응용통계연구 제 18권 3호, 2005년, pp. 701–712

로지스틱 회귀모형에서의 SUPPRESSION

홍종선[1] 김호일[2] 함주형[3]

요 약

로지스틱 회귀모형에서 suppression의 논의는 선형회귀의 논의보다 많지 않은데 그 이유 중의 하나는 회귀제곱합 또는 결정계수의 정의가 유일하지 않고 다양하기 때문이다. 여러 종류의 결정계수들 중에서 선호되는 두 종류의 결정계수와 Liao와 McGee(2003)가 제안한 두 종류의 수정 결정계수의 정의로부터 회귀제곱합을 유도하여 로지스틱 회귀모형에서의 suppression을 설명하고자 한다. 모의실험을 통하여 자료를 생성하여 어떤 경우에 suppression이 발생하는지를 살펴보고 그 결과를 선형회귀모형에서의 suppression 결과와 비교한다.

주요용어: 결정계수, 내재예측오차, 로그선형모형, 로짓모형, 편의, 회귀제곱합.

1. 서론

선형회귀분석에서 하나의 설명변수가 회귀모형에 추가되었을 경우 다른 설명변수의 중요성을 증가시켜주는 역할을 함으로써 회귀모형의 설명력을 높여주는 역할을 하는 변수를 suppressor 변수라고 한다. suppressor 변수는 반응변수와는 상관관계가 적지만 또 다른 설명변수와는 유의한 연관성이 존재한다. 그리고 suppressor 변수만이 설명변수가 되는 단순회귀모형의 설명력보다 이미 다른 설명변수가 포함된 모형에 suppressor 변수가 추가된 회귀모형에서 추가된 변수의 설명력이 높은 경우 그 현상을 suppression이라고 정의한다(Conger 1974, Cohen과 Cohen 1975, Velicer 1978). 선형회귀분석에서는 이 현상에 대해서 추정된 회귀의 자료에 대한 설명력 정도를 보여주는 회귀제곱합(sum of squares for regression : SSR)들의 관계를 사용하여 나타낼 수 있다. 선형회귀에서의 suppression을 살펴보기 위해 두 개의 설명변수 X_1, X_2 그리고 반응변수로 표현되는 다음과 같은 모형을 고려할 때

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i,\ i = 1, \cdots, n, \tag{1.1}$$

Horst(1941)는 다음의 조건을 만족하는 X_2를 suppressor 변수로 정의하였다.

$$SSR(X_2|X_1) > SSR(X_2), \tag{1.2}$$

여기에서 $SSR(X_2)$은 X_2 하나로만 표현되는 회귀제곱합이며 $SSR(X_2|X_1)$은 변수 X_1이 이미 포함되어있는 모형에 변수 X_2가 추가되었을 때 증가된 회귀제곱합이다. suppressor 변

1) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 경제학부 통계학전공, 교수
E-mail: cshong@skku.ac.kr
2) (137-070) 서울특별시 서초구 서초동 1303-22 교보타워 8층, 교보자동차보험, CS기획부
3) (100-736) 서울특별시 중구 서소문동 120-20 동화빌딩 9층, 한국후지제록스, MA 1 Team

수 X_2는 변수 Y와는 관계없이 변수 X_1에서의 분산의 일부를 억제시켜준다고 할 수 있고, 이로 인해 (1.1)식의 회귀모형에서 변수 X_1의 중요성을 증가시킨다(Hamilton, 1987). 위에서 제시한 suppression의 정의에 대해서 Hamilton (1987, 1988)은 $SSR(X_2|X_1)$과 $SSR(X_2)$ 사이의 관계, 그리고 X_1, X_2 그리고 Y에 대한 각각의 상관계수 r_{yx_1}, r_{yx_2}와 $r_{x_1x_2}$에 대해 고려하였고, Mitra(1988)와 Freud(1988)가 이를 더욱더 발전시켰다. Schey(1993)는 $SSR(X_2)$와 $SSR(X_2|X_1)$ 사이의 관계를 기하학적 표현으로 설명하였다. 특히 Sharpe와 Roberts(1997)는 suppression의 관계식인 (1.2)식을 X_1, X_2 그리고 Y변수들 사이의 상관계수들 $r_{x_1x_2}$, r_{yx_1}, r_{yx_2}로 다음과 같이 표현하고

$$r_{x_1x_2}(r_{x_1x_2} - 2\gamma/(1+\gamma^2)) > 0, \tag{1.3}$$

여기서 $\gamma = r_{yx_1}/r_{yx_2}$. 그리고 γ에 대응하는 $r_{x_1x_2}$의 관계를 그림 1.1과 같이 제시해서 시각적으로 suppression이 발생하는 조건을 보여주었다(그림에서 하얀 부분이 suppression이 발생하는 영역이다).

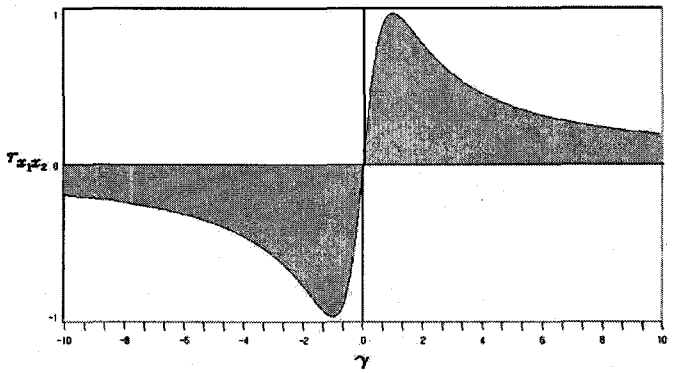

그림 1.1: suppression 발생조건

그래프에서 곡선은 원점에 대해 대칭이고, $r_{x_1x_2} > 0$일 때는 $\gamma = 1$ 즉, r_{yx_1}과 r_{yx_2}가 유사한 값을 가진 경우와 $r_{x_1x_2} < 0$일 때는 $\gamma = -1$ 즉, r_{yx_1}과 r_{yx_2}가 서로 다른 부호이며 절대값이 유사한 값을 가진 경우에는 suppression이 발생하지 않는다. 또한 $r_{x_1x_2}$와 γ의 부호가 반대이면 항상 suppression이 발생하고 부호가 같은 경우에도 발생할 수 있다.

