Journal of the Korean Data & Information Science Society
2010, **21**(1), 21–32

한국데이터정보과학회지

로지스틱모형에서 그래픽을 이용한 회귀와 모형평가[†]

강명욱[1] · 김부용[2] · 홍주희 [3]

[123]숙명여자대학교 통계학과

접수 2009년 10월 15일, 수정 2009년 12월 20일, 게재확정 2009년 12월 28일

요 약

그래픽적 회귀는 모형에 대한 가정을 하지 않고 회귀정보를 모두 포함하는 충분요약그림을 찾아내는 분석 방법으로 모든 회귀정보를 저차원의 그림으로 표현할 수 있게 하는 데에 그 목적이 있다. 잔차산점도를 이용한 모형의 평가는 적용 범위가 선형회귀모형에 국한되는 문제점이 있기 때문에 일반화선형모형에서는 그 대안으로 주변모형 산점도를 이용하여 모형의 적절성을 평가한다. 본 논문에서는 일반화선형모형 중에서 이진반응변수를 갖는 로지스틱모형에서의 그래픽적 회귀 방법과 주변모형 산점도를 이용한 모형평가 방법을 알아본다.

주요용어: 그래픽적 회귀, 로지스틱모형, 이항회귀, 주변모형 산점도, 차원축소.

1. 서 론

일반적으로 통계분석방법은 자료가 가지고 있는 정보를 하나의 숫자로 표현하는 요약통계량에 의존한다. 반면 그래픽적 회귀 (graphical regression)는 자료를 통해 얻을 수 있는 모든 정보를 분포에 대한 가정 없이 그림으로 나타내는 그래픽적인 접근이며 회귀정보 (regression information)를 모두 포함하는 충분요약그림 (sufficient summary plot)을 찾아내는 방법을 제시한다.

그래프를 이용한 회귀분석은 Ezekiel (1924)에 의해 처음 시도되었고, Cook과 Weisberg (1982), Chambers 등 (1983), Atkinson (1985), Cleveland (1987)에 의해 회귀진단에 사용되었다. 또한 Kahng (2005)에 의해 일반화선형모형에서 그래프를 이용한 분석이 시도되었다. Cook과 Weisberg (1994)가 그래픽적 회귀를 소개한 이후에 Cook (1998)은 이에 대한 수리적이고 정밀한 회귀분석 방법을 제시했다. 또한 Cook과 Weisberg (1999)는 그동안 제시된 그래픽적 회귀의 방법론을 종합적으로 정리하였다.

그래픽적 회귀를 통해 얻은 모형의 적절성에 대한 평가 역시 그래픽적으로 접근 할 수 있다. 일반적으로 선형회귀모형의 적절성을 평가하는 도구로 잔차산점도가 널리 이용되고 있으나 일반화선형모형의 적절성을 평가하기에는 부적합하다. Cook과 Weisberg (1997)는 잔차산점도의 대안으로써 주변모형 확인조건에 기초한 주변모형 산점도 (marginal model plot)를 제안하였다.

본 연구에서는 일반화선형모형 중에서 특히 이진반응변수를 가진 로지스틱모형에서의 Cook과 Weisberg (1999)이 제안한 그래픽적 회귀에 대한 방법을 확장하여 제시하고 로지스틱모형의 적절성을 평가하는 방법 중에서 그래픽적인 방법으로 주변모형 산점도를 이용한 모형평가 방법을 제시하고자 한다.

† 본 연구는 숙명여자대학교 2008년도 교내연구비 지원에 의해 수행되었음.
1 교신저자: (140-742) 서울특별시 용산구 청파동2가 53-12, 숙명여자대학교 통계학과, 교수.
E-mail: mwkahng@sm.ac.kr
2 (140-742) 서울특별시 용산구 청파동2가 53-12, 숙명여자대학교 통계학과, 교수.
3 (140-742) 서울특별시 용산구 청파동2가 53-12, 숙명여자대학교 통계학과, 대학원생.

2. 선형회귀모형에서의 그래픽적 회귀

2.1. 그래픽적 회귀

그래픽적 회귀는 설명변수를 이루는 공간의 차원축소 (dimension reduction)와 축소된 차원의 요약그림 (summary plot)을 통한 회귀정보의 설명을 기본개념으로 한다. 반응변수 y와 p개의 설명변수로 이루어진 벡터 $\boldsymbol{x} = (x_1, \cdots, x_p)^T$를 고려하자. p차원인 설명변수 $\boldsymbol{x}$가 주어졌을 때 반응변수 y의 조건부 분포가 p보다 작은 k차원의 새로운 변수 $H\boldsymbol{x} = (\boldsymbol{\eta}_1^T\boldsymbol{x}, \cdots, \boldsymbol{\eta}_k^T\boldsymbol{x})^T$가 주어졌을 때 y의 조건부 분포와 같은 경우를 다음과 같이 표현할 수 있다.

$$F(y|\boldsymbol{x}) = F(y|H\boldsymbol{x}). \tag{2.1}$$

여기서 $H = (\boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_k)^T$는 $k \times p$행렬, $\boldsymbol{\eta}_i^T$는 행렬 H의 i번째 행벡터이고 $F(y|\boldsymbol{x})$와 $F(y|\boldsymbol{\eta x})$는 y의 조건부 누적분포함수 (conditional cumulative distribution function)를 나타낸다.

그래픽적 회귀의 기본개념은 식 (2.1)을 만족하는 $\boldsymbol{H}$를 찾을 수 있다면 주어진 p차원인 설명변수 $\boldsymbol{x}$대신 축소된 k차원의 새로운 설명변수 $\boldsymbol{Hx}$에 의해 회귀정보를 설명할 수 있다는 것이다. 따라서 $(p+1)$차원의 산점도로 설명되어야할 회귀가 정보의 손실이 없이 축소된 $(k+1)$차원의 산점도로 충분히 설명이 된다는 것이다. 이러한 축소된 차원의 산점도를 충분요약그림이라고 하며 이러한 산점도를 찾는 것이 그래픽적 회귀의 목적 중의 하나가 된다.

