Journal of the Korean Data & Information Science Society
2015, **26**(1), 141–150

http://dx.doi.org/10.7465/jkdi.2015.26.1.141
한국데이터정보과학회지

로지스틱회귀에서 잔차산점도를 이용한 모형평가[†]

강명욱[1]

[1]숙명여자대학교 통계학과

접수 2014년 12월 16일, 수정 2015년 1월 2일, 게재확정 2015년 1월 10일

요 약

로지스틱회귀에서 모형을 평가하거나 진단할 때 가설검정이 주로 사용되지만 이것만으로는 놓칠 수 있는 부분이 많고 이에 대한 보완을 위하여 그래픽적 방법의 사용이 요구된다. 그래프를 이용한 모형의 적절성 평가를 위한 도구로 잔차산점도가 널리 이용되고 있으나 적용 범위가 선형회귀에 국한되는 문제점이 있다. 해결 방안으로 주변모형산점도를 이용하여 모형의 적절성을 평가하는 방법이 있으나 역시 문제점을 가지고 있다. 본 논문에서는 주변모형산점도의 대안으로 카이잔차산점도를 제안하고 그 효용성을 알아본다.

주요용어: 가중잔차, 로지스틱회귀, 이항회귀, 잔차산점도, 주변모형산점도, 카이잔차산점도.

1. 서 론

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), Cleveland (1987)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. 또한 Kahng (2005)에 의해 일반화선형모형에서 그래프를 이용한 분석이 시도되었다.

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), Kahng과 Shin (2012)의 연구가 있는데 로그-밀도비 (log-density ratio)의 개념과 그래픽적 방법을 근거로 하고 있다.

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. 하지만 주변모형산점도는 작성하기가 복잡하고 이를 시현할 수 있는 소프트웨어가 많지 않다는 단점이 있다. 본 연구에서는 일반화선형모형 중에서 특히 이항반응변수를 가진 로지스틱회귀모형의 적절성을 평가하는 그래픽적 방법으로 카이잔차산점도를 이용한 모형평가 방법을 제시하고자 한다.

[†] 본 연구는 숙명여자대학교 교내연구비지원에 의해 수행되었음 (과제번호 1-1303-0010).
[1] (140-742) 서울특별시 용산구 청파로 47길 100, 숙명여자대학교 통계학과, 교수.
E-mail: mwkahng@sm.ac.kr

2. 로지스틱회귀모형

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

$$E(y/m|\boldsymbol{x}) = \theta(\boldsymbol{x}) = \frac{exp(\boldsymbol{\beta}^T\boldsymbol{x})}{1 + exp(\boldsymbol{\beta}^T\boldsymbol{x})} = m(\boldsymbol{\beta}^T\boldsymbol{x}) \tag{2.1}$$

모형 (2.1)에서는 $\boldsymbol{x}$의 선형결합인 $\boldsymbol{\beta}^T\boldsymbol{x}$의 함수로 모형을 구성하고 있으나 Cook과 Weisberg (1999)에서와 같이 $\boldsymbol{u}$의 선형결합인 $\boldsymbol{\eta}^T\boldsymbol{u}$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. 모형 (2.1)에서와 같이 로지스틱함수를 커널평균함수로 하면 다음과 같이 로지스틱회귀모형이 되며

$$E(y/m|\boldsymbol{x}) = \theta(\boldsymbol{x}) = \frac{exp(\boldsymbol{\eta}^T\boldsymbol{u})}{1 + exp(\boldsymbol{\eta}^T\boldsymbol{u})} = m(\boldsymbol{\eta}^T\boldsymbol{u}) \tag{2.2}$$

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

$$log\left(\frac{\theta(\boldsymbol{x})}{1 - \theta(\boldsymbol{x})}\right) = \boldsymbol{\eta}^T\boldsymbol{u} \tag{2.3}$$