Lynn(2003)은 설명변수와 반응변수가 모두 범주형변수로 구성된 로짓(logit) 모형에서의 suppression을 최대가능도 통계량들의 관계를 통해서 다음과 같이 정의하였다.

$$L(X_2|X_1) < L(X_2), \tag{1.4}$$

여기서 $L(X_2|X_1) = L(X_1) - L(X_1, X_2)$이며, $L(X_1)$과 $L(X_1, X_2)$는 각각 설명변수가 X_1 인 모형과 X_1, X_2로 구성된 모형에 대한 -2배의 로그최대가능도 통계량이다. 홍종선(2004)은

Lynn(2003)이 사용한 축소모형과 포화모형의 로짓모형을 로그선형모형(log-linear model)으로 변환시켜, 각각의 로짓모형의 최대가능도에 대응하는 (1.4)식의 관계를 로그선형모형의 최대가능도로 전환하여 로그선형모형의 suppression을 정의하였다.

선형회귀모형, 로짓모형 그리고 로그선형모형과 달리 이항변수인 반응변수와 연속형 변수인 설명변수로 구성된 로지스틱 회귀모형(logistic regression model)에서는 suppression에 대한 연구는 많지 않다. 그 이유 중의 하나는 로지스틱회귀에서는 선형회귀와 같이 suppression을 규정하는 SSR과 SSR을 유도할 수 있는 결정계수(R^2)의 정의가 유일하지 않고 다양하기 때문이다(Mittlbock과 Schemper 1996, Menard 2000). 알려진 12종류의 결정계수들 중에서 Mittlbock과 Schemper(1996), Menard(2000)는 (2.2)식과 (2.3)식에서 정의한 R_l^2와 R_o^2를 선호하였다. 그리고 Liao와 McGee(2003)는 선형회귀에서의 수정된 결정계수(R_{adj}^2)의 수학적이며 개념적인 확장으로 로지스틱회귀에 있어서 (2.4)식과 (2.5)식에서 정의한 두 종류의 수정 결정계수 통계량 $R_{l,adj}^2$와 $R_{o,adj}^2$를 제안하였다. 새롭게 제안한 수정 결정계수들은 모형에 부적절한 설명변수들의 추가 혹은 표본크기의 변화에 민감하지 않는 것을 보였다. 따라서 본 논문에서는 R_l^2와 R_o^2 그리고 Liao와 McGee(2003)가 제안한 수정 결정계수 $R_{l,adj}^2$와 $R_{o,adj}^2$의 정의에서 SSR을 유도하여 로지스틱 회귀모형에서의 suppression 관계를 정의하고, 몬테칼로 방법으로 자료를 생성하여 suppression이 어떤 경우에 발생하는지에 대하여 살펴보고 그 결과를 선형회귀에서 Sharpe와 Roberts(1997)가 시각적으로 제시한 그림 1.1과 비교하는 데 연구목적이 있다.

논문의 구성은 다음과 같다. 2절은 선형회귀의 suppression 정의를 바탕으로 로지스틱 회귀에서 선호되고 새롭게 제안된 결정계수들로부터 네 종류의 SSR들을 유도하여 suppression을 정의한다. 3절에서는 모의실험을 통하여 자료를 생성하여 네 종류의 SSR을 각각 구하여 로지스틱 회귀모형에서의 suppression의 발생여부를 시각적으로 표현하여 선형회귀의 결과와 비교한다. 4절에서는 이를 토의하고 결론을 유도한다.

2. 로지스틱회귀에서의 Suppression

2.1. 로지스틱회귀에서 제안된 결정계수

일반적인 다변량 로지스틱 회귀모형을 고려한다.

$$Z_i \sim \text{Bernoulli}(\pi_i),\ \text{logit}(\pi_i) = \alpha + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \epsilon_i,\ i = 1, \ldots, n. \tag{2.1}$$

Kvalseth(1985)는 로지스틱 회귀모형에서 결정계수에 대한 여덟 가지의 기준을 제시하였는데 이를 기준으로 Mittlbock과 Schemper(1996), Menard(2000)는 12종류의 결정계수들 중에서 다음과 같은 R_l^2와 R_o^2를 선호했다.

$$R_l^2 = 1 - \frac{-l(\mathbf{z}, \hat{\pi})}{-l(\mathbf{z}, \hat{\pi}^0)} \tag{2.2}$$

$$R_o^2 = 1 - \frac{\sum_{i=1}^n (z_i - \hat{\pi}_i)^2}{\sum_{i=1}^n (z_i - \bar{z})^2}, \tag{2.3}$$

여기서 $\mathbf{z} = (z_1, \ldots, z_n)$는 이항반응변수벡터, $\pi = (\pi_1, \ldots, \pi_n)$는 확률벡터, 그리고 $\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_n)$은 최대가능도 추정에 기초를 둔 로지스틱 회귀모형 하에서 적합된 확률벡터 그리고 $\hat{\pi}^0 = (\bar{z}, \ldots, \bar{z})$, $\bar{z} = \sum_{i=1}^n z_i/n$는 절편하나만을 가진 모형에서 적합된 확률벡터이며, $l(\mathbf{z}, \pi)$는 로그최대가능도이며 $l(\mathbf{z}, \pi) = \sum_{i=1}^n \{z_i \log(\pi_i) + (1 - z_i)\log(1 - \pi_i)\}$로 정의되는데 이는 식 (1.4)의 $L(\cdot)$ 함수에서 -2배 하기 전의 로그최대가능도이다. $-l(\mathbf{z}, \hat{\pi})$값이 작을수록 더 좋은 적합을 의미하기 때문에 일반적으로 음의 로그가능도를 사용한다.