설명변수가 하나인 단순회귀의 그래픽적인 분석은 반응변수를 수직축으로 하고 설명변수를 수평축으로 하는 2차원 산점도에 기초하고 있고 이 산점도가 곧 충분요약그림이 된다. 설명변수가 2개인 경우에는 반응변수를 수직축인 V축 (vertical-axis)으로 하고, 설명변수를 V축에 수직이며 서로 직각인 두 개의 축인 H축 (horizontal-axis)과 O축 (out-of-page axis)축으로 하는 3차원 산점도가 충분요약그림이 된다. 통계분석의 일반적인 목적은 고차원의 정보를 가능한 간단하게 차원을 축소하는데 있다. 따라서 모든 회귀정보를 2차원 산점도가 포함하고 있다면 그것이 충분요약그림이 된다. 2차원의 충분요약그림을 찾기 위한 시각적인 방법은 회전을 시키는 것이다. 3차원 그림을 수직축인 V축을 중심으로 θ만큼 회전시키면 설명변수의 선형결합으로써 $h(\theta) = b(cos\theta)x_1 + c(sin\theta)x_2$가 화면상에서 수평축으로 보이는 수평 스크린 축 (horizontal screen axis)이 되는 2차원 그림을 얻을 수 있다. 이 때 b와 c는 임의의 상수이며 반응변수가 오직 선형결합 $h(\theta)$에만 의존한다면 이러한 2차원 그림은 충분요약그림이 될 것이다. 그러나 설명변수가 2보다 많은 p개인 경우, $(p+1)$차원의 산점도를 구현하기는 매우 힘들다. 그러므로 정보의 손실 없이 우리가 인지할 수 있는 2차원이나 3차원으로 차원을 축소하는 것이 그래픽적 회귀의 핵심이다.

2.2. 차원구조

정보의 손실 없이 회귀의 특성을 설명하는데 필요한 최소한의 설명변수들의 선형결합의 수를 차원구조 (structural dimension)라고 한다. 만약 반응변수 y가 설명변수 $\boldsymbol{x}$에 의존하지 않으면 0차원구조이다. 이 때 충분요약그림은 반응변수 y의 히스토그램이 되며, 평균함수 $E(y|\boldsymbol{x})$와 분산함수 $Var(y|\boldsymbol{x})$는 모두 상수가 된다.

만약 반응변수 y가 오직 선형결합 $\boldsymbol{\beta}^T\boldsymbol{x}$를 통해서만 설명변수 $\boldsymbol{x}$에 의존하면 1차원구조이고, 이 때 충분요약그림은 반응변수를 수직축으로 하고 $\boldsymbol{\beta}^T\boldsymbol{x}$를 수평축으로 하는 산점도가 되며, 평균함수와 분산함수는 모두 하나의 선형결합 $\boldsymbol{\beta}^T\boldsymbol{x}$에 의존한다. 회귀의 특성을 설명하는데 두 개의 선형결합 $\boldsymbol{\beta}_1^T\boldsymbol{x}$와 $\boldsymbol{\beta}_2^T\boldsymbol{x}$가 필요하면 2차원구조이다. 충분요약그림은 반응변수를 V축으로 하고 $\boldsymbol{\beta}_1^T\boldsymbol{x}$와 $\boldsymbol{\beta}_2^T\boldsymbol{x}$을 각각 H축과 O축으로 하는 3차원 산점도가 되며 평균함수와 분산함수는 두 개의 선형결합에 의존한다.

2.3. Arc를 이용한 그래픽적 회귀

그래픽적 회귀를 위한 차원축소를 위해서는 설명변수간의 선형성 조건을 만족해야한다 (Cook, 1998). 설명변수들을 짝을 지어 만든 산점도행렬 (scatter plot matrix)을 이용하여 선형성 조건을 점검할 수 있고 만약 이러한 조건이 만족되지 않으면 설명변수의 변환을 시도할 수 있다.

반응변수 y와 선형의 관계에 있는 p개의 설명변수 $\boldsymbol{x} = (x_1, \cdots, x_p)^T$를 갖는 회귀를 가정하자. 우리는 3개 이상의 변수에 대해서는 산점도를 한 번에 그릴 수가 없다. 그러므로 변수를 짝지어 살펴보는 방법을 생각할 수 있다. 반응변수가 y이고 p개의 선형관계에 있는 설명변수를 $\boldsymbol{x} = (x_1, \cdots, x_p)^T$라고 했을 때 그래픽적 회귀의 핵심은 변수 쌍 (x_1, x_2)를 정보의 손실 없이 식 (2.2)의 선형결합으로 대체 하는 것이다.

$$x_{12} = b_1 x_1 + b_2 x_2. \tag{2.2}$$

그래픽적 회귀는 설명변수의 변수 쌍을 식 (2.2)를 이용하여 결합하는 연속적인 절차이며 결합이 더 이상 불가능해 질 때까지 이 절차를 반복한다. 두 변수의 선형결합 가능여부는 3차원 추가변수그림 (added variable plot)을 이용하여 확인한다.

결합할 2개의 설명변수 x_1과 x_2를 제외한 나머지 $(p-2)$개의 설명변수를 $\boldsymbol{x}_3$라고 하자. 3차원 추가변수그림은 y를 $\boldsymbol{x}_3$에 적합 시킨 후의 잔차 $e(y|\boldsymbol{x}_3)$을 V축으로 하고 x_1과 x_2를 각각 $\boldsymbol{x}_3$에 적합 시킨 후의 잔차 $e(x_1|\boldsymbol{x}_3)$과 $e(x_2|\boldsymbol{x}_3)$을 각각 H축과 O축으로 하는 산점도이다. 만약 3차원 추가변수그림에서 0차원구조의 징후를 보인다면 회귀계수는 $b_1 = b_2 = 0$이므로 설명변수 x_1과 x_2는 영향력이 없다고 판단되어 삭제할 수 있다. 만약 3차원 추가변수그림에서 1차원구조를 갖는다고 판단되면 x_1과 x_2는 x_{12}로 결합할 수 있다. 2차원구조를 갖는다고 판단된다면 x_1과 x_2는 결합할 수 없으므로 다른 설명변수 쌍을 선택하여 이와 같은 절차를 반복한다.

