The Korean Communications
in Statistics Vol. 12 No. 2, 2005
pp. 313-322

Collapsibility and Suppression for Cumulative Logistic Model

Chong Sun Hong[1] and Kil Tae Kim[2]

Abstract

In this paper, we discuss suppression for logistic regression model. Suppression for linear regression model was defined as the relationship among sums of squared for regression as well as correlation coefficients of variables. Since it is not common to obtain simple correlation coefficient for binary response variable of logistic model, we consider cumulative logistic models with multinomial and ordinal response variables rather than usual logistic model. As number of category of a response variable for the cumulative logistic model gets collapsed into binary, it is found that suppressions for these logistic models are changed. These suppression results for cumulative logistic models are discussed and compared with those of linear model.

Keywords : Coefficient of determination, Log-linear model, Logit model.

1. 서 론

회귀분석에서 하나의 설명 변수가 회귀모형에 추가되었을 경우 다른 설명 변수의 중요성을 증가시켜주는 역할을 함으로써 회귀모형의 설명력을 높게 해주는 변수를 suppressor 변수라고 한다(Horst 1941). Suppressor 변수(이하 서프레서 변수라고 함)가 설명변수인 단순회귀모형의 회귀제곱합(sum of squares for regression : SSR)보다 다른 변수가 존재하는 회귀모형에 서프레서 변수가 추가되었을 때 증가된 회귀제곱합이 큰 경우의 현상을 suppression 이라고 정의한다(Conger 1974, Cohen과 Cohen 1975, Velicer 1978). Horst(1941)는 두 개의 설명변수 X_1, X_2 그리고 반응변수 Y로 구성되는 선형회귀모형에서의 suppression(이하 서프레션이라고 함) 조건을 만족하는 X_2를 서프레서 변수로 정의하였다.

$$SSR(X_2 \mid X_1) > SSR(X_2), \tag{1.1}$$

여기서 $SSR(X_2)$은 X_2 하나로만 설명되는 회귀제곱합이며 $SSR(X_2 \mid X_1)$은 변수 X_1이 이

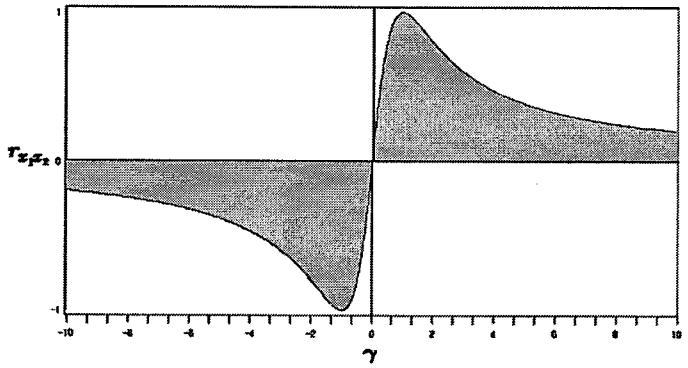
1) Professor, Department of Statistics, Sungkyunkwan University, 3-53 Myeongryun Dong, Chongro Gu, Seoul 110-745, Korea.
E-mail: cshong@skku.ac.kr
2) Graduate Student, Department of Statistics, Sungkyunkwan University, 3-53 Myeongryun Dong, Chongro Gu, Seoul 110-745, Korea.

미 포함되어있는 모형에 변수 X_2가 추가되었을 때 증가된 회귀제곱합이다. 서프레서 변수 X_2는 변수 Y와는 관계없이 변수 X_1에서의 분산의 일부를 억제시켜준다고 할 수 있고, 이로 인해 회귀모형에서 변수 X_1의 중요성을 증가시킨다(Hamilton 1987).

Schey(1993)는 $SSR(X_2)$와 $SSR(X_2 \mid X_1)$ 사이의 관계를 기하학적으로 표현하였으며, 특히 Sharpe와 Roberts(1997)는 선형회귀모형에서 서프레션의 관계식 (1.1)식을 X_1, X_2, Y 변수들 사이에 상관계수들 $r_{x_1x_2}$, r_{yx_1}, r_{yx_2} 로 다음과 같이 표현하였다.

$$r_{x_1x_2}(r_{x_1x_2}-2\gamma/(1+\gamma^2))>0, \tag{1.2}$$

여기서 $\gamma = r_{yx_1}/r_{yx_2}$ 이다. <그림 1.1>은 부등식 (1.2)에 포함된 γ와 $r_{x_1x_2}$의 관계를 이용하여 서프레션이 발생하는 영역을 하얀색으로 표현하였다.



