The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.6.1191
28(6), 1191–1208

Various Graphical Methods for Assessing a Logistic Regression Model

Kyung Jin Kim[a] · Myung Wook Kahng[a,1]

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

(Received October 19, 2015; Revised December 4, 2015; Accepted December 8, 2015)

Abstract

Most statistical methods are dependent on the summary statistic. However, with graphical approaches, it is easier to identify the characteristics of the data and detect information that cannot be obtained by the summary statistic. We present various graphical methods to assess the adequacy of models in logistic regression that include checking log-density ratio, structural dimension, marginal model plot, chi-residual plot, and CERES plot. Through simulation data, we investigate and compare the results of graphical approaches under diverse conditions.

Keywords: binary response plot, CERES plot, chi-residual plot, log-density ratio, marginal model plot, structural dimension

1. 서론

대부분의 통계분석방법은 자료가 가지고 있는 정보를 하나의 숫자로 표현하는 요약통계량에 의존한다. 선형회귀모형이나 일반화선형모형을 평가하고 진단할 때에도 통계량을 이용한 검정방법이 사용된다. 반면 그래픽적 방법을 이용하면 자료의 특성을 한눈에 파악하기 쉽고 통계량만으로는 알아낼 수 없는 부분까지도 접근이 가능하다.

그래프를 이용한 회귀분석은 Ezekiel (1924)에 의해 처음 시도되었고 Belsley 등 (1980), Cook과 Weisberg (1982), Atkinson (1985)에 의해 회귀진단에 사용되었다. Cook과 Weisberg (1994)가 그래픽적 회귀를 소개한 이후에 Cook (1998)은 이에 대한 수리적이고 정밀한 회귀분석 방법을 제시하였고 Cook과 Weisberg (1999)는 그동안 제시된 그래픽적 회귀의 방법론을 종합적으로 정리하였다.

선형회귀모형에서 반응변수의 기댓값은 설명변수들의 선형결합이라고 가정한다. 로지스틱회귀모형에서도 역시 설명변수들의 선형결합이 이용된다. 하지만 설명변수의 선형결합만으로는 충분히 설명이 되지 못하고 설명변수의 변환된 형태 등의 추가적인 요소의 포함이 필요한 경우가 있다. 이러한 연구는 Kay와 Little (1987)에 의하여 시작되었고 Scrucca (2003), Scrucca와 Weisberg (2004)의 연구가 있는데 로그-밀도비(log-density ratio)의 개념과 그래픽적 방법을 근거로 하고 있다.

This Research was supported by the Sookmyung Women's University Research Grants (1-1403-0001).
[1]Corresponding author: Department of Statistics, Sookmyung Women's University, Seoul 04310, Korea.
E-mail: mwkahng@sm.ac.kr

일반적으로 선형회귀모형의 적절성을 평가하는 도구로 잔차산점도가 널리 이용되고 있으나 일반화선형모형의 적절성을 평가하기에는 부적합하다. Cook과 Weisberg (1997)는 잔차산점도의 대안으로써 주변모형 확인조건에 기초한 주변모형산점도(marginal model plot)를 제안하였다. 일반화선형모형 중에서 특히 이항반응변수를 가진 로지스틱회귀모형의 적절성을 평가하는 그래픽적 방법으로 카이잔차산점도(chi-residual plot)를 이용한 모형평가 방법이 있다. 또한 Cook (1993), Cook과 Croos-Dabrera (1998)은 CERES 그림을 변수의 비선형성 진단방법으로 제시하였다.

본 연구에서는 로지스틱모형에서의 여러 가지 그래픽적인 방법을 통하여 모형을 평가하는 방법을 알아본다. 그리고 다양한 형태의 모의자료를 통하여 그래픽적 방법들에 의한 결론에 어떠한 차이가 나는지를 검토해 보고 실제자료에도 적용하여본다.

2. 로지스틱회귀모형

확률변수 y가 시행횟수가 m이고 성공확률이 θ인 이항분포를 따른다고 하자. 반응변수를 y/m로 설명변수를 $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)'$로 하는 이항회귀(binomial regression)의 모형은 $E(y/m|\boldsymbol{x}) = \theta(\boldsymbol{x}) = m(\boldsymbol{x}'\boldsymbol{\beta})$로 표현된다. 이항회귀모형은 일반화선형모형의 한 형태로 $m(\ )$는 커널평균함수(kernel mean function)이고 연결함수의 역함수이다. 커널평균함수로 로지스틱함수(logistic function)를 사용하는 로지스틱회귀모형은 다음과 같다.

$$E\left(\frac{y}{m}\Big|\boldsymbol{x}\right) = \theta(\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})} = m(\boldsymbol{x}'\boldsymbol{\beta}). \tag{2.1}$$

모형 (2.1)에서는 $\boldsymbol{x}$의 선형결합인 $\boldsymbol{x}'\boldsymbol{\beta}$의 함수로 모형을 구성하고 있으나 Cook과 Weisberg (1999)에서와 같이 $\boldsymbol{u}$의 선형결합인 $\boldsymbol{\eta}'\boldsymbol{u}$의 함수를 사용하면 변환 등 다양한 상황을 포함하는 모형을 구성할 수 있다. 여기서 $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$는 p개의 설명변수 $\boldsymbol{x}$로부터 구한 $q \times 1$벡터이다. 일반적으로 $\boldsymbol{u}$는 $\boldsymbol{x}$의 함수들로 구성된다. 이 경우 모형 (2.1)에서와 같이 로지스틱함수를 커널평균함수로 하면 다음과 같이 로지스틱회귀모형이 되며

$$E\left(\frac{y}{m}\Big|\boldsymbol{x}\right) = \theta(\boldsymbol{x}) = \frac{\exp(\boldsymbol{\eta}'\boldsymbol{u})}{1+\exp(\boldsymbol{\eta}'\boldsymbol{u})} = m(\boldsymbol{\eta}'\boldsymbol{u}). \tag{2.2}$$

모형 (2.2)는 로짓 연결함수를 통하여 다음과 같이 선형모형의 형태가 된다.

$$\log\left(\frac{\theta(\boldsymbol{x})}{1-\theta(\boldsymbol{x})}\right) = \phi = \boldsymbol{\eta}'\boldsymbol{u}.$$

로지스틱회귀에서 가능도함수를 최대화 시키는 $\boldsymbol{\eta}$의 최대가능도추정량 $\hat{\boldsymbol{\eta}}$을 찾을 수 있다. 여기서 $\hat{\boldsymbol{\eta}}$은 최소제곱추정량도 된다. $\theta(\boldsymbol{x}_i)$는 $\hat{\theta}(\boldsymbol{x}_i) = \exp(\hat{\boldsymbol{\eta}}'\boldsymbol{u}_i)/[1+\exp(\hat{\boldsymbol{\eta}}'\boldsymbol{u}_i)]$로 추정하고 로지스틱회귀에서 적합값(fitted value)은 $\hat{y}_i = m_i\hat{\theta}(\boldsymbol{x}_i)$가 된다.

선형회귀에서 모형의 비교에 대한 검정에 잔차제곱합이 기본적인 도구로 사용되는 것과 마찬가지로 로지스틱회귀에서 이탈도(deviance)가 사용된다. 이탈도 G^2은 다음과 같다.

$$G^2 = 2\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - \hat{y}_i}\right).$$

회귀 $y|\boldsymbol{x}$와 역회귀(inverse regression) $\boldsymbol{x}|y$사이의 관계를 알아보자. $f(\boldsymbol{x}|y=j)$를 $y=j$가 주어졌을 때, $\boldsymbol{x}$에 대한 확률밀도함수라 하자. 그리고 $f(\boldsymbol{x})$를 주변확률밀도함수라 하자. 반응변수가 이항변수이

므로 로지스틱회귀에서의 평균함수 $E(y|\boldsymbol{x}) = P(y=1|\boldsymbol{x}) = \theta(\boldsymbol{x})$와 $1-\theta(\boldsymbol{x})$의 로그비를 취하면 다음과 같이 로그-오즈(log-odds)를 얻을 수 있다.

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

이러한 로그-오즈는 두 항의 합이다. 첫 번째 항은 $\boldsymbol{x}$에 의존하지 않는 주변로그-오즈(marginal log-odds) $\log(c)$이고 두 번째 항 $h(\boldsymbol{x})$는 로그-밀도비라고 한다.