성공할 경우 "1"을, 실패할 경우 "0"을 가지는 이항반응변수(binary response variable) y와 p개의 설명변수 $\boldsymbol{x} = (x_1, ..., x_p)^T$를 가지는 이항회귀 (binomial regression)를 생각하자. 이항변수인 y의 조건부분포는 기댓값이 $E(y|\boldsymbol{x}) = P(y = 1|\boldsymbol{x}) = \theta(\boldsymbol{x})$이 되는 베르누이분포를 따른다. Kay와 Little (1987)은 설명변수의 조건부분포, 즉 $\boldsymbol{x}|y$의 분포에 따라 $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$를 적절하게 선택하는 과정을 제시하였다. Cook과 Weisberg (1999), Scrucca와 Weisberg (2004)에 따르면, 설명변수가 하나이고 그 조건부분포가 정규분포라 하면 $\boldsymbol{u}^T = (1, x, x^2)$를 사용하고 분산이 같을 때에는 $\boldsymbol{u}^T = (1, x)$를 사용한다. 또한 Kahng과 Shin (2012)은 조건부분포가 좌우대칭이 아니면 감마분포로 보고 $\boldsymbol{u}^T = (1, x, log(x))$를 사용한다. 일반적인 로지스틱회귀에서 우도함수는

$$L = \prod_{i=1}^{n} \binom{m_i}{y_i} [\theta(\boldsymbol{x}_i)]^{y_i} [1 - \theta(\boldsymbol{x}_i)]^{m_i - y_i} \propto \prod_{i=1}^{n} [\theta(\boldsymbol{x}_i)]^{y_i} [1 - \theta(\boldsymbol{x}_i)]^{m_i - y_i}$$

이고, 우도함수에 로그를 취한 로그우도함수는 다음과 같다.

$$log(L) \propto \sum_{i=1}^{n} \left[y_i log\left(\frac{\theta(\boldsymbol{x}_i)}{1 - \theta(\boldsymbol{x}_i)}\right) + m_i log\left(1 - \theta(\boldsymbol{x}_i)\right) \right] \tag{2.4}$$

이제, 식 (2.3)을 이용하여 $\theta(\boldsymbol{u}_i)$를 치환하면 로그우도함수 (2.4)를 다음과 같이 $\boldsymbol{\eta}$의 함수로 표현할 수 있고

$$log(L(\boldsymbol{\eta})) \propto \sum_{i=1}^{n} \left[(\boldsymbol{\eta}^T\boldsymbol{u}_i) y_i - m_i log\left(1 + exp(\boldsymbol{\eta}^T\boldsymbol{u}_i)\right) \right] \tag{2.5}$$

로그우도함수 (2.5)를 최대화 시키는 $\boldsymbol{\eta}$의 최대우도추정량 $\widehat{\boldsymbol{\eta}}$을 찾을 수 있다. 여기서 $\widehat{\boldsymbol{\eta}}$은 $\boldsymbol{\eta}$의 최소제곱추정량도 된다. 이를 이용하면 $\theta(\boldsymbol{x}_i)$는 $\widehat{\theta}(\boldsymbol{x}_i) = exp(\widehat{\boldsymbol{\eta}}^T\boldsymbol{u}_i)/[1 + exp(\widehat{\boldsymbol{\eta}}^T\boldsymbol{u}_i)]$로 추정하고 로지스틱회귀에서 적합값 (fitted value)은 $\hat{y}_i = m_i\widehat{\theta}(\boldsymbol{x}_i)$가 된다.

3. 그래프를 이용한 모형평가

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

3.1. 주변모형산점도

2장에서와 같이 반응변수 y와 $\boldsymbol{x}$의 함수들로 생성되는 $q \times 1$벡터 $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$를 설명변수로 하는 회귀모형을 생각하자. 주변모형산점도의 기본 개념은 회귀모형을 두 가지 관점에 근거한 조건부 누적밀도함수의 비교를 통하여 평가하는 것이다. 자료가 독립적이고 동일한 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 대상이 된다.