Liao와 McGee(2003)는 로지스틱 회귀모형에서 일반적으로 사용되는 결정계수인 R_l^2와 R_o^2도 역시 연관성 정도를 심각하게 과대 추정할 수 있기 때문에 (2.4)와 (2.5)식과 같은 결정계수에 대한 두 종류의 수정통계량 $R_{l,adj}^2$와 $R_{o,adj}^2$를 제안하였다. 평균이 π_i인 베르누이 분포를 따르는 z_i^{new}를 로지스틱회귀에서 반응변수의 독립적인 반복이라고 하면, R_l^2와 R_o^2에서의 π_i에 의한 z_i^{new}의 내재예측오차(IPE : Inherent Prediction Error)는 다음과 같이 정의된다.

$$IPE_l(\pi) \equiv n^{-1}E\{-l(\mathbf{z}^{new}, \pi)\} = -n^{-1}\sum_{i=1}^n \{\pi_i \log \pi_i + (1 - \pi_i)\log(1 - \pi_i)\}$$

$$IPE_o(\pi) \equiv n^{-1}E\sum_{i=1}^n \{(z_i^{new} - \pi_i)^2\} = n^{-1}\sum_{i=1}^n \pi_i(1 - \pi_i).$$

참의 π_i가 z_i^{new}를 예측하는 데 이용되기 때문에 이 오차를 내재예측오차라 부른다. 다음으로 로지스틱 모형에서의 추정량들을 고려하면 $IPE_l(\pi)$와 $IPE_o(\pi)$의 단순한 추정량은 각각 $-n^{-1}l(\mathbf{z}, \hat{\pi})$과 $n^{-1}\sum_{i=1}^n (z_i - \hat{\pi}_i)^2$이다. 두 추정량은 $\hat{\pi}$과 $\mathbf{z}$의 거리가 π와 $\mathbf{z}$의 거리보다 줄어드는 경향이 있기 때문에 0에 가까운 편의를 가지고 있다. 그 편의들은 각각 다음과 같다.

$$B_l(\pi) = n^{-1}E\{-l(\mathbf{z}^{new}, \hat{\pi}^{new})\} - IPE_l(\pi)$$

$$B_o(\pi) = n^{-1}E\sum_{i=1}^n \{(z_i^{new}, \hat{\pi}_i^{new})^2\} - IPE_o(\pi).$$

이후에 선형회귀에서의 R_{adj}^2를 고려하여 로지스틱회귀에 대한 수정 결정계수들을 다음과 같이 정의한다.

$$R_{l,adj}^2 = 1 - \frac{\widehat{IPE}_l^p}{\widehat{IPE}_l^0} \tag{2.4}$$

$$R_{o,adj}^2 = 1 - \frac{\widehat{IPE}_o^p}{\widehat{IPE}_o^0}, \tag{2.5}$$

여기서 $\widehat{IPE}_l^p = -n^{-1}l(\mathbf{z}, \hat{\pi}) - B_l(\hat{\pi})$, $\widehat{IPE}_o^p = n^{-1}\sum_{i=1}^n (z_i - \hat{\pi}_i)^2 - B_o(\hat{\pi})$이고 절편하나만을 갖는 모형 하에서 $IPE(\pi)$의 편의 수정된 추정량들 $\widehat{IPE}_l^0$과 $\widehat{IPE}_o^0$은 적합된 모형에서의 $\hat{\pi}$과 $\hat{\pi}^{new}$를 제외한 같은 공식을 사용하여 얻을 수 있다.

제안된 수정 결정계수들은 R_l^2와 R_o^2를 수정 보완했기에 그 유의함을 그대로 계승하였고 Kvalseth(1985)가 제시한 여덟 가지 기준에도 만족스럽다. 또한 모의실험을 바탕으로

R_l^2와 R_o^2보다 부적절한 설명변수의 추가 혹은 표본크기 변화에 민감하지 않다는 결론을 내렸다(Liao와 McGee, 2003). 2.2절에서는 이러한 이전 연구들에서 유의하다고 판단된 결정계수 R_l^2와 R_o^2 그리고 이를 보완한 수정 결정계수 $R_{l,adj}^2$, $R_{o,adj}^2$를 바탕으로 네 종류의 SSR을 유도하여 로지스틱회귀에서의 suppression의 조건을 정의해 보고자 한다.