3. 그래프를 이용한 모형평가

3.1. 조건부 밀도

일반적으로 회귀모형에서 설명변수에 대한 분포의 가정은 하지 않는다. Cook과 Weisberg (1999)는 반응변수의 형태가 특정 분포로 볼 수 있는 경우 $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$를 적절하게 선택할 수 있다고 하였다. 만약 $f(x|y=j)$, $j=0,1$가 평균 μ_i와 분산 σ_j^2를 가지는 정규밀도함수라면 로그-밀도비는

$$h(x) = \log\left(\frac{\sigma_0}{\sigma_1}\right) - \frac{\mu_1^2}{2\sigma_1^2} + \frac{\mu_0^2}{2\sigma_0^2} + \left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_0}{\sigma_0^2}\right)x + \frac{1}{2}\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right)x^2$$

이 되고 로그-오즈는

$$\log\left(\frac{\theta(\boldsymbol{x})}{1-\theta(\boldsymbol{x})}\right) = \eta_0 + \eta_1 x + \eta_2 x^2 = \boldsymbol{\eta}'\boldsymbol{u}$$

이고 $\eta_0 = \log(c) + \log(\sigma_0/\sigma_1) + (\mu_0^2/2\sigma_0^2 - \mu_1^2/2\sigma_1^2)$, $\eta_1 = \mu_1/\sigma_1^2 - \mu_0/\sigma_0^2$, $\eta_2 = (1/\sigma_0^2 - 1/\sigma_1^2)/2$이다. 조건부 분포가 정규분포의 형태를 보인다면 $y=0$인 경우와 $y=1$인 경우의 두 정규분포의 분산이 동일한가에 따라서 평균함수에서의 $\boldsymbol{u}$항이 달라질 수 있다. 두 정규분포의 분산이 같은 경우에는 로그-밀도비에서의 x^2이 포함된 항이 0이 된다. 따라서 두 정규분포의 분산이 같은 경우에는 평균함수의 $\boldsymbol{u}$항은 $\boldsymbol{u} = (1,x)'$가 되며 일차항만이 포함된 로지스틱회귀모형을 생각하면 된다. 두 정규분포의 분산이 다른 경우에는 $\boldsymbol{u}$항은 $\boldsymbol{u} = (1,x,x^2)'$가 되고 이차항이 포함된 로지스틱회귀모형을 생각해야 한다.

조건부 분포가 좌우대칭인 자료는 정규분포로 설명이 가능하다. 하지만 조건부 분포가 비대칭으로 치우친 경우(skewed)에는 감마분포를 사용할 수 있고 일차항 x나 로그항 $\log(x)$가 평균함수에 포함되어야 한다. $f(x|y=j)$, $j=0,1$가 형태모수(shape parameter) α_i와 척도모수(scale parameter) λ_j를 갖는 감마밀도함수라면 로그-밀도비는

$$h(x) = \log\left(\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)}\right) + [\alpha_0\log(\lambda_0) - \alpha_1\log(\lambda_1)] + \left(\lambda_0^{-1} - \lambda_1^{-1}\right)x + (\alpha_1 - \alpha_0)\log x$$

이 되고 로그-오즈는

$$\log\left(\frac{\theta(\boldsymbol{x})}{1-\theta(\boldsymbol{x})}\right) = \eta_0 + \eta_1 x + \eta_2\log(x) = \boldsymbol{\eta}'\boldsymbol{u}$$

이고 $\eta_0 = \log(c\Gamma(\alpha_0)/\Gamma(\alpha_1)) + [\alpha_0\log(\lambda_0) - \alpha_1\log(\lambda_1)]$, $\eta_1 = 1/\lambda_0 - 1/\lambda_1$, $\eta_2 = (\alpha_1 - \alpha_0)$가 된다. 따라서 평균함수의 $\boldsymbol{u}$항은 $\boldsymbol{u} = (1,x,\log(x))'$가 되며 로그항이 포함된 로지스틱회귀모형을 생각해야 한다. 두 감마분포의 형태모수 α_0와 α_1이 같은 경우에는 $\eta_2 = 0$이 되므로 평균함수의 $\boldsymbol{u}$항은 $\boldsymbol{u} = (1,x)'$가 되며 일차항만 포함된 로지스틱회귀모형으로 충분하다.

3.2. 차원 구조

반응변수 y와 p개의 설명변수들 $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)'$에 대한 회귀모형을 생각하자. Cook과 Weisberg (1999)에 의하면 회귀에서의 차원 구조(structural dimension)는 정보의 손실 없이 회귀의 특징을 표현할 수 있는 가장 적은 $\boldsymbol{x}$의 선형결합의 수이다. 구조적 차원은 항상 0과 p사이의 정수로써 회귀들이 0차원, 1차원, $\ldots$, p차원 구조를 가진다고 한다.

반응변수 y가 $\boldsymbol{x}$에 의존하지 않는다면, 즉 y와 $\boldsymbol{x}$가 아무런 관계가 없다면, $\boldsymbol{x}$에 대한 어떠한 선형결합도 y에 대한 정보를 주지 못하는 경우이기 때문에 0차원 구조(0D structure)라고 한다. 0차원 구조에서는 $\boldsymbol{x}$를 포함한다고 해서 y의 정보가 늘어나지 않는다. 0차원 구조에서는 평균함수 $E(y|\boldsymbol{x})$와 분산함수 $\mathrm{Var}(y|\boldsymbol{x})$가 모두 상수이다. 만약 반응변수 y가 $\boldsymbol{x}$에 대해 하나의 선형결합 $\boldsymbol{\beta}'\boldsymbol{x}$에만 의존한다면 1차원 구조(1D structure)라고 할 수 있다. 1차원 구조에서는 평균함수와 분산함수가 하나의 선형결합 $\boldsymbol{\beta}'\boldsymbol{x}$에 의존한다. 두 개의 선형결합 $\boldsymbol{\beta}_1'\boldsymbol{x}$, $\boldsymbol{\beta}_2'\boldsymbol{x}$가 회귀 $y|\boldsymbol{x}$를 표현하기 위해 필요하다면 2차원 구조(2D structure)가 된다. 2차원 구조에서는 평균함수와 분산함수가 두 개의 선형결합 $\boldsymbol{\beta}_1'\boldsymbol{x}$, $\boldsymbol{\beta}_2'\boldsymbol{x}$에 의존한다.

Cook과 Weisberg (1999)에 의하면 설명변수 사이의 관계가 선형이라는 조건이 만족되는 상태에서 반응변수가 적어도 하나의 설명변수에 관계되어 있다면 0차원 구조는 판단에서 배제하게 된다. 1차원 구조인지 그 이상인지를 판단하기 위하여 반응변수와 설명변수의 축을 바꾼 역회귀곡선(inverse regression curve)을 이용해야 한다. 만약 모형이 1차원이라면 p개의 x_j와 y를 축으로 하는 그림들은 다음 세 가지 중의 하나의 형태가 된다. 첫째, p개의 역회귀곡선의 형태가 같거나 둘째, 형태가 뒤집힌 경우이다. 즉, $E(x_j|y)$의 한 형태가 $\cup$형태라면 다른 형태들도 $\cup$형태가 되거나 $\cap$형태인 경우를 의미한다. 셋째, 역회귀곡선이 일정한 경우이다. 만약 이 3가지 중 어떤 경우에도 해당되지 않는다면 1차원 구조는 배제하게 된다.

반응변수 y가 0과 1을 가지는 경우의 로지스틱회귀모형을 생각해보자. x_1과 x_2를 설명변수로 하는 모형이 정확하다면 설명변수들의 적절한 선형결합 $\boldsymbol{\beta}'\boldsymbol{x}$가 주어졌을 때 반응변수 y와 x_1, x_2는 독립일 것이다. 만약 독립이 아니라면 이 모형은 회귀의 완전한 정보를 제공하는 것이 아닐 것이며 추가적인 정보가 요구될 것이다. 그래픽적인 판단을 하기 위해 x_1을 수평축으로, x_2를 수직축으로 하는 산점도에 반응변수 $y = 0$, $y = 1$을 다르게 나타내는 2차원 그림에 관측번호를 추가적인 축으로 하는 3차원 그림을 생각해보자. 여기서 추가된 축을 중심으로 회전시키다가 어떤 특정 각도에 해당되는 선형결합들의 2차원 그래프를 생각해보자. 이 그래프를 이항반응그림(binary response plot)이라고 한다. 이항반응그림의 수평축에 수직이 되도록 일정한 간격으로 분할하여 조각(slice)으로 나누었을 때, 그 조각 안에서 $y = 1$에 해당하는 점의 상대적 밀도가 모든 조각 내에서 일정하다면 반응변수와 수직축에 주어진 선형결합이 서로 독립임을 뜻하므로 수평축에 주어진 선형결합 하나만으로도 반응변수를 설명하기에 충분하다고 결론내릴 수 있다. 즉, 1차원 구조가 이 회귀모형에 적당하다는 뜻이 될 수 있다. 그러나 $y = 1$에 해당하는 점의 상대적 밀도가 위, 아래, 또는 중앙에 밀집되어 고르지 못하게 나타나는 조각들이 있다면 수평축에 주어진 선형결합만으로 반응변수의 모든 정보를 설명한다고 할 수 없다. 즉, 주어진 선형결합 외에 다른 선형결합이 필요하다는 뜻이므로 회귀모형이 1차원 구조가 아니라고 할 수 있다.