Cook과 Weisberg (1997)는 다음과 같은 주변모형 확인조건 (marginal model checking condition)을 제시하였다. $\boldsymbol{u}$의 모든 값에 대하여 $F(y|\boldsymbol{u}) = M(y|\boldsymbol{\eta}, \boldsymbol{u})$이 성립하기 위한 필요충분조건은 모든 $\boldsymbol{a}^T\boldsymbol{u}$에 대하여 $F(y|\boldsymbol{a}^T\boldsymbol{u}) = M(y|\boldsymbol{a}^T\boldsymbol{u})$인 경우이다. 이는 완전모형을 나타내는 $F(y|\boldsymbol{u})$가 내포하는 모든 정보를 주변모형 $F(y|\boldsymbol{a}^T\boldsymbol{u})$가 설명할 수 있다는 것을 의미한다. 따라서 $(q+1)$차원의 산점도 대신 반응변수를 수직축으로 하고 설명변수의 선형결합 $\boldsymbol{a}^T\boldsymbol{u}$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 벡터 $\boldsymbol{a}$를 적절히 선택해야 한다. 선택을 위한 몇 가지 표준적인 방법이 있지만 그 중에서 기본적으로 사용하는 방법은 다음과 같다. 첫째, 회귀모형이 선형모형인 경우, 즉 $\boldsymbol{\eta}^T\boldsymbol{u}$로 설명할 수 있다고 가정하는 경우에 $\boldsymbol{\eta}$의 최소제곱추정값 $\widehat{\boldsymbol{\eta}}$을 이용하여 $\boldsymbol{a} = \widehat{\boldsymbol{\eta}}$으로 선택한다. 따라서 y를 수직축으로 하고 $\widehat{\boldsymbol{\eta}}^T\boldsymbol{u}$를 수평축으로 하는 산점도를 이용하여 모형을 평가할 수 있다. 둘째, $\boldsymbol{a}^T\boldsymbol{u}$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) $E_F(y|\boldsymbol{a}^T\boldsymbol{u})$와 모형에 대한 구체적인 가정을 하는 경우의 주변평균함수 $E_M(y|\boldsymbol{a}^T\boldsymbol{u})$를 생각하자. y를 수직축으로 하고 $\boldsymbol{a}^T\boldsymbol{u}$를 수평축으로 하는 산점도에서 대표적인 평활 방법의 하나인 lowess (locally weighted scatterplot smoother; Cleveland와 Devlin, 1988)를 이용하여 주변평균함수를 $\widehat{E}_F = \widehat{E}_F(y|\boldsymbol{a}^T\boldsymbol{u})$와 $\widehat{E}_M = \widehat{E}_M(y|\boldsymbol{a}^T\boldsymbol{u})$로 추정할 수 있다. 또한, 모형에 대한 가정 여부에 따른 주변분산함수 (marginal variance function)를 $Var_F(y|\boldsymbol{a}^T\boldsymbol{u})$와 $Var_M(y|\boldsymbol{a}^T\boldsymbol{u})$라 하면 평활방법을 이용하여 $(SD_F)^2 = \widehat{Var}_F(y|\boldsymbol{a}^T\boldsymbol{u})$와 $(SD_M)^2 = \widehat{Var}_M(y|\boldsymbol{a}^T\boldsymbol{u})$를 추정할 수 있다.

모형의 평가는 y를 수직축으로 하고 $\boldsymbol{a}^T\boldsymbol{u}$를 수평축으로 하는 산점도에 모형에 대한 가정을 하지 않고 추정한 3개의 곡선 $\widehat{E}_F$, $\widehat{E}_F + SD_F$, $\widehat{E}_F - SD_F$과 모형에 대한 가정을 하고 추정한 $\widehat{E}_M$, $\widehat{E}_M + SD_M$, $\widehat{E}_M - SD_M$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.

3.2. 잔차산점도

평균함수가 $E(y_i|\boldsymbol{x}_i) = \boldsymbol{\eta}^T\boldsymbol{u}_i$이고 분산함수가 $Var(y_i|\boldsymbol{x}_i) = \sigma^2$인 다음의 선형모형에서

$$y_i|\boldsymbol{x}_i = \boldsymbol{\eta}^T\boldsymbol{u}_i + \epsilon_i, i = 1, ..., n \tag{3.1}$$

오차항 ϵ_i의 평균과 분산은 $E(\epsilon_i) = 0$과 $Var(\epsilon_i) = \sigma^2$이며 잔차 e_i는 다음과 같다.