<그림 1.1> 서프레션 발생 영역

회귀모형 이외의 모형에서 논의된 서프레션에 대하여 살펴보면 다음과 같다. 일반적으로 혼용하고 있지만, 저자는 범주형 반응변수에 대해 범주형 설명변수로 이루어진 모형을 로짓(logit) 모형이라 하고 연속형 설명변수로 이루어진 모형을 로지스틱(logistic) 회귀모형이라고 구별하여 설명하고자 한다. Lynn(2003)은 로짓모형에서의 서프레션을 최대가능도비(maximum likelihood ratio)의 관계를 통해서 다음과 같이 정의하였다.

$$L(X_2|X_1)<L(X_2), \tag{1.3}$$

여기서 $L(X_2|X_1)=-2\log(l_r/l_f)$이고 l_r, l_f는 각각 변수 X_1으로 구성된 축소(reduced) 모형과 변수 X_1과 X_2로 구성된 완전(full) 모형의 최대가능도비이며, Hong(2004)은 Lynn이 사용한 축소모형과 완전모형의 로짓모형을 로그선형모형(log-linear model)으로 변환시켜, 각각의 로짓모형의 최대가능도비에 대응하는 식 (1.3)의 관계를 로그선형모형의 최대가능도비로 전환하여 로그선형모형의 서프레션을 설명하였다.

로지스틱회귀모형에서는 서프레션에 대한 연구는 많지 않은데 그 이유 중의 하나는 로지스틱회귀에서는 선형회귀와 같이 서프레션을 규정하는 회귀제곱합과 회귀제곱합을 유도할 수 있는 결정계수(coefficient of determination)를 12가지 종류로 다양하게 정의할 수 있기 때문이다(Mittlbock

과 Schemper 1996, Menard 2000). 본 논문에서는 로지스틱회귀모형에서의 서프레션에 대하여 연구하고자 하며, 변수들 사이의 상관계수로 표현한 Sharpe와 Roberts(1997)가 얻은 선형회귀모형에서의 결과와 비교하고자 한다.

비교 연구를 수행하기 위해서는 이산형 반응변수와 연속형 설명변수와의 상관관계가 필요한데, 다항이며 순위형 범주를 갖는 반응변수와 연속형 설명변수와의 Pearson 상관관계식을 사용하여 계수를 구하는 것이 이항 반응변수가 포함된 단순 로지스틱모형에서 상관계수를 구하는 것보다 더욱 보편적이고 적절하기 때문에 누적로지스틱회귀모형(cumulative logistic regression model)을 고려한다. 본 연구에서는 누적로지스틱회귀모형에서 유도된 회귀제곱합으로 서프레션을 정의하고 그 결과를 변수들의 상관계수들의 관계로 설명하고자 한다. 또한 반응변수의 범주 수를 이항으로까지 점차적으로 축소(collapsing)하면서 설명변수와의 Pearson 상관계수를 사용하여 반응변수가 이항인 로지스틱회귀모형에서의 서프레션도 연구한다. 로지스틱 회귀모형에서의 서프레션 결과를 선형회귀모형의 결과를 표현한 <그림 1.1>과 같이 구현하여 (1.2)식과 비교하여 토론한다.

Murad와 그 외(2003)는 범주의 수가 5개이며 순서를 갖는 반응변수로 구성된 누적로지스틱회귀모형을 고려하고 범주의 수를 축소하면서 회귀계수의 유의성에 대한 연구를 하였다. 본 연구에서는 Murad와 그 외(2003)가 연구한 누적로지스틱회귀모형과 모형에 포함되어 있는 반응변수의 범주축소 방법을 적용하여, 모의실험을 통해 (2.1)식에 적합한 자료를 생성하고, 생성된 자료에서 범주의 수가 5개인 반응변수의 범주 수를 3개와 2개로 축소하면서 서프레션의 변화에 대하여 연구한다. 축소된 반응변수의 범주 수가 이항일 때는 일반적인 로지스틱회귀모형으로 변환된다는 사실에 주의하자. 모형에 포함된 여러 모수 값의 조합을 고려하여 자료를 생성하고, 각 경우에 대하여 상관계수와 회귀제곱합 등을 구한 후 서프레션 발생여부를 살펴보고자 한다. 누적로지스틱회귀모형과 회귀제곱합에 대하여는 2절에서 논의하고, 3절에서는 모의실험에 대해 설명한다. 4절에서는 반응변수의 범주 수가 많은 누적로지스틱회귀모형과 이항 반응변수로 구성된 로지스틱회귀모형에서의 서프레션 결과에 대하여 설명하고, 이를 선형회귀모형의 결과와 비교 토론을 5절에서 한다.

