Journal of Korean
Data & Information Science Society
2004, Vol. 15, No. 2, pp. 367~378

Comparative Study on Statistical Packages for Analyzing Logistic Regression
- MINITAB, SAS, SPSS, STATA -

Soon-kwi Kim[1)] · Dong-Bin Jeong[2)] · Young-Sool Park[3)]

Abstract

Recently logistic regression is popular in a variety of fields so that a number of statistical packages are developed for analyzing the logistic regression. This paper briefly considers the several types of logistic regression models used depending on different types of data. In addition, when four statistical packages (MINTAB, SAS, SPSS and STATA) are used to apply logistic regression models to the real fields respectively, their scope and characteristics are investigated.

keywords : Covariate pattern, Logistic regression model, Method of maximum likelihood, Proportional odds model, ROC curve

1. 서론

우리 주위에는 독립변수와 종속변수사이의 함수관계를 이용하여 설명할 수 있는 다양한 많은 문제들이 산적해 있으며, 이런 관계를 자료를 통하여 정확히 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 모형

1) First Author : Professor, Department of Information Statistics, Kangnung National University, Kangnung, 210-702, Korea,
E-mail: skkim@kangnung.ac.kr
2) Associate Professor, Department of Information Statistics, Kangnung National University, Kangnung, 210-702, Korea.
3) Professor, Division of Business Administration, Kwandong University, Kangnung, 210-701, Korea,

을 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?

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 통계패키지들을 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 EDA 기능에 관한 패키지 비교연구(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), 다변량 Q-기법의 사용을 위한 통계패키지의 비교연구(조영석, 문희정, 2003) 등이 있다. 참고로 로지스틱 회귀분석에 관한 저서로는 김순귀, 정동빈, 박영술(2003), 성웅현(2001) 등이 있으며 회귀분석에 관한 여러 저서에서 부분적으로 다루어지고 있다고 생각한다.

본 논문의 구성은 다음과 같다. 2절에서는 앞에서 제시한 여러 로지스틱 회귀모형(이분형, 다분형, 순서형)에 대해 간단히 소개하고, 3절에서는 여러 통계패키지들 중에서 SAS, SPSS, STATA, MINITAB을 중심으로 로지스틱 회귀분석을 행할 수 기능에 대해 각각 정리하고, 4절에서는 네 가지 통계패키지를 로지스틱 회귀분석에 적용할 경우 그 한계와 그 특성을 비교해 보았고, 4종류의 통계패키지에 대한 실제 예제를 참고할 수 있도록 간략히 정리하였다. 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 회귀모형(binary logistic regression model)을 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

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

여기에서 β_0와 β_1은 추정될 모수이고, x는 공변량을 나타낸다.

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

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

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 표현할 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.

$$\begin{aligned}\Pr(y=1|x_1,\cdots,x_p) &= \pi(x_1,\cdots,x_p)\\ &= \frac{\exp(\beta_0 + \beta_1 x + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x + \cdots + \beta_p x_p)}\end{aligned}$$

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 Hosmer와 Lemeshow(2000) 또는 Kleinbaum (1994)을 참조하기 바란다.

2.2 다항 로지스틱 회귀모형

다항 로지스틱 회귀모형(multinomial logistic regression model)이란 서론에서 언급하였듯이 종속변수가 세 수준이상의 명목척도로 측정되었을 때 사용하는 모형이다. 간단한 예로, 종속변수(Y)의 범주가 0, 1, 2 형태의 명목척도로 측정되었다고 하자. 이분형 로지스틱 회귀모형에서는 y=1 대 y=0의 로짓함수의 형태로 모수화시켰지만, 종속변수의 범주가 세 개인 로지스틱 모형에서는 두 개의 로짓함수를 필요로 한다. 편의상 y=0을 기준범주로 하여, y=1과 y=2인 범주를 y=0인 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.