$$e_i = y_i - \widehat{y}_i \tag{3.2}$$

잔차는 회귀진단의 도구로 사용되는 대표적인 통계량으로 잔차 e_i를 수직축으로 하고 적합값 $\widehat{y}_i$를 수평축으로 하는 잔차산점도는 그래프를 이용한 회귀진단에서 사용되는 대표적인 도구이다. 잔차의 기댓값은 오차항과 같이 $E(e_i) = 0$이지만 분산은 $Var(e_i) = \sigma^2(1 - h_{ii})$로 등분산이 되지 않는다. 따라서 모형이 적절하더라도 분산이 일정하지 않고 레버리지 값 (leverage value) h_{ii}에 의존하므로 잔차산점도의 형태가 왜곡될 수 있어 이에 대한 개선이 필요하다. 잔차를 표준화하는 스튜던트화잔차 (Studentized residual) $r_i = e_i/\left(\widehat{\sigma}\sqrt{1 - h_{ii}}\right)$의 이용으로 문제 해결이 가능하다. Belsley 등 (1980), Cook과 Weisberg (1982), Weisberg (2005) 등에서는 잔차산점도의 수직축에 스튜던트화잔차의 사용을 제안하고 있다. 모형 (3.1)과 달리 등분산 가정을 할 수 없는 상황에서는 다음의 선형모형을 생각해 볼 수 있고

$$y_i|\boldsymbol{x}_i = \boldsymbol{\eta}^T\boldsymbol{u}_i + \epsilon_i/\sqrt{w_i}, i = 1, ..., n \tag{3.3}$$

분산함수는 다음과 같이 가중치에 의존한다.

$$Var(y_i|\boldsymbol{x}_i) = Var((\epsilon_i/\sqrt{w_i})|\boldsymbol{u}_i) = \sigma^2/w_i \tag{3.4}$$

하지만 가중치와 관계없이 오차 ϵ_i는 등분산을 가진다. 우리는 모형 (3.3)의 다중선형회귀모형에서 일반적으로 오차항 ϵ_i와 설명변수 $\boldsymbol{x}_i$가 독립적이라는 가정을 한다. 이러한 가정하에서는 다음의 두 조건이 만족하게 될 것이다.

$$E(\epsilon_i|\boldsymbol{x}_i) = E(\epsilon_i) = 0 \tag{3.5}$$

$$Var(\epsilon_i|\boldsymbol{x}_i) = Var(\epsilon_i) = \sigma^2 \tag{3.6}$$

식 (3.5)과 (3.6)은 평균함수는 0이고 분산함수는 일정함을 나타내고 있다. 오차 ϵ_i를 알 수 있다면 오차 ϵ_i와 $\boldsymbol{x}_i$의 산점도를 통해 모형의 타당성을 파악할 수 있다. 만약 산점도가 두 조건 (3.5)과 (3.6)에 어긋남을 보인다면 모형의 타당성을 보장할 수 없을 것이다. 오차 ϵ_i가 관측할 수 있거나 계산하여 그 값들은 얻을 수 있다면 이를 이용하여 모형의 타당성의 검토할 수 있을 것이다. 하지만 오차의 값을 알 수 없으므로 우리는 다음과 같이 정의되는 잔차를 통해 오차를 추정할 수 있을 것이다. 우선 $\widehat{y}_i = \widehat{\boldsymbol{\eta}}^T\boldsymbol{u}_i$를 모형 (3.3)에서 가중최소제곱법에 의해 구한 i번째 적합값이라고 하고 잔차 (3.2)의 변형된 형태인 다음의 가중잔차 (weighted residual)를 정의하자.

$$e_i^w = \sqrt{w_i}(y_i - \widehat{y}_i) \tag{3.7}$$

이 잔차를 통해 어떻게 오차 ϵ_i의 추정이 가능하지를 알아보기 위해 y_i를 $\boldsymbol{\eta}^T\boldsymbol{u}_i + \epsilon_i/\sqrt{w_i}$로 대체시키면 가중잔차 e_i^w는 다음과 같다.

$$\begin{aligned} e_i^w &= \sqrt{w_i}(y_i - \widehat{y}_i) \\ &= \sqrt{w_i}([\boldsymbol{\eta}^T\boldsymbol{u}_i + \epsilon_i/\sqrt{w_i}] - \widehat{\boldsymbol{\eta}}^T\boldsymbol{u}_i) \\ &= \epsilon_i + \sqrt{w_i}(\boldsymbol{\eta}^T\boldsymbol{u}_i - \widehat{\boldsymbol{\eta}}^T\boldsymbol{u}_i) \\ &= \epsilon_i + \delta_i \end{aligned}$$

