 필요로 한다. 편의상 $y=0$을 기준범주로 하여, $y=1$과 $y=2$인 범주를 $y=0$인 범주와 각각 비교하기 위한 로짓함수를 사용한다.

상수항을 포함한 공변량을 벡터 $\boldsymbol{x}_{(p+1)\times 1}=(1,x_1,x_2,\cdots,x_p)'$라 할 때, 다음과 같이 두 개의 로짓함수로 나타낼 수 있다.

$$\begin{aligned} g_1(\boldsymbol{x}) &= \ln\left[\frac{P(Y=1\mid \boldsymbol{x})}{P(Y=0\mid \boldsymbol{x})}\right] \\ &= \beta_{10}+\beta_{11}x_1+\beta_{12}x_2+\cdots+\beta_{1p}x_p \\ &= \boldsymbol{x}'\boldsymbol{\beta}_1 \\ g_2(\boldsymbol{x}) &= \ln\left[\frac{P(Y=2\mid \boldsymbol{x})}{P(Y=0\mid \boldsymbol{x})}\right] \\ &= \beta_{20}+\beta_{21}x_1+\beta_{22}x_2+\cdots+\beta_{2p}x_p \\ &= \boldsymbol{x}'\boldsymbol{\beta}_2 \end{aligned}$$

여기에서, $\boldsymbol{\beta}_1' = (\beta_{10}, \beta_{11}, \cdots, \beta_{1p})$이고, $\boldsymbol{\beta}_2' = (\beta_{20}, \beta_{21}, \cdots, \beta_{2p})$이다.

공변량 벡터($\boldsymbol{x}$)가 주어진 상황에서 각 결과범주의 조건적 확률은 다음과 같다.

$$\pi_0(\boldsymbol{x}) = P(Y=0 \mid \boldsymbol{x}) = \frac{1}{1+e^{g_1(\boldsymbol{x})}+e^{g_2(\boldsymbol{x})}} \tag{2.5}$$

$$\pi_1(\boldsymbol{x}) = P(Y=1 \mid \boldsymbol{x}) = \frac{e^{g_1(\boldsymbol{x})}}{1+e^{g_1(\boldsymbol{x})}+e^{g_2(\boldsymbol{x})}} \tag{2.6}$$

$$\pi_2(\boldsymbol{x}) = P(Y=2 \mid \boldsymbol{x}) = \frac{e^{g_2(\boldsymbol{x})}}{1+e^{g_1(\boldsymbol{x})}+e^{g_2(\boldsymbol{x})}}. \tag{2.7}$$

모수의 추정방법으로는 이분형 로지스틱 회귀모형과 동일하게 최대가능도법을 사용하게 되며, 변수에 대한 계수들의 유의성을 검정하는 가능도비 검정에 대한 자유도는 (종속변수의 수준수-1) ×(공변량의 개수)가 됨에 유의하라.

2.3 순서 로지스틱 회귀모형

지금까지 이분형과 다항 로지스틱 회귀모형에 관하여 다루었다. 여기에서는 종속변수의 수준수가 세 개 이상인 순서형일 때 사용하는 모형을 살펴보려고 한다. 순서형 종속변수를 다루는 모형의 하나인 비례승산모형(proportional odds model)에 관하여 알아보자.

이 모형은 $y \le k$과 $y > k$의 확률을 비교하며, 다음과 같이 표현한다.

$$\begin{aligned} c_k(\boldsymbol{x}) &= \ln\left[\frac{P(y \le k \mid \boldsymbol{x})}{P(y > k \mid \boldsymbol{x})}\right] \\ &= \ln\left[\frac{\pi_1(\boldsymbol{x}) + \pi_2(\boldsymbol{x}) + \cdots + \pi_k(\boldsymbol{x})}{\pi_{k+1}(\boldsymbol{x}) + \pi_{k+2}(\boldsymbol{x}) + \cdots + \pi_K(\boldsymbol{x})}\right] \\ &= \alpha_k + \boldsymbol{x}'\boldsymbol{\beta}, \quad k=1,2,\cdots,K. \end{aligned} \tag{2.8}$$

여기에서, K는 종속변수의 수준수를, $c_k(\boldsymbol{x})$는 p개의 공변량 벡터 $\boldsymbol{x}$가 주어졌을 때 종속변수 y가 k이하일 로짓을, α_k는 k번째 범주에 대한 절편항을, $\boldsymbol{\beta}$는 공변량 $\boldsymbol{x}$의 회귀계수를 각각 나타낸다.