2. 누적로지스틱회귀모형과 회귀제곱합

순위를 갖는 범주의 수준 수가 3 이상인 m개이며(즉 $1,2,\cdots,m$ $(m\geq 3)$) 다항분포(multinomial distribution)를 따르는 반응변수 Z를 고려하자. i번째 관찰값에 대응하는 반응변수와 두 개의 설명변수를 각각 Z_i, X_{1i}, X_{2i}라고 하자 $(i=1,2,\cdots,n)$. Walker와 Duncan(1967)에 의해 처음 제안되었고 MaCullagh(1980)는 비례오즈(proportional odds) 모형으로도 정의하는 누적로지스틱모형을 다음과 같이 고려하여 보자.

$$\log[\frac{P(Z_i\leq j)}{P(Z_i>j)}]=\alpha_j+\beta_1 X_{1i}+\beta_2 X_{2i}\ ,\quad j=1,\cdots,m-1, \tag{2.1}$$

여기서 $\alpha_1\leq\alpha_3\leq\cdots\leq\alpha_{m-1}$이다.

일반적인 로지스틱회귀에서는 회귀제곱합을 유도할 수 있는 결정계수가 여러 가지 형태로 정의되어 있는데, 가장 선호되는 유사(pseudo) 결정계수는 Cox와 Snell(1989, pp. 208-209)이 제안한 R_c^2이며 다음과 같이 정의한다.

$$R_c^2 = \frac{1-[\,L(\hat{\beta}_0)}{L(\hat{\beta})}\Big]^{2/n}, \tag{2.2}$$

여기서 $L(\hat{\beta}_0)$는 절편항만을 포함한 로지스틱회귀모형에서 구한 가능도이며 $L(\hat{\beta})$는 설정된 모형에서 구한 가능도이다. 유사결정계수 R_c^2의 최대값은 1보다 작은 $1-\{L(\hat{\beta}_0)\}^{2/n}$이므로 Nagelkerke(1991)은 최대값이 1이 되도록 수정한 유사수정결정계수를 제안하였다. 또한 Mittlbock과 Schemper(1996), Menard(2000)는 여러 결정계수들 중에서 다음과 같은 로그가능도비(log likelihood ratio)결정계수 R_l^2을 선호하였다.

$$R_l^2 = 1 - \frac{-2\log L(\hat{\beta})}{-2\log L(\hat{\beta}_0)}. \tag{2.3}$$

일반적으로 결정계수는 $R^2 = SSR/SST$으로 정의하므로 회귀제곱합(SSR)은 결정계수의 분자로 구할 수 있다. SAS와 같은 통계패키지를 사용하여 얻은 로지스틱회귀분석의 결과는 (2.2)식과 (2.3)식 등의 유사결정계수, 유사수정결정계수, 그리고 로그가능도비결정계수 값을 제공한다. 유사결정계수와 유사수정결정계수의 값으로부터는 분자에 해당하는 회귀제곱합을 유도할 수 없으나, SAS의 결과에는 로그가능도비결정계수 R_l^2을 구할 수 있는 $-2\log L(\hat{\beta}_0)$와 $-2\log L(\hat{\beta})$의 값을 제공한다. 따라서 본 연구에서는 로그가능도비결정계수 R_l^2로부터 유도한 회귀제곱합 SSR_l을 다음과 같이 정의하기로 한다.

$$SSR_l = -2\log L(\hat{\beta}_0) + 2\log L(\hat{\beta}). \tag{2.4}$$

본 연구에서는 많이 사용하는 통계패키지인 SAS를 이용하여 하나의 설명변수 X_1과 X_2가 각각 포함된 모형과 모두 포함된 모형으로부터 회귀제곱합 $SSR_l(X_2)$와 $SSR_l(X_2 \mid X_1) = SSR_l(X_1, X_2) - SSR_l(X_1)$을 구하여, 선형회귀에서의 서프레션을 (1.1)식과 같이 정의하였듯이 로지스틱회귀모형에서의 서프레션을 다음과 같이 유도할 수 있다.

$$SSR_l(X_2 \mid X_1) > SSR_l(X_2). \tag{2.5}$$

3. 모의실험

3.1 연구방법

본 절에서는 (2.1)식에서 범주 수가 $m=5$인 경우의 누적 로지스틱 회귀모형을 따르는 자료를 몬테칼로 방법을 이용하여 생성하여 2절에서 논의한 $SSR_l(X_2)$와 $SSR_l(X_2 \mid X_1)$을 구한 다음 어떤 상황에서 서프레션이 발생하는지에 대하여 탐색하고자 한다.