Xlisp-Stat (Tierney, 1990) 언어에 기초한 Arc를 사용하면 그래픽적 회귀를 편리하게 수행할 수 있다. Arc는 웹 (http://www.stat.umn.edu/arc/software.html)에서 무료로 얻을 수 있다. 그래픽적 회귀에서는 선형의 관계에 있는 설명변수들이 필요하다. 이 조건을 만족시키기 위해서 원래의 설명변수 $\boldsymbol{x}$을 이용하여 생성한, 서로 상관관계가 없는 p개의 새로운 선형결합으로 대체할 수 있다. Arc에서는 p개의 무상관 (uncorrelated)의 선형결합을 GREG 설명변수라고 부르며 다음과 같은 특성을 갖는다. 가장 먼저 Fit이라는 선형결합을 생성한다. Fit은 최소제곱적합값과 절편인 $\widehat{\beta}_0$의 값만큼 차이가 난다. 즉, $Fit = \widehat{\boldsymbol{\beta}}^T \boldsymbol{x} = \widehat{y} - \widehat{\beta}_0$이다. 만약 1차원구조를 갖고 설명변수들이 선형의 관계에 있다면 반응변수를 수직축으로 하고 Fit을 수평축으로 하는 산점도가 충분요약그림이 된다. 그리고 만약 2차원이상의 구조를 갖는다면 Fit은 회귀정보를 설명하는데 필요한 선형결합들 중의 하나가 된다. 그 다음으로 생성되는 선형결합은 $gr1 = \boldsymbol{\beta}_1^T \boldsymbol{x}$이다. $gr1$은 첫 번째 선형결합인 Fit과는 무상관의 관계에 있다. 만약 2차원구조를 갖는 회귀라면 $gr1$은 Fit이외에 필요한 두 번째 선형결합의 좋은 후보가 된다. 세 번째로 생성되는 선형결합 $gr2 = \boldsymbol{\beta}_2^T \boldsymbol{x}$는 Fit과 $gr1$에 상관관계가 없는 선형결합이다. 나머지 선형결합도 이와 같은 방법으로 앞에서 생성된 선형결합들과는 서로 무상관의 관계에 있도록 생성된다.

3. 주변모형 산점도

선형회귀모형의 적절성 평가를 위한 도구로 잔차산점도가 널리 이용되고 있다. 모형이 적절하다면 잔차산점도의 수직축을 이루는 잔차와 수평축을 이루는 설명변수들의 선형결합 $\boldsymbol{a}^T \boldsymbol{x}$가 서로 독립적인 것으로 나타나야 한다는 것이 잔차산점도를 이용한 모형평가 방법의 기본 개념이다. 그러나 비선형회귀모형이나 대부분의 일반화선형모형에서 잔차산점도를 이용한 모형 평가 방법은 성공적이지 못하다. 4절에서 설명할 반응변수가 0 또는 1인 이항회귀 (binomial regression)에서 잔차산점도는 모형의 적절성과

는 관계없이 특정한 패턴을 갖게 된다. 잔차산점도를 이용하여 모형을 평가하는 방법의 적용 범위가 선형회귀모형에 국한되는 문제점이 있기 때문에 그 대안으로 주변모형 산점도를 이용하여 모형의 적절성을 평가한다.

주변모형 산점도의 기본 개념은 반응변수 y와 p개의 설명변수로 이루어진 벡터 $\boldsymbol{x}$를 가진 회귀모형의 특성을 다음의 두 가지 관점에서 생각한 조건부 누적밀도함수를 이용하여 비교하는 것이다. 첫째, 자료 $(y_i, \boldsymbol{x}_i^T)$, $i = 1, \cdots, n$가 독립적이고 같은 분포를 갖는다고 가정하고 모형에 대한 구체적인 가정 없이 오직 자료에서 얻어지는 미지의 누적밀도함수 $F(y|\boldsymbol{x})$를 추론한다. 둘째, 일반적인 의미의 회귀모형을 구체적으로 가정한 후에 회귀모형으로부터 형성되는 조건부 누적밀도함수 $M(y|\boldsymbol{\beta}, \boldsymbol{x})$를 추론한다.

모형의 평가는 Azzalini와 Bowman (1993)에 의해 제안된 검정법을 사용하는 수치적인 방법과 자료의 범위에서 두 개의 곡선을 비교하는 시각적인 방법이 있으나 본 논문에서는 시각적인 방법에 대해서만 언급하겠다. 구체적인 모형 가정 하에서의 조건부 누적밀도함수 $M(y|\boldsymbol{\beta}, \boldsymbol{x})$을 추론할 때 모수 대신 추정값을 대입하는 것은 대표본에서는 크게 문제되지 않기 때문에 모수 $\boldsymbol{\beta}$대신 모형하에서의 일치추정값 $\widehat{\boldsymbol{\beta}}$을 사용하고 표기의 편리를 위해서 $M(y|\widehat{\boldsymbol{\beta}}, \boldsymbol{x})$을 $M(y|\boldsymbol{x})$로 표현한다.

3.1. 주변모형 확인조건

Cook과 Weisberg (1997)는 다음과 같은 주변모형 확인조건 (marginal model checking condition)을 제시하였다. 표본공간 안에 있는 $\boldsymbol{x}$의 모든 값에 대하여 $F(y|\boldsymbol{x}) = M(y|\boldsymbol{x})$이 성립하기 위한 필요충분조건은 모든 $\boldsymbol{a}^T\boldsymbol{x}$에 대하여 $F(y|\boldsymbol{a}^T\boldsymbol{x}) = M(y|\boldsymbol{a}^T\boldsymbol{x})$인 경우이다. 이는 완전모형을 나타내는 $F(y|\boldsymbol{x})$가 내포하는 모든 정보를 주변모형 $F(y|\boldsymbol{a}^T\boldsymbol{x})$가 설명할 수 있다는 것을 의미한다. 따라서 $(p+1)$차원의 산점도 대신 반응변수를 수직축으로 하고 설명변수의 선형결합 $\boldsymbol{a}^T\boldsymbol{x}$를 수평축으로 하는 2차원 산점도를 이용하여 모형을 평가할 수 있다.