2.2. 로지스틱회귀에서의 Suppression

설명변수가 두개인 다음과 같은 로지스틱 회귀모형을 고려하여 보자.

$$Z_i \sim \text{Bernoulli}(\pi_i),\ \text{logit}(\pi_i) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i,\ i = 1, \ldots, n. \tag{2.6}$$

로지스틱회귀에서의 suppression 현상을 살펴보기 위해 2.1절에서 논의한 네 종류의 결정계수 R_l^2, R_o^2, $R_{l,adj}^2$, $R_{o,adj}^2$를 바탕으로 SSR을 유도하여 보자. 식 (2.2)과 (2.3)으로부터 SSR_l과 SSR_o를 각각 다음과 같이 유도할 수 있다.

$$SSR_l = -l(\mathbf{z}, \pi^0) + l(\mathbf{z}, \hat{\pi}) \tag{2.7}$$

$$SSR_o = \sum_{i=1}^{n}(z_i - \bar{z})^2 - \sum_{i=1}^{n}(z_i - \hat{\pi}_i)^2. \tag{2.8}$$

그리고 식 (2.4)와 (2.5)로부터 $SSR_{l,adj}$와 $SSR_{o,adj}$를 각각 다음과 같이 유도할 수 있다.

$$SSR_{l,adj} = \widehat{IPE}_l^0 - \widehat{IPE}_l^p = -n^{-1}l(\mathbf{z}, \pi^0) - B_l(\pi^0) + n^{-1}l(\mathbf{z}, \hat{\pi}) + B_l(\hat{\pi}) \tag{2.9}$$

$$SSR_{o,adj} = \widehat{IPE}_o^0 - \widehat{IPE}_o^p = n^{-1}\sum_{i=1}^{n}(z_i - \pi^0)^2 - B_o(\pi^0) - n^{-1}\sum_{i=1}^{n}(z_i - \hat{\pi}_i)^2 + B_o(\hat{\pi}). \tag{2.10}$$

식 (2.2)부터 (2.5)에 설명한 네 종류의 결정계수 값을 구하기 위하여 Liao와 McGee (2003)는 R로 작성된 함수를 http://www.geocities.com/jg_liao/software에 제공하였고, 우리는 식 (2.7)부터 (2.10)까지에서 언급한 네 종류의 SSR로부터 $SSR(X_2)$와 $SSR(X_2|X_1)$값을 구할 수 있다. 그러므로 회귀모형에서 정의한 식 (1.2)와 같은 suppression을 로지스틱회귀에서도 정의할 수 있다.

3. 모의실험

3.1. 연구방법

2절에서 제시한 일반적인 로지스틱 회귀모형 (2.1)에서의 suppression의 정의를 (1.2)식과 같이 정의할 수 있으나, suppression의 관계를 (1.3)식과 같이 유도하기는 쉽지 않다. 왜냐하면 로지스틱 회귀모형의 특성상 SSR을 여러 가지 형태로 정의할 수 있으나, 이항반응변수가 포함된 각각의 SSR을 (1.3)식과 같이 상관계수로 표현할 수 없기 때문이다. 따라서 본 절에서는 (2.1)식의 로지스틱 회귀모형을 따르는 자료를 몬테칼로 방법을 이용하여 생성하여 어떤 상황에서 suppression이 발생하는지에 대하여 탐색하고자 한다.

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}는 -3부터 3까지 간격 2.0의 크기로 변화시키고, 모분산 $\sigma^2_{x_1}$, $\sigma^2_{x_2}$는 1, 4, 9, 16으로 변화시킨다. 그리고 두 설명변수사이의 모상관계수 $\rho_{x_1x_2}$는 -0.8부터 0.8까지 간격 0.2의 크기로 변화시키면서 설명변수를 생성한다. 생성시킨 x_{1i}과 x_{2i}를 이용하여 다음의 식을 통하여 비율 π_i를 생성한다.

$$\pi_i = \exp(\alpha + \beta_1 x_{1i} + \beta_2 x_{2i})/(1 + \exp(\alpha + \beta_1 x_{1i} + \beta_2 x_{2i})),$$

여기에서 α값의 변화에 회귀제곱합(SSR)의 변동이 민감하지 않으므로 α =-1로 고정시키고 β_1과 β_2는 -3에서 3까지 변화시키면서 구한 비율 π_i의 베르누이 분포를 이용하여 이항변수 z_i들을 각각 생성한다. 다음으로는 생성된 x_{1i}, x_{2i}, z_i들로 구성된 자료를 로지스틱 회귀모형으로 분석하여 (2.7)식부터 (2.10)식에서 정의한 네 종류의 회귀제곱합(SSR)을 구한 후, 회귀제곱합들의 크기 비교로 suppression 발생여부를 살펴본다. 여기에서 모수 β_1, β_2, μ_{x_1}, μ_{x_2}, $\sigma^2_{x_1}$, $\sigma^2_{x_2}$, $\rho_{x_1x_2}$값의 변화에 따라 suppression 발생여부에 어떤 영향을 주는지 살펴보았다.

Sharpe와 Roberts(1997)가 연구한 선형회귀모형의 suppression 관계식 (1.3)식에서 γ는 반응변수 Y와 설명변수 X_1, X_2와의 상관계수의 함수로 나타난다. 로지스틱 회귀모형에서는 반응변수 Z가 연속형 변수가 아닌 이산형 변수인 이항변수이므로 보편적이고 일반적인 상관계수 통계량을 사용하는데 무리가 있을 수 있다. 그러나 2×2 분할표 자료의 연관성측도(measure of association)들 중에서 Bishop, Fienberg와 Holland(1975, pp. 380)는 두 개의 이항변수들의 상관계수를 각 변수의 첫번째 수준에서는 0, 두번째 수준에는 1의 값을 부여하여 Pearson 상관계수와 동일한 방법으로 사용하였다. 따라서 본 연구에서의 결과를 Sharpe와 Roberts(1997)의 연구결과와 비교하기 위해서 Z와 X_1, X_2와의 상관계수 r_{zx_1}, r_{zx_2}는 Pearson 상관계수 통계량을 사용하고자 한다.