3.3. 주변모형산점도

선형회귀분석에서 모형의 적절성 평가를 위한 도구로 잔차산점도가 널리 이용되고 있다. 모형이 적절하다면 잔차산점도의 수직축을 이루는 잔차와 수평축을 이루는 설명변수들의 선형결합이 서로 독립적인 것으로 나타나야 한다는 것이 잔차산점도를 이용한 모형평가 방법의 기본 개념이다. 그러나 대부분의 일반화선형모형에서는 잔차산점도를 이용한 모형평가 방법은 성공적이지 못하다. 특히 반응변수가 0

또는 1인 이항회귀에서 잔차산점도는 모형의 적절성과 관계없이 특정한 패턴을 갖게 된다. 잔차산점도를 이용하여 모형을 평가하는 방법의 적용 범위가 선형회귀모형에 국한되는 문제점이 있기 때문에 대안으로 Cook과 Weisberg (1997)가 제안한 주변모형산점도를 이용하여 모형의 적절성을 평가할 수 있다.

주변모형산점도의 기본 개념은 반응변수 y와 q개의 요소로 이루어진 벡터 $\boldsymbol{u}$로 구성된 회귀모형을 두 가지 관점에 근거한 조건부 누적밀도함수의 비교를 통하여 평가하는 것이다. 자료가 독립적이고 같은 분포를 갖는다고 가정하고 모형에 대한 구체적인 가정 없이 자료에서 얻어지는 미지의 누적밀도함수 $F(y|\boldsymbol{u})$와 회귀모형을 구체적으로 가정한 후에 회귀모형으로부터 형성되는 조건부 누적밀도함수 $M(y|\boldsymbol{\eta},\boldsymbol{u})$가 비교되는 대상이다.

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

주변모형 확인조건은 모든 주변모형이 참일 때에만 완전모형이 참이라는 것을 의미한다. 완전모형의 적절성을 평가하려면 고려해야하는 주변모형산점도의 수가 증가한다. 그러나 모든 주변모형산점도를 확인하는 것은 불가능하므로 방향 $\boldsymbol{a}$를 적절히 선택해야 한다. 방향 선택을 위한 몇 가지 표준적인 방법이 있지만 그 중에서 기본적으로 사용하는 방법은 다음과 같다. 첫째, 회귀모형이 선형모형인 경우, 즉 $\boldsymbol{\eta}'\boldsymbol{u}$로 설명할 수 있다고 가정하는 경우에 $\boldsymbol{\eta}$의 최소제곱추정값 $\hat{\boldsymbol{\eta}}$을 이용하여 $\boldsymbol{a} = \hat{\boldsymbol{\eta}}$으로 선택한다. 따라서 y를 수직축으로 하고 $\hat{\boldsymbol{\eta}}'\boldsymbol{u}$를 수평축으로 하는 산점도를 이용하여 모형을 평가할 수 있다. 둘째, $\boldsymbol{a}'\boldsymbol{u}$가 각각의 $\boldsymbol{u}$가 되도록 $\boldsymbol{a}$를 선택한다. 따라서 y를 수직축으로 하고 각각의 $\boldsymbol{u}$를 수평축으로 하는 산점도를 이용하여 모형을 평가할 수 있다. 이는 각각의 변수들의 적절성을 설명하여 변수변환이나 다른 처방을 요구하는 근거를 제시한다 (Cook과 Weisberg, 1997).

모형에 대한 구체적인 가정을 하지 않은 경우의 주변평균함수(marginal mean function) $E_F(y|\boldsymbol{a}'\boldsymbol{u})$와 모형에 대한 구체적인 가정을 하는 경우의 주변평균함수 $E_M(y|\boldsymbol{a}'\boldsymbol{u})$를 생각하자. y를 수직축으로 하고 $\boldsymbol{a}'\boldsymbol{u}$를 수평축으로 하는 산점도에서 대표적인 평활 방법의 하나인 lowess(locally weighted scatter-plot smoother; Cleveland와 Devlin, 1988)를 이용하여 주변평균함수를 $\hat{E}_F = \hat{E}_F(y|\boldsymbol{a}'\boldsymbol{u})$와 $\hat{E}_M = \hat{E}_M(y|\boldsymbol{a}'\boldsymbol{u})$로 추정할 수 있다. 또한, 모형에 대한 가정 여부에 따른 주변분산함수(marginal variance function)를 $\mathrm{Var}_F(y|\boldsymbol{a}'\boldsymbol{u})$와 $\mathrm{Var}_M(y|\boldsymbol{a}'\boldsymbol{u})$라 하면 평활방법을 이용하여 $(\mathrm{SD}_F)^2 = \widehat{\mathrm{Var}}_F(y|\boldsymbol{a}'\boldsymbol{u})$와 $(\mathrm{SD}_M)^2 = \widehat{\mathrm{Var}}_M(y|\boldsymbol{a}'\boldsymbol{u})$를 추정할 수 있다.

모형의 평가는 y를 수직축으로 $\boldsymbol{a}'\boldsymbol{u}$를 수평축으로 하는 산점도에 모형에 대한 가정을 하지 않고 추정한 3개의 곡선 $\hat{E}_F$, $\hat{E}_F+\mathrm{SD}_F$, $\hat{E}_F-\mathrm{SD}_F$과 모형에 대한 가정을 하고 추정한 $\hat{E}_M$, $\hat{E}_M+\mathrm{SD}_M$, $\hat{E}_M-\mathrm{SD}_M$ 등 총 6개의 곡선을 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 모형이 적절하지 않음을 나타낸다. 주변모형산점도의 작성은 Xlisp-Stat (Tierney, 1990) 언어에 기초한 Arc를 사용하면 편리하게 수행할 수 있다. Arc는 웹(http://www.stat.umn.edu/arc/software.html)에서 무료로 얻을 수 있다.

3.4. 카이잔차산점도

일반적인 선형모형에서 잔차는 회귀진단의 도구로 사용되는 대표적인 통계량으로 잔차를 수직축으로 하

고 적합값를 수평축으로 하는 잔차산점도는 그래프를 이용한 회귀진단에 사용되는 대표적인 도구이다. 하지만 이러한 방법은 등분산 가정이 가능하지 않은 경우에는 왜곡된 결과를 나타낸다.

등분산 가정을 할 수 없는 상황에서는 선형모형

$$y_i|\boldsymbol{x}_i = \boldsymbol{\eta}'\boldsymbol{u}_i + \frac{\epsilon_i}{\sqrt{\omega_i}}, \quad i = 1, \ldots, n$$

을 고려해 볼 수 있고 이 모형의 분산함수 $\mathrm{Var}(y_i|\boldsymbol{x}_i) = \mathrm{Var}(\epsilon_i/\sqrt{\omega_i}|\boldsymbol{u}_i) = \sigma^2/\omega_i$는 가중값 ω_i에 의존한다. 우선 $\hat{y}_i = \hat{\boldsymbol{\eta}}'\boldsymbol{u}_i$를 가중최소제곱법에 의해 구해진 i번째 적합값이라 하고 잔차의 변형된 형태인 가중잔차(weighted residual) $e_i^\omega = \sqrt{\omega_i}(y_i - \hat{y}_i)$를 생각해 보자. 가중잔차 e_i^ω와 $\boldsymbol{x}$의 함수인 $\boldsymbol{\eta}'\boldsymbol{u}_i$의 산점도에서 평균함수 $E(e_i^\omega|\boldsymbol{\eta}'\boldsymbol{u}_i)$의 형태를 알 수 있고 이를 통해 $E(e_i|\boldsymbol{\eta}'\boldsymbol{u}_i)$의 추정이 가능하고 따라서 이 산점도로 모형의 적절성 파악이 가능할 것이다.

로지스틱회귀모형에서 y_i의 분산은 $\mathrm{Var}(y_i|\boldsymbol{x}_i) = m_i\theta(\boldsymbol{x}_i)(1-\theta(\boldsymbol{x}_i))$로 등분산이 아니므로 가중잔차를 고려해야 한다. 분산함수를 $\mathrm{Var}(y_i|\boldsymbol{x}_i) = \sigma^2/\omega_i$라 하고 $\sigma^2 = 1$로 하면 가중값은 $\hat{\omega}_i = 1/[m_i\theta(\boldsymbol{x}_i)(1-\theta(\boldsymbol{x}_i))]$로 설정할 수 있다. 선형회귀모형에서와는 달리 이러한 가중치는 모수들에 의존하기 때문에 가중치는 $\hat{\omega}_i = 1/[m_i\hat{\theta}(\boldsymbol{x}_i)(1-\hat{\theta}(\boldsymbol{x}_i))]$로 추정하여야 한다. 이렇게 추정된 가중값을 이용하면 가중잔차 e_i^ω는 다음과 같이 추정할 수 있다.

$$e_i^* = \frac{y_i - \hat{y}_y}{\sqrt{m_i\hat{\theta}(\boldsymbol{x}_i)\left(1-\hat{\theta}(\boldsymbol{x}_i)\right)}}.$$

Cook과 Weisberg (1999)에 의하면 이러한 잔차를 카이잔차(chi-residuals)라고 부르는데 제곱해서 더하면 Pearson의 X^2통계량이 되기 때문이다. 카이잔차 e_i^*와 설명변수들의 선형결합인 $\boldsymbol{\eta}'\boldsymbol{u}$를 두 축으로 하는 그래프를 카이잔차산점도라고 한다. 이 산점도는 평균함수 $E(e_i^*|\boldsymbol{\eta}'\boldsymbol{u})$의 형태를 나타내고 있는데 매우 조심스런 해석이 요구된다. 만약 $\hat{\boldsymbol{\eta}}'\boldsymbol{u}$를 수평축으로 e_i^*를 수직축으로 갖는 카이잔차산점도에서 평균함수가 일정하게 나타나면 모형이 적절하다는 것을 뜻한다. 하지만 이 평균함수가 일정하지 않다면 모형을 적절하지 않다고 말할 수 있다. 평균함수의 형태를 알기 위하여 주변모형산점도에서와 같이 평활곡선을 이용하면 편리하다. 주변모형산점도와는 달리 이러한 카이잔차산점도의 장점은 Arc 등과 같이 특수한 기능을 가진 소프트웨어가 아니더라도 작성할 수 있다는 장점이 있다.