주변모형 확인 조건은 모든 주변모형이 참일 때에만 완전모형이 참이라는 것을 의미한다. 완전모형의 적절성을 평가하려면 고려해야하는 주변모형 산점도의 수가 증가한다. 그러나 모든 가능한 주변모형 산점도를 확인하는 것은 불가능하므로 방향 $\boldsymbol{a}$를 적절히 선택해야 한다. 방향 선택을 위한 몇 가지 표준적인 방법이 있지만 그 중에서 기본적으로 사용하는 두 가지 방법은 다음과 같다.

첫째, 만약 회귀모형이 선형모형인 경우, 즉 $\boldsymbol{\beta}^T\boldsymbol{x}$로 설명할 수 있다고 가정하는 경우에 $\boldsymbol{\beta}$의 최소제곱추정값 $\widehat{\boldsymbol{\beta}}$을 이용하여 $\boldsymbol{a} = \widehat{\boldsymbol{\beta}}$으로 취한다. 그러므로 y를 수직축으로 하고 $\widehat{\boldsymbol{\beta}}^T\boldsymbol{x}$를 수평축으로 하는 2차원 산점도를 이용하여 모형을 평가할 수 있다. 둘째, $\boldsymbol{a}^T\boldsymbol{x}$가 각각의 변수를 취하도록 $\boldsymbol{a}$를 정하도록 한다. 그러므로 y를 수직축으로 하고 각각의 설명변수를 수평축으로 하는 2차원 산점도를 이용하여 모형을 평가할 수 있다. 이는 각각의 변수들의 적절성을 설명하여 변수변환이나 다른 처방을 요구하는 근거를 제시한다 (Cook과 Weisberg, 1997). 또한 본 논문에서는 그래픽적 회귀에 의해 제공되는 선형결합에 대한 $\boldsymbol{a}$의 선택을 고려해본다.

3.2. 모형의 평가

모형에 대한 구체적인 가정을 하지 않은 주변평균함수 (marginal mean function) $E_F(y|\boldsymbol{a}^T\boldsymbol{x})$와 모형에 대한 구체적인 가정을 하는 주변평균함수 $E_M(y|\boldsymbol{a}^T\boldsymbol{x})$를 생각하자. y를 수직축으로 하고 $\boldsymbol{a}^T\boldsymbol{x}$를 수평축으로 하는 산점도에서 대표적인 평활 방법의 하나인 lowess (locally weighted scatterplot smoother; Cleveland와 Devlin, 1988)를 이용하여 $\widehat{E}_F(y|\boldsymbol{a}^T\boldsymbol{x})$와 $\widehat{E}_M(y|\boldsymbol{a}^T\boldsymbol{x})$를 추정할 수 있다. 또한, 모형에 대한 가정을 하지 않은 주변분산함수 (marginal variance function)를 $Var_F(y|\boldsymbol{a}^T\boldsymbol{x})$로 표현하고 모형을 가정한 주변분산함수를 $Var_M(y|\boldsymbol{a}^T\boldsymbol{x})$라고 표현하면 주변평균함수와 마찬가지로 평활법을

이용하여 $\widehat{Var}_F(y|\boldsymbol{a}^T\boldsymbol{x})$와 $\widehat{Var}_M(y|\boldsymbol{a}^T\boldsymbol{x})$를 추정할 수 있다. 또한 $SD_F(y|\boldsymbol{a}^T\boldsymbol{x}) = [\widehat{Var}_F(y|\boldsymbol{a}^T\boldsymbol{x})]^{1/2}$, $SD_M(y|\boldsymbol{a}^T\boldsymbol{x}) = [\widehat{Var}_M(y|\boldsymbol{a}^T\boldsymbol{x})]^{1/2}$이다.

주변모형 확인 조건에 기초한 모형의 평가는 y를 수직축으로 하고 $\boldsymbol{a}^T\boldsymbol{x}$를 수평축으로 하는 산점도에 $\widehat{E}_F(y|\boldsymbol{a}^T\boldsymbol{x})$, $\widehat{E}_F(y|\boldsymbol{a}^T\boldsymbol{x}) + SD_F(y|\boldsymbol{a}^T\boldsymbol{x})$, $\widehat{E}_F(y|\boldsymbol{a}^T\boldsymbol{x}) - SD_F(y|\boldsymbol{a}^T\boldsymbol{x})$의 3개의 곡선, 그리고 모형에 대한 가정이 없는 경우의 $\widehat{E}_M(y|\boldsymbol{a}^T\boldsymbol{x})$, $\widehat{E}_M(y|\boldsymbol{a}^T\boldsymbol{x}) + SD_M(y|\boldsymbol{a}^T\boldsymbol{x})$, $\widehat{E}_M(y|\boldsymbol{a}^T\boldsymbol{x}) - SD_M(y|\boldsymbol{a}^T\boldsymbol{x})$의 3개의 추정곡선을 추가한 요약그림을 통해 가능하다.

추정값과 상한, 하한을 나타내는 3쌍의 추정곡선들의 비교에서 모형의 가정이 없는 경우와 모형의 가정이 있는 경우의 추정곡선들이 근사적으로 일치하면 가정한 모형이 적절하다고 평가한다. 만약 일치하지 않으면 가정한 모형이 적절하지 않음을 나타낸다. 모형이 적절하지 않다고 판단되면 설명변수의 변환을 통하여 새로운 모형을 설정할 수 있다. 만약 반응변수가 이진변수인 경우에는 추정값에 해당하는 추정곡선인 $\widehat{E}_F(y|\boldsymbol{a}^T\boldsymbol{x})$와 $\widehat{E}_M(y|\boldsymbol{a}^T\boldsymbol{x})$만을 비교한다 (Cook과 Weisberg, 1999).

4. 로지스틱모형의 그래픽적 회귀

4.1. 이항회귀

회귀분석에서 반응변수는 일반적으로 특정한 구간 안에 있는 값이라면 어느 값이라도 가질 수 있는 연속형 자료이다. 그러나 반응변수가 이산형인 경우, 특히 몇 번의 시도 중에서 성공의 횟수를 반응변수 y로 하는 이항변수인 경우도 있다. 이런 경우의 회귀분석을 이항회귀라고 한다. 또한 y는 한 번의 시도에서 성공 혹은 실패를 나타내는 이진변수 (binary variable)인 경우도 있다.