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

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(x)가 주어진 상황에서 각 결과범주의 조건적 확률은 다음과 같다.

$$\begin{aligned} \pi_0(\boldsymbol{x}) &= \Pr(Y=0 \mid \boldsymbol{x}) = \frac{1}{1+\exp\left(g_1(\boldsymbol{x})\right)+\exp\left(g_2(\boldsymbol{x})\right)} \\ \pi_1(\boldsymbol{x}) &= \Pr(Y=1 \mid \boldsymbol{x}) = \frac{\exp\left(g_1(\boldsymbol{x})\right)}{1+\exp\left(g_1(\boldsymbol{x})\right)+\exp\left(g_2(\boldsymbol{x})\right)} \\ \pi_2(\boldsymbol{x}) &= \Pr(Y=2 \mid \boldsymbol{x}) = \frac{\exp\left(g_2(\boldsymbol{x})\right)}{1+\exp\left(g_1(\boldsymbol{x})\right)+\exp\left(g_2(\boldsymbol{x})\right)} \end{aligned}$$

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.

$$\begin{aligned} c_k(\boldsymbol{x}) &= \ln\left[\frac{\mathrm{P}(y \le k|\boldsymbol{x})}{\mathrm{P}(y > k|\boldsymbol{x})}\right] \\ &= \ln\left[\frac{\pi_1(\boldsymbol{x}) + \pi_2(\boldsymbol{x}) + \cdots + \pi_k(\boldsymbol{x})}{\pi_{k+1}(\boldsymbol{x}) + \pi_{k+2}(\boldsymbol{x}) + \cdots + \pi_K(\boldsymbol{x})}\right] \\ &= \alpha_k + \boldsymbol{x}'\boldsymbol{\beta}, \quad k = 1, 2, \cdots, K \end{aligned} \tag{2.2}$$

여기에서, K는 종속변수의 수준수를, $c_k(\boldsymbol{x})$는 $(p+1)\times 1$의 공변량 벡터 $\boldsymbol{x}$가 주어졌을 때 종속변수 y가 k이하일 로짓을, α_k는 k번째 로짓에 대한 절편항을, $\boldsymbol{\beta}$는 공변량 $\boldsymbol{x}$의 회귀계수를 각각 나타낸다.

식 (2.2)로부터 종속변수(Y)가 특정한 순서범주를 취할 확률을 다음과 같이 구할 수 있다.

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

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

$$\cdots$$

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

$$\pi_K(\boldsymbol{x}) = 1 - \pi_0(\boldsymbol{x}) - \cdots - \pi_{K-1}(\boldsymbol{x})$$

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 반응변수와 관련이 있다고 예상하는 변수들에 근거하여 대응시킨다. 대응된 각 층 마다 사례 (y=1)와 대조 (y=0)의

표본을 선택한다. 대부분의 대응계획은 한 사례(case)에 1-5개의 대조(control)를 대응시키게 된다. 이런 경우를 1-M 대응연구라 한다. 예를 들어, 1-1 대응계획은 각 층마다 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

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

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

3. 로지스틱 회귀분석을 위한 통계 프로그램의 고찰

앞 절에서는 범주형 종속변수를 위한 여러 형태의 로지스틱 회귀모형들을 정리하였다. 본 절에서는 통계 패키이지 SAS 8.1, SPSS 10.0, STATA 8.0, MINITAB 13.1을 중심으로, 로지스틱 회귀모형을 다룰 수 있는 프로그램들을 간략히 소개하고, 그 특징을 열거해 보도록 한다.

3.1 SAS

SAS에서 종속변수가 범주형인 자료를 처리할 수 있는 통계 프로그램은 Proc Logistic, Proc Probit, Proc Catmod 등이 있다.