먼저 표본크기가 100인 이변량 정규분포를 따르는 설명변수 값 x_{1i}과 x_{2i}, $i=1,\cdots,100$을 생성한다. 이때 각각의 모평균 μ_{x_1}, μ_{x_2}은 0으로 고정시키고, 모분산 $\sigma_{x_1}^2$, $\sigma_{x_2}^2$은 1, 4인 경우를 고려한다. 그리고 두 설명변수 x_1과 x_2의 모상관계수 $\rho_{x_1x_2}$는 -0.8부터 0.8까지 간격 0.2의

크기로 변화시킨다. 또한 회귀계수 β_1, β_2는 -1에서 1까지 0.2 간격으로 변화시킨다.

다음으로는 순위형 종속변수 z를 생성하기 위하여 Murad와 그 외(2003)가 연구한 방법과 같이 우선 다항분포를 따르는 확률벡터가 균일인 $\boldsymbol{p}_1$=(0.2, 0.2, 0.2, 0.2, 0.2), 대칭적인 $\boldsymbol{p}_2$=(0.05, 0.3, 0.3, 0.3, 0.05), 그리고 비대칭적인 $\boldsymbol{p}_3$=(0.05, 0.05, 0.3, 0.3, 0.3)와 같은 세가지 경우를 고려한다. $\beta_1 = \beta_2 = 0$을 가정할 때 $P(Z \le j) = [1+\exp(-\alpha_j)]^{-1}$이므로 세 가지 경우의 다항분포를 따를 때 상수 α_j를 추정하여 <표 3.1>에 나열하였다.

<표 3.1> Z의 분포에 따라 추정된 α_j

	균일	대칭	비대칭
$\hat{\alpha}_1$	-1.386	-2.945	-2.945
$\hat{\alpha}_2$	-0.405	-0.619	-2.197
$\hat{\alpha}_3$	0.405	0.619	-0.405
$\hat{\alpha}_4$	1.386	2.944	0.847

생성된 이변량 정규난수 x_{1i}과 x_{2i}와 변화시키는 모수 β_1, β_2, 그리고 α_j의 추정값을 대입하여 다음과 같이 누적로지스틱모형에서의 범주별 누적확률을 구한다.

$$\hat{P}(Z \le j) = \exp(\hat{\alpha}_j + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}) / [1 + \exp(\hat{\alpha}_j + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i})] \quad (3.1)$$

(3.1)식으로부터 $\hat{p}_j = \hat{P}(Z_i = j)$의 확률을 구하고 다항분포, multinom$l$ (100, $\hat{\boldsymbol{p}} = (\hat{p}_1, \cdots, \hat{p}_5)$)를 따르는 z값을 생성한다.

반응변수 Z의 범주 수는 5이며 1부터 5까지의 값을 갖는데, Murad와 그 외(2003)가 연구한 방법과 같이 반응변수의 범주 수가 3(1-2, 3, 4-5)인 경우와 2(1-2, 3-5)인 경우로 축소(collapsing)하여, m=5, 3, 2인 경우에 (2.1)식의 누적 로지스틱 회귀모형을 따르는 자료를 생성한다. 따라서 모수 $\sigma^2_{x_1}$, $\sigma^2_{x_2}$, $\rho_{x_1x_2}$, β_1, β_2의 값을 변화시켜 총 $2\times2\times9\times11\times11 = 4356$ 경우의 조합으로 표본크기 100인 자료를 생성하여 z_i, x_{1i}, x_{2i}값을 얻는다.

생성된 순위형 종속변수 z_i와 연속형 설명변수 x_{1i}과 x_{2i}, $i = 1, \cdots, 100$로 구성된 자료들을 SAS의 'Proc Logistic'을 이용하고 Model 문장의 옵션에 'link=clogit'을 이용하여 누적로지스틱분석을 실시한다. (2.4)식에서 정의한 SSR_I을 구한 후, $SSR_I(X_2)$과 $SSR_I(X_2 \mid X_1)$의 크기 비교에 의해서 서프레션의 발생여부를 살펴본다. 이러한 모의실험을 통하여 Z가 따르는 다항분포와 범주의 축소에 따라서 각 모수 상황에서 서프레션 현상이 어떻게 변화되는지 탐색하고 비교분석하였다.