이제 i번째 가중잔차 e_i^w는 오차 ϵ_i와 δ_i=$\sqrt{w_i}(\boldsymbol{\eta}-\widehat{\boldsymbol{\eta}})^T\boldsymbol{u}_i$의 합이라고 할 수 있다. 모형의 적절성과 상관없이 $E(\delta_i|\boldsymbol{x}_i)$=0이므로 가중잔차 e_i^w의 평균함수와 오차의 ϵ_i평균함수는 같다. 즉 $E(e_i^w|\boldsymbol{x}_i)$=$E(\epsilon_i|\boldsymbol{x}_i)$이다. 따라서 가중잔차 e_i^w와 $\boldsymbol{x}$의 함수인 $\boldsymbol{\eta}^T\boldsymbol{u}_i$ 잔차산점도에서 평균함수 $E(e_i^w|\boldsymbol{\eta}^T\boldsymbol{u}_i)$의 형태를 알 수 있고 이를 통해 $E(\epsilon_i|\boldsymbol{\eta}^T\boldsymbol{u}_i)$의 추정이 가능할 것이다. 따라서 이 잔차산점도로 모형의 적절성의 파악이 가능해진다.

3.3. 카이잔차산점도

로지스틱회귀모형에서 y_i의 분산은 $Var(y_i|\boldsymbol{x}_i) = m_i\theta(\boldsymbol{x}_i)(1-\theta(\boldsymbol{x}_i))$로 등분산이 아니므로 가중잔차를 고려해야 한다. 식 (3.4)에서와 같이 분산함수를 $var(y_i|\boldsymbol{x}_i) = \sigma^2/w_i$라 하고 $\sigma^2 = 1$로 하면 가중치는 $w_i = 1/[m_i\theta(\boldsymbol{x}_i)(1-\theta(\boldsymbol{x}_i))]$로 설정할 수 있다. 선형회귀모형에서와는 달리 이러한 가중치들은 모수들에 의존하기 때문에 가중치는 $\hat{w}_i = 1/[m_i\widehat{\theta}(\boldsymbol{x}_i)(1-\widehat{\theta}(\boldsymbol{x}_i))]$로 추정하여야 한다. 이렇게 추정된 가중치를 이용하면 식 (3.7)의 가중잔차 e_i^w는 다음과 같이 추정할 수 있다.

$$e_i^* = \frac{y_i - \widehat{y}_i}{\sqrt{m_i\widehat{\theta}(\boldsymbol{x}_i)(1-\widehat{\theta}(\boldsymbol{x}_i))}} \tag{3.8}$$

Cook과 Weisberg (1999)에 의하면 이러한 잔차 e_i^w를 카이잔차 (chi-residuals)라고 하는데 제곱해서 더하면 Pearson의 X^2통계량이 되기 때문이다.

카이잔차 e_i^*와 설명변수들의 선형결합인 $\boldsymbol{\eta}^T\boldsymbol{u}$를 두 축으로 하는 그래프를 카이잔차산점도 (chi-residual plot)라고 한다. 이 산점도는 평균함수 $E(e_i^*|\boldsymbol{\eta}^T\boldsymbol{u})$의 형태를 나타내고 있다고 할 수 있는데 이에 대한 매우 조심스런 해석이 요구된다. 만약 $\widehat{\boldsymbol{\eta}}^T\boldsymbol{u}$를 수평축으로, e_i^*를 수직축으로 갖는 카이잔차산점도에서 이 평균함수 $E(e_i^*|\boldsymbol{\eta}^T\boldsymbol{u})$가 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 아니더라고 작성할 수 있다는 장점이 있다.

4. 예제

Cook과 Weisberg (1994)에 제시된 오스트레일리아 스포츠 선수촌에서 훈련하는 운동선수 남녀 각각 102명과 100명의 신체지수와 혈액검사 자료에서 변수로는 성별 (Sex: 0=male, 1=female), 몸무게 (Wt), 키 (Ht), 적혈구수치 (RCC), 피부주름수치 (SSF) 등이 있다. 이 중 Sex를 반응변수로 사용하고, Wt와 RCC를 설명변수로 하는 모형을 생각하자. 모형의 적합결과인 Table 4.1에서 두 설명변수가 충분한 설명력이 있고 모형이 적절하다고 할 수 있다.