3.2. Suppression의 발생결과

Sharpe와 Roberts(1997)는 suppression이 발생하는 조건을 $r_{x_1x_2}$와 $\gamma = r_{zx_1}/r_{zx_2}$의 관계하에 나타낸 그림을 통하여 설명하였듯이 본 연구에서도 로지스틱 회귀모형에서 suppression 발생의 조건을 살펴보기 위하여 $r_{x_1x_2}$와 r_{zx_1}, r_{zx_2} 그리고 γ와의 관계에 따른 그림을 시각적으로 표현하였다. γ와의 관계를 살펴보기에 앞서 γ를 구성하고 있는 r_{zx_1}와 r_{zx_2} 각각에 대하여 $r_{x_1x_2}$와의 관계 하에서 suppression 발생여부를 살펴보고자 하는데 그림 3.1과 그림 3.2는 발생하는 경우는 'O'로 발생하지 않는 경우에는 'X'로 표현하였다. 2절에서 논의한 네 종류의 SSR 중에서 (2.7)식의 SSR_l을 기준으로 한 suppression 발생 여부에 대하여 먼저 살펴보기로 하자. 그림 3.1에서 살펴보면, $r_{x_1x_2}$와 r_{zx_1}의 관계에서는 공통적으로 suppression이 발생하는 지역은 $|r_{zx_1}|$이 큰 값을 가지며 $|r_{x_1x_2}|$가 0에 가까운 값을 가질 경우에 suppression은 더 많이 발생하였다. $r_{x_1x_2}$와 r_{zx_2}의 관계를 나타낸 그림 3.2를 살펴보면, suppression이 발생하는 지역은 원점을 중심으로 다이아몬드 형태로 나타나는 것을 파악할 수 있다. 흥미로운 것은 suppression이 발생하는 지역보다 발생하지 않는 지역이다. 발생하지 않는 지역들은 공통적으로 네 모서리 부분에 집중되어 있는 것을 볼 수 있다. 그림

3.1과 그림 3.2에서도 이와 유사한 현상이 발생하였다. 특히 그림 3.2에서는 suppression이 발생하지 않는 지역이 원점을 중심으로 'X'자 형태로 나타났으며 특히 네 모퉁이부분에 집중되어 있음을 파악할 수 있다. r_{zx_1}와 r_{zx_2} 각각에 대하여 $r_{x_1x_2}$와의 관계 하에서 나타난 suppression 발생여부에 대한 현상은 모분산이 큰 값을 가질 때 더욱 확실하게 나타났다. 즉 $r_{x_1x_2}$의 절대값이 크고, r_{zx_1}와 r_{zx_2}의 절대값이 모두 큰 경우에 suppression이 발생하지 않는다. 상관계수들 간의 관계 하에서 suppression의 발생여부에 대하여는 SSR_l뿐만 아니라 SSR_o, $SSR_{l,adj}$ 그리고 $SSR_{o,adj}$ 모두 유사한 결과를 나타내고 있다. 따라서 다른 종류의 SSR에 대한 설명과 그림은 생략한다.

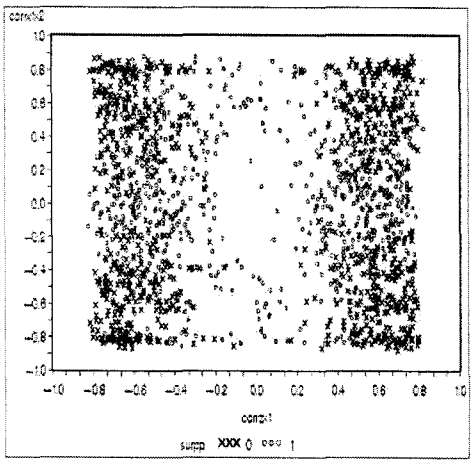

그림 3.1: $r_{x_1x_2}$와 r_{zx_1}의 그래프($\sigma^2_{x_1} = 4$)

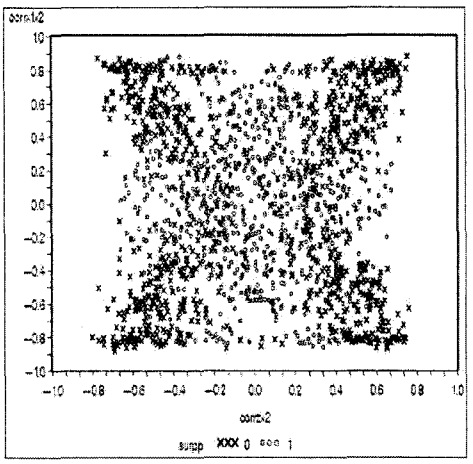

그림 3.2: $r_{x_1x_2}$와 r_{zx_2}의 그래프($\sigma^2_{x_1} = 4$)