3.2 범주 축소와 서프레션 결과

생성된 자료를 누적로지스틱 분석하여 SSR_I을 구하고, 서프레션이 발생하는 관계에 만족하는

지에 대한 결과를 얻었다. Sharpe와 Roberts(1997)는 서프레션 관계를 (1.2)식과 같이 유도하고 <그림 1.1>과 같이 시각적으로 표현하였는데, 본 연구에서도 $\gamma = r_{zx_1}/r_{zx_2}$와 $r_{x_1x_2}$에 대하여 그림으로 표현하였는데 서프레션이 발생하는 경우는 'o'로 발생하지 않는 경우에는 'x'로 나타내었다. 그리고 회귀모형에서 유도한 (1.2)식을 보조선으로 추가하였다.

먼저 다항분포를 따르는 확률변수 Z의 확률이 균일한 경우 (p_1), 대칭인 경우 (p_2) 그리고 비대칭인 경우 (p_3)에 γ와 $r_{x_1x_2}$에 대하여 서프레션의 발생에 관한 그림을 각각 <그림 3.1>부터 <그림 3.3>에 그리고 범주 수가 5와 2인 경우의 그림을 각각 (A)와 (B)로 구분하여 나타내었다(범주 수가 3인 경우는 생략함). 아래 그림은 모분산 $\sigma^2_{x_1} = \sigma^2_{x_2} = 1$인 경우에 대한 결과를 표현한 그림이며, 모분산이 모두 1이 아닌 경우에 대하여는 산포정도가 조금 심해지지만 매우 유사한 현상을 나타내고 있기 때문에 생략하였다.

<그림 3.1>부터 <그림 3.3>에서 서프레션이 발생하지 않는 영역을 살펴보면, 설명변수들의 상관계수 $r_{x_1x_2}$가 양의 값을 가진 경우 반응변수와 설명변수들의 상관계수 r_{zx_1}, r_{zx_2} 들의 비율로 표현되는 $\gamma = r_{zx_1}/r_{zx_2}$는 +1의 값 근처에 집중되어 있고, $r_{x_1x_2}$가 음의 값을 가진 경우에는 γ는 -1의 값 근처에 집중되어있다. 즉 반응변수와 설명변수들의 상관계수들이 모두 유사한 절대값을 갖고 있는 경우에 서프레션이 발생하지 않는다는 것을 파악할 수 있다. 이것은 선형회귀에서의 서프레션 발생 관계식 (1.2)와 <그림 1.1>에서 논의되고 설명된 결과와 매우 유사함을 파악할 수 있다. 로지스틱회귀모형에서 서프레션이 발생하지 않는 지역이 1, 3사분면에 그리고 γ의 절대값이 1인 경우에 집중하고 있으며 원점을 중심으로 대칭적으로 나타난다는 것과 2와 4사분면에는 서프레션이 발생한다는 사실은 선형회귀모형에서 서프레션이 발생하지 않는 영역을 시각적으로 표현한 <그림 1.1>과 유사하다는 결론을 내릴 수 있다.

균일확률, 대칭, 비대칭인 경우를 각각 나타낸 <그림 3.1 (A)>부터 <그림 3.3 (A)>를 자세히 살펴보면(즉, m=5인 경우) 약간씩 차이를 발견할 수 있는데 균일확률인 경우보다는 대칭인 경우 그리고 대칭인 경우보다는 비대칭인 경우에, 선형회귀의 서프레션 발생 영역 내에서 로지스틱의 서프레션이 발생하지 않는 경우는 감소하고 선형회귀의 서프레션이 발생하지 않는 영역 내에서 로지스틱의 서프레션이 발생하는 경우가 증가한다는 것을 발견할 수 있으며, 이를 <표 3.2>부터 <표 3.4>에서 설명하고자 한다.