이 모형에 대한 카이잔차산점도 Figure 4.1에서 평균함수를 나타내는 평활곡선이 직선으로 나타나고 있어 모형을 적절성을 보여주고 있다. 한편 Figure 4.2는 Arc를 이용하여 작성한 산점도로 각각 적합값

$\widehat{\boldsymbol{\eta}}^T\boldsymbol{u}$과 설명변수 Wt, RCC각각에 대한 주변모형산점도이다. 모든 산점도에서 모형의 가정을 하지 않는 경우와 모형의 가정을 하는 경우의 추정값과 상한, 하한을 나타내는 평활곡선들이 거의 일치하므로 모형이 적절하다고 할 수 있다.

Table 4.1 Logistic Regression Summary with $\boldsymbol{u} = (Wt, RCC)^T$

Coefficients	Estimate	Std.Error	Est / SE	p-value
Intercept	-32.3526	4.41318	7.331	<.00001
Wt	-0.10834	0.02307	-4.696	<.00001
RCC	-5.15424	0.78450	-6.570	<.00001
Number of cases:				202
Degrees of freedom:				199
Pearson X^2:				208.850
Deviance:				124.289

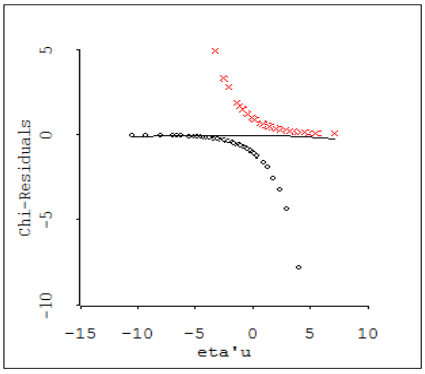

Figure 4.1 Chi-residual plot with $\boldsymbol{u} = (Wt, RCC)^T$

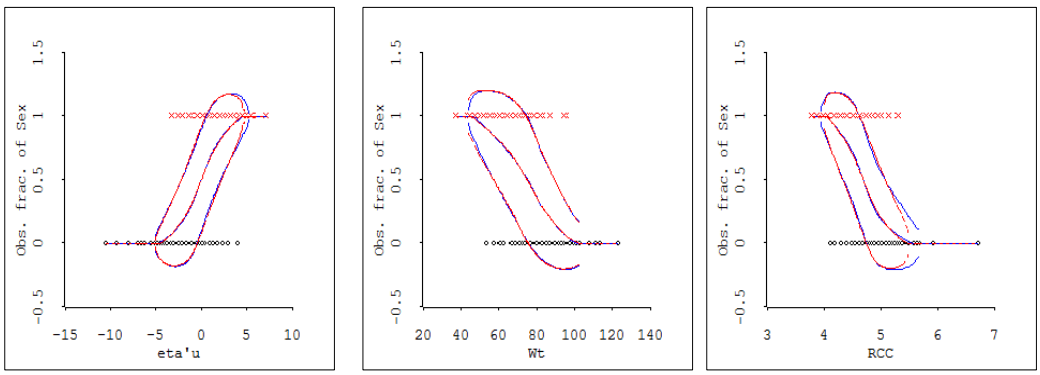

Figure 4.2 Marginal model plots for $\widehat{\boldsymbol{\eta}}^T\boldsymbol{u}$, Wt and RCC

이번에는 Ht와 SSF를 설명변수로 하는 모형을 생각해보자. 이 모형의 적합결과가 Table 4.2에 있다. 앞의 모형과 같이 두 개의 설명변수가 충분한 설명력이 있고 모형이 적절한 것으로 보인다. 하지만 Figure 4.3의 카이잔차산점도를 보면 평균함수를 나타내는 평활곡선의 가운데 부분이 조금 휘어져 있

다. 따라서 모형이 적절하지 않다고 할 수 있다. Figure 4.4에서 $\widehat{\boldsymbol{\eta}}^T\boldsymbol{u}$과 SSF의 주변모형산점도를 보면 3쌍의 평활곡선들이 일치하지 않으므로 역시 모형이 적절하다고 할 수 없다.