3.5. CERES 그림

CERES(combining conditional expectation and residual) 그림은 회귀진단에서 추가변수의 필요성과 함수의 형태를 파악하는데 사용되는 방법이다. 로지스틱회귀모형에서 반응변수에 영향을 주는 $q \times 1$ 벡터 $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$의 q개의 원소 중의 하나인 u_2의 변환 $\tau(u_2)$을 고려한 로지스틱회귀모형은 다음과 같다.

$$\log\left(\frac{\theta(\boldsymbol{x})}{1-\theta(\boldsymbol{x})}\right) = \phi = \boldsymbol{\eta}_1^{*\prime}\boldsymbol{u}_1 + \tau(u_2), \tag{3.1}$$

여기서 $\boldsymbol{u}_1$는 u_2를 제외한 나머지 $q-1$개의 원소로 이루어진 벡터이다.

만약 $\tau(u_2)$의 함수형태가 u_2에 대한 선형함수라면 식 (3.1)의 모형은

$$\phi = \boldsymbol{\eta}_1'\boldsymbol{u}_1 + \tau_2 u_2 \tag{3.2}$$

으로 표현되고 각각의 관측값에 대해 선형결합 $\phi_i = \boldsymbol{\eta}_1'\boldsymbol{u}_{1i} + \tau_2 u_{2i}$를 이용한 로지스틱회귀의 적합값을 $\hat{y}_i$라고 하자. Cook과 Croos-Dabrera (1998)는 편잔차 $\mathrm{pres}_i = \hat{e}_i + \hat{\eta}_2 u_{2i}$를 정의하였다. 여기서 $\hat{e}_i = (y_i - \hat{y}_i)/\hat{y}_i(1-\hat{y}_i)$이다.

추가변수 u_2의 편잔차그림(partial residual plot)은 pres의 값을 세로축으로 하고 추가변수 u_2를 가로축으로 하는 $\{\text{pres}_i, u_{2i}\}$의 산점도이다. 모형 (3.2)에서의 $\boldsymbol{\eta}_1$이 모형 (3.1)에서의 실제모수 $\boldsymbol{\eta}_1^*$에 근접하다면 편잔차그림은 $\tau(u_2)$의 형태를 잘 묘사하게 된다. 하지만 $E(\boldsymbol{u}_1|u_2)$가 u_2에 대해 선형함수가 아니면 편잔차그림은 $\tau(u_2)$의 형태를 파악하기에 적합하지 못하다. 만약 조건부 기대값 $E(\boldsymbol{u}_1|u_2)$가 선형형태로 나타나지 않는 경우에는 변환의 필요성을 확인하기 위하여 CERES 그림이 사용된다.

각각의 관측값에 대해 다음의 선형결합

$$\phi_i = \boldsymbol{\eta}_1' \boldsymbol{u}_{1i} + \boldsymbol{\alpha}' \boldsymbol{m}(u_{2i}), \quad i = 1, \ldots, n \tag{3.3}$$

을 생각해보자. $\boldsymbol{m}(u_2) = \{m_j(u_2)\}$는 $(q-1) \times 1$의 벡터이고 조건부기대값 $E(\boldsymbol{u}_1|u_2)$로 정의하며 $\boldsymbol{m}(u_2)$의 값은 사전에 알려져 있지 않았기 때문에 $m_j(u_2)$, $j = 1, \ldots, q-1$는 $\boldsymbol{u}_1$의 각 원소인 u_{1j}를 반응변수로 하고 u_2에 의해 비모수회귀시켜 구한 적합값으로 추정한다. 즉, $\hat{\boldsymbol{m}}(u_2) = \{\hat{m}_j(u_2)\} = \{\hat{E}(u_{1j}|u_2)\}$, $j = 1, \ldots, q-1$이다. 이러한 추정은 lowess 방법을 사용하여 구할 수 있다. $\boldsymbol{m}(u_2)$를 $\hat{\boldsymbol{m}}(u_2)$로 대체하면 식 (3.3)은 다음과 같이 표현된다.

$$\phi_i = \boldsymbol{\eta}_1' \boldsymbol{u}_{1i} + \boldsymbol{\alpha}' \hat{\boldsymbol{m}}(u_{2i}), \quad i = 1, \ldots, n. \tag{3.4}$$

이 선형결합을 이용한 로지스틱회귀모형의 적합값을 $\hat{y}_i$라고 하고 $\tilde{e}_i = (y_i - \hat{y}_i)/\hat{y}_i(1-\hat{y}_i)$를 정의하면 CERES는 다음과 같다.

$$\text{ceres}_i = \tilde{e}_i + \hat{\boldsymbol{\alpha}}' \hat{\boldsymbol{m}}(u_{2i})$$

추가변수 u_2의 CERES 그림은 ceres의 값을 세로축으로 하고 추가변수 u_2를 가로축으로 하는 $\{\text{ceres}_i, u_{2i}\}$의 산점도이다. 이 때 $\hat{\boldsymbol{\alpha}}$는 $\hat{\boldsymbol{m}}(u_2)$의 회귀계수추정값 벡터이다.

모형 (3.3)이 참모형이라면 CERES 그림의 세로축의 기댓값은

$$E(\text{ceres}_i|u_{2i}) = E(\boldsymbol{u}_{1i}|u_{2i})'(\boldsymbol{\eta}_1^* - \boldsymbol{\eta}_1) + \boldsymbol{\tau}(u_{2i})$$

로 나타낼 수 있다. $\boldsymbol{\eta}_1$이 참모형에서의 실제모수 $\boldsymbol{\eta}_1^*$에 근접하면 CERES 그림이 $\tau(u_2)$의 형태를 잘 묘사하게 된다.

4. 모의자료를 통한 검토

여기서는 다양한 형태의 모의자료를 이용하여 3장에서 살펴본 여러 가지 그래픽적 방법이 어떠한 결과를 나타내는지 비교하여 본다. 또한 5장에서는 실제자료를 통해서도 그래픽적 방법의 결과를 알아본다. 로지스틱회귀모형의 설명변수의 수는 많은 경우도 있으나 본 논문에서는 두 개의 설명변수가 있는 경우만을 생각해본다.

4.1. 그래픽적 방법들과 추론의 결과가 일치하는 자료 (모의자료 1)

$N(2,2)$와 $N(8,9)$에서 각각 100개씩의 x_1을 생성하고 $G(1,2)$와 $G(1,5)$에서 각각 100개씩의 x_2를 생성한다. y는 선형결합 $\phi = 6x_1 - 2x_2 - x_1^2 + 28 \log x_2$를 이용하여 $\theta(\boldsymbol{x}) = \exp(\phi)/(1+\exp(\phi))$의 값을 계산하고 $\theta(\boldsymbol{x})$의 확률을 갖는 베르누이 시행으로 0 또는 1의 값을 생성한다. 먼저 일차항 x_1과 x_2만 포함하는 모형을 적용해보자. 로지스틱회귀모형의 분석결과는 Table 4.1과 같다. 두 변수 모두 유의하게 나타나지만 이탈도를 확인해 보면 모형이 적절하지 않다고 판단된다.