식(2.8)로부터 종속변수(Y)가 특정한 순서범주를 취할 확률을 다음과 같이 구할 수 있다.

$$\pi_1(\boldsymbol{x}) = \Pr(y=1 \mid \boldsymbol{x}) = \frac{\exp(\alpha_1 + \beta'\boldsymbol{x})}{1 + \exp(\alpha_1 + \beta'\boldsymbol{x})}$$

$$\pi_2(\boldsymbol{x}) = \Pr(y=2|\boldsymbol{x}) = \frac{\exp(\alpha_2+\beta'\boldsymbol{x})}{1+\exp(\alpha_2+\beta'\boldsymbol{x})} - \frac{\exp(\alpha_1+\beta'\boldsymbol{x})}{1+\exp(\alpha_1+\beta'\boldsymbol{x})}$$

$$\cdots$$

$$\pi_{k-1}(\boldsymbol{x}) = \Pr(y=k-1|\boldsymbol{x}) = \frac{\exp(\alpha_{k-1}+\beta'\boldsymbol{x})}{1+\exp(\alpha_{k-1}+\beta'\boldsymbol{x})} - \frac{\exp(\alpha_{k-2}+\beta'\boldsymbol{x})}{1+\exp(\alpha_{k-2}+\beta'\boldsymbol{x})}$$

$$\pi_k(\boldsymbol{x}) = 1-\pi_0(\boldsymbol{x})-\cdots-\pi_{k-1}(\boldsymbol{x}).$$

모수의 추정방법으로는 이항 로지스틱 회귀모형(또는 다항 로지스틱 회귀모형)과 동일하게 최대가능도법을 사용하게 되며, 공변량에 대한 계수들의 유의성을 검정하는 가능도비 검정에 대한 자유도는 p임에 유의하라. 여기에서 k는 종속변수 y가 취하는 수준수를, p는 공변량의 개수를 각각 나타낸다. 순서형 종속변수를 다루는 모형에서 비례승산모형 이외의 다른 형태의 모형에 관한 연구는 Hosmer와 Lemeshow(2000)를 참조하기 바란다.

2.4 조건 로지스틱 회귀모형

대응(matched)이란 두 개 혹은 그 이상의 그룹을 비교하려는 연구의 계획단계(design stage)에서 수행하는 절차이다. 각 개체들을 반응변수(the outcome variable)와 관련이 있다고 예상하는 변수들에 근거하여 대응시킨다. 대응된 각 층 마다 사례 ($y=1$)와 대조 ($y=0$)의 표본을 선택한다. 대부분의 대응 계획은 한 사례(case)에 1-5개의 대조(control)를 대응시키게 된다. 이런 경우를 1-M 대응연구라 한다. 예를 들어, 1-1 대응계획은 각 층 마다 사례와 대조가 각각 한 개로 두 개의 개체를 가지게 된다. 각 사례-대조 쌍에 대하여 다른 가능한 위험요인(risk factor 또는 unmatched variables)에 관한 정보를 이용하게 된다.

일반적으로 이분형 로지스틱 회귀모형이나 다항 로지스틱 회귀모형에서는 모수의 추정값을 구하기 위해서는 조건없는 최대가능도법을 사용하였지만, 추정하여야 할 모수의 수가 많고 대응된 자료를 분석하기 위해서는 조건(conditional) 최대가능도법을 이용하여 모수들을 추정하게 된다.

1-1 대응계획으로 만든 k개의 층이 있다고 하자. 이때 위험요인으로 두 개의 요소가 있다. 즉, 대응된 변수와 관련된 위험요인과 대응되지 않은 변수에 관련된 위험요인이다. k번째 층에서 한 개체가 사건을 경험할 확률은

$$\Pr(y=1|\boldsymbol{x}) = \pi_k(\boldsymbol{x}) = \frac{e^{\alpha_k+\sum\beta_i x_i}}{1+e^{\alpha_k+\sum\beta_i x_i}} \tag{2.9}$$

이다. 여기에서 α_k는 대응변수값에 근거한 k번째 층의 효과이며, β_i는 대응되지 않은 공변량 x_i의 회귀계수를 각각 나타낸다.

참고문헌

(1) Cohen, S. B. (1997), "An Evaluation of Alternative PC-Based Software Packages Developed for the Analysis of Complex Survey Data", *The American Statistician*, Vol. 51, No. 3, 285-292.

(2) David J. Vining and Gregory W. Gladish (1992), "Receiver Operating Characteristic Curves: A Basic Understanding", *RadioGraphics* 12 : 1147-1154.

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

(4) Kim, S. K, D. B. Jeong and K. S. Bang (2001), "Use of Stata(2)-an application to Logistic Regression Model and Graphical Techniques-", *The Journal of Natural Science Research Institute* 17 : 109-117.

(5) Kleinbaum, David G. (1994), *LOGISTIC REGRESSION : A Self-Learning Text*, Springer.

(6) Mittlbock, M. and M. Schemper (1999), "Computing measures of explained variation for logistic regression model, Computer Methods and Programs in Biomedicine, Vol. 58, 17-24.

(7) Stata Corporation (2001), *Getting Started with Stata for Windows*.