Table 4.2 Logistic Regression Summary with $\boldsymbol{u} = (Ht, SSF)^T$

Coefficients	Estimate	Std.Error	Est / SE	p-value
Intercept	42.5342	7.27904	5.843	<.0001
Ht	-0.268303	0.04336	-6.188	<.0001
SSF	0.087722	0.01457	6.022	<.0001
	Number of cases:			202
	Degrees of freedom:			199
	Pearson X^2:			245.185
	Deviance:			109.115

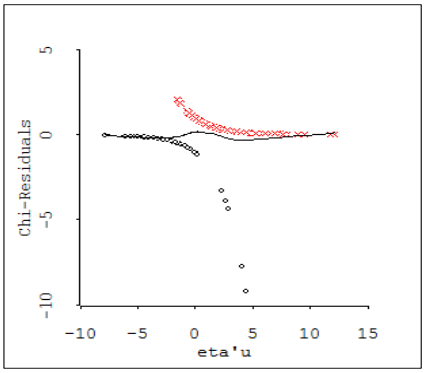

Figure 4.3 Chi-residual plot with $\boldsymbol{u} = (Ht, SSF)^T$

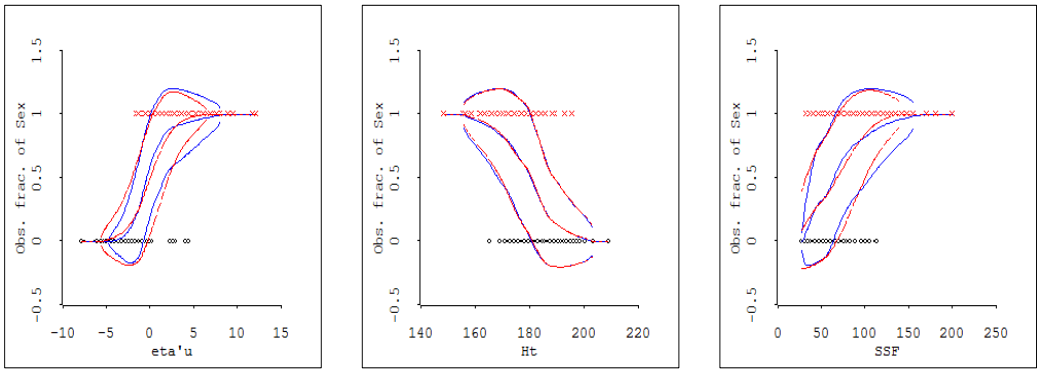

Figure 4.4 Marginal model plots for $\widehat{\boldsymbol{\eta}}^T\boldsymbol{u}$, Ht and SSF

Scrucca (2003)와 Kahng과 Shin (2012)에서 제안된 로그-밀도비와 그래픽적 방법을 근거로 하면 이 모형에 추가되어야할 변수로 $log(SSF)$이 선택된다. 따라서 이제 설명변수 Ht, SSF에 $log(SSF)$를 추가한 모형을 생각하자. 이 모형에 대한 카이잔차산점도 Figure 4.5에서는 Figure 4.3과는 달리 평균함수를 나타내는 평활곡선이 직선으로 나타나고 있어 모형이 적절함을 보여주고 있다. 또한 Figure 4.6의

주변모형산점도에서 $\hat{\boldsymbol{\eta}}^T\boldsymbol{u}$, Ht, SSF, $log(SSF)$의 모든 경우 모형의 가정을 여부에 따른 3쌍의 평활곡선들이 거의 일치하므로 모형이 적절하다고 할 수 있다.