Table 4.1. Fit of logistic regression with X_1, X_2

Coefficients	Estimate	Std.Error	EST/SE	P-value
Intercept	-0.0426	0.2739	-0.156	0.8764
X_1	-0.1842	0.0374	-4.925	<.0001
X_2	0.1736	0.0789	2.201	0.0277
Number of cases:				200
Degrees of freedom:				197
Pearson χ^2:				247.361
Deviance:				234.990

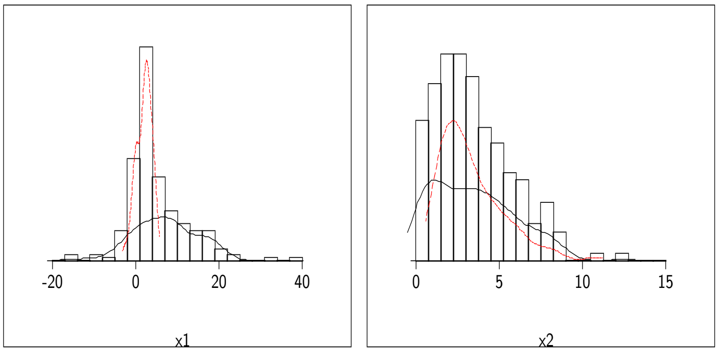

Figure 4.1. Histogram for x_1 and x_2 (solid line: $y=0$, dotted line: $y=1$).

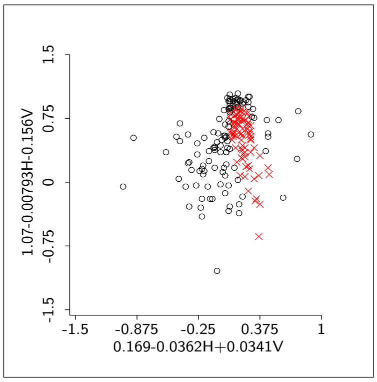

Figure 4.2. Binary response plot ($\circ$: $y=0$, $\times$: $y=1$).

Figure 4.1에서 x_1의 히스토그램은 y가 각각 1 또는 0인 경우 모두 정규분포의 형태를 보이지만 분산이 큰 차이를 보이므로 x_1^2이 포함되어야 한다고 볼 수 있다. x_2의 히스토그램에서는 감마분포의 형태로 나타나며 두 경우의 치우친 정도에 차이가 크므로 $\log x_2$가 포함되어야 한다고 볼 수 있다. Figure 4.2는 일차항만 포함되는 로지스틱회귀모형의 이항반응그림이다. 특정 각도에서 각 조각 내의 $y=1$에 해당하는 점들이 아래로 치우쳐 있는 형태로 나타나는 것을 확인할 수 있다. 따라서 1차원 구조라고 볼 수 없고 모형이 적절하지 않다고 판단된다. 모형이 적절하다면 카이잔차산점도에서 평균함수를 나타내는 평활곡선이 직선으로 나타나야 하는데 Figure 4.3에서 평활곡선이 휘어지게 나타나 모형이 적절하지 않다고 할 수 있다. 주변모형산점도에서 모형에 대한 가정을 하지 않고 추정한 3개의 곡선 $\hat{E}_F$,

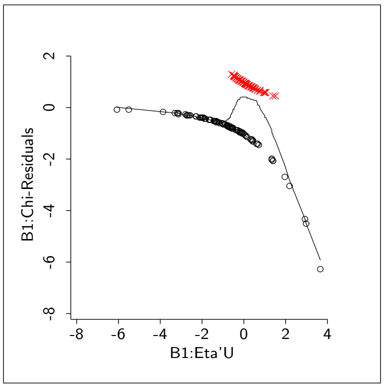

Figure 4.3. Chi residual plot (∘: $y = 0$, ×: $y = 1$).

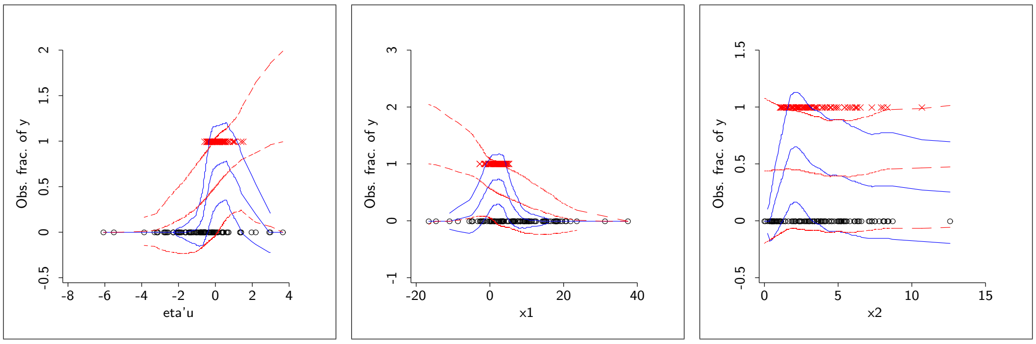

Figure 4.4. Marginal model plot (∘: $y = 0$, ×: $y = 1$).

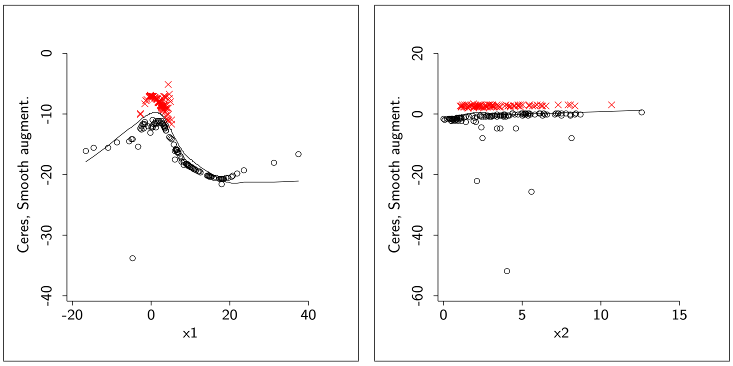

Figure 4.5. CERES plot for x_1, x_2 (∘: $y = 0$, ×: $y = 1$).

$\hat{E}_F + \mathrm{SD}_F$, $\hat{E}_F - \mathrm{SD}_F$과 모형에 대한 가정을 하고 추정한 3개의 곡선 $\hat{E}_M$, $\hat{E}_M + \mathrm{SD}_M$, $\hat{E}_M - \mathrm{SD}_M$이 일치한다면 모형이 적절하다고 할 수 있다. 하지만 Figure 4.4에서 차이가 많이 나타나므로 가정한 모형이 적절하지 않다고 판단된다. CERES 그림을 적용해 본 결과는 Figure 4.5와 같다. x_1의 CERES 그림을 보면 곡선의 형태가 보이므로 x_1의 경우 이차항이나 유사한 형태의 변환을 시도해 볼 필요가 있음을 시각적으로 확인할 수 있다.

Table 4.2. Fit of logistic regression with $X_1, X_2, X_1^2, \log X_2$

Coefficients	Estimate	Std.Error	EST/SE	P-value
Intercept	1.7544	2.2257	0.788	0.4306
X_1	4.8064	1.6825	2.857	0.0043
X_2	-4.6208	3.0037	-1.538	0.1240
X_1^2	-1.6078	0.5491	-2.928	0.0034
$\log X_2$	29.6447	12.1592	2.438	0.0148
Number of cases:				200
Degrees of freedom:				195
Pearson χ^2:				20.426
Deviance:				12.932

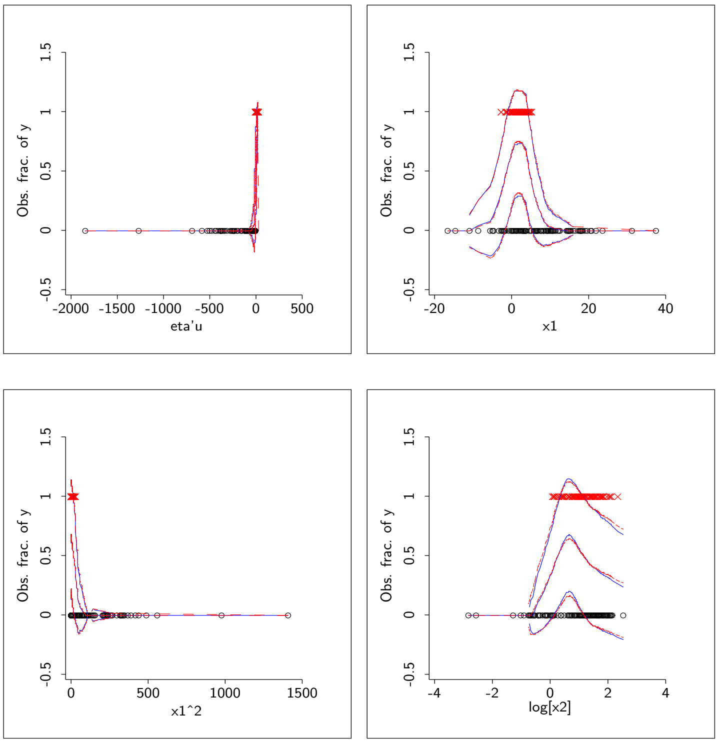

Figure 4.6. Marginal model plot (○: $y = 0$, ×: $y = 1$).