3.1.1 Proc Logistic

Proc Logistic은 로지스틱 회귀모형의 분석을 위한 가장 기본적이고 효율적인 프로시저이다. 종속변수 y가 단지 두 개의 값을 가진 이분형 로지스틱 회귀모형을 적합시킬 수 있을 뿐 아니라, 공변량들의 값이 같은(공변량 패턴이라고 함) 개체가 여러 개 있는 경우 각 공변량 패턴 내의 총 개체수와 그 패턴내의 사건 발생수로 표기된 자료를 분석할 수 있는 기능이 있다(events/trial model 이라고 함). 더 나아가 순서형 로지스틱 회귀모형을 적합시키는 기능이 있지만, 다항 로지스틱 회귀모형의 분석은 불가능하므로 Proc Logistic 보다는 Proc Catmod에서 다루어야 한다.

3.1.2 Proc Probit

Proc Probit은 프로빗모형뿐 아니라 이분형 로지스틱 회귀모형, 순서형 로지스틱 회귀모형 등을 분석하는 프로시저이다.

3.1.3 Proc Catmod

Proc Catmod는 범주형 자료를 분석하기 위한 표준적이고 더욱 일반적인 프로시저로 Proc Logistic 프로시저보다 더 많은 기능을 가진 프로그램이다. 선형모형, 로그선형모형, 반복측정자료뿐 아니라 로지스틱 회귀모형과 조건 로지스틱 회귀모형을 적합시키고 분석할 수 있는 기능을 가진 프로시저이다.

3.2 SPSS

이분형 로지스틱 회귀모형을 분석(logistic regression 프로시저 사용)할 뿐 아니라, ROC (Receiver Operating Characteristic) 곡선을 그려주고 변수를 선택할 수 있는 기능(roc 프로시저 사용)이 있다. 다항 로지스틱 회귀모형을 적합(nomreg 프로시저 사용)시킬 수 있는 기능이 있지만, 특이한 점은 잔차, 영향력 통계량 등의 값을 산출할 때 이분형 로지스틱 회귀분석의 명령을 사용할 시에는 각 개체에 대해서, 다항 로지스틱 회귀분석의 명령을 사용할 시에는 각 공변량 패턴에 대하여 계산하는 기능을 가지고 있다. 순서형 로지스틱 회귀모형(plum 프로시저 사용)뿐 아니라, SAS와 마찬가지로 자료의 변환을 통하여 조건 로지스틱 회귀모형을 분석할 수 있는 기능이 있다.

SPSS에서는 이런 일련의 작업들을 메뉴방식을 통하여 손쉽게 처리할 수 있도록 완전 메뉴화되어 있는 점이 그 특징이라고 할 수 있다.

3.3 STATA

STATA에서 logit명령은 종속변수가 이분형인 이분형 로지스틱 회귀모형을 분석하여 줄 뿐 아니라 변수선택의 기능이 있다. 또한 ROC 곡선을 그려주고 모형을 판별하기 위한 통계량인 그 곡선 아래의 면적을 산출하는 기능도 있다. SAS와 마찬가지로 각 공변량 패턴내의 총 개체수와 사건 발생수로 표기된 자료를 분석할 수 있다.

mlogit 명령은 다항 로지스틱 모형을 분석하여 준다. 이때 종속변수 y가 가지는 제일 작은 값을 기준범주로 택하게 되지만, 자료의 변환없이 기준범주를 다른 값으로 바꿀 수 있는 옵션을 가지고 있다.

ologit 문은 순서형 로지스틱 회귀모형을 처리 분석할 수 있는 명령이며, 자료를 변환하지 않고 조건 로지스틱 회귀모형을 직접 적합시킬 수 있는 clogit 명령이 있다.