<표 3.2>부터 <표 3.4>는 다항분포를 따르는 확률변수 Z의 확률이 균일한 경우 (p_1), 대칭인 경우 (p_2) 그리고 비대칭인 경우 (p_3)에 대하여 선형회귀에서의 (1.3)식에 관한 서프레션 발생 영역을 기준으로 하여 SSR_l을 이용한 누적로지스틱모형에서의 서프레션 발생결과와 비교하여 분할표로 나타내었다. 앞에서 논의한 그림들에서는 범주 수가 축소되는 과정에서 범주 수(m)가 3인 경우의 그림을 생략하였지만 <표 3.2>부터 <표 3.4>에서는 범주 수가 5, 3, 그리고 2인 모든 경우에 대하여 상세히 표를 만들었다. 예를 들어 <표 3.2>의 범주 수가 5인 경우에서 선형회귀에서 서프레션이 발생하는 영역을 조건으로 할 때, 누적로지스틱회귀에서의 서프레션이 발생하는 경우는 전체 경우 중에서 94.74% (= 29.34/(29.34+1.63))이며, 선형회귀와 누적로지스틱회귀와 서로 상반되는 결과가 발생하는 비율은 4.41% (= 1.63 + 2.78)인 것을 파악할 수 있다.

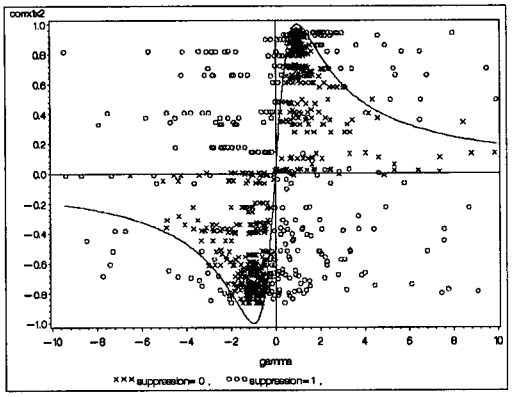


그림 3.1 (A) p_1과 m=5인 경우의 결과

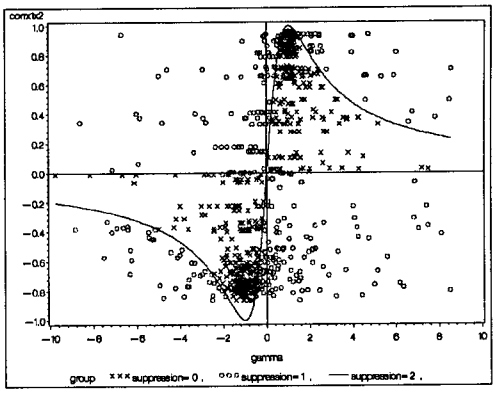


그림 3.1 (B) p_1과 m=2인 경우의 결과

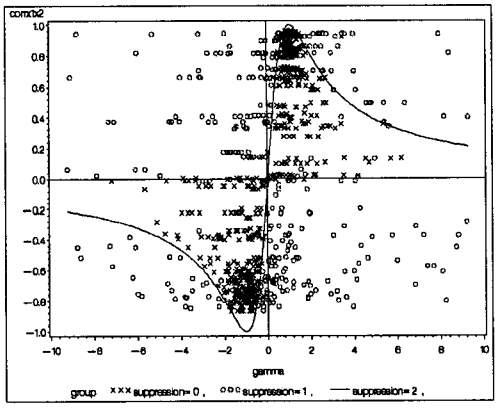


그림 3.2 (A) p_2와 m=5인 경우의 결과

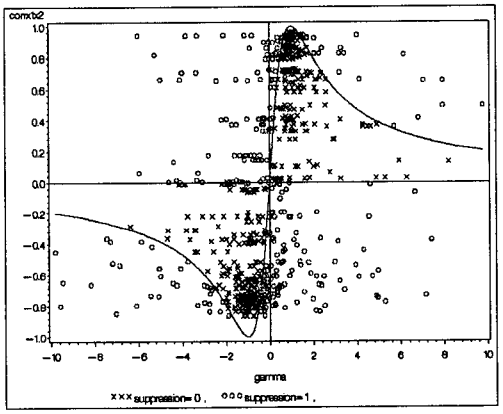


그림 3.2 (B) p_2와 m=2인 경우의 결과

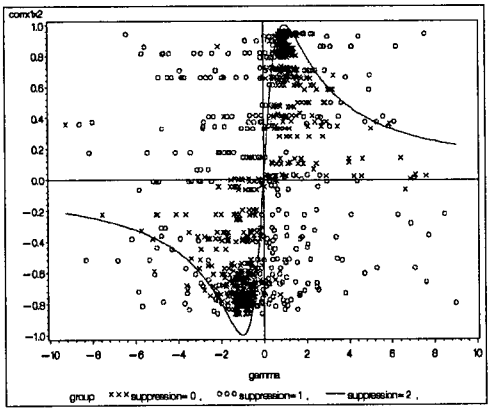


그림 3.3 (A) p_3와 m=5인 경우의 결과

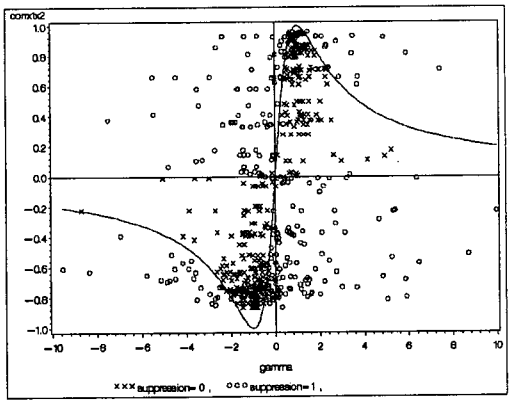


그림 3.3 (B) p_3와 m=2인 경우의 결과

<표 3.2> p_1 인 경우의 서프레션 발생 결과 비교(단위=%)

누적로지스틱의 서프레션		선형회귀의 서프레션	
		미발생	발생
m= 5	미발생	66.25	1.63
	발생	2.78	29.34
m= 3	미발생	65.15	1.38
	발생	2.27	31.20
m= 2	미발생	65.36	0.12
	발생	0.87	33.65

<표 3.3> p_2인 경우의 서프레션 발생 결과 비교(단위=%)