이 자료에서는 5가지 그래프가 모두 모형이 적절하지 않음을 나타내고 있다. 일차항 x_1, x_2에 x_1^2, $\log x_2$를 추가한 로지스틱회귀모형의 분석결과인 Table 4.2를 보면 x_2를 제외한 세 변수가 모두 유의적이고 이탈도에서도 모형의 적절함을 확인할 수 있다. 이항반응그림은 Figure 4.2와 달리 $y = 1$을 나타내는 점들의 상대적 밀도가 모든 조각 안에서 거의 일정하게 나타나는 것을 확인할 수 있으므로 1차원 구조라고 볼 수 있으며 모형이 적절하다고 판단된다. 카이잔차산점도에서는 평활곡선이 거의 완벽한 직선으로 나타나는 것을 확인할 수 있다. 주변모형산점도인 Figure 4.6에서도 구체적인 가정이 있는 경우와 없는 경우의 3개의 곡선이 모든 경우에서 일치하는 것을 확인할 수 있다. CERES 그림인 Figure 4.7을 보면 모두 직선으로 더 이상 추가적인 변환의 필요성이 없음을 나타낸다.

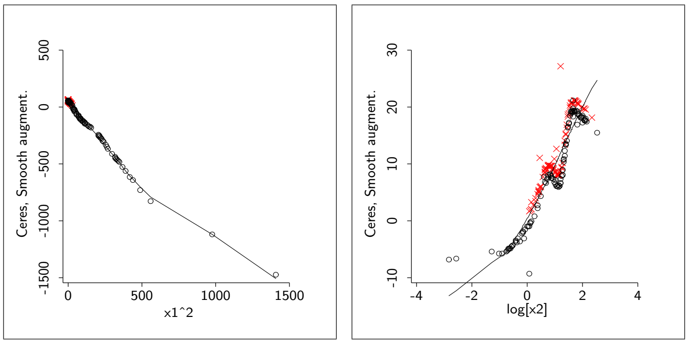

Figure 4.7. CERES plot for x_1^2, $\log x_2$ (○: $y = 0$, ×: $y = 1$).

4.2. 그래픽적 방법들과 추론의 결과가 다른 자료 (모의자료 2)

$N(3,2)$와 $N(8,9)$에서 각각 100개씩의 x_1을 생성하고 $G(1,2)$와 $G(1,5)$에서 각각 100개씩의 x_2를 생성한다. y는 선형결합 $\phi = 12x_1 - 3x_2 - 2x_1^2 + 5\log x_2$로 하고 $\theta(\boldsymbol{x})$의 확률을 갖는 베르누이 시행으로 0 또는 1의 값을 생성한다. 일차항 x_1과 x_2만 포함하는 로지스틱회귀모형의 분석결과는 두 변수 모두 유의하며 이탈도를 보면 대체로 모형이 적절하다고 볼 수 있다. 하지만 모형의 적절성을 판단하기 위한 그래픽적 방법들을 적용해 보면 다른 결과가 나타난다.

x_1과 x_2의 히스토그램은 Figure 4.1과 유사한 형태를 보이므로 x_1^2과 $\log x_2$가 포함되어야 한다고 볼 수 있다. 이항반응그림, 카이잔차산점도, 주변모형산점도은 각각 Figure 4.2, Figure 4.3, Figure 4.4와 유사한 형태를 보인다. 따라서 일차항만 포함하는 모형이 적절하지 않다는 것을 확인할 수 있다. CERES 그림에서 곡선들이 비선형함수 형태로 나타나는 것을 볼 수 있으므로 설명변수에 적절한 변환이 필요하다고 말할 수 있다. 특히 x_1의 CERES 그림은 평활곡선이 이차함수 형태로 나타나는 것으로 보아 히스토그램으로 판단한 결과와 마찬가지로 이차항을 포함하는 모형을 생각해 볼 수 있다. 일차항에 x_1^2, $\log x_2$를 추가한 로지스틱회귀모형에서 검정결과와 이탈도면에서 상당히 개선이 되고 모의자료1에서와 같이 5가지 그래프가 모두 모형이 적절함을 나타내고 있다.

일반적으로 회귀모형의 적절성을 판단할 때 가설검정 등의 추론을 통해 모형을 판단하는 것이 대부분이며 그래픽을 통한 검토는 생략하는 경우가 많다. 이 자료는 그러한 경우에 문제가 발생될 수 있음을 보여주고 있다. 수치를 통한 추론의 결과에서는 모형이 적절하다고 판단되더라도 그래픽적 방법들을 통해서도 모형이 적절한지 다시 한 번 판단해 보아야 한다. 이 모의자료에서와 같이 가설검정을 통해서는 찾아낼 수 없는 것들을 그래픽적인 방법들에서는 발견할 수 있는 경우가 있다. 따라서 여러 그래픽적 방법들을 적용하면 모형을 더욱 효과적으로 개선시킬 수 있을 것이다.

4.3. 그래픽적 방법들의 결과가 서로 다른 자료 (모의자료 3)

$N(4,1)$을 따르는 200개의 x_1과 $G(1,5)$를 따르는 200개의 x_2를 생성하고, y는 $\phi = 4x_1 - 3x_2 - x_1^2 + 10\log x_2$로 하고 $\theta(\boldsymbol{x})$의 확률을 갖는 베르누이 시행으로 0 또는 1의 값을 생성한다.

x_1의 히스토그램에서 y가 각각 1 또는 0인 경우 모두 정규분포의 형태를 보이며 분산이 크게 차이나지 않았다. 또한 x_2의 히스토그램에서 모두 감마 분포의 형태로 나타나고 치우친 정도가 비슷하다. 히스토그램만으로 판단한다면 모형에 변환된 형태를 포함시킬 필요가 없이 x_1과 x_2만으로 충분하다고 할 수 있다. 카이잔차산점도의 경우 평균함수가 직선에 가깝다고 볼 수 있으므로 일차항만 포함한 모형이 적

절하다고 할 수 있다.

하지만 로지스틱회귀분석의 결과를 보면 일차항만 포함한 모형이 적절하지 않은 것으로 나타난다. 차원구조를 판단하기 위한 이항반응그림에서 $y = 1$에 해당하는 점들이 상대적으로 중간 부분에 밀집되어 나타나므로 모형이 적절하지 않다고 판단된다. 주변모형산점도에서도 일차항만을 포함한 모형은 적절하지 못하다는 것을 알 수 있었다. CERES 그림을 보면 x_1과 x_2 모두 변수 변환이 필요하다는 것을 나타낸다.

일부 그래프를 이용한 방법에서 일차항만 포함하는 모형이 적절하다고 나타나지만 다른 방법에서는 적절하지 않은 것으로 나타나므로 CERES그림이 제시하는 x_1^2과 $\log x_2$를 추가하는 모형을 생각해보았다. 이 모형에 대한 검토결과 그래프를 이용하는 다섯 가지 방법 모두에서 모형이 적절한 것으로 나타난다.

이와 같이 여러 가지 그래픽적 방법들의 결과가 모두 같게 나오지 않을 수도 있으며 복수의 그래픽적 방법들을 통한 확인이 필요하다는 것을 알 수 있다.

4.4. 히스토그램을 통한 조건부 밀도의 결과가 오류인 자료 (모의자료 4)

$N(8, 2)$와 $N(7, 9)$에서 각각 100개씩의 x_1을 생성하고 $G(1, 2)$와 $G(1, 10)$에서 각각 100개씩의 x_2를 생성한다. y는 선형결합을 $\phi = 5x_1 - 11x_2$로 하고 $\theta(\boldsymbol{x})$의 확률을 갖는 베르누이 시행으로 0 또는 1의 값을 생성한다.

Figure 4.1에서와 유사하게 x_1의 히스토그램은 y가 각각 1 또는 0인 경우 모두 정규분포의 형태를 보이지만 분산이 큰 차이를 보이므로 x_1^2이 포함되어야 한다고 볼 수 있다. x_2의 히스토그램에서는 감마분포의 형태로 나타나며 두 경우의 치우친 정도에 차이가 크므로 $\log x_2$가 포함되어야 한다고 볼 수 있다.

x_1, x_2만을 포함하는 로지스틱회귀모형의 분석결과는 두 변수 모두 유의하게 나타나고 모형의 적합도를 확인해 보더라도 모형은 적절하다고 판단된다. 이항반응그림, 카이잔차산점도, 주변모형산점도, CERES 그림에서도 모형은 적절한 것으로 나타난다.

이 자료는 x_1, x_2만이 포함되었을 경우에 모형이 적절하게 나타나도록 생성된 자료였으므로 두 설명변수를 포함한 모형이 당연히 적절한 모형이라 할 수 있다. 히스토그램으로 조건부 밀도를 확인했을 때는 변수의 변환이 필요하다고 생각될 수 있었지만 다른 그래픽적 방법들은 모형의 적절성을 나타내고 있다. 이와 같이 모형이 적절한 경우에서도 조건부 밀도로 확인해 볼 때에는 적절하지 않게 나올 수 있고 이것만으로 모형의 적절성을 평가하는 것은 오류를 범할 수 있다. 따라서 히스토그램으로 조건부 밀도를 보는 방법 외에도 다른 그래픽적 방법들을 사용하여 모형을 평가해야 한다. 한 가지 방법만을 통해 모형을 평가하기 보다는 여러 방법들을 모두 적용시키는 것이 모형평가의 신뢰를 높일 수 있다.