이항확률변수 y를 성공확률이 θ인 m번의 독립적인 시행 중 성공의 횟수라고 하면, y는 이항분포 $Bin(m,\theta)$를 따른다. 특별히 $m = 1$인 경우에는 베르누이분포를 따른다고 한다. 이항분포의 평균과 분산은 θ와 m에 의존하고 시행 횟수 m이 고정되어 있는 이항분포는 성공확률 θ에 의해 결정된다. n개의 이항확률변수 y_i, $i = 1, \cdots, n$를 고려하자. 각 y_i에서 시행횟수는 m_i이고 관련된 p개의 설명변수를 $\boldsymbol{x}_i$라고 하면 이항확률변수의 조건부 분포는 시행횟수가 m_i이고 성공확률이 $\theta(\boldsymbol{x}_i)$인 이항분포를 따른다. 분포의 평균과 분산이 $\theta(\boldsymbol{x}_i)$에 의해 결정되므로 분포의 특성을 파악하기 위해서는 $\theta(\boldsymbol{x}_i)$를 추정해야 한다. 만약 모든 m_i가 충분히 크다면 $\theta(\boldsymbol{x}_i)$는 y_i/m_i으로 추정할 수 있다.

이항분포 y_i/m_i의 평균함수 $\theta(\boldsymbol{x}_i) = M(\boldsymbol{\beta}^T\boldsymbol{x}_i)$는 자료로 부터 추정될 회귀계수 $\boldsymbol{\beta}$와 커널평균함수 M에 의존한다. 커널평균함수 M은 $\boldsymbol{\beta}^T\boldsymbol{x}_i$의 함수이며 그 형태는 알려져 있을 수도 있고 그렇지 않을 수도 있다. 선형회귀모형에서와는 달리 이항회귀모형에서의 평균함수는 성공확률을 나타내므로 0과 1사이의 값을 가져야한다. 일반적으로 이항회귀에서 가장 많이 쓰이는 평균함수 M은 로지스틱함수 (logistic function)이며 다음과 같다 (Lee와 Rhee, 2003).

$$\theta(\boldsymbol{x}_i) = M(\boldsymbol{\beta}^T\boldsymbol{x}_i) = \frac{exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)}{1 + exp(\boldsymbol{\beta}^T\boldsymbol{x}_i).} \tag{4.1}$$

로지스틱회귀분석에서 회귀계수 $\boldsymbol{\beta}$의 최대우도추정값인 $\hat{\boldsymbol{\beta}}$을 얻을 수 있다. 추정한 회귀계수와 식 (4.1)을 이용하여 $\widehat{\theta}(\boldsymbol{x}_i) = M(\widehat{\boldsymbol{\beta}}^T\boldsymbol{x}_i)$과 $\widehat{y}_i = m_i\widehat{\theta}(\boldsymbol{x}_i)$을 추정할 수 있다. 식 (4.1)을 $\boldsymbol{\beta}^T\boldsymbol{x}_i$에 대해서 풀면 $\theta(\boldsymbol{x}_i)$의 함수인 식 (4.2)를 얻을 수 있는데 이것은 로짓 (logit)이라고 불리는 연결함수 (link function)이다 (Seo와 Kim, 2006).

$$log\left(\frac{\theta(\boldsymbol{x}_i)}{1-\theta(\boldsymbol{x}_i)}\right) = \boldsymbol{\beta}^T\boldsymbol{x}_i. \tag{4.2}$$

이 연결함수를 이용하여 이항반응변수 y의 기대값과 설명변수의 선형결합을 연결하여 선형회귀모형의 형태로 표현할 수 있다 (Kahng과 Kim, 2004).

설명변수가 하나일 때, 이항반응변수인 경우에도 연속반응변수일 때와 마찬가지로 반응변수를 수직축으로 하고 설명변수를 수평축으로 하는 2차원 산점도가 충분요약그림이지만 y의 조건부 분포 $y|x$의 변화를 시각적으로 보는 데는 유용하지 못하다. 그러므로 y의 x에 대한 의존성의 정보는 비율이나 상대밀도를 통해서 얻어진다. 조건부 밀도함수 (conditional density function) $f(x|y=j),\ j=0,1$를 추정함으로써 상대밀도의 좀 더 완전한 형태를 얻을 수 있다. 다음의 식은 조건부 밀도함수가 회귀정보를 어떻게 설명하는지 말해준다.

$$\frac{\theta(\boldsymbol{x})}{1-\theta(\boldsymbol{x})} = \frac{P(y=1)}{1-P(y=1)} \times \frac{f(\boldsymbol{x}|y=1)}{f(\boldsymbol{x}|y=0)} = c \times \frac{f(\boldsymbol{x}|y=1)}{f(\boldsymbol{x}|y=0)}. \tag{4.3}$$

성공의 오즈 (odds) $P(y=1)/[1-P(y=1)]$는 상수이므로 c 로 대체하고 식 (4.3)에 로그를 취하면 다음과 같고 이는 이항회귀의 로짓 연결함수와 같다.

$$log\left(\frac{\theta(\boldsymbol{x})}{1-\theta(\boldsymbol{x})}\right) = log(c) + log\left(\frac{f(\boldsymbol{x}|y=1)}{f(\boldsymbol{x}|y=0)}\right) = \boldsymbol{\beta}^T\boldsymbol{x}.$$

설명변수의 조건부 밀도함수는 모형에 포함될 설명변수를 평가하는 데에도 유용한 정보를 제공한다. $f(\boldsymbol{x}|y=0)$과 $f(\boldsymbol{x}|y=1)$이 등분산의 정규분포를 따른다면 로지스틱회귀모형에 포함될 설명변수 항은 $\boldsymbol{x}^T=(1,x)$이 된다. 반면에 만약 이분산의 정규분포를 따른다면 $\boldsymbol{x}^T=(1,x,x^2)$이 회귀모형의 설명변수가 된다 (Cook과 Weisberg, 1999).