다음으로 r_{zx_1}와 r_{zx_2}의 비율로 정의하는 γ와 x_{1i}와 x_{2i}의 상관계수 $r_{x_1x_2}$의 관계 하에서 suppression의 발생여부를 살펴보자. suppression이 발생하는 지역은 전체적으로 넓게 산포하고 있는 반면, 발생하지 않는 지역은 $|r_{x_1x_2}|$이 크고 $|\gamma|$ 값이 작은 지역에 집중되고 있는 것을 볼 수 있다. 그림 3.3부터 그림 3.8까지는 γ와 $r_{x_1x_2}$관계 하에서 suppression이 발생하지 않은 경우만을 'X'로 표현하였으며, 네 종류의 $SSR(SSR_l, SSR_o, SSR_{l,adj}, SSR_{o,adj})$을 사용하여 (2.6)식과 같은 로지스틱 회귀모형에서의 suppression 정의로부터 상관계수들의 함수로 나타나는 (1.3)식과 같은 관계식을 유도할 수 없기 때문에 선형회귀에서 suppression 관계식 (1.3)을 그림에 보조선으로 추가하였다. SSR_l에 대하여 suppression이 발생하지 않는 경우만 표현한 그림 3.3과 그림 3.4를 살펴보면, $r_{x_1x_2}$가 양의 값을 가진 경우 γ는 +1의 값 근처에 집중되어 있으며, $r_{x_1x_2}$가 음의 값을 가진 경우에는 γ는 -1의 값 근처에 집중되어 있다. 이것은 선형회귀에서 suppression 발생관계식 (1.3)과 그림 1.1에서 논의되고 설명된 결과와 매우 유사함을 파악할 수 있다. 그림 3.3과 그림 3.4는 각각 $\sigma^2_{x_1}$이 1과 16인 경우인데, $\sigma^2_{x_1}$의 값이 클수록 suppression이 발생하지 않는 지역이 보다 산포되어 나타난다는 것을 그림 3.3과 그림 3.4를 비교하면 알 수 있다. 로지스틱 회귀모형에서 발생하지 않는 지역이 1, 3사분면에 집중하고 있으며 원점을 중심으로 대칭적으로 나타난다는 사실은 선형

회귀모형에서 suppression이 발생하지 않는 영역을 시각적으로 표현한 그림 1.1과 유사하다는 결론을 내릴 수 있다. 그리고 이러한 결론은 통계량 SSR_o, $SSR_{l,adj}$, $SSR_{o,adj}$를 사용한 모의실험에서도 유사한 결론을 유도할 수 있다. 그림 3.5와 그림 3.6은 SSR_o를 기준으로 한 $\sigma^2_{x_1}$이 각각 1과 16일 경우이고 그림 3.7과 그림 3.8은 $\sigma^2_{x_1}$이 16일 때 각각 $SSR_{l,adj}$과 $SSR_{o,adj}$를 기준으로 한 경우이다.

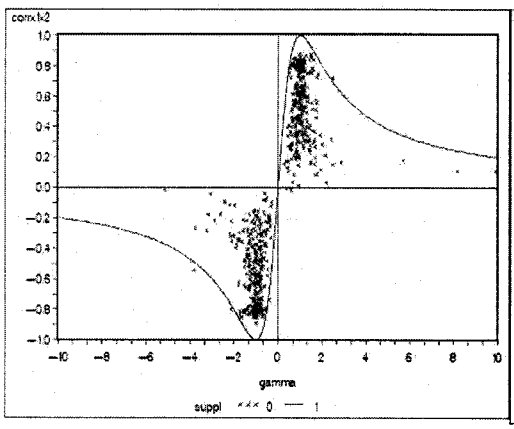

그림 3.3: Suppression미발생(SSR_l,$\sigma^2_{x_1} = 1$)

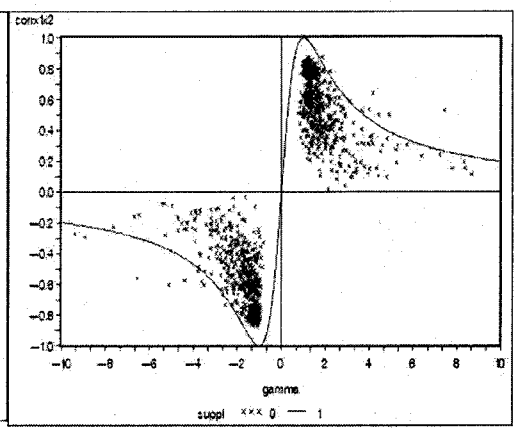

그림 3.4: Suppression미발생(SSR_l,$\sigma^2_{x_1} = 16$)

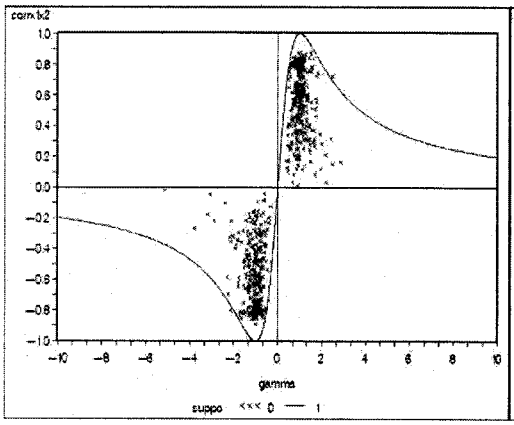

그림 3.5: Suppression미발생(SSR_o,$\sigma^2_{x_1} = 1$)

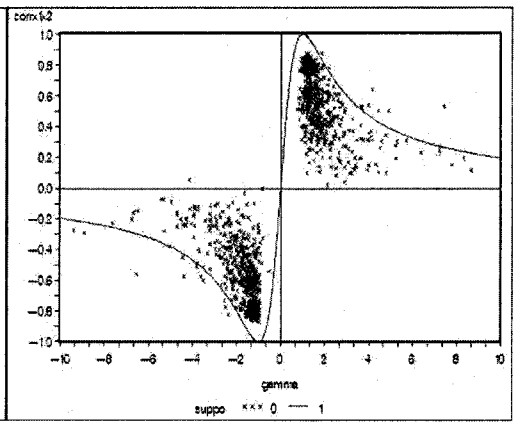

그림 3.6: Suppression미발생(SSR_o,$\sigma^2_{x_1} = 16$)

4. 결론

통계량 SSR_l과 SSR_o를 사용하여 suppression이 발생하지 않은 경우를 표현한 그림 3.3부터 그림 3.6까지를 비교하여 보면 큰 차이점을 발견할 수 없고 매우 유사한 현상이 나

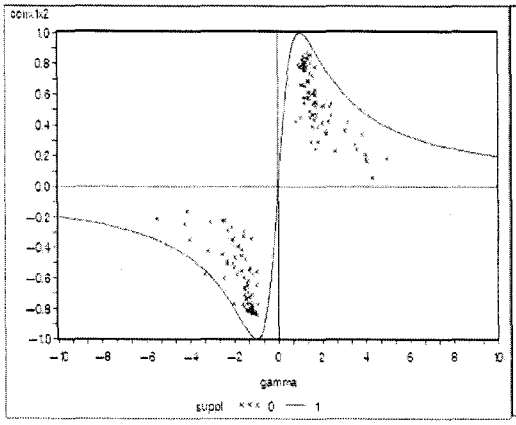

그림 3.7: Suppression미발생 ($SSR_{l,adj}$)

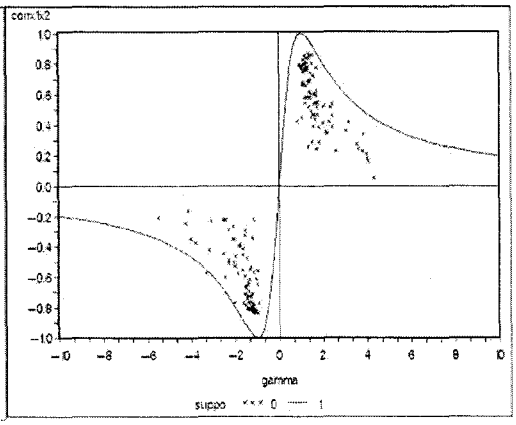

그림 3.8: Suppression미발생 ($SSR_{o,adj}$)

타남을 파악할 수 있다. 그리고 $SSR_{l,adj}$와 $SSR_{o,adj}$ 통계량을 사용하여 구한 결과를 표현한 그림 3.7과 그림 3.8도 매우 유사하다는 것을 발견할 수 있으나, 동일한 분산 $\sigma^2_{x_1} = 16$일 때 SSR_l과 SSR_o에 대한 그림 3.4, 그림 3.6과 비교하면 suppression이 발생하지 않는 지역의 산포가 넓게 퍼져있지 않으며 경계선(보조선)에 근접하거나 지나친 부분이 적기 때문에 매우 안정적이라고 할 수 있다. 따라서 Liao와 McGee(2003)가 제안한 $SSR_{l,adj}$과 $SSR_{o,adj}$ 통계량이 SSR_l과 SSR_o 통계량보다 로지스틱 회귀모형에서 suppression을 설명하는 데 민감하지 않으며 안정적이다고 판단할 수 있다.

$\sigma^2_{x_1} = 1$인 상황에서 여러 모수의 주어진 값을 조합하여 총 2304경우의 자료를 발생하였는데 통계량 SSR_o값을 구할 수 없는 경우가 216건으로 9.38%인데 반해, 통계량 SSR_l값을 구할 수 없는 경우가 SSR_o보다 상대적으로 많은 1015건이 되어 (1015/2304=0.4405) 약 44% 경우에 SSR_l값을 구할 수 없었다. 그리고 $\sigma^2_{x_1} = 16$으로 분산이 커진 상황에서는 통계량 SSR_l값을 구할 수 없는 경우가 215/2304=0.0933으로 약 10%미만으로 줄어든 사실을 파악할 수 있다. 통계량 SSR_l과 SSR_o을 이용하여 suppression 발생여부의 결과를 요약한 것이 표 4.1과 표 4.2에 정리되어 있는데 회귀제곱합을 구할 수 있는 경우를 조건으로 서로 상반되는 결과를 나타내는 비율은 $\sigma^2_{x_1} = 1$인 경우에 $(17+17)/(2304-1015)$=0.0264로 약 2.6% 정도이며, $\sigma^2_{x_1} = 16$인 경우에는 $(18+21)/(2304-215)$=0.0187으로 1.8% 정도로 감소된다. SSR_l과 SSR_o를 구하는 데도 많은 시간이 소요되나 Liao와 McGee(2003)가 제안한 $SSR_{l,adj}$과 $SSR_{o,adj}$를 구하는 데는 약 5배 정도의 시간이 더 필요하다. 따라서 보다 명확한 결과가 나타나는 분산 $\sigma^2_{x_1}$이 16인 경우에만 모의실험을 실시하여 표 4.3에 나열하였다. 모수조합의 종류도 축소하여 총 313건의 경우만 살펴보았는데, 회귀제곱합을 구하는 경우에 대하여 발생여부가 서로 상반되는 결과를 나타내는 비율은 $(1+1)/(313-55) = 0.0078(0.78\%)$으로 이는 SSR_l과 SSR_o의 경우보다 감소한 것이다. $SSR_{l,adj}$과 $SSR_{o,adj}$인 경우의 suppression 발생결과는 SSR_l과 SSR_o에 비해 더욱 명확하고 안정적이라고 판단할 수 있다.

표 4.1: SSR_l과 SSR_o의 suppression($\sigma^2_{x_1} = 1$)

		SSR_o			
		미발생	발생	NA	전체
SSR_l	미발생	619	17		636
	발생	17	636		653
	NA	372	427	216	1015(44.1%)
	전체	1008	1080	216	2304

표 4.2: SSR_l과 SSR_o의 suppression($\sigma^2_{x_1} = 16$)

		SSR_o			
		미발생	발생	NA	전체
SSR_l	미발생	1214	21		1235
	발생	18	836		854
	NA	34	166	15	215(9.3%)
	전체	1266	1023	15	2304

표 4.3: $SSR_{l,adj}$과 $SSR_{o,adj}$의 suppression($\sigma^2_{x_1} = 16$)

		$SSR_{o,adj}$			
		미발생	발생	NA	전체
$SSR_{l,adj}$	미발생	159	1		160
	발생	1	97		98
	NA	13	41	1	55(17.6%)
	전체	173	139	1	313

본 연구에서는 로지스틱 회귀모형에서 선호되는 네 종류의 SSR을 유도하여, 각각의 SSR로 표현되는 suppression의 정의에 따라 suppression이 발생하는 영역을 살펴보면서 다음과 같은 결론을 유도하였다. 첫째, 모분산 $\sigma^2_{x_1}$이 증가할수록 상관계수들의 관계가 명확하게 표현되기 때문에, 회귀제곱합 통계량 SSR_l과 SSR_o을 구할 수 있는 경우는 증가한다($\sigma^2_{x_1}$이 1에서 16으로 증가할 때 55.9%에서 90.7%로 증가). 또한 SSR_l과 SSR_o를 사용하여 suppression이 발생하는 결과가 상반되는 경우의 비율은 줄어든다(2.6%에서 1.8%로 감소). 이 현상은 $SSR_{l,adj}$과 $SSR_{o,adj}$의 경우도 유사하다. 둘째, Liao와 McGee(2003)가 제안한 $SSR_{l,adj}$과 $SSR_{o,adj}$를 사용하여 suppression 발생결과를 살펴보면, 값을 구할 수 없는 경우($SSR_{l,adj}$의 NA)가 줄어들고 있다고 단언할 수는 없으나 $SSR_{l,adj}$과 $SSR_{o,adj}$에 대한 suppression이 발생하는 결과가 서로 일치하지 않는 경우의 비율은 감소한다(1.8%에서 0.78%로 감소). 그러므로 설명력의 과대추정을 수정한 $SSR_{l,adj}$과 $SSR_{o,adj}$ 통계량을 사용하면 계산시간이 많이 소요되는 단점에도 불구하고, 선형 회귀모형에서 suppression 관계를 나타낸 그림 1.1과 관계식 (1.3)과 비교해 보면 비교적 안정적인 결과를 얻을 수 있다고

결론내릴 수 있다.

참고문헌

홍종선 (2004). Suppression and collapsibility for log-linear models, *The Korean Communications in Statistics*, **11**, 519-527.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis*, Cambridge, Massachusetts: MIT Press.

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

Conger, A. J. (1974). A revised definition for suppressor variables: A guide to their identification and interpretation, *Educational and Psychological Measurement*, **34**, 35-46.

Freud, R. J. (1988). when is $R^2 > r_{yx_1} + r_{yx_2}$(Revisited), *The American Statistician,*, **42**, 89-90.

Hamilton, D. (1987). Sometimes $R^2 > r_{yx_1} + r_{yx_2}$ correlated variables are not always redundant, *The American Statistician,*, **41**, 129-132.

Hamilton, D. (1988). Reply to Freund and Mitra, *The American Statistician,*, **42**, 90-91.

Horst, P. (1941). The role of prediction variables which are independent of the criterion, *in The Prediction of Personal Adjustment*, ed. P. Horst, New York: Social Science Research Council, 431-436.

Kvalseth, T. O. (1985). Cautionary note about R^2, *The American Statistician*, **39**, 279-285.

Liao, J. G. and McGee, D. (2003). Adjusted coefficients of determination for logistic regression, *The American Statistician*, **57**, 161-165.

Lynn, H. S. (2003). Supperssion and confounding in action, *The American Statistician*, **57**, 58-61.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis, *The American Statistician*, **54**, 17-24.

Mitra, S. (1988). The relationship between the multiple and the zero-order correlation coefficients, *The American Statistician*, **42**, 89.

Mittlbock, M. and Schemper, M. (1996). Explained variation for logistic regression, *Statistics in Medicine*, **15**, 1987-1997.

Schey, H. M. (1993). The relationship between the magnitudes of $SSR(x_2)$ and $SSR(x_2|x_1)$: A geometric description, *The American Statistician*, **47**, 26-30.

Sharpe, N. R. and Roberts, R. A. (1997). The relationship among sums of squares, correlation coefficients, and suppression, *The American Statistician*, **51**, 46-48.

Velicer, W. F. (1978). Suppressor variables and the semipartial correlation coefficient, *Educational and Psychological Measurement*, **38**, 953-958.

[2004년 12월 접수, 2005년 7월 채택]

Suppression for Logistic Regression Model

C. S. Hong[1] H. I. Kim[2] J. H. Ham[3]

ABSTRACT

The suppression for logistic regression models has been debated no longer than that for linear regression models since, among many other reasons, sum of squares for regression (SSR) or coefficient of determination (R^2) could be defined into various ways. Based on four kinds of R^2's: two kinds are most preferred, and the other two are proposed by Liao & McGee (2003), four kinds of SSR's are derived so that the suppression for logistic models is explained. Many data fitted to logistic models are generated by Monte Carlo method. We explore when suppression happens, and compare with that for linear regression models.

Keywords: Bias, Coefficient of determination, Inherent prediction error, Log-linear model Logit model, SSR.

1) Professor, Department of Statistics, Sungkyunkwan University, 3-53 Myeongnyun-Dong, Jongno-Gu, Seoul 110-745, Korea. E-mail: cshong@skku.ac.kr
2) CSC, KYOBO Auto Insurance Co., 1303-22 Seocho-Dong, Seocho-Gu, Seoul 137-070, Korea
3) Marketing Business, FUJI XEROX, 120-20 Seosomun-Dong, Jung-Gu, Seoul 100-736, Korea.