4.5. 히스토그램을 통한 조건부 밀도의 결과가 오류인 자료 (모의자료 5)

$N(2, 2)$을 따르는 200개의 x_1과 $G(1, 2)$를 따르는 200개의 x_2를 생성하고, y는 선형결합을 $\phi = 2x_1^2 + 2\log x_2$로 하고 $\theta(\boldsymbol{x})$의 확률을 갖는 베르누이 시행으로 0 또는 1의 값을 생성한다.

모의자료 3과 같이 x_1의 히스토그램에서 y가 각각 1 또는 0인 경우 모두 정규분포의 형태를 보이며 분산이 크게 차이나지 않는다. 또한 x_2의 히스토그램에서 모두 감마 분포의 형태로 나타나고 치우친 정도가 비슷하다. 히스토그램만으로 판단한다면 모형에 변환된 형태를 포함시킬 필요가 없이 x_1과 x_2만으로 충분하다고 할 수 있다.

하지만 로지스틱회귀분석의 결과를 보면 모형이 적절하지 않다고 나타난다. 또한 차원확인, 카이잔차산점도, 주변모형산점도에서도 x_1과 x_2만 포함한 로지스틱회귀모형이 적절하지 않다는 것을 알 수 있다.

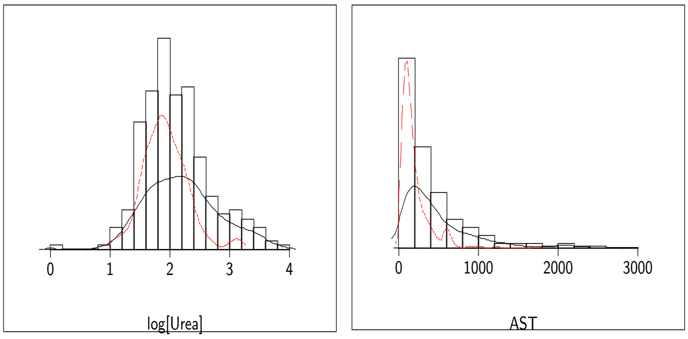

Figure 5.1. Histogram for log(Urea) and AST (solid line: $y = 0$, dotted line: $y = 1$).

Table 5.1. Fit of logistic regression with log(Urea), AST

Coefficients	Estimate	Std.Error	EST/SE	P-value
Intercept	2.6085	0.6585	3.961	0.0001
log(Urea)	−0.9611	0.2990	−3.214	0.0013
AST	−0.0044	0.0009	−4.953	<.0001
Number of cases:				263
Degrees of freedom:				260
Pearson χ^2:				591.785
Deviance:				288.938

x_1의 CERES 그림에서 평활곡선이 이차함수의 형태로 나타나고 x_2에 어떠한 변환이 필요한지는 분명하지 않지만 변환이 필요하다는 것을 확인할 수 있다. 위의 결과를 바탕으로 본다면 히스토그램을 확인했을 때는 모형이 적절한 것처럼 보이더라도 다른 방법들을 꼭 확인해봐야 한다는 것을 알 수 있다.

5. 실제자료를 이용한 분석

여기서는 실제자료를 이용하여 3절에서 알아본 여러 가지 그래픽적 방법들이 어떠한 결과를 나타내는지 비교해본다. Cook과 Weisberg (1994)는 1983년과 1984년에 뉴질랜드에서 광우병에 걸린 소 435마리의 생존 여부에 관한 자료를 제시하고 분석하였다. 변수들은 outcome($0 =$ died or killed, $1 =$ survived), 아스파르테이트 아미노전이효소(AST), 측정이 이루어졌을 때의 광우병이 지속된 일 수(Daysrec), 측정된 크레아틴포스포키네이스 혈청(CK), 혈청 요소(Urea) 등이 있으며, 이 중 0과 1의 값을 갖는 outcome을 반응변수로 사용하고, log(Urea)와 AST을 설명변수로 사용한다.

먼저 outcome에 따른 두 변수의 히스토그램은 Figure 5.1과 같다. log(Urea)의 경우 반응변수인 outcome결과에 따라 나누어 보았을 때 두 경우 모두 정규분포의 형태를 보이며 분산의 차이를 보인다. AST의 히스토그램은 두 경우에 모두 감마분포 형태이나 치우쳐진 정도에 차이가 나타난다. 따라서 두 설명변수만을 포함한 로지스틱회귀모형은 부족할 것으로 판단된다. 일차항만을 포함하는 경우 로지스틱회귀분석의 결과는 Table 5.1과 같다. 두 설명변수 모두 유의하지만 모형은 적절하지 않은 것으로 판단된다. Figure 5.2의 이항반응그림, Figure 5.3의 카이잔차산점도, Figure 5.4의 주변모형산점도 모두에서 두 설명변수만을 포함한 로지스틱회귀모형이 적절하지 않음을 볼 수 있다. 또한 Figure 5.5의 CERES 그림도 직선이 아니므로 모형에 개선이 필요하다는 것은 알 수 있지만 정확한 변환의 형태를 파악하기는 쉽지 않았다.

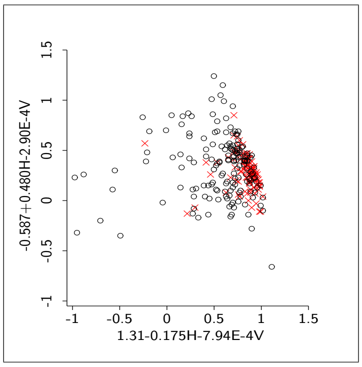

Figure 5.2. Binary response plot (○: $y = 0$, ×: $y = 1$).

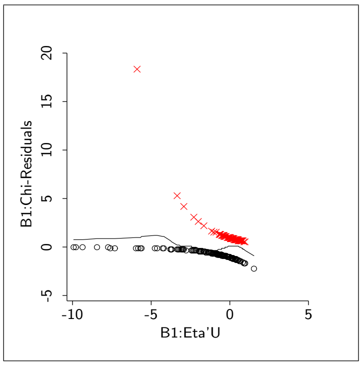

Figure 5.3. Chi residual plot (○: $y = 0$, ×: $y = 1$).

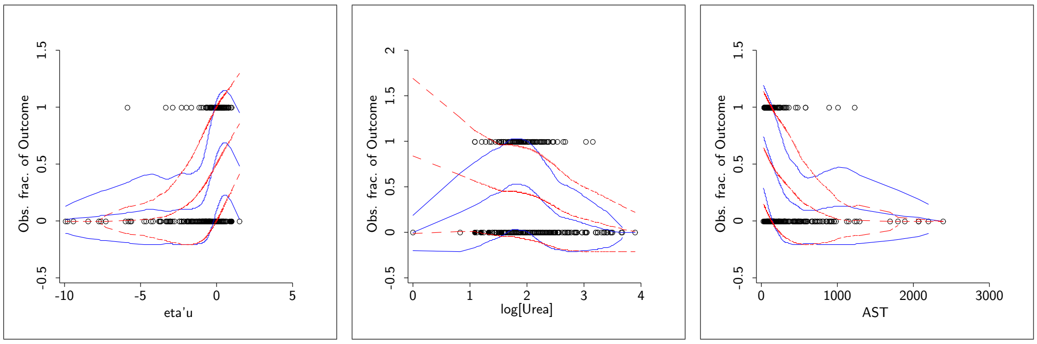

Figure 5.4. Marginal model plot (○: $y = 0$, ×: $y = 1$).

위의 결과들을 종합하여 판단하면 변환된 $[\log(\text{Urea})]^2$과 log(AST)항이 필요한 것으로 판단된다. 이러한 변환된 항을 추가한 모형을 적합한 결과 이탈도에서 많은 개선이 있었고 5가지 그래프를 이용한 방법에서 모두에서 이 모형에 적절하다고 결론내릴 수 있다.

이번에는 같은 자료에서 log(CK)와 Daysrec을 설명변수로 선택하여 모형을 적용해 보았다. log(CK)의 히스토그램은 outcome 결과에 따라 모두 정규분포의 모습을 보이며 분산은 차이가 없어 보인다.

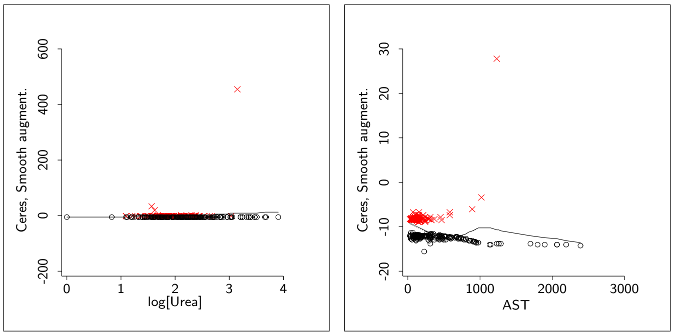

Figure 5.5. CERES plot for log(Urea) and AST (∘: $y = 0$, ×: $y = 1$).