4.2. 로지스틱모형에 대한 그래픽적 회귀

로지스틱모형에서도 선형모형에서와 같이 그래픽적 회귀를 할 수 있다. 다만 반응변수가 이진변수인 경우에는 추가변수그림 대신 이진반응그림 (binary response plot)을 통한 시각적인 평가가 가능하다.

x_1이 주어졌을 때 x_2의 추가적인 설명력을 평가하려면 x_1을 수평축으로 하고 x_2를 수직축으로 하는 이진반응그림을 그린다. 그리고 반응변수의 두 범주가 구분이 가능하도록 산점도 위에 표시한다. 이진반응그림은 반응변수를 x_1에 적합 시킨 후에 반응변수에 대한 추가적인 정보를 x_2가 가지고 있는지 알아보기 위해 사용한다. 즉 x_1이 주어졌을 때, 반응변수와 x_2의 독립여부에 대하여 평가하는 데에 이용된다. 만약 $y|x_1$과 x_2가 독립이면 반응변수의 상대 밀도가 수평축에 수직인 조각 (slice) 안에서 상수가 되어야 한다. 즉, $y=0$과 $y=1$의 분포가 같아야 한다는 의미이다.

로지스틱모형을 평가하는 데에도 이진반응그림이 이용된다. 특정한 로지스틱모형 $\boldsymbol{\beta}^T\boldsymbol{x} = \beta_0 + \beta_1x_1 + \beta_2x_2$를 생각해 보자. $\boldsymbol{\beta}^T\boldsymbol{x}$가 주어졌을 때 반응변수 y가 (x_1,x_2)에 독립적이라면 모형이 적절하다고 볼 수 있지만, $\boldsymbol{\beta}^T\boldsymbol{x}$가 주어졌을 때 반응변수 y가 (x_1,x_2)에 의존한다면 모형이 충분하지 않으며 추가적인 항이 필요함을 의미한다.

로지스틱회귀분석에서 그래픽적 회귀를 수행하는 방법은 선형회귀분석의 경우와 유사하다. 그래픽적 회귀에서는 식 (2.2) 형태로 표현되는 두 변수의 결합 가능여부를 판단해야 하는데 로지스틱회귀에서는 3차원 추가변수그림 대신 이진반응그림을 이용해야한다. 수행 방법은 다음과 같다. x_1을 H축, x_2를 V축으로 하고 회귀에 대한 아무런 정보가 없는 관측번호를 O축으로 하여 3차원 산점도를 그린다. O축을 중심으로 회전시켜서 수평 스크린 축에 상응하는 $\boldsymbol{b}^T\boldsymbol{x} = x_{12} = b_1x_1 + b_2x_2$를 얻는다. 여기서 얻어진 2차원 산점도가 이진반응그림이다. 만약 이진반응그림에서 0차원구조의 징후를 보인다면 x_1과 x_2는 영향력이 없다고 판단되어 삭제할 수 있다. 만약 3차원 추가변수그림에서 1차원구조를 갖는다고 판단되면 x_1과 x_2는 x_{12}로 결합할 수 있고 2차원구조를 갖는다고 판단된다면 x_1과 x_2는 결합할 수 없으므로

다른 설명변수 쌍을 선택하여 이와 같은 절차를 반복한다. Arc를 이용하면 차원축소를 위한 GREG 설명변수들을 결합하는 과정을 통해 회귀정보를 모두 포함하는 선형결합을 찾아낼 수 있다.

5. 예 제

반응변수가 이진변수인 경우에도 선형회귀모형에서와 같이 그래픽적인 회귀가 가능하지만 반응변수의 특성상 선형회귀모형에 적용했던 방법과는 다른 방법으로 접근하여야 한다. Weisberg (2005)에 소개된 사과나무 자료를 로지스틱모형인 경우의 그래픽적 회귀를 적용하고 모형의 적절성을 평가하고자 한다.

줄기의 평균 길이 (len), 휴면 일수 (day), 추출된 어린 싹의 수 (m), 휴면 일수 내 표준편차 ($stdev$)를 설명변수로 하고, 어린 싹의 길이가 길면 1, 짧으면 0으로 하는 이진변수를 반응변수 ($type$)로 하는 자료이다. 4절에서와 설명한 바와 같이 설명변수의 조건부 밀도함수는 모형에 포함될 설명변수를 평가하는 데에도 유용한 정보를 제공한다. $f(\boldsymbol{x}|y=0)$과 $f(\boldsymbol{x}|y=1)$는 반응변수의 값을 구분하여 표시하는 그림 5.1의 히스토그램을 통하여 확인이 가능하다.

그림 5.1을 보면 설명변수 len과 day는 등분산의 정규분포로 보아도 무방하다고 판단되나 m과 $stdev$는 치우친 분포로써 변수변환을 필요로 한다. 이 두 설명변수를 로그변환하면 그림 5.2에서와 같이 $log(m)$과 $log(stdev)$는 등분산의 정규분포로 볼 수 있다. 따라서 4개의 설명변수 len, day, $log(m)$, $log(stdev)$를 포함하는 로지스틱회귀를 실시해 본다. 적합시킨 결과 설명변수 $log(m)$와 $log(stdev)$는 유의하지 않은 것으로 판단되어 이들을 제외한 2개의 설명변수만을 포함하는 다음의 모형을 고려하기로 한다.

$$logit[\theta(\boldsymbol{x})] = \beta_0 + \beta_1 len + \beta_2 day. \tag{5.1}$$

로지스틱모형 (5.1)을 적합시킨 결과는 표 5.1과 같다. 이 모형의 평가를 위하여 4절에서 설명한 주변모형 산점도를 그려보면 그림 5.3과 같다. 실선과 점선은 각각 모형을 가정하지 않고 추정한 주변평균함수와 모형을 가정하고 추정한 주변평균함수를 나타낸다. 주변모형 산점도는 적합값에 대해서는 주변평균함수가 대체로 일치하고 있음을 보여주나 모형에 포함되는 2개의 설명변수들에 대해서는 주변평균함수가 일치하지 않음을 볼 수 있다. 그러므로 로지스틱모형 (5.1)이 적절하지 않음을 알 수 있다.