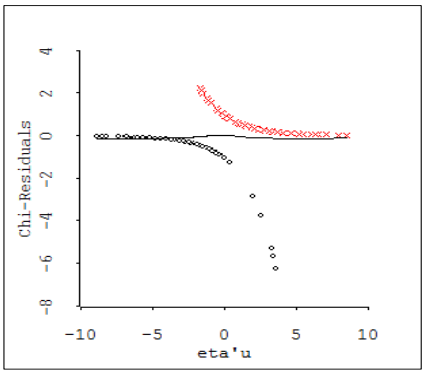

Figure 4.5 Chi-residual plot with $\boldsymbol{u} = (Ht, SSF, log(SSF))^T$

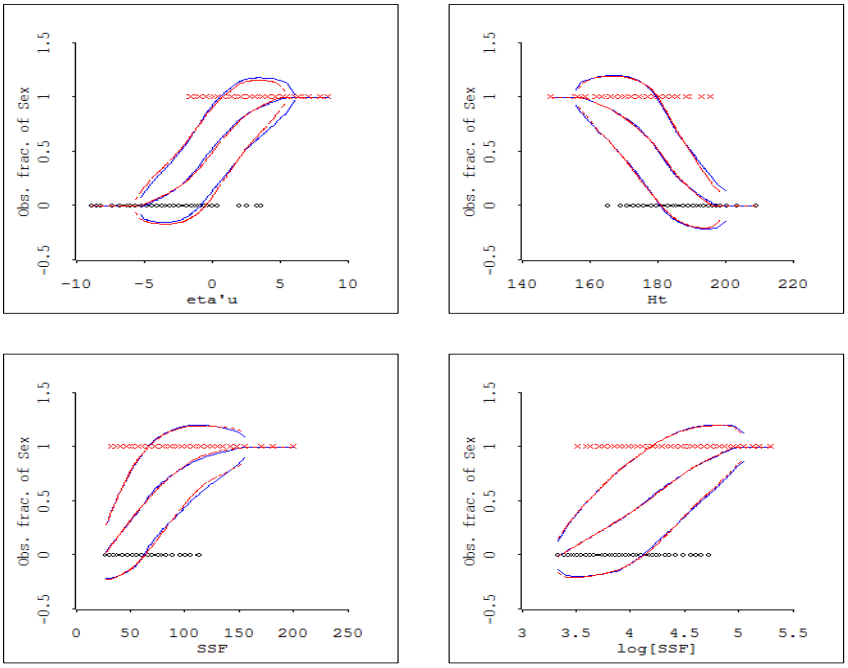

Figure 4.6 Marginal model plots for $\hat{\boldsymbol{\eta}}^T\boldsymbol{u}$, Ht, SF and $log(SSF)$

5. 결론

선형회귀분석에서 잔차산점도는 모형의 진단이나 적절성 평가를 위한 도구로 다양하게 이용되고 있다. 모형이 적절하다면 잔차산점도에 특별한 패턴이 없어야 한다. 그러나 로지스틱회귀나 이항회귀에

서 잔차산점도는 모형의 적절성과는 관계없이 특정한 패턴을 갖게 되어 모형평가의 도구로 적절하지 않다. 대안으로 제시된 주변모형산점도는 그 효용성은 인정되나 그래프 작성이 어렵고 판단이 복잡하고 모호할 수 있다는 단점이 있다. 본 논문에서는 등분산 가정을 할 수 없는 로지스틱회귀에서 카이잔차를 이용하여 잔차가 가지고 있는 문제점이 해결됨을 보이고 카이잔차산점도를 이용하는 모형의 적절성 평가방법을 제안하고 그 기능을 알아보았다. 그 결과 주변모형산점도보다 작성이 수월하고 판단도 간편한 카이잔차산점도가 로지스틱회귀에서 유용하게 사용될 수 있음을 확인하였다.

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

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.

Cook, R. D. (1998). *Regression graphics: Idea for studying regressions through graphics*, Wiley, 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*, 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 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 the Korean Data & Information Science Society*, **16**, 13-18.

Kahng, M. and Shin, E. (2012). A study on log-density with log-odds graph for variable selection in logistic regression. *Journal of Korean Data & Information Science Society*, **23**, 99-111.

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.

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

Journal of the Korean Data & Information Science Society
2015, **26**(1), 141–150

http://dx.doi.org/10.7465/jkdi.2015.26.1.141
한국데이터정보과학회지

Model assessment with residual plot in logistic regression[†]

Myung Wook Kahng[1]

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

Received 16 December 2014, revised 2 January 2015, accepted 10 January 2015

Abstract

Graphical paradigms for assessing the adequacy of models in logistic regression are discussed. The residual plot has been widely used as a graphical tool for evaluating the adequacy of the model. However, this approach works well only for linear models with constant variance, and the alternative approach, the marginal model plot, has its defects as well. We suggest a Chi-residual plot that overcomes the potential shortcomings of the marginal model plot.

Keywords: Binary regression, chi-residual plot, logistic regression model, marginal model plot, residual plot, weighted residual.

[†] This research was supported by the Sookmyung Women's University Research Grants (1-1303-0010).
[1] Professor, Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea.
E-mail: mwkahng@sm.ac.kr