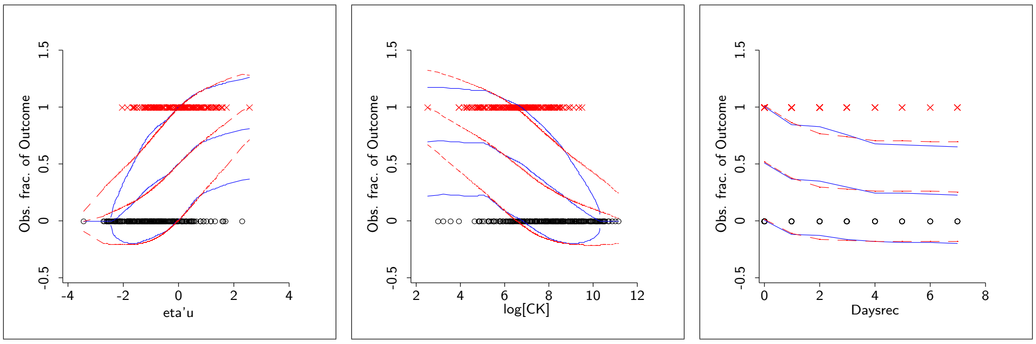

Figure 5.6. Marginal model plot (∘: $y = 0$, ×: $y = 1$).

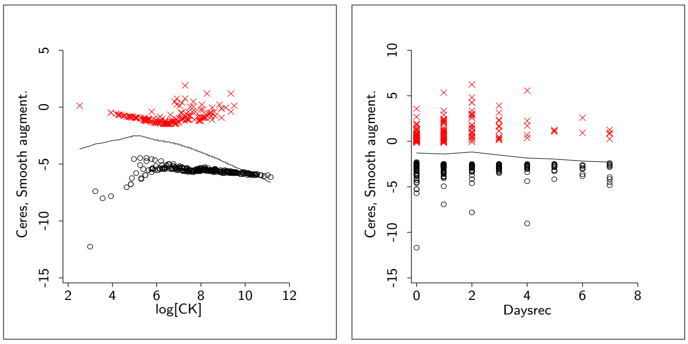

Figure 5.7. CERES plot for log(CK) and Daysrec (∘: $y = 0$, ×: $y = 1$).

Daysrec의 히스토그램은 감마분포 형태이며 치우쳐진 정도에 차이가 없는 것으로 보인다. 따라서 히스토그램을 통한 방법으로는 두 설명변수만을 포함한 로지스틱회귀모형이 충분할 것으로 판단된다. 하지만 로지스틱회귀분석의 추론 결과에서는 모형은 적절하지 않은 것으로 나타났다.

이항반응그림으로 차원 구조를 확인해 본 결과 1차원 구조라 볼 수 없었고 카이잔차산점도에서도 두 설명변수만을 포함한 모형이 적절하지 않음을 볼 수 있었다. 주변모형산점도인 Figure 5.6을 보면 모형이 적절하지 않은 것으로 나타나고 특히 log(CK)의 변수에 개선이 필요한 것으로 보인다. CERES 그림인 Figure 5.7에서도 log(CK)의 변수에 변환이 요구된다.

히스토그램을 통해 조건부 밀도를 확인했을 때는 변수에 변환이 필요 없다고 보았지만 위의 결과들을 종합하여 판단하면 두 설명변수만을 포함한 로지스틱회귀모형으로는 자료를 설명하기에 부족하다고 판단되며 log(CK)의 제곱항이 추가되어야함을 알 수 있다. 개선된 모형의 분석 결과는 다섯 가지 모든 그래프를 이용하는 방법에서 모형의 적절함을 보였고 추론의 결과에서도 문제가 없음이 확인되었다.

6. 결론

일반적으로 로지스틱회귀모형을 평가할 때 검정통계량을 사용한 검정방법이 주로 사용되지만 이와 같은 통계적 추론만으로 놓칠 수 있는 부분이 많다. 이에 대한 보완을 위해서 수치를 이용한 추론 외에 모형을 진단하기 위한 그래픽적 방법들의 사용이 요구된다. 그래픽적 방법들을 사용함에 있어서 여러 방법의 결과가 항상 모두 일치하는 것은 아니다. 따라서 정확한 회귀모형의 진단을 위해서는 여러 가지 그래픽적 방법들을 다양하게 적용하는 것이 필요하다.

본 논문에서는 로지스틱회귀모형을 평가하기 위해 사용되는 그래픽적 방법들인 히스토그램을 사용한 조건부 밀도 확인, 이항반응그림을 통한 차원 구조 확인, 카이잔차산점도, 주변모형산점도, CERES 그림을 소개하였다. 히스토그램으로 조건부 밀도를 확인하는 경우 변환된 변수의 추가여부를 판단할 수 있다. 이항반응그림으로 차원 구조를 확인하는 경우에는 주어진 모형의 차원 구조를 확인함으로써 모형의 적절성을 판단할 수 있고, 카이잔차산점도와 주변모형산점도를 이용하는 경우에도 그래프를 통하여 주어진 모형의 적절성을 판단할 수 있다. 또한 CERES 그림을 사용하면 변수에 어떠한 변환을 시도해야 하는지를 판단할 수 있다.

여러 상황의 모의자료를 통하여 다섯 가지 그래픽적 모형진단 방법들의 결과가 각각의 조건 하에서 다르게 나타날 수 있다는 것을 확인하였고 실제자료들에 적용해 보았다. 추론의 결과와 그래픽적 방법들의 결과가 일치할 때도 있었지만 다르게 나오는 경우도 있었다. 주어진 모형이 적절하지 않은 상황에서 추론의 결과는 모형이 적절하다고 나타냈지만 그래픽적 방법들은 모형이 적절하지 않음을 나타내는 경우가 있었다. 이를 통해 일반적으로 사용되는 가설검정의 결과만으로는 정확한 진단이라고 결론내릴 수 없으므로 여러 그래픽적 방법을 적용해야 한다는 것을 확인하였다. 또한 여러 그래픽적 방법들도 서로 다른 결과들을 보일 수도 있음을 확인하였으며 하나의 그래픽적 방법만을 사용한다면 잘못된 결론을 내리게 되는 위험이 따를 수 있다는 것을 확인하였다. 결론적으로 통계적 추론과 여러 그래픽적 방법의 상호보완적인 사용이 추천된다.

References

Atkinson, A. C. (1985). *Plots, Transformations and Regression*, Oxford University Press, Oxford.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics*, Wiley, New York.

Cleveland, W. S. 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.

Cook, R. D. (1993). Exploring partial residual plots, *Technometrics*, **35**, 351–362.

Cook, R. D. (1998). *Regression Graphics: Idea for Studying Regressions through Graphics*, Wiley, New York.

Cook, R. D. and Croos-Dabrera, R. (1998). Partial residual plots in generalized linear models, *Journal of the American Statistical Association*, **93**, 730–739.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*, Chapman & Hall, London.

Cook, R. D. and Weisberg, S. (1994). *An Introduction to Regression Graphics*, Wiley, 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*, Wiley, New York.

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

Kay, R. and Little, S. (1987). Transformations of the explanatory variables in the logistic regression model for binary data, *Biometrika*, **74**, 495–501.

Scrucca, L. (2003). Graphics for studying logistic regression models, *Statistical Methods and Applications*, **11**, 371–394.

Scrucca, L. and Weisberg, S. (2004). A simulation study to investigate the behavior of the log-density ratio under normality, *Communication in Statistics Simulation and Computation*, **33**, 159–178.

Tierney, L. (1990). *Lisp-Stat: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics*, Wiley, New York.

로지스틱회귀모형의 평가를 위한 그래픽적 방법

김경진[a] · 강명욱[a,1]

[a]숙명여자대학교 통계학과

(2015년 10월 19일 접수, 2015년 12월 4일 수정, 2015년 12월 8일 채택)

요 약

대부분의 통계분석방법은 요약통계량에 의존하지만 그래픽적 방법을 이용하면 자료의 특성을 파악하기 쉽고 통계량만으로는 알아낼 수 없는 부분까지도 접근이 가능하다. 그래프를 통한 로지스틱회귀모형의 평가 방법으로 로그-밀도비를 통한 검토, 차원 검토, 주변모형산점도, 카이잔차산점도, CERES 그림을 알아보고 모의자료들을 통해 다양한 상황에서 그래픽적 방법들 어떠한 결과를 나타내지를 비교 검토한다.

주요용어: 로그-밀도비, 이항반응그림, 주변모형산점도, 차원구조, 카이잔차산점도, CERES 그림

본 연구는 숙명여자대학교 교내연구비 지원에 의해 수행되었음 (1-1403-0001).
[1]교신저자: (04310) 서울시 용산구 청파로47길 100, 숙명여자대학교 통계학과. E-mail: mwkahng@sm.ac.kr