표 5.1 로지스틱모형의 적합 결과

Parameter	Estimate	std. Error	Est/SE	p-value
Constant	-30.1471	8.76487	-3.440	0.0006
Len	3.20983	0.927324	3.461	0.0005
Day	-0.669506	0.194231	-3.447	0.0006
Number of cases:			50	
Degrees of freedom:			47	
Pearson X2:			50.979	
Deviance:			22.301	

그래픽적 회귀를 하기 위해 먼저 설명변수들이 선형의 관계를 만족하고 있는지 확인해야 한다. 산점도행렬을 그려보면 len와 day를 제외한 나머지 설명변수들 간에는 선형성의 조건을 만족하지 못하는 것을 알 수 있다. 따라서 변수변환을 통해 설명변수들 사이에 선형성을 만족하도록 해야 한다. 위에서와 같이 m과 $stdev$를 로그변환을 시키고 4개의 변수 len, day, $log(m)$, $log(stdev)$에 대한 산점도행렬을

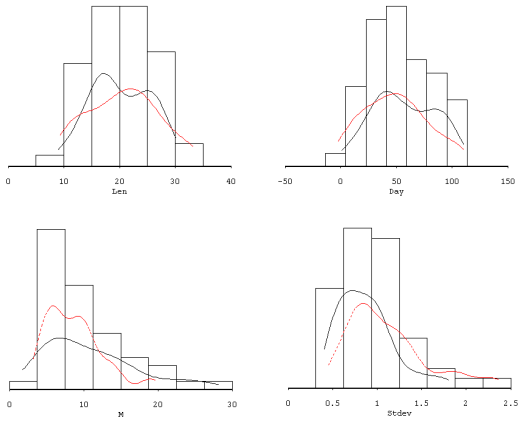

그림 5.1 설명변수의 히스토그램과 조건부 밀도함수 추정

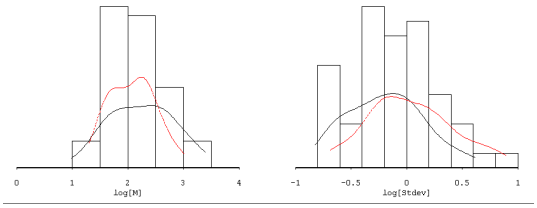

그림 5.2 변환된 설명변수의 히스토그램과 조건부 밀도함수 추정

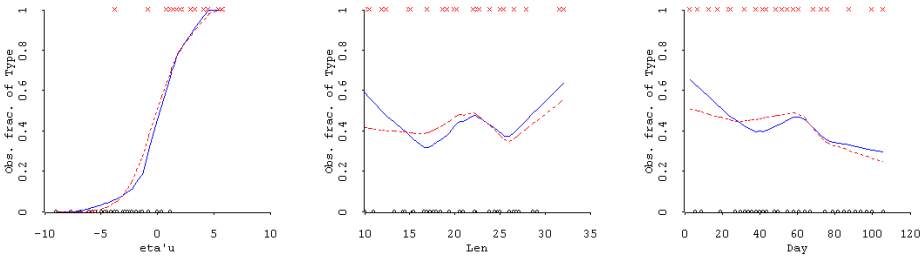

그림 5.3 모형 (5.1)의 주변모형 산점도

보면 그림 5.4와 같고 설명변수들 간의 관계가 선형에서 크게 벗어나지 않는다고 볼 수 있다. 따라서 4개의 변수를 선형의 관계에 있는 설명변수로 보고 그래픽적 회귀를 수행한다.

Arc를 이용하여 그래픽적 회귀를 수행한 결과가 표 5.2와 같다. 원래의 설명변수로부터 서로 무상관의 4개의 선형결합인 Fit, $gr1$, $gr2$, $gr3$가 만들어졌다. 이 회귀의 차원구조를 알아보기 위해 위의 4개의 선형결합 중에서 가장 영향력이 작다고 판단되는 선형결합들로 두 개씩 짝을 지어 차원구조를 알아보

는 절차를 반복한다.

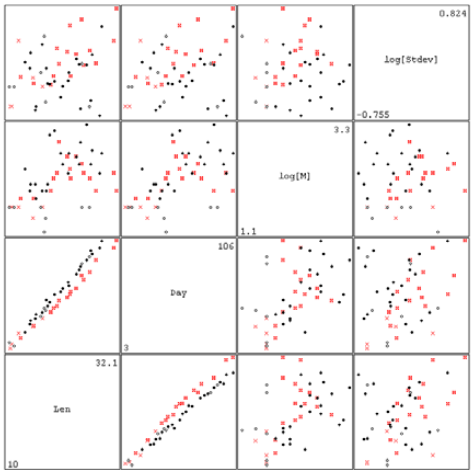

그림 5.4 설명변수의 산점도행렬

표 5.2 Arc를 이용한 그래픽적 회귀 결과

	len	day	$log(m)$	$log(stdev)$
Fit	0.909	-0.184	-0.308	0.214
$gr1$	-0.058	0.024	0.025	0.998
$gr2$	0.159	-0.021	0.071	-0.985
$gr3$	0.023	-0.015	0.994	0.102

반응변수가 이항변수이기 때문에 4절에서 설명한 바와 같이 이진반응그림을 그린다. 먼저 $gr2$와 $gr3$를 선택하여 $gr2$를 H축으로 하고 $gr3$를 V축으로 하는 이진반응그림를 얻을 수 있다. $gr2$와 $gr3$를 이용한 이진반응그림은 0차원구조를 갖는다고 판단할 수 있고 반응변수에 아무런 설명력이 없는 $gr2$와 $gr3$는 회귀모형에서 제거될 수 있다. 남아있는 두 GREG 설명변수인 Fit과 $gr1$을 이용하여 이진반응그림을 그린다. 이 이진반응그림을 O축을 중심으로 회전시켜 보면 수평축에 수직인 조각 내에서 반응변수의 상대적 밀도가 일정하게 분포하고 있지 않음을 볼 수 있다. 이와 같이 회전시키는 동안 반응변수의 상대밀도가 일정하지 않은 분포를 보이다가 어느 특정한 각도에서 그림 5.5과 같이 어느 정도 일정한 분포를 보이는 이진반응그림을 찾을 수 있다.