3.4 MINITAB

MINITAB 역시 하나 또는 그 이상의 설명변수와 범주형인 반응변수간의 관계를 평가하여 사용할 수 있는 세 가지 기본 로지스틱 회귀 프로시저가 있으며, SPSS와 같이 메뉴방식을 통해 사용자가 쉽게 사용하도록 했다. 세 가지 로지스틱 회귀모형의 모수들을 추정하기 위해 반복적인 가중 최소제곱 알고리즘을 적용한 최대가능도추정

법을 사용한다.

blogistic 프로시저는 이분형 반응변수에 대한 로지스틱 회귀를, nlogistic 프로시저는 명목형 반응변수에 대한 로지스틱 회귀를, ologistic 프로시저는 순서형 반응변수에 대한 로지스틱 회귀분석을 수행한다. 평행회귀선이 가정되기 때문에 각 공변량에 대해서 단일영역이 계산된다. 만일 이 가정이 성립하지 않을 경우 개별적인 로짓함수를 생성시켜 분석하는 것이 타당하다.

4. 통계패키지간의 비교 및 실례

3절에서 주로 많이 사용하는 네 개의 통계패키지를 여러 형태의 로지스틱회귀모형에 적합할 때 사용하는 각 프로시저에 대해 간략하게 살펴보았고, 4종류의 통계패키지를 실제적인 자료에 적용하여, 비교 분석하는 결과를 제시해 주었다.

<표 1> 로지스틱 회귀분석을 할 수 있는 여러 통계 프로시저

Package / Model	MINITAB	SAS	SPSS	STATA
Binary(multiple) logistic	BLogistic	proc logistic proc catmod proc probit	logistic regression NOMREG	.logit .logistic .mlogit
Grouped Data (events/trial)	NLogistic OLogistic	proc logistic proc probit proc catmod (need data transformation)	logistic regression (need data transformation)	.blogit .glogit
Multinomial logistic	NLogistic	proc catmod proc probit	NOMREG	.mlogit .mlogistic
Ordinal logistic	OLogistic	proc logistic	PLUM	.ologit
Conditional logistic	need data transformation	proc phreg proc logistic (need data transformation)	need data transformation	.clogit

<표 1>은 3절에서 다룬 세 가지 로지스틱회귀모형과 집단화된 자료를 다룰 수 있는 프로시저들을 통계패키지 별로 분류하여 정리한 것이다.

<표 2>에서는 각 통계패키지 내의 프로시저 별로 분석할 수 있는 몇 가지 기능에 대해 비교하여 요약하였다. 첫 번째 행은 이항 로지스틱 회귀모형을 적합시킬 때 ROC 곡선과 곡선이하의 영역을 시각적으로 나타내 주는 프로시저들을 각 패키지 별로 정리한 것이다. MINITAB을 제외한 세 개의 패키지는 이를 수행할 수 있음을 알

수 있다.

<표 2> 각 패키지의 특성 비교

Package / Model	MINITAB	SAS	SPSS	STATA
for Binary logistic				
ROC curve	N/A	proc plot	roc	.roc
Area under the curve	N/A	proc logistic	roc	.roc
for individuals				
predicted value	BLogistic	proc logistic	logistic reg., PLUM	N/A
Residuals	BLogistic	("output" option)	logistic reg., PLUM	N/A
for covariate pattern				
predicted value	N/A	N/A	NOMREG	.predict
Residuals	N/A	N/A	N/A	.predict
Hosmer & Lemeshow	BLogistic	proc logistic	PLUM	.ologit
Variables Selection	N/A	proc phreg proc logistic	logistic regression	.sw logit .sw logit .sw logit

* N/A: Not Available

두 번째 행은 각 개체와 공변량에 따라 예측값과 잔차를 산출하는 프로시저이다. STATA를 제외한 나머지 통계패키지들은 각 개체에 근거하여 두 개의 통계량을 산출하고, 추가로 SPSS는 nomreg 프로시저를 사용하여 공변량 패턴에 대한 예측값을 생성할 수 있음을 알 수 있다. 세 번 째 행에 나타낸 Hosmer와 Lemeshow 통계량은 모형의 적합도를 검정하는 대표적인 측도로 네 개의 패키지 모두 이에 해당하는 프로시저를 제공하고 있다. 참고로 주어진 자료에 대해 계산된 Hosmer와 Lemeshow 통계량 값은 서로 다른 값을 산출한다. 이는 각 통계 패키지마다 개체들을 각각 다르게 집단을 형성하기 때문이다. 마지막 네 번째 행은 변수선택에 관한 프로시저를 소개한 것으로 MINITAB을 제외한 나머지 패키지들은 이를 행할 수 있다.

<표 3>에서는 각 패키지별로 종속변수의 기준그룹(비교그룹)을 디폴트로 설정해 주는 값을 보여주고 있을 뿐 아니라, 그 기준그룹을 임의로 설정해 주는 기능이 있는지를 보여주고 있다. 주어진 기준그룹은 결과분석에 중요한 역할을 한다. STATA는 가장 적은 값이 기준그룹이 되고, STATA를 제외한 모든 패키지의 기준그룹은 가장 큰 값이 기준그룹이 된다.

<표 3> 각 패키지의 기준그룹 기능 비교

Package / Model	MINITAB	SAS	SPSS	STATA
Binary(multiple) logistic	BLogistic - OK ("reference" option) - default : hightest	proc logistic - OK ("descending" option) - default : hightest	logistic regression - N/A - default : hightest	.logit .logistic - N/A - default : lowest
Multinomial logistic	NLogistic - OK ("reference" option) - default : hightest	proc catmod - N/A - default : hightest proc probit - N/A - default : hightest	NOMREG - N/A - default : hightest	. mlogit - OK ("base(#)" option) - default : lowest

이항 로지스틱에서는 SAS와 MINITAB은 기준그룹을 임의로 지정해 주는 descending, reference 옵션 기능이 있다. 그리고 다항 로지스틱에서는 STATA와 MINITAB 패키지인 경우 base(#), reference 옵션으로 기준그룹을 변경해 주는 기능이 있다. 물론 임의로 지정해 줄 수 있는 기능이 없는 패키지일지라도 변수값의 코딩에 의해 기준그룹을 임의로 설정할 수 있을 것이다. 각 패키지마다 출력 결과에서 MINITAB의 기준그룹을 높은 값으로 지정한 결과와 다른 패키지에서 기준그룹을 낮은 값으로 지정한 결과가 동일하다.

<표 4> 각 패키지의 순서지정 기능 비교

Package / Model	MINITAB	SAS	SPSS	STATA
Ordinal logistic	OLogistic - OK ("order" option) - default : ascending - comparison group : lowest	proc logistic - OK ("descending" option) - default : ascending - comparison group : hightest	PLUM - N/A - default : ascending - comparison group : hightest	.ologit - N/A - default : ascending - comparison group : lowest

<표 4>는 순서 로지스틱에서 순서의 형태와 기준그룹을 보여주고 있다. 모든 패키지는 디폴트로 오름차순을 지정하고, 기준그룹은 SAS와 SPSS는 가장 큰 값, STATA와 MINITAB은 가장 낮은 값을 지정한다. SAS는 descending 옵션에 의해 역 순서만 추가 지정하여 가장 낮은 값을 기준그룹으로 정할 수 있다. MINITAB은 order 옵션을 사용하여 모든 경우의 수만큼 순서를 직접 지정하여 활용할 수 있다. 여

기서도 자료의 재기호화를 통해 원하는 순서를 만들어 기준그룹에 적용하여 결과를 생성할 수 있다. 아래 <표 5>는 4종류의 통계패키지에 대한 <표 1>-<표 3>에 관련된 예제를 나타낸 것이다.

<표 5> 관련된 예제

Package	<표 1> - <표 3>에 관련된 예제
MINITAB	데이터세트 "Exh_regr.MTB"를 사용하여 Menu>Stat>Regression의 대화상자에서 Help>Example 선택
SAS	www.ncsu.edu/it/sas/help/의 SAS OnlineDoc Version 8 참고
PSS	SPSS를 활용한 로지스틱 회귀모형의 이해와 응용(김순귀 외 2인(2003))의 7장 "로지스틱 회귀모형의 사례연구" 참고
STATA	Applied Logistic Regression(Hosmer & Lemeshow(2000)) 참고

5. 결론

앞에서 살펴본 네 개의 통계패키지 모두가 기본적인 세 가지의 유형(이분형, 다항, 순서형)의 로지스틱 회귀모형을 처리할 수 있는 프로시저를 구비하고 있다.

SAS는 로지스틱 회귀모형 분석시 다양한 입력 옵션과 통계량의 출력결과를 자세히 제공하였다. SPSS와 MINITAB은 완전 메뉴 방식을 채택하여 사용자들이 편리하고 손쉽게 분석할 수 있는 장점을 갖추었다. STATA 역시 여러 명령을 통하여 많은 통계량을 제공하여 줄 뿐 아니라 그래프 기능 역시 다양하였다.

예측값과 피어슨 잔차 등의 산출시 STATA는 각 공변량 패턴을 기준으로, SAS MINITAB SPSS는 각 개체를 기준으로 하였다. SPSS는 이항 로지스틱 회귀모형에서는 각 개체를 기준으로, 다항 로지스틱 회귀모형에서는 각 공변량 패턴을 기준으로 잔차 통계량을 산출하는 특징이 돋보였다.

조건 로지스틱 회귀모형 분석시 SAS MINITAB SPSS 모두 변수변환을 통하여 가능하였지만, STATA는 변수변환없이 분석할 수 있는 기능이 있어 편리하게 분석할 수 있었다. 조건로지스틱 회귀모형을 적합시키고자 할 때 변수변환 절차 없이 간단하게 처리할 수 있는 기능 개발이 요구된다.

향후 실제적인 상황에서 발생할 수 있는 반응변수가 독립이라는 가정이 위배된 서로 상관된 자료와 반응변수가 다변량인 경우 로지스틱 회귀모형에 적합시키는 프로시져의 개발이 요청된다. 또한 단순임의표집뿐만 아니라 복잡한 표집 계획에 의해 얻은 자료를 처리할 수 있는 기법이 연구되어져야 할 것이다.

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*. 51, 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, 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. *MINITAB Help Release 13 for Window* (2001). MINITAB Inc., State College, PA.
7. *MINITAB StatGuide Release 13 for Window* (2001). MINITAB Inc., State College, PA.
8. Mittlbock, M. and M. Schemper (1999). Computing measures of explained variation for logistic regression model, *Computer Methods and Programs in Biomedicine,* 58, 17-24.
9. *SAS OnlineDoc Version 8* (1999). SAS Institute Inc., Cary, NC.
10. *SPSS Advanced Model 10.0* (1999). SPSS Inc., Chicago, Illinois.
11. *STATA Reference Manual Release 8* (2003). STATA Press, College Station, TX.
12. 김수화, 김승희, 조신섭(1994). 통계패키지에서의 시계열 분석방법의 비교연구, 한국통계학회논문집, 제1권 1호, 119-130.
13. 김순귀, 정동빈, 박영술(2003). SPSS를 활용한 로지스틱 회귀모형의 이해와 응용, SPSS 아카데미, 서울.
14. 성웅현(2001). 응용로지스틱 회귀분석, 탐진, 서울.
15. 조신섭, 신봉섭(1997). 통계적 공정관리를 위한 주요 통계패키지의 비교, 응용통계연구, 제10권 1호, 29-36.
16. 조영석, 문희정(2003). 다변량 Q-기법의 사용을 위한 통계패키지의 비교연구, 한국통계학회논문집, 제10권 2호, 433-443.
17. 최은숙, 박태성, 문경미(1998). 반복측정 자료를 분석하기 위한 통계패키지의 고찰, 한국통계학회논문집, 제5권 1호, 167-176.
18. 허명회, 장진환(1990). 탐색적 데이터분석(EDA) 기능에 관한 통계패키지 프로그램의 비교검토, 응용통계연구, 제3권 2호, 17-25.

[2003년 12월 접수, 2004년 4월 채택]