누적로지스틱의 서프레션		선형회귀의 서프레션	
		미발생	발생
m= 5	미발생	65.24	2.02
	발생	3.10	29.64
m= 3	미발생	66.17	1.13
	발생	1.81	30.89
m= 2	미발생	65.13	0.14
	발생	1.08	33.65

<표 3.4> p_3인 경우의 서프레션 발생 결과 비교(단위=%)

누적로지스틱의 서프레션		선형회귀의 서프레션	
		미발생	발생
m= 5	미발생	64.37	3.10
	발생	4.55	27.98
m= 3	미발생	63.43	2.80
	발생	4.27	29.50
m= 2	미발생	63.04	0.92
	발생	1.86	34.18

4. 결론 및 토의

본 연구를 통해서 누적로지스틱회귀모형에서의 서프레션에 관한 연구결과가 Sharpe와 Roberts (1997)가 선형회귀모형에서 연구한 서프레션의 발생영역과 전체적으로 매우 유사하다는 것을 알 수 있었다. 이것은 누적로지스틱모형에서의 서프레션이 선형회귀모형에서의 서프레션과 동일하게 설명할 수 있다는 것을 의미한다.

다항분포를 따르는 반응변수 Z의 확률이 균일확률, 대칭, 비대칭인 경우에 대하여 설명변수들의 상관계수 $r_{x_1x_2}$와 반응변수와 설명변수들의 상관계수 r_{zx_1}, r_{zx_2} 들의 비율로 표현되는 $\gamma = r_{zx_1}/r_{zx_2}$와의 관계 하에서 서프레션의 발생 결과를 살펴보면, 균일확률인 경우보다는 대칭인 경우 그리고 대칭인 경우보다는 비대칭인 경우에 선형회귀의 서프레션 발생 영역 내에서 로지스틱의 서프레션 발생하는 경우가 작아지지만 범주 수가 축소될수록 동일한 형태로 수렴한다는 것을 발견하였다.

구체적으로 살펴보면 Z의 범주수가 5개 일 때, Z가 균일 다항분포를 따를 때는 동일 발생비율이 95.59%(= 66.25+29.34)를 보였고, 대칭 다항분포인 경우에는 94.88% (= 65.24+29.64)를 보였으며, 비대칭 다항분포를 따를 때는 92.35% (= 64.37+27.98)를 보였다. 따라서 종속변수 Z가 균일 다항분포를 따를 때 선형회귀 모형과의 유사정도가 가장 높다고 할 수 있고, Z가 대칭 분포를 따를 때는 그 다음 높다고 할 수 있으며, Z가 비대칭 분포를 따를 때는 선형회귀와의 유사정도가 가장 낮다고 할 수 있다.

Z의 범주 수를 3개, 2개로 축소했을 때의 동일 발생비율을 살펴보면, Z가 균일 다항분포인 경우에는 동일 발생비율이 95.59%에서 96.35%, 99.01%로 높아졌고, 대칭 다항분포인 경우에는 94.88%에서 97.06%, 98.78%로 높아졌으며, 비대칭 다항분포인 경우에는 92.35%에서 92.93%, 97.22%로 높아졌다. 그리고 범주 수가 2개로 축소되어 이항 반응변수로 구성된 일반적인 로지스틱회귀모형인 경우에 선형회귀에서 서프레션 발생하는 영역을 조건으로 할 때, 로지스틱회귀의 서프레션이 발생하는 비율은 반응변수 Z의 확률이 대칭적이 아니라 할지라도 99.6% (= 34.65/(34.65+0.14)정도 수렴하고 있다는 것을 발견할 수 있다. 따라서 선형회귀와 비교한 서프레션의 동일 발생 비율은 Z의 범주 수에 영향을 받는다고 할 수 있으며, 반응변수의 범주 수가 두 개일 때, 즉 일반적인 이항 로지스틱모형인 경우 선형회귀에서의 서프레션 발생영역과 가장 유사하다고 결론내릴 수 있다. 이것은 단 두 개의 결과범주만을 갖는 이산형 반응변수와 연속형 설명변수들의 상관계수 r_{zx_1}, r_{zx_2}들을 일반적인 상관계수로 간주해도 차이가 없다는 것을 의미한다.

그러므로 본 연구에서는 누적로지스틱모형에서 (2.4)식의 SSR_I을 이용해서 유도한 서프레션 정의 (2.5)식을 회귀제곱합을 이용해서 정의한 선형회귀에서의 서프레션과 비교했을 때, 반응변수와 설명변수들 사이의 상관계수들로 표현된 로지스틱모형에서의 서프레션은 매우 유사한 영역에서 발생하며, 특히 이항 로지스틱모형에서의 유사정도가 가장 높다는 것을 발견하였다.

참 고 문 헌

[1] Cohen, J. and Cohen, P.(1975). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences,* New Jersey: Lawrence Erlbaum Associates.

[2] Conger, A. J.(1975). A Revised Definition for Suppressor Variables: A Guide to Their Identification and Interpretation, *Educational and Psychological Measurement,* 34, 35-46

[3] Cox, D. R and Snell, E. J.(1989). A*nalysis of Binary Data,* Chapman and Hall.

[4] Hamilton, D.(1987). Sometimes $R^2 > r^2_{yx_1} + r^2_{yx_2}$ Correlated Variables are not Always

Redundant, *The American Statistician*, 41, 129-132.

[5] Horst, P.(1941). *The Role of Prediction Variables Which are Independent of the Criterion, in The Prediction Adjustment*, ed. P. Horst, New York: Social Science Research Council.

[6] Hong, C. S.(2004). Suppression and Collapsibility for Log-linear Model, *The Korean Communication in Statistics*, 11, 3, 519-527

[7] Lynn, H. S.(2003). Suppression and Confounding in Action, *The American Statistician*, 57, 58-61.

[8] McCullagh, P.(1980), Regression Models for Ordinal Data (with discussion), *Journal of Royal Statistical Society*, Ser. B, 42, 109-142.

[9] Menard, S.(2000). Coefficients of Determination for Multiple Logistic Regression Analysis, *The American Statistician*, 54, 17-24.

[10] Mittlebock, M. and Schemper, M.(1996). Explained Variation for Logistic Regression, *Statistics in Medicine*, 15, 1987-1997

[11] Murad, H., Fleischman, A., Sadetzki, S., Geyer, O., and Freedman, L. S.(2003). Small Samples and Ordered Logistic Regression: Does it Help to Collapse Categories of Outcome?" *The American Statistician*, 57, 3, 155-160.

[12] Nagelkerke, N. J. D.(1991). A Note on a General Definition of the Coefficient of Determination. *Biometrika*. 78: 691-692

[13] Schey, H. M.(1993). The Relationship Between the Magnitudes of $SSR(x_2)$ and $SSR(x_2 \mid x_1)$: A Geometric Description, *The American Statistician*, 47, 26-30.

[14] Sharpe, N. R., and Roberts, R. A.(1997). The Relationship Among Sums of Squares, Correlation Coefficients, and Suppression, *The American Statistician*, 51, 46-48.

[15] Velicer, W. F.(1978). Suppressor Variables and the Semipartial Correlation Coefficient. *Educational and Psychological Measurement*, 38: 953-958.

[16] Walker. S. H. and Duncan, D. B.(1967). Estimation of the Probability of an Event as a Function of Several Independent Variables. *Biometrika*. 54, 167-179.

[2004년 12월 접수, 2005년 4월 채택]