수평축에 수직인 조각이 좌우로 움직임에 따라 반응변수의 두 범주의 상대 밀도가 변화하지만 각 조각 안에서는 세로축의 값에 의존하지 않으므로 수평축을 이루는 선형결합 $gr4 = \widehat{\boldsymbol{\eta}}^T\boldsymbol{x} = 0.924len - 0.188day - 0.313log(m) + 0.114log(stdev)$에만 의존하는 1차원구조를 갖는다고 판단한다. 이러한 선형결합은 Arc를 통해 표 5.3과 같은 결과를 얻을 수 있다.

표 5.3 Arc를 이용한 그래픽적 회귀 결과

	len	day	$log(m)$	$log(stdev)$
$gr4$	0.924	-0.188	-0.313	0.114

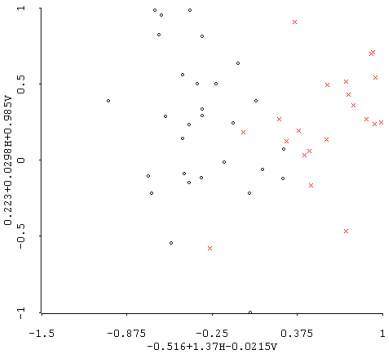

그림 5.5 이진반응그림

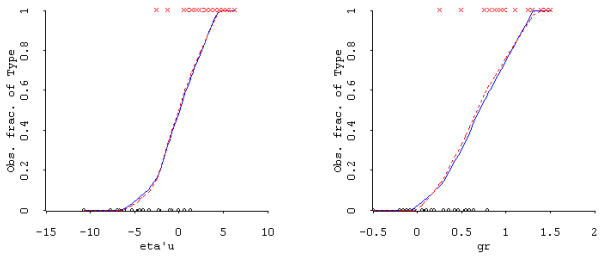

그림 5.6 주변모형 산점도

이제 그래픽적 회귀를 통해 얻은 선형결합 $gr4$가 과연 회귀의 정보를 모두 포함하는지 평가하기 위해 주변모형 산점도를 이용하기로 한다. 위에서 얻은 $gr4$의 회귀계수를 $\boldsymbol{a}$로 취하여 주변모형 산점도를 그리면 그림 5.6과 같다. 이 주변모형 산점도는 주변평균함수가 일치하고 있음을 보여준다. 따라서 위에서 얻은 선형결합이 회귀의 정보를 모두 설명하는데 적절하다고 판단된다.

6. 결론

그래픽적 회귀는 모형에 대한 가정을 하지 않고 회귀정보를 모두 포함하는 충분요약그림을 찾아내는 분석 방법으로 모든 회귀정보를 저차원의 그림으로 표현할 수 있게 하는 데에 그 목적이 있다. 본 논문에서는 그래픽적 회귀방법을 로지스틱회귀모형에 확장하여 적용해 보았다. 또한 잔차산점도를 이용한 모형의 평가는 적용 범위가 선형회귀모형에 국한되는 문제점이 있기 때문에 그 대안으로 주변모형 산점도를 이용하여 모형의 적절성을 평가해 보았다.

참고문헌

Atkinson, A. C. (1985). *Plots, transformations, and regression*, Oxford University Press, Oxford.

Azzalini, A. and Bowman, A. W. (1993). On the use of nonparametric regression for checking linear relationships. *Journal of the Royal Statistical Society, Ser. B,* **55**, 549-557.

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. (1983). *Graphical methods for data Analysis*, Chapman and Hall, New York.

Cleveland, W. and Devlin, D. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association,* **83**, 596-610.

Cleveland, W. S. (1987). Research in statistical graphics. *Journal of the American Statistical Association,* **82**, 419-423.

Cook, R. D. (1998). *Regression graphics: Idea for studying regressions through graphics*, John Wiley & Sons, New York.

Cook, R. D. and Weisberg, S. (1982). *Residuals and influence in regression*, Chapman and Hall, New York.

Cook, R. D. and Weisberg, S. (1994). *An introduction to regression graphics,* John Wiley & Sons, New York.

Cook, R. D. and Weisberg, S. (1997). Graphics for assessing the adequacy of regression models. *Journal of the American Statistical Association,* **92**, 490-499.

Cook, R. D. and Weisberg, S. (1999). *Applied regression including computing and graphics*, John Wiley & Sons, New York.

Ezekiel, M. (1924). A method for handling curvilinear correlation for any number of variables. *Journal of the American Statistical Association,* **19**, 431-453.

Kahng, M. (2005). Exploring interaction in generalized linear models. *Journal of Korean Data & Information Science Society,* **16**, 13-18.

Kahng, M. and Kim, M. (2004). A score test for detection of outliers in generalized linear models. *Journal of Korean Data & Information Science Society,* **15**, 129-139.

Lee, J. and Rhee, S. (2003). Logistic model for normality by neural networks. *Journal of Korean Data & Information Science Society,* **14**, 119-129.

Seo, M. and Kim, J. (2006). Estimation of odds ratio in proportional odds model. *Journal of Korean Data & Information Science Society,* **17**, 1067-1076.

Tierney, L. (1990). *Lisp-Stat: An object-oriented environment for statistical computing and dynamic graphics,* John Wiley & Sons, New York.

Weisberg, S. (2005). *Applied linear regression,* 3rd Ed., John Wiley & Sons, New York.

Journal of the Korean Data & Information Science Society
2010, **21**(1), 21–32

한국데이터정보과학회지

Graphical regression and model assessment in logistic model[†]

Myung-Wook Kahng[1] · Bu-Yong Kim[2] · Ju-Hee Hong[3]

[123]Department of Statistics, Sookmyung Women's University

Received 15 October 2009, revised 20 December 2009, accepted 28 December 2009

Abstract

Graphical regression is a paradigm for obtaining regression information using plots without model assumptions. The general goal of this approach is to find low-dimensional sufficient summary plots without loss of important information. Model assessments using residual plots are less likely to be successful in models that are not linear. As an alternative approach, marginal model plots provide a general graphical method for assessing the model. We apply the methods of graphical regression and model assessment using marginal model plots to the logistic regression model.

Keywords: Binomial regression, dimension reduction, graphical regression, logistic model, marginal model plot.

[†] This research was supported by the Sookmyung Women's University Research Grants 2008.
[1] Corresponding author: Professor, Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea. E-mail: mwkahng@sm.ac.kr
[2] Professor, Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea.
[3] Graduate Student, Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea.